CONSTRUCTIVE SOFTMAX TRAINING FOR CERTIFIED IN-CONTEXT POLICY IMPROVEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

We establish a constructive connection between finite noisy training and policy improvement with frozen Transformer parameters. For finite discounted Markov decision processes sharing a state–action space and discount factor, we train a structured Transformer with standard softmax attention to reproduce an explicit temporal-difference teacher update. A two-stage projected curriculum of designed probes learns three update coefficients and the free attention scores from arbitrary initialization within prescribed parameter boxes. The resulting uniform operator-error bound controls repeated evaluation as target policies and value estimates change using the same deployment record. An outer greedy policy-improvement loop repeatedly calls the frozen evaluator, reusing a single trajectory collected by a fixed behavior policy. Under the stated coverage, mixing, and training conditions, the complete procedure returns an ϵ-optimal policy with a high-probability certificate bounding its suboptimality. At fixed confidence and structural and conditioning constants, sufficient budgets are $N = O(1+\sigma^2\epsilon^{-2})$ labeled operator examples and $m = O(\epsilon^{-2})$ unique deployment transitions, where σ bounds label noise. Training iterations, evaluation depth, and improvement rounds each grow logarithmically in $1/\epsilon$. A two-task lower bound matches the deployment accuracy exponent when pretraining provides no information about the new task. Controlled experiments recover the teacher update from multiple initializations, show that training accuracy matters for repeated evaluation, and demonstrate policy improvement on unseen tasks. The RMS policy-uniform teacher Bellman residual decays approximately as $m^{-1/2}$.

1 INTRODUCTION

In-context reinforcement learning (ICRL) allows an agent to adapt to a new task using its interaction history while keeping the model's parameters fixed. Algorithm Distillation learns from an agent's learning history, while Decision-Pretrained Transformers use contextual information to predict actions (Laskin et al., 2022; Lee et al., 2023). These approaches raise a theoretical question: how can finite pretraining teach a reusable update rule that produces a provably near-optimal policy from a limited interaction history? Answering this question requires connecting the training process to the accuracy of the learned update, its behavior when applied repeatedly, and the resulting policy quality.

Pretraining and deployment provide different kinds of information. Pretraining can teach a model a temporal-difference (TD) update rule without revealing the reward function or transition dynamics of a new task. During deployment, repeatedly applying this rule to the same interaction history allows further policy evaluation, but provides no new observations. We therefore account separately for the number of training examples, optimization steps, distinct deployment transitions, evaluation iterations, and policy improvement rounds.

Existing theory has connected supervised pretraining to decision performance under assumptions such as exact prediction or optimal solutions to pretraining objectives (Lee et al., 2023; Lin et al., 2024; Dong et al., 2025). Other studies have shown how attention models can implement TD updates and have analyzed their training dynamics or evaluation accuracy (Wang et al., 2025; 2026; Xie et al., 2026; Liang & Lai, 2026). These results, however, do not jointly account for (i) the learning error due to finite optimization iterations with noisy supervision and (ii) how this error affects policy quality and its certification when the same deployment data are repeatedly used to evaluate and improve policies.

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.

Both issues are important for understanding the performance of in-context reinforcement learning under limited training and interaction resources. An imperfectly learned update may be applied many times, and its accuracy alone does not establish the decision quality it ultimately produces. Hence, our work aims to connect finite, noisy training to the performance of the returned policy, clarifying when near-optimal behavior can be achieved and how the remaining suboptimality can be certified.

To establish this connection, we study operator distillation, in which a structured Transformer with standard softmax attention is trained to reproduce a specified TD update. We use carefully designed training examples, called probes, to identify the update parameters and analyze how errors from finite, noisy training affect subsequent policy improvement. We train three TD coefficients and the free attention scores, with the rest of the architecture fixed in advance. Our guarantees apply to finite discounted Markov decision processes (MDPs) that share a state–action space and discount factor.

Training proceeds in two stages, using projected updates to keep the parameters within prescribed ranges. We first fit the TD coefficients using inputs containing a single record, then fit the free attention scores using probes containing two records. The parameters can be initialized anywhere within these ranges. With finite training and noisy labels, the learned update approximates the teacher update uniformly over bounded inputs. This guarantee also covers value estimates produced by earlier applications of the learned update during deployment.

At deployment, a greedy policy-improvement procedure repeatedly calls the trained evaluator with fixed parameters. Each round estimates the target policy's action values by iterating the learned update on the same trajectory from a fixed behavior policy. Under coverage and mixing conditions, our statistical guarantee holds simultaneously for all deterministic target policies, including those selected using that trajectory. Each evaluation provides a confidence set for the same optimal action-value function. Under this simultaneous guarantee, intersecting these sets preserves validity and may tighten the returned policy's suboptimality bound. Value estimates guide action selection, while the confidence sets certify the resulting policy's quality.

Our Contributions. We summarize our main contributions as follows.

Finite training of a structured softmax Transformer. We use designed probes that exploit the softmax attention structure to learn three TD coefficients and the free attention scores from noisy teacher labels, with the remaining architecture fixed. We give explicit bounds on the number of training examples and optimization steps sufficient to achieve uniform update accuracy over bounded inputs, from any initialization within the prescribed parameter ranges.

Policy improvement and certification with reused context. We track how training error and estimation error from a finite deployment record propagate through repeated policy evaluation. Combining these bounds with classical approximate policy iteration yields a high-probability near-optimality guarantee and a certificate bounding the returned policy's suboptimality. The analysis covers successive policies chosen using the same trajectory.

Separate sample and computation guarantees. We specify how many training examples, distinct deployment transitions, evaluation iterations, and policy improvement rounds suffice to obtain an ϵ-optimal policy. Under task-independent pretraining, a lower bound based on two tasks matches the ϵ^{-2} dependence of our deployment sample complexity bound.

In controlled experiments, training recovers the teacher update across three classes of attention initialization, and the frozen evaluator transfers to unseen tasks. With target policies and deployment records held fixed, increasing evaluation depth reduces mean evaluation error for final training checkpoints but increases it for early checkpoints. This shows why learning an accurate update matters when that update is applied repeatedly (Appendix C.1.2). With sufficient evaluation depth, the mean suboptimality gaps of the final policies match those of the teacher and an empirical planner using the same records, to the reported precision. Larger deployment contexts improve policy performance, and the RMS policy-uniform teacher Bellman residual decreases approximately as $m^{-1/2}$. Appendix C.1 also reports experiments with noisy training and larger state–action spaces.

2 Related work

Table 1 compares the theoretical guarantees of our work and prior studies, along with the assumptions under which these guarantees hold.

Work	Realization / training premise	Proved guarantee
Lee et al. (2023)	Exact distribution-consistent prediction	Posterior-sampling regret implications
Lin et al. (2024)	ReLU attention; empirical likelihood optimum	Regret with finite pretraining size
Dong et al. (2025)	Advantage-weighted population optimum	Conditional return-difference bound
Wang et al. (2025)	Linear TD; expected pretraining updates	TD realization; invariant parameter set
Wang et al. (2026)	Linear TD; expected-successor prompts	Prediction convergence; asymptotic loss optimality
Xie et al. (2026)	Softmax TD; prescribed parameters	Finite-context/depth evaluation; asymptotic loss optimality
Liang & Lai (2026)	Linear attention; teacher supervision	Local population gradient-flow convergence; actor–critic construction
This work	Softmax; finite noisy empirical GD on probes	Returned-policy bounds and certificates on reused context

Table 1: Selected theoretical guarantees under distinct architectures, supervision, and coverage assumptions. Our row learns prescribed free coordinates from designed probes and certifies an outer policy-improvement procedure using one covered deployment trajectory.

Supervised pretraining and control. Algorithm Distillation learns from RL histories (Laskin et al., 2022). Decision-Pretrained Transformers link exact distribution-consistent prediction to posterior sampling (Lee et al., 2023). Lin et al. (2024) combine ReLU-attention constructions with supervised empirical-likelihood solutions to obtain regret bounds tracking pretraining size. Dong et al. (2025) analyze advantage-weighted pretraining from suboptimal trajectories. These works already connect supervised pretraining to decision performance; our focus is finite optimization of a specified update and its deployment error.

Attention constructions and training dynamics. Wang et al. (2025) construct linear-attention TD updates and identify an invariant set of single-layer expected pretraining dynamics. Wang et al. (2026) establish prediction convergence and asymptotic pretraining-loss optimality using enumerated states and expected successor features. Xie et al. (2026) obtain softmax TD, finite-context/depth evaluation under contraction, and asymptotic optimality of prescribed parameters. Liang & Lai (2026) prove local population gradient-flow convergence for SARSA training and also construct actor–critic updates. Our designed probes yield finite noisy empirical optimization and uniform operator accuracy, which we propagate into returned-policy guarantees.

Looped execution, context reuse, and certification. Looped Transformers implement iterative computations through repeated execution of fixed blocks (Giannou et al., 2023). Our outer greedy loop uses classical approximate policy iteration (Bertsekas, 2011; Scherrer, 2014). We supply its evaluation-error bounds through finite training and simultaneous concentration over target policies on reused Markov context. Confidence sets preserve common-$Q^\star$ containment across changing policy-value centers and certify the returned policy. Our uniform state–action coverage is stronger than comparator-policy coverage studied in pessimistic offline RL (Jin et al., 2021). The deployment lower bound specializes classical testing (Garivier & Kaufmann, 2016) to task-independent pretraining.

3 Setup and algorithm

Fix a finite discounted MDP $M^\star = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$ with $0 < \gamma < 1$ and rewards $|R| \le R_{\max}$ almost surely, where $R_{\max} > 0$. Identify $\mathcal{X} = \mathcal{S} \times \mathcal{A}$ with $[n]$, $n \ge 2$, and set $B = R_{\max}/(1-\gamma)$ and $\mathcal{Q}_B = [-B, B]^n$. The supplied state is Markov. For a deterministic stationary policy π,

$$P_X^\pi((s,a),(s',a')) = P(s' \mid s,a)\mathbf{1}\{a' = \pi(s')\}, \qquad Q^\pi = r + \gamma P_X^\pi Q^\pi, \quad V^\pi(s) = Q^\pi(s, \pi(s)).$$

Let $Q^\star, V^\star$ denote optimal values and $\Pi_{\det}$ denote the deterministic stationary policies, with $P_{\det} = |\Pi_{\det}| = |\mathcal{A}|^{|\mathcal{S}|}$. For transfer across tasks, we assume a shared finite state–action space with consistent indexing and the same discount factor. Guarantees that hold uniformly over a family of tasks also require common bounds on rewards, coverage, and mixing.

Procedure and deployment context. During pretraining, we learn an update rule from N independent operator examples with noisy teacher labels. At deployment, an outer loop alternates between policy evaluation and greedy policy improvement. The value estimates and target policies change, while the evaluator's parameters and the data-collection policy remain fixed. A stationary behavior policy b collects transitions $(S_t, A_t, R_{t+1}, S_{t+1})$ from the new task. To evaluate a target policy π, we reuse these transitions by setting $x_t = (S_t, A_t)$ and assigning the target policy's action to each successor state, giving $y_t = (S_{t+1}, \pi(S_{t+1}))$. This forms the evaluation context $D_m^\pi = \{(x_t, R_{t+1}, y_t)\}_{t<m}$. Successor actions follow the target policy and stay fixed within each evaluation. The outer loop also greedifies the evaluator's output using fixed deterministic ties and computes its certificate. Target policies may depend on the observations.

Deployment sampling. Let $\mathcal{I}_0$ contain the training data, learned parameters, selected behavior and, for analysis only, the fixed deployment MDP. With the filtration $\mathcal{H}_t = \sigma(\mathcal{I}_0, S_0, A_0, R_1, S_1, A_1, \ldots, R_t, S_t, A_t)$, the sampling protocol is

$$\mathbb{E}[R_{t+1} \mid \mathcal{H}_t] = r(X_t), \qquad \mathbb{P}(S_{t+1} = s' \mid \mathcal{H}_t) = P(s' \mid X_t), \quad X_t = (S_t, A_t). \tag{1}$$

The next behavior action has conditional law $b(\cdot \mid S_{t+1})$, including after conditioning on the observed reward and transition. Thus X_t has kernel $K_b((s,a),(s',a')) = P(s' \mid s,a)b(a' \mid s')$. Its initial law may depend on $\mathcal{I}_0$; independence of pretraining and behavior selection is not required. Episodic resets must be included in the homogeneous transition law. We count N labeled operator examples, training gradient steps, m unique deployment transitions, L evaluation loops per round, and J policy improvements separately. Physical transitions used to construct training examples are additional. Our guarantees concern the returned policy's suboptimality, rather than the behavior policy's interaction regret.

3.1 Structured softmax evaluator and probe curriculum

Let $\mathcal{D} = \{(x_t, R_{t+1}, y_t)\}_{t<m}$ be any nonempty finite operator record with $x_t, y_t \in [n]$ and $|R_{t+1}| \le R_{\max}$. The memory $Q \in \mathcal{Q}_B$ stores current action-value estimates. The raw update, before clipping, is

$$\begin{aligned} G_\theta(\mathcal{D}, Q)(x) &= Q(x) + \sum_{t<m} w_t^A(x)[c_r R_{t+1} + c_p Q(x_t) + c_s Q(y_t)], \\ w_t^A(x) &= \exp(A_{x,x_t}) \Big/ \sum_{v<m} \exp(A_{x,x_v}). \end{aligned} \tag{2}$$

Attention determines each transition's contribution at query address x. The trainable coordinates are $\theta = (A, c)$ with $A_{xn} = 0$, $|A_{xu}| \le M_A$ for $u < n$, and $c = (c_r, c_p, c_s) \in [-C, C]^3$. Adding a constant to a score row leaves its softmax weights unchanged; fixing its last entry to zero removes this ambiguity. These constraints define Θ in $p = n(n-1)+3$ trainable score and coefficient coordinates. For $M_A > \tau > 0$ and $C > 1$, the prescribed teacher is

$$A^\star_{xu} = \tau \mathbf{1}\{x = u\} - \tau \mathbf{1}\{x = n\}, \qquad c^\star = (1, -1, \gamma).$$

At these coefficients, the bracket in (2) is the TD error $R_{t+1} - Q(x_t) + \gamma Q(y_t)$. Deployment applies coordinatewise clipping Π_B to keep the updated memory in $\mathcal{Q}_B$. Each loop uses n memory tokens, m transition tokens, and a zero dummy, of width $2n + 4$. Address masks, read/reset heads, and ReLU clipping are fixed; only (A, c) in the aggregation update are trained. Two attention sublayers are reused per loop (see Appendix B.1 for construction). The headline theorem uses exact address masks.

Why a structured evaluator? The structured block makes the learned computation explicit: fixed masks and read/reset heads route stored values, trainable scores weight transitions, and trainable coefficients combine rewards and value estimates. Clipping keeps the memory bounded. This design retains nonlinear softmax normalization while allowing finite optimization guarantees and uniform operator-error control through repeated evaluation and changing target policies. Comparable theoretical analyses also use architectural simplifications: ReLU attention and clipping (Lin et al., 2024), linear attention with fixed masks (Wang et al., 2025), or programmed weights in looped constructions (Giannou et al., 2023). Here the supplied structure and designed probes make the path from noisy operator supervision to frozen policy improvement mathematically traceable. The training guarantee concerns the specified score and coefficient coordinates within this evaluator.

Algorithm 1 Frozen softmax evaluation with reused context and greedy improvement

Require: Labeled data, operator tolerance $e_{\mathrm{op}} > 0$, behavior b, deterministic π_0, context length m, depth L, rounds J, and radius E computed from the training bound and Theorem 4.1

1: Train $\widehat{\theta}$ by the probe curriculum with its proved budgets; freeze it. Set $C_0 = 2B$ and initialize the certificate as in Appendix A.3.
2: Collect and store m transitions under behavior b.
3: **for** $j = 0, \dots, J-1$ **do**
4: Form $\mathcal{D}_m^{\pi_j}$ by re-encoding stored successors for π_j; set $Q_{j,0} = 0$.
5: **for** $\ell = 0, \dots, L-1$ **do**
6: $Q_{j,\ell+1} = \Pi_B G_{\widehat{\theta}}(\mathcal{D}_m^{\pi_j}, Q_{j,\ell})$.
7: **end for**
8: Set $\widehat{Q}_j = Q_{j,L}$; greedify with fixed ties to obtain π_{j+1}.
9: Compute C_{j+1} using $E_j = E$ and (27)–(28), with the stated fallback.
10: **end for**
11: **return** π_J, C_J.

Probe curriculum. Designed probes separate coefficient and score identification. Singletons give attention weight one and isolate coefficients in the increment $G_\theta(\mathcal{D}, Q) - Q$; two-record probes identify the scores. For $0 < r_0 \le R_{\max}$ and coordinate vectors e_k, use

$$\begin{array}{llll} U_R:\ \mathcal{D} = \{(1, r_0, n)\}, & Q = 0; & U_P:\ \mathcal{D} = \{(1, 0, n)\},\ Q = Be_1; \\ U_S:\ \mathcal{D} = \{(1, 0, n)\}, & Q = Be_n; & U_u:\ \mathcal{D} = \{(u, r_0, n), (n, 0, n)\},\ Q = 0\ (u < n). \end{array} \tag{3}$$

These are algebraic inputs and need not be physical trajectories. For U_u, output coordinate x equals $r_0 c_r \ell(A_{xu})$, where $\ell(a) = (1 + e^{-a})^{-1}$. The curriculum first uses projected gradient descent (GD) for the coefficients, then freezes them and applies projected GD to the scores, projecting onto the respective parameter boxes and using

$$\mathcal{L}_C(c) = \tfrac{1}{2}\|c - \widehat{c}\|_2^2, \qquad \mathcal{L}_A(A) = \tfrac{1}{2}\sum_{x=1}^{n}\sum_{u=1}^{n-1}[\bar{c}_r \ell(A_{xu}) - \bar{Y}_{U_u}(x)/r_0]^2. \tag{4}$$

Here $\widehat{c}$ is the empirical target obtained from singleton label means, $\bar{c}$ is the trained coefficient checkpoint, and $\bar{Y}_{U_u}(x)$ is the mean label coordinate for probe U_u. The losses balance probe groups independently of their observed frequencies (Appendix A.1). The headline theorem analyzes these staged losses. Both stages reuse the same N examples with raw labels before clipping; other inputs are not used in these staged objectives.

For a prescribed tolerance $e_{\mathrm{op}} > 0$, Theorem A.2 and Corollary A.3 provide the sample complexity and step costs giving high-probability uniform raw-operator accuracy:

$$\sup_{\mathcal{D},\, Q \in \mathcal{Q}_B} \|G_{\widehat{\theta}}(\mathcal{D}, Q) - G_{\theta^\star}(\mathcal{D}, Q)\|_\infty \le e_{\mathrm{op}},$$

where $\mathcal{D}$ ranges over all bounded nonempty records defined above. The tolerance is certified by the training bounds and remains valid on adaptively generated deployment inputs.

Frozen deployment and the returned certificate. Write $D_j = \|V^\star - V^{\pi_j}\|_\infty$ for the policy gap and let E_j bound $\|\widehat{Q}_j - Q^{\pi_j}\|_\infty$. Algorithm 1 returns a policy and a computable certificate C_J satisfying $D_J \le C_J$ on the theorem's confidence event. The certificate uses the proved training radius and supplied coverage and mixing bounds; its full recursion and inputs are specified in Appendices A.3 and A.4.2. The headline procedure uses one context length m and evaluation depth L throughout.

3.2 Assumptions

The deployment conditions ensure that the reused record covers every address. The training conditions specify noisy teacher supervision and recognizable probes.

Assumption 3.1 (Covered geometrically mixing behavior)**.** Conditionally on the initial information $\mathcal{I}_0$, K_b has stationary law μ_b and deterministic bounds $\mu_0 > 0$, $C_{\mathrm{mix}} < \infty$, $r_{\mathrm{mix}} \in [0, 1)$ satisfying

$$\min_x \mu_b(x) \ge \mu_0, \quad \sup_z \|K_b^k(z, \cdot) - \mu_b\|_{\mathrm{TV}} \le C_{\mathrm{mix}} r_{\mathrm{mix}}^k\ (k \ge 0), \qquad H = \frac{C_{\mathrm{mix}}}{1 - r_{\mathrm{mix}}}. \tag{5}$$

In the above assumption, coverage gives every state–action pair positive stationary probability; mixing controls dependence along the behavior trajectory. Such conditions also underpin finite-time asynchronous Q-learning (Li et al., 2020).

Choosing the teacher score. Choose τ within the prescribed score box to give matching addresses enough attention for contraction:

$$d_\star = \left[1 + \frac{1-\mu_0}{\mu_0} e^{-\tau}\right]^{-1}, \qquad \kappa = d_\star - \frac{1+\gamma}{2} > 0, \qquad \rho = 1 - \kappa. \tag{6}$$

Here $d_\star$ lower-bounds stationary population attention on matching addresses. After the context-length threshold, the margin $\kappa > 0$ yields empirical-teacher contraction with factor at most ρ on the high-probability context event. This score choice parallels Xie et al. (2026, Assumption 5.1).

Assumption 3.2 (Independent bounded operator labels)**.** The pairs $(U_i, Y_i)_{i=1}^N$ are i.i.d., with bounded nonempty operator inputs $U_i = (\mathcal{D}_i, Q_i)$, $Q_i \in \mathcal{Q}_B$, and labels and noise vectors $Y_i, \zeta_i \in \mathbb{R}^n$ satisfying

$$Y_i = G_{\theta^\star}(U_i) + \zeta_i, \qquad \mathbb{E}[\zeta_i \mid U_i] = 0, \qquad \|\zeta_i\|_2 \le \sigma \quad \text{a.s.} \tag{7}$$

Independence is across labeled examples; transitions within an input may be dependent. The noise may depend on its input but is conditionally centered, and σ bounds its Euclidean norm; $\sigma = 0$ gives exact teacher labels. The regression formulation is standard (Hsu et al., 2012); here the teacher exactly realizes the target update. Labels supervise that update, rather than optimal actions or the deployment $Q^\star$.

Assumption 3.3 (Probe identification)**.** Every input in (3) is recognizable from the input alone and has probability at least $p_{\min} > 0$. Other bounded inputs may have arbitrary remaining probability.

Theorem 4.1 uses this condition with the staged losses in (4); the probes identify coefficients and attention scores separately. Their identification role is related to the excitation condition of Liang & Lai (2026, Assumption 1).

4 Main results

The result connects finite training to the returned policy's suboptimality. For $\delta \in (0,1)$ and one predetermined context length m, write $b_m = 2Be^\tau\sqrt{2\log(6nP_{\mathrm{det}}/\delta)/m}$ for the finite-context Bellman-residual radius, uniform over deterministic target policies. Let $m_{\mathrm{ctr}}^{\mathrm{mix}}$ denote the sufficient context-length threshold for empirical-teacher contraction on the confidence event, defined in (25).

Theorem 4.1 (Policy improvement and sample complexity)**.** *Assume Assumptions 3.1, 3.2, and 3.3, together with* (1) *and* (6)*. Train* (4) *from any $\theta_0 \in \Theta$, using the step sizes and finite budgets of Theorem A.2 and Corollary A.3 for a uniform operator tolerance $e_{\mathrm{op}} > 0$ and failure probability $\delta/3$. Freeze the checkpoint and run Algorithm 1 with exact address masks, one predetermined context length $m \ge m_{\mathrm{ctr}}^{\mathrm{mix}}(\delta/3)$ reused in all $J \ge 1$ rounds, and a common evaluation depth $L \ge 1$. Use the common evaluation radius*

$$E_j = E = B\rho^L + \frac{1-\rho^L}{\kappa}(e_{\mathrm{op}} + b_m).$$

With probability at least $1-\delta$ over the training sample and deployment trajectory, the returned policy's actual gap D_J and its certificate C_J satisfy

$$D_J \le C_J \le 2B\gamma^J + \frac{2}{(1-\gamma)^2}\left[B\rho^L + \frac{e_{\mathrm{op}} + b_m}{\kappa}\right]. \tag{8}$$

For any $0 < \epsilon < 2B$, the explicit training and deployment choices in Appendix A.4 ensure $D_J \le C_J \le \epsilon$. Let K_C, K_A count the coefficient and attention gradient steps. At fixed confidence and fixed structural and conditioning constants (including dimension, discount, coverage, mixing, and training conditioning), sufficient budgets as $\epsilon \downarrow 0$ are

$$\begin{gathered} N = O(1+\sigma^2\epsilon^{-2}) \quad \textit{operator examples}, \qquad m = O(\epsilon^{-2}) \quad \textit{unique deployment transitions}, \\ K_C + K_A,\ L,\ J = O(\log(1/\epsilon)). \end{gathered} \tag{9}$$

These budgets include the probe-presence and mixing thresholds.

Theorem 4.1 connects finite, noisy training to certified policy quality through the joint bound $D_J \leq C_J$ and an explicit upper bound on C_J. The uniform update accuracy obtained during training continues to apply as bounded value estimates change, allowing the frozen evaluator to be used repeatedly. Simultaneous control over all deterministic target policies preserves the guarantee when the same m transitions are used both to select and to evaluate successive policies. In (8), the initial-policy term $2B\gamma^J$ and the finite-depth term $B\rho^L$ decrease geometrically with J and L, while training and new-task observations control the learned-update error e_{op} and finite-context error b_m, respectively. Choosing these resources as prescribed therefore ensures $D_J \leq C_J \leq \varepsilon$ with probability at least $1-\delta$. This establishes both near-optimality and an informative certificate for the returned policy, without retraining the evaluator or collecting a fresh trajectory for each improvement round. Under the stated sufficient resource bounds, deployment executes $JL = O(\log^2(1/\varepsilon))$ evaluator loops, each processing the same context. With noiseless labels, the training sample requirement reduces to the probe-presence threshold, while the stated optimization-step bounds still apply.

The Transformer architecture enters Theorem 4.1 through softmax learnability, attention concentration, and repeated computation with shared parameters. Singleton and two-record probes identify the TD coefficients and free attention scores, enabling finite noisy training to control e_{op}. The teacher's score gap τ determines the contraction margin κ and rate $\rho = 1-\kappa$ by concentrating attention on matching state–action pairs. Increasing τ improves the decay $B\rho^L$ and reduces the amplification factor $1/\kappa$, while increasing the factor e^τ in b_m, providing a quantitative basis for balancing evaluation stability and statistical accuracy. Increasing L repeats the same block without adding learned parameters, reducing the finite-depth contribution while training and finite-context errors remain.

Appendices A.1–A.4 provide the training guarantee, simultaneous context event, certificate induction, and explicit budgets. The appendix also covers multiple prefixes; Proposition A.12 refines the evaluation radius using the empirical teacher norm. Theorem A.21 matches the deployment ϵ^{-2} accuracy exponent under task-independent pretraining.

5 Proof outline

For the fixed context length m and depth L of Theorem 4.1, finite training yields uniform evaluation accuracy, which supports policy improvement and certification.

Finite training gives uniform operator accuracy. Singleton probes isolate the three TD coefficients independently of attention, giving the grouped quadratic coefficient loss in (4). After coefficient fitting, two-record reward probes identify each free score through its sigmoid response. On the prescribed score box, the sigmoid derivative is bounded away from zero, giving a contracting score-distance recursion perturbed by coefficient error and noisy label means. Projection keeps the iterates in the boxes, allowing arbitrary initialization within them. One training event controls probe counts and label means, so both stages reuse the same examples. Theorem A.2 and Corollary A.3 convert parameter accuracy into the uniform raw-operator bound e_{op}. This all-input bound covers adaptively generated deployment memories without a distributional match to training.

One context supports every selected policy. The target policies depend on the observations used to evaluate them. We therefore construct a single event covering every deterministic target policy. Concentration of state–action visit counts gives sufficient empirical coverage for teacher contraction under the score-margin condition. For a fixed policy π, define the TD error $\xi_t^\pi = R_{t+1} + \gamma Q^\pi(S_{t+1}, \pi(S_{t+1})) - Q^\pi(X_t)$ and the unnormalized attention weight $a_t(x) = \exp(\tau \mathbf{1}\{x = X_t\})$. The raw teacher residual satisfies

$$G_{\theta^\star}(\mathcal{D}_m^\pi, Q^\pi)(x) - Q^\pi(x) = \frac{\sum_{t<m} a_t(x)\xi_t^\pi}{\sum_{t<m} a_t(x)}.$$

The Bellman equation gives $\mathbb{E}[\xi_t^\pi \mid \mathcal{H}_t] = 0$, and $a_t(x)$ is $\mathcal{H}_t$-measurable, so the numerator is a martingale sum. Since $|\xi_t^\pi| \leq 2B$ and $1 \leq a_t(x) \leq e^\tau$, its increments are bounded by $2Be^\tau$, while the denominator is at least m. Concentrating the numerator before normalization therefore controls the residual despite dependence along the trajectory (Lemma B.15). This order matters because the full-record normalizer also depends on later observations. Intersecting these events over deterministic policies gives the uniform radius b_m before substituting data-dependent targets. The policy-class contribution to the concentration bound is $\log P_{\mathrm{det}} = |\mathcal{S}| \log |\mathcal{A}|$. This simultaneous guarantee is an analytical device: the algorithm evaluates only the policies generated by greedy improvement and

does not enumerate the policy class. Once the event holds, all selected targets are covered, with no additional union over evaluation loops or improvement rounds. Proposition A.10, specialized to the fixed context length, combines this context event with the training event.

Teacher contraction controls finite learned evaluation. Fix a round j and write $e_\ell = \|Q_{j,\ell} - Q^{\pi_j}\|_\infty$. Compare the learned update with the raw teacher update at the current memory and at Q^{π_j}. Teacher contraction propagates the previous evaluation error, uniform operator accuracy contributes e_{op}, and the finite-context residual contributes b_m. Clipping keeps the memories in $\mathcal{Q}_B$, is nonexpansive, and fixes Q^{π_j}, giving $e_{\ell+1} \le \rho e_\ell + e_{\mathrm{op}} + b_m$. Zero initialization gives $e_0 \le B$. Unrolling yields $e_L \le E$, with E as defined in Theorem 4.1. Geometric discounting keeps the accumulated training and context errors bounded by $(e_{\mathrm{op}} + b_m)/\kappa$, uniformly in the evaluation depth. The learned map itself need not contract; teacher contraction suffices. Proposition A.11 establishes the bound for every round.

Evaluation accuracy yields policy improvement and certification. Greedifying action values with error at most E incurs at most $2E$ in the one-step improvement comparison. Classical approximate policy iteration (Bertsekas, 2011; Scherrer, 2014) therefore gives $D_{j+1} \le \gamma D_j + 2E/(1-\gamma)$ (Lemma B.17). Certification maintains $D_j \le C_j$ on the same confidence event. The Bellman equations give $\|Q^{\pi_j} - Q^\star\|_\infty \le \gamma C_j$, hence $\|\widehat{Q}_j - Q^\star\|_\infty \le E + \gamma C_j$. Intersecting the ball of radius $a_j = E + \gamma C_j$ around $\widehat{Q}_j$ with the preceding set $\mathcal{F}_{j-1}$, starting from $\mathcal{F}_{-1} = \mathcal{Q}_B$, therefore preserves $Q^\star \in \mathcal{F}_j$. Its coordinatewise envelopes bound the optimal-action disadvantage of the actual critic-greedy action, yielding $D_{j+1} \le Z_j$ as defined in (28). The universal, policy-improvement, and envelope bounds are all valid, so their minimum preserves $D_{j+1} \le C_{j+1}$ and gives $C_{j+1} \le \gamma C_j + 2E/(1-\gamma)$. Theorem A.13 proves this coupled induction. Unrolling from $C_0 = 2B$ and using $1 - \rho^L \le 1$ in E yields (8), without assuming monotone policy returns.

Why deployment observations remain necessary. The lower-bound construction reverses a small reward advantage between two actions in a one-state MDP. An ϵ-optimal returned deterministic policy must identify the better action, while the per-observation KL divergence is bounded by a constant times the squared reward separation. Task-independent pretraining and the common exact teacher reveal no information about the selected task. The testing argument gives the matching ϵ^{-2} deployment accuracy exponent at fixed confidence (Theorem A.21).

Completing the proof of Theorem 4.1. The training, occupation, and uniform residual events each have failure probability at most $\delta/3$. Their intersection therefore has probability at least $1-\delta$, and the evaluation and certificate recursions give (8) on this event. For $0 < \epsilon < 2B$, set $t_\epsilon = \epsilon(1-\gamma)^2/12$. Choose $B\rho^L \le t_\epsilon$, $e_{\mathrm{op}}, b_m \le \kappa t_\epsilon$, and $2B\gamma^J \le \epsilon/2$. Then $E \le 3t_\epsilon$, so

$$D_J \le C_J \le 2B\gamma^J + \frac{2E}{(1-\gamma)^2} \le \frac{\epsilon}{2} + \frac{6t_\epsilon}{(1-\gamma)^2} = \epsilon.$$

At fixed confidence, structural parameters, and conditioning, geometric convergence gives the logarithmic iteration budgets, while concentration gives the sample budgets in (9). Theorem A.2 and Corollary A.17 provide the explicit training and deployment choices, including probe-presence and mixing thresholds. All rounds reuse the same m transitions. Together, these bounds establish Theorem 4.1.

6 Experimental results

Setup. We use four-state, two-action MDPs with $\gamma = 0.7$. Five uniform-start curriculum models are frozen before evaluation on 100 unseen tasks (50 ring, 50 random-kernel), with three behavior records per task. Methods share records; context sizes use nested prefixes. We compare the learned evaluator, prescribed teacher, and empirical-model planner using $D_j = \|V^\star - V^{\pi_j}\|_\infty$.

Learning and repeated evaluation. The curriculum recovers the teacher update from uniform, boundary, and zero initializations (Figure 1, left). On common fixed target policies, increasing L from 1 to 128 reduces the final checkpoint's mean evaluation error from 0.899 to 0.0478; the early checkpoint's mean instead rises from 1.697 to 4.328 (Appendix C.1.2). These checkpoints use 262,144 and 64 attention updates after coefficient fitting. This comparison exposes the effect of training accuracy on repeated evaluation.

Policy improvement. At $m = 1024$, the mean final gap after $J = 16$ rounds decreases from 0.1293 for $L = 1$ to 0.00147 for $L = 8$ or 64 (Figure 1, right).

Unique deployment observations. At $L = 128$, $J = 8$, the learned, teacher, and planner mean gaps coincide to displayed precision, falling from 0.0168 at $m = 64$ to 7.04×10^{-5} at $m = 16384$ (Figure 2, left). The RMS policy-uniform teacher Bellman residual has log–log slope -0.510, close to the $m^{-1/2}$ sampling term (right). The residual diagnoses the statistical error entering evaluation; greedy actions can stabilize while residual error remains.

Additional evidence. Under noisy labels, held-out operator RMSE has fitted slope -0.592 over $N = 256$–8192 (Appendix C.1.4). Separate models at $n = 8, 16, 32$ reproduce teacher policies using fixed coefficients and exact clean attention targets (Appendix C.1.6). The two-task testing witness and broader frozen transfer are documented in Appendices C.1.5 and C.2.3.

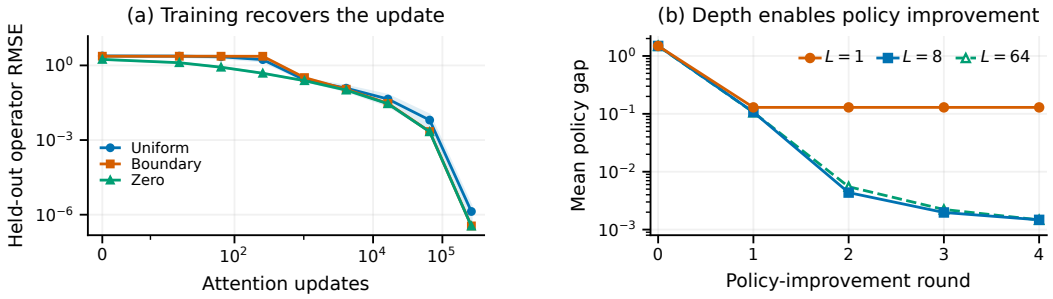

Figure 1: Training and frozen control. Left: held-out operator RMSE on 256 inputs, with mean and seed SD over five training datasets ($N = 1024$); attention step zero follows coefficient fitting. Right: mean gaps over 1,500 paired model–task–record cells (five models, 100 tasks, three records) at $m = 1024$, varying L; rounds 0–4 are shown, with unchanged means through round 16. Protocols: Appendix C.2 (left) and Appendix C.1.2 (right).

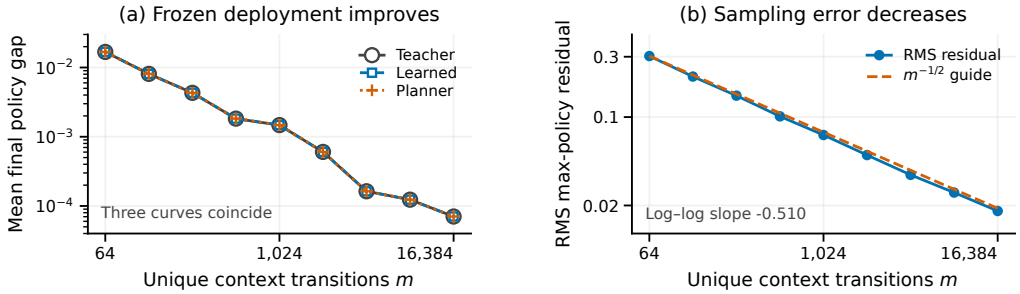

Figure 2: Unique deployment observations at fixed $J = 8$, $L = 128$. Left: mean final policy gaps over paired task–record observations, also averaged over five models for the learned evaluator; the three methods coincide to displayed precision. Right: teacher residual maximized over 16 deterministic policies for each record, then summarized by RMS across the 300 task–record pairs, with an $m^{-1/2}$ guide. Contexts are reused throughout deployment (Appendix C.1.3).

7 Conclusion

We connect constructive finite training of a structured softmax evaluator to frozen policy improvement and optimal-value certificates from reused Markov context. Designed probes permit arbitrary initialization within prescribed parameter boxes. The analysis separates learning the update rule from identifying the deployment task and matches the deployment accuracy exponent with a testing lower bound. Experiments link operator recovery to policy improvement and distinguish evaluation depth from deployment sample size. Residual scaling, noisy-training diagnostics, and teacher-matched transfer across rewards and transition kernels support the analysis.

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.

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.

References

Dimitri P Bertsekas. Approximate policy iteration: A survey and some new methods. *Journal of Control Theory and Applications*, 9:310–335, 2011.

Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. *Operations Research*, 69(3):950–973, 2021.

Juncheng Dong, Moyang Guo, Ethan X Fang, Zhuoran Yang, and Vahid Tarokh. In-context reinforcement learning from suboptimal historical data. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 14021–14039. PMLR, 13–19 Jul 2025.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pp. 998–1027. PMLR, 2016.

Angeliki Giannou, Shashank Rajput, Jy-yong Sohn, Kangwook Lee, Jason D Lee, and Dimitris Papailiopoulos. Looped transformers as programmable computers. In *International Conference on Machine Learning*, pp. 11398–11442. PMLR, 2023.

Daniel Hsu, Sham M Kakade, and Tong Zhang. Random design analysis of ridge regression. In *Conference on learning theory*, pp. 9–1. JMLR Workshop and Conference Proceedings, 2012.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International conference on machine learning*, pp. 5084–5096. PMLR, 2021.

Michael Laskin, Luyu Wang, Junhyuk Oh, Emilio Parisotto, Stephen Spencer, Richie Steigerwald, DJ Strouse, Steven Hansen, Angelos Filos, Ethan Brooks, et al. In-context reinforcement learning with algorithm distillation. *arXiv preprint arXiv:2210.14215*, 2022.

Jonathan Lee, Annie Xie, Aldo Pacchiano, Yash Chandak, Chelsea Finn, Ofir Nachum, and Emma Brunskill. Supervised pretraining can learn in-context reinforcement learning. *Advances in Neural Information Processing Systems*, 36:43057–43083, 2023.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Sample complexity of asynchronous q-learning: Sharper analysis and variance reduction. *Advances in neural information processing systems*, 33:7031–7043, 2020.

Haodong Liang and Lifeng Lai. Transformers provably implement in-context reinforcement learning with policy improvement. *arXiv preprint arXiv:2605.05755*, 2026.

Licong Lin, Yu Bai, and Song Mei. Transformers as decision makers: Provable in-context reinforcement learning via supervised pretraining. In *International Conference on Learning Representations*, volume 2024, pp. 31176–31233, 2024.

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

Jiuqi Wang, Ethan Blaser, Hadi Daneshmand, and Shangtong Zhang. Transformers can learn temporal difference methods for in-context reinforcement learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Pj06mxCXPl.

Jiuqi Wang, Rohan Chandra, and Shangtong Zhang. Towards provable emergence of in-context reinforcement learning. *Advances in Neural Information Processing Systems*, 38:25873–25908, 2026.

Zixuan Xie, Xinyu Liu, Claire Chen, Shuze Daniel Liu, Rohan Chandra, and Shangtong Zhang. Beyond linear attention: Softmax transformers implement in-context reinforcement learning. *arXiv preprint arXiv:2605.07333*, 2026.

APPENDIX CONTENTS

The contents below list sections and subsections. The following reading guide identifies the core proof, special case, and experimental details.

		Page
A	**Main results**	**14**
A.1	Constructive operator training	14
A.2	Simultaneous training and reused-context validity	17
A.3	Evaluation, policy improvement, and the joint invariant	19
A.4	Separate training, context, and computational resources	21
A.5	Detailed statements supporting the headline guarantee	26
B	**Supporting results**	**28**
B.1	Exact architecture and finite routing	28
B.2	Derivative bounds, coefficient design, and identifying probes	30
B.3	One-sample concentration events	32
B.4	Optimization steps and the need for a curriculum	33
B.5	Sampling, concentration, and empirical contraction	35
B.6	Greedification and envelope refinement	39
B.7	Testing and the scope of necessary resources	40
C	**Experimental details and reproducibility**	**44**
C.1	Policy improvement, resource scaling, and testing witnesses	44
C.2	Probe curriculum, frozen transfer, and finite-sample diagnostics	49
C.3	Prescribed-attention special case: coefficient-only training	54

GUIDE TO THE APPENDIX

Appendix A develops the main results, Appendix B supplies their supporting lemmas, and Appendix C gives the experimental protocols and additional results. The following routes let readers follow a particular claim without reading every technical detail in sequence.

The headline guarantee. Start with Appendix A.5 for the derivation of Theorem 4.1 and the detailed training, certification, and resource statements. The core proof then has four steps:

1. **Train a uniformly accurate operator.** Appendix A.1 proves the training guarantees; the probe curriculum begins in Appendix A.1.2. Appendices B.1–B.4 provide the architecture, derivative bounds, sample events, and optimization arguments.
2. **Justify reuse of deployment observations.** Appendix A.2 constructs an event that holds simultaneously over the permitted policies and context prefixes. Appendix B.5 supplies the sampling, concentration, and empirical contraction results.
3. **Convert evaluation accuracy into certified policy improvement.** Appendix A.3 proves the coupled policy and certificate invariant. Appendix B.6 gives the greedification, envelope, and certificate-calibration arguments.
4. **Separate the resource budgets.** Appendix A.4 combines the bounds for training examples, unique deployment observations, and computation. Appendix A.4.2 specifies the inputs needed to compute the certificates.

Special case and scope of the guarantees. Appendix A.1.3 studies coefficient training when the teacher attention is supplied, with its resource composition in Corollary A.18. For the deployment-sample lower bound, read Appendix A.4.4 and its testing proof in Appendix B.7. That appendix also gives the population depth witness and explains the scope of the score-margin requirement. Proposition B.10 explains why teacher excitation alone does not give global convergence of unrestricted joint training.

The experiments in the main text. For the training and operator-recovery panel of Figure 1, read Appendix C.2. For its policy-improvement panel, read Appendix C.1.2; for the context and residual curves in Figure 2, read Appendix C.1.3. These deployment studies reuse the curriculum-trained models on a separate task bank. The remaining studies in Appendix C.1 examine noisy training, the two-task testing construction, and larger address spaces. Appendix C.2 also reports broader frozen transfer. Appendix C.3 studies coefficient-only with prescribed attention under a separate protocol.

A MAIN RESULTS

The headline proof follows four steps: the probe curriculum and its all-input operator bound (Appendix A.1); the simultaneous policy-and-prefix event (Appendix A.2); evaluation and common-$Q^\star$ certification (Appendix A.3); and resource composition and the deployment lower bound (Appendix A.4). Appendix A.1.3 gives the prescribed-attention special case. Appendix B supplies the supporting lemmas, and Appendix A.5 collects the theorem composition and its finite-prefix version.

A.1 CONSTRUCTIVE OPERATOR TRAINING

This subsection proves the probe-curriculum training guarantee used by Theorem 4.1 and the prescribed-attention special case in Appendix A.1.3. Both give a uniform operator-error bound for frozen evaluation and policy certification.

The raw update is (2), with prescribed teacher $\theta_\star = (A^\star, c^\star)$ and labels satisfying Assumption 3.2. All norms on free parameter vectors are Euclidean. Appendix B.1 verifies the token implementation; Appendices B.2–B.4 provide the derivative, concentration, and optimization ingredients. Training examples are reused at every optimization step.

A.1.1 SHARED NUMERICAL INTERFACE

Write $d = n(n-1)$, $p = d + 3$, $\mathcal{C} = [-C, C]^3$, and $J_\theta(U) = D_\theta G_\theta(U)$ in the free score and coefficient coordinates. Define

$$H_v = C(R_{\max} + 2B), \qquad V_0 = \sqrt{R_{\max}^2 + 2B^2}, \qquad L_G = 2H_v + V_0. \tag{10}$$

Lemma B.3 proves the uniform parameter-to-operator bound without a record-length factor. The symbol H_v bounds a TD value projection; H in (5) is the separate mixing constant.

A.1.2 PROBE CURRICULUM FROM ARBITRARY BOUNDED SCORES AND COEFFICIENTS

The explicit probes and their logistic identities are in (61); write $P = n + 2$ for their number and retain the score slope lower bound s_M defined there.

Assumption A.1 (Recognizable probe coverage)**.** Under Assumption 3.2, each probe in (61) has probability at least $p_{\min} > 0$, as required by Assumption 3.3. Probe membership is read from the input before using its label. Remaining inputs may follow any bounded law.

The groups I_a, counts N_a, label means $\bar{Y}_a$, and the single probe count/noise event are defined in Appendix B.3. All groups use the same N examples. Ordinary examples identical to a probe may be included in its group; conditional centering given the input remains sufficient.

The curriculum balances the probe groups instead of weighting them by their observed frequencies. Consequently its first two objectives are different from the original full-data empirical loss. All group means are computed from the same N examples once. If any required group is empty, the procedure declares that the curriculum certificate is unavailable; it can output the initial parameter for definiteness. The theorem's event excludes that branch.

Define the singleton least-squares targets

$$\begin{aligned} \widehat{c}_r &= \frac{1}{nr_0} \sum_x \bar{Y}_R(x), \\ \widehat{c}_p &= \frac{1}{nB} \sum_x [\bar{Y}_P(x) - B\mathbf{1}\{x = 1\}], \\ \widehat{c}_s &= \frac{1}{nB} \sum_x [\bar{Y}_S(x) - B\mathbf{1}\{x = n\}]. \end{aligned} \tag{11}$$

Their vector is denoted $\widehat{c}$. Stage C keeps the arbitrary initial attention fixed and uses

$$\mathcal{L}_C(c) = \tfrac{1}{2}\|c - \widehat{c}\|_2^2, \qquad c_{k+1} = \Pi_{[-C,C]^3}\{c_k - \eta_C(c_k - \widehat{c})\}. \tag{12}$$

Up to a constant independent of c, this is the sum of the three empirical raw singleton losses, each averaged within its group and divided by nr_0^2, nB^2, or nB^2, respectively. Thus it uses the supplied labels, not teacher parameters as optimization targets. Its normalization is part of the declared algorithm.

Let $\bar{c} = c_{K_C}$ be the coefficient checkpoint and put $\widehat{t}_{xu} = \bar{Y}_{U_u}(x)/r_0$. Stage A freezes $\bar{c}$ and minimizes

$$\begin{aligned} \mathcal{L}_A(A;\bar{c}) &= \tfrac{1}{2}\sum_{x,u<n}[\bar{c}_r\ell(A_{xu}) - \widehat{t}_{xu}]^2, \\ A_{k+1} &= \Pi_{[-M_A,M_A]^d}\{A_k - \eta_A\nabla_A\mathcal{L}_A(A_k;\bar{c})\}. \end{aligned} \tag{13}$$

The gauge coordinates remain zero. This objective is, up to a constant, the sum of the empirical raw two-record probe losses, each averaged within its group and divided by r_0^2. Stage A begins at the same arbitrary attention used in Stage C. Both stages differentiate raw outputs before deployment clipping.

Theorem A.2 (Finite-sample curriculum from every initialization in the box)**.** *Under Assumption A.1, fix a parameter tolerance $e > 0$ and confidence $1 - u$. Define*

$$\begin{aligned} D_C &= \sqrt{2(C+1)^2 + (C+\gamma)^2}, \qquad D_A = 2M_A\sqrt{d}, \qquad S_C = \sqrt{r_0^{-2} + 2B^{-2}}, \\ e_A &= e/2, \qquad e_C = \min\left\{\tfrac{1}{2}, e/2, \frac{e_A s_M^2}{2\sqrt{d}}\right\}, \\ t &= \min\left\{\frac{e_C}{2S_C}, \frac{r_0 e_A s_M^2}{2\sqrt{d}}\right\}, \qquad m_0 = s_M^2/4. \end{aligned} \tag{14}$$

Take $N \geq N_{\mathrm{probe}}(t,u)$, deterministic step sizes $0 < \eta_C \leq 1$, $0 < \eta_A \leq 64/9$, and deterministic integer budgets

$$\begin{aligned} K_C &\geq \left\lceil \eta_C^{-1}\log\max\{1, 2D_C/e_C\}\right\rceil, \\ K_A &\geq \left\lceil (\eta_A m_0)^{-1}\log\max\{1, 2D_A/e_A\}\right\rceil. \end{aligned} \tag{15}$$

Start with any $(A_0, c_0) \in \Theta$, possibly selected from the sample. Run (12) *and then* (13)*. With probability at least $1 - u$, simultaneously for every such initialization,*

$$\|\bar{c} - c^\star\|_2 \leq e_C, \quad \|A_{K_A} - A^\star\|_2 \leq e_A, \quad \|\theta^{\mathrm{cur}} - \theta_\star\|_2 \leq e, \qquad \theta^{\mathrm{cur}} = (A_{K_A}, \bar{c}). \tag{16}$$

All curriculum iterates lie in Θ by projection. Neither stage assumes local initialization. There are N examples, reused throughout, and $K_C + K_A$ gradient steps on the declared grouped objectives.

Proof. Work on the event of Lemma B.7, which has probability at least $1 - u$. From the singleton identities and (11), $\|\widehat{c} - c^\star\|_2 \leq tS_C \leq e_C/2$. The coefficient phase is a projected quadratic step with identity Hessian. Lemma B.8 gives

$$d_{k+1}^C \leq (1 - \eta_C)d_k^C + \eta_C t S_C, \qquad d_k^C = \|c_k - c^\star\|_2.$$

Its unrolled solution and $(1 - \eta_C)^k \leq e^{-\eta_C k}$ show $d_{K_C}^C \leq e_C/2 + e_C/2 = e_C$ under (15). When $\eta_C = 1$, the transient vanishes after one step; the stated positive budget covers that case. In particular $1/2 \leq \bar{c}_r \leq 3/2$.

For Stage A, define $\nu_{xu} = \bar{\zeta}_{U_u}(x)/r_0$. The same event gives $\|\nu\|_2 \leq \sqrt{d}\, t/r_0$. Lemma B.9 gives the projected distance recurrence (72), with

$$m_A = \bar{c}_r^2 s_M^2 \geq m_0, \qquad E_A \leq \frac{|\bar{c}_r|\sqrt{d}}{4}(|\bar{c}_r - 1| + t/r_0).$$

Every iterate is in the box by projection, so the score mean-value bounds used in that lemma are valid before the next step. The chosen tolerances imply

$$\frac{E_A}{m_A} \leq \frac{\sqrt{d}}{2s_M^2}(e_C + t/r_0) \leq e_A/2.$$

Unrolling from $d_0^A \le D_A$ and using the attention budget yields

$$\|A_{K_A} - A^\star\|_2 \le e^{-\eta_A m_0 K_A} D_A + E_A/m_A \le e_A.$$

Finally $e_C, e_A \le e/2$, so the full distance is at most $\sqrt{e_C^2 + e_A^2} \le e$. All steps were pathwise on one event for fixed group means. It consequently covers all sample-dependent initializations and every reuse of those means without stage independence. □

Corollary A.3 (Direct all-input operator accuracy and control substitution)**.** *With L_G from* (10)*, the curriculum checkpoint in Theorem A.2,*

$$\sup_{U \text{ bounded, nonempty}} \|G_{\theta^{\mathrm{cur}}}(U) - G_{\theta_\star}(U)\|_\infty \le L_G e \tag{17}$$

on its event of probability $1 - u$. The same bound holds after clipping both outputs. Thus any desired operator tolerance $e_{\mathrm{op}} > 0$ is achieved by setting $e = e_{\mathrm{op}}/L_G$ in the curriculum theorem. Its checkpoint is frozen directly, supplying the training-error radius e_{op} to the evaluation and control theorem.

Proof. Lemma B.3 supplies the deterministic all-input Lipschitz inequality with constant L_G. Apply it to the parameter conclusion (16). Clipping is nonexpansive. The frozen output satisfies exactly the bounded-parameter, uniform-operator interface used by Proposition A.11, so its radius can be substituted without another training failure allocation. □

Remark A.4 (Curriculum resources and scope). For fixed conditioning constants, (69) supplies $O(\sigma^2 e^{-2})$ examples beyond probe presence, and (15) supplies logarithmic phase budgets as $e \downarrow 0$. The explicit s_M^{-1} and $p_{\min}^{-1}$ factors remain. Noiseless labels eliminate the statistical-noise term. The teacher is an explicit update construction, so these are sufficient training resources, not an information-theoretic need to estimate a known teacher. The curriculum changes the optimized losses and does not claim global joint-GD convergence on an arbitrary mixture; Proposition B.10 shows why that stronger claim would fail.

A.1.3 Special case: coefficient training with prescribed attention

When attention is supplied as $A = A^\star$, only the three coefficients must be learned. The raw evaluator is affine in these coefficients: $G_{A^\star,c}(U) = Q + F(U)c$. Thus the full empirical loss is quadratic, and coefficient excitation suffices without probes that identify the attention scores. This special case separates coefficient learning from the score-identification task addressed by the main curriculum. The feature matrix $F(U)$ and its bound $L_c = nV_0^2$ are given in (59)–(60); the loss, gradient, and concentration thresholds are (65)–(67).

Assumption A.5 (Coefficient-training excitation)**.** Under the common label Assumption 3.2, keep attention fixed at $A^\star$ and require

$$H_c = \mathbb{E}[F(U)^\top F(U)] \succeq \lambda_c I_3 \quad (\lambda_c > 0). \tag{18}$$

Only examples, not individual transitions inside an example, must be independent. Necessarily $0 < \lambda_c \le L_c$. Proposition B.5 gives a concrete sufficient law. Its excitation requirement is weaker than full score-and-coefficient excitation because attention is supplied; the prescribed-attention premise is an additional structural restriction. The coefficient Gram plays the identification role of a full-rank feature covariance in linear TD analysis (Bhandari et al., 2021).

Theorem A.6 (Global coefficient training with bounded noisy labels)**.** *Under Assumption A.5, take $N \ge N_c(u)$, any $c_0 \in \mathcal{C}$, and $0 < \eta_c \le 1/L_c$. The initial vector may be data dependent. Run projected full-batch GD on the raw loss:*

$$c_{k+1} = \Pi_{\mathcal{C}}\big(c_k - \eta_c \nabla \widehat{L}_c(c_k)\big), \qquad q_c = 1 - \eta_c \lambda_c/2. \tag{19}$$

With probability at least $1 - u$, simultaneously for every $k \ge 0$,

$$\begin{aligned} \|c_k - c^\star\|_2 &\le q_c^k \|c_0 - c^\star\|_2 + \frac{2h_c(N,u)}{\lambda_c}(1 - q_c^k), \\ \sup_{U \text{ bounded, nonempty}} \|G_{A^\star,c_k}(U) - G_{A^\star,c^\star}(U)\|_\infty &\le \varepsilon_c(N,k,u) \\ &:= V_0 \left[q_c^k D_c + \frac{2h_c(N,u)}{\lambda_c}(1 - q_c^k) \right], \end{aligned} \tag{20}$$

where a deterministic bound for every initialization in the cube is

$$D_c = \sup_{c \in \mathcal{C}} \|c - c^\star\|_2 = \sqrt{2(C+1)^2 + (C+\gamma)^2}. \tag{21}$$

The operator bound also holds after clipping both outputs to $[-B, B]^n$. No local-initialization or small-noise condition is needed for this theorem.

Proof. Lemma B.6 gives one event controlling the constant matrix $\widehat{H}_c$ and noise vector g_c. On that event, Lemma B.8, with $m = \lambda_c/2$, $L = L_c$, and noise bound $h_c(N, u)$, gives $d_{k+1} \le q_c d_k + \eta_c h_c(N, u)$. Unrolling this scalar recurrence yields

$$d_k \le q_c^k d_0 + \frac{2h_c(N,u)}{\lambda_c}(1 - q_c^k).$$

Projection keeps all iterates in $\mathcal{C}$ and fixes $c^\star$. The deterministic bound $d_0 \le D_c$ follows by maximizing the three coordinate distances from $(1, -1, \gamma)$ over the cube. For every input, (60) gives $\|F(U)(c_k - c^\star)\|_\infty \le V_0 d_k$. This bound is uniform even for inputs selected after training. The event does not depend on initialization, and hence covers every sample-dependent initialization and all iterates simultaneously. Nonexpansiveness proves the clipped-output assertion. □

Remark A.7 (Scope of the global coefficient guarantee). The guarantee is global in the coefficient cube with prescribed attention. For noiseless labels the iterates converge to $c^\star$. For noisy labels the projected-gradient map is a contraction by Lemma B.8; its unique fixed point satisfies the variational optimality condition for the strongly convex quadratic on $\mathcal{C}$. It is therefore the unique constrained empirical minimizer, which need not equal the teacher. Corollary A.16 gives explicit sufficient sample and iteration budgets. The mixture-law Gram remains positive whenever a positive-mass component supplies coefficient excitation, since other Gram contributions are positive semidefinite.

A.1.4 A COMMON TRAINING INTERFACE FOR CONTROL

Corollary A.8 (Training substitutions in the control theorem)**.** *Retain the deployment sampling, covered mixing behavior, exact architecture masks, finite deterministic target-policy class, and context-length conditions. Either of the following frozen checkpoints supplies the training radius in the evaluation, shrinking-set, and final-policy guarantees:*

1. *Use the curriculum checkpoint in Corollary A.3 with $u = \delta/3$, radius $e_{\mathrm{op}} = L_G e$, and the sample and stage budgets of Theorem A.2.*
2. *In the prescribed-attention special case, use Theorem A.6 with $u = \delta/3$ and radius $e_{\mathrm{op}} = \varepsilon_c(N, K_c, \delta/3)$, under Assumption A.5.*

The teacher, context error, contraction margin, feasible-set updates, and deployed policy rule are the same in both cases.

Proof. Each checkpoint is an admissible frozen parameter with a uniform raw-operator error e_{op} relative to the identical teacher. Proposition A.11 uses only this interface and the simultaneous context event, giving

$$E_j = B\rho^{L_j} + \frac{1 - \rho^{L_j}}{\kappa}(e_{\mathrm{op}} + b_j). \tag{22}$$

The joint feasible-set and policy proof in Theorem A.13 then applies unchanged. Uniformity over bounded inputs permits adaptive policies and prefixes without another training event. The allocated training failure and the two context-event failures total at most δ. No independence between these events is needed. □

A.2 SIMULTANEOUS TRAINING AND REUSED-CONTEXT VALIDITY

The deployment sampling law and concentration tools are specified in Appendix B.5. The curriculum or prescribed-attention special case supplies the following uniform raw-operator interface.

Theorem A.9 (Uniform validity on reused context prefixes)**.** *Assume* (73)–(76) *and let* $\mathcal{E}_{\mathrm{train}} \in \mathcal{I}_0$ *be a pretraining event with probability at least* $1 - \delta/3$ *on which the learned raw block satisfies*

$$\sup_{\substack{\mathsf{D}\ \textit{bounded, nonempty} \\ Q \in \mathcal{Q}_B}} \|G_{\widehat{\theta}}(\mathsf{D}, Q) - G_{\theta_\star}(\mathsf{D}, Q)\|_\infty \le e_{\mathrm{op}}. \tag{23}$$

The supremum need only range over the allowed context lengths $m \le M_{\max}$*. Put*

$$u_{\mathrm{occ}} = \frac{\delta}{3M_{\max}}, \qquad m_0 = \lceil m_{\mathrm{ctr}}(u_{\mathrm{occ}}) \rceil .$$

With probability at least $1 - \delta$*, all of the following hold jointly:*

1. *The operator-training bound* (23) *holds.*
2. *For every* $m \in \{m_0, \dots, M_{\max}\}$ *and every assignment of successor addresses, the corresponding ideal empirical map and its clipped version are* ρ*-contractions.*
3. *For every* $m \in \{1, \dots, M_{\max}\}$ *and every* $\pi \in \Pi_{\mathrm{det}}$*, the residual bound* (80) *holds.*

Consequently, a policy, a prefix length, and an evaluation depth may be chosen using the observed trajectory and all earlier computations, provided the selected policy is deterministic and stationary and the selected length lies in $\{m_0, \dots, M_{\max}\}$*. Re-encode that prefix according to* (74)*, initialize* $Q_0 = 0$*, and iterate the same learned clipped block for* $L \ge 0$ *loops. For exact address routing the output obeys*

$$\|Q_L - Q^\tau\|_\infty \le B\rho^L + \frac{1 - \rho^L}{\kappa}\{e_{\mathrm{op}} + b_m^{\mathrm{unif}}(\delta)\}. \tag{24}$$

If each implemented raw block has a uniform routing error at most $\varepsilon_{\mathrm{route}}$*, add this term inside braces. The conclusion is simultaneous over any number of improvement rounds using these policies and prefixes; there is no further failure penalty for the number of inner loops or repeated computations on a prefix.*

Proof. Apply Lemma B.12 to each fixed prefix with $u = u_{\mathrm{occ}}$ and take a union over $1 \le m \le M_{\max}$. The resulting conditional failure probability is at most $\delta/3$. On this event Lemma B.13 applies simultaneously to all lengths $m \ge m_0$; it is already uniform over successor addresses. Lemma B.15 contributes conditional failure probability at most $\delta/3$. Integrate these conditional probabilities over $\mathcal{I}_0$ and combine them with $\mathbb{P}(\mathcal{E}_{\mathrm{train}}^c) \le \delta/3$. This proves the joint event without assuming independence between the training event and the context events.

On the joint event, Proposition A.11 applies with $\varepsilon = e_{\mathrm{op}} + \varepsilon_{\mathrm{route}}$ and the stated residual radius. Its deterministic scalar recurrence proves (24), simultaneously for every selected policy, prefix, and depth. □

Consequences and scope. The threshold can be displayed explicitly as

$$m_0 = \left\lceil \max\left\{ \frac{16ne^\tau H}{\kappa}, \frac{512n^2e^{2\tau}H^2}{\kappa^2} \log \frac{6nM_{\max}}{\delta} \right\} \right\rceil .$$

If $m_0 > M_{\max}$, there is no nonvacuous contraction-certified prefix at that budget; the algorithm must collect more data or report only a trivial bound. Prefixes shorter than m_0 retain the trivial value-gap bound $2B$. These results justify repeated evaluation and improvement on a growing same-task prefix with a fixed covered behavior policy. They do not establish coverage for an arbitrary learned or changing behavior policy, and they do not transform a frozen-policy value-gap sum into online interaction regret. With reused prefixes, the unique deployment-transition count is the largest stored prefix length; the computation spent on repeated evaluation must be counted separately.

Proposition A.10 (Uniformity over predetermined endpoints)**.** *Let* $\mathcal{M} = \{m_1, \dots, m_R\}$ *be a nonempty deterministic finite set of positive prefix lengths, specified before observing the behavior trajectory. Under the modular operator event* (23)*, the joint training/context conclusions hold with*

probability at least $1-\delta$ *simultaneously over these endpoints and every deterministic policy, using*

$$m \geq m_{\mathrm{ctr}}^{\mathrm{mix}}(\delta/(3R)), \qquad b_m^{(R)} = \frac{\beta_{\mathrm{res}}(R,\delta)}{\sqrt{m}},$$
$$\beta_{\mathrm{res}}(R,\delta) = 2Be^{\tau}\sqrt{2\log\frac{6nRP_{\mathrm{det}}}{\delta}}, \qquad m_{\mathrm{ctr}}^{\mathrm{mix}}(u) = \max\left\{\frac{16ne^{\tau}H}{\kappa}, \frac{512n^2e^{2\tau}H^2}{\kappa^2}\log\frac{2n}{u}\right\}. \tag{25}$$

The number of repeated evaluations at an endpoint is unrestricted. It is R, not the largest numerical endpoint, that appears in this union bound.

Proof. Repeat the fixed-prefix occupation proof at each of the R endpoints with failure $\delta/(3R)$. Repeat the fixed-policy residual proof at each endpoint and each of the P_{det} policies with failure $\delta/(3RP_{\mathrm{det}})$. Each group has total failure at most $\delta/3$. Include the training event with failure $\delta/3$. These arguments do not use independence between prefixes. Adaptive policies or adaptive selection among these predetermined endpoints can be inserted after intersecting the events. No union is taken over the number of repeated computations. □

Empirical refinement on the same event. The direct matrices in Lemma B.14 yield a realized factor $q_j = \min\{\rho, \|M_{m_j}^{\pi_j}\|_\infty\}$. Proposition A.12 substitutes it in the finite-loop radius without another union bound. The event and prefix thresholds above remain the statistical premises; a small measured norm alone does not establish them.

A.3 Evaluation, policy improvement, and the joint invariant

For this section let b_j be the prescribed residual radius evaluated at the chosen prefix m_j, valid for the target policy π_j on the simultaneous context event. The radius is a deterministic function of prefix length; its evaluation may be random when the prefix is selected adaptively. Use the uniform-prefix radius from the preceding section, or the predetermined-endpoint radius in Proposition A.10. All inequalities below hold pathwise on that simultaneous event. Let $\varepsilon = e_{\mathrm{op}} + \varepsilon_{\mathrm{route}}$ and

$$E_j = B\rho^{L_j} + \frac{1-\rho^{L_j}}{\kappa}(\varepsilon + b_j). \tag{26}$$

Exact address masks give $\varepsilon_{\mathrm{route}} = 0$. For the finite-logit variant, $\varepsilon_{\mathrm{route}} = 4BC(n-1)e^{-a_{\mathrm{addr}}}$, a per-loop perturbation.

A.3.1 Certificate definitions and protocol

This protocol extends Algorithm 1 to permitted prefix lengths m_j and depths L_j. The headline theorem uses $m_j = m$ and $L_j = L$.

1. Train once using any proved route satisfying (23); freeze $\widehat{\theta}$. Choose a deterministic π_0. Initialize $C_0 = 2B$ and $\mathcal{F}_{-1} = \mathcal{Q}_B$. Fix the covered behavior policy and the admissible endpoint set before collecting the data. Adaptive choices among predetermined endpoints use their simultaneous event; arbitrary choices within a bounded prefix range use the all-prefix event.
2. At improvement j, append behavior transitions until at least m_j records are stored, and use the first m_j records. Construct the target-specific records with $y_t = (S_{t+1}, \pi_j(S_{t+1}))$ from the stored raw successor states.
3. Start $Q_{j,0} = 0$ and perform L_j shared clipped block updates on those records. Set $\widehat{Q}_j = Q_{j,L_j}$ and $\pi_{j+1}(s) \in \arg\max_a \widehat{Q}_j(s,a)$ with fixed tie-breaking.
4. Define $a_j = E_j + \gamma C_j$ and intersect

$$\mathcal{F}_j = \mathcal{F}_{j-1} \cap \{Q \in \mathcal{Q}_B : \left\|Q - \widehat{Q}_j\right\|_\infty \leq a_j\}. \tag{27}$$

For a nonempty $\mathcal{F}_j$, define its coordinatewise envelopes by

$$U_j(x) = \max_{Q\in\mathcal{F}_j} Q(x), \qquad L_j^{\mathrm{env}}(x) = \min_{Q\in\mathcal{F}_j} Q(x).$$

These extrema exist because $\mathcal{F}_j$ is a closed subset of the compact box $\mathcal{Q}_B$. Set

$$\begin{aligned} W_j(s) &= \max_a U_j(s,a) - \max_a L_j^{\mathrm{env}}(s,a), \\ S_j(s) &= \max_a L_j^{\mathrm{env}}(s,a) - L_j^{\mathrm{env}}(s,\pi_{j+1}(s)), \\ Z_j &= \|W_j + S_j\|_\infty / (1-\gamma), \\ C_{j+1} &= \min\{2B, \gamma C_j + 2E_j/(1-\gamma), Z_j\}. \end{aligned} \tag{28}$$

If an intersection is empty, retain it and mark all subsequent accumulated set certificates and widths unavailable. The policy algorithm may continue, but only the reward-bound fallback $2B$ is reported unless a separate bound is independently justified. The usual inherited certificate induction is not continued after this failure. On the joint event proved here, no intersection is empty.

With valid input radii these sets yield confidence certificates; their computation does not alter the greedy actions. Coordinatewise endpoints use max/min updates and require $O(n)$ additional storage. Raw context storage after round j is $O(\max_{i\le j} m_i)$ records, which is $O(m_j)$ for a growing-prefix schedule. Choosing a shorter stored prefix does not collect new transitions.

Proposition A.11 (Finite-loop evaluation on reused data)**.** *On the training event and the uniform context event, $\left\|\widehat{Q}_j - Q^{\pi_j}\right\|_\infty \le E_j$ simultaneously for every chosen policy/prefix pair. The policy may have been chosen using that same prefix, provided the uniform event covers its policy class and length.*

Proof. Fix the realized pair. Let $F = G_{\theta_\star}$ on its target-specific record. The context event gives $\left\|F(Q) - F(\widetilde{Q})\right\|_\infty \le \rho \left\|Q - \widetilde{Q}\right\|_\infty$ and $\|F(Q^{\pi_j}) - Q^{\pi_j}\|_\infty \le b_j$. Clipping is nonexpansive and fixes Q^{π_j}. The learned block and routing perturbations therefore yield $e_{\ell+1} \le \rho e_\ell + \varepsilon + b_j$ with $e_0 \le B$. Unrolling proves (26). No independence between loops is used; the uniform event permits substituting the adaptively chosen policy. □

Proposition A.12 (Evaluation using the empirical teacher norm)**.** *On the existing simultaneous training/context event $\mathcal{E}$, let*

$$\widehat{\rho}_j = \|M^{\pi_j}_{m_j}\|_\infty, \qquad q_j = \min\{\rho, \widehat{\rho}_j\}, \qquad d_j = \varepsilon + b_j.$$

For zero initial memory and any finite depth $L_j \ge 1$, the frozen evaluator satisfies, simultaneously for all permitted policy/prefix choices,

$$\|\widehat{Q}_j - Q^{\pi_j}\|_\infty \le E_j^{\mathrm{emp}} := B q_j^{L_j} + d_j \sum_{\ell=0}^{L_j-1} q_j^\ell \le E_j^{\mathrm{base}} := B\rho^{L_j} + d_j \sum_{\ell=0}^{L_j-1} \rho^\ell = E_j. \tag{29}$$

Here $\varepsilon = e_{\mathrm{op}}$ for the exact masks of the headline theorem; a separately justified uniform routing error may be added as above.

Proof. Lemma B.14 makes q_j a valid teacher Lipschitz factor on $\mathcal{E}$. Clipping fixes Q^{π_j} and is nonexpansive. Adding and subtracting the teacher at the current iterate and at Q^{π_j} therefore gives

$$e_{\ell+1} \le e_{\mathrm{op}} + \varepsilon_{\mathrm{route}} + q_j e_\ell + b_j = d_j + q_j e_\ell, \qquad e_0 \le B.$$

All iterates lie in $\mathcal{Q}_B$, so the all-input operator bound applies at each loop even if the learned map is not contractive. Finite induction gives the first radius. The polynomial $Bq^L + d\sum_{\ell<L} q^\ell$ has nonnegative coefficients for $B, d \ge 0$ and is nondecreasing for $q \ge 0$, proving the ordering. The original event already covers every allowed policy and prefix; the remaining argument is deterministic for every finite depth. □

Theorem A.13 (Joint invariant and finite-resource near-optimality)**.** *Under the preceding procedure and joint confidence event, use either E_j^{base} or E_j^{emp} as E_j. Then for all $j < J$,*

$$D_j := \|V^\star - V^{\pi_j}\|_\infty \le C_j, \qquad \left\|\widehat{Q}_j - Q^\star\right\|_\infty \le a_j = E_j + \gamma C_j, \tag{30}$$

$$Q^\star \in \mathcal{F}_j \subseteq \mathcal{F}_{j-1}, \qquad \mathrm{diam}_\infty \mathcal{F}_j \le 2 \min_{i\le j} a_i, \tag{31}$$

$$0 \le W_j(s) + S_j(s) \le 2a_j, \qquad D_{j+1} \le C_{j+1}. \tag{32}$$

The decision widths $W_j(s)$ are nonincreasing. Moreover,

$$D_J \le \min\left\{ C_J, \gamma^J D_0 + \frac{2}{1-\gamma}\sum_{j=0}^{J-1} \gamma^{J-1-j} E_j \right\}, \tag{33}$$

and the same weighted sum bounds C_J with D_0 replaced by $2B$.

Proof. Initially $D_0 \le 2B = C_0$ and $Q^\star \in \mathcal{Q}_B$. If $D_j \le C_j$, the Bellman equations yield $0 \le Q^\star - Q^{\pi_j} = \gamma P(V^\star - V^{\pi_j}) \le \gamma C_j$ coordinatewise. Proposition A.11 or Proposition A.12, for the selected radius, proves (30), so the current intersection contains $Q^\star$. It is nonempty and compact, and any two points are within $2a_i$ for every past constraint. Nested sets have decreasing upper envelopes and increasing lower envelopes, giving nonincreasing nonnegative W_j.

Containment gives $\Delta_{j+1}(s) := V^\star(s) - Q^\star(s, \pi_{j+1}(s)) \le W_j(s) + S_j(s)$. The newest ball gives $U_j \le \widehat{Q}_j + a_j$ and $L_j^{\mathrm{env}} \ge \widehat{Q}_j - a_j$. Since the deployed action maximizes $\widehat{Q}_j$, $W_j + S_j = \max_a U_j - L_j^{\mathrm{env}}(\cdot, \pi_{j+1}) \le 2a_j$. For the frozen policy, $V^\star - V^{\pi_{j+1}} = (I - \gamma P_S^{\pi_{j+1}})^{-1}\Delta_{j+1}$; the inverse has nonnegative entries and row sums $(1-\gamma)^{-1}$. This proves $D_{j+1} \le Z_j$. Lemma B.17 gives the second candidate certificate, and bounded rewards give $D_{j+1} \le 2B$. Their minimum closes the induction. Unrolling (81) proves the actual-gap bound; unrolling the certificate recursion gives its version initialized at $2B$. The current C_j, not the not-yet-justified C_{j+1}, justified the current set. □

Corollary A.14 (Validity and comparison on a common recorded trace)**.** *Run the recursion* (27)–(28) *with radii E_j^{base} and E_j^{emp}, each using its own incoming C_j. Initialize both at $C_0 = 2B$ and $\mathcal{F}_{-1} = \mathcal{Q}_B$. On $\mathcal{E}$, the empirical-radius recursion retains $Q^\star \in \mathcal{F}_j^{\mathrm{emp}}$ and $D_j \le C_j^{\mathrm{emp}}$. If critics, actual actions, prefixes, and depths are identical in both recursions, then for every round*

$$\mathcal{F}_j^{\mathrm{emp}} \subseteq \mathcal{F}_j^{\mathrm{base}}, \qquad C_{j+1}^{\mathrm{emp}} \le C_{j+1}^{\mathrm{base}}.$$

Thus the existing worst-case resource bound is retained.

Proof. Validity follows from Proposition A.12 and Theorem A.13. Lemma B.18 proves the comparison by induction, using (29). □

Scope and numerical implementation. The common-trace comparison does not order different action trajectories or stopping schedules. Validity still holds for adaptive choices covered by $\mathcal{E}$. The norm can vary with the policy but need not; balanced symmetric contexts can make it policy invariant. No strict improvement or better worst-case exponent is asserted. A certified implementation needs an upward enclosure $\overline{\rho}_j \ge \|M_{m_j}^{\pi_j}\|_\infty$ accounting for numerical error, uses $q_j = \min\{\rho, \overline{\rho}_j\}$ on $\mathcal{E}$, and rounds radii and interval endpoints outward. Compute the geometric polynomial by a stable recurrence near one. A certified lower norm bound exceeding ρ would flag a mismatch with the event; an upper bound exceeding ρ alone is inconclusive. Ordinary floating-point norms and unproved tolerances are diagnostics. Failed certification or empty intersections retain the unavailable-certificate and universal reward-bound $2B$ conventions.

A.4 Separate training, context, and computational resources

Set $\log_+(x) = \log\max\{1, x\}$. All core deployment statements retain the sampling and positive score-margin conditions of Appendix B.5. This section separates sufficient resources for the stated learning and deployment procedure from necessary resources for a specified statistical or computational problem. In particular, a nonzero term in an upper bound is not an actual error floor. Throughout, $R_{\max} > 0$, $0 < \gamma < 1$, $B = R_{\max}/(1-\gamma)$, and $\mathcal{Q}_B = [-B, B]^n$. All deployment statements below use exact address masks. Any separately proved uniform routing error can be added to the operator-error term.

A.4.1 A complete interface from operator error to policy error

For a deterministic stationary policy π, let Q^π and V^π be its action-value and value functions. Write $V^\star$ for the optimal value. Fix one raw behavior prefix of length m before inspecting its outcomes. When evaluating a new policy, re-encode each stored successor state with that policy's action. Let

F_m^π be the raw teacher operator on the resulting record and let $\widehat{F}_m^\pi$ be the frozen learned operator. The following event is the precise interface needed from the operator-training and context results:

$$\begin{aligned} \|F_m^\pi(Q) - F_m^\pi(Q')\|_\infty &\le \rho\|Q - Q'\|_\infty, \qquad 0 \le \rho < 1, \\ \|F_m^\pi(Q^\pi) - Q^\pi\|_\infty &\le b_m, \\ \|\widehat{F}_m^\pi(Q) - F_m^\pi(Q)\|_\infty &\le e_{\mathrm{op}}. \end{aligned} \tag{34}$$

Every inequality holds simultaneously for all deterministic policies and all $Q, Q' \in \mathcal{Q}_B$. In the last line, uniformity over all admissible operator inputs is a sufficient stronger condition. Neither the chosen policies nor the successive memories must be independent of the prefix.

Proposition A.15 (Fixed-prefix resource decomposition)**.** *On the event* (34)*, start every policy evaluation at $Q_{j,0} = 0$, perform $L \ge 1$ clipped learned updates, and greedify their output to obtain π_{j+1}. Use this same prefix and the same frozen parameters in all $J \ge 1$ improvement rounds. Define*

$$E_L = B\rho^L + \frac{1-\rho^L}{1-\rho}(e_{\mathrm{op}} + b_m), \qquad D_j = \|V^\star - V^{\pi_j}\|_\infty.$$

Then every evaluator has error at most E_L, and

$$\begin{aligned} D_J &\le \gamma^J D_0 + \frac{2(1-\gamma^J)}{(1-\gamma)^2}\left[B\rho^L + \frac{1-\rho^L}{1-\rho}(e_{\mathrm{op}} + b_m)\right] \\ &\le 2B\gamma^J + \frac{2B\rho^L}{(1-\gamma)^2} + \frac{2e_{\mathrm{op}}}{(1-\rho)(1-\gamma)^2} + \frac{2b_m}{(1-\rho)(1-\gamma)^2}. \end{aligned} \tag{35}$$

The four terms respectively control the initial policy, finite evaluation depth, operator training, and the finite deployment context.

Proof. Proposition A.11, with the uniform interface (34), gives the stated E_L on every reused policy/prefix pair. Lemma B.17 gives $D_{j+1} \le \gamma D_j + 2E_L/(1-\gamma)$. Summing this recurrence proves (35); the second bound uses $D_0 \le 2B$ and drops factors at most one. No new data or independence assumption enters this summation. □

Corollary A.16 (Training resources for the prescribed-attention special case)**.** *Let $e > 0$ be the required uniform operator-error tolerance. It suffices to choose a positive integer N and a nonnegative integer K_c satisfying*

$$\begin{aligned} N &\ge \max\left\{N_c(u), \left\lceil \frac{128 L_c \sigma^2 V_0^2}{\lambda_c^2 e^2}[3\log 5 + \log(4/u)] \right\rceil\right\}, \\ K_c &\ge \left\lceil \frac{2}{\eta_c \lambda_c} \log\left(\max\left\{1, \frac{2V_0 D_c}{e}\right\}\right)\right\rceil. \end{aligned} \tag{36}$$

Then $\varepsilon_c(N, K_c, u) \le e$ with confidence $1 - u$. For fixed conditioning constants, this requires $N = O(\sigma^2 e^{-2}\log(1/u))$ beyond the excitation threshold and $K_c = O(\log(1/e))$ as $e \downarrow 0$.

Proof. The sample bound gives $h_c(N, u) \le \lambda_c e/(4V_0)$. The inequality $(1-a)^k \le \exp(-ak)$ for $0 \le a \le 1$ and $k \ge 0$ gives $V_0 q_c^{K_c} D_c \le e/2$ under the iteration bound. The noise term in (20) is at most $e/2$. The expression uses $\log\max\{1, \cdot\}$, so it also covers a zero-step choice without evaluating the logarithm at zero. □

Corollary A.17 (One reused context yields a prescribed final-policy error)**.** *Fix $0 < \epsilon < 2B$ and $0 < \delta < 1$. Define*

$$t_\epsilon = \frac{\epsilon(1-\gamma)^2}{12}, \qquad e_\epsilon = \kappa t_\epsilon. \tag{37}$$

Use a uniform operator guarantee with $e_{\mathrm{op}} \le e_\epsilon$ on its $1 - \delta/3$ event, supplied by Corollary A.3 or the prescribed-attention special case in Theorem A.6. Choose integer budgets

$$\begin{aligned} J &= \max\left\{1, \left\lceil \frac{\log(4B/\epsilon)}{-\log\gamma} \right\rceil\right\}, \\ L &= \max\left\{1, \left\lceil \frac{\log_+(B/t_\epsilon)}{-\log\rho} \right\rceil\right\}, \\ m &\ge \left\lceil \max\left\{1, m_{\mathrm{ctr}}^{\mathrm{mix}}(\delta/3), \frac{\beta_{\mathrm{res}}(1,\delta)^2}{e_\epsilon^2}\right\}\right\rceil. \end{aligned} \tag{38}$$

Collect one raw behavior context of length m. Reuse it at every one of the J improvements, re-encoding successor addresses for each target policy. Then, with probability at least $1-\delta$, $D_J \le C_J \le \epsilon$.

Proof. Proposition A.10 applies with $R = 1$ to the single fixed endpoint m, uniformly over all deterministic policies. Training gives $e_{\mathrm{op}} \le e_\epsilon$, the context radius is at most e_ϵ, and $B\rho^L \le t_\epsilon$. Hence every evaluation radius is at most $t_\epsilon + 2e_\epsilon/\kappa = 3t_\epsilon$. The joint invariant implies

$$C_J \le 2B\gamma^J + \frac{2\max_j E_j}{(1-\gamma)^2} \le \epsilon/2 + \frac{6t_\epsilon}{(1-\gamma)^2} = \epsilon.$$

The event was uniform over policies before they were adaptively chosen; no further union over J is needed. □

For the curriculum, with fixed finite-dimensional conditioning constants and fixed confidence, $N = O(1+\sigma^2\epsilon^{-2})$, $m = O(\epsilon^{-2})$ beyond the mixing threshold, and $K_C + K_A, L, J = O(\log(1/\epsilon))$, with their discount, contraction, and training constants explicit in the stated budgets. There are m *unique deployment transitions*, not Jm. There are still JL macro-block evaluations and $2JL$ attention sublayers, each operating on its context tokens. The logarithmic policy-class cost $\log P_{\mathrm{det}} = |\mathcal{S}|\log|\mathcal{A}|$ is included in $\beta_{\mathrm{res}}(1,\delta)$. These choices are noncircular: the sample threshold contains neither $\log m$ nor $\log J$.

Curriculum training budgets. For operator tolerance e, apply Theorem A.2 with parameter tolerance e/L_G, the tolerances in (14), the sample threshold (69), and phase budgets (15). Corollary A.3 then supplies $e_{\mathrm{op}} \le e$ for the preceding deployment budgets. These choices separate the finite optimization residual from the noisy-label error; with noiseless labels, the sample condition only ensures probe presence.

Context specialization and coefficient composition. For a single predetermined endpoint, the context inputs already proved in Proposition A.10 are

$$m \ge m_{\mathrm{ctr}}(\delta/3), \qquad m_{\mathrm{ctr}}(u) := m_{\mathrm{ctr}}^{\mathrm{mix}}(u), \tag{39}$$

$$b_m = \beta_{\mathrm{res}}(1,\delta)/\sqrt{m}. \tag{40}$$

The constants, sampling law, and score margin are those in (5), (1), and (6); stationary coverage alone is insufficient. The two context events and the training event have total failure at most δ, without a union over J or L.

Corollary A.18 (Resources for the prescribed-attention special case)**.** *Under Theorem A.6's coefficient law and the deployment conditions of Corollary A.17, fix $0 < \epsilon < 2B$, $u = \delta/3$, $t_\epsilon = \epsilon(1-\gamma)^2/12$, and $e_\epsilon = \kappa t_\epsilon$. Use the explicit common budgets*

$$\begin{aligned} &(N, K_c) \text{ satisfying (36) with } e = e_\epsilon,\ u = \delta/3, \\ &(L, J, m) \text{ satisfying (38).} \end{aligned} \tag{41}$$

Then $D_J \le C_J \le \epsilon$ with probability at least $1-\delta$. Only one training sample and one deployment prefix are used.

Proof. Corollary A.16 supplies $e_{\mathrm{op}} \le e_\epsilon$ on a single event with failure at most $\delta/3$. The common deployment budget gives $b_m \le e_\epsilon$ and $B\rho^L \le t_\epsilon$, hence $E_L \le 3t_\epsilon$. Corollary A.17 then gives $C_J \le \epsilon/2 + 6t_\epsilon/(1-\gamma)^2 = \epsilon$. All coefficient updates reuse the same examples, and every improvement reuses the same prefix. The empirical refinement can only reduce these certificates on a common trace (Corollary A.14). □

Accounting and rates. The resources in (41) have different units:

Resource	Meaning
N	Independent one-step operator-training examples, not necessarily individual environment transitions.
$\sum_i m_i^{\mathrm{phys}}$	Physical transitions used to create those training examples; algebraic probes are counted separately. This count presumes independent physical records rather than overlapping windows.
$K_C + K_A$ or K_c	Optimization iterations on the same fixed training sample: curriculum or prescribed-attention coefficient training.
m	Unique behavior transitions in the single deployment prefix.
L	Shared evaluator loops for each target policy.
J	Greedy policy-improvement rounds; deployment executes JL loops and, for the specified construction, $2JL$ attention sublayers.

At fixed confidence and fixed conditioning constants, the prescribed-attention special case has sufficient $N = O(\sigma^2\epsilon^{-2})$ beyond the fixed excitation threshold, $m = O(\epsilon^{-2})$ beyond the fixed context threshold, and $K_c, L, J = O(\log(1/\epsilon))$. When $\sigma = 0$, the same theorem requires only $N \geq N_c(u)$, independent of ϵ; the additional accuracy comes from more GD steps. Repeated optimization does not multiply the example count by its number of steps. The operator-training transition cost depends separately on the record-length design. None of these fixed-constant statements suppresses that the constants can be large or can change if the conditioning or architecture changes with ϵ.

A.4.2 Certificate inputs and computability

Computability is conditional on the supplied teacher and justified constants listed below. An observed sample maximum or an uncalibrated Gram eigenvalue does not automatically provide the corresponding population bound.

Table 2: Inputs to a computable policy certificate and their scope.

Quantity	Source	Qualification
$\widehat{\theta}$, critics, policies, W_m, T_m^π	Training, stored records, known teacher scores and target policy	No unknown deployment transition matrix is needed for the empirical matrices.
$\widehat{\rho}_j$ and interval endpoints	Empirical teacher matrix and coordinatewise max/min updates	Use upper norm/radius enclosures and outward rounding of interval endpoints for certification.
$n, \gamma, R_{\max}, B, \tau, C, M_A$	Task family and architecture	Supplied justified bounds, not maxima fitted to one sample.
$\mu_0, C_{\mathrm{mix}}, r_{\mathrm{mix}}$	Assumed or externally justified coverage/mixing bounds	Not estimated from the same deployment record by the present proof.
$\sigma, p_{\min}, \lambda_c$	Bounded label noise and training design	The curriculum uses $\sigma, p_{\min}$; the prescribed-attention special case uses σ, λ_c.
$\rho, b_m^{(R)}, e_{\mathrm{op}}(N, K)$	Confidence level, budgets, supplied constants and formulas	A proved training route supplies the operator radius.
$L_G\|\widehat{\theta} - \theta^\star\|_2$	Known teacher and trained parameters in the prescribed box	Deterministic alternative all-input operator radius; no population excitation estimate is needed to compute it.
C_J, Z_j	Certified evaluation radii and interval recursions	Inherit the validity conditions of their inputs.
$Q^\pi, Q^\star, D_j$, oracle radii	Exact experimental MDP solver	Privileged evaluation quantities; not deployed decision inputs.

With prescribed teacher attention one may use the deterministic radius

$$e_{\mathrm{op}}^{\mathrm{par}} = V_0\|\widehat{c} - c^\star\|_2, \qquad V_0 = \sqrt{R_{\max}^2 + 2B^2}.$$

For the full parameter box use $L_G = 2C(R_{\max} + 2B) + V_0$. These known-teacher bounds document achieved operator accuracy but do not supply a missing context confidence event or a training-sample lower bound.

Remark A.19 (Upper-bound residuals are not actual floors). Holding N, K_c, or L fixed can leave a positive term in (35). This says the displayed upper bound does not vanish under that resource choice; it does not prove a positive lower bound on the algorithm's actual error. For example, on a zero-reward MDP the teacher initialized at zero remains exactly zero at every loop and every policy has zero gap, for every nonempty context and every finite loop count. Necessity requires a separate hard-instance argument, such as the deployment lower bound below.

A.4.3 One explicit score design and the resulting context budget

Assume $n \geq 2$, $0 < \gamma < 1$, and $0 < \mu_0 \leq 1/n$. Write $g = 1 - \gamma$ and $a_\mu = (1 - \mu_0)/\mu_0$. Retain $B = R_{\max}/g$, $P_{\det} = |\mathcal{A}|^{|\mathcal{S}|}$, and the mixing constant H of the context analysis. The following display makes a sufficient deployment budget explicit; it is not a necessary sample-complexity claim.

Corollary A.20 (An explicit score and single-prefix budget)**.** *Choose*

$$e^\tau = a_\mu \frac{3+\gamma}{g}, \qquad \kappa = \frac{g}{4}, \qquad \rho = 1 - \frac{g}{4}, \qquad M_A > \tau. \tag{42}$$

For $0 < \epsilon < 2B$ and $0 < \delta < 1$, require the trained operator error to be at most $e_\epsilon = \epsilon g^3/48$ on its event of probability $1 - \delta/3$. Take

$$\begin{aligned} J &\geq \max\left\{1, \left\lceil \frac{\log(4B/\epsilon)}{-\log\gamma} \right\rceil \right\}, \\ L &\geq \max\left\{1, \left\lceil \frac{\log_+(12R_{\max}/(\epsilon g^3))}{-\log(1-g/4)} \right\rceil \right\}, \qquad (43) \\ m &\geq \left\lceil \max\left\{1, \frac{64na_\mu(3+\gamma)H}{g^2}, \frac{8192n^2a_\mu^2(3+\gamma)^2H^2}{g^4}\log\frac{6n}{\delta}, \right.\right. \\ &\qquad \left.\left. \frac{18432R_{\max}^2a_\mu^2(3+\gamma)^2}{\epsilon^2 g^{10}}\log\frac{6nP_{\det}}{\delta} \right\} \right\rceil. \qquad (44) \end{aligned}$$

Under the sampling, training, and exact-mask conditions of the core route, one prefix of m unique behavior transitions can be reused throughout these J improvements, and the returned policy and its certificate obey $\|V^\star - V^{\pi_J}\|_\infty \leq C_J \leq \epsilon$ with probability at least $1 - \delta$.

Proof. The score choice yields

$$d_\star = \left[1 + a_\mu e^{-\tau}\right]^{-1} = \frac{3+\gamma}{4}, \qquad d_\star - \frac{1+\gamma}{2} = \frac{g}{4}.$$

The occupation threshold in (78) for a single endpoint is

$$m_{\mathrm{ctr}}(\delta/3) = \max\left\{\frac{16ne^\tau H}{\kappa}, \frac{512n^2e^{2\tau}H^2}{\kappa^2}\log\frac{6n}{\delta}\right\}.$$

Substitution gives the second and third terms in (44). The residual radius is

$$b_m = 2Be^\tau\sqrt{2\log(6nP_{\det}/\delta)/m}.$$

Requiring $b_m \leq e_\epsilon$ gives

$$m \geq \frac{8B^2e^{2\tau}}{e_\epsilon^2}\log\frac{6nP_{\det}}{\delta} = \frac{18432R_{\max}^2a_\mu^2(3+\gamma)^2}{\epsilon^2 g^{10}}\log\frac{6nP_{\det}}{\delta}.$$

Let $t_\epsilon = \epsilon g^2/12$, so $e_\epsilon = \kappa t_\epsilon$. The choice of L gives $B\rho^L \leq t_\epsilon$. Thus every evaluation error radius is bounded by $t_\epsilon + 2e_\epsilon/\kappa = 3t_\epsilon$. The simultaneous policy/context event and the certificate recursion (28) give

$$C_J \leq 2B\gamma^J + \frac{2\max_j E_j}{g^2} \leq \frac{\epsilon}{2} + \frac{6t_\epsilon}{g^2} = \epsilon.$$

The endpoint is fixed before data collection and is reused, so no union over J or L occurs. □

What the dependence does and does not say. For balanced coverage $\mu_0 = 1/n$, $a_\mu = n - 1$. The displayed occupation term then scales as $n^4 H^2/g^4$ times its logarithm, while the residual term scales as $n^2 R_{\max}^2/(\epsilon^2 g^{10})$ times $\log(6nP_{\det}/\delta)$. This last logarithm includes $|\mathcal{S}| \log |\mathcal{A}|$ and must not be silently treated as a logarithm of n. If a formula involving only n is desired, the safe bound $\log P_{\det} \le n$ replaces it by $\log(6n/\delta) + n$. With B held fixed instead of $R_{\max}$, the residual term has B^2/g^8. These are conservative sufficient bounds from the present proof; the occupation threshold, noise conditioning, and full training budget remain separate quantities. Larger score separation helps contraction but can worsen the sigmoid-slope conditioning of the curriculum. Prescribing attention does not remove the n, coverage, or discount dependence of deployment.

A.4.4 A STATISTICAL LOWER BOUND ON UNIQUE DEPLOYMENT TRANSITIONS

The following two-task theorem gives the deployment lower bound used in the main text. Its information restriction is specific to this lower bound; the tasks satisfy the core bounded-reward, coverage, and mixing assumptions, and the same exact teacher is available under both tasks.

Theorem A.21 (Two-task deployment lower bound)**.** *Fix $0 < \delta < 1/2$ and $0 < \epsilon \le R_{\max}/[4(1-\gamma)]$. Consider any algorithm that observes m transitions collected by fixed uniform behavior and returns a deterministic stationary policy. The algorithm may have arbitrary side information with a common law under the two tasks, independent of the subsequent context conditional on the task, and unlimited computation. If it is ϵ-optimal with probability at least $1 - \delta$ on every bounded-reward, finite covered mixing MDP in the class below, then necessarily*

$$m \ge \frac{R_{\max}^2}{16(1-\gamma)^2\epsilon^2}(1-2\delta)\log\frac{1-\delta}{\delta}. \tag{45}$$

The class consists of only two one-state, two-action MDPs. Thus the bound already holds at a fixed dimension, with stationary coverage $1/2$ and immediate behavior mixing. Exact teacher parameters and task-independent operator-training data can be included in the side information.

The testing proof is given in Appendix B.7.

Remark A.22 (Scope of the statistical comparison). At fixed confidence, Theorem A.21 matches the ϵ^{-2} exponent of the sufficient deployment bound. It does not match its full dimension, discount, coverage, or mixing dependence. It applies to the stated class allowing bounded stochastic rewards, not specifically to the deterministic-reward experimental ring. It counts unique observations: reusing a prefix in more layers or improvements supplies no further task information. The result does not apply if side information identifies the deployment task or reveals its reward means.

For randomized stationary output policies the same argument works with denominator 64 in place of 16, provided $\epsilon \le R_{\max}/[8(1-\gamma)]$. Set $\Delta = 2\epsilon(1-\gamma)/R_{\max}$. If p is the output probability of action 1, the two gaps are $4\epsilon(1-p)$ and $4\epsilon p$. Successful outputs therefore have $p \ge 3/4$ under M_+ and $p \le 1/4$ under M_-. Testing $p > 1/2$ and repeating the proof yields the asserted constant. This extension concerns stationary policies, not an unspecified adaptive future exploration protocol.

A.5 DETAILED STATEMENTS SUPPORTING THE HEADLINE GUARANTEE

We first obtain a frozen operator, then use its uniform error as the only training input to the deployment theorem. This keeps operator examples distinct from observations of the deployment task.

Derivation of Theorem 4.1. Corollary A.3 supplies the uniform trained-operator guarantee. Intersect its event with the context events in Proposition A.10 for one fixed prefix. The coupled induction of Theorem A.13 and the single-prefix resource bound in Corollary A.17 give (8). The explicit choices in Appendix A.4 then yield (9). The statements below record the operator interface and allow selection among finitely many predetermined prefixes.

Uniform operator interface. Under Assumptions 3.2 and 3.3, apply Theorem A.2 with parameter tolerance $e > 0$, failure probability u, and its explicit sample and phase budgets. From any $(A_0, c_0) \in \Theta$, Corollary A.3 gives, with probability at least $1 - u$,

$$\|\widehat{\theta} - \theta^\star\|_2 \le e, \qquad \sup_{\substack{\mathcal{D} \text{ bounded, nonempty} \\ Q \in \mathcal{Q}_B}} \|G_{\widehat{\theta}}(\mathcal{D}, Q) - G_{\theta^\star}(\mathcal{D}, Q)\|_\infty \le L_G e =: e_{\mathrm{op}}, \tag{46}$$

where $L_G = 2C(R_{\max}+2B)+\sqrt{R_{\max}^2+2B^2}$. This uniformity includes data-dependent memories; it is the only training input needed for the deployment theorem below.

Corollary A.23 (Prescribed-attention special case)**.** *Under Assumption A.5, projected GD with fixed $A^\star$ converges from every $c_0 \in [-C,C]^3$. For $N \ge N_c(u)$ and $0 < \eta_c \le 1/L_c$, its uniform operator radius is*

$$e_{\mathrm{op}} \le V_0 \left[(1-\eta_c\lambda_c/2)^{K_c} D_c + 2h_c(N,u)/\lambda_c\right]$$

with probability at least $1-u$, with constants from Theorem A.6. This bound can replace the operator radius in (46) *in the deployment theorem, without learning attention.*

Fix a deterministic nonempty set $\mathcal{M}$ of R permitted positive prefix lengths and $\delta \in (0,1)$. Define

$$\begin{aligned} \beta_{\mathrm{res}}(R,\delta) &= 2Be^{\tau}\sqrt{2\log(6nRP_{\det}/\delta)}, \qquad b_m^{(R)} = \beta_{\mathrm{res}}(R,\delta)/\sqrt{m}, \\ m_{\mathrm{ctr}}(u) &= \max\left\{\frac{16ne^{\tau}H}{\kappa}, \frac{512n^2e^{2\tau}H^2}{\kappa^2}\log\frac{2n}{u}\right\}, \qquad E_j = B\rho^{L_j} + \frac{1-\rho^{L_j}}{\kappa}\left[e_{\mathrm{op}} + b_{m_j}^{(R)}\right]. \end{aligned} \tag{47}$$

Policies and endpoint selections may depend on context. All-prefix uniformity follows by choosing $\mathcal{M} = \{1,\dots,M_{\max}\}$.

Theorem A.24 (Reused evaluation, policy improvement, and joint certification)**.** *Under Assumption 3.1,* (1) *and* (6)*, use a proved training route satisfying the raw-operator inequality in* (46) *with failure at most $\delta/3$. Run the protocol of Appendix A.3.1, extending Algorithm 1, with exact address masks and $m_j \in \mathcal{M}$, $m_j \ge m_{\mathrm{ctr}}(\delta/(3R))$. With probability at least $1-\delta$, simultaneously for all rounds,*

$$\begin{aligned} \|\widehat{Q}_j - Q^{\pi_j}\|_\infty &\le E_j, & D_j &\le C_j, \\ Q^\star \in \mathcal{F}_j &\subseteq \mathcal{F}_{j-1}, & \operatorname{diam}_\infty \mathcal{F}_j &\le 2\min_{i\le j} a_i, \\ 0 \le W_j(s) + S_j(s) &\le 2a_j, & D_{j+1} &\le C_{j+1}, \end{aligned} \tag{48}$$

and the decision widths $W_j(s)$ are nonincreasing. In particular,

$$D_J \le \min\left\{C_J, \gamma^J D_0 + \frac{2}{1-\gamma}\sum_{j=0}^{J-1}\gamma^{J-1-j}E_j\right\}. \tag{49}$$

The same weighted expression bounds C_J when D_0 is replaced by $2B$.

Appendices A.2–A.3 prove the uniform event and the coupled induction. Repeated computations at an endpoint incur no additional failure penalty in J or L_j. Nesting does not imply monotone returns or monotone full certificates. The sets certify the critic-greedy actions; they do not select a different policy. The empirical teacher norm $\widehat{\rho}_j = \|I - W_{m_j} + \gamma T_{m_j}^{\pi_j}\|_\infty$ refines E_j by replacing ρ with $q_j = \min\{\rho, \widehat{\rho}_j\}$ in the geometric polynomial (Proposition A.12). On the same event and trace, this cannot worsen C_j (Corollary A.14); Appendix A.4.2 identifies the supplied constants needed.

Resources and necessity. Proposition A.15 separates the initial-policy, finite-depth, operator-training, and context errors. Corollary A.17 gives explicit budgets for one reused prefix, and Corollary A.20 displays a sufficient score choice and its dimension, coverage, mixing, and discount dependence. Theorem A.21 matches only the deployment accuracy exponent under task-uninformative side information; it does not assert a lower bound for learning the prescribed teacher's known parameters.

B SUPPORTING RESULTS

The following lemmas support Appendix A, in the order architecture, training, reused-context concentration, control, and the statistical lower bound. The final scope results clarify the depth and score requirements of the prescribed teacher.

B.1 EXACT ARCHITECTURE AND FINITE ROUTING

This subsection proves the exact architecture used by the finite-state result; uniform derivative bounds are given in Appendix B.2. All attention operations below use the ordinary normalized exponential softmax. The construction uses prescribed token masks, residual connections, and a fixed ReLU clipping network; layer normalization is not part of the analyzed block. Thus it is a concrete structured Transformer, with a specified computation, rather than a representation claim for unrestricted decoders.

B.1.1 A COMPLETE TWO-SUBLAYER EVALUATION LOOP

Let $n = |\mathcal{S}||\mathcal{A}| \geq 2$, identify the state-action pairs with $[n]$, and write $B = R_{\max}/(1-\gamma)$, where $R_{\max} > 0$ and $0 < \gamma < 1$. For a nonempty context $\mathsf{D} = \{(x_t, R_{t+1}, y_t)\}_{t=0}^{m-1}$ and $Q \in [-B, B]^n$, define the raw block

$$G_\theta(\mathsf{D}, Q)(x) = Q(x) + \sum_{t=0}^{m-1} w_t^A(x)\{c_r R_{t+1} + c_p Q(x_t) + c_s Q(y_t)\}, \tag{50}$$

$$w_t^A(x) = \frac{\exp(A_{x,x_t})}{\sum_{k=0}^{m-1} \exp(A_{x,x_k})}, \qquad \theta = (A, c_r, c_p, c_s). \tag{51}$$

The parameter set imposes $A_{xn} = 0$, $|A_{xu}| \leq M_A$, and $|c_r|, |c_p|, |c_s| \leq C$. Consequently it has $p = n(n-1) + 3$ free scalar parameters. The gauge removes the invariance under adding a common scalar to all scores in one row. The ideal parameter is

$$A^\star_{xu} = \tau \mathbf{1}\{x = u\} - \tau \mathbf{1}\{x = n\}, \qquad (c_r^\star, c_p^\star, c_s^\star) = (1, -1, \gamma), \tag{52}$$

where $0 < \tau < M_A$ and $C > 1$. Subtracting the row constant leaves its softmax probabilities unchanged. We denote $G_{\theta_\star}$ by F_{D}^π when the successor addresses were re-encoded for target policy π.

Tokens and coordinate projections. Each token is a column vector of width $d = 2n + 4$ with coordinates

$$z = (u, v, r, p, s, q) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^4.$$

Write $E_u, E_v, E_r, E_p, E_s, E_q$ for the corresponding coordinate projections, and $e_p, e_s, e_q \in \mathbb{R}^d$ for the unit vectors that inject into the three indicated scalar coordinates. There are n memory tokens, m transition tokens, and one dummy token:

$$M_x = (e_x, 0, 0, 0, 0, Q(x)), \qquad T_t = (e_{x_t}, e_{y_t}, R_{t+1}, 0, 0, 0), \qquad Z_0 = 0. \tag{53}$$

The u, v, r coordinates are read-only. The two scratch coordinates p, s are overwritten in each loop. The q coordinate contains the persistent value memory on the memory tokens and is zero on all other tokens.

For precision, a scalar-valued attention head at query token i has output

$$o_i = W_O \sum_{k \in \mathcal{M}_A(i)} \frac{\exp((W_Q z_i)^\top (W_K z_k)/\sqrt{d_k})}{\sum_{\ell \in \mathcal{M}_A(i)} \exp((W_Q z_i)^\top (W_K z_\ell)/\sqrt{d_k})} W_V z_k. \tag{54}$$

All masks below are nonempty. Coordinates can be padded with zeros to use a common head dimension $d_k = n$; the scalar notation for W_V is just a convenient description of this padding.

First sublayer: four fixed heads. All four heads use $W_Q = W_K = 0$. For a transition query T_t, their allowed-key sets and value/output maps are

Head	Allowed keys at T_t	W_V	W_O
Predecessor read	$\{M_{x_t}\}$	E_q	e_p
Successor read	$\{M_{y_t}\}$	E_q	e_s
Predecessor reset	$\{T_t\}$	E_p	$-e_p$
Successor reset	$\{T_t\}$	E_s	$-e_s$

At memory and dummy queries, all four heads attend only to Z_0. The outputs are added in one residual sublayer. Each singleton softmax has weight exactly one, independently of its score. If the old scratch values are p_t, s_t, the new scratch values are therefore

$$p_t + Q(x_t) - p_t = Q(x_t), \qquad s_t + Q(y_t) - s_t = Q(y_t).$$

The reset heads are essential: without them, a shared residual layer would accumulate scratch values from previous evaluation loops. Memory values, identifiers, and rewards are unchanged in this sublayer. The masks use only token types and the supplied addresses, so they require no transition model.

Second sublayer: one learned TD aggregation head. A memory query M_x attends to all transition tokens, and every other query attends only to Z_0. Use head dimension n and the explicit matrices

$$W_Q = \sqrt{n}\, A^\top E_u, \qquad W_K = E_u, \qquad W_V = c_r E_r + c_p E_p + c_s E_s, \qquad W_O = e_q. \tag{55}$$

At a memory/transition pair, the scaled dot product is exactly $e_x^\top A e_{x_t} = A_{x,x_t}$. After the first sublayer, the value projection at T_t equals $c_r R_{t+1} + c_p Q(x_t) + c_s Q(y_t)$. Hence the second residual sublayer sets the memory coordinate to (50). Nonmemory tokens receive zero from the dummy and retain memory coordinate zero.

Fixed feed-forward clipping. The final feed-forward residual map leaves every coordinate except q unchanged. With $\mathrm{ReLU}(a) = \max\{a, 0\}$, take four hidden units with outputs

$$h(q) = \big(\mathrm{ReLU}(q + B), \mathrm{ReLU}(q - B), \mathrm{ReLU}(q), \mathrm{ReLU}(-q)\big).$$

Their output into coordinate q is $-B + h_1 - h_2 - h_3 + h_4$. Since $h_3 - h_4 = q$, the residual connection produces

$$q - B + h_1 - h_2 - h_3 + h_4 = -B + \mathrm{ReLU}(q + B) - \mathrm{ReLU}(q - B) = \mathrm{clip}_{[-B,B]}(q). \tag{56}$$

In particular, the dummy and transition memory coordinates stay zero.

Proposition B.1 (Exact loop closure)**.** *Starting with tokens* (53)*, one macro-block described above implements $Q^+ = \Pi_B G_\theta(\mathsf{D}, Q)$ exactly, where Π_B is coordinatewise clipping to $[-B, B]$. The nonmemory tokens keep their identifiers and rewards, and the next loop can use the same matrices and masks. Thus L shared macro-blocks implement L iterations of this map, using token width $2n+4$, four fixed routing heads, one learned aggregation head, and $2L$ attention sublayers. The number of distinct learned parameters is $n(n-1)+3$, independently of m and L.*

Proof. The first sublayer refreshes both scratch values exactly, even when they were left nonzero by a previous loop. Equation (55) then realizes the required logits and TD values, and the residual connection adds them to the current memory. Equation (56) clips precisely this memory. All output projections vanish in the identifier and reward coordinates, and the dummy is invariant. These invariants allow induction on the number of loops. The only free entries are the gauge-fixed score matrix and the three value coefficients. □

B.1.2 Finite-logit address routing

There is also a variant in which the two memory reads use all memory tokens as candidates. Structural token-type masks and the two exact self-reset masks are retained. For the predecessor read, use $W_Q = \sqrt{n}\, a E_u$, $W_K = E_u$; for the successor read, use $W_Q = \sqrt{n}\, a E_v$, $W_K = E_u$, where $a > 0$. For transition queries the matching address has score a and the other $n - 1$ addresses have score zero. At all nontransition queries these heads still attend only to the dummy. Their value/output matrices remain as in the preceding table.

Proposition B.2 (Soft address-routing error)**.** *For either memory read and every $Q \in [-B, B]^n$, its error from the addressed value is at most $2B(n-1)e^{-a}$. If $\widetilde{G}_\theta$ denotes the raw block with these two soft reads, then, uniformly over contexts, bounded memories, and allowed parameters,*

$$\left\|\widetilde{G}_\theta(\mathsf{D}, Q) - G_\theta(\mathsf{D}, Q)\right\|_\infty \le \varepsilon_{\mathrm{route}}(a) := 4BC(n-1)e^{-a}. \tag{57}$$

The same bound holds after applying Π_B. Exact address masks give $\varepsilon_{\mathrm{route}} = 0$.

Proof. For an addressed memory entry v, the total softmax mass on other entries is $(n-1)/(e^a + n - 1) \le (n-1)e^{-a}$. Subtracting $Q(v)$ inside the weighted sum bounds the error by twice B times that mass. The reset heads remove the old scratch entries, so this bound does not accumulate inside a single refresh. The learned aggregation head uses nonnegative weights summing to one; its two scratch coefficients have absolute value at most C. Its error is consequently at most $C \cdot 2B(n-1)e^{-a} + C \cdot 2B(n-1)e^{-a}$, proving (57). Coordinatewise clipping is nonexpansive in sup norm. Across multiple loops, this is a per-loop perturbation; its accumulation is controlled by the evaluation recurrence, rather than being asserted to stay at the one-loop value. □

Optimizer coordinates and metric. The trainable leaf variables are precisely the free entries of A and c. At each forward pass build $W_Q = \sqrt{n}A^\top E_u$ and $W_V = c_r E_r + c_p E_p + c_s E_s$ from these variables; W_K, W_O, all routing maps, and clipping are fixed. Differentiating this computation with respect to (A, c) gives the same loss gradient as the direct raw map, so a full-batch GD step agrees in both implementations. The training loss reads the raw aggregation output *before* the fixed clipping network; clipping is used in the deployment loops. Training on clipped outputs would be a different objective and is not covered by the curvature argument below. Forward equivalence alone does not justify optimizer equivalence. Restricted to the structured subspace $W_Q = \sqrt{n}A^\top E_u$, the Frobenius metric pulls back to $n\,\|dA\|_F^2$, so equal-step GD in those structured W_Q coordinates scales the induced A step by $1/n$. An unrestricted trainable W_Q can additionally acquire entries on the value-memory coordinate, producing logits that depend on Q. Such a parametrization, as well as jointly trained query/key factors, is outside the stated optimizer theorem.

B.2 Derivative bounds, coefficient design, and identifying probes

Use the constants in (10) and the raw-block Jacobian in the free coordinates.

Lemma B.3 (Uniform parameter-to-operator bound)**.** *For every admissible input and $\theta, \vartheta \in \Theta$,*

$$\|G_\theta(U) - G_\vartheta(U)\|_\infty \le L_G \|\theta - \vartheta\|_2. \tag{58}$$

Proof. For a fixed output row x, use expectation under its softmax probabilities and write $v_t = (R_{t+1}, Q(x_t), Q(y_t))$ and $z_t = c^\top v_t$. Then $\|v_t\|_2 \le V_0$ and $|z_t| \le H_v$. An attention perturbation in row x gives scores $a_t = h_{A,x,x_t}$, with the gauge entry interpreted as zero; hence $|a_t| \le \|h_{A,x,\cdot}\|_2$. The first derivative of an expectation is its covariance with a_t. Consequently

$$|DG_\theta(U)(x)[h]| \le 2H_v \|h_{A,x,\cdot}\|_2 + V_0 \|h_c\|_2 \le L_G \|h\|_2.$$

Integrating along the line segment between θ and ϑ in the convex parameter box proves the claim. All expectations are normalized finite sums, so the bound has no record-length factor. □

The bound holds uniformly over bounded nonempty records and does not require independence of transitions within a record.

B.2.1 Exact coefficient design

Fix $A = A^\star$ and write $w_t^\star(x) = w_t^{A^\star}(x)$. Define the $n \times 3$ design matrix by rows:

$$F(U)_{x,:} = \left(\sum_{t<m} w_t^\star(x) R_{t+1}, \quad \sum_{t<m} w_t^\star(x) Q(x_t), \quad \sum_{t<m} w_t^\star(x) Q(y_t) \right). \tag{59}$$

The normalized nonnegative attention weights give

$$\begin{gathered} G_{A^\star,c}(U) = Q + F(U)c, \qquad L_c = nV_0^2, \\ \max_x \|F(U)_{x,:}\|_2 \le V_0, \qquad \|F(U)\|_{\mathrm{op}} \le \|F(U)\|_{\mathrm{F}} \le \sqrt{L_c}. \end{gathered} \tag{60}$$

In particular L_c is a uniform Hessian upper bound, independent of record length. The coefficient excitation condition is (18).

B.2.2 IDENTIFYING PROBES AND NUMERICAL EXCITATION

Choose $r_0 \in (0, R_{\max}]$, and let e_u denote a coordinate vector. Write

$$\ell(a) = (1 + e^{-a})^{-1}, \qquad s_M = \frac{e^{-M_A}}{(1 + e^{-M_A})^2}, \qquad P = n + 2.$$

Then $s_M \le \ell'(a) \le 1/4$ on $[-M_A, M_A]$. The P distinct algebraic inputs are

$$\begin{array}{lll} U_R: & \mathcal{D} = \{(1, r_0, n)\}, & Q = 0, \\ U_P: & \mathcal{D} = \{(1, 0, n)\}, & Q = Be_1, \\ U_S: & \mathcal{D} = \{(1, 0, n)\}, & Q = Be_n, \\ U_u\ (1 \le u < n): & \mathcal{D} = \{(u, r_0, n), (n, 0, n)\}, & Q = 0. \end{array} \tag{61}$$

Let $\mathcal{P} = \{U_R, U_P, U_S, U_1, \dots, U_{n-1}\}$. After subtracting the input memory, the singleton outputs are respectively $r_0 c_r \mathbf{1}$, $B c_p \mathbf{1}$, and $B c_s \mathbf{1}$. On U_u, output x is $r_0 c_r \ell(A_{xu})$. These identities hold throughout Θ; the inputs need not be physical deployment trajectories. Set

$$C_{\mathrm{id}} = \max\left\{ \frac{1 + 2n(n-1)/s_M^2}{n r_0^2}, \quad \frac{1}{nB^2}, \quad \frac{2}{r_0^2 s_M^2} \right\}. \tag{62}$$

Proposition B.4 (Identification and teacher excitation from probes)**.** *For every $\theta \in \Theta$,*

$$\|\theta - \theta_\star\|_2^2 \le C_{\mathrm{id}} \sum_{U \in \mathcal{P}} \|G_\theta(U) - G_{\theta_\star}(U)\|_2^2. \tag{63}$$

If a training law assigns each probe mass at least $w > 0$, then $\mathbb{E}[J_{\theta_\star}(U)^\top J_{\theta_\star}(U)] \succeq (w/C_{\mathrm{id}}) I_p$.

Proof. Denote the squared prediction discrepancy on a probe U by E_U. The three singleton inputs give exactly

$$E_{U_R} = n r_0^2 (c_r - 1)^2, \quad E_{U_P} = nB^2 (c_p + 1)^2, \quad E_{U_S} = nB^2 (c_s - \gamma)^2.$$

For a two-record input U_u, let e_{xu} denote row x's prediction difference. The gauge gives $e_{xu} = r_0[c_r \ell(A_{xu}) - \ell(A^\star_{xu})]$. Since $\ell'(a) \ge s_M$ for $a \in [-M_A, M_A]$, the mean value theorem and $0 \le \ell \le 1$ imply

$$|A_{xu} - A^\star_{xu}|^2 \le \frac{2}{s_M^2} \left(\frac{|e_{xu}|^2}{r_0^2} + |c_r - 1|^2 \right).$$

Sum over the $n(n-1)$ free entries and add the three coefficient errors. Substitution of the singleton identities yields (63) with the stated maximum constant. Taking expectations bounds the squared prediction difference below by $(w/C_{\mathrm{id}})\|\theta - \theta_\star\|_2^2$. For any unit direction v, divide this inequality at $\theta = \theta_\star + tv$ by t^2 and let $t \downarrow 0$. Interiority permits this perturbation; the uniform Jacobian bound permits dominated convergence. The resulting bound is $\mathbb{E}\|J_{\theta_\star}(U)v\|_2^2 \ge w/C_{\mathrm{id}}$ for all v, which proves local excitation. □

Proposition B.5 (Singleton inputs suffice for coefficient excitation)**.** *If U_R, U_P, U_S in* (61) *have probabilities at least $w_R, w_P, w_S > 0$, respectively, then* (18) *holds with*

$$\lambda_c = n \min\{w_R r_0^2, w_P B^2, w_S B^2\}. \tag{64}$$

The remaining inputs may follow any bounded law.

Proof. The three singleton design matrices have every row equal to $(r_0, 0, 0)$, $(0, B, 0)$, and $(0, 0, B)$, respectively. Their Gram contributions dominate $n\,\mathrm{diag}(w_R r_0^2, w_P B^2, w_S B^2)$; every other contribution is positive semidefinite. □

Singleton attention Jacobians vanish, so these inputs identify the three coefficients without identifying any score coordinate. Coefficient excitation is therefore strictly weaker than full-parameter excitation.

Nondegeneracy and interpretation. The construction assumes $n \geq 2$ and $R_{\max} > 0$. If $R_{\max} = 0$, all value and control losses vanish and full excitation is unnecessary. If $n = 1$, predecessor and successor values coincide and c_p, c_s are not separately identifiable. For a constructive training law, probe mass and its sample cost must be reported, including any replacement of ordinary examples at fixed total N. A positive empirical Gram eigenvalue alone does not establish a specified population lower bound.

B.3 One-sample concentration events

The parameter box, teacher, derivative constants, and training law are fixed when the independent examples are drawn. The following events depend only on that sample and hold before any data-dependent initialization or iterate is selected. No additional independence is required when the sample is reused across optimization phases.

B.3.1 Coefficient Gram and centered noise

Write $F_i = F(U_i)$ and define

$$\begin{aligned} \widehat{L}_c(c) &= \frac{1}{2N} \sum_{i=1}^{N} \|Q_i + F_i c - Y_i\|_2^2, \\ \widehat{H}_c &= \frac{1}{N} \sum_{i=1}^{N} F_i^\top F_i, \qquad g_c = \frac{1}{N} \sum_{i=1}^{N} F_i^\top \zeta_i. \end{aligned} \tag{65}$$

The exact identities

$$\nabla \widehat{L}_c(c) = \widehat{H}_c(c - c^\star) - g_c, \qquad \nabla^2 \widehat{L}_c(c) = \widehat{H}_c \tag{66}$$

hold throughout $\mathbb{R}^3$. No local Taylor approximation is used. For $u \in (0, 1)$, let

$$\begin{aligned} N_c(u) &= \left\lceil \frac{8L_c^2}{\lambda_c^2} \left[3 \log 9 + \log(4/u)\right] \right\rceil, \\ h_c(N, u) &= 2\sqrt{L_c}\, \sigma \sqrt{\frac{2}{N} \left[3 \log 5 + \log(4/u)\right]}. \end{aligned} \tag{67}$$

The fixed feature matrix makes the noise vector independent of c, so h_c requires no parameter net and no $\log N$ factor.

Lemma B.6 (One coefficient-training event)**.** *Under Assumption A.5, if $N \geq N_c(u)$, then with probability at least $1 - u$,*

$$\frac{\lambda_c}{2} I_3 \preceq \widehat{H}_c \preceq L_c I_3, \qquad \|g_c\|_2 \leq h_c(N, u). \tag{68}$$

Proof. For a fixed unit vector $v \in \mathbb{R}^3$, $\|F_i v\|_2^2 \in [0, L_c]$. Hoeffding's inequality gives

$$\mathbb{P}\{|v^\top (\widehat{H}_c - H_c) v| > \lambda_c/4\} \leq 2 \exp[-N\lambda_c^2/(8L_c^2)].$$

A deterministic $1/4$-net of the unit sphere has at most 9^3 points. Indeed, a maximal ε-separated set is a net; the disjoint balls of radius $\varepsilon/2$ around its points lie in the ball of radius $1 + \varepsilon/2$, giving cardinality at most $(1 + 2/\varepsilon)^3$ by volume comparison. For a symmetric matrix M, approximation of a maximizing unit vector by a point in this net yields

$$\|M\|_{\mathrm{op}} \leq 2 \max_{v \text{ in the } 1/4\text{-net}} |v^\top M v|,$$

because the discrepancy of the two quadratic forms is at most $2(1/4)\|M\|_{\mathrm{op}}$. A union bound therefore gives $\|\widehat{H}_c - H_c\|_{\mathrm{op}} \leq \lambda_c/2$ with failure at most $u/2$ when $N \geq N_c(u)$. The upper bound $\widehat{H}_c \preceq L_c I_3$ holds deterministically.

For the noise term, $v^\top F_i^\top \zeta_i$ is centered, by conditional centering of ζ_i, and is bounded in absolute value by $\sqrt{L_c}\sigma$. For $\sigma > 0$, Hoeffding's inequality gives

$$\mathbb{P}\{|v^\top g_c| > t\} \leq 2 \exp[-Nt^2/(2L_c\sigma^2)].$$

Use a deterministic $1/2$-net with at most 5^3 points and take $t = \sqrt{L_c}\sigma\sqrt{(2/N)[3\log 5 + \log(4/u)]}$. Except on an event of probability $u/2$, every net projection is at most t. Approximating $g_c/\|g_c\|_2$ gives $\|g_c\|_2 \le 2t = h_c(N,u)$. When $\sigma = 0$, the same assertion holds deterministically. A union bound combines the two events; their independence is unnecessary. □

B.3.2 Recognizable probe counts and label means

For each probe a, let $I_a = \{i : U_i = U_a\}$ and $N_a = |I_a|$. When $N_a > 0$, let $\bar{Y}_a = N_a^{-1}\sum_{i\in I_a} Y_i$ and $\bar{\zeta}_a = N_a^{-1}\sum_{i\in I_a}\zeta_i$.

Lemma B.7 (Simultaneous probe counts and label means)**.** *Under Assumption A.1, fix $t > 0$ and $u \in (0,1)$. Set*

$$N_{\mathrm{probe}}(t,u) = \left\lceil \max\left\{1, \frac{8}{p_{\min}}\log\frac{2P}{u}, \frac{4\sigma^2}{p_{\min}t^2}\log\frac{4nP}{u}\right\}\right\rceil. \tag{69}$$

If $N \ge N_{\mathrm{probe}}(t,u)$, then with probability at least $1-u$,

$$N_a \ge Np_{\min}/2 > 0, \qquad \max_{a,x}|\bar{\zeta}_a(x)| \le t \tag{70}$$

simultaneously for all P probes and n output coordinates. For $\sigma = 0$, the last term in (69) *is zero and the noise assertion holds deterministically.*

Proof. Write $p_a = \mathbb{P}(U = U_a) \ge p_{\min}$. The binomial lower-tail Chernoff bound gives

$$\mathbb{P}\{N_a < Np_{\min}/2\} \le \mathbb{P}\{N_a < Np_a/2\} \le e^{-Np_a/8} \le e^{-Np_{\min}/8}.$$

The count threshold makes the union over the P probes at most $u/2$. Conditional on the input-membership indicators, noises in each probe group are independent, centered in every coordinate, and bounded in absolute value by σ. Therefore, for $\sigma > 0$, Hoeffding's inequality gives

$$\mathbb{P}\{|\bar{\zeta}_a(x)| > t \mid (I_b)_b\} \le 2\exp[-N_a t^2/(2\sigma^2)].$$

On the count event this is at most $2\exp[-Np_{\min}t^2/(4\sigma^2)]$. The last sample threshold makes its union over nP coordinates at most $u/2$, uniformly over the conditioned memberships. Integrating and adding the count failure proves (70). No independence between the count and noise events is asserted or needed. □

B.4 Optimization steps and the need for a curriculum

B.4.1 Projected quadratic and attention steps

Lemma B.8 (Projected quadratic step)**.** *Let $\mathcal{K} \subset \mathbb{R}^q$ be closed and convex, and let $x_\star \in \mathcal{K}$. Suppose H is symmetric with $mI_q \preceq H \preceq LI_q$, where $0 < m \le L$, and $\|g\|_2 \le h$. For $0 < \eta \le 1/L$, define*

$$T_g(x) = \Pi_{\mathcal{K}}\{x - \eta[H(x - x_\star) - g]\}.$$

Then for every $x, y \in \mathcal{K}$,

$$\|T_g(x) - x_\star\|_2 \le (1-\eta m)\|x - x_\star\|_2 + \eta h,$$
$$\|T_g(x) - T_g(y)\|_2 \le (1-\eta m)\|x - y\|_2.$$

Proof. Projection is nonexpansive and fixes $x_\star$. The eigenvalues of $I_q - \eta H$ lie in $[0, 1-\eta m]$, so its operator norm is at most $1-\eta m$. Apply this bound to $(I_q - \eta H)(x - x_\star) + \eta g$ for the first claim and to $(I_q - \eta H)(x - y)$ for the second. □

Lemma B.9 (Normalized attention step)**.** *Fix $\bar{c}_r \in [1/2, 3/2]$, a vector $\nu \in \mathbb{R}^d$, and $\widehat{t}_j = \ell(A_j^\star) + \nu_j$ for the free score coordinates $j = (x,u)$, $1 \le x \le n$, $1 \le u < n$. For the normalized objective $\mathcal{L}_A(A) = \frac{1}{2}\sum_j[\bar{c}_r\ell(A_j) - \widehat{t}_j]^2$ on the score box, its gradient decomposes as*

$$g_j(A) = \bar{c}_r^2\ell'(A_j)[\ell(A_j) - \ell(A_j^\star)] + \bar{c}_r\ell'(A_j)[(\bar{c}_r - 1)\ell(A_j^\star) - \nu_j]. \tag{71}$$

Set

$$m_A = \bar{c}_r^2 s_M^2, \qquad E_A = \frac{|\bar{c}_r|}{4}\left(\sqrt{d}\,|\bar{c}_r - 1| + \|\nu\|_2\right).$$

For $A_k \in [-M_A, M_A]^d$ *and* $0 < \eta_A \le 64/9$*, the projected step* $A_{k+1} = \Pi_{[-M_A,M_A]^d}\{A_k - \eta_A \nabla \mathcal{L}_A(A_k)\}$ *satisfies*

$$d_{k+1}^A \le (1 - \eta_A m_A) d_k^A + \eta_A E_A, \qquad d_k^A = \|A_k - A^\star\|_2. \tag{72}$$

Proof. Differentiate each squared residual and add and subtract $\bar{c}_r \ell(A_j^\star)$ to obtain (71). For a point ξ_j between A_j and $A_j^\star$, the mean value theorem gives

$$\ell'(A_j)[\ell(A_j) - \ell(A_j^\star)] = \ell'(A_j)\ell'(\xi_j)(A_j - A_j^\star), \qquad s_M^2 \le \ell'(A_j)\ell'(\xi_j) \le 1/16.$$

Both points lie in the declared interval. Thus the clean part of the unprojected step multiplies each teacher displacement by a factor in $[0, 1 - \eta_A m_A]$, since $\eta_A \bar{c}_r^2/16 \le (64/9)(9/4)/16 = 1$. The remaining gradient vector has norm at most E_A, by $\ell' \le 1/4$, $0 < \ell < 1$, and the triangle inequality. Finally, projection fixes $A^\star$ and is nonexpansive, proving (72). Projection supplies score-box containment at each step before the next mean value estimate is invoked. □

B.4.2 Why teacher excitation does not imply global joint PL

Proposition B.10 (Teacher excitation does not imply global joint PL)**.** *There is a bounded finite-support i.i.d. noiseless teacher-label law that assigns positive mass to every probe in* (61)*, has a positive definite full teacher Gram matrix, and has a nonoptimal interior stationary point of the raw population loss. Its score choice can simultaneously satisfy the manuscript's deployment contraction-margin condition.*

Proof. Set $n = 2$, $R_{\max} = r_0 = 1$, $\gamma = 1/4$, $B = 4/3$, and $\tau = \log 3$. Take any $M_A > \tau$ and $C > 1$. Give each of U_R, U_P, U_S, U_1 probability $4/35$, and give probability $19/35$ to

$$V: \quad \mathcal{D} = \{(1, 1, 2), (2, -1, 2)\}, \qquad Q = 0.$$

Use labels $Y = G_{\theta_\star}(U)$ and define $f(\theta) = \frac{1}{2}\mathbb{E}\|G_\theta(U) - G_{\theta_\star}(U)\|_2^2$. The two free teacher scores are $(\tau, -\tau)$. Consider

$$\bar{A} = (-\tau, \tau), \qquad \bar{c} = (0, -1, 1/4).$$

This point lies in the interior of the free-coordinate box. Its memory coefficients agree with the teacher. On the reward-only inputs U_R, U_1, V, the reward coefficient is zero, so all score derivatives vanish. All singleton score derivatives vanish independently of the coefficients. Thus the score gradient is zero. The memory-coefficient gradient is zero because only U_P, U_S depend on those coefficients and their residuals are zero.

Let $v_A(U)$ denote the reward-feature vector on a reward-only input. The derivative in c_r at this point is $-\mathbb{E}\langle v_{\bar{A}}(U), v_{A^\star}(U)\rangle$. Since $\ell(\tau) = 3/4$ and $\ell(-\tau) = 1/4$, the inner products on U_R, U_1, V are 2, $3/8$, and $-1/2$, respectively. Hence

$$\partial_{c_r} f(\bar{\theta}) = -\left(\frac{4}{35}\,2 + \frac{4}{35}\frac{3}{8} - \frac{19}{35}\frac{1}{2}\right) = 0.$$

On the same three inputs the squared residuals are 2, $5/8$, and $1/2$. Consequently

$$f(\bar{\theta}) = \frac{1}{2}\left(\frac{4}{35}\,2 + \frac{4}{35}\frac{5}{8} + \frac{19}{35}\frac{1}{2}\right) = \frac{2}{7} > 0, \qquad f(\theta_\star) = 0.$$

To verify full teacher excitation directly, suppose a direction h has zero quadratic form under the teacher Gram. Each positive-probability probe must then have $J_{\theta_\star}(U)h = 0$. The three singleton identities imply $h_r = h_p = h_s = 0$. The two rows of U_1 then give $r_0 \ell'(A_{x1}^\star) h_{x1} = 0$, so both score coordinates of h vanish. The Gram is therefore positive definite despite the nonoptimal stationary point.

For compatibility with the deployment assumptions, choose a one-state, two-action deployment MDP with uniform independent behavior actions and $\mu_0 = 1/2$. Its teacher diagonal-mass bound is $d_\star = 3/4$, and $d_\star - (1 + \gamma)/2 = 1/8 > 0$. Training inputs need not be physical trajectories of this deployment MDP. Finally, a PL inequality with positive constant fails at $\bar{\theta}$, and ordinary joint GD initialized exactly there remains there for every step size. □

Remark B.11 (What the counterexample excludes). The obstruction shows that bounded teacher labels and positive probe mass do not imply global joint PL or joint-GD convergence from every initialization for the full mixture loss. It does not rule out success from many random initializations or an enlarged, explicitly restricted region. Positive probe mass establishes teacher excitation, but the gradients of the probe and ordinary losses can cancel away from the teacher. Their positive-semidefinite teacher Gram matrices do not prevent this cancellation.

B.5 Sampling, concentration, and empirical contraction

B.5.1 Precisely specified deployment experiment

Let $\mathcal{S}$ and $\mathcal{A}$ be finite sets, let $n = |\mathcal{S}||\mathcal{A}| \geq 2$, and identify state–action pairs with $[n]$. Fix one discounted MDP throughout deployment, with discount $\gamma \in (0,1)$, transition kernel P, and rewards satisfying $|R_{t+1}| \leq R_{\max}$. Write

$$r(s,a) = \mathbb{E}[R_{t+1} \mid S_t = s, A_t = a], \qquad B = \frac{R_{\max}}{1-\gamma}, \qquad \mathcal{Q}_B = [-B, B]^n.$$

The statements below hold conditionally on an initial sigma-field $\mathcal{I}_0$ containing pretraining data, the trained weights, the chosen behavior policy, and, for purposes of analysis, the fixed deployment MDP. Conditional on $\mathcal{I}_0$, subsequent environment randomness obeys the stated MDP sampling law. In particular, after the current pair $X_t = (S_t, A_t)$ is included in the history $\mathcal{H}_t$,

$$\mathbb{E}[R_{t+1} \mid \mathcal{H}_t] = r(X_t), \qquad \mathbb{P}(S_{t+1} = s' \mid \mathcal{H}_t) = P(s' \mid X_t).$$

The initial pair may be arbitrarily distributed and may depend on $\mathcal{I}_0$. A fixed stationary behavior b draws the next behavior action with conditional law $b(\cdot \mid S_{t+1})$. Hence X_t is a homogeneous chain with kernel

$$K_b((s,a),(s',a')) = P(s' \mid s,a)b(a' \mid s').$$

More explicitly, one can take $\mathcal{H}_t = \sigma(\mathcal{I}_0, X_0, (R_v, S_v, A_v)_{1 \leq v \leq t})$. For every bounded function f on $[n]$, the stated transition and behavior sampling laws imply $\mathbb{E}[f(X_{t+1}) \mid \mathcal{H}_t] = (K_b f)(X_t)$. The behavior may be selected before deployment using pretraining, provided all bounds asserted below hold for that selection. It is not silently replaced by each newly improved target policy.

Assume that K_b has a stationary distribution μ_b and constants $\mu_0 > 0$, $C_{\mathrm{mix}} < \infty$, and $r_{\mathrm{mix}} \in [0,1)$ such that

$$\min_x \mu_b(x) \geq \mu_0, \qquad \sup_z \left\| K_b^k(z,\cdot) - \mu_b \right\|_{\mathrm{TV}} \leq C_{\mathrm{mix}} r_{\mathrm{mix}}^k \quad (k \geq 0), \qquad H = \frac{C_{\mathrm{mix}}}{1 - r_{\mathrm{mix}}}. \tag{73}$$

Total variation here is the supremum over measurable events. The constants are deterministic bounds valid conditional on $\mathcal{I}_0$; one can instead condition throughout on any fixed choice satisfying them. For $r_{\mathrm{mix}} = 0$, the $k = 0$ term is interpreted as one. In finite state space, irreducibility and aperiodicity guarantee the existence of geometric-mixing constants; the explicit lower bound μ_0 and quantitative mixing rate are still substantive coverage requirements.

Episode resets and state definition. Episodes may refer to the same fixed task without changing its unknown parameters. Nevertheless, concatenating episodes does not by itself produce the chain above. This theorem applies to a continuing trajectory, or to a reset-aware homogeneous Markov formulation in which the reset convention is included in the transition law and the value function being analyzed. An externally imposed reset jump must not be recorded as an ordinary transition of an unchanged kernel P. A protocol using separate finite blocks and an unchanged within-episode kernel needs a blockwise occupation proof, with its boundary terms accounted for. Likewise, a partially observed task requires an explicitly sufficient Markov state; finiteness of the raw observation space alone is insufficient.

Let $M_{\max} \geq 1$ be a deterministic upper bound on the number of deployment transitions used. Store the raw records

$$\mathsf{D}_m^{\mathrm{raw}} = \{(S_t, A_t, R_{t+1}, S_{t+1}) : 0 \leq t < m\}, \qquad 1 \leq m \leq M_{\max}.$$

Define the deterministic stationary policy class

$$\Pi_{\mathrm{det}} = \{\pi : \mathcal{S} \longrightarrow \mathcal{A}\}, \qquad |\Pi_{\mathrm{det}}| = |\mathcal{A}|^{|\mathcal{S}|}.$$

For each target $\pi \in \Pi_{\mathrm{det}}$, re-encode the same raw records using successor addresses

$$Y_t^\pi = (S_{t+1}, \pi(S_{t+1})), \qquad \mathsf{D}_m^\pi = \{(X_t, R_{t+1}, Y_t^\pi) : 0 \le t < m\}. \tag{74}$$

This operation requires the observed successor state and the current policy table, but no transition model and no new environment samples. The successor address is not generally the behavior pair X_{t+1}. Keep these addresses fixed during all inner evaluation loops for a given target. Fix deterministic tie-breaking in every greedy improvement so that all deployed targets stay in Π_{det}.

For $\tau > 0$, put

$$k_x(u) = \exp(\tau \mathbf{1}\{x = u\}), \qquad w_{t,m}(x) = \frac{k_x(X_t)}{\sum_{v=0}^{m-1} k_x(X_v)},$$

$$F_m^\pi(Q)(x) = Q(x) + \sum_{t=0}^{m-1} w_{t,m}(x)\{R_{t+1} + \gamma Q(Y_t^\pi) - Q(X_t)\}, \qquad \overline{F}_m^\pi = \Pi_B \circ F_m^\pi. \tag{75}$$

Here Π_B clips each coordinate to $[-B, B]$. This is exactly the ideal softmax TD block of our architecture: subtracting its gauge row constant from the logits does not change $w_{t,m}$. Choose finite τ such that

$$d_\star = \left[1 + \frac{1-\mu_0}{\mu_0} e^{-\tau}\right]^{-1}, \qquad \kappa = d_\star - \frac{1+\gamma}{2} > 0, \qquad \rho = 1 - \kappa. \tag{76}$$

Since $\mu_0 > 0$ and $\gamma < 1$, such a finite τ exists. Moreover $0 < \kappa < (1-\gamma)/2$ and $0 < \rho < 1$.

B.5.2 Occupation concentration and empirical contraction

Lemma B.12 (Geometric-mixing occupation bound)**.** *Under* (73)*, for every fixed* $m \ge 1$ *and* $u \in (0,1)$*, conditional on* $\mathcal{I}_0$*, with probability at least* $1-u$*,*

$$\max_x |\widehat{\mu}_m(x) - \mu_b(x)| \le d_m(u) := \frac{2H}{m} + 2H\sqrt{\frac{2\log(2n/u)}{m}}, \qquad \widehat{\mu}_m(x) = \frac{1}{m}\sum_{t=0}^{m-1} \mathbf{1}\{X_t = x\}. \tag{77}$$

This conclusion is uniform over the initial distribution of X_0*.*

Proof. Fix x and let $f_x(z) = \mathbf{1}\{z = x\}$. Define

$$h_x(z) = \sum_{k=0}^{\infty} \{K_b^k f_x(z) - \mu_b(f_x)\}.$$

By the total-variation bound and $0 \le f_x \le 1$, the series converges absolutely and uniformly, and $\|h_x\|_\infty \le H$. Termwise summation is valid and yields the Poisson identity $h_x - K_b h_x = f_x - \mu_b(f_x)$. Consequently,

$$\sum_{t=0}^{m-1} \{f_x(X_t) - \mu_b(f_x)\} = h_x(X_0) - h_x(X_m) + \sum_{t=0}^{m-1} \{h_x(X_{t+1}) - K_b h_x(X_t)\}.$$

The boundary term has absolute value at most $2H$. The summands in the remaining sum are martingale differences conditional on $\mathcal{H}_t$, since the next behavior pair has kernel K_b; each has absolute value at most $2H$. Conditional Azuma–Hoeffding therefore gives

$$\mathbb{P}\left(\left|\sum_{t=0}^{m-1} \{h_x(X_{t+1}) - K_b h_x(X_t)\}\right| > 2H\sqrt{2m\log(2n/u)} \,\middle|\, \mathcal{I}_0\right) \le \frac{u}{n}.$$

A union bound over x, followed by division by m, proves the claim. If the stored record ends at S_m, one may draw the final behavior action A_m solely to define X_m for this argument. This changes none of the records, occupancies, or empirical maps. No stationarity assumption on X_0 was used. □

Lemma B.13 (Population fixed point and uniform empirical contraction)**.** *Define, for a probability vector η,*

$$W(\eta)_{xu} = \frac{k_x(u)\eta(u)}{Z_x(\eta)}, \qquad Z_x(\eta) = \sum_v k_x(v)\eta(v).$$

For each $\pi \in \Pi_{\mathrm{det}}$, let P_X^π be the target state–action transition matrix. The population map

$$F_{\mu_b}^\pi(Q) = Q + W(\mu_b)\{r + \gamma P_X^\pi Q - Q\}$$

fixes Q^π and contracts in sup norm with factor at most $1 - 2\kappa$. For $u \in (0,1)$, define

$$m_{\mathrm{ctr}}(u) = \max\left\{\frac{16ne^\tau H}{\kappa}, \frac{512n^2e^{2\tau}H^2}{\kappa^2}\log\frac{2n}{u}\right\}. \tag{78}$$

Whenever $m \geq m_{\mathrm{ctr}}(u)$, the occupation event in Lemma B.12 implies that F_m^π and $\overline{F}_m^\pi$ contract in sup norm with factor at most ρ, simultaneously for every $\pi \in \Pi_{\mathrm{det}}$. In fact the empirical contraction holds for every assignment of successor addresses, including assignments chosen using the whole record.

Proof. Each $W(\eta)$ is stochastic, $Z_x(\eta) \geq 1$, and

$$W(\mu_b)_{xx} = \left[1 + \frac{1-\mu_b(x)}{\mu_b(x)}e^{-\tau}\right]^{-1} \geq d_\star.$$

The Bellman equation gives the population fixed-point identity. The induced matrix sup norm obeys

$$\|I - W(\mu_b) + \gamma W(\mu_b)P_X^\pi\|_\infty \leq 2\max_x\{1 - W(\mu_b)_{xx}\} + \gamma \leq 2(1-d_\star) + \gamma = 1 - 2\kappa.$$

For any two probability vectors η, η', adding and subtracting $k_x(u)\eta'(u)/Z_x(\eta)$ gives

$$\begin{aligned}\sum_u |W(\eta)_{xu} - W(\eta')_{xu}| &\leq \frac{\sum_u k_x(u)|\eta(u) - \eta'(u)|}{Z_x(\eta)} + \frac{|Z_x(\eta') - Z_x(\eta)|}{Z_x(\eta)} \\ &\leq 2e^\tau\|\eta - \eta'\|_1.\end{aligned}$$

Under (78), the two terms in $d_m(u)$ are each at most $\kappa/(8ne^\tau)$. Thus the occupation event implies

$$W(\widehat{\mu}_m)_{xx} \geq W(\mu_b)_{xx} - 2ne^\tau d_m(u) \geq d_\star - \frac{\kappa}{2}.$$

For any assigned successor addresses $y_t \in [n]$, set

$$W_{xu} = \sum_t w_{t,m}(x)\mathbf{1}\{X_t = u\} = W(\widehat{\mu}_m)_{xu}, \qquad T_{xv} = \sum_t w_{t,m}(x)\mathbf{1}\{y_t = v\}.$$

Both W and T are stochastic. The linear part of the corresponding raw empirical map is $I - W + \gamma T$, whose norm is at most

$$2\max_x(1 - W_{xx}) + \gamma \leq 2(1 - d_\star + \kappa/2) + \gamma = 1 - \kappa = \rho.$$

No probability statement about the successor addresses is needed in this deterministic calculation. Coordinatewise clipping is nonexpansive in sup norm, so it preserves the contraction factor. □

B.5.3 The empirical teacher matrix and its exact norm

Lemma B.14 (Direct matrix representation of the empirical teacher)**.** *For a nonempty prefix and deterministic target π, use the teacher weights $w_{t,m}(x)$ in* (75) *and form*

$$\begin{aligned}(W_m)_{xu} &= \sum_{t<m} w_{t,m}(x)\mathbf{1}\{X_t = u\}, & (T_m^\pi)_{xv} &= \sum_{t<m} w_{t,m}(x)\mathbf{1}\{Y_t^\pi = v\}, \\ (r_m^w)_x &= \sum_{t<m} w_{t,m}(x)R_{t+1}, & M_m^\pi &= I - W_m + \gamma T_m^\pi.\end{aligned} \tag{79}$$

Then W_m, T_m^π are row stochastic and $F_m^\pi(Q) = r_m^w + M_m^\pi Q$. Its exact sup-norm Lipschitz constant is $\widehat{\rho}_m^\pi = \|M_m^\pi\|_\infty = \max_x\sum_v|(M_m^\pi)_{xv}|$. On the occupation event and threshold of Lemma B.13, $\widehat{\rho}_m^\pi \leq \rho < 1$ simultaneously over target policies.

Proof. Weights are nonnegative and sum to one for each query, which proves row stochasticity. Collect predecessor and successor coefficients in (75) to obtain the affine identity. The induced norm is the maximum absolute row sum; choosing the coordinate signs of a maximizing row attains it. Differences of points in the box contain a scaled copy of each such sign vector, so this is also its Lipschitz constant on $\mathcal{Q}_B$. The final inequality is the matrix estimate in Lemma B.13. □

Direct weighted-address sums are defined even when an address is unvisited; no conditional transition row or unknown deployment transition matrix is needed. This is a property of the empirical teacher, not a contraction guarantee for the learned update or a substitute for the residual event.

B.5.4 Uniform residuals permit target-policy selection after observation

Lemma B.15 (A finite-policy, finite-prefix Bellman event)**.** *For any $\delta \in (0,1)$, conditional on $\mathcal{I}_0$, with probability at least $1-\delta/3$, the following holds simultaneously for every $1 \le m \le M_{\max}$ and every $\pi \in \Pi_{\det}$:*

$$\|F_m^\pi(Q^\pi) - Q^\pi\|_\infty \le b_m^{\mathrm{unif}}(\delta) := 2Be^\tau \sqrt{\frac{2\log(6nM_{\max}|\mathcal{A}|^{|\mathcal{S}|}/\delta)}{m}}. \tag{80}$$

The same bound holds with F_m^π replaced by $\overline{F}_m^\pi$. This lemma requires the MDP sampling law but does not require mixing or stationary coverage.

Proof. First fix a deterministic stationary policy π, before taking any union bound, and define

$$\xi_t^\pi = R_{t+1} + \gamma Q^\pi(S_{t+1}, \pi(S_{t+1})) - Q^\pi(X_t).$$

Conditional on $\mathcal{I}_0$, the function Q^π is fixed. Its Bellman equation and the MDP sampling law give $\mathbb{E}[\xi_t^\pi \mid \mathcal{H}_t] = 0$. Since $\|Q^\pi\|_\infty \le B$, also

$$|\xi_t^\pi| \le R_{\max} + (1+\gamma)B = 2B.$$

For a fixed query x, $k_x(X_t)$ is $\mathcal{H}_t$-measurable and belongs to $[1, e^\tau]$. Therefore $\sum_{t<m} k_x(X_t)\xi_t^\pi$ is a martingale sum with increment bound $2Be^\tau$. Conditional Azuma–Hoeffding and a union over the n queries imply, for each fixed (m,π) and $u \in (0,1)$,

$$\max_x \left| \frac{1}{m} \sum_{t<m} k_x(X_t)\xi_t^\pi \right| \le 2Be^\tau \sqrt{\frac{2\log(2n/u)}{m}}$$

except on a conditional event of probability at most u. But

$$F_m^\pi(Q^\pi)(x) - Q^\pi(x) = \frac{m^{-1}\sum_{t<m} k_x(X_t)\xi_t^\pi}{m^{-1}\sum_{t<m} k_x(X_t)}, \qquad m^{-1}\sum_{t<m} k_x(X_t) \ge 1.$$

Thus the same upper bound controls the residual. This argument applies Azuma to the unnormalized numerator: the final normalized weights contain future observations in their denominator and are not asserted to be predictable.

Set $u = \delta/(3M_{\max}|\Pi_{\det}|)$ and take a finite union bound over all m and all π. No independence among these events is required, although every event is evaluated using the same raw trajectory. The resulting total conditional failure probability is at most $\delta/3$, and substitution yields (80). Finally, $\Pi_B Q^\pi = Q^\pi$ and clipping is nonexpansive, proving the clipped claim. □

What the policy union means. The proof fixes each π when establishing its martingale bound and only then intersects the events. On their intersection, one may insert any data-dependent deterministic policy, including one selected from the entire prefix being reused. It would be invalid to assert the same martingale property directly for a policy selected from future samples. The additional logarithmic term is $\log|\Pi_{\det}| = |\mathcal{S}|\log|\mathcal{A}|$. It is a statistical dimension cost, not a claim that the algorithm enumerates or evaluates every policy. This argument does not cover an unrestricted class of stochastic policies; deterministic greedy targets are sufficient for the headline algorithm.

Remark B.16 (A primitive sufficient mixing condition). If a one-step Doeblin condition $K_b(x,\cdot) \ge \alpha\nu(\cdot)$ holds for $\alpha \in (0,1]$, a common refresh couples two chains with probability at least α per step. Hence the geometric bound holds with $C_{\mathrm{mix}} = 1$, $r_{\mathrm{mix}} = 1-\alpha$, and $H = 1/\alpha$, together with the stated stationary coverage assumption. The reused-prefix theorem only requires the more general quantitative geometric-mixing condition.

B.6 GREEDIFICATION AND ENVELOPE REFINEMENT

The following standard API propagation argument specializes approximate greedification to a uniform action-value error; see Bertsekas (2011); Scherrer (2014). We retain a self-contained proof with the normalization used here.

Lemma B.17 (Approximate greedification)**.** *If $\left\|\widehat{Q} - Q^\pi\right\|_\infty \le E$ and π' is greedy in $\widehat{Q}$, then*

$$\left\|V^\star - V^{\pi'}\right\|_\infty \le \gamma \left\|V^\star - V^\pi\right\|_\infty + \frac{2E}{1-\gamma}. \tag{81}$$

Proof. Write $T^\pi V = r^\pi + \gamma P_S^\pi V$ and $TV = \max_a(r_a + \gamma P_a V)$. Comparing maximizers of $\widehat{Q}$ and Q^π gives $T^{\pi'}V^\pi \ge TV^\pi - d\mathbf{1}$ with $d = 2E$. Since $TV^\pi \ge V^\pi$, the nonnegative resolvent gives $V^{\pi'} - V^\pi = (I - \gamma P_S^{\pi'})^{-1}(T^{\pi'}V^\pi - V^\pi) \ge -d\mathbf{1}/(1-\gamma)$. Decompose $V^\star - V^{\pi'} = (TV^\star - TV^\pi) + (TV^\pi - T^{\pi'}V^\pi) + \gamma P_S^{\pi'}(V^\pi - V^{\pi'})$. Each term is bounded above by $\gamma D\mathbf{1}$, $d\mathbf{1}$, and $\gamma d\mathbf{1}/(1-\gamma)$, respectively, where $D = \left\|V^\star - V^\pi\right\|_\infty$. Nonnegativity of the total gap gives the norm bound. □

B.6.1 ORDERING CERTIFICATES ON A COMMON TRACE

Lemma B.18 (Ordered radii and identical critics)**.** *For two nonempty certificate recursions with the same initial set and certificate, identical critics and selected actions, and $0 \le E_j^1 \le E_j^2$ for every round, one has $\mathcal{F}_j^1 \subseteq \mathcal{F}_j^2$ and $C_{j+1}^1 \le C_{j+1}^2$.*

Proof. Assume the ordering through the preceding round. Then $a_j^1 = E_j^1 + \gamma C_j^1 \le E_j^2 + \gamma C_j^2 = a_j^2$. The newest balls have the same center; intersection preserves inclusion. Consequently $U_j^1 \le U_j^2$ and $L_j^{1,\mathrm{env}} \ge L_j^{2,\mathrm{env}}$. For the same selected action this orders the nonnegative envelope gaps and hence $Z_j^1 \le Z_j^2$. The API candidates are also ordered. Taking their minima with $2B$ closes the induction from the equal initial values. In Corollary A.14, the confidence event ensures both intersections remain nonempty. □

B.6.2 WHAT HISTORY CONTRIBUTES TO THE ACTUAL CERTIFICATE

For the newest confidence ball alone, let

$$\mathcal{G}_j = [-B, B]^n \cap \{Q : \|Q - \widehat{Q}_j\|_\infty \le a_j\},$$

and write $U_j^{\mathrm{new}}, L_j^{\mathrm{new}}$ for its coordinatewise envelopes. Let $\mathcal{F}_j \subseteq \mathcal{G}_j$ be the nonempty history intersection, with envelopes U_j, L_j. For the fixed actual greedy action $\widehat{a}_j(s)$, define

$$Z_j^{\mathrm{new}} = \frac{1}{1-\gamma} \max_s \{\max_a U_j^{\mathrm{new}}(s,a) - L_j^{\mathrm{new}}(s, \widehat{a}_j(s))\}.$$

Proposition B.19 (History cannot worsen the envelope certificate for a fixed readout)**.** *For every nonempty intersection and the same chosen actions,*

$$Z_j \le Z_j^{\mathrm{new}} \le \frac{2a_j}{1-\gamma}.$$

The first inequality can be strict even when every historical radius is at least the newest radius. The history certificate can also be smaller than the approximate-policy-iteration candidate.

Proof. Set inclusion implies $U_j \le U_j^{\mathrm{new}}$ and $L_j \ge L_j^{\mathrm{new}}$. Substitute these coordinatewise inequalities in the definition of Z_j. The newest-ball bound follows from $U_j^{\mathrm{new}} \le \widehat{Q}_j + a_j$, $L_j^{\mathrm{new}} \ge \widehat{Q}_j - a_j$, and greediness in $\widehat{Q}_j$. The strict possibilities are witnessed by the deterministic construction below. □

A deterministic certificate-geometry example. Take one state with two actions, $B = 1$, $\gamma = 1/2$, and a zero-reward self-transition MDP, so $Q^\pi = Q^\star = (0, 0)$ for every policy. Consider an available historical ball centered at $(-1/5, -1/5)$ with radius $3/10$, and a newest ball centered at $(1/5, 1/5)$ with radius $1/4$. Both centers are valid critics with sup-norm evaluation error $E = 1/5$. The historical radius is $E+\gamma(1/5)$ and the newest radius is $E+\gamma C$ with $C = 1/10$; both supplied policy certificates are valid because the true policy gaps are zero. The history intersection is $[-1/20, 1/10]^2$, while the newest ball alone is $[-1/20, 9/20]^2$. Whichever action the fixed tie rule selects,

$$Z_{\text{hist}} = \frac{3}{10}, \qquad Z_{\text{new}} = 1, \qquad \gamma C + \frac{2E}{1-\gamma} = \frac{17}{20}.$$

Thus history strictly improves the newest-ball certificate and is strictly tighter than the API candidate, although the newest radius is smaller. This is a realizable example of the deterministic containment and evaluation premises. It does not assert that these supplied certificates and critics occur at consecutive rounds of the specific learned-Transformer algorithm analyzed here. Such an algorithm-level witness would require a separately verified data and training construction or experiment.

Worst-case and realized certificates. The generic upper bound is weaker than the API candidate, since

$$\frac{2a_j}{1-\gamma} - \left(\gamma C_j + \frac{2E_j}{1-\gamma}\right) = \frac{\gamma(1+\gamma)}{1-\gamma} C_j \geq 0;$$

this does not imply that the actual Z_j is always larger than that candidate. Likewise, decreasing radii do not make shifted historical balls redundant. History provides deterministic refinement of the certificate, while API provides the present worst-case rate. Intersection does not establish an improved worst-case sample-complexity exponent. Fixed E_j alone does not establish monotonicity of the implemented C_j recursion: a small envelope candidate can put it below the API error floor, and the slack depends on the next chosen action. Only W_j, for the nested nonempty sets, has the presently proved monotonicity statement.

A check on the implemented certificate recurrence. Nonmonotonicity can also be demonstrated within the deterministic evaluation premises, using the recurrence exactly as written. Take $B = 1$, $\gamma = 1/10$, $E_j = 1/5$ for all j, the same zero-reward one-state two-action MDP, $C_0 = 2$, and the full initial box. Use critics $\widehat{Q}_0 = (-1/5, -1/5)$, $\widehat{Q}_1 = (-1/5, 1/5)$, and $\widehat{Q}_2 = (-1/5, -1/5)$, breaking ties in favor of action 1. Every critic has the required error E_j and every true policy gap is zero. Direct endpoint updates give

j	a_j	Z_j	$\gamma C_j + 2E_j/(1-\gamma)$	C_{j+1}
0	$2/5$	$8/9$	$29/45$	$29/45$
1	$119/450$	$119/405$	$229/450$	$119/405$
2	$929/4050$	$1858/3645$	$1919/4050$	$1919/4050$

In particular, $C_3 > C_2$. This calculation refutes monotonicity as a consequence of fixed E_j and the stated deterministic certificate recurrence alone. As above, it does not assert that a single trained Transformer and its sampled context generate precisely this critic sequence.

Corollary B.20 (Correct calibration and critic-distance consequences)**.** *For any benchmark stationary policy $\pi^\dagger_{j+1}$, the calibration residual satisfies*

$$\eta^{\text{cal}}_{j+1} := \sup_s [\Delta_{j+1}(s) - (V^\star(s) - \mathbb{E}_{\pi^\dagger_{j+1}} Q^\star(s, \cdot))]_+ \leq \min\{2a_j, C_{j+1}\}.$$

For any nonempty feasible class containing $Q^\star$, the distance of $\widehat{Q}_j$ to that class is at most $E_j + \gamma C_j$.

Proof. The benchmark gap is nonnegative, so calibration is at most $\sup_s \Delta_{j+1}(s) \leq 2a_j$. Monotonicity gives $T^{\pi_{j+1}} V^\star \geq T^{\pi_{j+1}} V^{\pi_{j+1}} = V^{\pi_{j+1}}$, hence $\Delta_{j+1} \leq V^\star - V^{\pi_{j+1}} \leq C_{j+1}$. Use $Q^\star$ as a candidate in the distance infimum for the last assertion. □

B.7 Testing and the scope of necessary resources

The deployment testing theorem grants task-uninformative side information and counts unique observations. The following proof uses exactly that information model.

Proof of Theorem A.21. Let $\Delta = \epsilon(1-\gamma)/R_{\max} \le 1/4$. Both actions return to the only state. In task M_+, independently at each visit the rewards divided by $R_{\max}$ are Bernoulli with means $1/2+\Delta$ at action 1 and $1/2-\Delta$ at action 2. In M_-, interchange these two means. Rewards lie in $[0, R_{\max}]$. Choose each behavior action independently with probability $1/2$, including the first action. The state-action kernel has every row equal to $(1/2, 1/2)$, so its stationary mass is $1/2$ and its total-variation distance to stationarity is zero after one step. For example $C_{\mathrm{mix}} = 1, r_{\mathrm{mix}} = 1/2$ satisfy the mixing assumption (73), including time zero.

The value gap of the deterministic policy choosing the worse action is $2R_{\max}\Delta/(1-\gamma) = 2\epsilon$. Consequently an ϵ-optimal output must select the better action. Let $\mathbb{P}_+$ and $\mathbb{P}_-$ be the laws of all observations and side information. Independent algorithmic randomness can be adjoined to these laws. Side information has the same law under both tasks and conveys no task information; conditional on it, the behavior observations have the specified identical-in-time product law. By the chain rule for relative entropy, or by multiplying the likelihoods, their divergence is

$$\mathrm{KL}(\mathbb{P}_+ \| \mathbb{P}_-) = m d_\Delta, \qquad d_\Delta = 2\Delta \log \frac{1+2\Delta}{1-2\Delta}.$$

Indeed, the two action-conditioned Bernoulli divergences are equal, and the action distribution is identical under both tasks. Moreover,

$$\log \frac{1+2\Delta}{1-2\Delta} = \log\left(1 + \frac{4\Delta}{1-2\Delta}\right) \le \frac{4\Delta}{1-2\Delta} \le 8\Delta,$$

so $d_\Delta \le 16\Delta^2$.

Let ϕ be the indicator that the algorithm returns action 1. The assumed guarantee implies $\mathbb{P}_+(\phi = 1) \ge 1-\delta$ and $\mathbb{P}_-(\phi = 1) \le \delta$. Partitioning the likelihood integral according to ϕ and applying the log-sum inequality gives

$$\mathrm{KL}(\mathbb{P}_+ \| \mathbb{P}_-) \ge \mathrm{kl}\big(\mathbb{P}_+(\phi=1), \mathbb{P}_-(\phi=1)\big) \ge \mathrm{kl}(1-\delta, \delta).$$

For $a > b$, binary relative entropy is increasing in a and decreasing in b, which proves the last inequality. Finally, $\mathrm{kl}(1-\delta,\delta) = (1-2\delta)\log((1-\delta)/\delta)$. Combining this with $m d_\Delta \le 16 m \Delta^2$ and substituting Δ proves (45). □

B.7.1 An exact evaluation-depth versus logit tradeoff

Proposition B.21 (Covered mixing witness with an exact clipped orbit)**.** *Fix an integer $k \ge 2$, let $n = 2k$, and take any teacher score gap $\tau > 0$. There is a k-state, two-action MDP satisfying the coverage and mixing hypotheses* (73) *with*

$$\mu_0 = 1/n, \qquad C_{\mathrm{mix}} = 2, \qquad r_{\mathrm{mix}} = 1/2, \qquad H = 4,$$

and a deterministic target policy such that its population teacher evaluation from $Q_0 = 0$, including clipping, has exactly

$$\|Q_L - Q^\pi\|_\infty = B q^L, \quad q = 1 - \beta(1-\gamma), \quad \beta = \frac{e^\tau - 1}{e^\tau + n - 1}. \tag{82}$$

In particular, for $0 < \epsilon < B$ its smallest sufficient and necessary integer depth is

$$L_{\min} = \left\lceil \frac{\log(B/\epsilon)}{-\log[1-\beta(1-\gamma)]} \right\rceil. \tag{83}$$

Proof. At each state, action Stay preserves the state and action Reset draws the next state uniformly from all k states. Behavior chooses each action with probability $1/2$; the target policy always chooses Stay. The behavior state kernel is $P_b = (I+U)/2$, where every row of U is uniform. Its stationary state-action distribution is uniform. Starting from any state-action pair, after the first transition let the state law be ν. For $t \ge 1$, the time-t state-action law is νP_b^{t-1} times uniform action probabilities. Since $P_b^h = U + 2^{-h}(I-U)$,

$$\sup_x \|K_b^t(x,\cdot) - \mu_b\|_{\mathrm{TV}} \le 2^{-(t-1)} = 2(1/2)^t.$$

At $t = 0$ the bound also holds, proving the asserted constants.

Choose $f \in [-1, 1]^k$ with zero sum and $\|f\|_\infty = 1$; for example $f = (1, -1, 0, \dots, 0)$. Define the address vector and deterministic rewards by

$$v(s, \text{Stay}) = f(s), \qquad v(s, \text{Reset}) = 0, \qquad r = R_{\max} v.$$

If $T = P_X^\pi$ is the target state-action transition matrix, then $Tv = v$: the Stay rows preserve f and the Reset rows average it to zero. Also $\mathbf{1}^\top v = 0$, and hence the Bellman solution is $Q^\pi = Bv$.

Uniform population occupation and the teacher diagonal score gap give

$$W = \beta I + (1 - \beta)\Pi, \qquad \Pi = \mathbf{1}\mathbf{1}^\top / n, \qquad Wv = \beta v.$$

The actual teacher map is $F(Q) = Q + W(r + \gamma TQ - Q)$; in particular its linear part is $I - W + \gamma WT$, with the W retained in the successor term. For any $0 \le a \le 1$, direct substitution yields

$$F(aBv) = \big[a + \beta(1 - \gamma)(1 - a)\big] Bv.$$

Because $0 < \beta < 1$, the displayed new coefficient remains in $[0, 1]$. Clipping is therefore inactive along this entire orbit. Induction from zero gives $Q_L = (1 - q^L)Bv$, proving (82). Taking logarithms gives (83). □

Scope of the depth witness. The exact orbit gives a depth requirement for this teacher on a covered population instance. For every positive score gap it converges as depth increases, so it does not imply a positive error floor. Its depth/logit tradeoff is distinct from strict one-step supremum-norm contraction uniformly over tasks. The experiment in Appendix C.1.2 instantiates this witness with a balanced record representing population frequencies.

A supplied teacher and training sample complexity. The prescribed teacher parameters can be installed without inspecting any training examples. Our training theorem therefore gives finite optimization and estimation guarantees for the stated data-based procedure, rather than an information-theoretic lower bound for learning supplied parameters. The deployment lower bound remains meaningful because the exact teacher does not reveal the new task's rewards or transition law.

B.7.2 The scope of the score-margin requirement

Proposition B.22 (Necessary score separation for this teacher's uniform contraction)**.** *Fix $n \ge 2$, $\gamma \in (0, 1)$, and $0 < \mu_0 \le 1/n$. Consider the population teacher update*

$$F^\pi(Q) = Q + W_\tau(\mu)\{r + \gamma P_X^\pi Q - Q\}, \qquad W_\tau(\mu)_{xu} = \frac{e^{\tau \mathbf{1}\{x=u\}} \mu(u)}{\sum_v e^{\tau \mathbf{1}\{x=v\}} \mu(v)}.$$

If this map is required to be a strict sup-norm contraction for every covered finite MDP and every deterministic target policy, then necessarily

$$e^\tau > \frac{1 - \mu_0}{\mu_0} \frac{1 + \gamma}{1 - \gamma}. \tag{84}$$

The same necessity holds if the requirement is imposed on the clipped map $\Pi_B F^\pi$ on $[-B, B]^n$, for any $B > 0$. Consequently, this particular uniform-contraction requirement forces $e^\tau = \Omega(n/(1-\gamma))$ as n and the discount vary.

Proof. Take an MDP with one state, n actions, a deterministic self-transition, and identically zero rewards. Choose a stationary behavior distribution μ with $\mu(x) = \mu_0$ for one action x, and $\mu(u) = (1 - \mu_0)/(n - 1)$ for all $u \ne x$. All actions have stationary mass at least μ_0. The behavior action sequence is independent across times and therefore satisfies the covered mixing condition. Let the deterministic target always choose action x. Then $P_X^\pi = \mathbf{1} e_x^\top$, and stochasticity of $W := W_\tau(\mu)$ gives $W P_X^\pi = \mathbf{1} e_x^\top$.

The raw map is linear with matrix $M = I - W + \gamma \mathbf{1} e_x^\top$. Its row x has diagonal entry $1 - W_{xx} + \gamma > 0$ and off-diagonal entries $-W_{xu} < 0$, so

$$\sum_u |M_{xu}| = 2(1 - W_{xx}) + \gamma, \qquad W_{xx} = \left[1 + \frac{1 - \mu_0}{\mu_0} e^{-\tau}\right]^{-1}.$$

Since the induced sup norm is the maximum absolute row sum, strict contraction requires $2(1 - W_{xx}) + \gamma < 1$, equivalently $W_{xx} > (1 + \gamma)/2$. Solving this inequality gives (84). For clipping, $F^\pi(0) = 0$ and there is a neighborhood of zero on which clipping is inactive. The derivative of the clipped map there is still M, so a global Lipschitz constant below one would again require $\|M\|_\infty < 1$. Finally, $(1 - \mu_0)/\mu_0 \geq n - 1 \geq n/2$. □

Scope of the necessity statement. This proposition concerns strict sup-norm contraction of the specified softmax-weighted TD teacher, uniformly over covered tasks and target policies. It is not a lower bound for arbitrary softmax architectures, other norms or evaluation algorithms, convergence on each individual task, or necessary sample complexity. A failed contraction test alone does not prove that the iteration diverges. The sufficient budget above must not be converted into a statistical lower bound by substituting (84).

C EXPERIMENTAL DETAILS AND REPRODUCIBILITY

This section documents the experiments connecting operator training to repeated evaluation and frozen policy improvement. Appendix C.1 compares early and final checkpoints on common target policies, varies evaluation depth and unique deployment observations, and examines noisy training, two-task testing, and size-specific models on larger address spaces.

Appendix C.2 gives the probe-curriculum protocol, operator-recovery results, and broader frozen-transfer study. Appendix C.3 evaluates the prescribed-attention special case of Appendix A.1.3 under a separate protocol. The code was ran on a 12700k processor with an RTX 5090 and 32GB RAM.

C.1 POLICY IMPROVEMENT, RESOURCE SCALING, AND TESTING WITNESSES

The following studies vary the resources that enter the returned-policy bound: evaluation depth, policy-improvement rounds, deployment observations, and noisy training examples. T1 and T2 provide the main-text control and context curves. T3 separates statistical training error from incomplete optimization; T4 examines the two-task deployment difficulty; T5 checks larger address spaces. The original curriculum and broader transfer study are detailed in Appendix C.2.

C.1.1 SHARED SETUP, IMPLEMENTATION, AND REPORTING

Unless specified otherwise, tasks have four states, actions $\{-1, +1\}$, $\gamma = 0.7$, and $R_{\max} = 1$. Thus $n = 8$, $B = 10/3$, and the structured token evaluator has width $2n + 4 = 20$. Its trainable parameters are the 56 free attention scores and three TD coefficients, with the last score column fixed to zero. The score box is $[-6, 6]$, the coefficient box is $[-2, 2]^3$, and the teacher uses $\tau = 5$ and $c^\star = (1, -1, 0.7)$. T1 and T2 freeze the five clean, uniform-start curriculum models trained with $N = 1024$ in Appendix C.2: one coefficient update followed by 262,144 attention updates. The early checkpoint has 64 attention updates after coefficient fitting. The five independent training-bank seeds are 19001–19005.

Fresh tasks and paired observations. The deployment bank contains 50 ring tasks and 50 random-kernel tasks. For each task, the reward entries are independent $\mathrm{Unif}[-1, 1]$ draws and subsequently deterministic given the state and action. The kernels are

$$P_{\mathrm{ring}}(s' \mid s, a) = 0.8\mathbf{1}\{s' = (s + a) \bmod 4\} + 0.2/4,$$
$$P_{\mathrm{random}}(s' \mid s, a) = 0.6\widetilde{P}(s' \mid s, a) + 0.4/4,$$

where each row of $\widetilde{P}$ is an independent $\mathrm{Dirichlet}(1, 1, 1, 1)$ draw. Each task receives three independent behavior records, starting uniformly and choosing fair independent actions. Each record has 16,384 transitions, and smaller contexts use nested prefixes. All methods receive the same record. Consequently, a learned-model mean averages 100 tasks, three records per task, and five training banks; the resulting 1,500 cells are paired observations, not 1,500 independent tasks. The teacher and empirical planner have 300 task–record cells.

At every improvement round, the evaluator resets its memory to zero, re-encodes each recorded successor using the current target policy, executes L clipped updates, and greedifies the resulting critic. The initial policy is $\pi_0(s) = -1$; computed greedy ties select -1. The empirical-model comparator estimates rewards and transition frequencies from the same record and performs exact policy evaluation in that fitted model. Unvisited pairs receive reward zero and uniform successors. Neither comparator receives the true task model for choosing actions. We use true-model Bellman solves and value iteration, with stopping tolerance 10^{-13} for value iteration, only to measure the returned policies.

Exact cached evaluation. For fixed scores, coefficients, context, and target policy, the raw token update is affine in the memory. Writing $x_t = (s_t, a_t)$, $y_t = (s_{t+1}, \pi(s_{t+1}))$, and $w_t^A(x)$ for its

normalized attention, the campaign computes

$$h_\theta(x) = c_r \sum_{t<m} w_t^A(x) R_{t+1},$$
$$M_\theta^\pi(x,z) = \mathbf{1}\{x=z\} + c_p \sum_{t<m} w_t^A(x)\mathbf{1}\{x_t=z\} + c_s \sum_{t<m} w_t^A(x)\mathbf{1}\{y_t=z\},$$
$$Q_{\ell+1} = \Pi_B\big(h_\theta + M_\theta^\pi Q_\ell\big).$$

Caching these coefficients changes the arithmetic organization of the declared structured evaluator. A numerical check compared all eight iterates against literal token execution on one 37-transition input and one target policy for three models: the teacher, the final checkpoint from bank 19001, and the early checkpoint from bank 19003. The largest absolute discrepancy was 3.77×10^{-15}, below the 10^{-10} check tolerance. The checked cases accompany the algebraic identity above; the check was not an exhaustive test over every task or width.

Metrics and resource units. The measured policy gap, evaluation error, and teacher residual are

$$D_j = \|V^\star - V^{\pi_j}\|_\infty, \qquad E_j^{\mathrm{act}} = \|\widehat{Q}_j - Q^{\pi_j}\|_\infty, \tag{85}$$
$$r_m(\pi) = \|G_{\theta^\star}(D_m, Q^\pi) - Q^\pi\|_\infty, \qquad r_m^{\max} = \max_{\pi\in\{-1,+1\}^4} r_m(\pi). \tag{86}$$

The maximum in (86) enumerates all 16 deterministic policies, including policies selected after seeing the record. It is an observable-in-simulation diagnostic of the policy-uniform residual in the analysis, not a statistic available without the true model at deployment. Here N counts independent operator-training examples, m counts unique deployment transitions, L counts evaluator loops per round, and J counts policy improvements. Reusing a prefix for all JL loops still costs m observations. The T4 and T5 models instead use exact clean probe targets, as specified below.

Curves report descriptive means, except for the explicitly labeled RMS residual. T3 error bars are sample standard deviations across five independent training banks. The original campaign records aggregate means and descriptive standard deviations for the other studies; shared tasks, nested prefixes, and repeated use of a training bank are not treated as independent replications for confidence intervals. Fits are ordinary least-squares lines on the log of the plotted grid-level statistics. Rate guides have the indicated exponent and are normalized to a plotted endpoint. At the tested budgets, the sufficient certificate bounds remain uninformative; the experiments assess policy performance and resource-scaling trends.

C.1.2 T1: Training accuracy, evaluation depth, and policy improvement

This study fixes $m = 1024$, runs $J = 16$ improvement rounds, and compares $L \in \{1, 8, 64\}$ for the teacher, final learned models, and early checkpoints. It also runs the matched-record empirical planner. The main-text panel emphasizes the first few rounds, while Figure 3 retains the full horizon. For the final learned model, the mean gap changes from $D_0 = 1.48908$ to $D_{16} = 0.129255$ at $L = 1$, and to 0.00147309 at both $L = 8$ and $L = 64$. The latter final gap matches the teacher and the empirical planner on this bank. Additional evaluation depth changes the quality of greedy improvement when evaluation is too short; after the selected actions stabilize, greater accuracy need not change the policy.

Training accuracy on common target policies. To distinguish evaluation depth from policy changes, each final learned model generates an $L = 64$ trace on each record. Its policies at rounds $0, 1, 2$ serve as common targets for that model, its early checkpoint, and the teacher. All three evaluators then start from zero memory on the same record and target, using $L \in \{1, 2, 4, 8, 16, 32, 64, 128\}$. Each plotted method therefore averages 4,500 target cells, with the same pairing described above. The final learned and teacher curves nearly coincide: the mean evaluation error decreases from 0.89939 at $L = 1$ to 0.04780 at $L = 128$. The early model's mean error instead grows from 1.69666 to 4.32757. Records, target policies, and zero initial memories are shared across checkpoints. The contrasting mean curves therefore expose the effect of training accuracy on repeated evaluation: the final checkpoint benefits from additional depth, while the early checkpoint's error increases.

The campaign also records the unrolled API comparison $\gamma^j D_0 + \frac{2}{1-\gamma}\sum_{i<j}\gamma^{j-1-i}E_i^{\mathrm{act}}$. Its inputs are measured true-model errors. The stored depth envelope uses the evaluator-specific matrix norm

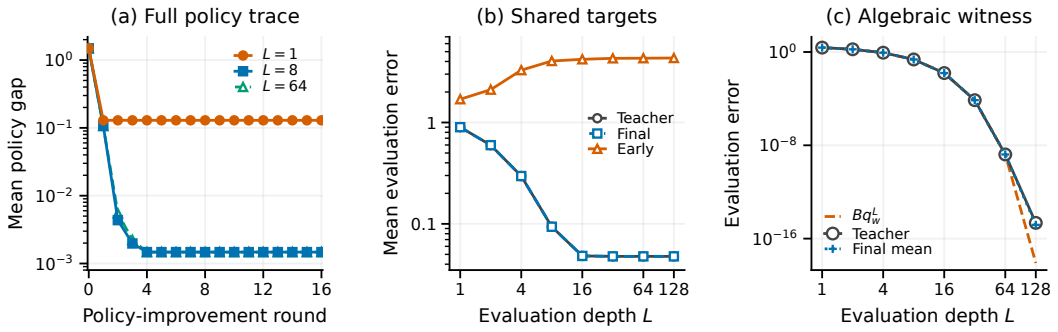

Figure 3: T1 depth diagnostics. Left: the full 16-round policy-gap traces at three evaluator depths. Center: evaluation errors on common fixed targets from each final model's $L = 64$ trace; the teacher and final checkpoint overlap. Right: the balanced population witness and its exact teacher depth formula. The first two panels show descriptive means over the paired task bank; the right panel uses the deterministic 32-entry input.

$\|M_\theta^\pi\|_\infty$ and its true-model residual, with an additional nonnegative $L_G\|\theta - \theta^\star\|_2$ term, and clips the result at $2B$. These are retrospective diagnostics, distinct from the theorem's teacher-based empirical contraction refinement and high-probability radii.

A population depth witness. The four-state Stay/Reset construction in Proposition B.21 removes context sampling error. Stay preserves the state, Reset selects a uniform next state, and the target policy always chooses Stay. Let $f = (1, -1, 0, 0)$, set the Stay rewards to f, and set the Reset rewards to zero. Then $Q^\pi = Bv$, where $v(s, \text{Stay}) = f(s)$ and $v(s, \text{Reset}) = 0$. A balanced 32-entry input has four Stay entries per state and one Reset entry for each possible successor. It represents the exact population frequencies, rather than a sampled behavior trajectory. For the teacher, the zero-memory error is exactly

$$\|Q_L - Q^\pi\|_\infty = Bq_{\mathrm{w}}^L, \qquad q_{\mathrm{w}} = 1 - \frac{e^5 - 1}{e^5 + 7}(1 - \gamma).$$

Across the eight tested depths, the largest discrepancy from this expression is 2.49×10^{-15}. The trained models closely track the teacher; at the largest depth, both reach numerical precision. This isolates the geometric depth mechanism without requiring observed policy gaps to follow an exact geometric sequence.

C.1.3 T2: Unique deployment observations

T2 fixes $J = 8$ and $L = 128$ and varies $m \in \{64, 128, 256, 512, 1024, 2048, 4096, 8192, 16384\}$ on the same 100 tasks and three records per task. Each smaller context is a prefix of the same maximum-length record. The teacher, final learned models, and empirical planner have matching aggregate final-gap curves. Their mean gap decreases from 0.0167907 at $m = 64$ to 7.04324×10^{-5} at $m = 16384$. This comparison changes the number of distinct observations while keeping the model frozen and computational budgets fixed.

For each of the 300 task–record pairs, we additionally compute $r_m^{\max}$ from (86). The main-text panel plots $[\frac{1}{300}\sum_i (r_{m,i}^{\max})^2]^{1/2}$, which decreases from 0.30381 to 0.01809 and has fitted log–log slope -0.510. The slope -0.509 reported in the original campaign summary instead fits the arithmetic mean of $r_m^{\max}$; the two statistics are kept distinct. The RMS curve is close to the $m^{-1/2}$ guide, supporting the sampling-error dependence that enters finite evaluation. Policy gaps are less suitable for estimating that dependence: a finite action gap can make a greedy policy optimal even when its critic still has nonzero residual. The measured trends illustrate the sampling and control mechanisms, without calibrating the theorem's worst-case sufficient constants.

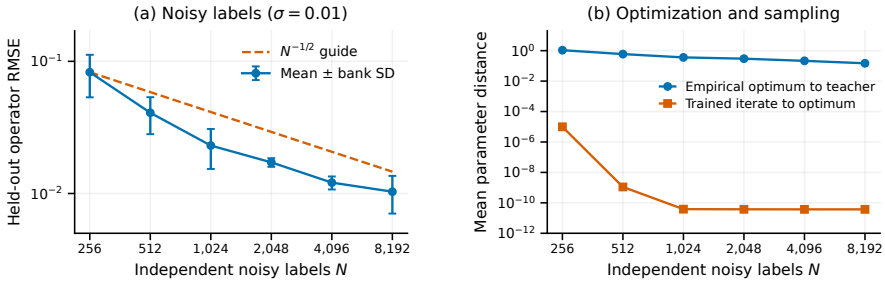

Figure 4: T3 noisy-label training ($\sigma = 0.01$, 1,048,576 attention updates). Left: operator RMSE on the common clean test bank, with one sample SD across five training banks and an $N^{-1/2}$ guide. Right: mean attention distance from the empirical minimizer to the teacher, and from the trained iterate to that minimizer, separating statistical and optimization errors.

C.1.4 T3: Noisy training and optimization error

This study examines noisy-training sample dependence while separately checking optimization error. It extends the sample-size sweep in Appendix C.2.4 to $N \in \{256, 512, 1024, 2048, 4096, 8192\}$ at fixed $\sigma = 0.01$, with five training banks and nested prefixes within each bank. It uses the same probe mixture, independent bounded Rademacher label noise, parameter boxes, and uniform initialization rule as Appendix C.2. The coefficient stage is one projected update, and attention receives 1,048,576 projected updates of step size one. Existing checkpoints at that budget are reused where available; the remaining cells use the identical grouped update. The clean held-out operator bank contains the same 256 inputs for every cell.

Separating optimization and sampling error. For attention-group means $\bar{Y}_{U_u}(x)$ and fitted coefficient $\bar{c}_r$, the constrained empirical minimizer can be computed coordinatewise as

$$A^{\mathrm{emp}}_{xu} = \mathrm{logit}\left(\mathrm{clip}_{[\ell(-6),\ell(6)]}\left[\frac{\bar{Y}_{U_u}(x)}{\bar{c}_r}\right]\right), \qquad \ell(a) = (1 + e^{-a})^{-1}.$$

This solution is used only to diagnose optimization, while the deployed parameters are produced by projected gradient descent. The recorded criterion is $\|\widehat{A} - A^{\mathrm{emp}}\|_F \le 0.1\|A^{\mathrm{emp}} - A^{\star}\|_F$. All 30 cells satisfy it. At $N = 256$, the mean optimization residual is 9.92×10^{-6} compared with empirical statistical distance 1.07449; at $N = 8192$, the corresponding values are 3.73×10^{-11} and 0.14836.

Operator error versus sample size. Mean clean-operator RMSE decreases from 0.08272 to 0.01032, with fitted log–log slope -0.592 across the six grid means (Figure 4). The deterministic all-input bound expression $L_G\|\widehat{\theta} - \theta^{\star}\|_2$ also decreases, from 38.13 to 5.265, but is much larger than the measured RMSE. The latter averages errors on a finite held-out bank and is not an estimate of a uniform supremum. Together with the small optimization residual, the trend supports the statistical training term under positive label noise.

C.1.5 T4: Two-task deployment testing witness

The testing task has one state and two actions. For task sign $z \in \{-1, +1\}$, rewards are Bernoulli with means $\mathbb{E}[R \mid a] = 1/2 + za\Delta$, where $\Delta \in \{1/8, 1/16, 1/32, 1/64\}$. Behavior actions are independent fair draws. We generate 1,000 records per sign and gap, with nested prefixes $m \in \{8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096, 8192, 16384\}$. The same records are supplied to all four methods. Returning the wrong action incurs value gap $2\Delta/(1-\gamma) = 2\epsilon_\Delta$, where $\epsilon_\Delta = \Delta/(1-\gamma)$.

A separate width-8 evaluator, with $n = 2$, $\tau = 5$, and score bound 6, is trained once independently of the task sign. Its coefficients are fixed to the exact teacher values, and 262,144 projected attention updates fit exact clean paired-probe targets from a uniform random start. Thus this study isolates deployment information requirements; it does not use a sampled finite-N pretraining bank. Learned

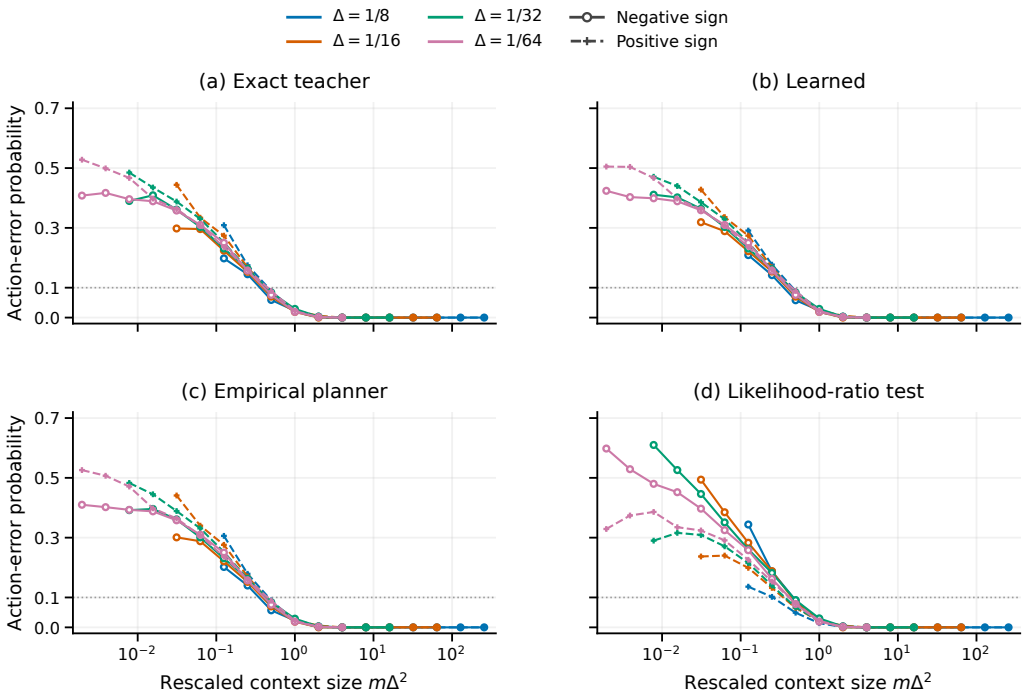

Figure 5: T4 testing witness. Each panel shows one method's observed error probability for all four reward gaps and both task signs. The horizontal axis is logarithmic in the rescaled context size $m\Delta^2$. Each point uses 1,000 records; prefixes and methods share records within a sign and gap. Curves show point estimates; uncertainty and grid-crossing interpretation are described in the text.

and exact-teacher evaluation use $L = 128$ and $J = 8$. The empirical planner chooses the larger per-action sample mean, assigning zero to an unobserved action. The likelihood-ratio reference uses $T = \sum_{t<m} a_t(2R_{t+1} - 1)$ and selects $+1$ when $T \geq 0$. Its zero-statistic tie selects $+1$; greedy evaluator and empirical-mean ties select -1. Both task signs are shown separately.

For every method and each sign, the first tested context size with observed error probability below 0.1 is as follows:

Reward half-gap Δ	$1/8$	$1/16$	$1/32$	$1/64$
First tested m below 10% error	32	128	512	2048

The error curves align approximately when displayed against $m\Delta^2$ (Figure 5). The fourfold increase in grid crossings as Δ halves is consistent with inverse-square deployment difficulty, including when the exact teacher is supplied. These are empirical crossings on a powers-of-two grid. The campaign also computed pointwise normal-binomial intervals based on 1,000 records; neither those intervals nor the crossing table constitutes a simultaneous guarantee or an experimental proof of the information lower bound.

C.1.6 T5: Size-specific models on larger address spaces

T5 repeats the ring construction with $S \in \{4, 8, 16\}$ states and two actions, giving $n = 2S \in \{8, 16, 32\}$ addresses and widths $20, 36, 68$. For each size, it draws 50 fresh reward tasks and one fair-behavior record per task. Context sizes are $m = 1024, 2048, 4096$, respectively; all deployments use $L = 128$ and $J = 8$. The score scale and box are

$$\tau_n = 5 + \log \frac{n-1}{7}, \qquad M_{A,n} = \tau_n + 1.$$

Table 3: T5 size-specific attention training with exact clean targets. Gaps average 50 tasks and five initialization seeds. The teacher has the same final gap at each size. Agreement is the fraction of evaluated states on which the learned final policy matches the teacher.

States	Width	Context m	Mean final gap	Agreement	Mean $\|A - A^\star\|_F$
4	20	1024	0.0002271	1.000	1.758×10^{-5}
8	36	2048	0.0011081	1.000	0.20714
16	68	4096	0.0087599	1.000	2.44582

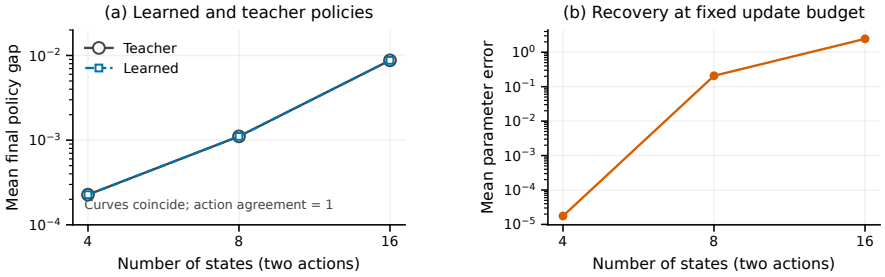

Figure 6: T5 state-space scaling. Left: mean final policy gaps for separately trained evaluators and the teacher; the curves coincide. Right: attention parameter error after the common optimization budget. Control means use 50 tasks per size; learned results average five attention initialization seeds on the same tasks and records.

For the ring's uniform state–action occupation, this keeps the teacher's population diagonal attention mass constant as n changes.

Each size has five independent uniform attention initializations. The coefficients are fixed at their teacher values, and attention is trained for 262,144 projected steps against exact clean probe targets. These are separate size-specific models; they do not transfer a fixed-width network between state spaces. The runner's metadata also records $N = m$, but no independent sampled training examples are generated in this study, so these numbers are reported as deployment budgets only.

The learned policies agree with the teacher on every evaluated state at all three sizes (Table 3). Parameter recovery is less complete at larger widths under the common optimization budget, while final greedy decisions still agree. The larger teacher and learned policy gaps show that preserving the population attention mass does not remove the other effects of dimension and finite context. This supports operation of the construction at modestly larger address counts, without attributing the trend to a fixed-dimension finite-sample training rate.

C.2 Probe curriculum, frozen transfer, and finite-sample diagnostics

This study evaluates the two-stage curriculum of (4) and Theorem A.2 from three prespecified attention initializations. Its operator-recovery curves supply the training panel of Figure 1; the complete original diagnostics and transfer results are retained below. Appendix C.1 uses these frozen checkpoints on a separate fresh task bank. The coefficient-only study in Appendix C.3 has a separate task bank and protocol.

C.2.1 Environment, model, and training protocol

Tasks have four states $\{0, 1, 2, 3\}$, two actions $\{-1, +1\}$, $\gamma = 0.7$, $R_{\max} = 1$, $n = 8$, and $B = 10/3$. The training-task kernel is

$$P_{\text{ring}}(s' \mid s, a) = 0.8\mathbf{1}\{s' = (s + a) \bmod 4\} + 0.2/4.$$

Each random task draws the eight reward entries independently from $\text{Unif}[-1, 1]$; rewards are deterministic given (s, a). The continuing behavior record starts uniformly and uses independent fair actions, with no additional resets inside a record. The model implements the width-20 token

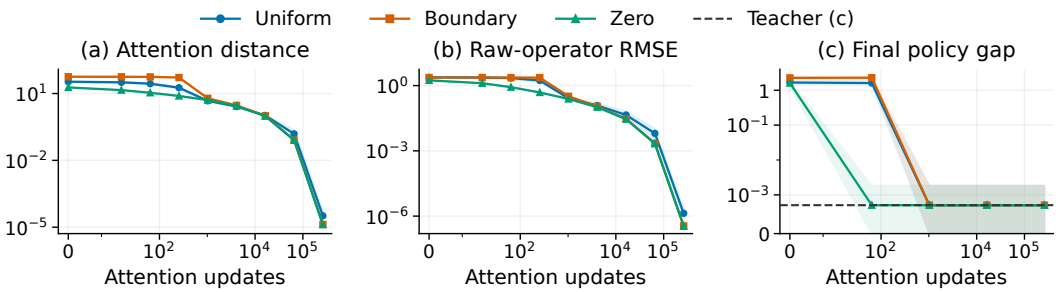

Figure 7: Full original clean-label curriculum ($N = m = 1024$, $L = 64$, $J = 8$). Step zero follows coefficient fitting. Bands show seed SD for attention distance and held-out RMSE (a,b), and saved 95% crossed seed/task intervals on 50 ring tasks (c). The teacher's mean final gap is 7.34×10^{-4}.

construction: eight memory tokens, transition tokens and a dummy token; fixed singleton read/reset heads; one learned softmax aggregation head; and fixed four-ReLU clipping. There are two attention sublayers per macro-loop. Only the 56 free scores and three TD coefficients are trainable; there is no layer normalization, dropout, learned action head, or additional learned projection. We use the fixed gauge $A_{xn} = 0$, score box $[-6, 6]^{56}$, coefficient box $[-2, 2]^3$, and teacher parameters with $\tau = 5$ and $c^\star = (1, -1, 0.7)$.

For each of five independent seed banks (19001–19005), draw i.i.d. operator examples from a mixture with probability $3/4$ on ordinary inputs and probability $1/4$ uniformly distributed among the ten probes in (3). Thus each designated probe has probability $1/40$. Ordinary inputs contain a fresh random reward task, a uniformly sampled deterministic target policy, random memory in $[-B, B]^8$, and a fresh length-16 behavior record. The probe amplitude is $r_0 = 1$; probes are synthetic algebraic inputs, counted separately from physical transitions. Labels are raw one-step teacher outputs before clipping. Primary runs use $N = 1024$ and clean labels. Probe membership is recognized from the input, then group means are computed once. Ordinary nonprobe inputs do not enter the two grouped objectives.

Specifically, using the singleton targets of (11), the implemented objectives are

$$\mathcal{L}_C(c) = \tfrac{1}{2}\|c - \widehat{c}\|_2^2, \qquad \mathcal{L}_A(A; \bar{c}) = \tfrac{1}{2} \sum_{x=1}^{8} \sum_{u=1}^{7} [\bar{c}_r \ell(A_{xu}) - \bar{Y}_{U_u}(x)]^2, \quad \ell(a) = (1 + e^{-a})^{-1}.$$

Both step sizes are one, and every update is projected onto its declared box. Stage C finishes in one update; Stage A freezes $\bar{c}$ and preserves the zero gauge column. The targets come from labels, rather than from direct access to teacher parameters during optimization. If any probe group is missing, the prescribed branch retains the initialization and marks the curriculum certificate unavailable; no resampling is permitted. Across all 55 curriculum runs, this branch is taken zero times. The minimum probe count is 2 at $N = 256$, 14 at $N = 1024$, and 82 at $N = 4096$.

Uniform starts draw all free scores independently in $[-6, 6]$ and coefficients in $[-2, 2]^3$. Boundary starts oppose each nonzero teacher-score sign and alternate $-6, +6$ on its zero-score coordinates, with random coefficients. Zero starts set both scores and coefficients to zero. Boundary and zero attention configurations are identical across seeds; their datasets remain independent. A separate validation pilot (seed 18000) selected 262,144 attention updates before final training. Saved attention checkpoints are $0, 16, 64, 256, 1024, 4096, 16384, 65536, 262144$; a longer reference continues to 1,048,576 updates. Ring deployment is measured at attention checkpoints $0, 64, 1024, 16384, 262144$, as well as the genuinely untrained checkpoint. *Attention step zero is after the coefficient update.* All initialization classes, seeds, and slow intermediate checkpoints are retained.

Table 4: Curriculum protocol. Every row has five seed banks, one coefficient update, primary attention budget 262,144, and reference budget 1,048,576. The noise and sample-size grid uses paired prefixes/signs within banks; its clean uniform $N = 1024$ row is shared with the primary comparison.

Initialization	(N, σ)	Runs	Min. probe count	Missing groups
Boundary	$(1024, 0)$	5	14	0
Zero	$(1024, 0)$	5	14	0
Uniform	$(256, \{0, 0.001, 0.01\})$	15	2	0
Uniform	$(1024, \{0, 0.001, 0.01\})$	15	14	0
Uniform	$(4096, \{0, 0.001, 0.01\})$	15	82	0

Table 5: Clean-label curriculum at the selected budget. Means are over five training seeds. INIT denotes deployment using the untrained evaluator, not the initial all-minus policy; step zero follows coefficient fitting. Every final ring gap has saved 95% interval $[0, 0.00183921]$.

Start	$\|A - A^\star\|_F$	Operator RMSE	INIT ring gap	Final ring gap
Uniform	$3.25314\,10^{-5}$	$1.34619\,10^{-6}$	1.611152	0.000733741
Boundary	$1.33052\,10^{-5}$	$3.53507\,10^{-7}$	1.898877	0.000733741
Zero	$1.30554\,10^{-5}$	$3.46870\,10^{-7}$	1.610674	0.000733741

C.2.2 Operator recovery and frozen control

The common held-out operator bank has 256 clean-labeled inputs, independent of training (seed 19000). For each trained model, raw-operator RMSE is

$$\left[\frac{1}{256n}\sum_{i=1}^{256}\|G_{\widehat{\theta}}(U_i) - G_{\theta^\star}(U_i)\|_2^2\right]^{1/2}.$$

We report the mean of these seedwise RMSEs and sample SD conditional on the fixed bank. This finite-bank measure and its maximum discrepancy are distinct from the theorem's all-input supremum. Because the teacher is explicitly known, the deterministic construction bound $L_G\|\widehat{\theta} - \theta^\star\|_2$ can also be evaluated inside the box, where $L_G = 2C(1 + 2B) + \sqrt{1 + 2B^2}$. Its final mean values are 0.00115440, 0.000472143, and 0.000463279 for uniform, boundary, and zero starts. These are ordinary float64 evaluations of a deterministic bound expression, not certified interval enclosures or evaluations of the finite-sample theorem's confidence radius.

The coefficient error after the single clean-label update is 2.78×10^{-16} on average. Attention and operator errors decline from all three starts (Figure 7), supporting finite learning of the prescribed update rule. Deployment freezes all parameters and gives every method the identical single length-1024 record on each task. Starting at $\pi_0(s) = -1$, each of $J = 8$ rounds re-encodes stored successors using the current policy, resets memory to zero, executes $L = 64$ shared clipped loops, and greedifies the critic. Computed exact ties choose action -1. The same transitions are reused at every loop and round. Reference values solve the true-MDP Bellman systems and enumerate the 16 deterministic policies; these references are used only to assess saved policies, never to select learned-model actions. The reported final policy gap is $D_8 = \|V^\star - V^{\pi_8}\|_\infty$.

By the saved 1024-update checkpoint, every initialization class matches the teacher's final ring actions and mean gap, even though the uniform-start operator RMSE is still 0.261796. Further training reduces operator error without changing those final actions. The teacher's mean gap is the positive 7.33741×10^{-4} shown in Table 5; matching it does not imply true-MDP optimality. The finite tested starts support the specified mechanism without establishing convergence from every point in the initialization box. A matched-data coefficient-only control fixes teacher attention and uses 4096 full-data projected gradient steps with $\eta = 1/\lambda_{\max}(H)$ for its cached quadratic Hessian. Its mean operator RMSE is 9.99120×10^{-15} and its final actions/gaps match the teacher in all four families. Grouped curriculum updates and full-data coefficient updates have different costs, so their step counts are not a runtime comparison.

Table 6: Fresh-task final gaps. Each listed mean, 95th percentile, and maximum is shared by all three final curriculum starts, the teacher, the coefficient control, and the empirical-model planner. Intervals shown are the saved crossed-bootstrap intervals for the curriculum; teacher/planner use task-only bootstrap intervals. Percentiles use 50 taskwise seed means, and maxima use all seed/task cells.

Family	Mean gap	95% interval	95th percentile	Maximum
Ring	0.000733741	$[0, 0.00183921]$	0	0.0185863
Dirichlet, $\alpha = 0.2$	0.00336813	$[0.000317749, 0.00857300]$	0.0164447	0.112677
Dirichlet, $\alpha = 0.4$	0.00112071	$[0, 0.00328554]$	0	0.0522064
Dirichlet, $\alpha = 0.8$	0.00560340	$[0, 0.01353048]$	0.0183022	0.155524

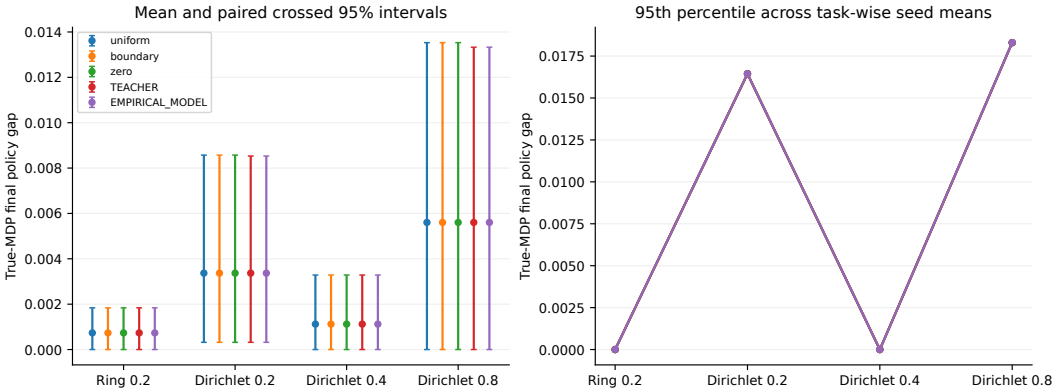

Figure 8: Complete frozen-transfer readouts on the original 200-task bank. Left: means and 95% bootstrap intervals; markers are horizontally offset for readability. Right: descriptive 95th percentiles of the 50 taskwise seed means; identical curves overlap. These are not tail confidence bounds. The Dirichlet $\alpha = 0.2$ family is an empirical stress test that fails the conservative score condition at $\tau = 5$.

C.2.3 Transfer across transition kernels and classical planning

The frozen checkpoint manifest precedes generation of the final deployment banks. There are 50 fresh random tasks in each of four families (200 total), plus one separate planning task in the ring family. Alongside fresh ring rewards, three families use

$$P_{\alpha,\widetilde{P}}(s' \mid s, a) = (1 - \alpha)\widetilde{P}(s' \mid s, a) + \alpha/4, \qquad \alpha \in \{0.2, 0.4, 0.8\},$$

with each row of $\widetilde{P}$ independently drawn from $\mathrm{Dirichlet}(1, 1, 1, 1)$ and fresh independent uniform reward entries. The policy, memory, and context budgets are unchanged. This tests transfer across rewards and transition kernels on the same address space; it does not test transfer to other state counts.

The classical comparator estimates rewards and successor frequencies from *the same behavior record*. For any unvisited pair it assigns reward zero and uniform successors, then performs eight policy-improvement rounds with exact empirical-model policy evaluation. No true transition matrix or full reward table is provided to this comparator or to the learned evaluator. All four final task banks happen to visit every pair, so the prespecified fallback is unused. This comparison matches observation access, not arithmetic operations or optimization time.

Every final curriculum model matches the teacher's actions on all 200 random tasks and the separate planning task. The empirical planner has the same final policies/gaps, establishing matched performance on these banks rather than an advantage over classical planning. In the separate planning task the reward rows in action order $(-1, +1)$ are $(0.1, 0)$, $(1, 1)$, $(0, 0.1)$, and $(-1, -1)$. The trained models and both references have zero recorded final gap; this task is excluded from random-task averages.

For random kernels, uniform behavior gives conservative common constants $\mu_0 = \alpha/8$, $C_{\mathrm{mix}} = 1$, $r_{\mathrm{mix}} = 1 - \alpha$, and $H = 1/\alpha$. The ring admits the sharper $\mu_0 = 1/8$. The respective sufficient score

Table 7: Mean clean-test operator RMSE over five seeds for the paired sample/noise sweep. Primary and reference attention budgets are 262,144 and 1,048,576. A reference run changes computation without collecting new samples.

N	σ	Primary RMSE	Longer-reference RMSE
256	0	$1.34619\,10^{-6}$	$1.62453\,10^{-12}$
1024	0	$1.34619\,10^{-6}$	$1.62458\,10^{-12}$
4096	0	$1.34619\,10^{-6}$	$1.60498\,10^{-12}$
256	0.001	0.00754296	0.00754345
1024	0.001	0.00230616	0.00230633
4096	0.001	0.00120522	0.00120525
256	0.01	0.0812539	0.0827228
1024	0.01	0.0230468	0.0230522
4096	0.01	0.0120890	0.0120899

thresholds are 3.68051 (ring), 5.39816 ($\alpha = 0.2$), 4.67904 ($\alpha = 0.4$), and 3.93183 ($\alpha = 0.8$). Hence the random-kernel $\alpha = 0.2$ slice fails this sufficient condition at $\tau = 5$ and is an empirical stress test. The other slices pass the score condition, which alone does not establish the training or context confidence events. No statistical certificate experiment was conducted in that earlier campaign.

C.2.4 Sample size, bounded noise, and action margins

The uniform-initialization sweep uses $N \in \{256, 1024, 4096\}$ and $\sigma \in \{0, 0.001, 0.01\}$. Noise has independent Rademacher coordinates $\zeta_i(x) = \sigma\varepsilon_{i,x}/\sqrt{8}$, so it is conditionally centered and has Euclidean norm σ. The sample-size settings are prefixes of each of five independent 4096-example banks; noise settings share inputs and coordinate signs. Consequently the 45 grid runs are paired, not 45 independent datasets. The longer reference reuses exactly the same inputs and labels with four times as many attention updates.

More samples reduce operator error under noisy supervision, while additional optimization chiefly removes the clean-label residual. Identical clean probe labels need not benefit from more examples after every group appears. Optimizing a noisy grouped target more accurately can slightly increase its clean-test error. These observations separate optimization residue from finite-sample noise. Appendix C.1.4 extends the positive-noise comparison to six sample sizes and reports the fitted sample dependence together with an explicit diagnostic separating optimization and statistical errors.

For the action diagnostic, both critics evaluate the same initial all-minus policy on the same context with zero initial memory and 64 loops. Write $\Delta_T(s)$ for the teacher critic's gap between its two actions and $d = \|\widehat{Q}_{\text{learned}} - \widehat{Q}_T\|_\infty$. The strict sufficient implication $d < \Delta_T(s)/2$ ensures identical greedy actions at s, excluding teacher ties. There are zero violations among the saved diagnostic cells. Failure of this sufficient condition does not imply disagreement. These common-policy comparisons are separate from each method's own eight-round control trace; final teacher agreement is neither an estimate of the all-input operator error nor a claim of true-MDP optimality.

C.2.5 Uncertainty, resource accounting, and implementation scope

Intervals use 2000 percentile bootstrap replicates, independently resampling the five seed IDs and 50 task IDs within each family and preserving method pairing. The teacher and empirical planner have no training randomness, so their intervals resample tasks only. Their saved upper endpoints are 0.00183921, 0.00853098, 0.00328554, and 0.01332982 in family order; the small differences from the curriculum endpoints reflect resampling streams, not different observed gaps. Absolute intervals reflect task variability even when every trained seed chooses the same actions. The 95th percentiles are computed from 50 taskwise seed means, avoiding quantile changes from duplicating a teacher task across seed IDs; maxima use all cells. Tail summaries are descriptive. Bootstrap intervals are separate from the failure probability in the theoretical statements.

The five largest training banks contain 20,480 unique operator examples, 247,488 physical transitions, and 5,012 synthetic probes. The smaller prefixes, three initializations, noise levels, and repeated updates reuse these inputs. The final 201 deployment records contain 205,824 unique transitions. The

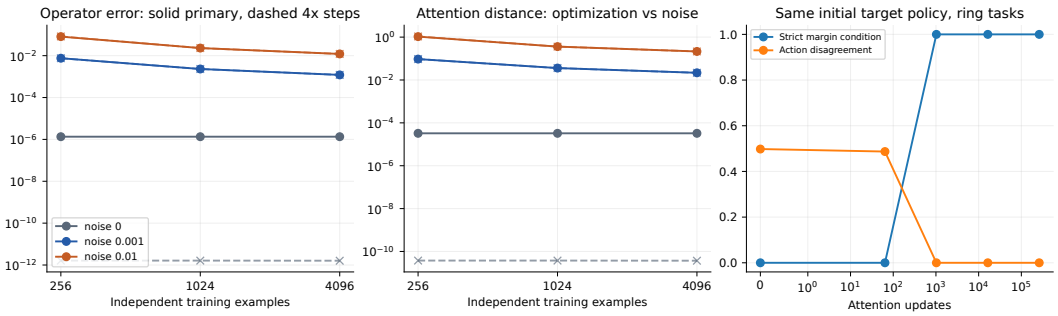

Figure 9: Sample/noise and action-margin diagnostics. Left and middle: seed means on the shared clean operator bank; solid curves use the primary budget and dashed curves use the four-times-longer reference on the same examples. Right: fractions satisfying the strict sufficient margin condition and fractions of disagreeing actions on the common initial-policy ring diagnostic at saved checkpoints. Step zero follows coefficient fitting.

pilot uses 12,560 physical training and 9,216 deployment transitions; operator test and validation each use 4,642 physical transitions and are counted separately. The primary grouped computation processes three singleton representatives and seven pair representatives per attention step, for $3 + 7(262144) = 1,835,011$ representative evaluations per run, versus $7,340,035$ for the longer reference. These are computation counts, not new training examples. The report records 208 logical deployment runs and 307.93 seconds of deployment computation; methods and checkpoints share records.

The implementation uses CPU float64 with one Torch thread. It caches immutable query/key token projections while retaining token reads, scratch reset, TD aggregation, and four-ReLU clipping. Group means are cached once; no inverse logistic solve replaces the attention optimizer. The saved report records 12 new tests, 117 architecture checks, nine original extensive-study tests, and 4260 independent saved-artifact checks, all passing. The maximum independent 64-loop replay discrepancy is 6.33×10^{-10} (rounded upward), satisfying the declared coordinatewise tolerance $10^{-10} + 10^{-9} \max(|a|, |b|)$ in every check. These finite numerical checks support implementation equivalence without verifying statistical coverage.

The experiment instantiates the prescribed finite-address construction with operator supervision and identifiable probes; it does not train a generic Transformer from ordinary trajectories alone. The optimizer and deployment budgets are pilot-selected practical budgets, not sufficient worst-case budgets evaluated from the theorem. The matched-data coefficient control optimizes the full mixture loss with prescribed attention.

C.3 Prescribed-attention special case: coefficient-only training

This study evaluates the special case in Appendix A.1.3: attention is fixed at its prescribed teacher value, and gradient descent learns the three TD coefficients before frozen policy improvement. It complements the full coefficient-and-attention curriculum in Appendix C.2 by isolating learning with supplied attention.

C.3.1 Environment, architecture, and protocol

Tasks have four states, actions $\{-1, +1\}$, discount $\gamma = 0.7$, and

$$P(s' \mid s, a) = 0.8\mathbf{1}\{s' = (s + a) \bmod 4\} + 0.2/4.$$

Each random task has eight independent rewards drawn uniformly from $[-1, 1]$, deterministic conditional on the state-action pair. Behavior starts uniformly and selects uniform random actions, with no additional within-record resets. The stationary state-action mass is $1/8$, and the state dynamics contract in total variation by at most 0.8 per step. We use $n = 8$, $B = 10/3$, and addresses $x(s, a) = 2s + \mathbf{1}\{a = +1\}$.

The width-20 token evaluator uses eight memory tokens, the observed transition tokens, and one dummy token. Each shared loop has four fixed singleton read/reset heads, one softmax aggregation

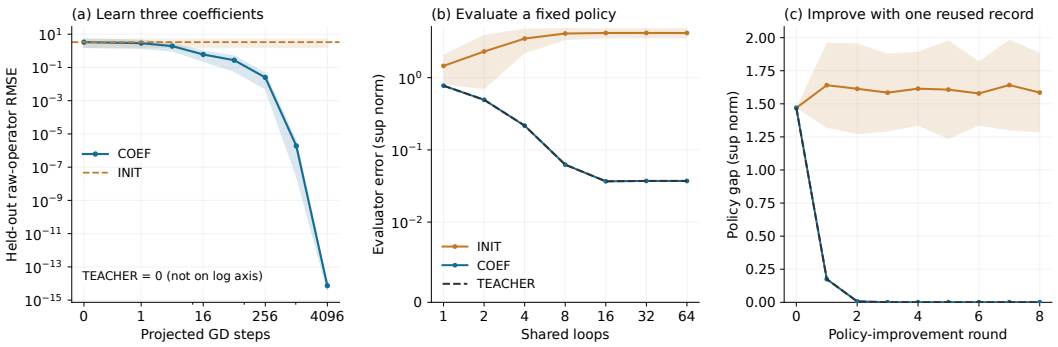

Figure 10: Coefficient-only training: (a) held-out raw-operator RMSE at prespecified GD checkpoints, (b) fixed-policy error through 64 shared loops, and (c) policy gap over eight improvements on the same context prefix. Shading is one sample SD across five training seeds, conditional on the shared task bank; panels (b,c) first average over its 50 random tasks. TEACHER has analytically zero operator error and no training randomness. Learned and teacher curves overlap. Nonnegative bands are truncated at zero, and the planning task is reported separately.

head, and fixed four-ReLU clipping, implemented in two attention sublayers. There are no extra learned projections or action-classification head. Attention remains at the teacher setting with gap $\tau = 5$; only the three coefficients are trained in $[-2, 2]^3$, with teacher value $(1, -1, 0.7)$.

For each of five training seeds (1001–1005), $N = 1024$ independent clean-label examples are sampled from a mixture: ordinary examples with probability $3/4$ and singleton coefficient probes with probability $1/4$, uniformly over the three probe types. An ordinary example draws a fresh reward task, a uniform deterministic target policy, memory in $[-B, B]^8$, and a length-16 behavior record. Probe inputs are synthetic and counted separately from physical transitions. Labels are raw one-step teacher outputs. The loss is $(2N)^{-1} \sum_i \|G_\theta(U_i) - Y_i\|_2^2$, with no additional division by the eight output coordinates. This study uses the mixture loss; the grouped curriculum loss is evaluated separately in Appendix C.2.

Coefficients start independently and uniformly in $[-2, 2]^3$. Projected full-batch gradient descent runs for 4096 steps with $\eta_c = [8(1 + 2(10/3)^2)]^{-1}$. Cached token features reproduce the checked optimization iterates. A separate bank of 256 ordinary inputs measures raw operator RMSE. INIT uses the initial coefficients, COEF the final learned coefficients, and TEACHER the prescribed coefficients; all use the same attention and deployment procedure.

The frozen evaluators share 50 held-out random tasks and their length-1024 behavior records, plus one separate planning task. Starting from $\pi_0(s) = -1$, each of eight improvement rounds re-encodes the stored successors under the current target policy, resets memory to zero, applies 64 clipped loops, and greedifies the critic. Exact computed ties select -1. The same record is reused without new observations or weight updates. We report $e_j = \|\widehat{Q}_j - Q^{\pi_j}\|_\infty$ and $D_j = \|V^\star - V^{\pi_j}\|_\infty$; final evaluator error is e_7 and final policy gap is D_8. Reference values enumerate the 16 deterministic policies and solve their Bellman equations, solely for evaluation.

C.3.2 Coefficient recovery and frozen policy improvement

Mean held-out raw-operator RMSE decreases from 3.32203 ± 1.73193 to $(7.3837 \pm 3.2042) \times 10^{-15}$, with sample SD across training seeds. After 64 loops on the initial fixed policy, COEF and TEACHER both have mean evaluator error 0.0371507. The initial policy gap is $D_0 = 1.469115$. After eight improvements, the mean final gap is 1.584943 ± 0.301331 for INIT and 0.000873190 for both COEF and TEACHER. The COEF seed SD is zero conditional on this common task bank. Figure 10 connects coefficient recovery, repeated fixed-policy evaluation, and policy improvement using the frozen update.

For the planning task, reward rows in action order $(-1, +1)$ are $(0.1, 0)$, $(1, 1)$, $(0, 0.1)$, and $(-1, -1)$. At states 0 and 2, the optimal action moves toward state 1 despite the opposing immediate-reward preference. COEF selects $+1$ at state 0 and -1 at state 2 in every seed, matching TEACHER, with

final gap at most 3.89×10^{-16}. The task and its trajectory were fixed in advance and excluded from random-task averages.

Implementation checks and reproducibility. All 659 float64 implementation checks passed at coordinatewise tolerance $10^{-10} + 10^{-9} \max(|a|, |b|)$. Independent token/algebraic comparisons cover raw and clipped outputs, repeated loops, scratch refresh, gradients, projected updates, padding, and frozen attention. Maximum discrepancies are 8.44×10^{-15} for raw outputs, 5.68×10^{-14} for coefficient gradients, and 2.22×10^{-16} for cached projected iterates; the maximum reference Bellman residual is 1.33×10^{-15}. These finite checks assess implementation accuracy rather than statistical coverage.