DIFFERENTIATING THROUGH DELAY: STOCHASTIC CONTROL WITHOUT FULL HISTORY HESSIANS

Anonymous authors
Paper under double-blind review

ABSTRACT

In stochastic delay control, the current state re-enters future dynamics and costs through the history. The stochastic maximum principle represents these effects through anticipated first-order adjoints and, with control-dependent diffusion, a coupled second-order family carrying current–history cross-curvature. We recover the information needed for constrained control refinement directly from a learned history policy's simulated continuations. Future control values are detached under state differentiation, while all physical delay paths remain active. Current-injection duality identifies the required diffusion curvature with an $n \times n$ fixed-control Hessian block, eliminating full history-Hessian assembly and a separate numerical solution of the coupled adjoint family. The stored curvature is $O(n^2)$; trajectory propagation retains its history dependence. We establish exact discrete adjoint correspondence and convergence of the population recovery inputs under the stated refinement assumptions. Building on Pontryagin-guided direct policy optimization (PGDPO), an anchored local solve then refines the control, with localized error bounds and sufficient conditions for objective improvement. Across five benchmarks, refinement reduces its learned reference policy's control error by 24–91 percent where optimizer references exist and significantly reduces paired cost on two constrained applications. Delay-model and mesh diagnostics test the computation separately from these performance comparisons.

1 INTRODUCTION

Stochastic systems with memory arise in epidemic intervention, trading with transient impact, and control with delayed feedback. Their drift, diffusion, or cost depends on past states, so the current physical state generally does not close the dynamics. Stochastic delay differential equations provide a natural formulation (Mohammed, 1984), but optimal control then involves a state history whose discretized dimension grows as the time mesh is refined. We study state-delay control with known differentiable dynamics, observed history, and control-dependent diffusion.

Learning has made such problems accessible to computation. Recurrent networks represent history-dependent controls and optimize their expected performance directly (Han & Hu, 2021); neural methods also learn delayed forward–backward systems (Fouque & Zhang, 2020) and stabilize stochastic delay dynamics with safety guarantees (Zhang et al., 2023). However, a learned policy does not by itself provide the optimality information needed to refine its control at a particular history. Small shared-parameter gradients do not imply pointwise control optimality, and strictly interior outputs cannot attain active bounds. We call sharp changes in the control profile, including boundary entry and exit, kinks.

The remaining computation is substantial. A current-state perturbation affects future rewards both through direct propagation and through its later re-entry into delayed coefficients. The stochastic maximum principle (SMP) represents this dependence by anticipated first-order adjoints (Chen & Wu, 2010; Øksendal et al., 2011) and, with control-dependent diffusion, a coupled second-order BSDE–BSVIE family carrying current–history cross-curvature (Meng et al., 2025). Finite-buffer automatic differentiation yields discrete sensitivities, but does not by itself establish their correspondence with the continuous-time inputs required for control refinement. Our question is how to compute the inputs needed for constrained refinement from a learned history policy while retaining these delayed effects.

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.

We recover the delay-control information directly from simulated continuations of a learned history policy. Future control values are fixed under state differentiation, while all physical delay paths remain active. The learning-and-recovery architecture builds on PGDPO (Huh et al., 2025; Jeon et al., 2025; 2026). Delay introduces a distinct obstacle: second-order cross-curvature does not close in a single current-state adjoint equation. We resolve the required control comparison through current-injection duality, which identifies the effective diffusion curvature of the coupled delay family with a directly computable current-state Hessian block. Our contributions are:

- **Delay sensitivities without a separate adjoint solver.** Fixed-control differentiation and conditional estimation recover the first-order inputs and effective diffusion curvature from learned continuations, retaining all delay re-entry paths without training a value or adjoint network.
- **Current-state curvature with full-history propagation.** We establish exact discrete correspondence and population-input consistency as the history dimension grows. The curvature output uses $O(n^2)$ storage and n Hessian–vector queries; propagation and each query retain their horizon and history costs.
- **Constrained recovery and computational evidence.** We bound the recovered-control error and give sufficient conditions for objective improvement. Five benchmarks and separate delay-model, estimation, and mesh checks evaluate the method's accuracy and computational trade-offs.

2 Related work

Learning with delay dynamics. Neural DDEs use delayed adjoints to train continuous-depth models (Zhu et al., 2021). In stochastic control, recurrent policies directly optimize the objective (Han & Hu, 2021), while mean-field delay control admits both direct policy learning and neural forward–anticipated-backward solvers (Fouque & Zhang, 2020). SYNC addresses stabilization and safety (Zhang et al., 2023). We instead compute the conditional first- and second-order inputs needed to refine a learned controller under stochastic optimality conditions.

Optimality systems and neural solvers. Delay dynamic programming generally leads to path-space HJB equations (Larssen, 2002; Larssen & Risebro, 2003; Bauer & Rieder, 2005). The SMP instead uses anticipated adjoints and coupled second-order systems (Peng & Yang, 2009; Chen & Wu, 2010; Øksendal et al., 2011; Meng et al., 2025). Neural PDE and forward–backward methods approximate values and adjoints (Saporito & Zhang, 2021; Ji et al., 2022; Andersson et al., 2023; 2026). We evaluate fixed-control sensitivities rather than train another equation solver.

Policy refinement and other forms of delay. Stochastic value gradients differentiate stochastic dynamics (Heess et al., 2015), and delay-free PGDPO separates policy learning from adjoint-based control recovery (Huh et al., 2025; Jeon et al., 2025; Huh & Koo, 2026; Jeon et al., 2026). The new delay step identifies the required current-injection curvature of the coupled second-order family and establishes correspondence as the history dimension grows. The local solve is a Hamiltonian update related to successive approximations (Chernousko & Lyubushin, 1982; Kerimkulov et al., 2021); future controls come from the learned continuation rather than a new trajectory optimization as in model predictive control. Observation, action, and inference delays in reinforcement learning (Wu et al., 2024; 2025; Anokhin et al., 2025; Riemer et al., 2025) concern a different source of memory: here the state history enters the dynamics themselves, and control delay is outside scope.

3 Problem formulation

Stochastic delay dynamics. Let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ be a complete filtered probability space supporting a d-dimensional Brownian motion B, with $(\mathcal{F}_t)_{t\geq 0}$ the usual augmented Brownian filtration. We write $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_t]$. Fix a delay length $\delta > 0$ and an initial history $x_0 \in C([-\delta, 0]; \mathbb{R}^n)$. We consider an $\mathbb{R}^n$-valued controlled state process X governed by the SDDE

$$\begin{aligned} \mathrm{d}X(t) &= b\big(\Xi(t)\big)\,\mathrm{d}t + \sigma\big(\Xi(t)\big)\,\mathrm{d}B(t), \quad t \in [0, T], \\ X(t) &= x_0(t), \quad t \in [-\delta, 0]. \end{aligned} \tag{1}$$

Here $\Xi(t) := (t, X(t), Y(t), M(t), u(t))$ is the aggregate process, where u takes values in a nonempty Borel set $\mathcal{U} \subseteq \mathbb{R}^m$, *not necessarily convex*, and the delay variables $Y(t)$, $M(t)$ are defined below. The coefficients depend on the control only through its current value $u(t)$, so control delay is outside the present scope, while the diffusion may depend on the control throughout.

Delay variables and history. For $t \geq 0$, the state history over the memory window is the trajectory segment $X_t(r) := X(t + r)$, $r \in [-\delta, 0]$ (Mohammed, 1984). Delays enter through a *point delay* and an exponentially weighted *distributed delay*,

$$Y(t) := X_t(-\delta) = X(t - \delta), \quad M(t) := \int_{-\delta}^{0} e^{\rho r} X_t(r) \, \mathrm{d}r, \tag{2}$$

with decay parameter $\rho \geq 0$. The theory is stated for this exponential kernel, and other kernels need the corresponding Volterra assumptions.

Objective and admissible controls. Given running reward f and terminal reward g, the performance functional is

$$J(u) := \mathbb{E}\left[\int_0^T f\big(\Xi(t)\big) \, \mathrm{d}t + g\big(X(T)\big)\right], \tag{3}$$

and the admissible class $\mathcal{A}$ consists of the progressively measurable processes $u : [0, T] \times \Omega \to \mathcal{U}$ with $\mathbb{E} \int_0^T \|u(t)\|^2 \, \mathrm{d}t < \infty$. The problem is to attain $\sup_{u \in \mathcal{A}} J(u)$, and minimization problems follow by replacing (f, g) with $(-f, -g)$. Well-posedness of the adjoint systems, the mesh-consistency conditions, and the localization used for recovery are stated in Appendices C and D.

4 Delay optimality and the inputs needed for recovery

From delayed variations to recovery inputs. The computation must account for both state propagation and delayed re-entry, including the second-order effect of a control-induced change in diffusion. We specialize the spike-variation principle for stochastic delay systems (Meng et al., 2025) to Section 3, retaining this diffusion effect (Peng, 1990). This section identifies the inputs needed for a local control comparison; Section 5 computes them from learned continuations. The full BSDE–BSVIE systems are in Appendix C.

First-order Hamiltonian. With the notation of Section 3, define the first-order Hamiltonian

$$H(t, x, y, m, u, p, q) := f(t, x, y, m, u) + p^\top b(t, x, y, m, u) + \langle q, \, \sigma(t, x, y, m, u) \rangle_F, \tag{4}$$

where $\langle A, B \rangle_F := \operatorname{tr}(A^\top B)$. Along an optimal pair (X^*, u^*) with first-order adjoints (p, q) below, we abbreviate the evaluation tuple by $\Theta^*(t) := \big(\Xi^*(t), p_t, q_t\big) = \big(t, X^*(t), Y^*(t), M^*(t), u^*(t), p_t, q_t\big)$.

Adjoint systems. The first-order adjoint pair (p, q) solves the anticipated backward SDE

$$\begin{aligned} -\mathrm{d}p_t = \Big[& H_x\big(\Theta^*(t)\big) + \mathbf{1}_{\{t+\delta \leq T\}} \mathbb{E}_t[H_y\big(\Theta^*(t+\delta)\big)] \\ & + \mathbb{E}_t \int_t^{(t+\delta)\wedge T} e^{-\rho(s-t)} H_m\big(\Theta^*(s)\big) \, \mathrm{d}s \Big] \mathrm{d}t - q_t \, \mathrm{d}B_t, \qquad p_T = g_x\big(X^*(T)\big). \end{aligned} \tag{5}$$

The three drift terms account for the current state, its point-delay re-entry, and its distributed-memory re-entry. The adjoint p_t measures how a perturbation of the current state changes the conditional remaining reward under fixed-control differentiation, q_t is its martingale input, and control-dependence of the diffusion enters through the pairings $\langle q, \sigma_x \rangle_F$, $\langle q, \sigma_y \rangle_F$, $\langle q, \sigma_m \rangle_F$ inside H_x, H_y, H_m.

Equation (5) is a compact anticipated specialization of the first-order BSDE–BSVIE system of Meng et al. (2025) to terminal reward $g(X(T))$ and no control delay. For the second-order object, let $(\mathfrak{P}, \mathfrak{Q})$ denote the second-order adjoint family of Meng et al. (2025), a backward SDE coupled with backward stochastic Volterra integral equations and their martingale components. A single $n \times n$

backward equation does not close, because the second variation of a delay system carries current–history cross-curvature. We write $\Pi_t := \mathfrak{A}_t[\mathfrak{P}] \in \mathbb{S}^n$ for the *effective diffusion curvature* assembled from the $\mathfrak{P}$-family by the map recalled in Appendix C. Its role is to price the quadratic effect of a diffusion perturbation $\Delta\sigma_t(u) := \sigma(t, X^*(t), Y^*(t), M^*(t), u) - \sigma_t^*$, where $\sigma_t^* := \sigma(\Xi^*(t))$, in the maximum condition below, so that a change of the control is credited for its change of the noise exposure as well as for its change of the drift.

Effective curvature and generalized Hamiltonian. Define the shifted martingale input

$$\zeta_t := q_t - \Pi_t \, \sigma_t^* \ \in \ \mathbb{R}^{n\times d}, \tag{6}$$

which removes from q the part of the shock response explained by the curvature Π acting on the reference diffusion, and the generalized Hamiltonian

$$\begin{aligned}\mathcal{H}(t, x, y, m; u, p, \zeta, \Pi) :=& f(t, x, y, m, u) + p^\top b(t, x, y, m, u) + \langle \zeta, \, \sigma(t, x, y, m, u)\rangle_F \\ &+ \tfrac{1}{2}\,\mathrm{tr}\big[\sigma(t, x, y, m, u)^\top \, \Pi \, \sigma(t, x, y, m, u)\big].\end{aligned} \tag{7}$$

By construction, $\mathcal{H}(t; u, \cdot) - \mathcal{H}(t; u^*(t), \cdot) = H(t; u, \cdot) - H(t; u^*(t), \cdot) + \frac{1}{2}\,\mathrm{tr}[\Delta\sigma_t(u)^\top \Pi_t \, \Delta\sigma_t(u)]$, so $\mathcal{H}$ turns the first-order sensitivities and the diffusion curvature into a local comparison of candidate controls, with the shift (6) taken from the delay-free setting of Jeon et al. (2026). The same adjoint systems define $(p^{\mathrm{ref}}, q^{\mathrm{ref}}, \Pi^{\mathrm{ref}}, \zeta^{\mathrm{ref}})$ along any admissible reference control u^{ref}, and when the reference is optimal its tuple also satisfies the maximum condition below.

General maximum condition. Along an optimal pair, the stochastic maximum principle gives the necessary condition

$$u^*(t) \ \in \ \arg\max_{u\in\mathcal{U}} \ \mathcal{H}\big(t, X^*(t), Y^*(t), M^*(t); \, u, \, p_t, \, \zeta_t, \, \Pi_t\big), \quad \mathrm{d}t \otimes \mathrm{d}\mathbb{P}\text{-a.e.} \tag{8}$$

When $\mathcal{U}$ is convex, (8) implies the local variational inequality

$$\big\langle H_u\big(\Theta^*(t)\big), \, u - u^*(t)\big\rangle \le 0 \quad \forall u \in \mathcal{U}, \quad (H_u)_\ell = f_{u_\ell} + p_t^\top b_{u_\ell} + \langle q_t, \, \sigma_{u_\ell}\rangle_F, \quad \ell = 1, \dots, m, \tag{9}$$

so that H_u collects the immediate, drift-mediated, and diffusion-mediated marginal effects of the control, with $H_u(\Theta^*(t)) = 0$ at interior optima, consistently with (8) because $\nabla_u \mathcal{H}(u^*(t)) = \nabla_u H(u^*(t))$ by (6).

5 Computing delay adjoints from learned continuations

Two direct computational routes are a numerical solution of the coupled delay-adjoint family or naive AD of the full fixed-control buffer Hessian. The latter is the dense assembly in Remark C.3's storage/query comparison. We instead extract the required current block on the full delay graph (Figure 1), estimate the conditional inputs, and solve the anchored control problem. Algorithm 1 gives pseudocode.

Refining history graphs. For a mesh size $\Delta t \equiv h > 0$, let $t_k^h = kh$, with $N_h = T/h$ and $H_h = \delta/h$ integers along a refining sequence. The simulator state is the buffer $\mathsf{S}_k^h = (X_k^h, X_{k-1}^h, \dots, X_{k-H_h}^h) \in \mathbb{R}^{n(H_h+1)}$, from which $Y_k^h = X_{k-H_h}^h$ and $M_k^h = \sum_{j=0}^{H_h} w_j^h X_{k-j}^h$ are read with weights (w_j^h) that consistently discretize (2), and the superscript h is suppressed below.

Stage I: learning a reference policy. We parameterize a history-dependent control by $u_k^\theta = \pi_\theta(t_k, \mathsf{S}_k)$, with an LSTM as the default history encoder. The warm start optimizes $\widehat{\mathbb{E}}[J(u^\theta)]$ over θ by ordinary closed-loop BPTT and keeps the checkpoint $\widehat{\theta}$ with the best validation cost, after which the actor is frozen and $u^{\mathrm{ref}} = \pi_{\widehat{\theta}}$ serves as the fixed reference, or anchor, control for Stage II.

Stage II-A: differentiating the delayed continuation. At decision time t_k, we fix the observed buffer S_k and replicate it across N_{mc} future branches: all branches share the same current state and history, and only their future Brownian increments differ. Along branch m, future reference controls are generated by the frozen actor but detached under state differentiation, $u_\ell^{\mathrm{ref},(m)} = \mathrm{sg}\big(\pi_{\widehat{\theta}}(t_\ell, \mathsf{S}_\ell^{(m)})\big)$

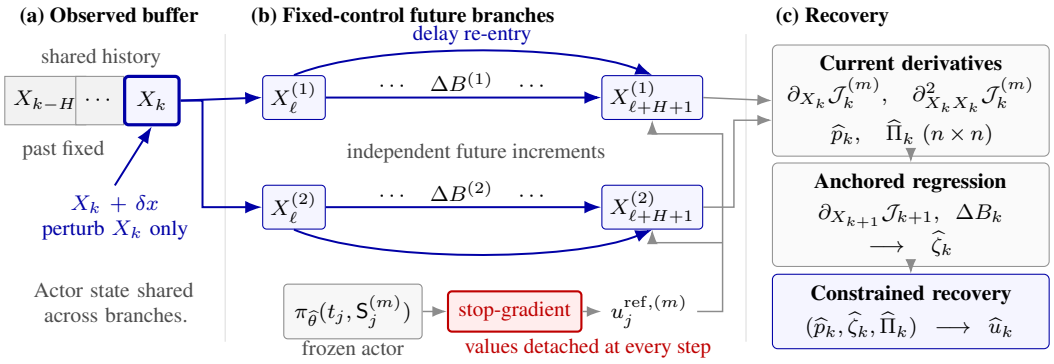


Figure 1: **Full-history propagation, current-state extraction.** Branches share the observed buffer and differ in future noise. Future controls are generated by the frozen actor and detached only under state differentiation; physical delay re-entry remains active. Differentiating with respect to the current state produces the needed curvature block without constructing the full history Hessian. Conditional averaging and anchored regression supply the inputs for constrained recovery.

for $\ell \geq k$, where sg denotes stop-gradient, while physical state propagation and all point- and distributed-delay re-entry paths remain differentiable. Write $\mathbb{E}_k[\cdot]$ for the expectation over future Brownian increments given the frozen buffer and actor state shared by the branches, as formalized in Assumption C.9, and D^{ol} for differentiation on the detached graph with respect to the physical buffer only, which is open-loop backpropagation through time (OL-BPTT) (Werbos, 1990). Writing $\mathcal{J}_k^{(m)} = \sum_{\ell=k}^{N-1} f_\ell^{(m)} \Delta t + g(X_N^{(m)})$ for the branch continuation payoff, the pathwise sensitivities are

$$\widetilde{\Lambda}_k^{(m)} := D^{\mathrm{ol}}_{\mathsf{S}_k} \mathcal{J}_k^{(m)}, \quad \widetilde{\mathbf{P}}_k^{(m)} := D^{2,\mathrm{ol}}_{\mathsf{S}_k \mathsf{S}_k} \mathcal{J}_k^{(m)}. \tag{10}$$

With $\iota_0 : \mathbb{R}^n \to \mathbb{R}^{n(H+1)}$ the current-coordinate insertion, the reference-indexed population targets are

$$\begin{aligned} p_k^{\mathrm{ref}} &:= \iota_0^\top \mathbb{E}_k[\widetilde{\Lambda}_k], & \Pi_k^{\mathrm{ref}} &:= \iota_0^\top \mathbb{E}_k[\widetilde{\mathbf{P}}_k] \iota_0, \\ q_k^{\mathrm{ref}} &:= \frac{1}{\Delta t} \mathbb{E}_k[(\iota_0^\top \widetilde{\Lambda}_{k+1}) \Delta B_k^\top], & \zeta_k^{\mathrm{ref}} &:= q_k^{\mathrm{ref}} - \Pi_k^{\mathrm{ref}} \sigma_k^{\mathrm{ref}}, \end{aligned} \tag{11}$$

where $\sigma_k^{\mathrm{ref}} = \sigma(t_k, X_k, Y_k, M_k, u_k^{\mathrm{ref}})$. For computation, define $F_{k,m}(v) := \mathcal{J}_k^{(m)}(\mathsf{S}_k + \iota_0 v)$, $v \in \mathbb{R}^n$, holding each branch's noise and realized future reference controls fixed at their $v = 0$ values. Before input projection,

$$\widehat{\Pi}_k^{\mathrm{ref}} = \frac{1}{N_{\mathrm{mc}}} \sum_{m=1}^{N_{\mathrm{mc}}} \nabla_v^2 F_{k,m}(0).$$

This is fixed-control continuation curvature, not the closed-loop policy-value Hessian. It equals the current block of the full-buffer estimator on the same branches (Corollary C.2), while every physical delay path remains active. Branch averages similarly estimate p; nested antithetic common-random-number regression estimates ζ with plug-in anchor $\widehat{\Pi}_k^{\mathrm{ref}} \sigma_k^{\mathrm{ref}}$ (Jeon et al., 2026) (Appendix C.9).

Stage II-B: constrained control refinement. Given the raw conditional estimates $(\widehat{p}_k^{\mathrm{ref}}, \widehat{\zeta}_k^{\mathrm{ref}}, \widehat{\Pi}_k^{\mathrm{ref}})$, Stage II-B first reconstructs the anchored Brownian input at the reference anchor and projects the adjoint inputs blockwise,

$$\widehat{q}_k^{\mathrm{anc}} := \widehat{\zeta}_k^{\mathrm{ref}} + \widehat{\Pi}_k^{\mathrm{ref}} \, \sigma_k^{\mathrm{ref}}, \qquad x^{\mathcal{N}} := \mathrm{Proj}_{\mathcal{N}_x}(x), \quad x \in \{p, q, \Pi\},$$

where the analysis assumes fixed compact convex sets satisfying Assumption D.5, with containment of the optimal inputs and uniform strong concavity, and the numerical ranges and the ablations without input clipping are in Appendices A and B.3. We symmetrize $\widehat{\Pi}$ before projection. Reconstructing $\widehat{q}^{\mathrm{anc}}$ *before* projecting keeps the curvature plug-in error within the first-order estimation layer; the blockwise projections are nonexpansive. Appendix D.1 gives the full construction.

With the solver coordinate $\widehat{\zeta}_k^{\mathcal{N}} := \widehat{q}_k^{\mathrm{anc},\mathcal{N}} - \widehat{\Pi}_k^{\mathcal{N}}\sigma_k^{\mathrm{ref}}$ and the projected objective $\mathcal{H}_k^{\mathcal{N}}(u) := \mathcal{H}\big(t_k, X_k, Y_k, M_k;\, u, \widehat{p}_k^{\mathcal{N}}, \widehat{\zeta}_k^{\mathcal{N}}, \widehat{\Pi}_k^{\mathcal{N}}\big)$, the recovery map is

$$\widetilde{u}_k := \arg\max_{u\in\mathcal{U}_{\mathrm{rec}}} \mathcal{H}_k^{\mathcal{N}}(u), \tag{12}$$

over a closed convex recovery set $\mathcal{U}_{\mathrm{rec}} \subseteq \mathcal{U}$ containing u_k^{ref}. PGDPO returns the feasible control $\widehat{u}_k \in \mathcal{U}_{\mathrm{rec}}$ approximating $\widetilde{u}_k$, warm-started at u_k^{ref}. For the returned control $\widehat{u}_k$, define the stationarity residual

$$r_{\mathrm{num},k} := \mathrm{dist}\big(0,\, -D_u\mathcal{H}_k^{\mathcal{N}}(\widehat{u}_k) + N_{\mathcal{U}_{\mathrm{rec}}}(\widehat{u}_k)\big),\ N_{\mathcal{U}_{\mathrm{rec}}}(u) := \{\nu : \langle \nu, v-u\rangle \le 0\ \forall v \in \mathcal{U}_{\mathrm{rec}}\}, \tag{13}$$

so that $r_{\mathrm{num},k}$ tests whether the deployed control is stationary for the supplied local objective, equal to the gradient norm at interior points and to the norm of the projection of the gradient onto the tangent cone at the boundary. Its mesh aggregate $\delta_{\mathrm{num}}^h := \big(h\sum_k \mathbb{E}\, r_{\mathrm{num},k}^2\big)^{1/2} = \|r_{\mathrm{num}}^h\|_{\mathbb{H}^2}$, with $\|Z\|_{\mathbb{H}^2} := (\mathbb{E}\int_0^T \|Z_t\|_F^2\,\mathrm{d}t)^{1/2}$ the $\mathbb{H}^2$ norm used throughout, is the residual layer of Proposition 5.2.

Computational correspondence and growing-history consistency. Stage II-A differentiates a mesh simulator rather than solving (5), so Theorem 5.1 verifies that what it computes is the adjoint of the buffer dynamics on that mesh and that the extracted inputs converge to those of Section 4 as the mesh is refined.

Theorem 5.1 (OL-BPTT correspondence and consistency of Pontryagin recovery inputs)**.** *Let Assumptions C.4, C.8, and C.9 hold, with the part-specific additional conditions stated in Appendix C.1, and let* $(p^{h,\mathrm{ref}}, q^{h,\mathrm{ref}}, \Pi^{h,\mathrm{ref}})$ *and* $\zeta^{h,\mathrm{ref}} = q^{h,\mathrm{ref}} - \Pi^{h,\mathrm{ref}}\sigma^{h,\mathrm{ref}}$ *be the recovery inputs extracted from the fixed-control OL-BPTT sensitivities along* $u^{h,\mathrm{ref}}$.

(a) ***Exact gridwise correspondence.*** *On every fixed mesh, and without any refinement assumption, the OL-BPTT sensitivities satisfy the exact discrete first- and second-order reverse recursions of Lemma C.10, and by Lemma C.11 their adapted projections are the unique square-integrable solution of the discrete buffer-adjoint system.*

(b) ***Consistency under growing history.*** *As* $h \downarrow 0$*, so that* $H_h = \delta/h \uparrow \infty$*,*

$$p^{h,\mathrm{ref}} \to p^{\mathrm{ref}}, \qquad q^{h,\mathrm{ref}} \to q^{\mathrm{ref}}, \qquad \Pi^{h,\mathrm{ref}} \to \Pi^{\mathrm{ref}}, \qquad \zeta^{h,\mathrm{ref}} \to \zeta^{\mathrm{ref}} \quad \text{in } \mathbb{H}^2,$$

and the anchored generalized-Hamiltonian difference $\mathcal{H}(u) - \mathcal{H}(u^{\mathrm{ref}})$ *of Appendix C.7 converges uniformly on compact control sets in integrated* L^1*. Here* $(p^{\mathrm{ref}}, q^{\mathrm{ref}})$ *solves the anticipated adjoint system* (5) *and* Π^{ref} *is the effective diffusion curvature identified from the second-order BSDE–BSVIE family through the current-injection duality of Lemma C.7.*

Part (a) gives exact discrete adjoint correspondence. The fixed-mesh current-block identity above is the chain rule; the substantive identification is that its population limit supplies the coupled delay family's effective diffusion curvature through current-injection duality, with growing-history consistency in part (b). No optimality of u^{ref} is used, and finite-branch and regression errors are treated in Propositions C.25 and C.26. Remark C.3 distinguishes the $O(n^2)$ stored curvature and n current-direction Hessian–vector queries from full-buffer assembly in dimension $n(H_h+1)$; propagation and each query retain their history and horizon costs.

Recovery accuracy from approximate delay inputs. Write $\chi = (X, Y, M)$ for the current state with its two delay channels, $\widehat{\mathcal{H}}_t$ for the projected objective of (12) built from the supplied inputs with $\widehat{\zeta}^{\mathcal{N}}$ anchored at u_t^{ref}, and $\mathcal{H}_t^*$ for the same objective built from the optimal tuple with $\zeta^* = q^* - \Pi^*\sigma(\chi^*, u^*)$. By (6), $D_w\widehat{\mathcal{H}}_t(u_t^{\mathrm{ref}}) = D_w H(\widehat{\chi}_t, u_t^{\mathrm{ref}}; \widehat{p}_t^{\mathcal{N}}, \widehat{q}_t^{\mathrm{anc},\mathcal{N}})$: the curvature enters only through a quadratic term that is stationary at the anchor, which is why the solver works in (p, ζ, Π) while the error analysis is carried out in (p, q). Two constants recur below: μ is a strong-concavity constant in w of the objective on $\mathcal{U}_{\mathrm{rec}}$ common to all $(p, q, \Pi) \in \mathcal{N}_p \times \mathcal{N}_q \times \mathcal{N}_\Pi$ and anchors in the localized set K_0 of Assumption D.1, and κ_{qg} is the quadratic-growth constant of Assumption D.6, $J(u^*) - J(u) \ge \kappa_{\mathrm{qg}}\|u - u^*\|_{\mathbb{H}^2}^2$. Stage I supplies an imperfect reference, and its inputs reach the solver after estimation, discretization, and an inexact solve. By Proposition D.17, the reference performance gap $\varepsilon_{\mathrm{ref}} := J(u^*) - J(u^{\mathrm{ref}})$ transfers to the first-order inputs at rate $(\varepsilon_{\mathrm{ref}}/\kappa_{\mathrm{qg}})^{1/2}$ and to the curvature and shifted input only at rate $(\varepsilon_{\mathrm{ref}}/\kappa_{\mathrm{qg}})^{1/4}$. The remaining layers

compare the deployed rollout, with state $\widehat{\chi}$ and harvested inputs $\widehat{p} = \widehat{p}^{\mathrm{ref}}$ and $\widehat{q}^{\mathrm{anc}}$, with the mesh reference $x^{h,\mathrm{ref}}$ and the continuous reference x^{ref} in $\mathbb{H}^2$,

$$\delta_{\mathrm{est}} := \sum_{x \in \{\chi, p, q^{\mathrm{anc}}\}} \|\widehat{x} - x^{h,\mathrm{ref}}\|, \quad \delta_{\mathrm{disc}} := \sum_{x \in \{\chi, p, q, u\}} \|x^{h,\mathrm{ref}} - x^{\mathrm{ref}}\|, \quad \delta_{\mathrm{num}} := \|r_{\mathrm{num}}\|, \tag{14}$$

where the population counterpart of $\widehat{q}^{\mathrm{anc}}$ is $q^{h,\mathrm{ref}}$. Thus estimation/deployment, discretization, and the local solve form separate layers.

Proposition 5.2 (Recovered-policy error budget)**.** *Under the hypotheses of Proposition D.18, including localized uniformly strongly concave recovery and quadratic growth, the deployed control anchored at $u^{h,\mathrm{ref}}$ satisfies*

$$\|\widehat{u} - u^*\|_{\mathbb{H}^2} \le \mu^{-1} C \Big[\delta_{\mathrm{est}} + \delta_{\mathrm{disc}} + (\varepsilon_{\mathrm{ref}}/\kappa_{\mathrm{qg}})^{1/2}\Big] + \mu^{-1}\delta_{\mathrm{num}}.$$

The constant is uniform over admissible curvature blocks in the common strongly concave class, with no separate curvature-mismatch term.

Anchoring makes the curvature correction stationary at the reference control, so its slower input-error rate does not enter as a separate policy-error term. The full statement also bounds the true-input local Hamiltonian loss (Appendix D.4); this is distinct from the full objective gap. Exact-anchor invariance and the supplied-objective residual certificate are in Corollary D.16. Improvement of J requires the margin below.

Theorem 5.3 (Conditional improvement of the corrected control)**.** *Let $\widehat{u}$ be the deployed Stage-II control anchored at u^{ref} and applied undamped, rather than through the damped update of the method of successive approximations, and suppose that the supplied objectives are μ-strongly concave as above and that the comparison-segment conditions of Lemma D.21 hold with quadratic-remainder constant C_J. Write $D := \|\widehat{u} - u^{\mathrm{ref}}\|_{\mathbb{H}^2}$ for the size of the correction and $\eta_\nabla := \|D_w \widehat{\mathcal{H}}(u^{\mathrm{ref}}) - g^{\mathrm{ref}}\|_{\mathbb{H}^2}$ for the gradient defect, where $g_t^{\mathrm{ref}} := D_w H(\chi_t^{\mathrm{ref}}, u_t^{\mathrm{ref}}; p_t^{\mathrm{ref}}, q_t^{\mathrm{ref}})$ is the exact reference gradient. Then*

$$J(\widehat{u}) - J(u^{\mathrm{ref}}) \;\ge\; (\mu - C_J)\, D^2 - (\eta_\nabla + \delta_{\mathrm{num}})\, D,$$

so the corrected control strictly improves J whenever $D > 0$, $\mu > C_J$, and $\eta_\nabla + \delta_{\mathrm{num}} < (\mu - C_J)D$. Along the reference trajectory, with the matching reference anchor, exact unclipped first-order inputs, and an exact solve, $\eta_\nabla = \delta_{\mathrm{num}} = 0$ for every admissible curvature block.

Proof. See Appendix D.5. □

The margin $\mu > C_J$ and defect condition distinguish a local control improvement from improvement of J; a good warm start alone is insufficient. These guarantees concern the localized model, and recursive deployment additionally requires Assumption D.7. Appendix C maps the claims to their assumptions and proofs.

Finite-step implementation. The experiments use the mesh block (16): drift pairs with the next-step conditional adjoint $\bar{p}_{k+1|k}$, and an h-scaled drift-curvature correction accounts for the finite step. Appendix A specifies this realization and retains its difference from the continuous block in the residual budget. All five control blocks are quadratic and admit a linear or box solve.

6 Numerical evaluation of delay-control recovery

Benchmarks. We use five stochastic state-delay control problems with control-dependent diffusion, specified in Appendix A. **P1** is a point-delay LQ tracking problem with an exact discrete-time optimizer. **P2** is a distributed-delay network LQ problem in dimension $n \in \{1, 5, 20\}$ with an exact Riccati reference obtained by modal decomposition. **P3** is a nonlinear epidemic problem with box control $[0, 1]$ and exponentially fading memory, whose learned methods see the raw infection history and the initial memory while the HJB reference uses the exact two-state lift: it tests nonlinear constrained recovery, while P1 and P2 test computation on fixed-window delay buffers. **P4** is a signal-aware optimal-execution problem with a one-sided box. **P5** is a torque-limited speed-tracking servo with nonlinear friction and delayed feedback, whose bounded correction torque enters both the drift and the actuator noise.

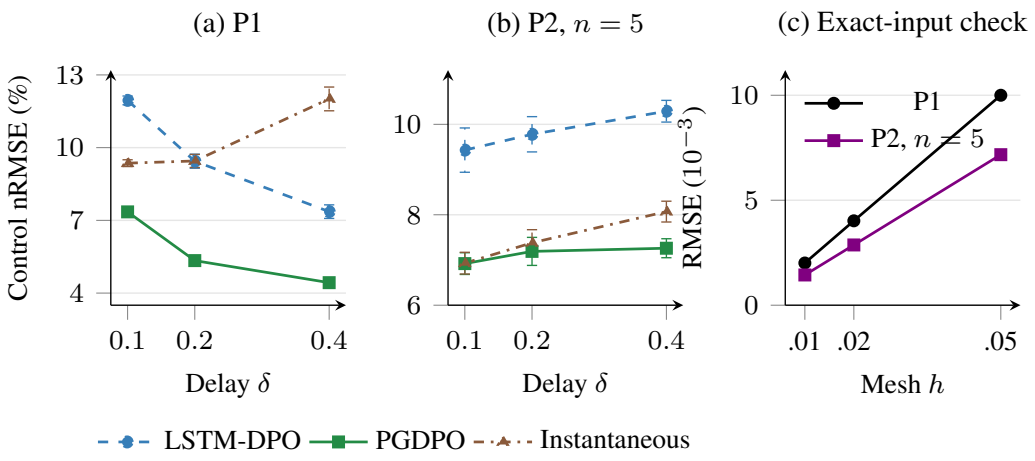


Figure 2: **Testing the delay computation.** Left and center: control nRMSE against same-mesh references, five-seed mean ± SD, as physical delay increases (Table 14). The instantaneous-memory variant changes both the continuation model and its derivatives. Right: exact-input local recovery RMSE against the exact same-mesh continuation minimizer (Table 12); this removes learning and estimation error and is not a test of full-pipeline convergence.

Methods. **LSTM-DPO** is the Stage-I actor alone and **PGDPO** the same actor followed by the Stage-II recovery of Section 5. **PDGM** learns the delay value functional from a path-dependent PDE residual (Saporito & Zhang, 2021), and **DeepSMP** learns the delay adjoint pair by multiple shooting of the anticipated BSDE with an LSTM history encoder (Ji et al., 2022; Andersson et al., 2023). Appendix A gives variants, budgets, hardware, and the mesh realization of Stage II, and Appendix B the sweeps and ablations.

What the computation checks establish. In Figure 2, retaining the delayed continuation improves control accuracy over the instantaneous-memory variant in five of six settings at the reported significance levels. At $\delta = 0.4$ in P1, omitting memory even increases error above the learned actor. Their paired cost differences are not statistically resolved, and the comparison diagnoses the continuation model and its derivatives together, not individual derivative paths. At $\delta = 0.2$ in P1 and in P2 at $n = 5$, the joint tests resolve neither a control-error nor a cost difference between full and re-entry-detached recovery, and Appendix B.3 compares the alignment and magnitude of the two corrections at shared initial histories. With exact inputs the local recovery discrepancy decreases with the mesh size, and Appendix B checks estimation and common-fine-mesh replay of the learned pipeline. Stage II costs an amortized 0.05–0.45 seconds per decision over the batched evaluation (Table 2); current-block extraction reduces curvature assembly, not the full continuation cost (Appendix B.4), while at nearly

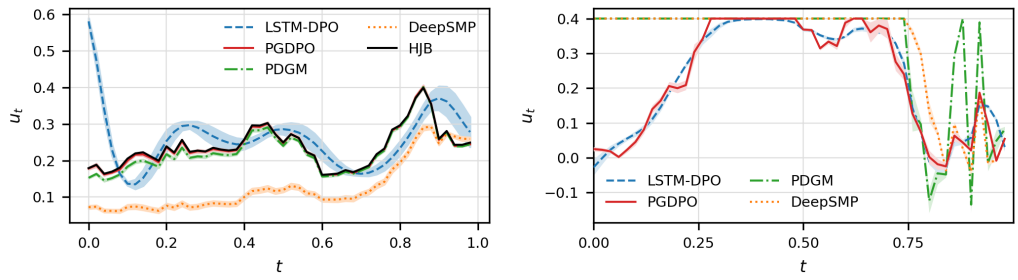


Figure 3: **Selected constrained-control paths**, five-seed mean ± SD on one evaluation path per problem. Left: P3 intervention, tracking local variations of the HJB control in black. Right: P5 correction torque, entering and leaving its bound without an optimizer reference.

Table 1: **Main results**, with (n, m) the physical state and control dimensions. Control nRMSE is the relative L^2 error against the optimizer reference, J is realized cost, and ΔJ is the paired realized-cost gap to the reference in Panel A and to LSTM-DPO in Panel B. Entries are five-seed means $\pm$ the SD of seed-level estimates, and the two-sided p-values test $\Delta J = 0$ using both seed and paired path-sampling variance (Appendix A), so the displayed SD alone does not determine p. The negative P2 gap to the exact reference is not resolved ($p = 0.67$), and finite-path comparisons do not establish an improvement over the optimum in expected cost. P2 at $n \in \{1, 5\}$ is in Appendix A.2.

Panel A: reference problems

Problem	(n, m)	Algorithm	Control nRMSE	ΔJ vs reference	p
P1	$(1, 1)$	PGDPO	$(5.340 \pm 0.175) \times 10^{-2}$	$(1.202 \pm 0.385) \times 10^{-3}$	0.17
		LSTM-DPO	$(9.435 \pm 0.292) \times 10^{-2}$	$(2.816 \pm 0.663) \times 10^{-3}$	0.089
		DeepSMP	$(5.704 \pm 0.094) \times 10^{-1}$	$(6.536 \pm 0.234) \times 10^{-2}$	< 0.001
		PDGM	$(1.521 \pm 0.113) \times 10^{-1}$	$(3.069 \pm 0.423) \times 10^{-3}$	0.009
P2	$(20, 20)$	PGDPO	$(1.189 \pm 0.015) \times 10^{-1}$	$(-6.741 \pm 1.570) \times 10^{-3}$	0.67
		LSTM-DPO	$(1.567 \pm 0.025) \times 10^{-1}$	$(1.006 \pm 0.237) \times 10^{-2}$	0.58
		DeepSMP	$(6.042 \pm 0.092) \times 10^{-1}$	$(5.085 \pm 0.299) \times 10^{-1}$	< 0.001
		PDGM	$(9.208 \pm 0.617) \times 10^{-2}$	$(1.520 \pm 0.113) \times 10^{-2}$	0.022
P3	$(1, 1)$	PGDPO	$(6.440 \pm 0.421) \times 10^{-3}$	$(2.246 \pm 3.229) \times 10^{-5}$	0.63
		LSTM-DPO	$(7.058 \pm 0.433) \times 10^{-2}$	$(5.809 \pm 1.503) \times 10^{-4}$	0.28
		DeepSMP	$(2.489 \pm 0.119) \times 10^{-1}$	$(4.247 \pm 0.872) \times 10^{-3}$	0.054
		PDGM	$(2.776 \pm 0.054) \times 10^{-2}$	$(1.015 \pm 0.183) \times 10^{-4}$	0.56

Panel B: constrained problems without an optimizer reference

Problem	(n, m)	Algorithm	Cost J	ΔJ vs LSTM-DPO	p
P4	$(2, 1)$	PGDPO	$(1.0515 \pm 0.0007) \times 10^{-1}$	$(-2.493 \pm 0.567) \times 10^{-4}$	0.001
		LSTM-DPO	$(1.0540 \pm 0.0002) \times 10^{-1}$	—	—
		DeepSMP	$(1.1494 \pm 0.0012) \times 10^{-1}$	$(+9.541 \pm 0.127) \times 10^{-3}$	< 0.001
		PDGM	$(1.1266 \pm 0.0129) \times 10^{-1}$	$(+7.263 \pm 1.290) \times 10^{-3}$	< 0.001
P5	$(1, 1)$	PGDPO	$(8.068 \pm 0.026) \times 10^{-3}$	$(-6.635 \pm 1.069) \times 10^{-4}$	< 0.001
		LSTM-DPO	$(8.731 \pm 0.092) \times 10^{-3}$	—	—
		DeepSMP	$(1.686 \pm 0.005) \times 10^{-2}$	$(+8.132 \pm 0.123) \times 10^{-3}$	< 0.001
		PDGM	$(1.797 \pm 0.018) \times 10^{-2}$	$(+9.242 \pm 0.190) \times 10^{-3}$	< 0.001

matched forward-simulation work a one-round derivative-free current-control search has higher control error in P1 and in P2 at $n = 5$ (Appendix B.5).

Control accuracy and cost. PGDPO reduces its actor's control error on all reference problems in Table 1; its reported cost gaps to the optimizer references are not significant at the 5% level. Control error and cost need not rank methods identically: PDGM has the lowest control error in P2 at $n = 20$ but a significant cost gap. Direct paired comparisons to the actor detect cost improvements in P2 at $n = 5, 20$ (Appendix A.2) and in both constrained applications, P4 and P5.

Figure 3 shows local tracking of the HJB control on the selected P3 path and entry to and exit from the torque bound in P5, where no optimizer reference is available, and Figure 4 of Appendix A gives the complete profiles.

7 Conclusion

Learned history continuations make the sensitivities needed for stochastic delay-control refinement computable by fixed-control differentiation. Current-state extraction avoids full history-Hessian assembly, with exact discrete correspondence and population-input consistency under the stated assumptions. Anchored recovery connects these inputs to constrained controls, error bounds, and conditional improvement guarantees, with empirical gains across five benchmarks.

Limitations and future work. Known differentiable dynamics and observed history are required; control delay is outside scope. Guarantees are local and conditional, and model errors can distort sensitivities. Longer horizons, learned dynamics, and cheaper online recovery remain open.

REFERENCES

Mihai Alexe and Adrian Sandu. On the discrete adjoints of adaptive time stepping algorithms. *Journal of Computational and Applied Mathematics*, 233(4):1005–1020, 2009. doi: 10.1016/j.cam.2009.08.109.

Kristoffer Andersson, Adam Andersson, and Cornelis W Oosterlee. Convergence of a robust deep fbsde method for stochastic control. *SIAM Journal on Scientific Computing*, 45(1):A226–A255, 2023. doi: 10.1137/22M1478057.

Kristoffer Andersson, Adam Andersson, and Cornelis W Oosterlee. The deep multi-FBSDE method: A robust deep learning method for coupled FBSDEs. *Journal of Scientific Computing*, 106(3):77, 2026. doi: 10.1007/s10915-026-03202-1.

Ivan Anokhin, Rishav Rishav, Matt Riemer, Stephen Chung, Irina Rish, and Samira Ebrahimi Kahou. Handling delay in real-time reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2025.

Harald Bauer and Ulrich Rieder. Stochastic control problems with delay. *Mathematical Methods of Operations Research*, 62(3):411–427, 2005. doi: 10.1007/s00186-005-0042-4.

Li Chen and Zhongfei Wu. Maximum principle for the stochastic optimal control problem with delay and application. *Automatica*, 46(6):1074–1080, 2010. doi: 10.1016/j.automatica.2010.03.005.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Felix L Chernousko and AA Lyubushin. Method of successive approximations for solution of optimal control problems. *Optimal Control Applications and Methods*, 3(2):101–114, 1982. doi: 10.1002/oca.4660030201.

Jean-Pierre Fouque and Zhaoyu Zhang. Deep learning methods for mean field control problems with delay. *Frontiers in Applied Mathematics and Statistics*, 6:11, 2020. doi: 10.3389/fams.2020.00011.

Andreas Griewank and Andrea Walther. *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*. SIAM, Philadelphia, PA, 2 edition, 2008.

Jiequn Han and Ruimeng Hu. Recurrent neural networks for stochastic control problems with delay. *Mathematics of Control, Signals, and Systems*, 33:775–795, 2021. doi: 10.1007/s00498-021-00300-3.

Nicolas Heess, Gregory Wayne, David Silver, Timothy Lillicrap, Tom Erez, and Yuval Tassa. Learning continuous control policies by stochastic value gradients. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL `https://proceedings.neurips.cc/paper_files/paper/2015/file/148510031349642de5ca0c544f31b2ef-Paper.pdf`.

Jeonggyu Huh and Hyeng Keun Koo. Breaking the dimensional barrier: Dynamic portfolio choice with parameter uncertainty via Pontryagin projection. arXiv preprint arXiv:2601.03175, 2026.

Jeonggyu Huh, Jaegi Jeon, Hyeng Keun Koo, and Byung Hwa Lim. Breaking the dimensional barrier: A Pontryagin-guided direct policy optimization for continuous-time multi-asset portfolio choice. arXiv preprint arXiv:2504.11116, 2025.

Jaegi Jeon, Jeonggyu Huh, Hyeng Keun Koo, and Byung Hwa Lim. Breaking the dimensional barrier for constrained dynamic portfolio choice. arXiv preprint arXiv:2501.12600, 2025.

Jaegi Jeon, Jeonggyu Huh, Hyeng Keun Koo, and Byung Hwa Lim. Scalable pontryagin-guided adjoint-to-control recovery for constrained dynamic portfolio choice. *arXiv preprint arXiv:2608.15667*, 2026.

Shaolin Ji, Shige Peng, Ying Peng, and Xichuan Zhang. Solving stochastic optimal control problem via stochastic maximum principle with deep learning method. *Journal of Scientific Computing*, 93 (1):30, 2022. doi: 10.1007/s10915-022-01979-5.

Bekzhan Kerimkulov, David Šiška, and Lukasz Szpruch. A modified msa for stochastic control problems. *Applied Mathematics & Optimization*, 84(3):3417–3436, 2021. doi: 10.1007/s00245-021-09750-2.

Bjørnar Larssen. Dynamic programming in stochastic control of systems with delay. *Stochastics and Stochastics Reports*, 74(3-4):651–673, 2002. doi: 10.1080/1045112021000060764.

Bjørnar Larssen and Nils Henrik Risebro. When are HJB-equations for control problems with stochastic delay equations finite dimensional? *Stochastic Analysis and Applications*, 21(3): 643–671, 2003. doi: 10.1081/SAP-120020430.

Xuechen Li, Ting-Kam Leonard Wong, Ricky T. Q. Chen, and David Duvenaud. Scalable gradients for stochastic differential equations. In *Proceedings of the Twenty-Third International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *Proceedings of Machine Learning Research*, 2020.

Weijun Meng, Jingtao Shi, Tianxiao Wang, and Ji-Feng Zhang. A general maximum principle for optimal control of stochastic differential delay systems. *SIAM Journal on Control and Optimization*, 63(1):175–205, 2025. doi: 10.1137/23M1552024.

Salah-Eldin A. Mohammed. *Stochastic Functional Differential Equations*, volume 99 of *Research Notes in Mathematics*. Pitman Advanced Pub. Program, 1984. ISBN 9780273085935.

Bernt Øksendal, Agnès Sulem, and Tusheng Zhang. Optimal control of stochastic delay equations and time-advanced backward stochastic differential equations. *Advances in Applied Probability*, 43(2):572–596, 2011. doi: 10.1239/aap/1308662493.

Shige Peng. A general stochastic maximum principle for optimal control problems. *SIAM Journal on control and optimization*, 28(4):966–979, 1990. doi: 10.1137/0328054.

Shige Peng and Zhen Yang. Anticipated backward stochastic differential equations. *The Annals of Probability*, 37(3):877–902, 2009. doi: 10.1214/08-AOP423.

Matthew Riemer, Gopeshh Subbaraj, Glen Berseth, and Irina Rish. Enabling realtime reinforcement learning at scale with staggered asynchronous inference. In *International Conference on Learning Representations (ICLR)*, 2025.

Yuri F. Saporito and Zhaoyu Zhang. Path-dependent deep galerkin method: a neural network approach to solve path-dependent partial differential equations. *SIAM Journal on Financial Mathematics*, 12 (3):912–940, 2021. doi: 10.1137/20M1329597.

Paul J. Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, 1990. doi: 10.1109/5.58337.

Qingyuan Wu, Simon Zhan, Yixuan Wang, Yuhui Wang, Chung-Wei Lin, Chen Lv, Qi Zhu, and Chao Huang. Variational delayed policy optimization. In *Advances in Neural Information Processing Systems*, volume 37, pp. 54330–54356, 2024. doi: 10.52202/079017-1721.

Qingyuan Wu, Yuhui Wang, Simon Sinong Zhan, Yixuan Wang, Chung-Wei Lin, Chen Lv, Qi Zhu, Jürgen Schmidhuber, and Chao Huang. Directly forecasting belief for reinforcement learning with delays. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 67810–67832. PMLR, 2025.

Jingdong Zhang, Qunxi Zhu, Wei Yang, and Wei Lin. SYNC: Safety-aware neural control for stabilizing stochastic delay-differential equations. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=_8mS2NE-HXN.

Qunxi Zhu, Yao Guo, and Wei Lin. Neural delay differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Q1jmmQz72M2.

AI USE STATEMENT

Generative AI tools assisted with manuscript organization and English-language editing, LaTeX and figure preparation, proof review, and experimental code development. The authors are responsible for the mathematical arguments, implementation, experimental results, and final content of this work.

ETHICS STATEMENT

This paper develops methodological tools for delayed stochastic control, with potential applications in operations, finance, and epidemic modeling. Because model misspecification may lead to unreliable decisions, we view the method as a research tool rather than a directly deployable system.

REPRODUCIBILITY STATEMENT

Appendix A specifies the benchmark dynamics, reference solutions, discretizations, architectures, training and branching budgets, hardware, and paired evaluation protocol. Appendix B documents the estimator, mesh, and delay-model checks. Appendices C and D give the precise assumptions, discrete constructions, and proofs underlying the theoretical claims. Algorithm 1 and the mesh realization (16) specify the deployed recovery. Source code, configurations, and the scripts that regenerate every table from the recorded per-seed results are available at `https://anonymous.4open.science/r/PGDPO_with_delay-3AD0/`.

A BENCHMARK PROBLEMS: MODELS, RECOVERED CONTROLS, AND EXPERIMENT SETTINGS

Algorithm 1 PGDPO for stochastic state-delay control

Stage I: train the history policy π_θ by closed-loop BPTT and freeze $u^{\mathrm{ref}} = \pi_{\widehat{\theta}}$ as the reference control.
for each decision time t_k **do**
  Fix the observed buffer S_k, simulate N_{mc} future-noise branches under the detached actor, and estimate $(\widehat{p}_k^{\mathrm{ref}}, \widehat{\Pi}_k^{\mathrm{ref}})$ by OL-BPTT and $\widehat{\zeta}_k^{\mathrm{ref}}$ by the anchored regression (Appendix C.9).
  Reconstruct $\widehat{q}_k^{\mathrm{anc}}$, project blockwise onto $\mathcal{N}_p, \mathcal{N}_q, \mathcal{N}_\Pi$, and form $\widehat{\zeta}_k^{\mathcal{N}}$ (Stage II-B).
  Assemble and minimize the finite-step form (16) of (12) with the conditional next-step adjoint over $\mathcal{U}_{\mathrm{rec}}$, apply $\widehat{u}_k$, and advance the buffer.
end for

Scope. Every benchmark is a stochastic state-delay control problem with control-dependent diffusion, $\partial_u \sigma \not\equiv 0$, of the form

$$\begin{aligned} \mathrm{d}X_t &= b(t, X_t, Y_t, M_t, u_t)\,\mathrm{d}t + \sigma(t, X_t, Y_t, M_t, u_t)\,\mathrm{d}W_t, \\ Y_t &= X_{t-\delta}, \qquad M_t = \int_0^\delta K(a) X_{t-a}\,\mathrm{d}a, \end{aligned} \tag{15}$$

with a point delay Y_t and a distributed delay M_t with kernel K, except that Problem 3 uses exponentially fading memory with a separately initialized memory state, which admits the exact lift of Appendix A.3. In this appendix the Brownian motion is written W, since B is used for coefficient matrices, and N_{mc}, N_{out}, and N_{in} are the branch budgets of Algorithm 1 and Appendix C.9. All benchmarks minimize a cost, so the projection results stated for Hamiltonian maximization in the main text apply verbatim to $-H$, with the adjoint signs reversed and strong concavity replaced by strong convexity. Each subsection states the model and cost, the first-order Hamiltonian and anticipated adjoint, the continuous-time generalized-Hamiltonian control characterization, the reference where one exists, the discretization, the experimental settings, and the results not shown in the main text.

Recovery inputs. In the minimization convention the recovery inputs are the sign-reversed adjoints $(p, q, \Pi, \zeta)_{\min} = -(p, q, \Pi, \zeta)_{\max}$, and the superscript is suppressed throughout this appendix. The generalized-Hamiltonian recovery block in the minimization convention is

$$\mathcal{H}^{\min}(u) = \ell(u) + p^\top b(u) + \langle \zeta, \sigma(u) \rangle_F + \tfrac{1}{2} \operatorname{tr}\big(\sigma(u)^\top \Pi\, \sigma(u)\big),$$

so each subsection records what enters it: the costate p, the current-injection curvature block Π, and the shifted input $\zeta = q - \Pi\sigma^{\mathrm{ref}}$, which solves no adjoint equation of its own and is estimated by nested common-random-number regression rather than by subtraction. Along the optimal anchor $q = \zeta + \Pi\sigma(u^*)$, so the first-order and generalized Hamiltonians have the same control gradient there, and each subsection displays both forms together with the linear or box minimization rule that the quadratic block uses. On the mesh Π is the current block of the detached fixed-control OL-BPTT Hessian of the continuation payoff, not the value Hessian of a Riccati or HJB reference, and the full second-order family of Appendix C is not assembled, and each subsection records only its terminal Hessian and the active part of $D^2_{\chi\chi} H$.

Evaluation and reporting. Every method and every reference is rolled out on one evaluation set of $N_p = 256$ paths per problem, and each evaluation path fixes an initial history and a sequence of Brownian increments that all methods and the reference share, so that costs are compared path by path. Every learned method is trained with five seeds, and tables report the mean $\pm$ standard deviation across seeds on the fixed evaluation set. The control nRMSE of a method is $\big(\sum_{i,k} \|u^{\mathrm{m}}_{i,k} - u^{\mathrm{ref}}_{i,k}\|^2 \big/ \sum_{i,k} \|u^{\mathrm{ref}}_{i,k}\|^2\big)^{1/2}$, pooled over the evaluation paths i and decision times k, where $u^{\mathrm{ref}}_{i,k}$ is the reference control evaluated at the state that the method itself reached. Because the reference is evaluated at method-dependent states, this diagnostic does not directly estimate the control-process norm of Proposition 5.2. The paired gap ΔJ is the mean over paths of the realized cost of the method minus that of the reference on the same path, against the optimizer reference where one exists and against the Stage-I actor otherwise. Both statistics are computed per seed before the seed mean and standard deviation are taken. The reported p-value is a two-sided test of a zero paired gap with the standard error $\mathrm{se}^2 = \mathrm{sd}^2_{\mathrm{seed}}/5 + \mathrm{se}^2_{\mathrm{MC}}$, where $\mathrm{sd}_{\mathrm{seed}}$ is the standard deviation of the per-seed gaps and $\mathrm{se}_{\mathrm{MC}}$ the mean over seeds of the per-seed paired path standard errors of the gap on the evaluation set, with p taken from the standard normal distribution, so that both training-seed and path-sampling variation are treated as noise. Figure bands show the standard deviation across seeds of the displayed statistic, the per-seed path means for averaged curves and the same selected evaluation path for the single-path panels.

Mesh realization of Stage II. Equation (12) specifies the generalized-Hamiltonian recovery block, and on the Euler mesh the experiments minimize at each decision the finite-step quadratic

$$\mathcal{H}^{\mathrm{mesh}}_k(u) = \ell_k(u) + \bar{p}^\top_{k+1|k} b_k(u) + \langle \zeta_k, \sigma_k(u) \rangle_F + \tfrac{1}{2} \operatorname{tr}\big(\sigma_k(u)^\top \Pi_k\, \sigma_k(u)\big) + \tfrac{h}{2}\, \Delta b_k(u)^\top \Pi_k\, \Delta b_k(u), \tag{16}$$

with $\Delta b_k(u) := b_k(u) - b_k(u^{\mathrm{ref}}_k)$, assembled from the same projected inputs with two changes: the drift is paired with $\bar{p}_{k+1|k}$, the estimated adjoint at the next mesh point conditioned on the history at t_k and projected like p_k, since the current control reaches the future through the next state, and the drift increment $h\,\Delta b_k(u)$ enters the curvature through the estimated current block Π_k. With the history and the supplied inputs fixed, ℓ_k is quadratic and b_k and σ_k are affine in u in all five problems, so (16) is a quadratic in u and is minimized exactly, by a linear solve in Problems 1 and 2 and by box minimization in Problems 3 to 5, so the deployed controls are feasible by construction. On the recorded PGDPO recovery queries of Problems 1, 3, 4, and 5 and of Problem 2 at $n = 5$, the pooled root-mean-square KKT residual of the deployed controls for this finite-step quadratic is below 10^{-7}. The residual (13) is instead evaluated for the continuous-form block, so exact minimization of (16) need not make it vanish: the difference is retained in the residual layer δ_{num} of Proposition 5.2. At the same deployed controls and with the corresponding recorded input estimates, the continuous-form pooled RMS residuals are 1.28×10^{-2}, 5.42×10^{-3}, 9.28×10^{-4}, 7.46×10^{-3}, and 1.79×10^{-2} in Problems 1, 2 ($n = 5$), 3, 4, and 5, respectively.

Shared method settings. Stage I is a two-layer LSTM of width 256 trained with Adam at batch 1024 and learning rate 5×10^{-5} on the running history buffer, keeping the checkpoint with the smallest cost on a fixed validation batch, evaluated every 100 updates and not involving the optimizer reference. Stage II estimates p and Π from $N_{\mathrm{mc}} = 2048$ continuation branches per decision, and

the nested regression for ζ uses $N_{\text{out}} = 256$ outer increments with two antithetic legs of $N_{\text{in}} = 8$ inner continuations each, without ridge regularization, $\kappa_Z = 0$. The projection intervals $\mathcal{N}_p$, $\mathcal{N}_q$, and $\mathcal{N}_\Pi$ of each problem were calibrated empirically from preliminary runs without input clipping, with problem-specific margins at the observed numerical scales of the estimates. The curvature intervals have nonnegative lower bounds, which, together with the positive control-effort weights and the affine control loadings, keep the supplied recovery quadratics strongly convex in every problem. On the recorded Stage-II queries of the runs in Table 19, no finite-range clipping was observed in Problems 1, 3, 4, and 5 or in Problem 2 at $n = 5$, so the selected intervals did not bind there and the spectral projector of Problem 2 introduced only floating-point reconstruction differences.

PDGM learns the delay value functional through a convex head on the current state and deploys the control that its gradient prescribes, and DeepSMP learns the first-order pair (p, q) by multiple shooting of the anticipated BSDE, propagating the adjoint over short blocks of steps whose boundary values are predicted from the encoded history and matched across blocks, and deploys the control that the stationarity condition prescribes. Both use the encoder width and depth of Stage I and are trained for 3000 updates at batch 1024 and learning rate 10^{-4}, keeping the checkpoint with the smallest value of their own training loss on a fixed validation batch in the same way as Stage I, and their problem-specific variants are listed in each subsection.

Hardware and runtime. All runs use one NVIDIA RTX A6000 (48 GB) on an Ubuntu 22.04 server with an Intel Xeon Gold 6430 CPU, under Python 3.10 and PyTorch 2.4.1 with CUDA 12.4. Table 2 reports wall-clock times. Stage I, PDGM, and DeepSMP are trained once per seed, whereas Stage II performs no training and its entry is the online recovery over all $256 \times N_h$ decisions of the evaluation set, between 0.05 and 0.45 seconds per decision across the problems. The per-decision figures are amortized averages over the batched evaluation rather than batch-one latency measurements or a real-time control demonstration, and a batch-one timing of Problem 5 is in Table 10.

Table 2: **Runtime** in minutes, five-seed mean, with (n, m) the physical state and control dimensions. Training time for Stage I, PDGM, and DeepSMP, and online recovery time over the evaluation set for Stage II.

Problem	(n, m)	Stage I	Stage II	PDGM	DeepSMP
P1	$(1, 1)$	15.5	29.2	27.4	17.3
	$(1, 1)$	17.9	45.7	34.5	38.1
P2	$(5, 5)$	21.8	49.9	122.0	38.8
	$(20, 20)$	21.9	97.0	462.7	39.1
P3	$(1, 1)$	5.3	11.1	72.3	26.6
P4	$(2, 1)$	13.8	60.5	73.9	24.9
P5	$(1, 1)$	19.1	91.6	32.7	24.3

A.1 Problem 1: Point-delay LQ tracking with controlled diffusion

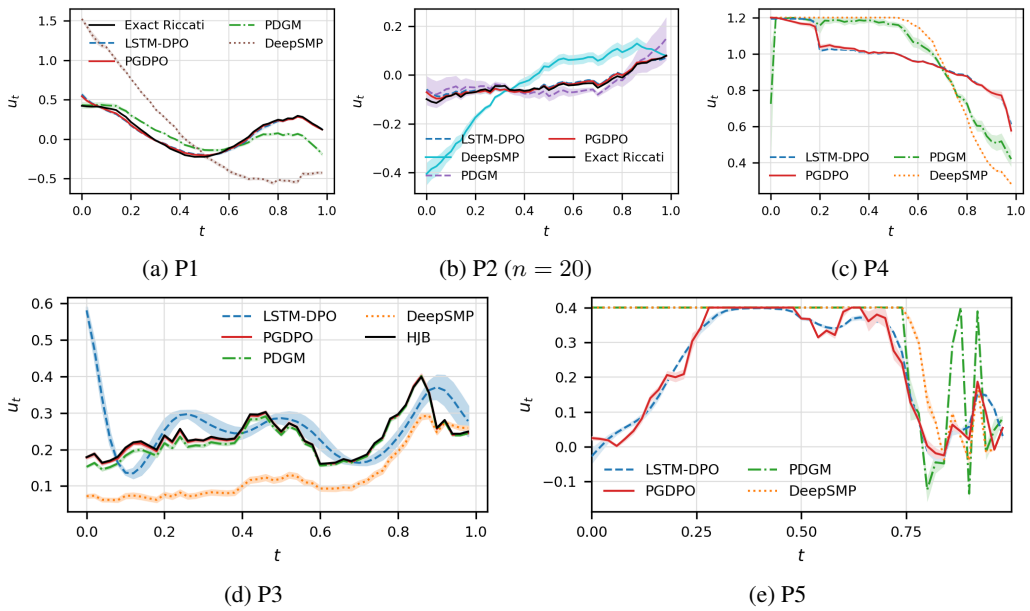
Model and cost. The scalar state follows the linear SDDE

$$\begin{aligned} \mathrm{d}X_t &= \big(aX_t + a_d X_{t-\delta} + b\,u_t\big)\,\mathrm{d}t + \big(\sigma_0 + c_x X_t + c_y X_{t-\delta} + \gamma_u u_t\big)\,\mathrm{d}W_t, \quad t \in [0, T], \\ X_t &= x_0(t), \quad t \in [-\delta, 0], \end{aligned} \tag{17}$$

so the delayed state enters the drift through a_d and the diffusion through c_y, and the control enters the diffusion through $\gamma_u \neq 0$. The control is unconstrained and the objective is

$$\min_u \; J(u) = \mathbb{E}\left[\int_0^T \left\{ \tfrac{Q}{2}\big(X_t - x_t^{\text{ref}}\big)^2 + \tfrac{R}{2} u_t^2 \right\} \mathrm{d}t + \tfrac{Q_T}{2}\big(X_T - x^{\text{tar}}\big)^2 \right], \tag{18}$$

with a deterministic reference path x_t^{ref} and terminal target x^{tar}. Initial histories are drawn from constant, ramp, and cosine templates on $[-\delta, 0]$ with amplitude $\mathrm{U}[-1.2, 1.2]$. The terminal costate is $p_T = Q_T(X_T - x^{\text{tar}})$, and the delayed drift and diffusion make the optimal control depend on the state history rather than on time alone.



(a) P1 (b) P2 ($n = 20$) (c) P4

(d) P3 (e) P5

Figure 4: **Deployed controls**, five-seed mean ± SD band. Top: averages over the evaluation paths for P1, P2 ($n = 20$), and P4, reference in black where available. Bottom: one selected evaluation path for P3 and P5, with the P5 correction torque u in its box.

Hamiltonian and adjoint. The first-order Hamiltonian is

$$H(t,x,y,u,p,q) = \tfrac{Q}{2}\big(x - x_t^{\mathrm{ref}}\big)^2 + \tfrac{R}{2}u^2 + \big(ax + a_d y + bu\big)p + \big(\sigma_0 + c_x x + c_y y + \gamma_u u\big)q, \quad (19)$$

and the costate solves the anticipated BSDE

$$\begin{aligned} -\mathrm{d}p_t &= \Big(Q\big(X_t - x_t^{\mathrm{ref}}\big) + a\,p_t + c_x\,q_t + \mathbb{E}_t\big[a_d\,p_{t+\delta} + c_y\,q_{t+\delta}\big]\mathbf{1}_{\{t+\delta\le T\}}\Big)\,\mathrm{d}t - q_t\,\mathrm{d}W_t, \\ p_T &= Q_T\big(X_T - x^{\mathrm{tar}}\big). \end{aligned} \quad (20)$$

The delayed diffusion loading c_y places the anticipated martingale component $\mathbb{E}_t[q_{t+\delta}]$ in the adjoint drift, and the anticipated block is absent after $t = T - \delta$, a structural change of the adjoint equation at the time where the control in Figure 4 changes slope. The second-order family of Appendix C is driven by $H_{xx} = Q$ and the terminal datum Q_T alone, so it is deterministic, and only its current-injection block Π_t enters the solver.

Recovered control. With $\bar\sigma_t := \sigma_0 + c_x X_t + c_y Y_t$, the control-dependent block of the generalized Hamiltonian is $\frac{R}{2}u^2 + b\,p_t\,u + \zeta_t(\bar\sigma_t + \gamma_u u) + \frac{1}{2}\Pi_t(\bar\sigma_t + \gamma_u u)^2$. Stationarity in the first-order pair reads $Ru_t^* + b\,p_t + \gamma_u q_t = 0$, and substituting $q_t = \zeta_t + \Pi_t(\bar\sigma_t + \gamma_u u_t^*)$ gives the generalized-Hamiltonian form,

$$u_t^* = -\frac{b\,p_t + \gamma_u\zeta_t + \gamma_u\Pi_t\bar\sigma_t}{R + \gamma_u^2\Pi_t}. \quad (21)$$

The diffusion-control channel raises the effective control weight from R to $R + \gamma_u^2\Pi_t$ and adds the shifted input $\gamma_u\zeta_t$ and the cross term $\gamma_u\Pi_t\bar\sigma_t$ to the numerator, and the projection interval $\mathcal{N}_\Pi$ of Table 3 keeps the denominator positive. In the experiments this block is assembled on the mesh as in (16).

Reference. On the buffer $Z_k = (X_k, \dots, X_{k-H_h})$ the discretized problem is a stochastic LQ problem with control-dependent diffusion, and its generalized Riccati recursion on the same mesh is the exact optimizer against which Table 1 reports the control error and the paired cost gap.

Protocol and settings. All methods share the mesh $h = 0.02$, $H_h = 10$, $N_h = 50$. Stage I trains for 5000 updates, and Stage II clips to the intervals of Table 3. PDGM uses the value head convex

Table 3: **Problem 1:** model parameters and experimental settings.

Description	Symbol	Value
Horizon / delay / mesh	$T, \delta, (h, H_h, N_h)$	$1.0, 0.2, (0.02, 10, 50)$
Drift (current / delayed / control)	a, a_d, b	$-0.3, 0.9, 1.0$
Diffusion loadings	$\sigma_0, c_x, c_y, \gamma_u$	$0.2, 0.1, 0.3, 0.4$
Tracking / control / terminal weights	Q, R, Q_T	$1.0, 0.1, 2.0$
Reference path / terminal target	$x_t^{\mathrm{ref}}, x^{\mathrm{tar}}$	$0.5\sin(2\pi t/T), 0.3$
Initial-history templates	$x_0(\cdot)$	constant/ramp/cosine, amplitude $\mathrm{U}[-1.2, 1.2]$
Stage-I updates	—	5000
Projection intervals	$\mathcal{N}_p, \mathcal{N}_q, \mathcal{N}_\Pi$	$[-4.2, 0.7], [-0.6, 4.6], [1.8, 6.6]$

in the current state and DeepSMP the multiple-shooting adjoint with an LSTM history encoder and five-step blocks.

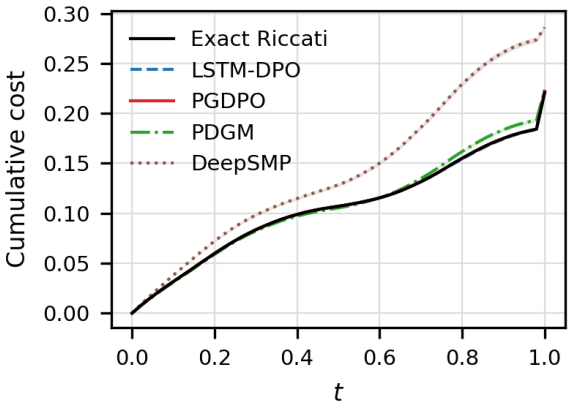


Figure 5: **Problem 1:** cumulative cost averaged over the evaluation paths, five-seed mean, with the exact discrete-time optimizer in black. The jump at $t = T$ is the terminal penalty.

Results. The paired gap of the Stage-II control to its own Stage-I actor is -1.61×10^{-3} with $p = 0.06$, and the projection of Table 3 is not active on the evaluation set. Figure 5 shows the cumulative cost: PGDPO and LSTM-DPO lie on the reference at this scale, PDGM departs from it after $t \approx 0.6$, and DeepSMP accumulates excess cost from the start.

A.2 Problem 2: Distributed-delay network LQ with multi-channel controlled diffusion

Model and cost. Let $X_t, u_t \in \mathbb{R}^n$ and let $W_t \in \mathbb{R}^r$ collect r independent Brownian channels. The memory is the truncated exponential average

$$M_t = \int_0^\delta K(a)\, X_{t-a}\, \mathrm{d}a, \qquad K(a) = \frac{\rho\, e^{-\rho a}}{1 - e^{-\rho\delta}}\, \mathbf{1}_{[0,\delta]}(a), \tag{22}$$

discretized on the mesh by the normalized trapezoidal weights $M_k = \sum_{j=0}^{H_h} w_j X_{k-j}$ with $w_j \propto c_j K(jh)$, $c_0 = c_{H_h} = \frac{1}{2}$ and $c_j = 1$ otherwise. Differentiating the truncated memory re-introduces the point delay $K(\delta)X_{t-\delta}$, so the dynamics do not close in (X_t, M_t) alone, the numerical representation retains the history buffer, and the buffer dimension $n(H_h{+}1)$ is the quantity this problem scales. The dynamics and objective are

$$\begin{aligned} \mathrm{d}X_t &= \big(AX_t + A_M M_t + Bu_t\big)\,\mathrm{d}t + \sum_{\ell=1}^{r}\big(C_\ell X_t + C_{M,\ell} M_t + D_\ell u_t + \sigma_\ell^0\big)\,\mathrm{d}W_t^\ell, \\ \min_u\ J(u) &= \mathbb{E}\left[\int_0^T \Big(\tfrac{1}{2}X_t^\top Q X_t + \tfrac{1}{2}u_t^\top R u_t\Big)\mathrm{d}t + \tfrac{1}{2}X_T^\top Q_T X_T\right], \end{aligned} \tag{23}$$

with u unconstrained, control loadings $D_\ell \neq 0$ in every channel, and additive loadings σ_ℓ^0. Every coefficient matrix is a polynomial $a_0 I + a_1 L$ in the Laplacian L of the ring graph on n nodes, so the system is coupled in the observed coordinates while all matrices share one orthonormal eigenbasis. Initial histories are drawn in each eigenmode from constant, linear, and cosine templates with amplitude $\mathrm{U}[-1,1]$ and uniform phase, and the dimension is varied over $n \in \{1,5,20\}$ at fixed $r = 4$.

Hamiltonian and adjoint. Writing $\bar{\sigma}_{\ell,t} := C_\ell X_t + C_{M,\ell} M_t + \sigma_\ell^0$ for the control-independent part of channel ℓ and $(p, q_1, \dots, q_r)$ for the costate and its martingale components,

$$\begin{aligned} H(t,x,m,u,p,q) = \tfrac{1}{2} x^\top Q x + \tfrac{1}{2} u^\top R u + \big(Ax + A_M m + Bu\big)^\top p \\ + \sum_{\ell=1}^{r} q_\ell^\top \big(C_\ell x + C_{M,\ell} m + D_\ell u + \sigma_\ell^0\big), \end{aligned} \tag{24}$$

and since X_t re-enters M_s for $s \in [t, t+\delta]$ with weight $K(s-t)$ through the drift and every diffusion channel,

$$\begin{aligned} -\mathrm{d}p_t = \Big(& Q X_t + A^\top p_t + \sum_{\ell=1}^{r} C_\ell^\top q_{\ell,t} \\ & + \int_0^\delta K(a) \mathbb{E}_t \left[A_M^\top p_{t+a} + \sum_{\ell=1}^{r} C_{M,\ell}^\top q_{\ell,t+a} \right] \mathbf{1}_{\{t+a \le T\}} \, \mathrm{d}a \Big) \mathrm{d}t - \sum_{\ell=1}^{r} q_{\ell,t} \, \mathrm{d}W_t^\ell, \\ p_T = {} & Q_T X_T. \end{aligned} \tag{25}$$

The memory diffusion loadings $C_{M,\ell}$ place the anticipated martingale components in the adjoint drift, and the second-order family is driven by $H_{xx} = Q$ and Q_T alone, so it is deterministic and separable across the eigenmodes of the common eigenbasis.

Recovered control. First-order stationarity reads $R u_t^* + B^\top p_t + \sum_\ell D_\ell^\top q_{\ell,t} = 0$, and substituting $q_{\ell,t} = \zeta_{\ell,t} + \Pi_t(\bar{\sigma}_{\ell,t} + D_\ell u_t^*)$ gives

$$u_t^* = -\mathcal{R}_t^{-1} g_t, \qquad \mathcal{R}_t := R + \sum_{\ell=1}^{r} D_\ell^\top \Pi_t D_\ell, \qquad g_t := B^\top p_t + \sum_{\ell=1}^{r} D_\ell^\top \big(\zeta_{\ell,t} + \Pi_t \bar{\sigma}_{\ell,t}\big), \tag{26}$$

so that the contributions of the r diffusion channels enter both the curvature and the gradient, and the spectral interval $\mathcal{N}_\Pi$ of Table 4 keeps $\mathcal{R}_t \succ 0$. In the experiments this block is assembled on the mesh as in (16).

Reference. The generalized Riccati recursion on the $n(H_h{+}1)$-dimensional buffer is the exact same-mesh optimizer against which Table 1 reports the control error and the paired cost gap. Because all matrices and the memory weights are simultaneously diagonal in the ring eigenbasis, the recursion decouples into n scalar $(H_h{+}1)$-dimensional delay-LQ problems, which keeps the reference exact and cheap at $n = 20$.

Protocol and settings. All methods share the mesh $h = 0.02$, $H_h = 10$, $N_h = 50$ at every n. Stage I trains for 5000 updates, and both stages work with the objective divided by n, so that the recovery inputs and the n-dependent projection sets of Table 4, Euclidean balls for $\widehat{p}$ and $\widehat{q}^{\,\mathrm{anc}}$ and a spectral interval for $\widehat{\Pi}$, are in units of J/n, whereas the reported costs are totals over the coordinates. Since every term of the local objective is scaled by the same constant, the recovered control is unchanged by this normalization. PDGM uses the value head convex in the current state and DeepSMP the multiple-shooting adjoint with one-step blocks, with the one-step adjoint operator rescaled by a fixed diagonal balancing that equalizes the numerical scales of the current and memory coordinates.

Results. Table 1 reports $n = 20$ and Table 5 the lower dimensions. PGDPO reduces the control error of its actor by 35%, 27%, and 24% at $n = 1, 5, 20$ and has five- to elevenfold lower control error than DeepSMP at every n. At each tested dimension the cost gap of PGDPO to the exact optimizer

Table 4: **Problem 2:** model parameters and experimental settings. All matrices are $a_0 I + a_1 L$ with L the ring-graph Laplacian, listed as $(a_0; a_1)$ for scalars and as the vector of a_0 followed by the vector of a_1 for the channel loadings.

Description	Symbol	Value
Horizon / delay / mesh	T, δ, (h, H_h, N_h)	1.0, 0.2, $(0.02,\ 10,\ 50)$
Dimensions / channels	n, r	$\{1, 5, 20\}$, 4
Kernel	K	truncated exponential, $\rho\delta = 2.5$
Drift (state / memory / control)	A; A_M; B	$(-0.5; -0.15)$; $(0.6; 0.15)$; $(1.0; 0.2)$
Diffusion state loadings	$C_{1:4}$ $(a_0; a_1)$	$0.15, 0.10, 0.08, 0.12$; $0.05, -0.03, 0.02, 0$
Diffusion memory loadings	$C_{M,1:4}$ $(a_0; a_1)$	$0.10, 0, 0.05, 0.03$; $0, 0.05, 0, 0.02$
Diffusion control loadings	$D_{1:4}$ $(a_0; a_1)$	$0.30, 0.15, 0.20, 0.10$; $0.10, 0, 0.05, 0$
Additive loadings	σ_ℓ^0	fixed $\mathcal{N}(0, 0.2^2)$ draw per eigenmode
Cost weights	Q; R; Q_T	$(1.0; 0.2)$; $(0.1; 0.02)$; $(2.0; 0)$
Initial-history templates	$x_0(\cdot)$	constant/linear/cosine, amplitude $\mathrm{U}[-1, 1]$
Stage-I updates	—	5000
Ball radii, $n = 1$ / 5 / 20	$\mathcal{N}_p, \mathcal{N}_q$	$(1.10, 2.10)$ / $(0.40, 0.70)$ / $(0.30, 0.38)$
Spectral interval, $n = 1$ / 5 / 20	$\mathcal{N}_\Pi$	$[1.8, 4.2]$ / $[0.35, 1.2]$ / $[0.09, 0.30]$

Table 5: **Problem 2:** results at $n \in \{1, 5\}$ in the format of Panel A of Table 1. Costs are totals over the n coordinates, and the paired gaps of the Stage-II control to its Stage-I actor at all three dimensions are reported in the text.

n	Algorithm	Control nRMSE	ΔJ vs reference	p
1	PGDPO	$(2.258 \pm 0.134) \times 10^{-2}$	$(0.500 \pm 0.532) \times 10^{-4}$	0.75
	LSTM-DPO	$(3.465 \pm 0.527) \times 10^{-2}$	$(1.397 \pm 0.652) \times 10^{-4}$	0.51
	DeepSMP	$(2.542 \pm 0.418) \times 10^{-1}$	$(5.346 \pm 1.531) \times 10^{-3}$	< 0.001
	PDGM	$(8.241 \pm 0.647) \times 10^{-2}$	$(2.534 \pm 1.300) \times 10^{-4}$	0.50
5	PGDPO	$(7.186 \pm 0.314) \times 10^{-2}$	$(0.845 \pm 0.223) \times 10^{-3}$	0.53
	LSTM-DPO	$(9.782 \pm 0.388) \times 10^{-2}$	$(2.483 \pm 0.214) \times 10^{-3}$	0.13
	DeepSMP	$(5.417 \pm 0.136) \times 10^{-1}$	$(9.142 \pm 0.565) \times 10^{-2}$	< 0.001
	PDGM	$(9.889 \pm 1.556) \times 10^{-2}$	$(4.794 \pm 0.906) \times 10^{-3}$	0.001

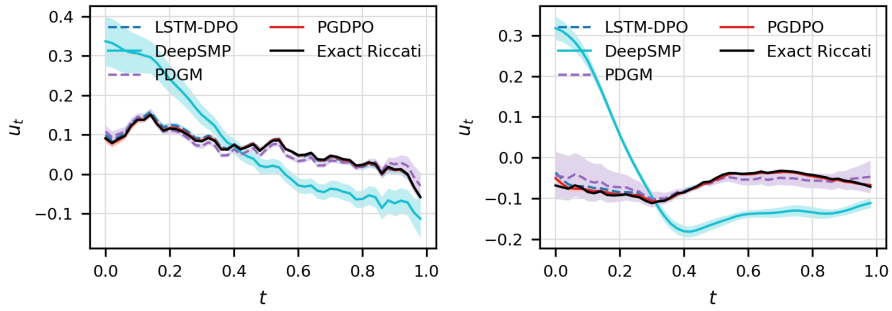
is not significant, with $p \geq 0.53$, and the gaps of the actor are also not significant. The paired gap between the two, which cancels the path-sampling noise of the reference cost, is -0.9×10^{-4}, -1.6×10^{-3}, and -1.7×10^{-2} at $n = 1, 5, 20$, not resolved at $n = 1$ with $p = 0.46$ and significant at $n = 5$ and $n = 20$ with $p < 0.001$. At $n = 1$ the actor is already within 3.5% of the reference control and the gap is not resolved. PDGM has the smallest control error at $n = 20$ but a significantly worse cost at $n = 5$ and $n = 20$, with $p = 0.001$ and 0.022. Figures 6 and 7 show the averaged controls and the cumulative cost, where PGDPO, LSTM-DPO, and PDGM lie on the reference at every n at this scale and the excess cost of DeepSMP grows with n.

A.3 Problem 3: Nonlinear renewal epidemic with intervention-dependent volatility

Model and cost. Let $I_t \geq 0$ be the infectious burden in a population of size N_{pop}, driven by the delayed infection pressure M_t and controlled by an intervention intensity $u_t \in [0, 1]$,

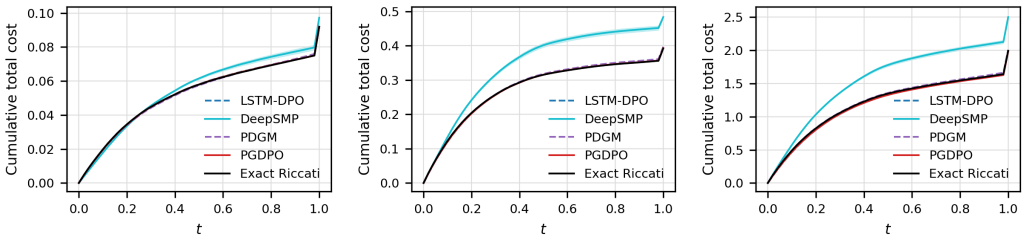
$$\begin{aligned} \mathrm{d}I_t &= \Big[\beta\Big(1 - \frac{I_t}{N_{\mathrm{pop}}}\Big) M_t - \gamma I_t - b\, u_t I_t\Big]\,\mathrm{d}t + \sigma_0\big(1 - \eta_\sigma u_t\big) I_t\,\mathrm{d}W_t, \\ \min_u\ J(u) &= \mathbb{E}\left[\int_0^T \Big(\tfrac{c_I}{2} I_t^2 + \tfrac{R}{2} u_t^2\Big)\mathrm{d}t + \tfrac{c_T}{2} I_T^2\right], \end{aligned} \tag{27}$$

where β is the transmission rate, γ the recovery rate, b the treatment efficacy, σ_0 the outbreak volatility, and $\eta_\sigma \in (0, 1)$ the intervention loading of the volatility, so the control enters the diffusion multiplicatively through the state and lowers the mean burden and its volatility at once. The infectivity



(a) $n = 1$ (b) $n = 5$

Figure 6: **Problem 2:** first coordinate of the deployed control averaged over the evaluation paths, five-seed band, with the exact Riccati control in black. The $n = 20$ panel is in Figure 4.



(a) $n = 1$ (b) $n = 5$ (c) $n = 20$

Figure 7: **Problem 2:** cumulative total cost averaged over the evaluation paths, five-seed mean, with the exact Riccati optimizer in black and the terminal penalty appearing as the jump at $t = T$.

memory fades exponentially from a prescribed initial memory M_0,

$$M_t = e^{-\rho t} M_0 + \int_0^t \rho\, e^{-\rho(t-s)}\, I_s \, \mathrm{d}s, \qquad \text{equivalently} \qquad \mathrm{d}M_t = \rho\,(I_t - M_t)\,\mathrm{d}t, \tag{28}$$

with I_0 and M_0 drawn independently from the uniform laws of Table 6. This memory lifts to the Markov state (I_t, M_t), which is what makes a numerical HJB reference available for this nonlinear problem.

Hamiltonian and adjoint. The first-order Hamiltonian is

$$H(t, i, m, u, p, q) = \tfrac{c_I}{2} i^2 + \tfrac{R}{2} u^2 + \Big[\beta\Big(1 - \frac{i}{N_{\mathrm{pop}}}\Big) m - \gamma i - b\, u\, i\Big] p + \sigma_0 (1 - \eta_\sigma u)\, i\, q, \tag{29}$$

and with $K(a) = \rho e^{-\rho a}$ the costate solves

$$\begin{aligned} -\mathrm{d}p_t = \Big(c_I I_t - \Big(\frac{\beta M_t}{N_{\mathrm{pop}}} + \gamma + b u_t\Big) p_t + \sigma_0 (1 - \eta_\sigma u_t) q_t \\ + \int_0^\infty K(a) \beta \mathbb{E}_t \Big[\Big(1 - \frac{I_{t+a}}{N_{\mathrm{pop}}}\Big) p_{t+a}\Big] \mathbf{1}_{\{t+a \le T\}}\, \mathrm{d}a\Big) \mathrm{d}t - q_t\, \mathrm{d}W_t, \\ p_T = c_T I_T. \end{aligned} \tag{30}$$

Unlike the LQ problems, the anticipated integrand couples the future costate to the future state nonlinearly. The second-order source Hessian on (i, m) has the off-diagonal entry $-\beta p_t / N_{\mathrm{pop}}$, so the second-order objects are stochastic even though the coefficients are smooth, and on the lifted state they close as a 2×2 matrix BSDE whose (I, I) entry is the recovery curvature Π_t.

Recovered control. At an interior optimum, first-order stationarity reads $R u_t^* - b\, I_t\, p_t - \eta_\sigma \sigma_0 I_t\, q_t = 0$, and substituting $q_t = \zeta_t + \Pi_t\, \sigma_0 (1 - \eta_\sigma u_t^*) I_t$ and projecting onto the box gives

$$u_t^* = \mathrm{clip}\Big(\frac{b\, I_t\, p_t + \eta_\sigma \sigma_0 I_t\, \zeta_t + \eta_\sigma \sigma_0^2 I_t^2\, \Pi_t}{R + \eta_\sigma^2 \sigma_0^2 I_t^2\, \Pi_t},\ 0,\ 1\Big) \tag{31}$$

when the quadratic coefficient $R + \eta_\sigma^2 \sigma_0^2 I_t^2 \, \Pi_t$ is positive. Clipping at 0 and 1 yields state-dependent switching surfaces in (t, I_t, M_t). The HJB reference carries the value curvature V_{II} in the same place, and the agreement of the two first-order conditions under $p = V_I$ and $q = V_{II}\sigma$ does not identify Π with V_{II}. In the numerical recovery, the scalar current-infection curvature estimate is projected onto the nonnegative interval $\mathcal{N}_\Pi = [0, 7.5]$ of Table 6 before it enters the mesh recovery block (16), and the lifted 2×2 matrix is not assembled. Since the drift and diffusion control loadings act only in the infection coordinate, both the continuous and the finite-step quadratic coefficients are bounded below by $R = 0.22$, so the box recovery clips the unconstrained minimizer to $[0, 1]$ without the nonpositive-curvature endpoint fallback.

Discretization and reference. All methods share the full-truncation Euler scheme on the same mesh, with coefficients, memory update $M_{k+1} = M_k + \rho h \, (I_k^+ - M_k)$, and costs evaluated at I_k^+. No negative grid-point states were observed on the evaluation set.

The learned methods encode the raw infection prefix $(I_0, \ldots, I_k)$ together with the initial memory M_0 and the time, and the current memory summary M_k is not supplied as a network input, while the simulator propagates the memory recurrence above for every method. Stage II holds the past prefix and M_0 fixed, differentiates the continuation with respect to the current infection coordinate, and retains the physical memory propagation under the detached future controls. The reference is a monotone upwind finite-difference solution of the HJB equation on $(I, M) \in [0, 1.5]^2$ with exact minimization of the discretized control objective at every grid node.

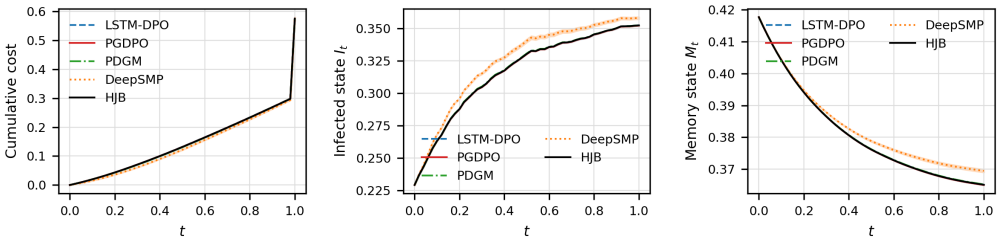
Refining the grid from 61^2 to the 151^2 reference grid changes the initial value by a relative RMSE of 1.8×10^{-3} over 512 probe states, and the largest policy RMSE across the five checked time slices is 4.0×10^{-3}. At the coarser spacing, expanding the domain from $[0, 1.5]^2$ on a 61^2 grid to $[0, 1.8]^2$ on a 73^2 grid gives an initial-value relative RMSE of 3.9×10^{-5} and a maximum time-slice policy RMSE of 9.5×10^{-6}. At the checked nodes, re-minimization reduces the discretized HJB control objective by at most 3×10^{-11}. On a separate set of 8000 validation paths, the mean rollout-minus-initial-value discrepancy is -4.8×10^{-4}, with a paired Monte Carlo standard error of 1.1×10^{-3}.

Table 6: **Problem 3:** model parameters and experimental settings.

Description	Symbol	Value
Horizon / mesh	T, (h, N_h)	1.0, $(0.02,\ 50)$
Transmission / recovery / treatment	β, γ, b	2.0, 0.6, 0.6
Memory rate / population	ρ, N_{pop}	0.8, 1.0
Volatility / intervention loading	σ_0, η_σ	0.7, 0.7
Infection / effort / terminal weights	c_I, R, c_T	4.0, 0.22, 4.0
Control box	$\mathcal{U}$	$[0, 1]$
Initial law	I_0, M_0	U$[0.05, 0.4]$, U$[0.02, 0.8]$
HJB reference grid	—	151^2 on $[0, 1.5]^2$, CFL 0.8
Stage-I updates	—	3000
Projection intervals	$\mathcal{N}_p, \mathcal{N}_q, \mathcal{N}_\Pi$	$[-0.5, 3.0]$, $[-0.25, 0.75]$, $[0, 7.5]$

Protocol and settings. All methods share the mesh $h = 0.02$, $N_h = 50$. Stage I trains through a logistic squashing onto the box for 3000 updates, and Stage II clips to the intervals of Table 6 and solves the box quadratic exactly. PDGM's value head is convex in the current burden I, and DeepSMP uses five-step blocks with a raw-history encoder and a kernel-structured adjoint head, whose lag components are constrained by the known exponential-memory weights. Its adjoint and terminal losses sum over active coordinates, while block-boundary matching averages over them.

Results. Table 1 reports the comparison against the HJB reference. PGDPO reduces the control nRMSE of its actor from 0.0706 to 0.0064, a reduction of 91%, while PDGM and the kernel-structured DeepSMP reach 0.0278 and 0.2489. Its paired realized-cost gap to the HJB policy is 2.2×10^{-5} with $p = 0.63$, and its paired gap to the Stage-I actor is -5.6×10^{-4} with $p = 0.30$, so the control-accuracy improvement is not accompanied by a cost improvement over the actor that is resolved at the 5% level. The selected control path in Figure 4 shows close tracking by PGDPO, while Figure 8 shows higher mean infection and memory under DeepSMP and much smaller differences among the cumulative-cost curves.



(a) Cumulative cost (b) Infectious burden I_t (c) Infectivity memory M_t

Figure 8: **Problem 3:** averages over the evaluation paths, five-seed band, with the HJB policy in black. The jump at $t = T$ is the terminal penalty, and the selected control path is in Figure 4.

A.4 Problem 4: Signal-aware optimal execution with finite-memory impact and stochastic fills

Model and cost. Problems 1 to 3 place the delay in the dynamics, so with the adjoint pair held fixed the past does not appear in the stationarity condition. Here the delay sits in the running cost: past trading changes the price at which the current trade executes, so the memory enters the pointwise stationarity condition directly. The state is the inventory Q_t, traded at the intended rate u_t with stochastic fills, together with a mean-reverting predictive signal α_t,

$$\mathrm{d}Q_t = -u_t\,\mathrm{d}t - \sigma_Q u_t\,\mathrm{d}W_t^Q, \quad \mathrm{d}\alpha_t = -\kappa_\alpha\,\alpha_t\,\mathrm{d}t + \sigma_\alpha\,\mathrm{d}W_t^\alpha, \quad Q_s = q_0,\ u_s = 0 \ \text{ for } s \le 0, \tag{32}$$

with independent W^Q, W^α and $\alpha_0 = 0$, so the diffusion is control-driven and state-independent, the pattern complementary to Problems 1 to 3. Realized fills $\mathrm{d}L_t = -\mathrm{d}Q_t$ leave a transient impact over a finite window,

$$I_t^\delta = \int_{t-\delta}^{t} G(t-s)\,\mathrm{d}L_s, \qquad G(a) = \gamma\, e^{-\rho_G a}, \quad a \in [0,\delta], \tag{33}$$

and the sell-only trader with $u_t \in [0, \bar{u}]$ minimizes

$$J(u) = \mathbb{E}\left[\int_0^T \left\{\tfrac{\phi}{2}Q_t^2 + \tfrac{\eta}{2}u_t^2 + u_t\,I_t^\delta - \alpha_t\,u_t\right\}\mathrm{d}t + \tfrac{\kappa}{2}\,Q_T^2\right], \tag{34}$$

which charges inventory risk, temporary impact, trading into one's own past impact, the gain from trading with the signal, and a terminal inventory penalty. With $K_I(a) := -G'(a) = \rho_G\gamma e^{-\rho_G a}$ and $M_t^Q := \int_0^\delta K_I(a) Q_{t-a}\,\mathrm{d}a$, integration by parts gives the pathwise identity

$$I_t^\delta = G(\delta)\,Q_{t-\delta} - G(0)\,Q_t + M_t^Q, \tag{35}$$

a point delay plus a distributed delay of the inventory, so the problem is a state-delay problem of the form (15) with the delay operators acting on Q.

Hamiltonian and adjoint. With costates (p^Q, p^α), martingale components indexed by the two Brownian columns, and $I = G(\delta)y - G(0)Q + m$ for $y = Q_{t-\delta}$ and $m = M_t^Q$,

$$H = \tfrac{\phi}{2}Q^2 + \tfrac{\eta}{2}u^2 + u\,I - \alpha\,u - u\,p^Q - \kappa_\alpha\,\alpha\,p^\alpha - \sigma_Q\,u\,q^{Q,Q} + \sigma_\alpha\,q^{\alpha,\alpha}. \tag{36}$$

Since the inventory at time t enters I_s for $s \in [t, t+\delta]$ through the current term, the point-delay term, and the distributed term, the inventory costate solves

$$\begin{aligned} -\mathrm{d}p_t^Q = \Big(&\phi Q_t - G(0)u_t + G(\delta)\mathbb{E}_t[u_{t+\delta}]\mathbf{1}_{\{t+\delta\le T\}} \\ &+ \int_0^\delta K_I(a)\mathbb{E}_t[u_{t+a}]\mathbf{1}_{\{t+a\le T\}}\,\mathrm{d}a\Big)\mathrm{d}t - q_t^Q\cdot\mathrm{d}W_t, \qquad p_T^Q = \kappa Q_T, \end{aligned} \tag{37}$$

which prices the impact created now against one's own expected future trading, while the signal costate solves the linear BSDE $-\mathrm{d}p_t^\alpha = (-u_t - \kappa_\alpha p_t^\alpha)\,\mathrm{d}t - q_t^\alpha\cdot\mathrm{d}W_t$, $p_T^\alpha = 0$, and does not enter the control. Because the diffusion is state-independent, no martingale component enters the adjoint drifts, and the fill-noise component $q^{Q,Q}$ enters the stationarity condition directly instead. The second-order source is $\mathrm{diag}(\phi, 0)$ with terminal datum $\mathrm{diag}(\kappa, 0)$, and the fill noise still generates the control curvature $\frac{1}{2}\Pi_t^{QQ}\sigma_Q^2 u^2$ in the generalized block.

Recovered control. At an interior optimum, stationarity in the first-order pair reads $\eta\, u_t^* + I_t^\delta - \alpha_t - p_t^Q - \sigma_Q q_t^{Q,Q} = 0$, and substituting $q_t^{Q,Q} = \zeta_t^{Q,Q} - \Pi_t^{QQ}\sigma_Q u_t^*$ and projecting onto the box gives

$$u_t^* = \mathrm{clip}\left(\frac{\alpha_t + p_t^Q + \sigma_Q \zeta_t^{Q,Q} - I_t^\delta}{\eta + \sigma_Q^2\, \Pi_t^{QQ}},\ 0,\ \bar{u} \right). \tag{38}$$

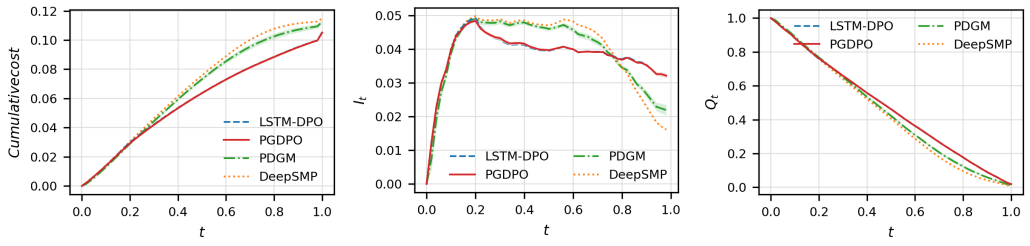
The trader sells faster when the signal or the shadow value of inventory is high, slows down against outstanding impact, and treats fill uncertainty as an increase of the effective trading cost from η to $\eta + \sigma_Q^2 \Pi_t^{QQ}$, with the shifted input $\sigma_Q \zeta_t^{Q,Q}$ pricing the fill-noise covariance. Since part of the memory reaches the control through I_t^δ rather than through the costate, identical current states with different trading histories demand different trades. In the experiments this block is assembled on the mesh as in (16).

Impact convention and reference. On the mesh the impact is defined pre-trade over realized fills, $\Delta L_j = u_j h + \sigma_Q u_j \Delta W_j^Q$ and $I_k = \sum_{j=1}^{\min(H_h,k)} G(jh)\, \Delta L_{k-j}$, so that $\partial_{u_k} I_k = 0$ as in (33), and this convention is shared by every method. We do not use an optimizer reference for this constrained benchmark, and Table 1 reports paired cost gaps to the Stage-I actor.

Table 7: **Problem 4:** model parameters and experimental settings.

Description	Symbol	Value
Horizon / impact window / mesh	$T,\ \delta,\ (h, H_h, N_h)$	$1.0,\ 0.2,\ (0.02,\ 10,\ 50)$
Initial inventory / signal	$q_0,\ \alpha_0$	$1.0,\ 0$
Fill-noise loading	σ_Q	0.10
Signal reversion / volatility	$\kappa_\alpha,\ \sigma_\alpha$	$2.0,\ 0.15$
Impact magnitude / decay	$\gamma,\ \rho_G$	$0.75,\ 15$
Cost weights (inventory / effort / terminal)	$\phi,\ \eta,\ \kappa$	$0.10,\ 0.10,\ 10$
Trading-rate box	$[0, \bar{u}]$	$[0, 1.2]$
Stage-I updates	—	3000
Projection intervals	$\mathcal{N}_p, \mathcal{N}_q, \mathcal{N}_\Pi$	$[-0.5, 3.0],\ [-1.5, 0.25],\ [9.5, 10.5]$

Protocol and settings. The mesh is that of Problem 1, $h = 0.02$, $H_h = 10$, $N_h = 50$. Stage I trains for 3000 updates on the inventory buffer, squashed logistically onto the box, and Stage II clips to the intervals of Table 7 before the exact box solve. The PDGM head is convex in the current inventory Q, and DeepSMP again uses five-step blocks.



(a) Cumulative cost (b) Transient impact I_t^δ (c) Inventory Q_t

Figure 9: **Problem 4:** averages over the evaluation paths, five-seed band. The averaged trading rate is the P4 panel of Figure 4.

Results. Table 1 reports the paired comparison to the Stage-I actor. PGDPO improves on the actor by 2.5×10^{-4}, or 0.24% of its cost, with $p = 0.001$, and both equation-based baselines are worse than the actor by 7–9% of the cost with $p < 0.001$. Figure 9 shows where the difference comes from: PGDPO leaves the trading-rate ceiling at $t = \delta$, while PDGM and DeepSMP stay on it until about $t = 0.5$, so they liquidate more completely, with a mean terminal inventory of 0.014 and 0.011 against 0.020 for PGDPO, but trade into their own transient impact and pay for it in the running cost.

A.5 Problem 5: Torque-limited speed tracking with nonlinear friction and delayed feedback

Model and cost. Let $r(t)$ be a deterministic speed command defined also on $[-\delta, 0]$, and with rotational inertia I_r write $\beta = 1/I_r$, $\sigma_b = \sigma_0/I_r$, and $\gamma = \gamma_\tau/I_r$. A fixed base loop applies the torque $s(t, v_{t-\delta})$ on the lagged speed error and a smooth Stribeck friction law F_{fr} resists motion,

$$s(t,y) = -\bar{\tau}_b \tanh\Big(\frac{y - r(t-\delta)}{v_b}\Big), \quad F_{\mathrm{fr}}(v) = c_v v + \big[\tau_c + (\tau_s - \tau_c)e^{-(v/v_s)^2}\big]\tanh(v/\varepsilon_v), \tag{39}$$

and the optimized input is a bounded correction torque $u_t \in \mathcal{U} = [-u_{\max}, u_{\max}]$ added to the base loop,

$$\begin{aligned} \mathrm{d}v_t &= \beta[s(t, v_{t-\delta}) + u_t - F_{\mathrm{fr}}(v_t)]\,\mathrm{d}t + \sigma_b\,\mathrm{d}W_t^{(0)} + \gamma[s(t, v_{t-\delta}) + u_t]\,\mathrm{d}W_t^{(1)}, \\ v_t &= \varphi(t), \qquad t \in [-\delta, 0], \end{aligned} \tag{40}$$

with independent $W^{(0)}, W^{(1)}$ and a constant initial history $\varphi \equiv V_0$. The delayed state enters the drift, the actuator noise, and the cost through the base torque, and since $u_{\max} < \bar{\tau}_b$ the correction cannot cancel the base loop over its full range, which keeps the delay operative. The cost is

$$J(u) = \mathbb{E}\Big[\int_0^T \Big\{\tfrac{Q_v}{2}(v_t - r(t))^2 + \tfrac{R_\tau}{2}\big[s(t, v_{t-\delta}) + u_t\big]^2\Big\}\mathrm{d}t + \tfrac{Q_T}{2}(v_T - r(T))^2\Big], \tag{41}$$

where the torque term charges the total commanded effort and the command r is a prescribed trajectory, not an optimizer, so no control error is reported.

Hamiltonian and adjoint. With $y = v_{t-\delta}$, $s = s(t, y)$, and $q = (q^{(0)}, q^{(1)})$,

$$H(t, v, y, u, p, q) = \tfrac{Q_v}{2}(v - r(t))^2 + \tfrac{R_\tau}{2}(s+u)^2 + \beta p\big[s + u - F_{\mathrm{fr}}(v)\big] + \sigma_b q^{(0)} + \gamma(s+u)q^{(1)}, \tag{42}$$

so that $H_u = R_\tau(s+u) + \beta p + \gamma q^{(1)}$ and $H_y = s_y(t,y)H_u$ with $s_y = \partial_y s$, and the costate solves

$$\begin{aligned} -\mathrm{d}p_t = \Big[&Q_v(v_t - r(t)) - \beta F_{\mathrm{fr}}'(v_t)p_t \\ &+ \mathbf{1}_{\{t+\delta \le T\}}\mathbb{E}_t\Big[s_y(t+\delta, v_t)\{R_\tau \tau_{t+\delta} + \beta p_{t+\delta} + \gamma q_{t+\delta}^{(1)}\}\Big]\Big]\mathrm{d}t \\ &- q_t^{(0)}\,\mathrm{d}W_t^{(0)} - q_t^{(1)}\,\mathrm{d}W_t^{(1)}, \qquad p_T = Q_T(v_T - r(T)), \end{aligned} \tag{43}$$

with $\tau_t = s(t, v_{t-\delta}) + u_t$ the total torque. The anticipated term carries delayed cost, drift, and diffusion contributions at once. The second-order source has $H_{vv} = Q_v - \beta p F_{\mathrm{fr}}''(v)$ and a nonzero H_{yy}, so a scalar physical state does not remove the current–history cross effects of the delay family, and only the current-injection block Π_t enters recovery.

Table 8: **Problem 5:** model parameters and experimental settings. The model is a declared regularized-friction benchmark, not an identified motor.

Description	Symbol	Value
Horizon / delay / mesh	T, δ, (h, H_h, N_h)	1.0, 0.2, (0.02, 10, 50)
Rotational inertia	I_r	0.25 ($\beta = 4$)
Viscous, Coulomb, static friction	c_v, τ_c, τ_s	0.10, 0.05, 0.08
Stribeck speed / smoothing width	v_s, ε_v	0.40, 0.15
Base torque limit / speed scale	$\bar{\tau}_b$, v_b	0.50, 0.50
Torque noise (baseline / actuator)	σ_0, γ_τ	0.03, 0.15 ($\sigma_b = 0.12$, $\gamma = 0.6$)
Correction bound	$u_{\max}$	0.40
Running / terminal weights	Q_v, R_τ, Q_T	1.0, 0.10, 2.0
Speed command	$r(t)$	$0.3[1 + \tanh((t - 0.4)/0.15)]$
Initial history	$\varphi \equiv V_0$	$V_0 \sim \mathrm{U}[-0.05, 0.05]$
Stage-I updates	—	3000
Projection intervals	$\mathcal{N}_p, \mathcal{N}_q, \mathcal{N}_\Pi$	$[-0.5, 0.5]$, $[-1, 1]$, $[0, 2.5]$

Recovered control. The box variational inequality in the first-order pair is equivalent to

$$u_t^* = \mathrm{clip}_{\mathcal{U}}\Big(-s_t - \frac{\beta p_t + \gamma q_t^{(1)}}{R_\tau}\Big), \tag{44}$$

and with $\sigma_t^{\mathrm{ref}} = (\sigma_b, \gamma(s_t + u_t^{\mathrm{ref}}))$ and $\zeta_t = q_t - \Pi_t \sigma_t^{\mathrm{ref}}$ the control block of the generalized Hamiltonian is $\frac{1}{2} D_t u^2 + C_t u$ with

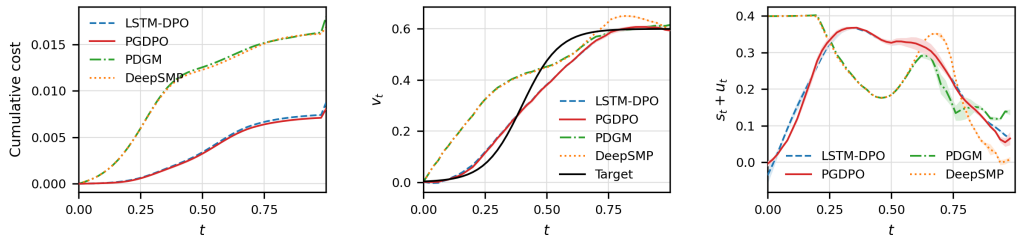
$$D_t = R_\tau + \gamma^2 \Pi_t, \qquad C_t = D_t\, s_t + \beta p_t + \gamma \zeta_t^{(1)}, \tag{45}$$

whose box minimizer is $\mathrm{clip}_{\mathcal{U}}(-C_t/D_t)$ when $D_t > 0$, and otherwise the better of the two endpoints, as in the exact box minimization of (16). At the optimal anchor $u_t^{\mathrm{ref}} = u_t^*$ it agrees with (44) through $q_t^{(1)} = \zeta_t^{(1)} + \gamma \Pi_t (s_t + u_t^*)$. In the experiments this block is assembled on the mesh as in (16).

Table 9: **Problem 5:** cost decomposition on the evaluation paths, five-seed mean.

Algorithm	Running tracking	Running torque	Terminal tracking	Total J
PGDPO	3.35×10^{-3}	3.80×10^{-3}	0.92×10^{-3}	8.07×10^{-3}
LSTM-DPO	3.66×10^{-3}	3.78×10^{-3}	1.29×10^{-3}	8.73×10^{-3}
DeepSMP	11.65×10^{-3}	4.53×10^{-3}	0.68×10^{-3}	16.86×10^{-3}
PDGM	11.63×10^{-3}	4.74×10^{-3}	1.60×10^{-3}	17.97×10^{-3}

Protocol and settings. All methods share the Euler–Maruyama discretization of (40) on the mesh $h = 0.02$, $H_h = 10$, $N_h = 50$, with the lagged speed read from the history buffer as v_{k-H_h}, and Table 1 reports the cost and the paired gap to the Stage-I actor and Table 9 the cost decomposition. Stage I trains through a $\tanh$ squashing for 3000 updates, and Stage II clips to the intervals of Table 8. Here the PDGM head is convex in the current speed, and DeepSMP keeps five-step blocks.



(a) Cumulative cost (b) Speed v_t and command $r(t)$ (c) Total torque $s_t + u_t$

Figure 10: **Problem 5:** averages over the evaluation paths, five-seed band, with the speed command in black. The single-path correction torque u_t is the P5 panel of Figure 4.

Results. We do not use an optimizer reference for this benchmark, so the comparison in Table 1 is again paired against the Stage-I actor. PGDPO improves on the actor by 6.6×10^{-4}, or 7.6% of its cost, with $p < 0.001$, while both equation-based baselines cost roughly twice the actor with $p < 0.001$. Table 9 locates the differences. The Stage-II gain comes from tracking, the running and terminal tracking costs falling by 8.5% and 29% against a torque cost higher by 0.5%, and the excess of the baselines is a threefold running tracking cost. Figure 10 shows the averaged speed, torque, and cumulative cost, where the equation-based baselines apply larger total torque early, run ahead of the command, and accumulate most of their excess cost before $t = 0.5$, while PGDPO and its actor follow the command with a smaller lag and separate only in the cost, and the single-path correction torque of Figure 4 shows the boundary contact and switching that the strictly interior parameterization of Stage I smooths. Table 10 separates batch-one recovery latency from batched throughput, whereas Table 2 reports only the amortized recovery time over the full evaluation set. Batch-one latency decreases strongly with the remaining continuation horizon: the per-GPU median is 1.73 to 1.79 s at $k = 0$ and 0.030 to 0.031 s at $k = 49$. All recorded batch-one calls exceed the model step $h = 0.02$ s, so these measurements characterize simulated closed-loop evaluation rather than real-time hardware control.

Table 10: **Problem 5:** PGDPO decision timing on eight saved histories at each of the decision indices $k \in \{0, 12, 25, 37, 49\}$, corresponding to $t/T \in \{0, 0.24, 0.50, 0.74, 0.98\}$ for $T = 1$ and $h = 0.02$, at the main budget of Appendix A. One training seed is measured on each RTX A6000 GPU after two warm-up calls and with five measured repetitions. Entries are ranges across the five per-GPU statistics, not confidence intervals. For batch size one, p50 and p95 denote the median and 95th percentile of the synchronous wall time over the 200 recorded calls per GPU. For batch size eight, the reported time per decision is the batch wall time divided by eight and is therefore an amortized throughput measure rather than a single-decision latency. Timing includes the reference action, branch simulation, differentiation, nested regression, projection, the mesh solve, and native host and device transfers, and excludes checkpoint loading and physical-environment advancement.

Batch	p50 latency (s)	p95 latency (s)	Amortized time per decision (s)	Throughput (decisions/s)
1	0.821–0.888	1.738–1.799	0.885–0.896	1.12–1.13
8	—	—	0.337–0.341	2.94–2.96

B Additional experiments

This appendix examines the accuracy of the recovery inputs, the dependence on the mesh, and the sensitivity to the delay length, together with two delay-related ablations of Stage II, an instantaneous-memory continuation at every delay and a re-entry-detached recovery at the main delay, on the two problems with an exact same-mesh reference, Problem 1 with $n = 1$ and Problem 2 with $n = 5$, using the models and training settings of Appendix A. It also checks the current-block extraction of Stage II against full-buffer differentiation on all five benchmarks, compares PGDPO with a Monte Carlo current-control search at matched forward simulation work on the two reference problems, and measures the contribution of the diffusion-curvature term of the recovery block on all five benchmarks. The branch budget is swept over $N_{\mathrm{mc}} \in \{128, 512, 2048\}$ with $(N_{\mathrm{out}}, N_{\mathrm{in}}) = (16, 8)$, $(64, 8)$, and $(256, 8)$, the mesh over $h \in \{0.05, 0.02, 0.01\}$ with $H_h = \delta/h$, and the delay over $\delta \in \{0.1, 0.2, 0.4\}$, with the remaining factors held at the main values $N_{\mathrm{mc}} = 2048$, $h = 0.02$, and $\delta = 0.2$. Changing the mesh retrains the actor and changing the delay changes the problem, so the sweeps vary the factors jointly with the warm start rather than in isolation. The results for LSTM-DPO, PGDPO, the recovery variants, and the Monte Carlo search use five training seeds, the current-block diagnostic of Appendix B.4 uses the seed-1 actor only, the estimator accuracy check and the exact-reference input check of Appendix B.3 use five independent Monte Carlo replicates at fixed states, and the exact-input mesh check is deterministic.

B.1 Estimator accuracy

Along the mesh Riccati reference the population recovery inputs (p, ζ, Π) of (11) are available in closed form, the costate along the reference and the current block of the fixed-control curvature of the continuation payoff, computed by the exact same-mesh recursions rather than as derivatives of the Riccati value, so the accuracy of the Stage-II estimates can be measured directly.

Table 11: **Estimator accuracy:** relative RMSE of the estimated recovery inputs against their exact same-mesh targets at 80 fixed states on 16 reference trajectories, with (n, m) the physical state and control dimensions, averaged over five independent Monte Carlo replicates as the branch budget N_{mc} grows.

Problem	(n, m)	N_{mc}	$\widehat{p}$	$\widehat{\zeta}$	$\widehat{\Pi}$
P1	$(1, 1)$	128	0.070	0.048	0.039
		512	0.033	0.021	0.019
		2048	0.016	0.011	0.009
P2	$(5, 5)$	128	0.087	0.086	0.061
		512	0.043	0.041	0.030
		2048	0.022	0.024	0.016

The estimates are evaluated at the 80 states that 16 reference trajectories occupy at the decision times $t \in \{0, 0.24, 0.5, 0.74, 0.98\}$. Table 11 reports the relative root-mean-square error of the estimated $\widehat{p}$, $\widehat{\zeta}$, and $\widehat{\Pi}$, the pooled RMSE divided by the pooled root-mean-square magnitude of the exact target over the same states and entries, averaged over five independent Monte Carlo replicates of the branch samples at the same states, as the branch budget grows. Every error falls by close to a factor of two for each fourfold increase of N_{mc}. For the branch means $\widehat{p}$ and $\widehat{\Pi}$ this is the $N_{\mathrm{mc}}^{-1/2}$ rate of Proposition C.25, whereas for the nested estimate $\widehat{\zeta}$ it is an empirical rate under the fixed inner budget $N_{\mathrm{in}} = 8$, and at every budget the squared error of the mean of the five replicates accounts for between 10% and 25% of the empirical mean squared error, with the remainder given by the dispersion of the replicates around that mean.

On the same budgets the control error of the deployed PGDPO control is 0.059, 0.054, and 0.053 in Problem 1 and 0.075, 0.073, and 0.072 in Problem 2, so the step from $N_{\mathrm{mc}} = 512$ to the main budget $N_{\mathrm{mc}} = 2048$ lowers it by only 2% and 1% although the input errors keep falling.

B.2 Mesh refinement

Two checks look at the discretization from different sides. The first removes estimation and learning altogether. On a common set of 80 smooth histories and decision times, with the current control displaced from the optimum by 0.2 in each coordinate and the future reference controls frozen, the exact same-mesh inputs are supplied to the Stage-II solve, and the minimizer of the assembled quadratic (16) is compared with that of the exact discrete continuation objective under the same history, anchor, and frozen future controls, computed by the exact same-mesh recursions. The two quadratics have the same gradient at the anchor by construction of the assembly, and they differ in the coefficient of the quadratic term, which (16) builds from the current curvature block where the exact objective carries the curvature at the next mesh point. Table 12 shows that the absolute error of the recovered control and the relative error of the coefficient of the quadratic term in the current control both decrease in proportion to h, while the first-order control terms of the assembled and of the exact objective agree at the anchor up to floating-point precision on every mesh. This is a test of the mesh dependence of the local solve, not of the full discretization error of the learned pipeline.

Table 12: **Mesh dependence with exact inputs:** absolute RMSE of the control recovered from the exact same-mesh inputs against the minimizer of the exact discrete continuation objective, relative error of the quadratic coefficient in the current control, and the mismatch between the first-order control terms of the assembled and the exact objectives at the anchor, on 80 smooth histories and decision times with displaced anchors and frozen future controls, with (n, m) the physical state and control dimensions.

Problem	(n, m)	h	Control RMSE	Curvature error	First-order mismatch at anchor
P1	$(1, 1)$	0.05	1.00×10^{-2}	4.8×10^{-2}	1.3×10^{-16}
		0.02	4.02×10^{-3}	1.9×10^{-2}	1.2×10^{-16}
		0.01	2.01×10^{-3}	0.9×10^{-2}	2.0×10^{-16}
P2	$(5, 5)$	0.05	7.17×10^{-3}	4.2×10^{-2}	5.6×10^{-17}
		0.02	2.87×10^{-3}	1.6×10^{-2}	5.0×10^{-17}
		0.01	1.44×10^{-3}	0.8×10^{-2}	5.7×10^{-17}

The second check keeps the full pipeline and asks whether the improvement of PGDPO over LSTM-DPO survives refinement. Both stages are run at each h, and the control values generated on each method's coarse simulated histories are held fixed within each coarse interval while the state dynamics are re-simulated on one common fine mesh with $h_{\mathrm{ref}} = 0.0025$, whose Brownian increments sum to the coarse increments of the same path. Neither the actor nor Stage II is re-evaluated during the replay. The control nRMSE uses the fine-mesh reference control evaluated at the replayed state of each method, the costs are evaluated on the same fine-grid paths, and Table 13 reports the paired gap of PGDPO to LSTM-DPO. Table 13 shows that the control error of both methods decreases with h and that PGDPO has the lower control error and the lower paired cost on every mesh and for every training seed, in all thirty seed and mesh combinations. Refining the common replay mesh from $h_{\mathrm{ref}} = 0.0025$ to 0.00125 changes each five-seed mean paired gap by less than 6×10^{-6}. These are

Table 13: **Refinement on a common fine mesh:** LSTM-DPO and PGDPO trained and deployed at mesh h and re-simulated on the common mesh $h_{\mathrm{ref}} = 0.0025$, five-seed mean $\pm$ SD, with (n, m) the physical state and control dimensions. Control nRMSE is measured against the fine-mesh reference control evaluated at the replayed state of each method, and ΔJ is the paired total cost gap of PGDPO to LSTM-DPO on the same paths.

Problem	(n, m)	h	Control nRMSE ($\times 10^{-2}$)		ΔJ vs LSTM-DPO ($\times 10^{-3}$)
			LSTM-DPO	PGDPO	
P1	$(1, 1)$	0.05	23.40 ± 0.04	22.63 ± 0.05	-1.980 ± 0.174
		0.02	13.82 ± 0.05	13.37 ± 0.04	-0.599 ± 0.186
		0.01	10.44 ± 0.54	9.41 ± 0.63	-1.608 ± 0.383
P2	$(5, 5)$	0.05	22.36 ± 0.07	21.87 ± 0.05	-0.927 ± 0.317
		0.02	16.59 ± 0.24	15.06 ± 0.15	-1.505 ± 0.241
		0.01	14.10 ± 0.20	11.74 ± 0.12	-2.480 ± 0.200

refinement results on a common numerical reference rather than a certificate of the continuous-time optimum.

B.3 Delay sensitivity and Stage-II ablations

Protocol. Table 14 varies the delay of Problem 1 and of Problem 2 at $n = 5$ with the mesh fixed, so that the history length $H_h = \delta/h$ grows from 5 to 20, and in Problem 2 the kernel is rescaled with the delay so that $\rho\delta = 2.5$ is preserved. Each delay uses its own Stage-I actors, so the sweep changes the problem and the warm start together.

At every delay the table adds an *instantaneous-memory* variant of Stage II that tests whether a history-aware reference policy makes the delay structure of the Stage-II continuation redundant. Only the continuation that Stage II simulates and differentiates is changed: in Problem 1 the point delay $Y_t = X_{t-\delta}$ is replaced by X_t, and in Problem 2 the normalized memory $M_k = \sum_{j=0}^{H_h} w_j^h X_{k-j}$ is replaced by X_k, with all coefficients unchanged and the substituted model differentiated consistently. The variant uses the same trained actors, the same branch budgets, and the same set of 256 evaluation paths as PGDPO at that delay, detaches the actor under state differentiation in the same way, and, like every method, is evaluated in the original delayed environment rather than in the substituted one. Since the variant alters the continuation model and its sensitivities together, it does not isolate the contribution of any single delay re-entry derivative.

At the main delay $\delta = 0.2$, Table 15 therefore adds a *re-entry-detached* variant that tests the combined effect of removing the differentiation through the delayed reads while keeping the forward continuation model fixed. The reads of strictly past states through the delay channels are detached from the computational graph, while the propagation through the current state and the transport of the history buffer are retained, including the current memory tap of Problem 2, so that the forward values of the continuation at the same anchor, noise, and future controls are unchanged and the recovery inputs $\widehat{p}$, $\widehat{\zeta}$, and $\widehat{\Pi}$ are re-estimated under the modified differentiation. The three recovery variants at this delay share the trained actors, the branch budgets, and the evaluation set of Table 14 at $\delta = 0.2$. The learned-actor Stage-II runs of this subsection use the estimated adjoint inputs without the finite-range projections specified in Appendix A, and the curvature symmetrization and the anchored reconstruction of $\widehat{q}^{\,\mathrm{anc}}$ and $\widehat{\zeta}$ are unchanged.

Control error. PGDPO lowers the control error of its Stage-I reference by 38 to 43% in Problem 1 and by 27 to 29% in Problem 2 at every tested delay. The instantaneous-memory variant separates from it increasingly with the delay. In Problem 1 its control error exceeds that of PGDPO by 27%, 77%, and 171% at $\delta = 0.1$, 0.2, and 0.4, it no longer improves on the Stage-I reference on average at $\delta = 0.2$, and at $\delta = 0.4$ it is worse than that reference for every seed, by 63% on average. In Problem 2 the two variants have close mean errors at $\delta = 0.1$, and the excess of the instantaneous-memory variant grows to 2.6% at $\delta = 0.2$ and 11% at $\delta = 0.4$, while both variants remain below the Stage-I error at every delay. The spreads in Table 14 are standard deviations across the five training seeds on the fixed evaluation set, computed for the difference columns from the per-seed

paired differences, and its p-values come from a two-sided crossed bootstrap test of a zero difference between PGDPO and the instantaneous-memory variant that also accounts for path sampling. Each of the 10^4 replicates resamples the training seeds and the evaluation paths independently, applies the same selected paths to every seed and method, recomputes the control nRMSE of each selected seed from its resampled paths before averaging the paired differences, and the test inverts the basic bootstrap interval of the centered replicate errors. The p-values are not adjusted for multiplicity, and values below 10^{-3} are displayed as $< 10^{-3}$. This joint test resolves the control-error difference between the two variants at every delay in Problem 1 and at $\delta = 0.2$ and 0.4 in Problem 2, but not at $\delta = 0.1$ in Problem 2.

Table 14: **Delay sensitivity and instantaneous-memory ablation** in Problem 1 with $n = 1$ and Problem 2 with $n = 5$, five-seed mean $\pm$ SD, with subscripts L, P, and I for LSTM-DPO, PGDPO, and the instantaneous-memory variant. Panel A reports the control nRMSE e against the exact same-mesh optimizer and the paired difference $e_P - e_I$, Panel B the paired total cost gaps $\Delta J_{A,B} = J_A - J_B$, negative when A is cheaper, and the 95% CI and p-value of the joint bootstrap test of the text for the difference between PGDPO and the variant.

Panel A: control nRMSE ($\times 10^{-2}$)

Problem	δ	LSTM-DPO	PGDPO	Instantaneous	$e_P - e_I$	95% CI	p
P1	0.1	11.95 ± 0.18	7.35 ± 0.20	9.36 ± 0.14	-2.00 ± 0.29	$[-2.82, -1.29]$	$< 10^{-3}$
	0.2	9.44 ± 0.29	5.34 ± 0.17	9.45 ± 0.27	-4.11 ± 0.21	$[-5.04, -3.17]$	$< 10^{-3}$
	0.4	7.36 ± 0.28	4.43 ± 0.18	12.01 ± 0.49	-7.58 ± 0.38	$[-8.75, -6.25]$	$< 10^{-3}$
P2	0.1	9.43 ± 0.49	6.92 ± 0.24	6.93 ± 0.24	-0.01 ± 0.01	$[-0.03, +0.01]$	0.45
	0.2	9.78 ± 0.39	7.19 ± 0.31	7.38 ± 0.29	-0.19 ± 0.06	$[-0.25, -0.14]$	$< 10^{-3}$
	0.4	10.29 ± 0.24	7.26 ± 0.21	8.07 ± 0.23	-0.81 ± 0.06	$[-0.89, -0.73]$	$< 10^{-3}$

Panel B: paired total cost gaps ($\times 10^{-3}$)

Problem	δ	$\Delta J_{P,L}$	$\Delta J_{I,L}$	$\Delta J_{P,I}$	95% CI	p
P1	0.1	-3.09 ± 0.26	-2.20 ± 0.23	-0.89 ± 0.04	$[-2.60, +1.69]$	0.43
	0.2	-1.61 ± 0.32	-0.66 ± 0.28	-0.96 ± 0.10	$[-2.54, +1.38]$	0.35
	0.4	-0.66 ± 0.32	$+0.08 \pm 0.26$	-0.74 ± 0.14	$[-2.70, +1.12]$	0.45
P2	0.1	-1.47 ± 0.32	-1.51 ± 0.31	$+0.04 \pm 0.01$	$[-0.16, +0.24]$	0.67
	0.2	-1.64 ± 0.17	-1.61 ± 0.15	-0.03 ± 0.04	$[-0.40, +0.34]$	0.88
	0.4	-2.28 ± 0.19	-1.90 ± 0.18	-0.38 ± 0.04	$[-1.12, +0.35]$	0.31

Realised cost. The paired cost gap of PGDPO to LSTM-DPO is negative in each of the thirty seed and delay combinations, with the per-seed gaps ranging from -0.2×10^{-3} to -3.4×10^{-3} in Problem 1 and from -1.1×10^{-3} to -2.5×10^{-3} in Problem 2, and in Problem 2 its magnitude increases with the delay. In Problem 1 the cost of PGDPO is below that of the instantaneous-memory variant for all fifteen seed and delay combinations, although the size of this difference is not monotone in the delay, and at $\delta = 0.4$ the variant does not lower the cost of the Stage-I reference on average. In Problem 2 the mean cost difference between the two variants is $+4.1 \times 10^{-5}$, -3.1×10^{-5}, and -3.8×10^{-4} at the three delays, so that the instantaneous-memory variant is slightly cheaper at the shortest delay and PGDPO is cheaper for every seed at the longest. The joint test resolves the cost gap of PGDPO to LSTM-DPO at every delay in Problem 2, but resolves the cost difference between the two variants at no delay in either problem. Across the tested delays, explicit delay modeling in the Stage-II continuation therefore yields a statistically resolved control-error advantage in five of the six settings, whereas the paired cost difference between the two variants is not resolved at the current sample size, which does not establish equivalence of their expected costs.

Re-entry-detached recovery at $\delta = 0.2$. Table 15 compares the three recovery variants at the main delay with the joint bootstrap test of Table 14 applied to every pair. PGDPO and the re-entry-detached variant lower the control error of the Stage-I reference by 42 to 43% in Problem 1 and by 27 to 28% in Problem 2, and the joint test resolves neither their control-error difference, with $p = 0.55$ in Problem 1 and $p = 0.13$ in Problem 2, nor their cost difference, with $p = 0.57$ and $p = 0.34$, which does not establish equivalence of the two variants. Both have a lower control error than the

instantaneous-memory variant, with $p < 10^{-3}$ in each of the four comparisons, while none of the three pairwise cost differences is resolved in either problem. In these two settings, both variants that retain the delayed continuation are therefore more accurate than the instantaneous-memory variant, while the joint analysis does not resolve an additional control-accuracy or cost benefit from retaining the delayed re-entry derivatives, which separates the approximation of the continuation from the removal of its derivative paths at this delay.

On the probe set of Table 11 at $N_{\mathrm{mc}} = 2048$, re-entry-detached differentiation under the exact same-mesh optimizer control gives relative RMSE of 57% and 54% for $\widehat{\Pi}$ and 70% and 71% for $\widehat{\zeta}$ in Problems 1 and 2, whereas differentiation through the full recovery graph gives 0.9% and 1.6% for $\widehat{\Pi}$ and 1.1% and 2.4% for $\widehat{\zeta}$, averaged over the same five Monte Carlo replicates. These diagnostics assess fixed-control input accuracy under the exact reference, rather than the learned-actor deployment performance compared above.

Correction alignment and magnitude at shared initial histories. For the $\delta = 0.2$ comparison in Problem 1 and in Problem 2 at $n = 5$, we examine the saved controls at $k = 0$ of the runs of Table 15, where the 256 evaluation histories are shared by the two variants for each of the five training seeds, so that each problem has $1{,}280$ paired queries and no additional training or rollouts are involved. At each query let w be the common Stage-I control, $u^\star$ the exact same-mesh optimizer control at that history, and u the recovered control, and write $v = u^\star - w$ for the oracle correction and $c = u - w$ for the applied correction. Concatenating all paired queries and control coordinates into single vectors, we report the pooled alignment $\cos \angle(v, c) = \langle v, c\rangle/(\|v\|_2\|c\|_2)$ and the pooled magnitude ratio $\|c\|_2/\|v\|_2$, which is a ratio of root-mean-square magnitudes rather than an average of querywise ratios. Table 16 shows that both corrections are short relative to the oracle correction at these initial histories. Full recovery has the higher alignment in both problems but the smaller magnitude, whereas the re-entry-detached variant applies a larger and less aligned correction and has a slightly lower initial control nRMSE. Alignment alone therefore does not determine the remaining control error $\|v - c\|_2$, which depends on the magnitude of the correction as well. This descriptive initial-time comparison does not identify the cause of the unresolved control-error and cost differences over the full rollouts.

B.4 Current-block extraction versus full-buffer differentiation

Protocol. Corollary C.2 states that the current block of the fixed-control curvature is obtained by differentiating with respect to the current state alone, and Remark C.3 counts the dense output and the Hessian–vector queries that this saves. Table 17 checks both numerically on the five benchmarks. For each problem we take the Stage-I actor of training seed 1, one evaluation path, and the initial, middle, and final decision times. Each of these queries has four independent future-noise branches that extend to the terminal time. On every branch the frozen actor generates a branch-specific control sequence, which is held fixed under state differentiation and reused in both computations. The gradient and the Hessian of the continuation payoff are then computed twice, once with respect to the current-state coordinates only and once with respect to every physical coordinate of the history buffer, from which the corresponding current block is extracted by the injection of Corollary C.2. The two computations are compared before any projection, in single precision, with one Hessian row per reverse pass.

Dimensions count differentiated coordinates rather than the physical state dimension. In Problem 3 the current computation differentiates the scalar infection coordinate I_k, whereas the full computation differentiates the raw prefix $(I_0, \dots, I_k)$ with M_0 and the time held fixed, so the full dimension is 1, 26, and 50 at the initial, middle, and final decision queries $k = 0$, 25, and 49, and only the initial query has equal differentiation dimensions. In Problem 4 only the inventory direction enters the control curvature, so the current computation has one direction against the eleven buffered inventory values and the current signal. In Problem 5 the clock coordinate is excluded from both computations while the delayed physical feedback remains differentiated.

Times are wall-clock means over the three queries for the re-simulation and differentiation of the continuation after the control sequences have been generated, with one measurement per query and computation after a warm-up call. Memory entries are the largest increase in peak allocated GPU memory over the same operation. Neither includes the regression for ζ, the projection, the local solve, or the recursive deployment.

Table 15: **Recovery variants at** $\delta = 0.2$ in Problem 1 with $n = 1$ and Problem 2 with $n = 5$, with subscripts L, P, D, and I for LSTM-DPO, PGDPO, the re-entry-detached variant, and the instantaneous-memory variant. The three Stage-II variants use no adjoint-input clipping and share the trained actors and the evaluation set of Table 14, and LSTM-DPO is the unrefined reference policy. Panel A reports the five-seed mean $\pm$ SD of the control nRMSE e and of the paired total cost gap $\Delta J_{\cdot,L}$ to LSTM-DPO, and Panel B the pairwise differences of the three recovery variants with the 95% CI and p-value of the joint bootstrap test of the text, unadjusted for multiplicity.

Panel A: method summaries

Statistic	Problem	Method			
		LSTM-DPO	PGDPO	Re-entry-detached	Instantaneous
Control nRMSE ($\times 10^{-2}$)	P1	9.44 ± 0.29	5.34 ± 0.17	5.46 ± 0.21	9.45 ± 0.27
	P2	9.78 ± 0.39	7.19 ± 0.31	7.00 ± 0.28	7.38 ± 0.29
$\Delta J_{\cdot,L}$ ($\times 10^{-3}$)	P1	—	-1.61 ± 0.32	-1.87 ± 0.39	-0.66 ± 0.28
	P2	—	-1.64 ± 0.17	-2.29 ± 0.32	-1.61 ± 0.15

Panel B: pairwise joint contrasts

Problem	Pair (A, B)	$e_A - e_B$ ($\times 10^{-2}$)			$\Delta J_{A,B}$ ($\times 10^{-3}$)		
		Difference	95% CI	p	Difference	95% CI	p
P1	(P, D)	-0.12	$[-0.32, +0.19]$	0.55	$+0.25$	$[-0.85, +1.09]$	0.57
	(P, I)	-4.11	$[-5.04, -3.17]$	$< 10^{-3}$	-0.96	$[-2.54, +1.38]$	0.35
	(D, I)	-3.99	$[-5.14, -2.87]$	$< 10^{-3}$	-1.21	$[-3.36, +2.12]$	0.40
P2	(P, D)	$+0.18$	$[-0.05, +0.40]$	0.13	$+0.65$	$[-0.78, +1.95]$	0.34
	(P, I)	-0.19	$[-0.25, -0.14]$	$< 10^{-3}$	-0.03	$[-0.40, +0.34]$	0.88
	(D, I)	-0.37	$[-0.57, -0.18]$	$< 10^{-3}$	-0.68	$[-1.85, +0.57]$	0.27

Table 16: **Correction alignment and magnitude at shared initial histories** at $k = 0$ and $\delta = 0.2$ in Problem 1 with $n = 1$ and Problem 2 with $n = 5$. The alignment and the magnitude ratio pool the five training seeds, the 256 shared histories per seed, and all control coordinates, and the last column is the control nRMSE at the initial histories rather than the rollout statistic of Table 15. Entries are descriptive statistics without confidence intervals or tests.

Problem	Method	$\cos \angle(v, c)$	$\|c\|_2 / \|v\|_2$	Initial control nRMSE ($\times 10^{-2}$)
P1	PGDPO	0.940	0.175	7.36
	Re-entry-detached	0.624	0.482	6.98
P2	PGDPO	0.889	0.215	11.27
	Re-entry-detached	0.734	0.379	10.59

Table 17: **Current-block extraction versus full-buffer differentiation** on one evaluation path of the seed-1 actor at three decision times with four branches each, with Problem 2 at $n = 5$. Coordinates are the differentiated state coordinates, the entry ratio compares the dense Hessian sizes, differences are maxima over the twelve branch-level comparisons, times are means over the three queries of the re-simulation and differentiation of the continuation, the ratio is that of the mean times before rounding, and memory is the largest baseline-subtracted peak GPU allocation in KiB. For Problem 3 the full coordinate counts and the entry ratios are listed in the order $k = 0, 25, 49$, and time and memory use the same three-query aggregation as the other rows.

Problem	Coordinates cur. / full	Entry ratio full / cur.	Max. abs. diff. Hessian / gradient	Time (ms) cur. / full	Time ratio full / cur.	Peak Δ memory cur. / full (KiB)
P1	1 / 11	121	0 / 0	21.0 / 92.6	4.42	56.0 / 85.5
P2	5 / 55	121	0 / 0	232.1 / 1676.3	7.22	164.0 / 228.5
P3	1 / (1, 26, 50)	(1, 676, 2500)	4.77×10^{-7} / 0	54.4 / 466.0	8.56	271.0 / 276.0
P4	1 / 12	144	0 / 0	25.1 / 124.0	4.95	105.5 / 126.5
P5	1 / 11	121	0 / 0	100.5 / 540.1	5.38	683.0 / 911.5

Results. The directly computed current blocks agree with the blocks extracted from the full-buffer Hessians to the stored single-precision values in fourteen of the fifteen queries, the initial query in Problem 3 differs by 4.77×10^{-7}, and the first-order blocks agree exactly in all fifteen.

The ratio of dense output entries is $(H_h + 1)^2 = 121$ in Problems 1, 2, and 5, where the direction-count ratio is $H_h + 1 = 11$ as in Remark C.3. In Problem 4 the inventory-only extraction uses one direction against twelve full coordinates, so the direction ratio is 12 and the entry ratio is 144. At the sampled initial and midpoint queries the full-buffer Hessians of Problems 1, 2, and 5 have nonzero entries throughout, whereas in Problem 4 only the current inventory diagonal is nonzero, so that case compares the extraction with a dense assembly of a structurally sparse matrix.

The ratio of the mean re-simulation and differentiation times of the full to the current computation is 4.42 in Problem 1, 7.22 in Problem 2 at $n = 5$, 4.95 in Problem 4, and 5.38 in Problem 5. These ratios are smaller than the corresponding direction-count ratios, and the measurements include the state replay and the derivative-graph work in addition to the Hessian-row extraction, whose separate contributions they do not isolate. In Problem 3 the ratio of the three-query mean times is 8.56. The ratios of the baseline-subtracted peak GPU allocations are 1.20 to 1.53 in Problems 1, 2, 4, and 5 and 1.02 in Problem 3, which are not ratios of total process memory or of the full online recovery. The entry-count and direction-count ratios therefore describe reductions in stored curvature and in Hessian-row queries, respectively, not proportional reductions in total time or memory, and the three decision times are distinct inputs rather than repeated measurements of one input.

B.5 Comparison with Monte Carlo current-control search

Protocol. The ablations of Appendix B.3 change what Stage II differentiates. This subsection instead compares the adjoint-based recovery with a derivative-free alternative, a Monte Carlo search over the current control that uses the same trained actor and the same delayed continuation model as PGDPO, under a forward-work-matched allocation and, separately, a validation-time-calibrated allocation. At each decision the search proposes a finite set of candidate controls around the actor output, which is itself the first candidate. Each candidate is applied for one step, after which the frozen actor supplies the controls over the remaining horizon, the candidates share the same future-noise branches, and the candidate with the smallest branch-averaged continuation cost is deployed. The deployed policy is evaluated recursively on the original delayed dynamics with the same five actors and the same 256 paired evaluation paths as PGDPO.

Two one-round search profiles are available, nine candidates at proposal scale 0.5 and seventeen candidates at scale 1.0. In Problem 1 both profiles place the candidates on a symmetric grid of spacing 0.125 around the actor control, so the larger profile widens the range without refining it. In Problem 2 the non-actor candidates use antithetic random directions at radius 0.5 or 1.0 from the actor control, so neither profile includes arbitrarily small nonzero corrections. For each trained actor and allocation the profile is selected by the recursive cost on four independent validation paths, and the selected profile and branch schedule are frozen before evaluation.

Two allocations are compared. The forward-work-matched allocation F chooses at every decision time the largest number of branches per candidate whose simulated state updates do not exceed those of PGDPO at its main budget, counting the direct branches, the antithetic outer steps, and the nested inner continuations of PGDPO, and it uses 99.87 to 99.92% of the PGDPO work in every seed. The time-calibrated allocation T instead chooses the branch counts so that the decision time of the search, measured on one validation history at each of the decision times $k \in \{0, 25, 49\}$ with one warm-up call and the median of three repetitions, lies within 25% of the PGDPO decision time at those three queries, with the counts interpolated in between. This calibration is met at all three queries in every seed, but it does not equalize the total deployment time, and the T schedules use 3.57 and 6.90 times the forward work of PGDPO in Problems 1 and 2. Forward-work ratios and total decision times are five-seed means, whereas validation decision-time ratios are ranges across seeds of the per-seed medians over the three calibration queries. The timers exclude environment advancement, reference evaluation, and reporting. The PGDPO runs of this subsection use the projection intervals of Appendix A. The two allocations share the same PGDPO and Stage-I evaluation rollouts for each of the five training seeds, which are not additional independent actors, and differences are tested with the crossed seed and path bootstrap of Table 14.

Results. Table 18 reports the comparison. Under the forward-work-matched allocation the control nRMSE of PGDPO is 0.0534 against 0.193 for the search in Problem 1 and 0.0719 against 0.154 in Problem 2 at $n = 5$, and the joint test resolves both differences with $p < 10^{-3}$. The paired cost difference of PGDPO to the search is -4.0×10^{-3} in Problem 1, which is not resolved with $p = 0.10$, and -6.0×10^{-3} in Problem 2, which is resolved with $p = 0.011$. Under the time-calibrated allocation the mean search errors are 0.141 and 0.123, PGDPO retains the lower control error with $p < 10^{-3}$ in both problems, and neither cost difference is resolved at the 5% level.

Table 18: **Monte Carlo current-control search versus PGDPO** in Problem 1 with $n = 1$ and Problem 2 with $n = 5$, with L the Stage-I actor, P PGDPO, and S_F and S_T the search under the forward-work-matched and the time-calibrated allocation of the text. Panel A reports the five-seed mean $\pm$ SD of the control nRMSE e and of the total cost J, Panel B the paired differences with the 95% CI and p-value of the joint bootstrap test of Appendix B.3, unadjusted for multiplicity, and Panel C the forward-simulation work and the per-decision time of the search relative to PGDPO, the latter as the range over seeds of the per-seed median at the three calibration queries, together with the five-seed mean total decision time of the search over the evaluation set.

Panel A: method summaries

Problem	Method	Control nRMSE ($\times 10^{-2}$)	Total cost J ($\times 10^{-3}$)
P1	Stage I	9.44 ± 0.29	223.47 ± 0.66
	PGDPO	5.34 ± 0.17	221.86 ± 0.38
	Search S_F	19.27 ± 3.36	225.86 ± 2.45
	Search S_T	14.11 ± 1.72	221.92 ± 2.35
P2	Stage I	9.78 ± 0.39	394.46 ± 0.21
	PGDPO	7.19 ± 0.31	392.82 ± 0.22
	Search S_F	15.37 ± 2.15	398.78 ± 2.63
	Search S_T	12.30 ± 1.13	396.16 ± 2.73

Panel B: pairwise joint contrasts

Problem	Pair (A, B)	$e_A - e_B$ ($\times 10^{-2}$)			$\Delta J_{A,B}$ ($\times 10^{-3}$)		
		Diff.	95% CI	p	Diff.	95% CI	p
P1	(P, S_F)	-13.93	$[-16.70, -10.10]$	$< 10^{-3}$	-4.00	$[-8.93, +1.10]$	0.10
	(S_F, L)	$+9.84$	$[+3.55, +14.70]$	$< 10^{-3}$	$+2.39$	$[-2.52, +8.11]$	0.33
	(P, S_T)	-8.77	$[-9.96, -7.11]$	$< 10^{-3}$	-0.06	$[-4.98, +3.98]$	0.96
	(S_T, L)	$+4.67$	$[+0.55, +7.68]$	0.016	-1.55	$[-5.84, +4.62]$	0.51
P2	(P, S_F)	-8.18	$[-9.73, -6.71]$	$< 10^{-3}$	-5.96	$[-9.37, -1.75]$	0.011
	(S_F, L)	$+5.59$	$[+4.17, +7.07]$	$< 10^{-3}$	$+4.32$	$[+0.04, +7.82]$	0.047
	(P, S_T)	-5.11	$[-6.07, -4.19]$	$< 10^{-3}$	-3.34	$[-6.46, +0.19]$	0.063
	(S_T, L)	$+2.52$	$[+1.70, +3.37]$	$< 10^{-3}$	$+1.70$	$[-2.08, +5.10]$	0.33

Panel C: search budget relative to PGDPO

Problem	Search	Forward-work ratio	Validation decision-time ratio	Total decision time (min)
P1	S_F	0.999	0.58–0.61	43.1
	S_T	3.567	0.98–1.01	146.6
P2	S_F	0.999	0.59–0.61	43.9
	S_T	6.903	0.98–1.01	283.4

The search has a higher control nRMSE than the Stage-I actor in all four settings, with $p \leq 0.016$. Its paired cost is also higher than that of the actor in Problem 2 under the forward-work-matched allocation at the unadjusted 5% level, with $p = 0.047$, and the other three cost differences to the actor are not resolved. Exact conditional scoring with the actor control among the candidates would prevent an increase in expected cost relative to the actor, but would not guarantee a lower control nRMSE, and finite-sample scoring removes that expected-cost guarantee. The present data do not separate the effects of the scoring noise and of the proposal resolution on these results.

Table 19: **Diffusion-curvature ablation** on the five benchmarks, with Problem 2 at $n = 5$, with P PGDPO and V the variant with the diffusion-curvature correction removed. The variant replaces the diffusion terms of (16) by their first-order part at the anchor and retains the estimation and projection procedures, the drift-curvature correction, and the original dynamics. Panel A reports the five-seed mean $\pm$ SD of the control nRMSE e, available only with an optimizer reference, and of the total cost J, and Panel B the paired differences $e_V - e_P$ and $\Delta J_{V,P} = J_V - J_P$, positive when the variant is worse, with the 95% CI and p-value of the joint bootstrap test of Appendix B.3, unadjusted for multiplicity.

Panel A: method summaries

Problem	Method	Control nRMSE ($\times 10^{-2}$)	Total cost J ($\times 10^{-3}$)
P1	Stage I	9.44 ± 0.29	223.47 ± 0.66
	PGDPO	5.34 ± 0.17	221.86 ± 0.38
	Diffusion-curvature off	8.10 ± 0.58	220.99 ± 0.83
P2	Stage I	9.78 ± 0.39	394.46 ± 0.21
	PGDPO	7.19 ± 0.31	392.82 ± 0.22
	Diffusion-curvature off	14.33 ± 3.54	401.15 ± 5.30
P3	Stage I	7.06 ± 0.43	575.74 ± 0.15
	PGDPO	0.64 ± 0.04	575.18 ± 0.03
	Diffusion-curvature off	0.89 ± 0.09	575.19 ± 0.05
P4	Stage I	—	105.40 ± 0.02
	PGDPO	—	105.15 ± 0.07
	Diffusion-curvature off	—	105.15 ± 0.09
P5	Stage I	—	8.73 ± 0.09
	PGDPO	—	8.07 ± 0.03
	Diffusion-curvature off	—	8.38 ± 0.08

Panel B: paired joint contrasts

Problem	$e_V - e_P$ ($\times 10^{-2}$)			$\Delta J_{V,P}$ ($\times 10^{-3}$)		
	Diff.	95% CI	p	Diff.	95% CI	p
P1	$+2.76$	$[+1.64, +3.84]$	$< 10^{-3}$	-0.862	$[-3.988, +2.136]$	0.58
P2	$+7.14$	$[+4.38, +10.23]$	$< 10^{-3}$	$+8.331$	$[+3.127, +13.459]$	0.001
P3	$+0.25$	$[+0.18, +0.31]$	$< 10^{-3}$	$+0.003$	$[-0.087, +0.094]$	0.97
P4	—	—	—	$+0.001$	$[-0.034, +0.034]$	0.90
P5	—	—	—	$+0.309$	$[+0.153, +0.447]$	$< 10^{-3}$

For the forward-work-matched allocation, the per-seed median ratio of the search to the PGDPO decision time across the three validation queries ranges from 0.58 to 0.61, and the individual query ratios vary more widely. Within this comparison the structured recovery of PGDPO therefore recovers the control more accurately than a one-round finite-candidate search at nearly the same forward simulation work, with a paired cost advantage resolved in Problem 2, which does not establish superiority over general trajectory-optimization methods.

B.6 Contribution of the diffusion-curvature correction

Protocol. The comparisons above change the continuation, its differentiation, or the use of the online simulation. This subsection instead leaves the adjoint-estimation and projection procedures of PGDPO unchanged and removes one term of the recovery block. In the mesh objective (16), with its projected input estimates, the diffusion terms $\langle \zeta_k, \sigma_k(u) \rangle_F + \frac{1}{2} \operatorname{tr}\big(\sigma_k(u)^\top \Pi_k \, \sigma_k(u)\big)$ are replaced by the first-order term $\langle \widehat{q}_k^{\mathrm{anc},\mathcal{N}}, \sigma_k(u) \rangle_F$, which has the same gradient at the anchor u_k^{ref} and differs from the original terms by the anchored quadratic $\frac{1}{2} \operatorname{tr}\big(\Delta\sigma_k(u)^\top \widehat{\Pi}_k^{\mathcal{N}} \Delta\sigma_k(u)\big)$ with $\Delta\sigma_k(u) = \sigma_k(u) - \sigma_k(u_k^{\mathrm{ref}})$ up to a control-independent constant. The estimation and projection of $\widehat{p}$, $\widehat{q}^{\mathrm{anc}}$, and $\widehat{\Pi}$, the drift term $\frac{h}{2} \Delta b_k(u)^\top \Pi_k \Delta b_k(u)$, and the controlled diffusion of the simulator are unchanged, so the variant is neither a curvature-free estimator nor a change of the dynamics. It is

Table 20: **Local diagnostics for the diffusion-curvature ablation**, with the runs of Table 19. Ratio and boundary entries are five-seed means ± SD, Defined is the percentage of all recovery queries that pass the denominator test, boundary rates use all recovery queries, and — marks unconstrained controls. No unbounded local problem or nonpositive supplied control curvature was recorded in these runs.

Problem	Method	Curvature/linear ratio ($\times 10^{-2}$)	Defined (%)	Boundary (%)
P1	PGDPO	37.5186 ± 0.0004	99.998	—
	Diffusion-curvature off	155.74 ± 0.01	100.000	—
P2	PGDPO	40.28 ± 0.54	99.994	—
	Diffusion-curvature off	224.25 ± 14.27	100.000	—
P3	PGDPO	2.45 ± 0.03	96.292	54.31 ± 0.05
	Diffusion-curvature off	2.77 ± 0.03	96.494	54.91 ± 0.08
P4	PGDPO	9.25 ± 0.32	91.833	28.26 ± 2.58
	Diffusion-curvature off	12.16 ± 0.40	92.870	29.17 ± 2.59
P5	PGDPO	16.55 ± 0.07	94.542	23.57 ± 0.30
	Diffusion-curvature off	28.28 ± 0.07	95.648	26.42 ± 0.28

also not a surrogate for DeepSMP, whose deployed control is coupled to its own learned first-order adjoints through a stationarity condition that retains the control-dependent diffusion.

For the same history, anchor, and supplied estimates, the two objectives have the same anchor gradient and differ in the supplied curvature, and during the recursive evaluation each variant re-estimates its inputs along its own history. For each problem the two variants use the same five Stage-I checkpoints, the same 256 paired evaluation initial histories and Brownian paths, the main budget $(N_{\mathrm{mc}}, N_{\mathrm{out}}, N_{\mathrm{in}}) = (2048, 256, 8)$, and the projection intervals of Appendix A, with Problem 2 at $n = 5$. Control errors are available only for the three problems with an optimizer reference, and differences are tested with the crossed seed and path bootstrap of Table 14.

Results. Table 19 reports the comparison. Removing the term increases the control nRMSE in all three reference problems, by 52%, 99%, and 39% of the PGDPO value in Problems 1, 2, and 3, with joint $p < 10^{-3}$ in each case. The paired cost increase is resolved in Problem 2, by 8.3×10^{-3} with $p = 0.001$, and in Problem 5, by 3.1×10^{-4} with $p < 10^{-3}$, but not in Problems 1, 3, or 4. In Problem 1 the variant has a slightly lower mean cost than PGDPO, without a resolved difference. Relative to the Stage-I actor, the variant is worse in both control error and cost in Problem 2, with $p < 10^{-3}$ and $p = 0.022$, whereas in Problem 5 it still lowers the cost of the actor with $p = 0.009$ while retaining only part of the mean improvement of PGDPO.

Local diagnostic quantities. To quantify the size of the diffusion-curvature term, we evaluate at each deployed control u_k of a variant and its anchor u_k^{ref} the ratio

$$\frac{\left|\frac{1}{2}\operatorname{tr}\left(\Delta\sigma_k(u_k)^\top \widehat{\Pi}_k^{\mathcal{N}} \Delta\sigma_k(u_k)\right)\right|}{\left|\left\langle \nabla_u \mathcal{H}_k^{\mathrm{mesh}}(u_k^{\mathrm{ref}}),\, u_k - u_k^{\mathrm{ref}} \right\rangle\right|},$$

where the mesh gradient uses the input estimates recorded at that query. For fixed inputs this gradient is unchanged at the anchor when the correction is removed, and for the variant the numerator is the omitted term evaluated at its deployed control. Writing d_k for the absolute value of the denominator, a ratio is retained when $d_k > 64\,\epsilon_{\mathrm{mach}} \max(1, d_k)$, with ϵ_{mach} the machine epsilon of the coefficient precision. Ratios are averaged within each seed over the retained queries and then summarized across seeds in Table 20, together with the share of retained queries and the share of all recovery queries at which the deployed control lies on the box boundary.

The ablated runs were feasible and had small KKT residuals for their supplied mesh objectives. For the same anchor and supplied estimates, including or removing the correction preserves the anchor gradient but changes the control Hessian and hence the finite update, including potentially its direction for multidimensional controls. The intervention acts on the recovery objective rather than on the

differentiation graph, so it is complementary to the re-entry-detached comparison of Appendix B.3. These results establish a control-accuracy benefit of the correction in the three reference problems and a cost benefit in Problems 2 and 5 at the tested settings, not a uniform cost advantage across the benchmarks.

C PROOFS FOR THEOREM 5.1

PROOF GUIDE: CLAIMS, ASSUMPTIONS, AND DEPENDENCIES

The table separates exact computation, continuous-time identification, and the conditions needed for recovery. References link to the full statements and proofs; the assumptions column is a guide to those statements, not a replacement for them.

Claim	**Assumptions and scope**	**Proof route / where to read**
Exact discrete adjoints and current-block extraction	Fixed-grid C^2 differentiability and square integrability; Assumption C.9. No refinement hypothesis.	Lemmas C.10–C.11 give Theorem C.1(a). Corollary C.2 (p. 38) gives the current-only reduction.
Identification of the continuous effective curvature	Assumption C.4; fixed-control current-injection variations of one reference policy.	Moment bounds and restarted duality, Lemmas C.5–C.6, lead to Lemma C.7 (p. 44).
Population input convergence under growing history	Assumptions C.4, C.8, C.9; curvature integrability (79); product and compact-control conditions in Proposition C.23.	(p, q): Proposition C.18; Π: Proposition C.21; ζ and objective differences: Proposition C.23. Appendix C.8 (p. 57) assembles Theorem C.1(b)–(d), i.e., Theorem 5.1(b).
Finite-branch and regression consistency	Fixed h, k; Assumptions C.9, C.24, and the regression conditions of Proposition C.26. Process limits are separate.	Appendix C.9 (p. 57): Propositions C.25, C.26; Corollary C.27 gives anchor cancellation and ridge bias. Remark C.28 states the limit order.
Recovered-policy error Proposition 5.2	Assumptions D.1–D.6; projected inputs and square-integrable errors as in Corollary D.15; matching mesh-reference anchor.	Input-to-control stability, Proposition D.14 and Corollary D.15, plus performance transfer, Proposition D.17, give Proposition D.18(i)–(ii) (p. 67).
Conditional improvement of J Theorem 5.3	Comparison-segment conditions of Lemma D.21; supplied strong concavity; the theorem's margin and defect conditions. No quadratic-growth hypothesis.	Lemma D.9 and Proposition D.10 support the first variation and remainder in Lemma D.21 (p. 69); the theorem's proof follows it.

Two separate local certificates. Corollary D.16 concerns the supplied local objective. Lemma D.20 adds a local-gradient Lipschitz condition and provides the true-input local-loss branch in Proposition D.18(iii). Neither certificate alone proves improvement of J.

Reading the dependencies. The recovery results concern the localized model and paired controls, with Assumption C.4 imposed policywise. Recursive deployment additionally requires Assumption D.7. Quadratic growth converts a reference performance gap into the policy-error rate; the one-update proof instead combines a first-variation remainder with the anchored variational inequality. Thus the three main results are not a single implication chain.

C.1 FULL STATEMENTS OF THE MAIN-TEXT RESULTS

This appendix identifies the conditional reward sensitivities that Stage II uses. Appendix C.2 recalls the continuous delay-adjoint systems, Appendices C.3 to C.7 connect them to fixed-control differentiation on refining history graphs, and Appendix C.9 treats finite-sample estimation, with the proof of the full theorem assembled in Appendix C.8. The population inputs used in Section 5 are the block extractions (68) of the fixed-control OL-BPTT sensitivities.

Scope and notation. Taking the coefficient consistency in Assumption C.8 as a separate hypothesis on the forward approximation, we prove refinement consistency of the required delay-adjoint inputs for one fixed reference under the stated additional assumptions. Finite-branch limits are treated separately in Appendix C.9. The starting points are the correspondence between automatic differentiation and adjoints established by Griewank & Walther (2008); Alexe & Sandu (2009); Chen et al. (2018); Li et al. (2020); Jeon et al. (2026) and the delay family of Meng et al. (2025). Throughout Appendices C and D, L^r means $L^r([0,T]\times\Omega, \mathrm{d}t\otimes\mathrm{d}\mathbb{P})$ unless another domain is shown, and $L^r_{\mathbb{F}}$ is its progressively measurable subspace. Vectors use Euclidean norms and matrix-valued processes use the Frobenius norm $\|\cdot\|_F$, and bare $\|\cdot\|$ on linear maps, including Gram matrices and their inverses, denotes the induced operator norm. Write $\mathbb{S}^n$ for symmetric $n\times n$ matrices, I_n for the identity, and e_i for the standard basis. For $r\geq 2$,

$$\|Y\|_{\mathcal{S}^r} := (\mathbb{E}\sup_{t\leq T}|Y_t|^r)^{1/r}, \qquad \|Z\|_{\mathbb{H}^r} := \left[\mathbb{E}\left(\int_0^T \|Z_t\|_F^2\,\mathrm{d}t\right)^{r/2}\right]^{1/r}.$$

Thus $\mathbb{H}^2 = L^2$, but $\mathbb{H}^4$ and L^4 are different norms, and $\|Y\|_{\mathcal{S}^\infty} := \|\sup_{t\leq T}|Y_t|\|_{L^\infty(\Omega)}$. Unspecified constants C may change between estimates, and their uniformity is stated with each result. No convergence of the full second-order family is claimed.

Theorem C.1 (Full form of Theorem 5.1)**.** *Suppose Assumptions C.4, C.8, and C.9 hold, where part (a) uses only the fixed-grid C^2 differentiability and square-integrability conditions together with Assumption C.9, and parts (c) and (d) add the additional conditions stated with them. Let $(\widetilde{\Lambda}^h_k, \widetilde{\mathbf{P}}^h_k)$ be the fixed-control OL-BPTT sensitivities along the mesh-indexed reference control $u^{h,\mathrm{ref}}$, let $(\Lambda^h_k, \mathbf{Z}^h_k, \mathbf{P}^h_k, \mathbf{R}^h_k)$ be their adapted conditional projections, with the mesh conventions of Appendix C, and let $(p^{h,\mathrm{ref}}, q^{h,\mathrm{ref}}, \Pi^{h,\mathrm{ref}})$ and $\zeta^{h,\mathrm{ref}} = q^{h,\mathrm{ref}} - \Pi^{h,\mathrm{ref}}\sigma^{h,\mathrm{ref}}$ denote the extracted recovery inputs and their piecewise-constant extensions.*

(a) ***Exact gridwise correspondence.*** *For every fixed mesh, and without any refinement assumption, the raw sensitivities satisfy the exact first- and second-order reverse recursions of Lemma C.10, and their conditional projections form the unique square-integrable adapted solution of the corresponding discrete buffer-adjoint system, including its one-step Brownian coefficients and orthogonal projection residuals of Lemma C.11.*

(b) ***First-order consistency.*** *As $h\downarrow 0$,*

$$\mathbb{E}\int_0^T \left(\left|p_t^{h,\mathrm{ref}} - p_t^{\mathrm{ref}}\right|^2 + \left\|q_t^{h,\mathrm{ref}} - q_t^{\mathrm{ref}}\right\|_F^2\right) dt \longrightarrow 0, \quad \mathbb{E}\left|p_T^{h,\mathrm{ref}} - p_T^{\mathrm{ref}}\right|^2 \longrightarrow 0,$$

where $(p^{\mathrm{ref}}, q^{\mathrm{ref}})$ is the solution pair of the compact anticipated adjoint system (5) *assembled in* (50)*.*

(c) ***Effective curvature consistency.*** *Under the current-curvature uniform integrability* (79) *of Appendix C.6,*

$$\mathbb{E}\int_0^T \left\|\Pi_t^{h,\mathrm{ref}} - \Pi_t^{\mathrm{ref}}\right\|_F^2 dt \longrightarrow 0,$$

where Π^{ref} is the effective diffusion curvature, characterized as the matrix representation of the current-injection quadratic form by Lemma C.7, so that for every $a, b\in\mathbb{R}^n$,

$$\mathbb{E}\int_0^T \left|a^\top\left(\Pi_t^{h,\mathrm{ref}} - \Pi_t^{\mathrm{ref}}\right)b\right|^2 dt \longrightarrow 0,$$

and in fixed dimension it suffices to verify this for the basis pairs (e_i, e_j).

(d) ***Recovery-objective consistency.*** *Assume in addition the product integrability* $\sup_h \|\Pi^{h,\mathrm{ref}}\|_{L^{4+\epsilon}} + \|\Pi^{\mathrm{ref}}\|_{L^{4+\epsilon}} < \infty$ *for some* $\epsilon > 0$*, the convergence* $\sigma^{h,\mathrm{ref}} \to \sigma^{\mathrm{ref}}$ *in* L^4*, and the compact-control uniformity condition* (ii) *of Proposition C.23. Then*

$$\mathbb{E}\int_0^T \big\|\zeta_t^{h,\mathrm{ref}} - \zeta_t^{\mathrm{ref}}\big\|_F^2\, dt \longrightarrow 0, \quad \mathbb{E}\int_0^T \sup_{u\in K} \big|\Delta\mathcal{H}_t^h(u) - \Delta\mathcal{H}_t(u)\big|\, dt \longrightarrow 0$$

for every compact $K \subset \mathcal{U}$*, that is, compact-control uniform consistency of the recovery objective in integrated* L^1*, where* $\Delta\mathcal{H}_t^h(u) := \Delta\mathcal{H}_{k_h(t)}^h(u)$ *is the mesh-level anchored generalized-Hamiltonian difference at the cell index* $k_h(t)$ *and* $\Delta\mathcal{H}_t(u) := \mathcal{H}\big(t;u,p_t^{\mathrm{ref}},\zeta_t^{\mathrm{ref}},\Pi_t^{\mathrm{ref}}\big) - \mathcal{H}\big(t;u_t^{\mathrm{ref}},p_t^{\mathrm{ref}},\zeta_t^{\mathrm{ref}},\Pi_t^{\mathrm{ref}}\big)$ *is its continuous counterpart, both defined in Appendix C.7.*

No optimality of u^{ref} is required for this convergence. The limit is its policy-indexed adjoint tuple, and when the reference is optimal it also satisfies (8). Appendix C.9 treats sampling, and Appendix D treats reference mismatch and recovery.

Corollary C.2 (Exact current-injection reduction)**.** *Under the fixed-mesh hypotheses of Theorem 5.1(a), let* $F_{k,m}^h(v) := \mathcal{J}_k^{h,(m)}(\mathsf{S}_k^h + \iota_0^h v)$*,* $v \in \mathbb{R}^n$*, with the branch noise and the realized future reference controls held fixed. Then, before input projection,*

$$\widehat{\Pi}_k^{h,\mathrm{ref}} = \frac{1}{N_{\mathrm{mc}}}\sum_m \nabla_v^2 F_{k,m}^h(0) = (\iota_0^h)^\top\Big(\frac{1}{N_{\mathrm{mc}}}\sum_m \widetilde{\mathbf{P}}_k^{h,(m)}\Big)\iota_0^h,$$

so current-state differentiation reproduces the full-buffer curvature estimator on the same branches, and for $\Delta\sigma_k^h(u) := \sigma_k^h(u) - \sigma_k^{h,\mathrm{ref}}$*,*

$$\tfrac{1}{2}\,\mathrm{tr}\big[(\iota_0^h\Delta\sigma_k^h(u))^\top \mathbf{P}_k^h(\iota_0^h\Delta\sigma_k^h(u))\big] = \tfrac{1}{2}\,\mathrm{tr}\big[\Delta\sigma_k^h(u)^\top \Pi_k^{h,\mathrm{ref}}\Delta\sigma_k^h(u)\big],$$

so the generalized-Hamiltonian correction is unchanged. Under the hypotheses of Theorem C.1(c), the population block converges to the effective curvature of the full family without a separate numerical solution of that family.

Proof. The fixed-mesh identities are the chain rule on the detached graph and Lemma C.22, and the continuous identification is Lemma C.7 with Theorem C.1(c). □

Remark C.3 (Curvature storage and computational scope). Let $d_h = n(H_h+1)$. Relative to dense full-buffer Hessian assembly, the explicitly stored curvature is $O(n^2)$ rather than $O(d_h^2) = O(n^2(H_h+1)^2)$, and a columnwise Hessian–vector-product construction on the same continuation graph queries n current directions instead of d_h: the dense-output entry-count ratio is $(H_h+1)^2$ and the query-count ratio is H_h+1, both unbounded under mesh refinement. With fixed per-decision budgets, direct and nested branching require $O\big((N_{\mathrm{mc}} + 2N_{\mathrm{out}}N_{\mathrm{in}})N_h^2\big)$ simulated Euler steps per deployed trajectory, as detailed in Appendix C.9. These counts exclude differentiation, regression, and local-solver work, and neither they nor the storage reduction imply history-independent total runtime or memory.

C.2 The continuous first- and second-order delay adjoint systems

Throughout the appendix, B is the standard d-dimensional Brownian motion of Section 3, $(\mathcal{F}_t)$ its augmented filtration, and $\langle A_1, A_2\rangle_F := \mathrm{tr}(A_1^\top A_2)$ the Frobenius inner product. The continuous-time adjoint objects below are recalled from Meng et al. (2025), specialized to state delay with terminal reward $g(X(T))$ and evaluated along an admissible reference control u^{ref}. No optimality of u^{ref} is assumed, and well-posedness of the resulting policy-indexed families is part of Assumption C.4. The results of Meng et al. (2025) are stated for cost minimization, and the displays below record the equivalent maximization form, obtained by applying them with running and terminal costs $(-f, -g)$ and negating all adjoints.

Reference jets, terminal data, and Volterra kernels. Write $b_x(s) := \partial_x b(\Xi^{\mathrm{ref}}(s)) \in \mathbb{R}^{n\times n}$ and analogously b_y, b_m and, columnwise, $\sigma_x^j, \sigma_y^j, \sigma_m^j$ for $j = 1,\dots,d$, together with $f_x, f_y, f_m \in \mathbb{R}^n$

and $\mathbf{f}(s) := (f_x(s)^\top, f_y(s)^\top, f_m(s)^\top)^\top \in \mathbb{R}^{3n}$, all evaluated along Ξ^{ref}. Because the terminal reward depends on $X(T)$ only, the lifted terminal data reduce to

$$\mathbf{g}^{(1)} := \big(g_x(X^{\mathrm{ref}}(T))^\top,\, 0,\, 0\big)^\top \in \mathbb{R}^{3n}, \quad \mathbf{g}^{(2)} := \mathrm{blkdiag}\big(g_{xx}(X^{\mathrm{ref}}(T)),\, 0_{n\times n},\, 0_{n\times n}\big) \in \mathbb{S}^{3n},$$

with vanishing point-delay and memory blocks. The lifted forward variational system uses the two-time kernels

$$\begin{aligned} A(t,s) &:= \begin{pmatrix} b_x(s) & b_y(s) & b_m(s) \\ \mathbf{1}_{\{t-s>\delta\}}\, b_x(s) & \mathbf{1}_{\{t-s>\delta\}}\, b_y(s) & \mathbf{1}_{\{t-s>\delta\}}\, b_m(s) \\ I_n & -e^{-\rho\delta} I_n & -\rho I_n \end{pmatrix}, \\ C^j(t,s) &:= \begin{pmatrix} \sigma_x^j(s) & \sigma_y^j(s) & \sigma_m^j(s) \\ \mathbf{1}_{\{t-s>\delta\}}\, \sigma_x^j(s) & \mathbf{1}_{\{t-s>\delta\}}\, \sigma_y^j(s) & \mathbf{1}_{\{t-s>\delta\}}\, \sigma_m^j(s) \\ 0 & 0 & 0 \end{pmatrix}, \end{aligned} \tag{46}$$

both $\mathbb{R}^{3n\times 3n}$-valued for $0 \le s \le t \le T$. The third row of A is the pathwise identity $\mathrm{d}M(t) = [X(t) - e^{-\rho\delta}X(t-\delta) - \rho M(t)]\,\mathrm{d}t$ from (2), and the boundary convention of the indicators is $\mathrm{d}t \otimes \mathrm{d}s$-negligible. (These two-time kernels are distinct from the one-step transition Jacobians $A_k^h, B_{k,a}^h$ of Lemma C.10.)

Assumption C.4 (Continuous reference system and adjoint regularity)**.** Let $\chi := (x, y, m)$. The coefficients b, σ and the running reward f are measurable in t, continuous in u, and twice continuously differentiable in χ. The first and second χ-derivatives of b and σ are bounded and Lipschitz in χ, uniformly in t and in u on the range of the reference controls. The second χ-derivatives of f are bounded and Lipschitz in the same uniform sense, the terminal reward g is C^2 with bounded Lipschitz second derivative, and f and f_χ, evaluated at $\chi = 0$ along the reference controls, are square integrable in (t, ω). The initial history satisfies $x_0 \in C^{\alpha_0}([-\delta, 0]; \mathbb{R}^n)$ for some $\alpha_0 > 0$.

The continuous-time reference control satisfies $u^{\mathrm{ref}} \in L^4_{\mathbb{F}}(0, T; \mathbb{R}^m)$, and the corresponding SDDE admits a unique strong solution with moments of every required order. Along this reference trajectory, the first- and second-order BSDE–BSVIE adjoint systems displayed below admit unique square-integrable adapted solutions, with effective quantities

$$(p^{\mathrm{ref}}, q^{\mathrm{ref}}, \Pi^{\mathrm{ref}}, \zeta^{\mathrm{ref}}), \quad \zeta^{\mathrm{ref}} = q^{\mathrm{ref}} - \Pi^{\mathrm{ref}}\sigma^{\mathrm{ref}}.$$

No optimality of u^{ref} is assumed. Whenever the control set, or a localization of it containing the range of the reference controls, renders the cost pair $(-f, -g)$ compliant with the standing assumption of Meng et al. (2025), this adjoint well-posedness is available by citation, and for unbounded control sets with control-coercive rewards it is a reference-path assumption, which suffices because Theorem 5.1 follows a fixed reference control. Coefficients of polynomial growth in χ would require an additional localization argument and are outside the scope of the present statements.

First-order family. The first-order adjoint family couples a backward SDE with a backward stochastic Volterra integral equation (BSVIE), whose assembled current-coordinate component is the pair (p, q) of Section 4. The anticipated equation (5) is the differential form of this assembly, obtained from Meng et al. (2025) after eliminating the auxiliary memory adjoint. The family $\mathfrak{Y} := (\eta, \psi, \mathcal{Y}, \mathcal{Z})$ consists of $\mathbb{R}^{3n}$-valued $\eta(t)$, $\mathcal{Y}(t)$ and, per Brownian index j, $\mathbb{R}^{3n}$-valued $\psi^j(t)$, $\mathcal{Z}^j(t,s)$, split into current/point/memory blocks $\eta = (\eta^0, \eta^1, \eta^2)$ and analogously for $\psi^j, \mathcal{Y}, \mathcal{Z}^j$, and all components are square integrable over their respective one- and two-time domains. Along u^{ref}, for $t \in [0, T]$,

$$\eta(t) = \mathbf{g}^{(1)} - \sum_{j=1}^{d} \int_t^T \psi^j(s)\,\mathrm{d}B^j(s), \tag{47}$$

$$\begin{aligned} \mathcal{Y}(t) = \mathbf{f}(t) + A(T,t)^\top \mathbf{g}^{(1)} + \sum_{j=1}^{d} C^j(T,t)^\top \psi^j(t) \\ + \int_t^T \Big[A(s,t)^\top \mathcal{Y}(s) + \sum_{j=1}^{d} C^j(s,t)^\top \mathcal{Z}^j(s,t)\Big]\mathrm{d}s \\ - \sum_{j=1}^{d} \int_t^T \mathcal{Z}^j(t,s)\,\mathrm{d}B^j(s), \end{aligned} \tag{48}$$

subject to the M-solution constraint

$$\mathcal{Y}(t) = \mathbb{E}\big[\mathcal{Y}(t)\big] + \sum_{j=1}^{d} \int_0^t \mathcal{Z}^j(t,s)\,\mathrm{d}B^j(s), \quad t \in [0,T]. \tag{49}$$

The assembled pair is, for $j = 1, \dots, d$,

$$\begin{aligned} p(t) &= \eta^0(t) + \eta^1(t)\,\mathbf{1}_{\{t<T-\delta\}} + \mathbb{E}_t\Big[\int_t^T \mathcal{Y}^0(s)\,\mathrm{d}s + \mathbf{1}_{\{t<T-\delta\}} \int_{t+\delta}^T \mathcal{Y}^1(s)\,\mathrm{d}s\Big], \\ q^j(t) &= \psi^{0j}(t) + \psi^{1j}(t)\,\mathbf{1}_{\{t<T-\delta\}} + \int_t^T \mathcal{Z}^{0j}(s,t)\,\mathrm{d}s + \mathbf{1}_{\{t<T-\delta\}} \int_{t+\delta}^T \mathcal{Z}^{1j}(s,t)\,\mathrm{d}s. \end{aligned} \tag{50}$$

Because the terminal reward depends on $X(T)$ only, (47) gives $\eta(t) = \mathbb{E}_t[\mathbf{g}^{(1)}]$, whence $\eta^1 \equiv \eta^2 \equiv 0$ and $\psi^{1j} \equiv \psi^{2j} \equiv 0$. The corresponding terms of (50) vanish identically, while the point and memory blocks of $(\mathcal{Y}, \mathcal{Z})$ remain active through the kernel couplings in (48).

Second-order family. The second-order adjoint family $(\mathfrak{P}, \mathfrak{Q})$, with $\mathfrak{P} := (P_1, P_2, P_3, P_4)$ and martingale family $\mathfrak{Q} := (Q_1, Q_2, Q_3, Q_4)$, consists of a backward SDE together with three coupled BSVIEs, and the two-time components carry the cross curvature between the current state and its delayed re-entries, which is why the family does not close as a single $n \times n$ backward equation. Here P_1, P_3 are $\mathbb{S}^{3n}$-valued, and P_2 and the two-time component P_4 are $\mathbb{R}^{3n\times 3n}$-valued. Let $\chi := (x, y, m)$ and let $D^2_{\chi\chi} H(r) \in \mathbb{S}^{3n}$ denote the Hessian of the first-order Hamiltonian (4) with respect to χ, evaluated at $(\Xi^{\mathrm{ref}}(r), p(r), q(r))$ with (p, q) from (50). Along u^{ref}, for $0 \le r \le T$,

$$P_1(r) = \mathbf{g}^{(2)} - \sum_{j=1}^{d} \int_r^T Q_1^j(\theta)\,\mathrm{d}B^j(\theta), \tag{51}$$

$$\begin{aligned} P_2(r) = {} & A(T,r)^\top P_1(r) + \sum_{j=1}^{d} C^j(T,r)^\top Q_1^j(r) \\ & + \int_r^T \Big[A(\theta,r)^\top P_2(\theta) + \sum_{j=1}^{d} C^j(\theta,r)^\top Q_2^j(\theta,r)\Big]\mathrm{d}\theta \\ & - \sum_{j=1}^{d} \int_r^T Q_2^j(r,\theta)\,\mathrm{d}B^j(\theta), \end{aligned} \tag{52}$$

$$\begin{aligned} P_3(r) = {} & D^2_{\chi\chi} H(r) + \sum_{j=1}^{d} C^j(T,r)^\top P_1(r)\, C^j(T,r) \\ & + \sum_{j=1}^{d} \int_r^T \Big[C^j(T,r)^\top P_2(\theta)^\top C^j(\theta,r) + C^j(\theta,r)^\top P_2(\theta)\, C^j(T,r)\Big]\mathrm{d}\theta \\ & + \sum_{j=1}^{d} \int_r^T C^j(\theta,r)^\top P_3(\theta)\, C^j(\theta,r)\,\mathrm{d}\theta + \sum_{j=1}^{d} \int_r^T \int_r^T C^j(\theta,r)^\top P_4(\theta',\theta)\, C^j(\theta',r)\,\mathrm{d}\theta\,\mathrm{d}\theta' \\ & - \sum_{j=1}^{d} \int_r^T Q_3^j(r,\theta)\,\mathrm{d}B^j(\theta), \end{aligned} \tag{53}$$

and, for $0 \le r \le \theta \le T$,

$$\begin{aligned} P_4(\theta,r) &= A(T,r)^\top P_2(\theta)^\top + \sum_{j=1}^{d} C^j(T,r)^\top Q_2^j(\theta,r)^\top + A(\theta,r)^\top P_3(\theta) \\ &\quad + \sum_{j=1}^{d} C^j(\theta,r)^\top Q_3^j(\theta,r) + \int_r^T \Big[\sum_{j=1}^{d} C^j(\theta',r)^\top Q_4^j(\theta,\theta',r) + A(\theta',r)^\top P_4(\theta,\theta')\Big]\mathrm{d}\theta' \\ &\quad - \sum_{j=1}^{d} \int_r^T Q_4^j(\theta,r,\theta')\,\mathrm{d}B^j(\theta'), \end{aligned} \tag{54}$$

with the symmetry extension

$$P_4(\theta,r) = P_4(r,\theta)^\top, \quad Q_4^j(\theta,r,\theta') = Q_4^j(r,\theta,\theta')^\top, \quad 0 \le \theta < r \le T, \tag{55}$$

subject to the adapted M-solution constraints

$$\begin{aligned} P_2(r) &= \mathbb{E}_\theta\big[P_2(r)\big] + \sum_{j=1}^{d} \int_\theta^r Q_2^j(r,\theta')\,\mathrm{d}B^j(\theta'), && 0 \le \theta \le r \le T, \\ P_3(r) &= \mathbb{E}_\theta\big[P_3(r)\big] + \sum_{j=1}^{d} \int_\theta^r Q_3^j(r,\theta')\,\mathrm{d}B^j(\theta'), && 0 \le \theta \le r \le T, \\ P_4(\theta,r) &= \mathbb{E}_{\theta'}\big[P_4(\theta,r)\big] + \sum_{j=1}^{d} \int_{\theta'}^{r\wedge\theta} Q_4^j(\theta,r,s)\,\mathrm{d}B^j(s), && 0 \le \theta' \le (\theta\wedge r) \le T. \end{aligned} \tag{56}$$

We write the Brownian-column sum in the double-integral term of (53) explicitly, consistently with the block expansion of Meng et al. (2025).

Effective curvature assembly. With the current-noise injection

$$\aleph(t,s) := \big(I_n,\ \mathbf{1}_{\{t-s>\delta\}} I_n,\ 0\big)^\top \in \mathbb{R}^{3n\times n},$$

define the maximization-form raw assembly

$$\begin{aligned} \mathcal{P}_0^{\max}(t) &:= \aleph(T,t)^\top P_1(T)\,\aleph(T,t) + \int_t^T \Big[\aleph(T,t)^\top P_2(\theta)^\top \aleph(\theta,t) + \aleph(\theta,t)^\top P_2(\theta)\,\aleph(T,t)\Big]\mathrm{d}\theta \\ &\quad + \int_t^T\!\!\int_t^T \aleph(\theta,t)^\top P_4(\theta',\theta)\,\aleph(\theta',t)\,\mathrm{d}\theta\,\mathrm{d}\theta' + \int_t^T \aleph(\theta,t)^\top P_3(\theta)\,\aleph(\theta,t)\,\mathrm{d}\theta, \end{aligned} \tag{57}$$

the sign-flip of the compact second-order matrix $\mathcal{P}_0$ of Meng et al. (2025). The matrix $\mathcal{P}_0^{\max}(t)$ is symmetric because $P_1(T)$ and $P_3(\theta)$ are symmetric, the two P_2 terms occur as a transpose pair, and the double integral is symmetric after interchanging (θ,θ') and using (55). The adapted effective curvature of Section 4 is

$$\Pi_t := \mathfrak{A}_t[\mathfrak{P}] := \mathbb{E}_t\big[\mathcal{P}_0^{\max}(t)\big] \ \in\ \mathbb{S}^n. \tag{58}$$

For every $\mathcal{F}_t$-measurable $D_t \in \mathbb{R}^{n\times d}$,

$$\mathbb{E}_t\big[\mathrm{tr}\big(D_t^\top\, \mathcal{P}_0^{\max}(t)\, D_t\big)\big] = \mathrm{tr}\big(D_t^\top\, \Pi_t\, D_t\big),$$

so the conditional projection leaves the spike-variation quadratic unchanged and is the curvature entering (8) and the Stage-II population target. Throughout, we fix a jointly measurable adapted version of $(t,\omega)\mapsto \Pi_t$, which exists because Π is the conditional expectation of the jointly measurable integrable field $\mathcal{P}_0^{\max}$, and conditional expectations of such fields admit jointly measurable versions. We adopt the normalization used in the variational expansions of Meng et al. (2025), which carries the factor $\frac{1}{2}$ in front of $\mathrm{tr}\big[\Delta\sigma_t(u)^\top\, \Pi_t\, \Delta\sigma_t(u)\big]$, and no rescaling of $\mathfrak{A}$ is made.

Current-injection variational flow and the effective quadratic form. For $t \in [0,T)$ and $a \in \mathbb{R}^n$, write $\phi^{t,a}(s) := \aleph(s,t)\,a$ and let $\mathcal{V}^{t,a}$ denote the unique solution of the linear lifted variational equation

$$\mathcal{V}^{t,a}(s) = \phi^{t,a}(s) + \int_t^s A(s,r)\,\mathcal{V}^{t,a}(r)\,\mathrm{d}r + \sum_{j=1}^{d}\int_t^s C^j(s,r)\,\mathcal{V}^{t,a}(r)\,\mathrm{d}B^j(r), \quad s \in [t,T], \tag{59}$$

driven by the current-injection profile $\aleph(\cdot,t)$: the injected direction persists in the current block, matures into the point-delay block after the lag δ, and feeds the distributed-memory block only through the kernels. For $t \le \tau \le s \le T$, the *restarted partial flow* freezes the row time and integrates only up to the clock τ,

$$\mathcal{V}^{t,a}(s;\tau) := \phi^{t,a}(s) + \int_t^\tau A(s,r)\,\mathcal{V}^{t,a}(r)\,\mathrm{d}r + \sum_{j=1}^{d}\int_t^\tau C^j(s,r)\,\mathcal{V}^{t,a}(r)\,\mathrm{d}B^j(r).$$

Thus $\mathcal{V}^{t,a}(s;s) = \mathcal{V}^{t,a}(s)$, mirroring the auxiliary process of the spike analysis in Meng et al. (2025).

Lemma C.5 (Current-injection channels and conditional moments)**.** *Under the bounded first reference jets of Assumption C.4, write $\mathcal{V}^{t,a} = (v^x, v^y, v^m)$ and set*

$$L_b = (b_x, b_y, b_m), \qquad L_\sigma^j = (\sigma_x^j, \sigma_y^j, \sigma_m^j), \qquad \kappa_{\rho,\delta} := \int_0^\delta e^{-\rho r}\,\mathrm{d}r \le \delta.$$

For the strict indicator convention of (46)*,*

$$\begin{aligned} v^y(s) &= \mathbf{1}_{\{s-t>\delta\}} v^x(s-\delta), \qquad & v^m(s) &= (\mathcal{W}_t v^x)(s), \\ (\mathcal{W}_t v)(s) &:= \int_{(s-\delta)\vee t}^{s} e^{-\rho(s-r)} v(r)\,\mathrm{d}r. \end{aligned} \tag{60}$$

Consequently the current channel is the semimartingale

$$\mathrm{d}v^x(s) = L_b(s)\mathcal{V}^{t,a}(s)\,\mathrm{d}s + \sum_{j=1}^{d} L_\sigma^j(s)\mathcal{V}^{t,a}(s)\,\mathrm{d}B^j(s), \qquad v^x(t) = a, \tag{61}$$

closed by (60)*. For every $r \ge 2$, there is a constant C_r, depending only on the jet bounds, ρ, δ, T and the dimensions, such that, for deterministic $t \le s \le s' \le T$,*

$$\mathbb{E}_t\left[\sup_{t\le u\le T} |\mathcal{V}^{t,a}(u)|^r\right] \le C_r|a|^r \quad \textit{a.s.}, \qquad \mathbb{E}|v^x(s') - v^x(s)|^r \le C_r|s'-s|^{r/2}|a|^r. \tag{62}$$

In particular, for every $0 < \alpha < 1/2$, $\mathbb{E}\|v^x\|^4_{C^\alpha([t,T])} \le C_\alpha|a|^4$. The conditional moment field is taken in a jointly measurable version, and its $L^\infty(\mathrm{d}t \otimes \mathrm{d}\mathbb{P})$ bound is the same $C_r|a|^r$. Constants are uniform over any policy class with common jet bounds.

Proof. The second row of (59) is its first row evaluated at $s-\delta$ when $s-t > \delta$, and is zero otherwise. The memory row solves $\mathrm{d}v^m = (v^x - e^{-\rho\delta}v^y - \rho v^m)\,\mathrm{d}s$, $v^m(t) = 0$. Variation of constants and the substitution $u = r - \delta$ in its delayed term give (60). Thus, with $M(s) := \sup_{t\le u\le s}|v^x(u)|$,

$$|\mathcal{V}(s)| \le (2 + \kappa_{\rho,\delta})M(s). \tag{63}$$

This reduction is essential: the stochastic integral in the lifted equation is not a martingale in its row time, whereas (61) is.

Stop v^x at $\tau_R := \inf\{s \ge t : |v^x(s)| \ge R\} \wedge T$, and put $M_R(s) := \sup_{t\le u\le s}|v^x(u\wedge\tau_R)|$. Conditional Burkholder–Davis–Gundy, Hölder and conditional Fubini give

$$F_R(s) := \mathbb{E}_t[M_R(s)^r] \le C_r|a|^r + C_r\int_t^s F_R(\ell)\,\mathrm{d}\ell.$$

A continuous version of F_R follows by bounded convergence for the stopped supremum. Gronwall and conditional monotone convergence as $R \uparrow \infty$ yield the first bound in (62). Applying the same inequalities on $[s,s']$ gives the increment bound. For $\alpha < 1/2$, choose $r > 4$ with $\alpha < 1/2 - 1/r$. The moment form of Kolmogorov–Chentsov and the supremum bound then give the asserted fourth moment of the Hölder norm. Finally, take jointly measurable conditional versions for the finitely many basis directions, use Fubini to discard a common product-measure null set, and extend by linearity. Dependence only on the common jet bounds gives the policy-uniform assertion, without selecting a simultaneous version over an uncountable policy class. □

The raw and conditional *effective quadratic forms* are

$$\begin{aligned}\widehat{\mathcal{Q}}_t(a,b) &:= (\mathcal{V}^{t,a}(T))^\top \mathbf{g}^{(2)} \mathcal{V}^{t,b}(T) + \int_t^T (\mathcal{V}^{t,a}(s))^\top D^2_{\chi\chi} H(s) \mathcal{V}^{t,b}(s)\,\mathrm{d}s, \\ \mathcal{Q}_t(a,b) &:= \mathbb{E}_t[\widehat{\mathcal{Q}}_t(a,b)].\end{aligned} \tag{64}$$

with $D^2_{\chi\chi}H(s)$ the Hamiltonian Hessian along the reference tuple $(\Xi^{\mathrm{ref}}(s), p(s), q(s))$, and the boundary bilinear form of the second-order family is

$$\begin{aligned}\mathfrak{B}_t(\phi^{t,a}, \phi^{t,b}) := \mathbb{E}_t\Big[&\phi^{t,a}(T)^\top P_1(T) \phi^{t,b}(T) \\ &+ \int_t^T \{\phi^{t,a}(T)^\top P_2(\theta)^\top \phi^{t,b}(\theta) + \phi^{t,a}(\theta)^\top P_2(\theta) \phi^{t,b}(T)\}\,\mathrm{d}\theta \\ &+ \int_t^T \phi^{t,a}(\theta)^\top P_3(\theta) \phi^{t,b}(\theta)\,\mathrm{d}\theta \\ &+ \int_t^T \int_t^T \phi^{t,a}(\theta)^\top P_4(\theta', \theta) \phi^{t,b}(\theta')\,\mathrm{d}\theta\,\mathrm{d}\theta'\Big].\end{aligned} \tag{65}$$

Lemma C.6 (Restarted conditional quadratic duality)**.** *Under Assumption C.4, after choosing jointly measurable versions, for Lebesgue-almost every $t \in [0,T)$ and all $a, b \in \mathbb{R}^n$,*

$$\mathcal{Q}_t(a,b) = \mathfrak{B}_t(\phi^{t,a}, \phi^{t,b}) \quad \mathbb{P}\text{-a.s.}$$

This is a policy-indexed linear-quadratic duality identity: no optimality of the admissible reference control is used, and the injected profiles need not arise from admissible control perturbations.

Proof. Localize first. Lemma C.5, square integrability of the adjoint family and stochastic Fubini justify the computations and removal of localization. Symmetry of P_1, P_3 also permits symmetric Q_1^j, Q_3^j by uniqueness of martingale representations, and (55) is used for P_4, Q_4. Fix a, abbreviate $V(s;\tau) = \mathcal{V}^{t,a}(s;\tau)$ and $\phi = \phi^{t,a}$, and define

$$\begin{aligned}\Psi(\tau) = V(T;\tau)^\top P_1(\tau) V(T;\tau) &+ 2\int_\tau^T V(\theta;\tau)^\top \mathbb{E}_\tau[P_2(\theta)] V(T;\tau)\,\mathrm{d}\theta \\ &+ \int_\tau^T V(\theta;\tau)^\top \mathbb{E}_\tau[P_3(\theta)] V(\theta;\tau)\,\mathrm{d}\theta + \int_{[\tau,T]^2} V(\theta;\tau)^\top \mathbb{E}_\tau[P_4(\theta',\theta)] V(\theta';\tau)\,\mathrm{d}\theta\,\mathrm{d}\theta'.\end{aligned}$$

At the endpoints, $\Psi(t) = \mathfrak{B}_t(\phi,\phi)$ because $V(s;t) = \phi(s)$ and $P_1(t) = \mathbb{E}_t[\mathbf{g}^{(2)}]$, whereas $\Psi(T) = V(T)^\top \mathbf{g}^{(2)} V(T)$.

To compute the drift, write $A_s = A(s,\tau)$, $C_s^j = C^j(s,\tau)$. Taking $\mathbb{E}_\tau$ in (52)–(54) and using (56) gives their adapted row identities, with future P_i replaced by $\mathbb{E}_\tau P_i$ and the Q_i densities unchanged. For example,

$$P_2(\tau) = A_T^\top P_1(\tau) + \sum_j C_T^{j\top} Q_1^j(\tau) + \int_\tau^T \left[A_\theta^\top \mathbb{E}_\tau P_2(\theta) + \sum_j C_\theta^{j\top} Q_2^j(\theta,\tau) \right] \mathrm{d}\theta.$$

Apply Itô and the moving-boundary Leibniz rule to the four groups of Ψ, using $\mathrm{d}_\tau V(s;\tau) = A_s V(\tau)\,\mathrm{d}\tau + \sum_j C_s^j V(\tau)\,\mathrm{d}B^j(\tau)$ and $\mathrm{d}_\tau \mathbb{E}_\tau P_i = \sum_j Q_i^j(\cdot,\tau)\,\mathrm{d}B^j(\tau)$. The $P_2(\tau)$ row cancels the terminal drift/martingale cross terms and the interior $A^\top P_2, C^\top Q_2$ terms, the $P_3(\tau)$ row cancels all quadratic diffusion terms, and the $P_4(\theta,\tau)$ row cancels the remaining cross terms from the two moving faces, paired by (55). Only the forcing $D^2_{\chi\chi}H$ remains:

$$\mathrm{d}\Psi(\tau) = -V(\tau)^\top D^2_{\chi\chi} H(\tau) V(\tau)\,\mathrm{d}\tau + \mathrm{d}M_\tau.$$

Here M is a martingale after the stated integrability argument. Integrating and conditioning at t proves the diagonal identity. Both sides are symmetric bilinear forms by linearity of the flow, so polarization gives the assertion for every (a,b). □

Lemma C.7 (Current-injection identification of the effective curvature)**.** *Under Assumption C.4,* $\mathrm{d}t \otimes \mathrm{d}\mathbb{P}$*-almost everywhere and simultaneously for all* $a, b \in \mathbb{R}^n$*,*

$$\mathcal{Q}_t(a,b) = \mathbb{E}_t\big[a^\top \mathcal{P}_0^{\max}(t)\, b\big] = a^\top\, \Pi_t\, b.$$

Consequently $\mathcal{Q}_t$ *is a symmetric bilinear form whose matrix representation is the effective diffusion curvature* $\Pi_t = \mathfrak{A}_t[\mathfrak{P}]$*. The form itself carries no rescaling, the factor* $\frac{1}{2}$ *belonging to the outer second-order expansions in which it is used.*

Proof. Insert $\phi^{t,a}(s) = \aleph(s,t)a$ into (65). Its four groups become the four terms of (57) sandwiched between a, b, hence $\mathfrak{B}_t(\phi^{t,a}, \phi^{t,b}) = \mathbb{E}_t[a^\top \mathcal{P}_0^{\max}(t) b] = a^\top \Pi_t b$. Apply Lemma C.6. For jointly measurable versions, discard the finite union of the basis-pair null sets and extend by bilinearity. The form is unhalved, and the factor $1/2$ remains in the outer spike expansion. □

C.3 Refining history graphs and mesh conventions

We write $h \equiv \Delta t$ and make the mesh index explicit throughout the appendix. Let $h \downarrow 0$ along a sequence such that

$$N_h := T/h \in \mathbb{N}, \quad H_h := \delta/h \in \mathbb{N}, \quad t_k^h := kh,$$

and let $\Delta B_k^h := B(t_{k+1}^h) - B(t_k^h)$ denote the standard Brownian increments, so that $\mathbb{E}[\Delta B_k^h] = 0$ and $\mathbb{E}[\Delta B_k^h (\Delta B_k^h)^\top] = I_d\, h$. Mesh-indexed quantities are extended to $[0,T]$ piecewise constantly through the cell index

$$k_h(t) := \min\big\{\lfloor t/h \rfloor,\, N_h - 1\big\}, \quad t \in [0,T],$$

so that $t \in J_{k_h(t)}^h$ for $t < T$, where $J_k^h := [t_k^h, t_{k+1}^h)$. Terminal quantities retain their terminal values, in particular $p_T^h = p_{N_h}^h$, and the convention $k_h(T) = N_h - 1$ applies only to cell-indexed expressions. The subscript is abbreviated to k when the argument is clear. The physical history buffer is

$$\mathsf{S}_k^h := (X_k^h, X_{k-1}^h, \dots, X_{k-H_h}^h) \in \mathbb{R}^{d_h}, \quad d_h := n(H_h + 1),$$

with coordinates $\mathsf{s} = (x^{(0)}, x^{(1)}, \dots, x^{(H_h)})$ and

$$Y_k^h = X_{k-H_h}^h, \quad M_k^h = m_h(\mathsf{S}_k^h), \quad m_h(\mathsf{s}) := \sum_{j=0}^{H_h} w_j^h\, x^{(j)}.$$

Writing $b_k^h(u) = b(t_k^h, X_k^h, Y_k^h, M_k^h, u)$ and $\sigma_k^h(u) = \sigma(t_k^h, X_k^h, Y_k^h, M_k^h, u)$, the Euler history map is

$$\mathsf{S}_{k+1}^h = \Phi_k^h(\mathsf{S}_k^h, u_k^h, \Delta B_k^h),$$

whose first block is $X_{k+1}^h = X_k^h + b_k^h(u_k^h)\, h + \sigma_k^h(u_k^h)\, \Delta B_k^h$ and whose remaining blocks perform the deterministic history shift. Set

$$r_k^h(\mathsf{s}, u) := h\, f\big(t_k^h, x^{(0)}, x^{(H_h)}, m_h(\mathsf{s}), u\big), \quad G_h(\mathsf{s}) := g(x^{(0)}).$$

For $j = 0, \dots, H_h$, let $\iota_j^h : \mathbb{R}^n \to \mathbb{R}^{d_h}$ denote insertion into the jth buffer block, and let $\pi_j^h := (\iota_j^h)^\top$ denote the corresponding block extraction. In particular, ι_0^h inserts into the current-state coordinate and $\iota_{H_h}^h$ into the point-delay coordinate. For a detached reference control $u^{h,\mathrm{ref}}$, write $\mathbb{E}_k^h[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_{t_k^h}]$ for the adapted conditional expectation. Its identification with conditioning on the conditioning state C_k^h defined in Assumption C.9 is part of that assumption. With the detached continuation payoff

$$\mathcal{J}_k^h := \sum_{\ell=k}^{N_h-1} r_\ell^h\big(\mathsf{S}_\ell^h, u_\ell^{h,\mathrm{ref}}\big) + G_h\big(\mathsf{S}_{N_h}^h\big)$$

of the reference rollout started from S_k^h, define the raw fixed-control OL-BPTT sensitivities

$$\widetilde{\Lambda}_k^h := D_{\mathsf{S}_k^h}^{\mathrm{ol}} \mathcal{J}_k^h, \quad \widetilde{\mathbf{P}}_k^h := D_{\mathsf{S}_k^h \mathsf{S}_k^h}^{2,\mathrm{ol}} \mathcal{J}_k^h.$$

Lag cells and the atom-priority endpoint convention. Set

$$r_j^h := -jh, \quad j = 0, \dots, H_h.$$

For all sufficiently fine meshes, define the interior Voronoi cells

$$I_j^h := \big[-(j+\tfrac{1}{2})h,\, -(j-\tfrac{1}{2})h\big), \quad j = 1, \dots, H_h - 1,$$

and the two endpoint strips

$$I_0^h := (-\tfrac{h}{2}, 0), \quad I_{H_h}^h := (-\delta, -\delta + \tfrac{h}{2}).$$

Under the *atom-priority convention* adopted throughout, the endpoint lags $r_0^h = 0$ and $r_{H_h}^h = -\delta$ are treated as atoms carrying the current and point-delay channels, and the endpoint strips I_0^h and $I_{H_h}^h$ are not split into further history-density degrees of freedom. The endpoint quadrature contributions carried by w_0^h and $w_{H_h}^h$ therefore remain in the Euler graph, and the resulting $O(h)$ discrepancies are absorbed into the quadrature moduli of Assumption C.8.

Assumption C.8 (Euler, quadrature, and reference-control consistency)**.** The refining meshes satisfy $N_h = T/h \in \mathbb{N}$ and $H_h = \delta/h \in \mathbb{N}$. The memory weights obey $W := \sup_h \sum_{j=0}^{H_h} |w_j^h| < \infty$ and $\max_{0 \le j \le H_h} |w_j^h| \le C_w h$ for a constant C_w independent of h, and consistently discretize the exponential kernel: for some modulus $\omega_K(h) \downarrow 0$ and some exponent $\alpha \in \big(0, \min\{\alpha_0, \tfrac{1}{2}\}\big)$, with α_0 the Hölder exponent of the initial history in Assumption C.4,

$$\Big| \sum_{j=0}^{H_h} w_j^h \, \varphi(-jh) - \int_{-\delta}^{0} e^{\rho r} \varphi(r)\, dr \Big| \le \omega_K(h)\, \|\varphi\|_{C^\alpha([-\delta,0])}$$

for every α-Hölder test path φ. Without loss of generality we enlarge ω_K so that $\omega_K(h) \ge h$, which absorbs the $O(h)$ endpoint contributions of the atom-priority convention of Appendix C.3. The weights are moreover partial-window consistent: with the same exponent α, there is a modulus $\omega_{\mathcal{W}}(h) \downarrow 0$ such that

$$\sup_{0 \le \ell \le H_h} \sup_{\substack{\varphi \in C^\alpha([-\ell h, 0]) \\ \|\varphi\|_{C^\alpha} \le 1}} \Big| \sum_{j=0}^{\ell} w_j^h \, \varphi(-jh) - \int_{-\ell h}^{0} e^{\rho r} \, \varphi(r)\, \mathrm{d}r \Big| \;\le\; \omega_{\mathcal{W}}(h); \tag{66}$$

the full-window clause is the case $\ell = H_h$ up to the endpoint convention, and cell-integrated exponential weights satisfy the clause with $\omega_{\mathcal{W}}(h) \le C(h^\alpha + h)$, as shown in Appendix C.6. The predictable piecewise-constant interpolants of the detached reference controls satisfy

$$u^{h,\mathrm{ref}} \longrightarrow u^{\mathrm{ref}} \quad \text{in} \quad L^4_{\mathbb{F}}(0, T; \mathbb{R}^m).$$

Along the reference Euler trajectories, the coefficients and their first two state-delay-memory derivatives converge in the corresponding L^2 spaces to their continuous-time counterparts.

Assumption C.9 (Detached graph and conditional branching)**.** For every k and every $\ell \ge k$, the open-loop derivative holds the future reference controls fixed:

$$D^{\mathrm{ol}}_{\mathsf{S}_k^h} u_\ell^{h,\mathrm{ref}} = D^{2,\mathrm{ol}}_{\mathsf{S}_k^h \mathsf{S}_k^h} u_\ell^{h,\mathrm{ref}} = 0,$$

while all derivatives through the physical state transition, the point-delay re-entry, and the distributed-memory re-entry remain active, and $D^{\mathrm{ol}}_{\mathsf{S}_k^h}$ acts only on the physical history buffer. The conditioning state C_k^h is the physical buffer together with every additional frozen actor variable needed to determine future reference controls such as the recurrent state of an LSTM actor, and for every square-integrable continuation functional generated by the detached reference-control graph, that is, every $\Psi_k^h \in L^2\big(\sigma(\mathsf{C}_k^h, \Delta B_k^h, \dots, \Delta B_{N_h-1}^h)\big)$,

$$\mathbb{E}[\Psi_k^h \mid \mathcal{F}_{t_k^h}] = \mathbb{E}[\Psi_k^h \mid \mathsf{C}_k^h].$$

Conditional on C_k^h, the Stage-II future branches are independent copies driven only by independent future Brownian increments.

C.4 EXACT IDENTITIES AND CONSISTENCY LEMMAS

Throughout this subsection, $\mathcal{J}_k^h$ denotes the detached continuation payoff of Appendix C.3 and $J_k^h = [t_k^h, t_{k+1}^h)$ the corresponding time cell.

Lemma C.10 (Exact fixed-control OL-BPTT identities)**.** *Define*

$$A_k^h := D^{\mathrm{ol}}_{\mathsf{S}_k^h} \Phi_k^h\big(\mathsf{S}_k^h, u_k^{h,\mathrm{ref}}, \Delta B_k^h\big), \quad B_{k,a}^h := D^{2,\mathrm{ol}}_{\mathsf{S}_k^h \mathsf{S}_k^h}\big[\Phi_k^h\big(\mathsf{S}_k^h, u_k^{h,\mathrm{ref}}, \Delta B_k^h\big)\big]_a, \quad a = 1, \dots, d_h.$$

Then, almost surely,

$$\widetilde{\Lambda}_{N_h}^h = \iota_0^h \, g_x(X_{N_h}^h), \quad \widetilde{\mathbf{P}}_{N_h}^h = \iota_0^h \, g_{xx}(X_{N_h}^h)\,(\iota_0^h)^\top,$$

and, for $k = N_h - 1, \dots, 0$,

$$\widetilde{\Lambda}_k^h = D^{\mathrm{ol}}_{\mathsf{S}_k^h} r_k^h + (A_k^h)^\top \widetilde{\Lambda}_{k+1}^h,$$

$$\widetilde{\mathbf{P}}_k^h = D^{2,\mathrm{ol}}_{\mathsf{S}_k^h \mathsf{S}_k^h} r_k^h + (A_k^h)^\top \widetilde{\mathbf{P}}_{k+1}^h A_k^h + \sum_{a=1}^{d_h} \big[\widetilde{\Lambda}_{k+1}^h\big]_a \, B_{k,a}^h,$$

where r_k^h and its derivatives are evaluated at $(\mathsf{S}_k^h, u_k^{h,\mathrm{ref}})$. All current, point-delay, and distributed-memory re-entry effects are contained exactly in A_k^h and $B_{k,a}^h$.

Lemma C.11 (Adapted conditional projections)**.** *Define*

$$\Lambda_k^h := \mathbb{E}_k^h\big[\widetilde{\Lambda}_k^h\big], \quad \mathbf{P}_k^h := \mathbb{E}_k^h\big[\widetilde{\mathbf{P}}_k^h\big],$$

with one-step Brownian coefficients

$$\mathbf{Z}_k^h := \frac{1}{h}\,\mathbb{E}_k^h\big[\Lambda_{k+1}^h (\Delta B_k^h)^\top\big], \quad \mathbf{R}_k^{h,\ell} := \frac{1}{h}\,\mathbb{E}_k^h\big[\mathbf{P}_{k+1}^h \, \Delta B_k^{h,\ell}\big], \quad \ell = 1, \dots, d.$$

The same coefficients are obtained by replacing $\Lambda_{k+1}^h, \mathbf{P}_{k+1}^h$ with their raw counterparts inside the conditional covariance. There exist residuals ξ_{k+1}^h and $\xi_{k+1}^{P,h}$ such that

$$\Lambda_{k+1}^h = \mathbb{E}_k^h[\Lambda_{k+1}^h] + \mathbf{Z}_k^h \, \Delta B_k^h + \xi_{k+1}^h, \quad \mathbf{P}_{k+1}^h = \mathbb{E}_k^h[\mathbf{P}_{k+1}^h] + \sum_{\ell=1}^{d} \mathbf{R}_k^{h,\ell} \, \Delta B_k^{h,\ell} + \xi_{k+1}^{P,h},$$

where

$$\mathbb{E}_k^h\big[\xi_{k+1}^h\big] = 0, \quad \mathbb{E}_k^h\big[\xi_{k+1}^h (\Delta B_k^h)^\top\big] = 0,$$

and analogously for $\xi_{k+1}^{P,h}$. Taking $\mathbb{E}_k^h$ in Lemma C.10 yields the adapted backward recursions

$$\Lambda_k^h = \mathbb{E}_k^h\Big[D^{\mathrm{ol}}_{\mathsf{S}_k^h} r_k^h + (A_k^h)^\top \widetilde{\Lambda}_{k+1}^h\Big], \quad \mathbf{P}_k^h = \mathbb{E}_k^h\Big[D^{2,\mathrm{ol}}_{\mathsf{S}_k^h \mathsf{S}_k^h} r_k^h + (A_k^h)^\top \widetilde{\mathbf{P}}_{k+1}^h A_k^h + \sum_{a=1}^{d_h} \big[\widetilde{\Lambda}_{k+1}^h\big]_a B_{k,a}^h\Big],$$

with the terminal values of Lemma C.10, and $(\Lambda^h, \mathbf{Z}^h, \xi^h)$ and $(\mathbf{P}^h, \mathbf{R}^h, \xi^{P,h})$ form their unique square-integrable adapted solution. Under Assumption C.9, these adapted projections coincide with the C_k^h-conditional branch averages computed in Stage II-A.

Proof of Lemmas C.10 and C.11. Apply the first and second chain rules to $\mathcal{J}_k^h = r_k^h + \mathcal{J}_{k+1}^h \circ \Phi_k^h$ on the detached graph. Conditioning gives the displayed recursions. The transition derivatives and ΔB_k^h are $\mathcal{F}_{t_{k+1}^h}$-measurable, so the tower property allows the next raw sensitivities to be replaced by their adapted projections inside these products. Orthogonal projection onto the constants and the Brownian increment gives the covariance formulas and the centered residuals, using $\mathbb{E}_k^h[\Delta B_k^h \Delta B_k^{h\top}] = h I_d$. Backward induction from the terminal values gives uniqueness, including the orthogonal residuals. Assumption C.9 identifies these conditional expectations with the population branch targets. □

Channel structure and recovery-input extraction. All coefficient jets in this paragraph are evaluated along the detached reference Euler rollout at $(t_k^h, X_k^h, Y_k^h, M_k^h, u_k^{h,\mathrm{ref}})$. Write

$$\mathbf{T}_h := \begin{pmatrix} \pi_0^h \\ \pi_{H_h}^h \\ \sum_{j=0}^{H_h} w_j^h \, \pi_j^h \end{pmatrix} \in \mathbb{R}^{3n \times d_h}, \quad \mathbf{S}_h := \iota_0^h \pi_0^h + \sum_{i=1}^{H_h} \iota_i^h \pi_{i-1}^h,$$

so that $\mathbf{T}_h \mathsf{s} = (x^{(0)}, x^{(H_h)}, m_h(\mathsf{s}))^\top$, and the Euler Jacobian decomposes exactly as

$$A_k^h = \mathbf{S}_h + \iota_0^h \Big(h \, J_{b,k}^h + \sum_{\ell=1}^{d} J_{\sigma,k}^{h,\ell} \Delta B_k^{h,\ell} \Big) \mathbf{T}_h, \quad J_{b,k}^h := D_\chi b(k), \quad J_{\sigma,k}^{h,\ell} := D_\chi \sigma^\ell(k), \tag{67}$$

with $\mathbf{S}_h \iota_0^h = \iota_0^h + \iota_1^h$, $\mathbf{S}_h \iota_j^h = \iota_{j+1}^h$ for $1 \le j < H_h$, and $\mathbf{S}_h \iota_{H_h}^h = 0$. The discrete recovery inputs are

$$p_k^h := (\iota_0^h)^\top \Lambda_k^h, \quad q_k^{h,\ell} := (\iota_0^h)^\top \mathbf{Z}_k^{h,\ell}, \quad \Pi_k^h := (\iota_0^h)^\top \mathbf{P}_k^h \, \iota_0^h, \quad \zeta_k^h := q_k^h - \Pi_k^h \, \sigma_k^{h,\mathrm{ref}}, \tag{68}$$

Here $q_k^{h,\ell}$ is column ℓ of q_k^h, and $q_k^{h,\ell,a}$ is its coordinate a. The discrete Hamiltonian jets are

$$\mathsf{H}_\kappa^h(k) := f_\kappa(k) + b_\kappa(k)^\top \mathbb{E}_k^h[p_{k+1}^h] + \sum_{\ell=1}^{d} \sigma_\kappa^\ell(k)^\top q_k^{h,\ell}, \quad \kappa \in \{x, y, m\}.$$

We also write $\widehat{p}_k^{h,\mathrm{raw}} := (\iota_0^h)^\top \widetilde{\Lambda}_k^h$ for the raw current gradient, the unconditioned counterpart of p_k^h.

Lemma C.12 (Exact lag recursions and the discrete anticipated identity)**.** *For every mesh h and $0 \le k < N_h$,*

$$\begin{aligned} \Lambda_{k,0}^h &= \mathbb{E}_k^h[\Lambda_{k+1,0}^h + \Lambda_{k+1,1}^h] + h \, \mathsf{H}_x^h(k) + h \, w_0^h \, \mathsf{H}_m^h(k), \\ \Lambda_{k,j}^h &= \mathbb{E}_k^h[\Lambda_{k+1,j+1}^h] + h \, w_j^h \, \mathsf{H}_m^h(k), \quad 1 \le j < H_h, \\ \Lambda_{k,H_h}^h &= h \, \mathsf{H}_y^h(k) + h \, w_{H_h}^h \, \mathsf{H}_m^h(k), \quad \Lambda_{N_h}^h = \iota_0^h \, g_x(X_{N_h}^h); \end{aligned}$$

consequently, following the state through successive lag slots and using that the terminal lag blocks vanish,

$$\begin{aligned} p_k^h =& \mathbb{E}_k^h[p_{k+1}^h] + h \, \mathsf{H}_x^h(k) + h \, w_0^h \, \mathsf{H}_m^h(k) + h \, \mathbf{1}_{\{k+H_h \le N_h - 1\}} \, \mathbb{E}_k^h[\mathsf{H}_y^h(k + H_h)] \\ &+ h \sum_{i=1}^{H_h} w_i^h \, \mathbf{1}_{\{k+i \le N_h - 1\}} \, \mathbb{E}_k^h[\mathsf{H}_m^h(k+i)], \end{aligned} \tag{69}$$

an exact identity at every fixed mesh, whose continuous counterpart is the compact anticipated system (5).

Proof. Substitute (67) in the raw reverse recursion, then condition. The drift term uses $\mathbb{E}_k^h p_{k+1}^h$ and the diffusion term uses $\mathbb{E}_k^h[\widehat{p}_{k+1}^{h,\mathrm{raw}} \Delta B_k^{h,\ell}] = h q_k^{h,\ell}$ by Lemma C.11. Thus

$$\Lambda_k^h = \mathbf{S}_h^\top \mathbb{E}_k^h \Lambda_{k+1}^h + h \mathbf{T}_h^\top \mathsf{H}^h(k), \qquad \mathsf{H}^h(k) := (\mathsf{H}_x^h(k)^\top, \mathsf{H}_y^h(k)^\top, \mathsf{H}_m^h(k)^\top)^\top.$$

Pair with ι_j^h and use the stated shift identities and $\mathbf{T}_h \iota_j^h = (\delta_{j0} I_n, \delta_{j H_h} I_n, w_j^h I_n)^\top$ where δ_{ij} is the Kronecker symbol. This gives the three lag recursions. Expand $\mathbb{E}_k^h \Lambda_{k+1,1}^h$ along $(k+i, i)$: each lag releases $h w_i^h \mathbb{E}_k^h \mathsf{H}_m^h(k+i)$, and reaching lag H_h also releases $h \mathbb{E}_k^h \mathsf{H}_y^h(k + H_h)$. If the terminal time is reached first, its zero lag blocks terminate the expansion. These alternatives give exactly the indicators in (69). □

Buffer tangent flow. With the product convention

$$A_{k:k}^h := I_{d_h}, \quad A_{k:m}^h := A_{m-1}^h \cdots A_k^h, \quad m > k,$$

let $\mathbf{X}_m^{h,\gamma} := \mathbf{T}_h A_{k:m}^h \gamma$ for $\gamma \in \mathbb{R}^{d_h}$ denote the channel-read tangent, which satisfies the exact channel-resolvent recursion

$$\mathbf{X}_m^{h,\gamma} = \mathbf{T}_h \mathbf{S}_h^{m-k} \gamma + \sum_{r=k}^{m-1} \mathbf{T}_h \mathbf{S}_h^{m-1-r} \iota_0^h \Big(h \, J_{b,r}^h + \sum_{\ell=1}^{d} J_{\sigma,r}^{h,\ell} \Delta B_r^{h,\ell} \Big) \mathbf{X}_r^{h,\gamma}; \tag{70}$$

for $\gamma = \iota_0^h a$ its solution is the discrete current-injection tangent, and the identification of its limit with the continuous current-injection flow (59) is carried out in the effective-curvature section.

Lemma C.13 (Raw and adapted quadratic master identities)**.** *Let* $\mathbf{g}_h^{(2)} := \mathrm{blkdiag}\big(g_{xx}(X_{N_h}^h), 0, 0\big)$, *recall the raw current gradient* $\widehat{p}_{m+1}^{h,\mathrm{raw}} = (\iota_0^h)^\top \widetilde{\Lambda}_{m+1}^h$, *and define the raw and adapted local quadratic sources*

$$\widetilde{\mathfrak{D}}_m^h := D^2_{\chi\chi} f(m) + \sum_{a=1}^n \widehat{p}_{m+1}^{h,\mathrm{raw},a}\, D^2_{\chi\chi} b^a(m) + \frac{1}{h}\sum_{a=1}^n \sum_{\ell=1}^d \widehat{p}_{m+1}^{h,\mathrm{raw},a}\, D^2_{\chi\chi}\sigma^{a,\ell}(m)\, \Delta B_m^{h,\ell},$$

$$\mathsf{D}_m^h := D^2_{\chi\chi} f(m) + \sum_{a=1}^n \mathbb{E}_m^h\big[p_{m+1}^{h,a}\big] D^2_{\chi\chi} b^a(m) + \sum_{a=1}^n \sum_{\ell=1}^d q_m^{h,\ell,a}\, D^2_{\chi\chi}\sigma^{a,\ell}(m), \quad \mathbb{E}_m^h\big[\widetilde{\mathfrak{D}}_m^h\big] = \mathsf{D}_m^h.$$

Then, for every $0 \le k < N_h$ *and* $\gamma, \gamma' \in \mathbb{R}^{d_h}$,

(a) (raw, pathwise)

$$\gamma^\top \widetilde{\mathbf{P}}_k^h\, \gamma' = (\mathbf{X}_{N_h}^{h,\gamma})^\top \mathbf{g}_h^{(2)}\, \mathbf{X}_{N_h}^{h,\gamma'} + h\sum_{m=k}^{N_h-1} (\mathbf{X}_m^{h,\gamma})^\top\, \widetilde{\mathfrak{D}}_m^h\, \mathbf{X}_m^{h,\gamma'};$$

(b) (adapted, conditional)

$$\gamma^\top \mathbf{P}_k^h\, \gamma' = \mathbb{E}_k^h\Big[(\mathbf{X}_{N_h}^{h,\gamma})^\top \mathbf{g}_h^{(2)}\, \mathbf{X}_{N_h}^{h,\gamma'} + h\sum_{m=k}^{N_h-1} (\mathbf{X}_m^{h,\gamma})^\top\, \mathsf{D}_m^h\, \mathbf{X}_m^{h,\gamma'}\Big];$$

in particular, for all $a, b \in \mathbb{R}^n$,

$$a^\top \Pi_k^h\, b = \mathbb{E}_k^h\Big[(\mathbf{X}_{N_h}^{h,\iota_0^h a})^\top \mathbf{g}_h^{(2)}\, \mathbf{X}_{N_h}^{h,\iota_0^h b} + h\sum_{m=k}^{N_h-1} (\mathbf{X}_m^{h,\iota_0^h a})^\top\, \mathsf{D}_m^h\, \mathbf{X}_m^{h,\iota_0^h b}\Big],$$

the discrete current-injection quadratic form matching (64)*.*

Proof. The terminal identity follows from $\widetilde{\mathbf{P}}_{N_h}^h = \iota_0^h g_{xx}(X_{N_h}^h)(\iota_0^h)^\top$. For one backward step, linearity of $\mathbf{T}_h$ gives $D^{2,\mathrm{ol}} r_k^h = h\mathbf{T}_h^\top D^2_{\chi\chi} f(k)\mathbf{T}_h$. Only the current transition coordinates have nonzero Hessians, namely

$$B_{k,a}^h = \mathbf{T}_h^\top \left[h D^2_{\chi\chi} b^a(k) + \sum_\ell D^2_{\chi\chi}\sigma^{a,\ell}(k)\Delta B_k^{h,\ell}\right]\mathbf{T}_h, \qquad 1 \le a \le n.$$

Their contraction with $\widetilde{\Lambda}_{k+1}^h$, together with the reward Hessian, is $h\mathbf{T}_h^\top \widetilde{\mathfrak{D}}_k^h \mathbf{T}_h$. Iterate the middle term of Lemma C.10 using $A_{k+1:m}^h A_k^h = A_{k:m}^h$ to obtain (a). For (b), condition (a), use $\mathcal{F}_{t_m^h}$-measurability of each tangent and the tower/covariance identities to replace $\widetilde{\mathfrak{D}}_m^h$ by D_m^h inside the running expectation. The terminal Hessian stays inside the expectation. The stated moment and bounded-jet hypotheses justify these products, and $\gamma = \iota_0^h a$, $\gamma' = \iota_0^h b$ gives the current-block formula. □

Lemma C.14 (Forward Euler state consistency)**.** *Under Assumptions C.4 and C.8, the detached reference Euler state satisfies*

$$\mathbb{E}\sup_{0\le t\le T}\big|X_t^{h,\mathrm{ref}} - X_t^{\mathrm{ref}}\big|^2 \longrightarrow 0.$$

Proof. Write $X = X^{\mathrm{ref}}$. For $t < T$, set

$$b_h(t) := b_{k_h(t)}^h\big(u_{k_h(t)}^{h,\mathrm{ref}}\big), \qquad \sigma_h(t) := \sigma_{k_h(t)}^h\big(u_{k_h(t)}^{h,\mathrm{ref}}\big),$$

and let $b_{\mathrm{ref}}(t) = b(\Xi^{\mathrm{ref}}(t))$ and $\sigma_{\mathrm{ref}}(t) = \sigma(\Xi^{\mathrm{ref}}(t))$. The coefficient-consistency clause of Assumption C.8 gives

$$\|b_h - b_{\mathrm{ref}}\|_{L^2}^2 + \|\sigma_h - \sigma_{\mathrm{ref}}\|_{L^2}^2 \longrightarrow 0,$$

where the L^2 norms are taken with respect to $\mathrm{d}t \otimes \mathrm{d}\mathbb{P}$. Define the auxiliary continuous interpolation

$$\widetilde{X}^h(t) = x_0(0) + \int_0^t b_h(s)\,\mathrm{d}s + \int_0^t \sigma_h(s)\,\mathrm{d}B(s).$$

It agrees with $X_k^{h,\mathrm{ref}}$ at every grid point, including T. The integral form of the reference SDDE, Cauchy–Schwarz, Doob's L^2 inequality and Itô's isometry imply

$$\mathbb{E}\sup_{0\le t\le T}|\widetilde{X}^h(t) - X(t)|^2 \le 2T\|b_h - b_{\mathrm{ref}}\|_{L^2}^2 + 8\|\sigma_h - \sigma_{\mathrm{ref}}\|_{L^2}^2 \longrightarrow 0.$$

To recover the piecewise-constant convention of Appendix C.3, let $\underline{t}_h(t) = h\lfloor t/h\rfloor$ for $t < T$ and $\underline{t}_h(T) = T$. Then $X^{h,\mathrm{ref}}(t) = \widetilde{X}^h(\underline{t}_h(t))$. The continuous reference has a square-integrable supremum, as also follows from the L^2 integrability of b_{ref} and σ_{ref}. Its modulus

$$m_X(h) := \sup_{\substack{s,t\in[0,T]\\ |s-t|\le h}} |X(s) - X(t)|$$

tends to zero almost surely and satisfies $m_X(h)^2 \le 4\sup_{t\le T}|X(t)|^2$. Dominated convergence therefore gives $\mathbb{E}\,m_X(h)^2 \to 0$. Consequently,

$$\mathbb{E}\sup_{0\le t\le T}|X^{h,\mathrm{ref}}(t) - X(t)|^2 \le 2\,\mathbb{E}\sup_{0\le t\le T}|\widetilde{X}^h(t) - X(t)|^2 + 2\,\mathbb{E}\,m_X(h)^2 \longrightarrow 0.$$

This proves the claim without a Hölder-norm moment for the reference state or a prescribed convergence rate. □

C.5 Direct first-order consistency

This subsection shows that the adapted first-order pair extracted on the mesh converges to the continuous anticipated adjoint pair as the history grows, which is part (b) of Theorem C.1. Throughout this subsection (p, q) denotes the continuous anticipated adjoint pair of (5), assembled from the family of Appendix C.2 through (50), and existence, square integrability, and a continuous version of p are inherited from the cited family. The driver is

$$\mathfrak{G}(t) := H_x(t) + \big(\mathcal{T}_\delta H_y\big)(t) + \big(\mathcal{K}H_m\big)(t), \quad H_\kappa(s) := f_\kappa(s) + b_\kappa(s)^\top p(s) + \sum_{\ell=1}^{d}\sigma_\kappa^\ell(s)^\top q^\ell(s), \tag{71}$$

with the anticipated shift and memory operators

$$(\mathcal{T}_\delta F)(t) := \mathbf{1}_{\{t+\delta\le T\}}\,\mathbb{E}_t\big[F(t+\delta)\big], \quad (\mathcal{K}F)(t) := \int_0^{\delta\wedge(T-t)} e^{-\rho r}\,\mathbb{E}_t\big[F(t+r)\big]\,\mathrm{d}r,$$

and their discrete counterparts

$$(\mathcal{T}_{\delta,h}F^h)_k := \mathbf{1}_{\{k+H_h\le N_h-1\}}\,\mathbb{E}_k^h\big[F_{k+H_h}^h\big], \quad (\mathcal{K}_hF^h)_k := \sum_{i=0}^{H_h} w_i^h\,\mathbf{1}_{\{k+i\le N_h-1\}}\,\mathbb{E}_k^h\big[F_{k+i}^h\big].$$

Since $\mathsf{H}_m^h(k)$ is $\mathcal{F}_{t_k^h}$-measurable, the w_0^h-term of (69) is the $i=0$ slot of $\mathcal{K}_h$, so Lemma C.12 reads, for every mesh h and $0 \le k < N_h$,

$$p_k^h = \mathbb{E}_k^h\big[p_{k+1}^h\big] + h\,\mathfrak{G}_k^h, \quad \mathfrak{G}_k^h := \mathsf{H}_x^h(k) + \big(\mathcal{T}_{\delta,h}\mathsf{H}_y^h\big)_k + \big(\mathcal{K}_h\mathsf{H}_m^h\big)_k, \quad p_{N_h}^h = g_x\big(X_{N_h}^h\big), \tag{72}$$

which carries the same three-operator shape as (71). The first-order comparison therefore reduces to replacing the continuous driver by its cell average, checking that the discrete operators reproduce their continuous counterparts, and closing the resulting error system by a stability estimate. The cell average is the predictable cell projection $(\mathsf{P}_h^{\mathrm{pred}}F)(t) := \bar{F}_{k_h(t)}$ with $\bar{F}_k := h^{-1}\mathbb{E}_k^h\int_{J_k^h}F(s)\,\mathrm{d}s$, which also supplies the comparator jets used below and later the cell-average inputs $(\bar{p}, \bar{q})$ used in Appendix C.6.

Lemma C.15 (Convergence of the predictable cell-average projection)**.** *For every predictable $F \in L^2(\mathrm{d}t\otimes\mathrm{d}\mathbb{P})$, $\|\mathsf{P}_h^{\mathrm{pred}}F - F\|_{L^2} \to 0$ as $h \downarrow 0$.*

Proof. For every $\mathcal{F}_{t_k^h}$-measurable v_k one has $\mathbb{E}\int_{J_k^h}\langle F(s)-\bar{F}_k, v_k\rangle\,\mathrm{d}s = 0$, so $\mathsf{P}_h^{\mathrm{pred}}$ is the orthogonal projection onto the closed subspace of processes that are constant on each cell and $\mathcal{F}_{t_k^h}$-measurable there. Elementary predictable step processes are dense in the predictable L^2 space and are asymptotically invariant under these projections as the mesh refines, and contractivity of orthogonal projections then yields strong convergence to the identity. No temporal path regularity of F is used. □

Lemma C.16 (Consistency of the anticipated operators and driver)**.** *Suppose Assumptions C.4, C.8, and C.9 hold, and recall from Assumption C.8 the weight bound $W := \sup_h \sum_{i=0}^{H_h}|w_i^h| < \infty$ and the partial-window modulus $\omega_{\mathcal{W}}$ of* (66)*. Then the discrete operators are uniformly L^2-bounded, $\|\mathcal{T}_{\delta,h}\| \le 1$ and $\|\mathcal{K}_h\| \le W$, and for every predictable $F \in L^2$,*

$$\big\|\mathcal{T}_{\delta,h}\mathsf{P}_h^{\mathrm{pred}}F - \mathsf{P}_h^{\mathrm{pred}}\mathcal{T}_\delta F\big\|_{L^2} \longrightarrow 0, \quad \big\|\mathcal{K}_h\mathsf{P}_h^{\mathrm{pred}}F - \mathsf{P}_h^{\mathrm{pred}}\mathcal{K}F\big\|_{L^2} \longrightarrow 0.$$

Consequently, with the comparator jets and driver

$$\begin{aligned} \mathsf{H}_{\kappa,k}^{h,c} &:= f_\kappa(k) + b_\kappa(k)^\top \pi_k^p + \sum_{\ell=1}^{d}\sigma_\kappa^\ell(k)^\top q_k^{\pi,\ell}, \quad \pi_k^p := \mathbb{E}_k^h\big[p(t_{k+1}^h)\big], \\ q_k^{\pi,\ell} &:= \frac{1}{h}\,\mathbb{E}_k^h\big[p(t_{k+1}^h)\,\Delta B_k^{h,\ell}\big], \quad \mathfrak{G}_k^{h,c} := \mathsf{H}_{x,k}^{h,c} + \big(\mathcal{T}_{\delta,h}\mathsf{H}_y^{h,c}\big)_k + \big(\mathcal{K}_h\mathsf{H}_m^{h,c}\big)_k, \end{aligned} \tag{73}$$

the consistency defect $\mathfrak{d}_k := \mathfrak{G}_k^{h,c} - \bar{\mathfrak{G}}_k$ satisfies $h\sum_k \mathbb{E}|\mathfrak{d}_k|^2 \to 0$.

Proof. Uniform boundedness follows from conditional Jensen and the weight bound. For the shift, the grid alignment $\delta = H_h h$ makes the discrete shift exact on grid points, and at the boundary cell $k = N_h - H_h$ the continuous indicator $\mathbf{1}_{\{s+\delta\le T\}}$ is nonzero only at the left endpoint of J_k^h, a $\mathrm{d}s$-null set, so both cell contributions vanish and the point-delay cutoff is aligned with the mesh. The remaining commutator with $\mathsf{P}_h^{\mathrm{pred}}$ vanishes by Lemma C.15. For the memory operator, full-window consistency alone does not suffice: near the horizon the discrete operator uses every truncated partial sum $\sum_{i=0}^{N_h-1-k} w_i^h$, and the last H_h cells span the fixed length δ, so the truncation is not a one-cell boundary effect. Both operators are therefore verified directly on elementary predictable step processes: on each constancy segment the prefix masses are controlled by (66) applied to the constant test function, the finitely many jump-straddling cells are bounded separately by $\max_j|w_j^h| \le C_w h$, only the single extra endpoint cell of the atom-priority convention is $O(h)$, and the conditional factors converge by Lemma C.15, since elementary predictable step processes are dense in the predictable L^2 space, and the uniform operator bounds extend the convergence to all predictable F by a three-epsilon argument. Three consistency defects remain. The first is the reference jet consistency of Assumption C.8, paired with the square-integrable inputs by an L^2 truncation. The second is the p-cell defect, which vanishes by the continuous version of p. The third is the q drift-covariance defect, for which the martingale representation of the cell increment of p gives the exact identity

$$\begin{aligned} q_k^{\pi,\ell} - \bar{q}_k^\ell &= -h^{-1}\mathbb{E}_k^h\left[\left(\int_{J_k^h}\mathfrak{G}(s)\,\mathrm{d}s\right)\Delta B_k^{h,\ell}\right], \\ h\sum_k \mathbb{E}\|q_k^\pi - \bar{q}_k\|_F^2 &\le Ch\,\mathbb{E}\int_0^T |\mathfrak{G}(s)|^2\,\mathrm{d}s. \end{aligned} \tag{74}$$

The covariance bound uses conditional orthogonal projection onto ΔB_k^h followed by Cauchy–Schwarz in the time integral, so it does not require higher moments of $\mathfrak{G}$. The operator commutators complete the list of defects. With the comparator built from (π^p, q^π), the q-part of the driver difference $\mathfrak{G}^h - \mathfrak{G}^{h,c}$ is exactly $z = q^h - q^\pi$, aligning the stability estimate below with the one-step martingale variance bound. □

Proposition C.17 (Stability of the discrete anticipated error system)**.** *Let $e_k := p_k^h - p(t_k^h)$ and, with q^π as in* (73)*, $z_k^\ell := q_k^{h,\ell} - q_k^{\pi,\ell}$. Then $z_k^\ell = h^{-1}\mathbb{E}_k^h[e_{k+1}\Delta B_k^{h,\ell}]$, with the one-step martingale variance bound $h\sum_\ell \mathbb{E}_k^h|z_k^\ell|^2 \le \mathbb{E}_k^h|\nu_{k+1}|^2$, $\nu_{k+1} := e_{k+1} - \mathbb{E}_k^h e_{k+1}$, and, for h small,*

$$\max_{0\le k\le N_h}\mathbb{E}|e_k|^2 + h\sum_{k=0}^{N_h-1}\sum_{\ell=1}^{d}\mathbb{E}\big|z_k^\ell\big|^2 \le C\Big(\mathbb{E}\big|e_{N_h}\big|^2 + h\sum_{k=0}^{N_h-1}\mathbb{E}\big|\mathfrak{d}_k\big|^2\Big).$$

Proof. The error equation $e_k = \mathbb{E}_k^h e_{k+1} + h(\mathfrak{G}_k^h - \mathfrak{G}_k^{h,c}) + h\,\mathfrak{d}_k$ and the bounded jets give – the q-part of each jet difference in $\mathfrak{G}_k^h - \mathfrak{G}_k^{h,c}$ being exactly z by the choice of comparator in (73) – with $\epsilon_j^2 := |\mathbb{E}_j^h e_{j+1}|^2 + \|z_j\|_F^2$, the convention $\epsilon_j := 0$ for $j \geq N_h$, and the weighted Cauchy–Schwarz inequality $|\sum_i w_i a_i|^2 \leq (\sum_i |w_i|) \sum_i |w_i||a_i|^2$,

$$\mathbb{E}\big|\mathfrak{G}_k^h - \mathfrak{G}_k^{h,c}\big|^2 \leq C\Big(\mathbb{E}\epsilon_k^2 + \mathbf{1}_{\{k+H_h \leq N_h-1\}}\mathbb{E}\epsilon_{k+H_h}^2 + \sum_{i=0}^{H_h} |w_i^h|\,\mathbf{1}_{\{k+i\leq N_h-1\}}\mathbb{E}\epsilon_{k+i}^2\Big).$$

Squaring the error equation, using Young's inequality with parameter β, and the orthogonal decomposition $\mathbb{E}|e_{k+1}|^2 = \mathbb{E}|\mathbb{E}_k^h e_{k+1}|^2 + \mathbb{E}|\nu_{k+1}|^2$ yields

$$\mathbb{E}|e_k|^2 + c_0\,\mathbb{E}|\nu_{k+1}|^2 \leq (1+C_\beta h)\,\mathbb{E}|e_{k+1}|^2 + c_\beta h\Big(\mathbb{E}\|z_k\|_F^2 + \mathbb{E}\epsilon_{k+H_h}^2 + \sum_i |w_i^h|\,\mathbb{E}\epsilon_{k+i}^2 + \mathbb{E}|\mathfrak{d}_k|^2\Big).$$

Summing backward, the delayed terms reindex with total weight at most $1 + W$, the reindexed z-terms are absorbed using $h\mathbb{E}\|z_j\|_F^2 \leq \mathbb{E}|\nu_{j+1}|^2$ after choosing β large and then h small, and the remaining level terms close by the standard backward discrete Gronwall inequality. No smallness of δ or of the horizon is used. □

Proposition C.18 (Direct first-order consistency)**.** *Under Assumptions C.4, C.8, and C.9,*

$$\mathbb{E}\int_0^T \Big(\big|p_t^{h,\mathrm{ref}} - p_t^{\mathrm{ref}}\big|^2 + \big\|q_t^{h,\mathrm{ref}} - q_t^{\mathrm{ref}}\big\|_F^2\Big)\mathrm{d}t \longrightarrow 0, \quad \mathbb{E}\big|p_T^{h,\mathrm{ref}} - p_T^{\mathrm{ref}}\big|^2 \longrightarrow 0.$$

Proof. Let (e, z) be as in Proposition C.17. The Lipschitz terminal gradient and Lemma C.14 give $\mathbb{E}|e_{N_h}|^2 \to 0$, and Lemma C.16 gives $h\sum_k \mathbb{E}|\mathfrak{d}_k|^2 \to 0$. Stability therefore yields

$$\max_k \mathbb{E}|e_k|^2 + h\sum_k \mathbb{E}\|z_k\|_F^2 \longrightarrow 0.$$

For p, split its piecewise-constant error into $e_{k_h(t)}$ and $p(t_{k_h(t)}^h) - p(t)$. The latter tends to zero in L^2 by continuity of the $\mathcal{S}^2$ version and dominated convergence. For q, with $q_t^{\pi,h} = q_{k_h(t)}^\pi$ and $\bar{q}^h = \mathsf{P}_h^{\mathrm{pred}} q$, use

$$\|q^h - q\|_{L^2}^2 \leq 3h\sum_k \mathbb{E}\|z_k\|_F^2 + 3h\sum_k \mathbb{E}\|q_k^\pi - \bar{q}_k\|_F^2 + 3\|\mathsf{P}_h^{\mathrm{pred}} q - q\|_{L^2}^2.$$

The middle term is at most $Ch\|\mathfrak{G}\|_{L^2}^2$ by (74), and the last vanishes by Lemma C.15. Together with the terminal estimate, this proves the claim without pointwise temporal regularity of q. □

C.6 Effective-curvature consistency

This subsection shows that the current block of the fixed-control OL-BPTT Hessian converges to the effective curvature of Section 4, which is part (c) of Theorem C.1, and it is where the point-delay maturity cell and the predictable projection of q are handled. Here Π is the identified curvature (58), and $\Pi_t^h = \Pi_{k_h(t)}^h$ is the extraction (68). We use its predictable version: predictable projection of truncations of the jointly measurable adapted Π agrees with Π almost everywhere, because $\mathcal{F}_{t-} = \mathcal{F}_t$ in the augmented Brownian filtration and the condition (79) supplies integrability. All conclusions are in product-measure norms, independently of this choice of version. For the tangents of (70), set

$$\mathbf{X}_{k\to r}^{h,\gamma} := \mathbf{T}_h A_{k:r}^h \gamma, \qquad \mathbf{X}_{k\to r}^{h,a} := \mathbf{X}_{k\to r}^{h,\iota_0^h a}, \qquad \mathbf{X}_{k\to r}^{h,i} := \mathbf{X}_{k\to r}^{h,e_i}.$$

Write their three channels as $(\mathsf{x}_r, \mathsf{y}_r, \mathsf{m}_r)$. The continuous channels (v^x, v^y, v^m), the first jet rows L_b, L_σ^j and $\mathcal{W}_t$ are those of Lemma C.5.

Partial-window quadrature. The prefix condition (66) is needed both by the anticipated operator and by the injection tangent. For cell-integrated exponential weights, Hölder comparison at the cell atoms and the half-cell endpoint overhang give $\omega_{\mathcal{W}}(h) \leq C(h^\alpha + h)$.

Lemma C.19 (Truncated exponential-window quadrature along the current tangent)**.** *Let Assumptions C.4, C.8, and C.9 hold, and let v^x be the current channel of $\mathcal{V}^{t,a}$, with the moments and Hölder regularity of Lemma C.5, extended by zero before t. For every anchor $t = t_k^h$, $k \le r \le N_h$, and $\ell = (r-k) \wedge H_h$,*

$$\mathbb{E}\left|\sum_{j=0}^{\ell} w_j^h v^x(t_r^h - jh) - (\mathcal{W}_{t_k^h} v^x)(t_r^h)\right|^4 \le C\omega_{\mathcal{W}}(h)^4 |a|^4.$$

Proof. Put $\varphi(\theta) = v^x(t_r^h + \theta)$ on $[-\ell h, 0]$. Since $t_r^h - \ell h = (t_r^h - \delta) \vee t_k^h$, the substitution $\theta = u - t_r^h$ identifies the integral with the truncated window. Apply (66) pathwise, raise to the fourth power, and use the Hölder-norm moment of Lemma C.5. □

Lemma C.20 (Current-injection tangent consistency)**.** *Let Assumptions C.4, C.8, and C.9 hold.*

(a) (channel resummation) *The continuous flow satisfies* (60)–(61)*. On the discrete side, for $a \in \mathbb{R}^n$ the channels of* (70) *obey, for $k \le r \le N_h$ and with the current recursion read for $r < N_h$, the exact identities*

$$\mathsf{y}_r = \mathbf{1}_{\{r-k \ge H_h\}}\, \mathsf{x}_{r-H_h}, \quad \mathsf{m}_r = \sum_{j=0}^{(r-k)\wedge H_h} w_j^h\, \mathsf{x}_{r-j},$$

$$\mathsf{x}_{r+1} - \mathsf{x}_r = \Big(h\, J_{b,r}^h + \sum_{\ell=1}^{d} J_{\sigma,r}^{h,\ell} \Delta B_r^{h,\ell}\Big) \mathbf{X}_{k\to r}^{h,a}, \quad \mathsf{x}_k = a,$$

so the discrete tangent is exactly the Euler and partial-quadrature discretization of (61)*, with the closed maturity convention $\mathbf{1}_{\{r-k\ge H_h\}}$ in the free term against the strict continuous convention $\mathbf{1}_{\{s-t>\delta\}}$: the kernel column indicators agree literally, $\mathbf{1}_{\{r-r'\ge H_h+1\}} = \mathbf{1}_{\{t_r^h - t_{r'}^h > \delta\}}$ for $k \le r' < r$, and the sole mismatch is the single maturity cell $r = k + H_h$ of the free term.*

(b) (moments and increments) *For every $p \ge 2$,*

$$\sup_h \max_{0\le k\le r\le N_h} \max_{1\le i\le n} \mathbb{E}\big|\mathbf{X}_{k\to r}^{h,i}\big|^{2p} < \infty,$$

and for $t \le s \le s' \le T$,

$$\mathbb{E}\big|v^x(s') - v^x(s)\big|^4 + \mathbb{E}\big|v^m(s') - v^m(s)\big|^4 \le C|s'-s|^2|a|^4,$$

while the same increment bound holds for v^y whenever $[s, s']$ does not contain the maturity time $t + \delta$, across which v^y has the single jump $v^x(t)$.

(c) (integrated tangent consistency)

$$\begin{aligned}\omega_{\tan}(h) := \max_{1\le i\le n}\Big\{ & h \sum_{k=0}^{N_h-1} \mathbb{E}|\mathbf{X}_{k\to N_h}^{h,i} - \mathcal{V}^{t_k^h, e_i}(T)|^4 \\ & + h^2 \sum_{k=0}^{N_h-1}\sum_{r=k}^{N_h-1} \mathbb{E}|\mathbf{X}_{k\to r}^{h,i} - \mathcal{V}^{t_k^h, e_i}(t_r^h)|^4\Big\} \to 0.\end{aligned} \tag{75}$$

(d) (anchor stability)

$$\begin{aligned}\omega_{\mathrm{anc}}(h) := \max_{1\le i\le n} \sum_{k=0}^{N_h-1} \int_{J_k^h} \Big[& \mathbb{E}|\mathcal{V}^{t_k^h, e_i}(T) - \mathcal{V}^{t, e_i}(T)|^4 \\ & + \int_t^T \mathbb{E}|\mathcal{V}^{t_k^h, e_i}(s) - \mathcal{V}^{t, e_i}(s)|^4\, \mathrm{d}s\Big] \mathrm{d}t \to 0.\end{aligned} \tag{76}$$

No pointwise or $\max_{k,r}$ *convergence of the tangent difference is asserted: under the strict boundary representative the free-term point-delay channel differs by the* $O(1)$ *value* $|a|$ *on the single maturity cell of each anchor and, for the single anchor* $t_k^h = T - \delta$*, at the terminal time, and both contributions vanish under the* h*- and* h^2*-weights of* (75)*.*

Proof. Channel identities and moments. The continuous identities were proved without mesh assumptions in Lemma C.5. In (70), $\mathbf{T}_h \mathbf{S}_h^j \iota_0^h$ has current row I_n, point row $\mathbf{1}_{\{j \geq H_h\}} I_n$, and memory row $(\sum_{i=0}^{j \wedge H_h} w_i^h) I_n$. Regrouping its free and kernel terms by lag gives (a). In the kernel column $j = r-1-r'$, so the point indicator is $\mathbf{1}_{\{r-r' \geq H_h+1\}}$, exactly the strict continuous indicator. Only the free term differs at $r = k + H_h$. The discrete current recursion, $\sum_j |w_j^h| \leq W$, discrete Burkholder–Davis–Gundy and Gronwall give the uniform $2p$-moments. Continuous current increments follow from Lemma C.5, the window formula gives the memory increments, and the delayed channel has the stated single maturity jump.

Current-channel comparison. For an anchor k, set $d_r = \mathsf{x}_r - v^x(t_r^h)$. Subtract the cell-integrated (61) from (a). The homogeneous error depends on $(d_r, \mathbf{1}_{\{r-k \geq H_h+1\}} d_{r-H_h}, \sum_{j \leq (r-k) \wedge H_h} w_j^h d_{r-j})$ through bounded jets. There are five forcing terms:

Defect	Control used in the fourth-moment estimate
Reference jets/state	Assumption C.8 and Lemma C.14; bounded jet differences paired with the uniform eighth tangent moments.
Truncated memory	Lemma C.19.
Maturity cell	One $O(\|a\|)$ discrepancy, entering the current increment as $hb_y a + \sum_j \sigma_y^j a \Delta B_r^{h,j}$; its fourth-moment contribution is $O(h^2)\|a\|^4$.
Within-cell evaluation	The increment bounds in (b), integrated over the cells, and coefficient consistency.
Endpoint quadrature	The $O(h)$ endpoint terms already included in the moduli of Assumption C.8.

For a bounded jet difference Δ_h, the product estimate is

$$\mathbb{E}|\Delta_h \mathbf{X}|^4 \leq (\mathbb{E}\|\Delta_h\|^8)^{1/2} (\mathbb{E}|\mathbf{X}|^8)^{1/2}, \qquad \mathbb{E}\|\Delta_h\|^8 \leq C\mathbb{E}\|\Delta_h\|^2.$$

Apply it in the time/anchor aggregates, where jet consistency holds. Burkholder–Davis–Gundy, weighted Cauchy–Schwarz and delay-Gronwall give

$$\max_{k \leq r \leq N_h} \mathbb{E}|d_r|^4 \leq C D_k(h), \qquad h \sum_{k=0}^{N_h - 1} D_k(h) \longrightarrow 0,$$

where $D_k(h)$ is the sum of the five forcing estimates over $[t_k^h, T]$. The first, second, fourth and fifth entries are controlled by mesh-global integrated defects, and the third is one cell. No uniform-in-anchor convergence of D_k is needed.

Integrated full-channel comparison. Away from the maturity cell the delayed error is d_{r-H_h}, and the memory error is a weighted average of d plus the quadrature defect. Thus the running current errors sum to at most $Th \sum_k \max_r \mathbb{E}|d_r|^4$, and the terminal errors to $h \sum_k \max_r \mathbb{E}|d_r|^4$. The $O(1)$ point-channel discrepancies have weights $h^2 \sum_k 1 = Th$ in the running sum and h at the single terminal anchor $t_k^h = T - \delta$. This proves (c).

Anchor stability. For $t \in J_k^h$, compare the two closed current equations anchored at t_k^h and t. Their discrepancies are the initial interval of length at most h (fourth moment $O(h^2)|a|^4$), the maturity-indicator strip $(t_k^h + \delta, t + \delta]$, and a memory lower-limit integral over an interval of length at most h. The same fourth-moment delay-Gronwall estimate and common jet bounds yield

$$\mathbb{E} \sup_{t \leq s \leq T} |v^x_{(t_k^h)}(s) - v^x_{(t)}(s)|^4 \leq \varpi_{\mathrm{anc}}(h)|a|^4, \qquad \varpi_{\mathrm{anc}}(h) \longrightarrow 0,$$

uniformly in these anchors. Here $v^x_{(\ell)}$ denotes the current flow started at ℓ, and no mesh-jet comparison is used in this restart estimate. The memory channel follows by integration, while the point channel

also has an $O(|a|)$ difference on the maturity strip. Its total (t,s)-weight is $O(h)$, including the terminal-anchor strip. Integrating gives (d). □

For the curvature comparison, use the quadratic representation of Lemma C.13(b): for all $a, b \in \mathbb{R}^n$,

$$a^\top \Pi_k^h\, b = \mathbb{E}_k^h\big[\widehat{\mathcal{Q}}_{k,h}(a,b)\big], \quad \widehat{\mathcal{Q}}_{k,h}(a,b) := (\mathbf{X}_{k\to N_h}^{h,a})^\top \mathbf{g}_h^{(2)}\, \mathbf{X}_{k\to N_h}^{h,b} + h \sum_{r=k}^{N_h-1} (\mathbf{X}_{k\to r}^{h,a})^\top \mathsf{D}_r^h\, \mathbf{X}_{k\to r}^{h,b},$$

its continuous counterpart being $a^\top \Pi_t\, b = \mathbb{E}_t[\widehat{\mathcal{Q}}_t(a,b)]$ by Lemma C.7 with $\widehat{\mathcal{Q}}_t$ of (64), whose Hamiltonian Hessian along the reference tuple is

$$D^2_{\chi\chi} H(s) = D^2_{\chi\chi} f(s) + \sum_{a=1}^n p^a(s)\, D^2_{\chi\chi} b^a(s) + \sum_{a=1}^n \sum_{\ell=1}^d q^{a,\ell}(s)\, D^2_{\chi\chi} \sigma^{a,\ell}(s) \;\in\; L^2(\mathrm{d}t \otimes \mathrm{d}\mathbb{P}),$$

square integrability holding because the Hessian jets are bounded and (p,q) is square integrable. The comparison targets are the predictable cell projections

$$\Pi_k^\pi := \frac{1}{h}\, \mathbb{E}_k^h \int_{J_k^h} \Pi_t\, \mathrm{d}t \;=\; \big(\mathsf{P}_h^{\mathrm{pred}}\Pi\big)\big|_{J_k^h}, \quad \overline{\mathsf{D}}_r := \frac{1}{h}\, \mathbb{E}_r^h \int_{J_r^h} D^2_{\chi\chi} H(s)\, \mathrm{d}s, \tag{77}$$

and the cell-source consistency established below is

$$h \sum_{r=0}^{N_h-1} \mathbb{E}\big\|\mathsf{D}_r^h - \overline{\mathsf{D}}_r\big\|_F^2 \longrightarrow 0, \tag{78}$$

not a pointwise comparison with $D^2_{\chi\chi} H(t_r^h)$, which is in general only an L^2-in-time object through q.

Proposition C.21 (Effective-curvature consistency)**.** *Let Assumptions C.4, C.8, and C.9 hold, including the partial-window clause* (66) *of Assumption C.8, and assume the current-curvature uniform integrability: for some $\eta > 0$,*

$$\sup_h\; \mathbb{E}\int_0^T \big\|\Pi_t^{h,\mathrm{ref}}\big\|_F^{2+\eta}\, \mathrm{d}t \;+\; \mathbb{E}\int_0^T \big\|\Pi_t^{\mathrm{ref}}\big\|_F^{2+\eta}\, \mathrm{d}t < \infty. \tag{79}$$

Then

$$\mathbb{E}\int_0^T \big\|\Pi_t^{h,\mathrm{ref}} - \Pi_t^{\mathrm{ref}}\big\|_F^2\, \mathrm{d}t \longrightarrow 0,$$

which is part (c) of Theorem C.1 and the curvature conclusion of main-text Theorem 5.1(b).

Proof. We work with a fixed basis pair (e_i, e_j) and suppress reference superscripts. *Projection and integrability.* Split $\Pi^h - \Pi = (\Pi^h - \mathsf{P}_h^{\mathrm{pred}}\Pi) + (\mathsf{P}_h^{\mathrm{pred}}\Pi - \Pi)$. The second term vanishes in L^2 by Lemma C.15. By (79) and conditional Jensen, the first is uniformly bounded in $L^{2+\eta}$. Therefore L^1 convergence suffices, since

$$\|F\|_{L^2} \le \|F\|_{L^1}^{\theta} \|F\|_{L^{2+\eta}}^{1-\theta}, \qquad \theta = \frac{\eta}{2(1+\eta)}.$$

Cell-source consistency. Put $M(s) = D^2_{\chi\chi} H(s)$ and $\bar{p}_r = h^{-1}\mathbb{E}_r^h \int_{J_r^h} p(s)\, \mathrm{d}s$, and let $\bar{q}_r$ be the analogous projection of q. With $e_{r+1} = p_{r+1}^h - p(t_{r+1}^h)$,

$$\mathbb{E}_r^h p_{r+1}^h - \bar{p}_r = \mathbb{E}_r^h e_{r+1} - \frac{1}{h}\mathbb{E}_r^h \int_{J_r^h} \int_s^{t_{r+1}^h} \mathfrak{G}(u)\, \mathrm{d}u\, \mathrm{d}s.$$

Its squared aggregate tends to zero by Proposition C.18, the last term contributing at most $h^2 \|\mathfrak{G}\|_{L^2}^2$. Also $h\sum_r \mathbb{E}\|q_r^h - \bar{q}_r\|_F^2 \to 0$ by (74) and that proposition. For any scalar multiplier component $m \in \{p^a, q^{a,\ell}\}$ and its Hessian jet $B \in \{D^2_{\chi\chi} b^a, D^2_{\chi\chi}\sigma^{a,\ell}\}$, let $m_r^h = \mathbb{E}_r^h p_{r+1}^{h,a}$ or $q_r^{h,\ell,a}$, respectively, and $\bar{m}_r = h^{-1}\mathbb{E}_r^h \int_{J_r^h} m$. The exact product split is

$$m_r^h B_r^h - \frac{1}{h}\mathbb{E}_r^h \int_{J_r^h} m(s) B(s)\, \mathrm{d}s = (m_r^h - \bar{m}_r) B_r^h + \frac{1}{h}\mathbb{E}_r^h \int_{J_r^h} \big[(\bar{m}_r - m)B + \bar{m}_r (B_r^h - B)\big]\, \mathrm{d}s.$$

The first two terms vanish in aggregate by the multiplier convergence, predictable projection and bounded jets. For the last, bounded jet defects Δ_h converge in integrated L^2, and they are paired with a square-integrable multiplier by truncation:

$$\mathbb{E}\int_0^T |m|^2\|\Delta_h\|^2\,\mathrm{d}s \le R^2\|\Delta_h\|_{L^2}^2 + C\mathbb{E}\int_0^T \mathbf{1}_{\{|m|>R\}}|m|^2\,\mathrm{d}s.$$

First let $h\downarrow 0$, then $R\uparrow\infty$. The same argument applies to $\bar{m}$, whose squared family is uniformly integrable by strong L^2 convergence of the projections. The pure $D^2 f$ term uses jet consistency alone. This proves (78). Crucially, adaptedness of the tangents gives the exact running identity

$$h\mathbb{E}_k^h[(\mathbf{X}_{k\to r}^{h,i})^\top \overline{\mathsf{D}}_r \mathbf{X}_{k\to r}^{h,j}] = \mathbb{E}_k^h\int_{J_r^h} (\mathbf{X}_{k\to r}^{h,i})^\top M(s)\mathbf{X}_{k\to r}^{h,j}\,\mathrm{d}s.$$

No pointwise temporal evaluation of q or M is required.

Five functional defects. For $t\in J_k^h$, define $U_i(s)=\mathbf{X}_{k\to k_h(s)}^{h,i}$ for $s<T$ and $U_i(T)=\mathbf{X}_{k\to N_h}^{h,i}$, and set $V_i(s)=\mathcal{V}^{t,e_i}(s)$. The tower property and stochastic Fubini give

$$e_i^\top(\Pi_k^h-\Pi_k^\pi)e_j = \frac{1}{h}\int_{J_k^h}\Delta_{k,t}^{ij}\,\mathrm{d}t, \qquad \Delta_{k,t}^{ij} := \mathbb{E}_k^h[\widehat{\mathcal{Q}}_{k,h}(e_i,e_j)-\widehat{\mathcal{Q}}_t(e_i,e_j)].$$

Here the moment bounds and $M\in L^2$ justify Fubini. Using the preceding running identity, $\Delta = R_{\mathrm{src}}+R_{\mathrm{init}}+R_{\tan}+T_{\mathrm{coef}}+T_{\tan}$, where

$$\begin{aligned}
R_{\mathrm{src}} &= \mathbb{E}_k^h\left[h\sum_{r=k}^{N_h-1} U_i(t_r^h)^\top(\mathsf{D}_r^h-\overline{\mathsf{D}}_r)U_j(t_r^h)\right],\\
R_{\mathrm{init}} &= \mathbb{E}_k^h\int_{t_k^h}^t U_i(s)^\top M(s)U_j(s)\,\mathrm{d}s,\\
R_{\tan} &= \mathbb{E}_k^h\int_t^T [U_i(s)^\top M(s)U_j(s)-V_i(s)^\top M(s)V_j(s)]\,\mathrm{d}s,\\
T_{\mathrm{coef}} &= \mathbb{E}_k^h[U_i(T)^\top(\mathbf{g}_h^{(2)}-\mathbf{g}^{(2)})U_j(T)],\\
T_{\tan} &= \mathbb{E}_k^h[U_i(T)^\top\mathbf{g}^{(2)}U_j(T)-V_i(T)^\top\mathbf{g}^{(2)}V_j(T)].
\end{aligned}$$

We estimate $\sum_k\int_{J_k^h}\mathbb{E}|\cdot|\,\mathrm{d}t$ for each term. The source term is bounded by $C(h\sum_r\mathbb{E}\|\mathsf{D}_r^h-\overline{\mathsf{D}}_r\|_F^2)^{1/2}\to 0$ using uniform fourth tangent moments. On the initial strip $U_i=\mathbf{T}_h\iota_0^h e_i=(e_i,0,w_0^h e_i)$, so its bound is $Ch\sqrt{T}\|M\|_{L^2}\to 0$. For the running tangent use

$$x^\top My - v^\top Mw = (x-v)^\top My + v^\top M(y-w)$$

and the three-term split

$$U_i(s)-V_i(s) = [\mathbf{X}_{k\to r}^{h,i}-\mathcal{V}^{t_k^h,e_i}(t_r^h)] + [\mathcal{V}^{t_k^h,e_i}(t_r^h)-\mathcal{V}^{t_k^h,e_i}(s)] + [\mathcal{V}^{t_k^h,e_i}(s)-\mathcal{V}^{t,e_i}(s)]$$

for $s\in J_r^h\cap[t,T]$. Hölder with exponents $(4,2,4)$ bounds the first and third contributions by $C\omega_{\tan}(h)^{1/4}$ and $C\omega_{\mathrm{anc}}(h)^{1/4}$. The middle contribution vanishes by Lemma C.20(b), and its single point-delay jump cell per anchor has aggregate weight $O(h)$. The terminal coefficient term vanishes by Lipschitz continuity of g_{xx}, Lemma C.14 and $(2,4,4)$ Hölder. The terminal tangent term uses the same bilinear split and the terminal clauses of (75)–(76), since there is no within-cell time error at T, and the single exceptional terminal anchor carries weight h. Thus the aggregate of $|\Delta_{k,t}^{ij}|$ tends to zero, proving L^1 convergence of $e_i^\top(\Pi^h-\Pi^\pi)e_j$. The first step upgrades it to L^2. Sum over the finitely many basis pairs and add the predictable-projection error to conclude. □

C.7 Effective recovery inputs

This subsection combines the previous two limits with the shifted input $\zeta = q-\Pi\sigma^{\mathrm{ref}}$ and shows that the mesh-level generalized Hamiltonian converges to its continuous counterpart, which is part (d) of Theorem C.1. Throughout this subsection the discrete recovery inputs $(p^{h,\mathrm{ref}}, q^{h,\mathrm{ref}}, \Pi^{h,\mathrm{ref}}, \zeta^{h,\mathrm{ref}})$

are the direct extractions (68), extended piecewise constantly in time, and $(p^{\mathrm{ref}}, q^{\mathrm{ref}}, \Pi^{\mathrm{ref}}, \zeta^{\mathrm{ref}})$ is the continuous recovery tuple of Section 4 along u^{ref}. For $u \in \mathcal{U}$ write $f_k^h(u) := f(t_k^h, X_k^h, Y_k^h, M_k^h, u)$ and define the mesh-level Hamiltonians

$$H_k^h(u; p, q) := f_k^h(u) + p^\top b_k^h(u) + \langle q, \sigma_k^h(u) \rangle_F,$$

$$\mathcal{H}_k^h(u; p, \zeta, \Pi) := f_k^h(u) + p^\top b_k^h(u) + \langle \zeta, \sigma_k^h(u) \rangle_F + \tfrac{1}{2} \operatorname{tr}\big[\sigma_k^h(u)^\top \, \Pi \, \sigma_k^h(u)\big],$$

the discrete counterparts of (4) and (7). The anchored generalized-Hamiltonian differences used in Theorem C.1(d) are

$$\Delta\mathcal{H}_k^h(u) := \mathcal{H}_k^h\big(u; p_k^{h,\mathrm{ref}}, \zeta_k^{h,\mathrm{ref}}, \Pi_k^{h,\mathrm{ref}}\big) - \mathcal{H}_k^h\big(u_k^{h,\mathrm{ref}}; p_k^{h,\mathrm{ref}}, \zeta_k^{h,\mathrm{ref}}, \Pi_k^{h,\mathrm{ref}}\big).$$

Set $\Delta\mathcal{H}_t^h(u) := \Delta\mathcal{H}_{k_h(t)}^h(u)$, and let

$$\Delta\mathcal{H}_t(u) := \mathcal{H}\big(t; u, p_t^{\mathrm{ref}}, \zeta_t^{\mathrm{ref}}, \Pi_t^{\mathrm{ref}}\big) - \mathcal{H}\big(t; u_t^{\mathrm{ref}}, p_t^{\mathrm{ref}}, \zeta_t^{\mathrm{ref}}, \Pi_t^{\mathrm{ref}}\big),$$

both anchored at the reference control, so that every u-independent term cancels and only reference-anchored coefficient differences remain.

Lemma C.22 (Current-noise injection and the anchored Hamiltonian identity)**.** *Recall from* (68) *that*

$$p_k^{h,\mathrm{ref}} = (\iota_0^h)^\top \Lambda_k^h, \quad q_k^{h,\mathrm{ref}} = (\iota_0^h)^\top \mathbf{Z}_k^h, \quad \Pi_k^{h,\mathrm{ref}} = (\iota_0^h)^\top \mathbf{P}_k^h \, \iota_0^h.$$

The augmented Euler diffusion is $\mathsf{G}_k^h(u) = \iota_0^h \, \sigma_k^h(u)$*, so for* $\Delta\sigma_k^h(u) := \sigma_k^h(u) - \sigma_k^{h,\mathrm{ref}}$*,*

$$\tfrac{1}{2} \operatorname{tr}\big[\Delta\mathsf{G}_k^h(u)^\top \, \mathbf{P}_k^h \, \Delta\mathsf{G}_k^h(u)\big] = \tfrac{1}{2} \operatorname{tr}\big[\Delta\sigma_k^h(u)^\top \, \Pi_k^{h,\mathrm{ref}} \, \Delta\sigma_k^h(u)\big],$$

so only the current block $\Pi_k^{h,\mathrm{ref}}$ *of* $\mathbf{P}_k^h$ *enters the generalized-Hamiltonian correction. With the extraction* $\zeta_k^{h,\mathrm{ref}} = q_k^{h,\mathrm{ref}} - \Pi_k^{h,\mathrm{ref}} \sigma_k^{h,\mathrm{ref}}$ *of* (68)*, for every* $u \in \mathcal{U}$*,*

$$\begin{aligned}
&\mathcal{H}_k^h\big(u; p_k^{h,\mathrm{ref}}, \zeta_k^{h,\mathrm{ref}}, \Pi_k^{h,\mathrm{ref}}\big) - \mathcal{H}_k^h\big(u_k^{h,\mathrm{ref}}; p_k^{h,\mathrm{ref}}, \zeta_k^{h,\mathrm{ref}}, \Pi_k^{h,\mathrm{ref}}\big) \\
&\quad = H_k^h\big(u; p_k^{h,\mathrm{ref}}, q_k^{h,\mathrm{ref}}\big) - H_k^h\big(u_k^{h,\mathrm{ref}}; p_k^{h,\mathrm{ref}}, q_k^{h,\mathrm{ref}}\big) + \tfrac{1}{2} \operatorname{tr}\big[\Delta\sigma_k^h(u)^\top \, \Pi_k^{h,\mathrm{ref}} \, \Delta\sigma_k^h(u)\big].
\end{aligned}$$

Proof. Use $\Delta G_k^h = \iota_0^h \Delta\sigma_k^h$ and cyclicity of the trace for the first identity. For the second, write $\sigma_k^h(u) = \sigma_k^{h,\mathrm{ref}} + \Delta\sigma_k^h(u)$. The constant quadratic term cancels, and symmetry of $\Pi_k^{h,\mathrm{ref}}$ makes the cross term combine with $\zeta_k^{h,\mathrm{ref}}$ into $q_k^{h,\mathrm{ref}} = \zeta_k^{h,\mathrm{ref}} + \Pi_k^{h,\mathrm{ref}} \sigma_k^{h,\mathrm{ref}}$. The remaining term is the displayed diffusion-increment quadratic. □

Proposition C.23 (Shifted-input and anchored-Hamiltonian consistency)**.** *Let Assumptions C.4, C.8, and C.9 hold, and assume in addition:*

(i) (product integrability) *for some* $\epsilon > 0$*,*

$$\sup_h \big\|\Pi^{h,\mathrm{ref}}\big\|_{L^{4+\epsilon}(\mathrm{d}t \otimes \mathrm{d}\mathbb{P})} + \big\|\Pi^{\mathrm{ref}}\big\|_{L^{4+\epsilon}(\mathrm{d}t \otimes \mathrm{d}\mathbb{P})} < \infty, \quad \sigma^{h,\mathrm{ref}} \longrightarrow \sigma^{\mathrm{ref}} \text{ in } L^4(\mathrm{d}t \otimes \mathrm{d}\mathbb{P});$$

(ii) (compact-control uniformity) *for every compact* $K \subset \mathcal{U}$ *the coefficients* $c \in \{f, b, \sigma\}$ *are Lipschitz in the state-delay-memory argument uniformly over* $u \in K$*, the families* $\sup_{u \in K}\big(|f_k^h(u)| + |b_k^h(u)| + |\sigma_k^h(u)|\big)$ *and their continuous counterparts are bounded in* $L^4(\mathrm{d}t \otimes \mathrm{d}\mathbb{P})$ *uniformly in* h*, and the coefficient consistency of Assumption C.8 holds uniformly on* K*:*

$$\mathbb{E}\int_0^T \sup_{u \in K} \big|c_{k_h(t)}^h(u) - c(t, u)\big|^2 \, \mathrm{d}t \longrightarrow 0,$$

the coefficients being evaluated along the Euler and continuous reference trajectories respectively. Only the candidate slot $u \in K$ *is governed by this condition, and the anchor slot* $u^{h,\mathrm{ref}}$ *is governed separately by the reference-control and coefficient clauses of Assumption C.8, at order* L^2 *for* f *and* b *and at order* L^4 *for* σ*, the latter supplied by (i). In particular* u^{ref} *need not take values in* K*.*

Then the shifted martingale inputs converge and the anchored generalized Hamiltonians are consistent, uniformly on compact control sets in integrated L^1: for every compact $K \subset \mathcal{U}$,

$$\mathbb{E}\int_0^T \left\|\zeta_t^{h,\mathrm{ref}} - \zeta_t^{\mathrm{ref}}\right\|_F^2 \,\mathrm{d}t \longrightarrow 0, \quad \mathbb{E}\int_0^T \sup_{u\in K}\left|\Delta\mathcal{H}_t^h(u) - \Delta\mathcal{H}_t(u)\right| \mathrm{d}t \longrightarrow 0.$$

Proof. All norms in this proof are on $\mathrm{d}t \otimes \mathrm{d}\mathbb{P}$, and reference superscripts are suppressed. Proposition C.21 and interpolation with the additional $L^{4+\epsilon}$ condition give

$$\|\Pi^h - \Pi\|_4 \le \|\Pi^h - \Pi\|_2^{\theta}\|\Pi^h - \Pi\|_{4+\epsilon}^{1-\theta} \to 0, \qquad \theta = \frac{\epsilon}{2(2+\epsilon)}.$$

Since $\zeta^h - \zeta = (q^h - q) - (\Pi^h - \Pi)\sigma^{h,\mathrm{ref}} - \Pi(\sigma^{h,\mathrm{ref}} - \sigma^{\mathrm{ref}})$, Hölder $(4,4)$ and Proposition C.18 imply $\|\zeta^h - \zeta\|_2 \to 0$.

For a compact $K \subset \mathcal{U}$, define the anchored coefficient differences and their envelopes by

$$\Delta c^h(u) = c_k^h(u) - c_k^h(u_k^{h,\mathrm{ref}}), \quad \Delta c(u) = c(t,u) - c(t,u_t^{\mathrm{ref}}),$$
$$B_{c,K}^h = \sup_{u\in K}|\Delta c^h(u)|, \qquad B_{c,K} = \sup_{u\in K}|\Delta c(u)|, \qquad E_{c,K}^h = \sup_{u\in K}|\Delta c^h(u) - \Delta c(u)|,$$

where $k = k_h(t)$ and $c \in \{f, b, \sigma\}$, with Frobenius norms for σ. The triangle inequality separates the candidate slot, controlled by (ii), from the anchor slot, controlled by Assumption C.8 and (i). Hence $\|E_{c,K}^h\|_2 \to 0$, $\sup_h \|B_{b,K}^h\|_2 < \infty$ and $\sup_h \|B_{\sigma,K}^h\|_4 + \|B_{\sigma,K}\|_4 < \infty$. In particular the anchor need not take values in K.

The anchored identity of Lemma C.22 reduces the Hamiltonian error to the reward, drift, diffusion and quadratic brackets. For the two linear brackets use $a_h^\top c_h - a^\top c = (a_h - a)^\top c_h + a^\top(c_h - c)$, and for the quadratic bracket use

$$\Delta\sigma^{h\top}(\Pi^h - \Pi)\Delta\sigma^h + (\Delta\sigma^h - \Delta\sigma)^\top\Pi\Delta\sigma^h + \Delta\sigma^\top\Pi(\Delta\sigma^h - \Delta\sigma).$$

Cauchy–Schwarz and Hölder $(2,4,4)$ therefore give

$$\begin{aligned}
&\mathbb{E}\int_0^T \sup_{u\in K}|\Delta\mathcal{H}_t^h(u) - \Delta\mathcal{H}_t(u)|\,\mathrm{d}t \\
&\quad\le \sqrt{T}\|E_{f,K}^h\|_2 + \|p^h - p\|_2\|B_{b,K}^h\|_2 + \|p\|_2\|E_{b,K}^h\|_2 \\
&\qquad + \|q^h - q\|_2\|B_{\sigma,K}^h\|_2 + \|q\|_2\|E_{\sigma,K}^h\|_2 \\
&\qquad + \tfrac{1}{2}\|\Pi^h - \Pi\|_2\|B_{\sigma,K}^h\|_4^2 + \tfrac{1}{2}\|\Pi\|_4\|E_{\sigma,K}^h\|_2(\|B_{\sigma,K}^h\|_4 + \|B_{\sigma,K}\|_4) \longrightarrow 0.
\end{aligned}$$

The curvature's L^2 limit suffices for this bracket, and its L^4 upgrade was needed only for the shifted-input product above. □

C.8 Completion of Theorem C.1

Proof. Part (a) is Lemmas C.10–C.11: these are fixed-grid chain-rule and projection identities, not refinement claims. Part (b) is Proposition C.18, from lag/operator consistency, the martingale energy estimate and terminal state convergence. Part (c) is Proposition C.21, combining the quadratic master identity, injection duality and integrated tangent consistency, and it uses (b) only in the cell-source comparison (78). Part (d) is Proposition C.23. Thus the main-text Theorem 5.1(a) is part (a) here, and its part (b) collects parts (b)–(d), with their stated additional conditions. Neither reference optimality nor convergence of the full second-order family is used. □

C.9 Finite-branch and nested-regression estimators

The two propositions below are fixed-mesh statements: h and k are fixed, all expectations are the conditional expectations $\mathbb{E}_k^h$ of Appendix C.3, and the separation from the population discretization limit of Theorem 5.1 is by design. Accordingly, they need only the following local integrability of the branch samples, and no mesh-uniform bound.

The Monte Carlo estimators replace $\mathbb{E}_k^h$ by branch averages over the shared buffer,

$$\widehat{p}_k^{\mathrm{ref}} = \frac{\iota_0^\top}{N_{\mathrm{mc}}} \sum_m \widetilde{\Lambda}_k^{(m)}, \quad \widehat{\Pi}_k^{\mathrm{ref}} = \frac{\iota_0^\top}{N_{\mathrm{mc}}} \Big(\sum_m \widetilde{\mathbf{P}}_k^{(m)} \Big) \iota_0, \quad \widehat{q}_k^{\mathrm{ref}} = \frac{1}{N_{\mathrm{mc}} \Delta t} \sum_m \big(\iota_0^\top \widetilde{\Lambda}_{k+1}^{(m)} \big) \Delta B_k^{(m)\top}, \tag{80}$$

Assumption C.24 (Fixed-mesh branch moments)**.** Fix a mesh h and an index k. The raw current gradients $\widehat{p}_k^{h,\mathrm{raw}}$ and $\widehat{p}_{k+1}^{h,\mathrm{raw}}$ have finite conditional fourth moments and the current block $(\iota_0^h)^\top \widetilde{\mathbf{P}}_k^h \, \iota_0^h$ of the raw Hessian has finite conditional second moment, $\mathbb{P}$-almost surely.

Proposition C.25 (Fixed-mesh conditional Monte Carlo consistency)**.** *Fix h and k, and suppose Assumptions C.9 and C.24 hold. The branch estimators $(\widehat{p}_k^{h,\mathrm{ref}}, \widehat{q}_k^{h,\mathrm{ref}}, \widehat{\Pi}_k^{h,\mathrm{ref}})$ of* (80) *are conditionally unbiased for the population targets $(p_k^{h,\mathrm{ref}}, q_k^{h,\mathrm{ref}}, \Pi_k^{h,\mathrm{ref}})$, and there is a finite $\mathcal{F}_{t_k^h}$-measurable C_k^h, equivalently C_k^h-measurable by Assumption C.9, such that*

$$\mathbb{E}_k^h \Big[\big| \widehat{p}_k^{h,\mathrm{ref}} - p_k^{h,\mathrm{ref}} \big|^2 + \big\| \widehat{q}_k^{h,\mathrm{ref}} - q_k^{h,\mathrm{ref}} \big\|_F^2 + \big\| \widehat{\Pi}_k^{h,\mathrm{ref}} - \Pi_k^{h,\mathrm{ref}} \big\|_F^2 \Big] \le \frac{C_k^h}{N_{\mathrm{mc}}}.$$

The constant C_k^h collects conditional second moments of the branch samples and may grow as h decreases, since the q-branch samples carry the normalization h^{-1}, so a joint statement in which $N_{\mathrm{mc}} \to \infty$ and $h \downarrow 0$ simultaneously requires the coupling $\mathbb{E}[C_k^h]/N_{\mathrm{mc}}(h) \to 0$ and is not claimed here.

Proof. By Assumption C.9, the branch samples are conditionally i.i.d. given C_k^h. Concatenate a branch's current gradient, vectorized current Hessian and vectorized covariance sample $h^{-1} \widehat{p}_{k+1}^{h,\mathrm{raw}} \Delta B_k^{h\top}$ into Y. Lemma C.11 identifies $\mathbb{E}_k^h Y$ with the concatenated population targets, so unbiasedness and

$$\mathbb{E}_k^h \left| N_{\mathrm{mc}}^{-1} \sum_m Y^{(m)} - \mathbb{E}_k^h Y \right|^2 = \frac{\mathbb{E}_k^h |Y - \mathbb{E}_k^h Y|^2}{N_{\mathrm{mc}}}$$

give the claim. The covariance sample is square integrable because

$$\mathbb{E}_k^h \| h^{-1} \widehat{p}_{k+1}^{h,\mathrm{raw}} \Delta B_k^{h\top} \|_F^2 \le \frac{\sqrt{d(d+2)}}{h} (\mathbb{E}_k^h |\widehat{p}_{k+1}^{h,\mathrm{raw}}|^4)^{1/2} < \infty.$$

Here $\mathbb{E}|\Delta B_k^h|^4 = d(d+2)h^2$, and the other moments are in Assumption C.24. The conditional variance sum defines C_k^h, and the h^{-1} factor explains why it is not mesh-uniform. □

Anchored nested regression for the shifted input. Fix the conditioning state C_k^h. For $i = 1, \dots, N_{\mathrm{out}}$, draw an outer first-step increment $d_i \sim \mathcal{N}(0, hI_d)$ and pair it with $-d_i$. At the two resulting next buffers, use the same N_{in} inner Brownian continuations with common random numbers, drawn independently across i, and form

$$\bar{\lambda}_{i,\pm} := \frac{1}{N_{\mathrm{in}}} \sum_{j=1}^{N_{\mathrm{in}}} (\iota_0^h)^\top \widetilde{\Lambda}_{k+1}^{h,(i,\pm,j)}, \quad y_i := \frac{\bar{\lambda}_{i,+} - \bar{\lambda}_{i,-}}{2}.$$

Writing

$$m_k(d) := \mathbb{E}_k^h \big[(\iota_0^h)^\top \widetilde{\Lambda}_{k+1}^h \mid \Delta B_k^h = d \big], \quad y_k(d) := \frac{m_k(d) - m_k(-d)}{2},$$

the symmetric Gaussian design gives

$$\mathbb{E}_k^h \big[y_k(d) \, d^\top \big] = \mathbb{E}_k^h \big[(\iota_0^h)^\top \widetilde{\Lambda}_{k+1}^h (\Delta B_k^h)^\top \big] = h \, q_k^{h,\mathrm{ref}}, \quad \mathbb{E}\big[d \, d^\top \big] = hI_d,$$

so the population regression coefficient is exactly $q_k^{h,\mathrm{ref}}$ and the regression carries no fixed-h bias. With the anchor $c_k := \widehat{\Pi}_k^{h,\mathrm{ref}} \sigma_k^{h,\mathrm{ref}}$ and residuals $r_i := y_i - c_k d_i$, the ridge estimator is

$$\widehat{\zeta}_k^{h,\mathrm{ref}} := \Big(\frac{1}{N_{\mathrm{out}}} \sum_{i=1}^{N_{\mathrm{out}}} r_i d_i^\top \Big) \Big(\frac{1}{N_{\mathrm{out}}} \sum_{i=1}^{N_{\mathrm{out}}} d_i d_i^\top + \kappa_Z I_d \Big)^{-1}, \quad \kappa_Z \ge 0.$$

The anchor implements the algebraic decomposition $q_k^{h,\mathrm{ref}} = \Pi_k^{h,\mathrm{ref}}\sigma_k^{h,\mathrm{ref}} + \zeta_k^{h,\mathrm{ref}}$ and does not impose a zero shift. With $\kappa_Z = 0$ the residual regression is algebraically equivalent to the ordinary least-squares estimate of $q_k^{h,\mathrm{ref}}$ followed by subtraction of the plug-in anchor, $\widehat{\zeta} = \widehat{q}_{\mathrm{OLS}} - \widehat{\Pi}\sigma^{h,\mathrm{ref}}$, even when the anchor is computed from the same samples, and sample splitting or cross-fitting between the anchor and the regression is optional but makes the proof and the replication check cleaner. Relative to the Markovian construction of Jeon et al. (2026), the conditioning is the fixed history buffer rather than a Markov state, and the anchor is the effective current block $\widehat{\Pi}\sigma^{\mathrm{ref}}$ rather than a wealth-row product.

Proposition C.26 (Consistency of the anchored nested regression)**.** *Fix h and k, and suppose Assumptions C.9 and C.24 hold. For the Gaussian design above, $G_{N_{\mathrm{out}}} := N_{\mathrm{out}}^{-1}\sum_i d_i d_i^\top \to hI_d$ almost surely. Assume that (i) $N_{\mathrm{out}}, N_{\mathrm{in}} \to \infty$, (ii) the ridge satisfies $\kappa_Z \downarrow 0$, with the inverse $(G_{N_{\mathrm{out}}} + \kappa_Z I_d)^{-1}$ almost surely defined along the asymptotic sequence, which is automatic for $\kappa_Z > 0$ and holds for $\kappa_Z = 0$ once $N_{\mathrm{out}} \geq d$ under the Gaussian design, and (iii) the anchor satisfies $\widehat{\Pi}_k^{h,\mathrm{ref}} \to \Pi_k^{h,\mathrm{ref}}$ in conditional L^2, for instance by Proposition C.25 as $N_{\mathrm{mc}} \to \infty$. Then*

$$\widehat{\zeta}_k^{h,\mathrm{ref}} \longrightarrow \zeta_k^{h,\mathrm{ref}} \quad \textit{in conditional probability},$$

and, if in addition the inverse Gram matrices have a uniform higher moment along the asymptotic sequence, that is, if for some $\epsilon > 0$ and some N_0

$$\sup_{N_{\mathrm{out}} \geq N_0} \mathbb{E}_k^h\Big[\big\|(G_{N_{\mathrm{out}}} + \kappa_Z I_d)^{-1}\big\|^{4+\epsilon}\Big] < \infty,$$

N_0 being chosen large enough for the displayed inverse moment to be finite, for which $N_0 \geq d$ is in particular necessary when the unregularized Gaussian Gram matrix must have full rank, the convergence holds in conditional L^2.

Proof. Condition throughout on C_k^h, and abbreviate $q = q_k^{h,\mathrm{ref}}$, $c = \Pi_k^{h,\mathrm{ref}}\sigma_k^{h,\mathrm{ref}}$, $\widehat{c} = \widehat{\Pi}_k^{h,\mathrm{ref}}\sigma_k^{h,\mathrm{ref}}$, $N = N_{\mathrm{out}}$ and $M = N_{\mathrm{in}}$. Inner averages satisfy $\mathbb{E}[y_i \mid d_i] = y_k(d_i)$, with conditional variance of order M^{-1}, and the two legs share their inner noise but are independent across outer draws. Symmetry gives $\mathbb{E}_k^h[y_k(d)d^\top] = hq$ and $\mathbb{E}_k^h[dd^\top] = hI_d$. Set

$$S_N = N^{-1}\sum_i y_i d_i^\top, \qquad G_N = N^{-1}\sum_i d_i d_i^\top, \qquad B_N = (G_N + \kappa_Z I_d)^{-1}.$$

The strong law gives $G_N \to hI_d$. Split

$$S_N - hq = N^{-1}\sum_i [y_i - y_k(d_i)]d_i^\top + \left[N^{-1}\sum_i y_k(d_i)d_i^\top - hq\right].$$

The bracket converges almost surely by the conditional strong law, while the first term has conditional mean square at most $\widetilde{C}_k^h/(NM)$ for a finite conditional constant $\widetilde{C}_k^h$. Thus $S_N \to hq$ in conditional probability, and arbitrary joint triangular-array sequences are not claimed to converge almost surely.

The estimator equals $(S_N - \widehat{c}G_N)B_N$. Its exact samplewise error is

$$\widehat{\zeta} - \zeta = (S_N - hq)B_N + q(hB_N - I_d) - (\widehat{c} - c)G_N B_N + c\kappa_Z B_N. \tag{81}$$

Because $G_N \succeq 0$, both $G_N B_N$ and $\kappa_Z B_N$ are contractions. No independence between the anchor and regression samples is used. The Gram limit and $\kappa_Z \downarrow 0$ give $B_N \to h^{-1}I_d$ in conditional probability, so the first two terms vanish. The third is bounded by $\|\widehat{c} - c\|_F \to 0$ in conditional L^2 and the fourth vanishes since $\kappa_Z B_N \to 0$. The latter conclusion uses the Gram limit, not just the contraction bound: on a singular direction of G_N, $\kappa_Z B_N$ would equal the identity.

For the conditional L^2 upgrade, the fourth response moments and Gaussian design moments bound S_N in conditional L^r for every $r < 4$, uniformly in N, M. Under the inverse-Gram condition choose $r < 4$ sufficiently close to 4 that $1/r + 1/(4+\epsilon) < 1/2$. Hölder then puts the first two terms of (81) in a uniformly bounded conditional $L^{2+\delta_0}$ for some $\delta_0 > 0$, so their squares are uniformly integrable. For the third term, conditional L^2 convergence of $\widehat{c} - c$ and $\|G_N B_N\| \leq 1$ suffice. The fourth is dominated by $\|c\|_F$. Vitali and dominated convergence complete the upgrade. □

Corollary C.27 (Consistency of the reconstructed anchored q-input)**.** *Under the hypotheses of Proposition C.26, define the reconstructed anchored input*

$$\widehat{q}_k^{\mathrm{anc}} := \widehat{\zeta}_k^{h,\mathrm{ref}} + \widehat{\Pi}_k^{h,\mathrm{ref}}\, \sigma_k^{h,\mathrm{ref}}.$$

Then $\widehat{q}_k^{\mathrm{anc}} \to q_k^{h,\mathrm{ref}}$ in conditional probability, and in conditional L^2 under the inverse-Gram moment condition of that proposition. Using $S_N, G_N, B_N, \widehat{c}$ from the preceding proof, samplewise,

$$\widehat{q}_k^{\mathrm{anc}} = S_N B_N + \widehat{c}\,\kappa_Z B_N,$$

so for $\kappa_Z = 0$ and invertible G_N one has $\widehat{q}_k^{\mathrm{anc}} = S_N G_N^{-1}$, the unanchored ordinary least-squares estimate of $q_k^{h,\mathrm{ref}}$, which does not involve the plug-in anchor at all. For $\kappa_Z > 0$ the anchor enters only through the attenuated term $\widehat{c}\,\kappa_Z B_N$, and the population limit at fixed h and fixed κ_Z carries the ridge bias

$$\lim_{N,M\to\infty} \widehat{q}_k^{\mathrm{anc}} - q_k^{h,\mathrm{ref}} = -\frac{\kappa_Z}{h+\kappa_Z}\,\zeta_k^{h,\mathrm{ref}},$$

so a joint schedule in which h also varies should satisfy $\kappa_Z(h)/h \to 0$.

Proof. Adding $\widehat{c}$ to $(S_N - \widehat{c}G_N)B_N$ and using $G_N B_N = I_d - \kappa_Z B_N$ gives the samplewise identity. Convergence follows from Proposition C.26 and the anchor's conditional L^2 convergence. For fixed ridge substitute $S_N \to hq$, $G_N \to hI_d$ and $\widehat{c} \to c$, obtaining $h(h+\kappa_Z)^{-1}q + \kappa_Z(h+\kappa_Z)^{-1}c - q = -\kappa_Z(h+\kappa_Z)^{-1}(q-c)$. □

Remark C.28 (From fixed-mesh estimators to process statements). The estimator constants may depend on h, k. The supported limit is sequential: at each fixed h, let $N_{\mathrm{mc}}, N_{\mathrm{out}}, N_{\mathrm{in}} \to \infty$ and $\kappa_Z \downarrow 0$, then let $h \downarrow 0$. A finite sum over $k < N_h$ combines the fixed-k results on that mesh. A joint process-norm limit additionally needs unconditional integrability and budget coupling. For example, if $\mathbb{E}C_k^h < \infty$, the direct branch layer requires

$$\frac{h}{N_{\mathrm{mc}}(h)} \sum_{k=0}^{N_h-1} \mathbb{E}C_k^h \longrightarrow 0.$$

The nested layer needs analogous h-weighted control of cross-moment, Gram and anchor errors, as well as vanishing relative ridge bias $\kappa_Z(h)/h \to 0$. No such joint coupling theorem is claimed.

D Recovery stability for the anchored Stage-II solve

D.1 Localization, projection, and proof dependencies

Appendix C identifies the sensitivities to be estimated, and this appendix studies how errors in those inputs and in the local solve affect the recovered control and when the resulting update improves the full objective. It compares paired controls from the same initial history and Brownian motion, with constants uniform over a localized class $\mathcal{U}_0$. It imposes Assumption C.4 policywise, including well-posedness and Lemma C.7, and only the two null sets for a fixed pair are combined. These assumptions are stronger than, and are not imposed on, the fixed-reference Theorem 5.1. Appendix C provides the claim-to-proof map, and the assumptions below specify the separate roles of localization, recovery regularity, quadratic growth, and deployment. The norms are defined in Appendix C.1. We write $\chi = (x, y, m)$ and let w denote a candidate control value.

Assumption D.1 (Common policy and state-history localization)**.** There exist $\mathcal{U}_0 \subset \mathcal{A}$, a closed feasible set $\mathcal{U}_{\mathrm{rec}} \subseteq \mathcal{U}$, a compact control set $K_0 \subset \mathcal{U}_{\mathrm{rec}}$ and a compact state-history region $\mathcal{X}_0$ such that:

(i) Every $u \in \mathcal{U}_0$ is K_0-valued almost everywhere, $u^{\mathrm{ref}}, u^* \in \mathcal{U}_0$, and all exact or numerical recovery outputs considered below remain in K_0.

(ii) Localized primitives $(b^{\mathrm{loc}}, \sigma^{\mathrm{loc}}, f^{\mathrm{loc}})$ agree with (b, σ, f) on a neighborhood of $\mathcal{X}_0 \times K_0$, and g^{loc} agrees with g near the terminal-current projection of $\mathcal{X}_0$. Every jet is an actual derivative of these primitives.

(iii) Localization acts only on state-history: the state coefficients and their jets are globally bounded and Lipschitz in χ, uniformly for $w \in K_0$. It does not remove the control-coercive or strongly concave recovery structure outside K_0.

(iv) All results, including the objective J, concern this localized model on $[0,T]$, uniformly over $\mathcal{U}_0$ and time. Any invocation of Appendix C also uses these primitives, and the superscript loc is suppressed.

Here u^* is a local optimum of the localized objective, and pointwise statements are $\mathrm{d}t \otimes \mathrm{d}\mathbb{P}$-almost everywhere.

Assumption D.2 (Policy-evaluation jet regularity)**.** On $\mathcal{X}_0 \times K_0$, with a common constant and uniformly in $t \in [0,T]$, the following hold. The coefficients b and σ are Lipschitz in (χ, w). The first jets f_χ, $D_\chi b$, $D_\chi \sigma$ are bounded and Lipschitz in (χ, w). The second jets $D^2_{\chi\chi} f$, $D^2_{\chi\chi} b$, $D^2_{\chi\chi}\sigma$ are bounded and Lipschitz in (χ, w). The terminal Hessian g_{xx} is bounded and Lipschitz. By Assumption D.1 these bounds hold globally in χ, uniformly for $w \in K_0$, for the localized primitives, of which the jets above are actual derivatives.

Assumption D.3 (Recovery-control derivative regularity)**.** On an open neighborhood of $\mathbb{R}^{3n} \times K_0$ the localized primitives f, b, σ are twice continuously differentiable in w and continuously differentiable jointly in (χ, w), the first control derivatives f_w, b_w, σ_w are bounded and Lipschitz in χ, and the second control derivatives f_{ww}, b_{ww}, σ_{ww} together with the mixed derivatives $D_\chi f_w$, $D_\chi b_w$, $D_\chi \sigma_w$ are bounded and locally Lipschitz, and all of these hold globally in χ, uniformly for $w \in K_0$ and uniformly in $t \in [0,T]$.

Assumption D.4 (First-adjoint multiplier regularity)**.** The policy-indexed first-order pair is uniformly essentially bounded on the localization:

$$\sup_{u \in \mathcal{U}_0} \left\| p^u \right\|_{\mathcal{S}^\infty} < \infty, \quad \sup_{u \in \mathcal{U}_0} \left\| q^u \right\|_{L^\infty(\mathrm{d}t \otimes \mathrm{d}\mathbb{P})} < \infty.$$

Uniform moments. Lemma C.5 applies policywise with one constant under Assumptions D.1–D.2, and it uses neither mesh hypotheses nor a simultaneous version over all policies.

Assumption D.5 (Uniform regularity of the recovery problem)**.** Let $\mathcal{N}_0 = \mathcal{N}_p \times \mathcal{N}_q \times \mathcal{N}_\Pi$ be a product of deterministic, nonempty, compact, convex sets in $\mathbb{R}^n \times \mathbb{R}^{n\times d} \times \mathbb{S}^n$, compactly contained in an open regularity neighborhood $\mathcal{N}$, and containing (p_t^*, q_t^*, Π_t^*) almost everywhere. Population and estimated inputs are projected blockwise by (84). The feasible set $\mathcal{U}_{\mathrm{rec}}$ is closed and convex, and $\Phi(\cdot; \chi, p, q, \Pi, \bar{w})$ of (82) is μ-strongly concave on it, uniformly in time, $\chi \in \mathbb{R}^{3n}$, $(p, q, \Pi) \in \mathcal{N}_0$, and $\bar{w} \in K_0$, outside a common product-measure null set. No fixed active set or strict complementarity is assumed. Every selected maximizer and numerical output remains in K_0. For $\mathcal{U}_{\mathrm{rec}} = \mathbb{R}^m$, maximizers lie in the interior of K_0. Compact $\mathcal{N}_\Pi$ supplies the curvature bound. Under the additional hypotheses of Proposition D.12(i), raw localized curvatures are also bounded, but containment in a regularity neighborhood with a common μ remains a separate requirement, not a consequence of boundedness.

Assumption D.6 (Local quadratic growth)**.** For some $\kappa_{\mathrm{qg}} > 0$, the localized objective satisfies $J(u^*) - J(u) \geq \kappa_{\mathrm{qg}} \|u - u^*\|^2_{\mathbb{H}^2}$ for all $u \in \mathcal{U}_0$.

Assumption D.7 (Additional assumption for recursive deployment)**.** The conditions above hold uniformly for the restarted problems obtained from every state-history in a compact deployment region, or the recovery map is a history-feedback field that is Lipschitz in state-history and generates a stable closed loop.

Remark D.8 (Transfer to the untruncated model)*.* Coupled forward paths and realized rewards agree with the original model until first exit from the agreement region. This does not imply equality of p, q, Π even on a path that does not exit: their conditional expectations include future paths on the exit event. An exit-tail transfer is separate and is not proved here, and localization is not asserted globally for the benchmarks.

Anchored objective and projection. Suppress the stage time t and write $\vartheta = (p, q, \Pi)$ and $\Delta\sigma(\chi, w; \bar{w}) = \sigma(\chi, w) - \sigma(\chi, \bar{w})$. We write $\Phi(w; \chi, \vartheta, \bar{w})$ also as $\Phi(w; \chi, p, q, \Pi, \bar{w})$. For symmetric Π and $\bar{w} \in K_0$,

$$\Phi(w; \chi, \vartheta, \bar{w}) = H(\chi, w; p, q) - H(\chi, \bar{w}; p, q) + \tfrac{1}{2} \operatorname{tr}[\Delta\sigma(\chi, w; \bar{w})^\top \Pi \Delta\sigma(\chi, w; \bar{w})]. \tag{82}$$

With $\zeta = q - \Pi\sigma(\chi, \bar{w})$, this is the generalized-Hamiltonian difference of Lemma C.22. Thus the main-text $\widehat{\mathcal{H}}$ and $\mathcal{H}^*$ differ from their corresponding Φ only by a control-independent constant. Its gradient is

$$\begin{aligned} D_w\Phi(w;\chi,\vartheta,\bar{w}) &= D_wH(\chi,w;p,q) + (D_w\sigma(\chi,w))^*[\Pi\Delta\sigma(\chi,w;\bar{w})], \\ D_w\Phi(\bar{w};\chi,\vartheta,\bar{w}) &= D_wH(\chi,\bar{w};p,q). \end{aligned} \tag{83}$$

Here $(D_w\sigma)^*$ is the Euclidean–Frobenius adjoint: $[(D_w\sigma)^*M]_a = \langle \partial_{w_a}\sigma, M\rangle_F$. All curvature inputs are symmetrized, since the skew part does not affect Φ. Only the adjoint block is projected:

$$\mathsf{P}(p,q,\Pi) = (\mathrm{Proj}_{\mathcal{N}_p}\, p, \mathrm{Proj}_{\mathcal{N}_q}\, q, \mathrm{Proj}_{\mathcal{N}_\Pi}\, \mathrm{sym}\,\Pi), \qquad \mathrm{sym}\,\Pi = \tfrac{1}{2}(\Pi + \Pi^\top). \tag{84}$$

For estimates, first reconstruct $\widehat{q}^{\,\mathrm{anc}} = \widehat{\zeta} + \widehat{\Pi}\sigma(\chi,\bar{w})$, then set $\widehat{\vartheta}^{\mathcal{N}} = \mathsf{P}(\widehat{p}, \widehat{q}^{\,\mathrm{anc}}, \widehat{\Pi})$ and $\widehat{\zeta}^{\mathcal{N}} = \widehat{q}^{\,\mathrm{anc},\mathcal{N}} - \widehat{\Pi}^{\mathcal{N}}\sigma(\chi,\bar{w})$. The policy and mesh layers use the same map, denoted $\vartheta^{u,\mathcal{N}} = \mathsf{P}(\vartheta^u)$ and $\vartheta^{h,\mathrm{ref},\mathcal{N}} = \mathsf{P}(\vartheta^{h,\mathrm{ref}})$. Each reconstruction uses that layer's own state and anchor: (χ^u, u) for a continuous policy and $(\chi^{h,\mathrm{ref}}, u^{h,\mathrm{ref}})$ for the common mesh comparison. For a different observed state, use that state consistently in both reconstructions and retain its induced defect in the error decomposition. Neither state-history nor anchor is projected.

The map P is measurable and nonexpansive in each block, and $\mathsf{P}(\vartheta^*) = \vartheta^*$. In particular

$$|p^{\mathcal{N}} - p^*| \le |p - p^*|, \qquad \|q^{\mathcal{N}} - q^*\|_F \le \|q - q^*\|_F, \qquad \|\Pi^{\mathcal{N}}\|_F \le \sup_{P\in\mathcal{N}_\Pi} \|P\|_F.$$

Also $\|\Pi^{\mathcal{N}} - \Pi^*\|_F \le \|\,\mathrm{sym}\,\Pi - \Pi^*\|_F$. This blockwise operation preserves separate input errors, whereas a joint projection could mix curvature error into the Brownian input.

Error layers and assumptions used. For any input ϱ, its total error splits into estimation, discretization and reference mismatch:

$$\|\widehat{\varrho} - \varrho^*\| \le \underbrace{\|\widehat{\varrho} - \varrho^{h,\mathrm{ref}}\|}_{\text{estimation}} + \underbrace{\|\varrho^{h,\mathrm{ref}} - \varrho^{\mathrm{ref}}\|}_{\text{discretization}} + \underbrace{\|\varrho^{\mathrm{ref}} - \varrho^*\|}_{\text{reference mismatch}}.$$

The state and first-adjoint policy bounds in $\mathcal{S}^2$ embed into $\mathbb{H}^2$ with factor $\sqrt{T}$. The first state term is zero only when the observed buffer is the same mesh reference buffer. Appendix C.9 controls estimation and Theorem 5.1 controls discretization, and neither controls reference mismatch or recursive deployment by itself.

Forward and injection stability use Assumptions D.1–D.2. First-adjoint and raw-curvature stability add Assumption D.4. Pointwise recovery uses Assumptions D.1, D.2, D.3, and D.5, not a quantitative raw-curvature modulus. Reference substitution adds first-adjoint stability and, only for a performance-gap rate, Assumption D.6. Local loss and dynamic improvement carry the separate conditions stated in Appendix D.5, and recursive deployment additionally requires Assumption D.7.

D.2 Policy-indexed forward and first-adjoint stability

This subsection supplies the reference-mismatch column of the error decomposition for the state history, the first-order pair, and the raw curvature and shifted-input block. The results compare policies from the same initial history and with the same Brownian motion, and constants are uniform over pairs in $\mathcal{U}_0$.

Lemma D.9 (Forward state-history stability under policy perturbation)**.** *Under Assumptions D.1–D.2, for paired $u, u' \in \mathcal{U}_0$ and $r \in \{2, 4\}$,*

$$\|\chi^u - \chi^{u'}\|_{\mathcal{S}^r} \le C_{\mathrm{fw},r}\|u - u'\|_{\mathbb{H}^r}.$$

Write $C_{\mathrm{fw}} = C_{\mathrm{fw},2}$. Constants depend only on the common jet bounds, $\kappa_{\rho,\delta}, T$ and dimensions, not on the policy pair, and no smallness of the delay is required.

Proof. Write $\Delta X = X^u - X^{u'}$, $\Delta\chi = \chi^u - \chi^{u'}$ and $\Delta u = u - u'$. Since initial histories coincide, $\Delta Y(s) = \mathbf{1}_{\{s>\delta\}}\Delta X(s-\delta)$ and $|\Delta M(s)| \le \kappa_{\rho,\delta}\sup_{v\le s}|\Delta X(v)|$. Thus $|\Delta\chi(s)| \le (2 + \kappa_{\rho,\delta})\sup_{v\le s}|\Delta X(v)|$. Subtract the state equations. Both coefficient differences are bounded

by $C(|\Delta\chi| + |\Delta u|)$. Burkholder–Davis–Gundy, Hölder and the channel bound yield, for $\phi_r(s) = \mathbb{E}\sup_{v\le s}|\Delta X(v)|^r$,

$$\phi_r(s) \le C\int_0^s \phi_r(\ell)\,\mathrm{d}\ell + C\|\Delta u\|_{\mathbb{H}^r}^r, \qquad r = 2, 4.$$

Gronwall gives the current-state estimate, and the channel bound gives the result. □

Proposition D.10 (Anticipated first-adjoint stability under policy perturbation)**.** *Let Assumptions D.1, D.2, and D.4 hold, let $u, u' \in \mathcal{U}_0$ be paired as in Lemma D.9, and let (p^u, q^u) and $(p^{u'}, q^{u'})$ be the anticipated first-order pairs of* (5) *along the two evaluation tuples for the localized primitives. With $\Delta p := p^u - p^{u'}$ and $\Delta q := q^u - q^{u'}$,*

$$\big\|\Delta p\big\|_{\mathcal{S}^2} + \big\|\Delta q\big\|_{\mathbb{H}^2} \le C_{\mathrm{adj}}\,\big\|\Delta u\big\|_{\mathbb{H}^2},$$

with C_{adj} depending only on the jet bounds, the multiplier bounds of Assumption D.4, $\kappa_{\rho,\delta}$, T, and the dimensions. In particular $\|\Delta p\|_{\mathbb{H}^2} \le \sqrt{T}\,\|\Delta p\|_{\mathcal{S}^2}$, which is the embedded bound used in the error decomposition. No smallness condition on the delay is used.

Proof. Use $\Delta\chi, \Delta u$ as in Lemma D.9 and $\Delta p = p^u - p^{u'}$, $\Delta q = q^u - q^{u'}$. For $c \in \{x, y, m\}$ split the difference of Hamiltonian jets into

$$a_c = D_c b(\chi^u, u)^\top \Delta p + \sum_j D_c\sigma^j(\chi^u, u)^\top \Delta q^j,$$
$$b_c = H_c(\chi^u, u; p^{u'}, q^{u'}) - H_c(\chi^{u'}, u'; p^{u'}, q^{u'}).$$

Here stage time is suppressed. Bounded jets imply $|a_c| \le C(|\Delta p| + \|\Delta q\|_F)$, and their Lipschitz bounds and the *raw* multiplier bounds of Assumption D.4 imply $|b_c| \le C(|\Delta\chi| + |\Delta u|)$. With the operators of Appendix C.5,

$$-\mathrm{d}\Delta p = (A + B)\,\mathrm{d}t - \Delta q\,\mathrm{d}B_t, \quad A = a_x + \mathcal{T}_\delta a_y + \mathcal{K}a_m, \quad B = b_x + \mathcal{T}_\delta b_y + \mathcal{K}b_m,$$

where A, B here are driver components, not the forward kernels or Brownian motion. The terminal difference satisfies $|\Delta p_T| \le C|\Delta X(T)|$.

For $\beta > 0$, the tower property and conditional Jensen give, on any backward interval $[t, T]$,

$$\mathbb{E}\int_t^T e^{\beta s}|\mathcal{T}_\delta Z(s)|^2\,\mathrm{d}s \le e^{-\beta\delta}\mathbb{E}\int_{(t+\delta)\wedge T}^T e^{\beta r}|Z(r)|^2\,\mathrm{d}r,$$
$$\mathbb{E}\int_t^T e^{\beta s}|\mathcal{K}Z(s)|^2\,\mathrm{d}s \le \kappa_{\rho,\delta}^2\mathbb{E}\int_t^T e^{\beta r}|Z(r)|^2\,\mathrm{d}r.$$

The second inequality is weighted Cauchy–Schwarz in the memory integral, then Fubini, and both reindexed weights are independent of a small-delay condition. Apply Itô to $e^{\beta t}|\Delta p_t|^2$. Its quadratic variation supplies $\int e^{\beta s}\|\Delta q_s\|_F^2\,\mathrm{d}s$ on the left. Young's inequality with small enough coefficients, followed by the two operator estimates, gives

$$2\mathbb{E}\int_t^T e^{\beta s}\langle\Delta p_s, A(s) + B(s)\rangle\,\mathrm{d}s \le \tfrac{1}{2}\mathbb{E}\int_t^T e^{\beta s}\|\Delta q_s\|_F^2\,\mathrm{d}s + C\mathbb{E}\int_t^T e^{\beta s}(|\Delta p_s|^2 + |\Delta\chi(s)|^2 + |\Delta u(s)|^2)\,\mathrm{d}s.$$

Choose $\beta > 2C$ to absorb the level term. The terminal bound and forward stability imply

$$\sup_t \mathbb{E}|\Delta p_t|^2 + \|\Delta q\|_{\mathbb{H}^2}^2 \le C(\|\Delta\chi\|_{\mathcal{S}^2}^2 + \|\Delta u\|_{\mathbb{H}^2}^2) \le C\|\Delta u\|_{\mathbb{H}^2}^2.$$

Finally use $\Delta p_t = \Delta p_T + \int_t^T (A + B)\,\mathrm{d}s - (M_T - M_t)$, $M_t = \int_0^t \Delta q_s\,\mathrm{d}B_s$. Burkholder–Davis–Gundy bounds its martingale supremum, and the operator bounds with $\beta = 0$ bound the driver integral, upgrading the level estimate to $\mathcal{S}^2$. The multiplier assumption makes the stochastic terms true martingales, and all constants are uniform over the paired policies. □

Lemma D.11 (Policy perturbation of the current-injection flow)**.** *Let Assumptions D.1 and D.2 hold, let $u, v \in \mathcal{U}_0$ be paired as in Lemma D.9, and write $\Delta u := u - v$, $\Delta\chi := \chi^u - \chi^v$. Then:*

(a) $\|\Delta\chi\|_{\mathcal{S}^4} \le C\,\|\Delta u\|_{\mathbb{H}^4}$.

(b) *for every basis direction e_i, the coupled current-injection flows of Lemma C.5 satisfy, with $\Delta\mathcal{V}^{t,e_i} := \mathcal{V}^{u,t,e_i} - \mathcal{V}^{v,t,e_i}$,*

$$\operatorname*{ess\,sup}_{t\in[0,T]} \mathbb{E}\Big[\sup_{t\le s\le T} \big|\Delta\mathcal{V}^{t,e_i}(s)\big|^2\Big] \le C\,\|\Delta u\|_{\mathbb{H}^4}^2,$$

the fixed-anchor estimate below holding for every deterministic t with the same constant, independent of i.

Both constants depend only on the jet bounds, $\kappa_{\rho,\delta}$, T, and the dimensions.

Proof. Part (a) is Lemma D.9 at $r = 4$. For (b), the common injection cancels, so the delayed and memory differences are the delayed value and window integral of the current difference, with zero initial value. Subtract (61), whose sources are $(L_b^u - L_b^v)\mathcal{V}^{v,t,e_i}$ and $(L_\sigma^{j,u} - L_\sigma^{j,v})\mathcal{V}^{v,t,e_i}$. Writing $g = |\Delta\chi| + |\Delta u|$, their energy is bounded by

$$C\mathbb{E}\left[\left(\int_t^T g(s)^2\,\mathrm{d}s\right)\sup_{t\le s\le T}|\mathcal{V}^{v,t,e_i}(s)|^2\right] \le C\left[\mathbb{E}\left(\int_t^T g^2\right)^2\right]^{1/2} \le C\|\Delta u\|_{\mathbb{H}^4}^2.$$

The first inequality uses conditional/unconditional fourth moments from Lemma C.5, and the last uses (a). Burkholder–Davis–Gundy and delay-Gronwall for the homogeneous current error give the stated second-moment supremum estimate, uniformly in t, i. The channel bound supplies the full-flow estimate. □

Proposition D.12 (Policy stability of the effective curvature and the shifted input)**.** *Let Assumptions D.1, D.2, and D.4 hold, and let $u, v \in \mathcal{U}_0$ be paired as above. For $w \in \{u, v\}$ write $D^2_{\chi\chi}H^w(s)$ for the lifted state-history Hessian of the Hamiltonian along $(\chi^w(s), w(s))$ with multipliers (p_s^w, q_s^w), $\mathbf{g}^{(2),w} := \mathrm{blkdiag}(g_{xx}(X^w(T)), 0, 0)$, and*

$$\widehat{\mathcal{Q}}_t^w(e_i, e_j) := \big(\mathcal{V}^{w,t,e_i}(T)\big)^\top \mathbf{g}^{(2),w}\,\mathcal{V}^{w,t,e_j}(T) + \int_t^T \big(\mathcal{V}^{w,t,e_i}(s)\big)^\top D^2_{\chi\chi}H^w(s)\,\mathcal{V}^{w,t,e_j}(s)\,\mathrm{d}s,$$

so that applying Lemma C.7 to each policy gives $e_i^\top \Pi_t^w e_j = \mathbb{E}_t\big[\widehat{\mathcal{Q}}_t^w(e_i, e_j)\big]$, and set $\zeta_t^w := q_t^w - \Pi_t^w\,\sigma(\chi^w(t), w(t))$. Then:

(i) $\sup_{w\in\mathcal{U}_0}\|\Pi^w\|_{L^\infty(\mathrm{d}t\otimes\mathrm{d}\mathbb{P})} \le C_\Pi$.

(ii) $\|\Pi^u - \Pi^v\|_{\mathbb{H}^2} \le C\,\|\Delta u\|_{\mathbb{H}^4}$.

(iii) $\|\zeta^u - \zeta^v\|_{\mathbb{H}^2} \le C\,\|\Delta u\|_{\mathbb{H}^4}$.

All constants depend only on the jet and multiplier bounds, $\kappa_{\rho,\delta}$, T, and the dimensions, and the difference estimates (ii)–(iii) are linear in $\|\Delta u\|_{\mathbb{H}^4}$.

Proof. Abbreviate $V_i^w = \mathcal{V}^{w,t,e_i}$ and $M^w = D^2_{\chi\chi}H^w$. Bounded second jets and raw multipliers give $\|M^w\|_{L^\infty} + \|\mathbf{g}^{(2),w}\|_{L^\infty} \le C$. The quadratic identification and conditional moments then yield

$$|e_i^\top\Pi_t^w e_j| \le C(1+T)\big(\mathbb{E}_t\sup_s|V_i^w(s)|^2\big)^{1/2}\big(\mathbb{E}_t\sup_s|V_j^w(s)|^2\big)^{1/2} \le C_\Pi,$$

proving (i) in fixed dimension. For either a running or terminal coefficient M, the difference of bilinear forms is

$$(V_i^u)^\top(M^u - M^v)V_j^u + (V_i^u - V_i^v)^\top M^v V_j^u + (V_i^v)^\top M^v(V_j^u - V_j^v).$$

The source coefficients satisfy

$$\|M^u - M^v\|_F \le C(|\Delta\chi| + |\Delta u| + |\Delta p| + \|\Delta q\|_F), \quad \|\mathbf{g}^{(2),u} - \mathbf{g}^{(2),v}\|_F \le C|\Delta X(T)|.$$

Conditional Cauchy–Schwarz bounds the running source contribution to $\mathbb{E}_t[\widehat{\mathcal{Q}}_t^u - \widehat{\mathcal{Q}}_t^v]$ by $C(\mathbb{E}_t \int_t^T \|M^u - M^v\|_F^2 \,\mathrm{d}s)^{1/2}$, and the terminal source by $C(\mathbb{E}_t|\Delta X(T)|^2)^{1/2}$. Each tangent leg is bounded by $C(\mathbb{E}_t \sup_{s\geq t} |V_i^u(s) - V_i^v(s)|^2)^{1/2}$ or its j analogue, using bounded coefficients and the common conditional moments. Square, integrate in t, and use the tower property. The source terms are controlled by Lemma D.9 and Proposition D.10, and the tangent legs are controlled by Lemma D.11. Thus

$$\|\Pi^u - \Pi^v\|_{\mathbb{H}^2}^2 \leq C(\|\Delta u\|_{\mathbb{H}^2}^2 + \|\Delta u\|_{\mathbb{H}^4}^2) \leq C\|\Delta u\|_{\mathbb{H}^4}^2.$$

Finally $\Delta\zeta = \Delta q - (\Pi^u - \Pi^v)\sigma(\chi^u, u) - \Pi^v[\sigma(\chi^u, u) - \sigma(\chi^v, v)]$. Bounded localized diffusion, (i), Lipschitz coefficients and the previous estimates give (iii). □

Corollary D.13 (Half-Hölder curvature modulus on the localized class)**.** *In the setting of Proposition D.12, with $D_0 := \operatorname{diam} K_0$,*

$$\|\Pi^u - \Pi^v\|_{\mathbb{H}^2} + \|\zeta^u - \zeta^v\|_{\mathbb{H}^2} \leq C\, D_0^{1/2}\, \|u - v\|_{\mathbb{H}^2}^{1/2}.$$

Proof. Since both controls are K_0-valued, $A := \int_0^T |u-v|^2 \,\mathrm{d}t \leq D_0^2 T$ and $\mathbb{E}A^2 \leq D_0^2 T\, \mathbb{E}A$. Hence $\|u - v\|_{\mathbb{H}^4} \leq D_0^{1/2} T^{1/4} \|u - v\|_{\mathbb{H}^2}^{1/2}$, and substituting in Proposition D.12(ii)–(iii) and absorbing $T^{1/4}$ gives the claim. □

D.3 Anchored recovery with input and numerical errors

This subsection bounds the error of the recovered control relative to the optimal control, accounting for discrepancies in the state history, the first-order inputs, and the anchor together with the numerical residual, and these bounds yield the input-to-control estimates used in Proposition 5.2. For a closed convex set $C \subseteq \mathbb{R}^m$ and $u \in C$, use the outward normal cone $N_C(u) := \{\nu : \langle \nu, w - u\rangle \leq 0 \text{ for all } w \in C\}$. Thus concave maximization is characterized by $D_w\Phi(u) \in N_C(u)$. The following estimate includes the exact solve as the case of zero numerical residual.

Proposition D.14 (Anchored recovery under perturbed inputs and an inexact solve)**.** *Let Assumptions D.1, D.2, D.3, and D.5 hold. For almost every stage (t, ω), every $\chi \in \mathbb{R}^{3n}$, $\vartheta = (p, q, \Pi) \in \mathcal{N}_0$, and $\bar{w} \in K_0$, the objective* (82) *has a unique maximizer*

$$\mathsf{R}_t(\chi, \vartheta, \bar{w}) := \operatorname*{arg\,max}_{w \in \mathcal{U}_{\mathrm{rec}}} \Phi(w; \chi, \vartheta, \bar{w}) \in K_0.$$

The same abbreviated notation is used for R_t. Write $\mathsf{R} = \mathsf{R}_t(\chi, \vartheta, \bar{w})$ and define the pointwise input mismatch with the optimal tuple of the localized model by

$$e_* := |\chi - \chi_t^*| + |p - p_t^*| + \|q - q_t^*\|_F + |\bar{w} - u_t^*|.$$

Then $|\mathsf{R} - u_t^*| \leq \mu^{-1} C e_*$. For any feasible numerical output $\widehat{u} \in \mathcal{U}_{\mathrm{rec}} \cap K_0$, let

$$r_{\mathrm{num}} := \operatorname{dist}\big(D_w\Phi(\widehat{u}; \chi, \vartheta, \bar{w}), N_{\mathcal{U}_{\mathrm{rec}}}(\widehat{u})\big).$$

The additional bounds are

$$\text{(i)} \quad |\widehat{u} - \mathsf{R}| \leq \mu^{-1} r_{\mathrm{num}}, \qquad \text{(ii)} \quad |\widehat{u} - u_t^*| \leq \mu^{-1}(C e_* + r_{\mathrm{num}}).$$

The constant C depends only on the common evaluation/recovery jet bounds, the radii of $\mathcal{N}_p, \mathcal{N}_q, \mathcal{N}_\Pi$, K_0, and the dimensions. There is no additive curvature difference $\|\Pi - \Pi_t^*\|_F$: the estimate is uniform over the admissible curvature block, under the common strong-concavity constant μ.

Proof. Existence and K_0-retention are part of Assumption D.5, and strong concavity gives uniqueness. The maximum condition (8) for the localized model makes u_t^* the unique maximizer for its optimal tuple. Suppress t and write $\Phi_* := \Phi(\,\cdot\,; \chi^*, \vartheta^*, u^*)$. The variational inequalities at R and u^* imply

$$\begin{aligned}\mu|\mathsf{R} - u^*|^2 &\leq \langle D_w\Phi(u^*) - D_w\Phi(\mathsf{R}), \mathsf{R} - u^*\rangle \\ &\leq \langle D_w\Phi(u^*) - D_w\Phi_*(u^*), \mathsf{R} - u^*\rangle.\end{aligned}$$

By (83), the gradient difference here is

$$\begin{aligned}&D_w H(\chi, u^*; p, q) - D_w H(\chi^*, u^*; p^*, q^*) \\ &\quad + \big(D_w\sigma(\chi, u^*)\big)^* \big[\Pi\{\sigma(\chi, u^*) - \sigma(\chi, \bar{w})\}\big].\end{aligned}$$

The optimal quadratic correction vanishes at its own anchor. The first line is bounded by $C(|\chi - \chi^*| + |p - p^*| + \|q - q^*\|_F)$, using the recovery jets and the compact multiplier blocks, and the second is bounded by $C|\bar{w} - u^*|$, using the bounded admissible Π and control Lipschitz continuity of σ. This gives Ce_* and hence the exact bound, with no comparison of two curvature matrices.

For (i), choose $\nu \in N_{\mathcal{U}_{\mathrm{rec}}}(\widehat{u})$ attaining the distance r_{num}, which exists because the cone is closed and nonempty. Strong concavity, the variational inequality at R, and the normal inequality tested at R yield

$$\mu|\widehat{u} - \mathsf{R}|^2 \le \langle D_w\Phi(\mathsf{R}) - D_w\Phi(\widehat{u}), \widehat{u} - \mathsf{R}\rangle \le \langle \nu - D_w\Phi(\widehat{u}), \widehat{u} - \mathsf{R}\rangle \le r_{\mathrm{num}}|\widehat{u} - \mathsf{R}|.$$

Division proves (i) when the displacement is nonzero, and (i) is trivial otherwise. The triangle inequality gives (ii). Finally, the maximization can be restricted to the deterministic compact set K_0. Joint measurability in time and continuity in the remaining arguments give a measurable unique-maximizer map. Fixing any K_0 value on the exceptional stage set yields a progressively measurable, square-integrable output on progressively measurable data. □

The anchored identity makes r_{num} exactly the generalized-Hamiltonian residual in (p, ζ, Π) coordinates, with $\zeta = q - \Pi\sigma(\chi, \bar{w})$ reconstructed from the same state and anchor. Since $0 \in N_{\mathcal{U}_{\mathrm{rec}}}(\widehat{u})$, $r_{\mathrm{num}} \le |D_w\Phi(\widehat{u})|$ is bounded under the localization.

Corollary D.15 (Integrated input error bounds and reference-tuple specialization)**.** *In the proposition's setting, let the observed state $\widehat{\chi}$, projected block $\widehat{\vartheta}^{\mathcal{N}} = (\widehat{p}^{\mathcal{N}}, \widehat{q}^{\mathrm{anc},\mathcal{N}}, \widehat{\Pi}^{\mathcal{N}}) \in \mathcal{N}_0$, anchor $\bar{w} \in K_0$, and output $\widehat{u} \in \mathcal{U}_{\mathrm{rec}} \cap K_0$ be progressively measurable. Use* (84) *with the same $(\widehat{\chi}, \bar{w})$ in both reconstructions, and suppose the displayed errors are square integrable. For the associated measurable residual process,*

$$\|\widehat{u} - u^*\|_{\mathbb{H}^2} \le \mu^{-1}C\big(\|\widehat{\chi} - \chi^*\|_{\mathbb{H}^2} + \|\widehat{p}^{\mathcal{N}} - p^*\|_{\mathbb{H}^2} + \|\widehat{q}^{\mathrm{anc},\mathcal{N}} - q^*\|_{\mathbb{H}^2} + \|\bar{w} - u^*\|_{\mathbb{H}^2}\big) + \mu^{-1}\|r_{\mathrm{num}}\|_{\mathbb{H}^2}.$$

Each input error splits into estimator, discretization, and reference mismatch as in the error decomposition, and blockwise projection does not enlarge the p and q errors relative to the optimal block. If Assumption D.4 also holds and $u^{\mathrm{ref}} \in \mathcal{U}_0$, then

$$\mathsf{R}_t^{\mathrm{ref}} := \mathsf{R}_t(\chi_t^{\mathrm{ref}}, \mathsf{P}(\vartheta_t^{\mathrm{ref}}), u_t^{\mathrm{ref}}), \qquad \|\mathsf{R}^{\mathrm{ref}} - u^*\|_{\mathbb{H}^2} \le C\|u^{\mathrm{ref}} - u^*\|_{\mathbb{H}^2}.$$

These are paired-trajectory statements, and recursive deployment additionally requires Assumption D.7.

Proof. Integrate Proposition D.14(ii) by Minkowski's inequality and use the nonexpansiveness of (84) with $\mathsf{P}(\vartheta^*) = \vartheta^*$. The three layers are a triangle inequality through the mesh and continuous reference tuples. For the reference specialization take zero residual, then apply Lemma D.9 and Proposition D.10, with $\|Y\|_{\mathbb{H}^2} \le \sqrt{T}\|Y\|_{\mathcal{S}^2}$. □

Corollary D.16 (Exact-anchor invariance and residual certificate)**.** *In the setting of Proposition D.14, with μ the common strong-concavity constant of $\Phi(\,\cdot\,; \chi, p, q, \Pi, \bar{w})$ on $\mathcal{U}_{\mathrm{rec}}$ over $(p, q, \Pi) \in \mathcal{N}_p \times \mathcal{N}_q \times \mathcal{N}_\Pi$ and anchors $\bar{w} \in K_0$, for $\mathrm{d}t \otimes \mathrm{d}\mathbb{P}$-almost every (t, ω):*

(i) *If $(\chi, p, q, \bar{w}) = (\chi_t^*, p_t^*, q_t^*, u_t^*)$, the maximizer of $\Phi(\,\cdot\,; \chi_t^*, p_t^*, q_t^*, \widetilde{\Pi}, u_t^*)$ over $\mathcal{U}_{\mathrm{rec}}$ is u_t^* for every $\widetilde{\Pi} \in \mathcal{N}_\Pi$.*

(ii) *For any inputs (χ, p, q, Π), the exact maximizer $\widetilde{u}$ of $\Phi = \Phi(\,\cdot\,; \chi, p, q, \Pi, u^{\mathrm{ref}})$ over $\mathcal{U}_{\mathrm{rec}}$ and a numerical output $\widehat{u}$ with residual r_{num} satisfy*

$$\Phi(\widetilde{u}) - \Phi(u^{\mathrm{ref}}) \ge \tfrac{\mu}{2}|\widetilde{u} - u^{\mathrm{ref}}|^2, \qquad |\widehat{u} - \widetilde{u}| \le \mu^{-1}r_{\mathrm{num}}, \qquad 0 \le \Phi(\widetilde{u}) - \Phi(\widehat{u}) \le \frac{r_{\mathrm{num}}^2}{2\mu}.$$

Proof. Part (i) is the exact recovery bound with $e_* = 0$. For (ii), strong concavity and the variational inequality at $\widetilde{u}$ give $\Phi(w) \le \Phi(\widetilde{u}) - \mu|w - \widetilde{u}|^2/2$, and $w = u^{\mathrm{ref}}$ gives the first inequality. The second is Proposition D.14(i). For the last one, a normal vector attaining the residual gives $\langle D_w\Phi(\widehat{u}), \widetilde{u} - \widehat{u}\rangle \le r_{\mathrm{num}}|\widetilde{u} - \widehat{u}|$. With $d_0 := |\widetilde{u} - \widehat{u}|$, strong concavity at $\widehat{u}$ therefore yields

$$0 \le \Phi(\widetilde{u}) - \Phi(\widehat{u}) \le r_{\mathrm{num}}d_0 - \frac{\mu}{2}d_0^2 \le \frac{r_{\mathrm{num}}^2}{2\mu}.$$

Subtracting this loss from the exact gain gives the supplied-input local-gain certificate

$$\Phi(\widehat{u}) - \Phi(u^{\mathrm{ref}}) \geq \frac{\mu}{2}|\widetilde{u} - u^{\mathrm{ref}}|^2 - \frac{r_{\mathrm{num}}^2}{2\mu}.$$

□

D.4 Performance-to-input transfer and the recovered-policy budget

The reference gap enters only through quadratic growth and the policy-stability results. We record the error layers once, prove the transfer of the reference gap to the inputs, and then assemble the proof of Proposition 5.2 from the recovery bound of Corollary D.15 and that transfer.

For the deployed tuple of Corollary D.15, compared through the mesh reference layer $(\chi^{h,\mathrm{ref}}, u^{h,\mathrm{ref}})$ and the continuous reference layer $(\chi^{\mathrm{ref}}, u^{\mathrm{ref}})$ with the anchor $\bar{w} = u^{h,\mathrm{ref}}$, the layer aggregates are

$$\begin{aligned}
\delta_\chi &:= \|\widehat{\chi} - \chi^{h,\mathrm{ref}}\|_{\mathbb{H}^2} + \|\chi^{h,\mathrm{ref}} - \chi^{\mathrm{ref}}\|_{\mathbb{H}^2} + \|\chi^{\mathrm{ref}} - \chi^*\|_{\mathbb{H}^2},\\
\delta_p &:= \|\widehat{p} - p^{h,\mathrm{ref}}\|_{\mathbb{H}^2} + \|p^{h,\mathrm{ref}} - p^{\mathrm{ref}}\|_{\mathbb{H}^2} + \|p^{\mathrm{ref}} - p^*\|_{\mathbb{H}^2},\\
\delta_q &:= \|\widehat{q}^{\mathrm{anc}} - q^{h,\mathrm{ref}}\|_{\mathbb{H}^2} + \|q^{h,\mathrm{ref}} - q^{\mathrm{ref}}\|_{\mathbb{H}^2} + \|q^{\mathrm{ref}} - q^*\|_{\mathbb{H}^2},\\
\delta_{\mathrm{anc}} &:= \|u^{h,\mathrm{ref}} - u^{\mathrm{ref}}\|_{\mathbb{H}^2} + \|u^{\mathrm{ref}} - u^*\|_{\mathbb{H}^2}, \qquad \delta_{\mathrm{num}} := \|r_{\mathrm{num}}\|_{\mathbb{H}^2},
\end{aligned} \tag{85}$$

in the norms of the error decomposition, and the estimator and discretization aggregates are

$$\begin{aligned}
\delta_{\mathrm{est}} &:= \|\widehat{\chi} - \chi^{h,\mathrm{ref}}\|_{\mathbb{H}^2} + \|\widehat{p} - p^{h,\mathrm{ref}}\|_{\mathbb{H}^2} + \|\widehat{q}^{\mathrm{anc}} - q^{h,\mathrm{ref}}\|_{\mathbb{H}^2},\\
\delta_{\mathrm{disc}} &:= \|\chi^{h,\mathrm{ref}} - \chi^{\mathrm{ref}}\|_{\mathbb{H}^2} + \|p^{h,\mathrm{ref}} - p^{\mathrm{ref}}\|_{\mathbb{H}^2} + \|q^{h,\mathrm{ref}} - q^{\mathrm{ref}}\|_{\mathbb{H}^2} + \|u^{h,\mathrm{ref}} - u^{\mathrm{ref}}\|_{\mathbb{H}^2}.
\end{aligned} \tag{86}$$

Proposition D.17 (Performance-to-input transfer)**.** *Let Assumptions D.1, D.2, D.4, and D.6 hold, let $u^{\mathrm{ref}} \in \mathcal{U}_0$, and write $\varepsilon_{\mathrm{ref}} := J(u^*) - J(u^{\mathrm{ref}}) \geq 0$ for the localized performance gap. Then*

$$\big\|u^{\mathrm{ref}} - u^*\big\|_{\mathbb{H}^2} \leq \big(\varepsilon_{\mathrm{ref}}/\kappa_{\mathrm{qg}}\big)^{1/2},$$

$$\big\|\chi^{\mathrm{ref}} - \chi^*\big\|_{\mathcal{S}^2} + \big\|p^{\mathrm{ref}} - p^*\big\|_{\mathcal{S}^2} + \big\|q^{\mathrm{ref}} - q^*\big\|_{\mathbb{H}^2} \leq C\big(\varepsilon_{\mathrm{ref}}/\kappa_{\mathrm{qg}}\big)^{1/2},$$

while for the curvature and shifted input the analysis yields the weaker reference-mismatch bound

$$\big\|\Pi^{\mathrm{ref}} - \Pi^*\big\|_{\mathbb{H}^2} + \big\|\zeta^{\mathrm{ref}} - \zeta^*\big\|_{\mathbb{H}^2} \leq C\big(\varepsilon_{\mathrm{ref}}/\kappa_{\mathrm{qg}}\big)^{1/4}$$

(Proposition D.12, Corollary D.13). All three are global process-norm estimates, and none implies a $\mathrm{d}t \otimes \mathrm{d}\mathbb{P}$-pointwise gap.

Proof. The first display rearranges Assumption D.6. The second follows by inserting it into Lemma D.9 and Proposition D.10. The third follows by inserting it into Corollary D.13. □

With the layers of (85) and the transfer of Proposition D.17 in place, the recovered-policy budget of the main text follows by substituting the reference-layer bounds into the recovery bound of Corollary D.15.

Proposition D.18 (Recovered-policy and true-local-loss error budgets)**.** *Let the assumptions of Proposition D.14 hold together with Assumptions D.4 and D.6, let $\widehat{u}$ be the deployed Stage-II control in the setting of Corollary D.15, anchored at $u^{h,\mathrm{ref}}$ with $u^{\mathrm{ref}} \in \mathcal{U}_0$, and let $\varepsilon_{\mathrm{ref}}$ and the layers $\delta_{\mathrm{est}}, \delta_{\mathrm{disc}}, \delta_{\mathrm{num}}$ be as in* (14)*.*

(i) (layerwise error budget)

$$\big\|\widehat{u} - u^*\big\|_{\mathbb{H}^2} \leq \mu^{-1} C\Big[\delta_{\mathrm{est}} + \delta_{\mathrm{disc}} + \big(\varepsilon_{\mathrm{ref}}/\kappa_{\mathrm{qg}}\big)^{1/2}\Big] + \mu^{-1}\delta_{\mathrm{num}}.$$

(ii) (curvature uniformity) *The constants in (i) and, under its local-gradient condition, (iii) can be chosen independently of the admissible block $\widehat{\Pi}^{\mathcal{N}} \in \mathcal{N}_\Pi$, with no separate term in $\|\widehat{\Pi}^{\mathcal{N}} - \Pi^*\|$: by the anchored cancellation of Remark D.19, the quarter-power curvature bound of Proposition D.17 does not degrade the recovered-policy bound.*

(iii) (integrated true-input local loss) *Let* $G_t^{\mathrm{loc},*} := \mathcal{H}_t^*(u_t^*) - \mathcal{H}_t^*(\widehat{u}_t)$ *be the local loss of the deployed control under the optimal tuple. Under the local-gradient condition of Lemma D.20,*

$$0 \le \mathbb{E}\int_0^T G_t^{\mathrm{loc},*}\,\mathrm{d}t \le C_G\,\eta^2, \qquad \eta := \delta_{\mathrm{est}} + \delta_{\mathrm{disc}} + \big(\varepsilon_{\mathrm{ref}}/\kappa_{\mathrm{qg}}\big)^{1/2} + \delta_{\mathrm{num}},$$

which is $O(\varepsilon_{\mathrm{ref}})$ *when the other layers vanish, for the local loss and not for* $J(u^*) - J(\widehat{u})$.

Proof of Proposition D.18. For (i), apply Corollary D.15 with anchor $\bar{w} = u^{h,\mathrm{ref}}$ and split its four input errors through the mesh and continuous reference layers of (85). The state and anchor give δ_χ and δ_{anc}, and projection together with $\mathsf{P}(\vartheta^*) = \vartheta^*$ gives $\|\widehat{p}^{\mathcal{N}} - p^*\|_{\mathbb{H}^2} \le \delta_p$ and $\|\widehat{q}^{\mathrm{anc},\mathcal{N}} - q^*\|_{\mathbb{H}^2} \le \delta_q$, so that

$$\big\|\widehat{u} - u^*\big\|_{\mathbb{H}^2} \le \mu^{-1}C\big(\delta_\chi + \delta_p + \delta_q + \delta_{\mathrm{anc}}\big) + \mu^{-1}\delta_{\mathrm{num}},$$

which holds without Assumptions D.4 and D.6. Under these two assumptions Proposition D.17 bounds every reference-layer term by $C(\varepsilon_{\mathrm{ref}}/\kappa_{\mathrm{qg}})^{1/2}$, using $\mathcal{S}^2 \hookrightarrow \mathbb{H}^2$ for the state and p, and the remaining terms are exactly (86). Part (ii) is the curvature-uniform anchored estimate of Proposition D.14, and any ridge-mediated curvature error in reconstructed q is already in δ_{est} by Corollary C.27, rather than a new additive term. For (iii), Lemma D.20 gives $\|e\|_{\mathbb{H}^2} \le \delta_\chi + \delta_p + \delta_q + \delta_{\mathrm{anc}}$ and $\|r\|_{\mathbb{H}^2} = \delta_{\mathrm{num}}$, and substituting the reference bound of (i) into the squared estimate of that lemma shows that one may take $C_G = C_0 \max\{C_a C, 1\}^2$. These are process-norm estimates for controls coupled through the same initial history and Brownian motion, and recursively deployed recovery additionally requires Assumption D.7. □

Constants and scope. The constant C in the input-wise bound of the proof above depends only on the common evaluation/recovery jet bounds, projected-input radii, K_0, T, and dimensions. The reference-rate form in Proposition D.18(i) also uses the raw multiplier bounds and delay mass $\kappa_{\rho,\delta}$, and C_G additionally uses μ^{-1} and L_Φ. Apart from the displayed μ^{-1} and κ_{qg} factors, the constants do not depend on the mesh (hence H_h), branch budgets, reference gap, or numerical residual. This is recovery-bound uniformity, not mesh-uniform Monte Carlo variance.

Remark D.19 (Why the curvature drops out). The quadratic correction is stationary at the anchor, and Assumption D.5 provides strong monotonicity uniformly over $\mathcal{N}_\Pi$, including across active-set changes. No quantitative curvature mismatch is needed in the policy budget, but the projected curvature still determines the supplied objective and its residual. Corollary C.27 shows that its plug-in effect on $\widehat{q}^{\mathrm{anc}}$ is only the ridge term, which vanishes samplewise for unregularized full-rank regression and is already included in δ_{est}.

When the adjoint acquisition is evaluated along the mesh reference rollout, $\widehat{\chi} = \chi^{h,\mathrm{ref}}$. The first layer of δ_χ and of δ_{est} then vanishes and $\widehat{q}^{\mathrm{anc}}$ is exactly the estimator of Corollary C.27, so Appendix C.9 supplies the remaining estimator layers under its fixed-mesh conditional consistency and sequential branch and ridge limits. The generic observed-state version keeps δ_χ and treats the induced anchor-reconstruction defect inside δ_q, while recursive deployment generally produces a different observed buffer and is covered only under Assumption D.7.

D.5 True-input local loss and improvement of the corrected control

This subsection supplies the two remaining ingredients of the main text. Lemma D.20 transfers the recovery budget to the local loss under the optimal tuple, which gives Proposition D.18(iii), and Lemma D.21 expands the objective J along the segment from the reference control to the corrected control, which combines with the strong concavity of the supplied objective into Theorem 5.3.

Additional local-gradient condition. Only for the local-loss extension, assume that the control gradients of the supplied and optimal-tuple objectives are uniformly L_Φ-Lipschitz on $\mathrm{co}\,K_0$, the convex hull of K_0. This follows from the recovery-control jet bounds if they hold on a common neighborhood of $\mathrm{co}\,K_0$, and otherwise it is an additional local hypothesis. Constants are uniform over the fixed input block $\mathcal{N}_0$.

Lemma D.20 (True-input residual and local-loss transfer)**.** *In the setting of Proposition D.14, let*

$$\widehat{\Phi}_t(w) := \Phi\big(w; \widehat{\chi}_t, \widehat{p}_t^{\mathcal{N}}, \widehat{q}_t^{\mathrm{anc},\mathcal{N}}, \widehat{\Pi}_t^{\mathcal{N}}, \bar{w}_t\big), \qquad \Phi_t^*(w) := \Phi\big(w; \chi_t^*, p_t^*, q_t^*, \Pi_t^*, u_t^*\big).$$

Suppose the additional local-gradient condition holds and define

$$e_t := |\widehat{\chi}_t - \chi_t^*| + |\widehat{p}_t^{\mathcal{N}} - p_t^*| + \|\widehat{q}_t^{\mathrm{anc},\mathcal{N}} - q_t^*\|_F + |\bar{w}_t - u_t^*|.$$

For a feasible numerical output $\widehat{u}_t \in \mathcal{U}_{\mathrm{rec}} \cap K_0$, let $r_t := \operatorname{dist}\big(D_w\widehat{\Phi}_t(\widehat{u}_t), N_{\mathcal{U}_{\mathrm{rec}}}(\widehat{u}_t)\big)$ and $r_t^* := \operatorname{dist}\big(D_w\Phi_t^*(\widehat{u}_t), N_{\mathcal{U}_{\mathrm{rec}}}(\widehat{u}_t)\big)$. There is a constant C_a, independent of the chosen admissible curvature block, such that, for $\mathrm{d}t \otimes \mathrm{d}\mathbb{P}$-almost every (t,ω),

$$r_t^* \le \Big(1 + \frac{2L_\Phi}{\mu}\Big)(C_a e_t + r_t), \qquad 0 \le G_t^{\mathrm{loc},*} \le \frac{1}{2\mu}\Big(1 + \frac{2L_\Phi}{\mu}\Big)^2 (C_a e_t + r_t)^2,$$

where $G_t^{\mathrm{loc},*} := \Phi_t^*(u_t^*) - \Phi_t^*(\widehat{u}_t)$. Consequently, with $C_0 := (2\mu)^{-1}(1 + 2L_\Phi/\mu)^2$,

$$\mathbb{E}\int_0^T G_t^{\mathrm{loc},*}\,\mathrm{d}t \le C_0\big(C_a\|e\|_{\mathbb{H}^2} + \|r\|_{\mathbb{H}^2}\big)^2.$$

Proof. Fix a stage outside the common null set and suppress t. The anchored-gradient comparison in the proof of Proposition D.14 gives $|D_w\widehat{\Phi}(u^*) - D_w\Phi^*(u^*)| \le C_a e$: at the optimal anchor the quadratic correction is stationary by (83), whereas the supplied correction at u^* is bounded by $C|\bar{w} - u^*|$ using the bounded admissible curvature and the control Lipschitz bound of σ, so no difference $\widehat{\Pi}^{\mathcal{N}} - \Pi^*$ is needed. The same strong-monotonicity comparison, including the numerical residual, yields $|\widehat{u} - u^*| \le \mu^{-1}(C_a e + r)$. Choose $\nu \in N_{\mathcal{U}_{\mathrm{rec}}}(\widehat{u})$ attaining the supplied residual. Then

$$\begin{aligned} r^* &\le |D_w\Phi^*(\widehat{u}) - \nu| \le r + |D_w\Phi^*(\widehat{u}) - D_w\widehat{\Phi}(\widehat{u})| \\ &\le r + C_a e + 2L_\Phi|\widehat{u} - u^*| \le (1 + 2L_\Phi/\mu)(C_a e + r). \end{aligned}$$

For a normal vector ν^* attaining r^*, strong concavity of Φ^* and feasibility of u^* give, with $d := |\widehat{u} - u^*|$,

$$G^{\mathrm{loc},*} \le \langle D_w\Phi^*(\widehat{u}), u^* - \widehat{u}\rangle - \tfrac{\mu}{2}d^2 \le r^* d - \tfrac{\mu}{2}d^2 \le \frac{(r^*)^2}{2\mu}.$$

Nonnegativity holds because u^* maximizes the true objective on the same feasible recovery set. Squaring the residual bound, integrating, and applying the $\mathbb{H}^2$ triangle inequality proves the process estimate. □

At a constraint boundary $D_w\Phi^*(u^*)$ need not vanish, so a loss bound quadratic in the control error cannot be inferred from the control error alone. The lemma instead transfers the small supplied-input residual of the recovered control to the true-input residual, which is why no active-set fixing or strict complementarity is needed.

Comparison segments for dynamic improvement. For the remaining statements, controls are compared on the same Brownian probability space and from the same initial history. For each pair (u, v) considered, require $u_s := u + s(v - u) \in \mathcal{U}_0$ for every $s \in [0,1]$, with Assumptions D.1, D.2, D.3, and D.4 holding uniformly along this segment, where the control-derivative bounds are required on its common control localization, including the control-coordinate line segments used below, and J is finite there. These are additional comparison conditions, not consequences of convexity of $\mathcal{U}_{\mathrm{rec}}$ alone. The interpolation is used only in the proof, and the algorithm applies v in full.

Lemma D.21 (First variation and a quadratic remainder)**.** *Under the comparison-segment conditions, define $g_t^a := D_w H(\chi_t^a, a_t; p_t^a, q_t^a)$ for a segment control a. For $\Delta := v - u$ and $u_s := u + s\Delta$, the localized objective has the feasible-segment derivative*

$$\frac{\mathrm{d}}{\mathrm{d}s}J(u_s) = \mathbb{E}\int_0^T \langle g_t^{u_s}, \Delta_t\rangle\,\mathrm{d}t,$$

with one-sided derivatives at the endpoints, and $\|g^{u_s} - g^{u_r}\|_{\mathbb{H}^2} \le L_J|s - r|\,\|\Delta\|_{\mathbb{H}^2}$ for a deterministic constant L_J uniform on the stated comparison localization. Hence

$$\Big|J(v) - J(u) - \mathbb{E}\int_0^T \langle g_t^u, \Delta_t\rangle\,\mathrm{d}t\Big| \le \frac{L_J}{2}\,\|\Delta\|_{\mathbb{H}^2}^2.$$

Proof. Fix a segment control $a = u_s$. The first state variation z in direction Δ has zero prehistory and, with $\mathfrak{z}_t = (z_t, z_{t-\delta}, \int_{(t-\delta)\vee 0}^{t} e^{-\rho(t-r)} z_r \, \mathrm{d}r)$, solves

$$\mathrm{d}z_t = (D_\chi b_t^a \mathfrak{z}_t + D_w b_t^a \Delta_t)\,\mathrm{d}t + \sum_j (D_\chi \sigma_t^{j,a} \mathfrak{z}_t + D_w \sigma_t^{j,a} \Delta_t)\,\mathrm{d}B_t^j.$$

Integral Taylor remainders, the bounded jets and compact control range, and the Burkholder–Davis–Gundy–Gronwall estimate of Lemma D.9 imply convergence of the state difference quotients in $\mathcal{S}^2$. Thus

$$DJ(a)[\Delta] = \mathbb{E}\left[\int_0^T \{\langle D_\chi f_t^a, \mathfrak{z}_t\rangle + \langle D_w f_t^a, \Delta_t\rangle\}\,\mathrm{d}t + \langle g_x(X_T^a), z_T\rangle\right].$$

Itô's formula for $\langle p_t^a, z_t\rangle$ cancels the current-state terms. The anticipated terms cancel by the two adjoint identities

$$\mathbb{E}\int_0^T \langle z_t, (\mathcal{T}_\delta F)(t)\rangle\,\mathrm{d}t = \mathbb{E}\int_0^T \langle z_{t-\delta}, F(t)\rangle\,\mathrm{d}t,$$
$$\mathbb{E}\int_0^T \langle z_t, (\mathcal{K} F)(t)\rangle\,\mathrm{d}t = \mathbb{E}\int_0^T \left\langle \int_{(t-\delta)\vee 0}^{t} e^{-\rho(t-r)} z_r\,\mathrm{d}r, F(t)\right\rangle \mathrm{d}t,$$

applied to $F = H_y^a$ and H_m^a, respectively. These follow by the tower property, Fubini, and zero prehistory. Localization and the moment bounds remove the martingale contribution. Since $p_T^a = g_x(X_T^a)$, substitution gives $DJ(a)[\Delta] = \langle g^a, \Delta\rangle_{\mathbb{H}^2}$ without assuming that a is optimal. This identity gives the marginal effect of a feasible control perturbation on the objective, including its propagation through the delay channels, and is the continuous-time form of the sensitivities that Stage II computes at each history.

For two segment controls a, b, the control-jet and multiplier bounds give

$$|g_t^a - g_t^b| \le C_{\mathrm{jet}}(|\chi_t^a - \chi_t^b| + |a_t - b_t|) + B_b|p_t^a - p_t^b| + B_\sigma \|q_t^a - q_t^b\|_F.$$

Here $C_{\mathrm{jet}} = L_{f_u} + L_{b_u} P_\infty + L_{\sigma_u} Q_\infty$, where the L's are the control-jet Lipschitz constants, P_∞, Q_∞ the common multiplier bounds, and B_b, B_σ bounds on $D_w b, D_w \sigma$. Lemma D.9 and Proposition D.10 permit the uniform choice

$$L_J = C_{\mathrm{jet}}(1 + \sqrt{T} C_{\mathrm{fw}}) + (B_b\sqrt{T} + B_\sigma) C_{\mathrm{adj}}.$$

This supplies the continuous $\mathbb{H}^2$ gradient along the segment, and no simultaneous pointwise adjoint version over all policies is needed. Finally,

$$|J(v) - J(u) - \langle g^u, \Delta\rangle_{\mathbb{H}^2}| \le \int_0^1 \|g^{u_s} - g^u\|_{\mathbb{H}^2} \|\Delta\|_{\mathbb{H}^2}\,\mathrm{d}s \le \frac{L_J}{2}\|\Delta\|_{\mathbb{H}^2}^2. \qquad \square$$

With the first variation and the quadratic remainder of Lemma D.21 in place, the improvement theorem follows by evaluating the first variation through the supplied objective at the anchor, where strong concavity turns the anchored variational inequality into a gain of order D^2 against which the remainder and the gradient and solver defects are measured.

Proof of Theorem 5.3. Set $u = u^{\mathrm{ref}}$, $v = \widehat{u}$, $\Delta = v - u$, $D = \|\Delta\|_{\mathbb{H}^2}$, $\tau = \delta_{\mathrm{num}}$, and $C_J = L_J/2$. Write $\widehat{\Phi}_t(w) = \widehat{\mathcal{H}}_t(w) - \widehat{\mathcal{H}}_t(u_t)$ for the anchored version of the supplied objective in the main text, whose control gradient is unchanged. Write $r_t := \mathrm{dist}\big(D_w \widehat{\Phi}_t(v_t), N_{\mathcal{U}_{\mathrm{rec}}}(v_t)\big)$ and, for almost every stage, select $\nu_t \in N_{\mathcal{U}_{\mathrm{rec}}}(v_t)$ attaining this distance. Since $u_t \in \mathcal{U}_{\mathrm{rec}}$, $\langle \nu_t, v_t - u_t\rangle \ge 0$. Strong concavity of the supplied objective thus gives

$$\langle D_w \widehat{\Phi}_t(u_t), \Delta_t\rangle \ge \mu|\Delta_t|^2 + \langle D_w \widehat{\Phi}_t(v_t), \Delta_t\rangle \ge \mu|\Delta_t|^2 - r_t|\Delta_t|.$$

This pointwise inequality needs no jointly selected normal-vector process, and its measurable upper-bound quantities are integrated directly. By the definition of η_∇ and Cauchy–Schwarz, $\mathbb{E}\int_0^T \langle g_t^u, \Delta_t\rangle\,\mathrm{d}t \ge \mu D^2 - (\eta_\nabla + \tau)D$, and the remainder estimate of Lemma D.21 proves $J(v) - J(u) \ge (\mu - C_J)D^2 - (\eta_\nabla + \tau)D$, from which the strict-improvement condition follows.

For the exact reference-trajectory specialization, take $v_t = \mathsf{R}_t(\chi_t^u, p_t^u, q_t^u, \widetilde{\Pi}_t, u_t)$ with $\widetilde{\Pi}_t \in \mathcal{N}_\Pi$, $(p_t^u, q_t^u, \widetilde{\Pi}_t) \in \mathcal{N}_0$, and the first-adjoint coordinates unaltered by input projection. The anchored identity (83) gives $D_w\widehat{\Phi}_t(u_t) = D_w H(\chi_t^u, u_t; p_t^u, q_t^u) = g_t^u$ independently of $\widetilde{\Pi}_t$, and exact maximization gives $r_t = 0$, so $DJ(u)[v-u] \geq \mu\|v-u\|_{\mathbb{H}^2}^2$ and $J(v) - J(u) \geq (\mu - C_J)\|v-u\|_{\mathbb{H}^2}^2$. The constants are uniform over the admissible curvature choices for which the common recovery and comparison-segment assumptions hold, and no quadratic-growth assumption is needed for this one-update result. □

Remark D.22 (A structural example for the margin). Suppose the localized objective has the form $J(u) = J_{\mathrm{state}}(u) - \frac{r}{2}\|u\|_{\mathbb{H}^2}^2$ with $r > 0$, where J_{state} collects the running and terminal rewards of the state path and its $\mathbb{H}^2$ gradient is L_{state}-Lipschitz on the comparison localization. Then $g^a = \nabla J_{\mathrm{state}}(a) - r\,a$ along the segment, so the gradient-stability bound of Lemma D.21 holds with $L_J = r + L_{\mathrm{state}}$ and the quadratic-remainder constant is $C_J = L_J/2 = (r + L_{\mathrm{state}})/2$. If moreover b and σ are affine in the control and the supplied curvature block satisfies $\widehat{\Pi}^{\mathcal{N}} \preceq 0$, the sign that does not weaken concavity, the control Hessian of $\widehat{\Phi}_t$ equals $-rI + \sum_j (D_w\sigma^j)^\top \widehat{\Pi}^{\mathcal{N}} D_w\sigma^j \preceq -rI$, so $\mu \geq r$ and $\mu - C_J \geq (r - L_{\mathrm{state}})/2$. Thus $r > L_{\mathrm{state}}$ secures a positive margin, and Theorem 5.3 yields strict improvement once $D > 0$ and $\eta_\nabla + \tau < (\mu - C_J)D$, where $\tau = \delta_{\mathrm{num}}$ in the notation of the main text.

Acquisition and deployment defects. Continue with $u = u^{\mathrm{ref}}$, $v = \widehat{u}$, $D = \|v-u\|_{\mathbb{H}^2}$, and $\tau = \delta_{\mathrm{num}}$, where $\bar{w}$ is the actual recovery anchor. By (83), the gradient defect is bounded by

$$\eta_\nabla \leq C_e\big(\|\widehat{\chi} - \chi^u\|_{\mathbb{H}^2} + \|\widehat{p}^{\mathcal{N}} - p^u\|_{\mathbb{H}^2} + \|\widehat{q}^{\mathrm{anc},\mathcal{N}} - q^u\|_{\mathbb{H}^2} + \|\bar{w} - u\|_{\mathbb{H}^2}\big).$$

There is no separate curvature-mismatch term, but projection bias in the first-adjoint coordinates and any bias in their anchored reconstruction remain in this bound, and the comparator is the exact reference gradient, not the optimal-tuple gradient. Recursive deployment generally changes the observed history and the reference anchor, even with exact conditional estimation, and these changes are retained in η_∇. On each fixed grid the same argument applies to the discrete objective J_h with its own remainder constant, and any mismatch between the discrete gradient and the supplied recovery gradient is likewise retained in η_∇.