POLICY IMPROVEMENT AND PRESERVATION IN CONTINUAL REINFORCEMENT LEARNING WITH LATENT FAMILIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Continual reinforcement learning requires agents to improve in recurring environments while preserving performance on environments that are currently unavailable. Sharing policies enables reuse, but updates for one environment can harm others. We introduce a continuous-state framework that separates identification of latent environment families from the decision to share a policy. Learned Bellman comparisons identify families, while one-sided certificates use finite stored data to bound loss relative to retained reference policies, called anchors. When a shared update cannot be certified, policy splitting allows learning to continue. Under stated separation, sampling, coverage, and local neural-geometry assumptions, we establish high-probability guarantees for correct identification, anchor-relative preservation, and final ε-optimality, using at most one deployed policy module per family. The complete observation bound retains coverage, conditioning, representation-complexity, and acquisition factors. Under the stronger specialization that these factors remain uniformly controlled, the prescribed exploration schedule gives sufficient observation complexity $\widetilde{\mathcal{O}}\big(K\varepsilon^{-2} + K(h-1)\varepsilon^{-3} + (S+K)\Delta_0^{-2}\big)$, where K is the number of families, h is module capacity, S counts segments, and Δ_0 is a diagnostic family separation lower bound. Under additional compatibility and critic-accuracy conditions, all families can remain in one deployed module without splitting; matched constructions show when separate policies are necessary. This module reduction does not imply lower observation or total-memory cost. In a recurrent experiment with learned neural evaluators and Bellman regressions, the method matches independent policies' final performance using two modules instead of three and substantially reduces forgetting relative to ungated sharing.

1 INTRODUCTION

Continual reinforcement learning (RL) requires improving behavior in recurring environments while retaining capabilities needed when earlier environments return (Ring, 1994; Thrun and Mitchell, 1995; Abel et al., 2023). Maintaining an independent policy for each recognized environment can avoid cross-family interference, but distinct environments may also benefit from a common policy. Sharing reduces the number of deployed policies, while creating a tension: an update that helps the active environment may harm an inactive one, whereas rejecting the update may prevent further learning. Neural continual-learning methods protect previous behavior through frozen components, consolidation, and replay (Rusu et al., 2016; Schwarz et al., 2018; Rolnick et al., 2019), while theory provides identification, transfer, and lifelong-learning guarantees (Brunskill and Li, 2013; Nguyen and Mehta, 2023; Lecarpentier et al., 2021; Amani et al., 2022). We study the setting in which several latent families may share one evolving policy, only the active family supplies new observations, and inactive families must be protected using stored data. We ask three linked questions: *Which environment has returned? Can its policy improve while protecting inactive environments? When can distinct environments continue sharing one policy?*

Prior work addresses these questions through different structural simplifications. For identification, finite-state analyses use model separation and reachability (Brunskill and Li, 2013; Nguyen and Mehta, 2023), while COMRLI treats continuous states using Lipschitz value structure, a task gap, and informative-state reachability (Liu et al., 2016). For protected improvement, SPIBB uses a known

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.

data-collecting baseline in a finite-MDP batch setting (Laroche et al., 2019), while safe lifelong policy search imposes shared latent structure and linear parameter constraints (Ammar et al., 2015). For sharing, Lipschitz lifelong RL transfers value information between tabular MDPs (Lecarpentier et al., 2021), and UCBlvd uses observed contexts, common dynamics, and linear representations with completeness (Amani et al., 2022). Appendix B gives a broader comparison.

What our work adds. We separate identification of latent families from the decision to share a policy. Learned Bellman comparisons identify returning families, while one-sided finite-data certificates protect named inactive-family anchors during shared policy updates. When no sufficiently large shared update can be certified, exact-copy splitting isolates the active family and permits learning to continue. We retain finite actions and observed segment boundaries; the guarantees additionally require the separation, coverage, sampling, local neural-geometry, and active-time conditions stated in Section 4.

Contributions. 1. Joint identification, preservation, and progress. Under the stated assumptions, we prove correct identification, bounded anchor loss, and final ε-optimality with at most K deployed modules and $K-1$ splits (Theorem 1). Comparing dormant changes directly with fixed named anchors avoids accumulating a separate preservation loss for every intervening update. A variable-step transport argument similarly controls accumulated critic error, while exact-copy splitting guarantees progress.

2. Compatible sharing and necessary separation. Under compatible-improvement conditions and tighter critic calibration, all families can remain in one deployed module with full shared steps (Theorem 3). Matched continuous-state constructions also show when separate policies are necessary. This result concerns the number of deployed policy modules; it does not imply lower observation or total-memory cost.

3. Finite-training guarantees for learned certificates. We analyze projected TD learning for evaluator hidden weights with fixed readout signs and Bellman regression training both hidden and readout parameters. Optimization, confidence amplification, concentration, and controlled replay disclosure yield adaptive finite-data certificates. The complete bounds retain the actual representation, conditioning, covering-complexity, and acquisition factors; the simplified rates therefore require these quantities to remain controlled rather than following from network width alone.

For the general accuracy guarantee, Theorem 1 gives the complete observation ledger with these factors explicit. Under the stronger specialization that they remain uniformly bounded or polylogarithmic in the target accuracies, an exploratory logging floor $p_{\log} = \Theta(\varepsilon)$ gives

$$\widetilde{\mathcal{O}}\big(K\varepsilon^{-2} + K(h-1)\varepsilon^{-3} + (S+K)\Delta_0^{-2}\big)\,, \tag{1}$$

where K is the number of families, $h \le K$ is the module capacity, S is the number of segments, and Δ_0 is the supplied diagnostic family separation lower bound (Corollary 2). If coverage, conditioning, or representation complexity deteriorates with accuracy, that dependence is retained in the complete ledger.

Empirical evidence. A recurrent learned-control experiment trains critics, references, snapshots, and Bellman images under matched interaction budgets. Across ten paired seeds, the full method matches independent policies' final performance using two deployed modules instead of three and substantially reduces forgetting relative to ungated sharing. A supporting CartPole study examines adaptation with shared policies on raw continuous states (Section 6).

2 Problem Setup

There are K fixed MDP families $\mathcal{M}_k = (\mathcal{S}, \mathcal{A}, P_k, r_k, \gamma)$, $k \in \{1, \dots, K\}$, on a common standard Borel state space and finite action set. Rewards are deterministic, $|r_k(s,a)| \le R_{\max}$, and $0 < \gamma < 1$; write $c = 1-\gamma$ and $Q_{\max} = R_{\max}/c$. The data stream contains S segments, each a contiguous occurrence of one family between observed boundaries. Family identities are hidden, only the active family supplies new transitions, and its reward and transition rules remain fixed on recurrence. The learner receives a diagnostic separation lower bound $\Delta_0 > 0$; Section 4 states its required property.

For a stationary policy π, let V_k^π and Q_k^π be its discounted state and action values. The reporting distribution ξ_k is a probability law on $\mathcal{S}$, and $\mathrm{Ret}_k(\pi) = \xi_k V_k^\pi$, where $\zeta f = \int f \, d\zeta$. For a target

accuracy $\varepsilon > 0$, we seek $\mathrm{Ret}_k(\pi_k^*) - \mathrm{Ret}_k(\pi_k^{\mathrm{report}}) \le \varepsilon$ for **every family**, where π_k^* is optimal and π_k^{report} is the final policy assigned to that family.

This setting presents three coupled challenges: recognizing returning environments without observing their identities, improving performance in the active environment, and preserving performance in environments that are currently unavailable. When distinct families share a policy, these goals can conflict: an update that benefits the active family may harm others, while rejecting updates to protect earlier performance may prevent further learning. Assessing this tradeoff is difficult because inactive families can be evaluated only through finite stored data. The learner must therefore distinguish environment identity from policy compatibility, determine when shared policy updates can preserve earlier performance, and retain a way to continue learning when sharing cannot be certified. Suppose policy updates use the exploratory class

$$\Pi_\alpha = \{(1-\alpha)\widetilde{\pi} + \alpha\nu_0 : \widetilde{\pi} \text{ stationary, measurable}\}, \qquad \nu_0(a \mid s) \ge \nu_{\min} > 0, \tag{2}$$

where ν_0 is a fixed exploration policy, $\alpha \in (0,1)$, and $p_{\min} = \alpha\nu_{\min}$ is the resulting action floor.

Write P_k^π for the state transition kernel under π and $T_k^\pi V = \sum_a \pi(a \mid \cdot) r_k(\cdot, a) + \gamma P_k^\pi V$ for its state-value Bellman operator. A fixed diagnostic policy π_{id} defines $Q_k^{\mathrm{id}} = Q_k^{\pi_{\mathrm{id}}}$ and $\mathcal{T}_k^{\mathrm{id}} q = r_k + \gamma P_k^{\mathrm{id}} q$, where P_k^{id} draws a successor state from P_k and its action from π_{id}. For any probability law ϱ, let $\|f\|_\varrho = (\int f^2 d\varrho)^{1/2}$; $\rho \otimes \pi$ denotes the joint law obtained by drawing a state from ρ, action from π.

3 Algorithm and notation

At a high level, the algorithm first routes each segment to a returning family record or creates a new one, then proposes an improvement to the policy assigned to that record. Several records may share one policy module, so an update for the active family can also change the policies used by inactive families. The learner therefore uses stored data to test whether the proposed update preserves each affected inactive family relative to its latest anchor. If a sufficiently large shared update can be certified, it is installed in the common module. Otherwise, the active record receives an independent exact copy of the incumbent policy and takes the full update there, allowing learning to continue without changing the other records' policies.

Records, modules, and anchors. The learner maintains a separate record for each identified family, while allowing several records to share one policy module. Each record points to one module, whose policy is used by every record assigned to it; write $g_t(i)$ for record i's module at decision t. Updating a module therefore changes the policy used by all of its records. Record i stores a diagnostic reference q_i estimating the family's Q^{id}, an anchor policy $\bar{\pi}_i$, and a snapshot $\widehat{V}_i$ estimating $V_i^{\bar{\pi}_i}$; the learned value objects are clipped to $[-Q_{\max}, Q_{\max}]$. The anchor is the record's initially assigned policy or its latest accepted own-update policy. An *own slot* is an opportunity for the active record to improve its policy and refresh its anchor and snapshot. These objects remain fixed during other records' updates and thereby provide a stable preservation baseline. There are K record slots and K module slots, with chosen capacity $1 \le h \le K$ records per module and quota J own updates per record. A record that completes its quota receives no further own slots, but remains protected whenever another record updates their shared module.

Learned functions and sampling laws. The algorithm uses *evaluators* to estimate policy values for improvement, diagnostic references, and anchor snapshots, and *Bellman regressors* to estimate one-step Bellman images for routing and preservation. For fixed features with $\|\phi(s,a)\|_2 \le 1$, the width-m evaluator is

$$f_W(s,a) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} b_j \sigma(w_j^\top \phi(s,a)), \qquad b_j \in \{-1, 1\}, \qquad \|W - W^0\|_F \le R. \tag{3}$$

Its hidden weights train within the displayed ball while the readout signs remain fixed; Appendix D.3 gives the evaluator training and confidence conversion. Bellman regressors use the network in (56), defined in Appendix D.4.2, and train both hidden and readout parameters. For regression call r, the selected class is

$$\mathcal{G}_r = \{f_\theta : \|\theta - \theta_r^0\|_2 \le R_r\}.$$

The preservation classes $\mathcal{G}_{i,e}$ and diagnostic classes $\mathcal{G}_\ell^{\mathrm{id}}$ are instances of $\mathcal{G}_r$. Preservation regressors take states as inputs, whereas diagnostic regressors take state–action pairs.

The sampling laws distinguish where a learned object is accurate from where it is used. Preservation scores use the state law μ_i, while snapshot error is measured under υ_i. Diagnostic scores in active family k use the state–action law μ_k^{id}, while reference error for record ℓ is measured under ν_ℓ^{id}. Preservation data are collected while family i is active under a logger β_i with invariant state law μ_i and known action probabilities bounded below by $p_{\log} > 0$. This logging floor is distinct from the deployed-policy exploration floor $p_{\min}$.

Identification and policy improvement. For each stored reference q_ℓ, use a common batch of active-environment transitions to fit $\widehat{g}_\ell$ to its current-family Bellman image $T_k^{\mathrm{id}} q_\ell$; the learner forms these targets without observing the analytical family label k. A matching reference approximately satisfies the active environment's Bellman equation. On a held-out monitor of n_{id} inputs from μ_k^{id}, write $\|f\|_n^2 = n^{-1}\sum_{b=1}^n f(U_b)^2$ and select the lowest-score record satisfying

$$\widehat{R}_\ell = \|q_\ell - \widehat{g}_\ell\|_{n_{\mathrm{id}}} \le \Delta_0/4. \tag{4}$$

If none passes, create a new record; ties are broken by creation order.

For an own update, evaluate the assigned incumbent π^- in the active environment to obtain $\widehat{Q}$. Form the exploratory greedy endpoint and its action-probability mixtures with the incumbent:

$$\pi^{\mathrm{g}} = (1-\alpha)\,\mathrm{Greedy}(\widehat{Q}) + \alpha\nu_0, \qquad \pi^\lambda = (1-\lambda)\pi^- + \lambda\pi^{\mathrm{g}}. \tag{5}$$

Here $\lambda \in [0,1]$ controls the update size: $\lambda = 0$ retains the incumbent and $\lambda = 1$ takes the full endpoint. Greedy ties use a fixed rule. Write $\pi^0 = \pi^-$, $\pi^1 = \pi^{\mathrm{g}}$, and let ϑ denote a tested mixture.

Preservation checks. Before changing a shared policy, the learner checks the proposal against every other record in the module. For protected record i, compare its frozen anchor snapshot $\widehat{V}_i$ with its Bellman update under the candidate policy. The population one-sided score is

$$R_i^+(\lambda) = \|(\widehat{V}_i - T_i^{\pi^\lambda}\widehat{V}_i)_+\|_{\mu_i}, \qquad x_+ = \max\{x, 0\}.$$

The positive part penalizes deterioration but not improvement. Under the snapshot-accuracy and coverage conditions in Section 4, controlling this one-step shortfall bounds return loss relative to the anchor.

To estimate the score, fit $\widehat{g}_i^e$ to $T_i^{\pi^e}\widehat{V}_i$ in $\mathcal{G}_{i,e}$ for $e \in \{0,1\}$ and interpolate

$$\widehat{g}_i^\lambda = (1-\lambda)\widehat{g}_i^0 + \lambda\widehat{g}_i^1.$$

A held-out solve monitor of m_i states from μ_i selects a step satisfying

$$U_i(\lambda) = \|(\widehat{V}_i - \widehat{g}_i^\lambda)_+\|_{m_i} + e_i \le \frac{3\tau_i}{4} \tag{6}$$

for every protected record, where $e_i \le \tau_i/8$ accounts for regression and monitor error. The solve guarantee is uniform over λ, so it also applies to the data-selected candidate $\widehat{\lambda}$. Algorithm 1 retains a fresh held-out audit as an operational check. On $m_{a,i}$ new states from μ_i, compute

$$\widehat{R}_{a,i}^+ = \|(\widehat{V}_i - \widehat{g}_i^{\widehat{\lambda}})_+\|_{m_{a,i}}$$

and require $\widehat{R}_{a,i}^+ + e_{a,i} \le \tau_i$, where $e_{a,i} \le \tau_i/8$. On the simultaneous certificate event used in the theory, every solve-feasible candidate passes this audit; outside that event, an audit failure triggers the conservative fallback, without an additional theorem-level guarantee.

Sharing and splitting. If a candidate passes the solve and audit checks for every protected record, it is installed in the shared module. If no candidate with $\lambda \ge \lambda_{\min}$ passes, a required reserve block is unavailable, or the operational audit fails, the active record is separated from the module. It receives an independent exact copy of the incumbent policy π^- and takes the full update $\lambda = 1$ there. The other records retain their previous policy, while the active family continues learning. The following definition specifies the disclosure and resource rules needed for the adaptive statistical guarantees.

Definition 1 (Certified execution)**.** *Prescribe capacities, quotas, minimum step $\lambda_{\min} \in (0,1]$, tolerances, budgets, and failure allocations. While active, each record collects a reserve containing $(h-1)J$ blocks for future inactive-family checks; unused blocks are charged. Each block contains raw transitions and disjoint solve and audit inputs, and one protected comparison consumes one*

block without replenishment. Freeze policies before evaluator initialization, targets, classes, and budgets before labels, and fitted regressors before monitors. Visible histories contain disclosed data, records, policies, anchors, counters, and frozen choices; schedules cannot inspect hidden reserves. Analytical conditioning may additionally fix the hidden source family without revealing it to the learner. Modules never merge. An independent exact copy preserves the incumbent action law pointwise and isolates every trainable dependency that could later affect it. The procedure stops before exceeding any declared cap, including on failure histories; Appendix D.1.4 gives the complete specification.

Algorithm 1 Routing, certified sharing, and exact-copy splitting

Route each segment by (4). A new record joins the first module with space, or creates one.
For each new record, fit and store its diagnostic reference, evaluate its assigned policy to form the anchor snapshot, and seal its reserve.
for each available own slot of active record ι below quota J **do**
  Freeze and evaluate π^-, form (5), and set $\lambda \leftarrow 1$.
  if there are protected records in the module **then**
    if all required reserve blocks are available **then**
      Disclose fit portions; fit and freeze both endpoint regressions for each protected record.
      Disclose solve monitors; test $\lambda = 1$ first; if it fails, seek $\lambda \in [\lambda_{\min}, 1]$ passing (6) for all protected records.
      If a candidate passes, freeze it; disclose fresh audit inputs and require $\widehat{R}^+_{a,i} + e_{a,i} \le \tau_i$ for all protected i.
    end if
    if a block is unavailable, no candidate passes, or an audit fails **then**
      Give ι an independent exact copy of π^-; reuse its critic and set $\lambda \leftarrow 1$.
    end if
  end if
  Install π^λ; evaluate it, refresh only ι's anchor, and increment its count.
end for

4 ASSUMPTIONS

The conditions below apply to the objects and execution rules of Section 3. Family-indexed kernels and laws refer to the actual data-generating family; indices i, k, ℓ in family comparisons range over $\{1, \dots, K\}$. The learner does not observe these analytical identities.

Assumption 1 (Family separation Bound)**.** *The supplied bound $\Delta_0 > 0$ satisfies*

$$\|Q^{\mathrm{id}}_\ell - \mathcal{T}^{\mathrm{id}}_k Q^{\mathrm{id}}_\ell\|_{\mu^{\mathrm{id}}_k} \ge \Delta_0 \qquad \textit{for every } k \ne \ell. \tag{7}$$

This separates families through their diagnostic Bellman equations (Brunskill and Li, 2013; Nguyen and Mehta, 2023). Boundaries determine when to test; Δ_0 sets the required precision without revealing identities. Any smaller valid lower bound is conservative and increases the routing budget; if the supplied value exceeds the actual separation, the routing guarantee need not hold.

For a pre-call analytical history $\mathcal{H}$, let $\mathsf{L}_{\mathcal{H}}$ be the prescribed law of one transition tuple with its required successor action or regression label, or of one monitor input. A block has N retained observations. For trajectory collection, M is the frozen collector's kernel, ϱ its required invariant input law, and g the number of transitions in each entry, spacing, or exit gap. A measure inequality $\zeta \le C\eta$ means domination on every measurable set; ζP is the law after one kernel step. The supplied envelopes H^{SA}_Q, H_M bound discounted L_2 density-ratio sums comparing relevant policy occupancies with critic and preservation sampling laws, respectively; their definitions are (16)–(17) in Appendix D.1.2.

Assumption 2 (Sampling, coverage, and active time)**.** *For every permitted history, including failure histories:*

(a) Each fresh block has conditional law $\mathsf{L}^{\otimes N}_{\mathcal{H}}$, or uses the correct conditional label law and the frozen collector above with supplied $\kappa_{\mathrm{mix}} < \infty$, $\beta \in (0, 1)$ satisfying

$$\sup_x \|M^n(x, \cdot) - \varrho\|_{\mathrm{TV}} \le \kappa_{\mathrm{mix}} \beta^n \qquad (n \ge 0),$$

where $\|\cdot\|_{\mathrm{TV}}$ is total variation. Targets, classes, and budgets are frozen before labels; fits precede disclosure of disjoint solve/audit monitors. Schedules cannot inspect hidden reserves. Spacing and disclosure follow Appendix D.1.1, Definition 2.

(b) Prescribed finite constants satisfy, for every i, k, ℓ and tested candidate ϑ,

$$\mu_i \le C_{0,i}\upsilon_i, \qquad \mu_i P_i^{\vartheta} \le C_{1,i}\upsilon_i,$$
$$\mu_k^{\mathrm{id}} \le D_{0,k\ell}\nu_\ell^{\mathrm{id}}, \qquad \mu_k^{\mathrm{id}} P_k^{\mathrm{id}} \le D_{1,k\ell}\nu_\ell^{\mathrm{id}}.$$

The discounted density-ratio sums just specified are bounded by $H_Q^{\mathrm{SA}}, H_M < \infty$. These bounds are uniform over permitted histories at the prescribed accuracy and calibration, but may depend on α, ε, and the selected widths, radii, and calls. Put

$$A_V = \sup_i\big(\sqrt{C_{0,i}} + \gamma\sqrt{C_{1,i}}\big), \qquad A_{\mathrm{id}} = \sup_{k,\ell}\big(\sqrt{D_{0,k\ell}} + \gamma\sqrt{D_{1,k\ell}}\big).$$

(c) Every family supplies ordered active intervals for all calibrated calls, its complete $(h-1)J$-block reserve, and J own updates. Each collection of N tuples with gap g fits its full $N + (N+1)g$ transitions before the next required operation (Appendix D.1.4).

These conditions support concentration and return transfer (Antos et al., 2008; Munos and Szepesvári, 2008; Ke et al., 2024); active-time quotas require enough time for the prescribed computations and data collection, beyond recurrence alone. A state-only coverage bound need not keep H_Q^{SA} bounded as the deployed-policy floor $p_{\min}$ decreases, and a fixed logging floor controls replay importance weights but does not by itself control on-policy evaluator coverage or conditioning. The accuracy-uniform control used for the simplified rate is an additional condition in Corollary 2.

For evaluator call v with pre-call history $\mathcal{H}_v$, let π_v, k_v be its frozen policy and true family, ρ_v its stationary state law, $\varrho_v = \rho_v \otimes \pi_v$, R_v its selected radius, and W_v^0 its fresh initialization, independent of data conditional on the history. With parameters vectorized, set

$$\psi_v = \nabla_W f_{W_v^0}, \qquad \Sigma_v = \mathbb{E}_{\varrho_v}[\psi_v\psi_v^\top].$$

An admissible initialization satisfies the input, activation, and initial-output bounds in Appendix D.3.1. The auxiliary classes $\mathcal{G}_{i,e}$ and $\mathcal{G}_\ell^{\mathrm{id}}$ are instances of $\mathcal{G}_r$ from Section 3, selected before label disclosure.

Assumption 3 (Evaluator geometry and Bellman approximation)**.** *For every permitted history and every width and radius selected by calibration:*

(a) Each evaluator replica has an admissible initialization event $\mathcal{E}_v$ with $\mathbb{P}(\mathcal{E}_v \mid \mathcal{H}_v) \ge 15/16$ on which, for supplied $\lambda_{\mathrm{cov},v} > 0$,

$$\exists u_v:\ \|u_v\|_2 \le R_v, \qquad Q_{k_v}^{\pi_v} = f_{W_v^0} + \langle \psi_v, u_v\rangle \quad \varrho_v\text{-a.e.},$$
$$z^\top \Sigma_v z \ge \lambda_{\mathrm{cov},v}\|z\|_2^2 \qquad \forall z \in \mathrm{range}(\Sigma_v).$$

Constants retain their call- and width-dependence.

(b) For every learned clipped snapshot $\widehat{V}_i$, reference q_ℓ, and tested endpoint π^e, $e \in \{0,1\}$, the selected classes satisfy

$$a_{i,e}^{\mathrm{pres}} := \inf_{g\in\mathcal{G}_{i,e}} \|g - T_i^{\pi^e}\widehat{V}_i\|_{\mu_i} \le \frac{\tau_i}{48},$$
$$a_{k\ell}^{\mathrm{id}} := \inf_{g\in\mathcal{G}_\ell^{\mathrm{id}}} \|g - \mathcal{T}_k^{\mathrm{id}} q_\ell\|_{\mu_k^{\mathrm{id}}} \le \frac{\Delta_0}{48} \qquad \forall k, \ell.$$

The diagnostic class is selected without knowing k, including for wrong-family comparisons. These approximation bounds are one third of the prescribed preservation and diagnostic regression root tolerances, $\tau_i/16$ and $\Delta_0/16$. The classes obey the uniform output and derivative bounds in Appendix D.1.3.

Part (a) is a local representation and conditioning premise at the selected initialization, width, and radius. It does not follow from width alone; width-independent conditioning under bounded gradients requires controlled effective tangent rank. Part (b) is an approximate Bellman-completeness condition for the targets reached by the adaptive execution. Its approximation error is absorbed into the calibrated regression radius together with optimization and sampling error, while the complete observation ledger retains the actual covering complexity. Appendix D.9 gives the corresponding budgets.

5 MAIN RESULTS

We first establish identification, preservation, progress, and final accuracy, and then specialize the complete observation ledger under accuracy-uniform regularity. We also characterize when one deployed module suffices and when separate policies are necessary.

Write $\mathrm{gap}_k = \mathrm{Ret}_k(\pi_k^*) - \mathrm{Ret}_k(\pi_k^{\mathrm{report}})$ and $G_0 = 2Q_{\max}$. For preservation, use the normalized occupancy $d_{i,\zeta}^{\vartheta} = c\sum_{t\geq 0}\gamma^t\zeta(P_i^{\vartheta})^t$ and the density-ratio norm $\mathsf{h}(\zeta \mid \rho) = \|d\zeta/d\rho\|_{L_2(\rho)}$, with value $+\infty$ unless $\zeta \ll \rho$. A return law ζ is covered for candidate ϑ when $\mathsf{h}(d_{i,\zeta}^{\vartheta} \mid \mu_i) < \infty$.

As specified in Table 2 of Appendix D.9.1, Calibration G sets the error targets, preservation tolerance τ, and budgets for the general guarantee, while Calibration S tightens active-critic accuracy for compatible sharing. For active evaluation call v and record i, the targets are

$$\|\widehat{Q}_v - Q_{k_v}^{\pi_v}\|_{\varrho_v} \leq \eta_Q, \qquad \|\widehat{V}_i - V_i^{\bar{\pi}_i}\|_{\upsilon_i} \leq \eta_{V,i} \leq \eta_V.$$

The uniform calibration uses $\tau_i = \tau$; put $A_{V,i} = \sqrt{C_{0,i}} + \gamma\sqrt{C_{1,i}}$. The trained subroutines meet these targets on the joint success event.

Theorem 1 (End-to-end guarantee)**.** *Under Assumptions 1–3, run Algorithm 1 with $0 < \varepsilon < 8Q_{\max}/c$, Calibration G, and failure budget δ. With probability at least $1-\delta$, simultaneously:*

(i) ***Identification.*** *Every segment is routed correctly, with one record per encountered family.*

(ii) ***Anchored preservation.*** *Every candidate ϑ installed while family i is protected satisfies, for every covered return law ζ,*

$$\zeta\big[(V_i^{\bar{\pi}_i} - V_i^{\vartheta})_+\big] \leq \frac{\mathsf{h}(d_{i,\zeta}^{\vartheta} \mid \mu_i)}{c}(\tau_i + A_{V,i}\eta_{V,i}). \tag{8}$$

This includes completed records and final reporting; incidental gains above the anchor need not be retained.

(iii) ***Progress.*** *Every own slot installs a step of weight at least $\lambda_{\min}$, using at most K modules and $K-1$ splits, without a compatibility assumption.*

(iv) ***Accuracy.*** *Every family completes J own updates, and*

$$\mathrm{gap}_k \leq \underbrace{G_0 e^{-c\lambda_{\min}J}}_{\textit{finite updates}} + \underbrace{\frac{3H_Q^{\mathrm{SA}}\eta_Q}{c^2}}_{\textit{critic}} + \underbrace{\frac{H_M(\tau + A_V\eta_V)}{c^2\lambda_{\min}}}_{\textit{interference}} + \underbrace{\frac{2\alpha Q_{\max}}{c}}_{\textit{exploration}} \leq \frac{7\varepsilon}{8}. \tag{9}$$

(v) ***Budget.*** *The explicit observation bound* (112) *includes unused reserves, evaluator geometry, auxiliary complexity, and all charged acquisition and spacing.*

Items (ii)–(iii) separate protection from progress: a certified shared update is controlled relative to every protected anchor, while failure to certify a sufficiently large step triggers an exact-copy full update for the active family. In item (iv), the first three terms control the gap to the best policy in the exploratory class and the final term converts this to the unrestricted optimum. Exact copying changes no existing policy and introduces no additional error term. The auxiliary Bellman approximation errors in Assumption 3(b) are included in the calibrated regression radii and therefore do not appear separately in (8)–(9). The proof is in Appendix C.1.

Why fixed anchors matter. Between two own updates of family i, its anchor and snapshot remain fixed. Every accepted update affecting its shared module is checked against this same baseline. Consequently, (8) bounds the current policy's loss directly, without summing a loss allowance over all intervening installations. After an own update, the new policy becomes the anchor for the next dormant interval; completed records remain protected through final reporting. Across successive own-update intervals, the contractive improvement operators and joint coverage bounds control the transported errors, yielding the interference term in (9) without an additional factor J. The one-sided comparison also permits beneficial shared changes while protecting the named anchor.

Corollary 2 (Observation rate under accuracy-uniform regularity)**.** *In Theorem 1, let $c, Q_{\max}, \nu_{\min}, \lambda_{\min} > 0$ have fixed positive bounds, and suppose $H_Q^{\mathrm{SA}}, H_M, A_V, A_{\mathrm{id}} = O(1)$*

uniformly over the calibrations as $\varepsilon, \Delta_0 \downarrow 0$. Suppose the actual evaluator factors, auxiliary covering complexities, and mixing/acquisition factors grow at most polylogarithmically in $1/\varepsilon$ and $1/\Delta_0$, with the prescribed evaluator representation, auxiliary approximation, initialization, and numerical conditions holding at every selected accuracy. Then

$$N_{\mathrm{total}} = \widetilde{\mathcal{O}}\Big(K\varepsilon^{-2} + K(h-1)p_{\log}^{-1}\varepsilon^{-2} + (S+K)\Delta_0^{-2}\Big). \tag{10}$$

An exploratory logger with $p_{\log} = \Theta(\varepsilon)$ gives a preservation term $K(h-1)\varepsilon^{-3}$; a fixed-floor logger gives $\widetilde{\mathcal{O}}(Kh\varepsilon^{-2} + (S+K)\Delta_0^{-2})$.

The proof is in Appendix C.2. This is an accuracy-uniform specialization of the complete ledger, not a consequence of finite coverage or conditioning at one fixed tolerance. If coverage, conditioning, representation complexity, or acquisition factors deteriorate as the exploration floor vanishes, their dependence remains in the general bound. Evaluation and preservation are indexed by distinct families, while explicit segment dependence enters through identification.

The next result characterizes when a single deployed module suffices and when separation is necessary. Compatibility requires that the active family's true greedy action have margin at least $a_{\min} > 0$ and also maximize every protected anchor's action value. The transfer condition $\mu_i \le \kappa_{\mathrm{tr}}\rho_{k,\pi}$ carries active-family critic accuracy to protected-family scores. These conditions hold uniformly over the histories in Assumption 4 of Appendix D.8; a common optimal policy alone is insufficient, and the algorithm does not test compatibility.

Theorem 3 (Compatible sharing and necessary separation)**.** *Under Assumptions 1–3:*

(a) ***One compatible group.*** *If all K families satisfy Assumption 4 with supplied margin $a_{\min} > 0$ and transfer bound κ_{tr}, then $h = K$ and Calibration S give Theorem 1's conclusions with probability at least $1-\delta$, with $\lambda = 1$ at every own slot, one deployed module, and no splits.*

(b) ***Necessary separation.*** *There are compatible and conflicting two-family instances on $[0,1]$ with common supplied bounds and a constant $L_{\mathrm{conf}} > 0$. On the conflicting instance, every common stationary policy has gap at least L_{conf} on one family. If the stated assumptions hold on both instances with common supplied budgets, then for $\varepsilon < L_{\mathrm{conf}}$ in the admissible range, Algorithm 1 with $K = h = 2$ and Calibration S is ε-accurate with probability at least $1-\delta$ for each instance separately. It uses one module and no split on the compatible instance, and two modules with exactly one split on the conflicting instance (Appendix D.10.2).*

The proof is in Appendix C.3. Part (a) shows that compatible directions make every full shared update certifiable; part (b) shows that one common policy cannot attain the target accuracy in the conflicting construction. These are structural module-count conclusions, not observation- or total-memory improvements. When $p_{\min} = \Theta(\varepsilon)$, Calibration S can require active-critic accuracy of order $\varepsilon^{3/2}$ and hence a sufficient $\widetilde{\mathcal{O}}(K\varepsilon^{-3})$ active-critic term under the corresponding uniformity conditions; this is an upper budget, not a lower bound on the cost of sharing.

Scope. Records, snapshots, reserves, and immutable mixture components remain, so fewer deployed modules need not reduce observations or retained memory. Evaluator geometry, Bellman approximation, and coverage remain assumptions, including for the continuous constructions; width alone does not establish them. Separation concerns a common stationary policy without family-dependent input, and the single-group result does not discover an optimal partition of multiple compatibility groups. A proof roadmap and the locations of all needed results are given in Appendix A.

6 Empirical evaluation

We evaluate the learned identification, preservation, sharing, and splitting pipeline in a controlled recurrent task, with CartPole as a supporting continuous-state demonstration. Both studies use practical estimators and budgets rather than the calibrated execution of Theorem 1.

Recurrent task and comparisons. Three latent families recur over 15 segments with supplied boundaries. Observations in $[0,1]$, with three actions: advance, retreat, and take. Families A and B share a goal but differ in rewards; C requires the opposite direction. Width-64 neural networks

learn critics, references, snapshots, and Bellman images using fitted policy evaluation and auxiliary regression. Ten randomly drawn seeds are paired across methods. We compare the full method with ungated sharing, no splitting, independent policies, a symmetric preservation gate, and a linear full-method reference. Ungated sharing installs updates without preservation checks. No splitting retains these checks but rejects proposals that cannot be certified. Independent policies allocate one module per inferred record. The symmetric gate tests Bellman residuals of both signs, isolating the effect of the one-sided criterion. The linear reference retains the outer algorithm while changing the fitted function classes. Exact model quantities are used only for assessment; the known cells make this a test of the decision pipeline in a supplied finite representation (Appendices E.1.1–E.1.2).

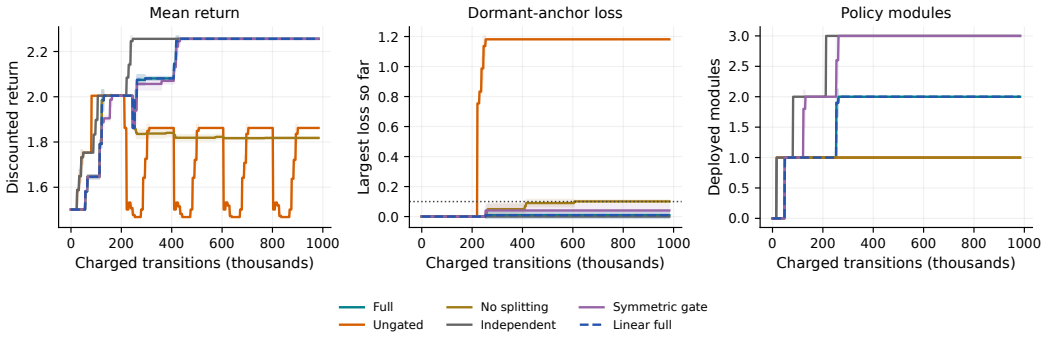


Figure 1: Recurrent learning over ten paired seeds: mean return across families, run-level maximum dormant-anchor loss, and deployed modules versus charged transitions. Shading gives pointwise 95% intervals; several curves overlap. Appendix E.1 reports complete outcomes and costs.

Performance, preservation, and sharing. The full method matches independent policies' final mean return of 2.256, reaching every family's exploration-constrained optimum in all ten runs with two deployed modules instead of three. Independent policies learn faster, and fewer modules do not imply lower total storage: records, snapshots, and reserves remain (Appendix E.1.5). The mean run-level maximum dormant-anchor loss is 0.0094, versus 1.181 for ungated sharing; no full-method run exceeds the 0.10 reporting tolerance. The method averages 7.3 positive-gain policy changes per run, including twelve updates across the cohort that improve both the active family and at least one protected family relative to their preceding policies. Removing splitting lowers final mean return to 1.818. Symmetric preservation reaches the optimum but uses three modules. Neural and linear full methods produce identical observed policy trajectories in this representation.

CartPole adaptation. An exploratory five-seed study uses raw four-dimensional states, four parent families, and four perturbed descendants. Descendants perturb parent gravity, pole mass, half-length, and force by 2.5–30%, while preserving rewards and action mappings. Family parameters and group assignments are withheld. The learner starts without heads or records and permits up to eight heads and sixteen records. All runs retain four shared heads and eight records; all twenty descendant–seed pairs improve over their pre-adaptation returns (mean normalized gain 0.1658). Independent assessment uses 128 matched initial states. Across all five seeds, the maximum measured losses relative to the designated original-family baselines and protected anchors are zero. Appendix E.3 gives the protocol, per-seed outcomes, rejected proposals, and the distinction between group routing and exact-record retrieval.

7 Conclusion and Limitations

We developed a continual learning framework that lets recurring environments share policies while protecting earlier performance. When sharing cannot be certified, splitting allows learning to continue independently. We provided the performance guarantee for our algorithm and characterized the end-to-end budget cost. Future directions include detecting unknown boundaries across environment changes, discovering compatible groups, and reducing total memory use.

AI USE STATEMENT

Generative AI tools were used for writing assistance. The authors verified all mathematical statements, proofs, citations, and experimental results, and take full responsibility for the content of the submission.

REPRODUCIBILITY STATEMENT

The attached code reproduces the recurrent experiments and reconstructs the figures and tables from scratch. Dependencies, commands, and reproduction scope are detailed in the accompanying README files.

REFERENCES

David Abel, André Barreto, Benjamin Van Roy, Doina Precup, Hado Van Hasselt, and Satinder Singh. A definition of continual reinforcement learning. *Advances in Neural Information Processing Systems*, 36:50377–50407, 2023.

Sanae Amani, Lin Yang, and Ching-An Cheng. Provably efficient lifelong reinforcement learning with linear representation. In *The Eleventh International Conference on Learning Representations*, 2022.

Haitham Bou Ammar, Rasul Tutunov, and Eric Eaton. Safe policy search for lifelong reinforcement learning with sublinear regret. In *International Conference on Machine Learning*, pages 2361–2369. PMLR, 2015.

András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71 (1):89–129, 2008.

André Barreto, Will Dabney, Rémi Munos, Jonathan J Hunt, Tom Schaul, Hado Van Hasselt, and David Silver. Successor features for transfer in reinforcement learning. *Advances in neural information processing systems*, 30, 2017.

Emma Brunskill and Lihong Li. Sample complexity of multi-task reinforcement learning. *arXiv preprint arXiv:1309.6821*, 2013.

Qi Cai, Zhuoran Yang, Jason D Lee, and Zhaoran Wang. Neural temporal-difference learning converges to global optima. *Advances in neural information processing systems*, 32, 2019.

Jonathan Chang, Kaiwen Wang, Nathan Kallus, and Wen Sun. Learning bellman complete representations for offline policy evaluation. In *International Conference on Machine Learning*, pages 2938–2971. PMLR, 2022.

Assaf Hallak, Dotan Di Castro, and Shie Mannor. Contextual markov decision processes. *arXiv preprint arXiv:1502.02259*, 2015.

Daniel Hsu and Sivan Sabato. Loss minimization and parameter estimation with heavy tails. *Journal of Machine Learning Research*, 17(18):1–40, 2016.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the nineteenth international conference on machine learning*, pages 267–274, 2002.

Christos Kaplanis, Murray Shanahan, and Claudia Clopath. Policy consolidation for continual reinforcement learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3242–3251. PMLR, 09–15 Jun 2019.

Zhifa Ke, Zaiwen Wen, and Junyu Zhang. An improved finite-time analysis of temporal difference learning with deep neural networks. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 23407–23429. PMLR, 21–27 Jul 2024.

Romain Laroche, Paul Trichelair, and Remi Tachet Des Combes. Safe policy improvement with baseline bootstrapping. In *International conference on machine learning*, pages 3652–3661. PMLR, 2019.

Erwan Lecarpentier, David Abel, Kavosh Asadi, Yuu Jinnai, Emmanuel Rachelson, and Michael L Littman. Lipschitz lifelong reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8270–8278, 2021.

Yao Liu, Zhaohan Guo, and Emma Brunskill. Pac continuous state online multitask reinforcement learning with identification. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 438–446, 2016.

Jorge Mendez, Boyu Wang, and Eric Eaton. Lifelong policy gradient learning of factored policies for faster training without forgetting. *Advances in Neural Information Processing Systems*, 33: 14398–14409, 2020.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5), 2008.

Quan Nguyen and Nishant Mehta. Adversarial online multi-task reinforcement learning. In *International Conference on Algorithmic Learning Theory*, pages 1124–1165. PMLR, 2023.

Samet Oymak and Mahdi Soltanolkotabi. Overparameterized nonlinear learning: Gradient descent takes the shortest path? In *International Conference on Machine Learning*, pages 4951–4960. PMLR, 2019.

Binghui Peng and Christos Papadimitriou. The complexity of non-stationary reinforcement learning. In *International Conference on Algorithmic Learning Theory*, pages 972–996. PMLR, 2024.

Dominic Richards and Mike Rabbat. Learning with gradient descent and weakly convex losses. In *International Conference on Artificial Intelligence and Statistics*, pages 1990–1998. PMLR, 2021.

Mark Ring and Tom Schaul. Q-error as a selection mechanism in modular reinforcement-learning systems. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, page 1452, 2011.

Mark Bishop Ring. *Continual learning in reinforcement environments*. The University of Texas at Austin, 1994.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

Bruno Scherrer. Approximate policy iteration schemes: A comparison. In *International Conference on Machine Learning*, pages 1314–1322. PMLR, 2014.

Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. In *International conference on machine learning*, pages 4528–4537. PMLR, 2018.

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In *International Conference on Machine Learning*, pages 2380–2388. PMLR, 2015.

Sebastian Thrun and Tom M Mitchell. Lifelong robot learning. *Robotics and autonomous systems*, 15(1-2):25–46, 1995.

Maciej Wołczyk, Michał Zając, Razvan Pascanu, Łukasz Kuciński, and Piotr Miłoś. Continual world: A robotic benchmark for continual reinforcement learning. *Advances in Neural Information Processing Systems*, 34:28496–28510, 2021.

Zhenghao Xu, Xiang Ji, Minshuo Chen, Mengdi Wang, and Tuo Zhao. Sample complexity of neural policy mirror descent for policy optimization on low-dimensional manifolds. *Journal of Machine Learning Research*, 25(226):1–67, 2024.

Qixin Zeng, Shuo Zhang, Hongyin Zhang, Renjie Wang, Han Zhao, Libang Zhao, Runze Li, Donglin Wang, and Chao Huang. Crl-vla: Continual vision-language-action learning. *arXiv preprint arXiv:2602.03445*, 2026.

Zhi Zhang, Chris Chow, Yasi Zhang, Yanchao Sun, Haochen Zhang, Eric Hanchen Jiang, Han Liu, Furong Huang, Yuchen Cui, and Oscar Hernan Madrid Padilla. Statistical guarantees for lifelong reinforcement learning using pac-bayes theory. *arXiv preprint arXiv:2411.00401*, 2024.

APPENDIX/PROOF ROADMAP

The appendix separates the proofs of the main statements from the supporting results they use. Appendix B compares theoretical objectives with related work. Appendix C follows the order of the main text: Theorem 1, Corollary 2, and Theorem 3. Appendix D collects the execution specifications, statistical and analytic ingredients, numerical budgets, and constructions. The outline below explains the ingredients and then shows explicitly how they combine into the final main statements, with links to their complete proofs. The linked appendix contents follow the outline.

A INGREDIENTS AND PROOF IDEAS

The proof first establishes a joint event on which all trained evaluations, Bellman fits, and monitors are accurate. On this event, routing identifies families, certificates preserve fixed anchors, and exact-copy fallback ensures progress. The final part of this outline, Appendix A.2, shows how these ingredients give the stated accuracy and observation bounds and the two sharing conclusions.

Reading routes and notation. For the complete main argument, read Appendix C. For the theoretical execution and its assumptions, begin with Appendix D.1, followed by the statistical and analytic results in Appendix D. Table 2 and Appendix D.9 collect the numerical requirements and confidence allocations; these remain part of the theorem premises. As in the main text, $c = 1 - \gamma$, J is the accepted own-update quota, and $\lambda_{\min}$ is the minimum step size. Supporting proofs also write $J_k(\pi) = \mathrm{Ret}_k(\pi)$ for a family's return.

Empirical and comparative material. Appendix E gives the recurrent experiment's setup, compared methods, complete results, costs, and execution diagnostics, followed by the supporting CartPole protocol and outcomes. Appendix F compares theoretical objectives with related work.

A.1 THE INGREDIENTS AND THEIR ROLES

Preservation through a fixed anchor. A one-step Bellman deficit controls long-term loss. For anchor $\bar{\pi}$ and candidate ϑ,

$$V^{\bar{\pi}} - V^{\vartheta} = (I - \gamma P^{\vartheta})^{-1}(V^{\bar{\pi}} - T^{\vartheta} V^{\bar{\pi}}). \tag{11}$$

Positivity of the discounted resolvent bounds the positive value loss by the propagated positive Bellman residual. Snapshot error contributes $A_V \eta_V$; coverage gives (8). Every protected update compares with the same frozen anchor, so dormant loss is charged once between own updates, including the final reporting interval. Beneficial changes are allowed, whereas a symmetric residual can reject them even when both families improve (Appendix D.10.2).

From accepted updates to final accuracy. Exact-copy fallback ensures every own update has weight at least $\lambda_{\min}$. Each update involves three policy-averaged critic errors: against the incumbent, fitted-greedy endpoint, and optimal exploratory policy. For fixed family k, let $\pi^*_{k,\alpha}$ be optimal in Π_α and put $P_* = P_k^{\pi^*_{k,\alpha}}$. A step λ_j contracts the value gap through $(1-\lambda_j)I + \gamma\lambda_j P_*$. These comparison operators use the same kernel even for adaptive steps. Define $u_j(z) = \prod_{\ell=j}^{J-1}(1 - \lambda_\ell + \gamma\lambda_\ell z)$, with $u_J = 1$. Telescoping gives

$$\sum_{j=0}^{J-1} \lambda_j u_{j+1}(z) = \frac{1 - u_0(z)}{1 - \gamma z} \preceq \frac{1}{1 - \gamma z}, \qquad \sum_{j=0}^{J-1} u_{j+1}(z) \preceq \frac{1}{\lambda_{\min}(1 - \gamma z)}. \tag{12}$$

Here $\preceq$ compares power-series coefficients. Nonnegativity and joint coverage control accumulated critic and dormant errors without an extra factor J. The initial gap contracts by at most $e^{-c\lambda_{\min} J}$, and exploration contributes $2\alpha Q_{\max}/c$, giving (9) (Appendix D.6).

Why compatible sharing persists. Under compatibility, the full endpoint's exact anchor Bellman residual is nonpositive wherever the fitted greedy action matches the common maximizing action. The action margin bounds disagreement mass by $4\eta_Q^2/(a_{\min}^2 p_{\min})$ under the active stationary state law; domination transfers it to protected monitor laws. Snapshot-error control and Calibration S

leave slack for solve and audit checks. Reserves supply all required checks before any first split, so induction gives one module and no splits. A positive-resolvent lower bound rules out one accurate common policy in the conflicting construction (Appendices D.8 and D.10.2).

Training value evaluators. Under the stated evaluator representation and conditioning conditions, projected neural TD learns policy values to root L_2 error η using $\widetilde{\mathcal{O}}(E\eta^{-2})$ transitions, where E retains geometry and collection factors. Derivative bounds control nonlinear perturbations, and comparison with the tangent model controls TD error. Independent clipped replicas, distance-based selection, and spaced-tuple coupling give conditional high-probability guarantees; Jensen's inequality gives value snapshots. Under uniform geometry, width $\widetilde{\mathcal{O}}(\eta^{-4})$ suffices provided the local representation condition holds at the selected width (Appendix D.3, Proposition 3).

Training and validating Bellman regressors. Bellman fits must remain accurate for adaptively selected mixtures. Ridge curvature controls optimization on the selected parameter ball; covering bounds control statistical error. The calibrated approximation condition $a \le u/3$ contributes at most $u^2/3$ to the squared fitted-error budget. With the prescribed width and optimization budgets, root error u costs $\widetilde{\mathcal{O}}(u^{-2})$ labels with logging-variance and covering factors retained (Proposition 6). Positive-part contraction and affine interpolation make endpoint-error bounds uniform over all candidate mixtures, including solve-selected candidates. Freeze targets before labels, fits before disjoint monitors, and candidates before fresh audits. Conditional guarantees and a finite call-wise failure allocation validate adaptive execution (Proposition 11); Appendices D.4 and D.5 give the complete bounds.

A.2 Obtaining the final theorem statements

The ingredients above close the main results in three steps. First, their simultaneous guarantees yield all five clauses of Theorem 1. Second, substituting Calibration G into the complete observation ledger gives Corollary 2 under its additional uniformity conditions. Third, Calibration S and the compatible and conflicting constructions give Theorem 3. The following derivations identify the final inequalities and the precise locations of the full proofs.

A.2.1 The end-to-end theorem: assembling all five conclusions

Full proof: Appendix C.1, starting on page 21. The proof's paragraphs (i)–(v) correspond directly to the five clauses of Theorem 1.

The common success event. The conditional complete-call guarantees and the adaptive union bound of Lemma 4, with the allocation in Appendix D.9.3, give an event $\mathcal{E}$ with $\Pr(\mathcal{E}) \ge 1-\delta$. It includes every reached evaluator, regression, routing monitor, solve monitor, audit, and required acquisition event under the declared caps. Choices are frozen before their corresponding observations are disclosed. All subsequent transport and counting arguments are deterministic on $\mathcal{E}$; they require neither independence of the adaptive updates nor a new failure budget for each theorem clause.

(i) Identification follows by separating the empirical scores. The diagnostic reference tolerance $\eta_{\mathrm{id}} = \Delta_0/(8A_{\mathrm{id}})$ and the two diagnostic dominations give transferred reference error at most $\Delta_0/8$, as shown in (13). The fitted Bellman comparison and its common monitor each contribute at most $\Delta_0/16$. Consequently,

$$\text{matching score} \le \frac{\Delta_0}{4}, \qquad \text{nonmatching score} \ge \Delta_0 - \frac{\Delta_0}{8} - \frac{\Delta_0}{16} - \frac{\Delta_0}{16} = \frac{3\Delta_0}{4}.$$

The threshold $\Delta_0/4$ therefore accepts the correct reference when it exists and rejects every stored reference for a new family. Induction over the S segments gives correct routing and exactly one record per encountered family. This is Proposition 10, used in paragraph (i) of Appendix C.1; the conclusion is independent of whether different records share a deployed module.

(ii) Preservation follows by transferring the certified residual. The uniform solve and fresh-audit guarantees certify the population snapshot residual $R_i^+ \le \tau_i$ for every installed protected candidate. Lemma 32 converts this to the exact-anchor bound

$$\|(V_i^{\bar\pi_i} - T_i^{\vartheta} V_i^{\bar\pi_i})_+\|_{\mu_i} \le \tau_i + A_{V,i}\eta_{V,i}.$$

Applying Theorem 33 to the positive resolvent in (11) gives, for every covered return law,

$$\zeta\big[(V_i^{\bar{\pi}_i} - V_i^{\vartheta})_+\big] \le \frac{\mathsf{h}(d_{i,\zeta}^{\vartheta} \mid \mu_i)}{c}(\tau_i + A_{V,i}\eta_{V,i}).$$

This is exactly the main preservation statement (8). The anchor remains fixed between accepted own updates, including the final reporting interval; completed records remain protected. Hence intermediate dormant losses are not added together. An exact-copy split changes no existing action law and introduces no additional preservation error. The complete combination is paragraph (ii) of Appendix C.1, using the certificate and return-transport results in Appendices D.5 and D.6.

(iii) Progress follows from certification or exact-copy fallback. Theorem 35 shows that every own slot either accepts a certified shared step with $\lambda \ge \lambda_{\min}$ or isolates the active record and installs the same critic's full step on its singleton copy. An unavailable reserve or a failed solve or audit uses this same fallback. A nonsingleton can split only when fewer modules than records are occupied, so one of the K slots is free. Every split adds one module, and modules are never merged. Thus there are at most K modules and $K-1$ splits. The ordered active-time premise supplies the required acquisition and J own slots per family, so all quotas complete. This proves clause (iii) and the completion part of clause (iv); see paragraph (iii) of Appendix C.1 and Appendix D.7.

(iv) Final accuracy follows by substituting Calibration G. The preceding preservation and progress conclusions supply the hypotheses of Theorem 30. Exact exploratory greedification makes its endpoint defect b_g zero. Its general bound (88) therefore becomes

$$\mathrm{gap}_k \le G_0 e^{-c\lambda_{\min} J} + \frac{3H_Q^{\mathrm{SA}}\eta_Q}{c^2} + \frac{H_M(\tau + A_V\eta_V)}{c^2\lambda_{\min}} + \frac{2\alpha Q_{\max}}{c}, \qquad G_0 = 2Q_{\max}.$$

The operator-polynomial bounds in (12) control the accumulated critic and dormant errors without an extra factor J, including the final dormant suffix. Table 2 sets

$$\alpha = \frac{c\varepsilon}{8Q_{\max}}, \qquad \eta_Q = \frac{c^2\varepsilon}{12H_Q^{\mathrm{SA}}}, \qquad \tau = \frac{c^2\lambda_{\min}\varepsilon}{10H_M}, \qquad \eta_V = \frac{\tau}{4A_V}.$$

Together with the prescribed ceiling in J, these choices give

$$G_0 e^{-c\lambda_{\min} J} \le \varepsilon/4, \qquad \frac{3H_Q^{\mathrm{SA}}\eta_Q}{c^2} = \varepsilon/4,$$
$$\frac{H_M(\tau + A_V\eta_V)}{c^2\lambda_{\min}} = \frac{5H_M\tau}{4c^2\lambda_{\min}} = \varepsilon/8, \qquad \frac{2\alpha Q_{\max}}{c} = \varepsilon/4.$$

Thus the final theorem inequality is obtained explicitly as

$$\boxed{\mathrm{gap}_k \le \frac{\varepsilon}{4} + \frac{\varepsilon}{4} + \frac{\varepsilon}{8} + \frac{\varepsilon}{4} = \frac{7\varepsilon}{8} < \varepsilon \quad \text{for every family } k.}$$

This is (9), with the final reporting policy identified in (14). Paragraph (iv) of Appendix C.1 contains the closing substitution; Appendix D.6 proves the transport bound and Appendix D.9 specifies its numerical budgets. The auxiliary approximation errors are already included in the calibrated regression radii through $a \le u/3$, so they are not added a second time to this four-term inequality.

(v) The observation bound follows by multiplying complete costs by counts. Equation (99), together with one diagnostic reference per family and one routing call per segment, bounds the execution by KJ active evaluations, $K(J+1)$ value snapshots, K diagnostic-reference evaluations, $K(h-1)J$ collected preservation blocks, and S routing calls. The preservation count includes unused reserves. A split retry reuses its active critic, and the two endpoint labels use the same retained transitions. Multiplying these counts by the complete evaluator costs of Corollary 11 and the preservation-block cost of Corollary 25 gives

$$N_{\mathrm{total}} \le N_{\mathrm{diag}} + N_{\mathrm{acq}} + \widetilde{\mathcal{O}}\Big(KJE_Q\eta_Q^{-2} + K(J+1)E_V\eta_V^{-2} + K(h-1)Jp_{\log}^{-1}Q_{\max}^2(\mathfrak{C}_{\mathrm{reg}}+1)\tau^{-2}\Big),$$
$$N_{\mathrm{diag}} = \widetilde{\mathcal{O}}\big([KE_{\mathrm{id}}A_{\mathrm{id}}^2 + SQ_{\max}^2(\mathfrak{C}_{\mathrm{id}}+1)]\Delta_0^{-2}\big).$$

These are the complete ledger (112) and diagnostic cost (113). Each routing call uses a common labeled block and monitor across stored references; there is no additional factor K in its retained-transition count. The term N_{acq} adds only acquisition and spacing not already charged in a complete call, according to (15). The actual evaluator geometry, auxiliary complexity, and collection factors remain in this bound. Paragraph (v) of Appendix C.1 performs the accounting; Appendix D.9.4 gives the full ledger and budgets.

Where the theorem is completed. All five conclusions now hold on the same event $\mathcal{E}$. Together with $\Pr(\mathcal{E}) \geq 1 - \delta$, this yields Theorem 1 exactly as stated in the main text. The final assembly is the end of Appendix C.1.

A.2.2 The observation-rate corollary: specializing the ledger

Full proof: Appendix C.2, page 23. Corollary 2 is obtained by substituting Calibration G into the preceding ledger, retaining its additional accuracy-uniform regularity assumptions. The prescribed quota satisfies $J = O(\log(1/\varepsilon))$ when the stated basic constants have fixed positive bounds. Before absorbing the coverage and geometry factors, the three control and preservation terms become

$$\widetilde{\mathcal{O}}\left(\frac{KE_Q(H_Q^{\mathrm{SA}})^2}{\lambda_{\min}\varepsilon^2} + \frac{KE_V A_V^2 H_M^2}{\lambda_{\min}^3\varepsilon^2} + \frac{K(h-1)(\mathfrak{C}_{\mathrm{reg}}+1)H_M^2}{p_{\log}\lambda_{\min}^3\varepsilon^2}\right),$$

with fixed discount and reward factors absorbed. Under the corollary's uniform bounds, the evaluator terms are $\widetilde{\mathcal{O}}(K\varepsilon^{-2})$, the preservation term is $\widetilde{\mathcal{O}}(K(h-1)p_{\log}^{-1}\varepsilon^{-2})$, and identification contributes $\widetilde{\mathcal{O}}((S+K)\Delta_0^{-2})$. The assumed acquisition inflation is absorbed in these same orders. Their sum gives the claimed rate:

$$\boxed{N_{\mathrm{total}} = \widetilde{\mathcal{O}}\Big(K\varepsilon^{-2} + K(h-1)p_{\log}^{-1}\varepsilon^{-2} + (S+K)\Delta_0^{-2}\Big).}$$

This is (10). Substituting $p_{\log} = \Theta(\varepsilon)$ yields the preservation term $\widetilde{\mathcal{O}}(K(h-1)\varepsilon^{-3})$; a fixed positive logging floor instead yields $\widetilde{\mathcal{O}}(Kh\varepsilon^{-2} + (S+K)\Delta_0^{-2})$. The final paragraph of Appendix C.2 specifies why these substitutions require the evaluator representation and conditioning, auxiliary approximation, and numerical conditions to hold at every selected accuracy. Finiteness at one tolerance, or a fixed logging floor by itself, does not establish this corollary.

A.2.3 The sharing theorem: certifying one module and forcing separation

Full proof: Appendix C.3, starting on page 23. Calibration S changes only the active-critic tolerance to

$$\eta_Q^{\mathrm{S}} = \min\left\{\eta_Q^{\mathrm{G}}, \frac{\tau\sqrt{p_{\min}}}{8C_{\mathrm{sh}}}\right\},$$

as specified in (118), and recomputes the complete evaluation budget at that tolerance. Since $\eta_Q^{\mathrm{S}} \leq \eta_Q^{\mathrm{G}}$, the general end-to-end proof continues to give preservation, progress, and accuracy whenever its execution premises hold at the tighter precision. Compatibility is needed for the additional sharing conclusion, not for this accuracy guarantee.

(a) From full-step slack to one module and no splits. Under Assumption 4, the direct certificate of Theorem 38 and Calibration S give, for every protected record,

$$R_i^+(1) \leq A_{V,i}\eta_{V,i} + \frac{C_{\mathrm{sh}}\eta_Q^{\mathrm{S}}}{\sqrt{p_{\min}}} \leq \frac{\tau}{4} + \frac{\tau}{8} = \frac{3\tau}{8}.$$

The calibrated solve and audit radii satisfy $e_i, e_{a,i} \leq \tau/8$. Hence the respective upper bounds at the full endpoint satisfy

$$\text{solve upper bound} \leq R_i^+(1) + 2e_i \leq \frac{5\tau}{8} < \frac{3\tau}{4},$$

$$\text{audit upper bound} \leq R_i^+(1) + 2e_{a,i} \leq \frac{5\tau}{8} < \tau.$$

Thus the endpoint-first rule selects $\lambda = 1$ and its audit passes whenever the required block is available.

It remains to rule out a split caused by exhausted reserves. Before a hypothetical first split, join-before-create and $h = K$ put every record in the first module. Each protected record can be tested at most $(K-1)J$ times, once for each possible own update of the other families. Its single initial reserve contains precisely this many complete blocks, so a block is available before every required test. Anchor refreshes do not refill the reserve. This is the inventory induction in Corollary 40. Neither reserve exhaustion nor failure of a solve or audit can cause the first split. Therefore

$\lambda = 1$ at every own slot,	1 deployed module,	0 splits.

Together with the general accuracy guarantee under the tighter tolerance, this is Theorem 3(a). The final induction is paragraph (a) of Appendix C.3; its direct-certificate and calibration ingredients are in Appendices D.8 and D.9.6.

(b) From a common-policy lower bound to exactly one split. The matched compatible and conflicting instances are defined in (123)–(126). Propositions 15 and 16 verify their environmental and diagnostic properties, and the latter also verifies compatibility for the compatible pair. For the conflicting pair, Proposition 17 proves that every common stationary policy satisfies

$$\max_{k=1,2}\{J_k(\pi_k^*) - J_k(\pi)\} \geq L_{\mathrm{conf}} := \frac{a_*(c+\gamma/2)}{2c} > 0, \qquad a_* = (1+\gamma)/4.$$

Retain the theorem's execution premises on both instances at the common supplied budgets, and take an admissible $\varepsilon < L_{\mathrm{conf}}$. The general guarantee under Calibration S gives gap at most $7\varepsilon/8$ for both families of each instance. Part (a) keeps the compatible pair in one module with no split. On the conflicting pair, one final module would supply a common stationary policy and would imply

$$L_{\mathrm{conf}} \leq \max_{k=1,2} \mathrm{gap}_k \leq 7\varepsilon/8 < L_{\mathrm{conf}},$$

a contradiction. There must be two final modules. The general $K = 2$ bound allows at most two modules and one split, while join-before-create initially puts both records in the same module. Consequently the conflicting pair has exactly one split. This proves the final comparison in Theorem 3(b):

compatible instance:	1 module,	0 splits,
conflicting instance:	2 modules,	1 split.

Paragraph (b) of Appendix C.3 contains this closing argument, using the constructions and lower bound in Appendix D.10.2. The probability is at least $1 - \delta$ for each execution separately. The constructions establish the stated environmental properties; evaluator realization and conditioning, auxiliary approximation, and the sampling and active-time conditions remain execution premises at the prescribed budgets. The conclusion concerns deployed policy modules, with one common stationary action law per module; records, snapshots, reserves, and mixture components remain part of the execution.

Locations of the completed main arguments. The final conclusions and their proof locations are collected below. The links lead directly to the complete proofs and to the final bounds used above.

Main result	Final assembly	Complete proof
Theorem 1	Clauses (i)–(v) on $\mathcal{E}$; preservation (8), accuracy (14), ledger (112)	Appendix C.1, p. 21
Corollary 2	Calibration G and the uniformity premises turn the ledger into (10)	Appendix C.2, p. 23
Theorem 3	Calibration S, the reserve induction, and the common-policy lower bound (129)	Appendix C.3, p. 23

APPENDIX CONTENTS

A Ingredients and proof ideas . . . **13**
A.1 The ingredients and their roles . . . 13
A.2 Obtaining the final theorem statements . . . 14
B Related work . . . **19**
B.1 Recognizing recurring tasks . . . 19
B.2 Transferring knowledge and sharing representations . . . 19
B.3 Retaining behavior and expanding capacity . . . 19
B.4 Safe improvement and candidate validation . . . 20
B.5 Trained value evaluators and Bellman regressors . . . 20
B.6 Comparison of guarantees . . . 20
C Proofs of the main results . . . **21**
C.1 Proof of Theorem 1 . . . 21
C.2 Proof of Corollary 2 . . . 23
C.3 Proof of Theorem 3 . . . 23
D Supporting theoretical results . . . **24**
D.1 Execution, observation, coverage, and representation specifications . . . 24
D.2 Conditional sampling, concentration, and sealed replay . . . 28
D.3 Continuous-state neural evaluation and confidence conversion . . . 31
D.4 Auxiliary Bellman regression and all-parameter optimization . . . 36
D.5 Uniform preservation certification and family identification . . . 42
D.6 Own-update improvement, return transport, and anchored preservation . . . 47
D.7 Module splitting with finite replay reserves . . . 53
D.8 Compatible sharing and zero splits . . . 55
D.9 Calibrations, simultaneous confidence, and observation accounting . . . 56
D.10 Concrete evaluator and compatible/conflicting constructions . . . 63
E Experimental results and protocols . . . **66**
E.1 Learned neural recurrent control . . . 66
E.2 Neural implementation and execution diagnostics . . . 70
E.3 CartPole family groups and descendants . . . 70
F Comparison of theoretical objectives . . . **72**

B Related work

Continual and lifelong reinforcement learning study agents that accumulate knowledge across an evolving sequence of experiences (Ring, 1994; Thrun and Mitchell, 1995; Abel et al., 2023). Our setting focuses on recurring, fixed environment families with continuous states, finite actions, observed segment boundaries, and hidden family identities. The relevant literature addresses task recognition, knowledge transfer, retention, safe improvement, and the training of Bellman estimators. We organize the comparison around these objectives.

B.1 Recognizing recurring tasks

Brunskill and Li (2013) analyze the sample complexity of identifying and reusing a finite collection of MDPs. Hallak et al. (2015) study contextual MDPs with a known finite number of latent contexts, each fixed within an episode. Their CECE framework combines model clustering, exploration, context classification, and exploitation, with regret bounds under conditions on these components. More recently, Nguyen and Mehta (2023) establish adversarial multitask regret guarantees and lower bounds, including separation-dependent clustering results. These works make recognition and the cost of obtaining informative observations explicit parts of multitask learning.

Continuous-state identification also has precedents. COMRLI uses Lipschitz optimal action-value functions, separation between tasks, and reachability of informative neighborhoods to obtain PAC guarantees (Liu et al., 2016). Bellman prediction error has also been used to select modules in reinforcement-learning systems (Ring and Schaul, 2011). Our contribution builds on this distinction between recognizing an environment and learning within it: diagnostic Bellman comparisons select a family record, while separate preservation tests determine whether that record can continue sharing a policy. Distinct records may therefore use the same deployed policy even when their rewards or transition kernels differ. The identification guarantee still requires diagnostic separation and sufficient access to informative data, as specified in Section 4.

B.2 Transferring knowledge and sharing representations

Successor features and generalized policy improvement transfer previously learned policies to tasks with common dynamics and different rewards (Barreto et al., 2017). Their guarantees bound target-task performance using stored policies and approximate value functions. Lipschitz lifelong RL instead uses distances between MDP models to transfer optimistic value bounds, with PAC-MDP guarantees and protection against negative transfer (Lecarpentier et al., 2021). UCBlvd combines observed contexts, common dynamics, and linear representations to obtain lifelong regret and planning-efficiency guarantees (Amani et al., 2022).

Sharing can also occur in policy parameters or distributions. LPG-FTW learns a shared dictionary of policy factors and task-specific coefficients during policy-gradient training; its analysis establishes asymptotic convergence for a surrogate multitask objective (Mendez et al., 2020). EPIC learns a shared policy distribution and provides PAC-Bayes guarantees relating generalization to retained prior tasks (Zhang et al., 2024). Our sharing result concerns one evolving state-to-action policy used by several separately identified families. An update for the active family must also satisfy return-preservation requirements for inactive families. The compatible case in Theorem 3 establishes when this joint requirement permits one deployed module throughout learning.

B.3 Retaining behavior and expanding capacity

Progressive networks preserve earlier components while adding capacity (Rusu et al., 2016); Progress & Compress transfers new knowledge into a consolidated model (Schwarz et al., 2018). CLEAR combines replay with behavioral cloning (Rolnick et al., 2019). Policy consolidation regularizes the current policy through a cascade of historical policies at different timescales and does not require supplied task boundaries (Kaplanis et al., 2019). These approaches provide mechanisms for retaining behavior as learning continues. Continual World and CRL-VLA examine retention and adaptation in robotic settings (Wołczyk et al., 2021; Zeng et al., 2026).

Our method protects each inactive family relative to a retained anchor, using learned Bellman tests on finite stored observations. It keeps sharing when a candidate passes these tests and creates

an independent exact copy when certification fails. Splitting isolates subsequent changes to the active family's policy and guarantees progress under the stated assumptions. The resulting resource guarantee counts deployed policy modules; records, snapshots, and preservation reserves remain part of the method.

B.4 Safe improvement and candidate validation

Conservative policy iteration uses controlled mixtures to obtain improvement and approximate optimality in a fixed MDP, with assumptions on approximate greedy updates and the sampling distribution (Kakade and Langford, 2002). High-confidence policy improvement uses off-policy return bounds to validate a candidate against a supplied performance threshold, separates candidate selection from a held-out safety test, and also provides an incremental repeated-improvement procedure (Thomas et al., 2015). SPIBB controls uncertainty by matching a known data-collecting baseline on poorly sampled state–action pairs in a finite-MDP batch setting (Laroche et al., 2019). Safe lifelong policy search combines shared latent policy structure with prescribed parameter constraints and sublinear regret (Ammar et al., 2015).

Our certificates apply simultaneously to multiple inactive-family anchors during an adaptive sequence of shared updates. The positive part of an anchor Bellman residual controls harmful changes while allowing beneficial ones. Repeated tests use the same frozen anchor until that family updates, so the preservation analysis charges dormant loss once between own updates, including the final reporting interval. Separate solve and audit data support candidate selection and validation; finite reserves account for the observations available while a family is inactive. Exact-copy splitting provides the fallback when a shared candidate cannot be certified.

B.5 Trained value evaluators and Bellman regressors

Fitted value and policy iteration connect Bellman approximation errors to control performance under coverage and function-class conditions (Antos et al., 2008; Munos and Szepesvári, 2008). Bellman-complete representation learning makes related closure conditions explicit for offline evaluation (Chang et al., 2022). Cai et al. (2019) establish convergence of overparameterized neural TD through local linearization, and Ke et al. (2024) give an improved finite-time analysis for deep neural TD under Markovian sampling. These results motivate the evaluator geometry used here. Weak-curvature and local-Jacobian analyses inform the auxiliary regression arguments (Richards and Rabbat, 2021; Oymak and Soltanolkotabi, 2019).

We analyze trained evaluators with adjustable hidden weights and fixed readout signs, and Bellman regressors with both hidden and readout parameters trained. Their error bounds are combined with policy-error propagation (Scherrer, 2014; Xu et al., 2024), confidence amplification (Hsu and Sabato, 2016), and concentration conditional on the visible history. This connects finite training to identification and preservation decisions for adaptively chosen policies and learned targets. The guarantees retain local realization, conditioning, coverage, and sampling assumptions. Comparisons of neural-TD rates must align root versus squared error, approximation terms, probability mode, and collection costs.

B.6 Comparison of guarantees

Table 1 summarizes the closest theoretical comparisons. The assumptions, performance criteria, and resource measures differ across rows.

Work	Established contribution	Relation to the present guarantee
Brunskill–Li; COMRLI (Brunskill and Li, 2013; Liu et al., 2016)	Task identification and reuse with reduced learning cost on subsequent tasks; COMRLI covers continuous states and discrete actions.	We protect inactive anchors during updates to a common policy used by distinct identified families.
Nguyen–Mehta (Nguyen and Mehta, 2023)	Adversarial multitask regret, separation-dependent clustering bounds, and corresponding lower bounds.	We analyze neural certificates and final family accuracy with preservation; no matching sample lower bound is claimed.
UCBlvd (Amani et al., 2022)	Linear representations, observed contexts, context-independent dynamics, sublinear regret, and sublinear planning calls.	We allow latent families with differing dynamics and protect inactive anchors; no planning-efficiency comparison is established.
Lipschitz lifelong RL (Lecarpentier et al., 2021)	MDP-similarity-based value transfer, PAC-MDP guarantees, and high-probability protection against negative transfer.	We bound subsequent policy changes relative to retained historical anchors.
EPIC (Zhang et al., 2024)	PAC-Bayes guarantees for shared policy distributions, with a relationship between retained tasks and generalization.	We require simultaneous individual-family anchor protection and final accuracy for the specified recurrent stream.
HCPI; SPIBB (Thomas et al., 2015; Laroche et al., 2019)	Candidate-performance certificates from off-policy data; SPIBB restricts deviation from a data-collecting baseline on poorly sampled pairs.	We protect multiple inactive anchors during adaptive shared updates, using finite reserves and independent splitting for progress.
Safe lifelong policy search (Ammar et al., 2015)	Shared latent policy structure, prescribed parameter constraints, and sublinear regret for lifelong policy search.	Our constraints arise from learned Bellman certificates for inactive-family return loss.

Table 1: Related theoretical guarantees and their relation to this paper. Each row should be read with its own assumptions and performance criterion.

The joint result combines family recognition, preservation of inactive anchors, continued active-family learning, and conditional savings in deployed modules. Our analysis connects fixed-anchor comparisons and adaptive update weights to final family accuracy while accounting for training, stored reserves, and data acquisition. The single-compatible-group result does not establish discovery of an optimal partition into sharing groups. Computational lower bounds for general local model changes (Peng and Papadimitriou, 2024) concern a broader nonstationary setting; our observation bounds rely on the stated recurrence and structural assumptions.

C Proofs of the main results

The proofs follow the order of the main statements. Detailed execution specifications and supporting results are collected in Appendix D; the numerical choices appear in Table 2 and Appendix D.9.4. Throughout, $J_k(\pi) = \mathrm{Ret}_k(\pi)$, whereas the unadorned J is the own-update quota. The simultaneous event $\mathcal{E}$ constructed in Appendix D.9.3 has probability at least $1 - \delta$. It includes every reached evaluator, regression, routing, solve-monitor, audit, and required acquisition event under the declared caps.

C.1 Proof of Theorem 1

Proof. Lemma 4, the conditional complete-call guarantees, and the separate acquisition couplings justify the event $\mathcal{E}$. All choices are frozen before the corresponding observations are disclosed, and unreached calls consume no failure probability. Work on this event for the five conclusions below.

(i) Identification. Corollary 11 supplies the reference errors at $\eta_{\mathrm{id}} = \Delta_0/(8A_{\mathrm{id}})$. For current family k and stored reference ℓ, Jensen's inequality and the two dominations in (19) give

$$\begin{aligned}\|(q_\ell - \mathcal{T}_k^{\mathrm{id}} q_\ell) - (Q_\ell^{\mathrm{id}} - \mathcal{T}_k^{\mathrm{id}} Q_\ell^{\mathrm{id}})\|_{\mu_k^{\mathrm{id}}} \\ \le (\sqrt{D_{0,k\ell}} + \gamma\sqrt{D_{1,k\ell}})\eta_{\mathrm{id}} \le \Delta_0/8.\end{aligned} \tag{13}$$

The diagnostic regression and common-monitor errors contribute at most $\Delta_0/16$ each. Consequently a matching empirical score is at most $\Delta_0/4$, while every wrong-family score is at least $\Delta_0 - \Delta_0/8 - \Delta_0/8 = 3\Delta_0/4$. The threshold $t = \Delta_0/4$ therefore accepts exactly the correct stored reference when it exists and rejects all references for a new family. Starting with the empty bank, induction over the S segments gives correct routing and one record per encountered family, as in Proposition 10. The argument does not depend on which records share a deployed module.

(ii) Anchored preservation. Every snapshot is evaluated for its named anchor, so Lemma 31 gives its value error in $L_2(\upsilon_i)$. The calibrated-regression and monitor budgets give the uniform solve radius $e_i \le \tau_i/8$ and the complete audit radius $e_{a,i} \le \tau_i/8$, each including both fitting and monitor error. Theorem 23 bounds the population snapshot residual of a feasible shared candidate by $3\tau_i/4$; its mandatory audit passes and certifies $R_i^+(\lambda) \le \tau_i$. This remains valid for the scalar selected using the solve monitor because the solve event is uniform in λ.

The snapshot dominations and Lemma 32 then give

$$\|(V_i^{\bar\pi_i} - T_i^{\vartheta} V_i^{\bar\pi_i})_+\|_{\mu_i} \le \tau_i + A_{V,i}\eta_{V,i}.$$

The positive-resolvent argument of Theorem 33 yields (8) for every covered return law on this same event. Between accepted own updates, the named anchor stays fixed. The last affecting shared candidate, including the reporting policy, is compared directly with that anchor; losses are not added across intermediate candidates. Completed records remain protected. An exact-copy split changes neither the copied action law nor the policies remaining in the old module, so it introduces no additional preservation error.

(iii) Progress and completion. By Theorem 35, a shared own slot either installs a certified step $\lambda \ge \lambda_{\min}$ or isolates the active record and installs the same critic's full step on its exact singleton copy. An unavailable reserve or a failed solve or audit invokes this same fallback. A singleton has no protected constraints. Whenever a nonsingleton splits, the number of occupied modules is strictly smaller than the number of records, so one of the K slots is free. Every split increases the module count by one, and modules are never merged. Thus there are at most K modules and $K-1$ splits. Each scheduled own slot completes one accepted own update; the ordered active-time condition supplies all acquisition work and J such slots per family. Hence every family completes its quota.

(iv) Accuracy at reporting. Index the accepted own updates of a fixed family, with the anchor refreshed immediately after each installation. Parts (ii)–(iii) supply the hypotheses of Theorem 30: all step weights are at least $\lambda_{\min}$, the active critic error is at most η_Q, and every harmful dormant change is controlled directly against the latest accepted-update anchor. Exact exploratory greedification sets the endpoint defect b_g to zero. Substitution in (88) gives the four-term inequality (9), including the final dormant suffix. Calibration G gives

$$\begin{gathered}G_0 e^{-c\lambda_{\min} J} \le \varepsilon/4, \qquad \frac{3H_Q^{\mathrm{SA}}\eta_Q}{c^2} = \varepsilon/4, \qquad \frac{2\alpha Q_{\max}}{c} = \varepsilon/4, \\ \frac{H_M(\tau + A_V\eta_V)}{c^2\lambda_{\min}} = \frac{5H_M\tau}{4c^2\lambda_{\min}} = \varepsilon/8.\end{gathered}$$

Here the first bound uses the prescribed ceiling in J and $G_0 = 2Q_{\max}$; the snapshot choice is $A_V\eta_V = \tau/4$. Consequently, simultaneously for every family,

$$J_k(\pi_k^*) - J_k(\pi^{\mathrm{report}}_{g_{\mathrm{fin}}(k)}) = \mathrm{gap}_k \le 7\varepsilon/8 < \varepsilon. \tag{14}$$

The absence of an extra factor J in the critic and dormant terms comes from the nonnegative operator-polynomial bounds in the transport theorem; it does not assume independence of the adaptive steps. The general achieved-error transport result remains applicable outside the prescribed calibration, without asserting arbitrary accuracy when its realization or numerical conditions fail.

(v) Complete observation accounting. There is one active incumbent evaluation per own slot, one initial snapshot per discovered family, one snapshot after each accepted own update, and one diagnostic-reference evaluation per family. Thus (99) gives at most KJ, $K(J+1)$, and K calls of these respective kinds. Corollary 11 bounds a complete active or snapshot call by $\widetilde{\mathcal{O}}(E_Q\eta_Q^{-2})$ or $\widetilde{\mathcal{O}}(E_V\eta_V^{-2})$, with actual geometry, confidence amplification, validation, and internal trajectory spacing retained. A split retry reuses its active critic; its singleton solve needs no protected fit.

Each family collects $(h-1)J$ complete preservation blocks once, so $K(h-1)J$ blocks are charged even when some are never opened. Both endpoint labels use the same retained transitions. Theorem 18 and the uniform solve and audit bounds, combined in Corollary 25, give per collected block

$$\widetilde{\mathcal{O}}\Big(p_{\log}^{-1}Q_{\max}^2(\mathfrak{C}_{\mathrm{reg}}+1)\tau^{-2}\Big).$$

Multiplying these three per-call costs by the counts above yields the three control and preservation terms in (112).

For identification, the K references cost $\widetilde{\mathcal{O}}(KE_{\mathrm{id}}A_{\mathrm{id}}^2\Delta_0^{-2})$. Each of the S routing calls uses one common labeled block and one common monitor for all stored references, costing $\widetilde{\mathcal{O}}(Q_{\max}^2(\mathfrak{C}_{\mathrm{id}}+1)\Delta_0^{-2})$. The reference count contributes confidence logarithms and computation, not an extra factor K in these retained transitions. This proves the diagnostic term (113).

Finally add only otherwise uncharged acquisition work to N_{acq}. For an outer trajectory collection, its N retained tuples are already in the appropriate term above, and (15) adds its entry, spacing, and exit cost $(N+1)g$ exactly once. Evaluator collection and coupling are already inside its complete-call cost and failure allocation. These counts prove the full ledger. Declared record, module, block, and call caps bound the same prescribed work on failure histories as well. This completes all five conclusions on $\mathcal{E}$. □

C.2 Proof of Corollary 2

Proof. Calibration G prescribes

$$\alpha=\frac{c\varepsilon}{8Q_{\max}},\qquad \tau=\frac{c^2\lambda_{\min}\varepsilon}{10H_M},\qquad \eta_Q=\frac{c^2\varepsilon}{12H_Q^{\mathrm{SA}}},\qquad \eta_V=\frac{\tau}{4A_V}.$$

With $c, Q_{\max}, \nu_{\min}, \lambda_{\min}$ fixed, its quota has $J=O(\log(1/\varepsilon))$. Substitution in the three terms of (112), absorbing fixed discount and reward factors and the logarithm in J, gives

$$\widetilde{\mathcal{O}}\left(\frac{KE_Q(H_Q^{\mathrm{SA}})^2}{\lambda_{\min}\varepsilon^2}+\frac{KE_VA_V^2H_M^2}{\lambda_{\min}^3\varepsilon^2}+\frac{K(h-1)(\mathfrak{C}_{\mathrm{reg}}+1)H_M^2}{p_{\log}\lambda_{\min}^3\varepsilon^2}\right).$$

The uniform evaluator and auxiliary-complexity conditions turn these into $\widetilde{\mathcal{O}}(K\varepsilon^{-2})$ and $\widetilde{\mathcal{O}}(K(h-1)p_{\log}^{-1}\varepsilon^{-2})$. Equation (113) contributes $\widetilde{\mathcal{O}}((S+K)\Delta_0^{-2})$, and the assumed acquisition inflation is absorbed in the same orders. Their sum is (10). For $p_{\log}=\Theta(\varepsilon)$ the preservation term is $\widetilde{\mathcal{O}}(K(h-1)\varepsilon^{-3})$; for a fixed positive floor it combines with the evaluator term to give $\widetilde{\mathcal{O}}(Kh\varepsilon^{-2})$. The substitutions require the evaluator's local exact tangent-realization and conditioning assumptions, the calibrated auxiliary approximation bounds $a\le u/3$, and the numerical budgets to hold uniformly at the selected widths and radii as ε and Δ_0 vary. Finiteness at one fixed tolerance or a logger's action floor alone does not establish these uniform premises. □

C.3 Proof of Theorem 3

Proof. Calibration S changes only the active root tolerance to (118). Since $\eta_Q^{\mathrm{S}}\le\eta_Q^{\mathrm{G}}$, the proof of Theorem 1 applies with the complete evaluation budget recomputed at the tighter tolerance. In particular, the general accuracy and progress conclusions require no compatibility. The same failure allocation gives probability at least $1-\delta$ for each execution separately.

(a) One compatible group. On this event, Theorem 38 and Calibration S give, for every protected family,

$$R_i^+(1)\le A_{V,i}\eta_{V,i}+C_{\mathrm{sh}}\eta_Q^{\mathrm{S}}/\sqrt{p_{\min}}\le\tau/4+\tau/8=3\tau/8.$$

Since $e_i, e_{a,i} \le \tau/8$, the solve upper bound at the full endpoint is at most $3\tau/8+2e_i \le 5\tau/8 < 3\tau/4$, and its mandatory audit upper bound is at most $3\tau/8 + 2e_{a,i} \le 5\tau/8 < \tau$. Thus the endpoint-first rule returns $\lambda = 1$ whenever its block is available.

Before a hypothetical first split, join-before-create puts every new record in the first module, whose capacity is $h = K$. A protected family is tested at most $(K-1)J$ times, once for each possible own update of the other families. Before each required test, fewer than this many blocks have been consumed, so its single initial reserve still contains a complete block. This is the inventory induction of Corollary 40; anchor refreshes do not reset the reserve. Neither reserve shortage nor failed movement or audit can cause the first split. Therefore every own update takes its full step, all records use one deployed module, and no split occurs. Accuracy follows from the tighter-tolerance argument above, with the ledger evaluated at η_Q^{S}.

(b) Compatible and conflicting instances. Use the continuous-state families (123) with parameters (126). They have the common supplied reward bound $R_{\max} = 9/8$. Proposition 15 supplies their environmental coverage, mixing, and logger bounds, and Proposition 16 gives common supplied diagnostic separation $\Delta_0 = 1/4$ and $A_{\mathrm{id}} = 1 + \gamma$. For the compatible pair it also proves the uniform alignment and margin conditions with $a_{\min} = a_*/(1 + \gamma/\pi)$, $\kappa_{\mathrm{tr}} = 2$, and $a_* = (1 + \gamma)/4$. Use these same supplied constants and budgets on the conflicting instance; no compatibility claim is needed there for accuracy.

For every common measurable stationary policy on the conflicting pair, Proposition 17 proves

$$\max_{k=1,2}\{J_k(\pi_k^*) - J_k(\pi)\} \ge L_{\mathrm{conf}} := \frac{a_*(c + \gamma/2)}{2c} > 0.$$

Retain the stated neural, sampling, and active-time assumptions on each instance at the common prescribed budgets. When $\varepsilon < L_{\mathrm{conf}}$ is admissible, the general accuracy conclusion under Calibration S gives gap at most $7\varepsilon/8$ on both families of each instance. Part (a) gives one module and no split for the compatible pair. On the conflicting pair a single final module would supply one common stationary policy and contradict the displayed lower bound. There must therefore be two modules. The general cap is two modules and one split; join-before-create initially puts both records in the same module, so the second module must arise from exactly one split. These are per-instance probability statements. A simultaneous claim for two separate runs would instead allocate $\delta/2$ to each. The construction verifies the environmental properties. The evaluator's local exact tangent realization and conditioning, together with the calibrated auxiliary approximation bounds $a \le u/3$, remain the stated execution premises at the prescribed widths, radii, and tolerances. □

D SUPPORTING THEORETICAL RESULTS

This section supplies the specifications and intermediate results used in Appendix C. Its first eight subsections develop the sampling, learning, certification, transport, and module-management tools; Appendix D.9 collects numerical choices and complete observation accounting, and Appendix D.10 gives concrete continuous-state constructions.

D.1 EXECUTION, OBSERVATION, COVERAGE, AND REPRESENTATION SPECIFICATIONS

D.1.1 OBSERVATION AND DISCLOSURE PROTOCOL

Definition 2 (Observation protocol)**.** *Only the active family supplies new transitions, with its identity used only as an analytical index. Use actual conditional independent access under the specified input laws, or frozen-policy trajectory collectors with the charged entry, spacing, and exit intervals below. A logger $\beta_i(a \mid s) \ge p_{\log} > 0$ has recorded probabilities and invariant state law μ_i; the logging floor is separate from $p_{\min}$. Freeze policies before evaluator initialization; freeze auxiliary targets, classes, and budgets before label disclosure, and fitted functions before monitor disclosure. Fit, solve-monitor, and audit portions are disjoint. Two endpoint labels may share a fit transition, and all diagnostic references may share a current-family fit block. Exposed blocks are never reused as fresh data.*

Each family collects one reserve of $(h-1)J$ complete preservation blocks while active; unused inventory is charged. Schedules can use visible history and independent external randomness, but

not hidden reserves. Their later evolution uses collection data only through the exposed exit state, deliberately disclosed tuples, and independent future randomness. All caps and the ordered active-time requirement of Assumption 2(c) apply, including on failure histories. No access to an inactive family is presumed.

An analytical pre-call history additionally fixes the current true family and already used exogenous schedule randomness, but excludes hidden reserves and future adaptive schedule outcomes. Conditional assertions do not reveal this family to the learner. More explicitly, for an input law η on state–action pairs define the one-transition law

$$\mathsf{L}_k(\eta)(ds, da, dy, ds') = \eta(ds, da)\, \delta_{r_k(s,a)}(dy)\, P_k(ds' \mid s, a).$$

Evaluator tuples use $\eta = \rho_v \otimes \pi_v$ and append $a' \sim \pi_v(\cdot \mid s')$; preservation tuples use $\eta = \mu_i \otimes \beta_i$; diagnostic tuples use $\eta = \mu_k^{\mathrm{id}}$. Monitor inputs have their prescribed marginal laws and no outcome labels. For each disjoint block B of N_B tuples, the independent-access alternative requires

$$\mathcal{L}(B \mid \mathcal{H}_B) = \mathsf{L}_B^{\otimes N_B},$$

where $\mathcal{H}_B$ fixes the frozen choices and L_B is the corresponding law just defined (or its monitor marginal). This is a conditional sampling requirement, including blocks disclosed from a reserve; it is not an assertion of independence between two endpoint labels formed from the same tuple. Under trajectory access, the equality instead holds in the coupled ideal experiment below. Each frozen collector has invariant law ϱ and transition kernel M satisfying $\sup_x \|M^n(x, \cdot) - \varrho\|_{\mathrm{TV}} \le \kappa_{\mathrm{mix}} \beta^n$ for prescribed $\kappa_{\mathrm{mix}} < \infty$, $0 < \beta < 1$, and every $n \ge 0$. For N retained tuples under this mixing envelope, prescribe a gap g satisfying

$$(N+1)\kappa_{\mathrm{mix}} \beta^g \le \delta_{\mathrm{mix}}, \qquad N_{\mathrm{collected}} = N + (N+1)g. \tag{15}$$

Proposition 1 justifies this coupling, including the exit state. Evaluator collection has its own complete-call allocation and explicit formula (38); its gaps are already included in (39).

D.1.2 Coverage and transfer constants

For probability laws ζ, ϱ, define

$$\mathsf{h}(\zeta \mid \varrho) = \begin{cases} \|d\zeta/d\varrho\|_{L_2(\varrho)}, & \zeta \ll \varrho, \\ +\infty, & \text{otherwise}, \end{cases} \qquad d^{\pi}_{k,\zeta} = c \sum_{t \ge 0} \gamma^t \zeta (P_k^{\pi})^t.$$

Here P_k^{π} is the state transition kernel under π and $(\zeta \otimes \pi)(ds, da) = \zeta(ds)\pi(da \mid s)$. Let $\pi^*_{k,\alpha}$ be an optimal policy in Π_α and set $\xi_{k,n} = \xi_k (P_k^{\pi^*_{k,\alpha}})^n$. At own-update stage j, write $\pi^-_{k,j}$ for the current policy, $\pi^{\mathrm{g}}_{k,j}$ for the exploratory greedy policy, and $\pi^+_{k,j} = (1-\lambda_j)\pi^-_{k,j} + \lambda_j \pi^{\mathrm{g}}_{k,j}$ for the installed mixture. Its critic is evaluated under the stationary law $\rho_{k,j} \otimes \pi^-_{k,j}$.

The coverage clause of Assumption 2 means that deterministic envelopes $H_Q^{\mathrm{SA}}, H_M < \infty$ satisfy, for every k and every permitted history,

$$c \sum_{n \ge 0} \gamma^n \sup_j \max_{p \in \{\pi^-_{k,j}, \pi^{\mathrm{g}}_{k,j}, \pi^*_{k,\alpha}\}} \mathsf{h}\left(d^{\pi^+_{k,j}}_{k,\xi_{k,n}} \otimes p \,\middle|\, \rho_{k,j} \otimes \pi^-_{k,j} \right) \le H_Q^{\mathrm{SA}}, \tag{16}$$

$$c \sum_{n \ge 0} \gamma^n \sup_u \mathsf{h}\left(d^{\vartheta_{k,u}}_{k,\xi_{k,n}} \,\middle|\, \mu_k \right) \le H_M. \tag{17}$$

The policies $\vartheta_{k,u}$ range over all protected anchor-to-candidate comparisons required by the transport proof, including the final reporting suffix. The fixed law μ_k is that family's preservation-monitor law. Use $H_M = 1$ when no protected comparison is needed. Deterministic bounds are used to prescribe budgets before observations are exposed; finite realized coefficients alone yield only the achieved-error statement.

An anchor policy $\bar{\pi}_i$ is the initial certified policy or the latest accepted own-update policy. Its snapshot $\widehat{V}_i$ is evaluated under υ_i and remains fixed while the family is inactive. A stored diagnostic reference q_ℓ is evaluated under ν_ℓ^{id}. The remaining coverage requirements are the measure inequalities

$$\mu_i \le C_{0,i} \upsilon_i, \qquad \mu_i P_i^{\vartheta} \le C_{1,i} \upsilon_i, \qquad \text{for every tested candidate } \vartheta, \tag{18}$$

$$\mu_k^{\mathrm{id}} \le D_{0,k\ell}\nu_\ell^{\mathrm{id}}, \qquad \mu_k^{\mathrm{id}} P_k^{\mathrm{id}} \le D_{1,k\ell}\nu_\ell^{\mathrm{id}}, \qquad \text{for all } k,\ell. \tag{19}$$

In the second line, P_k^{id} is the state–action transition kernel that uses P_k followed by π_{id}. All constants have finite prescribed envelopes over histories. Define

$$A_V = \sup_i(\sqrt{C_{0,i}} + \gamma\sqrt{C_{1,i}}), \qquad A_{\mathrm{id}} = \sup_{k,\ell}(\sqrt{D_{0,k\ell}} + \gamma\sqrt{D_{1,k\ell}}). \tag{20}$$

Jensen's inequality derives the snapshot and diagnostic error-transfer bounds from these dominations. They are not additional assumptions.

D.1.3 Evaluator realization and auxiliary target approximation

For evaluation call v, the analytical history $\mathcal{H}_v$ fixes the policy π_v and true family k_v without revealing its identity to the learner. The stationary joint law is $\varrho_v = \rho_v \otimes \pi_v$. Define

$$\psi_v = \nabla f_{W_v^0}, \quad \mathcal{F}_v = \{f_{W_v^0} + \langle \psi_v, u\rangle : \|u\|_2 \le R_v\}, \quad \Sigma_v = \mathbb{E}_{\varrho_v}[\psi_v\psi_v^\top].$$

Let $\mathfrak{H}_v$ be the set of permitted analytical pre-call histories. The quantifiers in Assumption 3(a) are: for every $H \in \mathfrak{H}_v$ and every width/radius pair (m_v, R_v) selected by the calibration at H, the prescribed fresh initialization law, independent of collection conditional on H, admits an event $\mathcal{E}_v(H)$ such that

$$\mathbb{P}(\mathcal{E}_v(H) \mid \mathcal{H}_v = H) \ge 15/16,$$

and every initialization on this same event is admissible and satisfies

$$\exists u_v \in \mathbb{R}^{\dim W_v}: \quad \|u_v\|_2 \le R_v, \qquad \|Q_{k_v}^{\pi_v} - f_{W_v^0} - \langle \psi_v, u_v\rangle\|_{\varrho_v} = 0,$$
$$\forall z \in \mathrm{range}(\Sigma_v): \quad z^\top \Sigma_v z \ge \lambda_{\mathrm{cov},v}\|z\|_2^2, \qquad \lambda_{\mathrm{cov},v} > 0.$$

Admissibility is exactly the input, activation, and initial-output specification of Appendix D.3.1. The supplied positive covariance bound, derivative envelopes, and radius retain their actual call/width dependence; positive definiteness outside the covariance range is not required. The history quantifier includes policies reached after sharing or splitting and permitted failure histories. Each replica has fresh initialization randomness. This event is amplified inside the complete evaluator, rather than charged as constant failure in the continual theorem.

For each reached auxiliary call r, prescribe a root-error tolerance $u_r > 0$ and choose before label disclosure

$$\Theta_r = \{\theta_r^0 + v : \|v\|_2 \le R_r\}, \qquad \mathcal{G}_r = \{f_\theta : \theta \in \Theta_r\},$$

and let ϱ_r be its input law. Define the actual conditional target $B_r(u) = \mathbb{E}[Z_r \mid U_r = u, \mathcal{H}_r]$. The visible pre-call history selects the class and tolerance without the true current-family label. For every permitted analytical history H and the width and radius selected by the calibration at H, Assumption 3(b) requires

$$a_r := \inf_{\theta\in\Theta_r} \|f_\theta - B_r\|_{\varrho_r} \le u_r/3.$$

This approximation bound must hold in the same nonlinear class, at the same width and radius, used for fitting and for the statistical and optimization budgets. Each diagnostic class must satisfy it for every possible current-family comparison target under the corresponding input law. Require the stated envelopes on the whole ball: $|f_\theta| \le Q_{\max}$, $\|\nabla f_\theta\|_2 \le G_r$, and $\|\nabla^2 f_\theta\|_{\mathrm{op}} \le \chi_r$. These are architecture requirements whose selected values must meet the calibration below. All approximation bounds below are uniform over permitted learned snapshots, references, endpoint policies, and pre-call histories.

1. For preservation of family i, input $U = s$ has law μ_i. With logging policy $\beta_i(a \mid s) \ge p_{\log} > 0$, an endpoint π^e, $e \in \{0,1\}$, has label

$$Z_e = \frac{\pi^e(a \mid s)}{\beta_i(a \mid s)}[r_i(s,a) + \gamma\widehat{V}_i(s')], \qquad B_e = T_i^{\pi^e}\widehat{V}_i,$$

where $(T_i^\pi V)(s) = \sum_a \pi(a \mid s)[r_i(s,a) + \gamma P_i(\cdot \mid s,a)V]$. For the actual target B_e, formed from the learned clipped snapshot, use $u_r = \tau_i/16$, where τ_i is the prescribed preservation threshold for family i ($\tau_i = \tau$ in the common calibration), and require

$$\inf_{f\in\mathcal{G}_r} \|f - B_e\|_{\mu_i} \le \tau_i/48.$$

2. For comparison with stored record ℓ while family k is active, input $U = (s, a)$ has law μ_k^{id} and
$$Z = r_k(s, a) + \gamma \sum_{a'} \pi_{\mathrm{id}}(a' \mid s') q_\ell(s', a'), \qquad B = \mathcal{T}_k^{\mathrm{id}} q_\ell.$$
Use $u_r = \Delta_0/16$. The chosen class for reference ℓ must satisfy
$$\inf_{f \in \mathcal{G}_r} \|f - \mathcal{T}_k^{\mathrm{id}} q_\ell\|_{\mu_k^{\mathrm{id}}} \le \Delta_0/48$$
for every possible current family k, including wrong-family comparisons. The class-selection rule cannot use k; the same chosen class must meet the bound under each corresponding μ_k^{id}.

For preservation labels take $\Lambda = p_{\log}^{-1}$; for diagnostic labels take $\Lambda = 1$. With $u = u_r$, $a = a_r$, and the corresponding class parameters, the remaining requirements in (53) yield $\|\widehat{g}_r - B_r\|_{\varrho_r} \le u_r$ by Theorem 18, at the allocated regression failure probability. In particular, the approximation contribution $3a_r^2 \le u_r^2/3$ is included in the fitted-error budget. The total fitted-error targets are therefore $\tau_i/16$ for preservation and $\Delta_0/16$ for routing, including approximation, statistical, curvature/ridge, and optimization errors. The tangent classes in Appendix D.4.2 control statistical complexity, with their dependence on the selected width retained. Auxiliary approximation is imposed on the nonlinear trained class itself. Exact auxiliary realization is the special case $a_r = 0$; the evaluator continues to satisfy the local tangent-realization condition of Assumption 3(a) stated above.

D.1.4 Visible history, resources, and acquisition

Fix the K record and module slots, module capacity $h \le K$, own quota J, segment count S, movement threshold $\lambda_{\min}$, logger choices, tolerance envelopes, and finite subroutine budgets. A permitted history records the observed segment, existing records, g_t, occupied modules, installed policies, named anchors and their target identities, exposed data, disclosed-block counters, and completed own slots. It excludes every unexposed reserve tuple and the collection gaps. All choices and failure allocations are measurable with respect to the relevant visible history before their data are revealed. The schedule may depend on visible history, but cannot inspect hidden reserve values.

A new record joins the first existing module with fewer than h assigned records, in creation order, leaving its installed policy unchanged; if none has space, it activates an unused slot. Existing records can move only by splitting the currently active record off a nonsingleton. Every occupied module has at least one assigned record. Thus, before a split, the number of occupied modules is strictly less than the number of records and an unused slot among the K is available. Never merge existing modules. Enforce all caps before collecting or fitting further data, including on statistically invalid histories.

For family i, freeze a logger β_i with state law μ_i and known action probabilities at least $p_{\log}$. At discovery, collect one inventory of $B_{\mathrm{res}} = (h-1)J$ complete blocks, then keep this same finite inventory through every later anchor refresh. A fast-backend block contains n_i labeled tuples shared by its two endpoint regressions, m_i solve-monitor inputs, and $m_{a,i}$ audit-monitor inputs. These portions are disjoint and disclosed in this order. The actual prescribed worst-case budgets over permitted histories determine block sizes. For $h = 1$, there are no preservation blocks.

Independent acquisition gives these input/label blocks their product laws directly. For trajectory acquisition, if a whole family reserve requires N_i retained tuples (including tuples retained only for their monitor inputs), use the entry, inter-tuple, and exit construction of the replay-isolation proposition. With
$$\sup_s \mathrm{TV}(P_{\beta_i}^{b_i}(s, \cdot), \mu_i) \le \kappa_{\mathrm{mix},i} \rho_i^{b_i}, \qquad (N_i + 1)\kappa_{\mathrm{mix},i} \rho_i^{b_i} \le \delta_{\mathrm{mix},i},$$
collection costs $N_i + (N_i + 1)b_i$ transitions. Expose only the exit state at collection completion. All subsequent visible evolution uses that exit state, deliberately disclosed tuples, and independent future randomness. The coupled ideal experiment has independent retained tuples and exit state, so remaining blocks stay independent of adaptive visible choices. Charge this coupling error once to the entire collection. No dormant-family query is needed to use the reserve. Independent resets, if assumed instead, must be an actual observation interface.

Route blocks similarly require current-family input acquisition. Each complete evaluator uses the spaced frozen-policy collector in Appendix D.3.6; its training, validation, and all intervening

transitions are charged internally. Preserving labels from a trajectory is not alone a product-law guarantee. Input laws must actually be sampled; knowledge of a density does not provide sample access. Any frozen logger may supply its own stationary monitor law, but all later anchors must satisfy the stated transfer inequalities from that fixed law. The logger's action floor alone supplies none of these state-law properties.

A finite opportunity condition. The observation requirement means that the given S-segment stream contains, for each family, enough active time to finish the first-discovery diagnostic and initial-anchor work, collect its entire reserve, and execute J own slots. Each slot includes the incumbent and post-update evaluations plus all acquisition charged by its prescribed subroutines. Segment-start routing also consumes its stated active budget. For trajectory collection the full $N_i + (N_i + 1)b_i$, rather than N_i retained records alone, must fit an active interval. The stream can supply multiple slots in a segment; no assertion that mere repeated appearance supplies these budgets is made. Formally, let the observed segment intervals be $[b_s, b_{s+1})$, $s = 1, \dots, S$, with analytical family labels k_s. For any permitted execution, list its required acquisition jobs in execution order as $r = 1, \dots, R_{\mathrm{jobs}}$, with required family $k(r)$ and prescribed complete cost c_r^{acq}. A trajectory job includes its entry, inter-tuple, and exit intervals, so $c_r^{\mathrm{acq}} = N_r + (N_r + 1)g_r$. The opportunity condition requires placements t_r and segment indices $s(r)$ satisfying

$$\begin{aligned} t_{r+1} &\geq t_r + c_r^{\mathrm{acq}} \quad (r < R_{\mathrm{jobs}}), \\ b_{s(r)} \leq t_r < t_r + c_r^{\mathrm{acq}} &\leq b_{s(r)+1}, \qquad k_{s(r)} = k(r) \quad (1 \leq r \leq R_{\mathrm{jobs}}). \end{aligned}$$

Placements respect segment-start routing and discovery-before-use. Monitor disclosure, rather than its acquisition, follows fitting. The jobs include exactly $(h-1)J$ complete reserve blocks for each discovered family and at least J available own slots for every $k \in \{1, \dots, K\}$, with all prescribed diagnostic and anchor-evaluation jobs. Zero-cost jobs can be omitted from this list. The requirement holds for every permitted history under the declared worst-case budgets, rather than only for successful statistical outcomes. If a required interval ends before its protocol can finish, the capped procedure reports that the accuracy theorem's opportunity condition was not met.

D.2 Conditional sampling, concentration, and sealed replay

D.2.1 Conditional probability and sample acquisition

All certificates below refer to the actual, adaptive execution. Let $\mathcal{H}_j$ contain everything observed before statistical call j, including the time-dependent module assignment, installed policies, stored references, previous tests, and the next call's prescribed budget. A call may depend on $\mathcal{H}_j$, but its new observations are drawn afterward. In particular, the policy stays fixed throughout a critic evaluation; endpoint policies and references stay fixed throughout a preservation solve; and the selected candidate is sealed before its independent audit. An unsuccessful audit is charged to the execution and is not reused to certify a modified candidate.

For probability analysis, enlarge $\mathcal{H}_j$ by the current true family and exogenous schedule randomness already used; these are not learner observations. Conditional bounds are uniform over these values. For a fixed family schedule one can instead condition on that schedule throughout. With an adaptive schedule, no future schedule outcomes depending on later disclosures are included. Unexposed reserve tuples remain excluded. The model and exposed fitted functions determine prior certificate-validity events in this analytical filtration. Collection coupling is handled separately in the ideal product experiment.

Lemma 4 (Adaptive certificate accounting)**.** *Suppose each reached call has a nonnegative $\mathcal{H}_j$-measurable failure allocation δ_j, with $\sum_j \delta_j \leq \delta$ on every execution. Conditional on every history on which preceding required certificates hold, call j fails its stated certificate with probability at most δ_j. Then all reached certificates hold jointly with probability at least $1 - \delta$. This conclusion includes an adaptive number of calls.*

Proof. Let A_j be the event that call j is reached with all preceding required certificates valid, and that this call is the first failure. The event describing its valid pre-call history is $\mathcal{H}_j$-measurable. Conditional expectation gives $\mathbb{P}(A_j) \leq \mathbb{E}[\mathbf{1}\{\text{valid history at } j\}\delta_j]$. The first-failure events are disjoint, so their total probability is at most $\mathbb{E}\sum_j \delta_j \leq \delta$. An unreached call contributes zero. □

Appendix D.9.3 gives the concrete five-pool allocation, covering initialization, training/selection, regression, outer monitors, additional primitive-envelope events, and acquisition coupling. Internal replica and selection events are included in their parent call allocation. Caps apply on failure histories as well; a confidence allocation does not bound unlimited retries.

Lemma 5 (Stationary starts and independent retained records)**.** *Suppose a frozen-policy state kernel P_π has invariant law ρ_π and*

$$\sup_s \mathrm{TV}(P_\pi^t(s,\cdot),\rho_\pi) \le \kappa_{\mathrm{mix}}\beta^t, \qquad 0<\beta<1. \tag{21}$$

For at most F_{traj} restarted training trajectories, a burn-in

$$b_{\mathrm{train}} = \left\lceil \frac{\log \max\{1, \kappa_{\mathrm{mix}} F_{\mathrm{traj}}/\delta_b\}}{-\log\beta} \right\rceil \tag{22}$$

allows all subsequent training paths to be coupled to stationary-start paths outside probability δ_b. For a fresh validation block with n retained states and at most F_{val} such blocks, use spacing

$$b_{\mathrm{val}} = \max\left\{1, \left\lceil \frac{\log \max\{1, \kappa_{\mathrm{mix}} n F_{\mathrm{val}}/\delta_v\}}{-\log\beta} \right\rceil \right\}. \tag{23}$$

Retain the states at times $b_{\mathrm{val}}, 2b_{\mathrm{val}}, \ldots, nb_{\mathrm{val}}$, so the first retained state also receives the required mixing steps. The retained states can be coupled to independent ρ_π draws, jointly over all blocks, outside probability δ_v. The costs are $F_{\mathrm{traj}} b_{\mathrm{train}}$ and $\sum_{\mathrm{blocks}} nb_{\mathrm{val}}$ transitions. Budgets depending on the call can instead use their allocated per-call coupling failures in these formulas.

Proof. After burn-in, maximal coupling to ρ_π fails with probability at most $\kappa_{\mathrm{mix}}\beta^{b_{\mathrm{train}}} \le \delta_b/F_{\mathrm{traj}}$. Once coupled, use the same subsequent transition randomness, preserving the entire trajectory. For retained records, conditional on any preceding retained history, the next state has total-variation distance at most $\kappa_{\mathrm{mix}}\beta^{b_{\mathrm{val}}}$ from ρ_π. Sequential maximal couplings therefore fail on at most $n\kappa_{\mathrm{mix}}\beta^{b_{\mathrm{val}}}$ of the probability mass of one block. Union bounds, or Lemma 4 for adaptive blocks, give the claims. These arguments compare distributions; they do not assert that discarding finitely many transitions produces exact stationarity. □

At each retained validation state draw an auxiliary action from the frozen policy, independently of the action used to continue the state chain. The validation law is then $\rho_\pi(ds)\pi(da \mid s)$. If actual trajectory actions are retained, add one transition to the spacing before applying the state-chain bound. The present TD proof uses the full-tuple spacing construction in Appendix D.3.6 for training as well as validation. Stationary initialization alone does not make consecutive TD observations independent.

The analogous correction for diagnostic or monitoring records requires a sampler for that particular input law. The active control chain's mixing bound cannot certify an unrelated monitor law. Alternatively, if an independent input draw costs c_i^{state} observations, charge that cost for each retained input. One current-family diagnostic tuple provides one label for each frozen reference; one logged preservation tuple provides both endpoint labels through their known importance ratios. Certification and optional audit monitors require input acquisition but no new outcome labels once the endpoint regressors are frozen. For retained trajectory reserves, the stronger entry, inter-record, and exit isolation statement in Proposition 1 applies; this validation-thinning lemma alone does not justify later adaptive disclosure of hidden outcomes.

With L attempted control transactions, R_{snap} stored-policy evaluations, and K diagnostic-reference evaluations, logarithmic confidence replication uses

$$F_{\mathrm{traj}} = O\left((L + R_{\mathrm{snap}} + K)\log\frac{L + R_{\mathrm{snap}} + K}{\delta}\right).$$

Uniform mixing makes the acquisition corrections logarithmic in accuracy; otherwise their dependence is retained explicitly. Every acquisition cost is either included in the relevant evaluation/record cost or displayed in $N_{\mathrm{acquisition}}$, once. Neural forward passes, the scalar quadratic solve, and gradient operations consume computation rather than additional environment observations.

D.2.2 Conditional observation laws and one-transition labels

Let the state space be standard Borel, the action set finite, $0 < \gamma < 1$, $c = 1 - \gamma$, and $|R| \le R_{\max}$. Set $Q_{\max} = R_{\max}/c > 0$. For family i, define

$$(T_i^\pi V)(s) = \mathbb{E}_i[R + \gamma V(S') \mid S = s,\ A \sim \pi(\cdot \mid s)].$$

All snapshots and regressors are measurable and bounded in $[-Q_{\max}, Q_{\max}]$. Thus $|T_i^\pi V| \le Q_{\max}$. A candidate policy is a distribution on actions, not an interpolation of neural parameters.

Definition 3 (Conditional regression block)**.** *Condition on a sigma-field $\mathcal{F}$ fixing the target, input law μ, regression class, budget, and confidence allocation. The n observed input–label pairs (U_t, Z_t) are conditionally independent and identically distributed, with $U_t \sim \mu$ and*

$$m(u) = \mathbb{E}[Z_t \mid U_t = u, \mathcal{F}], \qquad |m| \le Q_{\max}, \qquad |Z_t| \le \Lambda Q_{\max}, \qquad \mathbb{E}[Z_t^2 \mid U_t, \mathcal{F}] \le \Lambda Q_{\max}^2, \tag{24}$$

where $\Lambda \ge 1$ is a specified envelope. Monitor blocks are fresh conditional on the fitted functions. Any discrepancy from this product law must be bounded and charged through a coupling or total-variation argument.

Lemma 6 (One-transition replay labels)**.** *Fix family i, V, and π. A logged record is (S, A, R, S', β) with state law μ_i, conditional action law $\beta(\cdot \mid S)$, and the correct family outcome kernel given (S, A). Suppose the logging probabilities are known and $\beta(a \mid s) \ge p_{\log} > 0$. Set*

$$w = \frac{\pi(A \mid S)}{\beta(A \mid S)}, \qquad Y = R + \gamma V(S'), \qquad Z = wY.$$

Then (24) *holds with $U = S$, $m = T_i^\pi V$, and $\Lambda = p_{\log}^{-1}$. For an outcome generated directly under π, the unweighted label $Z = Y$ instead satisfies* (24) *with $\Lambda = 1$.*

Proof. Since $|Y| \le R_{\max} + \gamma Q_{\max} = Q_{\max}$, the range bound follows from $w \le 1/p_{\log}$. Conditional expectation over the logged action gives

$$\mathbb{E}[Z \mid S = s] = \sum_a \pi(a \mid s)\mathbb{E}_i[R + \gamma V(S') \mid s, a] = (T_i^\pi V)(s).$$

Furthermore,

$$\mathbb{E}[Z^2 \mid S = s] \le Q_{\max}^2 \sum_a \frac{\pi(a \mid s)^2}{\beta(a \mid s)} \le \frac{Q_{\max}^2}{p_{\log}} \sum_a \pi(a \mid s)^2 \le \frac{Q_{\max}^2}{p_{\log}}.$$

The same identities can be conditioned on a known logger identity in a mixture of loggers. Independence and the input law still have to be established for the resulting records. Direct sampling has $|Z| \le Q_{\max}$. □

The variance bound includes variability from the logged action, even when rewards are deterministic. The action floor supplies action overlap; it does not establish state coverage or the snapshot-transfer inequalities. Neither clipping nor self-normalizing the importance weights is needed for the following theorem.

D.2.3 When retained replay is statistically valid

Conditional fresh blocks and global confidence Lemma 4 supplies a single simultaneous event for all adaptively reached calls. A chosen budget is fixed before the relevant block is disclosed; an initial history-dependent target may be otherwise arbitrary under the conditional product interface.

The primary replay protocol is a retained inventory whose unused blocks, conditional on the visible history, still have the product law in Definition 3. Freeze the current target before exposing the next labeled block; freeze its fits before exposing its monitor block. Previously exposed blocks are not treated as fresh. The inventory must have been acquired while the family was active, and exhaustion invokes the immediate singleton-splitting branch of Algorithm 1; it never turns exposed observations into new data. All blocks, including those spent on failures, are charged when acquired.

Sealing alone does not establish this interface. A later visible state from the same collection trajectory can reveal information about an undisclosed earlier transition. Independent collection episodes with independent resets are one sufficient implementation. A precise mixing-based isolation alternative follows.

Proposition 1 (Collection and exit isolation under a frozen logger)**.** *While family i is active, freeze a logger β_i with invariant state law μ_i and*

$$\sup_s \mathrm{TV}(P^b_{\beta_i}(s,\cdot),\mu_i) \le \kappa_{\mathrm{mix},i}\rho_i^b, \qquad 0<\rho_i<1.$$

Starting at any entry state, run b logger transitions before the first retained input. Retain its actual tuple (S,A,R,S'), then run b further logger transitions from that next state before each subsequent retained input. After the Nth tuple, run a final b transitions and expose only the exit state. Keep all collection and gap observations hidden except when a retained tuple is later deliberately revealed. Then the joint law of the N retained tuples and exit state is within

$$(N+1)\kappa_{\mathrm{mix},i}\rho_i^b \tag{25}$$

in total variation of N independent logged tuples with state law μ_i and an independent exit state with law μ_i. Collection costs $N+(N+1)b$ transitions.

Proof. After the entry gap, the first retained state is within $\kappa_{\mathrm{mix},i}\rho_i^b$ of μ_i, uniformly over the entry state. Applying the same action–outcome kernel to coupled states preserves equality of their entire retained tuples. Conditional on all previous retained tuples, the next retained input has law $P^b_{\beta_i}(S'_{\mathrm{previous}},\cdot)$, again uniformly close to μ_i. Couple it to a fresh independent μ_i draw and use common action–outcome randomness. Repeat for N tuples and for the final exit state. The union bound gives (25). There are N retained transitions and $N+1$ gaps of length b. □

Assume future visible evolution depends on the hidden collection only through that exit state, revealed tuples, and independent future randomness; in particular, the family schedule does not read hidden reserve values. Applying the same future execution kernel cannot increase total variation. In the ideal product experiment, unused blocks remain independent of the visible history even after earlier blocks are revealed adaptively. Lemma 4 therefore applies there. The actual failure probability is at most its bound plus (25). Allocate this coupling error once to the whole collection, not again at every future reveal.

For finitely many adaptive collections, condition on each entry history, allocate coupling errors $\epsilon_{\mathrm{mix},j}$ with a deterministic summable envelope, and sum them. Choose b so that the displayed bound is at most its collection allocation. The proof does not require dormant-family queries. It does require sufficient active time to collect the inventory and the stated isolation discipline.

D.3 Continuous-state neural evaluation and confidence conversion

We prove the evaluator bound directly. The class has one hidden layer, fixed readout signs, and trainable hidden weights. The proof uses spaced frozen-policy tuples and does not import a convergence theorem for a different normalization or consecutive-transition implementation.

D.3.1 Architecture and initialization specification

Use the single-hidden-layer evaluator

$$f_W(X) = \frac{1}{\sqrt{m}}\sum_{i=1}^m b_i\sigma(w_i^\top\phi(X)), \qquad b_i\in\{-1,1\}, \tag{26}$$

where all hidden weights $W=(w_1,\dots,w_m)$ train and the readout signs stay fixed. The input representation $\phi(X)\in\mathbb{R}^d$ is fixed during evaluation and satisfies $\|\phi(X)\|_2\le 1$. The output used in TD is the nonlinear network, not its linearization. Only the final returned estimate is clipped.

Parameter domain and activation. Choose an initialization W^0 and a training radius $R>0$, and define $\mathcal{B}=\{W:\|W-W^0\|_F\le R\}$. Select a twice continuously differentiable activation with prescribed bounds

$$|\sigma'(z)|\le L_1, \qquad |\sigma''(z)|\le L_2 \quad \text{on all preactivations reachable in } \mathcal{B}. \tag{27}$$

The initialization specification also requires the known output envelope

$$\sup_X |f_{W^0}(X)| \le B_0. \tag{28}$$

Call an initialization admissible when it satisfies these network requirements. They remain premises of the evaluator theorem; their placement here makes the chosen network explicit before stating the conditions on the value functions and sampling laws.

An explicit initialization rule. For even width, initialize pairs with identical hidden weights and opposite fixed readout signs:

$$w^0_{2j-1} = w^0_{2j} = v_j, \qquad b_{2j-1} = 1, \quad b_{2j} = -1, \qquad 1 \le j \le m/2.$$

This gives $f_{W^0} = 0$ identically, so $B_0 = 0$. One can choose the pair centers with $\|v_j\|_2 \le W_{\mathrm{init}}$. Then every reachable preactivation lies in $[-W_{\mathrm{init}} - R, W_{\mathrm{init}} + R]$, and the finite constants

$$L_q = \sup_{|z| \le W_{\mathrm{init}} + R} |\sigma^{(q)}(z)|, \qquad q = 1, 2,$$

verify (27) for the selected activation. The pair centers may be deterministic or freshly sampled independently of the training data; after initialization all hidden weights train freely. This construction verifies the network bounds, while realizability and covariance conditioning remain the separate requirements stated below. Other initialization rules are permitted if their admissibility and the joint probability in Assumption 3 are justified.

D.3.2 Evaluation setup, structural assumptions, and TD update

Let $\mathcal{H}$ denote a pre-call history fixing a family and policy π, and condition on that history. The family label is an analytical index, not an input supplied to the learner. Write $X = (s, a)$, let $\mathcal{P}$ be the state–action transition kernel obtained by following this fixed policy, and let μ be an invariant probability law for $\mathcal{P}$. A one-step tuple is (X, Y, X'), where $|Y| \le R_{\max}$ and the conditional reward mean and transition kernel are those of the family. Put

$$c = 1 - \gamma > 0, \qquad Q_{\max} = R_{\max}/c, \qquad Q(X) = \mathbb{E}[Y + \gamma Q(X') \mid X].$$

In the ideal experiment, tuples are independent with $X \sim \mu$ and the correct joint conditional law of (Y, X') given X. In particular X' also has law μ. Section D.3.6 supplies these ideal laws by coupling a permitted trajectory collector; no generative access is silently introduced.

Use the network and parameter ball of Section D.3.1. Write the stacked hidden-weight displacement as $u = W - W^0$ and set $f_u = f_{W^0+u}$. Define

$$f_0 = f_{W^0}, \qquad \psi(X) = \nabla_W f_{W^0}(X), \qquad \ell_u(X) = f_0(X) + \langle \psi(X), u \rangle, \qquad \Sigma = \mathbb{E}_\mu[\psi\psi^\top].$$

Let $\mathcal{S} = \mathrm{range}(\Sigma)$ and let P be its orthogonal projector. Fix a supplied covariance lower bound $\lambda_{\mathrm{cov}} > 0$.

Assumption 3(a) places exact tangent realization and conditioning on the same admissible initialization event of conditional probability at least $15/16$. Condition on this event for the next two lemmas and the TD theorem. Its probability is amplified inside the complete evaluator; it is not a constant failure of the continual theorem. The specified bounds $B_0, L_1, L_2, R, \lambda_{\mathrm{cov}}^{-1}$ retain their actual call and width dependence.

TD update and parameter choices. Set

$$G = L_1, \quad B = B_0 + GR, \quad D = R_{\max} + (1+\gamma)B, \quad U = DG, \quad a = c\lambda_{\mathrm{cov}}. \tag{29}$$

For $t = 0, \dots, T-1$, start with $u_0 = 0$ and use

$$\begin{aligned} \alpha_t &= \frac{1}{a(t+2)}, \\ d_t(u) &= f_u(X_t) - Y_t - \gamma f_u(X'_t), \\ u_{t+1} &= \mathrm{Proj}_{\|u\|_2 \le R}\big(u_t - \alpha_t d_t(u_t) \nabla f_{u_t}(X_t)\big). \end{aligned} \tag{30}$$

The projection, gradient, and step size are implementable without knowing $\mathcal{S}$, P, Σ, or u_*. The supplied bound λ_{cov} tunes the step size. The policy remains frozen during collection and training.

D.3.3 Derivative bounds and the linearized comparison

Lemma 7 (Uniform local estimates). *For an admissible network from Section D.3.1, throughout the parameter ball,*

$$\|\nabla f_u(X)\|_2 \le G, \qquad \left\|\nabla^2 f_u(X)\right\|_{\mathrm{op}} \le L_2/\sqrt{m},$$
$$|f_u(X) - \ell_u(X)| \le r_m := L_2R^2/(2\sqrt{m}), \quad \|\nabla f_u(X) - \psi(X)\|_2 \le j_m := L_2R/\sqrt{m}. \tag{31}$$

Moreover, $|f_u|, |\ell_u| \le B$. For a tuple $z = (X, Y, X')$, define

$$g(u; z) = [f_u(X) - Y - \gamma f_u(X')]\nabla f_u(X), \qquad \widetilde{g}(u; z) = [\ell_u(X) - Y - \gamma \ell_u(X')]\psi(X).$$

Then

$$\|g(u; z)\|_2 \le U, \qquad \|g(u; z) - \widetilde{g}(u; z)\|_2 \le e_m := A/\sqrt{m}, \quad A = L_2\left(\frac{(1+\gamma)GR^2}{2} + DR\right). \tag{32}$$

Proof. The ith gradient block is $b_i\sigma'(w_i^\top\phi)\phi/\sqrt{m}$; summing its squared norms gives G^2. The Hessian is block diagonal across neurons, with blocks $b_i\sigma''(w_i^\top\phi)\phi\phi^\top/\sqrt{m}$, proving its operator bound. Taylor's theorem and the mean-value theorem prove the next two bounds and the output envelope. Write the nonlinear and linear TD residuals as d and $\widetilde{d}$. Then $|d| \le D$ and $|d - \widetilde{d}| \le (1+\gamma)r_m$, so

$$g - \widetilde{g} = (d - \widetilde{d})\psi + d(\nabla f_u - \psi).$$

The displayed perturbation bound follows. □

Let $\widetilde{u}_*$ be a tangent-realizing displacement with $\|\widetilde{u}_*\|_2 \le R$. Choose the minimum-norm representative $u_* = P\widetilde{u}_*$. It has norm at most R and still represents Q: every vector in $\ker\Sigma$ is orthogonal to $\psi(X)$ for μ-almost every X. Let $P_\perp = I - P$ and $\overline{g}(u) = \mathbb{E}[\widetilde{g}(u; Z)]$.

Lemma 8 (Mean contraction and null-space drift). *For every u in the ball,*

$$\langle u - u_*, \overline{g}(u)\rangle \ge c\,\|\ell_u - Q\|^2_{L_2(\mu)} \ge a\,\|P(u - u_*)\|_2^2. \tag{33}$$

For the actual nonlinear iterates, almost surely in the ideal experiment,

$$\|P_\perp u_t\|_2 \le e_m \sum_{s=0}^{t-1}\alpha_s \le \frac{e_m H_{t+1}}{a}, \qquad H_n = \sum_{j=1}^{n} j^{-1}. \tag{34}$$

Proof. The Bellman identity implies $\overline{g}(u_*) = 0$. For $v = u - u_*$ and $z_v(X) = \langle\psi(X), v\rangle$, stationarity and Cauchy–Schwarz give

$$\langle v, \overline{g}(u)\rangle = \mathbb{E}[z_v(X)^2 - \gamma z_v(X)z_v(X')] \ge (1-\gamma)\mathbb{E}[z_v(X)^2].$$

Apply the covariance lower bound on its range. This argument permits bounded stochastic rewards with the correct conditional mean.

For the second assertion, $\widetilde{g}(u; Z)$ belongs to $\mathcal{S}$ almost surely. Thus $\|P_\perp g(u; Z)\| \le e_m$. Projection onto the ball centered at zero multiplies its input by a scalar in $[0, 1]$. It follows that $\|P_\perp u_{t+1}\| \le \|P_\perp u_t\| + \alpha_t e_m$. Sum from $u_0 = 0$. No unknown subspace is projected onto by the algorithm. □

D.3.4 A direct finite-sample neural TD theorem

Theorem 9 (Nonlinear projected TD with a single normalization). *Use the network specification of Section D.3.1 and condition on the event in Assumption 3. For $T \ge 1$, the independent-tuple algorithm* (30) *satisfies*

$$\mathbb{E}\,\|f_{u_T} - Q\|^2_{L_2(\mu)} \le \frac{2G^2U^2}{a^2(T+1)} + \frac{4G^2Re_m}{a} + \frac{2G^2e_m^2H_{T+1}^2}{a^2} + 2r_m^2. \tag{35}$$

In particular, with the displayed structural constants fixed, this is $O(T^{-1} + m^{-1/2} + H_{T+1}^2/m)$.

Proof. Put $D_t = \mathbb{E}\,\|u_t - u_*\|_2^2$ and $z_t = \|P_\perp u_t\|_2$. The comparator u_* belongs to the projection ball, so projection nonexpansiveness and independent tuples give

$$D_{t+1} \le D_t - 2\alpha_t \mathbb{E}\,\langle u_t - u_*, g(u_t; Z_t)\rangle + \alpha_t^2 U^2.$$

Lemma 8, $\|u_t - u_*\| \le 2R$, and Lemma 7 imply

$$D_{t+1} \le (1 - 2a\alpha_t) D_t + 2a\alpha_t \mathbb{E} z_t^2 + 4R\alpha_t e_m + \alpha_t^2 U^2.$$

Fix T and let $Z_T = e_m H_{T+1}/a$ and $C_T = Z_T^2 + 2Re_m/a$. The null-space bound gives, for $0 \le t < T$,

$$D_{t+1} - C_T \le \frac{t}{t+2}(D_t - C_T) + \frac{U^2/a^2}{(t+2)^2}.$$

Multiply by $(t+1)(t+2)$ and sum from zero to $T-1$. The initial coefficient is zero and all other coefficients telescope, yielding

$$T(T+1)(D_T - C_T) \le \frac{U^2}{a^2} \sum_{t=0}^{T-1} \frac{t+1}{t+2} \le \frac{U^2 T}{a^2}.$$

Hence $D_T \le C_T + U^2/[a^2(T+1)]$. Finally,

$$\|f_{u_T} - Q\|_{L_2(\mu)}^2 \le 2\,\|\ell_{u_T} - \ell_{u_*}\|_{L_2(\mu)}^2 + 2r_m^2 \le 2G^2\,\|u_T - u_*\|_2^2 + 2r_m^2.$$

Substitution proves the theorem. The proof controls actual null-space movement and never uses projection nonexpansiveness with a comparator outside the permitted ball. □

What this result assumes and proves. Exact representation is imposed on the initial local tangent class. Optimization is performed on the nonlinear network with all hidden weights trainable. The theorem proves its evaluation accuracy from the stated geometry and sample law; it does not assume an accurate output. It requires no positive definiteness on the full, redundant parameter space. Clipping the final output to $[-Q_{\max}, Q_{\max}]$ can only reduce its error, but clipping is not inserted into the TD updates.

D.3.5 Confidence, numerical budgets, and trajectory costs

Proposition 2 (Numerical evaluator calibration)**.** *For explicit calibration write the right side of* (35) *as*

$$\frac{A_T}{T+1} + \frac{A_1}{\sqrt{m}} + \frac{A_2 H_{T+1}^2 + A_3}{m},$$

where

$$A_T = \frac{2G^2U^2}{a^2}, \quad A_1 = \frac{4G^2RA}{a}, \quad A_2 = \frac{2G^2A^2}{a^2}, \quad A_3 = \frac{L_2^2 R^4}{2}.$$

For requested root error η, set $r = \eta/5$. Choose integers $T \ge 1$ and m (also meeting all geometry requirements) so that

$$T + 1 \ge \frac{48A_T}{r^2}, \qquad m \ge \max\left\{\left(\frac{48A_1}{r^2}\right)^2, \frac{48(A_2H_{T+1}^2 + A_3)}{r^2}\right\}. \tag{36}$$

Then a clipped replica has root error at most r with probability at least $7/8$.

Proof. Each initialization in $\mathcal{E}$ then gives mean squared error at most $r^2/16$. Assumption 3 supplies $\mathbb{P}(\mathcal{E} \mid \mathcal{H}) \ge 15/16$. Markov's inequality and the initialization failure allocation show that a clipped replica has root error at most r with probability at least $7/8$. A deterministic good initialization also meets this requirement.

□

Lemma 10 (Confidence conversion without true value labels)**.** *For desired conditional failure $\delta \in (0,1)$, run an odd number $M \geq 32\log(4/\delta)$ of independent replicas satisfying the preceding constant-confidence guarantee. On a fresh independent block of n inputs from μ, measure all pairwise empirical squared distances, including zero distance to self. If*

$$n \geq \frac{32Q_{\max}^2}{r^2}\log\frac{8M^2}{\delta}, \tag{37}$$

return a replica having the smallest median squared distance. Its root error is at most $5r = \eta$ except with probability $\delta/2$ in the ideal independent experiment.

Proof. A strict majority of replicas is r-accurate except with probability at most $\exp(-9M/32) \leq \delta/4$. Conditional on all replicas, a squared pairwise difference is in $[0, 4Q_{\max}^2]$ and its variance is at most $4Q_{\max}^2$ times its mean. Bernstein's inequality, the inequality $\sqrt{2tz} \leq t/2 + z$, and a union bound over pairs give

$$|\widehat{d}_{ij}^2 - d_{ij}^2| \leq \tfrac{1}{2}d_{ij}^2 + r^2$$

except with probability $\delta/4$, under (37). Good pairs have empirical squared distance at most $7r^2$, so the selected median is at most $7r^2$. Its set of at least $(M+1)/2$ neighbors at this distance intersects the good majority. For a good neighbor the true squared distance is at most $16r^2$. The triangle inequality gives $4r + r = 5r$. This is the manuscript's bounded-distance selection argument with an explicit allocation for the later coupling. □

D.3.6 Obtaining the independent experiment from a trajectory

Suppose the frozen state–action chain is uniformly geometrically mixing:

$$\sup_x \|\mathcal{P}^g(x,\cdot) - \mu\|_{\mathrm{TV}} \leq \kappa_{\mathrm{mix}}\beta^g, \qquad 0 < \beta < 1.$$

At each retained observation, collect a full one-step tuple and then run g intervening transitions before the next tuple begins. Run an initial g transitions as well. Use disjoint blocks for the M replicas and the final validation inputs. For counting, validation inputs may also be collected as one-step tuples whose labels are discarded. All observations are collected while this family is active, with the same frozen policy. There are $N = MT + n$ retained tuples.

Conditional on the entire past at the end of each retained tuple, g intervening transitions make the next input law within $\kappa_{\mathrm{mix}}\beta^g$ of μ. Appending the correct reward and transition kernel cannot increase total variation. Sequential maximal coupling therefore compares the retained tuples to independent stationary tuples with total failure at most $(N+1)\kappa_{\mathrm{mix}}\beta^g$. Choose

$$g = \max\left\{1, \left\lceil \frac{\log\max\{1, 2\kappa_{\mathrm{mix}}(N+1)/\delta\}}{-\log\beta} \right\rceil\right\}. \tag{38}$$

With a final exit interval included, the total transition cost is

$$N_{\mathrm{eval}} \leq N + (N+1)g. \tag{39}$$

This couples both training and validation blocks; it is not a claim that finitely spaced samples are exactly independent. Initialization randomness is independent across replicas and independent of collection. The target and evaluation policy are fixed before that randomness.

Combining Lemma 10 with the coupling gives conditional failure at most δ. All budgets are prescribed in advance, including on failure histories. Under uniform geometry and mixing factors,

$$T = \widetilde{\mathcal{O}}(\eta^{-2}), \qquad m = \widetilde{\mathcal{O}}(\eta^{-4}) \text{ is sufficient when geometry also holds,} \qquad N_{\mathrm{eval}} = \widetilde{\mathcal{O}}(\eta^{-2}). \tag{40}$$

The notation suppresses logarithms in accuracy and confidence only. The constants $a, G, R, B_0, \kappa_{\mathrm{mix}}, \beta$ must stay explicit whenever they depend on width, accuracy, exploration, or family; use (36)–(39) in that case. No claim of computational efficiency follows from the observation rate: network evaluation and training still cost time proportional to width.

Corollary 11 (Amplified fixed-policy evaluation)**.** *Under the network specification, Assumptions 2(a) and 3(a), and budgets* (36)*,* (37)*, and* (38)*, the returned clipped critic has root error at most η with conditional probability $1 - \delta_q$. Use the preceding formulas with $\delta = \delta_q$. Its complete observation cost is* (39)*, including every training and validation gap. In particular, $N_{\mathrm{eval}} = \widetilde{\mathcal{O}}(\mathsf{E}_{\mathrm{eval}}\eta^{-2})$ only with actual geometry, mixing, and width factors retained in $\mathsf{E}_{\mathrm{eval}}$.*

Proof. Lemma 10 spends at most $\delta_q/2$ on training and selection in the ideal product experiment; (38) spends at most $\delta_q/2$ on coupling. Sum these failures and count $MT + n$ retained tuples with all gaps. The network and geometry requirements must hold at the same chosen width. □

For snapshots use $\widehat{V}(s) = \sum_a \pi(a \mid s)\widehat{Q}(s,a)$; Jensen gives $\|\widehat{V} - V^\pi\|_{L_2(\rho)} \le \|\widehat{Q} - Q^\pi\|_{L_2(\rho\otimes\pi)}$. Diagnostic references and named anchors receive separate evaluations. All complete-call failures, including the internal trajectory coupling, enter the evaluator pool exactly once (Appendix D.9.3).

D.3.7 Evaluator interface used by the main theorem

Single-normalization evaluator. The network is defined in (26), and its initialization is specified in Appendix D.3.1. For the supplied covariance-range lower bound λ_{cov}, projected TD at iteration n uses step $[c\lambda_{\mathrm{cov}}(n+2)]^{-1}$, clipping only the returned critic.

Proposition 3 (Direct neural evaluation)**.** *Under Assumptions 2–3 and the evaluator specification, the budgeted projected-TD routine returns, for a frozen π,*

$$\|\widehat{Q} - Q^\pi\|_{L_2(\rho\otimes\pi)} \le \eta \tag{41}$$

with probability at least $1 - \delta_q$, using $\widetilde{\mathcal{O}}(E\eta^{-2})$ transitions with actual geometry and collection factors retained. Under uniform geometry, width $\widetilde{\mathcal{O}}(\eta^{-4})$ suffices at which the realization condition holds.

Proof idea. Width-uniform derivatives control nonlinear TD perturbations. Tangent comparison separates covariance-range contraction from null-space drift. Independent clipped replicas, distance-based selection, and spaced-tuple coupling yield conditional confidence (Appendix D.3). Jensen gives value snapshots; Appendix D.10.1 illustrates nonzero local geometry.

D.4 Auxiliary Bellman regression and all-parameter optimization

D.4.1 Bellman regression with approximation and optimization error

Definition 4 (Approximation of the required Bellman targets)**.** *For every certification call, family i, and frozen pair (V,π) required at that call, the specified regression class $\mathcal{H}_i \subset [-Q_{\max}, Q_{\max}]^{\mathcal{S}}$ satisfies*

$$a_i(V,\pi) := \inf_{g\in\mathcal{H}_i} \|g - T_i^\pi V\|_{L^2(\mu_i)} \le \varepsilon_{\mathrm{BE},i}. \tag{42}$$

The bounds are deterministic envelopes over the permitted histories, or are themselves established on a separately allocated event. The analogous condition is imposed on the state–action diagnostic targets in Section D.5.2. For a finite sieve, the approximation bound must include its discretization error.

This is an approximation condition on Bellman images, related to inherent Bellman error (Munos and Szepesvári, 2008) and approximate Bellman completeness (Chang et al., 2022). It is not implied by approximability of the fixed points Q_i^π. Requiring it for all functions in a specified value class and all policies would give a stronger uniform condition for the corresponding classes. Here it must cover every target actually needed, including cross-family diagnostic references. It does not assume that $\|V - T_i^\pi V\|_{\mu_i}$ is already small.

Finite and covered classes For a function g, let $\widehat{L}_n(g) = n^{-1}\sum_t (g(U_t) - Z_t)^2$ and $r(g) = \|g - m\|^2_{L^2(\mu)}$. An empirical solver with error $\varepsilon_{\mathrm{opt}} \ge 0$ returns $\widehat{g}$ satisfying

$$\widehat{L}_n(\widehat{g}) \le \inf_{g\in\mathcal{H}} \widehat{L}_n(g) + \varepsilon_{\mathrm{opt}}. \tag{43}$$

The all-parameter solver in Theorem 18 establishes this requirement constructively. In the continual application, Assumption 3(b) requires nonlinear target approximation error $a \le u/3$ for the prescribed fitted root-error tolerance u. This bound holds at the same selected width and radius as the statistical and optimization budgets in Corollary 19, so approximation error is included in the total fitted-error guarantee. The general supporting bounds below retain approximation and optimization errors explicitly.

Lemma 12 (Bernstein inequality used in the proofs)**.** *For independent centered variables W_t with $|W_t| \le b$ and $\mathbb{E}W_t^2 \le v$, with probability at least $1 - 2e^{-x}$,*

$$\left| \frac{1}{n} \sum_{t=1}^{n} W_t \right| \le \sqrt{\frac{2vx}{n}} + \frac{bx}{3n}, \qquad x > 0.$$

Proof. For $0 \le z < 3/b$, the power series of the exponential and $k! \ge 2\,3^{k-2}$ for $k \ge 2$ give

$$\log \mathbb{E}e^{zW_t} \le \frac{z^2 v}{2(1 - zb/3)}.$$

Independence gives the same bound with v replaced by nv for the sum. If $v > 0$, insert $z = \sqrt{2x/(nv)}/(1 + (b/3)\sqrt{2x/(nv)})$ into the exponential Markov bound at threshold $\sqrt{2nvx} + (b/3)x$. The exponent is at most $-x$. Apply the same argument to $-W_t$ and divide by n. If $v = 0$, the variables vanish almost surely and the conclusion is immediate. □

Theorem 13 (Finite-class replay regression oracle inequality)**.** *Under Definition 3, let $\mathcal{G}$ be a nonempty finite class of functions bounded by $Q_{\max}$ and fixed by $\mathcal{F}$. Any $\widehat{g} \in \mathcal{G}$ satisfying* (43) *obeys, with conditional probability at least $1 - \delta$,*

$$\boxed{r(\widehat{g}) \le 3 \inf_{g \in \mathcal{G}} r(g) + \frac{48\Lambda Q_{\max}^2}{n} \log \frac{2|\mathcal{G}|}{\delta} + 2\varepsilon_{\mathrm{opt}}.} \tag{44}$$

In particular, the replay statistical term is proportional to $p_{\log}^{-1}$, including when the approximation error is nonzero.

Proof. Condition on $\mathcal{F}$. For $D_g = g - m$ and $\epsilon = Z - m(U)$, define the excess loss relative to the conditional mean, used only in the proof:

$$\ell_g = (g(U) - Z)^2 - (m(U) - Z)^2 = D_g(U)^2 - 2D_g(U)\epsilon.$$

Then $\mathbb{E}\ell_g = r(g)$, and conditional centering of ϵ yields

$$\mathbb{E}\ell_g^2 = \mathbb{E}[D_g^4 + 4D_g^2 \operatorname{Var}(Z \mid U)] \le 8\Lambda Q_{\max}^2 r(g).$$

Here $D_g^2 \le 4Q_{\max}^2$ and $\operatorname{Var}(Z \mid U) \le \Lambda Q_{\max}^2$. The identity $\ell_g = (g - m)(g + m - 2Z)$ gives $|\ell_g| \le 8\Lambda Q_{\max}^2$ and consequently $|\ell_g - r(g)| \le 12\Lambda Q_{\max}^2$. By Lemma 12 and $4\sqrt{Ar} \le r/2 + 8A$, with $A = \Lambda Q_{\max}^2 x/n$, we have

$$|\mathbb{P}_n \ell_g - r(g)| \le 4\sqrt{\frac{\Lambda Q_{\max}^2 r(g) x}{n}} + \frac{4\Lambda Q_{\max}^2 x}{n} \le \frac{1}{2} r(g) + \frac{12\Lambda Q_{\max}^2 x}{n}.$$

Take $x = \log(2|\mathcal{G}|/\delta)$ and union bound over $\mathcal{G}$. Let g_0 minimize r over the finite class and put $A_n = 12\Lambda Q_{\max}^2 x/n$. On this event,

$$\frac{1}{2} r(\widehat{g}) - A_n \le \mathbb{P}_n \ell_{\widehat{g}} \le \mathbb{P}_n \ell_{g_0} + \varepsilon_{\mathrm{opt}} \le \frac{3}{2} r(g_0) + A_n + \varepsilon_{\mathrm{opt}}.$$

Rearrangement proves (44). In particular, no well-specified-model Bernstein condition has been assumed: it follows from comparison with the conditional mean in the proof. □

Corollary 14 (A prescribed root-accuracy budget)**.** *For a target root accuracy $u > 0$, it is sufficient that*

$$\inf_{g \in \mathcal{G}} \|g - m\|_\mu \le u/3, \qquad \varepsilon_{\mathrm{opt}} \le u^2/6, \qquad n \ge \frac{144\Lambda Q_{\max}^2}{u^2} \log \frac{2|\mathcal{G}|}{\delta}.$$

Then $\|\widehat{g} - m\|_\mu \le u$ with probability at least $1 - \delta$. Without these calibrations, the square root of the right side of (44) *is the valid radius.*

Proof. The three terms in (44) are respectively at most $u^2/3$, $u^2/3$, and $u^2/3$. □

Uniform covering extension Let $N_\infty(\zeta, \mathcal{H})$ be the size of a finite ζ-net of $\mathcal{H}$ in the uniform norm, with centers in $\mathcal{H}$.

Theorem 15 (Covering extension for an actual empirical optimizer)**.** *Suppose $\mathcal{H} \subset [-Q_{\max}, Q_{\max}]^{\mathcal{U}}$ is fixed by $\mathcal{F}$ and has a finite uniform ζ-net. Any measurable $\widehat{g} \in \mathcal{H}$ satisfying* (43) *obeys, with conditional probability at least $1-\delta$,*

$$r(\widehat{g}) \le 3 \inf_{g \in \mathcal{H}} r(g) + \frac{48\Lambda Q_{\max}^2}{n} \log \frac{2N_\infty(\zeta, \mathcal{H})}{\delta} + 40\Lambda Q_{\max}\zeta + 2\varepsilon_{\mathrm{opt}}. \tag{45}$$

Proof. Apply the simultaneous excess-loss deviation in the proof of Theorem 13 to the net. If $\|g - g'\|_\infty \le \zeta$, then

$$|\ell_g - \ell_{g'}| \le 4\Lambda Q_{\max}\zeta, \qquad |r(g) - r(g')| \le 4Q_{\max}\zeta.$$

Consequently the net event implies, simultaneously for every $g \in \mathcal{H}$,

$$|\mathbb{P}_n \ell_g - r(g)| \le \frac{1}{2} r(g) + \frac{12\Lambda Q_{\max}^2}{n} \log \frac{2N_\infty(\zeta, \mathcal{H})}{\delta} + 10\Lambda Q_{\max}\zeta.$$

Compare $\widehat{g}$ with any population near-minimizer as in Theorem 13, then take the infimum. Multiplication of the additive deviation by four gives the displayed constant 40. □

D.4.2 A PROVED FULLY TRAINABLE REGRESSION ALGORITHM

Conditional data and neural class Condition on a history fixing the family, target, sampling law, network initialization, architecture, and all numerical budgets. Let (U_t, Z_t) be conditionally iid with

$$m(u) = \mathbb{E}[Z_t \mid U_t = u], \quad |m| \le Q_{\max}, \quad |Z_t| \le \Lambda Q_{\max}, \quad \mathbb{E}[Z_t^2 \mid U_t] \le \Lambda Q_{\max}^2, \quad \Lambda \ge 1. \tag{46}$$

Here $Q_{\max} = R_{\max}/(1-\gamma)$. For replay,

$$Z_t = \frac{\pi(A_t \mid S_t)}{\beta_i(A_t \mid S_t)}\{R_t + \gamma V(S_t')\}, \qquad U_t = S_t, \qquad \Lambda = p_{\log}^{-1},$$

where the known logging law has $\beta_i(a \mid s) \ge p_{\log} > 0$. Then $m = T_i^\pi V$. This is Lemma 6. For direct target-policy outcomes and for current-family diagnostic regression on state–action inputs, $\Lambda = 1$.

The logging floor $p_{\log}$ and deployed-policy exploration floor $p_{\min} = \alpha\nu_{\min}$ are different quantities. Using only the bound $\beta_i \in \Pi_\alpha$ permits $p_{\log} = p_{\min}$, but a dedicated logger can have a larger, accuracy-independent floor.

Definition 5 (Local smooth auxiliary regression class)**.** *Let $\Theta = \{\theta_0 + v : \|v\|_2 \le R\}$, $R > 0$, and suppose $f_\theta(u)$ is twice continuously differentiable in all its parameters on a neighborhood of this ball. Throughout it and on the input domain,*

$$|f_\theta(u)| \le Q_{\max}, \qquad \|\nabla f_\theta(u)\|_2 \le G, \qquad \left\|\nabla^2 f_\theta(u)\right\|_{\mathrm{op}} \le \chi. \tag{47}$$

Take $G > 0$, $\chi \ge 0$, and define the Bellman approximation error

$$a = \inf_{\theta \in \Theta} \|f_\theta - m\|_{L^2(\mu)} \le \varepsilon_{\mathrm{BE}}. \tag{48}$$

The constants and approximation bound are deterministic envelopes for each call, or hold on an explicitly allocated initialization event.

Assumption 3(b) requires $a \le u/3$ for the prescribed fitted root-error tolerance u. This approximation bound is measured in the nonlinear trained class at the same selected width and radius used for the statistical and optimization budgets, uniformly over the auxiliary targets and histories specified in Appendix D.1.3. Exact realization is the special case $a = 0$. No additional realizability of a tangent model is imposed. The tangent class below is used to control the class size:

$$\mathcal{K} = \{k_v = f_{\theta_0} + \langle \nabla f_{\theta_0}, v\rangle : \|v\|_2 \le R\}.$$

Write $N_{\mathcal{K}}(\zeta)$ for a finite uniform ζ-cover with centers in $\mathcal{K}$. A general finite network always has a cover controlled by its parameter dimension under (47); a width-independent bound needs additional structure and is not automatic.

Regularized all-parameter projected gradient descent Define

$$L_n(\theta) = \frac{1}{n}\sum_{j=1}^{n}(f_\theta(U_j) - Z_j)^2, \qquad \omega = 8\Lambda Q_{\max}\chi, \qquad F_n(\theta) = L_n(\theta) + \frac{\omega}{2}\left\|\theta - \theta_0\right\|_2^2,$$

and $M = 2G^2 + 3\omega/2$. Starting at $\theta^{(0)} = \theta_0$, perform

$$\begin{aligned}\theta^{(t+1)} &= \mathrm{Proj}_\Theta\left[\theta^{(t)} - \frac{1}{M}\left\{\frac{2}{n}\sum_{j=1}^{n}(f_{\theta^{(t)}}(U_j) - Z_j)\nabla f_{\theta^{(t)}}(U_j) + \omega(\theta^{(t)} - \theta_0)\right\}\right], \\ \widehat{\theta} &= \frac{1}{T}\sum_{t=1}^{T}\theta^{(t)}, \qquad \widehat{g} = f_{\widehat{\theta}}.\end{aligned} \tag{49}$$

The implemented predictor is the nonlinear network at the averaged parameter vector. Every parameter is included in the gradient and the projection. The same observations are reused for these optimization passes. Smoothness and curvature follow from the stated architecture; the target-approximation requirement remains structural. Related local curvature arguments appear in Richards and Rabbat (2021) and Oymak and Soltanolkotabi (2019); the constants used here are derived below.

Lemma 16 (Derived curvature and optimization guarantee)**.** *Under Definition 5, F_n is convex and M-smooth on Θ. If $\chi > 0$, it is at least $\omega/2$-strongly convex there. The output of* (49) *satisfies*

$$L_n(\widehat{\theta}) \le \inf_{\theta\in\Theta} L_n(\theta) + \frac{\omega R^2}{2} + \frac{MR^2}{2T}. \tag{50}$$

Proof. For one observation the Hessian of squared loss is

$$2\nabla f_\theta \nabla f_\theta^\top + 2(f_\theta - Z)\nabla^2 f_\theta.$$

Because $|f_\theta - Z| \le (1+\Lambda)Q_{\max} \le 2\Lambda Q_{\max}$,

$$-4\Lambda Q_{\max}\chi I \preceq \nabla^2 L_n \preceq (2G^2 + 4\Lambda Q_{\max}\chi)I.$$

Adding ωI gives lower bound $\omega I/2$ and upper bound MI. For $\chi = 0$ this is a convex quadratic-in-prediction composition with an affine parameter map on the ball, and the same inequalities hold.

Let θ_* minimize the continuous F_n on the compact ball. Projection optimality, convexity, and smoothness imply

$$F_n(\theta^{(t+1)}) - F_n(\theta_*) \le \frac{M}{2}\left(\left\|\theta^{(t)} - \theta_*\right\|_2^2 - \left\|\theta^{(t+1)} - \theta_*\right\|_2^2\right).$$

For completeness, insert the projection inequality

$$\left\langle \nabla F_n(\theta^{(t)}), \theta^{(t+1)} - \theta_* \right\rangle \le M\left\langle \theta^{(t)} - \theta^{(t+1)}, \theta^{(t+1)} - \theta_* \right\rangle$$

into the smoothness upper bound, and use the convexity lower bound at $\theta^{(t)}$. The squared step terms cancel by the polarization identity. Summing and applying convexity at the returned parameter average gives $F_n(\widehat{\theta}) - F_n(\theta_*) \le MR^2/(2T)$. Compare $F_n(\theta_*)$ with any unregularized-loss minimizer and bound its ridge term by $\omega R^2/2$. Dropping the nonnegative output ridge term proves (50). □

Thus the empirical optimization guarantee is proved for the algorithm. The regularizer's bias is retained; convexity or PL of the unregularized neural loss is not assumed.

Lemma 17 (Transfer of tangent covers to nonlinear covers)**.** *For every v in the radius-R ball,*

$$\left\|f_{\theta_0+v} - k_v\right\|_\infty \le \chi R^2/2.$$

Consequently the nonlinear class $\mathcal{H} = \{f_\theta : \theta \in \Theta\}$ satisfies

$$N_\infty(\zeta + \chi R^2, \mathcal{H}) \le N_{\mathcal{K}}(\zeta). \tag{51}$$

Proof. Taylor's formula with the uniform Hessian bound gives the first claim. Replace each tangent cover center k_{v_ℓ} by the actual network $f_{\theta_0+v_\ell}$. If k_v is within ζ of that center, two Taylor remainders and the triangle inequality give distance at most $\zeta + \chi R^2$ between the corresponding networks. These centers are members of the actual bounded nonlinear class. □

Theorem 18 (Fully trainable replay regression)**.** *Under* (46) *and Definition 5, the algorithm* (49) *obeys, with conditional probability at least* $1-\delta$*,*

$$\boxed{\begin{aligned}\|\widehat{g}-m\|_\mu^2 \le 3a^2 + \frac{48\Lambda Q_{\max}^2}{n}\log\frac{2N_{\mathcal{K}}(\zeta)}{\delta} \\ + 40\Lambda Q_{\max}\zeta + 48\Lambda Q_{\max}\chi R^2 + \frac{MR^2}{T}.\end{aligned}} \tag{52}$$

In particular, root error at most u *is guaranteed if*

$$\begin{gathered}a \le u/3, \qquad \zeta = Q_{\max}/n, \qquad \chi \le \frac{u^2}{144\Lambda Q_{\max}R^2}, \qquad T \ge \frac{6MR^2}{u^2}, \\ n \ge \frac{6\Lambda Q_{\max}^2}{u^2}\left[48\log\frac{2N_{\mathcal{K}}(Q_{\max}/n)}{\delta} + 40\right].\end{gathered} \tag{53}$$

Proof. Theorem 15 states, for any bounded class with a finite uniform η-cover and an empirical squared-loss error ϵ_{opt},

$$\|\widehat{g}-m\|_\mu^2 \le 3\inf_{g\in\mathcal{H}}\|g-m\|_\mu^2 + \frac{48\Lambda Q_{\max}^2}{n}\log\frac{2N_\infty(\eta,\mathcal{H})}{\delta} + 40\Lambda Q_{\max}\eta + 2\epsilon_{\mathrm{opt}}.$$

It is a consequence of excess-loss Bernstein concentration with label second moment $\Lambda Q_{\max}^2$, and allows nonzero approximation error. Apply it with $\eta = \zeta + \chi R^2$, using Lemma 17 and the proved error $\epsilon_{\mathrm{opt}} = \omega R^2/2 + MR^2/(2T)$. The curvature remainder contributes $40\Lambda Q_{\max}\chi R^2$ and ridge contributes $8\Lambda Q_{\max}\chi R^2$. This proves (52). Under (53), approximation, curvature plus ridge, statistical plus discretization, and optimization contribute at most $u^2/3, u^2/3, u^2/6, u^2/6$, respectively. □

Corollary 19 (Calibrated auxiliary approximation)**.** *Under Assumption 3(b), let the prescribed root tolerance be* $u > 0$ *and require* $a \le u/3$ *in the selected nonlinear class. In the setting of Theorem 18, the following budgets guarantee* $\|\widehat{g}-m\|_\mu \le u$ *with conditional probability at least* $1-\delta_r$*:*

$$\begin{gathered}\zeta = Q_{\max}/n, \qquad \chi \le \frac{u^2}{144\Lambda Q_{\max}R^2}, \qquad T_{\mathrm{reg}} \ge \frac{6MR^2}{u^2}, \\ n \ge \frac{6\Lambda Q_{\max}^2}{u^2}\left[48\log\frac{2N_{\mathcal{K}}(Q_{\max}/n)}{\delta_r} + 40\right].\end{gathered} \tag{54}$$

For the paired network in Proposition 4, also require even width satisfying (58) *and* $GR \le Q_{\max}$*. The approximation bound* $a \le u/3$ *must hold at this same width and radius, and* $N_{\mathcal{K}}$ *is the tangent cover of this selected class. The fitted-error guarantee includes approximation, statistical, curvature/ridge, and optimization errors. Gradient passes reuse the fitted block and do not multiply its observation cost.*

Proof. Apply (52) with $\delta = \delta_r$ and $T = T_{\mathrm{reg}}$. The approximation term satisfies $3a^2 \le u^2/3$. Under (54), the statistical plus discretization, curvature plus ridge, and optimization contributions are at most $u^2/6$, $u^2/3$, and $u^2/6$, respectively. Their total, including approximation, is at most u^2, which gives the claimed root-error bound. □

The budget inequality in n can be solved by deterministic doubling before data are exposed. Denote a supplied tangent-entropy complexity by $\mathfrak{C}(u,\Lambda)$, retaining its dependence on the chosen network. The theorem gives a labeled-record cost

$$N_{\mathrm{reg}}(u,\delta) = \widetilde{\mathcal{O}}\big(\Lambda Q_{\max}^2(\mathfrak{C}(u,\Lambda)+1)u^{-2}\big), \tag{55}$$

provided the corresponding approximation, width, and iteration conditions hold. Logarithms alone are suppressed by $\widetilde{\mathcal{O}}$; a growing dimension or covering complexity is not suppressed.

Concrete network: hidden weights and readout both trained Let $\|x(u)\|_2 \le 1$ for a fixed input representation, and consider

$$f_\theta(u) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} a_j \sigma(w_j^\top x(u)), \qquad \theta = (a_1, w_1, \dots, a_m, w_m). \tag{56}$$

Use even m and paired initialization with equal hidden weights and opposite readouts, so $f_{\theta_0} = 0$. Assume $|a_j^0| \le A_0$, $\|w_j^0\| \le W_0$. On the full parameter ball of radius R, put $A = A_0 + R$, $W = W_0 + R$, and let S_0, S_1, S_2 bound $|\sigma|, |\sigma'|, |\sigma''|$ on $[-W, W]$.

Proposition 4 (Explicit all-parameter derivative and width bounds)**.** *The network* (56) *satisfies*

$$G = \sqrt{S_0^2 + A^2 S_1^2}, \qquad \chi = \frac{H}{\sqrt{m}}, \qquad H = 2S_1 + AS_2, \qquad |f_\theta| \le GR. \tag{57}$$

If $GR \le Q_{\max}$, Definition 5's envelope holds. A sufficient width for the curvature condition of (53) *is an even integer satisfying*

$$m \ge \left(\frac{144 \Lambda Q_{\max} H R^2}{u^2} \right)^2. \tag{58}$$

Proof. The readout derivative is $\sigma(w_j^\top x)/\sqrt{m}$ and the hidden derivative is $a_j \sigma'(w_j^\top x) x/\sqrt{m}$. Summing their squared norms gives the displayed G. The full Hessian is block diagonal across neurons, with each block

$$\frac{1}{\sqrt{m}} \begin{pmatrix} 0 & \sigma'(w_j^\top x) x^\top \\ \sigma'(w_j^\top x) x & a_j \sigma''(w_j^\top x) x x^\top \end{pmatrix}.$$

Its operator norm is at most $H/\sqrt{m}$. The mean-value formula from $f_{\theta_0} = 0$ gives $|f_\theta| \le GR$. Substitution proves the width condition. □

Pairing is an initialization device; it is not imposed on later iterates. Both a_j and w_j receive gradients. The proof applies to twice differentiable activations such as smooth bounded activations or polynomials on the bounded preactivation range. It does not establish the same Hessian bound for ReLU.

If $GR \le Q_{\max}$ is inconvenient, a specified common envelope $B \ge \max\{Q_{\max}, GR\}$ can replace $Q_{\max}$ throughout the regression theorem. After smooth training, clipping only the final predictor to $[-Q_{\max}, Q_{\max}]$ does not increase its error relative to m and permits the monitor bounds of Lemma 20 with $Q_{\max}$. This retains the dependence on B in the regression and width costs. Importance-weighted labels are not clipped.

Proposition 5 (Tangent rank and a verified continuous-input example)**.** *Let*

$$r = \dim \operatorname{span}\{\nabla f_{\theta_0}(u) : u \in \mathcal{U}\}.$$

Then

$$N_{\mathcal{K}}(\zeta) \le (1 + 2GR/\zeta)^r. \tag{59}$$

For $\sigma(z) = z^2$ on augmented inputs $x \in \mathbb{R}^d$, $\|x\| \le 1$, $r \le d(d+1)/2$, independently of width. Moreover, a concrete paired initialization exactly represents every target $x^\top C x$ with $C = C^\top$ and $\|C\|_{\mathrm{F}} \le 2R/\sqrt{d}$ inside the nonlinear radius-R ball.

Proof. Project any parameter displacement orthogonally onto the span of the initial gradients. This leaves its tangent function unchanged and does not increase its norm. An r-dimensional parameter-ball cover at scale ζ/G proves (59). With quadratic activation all network outputs and tangent functions are homogeneous quadratic polynomials in the augmented input, a space of dimension $d(d+1)/2$.

For the last claim take $m = 2Ld$ and initialize L copies of $(a^+, w^+) = (1, e_\ell)$ and $(a^-, w^-) = (-1, e_\ell)$ for every coordinate ℓ. For a symmetric C, keep the comparator readouts at ± 1 and set

$$w_\ell^\pm = e_\ell \pm t_\ell, \qquad t_\ell = \frac{\sqrt{m}}{4L} C_{\ell,:}^\top.$$

Then

$$f_\theta(x) = \frac{L}{\sqrt{m}} \sum_{\ell=1}^{d} \left[\{(e_\ell + t_\ell)^\top x\}^2 - \{(e_\ell - t_\ell)^\top x\}^2 \right] = x^\top C x,$$

and $\|\theta - \theta_0\|_2^2 = 2L \sum_\ell \|t_\ell\|_2^2 = (d/4)\,\|C\|_{\mathrm{F}}^2 \le R^2$. This proves actual nonlinear realizability independent of width. □

An augmented constant coordinate supplies affine and constant terms. If a Bellman target is only within a in $L^2(\mu)$ of one of these quadratics, its nonlinear approximation error is at most a. This verifies approximation and entropy for a continuous-input regression example; it does not verify all assumptions of a complete continual-RL environment. For generic activations one may only have $r \le m(d+1)$, and the rank's width dependence must remain in (55). The fast rate is not universal for arbitrary fully trainable networks.

D.4.3 Bellman learner interface used by the main theorem

Auxiliary Bellman learner. For frozen V, π, a logged transition gives

$$Z = \frac{\pi(A \mid S)}{\beta(A \mid S)}[R + \gamma V(S')], \quad \mathbb{E}[Z \mid S] = T^\pi V, \quad \mathbb{E}[Z^2 \mid S] \le \Lambda Q_{\max}^2, \quad \Lambda = p_{\log}^{-1}. \tag{60}$$

Diagnostic state–action labels use $\Lambda = 1$. For the auxiliary class defined in Section 3, write θ_0, R for its center and radius and G, χ for its gradient/Hessian bounds. Projected descent minimizes

$$F_n(\theta) = n^{-1} \sum_t (f_\theta(U_t) - Z_t)^2 + \tfrac{\omega}{2}\|\theta - \theta_0\|_2^2, \qquad \omega = 8\Lambda Q_{\max}\chi, \tag{61}$$

using step $(2G^2 + 3\omega/2)^{-1}$ and averaged parameters.

Proposition 6 (All-parameter Bellman regression)**.** *Under Assumptions 2–3, let $u > 0$ be the prescribed root tolerance and require the selected nonlinear auxiliary class to satisfy*

$$a := \inf_{\theta : \|\theta - \theta_0\|_2 \le R} \|f_\theta - T^\pi V\|_\mu \le u/3.$$

With the auxiliary specification and the width/iteration budgets of Corollary 19, a valid block of $\widetilde{\mathcal{O}}(\Lambda Q_{\max}^2(\mathfrak{C}+1)u^{-2})$ labels yields $\|\widehat{g} - T^\pi V\|_\mu \le u$ with probability at least $1 - \delta_r$. Here $\mathfrak{C}$ retains the actual tangent-cover complexity at the selected width. The approximation bound and the statistical and optimization budgets must hold at this same width and radius.

Proof idea. The label mean is the Bellman image, with variance factor Λ. Ridge curvature makes the objective convex on the ball. Theorem 18 bounds the squared fitted error by $3a^2$ plus statistical, curvature/ridge, and optimization terms. The approximation condition gives $3a^2 \le u^2/3$, and the remaining budgets in Corollary 19 contribute at most $2u^2/3$. Their sum therefore gives the stated root-error guarantee (Appendix D.4.2).

D.5 Uniform preservation certification and family identification

D.5.1 Uniform routing and one-sided preservation certificates

Lemma 20 (Root empirical norm concentration)**.** *For independent $U_1, \dots, U_m \sim \mu$ and a fixed function $|h| \le B_0$, write $\|h\|_m = (m^{-1}\sum_t h(U_t)^2)^{1/2}$. With probability at least $1 - \delta$,*

$$\left| \|h\|_m - \|h\|_\mu \right| \le 2B_0 \sqrt{\frac{\log(2/\delta)}{m}}. \tag{62}$$

The assertion also holds conditionally when h is fixed before the block.

Proof. Put $a = \|h\|_\mu$, $\widehat{a} = \|h\|_m$, $x = \log(2/\delta)$, and $b = B_0\sqrt{x/m}$. For $X = h(U)^2$, $0 \le X \le B_0^2$ and $\mathrm{Var}(X) \le B_0^2 a^2$. Lemma 12, with a conservative enlargement of its linear term, gives

$$|\widehat{a}^2 - a^2| \le \sqrt{2}\, ab + \tfrac{2}{3}b^2.$$

The upper inequality implies $\widehat{a} \le a + b$. If $a \le 2b$, the lower deviation is already at most $2b$. If $a > 2b$, then $a^2 - \sqrt{2}\, ab - \frac{2}{3}b^2 \ge (a - 2b)^2$, since $(4 - \sqrt{2})a \ge (14/3)b$. Thus $\widehat{a} \ge a - 2b$. □

For a frozen snapshot V and policy π, a fitted Bellman target with $\|\widehat{g} - T^\pi V\|_\mu \le r$ therefore gives

$$\left| \|V - \widehat{g}\|_m - \|V - T^\pi V\|_\mu \right| \le r + 4Q_{\max}\sqrt{\frac{\log(2/\delta_{\mathrm{mon}})}{m}}, \tag{63}$$

on its regression event and a fresh monitor event. The radius contains the approximation and optimization contributions through r.

Lemma 21 (Positive-part contraction)**.** *For either an empirical or population law m, $|\,\|f_+\|_m - \|g_+\|_m\,| \le \|f-g\|_m$. If $f_\lambda = a + \lambda b$ is bounded, then $\lambda \mapsto \|(f_\lambda)_+\|_m$ is continuous and convex.*

Proof. The scalar positive-part map is 1-Lipschitz. Apply the reverse triangle inequality. Convexity follows from $(tf+(1-t)g)_+ \le tf_+ + (1-t)g_+$, monotonicity of the norm on nonnegative functions, and the triangle inequality. □

Theorem 22 (Uniform one-sided certificate on the scalar policy hull)**.** *Freeze V, π^0, and π^1, with $|V| \le Q_{\max}$. For $a = 0, 1$, let $g^a = T^{\pi^a}V$ and obtain fits $|\widehat{g}^a| \le Q_{\max}$ satisfying $\|\widehat{g}^a - g^a\|_\mu \le r_a$. The two fits may use the same valid labeled block, with confidence allocated to both targets. For $0 \le \lambda \le 1$, put*

$$\pi^\lambda = (1-\lambda)\pi^0 + \lambda\pi^1, \qquad \widehat{g}^\lambda = (1-\lambda)\widehat{g}^0 + \lambda\widehat{g}^1,$$
$$R^+(\lambda) = \left\|(V - T^{\pi^\lambda}V)_+\right\|_\mu, \qquad \widehat{R}^+(\lambda) = \left\|(V - \widehat{g}^\lambda)_+\right\|_m.$$

For $u > 0$, let $N = \max\{1, \lceil 4Q_{\max}/u \rceil\}$ and take a fresh monitor block of size

$$m \ge \frac{64Q_{\max}^2}{u^2}\log\frac{2(N+1)}{\delta_{\mathrm{mon}}}. \tag{64}$$

With conditional probability at least $1 - \delta_{\mathrm{mon}}$, on the endpoint regression events,

$$\boxed{|\widehat{R}^+(\lambda) - R^+(\lambda)| \le (1-\lambda)r_0 + \lambda r_1 + u \le e := \max\{r_0, r_1\} + u \quad \textit{for every } \lambda \in [0,1].} \tag{65}$$

Proof. Bellman affinity in action probabilities gives $T^{\pi^\lambda}V = (1-\lambda)g^0 + \lambda g^1$. Hence $\left\|\widehat{g}^\lambda - T^{\pi^\lambda}V\right\|_\mu \le (1-\lambda)r_0 + \lambda r_1$. Lemma 21 transfers this bound to the positive-part norms. For the uniform monitor argument put $h_\lambda = (V - \widehat{g}^\lambda)_+$. Then $|h_\lambda| \le 2Q_{\max}$ and $\|h_\lambda - h_{\lambda'}\|_\infty \le 2Q_{\max}|\lambda - \lambda'|$. The empirical and population norms are both $2Q_{\max}$-Lipschitz in λ. On the grid $\{j/N : 0 \le j \le N\}$, Lemma 20 and a union bound give maximum norm discrepancy at most $4Q_{\max}\sqrt{\log(2(N+1)/\delta_{\mathrm{mon}})/m} \le u/2$. Every point is within $1/(2N)$ of a grid point, so the additional empirical-minus-population discrepancy is at most $2Q_{\max}/N \le u/2$. Combine this uniform monitor event with the population approximation bound. □

Theorem 23 (Certified scalar solve, movement, and mandatory audit)**.** *For each protected family i, suppose Theorem 22 supplies $|\widehat{R}_i^+(\lambda) - R_i^+(\lambda)| \le e_i \le \tau_i/8$ uniformly in λ. Set $U_i^+ = \widehat{R}_i^+ + e_i$ and solve*

$$\widehat{\lambda} = \max\{\lambda \in [0,1] : U_i^+(\lambda) \le 3\tau_i/4 \textit{ for every protected } i\}. \tag{66}$$

Test 1 first. With no protected records return 1; an empty feasible set, or a returned weight below $\lambda_{\min}$, triggers singleton fallback. On the simultaneous certificate event:

- *(i) A feasible output has $R_i^+(\widehat{\lambda}) \le 3\tau_i/4$.*
- *(ii) If $R_i^+(\lambda_*) \le \tau_i/2$ for every i, then λ_* is feasible and $\widehat{\lambda} \ge \lambda_*$.*
- *(iii) Seal the candidate, and take fresh audit inputs with $|\widehat{R}_{a,i}^+ - R_i^+(\widehat{\lambda})| \le e_{a,i} \le \tau_i/8$. The required audit $\widehat{R}_{a,i}^+ + e_{a,i} \le \tau_i$ passes; every audited accepted output satisfies $R_i^+(\widehat{\lambda}) \le \tau_i$.*

Proof. Lemma 21 makes each feasible sublevel set a compact interval, possibly empty. Their intersection has a largest point if nonempty. Uniform accuracy gives $R_i^+(\widehat{\lambda}) \le U_i^+(\widehat{\lambda}) \le 3\tau_i/4$. At a comparison point, $U_i^+(\lambda_*) \le R_i^+(\lambda_*) + 2e_i \le 3\tau_i/4$. Finally $\widehat{R}_{a,i}^+ + e_{a,i} \le R_i^+(\widehat{\lambda}) + 2e_{a,i} \le \tau_i$. An accepted audit directly upper-bounds the true score. □

Proposition 7 (Per-transaction permissiveness)**.** *Fix one transaction's incumbent, endpoint, protected snapshots, endpoint fits, solve-monitor samples, and pre-attempt reserve inventory. Compare the symmetric and one-sided rules with the same thresholds τ_i, solve radii $e_i \le \tau_i/8$, audit radii $e_{a,i} \le \tau_i/8$, lower step $\lambda_{\min}$, and operational guards. Writing V_i for the frozen protected snapshot, define the symmetric scores by*

$$R_i(\lambda) = \left\| V_i - T^{\pi^\lambda} V_i \right\|_{\mu_i}, \qquad \widehat{R}_i(\lambda) = \left\| V_i - \widehat{g}_i^{\lambda} \right\|_{m_i}.$$

For every $\lambda \in [0,1]$, $R_i^+(\lambda) \le R_i(\lambda)$ and $\widehat{R}_i^+(\lambda) \le \widehat{R}_i(\lambda)$. Consequently:

(i) Every weight feasible for the symmetric solve is feasible for the one-sided solve. If the symmetric feasible set is nonempty, the largest one-sided feasible weight is at least the largest symmetric one.

(ii) On the one-sided uniform certificate event and the audit event of Theorem 23(iii) for its selected candidate, whenever the symmetric rule installs a shared step of weight at least $\lambda_{\min}$, the one-sided rule also installs a shared step of weight at least $\lambda_{\min}$ rather than splitting.

Proof. The scalar inequality $0 \le x_+ \le |x|$ and monotonicity of the L_2 norm on nonnegative functions give both score comparisons. Adding the same e_i preserves the empirical comparison, so every constraint $\widehat{R}_i(\lambda) + e_i \le 3\tau_i/4$ implies its one-sided counterpart. This proves (i), including the endpoint-first convention. For (ii), an installed symmetric step implies that the common pre-attempt reserve inventory permits the required transaction. Its weight is feasible for the one-sided solve, whose largest feasible weight is therefore at least $\lambda_{\min}$. On the stated events, Theorem 23(iii) makes the one-sided candidate's mandatory audit pass. No movement, audit, or reserve-shortage fallback is then triggered. With no protected records both rules return the full step. □

The comparison is per transaction. Once the rules install different policies their histories differ; no trajectory-wide ordering of split counts is asserted. In particular, the symmetric candidate's realized audit success alone does not imply the one-sided candidate's audit success: part (ii) uses the latter's stated audit event. The uniform fit and monitor radii can be chosen identically for both rules by the same contraction and grid argument in Theorem 22.

Proposition 8 (Finite scalar solve)**.** *For finite monitor blocks, the largest feasible weight is computable by finitely many piecewise-quadratic interval calculations.*

Proof. Write each fitted residual as $a_{it} + b_{it}\lambda$ and form

$$F_i(\lambda) = m_i^{-1} \sum_t (a_{it} + b_{it}\lambda)_+^2.$$

Sort $0, 1$ and all breakpoints $-a_{it}/b_{it} \in (0,1)$ with $b_{it} \neq 0$. On each successive closed interval, each summand is zero or a fixed quadratic, so all constraints become convex quadratic inequalities $F_i(\lambda) \le (3\tau_i/4 - e_i)^2$. Intersect their quadratic solution intervals with that interval, retain the feasible pieces, and take the largest endpoint. The formulas agree at every breakpoint. This is a finite exact-arithmetic construction; interval arithmetic can conservatively enclose roots and scores. Zero need not be feasible, so bisection from zero is not justified. Numerical score error must be added to the upper confidence bounds; conservative feasibility alone does not prove exact maximum movement. The sharing result uses the explicit test at 1, with positive slack in (119), and requires no boundary search. □

The mandatory audit reuses the fitted endpoints on fresh monitor states, with the selected candidate fixed before their disclosure. Positive-part contraction and Lemma 20 give the complete radius

$$e_{a,i} = \max_a r_{i,a} + 4Q_{\max}\sqrt{\log(2/\delta_{a,i})/m_{a,i}}. \tag{67}$$

Endpoint, solve-monitor, and audit-monitor root errors at most $\tau_i/16$ give $e_i, e_{a,i} \le \tau_i/8$. The audit needs fresh states but no new transition labels; it depends on the endpoint regression event. An audit failure never authorizes a shared installation.

Remark 24 (Symmetric routing and optional symmetric preservation)**.** *Routing continues to use the ordinary root residual in* (63)*; its two-sided diagnostic separation is unchanged. Omitting the positive parts throughout Theorems 22–23 gives the symmetric preservation variant and a single quadratic per family. Both versions have the same monitor budget: the grid argument already accounts for uniform selection of* λ*.*

Corollary 25 (Quadratic residual cost with an explicit replay factor)**.** *For each protected family, use the all-parameter neural or finite-class regression theorem with both endpoint errors at most* $\tau/16$*. Take the uniform solve monitor error and the monitor-only audit error at most* $\tau/16$*. If the approximation and optimization budgets satisfy those root-accuracy requirements, the sufficient per-family cost is*

$$\begin{array}{ll} \widetilde{\mathcal{O}}\big(\Lambda Q_{\max}^2(\mathfrak{C}+1)\tau^{-2}\big) & \textit{labeled records and monitor inputs together,} \\ \Lambda = p_{\log}^{-1} & \textit{for retained replay,} \\ \Lambda = 1 & \textit{for direct endpoint sampling,} \end{array} \tag{68}$$

before any additional input-acquisition or mixing factor. For a finite sieve, $\mathfrak{C} = \log|\mathcal{G}|$*; for the neural solver it is the actual tangent covering complexity in* (54)*.*

Proof. Apply Theorem 18, Corollary 14, or (54) at $u = \tau/16$, allocating confidence to two endpoint targets. One replay tuple supplies both weighted labels, so both fits use the same n tuples. Direct endpoint sampling uses two sets of outcomes, a constant factor. Equation (64) gives $\widetilde{\mathcal{O}}(Q_{\max}^2\tau^{-2})$ monitor inputs. The monitor-only audit has the same order from (63). These are dominated by the displayed expression since $\Lambda \geq 1$ and $\mathfrak{C} + 1 \geq 1$. □

If $p_{\log} = \Theta(\varepsilon)$ and $\tau = \Theta(\varepsilon)$ while all other displayed factors are bounded, (68) has order $\widetilde{\mathcal{O}}(\varepsilon^{-3})$ for replay and $\widetilde{\mathcal{O}}(\varepsilon^{-2})$ for direct sampling. These are *per-comparison* consequences. The complete ledger in Theorem 1 combines the number of attempts, active critics, snapshot acquisition, routing, and all collected reserves. In particular, a growing feature dimension or sieve entropy changes these exponents.

D.5.2 Bellman-family identification with certified references

Error transfer from primitive law domination

Proposition 9 (Law domination implies reference-error transfer)**.** *Let* P *be a Markov kernel,* $\mu \leq C_0\upsilon$*, and* $\mu P \leq C_1\upsilon$*. For* $e \in L^2(\upsilon)$*,*

$$\|e - \gamma Pe\|_{L^2(\mu)} \leq (\sqrt{C_0} + \gamma\sqrt{C_1})\|e\|_{L^2(\upsilon)}. \tag{69}$$

This applies to state kernels for value snapshots and state–action kernels for diagnostic references.

Proof. Domination gives $\|e\|_\mu \leq \sqrt{C_0}\|e\|_\upsilon$. Jensen gives $\|Pe\|_\mu^2 \leq (\mu P)e^2 \leq C_1\|e\|_\upsilon^2$. The triangle inequality proves the claim. □

For a stored reference q_ℓ with $\|q_\ell - Q_\ell^{\mathrm{id}}\|_{\nu_\ell^{\mathrm{id}}} \leq \eta_{\mathrm{id}}$, Assumption 2(b) and (19) thus supplies

$$A_{\mathrm{id},k\ell} = \sqrt{D_{0,k\ell}} + \gamma\sqrt{D_{1,k\ell}}, \qquad a_{\mathrm{id}} = \max_{k,\ell} A_{\mathrm{id},k\ell}\eta_{\mathrm{id}}. \tag{70}$$

A fixed deterministic envelope is used when prescribing budgets. Stationarity in family ℓ alone does not imply the needed cross-family law domination. Diagnostic references are clipped to $[-Q_{\max}, Q_{\max}]$, which cannot increase their error.

Root-regression scores and routing Use diagnostic notation on state–action inputs $X = (S, A)$:

$$\mathcal{T}_k^{\mathrm{id}} q = r_k + \gamma P_k^{\mathrm{id}} q, \qquad Q_k^{\mathrm{id}} = \mathcal{T}_k^{\mathrm{id}} Q_k^{\mathrm{id}}.$$

The current family k is unknown, and its diagnostic input law is μ_k^{id}. The fixed diagnostic policy is π_{id}. Use the operator separation from Assumption 1

$$\big\|Q_\ell^{\mathrm{id}} - \mathcal{T}_k^{\mathrm{id}} Q_\ell^{\mathrm{id}}\big\|_{L^2(\mu_k^{\mathrm{id}})} \geq \Delta_0 > 0 \quad (k \neq \ell) \tag{71}$$

and realized reference transfer

$$\left\|(q_\ell - \mathcal{T}_k^{\mathrm{id}} q_\ell) - (Q_\ell^{\mathrm{id}} - \mathcal{T}_k^{\mathrm{id}} Q_\ell^{\mathrm{id}})\right\|_{L^2(\mu_k^{\mathrm{id}})} \le a_{\mathrm{id}} = A_{\mathrm{id}}\eta_{\mathrm{id}}. \tag{72}$$

The bound holds for every current family and stored reference, including the matching pair. Proposition 9 derives this bound from Assumption 2(b) and (19).

Before observing a fresh diagnostic block, freeze the reference bank $q_1, \dots, q_M$. From each one-step tuple of the current environment, form, for every reference,

$$Y_{t,\ell} = R_t + \gamma \sum_a \pi_{\mathrm{id}}(a \mid S_t') q_\ell(S_t', a).$$

Regress this label on $X_t = (S_t, A_t)$ to fit $g_{k,\ell} = \mathcal{T}_k^{\mathrm{id}} q_\ell$. The inputs include the current action; no target-action importance ratio is needed. The label is bounded by $Q_{\max}$, so the regression theorems apply with $\Lambda = 1$. Use the same labeled block for all references and a fresh common monitor block for their root scores

$$\widehat{R}_{k,\ell} = \left\|q_\ell - \widehat{g}_{k,\ell}\right\|_m.$$

The regression class used by the algorithm must be available without knowing k; it may depend on the frozen reference or visible history. For the primary calibration, the selected nonlinear class $\mathcal{G}_\ell^{\mathrm{id}}$ must satisfy

$$\inf_{f \in \mathcal{G}_\ell^{\mathrm{id}}} \|f - \mathcal{T}_k^{\mathrm{id}} q_\ell\|_{\mu_k^{\mathrm{id}}} \le \Delta_0/48,$$

uniformly over possible current families, frozen references, and permitted pre-call histories, including wrong-family comparisons. This bound must hold at the same selected width and radius used for the regression budgets, as specified in Appendix D.1.3. The general score bound below allows any certified regression errors.

Theorem 26 (Simultaneous root scores and correct routing)**.** *Suppose simultaneous regression events give $\|\widehat{g}_{k,\ell} - g_{k,\ell}\|_{\mu_k^{\mathrm{id}}} \le r_\ell$. For $M \ge 1$, with monitor failure at most δ_{mon},*

$$\max_{\ell \le M} \left| \widehat{R}_{k,\ell} - \left\|q_\ell - \mathcal{T}_k^{\mathrm{id}} q_\ell\right\|_{\mu_k^{\mathrm{id}}} \right| \le \rho := \max_\ell r_\ell + 4Q_{\max}\sqrt{\frac{\log(2M/\delta_{\mathrm{mon}})}{m}}. \tag{73}$$

Accept a lowest-score stored reference if its score is at most $a_{\mathrm{id}} + \rho$; otherwise create a new record. On the reference and score events, every decision is correct whenever

$$\Delta_0 > 2a_{\mathrm{id}} + 2\rho. \tag{74}$$

Starting from an empty bank, this yields one record per encountered family, independently of its assignment to a policy module.

Proof. Conditional on the frozen fits and references, apply Lemma 20 to $q_\ell - \widehat{g}_{k,\ell}$, bounded by $2Q_{\max}$, and union bound over ℓ. Add the regression errors by the reverse triangle inequality to obtain (73). The exact matching value Q_k^{id} has zero Bellman residual, so (72) gives $R_{k,k} \le a_{\mathrm{id}}$. For a wrong reference, reverse triangle inequality and (71) give $R_{k,\ell} \ge \Delta_0 - a_{\mathrm{id}}$. The correct reference, when present, has empirical score at most $a_{\mathrm{id}} + \rho$; each wrong reference has empirical score at least $\Delta_0 - a_{\mathrm{id}} - \rho > a_{\mathrm{id}} + \rho$. Thus the correct reference is the unique accepted minimizer when present, and all references are rejected for an unseen family. An empty bank creates its first record without a score test. Induct over segment decisions and use Lemma 4 for the simultaneous probability statement. □

Proposition 10 (Adequacy from the known separation bound)**.** *Use Assumptions 1, 2, and 3, and the numerical choices in* (110)*. On the joint successful event of Appendix D.9.3, all S routing decisions are correct using threshold $t = \Delta_0/4$.*

Proof. Reference transfer is at most $A_{\mathrm{id}}\eta_{\mathrm{id}} = \Delta_0/8 = t/2$. For each diagnostic fit, set $u = \Delta_0/16$ and $\Lambda = 1$. Assumption 3(b) and Appendix D.1.3 give nonlinear target approximation error $a \le \Delta_0/48 = u/3$, uniformly over stored references and possible current families, including wrong-family comparisons. Corollary 19, with the prescribed width, iteration, and sample budgets, therefore gives total regression error at most $\Delta_0/16 = t/4$, including approximation error. On the joint successful event these regression bounds hold simultaneously.

With at most K references, (73) and the monitor budget give another $\Delta_0/16 = t/4$. Thus matching scores are at most t and wrong scores at least $\Delta_0 - t = 3\Delta_0/4 > t$. Induction proves correct record creation and recurrence identification. Lemma 4 gives the probability statement under the complete allocation; no independence between routing calls is needed. □

The input Δ_0 may conservatively underestimate the actual minimum gap. Reference training costs $\widetilde{\mathcal{O}}(KE_{\mathrm{id}}A_{\mathrm{id}}^2\Delta_0^{-2})$; S common fit and monitor blocks cost $\widetilde{\mathcal{O}}(SQ_{\max}^2(\mathfrak{C}_{\mathrm{id}}+1)\Delta_0^{-2})$, plus any acquisition not already included. Comparisons against stored references affect computation and confidence logarithms, without an extra factor K in retained transitions. Knowing a separation lower bound does not supply an identity or assume a successful route.

D.5.3 Joint certificate interface used by the main theorem

Proposition 11 (Uniform routing and preservation certificates)**.** *Under Propositions 3 and 6, and the root-monitor budgets of Appendix D.9.2, the following hold on their joint event.*

(a) With diagnostic fit and monitor errors $\Delta_0/16$ each, every fitted-reference routing score has error at most $\Delta_0/8$. Reference transfer error at most $\Delta_0/8$ then makes threshold $\Delta_0/4$ separate matching from wrong families.

(b) With endpoint fit errors at most u and a fresh monitor of size $\widetilde{\mathcal{O}}(Q_{\max}^2u^{-2})$, the empirical one-sided preservation score differs from its population score by at most $2u$, uniformly over $\lambda \in [0,1]$. The calibrated solve, audit, and full-step test are simultaneously valid.

Proof idea. Positive-part projection is a contraction in L_2. Endpoint fit errors therefore control their entire affine hull, and a finite monitor grid covers data-dependent scalar selection. The empirical squared one-sided score is convex and piecewise quadratic, yielding a finite conservative solve that checks the full endpoint first (Appendix D.5.1). Routing uses the symmetric residual and diagnostic separation (Appendix D.5.2).

D.6 Own-update improvement, return transport, and anchored preservation

D.6.1 Partial BCPI and exact return transport

Fix one family and suppress its index. Throughout this subsection, $c = 1-\gamma$, $Q_{\max} = R_{\max}/c$, ξ is the initial state law, and $J(\pi) = \xi V^\pi$. Fix $\alpha \in (0,1)$ for the entire run, with $\nu_0(a \mid s) \geq \nu_{\min} > 0$, and let

$$\Pi_\alpha = \{(1-\alpha)\widetilde{\pi} + \alpha\nu_0 : \widetilde{\pi} \text{ is a measurable policy}\}, \qquad p_{\min} = \alpha\nu_{\min}.$$

With a standard Borel state space, finitely many actions, measurable kernels, and bounded rewards, the constrained Bellman maximization operator is a contraction on bounded measurable functions. Measurable finite-action maximization supplies a stationary optimizer π_α^* and its value V_α^*. Write $P_* = P^{\pi_\alpha^*}$; this comparator kernel is fixed throughout the family's analysis. All installed modules, including modules changed while this family is dormant, belong to Π_α.

Policy-averaged pointwise improvement Let $\Pi_\alpha = \{(1-\alpha)\widetilde{\pi} + \alpha\nu_0\}$ with $\nu_0(a \mid s) \geq \nu_{\min} > 0$. Fix α for the run. For one family let $\pi_* = \pi_\alpha^*$ be its stationary optimum in this class, $P_* = P^{\pi_*}$, and $V_* = V^{\pi_*}$. At own stage j, freeze the incumbent π_j^- and obtain a critic

$$\left\| \widehat{Q}_j - Q^{\pi_j^-} \right\|_{L^2(\rho_j \otimes \pi_j^-)} \leq \eta_{Q,j}. \tag{75}$$

The endpoint $\pi_j^{\mathrm{g}} \in \Pi_\alpha$ is greedy for the fitted critic up to pointwise error $0 \leq b_j(s) \leq b_g$, meaning

$$\max_{p \in \Pi_\alpha(s)} \sum_a p(a)\widehat{Q}_j(s,a) - \sum_a \pi_j^{\mathrm{g}}(a \mid s)\widehat{Q}_j(s,a) \leq b_j(s). \tag{76}$$

Define

$$e_j^p(s) = \sum_a p(a \mid s)|\widehat{Q}_j(s,a) - Q^{\pi_j^-}(s,a)|, \qquad a_j = b_j + e_j^{\pi_j^{\mathrm{g}}} + e_j^{\pi_j^-} + e_j^{\pi_*}.$$

For a chosen $\lambda_j \in [0,1]$, write

$$\pi_j^+ = (1-\lambda_j)\pi_j^- + \lambda_j \pi_j^{\mathrm{g}}, \quad P_j = P^{\pi_j^+}, \quad R_j = (I-\gamma P_j)^{-1}, \quad A_j = (1-\lambda_j)I + \gamma\lambda_j P_*.$$

Lemma 27 (Pointwise improvement with averaged defects)**.** *For* $\Delta_j^{\pm} = V_* - V^{\pi_j^{\pm}}$,

$$V^{\pi_j^-} - V^{\pi_j^+} \le \lambda_j R_j a_j, \qquad \Delta_j^+ \le A_j \Delta_j^- + \lambda_j R_j a_j. \tag{77}$$

Proof. Put $V = V^{\pi_j^-}$. Comparing the endpoint's fitted-critic expectation directly against π_j^- and against π_* yields

$$T^{\pi_j^{\mathrm{g}}} V \ge V - b_j - e_j^{\pi_j^{\mathrm{g}}} - e_j^{\pi_j^-} \ge V - a_j,$$

$$T^{\pi_j^{\mathrm{g}}} V \ge T^{\pi_*} V - b_j - e_j^{\pi_j^{\mathrm{g}}} - e_j^{\pi_*} \ge T^{\pi_*} V - a_j.$$

Bellman affinity gives $T^{\pi_j^+} V \ge V - \lambda_j a_j$, and the positive resolvent identity proves the first inequality. The second direct comparison gives $V_* - T^{\pi_j^+} V \le A_j(V_* - V) + \lambda_j a_j$. Add $\gamma P_j (V - V^{\pi_j^+})$, use the first inequality, and apply $I + \gamma P_j R_j = R_j$. □

Transport from the specified initial law Consider $N \ge 1$ own stages, indexed $0, \dots, N-1$. The horizon N may count attempted stages, including rollbacks, or accepted stages when so specified. The finite splitting algorithm has accepted own stages on its joint event; this deterministic statement also permits zero steps in other executions, with the actual harmful drift retained below. Define

$$q_j = 1 - c\lambda_j, \qquad U_j = A_{N-1} \cdots A_j, \qquad W_j = \prod_{\ell=j}^{N-1} q_\ell,$$
$$U_N = I, \qquad W_N = 1, \qquad \omega_j = \xi U_j / W_j. \tag{78}$$

All A_j are polynomials in the same kernel P_* and hence commute. They are positive operators with constant mass $q_j \ge \gamma > 0$; therefore each ω_j is a probability law. The harmful dormant changes and reporting suffix are

$$D_i = (V^{\pi_i^+} - V^{\pi_{i+1}^-})_+, \qquad 0 \le i \le N-2,$$
$$D_{\mathrm{fin}} = (V^{\pi_{N-1}^+} - V^{\pi^{\mathrm{report}}})_+. \tag{79}$$

For a probability law ζ, use the normalized discounted occupancy

$$d_\zeta^\pi = c\zeta(I-\gamma P^\pi)^{-1} = c\sum_{t\ge 0} \gamma^t \zeta (P^\pi)^t. \tag{80}$$

For probability laws σ, μ, set

$$\mathsf{h}(\sigma \mid \mu) = \begin{cases} \|d\sigma/d\mu\|_{L_2(\mu)}, & \sigma \ll \mu \text{ with an } L_2 \text{ density}, \\ +\infty, & \text{otherwise}. \end{cases} \tag{81}$$

When finite, $\mathsf{h}^2 = 1 + \chi^2(\sigma\|\mu)$ and $\sigma|f| \le \mathsf{h}(\sigma \mid \mu)\, \|f\|_{L_2(\mu)}$. These quantities describe the analysis, without assuming that the algorithm knows any transition probabilities.

Theorem 28 (Exact return transport for arbitrary steps)**.** *Put* $\zeta_j = d_{\omega_{j+1}}^{\pi_j^+}$. *For every realized sequence of installed steps and dormant changes,*

$$J(\pi_\alpha^*) - J(\pi^{\mathrm{report}}) \le \xi U_0 \Delta_0^- + \sum_{j=0}^{N-1} \frac{\lambda_j W_{j+1}}{c}\, \zeta_j a_j + \sum_{i=0}^{N-2} W_{i+1}\omega_{i+1} D_i + \xi D_{\mathrm{fin}}. \tag{82}$$

If $\left\|\Delta_0^-\right\|_\infty \le G_0$*, then*

$$\xi U_0 \Delta_0^- \le W_0 G_0 \le \exp\left(-c \sum_{j=0}^{N-1} \lambda_j\right) G_0. \tag{83}$$

For each nonzero step with finite indicated coefficient,

$$\zeta_j a_j \le \eta_{Q,j} \sum_{p \in \{\pi_j^{\mathrm{g}}, \pi_j^-, \pi_*\}} \mathsf{h}(\zeta_j \otimes p \mid \rho_j \otimes \pi_j^-) + \zeta_j b_j. \tag{84}$$

Zero-step active contributions are omitted before evaluating any coverage coefficient; no coverage condition is required for these contributions.

Proof. The pointwise inequality $\Delta_{i+1}^- \le \Delta_i^+ + D_i$ and Lemma 27 give, by induction and positivity,

$$\Delta_{N-1}^+ \le U_0 \Delta_0^- + \sum_{j=0}^{N-1} \lambda_j U_{j+1} R_j a_j + \sum_{i=0}^{N-2} U_{i+1} D_i.$$

Adding D_{fin} bounds the reported gap. Integrating against ξ and using (78)–(80) gives (82). The mass of U_0 is W_0 and $1 - c\lambda_j \le e^{-c\lambda_j}$, proving (83). Finally, for each of the three action laws, Cauchy–Schwarz on state–action pairs yields

$$\zeta_j e_j^p = (\zeta_j \otimes p)|\widehat{Q}_j - Q^{\pi_j^-}| \le \mathsf{h}(\zeta_j \otimes p \mid \rho_j \otimes \pi_j^-)\eta_{Q,j}.$$

Sum these three bounds and add $\zeta_j b_j$ to obtain (84). □

The laws ω_j and ζ_j can depend on later fitted critics and selected steps. This creates no independence requirement: the transport and norm comparisons hold deterministically on the simultaneous certificate event. Conditional confidence allocations are made before each fresh call. A deterministic numerical sample budget additionally needs coefficient envelopes valid over its capped class of admissible histories. The terms D_i are actual value changes: a certificate against an older anchor does not compare two arbitrary dormant candidates. The accepted-stage specialization below bounds D_i only through the unchanged anchor installed at the preceding own stage, including the reporting suffix.

Why preservation and progress are distinct Two elementary obstructions motivate the lower-step fallback and the one-sided score. In a one-state, two-action family with rewards one and zero, an exact greedy endpoint has action-one probability $p_* = 1 - \alpha/2$. Mixture steps satisfy

$$p_* - p_N = (p_* - p_0) \prod_{j<N} (1 - \lambda_j).$$

Taking $\lambda_j = 2^{-(j+2)}$ leaves a positive limiting product, since $\log(1-x) \ge -2x$ for $0 \le x \le 1/2$. Positive steps alone therefore need not give arbitrary accuracy. The splitting fallback instead ensures $\lambda_j \ge \lambda_{\min}$ at every accepted own stage.

For two one-state families with $r_i(1) = v_i > 0$, $r_i(0) = 0$, an exact anchor at probability p_0 has residual $V_i^{p_0} - T_i^{p'} V_i^{p_0} = v_i(p_0 - p')$. The optional symmetric score would restrict even a beneficial move to $|p' - p_0| \le \tau_i / v_i$. The one-sided score used in the analyzed algorithm vanishes for $p' \ge p_0$ and avoids this obstruction. Compatibility and certified fitting errors are still needed to prove full-step acceptance; these are supplied in the sharing result.

D.6.2 Joint discounted coverage and accepted-stage accuracy

The relevant laws consist of a comparator-policy prefix followed by one candidate-policy discounted occupancy. Assumption 2 requires the corresponding joint critic and state monitor envelopes. No closure under arbitrary sequences of candidate kernels is needed.

For a probability law ζ define $d_\zeta^\pi = c\zeta(I-\gamma P^\pi)^{-1}$ and $\mathsf{h}(\sigma \mid \mu) = \|d\sigma/d\mu\|_{L^2(\mu)}$, with value $+\infty$ when that density condition fails. Write $\xi_n = \xi P_*^n$ and

$$\begin{aligned}\kappa_Q^{\mathrm{SA}}(n) &= \sup_j \max_{p\in\{\pi_j^g,\pi_j^-,\pi_*\}} \mathsf{h}(d_{\xi_n}^{\pi_j^+}\otimes p \mid \rho_j\otimes\pi_j^-),\\ H_Q^{\mathrm{SA}} &= c\sum_{n\ge 0}\gamma^n\kappa_Q^{\mathrm{SA}}(n),\\ \kappa_M(n) &= \sup_i \mathsf{h}(d_{\xi_n}^{\vartheta_i} \mid \mu_i), \qquad H_M = c\sum_{n\ge 0}\gamma^n\kappa_M(n).\end{aligned} \tag{85}$$

The monitor supremum includes the anchor–candidate comparisons needed between own updates and at reporting. Uniform numerical budgets use deterministic envelopes over all permitted histories, including those created by splitting.

A coefficientwise identity for variable steps

Lemma 29 (Polynomial budget)**.** *For the horizon N in Theorem 28, let*

$$f_j(z) = 1-\lambda_j+\gamma\lambda_j z, \qquad u_j(z) = \prod_{\ell=j}^{N-1} f_\ell(z), \qquad u_N(z) = 1.$$

For coefficientwise power-series comparison $\preceq$,

$$S_N(z) := \sum_{j=0}^{N-1}\lambda_j u_{j+1}(z) = \frac{1-u_0(z)}{1-\gamma z} \preceq \frac{1}{1-\gamma z}. \tag{86}$$

The coefficients of S_N are nonnegative. If $\lambda_j \ge \underline{\lambda} > 0$ for all stages, then

$$\sum_{i=0}^{N-2} u_{i+1}(z) + 1 = \sum_{j=0}^{N-1} u_{j+1}(z) \preceq \frac{1}{\underline{\lambda}(1-\gamma z)}. \tag{87}$$

Proof. Since $(1-\gamma z)\lambda_j u_{j+1}(z) = u_{j+1}(z) - u_j(z)$, summation proves the equality. Every coefficient in the defining sum for S_N is nonnegative. Moreover, the difference between $(1-\gamma z)^{-1}$ and S_N equals $u_0(z)/(1-\gamma z)$ and also has nonnegative coefficients. Finally, $u_{j+1} \preceq \lambda_j u_{j+1}/\underline{\lambda}$ term by term. These statements compare individual coefficients, rather than only values at $z = 1$. □

Theorem 30 (Averaged-error accuracy bound for accepted own stages)**.** *Suppose the family has $J \ge 1$ accepted own updates, $\lambda_j \ge \lambda_{\min} > 0$, critic certificates* (75) *with $\eta_{Q,j} \le \eta_Q$, endpoint defects bounded by b_g, and a certified new anchor immediately after each. Every later dormant policy change affecting this family is certified directly against that named anchor with one-sided root threshold τ. Retain snapshot error η_V and transfer constant A_V from Theorem 33, and set $z = \tau + A_V\eta_V$. If* (85) *is finite, then at final reporting*

$$\boxed{\begin{aligned}J(\pi^*) - J(\pi^{\mathrm{report}}) \le G_0 e^{-c\lambda_{\min}J} + \frac{3H_Q^{\mathrm{SA}}\eta_Q}{c^2} + \frac{b_g}{c^2}\\ + \frac{H_M z}{c^2\lambda_{\min}} + \frac{2\alpha Q_{\max}}{c},\end{aligned}} \tag{88}$$

where G_0 bounds $\left\|V_* - V^{\pi_0^-}\right\|_\infty$ and one may use $G_0 = 2Q_{\max}$. Policy-preserving splits do not add an error term.

Proof. Let

$$f_j(t) = 1-\lambda_j+\gamma\lambda_j t, \qquad u_j(t) = \prod_{\ell=j}^{J-1} f_\ell(t), \qquad u_J = 1, \qquad U_j = u_j(P_*).$$

Lemma 29, applied with $N = J$, supplies the coefficientwise bounds (86) and (87). These identities hold for the realized step sequence, without independence assumptions on its coefficients.

Define harmful dormant changes $D_i = (V^{\pi_i^+} - V^{\pi_{i+1}^-})_+$ for $0 \le i < J-1$ and $D_{\mathrm{fin}} = (V^{\pi_{J-1}^+} - V^{\pi^{\mathrm{report}}})_+$. Iterate Lemma 27, insert $\Delta_{i+1}^- \le \Delta_i^+ + D_i$, and use positivity to get

$$J(\pi_*) - J(\pi^{\mathrm{report}}) \le \xi U_0 \Delta_0^- + \sum_{j=0}^{J-1} \lambda_j \xi U_{j+1} R_j a_j + \sum_{i=0}^{J-2} \xi U_{i+1} D_i + \xi D_{\mathrm{fin}}.$$

The initial term is at most $G_0 \prod_j (1 - c\lambda_j) \le G_0 e^{-c\lambda_{\min} J}$. For each of the three action laws, Cauchy–Schwarz on state–action pairs gives

$$\xi_n R_j e_j^p = c^{-1} (d_{\xi_n}^{\pi_j^+} \otimes p) |\widehat{Q}_j - Q^{\pi_j^-}| \le c^{-1} \kappa_Q^{\mathrm{SA}}(n) \eta_Q.$$

Expand U_{j+1} in P_* and use the first coefficientwise bound in (86); the three critic terms sum to at most $3H_Q^{\mathrm{SA}} \eta_Q / c^2$. The endpoint term is at most $b_g(1 - \prod_j (1 - c\lambda_j))/c^2 \le b_g / c^2$.

The anchor after own stage i is exactly π_i^+. At the next own incumbent, either the family policy is unchanged or the last affecting shared update supplied a direct certificate against that same anchor. Theorem 33 therefore gives $\xi_n D_i \le z\kappa_M(n)/c$; the same holds for the final suffix. The coefficientwise bound in (87) bounds their total by $H_M z / (c^2 \lambda_{\min})$. Intermediate dormant updates are not summed individually. A split changes neither copied nor remaining policies, so it leaves these identities and anchor relations unchanged.

Finally, the mixture of an unrestricted optimum with $\alpha\nu_0$ has Bellman loss at V^* at most $2\alpha Q_{\max}$. The positive resolvent has sup-norm operator norm $1/c$, giving $0 \le V^* - V_* \le 2\alpha Q_{\max}/c$. Add this exploration restriction cost. □

The action factor has been localized rather than eliminated numerically:

$$\mathsf{h}(d \otimes p \mid \rho \otimes \pi^-)^2 = \int \left(\frac{dd}{d\rho}\right)^2 \sum_a \frac{p(a \mid s)^2}{\pi^-(a \mid s)} \, d\rho(s). \tag{89}$$

A state-density bound alone can leave H_Q^{SA} of order $\alpha^{-1/2}$. With fixed $\alpha > 0$, old state-only coverage and the action floor imply finiteness of the new coefficient. A bound uniform as $\alpha \downarrow 0$ is a stronger numerical requirement. The proof has no independence requirement for future-dependent transport laws: all density comparisons are deterministic on the joint certificate event.

Proposition 12 (Discounted active-error bound without a lower step)**.** *For a finite arbitrary-step execution as in Theorem 28, suppose $\eta_{Q,j} \le \eta_Q$ and $\|b_j\|_\infty \le b_g$. Use the joint coefficient in* (85)*, omitting zero-step active terms before evaluating density coefficients. Then*

$$\sum_{j=0}^{N-1} \lambda_j \xi U_{j+1} R_j a_j \le \frac{3H_Q^{\mathrm{SA}} \eta_Q}{c^2} + \frac{b_g(1 - W_0)}{c^2}. \tag{90}$$

Proof. Expand $U_{j+1} = \sum_n \beta_{j+1,n} P_*^n$, with nonnegative coefficients. For each of the three action laws, $\xi_n R_j e_j^p \le \kappa_Q^{\mathrm{SA}}(n) \eta_Q / c$ by joint Cauchy–Schwarz. Lemma 29 gives $\sum_j \lambda_j \beta_{j+1,n} \le \gamma^n$, so the three errors sum to at most $3H_Q^{\mathrm{SA}} \eta_Q / c^2$. For b_j, use $R_j \mathbf{1} = c^{-1}\mathbf{1}$ and

$$\sum_j \lambda_j W_{j+1} = (1 - W_0)/c, \tag{91}$$

which follows by evaluating the polynomial identity at one. All expansions have nonnegative coefficients and bounded functions. □

State-only coefficients as sufficient specializations To display the action-floor dependence, define

$$\kappa_Q^{\mathrm{state}}(n) = \sup_{j:\lambda_j > 0} \mathsf{h}(d_{\xi_n}^{\pi_j^+} \mid \rho_j), \qquad H_Q^{\mathrm{state}} = c \sum_{n \ge 0} \gamma^n \kappa_Q^{\mathrm{state}}(n). \tag{92}$$

The monitor coefficient remains

$$H_M = c \sum_{n \geq 0} \gamma^n \kappa_M(n), \tag{93}$$

with κ_M defined in (85). An incumbent action floor $p_{\min} > 0$ gives, by (89),

$$H_Q^{\mathrm{SA}} \leq H_Q^{\mathrm{state}} / \sqrt{p_{\min}}. \tag{94}$$

Indeed $\sum_a p(a \mid s)^2 / \pi^-(a \mid s) \leq p_{\min}^{-1}$; insert this in the joint density identity and sum discounted coefficients. This proves finiteness at fixed exploration, not a bound independent of accuracy when $p_{\min} = \Theta(\varepsilon)$. A fixed-floor logger changes replay overlap, but does not change this incumbent-law critic factor.

D.6.3 Value snapshots and anchor-based return preservation

Lemma 31 (Value snapshot from an anchor critic)**.** *Freeze the anchor policy $\bar{\pi}$ and let $\bar{\rho}$ be its evaluation state law. If the certified critic obeys $\left\| \widehat{Q} - Q^{\bar{\pi}} \right\|_{L_2(\bar{\rho}\bar{\pi})} \leq \eta_V$, then the stored value $\widehat{V}(s) = \sum_a \bar{\pi}(a \mid s) \widehat{Q}(s,a)$ satisfies*

$$\left\| \widehat{V} - V^{\bar{\pi}} \right\|_{L_2(\bar{\rho})} \leq \eta_V. \tag{95}$$

Proof. Since $V^{\bar{\pi}}(s) = \sum_a \bar{\pi}(a \mid s) Q^{\bar{\pi}}(s,a)$, Jensen's inequality bounds the square of the value error at s by $\sum_a \bar{\pi}(a \mid s) |\widehat{Q}(s,a) - Q^{\bar{\pi}}(s,a)|^2$. Integration proves the claim. The action probabilities are those of the anchor itself, so no inverse action-probability factor is needed. □

For a candidate policy ϑ and monitor law μ, retain the separate snapshot-transfer conditions

$$\mu \leq C_0 \bar{\rho}, \qquad \mu P^{\vartheta} \leq C_1 \bar{\rho}, \qquad A_V = \sqrt{C_0} + \gamma \sqrt{C_1}. \tag{96}$$

These are domination statements between measures. Discounted occupancy coverage for return transport does not by itself imply them.

Lemma 32 (Snapshot residual transfer)**.** *Under* (96)*, if $\|e\|_{\bar{\rho}} \leq \eta_V$, then $\left\|(I - \gamma P^{\vartheta}) e\right\|_{\mu} \leq A_V \eta_V$.*

Proof. Measure domination gives $\|e\|_{\mu} \leq \sqrt{C_0} \eta_V$. Jensen's inequality gives $\left\|P^{\vartheta} e\right\|_{\mu}^2 \leq \mu P^{\vartheta}(e^2) \leq C_1 \eta_V^2$. Apply the triangle inequality. □

Theorem 33 (One-sided anchored return preservation)**.** *Let $\bar{\pi}$ have snapshot $\left\| \widehat{V} - V^{\bar{\pi}} \right\|_{\bar{\rho}} \leq \eta_V$, and suppose* (96) *holds for the candidate ϑ. If $\left\| (\widehat{V} - T^{\vartheta} \widehat{V})_+ \right\|_{\mu} \leq \tau$, then for every probability law ζ with finite displayed density norm,*

$$\begin{aligned} \left\| (V^{\bar{\pi}} - T^{\vartheta} V^{\bar{\pi}})_+ \right\|_{\mu} &\leq z := \tau + A_V \eta_V, \\ \zeta (V^{\bar{\pi}} - V^{\vartheta})_+ &\leq \frac{z}{c} \, \mathsf{h}(d_\zeta^{\vartheta} \mid \mu), \qquad d_\zeta^{\vartheta} = c \zeta (I - \gamma P^{\vartheta})^{-1}. \end{aligned} \tag{97}$$

Here $\mathsf{h}(\sigma \mid \mu) = \|d\sigma / d\mu\|_{L^2(\mu)}$.

Proof. The exact-anchor and snapshot residuals differ by $(I - \gamma P^{\vartheta})(V^{\bar{\pi}} - \widehat{V})$. Lemmas 32 and 21 prove the first bound. With $b = V^{\bar{\pi}} - T^{\vartheta} V^{\bar{\pi}}$, Bellman identities give $V^{\bar{\pi}} - V^{\vartheta} = (I - \gamma P^{\vartheta})^{-1} b$. The inverse is a positive kernel series, so $(V^{\bar{\pi}} - V^{\vartheta})_+ \leq (I - \gamma P^{\vartheta})^{-1} b_+$. Integrate against ζ and use Cauchy–Schwarz with respect to μ. □

Remark 34 (Absolute change under the symmetric variant)**.** *If the certificate instead bounds $\left\| \widehat{V} - T^{\vartheta} \widehat{V} \right\|_{\mu}$, the same proof with absolute values gives $\zeta |V^{\bar{\pi}} - V^{\vartheta}| \leq c^{-1} \mathsf{h}(d_\zeta^{\vartheta} \mid \mu)(\tau + A_V \eta_V)$. The primary one-sided gate controls deterioration and permits gains.*

The statistical premise is supplied by the uniform root certificate in Theorem 23. It protects the named anchor and holds for every covered test law on the same event; no extra union bound over those laws is needed. Approximation error is already included in the regression radius used to certify τ. It must not be omitted, and must not be added a second time after it has been budgeted into that radius. The replay law need not equal the snapshot evaluation law; $C_0 = 1$ is available only if the laws actually coincide.

Named anchors and final reporting At discovery, certify the actual assigned policy before its first dormant interval. Immediately after every installed own update, freeze and evaluate the post-update policy and replace only that family's named anchor, snapshot, and evaluation law. A pre-update critic is not a post-update value certificate. During later updates by other families, keep this anchor fixed and test every affected dormant record directly against it. Completed records remain protected until final reporting.

Let π_i^+ be the anchor installed after own update i. At the next own incumbent, or at reporting, the policy is either unchanged or the last accepted dormant candidate tested against that same anchor. Thus the corresponding harmful change is bounded by Theorem 33. Intermediate changes are not accumulated one by one. The polynomial proof in Theorem 30 includes both all between-update changes and the final dormant suffix. An exact-copy split changes neither the copied action law nor any remaining module's policy, and introduces no additional preservation error.

The required monitor-only audit precedes installation and uses fresh states with the full radius (67). It depends on the endpoint regression event. A failed audit invokes singleton fallback.

Accepted-update indexing is essential for the one-sided guarantee. Two policies can both improve an older anchor while losing value between themselves. For example, in a one-state MDP with rewards $r(1) = 1, r(0) = 0$, an anchor with action-one probability $\alpha/2$ has zero positive residual for both probabilities $1 - \alpha/2$ and $\alpha/2$, although moving between them loses $(1 - \alpha)/c$. Refreshing the named anchor immediately after each accepted own update prevents this stale-anchor comparison in the accuracy proof; it does not protect every historical policy simultaneously.

D.7 Module splitting with finite replay reserves

Provide K available policy-module slots and set the record cap $K_{\mathrm{cap}} = K$. Prescribe a per-module family capacity $1 \le h \le K$, an own-update quota J, and a movement threshold $\lambda_{\min} \in (0, 1]$. Here h is a *chosen capacity bound*, not a post hoc realized maximum. Let $g_t(k)$ be the time-dependent module of family record k. A newly discovered family joins the first existing module with spare capacity, in creation order; only if none exists is a new module created. Joining leaves the occupied module's policy unchanged and requires a certified initial anchor for the new family. Existing modules are never merged. This assignment uses no group labels or compatibility oracle.

Modules must have independent trainable dependencies for their deployed policies. An unchanged frozen feature map may be common. Splitting a head while continuing to train a shared trunk does not isolate the policies. All certificate-regression networks are separate from deployed actor parameters; fitting a certificate must not change the tested policies.

D.7.1 One finite collection per family

For each family, choose a fixed logger β_i with floor at least $p_{\log}$ and fixed monitor law μ_i. At discovery, while the family is active, first certify the assigned policy's initial named anchor and snapshot before any dormant interval. Also acquire and isolate

$$B_{\mathrm{res}} = (h - 1)J \tag{98}$$

complete sealed certification blocks. One block contains the labeled records for its two endpoint regressions, solve monitor inputs, and mandatory audit monitor inputs. For $h = 1$, collect no preservation reserve. The family uses this inventory across all later anchors. It is not refilled wholesale after every snapshot refresh.

Use Lemma 4 and Proposition 1, including entry, inter-record, and exit isolation for trajectory collection. In particular, a collection of N retained tuples under a frozen logger with $\sup_s \mathrm{TV}(P_\beta^b(s, \cdot), \mu) \le \kappa_{\mathrm{mix}} \rho^b$ costs $N + (N + 1)b$ transitions and has joint coupling error at

most $(N+1)\kappa_{\mathrm{mix}}\rho^b$, including its exit state. Future visible evolution may not otherwise access the hidden collection. Unused blocks are charged at acquisition. The snapshot evaluation law may change, so the state-law transfer constants must be controlled between this fixed μ_i and every later anchor's evaluation law.

D.7.2 AN ACTIVE SLOT

Use Algorithm 1. The active incumbent, its critic, the exploratory greedy endpoint, all protected snapshots, and both endpoint regression classes are fixed before opening the next sealed blocks. Endpoint fits are frozen before solve monitors are disclosed. A shared attempt first checks $U_i(1) \le 3\tau_i/4$ for every protected record and returns $\lambda = 1$ whenever these inequalities are certified. Otherwise the finite convex piecewise-quadratic procedure of Theorem 23 seeks a conservatively feasible $\lambda \ge \lambda_{\min}$. No such output, a required reserve block unavailable before the attempt, or a failed mandatory audit triggers an immediate independent singleton copy and a full step. Previously consumed data and computation remain charged.

For a sealed shared candidate, the fresh monitor-only audit evaluates $\widehat{R}^{+}_{a,i} = \|(\widehat{V}_i - \widehat{g}_i^{\lambda})_+\|_{m_{a,i}}$ and accepts only if $\widehat{R}^{+}_{a,i} + e_{a,i} \le \tau_i$ for every protected record. Here $e_{a,i} \le \tau_i/8$ includes the endpoint regression error as well as monitor error; it is not solely a sampling radius for the fitted function. The solver slack guarantees audit passage on the simultaneous events of Theorem 23. After a shared or singleton installation, evaluate the new policy in the active family, refresh only its named anchor and snapshot, and increment its own-update count.

The critic and endpoint can be reused after a split because the copied incumbent is the same policy in the same active family. The copy retains the old named anchor until the ensuing accepted own update is evaluated. An older snapshot is not relabeled as the value of a different incumbent.

Completed records receive no additional own training slots, but every later shared update affecting them continues to protect their named anchors until final reporting. The final certificate is the last such direct anchor certificate, not a fresh dormant-family query.

Theorem 35 (Progress and bounded splitting without compatibility)**.** *On the joint statistical certificate event, every active slot above installs one update with step at least $\lambda_{\min}$. With K correctly identified families, there are at most $K-1$ splits, at most K occupied modules, and no policy change caused by the copy itself. If every family receives J sufficiently long active slots, then*

$$\begin{gathered} L_{\mathrm{critic}} \le KJ, \qquad L_{\mathrm{solve}} \le KJ + K - 1, \qquad P_{\mathrm{protected}} \le K(h-1)J, \\ R_{\mathrm{snapshot}} \le K(J+1), \qquad B_{\mathrm{collected}} = K(h-1)J. \end{gathered} \tag{99}$$

The last quantity counts all collected certification blocks, including unused ones. No common feasible-margin or compatible-partition assumption is required.

Proof. A singleton has no dormant constraints and hence admits $\lambda = 1$. A shared slot either certifies $\lambda \ge \lambda_{\min}$ or becomes singleton. The margins of Theorem 23 guarantee audit passage on the joint event. Each slot therefore has exactly one accepted own update.

Let M_t and K_t be the occupied module and record counts. A split of a nonsingleton increases M_t by one and does not empty its old module. Discovery increases K_t by one and increases M_t by zero or one. No operation decreases the module count by merging. Whenever a shared module must split, $M_t < K_t \le K$, so a free slot exists. Starting from the first record, $1 \le M_t \le K_t$; thus no more than $K-1$ splits can occur. This argument allows a previously split family to share with a later discovery and split again: the total potential still bounds all splits.

Each family has at most J slots and one critic per slot. A split causes at most one extra, vacuous singleton solve and uses the same critic. At most one nonvacuous shared solve occurs per slot, and it protects at most $h-1$ records. Initial anchors and post-own-update anchors give $K + KJ$ snapshot evaluations. Finally (98) is collected exactly once per record, so its total is exactly $K(h-1)J$ even if blocks remain unused. If a particular family consumes its own allotment early, the exhaustion branch prevents further demand on that exhausted reserve. No future refill or hidden K^2 inventory is needed. Exact copying leaves the action law unchanged pointwise and hence preserves its value; remaining policies on the old module are unchanged. □

Any conservatively certified scalar above $\lambda_{\min}$ suffices. Numerical failure to find such a scalar may cause an unnecessary split, but does not invalidate progress or the split bound. Thus exact maximization over the feasible interval is unnecessary for this theorem. In contrast, the sharing guarantee uses the explicit endpoint-first test: under Theorem 38, the full endpoint satisfies every solve inequality and is returned before an interior search is attempted. That conditional conclusion requires a successful implementation of the certified endpoint test; an arbitrary numerical failure still permits progress by splitting, but cannot establish a no-split claim. With $h = K$, discovery also retains a single module in the compatible setting of Corollary 40.

The theorem trades a compatibility assumption for storage and architectural freedom up to one module per family. It does not guarantee a small final number of modules, or infer sufficient active time from recurrence alone. For failure histories use the declared $K_{\mathrm{cap}} = K$ and enforce the capacity and call caps before collecting or fitting additional data.

D.8 Compatible sharing and zero splits

The general progress theorem requires no compatibility. The following additional condition permits all own updates to share a full step.

Assumption 4 (S1: compatible directions and cross-family transfer)**.** *With supplied constants $a_{\min} > 0$ and $0 < \kappa_{\mathrm{tr}} < \infty$, for every distinct active/protected pair k, i in a proposed group, every permitted incumbent π and named anchor $\bar{\pi}_i$, there is a measurable action $a^{\dagger}_{k,\pi}$ satisfying*

$$Q^{\pi}_k(s, a^{\dagger}_{k,\pi}(s)) - \max_{b \neq a^{\dagger}_{k,\pi}(s)} Q^{\pi}_k(s, b) \geq a_{\min} \qquad \rho_{k,\pi}\text{-a.e.}, \tag{100}$$

$$a^{\dagger}_{k,\pi}(s) \in \arg\max_a Q^{\bar{\pi}_i}_i(s, a) \qquad \mu_i\text{-a.e.}, \tag{101}$$

$$\mu_i \leq \kappa_{\mathrm{tr},i,k}\rho_{k,\pi}, \qquad \kappa_{\mathrm{tr},i,k} \leq \kappa_{\mathrm{tr}}. \tag{102}$$

These are prescribed uniform envelopes over histories before a first split under the grouping. Requiring them over all policies and anchors in Π_α is a sufficient stronger condition. All members use the same exploration component $\alpha\nu_0$.

The condition aligns active greedy actions with protected anchor values; a common optimal policy alone does not assert this alignment or cross-law transfer. The latter is separate from within-family coverage.

Supplied compatibility constants. The algorithm does not test Assumption 4. When the module-count conclusion is requested, it is supplied with a margin lower bound $a_{\min} > 0$ and a transfer upper bound $\kappa_{\mathrm{tr}} < \infty$, in the same role as the diagnostic separation Δ_0. These constants enter the algorithm only through Calibration S's active-critic precision (118).

Remark 36 (Misspecified compatibility constants)**.** *If Assumption 4 fails, or holds only with a smaller margin or a larger transfer constant than supplied, the no-split conclusion of Theorem 39 can fail. Under the general execution and evaluator assumptions at the selected budgets, accuracy, completion of own slots, and the caps of K modules and $K-1$ splits are unaffected: Theorem 1 does not use Assumption 4, and Calibration S only tightens the active-critic precision. This does not waive realization or conditioning requirements at that tighter precision.*

Lemma 37 (Greedy disagreement)**.** *Let $\widehat{a}$ be the active critic's greedy action and suppose $\left\|\widehat{Q}_k - Q^{\pi^-}_k\right\|_{L_2(\rho_{k,\pi^-} \otimes \pi^-)} \leq \eta_Q$. Under Assumption 4, the set $\mathcal{B} = \{s : \widehat{a}(s) \neq a^{\dagger}_{k,\pi^-}(s)\}$ satisfies*

$$\rho_{k,\pi^-}(\mathcal{B}) \leq \frac{4\eta_Q^2}{a_{\min}^2 p_{\min}}, \qquad \mu_i(\mathcal{B}) \leq \frac{4\kappa_{\mathrm{tr},i,k}\eta_Q^2}{a_{\min}^2 p_{\min}}. \tag{103}$$

Proof. On $\mathcal{B}$, true action separation and fitted greediness imply $a_{\min} \leq |e_Q(s, a^{\dagger})| + |e_Q(s, \widehat{a})|$, where $e_Q = \widehat{Q}_k - Q^{\pi^-}_k$. Consequently

$$\mathbf{1}_{\mathcal{B}} \leq 4a_{\min}^{-2} \max_a |e_Q(s, a)|^2 \leq \frac{4}{a_{\min}^2 p_{\min}} \sum_a \pi^-(a \mid s)|e_Q(s, a)|^2.$$

Integrate under ρ_{k,π^-}, then apply (102). □

Theorem 38 (Direct full-step certificate)**.** *Under Assumption 4 and the prescribed critic and snapshot certificates, the full endpoint satisfies*

$$R_i^+(1) \le A_{V,i}\eta_{V,i} + \frac{4(1-\alpha)Q_{\max}\sqrt{\kappa_{\mathrm{tr},i,k}}}{a_{\min}\sqrt{p_{\min}}}\eta_Q. \tag{104}$$

Proof. Put $\bar{q}_i = Q_i^{\bar{\pi}_i}$ and $\bar{\pi}_i = (1-\alpha)\widetilde{\bar{\pi}}_i + \alpha\nu_0$. The exact anchor residual of the full endpoint is

$$V_i^{\bar{\pi}_i} - T_i^{\pi^{\mathrm{g}}} V_i^{\bar{\pi}_i} = (1-\alpha)\left\{\sum_a \widetilde{\bar{\pi}}_i(a \mid s)\bar{q}_i(s,a) - \bar{q}_i(s,\widehat{a}(s))\right\}. \tag{105}$$

By alignment this is nonpositive off $\mathcal{B}$, and on $\mathcal{B}$ it is at most $2(1-\alpha)Q_{\max}$. Its positive-part root norm is therefore at most $2(1-\alpha)Q_{\max}\sqrt{\mu_i(\mathcal{B})}$. Replacing the exact value by the snapshot changes the residual by $(I - \gamma P_i^{\pi^{\mathrm{g}}})(\widehat{V}_i - V_i^{\bar{\pi}_i})$. Apply Lemmas 32, 21, and 37. This compares the endpoint directly to the anchor; there is no accumulation over intervening dormant updates. □

Set

$$C_{\mathrm{sh}} = 4(1-\alpha)Q_{\max}\sqrt{\kappa_{\mathrm{tr}}}/a_{\min}. \tag{106}$$

Theorem 39 (No splits under compatible assignments)**.** *Suppose the encountered families form G nonempty groups of size at most h satisfying Assumption 4. Each group's first record creates one module and its later members join that module; modules do not mix groups. Suppose routing, evaluations, uniform solve certificates, and mandatory audits all hold, each family supplies its $(h-1)J$ complete sealed blocks and at most J own slots, and*

$$A_{V,i}\eta_{V,i} + C_{\mathrm{sh}}\eta_Q/\sqrt{p_{\min}} \le \tau_i/2 \quad \textit{for every protected } i. \tag{107}$$

Then every own slot installs $\lambda = 1$, no split occurs, and the final number of deployed modules is G.

Proof. Consider a hypothetical first split. Before it, each module contains only its assigned group's members. A protected record can be tested at most $(|\mathcal{G}_g| - 1)J \le (h-1)J$ times across the entire run. Before any required test, fewer than this many blocks have been consumed, so a complete unused block exists. This inventory is not reset at anchor refreshes. By Theorem 38, every protected full-endpoint score is at most $\tau_i/2$. Theorem 23 makes 1 feasible; the endpoint-first rule chooses it and its audit passes. Singletons also take a full step. Thus neither insufficient movement, audit failure, nor reserve shortage can trigger that first split. Each group creates exactly one module at discovery and no split creates another. A reserve may become empty after its last required use. □

Corollary 40 (One compatible group without a grouping oracle)**.** *If all K encountered families form one compatible group, choose $h = K$ and assign a new record to an existing module with spare capacity before creating a module. Under the remaining conditions of Theorem 39, one module serves all K families with full own steps and zero splits.*

Proof. The first record creates a module. Until any first split, its capacity accommodates every subsequent record, so the rule puts them all there. The assignment premise holds with $G = 1$, ruling out the first split. □

For multiple groups the assignment premise is conditional: a passing discovery test does not identify compatibility over future histories. The single-group result avoids this premise. Its resource benefit is the number of deployed common policy maps. Family records, diagnostic references, anchors, snapshots, and provisioned reserves remain in the ledger; a module count does not determine total memory or sample cost.

D.9 Calibrations, simultaneous confidence, and observation accounting

D.9.1 Calibrations G and S

The selected widths, radii, initialization rules and sample sizes must jointly meet the structural requirements. Set $G_0 = 2Q_{\max}$ and $0 < \varepsilon < \min\{1, 8Q_{\max}/c\}$. The main theorem uses G; the sharing theorem uses S, changing only the active-critic precision.

Quantity	Calibration G	Calibration S
Exploration α	$c\varepsilon/(8Q_{\max})$	Same
Incumbent floor $p_{\min}$	$\alpha\nu_{\min}$	Same
Preservation tolerance τ	$c^2\lambda_{\min}\varepsilon/(10H_M)$	Same
Snapshot root error η_V	$\tau/(4A_V)$	Same
Active root error η_Q	$c^2\varepsilon/(12H_Q^{\mathrm{SA}})$	$\min\{\eta_Q^{\mathrm{G}}, \tau\sqrt{p_{\min}}/(8C_{\mathrm{sh}})\}$
Own-update quota J	$\max\left\{1, \left\lceil \frac{\log\max\{1, 4G_0/\varepsilon\}}{c\lambda_{\min}} \right\rceil\right\}$	
Diagnostic reference error	$\Delta_0/(8A_{\mathrm{id}})$	Same
Diagnostic approximation bound	$\Delta_0/48$	Same
Diagnostic fit/monitor errors	$\Delta_0/16$ each	Same
Endpoint approximation bound	$\tau/48$	Same
Protected endpoint/monitor errors	$\tau/16$ each	Same

Table 2: The two prescribed calibrations. Here $C_{\mathrm{sh}} = 4(1-\alpha)Q_{\max}\sqrt{\kappa_{\mathrm{tr}}}/a_{\min}$; only S uses the supplied sharing constants. Approximation bounds are included within the fit tolerances. Complete primitive budgets and failure allocations below are used at these tolerances.

D.9.2 Transaction and root-monitor budgets

Algorithm 1 freezes each target and class before label disclosure and each fitted predictor before monitor disclosure. Its full-endpoint-first solve and mandatory audit use the following budgets.

Exact policy mixtures and copies. Every deployed mixture is a mixture of action probabilities. Its exact expression can retain immutable policy components; finite-action expectations of network predictions require computation, not environment transitions. An exact copy reproduces the incumbent action law pointwise and isolates all trainable parameters that could later affect it, including a trainable trunk and mutable normalization statistics. A shared frozen feature map is allowed. Auxiliary fits cannot update deployed policy parameters. A numerical or approximate copy, actor-parameter interpolation, or distilled mixture needs a separate defect bound and is not covered by the zero-copy-error assertion.

For protected family i, require each endpoint target's nonlinear approximation error to be at most $\tau_i/48$, uniformly as specified in Appendix D.1.3. Apply Corollary 19 with root tolerance $\tau_i/16$ and its prescribed budgets at the same selected width and radius. This gives total fitted errors $r_{i,0}, r_{i,1} \le \tau_i/16$, including approximation, statistical, curvature/ridge, and optimization contributions. With

$$u_i = \tau_i/16, \quad N_i^{\mathrm{grid}} = \max\{1, \lceil 4Q_{\max}/u_i \rceil\}, \quad m_i \ge \frac{64Q_{\max}^2}{u_i^2}\log\frac{2(N_i^{\mathrm{grid}}+1)}{\delta_{\mathrm{hull},i}}, \tag{108}$$

the solve radius is $e_i = \max_a r_{i,a} + u_i \le \tau_i/8$. Thus the statistical event is uniform over every candidate scalar, including the scalar selected using the same monitor. There is no post-selection independence assumption about that scalar. For the default monitor-only audit, let

$$m_{a,i} \ge \frac{16Q_{\max}^2}{u_i^2}\log\frac{2}{\delta_{\mathrm{audit},i}}, \qquad e_{a,i} = \max_a r_{i,a} + 4Q_{\max}\sqrt{\frac{\log(2/\delta_{\mathrm{audit},i})}{m_{a,i}}} \le \tau_i/8. \tag{109}$$

The score is $\widehat{R}_{a,i} = \|(\widehat{V}_i - \widehat{g}_i^{\widehat{\lambda}})_+\|_{m_{a,i}}$ and the acceptance threshold is $\widehat{R}_{a,i} + e_{a,i} \le \tau_i$. A uniform feasible solve already implies true residual at most $3\tau_i/4$, so every such candidate passes on the common event. This audit uses new monitor states but the same regression events; it cannot independently detect class misspecification. The primary algorithm always performs this audit. Omitting it is a separate variant that accepts directly from the uniform solve.

For routing use the supplied Δ_0 and prescribe

$$t = \frac{\Delta_0}{4}, \quad \eta_{\mathrm{id}} = \frac{\Delta_0}{8A_{\mathrm{id}}}, \quad u_{\mathrm{id}} = \frac{\Delta_0}{16}, \quad n_{\mathrm{id}} \ge \frac{4096Q_{\max}^2}{\Delta_0^2}\log\frac{2\max\{1,K\}}{\delta_{\mathrm{route,mon}}}. \tag{110}$$

Corollary 11 provides the reference tolerance. For each stored reference, require nonlinear target approximation error at most $\Delta_0/48$ for every possible current family, including wrong-family

comparisons. With the prescribed budgets at the same selected width and radius, Corollary 19 gives total regression error at most $u_{\mathrm{id}} = \Delta_0/16$. The common monitor has root error at most $\Delta_0/16$. All fits share one labeled block, with a union allocation across references and a prescribed worst-case block size; this saves transitions, not regression computation. Proposition 10 derives simultaneous adequacy rather than assuming successful tests.

D.9.3 One simultaneous event with explicit failure allocations

Put $P_{\mathrm{cap}} = K(h-1)J$ and prescribe

$$E_{\mathrm{cap}} = K(2J+2), \qquad R_{\mathrm{cap}} = SK + 2P_{\mathrm{cap}}, \qquad M_{\mathrm{cap}} = S + 2P_{\mathrm{cap}}. \tag{111}$$

These bound complete evaluator calls (own critics, initial/post-update snapshots, and diagnostic references), auxiliary regression calls (diagnostic fits and preservation endpoints), and outer monitor events (common routing, hull, and audit), respectively. Unreached calls consume no probability. Evaluator replicas, initialization, median-distance validation, and the coupling of all retained training and validation tuples are internal to a complete evaluator allocation. They are not charged again to the separate acquisition pool. In particular, they are not silently omitted or counted as unamplified expected-error guarantees.

One concrete allocation for total confidence $1-\delta$ is:

Event class	Per-event failure allocation
Complete evaluator call	$\delta/(5\max\{1, E_{\mathrm{cap}}\})$
Complete auxiliary regression call	$\delta/(5\max\{1, R_{\mathrm{cap}}\})$
Outer routing, hull, or audit monitor	$\delta/(5\max\{1, M_{\mathrm{cap}}\})$
Additional auxiliary-envelope event	$\delta/(5\max\{1, A_{\mathrm{cap}}\})$
Separate acquisition $q = 1, 2, \ldots$	$\delta/[5q(q+1)]$

Here A_{cap} is any predeclared cap on additional random auxiliary architecture-envelope events not already included in a subroutine; set it to zero for deterministic envelopes. This row does not create an assumption that random envelopes succeed. Their probability must be separately established if they are used. The capped subroutine plan bounds the number of actual acquisitions. Since $\sum_{q\ge 1} 1/[q(q+1)] = 1$, their coupling allocations sum to at most $\delta/5$. This convenient allocation need not be tight.

Use the ideal product experiment for separately coupled auxiliary collectors. Complete evaluator guarantees already include their trajectory coupling and hold conditionally on each permitted history. The adaptive first-failure argument bounds evaluations, fits, and outer monitors by their first three pools. Any separately proved random auxiliary envelope uses the fourth pool; deterministic structural assumptions consume no probability. The isolation coupling for each outer acquisition uses the final pool, once for the entire corresponding collection. Thus all required events hold jointly on $\mathcal{E}$ with $\mathbb{P}(\mathcal{E}) \ge 1-\delta$. Hidden reserves are never conditioned on successful future tests. Uniformity over candidate scalars belongs to the solve event; deterministic transport requires no further failure allocation.

D.9.4 Complete numerical conditions and observation ledger

Dependencies and numerical calibration The three structural groups are Assumptions 3, 2, and 1. The observation protocol of Definition 2 and execution requirements of Definition 1 remain explicit premises. For clarity, the proof uses the following consequences and numerical conditions:

1. The frozen-policy TD routine of Corollary 11 provides active, diagnostic-reference, and snapshot errors at their prescribed tolerances. Its single-normalization architecture, local exact tangent realization, joint initialization event, conditioning, width, and spaced-trajectory requirements are retained. Its complete cost includes all training and validation gaps. Write its active and snapshot costs as $\widetilde{\mathcal{O}}(E_Q\eta^{-2})$ and $\widetilde{\mathcal{O}}(E_V\eta^{-2})$ with actual factors.
2. Each auxiliary class and Bellman target satisfy Definition 5, with nonlinear approximation error $a \le u/3$ for the prescribed fitted root-error tolerance u, together with the curvature/width, iteration, and sample inequalities in (54). These conditions hold at the same selected width and

radius, uniformly over the targets and permitted histories specified in Appendix D.1.3. These are inequalities on the selected architecture and budgets, not an assumed successful regression oracle. Actual tangent complexity is retained.

3. Joint critic coverage H_Q^{SA}, monitor coverage H_M, and the snapshot/diagnostic law dominations have the declared envelopes over every permitted history. The transfer constants are derived by the corresponding Jensen inequalities. Set $H_M = 1$ when no protected comparison is needed.
4. The known lower bound Δ_0 determines threshold $t = \Delta_0/4$ and the explicit budgets (110). Proposition 10 proves the resulting strict score separation. The auxiliary approximation bound $a \le \Delta_0/48$ holds for every possible current family, including wrong-family comparisons, using classes selected without its identity.
5. The observation protocol supplies enough active time for discovery, routing, references, the one-time isolated reserve collection, and J own slots per family with all their evaluation and acquisition work. The architecture implements exact probability mixtures and independent exact copies. Modules are never merged; K slots are available and $K_{\mathrm{cap}} = K$. All statistical and coupling failures are allocated before their observations are exposed and sum to at most δ.

No sharing-compatibility premise remains: Theorem 35 supplies the progress conclusion. If the calibrated auxiliary approximation bound or another calibration requirement fails, this primary theorem does not apply. The general regression and deterministic transport lemmas remain available with their actual achieved errors; no arbitrary-accuracy extension is asserted for evaluators outside Proposition 3.

Complete primitive budgets Under Assumptions 1, 2, and 3, the observation protocol, certified execution, and the single-normalization evaluator specification, use exact exploratory greedification. Use Calibration G in Table 2. Use the splitting/reserve algorithm of Appendix D.7. The following choices are prescribed before their data are exposed and must hold jointly at the selected widths and radii.

(a) **Complete evaluation.** At each root tolerance $\eta \in \{\eta_Q, \eta_V, \eta_{\mathrm{id}}\}$ set $r = \eta/5$. With A_T, A_1, A_2, A_3 defined in Appendix D.3.5, choose integers $T \ge 1$ and even m with

$$T + 1 \ge 48A_T/r^2, \qquad m \ge \max\{(48A_1/r^2)^2,\ 48(A_2H_{T+1}^2 + A_3)/r^2\}.$$

These choices additionally satisfy Assumption 3(a), admissible initialization, and the derivative bounds. For complete-call failure δ_q, use an odd $M \ge 32\log(4/\delta_q)$ and $n_{\mathrm{val}} \ge 32Q_{\max}^2 r^{-2}\log(8M^2/\delta_q)$. With $N = MT + n_{\mathrm{val}}$ choose g by (38). Charge $N + (N+1)g$ transitions, with its coupling inside δ_q.

(b) **Auxiliary fitting.** Use $u_{\mathrm{p}} = \tau/16$, $u_{\mathrm{id}} = \Delta_0/16$, and $\Lambda = p_{\log}^{-1}$ for preservation or $\Lambda = 1$ for diagnostics. For each call, require $a \le u/3$ with $u = u_{\mathrm{p}}$ or $u = u_{\mathrm{id}}$, respectively: the nonlinear approximation allowance is $\tau/48$ for each preservation endpoint and $\Delta_0/48$ for each diagnostic target. Set $\zeta = Q_{\max}/n$, $\omega = 8\Lambda Q_{\max}\chi$, and $M_{\mathrm{reg}} = 2G^2 + 3\omega/2$, and require

$$\chi \le \frac{u^2}{144\Lambda Q_{\max}R^2}, \quad T_{\mathrm{reg}} \ge \frac{6M_{\mathrm{reg}}R^2}{u^2}, \quad n \ge \frac{6\Lambda Q_{\max}^2}{u^2}\left[48\log\frac{2N_{\mathcal{K}}(Q_{\max}/n)}{\delta_r} + 40\right].$$

For the concrete auxiliary network require even $m \ge (144\Lambda Q_{\max}HR^2/u^2)^2$ and $GR \le Q_{\max}$ (Proposition 4), together with $a \le u/3$ at this same width and radius. Actual cover complexity is retained; deterministic doubling resolves the implicit sample inequality before data disclosure. Corollary 19 includes the approximation contribution $3a^2 \le u^2/3$ and gives total fitted root error at most u.

(c) **Monitors and routing.** Set $u_{\mathrm{mon}} = \tau/16$ and $N_{\mathrm{grid}} = \max\{1, \lceil 4Q_{\max}/u_{\mathrm{mon}}\rceil\}$. Use

$$n_{\mathrm{solve}} \ge \frac{64Q_{\max}^2}{u_{\mathrm{mon}}^2}\log\frac{2(N_{\mathrm{grid}}+1)}{\delta_{\mathrm{solve}}}, \qquad n_{\mathrm{audit}} \ge \frac{16Q_{\max}^2}{u_{\mathrm{mon}}^2}\log\frac{2}{\delta_{\mathrm{audit}}}.$$

Supply Δ_0, set $t = \Delta_0/4$, and use $\eta_{\mathrm{id}} = \Delta_0/(8A_{\mathrm{id}})$ with the diagnostic regression and common-monitor choices (110). The solve event is uniform over the scalar; the audit candidate is frozen before its own block.

(d) **Confidence and acquisition.** Impose $P_{\mathrm{cap}} = K(h-1)J$, $E_{\mathrm{cap}} = K(2J+2)$, $R_{\mathrm{cap}} = SK + 2P_{\mathrm{cap}}$, and $M_{\mathrm{cap}} = S + 2P_{\mathrm{cap}}$. Allocate $\delta/(5\max\{1, E_{\mathrm{cap}}\})$, $\delta/(5\max\{1, R_{\mathrm{cap}}\})$, and $\delta/(5\max\{1, M_{\mathrm{cap}}\})$ to complete evaluations, fits, and outer monitors. Reserve at most $\delta/5$ for separately proved random auxiliary envelopes and $\delta/[5q(q+1)]$ for outer acquisition q. Choose outer gaps by (15); do not charge evaluator coupling twice. The ordered active-time condition covers every collected block and all gaps, even unused inventory.

These are the complete numerical requirements used in the proof of Theorem 1; Appendix C.1 derives its simultaneous conclusions and the $7\varepsilon/8$ accuracy bound.

Let $\mathfrak{C}_{\mathrm{reg}}$ be a uniform bound on the actual certification complexity $\mathfrak{C}(u_{\mathrm{p}}, \Lambda)$ with $\Lambda = p_{\log}^{-1}$. A sufficient observation ledger, including unused replay reserves, is

$$\boxed{\begin{aligned} N_{\mathrm{total}} &\le N_{\mathrm{diag}} + N_{\mathrm{acq}} \\ &\quad + \widetilde{\mathcal{O}}\left(KJE_Q\eta_Q^{-2} + K(J+1)E_V\eta_V^{-2} + K(h-1)J\,\Lambda Q_{\max}^2(\mathfrak{C}_{\mathrm{reg}}+1)\tau^{-2}\right). \end{aligned}} \tag{112}$$

Here N_{acq} counts only acquisition costs not already included in the other terms. Off the joint event, the same deterministic cap uses the declared cap $K_{\mathrm{cap}} = K$ and enforces the prescribed block, record, module-capacity, and call limits.

An endpoint defect $b_g \le c^2\varepsilon/8$ may be allowed instead of exact greedification, using the remaining $\varepsilon/8$ in Theorem 30. This is distinct from the auxiliary regression optimization error, already included in its certificate.

Diagnostic and acquisition costs The supplied bound and (110) give reference error $\eta_{\mathrm{id}} = \Delta_0/(8A_{\mathrm{id}})$ and regression/monitor root errors $\Delta_0/16$. The actual gap need not equal the supplied lower bound; the budget is stated in Δ_0. Let $\mathfrak{C}_{\mathrm{id}}$ bound the trainable regression complexity for these state–action targets, with $\Lambda = 1$. Then

$$N_{\mathrm{diag}} = \widetilde{\mathcal{O}}\left[\{KE_{\mathrm{id}}A_{\mathrm{id}}^2 + SQ_{\max}^2(\mathfrak{C}_{\mathrm{id}}+1)\}\Delta_0^{-2}\right], \tag{113}$$

with any uncharged diagnostic input acquisition added to N_{acq}. All stored references use the same current-family labeled block. They incur a confidence logarithm and separate computation, not an extra multiplicative K in the transition count.

For trajectory-collected replay let the full family-i reserve use N_i retained one-step tuples, including tuples retained only for their monitor inputs. Its actual collection costs $N_i + (N_i+1)b_i$ transitions. Choose

$$(N_i+1)\kappa_{\mathrm{mix},i}\rho_i^{b_i} \le \delta_{\mathrm{mix},i}.$$

The N_i terms are already counted in (112); only excess spacing/entry/exit cost is added. With uniformly geometric mixing this inflates the reserve order by logarithms. Count training-start and diagnostic acquisition corrections once in the same manner. If these factors depend on accuracy or exploration, keep that dependence explicit. The assumption of enough active time includes the entire collection, not just the number of retained tuples.

Explicit rate and uniform-regularity corollaries For fixed $c, Q_{\max}, \nu_{\min}$, substituting Table 2 into (112) yields

$$\boxed{N_{\mathrm{total}} \le N_{\mathrm{diag}} + N_{\mathrm{acq}} + \widetilde{\mathcal{O}}\left(\frac{KE_Q(H_Q^{\mathrm{SA}})^2}{\lambda_{\min}\varepsilon^2} + \frac{KE_VA_V^2H_M^2}{\lambda_{\min}^3\varepsilon^2} + \frac{K(h-1)(\mathfrak{C}_{\mathrm{reg}}+1)H_M^2}{p_{\log}\,\lambda_{\min}^3\varepsilon^2}\right).} \tag{114}$$

Equation (112) retains the unabsorbed discount and reward dependence. Equation (114) retains the actual neural complexity and conditioning factors.

Corollary 41 (Conservative exploration-floor replay rate)**.** *Suppose $p_{\log} = \alpha\nu_{\min} = \Theta(\varepsilon)$, $\lambda_{\min}$ is fixed, and $H_Q^{\mathrm{SA}}, H_M, E_Q, E_V, A_V, \mathfrak{C}_{\mathrm{reg}}, \mathfrak{C}_{\mathrm{id}}, E_{\mathrm{id}}, A_{\mathrm{id}}$ and acquisition factors remain bounded*

independently of the accuracy parameters. Assume the evaluator's local exact tangent-realization condition and the auxiliary approximation bounds $a \le u/3$, with $u = u_{\mathrm{p}}$ for preservation and $u = u_{\mathrm{id}}$ for diagnostics, together with initialization success, derivative/radius envelopes, cover complexity, and width/training requirements, jointly at the selected widths and radii as both ε and Δ_0 vary. Also require sufficient ordered active time. These accuracy-uniform premises are stronger than finite bounds at one fixed tolerance. Then a sufficient total is

$$\boxed{N_{\mathrm{total}} = \widetilde{\mathcal{O}}\left(K\varepsilon^{-2} + K(h-1)\varepsilon^{-3} + (S+K)\Delta_0^{-2}\right).} \tag{115}$$

For the trainable two-layer regressor with fixed remaining constants, the preservation width is sufficient at order $m_{\mathrm{reg}} = \widetilde{\mathcal{O}}(\varepsilon^{-6})$ and diagnostic width at order $m_{\mathrm{id}} = \widetilde{\mathcal{O}}(\Delta_0^{-4})$.

Proof. Use (113) and (114); fixed mixing absorbs N_{acq} into the same orders. In (58), $u_{\mathrm{p}} = \Theta(\varepsilon)$ and $\Lambda = \Theta(\varepsilon^{-1})$ give $m_{\mathrm{reg}} = O(\varepsilon^{-6})$. Diagnostic $\Lambda = 1$ and $u_{\mathrm{id}} = \Theta(\Delta_0)$ give the other width. These widths satisfy the curvature requirement. The calibrated nonlinear approximation bounds and controlled tangent entropy must also hold at the selected widths and radii. □

Corollary 42 (Fixed-floor replay rate)**.** *Under the same uniform conditions, if a dedicated logger has a fixed $p_{\log} > 0$ independent of α, then*

$$N_{\mathrm{total}} = \widetilde{\mathcal{O}}\left(Kh\varepsilon^{-2} + (S+K)\Delta_0^{-2}\right). \tag{116}$$

The sufficient preservation regression width is then $\widetilde{\mathcal{O}}(\varepsilon^{-4})$.

Proof. In (114) the logging factor is now bounded, and the three control/certification terms have inverse-square accuracy order. Equation (58) gives the width. A logger such as $\beta_i = \nu_0$ has an accuracy-independent action floor, but its actual state-law coverage, mixing, and active collection must satisfy the stated conditions. This corollary does not infer those conditions from the action floor. □

The active neural TD and snapshot widths are those of the evaluator of Proposition 3 at η_Q and η_V. Its schematic root finite-width error $O(m^{-1/4})$ requires, for fixed structural constants, $m_Q \gtrsim (H_Q^{\mathrm{SA}})^4\varepsilon^{-4}$ and $m_V \gtrsim A_V^4 H_M^4 \lambda_{\min}^{-4}\varepsilon^{-4}$, together with its simultaneous local exact-realization, conditioning, derivative, and initialization requirements. These are separate networks and assumptions from the fully trainable auxiliary regressors.

D.9.5 Dependence audit and interpretation

The preceding theorems have the following scope and conditions, with different types of conditions:

- Fully trainable regression uses a bounded smooth network, controlled local parameter movement, enough width, explicit ridge, and a proved number of gradient steps. The prescribed root-error calibration uses $a \le u/3$ at the selected width and radius. The fast statistical corollary additionally requires controlled tangent entropy. Bellman approximation alone does not imply these optimization or complexity properties.
- Splitting removes the compatibility assumption by allowing up to K independent modules and by retrying in the same active slot. It requires actual isolation of every trainable policy dependency.
- The final theorem retains active evaluator regularity, diagnostic separation, state–action coverage, anchor transfer, and sufficient active-time acquisition. These do not follow from splitting.

The quadratic example of Proposition 5 shows that all weights may train while the regression function rank is independent of width. It does not establish that a generic smooth neural regressor has that property. For a general network, use its actual $N_{\mathcal{K}}$ or r in (53) and (114). A polynomially growing entropy factor changes the fast theorem's headline exponent.

For example, under $\alpha = \Theta(\varepsilon)$ suppose

$$H_Q^{\mathrm{SA}} = O(\alpha^{-u}), \quad H_M = O(\alpha^{-v}), \quad E_Q = O(\alpha^{-e_Q}), \quad E_V = O(\alpha^{-e_V}),$$
$$A_V = O(\alpha^{-a}), \quad \mathfrak{C}_{\mathrm{reg}} + 1 = O(\varepsilon^{-d_c}), \quad \lambda_{\min} = \varepsilon^r, \quad p_{\log}^{-1} = O(\varepsilon^{-\ell}).$$

Before additional acquisition factors, the sufficient exponents are

$$\begin{array}{ll} 2 + 2u + e_Q + r & \text{active critics,} \\ 2 + e_V + 2a + 2v + 3r & \text{snapshots,} \\ 2 + \ell + d_c + 2v + 3r & \text{preservation reserves.} \end{array} \tag{117}$$

This is an upper-budget calculation, not a lower bound. A state-density envelope that only yields $H_Q^{\mathrm{SA}} = O(\alpha^{-1/2})$ and $E_Q, E_V = O(\alpha^{-2})$ permits conservative active ε^{-5} and snapshot ε^{-4} terms, even with bounded regression complexity. It therefore does not establish Corollary 41 by itself.

The cubic replay term is likewise not an intrinsic consequence of having a replay buffer. It arises from the specified shrinking logging floor. Corollary 42 states exactly what improves when the logger has fixed action overlap.

D.9.6 Calibration S and the sharing cost

Calibration S uses all of Calibration G, including its α, τ, η_V, J, diagnostic threshold, confidence allocation, initialization requirements, and acquisition rules. Its sole change, for $h \geq 2$, is

$$\eta_Q^{\mathrm{S}} = \min \left\{ \underbrace{\frac{c^2 \varepsilon}{12 H_Q^{\mathrm{SA}}}}_{\eta_Q^{\mathrm{G}}}, \frac{\tau \sqrt{p_{\min}}}{8 C_{\mathrm{sh}}} \right\}, \qquad C_{\mathrm{sh}} = \frac{4(1-\alpha) Q_{\max} \sqrt{\kappa_{\mathrm{tr}}}}{a_{\min}}. \tag{118}$$

For $h = 1$ omit the sharing constraint. Here $a_{\min}$ and κ_{tr} are supplied positive envelopes used to set the budget; correctness of the general fallback does not require the instance to satisfy compatibility. The evaluator's local exact tangent-realization and conditioning premises apply at the newly selected active-critic width, radius, and precision. Auxiliary root tolerances remain $u_{\mathrm{p}} = \tau/16$ for preservation and $u_{\mathrm{id}} = \Delta_0/16$ for diagnostics. Each auxiliary call satisfies $a \leq u/3$ at its corresponding tolerance u, with the approximation, curvature, optimization, and statistical requirements holding jointly at the selected auxiliary width and radius. These conditions cover all permitted targets and histories under Calibration S as specified in Appendix D.1.3; the total fitted-error tolerances include the approximation contribution.

Proposition 13 (Calibration S preserves accuracy and ensures sharing)**.** *Under the general execution assumptions, Calibration S has the same simultaneous $7\varepsilon/8$ accuracy, routing, and progress conclusions as Calibration G. Under compatible assignments it additionally has full own steps, zero splits, and G deployed modules; for one compatible group join-before-create with $h = K$ suffices.*

Proof. The tighter active root tolerance preserves every inequality in Theorem 1. No other budget or failure class changes, and its $1 - \delta$ event remains valid under the prescribed new evaluator budgets. The four error terms are bounded by $\varepsilon/4, \varepsilon/4, \varepsilon/8, \varepsilon/4$, respectively. Under compatibility, Theorem 38 and Calibration S give

$$R_i^+(1) \leq \tau/4 + \tau/8 = 3\tau/8, \qquad U_i^+(1) \leq R_i^+(1) + 2e_i \leq 5\tau/8 < 3\tau/4. \tag{119}$$

The fresh mandatory audit has upper bound at most $R_i^+(1) + 2e_{a,i} \leq 5\tau/8 < \tau$. Thus Theorem 39 applies, or Corollary 40 for one group. The full-endpoint solve has at least $\tau/8$ slack; any numerical score allowance must be included within that slack. □

The same complete observation ledger (112) applies, including $K(h-1)J$ collected blocks even when none is needed after an early split. The changed active-critic cost is explicit:

$$(\eta_Q^{\mathrm{S}})^{-2} = \max \left\{ \frac{144 (H_Q^{\mathrm{SA}})^2}{c^4 \varepsilon^2}, \frac{64 C_{\mathrm{sh}}^2}{\tau^2 p_{\min}} \right\}, \qquad \eta_V^{-2} = 16 A_V^2 / \tau^2. \tag{120}$$

All training, confidence amplification, validation, and spacing costs are recomputed at η_Q^{S}; the number of active evaluations is still KJ. Auxiliary labels remain unbiased labels for the full Bellman target. Only the preservation score takes a positive part.

Sharing precision versus coverage. For $p_{\min} = \Theta(\varepsilon)$ and bounded C_{sh}, the second term in (120) is $O(\varepsilon^{-3})$, since Calibration G has $\tau = \Theta(\varepsilon)$ when the other calibration constants are fixed. With KJ active calls and uniformly controlled evaluator and acquisition factors, this adds a sufficient active-critic budget $\widetilde{\mathcal{O}}(K\varepsilon^{-3})$ independently of the coverage envelope H_Q^{SA}. In particular, bounded joint coverage does not remove this sharing precision requirement. The identity in (120) is a maximum: the combined active-critic cost can be larger when $(H_Q^{\mathrm{SA}})^2\varepsilon^{-2}$ is larger. Thus a total $\widetilde{\mathcal{O}}(K\varepsilon^{-3})$ active-critic bound also needs $(H_Q^{\mathrm{SA}})^2 = O(p_{\min}^{-1})$, as in the next corollary. These statements concern sufficient budgets, not lower bounds on samples needed for sharing.

Corollary 43 (Conditional rate for guaranteed sharing)**.** *Suppose $c, Q_{\max}, \nu_{\min}, \lambda_{\min}, a_{\min}$ have fixed positive bounds, $\kappa_{\mathrm{tr}}, H_M, A_V, A_{\mathrm{id}}$ are uniformly bounded, and $(H_Q^{\mathrm{SA}})^2 = O(p_{\min}^{-1})$. Suppose the actual evaluator factors, regression complexities, and acquisition inflation grow at most polylogarithmically in the inverse accuracies. Require the evaluator's local exact tangent-realization, conditioning, and initialization conditions, together with the auxiliary bounds $a \le u/3$ at $u \in \{u_{\mathrm{p}}, u_{\mathrm{id}}\}$ for the corresponding calls and their calibrated regression budgets, uniformly at the selected widths and radii as ε and Δ_0 vary. For $p_{\min} = \Theta(\varepsilon)$,*

$$N_{\mathrm{total}} = \widetilde{\mathcal{O}}\Big(K\varepsilon^{-3} + K(h-1)p_{\log}^{-1}\varepsilon^{-2} + (S+K)\Delta_0^{-2}\Big). \tag{121}$$

Proof. Calibration G gives $J = O(\log(1/\varepsilon))$, $\tau = \Theta(\varepsilon)$, $\eta_V = \Theta(\varepsilon)$. Equation (120) is $O(\varepsilon^{-3})$ under the stated coverage bound. Substitute in the complete ledger, retaining the actual factors and using their uniform bounds. The snapshot contribution $K\varepsilon^{-2}$ is dominated by $K\varepsilon^{-3}$. □

With an exploratory logger, (121) becomes $\widetilde{\mathcal{O}}(Kh\varepsilon^{-3} + (S+K)\Delta_0^{-2})$; with a fixed-floor logger it is $\widetilde{\mathcal{O}}(K\varepsilon^{-3} + K(h-1)\varepsilon^{-2} + (S+K)\Delta_0^{-2})$. The fixed logger changes $p_{\log}$, not the on-policy critic's $p_{\min}$. The extra precision is sufficient for this sharing proof; it is not a sample lower bound. Uniform complexity bounds are additional requirements beyond evaluator realization and auxiliary target approximation.

Deployed policies and retained memory. Only deployed modules are saved: records, snapshots, reserves, and immutable mixture components remain. Total-memory and observation savings do not follow. At $h = K$, the reserve factor $K(h-1)$ in the complete ledger is $K(K-1)$. Theorem 1 and Corollary 2 use Calibration G; the additional active-critic precision for guaranteed sharing is specific to Calibration S.

D.10 Concrete evaluator and compatible/conflicting constructions

D.10.1 A compatible single-normalization evaluator

This construction verifies that the evaluator geometry can hold for nonzero targets uniformly in width. It is an evaluator example, not a proof that all continually reached targets meet the assumptions.

Let the input dimension be d, use $\sigma(z) = z^2$, and let $m = 2nd$, where $n \ge 1$ is an integer. For every coordinate j, initialize n pairs at $w_j^+ = w_j^- = e_j$ with readouts $+1$ and -1. All hidden weights subsequently train freely. Then $f_0 = 0$. On a radius-R ball, $|w_i^\top x| \le 1 + R$, so $L_1 = 2(1+R)$ and $L_2 = 2$ are uniform in width.

Proposition 14 (Exact target representation and covariance)**.** *For any symmetric $C \in \mathbb{R}^{d\times d}$ with $\|C\|_F \le 2R/\sqrt{d}$, the target $x^\top Cx$ belongs to the initial radius-R tangent class and to the actual nonlinear network class on that ball. The tangent kernel is*

$$\langle \psi(x), \psi(y)\rangle = \frac{4}{d}(x^\top y)^2. \tag{122}$$

Let $z(x) = \mathrm{svec}(xx^\top)$, where the symmetric vectorization preserves Frobenius inner products, and let $\Gamma = \mathbb{E}_\mu[z(x)z(x)^\top]$. The nonzero eigenvalues of Σ are $4/d$ times the nonzero eigenvalues of Γ. In particular their positive lower bound is independent of m for a fixed input law.

Proof. For each repeated pair set $u_j^+ = t_j$, $u_j^- = -t_j$, with $t_j = \sqrt{m}\, C_{j,:}^\top/(4n)$. The tangent function is

$$\frac{4n}{\sqrt{m}} \sum_j x_j t_j^\top x = x^\top C x, \qquad \|u\|_2^2 = 2n \sum_j \|t_j\|_2^2 = \frac{d}{4}\|C\|_F^2 \le R^2.$$

For the nonlinear network, the same identity follows from $((e_j + t_j)^\top x)^2 - ((e_j - t_j)^\top x)^2 = 4 x_j t_j^\top x$. At initialization, each gradient block is $2 b_i x_j x/\sqrt{m}$. Summing inner products over the $2n$ copies of each coordinate yields (122). Hence the gradient feature map is $\sqrt{4/d}$ times an isometric embedding of $z(x)$ into parameter space, proving the covariance statement. □

For an explicit continuous distribution, take $d = 2$, $x(s) = (1, s)/\sqrt{2}$, and $s \sim \mathrm{Unif}[-1, 1]$. With $z = (1/2, s/\sqrt{2}, s^2/2)$,

$$\Gamma = \begin{pmatrix} 1/4 & 0 & 1/12 \\ 0 & 1/6 & 0 \\ 1/12 & 0 & 1/20 \end{pmatrix}, \qquad \lambda_{\min}(\Gamma) = \frac{9 - \sqrt{61}}{60} > 0.$$

Thus one can supply $\lambda_{\mathrm{cov}} = (9 - \sqrt{61})/30$ for every width $m = 4n$. For $Q(s) = 1 + s^2$, choose $C = 2I_2$ and $R \ge 2$. This is an actual policy-evaluation target in a continuous-state reset MDP: let every next state be independently uniform on $[-1, 1]$ and set $r(s) = 1 + s^2 - 4\gamma/3$. Since $\mathbb{E}Q = 4/3$, the Bellman equation holds exactly. Rewards are bounded and the state chain mixes after one step. One action suffices for this compatibility example; it makes no claim about difficult control or sharing. The example proves that the assumptions of Proposition 3 are jointly nonempty, including for a fixed nonzero value function as width grows.

Rank and exploration remain material. Since $\mathrm{tr}\,\Sigma \le G^2$, a uniform bound on all nonzero eigenvalues implies $\mathrm{rank}(\Sigma)\lambda_{\mathrm{cov}} \le G^2$. Consequently bounded G and a width-independent λ_{cov} require bounded effective feature rank. The quadratic example has that property. A generic feature class whose covariance rank grows with width does not automatically have it. Similarly, if an exploration floor tends to zero, the smallest positive eigenvalue under the actual policy law can tend to zero even at fixed feature rank. The full continual rate retains its existing uniform-coverage and uniform-conditioning qualifications.

D.10.2 Compatible and conflicting continuous-state families

This construction verifies the environmental premises of the sharing comparison, including compatibility for arbitrary state-dependent policies. The algorithmic conclusions below retain the evaluator's local exact tangent-realization and conditioning requirements in Assumption 3(a), together with the calibrated auxiliary approximation requirement in Assumption 3(b), at the prescribed budgets.

Let $\mathcal{S} = [0, 1]$, let ν be uniform, let $\mathcal{A} = \{0, 1\}$, and set $\phi(s) = \cos(2\pi s)$. For $|\theta_k| \le 1/2$, define

$$r_k(s, a) = B\phi(s) + b_k + w_k a, \qquad P_k(ds' \mid s, a) = [1 + \theta_k(2a - 1)\phi(s')]\nu(ds'). \tag{123}$$

Return uses initial law ν. Exploration is uniform over actions, so $\pi_p \in \Pi_\alpha$ has measurable action-one probability $p(s) \in [\alpha/2, 1 - \alpha/2]$.

Lemma 44 (Policy-dependent action advantage)**.** *Put $t_p = \nu(\phi p)$ and $A_k^0 = w_k + \gamma B\theta_k$. For every such policy,*

$$Q_k^{\pi_p}(s, 1) - Q_k^{\pi_p}(s, 0) = A_k(p) = \frac{A_k^0}{1 - 2\gamma\theta_k t_p}, \qquad |t_p| \le \frac{1 - \alpha}{\pi}. \tag{124}$$

The denominator lies in $[1 - \gamma(1 - \alpha)/\pi, 1 + \gamma(1 - \alpha)/\pi]$, so the advantage has the sign of A_k^0 and $|A_k(p)| \ge |A_k^0|/(1 + \gamma/\pi)$.

Proof. Let $M = \nu V_k^{\pi_p}$ and $z = \nu(\phi V_k^{\pi_p})$. Bellman's equation gives

$$Q_k^{\pi_p}(s, a) = B\phi(s) + b_k + w_k a + \gamma M + \gamma\theta_k(2a - 1)z,$$
$$A = w_k + 2\gamma\theta_k z, \qquad V_k^{\pi_p}(s) = B\phi(s) + (b_k + \gamma M - \gamma\theta_k z) + Ap(s).$$

Since $\nu\phi = 0$ and $\nu\phi^2 = 1/2$, $z = B/2 + At_p$, proving the formula. Write $p = \alpha/2 + (1 - \alpha)q$ with $0 \le q \le 1$. The extrema of $\nu(\phi q)$ are $\pm\nu(\phi_+) = \pm 1/\pi$. Since $|2\theta_k| \le 1$, the denominator bounds follow and are positive. □

The advantage equals A_k^0 for constant policies, but it is generally policy-dependent. This distinction is why the uniform margin below is $a_*/(1+\gamma/\pi)$ rather than a_*.

Proposition 15 (Uniform environmental envelopes)**.** *Each policy has invariant law $\rho_{k,\pi}$ with $\nu/2 \le \rho_{k,\pi} \le 3\nu/2$ and $\sup_s \|(P_k^\pi)^t(s,\cdot) - \rho_{k,\pi}\|_{\mathrm{TV}} \le 2^{-t}$. For monitor laws $\mu_i = \nu$ and invariant snapshot evaluation laws,*

$$\kappa_{\mathrm{tr}} = 2, \quad C_0 = 2, \quad C_1 = 3, \quad A_V = \sqrt{2} + \gamma\sqrt{3}, \quad H_M \le \sqrt{3/2}, \quad H_Q^{\mathrm{SA}} \le \sqrt{3/p_{\min}} \tag{125}$$

are valid envelopes. A uniform-action logger has floor $p_{\log} = 1/2$ and state transition law exactly ν.

Proof. Every policy transition density lies in $[1/2, 3/2]$. Minorization by $\nu/2$ gives contraction factor $1/2$, existence and uniqueness of the invariant law, and the mixing bound. Integrating the density bounds gives the invariant bounds and the same bounds after every positive number of transitions. Thus $\nu \le 2\rho_{k,\pi}$, and $\nu P_i^\vartheta \le (3/2)\nu \le 3\rho_{i,\bar\pi_i}$, proving cross-family and snapshot transfer. Every comparator prefix starting at ν, and every subsequent discounted occupancy d, is bounded by $(3/2)\nu$. Hence $\mathsf{h}(d \mid \nu)^2 \le 3/2$. For the active state law ρ, $d \le 3\rho$; writing $w = dd/d\rho$, we have $w \le 3$ and $\int w\, d\rho = 1$. For any action law q,

$$\mathsf{h}(d \otimes q \mid \rho \otimes \pi^-)^2 = \int w^2 \sum_a \frac{q(a \mid s)^2}{\pi^-(a \mid s)} d\rho \le 3/p_{\min}.$$

The discounted averages defining H_Q^{SA}, H_M preserve these bounds. Uniform action averaging cancels the cosine term in the logger kernel. □

Fix $B = 1/2$, $a_* = (1+\gamma)/4$, and define the two matched instances by

	θ	w	b
family 1 in both instances	$1/2$	$1/4$	$-1/8$
family 2: conflicting	$-1/2$	$-1/4$	$3/8$
family 2: compatible	$-1/2$	$1/4+\gamma/2$	$1/4-(1/4+\gamma/2)/2$

(126)

The common bound $R_{\max} = 9/8$ bounds all rewards in absolute value.

Proposition 16 (Matched identification and compatibility)**.** *The uniform diagnostic policy and input law $\nu \otimes \mathrm{Unif}\{0,1\}$ admit the same supplied separation $\Delta_0 = 1/4$ and $A_{\mathrm{id}} = 1+\gamma$ on both instances. The compatible pair satisfies Assumption 4 for all measurable incumbents and anchors, with*

$$a^\dagger(s) = 1, \qquad a_{\min} = a_*/(1+\gamma/\pi), \qquad \kappa_{\mathrm{tr}} = 2. \tag{127}$$

Proof. Let $m_k = b_k + w_k/2$. Uniform diagnostic evaluation gives $V_k^{\mathrm{id}} = B\phi + m_k/c$ and

$$Q_k^{\mathrm{id}}(s,a) = B\phi(s) + b_k + w_k a + \gamma m_k/c + \gamma B\theta_k(2a-1)/2.$$

Substituting reference ℓ into family k's diagnostic operator gives residual $(b_\ell - b_k) + a(w_\ell - w_k) + \gamma B(\theta_\ell - \theta_k)(2a-1)/2$. Its two action values are $(m_\ell - m_k) \pm (A_\ell^0 - A_k^0)/2$, so its squared norm is

$$(m_\ell - m_k)^2 + (A_\ell^0 - A_k^0)^2/4. \tag{128}$$

Here $m_1 = 0, m_2 = 1/4$ for both instances. The conflicting advantages are $a_*, -a_*$ and the compatible advantages are a_*, a_*. Their exact root separations are therefore $\sqrt{1/16 + a_*^2}$ and $1/4$, respectively, in either order. The diagnostic state-action kernel preserves the common law, and Jensen gives its L_2 contraction, hence $A_{\mathrm{id}} = 1+\gamma$. On the compatible pair, Lemma 44 makes action 1 the unique maximizer for every incumbent and anchor, with gap (127). Proposition 15 supplies cross-law domination. □

Proposition 17 (Conflicting families require two common policies)**.** *For the conflicting pair, every measurable common stationary policy π satisfies*

$$\max_{k=1,2}\{\mathrm{Ret}_k(\pi_k^*) - \mathrm{Ret}_k(\pi)\} \ge L_{\mathrm{conf}} := \frac{a_*(c+\gamma/2)}{2c} > 0. \tag{129}$$

Thus $\varepsilon < L_{\mathrm{conf}}$ requires two deployed policies when one module supplies one common state-to-action map with no family-dependent input.

Proof. Constant action 1 is optimal for family 1 and constant action 0 for family 2: their own advantages are a_* and $-a_*$, so they satisfy the optimal Bellman equations. At their optimal values a common action-one probability $p(s)$ has Bellman residuals $a_*(1-p)$ and a_*p. The positive resolvent identity yields gaps $c^{-1}d^{\pi}_{1,\nu}[a_*(1-p)]$ and $c^{-1}d^{\pi}_{2,\nu}[a_*p]$. Since each positive-time law dominates $\nu/2$, both occupancies dominate $(c+\gamma/2)\nu$. The larger gap is at least $a_*(c+\gamma/2)c^{-1}\max\{\nu(1-p), \nu p\} \geq L_{\mathrm{conf}}$. The argument covers all measurable $p \in [0,1]$, including policies outside Π_α. □

Theorem 3(b), proved in Appendix C.3, combines these constructions with the main accuracy guarantee at the same supplied budgets. Its probability guarantee applies to each instance separately; Table 3 records the deployed-policy comparison.

Two-family instance	Minimum policies	Our method	Separate policies
Compatible	1	1 module, 0 splits	2 by design
Conflicting, $\varepsilon < L_{\mathrm{conf}}$	2	2 modules, 1 split	2 by design

Table 3: Theorem 3(b): each module supplies one common state-to-action law. Algorithmic accuracy is conditional on the stated neural and sampling assumptions; the comparison concerns deployed policies.

The lower bound compares deployed common policies. Family-conditioned architectures and total parameter memory are different resource models. The model verifies environmental feasibility. The evaluator's local exact tangent realization and conditioning, and the auxiliary nonlinear approximation bounds $a \leq u/3$, remain explicit execution premises at the prescribed widths, radii, and tolerances. The auxiliary bounds cover the learned preservation endpoints and all cross-family diagnostic targets specified in Appendix D.1.3. Verifying these neural conditions for the complete execution requires additional arguments beyond this environmental construction.

Why a one-sided gate is needed. On the compatible pair start both records at constant $\pi_{1/2}$, take an exact active critic, and protect family 2's exact snapshot at that anchor. Put $d_\alpha = (1-\alpha)/2$. The full endpoint has action-one probability $1-\alpha/2$, so Bellman affinity gives

$$V_2^{\pi_{1/2}} - T_2^{\pi^\lambda} V_2^{\pi_{1/2}} = -a_*\lambda d_\alpha, \qquad R_2^+(\lambda) = 0, \quad R_2^{\mathrm{sym}}(\lambda) = a_*\lambda d_\alpha. \tag{130}$$

Constant-policy evaluation gives return gain $a_*\lambda d_\alpha/c$ for both families. A symmetric population gate requires $\lambda \leq \tau/(a_* d_\alpha)$ and eventually excludes every $\lambda \geq \lambda_{\min}$ when $\alpha = \Theta(\varepsilon)$ and $\tau = \Theta(\lambda_{\min}\varepsilon)$. The one-sided endpoint passes with the statistical slack in (119). This is a single-transaction obstruction for the symmetric gate; the trajectory-wide no-split result for the one-sided gate follows from Theorem 39.

For the same frozen transaction and radii, every symmetric-feasible step is also one-sided feasible (Proposition 7).

E Experimental results and protocols

Appendix E.1 expands the main recurrent experiment with its task, compared methods, complete outcomes, and resource costs. Appendix E.2 gives the learned implementation and execution diagnostics. Appendix E.3 presents the supporting CartPole study, including its protocol and all five seeds. Cohorts are separate; checkpoints, routing decisions, and assessment episodes within a run are not additional independent trials. The code was ran on a 12700k processor with an RTX 5090 and 32GB RAM.

E.1 Learned neural recurrent control

This primary study tests learned evaluation, routing, preservation, and splitting together. Ten randomly drawn seeds are paired across five neural methods and a linear reference, giving 60 runs. We expand the main-text outcomes here; Appendix E.2 supplies the neural implementation and execution diagnostics.

E.1.1 TASK AND EXPERIMENTAL PROTOCOL

. The three actions are *advance*, *retreat*, and *take*. Advance and retreat move one cell in the indicated direction, saturating at the boundary. Take stays in place except at the family's goal, where it moves to the opposite endpoint. With probability .025, this destination is replaced by a uniformly chosen cell. The next state is uniform within its destination cell. Families A and B have goal cell 5; family C has goal cell 0. Goal-taking pays .85, other taking pays .055, and movement pays $-.012$. Family B adds .12 to every reward. Thus A/B have aligned action preferences, whereas C requires the opposite direction of travel. The discount is $\gamma = .93$.

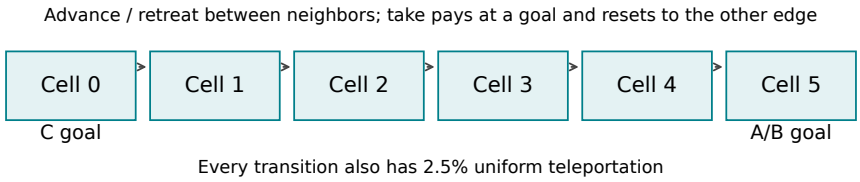


Figure 2: Recurrent control task. Reaching the rewarding endpoint requires state-dependent decisions; taking there resets the state to the other endpoint. The known cells provide the learner's input representation.

The stream is A,B,A,C,B,A,C,A,B,C,A,B,C,A,B. Its 15 supplied context intervals each permit 65,536 transitions, totaling 983,040 per run. Family labels are hidden. The state is initialized uniformly at each boundary, and a uniform-action collector continues across blocks within an interval. Exogenous random innovations are keyed by seed, interval, and absolute transition position to pair methods despite different block allocations. Each active critic, reference, or fresh snapshot uses 4,096 transitions; routing uses 8,192. Each record holds at most four sealed solve/audit reserve pairs, each with 4,096 transitions per side, acquired only while that record is active. All acquired observations are charged, including unused reserves and incomplete or unsuccessful operations. Methods with smaller preservation costs can use the remaining allowance for additional critic fits.

Policies are cellwise tables with exploration $\alpha = .12$, hence minimum action probability .04. Initially, take has probability .92. The endpoint assigns .92 to a greedy action of the learned critic, with fixed first-index tie breaking. The candidate is $\pi_\lambda = (1-\lambda)\pi^{\mathrm{old}} + \lambda\pi^{\mathrm{end}}$, where $\lambda \in \{1, .99, \ldots, .10\}$. Routing compares learned diagnostic Bellman residuals to .045 and selects the matching record with the smallest score, or creates a record if none matches. First-fit allocation allows three records per module; the independent baseline allows one. Neither exact values nor family labels select policies or repair routing.

E.1.2 COMPARED METHODS

Table 4 isolates the effects of preservation, splitting, sharing, and the one-sided score. The linear reference retains full's outer decisions while changing the fitted function classes.

All methods learn references, infer routes, and fit fresh snapshots. The five neural methods share width-64 learners; linear full replaces all critic, reference, snapshot, and auxiliary fits with linear fitted evaluation and cellwise conditional-mean regression. Ungated and independent omit preservation reserves and solve/audit fits but retain routing auxiliaries. All six share the stream, boundaries, collector, initial policy, exploration, greedy endpoint, paired seeds, and charged interaction allowance; optimizer steps and fresh proposals can differ. Appendix E.2.1 specifies the score and training rules.

E.1.3 ASSESSMENT AND UNCERTAINTY

Exact-model computations are used only for assessment. Let $J_k(\pi)$ be discounted return from the uniform initial state and $\overline{J} = \frac{1}{3}\sum_k J_k(\pi_k)$ the mean across the current family policies. The exploration-constrained optima are 1.684570, 3.398856, and 1.684570 for A, B, and C, respectively. Dormant-anchor loss is $[J_i(\pi_i^{\mathrm{anchor}}) - J_i(\pi_i^{\mathrm{after}})]_+$, maximized over dormant records and installations in each run; the old anchor is assessed before refresh. The .10 return-loss tolerance is a reporting criterion distinct from the .02 Bellman threshold. A genuine change has $\frac{1}{12}\sum_{\ell,a} |\pi^{\mathrm{new}}(a \mid \ell) -$

Method	Module allocation	Preservation test	Update or fallback
Full	First-fit sharing; capacity 3	One-sided solve and fresh audit	Accept a feasible mixture; otherwise split and install the full endpoint.
Ungated	First-fit sharing; capacity 3	None	Install the full endpoint in the shared module; no splitting.
No splitting	First-fit sharing; capacity 3	One-sided solve and fresh audit	Accept a feasible mixture; otherwise reject and retain the incumbent.
Independent	One module per inferred record	None	Install the full endpoint in the record's own module.
Symmetric gate	First-fit sharing; capacity 3	Symmetric solve and fresh audit	Use full's mixture and splitting rule with the symmetric residual.
Linear full	Same as full	Same as full	Use full's outer algorithm with linear fitted evaluators and regressions.

Table 4: Operational comparison. An unsuccessful protected proposal includes an infeasible solve, a failed fresh audit, or a missing reserve. Splitting copies the incumbent and isolates the active record before installing its endpoint. Independent allocation uses inferred records, not supplied family labels.

$\pi^{\mathrm{old}}(a \mid \ell)| > .01$; a positive change additionally gains more than .02 active-family return. Intervals are 95% Student-t intervals across the ten run-level outcomes. Curve bands are pointwise, and untruncated intervals for nonnegative statistics can have negative lower endpoints.

E.1.4 Recurrent performance, preservation, and sharing

Matching final performance with fewer deployed modules. Full reaches the exploration-constrained optimum in every run, with final mean return 2.255998 and worst-family return 1.684570. A/B share one module and C occupies a second in all ten runs. Independent policies reach the same endpoint using three modules. Table 5 gives the complete comparison, and Figure 3 shows the family-specific learning trajectories for the aligned pair and the conflicting family.

Method	Final mean return	Worst family	Mean max. anchor loss	Modules
Full	2.256	1.685	.009 [−.012,.031]	2
Ungated	1.862	.503	1.181	1
No splitting	1.818 [1.813,1.823]	.535 [.532,.539]	.101 [.092,.110]	1
Independent	2.256	1.685	.000	3
Symmetric gate	2.256	1.685	.041 [.001,.081]	3
Linear full	2.256	1.685	.009 [−.012,.031]	2

Table 5: Ten paired runs per method under the same 983,040-transition allowance. Brackets give run-level 95% intervals; identical run outcomes have no brackets. Worst-family return is assessed at the final policy.

Preservation during recurrent learning. Full's mean maximum dormant-anchor loss is $.009433$, compared with 1.181 for ungated sharing, a reduction of approximately 99.2%. Its largest individual loss is $.094328$. The numbers of runs exceeding the $.10$ reporting tolerance are 0/10 (full), 10/10 (ungated), 6/10 (no splitting), 0/10 (independent), 2/10 (symmetric), and 0/10 (linear full). These run maxima reflect preservation throughout learning. The tolerance is an empirical criterion; these observed outcomes do not assert a calibrated population failure rate.

Repeated progress and useful sharing. Full makes 7.3 [7.0,7.6] genuine positive-gain changes per run. Of its 13 audited shared installations across the cohort, 12 improve both the active family and at least one protected family relative to their immediately preceding policies, by more than .02 each. These simultaneous current-policy improvements demonstrate useful sharing during learning.

What the ablations explain. Ungated sharing loses performance on the conflicting family. The no-splitting method retains the one-sided test but stalls when a proposal cannot be accepted; splitting

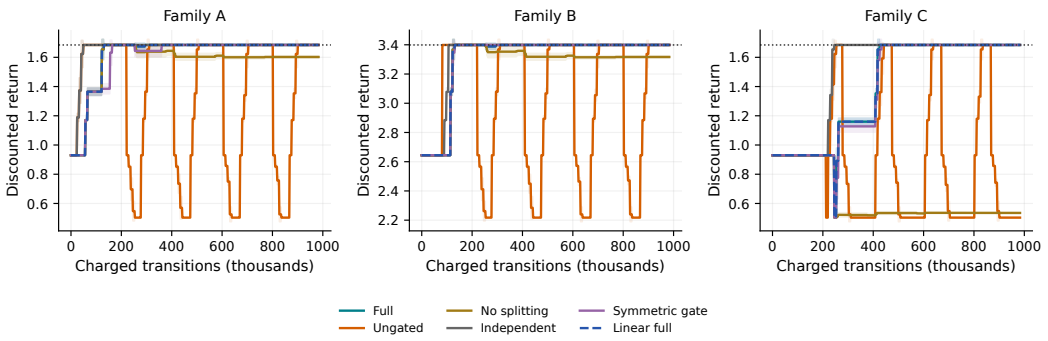


Figure 3: Per-family returns against charged interactions. Dotted lines mark the exploration-constrained optima; bands are pointwise 95% intervals over ten paired seeds.

allows full to continue learning. Symmetric preservation reaches full's final return using three modules, whereas the one-sided test retains the useful A/B sharing in this task.

Learning speed and the linear reference. The normalized area under the mean-return curve is 2.107 [2.105,2.110] for full and 2.163 [2.162,2.163] for independent policies: independent learning reaches high returns sooner despite using an additional module. The learned neural and paired linear full methods produce identical observed policy trajectories on this task. This comparison demonstrates that trained neural components recover the linear reference's continual control outcomes in the known feature space.

E.1.5 Resource costs

Table 6 gives measured cohort means. Each record retains 18 diagnostic-reference outputs, 18 anchor-policy probabilities, and six snapshot values. Full retains 162 policy/reference/anchor scalars, including 36 deployed probabilities in its two modules. A compressed solve/audit pair occupies 2,016 bytes. Preservation reserves increase full's peak retained numeric storage to 24.89 KiB, compared with 1.41 KiB for independent policies. Thus the reduction in deployed modules does not imply a reduction in total storage.

Method	Optimization steps	Time (s)	Retained peak (KiB)	Process peak (MiB)
Full	511,000	2.234	24.89	114.8
Ungated	542,000	2.291	1.12	115.3
No splitting	572,800	2.863	24.75	113.8
Independent	542,000	2.291	1.41	115.1
Symmetric gate	504,400	2.188	25.03	114.8
Linear full	43,180	.961	24.89	107.7

Table 6: Mean measured costs. Neural counts are empirical-gradient steps; linear counts are fitted-Bellman iterations and exclude direct auxiliary means. Process memory includes simulation, assessment, compilation, and logs.

Full averages 166.3 critic/reference/snapshot fits and 44.6 auxiliary fits per run, with summed fit times of 1.018 and .783 seconds, respectively. Training uses 1,152 critic or 1,216 auxiliary parameters plus transient arrays; cached outputs replace neural weights after each fit. Raw blocks can reach 256 KiB. CPU runs use an Intel Core i7-12700K, approximately 31.8 GiB RAM, three workers, one numerical thread per worker, Python 3.10.11, NumPy 2.2.6, and Numba 0.67.0. Runtime includes simulation, optimization, assessment, and first-use compiled-kernel loading, and excludes process launch and final serialization. Different execution batches and cache conditions make these descriptive costs rather than a controlled speed comparison. Interaction allowances, rather than optimizer steps, are matched. The known representation makes this a controlled test of learned continual control; the

widths, training budgets, consecutive sampling, and reporting tolerance are practical experimental choices.

E.2 Neural implementation and execution diagnostics

This subsection supports the recurrent outcomes in Appendix E.1. It records the implemented neural updates, preservation tests, and execution diagnostics from the same ten-seed recurrent cohort.

E.2.1 Neural implementation and preservation decisions

Preservation solve and audit. For a protected record i, the learned snapshot $\widehat{V}_i$ is frozen. An action-conditional regression estimates $q_i^{\mathrm{aux}}(\ell, a) = \mathbb{E}_i[R + \gamma \widehat{V}_i(S') \mid \ell, a]$. Its fitted output $\widehat{q}_i^{\mathrm{aux}}$ gives the one-sided score

$$\widehat{h}_i(\pi) = \left\{ \frac{1}{6} \sum_{\ell=0}^{5} \left[\widehat{V}_i(\ell) - \sum_a \pi(a \mid \ell) \widehat{q}_i^{\mathrm{aux}}(\ell, a) \right]_+^2 \right\}^{1/2}.$$

The solve selects the largest grid weight with every protected score plus .01 padding at most .02. The selected mixture is frozen and tested once using a fresh auxiliary fit on the separate audit block. Infeasibility, a failed audit, or exhausted reserves triggers a split and full-step retry; the no-split baseline rejects instead. Splitting copies the incumbent policy and isolates the active record. The symmetric baseline omits the positive part in the score. Accepted updates receive fresh snapshots. Unchanged proposals neither refresh anchors nor consume preservation reserves.

Trained evaluators and auxiliary regressions. The critic is the width-64 network $f_W(\ell, a) = 64^{-1/2} \sum_{j=1}^{64} b_j (w_j^\top e_{(\ell,a)})^2$, with 32 fixed readouts of each sign and all 1,152 hidden weights initialized to one and trained. Neural fitted policy evaluation uses 500 frozen-target rounds with four empirical-gradient steps per round, step size .8, and projection onto $\|W - W^0\|_F \le 40$. References and snapshots use this same learner. The auxiliary network trains both hidden weights and readouts (1,216 parameters), using 4,000 squared-loss gradient steps of size .1, radius 40, ridge 10^{-6} around initialization, and the average of its final 1,000 parameter vectors. Returned predictions are clipped to $[-1/(1-\gamma), 1/(1-\gamma)]$. One-hot counts and label sums compute the exact sampled-loss gradients. After training, the network's exact outputs on the finite input set are cached and the weights discarded; no fitted linear solution supplies neural weights. Deployed policies are probability tables on the supplied cells.

E.3 CartPole family groups and descendants

Four continuous-state, two-action CartPole families and their descendants are specified in Table 7. Descendants multiply parent gravity and pole mass by $1 + \delta$, half-length and force by $1 - \delta$, and inherit rewards and action mappings. Reversed actions exchange force labels. Configurations are withheld from the learner.

Table 7: Named original families and descendants. The perturbation δ is relative to the corresponding parent; original families have no entry.

Context	Gravity	Pole mass	Half-length	Force	Actions	Gain	δ
A	9.8	0.1	0.5	10	Normal	1	–
A'	10.045	0.1025	0.4875	9.75	Normal	1	2.5%
B	3	0.05	1.25	20	Normal	0.2	–
B'	3.15	0.0525	1.1875	19	Normal	0.2	5%
C	22	0.35	0.2	6	Reversed	0.75	–
C'	24.2	0.385	0.18	5.4	Reversed	0.75	10%
D	9.8	0.1	0.5	10	Reversed	1	–
D'	12.74	0.13	0.35	7	Reversed	1	30%

Five seeds (1–5) start without heads or records. Originals appear three times in shuffled order, followed by shuffled descendants and revisits to all eight contexts. Boundaries and the descendant

phase are supplied; identities, parameters, full schedule, family count, and group assignments are withheld. An eight-head/sixteen-record ceiling permits additional groups. Shared parents and descendants retain separate records; four heads is the observed outcome.

Rewards and learner. For $z = (x/2.4, \dot{x}/3, \theta/.20943951, \dot{\theta}/3.5)$, the current-state reward is

$$r(s) = .01g \ \mathrm{clip}(1 - .20z_1^2 - .05z_2^2 - .60z_3^2 - .15z_4^2, -1, 1),$$

where g is the tabulated gain; targets and additive offsets are zero. Initial states are uniform on $[-.05, .05]^4$. Collection uses up to 32 ordinary parallel reset trajectories, with separate training, monitor, and reserve calls. True termination has zero continuation; time-limit truncation bootstraps. Evaluators use float32 4–64–64–2 tanh FQE with zero output-layer initialization, scaled inputs clipped to $[-5, 5]$, and outputs clipped to $[-1, 1]$. Each fit uses 4,096 uniform-logger transitions, 30 frozen-target rounds, 40 Adam steps per round at 10^{-3}, batch size 512, and gradient cap 10; reward normalization uses the fitted block's maximum absolute reward. Diagnostic images use width-128 tanh networks, paired initialization, 1,000 projected-Adam steps at .01, radius 20, and target scale .1. Preservation uses a 4–64–64–2 conditional Bellman correction, 2,000 Adam steps at 10^{-3}, and gradient cap 10. Replay and solve-monitor sizes are 4,096 and 2,048.

Routing and adaptation. Two independent self probes for reference j give cutoff $\ell_j = \max\{.0045, 1.1 \max(e_{j,1}, e_{j,2})\}$; each probe uses 4,096 fit and 4,096 monitor observations. Initial routing reuses the lowest-scoring eligible reference or creates a group. The descendant phase and final revisits use nearest-reference Bellman group routing. Initial records receive six updates and descendants six opportunities. An anchor-relative Bellman-change gate ($\tau = .006$) admits descendant mixture weights in $[.125, .5]$. A proposal is installed only if a separate 64-episode paired active-return screen satisfies $\overline{\Delta R} - 1.645 \, \mathrm{SE}(\Delta R) > 0$; rejected proposals leave policies and anchors unchanged. Screen interactions are learning costs. Independent assessment uses 128 matched initial states, $\gamma = .99$, horizon 1500, and discounted returns divided by g.

Sharing, gains, and retention. All five runs finish with four heads and eight records. Initial discovery/reuse is correct in 60/60 decisions, descendant group assignment in 20/20, and final group routing in 40/40. All 20 original family–seed cases improve over uniform; all 20 descendant gains exceed .01 relative to their zero-shot returns. The mean gain is $+.1658$, with minimum $+.0447$ (Table 8).

Descendant	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5	Mean
A'	$+.2941$	$+.1728$	$+.1694$	$+.2619$	$+.2350$	$+.2266$
B'	$+.0618$	$+.0874$	$+.1870$	$+.0447$	$+.2481$	$+.1258$
C'	$+.2311$	$+.1017$	$+.0598$	$+.0608$	$+.0851$	$+.1077$
D'	$+.2573$	$+.1184$	$+.2594$	$+.2723$	$+.1083$	$+.2031$

Table 8: Normalized discounted-return gains after descendant adaptation, relative to each descendant's pre-adaptation zero-shot return. Means are arithmetic means over the fixed five-seed cohort.

There are 40 installed descendant proposals out of 120 opportunities (9, 8, 8, 8, and 7 by seed), all changing action probabilities. No installed gate violations or active changes below $-.01$ are observed. Maximum measured losses relative to designated original-family baselines and protected anchors are zero on every seed. The cohort charges 8,097,635 learning transitions, including screens, and 8,394,637 assessment transitions.

Scope and replication. Final revisits identify the exact record in 32/40 cases despite 40/40 correct policy-group routes. The 80 rejected proposals remain in the plotted curves. The five seeds were previously examined and retained without replacement, retuning, or checkpoint selection; this exploratory cohort has no comparative baseline, and its SEMs are descriptive. Retention refers to the named baselines and anchors. Manuscript labels (A, B, C, D) correspond to stored labels (A, C, D, B), likewise for descendants. .

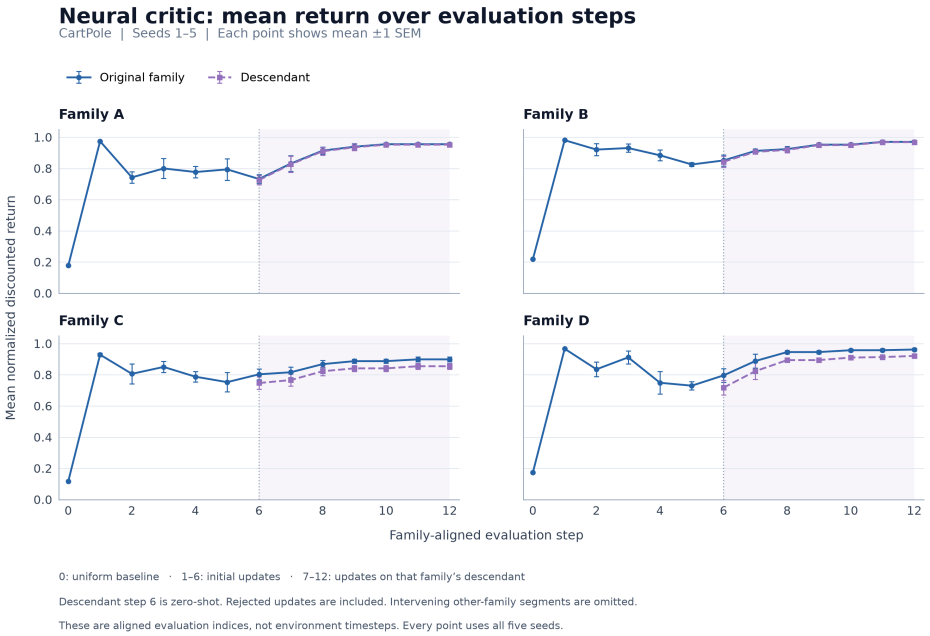


Figure 4: Complete CartPole curves: all five seeds, mean ± one SEM, including step 1. The axis counts family evaluation opportunities, not environment timesteps; intervening other-family segments are omitted. Step 6 is zero-shot, and steps 7–12 include rejected proposals. Retention concerns designated original-family baselines and protected anchors.

F Comparison of theoretical objectives

Table 9 separates established objectives in related work from the joint guarantee studied here. Continuous states, recognizing a returning task, and reusing earlier knowledge are established settings and ideas. The distinction is the finite-data control of changes to a policy shared by separately identified families, together with a fallback that lets the active family improve when sharing cannot be certified. The rows compare objectives and assumptions, rather than rank overall theoretical strength.

Work	Established contribution	Relation to the present guarantee
Brunskill–Li; COMRLI (Brunskill and Li, 2013; Liu et al., 2016)	Task identification and reuse with reduced learning cost on subsequent tasks; COMRLI covers continuous states and discrete actions.	We protect inactive anchors during updates to a common policy used by distinct identified families.
Nguyen–Mehta (Nguyen and Mehta, 2023)	Adversarial multitask regret, separation-dependent clustering bounds, and corresponding lower bounds.	We analyze neural certificates and final family accuracy with preservation; no matching sample lower bound is claimed.
UCBlvd (Amani et al., 2022)	Linear representations, observed contexts, context-independent dynamics, sublinear regret, and sublinear planning calls.	We allow latent families with differing dynamics and protect inactive anchors; no planning-efficiency comparison is established.
Lipschitz lifelong RL (Lecarpentier et al., 2021)	MDP-similarity-based value transfer, PAC-MDP guarantees, and high-probability protection against negative transfer.	We bound subsequent policy changes relative to retained historical anchors.
EPIC (Zhang et al., 2024)	PAC-Bayes guarantees for shared policy distributions, with a relationship between retained tasks and generalization.	We require simultaneous individual-family anchor protection and final accuracy for the specified recurrent stream.
SPIBB (Laroche et al., 2019)	Safe policy improvement relative to a data-collecting baseline from a fixed batch.	We use multiple inactive anchors, adaptive shared updates, finite reserves, and independent splitting for progress.
Safe lifelong policy search (Ammar et al., 2015)	Shared latent policy structure, prescribed parameter constraints, and sublinear regret for lifelong policy search.	Our constraints arise from learned Bellman certificates for inactive-family return loss.

Table 9: Different guarantees and their relation to this paper. Assumptions, performance criteria, and resource measures differ across rows.

What must be proved in combination. Safe improvement, replay, and architectural expansion each supply relevant ideas. Here repeated comparisons use the same retained anchor until that family updates, so intervening other-family steps do not incur an automatic sum of per-step dormant losses. The three-policy decomposition and variable-step propagation connect these comparisons to final active-family accuracy. Conditional concentration must remain valid for adaptively chosen comparisons, which motivates finite reserves with controlled disclosure and complete collection costs. The independent fallback then provides progress without assuming compatibility. The separate no-split theorem shows when one common policy remains sufficient; it does not discover an optimal partition into compatible groups.

Neural and computational scope. The neural contribution analyzes actual training: hidden evaluator weights with fixed readout signs, and hidden and readout parameters of the auxiliary regressor. Its local realization and conditioning assumptions do not imply unrestricted deep representation learning. Neural-TD rates such as those of Ke et al. (2024) require matching squared versus root error, probability mode, approximation terms, and sampling models before comparison. Peng and Papadimitriou (2024) give computational lower bounds for general local model changes under the strong exponential time hypothesis; our structured recurrence and observation upper budgets do not supersede them. Likewise, the robotic old/new-task tradeoffs studied by CRL-VLA (Zeng et al., 2026) concern a different empirical and theoretical setting.