NONUNIFORM MONTE CARLO OPTIMISTIC POLICY ITERATION CAN CYCLE

Anonymous authors
Paper under double-blind review

ABSTRACT

Tsitsiklis (2002) proved convergence of discounted Monte Carlo optimistic policy iteration under uniform state selection and left fixed nonuniform selection open. We give a negative answer for the unnormalized asynchronous state-value recursion with a shared scalar stepsize and model-based greedy improvement. For an explicit three-state, two-action MDP, outward-rounded interval certificates establish a nonconstant attracting periodic orbit of the mean dynamics that visits six greedy policies. We then prove nonconvergence of the original stochastic recursion with bounded unbiased returns: from a fixed nonempty open set of initial values, for every $\delta \in (0, 1)$, sufficiently small Robbins–Monro stepsizes yield nonconvergence with probability at least $1 - \delta$. The obstruction is geometric: uniform updates produce radial residual contraction, whereas nonuniform scalar updates introduce anisotropic distortion that can sustain policy switching. Inverse-frequency normalization removes this distortion and recovers a known almost-sure convergence guarantee for every strictly positive fixed sampling law. The orbit persists under small parameter perturbations and is independently certified at two additional sampling laws. Continuation and paired stochastic experiments complement the analysis.

1 INTRODUCTION

Optimistic policy iteration (OPI) interleaves partial policy evaluation and greedy policy improvement. In Monte Carlo OPI, the policy can change before its value estimate has converged; a contraction argument for evaluating a fixed policy therefore does not by itself establish convergence of the control algorithm (Sutton & Barto, 2018). Tsitsiklis (2002) proved convergence under uniform state selection and left fixed nonuniform selection open.

We answer that question negatively for the unnormalized asynchronous, tabular state-value recursion with one shared scalar stepsize. Greedy improvement uses the known MDP model; Monte Carlo returns are used only for partial policy evaluation.

Let J denote the current state-value estimate, J^μ the value of a fixed greedy policy μ, and q_i the probability of selecting state i among n states. Within that policy region, the mean dynamics are

$$\dot{J} = D(J^\mu - J), \qquad D = \mathrm{diag}(q_1, \ldots, q_n). \tag{1}$$

Uniform sampling gives $D = I/n$ and radial motion toward J^μ. Nonuniform sampling can change the direction of motion, not merely its speed. Every fixed-policy mode remains stable, yet the resulting policy switches can sustain an attracting cycle.

Contributions. First, for an explicit three-state, two-action MDP, we certify a nonconstant six-policy periodic orbit using outward-rounded interval arithmetic (Proposition 2). Second, a separate hybrid trapping argument proves that the original stochastic recursion fails to converge with probability at least $1 - \delta$, for any prescribed $\delta \in (0, 1)$, from a fixed nonempty open set of initial values under sufficiently small stepsizes (Theorem 1). Third, we prove local persistence under parameter perturbations and independently certify two nearby sampling laws (Section 4).

Relation to convergence results. Inverse-frequency normalization provides a matched comparison with a known almost-sure convergence guarantee (Liu, 2021, Section 5). Recent action-value

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.

results study initial-visit or first-visit updates with different sampling rules (Oliviers & Vinnicombe, 2026a;b). Section 6 gives the detailed comparison.

2 SETUP AND THE NORMALIZATION BOUNDARY

Let $\mathcal{M} = (S, A, P, g, \alpha)$ be a finite discounted MDP with $n = |S|$ and $0 < \alpha < 1$. We minimize costs: $g(i, a)$ is the one-step cost and $P(i, a)$ is the transition row, so $P(i, a)J = \sum_j P_{ij}(a)J(j)$. Policies are stationary and deterministic, $\mu : S \to A$; write $g_\mu(i) = g(i, \mu(i))$ and $P_\mu(i, j) = P_{ij}(\mu(i))$. For $J \in \mathbb{R}^n$, define

$$(T_\mu J)(i) = g(i, \mu(i)) + \alpha P(i, \mu(i))J, \qquad J^\mu = (I - \alpha P_\mu)^{-1} g_\mu, \tag{2}$$

and let $(TJ)(i) = \min_{a \in A}\{g(i, a) + \alpha P(i, a)J\}$. The Bellman operator T has the unique fixed point J^*, and $G(J) = \{\mu : T_\mu J = TJ\}$ is the set of policies greedy with respect to J.

Fix a measurable tie-breaking map τ with $\tau(J) \in G(J)$, and write $\mu(J) = \tau(J)$ and $\mu_t = \mu(J_t)$. At iteration t, draw I_t independently of the past from a fixed law $q \in \mathbb{R}^n_{>0}$, $\sum_i q_i = 1$, and apply

$$J_{t+1} = J_t + \gamma_t e_{I_t}\big(R_t - J_t(I_t)\big), \tag{3}$$

where e_i is the ith coordinate vector and $\gamma_t > 0$ is deterministic. Only coordinate I_t changes, and the same scalar γ_t is used for every state; unequal update frequencies enter only through q. If $\mathcal{F}_t$ is the history through J_t, then $\mathbb{P}(I_t = i \mid \mathcal{F}_t) = q_i$ and, using fresh rollout randomness, $\mathbb{E}[R_t \mid \mathcal{F}_t, I_t = i] = J^{\mu_t}(i)$. Greedy improvement uses the known model; sampling enters only through the partial evaluation return R_t.

A bounded unbiased return. Given (i, μ), draw a geometric horizon L independently of the transition randomness, with $\mathbb{P}(L = m) = (1 - \alpha)\alpha^m$ for $m \geq 0$. Run the chain under μ from $S_0 = i$, and set

$$R = \frac{g(S_L, \mu(S_L))}{1 - \alpha}. \tag{4}$$

The rollout is finite almost surely. With $G_{\max} = \max_{i,a} |g(i, a)|$ and $R_{\max} = G_{\max}/(1 - \alpha)$, Appendix A verifies

$$\mathbb{E}[R \mid S_0 = i, \mu] = J^\mu(i), \qquad |R| \leq R_{\max}. \tag{5}$$

This estimator obtains bounded unbiased returns from an almost-surely finite rollout; it is a proof convenience, not a source of cycling. The return assumptions in the theorems are precisely conditional unbiasedness and boundedness. The full return $\sum_{k \geq 0} \alpha^k g(S_k, \mu(S_k))$ also satisfies them, although it is not finitely observable from one rollout. A fixed finite truncation is generally biased; Section 4 separately treats sufficiently long truncations.

Mean dynamics. Let $D = \mathrm{diag}(q_1, \ldots, q_n)$ and $f_\mu(J) = D(J^\mu - J)$. Conditional unbiasedness gives

$$\mathbb{E}[J_{t+1} - J_t \mid \mathcal{F}_t] = \gamma_t D\big(J^{\mu_t} - J_t\big). \tag{6}$$

Away from policy ties, the mean ordinary differential equation (ODE) is

$$\dot{J} = f_{\mu(J)}(J) = D\big(J^{\mu(J)} - J\big). \tag{7}$$

At ties we use the closed differential inclusion

$$\dot{J} \in H(J), \qquad H(J) = D\ \mathrm{co}\{J^\mu - J : \mu \in G(J)\}, \tag{8}$$

where co denotes convex hull. This inclusion permits all greedy choices at a tie; the stochastic recursion itself retains the fixed rule τ.

Frequency-normalized control. For comparison, give selected coordinate i a weight w_i:

$$J_{t+1} = J_t + \gamma_t w_{I_t} e_{I_t}\big(R_t - J_t(I_t)\big), \qquad W = \mathrm{diag}(w_1, \ldots, w_n). \tag{9}$$

Its mean multiplier is DW. The scalar algorithm has $W = I$; inverse- frequency normalization takes

$$w_i = \frac{1}{nq_i}. \tag{10}$$

Then $DW = I/n$: within each policy region, motion is radial toward J^μ. The factor $1/n$ matches the mean-field rate under uniform sampling. Write $q_{\min} = \min_i q_i$.

The following matched guarantee is the discounted bounded-return specialization of the frequency-corrected update in Liu (2021, Section 5); we include it to make the comparison with the scalar rule explicit.

Proposition 1 (Inverse-frequency normalization)**.** *Consider any finite discounted MDP with n states, any fixed sampling law $q_i > 0$, and any bounded return satisfying*

$$\mathbb{E}[R_t \mid \mathcal{F}_t, I_t = i] = J^{\mu_t}(i), \qquad \mu_t \in G(J_t).$$

In the weighted recursion (9)*, set*

$$w_i = \frac{1}{nq_i}.$$

Suppose the deterministic stepsizes satisfy

$$\sum_t \gamma_t = \infty, \qquad \sum_t \gamma_t^2 < \infty, \qquad 0 < \gamma_t \le nq_{\min} \quad \textit{for all } t.$$

Then, from every initial condition, $J_t \to J^*$ almost surely. In particular, for $\gamma_t = (t+N)^{-1}$ it is enough to take $N \ge 1/(nq_{\min})$.

For completeness, Appendix A gives a direct proof in the present notation by rewriting the asynchronous update as the synchronous stochastic OPI recursion of Tsitsiklis (2002) plus a bounded martingale difference.

3 Counterexample and main results

Set $n = 3$, $A = \{0, 1\}$, $\alpha = 19/20$, and

$$q = (0.02, 0.18, 0.80) = \frac{1}{50}(1, 9, 40), \qquad D = \operatorname{diag}(q).$$

The exact rational MDP data appear in Appendix B. Define the affine greedy gaps

$$d_i(J) = Q_i(1; J) - Q_i(0; J), \qquad Q_i(a; J) = g(i, a) + \alpha P(i, a)J.$$

Because costs are minimized, $\operatorname{sgn} d_i(J)$ selects action 0 when positive and action 1 when negative. The hyperplane $d_i(J) = 0$ is therefore a policy-tie surface (a *guard* in hybrid-systems terminology). Within a region where the greedy policy μ is unique, the mean dynamics follow one affine vector field; crossing a tie surface switches to the field of the new greedy policy. Thus (7) is a piecewise-affine switched, or hybrid, system. Inside policy region μ, put $z = e^{-0.02t}$; its exact flow is

$$\Phi_\mu(x, z) = J^\mu + (x - J^\mu) \odot (z, z^9, z^{40}). \tag{11}$$

Here $\odot$ denotes componentwise multiplication.

The proof has two distinct parts. Certificates C1–C4 establish an exact attracting orbit of the mean dynamics. Section 3.3 then proves that, with probability at least $1 - \delta$ under Theorem 1's conditions, the stochastic iterates remain near the orbit and visit two separated neighborhoods infinitely often.

The counterexample follows the policy sequence

$$(0, 1, 1) \to (0, 0, 1) \to (1, 0, 1) \to (1, 0, 0) \to (1, 1, 0) \to (1, 1, 1) \to (0, 1, 1). \tag{12}$$

Label these six policies $\mu_0, \dots, \mu_5$ in the order shown. The finite certificates in Appendix C verify the following proposition.

Proposition 2 (Certified attracting hybrid cycle)**.** *For the MDP in Appendix B and the update distribution $q = (0.02, 0.18, 0.80)$, the hybrid ODE* (7) *has a nonconstant locally attracting periodic orbit C following the policy sequence* (12)*. Moreover, C is a strict local attractor for the local Filippov differential inclusion* (8)*.*

Proof. C1–C4, summarized in Section 3.2, jointly prove an exact closed orbit, its intended sequence of unique greedy policies, one-way crossings without sliding, and contraction of a nonzero-radius return-section box. The contraction mapping theorem and the hybrid Poincaré-map criterion (Wendel & Ames, 2010) give local attraction for the switched flow. The no-sliding check makes the same conclusion valid for the local Filippov differential inclusion (8). □

Theorem 1 (High-probability stochastic nonconvergence)**.** *For the MDP in Appendix B, put* $R_{\max} = G_{\max}/(1-\alpha)$*. There exist a nonempty open set* $\mathcal{O}$ *in the local basin of* C *and constants* $\eta > 0$ *and* $\bar{\gamma} > 0$ *such that the following holds. For every* $\delta \in (0,1)$*, every* $J_0 \in \mathcal{O}$*, and every deterministic stepsize sequence satisfying*

$$0 < \gamma_t \le \bar{\gamma}, \qquad \sum_t \gamma_t = \infty, \qquad \sum_t \gamma_t^2 \le \frac{\delta \eta^2}{48 R_{\max}^2}, \tag{13}$$

the scalar recursion (3)*, with any conditionally unbiased return bounded by* $R_{\max}$*, fails to converge with probability at least* $1-\delta$*.*

Consequently, for every $a > 0$ *and* $\beta \in (1/2, 1]$*, the schedule* $\gamma_t = a(t+N)^{-\beta}$ *has this property for all sufficiently large* N*; in particular this includes delayed harmonic stepsizes.*

Section 3.3 gives the proof idea, and Appendix E gives the full proof. The theorem establishes failure of convergence itself, not merely convergence to a suboptimal limit.

3.1 Mechanism and suboptimality

For $\mu^* = (0,1,1)$, exact computation gives

$$J^{\mu^*} \approx (-85.8270, -92.9773, -93.9064), \qquad d(J^{\mu^*}) \approx (6.3365, -13.6208, -5.4966).$$

Thus μ^* is self-greedy, so uniqueness of the discounted Bellman fixed point gives $J^{\mu^*} = J^*$. The cycle contains $x_0 \approx (-8.5937, -0.0178, 0.4942)$ and hence $\|x_0 - J^*\|_\infty > 90$: the attractor is far from the optimal value.

To expose the mechanism, put $X = T_\mu J - J$. Within a fixed policy region,

$$\dot{X} = -(I - \alpha P_\mu) D (I - \alpha P_\mu)^{-1} X. \tag{14}$$

Writing $[D, P_\mu] = DP_\mu - P_\mu D$, the coefficient in (14) decomposes as

$$(I - \alpha P_\mu) D (I - \alpha P_\mu)^{-1} = D + \alpha [D, P_\mu](I - \alpha P_\mu)^{-1}.$$

Relative to the uniform baseline $c = 1/n$, the term $D - cI$ gives direct anisotropic scaling, whereas the commutator term gives policy-dependent residual mixing. Both disappear when $D = cI$, yielding the radial contraction $\dot{X} = -cX$ shown in Figure 1. For nonuniform D, the combined directional distortion can change which policy-tie surface is reached next.

Every fixed-policy matrix in (14) is similar to D, so each individual policy-evaluation mode is stable. Nonconvergence arises from switching: before one mode reaches its target, the trajectory crosses a tie surface. The value-space target J^μ and the residual dynamics then change, even though the residual equilibrium remains $X = 0$. Six such switches form a closed orbit, and C4 proves contraction transverse to it.

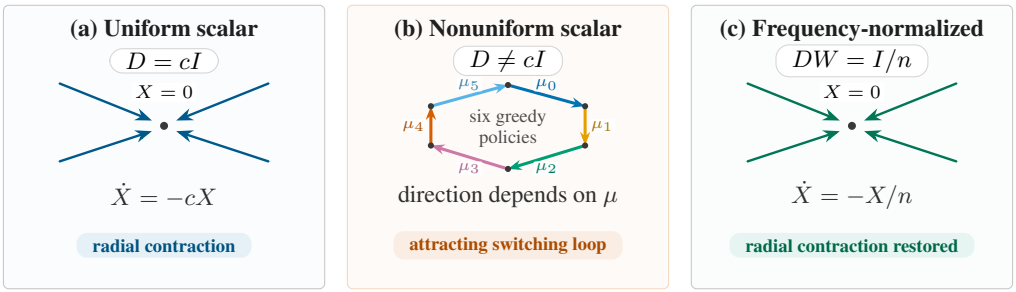


Figure 1: Schematic mean residual dynamics, $X = T_\mu J - J$. Uniform scalar updates (a) and inverse-frequency normalization (c) give radial contraction. Nonuniform scalar updates (b) change the direction of motion and can sustain a six-policy cycle; the loop indicates the policy itinerary.

The fast coordinate can track its current policy-value target while slower drift carries the trajectory to the next tie surface. Appendix B.1 gives the event data and a detailed timescale interpretation.

3.2 CERTIFICATION OVERVIEW

A high-precision numerical trajectory alone does not establish the existence or local attraction of a periodic orbit. We establish these properties by verifying four finite conditions (C1–C4) using outward-rounded interval arithmetic (Rump, 2010). Appendix C gives the event system and detailed calculations, and Table 1 summarizes the certified bounds.

Proof pipeline. For the specified MDP and sampling law, we use the candidate-orbit data in Appendix B, interpreting all displayed decimals as exact rational inputs. The C1–C4 interval checks establish the existence of a periodic orbit with the prescribed six-policy itinerary, one-way crossings, and local attraction.

C1: orbit existence. A Krawczyk inclusion is a verified interval-Newton step: when its interval image lies strictly inside the candidate box, an exact zero exists there; the companion derivative bound gives local uniqueness. Applying it on a radius-10^{-30} rational box proves the exact six-event orbit (Krawczyk, 1969; Krawczyk & Neumaier, 1986).

C2: guard signs. On each segment, the signed greedy gaps become polynomials on $[0, 1]$. A polynomial lies in the convex hull of its Bernstein coefficients, so a common coefficient sign proves that no gap changes sign anywhere on the segment (Garloff, 1986). This Bernstein-polynomial range certificate gives

k	μ_k	$\mathrm{sign}(d_1, d_2, d_3)$
0	$(0, 1, 1)$	$(+, -, -)$
1	$(0, 0, 1)$	$(+, +, -)$
2	$(1, 0, 1)$	$(-, +, -)$
3	$(1, 0, 0)$	$(-, +, +)$
4	$(1, 1, 0)$	$(-, -, +)$
5	$(1, 1, 1)$	$(-, -, -)$

Thus the orbit follows the intended greedy policies and hits the intended tie surfaces in order.

C3: one-way transversality and no sliding. At every switch, the incoming and outgoing fields have nonzero normal components with the same sign. Both fields therefore point across the tie surface in one direction, so every Filippov convex combination crosses rather than sliding along it.

C4: return-map contraction. A Poincaré return map sends one crossing of a two-dimensional section to the next crossing after a complete six-policy circuit. The cycle is its fixed point, and contraction of the map proves local attraction (Wendel & Ames, 2010). The verifier parameterizes $\Sigma = \{d_1 = 0\}$ from an exact rational anchor and, on a 4×4 subdivision of its radius 10^{-3} box, proves that the true oriented return maps the box strictly inside itself with derivative norm below one.

Table 1: Decisive margins in the nominal C1–C4 proof, rounded outward.

certificate	certified quantity	bound
C1	centered Krawczyk radius	$4.385 \times 10^{-45} < 10^{-30}$
C2	nonendpoint Bernstein margin	$3.675 \times 10^{-3} > 0$
C3	minimum outgoing crossing speed	$3.2308 > 0$
C4	$\sup \lVert D\Pi_\Sigma \rVert_\infty$	$0.623284 < 1$
C4	inward section margin	$3.767 \times 10^{-4} > 0$

Section 4 states the rigorous persistence and independent-replication results. Their complete margins are reported in Appendix C.5 so the main text can focus on what the certificate establishes.

3.3 FROM THE MEAN CYCLE TO STOCHASTIC NONCONVERGENCE

We prove stochastic nonconvergence by combining the return-map contraction in C4 with a bound on cumulative martingale noise. This is a local *lock-in* argument (Borkar, 2002; 2003; Faure & Roth,

2010; Dalal et al., 2018). The additional difficulty is that discrete updates can overshoot policy-tie surfaces. The trapping theorem below controls these crossings: from suitable initial conditions, sufficiently small stepsizes and cumulative perturbations keep the iterates near the cycle and force infinitely many visits to two separated neighborhoods. Under the stepsize conditions of Theorem 1, the required noise bound holds with probability at least $1 - \delta$.

The initial conditions lie in a *safe post-crossing flow box*: a full-dimensional set on the outgoing side of the initial return section, separated from it by a small positive margin. Following the outgoing affine flow backward maps each point into the compact section neighborhood V_*. The margin is chosen to prevent immediate recrossing under the allowed perturbations. Appendix D gives the construction and proof.

Theorem 2 (Hybrid trapping under small cumulative perturbations)**.** *Let C and its exact oriented return section satisfy C1–C4 as certified in Appendix C.4. Then there exist a compact tube U around C, two open sets $U_a, U_b \subset U$ with disjoint closures, centered at two distinct points of C, and constants $\eta > 0$ and $\bar{\gamma} > 0$ with the following property. Let*

$$Y_{t+1} = Y_t + \gamma_t\big(h(Y_t) + \xi_{t+1}\big), \tag{15}$$

where off the tie surfaces $h(J) = D(J^{\mu(J)} - J)$ for the unique greedy policy, and on a simple tie the chosen greedy policy may be either adjacent policy. Assume

$$0 < \gamma_t \le \bar{\gamma}, \qquad \sum_t \gamma_t = \infty,$$

and

$$\sup_{m \ge 0} \left\| \sum_{t=0}^{m-1} \gamma_t \xi_{t+1} \right\|_\infty \le \eta, \tag{16}$$

with the empty sum at $m = 0$ equal to zero. If Y_0 is chosen in a sufficiently small safe post-crossing outgoing-flow box over a compact section neighborhood V_* of the return-map fixed point, contained in the certified section box, then the entire path remains in U and visits both U_a and U_b infinitely often. In particular, Y_t does not converge.

Proof sketch of Theorem 1. Write the recursion as its mean field plus a martingale difference M_{t+1}. For $J_0 \in \mathcal{O}$ and $\bar{\gamma} \le 1$, convexity keeps every iterate and policy value in $[-R_{\max}, R_{\max}]^3$, giving the explicit bound $\|M_{t+1}\|_\infty \le B := 4R_{\max}$. Doob's inequality gives

$$\mathbb{P}\left(\sup_m \left\| \sum_{t<m} \gamma_t M_{t+1} \right\|_\infty > \eta \right) \le \frac{3B^2}{\eta^2} \sum_t \gamma_t^2. \tag{17}$$

Condition (13) makes this at most δ; on the complementary event, Theorem 2 directly forces infinitely many visits to two separated phases. Appendix E gives the full argument. For harmonic stepsizes it suffices symbolically that $N^{-1} \le \bar{\gamma}$ and $(N-1)^{-1} \le \delta\eta^2/(48R_{\max}^2)$. We report no numerical scalar offset because C1–C4 do not extract the additional tube and shadowing constants that determine η and $\bar{\gamma}$; $N \ge 17$ below concerns only the normalized rule.

4 Robustness

The persistence result and independent certificates below establish robustness in complementary senses: the first gives an open parameter neighborhood, while the second independently repeats every finite proof obligation at two nearby rational sampling laws.

Proposition 3 (Persistence on an open parameter set)**.** *Let $\theta_0 = (\alpha, P, g, q)$ denote the certified instance, and give the admissible parameter space the relative topology induced by $\alpha \in (0,1)$, stochastic transition rows, finite costs, and strictly positive q in the probability simplex. There is a relative open neighborhood $\mathcal{U}$ of θ_0 such that, for every $\theta \in \mathcal{U}$, the corresponding hybrid mean field has a locally unique attracting periodic orbit C_θ in a common tube around C. It follows the same six-policy itinerary, crosses every switching surface one-way, and depends continuously on θ. For every such θ, the conclusion of Theorem 1 holds with a θ-dependent nonempty open set $\mathcal{O}_\theta$ and constants $\eta_\theta, \bar{\gamma}_\theta > 0$.*

Proof. Transversality makes all six event maps C^1 in their section coordinate and parameters. The implicit-function theorem continues the fixed point of the contracting return map, while the strict guard and crossing margins preserve the itinerary and exclude sliding. The stochastic statement then follows from the same trapping argument as Theorem 1. Appendix E.2 gives the details. □

The conclusion is also unchanged for every sufficiently long ordinary unbootstrapped return $R_H = \sum_{k=0}^{H-1} \alpha^k g(S_k, \mu(S_k))$: uniformly over the eight policies, its mean differs from J^μ by at most $\alpha^H R_{\max}$, so the event-map persistence and trapping arguments apply for all sufficiently large H (with H-dependent constants). Indeed, the persistence proof uses the parameters only through the eight policy-value targets and the guard coefficients; it applies verbatim when the targets are treated as additional free parameters and the guards are held fixed.

Proposition 4 (Independent certificates at three sampling laws)**.** *Fix the MDP in Appendix B. The complete C1–C4 proof succeeds independently at each of*

$$q^- = \frac{1}{49}(1, 8, 40), \qquad q^0 = \frac{1}{50}(1, 9, 40), \qquad q^+ = \frac{1}{51}(1, 10, 40).$$

At every one of these sampling laws, the mean field has a locally attracting periodic orbit with the same six-policy itinerary. The conclusion of Theorem 1 holds at each law with law-dependent $\mathcal{O}_q, \eta_q$, and $\bar{\gamma}_q$.

Proof. The nominal law $q^0 = (0.02, 0.18, 0.80)$ is Proposition 2. At q^- and q^+, the supplementary verifier first reconstructs a high-precision event-system center and then treats its printed finite decimals as exact rationals. It independently reruns the unchanged C1–C4 obligations at each law. The reconstructed centers are only candidates; the interval inclusions supply the proofs. The stochastic conclusion follows from the same trapping and martingale argument as Theorem 1. □

Each of the three point certificates has its own positive-radius persistence neighborhood by Proposition 3; the points do not, by themselves, certify the entire line segment between q^- and q^+. Their ℓ_∞ distances from q^0 are 0.016735 and 0.016078, respectively; these spans quantify the replicated perturbations but are not a certified ball radius. Appendix C.5 records their complete margins.

5 Diagnostics

For the constructed MDP, we show how the mean-field cycle changes with the sampling law and compare the finite-horizon behavior of scalar and frequency-normalized OPI. Additional horizon and initialization studies appear in Appendix F.

5.1 Continuation toward uniform sampling

We examine how the certified six-policy orbit changes as the sampling law approaches uniformity along

$$q(\lambda) = (1 - \lambda)(1/3, 1/3, 1/3) + \lambda(0.02, 0.18, 0.80), \qquad 0 \le \lambda \le 1. \tag{18}$$

Starting from the certified orbit at $\lambda = 1$, we solve the physical-time form of the event and closure equations in Appendix C, with $D = \mathrm{diag}(q(\lambda))$, for the initial section point and six segment durations. A scaled pseudo-arclength predictor–corrector traces the periodic-orbit branch, with the policy itinerary and crossing directions checked at each computed point. The implementation is provided in `simulations/parameter_continuation.py` in the supplementary archive.

The double-precision results indicate that the return-map fixed point remains attracting at the sampled values from $\lambda = 1$ down to 0.9909, while the period increases from 33.1758 to 40.5005. Further continuation indicates a fold where the attracting and unstable branches meet, at

$$\lambda_{\mathrm{fold}} \approx 0.9908582957, \qquad q(\lambda_{\mathrm{fold}}) \approx (0.0228644, 0.1814017, 0.7957339).$$

The estimated period at the fold is 41.2417; the value 40.5005 corresponds instead to the last sampled stable grid point, $\lambda = 0.9909$. The traced six-policy branch turns at the fold rather than reaching $\lambda = 0.99$. These numerical observations concern only the connected branch traced from the certified orbit and do not exclude other invariant objects.

5.2 A SAMPLED TRAJECTORY NEAR THE CERTIFIED CYCLE

Figure 2 makes the switching mechanism concrete using one run with 200,000 updates, $\gamma_t = 0.2(t + 2000)^{-0.55}$, and seed 1. It accumulates 94.86 ODE-time units (2.86 certified periods) with no off-itinerary updates. Appendix F.1 explains why a direct harmonic-stepsize visualization has an impractical ODE-time budget.

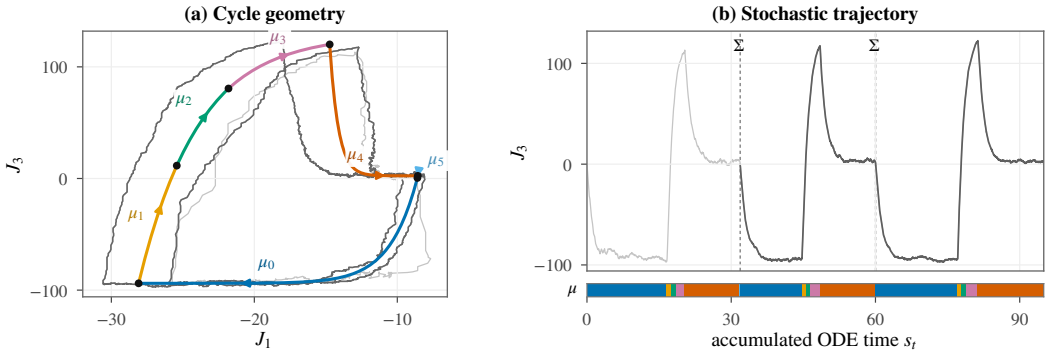


Figure 2: Finite-horizon dynamics near the certified cycle. (a) Mean-field orbit and stochastic sample path; colors denote $\mu_0, \ldots, \mu_5$. (b) Sample path in ODE time, with dashed lines marking section returns and the policy sequence shown below.

5.3 PAIRED STOCHASTIC COMPARISON

We compare scalar and frequency-normalized OPI at six sampling laws, using 200 pairs of 200,000-update runs per law with shared initial conditions and random draws. At $\lambda = 1$, post-burn-in circuit fractions are 30.0% versus 0% (Wilson 95% intervals $[0.241, 0.367]$ and $[0, 0.0188]$), with median tail errors 95.73 and 2.89, respectively. Full results appear in Appendix F.

Deterministic simulations classify $J_0 = 0$ and 86.9% of 512 Latin-hypercube starts as approaching the cycle (Appendix F.3).

6 RELATED WORK

The closest antecedent is Tsitsiklis (2002, Section 3, pp. 65–66; see also p. 72): its uniform-update proof uses scalar residual dynamics and leaves fixed nonuniform scalar updates open; it notes that per-state visit-count stepsizes would equalize nonuniform update frequencies. Liu (2021, Section 5) later proves convergence both for the adaptive coefficient $1/n_i(t)$ and for the deterministic frequency-corrected coefficient $\widehat{\gamma}_t/p(i)$. Taking $p(i) = q_i$ and $\widehat{\gamma}_t = \gamma_t/n$ gives exactly the normalized update in Section 2. Proposition 1 records its discounted bounded-return specialization and an explicit coefficient bound; neither result resolves the unnormalized scalar rule. Zoroddu (2023) extends this frequency-matched state-value line to lookahead and first-visit variants.

Earlier evidence that unnormalized updates can fail uses different regimes. The arbitrary-schedule, coordinate-stepsize construction in Bertsekas & Tsitsiklis (1996, Example 5.12) originates with Williams & Baird (1993); contraction-based enhanced policy-iteration variants provide one distinct remedy (Bertsekas & Yu, 2012). Liu (2021, Example 18) reports numerical nonconvergence under nonuniform sampling but leaves a concrete fixed-law example open. Positive MCES/OPI results instead use stochastic-shortest-path assumptions (Chen, 2018; Liu, 2021), feed-forward or acyclic structure (Wang et al., 2022; Lubars et al., 2021), sufficient lookahead (Winnicki & Srikant, 2023), or a small discount factor (Delattre & Fournier, 2026).

Recent model-free action-value work addresses a different setting. Oliviers & Vinnicombe (2026a) prove convergence of initial-visit Monte Carlo OPI under uniform updates over actions within each state and additional stepsize conditions. Oliviers & Vinnicombe (2026b) study sample-average

initial-visit and first-visit MCES with policy-dependent state-action start laws. In their first-visit construction, the 26-phase mean trajectory contracts to a suboptimal boundary point; the corresponding fully sampled behavior is investigated through simulation. Here we use a fixed i.i.d. state-selection law, certify a nonconstant mean orbit, and prove nonconvergence of the stochastic state-value recursion itself (Theorem 1).

Relative update frequencies enter asynchronous mean fields through diagonal scalings (Borkar, 1998). Differential-inclusion methods treat greedy ties (Benaïm et al., 2005; Perkins & Leslie, 2012), while lock-in and positive-probability attraction are standard stochastic-approximation tools (Borkar, 2002; 2003; Faure & Roth, 2010). Sampling weights also shape projected off-policy TD fixed points (Baird, 1995; Kolter, 2011), but the mechanisms differ: there D interacts with projection and target-policy dynamics, whereas here every tabular fixed-policy mode is stable and greedy switching creates the orbit. Thus the role of D is geometrically analogous, not an identification of the two failures. The contribution is a certified non-equilibrium attractor at the scalar nonuniform boundary and a lift from that attractor to the original stochastic recursion; this also differs from approximation-induced policy oscillation (Wagner, 2011; 2013).

7 Limitations and conclusion

We show that fixed nonuniform sampling can prevent convergence of scalar-stepsize state-value Monte Carlo OPI, even in the tabular setting with bounded unbiased returns. We certify an attracting periodic orbit of the mean dynamics and prove that the stochastic recursion can remain near it without converging. This gives a negative answer to the fixed-nonuniform state-value question posed by Tsitsiklis (2002).

The stochastic guarantee requires initialization in a nonempty open set near the cycle and sufficiently small stepsizes. We prove that the required trapping constants exist but do not compute them, so the experiments do not verify the theorem's probability bound. The three certified sampling laws and local persistence result do not characterize the global convergence boundary. Policy-dependent or time-varying sampling, unknown frequencies, and model-free action-value updates remain outside the scope of this analysis. A natural next step is to identify general conditions that rule out such switching cycles.

AI use statement

Generative AI tools (GPT-5.6 Sol) were used to assist with the literature search, manuscript editing, and the implementation of simulation and verification code. The authors take full responsibility for the final content, including all claims, code, and artifacts.

Reproducibility statement

The paper and appendices specify the MDP, sampling laws, assumptions, and certificate requirements, with full proofs in the appendices. The supplement provides the interval verifiers, simulation code, fixed seeds, results, tests, and rebuild script needed to reproduce all computational steps supporting the theorem from the specified MDP. Rigorous certificates are separated from exploratory continuation and finite-horizon simulations.

References

Leemon Baird. Residual Algorithms: Reinforcement Learning with Function Approximation. In *Machine Learning Proceedings 1995*, pp. 30–37. Elsevier, 1995.

Michel Benaïm, Josef Hofbauer, and Sylvain Sorin. Stochastic Approximations and Differential Inclusions. *SIAM Journal on Control and Optimization*, 44(1):328–348, 2005.

Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, Massachusetts, 1996.

Dimitri P. Bertsekas and Huizhen Yu. Q-Learning and Enhanced Policy Iteration in Discounted Dynamic Programming. *Mathematics of Operations Research*, 37(1):66–94, 2012. doi: 10.1287/moor.1110.0532.

Vivek S. Borkar. Asynchronous Stochastic Approximations. *SIAM Journal on Control and Optimization*, 36(3):840–851, 1998.

Vivek S. Borkar. On the Lock-In Probability of Stochastic Approximation. *Combinatorics, Probability and Computing*, 11(1):11–20, 2002. doi: 10.1017/S0963548301004953.

Vivek S. Borkar. Avoidance of Traps in Stochastic Approximation. *Systems & Control Letters*, 50 (1):1–9, 2003.

Yuanlong Chen. On the Convergence of Optimistic Policy Iteration for Stochastic Shortest Path Problem. *arXiv preprint arXiv:1808.08763*, 2018.

Gal Dalal, Gugan Thoppe, Balázs Szörényi, and Shie Mannor. Finite Sample Analysis of Two-Timescale Stochastic Approximation with Applications to Reinforcement Learning. In *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 1199–1233. PMLR, 2018.

Sylvain Delattre and Nicolas Fournier. Markov Decision Processes: On the Convergence of the Monte-Carlo First-Visit Algorithm. *SIAM Journal on Control and Optimization*, 64(3):1672–1697, 2026.

Mathieu Faure and Grégory Roth. Stochastic Approximations of Set-Valued Dynamical Systems: Convergence with Positive Probability to an Attractor. *Mathematics of Operations Research*, 35 (3):624–640, 2010.

Jürgen Garloff. Convergent Bounds for the Range of Multivariate Polynomials. In Karl Nickel (ed.), *Interval Mathematics 1985*, volume 212 of *Lecture Notes in Computer Science*, pp. 37–56. Springer, 1986.

J. Zico Kolter. The Fixed Points of Off-Policy TD. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

R. Krawczyk. Newton-Algorithmen zur Bestimmung von Nullstellen mit Fehlerschranken. *Computing*, 4(3):187–201, 1969. doi: 10.1007/BF02234767.

R. Krawczyk and Arnold Neumaier. An Improved Interval Newton Operator. *Journal of Mathematical Analysis and Applications*, 118(1):194–207, 1986.

Jun Liu. On the Convergence of Reinforcement Learning with Monte Carlo Exploring Starts. *Automatica*, 129:109693, 2021.

Joseph Lubars, Anna Winnicki, Michael Livesay, and R. Srikant. Optimistic Policy Iteration for MDPs with Acyclic Transient State Structure. *arXiv preprint arXiv:2102.00030*, 2021.

Octave Oliviers and Glenn Vinnicombe. Convergence of Monte Carlo Optimistic Policy Iteration: Beyond Uniform State-Action Updates. *arXiv preprint arXiv:2606.10580*, 2026a.

Octave Oliviers and Glenn Vinnicombe. Exploring Starts Are Not Enough: Counterexamples and a Fix for Monte Carlo Exploring Starts. *arXiv preprint arXiv:2606.15247*, 2026b.

Steven Perkins and David S. Leslie. Asynchronous Stochastic Approximation with Differential Inclusions. *Stochastic Systems*, 2(2):409–446, 2012.

Siegfried M. Rump. Verification Methods: Rigorous Results Using Floating-Point Arithmetic. *Acta Numerica*, 19:287–449, 2010.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, second edition, 2018.

John N. Tsitsiklis. On the Convergence of Optimistic Policy Iteration. *Journal of Machine Learning Research*, 3(Jul):59–72, 2002.

Paul Wagner. A Reinterpretation of the Policy Oscillation Phenomenon in Approximate Policy Iteration. In *Advances in Neural Information Processing Systems*, volume 24, pp. 2573–2581, 2011.

Paul Wagner. Optimistic Policy Iteration and Natural Actor-Critic: A Unifying View and a Non-Optimality Result. In *Advances in Neural Information Processing Systems*, volume 26, pp. 1592–1600, 2013.

Che Wang, Shuhan Yuan, Kai Shao, and Keith W. Ross. On the Convergence of the Monte Carlo Exploring Starts Algorithm for Reinforcement Learning. In *International Conference on Learning Representations*, 2022.

Eric D. B. Wendel and Aaron D. Ames. Rank Properties of Poincaré Maps for Hybrid Systems with Applications to Bipedal Walking. In *Proceedings of the 13th ACM International Conference on Hybrid Systems: Computation and Control*, pp. 151–160. ACM, 2010.

Ronald J. Williams and Leemon C. Baird, III. Analysis of Some Incremental Variants of Policy Iteration: First Steps Toward Understanding Actor-Critic Learning Systems. Technical Report NU-CCS-93-11, College of Computer Science, Northeastern University, Boston, Massachusetts, 1993.

Anna Winnicki and R. Srikant. On the Convergence of Policy Iteration-Based Reinforcement Learning with Monte Carlo Policy Evaluation. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 9852–9878. PMLR, 2023.

Giovanni Zoroddu. *On the Convergence of Three Applied Stochastic Models Related to Reflected Jump Diffusions, Fast-Slow Dynamical Systems, and Optimistic Policy Iteration*. PhD thesis, Concordia University, 2023. URL `https://spectrum.library.concordia.ca/id/eprint/993228/`.

Appendix roadmap. Appendix A collects the setup and normalization calculations. Appendix B gives the exact instance, Appendix C records its finite certificates, and Appendix D proves the hybrid trapping theorem. Appendix E then completes the stochastic and robustness arguments, Appendix F reports the diagnostics, and Appendix G maps the proof obligations to the verification artifact.

A SETUP AND NORMALIZATION PROOFS

Geometric-estimator properties. Conditioning on the random horizon and terminal state gives

$$\begin{aligned}\mathbb{E}[R \mid S_0 = i, \mu] &= \sum_{m=0}^{\infty} (1-\alpha)\alpha^m \sum_j (P_\mu^m)(i,j)\frac{g(j,\mu(j))}{1-\alpha} \\ &= \sum_{m=0}^{\infty} \alpha^m (P_\mu^m g_\mu)(i) = \big[(I-\alpha P_\mu)^{-1} g_\mu\big](i) = J^\mu(i).\end{aligned}$$

The bound follows immediately from (4).

Proof of Proposition 1. Put $\bar{\gamma}_t = \gamma_t/n$ and define

$$Z_{t+1} = q_{I_t}^{-1} e_{I_t}\big(R_t - J_t(I_t)\big) - (J^{\mu_t} - J_t).$$

Conditional unbiasedness and $\mathbb{P}(I_t = i) = q_i$ give $\mathbb{E}[Z_{t+1} \mid \mathcal{F}_t] = 0$, and the normalized asynchronous update is exactly

$$J_{t+1} = J_t + \bar{\gamma}_t\big(J^{\mu_t} - J_t + Z_{t+1}\big). \tag{19}$$

This is the synchronous stochastic OPI recursion in Proposition 1 of Tsitsiklis (2002), with scalar stepsize $\bar{\gamma}_t$.

When coordinate i is selected, its effective coefficient is $\gamma_t/(nq_i) \in (0,1]$, so that coordinate becomes a convex combination of its old value and the bounded return. Hence

$$\sup_t \|J_t\|_\infty \le \max\left\{\|J_0\|_\infty, \sup_t |R_t|\right\} < \infty \quad \text{almost surely.}$$

The finitely many J^μ are bounded, so Z_{t+1} has uniformly bounded conditional second moment. Finally, $\sum_t \bar{\gamma}_t = \infty$ and $\sum_t \bar{\gamma}_t^2 < \infty$. The cited proposition gives $J_t \to J^*$ almost surely. For harmonic stepsizes, the coefficient condition follows from $1/N \le nq_{\min}$. □

B COUNTEREXAMPLE DATA

The discount factor is

$$\alpha = 0.95 = \frac{19}{20}.$$

All decimals below are interpreted as exact rational numbers. The costs are

$$g = \begin{pmatrix} 2.8645365128 & 6.2604856687 \\ 8.8765458733 & -6.6644477357 \\ -4.4428868652 & -7.5034828348 \end{pmatrix}.$$

The transition rows are

$$\begin{aligned}P(1,0) &= (0.0083421983,\ 0.5161392971,\ 0.4755185046),\\ P(1,1) &= (0.4115295055,\ 0.3414919270,\ 0.2469785675),\\ P(2,0) &= (0.0176113658,\ 0.9549690854,\ 0.0274195488),\\ P(2,1) &= (0.3574261643,\ 0.1754143753,\ 0.4671594604),\\ P(3,0) &= (0.6791060489,\ 0.0358694514,\ 0.2850244997),\\ P(3,1) &= (0.3221209528,\ 0.3803472570,\ 0.2975317902).\end{aligned}$$

Each row sums exactly to one.

The approximate initial point and event parameters are

$$\bar{x}_0 = \begin{pmatrix} -8.59365549110427684060547352373810356279806687 \\ -0.01782901529947570479958611237140123183129413 36 \\ 0.49419938415962381038498664432023704303346133 6 \end{pmatrix}$$

and

$$\begin{aligned} \bar{z}_0 &= 0.74751889495739629806803796519241077974415863 2, \\ \bar{z}_1 &= 0.98392537687888308599954063213500499624638477 0, \\ \bar{z}_2 &= 0.97886049144597797441700548897887017524839198 9, \\ \bar{z}_3 &= 0.95724678294356402143515488654570779490148669 5, \\ \bar{z}_4 &= 0.74800543712818509760527013735869718942942261 9, \\ \bar{z}_5 &= 0.99908989163508063408586084033753455861069194 3. \end{aligned}$$

B.1 Numerical cycle anatomy

The following rounded values summarize one circuit of the certified orbit; the interval proof uses the exact inputs above.

μ_k	duration	segment start x_k	fixed-policy target J^{μ_k}
011	14.550	$(-8.59, -0.02, 0.49)$	$(-85.83, -92.98, -93.91)$
001	0.810	$(-28.09, -86.20, -93.91)$	$(137.99, 156.24, 127.10)$
101	1.068	$(-25.42, -53.30, 11.52)$	$(145.69, 158.94, 131.75)$
100	2.185	$(-21.81, -16.17, 80.60)$	$(143.73, 157.65, 128.44)$
110	14.517	$(-14.73, 40.34, 120.11)$	$(9.71, -2.78, 2.37)$
111	0.046	$(-8.57, 0.39, 2.37)$	$(-35.17, -49.05, -50.17)$

The two-dimensional nominal section derivative has eigenvalues approximately 3.44×10^{-11} and 0.507974, consistent with strong transverse contraction.

Timescale interpretation. For the certified instance, this mechanism has a simple timescale reading: $1/q_i = (50, 5.56, 1.25)$. The two long segments, μ_0 and μ_4, last 14.550 and 14.517, whereas the other four total 4.109. This separation is consistent with the fast coordinate tracking its current fixed-policy target while slower drift reaches the long-segment guards, with the short segments providing transitions between distant targets. Over the rounded period $T = 33.176$, $e^{-q_1 T} \approx 0.515$ is close to the dominant nominal return-map eigenvalue 0.508, while $e^{-q_3 T} \approx 3.0 \times 10^{-12}$ reflects near-complete fast-coordinate relaxation. Uniform q removes this timescale separation and recovers radial residual contraction.

C Detailed finite certificates

Event system. For $y = (x_0, z_0, \ldots, z_5)$ define $x_{k+1} = \Phi_{\mu_k}(x_k, z_k)$ with indices modulo six. The exit guards are

$$s_0 = 2, \quad s_1 = 1, \quad s_2 = 3, \quad s_3 = 2, \quad s_4 = 3, \quad s_5 = 1.$$

The event equations are $d_{s_k}(x_{k+1}) = 0$ for $k = 0, \ldots, 5$, together with $x_6 = x_0$. Equivalently,

$$G(y) = \begin{pmatrix} d_2(x_1) & d_1(x_2) & d_3(x_3) & d_2(x_4) & d_3(x_5) & d_1(x_6) & (x_6 - x_0)^\top \end{pmatrix}^\top = 0.$$

Because (11) uses only z, z^9, z^{40} and every printed datum is an exact rational decimal, this is a polynomial system over $\mathbb{Q}$.

C.1 Krawczyk certificate for the orbit (C1)

Let $\bar{y} = (\bar{x}_0, \bar{z}_0, \ldots, \bar{z}_5)$ and let

$$\mathcal{B} = \bar{y} + [-10^{-30}, 10^{-30}]^9.$$

The bundled verifier `certificates/verify_nominal_orbit.py` parses every MDP and orbit decimal as an exact rational number. It evaluates $G(\bar{y})$ and $DG(\bar{y})$ exactly over $\mathbb{Q}$. It constructs

a fixed 105-significant-digit finite-decimal preconditioner A from a high-precision numerical inverse and thereafter treats A as an exact point matrix. The script interval-encloses $DG(\mathcal{B})$ by forward derivative propagation with 120-digit outward-rounded decimal interval arithmetic.

The Krawczyk operator is

$$K(\bar{y}, \mathcal{B}) = \bar{y} - AG(\bar{y}) + (I - ADG(\mathcal{B}))(\mathcal{B} - \bar{y}).$$

The verifier establishes the following bounds:

$$\|G(\bar{y})\|_\infty \le 7.074240613034390477 \times 10^{-43},$$

$$\|A\|_\infty \le 12.28569267773200443, \qquad \|AG(\bar{y})\|_\infty \le 4.384319990214461726 \times 10^{-45},$$

and

$$\|I - ADG(\mathcal{B})\|_\infty \le 2.303711148043807580 \times 10^{-25}.$$

Hence every coordinate of the centered Krawczyk image has radius at most

$$4.384319990281492284 \times 10^{-45} < 10^{-30},$$

so

$$K(\bar{y}, \mathcal{B}) \subset \mathrm{int}(\mathcal{B}).$$

The Krawczyk theorem gives existence of a zero $y^* \in \mathcal{B}$. Moreover, $\|I - ADG(\mathcal{B})\|_\infty < 1$ implies, by the Neumann criterion, that $ADG(y)$ is nonsingular throughout the box. If two zeros lay in $\mathcal{B}$, the interval mean-value identity followed by multiplication by A would give a strict contraction of their difference. Thus the zero in $\mathcal{B}$ is unique.

C.2 Guard-sign certificate (C2)

For segment k, write

$$J(r) = \Phi_{\mu_k}(x_k, r), \qquad r \in [z_k, 1].$$

Each guard polynomial has the form

$$d_i(J(r)) = a + br + cr^9 + er^{40}.$$

Set

$$\sigma_{k,i} = \begin{cases} +1, & \mu_k(i) = 0, \\ -1, & \mu_k(i) = 1. \end{cases}$$

The required condition is

$$\sigma_{k,i} d_i(J(r)) \ge 0$$

throughout the segment. Substitute

$$r = z_k + (1 - z_k)u, \qquad u \in [0, 1],$$

and express the signed polynomial in degree-40 Bernstein form:

$$\sigma_{k,i} d_i(J(z_k + (1 - z_k)u)) = \sum_{m=0}^{40} b_m^{(k,i)} \binom{40}{m} u^m (1 - u)^{40-m}.$$

The Bernstein basis is nonnegative on $[0, 1]$. The verifier proves strict positivity of every nonendpoint coefficient uniformly over $\mathcal{B}$. The remaining coefficients correspond to the intended endpoint zeros: they vanish at the exact certified root y^* by $G(y^*) = 0$, not necessarily at every point of $\mathcal{B}$. Combining these interval inequalities with the exact endpoint identities proves the required signs along the entire orbit segment. The nonendpoint coefficients have the following conservative lower

bounds:

k	μ_k	guard	lower bound for non-endpoint Bernstein coefficients
0	$(0,1,1)$	d_1	5.86
0	$(0,1,1)$	d_2	0.22
0	$(0,1,1)$	d_3	0.14
1	$(0,0,1)$	d_1	0.51
1	$(0,0,1)$	d_2	0.85
1	$(0,0,1)$	d_3	11.74
2	$(1,0,1)$	d_1	0.68
2	$(1,0,1)$	d_2	20.53
2	$(1,0,1)$	d_3	0.26
3	$(1,0,0)$	d_1	19.77
3	$(1,0,0)$	d_2	0.66
3	$(1,0,0)$	d_3	0.51
4	$(1,1,0)$	d_1	0.46
4	$(1,1,0)$	d_2	10.53
4	$(1,1,0)$	d_3	0.13
5	$(1,1,1)$	d_1	0.011
5	$(1,1,1)$	d_2	17.60
5	$(1,1,1)$	d_3	0.0036

The same verifier prints the global certified lower bound

$$\min_{k,i,m \text{ non-endpoint}} b_m^{(k,i)} \geq 3.675551493546090391 \times 10^{-3}.$$

C.3 One-way transversality and no Filippov sliding (C3)

Let

$$n_k = \nabla d_{s_k}(x_{k+1}), \qquad f_k = f_{\mu_k}, \qquad f_{k+1} = f_{\mu_{k+1}},$$

where indices are read modulo six. The notation $f_{k+1}(x_{k+1})$ means the outgoing policy vector field evaluated at the same switching point x_{k+1}; it does not mean evaluation at x_{k+2}. Incoming transversality requires

$$n_k^\top f_k(x_{k+1}) \neq 0.$$

For the differential inclusion (8), we also require the outgoing vector field to cross the same switching surface in the same direction:

$$n_k^\top f_{k+1}(x_{k+1}) \neq 0, \qquad \operatorname{sgn}(n_k^\top f_k(x_{k+1})) = \operatorname{sgn}(n_k^\top f_{k+1}(x_{k+1})).$$

The verifier gives the following interval enclosures:

k	guard	$n_k^\top f_k(x_{k+1})$	$n_k^\top f_{k+1}(x_{k+1})$
0	$d_2 = 0$	$[0.5300, 0.5301]$	$[42.6155, 42.6156]$
1	$d_1 = 0$	$[-25.0820, -25.0819]$	$[-25.9101, -25.9100]$
2	$d_3 = 0$	$[9.6649, 9.6651]$	$[9.5711, 9.5713]$
3	$d_2 = 0$	$[-11.8302, -11.8300]$	$[-33.4407, -33.4406]$
4	$d_3 = 0$	$[-0.3103, -0.3101]$	$[-3.2309, -3.2308]$
5	$d_1 = 0$	$[10.0605, 10.0607]$	$[18.5810, 18.5811]$

Thus, in a sufficiently small neighborhood of each switching point, every vector in the convex hull of the adjacent fields has normal component bounded away from zero with the same sign. Hence the Filippov inclusion has no sliding branch on the switching surface near the certified orbit; all solutions cross the surface one-way.

C.4 Exact-section return-map contraction certificate (C4)

The initial and final section is the affine guard hyperplane

$$\Sigma = \{x : d_1(x) = 0\}.$$

Write

$$d_1(x) = c_0 + n_0^\top x, \qquad c_0 = g(1,1) - g(1,0),$$

with

$$n_0 = \alpha(P(1,1,\cdot) - P(1,0,\cdot)) = (0.38302794184, -0.165915001595, -0.217112940245).$$

All quantities in this subsection are interpreted as exact rationals before being embedded into outward-rounded decimal intervals.

Define the exact rational anchor $\widehat{x}_0 \in \Sigma$ by retaining the listed second and third coordinates and solving the affine guard equation for the first:

$$\widehat{x}_{0,2} = \bar{x}_{0,2}, \qquad \widehat{x}_{0,3} = \bar{x}_{0,3},$$

$$\widehat{x}_{0,1} = -\frac{c_0 + n_{0,2}\widehat{x}_{0,2} + n_{0,3}\widehat{x}_{0,3}}{n_{0,1}}.$$

Then $d_1(\widehat{x}_0) = 0$ as an exact identity over $\mathbb{Q}$. For reference, exact rational evaluation of the listed approximation gives

$$d_1(\bar{x}_0) = 1.621881434972 \times 10^{-45},$$

and the difference between the exact anchor and its first coordinate is

$$\widehat{x}_{0,1} - \bar{x}_{0,1} \approx -4.234368456726060349 \times 10^{-45}.$$

The exact construction ensures that both the domain and codomain of the certified return map lie on the same affine section.

Since $n_{0,1} \neq 0$, use the exact rational tangent basis

$$B = \begin{pmatrix} -n_{0,2}/n_{0,1} & -n_{0,3}/n_{0,1} \\ 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

The identities $n_0^\top B = 0$ hold exactly over $\mathbb{Q}$. The section is parameterized by

$$x = \widehat{x}_0 + B\begin{pmatrix} u \\ v \end{pmatrix}, \qquad |u|, |v| \le r_\Sigma, \qquad r_\Sigma = 10^{-3}.$$

Equivalently, $u = x_2 - \widehat{x}_{0,2}$ and $v = x_3 - \widehat{x}_{0,3}$. Denote this exact section box by

$$V_\Sigma(r_\Sigma) = \left\{\widehat{x}_0 + B(u,v)^\top : |u| \le r_\Sigma,\ |v| \le r_\Sigma\right\}.$$

The hyperplane Σ is crossed twice during one geometric period. At the cycle point near $\widehat{x}_0$, the time-oriented crossing is from $d_1 < 0$ to $d_1 > 0$ and the outgoing policy is μ_0. The other d_1 crossing has the opposite orientation. The map Π_Σ is therefore the oriented six-event return beginning with the outgoing branch μ_0; it is not the unoriented first intersection with the set Σ.

The verifier subdivides $[-10^{-3}, 10^{-3}]^2$ into a 4×4 grid. On each section subbox it interval-isolates the six roots of

$$\phi_k(x,r) = d_{s_k}(\Phi_{\mu_k}(x,r))$$

by interval Newton iteration. At both endpoints of every resulting root interval $R_k = [\ell_k, u_k]$, it proves strict opposite interval signs for ϕ_k, and it proves

$$0 \notin \partial_r \phi_k(X_k, R_k).$$

Hence, for every point in the section subbox, each event root exists and is unique. At the enclosed crossing box

$$Y_k = \Phi_{\mu_k}(X_k, R_k),$$

the verifier proves that the incoming and outgoing normal velocities are nonzero and have the same sign. It also proves strict signs for all nonactive guards. Degree-40 Bernstein certificates on the pre-event intervals prove that no unintended guard is hit first.

For an event with active guard normal n_k, the event derivative is enclosed using

$$D\Psi_k(x) = \left[I - \frac{f_{\mu_k}(y) n_k^\top}{n_k^\top f_{\mu_k}(y)}\right] \operatorname{diag}(r, r^9, r^{40}), \qquad y = \Phi_{\mu_k}(x,r).$$

The six event derivatives are multiplied and converted to exact section coordinates:

$$D\Pi_\Sigma(u,v) = S\, D\Pi(\widehat{x}_0 + B(u,v)^\top)B, \qquad S = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

The center residual is also computed by certified event-root isolation. In particular, the verifier starts from the degenerate section box $u = v = 0$ and isolates all six actual event roots; the printed $\bar{z}_k$ values are used only as initial guesses. It obtains

$$\Pi_\Sigma(0,0) \in \begin{pmatrix} [1.468934277679142191 \times 10^{-46}, 1.468934277679142192 \times 10^{-46}] \\ [-2.217996587438415528 \times 10^{-46}, -2.217996587438415527 \times 10^{-46}] \end{pmatrix},$$

and hence

$$\|\Pi_\Sigma(0,0)\|_\infty \le 2.217996587438415528 \times 10^{-46}.$$

The verifier produces the following global bounds over all 16 section subboxes:

quantity	certified value/bound
section identity $d_1(x) = 0$	exact over $\mathbb{Q}$
section radius r_Σ	10^{-3}
center return residual	2.2180×10^{-46}
maximum event-root interval width	2.0074×10^{-5}
$\min \lvert\partial_r \phi_k\rvert$	20.7336
$\min \lvert n_k^\top f_{\mu_k}(Y_k)\rvert$	0.310169
$\min \lvert n_k^\top f_{\mu_{k+1}}(Y_k)\rvert$	3.230812
minimum first-hit Bernstein lower bound	9.8967×10^{-10}
minimum nonactive crossing-guard lower bound	0.145575
row 1 absolute row-sum of $D\Pi_\Sigma$	0.623283830439181
row 2 absolute row-sum of $D\Pi_\Sigma$	0.439294078059216

Consequently,

$$\sup_{(u,v)\in[-10^{-3},10^{-3}]^2} \|D\Pi_\Sigma(u,v)\|_\infty \le \kappa_\Sigma := 0.623283830439181 < 1.$$

The mean-value theorem in the exact section coordinates gives

$$\begin{aligned} \|\Pi_\Sigma(u,v)\|_\infty &\le \|\Pi_\Sigma(0,0)\|_\infty + \kappa_\Sigma \|(u,v)\|_\infty \\ &\le 2.2180 \times 10^{-46} + 0.623283830439181 \times 10^{-3} \\ &< 10^{-3}. \end{aligned}$$

Thus

$$\Pi_\Sigma(V_\Sigma(10^{-3})) \subset \operatorname{int} V_\Sigma(10^{-3}),$$

and, in the certified (u,v) section coordinates, the invariant margin

$$m_\Sigma := \operatorname{dist}_\infty\big(\Pi_\Sigma([-10^{-3},10^{-3}]^2), \partial[-10^{-3},10^{-3}]^2\big)$$

satisfies

$$\begin{aligned} m_\Sigma &\ge 10^{-3} - \big(2.2180 \times 10^{-46} + 0.623283830439181 \times 10^{-3}\big) \\ &\ge 3.76716169560818 \times 10^{-4}. \end{aligned}$$

The contraction mapping theorem therefore gives a unique fixed point of the oriented return map in the section box. The C1 orbit lies in this box and follows the same certified itinerary, so it is this fixed point.

C.5 Certificates at three sampling laws

Table 2 records compact margins for the nominal C1–C4 certificate and the two independent replications in Proposition 4. Here r_K is the largest centered C1 Krawczyk radius, $b_{\min}$ is the C2 nonendpoint Bernstein lower bound, $v_{\min}^{\text{out}}$ is the C3 outgoing crossing-speed lower bound, L is the C4 section derivative infinity-norm bound, and m is the C4 inward margin. Every C4 calculation uses section radius 10^{-3}.

Table 2: Independent C1–C4 certificates at three sampling laws. The r_K and L columns are rounded-up upper bounds; the other three are rounded-down lower bounds.

q ratios	r_K	$b_{\min}$	$v_{\min}^{\text{out}}$	L	m
$(1, 8, 40)$	2.101×10^{-55}	4.660×10^{-3}	2.924	0.694409	3.055×10^{-4}
$(1, 9, 40)$	4.385×10^{-45}	3.675×10^{-3}	3.2308	0.623284	3.767×10^{-4}
$(1, 10, 40)$	2.522×10^{-55}	3.995×10^{-3}	3.484	0.635498	3.645×10^{-4}

D PROOF OF THE HYBRID TRAPPING THEOREM

The discrete recursion generally overshoots a switching hyperplane; it does not land exactly on the Poincaré section. The proof therefore applies the exact Poincaré maps only to auxiliary section representatives obtained by projecting a safely post-crossing iterate backward along the outgoing flow. This section supplies the complete construction.

Index the six switching sections cyclically. Let Σ_k be the oriented entry section of segment k, let $W_k \subset \Sigma_k$ be a compact section neighborhood, and let

$$\Psi_k : W_k \longrightarrow W_{k+1}$$

be the exact first-hit map generated by the field

$$f_k(x) = D(J^{\mu_k} - x) = Ax + b_k, \qquad A = -D.$$

The exact return is

$$\Pi = \Psi_5 \circ \cdots \circ \Psi_0.$$

Fix affine section charts $\chi_k : W_k \to \mathbb{R}^2$. At the certified initial section we use exactly the coordinates from Appendix C.4,

$$\chi_0(x) = (x_2 - \widehat{x}_{0,2}, x_3 - \widehat{x}_{0,3}),$$

so the certified coordinate return is $\Pi_\Sigma = \chi_0 \circ \Pi \circ \chi_0^{-1}$. Because the charts and their inverses are affine on compact patches, ambient and section-coordinate errors are related by uniform finite constants. Whenever a norm between points on a section is used below, it is the corresponding chart ℓ_∞ norm; ambient flow errors retain the ordinary three-dimensional ℓ_∞ norm. Orient the affine guard at Σ_{k+1} as

$$\ell_{k+1}(x) = n_{k+1}^\top x + c_{k+1},$$

with $\ell_{k+1} < 0$ on the incoming μ_k side and $\ell_{k+1} > 0$ on the outgoing μ_{k+1} side.

By the strict positive-radius first-hit, nonactive-guard, and transversality certificates, compactness gives constants

$$0 < \tau_- < \tau_+ < \infty, \qquad \nu > 0, \qquad M < \infty,$$

and pairwise disjoint crossing neighborhoods G_k such that:

1. every exact segment beginning in W_k first hits Σ_{k+1} at a time in $[\tau_-, \tau_+]$;
2. outside $G_k \cup G_{k+1}$ the unique greedy policy is μ_k;
3. in G_{k+1} the only possible greedy policies are μ_k and μ_{k+1}, and both fields cross in the positive oriented direction:
$$n_{k+1}^\top f_k(x) \geq \nu, \qquad n_{k+1}^\top f_{k+1}(x) \geq \nu;$$
4. every relevant field has sup norm at most M on the chosen compact tube.

All statements remain valid after shrinking the section and crossing neighborhoods, because the certified inequalities are strict. We also fix, for each segment, a buffered open flow tube $\mathcal{T}_k$ around the compact family of exact arcs from the outgoing face of G_k to the incoming face of G_{k+1}. The exact arc family has a positive distance δ_T from the complement of $\mathcal{T}_k$. Every use of finite-time shadowing below is understood first up to the minimum of the intended event index and the first exit from $\mathcal{T}_k$. If the latter occurred first, the shadowing bound, chosen below to be smaller than $\delta_T/2$,

would contradict the definition of that exit. Thus the tube hypothesis in the shadowing lemma is not a circular assumption.

Transversality also gives an outgoing flow box. For $p \in \Sigma_k$ and $|r|$ small, define

$$\Theta_k(p, r) = \varphi_k(r, p),$$

where φ_k is the flow of f_k. The derivative of Θ_k contains a basis of $T_p\Sigma_k$ together with $f_k(p)$; these vectors span $\mathbb{R}^3$ because $n_k^\top f_k(p) \neq 0$. The inverse function theorem, uniformly on the compact section patch, therefore yields Lipschitz maps π_k and ρ_k satisfying

$$y = \varphi_k(\rho_k(y), \pi_k(y)), \qquad \pi_k(y) \in \Sigma_k.$$

The map π_k is the projection to the section along the outgoing flow.

Lemma 1 (Local cumulative perturbation)**.** *Put*

$$\zeta_{t+1} = \gamma_t \xi_{t+1}, \qquad Z_m = \sum_{t=0}^{m-1} \zeta_{t+1}, \qquad Z_0 = 0.$$

If $\sup_m \|Z_m\|_\infty \leq \eta$*, then, for every* $a < b$*,*

$$\left\| \sum_{t=a}^{b-1} \zeta_{t+1} \right\|_\infty \leq 2\eta, \qquad \|\zeta_{t+1}\|_\infty \leq 2\eta.$$

Proof. The interval sum is $Z_b - Z_a$, so its norm is at most $\|Z_b\|_\infty + \|Z_a\|_\infty \leq 2\eta$. Taking $b = a+1$ gives the second claim. □

In particular, while the path remains in the compact tube,

$$\|Y_{t+1} - Y_t\|_\infty \leq M\gamma_t + \|\zeta_{t+1}\|_\infty \leq M\bar{\gamma} + 2\eta \leq C_{\text{step}}(\bar{\gamma} + 2\eta). \tag{20}$$

This bound will prevent a single update from jumping over a buffered crossing neighborhood.

Lemma 2 (Affine finite-time shadowing)**.** *Fix* $T < \infty$*. Suppose that, for* $a \leq t < m$*, the recursion uses one fixed field* $f_k(x) = Ax + b_k$*, remains in the compact tube, and*

$$s_j = \sum_{t=a}^{j-1} \gamma_t, \qquad s_m \leq T.$$

There are constants $C_{\text{sh}} < \infty$ *and* $\gamma_{\text{sh}} > 0$*, independent of* a *and* m*, such that, when* $\max_{a \leq t < m} \gamma_t \leq \gamma_{\text{sh}}$*,*

$$\max_{a \leq j \leq m} \|Y_j - \varphi_k(s_j, Y_a)\|_\infty \leq C_{\text{sh}}(\bar{\gamma} + 2\eta), \qquad \bar{\gamma} = \max_{a \leq t < m} \gamma_t.$$

Proof. Let $x_j = \varphi_k(s_j, Y_a)$ and $e_j = Y_j - x_j$. Uniform Taylor expansion of the exact affine flow on the compact tube gives

$$x_{j+1} = x_j + \gamma_j(Ax_j + b_k) + r_j, \qquad \|r_j\|_\infty \leq C_0\gamma_j^2.$$

Hence

$$e_{j+1} = (I + \gamma_j A)e_j + \zeta_{j+1} - r_j.$$

Define

$$\Phi_{m,j} = \prod_{\ell=j}^{m-1} (I + \gamma_\ell A), \qquad \Phi_{m,m} = I.$$

For sufficiently small γ_{sh} and $s_m \leq T$, $\|\Phi_{m,j}\|_\infty \leq C_A(T)$. Iteration gives

$$e_m = \sum_{j=a}^{m-1} \Phi_{m,j+1}\zeta_{j+1} - \sum_{j=a}^{m-1} \Phi_{m,j+1} r_j.$$

Let $Q_j = \sum_{\ell=a}^{j} \zeta_{\ell+1}$. Discrete summation by parts yields

$$\sum_{j=a}^{m-1} \Phi_{m,j+1}\zeta_{j+1} = Q_{m-1} + \sum_{j=a}^{m-2} (\Phi_{m,j+1} - \Phi_{m,j+2})Q_j.$$

By Lemma 1, $\|Q_j\|_\infty \le 2\eta$, and

$$\|\Phi_{m,j+1} - \Phi_{m,j+2}\|_\infty \le C_A(T)\|A\|_\infty \gamma_{j+1}.$$

Thus the accumulated perturbation is bounded by $C(T)2\eta$. Moreover,

$$\sum_{j=a}^{m-1} \|r_j\|_\infty \le C_0\bar{\gamma} \sum_{j=a}^{m-1} \gamma_j \le C_0 T\bar{\gamma}.$$

The same estimate applies to every intermediate index, proving the claim. □

Lemma 3 (One switching event with an exact section representative)**.** *There are constants $C_{\mathrm{ev}} < \infty$ and $\varepsilon_0 > 0$ with the following property. Put*

$$\varepsilon = \bar{\gamma} + 2\eta \le \varepsilon_0.$$

Assume that a safely post-crossing iterate can be written as

$$Y_a = \varphi_k(r, p), \qquad p \in W_k, \qquad 0 \le r \le C_{\mathrm{ev}}\varepsilon,$$

and that its entry guard satisfies

$$\ell_k(Y_a) \ge K_0\varepsilon,$$

where K_0 is a fixed sufficiently large constant. Then a finite later iterate Y_b lies safely on the outgoing side of Σ_{k+1}. Its outgoing-flow projection

$$p^+ = \pi_{k+1}(Y_b) \in \Sigma_{k+1}$$

satisfies

$$\|p^+ - \Psi_k(p)\|_\infty \le C_{\mathrm{ev}}\varepsilon, \qquad 0 \le \rho_{k+1}(Y_b) \le C_{\mathrm{ev}}\varepsilon.$$

After b, the path cannot cross back through Σ_{k+1} before leaving G_{k+1} on its outgoing side.

Proof. Let $q = \Psi_k(p)$, let $\theta_k(p)$ be the exact event time from p, and put

$$\vartheta = \theta_k(p) - r.$$

For small ε, ϑ remains uniformly positive. While the path is in the entry neighborhood G_k, both fields adjacent to Σ_k have positive oriented normal component. Hence, for every such m,

$$\begin{aligned}\ell_k(Y_m) - \ell_k(Y_a) &= \sum_{t=a}^{m-1} \gamma_t n_k^\top h(Y_t) + n_k^\top \sum_{t=a}^{m-1} \zeta_{t+1} \\ &\ge \nu \sum_{t=a}^{m-1} \gamma_t - 2\|n_k\|_1\eta.\end{aligned}$$

Thus $\ell_k(Y_m) > 0$ if $K_0 > \|n_k\|_1$; the process cannot recross the entry section. After leaving G_k, the certified positive entry-guard margin and Lemma 2 preserve this sign.

Extend the incoming affine field f_k through Σ_{k+1} only for the comparison trajectory. Uniform transversality gives a fixed time window about ϑ in which, after decreasing ν if necessary,

$$\ell_{k+1}(\varphi_k(\vartheta + u, Y_a)) \ge \nu u, \qquad \ell_{k+1}(\varphi_k(\vartheta - u, Y_a)) \le -\nu u$$

for $u \ge 0$ in that window. Let

$$c = \inf\{m \ge a : \ell_{k+1}(Y_m) \ge 0\}.$$

Before c, the target guard is negative; the certified nonactive-guard margins exclude every other switch, so every update producing $Y_a, \dots, Y_c$ uses f_k.

For completeness, put $s_j = \sum_{t=a}^{j-1} \gamma_t$ and let $E = C_{\mathrm{sh}}\varepsilon$ be the shadowing bound. Since the guard is affine, its shadowing error is at most $\|n_{k+1}\|_1 E$. The preceding two linear guard estimates show that no grid point with

$$s_j \le \vartheta - \frac{2\|n_{k+1}\|_1 E}{\nu}$$

can have nonnegative target guard. Conversely, if no crossing had occurred by the first grid point satisfying

$$s_j \ge \vartheta + \frac{2\|n_{k+1}\|_1 E}{\nu},$$

then fixed-field shadowing would make the target guard at that point strictly positive, a contradiction. The possible overshoot of this grid point is at most $\bar{\gamma} \le \varepsilon$. Therefore

$$\left| \sum_{t=a}^{c-1} \gamma_t - \vartheta \right| \le C_{\mathrm{hit}}\varepsilon,$$

and boundedness of the exact flow gives

$$\|Y_c - q\|_\infty \le C_c\varepsilon. \tag{21}$$

This is an overshoot estimate; it does not assert that Y_c lies on the section.

We now prove, without a bootstrap assumption, that the path reaches a safe outgoing level before it can leave the crossing neighborhood. Since the compact event set $\Psi_k(W_k)$ lies in the open set G_{k+1}, choose $\delta_G > 0$ such that every $q \in \Psi_k(W_k)$ has distance at least δ_G from G_{k+1}^c. Fix

$$a_\varepsilon = K\varepsilon$$

with

$$K > \max\{K_0, 2\max_j \|n_j\|_1\},$$

and define the stopped index

$$\sigma = \inf\{m \ge c : \ell_{k+1}(Y_m) \ge a_\varepsilon \text{ or } Y_m \notin G_{k+1}\}.$$

The index is finite: while neither stopping condition holds, only the two adjacent policies are possible and both have normal velocity at least ν; the divergence of the remaining stepsize sum would otherwise force the guard above a_ε.

For $c \le m < \sigma$, the same normal-drift calculation gives

$$\ell_{k+1}(Y_m) - \ell_{k+1}(Y_c) \ge \nu \sum_{t=c}^{m-1} \gamma_t - 2\|n_{k+1}\|_1\eta.$$

Because $\ell_{k+1}(Y_c) \ge 0$ and $\ell_{k+1}(Y_m) < K\varepsilon$ before stopping,

$$\sum_{t=c}^{m-1} \gamma_t \le \frac{K\varepsilon + 2\|n_{k+1}\|_1\eta}{\nu} \le C_T\varepsilon. \tag{22}$$

At the stopping step one additional increment contributes at most $\bar{\gamma} \le \varepsilon$. The field bound, the local cumulative-noise bound, (21), and (22) therefore imply

$$\|Y_\sigma - q\|_\infty \le C_G\varepsilon. \tag{23}$$

Choose ε_0 so that $C_G\varepsilon_0 < \delta_G$. Then the second stopping alternative $Y_\sigma \notin G_{k+1}$ is impossible. Hence $b := \sigma$ reaches the safe level inside G_{k+1}, after only $O(\varepsilon)$ additional ODE time, and

$$\|Y_b - q\|_\infty \le C_G\varepsilon. \tag{24}$$

The outgoing flow-box coordinates are defined at Y_b. Their uniform Lipschitz bounds and (24) imply

$$\|\pi_{k+1}(Y_b) - q\|_\infty \le C_{\mathrm{ev}}\varepsilon, \qquad |\rho_{k+1}(Y_b)| \le C_{\mathrm{ev}}\varepsilon.$$

After shrinking the flow box, the sign of ρ_{k+1} agrees with the sign of the oriented guard. Since $\ell_{k+1}(Y_b) > 0$, this improves to $0 \le \rho_{k+1}(Y_b) \le C_{\mathrm{ev}}\varepsilon$.

Finally, for every later m before leaving G_{k+1},

$$\ell_{k+1}(Y_m) \geq K\varepsilon + \nu \sum_{t=b}^{m-1} \gamma_t - 2\|n_{k+1}\|_1\eta \geq (K - \|n_{k+1}\|_1)\varepsilon > 0.$$

Thus no recrossing is possible. On this positive side the greedy policy is μ_{k+1}. Choosing G_{k+1} as a buffered outgoing flow-box slab and using Lemma 2 shows that the path exits through the outgoing face into the compact part of the intended μ_{k+1} segment. The strict guard margins then keep it in that policy region until the next crossing neighborhood. □

Proof of Theorem 2. Put

$$\varepsilon = \bar{\gamma} + 2\eta.$$

Let

$$\mathcal{V} = [-10^{-3}, 10^{-3}]^2$$

be the full certified coordinate box, and let u^* be the exact fixed point of Π_Σ. Choose $r_* > 0$ so small that

$$\mathcal{V}_* := \{u : \|u - u^*\|_\infty \leq r_*\} \subset \operatorname{int} \mathcal{V},$$

all six exact partial-orbit families issued from $V_* := \chi_0^{-1}(\mathcal{V}_*)$ lie in the selected local tubes, and two chosen interior phases on the periodic orbit retain disjoint neighborhoods throughout those families. Such an r_* exists by continuity of the finite partial flow/event compositions and because the orbit is nonconstant. The certified contraction gives

$$\|\Pi_\Sigma(u) - u^*\|_\infty \leq \kappa_\Sigma r_*, \qquad u \in \mathcal{V}_*,$$

so the exact return has inward coordinate margin at least

$$m_* := (1 - \kappa_\Sigma)r_* > 0$$

relative to $\partial\mathcal{V}_*$.

The safe initial flow box over V_* is nonempty. Indeed, in outgoing flow coordinates at Σ_0, uniform transversality gives

$$\ell_0(\varphi_0(r, p)) \geq \tfrac{1}{2}\nu r$$

for all p in V_* and all sufficiently small $r \geq 0$. Taking r in a short interval of order $2K_0\varepsilon/\nu$ gives both $\ell_0(Y_0) \geq K_0\varepsilon$ and $\rho_0(Y_0) = O(\varepsilon)$.

At every completed event, apply the outgoing-flow projection from Lemma 3. In one circuit this produces auxiliary section representatives

$$p_0, p_1, \ldots, p_6, \qquad p_k \in \Sigma_k,$$

although the actual Euler iterates need not lie on any section. By the uniform Lipschitz bounds of the affine section charts, enlarge C_{ev} if necessary so that each segment satisfies

$$\|\chi_{k+1}(p_{k+1}) - \chi_{k+1}(\Psi_k(p_k))\|_\infty \leq C_{\mathrm{ev}}\varepsilon. \tag{25}$$

The coordinate versions of the exact partial maps are Lipschitz on their compact domains. Iterating (25) through six events gives a finite constant C_R such that, with $u_0 = \chi_0(p_0)$ and $u_1 = \chi_0(p_6)$,

$$\|u_1 - \Pi_\Sigma(u_0)\|_\infty \leq C_R\varepsilon. \tag{26}$$

The left side is the projected return generated by that particular circuit; there is no claim that the raw Euler path itself lands on the section or that one fixed deterministic approximate return map governs all circuits.

Choose $\bar{\gamma}$ and η so small that all local constructions apply, all intermediate representatives stay in their compact section neighborhoods, and

$$C_R(\bar{\gamma} + 2\eta) < \frac{1}{2}m_*. \tag{27}$$

If $u_0 \in \mathcal{V}_*$, then (26)–(27) imply $u_1 \in \mathcal{V}_*$, equivalently $p_6 \in V_*$. Induction therefore keeps every full-cycle representative in V_*, while Lemma 3 and Lemma 2 keep all intervening iterates in a fixed compact tube U.

The contraction gives the sharper bound. Let $u_j = \chi_0(p_{6j})$ be the coordinate representative after the jth complete circuit. Then

$$\begin{aligned}\|u_{j+1} - u^*\|_\infty &\le \|u_{j+1} - \Pi_\Sigma(u_j)\|_\infty + \|\Pi_\Sigma(u_j) - \Pi_\Sigma(u^*)\|_\infty \\ &\le C_R\varepsilon + \kappa_\Sigma\|u_j - u^*\|_\infty.\end{aligned}$$

Hence

$$\|u_j - u^*\|_\infty \le \kappa_\Sigma^j\|u_0 - u^*\|_\infty + \frac{C_R\varepsilon}{1-\kappa_\Sigma}. \tag{28}$$

Every exact segment time is at most τ_+, and the safe-crossing adjustment uses only $O(\varepsilon)$ additional ODE time. Since every tail of $\sum_t \gamma_t$ diverges, every next safe post-crossing index is finite. Thus infinitely many circuits are completed. More precisely, the ODE time between two consecutive safe representatives is

$$\theta_k(p) - \rho_k(Y_a) + O(\varepsilon) \ge \tau_- - C_*\varepsilon.$$

Choose ε so small that this is at least $\tau_-/2$. Hence no two successive intended events can collapse into zero ODE time. The certified first-hit order, the pairwise disjoint buffered crossing neighborhoods, and (20) also make a one-step jump over an intervening guard impossible. This rules out skipped guards and Zeno accumulation.

Let z_a, z_b be the two separated interior orbit points used when r_* was chosen. By that choice, the exact partial trajectories issued from V_* pass through disjoint fixed neighborhoods of z_a and z_b. Shrink the safe initial flow box and ε once more so that affine shadowing and the ODE-time mesh bound are smaller than half of the remaining separation margin. Every complete circuit then contains an iterate in each of two open balls U_a and U_b with disjoint closures. Both are visited infinitely often, so Y_t cannot converge. □

E Stochastic nonconvergence and robustness proofs

E.1 Stochastic nonconvergence

Proof of Theorem 1. Put $R_{\max} = G_{\max}/(1-\alpha)$. Direct evaluation of the exact instance shows that the certified cycle lies strictly inside $[-R_{\max}, R_{\max}]^3$. Choose a nonempty open safe post-crossing flow box $\mathcal{O}$ supplied by Theorem 2, shrinking it so that its closure lies in this cube, and replace the theorem's $\bar\gamma$ by $\min\{\bar\gamma, 1\}$. Fix any $J_0 \in \mathcal{O}$ and write

$$J_{t+1} = J_t + \gamma_t\big(h(J_t) + M_{t+1}\big),$$

where the fixed measurable greedy tie-breaking rule is used in h and

$$h(J_t) = D(J^{\mu_t} - J_t), \qquad M_{t+1} = e_{I_t}\big(R_t - J_t(I_t)\big) - D(J^{\mu_t} - J_t).$$

By (6), $\mathbb{E}[M_{t+1} \mid \mathcal{F}_t] = 0$.

Every policy value satisfies $\|J^\mu\|_\infty \le R_{\max}$. Moreover, if J_t lies in the cube, the selected coordinate update is $(1-\gamma_t)J_t(I_t) + \gamma_t R_t$, a convex combination of two points in $[-R_{\max}, R_{\max}]$; the other coordinates do not change. Induction therefore keeps every J_t in the cube. Consequently,

$$\|M_{t+1}\|_\infty \le 4R_{\max} =: B \qquad \text{almost surely.} \tag{29}$$

Let η be supplied by Theorem 2. For each coordinate i, let

$$S_m(i) = \sum_{t=0}^{m-1} \gamma_t M_{t+1}(i), \qquad S_0(i) = 0.$$

Then $(S_m(i))_{m\ge 0}$ is an L^2 martingale, and orthogonality of its increments gives

$$\mathbb{E}[S_m(i)^2] \le B^2 \sum_{t=0}^{\infty} \gamma_t^2.$$

Doob's maximal inequality and a union bound over the three coordinates yield

$$\mathbb{P}\left(\sup_{m\ge 0}\|S_m\|_\infty > \eta\right) \le \frac{3B^2}{\eta^2}\sum_{t=0}^{\infty}\gamma_t^2.$$

Because $B = 4R_{\max}$, condition (13) makes the right-hand side at most δ. Hence the event

$$\mathcal{E} = \left\{ \sup_{m \geq 0} \left\| \sum_{t=0}^{m-1} \gamma_t M_{t+1} \right\|_\infty \leq \eta \right\}$$

has probability at least $1 - \delta$.

On $\mathcal{E}$, the original recursion directly satisfies Theorem 2; it remains in the trapping tube and visits U_a and U_b infinitely often, so J_t cannot converge. The same probability bound holds for each $J_0 \in \mathcal{O}$, with the same constants.

Finally, for $\gamma_t = a(t+N)^{-\beta}$ with $\beta \in (1/2, 1]$, the sum of the stepsizes diverges, while

$$\sup_t \gamma_t = aN^{-\beta} \longrightarrow 0, \qquad \sum_{t=0}^{\infty} \gamma_t^2 \leq a^2 \left(N^{-2\beta} + \frac{N^{1-2\beta}}{2\beta - 1} \right) \longrightarrow 0.$$

Thus both smallness requirements hold for all sufficiently large N. □

E.2 Open-set persistence

Proof of Proposition 3. For each fixed policy, both $J_\theta^\mu = (I - \alpha P_\mu)^{-1} g_\mu$ and the affine flow

$$\Phi_{\mu,\theta}(t, x) = J_\theta^\mu + \exp(-D_\theta t)(x - J_\theta^\mu)$$

are continuously differentiable in (t, x, θ) near the certified instance. Certificate C3 gives a nonzero time derivative of the active guard at every event. The implicit-function theorem therefore makes each local first-event time, and hence each event map, continuously differentiable in its incoming section coordinate and in θ. The strict C2 guard signs, nonactive-guard separation, and same-direction C3 crossings persist after shrinking the parameter neighborhood. Thus the composition of the six event maps defines a C^1 local return map Π_θ with the same itinerary and no Filippov sliding branch.

At the nominal fixed point, C4 gives $\|D\Pi_{\theta_0}\|_\infty < 1$. Hence $I - D\Pi_{\theta_0}$ is nonsingular. Applying the implicit-function theorem to $\Pi_\theta(x) - x = 0$ yields a locally unique continuous fixed-point branch x_θ. By continuity, the spectrum of $D\Pi_\theta(x_\theta)$ remains strictly inside the unit disk after another shrinkage, so the associated periodic orbit is locally attracting. The strict crossing conditions also retain a common compact local tube after shrinkage.

For each fixed $\theta \in \mathcal{U}$, the geometric-horizon return remains bounded and unbiased, and Theorem 2 applies to C_θ. Repeating the preceding martingale argument gives the conclusion of Theorem 1 with θ-dependent trapping and stepsize constants. □

F Additional stochastic diagnostics

F.1 Stepsize and ODE-time budget

Theorem 1 includes harmonic stepsizes $\gamma_t = 1/(t+N)$, but this schedule is poor for short simulations. Its ODE time by iteration m is approximately

$$\sum_{t=0}^{m} \frac{1}{t+N} \approx \log \frac{m+N}{N}.$$

Since the certified period is about 33.18, one period with large N needs roughly $m \approx N(e^{33.18} - 1)$ updates. The qualitative illustration therefore uses the still square-summable schedule $\gamma_t = 0.2(t + 2000)^{-0.55}$. Its 200,000-step release run accumulates 94.8585 ODE-time units and repeats the six-policy order.

F.2 Complete paired stochastic comparison

We compare scalar and frequency-normalized OPI at $\lambda \in \{0, 0.5, 0.9, 0.99, 0.995, 1\}$. At each sampling law, 200 pairs of 200,000-update runs share initial conditions and random draws, differing only in the weights $w_i = 1$ versus $w_i = 1/(3q_i)$. Table 3 reports post-burn-in circuit fractions and tail medians of optimal-value error and Bellman residual.

Table 3: Paired finite-horizon stochastic comparison. Here $\widehat{p}_{\mathrm{circ}}$ is the fraction with a post-burn-in complete circuit, $E_* = \|J - J^*\|_\infty$, and $B = \|TJ - J\|_\infty$; E_* and B are tail medians as described in Section 5.

	$\widehat{p}_{\mathrm{circ}}$		E_*		B	
λ	scalar	normalized	scalar	normalized	scalar	normalized
0	0.000	0.000	1.237	1.237	1.116	1.116
0.5	0.000	0.000	1.310	1.417	1.168	1.332
0.9	0.000	0.000	1.438	2.082	1.292	1.959
0.99	0.330	0.000	96.163	2.621	26.020	2.531
0.995	0.330	0.000	92.676	2.759	39.197	2.711
1	0.300	0.000	95.726	2.893	33.731	2.773

F.3 Broad-start basin and stochastic check

We construct the initialization box by taking the smallest axis-aligned box containing a dense reconstruction of the certified cycle and J^*, then extending each coordinate interval by 10% of its width at both ends:

$$[-93.553, -0.844] \times [-106.309, 53.677] \times [-115.308, 141.506].$$

A seeded 512-point Latin hypercube was integrated to ODE time 400 with Euler meshes 0.01 and 0.005. A cycle label requires an intended circuit begun in the final 80 time units and final ℓ_∞ distance at most one from the certified cycle; an optimal label requires both optimal-value error and Bellman residual at most 0.5. The two meshes agree on 507 points: 445 cycle and 62 optimal; they also agree that $J_0 = 0$ approaches the cycle. The remaining five starts are left unclassified because the two meshes give different labels.

Using the same accelerated schedule and 200,000-update horizon as the main comparison, we then ran 200 paired replicates from zero and 200 from a fresh Latin hypercube in the same box. Pair members share the initial point and all random draws. The circuit criterion is deliberately strict:

Table 4: Broad-start finite-horizon circuit fractions. Parentheses give Wilson 95% intervals; E_* and B are median final optimal-value error and Bellman residual.

initialization	rule	$\widehat{p}_{\mathrm{circ}}$	E_*	B
$J_0 = 0$	scalar	0.175 $(0.129, 0.234)$	95.64	45.85
$J_0 = 0$	normalized	0 $(0, 0.0188)$	2.54	2.62
Latin hypercube	scalar	0.275 $(0.218, 0.341)$	95.53	29.16
Latin hypercube	normalized	0 $(0, 0.0188)$	3.24	3.14

noise can disrupt the exact six-event order even while values remain far from optimum. At the final iterate, $E_* > 50$ for 100% of zero-start and 90% of Latin-hypercube scalar runs, versus 0% for both normalized groups. These fractions describe only this specified start distribution, noise law, and finite horizon; they are neither basin-volume estimates nor estimates of Theorem 1's infinite-time probability.

F.4 Subthreshold long-horizon diagnosis

To reconcile circuits at $\lambda = 0.99$ with the deterministic fold, we ran 100 paired replicates at 200,000 and 1,000,000 updates with a shared random prefix. A late circuit is a complete intended six-switch sequence beginning in the final 40 ODE-time units. At the two horizons, the scalar late-circuit fractions are 0.13 (Wilson interval $[0.0776, 0.210]$) and 0.15 ($[0.0931, 0.233]$). The normalized fraction is zero at both horizons, with Wilson upper endpoint 0.037. Scalar median tail error decreases from 74.96 to 70.75 and its Bellman residual from 32.04 to 24.30; the normalized pairs decrease from $(2.77, 2.79)$ to $(1.84, 1.82)$.

Thus noisy finite-step paths can exhibit long switching transients below the fold even though no point of the traced deterministic six-policy branch reaches $\lambda = 0.99$. The experiment proves neither eventual convergence nor infinite-time recurrence there.

F.5 Exact-harmonic offset, horizon, and initialization sensitivity

We ran $\gamma_t = (t+N)^{-1}$ at $N \in \{20, 200, 2000\}$ and horizons $10^4, 10^5, 5\times10^5$, with 40 paired replicates from each of three initialization families: the outgoing certified section, all six orbit phases, and perturbations of J^*. Every N satisfies the explicit $N \geq 17$ condition in Proposition 1. At 500,000 updates, the corresponding ODE times are only 10.1519, 7.8269, and 5.5257, all less than one third of a certified period.

Table 5: Exact-harmonic sensitivity at 500,000 updates (40 replicates per rule, initialization, and N).

N	initialization	ODE time	E_*		B	
			scalar	normalized	scalar	normalized
20	section	10.152	63.510	9.760	54.833	2.817
20	phase	10.152	97.133	18.301	20.190	3.180
20	near J^*	10.152	2.270	3.850	1.945	1.788
200	section	7.827	66.284	8.597	54.602	1.771
200	phase	7.827	102.255	14.166	21.589	3.635
200	near J^*	7.827	1.003	1.452	1.029	1.231
2000	section	5.526	69.076	15.610	51.219	1.971
2000	phase	5.526	108.994	25.814	28.860	7.478
2000	near J^*	5.526	0.660	0.863	0.486	0.879

For section initialization at $N = 200$, normalized median error falls from 30.62 to 14.28 to 8.60 as the horizon increases, whereas scalar error moves from 71.64 to 68.46 to 66.28. No run completes six intended switches because the ODE-time budget is too short.

G Verification artifact and proof boundary

The supplementary archive contains rigorous verifiers and their certificate outputs, diagnostic programs and generated summaries, and reproducibility metadata.

Proof scope and claim dependency. The nominal finite certificates establish the orbit, guard validity, one-way transversality, exact oriented hybrid return map, strict contraction, and inward margin required by Proposition 2. Appendix D uses these certified inequalities to prove Theorem 2 analytically, and Appendix E combines that theorem with the stochastic lock-in argument to prove Theorem 1. The nominal and nearby-law verifiers jointly establish Proposition 4. The computations in Section 5 and Appendix F are descriptive and have no proof status.

The analytic argument in Appendix D supplies the outgoing-flow projection, the summation-by-parts shadowing estimate, the safe-level/no-recrossing construction, and the per-circuit projected-return bound. The proof uses only finite local constants whose existence follows from compactness and the strict certificate inequalities; it then chooses $\bar{\gamma}$ and η sufficiently small. All numerical proof obligations appear in the reported finite certificates; the remaining steps are analytic.

Archive contents. The archive is organized by role. The `certificates/` folder contains the three rigorous verifiers, their JSON and log outputs, and certificate tests. The `checks/` folder contains the independent non-interval return-map diagnostic, while `simulations/` contains the descriptive studies and their data. The archive root contains the documentation, integrity manifest, and end-to-end `run_all.sh`.

Program-to-claim map. Within `certificates/`, the three rigorous entry points correspond to the earlier certificate sections as follows.

- `verify_nominal_orbit.py` verifies the orbit-side obligations for Proposition 2 (conditions C1–C3 in Appendices C.1–C.3): Krawczyk orbit inclusion, Bernstein guard signs, one-way transversality and guard separation, and the exact optimality check.
- `verify_nominal_return_map.py` verifies the return-map obligations for Proposition 2 (condition C4 in Appendix C.4): exact section identities, certified event isolation, the interval return-map derivative, contraction, and inward mapping.
- `verify_nearby_sampling_laws.py` verifies the two additional cases in Proposition 4 by rerunning all four conditions at the nearby sampling laws; together with the nominal certificates, these results establish the proposition. Full results appear in Appendix C.5.

Supporting checks. The programs `test_certificate_core.py` and `check_decimal_context_independence.py`, also in `certificates/`, test the exact interval implementation and its independence from the ambient decimal context. The separate `checks/check_high_precision_return_map.py` is a non-interval numerical cross-check and has no proof status.

Reproduction. Running `./run_all.sh` from the archive root regenerates the certificates, supporting checks, descriptive studies, and sampled-cycle figure. The pinned environment and expected outputs are documented in `REPRODUCIBILITY.md`, and `SHA256SUMS` records archive integrity.