A Contractive Bellman Update Can Solve the Wrong Bellman Equation

Anonymous authors
Paper under double-blind review

Abstract

A contractive Bellman update can have a unique fixed point that recommends the wrong action. We study product eligibility traces where one outcome determines both a temporal-difference error and the coefficient scaling the inherited trace. In finite discounted models with deterministic state–action rewards, the fixed point combines an environmental first transition with an induced continuation kernel. A model can preserve every action gap on its own, yet produce a strict policy cycle through this mixed use. For fixed, reward- and policy-independent continuation kernels, a finite linear test characterizes preservation of every action comparison across rewards and stationary policies. The test permits state-dependent value errors and yields a closest compatible coefficient with unchanged conditional means. Without knowing the transition matrix, we also construct policy-dependent coefficients that can retain outcome dependence while preserving one selected comparison. The induced gap is a strictly positive multiple of the true gap, allowing exact pairwise policy improvement after re-evaluation. A learned-likelihood study connects the failure mechanism to an implemented update: more independent fitting data can make the wrong population ordering more likely. Values need not be exact to support correct policy comparisons.

1 Introduction

A policy evaluator may converge reliably and still recommend a worse action. The difficulty is not always approximation error. It can arise from the order in which an update uses a sample. When one realized outcome determines both a temporal-difference (TD) error and its coefficient, averaging their product need not preserve the intended Bellman equation. A contraction then guarantees a unique answer, but not the answer needed for policy improvement.

SlipperyBridge makes this change concrete (Figure 1). The bridge has three crossings, each succeeding independently with probability $p = 0.8$. Completing all three yields $+1$; a slip yields -0.2. A safe action pays 0.55. After a deterministic zero-reward launch, all payoffs arrive at one deadline. Factoring out the common discount factor, the bridge is worth 0.4144. At each crossing, the realized outcome's likelihood rescales the inherited eligibility trace before the current TD error is applied. This induces success probability $\rho(p) = p^2(3 - 2p)$. At $p = 0.8$, the bridge is evaluated at 0.6632 and preferred to the safe action. The update has not diverged. It evaluates a different bridge.

With deterministic state–action rewards, transition-dependent product traces follow the true environment for the first step and an induced transition kernel thereafter. The reward function is unchanged. This differs from using the induced model for every transition: *a model can preserve every action gap on its own, yet reverse a policy update when its values follow a real first step.* We exhibit a strict policy cycle even when every evaluation is solved exactly. The reverse separation also holds: a model can make incorrect comparisons on its own but preserve every comparison after the real first step (Table 2).

Which continuation values remain suitable for policy improvement? Requiring every value to be correct is stronger than necessary. Actions are compared through their different successor distributions. A value error matters only if those distributions assign different expectations to it. Some errors therefore leave every comparison unchanged, even when they vary across states.

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.

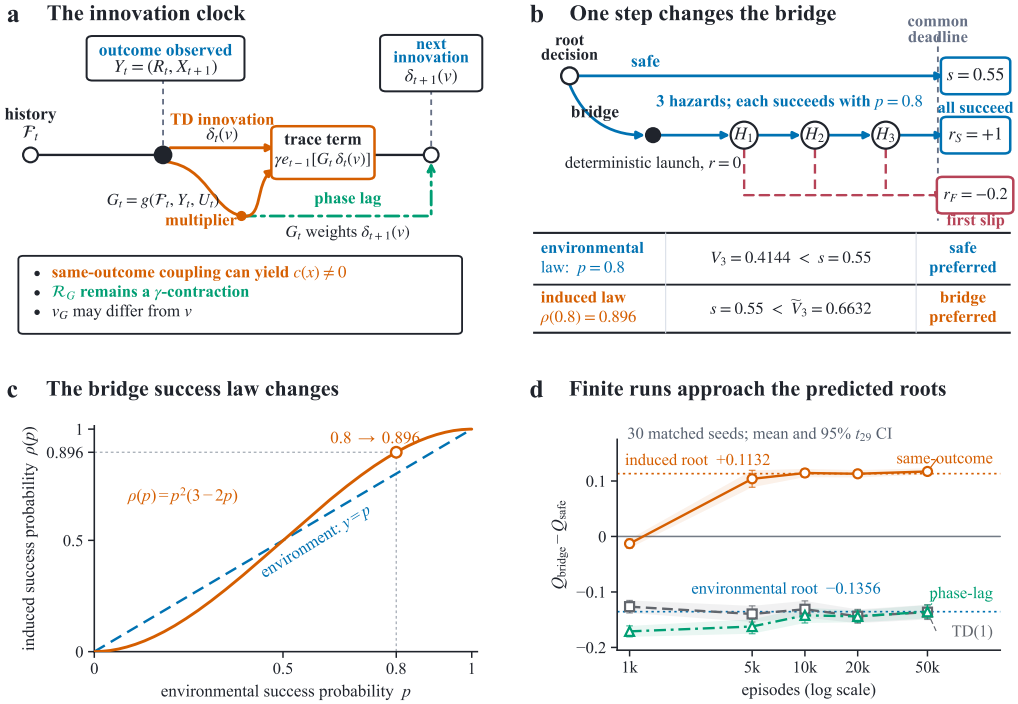

Figure 1: **One outcome, two jobs.** (a) Y_t determines both G_t and the current TD error. Same-outcome timing applies G_t to the inherited trace before this error is used; phase lag delays that scaling by one step. (b–c) In SlipperyBridge, the induced success law changes the root ordering from $0.4144 < 0.55$ to $0.55 < 0.6632$. (d) Thirty matched $\gamma = 1$ finite-horizon runs approach their predicted roots. Bands are 95% Student-t intervals across seeds (29 d.f.). The shared payoff deadline preserves the ordering for $0 < \gamma < 1$ as well.

Theorem 2 identifies the induced evaluator, and Theorem 3 shows how its local error propagates. The policy cycle then motivates the main question: which continuation kernels preserve every action comparison? For fixed, reward- and policy-independent kernels, Theorem 5 gives a finite linear test. A mean-preserving coefficient correction restores the cycling example's true gaps while retaining outcome dependence.

When the transition matrix is unknown, Theorem 8 protects only the comparison used in the next policy update. The coefficient changes on two selected successor actions, with zero mean change under the current policy. The resulting root's gap is a positive multiple of the true gap. Thus eliminating outcome dependence is sufficient, but not necessary, for a correct policy comparison. Each pairwise update uses a coefficient and evaluation for its selected pair.

The learned-likelihood study tests the mechanism in an implemented update. Independent fitting and freezing remove training dependence, but not the reuse of an evaluation outcome. In Accumulated-Charge, more fitting data makes the reversed ordering more reliable.

2 From outcome weighting to an induced evaluator

The local dependence. Let $\mathcal{F}_t$ contain the information available before $Y_t = (R_t, X_{t+1})$, and let $G_t \in [0, 1]$. For a value vector w, the TD error, or innovation, is $\delta_t(w) = R_t + \gamma w(X_{t+1}) - w(X_t)$.

Lemma 1 (Local mean preservation)**.** *The equality* $\mathbb{E}[G_t \Delta_t \mid \mathcal{F}_t] = 0$ *holds for every bounded,* $\sigma(\mathcal{F}_t, Y_t)$*-measurable random variable* Δ_t *with* $\mathbb{E}[\Delta_t \mid \mathcal{F}_t] = 0$ *if and only if*

$$\mathbb{E}[G_t \mid \mathcal{F}_t, Y_t] = \mathbb{E}[G_t \mid \mathcal{F}_t]. \tag{1}$$

Conditioning on $(\mathcal{F}_t, Y_t)$ proves sufficiency. For necessity, use $\Delta_t = \mathbb{E}[G_t \mid \mathcal{F}_t, Y_t] - \mathbb{E}[G_t \mid \mathcal{F}_t]$; the remaining expectation is a conditional variance. The test covers all conditionally centered random variables of the stated class. A particular task may preserve its value through cancellation even when the test fails (Appendix A.2).

The recursive consequence. Consider a finite fixed-policy Markov reward process (MRP) with one-hot features and the trace

$$e_t = \phi(X_t) + \gamma G_t e_{t-1}, \qquad w \leftarrow w + \alpha e_t \delta_t(w). \tag{2}$$

The coefficient G_t rescales only the inherited trace; it does not multiply the new entry $\phi(X_t)$. Assume $0 \le \gamma < 1$ and a fixed rule $G_t = g(X_t, Y_t, U_t)$ with conditionally fresh auxiliary randomness:

$$\Pr(du_{0:T} \mid X_{0:T+1}, R_{0:T}) = \prod_{t=0}^{T} q(du_t \mid X_t, Y_t). \tag{3}$$

Any persistent multiplier memory belongs in the Markov state. Let P, r be the environmental kernel and expected reward. Define

$$\begin{aligned} K(x, x') &= \mathbb{E}[G_t \mathbf{1}\{X_{t+1} = x'\} \mid X_t = x], \\ D &= \mathrm{diag}(\mathbb{E}[G_t \mid X_t = x]), \qquad u(x) = \mathbb{E}[G_t R_t \mid X_t = x]. \end{aligned} \tag{4}$$

Theorem 2 (Population operator)**.** *The expected forward-view operator induced by Eq.* (2) *is*

$$\mathcal{R}_G w = r + \gamma P (I - \gamma K)^{-1} u + \gamma P (I - \gamma K)^{-1} (I - D) w. \tag{5}$$

Its linear part is nonnegative with row sums at most γ. It is therefore a sup-norm contraction with a unique fixed point v_G.

The fixed point can nevertheless differ from the environmental value. Let $v = (I - \gamma P)^{-1} r$ and $c(x) = \mathbb{E}[G_t \delta_t(v) \mid X_t = x]$. The stochastic kernel

$$B = K + (I - D)P \tag{6}$$

propagates this local residual.

Theorem 3 (Exact error transport)**.** *Under the same assumptions,*

$$v_G - v = \gamma P (I - \gamma B)^{-1} c, \qquad \|v_G - v\|_\infty \le \frac{\gamma}{1 - \gamma} \|c\|_\infty. \tag{7}$$

The bound is sharp over finite tabular examples.

The term $(I - D)P$ returns the remaining probability mass to the environmental transition law. Thus small products of coefficients do not by themselves imply small value error. Appendix A.4 proves the identity; Appendix A.5 gives its layered extension and a sharp $\Theta(H)$ distortion at $\gamma = 1$.

Proposition 4 (Connection to online updates)**.** *For any stationary state law $\mu P = \mu$, initialize the trace in its stationary causal solution $e_t^{\leftarrow}$. For fixed w,*

$$\mathbb{E}_\mu[e_t^{\leftarrow} \delta_t(w)] = \mathrm{diag}(\mu)(\mathcal{R}_G w - w). \tag{8}$$

With full-support μ, the expected update vanishes exactly at v_G.

Appendix A.9 proves this identity and an almost-sure convergence result for finite-horizon tabular evaluation. That result uses i.i.d. reset episodes, includes time in the state, and requires a common step size within each episode. The population contraction alone is not an online convergence theorem.

3 Correct model comparisons, incorrect policy updates

We now specialize to deterministic state–action rewards and a fixed coefficient $g(s, a, s') \in [0, 1]$, independent of both reward and policy. There are n states and m state–action pairs. Each row of $P \in \mathbb{R}^{m \times n}$ corresponds to a state–action pair; each column corresponds to a successor state. The matrix $\Pi_\pi \in \mathbb{R}^{n \times m}$ averages over actions under a stationary policy. With $\odot$ denoting entrywise

multiplication, $K = P \odot g$ and $D = \mathrm{diag}(K\mathbf{1})$. Thus $B = K + (I - D)P$ is a policy-independent stochastic kernel.

Applying Theorem 2 to the state–action chain gives

$$\begin{aligned} Q_G^\pi &= r + \gamma P V_B^\pi, \quad V_B^\pi = (I - \gamma \Pi_\pi B)^{-1} \Pi_\pi r, \\ Q_P^\pi &= r + \gamma P V_P^\pi, \quad Q_B^\pi = r + \gamma B V_B^\pi. \end{aligned} \tag{9}$$

The first step of Q_G^π uses P, but its continuation uses B. These are generally not the ordinary action values Q_B^π. The derivation appears in Appendix C.1.

Ordinary models can preserve all action gaps without matching the environment's transition probabilities (Batra, 2026). Yet even that guarantee need not survive the mixed use in Eq. (9). Consider states x, y, actions a, b at x, and one action c at y. In row order $(x, a), (x, b), (y, c)$, take

$$P = \begin{pmatrix} 1/5 & 4/5 \\ 4/5 & 1/5 \\ 3/5 & 2/5 \end{pmatrix}, \qquad B = \begin{pmatrix} 17/162 & 145/162 \\ 125/162 & 37/162 \\ 109/162 & 53/162 \end{pmatrix}. \tag{10}$$

Use $\gamma = 9/10$ and $r = (-1, -27/100, 1)^\mathsf{T}$. Let π_a and π_b choose a and b at x, respectively; both choose c at y. The coefficient $g_{ij}^0 = 1/2 + (B_{ij} - P_{ij})/P_{ij}$ is in $[0, 1]$, has conditional row mean $\sum_j P_{ij} g_{ij}^0 = 1/2$, and induces this B. Ordinary evaluation in B preserves every true numerical action gap for all rewards and stationary policies.[1] Mixed evaluation reverses the comparison under π_a (Table 1).

Table 1: **The same model supports correct comparisons and incorrect updates.** Entries are the b-minus-a gap at x; positive selects b. The last column uses the mean-preserving correction from Section 5. All entries are exact.

Policy	True Q_P	Ordinary Q_B	Mixed Q_G	Corrected mixed
π_a	$-109/1700$	$-109/1700$	$+13/850$	$-109/1700$
π_b	$-109/1025$	$-109/1025$	$-931/41000$	$-109/1025$

Mixed evaluation sends π_a to π_b, then π_b back to π_a. The first switch lowers the true value at x from $10/17$ to $18/205$. Under either policy, the original trace operator has contraction modulus $9/11$; the cycle occurs at its exact roots. The strict inequalities persist over an open reward interval (Appendix B). Additional evaluation cannot correct this mismatch.

4 Which continuation models preserve improvement?

Let $z^\pi = V_B^\pi - V_P^\pi$. The error in comparing two actions is

$$\begin{aligned} &[Q_G^\pi(s, a) - Q_G^\pi(s, b)] - [Q_P^\pi(s, a) - Q_P^\pi(s, b)] \\ &\qquad = \gamma (P_{sa} - P_{sb}) z^\pi. \end{aligned} \tag{11}$$

An action comparison can therefore be correct without correct continuation values. We seek all kernels for which this holds across rewards and policies.

Fix stochastic P, B and $0 < \gamma < 1$ before varying r, π. Choose any deterministic baseline π_0, with action $a_0(s)$. Let C contain the rows $e_{sa}^\mathsf{T} - e_{s,a_0(s)}^\mathsf{T}$ for nonbaseline actions. Thus Cq lists action gaps and CP lists successor-distribution differences. First compute

$$R_0 = (I - \gamma \Pi_0 P)^{-1}, \qquad O = CPR_0, \tag{12}$$

where $\Pi_0 = \Pi_{\pi_0}$. The rows of O measure differences in discounted state visits after alternative actions, following π_0 thereafter. A zero column means these comparisons cannot distinguish visits to that state, not that the state is unreachable. Let $J = \{s : Oe_s \neq 0\}$ and let C_J retain the contrasts at those states. Define

$$F = \begin{pmatrix} O\Pi_0 \\ C_J \end{pmatrix}. \tag{13}$$

[1] For $L = \frac{1}{20}\left(\begin{smallmatrix} 19 & 1 \\ 1 & 19 \end{smallmatrix}\right)$, each row satisfies $B_{sa} L = P_{sa} + \gamma^{-1} e_s^\mathsf{T}(L - I)$. Hence $V_B^\pi = L V_P^\pi$ and $Q_B^\pi(s, a) - Q_P^\pi(s, a) = e_s^\mathsf{T}(L - I) V_P^\pi$, independent of the action.

Write $E = B - P$. The first block, $O\Pi_0 E = 0$, checks compatibility across rewards under the baseline policy. The second, $C_J E = 0$, requires the transition change E to agree across actions at each state in J. Together, these constraints suffice for every stationary randomized policy.

Theorem 5 (Exact comparison compatibility)**.** *The following statements are equivalent:*

1. *For every state–action reward $r \in \mathbb{R}^m$ and every stationary randomized policy π, Q_G^π and Q_P^π have the same complete action order at every state, including ties.*
2. *For every such r, π, their numerical action gaps agree: $CQ_G^\pi = CQ_P^\pi$.*
3. *$F(B - P) = 0$.*

No rank or transition-support assumption is required. For rational inputs, the test can be constructed and checked in polynomial time using one environmental matrix factorization and local action contrasts.

Proof idea. Let $O_\pi = CP(I - \gamma\Pi_\pi P)^{-1}$. The resolvent identity reduces gap preservation to $O_\pi\Pi_\pi(B-P) = 0$ for all π. The zero columns of O_π do not change with the policy. Comparing the baseline with one-state policy changes gives the two blocks of F. They also suffice for randomized policies: the remaining residual is supported on zero columns. Finally, any nonzero gap error can be turned into a strict reversal by changing action rewards while keeping $\Pi_\pi r$ unchanged. Appendix B proves each step.

What the condition permits. The condition allows more than constant value shifts. Its full structure makes the distinction from ordinary model equivalence explicit. Let $S \in \mathbb{R}^{m\times n}$ repeat each state coordinate over its available actions.

Corollary 6 (Compatible transition changes)**.** *Relative to the chosen baseline, every compatible difference $E = B - P$ has the unique form*

$$E = (S - \gamma P)Z + U, \qquad CPZ = 0, \tag{14}$$

where U is zero except on nonbaseline action rows at states outside J. Among compatible kernels, ordinary-model action gaps are also preserved for all rewards and policies if and only if $U = 0$.

Here $Z = R_0\Pi_0 E \in \mathbb{R}^{n\times n}$ and $U = E - (S - \gamma P)Z \in \mathbb{R}^{m\times n}$. The first term cancels in action comparisons. The second allows changes to nonbaseline actions at states with zero columns in O. These changes can retain state-dependent value error.

The reverse separation. In the three-state example of Appendix B.5, mixed evaluation preserves every gap across rewards and policies. For one reward and policy, its continuation-value error is $(1, 1, 3)^\intercal/72$, yet its action gap remains the true $-1/32$. Ordinary evaluation in B instead gives $+1/32$. Together with the two-state cycle, this shows that neither comparison property implies the other (Table 2).

Table 2: **Neither comparison property implies the other.** Gap preservation holds for every reward and stationary policy. A reversal occurs for at least one reward and policy. Both models are realized by legal transition-dependent coefficients.

Continuation model	Ordinary evaluation Q_B	Mixed evaluation Q_G
Two-state (Section 3)	Gaps preserved	Comparison reversed
Three-state (Appendix B.5)	Comparison reversed	Gaps preserved

Ordinary-model equivalence already admits a single-policy test (Batra, 2026, Corollary 3.5); its parameterization also yields an environmental linear test. Related zero-column conditions appear there in Eqs. (26)–(27). Mixed compatibility permits the additional term U but requires $CPZ = 0$. Appendix B develops this comparison of the two classes.

Approximate compatibility also admits a quantitative test. Appendix B.6 gives a continuous residual that vanishes exactly when $F(B-P) = 0$ and bounds comparison errors uniformly over policies. For

a coefficient fitted using estimated transitions, Corollary B.5 adds a transition-estimation allowance to the comparison bound. The resulting bound can be compared with the action gap to assess its sign in the true environment.

5 From compatible coefficients to policy updates

All comparisons, with known transitions. For a coefficient g^0, retain its row means $\mu_i = \sum_j P_{ij} g^0_{ij}$. Then $B - P = P \odot (g - \mu \mathbf{1}^\mathsf{T})$, so compatibility is linear in g.

Proposition 7 (Mean-preserving correction)**.** *For fixed positive weights ω_i, consider*

$$\begin{aligned} \min_{g} \quad & \sum_{i,j} \omega_i P_{ij} (g_{ij} - g^0_{ij})^2 \\ \textit{subject to} \quad & 0 \le g_{ij} \le 1, \qquad \sum_j P_{ij} g_{ij} = \mu_i, \\ & F[P \odot (g - \mu \mathbf{1}^\mathsf{T})] = 0. \end{aligned} \tag{15}$$

The problem is feasible and convex. The minimizer is unique on entries with $P_{ij} > 0$. Its induced mixed evaluator preserves every action gap for every reward and stationary policy.

The row-constant choice $g_{ij} = \mu_i$ is feasible and gives $B = P$. Strict convexity gives uniqueness on positive-probability transitions. The objective measures coefficient change, not learning error.

For the cycling example and $\omega_i = 1$, the unique correction induces

$$B^* = P + \mathbf{1}(d^*)^\mathsf{T}, \qquad d^* = \left(-\frac{91}{3240}, \frac{91}{3240} \right)^\mathsf{T}, \qquad g^*_{ij} = \frac{1}{2} + \frac{d^*_j}{P_{ij}}. \tag{16}$$

Each row remains nonconstant with mean $1/2$, and the contraction modulus stays $9/11$. The corrected gaps select a under either policy (Table 1); Appendix B proves optimality.

One comparison, without the transition matrix. We now allow the coefficient to depend on the policy and sampled successor action. Rather than preserving all numerical gaps, we protect one selected comparison without knowing P. Freeze a policy π, put $T = P\Pi_\pi$, and select actions a, b at state x. Write $p_a = \pi(a \mid x)$, $p_b = \pi(b \mid x)$, and $c^\mathsf{T} = e^\mathsf{T}_{(x,b)} - e^\mathsf{T}_{(x,a)}$. For each current state–action pair i, choose $\mu_i \in [0, 1]$ and $|\kappa_i| \le \min(\mu_i, 1 - \mu_i)$. Define

$$g(i, j) = \begin{cases} \mu_i - \kappa_i p_b, & j = (x, a), \\ \mu_i + \kappa_i p_a, & j = (x, b), \\ \mu_i, & \text{otherwise.} \end{cases} \tag{17}$$

This bounded rule uses π, not P. Its mean over the next action is μ_i at every successor state. With $K = T \odot g$ and $D = \mathrm{diag}(\mu_i)$, its continuation matrix satisfies $A = K + (I - D)T = T + wc^\mathsf{T}$, where $w_i = \kappa_i P_i(x) p_a p_b$. Unlike B, which maps state–action pairs to successor states, A acts on state–action pairs. Only the two selected successor-action columns change. The original trace operator remains contractive, and its root is

$$Q^\pi_G = [I + \gamma T (I - \gamma A)^{-1}] r, \qquad Q^\pi_P = (I - \gamma T)^{-1} r. \tag{18}$$

Theorem 8 (Preserving a selected comparison)**.** *Fix any finite MDP, $0 < \gamma < 1$, and policy π. For the values in Eq.* (18)*, let A be any stochastic matrix of the form $T + wc^\mathsf{T}$. Every reward vector satisfies*

$$c^\mathsf{T} Q^\pi_G = \lambda c^\mathsf{T} Q^\pi_P, \qquad \lambda = \frac{1 - \gamma c^\mathsf{T} w}{1 - \gamma c^\mathsf{T} (I - \gamma T)^{-1} w} > 0. \tag{19}$$

Thus the selected pair's ordering, including ties, is preserved.

The rank-one inverse formula gives the ratio. Its denominator equals $\det(I - \gamma A)/\det(I - \gamma T) > 0$; stochasticity bounds its numerator below by $1 - \gamma$. Allowing κ_i to vary across rows lets continuation actions depend on the preceding state–action pair, so A need not describe any stationary policy in P. Centering does not remove bias: $\mathbb{E}[(g - \mu_i)\delta_i(q) \mid i] = \gamma w_i c^\mathsf{T} q$ can be nonzero.

Other comparisons from the same critic need not be correct. With fixed probability floors, transfer all probability above the worse selected action's floor to the better action, and leave the policy unchanged on ties. Rebuild the coefficient and re-evaluate before comparing the next pair. Each exact update preserves or improves the true value at every state. Repeatedly visiting every action pair at every state reaches the floor-constrained optimum after finitely many policy changes (Appendix F.2).

Under reset access, independent rollouts and a confidence bound also permit conservative updates from sampled gaps (Appendix F.3). A four-task reset-rollout study found no harmful accepted updates, but no advantage over direct evaluation at matched budgets (Appendix F.4).

6 Learned likelihoods in an implemented update

We next test the outcome-reuse mechanism in the released Chunked-TD update (Ramesh et al., 2024), using its original coefficient rule. Chunked-TD uses the current transition's likelihood to rescale the inherited trace before applying its TD error. A phase-lag control delays this scaling by one step. An independent-outcome control evaluates the coefficient on a fresh conditional draw, retaining the target on the actual trajectory. It preserves the weight and target marginals but requires an additional sample. Appendix A.13 relates this timing to the published update identities.

In SlipperyBridge, 30 matched 50,000-episode runs select the bridge under same-outcome timing and the safe action under phase lag or TD(1), in every seed (Figure 1). A separate fixed-policy reset experiment uses 100,000 episodes per mode: all 90 estimated gaps have their predicted signs, and each analytic root lies within its corresponding interval (Appendix A.9).

Does independent fitting remove the reversal? We fit a likelihood on independent data and freeze it during value evaluation. AccumulatedCharge uses $\gamma = 1$ and ten independent increments $Z_j \sim \text{Binomial}(10, 0.2)$. The two root actions receive opposite halves of the centered total; action 1 also receives 0.1. Their environmental values are $(0, 0.1)$. After the root, each state has one action. For any fixed increment multiplier $g(z)$, Appendix A.7 gives

$$Q_G(0) = -10\eta_g, \qquad Q_G(1) = 0.1 + 10\eta_g, \qquad \eta_g = \tfrac{1}{2}\mathbb{E}[g(Z)(Z-2)]. \tag{20}$$

Using an independent $\widetilde{Z}$ for the coefficient instead gives $\mathbb{E}[g(\widetilde{Z})(Z-2)] = 0$, restoring the gap $+0.1$ for every fit.

For $g_{\widehat{q}}(z) = \Pr_{\widehat{q}}(Z = z)$, let $\widehat{q}$ be the maximum-likelihood estimate from N independent episodes, each with ten increments. The same-outcome gap changes sign at the unique threshold $q_\star \approx 0.2228$. As fitting data grows, the estimate converges to $p = 0.2 < q_\star$, where the population gap is approximately -0.2685. The exact reversal probability rises from 0.739 at $N = 1$ to 0.999963 at $N = 50$ (Figure 2). Independent fitting removes training dependence, but not the reuse of the evaluation outcome.

After full neural training. We train the released likelihood module for 10,000 episodes on $H \in \{100, 200\}$ and $p \in \{0.2, 0.35, 0.5\}$, with three training seeds per configuration. Each episode has ten charge events with increments $Z \sim \text{Binomial}(H/10, p)$ and $H + 1$ physical transitions. The grid includes the released implementation's default $(200, 0.5)$. Each model is then frozen and evaluated on 30 fresh paired tabular runs of 100,000 episodes. Overall prediction cross-entropy and charge-event negative log-likelihood decrease between training episodes 1,000 and 10,000 in all 18 models. Yet 11 have negative same-outcome mean gaps with individual 95% intervals below zero; ten rank the actions incorrectly in all 30 runs. Independent-outcome weighting gives intervals entirely above zero for 16 models; the other two span zero. Correct ordering need not imply an accurate gap: at $(100, 0.2)$, the independent-outcome mean gaps are about 0.02, versus the population gap 0.1. Appendix D reports every configuration, training schedule, paired result, and checkpoint error.

These studies test the outcome-reuse mechanism under specific tasks and policies. Appendices A.11, C, and A.10 give the complete matched controls, the policy-dependent sampled cycle, and the non-recursive WIMLE analysis.

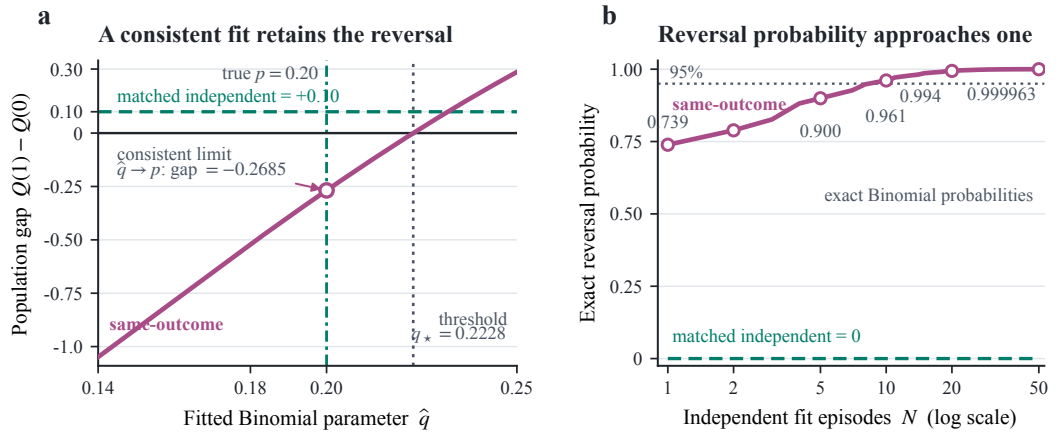

Figure 2: **More fit data makes the reversal more reliable.** (a) Population action gap for an independently fitted and then frozen Binomial likelihood. (b) Exact finite-sample reversal probabilities. Independent-outcome weighting recovers the environmental ordering for every fit while preserving the multiplier's marginal distribution.

7 Related work

Transition-dependent traces already permit a realized coefficient to affect later innovations (Sutton & Singh, 1994; White, 2017). Retrace, generalized stopping-time traces, and trajectory-aware traces use information available before the fresh reward and successor that form the weighted innovation (Munos et al., 2016; Yu et al., 2018; Daley et al., 2023). This permits the conditional Bellman factorization. Expected Eligibility Traces provides the corresponding factorization for state-predictable traces (van Hasselt et al., 2021). Appendix A.14 compares the timing assumptions directly.

Prioritized replay, robust losses, and sample-dependent steps already show that weighting can change an estimand (Schaul et al., 2016; Patterson et al., 2023; Sharifnassab et al., 2026). Our operator identifies the recursive consequence of the same-outcome product trace. The cubic $\rho(p) = 3p^2 - 2p^3$ is the classical restoring transformation (von Neumann, 1956); here it arises as the induced transition law.

Value equivalence matches Bellman updates on specified sets of policies and value functions (Grimm et al., 2020). Our criterion instead preserves action comparisons after a real first step and an induced continuation, while allowing state-dependent value errors. The way a planner uses a model also matters in policy-aware model learning and model-based value expansion (Abachi et al., 2021; Palenicek et al., 2022). Approximate policy iteration relates evaluation errors to progress (Metelli et al., 2021). Gated Q-learning gives a contractive biased operator for off-policy trace retention (Daley, 2026). Ordinary model equivalence characterizes a different use of B (Batra, 2026); Eq. (14) specifies its intersection with mixed compatibility.

Biased values can preserve decisions, as in advantage-learning operators (Bellemare et al., 2016). One-state policy changes also have a rank-one geometry (Dadashi et al., 2019). Theorem 8 concerns the fixed-policy root of a same-outcome trace: its first transition follows T, while later actions may depend on the preceding state–action pair. The coefficient realizes a positive rescaling of one comparison without knowing P; the scale need not exceed one.

8 Scope and conclusion

Theorem 5 characterizes fixed, reward- and policy-independent continuation kernels; its correction uses known transitions. The selected-pair construction changes the coefficient between evaluations. Its sampled updates require reset access and independent batches. Online convergence is established only for reset finite-horizon tabular evaluation. Function approximation and simultaneous online adaptation of the coefficient and policy require further analysis. The neural study freezes the likelihood before value evaluation.

Convergence and correct policy improvement ask different questions. Same-outcome weighting can change a contractive evaluator's continuation law. Correct comparisons inside that model need not survive a real first step. For fixed kernels, compatibility identifies exactly when every comparison survives. Without the transition matrix, a designed coefficient can instead preserve the pair being updated. Values need not all be correct; their errors must leave the intended policy comparison intact.

Reproducibility statement

The appendix gives assumptions and complete proofs. Exact-arithmetic checks test the comparison criterion on fixed finite models and reproduce the two-policy cycle, its correction, and examples with rank-deficient action-transition contrast matrices. The experimental records include source references, fixed protocols, per-seed results, and statistical recalculation code. Appendices D and F specify the supplied records and the larger inputs needed for full reanalysis.

AI use statement

Generative AI tools assisted with theoretical development and proof checking, experiment design and implementation, literature search, figure preparation, and writing. The authors take responsibility for the paper.

References

Romina Abachi, Mohammad Ghavamzadeh, and Amir-massoud Farahmand. Policy-aware model learning for policy gradient methods. *arXiv preprint arXiv:2003.00030*, 2021. URL https://arxiv.org/abs/2003.00030v2. Version 2.

Mehran Aghabozorgi, Alireza Moazeni, Yanshu Zhang, and Ke Li. WIMLE: Uncertainty-aware world models with IMLE for sample-efficient continuous control. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/html/2602.14351v2. arXiv:2602.14351v2.

Neal Batra. From optimal actions to world models: Identifiability of transition kernels in discounted MDPs. *arXiv preprint arXiv:2608.07301*, 2026. URL https://arxiv.org/abs/2608.07301v1. Version 1.

Marc G. Bellemare, Georg Ostrovski, Arthur Guez, Philip S. Thomas, and Rémi Munos. Increasing the action gap: New operators for reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016. URL https://arxiv.org/abs/1512.04860.

Robert Dadashi, Adrien Ali Taiga, Nicolas Le Roux, Dale Schuurmans, and Marc G. Bellemare. The value function polytope in reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1486–1495, 2019. URL https://proceedings.mlr.press/v97/dadashi19a.html.

Brett Daley. Gated Q-learning: Add off-policy bias to taste. *Reinforcement Learning Journal*, 7, 2026. URL https://rlj.cs.umass.edu/2026/papers/Paper141.html.

Brett Daley, Martha White, Christopher Amato, and Marlos C. Machado. Trajectory-aware eligibility traces for off-policy reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 6818–6835. PMLR, 2023. URL https://proceedings.mlr.press/v202/daley23a.html.

Christopher Grimm, Andre Barreto, Satinder Singh, and David Silver. The value equivalence principle for model-based reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5541–5552, 2020. URL https://arxiv.org/abs/2011.03506.

Harold J. Kushner and G. George Yin. *Stochastic Approximation and Recursive Algorithms and Applications*, volume 35 of *Stochastic Modelling and Applied Probability*. Springer, New York, NY, 2nd edition, 2003. doi: 10.1007/b97441. URL https://doi.org/10.1007/b97441.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009. URL https://arxiv.org/abs/0907.3740.

Alberto Maria Metelli, Matteo Pirotta, Daniele Calandriello, and Marcello Restelli. Safe policy iteration: A monotonically improving approximate policy iteration approach. *Journal of Machine Learning Research*, 22(97):1–83, 2021. URL https://www.jmlr.org/papers/v22/19-707.html.

Rémi Munos, Tom Stepleton, Anna Harutyunyan, and Marc G. Bellemare. Safe and efficient off-policy reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/c3992e9a68c5ae12bd18488bc579b30d-Abstract.html.

Daniel Palenicek, Michael Lutter, and Jan Peters. Revisiting model-based value expansion. *arXiv preprint arXiv:2203.14660*, 2022. URL https://arxiv.org/abs/2203.14660v1.

Andrew Patterson, Victor Liao, and Martha White. Robust losses for learning value functions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):6157–6167, 2023. doi: 10.1109/TPAMI.2022.3213503. URL https://arxiv.org/abs/2205.08464.

Aditya Ramesh, Kenny John Young, Louis Kirsch, and Jürgen Schmidhuber. Sequence compression speeds up credit assignment in reinforcement learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 42104–42126. PMLR, 2024. URL https://proceedings.mlr.press/v235/ramesh24b.html.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.05952.

Arsalan Sharifnassab, Mohamed Elsayed, Kris De Asis, A. Rupam Mahmood, and Richard S. Sutton. Intentional updates for streaming reinforcement learning. *arXiv preprint arXiv:2604.19033*, 2026. URL https://arxiv.org/abs/2604.19033.

Richard S. Sutton and Satinder P. Singh. On step-size and bias in temporal-difference learning. In *Proceedings of the Eighth Yale Workshop on Adaptive and Learning Systems*, pp. 91–96. Center for Systems Science, Yale University, 1994. URL https://web.eecs.umich.edu/~baveja/Papers/Yale94.pdf.

Hado van Hasselt, Sephora Madjiheurem, Matteo Hessel, David Silver, André Barreto, and Diana Borsa. Expected eligibility traces. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9997–10005, 2021. doi: 10.1609/aaai.v35i11.17200. URL https://ojs.aaai.org/index.php/AAAI/article/view/17200.

John von Neumann. Probabilistic logics and the synthesis of reliable organisms from unreliable components. In Claude E. Shannon and John McCarthy (eds.), *Automata Studies*, pp. 43–98. Princeton University Press, Princeton, NJ, 1956. URL https://www.smri.it/Publications/vonNeumannNew.pdf.

Martha White. Unifying task specification in reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3742–3750. PMLR, 2017. URL https://proceedings.mlr.press/v70/white17a.html.

Huizhen Yu, A. Rupam Mahmood, and Richard S. Sutton. On generalized Bellman equations and temporal-difference learning. *Journal of Machine Learning Research*, 19(48):1–49, 2018. URL https://jmlr.org/papers/v19/17-283.html.

A Proofs and assumptions

This appendix proves the population-level statements used in the main text. The local result allows an arbitrary pre-outcome history. The recursive results require a fixed Markov multiplier whose auxiliary law factorizes through a fixed local kernel, or an augmented state containing every persistent multiplier variable. Section A.9 also proves almost-sure convergence for reset finite-horizon tabular evaluation under its stated fixed-parameter assumptions.

A.1 Setting

Let X_t be the state of a finite tabular Markov reward process under a fixed policy. Assume that the one-step reward is integrable from every state. Let

$$Y_t = (R_t, X_{t+1})$$

denote the joint transition outcome. The pre-outcome sigma-field $\mathcal{F}_t$ contains the trajectory, parameters, traces, and auxiliary randomness available before Y_t is sampled. For a candidate value w, define

$$\delta_t(w) = R_t + \gamma w(X_{t+1}) - w(X_t). \tag{A.1}$$

The local theorem permits $G_t \in [0, 1]$ to depend on $(\mathcal{F}_t, Y_t)$ and on additional randomization. For the recursive results, we impose the more restrictive form

$$G_t = g(X_t, Y_t, U_t), \tag{A.2}$$

where g and its parameters are fixed. The auxiliary variables are conditionally fresh: for every finite environment trajectory,

$$\Pr(du_{0:T} \mid X_{0:T+1}, R_{0:T}) = \prod_{t=0}^{T} q_{X_t, Y_t}(du_t). \tag{A.3}$$

Equation (A.3) turns products of realized multipliers into products of a one-step kernel. A variable shared across an episode is allowed only after it is included in the Markov state. Infinite-horizon results assume $0 \le \gamma < 1$. The layered finite-horizon result also permits $\gamma = 1$.

A.2 The local timing condition

Every expectation and conditional expectation in this subsection is taken under the current environmental sampling law, denoted P_{env}. In particular, "centered" means centered under P_{env}. A density ratio that targets a declared reference law P_{ref} defines a different estimand, with its conditioning law, centering condition, and multiplier changed consistently. Proposition A.1 concerns preservation under P_{env}.

Define the effective multiplier and its pre-outcome mean by

$$m_t(Y_t) = \mathbb{E}[G_t \mid \mathcal{F}_t, Y_t], \qquad \bar{g}_t = \mathbb{E}[G_t \mid \mathcal{F}_t]. \tag{A.4}$$

Proposition A.1 (Universal local criterion)**.** *The following statements are equivalent:*

$$\mathbb{E}[G_t \Delta_t \mid \mathcal{F}_t] = 0 \quad \textit{for every bounded } \Delta_t \textit{ such that } \mathbb{E}[\Delta_t \mid \mathcal{F}_t] = 0, \tag{A.5}$$

where Δ_t is measurable with respect to $\sigma(\mathcal{F}_t, Y_t)$, and

$$\mathbb{E}[G_t \mid \mathcal{F}_t, Y_t] = \mathbb{E}[G_t \mid \mathcal{F}_t] \quad \textit{almost surely}. \tag{A.6}$$

Proof. Conditional expectation and centeredness give

$$\begin{aligned} \mathbb{E}[G_t \Delta_t \mid \mathcal{F}_t] &= \mathbb{E}[m_t(Y_t)\Delta_t \mid \mathcal{F}_t] \\ &= \mathbb{E}[(m_t(Y_t) - \bar{g}_t)\Delta_t \mid \mathcal{F}_t]. \end{aligned} \tag{A.7}$$

Thus Eq. (A.6) implies Eq. (A.5). Conversely, $m_t(Y_t) - \bar{g}_t$ is bounded, is measurable with respect to $\sigma(\mathcal{F}_t, Y_t)$, and has conditional mean zero. It is therefore an admissible choice of Δ_t in Eq. (A.5). Substitution into Eq. (A.7) yields

$$\mathbb{E}[(m_t(Y_t) - \bar{g}_t)^2 \mid \mathcal{F}_t] = 0,$$

which proves Eq. (A.6). □

If G_t is a deterministic function of $(\mathcal{F}_t, Y_t)$, then $m_t(Y_t) = G_t$, so Eq. (A.6) says that its effective value is already measurable before the current outcome is observed.

The universal quantifier in Proposition A.1 is essential. For one fixed MRP, preservation of its intended value can follow from the weaker, task-specific orthogonality

$$\mathbb{E}[G_t \delta_t(v) \mid X_t = x] = 0 \quad \text{for each state } x. \tag{A.8}$$

This condition may hold even when Eq. (A.6) fails. For example, if Y is uniform on $\{-1, 0, 1\}$, $\Delta(Y) = Y$, and $G(Y) = \mathbf{1}\{Y = 0\}$, then $\mathbb{E}[G\Delta] = 0$ although G is outcome-dependent. Proposition A.1 therefore characterizes preservation of *every* bounded random variable satisfying its measurability and conditional-centering requirements. A fixed task can still satisfy Eq. (A.8) through a task-specific cancellation. Under the recursive assumptions below, Eq. (A.8) makes the vector c in Eq. (A.21) zero, and Theorem A.4 then gives $v_G = v$.

A.3 Exact recursive population operator

Let P be the transition matrix and $r(x) = \mathbb{E}[R_t \mid X_t = x]$. Define

$$K(x, x') = \mathbb{E}[G_t \mathbf{1}\{X_{t+1} = x'\} \mid X_t = x], \tag{A.9}$$

$$a(x) = \mathbb{E}[G_t \mid X_t = x], \qquad D = \mathrm{diag}(a), \qquad u(x) = \mathbb{E}[G_t R_t \mid X_t = x]. \tag{A.10}$$

Then $K \geq 0$, $K\mathbf{1} = a$, and K is substochastic. Set $A = I - \gamma K$. Since $\|\gamma K\|_\infty \leq \gamma < 1$,

$$A^{-1} = \sum_{j=0}^{\infty} \gamma^j K^j \geq 0. \tag{A.11}$$

Consider the outcome-coupled product trace

$$e_t = \phi(X_t) + \gamma G_t e_{t-1}, \qquad w \leftarrow w + \alpha e_t \delta_t(w), \tag{A.12}$$

where ϕ is the one-hot state feature. Its conditional forward-view population operator is obtained as follows. Unrolling Eq. (A.12), the coefficient propagated from the trace injection $\phi(X_0)$ at time 0 is

$$\gamma^n \prod_{i=1}^{n} G_i.$$

The term inherited from a bounded initial trace vanishes geometrically. Because the state space is finite, rewards are integrable, and $0 \leq G_i \leq 1$, the expected absolute sum is dominated by a geometric series. We may therefore collect, for each starting state, all later residuals to obtain

$$(\mathcal{R}_G w)(x) = w(x) + \mathbb{E}\left[\sum_{n \geq 0} \gamma^n \left(\prod_{i=1}^{n} G_i \right) \delta_n(w) \,\middle|\, X_0 = x \right], \tag{A.13}$$

with the empty product equal to one. This definition concerns the expected operator associated with a fixed multiplier. Proposition A.8 below gives the exact link from this conditional forward view to the mean update of the online backward trace.

Lemma A.2 (Operator identity)**.** *Under Eq.* (A.3)*,*

$$\boxed{\mathcal{R}_G w = r + \gamma P (I - \gamma K)^{-1} u + \gamma P (I - \gamma K)^{-1} (I - D) w.} \tag{A.14}$$

Proof. The multiplier-weighted one-step residual is

$$q(w) := \mathbb{E}[G_t \delta_t(w) \mid X_t] = u + \gamma K w - D w. \tag{A.15}$$

The Markov property and Eq. (A.3) imply, for $n \geq 1$,

$$\mathbb{E}\left[\left(\prod_{i=1}^{n} G_i \right) \delta_n(w) \,\middle|\, X_0 \right] = P K^{n-1} q(w). \tag{A.16}$$

Indeed, integrate the last multiplier and residual to obtain $q(w)$, then integrate each preceding multiplier with K, and finally integrate the unweighted transition from X_0 to X_1 with P. Hence Eq. (A.13) becomes

$$\mathcal{R}_G w = r + \gamma P w + \gamma P(I - \gamma K)^{-1} q(w). \tag{A.17}$$

Using $A = I - \gamma K$,

$$Pw + PA^{-1}(\gamma K w - Dw) = PA^{-1}(I - D)w. \tag{A.18}$$

Substituting Eq. (A.15) into Eq. (A.17) and applying Eq. (A.18) proves Eq. (A.14). □

Proposition A.3 (Contraction and uniqueness)**.** *The affine operator $\mathcal{R}_G$ is a sup-norm contraction with modulus at most γ, and therefore has a unique fixed point v_G.*

Proof. The linear part of Eq. (A.14) is

$$T = \gamma P A^{-1}(I - D) \geq 0. \tag{A.19}$$

Moreover,

$$A\mathbf{1} = \mathbf{1} - \gamma a \geq \mathbf{1} - a = (I - D)\mathbf{1}. \tag{A.20}$$

Multiplication by the nonnegative inverse A^{-1} gives

$$A^{-1}(I - D)\mathbf{1} \leq \mathbf{1}.$$

Thus $T\mathbf{1} \leq \gamma\mathbf{1}$. Since T is nonnegative, its maximum row sum is at most γ, so $\|T\|_\infty \leq \gamma < 1$. Banach's fixed-point theorem proves uniqueness. □

Corollary (Explicit augmented-MRP realization). Write the affine operator as

$$\mathcal{R}_G w = b_G + L_G w, \qquad b_G = r + \gamma P(I - \gamma K)^{-1} u, \qquad L_G = \gamma P(I - \gamma K)^{-1}(I - D).$$

Proposition A.3 gives $L_G \geq 0$ and $L_G\mathbf{1} \leq \gamma\mathbf{1}$. Suppose first that $\gamma > 0$. Add an absorbing zero-value state $\perp$ to the original state space $\mathcal{X}$ and define

$$\bar{P}_G(x, x') = \frac{L_G(x, x')}{\gamma}, \qquad \bar{P}_G(x, \perp) = 1 - \sum_{x' \in \mathcal{X}} \frac{L_G(x, x')}{\gamma} \quad (x, x' \in \mathcal{X}),$$

with $\bar{P}_G(\perp, \perp) = 1$. Nonnegativity and the row-mass bound make $\bar{P}_G$ a stochastic matrix. Give state $x \in \mathcal{X}$ the deterministic one-step reward $b_G(x)$ and give $\perp$ reward zero. If $\iota w = (w, 0)$ embeds a value vector by assigning zero to $\perp$, then the ordinary Bellman operator $\bar{\mathcal{T}}_G$ of this augmented MRP satisfies

$$(\bar{\mathcal{T}}_G \iota w)(x) = b_G(x) + \gamma \sum_{x' \in \mathcal{X}} \frac{L_G(x, x')}{\gamma} w(x') = (\mathcal{R}_G w)(x), \qquad (\bar{\mathcal{T}}_G \iota w)(\perp) = 0.$$

Thus $\bar{\mathcal{T}}_G \iota = \iota \mathcal{R}_G$, so $(v_G, 0)$ is the ordinary value of the augmented MRP. If $\gamma = 0$, the assumptions imply $L_G = 0$, and the constant operator $\mathcal{R}_G w = b_G$ is realized by any transition kernel with discount zero and expected reward b_G.

This realization is not unique: the missing row mass may be distributed among any collection of zero-valued absorbing states, and any integrable reward law with conditional mean b_G gives the same Bellman operator. It is an algebraic realization of $\mathcal{R}_G$, not an identification of $\bar{P}_G$ with the environmental transition matrix P.

The proposition certifies the stability of the population operator. Additional sampling, step-size, visitation, and approximation assumptions would be needed to turn it into a stochastic-approximation convergence result.

A.4 Exact error transport and a sharp bound

Let

$$v = (I - \gamma P)^{-1} r$$

be the intended value, and define the local weighted residual

$$c = u + \gamma K v - D v = \mathbb{E}[G_t \delta_t(v) \mid X_t]. \tag{A.21}$$

The matrix that propagates this local weighted residual is

$$B = K + (I - D)P. \tag{A.22}$$

Because $B \geq 0$ and

$$B\mathbf{1} = K\mathbf{1} + (I - D)P\mathbf{1} = a + (\mathbf{1} - a) = \mathbf{1}, \tag{A.23}$$

B is stochastic. We call it an error-transport kernel. It need not, in a general problem, be interpreted as a newly defined environment.

Theorem A.4 (Exact fixed-point error)**.** *Under Eq.* (A.3)*,*

$$\boxed{v_G - v = \gamma P (I - \gamma B)^{-1} c.} \tag{A.24}$$

Consequently,

$$\boxed{\|v_G - v\|_\infty \leq \frac{\gamma}{1 - \gamma} \|c\|_\infty,} \tag{A.25}$$

and the constant is sharp over finite tabular examples.

Proof. Evaluating Eq. (A.14) at v, using the Bellman equation and Eq. (A.21), gives

$$\mathcal{R}_G v - v = \gamma P A^{-1} c. \tag{A.26}$$

Let $h = v_G - v$. The two fixed-point equations imply

$$h = \gamma P A^{-1} \{c + (I - D)h\}. \tag{A.27}$$

Set $z = A^{-1}\{c + (I - D)h\}$. Then $h = \gamma P z$, while

$$\begin{aligned} Az &= c + (I - D)h \\ &= c + \gamma (I - D) P z. \end{aligned}$$

Since $A = I - \gamma K$, this is

$$z = c + \gamma [K + (I - D)P] z = c + \gamma B z. \tag{A.28}$$

Solving Eq. (A.28) and substituting into $h = \gamma P z$ proves Eq. (A.24). Stochasticity gives $\|P\|_\infty = 1$ and

$$\left\| (I - \gamma B)^{-1} \right\|_\infty = \left\| \sum_{j \geq 0} \gamma^j B^j \right\|_\infty = \frac{1}{1 - \gamma},$$

which proves Eq. (A.25).

For sharpness, take a one-state recurrent MRP with equiprobable reward $R \in \{-1, +1\}$ and

$$G = \frac{1 + \eta R}{2}, \qquad 0 < \eta \leq 1.$$

Here $v = 0$, $P = B = [1]$, and $c = \eta/2$. Equation (A.24) gives

$$v_G - v = \frac{\gamma \eta}{2(1 - \gamma)},$$

which attains Eq. (A.25). □

Equation (A.22) explains why small products of multipliers do not by themselves protect the fixed point. The substochastic kernel K carries the weighted mass, while $(I - D)P$ returns the remaining mass through the environmental transition kernel. The resulting error transport preserves total mass.

A.5 Arbitrary finite horizon

Consider layers $t = 0, \dots, H$, with transitions $t = 0, \dots, H-1$, possibly different finite state spaces at each layer, and terminal value $v_H = 0$. Define layerwise P_t, K_t, D_t, c_t as above and

$$B_t = K_t + (I - D_t)P_t. \tag{A.29}$$

Each B_t is a rectangular stochastic kernel from layer t to layer $t+1$.

Theorem A.5 (Layered recursion)**.** *At the joint population root, let $h_t = v_{G,t} - v_t$. Then*

$$\boxed{z_H = 0, \qquad z_t = c_t + \gamma B_t z_{t+1}, \qquad h_t = \gamma P_t z_{t+1}.} \tag{A.30}$$

Equivalently,

$$\boxed{h_t = \gamma P_t \sum_{j=t+1}^{H-1} \gamma^{j-t-1} B_{t+1} \cdots B_{j-1} c_j,} \tag{A.31}$$

where an empty kernel product is the identity, and

$$\boxed{\|h_t\|_\infty \le \sum_{j=t+1}^{H-1} \gamma^{j-t} \|c_j\|_\infty.} \tag{A.32}$$

Proof. For a candidate value family $w = (w_0, \dots, w_H)$, define the multiplier-weighted residual tail

$$Q_t(w) = \mathbb{E}\left[\sum_{n=t}^{H-1} \gamma^{n-t} \left(\prod_{i=t}^{n} G_i\right) \delta_n(w) \,\middle|\, X_t\right], \qquad Q_H = 0. \tag{A.33}$$

One-step conditioning gives

$$Q_t(w) = q_t(w) + \gamma K_t Q_{t+1}(w), \tag{A.34}$$

where, at $w = v + h$,

$$q_t(v + h) = c_t + \gamma K_t h_{t+1} - D_t h_t. \tag{A.35}$$

The current residual followed by the weighted future tail yields the population root equation

$$h_t = \gamma P_t \{h_{t+1} + Q_{t+1}(v + h)\}. \tag{A.36}$$

Set $z_t = h_t + Q_t(v + h)$. Equation (A.36) gives $h_t = \gamma P_t z_{t+1}$. Combining Eqs. (A.34)– (A.36),

$$\begin{aligned} z_t &= c_t + (I - D_t)h_t + \gamma K_t z_{t+1} \\ &= c_t + \gamma[(I - D_t)P_t + K_t]z_{t+1} \\ &= c_t + \gamma B_t z_{t+1}, \end{aligned}$$

which proves Eq. (A.30). Backward substitution gives Eq. (A.31). Since every P_t and B_t is stochastic, applying the sup norm term by term proves Eq. (A.32). The system is triangular, so the same backward argument proves uniqueness, including when $\gamma = 1$. □

The bound is sharp. Take one state per layer, independent equiprobable rewards $R_t \in \{-1, +1\}$, and fresh multipliers $G_t = (1 + \eta R_t)/2$. Then $v_t = 0$, $c_t = \eta/2$, and every $P_t = B_t = [1]$. Hence

$$h_0 = \frac{\eta}{2} \sum_{j=1}^{H-1} \gamma^j, \tag{A.37}$$

with equality in Eq. (A.32). At $\gamma = 1$, the distortion is $(H - 1)\eta/2 = \Theta(H)$.

A.6 SlipperyBridge

We now specify the time-augmented construction. Every hazard reached by a trajectory, including the first, contributes its realized multiplier to the root trace. Fix $H \geq 1$ and a common payoff depth $d \geq H + 1$. Conditional on choosing the bridge at root $x_\star$, transition $t = 0$ is a deterministic zero-reward launch from $x_\star$ to the first hazard state. This launch establishes the predecessor trace: e_0 contains $\phi(x_\star)$, so at the first hazard transition $t = 1$, the root component of e_1 is $\gamma G_1 \phi(x_\star)$. At every reached hazard $j \in \{1, \ldots, H\}$, the same realized hazard outcome appears in both δ_j and the root-trace coefficient $\gamma^j \prod_{i=1}^{j} G_i$. On an all-success trajectory, all H hazard multipliers appear in the root trace before the terminal payoff.

Each hazard transition has zero reward and succeeds with environmental probability p or slips with probability $q = 1-p$. A slip enters a deterministic failure delay chain; H successes enter a deterministic success delay chain. The safe action enters its own deterministic delay chain. The lengths of these chains are chosen so that the safe payoff s, the early-slip payoff r_F, and the completion payoff r_S are all delivered on the common transition $t = d$, with every earlier reward zero. The deterministic launch, all delay-chain transitions, and the terminal payoff transitions use $G = 1$. Thus, for any $0 < \gamma < 1$, all root action values share the same positive factor γ^d. The experiments use the layered $\gamma = 1$ case, whose population system is uniquely solved by backward induction.

On every deterministic launch or delay-chain arc, P is a point mass and $G = 1$, hence $K = P$, $D = I$, and

$$B = K + (I - D)P = P.$$

Such arcs preserve transition mass exactly. At a hazard state, use the self-likelihood multiplier $G(Y) = P_{\text{env}}(Y \mid X)$. The success and failure entries of K are then p^2 and q^2, while

$$D = p^2 + q^2, \qquad 1 - D = 2pq.$$

In this bridge construction, the error-transport kernel $B = K + (I - D)P$ assigns success probability

$$\begin{aligned} \rho(p) &= p^2 + (1 - D)p \\ &= p^2 + 2p^2 q \\ &= \boxed{p^2(3 - 2p)}. \end{aligned} \tag{A.38}$$

The corresponding failure probability is $q^2(3 - 2q)$, and the two sum to one. Also

$$\rho(p) - p = p(1 - p)(2p - 1), \tag{A.39}$$

so the induced bridge has a strictly larger per-stage success probability whenever $1/2 < p < 1$. The cubic in Eq. (A.38) is the classical majority-of-three restoring map. Here it arises as the transition law of the Bellman population operator.

Completing all H stages yields r_S at the common deadline, while a slip yields $r_F < r_S$ at that deadline. Factor out the common discount factor γ^d. The environmental and induced recursions then give, by induction on H,

$$V_H = r_F + (r_S - r_F)p^H, \qquad \widetilde{V}_H = r_F + (r_S - r_F)\rho(p)^H. \tag{A.40}$$

Along the all-success branch, there are exactly H hazard kernels between the launch and the common deadline. The environmental all-success mass is therefore p^H, whereas the error-transport kernel assigns all-success mass $\rho(p)^H$; deterministic launch and delay-chain arcs contribute only $B = P$. Consequently the three root action values are

$$Q_{\text{env}}(x_\star, a_{\text{safe}}) = \gamma^d s, \qquad Q_{\text{env}}(x_\star, a_{\text{bridge}}) = \gamma^d V_H, \qquad Q_G(x_\star, a_{\text{bridge}}) = \gamma^d \widetilde{V}_H.$$

The deterministic launch creates the root trace before hazard 1. Each of the H hazards on the all-success branch therefore contributes one multiplier. No deterministic arc changes the induced branch probability, yielding the factor $\rho(p)^H$. Thus, for every $1/2 < p < 1$, the interval

$$V_H < s < \widetilde{V}_H \tag{A.41}$$

is a nonempty open set of safe payoffs that reverses the decision.

For the running parameters $p = 0.8$, $H = 3$, $r_S = 1$, and $r_F = -0.2$,

$$V_3 = -0.2 + 1.2(0.8)^3 = 0.4144,$$

whereas $\rho(0.8) = 0.896$ and

$$\widetilde{V}_3 = -0.2 + 1.2(0.896)^3 = 0.6631877632. \tag{A.42}$$

The safe payoff $s = 0.55$ lies strictly between them.

Robustness over a parameter neighborhood. Write the slip payoff as $r_F = -\kappa$. On

$$p \in [0.78, 0.82], \qquad \kappa \in [0.15, 0.25], \qquad s \in [0.52, 0.57], \tag{A.43}$$

both bridge values increase with p and decrease with κ; $\rho'(p) = 6p(1-p) > 0$. Consequently,

$$\max V_3 = 1.15(0.82)^3 - 0.15 = 0.4840732 < 0.52, \tag{A.44}$$

and

$$\min \widetilde{V}_3 = 1.25\rho(0.78)^3 - 0.25 = 0.590553005635666\ldots > 0.57. \tag{A.45}$$

Hence every point in the specified box satisfies $V_3 < s < \widetilde{V}_3$, with uniform margins at least 0.0359268 and 0.0205530, respectively. Because the inequalities are strict and the value maps are continuous, the reversal persists throughout an open neighborhood of the box.

A.6.1 Rare-hazard corollary

Set $p = 1 - \varepsilon$. Equation (A.38) becomes

$$\rho(1-\varepsilon) = 1 - \varepsilon', \qquad \varepsilon' = 3\varepsilon^2 - 2\varepsilon^3. \tag{A.46}$$

For a bridge of length H, the real and induced probabilities of at least one slip are

$$T_H(\varepsilon) = 1 - (1-\varepsilon)^H, \qquad E_H(\varepsilon) = 1 - (1-\varepsilon')^H. \tag{A.47}$$

Fix $\tau > 0$, take $0 < \varepsilon < \min\{\tau, 1/2\}$, and set $H_\varepsilon = \lfloor \tau/\varepsilon \rfloor$. Since $\tau - \varepsilon \le H_\varepsilon \varepsilon \le \tau$ and

$$1 - e^{-Hx} \le 1 - (1-x)^H \le 1 - e^{-Hx/(1-x)}, \tag{A.48}$$

we have

$$1 - e^{-(\tau-\varepsilon)} \le T_{H_\varepsilon}(\varepsilon) \le 1 - e^{-\tau/(1-\varepsilon)}. \tag{A.49}$$

Therefore

$$T_{H_\varepsilon}(\varepsilon) = 1 - e^{-\tau} + O_\tau(\varepsilon). \tag{A.50}$$

For the induced process, the union and second Bonferroni bounds give

$$0 \le H_\varepsilon \varepsilon' - E_{H_\varepsilon}(\varepsilon) \le \binom{H_\varepsilon}{2} (\varepsilon')^2. \tag{A.51}$$

Writing $H_\varepsilon = \tau/\varepsilon + d_\varepsilon$, with $-1 < d_\varepsilon \le 0$, yields

$$|H_\varepsilon \varepsilon' - 3\tau\varepsilon| \le (2\tau + 3)\varepsilon^2 + 2\varepsilon^3. \tag{A.52}$$

Combining Eqs. (A.51) and (A.52),

$$|E_{H_\varepsilon}(\varepsilon) - 3\tau\varepsilon| \le \varepsilon^2 \left[2\tau + 3 + 2\varepsilon + \frac{\tau^2}{2}(3 - 2\varepsilon)^2 \right]. \tag{A.53}$$

Thus

$$\boxed{T_{H_\varepsilon}(\varepsilon) \to 1 - e^{-\tau}, \qquad E_{H_\varepsilon}(\varepsilon) = 3\tau\varepsilon + O_\tau(\varepsilon^2) \to 0.} \tag{A.54}$$

This scaling limit follows from the same map in Eq. (A.38). It also yields decision reversals on an open set. Use the normalized values in Eq. (A.40). Let $d = r_S - r_F > 0$ and $a = (s - r_F)/d$. If $e^{-\tau} < a < 1$, then the environmental risky value tends to $r_F + de^{-\tau} < s$, while the induced value tends to $r_F + d = r_S > s$. These strict inequalities define an open bounded-reward reversal set.

A.7 Learned likelihoods on AccumulatedCharge

This section removes the perfect-likelihood restriction on AccumulatedCharge, a task in the public Chunked-TD repository. At $p = 0.5$, the exact symmetric likelihood has zero local weighted residual, so we study $p = 0.2$. We use the undiscounted setting ($\gamma = 1$). The task has $M = 10$ accumulation events, and at each event,

$$Z_j \overset{\text{iid}}{\sim} \text{Binomial}(n, p), \qquad n = 10, \quad p = 0.2, \tag{A.55}$$

and root action $a \in \{0, 1\}$ receives the terminal reward

$$R_a = \beta a + \frac{2a - 1}{2} \left(\sum_{j=1}^{M} Z_j - Mnp \right), \qquad \beta = 0.1. \tag{A.56}$$

Thus the environmental root values are $Q(0) = 0$ and $Q(1) = \beta$. The supplementary provenance table links Eqs. (A.55)– (A.56) to the archived public source. The state–action likelihood also contains the next-action probability. Every post-root AccumulatedCharge state exposes a single available action, so that policy factor is identically one and the multiplier reduces exactly to the transition likelihood below.

Let $g : \{0, \dots, n\} \to [0, 1]$ be any model likelihood fitted on data independent of value evaluation and then frozen. The same-outcome update uses $G_j = g(Z_j)$ on the innovation formed from that same Z_j. Define

$$\eta_g = \frac{1}{2} \mathbb{E}_p[g(Z)(Z - np)] \,. \tag{A.57}$$

Proposition A.6 (Frozen learned-likelihood root)**.** *For the time-augmented fixed-policy population system above,*

$$Q_G(0) = -M\eta_g, \qquad Q_G(1) = \beta + M\eta_g, \qquad \Delta_G := Q_G(1) - Q_G(0) = \beta + 2M\eta_g. \tag{A.58}$$

If an independent draw $\widetilde{Z}_j \sim \text{Binomial}(n, p)$ with $\widetilde{Z}_j \perp Z_j$ supplies the multiplier $G_j = g(\widetilde{Z}_j)$, then $Q_G = (0, \beta)$ for every frozen g.

Proof. Write $\sigma_a = 2a - 1$. Immediately before one accumulation event, the environmental continuation value centers the remaining charge increments at their mean. The zero-reward TD innovation at the environmental value is therefore

$$\delta_j(v) = \frac{\sigma_a}{2}(Z_j - np). \tag{A.59}$$

Hence the local weighted residual at every reachable state in that layer is the state-independent constant

$$c_j = \mathbb{E}[g(Z_j)\delta_j(v)] = \sigma_a \eta_g. \tag{A.60}$$

All deterministic transitions have multiplier one. On stochastic layers, $B_j \mathbf{1} = \mathbf{1}$ by Eq. (A.23). Applying the finite-horizon recursion in Eq. (A.30) backward from the terminal layer therefore preserves constant vectors and sums one copy of $\sigma_a \eta_g$ for each of the M accumulation events. The root error is $M\sigma_a\eta_g$, which gives Eq. (A.58).

Under the matched-independent construction,

$$\mathbb{E}[g(\widetilde{Z}_j)(Z_j - np)] = \mathbb{E}[g(\widetilde{Z}_j)] \, \mathbb{E}[Z_j - np] = 0 \tag{A.61}$$

conditionally on the frozen fit. Thus every local weighted residual vanishes and the environmental root is recovered. □

The result also describes the limiting population gap as the fitted likelihood becomes consistent. It does not establish stochastic-approximation convergence. Let $p_z = \Pr_p(Z = z)$, and let $\widehat{g}_N(z)$ be any independently fitted frozen likelihood satisfying $\max_z |\widehat{g}_N(z) - p_z| \to 0$ in probability. Since Eq. (A.58) is a continuous linear functional of the finite multiplier vector,

$$\Delta_{\widehat{g}_N} \xrightarrow{p} \beta + M \sum_{z=0}^{n} p_z^2 (z - np). \tag{A.62}$$

For the $H = 100$, $p = 0.2$ configuration used here, this limit is

$$\eta_p = -0.01842413576719958016, \tag{A.63}$$
$$Q_G = (0.18424135767199580 16, -0.08424135767199580 16), \tag{A.64}$$
$$\Delta_G = -0.26848271534399160 32. \tag{A.65}$$

Thus the population action ordering reverses the environmental ordering with probability tending to one as the frozen likelihood estimator becomes consistent.

For a finite-sample calculation, fit the correctly specified family

$$g_q(z) = \binom{n}{z} q^z (1-q)^{n-z}. \tag{A.66}$$

Substitution into Eq. (A.58) gives the degree-ten polynomial

$$\Delta(q) = \beta + M \sum_{z=0}^{n} p_z \binom{n}{z} q^z (1-q)^{n-z} (z - np). \tag{A.67}$$

The exact-arithmetic calculation constructs its rational coefficients and uses a Sturm sequence to certify exactly one root in $[0, 1]$:

$$q_\star = 0.22277182574584593956011165615 4 \ldots. \tag{A.68}$$

The polynomial is negative at zero and positive at one, so the fitted model reverses the action ordering exactly when $q < q_\star$.

If $\widehat{q}_N$ is the maximum-likelihood estimate from N independent model-fit episodes, then each episode contributes $Mn = 100$ Bernoulli trials, so

$$100N\widehat{q}_N \sim \text{Binomial}(100N, 0.2), \tag{A.69}$$
$$\Pr\{\Delta(\widehat{q}_N) < 0\} = \Pr\{S_N \le \lfloor 100N q_\star \rfloor\}, \quad S_N \sim \text{Binomial}(100N, 0.2). \tag{A.70}$$

High-precision recurrence gives:

Fit episodes N	Exact reversal probability
1	.7389327523309054
2	.7887383082854017
5	.8995602626586590
10	.9610244723274860
20	.9940421923552537
50	.9999628753525703

Because $\widehat{q}_N \to 0.2 < q_\star$, the probability tends to one. In contrast, Proposition A.6 gives zero reversal probability for matched-independent use at every N, conditional on every frozen fit. The supplementary exact-arithmetic calculation reproduces the polynomial and root count and evaluates the probabilities with 90-digit decimal arithmetic; invocation and environment details are provided in the supplementary documentation.

A.8 Why persistent randomness must be represented

The factorization in Eq. (A.3) cannot be replaced by one-step marginalization of a hidden episode-level variable.

Counterexample A.7. Consider one recurrent observed state, $\gamma = 1/2$, and i.i.d. equiprobable rewards $R_t \in \{-1, +1\}$. Draw $U \in \{0, 1\}$ once per infinite run, independently and uniformly, and set

$$G_t = \mathbf{1}\{U = 0,\ R_t = +1\}. \tag{A.71}$$

If U is omitted from the state, the one-step conditional means are

$$m(+1) = \frac{1}{2}, \qquad m(-1) = 0, \qquad a = \frac{1}{4}, \qquad c = \frac{1}{4}. \tag{A.72}$$

The marginal kernel has $B = P = [1]$. Incorrectly applying Eq. (A.24) to these marginalized quantities would predict

$$v_G - v = \frac{\gamma}{1-\gamma} c = \frac{1}{4}. \tag{A.73}$$

The actual expected-update root is different. In the stationary trace recursion $e_t = 1 + \gamma G_t e_{t-1}$, conditional independence of R_t and e_{t-1} given U gives

$$\mathbb{E}[e_t \mid U = 0] = \frac{1}{1-\gamma/2} = \frac{4}{3}, \qquad \mathbb{E}[e_t \mid U = 1] = 1. \tag{A.74}$$

Hence $\mathbb{E}[e_t] = 7/6$. Moreover,

$$\begin{aligned}\mathbb{E}[e_t R_t \mid U = 0] &= \gamma\, \mathbb{E}[G_t R_t \mid U = 0]\mathbb{E}[e_{t-1} \mid U = 0] \\ &= \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{4}{3} = \frac{1}{3},\end{aligned}$$

and therefore $\mathbb{E}[e_t R_t] = 1/6$ after averaging over U. In the one-state problem,

$$\delta_t(w) = R_t - (1-\gamma)w.$$

The scalar expected-update root solves

$$\mathbb{E}[e_t \delta_t(w)] = 0,$$

so

$$\boxed{w = \frac{1/6}{(1/2)(7/6)} = \frac{2}{7} \neq \frac{1}{4}.} \tag{A.75}$$

Products of multipliers share U, so their expectations cannot be replaced by powers of the marginalized K. If U is included in the state, the two augmented states satisfy the Markov assumptions and the recursive theorem applies statewise. A scalar approximation that aliases those two states is a projected function-approximation problem and is outside the tabular result.

A.9 From the online backward view to the population operator

The definition in Eq. (A.13) is conditional on a starting state, whereas the implemented update in Eq. (A.12) is applied at the current time. The following result connects this conditional forward view to the implemented backward update. The candidate w, the MRP, and the multiplier—including all its parameters—are fixed throughout the expectation.

Proposition A.8 (Stationary backward-view mean field)**.** *Under the assumptions of Section A.1, let $0 \le \gamma < 1$ and let $(X_t, R_t, U_t)_{t\in\mathbb{Z}}$ be a two-sided stationary realization with $X_t \sim \mu$. Initialize the eligibility trace at its stationary solution*

$$e_t^{\leftarrow} = \sum_{j=0}^{\infty} \gamma^j \left(\prod_{\ell=0}^{j-1} G_{t-\ell} \right) \phi(X_{t-j}), \tag{A.76}$$

where the empty product is one. Then, for every state x and every fixed candidate w,

$$\boxed{\mathbb{E}_\mu[e_t^{\leftarrow}(x)\delta_t(w)] = \mu(x)\big[(\mathcal{R}_G w)(x) - w(x)\big], \qquad \mathbb{E}_\mu[e_t^{\leftarrow}\delta_t(w)] = \mathrm{diag}(\mu)(\mathcal{R}_G w - w).} \tag{A.77}$$

Consequently, if μ has full support, the stationary expected update is zero if and only if $w = v_G$. Given the fixed-multiplier sampling assumptions above, stationarity—rather than ergodicity—is sufficient for Eq. (A.77)*; irreducibility and aperiodicity are not required.*

Proof. The series in Eq. (A.76) is absolutely bounded by $(1-\gamma)^{-1}$ coordinatewise and solves $e_t^{\leftarrow} = \phi(X_t) + \gamma G_t e_{t-1}^{\leftarrow}$. Integrability of the rewards and finiteness of the state space justify exchanging the expectation and the geometrically dominated sum. For a fixed coordinate x, stationarity and a

shift of j time steps give

$$\begin{aligned}\mathbb{E}_\mu[e_t^{\leftarrow}(x)\delta_t(w)] &= \sum_{j\geq 0}\gamma^j \mathbb{E}_\mu\left[\mathbf{1}\{X_{t-j}=x\}\left(\prod_{i=t-j+1}^{t} G_i\right)\delta_t(w)\right] \\ &= \mu(x)\sum_{j\geq 0}\gamma^j \mathbb{E}\left[\left(\prod_{i=1}^{j} G_i\right)\delta_j(w)\,\middle|\, X_0=x\right] \\ &= \mu(x)\big[(\mathcal{R}_G w)(x) - w(x)\big].\end{aligned}$$

Stacking the coordinates proves Eq. (A.77). Under full support, $\mathrm{diag}(\mu)$ is invertible, so a zero mean update is equivalent to $\mathcal{R}_G w = w$; Proposition A.3 then gives $w = v_G$. □

If the environment is stationary but the trace starts from any uniformly bounded initialization rather than Eq. (A.76), its difference from the stationary trace is multiplied by at most γ at each step. The corresponding expected-update transient therefore vanishes geometrically, although Eq. (A.77) is exact at finite time only for the stationary trace. When $\mu(x) = 0$, both sides of the coordinate identity vanish and provide no estimating equation for x. In a finite chain, stationary laws put zero mass on transient states and can put zero mass on entire recurrent classes. Thus full support, or a separate visitation argument supplying every coordinate, is needed to infer the global root from the online mean field.

Corollary A.9 (Episodic occupancy mean field). Fix a finite horizon $H \geq 1$ and $0 \leq \gamma \leq 1$, and augment the state with time, $\bar{X}_t = (t, X_t)$ for $t = 0, \ldots, H$, with $X_0 \sim \nu$ and terminal candidate $w_H = 0$. Assume the layerwise analogue of Eq. (A.3); keep w and the multiplier fixed within the expectation. At the start of every episode reset $e_{-1} = 0$, and within the episode use $e_t = \phi(\bar{X}_t) + \gamma G_t e_{t-1}$ for $t = 0, \ldots, H-1$. Define the unnormalized time-augmented occupancy

$$d_\nu(s,x) = \sum_{t=0}^{H-1} \Pr_\nu\{\bar{X}_t = (s,x)\} = \Pr_\nu\{X_s = x\}$$

and the truncated conditional operator

$$(\mathcal{R}_{G,H} w)(s,x) = w_s(x) + \mathbb{E}_\nu\left[\sum_{n=s}^{H-1}\gamma^{n-s}\left(\prod_{i=s+1}^{n} G_i\right)\delta_n(w)\,\middle|\, X_s = x\right]. \tag{A.78}$$

Then the expected total update in one episode satisfies

$$\boxed{\mathbb{E}_\nu\left[\sum_{t=0}^{H-1} e_t(s,x)\delta_t(w)\right] = d_\nu(s,x)\big[(\mathcal{R}_{G,H} w)(s,x) - w_s(x)\big].} \tag{A.79}$$

Proof. Because the state includes time, the coordinate (s,x) can enter the trace only at time s. For $t \geq s$, its coefficient is

$$e_t(s,x) = \mathbf{1}\{X_s = x\}\gamma^{t-s}\prod_{i=s+1}^{t} G_i,$$

and it is zero for $t < s$. Substitution into the episode sum, followed by conditioning on $X_s = x$, gives Eq. (A.79) directly. □

Hence the episodic mean update identifies the truncated-operator equation at every coordinate with $d_\nu(s,x) > 0$, but not at unreachable coordinates. No stationarity or ergodicity is needed: finite-horizon occupancy replaces stationary mass. If episodes are concatenated, the trace must be reset at the terminal-to-initial boundary; otherwise cross-episode eligibility terms are present and Eq. (A.79) does not apply.

Proposition A.10 (Episodic stochastic approximation reaches the induced root)**.** *Assume the setting of Corollary A.9, bounded rewards, i.i.d. reset episodes, and a fixed multiplier kernel. Restrict the parameter vector to the finite set of reachable nonterminal time–state coordinates, so $d_\nu(s,x) > 0$ for every retained coordinate. Let $F(w;\xi)$ be the total backward-view update from one episode ξ when the candidate w is held fixed throughout that episode. If*

$$w_{k+1} = w_k + \alpha_k F(w_k;\xi_{k+1}), \qquad \sum_k \alpha_k = \infty, \qquad \sum_k \alpha_k^2 < \infty, \tag{A.80}$$

with deterministic $\alpha_k > 0$, then

$$w_k \longrightarrow v_{G,H} \qquad \textit{almost surely}, \tag{A.81}$$

where $v_{G,H}$ is the unique reachable-coordinate solution of $\mathcal{R}_{G,H} v = v$.

The ordinary sequential implementation that uses the same α_k at each transition of episode k and updates w immediately after every innovation is algebraically identical at the episode boundary and therefore has the same almost-sure limit.

Proof. Write the truncated affine operator on reachable coordinates as $\mathcal{R}_{G,H} w = b + Lw$. Order coordinates by increasing time. The current coordinate $w_s(x)$ cancels against the $-w_s(x)$ term in the current innovation in Eq. (A.78); every remaining value term belongs to a later layer. Hence L is strictly block upper triangular. With

$$D_\nu = \operatorname{diag}\big(d_\nu(s,x)\big) > 0,$$

Corollary A.9 gives the episode mean field

$$h(w) = \mathbb{E}[F(w;\xi)] = D_\nu\{b + (L-I)w\} = -A(w - v_{G,H}), \qquad A = D_\nu(I-L). \tag{A.82}$$

The matrix A is block upper triangular and its diagonal entries are the strictly positive occupancies. Thus every eigenvalue of A has positive real part, and the linear ODE $\dot{w} = h(w)$ is globally exponentially stable at $v_{G,H}$.

For a fixed episode, $F(w;\xi)$ is affine in w. Finite horizon, $0 \le G_t \le 1$, finite state spaces, and bounded rewards make its random affine coefficients uniformly bounded. Therefore

$$M_{k+1} = F(w_k;\xi_{k+1}) - h(w_k)$$

is a martingale difference with

$$\mathbb{E}[\|M_{k+1}\|^2 \mid \mathcal{G}_k] \le C\big(1 + \|w_k - v_{G,H}\|^2\big), \tag{A.83}$$

where $\mathcal{G}_k$ is the sigma-field before episode $k+1$. We prove stability directly. Since every eigenvalue of A has positive real part, there is a symmetric $P > 0$ solving $A^\top P + PA = I$. Put $x_k = w_k - v_{G,H}$ and $V_k = x_k^\top P x_k$. Expanding one recursion step, using the martingale-difference property and Eq. (A.83), gives constants $c, C' > 0$ such that, for all sufficiently large k,

$$\mathbb{E}[V_{k+1} \mid \mathcal{G}_k] \le (1 + C'\alpha_k^2)V_k - c\alpha_k V_k + C'\alpha_k^2. \tag{A.84}$$

The Robbins–Siegmund almost-supermartingale argument therefore makes V_k convergent and $\sum_k \alpha_k V_k < \infty$ almost surely. Because $\sum_k \alpha_k = \infty$, its limit must be zero. Hence $x_k \to 0$ almost surely, proving Eq. (A.81). This is the standard linear martingale-difference stochastic-approximation proof; see, e.g., Kushner & Yin (2003) for the general framework.

For the sequential implementation, the trace does not depend on w, and time is part of the state. After transition t, the support of e_t contains only coordinates from layers $0, \dots, t$. The next innovation depends only on the value coordinates at layers $t+1$ and $t+2$, which have not yet been changed in the current episode. Inductively, every sequential innovation equals the innovation computed from the episode-start vector w_k. Therefore the sequential update at the end of each episode is exactly Eq. (A.80). □

This proposition establishes convergence for reset finite-horizon tabular evaluation with a fixed multiplier and one common step size per episode.

Finite-run verification. We instantiated the reset time-augmented SlipperyBridge MRP with the paper's central parameters, held the policy and multiplier fixed, and used $\alpha_k = 0.8(k + 50)^{-0.6}$ for 100,000 episodes. Hazard draws were matched across modes for seeds $0, \ldots, 29$. The ordinary sequential implementation was used; the proof above shows that its episode-boundary iterates equal the batch recursion exactly. The resulting root-action gaps were:

Mode	Exact gap	Mean	95% t interval	Sign agreement
Same outcome	$+0.113188$	$+0.112250$	$[+0.109410, +0.115091]$	$30/30$
Phase lag	-0.135600	-0.136491	$[-0.140904, -0.132077]$	$30/30$
TD(1)	-0.135600	-0.136368	$[-0.141129, -0.131608]$	$30/30$

Every exact population gap lies inside its interval. The supplementary material provides the fixed-policy configuration, implementation, per-seed results, and summary statistics. This verification separates the population effect from control and exploration in finite runs. In the separate 50,000-episode experiment in Figure 1, the endpoint intervals also contain the corresponding analytic roots.

A.10 Finite-support analysis of WIMLE's critic objective

WIMLE weights synthetic targets by uncertainty from finite world-model predictions (Aghabozorgi et al., 2026). Its positive-weight argument requires a weight fixed by state and action. The finite implementation instead derives target and weight from the same prediction block, which can make them dependent even under a perfect model. We analyze a bounded terminal construction under the published critic loss and target–weight computation, using the public implementation version archived with the supplementary material. In that implementation, one prediction tensor determines the synthetic reward, next-state prediction, and uncertainty weight. The weight multiplies a 100-quantile Huber critic loss with $\kappa = 1$, and real and model batches enter in a 1:1 ratio. The published configuration uses 7 ensemble models and 10 latent draws, giving $N = 70$ predictions; we also enumerate the configured 7×4 case, $N = 28$. Exact source locations and hashes appear in the supplementary provenance table.

A.10.1 Terminal law and finite prediction block

There is one risky action. A perfect model and the real environment share the following bounded terminal law for reward R and the scalar next-state prediction S':

R	S'	Probability
-2	3	$1/50$
0	-3	$91/100$
$4/7$	-3	$7/100$

The risky action has environmental value

$$\mathbb{E}[R] = \frac{1}{50}(-2) + \frac{91}{100}(0) + \frac{7}{100}\frac{4}{7} = 0. \tag{A.85}$$

A deterministic safe action has terminal value $v_{\text{safe}} = 0.005$.

For a synthetic transition, draw an iid block $\{(R_i, S'_i)\}_{i=1}^N$ from the law above and set

$$\bar{R}_N = \frac{1}{N}\sum_{i=1}^N R_i, \qquad \bar{S}'_N = \frac{1}{N}\sum_{i=1}^N S'_i, \tag{A.86}$$

$$W_N = \frac{1}{1 + \mathrm{sd}_N(R_{1:N}) + \mathrm{sd}_N(S'_{1:N})}, \tag{A.87}$$

where sd_N is the empirical standard deviation with divisor N, as in the corresponding `numpy.std` call. Because the transition is terminal, the synthetic critic target is $\bar{R}_N$; there is no bootstrap term.

For deterministic enumeration, let $c = (c_0, c_1, c_2)$ be the counts of the three outcomes, with $c_0 + c_1 + c_2 = N$. Writing their probabilities as p_k, rewards as r_k, and next-state predictions as s_k, define

$$\pi_N(c) = \frac{N!}{c_0!c_1!c_2!} \prod_{k=0}^{2} p_k^{c_k}, \tag{A.88}$$

$$\bar{r}_c = \sum_{k=0}^{2} \frac{c_k}{N} r_k, \qquad \bar{s}_c = \sum_{k=0}^{2} \frac{c_k}{N} s_k, \tag{A.89}$$

$$\sigma_{r,c} = \sqrt{\sum_{k=0}^{2} \frac{c_k}{N} (r_k - \bar{r}_c)^2}, \tag{A.90}$$

$$\sigma_{s,c} = \sqrt{\sum_{k=0}^{2} \frac{c_k}{N} (s_k - \bar{s}_c)^2}, \tag{A.91}$$

$$W_c = (1 + \sigma_{r,c} + \sigma_{s,c})^{-1}. \tag{A.92}$$

Thus exact enumeration sums over the finite multinomial support.

A.10.2 Three population measures

The real half of every 1:1 mixture has coefficient one. Let δ_z denote a unit point mass at z and $\bar{W}_N = \sum_c \pi_N(c) W_c$. The three positive, not necessarily normalized, loss measures are

$$\nu_N^{\text{same}} = \frac{1}{2} \sum_{k=0}^{2} p_k \delta_{r_k} + \frac{1}{2} \sum_c \pi_N(c) W_c \delta_{\bar{r}_c}, \tag{A.93}$$

$$\nu_N^{\text{ind}} = \frac{1}{2} \sum_{k=0}^{2} p_k \delta_{r_k} + \frac{1}{2} \bar{W}_N \sum_c \pi_N(c) \delta_{\bar{r}_c}, \tag{A.94}$$

$$\nu_N^{\text{unw}} = \frac{1}{2} \sum_{k=0}^{2} p_k \delta_{r_k} + \frac{1}{2} \sum_c \pi_N(c) \delta_{\bar{r}_c}. \tag{A.95}$$

Equation (A.93) uses one block for both $\bar{R}_N$ and W_N. In Eq. (A.94), one block supplies $\bar{R}_N$ and an iid block supplies W_N. Integrating out the latter gives $\bar{W}_N$; consequently, the target and weight marginals are unchanged while their shared-block dependence is removed. Equation (A.95) sets every coefficient to one. Global normalization of any measure would not change its minimizers.

A.10.3 Published quantile-Huber objective

For $\kappa = 1$, define

$$h_1(u) = \begin{cases} \frac{1}{2}u^2, & |u| \le 1, \\ |u| - \frac{1}{2}, & |u| > 1, \end{cases} \qquad \ell_\tau(u) = |\tau - \mathbf{1}\{u < 0\}|\, h_1(u). \tag{A.96}$$

The 100 quantile positions are

$$\tau_j = \frac{j + 1/2}{100}, \qquad j = 0, \ldots, 99. \tag{A.97}$$

For construction $C \in \{\text{same}, \text{ind}, \text{unw}\}$, the jth population critic coordinate and the reported action value are

$$q_{N,C,j} = \arg\min_{q \in \mathbb{R}} \int \ell_{\tau_j}(z - q)\, \nu_N^C(dz), \qquad Q_{N,C} = \frac{1}{100} \sum_{j=0}^{99} q_{N,C,j}. \tag{A.98}$$

In the terminal construction, broadcasting the scalar target over target quantiles and summing the two critic losses only multiplies separable terms by positive constants, so it does not change these coordinatewise minimizers.

Proposition A.11 (Every enumerated quantile minimizer is unique)**.** *For $N \in \{28, 70\}$, every $C \in \{\text{same}, \text{ind}, \text{unw}\}$, and every midpoint τ_j in Eq.* (A.97)*, the minimizer $q_{N,C,j}$ in Eq.* (A.98) *exists and is unique.*

Proof. For a finite positive measure ν, the derivative of its quantile-Huber objective is the continuous, nondecreasing function

$$\Psi_{\tau,\nu}(q) = (1-\tau)\int_{z<q} \min\{q-z, 1\}\,\nu(dz) - \tau\int_{z\geq q} \min\{z-q, 1\}\,\nu(dz). \tag{A.99}$$

All targets in Eqs. (A.93)– (A.95) lie in $[-2, 4/7]$. Each measure assigns positive mass to the real atoms -2, 0, and $4/7$. Because both block sizes are even, the count vector $(N/2, N/2, 0)$ also has positive probability and gives the model atom $\bar{r}_c = -1$. Its mass is positive in every construction: $W_c > 0$ for the same-block measure, $\bar{W}_N > 0$ for the matched-independent measure, and the unweighted coefficient is one.

The four atoms $-2, -1, 0, 4/7$ have positive mass and consecutive gaps no larger than one. Hence every $q \in [-2, 4/7]$ has a positive-mass atom strictly inside its unit Huber window $(q-1, q+1)$. Since $0 < \tau_j < 1$, that atom makes $\Psi_{\tau_j,\nu}$ locally strictly increasing at q. Therefore the derivative is strictly increasing across the entire target interval. It is negative below the minimum target and positive above the maximum target, so it has exactly one zero. Convexity of the objective makes that zero its unique minimizer. □

Proposition A.12 (The same-block root displacement is $O(N^{-1})$)**.** *For the fixed terminal law, real/model mixture, weight, and quantile grid above, there is a finite constant K, independent of N and j, such that for every $N \geq 1$ and $j = 0, \ldots, 99$,*

$$|q_{N,\text{same},j} - q_{N,\text{ind},j}| \leq \frac{K}{N}, \qquad |Q_{N,\text{same}} - Q_{N,\text{ind}}| \leq \frac{K}{N}. \tag{A.100}$$

This upper bound does not assert monotonicity, a matching lower bound, or a leading constant.

Proof. Write $\mu_X = \mathbb{E}[X]$, $\sigma_X^2 = \text{Var}(X)$, and $s_{X,N} = \text{sd}_N(X_{1:N})$ for $X \in \{R, S'\}$. The displayed law is bounded and has $\sigma_R^2 = 18/175 > 0$ and $\sigma_{S'}^2 = 441/625 > 0$. If $v_{X,N} = s_{X,N}^2$, then

$$v_{X,N} - \sigma_X^2 = \frac{1}{N}\sum_{i=1}^{N}\big((X_i - \mu_X)^2 - \sigma_X^2\big) - (\bar{X}_N - \mu_X)^2. \tag{A.101}$$

The iid fourth-moment expansion therefore gives a finite A_X, depending only on the fixed law, for which

$$\mathbb{E}[(v_{X,N} - \sigma_X^2)^2] \leq \frac{A_X}{N}, \qquad \mathbb{E}[(s_{X,N} - \sigma_X)^2] \leq \frac{A_X}{\sigma_X^2 N}. \tag{A.102}$$

For example, one may take $A_X = 2\,\text{Var}((X-\mu_X)^2) + 2\{\mathbb{E}[(X-\mu_X)^4] + 3\sigma_X^4\}$. The second inequality follows from $|s_{X,N} - \sigma_X| = |v_{X,N} - \sigma_X^2|/(s_{X,N} + \sigma_X)$. It consequently remains valid when an empirical standard deviation is exactly zero; no differentiability or samplewise lower bound on $s_{X,N}$ is used.

Let $w_\infty = (1 + \sigma_R + \sigma_{S'})^{-1}$. The map $(a, b) \mapsto (1 + a + b)^{-1}$ is one-Lipschitz in each nonnegative argument, so Eq. (A.102) implies

$$\text{Var}(W_N) \leq \mathbb{E}[(W_N - w_\infty)^2] \leq \frac{C_W}{N}, \quad C_W = 2\left(\frac{A_R}{\sigma_R^2} + \frac{A_{S'}}{\sigma_{S'}^2}\right). \tag{A.103}$$

For fixed q and τ, denote the single-target contribution to Eq. (A.99) by

$$g_{q,\tau}(z) = (1-\tau)\mathbf{1}\{z < q\}\min\{q-z, 1\} - \tau\mathbf{1}\{z \geq q\}\min\{z-q, 1\}. \tag{A.104}$$

As a function of z, this continuous clipped score is one-Lipschitz, uniformly in q and τ. Hence $\text{Var}(g_{q,\tau}(\bar{R}_N)) \leq \text{Var}(\bar{R}_N) = \sigma_R^2/N$. The real halves of the same-block and matched-independent

derivatives cancel, leaving exactly a covariance on the model half. Cauchy–Schwarz and Eq. (A.103) give the uniform bound

$$\left|\Psi_{\tau,\nu_N^{\text{same}}}(q) - \Psi_{\tau,\nu_N^{\text{ind}}}(q)\right| = \frac{1}{2}\left|\text{Cov}\left(W_N, g_{q,\tau}(\bar{R}_N)\right)\right| \tag{A.105}$$

$$\leq \frac{\sigma_R\sqrt{C_W}}{2N}. \tag{A.106}$$

It remains to convert derivative error into root error uniformly over the 100 quantiles. The fixed real half assigns masses $1/100$, $91/200$, and $7/200$ to -2, 0, and $4/7$. For $\tau_j \in [1/200, 199/200]$, the -2 atom supplies derivative slope at least $(1-\tau_j)/100$ on $(-2,-1)$, while the 0 atom supplies slope at least $\tau_j(91/200)$ on $(-1,0)$ and $(1-\tau_j)(91/200)$ on $(0,4/7)$. Thus, for every N and j, the matched-independent derivative is strongly increasing on the full target interval with the common constant

$$\Psi_{\tau_j,\nu_N^{\text{ind}}}(y) - \Psi_{\tau_j,\nu_N^{\text{ind}}}(x) \geq m(y-x), \qquad m = \frac{1}{20000}, \quad -2 \leq x < y \leq \frac{4}{7}. \tag{A.107}$$

The same bound holds for the same-block derivative because the bound comes entirely from the shared real half. Thus both minimizers exist and are unique for every N. Both roots lie in the interior of this interval because each measure retains positive real mass at both endpoints and at interior targets. Applying Eq. (A.107) between the two roots and then Eq. (A.106) yields Eq. (A.100) with $K = \sigma_R\sqrt{C_W}/(2m)$. Averaging the 100 coordinatewise bounds proves the aggregate bound. □

Deterministic enumeration gives the following aggregate critic values. The supplementary implementation reproduces the displayed digits with 70-digit decimal arithmetic and checks each derivative root to tolerance 10^{-60}.

N	Construction C	$Q_{N,C}$	$Q_{N,C} > 0.005$
28	same block	0.009189551762609	yes
28	matched independent	0.001571867243799	no
28	unweighted	0.001119066929582	no
70	same block	0.006492168821341	yes
70	matched independent	0.003096953137488	no
70	unweighted	0.003083085470212	no

At the published block size $N = 70$, the same-block critic value is 0.00649, versus 0.00310 and 0.00308 under independent-block and unweighted losses. Only same-block reuse prefers the stochastic action to the safe payoff 0.005, despite the stochastic action's zero environmental expected return. The same ordering holds at $N = 28$. The independent-block and unweighted roots need not equal one another or the environmental mean. Figure 3 illustrates the dependence control and these decision reversals.

A.10.4 Robustness to perturbations

Proposition A.13 (Open-neighborhood robustness)**.** *Keep the block sizes,* 1:1 *mixture, weight formula, quantile grid, and* $\kappa = 1$ *fixed. Parameterize the three-outcome terminal law by its positive probabilities, reward atoms, next-state atoms, and safe value, using the relative topology of the probability simplex. There is an open neighborhood of the displayed law in which, simultaneously for* $N = 28$ *and* $N = 70$*,*

$$\mathbb{E}[R] < v_{\text{safe}} < Q_{N,\text{same}}, \qquad Q_{N,\text{ind}} < v_{\text{safe}}, \qquad Q_{N,\text{unw}} < v_{\text{safe}}. \tag{A.108}$$

Proof. For fixed N, the set of count vectors is finite. Each multinomial mass, block mean, empirical standard deviation, and positive weight in Eq. (A.92) is a continuous function of the law parameters. The finite-sum derivative in Eq. (A.99) is therefore jointly continuous in the critic coordinate and those parameters.

At the displayed law, the positive-mass atoms used in the proof of Proposition A.11 have consecutive gaps strictly below two, the diameter of a $\kappa = 1$ Huber window. Positivity, their ordering, and these

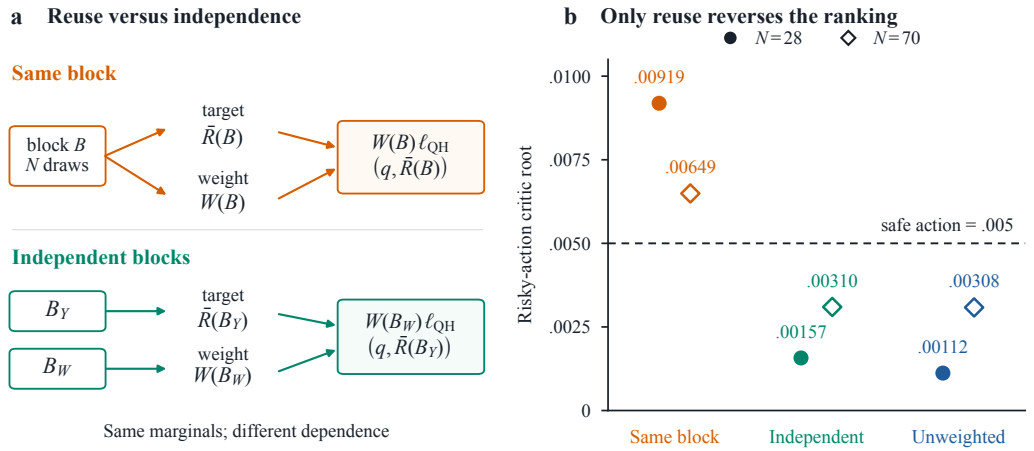

Figure 3: **Matched marginals isolate same-block reuse.** The independent condition preserves target and weight marginals and changes only their dependence. Deterministic enumeration shows that only same-block reuse crosses the safe value at both block sizes considered.

strict gap inequalities persist under sufficiently small perturbations. The corresponding objectives consequently retain unique coordinatewise roots. Joint continuity of the derivative and strict monotonicity then imply continuity of every root $q_{N,C,j}$, and hence of each finite average $Q_{N,C}$. The environmental mean and safe value are also continuous.

At the displayed law, Eq. (A.85) and the enumerated table satisfy every inequality in Eq. (A.108) strictly. There are only two block sizes, three constructions, and 100 coordinates, so the intersection of the finitely many continuity neighborhoods is open and preserves all inequalities simultaneously. □

Even with the terminal law held fixed, the roots certify the explicit common safe-value interval

$$0.003097 < v_{\mathrm{safe}} < 0.006492, \tag{A.109}$$

which contains $v_{\mathrm{safe}} = 0.005$ and lies strictly between every control aggregate and both same-block aggregates. Proposition A.13 is stronger: it also permits small perturbations of the law itself.

Reproduction. The accompanying material reproduces the table by deterministic finite-support enumeration; the supplementary documentation gives invocation and environment details. Proposition A.12 bounds the same-block versus matched-independent displacement by $O(N^{-1})$ for this fixed law, without asserting a matching lower bound. The example proves a decision reversal under the published loss; its prevalence and end-to-end impact are separate empirical questions.

A.11 Matched controls and task-dependent performance

RMSE AUC measures error accumulated throughout learning. Figure 4 separates this measure from endpoint correctness. With exact likelihoods, phase lag recovers the environmental ordering in 28/30 AccumulatedCharge and 30/30 SlipperyBridge seeds, although its trajectory-level RMSE AUC can be worse than a comparator. With frozen neural likelihoods, phase lag preserves the environmental population root but leaves negative finite-run mean gaps in six of the 18 fully trained models (Appendix D). Thus recovering the population root does not by itself guarantee lower error over a finite learning run.

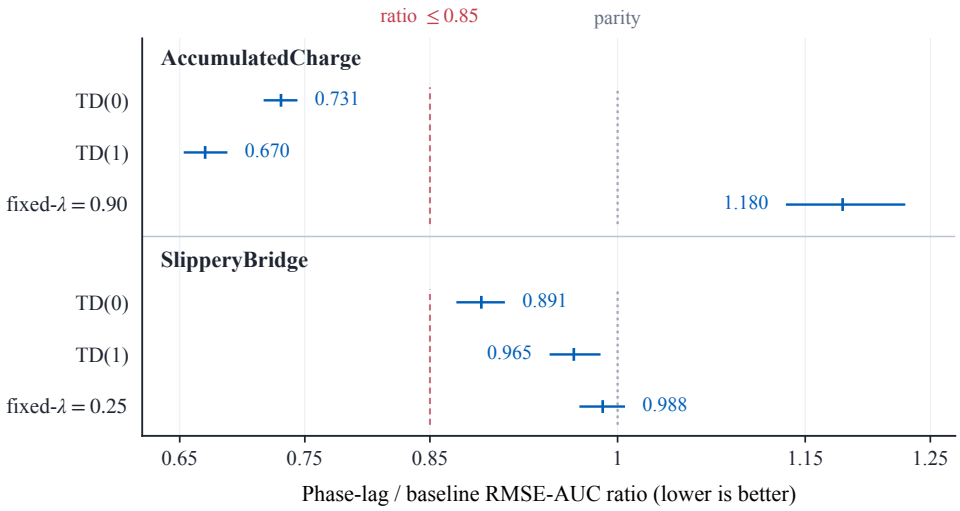

Figure 4: **Correct endpoint ordering need not reduce error throughout learning.** Each point is the ratio of mean RMSE AUC for phase lag to that of the indicated baseline, using 30 matched seeds per task. Horizontal segments are paired percentile-bootstrap 95% intervals from 100,000 resamples. The red dashed line marks the prespecified point-estimate target of a 15% reduction; the dotted line marks equal performance. Phase lag recovers the environmental endpoint ordering in 28/30 AccumulatedCharge and 30/30 SlipperyBridge seeds, but its RMSE AUC is higher than the fixed-λ baseline on AccumulatedCharge.

A.12 What the proofs establish

The appendix proves the local conditional-mean criterion, the induced operator and MRP, contraction and uniqueness, the exact error-transport identity and finite-horizon amplification, backward-view mean-field equivalence, and almost-sure convergence for i.i.d. reset finite-horizon tabular evaluation. SlipperyBridge realizes the transport kernel as an alternative transition law. A zero local weighted residual preserves the task value; otherwise, the error-transport identity determines its effect. The convergence result uses the reset finite-horizon, fixed-multiplier assumptions stated above. WIMLE is analyzed separately as a finite-support critic objective rather than a recursive Bellman update. Supplementary exact-arithmetic calculations verify the identities, sharp examples, finite-horizon expansion, and persistent-randomness example; invocation details are included there.

A.13 Published guarantees and outcome reuse

The two methods use weights for different purposes. Table 3 relates their published guarantees to the dependence studied here. Equality of two update implementations does not by itself identify their shared population fixed point. A weight fixed by state and action also differs from a random weight computed from the same target sample.

A.14 Comparison by information timing

Table 4 compares representative results by the information available to the coefficient multiplying a given innovation. It separates established ingredients—nonanticipation, sample weighting, and substochastic resolvents—from the recursive effect studied here.

Table 3: Published guarantees and the same-sample questions studied here.

Method	Published guarantee	Additional same-sample question
Chunked-TD	The paper defines environmental V_π/Q_π, and Proposition 1 proves equivalence between its offline and online updates (Ramesh et al., 2024).	Both forms use the realized transition likelihood to scale prior traces before those traces multiply the innovation from the same transition. Our analysis identifies the population fixed point shared by these equivalent updates.
WIMLE	The positive-weight argument shows that strictly positive weights fixed by state and action leave the Bellman fixed point unchanged (Aghabozorgi et al., 2026).	In the finite implementation, one prediction block determines both the synthetic target and its random uncertainty weight. We compare same-block reuse with a matched-independent construction to determine whether this dependence changes the critic minimizer.

Table 4: **Comparison at the current innovation.** "Pre-outcome" means measurable before the reward and successor state used by the current innovation are sampled. Transition-dependent coefficients can preserve the intended value when they are applied only to later innovations under the stated assumptions.

Source	Allowed dependence	Preserved value or target	Established consequence	What this work adds
Sutton–Singh; White (Sutton & Singh, 1994; White, 2017)	A realized transition coefficient enters only *later* innovations.	The specified acyclic prediction target or transition-discounted value.	Transition-dependent traces and substochastic operator/resolvent arguments already exist.	The one-innovation timing distinction for probability-based traces.
Retrace; generalized and trajectory-aware traces (Munos et al., 2016; Yu et al., 2018; Daley et al., 2023)	Arbitrary history dependence, but the coefficient on the current innovation is pre-outcome.	Q^π or v^π under the declared task and stated off-policy conditions.	Contraction toward the intended fixed point, stopping-time formulations, and trace resolvents are established.	The fixed point when the coefficient uses the same fresh outcome as the current innovation.
Expected Eligibility Traces (van Hasselt et al., 2021)	A past-measurable trace or its state-conditional projection.	The expected TD update and its fixed point.	Conditional factorization under nonanticipation is established.	A checkable timing criterion for the current innovation.
Prioritized replay, robust losses, intentional updates (Schaul et al., 2016; Patterson et al., 2023; Sharifnassab et al., 2026)	A multiplier may use the sampled item or residual; correction is relative to a declared reference measure.	The reference estimand only with the appropriate correction; otherwise a deliberately reweighted objective/update.	Same-sample and sample-weighting effects can change an estimand.	An exact recursive closure for a product trace.
This work	$G_t(Y_t)$ uses the fresh outcome to rescale the inherited trace before the innovation from that same Y_t is applied.	An alternative Bellman value; the operator admits an explicit induced MRP.	Strict contraction can coexist with a unique alternative fixed point.	Exact operator closure, error transport B, and arbitrary-H decision consequences.

B Exact compatibility of mixed evaluation and policy improvement

This appendix proves Theorem 5, Corollary 6, and Proposition 7, and verifies the example in Table 1. The exact compatibility results concern population evaluators in a known finite discounted environment; the final corollary also allows an estimated environmental kernel.

B.1 Setting and the all-reward quantifier

There are $n \geq 1$ states and m available state–action pairs, with at least one action per state. Fix $0 < \gamma < 1$ and row-stochastic $P, B \in \mathbb{R}^{m \times n}$, independently of reward and policy. Let $\Pi_\pi \in \mathbb{R}^{n \times m}$ represent an arbitrary stationary randomized policy, including deterministic policies. Let $S \in \mathbb{R}^{m \times n}$ repeat a state coordinate across its actions, so $\Pi_\pi S = I_n$. Write

$$E = B - P, \quad X^\pi = \Pi_\pi X, \quad R_X^\pi = (I - \gamma X^\pi)^{-1}, \quad V_X^\pi = R_X^\pi \Pi_\pi r \qquad (X = P, B).$$

The reward r ranges over all deterministic state–action vectors in $\mathbb{R}^m$, and is the same for both models. The true, mixed, and ordinary-model action values are, respectively,

$$Q_P^\pi = r + \gamma P V_P^\pi, \qquad Q_G^\pi = r + \gamma P V_B^\pi, \qquad Q_B^\pi = r + \gamma B V_B^\pi.$$

The middle identity is Eq. (9). Initial actions with zero probability under π are still evaluated. For a fixed transition-only coefficient $g \in [0,1]^{m \times n}$, $B = P \odot g + (I - \mathrm{diag}((P \odot g)\mathbf{1}))P$ is stochastic. The results below also apply to stochastic B not represented by such a coefficient. A coefficient involving the sampled next action, and changing with π, does not in general induce one common B and is not covered by the all-policy test.

Choose a deterministic baseline π_0, with action $a_0(s)$ at each state, and write $\Pi_0 = \Pi_{\pi_0}$. Let $C \in \mathbb{R}^{(m-n) \times m}$ contain $e_{sa}^\mathsf{T} - e_{s,a_0(s)}^\mathsf{T}$ for every nonbaseline action, and set $\mathcal{D}_P = CP$. These rows span every within-state action contrast. Same complete order means that every same-state pair has the same gap sign, including zero; it is stronger than selecting one common maximizing action for one reward.

Lemma B.1 (Order and numerical gaps)**.** *For fixed P, B, γ, π, the following are equivalent: every reward gives the same complete action order; no reward gives a strict reversal; every reward gives $CQ_G^\pi = CQ_P^\pi$; and*

$$\mathcal{D}_P R_B^\pi E^\pi = 0, \qquad E^\pi = \Pi_\pi E. \tag{B.1}$$

Proof. Let $z^\pi = V_B^\pi - V_P^\pi$. Subtracting the Bellman equations gives

$$z^\pi = \gamma R_B^\pi E^\pi V_P^\pi, \qquad CQ_G^\pi - CQ_P^\pi = \gamma^2 \mathcal{D}_P R_B^\pi E^\pi V_P^\pi. \tag{B.2}$$

The map $r \mapsto V_P^\pi$ is onto: for any $v \in \mathbb{R}^n$, take $r_0 = S(I - \gamma P^\pi)v$. Thus numerical gap equality for every reward is equivalent to Eq. (B.1).

If that matrix is nonzero, choose a contrast between a, b at s and v giving a nonzero gap error δ in Eq. (B.2). Under r_0, let the true $a - b$ gap be t. The perturbation

$$h = e_{sa} - \pi(a \mid s) S e_s$$

satisfies $\Pi_\pi h = 0$ and $h(s,a) - h(s,b) = 1$, also when either action has policy probability zero or one. Replace r_0 by $r_0 + (-t - \delta/2)h$. Neither continuation value changes, while the true and mixed gaps become $-\delta/2$ and $\delta/2$. This is a strict reversal. Numerical equality implies order equality and no reversal, which completes the equivalence. □

The independent action-reward coordinates are essential to this argument. It does not establish necessity for state-only rewards or a single specified reward. Unbounded rewards are not essential: a common shift and positive scaling put any finite witness in a fixed interval of positive length without changing its gap signs.

Linearization. Both forms of the resolvent identity yield

$$R_B^\pi - R_P^\pi = \gamma R_B^\pi E^\pi R_P^\pi = \gamma R_P^\pi E^\pi R_B^\pi.$$

Consequently,

$$\mathcal{D}_P R_B^\pi E^\pi = 0 \quad \Longleftrightarrow \quad \mathcal{D}_P R_P^\pi \Pi_\pi E = 0. \tag{B.3}$$

For example, the left condition first implies $\mathcal{D}_P R_B^\pi = \mathcal{D}_P R_P^\pi$; multiply by E^π. The other resolvent identity proves the converse. The right-hand condition is linear in E, with coefficients depending only on P.

B.2 One environmental resolvent gives the all-policy test

Put $R_0 = (I - \gamma\Pi_0 P)^{-1}$, $O = \mathcal{D}_P R_0$, $J = \{s : Oe_s \neq 0\}$, and let $\mathcal{I}$ be the complement of J.

Lemma B.2 (Policy-invariant invisible columns)**.** *For $i \in \mathcal{I}$ and every stationary policy π, $R_P^\pi e_i = R_0 e_i$ and $\mathcal{D}_P R_P^\pi e_i = 0$. Moreover, the zero-column set of $\mathcal{D}_P R_P^\pi$ is exactly $\mathcal{I}$ for every π.*

Proof. Each row of $P^\pi - P^{\pi_0}$ is a linear combination of rows of $\mathcal{D}_P$. Hence $(P^\pi - P^{\pi_0})R_0 e_i = 0$ whenever $Oe_i = 0$. It follows that $(I - \gamma P^\pi)R_0 e_i = e_i$, proving both identities. The same argument works with any stationary reference policy. Interchanging π and π_0 therefore proves the converse zero-column inclusion. □

These columns describe signed discounted-occupancy differences, not ordinary reachability. Related zero-column criteria for decision-irrelevant single-action states appear in Batra (2026, Eqs. (26)–(27)). Here the full action-contrast span is used to construct the mixed all-policy constraint.

Proof of Theorem 5. Let C_J retain the rows of C at states in J, and let $F = ((O\Pi_0)^\mathsf{T}, C_J^\mathsf{T})^\mathsf{T}$, as in Eq. (13). Let $f \in \mathbb{R}^m$ be any reward vector in the environment P. Write $V_f^\pi = R_P^\pi \Pi_\pi f$ for its value. We first establish the scalar certificate

$$\mathcal{D}_P R_P^\pi \Pi_\pi f = 0 \text{ for every } \pi \quad \Longleftrightarrow \quad O\Pi_0 f = 0, \quad C_J f = 0. \tag{B.4}$$

For necessity, set $v_0 = R_0\Pi_0 f$. The baseline condition gives $\mathcal{D}_P v_0 = 0$. Change only the action at s to a, obtaining policy σ. Its Bellman residual at v_0 is

$$e_s\big[f(s,a) - f(s,a_0(s)) + \gamma(P_{sa} - P_{s,a_0(s)})v_0\big] = e_s[f(s,a) - f(s,a_0(s))].$$

Thus $V_f^\sigma - v_0 = R_P^\sigma e_s[f(s,a) - f(s,a_0(s))]$. Both values have zero $\mathcal{D}_P$ output. At $s \in J$, Lemma B.2 says that $\mathcal{D}_P R_P^\sigma e_s \neq 0$; the reward difference must therefore vanish. This proves $C_J f = 0$.

Conversely, suppose the two right-hand conditions hold. For arbitrary, possibly randomized π, the residual at v_0 is

$$b_\pi = \Pi_\pi f - \Pi_0 f + \gamma(P^\pi - P^{\pi_0})v_0 = \Pi_\pi f - \Pi_0 f.$$

It is supported on $\mathcal{I}$, since action rewards agree at every state in J. Lemma B.2 gives $\mathcal{D}_P R_P^\pi b_\pi = 0$, and hence $\mathcal{D}_P V_f^\pi = 0$. Apply Eq. (B.4) to each column of E, then use Eq. (B.3) and Lemma B.1. This proves all equivalences in the theorem. □

Exact construction and its scope. The matrix F has at most $2(m-n)$ rows. Forming it uses one $n \times n$ environmental matrix factorization, the required multiple right-hand sides, and exact zero tests and matrix products. For rational P, B, γ, rational linear algebra has polynomial bit complexity in the dimensions and encoded input length, so construction of F and testing $FE = 0$ are polynomial-time operations. No policy enumeration is required. Equivalently, Eq. (B.4) identifies

$$\operatorname{span}_\pi \operatorname{row}(\mathcal{D}_P R_P^\pi \Pi_\pi) = \operatorname{row}(F).$$

If there are no action contrasts, F is empty and comparison preservation is vacuous. For $n = 1$, stochasticity forces $B = P$.

B.3 Structure and relation to ordinary-model comparisons

Write $\rho = \operatorname{rank}(\mathcal{D}_P)$ and $k_J = \sum_{s\in J}(|A_s| - 1)$, $k_I = \sum_{s\in\mathcal{I}}(|A_s| - 1)$. Then

$$\operatorname{rank}(F) = \rho + k_J, \qquad \dim\ker F = n - \rho + k_I. \tag{B.5}$$

Indeed, $O\Pi_0$ has rank ρ and is supported on baseline-action columns. Each row of C_J has a distinct nonbaseline column, so its k_J independent rows add rank without intersecting the first row space.

Proposition B.3 (Compatible perturbations)**.** *Relative to the chosen baseline, $FE = 0$ if and only if E has the unique decomposition*

$$E = (S - \gamma P)Z + U, \qquad \mathcal{D}_P Z = 0, \tag{B.6}$$

where U is supported only on nonbaseline-action rows at states in $\mathcal{I}$. Here $Z = R_0\Pi_0 E$ and $U = E - (S - \gamma P)Z$.

Proof. The first block of $FE = 0$ gives $\mathcal{D}_P Z = 0$. Consequently $PZ = SP^{\pi_0}Z$, and $(I - \gamma P^{\pi_0})Z = \Pi_0 E$ gives $(S - \gamma P)Z = S\Pi_0 E$: this part repeats the baseline row across actions at each state. The second block of $FE = 0$ makes these repetitions equal to E at every state in J. The difference U therefore has the claimed support. Conversely, the resulting E is annihilated by both blocks of F. Applying Π_0 recovers Z uniquely, then U. □

The row-sum condition separates as $E\mathbf{1} = 0$ if and only if $Z\mathbf{1} = 0$ and $U\mathbf{1} = 0$. For necessity apply Π_0, use $\Pi_0 U = 0$, and invert $I - \gamma P^{\pi_0}$. For full-support P, sufficiently small such perturbations are realized by the legal coefficient $g_{ij} = 1/2 + E_{ij}/P_{ij}$. Thus their local kernel dimension is $(n - \rho + k_I)(n - 1)$. For coefficient-realizable kernels with sparse P, additionally impose $E_{ij} = 0$ when $P_{ij} = 0$; the nullity of this combined linear system gives the local dimension. Small supported perturbations are again realizable by the same formula on positive entries. These are dimensions of kernels, not of unrestricted coefficient parameterizations.

At full action-contrast rank $\rho = n - 1$ and $n > 1$, all states belong to J: otherwise $R_0 e_s$ would be constant, and multiplying by $I - \gamma P^{\pi_0}$ would make e_s constant. Thus Eq. (B.6) reduces to $E = \mathbf{1}_m d^\mathsf{T}$, $d^\mathsf{T}\mathbf{1}_n = 0$. This common-row form need not hold when $\mathrm{rank}(CP) < n - 1$.

When support forces the row-mean correction. Suppose $n > 1$, $\rho = n - 1$, and each successor-state column j has some row i with $P_{ij} = 0$. For a compatible transition-dependent coefficient with fixed row means μ_i, the common-row form gives $E_{ij} = P_{ij}(g_{ij} - \mu_i) = d_j$. Each column's zero entry forces $d_j = 0$. Hence $B = P$ and $g_{ij} = \mu_i$ whenever $P_{ij} > 0$; coefficients on impossible transitions remain unrestricted. Under these conditions, the row-mean coefficient is the only compatible correction on the positive-probability support.

Ordinary comparison preservation. The ordinary class is characterized by the known transformation (Batra, 2026, Theorem 3.4 and Corollary 3.5), with an invertible L:

$$B = (P + \gamma^{-1} S(L - I))L^{-1}, \qquad L\mathbf{1} = \mathbf{1}. \tag{B.7}$$

It gives $(I - \gamma B^\pi)L = I - \gamma P^\pi$ for every randomized policy, and therefore $V_B^\pi = LV_P^\pi$ and $Q_B^\pi - Q_P^\pi = S(L - I)V_P^\pi$. For comparison, setting $Z = (I - L^{-1})/\gamma$ reparameterizes this known class as

$$E = (S - \gamma P)Z = (S - \gamma P)R_0\Pi_0 E. \tag{B.8}$$

Conversely, for stochastic $B = P + E$ satisfying the latter condition, $I - \gamma B^{\pi_0} = (I - \gamma P^{\pi_0})(I - \gamma Z)$ makes $I - \gamma Z$ invertible. Also $Z\mathbf{1} = R_0\Pi_0 E\mathbf{1} = 0$, so $L = (I - \gamma Z)^{-1}$ satisfies $L\mathbf{1} = \mathbf{1}$ and recovers Eq. (B.7). Thus the ordinary class also has an environmental linear test. The mixed class differs in its constraints on Z and its additional U term.

Among already mixed-compatible kernels, ordinary comparison preservation holds precisely when $CE = 0$, since $Q_B^\pi - Q_G^\pi = \gamma E V_B^\pi$ and arbitrary rewards realize every V_B^π. In Eq. (B.6), $C(S - \gamma P)Z = -\gamma \mathcal{D}_P Z = 0$; since U has zero baseline rows, $CU = 0$ is equivalent to $U = 0$. The intersection is therefore exactly

$$E = (S - \gamma P)Z, \qquad \mathcal{D}_P Z = 0.$$

Neither compatibility notion implies the other in general, as the two examples below show.

B.4 The ordinary-equivalent cycle and its exact correction

For the main-text example, take

$$P = \begin{pmatrix} 1/5 & 4/5 \\ 4/5 & 1/5 \\ 3/5 & 2/5 \end{pmatrix}, \quad \gamma = 9/10, \quad L = \begin{pmatrix} 19/20 & 1/20 \\ 1/20 & 19/20 \end{pmatrix}.$$

Equation (B.7) gives

$$B = \begin{pmatrix} 17/162 & 145/162 \\ 125/162 & 37/162 \\ 109/162 & 53/162 \end{pmatrix}, \qquad g^0 = \begin{pmatrix} 2/81 & 401/648 \\ 301/648 & 52/81 \\ 151/243 & 103/324 \end{pmatrix}.$$

The rows correspond to $(x,a),(x,b),(y,c)$. Every entry of g^0 is legal, every environmental row mean is $1/2$, and $B-P=P\odot(g^0-\frac{1}{2})$, so the same coefficient realizes the ordinary-equivalent model used in the mixed evaluation.

For $r=(-1,-27/100,1)^\mathsf{T}$, direct Bellman solution gives

$$V_P^{\pi_a}=(10/17,35/17)^\mathsf{T},\qquad V_P^{\pi_b}=(18/205,671/410)^\mathsf{T}.$$

For any policy write $v=V_P^\pi(x)-V_P^\pi(y)$. The true $b-a$ gap is $73/100+(27/50)v$. Since $V_B^\pi=LV_P^\pi$, its state difference is $(9/10)v$, and the mixed gap is $73/100+(243/500)v$. Substituting the two displayed values yields true gaps $-109/1700,-109/1025$ and mixed gaps $13/850,-931/41000$, respectively. Ordinary B gaps equal the true ones by Eq. (B.7). This proves the strict $\pi_a\to\pi_b\to\pi_a$ cycle and the harmful first switch reported in Table 1. More generally, keep $r_a=-1,r_c=1$ fixed and write $d=r_b-r_a$. Bellman subtraction in P gives state differences $-25/17$ under π_a and $(50/41)(d-2)$ under π_b. Multiplication by L scales each difference by $9/10$, so the two mixed $b-a$ gaps are

$$d-\frac{243}{340},\qquad \frac{653d-486}{410}.$$

They have the strict cycle signs exactly on $243/340<d<486/653$, an open interval containing $73/100$.

For either policy, let $P_\pi=P\Pi_\pi$, $K_\pi=K\Pi_\pi$. Since $D=I/2$ and $K_\pi\mathbf{1}=\mathbf{1}/2$, the nonnegative linear part of the original trace operator has row sum

$$\gamma P_\pi(I-\gamma K_\pi)^{-1}(I-D)\mathbf{1}=\frac{\gamma/2}{1-\gamma/2}\mathbf{1}=\frac{9}{11}\mathbf{1}.$$

Thus $9/11$ is its exact sup-norm contraction modulus.

Proof of Proposition 7. Fix legal g^0, positive row weights ω_i, and environmental means $\mu_i=\sum_j P_{ij}g^0_{ij}$. Under the mean constraints, $E=P\odot(g-\mu\mathbf{1}^\mathsf{T})$ is linear in g. Theorem 5 makes $FE=0$ exactly the universal comparison requirement. Together with the means and $0\le g\le 1$, it defines the nonempty compact polytope in Eq. (15): the row-constant choice $g_{ij}=\mu_i$ is feasible. The objective has positive diagonal Hessian on coordinates with $P_{ij}>0$, so its minimizer is unique on that support. Off-support coordinates affect neither the constraints nor the objective and may retain their original values. This proves the stated closest-coefficient result. □

Exact solution in the cycling example. Here $\operatorname{rank}(CP)=n-1=1$, so compatible kernels have $E=\mathbf{1}d^\mathsf{T}$. With means fixed, the coefficient is $g_{ij}=\mu_i+d_j/P_{ij}$. Its feasible coordinates satisfy

$$\sum_j d_j=0,\quad \ell_j=\max_i(-\mu_iP_{ij})\le d_j\le u_j=\min_i((1-\mu_i)P_{ij}).$$

For unit row weights the objective, apart from a constant, is $\sum_j H_jd_j^2-2b_jd_j$, with $H_j=\sum_i 1/P_{ij}$ and $b_j=\sum_i(g^0_{ij}-\mu_i)$. Here these coefficients and bounds are

$$H=(95/12,35/4)^\mathsf{T},\quad b=(-757/1944,17/216)^\mathsf{T},\quad \ell=(-1/10,-1/10)^\mathsf{T},\quad u=-\ell.$$

The minimizer and a scaled equality multiplier are

$$d^*=(-91/3240,91/3240)^\mathsf{T},\qquad \tau=-433/2592.$$

The multiplier for $\sum_j d_j=0$ is 2τ. Both coordinates are strictly inside their bounds and $H_jd_j^*-b_j+\tau=0$ for each j. These feasibility and stationarity equalities are sufficient KKT conditions for this strictly convex problem. They give

$$g^*=\begin{pmatrix}233/648 & 1387/2592\\ 1205/2592 & 415/648\\ 881/1944 & 739/1296\end{pmatrix}.$$

All rows remain nonconstant with mean $1/2$. The corrected kernel is $P+\mathbf{1}(d^*)^\mathsf{T}$, so every true action gap is recovered for all rewards and policies, while the modulus remains $9/11$. The correction loss is $14797/209952$, compared with $1646/19683$ for the row-mean coefficient. The relative reduction in weighted squared correction distance is $8281/52672$.

B.5 A compatible model with nonconstant value bias

The full-rank common-row restriction is not valid in general. Consider three states, one action at each of states $1, 2$, and actions a, b at state 3. In row order $(1, 2, 3a, 3b)$, take

$$P = \begin{pmatrix} 1/4 & 1/4 & 1/2 \\ 1/4 & 1/4 & 1/2 \\ 1/2 & 1/4 & 1/4 \\ 1/4 & 1/2 & 1/4 \end{pmatrix}, \qquad g = \begin{pmatrix} 1/2 & 1/2 & 1/2 \\ 1/2 & 1/2 & 1/2 \\ 5/8 & 1/4 & 1/2 \\ 1/4 & 5/8 & 1/2 \end{pmatrix}.$$

All means are $1/2$, and the rows of $E = B - P$ are $0, 0, e, -e$, where $e = (1/16, -1/16, 0)$. This is a legal, full-support example with $\operatorname{rank}(CP) = 1 < 2$. For arbitrary $r = (x, y, u, v)^\mathsf{T}$ and any probability of choosing a at state 3, equality of the first two transition rows gives

$$V_P^\pi(1) - V_P^\pi(2) = V_B^\pi(1) - V_B^\pi(2) = x - y.$$

It follows, for every $0 < \gamma < 1$, that

$$\begin{aligned} Q_P^\pi(3, a) - Q_P^\pi(3, b) &= Q_G^\pi(3, a) - Q_G^\pi(3, b) = u - v + \tfrac{\gamma}{4}(x - y), \\ Q_B^\pi(3, a) - Q_B^\pi(3, b) &= u - v + \tfrac{3\gamma}{8}(x - y). \end{aligned}$$

Thus all mixed comparisons are exact, although the rows of E are not all equal. Let π_a select a at state 3 and the sole available action at states $1, 2$. For $\gamma = 1/2$ and $r = (1, 0, -5/32, 0)^\mathsf{T}$, direct solution gives $V_B^{\pi_a} - V_P^{\pi_a} = (1, 1, 3)^\mathsf{T}/72$. The value bias is nonzero and nonconstant. Under every policy, the true and mixed $a - b$ gaps are $-1/32$, whereas the ordinary gap is $1/32$.

With $3a$ as the baseline, state 3 is invisible: $J = \{1, 2\}$ and C_J has no rows. In Eq. (B.6), the repeated baseline part has rows $0, 0, e, e$, and the only nonzero row of U is $U_{3b} = -2e$. This exhibits the extra perturbation allowed at the invisible state 3 but excluded from the intersection of the ordinary and mixed compatibility classes. Together with the ordinary-compatible but mixed-incompatible example above, it shows why the two uses of a continuation model must be distinguished.

B.6 A continuous bound for approximate compatibility

The exact test has a continuous counterpart. Keep the same fixed baseline and canonical contrasts C, and use induced matrix infinity norms. Set $A = O\Pi_0 E$ and

$$\begin{aligned} d_s &= \max_{a \neq a_0(s)} \|E_{sa} - E_{s,a_0(s)}\|_1, \\ \eta &= \|A\|_\infty + \max_j \sum_s |O_{js}| d_s, \qquad \kappa = 1 + \frac{\gamma \|\mathcal{D}_P\|_\infty}{1 - \gamma}. \end{aligned} \tag{B.9}$$

At a single-action state, set $d_s = 0$. Empty maxima and norms are zero.

Proposition B.4 (Uniform comparison error)**.** *For fixed action sets, baseline, and $0 < \gamma < 1$, η is continuous in the stochastic kernels P, B, and $\eta = 0$ exactly when $F(B - P) = 0$. For every reward and stationary randomized policy,*

$$\|C(Q_G^\pi - Q_P^\pi)\|_\infty \leq \frac{\gamma^2 \kappa \eta}{2(1 - \gamma)} \operatorname{span}(r), \qquad \operatorname{span}(r) = \max_i r_i - \min_i r_i. \tag{B.10}$$

This controls comparisons with the baseline action. An arbitrary pair of actions is covered by twice this bound; the extra factor is unnecessary at states with at most two actions.

Proof. Put $T_\pi(s, (s, a)) = \pi(a \mid s)$ for each nonbaseline action, with all other entries zero. Then $\Pi_\pi - \Pi_0 = T_\pi C$ and $\|T_\pi\|_\infty \leq 1$. The resolvent identity gives

$$\begin{aligned} \mathcal{D}_P R_P^\pi &= (I + \gamma \mathcal{D}_P R_P^\pi T_\pi) O, \\ M_\pi := \mathcal{D}_P R_P^\pi \Pi_\pi E &= (I + \gamma \mathcal{D}_P R_P^\pi T_\pi)(A + O T_\pi C E). \end{aligned}$$

Each row of $T_\pi C E$ is a nonnegative combination, of total weight at most one, of the action-difference rows at that state. Its 1-norm is at most d_s. Thus $\|A + O T_\pi C E\|_\infty \leq \eta$. Stochasticity gives $\|R_P^\pi\|_\infty = (1 - \gamma)^{-1}$, so $\|M_\pi\|_\infty \leq \kappa \eta$.

The other resolvent form gives $V_B^\pi - V_P^\pi = \gamma R_P^\pi \Pi_\pi E V_B^\pi$, and therefore $C(Q_G^\pi - Q_P^\pi) = \gamma^2 M_\pi V_B^\pi$. Combining this with $\|V_B^\pi\|_\infty \le \|r\|_\infty/(1-\gamma)$ proves the bound with $\|r\|_\infty$. Subtracting the common reward constant $(\max_i r_i + \min_i r_i)/2$ leaves gap errors unchanged and replaces that norm by $\mathrm{span}(r)/2$.

Finally, $\eta = 0$ holds exactly when $A = 0$ and $d_s = 0$ for every nonzero column of O, which is $O\Pi_0 E = 0$ and $C_J E = 0$. Continuity follows from continuity of the resolvent, finite sums, absolute values, and maxima. Subtracting two baseline gap errors gives the arbitrary-pair bound. □

The residual weights within-state action variation in E by its contribution to discounted comparisons. A column of O approaching zero therefore has a vanishing contribution, without a discrete change in the residual. This population bound uses known kernels. The following extension adds transition-estimation error; numerical evaluation error requires a separate allowance.

Corollary B.5 (Estimated transitions)**.** *Let $P, \widehat{P}$ be stochastic kernels on the same state and action sets. Fix $0 < \gamma < 1$, $r \in \mathbb{R}^m$, and a transition-only coefficient $g \in [0,1]^{m\times n}$ independent of reward and policy. Use the same complete coefficient array at both kernels, including legal entries on transitions absent from $\widehat{P}$. Denote the induced continuation kernels by B and $\widehat{B}$, respectively, and set*

$$S_r = \mathrm{span}(r), \qquad \epsilon = \max_i \|P_i - \widehat{P}_i\|_1, \qquad \ell = \max_i(\max_j g_{ij} - \min_j g_{ij}).$$

For a stationary policy and a same-state action pair, let Δ_P be the mixed action gap minus the environmental action gap. Denote the corresponding difference at $(\widehat{P}, \widehat{B})$ by $\Delta_{\widehat{P}}$. If $|\Delta_{\widehat{P}}| \le \widehat{\tau}$, then

$$\begin{aligned} |\Delta_P| &\le \widehat{\tau} + \epsilon S_r \min\{a_\gamma, b_{\gamma,\ell}\}, \\ a_\gamma &= \frac{\min\{\gamma(2+\gamma), \frac{7}{2}\gamma^2\}}{(1-\gamma)^2}, \qquad b_{\gamma,\ell} = \frac{\gamma^2\ell(3-\gamma+\gamma\ell)}{2(1-\gamma)^3}. \end{aligned} \tag{B.11}$$

Exact fitted compatibility gives $\widehat{\tau} = 0$. Otherwise Proposition B.4, evaluated at the fitted kernels, supplies $\widehat{\tau} = \gamma^2\widehat{\kappa}\widehat{\eta}S_r/[2(1-\gamma)]$ for a baseline comparison. Double only this fitted-residual term for two nonbaseline actions; the model-error term already covers any pair. The bound is uniform over stationary randomized policies whenever the fitted certificate is uniform. If every coefficient row is constant, then $\ell = 0$ and the model-error allowance vanishes, consistently with $B = P$.

Proof. We first establish the row estimates. For probability rows p, q and coefficient row h, write $E_h(p) = p \odot h - (ph^\mathsf{T})p$ and $B_h(p) = p + E_h(p)$. If $\mathrm{span}(h) \le \ell$, then

$$\begin{aligned} \|E_h(p) - E_h(q)\|_1 \le \ell\|p-q\|_1, \qquad \|B_h(p) - B_h(q)\|_1 \le (1+\ell)\|p-q\|_1, \\ \|E_h(p)\|_1 \le \ell/2. \end{aligned} \tag{B.12}$$

Constant h is immediate; otherwise shift and rescale h to $[0,1]$. For a subset A, put $t = p(A)$ and $\mu = ph^\mathsf{T}$. The derivative of $E_h(p)(A)$ has coefficients $h_j(1_{j\in A} - t) - \mu 1_{j\in A}$, lying in $[-\mu, 1-t-\mu]$ on A and $[-t, 0]$ outside. Their combined range is at most one. A zero-sum direction v therefore changes this subset mass by at most $\|v\|_1/2$. Integrating along $[p,q]$, using $\|w\|_1 = 2\max_A |w(A)|$ for zero-sum w, and scaling back proves the first bound. The second follows by addition. Finally, $\sum_j p_j|h_j - \mu| \le 2\mu(1-\mu) \le 1/2$; rescaling proves the last bound.

Fix the policy and suppress its index. Put $d = P_{sa} - P_{sb}$, $\widehat{d} = \widehat{P}_{sa} - \widehat{P}_{sb}$, $z = V_B - V_P$, and $\widehat{z} = V_{\widehat{B}} - V_{\widehat{P}}$. Then

$$\Delta_P - \Delta_{\widehat{P}} = \gamma(d - \widehat{d})\widehat{z} + \gamma d(z - \widehat{z}). \tag{B.13}$$

For a zero-row-sum matrix M, $\|Mv\|_\infty \le \|M\|_\infty \mathrm{span}(v)/2$. Stochasticity gives $\mathrm{span}(R_X^\pi v) \le \mathrm{span}(v)/(1-\gamma)$ and $\mathrm{span}(V_X) \le S_r/(1-\gamma)$. The resolvent identity and Eq. (B.12) consequently give

$$\begin{aligned} \|V_X - V_{\widehat{X}}\|_\infty \le \frac{\gamma\epsilon_X S_r}{2(1-\gamma)^2}, \qquad \epsilon_X = \max_i \|X_i - \widehat{X}_i\|_1, \\ \mathrm{span}(\widehat{z}) \le \min\left\{\frac{2S_r}{1-\gamma}, \frac{\gamma\ell S_r}{2(1-\gamma)^2}\right\}. \end{aligned}$$

Here the second bound uses $\widehat{z} = \gamma R^{\pi}_{\widehat{B}} \Pi_{\pi}(\widehat{B} - \widehat{P})V_{\widehat{P}}$. Since $\epsilon_B \le (1+\ell)\epsilon$, Eq. (B.13) implies

$$|\Delta_P - \Delta_{\widehat{P}}| \le \frac{\gamma \|d - \widehat{d}\|_1}{2} \operatorname{span}(\widehat{z}) + \frac{\gamma^2 \|d\|_1 (2+\ell)\epsilon S_r}{2(1-\gamma)^2}.$$

Using $\|d - \widehat{d}\|_1 \le 2\epsilon$, $\|d\|_1 \le 2$, and $\ell \le 1$ yields $a_\gamma \epsilon S_r$.

For the other bound, put $E = B - P$ and $\widehat{E} = \widehat{B} - \widehat{P}$. Subtracting the bias Bellman equations gives

$$\begin{aligned}(I - \gamma \Pi_\pi B)(z - \widehat{z}) = \gamma \Pi_\pi E(V_P - V_{\widehat{P}}) + \gamma \Pi_\pi (E - \widehat{E}) V_{\widehat{P}} \\ + \gamma \Pi_\pi (B - \widehat{B})\widehat{z}.\end{aligned}$$

Applying the same row and resolvent estimates to these three terms gives

$$\operatorname{span}(z - \widehat{z}) \le \frac{\gamma \ell \epsilon S_r}{(1-\gamma)^2} + \frac{\gamma^2 \ell (2+\ell)\epsilon S_r}{2(1-\gamma)^3}.$$

Substituting this and the second bound on $\operatorname{span}(\widehat{z})$ in Eq. (B.13) yields $b_{\gamma,\ell}\epsilon S_r$. Adding $|\Delta_{\widehat{P}}| \le \widehat{\tau}$ proves the claim. □

The coefficient is held fixed after fitting, so the result requires no continuity of the projection optimizer. A statistical guarantee additionally needs a justified bound on ϵ. If numerical action values have uniform absolute error at most ξ relative to the deployed mixed evaluator, an observed gap exceeding the bound in Eq. (B.11) plus 2ξ in magnitude certifies its environmental sign.

C From policy evaluation to policy improvement

This section develops the self-likelihood cycle family in Theorem C.1, then gives its successor–action extension and sampling protocol. The self-likelihood example uses $g = P$ and $r_b = -1/4$. It is distinct from the ordinary-equivalent model and $r_b = -27/100$ used in Table 1.

C.1 Representation and the two-state cycle

Evaluation operator. Consider a finite MDP with n states, m available state–action pairs, discount $0 < \gamma < 1$, and deterministic rewards $r(s,a)$. Write $P \in \mathbb{R}^{m \times n}$ for the environmental transition matrix and $\Pi_\pi \in \mathbb{R}^{n \times m}$ for a stationary policy, so that $(\Pi_\pi q)(s) = \sum_a \pi(a \mid s) q(s,a)$. Fix a multiplier $g(s,a,s') \in [0,1]$ that does not depend on the policy or the next action. Define

$$K_{sa,s'} = P_{sa,s'} g(s,a,s'), \qquad D = \operatorname{diag}(K\mathbf{1}), \qquad B = K + (I - D)P. \tag{C.1}$$

The matrix B is nonnegative and row-stochastic: its row sums are $D\mathbf{1} + (I - D)\mathbf{1} = \mathbf{1}$. It is independent of π. On the state–action chain, let $P_\pi = P\Pi_\pi$, $K_\pi = K\Pi_\pi$, and $B_\pi = B\Pi_\pi$. Since the reward is determined by the current state–action pair, $u = Dr$. Theorem 2 therefore gives

$$\mathcal{R}^{\pi}_{G} q = r + \gamma P_\pi (I - \gamma K_\pi)^{-1} \big[Dr + (I - D)q \big]. \tag{C.2}$$

This operator is a sup-norm contraction with coefficient at most γ. It evaluates every initial state–action pair, followed by π; the definition does not require the initial action to be selected by π.

First step and continuation. Let q be the unique fixed point of Eq. (C.2), and set $z = (I - \gamma K_\pi)^{-1}[Dr + (I - D)q]$. Then $q = r + \gamma P_\pi z$, and substitution gives

$$z = Dr + (I - D)r + \gamma \big[K_\pi + (I - D)P_\pi \big] z = r + \gamma B_\pi z. \tag{C.3}$$

Thus $z = Q^{\pi}_{B}$, the ordinary action value in the MDP with reward r and transition B. With $V^{\pi}_{B} = \Pi_\pi Q^{\pi}_{B}$, this proves Eq. (9). In particular,

$$Q^{\pi}_{G} = r + \gamma P V^{\pi}_{B}, \qquad Q^{\pi}_{G} - Q^{\pi}_{B} = \gamma (P - B) V^{\pi}_{B}. \tag{C.4}$$

The current action is evaluated through P, whereas its continuation is evaluated through B. In general, $\Pi_\pi Q^{\pi}_{G} \ne V^{\pi}_{B}$, so these are not the ordinary action values of the B-MDP.

Theorem C.1 (A family of strict policy cycles)**.** *Consider states x, y, actions a, b at x, and one action c at y. Write $p_i = P(x \mid i)$ and $\beta_i = B(x \mid i)$. Fix P, g, with $p_b > p_a$, and rewards $r_c > r_a$. There is a nonempty open interval of rewards r_b for which exact evaluation by Q_G^π, followed by greedy improvement, strictly alternates between the two deterministic policies if and only if*

$$\beta_b - \beta_a > p_b - p_a. \tag{C.5}$$

Thus a cycle is possible when the transformed continuation enlarges the difference between the actions' transition probabilities.

Proof. Put $\Delta = r_c - r_a > 0$ and

$$d = r_b - r_a, \qquad \eta = \gamma(p_b - p_a), \qquad h_i = 1 - \gamma(\beta_i - \beta_c), \quad i \in \{a, b\}. \tag{C.6}$$

Both $h_i \geq 1 - \gamma > 0$. Let π_a and π_b select a and b at x, respectively; both select c at y. Subtracting the two state Bellman equations under B gives

$$V_B^{\pi_a}(x) - V_B^{\pi_a}(y) = -\Delta/h_a, \qquad V_B^{\pi_b}(x) - V_B^{\pi_b}(y) = (d - \Delta)/h_b.$$

Equation (C.4) consequently yields

$$\begin{aligned} Q_G^{\pi_a}(x, b) - Q_G^{\pi_a}(x, a) &= d - \eta\Delta/h_a, \\ Q_G^{\pi_b}(x, b) - Q_G^{\pi_b}(x, a) &= (1 + \eta/h_b)d - \eta\Delta/h_b. \end{aligned} \tag{C.7}$$

Strict greedy improvement cycles from π_a to π_b and back exactly when the first gap is positive and the second negative, or

$$\frac{\Delta\eta}{h_a} < d < \frac{\Delta\eta}{h_b + \eta}. \tag{C.8}$$

This interval is nonempty exactly when $h_a > h_b + \eta$, equivalently $\beta_b - \beta_a > p_b - p_a$. Its width is

$$\frac{\Delta\gamma^2(p_b - p_a)[\beta_b - \beta_a - (p_b - p_a)]}{h_a(h_b + \eta)}. \tag{C.9}$$

This proves the stated family condition. At either endpoint of Eq. (C.8), one comparison is tied. If $\gamma = 0$, $p_b = p_a$, or $\Delta = 0$, the two strict inequalities cannot both hold under the remaining ordering assumptions. □

An exact instance. For true successor-state likelihoods, the two realized coefficients are p_i and $1 - p_i$. Equation (C.1) gives $\beta_i = \rho(p_i) = 3p_i^2 - 2p_i^3$. Take

$$\gamma = \frac{9}{10}, \qquad (p_a, p_b, p_c) = \left(\frac{1}{5}, \frac{4}{5}, \frac{3}{5}\right), \qquad (r_a, r_b, r_c) = \left(-1, -\frac{1}{4}, 1\right). \tag{C.10}$$

Then $(\beta_a, \beta_b, \beta_c) = (13, 112, 81)/125$, so the transformed action difference $99/125$ exceeds $p_b - p_a = 3/5$. The reward interval is $675/931 < d < 675/823$, equivalently $-256/931 < r_b < -148/823$, and contains $d = 3/4$. The two gaps are $93/3724 > 0$ and $-231/1942 < 0$, giving $\pi_a \to \pi_b \to \pi_a$. The first change lowers the true value at x from $10/17$ to $10/41$. Direct evaluation of Eq. (C.2) gives contraction coefficients $3168/3991$ and $3420/4391$, both below γ. Strict gaps and positive denominators also preserve the cycle under sufficiently small joint perturbations of these parameters.

No common MDP for these action values. There is no single MDP on these same states and actions, with discount γ and the same correspondence between policies, whose ordinary action values equal both $Q_G^{\pi_a}$ and $Q_G^{\pi_b}$. Otherwise, the positive first gap would make π_b a strict policy improvement over π_a at x, while the negative second gap would make π_a a strict improvement over π_b at x. The ordinary policy improvement theorem gives contradictory strict value inequalities. Equivalently, each improvement residual is positive at x and zero at y; multiplying by the nonnegative resolvent $(I - \gamma P^{\pi'})^{-1}$ preserves strict positivity at x. This conclusion concerns a common MDP across policies. Each individual evaluation operator still has the MRP representation given in the augmented-MRP construction in Appendix A.3.

Predictable coefficients. If $g(s, a, s') = \lambda(s, a)$, then $K = DP$ and $B = P$. Equation (C.4) reduces to $Q_G^\pi = Q^\pi$ for every stationary policy, restoring ordinary policy improvement. This is sufficient but not necessary for a particular task.

C.2 A full-support successor–action instance

To evaluate every action from continuing trajectories, replace the two deterministic policies by ϵ-greedy policies with $\epsilon = 0.1$. Policy π_a selects a at x with probability 0.95; π_b selects it with probability 0.05. Both select c at y. Keep the environment and rewards in Eq. (C.10).

Let X_t denote the sampled state–action pair. For the joint successor–action coefficient,

$$P_\pi((s, a), (s', a')) = P(s' \mid s, a)\pi(a' \mid s'),$$
$$G_t = P_\pi(X_t, X_{t+1}), \qquad K_\pi = P_\pi \odot P_\pi. \tag{C.11}$$

Here $\odot$ denotes entrywise multiplication. This coefficient includes the sampled next action, unlike the transition-only coefficient above. For each frozen policy, Eq. (C.2) still applies with $D_\pi = \mathrm{diag}(K_\pi \mathbf{1})$, but $B_\pi = K_\pi + (I - D_\pi)P_\pi$ need not be induced by a policy-independent state-transition kernel.

Table 5 gives exact rational calculations. All entries of P_π are positive under both policies. Their stationary distributions therefore have full support, and Proposition 4 identifies the stationary mean-update root with Q_G^π in all three coordinates. The deterministic mean drift is stable as well: $\mathrm{diag}(\mu_\pi)(L_\pi - I)$ has nonnegative off-diagonal entries and strictly negative row sums.

Table 5: Exact evaluation with joint successor–action likelihoods. The coordinate order is $(x, a), (x, b), (y, c)$, and the gap is $Q_G^\pi(x, b) - Q_G^\pi(x, a)$. L_π is the linear part of Eq. (C.2).

Policy	μ_π	Exact gap	$\|L_\pi\|_\infty$
π_a	$(57, 3, 77)/137$	$11509461501/735142908668$	$661409019/823123250$
π_b	$(3, 57, 23)/83$	$-6173351643/56083584476$	$400289481/499933250$

The gaps are approximately $+0.0156561$ and -0.1100741, while both contraction coefficients are below 0.804. Selecting the larger action value and forming the next ϵ-greedy policy therefore again alternates between π_a and π_b. In the true environment, the gaps are instead $-60/1333$ and $-60/847$. The change $\pi_a \to \pi_b$ lowers the two true state values by $(345600, 291600)/1129051$, respectively.

The coefficient and innovation match the sampled-SARSA branch of the released Chunked-TD code (Ramesh et al., 2024), commit `1849e14`: `chunkers.py` computes the joint likelihood, `algorithms/tabular_algo.py` applies it before the current innovation, and `scripts/train_acc_charge.py` supplies the sampled next action. Our environment and stagewise policy-improvement schedule are specified here; they are not the released joint model-training schedule.

C.3 Finite sampled evaluation and improvement

Protocol. All settings were fixed before sampling. We use 30 seeds (1909202600–1909202629), both initial policies, and six stages per run. Each stage freezes the policy and exact model for $N =$ 2,000,000 transitions, starting from $q = e = 0$, state x, and an action drawn from the policy. After sampling the next state and action, the pre-update q gives

$$\delta_t = r(X_t) + \gamma q(X_{t+1}) - q(X_t),$$
$$e_t = \phi(X_t) + \gamma G_t e_{t-1}, \qquad q \leftarrow q + \alpha_t e_t \delta_t, \tag{C.12}$$

where $\alpha_t = 0.1(1 + t/1000)^{-0.6}$ and t starts at zero in each stage. The next policy is ϵ-greedy with respect to the arithmetic mean of the last $N/2$ post-update vectors; ties select a. Values, traces, state, and the step-size counter are reset at the next stage. Final iterates are also recorded, but do not determine improvement.

The two controls replace G_t by its conditional mean $d_\pi(X_t) = \sum_j P_\pi(X_t, j)^2$, or by zero for TD(0). The former preserves the coefficient's conditional mean, not its full distribution. Since $K_\pi = D_\pi P_\pi$

for this predictable control, its root is the environmental Q^π; TD(0) has the same root. All other sampling, update, and averaging choices are shared.

Random streams depend on seed, initial policy, and stage, not method. Matching policies share trajectories; differing policies share random numbers. Checkpoints at 10k, 100k, 1M, and 2M record final iterates and second-half means; only the prespecified 2M mean determines improvement. Exact roots and population residuals are computed afterwards, never for stopping or policy selection. Supplementary code provides the configuration, per-seed records, and statistical recalculation. A separate Python implementation checks the compiled sampler step by step on short trajectories.

Policy paths and evaluation error. Same-outcome weighting makes all six policy changes in every seed, from both starting policies. Both controls choose a at every improvement. Counting a seed as successful only when the stated pattern holds from both starts gives $30/30$ for each rule, with a Wilson 95% interval of $[0.8865, 1]$. The six stages and two starts are not treated as additional independent seeds.

First-stage errors separate proximity to an induced root from proximity to the environmental value. At 2M transitions, the mean per-seed $\|\bar{q} - Q^\pi_G\|_\infty$ under same-outcome weighting is 0.006005 for π_a and 0.006118 for π_b. The corresponding distances to Q^π are 0.271019 and 0.892096. The root errors decrease across the four checkpoints from 0.074140 to 0.006005 for π_a, and from 0.090127 to 0.006118 for π_b. The conditional-mean control's mean errors at 2M transitions are 0.007059 and 0.007028; TD(0)'s are 0.006378 and 0.006127, all measured against the environmental root.

Averaging matters at this finite budget. For same-outcome weighting under π_a, the averaged gap is $+0.014364$, with a 95% Student-t interval $[0.013035, 0.015693]$ across 30 seeds (29 degrees of freedom). All 30 averaged gaps are positive. Only 18 of the 30 final-iterate gaps are positive; their mean interval is $[-0.001914, 0.023450]$. Under π_b, the averaged gap is -0.109135, with interval $[-0.110627, -0.107642]$. Policy improvement uses the averaged evaluator.

For an additional accuracy check, contraction gives

$$\|\bar{q} - Q^\pi_G\|_\infty \le \frac{\|\mathcal{R}^\pi_G \bar{q} - \bar{q}\|_\infty}{1 - \|L_\pi\|_\infty}. \tag{C.13}$$

An observed gap exceeding twice this bound in magnitude has the exact root's sign. At the first stage, 7/30 same-outcome averages under π_a and 30/30 under π_b pass; the other π_a signs agree without certification. Appendix C.2 establishes the corresponding exact-evaluation cycle.

D Neural likelihoods after full training

After the released 10,000-episode joint model/value training schedule, frozen likelihoods can still produce different value orderings under the paired update modes. This study uses 18 fits, with training and evaluation seeds distinct from the 1,000-episode studies in Appendix E.

D.1 Training and paired evaluation

We use the released Chunked-TD implementation (Ramesh et al., 2024) at commit `1849e14` on the full grid $H \in \{100, 200\}$, $p \in \{0.2, 0.35, 0.5\}$, with training seeds 4100, 4101, and 4102 in each cell. AccumulatedCharge has ten charge events, each with increment $Z \sim \text{Binomial}(H/10, p)$, and $H + 1$ physical transitions per episode. The two actions receive opposite halves of the centered total charge, with an additional 0.1 for action 1. Their environmental root values are therefore $(0, 0.1)$. The cell $(200, 0.5)$ uses the released launcher's default environment parameters.

Training follows the released schedule: uniform root actions for the first 1,000 episodes, then ϵ-greedy root actions with $\epsilon = 0.1$ through episode 10,000 while both the model and tabular values continue learning. The network has three hidden layers of 256 Tanh units and three categorical state-delta heads, with 11 categories per head at $H = 100$ and 21 at $H = 200$. Adam uses learning rate 10^{-4} and weight decay 10^{-6}. Training uses batches of 128 and a replay buffer of capacity 100,000. The original transition, model-training, likelihood, and value-update order is retained, with model updates every four frames after frame 128. The 10,000-episode snapshot is the primary endpoint; the paired 1,000-episode snapshot is secondary.

For each frozen snapshot, evaluation starts from zero values and uses 30 fresh seeds, 64000–64029, for 100,000 episodes each. Root actions remain uniform, $\gamma = 1$, and the constant value step size is 0.00625. Same-outcome weighting uses the actual transition for both the likelihood and TD innovation. Independent-outcome weighting obtains the likelihood from a fresh conditional environment draw at the actual state, keeping the actual reward and successor in the target. Phase lag applies the preceding transition's likelihood to the current innovation.

Evaluation likelihoods are products of the three softmax probabilities, computed with single-state neural predictions in float32. We use no masking, support renormalization, clipping, or trace truncation. The three modes and two snapshots of a fit share complete actual trajectories. Fits in the same cell share root-action and increment arrays, while retaining their own sampled event times. Thus 180 generated streams supply 1,080 snapshot–seed pairs and 3,240 mode runs. Population roots use the frozen state-dependent likelihoods in Eq. (E.1); all physical transitions remain in finite evaluation.

Intervals are 95% Student-t intervals over each fitted model's 30 evaluation seeds (29 degrees of freedom). Paired contrasts subtract matching-seed values before forming intervals. We do not treat a cell's 90 evaluations as 90 independent fits. Intervals are marginal, not simultaneous over the grid.

D.2 Fit quality and value ordering

Table 6 reports prediction losses at episodes 1,000 and 10,000. The three-head cross-entropy averages over coordinates and the exact environmental occupancy of all $H + 1$ transitions under uniform root actions. Charge NLL and KL divergence from the true Binomial law are also reported conditional on stochastic charge events, whose errors can be obscured by the more numerous deterministic transitions. These are population prediction losses under the stated evaluation law; the training replay distribution changes with the behavior policy. All three measures decrease from 1,000 to 10,000 episodes in all 18 fits. The diagnostic records also retain initialization, input-independent reference losses, and separate root-branch summaries.

Tables 7 and 8 report every primary fit and mode. The same-outcome population gap is negative in 13 of 18 fits. The finite mean gap is negative in 11, with all 30 evaluations negative in ten of those fits. Independent-outcome intervals are entirely positive for 16 fits; the other two, seeds 4100 and 4102 at $(200, 0.2)$, span zero. Phase lag gives mixed finite-run recovery. At the default environment parameters $(200, 0.5)$, two same-outcome fits have positive finite-gap intervals and the third spans zero, although one of their population gaps is negative. This cell therefore does not exhibit a consistent finite-run reversal across fits.

Table 9 compares same-outcome minus independent-outcome gaps across training snapshots. The contrast becomes less negative in all three fits at $(100, 0.2)$, but more negative in all three at $(200, 0.2)$, despite improved prediction losses in both cells. The neural results therefore show no uniform monotone relation.

The snapshots differ in both training exposure and behavior: episode 1,000 ends the uniform phase; episode 10,000 includes 9,000 adaptive episodes. This comparison does not isolate training duration. Independent-outcome weighting also requires conditional environment samples.

Ordering and numerical accuracy. For each frozen evaluator, the gap error is the finite-run gap minus its population gap. Within each fit, we average errors over its 30 evaluation seeds, then take the absolute value. At 100,000 episodes, the median of these 18 values is 0.01359 for same-outcome weighting, 0.04670 for independent-outcome weighting, and 0.07625 for phase lag. From 30,000 to 100,000 episodes, both controls' absolute mean gap errors and gap RMSE decrease in all 18 fits, while their sample standard deviations increase. For same-outcome weighting, these errors measure proximity to a potentially biased fixed point. For the controls, whose population gaps equal the environmental gap, they also measure environmental action-gap error. The checkpoint comparison neither establishes asymptotic convergence nor separates transient error from constant-step-size effects.

D.3 Verification and reproduction

The training and evaluation protocol was fixed and hashed before execution. Training used Python 3.12.13, NumPy 1.26.4, PyTorch 2.13.0, and Gymnasium 1.2.3 on CPU, with one Torch thread per worker. The C++17 evaluation kernel disables fast-math and floating-point contraction.

We checked 18 fits, 72 checkpoint hashes, and training records, and reaggregated diagnostics for 54 snapshots. An independent implementation regenerated 180 streams, checked hashes and root histories for 1,080 final value tables, and recomputed population roots and paired intervals with scalar `math.fsum`. Roots agreed within 6.4×10^{-16} and statistics within 2.3×10^{-16}. The supplementary code and per-seed records reproduce the gap-error statistics without new inference or runs. Checks of models and trajectories also require the archived checkpoints and random inputs.

Table 6: Frozen likelihood quality for every fit at 1,000 and 10,000 training episodes. CE averages the three head cross-entropies over exact uniform-root physical occupancy, including root and terminal transitions. Charge NLL and $\mathrm{KL}(P_{\text{true}} \| P_{\text{model}})$ condition on event transitions and use their physical occupancy weights; KL is computed per state and then averaged, not from an averaged predictive PMF. These population diagnostics are not losses on the adaptively collected training replay. All quantities use natural logarithms.

			All-frame CE		Event charge NLL		Event charge KL	
H	p	Fit seed	1k	10k	1k	10k	1k	10k
100	0.2	4100	0.1166	0.0544	2.8613	1.6341	1.2393	0.0122
100	0.2	4101	0.1204	0.0540	2.8783	1.6308	1.2563	0.0089
100	0.2	4102	0.1236	0.0547	2.9853	1.6412	1.3634	0.0192
100	0.35	4100	0.1401	0.0608	3.5175	1.8322	1.6935	0.0082
100	0.35	4101	0.1368	0.0611	3.3336	1.8460	1.5097	0.0221
100	0.35	4102	0.1482	0.0608	3.6353	1.8336	1.8114	0.0097
100	0.5	4100	0.1394	0.0624	3.5232	1.8886	1.6472	0.0127
100	0.5	4101	0.1371	0.0625	3.3696	1.8906	1.4937	0.0147
100	0.5	4102	0.1514	0.0624	3.6213	1.8852	1.7454	0.0092
200	0.2	4100	0.0863	0.0335	4.2449	2.0051	2.2566	0.0168
200	0.2	4101	0.0787	0.0350	3.7851	2.0359	1.7968	0.0476
200	0.2	4102	0.0763	0.0346	3.8464	2.0518	1.8581	0.0635
200	0.35	4100	0.0893	0.0368	4.4505	2.2063	2.2761	0.0319
200	0.35	4101	0.0811	0.0369	3.9353	2.2089	1.7609	0.0346
200	0.35	4102	0.0801	0.0369	4.0469	2.2119	1.8726	0.0375
200	0.5	4100	0.0866	0.0377	4.2600	2.2576	2.0366	0.0342
200	0.5	4101	0.0791	0.0376	3.8525	2.2558	1.6291	0.0324
200	0.5	4102	0.0801	0.0378	3.9572	2.2633	1.7338	0.0399

Table 7: All nine $H = 100$ fits after 10,000 training episodes, with all three frozen evaluation modes. Population gaps are computed for the frozen likelihoods. Finite gaps are means and 95% Student-t intervals over 30 evaluation seeds, each run for 100,000 episodes; intervals are conditional on one fit, descriptive, and not simultaneous. Positive counts use unrounded per-run gaps > 0. The true environmental gap is $+0.1$.

p	Fit seed	Timing	Population gap	Finite gap [95% CI]	Positive
0.2	4100	Same	-0.2775	-0.2852 $[-0.2966, -0.2737]$	0/30
0.2	4100	Independent	$+0.1000$	$+0.0218$ $[+0.0097, +0.0340]$	23/30
0.2	4100	Phase lag	$+0.1000$	-0.0141 $[-0.0290, +0.0009]$	12/30
0.2	4101	Same	-0.3581	-0.3592 $[-0.3708, -0.3476]$	0/30
0.2	4101	Independent	$+0.1000$	$+0.0226$ $[+0.0102, +0.0350]$	22/30
0.2	4101	Phase lag	$+0.1000$	-0.0161 $[-0.0312, -0.0010]$	12/30
0.2	4102	Same	-0.2353	-0.2463 $[-0.2578, -0.2348]$	0/30
0.2	4102	Independent	$+0.1000$	$+0.0197$ $[+0.0076, +0.0317]$	22/30
0.2	4102	Phase lag	$+0.1000$	-0.0132 $[-0.0281, +0.0018]$	12/30
0.35	4100	Same	-0.0559	-0.0670 $[-0.0786, -0.0555]$	0/30
0.35	4100	Independent	$+0.1000$	$+0.0580$ $[+0.0465, +0.0695]$	28/30
0.35	4100	Phase lag	$+0.1000$	$+0.0320$ $[+0.0178, +0.0462]$	24/30
0.35	4101	Same	-0.1269	-0.1230 $[-0.1346, -0.1113]$	0/30
0.35	4101	Independent	$+0.1000$	$+0.0554$ $[+0.0438, +0.0669]$	28/30
0.35	4101	Phase lag	$+0.1000$	$+0.0274$ $[+0.0131, +0.0418]$	23/30
0.35	4102	Same	-0.0210	-0.0364 $[-0.0479, -0.0248]$	3/30
0.35	4102	Independent	$+0.1000$	$+0.0541$ $[+0.0428, +0.0655]$	28/30
0.35	4102	Phase lag	$+0.1000$	$+0.0312$ $[+0.0169, +0.0454]$	24/30
0.5	4100	Same	$+0.0570$	$+0.0689$ $[+0.0563, +0.0814]$	29/30
0.5	4100	Independent	$+0.1000$	$+0.1052$ $[+0.0926, +0.1178]$	30/30
0.5	4100	Phase lag	$+0.1000$	$+0.0950$ $[+0.0791, +0.1109]$	30/30
0.5	4101	Same	-0.0036	$+0.0203$ $[+0.0076, +0.0330]$	22/30
0.5	4101	Independent	$+0.1000$	$+0.1070$ $[+0.0944, +0.1196]$	30/30
0.5	4101	Phase lag	$+0.1000$	$+0.0931$ $[+0.0773, +0.1089]$	30/30
0.5	4102	Same	$+0.1601$	$+0.1563$ $[+0.1437, +0.1689]$	30/30
0.5	4102	Independent	$+0.1000$	$+0.1073$ $[+0.0947, +0.1199]$	30/30
0.5	4102	Phase lag	$+0.1000$	$+0.1015$ $[+0.0857, +0.1173]$	30/30

Table 8: All nine $H = 200$ fits after 10,000 training episodes, with all three frozen evaluation modes. Population gaps are computed for the frozen likelihoods. Finite gaps are means and 95% Student-t intervals over 30 evaluation seeds, each run for 100,000 episodes; intervals are conditional on one fit, descriptive, and not simultaneous. Positive counts use unrounded per-run gaps > 0. The true environmental gap is $+0.1$.

p	Fit seed	Timing	Population gap	Finite gap [95% CI]	Positive
0.2	4100	Same	-0.1339	-0.1666 $[-0.1803, -0.1528]$	0/30
0.2	4100	Independent	$+0.1000$	-0.0031 $[-0.0166, +0.0104]$	15/30
0.2	4100	Phase lag	$+0.1000$	-0.0411 $[-0.0568, -0.0254]$	4/30
0.2	4101	Same	-0.2872	-0.2464 $[-0.2593, -0.2336]$	0/30
0.2	4101	Independent	$+0.1000$	$+0.0236$ $[+0.0110, +0.0362]$	23/30
0.2	4101	Phase lag	$+0.1000$	-0.0239 $[-0.0386, -0.0091]$	9/30
0.2	4102	Same	-0.2525	-0.2469 $[-0.2607, -0.2331]$	0/30
0.2	4102	Independent	$+0.1000$	$+0.0004$ $[-0.0128, +0.0135]$	16/30
0.2	4102	Phase lag	$+0.1000$	-0.0434 $[-0.0587, -0.0281]$	3/30
0.35	4100	Same	$+0.0409$	$+0.0072$ $[-0.0063, +0.0208]$	19/30
0.35	4100	Independent	$+0.1000$	$+0.0435$ $[+0.0301, +0.0569]$	25/30
0.35	4100	Phase lag	$+0.1000$	$+0.0183$ $[+0.0037, +0.0329]$	22/30
0.35	4101	Same	-0.1686	-0.1416 $[-0.1551, -0.1281]$	0/30
0.35	4101	Independent	$+0.1000$	$+0.0431$ $[+0.0302, +0.0560]$	25/30
0.35	4101	Phase lag	$+0.1000$	$+0.0066$ $[-0.0079, +0.0210]$	21/30
0.35	4102	Same	-0.0896	-0.0696 $[-0.0832, -0.0560]$	0/30
0.35	4102	Independent	$+0.1000$	$+0.0525$ $[+0.0395, +0.0655]$	27/30
0.35	4102	Phase lag	$+0.1000$	$+0.0201$ $[+0.0054, +0.0347]$	22/30
0.5	4100	Same	$+0.1601$	$+0.1638$ $[+0.1502, +0.1774]$	30/30
0.5	4100	Independent	$+0.1000$	$+0.1216$ $[+0.1083, +0.1349]$	30/30
0.5	4100	Phase lag	$+0.1000$	$+0.1064$ $[+0.0920, +0.1208]$	30/30
0.5	4101	Same	-0.0655	$+0.0007$ $[-0.0129, +0.0143]$	14/30
0.5	4101	Independent	$+0.1000$	$+0.1163$ $[+0.1034, +0.1291]$	30/30
0.5	4101	Phase lag	$+0.1000$	$+0.0888$ $[+0.0742, +0.1034]$	29/30
0.5	4102	Same	$+0.0198$	$+0.0790$ $[+0.0653, +0.0927]$	29/30
0.5	4102	Independent	$+0.1000$	$+0.1250$ $[+0.1120, +0.1379]$	30/30
0.5	4102	Phase lag	$+0.1000$	$+0.1033$ $[+0.0888, +0.1178]$	30/30

Table 9: Same-outcome minus independent-outcome finite-gap contrasts for all 18 fits. Each cell gives a paired mean and 95% Student-t interval across the same 30 evaluation seeds at 100,000 evaluation episodes. The last column first takes the 10k-minus-1k contrast difference within each seed and then forms its interval. Intervals condition on one fit and are not simultaneous. The checkpoints change both training budget and behavior policy, so this is not a pure training-duration causal comparison. Complete phase-lag contrasts are retained in the accompanying CSV.

H	p	Fit seed	Contrast at 1k	Contrast at 10k	Paired change
			-0.8991	-0.3070	$+0.5921$
100	0.2	4100	$[-0.9038, -0.8943]$	$[-0.3091, -0.3049]$	$[+0.5871, +0.5970]$
			-0.9059	-0.3819	$+0.5241$
100	0.2	4101	$[-0.9106, -0.9012]$	$[-0.3843, -0.3795]$	$[+0.5182, +0.5299]$
			-0.9051	-0.2659	$+0.6391$
100	0.2	4102	$[-0.9104, -0.8997]$	$[-0.2682, -0.2637]$	$[+0.6331, +0.6451]$
			-0.1854	-0.1250	$+0.0604$
100	0.35	4100	$[-0.1872, -0.1837]$	$[-0.1272, -0.1228]$	$[+0.0582, +0.0626]$
			-0.1878	-0.1783	$+0.0095$
100	0.35	4101	$[-0.1893, -0.1864]$	$[-0.1803, -0.1763]$	$[+0.0072, +0.0118]$
			-0.1782	-0.0905	$+0.0877$
100	0.35	4102	$[-0.1797, -0.1767]$	$[-0.0930, -0.0879]$	$[+0.0847, +0.0908]$
			-0.0355	-0.0364	-0.0008
100	0.5	4100	$[-0.0365, -0.0346]$	$[-0.0385, -0.0342]$	$[-0.0025, +0.0008]$
			-0.0331	-0.0867	-0.0536
100	0.5	4101	$[-0.0338, -0.0324]$	$[-0.0890, -0.0843]$	$[-0.0556, -0.0515]$
			-0.0281	$+0.0490$	$+0.0771$
100	0.5	4102	$[-0.0288, -0.0275]$	$[+0.0467, +0.0514]$	$[+0.0750, +0.0793]$
			-0.1371	-0.1635	-0.0264
200	0.2	4100	$[-0.1385, -0.1357]$	$[-0.1658, -0.1612]$	$[-0.0291, -0.0237]$
			-0.1442	-0.2700	-0.1258
200	0.2	4101	$[-0.1455, -0.1428]$	$[-0.2721, -0.2680]$	$[-0.1283, -0.1233]$
			-0.1691	-0.2472	-0.0781
200	0.2	4102	$[-0.1707, -0.1676]$	$[-0.2495, -0.2449]$	$[-0.0809, -0.0753]$
			-0.0053	-0.0363	-0.0310
200	0.35	4100	$[-0.0056, -0.0050]$	$[-0.0384, -0.0342]$	$[-0.0330, -0.0290]$
			-0.0088	-0.1847	-0.1759
200	0.35	4101	$[-0.0090, -0.0085]$	$[-0.1867, -0.1827]$	$[-0.1779, -0.1740]$
			-0.0196	-0.1221	-0.1025
200	0.35	4102	$[-0.0199, -0.0193]$	$[-0.1238, -0.1204]$	$[-0.1041, -0.1009]$
			$+0.0009$	$+0.0422$	$+0.0413$
200	0.5	4100	$[+0.0006, +0.0013]$	$[+0.0399, +0.0445]$	$[+0.0392, +0.0434]$
			$+0.0032$	-0.1155	-0.1187
200	0.5	4101	$[+0.0029, +0.0035]$	$[-0.1176, -0.1135]$	$[-0.1206, -0.1169]$
			-0.0083	-0.0460	-0.0376
200	0.5	4102	$[-0.0086, -0.0080]$	$[-0.0479, -0.0440]$	$[-0.0394, -0.0359]$

E Neural likelihoods after 1,000 training episodes

These studies use models trained for the released initial 1,000-episode uniform-root phase. Their training and evaluation seeds differ from those of the 10,000-episode study in Appendix D, and their model counts are reported separately. In each study, the model is frozen before fresh tabular evaluation, so changing the weight's input changes neither the fitted model nor the actual trajectory.

E.1 Training and frozen population calculations

We use the released Chunked-TD source at commit `1849e14`; the supplementary source manifest records the full hash. The initial study uses $H = 100$, ten accumulation events, bonus 0.1, and $p \in \{0.2, 0.5\}$. At each event, the charge increment is $Z \sim \text{Binomial}(10, p)$. There are $H+1 = 101$ transitions per episode, including the deterministic launch and terminal update. This is a controlled configuration of the released environment; the launcher defaults are $H = 200$ and $p = 0.5$.

For each probability, training seeds 3100, 3101, and 3102 produce three model instances. Each is trained for 1,000 episodes with the released initial uniform root-action policy. We retain the three hidden layers of 256 Tanh units and the three categorical state-delta heads, each with 11 categories. The transition likelihood is the product of the three head probabilities; we neither mask the outputs nor renormalize their legal-support mass. Adam uses learning rate 10^{-4} and weight decay 10^{-6}, with batch size 128 and replay capacity 100,000. The released call order is retained: add the transition, train when scheduled, compute its likelihood, then apply the value update. Training occurs every four frames after frame 128. Each run contains 101,000 transitions and 25,217 optimizer steps.

We save the initialization and episodes 100, 300, and 1,000, retaining all 24 snapshots. At each snapshot, we enumerate all 920 reachable stochastic event states and their 10,120 possible successors. The neural likelihood $g(x, z)$ depends on the full state, including branch and accumulated charge; it is not reduced to the state-independent $g(z)$ of Proposition A.6. At an event state x, the transport kernel is

$$B_x(z) = p_z\big(g(x, z) + 1 - d_x\big), \qquad d_x = \sum_z p_z g(x, z), \tag{E.1}$$

where p_z is the environmental increment probability. Direct backward value recursion and a separate residual-transport recursion agree to 2.2×10^{-14} over these snapshots. Deterministic transitions have $B = P$, regardless of their learned likelihood. They can therefore be collapsed for this population calculation, but remain in every finite run.

At $p = 0.2$, all three final frozen models have negative population gaps: -1.32895, -1.71701, and -1.25610. Their environmental gap is $+0.1$. At $p = 0.5$, the final population gaps are $+0.07303$, $+0.07581$, and $+0.07593$. Exact likelihoods cancel the local residual at this symmetric probability; finite neural fits need not retain that symmetry. The retained likelihood diagnostics include charge-only and full-delta negative log-likelihood. They do not establish a monotone accuracy–reversal relation.

E.2 Paired finite runs

We freeze each of the six final models and evaluate it on 30 new seeds, 61000–61029, for 100,000 episodes per update mode. Values start at zero, $\gamma = 1$, and the constant step size is the released 0.00625. Root actions remain uniformly random throughout evaluation. These runs measure learned value ordering under a fixed policy, not the return of an adaptive control policy.

Within each model, all three modes share the complete actual trajectories and root actions. Same-outcome weighting uses $g(x, Z)$ on the current innovation. Independent-outcome weighting uses $g(x, \widetilde{Z})$, where $\widetilde{Z}$ is an additional independent draw from the true conditional environment law at that same state; the target still uses Z. Phase lag first applies the preceding transition's weight to the current innovation, including the weight from the root transition. The independent control isolates dependence by using an extra conditional sample for the weight.

Table 10 reports every final model and mode. Intervals are marginal 95% Student-t intervals over the 30 evaluation seeds for a fixed model, with 29 degrees of freedom. Models A, B, and C correspond to training seeds 3100, 3101, and 3102. There are three model fits per probability, not 90 independent

fits. Evaluation statistics for different fitted models are dependent because they share evaluation streams. We therefore do not pool them to construct an overall confidence interval.

At $p = 0.2$, each same-outcome model reverses the ordering in 30/30 runs, whereas its independent-outcome counterpart gives the environmental sign in 30/30 runs. The per-model paired mean differences, same-outcome minus independent-outcome, are -0.8696, -0.8904, and -0.8379, with respective 95% intervals $[-0.8735, -0.8658]$, $[-0.8955, -0.8853]$, and $[-0.8417, -0.8340]$. Phase lag gives mixed finite-run recovery. At $p = 0.5$, every mode retains the correct sign. Across both probability settings, none of the 18 model–mode mean intervals contains its corresponding population gap. Correct ordering is therefore distinct from numerical agreement with a fixed point. The constant-step-size experiment does not instantiate the decreasing-step-size convergence theorem.

E.3 Implementation and numerical checks

The evaluator caches the frozen model's probabilities and stores eligibility traces sparsely, without truncation. Full value tables and episode-level outputs match the unchanged released dense updater for the first 200 episodes of every model. One complete 100,000-episode evaluation per model was repeated with identical value tables and all saved checkpoints. All actual and independent random inputs were regenerated from their seeds and matched exactly. The six model parameter hashes match the training snapshots, and all cached probabilities were regenerated.

The neural network computes probabilities in float32; population and value calculations use float64. Batched and single-state neural forward passes differ by at most 5.67×10^{-7} in the resulting weight. No fused multiply-add or fast-math option is used in the compiled evaluator. Tests cover terminal handling, dense/sparse equality, independent random streams, the all-one trace, zero trace, paired statistics, and a complete-path expected-update calculation. At the final mean value tables the expected update remains nonzero; at the environmental value the two decoupled controls have zero expected update to numerical precision.

The archived study records retain the protocol, all model checkpoints, training traces, evaluation inputs, value tables, and seven evaluation checkpoints: 0, 100, 1,000, 5,000, 10,000, 30,000, and 100,000 episodes. Repeat executions are verification runs, not additional statistical samples.

E.4 Prespecified configuration range

After the initial $H = 100$ study, we fixed a grid with $H \in \{100, 200\}$ and $p \in \{0.2, 0.35, 0.5\}$, three training seeds (3100, 3101, 3102) per configuration, and 1,000 training episodes per model. The six existing $H = 100$, $p \in \{0.2, 0.5\}$ models are reused with verified parameter hashes; the other twelve are newly trained. The grid, seeds, budgets, and reporting rules were fixed before obtaining these new results. All 18 models are evaluated on new seeds 62000–62029 under all three update modes, with the same constant step size, zero initialization, uniform root policy, and 100,000-episode budget as the initial study. Every configuration and run is retained.

The number of accumulation events remains ten. At $H = 200$, increments are Binomial$(20, p)$, each categorical head has 21 outputs, and an episode contains 201 transitions. Each model then sees 201,000 training transitions and 50,217 optimizer steps, compared with 101,000 transitions and 25,217 steps at $H = 100$. Thus this grid varies environment configurations, not horizon alone. It includes the released launcher's default environment parameters $(H, p) = (200, 0.5)$ while retaining the fixed-policy, frozen-model evaluation design.

Each configuration has 30 generated evaluation streams, shared across its three fitted models. There are 180 such streams, 540 model–stream pairs, and 1,620 method runs. Within each model–stream pair, the three modes use identical actual trajectories; the independent mode uses a separate conditional environment draw only for the weight. Statistical intervals remain conditional on each fitted model. Shared seeds across models and configurations are not counted as independent model replications.

Table 10: Frozen neural likelihoods: initial paired study ($H = 100$). The gap is $Q(1) - Q(0)$; its environmental value is $+0.1$. Each row uses 30 evaluation seeds for one fixed model. Correct counts refer to the environmental sign.

p	Fit	Weight timing	Population gap	Mean gap [95% CI]	Correct
0.2	A	Same outcome	-1.3289	-0.7932 $[-0.8025, -0.7839]$	$0/30$
0.2	A	Independent	$+0.1000$	$+0.0764$ $[+0.0659, +0.0869]$	$30/30$
0.2	A	Phase lag	$+0.1000$	$+0.0021$ $[-0.0093, +0.0134]$	$15/30$
0.2	B	Same outcome	-1.7170	-0.7628 $[-0.7695, -0.7560]$	$0/30$
0.2	B	Independent	$+0.1000$	$+0.1277$ $[+0.1182, +0.1372]$	$30/30$
0.2	B	Phase lag	$+0.1000$	$+0.0324$ $[+0.0228, +0.0419]$	$26/30$
0.2	C	Same outcome	-1.2561	-0.7672 $[-0.7755, -0.7589]$	$0/30$
0.2	C	Independent	$+0.1000$	$+0.0707$ $[+0.0605, +0.0808]$	$30/30$
0.2	C	Phase lag	$+0.1000$	-0.0013 $[-0.0119, +0.0094]$	$14/30$
0.5	A	Same outcome	$+0.0730$	$+0.3573$ $[+0.3484, +0.3662]$	$30/30$
0.5	A	Independent	$+0.1000$	$+0.3900$ $[+0.3813, +0.3987]$	$30/30$
0.5	A	Phase lag	$+0.1000$	$+0.3590$ $[+0.3498, +0.3681]$	$30/30$
0.5	B	Same outcome	$+0.0758$	$+0.4190$ $[+0.4111, +0.4270]$	$30/30$
0.5	B	Independent	$+0.1000$	$+0.4340$ $[+0.4261, +0.4418]$	$30/30$
0.5	B	Phase lag	$+0.1000$	$+0.3885$ $[+0.3803, +0.3967]$	$30/30$
0.5	C	Same outcome	$+0.0759$	$+0.3917$ $[+0.3842, +0.3992]$	$30/30$
0.5	C	Independent	$+0.1000$	$+0.4195$ $[+0.4121, +0.4270]$	$30/30$
0.5	C	Phase lag	$+0.1000$	$+0.3814$ $[+0.3738, +0.3891]$	$30/30$

Population and finite-run results. All 18 initialized models have positive population gaps. After training, all three models at each of $(H, p) = (100, 0.2)$, $(100, 0.35)$, and $(200, 0.2)$ have negative population gaps. All three at each remaining configuration retain positive gaps. At the default environment parameters $(200, 0.5)$, the final gaps are $+0.09487$, $+0.09871$, and $+0.07298$. These are finite-support population calculations for the actual frozen state-dependent model, not extrapolations from its learning curve.

Tables 11 and 12 report all final finite-run results. Only $(100, 0.2)$ has wrong same-outcome ordering at this budget, with 0/30 correct evaluations for each model. Independent-outcome weighting gives 30/30 for each, while phase lag gives 19/30, 26/30, and 17/30 on the new evaluation seeds. Every mode at every other configuration gives 30/30 correct signs. In particular, the newly trained models at $(100, 0.35)$ and $(200, 0.2)$ have reversed population roots but retain correct finite-run ordering. The twelve new fits add no finite-run reversal cases at this budget.

All 54 model–mode mean intervals exclude their respective population gaps. The mean value tables also have nonzero expected updates. A shifted population root and a finite-budget decision effect are therefore distinct outcomes. The exact-likelihood controls cancel the local residual at $p = 0.5$ for both horizons; neural fits need not preserve this cancellation.

We also evaluate prediction loss over the exact environmental transition distribution under the same uniform-root policy. Cross-entropy, averaged over all transitions and the three prediction coordinates, decreases from initialization to episode 1,000 in all 18 models. At stochastic accumulation events, however, charge-only negative log-likelihood increases in 16. All 18 final models have lower losses on both measures than a reference that ignores state and action and uses each coordinate's true full-frame marginal distribution. The fit diagnostics give the complete per-model losses and protocol. These neural results do not extend the consistent-Binomial-fit accuracy result to neural training.

Numerical consistency and verification. For the population-root calculations and finite-run evaluation, we use single-state neural forward passes for all 36 initial and final snapshots and all 18 evaluation caches. These predictions match the released single-state calculation exactly on every transition checked. Batched forward passes differ by up to 3.25×10^{-6} in the weight over the full grid. The maximum scalar-versus-batched population-root difference is 1.14×10^{-7}.

All 36 checkpoint payloads were reconstructed, all 49,320 stochastic-state records and 18 caches recomputed, and all 180 input streams regenerated. Each model's first evaluation seed was repeated for the full 100,000 episodes with identical value tables and checkpoint outputs. Its first 200 episodes

also matched the unchanged dense updater in the full value table and every episode-level root record. Independent recalculation from 11,340 seed–method–checkpoint records reproduced the final statistics and paired intervals to numerical precision. The archived study records retain the complete initialization and final diagnostics, intermediate evaluation checkpoints, numerical implementation record, and verification scripts.

Table 11: Neural likelihoods at $H = 100$: population targets and finite-run gaps after 1,000 training and 100,000 evaluation episodes. Intervals use 30 evaluation seeds conditional on each fitted model. A, B, and C correspond to training seeds 3100, 3101, and 3102. Correct counts use the environmental sign.

p	Fit	Weight timing	Population gap	Mean gap [95% CI]	Correct
0.2	A	Same outcome	-1.3289	-0.7916 $[-0.8009, -0.7823]$	$0/30$
0.2	A	Independent	$+0.1000$	$+0.0766$ $[+0.0655, +0.0876]$	$30/30$
0.2	A	Phase lag	$+0.1000$	$+0.0003$ $[-0.0114, +0.0121]$	$19/30$
0.2	B	Same outcome	-1.7170	-0.7624 $[-0.7688, -0.7559]$	$0/30$
0.2	B	Independent	$+0.1000$	$+0.1268$ $[+0.1177, +0.1358]$	$30/30$
0.2	B	Phase lag	$+0.1000$	$+0.0307$ $[+0.0211, +0.0403]$	$26/30$
0.2	C	Same outcome	-1.2561	-0.7678 $[-0.7757, -0.7600]$	$0/30$
0.2	C	Independent	$+0.1000$	$+0.0709$ $[+0.0613, +0.0804]$	$30/30$
0.2	C	Phase lag	$+0.1000$	-0.0032 $[-0.0136, +0.0073]$	$17/30$
0.35	A	Same outcome	-0.2434	$+0.0578$ $[+0.0491, +0.0664]$	$30/30$
0.35	A	Independent	$+0.1000$	$+0.2432$ $[+0.2344, +0.2520]$	$30/30$
0.35	A	Phase lag	$+0.1000$	$+0.2022$ $[+0.1926, +0.2118]$	$30/30$
0.35	B	Same outcome	-0.3027	$+0.1271$ $[+0.1199, +0.1344]$	$30/30$
0.35	B	Independent	$+0.1000$	$+0.2998$ $[+0.2924, +0.3073]$	$30/30$
0.35	B	Phase lag	$+0.1000$	$+0.2451$ $[+0.2372, +0.2530]$	$30/30$
0.35	C	Same outcome	-0.2251	$+0.0961$ $[+0.0886, +0.1036]$	$30/30$
0.35	C	Independent	$+0.1000$	$+0.2695$ $[+0.2619, +0.2772]$	$30/30$
0.35	C	Phase lag	$+0.1000$	$+0.2259$ $[+0.2179, +0.2338]$	$30/30$
0.5	A	Same outcome	$+0.0730$	$+0.3581$ $[+0.3488, +0.3674]$	$30/30$
0.5	A	Independent	$+0.1000$	$+0.3903$ $[+0.3810, +0.3996]$	$30/30$
0.5	A	Phase lag	$+0.1000$	$+0.3569$ $[+0.3473, +0.3666]$	$30/30$
0.5	B	Same outcome	$+0.0758$	$+0.4177$ $[+0.4096, +0.4258]$	$30/30$
0.5	B	Independent	$+0.1000$	$+0.4325$ $[+0.4243, +0.4406]$	$30/30$
0.5	B	Phase lag	$+0.1000$	$+0.3854$ $[+0.3769, +0.3938]$	$30/30$
0.5	C	Same outcome	$+0.0759$	$+0.3908$ $[+0.3827, +0.3989]$	$30/30$
0.5	C	Independent	$+0.1000$	$+0.4188$ $[+0.4108, +0.4269]$	$30/30$
0.5	C	Phase lag	$+0.1000$	$+0.3799$ $[+0.3719, +0.3878]$	$30/30$

Table 12: Neural likelihoods at $H = 200$: population targets and finite-run gaps after 1,000 training and 100,000 evaluation episodes. Intervals use 30 evaluation seeds conditional on each fitted model. A, B, and C correspond to training seeds 3100, 3101, and 3102. Correct counts use the environmental sign.

p	Fit	Weight timing	Population gap	Mean gap [95% CI]	Correct
0.2	A	Same outcome	-0.3364	$+0.1316$ $[+0.1246, +0.1386]$	30/30
0.2	A	Independent	$+0.1000$	$+0.2894$ $[+0.2822, +0.2966]$	30/30
0.2	A	Phase lag	$+0.1000$	$+0.2218$ $[+0.2145, +0.2291]$	30/30
0.2	B	Same outcome	-0.3241	$+0.1711$ $[+0.1655, +0.1766]$	30/30
0.2	B	Independent	$+0.1000$	$+0.2939$ $[+0.2882, +0.2997]$	30/30
0.2	B	Phase lag	$+0.1000$	$+0.2216$ $[+0.2159, +0.2273]$	30/30
0.2	C	Same outcome	-0.3214	$+0.0956$ $[+0.0886, +0.1026]$	30/30
0.2	C	Independent	$+0.1000$	$+0.2512$ $[+0.2439, +0.2585]$	30/30
0.2	C	Phase lag	$+0.1000$	$+0.1892$ $[+0.1819, +0.1965]$	30/30
0.35	A	Same outcome	$+0.0686$	$+0.5413$ $[+0.5340, +0.5485]$	30/30
0.35	A	Independent	$+0.1000$	$+0.5551$ $[+0.5478, +0.5624]$	30/30
0.35	A	Phase lag	$+0.1000$	$+0.4613$ $[+0.4542, +0.4684]$	30/30
0.35	B	Same outcome	$+0.0691$	$+0.5794$ $[+0.5732, +0.5856]$	30/30
0.35	B	Independent	$+0.1000$	$+0.5882$ $[+0.5819, +0.5945]$	30/30
0.35	B	Phase lag	$+0.1000$	$+0.4754$ $[+0.4694, +0.4813]$	30/30
0.35	C	Same outcome	$+0.0394$	$+0.3944$ $[+0.3868, +0.4021]$	30/30
0.35	C	Independent	$+0.1000$	$+0.4230$ $[+0.4154, +0.4307]$	30/30
0.35	C	Phase lag	$+0.1000$	$+0.3501$ $[+0.3427, +0.3575]$	30/30
0.5	A	Same outcome	$+0.0949$	$+0.7494$ $[+0.7420, +0.7568]$	30/30
0.5	A	Independent	$+0.1000$	$+0.7561$ $[+0.7487, +0.7635]$	30/30
0.5	A	Phase lag	$+0.1000$	$+0.6319$ $[+0.6246, +0.6392]$	30/30
0.5	B	Same outcome	$+0.0987$	$+0.7534$ $[+0.7471, +0.7597]$	30/30
0.5	B	Independent	$+0.1000$	$+0.7553$ $[+0.7489, +0.7616]$	30/30
0.5	B	Phase lag	$+0.1000$	$+0.6107$ $[+0.6047, +0.6168]$	30/30
0.5	C	Same outcome	$+0.0730$	$+0.5806$ $[+0.5726, +0.5887]$	30/30
0.5	C	Independent	$+0.1000$	$+0.5992$ $[+0.5912, +0.6072]$	30/30
0.5	C	Phase lag	$+0.1000$	$+0.5086$ $[+0.5007, +0.5164]$	30/30

F Policy-dependent coefficients for selected action comparisons

This appendix develops the selected-pair result for finite discounted MDPs with arbitrary finite action sets. Rewards are deterministic functions of the current state–action pair. The policy and coefficient are fixed during each evaluation; both may change between policy updates.

F.1 Positive scaling and a realizable coefficient

Let $T = P\Pi_\pi$, $R_T = (I - \gamma T)^{-1}$, and $0 \le \gamma < 1$. Choose distinct actions a, b at state x, write $p_a = \pi(a \mid x)$ and $p_b = \pi(b \mid x)$, and let $c^\top = e_{(x,b)}^\top - e_{(x,a)}^\top$. Other actions at x are allowed, so $p_a + p_b \le 1$. Theorem 8 concerns any stochastic continuation matrix $A = T + wc^\top$, not only those obtained from a bounded coefficient.

Proof of Theorem 8. Nonnegativity of the two modified columns implies

$$-p_b P(i,x) \le w_i \le p_a P(i,x) \quad \text{for every current state–action pair } i. \tag{F.1}$$

Put $R_A = (I - \gamma A)^{-1}$. The determinant lemma gives

$$d := 1 - \gamma c^\top R_T w = \frac{\det(I - \gamma A)}{\det(I - \gamma T)} > 0. \tag{F.2}$$

To see the sign, for any stochastic matrix M, $\det(I - tM)$ is continuous and nonzero on $0 \le t < 1$, since $\rho(tM) = t < 1$. It equals one at $t = 0$, so remains positive. The rank-one inverse identity is therefore applicable and yields $c^\top R_A = c^\top R_T / d$. Moreover,

$$\begin{aligned} c^\top [I + \gamma T R_A] &= (1 - \gamma c^\top w) c^\top R_A, \\ c^\top w = w_{(x,b)} - w_{(x,a)} &\le p_a P((x,b),x) + p_b P((x,a),x) \le 1. \end{aligned} \tag{F.3}$$

The first identity uses $T = A - wc^\top$; the second follows from Eq. (F.1). Thus the numerator $1 - \gamma c^\top w$ is at least $1 - \gamma > 0$. Multiplying by any reward vector r proves Eq. (19), with

$$\lambda = \frac{1 - \gamma c^\top w}{1 - \gamma c^\top R_T w} > 0. \tag{F.4}$$

Neither full support nor irreducibility is required. Strict comparisons and ties are both preserved. □

Realization by a bounded coefficient. For each current pair i, take $\mu_i \in [0,1]$ and $|\kappa_i| \le \min\{\mu_i, 1 - \mu_i\}$. The coefficient in Eq. (17) has values $\mu_i - \kappa_i p_b$ and $\mu_i + \kappa_i p_a$ at the two selected successor coordinates, and μ_i elsewhere. These values lie in $[0,1]$. Its conditional action mean is μ_i at every successor state, because $p_a(-\kappa_i p_b) + p_b(\kappa_i p_a) = 0$. Consequently, with $K = T \odot g$ and $D_\mu = \mathrm{diag}(\mu_i)$,

$$A = K + (I - D_\mu)T = T + wc^\top, \qquad w_i = \kappa_i P(i,x) p_a p_b. \tag{F.5}$$

The coefficient formula uses the policy but not P. If $p_a p_b = 0$, its perturbation vanishes on realized transitions. A bounded fitted score $h_i(x,u) \in [0,1]$ can supply $\kappa_i = \eta_i[h_i(x,b) - h_i(x,a)]$, where $0 \le \eta_i \le \min\{\mu_i, 1 - \mu_i\}$; the comparison guarantee does not require an accurate score. Row-dependent κ_i can produce different successor-action ratios in different rows, so A need not factor as $P\Pi_{\pi'}$ for any stationary policy π'.

Connection to the original trace operator. The coefficient still multiplies only the inherited trace. Applying Theorem 2 in state–action coordinates gives

$$\mathcal{R}_g q = r + \gamma T(I - \gamma K)^{-1} D_\mu r + \gamma T(I - \gamma K)^{-1}(I - D_\mu) q. \tag{F.6}$$

Here $K\mathbf{1} = D_\mu \mathbf{1}$. Since $(I - \gamma K)\mathbf{1} - (I - D_\mu)\mathbf{1} = (1 - \gamma) D_\mu \mathbf{1} \ge 0$, the nonnegative linear part has row sums at most γ. For a common mean μ, its exact sup-norm modulus is $\gamma(1 - \mu)/(1 - \gamma\mu)$. At a fixed point, set $z = (I - \gamma K)^{-1}[D_\mu r + (I - D_\mu) q]$. Then $q = r + \gamma T z$ and $z = r + \gamma A z$, giving exactly the mixed root in Eq. (18). This identifies a contractive population evaluator; it is not a convergence claim for arbitrary online TD iterates.

F.2 Exact pairwise policy improvement

Fix action-probability floors $\ell_s(u) \geq 0$ satisfying $\ell_s := \sum_u \ell_s(u) \leq 1$. Start with any policy above these floors. For a selected pair, evaluate its mixed root. Transfer all probability above the floor from the lower-valued action to the higher-valued action, leaving every other probability unchanged. Hold the policy on a tie. The next comparison uses a coefficient rebuilt for its selected pair and current policy.

Proposition F.1 (Exact pairwise improvement)**.** *Each update weakly increases true value at every state. An actual policy change strictly increases value at the selected state. If every state and every unordered pair of distinct actions at that state is selected infinitely often, there are finitely many actual policy changes, and the resulting policy is optimal among policies satisfying the floors. Zero floors recover the ordinary optimal value.*

Proof. Theorem 8 selects the same better action as Q_P^π. If positive mass d moves from v to u at x, the new policy's Bellman residual at V_P^π is zero outside x and equals $d[Q_P^\pi(x,u) - Q_P^\pi(x,v)] > 0$ at x. Writing $P_{\pi'} = \Pi_{\pi'} P$ and $\mathcal{T}_P^{\pi'}$ for its state-value Bellman operator,

$$V_P^{\pi'} - V_P^\pi = (I - \gamma P_{\pi'})^{-1}[\mathcal{T}_P^{\pi'} V_P^\pi - V_P^\pi] \geq 0. \tag{F.7}$$

The inverse is nonnegative and has diagonal entries at least one, giving strict improvement at x.

For finiteness, treat each initial excess probability $\pi_0(u \mid s) - \ell_s(u)$ as an indivisible mass. An update moves all masses at the losing action to the winning action; it never splits them. At a state with k_s actions there are at most $k_s^{k_s}$ assignments of these masses. Thus only finitely many policies can occur, even for arbitrary real initial probabilities. Strict value improvement prevents a policy from recurring after an actual change.

Let $\bar{\pi}$ be the eventually constant policy. Fair selection of all pairs implies that every action with positive excess mass maximizes $Q_P^{\bar{\pi}}$; otherwise a later visit would change it. Its value therefore satisfies

$$V(s) = \sum_u \ell_s(u)[r(s,u) + \gamma P_{s,u} V] + (1 - \ell_s) \max_u [r(s,u) + \gamma P_{s,u} V]. \tag{F.8}$$

The right side is the optimal Bellman operator for the constrained policy class and is a γ contraction. Its unique fixed point is the constrained optimal value. □

Only the selected comparison is protected. Using the same mixed critic for other action comparisons would require additional conditions. Under cyclic pair selection, a full unchanged sweep is a stopping rule for exact evaluation; the finiteness argument does not give a useful evaluation-complexity bound.

F.3 A rollout certificate without the transition matrix

Assume reset access to a selected state and either initial action, or equivalent generative access. For the bounded construction, define

$$\nu_i(u \mid y) = \pi(u \mid y)[1 + g(i,(y,u)) - \mu_i]. \tag{F.9}$$

Conditional centering makes ν_i a probability distribution, and $A(i,(y,u)) = P(i,y)\nu_i(u \mid y)$. One environment transition followed by an action drawn from ν_i therefore samples A without a transition model. This controller remembers the preceding state–action pair. To sample the mixed root, start with the forced initial action, draw the first successor action under π, and use ν_i thereafter. The expected return is $r + \gamma T(I - \gamma A)^{-1} r$. This is an alternative rollout evaluator of the same population root, not a confidence bound for a learned TD critic.

Suppose rewards lie in a known interval $[r_{\min}, r_{\max}]$. Fix $0 < \delta < 1$ and $H \geq 1$. Let $G_H^{(j)}(x,u)$ denote the discounted sum of the first H rewards in rollout j starting from (x,u). For $N \geq 2$ rollout pairs that are conditionally i.i.d. given the frozen selections, let $Z_j = G_H^{(j)}(x,b) - G_H^{(j)}(x,a)$ and define

$$\begin{aligned} \bar{Z} &= \frac{1}{N}\sum_{j=1}^N Z_j, \qquad S_N^2 = \frac{1}{N-1}\sum_{j=1}^N (Z_j - \bar{Z})^2, \\ W &= \frac{r_{\max} - r_{\min}}{1-\gamma}, \qquad L = W(1-\gamma^H). \end{aligned} \tag{F.10}$$

Then $Z_j \in [-L, L]$. A two-sided empirical Bernstein radius, including the unobserved tail, is

$$\rho_N = \sqrt{\frac{2S_N^2 \log(4/\delta)}{N}} + \frac{14L \log(4/\delta)}{3(N-1)} + W\gamma^H. \tag{F.11}$$

Proposition F.2 (Certified pairwise update)**.** *Freeze the policy, coefficient, selected pair, N, and H before sampling. With probability at least $1-\delta$, the mixed population gap lies in $[\bar{Z}-\rho_N, \bar{Z}+\rho_N]$. An update toward b when $\bar{Z} > \rho_N$, or toward a when $\bar{Z} < -\rho_N$, therefore has the correct environmental advantage sign on this event.*

Proof. Apply Theorem 4 of Maurer & Pontil (2009) to $(Z_j + L)/(2L)$ and its complement, each with failure probability $\delta/2$. Rescaling gives the first two terms of Eq. (F.11); the sample variance uses denominator $N-1$. The omitted expected tails differ by at most $W\gamma^H$. The positive factor in Eq. (19) transfers the certified sign to the true gap. If $L = 0$, both differences and the radius are zero. □

At round j, the selections may depend on earlier batches. Fresh samples and positive error allocations δ_j with $\sum_j \delta_j \le \delta_{\text{tot}} \in (0, 1)$ give probability at least $1 - \delta_{\text{tot}}$ that every accepted policy update is value-nondecreasing. This is a fixed-batch guarantee: the interval is checked once after its prescribed N samples, not repeatedly until it excludes zero. Zero observed variance does not remove the linear range term. The guarantee concerns accepted policies, not the exploratory rollout returns.

F.4 Fixed-budget comparison with direct evaluation

We compare the rollout construction with ordinary Q_P^π rollouts under the same interaction budget. Both use independent arms within each pair and Eq. (F.11); their random streams are paired across methods. The four fixed tasks have two binary-action states with full support, three binary-action states with full support, three binary-action states with sparse cyclic transitions, and a zero-reward version of the two-state task. Their discounts are $0.9, 0.9, 0.95, 0.9$, respectively. The accompanying code specifies all transition, reward and fixed score matrices.

Across the four tasks, all 24 initial task–policy combinations ($4+8+8+4$) with action probabilities in $\{0.1, 0.9\}$ are used, with probability floor 0.1 and seeds 0–15. The coefficient has mean $1/2$ and uses the fixed bounded score $h_i(x, u) = \widehat{P}(i, x)\pi(u \mid x)$ with $\eta_i = 1/2$. The matrices $\widehat{P}$ are prescribed bounded scores, not fitted models. Each run performs two complete state sweeps, with $N \in \{1{,}024, 8{,}192\}$ new pairs per comparison and $H = 200$ rewards per arm. These two budgets and the sampling protocol were fixed before obtaining the rollout results. The per-run error budget is 0.05, divided equally among its $2n$ comparisons. The control rule uses observed returns, known reward bounds, the policy and scores; exact values are used only afterward to measure outcomes.

Let $\bar{V}$ denote the mean true value over states. The performance measure is $(\bar{V}_{\text{final}} - \bar{V}_{\text{initial}})/(\bar{V}_\ell^* - \bar{V}_{\text{initial}})$, the fraction of the initial gap to the floor-constrained optimum removed. Initially optimal runs are excluded from this ratio but retained in all update and cost counts. Table 13 reports every task at both budgets.

At the larger budget, both methods reach the constrained optimum in every initially nonoptimal run of the two three-state tasks. The two-state task remains a bottleneck: only $3/64$ runs change policy, closing 3.69% of the initial gap on average. Final true value vectors agree in all 384 matched method pairs at this budget. The sparse task takes 252 mixed updates versus 254 direct updates, but both spend the same fixed sampling budget.

For method comparison, first average eligible starts within each task, then weight the three nonzero-reward tasks equally. At $N = 1{,}024$, this measure is 38.84% for mixed and 38.94% for direct. The mixed-minus-direct mean is -0.100 percentage points, with a 97.5% paired Student-t interval $[-1.211, 1.010]$ across 16 seed-level differences. The interval level accounts for the two budgets under the paired-t assumptions. At $N = 8{,}192$, both measures are 67.90% and all 16 observed paired differences are zero. The study produces accepted policy updates, but shows no advantage over direct evaluation at these budgets.

Table 13: Fixed-budget pairwise improvement. Counts include all initial policies and 16 seeds; gap closure averages only initially nonoptimal runs. The zero-reward control has no initial gap. Direct and mixed receive identical sampling budgets.

	Runs with an update		Gap closed (%)	
Task	Direct	Mixed	Direct	Mixed
$N = 1{,}024$ pairs per comparison				
Two-state, full support	$0/64$	$0/64$	0.00	0.00
Three-state, full support	$72/128$	$72/128$	54.54	54.24
Three-state, sparse cycle	$80/128$	$80/128$	62.29	62.29
Zero-reward ties	$0/64$	$0/64$	—	—
$N = 8{,}192$ pairs per comparison				
Two-state, full support	$3/64$	$3/64$	3.69	3.69
Three-state, full support	$112/128$	$112/128$	100.00	100.00
Three-state, sparse cycle	$112/128$	$112/128$	100.00	100.00
Zero-reward ties	$0/64$	$0/64$	—	—

Across all methods and budgets, 1,257 actual updates produced no observed decrease greater than 10^{-10} in any state's true value. No interval missed its population target. These observations are distinct from the per-run probability guarantee; the 1,536 runs do not share one joint 95% guarantee. The full comparison used 15,099,494,400 environment transitions ($2NH$ per comparison), including both methods, budgets, all starts and seeds. At the larger budget, a single two-state run uses 13,107,200 transitions and a three-state run uses 19,660,800. Thus the certificate's interaction cost is substantial even in these small environments.

An independent verifier recomputes all 8,192 batch means, unbiased variances, radii and decisions from the 37,748,736 saved paired differences. It also recovers mixed roots from the original affine operator, checks true-value changes and constrained optima, and reconstructs seed-level summaries. These checks connect every accepted decision to its sampled interval and the resulting environmental value change. The supplementary ZIP includes code and per-batch records, but not the full raw-difference archive.