LEARNING TO DRAW ON THE CURVED SURFACE: A RIEMANNIAN ADAPTIVE MOMENT ITERATIVE LEARNING CONTROL APPROACH

Anonymous authors
Paper under double-blind review

ABSTRACT

Precise trajectory tracking on curved surfaces is essential in robotic operations such as drawing, polishing, and manufacturing. While classical Euclidean iterative learning control (ILC) improves repeatable tracking accuracy through trial-to-trial error correction, it ignores the underlying surface geometry, often producing command updates that violate geometric feasibility. This paper proposes R-ILC, a Riemannian adaptive moment ILC framework that embeds surface constraints directly into the command update rule via Riemannian optimization. Theoretical analysis establishes almost-sure convergence of the Riemannian gradient norm to zero, with avoidance of strict saddles. Experiments on curved-surface robotic drawing validate the effectiveness of our approach while demonstrating its advantages over classical Euclidean ILC in tracking accuracy and geometric feasibility.

1 INTRODUCTION

Operations such as manufacturing, polishing, and drawing require a tool to follow a prescribed path while maintaining the intended contact with a curved workpiece. In robotic polishing, the tool must follow the curved surface and maintain contact to make it smooth (Tian et al., 2016); in surface drawing, maintaining contact is necessary to reproduce the desired contour (Song & Kim, 2019). A trajectory that departs from the surface can fail to perform the task even when it remains close to the desired path. Therefore, trajectory learning for curved-surface tasks must ensure both tracking accuracy and geometric feasibility.

To improve tracking accuracy over repeated executions of these tasks, iterative learning control (ILC) uses information from previous trials to refine the input sequence (Bristow et al., 2006). For curved-surface tasks, the surface geometry constrains the admissible set of end-effector positions (Song & Kim, 2019). However, the iterative control laws used in classical Euclidean ILC do not inherently preserve geometric feasibility (Chu & Owens, 2010; Chen et al., 2020). Even when tracking error decreases, the resulting trajectories may violate the surface constraints during execution, which can cause the task to fail. Because each updated command is used in the next trial, geometric feasibility must be maintained throughout the learning process.

Existing work on ILC has recognized the need to incorporate geometric constraints into learning. Geometric learning laws have been developed for surface-constrained motion (Arimoto et al., 2010). Related surface-following methods use surface models to guide robot motion, but errors in these models can limit tracking accuracy (Ubbink et al., 2024). Contact-control methods also account for interaction with the surface, yet can produce path deviations when the surface curvature changes (Wahballa et al., 2025). These findings show that accounting for surface constraints does not by itself ensure accurate tracking. This motivates a learning approach that reduces tracking error while preserving surface feasibility across trials.

Our central idea is to incorporate surface feasibility into the trajectory update itself. We formulate learning on a smooth manifold of admissible commands, where each point represents a full input sequence whose nominal output satisfies the surface constraints. The intrinsic geometry of this manifold guides both the update and the use of information from previous trials. Learning thus seeks to reduce tracking loss while keeping successive input sequences within the feasible domain (Figure 1).

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.

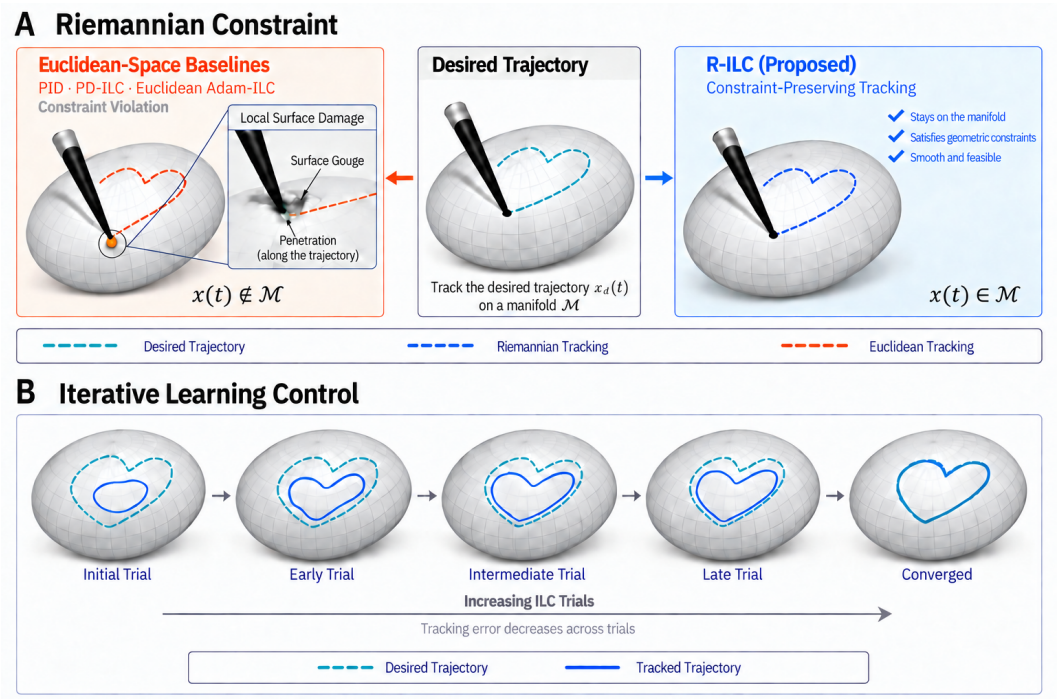

Figure 1: The schematic illustrates R-ILC trajectory learning across trials under surface constraints. (A) On the output surface $\mathcal{M}$, whose constraint induces the feasible command manifold, Euclidean updates can violate feasibility while geometric updates preserve it. (B) Repeated ILC trials progressively refine the pen-tip output $x(t)$ toward the desired output $x_d(t)$.

This perspective leads to the central question of this work:

How can we integrate intrinsic geometry into ILC updates to improve tracking accuracy while preserving geometric feasibility across trials?

The main contributions of our work are as follows:

1. We formulate surface-constrained ILC on a product manifold of admissible input sequences and develop R-ILC. By combining transported momentum, factorwise adaptive scaling, and geometric updates, the approach uses trial history to learn the full input sequence while preserving nominal surface feasibility at every trial (Sections 2.4 and 3.1).
2. We establish convergence guarantees for nonconvex stochastic R-ILC. Specifically, we prove almost-sure and mean-square convergence of the Riemannian gradient norms to zero, derive an almost-sure running-minimum rate, and establish almost-sure nonconvergence to a prescribed finite or countable set of strict saddles (Section 3.2).
3. We evaluate R-ILC on curved-surface drawing through simulation and physical execution. Experiments show that R-ILC achieves lower final tracking error and surface deviation than Euclidean Adam-ILC, PD-ILC, and PID while maintaining surface feasibility to numerical precision throughout learning (Section 4).

2 PRELIMINARIES AND SETUP

2.1 ITERATIVE LEARNING CONTROL

ILC updates an entire input sequence between successive executions using the tracking error from the completed trial (Bristow et al., 2006). For a trial $k \geq 1$ lasting N control steps, let $u_k(t) \in \mathbb{R}^{d_u}$ denote the input command, $h_k(t) \in \mathbb{R}^{d_h}$ the plant state, and $y_k(t) \in \mathbb{R}^{d_y}$ the measured output. The plant response is modeled by

$$h_k(t+1) = \mathcal{F}(h_k(t), u_k(t)) + w_k(t), \qquad y_k(t) = \mathcal{H}(h_k(t)) + \nu_k(t). \tag{2.1}$$

The terms $w_k(t)$ and $\nu_k(t)$ account for process and measurement noise, respectively. Each trial starts from $h_k(0) = h_0$; the possibly nonlinear maps $\mathcal{F}, \mathcal{H}$, the reference, and any within-trial feedback law remain fixed. The N commands are applied at $t = 0, \ldots, N-1$, and the resulting output record is sampled at $t = 1, \ldots, N$.

Collecting the commands and output samples over a trial gives the lifted trajectories:

$$\mathbf{u}_k := \left[u_k(0)^\top, \ldots, u_k(N-1)^\top\right]^\top \in \mathbb{R}^{d_u N}, \qquad \mathbf{y}_k := \left[y_k(1)^\top, \ldots, y_k(N)^\top\right]^\top \in \mathbb{R}^{d_y N}. \tag{2.2}$$

Let $\mathbf{y}^{\mathrm{des}} := \left[y^{\mathrm{des}}(1)^\top, \ldots, y^{\mathrm{des}}(N)^\top\right]^\top$ denote the fixed desired output trajectory, and define the trial tracking error $\mathbf{e}_k := \mathbf{y}^{\mathrm{des}} - \mathbf{y}_k$. An ILC law uses the completed-trial error to construct the next input sequence and may retain additional information from earlier trials.

2.2 MOMENTUM AND ADAPTIVE SCALING

Gradient-based ILC can retain information from earlier trials by aggregating estimates of the tracking-loss gradient when constructing the next command update. Optimization methods such as AdaGrad (Duchi et al., 2011), RMSProp (Hong & Lin, 2024), and Adam (Kingma & Ba, 2015) provide ways to use this gradient history. We introduce momentum and adaptive scaling into ILC to smooth gradient fluctuations across trials and account for differences in gradient magnitude along the input sequence.

Let $g_k \in \mathbb{R}^{d_u N}$ be a Euclidean gradient estimate of the trial tracking loss. Momentum replaces the current gradient by a weighted average (Polyak, 1964):

$$m_k = \beta m_{k-1} + (1-\beta) g_k, \qquad m_0 = 0, \quad \beta \in [0, 1).$$

The weight β controls how quickly contributions from earlier trials decay.

For adaptive scaling, a nonnegative statistic records the squared magnitude of each gradient component. It may accumulate the full history, $v_k = \sum_{j=1}^{k} g_j^{\odot 2}$, or use exponential averaging,

$$v_k = \theta v_{k-1} + (1-\theta) g_k^{\odot 2}, \qquad v_0 = 0, \quad \theta \in [0, 1),$$

where $g_k^{\odot 2}$ denotes coordinatewise squaring and θ controls the decay of past squared-gradient contributions. Using $m_k/(\sqrt{v_k}+\epsilon)$ in place of the raw gradient combines these mechanisms. Square roots and division are coordinatewise, and $\epsilon > 0$ prevents division by zero. This rescales the averaged direction so that coordinates with larger historical gradient magnitudes receive smaller multipliers.

2.3 RIEMANNIAN GEOMETRY

The preceding formulas accumulate and rescale gradients in a fixed vector space. When geometric constraints restrict admissible input sequences to a manifold, intrinsic gradients at successive input sequences lie in different tangent spaces, and a finite additive correction may leave the feasible set. Riemannian geometry provides the tools to measure these directions, compare them across trials, and construct updates within the manifold (Boumal, 2023).

To describe variables whose components satisfy separate manifold constraints, let $(\mathcal{M}_i, \langle\cdot,\cdot\rangle^i)$, $i = 1, \ldots, n$, be Riemannian manifolds and equip $\mathcal{M} = \prod_{i=1}^{n} \mathcal{M}_i$ with the product metric. At $x =$

$(x^1, \dots, x^n) \in \mathcal{M}$, tangent vectors are velocities of curves through x. For $a, b \in T_x\mathcal{M}$, the product structure gives

$$T_x\mathcal{M} = \bigoplus_{i=1}^{n} T_{x^i}\mathcal{M}_i, \qquad \langle a, b\rangle_x = \sum_{i=1}^{n} \langle a^i, b^i\rangle^i_{x^i}. \tag{2.3}$$

This metric defines $\|a\|_x^2 = \langle a, a\rangle_x$ and induces the geodesic distance $d(x, y)$.

For differentiable $f : \mathcal{M} \to \mathbb{R}$, its Riemannian gradient is the unique vector $\operatorname{grad} f(x) \in T_x\mathcal{M}$ satisfying

$$\mathrm{d}f_x[a] = \langle \operatorname{grad} f(x), a\rangle_x, \qquad a \in T_x\mathcal{M}.$$

Thus $-\operatorname{grad} f(x)$ is a descent direction whenever it is nonzero. For twice differentiable f, the Hessian $\operatorname{Hess} f(x)[a] = \nabla_a \operatorname{grad} f(x)$ describes how the gradient varies along a tangent direction, where ∇ is the Levi-Civita connection.

Directions from different trials must be expressed in the same tangent space before they can be combined. Parallel transport $P_{x\to y} : T_x\mathcal{M} \to T_y\mathcal{M}$ carries a tangent vector along the chosen geodesic while preserving inner products. To turn a tangent correction into a new manifold point, the exponential map $\operatorname{Exp}_x(a)$ follows the geodesic from x with initial velocity a to time one, wherever defined; Log_x is its local inverse. These operations act factorwise on the product manifold.

2.4 Problem Setup

Our objective is to minimize expected tracking loss while ensuring that the nominal output trajectory satisfies the prescribed geometric constraints. Let $\mathbf{u} \in \mathbb{R}^{d_u N}$ denote a full input sequence as in (2.2), and let υ collect the process and measurement noise over one trial. The lifted output of (2.1) is denoted by $\mathcal{Y}(\mathbf{u}; \upsilon)$. The nominal response $\mathcal{Y}_0(\mathbf{u}) := \mathcal{Y}(\mathbf{u}; 0)$ corresponds to zero process and measurement noise.

Let $\mathcal{S}_y \subset \mathbb{R}^{d_y}$ be the prescribed surface for each nominal output sample. We assume that the command domain $\mathcal{M}$ has the product structure and metric of Section 2.3. Each point $\mathbf{u} = (u^1, \dots, u^n)$, with $u^i \in \mathcal{M}_i$, is identified with the corresponding full input sequence in (2.2). Geometric feasibility requires

$$\mathcal{M} \subseteq \left\{\mathbf{u} \in \mathbb{R}^{d_u N} : \mathcal{Y}_0(\mathbf{u}) \in \mathcal{S}_y^N\right\}. \tag{2.4}$$

With the reference fixed, let $\mathcal{L}$ be a differentiable tracking loss and define $\ell(\mathbf{u}; \upsilon) := \mathcal{L}\big(\mathbf{y}^{\mathrm{des}} - \mathcal{Y}(\mathbf{u}; \upsilon)\big)$. The trajectory optimization problem is

$$\min_{\mathbf{u}\in\mathcal{M}} f(\mathbf{u}), \qquad f(\mathbf{u}) := \mathbb{E}_\upsilon[\ell(\mathbf{u}; \upsilon)], \tag{2.5}$$

where the expectation is over the fixed trial-noise law. For the noise realization υ_k in trial k, the observed output $\mathbf{y}_k = \mathcal{Y}(\mathbf{u}_k; \upsilon_k)$ yields the trial loss $\ell_k = \mathcal{L}(\mathbf{e}_k)$.

3 Algorithm and Analysis

3.1 Riemannian Adaptive Moment ILC

R-ILC addresses (2.5) by optimizing the full input sequence $\mathbf{u} \in \mathcal{M}$ across repeated trials. Executing $\mathbf{u}_k$ yields the observed output $\mathbf{y}_k$ and trial loss ℓ_k. These data, together with the required model or sensitivity information, are used to form $g_k \in T_{\mathbf{u}_k}\mathcal{M}$, a stochastic estimate of the Riemannian gradient $\operatorname{grad} f(\mathbf{u}_k)$ of the expected tracking loss.

R-ILC uses trial history to adapt the direction and scale of command corrections while preserving geometric feasibility. Transported momentum combines information from successive trials in a common tangent space, while one scalar second-moment estimate per factor $\mathcal{M}_i$ accounts for differences in gradient magnitude across command blocks (Bécigneul & Ganea, 2019).

The first- and second-moment weights are $\beta_k, \theta_k \in [0, 1)$, respectively, and no bias correction is applied. We express this factorwise adaptive scaling through the linear operator $D_{\mathbf{u}}(v) : T_{\mathbf{u}}\mathcal{M} \to T_{\mathbf{u}}\mathcal{M}$, defined for $\epsilon > 0$, $v \in [0, \infty)^n$, and $a \in T_{\mathbf{u}}\mathcal{M}$ by

$$(D_{\mathbf{u}}(v)a)^i = \frac{a^i}{\sqrt{v^i + \epsilon}}, \qquad i = 1, \dots, n.$$

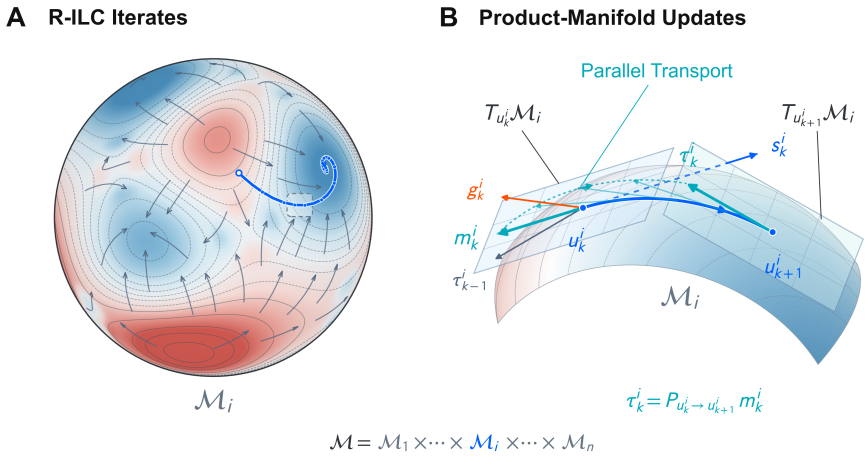

Figure 2: The schematic illustrates R-ILC updates on the product command manifold $\mathcal{M} = \prod_{i=1}^{n} \mathcal{M}_i$. (A) The blue curve shows successive iterates on a representative spherical factor $\mathcal{M}_i$, with the dashed rounded frame marking two consecutive iterates. (B) The local view shows the tangent step s_k^i, the update from u_k^i to u_{k+1}^i, and momentum transport between the two tangent spaces.

Algorithm 1 Riemannian Adaptive Moment ILC (R-ILC)

Input: Desired trajectory $\mathbf{y}^{\mathrm{des}}$, initial command $\mathbf{u}_1 \in \mathcal{M}$, trial budget $K \geq 1$, tracking loss $\mathcal{L}$, deterministic schedules $\{\eta_k, \beta_k, \theta_k\}_{k \geq 1}$, $\epsilon > 0$, and a stochastic Riemannian gradient estimator.
Output: Learned input sequence $\mathbf{u}^*$ and its corresponding observed output trajectory $\mathbf{y}^*$.

Initialize $m_0 = \tau_0 = 0 \in T_{\mathbf{u}_1}\mathcal{M}$ and $v_0 = 0 \in \mathbb{R}^n$.
for $k = 1, \ldots, K$ **do**
  Execute $\mathbf{u}_k$ and record the output trajectory $\mathbf{y}_k$.
  Compute $\mathbf{e}_k \leftarrow \mathbf{y}^{\mathrm{des}} - \mathbf{y}_k$.
  Obtain $g_k \in T_{\mathbf{u}_k}\mathcal{M}$ from the gradient estimator using $\mathbf{u}_k$ and the completed-trial data.
  $m_k \leftarrow \beta_k \tau_{k-1} + (1 - \beta_k) g_k$.
  $v_k^i \leftarrow \theta_k v_{k-1}^i + (1 - \theta_k) \|g_k^i\|_{u_k^i}^2$, $i = 1, \ldots, n$.
  $s_k \leftarrow -\eta_k D_{\mathbf{u}_k}(v_k) m_k$.
  $\mathbf{u}_{k+1} \leftarrow \mathrm{Exp}_{\mathbf{u}_k}(s_k)$.
  $\tau_k \leftarrow P_{\mathbf{u}_k \to \mathbf{u}_{k+1}} m_k$.
end for
return a selected command–output pair $(\mathbf{u}^*, \mathbf{y}^*)$ from the completed trials.

Algorithm 1 specifies the updates; Figure 2 illustrates their geometry. The distinct tangent spaces in Figure 2 explain the need to transport momentum before combining it with a new gradient. Scaling by one scalar per factor preserves tangency without selecting coordinates within a factor. The stabilizer inside the square root ensures smooth scaling near $v = 0$. The exponential map then converts the resulting tangent correction into an updated command in $\mathcal{M}$. Although the geometric operations act factorwise, each gradient block may depend on the full input sequence. R-ILC therefore optimizes the input sequence as a whole while preserving nominal output feasibility through (2.4).

3.2 Main Results

Gradient convergence and running-minimum rate

We first analyze the first-order convergence of the R-ILC updates for the expected tracking objective in (2.5). For example, differentiating individual trial losses under fresh sampling from the fixed disturbance law yields conditionally unbiased gradients when differentiation and expectation can be interchanged. Bounds on tracking errors, plant sensitivities, and their local variation within the operating region motivate the boundedness and smoothness conditions below.

We work on a filtered probability space $(\Omega, \mathscr{F}, (\mathcal{F}_k)_{k\geq 0}, \mathbb{P})$, where $\mathcal{F}_k$ records information before sampling g_k: the states $\mathbf{u}_k, \tau_{k-1}, v_{k-1}$ are $\mathcal{F}_k$-measurable, and $g_k \in T_{\mathbf{u}_k}\mathcal{M}$ is $\mathcal{F}_{k+1}$-measurable, with possibly dependent blocks. Set $\xi_k := g_k - \operatorname{grad} f(\mathbf{u}_k)$.

Assumption 1 (Smoothness and stochastic gradients). The function f is differentiable, each $f(\mathbf{u}_k)$ is integrable, and $f(\mathbf{u}_k) \geq f_* > -\infty$ almost surely for a deterministic f_*. There exist $L_0, L_1 \geq 0$ and deterministic $\rho \in (0, \infty]$ such that, on a common probability-one event, for every $k \geq 1$ and every y in the image of a minimizing normal domain of $\operatorname{Exp}_{\mathbf{u}_k}$ with $d(\mathbf{u}_k, y) \leq \rho$,

$$\|P_{y\to\mathbf{u}_k} \operatorname{grad} f(y) - \operatorname{grad} f(\mathbf{u}_k)\|_{\mathbf{u}_k} \leq (L_0 + L_1 \|\operatorname{grad} f(\mathbf{u}_k)\|_{\mathbf{u}_k})\, d(\mathbf{u}_k, y).$$

Here $P_{y\to\mathbf{u}_k}$ is reverse transport along the unique minimizing geodesic. For a deterministic $M < \infty$,

$$\mathbb{E}[g_k \mid \mathcal{F}_k] = \operatorname{grad} f(\mathbf{u}_k), \qquad \|g_k\|_{\mathbf{u}_k} \leq M \quad (k \geq 1), \quad \text{a.s.}$$

For every $k \geq 1$ almost surely, s_k lies in a minimizing normal domain of $\operatorname{Exp}_{\mathbf{u}_k}$. Set $L := L_0 + L_1 M$.

Theorem 1. Suppose Assumption 1 holds. Let $(\mathbf{u}_k)_{k\geq 1}$ be generated by Algorithm 1 with $\epsilon > 0$ and deterministic schedules satisfying $\eta_k > 0$, $\eta_{k+1} \leq \eta_k$, $0 \leq \beta_k \leq \beta < 1$, and $0 \leq \theta_k < 1$. If

$$\sum_k \eta_k = \infty, \qquad \sum_k \eta_k^2 < \infty, \qquad \sum_k \eta_k(1-\theta_k) < \infty,$$

then the following hold.

(i) The Riemannian gradient norms converge to zero almost surely and in mean square:

$$\|\operatorname{grad} f(\mathbf{u}_k)\|_{\mathbf{u}_k} \longrightarrow 0 \quad \text{a.s.}, \qquad \lim_{k\to\infty} \mathbb{E}\big[\|\operatorname{grad} f(\mathbf{u}_k)\|_{\mathbf{u}_k}^2\big] = 0.$$

(ii) The running minimum satisfies:

$$\min_{1\leq k\leq K} \|\operatorname{grad} f(\mathbf{u}_k)\|_{\mathbf{u}_k}^2 = o\left(\frac{1}{\sum_{k=1}^K \eta_k}\right) \qquad \text{a.s.}$$

Proof sketch. We first establish weighted summability. The descent estimate uses conditional unbiasedness with lagged scaling and bounds the change to the current scaling as an error. Bounded gradients give $O(\eta_k)$ steps, while local smoothness controls the change in gradient alignment when momentum is transported. The schedule conditions make the accumulated errors summable, yielding

$$\mathbb{E}\left[\sum_{k=1}^{\infty} \eta_k \|\operatorname{grad} f(\mathbf{u}_k)\|_{\mathbf{u}_k}^2\right] < \infty.$$

The weighted series is therefore finite almost surely (Appendix A.4).

The gradient norms converge to zero almost surely. Indeed, $\sum_k \eta_k = \infty$ and weighted summability give a zero lower limit, while their $O(\eta_k)$ variation rules out repeated excursions above a fixed positive level: such excursions would force the weighted series to diverge. Uniform boundedness then gives mean-square convergence by dominated convergence.

The same summability estimate yields the running-minimum rate via a sequence lemma (Liu & Yuan, 2022). Its ratio-divergence condition follows from the step-size assumptions (Appendix A.6).

STRICT-SADDLE NONCONVERGENCE

We further analyze whether R-ILC can converge to a strict saddle. Because each trial gradient also updates the adaptive scale, the local analysis must track its second moments. For a fixed estimator based only on the current command and fresh trial disturbances, these moments are functions of the command; smooth tracking sensitivities motivate a smooth local model. The following assumptions combine this model with noise excitation in unstable directions, under which Theorem 2 excludes convergence to the prescribed saddles.

Fix a strict saddle $\mathbf{u}^\dagger \in \mathcal{M}$, and write

$$\operatorname{grad} f(\mathbf{u}^\dagger) = 0, \qquad H^\dagger := \operatorname{Hess} f(\mathbf{u}^\dagger), \qquad \mathfrak{m} := n_-(H^\dagger) \geq 1. \tag{3.1}$$

The self-adjoint Hessian acts on $T_{\mathbf{u}^\dagger}\mathcal{M}$; n_- counts negative eigenvalues with multiplicity, and zero eigenvalues are allowed. Neighborhoods and constants in Assumptions 2–3 may depend on $\mathbf{u}^\dagger$.

To state the local conditions, consider a normal neighborhood U of $\mathbf{u}^\dagger$ and scalar functions $Q_i : U \to [0, M^2]$, $i = 1, \dots, n$. Set

$$Q(x) = (Q_1(x), \dots, Q_n(x)), \qquad v^\dagger := Q(\mathbf{u}^\dagger), \qquad D^\dagger := D_{\mathbf{u}^\dagger}(v^\dagger). \tag{3.2}$$

The noise condition accounts for adaptive scaling through the self-adjoint operator

$$K^\dagger := (D^\dagger)^{1/2} H^\dagger (D^\dagger)^{1/2}. \tag{3.3}$$

Let $\Pi_-^\dagger$ project orthogonally onto the negative spectral subspace $E_-^\dagger$ of $K^\dagger$. To express nearby noise vectors in the same tangent space, choose a smooth family of linear isometries $\mathcal{T}_x : T_x\mathcal{M} \to T_{\mathbf{u}^\dagger}\mathcal{M}$, $x \in U$, acting separately on each manifold factor and satisfying $\mathcal{T}_{\mathbf{u}^\dagger} = I$. Set $\widehat{\xi}_k := \mathcal{T}_{\mathbf{u}_k}\xi_k$ when $\mathbf{u}_k \in U$, and $\widehat{\xi}_k := 0$ otherwise.

Assumption 2 (Local regularity at strict saddles). The manifold $\mathcal{M}$ is finite-dimensional. There exist a neighborhood U and functions Q_i as above such that, for some $\nu \in [1/2, 1]$, U has compact closure, $f \in C^{2,\nu}(U)$, and each Q_i is locally $C^{1,\nu}$. Moreover, whenever $\mathbf{u}_k \in U$,

$$\mathbb{E}[\|g_k^i\|_{u_k^i}^2 \mid \mathcal{F}_k] = Q_i(\mathbf{u}_k). \tag{3.4}$$

Assumption 3 (Asymptotic noise excitation). With $E^\dagger := \{\mathbf{u}_k \to \mathbf{u}^\dagger\}$, assume

$$0 < \liminf_{k\to\infty} \mathbb{E}\left[\left\|\Pi_-^\dagger (D^\dagger)^{1/2}\widehat{\xi}_k\right\|^2 \middle| \mathcal{F}_k\right] \qquad \text{a.s. on } E^\dagger. \tag{3.5}$$

Only asymptotic excitation along trajectories approaching the prescribed saddle is required, and the conditional covariance may be singular. A sufficient condition is a positive uniform lower bound on the projected second moment whenever the iterates lie in a neighborhood of the saddle (Appendix D.3).

Theorem 2. Suppose the conditions of Theorem 1 hold, and let $\alpha_k := (1-\beta_k)\sum_{j=k}^\infty \eta_j \prod_{\ell=k+1}^j \beta_\ell$ denote the effective step induced by momentum, with empty products equal to one. Let $\mathcal{S}$ be a deterministic finite or countable set of strict saddles. For every $\mathbf{u}^\dagger \in \mathcal{S}$, suppose Assumptions 2–3 hold with exponent ν and there exists a constant $q > 0$ such that

$$\sum_k \left(\frac{1-\theta_k}{\alpha_k} - q\right)^2 < \infty, \qquad \sum_{k>t} \alpha_k^{1+\nu} = o\left(\sqrt{\sum_{k>t} \alpha_k^2}\right) \quad (t \to \infty). \tag{3.6}$$

Then

$$\mathbb{P}\left(\mathbf{u}_k \text{ converges and } \lim_{k\to\infty} \mathbf{u}_k \in \mathcal{S}\right) = 0. \tag{3.7}$$

The first condition in (3.6) places the second-moment update and the effective step on a common asymptotic time scale, with relative rate q. The second ensures that the tail of the higher-order errors is negligible relative to the accumulated noise scale.

Proof sketch. Fix $\mathbf{u}^\dagger \in \mathcal{S}$. In coordinates centered at this saddle, an $O(\eta_k)$ correction of the command iterate defines a shadow position that removes the leading momentum memory. On a tail remaining near the saddle, this position and the lagged second moment satisfy an exact stochastic approximation recursion with effective step $\alpha_k \asymp \eta_k$. The leading noise term uses the lagged scaling and is conditionally centered; the contribution of the difference between current and lagged scaling is retained exactly in the remainder. Geometric, scaling, and schedule errors together form a square-summable normalized remainder (Appendices C.1, C.5–C.6).

On the event $\mathbf{u}_k \to \mathbf{u}^\dagger$, this joint state converges almost surely to the equilibrium $(0, v^\dagger)$, with limiting scaling $D^\dagger$. This follows from $v_k \to v^\dagger = Q(\mathbf{u}^\dagger)$ and the vanishing shadow correction (Appendices C.3, C.5).

This equilibrium is unstable. The mean dynamics combine preconditioned gradient descent for the shadow position with relaxation of the second moment toward Q at relative rate q. Their linearization

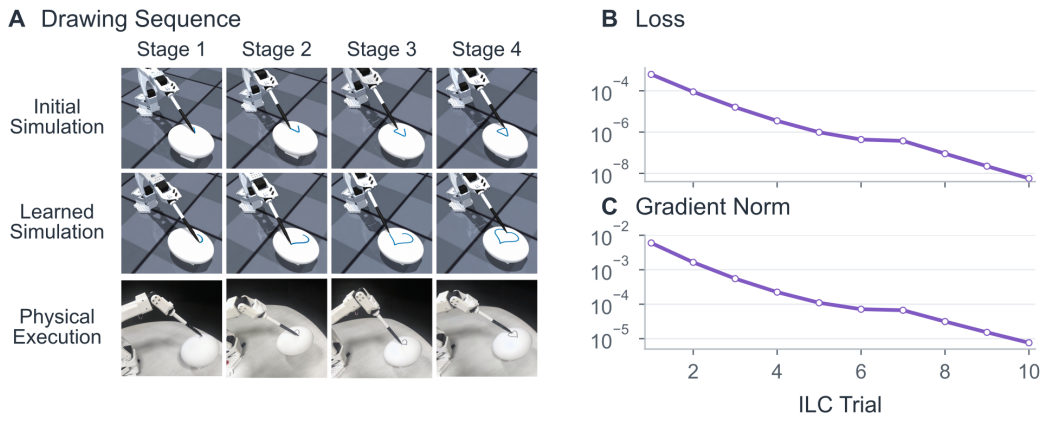

Figure 3: Representative drawing sequences and simulation learning curves illustrate the effectiveness of R-ILC. (A) shows initial and learned drawing sequences in simulation, together with physical execution of a learned trajectory. (B) shows the simulation loss L_k over ten ILC trials on a logarithmic scale. (C) shows the corresponding Riemannian gradient norm $\mathcal{G}_k$ on a logarithmic scale.

is block triangular, with diagonal blocks $-D^\dagger H^\dagger$ and $-qI$. The position block is similar to $-K^\dagger$ and has a positive eigenvalue because positive-definite preconditioning preserves Hessian inertia. The stable second-moment block cannot remove this instability (Appendices B.4–B.5).

On the candidate convergence event, Assumption 3 yields nonvanishing noise energy in the unstable subspace of the joint recursion (Appendix D.3). Local regularity, bounded noise moments, the remainder bound, and the tail condition in (3.6) permit a localized application of the nonconvergence theorem (Raimond & Tarrès, 2023); Appendix D gives the localization and verifies its hypotheses. Thus $\mathbb{P}(\mathbf{u}_k \to \mathbf{u}^\dagger) = 0$, and a countable union over $\mathcal{S}$ proves the result.

4 Experiments

4.1 Experimental Setup

We evaluate R-ILC on curved-surface robotic drawing in simulation and through physical execution. Physical execution uses a simulation-learned joint command sequence as a fixed reference on a convex workpiece (Appendix E). The target is a fixed closed heart-shaped contour on an ellipsoid defined by $\phi(y) = 0$, with $\phi(y) < 0$ inside. The full joint command sequence Q_k, represented by $\mathbf{u}_k$, has $N = 200$ nodes. In simulation, pen-tip positions are computed from the commanded joint configurations using forward kinematics in MuJoCo (Todorov et al., 2012).

In the main comparison, all methods use the same target and robot model and start from the same feasible joint trajectory. We run R-ILC, Euclidean Adam-ILC, and PD-ILC for ten trials, evaluating the full joint command sequence before each update. The two adaptive methods retain their optimizer states across trials. We also include a PID-based correction of the initial trajectory as a non-learning baseline. The baselines enforce joint limits without the proposed surface retraction. All configurations and schedules were fixed before comparison (Appendices E.3–E.4).

The objective combines metric-weighted tracking error with a penalty on differences between learned and desired trajectory increments (Appendix E.2). We monitor optimization using the loss $L_k = f_{\exp}(Q_k)$ and the norm $\mathcal{G}_k = \|g_k\|_{Q_k}$ of the evaluated Riemannian gradient. For comparison across all four methods, we use the same Cartesian tracking RMSE and maximum surface deviation:

$$E_k = \left(\frac{1}{N} \sum_{t=1}^{N} \|y_k(t) - y^{\mathrm{des}}(t)\|_2^2 \right)^{1/2}, \qquad P_k = \max_{1 \le t \le N} \frac{|\phi(y_k(t))|}{\|\nabla \phi(y_k(t))\|_2}.$$

Both metrics use raw simulated pen-tip positions and are expressed in meters; P_k approximates the largest absolute normal deviation from the surface at the sampled nodes (Appendix E.5).

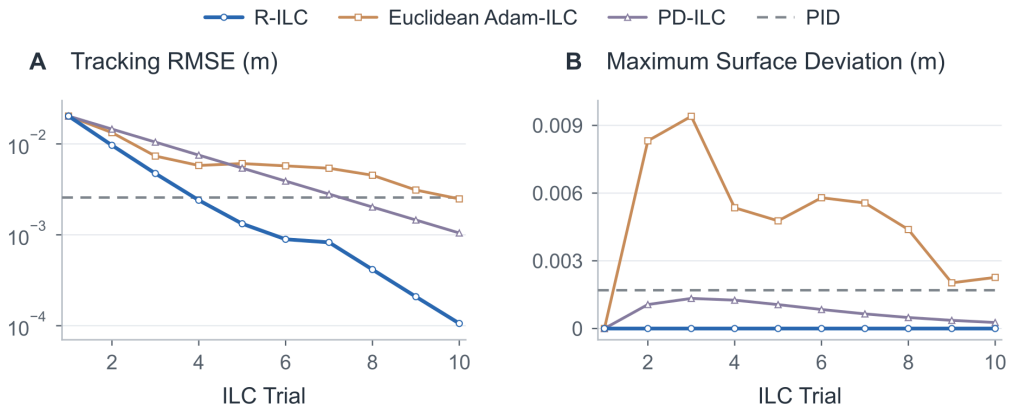

Figure 4: R-ILC is compared with Euclidean Adam-ILC, PD-ILC, and PID in simulation. (A) shows tracking RMSE E_k in meters on a logarithmic scale. (B) shows maximum surface deviation P_k in meters on a linear scale. Dashed lines show the PID baseline evaluated once.

4.2 Results and Analysis

We assess the learning behavior and comparative performance of R-ILC in simulation and demonstrate physical execution of a learned trajectory. Across the three tested initializations, we consistently find that R-ILC learns to track the target heart-shaped contour and achieves lower final tracking error and surface deviation than the baselines (Appendix E.6). We use the rounded triangular case to illustrate these findings in detail.

R-ILC reduces the trajectory loss by more than 99% by trial 4 and continues to refine the command in later trials (Figure 3(B)). The accompanying decrease in the Riemannian gradient norm links this improvement in the task objective to progress toward first-order stationarity (Figure 3(C)). In physical execution, a simulation-learned command, used as a fixed reference, produces the target heart-shaped contour on the curved workpiece (Figure 3(A)).

R-ILC achieves the lowest final tracking RMSE and reduces tracking error faster across ILC trials than the other ILC methods (Figure 4(A)). Its final RMSE is 1.06×10^{-4} m, 89.9% lower than that of PD-ILC, the strongest baseline in final RMSE. Tracking improvement alone, however, does not ensure surface feasibility: Euclidean Adam-ILC reduces RMSE while its surface deviation remains nonzero, and PD-ILC still produces off-surface trajectories as its tracking error decreases. The baseline corrections do not explicitly enforce the surface constraint, whereas R-ILC enforces it at each update and keeps surface deviation at the level of numerical residuals throughout learning (Figure 4(B)). Its lower tracking error is thus achieved while maintaining feasible intermediate trajectories.

5 Conclusion

R-ILC demonstrates that tracking improvement across trials is compatible with preserving nominal surface feasibility throughout learning. The theoretical analysis establishes convergence guarantees for R-ILC, while the curved-surface drawing experiments show lower tracking error and surface deviation than the baselines across the tested initializations. Together, these results support incorporating surface geometry directly into adaptive ILC, so that accumulated trial information refines the full input sequence while intermediate commands remain admissible.

Future work will explore how R-ILC can reuse experience across related tracking tasks. In particular, we will investigate meta-learning to construct feasible initial input sequences from previously learned tasks, followed by refinement through R-ILC. The aim is to reduce the number of learning trials required when the target contour or surface geometry changes. We will examine how task similarity and geometric variation affect the benefits of transferring prior learning experience while preserving nominal feasibility throughout adaptation.

AI USE STATEMENT

OpenAI Codex was used to assist with manuscript organization, English drafting and revision, mathematical derivations and proof checking, reference verification, and the preparation of supplementary materials. We have reviewed and verified all AI-assisted content and take full responsibility for the accuracy, reproducibility, and final content of this work.

REPRODUCIBILITY STATEMENT

The algorithm and assumptions are specified in Section 3, with complete proofs in Appendices A–D. Appendix E details the experimental setup, initialization procedures, numerical implementations, hyperparameter search and selection, and evaluation metrics. The anonymized supplementary material provides the simulation code, configurations, robot models, execution instructions, and demonstration videos.

REFERENCES

Suguru Arimoto, Masahiro Sekimoto, and Kenji Tahara. Iterative learning without reinforcement or reward for multijoint movements: A revisit of Bernstein's DOF problem on dexterity. *Journal of Robotics*, 2010:217867, 2010. doi: 10.1155/2010/217867. URL https://doi.org/10.1155/2010/217867.

Anas Barakat, Pascal Bianchi, Walid Hachem, and Sholom Schechtman. Stochastic optimization with momentum: Convergence, fluctuations, and traps avoidance. *Electronic Journal of Statistics*, 15(2):3892–3947, 2021. doi: 10.1214/21-EJS1880. URL https://doi.org/10.1214/21-EJS1880.

Gary Bécigneul and Octavian-Eugen Ganea. Riemannian adaptive optimization methods. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=r1eiqi09K7.

Nicolas Boumal. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, 2023. doi: 10.1017/9781009166164. URL https://www.nicolasboumal.net/book/.

Douglas A. Bristow, Marina Tharayil, and Andrew G. Alleyne. A survey of iterative learning control. *IEEE Control Systems Magazine*, 26(3):96–114, 2006. doi: 10.1109/MCS.2006.1636313. URL https://doi.org/10.1109/MCS.2006.1636313.

Yiyang Chen, Bing Chu, Christopher T. Freeman, and Yanhong Liu. Generalized iterative learning control with mixed system constraints: A gantry robot based verification. *Control Engineering Practice*, 95:104260, 2020. doi: 10.1016/j.conengprac.2019.104260. URL https://doi.org/10.1016/j.conengprac.2019.104260.

Bing Chu and David H. Owens. Iterative learning control for constrained linear systems. *International Journal of Control*, 83(7):1397–1413, 2010. doi: 10.1080/00207171003758752. URL https://doi.org/10.1080/00207171003758752.

Jacques Denavit and Richard S. Hartenberg. A kinematic notation for lower-pair mechanisms based on matrices. *Journal of Applied Mechanics*, 22(2):215–221, 1955. doi: 10.1115/1.4011045. URL https://doi.org/10.1115/1.4011045.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011. URL https://jmlr.org/papers/v12/duchi11a.html.

Rick Durrett. *Probability: Theory and Examples*. Cambridge University Press, 5th edition, 2019. doi: 10.1017/9781108591034. URL https://math.duke.edu/~rtd/PTE/PTE5_011119.pdf.

Morris W. Hirsch, Charles C. Pugh, and Michael Shub. *Invariant Manifolds*, volume 583 of *Lecture Notes in Mathematics*. Springer, 1977. doi: 10.1007/BFb0092042. URL https://doi.org/10.1007/BFb0092042.

Yusu Hong and Junhong Lin. On convergence of Adam for stochastic optimization under relaxed assumptions. In *Advances in Neural Information Processing Systems*, volume 37, pp. 10827–10877, 2024. doi: 10.52202/079017-0346. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/14bb27f680bee45d83bc769738e7f9b5-Abstract-Conference.html.

Ya-Ping Hsieh, Mohammad Reza Karimi Jaghargh, Andreas Krause, and Panayotis Mertikopoulos. Riemannian stochastic optimization methods avoid strict saddle points. In *Advances in Neural Information Processing Systems*, volume 36, pp. 29580–29601. Curran Associates, Inc., 2023. doi: 10.52202/075280-1286. URL https://papers.nips.cc/paper_files/paper/2023/hash/5e809ba53f34d9170386ebfc8b60300f-Abstract-Conference.html.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Haochuan Li, Alexander Rakhlin, and Ali Jadbabaie. Convergence of Adam under relaxed assumptions. In *Advances in Neural Information Processing Systems*, volume 36, pp. 52166–52196. Curran Associates, Inc., 2023. doi: 10.52202/075280-2272. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a3cc50126338b175e56bb3cad134db0b-Abstract-Conference.html.

Jun Liu and Ye Yuan. On almost sure convergence rates of stochastic gradient methods. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 2963–2983. PMLR, 2022. URL https://proceedings.mlr.press/v178/liu22d.html.

B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964. doi: 10.1016/0041-5553(64)90137-5. URL https://doi.org/10.1016/0041-5553(64)90137-5.

Olivier Raimond and Pierre Tarrès. Non-convergence to unstable equilibriums for continuous-time and discrete-time stochastic processes, 2023. URL https://arxiv.org/abs/2311.02978v1. arXiv:2311.02978v1.

Daeun Song and Young J. Kim. Distortion-free robotic surface-drawing using conformal mapping. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 627–633. IEEE, 2019. doi: 10.1109/ICRA.2019.8794034. URL https://doi.org/10.1109/ICRA.2019.8794034.

Fengjie Tian, Chong Lv, Zhenguo Li, and Guangbao Liu. Modeling and control of robotic automatic polishing for curved surfaces. *CIRP Journal of Manufacturing Science and Technology*, 14:55–64, 2016. doi: 10.1016/j.cirpj.2016.05.010. URL https://doi.org/10.1016/j.cirpj.2016.05.010.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033. IEEE, 2012. doi: 10.1109/IROS.2012.6386109. URL https://doi.org/10.1109/IROS.2012.6386109.

Johan Ubbink, Ruan Viljoen, Erwin Aertbeliën, Wilm Decré, and Joris De Schutter. Contactless surface following with acceleration limits: Enhancing robot manipulator performance through model predictive control. In *2024 European Control Conference (ECC)*, pp. 3594–3599. IEEE, 2024. doi: 10.23919/ECC64448.2024.10591035. URL https://www.paperhost.org/proceedings/controls/ECC24/files/0624.pdf.

Hosham Wahballa, Abubker Ahmed, Ghazally I. Y. Mustafa, Mohammednour Gibreel, and Lei Weining. Robotic contact on complex curved surfaces using adaptive trajectory planning through precise force control. *Machines*, 13(9):794, 2025. doi: 10.3390/machines13090794. URL https://doi.org/10.3390/machines13090794.

Charles W. Wampler. Manipulator inverse kinematic solutions based on vector formulations and damped least-squares methods. *IEEE Transactions on Systems, Man, and Cybernetics*, 16(1):93–101, 1986. doi: 10.1109/TSMC.1986.289285. URL https://doi.org/10.1109/TSMC.1986.289285.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2020. URL https://sls.csail.mit.edu/publications/2020/TianxingHe_ICLR-2020.pdf.

A PROOF OF THEOREM 1

We first collect the path notation and direct consequences used throughout the proof. For $t \in [0,1]$, set

$$\gamma_k(t) := \mathrm{Exp}_{\mathbf{u}_k}(ts_k), \qquad P_{k,t} : T_{\mathbf{u}_k}\mathcal{M} \to T_{\gamma_k(t)}\mathcal{M},$$

where $P_{k,t}$ is product parallel transport along γ_k. Assumption 1 makes this the unique minimizing update geodesic and gives $P_{k,1} = P_{\mathbf{u}_k \to \mathbf{u}_{k+1}}$, $\mathrm{Log}_{\mathbf{u}_k}(\gamma_k(t)) = ts_k$, and $d(\mathbf{u}_k, \gamma_k(t)) = t\|s_k\|_{\mathbf{u}_k}$; reverse transport is $P_{k,t}^{-1}$.

Write $m/\sqrt{v+\epsilon} := D_x(v)m$ for $m \in T_x\mathcal{M}$. To include $k = 1$ in the following identities, set $\mathbf{u}_0 = \mathbf{u}_1$ and $P_{\mathbf{u}_0 \to \mathbf{u}_1} = I$. Algorithm 1 gives $\tau_{k-1} = P_{\mathbf{u}_{k-1} \to \mathbf{u}_k} m_{k-1}$, so its momentum update also has the form $m_k = \beta_k P_{\mathbf{u}_{k-1} \to \mathbf{u}_k} m_{k-1} + (1-\beta_k) g_k$. The deterministic update maps and Assumption 1 make $m_k, v_k, s_k, \mathbf{u}_{k+1}, \tau_k$ $\mathcal{F}_{k+1}$-measurable and give $\mathbb{E}[\xi_k \mid \mathcal{F}_k] = 0$.

Assumption 1 and conditional Jensen's inequality give

$$\|\operatorname{grad} f(\mathbf{u}_k)\|_{\mathbf{u}_k} \le M \qquad \text{a.s.}$$

The gradient-dependent bound in Assumption 1 is related to Euclidean generalized smoothness (Zhang et al., 2020; Li et al., 2023); here it is imposed locally using parallel transport.

The transported-gradient condition in Assumption 1 is iterate-local and direction-independent: it holds at every point in the radius-ρ portion of a minimizing normal domain at $\mathbf{u}_k$, not only in the realized direction s_k. The step bound below places the update path in this region on a deterministic tail. A stronger sufficient condition is the same inequality for every pair of points at distance at most ρ in a strongly geodesically convex region containing all update paths. Taking $\rho = \infty$ removes the radius restriction within the relevant normal domains.

Using the constant L from Assumption 1, set

$$B_{\mathrm{step}} := \frac{M}{\sqrt{\epsilon}}.$$

Because $\sum_{k=1}^{\infty} \eta_k^2 < \infty$, one has $\eta_k \to 0$. Fix a deterministic index $K_\rho \ge 1$ such that

$$\eta_k B_{\mathrm{step}} \le \rho, \qquad k \ge K_\rho.$$

Such an index exists; when $\rho = \infty$, take $K_\rho = 1$. Lemma 3 below then implies

$$d(\mathbf{u}_k, \gamma_k(t)) = t\|s_k\|_{\mathbf{u}_k} \le t\eta_k B_{\mathrm{step}} \le \rho$$

for every $k \ge K_\rho$ and every $t \in [0,1]$ almost surely.

A.1 SUMMATION AND MOMENT BOUNDS

Lemma 1. Summation inequality. For any non-negative sequence (b_k) and any $\beta \in [0,1)$,

$$\sum_{k=1}^{K} \sum_{j=1}^{k} \beta^{k-j} b_j \le \frac{1}{1-\beta} \sum_{k=1}^{K} b_k.$$

Proof. Exchanging the summation order gives

$$\sum_{k=1}^{K}\sum_{j=1}^{k}\beta^{k-j}b_j = \sum_{j=1}^{K} b_j \sum_{k=j}^{K}\beta^{k-j}$$
$$\leq \sum_{j=1}^{K} b_j \sum_{\ell=0}^{\infty}\beta^{\ell} = \frac{1}{1-\beta}\sum_{j=1}^{K} b_j. \qquad \square$$

Lemma 2. Boundedness of moments. Under the conditions of Theorem 1, for every $k \geq 0$,

$$\|m_k\|_{\mathbf{u}_k} \leq M, \qquad \|\tau_k\|_{\mathbf{u}_{k+1}} \leq M, \qquad 0 \leq v_k^i \leq M^2,$$

and

$$\left\| \frac{m_k}{\sqrt{v_k+\epsilon}} \right\|_{\mathbf{u}_k} \leq \frac{M}{\sqrt{\epsilon}} \qquad \text{a.s.}$$

Proof. The claims hold at $k = 0$. Suppose $\|\tau_{k-1}\|_{\mathbf{u}_k} \leq M$. Since parallel transport is an isometry and $\|g_k\|_{\mathbf{u}_k} \leq M$,

$$\|m_k\|_{\mathbf{u}_k} \leq \beta_k\|\tau_{k-1}\|_{\mathbf{u}_k} + (1-\beta_k)\|g_k\|_{\mathbf{u}_k} \leq M.$$

Parallel transport then gives $\|\tau_k\|_{\mathbf{u}_{k+1}} = \|m_k\|_{\mathbf{u}_k} \leq M$. For each block,

$$v_k^i = \theta_k v_{k-1}^i + (1-\theta_k)\|g_k^i\|_{u_k^i}^2 \leq M^2$$

by induction. Since $v_k^i + \epsilon \geq \epsilon$,

$$\left\| \frac{m_k}{\sqrt{v_k+\epsilon}} \right\|_{\mathbf{u}_k}^2 = \sum_i \frac{\|m_k^i\|_{u_k^i}^2}{v_k^i+\epsilon}$$
$$\leq \frac{\|m_k\|_{\mathbf{u}_k}^2}{\epsilon} \leq \frac{M^2}{\epsilon}. \qquad \square$$

A.2 Step and denominator bounds

Lemma 3. Step bound. For every $k \geq 1$ and every $t \in [0,1]$, Algorithm 1 satisfies

$$\|s_k\|_{\mathbf{u}_k} \leq \eta_k B_{\text{step}}, \qquad d(\mathbf{u}_k, \gamma_k(t)) = t\|s_k\|_{\mathbf{u}_k} \leq t\eta_k B_{\text{step}}, \qquad \text{a.s.}$$

In particular,

$$d(\mathbf{u}_k, \mathbf{u}_{k+1}) = \|s_k\|_{\mathbf{u}_k} \leq \eta_k B_{\text{step}} \qquad \text{a.s.}$$

Proof. By Lemma 2 and the conditions of Theorem 1,

$$\|s_k\|_{\mathbf{u}_k} \leq \eta_k \left\| \frac{m_k}{\sqrt{v_k+\epsilon}} \right\|_{\mathbf{u}_k}$$
$$\leq \eta_k \left(\frac{M}{\sqrt{\epsilon}} \right) = \eta_k B_{\text{step}}.$$

The curve γ_k has constant speed $\|s_k\|_{\mathbf{u}_k}$. By Assumption 1, γ_k is the unique minimizing geodesic from $\mathbf{u}_k$ to $\mathbf{u}_{k+1}$, and every subsegment from $\mathbf{u}_k$ to $\gamma_k(t)$ is minimizing. Therefore,

$$d(\mathbf{u}_k, \gamma_k(t)) = \text{Length}(\gamma_k|_{[0,t]}) = t\|s_k\|_{\mathbf{u}_k} \leq t\eta_k B_{\text{step}}.$$

Setting $t = 1$ proves the endpoint identity and bound. $\square$

Lemma 4. Gradient norm lower bound. Under the conditions of Theorem 1, for every $k \geq 1$,

$$\sum_{i=1}^{n} \frac{\|(\operatorname{grad} f(\mathbf{u}_k))^i\|_{u_k^i}^2}{\sqrt{v_{k-1}^i + \epsilon}} \geq \frac{1}{\sqrt{M^2 + \epsilon}} \|\operatorname{grad} f(\mathbf{u}_k)\|_{\mathbf{u}_k}^2.$$

Proof. Lemma 2 gives $v_{k-1}^i \leq M^2$, and hence

$$\frac{1}{\sqrt{v_{k-1}^i + \epsilon}} \geq \frac{1}{\sqrt{M^2 + \epsilon}}.$$

Summing over the blocks proves the claim. □

Lemma 5. Adaptive denominator variation. Under the conditions of Theorem 1, for every $k \geq 1$,

$$\left\| \frac{m_k}{\sqrt{v_{k-1} + \epsilon}} - \frac{m_k}{\sqrt{v_k + \epsilon}} \right\|_{\mathbf{u}_k} \leq \frac{M^3}{\epsilon^{3/2}} (1 - \theta_k) \qquad \text{a.s.}$$

Proof. For $a, b \geq 0$,

$$\left| \frac{1}{\sqrt{a + \epsilon}} - \frac{1}{\sqrt{b + \epsilon}} \right| \leq \frac{|a - b|}{2\epsilon^{3/2}}.$$

Moreover,

$$\begin{aligned} |v_k^i - v_{k-1}^i| &= (1 - \theta_k) \left| \|g_k^i\|_{u_k^i}^2 - v_{k-1}^i \right| \\ &\leq 2M^2 (1 - \theta_k). \end{aligned}$$

Thus

$$\left| \frac{1}{\sqrt{v_{k-1}^i + \epsilon}} - \frac{1}{\sqrt{v_k^i + \epsilon}} \right| \leq \frac{M^2}{\epsilon^{3/2}} (1 - \theta_k).$$

Therefore

$$\begin{aligned} \left\| \frac{m_k}{\sqrt{v_{k-1} + \epsilon}} - \frac{m_k}{\sqrt{v_k + \epsilon}} \right\|_{\mathbf{u}_k}^2 &\leq \frac{M^4}{\epsilon^3} (1 - \theta_k)^2 \sum_i \|m_k^i\|_{u_k^i}^2 \\ &\leq \frac{M^6}{\epsilon^3} (1 - \theta_k)^2. \end{aligned}$$

Taking square roots proves the lemma. □

A.3 Descent estimates

Lemma 6. Descent inequality. Under the conditions of Theorem 1, for every $k \geq K_\rho$, Algorithm 1 satisfies

$$\begin{aligned} f(\mathbf{u}_{k+1}) \leq f(\mathbf{u}_k) - \eta_k \left\langle \operatorname{grad} f(\mathbf{u}_k), \frac{m_k}{\sqrt{v_k + \epsilon}} \right\rangle_{\mathbf{u}_k} \\ + \frac{L}{2} B_{\text{step}}^2 \eta_k^2. \end{aligned}$$

Proof. Along γ_k, one has $\dot{\gamma}_k(t) = P_{k,t}s_k$. Hence the chain rule and integration along the actual update path give

$$\begin{aligned} f(\mathbf{u}_{k+1}) - f(\mathbf{u}_k) &= \int_0^1 \langle \operatorname{grad} f(\gamma_k(t)), P_{k,t}s_k \rangle_{\gamma_k(t)}\, dt \\ &= \int_0^1 \left\langle P_{k,t}^{-1} \operatorname{grad} f(\gamma_k(t)), s_k \right\rangle_{\mathbf{u}_k} dt. \end{aligned}$$

Assumption 1, Lemma 3, and the definition of K_ρ imply

$$\left\| P_{k,t}^{-1} \operatorname{grad} f(\gamma_k(t)) - \operatorname{grad} f(\mathbf{u}_k) \right\|_{\mathbf{u}_k} \le Ld(\mathbf{u}_k, \gamma_k(t)) = Lt\|s_k\|_{\mathbf{u}_k}.$$

Consequently,

$$\begin{aligned} f(\mathbf{u}_{k+1}) &\le f(\mathbf{u}_k) + \left\langle \operatorname{grad} f(\mathbf{u}_k), \operatorname{Log}_{\mathbf{u}_k}(\mathbf{u}_{k+1}) \right\rangle_{\mathbf{u}_k} + \frac{L}{2} d(\mathbf{u}_k, \mathbf{u}_{k+1})^2 \\ &= f(\mathbf{u}_k) + \langle \operatorname{grad} f(\mathbf{u}_k), s_k \rangle_{\mathbf{u}_k} + \frac{L}{2}\|s_k\|_{\mathbf{u}_k}^2. \end{aligned}$$

Substitution of the update direction s_k yields

$$\begin{aligned} f(\mathbf{u}_{k+1}) &\le f(\mathbf{u}_k) - \eta_k \left\langle \operatorname{grad} f(\mathbf{u}_k), \frac{m_k}{\sqrt{v_k + \epsilon}} \right\rangle_{\mathbf{u}_k} \\ &\quad + \frac{L}{2}\|s_k\|_{\mathbf{u}_k}^2 \\ &\le f(\mathbf{u}_k) - \eta_k \left\langle \operatorname{grad} f(\mathbf{u}_k), \frac{m_k}{\sqrt{v_k + \epsilon}} \right\rangle_{\mathbf{u}_k} + \frac{L}{2} B_{\text{step}}^2 \eta_k^2, \end{aligned}$$

where the last step uses Lemma 3. Assumption 1 justifies $\operatorname{Log}_{\mathbf{u}_k}(\mathbf{u}_{k+1}) = s_k$ for the unique minimizing update geodesic. No globally defined logarithm map is required. □

Lemma 7. Adaptive inner-product lower bound. Under the conditions of Theorem 1, let

$$\Xi_k^{\text{align}} := \mathbb{E}\left[\langle \operatorname{grad} f(\mathbf{u}_k), D_{\mathbf{u}_k}(v_k) m_k \rangle_{\mathbf{u}_k} \right].$$

Then, for every $k \ge K_\rho + 1$,

$$\Xi_k^{\text{align}} \ge \left(\prod_{t=K_\rho+1}^{k} \beta_t \right) \Xi_{K_\rho}^{\text{align}} + \sum_{j=K_\rho+1}^{k} \left(\prod_{t=j+1}^{k} \beta_t \right) \Delta_j^{\text{align}},$$

where

$$\Delta_j^{\text{align}} := -\beta_j \frac{LM}{\sqrt{\epsilon}} B_{\text{step}} \eta_{j-1} + \frac{1-\beta_j}{\sqrt{M^2+\epsilon}} \mathbb{E}\| \operatorname{grad} f(\mathbf{u}_j)\|_{\mathbf{u}_j}^2 - \frac{M^4}{\epsilon^{3/2}}(1-\theta_j).$$

Moreover,

$$|\Xi_{K_\rho}^{\text{align}}| \le \frac{M^2}{\sqrt{\epsilon}}.$$

Proof. Since each v_{k-1}^i is a scalar and parallel transport acts within the corresponding tangent block,

$$D_{\mathbf{u}_k}(v_{k-1})P_{\mathbf{u}_{k-1}\to\mathbf{u}_k} = P_{\mathbf{u}_{k-1}\to\mathbf{u}_k}D_{\mathbf{u}_{k-1}}(v_{k-1}).$$

Split

$$\begin{aligned}\Xi_k^{\text{align}} &= \mathbb{E}[\langle \operatorname{grad} f(\mathbf{u}_k), D_{\mathbf{u}_k}(v_{k-1})m_k\rangle_{\mathbf{u}_k}] \\ &\quad + \mathbb{E}\Big[\big\langle \operatorname{grad} f(\mathbf{u}_k), \big(D_{\mathbf{u}_k}(v_k) - D_{\mathbf{u}_k}(v_{k-1})\big)m_k\big\rangle_{\mathbf{u}_k}\Big]\end{aligned}$$

By Lemma 5 and $\|\operatorname{grad} f(\mathbf{u}_k)\|_{\mathbf{u}_k} \le M$,

$$\mathbb{E}\Big[\big\langle \operatorname{grad} f(\mathbf{u}_k), \big(D_{\mathbf{u}_k}(v_k) - D_{\mathbf{u}_k}(v_{k-1})\big)m_k\big\rangle_{\mathbf{u}_k}\Big] \ge -\frac{M^4}{\epsilon^{3/2}}(1-\theta_k).$$

For the first term, use

$$m_k = \beta_k P_{\mathbf{u}_{k-1}\to\mathbf{u}_k} m_{k-1} + (1-\beta_k)g_k.$$

Because $\mathbf{u}_k$ and v_{k-1} are $\mathcal{F}_k$-measurable, the stochastic-gradient term satisfies

$$\begin{aligned}&\mathbb{E}[\,\langle \operatorname{grad} f(\mathbf{u}_k), D_{\mathbf{u}_k}(v_{k-1})g_k\rangle_{\mathbf{u}_k}\,|\,\mathcal{F}_k] \\ &\quad = \sum_i \frac{\|(\operatorname{grad} f(\mathbf{u}_k))^i\|_{u_k^i}^2}{\sqrt{v_{k-1}^i + \epsilon}} \\ &\quad \ge \frac{1}{\sqrt{M^2+\epsilon}}\|\operatorname{grad} f(\mathbf{u}_k)\|_{\mathbf{u}_k}^2,\end{aligned}$$

where the final step uses Lemma 4.

For $k \ge K_\rho + 1$, the preceding update index satisfies $k-1 \ge K_\rho$. Hence Lemma 3 and the definition of K_ρ give $d(\mathbf{u}_{k-1}, \mathbf{u}_k) \le \rho$, so Assumption 1 applies at the endpoint of the whole preceding unique minimizing update geodesic. Since $P_{\mathbf{u}_k\to\mathbf{u}_{k-1}} = P_{k-1,1}^{-1}$ along that geodesic, isometry of parallel transport gives

$$\begin{aligned}&\mathbb{E}\Big[\big\langle \operatorname{grad} f(\mathbf{u}_k), D_{\mathbf{u}_k}(v_{k-1})P_{\mathbf{u}_{k-1}\to\mathbf{u}_k}m_{k-1}\big\rangle_{\mathbf{u}_k}\Big] \\ &= \mathbb{E}\left[\left\langle P_{k-1,1}^{-1}\operatorname{grad} f(\mathbf{u}_k), D_{\mathbf{u}_{k-1}}(v_{k-1})m_{k-1}\right\rangle_{\mathbf{u}_{k-1}}\right] \\ &\ge \Xi_{k-1}^{\text{align}} - L\mathbb{E}\big[d(\mathbf{u}_{k-1},\mathbf{u}_k)\|D_{\mathbf{u}_{k-1}}(v_{k-1})m_{k-1}\|_{\mathbf{u}_{k-1}}\big] \\ &\ge \Xi_{k-1}^{\text{align}} - \frac{LM}{\sqrt{\epsilon}}B_{\text{step}}\eta_{k-1}.\end{aligned}$$

The last step uses Lemmas 2 and 3. Combining the preceding estimates yields, for every $k \ge K_\rho + 1$,

$$\Xi_k^{\text{align}} \ge \beta_k \Xi_{k-1}^{\text{align}} + \Delta_k^{\text{align}}.$$

Finally, Assumption 1 and Lemma 2 imply

$$\begin{aligned}|\Xi_{K_\rho}^{\text{align}}| &\le \mathbb{E}\Big[\|\operatorname{grad} f(\mathbf{u}_{K_\rho})\|_{\mathbf{u}_{K_\rho}}\|D_{\mathbf{u}_{K_\rho}}(v_{K_\rho})m_{K_\rho}\|_{\mathbf{u}_{K_\rho}}\Big] \\ &\le M\frac{M}{\sqrt{\epsilon}} = \frac{M^2}{\sqrt{\epsilon}}.\end{aligned}$$

Unrolling the recursion from $\Xi_{K_\rho}^{\text{align}}$ proves the lemma. □

A.4 WEIGHTED SUMMABILITY AND EXPECTED MINIMUM-GRADIENT BOUND

Proof. For $k \geq K_\rho + 1$, Lemma 6, followed by expectation and Lemma 7, gives

$$\mathbb{E}[f(\mathbf{u}_{k+1})] \leq \mathbb{E}[f(\mathbf{u}_k)] - \eta_k \left(\prod_{t=K_\rho+1}^{k} \beta_t \right) \Xi_{K_\rho}^{\text{align}} - \eta_k \sum_{j=K_\rho+1}^{k} \left(\prod_{t=j+1}^{k} \beta_t \right) \Delta_j^{\text{align}} + \frac{L}{2} B_{\text{step}}^2 \eta_k^2.$$

Use $|\Xi_{K_\rho}^{\text{align}}| \leq M^2/\sqrt{\epsilon}$, substitute Δ_j^{align}, retain only the $j = k$ gradient term in the descent sum, and use $0 \leq \beta_k \leq \beta < 1$. Then

$$\frac{1-\beta}{\sqrt{M^2+\epsilon}} \eta_k \mathbb{E}\| \operatorname{grad} f(\mathbf{u}_k)\|_{\mathbf{u}_k}^2 \leq \mathbb{E}[f(\mathbf{u}_k)] - \mathbb{E}[f(\mathbf{u}_{k+1})] + \frac{M^2}{\sqrt{\epsilon}} \eta_k \beta^{k-K_\rho} + \frac{L}{2} B_{\text{step}}^2 \eta_k^2 + \frac{LM}{\sqrt{\epsilon}} B_{\text{step}} \eta_k \sum_{j=K_\rho+1}^{k} \beta^{k-j+1} \eta_{j-1} + \frac{M^4}{\epsilon^{3/2}} \eta_k \sum_{j=K_\rho+1}^{k} \beta^{k-j}(1-\theta_j).$$

Summing over $k = K_\rho + 1, \ldots, K$ and using $f(\mathbf{u}_{K+1}) \geq f_*$ yields

$$\frac{1-\beta}{\sqrt{M^2+\epsilon}} \sum_{k=K_\rho+1}^{K} \eta_k \mathbb{E}\| \operatorname{grad} f(\mathbf{u}_k)\|_{\mathbf{u}_k}^2 \leq \mathbb{E}[f(\mathbf{u}_{K_\rho+1})] - f_* + \frac{M^2}{\sqrt{\epsilon}} \sum_{k=K_\rho+1}^{K} \eta_k \beta^{k-K_\rho} + \frac{L}{2} B_{\text{step}}^2 \sum_{k=K_\rho+1}^{K} \eta_k^2 + \frac{LM}{\sqrt{\epsilon}} B_{\text{step}} \sum_{k=K_\rho+1}^{K} \eta_k \sum_{j=K_\rho+1}^{k} \beta^{k-j+1} \eta_{j-1} + \frac{M^4}{\epsilon^{3/2}} \sum_{k=K_\rho+1}^{K} \eta_k \sum_{j=K_\rho+1}^{k} \beta^{k-j}(1-\theta_j).$$

Monotonicity of (η_k) gives

$$\sum_{k=K_\rho+1}^{K} \eta_k \beta^{k-K_\rho} \leq \frac{\beta}{1-\beta} \eta_{K_\rho+1},$$

and

$$\sum_{k=K_\rho+1}^{K} \eta_k \sum_{j=K_\rho+1}^{k} \beta^{k-j+1} \eta_{j-1} \leq \frac{\beta}{1-\beta} \sum_{k=K_\rho}^{K-1} \eta_k^2 \leq \frac{\beta}{1-\beta} \sum_{k=K_\rho}^{K} \eta_k^2.$$

Similarly, Lemma 1 and monotonicity of (η_k) give

$$\sum_{k=K_\rho+1}^{K} \eta_k \sum_{j=K_\rho+1}^{k} \beta^{k-j}(1-\theta_j) \le \frac{1}{1-\beta} \sum_{k=K_\rho+1}^{K} \eta_k(1-\theta_k).$$

Define

$$C_{\mathrm{start}}^{(K_\rho)} := \frac{\sqrt{M^2+\epsilon}}{1-\beta}\left[\mathbb{E}[f(\mathbf{u}_{K_\rho+1})] - f_* + \frac{\beta\eta_{K_\rho+1}}{1-\beta}\frac{M^2}{\sqrt{\epsilon}}\right], \tag{A.1}$$

$$C_\theta := \frac{\sqrt{M^2+\epsilon}}{(1-\beta)^2}\frac{M^4}{\epsilon^{3/2}}, \tag{A.2}$$

$$C_{\eta^2} := \frac{\sqrt{M^2+\epsilon}}{1-\beta}\frac{L}{2}B_{\mathrm{step}}^2 + \frac{\beta\sqrt{M^2+\epsilon}}{(1-\beta)^2}\frac{LM}{\sqrt{\epsilon}}B_{\mathrm{step}}. \tag{A.3}$$

Then, for every $K \ge K_\rho + 1$,

$$\begin{aligned}\sum_{k=K_\rho+1}^{K} \eta_k \mathbb{E}\|\operatorname{grad} f(\mathbf{u}_k)\|_{\mathbf{u}_k}^2 \le C_{\mathrm{start}}^{(K_\rho)} + C_\theta \sum_{k=K_\rho+1}^{K} \eta_k(1-\theta_k) \\ + C_{\eta^2}\sum_{k=K_\rho}^{K} \eta_k^2.\end{aligned} \tag{A.4}$$

Since

$$\min_{K_\rho+1\le k\le K} \mathbb{E}\|\operatorname{grad} f(\mathbf{u}_k)\|_{\mathbf{u}_k}^2 \sum_{k=K_\rho+1}^{K} \eta_k \le \sum_{k=K_\rho+1}^{K} \eta_k \mathbb{E}\|\operatorname{grad} f(\mathbf{u}_k)\|_{\mathbf{u}_k}^2,$$

(A.4) gives

$$\begin{aligned}&\min_{K_\rho+1\le k\le K} \mathbb{E}\|\operatorname{grad} f(\mathbf{u}_k)\|_{\mathbf{u}_k}^2 \\ &\quad\le \frac{C_{\mathrm{start}}^{(K_\rho)} + C_\theta \sum_{k=K_\rho+1}^{K} \eta_k(1-\theta_k) + C_{\eta^2}\sum_{k=K_\rho}^{K} \eta_k^2}{\sum_{k=K_\rho+1}^{K} \eta_k}.\end{aligned}$$

The numerator is uniformly bounded by the parameter conditions of Theorem 1, whereas the tail denominator diverges. Therefore,

$$\lim_{K\to\infty} \min_{K_\rho+1\le k\le K} \mathbb{E}\|\operatorname{grad} f(\mathbf{u}_k)\|_{\mathbf{u}_k}^2 = 0.$$

Because the minimum over $1 \le k \le K$ is no larger than the minimum over the tail $K_\rho+1 \le k \le K$, this also establishes the expected minimum-gradient bound over the full prefix. The stronger full-sequence conclusion and the running-minimum rate in Theorem 1 are proved in Appendix A.6. The equivalent blockwise expression uses the product-gradient identity

$$\|\operatorname{grad} f(\mathbf{u}_k)\|_{\mathbf{u}_k}^2 = \sum_{i=1}^{n} \|(\operatorname{grad} f(\mathbf{u}_k))^i\|_{u_k^i}^2. \qquad \square$$

A.5 SEQUENCE LEMMA FOR THE ALMOST-SURE RATE

Lemma 8. Sequence lemma (Liu & Yuan, 2022, Lemma 2). Let $(X_k)_{k\geq 1}$ be non-negative, and let $(a_k)_{k\geq 1}$ be decreasing and positive. If

$$\sum_{k=1}^{\infty} a_k X_k < \infty, \qquad \sum_{k=2}^{\infty} \frac{a_k}{\sum_{j=1}^{k-1} a_j} = \infty,$$

then

$$\min_{1\leq k\leq K} X_k = o\left(\frac{1}{\sum_{k=1}^{K} a_k}\right).$$

Remark (indexing convention). Liu and Yuan's published Lemma 2 is written with a denominator ending at the preceding index. Writing $S_\ell^\eta := \sum_{k=1}^{\ell} \eta_k$ and applying that statement at $t = K$ gives $\min_{1\leq k\leq K} X_k = o((S_{K-1}^\eta)^{-1})$. Since $S_K^\eta / S_{K-1}^\eta = 1 + \eta_K / S_{K-1}^\eta \to 1$, this is equivalent to $\min_{1\leq k\leq K} X_k = o((S_K^\eta)^{-1})$, the almost-sure rate formulation used in Theorem 1.

A.6 FULL-SEQUENCE CONVERGENCE AND ALMOST-SURE RATE

Proof. From (A.4) and the summability conditions in Theorem 1,

$$\sum_{k=K_\rho+1}^{\infty} \eta_k \mathbb{E}\| \operatorname{grad} f(\mathbf{u}_k)\|_{\mathbf{u}_k}^2 < \infty.$$

By the monotone convergence theorem,

$$\mathbb{E}\left[\sum_{k=K_\rho+1}^{\infty} \eta_k \| \operatorname{grad} f(\mathbf{u}_k)\|_{\mathbf{u}_k}^2\right] = \sum_{k=K_\rho+1}^{\infty} \eta_k \mathbb{E}\| \operatorname{grad} f(\mathbf{u}_k)\|_{\mathbf{u}_k}^2 < \infty.$$

Hence

$$\sum_{k=K_\rho+1}^{\infty} \eta_k \| \operatorname{grad} f(\mathbf{u}_k)\|_{\mathbf{u}_k}^2 < \infty \qquad \text{a.s.}$$

The omitted prefix is finite because Assumption 1 gives

$$\sum_{k=1}^{K_\rho} \eta_k \| \operatorname{grad} f(\mathbf{u}_k)\|_{\mathbf{u}_k}^2 \leq M^2 \sum_{k=1}^{K_\rho} \eta_k < \infty \qquad \text{a.s.}$$

Consequently,

$$\sum_{k=1}^{\infty} \eta_k \| \operatorname{grad} f(\mathbf{u}_k)\|_{\mathbf{u}_k}^2 < \infty \qquad \text{a.s.}$$

We next prove full-sequence gradient convergence. Set $b_k := \| \operatorname{grad} f(\mathbf{u}_k)\|_{\mathbf{u}_k}$ and work on the probability-one event on which the preceding summability and bounds hold. Then $b_k \leq M$. For $k \geq K_\rho$, Assumption 1, isometric parallel transport, and Lemma 3 give

$$\begin{aligned} |b_{k+1} - b_k| &\leq \left\| P_{\mathbf{u}_{k+1}\to\mathbf{u}_k} \operatorname{grad} f(\mathbf{u}_{k+1}) - \operatorname{grad} f(\mathbf{u}_k)\right\|_{\mathbf{u}_k} \\ &\leq L d(\mathbf{u}_k, \mathbf{u}_{k+1}) \leq C\eta_k, \qquad C := LB_{\text{step}}. \end{aligned}$$

Define $\sigma_{K_\rho} = 0$ and $\sigma_{k+1} - \sigma_k = \eta_k$, and linearly interpolate $b(\sigma_k) = b_k$. Since $\sum_k \eta_k = \infty$, this defines a nonnegative C-Lipschitz function on $[0, \infty)$. On each interval, $b(t) \le b_k + C\eta_k$, so

$$\int_0^\infty b(t)^2\, dt \le 2 \sum_{k \ge K_\rho} \eta_k b_k^2 + 2C^2 \sum_{k \ge K_\rho} \eta_k^3 < \infty.$$

The last series is finite because the steps are bounded and square-summable. A nonnegative Lipschitz function with finite squared integral tends to zero: otherwise uniform continuity supplies infinitely many disjoint intervals of fixed positive length on which it stays above a fixed positive level, contradicting integrability. Thus $b_k \to 0$ almost surely. Since $b_k^2 \le M^2$, dominated convergence also gives $\mathbb{E}[b_k^2] \to 0$.

Finally, the parameter conditions of Theorem 1 make (η_k) positive and non-increasing. It remains to verify the ratio-divergence condition in Lemma 8. Let

$$S_k^\eta := \sum_{j=1}^{k} \eta_j.$$

Theorem 1 requires $S_K^\eta \to \infty$. Moreover, for every $k \ge 2$,

$$\log S_k^\eta - \log S_{k-1}^\eta = \log\left(1 + \frac{\eta_k}{S_{k-1}^\eta}\right) \le \frac{\eta_k}{S_{k-1}^\eta}.$$

Therefore,

$$\sum_{k=2}^{K} \frac{\eta_k}{\sum_{j=1}^{k-1} \eta_j} \ge \log S_K^\eta - \log S_1^\eta \longrightarrow \infty.$$

Thus the ratio-divergence condition in Lemma 8 follows automatically from the parameter conditions of Theorem 1.

Apply Lemma 8 pathwise with

$$a_k = \eta_k, \qquad X_k = \|\operatorname{grad} f(\mathbf{u}_k)\|_{\mathbf{u}_k}^2.$$

This proves

$$\min_{1 \le k \le K} \|\operatorname{grad} f(\mathbf{u}_k)\|_{\mathbf{u}_k}^2 = o\left(\frac{1}{\sum_{k=1}^{K} \eta_k}\right) \qquad \text{a.s.} \qquad \square$$

B Local Geometry and Spectral Analysis

Lemma 9 provides the fixed-coordinate geometry used in the shadow reduction. Lemma 10 and Proposition 2 then establish instability of the reduced ODE required by Theorem 2. The full ODE comparison and center-stable geometry explain this instability; the stochastic nonconvergence argument is completed separately in Appendix D.

For each factor, the exponential map is defined on tangent vectors whose geodesics exist on $[0, 1]$, Log denotes the unique normal inverse on the selected minimizing normal neighborhood, and P denotes parallel transport along the specified geodesic. Their product versions are componentwise. These conventions require neither geodesic completeness nor a globally single-valued logarithm.

B.1 Local blockwise trivialization

Set $d := \dim \mathcal{M}$. Fix a product orthonormal basis $(e_a)_{a=1}^d$ of $T_{\mathbf{u}^\dagger}\mathcal{M}$ and use it, when Euclidean coordinates are required, to identify $T_{\mathbf{u}^\dagger}\mathcal{M}$ with $\mathbb{R}^d$.

Lemma 9. Local blockwise trivialization and bounded coordinate geometry. After shrinking U to a product normal neighborhood if necessary, choose open product neighborhoods

$$\mathbf{u}^\dagger \in U_{\mathrm{sh}} \Subset U_{\mathrm{geo}} \Subset U,$$

where U_{geo} is a product of strongly convex normal neighborhoods. On U there exist, for each factor, smooth local linear isometries

$$\mathcal{T}^i_{x^i} : T_{x^i}\mathcal{M}_i \longrightarrow T_{u^{\dagger,i}}\mathcal{M}_i, \qquad \mathcal{T}^i_{u^{\dagger,i}} = I,$$

and use the blockwise trivialization

$$\mathcal{T}_x := \bigoplus_{i=1}^n \mathcal{T}^i_{x^i} : T_x\mathcal{M} \longrightarrow T_{\mathbf{u}^\dagger}\mathcal{M}, \qquad \mathcal{T}_{\mathbf{u}^\dagger} = I. \tag{B.1}$$

In these coordinates, the representation of the Adam preconditioner is independent of the base point:

$$\begin{aligned} \mathcal{T}_x D_x(v)\mathcal{T}_x^{-1} &= \bigoplus_{i=1}^n \mathcal{T}^i_{x^i}\big((v^i+\epsilon)^{-1/2} I_{T_{x^i}\mathcal{M}_i}\big)(\mathcal{T}^i_{x^i})^{-1} \\ &= \bigoplus_{i=1}^n (v^i+\epsilon)^{-1/2} I_{T_{u^{\dagger,i}}\mathcal{M}_i} =: D(v). \end{aligned} \tag{B.2}$$

Thus every fixed-coordinate formula below uses the same block-diagonal operator $D(v)$ on $T_{\mathbf{u}^\dagger}\mathcal{M}$. Along any specified product geodesic, and in particular along each algorithmic update geodesic, product parallel transport has the fixed-coordinate representation

$$R_{x\to x'} := \mathcal{T}_{x'} P_{x\to x'} \mathcal{T}_x^{-1},$$

which is block diagonal. Hence, because $D(v)$ is scalar on each block,

$$R_{x\to x'} D(v) = D(v) R_{x\to x'}. \tag{B.3}$$

For $z = \mathrm{Log}_{\mathbf{u}^\dagger}(x)$, define

$$\mathcal{A}(z) := D(\mathrm{Log}_{\mathbf{u}^\dagger})_x \circ \mathcal{T}_x^{-1}. \tag{B.4}$$

Put $x(z) := \mathrm{Exp}_{\mathbf{u}^\dagger}(z)$. For a fixed-coordinate step $u \in T_{\mathbf{u}^\dagger}\mathcal{M}$, define the one-step map

$$\Phi(z,u) := \mathrm{Log}_{\mathbf{u}^\dagger}\Big(\mathrm{Exp}_{x(z)}(\mathcal{T}^{-1}_{x(z)}u)\Big). \tag{B.5}$$

If $x^+(z,u) := \mathrm{Exp}_{x(z)}(\mathcal{T}^{-1}_{x(z)}u)$, define the fixed-coordinate parallel-transport map along that update geodesic by

$$\mathscr{R}(z,u) := \mathcal{T}_{x^+(z,u)} P_{x(z)\to x^+(z,u)} \mathcal{T}^{-1}_{x(z)}. \tag{B.6}$$

For every $V \Subset U_{\mathrm{geo}}$, there exists $r_V > 0$ such that, whenever $x \in \overline{V}$ and $a \in T_x\mathcal{M}$ satisfies $\|a\|_x \le r_V$, the radial geodesic $t \mapsto \mathrm{Exp}_x(ta)$ is defined on $[0,1]$ and remains in U_{geo}. It follows that the maps in (B.5)–(B.6) are defined on an open neighborhood of the compact parameter set

$$\mathcal{K}_V := \big\{(z,u) : \mathrm{Exp}_{\mathbf{u}^\dagger}(z) \in \overline{V},\ \|u\| \le r_V\big\},$$

and all of their derivatives through order two are uniformly bounded on $\mathcal{K}_V$. The maps $\mathcal{A}$ and $\mathcal{A}^{-1}$, together with their derivatives through order two, are uniformly bounded on $\mathrm{Log}_{\mathbf{u}^\dagger}(\overline{U_{\mathrm{geo}}})$, and $\mathcal{A}(0) = I$.

Proof. In each factor tangent bundle over U, choose a smooth local orthonormal frame whose value at $u^{\dagger,i}$ agrees with the corresponding block of the fixed product basis chosen above. This gives the blockwise isometries in (B.1). Since $D_x(v)$ is a scalar multiple of the identity on each factor, direct conjugation gives (B.2), and the same blockwise scalar structure gives the commutation relation (B.3).

For $V \Subset U_{\mathrm{geo}}$, the zero section over $\overline{V}$ is compact and lies in the open domain of the geodesic flow. Smooth dependence of geodesics on their initial data and compactness therefore give a uniform $r_V > 0$ for which the indicated radial geodesics exist and remain in U_{geo}. The exponential map, the logarithm chart, the local trivialization, and parallel transport along these radial geodesics depend smoothly on their arguments. Hence Φ and $\mathscr{R}$ are C^2 on an open neighborhood of $\mathcal{K}_V$, and compactness gives the asserted uniform derivative bounds. Finally, the normal chart makes $D(\mathrm{Log}_{\mathbf{u}^\dagger})_x$ invertible on U_{geo}; compactness gives the uniform bounds for $\mathcal{A}$ and $\mathcal{A}^{-1}$, while $D(\mathrm{Log}_{\mathbf{u}^\dagger})_{\mathbf{u}^\dagger} = I$ gives $\mathcal{A}(0) = I$. Thus this is a local coordinate construction, not an additional global curvature assumption. □

B.2 Full Adam ODE family

Define the block squared-gradient map

$$G_i^{\mathrm{sq}}(x) := \|(\mathrm{grad}\, f(x))^i\|_{x^i}^2, \qquad G^{\mathrm{sq}}(x) := (G_1^{\mathrm{sq}}(x), \dots, G_n^{\mathrm{sq}}(x)). \tag{B.7}$$

Let $S : U \to [0,\infty)^n$, $S = (S_1, \dots, S_n)$, be any locally $C^{1,\nu}$ nonnegative second-moment forcing; call such a map admissible. For $\delta > 0$, define the S-forced member of the full Adam ODE family by

$$\mathsf{A}_{\delta,S}: \qquad \begin{aligned} \dot{x} &= -D_x(v)m, \\ \delta\nabla_{\dot{x}}m &= \mathrm{grad}\, f(x) - m, \\ \dot{v} &= q\big(S(x) - v\big). \end{aligned} \tag{B.8}$$

At a critical point $\mathbf{u}^\dagger$, the associated equilibrium is

$$X_{\delta,S}^\dagger = \big(\mathbf{u}^\dagger, 0, S(\mathbf{u}^\dagger)\big). \tag{B.9}$$

Two choices of S distinguish two members of this ODE family:

$$\mathsf{A}_{\delta,G^{\mathrm{sq}}} \quad \text{and} \quad \mathsf{A}_{\delta,Q}. \tag{B.10}$$

The first is the gradient-square-forced full Adam ODE and has $G^{\mathrm{sq}}(\mathbf{u}^\dagger) = 0$. The second is the Q-forced full Adam ODE; it uses the deterministic conditional second-moment field Q and has the generally nonzero equilibrium $Q(\mathbf{u}^\dagger)$. Thus the forcing and equilibrium change, while the resulting object remains a deterministic ODE.

The full Adam ODE family is auxiliary in the proof of Theorem 2. It is included because it proves a useful continuous statement: the strict-saddle spectral mechanism does not depend on whether the second-moment forcing is G^{sq}, Q, or another admissible S.

B.3 Reduced shadow ODE and relation to the full ODE

For $z = \mathrm{Log}_{\mathbf{u}^\dagger}(x)$ and $x = \mathrm{Exp}_{\mathbf{u}^\dagger}(z)$, use the frame from Lemma 9 to define

$$\mathcal{G}(z) := \mathcal{T}_x\, \mathrm{grad}\, f(x), \qquad \widetilde{Q}(z) := Q(x).$$

With $\mathcal{A}(z)$ and $D(v)$ from (B.4) and (B.2), the Reduced Shadow Mean-Field ODE is

$$\dot{Z} = h(Z), \qquad Z = (z, v), \qquad h(z, v) := \begin{pmatrix} -\mathcal{A}(z)D(v)\mathcal{G}(z) \\ q(\widetilde{Q}(z) - v) \end{pmatrix}.$$

Its equilibrium is $Z^\dagger = (0, v^\dagger)$, where $v^\dagger = Q(\mathbf{u}^\dagger)$. The second moment need not vanish at a critical point because Q includes the noise contribution (Appendix C.3). The field is regular on an open neighborhood of the equilibrium and has an unstable linearization there (Appendix B.5).

On the shadow clock, the momentum-relaxation scale $\delta_k = \alpha_k/(1 - \beta_k)$ tends to zero (Appendix C.1). Formally imposing $m = \operatorname{grad} f(x)$ in the Q-forced full Adam ODE gives the reduced ODE above. The continuous comparison is

$$\mathsf{A}_{\delta, G^{\mathrm{sq}}} \xrightarrow{G^{\mathrm{sq}} \mapsto Q} \mathsf{A}_{\delta, Q} \xrightarrow{\delta \downarrow 0, \text{ spectral level}} \dot{Z} = h(Z). \tag{B.11}$$

The first arrow changes the forcing and equilibrium. The second identifies the slow spectral limit after removal of the fast stable momentum mode. This comparison is spectral; no trajectory-level convergence is asserted or used.

B.4 Preconditioned Hessian inertia

The spectral argument is first proved for an arbitrary positive-definite Adam preconditioner. It then applies to every member of the full Adam ODE family and to the Reduced Shadow Mean-Field ODE.

For a real square matrix A, let $E^u(A)$ denote its generalized real spectral subspace associated with eigenvalues having positive real part.

With $x = \operatorname{Exp}_{\mathbf{u}^\dagger}(z)$ in the normal chart, the fixed-coordinate gradient from Appendix B.3 satisfies

$$\mathcal{G}(z) = H^\dagger z + O(\|z\|^{1+\nu}). \tag{B.12}$$

Indeed, $D\mathcal{G}(0) = H^\dagger$: the derivative of the trivialization multiplies $\operatorname{grad} f(\mathbf{u}^\dagger) = 0$, while $D \operatorname{Exp}_{\mathbf{u}^\dagger}(0) = I$. The Hölder remainder follows from Assumption 2 and the smooth trivialization furnished by Lemma 9.

Let $D_\star$ be a linear self-adjoint preconditioner satisfying

$$D_\star : T_{\mathbf{u}^\dagger}\mathcal{M} \to T_{\mathbf{u}^\dagger}\mathcal{M}, \qquad D_\star \succ 0. \tag{B.13}$$

Lemma 10. Positive-definite preconditioning preserves Hessian inertia. The operator $D_\star H^\dagger$ is diagonalizable with real spectrum, and

$$n_+(-D_\star H^\dagger) = n_-(H^\dagger) = \mathfrak{m}. \tag{B.14}$$

Proof. The operator $D_\star H^\dagger$ is similar to

$$K_\star := D_\star^{1/2} H^\dagger D_\star^{1/2}, \tag{B.15}$$

because

$$D_\star^{-1/2}(D_\star H^\dagger) D_\star^{1/2} = K_\star. \tag{B.16}$$

The operator $K_\star$ is self-adjoint. Sylvester's law of inertia gives

$$n_-(K_\star) = n_-(H^\dagger) = \mathfrak{m}. \tag{B.17}$$

Similarity transfers the real eigenvalues and their algebraic multiplicities to $D_\star H^\dagger$. Negation exchanges positive and negative eigenvalues, proving (B.14). □

Unless $D_\star$ commutes with $H^\dagger$, a Hessian eigenvector need not remain an eigenvector after preconditioning. The invariant statement is the equality of spectral counts in (B.14), not a canonical direction-by-direction identification in the original Hessian eigenbasis.

B.5 Instability of the full and reduced ODEs

Proposition 1. The full Adam ODE family has Morse-index instability. Let S be any admissible forcing in $\mathsf{A}_{\delta,S}$, set

$$S^\dagger := S(\mathbf{u}^\dagger), \qquad D_S^\dagger := D(S^\dagger), \tag{B.18}$$

and fix $\delta > 0$. In the state order $(z, m, v - S^\dagger)$, the Jacobian at $X_{\delta,S}^\dagger$ is

$$\mathcal{J}_{\delta,S}^\dagger = \begin{pmatrix} 0 & -D_S^\dagger & 0 \\ \delta^{-1}H^\dagger & -\delta^{-1}I & 0 \\ qDS(\mathbf{u}^\dagger) & 0 & -qI \end{pmatrix}. \tag{B.19}$$

It has exactly $\mathfrak{m}$ eigenvalues with positive real part. In particular,

$$\dim E^u(\mathcal{J}_{\delta,G^{\mathrm{sq}}}^\dagger) = \dim E^u(\mathcal{J}_{\delta,Q}^\dagger) = \mathfrak{m}. \tag{B.20}$$

Proof. The lower v block is triangular and contributes only the stable eigenvalue $-q$. By Lemma 10, the eigenvalues κ of $D_S^\dagger H^\dagger$ are real, and exactly $\mathfrak{m}$ of them are negative. Each mode of the (z, m) block satisfies

$$\delta r^2 + r + \kappa = 0. \tag{B.21}$$

If $\kappa < 0$, the roots have opposite signs and

$$r_+(\delta, \kappa) = \frac{-1 + \sqrt{1 - 4\delta\kappa}}{2\delta} > 0. \tag{B.22}$$

If $\kappa = 0$, the roots are 0 and $-\delta^{-1}$. If $\kappa > 0$, both roots have negative real part: they are either negative real numbers or a complex-conjugate pair with real part $-1/(2\delta)$. Thus precisely the $\mathfrak{m}$ negative modes of $D_S^\dagger H^\dagger$ generate positive roots. The lower-left coupling $qDS(\mathbf{u}^\dagger)$ changes eigenvectors but not eigenvalues. □

Proposition 1 explains why replacing the noise-free forcing G^{sq} by the true conditional second moment Q cannot remove strict-saddle instability. The preconditioner and the eigenvectors may change, but the unstable dimension does not.

To define the reduced field on an open neighborhood of its equilibrium, extend each scaling factor $(v^i + \epsilon)^{-1/2}$ to $v^i > -\epsilon$. Since $v^\dagger = Q(\mathbf{u}^\dagger) \in [0, M^2]^n$, this domain contains $v^\dagger$, even if one of its components is zero. The chart differential and blockwise isometries are smooth by Lemma 9, while the local regularity and conditional-moment conditions in Assumption 2 give the required regularity of $\mathcal{G}$ and $\widetilde{Q}$. Consequently, the field h defined in Appendix B.3 belongs to $C^{1,\nu}$ on an open neighborhood of $Z^\dagger$.

Proposition 2. The Reduced Shadow Mean-Field ODE has the same unstable dimension. For every negative eigenvalue $\kappa < 0$ of $D^\dagger H^\dagger$, the slow positive root in (B.22) satisfies

$$r_+(\delta, \kappa) \longrightarrow -\kappa > 0 \qquad (\delta \downarrow 0). \tag{B.23}$$

The other root is $-\delta^{-1} + O(1)$ and represents the fast stable momentum mode removed from the Reduced Shadow Mean-Field ODE.

The Jacobian of the vector field h of the Reduced Shadow Mean-Field ODE at $Z^\dagger$ is

$$J^\dagger := Dh(Z^\dagger) = \begin{pmatrix} -D^\dagger H^\dagger & 0 \\ qD\widetilde{Q}(0) & -qI \end{pmatrix}. \tag{B.24}$$

Write $E^u := E^u(J^\dagger)$. Then

$$\operatorname{spec}(J^\dagger) = \operatorname{spec}(-D^\dagger H^\dagger) \cup \{-q\}, \qquad \dim E^u = \mathfrak{m} \geq 1. \tag{B.25}$$

Proof. In fixed coordinates, the upper component is $-\mathcal{A}(z)D(v)\mathcal{G}(z)$. Its derivative with respect to v vanishes at $z = 0$ because $\mathcal{G}(0) = \operatorname{grad} f(\mathbf{u}^\dagger) = 0$. The Jacobian is therefore block triangular. Its lower diagonal block $-qI$ is strictly stable, while Lemma 10 gives exactly $\mathfrak{m}$ positive real eigenvalues in the upper block. Equation (B.23) gives the corresponding spectral consistency with the Q-forced full Adam ODE. □

The coupling $qD\widetilde{Q}(0)$ can change the augmented eigenvectors but cannot remove or create positive eigenvalues. The actual unstable subspace used in the stochastic argument is the subspace E^u of (B.24), not an unstable subspace copied from the gradient-square-forced full Adam ODE.

B.6 Local center-stable geometry

Proposition 3. Center-stable geometry of the Reduced Shadow Mean-Field ODE. Let φ_t be the local flow generated by h. There are neighborhoods $\mathcal{V}_{\text{cs}} \Subset \mathcal{U}_{\text{cs}}$ of $Z^\dagger$ and a local C^1 center-stable manifold

$$W^{cs}_{\text{loc}}(Z^\dagger) \subset \mathcal{U}_{\text{cs}} \tag{B.26}$$

such that:

1. its tangent space at the equilibrium is the generalized spectral subspace E^{cs} associated with eigenvalues having nonpositive real part;
2. every orbit starting in $\mathcal{V}_{\text{cs}}$ and remaining in $\mathcal{V}_{\text{cs}}$ for all positive time lies in $W^{cs}_{\text{loc}}(Z^\dagger)$;
3. its codimension is

$$\operatorname{codim} W^{cs}_{\text{loc}}(Z^\dagger) = \dim E^u = \mathfrak{m} \geq 1. \tag{B.27}$$

Consequently this manifold has zero Lebesgue measure in local augmented coordinates and zero measure for every smooth positive volume density there.

Proof. The local field h is $C^{1,\nu}$ and hence C^1. Apply the standard local center-stable manifold theorem (Hirsch et al., 1977) to the local flow, or equivalently to a restricted local time-one map, using

$$T_{Z^\dagger}\mathbb{R}^{d+n} = E^u \oplus E^{cs}. \tag{B.28}$$

The theorem gives the local invariant manifold, its tangent space, and the capture property. Proposition 2 gives its codimension. A C^1 embedded submanifold of positive codimension has zero ambient Lebesgue measure, and smooth coordinate changes preserve null sets. □

Proposition 3 provides an auxiliary deterministic interpretation of the unstable spectrum; the stochastic nonconvergence conclusion is obtained from the discrete recursion through Theorem D.1.

C Exact Shadow Reduction

C.1 Shadow-coefficient identities

Recall the effective step from Theorem 2:

$$\alpha_k := (1-\beta_k)\sum_{j=k}^{\infty}\eta_j\prod_{\ell=k+1}^{j}\beta_\ell,$$

where an empty product equals one. This coefficient collects each gradient's influence through subsequent momentum updates; the parameter conditions of Theorem 1 make the series finite. Define $c_k := \beta_k\alpha_k/(1-\beta_k)$. The corresponding momentum-tail series and the auxiliary gain notation are

$$c_k = \sum_{j=k}^{\infty}\eta_j\prod_{\ell=k}^{j}\beta_\ell, \qquad \omega_k := 1-\theta_k,$$

$$q_k := \frac{\omega_k}{\alpha_k}, \qquad \mathfrak{a}(t) := \left(\sum_{k>t}\alpha_k^2\right)^{1/2}.$$

The formula for c_k follows by multiplying the series for α_k by $\beta_k/(1-\beta_k)$, including when $\beta_k = 0$. Separating the initial term also gives $\alpha_k = (1-\beta_k)(\eta_k + c_{k+1})$. Thus these definitions preserve the exact momentum cancellation. Monotonicity of η_k gives

$$0 \le c_k \le \eta_k\sum_{\ell=1}^{\infty}\beta^\ell = \frac{\beta}{1-\beta}\eta_k \longrightarrow 0. \tag{C.1}$$

Separating the $j=k$ term in the series for c_k gives the exact backward identity

$$c_k = \beta_k\eta_k + \beta_k\sum_{j=k+1}^{\infty}\eta_j\prod_{\ell=k+1}^{j}\beta_\ell = \beta_k(\eta_k + c_{k+1}). \tag{C.2}$$

Together with the definition of α_k, this yields

$$(1-\beta)\eta_k \le \alpha_k \le \frac{\eta_k}{1-\beta}. \tag{C.3}$$

Consequently,

$$\sum_k \alpha_k = \infty, \qquad \sum_k \alpha_k^2 < \infty. \tag{C.4}$$

No convergence of β_k is required. The dimensionless momentum-relaxation scale is

$$\delta_k := \frac{\alpha_k}{1-\beta_k} = \eta_k + c_{k+1}, \tag{C.5}$$

and (C.1) gives

$$0 < \delta_k \le \frac{\eta_k}{1-\beta} \longrightarrow 0. \tag{C.6}$$

This scaling is a deterministic coefficient identity; the exact algorithmic reduction is proved in Lemma 12.

By the first condition in (3.6),

$$\sum_k |q_k - q|^2 < \infty.$$

Hence $q_k \to q > 0$ and $\omega_k = q_k \alpha_k = O(\eta_k)$. Together with (C.3) and the parameter conditions of Theorem 1, this gives

$$\sum_k \omega_k = \infty, \qquad \sum_k \omega_k^2 < \infty, \qquad \sum_k \eta_k \omega_k < \infty. \tag{C.7}$$

C.2 Schedule examples and tail estimates

For fixed momentum $\beta_k = \beta_0 \in [0, 1)$ and $\omega_k = q\eta_k$, both step-size schedules below satisfy $0 \le 1 - \eta_{k+r}/\eta_k \le Cr/k$ for all integers $r \ge 0$ and sufficiently large k, with a deterministic constant C. Averaging this bound with the geometric momentum weights gives

$$0 \le 1 - \frac{\alpha_k}{\eta_k} = (1 - \beta_0) \sum_{r=0}^{\infty} \beta_0^r \left(1 - \frac{\eta_{k+r}}{\eta_k}\right) = O(k^{-1}).$$

Thus $q_k - q = q(\eta_k/\alpha_k - 1) = O(k^{-1})$, which verifies the first condition in (3.6). A sufficiently large offset ensures $\theta_k \in [0, 1)$. For any momentum sequence allowed by Theorem 1, the choice $\theta_k = 1 - q\alpha_k$ on a tail gives $q_k = q$ exactly; a finite prefix can be adjusted to keep $\theta_k \in [0, 1)$.

If $\nu > 1/2$, then for any $\eta_0 > 0$ and fixed offset $k_{\mathrm{off}} \ge 0$, a sufficient example for the tail condition in (3.6) is

$$\eta_k = \eta_0 (k + k_{\mathrm{off}})^{-p}, \qquad p \in (1/2, 1], \qquad p\nu > 1/2. \tag{C.8}$$

At the endpoint $\nu = 1/2$, no polynomial schedule in (C.8) can satisfy the strict inequality while also having $p \le 1$. The endpoint is nevertheless admissible. For any $\eta_0 > 0$, $k_{\mathrm{off}} > e$, and $r \in (0, 1]$, take

$$\eta_k := \frac{\eta_0}{(k + k_{\mathrm{off}})[\log(k + k_{\mathrm{off}})]^r}. \tag{C.9}$$

Then $\sum_k \eta_k = \infty$, $\sum_k \eta_k^2 < \infty$, and, because $\alpha_k \asymp \eta_k$,

$$\sum_{k>t} \alpha_k^{3/2} = O\Big(t^{-1/2} (\log t)^{-3r/2}\Big), \qquad \mathfrak{a}(t) \asymp t^{-1/2} (\log t)^{-r}. \tag{C.10}$$

Their ratio is $O((\log t)^{-r/2}) = o(1)$, which verifies the tail condition in (3.6) at $\nu = 1/2$.

C.3 Conditional convergence of the second moment

The first discrete step identifies the only possible limiting preconditioner. This is necessary because the saddle equilibrium of the stochastic algorithm is not known before the v recursion is analyzed.

Whenever $\mathbf{u}_k \in U$, conditional unbiasedness removes the cross term:

$$\begin{aligned} Q_i(\mathbf{u}_k) &= \|(\operatorname{grad} f(\mathbf{u}_k))^i\|_{u_k^i}^2 + \mathbb{E}[\|\xi_k^i\|_{u_k^i}^2 \mid \mathcal{F}_k], \\ \mathbb{E}[v_k^i - v_{k-1}^i \mid \mathcal{F}_k] &= \omega_k \big(Q_i(\mathbf{u}_k) - v_{k-1}^i\big). \end{aligned}$$

The first identity explains why the equilibrium need not be zero when the mean gradient vanishes. The second identifies the stable averaging recursion whose conditional convergence is proved next.

Lemma 11. Conditional convergence of the second moment. Under Assumptions 1–2 and the parameter conditions of Theorems 1–2,

$$\{\mathbf{u}_k \to \mathbf{u}^\dagger\} \subseteq \{v_k \to v^\dagger\} \qquad \text{a.s.} \tag{C.11}$$

Proof. Since Q is only defined on U, its martingale residual must first be localized. Choose a normal neighborhood U_Q such that $\mathbf{u}^\dagger \in U_Q \Subset U$ and define the predictable indicator

$$I_k^Q := \mathbf{1}_{\{\mathbf{u}_k \in U_Q\}}.$$

Define the block residual piecewise, without evaluating Q outside U, by

$$\zeta_k^i := \begin{cases} \|g_k^i\|_{u_k^i}^2 - Q_i(\mathbf{u}_k), & I_k^Q = 1, \\ 0, & I_k^Q = 0. \end{cases} \tag{C.12}$$

Because I_k^Q is $\mathcal{F}_k$-measurable, (3.4) and the bounded-gradient clause of Assumption 1 give

$$\mathbb{E}[\zeta_k \mid \mathcal{F}_k] = 0, \qquad \|\zeta_k\| \le M^2\sqrt{n} \quad \text{a.s. for every } k. \tag{C.13}$$

The parameter conditions of Theorem 1 and (C.3) imply $\sum_k \alpha_k = \infty$ and $\sum_k \alpha_k^2 < \infty$. The first condition in (3.6) gives $q_k \to q > 0$ and $\omega_k = q_k\alpha_k$, hence, as recorded in (C.7),

$$\sum_k \omega_k = \infty, \qquad \sum_k \omega_k^2 < \infty, \tag{C.14}$$

and $\sum_k \omega_k\zeta_k$ is a globally defined martingale series whose conditional quadratic variation is bounded by a deterministic multiple of $\sum_k \omega_k^2$. It therefore converges almost surely (Durrett, 2019, Theorem 4.5.2).

Let

$$E_{k_0}^Q := \{\mathbf{u}_k \to \mathbf{u}^\dagger\} \cap \{\mathbf{u}_k \in U_Q \text{ for every } k \ge k_0\}. \tag{C.15}$$

Since U_Q is a neighborhood of $\mathbf{u}^\dagger$,

$$\{\mathbf{u}_k \to \mathbf{u}^\dagger\} = \bigcup_{k_0=1}^{\infty} E_{k_0}^Q. \tag{C.16}$$

Fix a deterministic k_0. On $E_{k_0}^Q$ and for every $k \ge k_0$, the localized residual in (C.12) equals the actual second-moment residual, so the second-moment recursion is

$$v_k - v^\dagger = (1-\omega_k)(v_{k-1} - v^\dagger) + \omega_k\big(Q(\mathbf{u}_k) - Q(\mathbf{u}^\dagger)\big) + \omega_k\zeta_k. \tag{C.17}$$

On $E_{k_0}^Q$, continuity of Q gives

$$Q(\mathbf{u}_k) - Q(\mathbf{u}^\dagger) \longrightarrow 0. \tag{C.18}$$

For completeness, the stable-recursion step is now proved directly. Choose a deterministic K so large that $0 < \omega_k < 1$ for $k \ge K$, set $K_{k_0} := \max\{K, k_0\}$, and put

$$\pi_{j,k}^\omega := \prod_{\ell=j}^{k}(1-\omega_\ell), \qquad \pi_{k+1,k}^\omega := 1.$$

Unrolling (C.17) gives

$$v_k - v^\dagger = \pi^{\omega}_{K_{k_0},k}\big(v_{K_{k_0}-1} - v^\dagger\big) + \sum_{j=K_{k_0}}^{k} \pi^{\omega}_{j+1,k}\omega_j\big(Q(\mathbf{u}_j) - Q(\mathbf{u}^\dagger)\big) + \sum_{j=K_{k_0}}^{k} \pi^{\omega}_{j+1,k}\omega_j\zeta_j \qquad \text{on } E^Q_{k_0}. \tag{C.19}$$

First, $\pi^{\omega}_{K_{k_0},k} \le \exp(-\sum_{\ell=K_{k_0}}^{k}\omega_\ell) \to 0$. Second, the deterministic weights are nonnegative and satisfy

$$\sum_{j=K_{k_0}}^{k} \pi^{\omega}_{j+1,k}\omega_j = 1 - \pi^{\omega}_{K_{k_0},k} \le 1. \tag{C.20}$$

For every fixed j, $\pi^{\omega}_{j+1,k} \to 0$; together with $Q(\mathbf{u}_j) - Q(\mathbf{u}^\dagger) \to 0$, (C.20) is the Toeplitz argument showing that the middle term in (C.19) tends to zero.

For the last term, set

$$\mathsf{M}^{\zeta}_{\ell} := \sum_{j=K}^{\ell} \omega_j\zeta_j.$$

By the almost-sure convergence established above, write $\mathsf{M}^{\zeta}_{\ell} \to \mathsf{M}^{\zeta}_{\infty}$ almost surely. Summation by parts, followed by splitting at any deterministic $J > K_{k_0}$, shows that the finite prefix vanishes and that the remaining tail has

$$\limsup_{k\to\infty} \left\| \sum_{j=J}^{k} \pi^{\omega}_{j+1,k}(\mathsf{M}^{\zeta}_{j} - \mathsf{M}^{\zeta}_{j-1}) \right\| \le 3 \sup_{\ell \ge J-1} \|\mathsf{M}^{\zeta}_{\ell} - \mathsf{M}^{\zeta}_{\infty}\|. \tag{C.21}$$

The right side tends to zero as $J \to \infty$. Hence all three terms in (C.19) vanish on $E^Q_{k_0}$. This proves $v_k - v^\dagger \to 0$ on every $E^Q_{k_0}$; taking the countable union in (C.16) proves (C.11). □

The neighborhood U_Q and the globally defined residual ζ_k in (C.12) are retained below. The shadow localization will choose its inner neighborhood inside U_Q, so every later use of ζ_k in an exact algorithmic identity coincides with the actual second-moment residual.

In particular,

$$D(v_k) \longrightarrow D^\dagger \quad \text{on } \{\mathbf{u}_k \to \mathbf{u}^\dagger\}. \tag{C.22}$$

C.4 Fixed-preconditioner Euclidean shadow cancellation

This section performs the algorithm-specific reduction. The argument does not replace β_k by a continuous relaxation rate and does not require β_k to converge.

First freeze the geometry and preconditioner at $D^\dagger$ and write

$$\mathbf{u}_{k+1} = \mathbf{u}_k - \eta_k D^\dagger m_k, \qquad m_k = \beta_k m_{k-1} + (1-\beta_k)g_k. \tag{C.23}$$

Define the Euclidean shadow point

$$y_k := \mathbf{u}_k - c_k D^\dagger m_{k-1}. \tag{C.24}$$

Using (C.2) and the definition of α_k in Appendix C.1,

$$\begin{aligned} y_{k+1} &= \mathbf{u}_k - (\eta_k + c_{k+1})D^\dagger m_k \\ &= \mathbf{u}_k - \beta_k(\eta_k + c_{k+1})D^\dagger m_{k-1} - (1-\beta_k)(\eta_k + c_{k+1})D^\dagger g_k \\ &= y_k - \alpha_k D^\dagger g_k. \end{aligned} \tag{C.25}$$

Moreover, (C.1) and Lemma 2 imply

$$\|y_k - \mathbf{u}_k\| \le \frac{\beta M \|D^\dagger\|}{1-\beta}\eta_k \longrightarrow 0. \tag{C.26}$$

Thus the shadow point and the original point have the same pointwise limits, while the first-moment state has disappeared from the one-step recursion.

C.5 Local Riemannian shadow reduction

Whenever $\mathbf{u}_k \in U$, let

$$z_k := \mathrm{Log}_{\mathbf{u}^\dagger}(\mathbf{u}_k) \in T_{\mathbf{u}^\dagger}\mathcal{M}. \tag{C.27}$$

Use the blockwise trivialization furnished by Lemma 9 and retain hats for all sample-dependent tangent vectors transported into the fixed space $T_{\mathbf{u}^\dagger}\mathcal{M}$. The coordinate noise $\widehat{\xi}_k$ is defined in Section 3.2. Define the remaining hatted variables by

$$\widehat{g}_k := \mathcal{T}_{\mathbf{u}_k} g_k, \qquad \widehat{m}_k := \mathcal{T}_{\mathbf{u}_k} m_k.$$

For $x = \mathrm{Exp}_{\mathbf{u}^\dagger}(z)$, use the coordinate maps from Appendices B.1 and B.3 to abbreviate the position drift as

$$F(z, v) := -\mathcal{A}(z)D(v)\mathcal{G}(z).$$

The corresponding noise coefficient is

$$\mathcal{B}_{\mathrm{noise}}(z, v) := -\mathcal{A}(z)D(v). \tag{C.28}$$

Then

$$\widehat{g}_k = \mathcal{G}(z_k) + \widehat{\xi}_k.$$

The shadow correction uses the preconditioner known before the current sample and the same chart differential that appears in the position update:

$$y_k := z_k - c_k\mathcal{A}(z_k)D(v_{k-1})\widehat{\tau}_{k-1}, \tag{C.29}$$

where

$$\widehat{\tau}_{k-1} := \mathcal{T}_{\mathbf{u}_k}\tau_{k-1}, \qquad \widehat{m}_k = \beta_k\widehat{\tau}_{k-1} + (1-\beta_k)\widehat{g}_k. \tag{C.30}$$

Lemma 12. Local Riemannian shadow reduction. Choose nested normal neighborhoods, shrinking U_{sh} from Lemma 9 if necessary, $U_{\mathrm{sh}} \Subset U_Q \Subset U_{\mathrm{stop}} \Subset U_{\mathrm{geo}}$, stop the process at the first exit from U_{stop}, and consider a tail on which $\mathbf{u}_k \in U_{\mathrm{sh}}$. Under Assumptions 1–2 and the parameter conditions of Theorems 1–2, the construction of Lemma 9 gives the following representation for all sufficiently large stopped indices:

$$\begin{aligned} y_{k+1} &= y_k + \alpha_k\Big(F(y_k, v_{k-1}) + \mathcal{B}_{\mathrm{noise}}(y_k, v_{k-1})\widehat{\xi}_k\Big) + e_{k+1}, \\ v_k &= v_{k-1} + \omega_k\Big(\widetilde{Q}(y_k) - v_{k-1} + \zeta_k\Big) + e^Q_{k+1}, \end{aligned} \tag{C.31}$$

The remainders can be chosen so that the following properties hold on every relatively compact localized neighborhood:

1. F is locally $C^{1,\nu}$, while $\mathcal{B}_{\mathrm{noise}}$ is smooth in v and locally at least $C^{1,\nu}$ in y;
2. $F(0, v) = 0$ for every nearby v;
3. in the normal frame,
$$D_y F(0, v^\dagger) = -D^\dagger H^\dagger; \tag{C.32}$$
4. $\mathcal{B}_{\mathrm{noise}}(0, v^\dagger) = -D^\dagger$;
5. y_k, z_k, v_{k-1} are $\mathcal{F}_k$-measurable, while $\widehat{\xi}_k, \zeta_k, e_{k+1}, e^Q_{k+1}$ are $\mathcal{F}_{k+1}$-measurable, and
$$\mathbb{E}[\widehat{\xi}_k \mid \mathcal{F}_k] = 0, \qquad \mathbb{E}[\zeta_k \mid \mathcal{F}_k] = 0. \tag{C.33}$$
No conditional-centering property is asserted or needed for e_{k+1} or e^Q_{k+1};
6. there is a deterministic local constant C_U such that
$$\|e_{k+1}\| + \|e^Q_{k+1}\| \le C_U \alpha_k^2 \tag{C.34}$$
for all sufficiently large stopped indices, and hence
$$\sum_k \left\| \frac{(e_{k+1}, e^Q_{k+1})}{\alpha_k} \right\|^2 < \infty. \tag{C.35}$$

Finally,

$$\|y_k - z_k\| = O(\eta_k), \tag{C.36}$$

so $\mathbf{u}_k \to \mathbf{u}^\dagger$ if and only if $y_k \to 0$.

Proof. With $U_{\mathrm{sh}} \Subset U_Q \Subset U_{\mathrm{stop}} \Subset U_{\mathrm{geo}}$ as in the statement, the bounds $c_k \to 0$, $\|m_k\| \le M$, and boundedness of $\mathcal{A}D$ first give the crude estimate $\|y_k - z_k\| \le C c_k \to 0$. Lemma 3 gives $\|\mathcal{T}_{\mathbf{u}_k} s_k\| = \|s_k\|_{\mathbf{u}_k} \le \eta_k B_{\mathrm{step}} \to 0$ and $d(\mathbf{u}_{k+1}, \mathbf{u}_k) \to 0$. Taking $r_{U_{\mathrm{stop}}}$ from Lemma 9 and increasing the tail index if necessary, one has $\|\mathcal{T}_{\mathbf{u}_k} s_k\| \le r_{U_{\mathrm{stop}}}$, while $\mathbf{u}_k, \mathbf{u}_{k+1}$ and the virtual point $\mathrm{Exp}_{\mathbf{u}^\dagger}(y_k)$ lie in U_{stop}. We prove the identities for those indices and stop before leaving U_{stop}. All constants below are deterministic and uniform on the compact parameter sets supplied by Lemma 9.

Step 1: exact chart and transport identities. Use the one-step map Φ and transport map $\mathscr{R}$ from (B.5)–(B.6). By construction,

$$\Phi(z, 0) = z, \qquad D_u \Phi(z, 0) = \mathcal{A}(z), \qquad \mathscr{R}(z, 0) = I.$$

Taylor's formula with integral remainder gives the exact identities

$$\begin{aligned} \Phi(z, u) &= z + \mathcal{A}(z)u + \rho^\Phi(z, u), \\ \rho^\Phi(z, u) &:= \int_0^1 (1 - t) D^2_{uu} \Phi(z, tu)[u, u]\, dt, \\ \mathscr{R}(z, u) &= I + \rho^R(z, u), \\ \rho^R(z, u) &:= \int_0^1 D_u \mathscr{R}(z, tu)[u]\, dt. \end{aligned} \tag{C.37}$$

Lemma 9 therefore supplies a deterministic local constant C such that

$$\|\rho^{\Phi}(z,u)\| \le C\|u\|^2, \qquad \|\rho^{R}(z,u)\| \le C\|u\|. \tag{C.38}$$

By the conjugacy formula (B.2), the fixed-coordinate tangent step is exactly

$$\widehat{s}_k := \mathcal{T}_{\mathbf{u}_k} s_k = -\eta_k D(v_k)\widehat{m}_k. \tag{C.39}$$

Lemma 2 and boundedness of $D(v_k)$ give

$$\|\widehat{s}_k\| \le C\eta_k. \tag{C.40}$$

The maps in (B.5)–(B.6), equations (C.37)–(C.39), and the definition of τ_k now give the exact relations

$$z_{k+1} = z_k + \mathcal{A}(z_k)\widehat{s}_k + \rho^{\Phi}_{k+1}, \qquad \widehat{\tau}_k = \mathscr{R}(z_k,\widehat{s}_k)\widehat{m}_k, \qquad \rho^{\Phi}_{k+1} := \rho^{\Phi}(z_k,\widehat{s}_k). \tag{C.41}$$

Step 2: exact shadow cancellation. Applying (C.29) at indices k and $k+1$, and then substituting (C.39) and (C.41), gives the fixed-coordinate identity

$$\begin{aligned} y_{k+1} - y_k = \; & -\eta_k \mathcal{A}(z_k)D(v_k)\widehat{m}_k + \rho^{\Phi}_{k+1} \\ & - c_{k+1}\mathcal{A}(z_{k+1})D(v_k)\mathscr{R}(z_k,\widehat{s}_k)\widehat{m}_k + c_k\mathcal{A}(z_k)D(v_{k-1})\widehat{\tau}_{k-1}. \end{aligned} \tag{C.42}$$

Define the one-step geometric operator mismatch

$$E^{\mathrm{geo}}_k := \mathcal{A}(z_{k+1})D(v_k)\mathscr{R}(z_k,\widehat{s}_k) - \mathcal{A}(z_k)D(v_k). \tag{C.43}$$

It has the exact decomposition

$$E^{\mathrm{geo}}_k = \big(\mathcal{A}(z_{k+1}) - \mathcal{A}(z_k)\big)D(v_k)\mathscr{R}(z_k,\widehat{s}_k) + \mathcal{A}(z_k)D(v_k)\big(\mathscr{R}(z_k,\widehat{s}_k) - I\big). \tag{C.44}$$

Notice the distinction between the operator mismatch and its contribution to the shadow update. Smoothness of $\mathcal{A}$, (C.38), and (C.40) give

$$\begin{aligned} \|\mathcal{A}(z_{k+1}) - \mathcal{A}(z_k)\| \le C\|z_{k+1} - z_k\| \le C\eta_k, \\ \|\mathscr{R}(z_k,\widehat{s}_k) - I\| \le C\eta_k, \\ \|E^{\mathrm{geo}}_k\| \le C\eta_k. \end{aligned} \tag{C.45}$$

Thus the operator mismatch is $O(\eta_k)$; only after multiplication by $c_{k+1} = O(\eta_k)$ does its contribution become $O(\eta_k^2)$.

Using (C.43), equation (C.42) becomes

$$\begin{aligned} y_{k+1} - y_k = \; & -(\eta_k + c_{k+1})\mathcal{A}(z_k)D(v_k)\widehat{m}_k + c_k\mathcal{A}(z_k)D(v_{k-1})\widehat{\tau}_{k-1} \\ & - c_{k+1}E^{\mathrm{geo}}_k\widehat{m}_k + \rho^{\Phi}_{k+1}, \end{aligned} \tag{C.46}$$

Substitute the exact momentum relation (C.30). The coefficient of $\widehat{\tau}_{k-1}$ is

$$c_k - \beta_k(\eta_k + c_{k+1}) = 0$$

by (C.2), except for the explicit lag $D(v_{k-1}) - D(v_k)$. The coefficient of $\widehat{g}_k$ is

$$-(1-\beta_k)(\eta_k + c_{k+1}) = -\alpha_k$$

by the equivalent step identity in Appendix C.1. Hence (C.46) is exactly

$$\begin{aligned} y_{k+1} - y_k = & -\alpha_k \mathcal{A}(z_k) D(v_k) \widehat{g}_k + c_k \mathcal{A}(z_k) \big(D(v_{k-1}) - D(v_k)\big) \widehat{\tau}_{k-1} \\ & - c_{k+1} E_k^{\mathrm{geo}} \widehat{m}_k + \rho_{k+1}^{\Phi}. \end{aligned} \tag{C.47}$$

Equation (C.47), rather than an order estimate, is the exact discrete momentum elimination identity.

Step 3: conversion to the predictable drift and noise coefficient. The cancellation in (C.47) still uses the current-sample preconditioner $D(v_k)$ and the original coordinate z_k. To identify a conditionally centered noise term and a drift evaluated at the shadow state, replace these by $D(v_{k-1})$ and y_k, retaining each difference exactly. Together with the chart and transport errors from Steps 1–2, these differences give six remainders:

$$\begin{aligned} e_{k+1}^{\mathrm{chart}} &:= \rho_{k+1}^{\Phi}, \\ e_{k+1}^{\mathrm{geo}} &:= -c_{k+1} E_k^{\mathrm{geo}} \widehat{m}_k, \\ e_{k+1}^{D,\mathrm{old}} &:= c_k \mathcal{A}(z_k) \big(D(v_{k-1}) - D(v_k)\big) \widehat{\tau}_{k-1}, \\ e_{k+1}^{D,\mathrm{new}} &:= -\alpha_k \mathcal{A}(z_k) \big(D(v_k) - D(v_{k-1})\big) \widehat{g}_k, \\ e_{k+1}^{y,\mathrm{drift}} &:= -\alpha_k \big(\mathcal{A}(z_k) D(v_{k-1}) \mathcal{G}(z_k) - \mathcal{A}(y_k) D(v_{k-1}) \mathcal{G}(y_k)\big), \\ e_{k+1}^{y,\mathrm{noise}} &:= -\alpha_k \big(\mathcal{A}(z_k) - \mathcal{A}(y_k)\big) D(v_{k-1}) \widehat{\xi}_k, \end{aligned} \tag{C.48}$$

and set

$$e_{k+1} := e_{k+1}^{\mathrm{chart}} + e_{k+1}^{\mathrm{geo}} + e_{k+1}^{D,\mathrm{old}} + e_{k+1}^{D,\mathrm{new}} + e_{k+1}^{y,\mathrm{drift}} + e_{k+1}^{y,\mathrm{noise}}. \tag{C.49}$$

Because $\widehat{g}_k = \mathcal{G}(z_k) + \widehat{\xi}_k$, adding and subtracting the terms in (C.48) transforms (C.47), with no approximation, into

$$y_{k+1} - y_k = \alpha_k \Big(F(y_k, v_{k-1}) + \mathcal{B}_{\mathrm{noise}}(y_k, v_{k-1}) \widehat{\xi}_k \Big) + e_{k+1}. \tag{C.50}$$

The use of $D(v_{k-1})$ in the leading noise coefficient is essential. The random operator $D(v_k)$ depends on the current sample g_k, so $D(v_k)\widehat{\xi}_k$ need not be conditionally centered. In contrast, $\mathcal{A}(y_k)D(v_{k-1})$ is $\mathcal{F}_k$-measurable.

It remains to estimate the explicitly listed terms. The second-moment update and the bounds of Lemma 2 imply

$$\|v_k - v_{k-1}\| \le C\omega_k, \qquad \|D(v_k) - D(v_{k-1})\| \le C\omega_k. \tag{C.51}$$

The shadow definition gives

$$\|z_k - y_k\| \le c_k \|\mathcal{A}(z_k)\| \|D(v_{k-1})\| \|\widehat{\tau}_{k-1}\| \le C c_k. \tag{C.52}$$

The maps $\mathcal{A}$ and $\mathcal{G}$ are locally Lipschitz. Using (C.38), (C.45), (C.51), (C.52), boundedness of $\widehat{g}_k, \widehat{m}_k, \widehat{\tau}_{k-1}, \widehat{\xi}_k$, and $c_k, c_{k+1}, \omega_k, \eta_k = O(\alpha_k)$ gives, term by term,

$$\begin{aligned} \|e_{k+1}^{\mathrm{chart}}\| &\le C\eta_k^2, \\ \|e_{k+1}^{\mathrm{geo}}\| &\le C c_{k+1} \eta_k, \\ \|e_{k+1}^{D,\mathrm{old}}\| &\le C c_k \omega_k, \\ \|e_{k+1}^{D,\mathrm{new}}\| &\le C \alpha_k \omega_k, \\ \|e_{k+1}^{y,\mathrm{drift}}\| + \|e_{k+1}^{y,\mathrm{noise}}\| &\le C \alpha_k \|z_k - y_k\|. \end{aligned} \tag{C.53}$$

Every quantity on the right of (C.53) is $O(\alpha_k^2)$. Therefore (C.34) holds for the position remainder once the second-moment remainder is included below.

Step 4: exact second-moment remainder. From (C.12) and (3.4), the coordinate second-moment update is exactly

$$v_k - v_{k-1} = \omega_k\big(\widetilde{Q}(z_k) - v_{k-1} + \zeta_k\big). \tag{C.54}$$

Define

$$e_{k+1}^Q := \omega_k\big(\widetilde{Q}(z_k) - \widetilde{Q}(y_k)\big). \tag{C.55}$$

Then (C.54) is the second line of (C.31) exactly. Local Lipschitz continuity of $\widetilde{Q}$, (C.52), and $\omega_k = O(\alpha_k)$ give

$$\|e_{k+1}^Q\| \le C\omega_k\|z_k - y_k\| \le C\alpha_k^2. \tag{C.56}$$

Combining (C.53) and (C.56) gives

$$\|e_{k+1}\| + \|e_{k+1}^Q\| \le C_U\alpha_k^2, \qquad \sum_k \left\|\frac{(e_{k+1}, e_{k+1}^Q)}{\alpha_k}\right\|^2 \le C_U^2\sum_k \alpha_k^2 < \infty. \tag{C.57}$$

Step 5: measurability and centering. The state variables z_k, y_k, v_{k-1} and the drift and noise coefficients evaluated at them are $\mathcal{F}_k$-measurable. Since g_k is $\mathcal{F}_{k+1}$-measurable, the deterministic update maps and the definitions in (C.48)–(C.49) and (C.55) make e_{k+1} and e_{k+1}^Q $\mathcal{F}_{k+1}$-measurable.

Since $\mathcal{T}_{\mathbf{u}_k}$ is $\mathcal{F}_k$-measurable and linear,

$$\mathbb{E}[\widehat{\xi}_k \mid \mathcal{F}_k] = \mathcal{T}_{\mathbf{u}_k}\mathbb{E}[\xi_k \mid \mathcal{F}_k] = 0. \tag{C.58}$$

The definition of ζ_k and (3.4) similarly give $\mathbb{E}[\zeta_k \mid \mathcal{F}_k] = 0$. This proves (C.33); the remainders need only adaptedness, not conditional centering.

Finally, (C.52), (C.1), and (C.3) prove (C.36). On a localized tail the normal chart is a homeomorphism and $\|y_k - z_k\| \to 0$, so $\mathbf{u}_k \to \mathbf{u}^\dagger$ if and only if $y_k \to 0$.

The regularity statements follow from $F(z, v) = -\mathcal{A}(z)D(v)\mathcal{G}(z)$, (B.12), Assumption 2, and Lemma 9. Since $\mathcal{G}(0) = 0$, one has $F(0, v) = 0$, and differentiation at $(0, v^\dagger)$ gives

$$D_y F(0, v^\dagger) = -D^\dagger H^\dagger, \qquad \mathcal{B}_{\mathrm{noise}}(0, v^\dagger) = -D^\dagger.$$

This completes (C.31)–(C.36). □

C.6 Exact drift compatibility

Define the augmented shadow state

$$Z_k := (y_k, v_{k-1}). \tag{C.59}$$

The equilibrium is $Z^\dagger = (0, v^\dagger)$ from Appendix B.3.

Recall the vector field h of the Reduced Shadow Mean-Field ODE in Appendix B.3; in the present notation it is evaluated at $Z_k = (y_k, v_{k-1})$.

Define the martingale innovation

$$\varepsilon_{k+1} := \begin{pmatrix} \mathcal{B}_{\mathrm{noise}}(y_k, v_{k-1})\widehat{\xi}_k \\ q_k\zeta_k \end{pmatrix} \tag{C.60}$$

and the normalized remainder

$$\rho_{k+1} := \begin{pmatrix} e_{k+1}/\alpha_k \\ e^Q_{k+1}/\alpha_k + (q_k - q)(\widetilde{Q}(y_k) - v_{k-1}) \end{pmatrix}. \tag{C.61}$$

Proposition 4. Continuous–discrete compatibility. On every localized tail, the transformed Algorithm 1 satisfies exactly

$$Z_{k+1} = Z_k + \alpha_k\big(h(Z_k) + \varepsilon_{k+1} + \rho_{k+1}\big). \tag{C.62}$$

Here Z_k is $\mathcal{F}_k$-measurable, while ε_{k+1} and ρ_{k+1} are $\mathcal{F}_{k+1}$-measurable. Moreover,

$$\mathbb{E}[\varepsilon_{k+1} \mid \mathcal{F}_k] = 0 \tag{C.63}$$

and

$$\sum_k \|\rho_{k+1}\|^2 \le C\left(\sum_k \alpha_k^2 + \sum_k |q_k - q|^2\right) < \infty. \tag{C.64}$$

Proof. Substitute $\omega_k = q_k\alpha_k$ into the second line of (C.31), then add and subtract $q(\widetilde{Q}(y_k) - v_{k-1})$. The two lines of (C.31) become (C.62) with (C.60)–(C.61).

The coefficients in (C.60) are $\mathcal{F}_k$-measurable and q_k is deterministic, so Lemma 12 gives the stated measurability and centering property (C.63). The definitions in (C.61) also give the $\mathcal{F}_{k+1}$-measurability of ρ_{k+1}, without a centering requirement. Lemma 12, the first condition in (3.6), and local boundedness of $\widetilde{Q}(y) - v$ give (C.64). □

Proposition 4 identifies the vector field h of the Reduced Shadow Mean-Field ODE with the drift of the exact shadow recursion.

Appendix D applies Theorem D.1 to this recursion after verifying its spectral, excitation, moment, remainder, tail, and localization conditions.

D Proof of Theorem 2

Stochastic nonconvergence results cover Euclidean momentum methods (Barakat et al., 2021) and Riemannian stochastic approximation (Hsieh et al., 2023). For Algorithm 1, the exact shadow recursion in Appendix C permits the verification below of the common-step stochastic nonconvergence theorem.

D.1 Raimond–Tarrès nonconvergence theorem

The final implication uses the mixed-spectrum branch of the nonconvergence theorem (Raimond & Tarrès, 2023, Theorem 2.2.4(ii)). We use the following common-step specialization for (C.62).

Theorem D.1. Common-step specialization of Raimond–Tarrès. Let $(\mathcal{F}_k)$ be a filtration and let X_k be an $\mathcal{F}_k$-measurable process in a finite-dimensional Euclidean space. Suppose, after an inessential index shift, that

$$X_{k+1} - X_k = a_k\big(\mathsf{h}(X_k) + \varepsilon_{k+1} + \rho_{k+1}\big), \tag{D.1}$$

where $a_k > 0$ is deterministic,

$$\sum_k a_k = \infty, \qquad \sum_k a_k^2 < \infty, \tag{D.2}$$

ε_{k+1} and ρ_{k+1} are $\mathcal{F}_{k+1}$-measurable, ε_{k+1} is conditionally square integrable, and

$$\mathbb{E}[\varepsilon_{k+1} \mid \mathcal{F}_k] = 0,$$

and no conditional-centering assumption is imposed on ρ_{k+1}. Let $X^\dagger$ be an unstable equilibrium of h. Assume that h is $C^{1,\nu}$ on a convex neighborhood of $X^\dagger$ for some $\nu \in (0, 1]$.

Choose an invertible real spectral coordinate map that splits the positive-real-part generalized spectral subspace of $D\mathsf{h}(X^\dagger)$ from its complementary nonpositive-real-part subspace. Write

$$\varepsilon_k = (\varepsilon_k^+, \varepsilon_k^-). \tag{D.3}$$

For an event Γ, suppose that for some $p_0 > 2$, almost surely on Γ,

$$0 < \liminf_k \mathbb{E}[\|\varepsilon_{k+1}^+\|^2 \mid \mathcal{F}_k], \tag{D.4}$$

$$\limsup_k \mathbb{E}[\|\varepsilon_{k+1}\|^{p_0} \mid \mathcal{F}_k] < \infty, \qquad \sum_k \|\rho_{k+1}\|^2 < \infty. \tag{D.5}$$

Set

$$A(t) := \left(\sum_{k>t} a_k^2\right)^{1/2}. \tag{D.6}$$

The mixed-spectrum branch further assumes

$$\sum_{k>t} a_k^2 = O(A(t)) \tag{D.7}$$

and

$$\sum_{k>t} a_k^{1+\nu} \|\varepsilon_{k+1}^-\|^{1+\nu} = o(A(t)) \tag{D.8}$$

on $\Gamma \cap \{X_k \to X^\dagger\}$. Then

$$\mathbb{P}\big(\Gamma \cap \{X_k \to X^\dagger\}\big) = 0. \tag{D.9}$$

This is Theorem 2.2.4(ii) whose recurrence is $X_{n+1} - X_n = \gamma_{n+1} G_n + c_{n+1}(\varepsilon_{n+1} + r_{n+1}^{\mathrm{RT}})$ specialized by setting both γ_{n+1} and c_{n+1} equal to a_k and $G_n = \mathsf{h}(X_k)$. Here r_{n+1}^{RT} denotes the remainder in that theorem; the same object is denoted ρ_{k+1} in (D.1) and in the application below. Raimond–Tarrès write $C^{1+\nu}$ for the same Hölder-Jacobian regularity denoted $C^{1,\nu}$ here. The mixed-spectrum branch permits zero and stable eigenvalues in the complementary block; it does not require the equilibrium to be purely repulsive. In their discrete-time setup, the noise, remainder, and drift sequences are adapted. Therefore, under the index convention in (D.1), ρ_{k+1} being $\mathcal{F}_{k+1}$-measurable is exactly the required adaptedness; neither Theorem 2.2.4(ii) nor Theorem D.1 requires ρ_{k+1} to be $\mathcal{F}_k$-measurable.

Theorem D.1 permits $\nu \in (0, 1]$. Assumption 2 takes $\nu \in [1/2, 1]$ for Theorem 2, and (C.9)–(C.10) verify the endpoint schedule.

D.2 Regularity and spectrum

The vector field used in Proposition 4 is exactly the one analyzed in Appendices B.3 and B.5. The open-neighborhood construction in Appendix B.5, using the local regularity and conditional-moment conditions in Assumption 2 and Lemma 9, gives

$$h \in C^{1,\nu} \quad \text{near } Z^\dagger. \tag{D.10}$$

Its Jacobian is (B.24), and Proposition 2 gives

$$\dim E^u = \mathfrak{m} \geq 1. \tag{D.11}$$

Thus $Z^\dagger$ is an unstable equilibrium in the sense of Theorem D.1. Possible zero Hessian eigenvalues lie in the complementary block and are allowed by branch (ii).

D.3 Excitation of the actual unstable coordinates

Assumption 3 imposes the asymptotic excitation condition (3.5) on $E^\dagger := \{\mathbf{u}_k \to \mathbf{u}^\dagger\}$. This restriction specifies the paths on which excitation is required; the conditional expectations remain conditioned on $\mathcal{F}_k$.

A sufficient condition for (3.5) is the existence of an open neighborhood U_{exc} of $\mathbf{u}^\dagger$, with $\overline{U_{\mathrm{exc}}} \subset U$, and a deterministic constant $\sigma_u > 0$ such that, for every k, whenever $\mathbf{u}_k \in U_{\mathrm{exc}}$,

$$\mathbb{E}\left[\left\|\Pi_-^\dagger (D^\dagger)^{1/2} \mathcal{T}_{\mathbf{u}_k} \xi_k\right\|^2 \middle| \mathcal{F}_k\right] \geq \sigma_u^2 \qquad \text{a.s.}$$

Indeed, on $E^\dagger$, the iterates eventually remain in U_{exc}, so this bound implies (3.5). The zero extension of $\widehat{\xi}_k$ outside U is immaterial on this event. The argument below uses (3.5) directly: we identify the unstable coordinates of the augmented recursion and verify the excitation required by Theorem D.1.

Choose an orthonormal eigenbasis $U_-^\dagger = (u_1, \ldots, u_{\mathfrak{m}})$ for $E_-^\dagger$, with $K^\dagger u_i = \kappa_i u_i$ and $\kappa_i < 0$. Define the left spectral-coordinate map for the unstable block

$$L_+ := (U_-^\dagger)^\top (D^\dagger)^{-1/2}. \tag{D.12}$$

For the position Jacobian $-D^\dagger H^\dagger$,

$$L_+(-D^\dagger H^\dagger) = \mathrm{diag}(-\kappa_1, \ldots, -\kappa_{\mathfrak{m}}) L_+, \qquad -\kappa_i > 0. \tag{D.13}$$

Because the upper-right block of (B.24) is zero and the second diagonal block is $-qI$, the rows $(L_+, 0)$ have full row rank and satisfy

$$(L_+, 0) J^\dagger = \Lambda_+ (L_+, 0), \qquad \Lambda_+ := \mathrm{diag}(-\kappa_1, \ldots, -\kappa_{\mathfrak{m}}) \succ 0. \tag{D.14}$$

To identify the complement explicitly, put $\mu_i := -\kappa_i > 0$ and

$$p_i := \begin{pmatrix} (D^\dagger)^{1/2} u_i \\ \dfrac{q}{\mu_i + q} D\widetilde{Q}(0)(D^\dagger)^{1/2} u_i \end{pmatrix}.$$

Equation (B.24) gives $J^\dagger p_i = \mu_i p_i$, while orthonormality of the u_i gives $(L_+, 0) p_i = e_i$, the i-th coordinate vector in $\mathbb{R}^{\mathfrak{m}}$. Thus the restriction of $(L_+, 0)$ to E^u is an isomorphism. Its common kernel is $J^\dagger$-invariant by (D.14), has the complementary dimension, and, by uniqueness of the invariant spectral splitting for the disjoint positive and nonpositive spectra, is exactly the generalized spectral

subspace for eigenvalues with nonpositive real part. The rows can consequently be completed by a block L_{cs} to an invertible real matrix

$$\mathcal{P}^{-1} := \begin{pmatrix} L_+ & 0 \\ L_{\mathrm{cs}} & \end{pmatrix}, \qquad \mathcal{P}^{-1} J^\dagger \mathcal{P} = \begin{pmatrix} \Lambda_+ & 0 \\ 0 & H^{\mathrm{cs}} \end{pmatrix}, \qquad \max \mathrm{Re}\,\mathrm{spec}(H^{\mathrm{cs}}) \le 0. \tag{D.15}$$

This is the invertible spectral coordinate map required in Theorem D.1, with $(L_+, 0)$ as its positive block. Define the spectral-coordinate decomposition of the innovation by

$$\begin{pmatrix} \varepsilon^+_{k+1} \\ \varepsilon^-_{k+1} \end{pmatrix} := \mathcal{P}^{-1} \varepsilon_{k+1}.$$

In particular, the second-moment innovation does not enter the positive coordinates, and those coordinates of the actual martingale innovation are

$$\varepsilon^+_{k+1} := L_+ \mathcal{B}_{\mathrm{noise}}(y_k, v_{k-1}) \widehat{\xi}_k. \tag{D.16}$$

At the equilibrium $\mathcal{B}^\dagger_{\mathrm{noise}} := \mathcal{B}_{\mathrm{noise}}(0, v^\dagger) = -D^\dagger$, so

$$L_+ \mathcal{B}^\dagger_{\mathrm{noise}} \widehat{\xi}_k = -(U^\dagger_-)^\top (D^\dagger)^{1/2} \widehat{\xi}_k, \qquad \|L_+ \mathcal{B}^\dagger_{\mathrm{noise}} \widehat{\xi}_k\| = \|\Pi^\dagger_- (D^\dagger)^{1/2} \widehat{\xi}_k\|. \tag{D.17}$$

On $E^\dagger$, the iterates eventually remain in U_{sh}, where the local chart and shadow construction apply. Along this tail, Lemma 11 and (C.36) give $v_{k-1} \to v^\dagger$ and $y_k \to 0$ almost surely. Consequently, $\mathcal{B}_{\mathrm{noise}}(y_k, v_{k-1}) \to \mathcal{B}^\dagger_{\mathrm{noise}}$. Conditional Minkowski's inequality and $\|\widehat{\xi}_k\| \le 2M$ yield

$$\begin{aligned} &\Big| \mathbb{E}[\|L_+ \mathcal{B}_{\mathrm{noise}}(y_k, v_{k-1}) \widehat{\xi}_k\|^2 \mid \mathcal{F}_k]^{1/2} - \mathbb{E}[\|L_+ \mathcal{B}^\dagger_{\mathrm{noise}} \widehat{\xi}_k\|^2 \mid \mathcal{F}_k]^{1/2} \Big| \\ &\quad \le \|L_+ \big(\mathcal{B}_{\mathrm{noise}}(y_k, v_{k-1}) - \mathcal{B}^\dagger_{\mathrm{noise}}\big)\| \mathbb{E}[\|\widehat{\xi}_k\|^2 \mid \mathcal{F}_k]^{1/2} \\ &\quad \le 2M \|L_+ \big(\mathcal{B}_{\mathrm{noise}}(y_k, v_{k-1}) - \mathcal{B}^\dagger_{\mathrm{noise}}\big)\| \longrightarrow 0. \end{aligned} \tag{D.18}$$

Assumption 3 and (D.17)–(D.18) therefore give the asymptotic excitation condition (D.4) almost surely on $E^\dagger$, evaluated along this tail.

The stochastic gradients are bounded, $\mathcal{B}_{\mathrm{noise}}(y_k, v_{k-1})$ is bounded on the localized region, and $\mathcal{T}_{\mathbf{u}_k}$ is an isometry. Hence $\widehat{\xi}_k$, ζ_k, and ε_{k+1} have uniformly bounded conditional moments of every finite order. This verifies the moment part of (D.5) for any fixed $p_0 > 2$.

Assumption 3 and the local sufficient condition above both permit singular, even rank-one, conditional covariance: they require only aggregate noise energy in the unstable subspace. Excitation in every unstable or tangent direction is not required. Condition (3.5) is independent of the chosen admissible blockwise isometric trivialization: two such choices differ by an orthogonal map converging to I at $\mathbf{u}^\dagger$, while boundedness of ξ_k makes the associated conditional L^2 projection norms differ by $o(1)$ on $E^\dagger$.

D.4 Remainder and tail conditions

Set $a_k = \alpha_k$. Equation (C.4) gives (D.2), and Proposition 4 gives

$$\sum_k \|\rho_{k+1}\|^2 < \infty. \tag{D.19}$$

This is square summability of the normalized remainder in (C.62). Absolute summability of the unnormalized increment is neither asserted nor substituted for (D.19).

Since $\sum_{k>t} \alpha_k^2 \to 0$, one has eventually

$$\sum_{k>t} \alpha_k^2 \le \left(\sum_{k>t} \alpha_k^2 \right)^{1/2} = \mathfrak{a}(t). \tag{D.20}$$

Thus (D.7) is automatic. The complementary noise ε_{k+1}^- is bounded. Using the tail condition in (3.6) gives

$$\sum_{k>t} \alpha_k^{1+\nu} \|\varepsilon_{k+1}^-\|^{1+\nu} \le C \sum_{k>t} \alpha_k^{1+\nu} = o(\mathfrak{a}(t)). \tag{D.21}$$

For $\nu > 1/2$ and the polynomial schedule (C.8), $\alpha_k \asymp k^{-p}$, and the ratio $\sum_{k>t} \alpha_k^{1+\nu} / \mathfrak{a}(t)$ is $O(t^{1/2-p\nu})$. Hence $p\nu > 1/2$ is sufficient (Raimond & Tarrès, 2023, Remark 2.10). At $\nu = 1/2$, (C.9)–(C.10) replace this power ratio by $O((\log t)^{-r/2})$, which still tends to zero.

D.5 Localization

Only local regularity is assumed. Choose convex augmented coordinate neighborhoods

$$\mathcal{V} \Subset \mathcal{U} \tag{D.22}$$

of $Z^\dagger$ with $\overline{\mathcal{U}}$ contained in the open domain of the reduced field specified in Appendix B.5 and its position projection compactly contained in $\mathrm{Log}_{\mathbf{u}^\dagger}(U_{\mathrm{stop}})$. The algorithmic base point will be restricted separately below. Choose $\chi \in C_c^\infty(\mathcal{U}; [0,1])$ with $\chi = 1$ on $\mathcal{V}$ and define $\bar{h} = \chi h$ on $\mathcal{U}$, extended by zero to the ambient Euclidean space. Then $\bar{h}$ is $C^{1,\nu}$, agrees with h near $Z^\dagger$, and has the same Jacobian and spectral splitting there.

To define the localization without ever evaluating a partial quantity, let

$$\mathscr{S} := T\mathcal{M} \times [0,\infty)^n, \qquad \Theta_k := (\mathbf{u}_k, \tau_{k-1}, v_{k-1}).$$

The variable Θ_k is $\mathcal{F}_k$-measurable. On the Borel local domain $\mathscr{D}_{\mathrm{loc}} := TU \times [0,\infty)^n \subset \mathscr{S}$, define the deterministic Borel map

$$\begin{aligned} &\Psi_k : \mathscr{D}_{\mathrm{loc}} \longrightarrow \mathbb{R}^{d+n}, \\ &\Psi_k(x,\tau,v) := \left(z - c_k \mathcal{A}(z) D(v) \mathcal{T}_x \tau,\ v\right), \qquad z := \mathrm{Log}_{\mathbf{u}^\dagger}(x). \end{aligned} \tag{D.23}$$

Whenever the local shadow state is defined, (C.29) and (C.59) give $Z_k = \Psi_k(\Theta_k)$. Use the nested neighborhoods $U_{\mathrm{sh}} \Subset U_Q \Subset U_{\mathrm{stop}} \Subset U_{\mathrm{geo}}$ from Lemma 12. On the probability-one event of bounded algorithmic moments, $\mathbf{u}_k \in U_{\mathrm{sh}}$ gives a uniform bound $\|y_k - z_k\| \le C c_k = O(\eta_k)$. The one-step bound from Lemma 3 and compact containment therefore give a deterministic K_{loc} such that, for $k \ge K_{\mathrm{loc}}$, the next base point and the shadow points needed by Lemma 12 remain in its coordinate region whenever $\mathbf{u}_k \in U_{\mathrm{sh}}$ and $\Psi_k(\Theta_k) \in \mathcal{U}$. All terms in (C.60)–(C.61) and their local bounds are then available. In particular, $\mathbf{u}_k \in U_Q$ ensures that ζ_k equals the actual second-moment residual, while the base and shadow coordinates lie where Q, $\mathcal{T}$, and the geometric remainder estimates apply.

Discarding the finite prefix before K_{loc} is permitted by Theorem D.1. For $k \ge K_{\mathrm{loc}}$, define

$$\begin{aligned} \Omega_k^{\mathcal{U}} &:= \{\mathbf{u}_k \in U_{\mathrm{sh}}\} \cap \Theta_k^{-1}\big(\Psi_k^{-1}(\mathcal{U})\big), \\ \Omega_k^{\mathcal{V}} &:= \{\mathbf{u}_k \in U_{\mathrm{sh}}\} \cap \Theta_k^{-1}\big(\Psi_k^{-1}(\mathcal{V})\big). \end{aligned} \tag{D.24}$$

These events belong to $\mathcal{F}_k$, and $\Omega_k^{\mathcal{V}} \subseteq \Omega_k^{\mathcal{U}}$. Write $\varepsilon_{k+1}^{\mathrm{loc},k}$ and $\rho_{k+1}^{\mathrm{loc},k}$ for the maps on the trace-measurable space $\Omega_k^{\mathcal{U}}$ furnished by the local formulas (C.60)–(C.61). Define total random variables, pointwise in $\varpi \in \Omega$, by

$$Z_k^{\mathcal{U}}(\varpi) := \begin{cases} \Psi_k(\Theta_k(\varpi)), & \varpi \in \Omega_k^{\mathcal{U}}, \\ Z^\dagger, & \varpi \notin \Omega_k^{\mathcal{U}}, \end{cases}$$

$$\varepsilon_{k+1}^{\mathcal{U}}(\varpi) := \begin{cases} \varepsilon_{k+1}^{\mathrm{loc},k}(\varpi), & \varpi \in \Omega_k^{\mathcal{U}}, \\ 0, & \varpi \notin \Omega_k^{\mathcal{U}}, \end{cases} \tag{D.25}$$

$$\rho_{k+1}^{\mathcal{U}}(\varpi) := \begin{cases} \rho_{k+1}^{\mathrm{loc},k}(\varpi), & \varpi \in \Omega_k^{\mathcal{U}}, \\ 0, & \varpi \notin \Omega_k^{\mathcal{U}}. \end{cases}$$

The first branches of (D.25) are therefore evaluated only on their stated domains. Hence $Z_k^{\mathcal{U}}$ is $\mathcal{F}_k$-measurable, while $\varepsilon_{k+1}^{\mathcal{U}}$ and $\rho_{k+1}^{\mathcal{U}}$ are $\mathcal{F}_{k+1}$-measurable. For every $A \in \mathcal{F}_k$, predictability of $\Omega_k^{\mathcal{U}}$ and the calculation proving (C.63) give

$$\mathbb{E}\big[\mathbf{1}_A \varepsilon_{k+1}^{\mathcal{U}}\big] = \int_{A \cap \Omega_k^{\mathcal{U}}} \varepsilon_{k+1}^{\mathrm{loc},k}\, d\mathbb{P} = 0, \qquad \mathbb{E}[\varepsilon_{k+1}^{\mathcal{U}} \mid \mathcal{F}_k] = 0. \tag{D.26}$$

For each deterministic $k_0 \geq K_{\mathrm{loc}}$, define

$$\sigma_{k_0}^{\mathcal{V}}(\varpi) := \inf\{k \geq k_0 : \varpi \notin \Omega_k^{\mathcal{V}}\}, \qquad \inf \varnothing := \infty. \tag{D.27}$$

This is a stopping time because every $\Omega_k^{\mathcal{V}}$ belongs to $\mathcal{F}_k$. Define the globally Euclidean, adapted recursion by

$$\begin{aligned} &\bar{Z}_{k_0}^{(k_0)} := Z_{k_0}^{\mathcal{U}}, \\ &\bar{Z}_{k+1}^{(k_0)} := \bar{Z}_k^{(k_0)} + \alpha_k \Big(\bar{h}(\bar{Z}_k^{(k_0)}) + \chi(\bar{Z}_k^{(k_0)}) \varepsilon_{k+1}^{\mathcal{U}} + \chi(\bar{Z}_k^{(k_0)}) \rho_{k+1}^{\mathcal{U}} \Big), \qquad k \geq k_0. \end{aligned} \tag{D.28}$$

The coefficient $\chi(\bar{Z}_k^{(k_0)})$ is $\mathcal{F}_k$-measurable. Therefore its product with $\varepsilon_{k+1}^{\mathcal{U}}$ is still conditionally centered, while its product with $\rho_{k+1}^{\mathcal{U}}$ is $\mathcal{F}_{k+1}$-measurable and square summable on the event used below. Moreover,

$$\{\sigma_{k_0}^{\mathcal{V}} = \infty\} = \bigcap_{k \geq k_0} \Omega_k^{\mathcal{V}}. \tag{D.29}$$

On this event $\Omega_k^{\mathcal{U}}$ also holds on the entire tail. An induction using $\chi = 1$ and $\bar{h} = h$ on $\mathcal{V}$ therefore gives $\bar{Z}_k^{(k_0)} = Z_k$ for every $k \geq k_0$. On this event the extended noise, remainder, excitation, and square-summability conditions are exactly those of (C.62). Thus Theorem D.1 may be applied to the globally defined recursion after a finite index shift. This piecewise-extension-and-exit-time construction is the only local-to-Euclidean step.

D.6 Exclusion of a fixed strict saddle

Proof of the pointwise step. Fix $\mathbf{u}^\dagger \in \mathcal{S}$ and recall the candidate convergence event from Appendix D.3:

$$E^\dagger = \{\mathbf{u}_k \to \mathbf{u}^\dagger\}. \tag{D.30}$$

On $E^\dagger$, the iterates eventually remain in every sufficiently small neighborhood of $\mathbf{u}^\dagger$. The localized chart and shadow construction can therefore be used on a tail.

Lemma 11 gives

$$v_k \to v^\dagger \qquad \text{on } E^\dagger, \tag{D.31}$$

while (C.36) gives

$$y_k \to 0 \qquad \text{on } E^\dagger. \tag{D.32}$$

Hence, up to a null event,

$$E^\dagger \subseteq \{Z_k \to Z^\dagger\}. \tag{D.33}$$

Fix $\mathcal{V} \Subset \mathcal{U}$ and K_{loc} as in Appendix D.5. On the common probability-one event for the convergence and reduction statements in Lemmas 11–12 and Proposition 4, define

$$E_{k_0} := E^\dagger \cap \{\sigma_{k_0}^{\mathcal{V}} = \infty\}. \tag{D.34}$$

On $E^\dagger$, both $\mathbf{u}_k \in U_{\mathrm{sh}}$ and $Z_k \in \mathcal{V}$ hold for all sufficiently large k: the first follows from $\mathbf{u}_k \to \mathbf{u}^\dagger$, and the second from (D.33). Thus both restrictions in (D.24) hold on an eventual tail, giving

$$E^\dagger \subseteq \bigcup_{k_0=K_{\mathrm{loc}}}^{\infty} E_{k_0} \qquad \text{up to a null event.} \tag{D.35}$$

Fix $k_0 \geq K_{\mathrm{loc}}$. Apply Theorem D.1 to the shifted and extended recursion from Appendix D.5 with

$$a_k = \alpha_k, \qquad \Gamma = E_{k_0}. \tag{D.36}$$

Every imported-theorem hypothesis has already been verified:

- (C.4), (D.20)–(D.21) give the step and mixed-spectrum tail conditions;
- (D.10)–(D.11) give a local $C^{1,\nu}$ field with a nonempty positive spectral subspace;
- Assumption 3 and (D.17)–(D.18) give unstable-coordinate excitation on $E_{k_0} \subseteq E^\dagger$; boundedness of the stochastic gradients and the localized coefficients gives a $p_0 > 2$ conditional moment;
- (D.19) gives the square-summable normalized remainder;
- the cutoff construction provides the required pathwise localization (Appendix D.5).

Theorem D.1 yields

$$\mathbb{P}\big(E_{k_0} \cap \{\bar{Z}_k^{(k_0)} \to Z^\dagger\}\big) = 0. \tag{D.37}$$

On E_{k_0}, the extended recursion satisfies $\bar{Z}_k^{(k_0)} = Z_k$ for every $k \geq k_0$ (Appendix D.5), while (D.33) gives $Z_k \to Z^\dagger$. Hence $E_{k_0} \subseteq \{\bar{Z}_k^{(k_0)} \to Z^\dagger\}$, so $\mathbb{P}(E_{k_0}) = 0$. Taking the countable union in (D.35) gives $\mathbb{P}(E^\dagger) = 0$.

This establishes the fixed-saddle exclusion used in Theorem 2.

D.7 Countable-union completion

For the prescribed deterministic finite or countable set $\mathcal{S}$, the pointwise argument with the local data associated with each $\mathbf{u}^\dagger \in \mathcal{S}$ gives $\mathbb{P}(\mathbf{u}_k \to \mathbf{u}^\dagger) = 0$ (Appendix D.6). The event of convergence to $\mathcal{S}$ is the union of these pointwise events, so countable subadditivity yields

$$\mathbb{P}\Big(\mathbf{u}_k \text{ converges and } \lim_{k\to\infty} \mathbf{u}_k \in \mathcal{S}\Big) = \mathbb{P}\left(\bigcup_{\mathbf{u}^\dagger \in \mathcal{S}} \{\mathbf{u}_k \to \mathbf{u}^\dagger\}\right) \leq \sum_{\mathbf{u}^\dagger \in \mathcal{S}} \mathbb{P}(\mathbf{u}_k \to \mathbf{u}^\dagger) = 0. \tag{D.38}$$

This proves Theorem 2. □

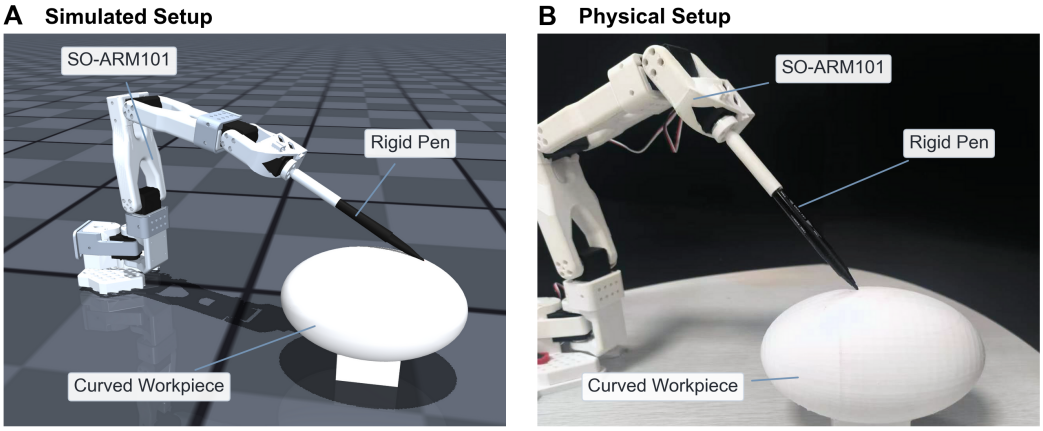

Figure E.1: The simulated and physical setups use a five-joint SO-ARM101 manipulator equipped with a rigid pen for curved-surface drawing. (A) shows the simulated setup in MuJoCo. (B) shows the physical setup.

E Experimental Details

Figure E.1 shows the simulated and physical setups used in Section 4.1.

E.1 Task geometry and initialization

The ellipsoid has semi-axes $(a_1, a_2, a_3) = (90, 64.8, 52.2)$ mm. For local pen-tip coordinates $\widetilde{y} = R^\top(y - c)$, where c and R specify the surface pose, the surface condition is

$$\phi(y) = \sum_{j=1}^{3} \frac{\widetilde{y}_j^2}{a_j^2} - 1 = 0.$$

Each simulated trial evaluates the $N = 200$-node joint command $Q_k = (q_{k,0}, \ldots, q_{k,N-1})$, where $q_{k,t} = u_k(t) \in \mathbb{R}^5$. MuJoCo forward kinematics gives $y_k(t+1) = \mathrm{FK}_{\mathrm{pen}}(q_{k,t})$. Numerical computations in both simulation and physical execution are subject to numerical noise from finite-precision arithmetic. No additional process or measurement noise is injected into the simulation. The surface constraint defines the joint set

$$\mathcal{M}_q = \left\{ q \in \mathbb{R}^5 : \phi(\mathrm{FK}_{\mathrm{pen}}(q)) = 0 \right\}.$$

The regular portion of this set, where the derivative of $\phi \circ \mathrm{FK}_{\mathrm{pen}}$ has full rank, is a manifold. Duplicating the first node gives the closed command domain

$$\mathcal{M}_{\mathrm{cl}} = \{ Q \in \mathcal{M}_q^N : q_{N-1} = q_0 \}.$$

The $n = N-1$ independent factors are $q_0, \ldots, q_{N-2}$, while $\mathbf{u}$ stacks the full N-node command. The metric induced by the node product metric in Appendix E.2 weights the first factor twice because of the duplicated endpoint.

Offline inverse kinematics followed by constraint correction generates a feasible initial command. For the main-text evaluation, the initial contour is a rounded equilateral triangle centered at the chart origin, with circumradius equal to half the maximum norm of the target chart coordinates. All four methods share this frozen initial joint trajectory and the same 200-point target. Analytical Denavit–Hartenberg kinematics (Denavit & Hartenberg, 1955) supply the position Jacobians for learning.

The anonymized supplementary material includes the simulation code, configurations, robot models, execution instructions, and eight demonstration videos.

E.2 Objective and Riemannian gradient

For $\widetilde{y} = R^\top (y - c)$ with $\widetilde{y}_3 > 0$, the normalized upper-ellipsoid chart is

$$z = \chi(y) = \left(\frac{\widetilde{y}_1}{a_1}, \frac{\widetilde{y}_2}{a_2} \right)^\top, \qquad z_1^2 + z_2^2 < 1.$$

All lengths are converted to meters before numerical evaluation. The induced metric is $G(z) = D\chi^{-1}(z)^\top D\chi^{-1}(z)$. Let $z_t = \chi(\mathrm{FK}_{\mathrm{pen}}(q_t))$ and $z_t^{\mathrm{des}} = \chi(y^{\mathrm{des}}(t+1))$. The local metric-based tracking distance and relative-increment error are

$$d_t^2 = (z_t - z_t^{\mathrm{des}})^\top G(z_t^{\mathrm{des}})(z_t - z_t^{\mathrm{des}}),$$
$$\delta_t = (z_{t+1} - z_t) - (z_{t+1}^{\mathrm{des}} - z_t^{\mathrm{des}}).$$

The two terms in the objective reported in Section 4.1 are

$$f_{\mathrm{track}}(Q) = \frac{1}{N} \sum_{t=0}^{N-1} \left(\frac{1}{2} d_t^2 + \frac{c_{\mathrm{tr}}}{3} d_t^3 \right),$$
$$f_{\mathrm{inc}}(Q) = \frac{1}{2(N-1)} \sum_{t=0}^{N-2} \delta_t^\top G(\bar{z}_t^{\mathrm{des}}) \delta_t,$$
$$f_{\mathrm{exp}}(Q) = f_{\mathrm{track}}(Q) + c_{\mathrm{sm}} f_{\mathrm{inc}}(Q),$$

where $\bar{z}_t^{\mathrm{des}} = (z_t^{\mathrm{des}} + z_{t+1}^{\mathrm{des}})/2$, $c_{\mathrm{tr}} = 100\,\mathrm{m}^{-1}$, and $c_{\mathrm{sm}} = 0.01$. The first term measures local tracking discrepancy. The second penalizes differences between learned and desired trajectory increments.

The joint-space metric is the regularized pullback

$$H(q) = B(q)^\top G(z) B(q) + \gamma I_5, \qquad B(q) = D(\chi \circ \mathrm{FK}_{\mathrm{pen}})(q), \qquad \gamma = 10^{-4}.$$

We record the gradient $g_k = (g_{k,0}, \ldots, g_{k,N-1})$ of f_{exp} on the node product $\mathcal{M}_q^N$, before tying the duplicate endpoint. Its norm is

$$\mathcal{G}_k = \|g_k\|_{Q_k} = \left(\sum_{t=0}^{N-1} g_{k,t}^\top H(q_{k,t}) g_{k,t} \right)^{1/2}.$$

The gradient on $\mathcal{M}_{\mathrm{cl}}$ is the orthogonal projection of g_k onto its tangent space, obtained by replacing both endpoint blocks with their average. Thus $\| \operatorname{grad}_{\mathcal{M}_{\mathrm{cl}}} f_{\mathrm{exp}}(Q_k) \|_{Q_k} \le \mathcal{G}_k$.

E.3 Numerical implementations

The numerical implementation follows the Riemannian update structure of Algorithm 1, combining transported momentum with nodewise adaptive scaling. Tangent projection and Newton constraint correction approximate the parallel transport and exponential-map updates considered in Section 3. Additional smoothing, step limiting, momentum restart, and backtracking (Boumal, 2023, Section 4.5) are used for numerical stabilization, with detailed procedures and settings provided in the accompanying code.

The update uses bias-uncorrected moments, one scalar second moment per node computed using the metric $H(q_t)$, and denominator $\sqrt{v_k^t + \varepsilon}$. The duplicated endpoints share the averaged gradient and tied moment states.

Backtracking and constraint correction introduce model evaluations in addition to the formal ILC trials.

Euclidean Adam-ILC applies bias-corrected Adam (Kingma & Ba, 2015) to $y_k(t+1) - y^{\mathrm{des}}(t+1)$. PD-ILC (Bristow et al., 2006) uses the Cartesian tracking error and its periodic central difference. PID computes proportional, integral, and difference corrections from the errors of the frozen nominal trajectory, updating controller states along the node index before evaluating the corrected trajectory once. The PD and PID difference terms use unit spacing in the trajectory index.

All baselines map Cartesian corrections into joint increments through a damped Jacobian inverse (Wampler, 1986) and enforce joint limits, without the proposed surface retraction. Hyperparameter selection and numerical settings are described in Appendix E.4.

E.4 Hyperparameter selection and settings

Under a fixed search budget, we evaluate 54 hyperparameter configurations per method, each under the triangular, circular, and scaled-heart initializations. This results in 162 runs per method and 648 runs in total. Each configuration is scored by the mean ratio of final to initial tracking RMSE across the three initializations. Each ILC run contains ten evaluated trials, while each PID run contains one corrected-trajectory evaluation. The final RMSE is taken from the last evaluation of each run. For each method, the configuration with the lowest score is selected, and the same parameter values and schedules are used across all three initializations. Surface deviation does not enter the selection score.

For R-ILC, the schedules are parameterized as

$$\eta_k = \eta_0 k^{-p}, \qquad \beta_k = \beta, \qquad \theta_k = 1 - c_\theta k^{-p}.$$

For Euclidean Adam-ILC, α denotes the constant learning rate. Table E.1 lists the candidate values and selected parameters for all four methods.

Table E.1: Each method evaluates 54 hyperparameter configurations over 162 runs, with the selected values used across all three initializations.

Method	Parameter	Candidate values	Selected value
R-ILC	η_0	0.016, 0.032, 0.064	0.064
	β	0.8, 0.9	0.8
	c_θ	0.02, 0.04, 0.08	0.08
	p	0.60, 0.75, 0.90	0.60
Euclidean Adam-ILC	α	0.00025, 0.0005, 0.001, 0.002, 0.004, 0.008	0.008
	β_1	0.7, 0.8, 0.9	0.7
	β_2	0.9, 0.99, 0.999	0.999
PD-ILC	L_p	0.07, 0.1, 0.13, 0.16, 0.19, 0.22, 0.24, 0.26, 0.28	0.28
	L_d	0.04, 0.05, 0.06, 0.08, 0.1, 0.12	0.04
PID	K_p	0.2125, 0.425, 0.6375, 0.85, 1.275, 1.7	0.85
	K_i	0, 0.04, 0.08	0.08
	K_d	0, 0.08, 0.16	0.16

R-ILC uses $\varepsilon = 10^{-16}$, while Euclidean Adam-ILC uses a fixed stabilizer of 10^{-8}. All three baselines use Cartesian and joint correction limits of 30 mm and 0.20 rad, respectively. The remaining numerical settings for each method are fixed across candidate configurations and provided in the accompanying code.

E.5 Evaluation details

Each initialization uses one deterministic run per method, with ten evaluated ILC trials and one corrected-trajectory evaluation for PID. The ILC metrics are recorded at Q_k before the next update. All plots and numerical summaries use raw forward-kinematics outputs without surface projection.

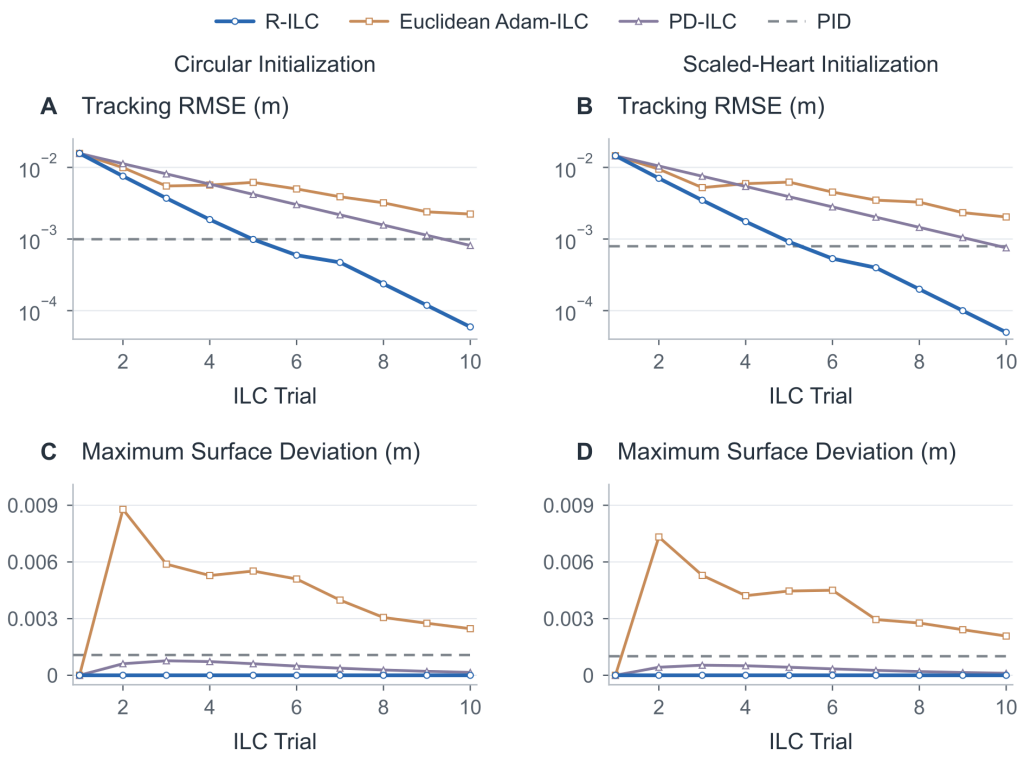

Figure E.2: R-ILC and the three baselines are compared from circular and scaled-heart initial contours. (A) shows tracking RMSE from the circular initialization. (B) shows tracking RMSE from the scaled-heart initialization. (C) shows maximum surface deviation from the circular initialization. (D) shows maximum surface deviation from the scaled-heart initialization.

The metric P_k in Section 4.1 estimates absolute normal distance to first order; it does not measure contact force or material deformation. The maximum absolute implicit residual for R-ILC across all three initializations, 200 nodes, and ten trials is 9.218×10^{-11}, below the configured tolerance of 10^{-10}.

The physical photographs in Figure 3(A) provide qualitative execution evidence and show local stroke irregularities near the upper indentation and closure region.

E.6 Additional Initializations

We additionally evaluate circular and scaled-heart initial contours for the same heart-shaped target. In the surface chart, the circle is centered at the origin with radius equal to half the maximum norm of the target coordinates; the scaled heart multiplies the target coordinates by 0.58 about the origin. For each case, all four methods share one feasible initial joint trajectory generated as in Appendix E.1. The robot, target, objective, method-specific parameters, and evaluation protocol are unchanged from Section 4.1 and Appendices E.2–E.5.

R-ILC attains final tracking RMSEs of 5.931×10^{-5} m and 4.985×10^{-5} m from the circular and scaled-heart initializations, respectively (Figure E.2(A,B)). R-ILC again achieves the lowest final RMSE, followed by PD-ILC. In both additional cases, R-ILC keeps $\max_k P_k < 2.6 \times 10^{-12}$ m throughout the ten evaluated trials, while the baselines exhibit larger surface deviations (Figure E.2(C,D)). These comparisons support the tracking and surface-feasibility advantages of R-ILC across the three tested initializations.