EFFICIENT ONLINE REINFORCEMENT LEARNING WITH ALL-POLICY LINEAR Q-REALIZABILITY AND SPARSE STOCHASTIC TRANSITIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study computationally efficient online reinforcement learning in potentially infinite state spaces. At each layer, every Markov policy's action-value function is linear in the same known d-dimensional features, and each transition has at most K possible successor states. Under these assumptions, we give a polynomial-time algorithm with high-probability regret

$$\widetilde{O}\Big(dH\sqrt{A\min\{A,d\}KH(H-1)T}\Big)$$

for A actions, horizon $H \geq 2$, and T episodes. The algorithm discovers a finite set of noninitial states on which to concentrate learning, then combines a linear contextual bandit at the first layer with tabular learning on stored states and a fixed default policy elsewhere. The guarantee allows rewards in $[0, 1]$ and arbitrary initial-state distributions. The algorithm uses only online episodes, without a simulator, resets to chosen states, or a policy-optimization oracle, and requires neither a linear transition model nor Bellman completeness.

1 INTRODUCTION

Online reinforcement learning in large state spaces can use state–action features to generalize beyond observed pairs. Common structural assumptions include linear MDPs, which impose linear structure on rewards and transitions, and Bellman completeness, which requires closure of the function class under Bellman backups (Jin et al., 2020; Zanette et al., 2020).

We study *all-policy linear Q-realizability*: at each layer, every Markov policy's action-value function is linear in the same known features. The representing weights may depend on the policy and layer, but are shared across all state–action pairs in that layer. Existing algorithms establish sample efficiency but can require expensive optimization (Weisz et al., 2023; Mhammedi, 2025). We seek computationally efficient learning in a setting covering both deterministic dynamics (Ke et al., 2026) and finite tabular MDPs, without a linear transition model or Bellman completeness.

This motivates *sparse stochastic transitions*: each state–action pair has a bounded number of possible successor states. This permits stochastic branching, potentially infinite state spaces, and arbitrary initial-state support. Beyond exact realizability, we also allow uniform value-approximation error and *soft sparsity*, where a bounded set of successors carries all but a prescribed tail mass.

Our approach concentrates learning on a finite set of noninitial states and bounds the loss from using a fixed default policy elsewhere. Making this reduction work online presents three difficulties. The relevant set is unknown, and states unimportant under one continuation policy may matter under another. Discovering new states changes the reachability targets used to find further states. Finally, updating the learned continuation policy changes the returns used to learn the initial action, introducing regression bias.

Contributions.

1. **Polynomial-time online learning.** Our algorithm has polynomial computational cost and uses only online episodes, without a simulator, resets to chosen states, or a policy-optimization oracle.

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.

2. **Square-root regret.** Bound (1) gives square-root regret under exact realizability and sparse transitions, up to logarithms, and quantifies the additional linear costs of approximation and transition tails.
3. **A support cover shared across policies.** Lemma 5.1 identifies a cover in each noninitial layer of size proportional to the feature dimension times the transition support bound. Outside it, action choices have uniformly limited expected effect across continuation policies, enabling discovery.
4. **Complementary lower bounds.** For $A = K = 2$, $B \geq H$, and $\kappa = \upsilon = 0$, our upper bound is $\widetilde{O}(dH^2\sqrt{T})$, while Appendix C gives an expected-regret lower bound of $\Omega(H^{3/2}\sqrt{dT})$. The bounds differ by only a factor $\sqrt{dH}$, up to logarithms. A linear approximation-error cost is also unavoidable.

2 Related work

All-policy linear realizability. Table 1 compares the closest online results under exact realizability. Prior work skips states with small action ranges to obtain sample efficiency, but uses computationally inefficient optimization (Weisz et al., 2023). Our support cover shared across policies enables polynomial-time online discovery and learning under sparse stochastic transitions. Our guarantee extends efficient online learning beyond the deterministic dynamics considered in Ke et al. (2026).

Table 1: Selected online guarantees under exact all-policy linear Q-realizability. PAC bounds give samples for an ε-optimal policy ($\varepsilon > 0$); regret uses T episodes. Here d, H, A, K are feature dimension, horizon, action count, and successor-support bound; CSC denotes cost-sensitive classification. Confidence and logarithmic factors are suppressed.

Work	Transitions	Computation	Statistical guarantee
Weisz et al. (2023)	General stochastic	Inefficient global optimization	PAC: $\widetilde{O}(H^{11}d^7/\varepsilon^2)$ steps
Mhammedi (2025)	General stochastic	Polynomial CSC calls; efficient for fixed d	PAC: $\mathrm{poly}(A, H, d, \varepsilon^{-1})$ episodes
Ke et al. (2026)	Deterministic	Polynomial	Regret: $\widetilde{O}(dH^3\sqrt{T})$
This work	K-sparse stochastic	Polynomial; no CSC oracle	Regret: $\widetilde{O}(dH^2\sqrt{A\min\{A,d\}KT})$

Weisz et al. (2023) use a fixed initial state; the other rows allow random initial states.

Bellman completeness. Under Bellman completeness, ELEANOR is statistically efficient but uses a nonconvex optimistic program (Zanette et al., 2020). PSDP-UCB is efficient for each fixed number of actions, with an exponent depending on the action count and an additional coefficient-norm assumption (Golowich & Moitra, 2024). Deterministic transitions admit algorithms whose complexity is polynomial in the action count as well (Mhammedi et al., 2026).

Online learning tools. We build on variance-sensitive finite-state methods (Zhang et al., 2025; Zhou et al., 2023). The root learner uses OFUL (Abbasi-Yadkori et al., 2011) or combines SquareCB (Foster & Rakhlin, 2020) with a sequential linear square-loss forecaster (Azoury & Warmuth, 2001; Vovk, 2001). Our analysis extends the finite-state arguments to changing first-hit targets during discovery and adaptively selected episodes during restricted learning. The analysis of restricted learning yields a squared continuation-error bound that controls the root learner's changing regression bias (Section 5.3). Corrupted linear bandits (He et al., 2022) provide a related treatment of biased observations.

3 Problem formulation and main result

States, actions, and observations. Fix positive integers A, d, K and an integer horizon $H \geq 2$. The layered MDP has standard Borel state spaces $\mathcal{S}_1, \ldots, \mathcal{S}_H$, terminal space $\mathcal{S}_{H+1} = \{\perp\}$, actions

$\mathcal{A} = [A]$, and an arbitrary unknown initial distribution μ on $\mathcal{S}_1$. State identities include their layers. At (s, a) in layer h, a Markov kernel generates $(R_h, S_{h+1}) \in [0, 1] \times \mathcal{S}_{h+1}$; reward and next state may be correlated. Write $r_h(s, a) = \mathbb{E}[R_h \mid s, a]$ and let $P_h(\cdot \mid s, a)$ be the next-state marginal.

Each episode starts with an independent draw $S_1 \sim \mu$. Conditional on the complete history and selected pair, each outcome is a fresh draw from its kernel. The learner observes states, chooses actions, and receives rewards and next states, with no other environment access. Write $S_h^{(t)}, A_h^{(t)}, R_h^{(t)}$ for the episode-t trajectory.

Values and regret. For a randomized Markov policy π, let

$$V_h^\pi(s) = \mathbb{E}_\pi\left[\sum_{j=h}^{H} R_j \mid S_h = s\right], \qquad Q_h^\pi(s, a) = r_h(s, a) + \mathbb{E}_{S' \sim P_h(\cdot|s,a)} V_{h+1}^\pi(S'),$$

with $V_{H+1}^\pi(\perp) = 0$, and write $J(\pi) = \mathbb{E}_{S_1 \sim \mu} V_1^\pi(S_1)$. Fix a deterministic Bellman-optimal policy $\pi^\star$ with fixed tie-breaking, and write $V_h^\star = V_h^{\pi^\star}$, $Q_h^\star = Q_h^{\pi^\star}$, and $J^\star = J(\pi^\star)$. The possibly randomized episode-t policy π_t is measurable with respect to preceding data. Define pseudo-regret by

$$\mathrm{Reg}_T := \sum_{t=1}^{T} \big(J(\pi^\star) - J(\pi_t)\big).$$

Assumption 3.1 (All-policy linear action values)**.** *There are known scalars $B > 0$ and $\kappa \in [0, H]$ and a known measurable feature map $\phi_h : \mathcal{S}_h \times \mathcal{A} \to \mathbb{R}^d$, with $\|\phi_h(s, a)\|_2 \le 1$, such that for every deterministic Markov policy π and layer h there is a vector $\theta_h^\pi \in \mathbb{R}^d$ satisfying*

$$\sup_{(s,a) \in \mathcal{S}_h \times \mathcal{A}} \left|Q_h^\pi(s, a) - \phi_h(s, a)^\top \theta_h^\pi\right| \le \kappa, \qquad \|\theta_h^\pi\|_2 \le B.$$

The constants B, κ are common to all policies and layers, while the representing vector may vary with both. The uniform bound includes policies differing at one state. Lemma B.1 in Appendix B.1 extends the assumption to randomized Markov policies with the same B, κ.

Assumption 3.2 (Sparse transition support)**.** *The learner knows bounds $K \in \mathbb{N}$ and $\upsilon \in [0, 1]$ such that, for every $h < H$ and $(s, a) \in \mathcal{S}_h \times \mathcal{A}$, there is a finite set $\mathcal{N}_h(s, a) \subseteq \mathcal{S}_{h+1}$ with $|\mathcal{N}_h(s, a)| \le K$ and $P_h(\mathcal{N}_h(s, a) \mid s, a) \ge 1 - \upsilon$. The sets and transition probabilities are unknown. The remaining mass may have arbitrary support, and μ is unrestricted.*

Define the discovery error scale $\chi = \kappa + H\upsilon/2 \in [0, 3H/2]$: κ measures value approximation and υ controls transition tails.

Computation model. The learner knows A, d, H and accesses features and state equality through oracles; all action features are available at observed states. Complexity counts scalar arithmetic, comparisons, square roots, logarithms, integer rounding, oracle operations, and draws from computed finite-action distributions. Fixed tie-breaking gives measurable selectors.

Theorem 3.3 (Efficient regret)**.** *Under Assumptions 3.1 and 3.2, for every integer episode budget $T \ge 1$ and failure probability $\delta \in (0, 1)$, there is an online algorithm using a number of real-arithmetic and evaluation-oracle operations polynomial in A, d, H, K, T, B and $\log(1/\delta)$ such that, with probability at least $1 - \delta$,*

$$\mathrm{Reg}_T \le \widetilde{O}\Big(dH\sqrt{A \min\{A, d\} KH(H-1)T} + A \min\{A, d\} H^2 (H-1) d^2 \chi T\Big). \tag{1}$$

The notation $\widetilde{O}$ hides universal constants and logarithmic factors in the displayed parameters, B, and $1/\delta$, but no polynomial dependence. The right side may be capped at HT; for $A = 1$, the regret is identically zero.

Reading the bound. When $\kappa = \upsilon = 0$, bound (1) gives square-root regret in T, up to logarithms. Discovery incurs the additional χT cost; restricted learning contributes only $\widetilde{O}(\kappa\sqrt{\min\{A, d\}T})$, independent of υ.

4 ALGORITHM

Algorithm 1 first uses Procedure 1 to build a finite set of noninitial states and certify the loss from using a fixed default policy elsewhere. Procedure 4 then combines a linear contextual bandit for the initial action with tabular learning on stored states.

Fix the deterministic default policy $\pi^0 = (\pi_h^0)_{h=1}^H$ with $\pi_h^0(s) = 1$ on $\mathcal{S}_h$. For finite sets $Z_h \subseteq \mathcal{S}_h$, $h \geq 2$, write $Z = (Z_2, \ldots, Z_H)$ and identify Z with its layer-tagged union. The class $\Pi(Z)$ consists of the Markov policies that may choose freely at the initial state (the root) and at states in Z, and use π^0 at every other noninitial state. Let $J_Z^\star = \sup_{\pi \in \Pi(Z)} J(\pi)$. Backward induction on the finite set Z in layer order, followed by finite-action maximization at the root with fixed tie-breaking, yields a deterministic policy attaining $J_Z^\star$.

The matrix count $L_{\mathrm{d}} \in \mathbb{N}$ and positive scalars $\lambda_{\mathrm{d}}, \eta, p_\eta, \delta_{\mathrm{d}}$ are fixed by (10)–(14) in Appendix A.1.

Algorithm 1 Main algorithm: discovery followed by restricted online learning

Require: Known A, d, H, feature oracles $(\phi_h)_{h=1}^H$, episode budget $T \in \mathbb{N}$, failure probability $\delta \in (0,1)$, and bounds $B > 0, K \in \mathbb{N}, \kappa \in [0,H], \upsilon \in [0,1]$.
Local variables: $N_{\mathrm{d}} \in \{0, \ldots, T\}$, $\sigma \in \{\textsc{Certified}, \textsc{BudgetExhausted}\}$.
Compute $L_{\mathrm{d}}, \lambda_{\mathrm{d}}, \eta, p_\eta, \delta_{\mathrm{d}}$ from (10)–(14), with $\chi = \kappa + H\upsilon/2$.
if $\eta > H$ **then**
  Execute π^0 for all T episodes and terminate.
$(Z, N_{\mathrm{d}}, \sigma) \leftarrow \textsc{SimultaneousDiscovery}(T, \delta_{\mathrm{d}}, p_\eta, L_{\mathrm{d}}, \lambda_{\mathrm{d}}, H, \pi^0)$.
if $\sigma = \textsc{Certified}$ and $N_{\mathrm{d}} < T$ **then**
  Call $\textsc{RestrictedLearn}(Z, T - N_{\mathrm{d}}, \delta/2, B, H, \kappa, \pi^0)$.

4.1 DISCOVERY PROCEDURES

Procedure 1 returns Z, the number N_{d} of episodes used, and the status σ. On the confidence event of Theorem 5.2, a CERTIFIED output guarantees $J^\star - J_Z^\star \leq \eta$. Discovery uses the formulas in Appendices A.2 and A.3. For a finite set E, let $\Delta(E)$ be its probability simplex.

Stored state and novelty test. The maintained discovery state is the tuple

$$\mathsf{D} = (Z, W, \mathcal{U}, \mathsf{A}^{\mathrm{d}}, \mathsf{F}^{\mathrm{d}}, m^{\mathrm{d}}, Q^+).$$

Here $W = (W_h^{(\iota)})_{h=2,\ldots,H;\, \iota \in [L_{\mathrm{d}}]}$ is an array of positive-definite matrices $W_h^{(\iota)} \in \mathbb{R}^{d \times d}$, and $\mathcal{U} = \{u_\circ\} \cup \{(z,a) : z \in Z, a \in \mathcal{A}\}$ is the finite set of controlled pairs. The synthetic start pair $u_\circ$ represents sampling $S_1 \sim \mu$ and $A_1 \sim \lambda_{S_1}$ at the root; Procedure 3 returns the distribution $\lambda_s \in \Delta(\mathcal{A})$. For $u \in \mathcal{U}$, the active batch $\mathsf{A}^{\mathrm{d}}(u)$ and frozen batch $\mathsf{F}^{\mathrm{d}}(u)$ are finite label sequences in $[0,1]$. New labels enter the active batch, whose target size is $m^{\mathrm{d}}(u) \in \mathbb{N}$; the frozen batch is the last completed batch. The score table is $Q^+ : \mathcal{U} \to [0,1]$, with induced value table $V^+ : Z \to [0,1]$, $V^+(z) = \max_{a \in \mathcal{A}} Q^+(z,a)$.

For $x \in \mathcal{S}_h$, define the feature-difference set

$$\mathcal{V}_h(x) = \{\phi_h(x,a) - \phi_h(x, \pi_h^0(x)) : a \in \mathcal{A}\} \subseteq \mathbb{R}^d \tag{2}$$

and define the novelty set by

$$\mathcal{G}(Z, W) = \bigcup_{h=2}^{H} \left\{ x \in \mathcal{S}_h \setminus Z_h : \max_{\iota \in [L_{\mathrm{d}}],\, v \in \mathcal{V}_h(x)} v^\top (W_h^{(\iota)})^{-1} v > 1 \right\}. \tag{3}$$

At an observed state $x \in \mathcal{S}_h \setminus Z_h$, the *novelty test* checks whether the maximum inverse-matrix squared norm in (3) exceeds one. This uses only Z, W, and known features. Procedure 1 initializes the matrices using the identity matrix I_d. The scales in (14) account for approximation error and transition tails.

First-hit observations and their target. For $u = (z,a)$ with $z \in \mathcal{S}_h$, let $\Omega_u = \prod_{j=h}^H ([0,1] \times \mathcal{S}_{j+1})$ be the suffix space. A *primitive continuation* is the complete random suffix

$$\Omega^u = ((R_j, S_{j+1}))_{j=h}^H \in \Omega_u \tag{4}$$

generated by taking a at z and then following π^0, with probability law P_u on Ω_u. For $u_\circ$, put $\Omega_{u_\circ} = \mathcal{S}_1 \times \mathcal{A} \times \prod_{j=1}^{H}([0,1] \times \mathcal{S}_{j+1})$ and define

$$\Omega^{u_\circ} = (S_1, A_1, ((R_j, S_{j+1}))_{j=1}^{H}) \in \Omega_{u_\circ},$$

where $S_1 \sim \mu$, $A_1 \sim \lambda_{S_1}$, and the process follows π^0 after the root; let $P_{u_\circ}$ be this law.

At the start of episode t, keep the frozen batches $\mathsf{F}_t^{\mathrm{d}}$, sets $Z_t, \mathcal{U}_t$, matrices W_t, and score table $Q_t^+ : \mathcal{U}_t \to [0,1]$ fixed for the episode. The corresponding value map is $V_t^+ : Z_t \to [0,1]$, with $V_t^+(y) = \max_a Q_t^+(y,a)$. On a primitive continuation $\omega \in \Omega_u$, stop at the first later state in Z_t, the first state in $\mathcal{G}(Z_t, W_t)$, or termination, and define the measurable label $F_t^u : \Omega_u \to [0,1]$ by

$$F_t^u(\omega) = \begin{cases} 1, & \text{the first event is a state in } \mathcal{G}(Z_t, W_t), \\ V_t^+(y), & \text{the first event is } y \in Z_t, \\ 0, & \text{the first event is termination.} \end{cases} \tag{5}$$

A visit to u reveals the prefix determining F_t^u and appends its observed label to $\mathsf{A}^{\mathrm{d}}(u)$. The conditional-mean target is $q_t^{\mathrm{d}} : \mathcal{U}_t \to [0,1]$ given by $q_t^{\mathrm{d}}(u) = \int F_t^u \, dP_u$. Lemma B.6 proves this conditional law.

Optimistic reachability and actions. Equation (21) computes a score from the frozen batch and sets $Q_t^+(u)$ to its minimum with the previous score. Thus $Q_t^+(u)$ optimistically estimates the maximum probability of reaching the current novelty set from u, choosing actions on Z_t and using π^0 elsewhere. The greedy map $a_t^+ : Z_t \to \mathcal{A}$ chooses the action with the largest estimate: $a_t^+(y) = \min \arg\max_{a \in \mathcal{A}} Q_t^+(y,a)$. Discovery learns Q^+ tabularly on $\mathcal{U}_t$: the Markov property and acyclic layer order justify the recursion without Bellman completeness. Assumption 3.1 enters the discovery certificate, which uses novelty-hit probabilities to bound the value lost outside Z.

Episode updates. Procedure 1 appends only to active batches during an episode and keeps its policy fixed even after a novelty hit. At the episode boundary, (21) promotes full batches and updates scores; incomplete batches carry over. A discovered state is then inserted, one matrix is updated, and all new pairs are initialized. Write (Z_{t+1}, W_{t+1}) for the resulting state, with (Z_1, W_1) denoting initialization.

The procedure's inputs are $N, L_{\mathrm{d}}, H \in \mathbb{N}$, $\delta_{\mathrm{d}} \in (0,1)$, $p_\eta \in (0,1]$, $\lambda_{\mathrm{d}} > 0$, and the deterministic Markov policy π^0; it returns a finite layer-tagged set Z, an episode count in $\{0, \dots, N\}$, and a status in $\{\textsc{Certified}, \textsc{BudgetExhausted}\}$.

Procedure 1 SIMULTANEOUSDISCOVERY

procedure SIMULTANEOUSDISCOVERY($N, \delta_{\mathrm{d}}, p_\eta, L_{\mathrm{d}}, \lambda_{\mathrm{d}}, H, \pi^0$)
  Initialize $Z_h \leftarrow \varnothing$ and $W_h^{(\iota)} \leftarrow \lambda_{\mathrm{d}} I_d$ for $h = 2, \dots, H$ and $\iota \in [L_{\mathrm{d}}]$.
  Set $\mathcal{U} \leftarrow \{u_\circ\}$ and initialize $(\mathsf{A}^{\mathrm{d}}(u_\circ), \mathsf{F}^{\mathrm{d}}(u_\circ), m^{\mathrm{d}}(u_\circ), Q^+(u_\circ)) \leftarrow (\varnothing, \varnothing, 1, 1)$.
  for $t = 1, \dots, N$ **do**
    if $Q^+(u_\circ) < p_\eta$ **then**
      return $(Z, t-1, \textsc{Certified})$.
    $(x, \mathsf{D}) \leftarrow \textsc{FirstHit}(t, \mathsf{D}, \pi^0)$ by Procedure 2.
    After episode t ends, update $(\mathsf{F}^{\mathrm{d}}, \mathsf{A}^{\mathrm{d}}, m^{\mathrm{d}}, Q^+)$ according to (21).
    if $x \in \mathcal{S}_h$ **then**
      Put $v_x(a) = \phi_h(x,a) - \phi_h(x, \pi_h^0(x)) \in \mathbb{R}^d$ for $a \in \mathcal{A}$.
      Choose the lexicographically first $(\iota_x, a_x) \in [L_{\mathrm{d}}] \times \mathcal{A}$ with $v_x(a_x)^\top (W_h^{(\iota_x)})^{-1} v_x(a_x) > 1$.
      Set $W_h^{(\iota_x)} \leftarrow W_h^{(\iota_x)} + v_x(a_x) v_x(a_x)^\top$ and $Z_h \leftarrow Z_h \cup \{x\}$.
      Add $\{(x,a) : a \in \mathcal{A}\}$ to $\mathcal{U}$.
      For each new pair u, set $(\mathsf{A}^{\mathrm{d}}(u), \mathsf{F}^{\mathrm{d}}(u), m^{\mathrm{d}}(u), Q^+(u)) \leftarrow (\varnothing, \varnothing, 1, 1)$.
  if $Q^+(u_\circ) < p_\eta$ **then**
    return $(Z, N, \textsc{Certified})$.
  return $(Z, N, \textsc{BudgetExhausted})$.

Procedure 2 takes $t \in \mathbb{N}$, the discovery tuple D, and π^0. It reads $Z, W, \mathcal{U}, Q^+$, appends the labels (5) only to A^{d}, and returns the updated tuple and a discovered state $x \in \bigcup_{h=2}^{H} \mathcal{S}_h \cup \{\perp\}$, with $x = \perp$ if none is found.

Procedure 2 FIRSTHIT

procedure FIRSTHIT($t \in \mathbb{N}, \mathsf{D}, \pi^0$)
  Start episode t and observe $S_1^{(t)} \sim \mu$.
  Set $\lambda_{S_1^{(t)}} \leftarrow$ ROOTDESIGN($S_1^{(t)}$) by Procedure 3 and draw $A_1^{(t)} \sim \lambda_{S_1^{(t)}}$.
  Execute $A_1^{(t)}$ and observe $(R_1^{(t)}, S_2^{(t)})$.
  Follow π^0 until reaching $Z \cup \mathcal{G}(Z, W)$ or termination.
  while the reached state is $z \in Z$ **do**
    Execute $a_z = \min \arg\max_{a \in \mathcal{A}} Q^+(z, a)$ and observe its reward and next state.
    From that next state, follow π^0 to the next event.
  For each visited $u \in \mathcal{U}$ before novelty or termination, append its observed first-event label (5) to $\mathsf{A}^{\mathrm{d}}(u)$.
  Set x to the reached state if it is in $\mathcal{G}(Z, W)$, and set $x \leftarrow \perp$ otherwise.
  if $x \neq \perp$ **then**
    Complete episode t using the frozen greedy map a_t^+ on Z and π^0 elsewhere; append no further discovery labels.
  return (x, D).

Procedure 3 reads $s \in \mathcal{S}_1$ and its feature menu, returns a distribution in $\Delta(\mathcal{A})$ satisfying (19), and maintains no state between calls.

Procedure 3 ROOTDESIGN

procedure ROOTDESIGN($s \in \mathcal{S}_1$)
  Set $x_a = \phi_1(s, a) \in \mathbb{R}^d$ for every $a \in \mathcal{A}$.
  Compute the deterministic distribution $\lambda_s \in \Delta(\mathcal{A})$ specified in Appendix A.2.
  return λ_s, which satisfies (19).

4.2 RESTRICTED LEARNING PROCEDURE

Procedure 4 takes a fixed finite layer-tagged set Z, integers $N \geq 1, H \geq 2$, scalars $\delta_0 \in (0, 1)$, $B > 0$, and $\kappa \in [0, H]$, and the policy π^0; it returns N randomized Markov policies in $\Pi(Z)$, indexed locally by $t = 1, \ldots, N$. It retains only Z from discovery and starts with empty batches. Appendix A.5 gives the update formulas. Define the controlled-pair set $\mathcal{U}_Z = \{(z, a) : z \in Z, a \in \mathcal{A}\}$ and its size $D_Z = |\mathcal{U}_Z|$.

Segment observations and their target. For $u = (z, a) \in \mathcal{U}_Z$ at layer h, take a and then follow π^0. Let $\tau \in \{h + 1, \ldots, H + 1\}$ be the stopping time of the first later visit to Z or termination. Store the joint outcome

$$O = (R, Y) = \left(\sum_{j=h}^{\tau-1} R_j, S_\tau \right) \in [0, H] \times (Z \cup \{\perp\}).$$

The segment includes the reward for taking a at z and ends before any action at Y. Let P_u^{r} be the probability law of O on this product space; the reward R and endpoint Y may be dependent. Stopping at Z makes the endpoint set finite even when original transitions have infinitely many successors. For $u \in \mathcal{U}_Z$, the active batch $\mathsf{A}_t^{\mathrm{r}}(u)$ and frozen batch $\mathsf{F}_t^{\mathrm{r}}(u)$ are finite sequences of joint outcomes. New outcomes enter the active batch, whose target size is $m_t^{\mathrm{r}}(u) \in \mathbb{N}$; the frozen batch is the last completed batch.

Write $V^t : Z \cup \{\perp\} \to [0, H]$ for the continuation value table, with $V^t(\perp) = 0$. Evaluate each outcome from (z, a) by adding the segment reward to the continuation estimate at its endpoint. The conditional-mean target on a fresh segment is $q_t^{\mathrm{r}} : \mathcal{U}_Z \to [0, 2H]$,

$$q_t^{\mathrm{r}}(z, a) = \int \{r + V^t(y)\} P_{(z,a)}^{\mathrm{r}}(dr, dy).$$

Optimistic values and actions. The frozen batch $\mathsf{F}_t^{\mathrm{r}}(z, a)$ supplies the outcomes used in (31) to compute the optimistic table $Q^t : \mathcal{U}_Z \to [0, H]$; when this batch is empty, $Q^t(z, a) = H$. The active

batch $\mathsf{A}_t^{\mathrm{r}}$ does not affect Q^t. The value table and greedy tail policy $\rho_t : Z \to \mathcal{A}$ are given by

$$V^t(z) = \max_{a \in \mathcal{A}} Q^t(z, a), \qquad \rho_t(z) = \min \arg\max_{a \in \mathcal{A}} Q^t(z, a).$$

Thus $Q^t(z, a)$ estimates the remaining reward. The Markov property makes the joint reward–endpoint law depend only on (z, a). Because every endpoint lies in a later layer, backward induction computes Q^t, V^t, ρ_t. This tabular recursion over Z does not require Bellman completeness.

Learning the initial decision. The root prediction state is the matrix–vector pair $\mathsf{P}_t = (\Sigma_t, b_t) \in \mathbb{R}^{d \times d} \times \mathbb{R}^d$ in (27). The root learner uses the initial state's features with SquareCB (33) if $A \le d$, or OFUL (34) otherwise. It updates from the chosen feature and the complete episode return, so learning at different initial states shares one predictor. All-policy realizability is used only for this root predictor in restricted learning.

Procedure 4 RESTRICTEDLEARN

procedure RESTRICTEDLEARN($Z, N, \delta_0, B, H, \kappa, \pi^0$)
  For every $u \in \mathcal{U}_Z$, set $(\mathsf{A}_1^{\mathrm{r}}(u), \mathsf{F}_1^{\mathrm{r}}(u), m_1^{\mathrm{r}}(u)) \leftarrow (\varnothing, \varnothing, 1)$; initialize P_1 by (27)–(30).
  for $t = 1, \dots, N$ **do**
    From $\mathsf{F}_t^{\mathrm{r}}$ compute Q^t, V^t, and ρ_t by (31).
    Observe $S_1^{(t)}$ and set $x_t(a) = \phi_1(S_1^{(t)}, a) \in \mathbb{R}^d$ for $a \in \mathcal{A}$.
    From P_t and $(x_t(a))_{a \in \mathcal{A}}$, choose $A_1^{(t)}$ by (33) if $A \le d$, and by (34) otherwise.
    Execute $A_1^{(t)}$, use ρ_t on Z, and use π^0 outside Z.
    For every visited $u \in \mathcal{U}_Z$, append its joint segment outcome to $\mathsf{A}_t^{\mathrm{r}}(u)$.
    Set $G_t = \sum_{h=1}^{H} R_h^{(t)} \in [0, H]$.
    Update P_{t+1} from P_t and $(x_t(A_1^{(t)}), G_t)$ by (28).
    After episode t ends, apply (26) to obtain $(\mathsf{F}_{t+1}^{\mathrm{r}}, \mathsf{A}_{t+1}^{\mathrm{r}}, m_{t+1}^{\mathrm{r}})$.
  return $(\pi_t)_{t=1}^N \in \Pi(Z)^N$.

Episode updates. The frozen batches and P_t determine $\pi_t \in \Pi(Z)$ before the initial state is observed: a root-action distribution for each initial state, ρ_t on Z, and π^0 elsewhere. During the episode, only active batches receive new segment outcomes. Afterward, (28) updates P_t, and (26) promotes completed batches while retaining incomplete ones.

5 ANALYSIS

We prove Theorem 3.3 by bounding the stored-set size, the restriction loss, and the regret within the restricted class.

5.1 A SUPPORT COVER AND THE DISCOVERY BUDGET

The state-creation step of Procedure 1 satisfies

$$|Z| \le M = (H-1) L_{\mathrm{d}} m_{\mathrm{d}} = \widetilde{O}((H-1) d^2 K), \tag{6}$$

with integer budgets from Appendix A.1. Appendix B.1 proves the structural and matrix arguments below.

For a Markov continuation policy ρ, define its measurable action-range function $a_h^\rho : \mathcal{S}_h \to [0, H]$ by

$$a_h^\rho(z) = \max_a Q_h^\rho(z, a) - \min_a Q_h^\rho(z, a).$$

This range measures the effect of the current action under fixed continuation.

Lemma 5.1 (Support cover for SIMULTANEOUSDISCOVERY)**.** *For each $h \in \{2, \dots, H\}$, there is a finite set $C_h \subseteq \mathcal{S}_h$ with $|C_h| \le (2d-1)K$ such that, for every Markov continuation policy ρ and predecessor pair $x \in \mathcal{S}_{h-1} \times \mathcal{A}$,*

$$\int_{\mathcal{S}_h \setminus C_h} a_h^\rho(z) P_{h-1}(dz \mid x) \le (2\sqrt{2d} + 2)\kappa + \sqrt{2d} H \upsilon. \tag{7}$$

The same unknown cover C_h works for every ρ and serves only the analysis; the algorithm constructs the stored set Z_h from observations. The cover is the union of concentration sets for at most $2d-1$ predecessor pairs selected by (36).

Procedure 1 uses several matrices because updating a matrix can reduce novelty scores at other important states. Assigning each discovery to one matrix limits updates from C_h to at most $|C_h|$ matrices. Over the episode prefixes in (42), insertions outside C_h with large optimal action range are also rare. Lemma B.5 therefore preserves a matrix updated only at states with small optimal action range. Its updates have small projections along the optimal value coefficient, so omitted states with sufficiently large optimal advantage still pass its novelty test. Each update more than doubles one determinant, allowing at most m_{d} updates per matrix and proving (6).

An exactly realizable stochastic example. Let $H = A = d = K = 2$, $B = 1$, $\mathcal{S}_1 = \{s\}$, and $\mathcal{S}_2 = \{z, y_1, y_2\}$. Action $a \in \{1, 2\}$ at s gives reward zero and reaches each of z, y_a with probability $1/2$. All layer-2 transitions terminate; the only nonzero reward is $r_2(z, 2) = 1$. For the coordinate vectors $e_1, e_2 \in \mathbb{R}^2$, use

$$\phi_1(s, a) = e_1, \quad \phi_2(z, a) = \mathbf{1}\{a = 2\}e_1, \quad \phi_2(y_1, a) = 0, \quad \phi_2(y_2, a) = e_2.$$

Every policy π has exact coefficients $\theta_1^\pi = \frac{1}{2}\pi_2(2 \mid z)e_1$ and $\theta_2^\pi = e_1$, both of norm at most one. The cover $C_2 = \{z\}$ satisfies (7) with zero right side: y_1, y_2 have zero action range, although half of each root transition's mass lies outside C_2.

For these features, a linear transition model is impossible because the equal root features have different next-state laws. Bellman completeness also fails: the max-linear function $f : \mathcal{S}_2 \to [0, 1]$, $f(x) = \max_a \phi_2(x, a)^\top e_2$, has values $f(z) = f(y_1) = 0$ and $f(y_2) = 1$, so its root backups are 0 and $1/2$, which cannot be represented by the equal root features.

5.2 Discovery certificate

Procedure 1 satisfies the episode bound (8) and certifies $J^\star - J_Z^\star \le \eta$. Assume $\eta \le H$ and put $\overline{\Gamma} = \min\{A, 2d\}$, with parameters from Appendix A.1. The episode-count and certificate proofs are in Appendices B.3 and B.4.

Discovery changes its own reachability targets: inserting a state changes segment endpoints and continuation labels. By Lemma B.6, shrinking novelty sets and nonincreasing scores make the labels in (5) decrease pointwise, so earlier confidence bounds remain valid. Lemmas B.11 and B.12 convert the reachability bounds into a restriction-loss certificate, accounting for states stored later.

Theorem 5.2 (Guarantee for SimultaneousDiscovery)**.** *With probability at least $1 - \delta/2$, discovery stores at most M states, every* Certified *output satisfies $J^\star - J_Z^\star \le \eta$, and discovery executes at most*

$$n_{\mathrm{d}} \le \widetilde{O}\left(\frac{\overline{\Gamma} H}{\eta}[H(1 + AM) + M]\right) \tag{8}$$

episodes. This episode bound also holds when the total budget is exhausted; n_{d} is defined in (15).

The episode bound follows from learning novelty reachability with at most $1 + AM$ controlled pairs.

To certify $J^\star - J_Z^\star \le \eta$, consider the comparison policy $\beta_Z \in \Pi(Z)$ that follows $\pi^\star$ at the root and on Z, and π^0 elsewhere. Since $J_Z^\star \ge J(\beta_Z)$, it suffices to bound $J^\star - J(\beta_Z)$. Lemma B.10 first bounds the continuation loss when the root action is drawn from the design distribution and all later actions follow β_Z. The domination inequality (35) transfers this bound to the optimal root action used by β_Z, giving $J^\star - J(\beta_Z) \le \eta$.

5.3 Regret after discovery

Procedure 4 satisfies (9), proved in Appendix B.6 with updates from Appendix A.5. Fix a finite set Z, an integer episode budget $N \ge 1$, and confidence input $\delta_0 \in (0, 1)$. Put $D_Z = A|Z|$ and define $\mathrm{Reg}_{N,Z} = \sum_{t=1}^{N}\{J_Z^\star - J(\pi_t)\}$.

Lemma 5.3 (Regret of RestrictedLearn)**.** *With probability at least $1 - \delta_0$,*

$$\mathrm{Reg}_{N,Z} \le \widetilde{O}\left(H\sqrt{d\min\{A, d\}N} + H\sqrt{\min\{A, d\}(D_Z + 1)N} + \kappa\sqrt{\min\{A, d\}}\,N\right). \tag{9}$$

Root regret and continuation loss. Fix $\theta_Z^\star \in \mathbb{R}^d$ approximating root action values under the optimal restricted tail, and define $f_t : \mathcal{A} \to [-\kappa, H+\kappa]$ by $f_t(a) = \phi_1(S_1^{(t)}, a)^\top \theta_Z^\star$. The root observation (125) satisfies

$$G_t = f_t(A_1^{(t)}) + \nu_t - c_t + e_{Z,t},$$

where, conditional on the history through the root action, $\nu_t \in \mathbb{R}$ is centered noise and $c_t \in [0, H]$ is the expected loss of the learned tail relative to the optimal restricted tail; $e_{Z,t} \in [-\kappa, \kappa]$ is the realizability residual. For the root-action distribution $p_t \in \Delta(\mathcal{A})$, define the root regret as in (121) by

$$\overline{R}_N^{\mathrm{root}} = \sum_{t=1}^{N} \left\{ \max_a f_t(a) - \sum_a p_t(a) f_t(a) \right\}.$$

Let $\bar{c}_t : \mathcal{S}_1 \to [0, H]$ be the continuation loss, obtained by averaging c_t over the root action at each initial state. Equation (123) gives

$$\mathrm{Reg}_{N,Z} \le \overline{R}_N^{\mathrm{root}} + \sum_{t=1}^{N} \bar{c}_t(S_1^{(t)}) + 2\kappa N + \widetilde{O}(H\sqrt{N}).$$

Using the squared-bias estimate (128), the root learners satisfy (132) and (137). Equation (127) bounds the continuation loss. Substituting these bounds and absorbing the remaining terms gives (9).

5.4 Proof of the main theorem

For Algorithm 1, Appendices B.8 and B.7 prove the operation bound and handle $\eta > H$. For $\eta \le H$, discovery and restricted learning conditional on its history each fail with probability at most $\delta/2$; a union bound gives joint success probability at least $1 - \delta$.

Discovery certifies before the budget is exhausted. Put $N_{\mathrm{r}} = T - N_{\mathrm{d}} > 0$. Theorem 5.2 and Lemma 5.3 give $D_Z \le AM$, $N_{\mathrm{d}} \le n_{\mathrm{d}}$, and

$$\mathrm{Reg}_T \le HN_{\mathrm{d}} + N_{\mathrm{r}}(J^\star - J_Z^\star) + \mathrm{Reg}_{N_{\mathrm{r}},Z} \le HN_{\mathrm{d}} + T\eta + \mathrm{Reg}_{N_{\mathrm{r}},Z}.$$

For $\kappa = \upsilon = 0$, balancing the $1/\eta$ discovery bound with the restriction loss $T\eta$ gives $\eta \propto T^{-1/2}$ and square-root regret, up to parameter and logarithmic factors. Equation (16) bounds all three terms.

Discovery uses the entire budget. If $N_{\mathrm{d}} = T$, Theorem 5.2 gives $T \le n_{\mathrm{d}}$, also for certification after episode T. Apply (16) with $N_0 = T$, $n = D_0 = 0$ to $\mathrm{Reg}_T \le HT$, which also gives the theorem's cap. For $A = 1$, all policies coincide and $\mathrm{Reg}_T = 0$.

6 Discussion

Conclusion. Under exact all-policy linear Q-realizability and K-sparse transitions, our polynomial-time online algorithm achieves high-probability regret $\widetilde{O}(dH^2\sqrt{A\min\{A,d\}KT})$, even with infinite state spaces and arbitrary initial-state support.

Statistical and structural gaps. Appendix C leaves open the necessary dependence on K and the cost of positive transition tails. Procedure 1 stores $\widetilde{O}(d^2K)$ states per layer, while Lemma 5.1 guarantees a cover of size $(2d-1)K$. Matching this linear dependence on d in online discovery also remains open.

Beyond soft sparsity. The sparse transition assumption is a limitation of our result, even in its soft form: it still requires concentration on finitely many successors. Even under exact realizability, a fixed positive υ contributes a linear term in (1). Decreasing υ may require a larger K, increasing statistical and computational costs. General kernels need not admit a favorable tradeoff: a nonatomic next-state law assigns zero mass to every finite set, forcing $\upsilon = 1$ in Assumption 3.2.

For general transitions, our proof provides no finite cover certifying small loss outside stored states, while Procedure 4 only generalizes through features at the root. An extension must learn across unseen noninitial states or replace the finite cover with another representation that controls their aggregate contribution to regret. Deriving such a representation with polynomial computation from all-policy realizability remains open.

AI USE STATEMENT

We used generative AI tools to search for related work, polish the writing, and assist in developing and writing proofs of mathematical propositions. The authors take full responsibility for the final content of this paper, including all mathematical claims and proofs.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24, pp. 2312–2320. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/hash/e1d5be1c7f2f456670de3d53c7b54f4a-Abstract.html.

Katy S. Azoury and M. K. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43(3):211–246, 2001. doi: 10.1023/A:1010896012157.

Dylan J. Foster and Alexander Rakhlin. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3199–3210. PMLR, 2020. URL https://proceedings.mlr.press/v119/foster20a.html.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975. doi: 10.1214/aop/1176996452.

Noah Golowich and Ankur Moitra. Linear Bellman completeness suffices for efficient online reinforcement learning with few actions. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 1939–1981. PMLR, 2024. URL https://proceedings.mlr.press/v247/golowich24a.html.

Jiafan He, Dongruo Zhou, Tong Zhang, and Quanquan Gu. Nearly optimal algorithms for linear contextual bandits with adversarial corruptions. In *Advances in Neural Information Processing Systems*, volume 35, pp. 34614–34625. Curran Associates, Inc., 2022. doi: 10.52202/068431-2508.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I. Jordan. Provably efficient reinforcement learning with linear function approximation. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2137–2143. PMLR, 2020. URL https://proceedings.mlr.press/v125/jin20a.html.

Yijing Ke, Zihan Zhang, and Ruosong Wang. Frozen policy iteration: Computationally efficient RL under linear Q^π realizability for deterministic dynamics. *arXiv preprint arXiv:2603.00716*, 2026. doi: 10.48550/arXiv.2603.00716.

J. Kiefer and J. Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366, 1960. doi: 10.4153/CJM-1960-030-4.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009. URL https://www.cs.mcgill.ca/~colt2009/papers/012.pdf.

Zakaria Mhammedi. Sample and oracle efficient reinforcement learning for MDPs with linearly-realizable value functions. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 4078–4165. PMLR, 2025. URL https://proceedings.mlr.press/v291/mhammedi25b.html.

Zakaria Mhammedi, Alexander Rakhlin, and Nneka Okolo. End-to-end efficient RL for linear Bellman complete MDPs with deterministic transitions. *arXiv preprint arXiv:2603.23461*, 2026. doi: 10.48550/arXiv.2603.23461.

Volodya Vovk. Competitive on-line statistics. *International Statistical Review*, 69(2):213–248, 2001. doi: 10.1111/j.1751-5823.2001.tb00457.x.

Gellért Weisz, András György, and Csaba Szepesvári. Online RL in linearly q^π-realizable MDPs is as easy as in linear MDPs if you learn what to ignore. In *Advances in Neural Information Processing Systems*, volume 36, pp. 59172–59205. Curran Associates, Inc., 2023. doi: 10.52202/075280-2583.

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent Bellman error. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10978–10989. PMLR, 2020. URL https://proceedings.mlr.press/v119/zanette20a.html.

Zihan Zhang, Yuxin Chen, Jason Lee, and Simon S. Du. Settling the sample complexity of online reinforcement learning. *Journal of the ACM*, 72(3):22:1–22:63, 2025. doi: 10.1145/3733592.

Runlong Zhou, Zihan Zhang, and Simon S. Du. Sharp variance-dependent bounds in reinforcement learning: Best of both worlds in stochastic and deterministic environments. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 42878–42914. PMLR, 2023. URL https://proceedings.mlr.press/v202/zhou23t.html.

A Algorithmic details

This appendix specifies the parameters and updates of Algorithm 1 and its four procedures.

A.1 Global parameters

Define the scalars $\overline{\Gamma} = \min\{A, 2d\}$ and $c_d = 2\sqrt{2d}+2$. The integer $b_d = 2d-1$ bounds the number of predecessor pairs used to construct each support cover. Let J_{d} be an upper bound on the number of dyadic advantage intervals defined after (38), $\lambda_{\mathrm{d}} > 0$ the matrix regularization, L_{d} the number of matrices per layer, and m_{d} an upper bound on the updates to each matrix:

$$\begin{aligned} J_{\mathrm{d}} &= 2 + \left\lceil \log_2(4\overline{\Gamma}H\sqrt{T}) \right\rceil, & \lambda_{\mathrm{d}} &= \frac{1}{256\overline{\Gamma}^2 B^2 T}, \\ L_{\mathrm{d}} &= 2Kb_d + \left\lceil 8 \log \frac{8(H-1)J_{\mathrm{d}}}{\delta} \right\rceil, & m_{\mathrm{d}} &= 1 + \left\lceil d \log_2 \left(1 + \frac{4T}{d\lambda_{\mathrm{d}}}\right) \right\rceil. \end{aligned} \tag{10}$$

Define the positive integer state and pair caps $M, \overline{D}$, the positive scalar Ξ, and the discovery confidence $\delta_{\mathrm{d}} \in (0,1)$ by

$$M = (H-1)L_{\mathrm{d}}m_{\mathrm{d}}, \qquad \overline{D} = 1 + AM, \qquad \Xi = H\overline{D} + M, \qquad \delta_{\mathrm{d}} = \frac{\delta}{8}.$$

For positive integers D, ℓ, T, a nonnegative integer R, and scalars $p \in (0,1]$, $\delta_0 \in (0,1)$, define the real-valued function L_{reach} and integer-valued function $\mathfrak{N}$ by

$$\begin{aligned} L_{\mathrm{reach}}(D, \ell, T, \delta_0) &= 1 + \log \frac{64(D+1)(\ell+1)(T+1)}{\delta_0}, \\ \mathfrak{N}(D, \ell, R, p, \delta_0, T) &= \left\lceil \frac{2^{46}\{D\ell + R + \log(4/\delta_0)\}}{p} L_{\mathrm{reach}}(D, \ell, T, \delta_0)^{17} \right\rceil. \end{aligned} \tag{11}$$

Define the positive scalar

$$\mathsf{S}_{\mathrm{d}} = 2^{46} \left\{ \Xi + \log \frac{4}{\delta_{\mathrm{d}}} \right\} L_{\mathrm{reach}}(\overline{D}, H, T, \delta_{\mathrm{d}})^{17}. \tag{12}$$

The positive scalars C_κ and C_{d} are

$$\begin{aligned} \mathsf{C}_\kappa &= \max\left\{ 256\overline{\Gamma}H\sqrt{m_{\mathrm{d}}}, \quad \frac{12288\overline{\Gamma}HJ_{\mathrm{d}}c_d\sqrt{m_{\mathrm{d}}}\mathsf{S}_{\mathrm{d}}}{L_{\mathrm{d}}} \right\}, \\ \mathsf{C}_{\mathrm{d}} &= 192\overline{\Gamma}H^2 J_{\mathrm{d}}\mathsf{S}_{\mathrm{d}}. \end{aligned} \tag{13}$$

Define the discovery accuracy $\eta > 0$, stopping threshold $p_\eta > 0$, and analysis threshold $\zeta > 0$ by

$$\eta = \max\left\{\frac{4H}{\sqrt{T}}, \quad \sqrt{\frac{\mathsf{C}_{\mathrm{d}}}{T}}, \quad \mathsf{C}_\kappa \chi\right\}, \qquad p_\eta = \frac{\eta}{64\overline{\Gamma} H J_{\mathrm{d}}}, \qquad \zeta = \frac{\eta}{16\overline{\Gamma} H}. \tag{14}$$

Only p_η enters the stopping test. The threshold ζ and count bound J_{d} are used for the dyadic advantage intervals in Appendix B.1. Here $\chi = \kappa + H\upsilon/2$ is the discovery error scale from Section 3; restricted learning continues to use κ. When $\eta \le H$, $p_\eta \in (0, 1)$; define the deterministic integer

$$n_{\mathrm{d}} = \min\left\{T, \left\lceil \frac{\mathsf{S}_{\mathrm{d}}}{p_\eta} \right\rceil\right\}, \qquad \mathfrak{N}(\overline{D}, H, M, p_\eta, \delta_{\mathrm{d}}, T) = \left\lceil \frac{\mathsf{S}_{\mathrm{d}}}{p_\eta} \right\rceil. \tag{15}$$

The discovery analysis is allocated failure probability $\delta/2$. Its procedure receives $\delta_{\mathrm{d}} = \delta/8$ for the discovery event, and the events in Lemma B.5 use at most $\delta/8$ in total. Restricted learning receives $\delta/2$.

The parameters satisfy the comparison used in Theorem 3.3. When $\eta \le H$, for integers $0 \le n \le T$, $0 \le D_0 \le AM$, and $0 \le N_0 \le n_{\mathrm{d}}$,

$$\begin{aligned} &HN_0 + T\eta + H\sqrt{d\min\{A,d\}n} + H\sqrt{\min\{A,d\}(D_0+1)n} + \kappa\sqrt{\min\{A,d\}\,n} \\ &\le \widetilde{O}\Big(dH\sqrt{A\min\{A,d\}KH(H-1)T} + A\min\{A,d\}H^2(H-1)d^2\chi T\Big). \end{aligned} \tag{16}$$

Appendix B.7 proves (16) and the corresponding bound when $\eta > H$.

A.2 Root design computation

For a realized root state $s \in \mathcal{S}_1$, put $x_a = \phi_1(s, a) \in \mathbb{R}^d$ and define the integer $r_s = \dim \operatorname{span}\{x_a : a \in \mathcal{A}\}$. If $A \le 2d$, use the uniform distribution. Otherwise, for $\lambda \in \Delta(\mathcal{A})$ define $\Sigma_s(\lambda) = \sum_{a\in\mathcal{A}} \lambda(a) x_a x_a^\top$ and let

$$\lambda_s^\star \in \operatorname*{arg\,min}_{\lambda\in\Delta(\mathcal{A}):\operatorname{range}(\Sigma_s(\lambda))=\operatorname{span}\{x_a:a\in\mathcal{A}\}} \max_{a\in\mathcal{A}} x_a^\top \Sigma_s(\lambda)^\dagger x_a. \tag{17}$$

The Kiefer–Wolfowitz equivalence theorem (Kiefer & Wolfowitz, 1960) gives

$$\max_a x_a^\top \Sigma_s(\lambda_s^\star)^\dagger x_a = r_s.$$

If $r_s = 0$, use the uniform distribution. If $r_s = 1$, put all mass on the lexicographically first action maximizing $\|x_a\|_2$; its maximum leverage is one. For $r_s \ge 2$, start from the uniform distribution $\lambda^{(0)}$ and repeat the design update. At iteration j, let

$$g_j = \max_a x_a^\top \Sigma_s(\lambda^{(j)})^\dagger x_a$$

and let a_j be the lexicographically first maximizer. If $g_j \le 2r_s$, return $\lambda_s = \lambda^{(j)}$. Otherwise put

$$\alpha_j = \frac{g_j - r_s}{r_s(g_j - 1)}, \qquad \lambda^{(j+1)} = (1-\alpha_j)\lambda^{(j)} + \alpha_j e_{a_j},$$

where $e_a \in \Delta(\mathcal{A})$ is the point mass at a. All covariances in this recursion have full rank on $\operatorname{span}\{x_a\}$.

The loop performs at most $\lceil 8r_s \log A \rceil$ updates and is measurable, as proved in Lemma B.2. In the branch $A > 2d$ the returned design satisfies

$$\max_a x_a^\top \Sigma_s(\lambda_s)^\dagger x_a \le 2r_s \le 2d. \tag{18}$$

Combining this bound in the branch $A > 2d$ with the uniform-distribution bound in the branch $A \le 2d$ gives

$$\max_a g_a \le \overline{\Gamma}\mathbb{E}_{a\sim\lambda_s} g_a + (\sqrt{\overline{\Gamma}} + 1)e, \quad g_a \ge 0, \quad |g_a - x_a^\top u| \le e \quad (a \in \mathcal{A}), \tag{19}$$

where $u \in \mathbb{R}^d$ and $e \ge 0$; its proof is Lemma B.3. Uniform sampling gives the factor A without an additive term.

A.3 DISCOVERY BATCH UPDATES

Fix a reachability confidence $\delta_0 \in (0,1)$, equal to the confidence input δ_{d} of Procedure 1, and the scalar $\rho = (32H)^{-1}$. The deterministic pair cap used to calibrate all discovery confidence radii is $D_{\max} = \overline{D}$. For a pair $u \in \mathcal{U}$, let $\mathsf{A}^{\mathrm{d}}_t(u)$ and $\mathsf{F}^{\mathrm{d}}_t(u)$ be its active and frozen label sequences at the beginning of episode t, let $m^{\mathrm{d}}_t(u) \in \mathbb{N}$ be the active target size, and let $Q^+_t(u) \in [0,1]$ be its optimistic estimate. On local visit i, let $X_i \in [0,1]$ be the scalar first-hit label and use the predictable forecast $\widehat{X}_i = X_{i-1}$, with $\widehat{X}_1 = 0$. The predecessor of the first label in a new active batch is the last label of the current frozen batch, or zero before the first completion; all other predecessors are in the active batch itself. Thus the maintained batches determine every forecast used by the update. For a completed sequence $B = (X_i)_{i \in I_B}$ with ordered index set $I_B = \{i_1 < \cdots < i_m\}$, set

$$\begin{aligned} \overline{X}(B) &= \frac{1}{m}\sum_{j=1}^{m} X_{i_j}, & A_B &= \sum_{j=1}^{m} (X_{i_j} - \widehat{X}_{i_j})^2, \\ \mathrm{rad}(B) &= c_1 \frac{\sqrt{(A_B + \Lambda)\Lambda}}{m} + c_2 \frac{\Lambda}{m}, & \Lambda &= 8 \log \frac{64 D_{\max}(T+1)}{\delta_0}. \end{aligned} \tag{20}$$

where $c_1 = 8$ and $c_2 = 32$. Define the function $g_{\mathrm{d}} : \mathbb{N} \to \mathbb{N}$ for the next target size to be $2m$ when $m < 1/\rho$ and $\lceil (1+\rho)m \rceil$ otherwise. Let $\widetilde{\mathsf{A}}^{\mathrm{d}}_t(u)$ be the active sequence after episode t has appended its label if u was visited. At initialization, $\mathsf{F}^{\mathrm{d}}_1(u_\circ) = \mathsf{A}^{\mathrm{d}}_1(u_\circ) = \varnothing$, $m^{\mathrm{d}}_1(u_\circ) = 1$, and $Q^+_1(u_\circ) = 1$. After episode t ends, the complete transition for each $u \in \mathcal{U}_t$ is

$$\begin{aligned} &\big(\mathsf{F}^{\mathrm{d}}_{t+1}(u), \mathsf{A}^{\mathrm{d}}_{t+1}(u), m^{\mathrm{d}}_{t+1}(u), Q^+_{t+1}(u)\big) \\ &\quad = \begin{cases} \Big(\widetilde{\mathsf{A}}^{\mathrm{d}}_t(u), \varnothing, g_{\mathrm{d}}(m^{\mathrm{d}}_t(u)), \min\Big\{Q^+_t(u), \overline{X}(\widetilde{\mathsf{A}}^{\mathrm{d}}_t(u)) + \mathrm{rad}(\widetilde{\mathsf{A}}^{\mathrm{d}}_t(u))\Big\}\Big), & |\widetilde{\mathsf{A}}^{\mathrm{d}}_t(u)| = m^{\mathrm{d}}_t(u), \\ \Big(\mathsf{F}^{\mathrm{d}}_t(u), \widetilde{\mathsf{A}}^{\mathrm{d}}_t(u), m^{\mathrm{d}}_t(u), Q^+_t(u)\Big), & |\widetilde{\mathsf{A}}^{\mathrm{d}}_t(u)| < m^{\mathrm{d}}_t(u). \end{cases} \end{aligned} \tag{21}$$

For every pair $u \in \mathcal{U}_{t+1} \setminus \mathcal{U}_t$ created after episode t, set

$$(\mathsf{F}^{\mathrm{d}}_{t+1}(u), \mathsf{A}^{\mathrm{d}}_{t+1}(u), m^{\mathrm{d}}_{t+1}(u), Q^+_{t+1}(u)) = (\varnothing, \varnothing, 1, 1).$$

Then set $V^+_{t+1}(z) = \max_{a \in \mathcal{A}} Q^+_{t+1}(z,a)$ for $z \in Z_{t+1}$. Thus an existing score can change only when a completed active batch becomes the new frozen sequence. The updated score is the minimum of its previous value and the empirical upper bound computed from that frozen sequence; the frozen sequence is replaced rather than augmented.

The primitive continuation law and the measurable first-hit label are defined in (4)–(5). The observed stopped prefix determines the label. Given its stopping state, extend that prefix by an unobserved default-policy suffix drawn from the corresponding Markov continuation law, conditionally independently of the preceding history. This gives the coupling $F^u_t(\Omega^u_t)$ for a fresh $\Omega^u_t \sim P_u$ described in the main algorithm section; no unobserved variable is used by the procedure. Updates and promotions occur after the episode. When a state is discovered, all its action pairs are added to the controlled set and receive the displayed initial batches, target, and score. Each new pair's confidence process starts strictly after the stopping time at which the pair is created; existing pairs retain their batches and scores. Procedure 4 initializes all final controlled pairs independently when restricted learning starts.

The proof allocates $\delta_0/4$ to each of: batch confidence, batch-mean conversion, variance conversion, and novelty-count bound. The two conversion families are internally split between their two applications, as specified in Appendix B.3. Appendix B.3 proves the episode bound (11) with these allocations.

A.4 FINITE-CONTROL MODEL AND ESTIMATOR

This subsection defines the estimator used by Procedure 4; its concentration and regret guarantees are proved in Appendix B.5.

Fix an integer budget $N \geq 1$, a confidence level $\delta \in (0,1)$, and let $\mathcal{U}$ be the $D = |\mathcal{U}| \geq 1$ feasible nonterminal state–action pairs of a finite directed acyclic graph. An outcome from u has probability law P_u on the measurable outcome space $\mathcal{O}_u$ and is a joint pair $O = (R, Y)$, where Y is later in a

fixed topological order or terminal and $R \in [0, H]$ is the segment reward; R and Y may be correlated. A pair occurs at most once on a trajectory, and the segment rewards on any trajectory sum to at most H.

Maintain disjoint doubling batches of sizes $1, 2, 4, \ldots$ for every pair. At the start of episode k, process batches completed in episode $k-1$, recompute once in reverse topological order, and freeze the value function V^k from graph states to $[0, H]$ and its deterministic greedy policy ρ_k, which maps nonterminal states to feasible actions, using a fixed tie-breaking rule. A frozen batch of size n is reused for at most $2n$ visits. Fix a scalar range $\mathcal{R} \geq 2H$. For a measurable target function $x : \mathcal{O}_u \to [0, \mathcal{R}]$, let

$$\widehat{E}_u x = \frac{1}{n}\sum_{j=1}^{n} x(O_j), \qquad \widehat{\mathrm{Var}}_u(x) = \frac{1}{n}\sum_{j=1}^{n}\{x(O_j) - \widehat{E}_u x\}^2$$

be its empirical mean and biased empirical variance on the frozen batch $(O_1, \ldots, O_n)$; in particular, the variance is zero when $n = 1$. Write $E_u x = \int x(o) P_u(do)$ for the population mean and define

$$f_{u,n}(x) = \widehat{E}_u x + \max\left\{\frac{20}{3}\sqrt{\frac{\widehat{\mathrm{Var}}_u(x)\ell}{n}}, \frac{400\mathcal{R}\ell}{9n}\right\}, \qquad \ell = 64\log\frac{8D(2+\lceil\log_2 N\rceil)}{\delta}. \tag{22}$$

Let $\mathcal{A}(x) = \{a : (x, a) \in \mathcal{U}\}$ be the feasible action set at state x. On a pair with a frozen batch of size n, the reverse dynamic program uses

$$X_u^k(O) = R + V^k(Y), \qquad Q^k(u) = \min\{f_{u,n}(X_u^k), H\}, \qquad V^k(x) = \max_{a\in\mathcal{A}(x)} Q^k(x, a). \tag{23}$$

Every nonterminal state has at least one feasible action, and every terminal value is zero, both for the learner and for policy evaluation. Pairs without a completed batch have value H. Here V^π denotes the expected remaining segment reward under a policy π. Consequently $0 \leq V^k \leq H$, and the total-return assumption gives $0 \leq V^\pi \leq H$. Entry at a terminal state is allowed and gives zero value and no segment observations.

A.5 Restricted-learning parameters and updates

Fix a budget $N \in \mathbb{N}$, a confidence level $\delta_0 \in (0, 1)$, known bounds $B > 0$ and $\kappa \in [0, H]$, and the integer $D = D_Z$. Allocate

$$\delta_{\mathrm{sq}} = \delta_{L_2} = \delta_{\mathrm{root}} = \delta_{\mathrm{pred}} = \delta_{\mathrm{fin}} = \delta_{\mathrm{l1}} = \delta_{\mathrm{full}} = \delta_{\mathrm{reg}} = \delta_0/10.$$

For $D \geq 1$, define the explicit deterministic bound

$$\begin{aligned} L_{\mathrm{sq}}(D, N, \delta_{\mathrm{sq}}) &= 1 + \log\frac{64(D+1)(N+1)}{\delta_{\mathrm{sq}}}, \\ B_2(D, N, H, \delta_{\mathrm{sq}}) &= H^2 \min\left\{N,\ 1 + 2^{26} D L_{\mathrm{sq}}(D, N, \delta_{\mathrm{sq}})^4\right\}. \end{aligned} \tag{24}$$

This is the budget for the sum of squared continuation gaps; its guarantee is proved in Corollary B.18. Define the associated conditional-loss budget by

$$\mathcal{B}_{\mathrm{cont}} = \min\left\{NH^2,\ 2B_2 + 2H^2\log(1/\delta_{L_2})\right\}; \tag{25}$$

when $D = 0$, set $B_2 = \mathcal{B}_{\mathrm{cont}} = 0$. Let $m_t^{\mathrm{r}}(u) \in \mathbb{N}$ be the target size of $\mathsf{A}_t^{\mathrm{r}}(u)$ and let $\widetilde{\mathsf{A}}_t^{\mathrm{r}}(u)$ be that batch after episode t has appended its outcome, if u was visited. Initially $\mathsf{F}_1^{\mathrm{r}}(u) = \mathsf{A}_1^{\mathrm{r}}(u) = \varnothing$ and $m_1^{\mathrm{r}}(u) = 1$. After episode t ends, the transition is

$$\begin{aligned} &\left(\mathsf{F}_{t+1}^{\mathrm{r}}(u), \mathsf{A}_{t+1}^{\mathrm{r}}(u), m_{t+1}^{\mathrm{r}}(u)\right) \\ &\quad = \begin{cases} \left(\widetilde{\mathsf{A}}_t^{\mathrm{r}}(u), \varnothing, 2m_t^{\mathrm{r}}(u)\right), & |\widetilde{\mathsf{A}}_t^{\mathrm{r}}(u)| = m_t^{\mathrm{r}}(u), \\ \left(\mathsf{F}_t^{\mathrm{r}}(u), \widetilde{\mathsf{A}}_t^{\mathrm{r}}(u), m_t^{\mathrm{r}}(u)\right), & |\widetilde{\mathsf{A}}_t^{\mathrm{r}}(u)| < m_t^{\mathrm{r}}(u). \end{cases} \end{aligned} \tag{26}$$

For episode i, write $x_i = \phi_1(S_1^{(i)}, A_1^{(i)}) \in \mathbb{R}^d$ and $G_i = \sum_{h=1}^{H} R_h^{(i)} \in [0, H]$ as in Procedure 4. Set the scalar $\lambda = H^2/B^2$ and

$$\Sigma_t = \lambda I + \sum_{i<t} x_i x_i^\top, \qquad b_t = \sum_{i<t} x_i G_i, \qquad \mathsf{P}_t = (\Sigma_t, b_t), \qquad \widehat{\theta}_t = \Sigma_t^{-1} b_t. \tag{27}$$

After episode t, this state changes by

$$\mathsf{P}_{t+1} = \big(\Sigma_t + x_t x_t^\top,\ b_t + x_t G_t\big). \tag{28}$$

The observable OFUL radius is

$$\beta_t = \frac{H}{2}\sqrt{2\log(1/\delta_{\mathrm{root}}) + \log\frac{\det\Sigma_t}{\det(\lambda I)}} + \sqrt{\lambda}B + \sqrt{\mathcal{B}_{\mathrm{cont}}} + \kappa\sqrt{t-1}. \tag{29}$$

For the SquareCB branch, define the square-loss budget $R_{\mathrm{pred}} > 0$, prediction-error budget $U_{\mathrm{scb}} > 0$, and sampling parameter $\gamma > 0$ by

$$R_{\mathrm{pred}} = H^2 + H^2 d\log\left(1 + \frac{NB^2}{dH^2}\right),$$
$$U_{\mathrm{scb}} = 4R_{\mathrm{pred}} + 32\mathcal{B}_{\mathrm{cont}} + 32N\kappa^2 + 16H^2\log(1/\delta_{\mathrm{pred}}), \qquad \gamma = \sqrt{\frac{8AN}{U_{\mathrm{scb}}}}. \tag{30}$$

Use the envelope $f_{u,n}$ from Appendix A.4, with $\mathcal{U} = \mathcal{U}_Z$, joint outcome law $P_u = P_u^{\mathrm{r}}$, $\mathcal{R} = 2H$, and $\delta = \delta_{\mathrm{fin}}$. If $D = 0$, no envelope is evaluated. Its analysis is in Appendix B.5. For each $u = (z,a) \in \mathcal{U}_Z$, put $n_t(u) = |\mathsf{F}_t^{\mathrm{r}}(u)|$. In reverse layer order, define

$$\begin{aligned} X_{z,a}^t(r,y) &= r + V^t(y), \\ Q^t(z,a) &= \begin{cases} \min\{f_{(z,a),n_t(z,a)}(X_{z,a}^t), H\}, & \mathsf{F}_t^{\mathrm{r}}(z,a) \neq \varnothing, \\ H, & \mathsf{F}_t^{\mathrm{r}}(z,a) = \varnothing, \end{cases} \\ V^t(z) &= \max_{a\in\mathcal{A}} Q^t(z,a), \qquad \rho_t(z) = \min\arg\max_{a\in\mathcal{A}} Q^t(z,a), \end{aligned} \tag{31}$$

where $V^t(\perp) = 0$ for the terminal endpoint $\perp$ and the envelope $f_{u,n}$ is defined in (22). If $Z = \varnothing$, the functions Q^t and ρ_t have empty domains, while $V^t(\perp) = 0$.

For scalars $l \le r$ and $y \in \mathbb{R}$, define $\mathrm{clip}_{[l,r]}(y) = \min\{r, \max\{l, y\}\}$. For the SquareCB branch, define the clipped Vovk–Azoury–Warmuth prediction for each $a \in \mathcal{A}$ by

$$\widehat{f}_t(a) = \mathrm{clip}_{[0,H]}\left(x_t(a)^\top\left[\Sigma_t + x_t(a)x_t(a)^\top\right]^{-1} b_t\right). \tag{32}$$

Let $a_t^0 = \min\arg\max_{a\in\mathcal{A}} \widehat{f}_t(a)$ and define $p_t \in \Delta(\mathcal{A})$ by

$$p_t(a) = \frac{1}{A + \gamma(\widehat{f}_t(a_t^0) - \widehat{f}_t(a))} \quad (a \neq a_t^0), \qquad p_t(a_t^0) = 1 - \sum_{a\neq a_t^0} p_t(a). \tag{33}$$

The procedure draws $A_1^{(t)} \sim p_t$. In the OFUL branch it uses the lexicographically first maximizer

$$A_1^{(t)} = \min\arg\max_{a\in\mathcal{A}}\left\{x_t(a)^\top\widehat{\theta}_t + \beta_t\|x_t(a)\|_{\Sigma_t^{-1}}\right\}. \tag{34}$$

After observing G_t, both root predictors append the typed observation $(x_t, G_t) = (x_t(A_1^{(t)}), G_t) \in \mathbb{R}^d \times [0,H]$; equivalently, (28) forms P_{t+1} from all samples with $1 \le i \le t$. Episode $t+1$ uses this updated state in (32) or (34).

B Proofs

This appendix contains the concentration, finite-state, and regret arguments used in the main analysis. Each subsection states the precise claim it proves and cites the algorithmic definitions from Appendix A.

B.1 Structural lemmas for discovery

This subsection analyzes Procedures 1 and 3. It proves the approximate domination inequality (35), the support-cover bound in Lemma 5.1, the matrix update and novelty-inclusion claims of Lemma B.5, and monotonicity of the discovery labels. These are the structural inputs to the later discovery-certificate proof. All parameters are specified in Appendix A.1.

Lemma B.1 (Deterministic policies suffice for all-policy realizability)**.** *Under the model of Section 3, Assumption 3.1 implies the same uniform approximation guarantee for every randomized Markov policy, with unchanged coefficient bound B and approximation error κ.*

Proof. Fix a randomized Markov policy π and a layer h. Let $U = (U_1, \dots, U_H)$ have independent uniform coordinates on $[0, 1]$. For each seed vector u, define the deterministic measurable policy π^u by applying the inverse cumulative distribution of $\pi_j(\cdot \mid s)$ to u_j at layer j; define the action at endpoint seeds by the fixed tie-breaking rule. Finite actions make this map jointly measurable in s, u. Backward integration against the Markov kernels makes $u \mapsto Q_h^{\pi^u}(x)$ measurable and bounded for every pair x. A trajectory uses each layer seed at most once, so successive conditioning on the unused seeds gives, for every pair x at layer h,

$$Q_h^{\pi}(x) = \mathbb{E}_U Q_h^{\pi^U}(x).$$

This identity concerns values, and makes no choice of a random coefficient vector.

For a finite set $E = \{x_1, \dots, x_m\}$ of layer-h pairs, let $\Phi_E \in \mathbb{R}^{m \times d}$ have rows $\phi_h(x_i)^\top$. The compact convex set

$$\mathcal{K}_E = \{\Phi_E \theta + e : \theta \in \mathbb{R}^d,\ \|\theta\|_2 \le B,\ e \in [-\kappa, \kappa]^m\}$$

contains $(Q_h^{\pi^u}(x_i))_{i=1}^m$ for every seed u, by Assumption 3.1. It therefore contains the expectation $(Q_h^{\pi}(x_i))_{i=1}^m$. Consequently every finite collection of the closed constraints

$$\|\theta\|_2 \le B, \qquad |Q_h^{\pi}(x) - \phi_h(x)^\top \theta| \le \kappa$$

has a solution. Compactness of the Euclidean ball of radius B and the finite-intersection property give one vector satisfying every constraint, even when the pair space is uncountable. This proves the assertion. □

Lemma B.2 (Computation of RootDesign)**.** *The iteration in Appendix A.2 terminates within $\lceil 8 r_s \log A \rceil$ updates and returns a measurable distribution using polynomially many real-arithmetic operations.*

Proof. If $A \le 2d$ or $r_s \le 1$, the procedure returns immediately. Otherwise take determinants on the r_s-dimensional feature span. The matrix determinant lemma gives

$$\log \frac{\det \Sigma_s(\lambda^{(j+1)})}{\det \Sigma_s(\lambda^{(j)})} = (r_s - 1) \log \frac{g_j (r_s - 1)}{r_s (g_j - 1)} + \log \frac{g_j}{r_s} \ge \frac{1}{8}$$

whenever $g_j \ge 2 r_s$. For $r = r_s \ge 2$, let $h_r(g)$ denote the displayed log ratio. Its derivative is $h_r'(g) = (g - r)/\{g(g-1)\} \ge 0$ for $g \ge 2r$. At $g = 2r$,

$$h_r(2r) = \log 2 + (r - 1) \log \left(1 - \frac{1}{2r - 1}\right) \ge \log 2 - \frac{1}{2} > \frac{1}{8},$$

where $\log(1 - x) \ge -x/(1 - x)$ for $0 \le x < 1$ follows by integrating $(1 - t)^{-1} \le (1 - x)^{-1}$ over $0 \le t \le x$. The compactness of $\Delta(\mathcal{A})$ and positivity of the uniform design on the r_s-dimensional span give a maximum-determinant design λ_s^{D}. Moreover, $\Sigma_s(\lambda^{(0)}) \succeq A^{-1} \Sigma_s(\lambda_s^{\mathrm{D}})$, so the initial log-determinant gap is at most $r_s \log A$. Consequently the loop performs at most $\lceil 8 r_s \log A \rceil$ updates. It uses only finite linear algebra, arithmetic, and comparisons. Hence the fixed tie-breaking makes $s \mapsto \lambda_s$ measurable and gives a number of real-arithmetic operations polynomial in A and d. □

Lemma B.3 (Approximate pointwise root domination)**.** *Let $s \in \mathcal{S}_1$, $u \in \mathbb{R}^d$, and $e \ge 0$. For any nonnegative real numbers $(g_a)_{a \in \mathcal{A}}$ satisfying $|g_a - \phi_1(s, a)^\top u| \le e$, the distribution returned by Procedure 3 satisfies*

$$\max_{a \in \mathcal{A}} g_a \le \overline{\Gamma} \mathbb{E}_{a \sim \lambda_s} g_a + (\sqrt{\overline{\Gamma}} + 1) e. \tag{35}$$

Proof. If $A \le 2d$, uniform sampling gives the multiplicative factor A without an additive term. Otherwise put $f_a = \phi_1(s, a)^\top u$, $F = \max_a |f_a|$, and $\overline{g} = \mathbb{E}_{a \sim \lambda_s} g_a$. Since $f_a \ge -e$, the approximation guarantee implies

$$f_a^2 \le e^2 + (F + e) g_a.$$

Indeed, if $f_a \le e$ then $|f_a| \le e$; if $f_a > e$ then $f_a^2 - e^2 = (f_a - e)(f_a + e) \le g_a(F + e)$. Put $x_a = \phi_1(s, a) \in \mathbb{R}^d$ and $\Sigma = \sum_a \lambda_s(a) x_a x_a^\top$. The construction in Appendix A.2 ensures $\mathrm{range}(\Sigma) = \mathrm{span}\{x_a : a \in \mathcal{A}\}$, including the rank-zero and rank-one branches. Thus $\Sigma^{\dagger/2}\Sigma^{1/2}x_a = x_a$, where $\Sigma^{\dagger/2}$ is the positive-semidefinite square root of $\Sigma^\dagger$. For every action, Euclidean Cauchy–Schwarz gives

$$\begin{aligned} f_a^2 &= \left\langle \Sigma^{\dagger/2} x_a, \Sigma^{1/2} u \right\rangle^2 \\ &\le (x_a^\top \Sigma^\dagger x_a)(u^\top \Sigma u) \le \overline{\Gamma} \sum_b \lambda_s(b) f_b^2. \end{aligned}$$

The last inequality uses (18) and $\overline{\Gamma} = 2d$ in this branch. Taking the maximum over a, then averaging the preceding pointwise bound on f_a^2, gives

$$F^2 \le \overline{\Gamma} \mathbb{E}_{a \sim \lambda_s} f_a^2 \le \overline{\Gamma}\{e^2 + (F + e)\overline{g}\}.$$

If $F \le \sqrt{\overline{\Gamma}}\, e$, then $F \le \overline{\Gamma}\overline{g} + \sqrt{\overline{\Gamma}}\, e$ immediately. Otherwise $F + e > 0$, and factoring the last display gives

$$F - \sqrt{\overline{\Gamma}}\, e \le \overline{\Gamma}\overline{g}\, \frac{F + e}{F + \sqrt{\overline{\Gamma}}\, e} \le \overline{\Gamma}\overline{g},$$

because $\overline{\Gamma} \ge 1$. Combining these two cases with $\max_a g_a \le F + e$ proves (35), including $F = e = 0$. □

Proof of Lemma 5.1. Fix $h \in \{2, \dots, H\}$ and let $r \le d$ be the dimension of the span of the predecessor feature vectors. For $r > 0$, fix orthonormal coordinates on this span and write $v(x) \in \mathbb{R}^r$ for the feature vector of a predecessor pair x. We first construct predecessor pairs $x_1, \dots, x_n$, with repetitions allowed and $n = 2r - 1$, whose empirical covariance satisfies

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} v(x_i) v(x_i)^\top \succ 0, \qquad \sup_x v(x)^\top \widehat{\Sigma}^{-1} v(x) \le 2r. \tag{36}$$

The closure $\mathcal{X} \subseteq \mathbb{R}^r$ of the predecessor feature set is compact and spans $\mathbb{R}^r$. Choose a tuple $(v_1^\star, \dots, v_n^\star) \in \mathcal{X}^n$ maximizing $\det M_\star$, where $M_\star = \sum_{i=1}^n v_i^\star (v_i^\star)^\top \in \mathbb{R}^{r \times r}$. Compactness gives a maximizer, and a spanning tuple exists because $n \ge r$, so $M_\star \succ 0$. For any $v \in \mathcal{X}$, define the scalars

$$a_i = (v_i^\star)^\top M_\star^{-1} v_i^\star, \qquad b = v^\top M_\star^{-1} v, \qquad c_i = (v_i^\star)^\top M_\star^{-1} v.$$

Replacing the ith tuple entry by v and applying the rank-two determinant identity gives, for each $i \in [n]$,

$$(1 - a_i)(1 + b) + c_i^2 = \frac{\det\{M_\star - v_i^\star (v_i^\star)^\top + v v^\top\}}{\det M_\star} \le 1.$$

To see the identity even when removal makes the matrix singular, apply $\det(I + UV^\top) = \det(I + V^\top U)$ with $U = M_\star^{-1}[-v_i^\star, v]$ and $V = [v_i^\star, v]$. The resulting two-by-two matrix is $\left(\begin{smallmatrix} 1 - a_i & c_i \\ -c_i & 1 + b \end{smallmatrix}\right)$. Furthermore,

$$\sum_i a_i = \mathrm{tr}(M_\star^{-1} M_\star) = r, \qquad \sum_i c_i^2 = v^\top M_\star^{-1} M_\star M_\star^{-1} v = b.$$

Summing the determinant inequalities gives $(n - r)(1 + b) + b \le n$, or $(n - r + 1)b \le r$. The positive-definite covariance $\Sigma_\star = M_\star / n$ therefore satisfies

$$\sup_{v \in \mathcal{X}} v^\top \Sigma_\star^{-1} v \le \frac{nr}{n - r + 1} = 2r - 1.$$

Density of the actual feature set in $\mathcal{X}$ and continuity of the finite covariance sum give actual predecessor pairs $x_1, \dots, x_n$ such that

$$\|\widehat{\Sigma} - \Sigma_\star\|_{\mathrm{op}} \le \frac{\lambda_{\min}(\Sigma_\star)}{2r}, \qquad \widehat{\Sigma} \succeq \left(1 - \frac{1}{2r}\right) \Sigma_\star.$$

Consequently the leverage is at most $(2r - 1)/(1 - 1/(2r)) = 2r$, proving (36). This construction also covers $r = 1$, where $n = 1$ and the covariance approximation loses at most a factor two. For each

of these finitely many predecessor pairs, choose a concentration set $\mathcal{N}_{h-1}(x_i)$ from Assumption 3.2, and put $C_h = \bigcup_{i=1}^{n} \mathcal{N}_{h-1}(x_i)$. This finite set is measurable because the state space is standard Borel, and $|C_h| \le nK \le b_d K$. Only finitely many sets are selected; no measurable selection over the predecessor space is required. For $r = 0$, put $C_h = \varnothing$.

Fix any Markov continuation policy ρ, and define its measurable full action-range function

$$a_h^\rho(z) = \max_{a \in \mathcal{A}} Q_h^\rho(z, a) - \min_{a \in \mathcal{A}} Q_h^\rho(z, a) \in [0, H].$$

Construct two policies that agree with ρ after layer h, agree with each other on C_h, and outside C_h choose respectively a maximizing and a minimizing action for Q_h^ρ. Fixed finite-action tie-breaking makes both policies measurable. Their predecessor action-value difference is the nonnegative function

$$g(x) = \int_{\mathcal{S}_h \setminus C_h} a_h^\rho(z) P_{h-1}(dz \mid x).$$

This identity cancels the common expected current reward and does not require reward–next-state independence. For $r > 0$, Assumption 3.1 and Lemma B.1 supply a vector $u \in \mathbb{R}^r$ with $|g(x) - v(x)^\top u| \le 2\kappa$ for every predecessor pair x. For each selected predecessor, the probability outside C_h is at most υ. Since $a_h^\rho \le H$, we have $0 \le g(x_i) \le H\upsilon$, and hence $|v(x_i)^\top u| \le 2\kappa + H\upsilon$. Consequently $u^\top \widehat{\Sigma} u \le (2\kappa + H\upsilon)^2$. For every predecessor x,

$$\begin{aligned} g(x) &\le |v(x)^\top u| + 2\kappa \\ &\le \sqrt{v(x)^\top \widehat{\Sigma}^{-1} v(x)} \sqrt{u^\top \widehat{\Sigma} u} + 2\kappa \le \sqrt{2r}(2\kappa + H\upsilon) + 2\kappa. \end{aligned}$$

Here the middle inequality applies Euclidean Cauchy–Schwarz to $\widehat{\Sigma}^{-1/2} v(x)$ and $\widehat{\Sigma}^{1/2} u$. Taking the supremum yields

$$\sup_x \int_{\mathcal{S}_h \setminus C_h} a_h^\rho(z) P_{h-1}(dz \mid x) \le c_d \kappa + \sqrt{2d} H \upsilon \le c_d \chi. \tag{37}$$

For $r = 0$, the feature term vanishes and the bound is at most 2κ. The cover depends only on the predecessor feature map and the selected transition concentration sets, so the same cover works for every continuation policy. It is an analysis object and need not be computed by the learner. All integrals use the original transition kernels, including their unrestricted residual mass. □

Remark B.4 (Why the support bound uses action ranges)**.** For $H = 2$, $A = 2$, $d = 1$, $B > 0$, and $\upsilon \in (0, 1)$, put $b = \min\{1, B\}$. There is one root state s, where both actions pay zero and have the same next-state law: probability $1 - \upsilon$ on z_0 and probability $\upsilon 2^{-i}$ on z_i for every integer $i \ge 1$. At z_0, both rewards vanish; at each z_i, $i \ge 1$, actions 1 and 2 pay respectively zero and b. Set $\phi_1(s, a) = 1$ and $\phi_2(z_i, a) = \mathbf{1}\{i \ge 1, a = 2\}$. For every Markov policy π, the coefficients $\theta_1^\pi = b\upsilon \sum_{i \ge 1} 2^{-i} \pi_2(2 \mid z_i)$ and $\theta_2^\pi = b$ have norm at most B and represent its action values exactly. The concentration set $\{z_0\}$ verifies Assumption 3.2 with $K = 1$. Thus infinitely many states can have action range $b > 0$ even when $\kappa = 0$. Their aggregate expected action range is $b\upsilon$, illustrating why (37) controls value-weighted mass rather than the number of such states.

For the next discovery lemmas, suppose $\eta \le H$, let $C_h \subseteq \mathcal{S}_h$ be the fixed support cover from Lemma 5.1, and define the measurable optimal default-advantage function $\alpha_h^0 : \mathcal{S}_h \to [0, H]$ by

$$\alpha_h^0(z) = V_h^\star(z) - Q_h^\star(z, \pi_h^0(z)).$$

The number $J \in \mathbb{N}$ of dyadic advantage intervals, positive scales r_j, w_j, p_j, ν_j, and integer prefix cap n_j below are analysis objects; the procedure does not evaluate them:

$$\begin{gathered} J = 1 + \lfloor \log_2(H/\zeta) \rfloor, \qquad r_j = 2^j \zeta, \quad w_j = \min\{1, 2r_j\}, \quad 0 \le j < J, \\ p_j = \frac{\eta}{16 \overline{\Gamma} H J_{\mathrm{d}} w_j}, \qquad \nu_j = \frac{r_j}{8\sqrt{m_{\mathrm{d}}}}, \\ n_j = \min\left\{ T, \left\lceil \frac{\mathsf{S}_{\mathrm{d}}}{p_j} \right\rceil \right\}. \end{gathered} \tag{38}$$

For $j \in \{0, \dots, J-1\}$, the *jth dyadic advantage interval* is $[r_j, 2r_j)$. States are grouped according to the interval containing their default advantage $\alpha_h^0(z)$. Since $r_{J-1} \le H < 2r_{J-1}$, these intervals

partition $[\zeta, H]$ after intersection with $[0, H]$, including the endpoint H. The integer n_j limits the number of completed discovery episodes considered for the jth dyadic advantage interval.

For $n \in \{1, \dots, T\}$, let (Z_{n+1}, W_{n+1}) denote the stored set and matrix array after all updates for discovery episode n, including state insertion and matrix updates in Procedure 1. For $n = 0$, (Z_1, W_1) is the initialized state. Write $Z_{h,n+1} = Z_{n+1} \cap \mathcal{S}_h$. After discovery terminates, keep these objects fixed through $n = T$.

For a discovered state $x \in \mathcal{S}_h$, let $\iota_x \in [L_{\mathrm{d}}]$ and $a_x \in \mathcal{A}$ be the matrix index and action selected when Procedure 1 inserts x, and define $\Delta_x = \phi_h(x, a_x) - \phi_h(x, \pi_h^0(x)) \in \mathbb{R}^d$. For $\iota \in [L_{\mathrm{d}}]$, define the finite assignment set

$$\mathcal{D}_{h,\iota}(n) = \{x \in Z_{h,n+1} : \iota_x = \iota\}. \tag{39}$$

Thus $\mathcal{D}_{h,\iota}(n)$ records exactly the states whose insertion updated matrix ι at layer h during the first n episodes. These sets are disjoint across ι, increase with n, and satisfy

$$Z_{h,n+1} = \bigcup_{\iota=1}^{L_{\mathrm{d}}} \mathcal{D}_{h,\iota}(n), \qquad W_{h,n+1}^{(\iota)} = \lambda_{\mathrm{d}} I_d + \sum_{x \in \mathcal{D}_{h,\iota}(n)} \Delta_x \Delta_x^\top. \tag{40}$$

Lemma B.5 (Matrix update counts and novelty inclusion)**.** *Suppose $\eta \le H$ and use the parameters of Appendix A.1. For every $h \in \{2, \dots, H\}$, $\iota \in [L_{\mathrm{d}}]$, and $n \in \{0, \dots, T\}$,*

$$|\mathcal{D}_{h,\iota}(n)| < m_{\mathrm{d}}, \qquad |Z_{n+1}| \le M = (H-1) L_{\mathrm{d}} m_{\mathrm{d}}.$$

With probability at least $1 - \delta/8$, simultaneously for all $h \in \{2, \dots, H\}$ and $j \in \{0, \dots, J-1\}$, there exists an index $\iota_{h,j} \in [L_{\mathrm{d}}]$ such that

$$\mathcal{D}_{h,\iota_{h,j}}(n_j) \subseteq \{z \in \mathcal{S}_h \setminus C_h : a_h^{\pi^\star}(z) \le \nu_j\}. \tag{41}$$

The same index satisfies this containment with n_j replaced by any $n \in \{0, \dots, n_j\}$; an empty assignment set also satisfies it. On this event, for every $h \in \{2, \dots, H\}$, $j \in \{0, \dots, J-1\}$, and $n \in \{0, \dots, n_j\}$,

$$\{z \in \mathcal{S}_h \setminus Z_{h,n+1} : \alpha_h^0(z) \ge r_j\} \subseteq \mathcal{G}(Z_{n+1}, W_{n+1}). \tag{42}$$

Proof. Fix h, ι, and n, and put $k = |\mathcal{D}_{h,\iota}(n)|$. List the selected vectors of these k updates in insertion order as $v_1, \dots, v_k \in \mathbb{R}^d$, and put $U_i = \lambda_{\mathrm{d}} I_d + \sum_{\ell=1}^{i} v_\ell v_\ell^\top$ for $i \in \{0, \dots, k\}$. Then $U_k = W_{h,n+1}^{(\iota)}$. The update rule in Procedure 1 gives $v_i^\top U_{i-1}^{-1} v_i > 1$ for each $i \ge 1$. Since $\|v_i\|_2 \le 2$ and $k \le T$, the determinant lemma and the arithmetic–geometric mean inequality for the eigenvalues give, when $k > 0$,

$$2^k < \frac{\det U_k}{\lambda_{\mathrm{d}}^d} \le \left(1 + \frac{4k}{d\lambda_{\mathrm{d}}}\right)^d \le \left(1 + \frac{4T}{d\lambda_{\mathrm{d}}}\right)^d.$$

Thus $k < m_{\mathrm{d}}$ by (10), also when $k = 0$. Summing over the disjoint assignment sets in (40) proves $|Z_{n+1}| \le M$.

Since $\eta \ge 4H/\sqrt{T}$, $J \le J_{\mathrm{d}}$. Also $\zeta = \eta/(16\overline{\Gamma}H) \le 1/(16\overline{\Gamma})$ and $w_0 = 2\zeta$, so

$$4p_\eta \le p_j \le \frac{1}{2J_{\mathrm{d}}} \le 1, \qquad n_j \le \frac{3\mathsf{S}_{\mathrm{d}}}{p_j}.$$

Fix $h \in \{2, \dots, H\}$ and $j \in \{0, \dots, J-1\}$, and define the measurable set

$$F_{h,j} = \{z \in \mathcal{S}_h \setminus C_h : a_h^{\pi^\star}(z) > \nu_j\}.$$

Let $\mathcal{F}_t$ be the learning history after all updates for discovery episode t, with the history held fixed after discovery terminates. Let $I_t^{h,j} \in \{0, 1\}$ indicate that episode t inserts a state in $F_{h,j}$, and set $I_t^{h,j} = 0$ after termination. Such an insertion requires $S_h^{(t)} \in F_{h,j}$. For each predecessor pair $x \in \mathcal{S}_{h-1} \times \mathcal{A}$, (37) gives

$$P_{h-1}(F_{h,j} \mid x) \le \frac{1}{\nu_j} \int_{\mathcal{S}_h \setminus C_h} a_h^{\pi^\star}(z) P_{h-1}(dz \mid x) \le \frac{c_d \chi}{\nu_j} =: q_j = \frac{8 c_d \sqrt{m_{\mathrm{d}}}\chi}{r_j}.$$

Conditioning first on the predecessor state and action and then on $\mathcal{F}_{t-1}$ therefore yields $\mathbb{E}[I_t^{h,j} \mid \mathcal{F}_{t-1}] \le q_j$. Over the first deterministic n_j episodes these conditional means sum to at most $q_j n_j$. The scale choices give

$$\begin{aligned} q_j n_j &\le \frac{384\overline{\Gamma} H J_{\mathrm{d}} c_d \sqrt{m_{\mathrm{d}}} \mathsf{S}_{\mathrm{d}} \chi w_j}{\eta r_j} \\ &\le \frac{768\overline{\Gamma} H J_{\mathrm{d}} c_d \sqrt{m_{\mathrm{d}}} \mathsf{S}_{\mathrm{d}} \chi}{\eta} \le \frac{L_{\mathrm{d}}}{16}. \end{aligned}$$

The final comparison follows from (13) and (14). Set $\ell_0 = \log\{8(H-1)J_{\mathrm{d}}/\delta\} > 0$. Since $\mathbb{E}[e^{I_t^{h,j}} \mid \mathcal{F}_{t-1}] \le \exp\{(e-1)\mathbb{E}[I_t^{h,j} \mid \mathcal{F}_{t-1}]\}$, iteration and Markov's inequality give

$$\mathbb{P}\left\{\sum_{t=1}^{n_j} I_t^{h,j} > 2q_j n_j + \ell_0\right\} \le \exp\{(e-3)q_j n_j - \ell_0\} \le \frac{\delta}{8(H-1)J_{\mathrm{d}}}.$$

By (10), $\ell_0 \le L_{\mathrm{d}}/8$, so $2q_j n_j + \ell_0 \le L_{\mathrm{d}}/4$. A union bound over the $(H-1)J$ choices of (h,j) gives an event $\mathcal{E}_{\mathrm{mat}}$ with $\mathbb{P}(\mathcal{E}_{\mathrm{mat}}) \ge 1 - \delta/8$ on which $\sum_{t=1}^{n_j} I_t^{h,j} \le L_{\mathrm{d}}/4$ for every (h,j). This event does not condition on certification.

Work on $\mathcal{E}_{\mathrm{mat}}$. For each fixed (h,j), define the two sets of matrix indices

$$\begin{aligned} \mathcal{I}_{h,j}^{C} &= \{\iota \in [L_{\mathrm{d}}] : \mathcal{D}_{h,\iota}(n_j) \cap C_h \neq \varnothing\}, \\ \mathcal{I}_{h,j}^{F} &= \{\iota \in [L_{\mathrm{d}}] : \mathcal{D}_{h,\iota}(n_j) \cap F_{h,j} \neq \varnothing\}. \end{aligned}$$

Each state is inserted at most once and is assigned to exactly one matrix. Hence

$$|\mathcal{I}_{h,j}^{C}| \le |C_h| \le b_d K \le L_{\mathrm{d}}/2, \qquad |\mathcal{I}_{h,j}^{F}| \le \sum_{t=1}^{n_j} I_t^{h,j} \le L_{\mathrm{d}}/4.$$

It follows that

$$\left|[L_{\mathrm{d}}] \setminus (\mathcal{I}_{h,j}^{C} \cup \mathcal{I}_{h,j}^{F})\right| \ge L_{\mathrm{d}}/4 > 0.$$

Choose $\iota_{h,j}$ to be the smallest index in this complement. Its assignment set is disjoint from both C_h and $F_{h,j}$, proving (41). For every $0 \le n \le n_j$, $\mathcal{D}_{h,\iota_{h,j}}(n) \subseteq \mathcal{D}_{h,\iota_{h,j}}(n_j)$, so this single choice of index has the stated property for every $n \in \{0,\dots,n_j\}$. No additional union bound over n is needed.

To prove (42), fix (h,j) and an integer $n \in \{0,\dots,n_j\}$, and write $W = W_{h,n+1}^{(\iota_{h,j})} \in \mathbb{R}^{d\times d}$. Let $\theta = \theta_h^{\pi^\star} \in \mathbb{R}^d$ be a coefficient vector supplied by Assumption 3.1, with $\|\theta\|_2 \le B$. For every $x \in \mathcal{D}_{h,\iota_{h,j}}(n)$,

$$|\Delta_x^\top \theta| \le |Q_h^\star(x,a_x) - Q_h^\star(x,\pi_h^0(x))| + 2\kappa \le a_h^{\pi^\star}(x) + 2\kappa \le \nu_j + 2\kappa.$$

Equations (13) and (14) imply $\eta \ge 256\overline{\Gamma} H\sqrt{m_{\mathrm{d}}}\kappa$ and $\eta \ge 4H/\sqrt{T}$. Using $r_j \ge \zeta$, these give $2\kappa \le \nu_j \le r_j/8$ and $\sqrt{\lambda_{\mathrm{d}}} B \le \zeta/4 \le r_j/4$. Thus (40) and the deterministic update bound yield

$$\begin{aligned} \|\theta\|_W^2 &= \lambda_{\mathrm{d}}\|\theta\|_2^2 + \sum_{x \in \mathcal{D}_{h,\iota_{h,j}}(n)} (\Delta_x^\top \theta)^2 \\ &\le \lambda_{\mathrm{d}} B^2 + 4 m_{\mathrm{d}} \nu_j^2 \le \frac{r_j^2}{8}. \end{aligned} \tag{43}$$

For any $z \in \mathcal{S}_h \setminus Z_{h,n+1}$ with $\alpha_h^0(z) \ge r_j$, the vector $v = \phi_h(z,\pi_h^\star(z)) - \phi_h(z,\pi_h^0(z)) \in \mathbb{R}^d$ satisfies $v^\top\theta \ge \alpha_h^0(z) - 2\kappa \ge 7r_j/8$. Matrix Cauchy–Schwarz gives

$$(7r_j/8)^2 \le (v^\top\theta)^2 \le (v^\top W^{-1} v)\|\theta\|_W^2 \le (v^\top W^{-1} v) r_j^2/8.$$

Therefore $v^\top W^{-1} v \ge 49/8 > 1$. Since $v \in \mathcal{V}_h(z)$, the novelty test (3) places z in $\mathcal{G}(Z_{n+1}, W_{n+1})$. This proves (42) for every integer $0 \le n \le n_j$, including initialization and the state after all updates for episode n_j (or the terminal state if discovery has already stopped). □

Procedure 1 retains labels collected before later discoveries. The next lemma shows that their conditional-mean targets can only decrease and identifies the conditional law of each observed label.

Lemma B.6 (Monotonicity and conditional law of discovery labels)**.** *Fix an executed discovery episode t and a pair $u \in \mathcal{U}_t$. The label functions in (5) and their means $q_t^{\mathrm{d}}(u) = \int_{\Omega_u} F_t^u \, dP_u$ satisfy, on every learning history,*

$$\begin{aligned} F_{t+1}^u(\omega) &\le F_t^u(\omega) \quad \text{for every } \omega \in \Omega_u, \\ q_{t+1}^{\mathrm{d}}(u) &\le q_t^{\mathrm{d}}(u). \end{aligned} \tag{44}$$

For a recorded visit to u in episode t, let $L_t^u \in [0,1]$ be the stored label and let $\mathcal{H}_{t,u}$ be the sigma-field generated by the history after selecting u but before observing its segment outcome. For $u = u_\circ$, this history precedes the initial-state and root-action draws. Then, for every $b \in [0,1]$,

$$\begin{aligned} \mathbb{P}\{L_t^u \le b \mid \mathcal{H}_{t,u}\} &= \int_{\Omega_u} \mathbf{1}\{F_t^u(\omega) \le b\} \, P_u(d\omega), \\ \mathbb{E}[L_t^u \mid \mathcal{H}_{t,u}] &= q_t^{\mathrm{d}}(u). \end{aligned} \tag{45}$$

Proof. Write $\mathcal{G}_t = \mathcal{G}(Z_t, W_t)$. The matrix updates in Procedure 1 and the score update (21) give

$$\begin{aligned} &Z_t \subseteq Z_{t+1}, \qquad W_{h,t}^{(\iota)} \preceq W_{h,t+1}^{(\iota)} \quad (h \in \{2,\dots,H\},\ \iota \in [L_{\mathrm{d}}]), \\ &Q_{t+1}^+(u) \le Q_t^+(u) \quad (u \in \mathcal{U}_t), \qquad V_{t+1}^+(y) \le V_t^+(y) \quad (y \in Z_t). \end{aligned}$$

Inversion reverses the positive-semidefinite order, so the novelty test (3) gives $\mathcal{G}_{t+1} \subseteq \mathcal{G}_t$. Every newly inserted state was in $\mathcal{G}_t$, and hence

$$Z_{t+1} \setminus Z_t \subseteq \mathcal{G}_t, \qquad Z_{t+1} \cup \mathcal{G}_{t+1} \subseteq Z_t \cup \mathcal{G}_t.$$

Fix $\omega \in \Omega_u$, and let $e_t^u(\omega) \in Z_t \cup \mathcal{G}_t \cup \{\perp\}$ denote its first-event endpoint under (5), with $\perp$ denoting termination. The last inclusion prevents an earlier event under the episode-$(t+1)$ rule. If $e_t^u(\omega) \in Z_t$ or equals $\perp$, that endpoint therefore remains unchanged. The three cases are

$$\begin{aligned} &F_{t+1}^u(\omega) \le 1 = F_t^u(\omega), && e_t^u(\omega) \in \mathcal{G}_t, \\ &F_{t+1}^u(\omega) = V_{t+1}^+(y) \le V_t^+(y) = F_t^u(\omega), && e_t^u(\omega) = y \in Z_t, \\ &F_{t+1}^u(\omega) = 0 = F_t^u(\omega), && e_t^u(\omega) = \perp. \end{aligned}$$

Integrating this pointwise inequality against the fixed law P_u proves the conditional-mean inequality in (44).

For (45), condition on $\mathcal{H}_{t,u}$. The functions F_t^u and V_t^+ and the sets $Z_t, \mathcal{G}_t$ are then fixed. Until the first event, the Markov property gives the observed segment the same law as the corresponding prefix of a primitive continuation with law P_u. Complete that segment with a fresh default-policy suffix conditional on its endpoint, obtaining $\Omega_t^u \in \Omega_u$ on an enlarged probability space with

$$\mathrm{Law}(\Omega_t^u \mid \mathcal{H}_{t,u}) = P_u, \qquad L_t^u = F_t^u(\Omega_t^u) \quad \text{almost surely.}$$

The added suffix does not change the first event or the stored label. Taking conditional probabilities and expectations proves (45). The completion preserves the joint law of the observed rewards and endpoint. □

Equation (44) allows an upper confidence bound computed from an earlier batch to bound a later target, as made explicit in (55). Equation (45) supplies the conditional means used in that concentration argument.

B.2 A CONDITIONAL-MEAN CONVERSION INEQUALITY

The two inequalities in (46) will be used in the analyses of Procedures 1 and 4. The intermediate sigma-field permits a context or action to be revealed before the outcome being averaged.

Lemma B.7 (Conditional-mean conversion)**.** *Fix a positive integer n and a deterministic scalar $R_0 \ge 0$. Let $\mathcal{F}_0 \subseteq \cdots \subseteq \mathcal{F}_n$ be a filtration, and let $\mathcal{G}_k$ be a sigma-field satisfying $\mathcal{F}_{k-1} \subseteq \mathcal{G}_k \subseteq \mathcal{F}_k$. For $k \in [n]$, let Z_k be $\mathcal{F}_k$-measurable with $0 \le Z_k \le R_0$, and set $\mu_k = \mathbb{E}[Z_k \mid \mathcal{G}_k]$. For every $\delta \in (0,1)$, each of the following inequalities holds with probability at least $1-\delta$, separately:*

$$\sum_{k=1}^n \mu_k \le 2\sum_{k=1}^n Z_k + 2R_0 \log(1/\delta), \qquad \sum_{k=1}^n Z_k \le 2\sum_{k=1}^n \mu_k + R_0 \log(1/\delta). \tag{46}$$

To obtain both directions on one event of probability at least $1-\delta$, apply each with failure probability $\delta/2$.

Proof. Suppose $R_0 > 0$, and write $z_k = Z_k/R_0$ and $m_k = \mu_k/R_0$. Convexity gives $e^{-x} \le 1-(1-e^{-1})x$ and $e^x \le 1+(e-1)x$ on $[0,1]$. The tower property, first through $\mathcal{G}_k$ and then through the previous sigma-field $\mathcal{F}_{k-1}$, therefore shows that

$$\exp\left\{(1-e^{-1})\sum_{k\le n} m_k - \sum_{k\le n} z_k\right\}, \qquad \exp\left\{\sum_{k\le n} z_k - (e-1)\sum_{k\le n} m_k\right\}$$

are nonnegative supermartingales in n, relative to $(\mathcal{F}_n)$, starting at one. Indeed, the conditional expectations of the one-step factors given $\mathcal{G}_k$ are at most $\exp\{(1-e^{-1})m_k\}\{1-(1-e^{-1})m_k\} \le 1$ and $\exp\{-(e-1)m_k\}\{1+(e-1)m_k\} \le 1$, respectively. Markov's inequality and $(1-e^{-1})^{-1} < 2$, $e-1<2$ give the first and second inequalities of (46), respectively. When $R_0 = 0$, both inequalities hold deterministically. □

B.3 Proof of the discovery episode bound

This subsection proves (47) from the update formulas in Appendix A.3. For positive integers D, ℓ, T, a nonnegative integer R, $p \in (0,1]$, and $\delta_0 \in (0,1)$, let $\mathfrak{N}(D,\ell,R,p,\delta_0,T)$ be the deterministic integer in (11).

Lemma B.8 (Episode bound for SimultaneousDiscovery)**.** *Fix the episode budget T, put $\ell = H$, and let D be the deterministic pair cap used in (20). The update schedule is (21). With probability at least $1-\delta_0$, the following holds simultaneously for every executed prefix of length $N \in \{0,\dots,T\}$, every integer $R \in \{0,\dots,T\}$, and every $p \in (0,1]$: if the prefix creates at most D pairs, visits at most ℓ pairs per episode, returns at most R novelty states, and has $Q_t^+(u_\circ) \ge p$ for $t = 1,\dots,N$, then*

$$N \le \mathfrak{N}(D,\ell,R,p,\delta_0,T) = \widetilde{O}\left(\frac{D\ell+R}{p}\right). \tag{47}$$

The event also includes (57) for $t = 1,\dots,N+1$, where $t = N+1$ denotes the state after all updates for episode N. No support restriction is imposed on the primitive continuation laws.

The proof uses batch confidence, Bellman telescoping, and bounds on the batch-mean errors and confidence radii. The conditional-mean inequality (46) was proved in Appendix B.2.

We use three indices and keep them distinct throughout:

$$\begin{aligned} k &= \text{discovery episode number},\\ j &= \text{local visit number to one fixed pair},\\ i &= \text{decision position within an episode}. \end{aligned} \tag{48}$$

B.3.1 Pair-local samples and the confidence event

We prove batch confidence in (54) and policy optimism in (57).

The prefix $N = 0$ satisfies the bound and optimism by initialization. Fix a prefix of $N \in \{1,\dots,T\}$ discovery episodes. Let $\mathcal{U}_{1:N} = \bigcup_{t=1}^{N} \mathcal{U}_t$ be the finite set of pairs available at the starts of these episodes. Pairs created after episode N have no visits in this prefix. Let D be the deterministic cap used in (20), so $|\mathcal{U}_{1:N}| \le D$; for Procedure 1, $D = D_{\max} = \overline{D}$. Let $\ell = H$ be the maximum number of visited pairs per episode, and let $I_t \in \{0,1\}$ indicate that Procedure 2 returns a novel state in episode t. Define the integer $R_N = \sum_{t=1}^{N} I_t$. Put $L_\circ = L_{\mathrm{reach}}(D,\ell,T,\delta_0)$. In this subsection, $\widetilde{O}$ suppresses universal constants and powers of $L_\circ$; explicit constants are retained for the final episode bound. All probability events constructed below are independent of a score threshold p. For each $u \in \mathcal{U}_{1:N}$, let

$$\tau_1^u < \tau_2^u < \cdots \tag{49}$$

be the global episodes in which u is visited, strictly after the stopping time at which u is created. All unqualified sums over j below are restricted to visits with $\tau_j^u \le N$. On local visit j, define

the measurable function, random variable, conditional mean, martingale difference, and conditional variance

$$f_j^u := F^u_{\tau_j^u}, \qquad \widetilde{\Omega}_j^u \sim P_u, \qquad X_j^u = f_j^u(\widetilde{\Omega}_j^u), \qquad \mu_j^u = P_u f_j^u, \\ \xi_j^u = X_j^u - \mu_j^u, \qquad (\sigma_j^u)^2 = \mathbb{E}[(\xi_j^u)^2 \mid \mathcal{F}^u_{j-1}], \tag{50}$$

where $P_u f_j^u := \int_{\Omega_u} f_j^u(\omega) P_u(d\omega)$ and $\mathcal{F}^u_{j-1}$ is the history immediately before the continuation used on the jth local visit. The lowercase f_j^u uses local visits, while F_t^u retains the episode index from (5). Thus $f_j^u : \Omega_u \to [0,1]$ is a predictable function, whereas $X_j^u \in [0,1]$ is the random label actually stored. Conditional on the history at which u is created, generate an infinite i.i.d. sequence with law P_u and reveal its next unused element on each visit. The visit decision precedes that revelation, so the conditional law remains P_u. Unused elements remain unobserved. Equivalently, extend each observed stopped prefix by a fresh default-policy completion conditional on its stopping state and layer. Only the stopped prefix is used by the algorithm. By Lemma B.6, $f^u_{j+1}(\omega) \le f_j^u(\omega)$ pointwise and hence

$$1 \ge \mu_1^u \ge \mu_2^u \ge \cdots \ge 0. \tag{51}$$

We first record the confidence statement used by the batch update.

Lemma B.9 (Predictable empirical–Bernstein confidence)**.** *Fix $n \in \mathbb{N}$ and $\delta \in (0,1)$. Let $(\mathcal{F}_i)_{i=0}^n$ be a filtration and let $X_i \in [0,1]$ be adapted to it, $\mu_i = \mathbb{E}[X_i \mid \mathcal{F}_{i-1}]$, and let $\widehat{X}_i \in [0,1]$ be predictable. Put $\xi_i = X_i - \mu_i$ and $A_n = \sum_{i=1}^n (X_i - \widehat{X}_i)^2$ and $L = \log\{8(n+1)/\delta\}$. With probability at least $1-\delta$,*

$$\left| \sum_{i=1}^n \xi_i \right| \le 4\sqrt{(A_n + L)L} + 4L. \tag{52}$$

The same statement holds conditionally after a stopping time.

Proof. For $Z_i = (X_i - \widehat{X}_i)^2 \in [0,1]$,

$$\mathbb{E}[Z_i \mid \mathcal{F}_{i-1}] = \mathrm{Var}(X_i \mid \mathcal{F}_{i-1}) + (\mu_i - \widehat{X}_i)^2 \ge \mathbb{E}[\xi_i^2 \mid \mathcal{F}_{i-1}]. \tag{53}$$

The first inequality in (46), applied to (Z_i) with failure probability $\delta/2$, gives

$$\sum_{i=1}^n \mathbb{E}[Z_i \mid \mathcal{F}_{i-1}] \le 2A_n + 2\log(2/\delta).$$

For the remaining allocation, partition the range of $V_n = \sum_i \mathbb{E}[\xi_i^2 \mid \mathcal{F}_{i-1}] \in [0,n]$ into the initial variance interval $[0,1]$ and the *dyadic variance intervals* $(2^{b-1}, 2^b]$ for $b = 1, \ldots, \lceil \log_2 n \rceil$, intersected with $[0,n]$. When V_n lies in a dyadic variance interval, that interval's upper endpoint is at most $2V_n$. There are at most $1 + \lceil \log_2 n \rceil \le n+1$ variance intervals. Freedman's inequality (Freedman, 1975), applied to both signs and unioned over these intervals, gives with total failure probability at most $\delta/2$

$$\left| \sum_{i=1}^n \xi_i \right| \le 2\sqrt{V_n L} + 2L.$$

Indeed, Freedman's inequality at a deterministic variance cap v gives the valid threshold $\sqrt{2vL} + 2L/3$. Substituting the upper endpoint $v \le 2V_n$ proves the bound on each dyadic variance interval; on the initial variance interval, $\sqrt{2L} + 2L/3 \le 2L$ because $L \ge \log 16 > 9/8$. The definition of L dominates the logarithm of the two signs and all variance intervals. On the intersection of these two events, (53) permits the substitution $V_n \le 2A_n + 2\log(2/\delta)$ and hence

$$\left| \sum_{i=1}^n \xi_i \right| \le 2\sqrt{\{2A_n + 2L\}L} + 2L,$$

which is bounded by the right side of (52). Conditioning on the stopping-time sigma-field before starting the same two supermartingales proves the final assertion. □

Apply Lemma B.9 to every completed batch, using the preceding label as the forecast even when it belongs to the preceding batch, as specified in (20). Number pairs by creation order and batches by completion order. There are at most $D(T+1)$ relevant pair–batch indices. Apply the lemma to each with $\delta_B = \delta_0/\{4D(T+1)\}$. The definition of Λ and a union bound then give, with failure probability at most $\delta_0/4$, the symmetric event

$$\left|\sum_{j\in I_B}(X_j^u - \mu_j^u)\right| \le |B|\,\mathrm{rad}(B) = c_1\sqrt{(A_B+\Lambda)\Lambda} + c_2\Lambda \tag{54}$$

for every completed discovery batch. To justify the adaptive indices, let $\mathcal{H}_B$ be the history when batch B starts: the history at which the pair is created for its first batch, and the preceding batch-completion history otherwise. Its target size n_B is $\mathcal{H}_B$-measurable. If discovery terminates before B completes, append deterministic zero labels after termination until there are n_B labels. This defines a length-n_B adapted process without conditioning on completion; its artificial labels have conditional mean zero. Conditional on $\mathcal{H}_B$, the lemma therefore gives

$$\mathbb{P}\{B \text{ completes and violates (54)} \mid \mathcal{H}_B\} \le \delta_B.$$

Indices never created have empty failure events; batches with target size exceeding T cannot complete and need no confidence bound. Taking expectations and unioning the $D(T+1)$ indices proves the assertion. Indeed, for a batch of size $n \le T$, the corresponding logarithm is

$$L_B = \log\{32D(T+1)(n+1)/\delta_0\}.$$

Its argument is at most $\{64D_{\max}(T+1)/\delta_0\}^2$, so $L_B \le \Lambda/4 < \Lambda$. Thus $c_1 = 8$ and $c_2 = 32$ in (20) dominate the explicit right side of (52).

Fix an episode-state index $t \in \{1,\dots,N+1\}$, a pair $u \in \mathcal{U}_t$, and a batch B of u completed before episode t. Every local visit $j \in I_B$ occurred at $\tau_j^u < t$, so (51) and the lower-tail direction in (54) give

$$q_t^{\mathrm{d}}(u) = P_u F_t^u \le \frac{1}{|B|}\sum_{j\in I_B}\mu_j^u \le \frac{1}{|B|}\sum_{j\in I_B}X_j^u + \mathrm{rad}(B). \tag{55}$$

For a pair without a completed batch, the score is one. Otherwise, both arguments of the minimum in (21) dominate $q_t^{\mathrm{d}}(u)$: the previous score does so by induction in time and target monotonicity, while the frozen-batch upper estimate does so by (55). Reverse induction over the acyclic first-hit control graph makes this simultaneous argument well defined and gives statewise optimism in the explicit form

$$Q_t^+(u) \ge \int_{\Omega_u} F_t^u(\omega)\,P_u(d\omega) \quad \text{for every } t \le N+1 \text{ and } u \in \mathcal{U}_t. \tag{56}$$

The post-episode state $t = N+1$ is covered as well. Let $\mathcal{E}_t^{\mathrm{d}}$ be the event that (54) holds for every batch completed before episode t. This event is measurable at that episode's start, and the induction proves

$$\mathbf{1}\{\mathcal{E}_t^{\mathrm{d}}\}\big(Q_t^+(u) - P_u F_t^u\big) \ge 0 \qquad (u \in \mathcal{U}_t).$$

On $\mathcal{E}_t^{\mathrm{d}}$, optimism holds for every current pair, including unvisited pairs. Every later martingale conversion is applied to the unconditioned bounded process and then intersected with $\bigcap_{t\le N+1}\mathcal{E}_t^{\mathrm{d}}$. No conditional expectation is taken conditional on this eventual event. Equation (56) also bounds the novelty-hit probability under every policy on the stored states. Let $\rho : Z_t \to \mathcal{A}$ be any deterministic stored-state policy, and let $p_t^\rho(u)$ be the probability of hitting $\mathcal{G}(Z_t, W_t)$ before termination when the process starts from a controlled pair $u = (z,a)$, first takes a, subsequently uses ρ on Z_t, and uses π^0 elsewhere. Define $p_t^\rho(u_\circ)$ analogously, with the fixed root design as the initial action rule. Reverse induction in the layer order gives

$$Q_t^+(z,a) \ge p_t^\rho(z,a) \quad (z \in Z_t,\ a \in \mathcal{A}), \qquad Q_t^+(u_\circ) \ge p_t^\rho(u_\circ). \tag{57}$$

Indeed, at the last controlled layer (56) is exactly the one-step domination. At an earlier layer, its stored-state continuation is $V_t^+(y) = \max_a Q_t^+(y,a) \ge Q_t^+(y,\rho(y))$, so the induction hypothesis dominates the continuation probability under ρ. The same argument at the synthetic root uses its fixed design distribution. In particular, if Procedure 1 stops because $Q^+(u_\circ) < p$, then the optimal current novelty-hitting probability is below p: the finite acyclic control problem has a deterministic maximizing policy, and randomized policies are convex combinations at each decision.

B.3.2 Episode notation and Bellman telescoping

We establish the identities (65) and the conditional-mean bound (66).

Fix an executed episode k. Let $\mathcal{F}_{k,0}$ be the history after the episode-k objects Z_k, W_k, Q_k^+ and V_k^+ have been frozen but before the initial state and root action are sampled. Let $1 \le \nu_k \le \ell$ be the number of pairs selected before the episode reaches the novelty set or terminates, and list them in path order as $u_{k,1}, \dots, u_{k,\nu_k}$, with $u_{k,1} = u_\circ$. Set $\mathcal{F}_{k,1} = \mathcal{F}_{k,0}$ for the synthetic start pair. For $2 \le i \le \nu_k$, let $\mathcal{F}_{k,i}$ be the history after $u_{k,i}$ is selected and before its stopped segment outcome is revealed. Because the control graph follows the acyclic layer order, a pair is selected at most once in an episode. Hence there is a unique local index $j(k,i)$ such that

$$\tau^{u_{k,i}}_{j(k,i)} = k. \tag{58}$$

The episode-path aliases for the local objects in (50) are

$$\begin{aligned} F_{k,i} &:= f^{u_{k,i}}_{j(k,i)} = F^{u_{k,i}}_k, \qquad \Omega_{k,i} := \widetilde{\Omega}^{u_{k,i}}_{j(k,i)}, \\ X_{k,i} &:= F_{k,i}(\Omega_{k,i}) = X^{u_{k,i}}_{j(k,i)}, \qquad \mu_{k,i} := P_{u_{k,i}} F_{k,i} = \mu^{u_{k,i}}_{j(k,i)}. \end{aligned} \tag{59}$$

If u is visited in episode k, let $i_k(u)$ be its unique path position and define

$$\mu_{k,u} := \mu_{k,i_k(u)}. \tag{60}$$

Thus $\mu_{k,u}$ is only an episode-indexed alias for the pair-local mean; it is not a new target.

The optimistic value at an actual decision point is

$$Y_{k,i} := Q_k^+(u_{k,i}) = \begin{cases} Q_k^+(u_\circ), & i = 1, \\ V_k^+(z_{k,i}) = Q_k^+(z_{k,i}, a_k^+(z_{k,i})), & i \ge 2, \end{cases} \tag{61}$$

where $z_{k,i} \in Z_k$ is the state at decision position $i \ge 2$ (position 1 is the synthetic start pair). The indicator I_k records whether episode k reaches $\mathcal{G}(Z_k, W_k)$. After the first novelty hit or termination, set $Y_{k,i} = I_k$ for $i = \nu_k + 1, \dots, \ell + 1$ and let $\mathcal{F}_{k,i}$ be the history through that hit or termination for all these indices. The definition of the stored label in (5) then gives the pathwise identity

$$X_{k,i} = Y_{k,i+1}, \qquad \mu_{k,i} = \mathbb{E}[X_{k,i} \mid \mathcal{F}_{k,i}] = \mathbb{E}[Y_{k,i+1} \mid \mathcal{F}_{k,i}] \quad (i \le \nu_k). \tag{62}$$

Indeed, both sides equal one at a novelty hit, $V_k^+(y)$ at the next stored state y, and zero at termination without a novelty hit.

Define $q_k := \mathbb{E}[I_k \mid \mathcal{F}_{k,0}]$. On the $\mathcal{F}_{k,0}$-measurable event $\mathcal{E}_k^{\mathrm{d}}$, for every actual decision, (56) and (62) imply $Y_{k,i} - \mathbb{E}[Y_{k,i+1} \mid \mathcal{F}_{k,i}] \ge 0$; the same difference is zero on the padded indices. Therefore define

$$\Delta_{k,i} := Y_{k,i} - \mathbb{E}[Y_{k,i+1} \mid \mathcal{F}_{k,i}], \qquad w_k := \mathbb{E}\left[\sum_{i=1}^{\ell} \Delta_{k,i} \,\middle|\, \mathcal{F}_{k,0}\right]. \tag{63}$$

The differences $\Delta_{k,i}$ are nonnegative on $\mathcal{E}_k^{\mathrm{d}}$; their definition applies on every history. Taking conditional expectations in this definition and telescoping gives the exact identities

$$w_k = Q_k^+(u_\circ) - q_k, \qquad \mathbb{E}[Y_{k,i} \mid \mathcal{F}_{k,0}] = q_k + \mathbb{E}\left[\sum_{r=i}^{\ell} \Delta_{k,r} \,\middle|\, \mathcal{F}_{k,0}\right] \quad (1 \le i \le \ell + 1). \tag{64}$$

In particular, put

$$Q := \sum_{k=1}^{N} q_k, \qquad W := \sum_{k=1}^{N} w_k = \sum_{k=1}^{N} \{Q_k^+(u_\circ) - q_k\}. \tag{65}$$

For the next bound only, extend $\mu_{k,i}$ by zero when $i > \nu_k$. On an actual decision it equals the conditional mean of $Y_{k,i+1}$ by (62); on a padded index it is at most that conditional mean because $Y_{k,i+1} \ge 0$. The tower property and (64), now used at index $i + 1$, therefore give

$$\begin{aligned} \mathbb{E}\left[\sum_{u \text{ visited in } k} \mu_{k,u} \,\middle|\, \mathcal{F}_{k,0}\right] &= \sum_{i=1}^{\ell} \mathbb{E}[\mu_{k,i} \mid \mathcal{F}_{k,0}] \\ &\le \sum_{i=1}^{\ell} \mathbb{E}[Y_{k,i+1} \mid \mathcal{F}_{k,0}] \le \ell(q_k + w_k). \end{aligned} \tag{66}$$

B.3.3 Error from earlier batch means

We bound the accumulated batch-mean error by (73).

Let $\mathcal{I}_{\text{batch}}$ be the set of visits (k,i) for which $u_{k,i}$ has a completed frozen batch $B_{k,i}$. If $I_{B_{k,i}}$ is that batch's set of pair-local indices, define

$$\begin{aligned} \overline{\mu}_{k,i} &:= \frac{1}{|B_{k,i}|} \sum_{j \in I_{B_{k,i}}} \mu_j^{u_{k,i}}, \qquad d_{k,i} := \overline{\mu}_{k,i} - \mu_{k,i} \geq 0, \\ r_{k,i} &:= \mathrm{rad}(B_{k,i}), \qquad S_{\text{mean}} := \sum_{(k,i)\in\mathcal{I}_{\text{batch}}} d_{k,i}, \qquad S_{\text{rad}} := \sum_{(k,i)\in\mathcal{I}_{\text{batch}}} r_{k,i}. \end{aligned} \tag{67}$$

The inequality $d_{k,i} \geq 0$ follows from (51): every index in the frozen batch precedes the current local visit.

Fix one pair u. Index its completed batches by b. For consecutive batches B_b, B_{b+1}, put

$$m_b := |B_b|, \qquad M_b := \sum_{j \in I_{B_b}} \mu_j^u, \qquad \gamma_b := m_{b+1}/m_b. \tag{68}$$

While B_{b+1} is filled, B_b is the frozen batch. Consequently its total error from earlier batch means is exactly

$$\sum_{j \in I_{B_{b+1}}} \left(\frac{M_b}{m_b} - \mu_j^u \right) = \gamma_b M_b - M_{b+1}. \tag{69}$$

After the initial doubling batches, $\gamma_b \leq 1 + 2\rho$, and hence

$$\gamma_b M_b - M_{b+1} = (M_b - M_{b+1}) + (\gamma_b - 1) M_b \leq M_b - M_{b+1} + 2\rho M_b. \tag{70}$$

Fewer than $4/\rho$ visits occur before a pair has a completed slow-growth batch: the completed doubling sizes sum to less than $2/\rho$, and the final incomplete target is smaller than $2/\rho$. Each such visit contributes at most one. Let B_{b_0} be the first completed batch in the slow-growth regime. Its size is at most $2/\rho$, so $M_{b_0} \leq 2/\rho$. Let $b_{\max}$ be the last completed batch number, and let C be the set of local visit indices in the possibly empty active batch after $B_{b_{\max}}$. Its target size is $m_{b_{\max}+1} = g_{\mathrm{d}}(m_{b_{\max}})$. Then $|C| < m_{b_{\max}+1} \leq (1+2\rho) m_{b_{\max}}$ and

$$\sum_{j \in C} \left(\frac{M_{b_{\max}}}{m_{b_{\max}}} - \mu_j^u \right) \leq M_{b_{\max}} - \sum_{j \in C} \mu_j^u + 2\rho M_{b_{\max}}. \tag{71}$$

Summing (70) from b_0 through $b_{\max} - 1$ and adding (71) gives, for this pair,

$$\sum_{b=b_0}^{b_{\max}-1} (\gamma_b M_b - M_{b+1}) + \sum_{j \in C} \left(\frac{M_{b_{\max}}}{m_{b_{\max}}} - \mu_j^u \right) \leq M_{b_0} - \sum_{j \in C} \mu_j^u + 2\rho \sum_{b=b_0}^{b_{\max}} M_b.$$

Including the incomplete batch cancels $M_{b_{\max}}$; the remaining initial term is $M_{b_0} \leq 2/\rho$. Every M_b is the sum of the conditional means on the visits that formed B_b, and these batches are disjoint. The initial visits contribute less than $4/\rho$, while the initial term M_{b_0} contributes at most $2/\rho$. Bounding their sum by $8/\rho$ and summing over pairs gives

$$S_{\text{mean}} \leq 8D/\rho + 2\rho \sum_{k=1}^{N} \sum_{u \text{ visited in } k} \mu_{k,u}. \tag{72}$$

If a pair never completes a slow-growth batch, all of its visits belong to the preceding initial contribution, so the same display covers that edge case without defining b_0 for that pair.

Applying the observed-to-predictable martingale conversion to (72) and using (66) gives, with failure probability at most $\delta_0/8$, using half of the batch-mean conversion allocation,

$$S_{\text{mean}} \leq 8D/\rho + 4\rho\ell(Q+W) + 2\rho\ell \log \frac{16(T+1)}{\delta_0}. \tag{73}$$

The first term covers the initial doubling batches and the initial term in each telescoping sum. For the remaining term, the realized episode total of conditional means lies in $[0, \ell]$ and has conditional expectation at most $\ell(q_k + w_k)$ by (66); the second inequality in (46), unioned over all prefixes, gives the displayed constants.

B.3.4 SUM OF CONFIDENCE RADII

We prove the radius bound (80).

The sampled labels need not be monotone. With the local indexing in (50), set $X_0^u = \mu_0^u = \xi_0^u = 0$ to represent the initial forecast. Then

$$X_j^u - X_{j-1}^u = \xi_j^u + (\mu_j^u - \mu_{j-1}^u) - \xi_{j-1}^u, \tag{74}$$

and the square of the right side is at most three times the sum of the three squared terms. Since $(\mu_j^u)_{j\ge 1}$ is decreasing but the artificial initial value is $\mu_0^u = 0$,

$$\sum_j (\mu_j^u - \mu_{j-1}^u)^2 \le (\mu_1^u)^2 + \left(\sum_{j\ge 2}(\mu_{j-1}^u - \mu_j^u)\right)^2 \le 2. \tag{75}$$

The second inequality in (46), applied to $(\xi_j^u)^2 \in [0,1]$, followed by a union bound over pair creation order and every deterministic local prefix length $0 \le m \le T$, gives, with failure probability at most $\delta_0/8$, using half of the variance-conversion allocation,

$$\sum_{u,j}(\xi_j^u)^2 \le 2U + 2D\log\frac{16(D+1)(T+1)}{\delta_0}, \qquad U := \sum_{u,j}(\sigma_j^u)^2. \tag{76}$$

The forecast uses the preceding label, including at the first visit of each new batch, so (74)–(76) include every increment $X_j^u - X_{j-1}^u$. Summing the elementary three-term bound first gives $\sum_B A_B \le 6\sum_{u,j}(\xi_j^u)^2 + 6D$. Since the logarithm below is at least one, (76) then implies

$$\sum_B A_B \le 24\left\{U + D\log\frac{16(D+1)(T+1)}{\delta_0}\right\}. \tag{77}$$

The quantity S_{rad} in (67) is the sum of confidence radii over visits with a completed batch. After the initial doubling batches, a completed batch B is reused for at most $(1+2\rho)|B|$ later visits, so by the confidence radius (20) its total contribution is at most

$$16\sqrt{(A_B+\Lambda)\Lambda} + 64\Lambda. \tag{78}$$

The initial doubling batches contribute at most $2^{10}DL_\circ^2/\rho$. Indeed, a size-m doubling batch is reused for at most $2m$ visits, $A_B \le m$, and its total contribution is at most $16\sqrt{(m+\Lambda)\Lambda} + 64\Lambda$. Before the slow-growth regime, the batch sizes form a geometric sequence with $\sum_m \sqrt{m} \le 4/\sqrt{\rho}$. There are at most $1 + \lceil \log_2(1/\rho)\rceil \le L_\circ/(2\rho)$ such batches because $\rho = (32H)^{-1} \le 1/64$, and $\Lambda \le 8L_\circ$. Using $\sqrt{m+\Lambda} \le \sqrt{m} + \sqrt{\Lambda}$ and summing the preceding display gives the stated bound. There are at most $2DL_\circ/\rho$ completed batches: per pair there are at most $\log_2(1/\rho)+1$ doubling batches and at most $\log(T+1)/\log(1+\rho)+1$ slow-growth batches. Since $\log(1+\rho) \ge \rho/(1+\rho)$ and $\rho \le 1$, the latter count is at most $2\rho^{-1}\log(T+1)+1$. Moreover, $\log_2(1/\rho)+2 \le \rho^{-1}+1 \le 2\rho^{-1}$, so the sum of the two counts is at most $2\rho^{-1}\{1+\log(T+1)\} \le 2L_\circ/\rho$, as claimed. If $\mathcal{B}_{\mathrm{slow}}$ is the collection of completed slow-growth batches and $K_{\mathrm{slow}} := |\mathcal{B}_{\mathrm{slow}}|$, then

$$\sum_{B\in\mathcal{B}_{\mathrm{slow}}}\sqrt{(A_B+\Lambda)\Lambda} \le \sqrt{K_{\mathrm{slow}}\Lambda\left\{\sum_B A_B + K_{\mathrm{slow}}\Lambda\right\}}, \qquad K_{\mathrm{slow}} \le \frac{2DL_\circ}{\rho}. \tag{79}$$

Combining this inequality with the initial contribution, (77), and $\Lambda \le 8L_\circ$ gives

$$S_{\mathrm{rad}} \le 2^{13}L_\circ^4\left(\sqrt{\frac{D}{\rho}(U+D)} + \frac{D}{\rho}\right). \tag{80}$$

B.3.5 VARIANCE OF THE STORED LABELS

We bound the conditional-variance sum U by (85). From (63), exact expansion gives

$$\mathrm{Var}(Y_{k,i+1} \mid \mathcal{F}_{k,i}) = \mathbb{E}[Y_{k,i+1}^2 \mid \mathcal{F}_{k,i}] - Y_{k,i}^2 + 2Y_{k,i}\Delta_{k,i} - \Delta_{k,i}^2. \tag{81}$$

For every actual visit (k, i), the correspondence (58) and the identity (62) give

$$(\sigma^{u_{k,i}}_{j(k,i)})^2 = \mathrm{Var}(X_{k,i} \mid \mathcal{F}_{k,i}) = \mathrm{Var}(Y_{k,i+1} \mid \mathcal{F}_{k,i}). \tag{82}$$

Thus the pair-local variance sum has the episode-path representation

$$U = \sum_{k=1}^{N} \sum_{i=1}^{\nu_k} \mathrm{Var}(Y_{k,i+1} \mid \mathcal{F}_{k,i}). \tag{83}$$

The padded indices contribute zero. Taking conditional expectations over an episode and summing in i telescopes the square terms exactly:

$$\mathbb{E}\left[\sum_{i=1}^{\ell} \{\mathbb{E}[Y^2_{k,i+1} \mid \mathcal{F}_{k,i}] - Y^2_{k,i}\} \middle| \mathcal{F}_{k,0}\right] = q_k - Q^+_k(u_\circ)^2 \le q_k.$$

On the start-measurable confidence event $\mathcal{E}^{\mathrm{d}}_k$, $0 \le Y_{k,i} \le 1$ and $\Delta_{k,i} \ge 0$, so $Y_{k,i}\Delta_{k,i} \le \Delta_{k,i}$ and

$$\mathbb{E}\left[\sum_{i=1}^{\ell} \mathrm{Var}(Y_{k,i+1} \mid \mathcal{F}_{k,i}) \middle| \mathcal{F}_{k,0}\right] \le q_k + 2w_k. \tag{84}$$

Let $Z^{\mathrm{v}}_k = \sum_{i=1}^{\ell} \mathrm{Var}(Y_{k,i+1} \mid \mathcal{F}_{k,i})$ along the realized episode. Then $0 \le Z^{\mathrm{v}}_k \le \ell$, its conditional mean given $\mathcal{F}_{k,0}$ is at most $q_k + 2w_k$ by (84), and $U = \sum_k Z^{\mathrm{v}}_k$. The second inequality in (46), applied chronologically to Z^{v}_k/ℓ and unioned over the T possible prefixes, therefore yields, on the remaining half of the variance-conversion allocation,

$$U \le 2Q + 4W + \ell \log \frac{16(T+1)}{\delta_0} \le \widetilde{O}(Q + 2W + \ell). \tag{85}$$

B.3.6 Solving the Bellman-gap inequality

We combine the preceding bounds to prove (47).

We next connect the two contributions in (67) to the Bellman gaps. At a visit with a completed batch (k, i), (62) gives $\Delta_{k,i} = Q^+_k(u_{k,i}) - \mu_{k,i}$. Because Q^+_k is the minimum of its previous value and the upper estimate from the current frozen batch, the confidence event (54) gives

$$Q^+_k(u_{k,i}) \le \frac{1}{|B_{k,i}|} \sum_{j \in I_{B_{k,i}}} X_j^{u_{k,i}} + r_{k,i} \le \overline{\mu}_{k,i} + 2r_{k,i}.$$

Consequently,

$$\Delta_{k,i} \le d_{k,i} + 2r_{k,i}. \tag{86}$$

At the first visit to each pair, $\Delta_{k,i} \le 1$. Every later visit has a completed batch because the initial target size is one and the batch is promoted at the end of the first visit's episode.

Let C_k be the sum along the realized episode-k path of $\min\{1, d_{k,i} + 2r_{k,i}\}$, with value one at a first visit. Then $0 \le C_k \le \ell$, and Bellman telescoping with (86) gives $w_k \le \mathbb{E}[C_k \mid \mathcal{F}_{k,0}]$. Hence the definitions of the two contributions and the number of first visits give

$$\sum_{k=1}^{N} C_k \le S_{\mathrm{mean}} + 2S_{\mathrm{rad}} + D. \tag{87}$$

Applying the first inequality in (46) to C_k/ℓ, simultaneously over all prefixes, gives on the remaining half of the batch-mean conversion allocation

$$W \le 2S_{\mathrm{mean}} + 4S_{\mathrm{rad}} + 2D + 2\ell \log \frac{16(T+1)}{\delta_0}. \tag{88}$$

Substituting (73), (80), and (85) into (88) gives

$$W \le 2^{18} L^8_\circ \sqrt{(D/\rho)(Q + W + D\ell)} + 2^{18} L^8_\circ (D/\rho + \ell) + 8\rho\ell(Q + W). \tag{89}$$

To verify the displayed constants, every logarithm in (73), (80), (85), and (88) is at most $L_\circ$. Hence

$$U + D \le 4L_\circ(Q + W + D\ell).$$

The statistical term in (88) is therefore at most the first term in (89) plus $2^{16}L_\circ^4 D/\rho$. The remaining terms are at most

$$8\rho\ell(Q+W) + 16D/\rho + 4\rho\ell L_\circ + 2D + 2\ell L_\circ,$$

which are bounded by the last two terms of (89) because $D, \ell, L_\circ \ge 1$ and $\rho = (32\ell)^{-1} \le 1$.

Here $\ell = H$ and $\rho = (32H)^{-1}$, so the final term is $(Q+W)/4$. Put $X = D\ell + Q + 1$. Splitting the square root into its W and X parts and using Young's inequality gives

$$2^{18}L_\circ^8\sqrt{32D\ell W} \le \frac{W}{4} + 2^{41}L_\circ^{16}D\ell, \qquad 2^{18}L_\circ^8\sqrt{32D\ell X} \le 2^{21}L_\circ^8 X.$$

Moving the resulting $W/4$ and the existing $W/4$ to the left gives

$$W \le 2^{44}L_\circ^{16}\{D\ell + Q + 1\}. \tag{90}$$

Fix any $p \in (0,1]$ such that every executed episode satisfies $Q_k^+(u_\circ) \ge p$. On the same event,

$$Np \le Q + W. \tag{91}$$

The first inequality in (46), unioned over all prefixes and applied to the novelty-hit indicators, gives on the remaining event of failure probability $\delta_0/4$

$$Q \le 2\sum_k I_k + 2\log\frac{8(T+1)}{\delta_0} = \widetilde{O}(R_N + 1).$$

To obtain the deterministic episode bound, put $B_\circ = D\ell + R_N + \log(4/\delta_0)$. The logarithm in the novelty-count bound is at most $L_\circ$, so $Q \le 2R_N + 2L_\circ \le 2L_\circ B_\circ$. Also $D\ell + 1 \le B_\circ$ and $L_\circ \ge 1$. Consequently (90) gives

$$Q + W \le 2L_\circ B_\circ + 2^{44}L_\circ^{16}(B_\circ + 2L_\circ B_\circ) \le 2^{46}L_\circ^{17}B_\circ.$$

Combining this inequality with (91) and the ceiling in (11), using $R_N \le R$, gives the claimed integer bound simultaneously for every such p: all preceding concentration events and estimates are independent of p. The four allocated failure probabilities sum to at most δ_0. Within the batch-mean allocation, split the budget equally between (73) and (88); within the variance allocation, split it equally between (76) and (85). The factors 16 in the displayed logarithms already dominate these substitutions.

B.4 Proof of the discovery certificate

For Procedure 1, this subsection proves (97), the design-weighted loss bound used in Lemma B.10. The dyadic advantage intervals and novelty inclusion are defined in Appendix B.1; the episode bound and statewise optimism were proved in Appendix B.3.

Throughout this subsection, $S_h \in \mathcal{S}_h$ denotes the state at layer $h \in [H+1]$ of a single fresh evaluation trajectory, with the episode superscript in $S_h^{(t)}$ omitted. Thus h indexes layers, while $S_1 \sim \mu$ is the initial state and $S_{H+1} = \perp$ is terminal. The trajectory law is specified below for each comparison policy.

A policy for measuring the restriction loss. Assume $\eta \le H$ and use the parameters in Appendix A.1. For a realized final set Z, let the policy $\beta_Z \in \Pi(Z)$ use $\pi^\star$ at the root and on Z. Define the nonnegative functions $D_h, \alpha_h : \mathcal{S}_h \to [0,H]$ by $D_h = V_h^\star - V_h^{\beta_Z}$ and $\alpha_h(s) = V_h^\star(s) - Q_h^\star(s, \beta_{Z,h}(s))$. The Markov property decomposes the remaining loss D_h into the current action's loss α_h and the subsequent loss:

$$D_h(s) = \alpha_h(s) + \mathbb{E}[D_{h+1}(S_{h+1}) \mid s, \beta_{Z,h}(s)], \qquad D_{H+1} = 0. \tag{92}$$

Let $\mathbb{P}_{\mathrm{des},\beta_Z}$ be the law of a fresh trajectory using root design and then β_Z, and $\mathbb{E}_{\mathrm{des},\beta_Z}$ its expectation.

Lemma B.10 (Certificate for SimultaneousDiscovery)**.** *On the events of Lemmas B.5 and B.8, every* Certified *output satisfies*

$$\mathbb{E}_{\mathrm{des},\beta_Z} D_2(S_2) \le \frac{\eta}{8\overline{\Gamma}}. \tag{93}$$

Lemmas B.11 and B.12 prove this by grouping omitted states according to the dyadic advantage intervals defined after (38) and bounding the probability of hitting each resulting state set; (97) sums the resulting losses.

Lemma B.11 (Hitting probabilities of frozen novelty sets)**.** *On the event of Lemma B.5 and the uniform discovery event of Lemma B.8, suppose Procedure 1 certifies a final set Z_f after N_{d} episodes. For every $j \in \{0, \dots, J-1\}$ there is a measurable layer-tagged set $B_j \subseteq \bigcup_{h=2}^{H} \mathcal{S}_h$ with $\{z \in \mathcal{S}_h \setminus Z_f : \alpha_h^0(z) \geq r_j\} \subseteq B_j$ for every $h \in \{2, \dots, H\}$, and satisfying*

$$\mathbb{P}_{\mathrm{des},\beta_{\mathrm{f}}}\{\textit{hit } B_j\} \leq p_j, \tag{94}$$

where β_f uses $\pi^\star$ at the root and on Z_f, and π^0 elsewhere. Here $\mathbb{P}_{\mathrm{des},\rho}$ denotes the law of a fresh evaluation trajectory using root design and then a fixed Markov policy ρ.

Proof. After all updates for episode n, the discovery state is $(Z_{n+1}, W_{n+1}, Q^+_{n+1})$, with $n = 0$ denoting initialization. For the filtration of completed-episode histories, define the stopping time

$$\tau_j = \min\{n \in \{0, \dots, N_{\mathrm{d}}\} : Q^+_{n+1}(u_\circ) \leq p_j\}.$$

It exists because certification gives $Q^+_{N_{\mathrm{d}}+1}(u_\circ) < p_\eta < p_j$. Every one of the τ_j preceding episodes has pre-episode score greater than p_j. The uniform episode bound therefore gives

$$\tau_j \leq \left\lceil \frac{\mathsf{S}_{\mathrm{d}}}{p_j} \right\rceil, \qquad \tau_j \leq n_j.$$

This episode bound uses only scores and the discovery event; it does not assume (42). Freeze $B_j = \mathcal{G}(Z_{\tau_j+1}, W_{\tau_j+1})$. Equation (42), applied with $n = \tau_j$, implies that every $z \in \mathcal{S}_h \setminus Z_{\tau_j+1}$ with $\alpha_h^0(z) \geq r_j$ lies in B_j.

The novelty sets shrink pointwise, and each newly stored state was novel immediately before insertion. Consequently

$$Z_f \setminus Z_{\tau_j+1} \subseteq B_j. \tag{95}$$

Let the comparison policy β_j use $\pi^\star$ at the root and on Z_{τ_j+1} and π^0 elsewhere. The policies β_j and β_f can differ only on the left side of (95). Under the same root design their stopped trajectory laws up to the first B_j hit are therefore equal. Discovery optimism, proved in Appendix B.3 at (57), yields

$$\mathbb{P}_{\mathrm{des},\beta_{\mathrm{f}}}\{\text{hit } B_j\} = \mathbb{P}_{\mathrm{des},\beta_{\mathrm{j}}}\{\text{hit } B_j\} \leq Q^+_{\tau_j+1}(u_\circ) \leq p_j.$$

The equality is a deterministic statement after fixing the complete learning history: the comparison concerns fresh evaluation trajectories, and statewise optimism applies to β_j after episode τ_j. No concentration statement is conditioned on the future final set. Finally, $Z_{\tau_j+1} \subseteq Z_f$, so every state omitted from Z_f is also omitted from Z_{τ_j+1}, proving the asserted containment. □

Lemma B.12 (Loss bound for one dyadic advantage interval)**.** *Fix a deterministic Markov policy β, and define the nonnegative functions $\alpha_h^\beta, D_h^\beta : \mathcal{S}_h \to [0, H]$ by*

$$\alpha_h^\beta(s) = V_h^\star(s) - Q_h^\star(s, \beta_h(s)), \qquad D_h^\beta(s) = V_h^\star(s) - V_h^\beta(s).$$

Let $r > 0$, and let $E, B \subseteq \bigcup_{h=2}^{H} \mathcal{S}_h$ be measurable layer-tagged sets with $E \subseteq B$ and $r \leq \alpha_h^\beta(s) < 2r$ on $E \cap \mathcal{S}_h$. On a fresh trajectory with law $\mathbb{P}_{\mathrm{des},\beta}$ define the nonnegative random variable $L_E = \sum_{h=2}^{H} \alpha_h^\beta(S_h)\mathbf{1}\{S_h \in E\}$. Then

$$\mathbb{E}_{\mathrm{des},\beta} L_E \leq H \min\{1, 2r\} \mathbb{P}_{\mathrm{des},\beta}\{\textit{hit } B\}. \tag{96}$$

Proof. Let σ be the first layer at which the fresh trajectory hits B, with $\sigma = H + 1$ if it never does. Since $E \subseteq B$, there is no loss from states in E before σ, and pathwise $L_E \leq 2Hr\mathbf{1}\{\sigma \leq H\}$. A second bound uses the standing return scale. Nonnegativity gives

$$L_E \leq \mathbf{1}\{\sigma \leq H\} \sum_{h=\sigma}^{H} \alpha_h^\beta(S_h).$$

Let $\mathcal{H}_\sigma$ be the stopped history through the observation of S_σ, before the action at that layer. The Markov property and the Bellman loss recursion (92) give, on $\{\sigma \le H\}$,

$$\mathbb{E}_{\mathrm{des},\beta}\left[\sum_{h=\sigma}^{H} \alpha_h^\beta(S_h) \,\middle|\, \mathcal{H}_\sigma\right] = D_\sigma^\beta(S_\sigma) \le H.$$

Thus $\mathbb{E}L_E \le H\mathbb{P}\{\sigma \le H\}$ as well as $\mathbb{E}L_E \le 2Hr\mathbb{P}\{\sigma \le H\}$. Taking the smaller of these two bounds proves the claim. Only the expected suffix advantage sum is bounded by H; no pathwise bound on that sum is assumed. □

Proof of Lemma B.10. Fix a certified output Z_f on the events in Lemma B.11, and let β_f be its comparison policy. Write $D_h = D_h^{\beta_f}$ for its nonnegative loss function. Its default advantage equals α_h^0 outside Z_f and is zero on Z_f. For each $j \in \{0, \dots, J-1\}$ define the measurable set

$$E_j = \bigcup_{h=2}^{H} \{s \in \mathcal{S}_h \setminus Z_f : r_j \le \alpha_h^0(s) < 2r_j\}.$$

Each E_j collects the omitted states whose advantages lie in the jth dyadic advantage interval. These sets contain all omitted states of advantage at least ζ; the last interval includes H, even when H/ζ is a power of two. By Lemma B.11, $E_j \subseteq B_j$ and the fresh trajectory hits B_j with probability at most p_j. Therefore Lemma B.12 gives $\mathbb{E}_{\mathrm{des},\beta_f} L_{E_j} \le H w_j p_j$. The remaining advantage is less than ζ per visited layer. Applying the Bellman loss recursion (92) under root design and then summing yields

$$\begin{aligned} \mathbb{E}_{\mathrm{des},\beta_f} D_2(S_2) &= \mathbb{E}_{\mathrm{des},\beta_f} \sum_{h=2}^{H} \alpha_h^{\beta_f}(S_h) \\ &\le H\zeta + \sum_{j=0}^{J-1} H w_j p_j \le \frac{\eta}{16\overline{\Gamma}} + \frac{J\eta}{16\overline{\Gamma} J_{\mathrm{d}}} \le \frac{\eta}{8\overline{\Gamma}}. \end{aligned} \tag{97}$$

The events in Lemmas B.5 and B.8 each fail with probability at most $\delta/8$. Their intersection proves the certificate with probability at least $1 - \delta/4$. Every omitted state belongs to some E_j or has advantage below ζ. The support cover is used only to prove (43). □

Proof of Theorem 5.2. The confidence allocation in Appendix A.1 gives the joint event of Lemmas B.5 and B.8. The latter lemma, with $D = 1 + AM$, $\ell = H$, $R = M$, and $p = p_\eta$, proves the episode bound. Let $\beta_Z \in \Pi(Z)$ use $\pi^\star$ at the root and on Z. Lemma B.10 bounds its expected loss after root-design sampling by $\eta/(8\overline{\Gamma})$. Applying (35) to the nonnegative differences $Q_1^\star - Q_1^{\beta_Z}$ transfers this bound to the optimal root action:

$$J^\star - J_Z^\star \le J^\star - J(\beta_Z) \le \frac{\eta}{8} + 2(\sqrt{\overline{\Gamma}} + 1)\kappa \le \eta.$$

The final inequality uses $\eta \ge \mathsf{C}_\kappa \chi \ge \mathsf{C}_\kappa \kappa$ from (14). □

B.5 Batch-profile analysis for restricted learning

For Procedure 4, we prove the finite-state regret bound (114) and squared-gap bound (117). The proof extends the batch-profile analysis of Zhang et al. (2025) to finite acyclic control graphs indexed by feasible state–action pairs. All statements hold conditionally on a history fixing the graph, its outcome laws, and the budget, provided each subsequent outcome has its prescribed law conditional on the preceding history and selected pair. Averaging the conditional failure probability therefore covers the random graph returned by discovery.

B.5.1 Estimator and optimistic backup

For Procedure 4, we establish (98) and hence optimism of the value estimates.

Use the finite acyclic control model, joint outcome laws P_u, disjoint doubling batches, range $\mathcal{R} \ge 2H$, and estimator (22) from Appendix A.4. In particular, $\widehat{E}_u$, E_u, and $\widehat{\mathrm{Var}}_u$ denote the empirical mean, population mean, and biased empirical variance on the specified frozen batch. The estimates V^k and graph policies ρ_k are given by (23). Fix the graph, $D \ge 1$, budget $N \ge 1$, and confidence $\delta \in (0, 1)$ throughout this subsection.

Monotonicity of the envelope. The envelope in (22) is coordinatewise nondecreasing. Indeed, where the constant bonus branch is active its derivative is zero and the empirical-mean term has derivative $1/n$; on the variance branch the activation condition gives $\sqrt{\widehat{\mathrm{Var}}_u(x)} \geq (20/3)\mathcal{R}\sqrt{\ell/n}$. Hence, at a differentiability point, its derivative in observation x_j is bounded below by

$$\frac{1}{n} - \frac{20}{3}\sqrt{\frac{\ell}{n}}\frac{\mathcal{R}}{n\sqrt{\widehat{\mathrm{Var}}_u(x)}} \geq 0.$$

Continuity handles equality of the two bonus branches and zero-variance points.

Scalar concentration and optimism. For $n \geq 2$, the scalar empirical Bernstein inequality of Maurer & Pontil (2009, Theorem 4), rescaled to $[0, \mathcal{R}]$, gives for a fixed target x and $a = \log(2/\delta_{\mathrm{eb}})$, where $\delta_{\mathrm{eb}} \in (0, 1)$,

$$E_u x \leq \widehat{E}_u x + \sqrt{\frac{2\widehat{\mathrm{Var}}_u(x)a}{n-1}} + \frac{7\mathcal{R}a}{3(n-1)}$$

with probability at least $1 - \delta_{\mathrm{eb}}$. Since $n/(n-1) \leq 2$ and ℓ dominates every allocated a, each of the two added terms is at most half of the corresponding branch in (22); their sum is bounded by the maximum of those branches. For $n = 1$, the constant branch exceeds $\mathcal{R}$. Let $V^\star$ be the optimal value function of the fixed control graph and put $x_u^\star(r, y) = r + V^\star(y)$. These D targets are deterministic conditional on the fixed control graph and its outcome laws. Each pair can be represented by an independent i.i.d. sequence of outcomes, with disjoint blocks of sizes $1, 2, 4, \ldots$. Concentration is applied to these fixed blocks, without conditioning on whether the algorithm completes them. With $\delta_{\mathrm{eb}} = \delta/[4D(1 + \lceil \log_2 N \rceil)]$, a union bound over these targets and all batch levels gives an event $\mathcal{E}_{\mathrm{opt}}$ of probability at least $1 - \delta/4$ on which

$$E_u x_u^\star \leq f_{u,n}(x_u^\star) \qquad (u \in \mathcal{U},\ n \text{ a completed batch size}). \tag{98}$$

Monotonicity and reverse induction then give $V^k \geq V^\star$. This does not assert the generally false inequality $V^k(x) \geq E_u[R + V^k(Y)]$.

B.5.2 Batch-profile bound

We prove (104) for the estimates and greedy policies used by Procedure 4.

Episode histories and the three totals. Throughout this subsection, $k \in [N]$ indexes episodes. Let S_k be the random state at which episode k enters the control graph, with terminal entry allowed. For Procedure 4, $S_k \in Z \cup \{\perp\}$ is the first state in Z reached after the root, or $\perp$ if none is reached. Thus S_k is distinct from the initial state $S_1^{(k)} \sim \mu$, and its layer may vary across episodes.

The value function and greedy policy are fixed before the episode enters the control graph. Let $\mathcal{H}_k^{\mathrm{pre}}$ be the complete external prefix ending at the entry state S_k, and put $\mathcal{G}_{k,0} = \mathcal{F}_{k-1} \vee \sigma(\mathcal{H}_k^{\mathrm{pre}})$, where $\mathcal{F}_{k-1}$ contains the completed first $k-1$ episodes. The state S_k is $\mathcal{G}_{k,0}$-measurable. The external prefix is nonanticipating: conditional on the complete history before any segment outcome and the selected pair u, that outcome still has law P_u. In the table representation below, the prefix may use consumed entries and independent external randomness, but it reveals no information about unused table entries. Let $\mathcal{F}_{k,i}$ be the history before the ith segment outcome, $X_{k,i}$ the visited state, and $u_{k,i}$ the selected pair. All sums over i range over the segment visits actually made in episode k, and sums over (k, i) also range over $k \in [N]$. Let $n_k(u) \in \mathbb{N} \cup \{0\}$ be the size of the frozen batch used for u in episode k, with $n_k(u) = 0$ before the first completion. At the start of episode k, define $\iota_k(u) = 0$ when u has no frozen batch and $\iota_k(u) = r + 1$ when its frozen batch is the size-2^r batch. A batch used in an episode was completed in a preceding episode, so $\iota_k(u) \leq \lceil \log_2 N \rceil$. The realized *batch profile* is the random array $P = (\iota_k(u) : k \leq N, u \in \mathcal{U})$. Every coordinate is nondecreasing and can increase only to the next level. Thus P belongs to the finite set $\mathcal{I}$ of monotone paths starting at zero in $\{0, 1, \ldots, \lceil \log_2 N \rceil\}^D$, with

$$|\mathcal{I}| \leq (4DN)^{D\lceil \log_2 N \rceil + 1}. \tag{99}$$

Indeed, the path has at most $D\lceil \log_2 N \rceil$ coordinate increments and can be encoded, after padding, by their episode–coordinate pairs. Define the Bellman residual

$$\Delta_{k,i} = V^k(X_{k,i}) - E_{u_{k,i}}[R + V^k(Y)]. \tag{100}$$

For a scalar x, write $(x)_+ = \max\{x, 0\}$. Define the value-estimation gap, positive-residual total, and conditional-variance total by

$$L_N = \sum_{k=1}^{N} \big(V^k(S_k) - V^{\rho_k}(S_k)\big), \tag{101}$$

$$W_N = \sum_{k=1}^{N} \mathbb{E}_{\rho_k}\left[\sum_i (\Delta_{k,i})_+ \mid \mathcal{G}_{k,0}\right], \tag{102}$$

$$U_N = \sum_{k=1}^{N} \mathbb{E}_{\rho_k}\left[\sum_i \mathrm{Var}_{u_{k,i}}(R + V^k(Y)) \mid \mathcal{G}_{k,0}\right]. \tag{103}$$

Put

$$L_{\mathrm{prof}} = 1 + \log \frac{64D(N+1)}{\delta}.$$

Lemma B.13 (Batch-profile bound)**.** *With probability at least* $1 - \delta$*,*

$$0 \le L_N \le W_N \le \widetilde{O}(\sqrt{DU_N} + \mathcal{R}D). \tag{104}$$

There is no multiplicative graph-depth or endpoint-support factor beyond the feasible-pair count D*. The condition* $\mathcal{R} \ge 2H$ *ensures that* $R + V^k(Y) \in [0, \mathcal{R}]$*. Each pair contributes at most* H *before its first batch is completed.*

We prove Lemmas B.14 and B.15, then combine them with the conditional-mean conversion in Lemma B.7 to obtain Lemma B.13.

Profiles and candidate targets. For every $u \in \mathcal{U}$, prepare an independent table of i.i.d. joint outcomes, each with law P_u, and consume its next unused entry on each visit. The nonanticipating external prefix reveals no unused entry. Number the disjoint table blocks of sizes $1, 2, 4, \ldots$ by $0, 1, 2, \ldots$. Fix a deterministic profile $I \in \mathcal{I}$. In episode k, the auxiliary dynamic program uses the table block selected by coordinate $I_{k,u}$: block $I_{k,u} - 1$ of size $2^{I_{k,u}-1}$ if $I_{k,u} > 0$, or the initial value H if $I_{k,u} = 0$, regardless of whether I is realized. Its active counts are not inputs to the value recursion. They are reconstructed from the visits only when coupling this auxiliary execution to the actual one. For each pair u, index the candidate targets by the deterministic labels $0, 1, \ldots, N$, with label zero denoting the zero function. The set

$$\mathcal{X}_{u,I} = \{x_u^{k,I} : k \le N\} \cup \{0\}, \qquad x_u^{k,I}(r, y) = r + V^{k,I}(y),$$

contains at most $N + 1$ functions of joint outcomes. Write $X_u^{k,I} = x_u^{k,I}$ for the corresponding backup target. For each fixed label, the target is measurable with respect to the tables of pairs reachable strictly after u in the acyclic control graph (the strict-successor tables). Conditional on these tables, it is therefore independent of the current pair's batch. This uses independence between pair tables and preserves dependence between reward and endpoint within an outcome. To control positive deviations, allow a separate target-or-zero choice for every pair. The resulting selector class is

$$\mathcal{C}_I = \prod_{u \in \mathcal{U}} \mathcal{X}_{u,I}, \qquad |\mathcal{C}_I| \le (N+1)^D, \qquad \log|\mathcal{C}_I| \le D \log(N+1). \tag{105}$$

Lemma B.14 (Uniform deviations over batch profiles)**.** *There is an event* $\mathcal{E}_{\mathrm{prof}}$ *with probability at least* $1 - \delta/4$ *on which, for every* $I \in \mathcal{I}$*, every selector* $x = (x_u)_{u \in \mathcal{U}} \in \mathcal{C}_I$*, and every common doubling batch size* $n \le N$*,*

$$\sum_{u \in \mathcal{U}} [(\widehat{E}_u - E_u) x_u]_+ \le 16 L_{\mathrm{prof}} \sqrt{\frac{D}{n} \sum_u \mathrm{Var}_u(x_u)} + 64 \frac{\mathcal{R} D L_{\mathrm{prof}}^2}{n}. \tag{106}$$

Here $\widehat{E}_u$ *is computed on pair* u*'s unique size-*n *table block, whether or not that block is completed in the actual execution. The same bound holds with* $\widehat{E}_u - E_u$ *replaced by* $E_u - \widehat{E}_u$*. Both bounds also hold for the squared targets, with range* $\mathcal{R}^2$ *and variances* $\mathrm{Var}_u(x_u^2)$*.*

Proof. Step 1. Reverse exposure for a fixed profile and selector. Fix a deterministic label vector in $\{0,\dots,N\}^{\mathcal{U}}$, producing a selector $x=(x_u)_{u\in\mathcal{U}}\in\mathcal{C}_I$, and one batch size n. Under reverse exposure, expose the n entries of the size-n block of each pair one at a time, ordering pairs in reverse topological order. After this block, expose the rest of that pair's table needed by the auxiliary dynamic program, assigning zero martingale increments to those additional entries. Thus all strict-successor tables needed to define x_u have been exposed before the block of u, while no entry of that block has been exposed. Only finitely many table entries occur in a fixed N-episode profile. Let $\mathcal{H}^{\mathrm{exp}}_{u,j-}$ be the sigma-field immediately before the jth entry of this block of u is exposed. The increment $Z_{u,j}=n^{-1}\{x_u(O_{u,j})-E_u x_u\}$ satisfies

$$\mathbb{E}[Z_{u,j}\mid\mathcal{H}^{\mathrm{exp}}_{u,j-}]=0,\qquad |Z_{u,j}|\le\mathcal{R}/n,\qquad \mathbb{E}[Z^2_{u,j}\mid\mathcal{H}^{\mathrm{exp}}_{u,j-}]=\mathrm{Var}_u(x_u)/n^2.$$

Thus $\sum_u(\widehat{E}_u-E_u)x_u$ is a martingale sum with total predictable variance $\mathcal{V}_x=n^{-1}\sum_u\mathrm{Var}_u(x_u)$. Fix a scalar $\alpha\ge1$. Partition the variance range into the initial variance interval $[0,\mathcal{R}^2/n^2]$ and the *dyadic variance intervals* $(2^{b-1}\mathcal{R}^2/n^2,2^b\mathcal{R}^2/n^2]$ for $b=1,\dots,\lceil\log_2(Dn)\rceil$. When $\mathcal{V}_x$ lies in a dyadic variance interval, that interval's upper endpoint is at most $2\mathcal{V}_x$. Bernstein–Freedman with that deterministic upper endpoint gives, with failure probability $e^{-\alpha}$,

$$\sum_u(\widehat{E}_u-E_u)x_u\le 2\sqrt{\frac{\alpha}{n}\sum_u\mathrm{Var}_u(x_u)}+\frac{3\mathcal{R}\alpha}{n}. \tag{107}$$

The same display holds on the initial variance interval: Freedman's square-root term is at most $\sqrt{2\alpha}\mathcal{R}/n$, which is absorbed by the displayed linear term because $\alpha\ge1$. Since $\mathcal{V}_x\le D\mathcal{R}^2/n$, these $1+\lceil\log_2(Dn)\rceil$ intervals cover every possibility. For a sigma-field $\mathcal{H}$, scalar $b>0$, and random variable Z with $\mathbb{E}[Z\mid\mathcal{H}]=0$ and $Z\le b$, the conditional moment bound $\mathbb{E}[e^{\lambda Z}\mid\mathcal{H}]\le\exp\{\lambda^2\mathbb{E}[Z^2\mid\mathcal{H}]/[2(1-\lambda b/3)]\}$ proves the Bernstein–Freedman bound on each variance interval by iteration, Markov's inequality, and optimization over $0\le\lambda<3/b$. The displayed constants account for replacing the variance by the interval's upper endpoint.

Step 2. A simultaneous event over profiles and selectors. Union-bound over the deterministic selector labels underlying $\mathcal{C}_I$, the batch profiles in (99), at most $1+\lceil\log_2 N\rceil$ batch levels, and at most $2+\lceil\log_2(DN)\rceil$ variance intervals, including the initial interval. Include both signs for the targets and for their squares, a factor of four. No additional union bound over reuse positions is needed: every such position selects one of the already included label vectors. The logarithm of this event count, including the factor $4/\delta$ for its failure allocation, is at most

$$D\{\lceil\log_2 N\rceil+1\}\log(4DN)+D\log(N+1)+3\log\{4D(N+1)\}+\log(16/\delta)\le 16DL^2_{\mathrm{prof}}.$$

Taking $\alpha=16DL^2_{\mathrm{prof}}$ in (107) constructs an event $\mathcal{E}_{\mathrm{prof}}$ of probability at least $1-\delta/4$ on which all profiles, selectors, levels, and variance intervals are covered for both signs of the targets and their squares. The remainder of the profile calculation is pathwise on $\mathcal{E}_{\mathrm{prof}}$. Choose the zero target at coordinates with negative deviation and retain the selected target at the other coordinates. Equation (107), with the preceding choice of α, gives (106): zeroing a coordinate also removes its nonnegative variance. Enlarging the remaining variance sum proves the claim for every selector. Reversing the sign gives the lower tail; applying the argument to squared targets uses $\mathrm{Var}(x^2)\le4\mathcal{R}^2\,\mathrm{Var}(x)$. □

Applying the event to the realized profile. Couple the auxiliary execution for I and the actual execution using the same external randomness and outcome tables. On $\{P=I\}$, their frozen batches agree initially. If they agree before episode k, reverse induction gives equal values and greedy actions; the coupled executions then consume the same entries, visit the same pairs, and promote the same batches after episode k. Induction over episodes proves equality of their values, actions, active counts, and frozen batches. Therefore the uniform event in Lemma B.14 applies with $I=P$. This is a pathwise substitution into a simultaneous event; it does not condition the concentration argument on $\{P=I\}$.

Lemma B.15 (Positive residuals along realized visits)**.** *Define the nonnegative scalar* $U^{\mathrm{path}}_N=\sum_{k,i}\mathrm{Var}_{u_{k,i}}(X^k_{u_{k,i}})$ *over all realized visits. This is a sum of population variances, not an empirical variance estimate. On* $\mathcal{E}_{\mathrm{prof}}$,

$$\sum_{k,i}(\Delta_{k,i})_+\le 2^7L^{3/2}_{\mathrm{prof}}\sqrt{DU^{\mathrm{path}}_N}+2^{14}L^3_{\mathrm{prof}}\mathcal{R}D. \tag{108}$$

Proof. Throughout this proof, X^k at visit (k,i) abbreviates $X^k_{u_{k,i}} = X^{k,P}_{u_{k,i}}$. Sums of empirical statistics, bonuses, and $1/n_k(u_{k,i})$ range only over visits with a completed batch; U_N^{path} includes all visits. First visits contribute at most HD in total.

Step 1. Sum deviations over batch reuse. At level r, with $n_r = 2^r$, enumerate the at most $2n_r$ reuse positions of every frozen batch. For $j \le 2n_r$ and $u \in \mathcal{U}$, let $k_{r,j,u}$ be the episode of the jth visit to u while its size-n_r batch is frozen, when that visit exists, and set $x_{r,j,u} = x_u^{k_{r,j,u},P}$; use the zero candidate otherwise. For fixed (r,j) this is a selector from (105), so (106) applies across all pairs. Put $v_{r,j} = \sum_u \mathrm{Var}_u(x_{r,j,u})$ and $U^{\mathrm{path}}_{N,r} = \sum_{j=1}^{2n_r} v_{r,j}$. Each actual visit at level r occurs exactly once in this last sum, and zero candidates contribute zero. Hence

$$\begin{aligned}\sum_{j=1}^{2n_r} 16L_{\mathrm{prof}}\sqrt{Dv_{r,j}/n_r} &\le 16L_{\mathrm{prof}}\sqrt{\frac{D}{n_r}\,2n_r\sum_{j=1}^{2n_r} v_{r,j}}\\ &= 16\sqrt{2}L_{\mathrm{prof}}\sqrt{DU^{\mathrm{path}}_{N,r}},\\ \sum_{j=1}^{2n_r}\frac{64\mathcal{R}DL^2_{\mathrm{prof}}}{n_r} &= 128\mathcal{R}DL^2_{\mathrm{prof}}.\end{aligned}$$

There are at most $1+\lceil\log_2 N\rceil$ levels. Apply $\sum_r\sqrt{U^{\mathrm{path}}_{N,r}} \le \sqrt{(1+\lceil\log_2 N\rceil)U^{\mathrm{path}}_N}$ to obtain

$$\sum_{k,i}[(\widehat{E}-E)X^k]_+,\quad \sum_{k,i}[(E-\widehat{E})X^k]_+ \le 32L^{3/2}_{\mathrm{prof}}\sqrt{DU^{\mathrm{path}}_N} + 256\mathcal{R}DL^3_{\mathrm{prof}}, \tag{109}$$

$$\sum_{k,i}[(\widehat{E}-E)(X^k)^2]_+ \le 64\mathcal{R}L^{3/2}_{\mathrm{prof}}\sqrt{DU^{\mathrm{path}}_N} + 256\mathcal{R}^2DL^3_{\mathrm{prof}}. \tag{110}$$

For the squared-target statement, let X' be an independent copy of $X \in [0,\mathcal{R}]$. Then

$$\mathrm{Var}(X^2) = \tfrac{1}{2}\mathbb{E}[(X^2-(X')^2)^2] \le 4\mathcal{R}^2\,\tfrac{1}{2}\mathbb{E}[(X-X')^2] = 4\mathcal{R}^2\,\mathrm{Var}(X).$$

Thus the same reuse calculation applies with the squared range and variance factor displayed in (110); both displays use $1+\lceil\log_2 N\rceil \le 2L_{\mathrm{prof}}$. Each deviation concerns the scalar $X^k(O)$, so these bounds contain no endpoint-cardinality factor.

Step 2. Empirical variances and bonuses. Let $b_{k,i} \ge 0$ be the maximum of the two bonus branches in (22) for the frozen batch at visit (k,i). For $X \in [0,\mathcal{R}]$, put $m = EX$ and $\widehat{m} = \widehat{E}X$. Expanding the two variances and using $m+\widehat{m} \le 2\mathcal{R}$ gives

$$\begin{aligned}\widehat{\mathrm{Var}}(X) - \mathrm{Var}(X) &= (\widehat{E}-E)X^2 + (m-\widehat{m})(m+\widehat{m})\\ &\le [(\widehat{E}-E)X^2]_+ + 2\mathcal{R}[(E-\widehat{E})X]_+.\end{aligned}$$

Disjoint doubling gives $\sum_{k,i}(n_k(u_{k,i})\vee 1)^{-1} \le D(1+2\log_2(2N))$. Equations (109)–(110) control the deviations in the variance comparison. Sum the constant bonus branch using the inverse-batch-size bound. For the variance branch, Cauchy–Schwarz gives

$$\sum_{k,i}\sqrt{\frac{\widehat{\mathrm{Var}}_{u_{k,i}}(X^k)\ell}{n_k(u_{k,i})}} \le \sqrt{\ell\sum_{k,i}\widehat{\mathrm{Var}}_{u_{k,i}}(X^k)}\sqrt{\sum_{k,i}\frac{1}{n_k(u_{k,i})}}.$$

Put $v = U^{\mathrm{path}}_N$ and $h = \mathcal{R}\sqrt{D}$. Substituting (109)–(110) into the variance comparison gives

$$\begin{aligned}\sum_{k,i}\widehat{\mathrm{Var}}_{u_{k,i}}(X^k) &\le v + 128L^{3/2}_{\mathrm{prof}}h\sqrt{v} + 768L^3_{\mathrm{prof}}h^2\\ &\le (\sqrt{v} + 64L^{3/2}_{\mathrm{prof}}h)^2.\end{aligned}$$

Using $\ell \le 64L_{\mathrm{prof}}$ and $1+2\log_2(2N) \le 4L_{\mathrm{prof}}$, the variance branch contributes at most

$$\frac{320}{3}L_{\mathrm{prof}}\sqrt{D}(\sqrt{v} + 64L^{3/2}_{\mathrm{prof}}h).$$

The constant branch contributes at most $(102400/9)\mathcal{R}DL_{\text{prof}}^2$. Bounding their maximum by their sum gives

$$\sum_{k,i} b_{k,i} \le \frac{320}{3} L_{\text{prof}} \sqrt{DU_N^{\text{path}}} + \left(\frac{20480}{3} L_{\text{prof}}^{5/2} + \frac{102400}{9} L_{\text{prof}}^2 \right) \mathcal{R}D. \tag{111}$$

Step 3. Positive residuals and first visits. Since the chosen action is greedy and the backup is clipped at H, every completed-batch visit satisfies

$$(\Delta_{k,i})_+ \le [(\widehat{E} - E)X^k]_+ + \min\{b_{k,i}, H\}. \tag{112}$$

At a first visit $(\Delta_{k,i})_+ \le H$, and each pair has only one such visit. Combining (109), (111), (112), and $HD \le \mathcal{R}D/2$ proves (108). Indeed, for $L = L_{\text{prof}} \ge 2$, the coefficients of $\sqrt{DU_N^{\text{path}}}$ and $\mathcal{R}D$ are bounded by

$$32L^{3/2} + \frac{320}{3}L \le 2^7 L^{3/2}, \qquad 256L^3 + \frac{20480}{3}L^{5/2} + \frac{102400}{9}L^2 + \frac{1}{2} \le 2^{14}L^3.$$

□

Proof of Lemma B.13. Work on $\mathcal{E}_{\text{opt}}$ and $\mathcal{E}_{\text{prof}}$, so scalar optimism and Lemma B.15 hold.

Step 1. Conditional variance and residual totals. Apply Lemma B.7 to the episode variance total, bounded by $\mathcal{R}^2 D$, and to the positive residual total, bounded by HD. With failure probability $\delta/4$ for each conversion,

$$U_N^{\text{path}} \le 2U_N + \mathcal{R}^2 D \log(4/\delta), \qquad W_N \le 2\sum_{k,i} (\Delta_{k,i})_+ + 2HD \log(4/\delta).$$

Substitute the first inequality into (108), then use the second inequality. Using $\log(4/\delta) \le L_{\text{prof}}$ and $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$ gives

$$\begin{aligned} W_N &\le a_{\text{prof}} \sqrt{DU_N} + b_{\text{prof}} \mathcal{R}D, \\ a_{\text{prof}} &= 2^8 \sqrt{2}\, L_{\text{prof}}^{3/2}, \qquad b_{\text{prof}} = 2^{16} L_{\text{prof}}^3. \end{aligned} \tag{113}$$

Here $a_{\text{prof}}, b_{\text{prof}} > 0$ are deterministic scalars. The additive coefficient before the final bound is at most $2^8 L_{\text{prof}}^2 + 2^{15} L_{\text{prof}}^3 + L_{\text{prof}} \le b_{\text{prof}}$. *Step 2. Loss and probability of success.* Bellman telescoping along the realized sequence of segments gives

$$V^k(S_k) - V^{\rho_k}(S_k) = \mathbb{E}_{\rho_k}\left[\sum_i \Delta_{k,i} \mid \mathcal{G}_{k,0} \right] \le \mathbb{E}_{\rho_k}\left[\sum_i (\Delta_{k,i})_+ \mid \mathcal{G}_{k,0} \right].$$

On $\mathcal{E}_{\text{opt}}$, $V^k \ge V^\star \ge V^{\rho_k}$, so $0 \le L_N \le W_N$. The events $\mathcal{E}_{\text{opt}}$, $\mathcal{E}_{\text{prof}}$, and the two conversions each have failure probability at most $\delta/4$. Their intersection proves (104) with probability at least $1 - \delta$.

□

B.5.3 Regret in the finite control problem

For Procedure 4, the total-return bound turns Lemma B.13 into the regret guarantee (114).

Corollary B.16 (Regret bound for the first-hit control problem)**.** *For a finite acyclic first-hit control graph with $D \ge 1$ feasible pairs, where every outcome contains the nonnegative segment reward jointly with its endpoint, let $\text{Reg}_N = \sum_{k=1}^N \{V^\star(S_k) - V^{\rho_k}(S_k)\}$. The learner satisfies, with probability at least $1 - \delta$,*

$$\text{Reg}_N \le \widetilde{O}(H\sqrt{DN}) \tag{114}$$

for any entry state revealed after the episode policy is fixed.

Proof. Step 1. Bound conditional variance by positive residuals. Run (23) with target $\mathsf{X}_i = R_i + V^k(Z_{i+1})$ and range $\mathcal{R} = 2H$, where Z_i is the ith state in the control graph. For one episode put $Y_i = V^k(Z_i)$ and $\Delta_i = Y_i - \mathbb{E}[R_i + Y_{i+1} \mid \mathcal{F}_{k,i}]$. Since $\sum_i R_i \le H$ and $0 \le Y_i \le H$,

$$\sum_i \mathbb{E}[R_i^2 + 2R_iY_{i+1} \mid \mathcal{G}_{k,0}] \le 3H^2.$$

Pad the path after termination with $R_i = Y_i = \Delta_i = 0$. The exact conditional-variance identity is

$$\mathrm{Var}(R_i + Y_{i+1} \mid \mathcal{F}_{k,i}) = \mathbb{E}[R_i^2 + 2R_iY_{i+1} + Y_{i+1}^2 \mid \mathcal{F}_{k,i}] - Y_i^2 + 2Y_i\Delta_i - \Delta_i^2.$$

Take expectations given the history at entry to the control graph, sum over the padded path, and telescope the $Y_{i+1}^2 - Y_i^2$ terms. The terminal value is zero, so their total is $-Y_1^2 \le 0$; also $2Y_i\Delta_i - \Delta_i^2 \le 2H(\Delta_i)_+$. This proves the episodewise inequality

$$\mathbb{E}_{\rho_k}\left[\sum_i \mathrm{Var}(R_i + Y_{i+1} \mid \mathcal{F}_{k,i}) \mid \mathcal{G}_{k,0}\right] \le 3H^2 + 2H\mathbb{E}_{\rho_k}\left[\sum_i (\Delta_i)_+ \mid \mathcal{G}_{k,0}\right].$$

The mixed term R_iY_{i+1} was retained throughout, so no independence between segment reward and endpoint is required. Thus $U_N \le 3H^2N + 2HW_N$.

Step 2. Solve the profile bound and cap regret. Applying $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$ to (113) gives

$$W_N \le a_{\mathrm{prof}} H\sqrt{3DN} + a_{\mathrm{prof}}\sqrt{2HDW_N} + 2b_{\mathrm{prof}}HD,$$
$$a_{\mathrm{prof}}\sqrt{2HDW_N} \le \tfrac{1}{2}W_N + a_{\mathrm{prof}}^2 HD.$$

Subtracting $W_N/2$ and multiplying by two yields $W_N \le 2\sqrt{3}a_{\mathrm{prof}}H\sqrt{DN} + (2a_{\mathrm{prof}}^2 + 4b_{\mathrm{prof}})HD$. Both coefficients are at most $2^{19}L_{\mathrm{prof}}^3$, so

$$W_N \le 2^{19}L_{\mathrm{prof}}^3\{H\sqrt{DN} + HD\}.$$

Bellman telescoping and optimism give $0 \le \mathrm{Reg}_N \le W_N$, while deterministically $\mathrm{Reg}_N \le HN$. If $D \le N$, then $HD \le H\sqrt{DN}$; if $D > N$, then $HN \le H\sqrt{DN}$. Hence in both cases $\mathrm{Reg}_N \le 2^{20}L_{\mathrm{prof}}^3 H\sqrt{DN}$, proving the displayed bound without an additive HD term. □

B.5.4 Selected episodes and squared continuation gaps

For Procedure 4, we prove the selected-episode regret bound (115) and use it to establish the squared-gap bound (117).

Lemma B.17 (Regret on predictably selected episodes)**.** *Let $b_k \in \{0,1\}$ be $\mathcal{G}_{k,0}$-measurable for every $k \in [N]$, and set $n_b = \sum_k b_k$. All episodes continue to update the batches; b_k selects only the terms included in the regret sum. For every fixed predictable rule, with probability at least $1-\delta$,*

$$\sum_{k=1}^{N} b_k\{V^\star(S_k) - V^{\rho_k}(S_k)\} \le \widetilde{O}(H\sqrt{Dn_b}). \tag{115}$$

The event of Lemma B.14 is uniform over the target selections in (105); only the final martingale conversion requires a union bound over a finite family of episode-selection rules.

Proof. Step 1. Select targets on the uniform profile event. Use the segment-reward target from Corollary B.16, so $\mathcal{R} = 2H$. On the event of Lemma B.14, use the realized profile P and the reuse indices $k_{r,j,u}$ from the proof of Lemma B.15. For each level r and reuse position j, select

$$x_{r,j,u} = \begin{cases} b_{k_{r,j,u}} x_u^{k_{r,j,u},P}, & k_{r,j,u} \text{ exists,} \\ 0, & \text{otherwise.} \end{cases}$$

This is one selector across pairs in (105), and its variance sum includes exactly the selected visits at that level and reuse position. Hence no union bound over the 2^N possible indicator sequences is needed. Selection may depend on the frozen batches and the revealed entry state, but must precede

the episode's segment outcomes. The concentration event is uniform over the selected targets and the zero target. The reuse and bonus calculations in Lemma B.15, restricted to selected visits, give

$$\sum_{k,i} b_k(\Delta_{k,i})_+ \le 2^7 L_{\mathrm{prof}}^{3/2}\sqrt{DU_b^{\mathrm{path}}} + 2^{14}L_{\mathrm{prof}}^3\mathcal{R}D, \quad U_b^{\mathrm{path}} = \sum_{k,i} b_k \operatorname{Var}_{u_{k,i}}(X_{u_{k,i}}^{k,P}).$$

Step 2. Convert the selected episode totals. Predictability permits applying Lemma B.7 to the selected episode totals. Define

$$W_b = \sum_k b_k \mathbb{E}_{\rho_k}\left[\sum_i (\Delta_{k,i})_+ \mid \mathcal{G}_{k,0}\right], \qquad U_b = \sum_k b_k \mathbb{E}_{\rho_k}\left[\sum_i \operatorname{Var}_{u_{k,i}}(R + V^k(Y)) \mid \mathcal{G}_{k,0}\right].$$

Let $Z_k^\Delta = b_k \sum_i (\Delta_{k,i})_+$ and $Z_k^v = b_k \sum_i \operatorname{Var}_{u_{k,i}}(R + V^k(Y))$. Conditional on $\mathcal{G}_{k,0}$, their means are the episode summands of W_b and U_b; moreover $0 \le Z_k^\Delta \le HD$ and $0 \le Z_k^v \le \mathcal{R}^2 D$. Allocate $\delta/4$ to each conversion, in addition to the $\delta/4$ allocations for $\mathcal{E}_{\mathrm{opt}}$ and $\mathcal{E}_{\mathrm{prof}}$. Then (46) gives

$$\sum_k Z_k^v \le 2U_b + \mathcal{R}^2 D \log(4/\delta), \qquad W_b \le 2\sum_k Z_k^\Delta + 2HD\log(4/\delta).$$

Substitution into the preceding pathwise selector bound gives $W_b \le a_{\mathrm{prof}}\sqrt{DU_b} + b_{\mathrm{prof}}\mathcal{R}D$, with the coefficients in (113).

Step 3. Apply the segment-reward variance identity. The episode variance identity from Corollary B.16 remains valid after multiplication by b_k and gives $U_b \le 3H^2 n_b + 2HW_b$. The calculation in Corollary B.16 therefore gives

$$\sum_k b_k\{V^\star(S_k) - V^{\rho_k}(S_k)\} \le W_b \le 2\sqrt{3}a_{\mathrm{prof}}H\sqrt{Dn_b} + (2a_{\mathrm{prof}}^2 + 4b_{\mathrm{prof}})HD. \tag{116}$$

The selected regret is also at most Hn_b. Splitting into $D \le n_b$ and $D > n_b$ as in Corollary B.16 proves (115), with explicit right side $2^{20}L_{\mathrm{prof}}^3 H\sqrt{Dn_b}$. □

Corollary B.18 (Squared continuation gaps)**.** *Let $g_k = V^\star(S_k) - V^{\rho_k}(S_k) \in [0, H]$ be the continuation gap at the adaptively revealed episode-k entry state. With probability at least $1-\delta$,*

$$\sum_{k=1}^N g_k^2 \le \widetilde{O}(H^2(D+1)). \tag{117}$$

For the algorithmic invocation $\delta = \delta_{\mathrm{sq}}$, the right side is bounded by (24).

Proof. Step 1. Define dyadic gap intervals and predictable membership. If $N = 1$, then $g_1^2 \le H^2 \le B_2(D, 1, H, \delta)$. Assume henceforth that $N \ge 2$, put $J_{\mathrm{sq}} = \lceil \log_2 \sqrt{N} \rceil$. For each $j \in \{0, \dots, J_{\mathrm{sq}} - 1\}$, the *jth dyadic gap interval* is $(2^{-j-1}H, 2^{-j}H]$; define its predictable membership indicators and episode count by

$$b_{k,j} = \mathbf{1}\{2^{-j-1}H < g_k \le 2^{-j}H\}, \qquad N_j^{\mathrm{sq}} = \sum_k b_{k,j}, \quad 0 \le j < J_{\mathrm{sq}}.$$

Although g_k need not be known to the algorithm, it is $\mathcal{G}_{k,0}$-measurable: the policy is fixed and the entry state is revealed before the segment outcomes of episode k. Hence each $b_{k,j}$ satisfies the predictability hypothesis of Lemma B.17.

Step 2. Count episodes with gaps in each dyadic gap interval. Use $\mathcal{E}_{\mathrm{opt}}$ and $\mathcal{E}_{\mathrm{prof}}$ once, each with failure probability $\delta/4$. For each of the J_{sq} predictable rules, allocate $q = \delta/(4J_{\mathrm{sq}})$ to each of the two conversions in (46). The intersection has probability at least $1 - \delta$. Define the deterministic scalars

$$L_* = 1 + \log\frac{64D(N+1)}{q}, \qquad a_* = 2^8\sqrt{2}\,L_*^{3/2}, \qquad b_* = 2^{16}L_*^3.$$

The envelope retains its original confidence parameter δ; only the conversion logarithms use q. Since $q \le \delta$ and $\ell \le 64L_*$, (116) holds with coefficients a_*, b_* simultaneously for these rules. Put $\epsilon_j = 2^{-j-1} \in (0, 1/2]$. It follows that

$$N_j^{\mathrm{sq}}\epsilon_j \le 2\sqrt{3}a_*\sqrt{DN_j^{\mathrm{sq}}} + (2a_*^2 + 4b_*)D.$$

Young's inequality bounds the square-root term by $N_j^{\mathrm{sq}}\epsilon_j/2 + 6a_*^2 D/\epsilon_j$. Rearranging gives

$$N_j^{\mathrm{sq}}\epsilon_j^2 \le 12a_*^2 D + (4a_*^2 + 8b_*)D\epsilon_j \le (14a_*^2 + 4b_*)D.$$

Step 3. Sum squared gaps and check the algorithmic scale. Since the upper endpoint of the jth dyadic gap interval is $2\epsilon_j H$, the squared gaps of episodes assigned to this interval sum to at most $(56a_*^2 + 16b_*)H^2 D = 2^{23}L_*^3 H^2 D$. Keeping the two coefficients separate in (116) avoids squaring its additive coefficient. All remaining gaps satisfy $g_k \le 2^{-J_{\mathrm{sq}}}H \le H/\sqrt{N}$, so

$$\sum_{k=1}^{N} g_k^2 \le H^2 + 2^{23}J_{\mathrm{sq}}L_*^3 H^2 D = \widetilde{O}(H^2(D+1)).$$

For the algorithmic invocation, $\delta = \delta_{\mathrm{sq}} = \delta_{\mathrm{fin}}$ by Appendix A.5; hence the original envelope is exactly the one used by Procedure 4. Write $L_{\mathrm{sq}} = L_{\mathrm{sq}}(D, N, \delta_{\mathrm{sq}})$ as in (24). Since $J_{\mathrm{sq}} \le L_{\mathrm{sq}}$ and $L_* \le 2L_{\mathrm{sq}}$, and also $\sum_k g_k^2 \le NH^2$ deterministically,

$$\sum_{k=1}^{N} g_k^2 \le H^2 \min\{N,\, 1 + 2^{26}DL_{\mathrm{sq}}^4\} = B_2(D, N, H, \delta_{\mathrm{sq}}).$$

□

B.6 Proof of restricted-class regret

This subsection gives the full proof of Lemma 5.3 for Procedure 4, including its $\widetilde{O}(\kappa\sqrt{\min\{A,d\}N})$ term. We first separate root-action regret from continuation loss, then bound the continuation loss and its effect on the two root learners. The tail estimates use Lemma B.13 and Corollaries B.16 and B.18. For an invocation after discovery, condition throughout on its complete terminal history, so Z and the remaining budget N are fixed. Include this history in the initial sigma-field $\mathcal{F}_0$; the conditional Markov kernels of future episodes remain unchanged by the conditional sampling assumption in Section 3. Let $\mathcal{E}_{\mathrm{res}}$ be the intersection of the tail-profile event, the squared-gap profile event, the continuation L_1 and L_2 conversion events, the root self-normalization event, the SquareCB selected- and full-action prediction events when that branch is used, and the final conversion from realized root states to their conditional expectations. The allocations in Appendix A.5 and a union bound give

$$\mathbb{P}(\mathcal{E}_{\mathrm{res}}) \ge 1 - \delta_0. \tag{118}$$

Each concentration inequality below is proved under the conditional law given the discovery history. The resulting inequalities are then combined on $\mathcal{E}_{\mathrm{res}}$; expectations are not conditioned on that event.

B.6.1 Benchmark and regret decomposition

Let $\rho_Z^\star : Z \to \mathcal{A}$ be a deterministic policy maximizing the value at every state of the finite acyclic segment control problem on Z, and let $V_Z^\star : Z \cup \{\perp\} \to [0, H]$ be its value function, with $V_Z^\star(\perp) = 0$. Extend $\rho_Z^\star$ by π^0 outside Z and by any measurable root rule to obtain a full Markov policy; its root action-value function does not depend on the chosen root rule. Use this extension convention also for each learned tail $\rho_t : Z \to \mathcal{A}$. Assumption 3.1 therefore supplies a vector $\theta_Z^\star \in \mathbb{R}^d$, $\|\theta_Z^\star\|_2 \le B$, and a measurable residual function $e_Z : \mathcal{S}_1 \times \mathcal{A} \to [-\kappa, \kappa]$ such that

$$Q_1^{\rho_Z^\star}(s, a) = \phi_1(s, a)^\top \theta_Z^\star + e_Z(s, a). \tag{119}$$

This is the sole invocation of Assumption 3.1 in restricted learning. The finite-state profile bounds use the true joint segment laws and incur no approximation error.

Let $\mathcal{F}_{t-1}$ be the sigma-field generated by $\mathcal{F}_0$ and the first $t-1$ restricted episodes, put $\mathcal{K}_t = \mathcal{F}_{t-1} \vee \sigma(S_1^{(t)})$, and put $\mathcal{H}_t = \mathcal{K}_t \vee \sigma(A_1^{(t)})$.

The continuation loss $c_t \in [0, H]$ is

$$c_t = Q_1^{\rho_Z^\star}(S_1^{(t)}, A_1^{(t)}) - Q_1^{\rho_t}(S_1^{(t)}, A_1^{(t)}). \tag{120}$$

Statewise optimality of $\rho_Z^\star$ on the acyclic tail implies $Q_1^{\rho_Z^\star}(s, a) \ge Q_1^{\rho_t}(s, a)$ for every root pair, so this definition indeed gives $c_t \in [0, H]$. For a root context s and action a, write $c_t(s, a)$ for the

analogue of (120). Let $p_t(\cdot \mid s) \in \Delta(\mathcal{A})$ be the measurable root-action kernel obtained from the frozen predictor: it is the SquareCB distribution when $A \le d$ and the point mass on the OFUL action when $d < A$. Abbreviate $p_t(a) = p_t(a \mid S_1^{(t)})$ and put $\bar{c}_t(s) = \sum_a p_t(a \mid s) c_t(s,a)$.

Define the fixed linear approximation $f_Z : \mathcal{S}_1 \times \mathcal{A} \to [-\kappa, H+\kappa]$ by $f_Z(s,a) = \phi_1(s,a)^\top \theta_Z^\star$. For episode t, define $f_t : \mathcal{A} \to [-\kappa, H+\kappa]$ by $f_t(a) = f_Z(S_1^{(t)}, a)$. Let $a_Z^\star : \mathcal{S}_1 \to \mathcal{A}$ be the measurable maximizing map $a_Z^\star(s) = \min \arg\max_{a \in \mathcal{A}} f_Z(s,a)$, and write $a_t^\star = a_Z^\star(S_1^{(t)})$. With the preceding convention for p_t, define the root regret for the linear approximation

$$\overline{R}_N^{\mathrm{root}} = \sum_{t=1}^{N} \sum_{a \in \mathcal{A}} p_t(a)\{f_t(a_t^\star) - f_t(a)\}. \tag{121}$$

This is the cumulative loss from the root-action choices, evaluated with one fixed approximation to the optimal restricted tail's action values. The average over p_t is conditional on the history and root state; for $\kappa = 0$, the approximation equals the true action-value function.

Define the actual value-gap function $r_t : \mathcal{S}_1 \to [0, H]$ by

$$r_t(s) = \max_a Q_1^{\rho_Z^\star}(s,a) - \sum_a p_t(a \mid s) Q_1^{\rho_t}(s,a).$$

Tail optimality gives $Q_1^{\rho_t}(s,a) \le Q_1^{\rho_Z^\star}(s,a)$, so $0 \le r_t(s) \le H$. Adding and subtracting the value of the chosen root action under the optimal restricted tail gives the exact decomposition

$$r_t(s) = \left\{ \max_a Q_1^{\rho_Z^\star}(s,a) - \sum_a p_t(a \mid s) Q_1^{\rho_Z^\star}(s,a) \right\} + \bar{c}_t(s).$$

The term in braces is the loss from the root-action choice, and $\bar{c}_t(s)$ is the loss from executing the learned tail. Equation (119) and $|e_Z(s,a)| \le \kappa$ give

$$r_t(s) \le \sum_a p_t(a \mid s)\{f_Z(s, a_Z^\star(s)) - f_Z(s,a) + c_t(s,a)\} + 2\kappa. \tag{122}$$

The two residuals in this comparison are evaluated at the best true root action and at the selected action; neither is assumed centered. Since $S_1^{(t)} \sim \mu$ independently of $\mathcal{F}_{t-1}$,

$$\mathbb{E}[r_t(S_1^{(t)}) \mid \mathcal{F}_{t-1}] = J_Z^\star - J(\pi_t).$$

The conditional Hoeffding inequality gives, with failure probability δ_{reg},

$$\mathrm{Reg}_{N,Z} \le \sum_{t=1}^{N} r_t(S_1^{(t)}) + H\sqrt{2N \log(1/\delta_{\mathrm{reg}})}.$$

Using (122), (121), and the definition of $\bar{c}_t$ gives

$$\mathrm{Reg}_{N,Z} \le \overline{R}_N^{\mathrm{root}} + \sum_{t=1}^{N} \bar{c}_t(S_1^{(t)}) + 2\kappa N + \widetilde{O}(H\sqrt{N}). \tag{123}$$

We next bound the direct continuation-loss term $\sum_t \bar{c}_t(S_1^{(t)})$ and the observation bias that affects $\overline{R}_N^{\mathrm{root}}$. These are separate effects of learning the tail.

B.6.2 Continuation loss and root-observation bias

Let $\bar{Z}_t \in Z \cup \{\perp\}$ be the first state in Z reached after the root in episode t, with $\bar{Z}_t = \perp$ if no such state is reached. Define the tail gap $g_t = V_Z^\star(\bar{Z}_t) - V^{\rho_t}(\bar{Z}_t) \in [0, H]$. If $D_Z = 0$, then $\bar{Z}_t = \perp$ and $g_t = 0$. For $D_Z \ge 1$, apply Corollary B.16. Thus in both cases

$$\sum_{t=1}^{N} \big(V_Z^\star(\bar{Z}_t) - V^{\rho_t}(\bar{Z}_t)\big) \le \widetilde{O}(H\sqrt{D_Z N}). \tag{124}$$

Root observations. Define the centered noise $\nu_t = G_t - \mathbb{E}[G_t \mid \mathcal{H}_t] \in \mathbb{R}$ and the residual $e_{Z,t} = e_Z(S_1^{(t)}, A_1^{(t)}) \in [-\kappa, \kappa]$. Since ρ_t is $\mathcal{F}_{t-1}$-measurable, $\mathbb{E}[G_t \mid \mathcal{H}_t] = Q_1^{\rho_t}(S_1^{(t)}, A_1^{(t)})$. Substitution of (119) gives

$$G_t = \phi_1(S_1^{(t)}, A_1^{(t)})^\top \theta_Z^\star + \nu_t - c_t + e_{Z,t}. \tag{125}$$

The prefix before the first controlled state is common to both tails, so

$$c_t = \mathbb{E}[V_Z^\star(\bar{Z}_t) - V^{\rho_t}(\bar{Z}_t) \mid \mathcal{H}_t]. \tag{126}$$

Direct continuation loss. If $D_Z = 0$, then $c_t = \bar{c}_t(s) = 0$ for all t, s. For $D_Z \geq 1$, proceed as follows. Apply the first inequality in (46) first to g_t/H with conditional mean c_t/H given $\mathcal{H}_t$, and then to c_t/H with conditional mean $\bar{c}_t(S_1^{(t)})/H$ given $\mathcal{K}_t$. Allocating half of δ_{l1} to each conversion gives

$$\sum_{t=1}^N \bar{c}_t(S_1^{(t)}) \leq 4 \sum_{t=1}^N g_t + 6H \log(4/\delta_{\mathrm{l1}}).$$

Together with (124), this yields

$$\sum_{t=1}^N \bar{c}_t(S_1^{(t)}) \leq \widetilde{O}(H\sqrt{D_Z N}). \tag{127}$$

Squared observation bias. Apply Corollary B.18 with failure probability δ_{sq} when $D_Z \geq 1$; when $D_Z = 0$, use $g_t = c_t = 0$. Jensen's inequality gives $c_t^2 \leq v_t := \mathbb{E}[g_t^2 \mid \mathcal{H}_t]$, where $v_t \in [0, H^2]$. The first inequality in (46), applied to g_t^2/H^2, gives, with probability at least $1 - \delta_{L_2}$,

$$\sum_{t=1}^N c_t^2 \leq \sum_{t=1}^N v_t \leq 2 \sum_{t=1}^N g_t^2 + 2H^2 \log(1/\delta_{L_2}).$$

Combining this display with Corollary B.18, the deterministic bound $\sum_t c_t^2 \leq NH^2$, and the definition (25) gives

$$\sum_{t=1}^N c_t^2 \leq \mathcal{B}_{\mathrm{cont}} = \widetilde{O}(H^2(D_Z + 1)). \tag{128}$$

The two bounds have different roles in (123): (127) controls the direct continuation loss, whereas (128) controls the effect of $-c_t$ on learning the root action values. Below, $\sqrt{\mathcal{B}_{\mathrm{cont}}}$ enters the OFUL confidence radius and $\mathcal{B}_{\mathrm{cont}}$ enters the SquareCB prediction-error budget. Their respective conversions to root regret give $\widetilde{O}(\sqrt{dN\mathcal{B}_{\mathrm{cont}}})$ and $\widetilde{O}(\sqrt{AN\mathcal{B}_{\mathrm{cont}}})$. Using only (124) and its conditional-mean conversion would instead give $\sum_t c_t^2 \leq H \sum_t c_t = \widetilde{O}(H^2\sqrt{D_Z N})$, leading to an $N^{3/4}$ term in these bounds.

B.6.3 Root regret for OFUL

Recall $x_t = \phi_1(S_1^{(t)}, A_1^{(t)}) \in \mathbb{R}^d$ and the ridge state (27). For OFUL, let $X = [x_1, \ldots, x_{t-1}] \in \mathbb{R}^{d \times (t-1)}$. A singular-value decomposition $X = U \operatorname{diag}(s_j) V^\top$ shows that the nonzero eigenvalues of $X^\top(\lambda I + XX^\top)^{-1}X$ are $s_j^2/(\lambda + s_j^2) \leq 1$; the remaining eigenvalues are zero. Thus for every vector $v \in \mathbb{R}^{t-1}$,

$$\|Xv\|_{\Sigma_t^{-1}}^2 = v^\top X^\top (\lambda I + XX^\top)^{-1} Xv \leq \|v\|_2^2. \tag{129}$$

Applying this inequality separately to $(c_i)_{i<t}$ and $(e_{Z,i})_{i<t}$, then using the triangle inequality, gives

$$\left\| \sum_{i<t} x_i (c_i - e_{Z,i}) \right\|_{\Sigma_t^{-1}} \leq \sqrt{\mathcal{B}_{\mathrm{cont}}} + \kappa\sqrt{t-1}. \tag{130}$$

Define the event

$$\mathcal{E}_{\mathrm{sn}} = \left\{ \forall t \leq N : \left\| \sum_{i<t} x_i \nu_i \right\|_{\Sigma_t^{-1}} \leq \frac{H}{2} \sqrt{2 \log(1/\delta_{\mathrm{root}}) + \log \frac{\det \Sigma_t}{\det(\lambda I)}} \right\}. \tag{131}$$

This event has probability at least $1-\delta_{\mathrm{root}}$. Indeed, x_t is $\mathcal{H}_t$-measurable and is selected before the episode return is observed, whereas $\mathbb{E}[\nu_t \mid \mathcal{H}_t] = 0$. Conditional on $\mathcal{H}_t$, the return $G_t \in [0, H]$, and subtracting its conditional mean does not change its range length. Hoeffding's lemma therefore makes ν_t conditionally $H/2$-sub-Gaussian. To derive the self-normalized inequality directly, for every fixed vector $v \in \mathbb{R}^d$ define

$$\mathcal{M}_t(v) = \exp\left\{ \frac{2}{H} v^\top \sum_{i<t} x_i \nu_i - \frac{1}{2} v^\top (\Sigma_t - \lambda I) v \right\}.$$

Conditional Hoeffding followed by conditioning from $\mathcal{H}_t$ back to $\mathcal{F}_{t-1}$ makes $\mathcal{M}_t(v)$ a nonnegative supermartingale starting at one. Integrating v against the probability law $\mathcal{N}(0, \lambda^{-1} I)$ preserves this property. Completing the Gaussian square evaluates the mixture as

$$\overline{\mathcal{M}}_t = \sqrt{\frac{\det(\lambda I)}{\det \Sigma_t}} \exp\left\{ \frac{2}{H^2} \left\| \sum_{i<t} x_i \nu_i \right\|^2_{\Sigma_t^{-1}} \right\}.$$

Stopping this nonnegative supermartingale at its first crossing of $1/\delta_{\mathrm{root}}$ gives a crossing probability at most δ_{root}. Rearranging $\overline{\mathcal{M}}_t \le 1/\delta_{\mathrm{root}}$ proves (131) simultaneously for all $t \le N$. This is the Gaussian-mixture proof of the self-normalized inequality in Abbasi-Yadkori et al. (2011, Theorem 1); it uses no independence between the selected feature and previous returns. On $\mathcal{E}_{\mathrm{sn}}$, (130) and the ridge bias bound $\sqrt{\lambda}\|\theta_Z^\star\|_2 \le \sqrt{\lambda} B$ show that the confidence set with radius (29) contains $\theta_Z^\star$. OFUL optimism therefore gives, for the selected root feature,

$$f_t(a_t^\star) - f_t(A_1^{(t)}) \le 2\beta_t \|x_t\|_{\Sigma_t^{-1}}.$$

The gap between the linear action values is at most $H + 2\kappa \le 3H$, while $\beta_t \ge \sqrt{\lambda} B = H$. Consequently the left side is at most $3\beta_t \min\{1, \|x_t\|_{\Sigma_t^{-1}}\}$. Put $a_t = \|x_t\|^2_{\Sigma_t^{-1}}$. The determinant lemma gives $\det \Sigma_{t+1} / \det \Sigma_t = 1 + a_t$. Since $\min\{1, a\} \le 2\log(1+a)$ for $a \ge 0$, summation gives

$$\sum_{t=1}^{N} \min\{1, \|x_t\|^2_{\Sigma_t^{-1}}\} \le 2 \log \frac{\det \Sigma_{N+1}}{\det(\lambda I)},$$

and the trace–determinant inequality gives

$$\log \frac{\det \Sigma_{N+1}}{\det(\lambda I)} \le d \log\left(1 + \frac{NB^2}{dH^2}\right).$$

The radius β_t is nondecreasing. Consequently Cauchy–Schwarz and the preceding determinant bound give

$$\overline{R}_N^{\mathrm{root}} \le 3\beta_N \sum_{t=1}^{N} \sqrt{\min\{1, a_t\}} \le 3\beta_N \sqrt{2Nd \log\left(1 + \frac{NB^2}{dH^2}\right)}.$$

By (29) and (25), $\beta_N = \widetilde{O}(H\sqrt{d} + H\sqrt{D_Z + 1} + \kappa\sqrt{N})$. The radius term $\sqrt{\mathcal{B}_{\mathrm{cont}}}$ therefore contributes $\widetilde{O}(\sqrt{dN\mathcal{B}_{\mathrm{cont}}})$ to root regret. Multiplying all terms gives

$$\overline{R}_N^{\mathrm{root}} \le \widetilde{O}\left(Hd\sqrt{N} + H\sqrt{d(D_Z+1)N} + \kappa\sqrt{d}\, N \right), \qquad d < A. \tag{132}$$

B.6.4 Root regret for SquareCB

For SquareCB, define the scalar $e_t = \widehat{f}_t(A_1^{(t)}) - f_t(A_1^{(t)}) \in \mathbb{R}$. The clipped Vovk–Azoury–Warmuth predictor obeys

$$\sum_{t=1}^{N} \{(\widehat{f}_t(A_1^{(t)}) - G_t)^2 - (f_t(A_1^{(t)}) - G_t)^2\} \le R_{\mathrm{pred}}. \tag{133}$$

Here is a direct proof of the forward-ridge guarantee (Azoury & Warmuth, 2001; Vovk, 2001). Define the scalars $a_t = x_t^\top \Sigma_t^{-1} x_t \ge 0$ and $q_t = x_t^\top \Sigma_t^{-1} b_t$, and let $\widetilde{f}_t = q_t/(1+a_t)$ be the prediction

before clipping. Sherman–Morrison gives precisely this prediction in (32). Define the regularized least-squares minimum

$$M_t = \min_{\theta\in\mathbb{R}^d}\left\{\lambda\|\theta\|_2^2 + \sum_{i=1}^{t}(G_i - x_i^\top\theta)^2\right\}$$
$$= \sum_{i=1}^{t} G_i^2 - b_{t+1}^\top \Sigma_{t+1}^{-1} b_{t+1}, \qquad M_0 = 0.$$

Another application of Sherman–Morrison and direct expansion give

$$M_t - M_{t-1} = \frac{(G_t - q_t)^2}{1+a_t},$$
$$(G_t - \widetilde{f}_t)^2 - (M_t - M_{t-1}) = \frac{a_t G_t^2}{1+a_t} - \frac{a_t q_t^2}{(1+a_t)^2} \le H^2\log(1+a_t).$$

The last step uses $G_t \in [0,H]$ and $a/(1+a) \le \log(1+a)$ for $a \ge 0$. Sum over t, use the matrix determinant lemma, and upper-bound M_N by its objective at any $\theta \in \mathbb{R}^d$. Projection onto $[0,H]$ cannot increase square loss, so the clipped predictor satisfies the pathwise inequality

$$\sum_{t=1}^{N}\{(\widehat{f}_t(A_1^{(t)}) - G_t)^2 - (x_t^\top\theta - G_t)^2\} \le \lambda\|\theta\|_2^2 + H^2\log\frac{\det\Sigma_{N+1}}{\det(\lambda I)}.$$

Taking $\theta = \theta_Z^\star$, using $\lambda B^2 = H^2$, and applying the trace–determinant bound proves (133) with the stated R_{pred}. Substitute (125) and use Young's inequality:

$$-2e_t(c_t - e_{Z,t}) \le \frac{1}{4}e_t^2 + 4(c_t - e_{Z,t})^2, \qquad \sum_{t=1}^{N}(c_t - e_{Z,t})^2 \le 2\mathcal{B}_{\mathrm{cont}} + 2N\kappa^2.$$

The conditional Hoeffding supermartingale applies because e_t is $\mathcal{H}_t$-measurable and ν_t has conditional range length at most H. Specifically, for every $\alpha > 0$, $\mathbb{E}[\exp(2\alpha e_t\nu_t) \mid \mathcal{H}_t] \le \exp(\alpha^2H^2e_t^2/2)$. Iteration and Markov's inequality with $\alpha = 1/(2H^2)$ give, with failure probability at most δ_{pred},

$$2\sum_{t=1}^{N} e_t\nu_t \le \frac{1}{4}\sum_{t=1}^{N} e_t^2 + 2H^2\log(1/\delta_{\mathrm{pred}}).$$

Moving the two quarter-square terms to the left and using (128) gives

$$\sum_{t=1}^{N} e_t^2 \le 2R_{\mathrm{pred}} + 16\mathcal{B}_{\mathrm{cont}} + 16N\kappa^2 + 4H^2\log(1/\delta_{\mathrm{pred}}). \tag{134}$$

Put $m_t = \sum_a p_t(a)(\widehat{f}_t(a) - f_t(a))^2 = \mathbb{E}[e_t^2 \mid \mathcal{K}_t]$. The first inequality in (46), applied to $e_t^2/(4H^2)$ under the root-action draw, gives with failure probability δ_{full}

$$\sum_{t=1}^{N} m_t \le 2\sum_{t=1}^{N} e_t^2 + 8H^2\log(1/\delta_{\mathrm{full}}).$$

Here $|e_t| \le H + \kappa \le 2H$ because the predictor is clipped and the linear approximation lies in $[-\kappa, H+\kappa]$. Since $\delta_{\mathrm{full}} = \delta_{\mathrm{pred}}$, (134) and the definition of U_{scb} yield

$$\sum_{t=1}^{N}\sum_{a} p_t(a)(\widehat{f}_t(a) - f_t(a))^2 \le U_{\mathrm{scb}}. \tag{135}$$

The one-step SquareCB inequality is

$$\sum_{a} p_t(a)\{f_t(a_t^\star) - f_t(a)\} \le \frac{2A}{\gamma} + \frac{\gamma}{4}\sum_{a} p_t(a)(\widehat{f}_t(a) - f_t(a))^2. \tag{136}$$

It is algebraic and remains valid when the linear approximation lies outside $[0, H]$. Write $\widehat{\Delta}_t(a) = \widehat{f}_t(a_t^0) - \widehat{f}_t(a) \geq 0$, where $a_t^0 = \min \arg\max_a \widehat{f}_t(a)$, and put $e_t(a) = \widehat{f}_t(a) - f_t(a)$. If $a_t^\star$ maximizes f_t, then exact expansion gives

$$\sum_a p_t(a)\{f_t(a_t^\star) - f_t(a)\} = -e_t(a_t^\star) - \widehat{\Delta}_t(a_t^\star) + \sum_a p_t(a)\widehat{\Delta}_t(a) + \sum_a p_t(a)e_t(a).$$

Weighted Young's inequality, applied once to the complete error term, gives

$$\begin{aligned}
&\sum_a \{p_t(a) - \mathbf{1}\{a = a_t^\star\}\}e_t(a) \\
&\quad \leq \frac{\gamma}{4}\sum_a p_t(a)e_t(a)^2 + \frac{1}{\gamma}\sum_a \frac{(p_t(a) - \mathbf{1}\{a = a_t^\star\})^2}{p_t(a)} \\
&\quad = \frac{\gamma}{4}\sum_a p_t(a)e_t(a)^2 + \frac{1}{\gamma}\left(\frac{1}{p_t(a_t^\star)} - 1\right).
\end{aligned}$$

The sampling rule has $p_t(a_t^0) \geq 1/A$ and, for $a \neq a_t^0$, $p_t(a)^{-1} = A + \gamma\widehat{\Delta}_t(a)$. Hence

$$\frac{1}{\gamma p_t(a_t^\star)} - \widehat{\Delta}_t(a_t^\star) \leq \frac{A}{\gamma}, \qquad \sum_a p_t(a)\widehat{\Delta}_t(a) \leq \frac{A-1}{\gamma}.$$

Combining the last three displays bounds the non-square terms by $2(A-1)/\gamma \leq 2A/\gamma$, proving (136). Summing gives $\overline{R}_N^{\mathrm{root}} \leq 2AN/\gamma + \gamma U_{\mathrm{scb}}/4$. The choice $\gamma = \sqrt{8AN/U_{\mathrm{scb}}}$ in (30) minimizes this upper bound at $\sqrt{2ANU_{\mathrm{scb}}}$. The $\mathcal{B}_{\mathrm{cont}}$ term in U_{scb} therefore contributes $\widetilde{O}(\sqrt{AN\mathcal{B}_{\mathrm{cont}}})$ to root regret. Using (128) for that term gives

$$\overline{R}_N^{\mathrm{root}} \leq \widetilde{O}\Big(H\sqrt{AdN} + H\sqrt{A(D_Z+1)N} + \kappa\sqrt{A}\,N\Big), \qquad A \leq d. \tag{137}$$

B.6.5 Combining root regret and continuation loss

The OFUL bound (132) and the SquareCB bound (137) each consist of statistical estimation error, the effect of continuation bias on root estimation, and realizability error. Using the prescribed branch, the middle term is

$$\widetilde{O}\Big(\sqrt{\min\{A,d\}N\mathcal{B}_{\mathrm{cont}}}\Big) = \widetilde{O}\Big(H\sqrt{\min\{A,d\}(D_Z+1)N}\Big).$$

Equation (127) bounds the additional direct continuation loss by $\widetilde{O}(H\sqrt{D_Z N})$, which is absorbed by this term. The $2\kappa N$ and $\widetilde{O}(H\sqrt{N})$ terms in (123) are absorbed by the realizability and statistical terms, respectively. Substitution of (127) and either (132) or (137) into (123) proves (9). Its failure event is measurable in the observed history and executed policies. Averaging its conditional failure probability over the discovery history proves the unconditional claim; no measurable choice of $\theta_Z^\star$ across histories is needed.

B.7 Parameter comparison for the main theorem

For Algorithm 1, we prove (16) from the parameters in Appendix A.1. We also verify the regret bound for the branch $\eta > H$.

Proof. Put $\eta_0 = \sqrt{\mathsf{C}_{\mathrm{d}}/T}$. The definitions give

$$\begin{gathered}
L_{\mathrm{d}} = \widetilde{O}(dK), \qquad m_{\mathrm{d}} = \widetilde{O}(d), \qquad M = \widetilde{O}((H-1)d^2K), \\
\Xi = \widetilde{O}(HA(H-1)d^2K), \qquad \mathsf{S}_{\mathrm{d}} = \widetilde{O}(\Xi), \\
\mathsf{C}_{\mathrm{d}} = \widetilde{O}(\overline{\Gamma}AH^3(H-1)d^2K).
\end{gathered}$$

Thus $T\eta_0$ and both restricted statistical terms are bounded by the first term of (16), as is $4H\sqrt{T}$. Since $\eta \geq \eta_0$ and $\mathsf{S}_{\mathrm{d}}/p_\eta \geq 1$,

$$HN_0 \leq \frac{3H\mathsf{S}_{\mathrm{d}}}{p_\eta} = \frac{\mathsf{C}_{\mathrm{d}}}{\eta} \leq T\eta_0.$$

Furthermore $\Xi/L_{\mathrm{d}} = O(HA(H-1)m_{\mathrm{d}})$, so

$$\mathsf{C}_\kappa = \widetilde{O}\big(\overline{\Gamma}AH^2(H-1)d^2\big)\,.$$

This also dominates $\sqrt{\min\{A,d\}}$ because $A,d,K \geq 1$ and $H \geq 2$. Moreover $\overline{\Gamma} \leq 2\min\{A,d\}$. The discovery error term is $\mathsf{C}_\kappa\chi T$, while $\kappa \leq \chi$ bounds the restricted approximation term by the same scale. Using $\max\{a,b,c\} \leq a+b+c$ proves (16). If $\eta > H$, the same comparison of $T\eta$ bounds $\mathrm{Reg}_T \leq HT < T\eta$ by (1); no episode bound or definition of n_{d} is needed in that branch. □

B.8 Computation of the main algorithm

For Algorithm 1, we establish the polynomial operation claim in Theorem 3.3 under the exact computation model of Section 3. A direct implementation uses at most

$$O\big(HL_{\mathrm{d}}d^2 + THAL_{\mathrm{d}}d^2 + T^2(H+A) + Td\log(A+1)(Ad^2+d^3)\big) \tag{138}$$

operations. Initialization stores $(H-1)L_{\mathrm{d}}$ matrices. There are at most T discovered states, $1+AT$ controlled pairs, and TH recorded labels or segment outcomes across the two phases. Linear scans for state equality therefore cost $O(T^2H)$. At most TH novelty tests each evaluate AL_{d} quadratic forms at cost $O(d^2)$ per form. Maintain matrix inverses by Sherman–Morrison after each discovery; its denominator is $1 + v^\top W^{-1}v \geq 1$ because $W \succ 0$.

Discovery batches are disjoint and their statistics can be computed once at promotion. Store the index of each observed segment endpoint when it is recorded. Every restricted backward pass then scans at most TH frozen outcomes and AT pairs, costing $O(T^2(H+A))$ over all episodes. Lemma B.2 bounds each of the at most T root-design calls by $O(d\log(A+1))$ iterations, each using $O(Ad^2+d^3)$ operations. The restricted root predictors use at most this many operations as well. Finally, (10) makes L_{d} polynomial in the stated parameters. Thus (138) proves the computational claim without enumerating the state spaces or transition supports. Only observed states and finite tables are stored; the support covers and profile classes occur solely in the proof.

C An information-theoretic lower bound

This appendix proves the statistical lower bound (139) and the approximation lower bound (145). Both constructions satisfy Assumption 3.2 for every prescribed concentration error $\upsilon \in [0,1]$; the latter also uses unrestricted initial support. For each positive integer r, let $\mathcal{X}_r = \{-1,+1\}^r$ be the set of sign vectors. The tail-mass parameter is an upper bound, so a construction with zero residual mass belongs to the class for any prescribed υ. These results therefore give no lower bound on the cost of positive tail mass.

Theorem C.1 (Expected-regret lower bound). *There are universal constants $c, c_0 > 0$ such that, for $H, d \geq 8$, $A \geq 2$, $B > 0$, $\kappa \in [0,H]$, $\upsilon \in [0,1]$, and $T \geq c_0 Hd^2$, if*

$$r = \min\{\lfloor \log_2 A \rfloor, d-4\},$$

then every possibly randomized, computationally unbounded online algorithm has an exactly realizable instance with $K = 2$ satisfying Assumptions 3.1 and 3.2 for which

$$\begin{aligned} \mathbb{E}\,\mathrm{Reg}_T &\geq c\min\{H,B\}\sqrt{HdrT}, \\ \mathbb{E}\,\mathrm{Reg}_T &\geq c\min\{H,B\}\sqrt{Hd\min\{d,1+\log_2 A\}T}. \end{aligned} \tag{139}$$

We begin with a finite-action contextual-bandit fact.

Lemma C.2 (Hypercube bandits in independent contexts). *For positive integers n, N, r, on each of N rounds, a context X_t is drawn independently and uniformly from $[n]$. Context x has an unknown sign vector $\sigma_x \in \mathcal{X}_r$. For a positive scalar γ specified below, choosing $Z_t \in \mathcal{X}_r$ produces a Bernoulli observation with mean*

$$q_x(z) = \frac{1}{2} + \frac{\gamma}{r}\langle z, \sigma_x\rangle.$$

There are universal constants $\alpha, c > 0$ such that, whenever $\gamma = \alpha r\sqrt{n/N} \leq 1/8$, every algorithm satisfies

$$\sup_{(\sigma_x)_{x=1}^n} \mathbb{E}\sum_{t=1}^N \{q_{X_t}(\sigma_{X_t}) - q_{X_t}(Z_t)\} \geq cr\sqrt{nN}. \tag{140}$$

Proof. Put the independent uniform prior on all nr signs. Write KL and TV for relative entropy and total variation distance between probability laws, and Ham for the Hamming distance between sign vectors. Fix a context x, a coordinate ℓ, and a deterministic configuration of all signs other than $\sigma_{x,\ell}$. Let P_+ and P_- be the full interactive laws when the remaining sign is respectively $+1$ and -1; the learner's internal randomness is included in both laws with the same distribution. Conditional on any common pre-action history, the two Bernoulli means agree unless the current context is x; when it is x, they differ in absolute value by $2\gamma/r$ and both lie in $[3/8, 5/8]$. Let $N_x = \sum_{t=1}^{N} \mathbf{1}\{X_t = x\}$ be the number of visits to context x. For $p, q \in (0,1)$, applying $\log z \le z - 1$ to the two terms of the Bernoulli relative entropy gives

$$\mathrm{kl}(p,q) \le p\left(\frac{p}{q} - 1\right) + (1-p)\left(\frac{1-p}{1-q} - 1\right) = \frac{(p-q)^2}{q(1-q)}.$$

Here $q \in [3/8, 5/8]$ implies $q(1-q) \ge 15/64$. Thus an informative observation contributes at most $(256/15)\gamma^2/r^2 \le 32\gamma^2/r^2$. In the chain rule, context and action kernels contribute zero because their conditional laws, as functions of the observed history, are the same under both hypotheses. Only the Bernoulli observations contribute, and hence

$$\mathrm{KL}(P_+, P_-) \le 32\frac{\gamma^2}{r^2}\mathbb{E}_+ N_x = 32\frac{\gamma^2 N}{r^2 n} = 32\alpha^2.$$

Here $\mathbb{E}_+ N_x = N/n$ because the contexts are independent and uniform. For $\alpha = 1/32$, Pinsker's inequality gives $\mathrm{TV}(P_+, P_-) \le 1/8$. At round t, $Z_{t,\ell}$ is measurable with respect to the learner's randomness, the pre-round history, and X_t. The random variable X_t is independent of the learner's random seed and pre-round history and has the same law under P_+ and P_-. Conditional on $X_t = x$, the joint law of that seed and history is therefore unchanged under each hypothesis. Its total variation distance is at most that of the two full interactive laws. Data processing for the action and Le Cam's inequality therefore give

$$\frac{1}{2}P_+\{Z_{t,\ell} = -1 \mid X_t = x\} + \frac{1}{2}P_-\{Z_{t,\ell} = +1 \mid X_t = x\} \ge \frac{1 - \mathrm{TV}(P_+, P_-)}{2} \ge \frac{7}{16}.$$

Averaging this paired inequality over every configuration of the other signs, then over x and ℓ, yields $\mathbb{E}\,\mathrm{Ham}(Z_t, \sigma_{X_t}) \ge 7r/16$ under the stated prior. Since $q_x(\sigma_x) - q_x(z) = (2\gamma/r)\,\mathrm{Ham}(z, \sigma_x)$, summing over t gives Bayes regret at least $7\gamma N/8 = (7/256)r\sqrt{nN}$. One deterministic sign collection has expected regret at least this Bayes average. □

Proof of Theorem C.1. Define the positive integers

$$S = \left\lfloor \frac{d-3}{r+1} \right\rfloor, \qquad m = \left\lfloor \frac{H}{4} \right\rfloor.$$

For each $i \in [S]$ and $j \in \{2, \dots, m+1\}$ choose an unknown $\sigma_{i,j} \in \mathcal{X}_r$. Identify the first 2^r actions with sign vectors $z(a) \in \mathcal{X}_r$; any remaining action copies a fixed effective action. Put

$$G = \min\{H, B\}/3, \qquad \rho = G/H, \qquad p = 1/2, \qquad \gamma = \alpha r\sqrt{Sm/T}. \tag{141}$$

Because $S \le d/(r+1)$, $m \le H/4$, and $r \le d$,

$$\gamma^2 = \frac{\alpha^2 r^2 Sm}{T} \le \frac{\alpha^2}{4c_0}, \qquad S\gamma^2 = \frac{\alpha^2 (rS)^2 m}{T} \le \frac{\alpha^2}{4c_0}.$$

With $\alpha = 1/32$ as in Lemma C.2, taking $c_0 = 1$ gives $\gamma \le 1/64 \le 1/8$ and $S\gamma^2 \le 1/4096 \le 1$.

At layer $h \in [H]$, include the chain states g_h, b_h, the waiting states $w_{i,h}$ for $i \in [S]$ if $h \le m$, and the active states $x_{i,h}$ for $i \in [S]$ if $2 \le h \le m+1$. There are no other states; both chains terminate at $\perp$ after layer H. The initial state is uniform on waiting states $w_{i,1}$, $i \in [S]$. From a waiting state $w_{i,h}$, $1 \le h \le m$, every action moves to the active state $x_{i,h+1}$ with probability $1/(m-h+1)$ and, when $h < m$, otherwise moves to $w_{i,h+1}$. At $h = m$ the activation probability is one. Let J be the random layer at which the active state is reached. Then, for $j \in \{2, \dots, m+1\}$,

$$\mathbb{P}(J = j) = \prod_{h=1}^{j-2} \frac{m-h}{m-h+1} \cdot \frac{1}{m-j+2} = \frac{1}{m}.$$

Thus J is uniform on $\{2,\dots,m+1\}$. At $x_{i,j}$, action a has reward zero and moves to the rewarding state g_{j+1} with probability

$$p+\frac{\gamma}{r}\langle z(a),\sigma_{i,j}\rangle, \tag{142}$$

and to the zero-reward state b_{j+1} otherwise. The rewarding and zero-reward chains are deterministic; every rewarding state pays deterministic reward ρ, while all other rewards vanish. Thus success at active layer j produces subsequent return $L_j=(H-j)\rho$. The one or two destinations specified for each pair form a concentration set of mass $1\ge 1-\upsilon$.

Let $I\in[S]$ be the index of the sampled initial state. The random pair (I,J) is uniform on $[S]\times\{2,\dots,m+1\}$ independently across episodes. Conditional on this pair, the waiting prefix has a law independent of every hidden sign; actions do not affect this law. The active transition is the only informative Bernoulli observation, and the remaining trajectory is a deterministic function of that observation and J. Consequently a bandit learner given (I,J) can simulate the entire episode, including all noninformative observations and the MDP learner's randomization. Therefore the active decisions instantiate Lemma C.2 with $n=Sm, N=T$. Define $q_{i,j}:\mathcal{X}_r\to[0,1]$ by $q_{i,j}(z)=p+(\gamma/r)\langle z,\sigma_{i,j}\rangle$. Let $\mathcal{F}_{t-1}$ be the sigma-field of preceding episode data including the learner's chosen episode policy π_t, and let (I,J) and Z_t be the random context and active action vector of episode t. Conditional on each fixed sign collection,

$$J^\star-J(\pi_t)=\mathbb{E}[L_J\{q_{I,J}(\sigma_{I,J})-q_{I,J}(Z_t)\}\mid\mathcal{F}_{t-1}].$$

Since $j\le m+1$ and $H\ge 8$ imply $L_j\ge G/2$, taking expectations and using the lemma's constant $7/256$ gives

$$\sup_{(\sigma_{i,j})}\mathbb{E}\,\mathrm{Reg}_T\ge\frac{7G}{512}r\sqrt{SmT}. \tag{143}$$

Indeed, $r\le d-4$ implies $(d-3)/(r+1)\ge 1$, so

$$Sr\ge\frac{r(d-3)}{2(r+1)}\ge\frac{d-3}{4}\ge\frac{d}{8},\qquad m=\lfloor H/4\rfloor\ge H/8.$$

These bounds show that $r\sqrt{SmT}\ge\frac{1}{8}\sqrt{HdrT}$, proving the first inequality in (139).

It remains to verify Assumption 3.1. Introduce mutually orthogonal coordinates

$$e_{c_x},e_{c_w},e_{c_g},\quad e_{x_i,\ell}\ (i\in[S],\ell\in[r]),\quad e_{w_i}\ (i\in[S]),$$

whose number is $Sr+S+3\le d$, and pad with zeros. Define

$$\begin{aligned}\phi_h(x_{i,h},a)&=\frac{1}{\sqrt{2}}e_{c_x}+\frac{1}{\sqrt{2r}}\sum_{\ell=1}^{r}z_\ell(a)e_{x_i,\ell},\\ \phi_h(w_{i,h},a)&=\frac{1}{\sqrt{2}}(e_{c_w}+e_{w_i}),\qquad\qquad \phi_h(g_h,a)=e_{c_g},\quad \phi_h(b_h,a)=0.\end{aligned} \tag{144}$$

The known feature map has norm at most one and does not depend on the hidden signs.

Fix an arbitrary randomized Markov policy π. Let $G_h=(H-h+1)\rho$ and $\widetilde{R}_h=(H-h)\rho$. At an active state,

$$Q_h^\pi(x_{i,h},a)=p\widetilde{R}_h+\frac{\gamma\widetilde{R}_h}{r}\langle z(a),\sigma_{i,h}\rangle.$$

Conditional on a visit to $w_{i,h}$, the activation layer is uniform on $\{h+1,\dots,m+1\}$, independently of the current action. For each active state, define the scalar

$$\beta_{i,j}^\pi=\mathbb{E}_{a\sim\pi_j(\cdot|x_{i,j})}\frac{1}{r}\langle z(a),\sigma_{i,j}\rangle,$$

and

$$W_h^0=\frac{p}{m-h+1}\sum_{j=h+1}^{m+1}L_j,\qquad\Delta_{i,h}^\pi=\frac{\gamma}{m-h+1}\sum_{j=h+1}^{m+1}L_j\beta_{i,j}^\pi.$$

These two quantities are used for $1\le h\le m$; set them to zero outside that range. Also set $\sigma_{i,h}=0\in\mathbb{R}^r$ outside $2\le h\le m+1$. Then $Q_h^\pi(w_{i,h},a)=W_h^0+\Delta_{i,h}^\pi$, $0\le W_h^0\le pG$,

$|\Delta^\pi_{i,h}| \le \gamma G$, $Q^\pi_h(g_h, a) = G_h$, and $Q^\pi_h(b_h, a) = 0$. At every layer these values are represented by the coefficient vector

$$(\theta^\pi_h)_{c_x} = \mathbf{1}\{2 \le h \le m+1\}\sqrt{2}p\widetilde{R}_h, \quad (\theta^\pi_h)_{x_i,\ell} = \sqrt{2}\widetilde{R}_h \frac{\gamma}{\sqrt{r}}\sigma_{i,h,\ell},$$
$$(\theta^\pi_h)_{c_w} = \mathbf{1}\{1 \le h \le m\}\sqrt{2}W^0_h, \qquad (\theta^\pi_h)_{w_i} = \sqrt{2}\Delta^\pi_{i,h}, \qquad (\theta^\pi_h)_{c_g} = G_h.$$

The coordinate groups are orthogonal. By $p = 1/2$ and $G_h, \widetilde{R}_h \le G$, the three shared coordinates contribute at most $(4p^2+1)G^2 = 2G^2$, while the two i-indexed groups contribute at most $4S\gamma^2G^2$. Therefore

$$\|\theta^\pi_h\|_2^2 \le (2 + 4S\gamma^2)G^2 \le 6G^2 = \frac{2}{3}\min\{H, B\}^2 \le B^2.$$

For this family, the displayed coefficients satisfy

$$\sup_{s,a}\left|Q^\pi_h(s,a) - \phi_h(s,a)^\top\theta^\pi_h\right| = 0$$

for every policy and layer, verifying Assumption 3.1 with zero approximation error. The concentration sets above verify Assumption 3.2 with $K = 2$ for the prescribed υ. Finally, if $\log_2 A < d - 4$, then $r = \lfloor \log_2 A \rfloor \ge (\log_2 A)/2$; otherwise $r = d - 4 \ge d/2$. Hence, in both cases,

$$r \ge \frac{1}{4}\min\{d, 1 + \log_2 A\}.$$

Thus $c_0 = 1$ and $c = 7/24576$ suffice for both inequalities in (139). □

Proposition C.3 (Approximation lower bound with unrestricted initial support)**.** *For $H \ge 2$, $d \ge 1$, $A \ge 2$, $B > 0$, $\kappa \in [0, H]$, $\upsilon \in [0, 1]$, and positive integers T, n, define the scalar $L_\kappa = \min\{H, 2\kappa, B + \kappa\}$. Every possibly randomized online algorithm has an instance satisfying Assumptions 3.1 and 3.2 with $K = 1$ and initial support of size n for which*

$$\mathbb{E}\,\mathrm{Reg}_T \ge \left(1 - \frac{1}{A}\right) L_\kappa n \left\{1 - \left(1 - \frac{1}{n}\right)^T\right\}. \tag{145}$$

In particular, $n = 2T$ gives $\mathbb{E}\,\mathrm{Reg}_T \ge \frac{3}{4}(1 - 1/A)L_\kappa T \ge 3\kappa T/8$. More generally, for every $\xi \in (0, 1)$, finite initial support suffices for $\mathbb{E}\,\mathrm{Reg}_T \ge (1 - \xi)(1 - 1/A)L_\kappa T$.

Proof. The initial state is uniform on n distinct states $x_1, \dots, x_n$. Each root state x_i has an unknown label $\sigma_i \in [A]$. At x_i, action a pays reward L_κ/H and moves deterministically to g_2 if $a = \sigma_i$; otherwise it pays zero and moves deterministically to b_2. The rewarding chain $g_2, \dots, g_H$ and zero-reward chain $b_2, \dots, b_H$ have deterministic transitions independent of actions. Each rewarding state pays $L_\kappa/H \in [0, 1]$; each zero-reward state pays zero. The terminal transition is deterministic. Each pair's destination is a singleton concentration set of mass $1 \ge 1 - \upsilon$, verifying Assumption 3.2 with $K = 1$. A successful root action gives total episode return exactly L_κ, including its immediate reward.

Let $e_1 \in \mathbb{R}^d$ be the first coordinate vector and set the known feature map to $\phi_h(s, a) = e_1$ at every layer and pair, independently of the hidden labels. Define the scalar $c_\kappa = L_\kappa - \kappa$. The definition of L_κ gives $L_\kappa \ge \kappa$ and $0 \le c_\kappa \le \min\{B, \kappa\}$. For every policy π and layer h, set $\theta^\pi_h = c_\kappa e_1$. Every possible remaining return lies in $[0, L_\kappa]$, so

$$\sup_{s,a}\left|Q^\pi_h(s,a) - \phi_h(s,a)^\top\theta^\pi_h\right| \le \max\{c_\kappa, L_\kappa - c_\kappa\} = \kappa, \qquad \|\theta^\pi_h\|_2 = c_\kappa \le B.$$

Thus Assumption 3.1 holds for every randomized Markov policy.

Put independent uniform priors on the n labels. Before an initial state is first visited, its label is independent of the learner's complete history and internal randomness: observations at other root states depend only on their own labels. Conditional on such a first visit, any action therefore selects the rewarding chain with probability $1/A$, whereas an optimal policy does so with probability one. The expected regret on that visit is $(1 - 1/A)L_\kappa$.

The probability that the initial state of episode t has not appeared in the preceding $t-1$ episodes is exactly $(1-1/n)^{t-1}$. Regret on repeat visits is nonnegative, so the Bayes expected regret is at least

$$\left(1-\frac{1}{A}\right) L_\kappa \sum_{t=1}^{T} \left(1-\frac{1}{n}\right)^{t-1} = \left(1-\frac{1}{A}\right) L_\kappa n \left\{1-\left(1-\frac{1}{n}\right)^{T}\right\}.$$

Here the summand for $t=1$ equals one, including when $n=1$. The same calculation gives expected pseudo-regret: condition first on the fixed environment and selected episode policy, then average over the fresh initial draw and trajectory. Some deterministic label collection attains at least the Bayes lower bound, proving (145).

For the two consequences in the statement, Bernoulli's inequality gives

$$\sum_{t=1}^{T} \left(1-\frac{1}{n}\right)^{t-1} \geq T - \frac{T(T-1)}{2n}.$$

With $n=2T$, the right side is $(3T+1)/4 \geq 3T/4$. For the last claim, choose any positive integer $n \geq (T-1)/(2\xi)$; the right side is then at least $(1-\xi)T$. □