ASYMPTOTICALLY OPTIMAL MODEL-FREE REINFORCEMENT LEARNING FOR AVERAGE-REWARD MDPS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study average-reward reinforcement learning in finite weakly communicating Markov decision processes with access to a generative model under an $\widetilde{O}(SA)$ working-space constraint, where S and A denote the numbers of states and actions, respectively. Let $H = \max\{1, \mathrm{sp}(h^\star)\}$, where $h^\star$ is an optimal bias function. Our basic algorithm requires no prior bound on H and terminates almost surely. With high probability, it returns a stationary deterministic policy that is ϵ-optimal from every initial state using

$$\widetilde{O}\left(\frac{SAH}{\epsilon^2} + \frac{S^{14/3}A^{4/3}H^{4/3}}{\epsilon^{4/3}}\right)$$

samples. For fixed S, A, H, this bound matches the span-based statistical lower bound up to logarithmic factors as $\epsilon \to 0$.

Our approach has two key ingredients: (i) reward shaping based on a coarse reference bias to reduce the value scale of the planning problem; and (ii) space-efficient linear programming coupled with a randomized square-grid decoder that reconstructs transition probability on demand.

1 INTRODUCTION

The classical theory of Markov decision processes (MDPs) and dynamic programming provides a foundation for sequential decision-making under uncertainty (Puterman, 1994), while temporal-difference learning and Q-learning established influential approaches to learning from experience (Sutton, 1988; Watkins & Dayan, 1992). A central distinction is between model-based methods, which learn an environment model and use it for planning, and model-free methods, which update value or policy estimates directly from sampled experience (Sutton et al., 1998). Model-free learning offers flexibility by allowing representations to focus on values or policies rather than the full environment dynamics, and supports incremental updates without maintaining a complete transition model. This distinction also has a concrete memory consequence: in a tabular MDP with S states and A actions, a dense transition table contains $O(S^2A)$ entries, whereas a state–action value table requires only $O(SA)$. These advantages motivate the development of model-free algorithms that combine compact storage with strong statistical guarantees.

The average-reward criterion measures a policy's long-run performance without specifying a finite horizon or a discount factor (Puterman, 1994; Abounadi et al., 2001). We study finite weakly communicating MDPs, a general setting that allows transient states without requiring every policy to induce an irreducible Markov chain (Puterman, 1994). Given access to a generative model, our goal is to learn a stationary deterministic policy that is ϵ-optimal from every initial state, while retaining only $\widetilde{O}(SA)$ working space.

The information-theoretic lower bound is $\widetilde{\Omega}(SAH/\epsilon^2)$ (Wang et al., 2022). For model-based learning, Zurek & Chen (2025) achieve a matching upper bound by planning with an empirical transition kernel. Model-free methods instead maintain value estimates without storing a transition kernel. However, the sample-complexity bounds of Zhang & Xie (2023) and Lee et al. (2025) with a generative model have leading terms quadratic in H. This gap raises a natural question: *can a model-free learner attain the optimal span-based sample complexity, without knowing H in advance?*

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.

Table 1: Representative span-based simulator guarantees for an ϵ-optimal policy in weakly communicating MDPs. Logarithmic factors are suppressed by $\tilde{O}$ and $\tilde{\Omega}$. Assume $H \geq 1$.

Method	Sample complexity	H knowledge	Working memory
Wang et al. (2022)	$\tilde{O}(SAH/\epsilon^3)$	Yes	$\tilde{O}(S^2A)$
Zhang & Xie (2023)	$\tilde{O}(SAH^2/\epsilon^2 + S^2AH/\epsilon)$	Yes	$\tilde{O}(SA)$
Zurek & Chen (2024)	$\tilde{O}(SAH/\epsilon^2)$	Yes	$\tilde{O}(S^2A)$
Lee et al. (2025)	$\tilde{O}(SAH^2/\epsilon^2)$	No	$\tilde{O}(SA)$
Zurek & Chen (2025)	$\tilde{O}(SAH/\epsilon^2)$	No	$\tilde{O}(S^2A)$
This work	$\tilde{O}(SAH/\epsilon^2 + S^{14/3}A^{4/3}H^{4/3}/\epsilon^{4/3})$	No	$\tilde{O}(SA)$
Lower bound	$\tilde{\Omega}(SAH/\epsilon^2)$	–	–

Our contribution. We answer this question affirmatively in the asymptotic regime. Specifically, we design an algorithm that, with probability at least $1-\delta$, returns a policy that is ϵ-optimal from every initial state using

$$\widetilde{O}\left(\frac{SAH}{\epsilon^2} + \frac{S^{14/3}A^{4/3}H^{4/3}}{\epsilon^{4/3}}\right)$$

simulator samples and $\widetilde{O}(SA)$ working space (see Theorem 4.1). For fixed S, A, H, this bound matches the statistical lower bound up to logarithmic factors as $\epsilon \to 0$. The algorithm requires no prior knowledge of H and terminates almost surely.

We have two key technical ingredients.

Reward shaping with a coarse reference. We use the anchored iteration of Lee et al. (2025) to learn a coarse reference h with a small Bellman residual. We then shape the reward using h, converting the perturbed discounted problem into a nonnegative-cost correction problem whose optimal value is controlled by this residual. This separates the leading cost of scalar estimation from the cost of correction planning.

Space-efficient linear programming with on-demand transition reconstruction. The correction solver reconstructs one probability vector at a time and discards it after use, avoiding storage of the full kernel. Calibration ensures that repeated reconstructions agree with a single rounded kernel (Lemma B.1). A centered total-variance argument applied to this kernel's optimal value function transfers the numerical policy certificate to the true MDP. We explicitly count all row queries made by the numerical planner. Finally, an observable value-span test certifies the output policy without first verifying that the trial scale is at least H.

1.1 Related work

Simulator-based sample complexity guarantees. A line of work studies the sample complexity of average-reward MDPs with a generative model (Wang, 2017; Jin & Sidford, 2020; 2021; Wang et al., 2024). Table 1 summarizes the span-based guarantees, distinguishing whether they require prior span knowledge and retain an empirical transition kernel. Wang et al. (2022) use an average-to-discounted reduction to obtain an ϵ^{-3} upper bound linear in the bias span. Zurek & Chen (2024) achieve the optimal ϵ^{-2} rate given a span upper bound, and Zurek & Chen (2025) remove this requirement through horizon calibration and a data-dependent stopping rule. The latter result settles the span-agnostic sample-complexity question when empirical-kernel storage is allowed.

In the model-free setting, Zhang & Xie (2023) use reference–advantage decomposition to achieve a sample complexity of $\widetilde{O}(SAH^2/\epsilon^2 + S^2AH/\epsilon)$, with parameter choices that depend on the bias span. Lee et al. (2025) develop SAVIA+, which requires no prior span knowledge and achieves a sample-complexity bound quadratic in H. In addition to an ϵ-optimal policy, SAVIA+ produces a bias estimate h with a small Bellman residual, making it an efficient coarse reference learner for our algorithm.

Online regret guarantees. A related line of work studies average-reward reinforcement learning through online interaction, where the learner follows a single trajectory rather than querying arbitrary state–action pairs. Jaksch et al. (2010) develop UCRL2 and establish diameter-dependent regret bounds for communicating MDPs, while REGAL and SCAL obtain bias-span-dependent bounds for weakly communicating MDPs (Bartlett & Tewari, 2009; Fruit et al., 2018). Zhang & Ji (2019) establish a near-optimal span-based regret bound given a span upper bound. Boone & Zhang (2024) achieve the leading rate $\widetilde{O}(\sqrt{SAHT})$, where T is the number of interaction steps, with computationally efficient planning and no prior span knowledge. In the model-free setting, Wei et al. (2020) obtain a regret bound with $T^{2/3}$ dependence under weak communication. Using reference–advantage decomposition, Zhang & Xie (2023) improve the dependence on T to the optimal $\sqrt{T}$ rate, with regret $\widetilde{O}(S^5A^2H\sqrt{T})$.

Low-memory linear programming (LP). For discounted MDPs, Jin et al. (2024) give a generative-model planner based on truncated variance-reduced value iteration with near-linear working space. Sampled primal–dual methods provide another route to MDP optimization (Wang, 2020; Jin & Sidford, 2020). Matrix-free interior-point methods use implicit matrix-vector products (Gondzio, 2012), and Liu et al. (2026) give near-linear working-bit linear programming bounds under explicit rank and numerical assumptions. In our construction, the additional requirement is to supply consistent coefficients from fresh simulator samples and account for every reconstruction.

2 PRELIMINARIES

We consider a finite average-reward MDP $(\mathcal{S}, \mathcal{A}, P, r)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, $P(\cdot \mid s, a)$ is the transition distribution, and $r(s, a) \in [0, 1]$ is the mean reward. Write $S = |\mathcal{S}|$ and $A = |\mathcal{A}|$. A stationary policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ assigns a distribution over actions to each state, where $\Delta(\mathcal{A})$ denotes the probability simplex over $\mathcal{A}$. A stationary deterministic policy is identified with a map $\pi : \mathcal{S} \to \mathcal{A}$.

Average reward and weak communication. The average reward of a stationary policy π from an initial state s is

$$\rho^\pi(s) := \lim_{T\to\infty} \frac{1}{T}\mathbb{E}_\pi\left[\sum_{t=0}^{T-1} r(s_t, a_t) \,\middle|\, s_0 = s\right], \tag{1}$$

where the expectation is over actions $a_t \sim \pi(\cdot \mid s_t)$ and transitions $s_{t+1} \sim P(\cdot \mid s_t, a_t)$. For finite MDPs and stationary policies, this limit exists, but it may depend on the initial state (Puterman, 1994).

Weakly communicating MDPs. An MDP is *weakly communicating* if its state space can be partitioned into a communicating set, in which every state is reachable from every other state under some stationary policy, and a set of states that are transient under every stationary policy (Bartlett & Tewari, 2009). Under this assumption, the optimal average reward $\rho^\star = \max_\pi \rho^\pi(s)$ is independent of s, and a stationary deterministic policy attains it from every initial state (Puterman, 1994). An optimal bias function $h^\star \in \mathbb{R}^S$ satisfies the average-reward Bellman optimality equation

$$\rho^\star + h^\star(s) = \max_{a\in\mathcal{A}} \left\{ r(s, a) + \sum_{s'\in\mathcal{S}} P(s' \mid s, a)h^\star(s') \right\}, \qquad s \in \mathcal{S}. \tag{2}$$

For a vector $v \in \mathbb{R}^S$, define its span by $\mathrm{sp}(v) := \max_{s\in\mathcal{S}} v(s) - \min_{s\in\mathcal{S}} v(s)$. We fix an optimal bias function $h^\star$ and write $H := \max\{1, \mathrm{sp}(h^\star)\}$. The learner is given neither H nor an upper bound on it; H appears only in the analysis.

Generative model. The transition kernel and mean rewards are unknown. The learner has access to a generative model, or simulator, which can be queried at any state–action pair (s, a). Each call returns a fresh reward and next-state sample (R, s') satisfying $R \in [0, 1]$, $\mathbb{E}[R \mid s, a] = r(s, a)$ and $s' \sim P(\cdot \mid s, a)$.

PAC learning objective and resource constraints. Given $\epsilon, \delta \in (0,1)$, the probably approximately correct (PAC) objective is to return a stationary deterministic policy $\widehat{\pi}$ such that

$$\mathbb{P}\Big(\rho^{\widehat{\pi}}(s) \geq \rho^\star - \epsilon \text{ for every } s \in \mathcal{S}\Big) \geq 1 - \delta. \tag{3}$$

The probability is over the simulator samples and the learner's internal randomness. We require the algorithm to terminate almost surely and measure sample complexity by a high-probability bound on its total number of simulator calls.

Notation. Let $P_{s,a} = P(\cdot \mid s,a)$ denote the transition row for (s,a), and write $P_{s,a}v := \sum_{s' \in \mathcal{S}} P(s' \mid s,a)v(s')$ for $v \in \mathbb{R}^S$. For a stationary policy π, define its transition kernel and mean reward by $P_\pi(s,s') := \sum_{a \in \mathcal{A}} \pi(a \mid s)P(s' \mid s,a)$ and $r_\pi(s) := \sum_{a \in \mathcal{A}} \pi(a \mid s)r(s,a)$, respectively. For a probability vector $p \in \Delta(\mathcal{S})$ and $v \in \mathbb{R}^S$, let $\mathbb{V}(p,v) := \sum_{s \in \mathcal{S}} p(s)v(s)^2 - \big(\sum_{s \in \mathcal{S}} p(s)v(s)\big)^2$ denote the variance of $v(s)$ when $s \sim p$. We write $\mathbf{1}$ for the S-dimensional all-ones vector and interpret vector inequalities componentwise. The notation $\widetilde{O}(\cdot)$ suppresses logarithmic factors. We use Γ to denote the Bellman operator, i.e., $(\Gamma h)_s = \max_a (r(s,a) + P_{s,a}h)$ for any $s \in \mathcal{S}$.

3 The Algorithm

In this section, we present our main algorithm, given in Algorithm 1. The algorithm tries the scales $G = 1, 2, 4, \ldots$ to adapt to the unknown optimal bias span H.

Each trial consists of three steps: (i) compute a coarse reference h using SAVIA (Lee et al., 2025), with accuracy chosen according to the current scale; (ii) compute a frozen lookahead estimate $\widehat{q}$ and use it to construct a reward-shaped correction problem; and (iii) solve the correction problem using space-efficient linear programming.

After obtaining an LP solution w, the algorithm checks whether $\mathrm{sp}(h - w)$ is small enough to certify that the extracted policy is ϵ-optimal. If the check passes, the algorithm returns the policy. If any check fails or a prescribed resource limit is reached before completion, the trial returns FAIL, and the algorithm proceeds to the next scale.

3.1 Coarse Reference Learning

The subroutine $\mathrm{Coarse}(G, t, \eta)$ is a capped version of SAVIA (Lee et al., 2025), with tolerance t and failure budget η. Starting from zero, it uses an iteration budget of $\lceil 128G/t \rceil$ and targets a uniform Bellman-oracle error of at most $t/256$. It returns a reference–policy pair (h, π_0) if its empirical residual and span checks pass, and FAIL if a check fails or the prescribed sample cap would be exceeded. We normalize h so that $\max_s h(s) = 0$ and choose $\pi_0(s) \in \arg\max_{a \in \mathcal{A}} \widehat{Q}(s,a)$, where $\widehat{Q}$ is the final action-value table.

The subroutine uses $\widetilde{O}(SAG^2/t^2)$ simulator samples and $\widetilde{O}(SA)$ working space. When $G \geq H$, it returns, with probability at least $1 - \eta$, a reference–policy pair (h, π_0) such that $\mathrm{sp}(h) \leq 8G$, $\mathrm{sp}(\Gamma h - h) \leq t$, and π_0 is t-optimal from every initial state. The implementation and analysis are given in Appendix A, with the formal guarantees stated in Lemma A.1.

Remark 3.1. SAVIA provides Bellman-residual and iterate-span bounds (Lee et al., 2025, Theorem 3.2), making it well suited to our reference-learning step. Policy-optimality guarantees alone, such as that of Zhang & Xie (2023), do not ensure these properties.

3.2 Reward-Shaping and Efficient Linear Programming

Given a coarse reference h satisfying $\mathrm{sp}(h) \leq 8G$ and $\mathrm{sp}(\Gamma h - h) \leq \tau$, we next solve a discounted planning problem for the remaining correction, with discount factor $\gamma = 1 - \epsilon/(64G)$.

To illustrate how the reference h facilitates planning, suppose first that $q(s,a) = r(s,a) + \gamma P_{s,a}h$ is known exactly. Define the shaped cost as $c(s,a) = \bar{c} + h(s) - q(s,a)$, where $\bar{c} = \max_{s,a}\{q(s,a) -$

Algorithm 1 RECOVER-AND-DISCARD: unknown-span planning

Input: Simulator; S states, A actions; accuracy ϵ; failure budget δ
Output: An ϵ-optimal policy;

for $i = 0, 1, 2, \ldots$ **do**
  $G \leftarrow 2^i$, $\delta_i \leftarrow \delta/[2(i+1)^2]$
  $\tau \leftarrow \left(G\epsilon^2/[S^5\sqrt{SA}]\right)^{1/3}$
  ▷ *Learn a reference*
  if $\tau \le \epsilon$ **then**
    $(h, \pi_0) \leftarrow$ COARSE$(G, \epsilon, \delta_i/4)$
    if the call succeeds **then**
      return π_0
    end if
    continue
  else if $\tau \ge 1$ **then**
    $\tau \leftarrow 1$, $h \leftarrow 0$ ▷ *no reference samples*
  else
    $(h, \pi_0) \leftarrow$ COARSE$(G, \tau, \delta_i/4)$
    if the call fails **then**
      continue
    end if
  end if
  ▷ *Reshape the reward*
  $\alpha \leftarrow \epsilon/(64G)$, $\gamma \leftarrow 1 - \alpha$
  Compute $\widehat{q}$ by equation 6, using fresh samples and budget $\delta_i/4$
  $\bar{c} \leftarrow \max_{s,a}\{\widehat{q}(s,a) - h(s)\}$
  $c(s,a) \leftarrow \bar{c} + h(s) - \widehat{q}(s,a)$ for all (s,a); $R \leftarrow 4\tau/\alpha$
  ▷ *Solve the correction LP with $\tilde{O}(SA)$ space*
  $(w, \pi) \leftarrow$ CORRECTIONLP$(c, \gamma, R, G, \delta_i/2)$
  if the call fails **then**
    continue
  end if
  ▷ *Certify the current policy*
  if $\alpha\,\mathrm{sp}(h - w) \le \epsilon/2$ **then**
    return π
  end if
end for

$h(s)\}$. This transformation preserves discounted policy suboptimality. Indeed, for every policy π,

$$V^{\pi}_{\gamma,c}(s) = \mathbb{E}_{\pi}\left[\sum_{t\ge 0}\gamma^t c(s_t,a_t)\,\middle|\, s_0 = s\right]$$
$$= \frac{\bar{c}}{1-\gamma} - \mathbb{E}_{\pi}\left[\sum_{t\ge 0}\gamma^t\big(r(s_t,a_t) + \gamma P_{s_t,a_t}h - h(s_t)\big)\,\middle|\, s_0 = s\right] = \frac{\bar{c}}{1-\gamma} + h(s) - V^{\pi}_{\gamma,r}(s), \tag{4}$$

where the last equality follows by telescoping the potential terms in expectation. Since the first two terms on the right-hand side are independent of π, minimizing the correction cost is equivalent to maximizing the original discounted reward. Thus, reward shaping changes the quantity we plan for without changing the underlying policy objective.

The key benefit is that the shaped problem has a *small optimal cumulative cost*. By construction, $c(s,a) \ge 0$ for every (s,a). Moreover, writing $\Gamma_\gamma h(s) = \max_a q(s,a)$ for the discounted Bellman operator applied to h, we obtain

$$\min_a c(s,a) = \bar{c} + h(s) - \Gamma_\gamma h(s) \le \mathrm{sp}(\Gamma_\gamma h - h) \le \mathrm{sp}(\Gamma h - h) + (1-\gamma)\mathrm{sp}(h) \le \tau + \frac{\epsilon}{8}.$$

Here we use the fact that $\|\Gamma_\gamma h - \Gamma h\|_\infty \le (1-\gamma)\mathrm{sp}(h)$. Hence, every state admits an action with cost at most $\tau + \epsilon/8$. A policy that selects such an action in each state incurs discounted cumulative cost at most $R := (\tau + \epsilon/8)/(1-\gamma)$. Consequently, the optimal correction value $w^\star(s) := \min_\pi V^{\pi}_{\gamma,c}(s)$ satisfies $0 \le w^\star(s) \le R$ for every state s.

To compute $w^\star$, we introduce the following discounted-cost linear program $\mathrm{LP}(p)$, parameterized by a transition kernel p:

$$\begin{aligned} \max_{w\in\mathbb{R}^S} \quad & \sum_s w(s) \\ \text{subject to} \quad & w(s) \le c(s,a) + \gamma p_{s,a} w, \quad \forall (s,a), \\ & -2R \le w(s) \le 2R, \qquad \forall s. \end{aligned} \tag{5}$$

Since the true transition kernel P is unknown, the coefficients of $\mathrm{LP}(P)$ are not directly available. We instead invoke $\mathrm{CorrectLP}(c,\gamma,R,G,\eta)$ to solve $\mathrm{LP}(\overline{P})$, where $\overline{P}$ is a sample-based approximation to P. For the moment, suppose that $\overline{P}$ is stored explicitly. Our space-efficient implementation replaces this explicit storage with row-query access: each requested row is reconstructed, used, and then discarded. The construction and guarantees of this interface are presented in Section 3.3.

On success, $\mathrm{CorrectLP}(c,\gamma,R,G,\eta)$ returns a nonnegative feasible near-optimal solution $\widehat{w}$ to $\mathrm{LP}(\overline{P})$. A final pass over the state-action pairs then extracts the stationary deterministic policy $\pi(s) \in \arg\min_a \{c(s,a) + \gamma \overline{P}_{s,a}\widehat{w}\}$, with ties broken according to a fixed rule.

Given $(\widehat{w},\pi)$, we accept the current trial and return π if $\mathrm{sp}(h-\widehat{w}) \le C_{\mathrm{stop}} G$; otherwise, we proceed to the next scale. Here $C_{\mathrm{stop}} = 32$ is a fixed numerical constant specified in the analysis. Its validity relies on the joint value-and-policy certificate for $(\widehat{w},\pi)$ after accounting for estimation errors, rather than on the span of an arbitrary vector.

Remark 3.2. A key advantage of correction planning is that the transition perturbation $\gamma(P_{s,a} - \overline{P}_{s,a})w$ acts on a value vector of scale R. In particular, for any $w \in [0,R]^S$, its magnitude is at most $\gamma R \|P_{s,a} - \overline{P}_{s,a}\|_1/2$. With an appropriate choice of R, the correction-planning stage contributes only a lower-order term to the overall sample complexity.

Estimation of $q(s,a)$. The preceding discussion assumes exact knowledge of $q(s,a)$. To remove this assumption, we draw a fresh batch of n_f independent samples from the generative model for each state-action pair and construct the estimate

$$n_f = \left\lceil C \log \frac{CSA(G+1)}{\epsilon\eta} \left[\frac{G}{\epsilon^2} + \frac{(1+\mathrm{sp}(h))^2}{\tau^2} \right] \right\rceil, \qquad \widehat{q}(s,a) = \frac{1}{n_f} \sum_{j=1}^{n_f} [Z_j + \gamma h(Y_j)], \tag{6}$$

where Z_j and Y_j denote the reward and next state in the fresh samples. We then set

$$\bar{c} = \max_{s,a} \{\hat{q}(s,a) - h(s)\}, \qquad c(s,a) = \bar{c} + h(s) - \hat{q}(s,a), \forall (s,a).$$

3.3 Calibrated Empirical Probability Kernel

The correction LP in Section 3.2 uses transition probabilities as coefficients, which the numerical solver accesses repeatedly across multiple passes. Our implementation must therefore supply the same approximate kernel $\overline{P}$ on every pass. Storing $\overline{P}$ explicitly would require $O(S^2A)$ entries. Instead, we retain a bias for each state-action pair and reconstruct only the transition row currently requested by the solver. Each reconstruction uses fresh simulator samples together with the rounding rule determined by the stored bias. We outline the idea below and defer the details to Appendix B.

Fix a state-action pair (s,a). A natural starting point is to project empirical transition estimates onto a discrete grid. However, accurate empirical evaluation need not produce identical projections, because the projection is discontinuous at the boundaries between rounding regions.

The key is therefore to make the rounding rule locally constant throughout a confidence region containing the true transition row. Our construction achieves this through a one-time calibration of bias-based rounding thresholds. Specifically, we apply the decoder to the square roots of the transition probabilities. A calibration batch estimates these roots, and we select a bias whose decoding rule is constant throughout the confidence interval around every estimated root. Once this bias is fixed, subsequent sufficiently accurate root estimates produce exactly the same decoded integer levels. Squaring and normalizing these levels therefore reconstructs the same probability row $\overline{P}_{s,a}$. On a joint high-probability event covering the prescribed finite number of queries, every reconstruction returns this fixed row. Only the selected bias for each state-action pair is retained between queries, requiring $\widetilde{O}(SA)$ space.

4 SAMPLE COMPLEXITY ANALYSIS

Theorem 4.1 (Span-adaptive, near-linear-space planning)**.** *With probability at least $1-\delta$, Algorithm 1 returns a stationary deterministic policy $\widehat{\pi}$ satisfying $\rho^{\widehat{\pi}}(s) \geq \rho^\star - \epsilon$ for every state, at a trial scale $G < 2H$, using*

$$\widetilde{O}\Big(\frac{SAH}{\epsilon^2} + \frac{S^{14/3}A^{4/3}H^{4/3}}{\epsilon^{4/3}}\Big) \tag{7}$$

samples and $\widetilde{O}(SA)$ working space.

In the rest of this section, we provide the sketched proof.

Recall $H = \max\{1, \mathrm{sp}(h^\star)\}$. Write $V_\gamma^\pi, V_\gamma^\star$ for the original discounted reward values under P.

4.1 A TRIAL WITH $G \geq H$ RETURNS A POLICY

Fix a reached trial with $G \geq H$, and first consider the correction branch $\epsilon < \tau < 1$. Recall $\alpha = 1-\gamma = \epsilon/(64G)$. Write V_γ^π and $V_\gamma^\star$ for the original discounted policy and optimal reward values under P.

By Lemma A.1, the call $\mathrm{Coarse}(G, \tau, \eta/3)$ succeeds with probability at least $1-\eta/3$, using $\widetilde{O}(SAG^2/\tau^2)$ samples. It returns a reference h satisfying $\max_s h(s) = 0$, $\mathrm{sp}(h) \leq 8G$, and $\mathrm{sp}(\Gamma h - h) \leq \tau$. We first work on this success event.

Given h, we recall that $\widehat{q}(s,a) = n_f^{-1}\sum_{j=1}^{n_f}[Z_j + \gamma h(Y_j)]$ for each pair (s,a), where Z_j and Y_j are its sampled reward and next state. Define the estimation error by $e(s,a) = \widehat{q}(s,a) - r(s,a) - \gamma P_{s,a}h$.

Lemma 4.2. *Conditional on the coarse-learning history, the prescribed batch size $n_f = \widetilde{O}(G/\epsilon^2 + G^2/\tau^2)$ ensures, with probability at least $1-\eta/3$, that*

$$\|e\|_\infty \leq \tau, \qquad \sup_\pi \big\|(I-\gamma P_\pi)^{-1}|e_\pi|\big\|_\infty \leq G/16, \tag{8}$$

where $e_\pi(s) = e(s, \pi(s))$.

On this event, define the fixed cost table by $c(s,a) = \bar{c} + h(s) - \widehat{q}(s,a)$, where $\bar{c} = \max_{s,a}\{\widehat{q}(s,a) - h(s)\}$. The costs are nonnegative by construction. Moreover, comparing $\max_a \widehat{q}(s,a) - h(s)$ with $\Gamma h(s) - h(s)$ gives

$$\begin{aligned}\min_a c(s,a) &\leq \mathrm{sp}(\Gamma h - h) + \alpha\,\mathrm{sp}(h) + 2\|e\|_\infty \\ &\leq 3\tau + \epsilon/8 < 4\tau \qquad \text{for every } s.\end{aligned}$$

Thus, selecting an immediate-cost-minimizing action in each state incurs discounted cost at most $R := 4\tau/\alpha$, regardless of the transition kernel. Writing W_p^π and $w_p^\star$ for the discounted policy and optimal values of this cost problem under kernel p, we obtain $0 \leq w_p^\star \leq R\mathbf{1}$ for every p. In particular, the true optimal correction value $w^\star := w_P^\star$ satisfies $\|w^\star\|_\infty \leq R$, and the box constraints in equation 5 do not exclude the optimum.

To relate this correction problem to the original rewards, let $u = h + (\bar{c}/\alpha)\mathbf{1}$. The identity $c(s,a) = u(s) - r(s,a) - e(s,a) - \gamma P_{s,a}u$ and discounted telescoping give $W_P^\pi = u - V_\gamma^\pi - (I-\gamma P_\pi)^{-1}e_\pi$. Consequently, Lemma 4.2 implies

$$\sup_\pi \|W_P^\pi - (u - V_\gamma^\pi)\|_\infty \leq G/16, \qquad \|w^\star - (u - V_\gamma^\star)\|_\infty \leq G/16. \tag{9}$$

Since $G \geq H$ and the discounted-span comparison gives $\mathrm{sp}(V_\gamma^\star) \leq 2H$, we also have $\mathrm{sp}(w^\star) \leq \mathrm{sp}(h) + 2H + G/8 < 11G$.

We now apply the complete correction-planning subroutine. The following lemma summarizes its statistical and numerical guarantees for the correction problem just constructed.

Lemma 4.3. *The call $\mathrm{CorrectLP}(c, \gamma, R, G, \eta/3)$ succeeds with probability at least $1-\eta/3$. With objective tolerance $\zeta = \alpha G/256$, it returns $\widehat{w} \in [0,R]^S$ and a stationary deterministic policy π satisfying*

$$\|\widehat{w} - w^\star\|_\infty \leq G/128 + \zeta, \qquad 0 \leq W_P^\pi - w^\star \leq (G/64 + 2\zeta/\alpha)\,\mathbf{1}. \tag{10}$$

Moreover, the sample cost in this call is bounded by $N_{\mathrm{LP}} = \tilde{O}(S^5(SA)^{3/2}G\tau/\epsilon^2)$.

Remark 4.4. The proof of Lemma 4.3 uses the recovery guarantee in Lemma B.1. Internally, the routine solves the LP for a fixed recovered kernel $\overline{P}$ and extracts a greedy policy. The optimization error is retained through ζ.

It remains to check that these outputs satisfy the stopping criterion. Combining equation 9 with Lemma 4.3 yields

$$\|u - \widehat{w} - V_\gamma^\star\|_\infty \le G/16 + G/128 + \zeta < G,$$
$$\|u - \widehat{w} - V_\gamma^\pi\|_\infty \le G/16 + 3G/128 + \zeta + 2\zeta/\alpha < G.$$

Since $u - h$ is constant across states, $\mathrm{sp}(h - \widehat{w}) = \mathrm{sp}(u - \widehat{w}) \le 2H + 2G \le 4G$. Therefore, $\alpha\,\mathrm{sp}(h - \widehat{w}) \le \epsilon/16 < \epsilon/2$, and the trial returns π. The intersection of the three success events has probability at least $1 - \eta$.

We now show that the returned policy π is ϵ-optimal for the original average-reward problem. Let d be the stationary distribution of any recurrent class of P_π, and write $r_\pi(s) = r(s, \pi(s))$. Multiplying the discounted Bellman equation $V_\gamma^\pi = r_\pi + \gamma P_\pi V_\gamma^\pi$ by $d^\top$ gives $d^\top r_\pi = \alpha d^\top V_\gamma^\pi \ge \alpha \min_x V_\gamma^\pi(x)$. The average reward from any initial state is a convex combination of the recurrent-class rewards, so $\rho^\pi(s) \ge \alpha \min_x V_\gamma^\pi(x)$ for every s. Applying the same stationary-distribution identity to an average-reward optimal stationary policy $\pi^\star$, and using $V_\gamma^{\pi^\star} \le V_\gamma^\star$, gives $\rho^\star \le \alpha \max_x V_\gamma^\star(x)$.

The preceding bounds $\|u - \hat{w} - V_\gamma^\star\|_\infty \le G$ and $\|u - \hat{w} - V_\gamma^\pi\|_\infty \le G$ therefore imply, for every state s,

$$\begin{aligned}\rho^\star - \rho^\pi(s) &\le \alpha\left(\max_x V_\gamma^\star(x) - \min_x V_\gamma^\pi(x)\right) \le \alpha\left(\max_x(u(x) - \hat{w}(x)) - \min_x(u(x) - \hat{w}(x)) + 2G\right) \\ &= \alpha\,\mathrm{sp}(h - \widehat{w}) + 2\alpha G \le \frac{\epsilon}{2} + \frac{\epsilon}{32} < \epsilon,\end{aligned}$$

where the last inequality uses the stopping condition and $\alpha = \epsilon/(64G)$. Thus, the trial returns a stationary deterministic policy that is ϵ-optimal from every initial state.

The three stages require $N_{\mathrm{coarse}} = \widetilde{O}(SAG^2/\tau^2)$ samples to learn the coarse reference, $N_{\mathrm{freeze}} = \widetilde{O}\big(SA[G/\epsilon^2 + G^2/\tau^2]\big)$ samples to compute $\widehat{q}$, and $N_{\mathrm{LP}} = \widetilde{O}\big(S^5(SA)^{3/2}G\tau/\epsilon^2\big)$ samples to solve the correction LP. Summing these terms gives the following bound

$$N_{\mathrm{total}} = \widetilde{O}\left(\frac{SAG}{\epsilon^2} + \frac{SAG^2}{\tau^2} + \frac{S^5(SA)^{3/2}G\tau}{\epsilon^2}\right) = \widetilde{O}\left(\frac{SAH}{\epsilon^2} + \frac{S^{14/3}A^{4/3}H^{4/3}}{\epsilon^{4/3}}\right)$$

4.2 Validity of an early return

Now fix a reached trial with $1 \le G < H$. Assume the good event in Lemma A.1. If the trial returns the coarse policy directly at tolerance ϵ, its ϵ-optimality already holds.

It remains to consider an accepted correction trial, with $\epsilon < \tau \le 1$, $\alpha = \epsilon/(64G)$, and $R = 4\tau/\alpha$. For $\tau < 1$, any returned reference satisfies $\max_s h(s) = 0$, $\mathrm{sp}(h) \le 8G$, and $\mathrm{sp}(\Gamma h - h) \le \tau$; for $\tau = 1$, these properties hold for $h = 0$. The concentration result in Lemma 4.2 requires only these reference bounds, not $G \ge H$. Consequently, the same construction of $\widehat{q}$ and c gives $0 \le w^\star \le R\mathbf{1}$ and the shaping comparison equation 9.

Unlike the preceding case, we cannot use $G \ge H$ to control the span of the optimal discounted value. Instead, acceptance gives $\alpha\,\mathrm{sp}(h - \widehat{w}) \le \epsilon/2$, and hence

$$\mathrm{sp}(\widehat{w}) \le \mathrm{sp}(h) + \mathrm{sp}(h - \widehat{w}) \le 8G + \frac{\epsilon}{2\alpha} = 40G.$$

This observable bound allows us to apply the following output-dependent form of the correction-planning guarantee.

Lemma 4.5. *Consider the call* $\mathrm{CorrectLP}(c, \gamma, R, G, \eta/3)$. *There is an event of probability at least* $1 - \eta/3$ *on which every returned pair* $(\widehat{w}, \pi)$ *satisfying* $\mathrm{sp}(\widehat{w}) \le 40G$ *obeys*

$$\|\widehat{w} - w^\star\|_\infty \le G/128 + \zeta, \qquad 0 \le W_P^\pi - w^\star \le (G/64 + 2\zeta/\alpha)\,\mathbf{1}, \tag{11}$$

where $\zeta = \alpha G/256$. *The sample and space bounds are the same as in Lemma 4.3.*

Since the stopping test ensures the hypothesis of Lemma 4.5, we recover the same value-and-policy certificate as in the preceding case. Specifically, let $u = h + (\bar{c}/\alpha)\mathbf{1}$. Combining equation 9 with equation 11 yields

$$\|u - \widehat{w} - V_\gamma^\star\|_\infty \le G/16 + G/128 + \zeta < G,$$
$$\|u - \widehat{w} - V_\gamma^\pi\|_\infty \le G/16 + 3G/128 + \zeta + 2\zeta/\alpha < G.$$

Therefore, for every state s,

$$\rho^\star - \rho^\pi(s) \le \alpha \left(\max_x V_\gamma^\star(x) - \min_x V_\gamma^\pi(x) \right)$$
$$\le \alpha \,\mathrm{sp}(u - \widehat{w}) + 2\alpha G = \alpha \,\mathrm{sp}(h - \widehat{w}) + 2\alpha G \le \frac{\epsilon}{2} + \frac{\epsilon}{32} < \epsilon.$$

The sample-complexity analysis is identical to that for $G \ge H$: the prescribed sample caps depend only on the current trial parameters. Thus, the same per-trial sample bound applies to an early return.

5 Numerical Experiments

Table 1 ranks six guarantees by asymptotic sample cost; we ask whether that order is visible at a budget every arm can afford. We run all six arms on one simulator at $n = 4000$ samples per state–action pair, three seeds per instance (Table 2). For clearer comparison, we use *consistency* as the metric, which is the fraction of states at which the returned policy attains $\max_a\{r(s,a) + P_{sa}h^\star\}$; an arm that does not return a policy is recorded and is not scored. Please see Appendix E for details of the true loss.

Table 2: Consistency at a fixed budget of $n = 4000$ samples per state–action pair, three seeds per instance, sorted by the suite-wide mean. "all-18" averages over the whole suite; "returns" counts the 54 seed–instance pairs that produced a policy at all.

Method	H input	kernel	all-18	returns	sd
RAD (this work)	No	No	**0.931**	54/54	0.045
(Zurek & Chen, 2024)	Yes	Yes	0.819	54/54	0.062
(Zurek & Chen, 2025)	No	Yes	0.819	54/54	0.068
(Wang et al., 2022)	Yes	Yes	0.817	54/54	0.061
(Lee et al., 2025)	No	No	0.761	51/54	0.000
(Zhang & Xie, 2023)	Yes	No	0.757	54/54	0.000

The suite consists of 18 instances which are H-controlled. RAD leads at 0.931, by 0.112 over the next best and 0.174 over the least accurate, and the resource columns make the same point in reverse: the three arms within 0.12 of it retain a transition kernel, two of them are handed H, and SAVIA, the only other arm that takes neither, scores 0.761 and is the only arm that fails to return a policy on some runs (51/54). RAD returns on all 54 seed–instance pairs with the smallest spread of any arm that moves (0.045; the others run 0.061 to 0.068): the arm given the least information and the least storage is the most accurate and the most reliable of the six.

Appendix E resolves the same runs per instance and per budget, where RAD's margin is widest; the comparison remains one simulator, one suite and one budget.

6 Discussion

In this work, we develop an algorithm that uses $\widetilde{O}(SA)$ working space and attains the optimal sample complexity $\widetilde{O}(SAH/\epsilon^2)$ asymptotically as $\epsilon \to 0$ for fixed S, A, H. Several directions merit further investigation. First, reducing or eliminating the lower-order burn-in terms would strengthen the guarantee beyond the asymptotic regime. Second, extending these ideas to online learning may help improve the dependence on S, A, H in model-free regret bounds while retaining near-linear working space. Finally, developing simpler and faster alternatives to the LP-based correction solver, without sacrificing the sample and space guarantees, would help translate the theoretical framework into practical algorithms.

AI USE STATEMENT

Generative AI tools were used to assist with exploring and refining algorithmic ideas, developing and auditing proposed proofs, locating and checking references, constructing numerical sanity-check scripts, and improving the exposition and LaTeX source. Specifically, AI tools were used to analyze the error guarantees and space complexity of the LP subroutine. The authors take full responsibility for the final manuscript and associated artifacts, including the correctness of the mathematical claims, the accuracy of the references, and the presentation of any numerical results.

REPRODUCIBILITY STATEMENT

Section 2 specifies the MDP assumptions, simulator access, learning objective, and resource constraints. Algorithm 1 and Section 3 describe the span-adaptive procedure, reward shaping, and correction-planning framework. Appendix A details the capped coarse-reference routine, Appendix B specifies the calibration and recovery construction, and Appendix C proves the frozen-estimation guarantee. Appendix D presents the numerical planning and kernel-transfer analysis, including the accounting of transition-row queries and simulator samples. Section 4 assembles these components in the analysis of Theorem 4.1.

ETHICS STATEMENT

This theoretical study uses no human-subject data, private datasets, or deployed decision system. Its potential benefits concern memory-constrained planning. We state the assumptions, asymptotic scope, model-reconstruction character, and substantial computational overhead explicitly to avoid suggesting an immediately deployable reinforcement-learning system.

REFERENCES

Jinane Abounadi, Dimitrib Bertsekas, and Vivek S Borkar. Learning algorithms for markov decision processes with average cost. *SIAM Journal on Control and Optimization*, 40(3):681–698, 2001.

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J. Kappen. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine Learning*, 91(3): 325–349, 2013. doi: 10.1007/s10994-013-5368-1.

Peter L. Bartlett and Ambuj Tewari. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *Uncertainty in Artificial Intelligence: Proceedings of the 25th Conference*, pp. 35–42. AUAI Press, 2009.

Victor Boone and Zihan Zhang. Achieving tractable minimax optimal regret in average reward mdps. *Advances in Neural Information Processing Systems*, 37:26728–26769, 2024.

Li Chen, Wayne Eberly, Erich Kaltofen, B. David Saunders, William J. Turner, and Gilles Villard. Efficient matrix preconditioners for black box linear algebra. *Linear Algebra and its Applications*, 343–344:119–146, 2002. doi: 10.1016/S0024-3795(01)00472-4.

Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Ronald Ortner. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *International Conference on Machine Learning*, pp. 1578–1586. PMLR, 2018.

Jacek Gondzio. Matrix-free interior point method. *Computational Optimization and Applications*, 51 (2):457–480, 2012. doi: 10.1007/s10589-010-9361-3.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(51):1563–1600, 2010.

Yujia Jin and Aaron Sidford. Efficiently solving MDPs with stochastic mirror descent. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4890–4900. PMLR, 2020.

Yujia Jin and Aaron Sidford. Towards tight bounds on the sample complexity of average-reward mdps. In *International Conference on Machine Learning*, pp. 5055–5064. PMLR, 2021.

Yujia Jin, Ishani Karmarkar, Aaron Sidford, and Jiayi Wang. Truncated variance reduced value iteration. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Jongmin Lee, Mario Bravo, and Roberto Cominetti. Near-optimal sample complexity for MDPs via anchoring. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 32907–32929. PMLR, 2025.

Yiping Liu, Hoai An Nguyen, and Junzhao Yang. Numerical linear algebra in linear space. In *Proceedings of the 2026 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 6392–6435. SIAM, 2026. doi: 10.1137/1.9781611978971.230. Appendix A cites Corollary 10 in the complete version, arXiv:2507.02433v1; that numbering is not the SODA one.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization, 2009.

Arkadi S. Nemirovski and Michael J. Todd. Interior-point methods for optimization. *Acta Numerica*, 17:191–234, 2008. doi: 10.1017/S0962492906370018.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, New York, 1994.

J. Barkley Rosser and Lowell Schoenfeld. Approximate formulas for some functions of prime numbers. *Illinois Journal of Mathematics*, 6(1):64–94, 1962. doi: 10.1215/ijm/1255631807.

Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3 (1):9–44, 1988.

Richard S Sutton, Andrew G Barto, and Andrew Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Jinghan Wang, Mengdi Wang, and Lin F. Yang. Near sample-optimal reduction-based policy learning for average reward MDP, 2022.

Mengdi Wang. Primal-dual π learning: Sample complexity and sublinear run time for ergodic markov decision problems. *arXiv preprint arXiv:1710.06100*, 2017.

Mengdi Wang. Randomized linear programming solves the Markov decision problem in nearly linear (sometimes sublinear) time. *Mathematics of Operations Research*, 45(2):517–546, 2020. doi: 10.1287/moor.2019.1000.

Shengbo Wang, Jose Blanchet, and Peter Glynn. Optimal sample complexity for average reward markov decision processes. In *International Conference on Learning Representations*, volume 2024, pp. 29843–29861, 2024.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3):279–292, 1992.

Chen-Yu Wei, Mehdi Jafarnia Jahromi, Haipeng Luo, Hiteshi Sharma, and Rahul Jain. Model-free reinforcement learning in infinite-horizon average-reward markov decision processes. In *International conference on machine learning*, pp. 10170–10180. PMLR, 2020.

Douglas H. Wiedemann. Solving sparse linear equations over finite fields. *IEEE Transactions on Information Theory*, 32(1):54–62, 1986. doi: 10.1109/TIT.1986.1057137.

Zihan Zhang and Xiangyang Ji. Regret minimization for reinforcement learning by evaluating the optimal bias function. *Advances in Neural Information Processing Systems*, 32, 2019.

Zihan Zhang and Qiaomin Xie. Sharper model-free reinforcement learning for average-reward Markov decision processes. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 5476–5477. PMLR, 2023. Extended abstract; full version arXiv:2306.16394.

Matthew Zurek and Yudong Chen. Span-based optimal sample complexity for weakly communicating and general average reward MDPs. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Matthew Zurek and Yudong Chen. Span-agnostic optimal sample complexity and oracle inequalities for average-reward RL. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 6156–6209. PMLR, 2025.

ADDITIONAL NOTATIONS

We use $\log$ for the natural logarithm and $\log_2$ for binary precision. All unspecified constants denoted by C can be replaced by fixed sufficiently large universal constants. Only logarithmic powers are suppressed in $\widetilde{O}$. There is no hidden transition-model storage or access to a random tape that reproduces simulator outputs. A simulator observation is processed online and discarded. Standard integer and fixed-point arithmetic count their actual bit storage.

A LEARNING THE COARSE REFERENCE

We use a capped version of SAVIA (Lee et al., 2025, Algorithm 2), adapted to sampled rewards. We treat arithmetic as exact and measure space by the number of stored scalar entries. For a state-action table Q, write $h_Q(s) = \max_{a\in\mathcal{A}} Q(s,a)$ and $\mathcal{T}Q(s,a) = r(s,a) + P_{s,a}h_Q$. The corresponding state-value Bellman operator is $(\Gamma v)(s) = \max_{a\in\mathcal{A}}\{r(s,a) + P_{s,a}v\}$.

Parameters. For $G \geq 1$, $t \in (0,1]$, and $\eta \in (0,1)$, set $K = \lceil 128G/t \rceil$, $\zeta = t/256$, $\xi = \zeta/4$, $\ell = \log \frac{16SA(K+1)}{\eta}$ and $n_0 = \left\lceil 32\zeta^{-2} \log \frac{8SA}{\eta} \right\rceil$. For each $k \geq 0$, let $\beta_k = k/(k+2)$ and $a_k = 5(k+2)\log^2(k+2)$. We use the deterministic sample cap

$$N_{\max} = \left\lceil CSA\frac{G^2}{t^2}\left[1 + \log\frac{CSA(G+1)}{t\eta}\right]^6 \right\rceil, \tag{12}$$

where C is a sufficiently large universal constant. The algorithm returns FAIL before drawing any batch that would exceed this cap.

The reward estimate $\widehat{r}(s,a)$ is the mean of n_0 fresh reward observations at (s,a). Each observation lies in $[0,1]$ and has mean $r(s,a)$. This estimate is fixed throughout the call. At iteration k, the vector $d^k = h^k - h^{k-1}$ is the change in the state-value estimate, and $D^k(s,a)$ is a sample mean estimating $P_{s,a}d^k$. All batches use fresh simulator calls.

Lemma A.1 (Coarse reference guarantee)**.** *For every $G \geq 1$, $t \in (0,1]$, and $\eta \in (0,1)$, Algorithm 2 terminates after at most $N_{\max}$ simulator calls and stores $O(SA)$ scalar entries.*

There is an event of probability at least $1-\eta$ on which any returned pair (h, π_0) consists of a reference vector and a stationary deterministic policy satisfying

$$\begin{aligned} \max_s h(s) = 0, \qquad & \mathrm{sp}(h) \leq 8G, \\ \mathrm{sp}(\Gamma h - h) \leq t, \qquad & \rho^{\pi_0}(s) \geq \rho^\star - t \quad \forall s \in \mathcal{S}. \end{aligned} \tag{13}$$

On the same event, whenever $G \geq H$, the algorithm returns (h, π_0) rather than FAIL*. In particular, for $G \geq H$, $\mathbb{P}(\mathrm{Coarse}(G,t,\eta) \neq \mathrm{FAIL}) \geq 1-\eta$.*

Proof. Because $h^{-1} = 0$ and $d^j = h^j - h^{j-1}$, we have $h^k = \sum_{j=0}^k d^j$. Combining this identity with the recursive update of B^k gives

$$B^k - \mathcal{T}Q^k = \widehat{r} - r + \sum_{j=0}^k D^j - P\sum_{j=0}^k d^j = (\widehat{r} - r) + \sum_{j=0}^k (D^j - Pd^j), \tag{14}$$

where $(Pd^j)(s,a) = P_{s,a}d^j$.

The choice of n_0 and Hoeffding's inequality imply $\|\widehat{r} - r\|_\infty \leq \zeta/4$ with failure probability at most $\eta/4$. At each iteration j, the vector d^j and batch size m_j are determined before the fresh transition samples are drawn. Hence $D^j - Pd^j$ is conditionally centered. Since $2\sum_{j\geq 0} a_j^{-1} \leq 1$, the conditional sampling argument of Lee et al. (2025, Proposition 3.1), applied with accuracy ξ, yields $\max_{k\leq K} \| \sum_{j=0}^k (D^j - Pd^j)\|_\infty \leq \xi$ with failure probability at most $\eta/8$.

Algorithm 2 $\mathrm{Coarse}(G, t, \eta)$

Input: Scale $G \geq 1$, tolerance $t \in (0, 1]$, failure budget $\eta \in (0, 1)$

Set $K, \zeta, \xi, \ell, n_0, N_{\max}$ as above, $h^{-1} \leftarrow 0$
if $SAn_0 > N_{\max}$ **then**
  return FAIL
end if
for each $(s, a) \in \mathcal{S} \times \mathcal{A}$ **do**
  Draw n_0 fresh rewards $Z_1, \ldots, Z_{n_0}$ at (s, a)
  $\widehat{r}(s, a) \leftarrow n_0^{-1} \sum_{j=1}^{n_0} Z_j$
end for
$B^{-1} \leftarrow \widehat{r}$, $h^{-1} \leftarrow 0$, $N \leftarrow SAn_0$
for $k = 0, \ldots, K$ **do**
  $Q^k \leftarrow \beta_k B^{k-1}$
  $h^k(s) \leftarrow \max_a Q^k(s, a)$ for every s; $d^k \leftarrow h^k - h^{k-1}$
  $m_k \leftarrow \max\{1, \lceil \ell a_k \, \mathrm{sp}(d^k)^2 / \xi^2 \rceil\}$
  if $N + SAm_k > N_{\max}$ **then**
    return FAIL
  end if
  for each $(s, a) \in \mathcal{S} \times \mathcal{A}$ **do**
    Draw m_k fresh next states $Y_1, \ldots, Y_{m_k}$ at (s, a)
    $D^k(s, a) \leftarrow m_k^{-1} \sum_{j=1}^{m_k} d^k(Y_j)$
    $B^k(s, a) \leftarrow B^{k-1}(s, a) + D^k(s, a)$
  end for
  $N \leftarrow N + SAm_k$
end for
$h \leftarrow h^K - (\max_s h^K(s))\mathbf{1}$
Choose $\pi_0(s) \in \arg\max_a Q^K(s, a)$ for every s, with fixed tie-breaking
if $\mathrm{sp}(Q^K - B^K) \leq t/4$ and $\mathrm{sp}(h) \leq 8G$ **then**
  return (h, π_0)
end if
return FAIL

Consequently, there is an event $\mathcal{E}$ with $\mathbb{P}(\mathcal{E}) \geq 1 - \eta$ on which

$$\max_{k \leq K} \|B^k - \mathcal{T}Q^k\|_\infty \leq \|\widehat{r} - r\|_\infty + \max_{k \leq K} \left\| \sum_{j=0}^{k} (D^j - Pd^j) \right\|_\infty \leq \zeta/4 + \xi \leq \zeta. \tag{15}$$

All errors are measured relative to the original reward r. The initialization error $\widehat{r} - r$ appears only once in equation 14; it does not accumulate over iterations.

Work on $\mathcal{E}$. Every entry of D^k is an average of coordinates of d^k, so $\mathrm{sp}(D^k) \leq \mathrm{sp}(d^k)$. Together with equation 15 and the exact update $Q^k = \beta_k B^{k-1}$, this is the structure required by the deterministic SAVIA analysis.

Let $Q^\star$ be an optimal action-bias table satisfying $\mathcal{T}Q^\star = Q^\star + \rho^\star \mathbf{1}$, and write $\mu = \mathrm{sp}(Q^\star)$. Since $Q^0 = 0$, the bounds of Lee et al. (2025, Lemmas A.1–A.2 and Theorem 3.2) give

$$\mathrm{sp}(Q^k - Q^\star) \leq \mu + \frac{2}{3} k\zeta, \tag{16}$$

$$\mathrm{sp}(d^k) \leq \frac{4\mu}{k+1} + 2\zeta, \qquad k \geq 1, \tag{17}$$

$$\mathrm{sp}(Q^k - \mathcal{T}Q^k) \leq \frac{8\mu}{k+2} + 4\zeta. \tag{18}$$

These deterministic arguments use the uniform oracle-error bound and the span bound on the sampling increments. Thus the sampled-reward initialization is already accounted for by equation 15.

Validity of an accepted pair. For any table Q, taking statewise maxima in the coordinatewise bounds on $\mathcal{T}Q - Q$ gives

$$\mathrm{sp}(\Gamma h_Q - h_Q) \leq \mathrm{sp}(\mathcal{T}Q - Q). \tag{19}$$

If the algorithm accepts, then

$$\mathrm{sp}(\mathcal{T}Q^K - Q^K) \le \mathrm{sp}(B^K - Q^K) + 2\|B^K - \mathcal{T}Q^K\|_\infty \le t/4 + 2\zeta < t.$$

Subtracting a constant from h^K does not change its span or its Bellman residual. Hence the returned h satisfies equation 13, including the normalization and the explicit span check. Moreover, π_0 is greedy with respect to Q^K, so Lee et al. (2025, Proposition 2.1) gives $\rho^\star - \rho^{\pi_0}(s) \le \mathrm{sp}(\mathcal{T}Q^K - Q^K) < t$ for every initial state s. This argument does not require $G \ge H$.

Success when $G \ge H$. Choose $Q^\star(s,a) = r(s,a) + P_{s,a}h^\star - \rho^\star$. Then $\mu \le 1 + \mathrm{sp}(h^\star) \le 2G$. The bounds above imply

$$\mathrm{sp}(Q^K - B^K) \le \frac{16G}{K+2} + 6\zeta \le \frac{t}{8} + \frac{6t}{256} < \frac{t}{4},$$
$$\mathrm{sp}(h) \le \mathrm{sp}(Q^K) \le 2\mu + \frac{2}{3}K\zeta \le 4G + \frac{2}{3}K\zeta < 8G.$$

Thus both acceptance checks pass if the final iteration is reached.

It remains to verify that the sample cap does not interrupt this execution. Since $d^0 = 0$, we have $m_0 = 1$. For $k \ge 1$, equation 17 gives $\mathrm{sp}(d^k) \le 8G/(k+1) + 2\zeta$. Using the definition of a_k, we have $\sum_{k=1}^K a_k/(k+1)^2 = O(\log^3(K+2))$ and $\sum_{k=1}^K a_k = O(K^2\log^2(K+2))$. Therefore, the same batch-size summation as in Lee et al. (2025, Theorem 3.2) yields

$$\begin{aligned} n_0 + \sum_{k=0}^{K} m_k &\le n_0 + K + 1 + \frac{C_1\ell}{\xi^2}\left[G^2\log^3(K+2) + \zeta^2K^2\log^2(K+2)\right] \\ &\le C_2\frac{G^2}{t^2}\ell\log^3(K+2), \\ &\le N_{\max}. \end{aligned}$$

As a result, the execution will not stop until the K-th round. The proof is completed.

□

B SQUARE-GRID RECOVERY AND DIRECTIONAL ACCURACY

We define the algorithm and the recovery lemma for a general multinomial distribution.

Let $p \in \Delta(\mathcal{S})$ be accessible through independent samples. Fix a failure probability $\delta \in (0,1)$ and an upper bound $T \ge 1$ on the number of recovery calls, and write $\mathrm{Log} := \log(2\mathcal{S}(T+1)/\delta)$. We construct a discretized distribution that can be recovered repeatedly using N fresh samples per call, while retaining only one distribution-dependent scalar b. The construction rounds square-root probabilities after adding a common shift, then squares and normalizes the rounded coordinates. For $\lambda > 0$, define $R_\lambda(x) := \lambda\lfloor x/\lambda + 1/2\rfloor$ and the set of rounding midpoints $\mathcal{M}_\lambda := \{(k+1/2)\lambda : k \in \mathbb{Z}\}$. Recall $\mathrm{sp}(v) := \max_i v_i - \min_i v_i$ and $\mathrm{Var}_p(v) := \sum_i p_i(v_i - pv)^2$.

Lemma B.1. *Let b be the shift selected by Algorithm 3, and define its canonical distribution by $\overline{p}_i := z_i^2/Z$, where $z_i := R_\lambda(\sqrt{p_i} + b)$ and $Z := \sum_j z_j^2$. With probability at least $1-\delta$ over calibration and all recovery samples, every one of the T recovery calls returns exactly $\overline{p}$, using N fresh samples per call.*

Proof. For an empirical distribution $\widehat{p}$ based on N independent samples, the binomial Chernoff bound gives $\Pr(|\sqrt{\widehat{p}_i} - \sqrt{p_i}| > \rho) \le 2e^{-N\rho^2}$. Indeed, the Bernoulli relative entropy satisfies $D_{\mathrm{Ber}}(q\|p_i) \ge (\sqrt{q} - \sqrt{p_i})^2$. A union bound over the S coordinates and the $T+1$ batches therefore shows that, with probability at least $1-\delta$, every empirical distribution used by the algorithm satisfies $|\sqrt{\widehat{p}_i} - \sqrt{p_i}| \le \rho$ for all i. Work on this event.

The residues θ_i are exactly the shifts modulo λ that place a calibrated coordinate $\sqrt{\widehat{p}_i^0} + b$ on a rounding midpoint. Their S circular gaps sum to λ, so the longest gap has length at least λ/S. Choosing its

Algorithm 3 Single-shift grid calibration and recovery

Input: Sample budget N and sampling access to p; public parameters S, T, δ.

1: Set $\rho \leftarrow \sqrt{\mathrm{Log}/N}$ and $\lambda \leftarrow 8S\rho$.
Calibration: performed once using N samples.
2: Draw $X_1, \ldots, X_N \sim p$ independently and set $\widehat{p}_i^0 \leftarrow N^{-1} \sum_{n=1}^N \mathbf{1}\{X_n = i\}$ for $i \in [S]$.
3: Compute $\theta_i \leftarrow (\lambda/2 - \sqrt{\widehat{p}_i^0}) \bmod \lambda$ for $i \in [S]$.
4: Sort these residues as $\theta_{(1)} \le \cdots \le \theta_{(S)}$ and set $\theta_{(S+1)} \leftarrow \theta_{(1)} + \lambda$.
5: Choose $j \in \arg\max_{k \in [S]} (\theta_{(k+1)} - \theta_{(k)})$.
6: Let b be the unique number in $[\lambda/2, 3\lambda/2)$ congruent to $(\theta_{(j)} + \theta_{(j+1)})/2$ modulo λ.
7: Retain b and discard the empirical distribution and residues.
Recovery: performed on each query using N fresh samples.
8: Draw $Y_1, \ldots, Y_N \sim p$ independently and set $\widehat{p}_i \leftarrow N^{-1} \sum_{n=1}^N \mathbf{1}\{Y_n = i\}$ for $i \in [S]$.
9: Set $\widehat{z}_i \leftarrow R_\lambda(\sqrt{\widehat{p}_i} + b)$ for $i \in [S]$.
10: **return** $\overline{p}^{\mathrm{rec}}$, where $\overline{p}_i^{\mathrm{rec}} := \widehat{z}_i^2 / \sum_{j=1}^S \widehat{z}_j^2$.

midpoint yields $\mathrm{dist}(\sqrt{\widehat{p}_i^0} + b, \mathcal{M}_\lambda) \ge \lambda/(2S) = 4\rho$ for every i. Since $|\sqrt{\widehat{p}_i^0} - \sqrt{p_i}| \le \rho$, the true coordinate lies in the same rounding cell and satisfies $\mathrm{dist}(\sqrt{p_i} + b, \mathcal{M}_\lambda) \ge 3\rho$.

Every recovery estimate is within ρ of the true square-root coordinate. Consequently, no recovery coordinate crosses a rounding boundary, and $\widehat{z}_i = R_\lambda(\sqrt{p_i} + b) = z_i$ in every call. Squaring and normalizing thus returns exactly $\overline{p}$. Finally, $b \ge \lambda/2$ implies $R_\lambda(x + b) \ge x$ for $x \ge 0$, so $Z \ge \sum_i p_i = 1$. □

Lemma B.2. *For every calibration outcome, the canonical distribution $\overline{p}$ in Lemma B.1 satisfies, simultaneously for every $v \in \mathbb{R}^S$, the bound $|(p - \overline{p})v| \le 32\sqrt{S^3 \mathrm{Log}\, \mathrm{Var}_p(v)/N} + 256(S^3 \mathrm{Log}/N)\,\mathrm{sp}(v)$. In particular, with probability at least $1 - \delta$, this bound holds for every recovered output and every v.*

Proof. Let $e_i := z_i - \sqrt{p_i}$ and $\mu := pv$. Nearest-grid rounding has error at most $\lambda/2$, and $b \in [\lambda/2, 3\lambda/2)$, so $0 \le e_i \le 2\lambda$. Hence $\|e\|_2 \le 2\sqrt{S}\,\lambda$ and $Z \ge 1$. Using $\sum_i p_i(v_i - \mu) = 0$ and expanding $z_i^2 = (\sqrt{p_i} + e_i)^2$ gives $(\overline{p} - p)v = Z^{-1} \sum_i (2\sqrt{p_i} e_i + e_i^2)(v_i - \mu)$. Cauchy–Schwarz bounds the linear contribution by $|\sum_i \sqrt{p_i} e_i (v_i - \mu)| \le \|e\|_2 \sqrt{\mathrm{Var}_p(v)}$, while $|v_i - \mu| \le \mathrm{sp}(v)$ bounds the quadratic contribution by $|\sum_i e_i^2 (v_i - \mu)| \le \|e\|_2^2\,\mathrm{sp}(v)$. Therefore $|(\overline{p} - p)v| \le 2\|e\|_2 \sqrt{\mathrm{Var}_p(v)} + \|e\|_2^2\,\mathrm{sp}(v) \le 4\sqrt{S}\,\lambda \sqrt{\mathrm{Var}_p(v)} + 4S\lambda^2\,\mathrm{sp}(v)$. Substituting $\lambda = 8S\sqrt{\mathrm{Log}/N}$ proves the claimed bound. The argument is deterministic and uniform in v; Lemma B.1 transfers it to all recovered outputs. □

C Proof of Lemma 4.2

Condition on a coarse-learning history for which $\max_s h(s) = 0$, $\mathrm{sp}(h) \le 8G$, and $\mathrm{sp}(\Gamma h - h) \le \tau$. Write $B_h = \mathrm{sp}(h)$, so that $-B_h \mathbf{1} \le h \le 0$, and let $\ell = \log(12SA/\eta)$. It suffices to take $n_f \ge C\ell[G/\epsilon^2 + (1 + B_h)^2/\tau^2]$ for a sufficiently large universal constant C. Since $B_h \le 8G$ and $G \ge 1$, this has the stated order.

We first bound the transition variance of the reference. Let $c_0 = \max_s (\Gamma h - h)(s)$. The Bellman-residual comparison gives $0 \le c_0 \le \rho^\star + \tau \le 2$. For each (s, a), put $\Delta(s, a) = P_{s,a} h - h(s)$. Since $\Delta(s, a) \le c_0 - r(s, a) \le c_0$ and $h(s) \le 0$,

$$\begin{aligned} \mathrm{Var}_{P_{s,a}}(h) &= P_{s,a} h^2 - h(s)^2 - 2h(s)\Delta(s, a) - \Delta(s, a)^2 \\ &\le P_{s,a} h^2 - h(s)^2 + 2B_h c_0. \end{aligned} \tag{20}$$

Fix any stationary deterministic policy π, and write $\mathcal{R}_\pi = (I - \gamma P_\pi)^{-1}$. The matrix $\alpha \mathcal{R}_\pi$ has nonnegative entries and every row sums to one. Applying it to equation 20 along π gives the

coordinatewise bound

$$\begin{aligned}\alpha\mathcal{R}_\pi \operatorname{Var}_{P_\pi}(h) &\le \alpha\mathcal{R}_\pi(P_\pi h^2 - h^2) + 2B_h c_0 \mathbf{1} \\ &= \frac{\alpha}{\gamma}\big(\alpha\mathcal{R}_\pi h^2 - h^2\big) + 2B_h c_0 \mathbf{1} \\ &\le \left(2B_h c_0 + \frac{\alpha B_h^2}{\gamma}\right)\mathbf{1} \le 34G\mathbf{1}. \end{aligned} \tag{21}$$

The equality follows from $\mathcal{R}_\pi P_\pi = (\mathcal{R}_\pi - I)/\gamma$. The last inequality uses $B_h \le 8G$, $c_0 \le 2$, $\alpha = \epsilon/(64G)$, and $\gamma \ge 63/64$.

For one fresh observation at (s,a), let $X = Z + \gamma h(Y)$ and $\sigma^2(s,a) = \operatorname{Var}(X \mid s,a,h)$. Then $\mathbb{E}[X \mid s,a,h] = r(s,a) + \gamma P_{s,a}h$, and X lies in an interval of length at most $1 + B_h$. Moreover,

$$\sigma^2(s,a) \le 2\operatorname{Var}(Z \mid s,a) + 2\gamma^2 \operatorname{Var}_{P_{s,a}}(h) \le \frac{1}{2} + 2\operatorname{Var}_{P_{s,a}}(h).$$

This bound does not require Z and Y to be independent. Writing $\sigma_\pi^2(s) = \sigma^2(s,\pi(s))$, equation 21 consequently gives $\alpha\mathcal{R}_\pi\sigma_\pi^2 \le 69G\mathbf{1}$ for every π.

Recall that $e(s,a) = \widehat{q}(s,a) - r(s,a) - \gamma P_{s,a}h$. Bernstein's and Hoeffding's inequalities, followed by a union bound over the SA pairs, give an event $\mathcal{E}$ of conditional probability at least $1 - \eta/3$ on which

$$\begin{aligned} |e(s,a)| &\le \sqrt{\frac{2\sigma^2(s,a)\ell}{n_f}} + \frac{2(1+B_h)\ell}{3n_f} \qquad \forall (s,a), \\ \|e\|_\infty &\le (1+B_h)\sqrt{\frac{\ell}{2n_f}} \le \tau. \end{aligned} \tag{22}$$

Indeed, each concentration inequality fails with probability at most $2e^{-\ell}$ per pair, so their combined failure probability is at most $4SAe^{-\ell} = \eta/3$. The last inequality follows from the $(1+B_h)^2/\tau^2$ term in the sample size.

Work on $\mathcal{E}$ and fix any π. Let $\sigma_\pi(s) = \sqrt{\sigma_\pi^2(s)}$. Since $\alpha\mathcal{R}_\pi$ is row-stochastic, Cauchy–Schwarz applied to each row yields $\alpha\mathcal{R}_\pi\sigma_\pi \le \sqrt{\alpha\mathcal{R}_\pi\sigma_\pi^2} \le \sqrt{69G}\mathbf{1}$. Applying $\alpha\mathcal{R}_\pi$ to equation 22 therefore gives

$$\alpha\mathcal{R}_\pi|e_\pi| \le \sqrt{\frac{2\ell}{n_f}}\,\alpha\mathcal{R}_\pi\sigma_\pi + \frac{2(1+B_h)\ell}{3n_f}\mathbf{1} \le \left(\sqrt{\frac{138G\ell}{n_f}} + \frac{6G\ell}{n_f}\right)\mathbf{1} \le \frac{\epsilon}{1024}\mathbf{1} = \frac{\alpha G}{16}\mathbf{1}.$$

For the final inequality, use $n_f \ge C\ell G/\epsilon^2$ and $\epsilon \le 1$, choosing C large enough that $\sqrt{138/C} + 6/C \le 1/1024$. Dividing by α proves $\|(I - \gamma P_\pi)^{-1}|e_\pi|\|_\infty \le G/16$.

Taking the supremum over π completes the proof.

D Proof of Lemma 4.3 and Lemma 4.5

Fix $\epsilon, \eta \in (0,1)$ and any trial with $G \ge 1$, $\epsilon < \tau \le 1$, $\alpha = 1 - \gamma = \epsilon/(64G)$, and $R = 4\tau/\alpha$. Condition on a preceding coarse-learning and frozen-estimation history satisfying their guarantees. In particular, $\operatorname{sp}(h) \le 8G$, and the fixed cost table satisfies $c \ge 0$ and $\max_s \min_a c(s,a) \le \alpha R$. Write W^π and $w^\star$ for the discounted policy and optimal costs under kernel L, and set $\zeta = \alpha G/256$. Following a minimum-cost action in each state gives $0 \le w^\star \le R\mathbf{1}$ for every stochastic kernel L. We first prove numerical and statistical bounds without assuming $G \ge H$, and then specialize them to the two cases. Let $\mathbf{1}_s$ denote the one-hot vector with the only one element at the s-th dimension.

Recovery accuracy. Let $v = w^\star_{\overline{P}}$, $\check{C} = 2^{-20}$, $\vartheta = \check{C}\sqrt{\alpha G/R}$, and $d_0 = \check{C}^2\alpha G$. For each query to $\overline{P}_{s,a}$ in the LP process, we use $n_r = \widetilde{O}(S^3R/(\alpha G))$ fresh samples, together with Algorithm 3 to recover $\overline{P}_{s,a}$. The recovery metadata and a temporary probability vector use $\widetilde{O}(SA)$ space. By choosing the constant in n_r large enough and Lemma B.2, it holds that

$$\big|(\overline{P}_{s,a} - P_{s,a})v\big| \le 8\vartheta\sqrt{\operatorname{Var}_{P_{s,a}}(v)} + 4\vartheta^2\operatorname{sp}(v) + d_0, \qquad \forall (s,a). \tag{23}$$

Numerical precision and number of queries. The LP has S variables and $m = SA + 2S$ inequalities, including $-2R\mathbf{1} \le x \le 2R\mathbf{1}$. Every feasible x satisfies $x \le v$ by monotonicity and contraction of the discounted cost Bellman operator, and v is feasible. Thus v is the LP optimum. Write its constraints as $a_j^\top x \le b_j$ and their slacks as $\sigma_j(x) = b_j - a_j^\top x$. The Bellman normals are $\mathbf{1}_s - \gamma \overline{P}_{s,a}^\top$, where $\mathbf{1}_s$ is a coordinate vector; the box normals are $\pm\mathbf{1}_s$. Hence $\|a_j\|_2 \le 2$, and $x_0 = -R\mathbf{1}$ has all slacks at least αR.

Apply the logarithmic-barrier path-following method of Nemirovski & Todd (2008, Theorem 2.1 and Remark A) to $\phi(x) = -\sum_j \log \sigma_j(x)$. Its backward-path initialization applies because the feasible set is bounded and x_0 is strictly feasible. Using a conservative short-step schedule and stopping at barrier coefficient $t \ge 4m/\zeta$, where $\zeta = \alpha G/256$, requires $K = \widetilde{O}(\sqrt{m})$ Newton steps and returns a strictly feasible point with summed objective gap at most $\zeta/2$.

We use inexact, rather than exact, Newton directions. If $M = \nabla^2\phi(x)$ and $g = \nabla\phi(x) - \ell$ for the current linear coefficient ℓ, a sufficient combined direction and update error is $\|\widetilde{d} + M^{-1}g\|_M \le 10^{-3}$, where $\|z\|_M = (z^\top M z)^{1/2}$. The local barrier estimates in Nemirovski & Todd (2008, Section 2.1) justify this tolerance: with pre-step decrement at most $3/16$, the actual step norm is $r \le 3/16 + 10^{-3} < 0.19$, and the new decrement is at most $10^{-3}/(1-r) + r^2/(1-r)^2 < 1/8$. Consequently, strict feasibility and the objective-gap bound are preserved by the inexact implementation.

For clarity, this local tolerance requires only inverse-polynomial absolute precision in the trial parameters. The box gives $M \succeq \mu I$ with $\mu = 1/(2R^2)$. Along the backward and forward paths, $\|\ell\|_1 \le B_{\mathrm{lin}} := C(m/(\alpha R) + mS/\zeta)$. At a point with decrement at most $1/4$, put $z = x - x_0$ and $T_x = \sum_j \sigma_j(x_0)/\sigma_j(x)$. Then $T_x = m + z^\top \nabla\phi(x)$, $\|z\|_M \le T_x + m$, and $\|z\|_\infty \le 3R$, giving $T_x \le 2m + 4RB_{\mathrm{lin}}$. It follows that every slack is at least $L = \alpha R/(2m + 4RB_{\mathrm{lin}})$ and that $M \preceq \Lambda I$ with $\Lambda = 4m/L^2$. Thus coordinate accuracy of order $1/(10^4\sqrt{S\Lambda})$ for the Newton solve and the iterate update suffices, after reserving part of the local error budget for gradient and Hessian approximation.

To implement these solves in small space, round each factor $a_j/\sigma_j(x)$ on one fixed dyadic grid. This yields a fixed dyadic Hessian and gradient satisfying $\|\widetilde{M} - M\|_{\mathrm{op}} \le 10^{-6}\mu$ and $\|\widetilde{g} - g\|_2 \le 10^{-6}\sqrt{\mu}$; the bounds above show that logarithmic-length precision suffices. Whitening these inequalities bounds the resulting direction perturbation by $3 \cdot 10^{-6}$ in the M-norm. After common scaling, the system is an invertible integer system. Invoke Liu et al. (2026, Corollary 10), which gives entrywise e^θ-multiplicative accuracy with $O(S^2 \log(1/\theta))$ matrix–vector queries and linear working space up to logarithmic factors. Choose θ to imply the coordinate tolerance above, using $\|\widetilde{M}^{-1}\widetilde{g}\|_\infty \le 2\|\widetilde{g}\|_2/\mu$. This needs only $\widetilde{O}(S^2)$ products per Newton step; amplifying each solve to failure probability $\eta/(6K)$ adds logarithmic factors. For $S = 1$, solve the scalar system directly.

Each product with the rounded Hessian is a sum of factor products $\sum_j f_j(f_j^\top z)$. One pass through the SA recovered rows regenerates these factors and computes the product using $O(S)$ temporary entries; box factors require no simulator calls. Therefore the total number of row queries is

$$J = \widetilde{O}(\sqrt{SA})\,\widetilde{O}(S^2)\,SA = \widetilde{O}\big(S^2(SA)^{3/2}\big). \tag{24}$$

Gradient passes and the final policy scan are included in this bound. This reduction uses the cited linear-system theorem, not a black-box LP theorem. It stores no Hessian or transition kernel and uses $\widetilde{O}(SA)$ working space overall.

The computed value and policy. Clip the feasible numerical output to its nonnegative part $\widehat{w}$. This preserves Bellman feasibility because $c \ge 0$ and improves the objective. Since every feasible vector is below v, we obtain $0 \le \widehat{w} \le v \le R\mathbf{1}$ and $0 \le v - \widehat{w} \le \zeta\mathbf{1}$. A final row scan chooses $\pi(s) \in \arg\min_a\{c(s,a) + \gamma\overline{P}_{s,a}\widehat{w}\}$, with fixed tie-breaking. Bellman monotonicity gives

$$0 \le c_\pi + \gamma\overline{P}_\pi\widehat{w} - \widehat{w} \le \zeta\mathbf{1}, \tag{25}$$

$$0 \le c_\pi + \gamma\overline{P}_\pi v - v \le 2\zeta\mathbf{1}. \tag{26}$$

The second bound follows by adding $\gamma\overline{P}_\pi(v - \widehat{w}) - (v - \widehat{w})$. All feasibility statements here concern $\overline{P}$, not P.

Transfer to the true kernel. For a stationary deterministic policy σ, write $c_\sigma(s) = c(s, \sigma(s))$ and define $s_\sigma = c_\sigma + \gamma \overline{P}_\sigma v - v \geq 0$, $E_\sigma = \gamma(P_\sigma - \overline{P}_\sigma)v$, and $\mathcal{R}_\sigma = (I - \gamma P_\sigma)^{-1}$. Write $B = \mathrm{sp}(v)$, $M_\sigma = \|W_P^\sigma\|_\infty$, and $T_\sigma = \||\mathcal{R}_\sigma| E_\sigma|\|_\infty$. The exact policy equation retains the model error: $W_P^\sigma - v = \mathcal{R}_\sigma(s_\sigma + E_\sigma)$.

To control T_σ, let $m_v = \min_s v(s)$, $x = v - m_v \mathbf{1}$, and $d = x - \gamma P_\sigma x$. Then $0 \leq x \leq B\mathbf{1}$ and $d = c_\sigma - s_\sigma - E_\sigma - \alpha m_v \mathbf{1}$, so $xd \leq B(c_\sigma + |E_\sigma|)$ coordinatewise. Expanding $2xd - d^2 = x^2 - \gamma^2 (P_\sigma x)^2$ gives

$$\gamma \mathcal{R}_\sigma \operatorname{Var}_{P_\sigma}(v) = -x^2 + 2\mathcal{R}_\sigma(xd) - \mathcal{R}_\sigma(d^2) - \alpha\gamma \mathcal{R}_\sigma[(P_\sigma x)^2].$$

Here products, squares, and the variance vector are coordinatewise. Dropping nonpositive terms and applying rowwise Cauchy–Schwarz, using $\mathcal{R}_\sigma \mathbf{1} = \alpha^{-1}\mathbf{1}$, yields

$$\|\gamma \mathcal{R}_\sigma \operatorname{Var}_{P_\sigma}(v)\|_\infty \leq 2B(M_\sigma + T_\sigma),$$
$$\|\gamma \mathcal{R}_\sigma \sqrt{\operatorname{Var}_{P_\sigma}(v)}\|_\infty \leq \sqrt{2\gamma B(M_\sigma + T_\sigma)/\alpha}.$$

Consequently, equation 23 implies $T_\sigma \leq 8\vartheta\sqrt{2\gamma B(M_\sigma + T_\sigma)/\alpha} + 4\gamma\vartheta^2 B/\alpha + \gamma d_0/\alpha$.

Use this inequality for a true optimal policy π_P, a recovered-kernel optimal policy $\pi_{\overline{P}}$, and the returned π. For π_P, $M_{\pi_P} \leq R$. For the other two, the exact policy equation and equation 26 give $M_\sigma \leq R + T_\sigma + 2\zeta/\alpha$. Since $2\zeta/\alpha \leq R$, all three satisfy $M_\sigma + T_\sigma \leq 2R + 2T_\sigma$. Set $\chi = \vartheta^2 B/\alpha$; then

$$\begin{aligned} T_\sigma &\leq 16\sqrt{\chi R} + 16\sqrt{\chi T_\sigma} + 4\chi + d_0/\alpha \\ &\leq 16\sqrt{\chi R} + T_\sigma/2 + 132\chi + d_0/\alpha. \end{aligned}$$

Rearranging and substituting ϑ, d_0 gives $T_\sigma \leq F(B) := 32\check{C}\sqrt{GB} + 264\check{C}^2 GB/R + 2\check{C}^2 G$.

For π_P, nonnegative recovered slack implies $v \leq w^\star + T_{\pi_P}\mathbf{1}$. For $\pi_{\overline{P}}$, zero recovered slack implies $w^\star \leq W_P^{\pi_{\overline{P}}} \leq v + T_{\pi_{\overline{P}}}\mathbf{1}$. For the returned policy, its recovered slack is at most $2\zeta\mathbf{1}$. Hence

$$\|v - w^\star\|_\infty \leq F(B), \qquad \|W_P^\pi - v\|_\infty \leq F(B) + 2\zeta/\alpha. \tag{27}$$

Together with $0 \leq v - \widehat{w} \leq \zeta\mathbf{1}$, these yield

$$\begin{aligned} \|\widehat{w} - w^\star\|_\infty &\leq F(B) + \zeta, \\ 0 \leq W_P^\pi - w^\star &\leq \big(2F(B) + 2\zeta/\alpha\big)\mathbf{1}. \end{aligned} \tag{28}$$

It remains only to show $F(B) < G/128$ in each case.

Case $G \geq H$. The preceding shaping comparison and $\mathrm{sp}(V_\gamma^\star) \leq 2H$ imply $\mathrm{sp}(w^\star) \leq 8G + 2H + G/8 < 11G$. Here $V_\gamma^\star$ denotes the original optimal discounted reward value under P. Equation 27 therefore gives $B \leq 11G + 2F(B)$. Writing $z = B/G$ and using $R \geq 256G$,

$$\begin{aligned} z &\leq 11 + 64\check{C}\sqrt{z} + 528\check{C}^2(G/R)z + 4\check{C}^2 \\ &\leq 11 + z/2 + 4100\check{C}^2. \end{aligned}$$

The second step uses $64\check{C}\sqrt{z} \leq z/4 + 4096\check{C}^2$ and $528\check{C}^2 G/R \leq 1/4$. Thus $z \leq 22 + 8200\check{C}^2 < 24$, so $F(B)/G \leq 160\check{C} + 27\check{C}^2 < 1/128$. This conclusion is established before applying the stopping test.

Case of early acceptance. Now suppose $1 \leq G < H$ and the returned pair passes $\alpha\,\mathrm{sp}(h - \widehat{w}) \leq \epsilon/2$. Instead of a bound involving H, use the observable test:

$$\begin{aligned} \mathrm{sp}(\widehat{w}) &\leq \mathrm{sp}(h) + \mathrm{sp}(h - \widehat{w}) \leq 8G + \frac{\epsilon}{2\alpha} = 40G, \\ B &\leq \mathrm{sp}(\widehat{w}) + 2\|v - \widehat{w}\|_\infty \leq 40G + 2\zeta < 41G. \end{aligned}$$

Consequently, $F(B)/G \leq 224\check{C} + 45\check{C}^2 < 1/128$. More generally, this argument applies to any returned pair with $\mathrm{sp}(\widehat{w}) \leq 40G$, regardless of whether $G < H$.

In either case, substitution into equation 28 gives

$$\|\widehat{w} - w^\star\|_\infty \leq G/128 + \zeta, \qquad 0 \leq W_P^\pi - w^\star \leq \big(G/64 + 2\zeta/\alpha\big)\mathbf{1}. \tag{29}$$

Thus the large-scale and early-acceptance guarantees follow from the same numerical and statistical comparison, with only the final span bound differing.

Probability, simulator samples, and space. The numerical guarantee holds for every fixed canonical kernel. Analyze this ideal fixed-row execution first and intersect its success event with the finite-query recovery event. On the intersection, the actual and ideal executions coincide, including their feasibility checks and policy extraction. The total failure probability is at most $\eta/6+\eta/6 = \eta/3$. This event is constructed before imposing acceptance. Hence for $G \geq H$ the call succeeds and satisfies equation 29 with probability at least $1 - \eta/3$; for early stopping, the probability of returning an accepted pair that violates these bounds is at most $\eta/3$.

The prescribed numerical and sampling caps depend only on the current trial parameters, not on H or on whether the trial is accepted. Including calibration and policy extraction, the sample count in either case is

$$N_{\mathrm{CorrectLP}} \leq (SA + J)n_r = \widetilde{O}\left(\frac{S^5(SA)^{3/2}R}{\alpha G}\right).$$

The stored inputs, recovery metadata, and streamed numerical workspaces use $\widetilde{O}(SA)$ working space.

E NUMERICAL EXPERIMENTS: PER-INSTANCE AND BUDGET-RESOLVED DETAIL

This appendix gives the same runs as in Section 5 at finer resolution: Figure 1 per instance and arm at that budget, and Figure 2 as the budget varies.

The suite. The 18 instances are H-controlled and come from four families. The ring ladder sweeps the decision margin over four orders of magnitude (drift δ from 10^{-4} to 0.5), hardest first; the escape traps occur at three rates; and the padding and the action duplication add states that no policy can improve and actions that duplicate one another, which are the structures a row-reconstruction planner is most exposed to. Ground truth—the optimal bias and the optimal gain—is computed exactly on every instance, so the consistency and loss of Figure 1 are measured against the true optimum rather than against a longer run of one of the arms.

On the band where the arms actually separate, RAD leads by 0.225, in every seed. Its mean score moves by 0.093 across a $64\times$ range of budgets, against 0.36 for the horizon-aware arms, so the advantage is widest where samples bind; to match the score RAD already reaches at $n = 10^3$ the other arms need sixteen times the samples.

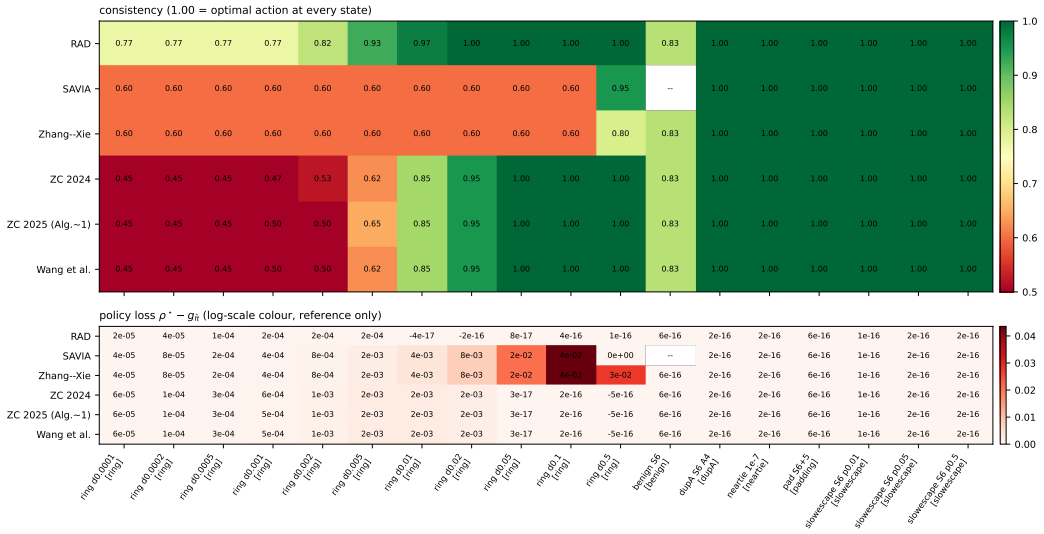


Figure 1: Per-instance outcome at $n = 4000$ samples per state–action pair, every arm, all 18 instances. Ring instances are ordered left to right by decreasing decision margin (drift δ from 10^{-4} to 0.5), hardest first; the remaining families follow. Upper panel: consistency. Lower panel: true loss $\rho^\star - g_{\hat{\pi}}$, logarithmic colour scale. Cells average four runs per instance (the single-seed budget grid and a three-seed rerun under the same configuration). RAD's row is the only one whose cells never fall below 0.77 and the only one whose loss never exceeds 2×10^{-4}; on easy instances every arm is optimal and every loss sits at the floating-point floor.

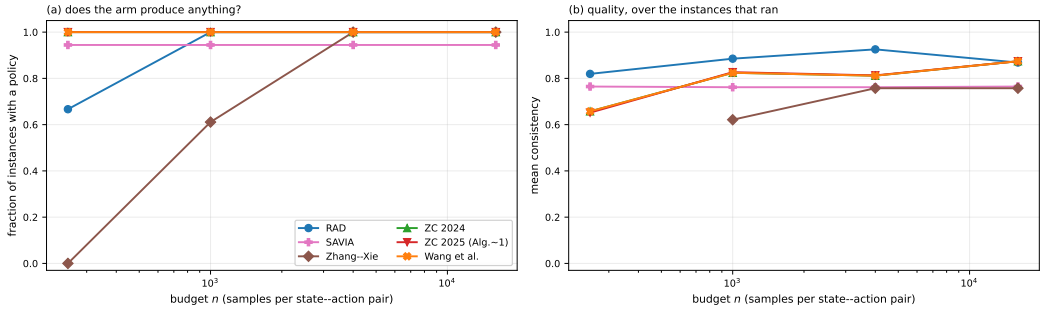


Figure 2: The same runs as a function of the budget. (a) Fraction of the 18 instances on which an arm returned a policy at all. (b) Mean consistency over the instances that ran, which are not the same set for every arm: RAD covers 12 of 18 at $n = 250$ and Zhang–Xie covers none. Zhang–Xie is absent from (b) at $n = 250$ because it returns nothing. The budget grid is single-seed except at $n = 1000$ and $n = 4000$, where three seeds were run; (b) averages over the instances each arm actually returned on.