FINITE-REFERENCE ERROR BOUNDS FOR RECURRENT VALUE EVALUATORS IN POLICY IMPROVEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

We bound value error for finite-depth recurrent evaluators in policy improvement without assuming recurrent convergence. We compare the endpoint U_n, the value read at the deployed depth n, with a deeper finite endpoint U_m of the same evaluator. For a fixed policy, the standard Bellman-residual bound and triangle inequality give an error radius of $\|U_n - U_m\|_\infty$ plus the K-step residual of U_m divided by $1 - \gamma^K$. It is tighter than the direct residual bound on U_n exactly when the residual reduction exceeds $(1 - \gamma^K)\|U_n - U_m\|_\infty$; depth alone does not suffice. When the latent is carried across environment steps, the state includes the latent and the remaining edit budget; fixing the deployed policy and the shared depth-n recurrent map keeps both endpoints in one MDP, so m changes only endpoint evaluation. Exactly centered advantages connect the radius to a conservative policy improvement lower bound with signed evaluation error. An exact three-state example gives a nonvacuous radius and certifies improvement over the base. On a fixed 128-state Sudoku census, finite-reference proxies are 0.9 to 1.9% smaller than direct proxies across eight training seeds sharing one base initialization. Both remain uninformative as radii at roughly 70 times the checker scale; this is neither a uniform certificate nor a task-performance result. We analyze one candidate mixed exactly, state by state, with a frozen base policy at a fixed parameter snapshot.

1 INTRODUCTION

We study value error for recurrent evaluators read after a fixed number n of internal steps. When the output U_n is used to estimate advantages, its error relative to the base-policy value V^π contributes to the policy-improvement penalty. We ask when a deeper finite endpoint gives a tighter error bound without assuming recurrent convergence. Our setting is Unrolled Policy Iteration with TRMs (UPI-TRM), in which a Tiny Recursive Model (TRM) (Jolicoeur-Martineau, 2025) evaluates plan edits rewarded by a checker. We analyze one candidate's exact probability mixture with a frozen base policy at a fixed parameter snapshot.

The direct bound and a finite reference. Let $\mathcal{T}_K^\pi$ be the K-step Bellman operator for the fixed policy, with discount $\gamma \in [0, 1)$ and integer $K \geq 1$, and write $\varepsilon_{\mathrm{res},q} := \|U_q - \mathcal{T}_K^\pi U_q\|_\infty$. All norms below use the same policy-invariant domain under the conditions of Theorem 5.1. Standard residual analysis (Puterman, 1994; Sutton & Barto, 2018) gives the direct bound $\|U_n - V^\pi\|_\infty \leq \varepsilon_{\mathrm{res},n}/(1 - \gamma^K)$. It already needs no recurrent fixed point, but can be loose.

Instead, compare U_n with the same evaluator read at depth $m > n$. Applying the residual bound to U_m and the triangle inequality gives

$$\|U_n - V^\pi\|_\infty \leq \|U_n - U_m\|_\infty + \frac{\varepsilon_{\mathrm{res},m}}{1 - \gamma^K}.$$

The discrepancy measures sensitivity to additional computation; the reference residual measures agreement with the base policy's value equation. This radius is strictly smaller than the direct bound exactly when

$$(1 - \gamma^K)\|U_n - U_m\|_\infty < \varepsilon_{\mathrm{res},n} - \varepsilon_{\mathrm{res},m}.$$

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.

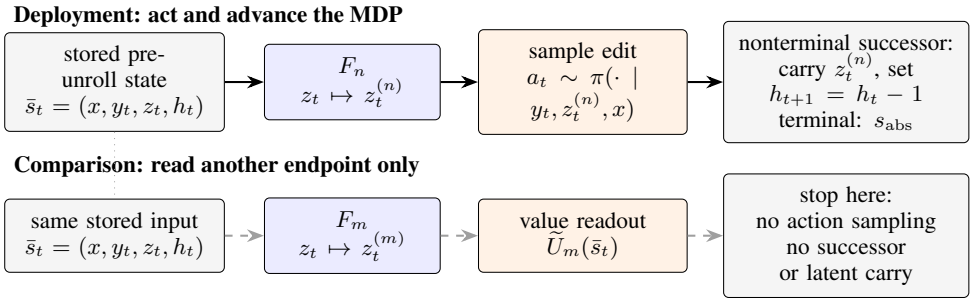

Figure 1: A deeper comparison is not a deeper deployment. Both lanes start from the same stored state $\bar{s}_t$ (instance x, plan y_t, pre-unroll latent z_t, remaining budget h_t). The upper lane applies F_n, samples an edit, and carries the new latent only after a nonterminal transition. The lower lane applies F_m, reads $\widetilde{U}_m$, and stops. Changing m leaves the deployed MDP and trajectory unchanged.

The residual reduction must outweigh the discrepancy penalty; depth alone guarantees nothing (Eq. (7)).

Two depths in one deployed MDP. For a carried latent, the depth-n map F_n determines the latent stored after a nonterminal transition and is therefore part of the transition kernel. Replacing it with F_m would generally change the MDP. We include the pre-unroll latent and remaining edit budget in the state; both policies share one frozen F_n. The comparison starts from the same stored state but changes only endpoint evaluation. Varying m leaves the deployed policy, action sampling, stored latent, and trajectory unchanged (Section 8).

Standard machinery.

- The K-step Bellman-residual estimate and triangle inequality give the displayed value bound.
- Conservative policy improvement (CPI) uses the surrogate and exact-mixture occupancy argument of Kakade & Langford (2002), restated in Lemma 7.2.

Contributions.

- A finite-reference formulation of recurrent value error, with an explicit comparison against the direct bound (Theorem 5.1, Eq. (7)).
- A clock-complete augmented MDP with a shared frozen recurrent map, so reference depth changes evaluation without changing deployment (Corollary C.2).
- Composition with exactly centered advantages and signed-error CPI, with target-network lag and approximate deployment treated separately (Corollary 5.3, Theorem 7.3, Proposition 6.1, Section 8).
- An exact example and a finite-census diagnostic that distinguish a tighter comparison from an informative value radius (Section 9).

Evidence. The exact three-state example reduces the direct radius from $3/4$ to $1/2$ and gives a signed CPI lower bound of $2/3$ above the base return of zero. On a fixed 128-state Sudoku census, eight training seeds sharing one base yield finite-reference proxies 0.9 to 1.9% smaller than the direct proxies. Both are roughly 70 times the checker scale and uninformative as value radii. This diagnostic establishes a finite-census comparison, not a uniform certificate or a task-performance result (Appendix E).

2 Related Work

CPI and safe policy iteration bound occupancy shift and approximate improvement (Kakade & Langford, 2002; Kakade, 2003; Pirotta et al., 2013; Metelli et al., 2021). We retain that exact-

mixture argument and supply its evaluation term from two finite recurrent endpoints. Classical approximate dynamic programming instead controls value error and its propagation through policy or value updates (Bertsekas & Tsitsiklis, 1996; Munos, 2003; Munos & Szepesvári, 2008; Scherrer, 2014); our radius is a fixed-policy population sup-norm statement, not a policy-sequence result. TRMs and Hierarchical Reasoning Models (HRMs) use recurrent latent refinement for structured prediction (Jolicoeur-Martineau, 2025; Wang et al., 2025), while deep equilibrium models analyze equilibrium representations (Bai et al., 2019). We read the evaluator at finite depth and reserve contraction for an optional fixed-point specialization. Learning to search (Daumé III et al., 2009; Ross et al., 2011; Chang et al., 2015), verifiable-reward systems (DeepSeek-AI, 2025), and value-iteration networks (Tamar et al., 2016) provide task context; our inner depth n is computation within one evaluation, not planning over transitions or the Bellman horizon K.

3 Background and Setup

The working setup has two clocks: environment transitions form the outer edit MDP, while recurrent steps refine one endpoint without advancing that MDP.

The outer edit MDP. An edit a_t changes plan y_t for a fixed instance x. With h_t edits remaining, the episodic state and nonterminal successor are

$$s_t = (x, y_t, h_t), \qquad s_{t+1} = \big(x, \mathrm{edit}(y_t, a_t; x), h_t - 1\big).$$

The remaining-budget clock h makes the displayed state clock-complete. Solving, terminal STOP, or budget exhaustion reaches the shared absorber s_{abs}. A bounded checker score $c(x, y)$ supplies reward, optionally shaped as $r = r_0 + \gamma\Phi(s') - \Phi(s)$ with $\Phi = c$ (Ng et al., 1999). We maximize discounted return $\eta(\pi)$ from initial law ρ, with $\gamma \in [0, 1)$. The shaped boundary is $V^\pi(s_{\mathrm{abs}}) = b_{\mathrm{abs}} = -C_{\max}$; folding its tail into the terminal reward with zero terminal bootstrap preserves returns (Appendices A.2 and F).

The inner evaluator. A recurrent map refines a latent while the instance and current plan remain fixed. Starting from $z^{(0)} = z_{\mathrm{init}}(x, y)$ in episodic mode,

$$z^{(k+1)} = \widetilde{f}_\theta(z^{(k)}, y, x), \qquad U_n(s) = V_\psi(z^{(n)}, x, y).$$

Depth n counts internal computation; the Bellman horizon $K \geq 1$ counts environment transitions. Here $\widetilde{f}_\theta = \mathsf{P}_{\mathrm{rec}} \circ f_\theta$, where $\mathsf{P}_{\mathrm{rec}} = \Pi_R$ projects onto a radius-$R > 0$ ball (Eq. (33)) or is the identity. The head reads (z, x, y); rewards, termination, and successors still require h.

Reset or carried latent state. Resetting z at each edit makes (x, y, h) Markov. Carrying z requires $\bar{s} = (x, y, z, h)$. This state stores the *pre-unroll* latent; deployment computes $F_n(x, y, z) = \widetilde{f}_\theta^{\circ n}(z, y, x)$, selects an action, and carries the result only after a nonterminal transition. Both CPI policies share this frozen map. The endpoint $\widetilde{U}_m$ starts from the same stored z and changes neither F_n, action sampling, nor successors (Figure 1).

Values and domains. Fix the MDP, recurrence, initialization, value head, and both policies. Policies are measurable kernels and endpoints are bounded measurable functions. On $\mathcal{C}$, $\|f\|_\infty = \sup_{s \in \mathcal{C}} |f(s)|$. Every endpoint equals b_{abs} at the absorber; latent paths exist only off the absorber. Section 5 introduces the needed advantage-evaluation domain, and Appendix A.1 gives the full taxonomy.

4 UPI-TRM Algorithm

Algorithm 1 constructs one proposal against a frozen base, not an iterated policy sequence. Before collection it freezes the base policy and recurrent map. Checker rewards and K-step targets fit the value head, and centered advantages from U_n fit one candidate head. The proposal is the exact probability mixture $\pi_\alpha = (1 - \alpha)\pi + \alpha\pi_{\mathrm{cand}}$; the base is never replaced.

Algorithm 1 One fixed-base UPI-TRM proposal on the clock-complete edit MDP

1: Freeze the base policy and recurrent map; sample x, initialize y_0, and set $h_0 = T \geq 1$.
2: Use the mode-appropriate state s_t: (x, y_t, h_t) in episodic mode, (x, y_t, z_t, h_t) in persistent mode; initialize persistent z_0.
3: **for** $t = 0, \ldots, T-1$ **do**
4: Use $z_t = z_{\mathrm{init}}(x, y_t)$ in episodic mode or stored z_t in persistent mode; compute $z_t^{(n)} = \widetilde{f}_\theta^{\circ n}(z_t, y_t, x)$ and $V_\psi(z_t^{(n)}, x, y_t)$.
5: Sample $a_t \sim \pi_\phi(\cdot \mid y_t, z_t^{(n)}, x)$; observe r_t, y_{t+1} and termination flag d_t; set $h_{t+1} = h_t - 1$.
6: Set $s_{t+1} = s_{\mathrm{abs}}$ if d_t or $h_{t+1} = 0$; otherwise construct the successor, carrying $z_t^{(n)}$ only in persistent mode.
7: Store (s_t, a_t, r_t, s_{t+1}); stop collection immediately if $s_{t+1} = s_{\mathrm{abs}}$.
8: **end for**
9: Form exact-K or terminal folded targets and fit the mode-appropriate value head.
10: Compute advantages by Eq. (2), conditional on the complete state.
11: Fit one candidate policy head while the base policy and recurrent map remain frozen.
12: Form the exact probability mixture $\pi_\alpha = (1-\alpha)\pi_\phi + \alpha\pi_{\mathrm{cand}}$.

For a bounded value function V, define the exact one-step estimator

$$\widehat{Q}_V(s,a) = \mathbb{E}_{s'}[r(s,a,s') + \gamma V(s')], \tag{1}$$

$$\widehat{A}_V(s,a) = \widehat{Q}_V(s,a) - \mathbb{E}_{a' \sim \pi(\cdot|s)}[\widehat{Q}_V(s,a')]. \tag{2}$$

The candidate uses $V = U_n$ and, with persistent latents, conditions on the full $\bar{s}$. Depth m is used only for evaluation, never for collection or action selection. A single-network approximation to the mixture requires the separate deployment correction in Section 8.

Appendix F gives folded targets and replay details. Exact baseline enumeration ensures statewise centering; approximate integration instead contributes the signed defect in Theorem 7.3. One depth-n unroll per action term over T states costs $O(T|\mathcal{A}|n)$ recurrent-map applications.

5 Finite-Reference Evaluation and Advantage Error

Domain-wide endpoint discrepancies and Bellman residuals bound value and advantage error. We first use episodic notation; Section 8 applies the same argument to one fixed augmented MDP.

5.1 The finite-reference value certificate

The certificate must cover the successors used to construct advantages. Let $\mathcal{C}_\pi$ carry the initial law and be invariant under π, and let $\mathcal{C} = \mathcal{R}^{\mathrm{adv}}_{\pi,\pi_{\mathrm{cand}}}$ contain $\mathcal{C}_\pi$, the absorber, and the successors of one current or candidate action followed by any current-policy continuation (Definition A.3). Write $\|f\|^{\mathrm{adv}}_{\infty,\pi,\pi_{\mathrm{cand}}} = \sup_{s\in\mathcal{C}} |f(s)|$. On this invariant domain the residual bound is

$$\|V - V^\pi\|^{\mathrm{adv}}_{\infty,\pi,\pi_{\mathrm{cand}}} \leq \frac{1}{1-\gamma^K} \|V - \mathcal{T}_K^\pi V\|^{\mathrm{adv}}_{\infty,\pi,\pi_{\mathrm{cand}}}. \tag{3}$$

Apply this estimate to $V = U_m$ and add the endpoint discrepancy:

Theorem 5.1 (Finite-reference bound)**.** *Let $\mathcal{C}$ be a measurable state set with its trace σ-algebra, let P_π be a Markov kernel satisfying $P_\pi(\mathcal{C} \mid s) = 1$ for every $s \in \mathcal{C}$, and fix $\gamma \in [0,1)$ and an integer $K \geq 1$. If an initial law is used, assume $\rho(\mathcal{C}) = 1$. Assume the one-step reward is uniformly bounded and its fixed-policy conditional mean $r^\pi(s) := \mathbb{E}^\pi_s[r_0]$ is measurable on $\mathcal{C}$. Then the K-step expected reward*

$$r^\pi_K(s) := \mathbb{E}^\pi_s\left[\sum_{i=0}^{K-1} \gamma^i r_i\right] \tag{4}$$

is bounded and measurable on $\mathcal{C}$. On the Banach space $B_b(\mathcal{C})$ of bounded measurable functions with the ordinary pointwise sup norm, define

$$(\mathcal{T}^\pi_K V)(s) := r^\pi_K(s) + \gamma^K (P^K_\pi V)(s). \tag{5}$$

Then $\mathcal{T}_K^\pi : B_b(\mathcal{C}) \to B_b(\mathcal{C})$ is a γ^K-contraction and has a unique fixed point V^π. For arbitrary endpoint functions $U_n, U_m \in B_b(\mathcal{C})$ and finite integers $0 \le n < m$,

$$\|U_n - V^\pi\|_\infty \le \underbrace{\|U_n - U_m\|_\infty}_{\textit{finite-depth discrepancy}} + \underbrace{\frac{\|U_m - \mathcal{T}_K^\pi U_m\|_\infty}{1-\gamma^K}}_{\textit{reference-depth residual}}. \tag{6}$$

The first term measures sensitivity to extra recurrent computation. The second measures the reference endpoint's disagreement with the frozen policy's value equation. Both use the same policy, kernel, domain, norm, and Bellman horizon as V^π. The direct comparator is $\varepsilon_{\mathrm{res},n}/(1-\gamma^K)$ (Corollary C.3). With both residuals measured on the same domain, the finite-reference radius is strictly smaller if and only if

$$(1-\gamma^K)\,\|U_n - U_m\|_\infty < \varepsilon_{\mathrm{res},n} - \varepsilon_{\mathrm{res},m}. \tag{7}$$

Both bounds remain valid when Eq. (7) fails; the condition compares radii rather than asserting that either radius is useful. Large residuals can make both vacuous. Appendix C.4.1 gives the proof and a nonconvergent endpoint example; Corollary C.8 gives the latent-path inequalities. Appendix D.1 gives the optional contraction specialization with its geometric L_z^n limit, and Appendix D.2 a weighted-L_p version, which requires domination of the fixed-policy occupancy by the reference measure.

5.2 From value error to candidate advantage error

Successor value error E contributes at most γE separately to the action value and to the current-policy baseline, giving $2\gamma E$. Exact centering then permits a candidate-specific refinement by policy overlap. Define

$$\tau(s) := \mathrm{TV}(\pi_{\mathrm{cand}}(\cdot \mid s), \pi(\cdot \mid s)), \qquad \tau_\infty := \sup_{s \in \mathcal{C}_\pi} \tau(s) \le 1. \tag{8}$$

We use $b(s) = \mathbb{E}_{\pi_{\mathrm{cand}}}[\widehat{A}(s,a) - A^\pi(s,a)]$ for the signed candidate defect, and $\varepsilon_{A,\mathrm{cand}} = \sup_{s \in \mathcal{C}_\pi} |b(s)|$. The normalized discounted occupancy is $d_\pi = (1-\gamma)\sum_{t \ge 0} \gamma^t \rho P_\pi^t$.

Value error is controlled on $\mathcal{C}$, while advantage error is evaluated on $\mathcal{C}_\pi$. Because $\mathcal{C}$ contains candidate successors, this is a post-candidate check.

Corollary 5.2 (Advantage-error bridge)**.** *Under the hypotheses of Theorem 5.1, define*

$$B_{n,m} := \|U_n - U_m\|^{\mathrm{adv}}_{\infty,\pi,\pi_{\mathrm{cand}}} + \frac{\|U_m - \mathcal{T}_K^\pi U_m\|^{\mathrm{adv}}_{\infty,\pi,\pi_{\mathrm{cand}}}}{1-\gamma^K}. \tag{9}$$

Let $\widehat{A} = \widehat{A}_{U_n}$ be the exact one-step centered estimator of Eq. (2)*, with the transition expectation evaluated exactly (or applied to a deterministic transition) and the baseline computed exactly under π. Then Assumption 7.1 holds with the valid choice*

$$\varepsilon_A = 2\gamma B_{n,m}. \tag{10}$$

The reference assesses the deployed depth-n estimator without changing action selection. Exact centering gives the sharper candidate-average bound:

Corollary 5.3 (Candidate evaluation bias from policy overlap)**.** *Let $\widehat{A}_V$ be the exactly centered one-step estimator, $e = V - V^\pi$, and $\delta_V = \|e\|^{\mathrm{adv}}_{\infty,\pi,\pi_{\mathrm{cand}}}$. Use τ and τ_∞ from Eq.* (8)*. Then, for every $s \in \mathcal{C}_\pi$,*

$$\left|\mathbb{E}_{a \sim \pi_{\mathrm{cand}}(\cdot|s)}[\widehat{A}_V(s,a) - A^\pi(s,a)]\right| \le 2\gamma\delta_V \tau(s), \tag{11}$$

and hence

$$\varepsilon_{A,\mathrm{cand}} \le 2\gamma\delta_V\tau_\infty, \tag{12}$$

$$\mathbb{E}_{s \sim d_\pi}|b(s)| \le 2\gamma\delta_V \mathbb{E}_{s \sim d_\pi}\tau(s). \tag{13}$$

For an exactly centered K-step estimator whose intervening actions follow π, replace γ by γ^K. The fallback $2\gamma\delta_V$ follows from $\tau \le 1$.

Appendix C.4.4 gives the proof. Exact centering cancels the shared part of the two action laws. The resulting $2\gamma\delta_V\tau(s)$ is zero where the policies agree and can be smaller than the $2\gamma\delta_V$ fallback. Under the exact-mixture and centering hypotheses of Theorem 7.3, substituting the finite-reference radius gives

$$\eta(\pi_\alpha) \geq \widehat{L}_\pi(\pi_\alpha) - \frac{2\alpha\gamma\tau_\infty}{1-\gamma} B_{n,m} - \frac{2\varepsilon_{\mathrm{CPI}}\gamma\alpha^2}{(1-\gamma)(1-\gamma+\gamma\alpha)}. \tag{14}$$

Section 7 defines the estimated surrogate $\widehat{L}_\pi$ and the candidate-advantage bound $\varepsilon_{\mathrm{CPI}}$. Estimators outside Corollary 5.3, including generalized advantage estimation (GAE), need a separate error bound (Remark C.10).

6 Target-network residuals

Value regression and certification use different residuals. Regression fits U_n to a target network $\bar{V}$, while the certificate needs the self-bootstrap backup of the evaluated endpoint. For a frozen current-policy trajectory, the target is $G^{(K)} = \sum_{k<K} \gamma^k r_k + \gamma^K \bar{V}(S_K)$, with loss $\mathcal{L}_{\mathrm{val}} = \mathbb{E}_{S_0\sim\mu}[(U_n(S_0) - \mathrm{sg}(G^{(K)}))^2]$. Here sg denotes stop-gradient, μ selects starting states, and the conditional trajectory law is π. A zero target residual can coexist with a nonzero self-bootstrap residual. Propagated target lag supplies the required correction:

Proposition 6.1 (Target-network residual)**.** *Let $\mathcal{E}$ be a measurable P_π-invariant state domain. Assume $r_K^\pi \in B_b(\mathcal{E})$ and $U_q, \bar{V} \in B_b(\mathcal{E})$ for a finite depth $q \geq 0$, and suppose the conditional K-step target trajectory follows the frozen policy π. Define*

$$\varepsilon_{\mathrm{targ},q} := \left\| U_q - \left(r_K^\pi + \gamma^K P_\pi^K \bar{V}\right) \right\|_{\infty,\mathcal{E}}. \tag{15}$$

Retain the propagated target lag by defining

$$R^P_{\mathrm{targ},q} := \varepsilon_{\mathrm{targ},q} + \gamma^K \|P_\pi^K(\bar{V} - U_q)\|_{\infty,\mathcal{E}}. \tag{16}$$

Then the self-bootstrap residual satisfies

$$\|U_q - \mathcal{T}_K^\pi U_q\|_{\infty,\mathcal{E}} \leq R^P_{\mathrm{targ},q}. \tag{17}$$

For the formal absorbing representation, require $U_q(s_{\mathrm{abs}}) = \bar{V}(s_{\mathrm{abs}}) = b_{\mathrm{abs}}$ and use the same absorbing reward and boundary in both population operators. The equivalent folded implementation may instead fold γb_{abs} into the terminal-transition reward and mask the terminal bootstrap to zero, but it must not add the absorbing tail again.

Appendix C.4.5 gives the proof. The ordinary sup-lag form $R^{\sup}_{\mathrm{targ},q} := \varepsilon_{\mathrm{targ},q} + \gamma^K \|\bar{V} - U_q\|_{\infty,\mathcal{E}}$ is a looser nonexpansive fallback: $R^P_{\mathrm{targ},q} \leq R^{\sup}_{\mathrm{targ},q}$. Corollary C.3 gives the value/CPI substitutions. A population mean-square loss controls only an $L^2(\mu)$ conditional residual, not the uniform quantity above.

7 Conservative Policy Improvement

Start with the exactly centered value-derived advantage of Eq. (2). The analyzed proposal is the exact statewise probability mixture $\pi_\alpha = (1-\alpha)\pi + \alpha\pi_{\mathrm{cand}}$, realizable by a Bernoulli selector at each decision. CPI separates the signed evaluation correction under d_π from the occupancy shift of repeated mixture decisions (Kakade & Langford, 2002). The first uses $\mathcal{C}_\pi$; the second requires $\mathcal{C}_{\mathrm{mix}}$ to contain full trajectories under both policies (Definition A.4). The signed bound (Theorem 7.3) allows a centering defect; the next assumption supplies a sufficient bound on the candidate-average error in its scalar specialization.

Assumption 7.1 (Uniform advantage error on $\mathcal{C}_\pi$)**.** For some $\varepsilon_A \geq 0$:

$$|\widehat{A}(s,a) - A^\pi(s,a)| \leq \varepsilon_A \tag{18}$$

for every $s \in \mathcal{C}_\pi$ and for $\pi(da \mid s)$- and $\pi_{\mathrm{cand}}(da \mid s)$-almost every a.

Using the normalized discounted occupancy d_π defined above, the true and estimated surrogates are

$$L_\pi(\pi') := \eta(\pi) + \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\pi,\, a\sim\pi'(\cdot|s)}[A^\pi(s,a)],$$
$$\widehat{L}_\pi(\pi') := \eta(\pi) + \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\pi,\, a\sim\pi'(\cdot|s)}[\widehat{A}(s,a)].$$

Lemma 7.2 (Exact-mixture occupancy)**.** *Let π and π_{cand} act in the same MDP from the same initial law, and let π_α be their exact pointwise probability mixture. For $\alpha \in [0,1]$,*

$$\mathrm{TV}(d_{\pi_\alpha}, d_\pi) \le \frac{\gamma\alpha}{1-\gamma+\gamma\alpha}. \tag{19}$$

For finite or countable state spaces, with $\mathrm{TV}(p,q) = \frac{1}{2}\|p-q\|_1$,

$$\|d_{\pi_\alpha} - d_\pi\|_1 \le \frac{2\gamma\alpha}{1-\gamma+\gamma\alpha}. \tag{20}$$

Appendix C.4.6 gives the proof. The order-α occupancy shift multiplies an order-α candidate term, so its correction is $O(\alpha^2)$ for a fixed domain and advantage span. The distinct evaluation correction combines centering defect $c(s)$ and candidate bias $b(s)$ under d_π.

Theorem 7.3 (CPI with signed evaluation error)**.** *Fix one discounted MDP with bounded rewards and measurable current and candidate policy kernels. Let $\alpha \in [0,1]$ and let $\pi_\alpha = (1-\alpha)\pi + \alpha\pi_{\mathrm{cand}}$ be their exact pointwise probability-space mixture, with no parameter interpolation, hidden-state interpolation, logit interpolation, or distillation. Let $\mathcal{C}_{\mathrm{mix}}$ be measurable with $\rho(\mathcal{C}_{\mathrm{mix}}) = 1$ and invariant under both π and π_α, as in Definition A.4. Assume that a measurable π-invariant domain $\mathcal{C}_\pi \subseteq \mathcal{C}_{\mathrm{mix}}$ also carries ρ. Assume the functions below are bounded and measurable on their declared domains. Define g on $\mathcal{C}_{\mathrm{mix}}$ and c, b on $\mathcal{C}_\pi$ by*

$$g(s) := \mathbb{E}_{a\sim\pi_{\mathrm{cand}}(\cdot|s)}[A^\pi(s,a)], \qquad \Delta_g := \sup_{s\in\mathcal{C}_{\mathrm{mix}}} g(s) - \inf_{s\in\mathcal{C}_{\mathrm{mix}}} g(s), \tag{21}$$

$$c(s) := \mathbb{E}_{a\sim\pi(\cdot|s)}[\widehat{A}(s,a)], \qquad b(s) := \mathbb{E}_{a\sim\pi_{\mathrm{cand}}(\cdot|s)}[\widehat{A}(s,a) - A^\pi(s,a)], \tag{22}$$

$$\Xi_\alpha := \mathbb{E}_{s\sim d_\pi}[(1-\alpha)c(s) + \alpha b(s)]. \tag{23}$$

Then the surrogate error is exactly

$$\widehat{L}_\pi(\pi_\alpha) - L_\pi(\pi_\alpha) = \frac{\Xi_\alpha}{1-\gamma}, \tag{24}$$

and

$$\eta(\pi_\alpha) \ge \widehat{L}_\pi(\pi_\alpha) - \frac{\Xi_\alpha}{1-\gamma} - \frac{\gamma\alpha^2\Delta_g}{(1-\gamma)(1-\gamma+\gamma\alpha)}. \tag{25}$$

Because Ξ_α is signed, a usable one-sided guarantee may replace it by any established $B_\alpha \ge \Xi_\alpha$.

Under exact statewise centering,

$$\mathbb{E}_{a\sim\pi(\cdot|s)}[\widehat{A}(s,a)] = 0 \qquad \textit{for every } s \in \mathcal{C}_\pi, \tag{26}$$

define the uniform candidate-policy bias

$$\varepsilon_{A,\mathrm{cand}} := \sup_{s\in\mathcal{C}_\pi} \left| \mathbb{E}_{a\sim\pi_{\mathrm{cand}}(\cdot|s)}\left[\widehat{A}(s,a) - A^\pi(s,a)\right] \right|. \tag{27}$$

If $\varepsilon_{A,\mathrm{cand}} < \infty$ and $\varepsilon_{\mathrm{CPI}} := \sup_{s\in\mathcal{C}_{\mathrm{mix}}} |g(s)| < \infty$, then

$$\eta(\pi_\alpha) \ge \widehat{L}_\pi(\pi_\alpha) - \frac{\alpha}{1-\gamma}\,\varepsilon_{A,\mathrm{cand}} - \frac{2\varepsilon_{\mathrm{CPI}}\gamma\alpha^2}{(1-\gamma)(1-\gamma+\gamma\alpha)}. \tag{28}$$

A negative signed defect credits underestimated advantage; an absolute bound loses that credit (Section 9). For example, $B_\alpha = (1-\alpha)\mathbb{E}_{d_\pi}|c(s)| + \alpha\mathbb{E}_{d_\pi}|b(s)|$ upper-bounds Ξ_α, and sampled or learned baselines contribute through $c(s)$. Appendix C.4.7 proves the result. Because the domain and span may depend on α, Corollary C.11 (Appendix C.4.8) gives a common-domain safe-step extension, with proof and boundary cases in Appendix C.4.9. Assumption 7.1 implies $\varepsilon_{A,\mathrm{cand}} \le \varepsilon_A$, but the signed theorem does not require it.

8 PERSISTENT LATENTS AND DEPLOYMENT

One augmented MDP. The required invariant is a clock-complete stored state $\bar{s} = (x, y, z, h)$ and one shared frozen map F_n. Fix its measurable initialization law $\bar{\rho}$, bounded rewards, measurable reward/edit and policy kernels, and value head. Both policies use F_n, and only a nonterminal transition carries its output; every terminal cause reaches the common absorber. A finite comparison endpoint starts from the same stored z:

$$\widetilde{U}_q(\bar{s}) := V_\psi(\widetilde{f}_\theta^{\circ q}(z, y, x), x, y), \qquad \widetilde{U}_q(s_{\mathrm{abs}}) := -C_{\max}. \tag{29}$$

The deployed action laws are

$$\bar{\pi}(a \mid \bar{s}) := \pi_\phi(a \mid y, F_n(x, y, z), x), \qquad \bar{\pi}_{\mathrm{cand}}(a \mid \bar{s}) := \pi_{\mathrm{cand}}(a \mid y, F_n(x, y, z), x). \tag{30}$$

Let $\bar{\mathcal{T}}_K^{\bar{\pi}}$ be the K-step Bellman operator of this augmented MDP, whose transitions use F_n. Changing q changes only endpoint evaluation; it leaves F_n, both action laws, the kernel, and the trajectory fixed (Figure 1, Appendix C.1.1).

Persistent-state consequence (informal). Corollary C.2 gives the complete statement. Let $\bar{\mathcal{C}}$ be a measurable augmented advantage domain containing the common absorber and a $\bar{\pi}$-invariant current-policy domain $\bar{\mathcal{C}}_\pi$ carrying $\bar{\rho}$. Assume bounded K-step expected rewards and $\widetilde{U}_n \in B_b(\bar{\mathcal{C}})$ with the common scalar boundary. Write

$$\bar{\varepsilon}_{\mathrm{res},n} := \|\widetilde{U}_n - \bar{\mathcal{T}}_K^{\bar{\pi}} \widetilde{U}_n\|_\infty,$$

with norms on $\bar{\mathcal{C}}$. On $B_b(\bar{\mathcal{C}})$, the operator has a unique fixed point $V^{\bar{\pi}}$. The self-bootstrap residual then gives

$$\|\widetilde{U}_n - V^{\bar{\pi}}\|_\infty \le \bar{\varepsilon}_{\mathrm{res},n}/(1 - \gamma^K).$$

With the exact transition expectation and exact current-policy baseline, both conditional on the full augmented state, the advantage error is at most $2\gamma\bar{\varepsilon}_{\mathrm{res},n}/(1 - \gamma^K)$ for every $\bar{s} \in \bar{\mathcal{C}}_\pi$ and for $\bar{\pi}$- and $\bar{\pi}_{\mathrm{cand}}$-almost every action. For the exact full-state mixture $\bar{\pi}_{\mathrm{new}} = (1 - \alpha)\bar{\pi} + \alpha\bar{\pi}_{\mathrm{cand}}$, a common mixture domain, and all remaining hypotheses of Theorem 7.3, let $\bar{\tau}_\infty$ be the uniform policy TV distance on the current-policy domain and let $\bar{\varepsilon}_{\mathrm{CPI}}$ bound the magnitude of the candidate's expected true advantage on the mixture domain. With $d_{\bar{\pi}} = (1 - \gamma) \sum_{t \ge 0} \gamma^t \bar{\rho} P_{\bar{\pi}}^t$ and $\widehat{A} = \widehat{A}_{\widetilde{U}_n}$, define the full-state surrogate

$$\widehat{L}_{\bar{\pi}}(\bar{\pi}') := \eta(\bar{\pi}) + \frac{1}{1 - \gamma} \mathbb{E}_{\bar{s} \sim d_{\bar{\pi}},\, a \sim \bar{\pi}'(\cdot \mid \bar{s})}[\widehat{A}(\bar{s}, a)].$$

The overlap refinement gives

$$\eta(\bar{\pi}_{\mathrm{new}}) \ge \widehat{L}_{\bar{\pi}}(\bar{\pi}_{\mathrm{new}}) - \frac{2\alpha\gamma\bar{\tau}_\infty}{1 - \gamma} \frac{\bar{\varepsilon}_{\mathrm{res},n}}{1 - \gamma^K} - \frac{2\bar{\varepsilon}_{\mathrm{CPI}}\gamma\alpha^2}{(1 - \gamma)(1 - \gamma + \gamma\alpha)}.$$

Target-network and finite-reference substitutions retain the same augmented domains, common boundary, frozen-policy trajectory, exact conditional transition expectation, and exact baseline.

For finite $0 \le n < m$ and bounded endpoints with the common boundary, the same-MDP finite-reference form (Appendix C.4.10) is

$$\|\widetilde{U}_n - V^{\bar{\pi}}\|_\infty \le \|\widetilde{U}_n - \widetilde{U}_m\|_\infty + \frac{\|\widetilde{U}_m - \bar{\mathcal{T}}_K^{\bar{\pi}} \widetilde{U}_m\|_\infty}{1 - \gamma^K}. \tag{31}$$

Approximating the mixture at deployment. Distillation or parameter, logit, or hidden-state interpolation generally produces a different law $\widetilde{\pi}$. If a common invariant domain gives uniform TV discrepancy δ_{dep} from π_α and reward range Δ_r, Theorem C.4 gives

$$|\eta(\widetilde{\pi}) - \eta(\pi_\alpha)| \le \frac{\Delta_r \delta_{\mathrm{dep}}}{(1 - \gamma)(1 - \gamma + \gamma\delta_{\mathrm{dep}})}. \tag{32}$$

Either-direction uniform KL control supplies a TV radius (Corollary C.6). This correction applies only within the same MDP. Changing the recurrent map requires a different-MDP analysis.

9 EXACT EXAMPLE AND LEARNED-EVALUATOR DIAGNOSTIC

Exact example: can both bounds be useful? Yes. The finite reference beats the direct radius and the signed CPI bound certifies improvement, although its absolute specialization is vacuous. Nothing is sampled or learned.

Let $\mathcal{S} = \{s_0, s_1, s_{\mathrm{abs}}\}$, actions $\{a_0, a_1\}$, $\rho = \delta_{s_0}$, $\gamma = 1/2$, and $K = 1$. The base always chooses a_0 and the candidate a_1. The deterministic reward-successor pairs are $(s_0, a_0) : (0, s_0)$, $(s_0, a_1) : (1, s_1)$, $(s_1, a_0) : (0, s_0)$, and $(s_1, a_1) : (0, s_1)$; absorber actions self-loop with zero reward. Here $\mathcal{C}_\pi = \{s_0, s_{\mathrm{abs}}\}$ and every other domain is $\mathcal{S}$ (Appendix B). Set $z^{(0)} = 0$, $T_s(z) = u_s - z/2$, $(u_{s_0}, u_{s_1}) = (1/4, -1/4)$, $u_{s_{\mathrm{abs}}} = 0$, and $V_\psi(z, s) = z$. Use $n = 1$, $m = 2$, and $\alpha = 1/2$ (Table 1).

Table 1: Exact two-nonabsorbing-state calculation. State vectors are ordered $(s_0, s_1, s_{\mathrm{abs}})$.

Quantity	Exact value
V^π; U_1; U_2	$(0,0,0)$; $(1/4, -1/4, 0)$; $(1/8, -1/8, 0)$
$\|U_1 - U_2\|_\infty$; $\|U_2 - \mathcal{T}_1^\pi U_2\|_\infty$	$1/8$; $3/16$
Finite-reference upper bound; actual $\|U_1 - V^\pi\|_\infty$	$1/2$; $1/4$
d_π; $d_{\pi_{1/2}}$; $\pi_{1/2}(a_0 \mid s)$, $\pi_{1/2}(a_1 \mid s)$	$(1,0,0)$; $(3/4, 1/4, 0)$; $(1/2, 1/2)$ at every state
g; Δ_g; $c(s)$; $b(s)$	$(1,0,0)$; 1; $(0,0,0)$; $(-1/4, -1/4, 0)$
$\widehat{L}_\pi(\pi_{1/2})$; $\Xi_{1/2}$	$3/4$; $-1/8$
Signed CPI lower bound; actual $\eta(\pi_{1/2})$	$2/3$; $3/4$

The finite-reference radius $B_{1,2} = 1/2$ bounds the actual error $1/4$ using two finite endpoints. The direct depth-1 bound (Corollary C.3) is $\varepsilon_{\mathrm{res},1}/(1-\gamma) = (3/8)/(1/2) = 3/4$, and Eq. (7) holds: $(1-\gamma)\|U_1 - U_2\|_\infty = 1/16 < 3/16 = \varepsilon_{\mathrm{res},1} - \varepsilon_{\mathrm{res},2}$. The fixed-point specialization Eq. (84) also gives $3/4$ (Appendix B). Deeper references are not monotone: $m = 3$ gives $5/8$ and $m = 4$ gives $9/16$.

The signed calculation answers a separate policy question. Because $b(s_0) = -1/4$ and $\Xi_{1/2} = -1/8$, correcting the surrogate $3/4$ adds $1/4$. After the exact-span occupancy penalty $1/3$, the result is $\eta(\pi_{1/2}) \geq 2/3$, versus actual return $3/4$ and base return zero. The absolute specialization instead subtracts $1/4$, replaces $\Delta_g = 1$ by $2\varepsilon_{\mathrm{CPI}} = 2$, and gives $-1/6$. Both are valid; only the signed-span version certifies improvement. Appendix B derives every quantity, and `verify_finite_mdp.py` checks them.

Learned diagnostic: does the comparison hold on one census? We test the sign of Eq. (7) using census maxima, not population norms. For seed j, the direct proxy is $\varepsilon_{n,j}/(1-\gamma)$ and the finite-reference proxy is $D_j + \varepsilon_{m,j}/(1-\gamma)$, where D_j and $\varepsilon_{q,j}$ are the census maxima of the endpoint discrepancy and the one-step residual (Eq. (91)). Eight training seeds share one base initialization and one fixed 128-state census of a Sudoku plan-editing task with $C_{\max} = 16$. With $n = 2$, $m = 8$, $K = 1$, $\gamma = 0.99$, and $\alpha = 0.1$, the finite-reference proxy is smaller for all eight seeds. The mean gap is 14.25 and the nominal 95% seed-bootstrap interval is $[12.19, 16.74]$, conditional on that base and census. The reduction is only 0.9 to 1.9% of direct proxies near 1.1×10^3, roughly 70 times $C_{\max}$. Neither proxy is informative as a radius at these snapshots. Replay reproduced all eight gaps bitwise from existing checkpoints; it reproduces measurements, not training or historical execution. Appendix E gives full-precision results and numerical checks, including a float32 tolerance widened after inspection of earlier aggregate output and before these measurements.

10 CONCLUSION AND LIMITATIONS

Finite-reference evaluation compares error radii within one unchanged deployed MDP. It needs no recurrent convergence, improves on the direct radius exactly under Eq. (7), and enters the signed CPI bound for one fixed-base proposal. The guarantees still require uniform residual, centering, boundedness, and invariant-domain premises at one snapshot. They do not establish training convergence, recursive improvement, or learned-task performance. The eight-seed result is a conditional finite-census comparison, not a uniform certificate, monotone-depth claim, or policy-improvement result.

AI USE STATEMENT

In this work, we used generative AI tools to help develop and review theoretical models and conceptual frameworks, to formulate and refine mathematical claims, to provide critical ingredients for proving mathematical claims, to assist in writing and revising proofs, to provide feedback on research methodology and experimental design, to implement and review methods and supporting software, and to interpret numerical results. We have not used generative AI tools to generate synthetic data sets, to propose or refine hypotheses, or to assist with translation. Cleaning and reformatting datasets, and qualitative and thematic data analysis, are not applicable to this work.

Additionally, we used generative AI tools for literature search and for summarising and analysing existing literature, for drafting and editing the manuscript, for LaTeX reorganisation, for figure editing, for preparing the audit artifacts that accompany the numerical diagnostic, for cross-reference checking, for exact-arithmetic validation checks, and for reviewing the submission and its compiled PDF.

Human verification covered the theoretical development, proofs, methodology, implementation and supporting code, numerical results, and final manuscript.

REPRODUCIBILITY STATEMENT

The assumptions are stated with each result, and proofs appear in Appendices C.4 and D.3. Appendix B gives the full rational calculation for Table 1; the standard-library script `verify_finite_mdp.py` reproduces its quantities. Appendix F specifies the implementation semantics. Appendix E documents the census, statistical analysis, numerical checks, and measurement reproduction. Reproducing the per-state comparison also requires the exports held separately from the source files.

ETHICS AND BROADER IMPACT

Ethics statement. The paper studies a mathematical model of checker-trained Tiny Recursive Models on synthetic plan-editing tasks. We use no personal data, no human subjects, no crowdsourcing, and no deployed decision system. The main ethical risk is rhetorical: conditional bounds can be overstated as evidence about general reasoning or high-stakes automation.

Broader impact. The methodological contribution is to separate internal evaluator depth from outer policy improvement, which may help analyses of verifier-feedback systems with bounded internal compute. We make no deployment recommendation and keep the claims conditional on the displayed assumptions.

REFERENCES

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep equilibrium models. In *Advances in Neural Information Processing Systems*, volume 32, pp. 690–701, 2019.

Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996. ISBN 1886529108.

Kai-Wei Chang, Akshay Krishnamurthy, Alekh Agarwal, Hal Daumé III, and John Langford. Learning to search better than your teacher. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2058–2066. PMLR, 2015.

Hal Daumé III, John Langford, and Daniel Marcu. Search-based structured prediction. *Machine Learning*, 75(3):297–325, 2009.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning, 2025. URL `https://arxiv.org/abs/2501.12948`.

Alexia Jolicoeur-Martineau. Less is more: Recursive reasoning with Tiny networks. *arXiv preprint arXiv:2510.04871*, 2025.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the 19th International Conference on Machine Learning*, pp. 267–274, 2002.

Sham Machandranath Kakade. *On the Sample Complexity of Reinforcement Learning*. PhD thesis, University College London, 2003. URL https://homes.cs.washington.edu/~sham/papers/thesis/sham_thesis.pdf.

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.

Rémi Munos. Error bounds for approximate policy iteration. In *Proceedings of the Twentieth International Conference on Machine Learning*, pp. 560–567, 2003. URL https://aaai.org/papers/icml03-074-error-bounds-for-approximate-policy-iteration/.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9:815–857, 2008. URL https://www.jmlr.org/papers/volume9/munos08a/munos08a.pdf.

Andrew Y. Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pp. 278–287. Morgan Kaufmann, 1999.

Matteo Pirotta, Marcello Restelli, Alessio Pecorino, and Daniele Calandriello. Safe policy iteration. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 307–315. PMLR, 2013.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley-Interscience, 1994. ISBN 0471619779.

Stephane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635. PMLR, 2011.

Bruno Scherrer. Approximate policy iteration schemes: A comparison. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1314–1322. PMLR, 2014.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2 edition, 2018.

Aviv Tamar, Yi Wu, Garrett Thomas, Sergey Levine, and Pieter Abbeel. Value iteration networks. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Guan Wang, Jin Li, Yuhao Sun, Xing Chen, Changling Liu, Yue Wu, Meng Lu, Sen Song, and Yasin Abbasi-Yadkori. Hierarchical Reasoning Model. *arXiv preprint arXiv:2506.21734*, 2025.

Table 2: Symbols used in statements

Symbol	Meaning
h, $\bar{s} = (x, y, z, h)$	Remaining edit budget / clock-complete persistent state
$\mathcal{D}^\circ$	Declared nonabsorbing domain for a contraction invocation
$\mathcal{C}$, $\mathcal{C}^\circ$	Value domain / nonabsorbing latent domain
$\mathsf{P}_{\mathrm{rec}}$	Declared recurrent projection operator: Π_R for $R > 0$, or $\mathrm{Id}_{\mathcal{Z}}$
$\mathcal{Z}_{\mathrm{bd}}$, $V_{\max}$	Declared bounded-head latent set / its uniform finite bound
$z^{(j)}(s)$, $s \in \mathcal{C}^\circ$	Latent after finite depth j
U_j, U_*	Finite-depth value / optional fixed-point value, with scalar boundary
b_{abs}	Common scalar value of every endpoint at the absorber
$B_{n,m}$	Endpoint discrepancy plus depth-m residual radius
$\varepsilon_{\mathrm{res},n}$	Episodic finite-depth Bellman residual
$\varepsilon_{\mathrm{targ},q}$	Target-network conditional population residual at endpoint depth q
$R^P_{\mathrm{targ},q}$, $R^{\sup}_{\mathrm{targ},q}$	Propagated-lag target bound / ordinary sup-lag fallback
$d^{\pi,K}_\rho$	Discounted occupancy of the fixed K-step current-policy skeleton
$C^{\pi,K}_{\rho,\mu}$, $C^{(0)}_{\rho,\mu}$	Fixed-kernel occupancy / initial-law density-ratio bounds for the weighted-L_p restatement
$\bar{\varepsilon}_{\mathrm{res},n}$	Persistent augmented-state finite-depth residual
ε_A, $\bar{\varepsilon}_A$	Episodic / augmented advantage error bounds
$\varepsilon^*_{\mathrm{res}}$	Architectural residual at U_*
g, Δ_g	Candidate expected advantage / span on the declared mixture or pair domain
$\widehat{g}$, $\overline{\beta}_d$, M_d	Estimated candidate slope / one-sided signed-defect bound / certified slope
$\mathcal{C}_{\mathrm{pair}}$	Common π/π_{cand}-invariant domain for step intervals
c, $b(s)$, Ξ_α	Signed centering and candidate estimator defects
$\tau(s)$, δ_{dep}	Policy overlap / deployment TV radius
$\varepsilon_{\mathrm{CPI}}$	Uniform scalar bound on $\lvert g\rvert$

A Notation, domains, and absorbing normalization

The definitions below separate value evaluation, one-deviation advantage estimation, repeated mixture decisions, and deployment. Table 2 lists the notation.

Assumption A.1 (Bounded rewards)**.** Potential-shaped rewards are uniformly bounded. For some $R_{\max} < \infty$,

$$|r(s, a, s')| \le R_{\max} \quad \text{for every transition.}$$

For $R > 0$, define the Euclidean projection

$$\begin{aligned} \Pi_R(z) &= \arg\min_{u:\|u\|_2 \le R} \|u - z\|_2 \\ &= \begin{cases} z, & \|z\|_2 \le R, \\ R\,\dfrac{z}{\|z\|_2}, & \|z\|_2 > R. \end{cases} \end{aligned} \tag{33}$$

The projection Π_R is nonexpansive and leaves B_R invariant. It does not by itself give contraction. Write $\mathcal{C}^\circ = \mathcal{C} \setminus \{s_{\mathrm{abs}}\}$ for the nonabsorbing part of a value domain. The first-increment constant used by the contraction specialization is

$$C_z := \sup_{s \in \mathcal{C}^\circ} \left\| \widetilde{f}_\theta(z_{\mathrm{init}}(x(s), y(s)), y(s), x(s)) - z_{\mathrm{init}}(x(s), y(s)) \right\| < \infty. \tag{34}$$

A.1 Evaluation and policy domains

Definition A.2 (π-invariant Bellman domain)**.** A measurable set $\mathcal{C}_\pi$ is a Bellman/value domain for π if $\rho(\mathcal{C}_\pi) = 1$ whenever the initial law is used and

$$P(\mathcal{C}_\pi \mid s, a) = 1 \quad \text{for } \pi(da \mid s)\text{-almost every } a \text{ and every } s \in \mathcal{C}_\pi.$$

Table 3: Domain and theorem map. Every domain carries the stated initial law. The superscript $^\circ$ removes the absorber only for latent-path, fixed-point, and head-Lipschitz assumptions; value norms include the absorber.

Domain	Required invariance	Paths included	α?	Used by
$\mathcal{C}_\pi$	P_π	Repeated current-policy decisions	No	Bellman values, d_π, centering
$\mathcal{R}^{\mathrm{adv}}_{\pi,\pi_{\mathrm{cand}}}$	P_π after the first action	One current or candidate deviation, then π	No	Finite-reference and value-to-advantage bounds
$\mathcal{C}_{\mathrm{mix}}$	P_π and P_{π_α}	Repeated exact-mixture decisions	May	Theorem 7.3
$\mathcal{C}_{\mathrm{pair}}$	P_π and $P_{\pi_{\mathrm{cand}}}$	Repeated decisions for every exact mixture weight	No	Step-interval corollary
$\mathcal{D}_{\mathrm{dep}}$	P_{π_α} and $P_{\widetilde{\pi}}$	Repeated reference-mixture and deployed-policy decisions	Through the selected π_α	Theorem C.4
Barred counterparts	The corresponding augmented kernels with one shared F_n	The same closures on (x, y, z, h)	As above	Persistent-state results

Equivalently, the policy-induced kernel leaves $\mathcal{C}_\pi$ invariant. Theorem 5.1 needs only such a domain and does not need a candidate policy.

Definition A.3 (Advantage-evaluation domain)**.** Fix current and candidate stochastic kernels π and π_{cand}. A measurable set $\mathcal{R}^{\mathrm{adv}}_{\pi,\pi_{\mathrm{cand}}}$ contains a π-invariant domain $\mathcal{C}_\pi$ and the common absorber and satisfies

1. for every $s \in \mathcal{C}_\pi$, $P(\mathcal{R}^{\mathrm{adv}}_{\pi,\pi_{\mathrm{cand}}} \mid s, a) = 1$ for $\pi(da \mid s)$- and $\pi_{\mathrm{cand}}(da \mid s)$-almost every action; and
2. for every $s \in \mathcal{R}^{\mathrm{adv}}_{\pi,\pi_{\mathrm{cand}}}$, $P(\mathcal{R}^{\mathrm{adv}}_{\pi,\pi_{\mathrm{cand}}} \mid s, a) = 1$ for $\pi(da \mid s)$-almost every action.

The domain contains current-policy evaluation states, one current or candidate action, and every subsequent current-policy continuation. We write $\|f\|^{\mathrm{adv}}_{\infty,\pi,\pi_{\mathrm{cand}}}$ for the ordinary supremum on this set, including the scalar absorbing boundary.

Definition A.4 (Mixture-occupancy and deployment domains)**.** For a fixed α, a measurable $\mathcal{C}_{\mathrm{mix}}$ is a mixture-occupancy domain if $\rho(\mathcal{C}_{\mathrm{mix}}) = 1$ and it is invariant under both π and $\pi_\alpha = (1-\alpha)\pi + \alpha\pi_{\mathrm{cand}}$. It therefore supports the state laws at every time under the current policy and under repeated mixture decisions, not merely a one-deviation path. A deployment domain $\mathcal{D}_{\mathrm{dep}}$ satisfies $\rho(\mathcal{D}_{\mathrm{dep}}) = 1$ and is recursively invariant under both π_α and $\widetilde{\pi}$. The CPI span theorem uses $\mathcal{C}_{\mathrm{mix}}$; the deployment theorem uses $\mathcal{D}_{\mathrm{dep}}$.

For a certificate simultaneous over more than one mixture weight, a measurable common policy-pair domain $\mathcal{C}_{\mathrm{pair}}$ satisfies

$$\rho(\mathcal{C}_{\mathrm{pair}}) = 1, \qquad P_\pi(\mathcal{C}_{\mathrm{pair}} \mid s) = P_{\pi_{\mathrm{cand}}}(\mathcal{C}_{\mathrm{pair}} \mid s) = 1 \quad \text{for every } s \in \mathcal{C}_{\mathrm{pair}}. \tag{35}$$

By linearity of policy-induced kernels, every exact mixture π_α also leaves $\mathcal{C}_{\mathrm{pair}}$ invariant. Thus $\mathcal{C}_{\mathrm{pair}}$ is a valid single choice of $\mathcal{C}_{\mathrm{mix}}$ for every $\alpha \in [0, 1]$.

In finite or countable spaces these domains can be constructed with the usual positive-probability reachable sets and action supports. On arbitrary measurable action spaces we use the policy-almost-every-action clauses above; no topological action support is presumed.

Why boundedness matters. On the countable zero-reward chain $s_i \mapsto s_{i+1}$ with one action and $\gamma \in (0, 1)$, add an unreachable zero-reward absorber with $U(s_{\mathrm{abs}}) = V^\pi(s_{\mathrm{abs}}) = 0$. The unbounded function $U(s_i) = \gamma^{-i}$ has zero K-step residual for every $K \geq 1$, while its sup-norm error from $V^\pi = 0$ is infinite. This is why the residual theorem works on bounded functions, not arbitrary pointwise solutions.

A.2 Shaping Convention and Absorbing-State Normalization

We use potential-based shaping with potential $\Phi(s) = c(x, y)$ for episodic $s = (x, y, h)$ and persistent $\bar{s} = (x, y, z, h)$ (the potential ignores the clock and latent), and $\Phi(s_{\mathrm{abs}}) = C_{\max}$ on their shared absorbing atom. We define

$$r(s, a, s') = r_0(s, a, s') + \gamma\,\Phi(s') - \Phi(s), \tag{36}$$

applying shaping on *all* transitions, including the absorbing self-loop.

For the absorbing self-loop, taking $r_0(s_{\text{abs}}, a, s_{\text{abs}}) = 0$ gives

$$r(s_{\text{abs}}, a, s_{\text{abs}}) = \gamma\Phi(s_{\text{abs}}) - \Phi(s_{\text{abs}}) = (\gamma - 1)C_{\max}.$$

Thus every policy satisfies the fixed shaped boundary value

$$V^\pi(s_{\text{abs}}) = \sum_{t=0}^{\infty} \gamma^t(\gamma - 1)C_{\max} = -C_{\max}.$$

If one prefers the standard normalization $V(s_{\text{abs}}) = 0$, use the shifted potential $\tilde{\Phi}(s) := \Phi(s) - C_{\max}$. This adds the constant reward $(1 - \gamma)C_{\max}$ to every transition and shifts all returns by $+C_{\max}$ independently of the policy. Accordingly, performance differences and optimal policies are unchanged.

B Exact Finite-MDP Derivation

This section derives Table 1. It uses exact rational arithmetic and no data, learned parameter, or random sample. The state and action definitions are

State	Action	(r, S')
s_0	a_0	$(0, s_0)$
s_0	a_1	$(1, s_1)$
s_1	a_0	$(0, s_0)$
s_1	a_1	$(0, s_1)$
s_{abs}	a_0 or a_1	$(0, s_{\text{abs}})$

The current policy is $\pi(a_0 \mid s) = 1$, the candidate is $\pi_{\text{cand}}(a_1 \mid s) = 1$, and $\rho = \delta_{s_0}$. These definitions also apply at the absorber, where both actions induce the same zero-reward self-loop. Thus

$$\mathcal{C}_\pi = \{s_0, s_{\text{abs}}\}, \qquad \mathcal{R}^{\text{adv}}_{\pi,\pi_{\text{cand}}} = \mathcal{C}_{\text{mix}} = \mathcal{C}_{\text{pair}} = \mathcal{D}_{\text{dep}} = \mathcal{S}.$$

The first set is π-invariant and carries ρ. One candidate deviation from s_0 reaches s_1, after which π returns to s_0. Repeated mixture decisions remain in $\mathcal{S}$, so this is the complete relevant closure.

Finite-reference value calculation. Set $\gamma = 1/2$, $K = 1$, $n = 1$, and $m = 2$. On the scalar latent space, use the common initialization $z^{(0)}(s) = 0$, head $V_\psi(z, s) = z$, and

$$T_s(z) = u_s - \frac{1}{2}z, \qquad (u_{s_0}, u_{s_1}, u_{s_{\text{abs}}}) = \left(\frac{1}{4}, -\frac{1}{4}, 0\right).$$

The absorber therefore has the common depth-independent boundary zero. The finite endpoints are

$$U_1 = \left(\frac{1}{4}, -\frac{1}{4}, 0\right), \qquad U_2 = \left(\frac{1}{8}, -\frac{1}{8}, 0\right).$$

Under π, all nonabsorbing rewards are zero and both nonabsorbing states move to s_0, so $V^\pi = (0, 0, 0)$. Direct computation gives

$$\mathcal{T}_1^\pi U_2 = \left(\frac{1}{16}, \frac{1}{16}, 0\right), \qquad U_2 - \mathcal{T}_1^\pi U_2 = \left(\frac{1}{16}, -\frac{3}{16}, 0\right).$$

Consequently,

$$\|U_1 - U_2\|_\infty = \frac{1}{8}, \qquad \|U_2 - \mathcal{T}_1^\pi U_2\|_\infty = \frac{3}{16},$$
$$\|U_1 - V^\pi\|_\infty \le \frac{1}{8} + \frac{3/16}{1 - 1/2} = \frac{1}{2}, \qquad \|U_1 - V^\pi\|_\infty = \frac{1}{4}.$$

The certificate gives the finite error radius $1/2$ without a recurrent fixed-point argument. It is not tight: the actual error is smaller by a factor of two. For comparison, the direct bound at depth 1 (Corollary C.3) is $\varepsilon_{\text{res},1}/(1 - \gamma) = (3/8)/(1/2) = 3/4$, and Eq. (7) holds because $(1 - \gamma)\|U_1 - U_2\|_\infty = 1/16 < 3/16 = \varepsilon_{\text{res},1} - \varepsilon_{\text{res},2}$. The fixed-point specialization Eq. (84) applies with $\mathcal{Z}_{\text{inv}} = [-1/2, 1/2]$, $L_z = 1/2$, $C_z = 1/4$, and $L_V = 1$; it gives $\varepsilon^*_{\text{res}}/(1-\gamma) + L_V L_z^n C_z/(1-L_z) = 1/2 + 1/4 = 3/4$.

Exact mixture and occupancies. Take $\alpha = 1/2$. At every state, $\pi_{1/2}(a_0 \mid s) = \pi_{1/2}(a_1 \mid s) = 1/2$. On the two nonabsorbing states,

$$P_{\pi_{1/2}} = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}, \qquad r_{\pi_{1/2}} = \begin{pmatrix} 1/2 \\ 0 \end{pmatrix}.$$

The absorber is a separate zero-reward self-loop. Solving $d = (1-\gamma)\rho + \gamma dP$ gives

$$d_\pi = (1, 0, 0), \qquad d_{\pi_{1/2}} = \left(\frac{3}{4}, \frac{1}{4}, 0\right).$$

Solving the mixture Bellman equations gives $V^{\pi_{1/2}} = (3/4, 1/4, 0)$ and hence $\eta(\pi_{1/2}) = 3/4$.

Advantage, centering, and CPI lower bound. Because $V^\pi = 0$, the candidate expected advantage is

$$g = (1, 0, 0), \qquad \Delta_g = 1.$$

For the value-derived one-step estimator based on U_1, the exact action values on (s_0, s_1) are

$$\widehat{Q}_{U_1} = \begin{pmatrix} 1/8 & 7/8 \\ 1/8 & -1/8 \end{pmatrix},$$

with columns (a_0, a_1). Exact centering under the deterministic current policy subtracts the first column. Including the absorber gives

$$\widehat{A}(\cdot, a_0) = (0, 0, 0), \qquad \widehat{A}(\cdot, a_1) = \left(\frac{3}{4}, -\frac{1}{4}, 0\right).$$

Thus the centering defect and candidate defect from Eq. (22) are

$$c = (0, 0, 0), \qquad b = \left(-\frac{1}{4}, -\frac{1}{4}, 0\right), \qquad \Xi_{1/2} = -\frac{1}{8}.$$

The estimated surrogate is

$$\widehat{L}_\pi(\pi_{1/2}) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\pi, a\sim\pi_{1/2}}\widehat{A}(s,a) = 2\left(\frac{1}{2}\cdot\frac{3}{4}\right) = \frac{3}{4}.$$

The exact span penalty is

$$\frac{\gamma\alpha^2\Delta_g}{(1-\gamma)(1-\gamma+\gamma\alpha)} = \frac{1}{3}.$$

Equation (25) therefore yields

$$\eta(\pi_{1/2}) \geq \frac{3}{4} - \frac{-1/8}{1/2} - \frac{1}{3} = \frac{2}{3},$$

while the exact return is $3/4$. The absolute uniform specialization instead uses $\varepsilon_{A,\mathrm{cand}} = 1/4$ and $\varepsilon_{\mathrm{CPI}} = 1$, producing

$$\frac{3}{4} - \frac{1}{4} - \frac{2}{3} = -\frac{1}{6}.$$

This valid but vacuous bound shows why the signed defect and the advantage span matter. The standard-library script `verify_finite_mdp.py` reproduces this calculation.

C Complete formal statements, constructions, and proofs

C.1 Persistent latents on the augmented MDP

C.1.1 Augmented state and endpoint construction

Corollary C.2 uses the following augmented MDP construction, which permits a continuous carried latent $z \in \mathbb{R}^d$.

Fix n, (θ, ψ), the policy parameters, and all environment kernels. Let $h \in \{1, 2, \ldots, H\}$ be the number of edits remaining before the current action. Define

$$F_n(x, y, z) := T^{\circ n}_{x,y}(z) = \widetilde{f}^{\circ n}_\theta(z, y, x), \qquad \bar{s} = (x, y, z, h). \tag{37}$$

For a nonabsorbing state, first compute F_n and sample an action. If the environment transition terminates for any reason (including a solution or terminal STOP), or if $h = 1$, the successor is the shared atom s_{abs}; budget exhaustion is terminal. Otherwise the clock-complete successor is

$$\bar{s}' = \big(x, \mathrm{edit}(y, a; x), F_n(x, y, z), h - 1\big).$$

The state stores the pre-unroll carried latent z, the policy uses the post-unroll latent $F_n(x, y, z)$, and a nonterminal successor carries that post-unroll latent. The current and candidate action kernels share this same frozen F_n and hence the same augmented MDP, as in Eq. (30). Changing the recurrent map between these two policies would change the transition kernel and is outside ordinary CPI.

Remark C.1 (Why the recurrent map is shared). A small Euclidean discrepancy between deterministic recurrent maps does not imply a small policy or augmented-kernel discrepancy. For example, take $F_\varepsilon(z) = -\varepsilon/2$ and $G_\varepsilon(z) = \varepsilon/2$ and a fixed threshold policy head that chooses disjoint actions below and above zero. Then $\sup_z \|F_\varepsilon(z) - G_\varepsilon(z)\| = \varepsilon$, but the action laws are disjoint and a fixed MDP can have an order-$1/(1-\gamma)$ return gap. The transition kernels are also far apart: $\mathrm{TV}(\delta_{F(z)}, \delta_{G(z)}) = 1$ whenever $F(z) \neq G(z)$. A different-map theorem therefore needs explicit reward and transition-kernel discrepancies and sufficient regularity to propagate them. Map distance alone is insufficient.

The initial law is

$$\bar{\rho}(dx, dy, dz, dh) = \rho(dx, dy, dh)\, \delta_{z_{\mathrm{init}}(x,y)}(dz), \tag{38}$$

or, for stochastic initialization, $\bar{\rho}(dx, dy, dz, dh) = \rho(dx, dy, dh)\nu_0(dz \mid x, y, h)$ with any recorded measurable initialization kernel. Ignoring h is an architectural specialization, not a Bellman requirement.

For the result below, freeze n, (θ, ψ), the initialization law, and all reward/edit and policy kernels. Assume the initialization law, recurrent map, value head, and these kernels are measurable. Together with the common absorber, they define one fixed augmented MDP. Its K-step Bellman operator $\bar{\mathcal{T}}^{\bar{\pi}}_K$ uses the deployed F_n at every environment transition; comparison depth q changes only the endpoint in Eq. (29). The direct residual and surrogate are Eqs. (39) and (44), with discounted return measured from $\bar{\rho}$.

Figure 1 shows the deployment lane and the comparison lane. Comparison with a memoryless fixed-point reference would require a drift assumption that this paper does not make.

C.1.2 Full persistent-state statement

The following is the complete formal statement summarized in Section 8.

Corollary C.2 (Persistent-state evaluation and improvement)**.** *Let $\gamma \in [0, 1)$ and $K \geq 1$. Let $\bar{\mathcal{C}}_\pi$ be a measurable $\bar{\pi}$-invariant domain carrying $\bar{\rho}$, and let $\bar{\mathcal{C}}$ be a measurable augmented advantage-evaluation domain as in Definition A.3, containing $\bar{\mathcal{C}}_\pi$ and the common absorber. Assume the preceding frozen-MDP and measurability conditions, bounded K-step expected rewards, and $\widetilde{U}_n \in B_b(\bar{\mathcal{C}})$ with the common scalar boundary. Use the ordinary pointwise norm on $\bar{\mathcal{C}}$ and define*

$$\bar{\varepsilon}_{\mathrm{res},n} := \|\widetilde{U}_n - \bar{\mathcal{T}}^{\bar{\pi}}_K \widetilde{U}_n\|^{\mathrm{adv}}_{\infty,\bar{\pi},\bar{\pi}_{\mathrm{cand}}}. \tag{39}$$

The Bellman operator has a unique $V^{\bar{\pi}} \in B_b(\bar{\mathcal{C}})$, and

$$\|\widetilde{U}_n - V^{\bar{\pi}}\|^{\mathrm{adv}}_{\infty,\bar{\pi},\bar{\pi}_{\mathrm{cand}}} \leq \frac{\bar{\varepsilon}_{\mathrm{res},n}}{1 - \gamma^K}. \tag{40}$$

If $\widehat{A}_{\widetilde{U}_n}$ uses the exact transition expectation and exact baseline and is therefore centered pointwise under $\bar{\pi}$ on the full augmented state, then Assumption 7.1 holds for every $\bar{s} \in \bar{\mathcal{C}}_\pi$ and for $\bar{\pi}$- and $\bar{\pi}_{\mathrm{cand}}$-almost every action, with the valid choice

$$\bar{\varepsilon}_A := \frac{2\gamma\, \bar{\varepsilon}_{\mathrm{res},n}}{1 - \gamma^K}. \tag{41}$$

If, additionally, the analyzed policy is the pointwise exact mixture $\bar{\pi}_{\mathrm{new}} = (1-\alpha)\bar{\pi} + \alpha\bar{\pi}_{\mathrm{cand}}$ with no parameter interpolation or distillation, let $\bar{\mathcal{C}}_{\mathrm{mix}}$ be a measurable augmented domain carrying $\bar{\rho}$, containing $\bar{\mathcal{C}}_\pi$, and invariant under both $\bar{\pi}$ and $\bar{\pi}_{\mathrm{new}}$. Define

$$\bar{\varepsilon}_{\mathrm{CPI}} := \sup_{\bar{s}\in\bar{\mathcal{C}}_{\mathrm{mix}}} \left|\mathbb{E}_{a\sim\bar{\pi}_{\mathrm{cand}}(\cdot|\bar{s})} A^{\bar{\pi}}(\bar{s},a)\right|, \tag{42}$$

$$\bar{\tau}_\infty := \sup_{\bar{s}\in\bar{\mathcal{C}}_\pi} \mathrm{TV}(\bar{\pi}_{\mathrm{cand}}(\cdot \mid \bar{s}), \bar{\pi}(\cdot \mid \bar{s})), \tag{43}$$

where $\bar{\mathcal{C}}_\pi \subseteq \bar{\mathcal{C}} \cap \bar{\mathcal{C}}_{\mathrm{mix}}$ is the common full-measure current-policy domain. With $d_{\bar{\pi}} = (1-\gamma)\sum_{t\geq 0}\gamma^t \bar{\rho} P_{\bar{\pi}}^t$ and $\widehat{A} = \widehat{A}_{\widetilde{U}_n}$, define the full-state surrogate

$$\widehat{L}_{\bar{\pi}}(\bar{\pi}') := \eta(\bar{\pi}) + \frac{1}{1-\gamma}\mathbb{E}_{\bar{s}\sim d_{\bar{\pi}},\, a\sim\bar{\pi}'(\cdot|\bar{s})}[\widehat{A}(\bar{s},a)]. \tag{44}$$

Theorem 7.3 gives

$$\eta(\bar{\pi}_{\mathrm{new}}) \geq \widehat{L}_{\bar{\pi}}(\bar{\pi}_{\mathrm{new}}) - \frac{2\alpha\gamma\bar{\tau}_\infty}{1-\gamma}\frac{\bar{\varepsilon}_{\mathrm{res},n}}{1-\gamma^K} - \frac{2\bar{\varepsilon}_{\mathrm{CPI}}\gamma\alpha^2}{(1-\gamma)(1-\gamma+\gamma\alpha)}. \tag{45}$$

For the target-network and finite-reference forms, apply Corollary C.3 with barred quantities: use a finite target depth $q \geq 0$, $\widetilde{U}_q, \bar{V} \in B_b(\bar{\mathcal{C}})$ with the common boundary, a target trajectory under the frozen augmented policy, and exact transition expectation and baseline conditional on the full augmented state. All corresponding domain and exact-mixture hypotheses remain in force.

C.2 Residual-to-CPI certificates

Corollary C.3 (Self-bootstrap and target-network residual certificates)**.** *Use the measurable Bellman domain of Theorem 5.1 with $\mathcal{C} = \mathcal{R}^{\mathrm{adv}}_{\pi,\pi_{\mathrm{cand}}}$. Take $U_n \in B_b(\mathcal{C})$ and use the common scalar boundary $U_n(s_{\mathrm{abs}}) = V^\pi(s_{\mathrm{abs}}) = b_{\mathrm{abs}}$. Define*

$$\varepsilon_{\mathrm{res},n} := \|U_n - \mathcal{T}_K^\pi U_n\|^{\mathrm{adv}}_{\infty,\pi,\pi_{\mathrm{cand}}}. \tag{46}$$

Then

$$\|U_n - V^\pi\|^{\mathrm{adv}}_{\infty,\pi,\pi_{\mathrm{cand}}} \leq \frac{\varepsilon_{\mathrm{res},n}}{1-\gamma^K}. \tag{47}$$

If $\widehat{A} = \widehat{A}_{U_n}$ is the exact one-step centered estimator of Eq. (2)*, with its transition expectation and baseline sum under π evaluated exactly, then Assumption 7.1 holds with*

$$\varepsilon_A := \frac{2\gamma\,\varepsilon_{\mathrm{res},n}}{1-\gamma^K}. \tag{48}$$

Under Theorem 7.3,

$$\eta(\pi_\alpha) \geq \widehat{L}_\pi(\pi_\alpha) - \frac{2\alpha\gamma\tau_\infty}{1-\gamma}\,\frac{\varepsilon_{\mathrm{res},n}}{1-\gamma^K} - \frac{2\varepsilon_{\mathrm{CPI}}\gamma\alpha^2}{(1-\gamma)(1-\gamma+\gamma\alpha)}. \tag{49}$$

For a finite depth $q \geq 0$, instead suppose Proposition 6.1 holds on the same domain with $\mathcal{E} = \mathcal{C}$. Then

$$\|U_q - V^\pi\|^{\mathrm{adv}}_{\infty,\pi,\pi_{\mathrm{cand}}} \leq \frac{R^P_{\mathrm{targ},q}}{1-\gamma^K}. \tag{50}$$

For finite depths $0 \leq n < m$, if $U_n, U_m \in B_b(\mathcal{C})$ use the same scalar boundary and the depth-m target-network quantities are established, then

$$\|U_n - V^\pi\|^{\mathrm{adv}}_{\infty,\pi,\pi_{\mathrm{cand}}} \leq \|U_n - U_m\|^{\mathrm{adv}}_{\infty,\pi,\pi_{\mathrm{cand}}} + \frac{R^P_{\mathrm{targ},m}}{1-\gamma^K}. \tag{51}$$

For the depth-q target statement, take $\widehat{A} = \widehat{A}_{U_q}$ from Eq. (2)*, with its transition expectation and current-policy baseline evaluated exactly, and define $\varepsilon_{A,\mathrm{cand}}$ by Eq.* (27)*. Then*

$$\varepsilon_{A,\mathrm{cand}} \leq 2\gamma\tau_\infty \frac{R^P_{\mathrm{targ},q}}{1-\gamma^K}. \tag{52}$$

Under the exact-mixture and scalar-span hypotheses of Theorem 7.3,

$$\eta(\pi_\alpha) \geq \widehat{L}_\pi(\pi_\alpha) - \frac{2\alpha\gamma\tau_\infty}{(1-\gamma)(1-\gamma^K)} R^P_{\mathrm{targ},q} - \frac{2\varepsilon_{\mathrm{CPI}}\gamma\alpha^2}{(1-\gamma)(1-\gamma+\gamma\alpha)}, \tag{53}$$

$$\eta(\pi_\alpha) \geq \widehat{L}_\pi(\pi_\alpha) - \frac{2\alpha\gamma\tau_\infty}{1-\gamma}\left(\|U_n - U_m\|^{\mathrm{adv}}_{\infty,\pi,\pi_{\mathrm{cand}}} + \frac{R^P_{\mathrm{targ},m}}{1-\gamma^K}\right) - \frac{2\varepsilon_{\mathrm{CPI}}\gamma\alpha^2}{(1-\gamma)(1-\gamma+\gamma\alpha)}. \tag{54}$$

The second display uses the same exact value-derived construction at depth n. Every target expression retains the propagated-lag quantity $R^P_{\mathrm{targ},q}$ before the ordinary sup-lag fallback $R^P_{\mathrm{targ},q} \leq R^{\mathrm{sup}}_{\mathrm{targ},q}$.

Proof in Appendix C.4.15. The common-domain safe-step result is Corollary C.11.

C.3 Deployment mismatch

Theorem C.4 (First-mismatch deployment perturbation)**.** *Fix $0 \leq \gamma < 1$ and $\pi_\alpha = (1-\alpha)\pi + \alpha\pi_{\mathrm{cand}}$ for $\alpha \in [0,1]$. Let $\widetilde{\pi}$ be a deployed stochastic kernel in the same fixed MDP as π_α, with the same initial law ρ. Assume the state and action spaces are standard Borel, and let $\mathsf{K}(dr, ds' \mid s, a)$ be the measurable joint reward-transition kernel.*

Let $\mathcal{D}_{\mathrm{dep}}$ be measurable with $\rho(\mathcal{D}_{\mathrm{dep}}) = 1$, and assume the policy-induced transition kernels of both π_α and $\widetilde{\pi}$ leave $\mathcal{D}_{\mathrm{dep}}$ invariant. For every $s \in \mathcal{D}_{\mathrm{dep}}$, assume K assigns rewards in $[r_{\min}, r_{\max}]$ for $\pi_\alpha(da \mid s)$- and $\widetilde{\pi}(da \mid s)$-almost every action. Write $\Delta_r := r_{\max} - r_{\min}$ and suppose

$$\sup_{s \in \mathcal{D}_{\mathrm{dep}}} \mathrm{TV}(\widetilde{\pi}(\cdot \mid s), \pi_\alpha(\cdot \mid s)) \leq \delta_{\mathrm{dep}}, \qquad \delta_{\mathrm{dep}} \in [0,1]. \tag{55}$$

Then

$$|\eta(\widetilde{\pi}) - \eta(\pi_\alpha)| \leq \frac{\Delta_r \delta_{\mathrm{dep}}}{(1-\gamma)(1-\gamma+\gamma\delta_{\mathrm{dep}})}. \tag{56}$$

The extra denominator factor satisfies $1-\gamma+\gamma\delta_{\mathrm{dep}} \geq 1-\gamma$, so the first-mismatch constant improves the generic perturbation constant.

Proof in Appendix C.4.12.

Remark C.5 (Attainment by a first-mismatch process). The constant in Theorem C.4 is attained by a two-state MDP. Let ℓ be initial and u absorbing. At ℓ, action a_0 gives reward 0 and returns to ℓ, while a_1 gives reward Δ_r and enters u; every action at u gives reward Δ_r and remains there. Let π_α choose a_0 at ℓ, and let $\widetilde{\pi}$ choose a_1 with probability δ_{dep} and a_0 otherwise. Let the policies agree at u. The reward difference at time t is exactly $\Delta_r[1-(1-\delta_{\mathrm{dep}})^{t+1}]$, so equality holds in Eq. (56).

Corollary C.6 (Either-direction uniform-KL deployment bound)**.** *Under Theorem C.4, suppose either*

$$\sup_{s \in \mathcal{D}_{\mathrm{dep}}} D_{\mathrm{KL}}(\widetilde{\pi}(\cdot \mid s) \| \pi_\alpha(\cdot \mid s)) \leq \kappa, \quad \textit{or}$$
$$\sup_{s \in \mathcal{D}_{\mathrm{dep}}} D_{\mathrm{KL}}(\pi_\alpha(\cdot \mid s) \| \widetilde{\pi}(\cdot \mid s)) \leq \kappa$$

for a finite κ. Define

$$\bar{\delta}_\kappa := \min\left\{1, \sqrt{\frac{\kappa}{2}}\right\}. \tag{57}$$

Then

$$|\eta(\widetilde{\pi}) - \eta(\pi_\alpha)| \leq \frac{\Delta_r \bar{\delta}_\kappa}{(1-\gamma)(1-\gamma+\gamma\bar{\delta}_\kappa)}. \tag{58}$$

Proof in Appendix C.4.13.

Corollary C.7 (Occupancy-aware CPI with deployment discrepancy)**.** *Assume Theorems 7.3 and C.4 hold in one fixed MDP with one initial law. Let $\mathcal{C}_{\mathrm{mix}}$ satisfy Definition A.4, and let $\mathcal{D}_{\mathrm{dep}}$ be measurable, carry the same initial law, and be recursively invariant under both π_α and $\widetilde{\pi}$. Assume the*

standard-Borel, bounded-reward, and uniform-TV conditions of Theorem C.4. For any established one-sided bound $B_\alpha \geq \Xi_\alpha$,

$$\eta(\widetilde{\pi}) \geq \widehat{L}_\pi(\pi_\alpha) - \frac{B_\alpha}{1-\gamma} - \frac{\gamma\alpha^2\Delta_g}{(1-\gamma)(1-\gamma+\gamma\alpha)} - \frac{\Delta_r\delta_{\mathrm{dep}}}{(1-\gamma)(1-\gamma+\gamma\delta_{\mathrm{dep}})}. \tag{59}$$

A distilled, logit-interpolated, hidden-state-interpolated, parameter-interpolated, or otherwise realized policy is covered only when it acts in this same MDP and its TV or KL discrepancy from the exact probability-space mixture is established uniformly on $\mathcal{D}_{\mathrm{dep}}$. For persistent state (x, y, z, h), all policies must use the same frozen recurrent map that determines the carried-latent transition. Changing that map changes the MDP.

Proof in Appendix C.4.14.

C.4 Proofs

C.4.1 Finite-reference value bound

Proof of Theorem 5.1. Write $\|\cdot\|_\infty$ for the ordinary sup-norm on $\mathcal{C}$. Invariance and the Markov-kernel measurability property imply $P_\pi^K V \in B_b(\mathcal{C})$ whenever $V \in B_b(\mathcal{C})$; bounded r_K^π therefore makes $\mathcal{T}_K^\pi$ a self-map. For $V, W \in B_b(\mathcal{C})$,

$$\|\mathcal{T}_K^\pi V - \mathcal{T}_K^\pi W\|_\infty \leq \gamma^K \|V - W\|_\infty.$$

The bounded-reward premise makes the ordinary discounted policy value well defined, bounded, and measurable, with the uniformly convergent representation

$$V^\pi = \sum_{t=0}^{\infty} \gamma^t P_\pi^t r^\pi.$$

Splitting this series after K terms gives $V^\pi = \mathcal{T}_K^\pi V^\pi$. Completeness and Banach's theorem then identify this policy value as the unique fixed point of $\mathcal{T}_K^\pi$. Thus

$$\begin{aligned}\|U_m - V^\pi\|_\infty &\leq \|U_m - \mathcal{T}_K^\pi U_m\|_\infty + \|\mathcal{T}_K^\pi U_m - \mathcal{T}_K^\pi V^\pi\|_\infty \\ &\leq \|U_m - \mathcal{T}_K^\pi U_m\|_\infty + \gamma^K \|U_m - V^\pi\|_\infty.\end{aligned}$$

Rearranging and then applying the triangle inequality to $U_n - U_m + U_m - V^\pi$ proves (6). □

With a finite edit budget, a common terminal boundary and uniformly controlled stagewise residuals, the same backward recursion gives a finite geometric sum, including the final partial block; this refinement is not used here.

Noncontractive endpoints. The scalar map $F(z) = -z$ is nonconvergent away from zero, while endpoints whose depths have the same parity agree exactly. The finite-reference bound still requires the reference endpoint's Bellman residual.

C.4.2 Recurrent endpoint paths

Corollary C.8 (Recurrent endpoint path bound)**.** *Under Theorem 5.1, take $\mathcal{C} = \mathcal{R}^{\mathrm{adv}}_{\pi,\pi_{\mathrm{cand}}}$ and $\mathcal{C}^\circ = \mathcal{C} \setminus \{s_{\mathrm{abs}}\}$. On $\mathcal{C}^\circ$, fix a jointly measurable recurrent map $(s, z) \mapsto T(s, z)$, a common measurable initialization into its declared latent domain, and a latent norm, and define $z^{(j)}$ by iterating that map and $U_j(s) = V_\psi(z^{(j)}(s), x(s), y(s))$. At the absorber define only the common scalar boundary $U_j(s_{\mathrm{abs}}) = b_{\mathrm{abs}}$, and assume the two endpoint functions lie in $B_b(\mathcal{C})$. If the measurable scalar head*

is L_V-Lipschitz in its latent argument uniformly while (x, y) is held fixed, then

$$\|U_n - U_m\|^{\mathrm{adv}}_{\infty,\pi,\pi_{\mathrm{cand}}} \le L_V \sup_{s\in\mathcal{C}^\circ} \|z^{(n)}(s) - z^{(m)}(s)\| \tag{60}$$

$$\le L_V \sup_{s\in\mathcal{C}^\circ} \sum_{j=n}^{m-1} \|z^{(j+1)}(s) - z^{(j)}(s)\| \tag{61}$$

$$\le L_V \sum_{j=n}^{m-1} \sup_{s\in\mathcal{C}^\circ} \|z^{(j+1)}(s) - z^{(j)}(s)\|. \tag{62}$$

If $\mathcal{C}^\circ$ is empty, the nonnegative latent suprema in this chain are defined to be zero.

Proof of Corollary C.8. For every nonabsorbing s, Lipschitzness of the scalar head and the telescoping latent path give

$$|U_n(s) - U_m(s)| \le L_V \|z^{(n)}(s) - z^{(m)}(s)\|$$
$$\le L_V \sum_{j=n}^{m-1} \|z^{(j+1)}(s) - z^{(j)}(s)\|.$$

The left side is zero at s_{abs} because its value is depth independent. Taking the supremum only over $\mathcal{C}^\circ$ for the latent endpoints proves Eq. (60). The triangle inequality for the telescoping latent path proves Eq. (61); using $\sup_s \sum_j g_j(s) \le \sum_j \sup_s g_j(s)$ proves Eq. (62). No absorbing latent is defined. □

C.4.3 VALUE-TO-ADVANTAGE ERROR

Lemma C.9 (Value to one-step advantage error)**.** *Let $\mathcal{R}^{\mathrm{adv}}_{\pi,\pi_{\mathrm{cand}}}$ be the advantage-evaluation closure from Definition A.3, and let $\delta_V := \|V - V^\pi\|^{\mathrm{adv}}_{\infty,\pi,\pi_{\mathrm{cand}}}$. Then for every $s \in \mathcal{C}_\pi$ and for $\pi(da \mid s)$- and $\pi_{\mathrm{cand}}(da \mid s)$-almost every a,*

$$|\widehat{A}_V(s,a) - A^\pi(s,a)| \le 2\gamma\,\delta_V.$$

Proof of Lemma C.9. For any (s, a),

$$\widehat{Q}_V(s,a) - Q^\pi(s,a) = \gamma\,\mathbb{E}_{s'\sim P(\cdot|s,a)}\big[V(s') - V^\pi(s')\big].$$

If $s \in \mathcal{C}_\pi$, Definition A.3 gives $P(\mathcal{R}^{\mathrm{adv}}_{\pi,\pi_{\mathrm{cand}}} \mid s, a) = 1$ for $\pi(da \mid s)$- and $\pi_{\mathrm{cand}}(da \mid s)$-almost every a. Thus $|V(s') - V^\pi(s')| \le \delta_V$ for transition-almost every s', and hence $|\widehat{Q}_V(s,a) - Q^\pi(s,a)| \le \gamma\delta_V$.

Similarly,

$$\Big|\mathbb{E}_{a'\sim\pi(\cdot|s)}\widehat{Q}_V(s,a') - \mathbb{E}_{a'\sim\pi(\cdot|s)}Q^\pi(s,a')\Big| \le \gamma\delta_V,$$

since $a' \sim \pi(\cdot|s)$ and the same almost-sure closure condition applies. Subtracting these two bounds yields $|\widehat{A}_V(s,a) - A^\pi(s,a)| \le 2\gamma\delta_V$. □

Proof of Corollary 5.2. Theorem 5.1 gives value error at most $B_{n,m}$; Lemma C.9 multiplies this radius by 2γ. □

Remark C.10 (Extensions to K-step and GAE-style advantages). Lemma C.9 is stated for the *one-step* estimator (2), hence the factor 2γ. If instead one defines a K-step action-value estimate

$$\widehat{Q}^{(K)}_V(s,a) := \mathbb{E}\Big[\sum_{t=0}^{K-1}\gamma^t r_t + \gamma^K V(s_K) \;\Big|\; s_0 = s, a_0 = a, \pi\Big],$$

and the corresponding centered advantage $\widehat{A}^{(K)}_V(s,a) := \widehat{Q}^{(K)}_V(s,a) - \mathbb{E}_{a'\sim\pi(\cdot|s)}[\widehat{Q}^{(K)}_V(s,a')]$ (with the baseline computed exactly), then the same argument yields

$$|\widehat{A}^{(K)}_V(s,a) - A^\pi(s,a)| \le 2\gamma^K\,\delta_V.$$

λ-return or GAE-style estimators lie outside the displayed bridge unless their estimator-specific uniform error is separately established. Such an error may be supplied directly as ε_A in Assumption 7.1; no GAE coefficient is derived here.

C.4.4 Proof of Corollary 5.3

Proof of Corollary 5.3. Let $q_e(s,a) = \gamma \mathbb{E}[e(S') \mid s,a]$. Exact centering gives $b(s) = \mathbb{E}_{\pi_{\mathrm{cand}}} q_e - \mathbb{E}_\pi q_e$. By Definition A.3, there is a common action set of full $\pi(\cdot \mid s)$- and $\pi_{\mathrm{cand}}(\cdot \mid s)$-measure on which $q_e(s,a) \in [-\gamma\delta_V, \gamma\delta_V]$. The total-variation expectation inequality on this common essential range proves Eq. (11); taking a supremum or a d_π expectation gives the two corollaries. For a K-step estimator, $q_e^{(K)}(s,a) = \gamma^K \mathbb{E}[e(S_K) \mid s,a]$ has the corresponding essential range $[-\gamma^K\delta_V, \gamma^K\delta_V]$. □

C.4.5 Target-network residuals

Proof of Proposition 6.1. Add and subtract the target-network population backup:

$$U_q - \mathcal{T}_K^\pi U_q = \left[U_q - \left(r_K^\pi + \gamma^K P_\pi^K \bar{V}\right)\right] + \gamma^K P_\pi^K (\bar{V} - U_q).$$

The triangle inequality proves Eq. (17). Define the ordinary sup-lag fallback by

$$R_{\mathrm{targ},q}^{\sup} := \varepsilon_{\mathrm{targ},q} + \gamma^K \|\bar{V} - U_q\|_{\infty,\mathcal{E}}. \tag{63}$$

Invariance makes P_π^K a Markov operator on $B_b(\mathcal{E})$, so nonexpansiveness gives $R_{\mathrm{targ},q}^P \le R_{\mathrm{targ},q}^{\sup}$. □

C.4.6 Proof of Lemma 7.2

Proof of Lemma 7.2. Realize each mixture decision with an independent Bernoulli selector. The event that decisions $0, \dots, t-1$ all select the current-policy component has probability $(1-\alpha)^t$ and, conditional on that event, produces the current-policy state law before decision t. Therefore

$$\mathrm{TV}(\rho P_{\pi_\alpha}^t, \rho P_\pi^t) \le 1 - (1-\alpha)^t.$$

Convexity of total variation and the geometric identity give

$$\begin{aligned}\mathrm{TV}(d_{\pi_\alpha}, d_\pi) &\le (1-\gamma) \sum_{t=0}^{\infty} \gamma^t [1 - (1-\alpha)^t] \\ &= \frac{\gamma\alpha}{1-\gamma+\gamma\alpha}.\end{aligned}$$

The exponent is t, not $t+1$, because occupancy is measured before decision t. The formula gives zero at $\alpha = 0$ or $\gamma = 0$ and γ at $\alpha = 1$, and improves the generic bound because $1-\gamma+\gamma\alpha \ge 1-\gamma$. In the countable case this gives $2\gamma\alpha/(1-\gamma+\gamma\alpha) \le 2\gamma\alpha/(1-\gamma)$. □

C.4.7 CPI with signed evaluation error

Proof of Theorem 7.3. Linearity of the exact mixture and $\mathbb{E}_{a\sim\pi(\cdot|s)} A^\pi(s,a) = 0$ give

$$\begin{aligned}\widehat{L}_\pi(\pi_\alpha) - L_\pi(\pi_\alpha) &= \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_\pi} \left[(1-\alpha)c(s) + \alpha b(s)\right] \\ &= \frac{\Xi_\alpha}{1-\gamma},\end{aligned}$$

which proves Eq. (24).

The performance-difference identity also gives

$$\eta(\pi_\alpha) - L_\pi(\pi_\alpha) = \frac{\alpha}{1-\gamma} \left(\mathbb{E}_{d_{\pi_\alpha}} g - \mathbb{E}_{d_\pi} g\right).$$

For any bounded measurable g and probability measures p, q, $|\mathbb{E}_p g - \mathbb{E}_q g| \le \mathrm{span}(g)\, \mathrm{TV}(p,q)$. Applying this inequality on $\mathcal{C}_{\mathrm{mix}}$ and then Lemma 7.2 yields

$$\eta(\pi_\alpha) - L_\pi(\pi_\alpha) \ge -\frac{\gamma\alpha^2 \Delta_g}{(1-\gamma)(1-\gamma+\gamma\alpha)}.$$

Combining the last display with the exact surrogate identity proves Eq. (25). Any $B_\alpha \ge \Xi_\alpha$ may replace the signed term. Under exact centering, $c = 0$ and $\Xi_\alpha = \alpha \mathbb{E}_{d_\pi} b \le \alpha \varepsilon_{A,\mathrm{cand}}$. Finally, $\Delta_g \le 2 \sup_s |g(s)| = 2\varepsilon_{\mathrm{CPI}}$, proving Eq. (28). An occupancy-averaged absolute choice is

$$B_\alpha = (1-\alpha)\mathbb{E}_{d_\pi}|c(s)| + \alpha \mathbb{E}_{d_\pi}|b(s)|, \tag{64}$$

which upper-bounds Ξ_α; substituting it in Eq. (25) gives the absolute-defect specialization. □

C.4.8 SAFE MIXTURE STEPS

Corollary C.11 (Span-based safe mixture step)**.** *In the setting of Theorem 7.3, fix one measurable common policy-pair domain $\mathcal{C}_{\mathrm{pair}}$ from Definition A.4, with $\mathcal{C}_\pi \subseteq \mathcal{C}_{\mathrm{pair}}$. Use one fixed estimator $\widehat{A}$ satisfying exact statewise centering on $\mathcal{C}_\pi$ for every $\alpha \in [0,1]$, and define its fixed candidate-defect function*

$$b(s) := \mathbb{E}_{a\sim\pi_{\mathrm{cand}}(\cdot|s)}[\widehat{A}(s,a) - A^\pi(s,a)]. \tag{65}$$

Neither the estimator nor this defect object depends on α. Define, once on $\mathcal{C}_{\mathrm{pair}}$,

$$g(s) := \mathbb{E}_{a\sim\pi_{\mathrm{cand}}(\cdot|s)}[A^\pi(s,a)], \qquad \Delta_g := \sup_{s\in\mathcal{C}_{\mathrm{pair}}} g(s) - \inf_{s\in\mathcal{C}_{\mathrm{pair}}} g(s), \tag{66}$$

and assume this function is bounded and measurable there. Let $\overline{\beta}_d$ be any finite, α-independent, established one-sided bound satisfying

$$\overline{\beta}_d \geq \mathbb{E}_{s\sim d_\pi} b(s). \tag{67}$$

Define

$$\widehat{g} := \mathbb{E}_{s\sim d_\pi,\, a\sim\pi_{\mathrm{cand}}(\cdot|s)}[\widehat{A}(s,a)], \qquad M_d := \widehat{g} - \overline{\beta}_d. \tag{68}$$

Then, simultaneously for every $\alpha \in [0,1]$,

$$\eta(\pi_\alpha) - \eta(\pi) \geq \frac{\alpha}{1-\gamma}\left[M_d - \frac{\gamma\alpha\Delta_g}{1-\gamma+\gamma\alpha}\right]. \tag{69}$$

Thus $\eta(\pi_\alpha) \geq \eta(\pi)$ is certified whenever

$$M_d \geq \frac{\gamma\alpha\Delta_g}{1-\gamma+\gamma\alpha}. \tag{70}$$

Three immediate choices preserve the same fixed defect object:

1. *the exact signed defect, $\overline{\beta}_d = \mathbb{E}_{d_\pi} b(s)$;*
2. *the occupancy-averaged absolute defect, $\overline{\beta}_d = \mathbb{E}_{d_\pi}|b(s)|$;*
3. *the uniform specialization already used above, $\overline{\beta}_d = \varepsilon_{A,\mathrm{cand}}$ from Eq.* (27)*.*

The third choice recovers the previous uniform statement as an immediate specialization. Appendix C.4.9 gives the exact positive-step threshold and all boundary cases. Under Corollaries 5.2 and 5.3, with the finite-reference estimator $\widehat{A} = \widehat{A}_{U_n}$, one may use $\mathbb{E}_{d_\pi}|b(s)| \leq 2\gamma B_{n,m}\mathbb{E}_{d_\pi}\tau(s)$ for the occupancy-averaged specialization, or $\varepsilon_{A,\mathrm{cand}} \leq 2\gamma B_{n,m}\tau_\infty$ for the uniform specialization. The bound $\tau_\infty \leq 1$ recovers the previous uniform fallback.

C.4.9 SAFE-STEP BOUNDARY CASES

For Corollary C.11, the zero step $\alpha = 0$ is equality. At $\alpha = 1$, the bound certifies the full candidate step exactly when $M_d \geq \gamma\Delta_g$. The remaining nonredundant cases are:

- if $\gamma = 0$ and $M_d \geq 0$, every $\alpha \in [0,1]$ is certified;
- if $M_d < 0$, no positive step is certified by this bound;
- if $\gamma > 0$, $M_d = 0$, and $\Delta_g > 0$, no positive step is certified;
- if $M_d = \Delta_g = 0$, every step has a zero lower bound;
- if $\gamma > 0$, $M_d > 0$, and $M_d \geq \gamma\Delta_g$, every $\alpha \in [0,1]$ is certified;
- if $\gamma > 0$ and $0 < M_d < \gamma\Delta_g$, every

$$\alpha \leq \frac{M_d(1-\gamma)}{\gamma(\Delta_g - M_d)} \tag{71}$$

  is certified. The denominator is positive and the threshold lies strictly between zero and one.

Proof of Corollary C.11. The invariance of $\mathcal{C}_{\mathrm{pair}}$ under both endpoint policies implies its invariance under every π_α. Hence Theorem 7.3 applies for each $\alpha \in [0,1]$ with the same choice $\mathcal{C}_{\mathrm{mix}} = \mathcal{C}_{\mathrm{pair}}$, the same Δ_g, estimator, defect function, and one-sided defect bound. Exact centering and the fixed candidate-defect definition give

$$\widehat{g} = \mathbb{E}_{s\sim d_\pi}[g(s) + b(s)]. \tag{72}$$

Consequently,

$$\mathbb{E}_{d_\pi} g = \widehat{g} - \mathbb{E}_{d_\pi} b \geq \widehat{g} - \overline{\beta}_d = M_d. \tag{73}$$

The performance-difference identity, the span/TV expectation inequality, and Lemma 7.2 now give

$$\eta(\pi_\alpha) - \eta(\pi) = \frac{\alpha}{1-\gamma}\mathbb{E}_{d_{\pi_\alpha}} g \geq \frac{\alpha}{1-\gamma}\left[\mathbb{E}_{d_\pi} g - \frac{\gamma\alpha\Delta_g}{1-\gamma+\gamma\alpha}\right] \geq \frac{\alpha}{1-\gamma}\left[M_d - \frac{\gamma\alpha\Delta_g}{1-\gamma+\gamma\alpha}\right].$$

This proves Eqs. (69) and (70). The exact signed choice meets Eq. (67) with equality, while $\mathbb{E}b \leq \mathbb{E}|b| \leq \varepsilon_{A,\mathrm{cand}}$ proves the other two specializations. For

$$\varphi(\alpha) := \frac{\gamma\alpha\Delta_g}{1-\gamma+\gamma\alpha}, \qquad \varphi'(\alpha) = \frac{\gamma\Delta_g(1-\gamma)}{(1-\gamma+\gamma\alpha)^2} \geq 0.$$

Thus $\varphi(0) = 0$ and $\max_{\alpha\in[0,1]}\varphi(\alpha) = \varphi(1) = \gamma\Delta_g$. If $\gamma > 0$ and $0 < M_d < \gamma\Delta_g$, then $M_d < \Delta_g$ because $\gamma < 1$, so $\Delta_g - M_d > 0$ and the threshold is positive. It is also below one:

$$\frac{M_d(1-\gamma)}{\gamma(\Delta_g - M_d)} < 1 \quad\Longleftrightarrow\quad M_d < \gamma\Delta_g,$$

which is the branch premise. Rearrangement gives Eq. (71); monotonicity and direct substitution give the listed boundary cases. □

C.4.10 Persistent-state evaluation and improvement

Proof of Corollary C.2. The clock-complete state, frozen shared map F_n, and specified initial law define one augmented MDP for both action kernels. On its bounded measurable functions, $\overline{\mathcal{T}}_K^{\bar\pi}$ is a γ^K-contraction in the augmented closure norm. In this proof, $\|\cdot\|_\infty^{\mathrm{adv}}$ abbreviates $\|\cdot\|_{\infty,\bar\pi,\bar\pi_{\mathrm{cand}}}^{\mathrm{adv}}$. Thus

$$\begin{aligned}\|\widetilde{U}_n - V^{\bar\pi}\|_\infty^{\mathrm{adv}} &\leq \|\widetilde{U}_n - \overline{\mathcal{T}}_K^{\bar\pi}\widetilde{U}_n\|_\infty^{\mathrm{adv}} + \|\overline{\mathcal{T}}_K^{\bar\pi}\widetilde{U}_n - \overline{\mathcal{T}}_K^{\bar\pi}V^{\bar\pi}\|_\infty^{\mathrm{adv}} \\ &\leq \bar\varepsilon_{\mathrm{res},n} + \gamma^K\|\widetilde{U}_n - V^{\bar\pi}\|_\infty^{\mathrm{adv}}.\end{aligned}$$

Rearrangement proves (40). Lemma C.9, applied on the augmented closure with exact centering conditional on $\bar{s}$, gives (41). Corollary 5.3 supplies the $\bar\tau_\infty$ refinement. Under the exact pointwise mixture, substitution into Theorem 7.3 proves (45).

For the target-network and finite-reference statements, apply Corollary C.3 on the same augmented MDP with its target-depth, bounded-endpoint, common-boundary, frozen-policy trajectory, and exact full-state baseline hypotheses. This gives every barred substitution while keeping F_n and the deployed kernel fixed. □

C.4.11 Weighted-L_p bound

Proof of Proposition D.4. Let $e_m := U_m - V^\pi$. Subtracting the fixed-point identity $V^\pi = \mathcal{T}_K^\pi V^\pi$ from the definition of r_m gives

$$e_m = r_m + \beta P e_m. \tag{74}$$

Because e_m and r_m are bounded and $\beta < 1$, iteration of this identity and uniform decay of the remainder yield

$$e_m = \sum_{t=0}^{\infty}\beta^t P^t r_m, \qquad |e_m| \leq \sum_{t=0}^{\infty}\beta^t P^t|r_m|. \tag{75}$$

Define the Markov kernel $Q := (1-\beta)\sum_{t\geq 0}\beta^t P^t$. Then $\rho Q = d_\rho^{\pi,K}$ and $|e_m| \leq Q|r_m|/(1-\beta)$. Jensen's inequality for the Markov kernel Q gives

$$\begin{aligned}\|e_m\|_{p,\rho}^p &\leq \frac{1}{(1-\beta)^p}\int_{\mathcal{C}} Q|r_m|^p\, d\rho \\ &= \frac{1}{(1-\beta)^p}\int_{\mathcal{C}} |r_m|^p\, dd_\rho^{\pi,K} \\ &\leq \frac{C_{\rho,\mu}^{\pi,K}}{(1-\beta)^p}\|r_m\|_{p,\mu}^p. \end{aligned} \tag{76}$$

Taking pth roots and applying the triangle inequality proves Eq. (88). Under Eq. (89), the same change-of-measure calculation gives $\|U_n - U_m\|_{p,\rho} \leq (C_{\rho,\mu}^{(0)})^{1/p}\|U_n - U_m\|_{p,\mu}$, which proves Eq. (90). □

C.4.12 Deployment perturbation

Proof of Theorem C.4. Let $\mu_t^\alpha, \widetilde{\mu}_t$ be the state laws and set $d_t = \mathrm{TV}(\widetilde{\mu}_t, \mu_t^\alpha)$. Their common sub-measure $\lambda_t = \widetilde{\mu}_t \wedge \mu_t^\alpha$ has mass $1 - d_t$; each remaining measure has mass d_t. Appending the common reward-transition kernel K to the two action laws at a state cannot increase total variation. Their joint action-reward-successor laws therefore differ by at most δ_{dep}.

Write $J_t^\alpha, \widetilde{J}_t$ for the full joint laws of state, action, reward, and successor. Integrating the statewise bound over λ_t and bounding the residual components by their mass gives

$$\mathrm{TV}(\widetilde{J}_t, J_t^\alpha) \leq d_t + (1-d_t)\delta_{\mathrm{dep}}.$$

Marginalization gives the same bound for d_{t+1}. Since $d_0 = 0$, induction yields $d_t \leq 1-(1-\delta_{\mathrm{dep}})^t$ and

$$\mathrm{TV}(\widetilde{J}_t, J_t^\alpha) \leq 1-(1-\delta_{\mathrm{dep}})^{t+1}.$$

The exponent is $t+1$ because the reward at time t follows decision t. Both rewards lie in an interval of width Δ_r, so

$$|\mathbb{E}[\widetilde{r}_t] - \mathbb{E}[r_t^\alpha]| \leq \Delta_r\,\mathrm{TV}(\widetilde{J}_t, J_t^\alpha) \leq \Delta_r\left[1-(1-\delta_{\mathrm{dep}})^{t+1}\right].$$

Sum these reward differences for $t = 0, \ldots, H-1$ and let $H \to \infty$. Bounded rewards justify the limit, and

$$\begin{aligned}\sum_{t=0}^{\infty}\gamma^t\left[1-(1-\delta_{\mathrm{dep}})^{t+1}\right] &= \frac{1}{1-\gamma} - \frac{1-\delta_{\mathrm{dep}}}{1-\gamma(1-\delta_{\mathrm{dep}})} \\ &= \frac{\delta_{\mathrm{dep}}}{(1-\gamma)(1-\gamma+\gamma\delta_{\mathrm{dep}})}.\end{aligned}$$ □

C.4.13 Uniform-KL deployment bound

Proof of Corollary C.6. Pinsker's inequality in either direction and $\mathrm{TV} \leq 1$ show that the actual uniform TV discrepancy is at most $\bar{\delta}_\kappa$. Apply Theorem C.4 with $\delta_{\mathrm{dep}} = \bar{\delta}_\kappa$; the right-hand side is increasing in this radius. If the selected KL divergence is infinite, this corollary does not apply; the TV theorem remains available when Eq. (55) is established directly. □

C.4.14 CPI with deployment discrepancy

Proof of Corollary C.7. Lower-bound $\eta(\widetilde{\pi})$ by $\eta(\pi_\alpha)$ minus Eq. (56), then apply Eq. (25) with $\Xi_\alpha \leq B_\alpha$. □

C.4.15 Residual-to-CPI substitutions

Proof of Corollary C.3. Apply Eq. (3) to $V = U_n$ to obtain Eq. (47). Lemma C.9 yields (48), and Corollary 5.3 sharpens the candidate bias to $2\gamma\tau_\infty\varepsilon_{\mathrm{res},n}/(1-\gamma^K)$. Substitution into Theorem 7.3 gives Eq. (49); $\tau_\infty \leq 1$ recovers the uniform fallback.

For the target-network forms, combine Proposition 6.1 with Eq. (3) to obtain Eq. (50). The triangle inequality through U_m gives Eq. (51). Corollary 5.3 applied at depth q proves Eq. (52); substitution into Theorem 7.3 gives Eq. (53). Applying the same overlap bound to the depth-n estimator and the radius in Eq. (51) gives Eq. (54). □

D Optional contraction and weighted norms

D.1 Recurrent-contraction specialization

Assumption D.1 (Domain-relative forward-invariant contraction)**.** Fix a declared nonabsorbing state domain $\mathcal{D}^\circ$. The latent space $\mathcal{Z}$ is a complete normed space (in our implementation, $\mathbb{R}^d$ with the Euclidean norm). For $s \in \mathcal{D}^\circ$, set

$$T_s(z) := \widetilde{f}_\theta(z, y(s), x(s)).$$

There exist a closed set $\mathcal{Z}_{\mathrm{inv}} \subseteq \mathcal{Z}$ and $L_z \in [0, 1)$ such that, uniformly for every $s \in \mathcal{D}^\circ$,

$$T_s(\mathcal{Z}_{\mathrm{inv}}) \subseteq \mathcal{Z}_{\mathrm{inv}}, \qquad \|T_s(z) - T_s(z')\| \le L_z \|z - z'\| \quad (z, z' \in \mathcal{Z}_{\mathrm{inv}}).$$

Every initialization used by the result satisfies $z_{\mathrm{init}}(x(s), y(s)) \in \mathcal{Z}_{\mathrm{inv}}$ for every $s \in \mathcal{D}^\circ$.

For the specialization below use $\mathcal{D}^\circ = \mathcal{C}^\circ$ and finite C_z from Eq. (34). We use $L_z^0 = 1$ for every $L_z \in [0, 1)$ and $L_z^n = 0$ for $n \ge 1$ when $L_z = 0$.

Proposition D.2 (Domain-relative fixed point and finite-unrolling bias (Banach))**.** *Under Assumption D.1, for every $s \in \mathcal{D}^\circ$ the map T_s has a unique fixed point $z^*(s) \in \mathcal{Z}_{\mathrm{inv}}$. For every $z^{(0)}(s) \in \mathcal{Z}_{\mathrm{inv}}$, its iterates converge linearly:*

$$\left\| z^{(t)}(s) - z^*(s) \right\| \;\le\; L_z^t \left\| z^{(0)}(s) - z^*(s) \right\|. \tag{77}$$

If $\mathcal{D}^\circ = \mathcal{C}^\circ$, C_z is finite as in Eq. (34)*, and $z^{(0)}(s) = z_{\mathrm{init}}(x(s), y(s))$, then uniformly on $\mathcal{C}^\circ$,*

$$\left\| z^{(n)}(s) - z^*(s) \right\| \;\le\; \frac{L_z^n}{1 - L_z}\, C_z. \tag{78}$$

Fixed-point existence is asserted only for input pairs represented by the declared domain. Assumption D.1's metric contraction and invariance clauses alone do not imply measurability of $s \mapsto z^*(s)$.

For a sufficient measurability condition, equip the closed set $\mathcal{Z}_{\mathrm{inv}}$ with its Borel σ-algebra and $\mathcal{D}^\circ$ with its declared σ-algebra. If

$$T : \mathcal{D}^\circ \times \mathcal{Z}_{\mathrm{inv}} \longrightarrow \mathcal{Z}_{\mathrm{inv}}, \qquad T(s, z) = T_s(z), \tag{79}$$

is jointly measurable and $z^{(0)} : \mathcal{D}^\circ \to \mathcal{Z}_{\mathrm{inv}}$ is measurable, then every iterate $z^{(j)}$ and the fixed-point map z^* are measurable.

For any result below that invokes L_V on a declared nonabsorbing domain $\mathcal{D}^\circ$, let $\mathcal{Z}_{\mathcal{D}}(s)$ be the set of latent arguments compared by that result at s. Assume the scalar head is uniformly Lipschitz while the instance and plan are held fixed: for every $s \in \mathcal{D}^\circ$ and all $z, z' \in \mathcal{Z}_{\mathcal{D}}(s)$,

$$\left| V_\psi(z, x(s), y(s)) - V_\psi(z', x(s), y(s)) \right| \;\le\; L_V \, \|z - z'\|. \tag{80}$$

For a contractive specialization one may take $\mathcal{Z}_{\mathcal{D}}(s) = \mathcal{Z}_{\mathrm{inv}}$; a pathwise result needs the condition only at the latent pairs it displays.

Assumption D.3 (Domain-explicit bounded value heads)**.** For an invocation on a declared nonabsorbing state domain $\mathcal{D}^\circ$, declare a latent set $\mathcal{Z}_{\mathrm{bd}} \subseteq \mathcal{Z}_{\mathrm{inv}}$. It is either all of $\mathcal{Z}_{\mathrm{inv}}$ or an explicitly identified smaller set containing every latent used by that result. There exists $V_{\max} < \infty$ such that, for every represented $s \in \mathcal{D}^\circ$ and every $z \in \mathcal{Z}_{\mathrm{bd}}$,

$$|V_\psi(z, x(s), y(s))| \le V_{\max}, \qquad |V_{\bar{\psi}}(z, x(s), y(s))| \le V_{\max}. \tag{81}$$

The absorber is handled by its separately declared finite scalar boundary; no absorbing latent is introduced. This assumption can supply boundedness in the optional contractive specializations, but it does not replace any explicit $B_b(\mathcal{C})$ premise in a Bellman theorem.

For the optional contractive specialization, invoke Assumption D.1 with $\mathcal{D}^\circ = \mathcal{C}^\circ$ and the finite constant C_z above. For each represented $s \in \mathcal{C}^\circ$, write $z^*(s)$ for the unique fixed point of T_s and $U_*(s) = V_\psi(z^*(s), x(s), y(s))$ on $\mathcal{C}^\circ$. The joint measurability and measurable initialization make $s \mapsto z^*(s)$ measurable by Proposition D.2. Invoke Assumption D.3 on $\mathcal{D}^\circ = \mathcal{C}^\circ$ with $\mathcal{Z}_{\mathrm{bd}} = \mathcal{Z}_{\mathrm{inv}}$; together with the separate finite absorbing boundary, this makes U_* bounded on the closure. Set $U_*(s_{\mathrm{abs}}) = -C_{\max}$ and define

$$\varepsilon^*_{\mathrm{res}} := \|U_* - \mathcal{T}^\pi_K U_*\|^{\mathrm{adv}}_{\infty,\pi,\pi_{\mathrm{cand}}}. \tag{82}$$

Consecutive latent differences satisfy $\|z^{(j+1)} - z^{(j)}\| \le L_z^j C_z$ uniformly. Hence Theorem 5.1 gives

$$\|U_n - V^\pi\|^{\mathrm{adv}}_{\infty,\pi,\pi_{\mathrm{cand}}} \le L_V C_z \frac{L_z^n(1 - L_z^{m-n})}{1 - L_z} + \frac{\|U_m - \mathcal{T}^\pi_K U_m\|^{\mathrm{adv}}_{\infty,\pi,\pi_{\mathrm{cand}}}}{1 - \gamma^K}. \tag{83}$$

Uniform convergence $U_m \to U_*$ and Bellman contraction imply convergence of the residual norm. Taking $m \to \infty$ recovers

$$\|U_n - V^\pi\|^{\mathrm{adv}}_{\infty,\pi,\pi_{\mathrm{cand}}} \le \underbrace{\frac{\varepsilon^*_{\mathrm{res}}}{1 - \gamma^K}}_{\text{fixed-point residual}} + \underbrace{L_V \frac{L_z^n}{1 - L_z} C_z}_{\text{geometric path bound}}. \tag{84}$$

Proofs of Proposition D.2 and Eqs. (83) and (84) appear in Appendix D.3.

D.2 A WEIGHTED-L_p FIXED-POLICY RESTATEMENT

A weighted norm can replace the population supremum when a reference measure covers the fixed policy's discounted occupancy. The following result states the corresponding change-of-measure cost. For a probability measure ν and $1 \le p < \infty$, write $\|f\|_{p,\nu} := (\int_{\mathcal{C}} |f|^p \, d\nu)^{1/p}$.

Proposition D.4 (Weighted-L_p finite-reference bound)**.** *Under the measurable fixed-policy hypotheses of Theorem 5.1, let ρ and μ be probability measures on $\mathcal{C}$, let $1 \le p < \infty$, and take bounded measurable endpoints U_n, U_m. Set*

$$\beta := \gamma^K, \qquad P := P^K_\pi, \qquad r_m := U_m - \mathcal{T}^\pi_K U_m, \tag{85}$$

and define the discounted occupancy of the K-step skeleton by

$$d^{\pi,K}_\rho := (1 - \beta) \sum_{t=0}^{\infty} \beta^t \rho P^t. \tag{86}$$

This is a probability measure on $\mathcal{C}$. Suppose $d^{\pi,K}_\rho \ll \mu$ and

$$C^{\pi,K}_{\rho,\mu} := \left\| \frac{d d^{\pi,K}_\rho}{d\mu} \right\|_{\infty,\mu} < \infty, \tag{87}$$

where the norm is the μ-essential supremum. Then

$$\|U_n - V^\pi\|_{p,\rho} \le \|U_n - U_m\|_{p,\rho} + \frac{(C^{\pi,K}_{\rho,\mu})^{1/p}}{1 - \gamma^K} \|U_m - \mathcal{T}^\pi_K U_m\|_{p,\mu}. \tag{88}$$

If also $\rho \ll \mu$ and

$$C^{(0)}_{\rho,\mu} := \left\| \frac{d\rho}{d\mu} \right\|_{\infty,\mu} < \infty, \tag{89}$$

then the endpoint discrepancy may instead be measured under μ:

$$\|U_n - V^\pi\|_{p,\rho} \le (C^{(0)}_{\rho,\mu})^{1/p} \|U_n - U_m\|_{p,\mu} + \frac{(C^{\pi,K}_{\rho,\mu})^{1/p}}{1 - \gamma^K} \|r_m\|_{p,\mu}. \tag{90}$$

Proof in Appendix C.4.11. The coefficient concerns the one fixed kernel P^K_π; the policy-sequence API/FVI comparisons are discussed in Section 2. For persistent deterministic latents, augmented occupancies can lie on graphs of zero mass under a product reference measure, so that measure need not dominate them. Choosing $\mu = d^{\pi,K}_\rho$ makes the coefficient one, but an exogenous training distribution need not have that coverage.

D.3 Optional recurrent contraction

Proof of Proposition D.2. For each $s \in \mathcal{D}^\circ$, Assumption D.1 makes T_s a contraction on the complete space $\mathcal{Z}_{\mathrm{inv}}$. Banach's theorem gives its unique fixed point and Eq. (77).

Under the additional measurability premises in Eq. (79), the iterates are measurable by induction. Indeed, if $z^{(j)}$ is measurable, then $s \mapsto (s, z^{(j)}(s))$ is measurable into the product space, so $z^{(j+1)}(s) = T(s, z^{(j)}(s))$ is measurable. Banach convergence gives $z^{(j)}(s) \to z^*(s)$ pointwise. For every nonempty closed $F \subseteq \mathcal{Z}_{\mathrm{inv}}$, continuity of distance and pointwise convergence give

$$d(z^*(s), F) = \lim_{j\to\infty} d(z^{(j)}(s), F).$$

The right side is a pointwise limit of measurable real functions, and

$$\{s : z^*(s) \in F\} = \{s : d(z^*(s), F) = 0\}.$$

The empty closed set has empty preimage. Since closed sets generate the Borel σ-algebra, z^* is measurable.

For the finite-unrolling bound, specialize to $\mathcal{D}^\circ = \mathcal{C}^\circ$ and take $z^{(0)}(s) = z_{\mathrm{init}}(x(s), y(s))$:

$$z^*(s) - z^{(n)}(s) = \sum_{k=n}^{\infty} (z^{(k+1)}(s) - z^{(k)}(s)).$$

Contraction gives $\|z^{(k+1)}(s) - z^{(k)}(s)\| \le L_z^k \|z^{(1)}(s) - z^{(0)}(s)\|$, so

$$\|z^*(s) - z^{(n)}(s)\| \le \|z^{(1)}(s) - z^{(0)}(s)\| \sum_{k=n}^{\infty} L_z^k = \frac{L_z^n}{1 - L_z} \|z^{(1)}(s) - z^{(0)}(s)\|.$$

Equation (34) bounds the first increment by C_z, giving Eq. (78). □

Proof of Eqs. (83) *and* (84). Under the $\mathcal{D}^\circ = \mathcal{C}^\circ$ invocation of Assumption D.1,

$$\|z^{(j+1)}(s) - z^{(j)}(s)\| \le L_z^j \|z^{(1)}(s) - z^{(0)}(s)\| \le L_z^j C_z.$$

Summing from $j = n$ through $m - 1$ yields (83). Proposition D.2 and the Lipschitz head give $\|U_m - U_*\|_\infty \to 0$, and hence

$$\left| \|U_m - \mathcal{T}_K^\pi U_m\|_\infty - \|U_* - \mathcal{T}_K^\pi U_*\|_\infty \right| \le (1 + \gamma^K) \|U_m - U_*\|_\infty \to 0.$$

Taking $m \to \infty$ therefore proves (84). □

E Finite-census learned-evaluator diagnostic

Let $\mathcal{S}_B$ be the fixed census of 128 full clock-complete augmented states (bars and tildes suppressed). For seed j and $q \in \{n, m\}$, let $U_{q,j}$ be the restored endpoint, r_f the folded reward, and $\mathbf{1}_{\mathrm{nt}}$ the nonterminal indicator. The one-step self-bootstrap exactly enumerates every valid action and its deterministic transition under the frozen base policy $\bar{\pi}$ and gives

$$D_j := \max_{s \in \mathcal{S}_B} |U_{n,j}(s) - U_{m,j}(s)|,$$

$$\varepsilon_{q,j} := \max_{s \in \mathcal{S}_B} \left| U_{q,j}(s) - \sum_{a \in \mathcal{A}(s)} \bar{\pi}(a \mid s) \left[r_f(s,a) + \gamma \mathbf{1}_{\mathrm{nt}} U_{q,j}(s') \right] \right|.$$

The folded reward r_f includes the absorbing tail; terminal transitions receive no bootstrap. Both endpoints share the depth-n base law, environment transition, and persistent successor latent from the depth-n deployed mixture. Only endpoint evaluation uses m. The separate maxima may have different witnesses; their sum is not a maximum of recordwise sums. These census quantities do not uniformly bound a theorem domain.

The diagnostic evaluates the finite-census quantity

$$G_j = \frac{\varepsilon_{n,j}}{1 - \gamma} - \left(D_j + \frac{\varepsilon_{m,j}}{1 - \gamma} \right) \tag{91}$$

for eight training seeds, with $n = 2$, $m = 8$, $K = 1$, $\gamma = 0.99$, $\alpha = 0.1$, and 128 fixed validation-bridge records. The task is Sudoku plan editing with checker scale $C_{\max} = 16$, a budget of 16 edits, and shaped rewards; each training run stops after 10,000 environment interactions. Table 4 reports the signed proxy-radius difference and two centering diagnostics. The constructed-centering roundoff ranges from $1.3344159524586459 \times 10^{-16}$ to $4.903119502940478 \times 10^{-16}$.

Table 4: Finite-census diagnostics in the listed seed order. The last two columns are maximum absolute errors over the 128-state census.

Seed	G_j	Trainer centering defect	Parity error
2081976412	21.880301018232558	3.156856×10^{-7}	1.277353×10^{-6}
781025396	12.585726740108612	1.428112×10^{-7}	1.385260×10^{-6}
1148619853	10.658699265691894	7.030317×10^{-8}	1.328253×10^{-6}
913535362	13.393999128162932	3.000817×10^{-7}	1.413822×10^{-6}
2143253519	17.061289531923650	4.780728×10^{-8}	1.507580×10^{-6}
2279379379	11.247952346485135	1.714248×10^{-7}	1.468691×10^{-6}
402312153	13.701769789184710	2.211111×10^{-7}	1.328038×10^{-6}
2736405725	13.459892571054297	1.098139×10^{-7}	1.375904×10^{-6}

Table 5 reports the components of G_j from the same result record. The direct proxy ranges from 1096.60 to 1179.66 and the finite-reference proxy from 1083.21 to 1157.78, so G_j is 0.94% to 1.85% of the direct proxy. In every seed one census state attains both residual maxima. With residuals near 11 against $C_{\max} = 16$ and $1/(1-\gamma) = 100$, both proxies are about 70 times $C_{\max}$: at these snapshots neither is an informative value radius, and the diagnostic speaks only to the sign of G_j.

Table 5: Components of G_j over the 128-state census, rounded. The direct proxy is $\varepsilon_{n,j}/(1-\gamma)$ and the finite-reference proxy is $D_j + \varepsilon_{m,j}/(1-\gamma)$; their difference is G_j in Table 4.

Seed	D_j	$\varepsilon_{n,j}$	$\varepsilon_{m,j}$	Direct proxy	Finite-ref. proxy	G_j/direct (%)
2081976412	0.6770	11.7966	11.5710	1179.66	1157.78	1.85
781025396	0.4689	11.1513	11.0207	1115.13	1102.54	1.13
1148619853	0.3171	11.2987	11.1890	1129.87	1119.21	0.94
913535362	0.5873	10.9660	10.8262	1096.60	1083.21	1.22
2143253519	0.4760	11.4423	11.2669	1144.23	1127.17	1.49
2279379379	0.4479	11.2946	11.1777	1129.46	1118.21	1.00
402312153	0.5830	11.1187	10.9758	1111.87	1098.17	1.23
2736405725	0.6017	11.2345	11.0939	1123.45	1109.99	1.20

The equal-weight mean of G_j is 14.248703798855473. Hyndman-Fan type-7 quantiles at 0.025 and 0.975 of 10,000 stored seed-resampling replicate means give the nominal interval $[12.185185177643433, 16.739775084598428]$. Seeds, not census records, are resampled. The interval is conditional on the single recorded base initialization and fixed census; it establishes neither finite-sample coverage nor independence across datasets or base-policy preparations.

The revised centering-consistency check retains the inherited absolute tolerance 10^{-6} and adds four float32 epsilons scaled by the advantage clip of 10:

$$10^{-6} + 4(2^{-23})(10) = 5.76837158203125 \times 10^{-6}.$$

All eight parity errors exceed the old tolerance and fall below the revised threshold. A one-epsilon allowance, $2.1920928955078122 \times 10^{-6}$, would accept the largest error with 1.45x headroom. Four epsilons are unnecessary for these observations and have no derived operation-level guarantee.

Earlier aggregate diagnostic output was inspected before the revised tolerance was fixed for the reported measurements. Its canonical digest is recorded in the audit documentation. The outcome no-inspection declaration concerns those measurements; the validation/test declarations concern access to examples for selection or tuning, not the earlier aggregate output.

The audit recovered eight training manifests, verified checkpoint bytes, and matched the training executable's source to its recorded revision. An identified read-only executable replayed every census state. A separate consumer made 108,648 exact deterministic comparisons and reproduced all

eight gaps, their mean, and the interval bitwise from exported values, without reevaluating every nonterminal action value. These checks reproduce measurements from existing checkpoints; they neither independently replicate training nor prove historical execution.

The audit separates training manifests from measurement payloads and records build identities, export digests and sizes; reproduction requires the external per-state exports. A repository-independent standard-library checker recomputes all 35 digest-chain links, eight gaps, mean, stored-replicate interval, and numerical parity bound. It passes under `python3 -O` and rejects a recorded one-byte result mutation.

F Implementation and original-TRM mapping

F.1 Collection, targets, and replay

Algorithm 1 collects episodes under one frozen base policy and recurrent map. Value fitting changes the policy-independent value head; candidate fitting changes one policy head. In episodic mode the evaluator reinitializes z at each edit; persistent updates consume the stored pre-unroll latent. A target-head update uses $\bar{\psi}^+ = \omega\bar{\psi} + (1-\omega)\psi^+$, where $\omega \in [0,1]$ retains the old target. For a collected segment of length $\ell \le K$, let d_{seg} indicate that its final transition terminates. The folded-reward implementation target is

$$G_{\mathrm{fold}} = \sum_{k=0}^{\ell-1} \gamma^k r_k + \mathbf{1}\{\neg d_{\mathrm{seg}}\}\gamma^\ell \bar{V}(\bar{s}_\ell). \tag{92}$$

Here $d_{\mathrm{seg}} = 0$ implies $\ell = K$; early nonterminal segments are not admissible exact-K targets. On a terminal segment, the stored final reward is the formal terminal-transition reward plus γb_{abs}, where $b_{\mathrm{abs}} = V^\pi(s_{\mathrm{abs}})$. The zero padding and zero terminal bootstrap in Eq. (92) are therefore an accounting representation of the formal absorbing tail, not an additional self-loop.

Under frozen-base collection the conditional population target is $r_K^\pi + \gamma^K P_\pi^K \bar{V}$, or its full augmented-state counterpart. Proposition 6.1 accounts for propagated target lag.

Replay records the carried input and successor latents, the remaining-edit clock, the terminal cause and mask, and the behavior log probability. Terminal records carry the folded absorbing tail in the reward and are never bootstrapped. Only nonterminal transitions carry the post-unroll latent. Exact baseline evaluation uses Eq. (2) conditional on the full stored state.

Replay means, finite-population maxima, and local finite-difference estimates do not establish the population residual, target-lag, centering, or deployment bounds assumed by the theorems.

F.2 Mapping to the original TRM

The edit-MDP formulation differs from the original TRM as follows:

1. **Training signal**: The original TRM is trained with a supervised loss; the analyzed UPI-TRM object uses checker-based rewards to construct one fixed-base candidate proposal and exact mixture.
2. **Episodic-z vs. persistent-z**: We analyze both; the original TRM is persistent-z.
3. **Halting**: The original TRM halts adaptively across its outer improvement steps; in the edit MDP that role belongs to the STOP action and the remaining budget h. The inner depth n is fixed in both settings and is treated here as an analysis parameter.