WHY BACKWARD INDUCTION DOES NOT WORK IN SAMPLE-BASED REINFORCEMENT LEARNING?

Anonymous authors
Paper under double-blind review

ABSTRACT

It is well known that, in exact finite-horizon dynamic programming, backward induction finds an optimal policy in a single backward sweep by selecting actions that maximize the action-value function. In sample-based reinforcement learning, however, the action-value function is unknown and must be estimated from data. Natural Policy Gradient (NPG) is widely used in this setting, updating the policy gradually according to estimated action values rather than directly selecting their maximizer. This raises a basic question: if hard maximization is sufficient under exact value information, why not use the same hard-max update in the sample-based setting? Equivalently, what explains the favorable behavior of NPG under noisy value estimates? We study this question in finite-horizon Markov decision processes by comparing sample-based hard-max updates with NPG under a common sampling model. We show that the two methods respond differently to estimation error: hard maximization is more sensitive to noisy value estimates, whereas NPG updates the policy more gradually. Our results clarify why the advantage of exact backward induction does not necessarily carry over to the sample-based setting.

1 INTRODUCTION

Markov decision processes (MDPs) provide a standard framework for sequential decision making under uncertainty (Puterman, 1994). Policy iteration is a classical solution method that alternates policy evaluation and greedy policy improvement. In a finite-horizon MDP, applying greedy updates from horizon H back to horizon 1, with each update evaluated under the already-updated continuation decision rules, gives backward induction. When the transition kernels and rewards are known, a single backward sweep returns an optimal policy (see Remark 1).

In sample-based reinforcement learning (RL), transition kernels are unknown and action values are estimated from observations (Kakade, 2003). For tabular softmax policies, natural policy gradient (NPG) multiplicatively reweights action probabilities using exponentiated action-value estimates (Kakade, 2001; Agarwal et al., 2021; Xiao, 2022). With fixed estimates and a full-support policy, the update concentrates on maximizing actions as the step size tends to infinity (Section 2.2). To investigate updates with finite step sizes under sampling noise, we compare NPG with one-step evaluation and backward induction with fresh Monte Carlo evaluation. The procedures differ in both evaluation and update but share a generative oracle permitting adaptive queries. We compare their transition-query requirements under the expected optimality-gap criterion at an initial distribution ρ.

Here, S and A denote the numbers of states and actions, respectively, and H is the horizon length. Therefore, our main contributions are as follows:

- Theorem 1 proves that any agent guaranteeing an expected optimality gap of at most ε uniformly over a constructed MDP family requires $\Omega(SAH^4/\varepsilon^2)$ transition queries. This includes agents that choose queries adaptively.
- Theorem 2 and Corollary 1 establish a sufficient transition count of $O(SAH^5 \log A/\varepsilon^2)$ for every MDP in the stated setting. The procedure uses one successor sample per state-action-horizon triple with $h < H$ at each iteration, a step size that is constant across iterations, and a uniformly selected output iterate.

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.

- Theorem 3 shows that one backward sweep using fresh complete Monte Carlo rollouts requires $\Omega(SAH^6/\varepsilon^2)$ transition queries to guarantee the same expected accuracy uniformly over this family. The bound concerns the specified rollout-based implementation.

The following summarizes the findings of this paper:

$$\underbrace{\frac{H^4}{\varepsilon^2}}_{\substack{\text{information theoretic}\\ \text{lower bound}\\ \text{Theorem 1}}} < \underbrace{\frac{H^5}{\varepsilon^2}}_{\substack{\text{NPG with one-step evaluation}\\ \text{upper bound}\\ \text{Theorem 2}}} < \underbrace{\frac{H^6}{\varepsilon^2}}_{\substack{\text{Backward induction with}\\ \text{fresh Monte Carlo evaluation}\\ \text{lower bound}\\ \text{Theorem 3}}} .$$

1.1 Related work

Prior sample-complexity results for finite-horizon MDPs consider both online episodic interaction and generative access. Table 1 summarizes representative guarantees for returning an ε-optimal policy and distinguishes their performance criteria.

Algorithm	Setting	Sample complexity
Information theoretic bound (Domingues et al., 2021)	Online; (ε, δ)-PAC BPI	$\Omega\left(\frac{SAH^4}{\varepsilon^2} \log \frac{1}{\delta}\right)$
BPI-UCBVI (Ménard et al., 2021)	Online; (ε, δ)-PAC BPI	$\widetilde{O}\left(\frac{SAH^4}{\varepsilon^2}\right)$
Modified MVP (Zhang et al., 2024)	Online; high probability	$\widetilde{O}\left(\frac{SAH^4}{\varepsilon^2}\right)$
Variance-reduced QVI (Sidford et al., 2018)	Generative; high probability	$\widetilde{O}\left(\frac{SAH^4}{\varepsilon^2}\right)$
Model-based planning (Yin et al., 2021)	Generative; high probability	$\widetilde{O}\left(\frac{SAH^4}{\varepsilon^2}\right)$
Model-based planning (Li et al., 2024)	Generative; high probability	$\widetilde{O}\left(\frac{SAH^4}{\varepsilon^2}\right)$
Information theoretic bound (Theorem 1)	Generative; expected gap	$\Omega\left(\frac{SAH^4}{\varepsilon^2}\right)$
NPG, one-step evaluation (Corollary 1)	Generative; expected gap	$O\left(\frac{SAH^5}{\varepsilon^2}\right)$
Backward induction, fresh Monte Carlo evaluation (Theorem 3)	Generative; expected gap	$\Omega\left(\frac{SAH^6}{\varepsilon^2}\right)$

Table 1: Sample complexity measured in transitions for finite-horizon tabular MDPs with horizon-dependent kernels and rewards in $[0, 1]$. PAC BPI denotes (ε, δ)-PAC best-policy identification. The notation $\widetilde{O}$ suppresses logarithmic factors, including confidence dependence. The Domingues et al. lower bound concerns the expected sample count under a high-probability policy-correctness requirement; our results concern the expected optimality gap at the initial distribution ρ. Prior-result assumptions and conversions are detailed in Appendix 9.1. Our rows hold under the conditions of the corresponding results.

The information-theoretic bounds under online episodic interaction and generative access have the same polynomial dependence on S, A, H, and ε, under their respective performance criteria. After converting its high-probability guarantee to the expected-gap criterion, empirical-model planning (Li et al., 2024) matches our information-theoretic lower bound up to logarithmic factors. This planner uses backward induction on empirical kernels, whereas our backward-induction lower bound concerns hard-max updates with fresh Monte Carlo evaluation. Our results therefore show that, in the sample-based setting considered here, gradual NPG updates with one-step evaluation can require asymptotically fewer transition samples than hard-max backward induction based on fresh Monte Carlo estimates. Related PMD or NPG analyses in discounted MDPs provide high-probability last-iterate guarantees (Johnson et al., 2023; Liu et al., 2025) or an expected-gap guarantee for a uniformly selected iterate (Lan, 2023); detailed differences in setting, evaluation, and performance criteria are discussed in Appendix 9.1.

2 PRELIMINARIES

2.1 FINITE-HORIZON MARKOV DECISION PROCESS

We consider a finite-horizon Markov decision process (MDP) represented by the tuple $(\mathcal{S}, \mathcal{A}, H, \mathcal{P}, \mathcal{R})$, where $\mathcal{S}$ and $\mathcal{A}$ are finite state and action sets, respectively, and H is the horizon length. We write $[H] = \{1, \dots, H\}$, $S = |\mathcal{S}|$, $A = |\mathcal{A}|$, and for a finite set $\mathcal{X}$, $\Delta(\mathcal{X})$ is the probability simplex over $\mathcal{X}$. The transition kernels are $\mathcal{P} = \big\{P^h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S}) \mid h \in [H-1]\big\}$, where $P^h(s' \mid s, a)$ is the probability of transitioning to state s' from state-action pair (s, a) at horizon h. The reward functions are $\mathcal{R} = \big\{R^h : \mathcal{S} \times \mathcal{A} \to [0,1] \mid h \in [H]\big\}$.

We represent a Markov policy as $\pi = (\pi^1, \dots, \pi^H)$, where π^h is the decision rule at horizon h and $\pi^h(\cdot \mid s) \in \Delta(\mathcal{A})$ is its action distribution in state s. Given $h \in [H]$ and $S^h = s$, a policy π generates the trajectory $(S^h, A^h, \dots, S^H, A^H)$ with $A^i \sim \pi^i(\cdot \mid S^i)$ for $i \in \{h, \dots, H\}$ and $S^{i+1} \sim P^i(\cdot \mid S^i, A^i)$ for $i \in \{h, \dots, H-1\}$. For $h \in [H]$ and $s \in \mathcal{S}$, the value function is

$$V^{\pi,h}(s) := \mathbb{E}\left[\sum_{i=h}^{H} R^i(S^i, A^i) \,\middle|\, S^h = s\right],$$

where the expectation is over the trajectory generated by π. Since $R^i(s,a) \in [0,1]$, we have $0 \le V^{\pi,h}(s) \le H - h + 1$. The corresponding action-value function is

$$Q^{\pi,h}(s,a) := R^h(s,a) + \sum_{s' \in \mathcal{S}} P^h(s' \mid s, a)\, V^{\pi,h+1}(s'), \qquad h \in [H-1],$$

and $Q^{\pi,H}(s,a) := R^H(s,a)$. For every $h \in [H]$ and $s \in \mathcal{S}$, $V^{\pi,h}(s) = \sum_{a \in \mathcal{A}} \pi^h(a \mid s)\, Q^{\pi,h}(s,a)$. The advantage function is $A^{\pi,h}(s,a) := Q^{\pi,h}(s,a) - V^{\pi,h}(s)$. The goal is to find an optimal policy $\pi^\star$ satisfying $V^{\pi^\star,h}(s) \ge V^{\pi,h}(s),\ \forall \pi,\ \forall h \in [H],\ \forall s \in \mathcal{S}$.

For $h \in [H]$, $i \in \{h, \dots, H\}$, and $s, \widetilde{s} \in \mathcal{S}$, the state visitation distribution induced by a policy π is

$$d_s^{h \to i, \pi}(\widetilde{s}) := \mathbb{P}\Big(S^i = \widetilde{s} \mid S^h = s,\ A^j \sim \pi^j(\cdot \mid S^j),\ S^{j+1} \sim P^j(\cdot \mid S^j, A^j),\ j = h, \dots, i-1\Big).$$

In particular, $d_s^{h \to h, \pi}(\widetilde{s}) = \mathbf{1}\{\widetilde{s} = s\}$. For a distribution $\rho \in \Delta(\mathcal{S})$, we write $V^{\pi,h}(\rho) := \sum_{s \in \mathcal{S}} \rho(s) V^{\pi,h}(s)$. Throughout the paper, a superscript $\star$ denotes a quantity associated with $\pi^\star$.

2.2 POLICY UPDATE RULES

We consider two policy update rules: backward induction and natural policy gradient (NPG), first with exact action values and then with estimates in Section 4. For $t \ge 0$, let $\pi_t = (\pi_t^1, \dots, \pi_t^H)$ denote the policy after t iterations; a backward induction iteration is one complete backward sweep of hard-max updates. Both rules start from the uniform policy $\pi_0^h(a \mid s) = 1/A$ for every $(h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}$.

Backward induction update: The action-value function $Q^{\pi,h}$ depends on π only through the continuation decision rules $\pi^{h+1}, \dots, \pi^H$; by definition, $Q^{\pi,H} = R^H$ does not depend on π. Starting from π_t, we update horizons in the order $H, H-1, \dots, 1$. Before updating horizon h, the intermediate policy is $\pi_{t,h} := (\pi_t^1, \dots, \pi_t^h, \pi_{t+1}^{h+1}, \dots, \pi_{t+1}^H)$, whose continuation decision rules have already been updated during the current sweep. For every $s \in \mathcal{S}$, we choose

$$a_{t+1}^h(s) \in \arg\max_{a \in \mathcal{A}} Q^{\pi_{t,h},h}(s,a), \tag{1}$$

and set $\pi_{t+1}^h(a \mid s) = \mathbf{1}\{a = a_{t+1}^h(s)\}$. We fix an ordering of $\mathcal{A}$, used at every state, horizon, and iteration; if several actions attain the maximum in equation 1, we choose the first maximizing action in this ordering.

NPG update: We use the softmax NPG update (Kakade, 2001); see Agarwal et al. (2021, Lemma 15) for its multiplicative form in the discounted setting. With a constant step size $\eta > 0$, the update at iteration t for each $h \in [H]$, $s \in \mathcal{S}$, and $a \in \mathcal{A}$ is

$$\pi_{t+1}^h(a \mid s) = \frac{\pi_t^h(a \mid s) \exp\big(\eta Q^{\pi_t,h}(s,a)\big)}{Z_t^h(s)}, \tag{2}$$

where the normalization constant is $Z_t^h(s) := \sum_{a' \in \mathcal{A}} \pi_t^h(a' \mid s) \exp\left(\eta Q^{\pi_t,h}(s,a')\right)$. For every finite $\eta > 0$, the NPG update depends smoothly on the action values and preserves the support of the current policy; as $\eta \to \infty$ with the action values held fixed, $\pi_{t+1}^h(\cdot \mid s)$ converges to $\pi_t^h(\cdot \mid s)$ restricted to the maximizing actions and renormalized. NPG evaluates all horizons under π_t and updates them jointly, whereas backward induction evaluates horizon h under $\pi_{t,h}$.

3 Exact backward induction

We first consider the setting in which the transition kernels and reward functions are known, so that the action-value functions can be evaluated exactly. In this scenario, the backward-induction update of Section 2.2 has the following optimality property.

Remark 1. *For any initial policy π_0, one complete backward sweep produces an optimal policy π_1; see Puterman (1994, Theorem 4.5.1). The claim follows by backward induction on h, using the Bellman optimality equation $V^{\star,h}(s) = \max_{a \in \mathcal{A}} Q^{\star,h}(s,a)$. At horizon H, $Q^{\pi_0,H} = R^H = Q^{\star,H}$, so maximizing it gives $V^{\pi_1,H} = V^{\star,H}$. If $V^{\pi_1,h+1} = V^{\star,h+1}$, then $Q^{\pi_{0,h},h} = Q^{\star,h}$, since $Q^{\pi_{0,h},h}$ depends on $\pi_{0,h}$ only through $\pi_1^{h+1}, \dots, \pi_1^H$; maximizing it gives $V^{\pi_1,h} = V^{\star,h}$.*

When action values are estimated from samples, an action maximizing an estimate need not maximize the corresponding exact action value. Section 4 specifies the sampling model, and Section 6 quantifies the consequence for backward induction with fresh Monte Carlo evaluation.

4 Sample-based setting and MDP construction

We consider unknown horizon-dependent transition kernels and known reward functions. We specify the sampling model and performance criterion, then construct the family used for both lower bounds and state the information-theoretic bound.

4.1 Generative access and performance criterion

The agent accesses the transition kernels through a generative oracle (Kakade, 2003, Definition 2.4.1). It can query any triple $(h, s, a) \in [H-1] \times \mathcal{S} \times \mathcal{A}$. Each queried triple is a function of the preceding interaction history and the agent's internal randomness. Conditional on the query and this information, the oracle returns $S' \sim P^h(\cdot \mid s, a)$. Each response counts as one transition query, and the agent makes at most N queries almost surely. No query is made at horizon H, where $Q^{\pi,H} = R^H$ is known.

For an initial distribution $\rho \in \Delta(\mathcal{S})$ and a possibly random output policy π_{out}, the expected optimality gap on an instance M is

$$\mathbb{E}_M\left[V^{\star,1}(\rho) - V^{\pi_{\mathrm{out}},1}(\rho)\right], \tag{3}$$

where the value functions are evaluated in M and the expectation is over the oracle observations and the agent's internal randomness. For a prescribed accuracy $\varepsilon > 0$, the target is to bound equation 3 by ε on every instance in the class under consideration. NPG returns $\pi_{\mathrm{out}} = \pi_{\widehat{T}}$, where, for a fixed integer $T \geq 1$, $\widehat{T} \sim \mathrm{Unif}\{0, \dots, T-1\}$ is independent of the randomness generating the policy iterates. Backward induction returns $\pi_{\mathrm{out}} = \pi_T$ after T backward sweeps.

4.2 A common family of finite-horizon MDPs

The following family adapts the lower-bound construction in Kakade (2003, proof of Theorem 2.5.2). Uniform mixing among decision states allows selection errors across states to contribute to the optimality gap under a single initial distribution.

Fix integers $K \geq 1$, $A \geq 2$, and $H \geq 2$. Let $\mathcal{S} = \{s_1, \dots, s_K, g\}$ and $\mathcal{A} = \{a_1, \dots, a_A\}$, where $s_1, \dots, s_K$ are decision states and g is an absorbing reward state. Thus, $S = K + 1$. The initial distribution is $\rho(s_i) = 1/K$ for $i \in [K]$ and $\rho(g) = 0$. For every $h \in [H]$, $i \in [K]$, and $a \in \mathcal{A}$, set $R^h(s_i, a) = 0$ and $R^h(g, a) = 1$. For $p \in (0,1)$, define the distribution B_p on $\mathcal{S}$ by

$$B_p(g) = p, \qquad B_p(s_i) = \frac{1-p}{K}, \quad i \in [K]. \tag{4}$$

For each $h \in [H-1]$, choose parameters p^h and α^h satisfying

$$0 < \alpha^h < p^h, \qquad p^h + \alpha^h \le \frac{1}{2H}. \tag{5}$$

Let Θ be the set of label arrays $\theta = (\theta_i^h)_{h\in[H-1],\, i\in[K]}$ with $\theta_i^h \in \{0, 1, \ldots, A\}$. The label $\theta_i^h = j \in [A]$ identifies a_j as the favorable action at (h, s_i); $\theta_i^h = 0$ indicates that all actions have the same probability of entering g. For $\theta \in \Theta$, $h \in [H-1]$, $i \in [K]$, and $j \in [A]$, define

$$p^h_{\theta,i,a_j} := p^h - \alpha^h + 2\alpha^h \mathbf{1}\{\theta_i^h = j\}, \tag{6}$$

and set

$$P_\theta^h(\cdot \mid s_i, a_j) = B_{p^h_{\theta,i,a_j}}. \tag{7}$$

At g, set $P_\theta^h(g \mid g, a) = 1$ for every $h \in [H-1]$ and $a \in \mathcal{A}$. For fixed parameters $(p^h, \alpha^h)_{h\in[H-1]}$, let M_θ denote the MDP indexed by θ. The family $\mathcal{M}_{K,A,H}$ is the set of all such MDPs as the labels and parameters vary subject to equation 5. The lower bounds remain valid when the agent knows the construction and parameters, with only θ unknown.

A favorable action, when present, uniquely maximizes the action-value function under every continuation policy. The optimality gap is bounded below by a weighted sum of the probabilities assigned to suboptimal actions. These properties are established in Lemmas 1 and 2, respectively, in Appendix 10.

4.3 Information theoretic lower bound

The following theorem gives a lower bound on the transition query budget N of any agent that guarantees an expected optimality gap of at most ε on every instance in $\mathcal{M}_{K,A,H}$.

Theorem 1. *There exists a universal constant $c > 0$ such that the following holds for all integers $K \ge 1$, $A \ge 2$, $H \ge 2$, and $N \ge 1$. For every agent using at most N transition queries under the sampling model of Section 4.1 and returning a policy π_{out}, there exists an instance $M \in \mathcal{M}_{K,A,H}$ such that*

$$\mathbb{E}_M\big[V^{\star,1}(\rho) - V^{\pi_{\mathrm{out}},1}(\rho)\big] \ge c \min\left\{H,\, H^2\sqrt{\frac{KA}{N}}\right\}. \tag{8}$$

Consequently, for any $0 < \varepsilon < cH$, an agent satisfying

$$\sup_{M\in\mathcal{M}_{K,A,H}} \mathbb{E}_M\big[V^{\star,1}(\rho) - V^{\pi_{\mathrm{out}},1}(\rho)\big] \le \varepsilon \tag{9}$$

must satisfy $N \ge c^2 KAH^4/\varepsilon^2$.

Since $K = S - 1 \ge S/2$, the theorem yields a necessary query budget of $\Omega(SAH^4/\varepsilon^2)$ for $0 < \varepsilon < cH$. The proof appears in Appendix 11.

5 NPG with one-step evaluation

We establish an upper bound for NPG under the sampling model of Section 4.1, for every finite-horizon MDP satisfying Section 2.1.

5.1 Policy evaluation and the NPG update

Fix integers $T, m \ge 1$ and a constant step size $\eta > 0$, where m is the number of successor samples drawn at each triple (h, s, a) with $h \in [H-1]$ in one iteration, and write $L_h := H - h + 1$ for $h \in [H]$. Let $\mathcal{F}_t$ be the sigma-field representing the information available before sampling at iteration t, so π_t is $\mathcal{F}_t$-measurable. For every $h \in [H-1]$, $s \in \mathcal{S}$, $a \in \mathcal{A}$, and $\ell \in [m]$, the procedure queries the oracle at (h, s, a) and obtains $S_{t,\ell}^{h+1}(s, a) \sim P^h(\cdot \mid s, a)$. Conditional on $\mathcal{F}_t$, these samples are mutually independent across horizons, states, actions, and sample indices. With π_t held fixed, set $\widehat{Q}^{\pi_t,H}(s, a) := R^H(s, a)$ and evaluate the policy in backward order using

$$\widehat{Q}^{\pi_t,h}(s, a) := R^h(s, a) + \frac{1}{m}\sum_{\ell=1}^{m} \widehat{V}^{\pi_t,h+1}\big(S_{t,\ell}^{h+1}(s, a)\big), \qquad h \in [H-1], \tag{10}$$

Algorithm 1 NPG with one-step evaluation

Require: Generative access, known rewards, integers $T, m \geq 1$, and constant step size $\eta > 0$.
Initialize $\pi_0^h(a \mid s) = 1/A$ for every (h, s, a).
for $t = 0, \ldots, T-1$ **do**
  Set $\widehat{Q}^{\pi_t,H} = R^H$ and compute $\widehat{V}^{\pi_t,H}$ by equation 11.
  for $h = H-1, \ldots, 1$ **do**
    For each (s, a), draw fresh samples $S_{t,\ell}^{h+1}(s, a) \sim P^h(\cdot \mid s, a)$, $\ell \in [m]$, and compute $\widehat{Q}^{\pi_t,h}(s, a)$ by equation 10.
    Compute $\widehat{V}^{\pi_t,h}(s)$ for every s by equation 11.
  end for
  Update $\pi_{t+1}^h(\cdot \mid s)$ for every (h, s) by equation 12.
end for
Draw $\widehat{T} \sim \mathrm{Unif}\{0, \ldots, T-1\}$, independently of the randomness generating the iterates.
return $\pi_{\widehat{T}}$.

$$\widehat{V}^{\pi_t,h}(s) := \sum_{a \in \mathcal{A}} \pi_t^h(a \mid s)\, \widehat{Q}^{\pi_t,h}(s, a), \qquad h \in [H]. \tag{11}$$

After evaluation, update all decision rules by

$$\pi_{t+1}^h(a \mid s) = \frac{\pi_t^h(a \mid s) \exp\Big(\eta \widehat{Q}^{\pi_t,h}(s, a)\Big)}{\sum_{a' \in \mathcal{A}} \pi_t^h(a' \mid s) \exp\Big(\eta \widehat{Q}^{\pi_t,h}(s, a')\Big)}, \tag{12}$$

for every $(h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}$. This is equation 2 with estimated action values. The procedure starts from the uniform policy π_0 of Section 2.2 and returns $\pi_{\widehat{T}}$ with $\widehat{T}$ as defined in Section 4.1. Algorithm 1 summarizes the procedure. Each iteration uses $mSA(H-1)$ transition queries, giving a total query count of $N = mSA(H-1)T$.

5.2 Expected optimality gap and sample requirement

Theorem 2. *For every MDP M satisfying Section 2.1, integers $T, m \geq 1$, constant step size $\eta > 0$, and initial distribution $\rho \in \Delta(\mathcal{S})$, the policy $\pi_{\widehat{T}}$ returned by Algorithm 1 satisfies*

$$\mathbb{E}_M\big[V^{\star,1}(\rho) - V^{\pi_{\widehat{T}},1}(\rho)\big] \leq \frac{H \log A}{\eta T} + \frac{3\eta H(H+1)(2H+1)}{16}. \tag{13}$$

The proof bounds the estimation error using the value range $[0, L_h]$, without exploiting its dependence on m. The stated guarantee therefore does not improve with m. The following corollary takes $m = 1$, which minimizes the query count for this guarantee.

Corollary 1. *Let $H \geq 2$ and $A \geq 2$. For an integer $T \geq 1$, run Algorithm 1 with $m = 1$ and constant step size*

$$\eta = \sqrt{\frac{16 \log A}{3(H+1)(2H+1)T}}. \tag{14}$$

Then, for any $0 < \varepsilon < H$, choosing $T = \big\lceil 3H^2(H+1)(2H+1)\log A/(4\varepsilon^2) \big\rceil$ guarantees, by equation 13, an expected optimality gap of at most ε using $N = SA(H-1)T = O\big(SAH^5 \log A/\varepsilon^2\big)$ transition queries (see Appendix 12).

6 A lower bound for backward induction with fresh Monte Carlo evaluation

We analyze one backward sweep ($T = 1$ in Section 4.1) on the family of Section 4.2 and derive lower bounds on its expected optimality gap and transition-query requirement.

6.1 Backward induction with fresh Monte Carlo evaluation

The procedure replaces the exact action values $Q^{\pi_{0,h},h}$ in the backward induction update of Section 2.2 with Monte Carlo estimates. Fix an integer $m \geq 1$, the number of fresh rollouts generated at each triple (h, s_i, a) with $h \in [H-1]$, $i \in [K]$, and $a \in \mathcal{A}$.

Let $\mathcal{F}_h$ be the sigma-field representing the information available before collecting rollouts at horizon h. The continuation rules $\pi_1^{h+1}, \ldots, \pi_1^H$ are $\mathcal{F}_h$-measurable and remain fixed during evaluation. For each triple, rollout $\ell \in [m]$ has trajectory $(S_\ell^u, A_\ell^u)_{u=h}^H$, starting from $S_\ell^h = s_i$ with $A_\ell^h = a$ and following π_1^u at horizons $u = h+1, \ldots, H$. Transitions are sampled through the generative oracle, and the return is $G_\ell^h(s_i, a) := \sum_{u=h}^H R^u(S_\ell^u, A_\ell^u)$. Conditional on $\mathcal{F}_h$, all rollouts collected at horizon h are independent. Then, we define

$$\widehat{Q}^{\pi_{0,h},h}(s_i, a) := \frac{1}{m} \sum_{\ell=1}^m G_\ell^h(s_i, a). \tag{15}$$

Each estimate uses only the rollouts generated for its own triple. The known values $Q^{\pi,h}(g, a) = H - h + 1$ and $Q^{\pi,H}(s, a) = R^H(s, a)$ are used directly. At each state, the update assigns probability one to the first maximizing action in the ordering fixed in Section 2.2. Algorithm 2 summarizes the sweep.

Algorithm 2 Backward induction with fresh Monte Carlo evaluation

Require: Generative access to an instance in $\mathcal{M}_{K,A,H}$, known rewards, and integer $m \geq 1$.

1: Set $\widehat{Q}^{\pi_{0,H},H} = R^H$ and $\widehat{Q}^{\pi_{0,h},h}(g, a) = H - h + 1$ for every $h \in [H-1]$ and $a \in \mathcal{A}$.
2: **for** $h = H, H-1, \ldots, 1$ **do**
3: If $h < H$: for each $(i, a) \in [K] \times \mathcal{A}$, generate m fresh rollouts as specified above and compute $\widehat{Q}^{\pi_{0,h},h}(s_i, a)$ by equation 15.
4: For every $s \in \mathcal{S}$, choose $a_1^h(s) \in \arg\max_{a \in \mathcal{A}} \widehat{Q}^{\pi_{0,h},h}(s, a)$ using the fixed tie-breaking ordering.
5: Set $\pi_1^h(a \mid s) = \mathbf{1}\{a = a_1^h(s)\}$ for every $(s, a) \in \mathcal{S} \times \mathcal{A}$.
6: **end for**
7: **return** π_1.

If every transition is queried, including after entry into g, a rollout launched at horizon h uses $H - h$ transition queries. One sweep therefore uses

$$N = KAm \sum_{h=1}^{H-1} (H - h) = KAm \, \frac{H(H-1)}{2}. \tag{16}$$

Remark 2 treats stopping oracle queries upon entry into g.

6.2 Necessary sample count

Theorem 3. *There exists a universal constant $c_{\mathrm{BI}} > 0$ such that, for all integers $K \geq 1$, $A \geq 2$, $H \geq 2$, and $m \geq 1$, there is an instance $M \in \mathcal{M}_{K,A,H}$ on which the policy π_1 returned by Algorithm 2 satisfies*

$$\mathbb{E}_M\big[V^{\star,1}(\rho) - V^{\pi_1,1}(\rho)\big] \geq c_{\mathrm{BI}} \min\left\{H, \, \frac{H^2}{\sqrt{m}}\right\}. \tag{17}$$

Consequently, for any $0 < \varepsilon < c_{\mathrm{BI}} H$, if the algorithm satisfies equation 9 with $\pi_{\mathrm{out}} = \pi_1$, then $m \geq c_{\mathrm{BI}}^2 H^4 / \varepsilon^2$. If every rollout transition is queried, this implies $N \geq (c_{\mathrm{BI}}^2/4) KAH^6/\varepsilon^2$.

The proof in Appendix 13 lower-bounds the conditional probability that an inferior action has a strictly larger estimate than the favorable action, uniformly over continuation policies (Lemmas 10 and 11). Such events force selection errors under every fixed tie-breaking ordering; Lemma 2 converts these errors into an optimality-gap lower bound.

Remark 2. *A rollout can stop sampling when it first enters g: if entry occurs at horizon u, the remaining reward $H - u + 1$ is added directly to the return. This preserves the return and hence the distribution of π_1, so equation 17 continues to hold. Let N_{stop} denote the total transition-query count for one sweep with stopping. On the instance used to prove Theorem 3, equation 34 shows that, conditional on $\mathcal{F}_h$, a rollout launched at horizon h avoids g through horizon H with probability at least $5/8$. On this event all $H-h$ transitions are sampled, so the conditional expected query count is at least $5(H-h)/8$. Summing over the rollouts gives $\mathbb{E}_M[N_{\mathrm{stop}}] \geq \frac{5}{8} KAm\, H(H-1)/2$. If this implementation satisfies equation 9 with $\pi_{\mathrm{out}} = \pi_1$ and $0 < \varepsilon < c_{\mathrm{BI}} H$, the theorem requires $m \geq c_{\mathrm{BI}}^2 H^4/\varepsilon^2$. Using $H(H-1)/2 \geq H^2/4$ therefore yields $\mathbb{E}_M[N_{\mathrm{stop}}] \geq (5c_{\mathrm{BI}}^2/32) KAH^6/\varepsilon^2$.*

The lower bound arises from rollout fluctuations that give an inferior action a strictly larger estimate than the favorable action, which maximizes the exact action value under every continuation policy (Lemma 1). One-sweep optimality with exact action values (Remark 1) therefore does not imply sample efficiency for this Monte Carlo implementation.

6.3 Comparison of the sample requirements

Fix $0 < \varepsilon < \min\{1, c, c_{\mathrm{BI}}\}H$, so that all three results apply. On $\mathcal{M}_{K,A,H}$, one backward sweep with fresh Monte Carlo evaluation requires $N \geq (c_{\mathrm{BI}}^2/4) KAH^6/\varepsilon^2$ to satisfy equation 9 when every rollout transition is queried (Theorem 3). NPG with one-step evaluation satisfies the same criterion using $N = O(SAH^5 \log A/\varepsilon^2)$ transition queries (Corollary 1). Since $S = K + 1 \leq 2K$, for fixed A and sufficiently large H, there exists a query budget sufficient for NPG but insufficient for this backward induction implementation to guarantee the same accuracy uniformly over the family. Every agent satisfying equation 9 requires $\Omega(SAH^4/\varepsilon^2)$ transition queries (Theorem 1). The NPG upper bound remains a factor of $H \log A$ above this information-theoretic benchmark, so these results do not establish sample optimality of NPG.

The compared procedures differ in both policy evaluation and policy update. On the instance constructed for backward induction, the conditional probability of selecting an inferior action is bounded below by a positive universal constant at every decision state and horizon $1 \leq h \leq \lfloor H/2 \rfloor$ (Lemma 11). Each selected action is retained in the returned policy. For NPG, the backward evaluation recursion yields conditionally unbiased estimates (Lemma 6), and Lemma 8 bounds the conditional expected estimation-error contribution to the local update inequality by ηL_h^2. Thus, the step size controls this contribution across iterations.

7 Simulations

We compare backward induction with fresh Monte Carlo evaluation and NPG with one-step evaluation on one fixed MDP and generative oracle under approximately matched transition-query budgets.

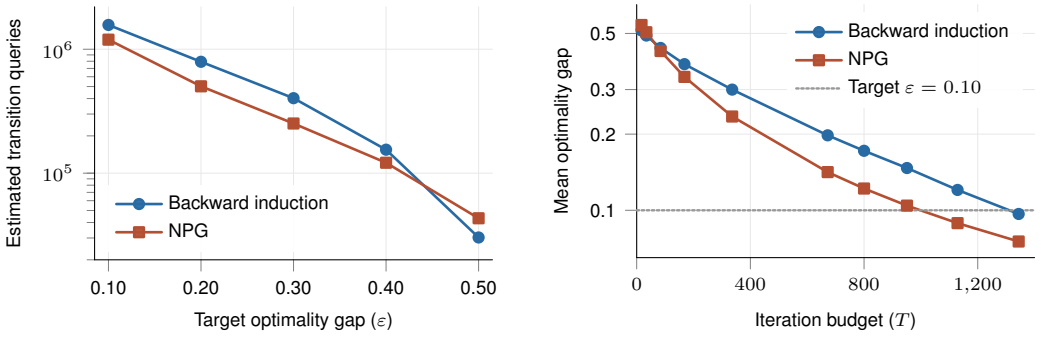
7.1 Setup

We instantiate the family of Section 4.2 with $K = 24$, $A = 8$, and $H = 7$, giving $S = 25$, and set $p^h = 1/(4H)$ and $\alpha^h = 1/(8H)$ for every $h < H$. The initial distribution is uniform over the decision states. Favorable-action labels are drawn uniformly from $[A]$ once using a fixed seed and retained across all budgets and runs. Backward induction follows Algorithm 2 with m rollouts per triple and ties broken by the smallest action index. NPG follows Algorithm 1 with one successor sample per triple per iteration and $\eta = (\log A)/H \approx 0.297$. We fix this step size across budgets so that each NPG run supplies all checkpoints. The query counts are $N_{\mathrm{BI}} = KAm\, H(H-1)/2 = 4032m$ and $N_{\mathrm{NPG}} = SA(H-1)T = 1200T$. We pair budgets using $T = \lceil 3.36m \rceil$ for $m \in \{5, 10, 25, 50, 100, 200, 238, 283, 336, 400\}$. Rounding adds at most 480 NPG queries on this grid; query estimates use the actual counts. Counts include transitions after absorption and NPG queries at g, with no transition queries at horizon H. The displayed grid ends at the first paired budget where both empirical mean gaps are at most 0.10.

7.2 Evaluation and results

Policy values are computed by exact dynamic programming for evaluation only. For backward induction, we record $V^{\star,1}(\rho) - V^{\pi_1,1}(\rho)$; for NPG, we record $T^{-1}\sum_{t=0}^{T-1}[V^{\star,1}(\rho) - V^{\pi_t,1}(\rho)]$, the conditional expected gap of the uniformly selected output. We average over 500 independent sampling runs; backward induction uses fresh sampling at each budget.

The accompanying data report sample standard deviations and approximate pointwise 95% Student-t confidence intervals for the means, with 499 degrees of freedom. Intervals are omitted from the figure for clarity. For $\varepsilon \in \{0.10, 0.20, 0.30, 0.40, 0.50\}$, query requirements are estimated by log–log interpolation of the first crossing of the empirical mean curve. Pointwise 95% percentile intervals use 10,000 bootstrap resamples of complete run vectors, preserving dependence across budgets, and are conditional on the grid and interpolation rule.[1]



(a) Estimated queries to reach target gap ε.

(b) Mean optimality gap versus iteration budget T.

Figure 1: Backward induction with fresh Monte Carlo evaluation and NPG with one-step evaluation on a fixed MDP with $S = 25$, $A = 8$, and $H = 7$. (a) Estimated transition queries to attain target mean gap ε. (b) Mean gaps over 500 runs at paired budgets. Here, T counts NPG updates, each backward induction point represents one sweep, and NPG's gap averages the iterates $\pi_0, \ldots, \pi_{T-1}$. Both vertical axes are logarithmic.

In Figure 1(a), NPG uses an estimated 21.5%–37.5% fewer queries for the tested targets 0.10–0.40, while backward induction uses fewer at $\varepsilon = 0.50$. At $\varepsilon = 0.10$, the estimates are 1.195×10^6 queries for NPG and 1.570×10^6 for backward induction.

In Figure 1(b), NPG has the smaller mean gap at every displayed budget from $T = 84$ onward. Its recorded gap also has the smaller sample standard deviation at every displayed budget. At $N = 1{,}612{,}800$ ($m = 400$, $T = 1344$), backward induction and NPG attain mean gaps of 0.0967 and 0.0753, with standard deviations 0.0219 and 0.00671, respectively. At larger budgets reported in the supplement, backward induction attains the smaller mean gap. These experiments characterize the specified procedures on one instance; the worst-case query comparison follows from the theoretical results.

8 Conclusion

Under a common generative oracle and expected-gap criterion, our bounds establish an asymptotic separation between NPG with one-step evaluation and one-sweep backward induction with fresh Monte Carlo evaluation. Overall, this paper clarifies the significance of NPG in sample-based RL. Closing the $H \log A$ gap between the NPG upper bound and the information-theoretic lower bound, and extending the comparison to online episodic interaction, remain open.

[1] Five backward induction bootstrap curves do not reach 0.10. Their crossings are retained as exceeding the largest displayed budget; both percentile endpoints remain within the grid.

AI USE STATEMENT

We used large language models to assist with drafting and polishing the manuscript, including improving clarity, conciseness, and organization. The authors reviewed and revised the AI-assisted text for accuracy and consistency with the theoretical analysis and experimental results (to check the code). We take responsibility for the final content of this work, including all AI-assisted material.

ETHICS STATEMENT

This work concerns theoretical analysis and simulations on constructed finite-horizon Markov decision processes. It does not involve human participants, personal data, or deployment in real-world decision-making systems.

REPRODUCIBILITY STATEMENT

The main text states the assumptions and sampling model, and the appendices provide proofs of the theoretical results. Algorithms 2 and 1 specify the compared procedures. The simulations provide empirical support for the theoretical comparison. Section 7 details the MDP construction, parameter choices, query budgets, and evaluation procedures, with additional simulation results in the supplementary material. The simulation code will be available upon request.

REFERENCES

Alekh Agarwal, Sham Kakade, and Lin F. Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Proceedings of the 33rd Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 67–83. PMLR, 2020.

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021. URL https://jmlr.org/papers/v22/19-736.html.

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. URL https://doi.org/10.1007/s10994-013-5368-1.

J. Bagnell, Sham M Kakade, Jeff Schneider, and Andrew Ng. Policy search by dynamic programming. In S. Thrun, L. Saul, and B. Schölkopf (eds.), *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2003. URL https://proceedings.neurips.cc/paper_files/paper/2003/file/3837a451cd0abc5ce4069304c5442c87-Paper.pdf.

Asha Barua and Sajad Khodadadian. Finite-time analysis of the natural policy gradient in finite-horizon Markov decision processes. In *Reinforcement Learning Conference*, 2026. arXiv:2607.22982.

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1283–1294. PMLR, 2020. URL https://proceedings.mlr.press/v119/cai20d.html.

Omar Darwiche Domingues, Pierre Ménard, Emilie Kaufmann, and Michal Valko. Episodic reinforcement learning in finite MDPs: Minimax lower bounds revisited. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pp. 578–598. PMLR, 2021. URL https://proceedings.mlr.press/v132/domingues21a.html.

Chi Jin, Zeyuan Allen-Zhu, Sébastien Bubeck, and Michael I. Jordan. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Emmeran Johnson, Ciara Pike-Burke, and Patrick Rebeschini. Optimal convergence rate for exact policy mirror descent in discounted Markov decision processes. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f0d7b528c31bc3f9a0d5bab515ed6ed5-Abstract-Conference.html.

Sham M. Kakade. A natural policy gradient. In *Advances in Neural Information Processing Systems*, volume 14, 2001. URL https://proceedings.neurips.cc/paper/2001/hash/4b86abe48d358ecf194c56c69108433e-Abstract.html.

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

Michael Kearns, Yishay Mansour, and Andrew Y. Ng. A sparse sampling algorithm for near-optimal planning in large Markov decision processes. *Machine Learning*, 49(2–3):193–208, 2002. doi: 10.1023/A:1017932429737.

Sara Klein, Simon Weissmann, and Leif Döring. Beyond stationarity: Convergence analysis of stochastic softmax policy gradient methods. In *International Conference on Learning Representations*, 2024. arXiv:2310.02671.

Guanghui Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical Programming*, 198(1):1059–1106, 2023. doi: 10.1007/s10107-022-01816-5.

Alessandro Lazaric, Mohammad Ghavamzadeh, and Rémi Munos. Analysis of classification-based policy iteration algorithms. *Journal of Machine Learning Research*, 17(19):1–30, 2016.

Gen Li, Yuting Wei, Yuejie Chi, and Yuxin Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Operations Research*, 72(1):203–221, 2024. doi: 10.1287/opre.2023.2451. URL https://doi.org/10.1287/opre.2023.2451.

Jiacai Liu, Wenye Li, and Ke Wei. On the convergence of policy mirror descent with temporal difference evaluation, 2025. arXiv:2509.18822.

Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Emilie Kaufmann, Edouard Leurent, and Michal Valko. Fast active learning for pure exploration in reinforcement learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7599–7608. PMLR, 2021. URL https://proceedings.mlr.press/v139/menard21a.html.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 1994. doi: 10.1002/9780470316887. URL https://doi.org/10.1002/9780470316887.

Aaron Sidford, Mengdi Wang, Xian Wu, Lin Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving markov decision processes with a generative model. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/bb03e43ffe34eeb242a2ee4a4f125e56-Paper.pdf.

Kevin Tan, Wei Fan, and Yuting Wei. Actor-critics can achieve optimal sample efficiency. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 58545–58590. PMLR, 2025. arXiv:2505.03710.

Lin Xiao. On the convergence rates of policy gradient methods. *Journal of Machine Learning Research*, 23(282):1–36, 2022. URL https://jmlr.org/papers/v23/22-0056.html.

Ming Yin, Yu Bai, and Yu-Xiang Wang. Near-optimal provable uniform convergence in offline policy evaluation for reinforcement learning. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1567–1575. PMLR, 2021. URL https://proceedings.mlr.press/v130/yin21a.html.

Zihan Zhang, Yuxin Chen, Jason D. Lee, and Simon S. Du. Settling the sample complexity of online reinforcement learning. In *Proceedings of the 37th Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*. PMLR, 2024. arXiv:2307.13586.

9 APPENDIX

9.1 EXTENDED RELATED WORK

Assumptions and sample accounting: Domingues et al. (2021) establish their lower bound under $H \geq 4$, $0 < \varepsilon \leq H/24$, and $0 < \delta \leq 1/16$. Their Assumption 1 additionally requires $S \geq 6$, $A \geq 2$, and an integer $d \geq 1$ such that $S = 3 + (A^d - 1)/(A - 1)$ and $H \geq 3d$. The result allows randomized output policies and random stopping times, and lower-bounds the expected number of episodes. For $0 < \varepsilon \leq 1/S^2$, Ménard et al. (2021) establish PAC correctness and a high-probability stopping-time bound for BPI-UCBVI. The modified MVP row uses the sample-complexity consequence stated in Zhang et al. (2024), obtained from their regret bound through the regret-to-policy-identification reduction of Jin et al. (2018). The random-iterate reduction directly yields a constant-confidence guarantee; independent repetition and policy evaluation extend it to arbitrary confidence with additional logarithmic factors. Under our transition-query convention, we omit the transition after the terminal reward and therefore multiply episode counts by $H - 1$.

Sidford et al. (2018) analyze finite-horizon MDPs with horizon-independent transition kernels. Applying their result to horizon-dependent kernels by augmenting the state with the horizon index introduces an additional factor of H in the state count. The bound cited from Yin et al. (2021) is an offline policy-optimization guarantee. Its generative-model specialization, stated in their Section 4, replaces the inverse coverage parameter by SA. For these generative-model comparisons, we use the accuracy ranges reported in Table 2 of Li et al. (2024): $0 < \varepsilon \leq 1$ for variance-reduced QVI, $0 < \varepsilon \leq \sqrt{H}$ for the specialization of Yin et al. (2021), and $0 < \varepsilon \leq H$ for Li et al. (2024).

The guarantees of Sidford et al. (2018); Li et al. (2024) hold uniformly over states and horizons. The BPI-UCBVI guarantee of Ménard et al. (2021) concerns the value at a fixed initial state. The cited results of Domingues et al. (2021); Zhang et al. (2024); Yin et al. (2021) use an optimality-gap criterion under the initial distribution.

Conversion to the expected-gap criterion: For $0 < \varepsilon < H$, apply Li et al. (2024) with accuracy $\varepsilon/2$ and failure probability $\delta = \varepsilon/(2H)$. Let π_{out} denote the policy returned by this planner. Since the optimality gap is bounded by H, on every instance M,

$$\mathbb{E}_M\big[V^{\star,1}(\rho) - V^{\pi_{\mathrm{out}},1}(\rho)\big] \leq \frac{\varepsilon}{2} + H\delta = \varepsilon.$$

This conversion changes the sample requirement only through constant and logarithmic factors. The planner of Li et al. (2024) uses uniform sample allocation, drawing the same number of independent transition samples for each state-action-horizon triple. Theorem 1 permits each queried triple to depend on preceding observations and the agent's internal randomness, so it also covers adaptive sample allocation.

Hard-max updates in sample-based planning: Sparse sampling (Kearns et al., 2002) approximates Bellman maximization through recursively sampled lookahead trees in discounted MDPs. Its estimates differ from averages of complete rollouts following a fixed continuation policy. Policy search by dynamic programming (Bagnell et al., 2003) constructs a finite-horizon policy in backward order, optimizing each decision rule over a policy class while holding the continuation fixed.

Classification-based policy iteration (Lazaric et al., 2016) estimates action values through rollouts and approximates greedy policy improvement through classification. Its high-probability finite-sample analysis accounts for sampling error, policy-class approximation, and distribution mismatch in discounted MDPs. These analyses provide sufficient sample counts for their respective planning and policy-improvement procedures.

Finite-horizon policy optimization: Klein et al. (2024) analyze softmax policy gradient under simultaneous updates and dynamic training, which optimizes the decision rules in backward or-

der. They establish convergence with exact gradients and high-probability guarantees with stochastic gradients. Their stochastic guarantees assume full-support training distributions, together with specified step sizes and batch sizes. Barua & Khodadadian (2026) study finite-time convergence of finite-horizon NPG with exact action values under constant and increasing step sizes.

For online interaction, Cai et al. (2020) combine multiplicative policy updates with optimistic value estimation and establish high-probability regret guarantees under linear transition structure. Tan et al. (2025) establish high-probability regret guarantees for actor-critic methods under function-approximation assumptions. Their principal sample-efficient algorithm combines optimistic off-policy estimation of the optimal action-value function with infrequent critic updates and policy resets. These results address exploration together with policy optimization.

PMD or NPG in discounted MDPs: Generative-model sample complexity in discounted MDPs, with discount factor $\gamma \in (0,1)$, has been studied through minimax lower bounds and model-based planning (Azar et al., 2013; Agarwal et al., 2020). For unregularized stochastic policy mirror descent, Lan (2023) bounds the expected gap of a uniformly selected iterate under a step size held constant over the run. The objective is weighted by a stationary distribution of an optimal policy. The corresponding generative-model implementation uses one truncated rollout per state-action pair and iteration. The rollout length contributes to the transition count and must be retained when comparing dependence on the discount factor.

Xiao (2022) analyzes unregularized policy mirror descent with truncated Monte Carlo evaluation and geometrically increasing step sizes. The resulting high-probability guarantee depends on a distribution-mismatch factor. Johnson et al. (2023) remove this dependence and obtain a high-probability sup-norm guarantee for the last iterate using $\widetilde{O}(SA(1-\gamma)^{-8}\varepsilon^{-2})$ samples, with step sizes satisfying an adaptive condition. More recently, Liu et al. (2025) obtain a high-probability sup-norm guarantee for the last iterate using $\widetilde{O}(SA(1-\gamma)^{-7}\varepsilon^{-2})$ transition samples. Their method uses one-step temporal-difference evaluation based on the previous value estimate, with step sizes satisfying an adaptive condition.

These discounted results use different objectives, evaluation procedures, and output guarantees. Their discount-factor dependence therefore does not translate directly into finite-horizon sample complexity by substituting H for $(1-\gamma)^{-1}$.

10 Properties of the MDP family

Throughout this section, fix an instance $M_\theta \in \mathcal{M}_{K,A,H}$, defined in Section 4.2.

Lemma 1. *For a policy π, define the average value and the average probability of transitioning to g by*

$$\overline{V}^{\pi,h} := \frac{1}{K}\sum_{i=1}^{K} V^{\pi,h}(s_i), \qquad h \in [H],$$

$$\overline{q}^{\pi,h} := \frac{1}{K}\sum_{i=1}^{K}\sum_{j=1}^{A} \pi^h(a_j \mid s_i)\, p^h_{\theta,i,a_j}, \quad h \in [H-1].$$

Then $V^{\pi,h}(g) = H - h + 1$ for every $h \in [H]$, and $\overline{V}^{\pi,H} = 0$. For every $h \in [H-1]$, $i \in [K]$, and $j \in [A]$,

$$Q^{\pi,h}(s_i,a_j) = p^h_{\theta,i,a_j}(H-h) + \big(1 - p^h_{\theta,i,a_j}\big)\overline{V}^{\pi,h+1}, \tag{18}$$

and

$$\overline{V}^{\pi,h} = \overline{q}^{\pi,h}(H-h) + \big(1-\overline{q}^{\pi,h}\big)\overline{V}^{\pi,h+1}. \tag{19}$$

If $\theta^h_i \neq 0$, then every $j \neq \theta^h_i$ satisfies

$$Q^{\pi,h}(s_i,a_{\theta^h_i}) - Q^{\pi,h}(s_i,a_j) = 2\alpha^h\left[(H-h) - \overline{V}^{\pi,h+1}\right] \geq 2\alpha^h > 0. \tag{20}$$

Consequently, every optimal policy $\pi^\star$ satisfies $\pi^{\star,h}(a_{\theta^h_i} \mid s_i) = 1$ whenever $\theta^h_i \neq 0$. The function $Q^{\star,h}(s_i,\cdot)$ is constant on $\mathcal{A}$ when $h = H$, or when $h \in [H-1]$ and $\theta^h_i = 0$.

Proof. Since g is absorbing and yields reward one at each remaining horizon, $V^{\pi,h}(g) = H-h+1$. At horizon H, every decision state yields zero reward, so $V^{\pi,H}(s_i) = 0$ and $\overline{V}^{\pi,H} = 0$.

Fix $h \in [H-1]$, $i \in [K]$, and $j \in [A]$. The immediate reward at s_i is zero. Substituting equation 7 into the Bellman equation gives

$$\begin{aligned} Q^{\pi,h}(s_i, a_j) &= p^h_{\theta,i,a_j} V^{\pi,h+1}(g) + \frac{1 - p^h_{\theta,i,a_j}}{K} \sum_{\ell=1}^{K} V^{\pi,h+1}(s_\ell) \\ &= p^h_{\theta,i,a_j}(H-h) + \big(1 - p^h_{\theta,i,a_j}\big)\overline{V}^{\pi,h+1}, \end{aligned}$$

which proves equation 18. Averaging over actions according to π^h and then uniformly over decision states yields

$$\begin{aligned} \overline{V}^{\pi,h} &= \frac{1}{K} \sum_{i=1}^{K} \sum_{j=1}^{A} \pi^h(a_j \mid s_i) Q^{\pi,h}(s_i, a_j) \\ &= \overline{q}^{\pi,h}(H-h) + \big(1 - \overline{q}^{\pi,h}\big)\overline{V}^{\pi,h+1}, \end{aligned}$$

where the last equality uses the definition of $\overline{q}^{\pi,h}$ and $\sum_{j=1}^{A} \pi^h(a_j \mid s_i) = 1$. This proves equation 19.

Suppose $\theta^h_i \neq 0$ and $j \neq \theta^h_i$. By equation 6,

$$p^h_{\theta,i,a_{\theta^h_i}} - p^h_{\theta,i,a_j} = 2\alpha^h.$$

Subtracting the corresponding expressions in equation 18 gives the equality in equation 20. Starting from a decision state at horizon $h+1$, the reward at that horizon is zero, and each subsequent reward is at most one. Hence,

$$\overline{V}^{\pi,h+1} \leq H-h-1, \qquad (H-h) - \overline{V}^{\pi,h+1} \geq 1.$$

Together with $\alpha^h > 0$, this proves equation 20.

Let $\pi^\star$ be optimal. By Bellman optimality (Remark 1),

$$\sum_{j=1}^{A} \pi^{\star,h}(a_j \mid s_i) Q^{\star,h}(s_i, a_j) = V^{\star,h}(s_i) = \max_{j \in [A]} Q^{\star,h}(s_i, a_j).$$

When $\theta^h_i \neq 0$, applying equation 20 with $\pi = \pi^\star$ shows that $a_{\theta^h_i}$ is the unique maximizing action. The displayed equality therefore requires $\pi^{\star,h}(a_{\theta^h_i} \mid s_i) = 1$.

When $\theta^h_i = 0$, all actions have transition probability $p^h_{\theta,i,a_j} = p^h - \alpha^h$, so equation 18 implies that $Q^{\star,h}(s_i, \cdot)$ is constant. At horizon H, $Q^{\star,H}(s_i, a) = R^H(s_i, a) = 0$ for every $a \in \mathcal{A}$, completing the proof. □

The next lemma expresses the optimality gap in terms of the probability assigned to suboptimal actions. For a policy π, define

$$w^h_\pi(i) := \begin{cases} 0, & \theta^h_i = 0, \\ 1 - \pi^h(a_{\theta^h_i} \mid s_i), & \theta^h_i \neq 0, \end{cases} \qquad \overline{w}^h_\pi := \frac{1}{K} \sum_{i=1}^{K} w^h_\pi(i), \qquad h \in [H-1],\ i \in [K].$$

Write $\overline{V}^{\star,h}$ and $\overline{q}^{\star,h}$ for the corresponding averages under an optimal policy. By Lemma 1,

$$\overline{q}^{\star,h} = p^h - \alpha^h + \frac{2\alpha^h}{K} \sum_{i=1}^{K} \mathbf{1}\{\theta^h_i \neq 0\}, \qquad h \in [H-1],$$

which is independent of the choice of optimal policy. Define

$$\omega_h := \prod_{u=1}^{h-1} \big(1 - \overline{q}^{\star,u}\big), \qquad h \in [H-1],$$

where the empty product gives $\omega_1 = 1$. The following lemma holds for every policy and hence for each realization of π_{out}.

Lemma 2. *For every policy π, we have*

$$V^{\star,1}(\rho) - V^{\pi,1}(\rho) = \sum_{h=1}^{H-1} 2\omega_h \alpha^h \left[(H-h) - \overline{V}^{\pi,h+1}\right] \overline{w}_\pi^h, \tag{21}$$

and

$$V^{\star,1}(\rho) - V^{\pi,1}(\rho) \geq \frac{3}{4} \sum_{h=1}^{H-1} \alpha^h (H-h)\, \overline{w}_\pi^h. \tag{22}$$

Proof. Fix a policy π and a horizon $h \in [H-1]$. If $\theta_i^h \neq 0$, then equation 6 and the definition of $w_\pi^h(i)$ give

$$\begin{aligned}\sum_{j=1}^{A} \pi^h(a_j \mid s_i)\, p^h_{\theta,i,a_j} &= p^h - \alpha^h + 2\alpha^h \pi^h(a_{\theta_i^h} \mid s_i) \\ &= p^h + \alpha^h - 2\alpha^h w_\pi^h(i).\end{aligned}$$

At this state and horizon, every optimal policy enters g with probability $p^h + \alpha^h$. If $\theta_i^h = 0$, every policy enters g with probability $p^h - \alpha^h$, and $w_\pi^h(i) = 0$. Averaging over the decision states therefore yields

$$\overline{q}^{\star,h} - \overline{q}^{\pi,h} = 2\alpha^h\, \overline{w}_\pi^h.$$

For $h \in [H]$, define $\Delta^h := \overline{V}^{\star,h} - \overline{V}^{\pi,h}$. Subtracting the two recursions in equation 19 gives

$$\begin{aligned}\Delta^h &= \left(\overline{q}^{\star,h} - \overline{q}^{\pi,h}\right)\left[(H-h) - \overline{V}^{\pi,h+1}\right] + \left(1 - \overline{q}^{\star,h}\right)\Delta^{h+1} \\ &= 2\alpha^h \overline{w}_\pi^h \left[(H-h) - \overline{V}^{\pi,h+1}\right] + \left(1 - \overline{q}^{\star,h}\right)\Delta^{h+1}.\end{aligned}$$

Since $\Delta^H = 0$ by Lemma 1, iterating this recursion and using the definition of ω_h yields

$$\Delta^1 = \sum_{h=1}^{H-1} 2\omega_h \alpha^h \left[(H-h) - \overline{V}^{\pi,h+1}\right] \overline{w}_\pi^h.$$

The initial distribution ρ is uniform over the decision states, so $\Delta^1 = V^{\star,1}(\rho) - V^{\pi,1}(\rho)$. This proves equation 21.

We next bound the coefficients in this identity. By equation 5, the probability of entering g from any decision state under any action is at most $1/(2H)$. Thus, $\overline{q}^{\star,u} \leq 1/(2H)$, and Bernoulli's inequality gives

$$\omega_h \geq \left(1 - \frac{1}{2H}\right)^{h-1} \geq 1 - \frac{h-1}{2H} \geq \frac{1}{2}, \qquad h \in [H-1].$$

Fix $h \in [H-1]$ and $i \in [K]$. For $u \in \{h+1, \ldots, H\}$, $d_{s_i}^{h+1 \to u, \pi}(g)$ is the probability of being at g at horizon u, starting from s_i at horizon $h+1$ and following π. Since g is absorbing, this event requires a first entry during one of the $u-h-1$ preceding transitions. The conditional entry probability at each transition is at most $1/(2H)$, so a union bound gives

$$d_{s_i}^{h+1 \to u, \pi}(g) \leq \frac{u-h-1}{2H}.$$

Using $R^u(s,a) = \mathbf{1}\{s = g\}$, we obtain

$$\begin{aligned}V^{\pi,h+1}(s_i) &= \sum_{u=h+1}^{H} d_{s_i}^{h+1 \to u, \pi}(g) \\ &\leq \frac{1}{2H} \sum_{u=h+1}^{H} (u-h-1) \\ &= \frac{(H-h-1)(H-h)}{4H} \leq \frac{H-h}{4}.\end{aligned}$$

Averaging over i gives $\overline{V}^{\pi,h+1} \le (H-h)/4$. Together with $\omega_h \ge 1/2$, this implies

$$2\omega_h \left[(H-h) - \overline{V}^{\pi,h+1}\right] \ge 2 \cdot \frac{1}{2} \cdot \frac{3}{4}(H-h) = \frac{3}{4}(H-h).$$

Since $\alpha^h \overline{w}_\pi^h \ge 0$, substituting this bound into equation 21 proves equation 22. □

11 Proof of the information-theoretic lower bound

Now, considering the MDP family defined in Section 4.2, we relate action-selection errors to transition-query counts. Throughout the following lemmas, fix the parameters $(p^h, \alpha^h)_{h\in[H-1]}$ and an agent using at most N transition queries. For $\theta \in \Theta$, let $\mathbb{P}_\theta$ denote the joint law of the agent's queries, oracle responses, internal randomness, and returned policy under M_θ. Write $\mathbb{E}_\theta$ for the corresponding expectation, so $\mathbb{E}_\theta = \mathbb{E}_{M_\theta}$ in equation 3. For each pair (h, i), we use the instance with $\theta_i^h = 0$ as the reference while holding all other labels fixed.

For $h \in [H-1]$, $i \in [K]$, and $j \in [A]$, let $C_i^h(a_j)$ be the number of queries at (h, s_i, a_j), and define

$$C_i^h := \sum_{j=1}^{A} C_i^h(a_j).$$

The query budget implies

$$\sum_{h=1}^{H-1}\sum_{i=1}^{K} C_i^h \le N \qquad \text{almost surely.} \tag{23}$$

For probability measures $\mathbb{P}$ and $\mathbb{Q}$, write $D_{\mathrm{KL}}(\mathbb{P}\|\mathbb{Q})$ for their Kullback–Leibler divergence, using natural logarithms, and define their total variation distance by

$$\|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}} := \sup_E |\mathbb{P}(E) - \mathbb{Q}(E)|,$$

where the supremum is over measurable events. For $p, q \in (0,1)$, the response distributions B_p and B_q defined in equation 4 satisfy

$$\begin{aligned} D_{\mathrm{KL}}(B_p\|B_q) &= p\log\frac{p}{q} + \sum_{\ell=1}^{K}\frac{1-p}{K}\log\frac{(1-p)/K}{(1-q)/K} \\ &= p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}, \end{aligned}$$

because both are uniform over the decision states conditional on not entering g. For later use, define

$$\kappa_h := D_{\mathrm{KL}}\left(B_{p^h-\alpha^h}\middle\|B_{p^h+\alpha^h}\right), \qquad h \in [H-1].$$

Lemma 3. *For any $\theta, \theta' \in \Theta$,*

$$D_{\mathrm{KL}}(\mathbb{P}_\theta\|\mathbb{P}_{\theta'}) = \sum_{h=1}^{H-1}\sum_{i=1}^{K}\sum_{j=1}^{A}\mathbb{E}_\theta[C_i^h(a_j)]D_{\mathrm{KL}}\left(B_{p^h_{\theta,i,a_j}}\middle\|B_{p^h_{\theta',i,a_j}}\right). \tag{24}$$

Proof. Let $\tau \le N$ be the number of queries made by the agent. For $n \le \tau$, write $X_n = (h_n, s_n, a_n)$ for the n-th queried triple and Y_n for its response. For each $n \in [N]$, let $\mathcal{G}_{n-1}$ be the σ-field generated by the preceding queries and responses, together with the internal randomness used to decide whether and where to make the n-th query. Then $\{n \le \tau\} \in \mathcal{G}_{n-1}$, and X_n is $\mathcal{G}_{n-1}$-measurable on this event. Conditional on $\mathcal{G}_{n-1}$ and on making the query, Y_n has distribution $P_\theta^{h_n}(\cdot \mid s_n, a_n)$.

For a fixed history, the agent uses the same conditional rules for internal randomization, query selection, stopping, and output under both instances. These factors therefore cancel in the likelihood ratio, giving

$$\log\frac{d\mathbb{P}_\theta}{d\mathbb{P}_{\theta'}} = \sum_{n=1}^{\tau}\log\frac{P_\theta^{h_n}(Y_n \mid s_n, a_n)}{P_{\theta'}^{h_n}(Y_n \mid s_n, a_n)}.$$

The ratio is well defined on the common support. Indeed, the parameter restrictions ensure that every next state has positive probability following a query at a decision state, whereas a query at g returns g under both instances.

Taking $\mathbb{E}_\theta$ and conditioning each summand on $\mathcal{G}_{n-1}$ yields

$$D_{\mathrm{KL}}(\mathbb{P}_\theta \| \mathbb{P}_{\theta'}) = \mathbb{E}_\theta \left[\sum_{n=1}^{\tau} D_{\mathrm{KL}}\Big(P_\theta^{h_n}(\cdot \mid s_n, a_n) \Big\| P_{\theta'}^{h_n}(\cdot \mid s_n, a_n) \Big) \right].$$

This conditioning is valid because $\tau \le N$ and $\mathcal{G}_{n-1}$ determines both whether the n-th query is made and which triple is queried.

Queries at g contribute zero. For a query at (h, s_i, a_j), the response laws are $B_{p^h_{\theta,i,a_j}}$ and $B_{p^h_{\theta',i,a_j}}$ by equation 7. Grouping the remaining terms by their queried triples gives equation 24. □

Lemma 4. *Fix $h \in [H-1]$, $i \in [K]$, and all labels other than θ_i^h. For $j \in \{0, 1, \dots, A\}$, let $\mathbb{P}_j$ and $\mathbb{E}_j$ denote the law and expectation under the instance with $\theta_i^h = j$. Then*

$$\frac{1}{A} \sum_{j=1}^{A} \mathbb{E}_j\big[1 - \pi_{\mathrm{out}}^h(a_j \mid s_i)\big] \ge 1 - \frac{1}{A} - \sqrt{\frac{\kappa_h \, \mathbb{E}_0[C_i^h]}{2A}}. \tag{25}$$

Proof. Fix $j \in [A]$. Changing θ_i^h from 0 to j changes only the response distribution at (h, s_i, a_j), from $B_{p^h - \alpha^h}$ to $B_{p^h + \alpha^h}$. Lemma 3 and the definition of κ_h therefore give

$$D_{\mathrm{KL}}(\mathbb{P}_0 \| \mathbb{P}_j) = \kappa_h \, \mathbb{E}_0[C_i^h(a_j)].$$

Define $f_j := 1 - \pi_{\mathrm{out}}^h(a_j \mid s_i)$. Since $f_j \in [0,1]$, its expectations under two probability measures differ by at most their total variation distance. Applying this fact and Pinsker's inequality yields

$$\begin{aligned} \mathbb{E}_j[f_j] &\ge \mathbb{E}_0[f_j] - \|\mathbb{P}_0 - \mathbb{P}_j\|_{\mathrm{TV}} \\ &\ge \mathbb{E}_0[f_j] - \sqrt{\frac{1}{2} D_{\mathrm{KL}}(\mathbb{P}_0 \| \mathbb{P}_j)} \\ &= \mathbb{E}_0[f_j] - \sqrt{\frac{\kappa_h \, \mathbb{E}_0[C_i^h(a_j)]}{2}}. \end{aligned}$$

The normalization $\sum_{j=1}^{A} \pi_{\mathrm{out}}^h(a_j \mid s_i) = 1$ implies

$$\frac{1}{A} \sum_{j=1}^{A} \mathbb{E}_0[f_j] = 1 - \frac{1}{A}.$$

Averaging the preceding inequality over j and applying Jensen's inequality to the concave square-root function gives

$$\begin{aligned} \frac{1}{A} \sum_{j=1}^{A} \mathbb{E}_j[f_j] &\ge 1 - \frac{1}{A} - \frac{1}{A} \sum_{j=1}^{A} \sqrt{\frac{\kappa_h \, \mathbb{E}_0[C_i^h(a_j)]}{2}} \\ &\ge 1 - \frac{1}{A} - \sqrt{\frac{\kappa_h}{2A} \sum_{j=1}^{A} \mathbb{E}_0[C_i^h(a_j)]} \\ &= 1 - \frac{1}{A} - \sqrt{\frac{\kappa_h \, \mathbb{E}_0[C_i^h]}{2A}}, \end{aligned}$$

where the last equality uses $C_i^h = \sum_{j=1}^{A} C_i^h(a_j)$. This proves equation 25. □

Lemma 4 uses a separate reference instance for each pair (h, i). To combine these comparisons under a common query budget, we average over a distribution on the instance labels.

Lemma 5. *Let ν be the distribution on Θ under which the labels are independent and, for every $h \in [H-1]$ and $i \in [K]$,*

$$\nu(\theta_i^h = 0) = \frac{1}{2}, \qquad \nu(\theta_i^h = j) = \frac{1}{2A}, \quad j \in [A].$$

Let $\mathbb{E}_\nu$ denote expectation over $\theta \sim \nu$, the oracle observations, and the agent's internal randomness. All value functions inside this expectation are evaluated in the sampled instance M_θ. For $x \in \mathbb{R}$, write $[x]_+ := \max\{x, 0\}$. Then

$$\begin{aligned}&\mathbb{E}_\nu\big[V^{\star,1}(\rho) - V^{\pi_{\mathrm{out}},1}(\rho)\big] \\ &\quad\geq \frac{3}{8}\left[\left(1-\frac{1}{A}\right)\sum_{h=1}^{H-1}\alpha^h(H-h) - \sqrt{\frac{N}{AK}\sum_{h=1}^{H-1}\kappa_h(\alpha^h)^2(H-h)^2}\right]_+ .\end{aligned} \tag{26}$$

Consequently, some $\theta \in \Theta$ has expected optimality gap at least the right-hand side of equation 26.

Proof. Fix $h \in [H-1]$ and $i \in [K]$. Let $\theta_{-(h,i)}$ denote all labels other than θ_i^h, and write $\mathbb{E}_{\theta_{-(h,i)}}$ for expectation over their marginal distribution under ν. For each fixed $\theta_{-(h,i)}$, use the expectations $\mathbb{E}_j$ from Lemma 4.

By definition, $w^h_{\pi_{\mathrm{out}}}(i) = 0$ when $\theta_i^h = 0$. Since each positive label has probability $1/(2A)$,

$$\mathbb{E}_\nu[w^h_{\pi_{\mathrm{out}}}(i)] = \frac{1}{2}\mathbb{E}_{\theta_{-(h,i)}}\left[\frac{1}{A}\sum_{j=1}^{A}\mathbb{E}_j\big[1-\pi^h_{\mathrm{out}}(a_j \mid s_i)\big]\right].$$

Applying Lemma 4 and then Jensen's inequality yields

$$\begin{aligned}\mathbb{E}_\nu[w^h_{\pi_{\mathrm{out}}}(i)] &\geq \frac{1}{2}\left[1-\frac{1}{A}-\mathbb{E}_{\theta_{-(h,i)}}\left[\sqrt{\frac{\kappa_h\,\mathbb{E}_0[C_i^h]}{2A}}\right]\right] \\ &\geq \frac{1}{2}\left[1-\frac{1}{A}-\sqrt{\frac{\kappa_h}{2A}\mathbb{E}_\nu[C_i^h \mid \theta_i^h = 0]}\right].\end{aligned}$$

The last step uses independence of the labels, which gives

$$\mathbb{E}_{\theta_{-(h,i)}}\mathbb{E}_0[C_i^h] = \mathbb{E}_\nu[C_i^h \mid \theta_i^h = 0].$$

For every $h \in [H-1]$ and $i \in [K]$, define $n_i^h := \mathbb{E}_\nu[C_i^h]$. Nonnegativity of the query counts implies

$$n_i^h \geq \mathbb{E}_\nu\big[C_i^h \mathbf{1}\{\theta_i^h = 0\}\big] = \frac{1}{2}\mathbb{E}_\nu[C_i^h \mid \theta_i^h = 0].$$

Substituting into the preceding bound gives

$$\mathbb{E}_\nu[w^h_{\pi_{\mathrm{out}}}(i)] \geq \frac{1}{2}\left[1-\frac{1}{A}-\sqrt{\frac{\kappa_h n_i^h}{A}}\right].$$

Moreover, equation 23 implies

$$\sum_{h=1}^{H-1}\sum_{i=1}^{K} n_i^h \leq N.$$

The policy-loss bound equation 22 holds for every instance and every realization of π_{out}. Taking expectations and applying the decision-error bound above, with α^h fixed across instances, yields

$$\begin{aligned}&\mathbb{E}_\nu\big[V^{\star,1}(\rho) - V^{\pi_{\mathrm{out}},1}(\rho)\big] \\ &\quad\geq \frac{3}{4}\sum_{h=1}^{H-1}\alpha^h(H-h)\mathbb{E}_\nu[\overline{w}^h_{\pi_{\mathrm{out}}}] \\ &\quad\geq \frac{3}{8}\left[\left(1-\frac{1}{A}\right)\sum_{h=1}^{H-1}\alpha^h(H-h) - \frac{1}{K\sqrt{A}}\sum_{h=1}^{H-1}\sum_{i=1}^{K}\alpha^h(H-h)\sqrt{\kappa_h n_i^h}\right].\end{aligned}$$

By Cauchy–Schwarz and the query budget,

$$\begin{aligned}
&\sum_{h=1}^{H-1}\sum_{i=1}^{K}\alpha^h(H-h)\sqrt{\kappa_h n_i^h} \\
&\quad\le \left[\sum_{h=1}^{H-1}\sum_{i=1}^{K}\kappa_h(\alpha^h)^2(H-h)^2\right]^{1/2}\left[\sum_{h=1}^{H-1}\sum_{i=1}^{K}n_i^h\right]^{1/2} \\
&\quad\le \sqrt{KN\sum_{h=1}^{H-1}\kappa_h(\alpha^h)^2(H-h)^2}.
\end{aligned}$$

Multiplying by $1/(K\sqrt{A})$ and substituting gives equation 26 without the positive part. The positive part follows because the optimality gap is nonnegative.

Finally, $\mathbb{E}_\nu$ averages the expected optimality gaps over the finite set Θ. At least one instance has expected gap no smaller than this average which concludes the result. □

Proof of Theorem 1. Fix an agent using at most N transition queries. For every $h \in [H-1]$, choose

$$p^h = p := \frac{1}{4H}, \qquad \alpha^h = \alpha := \min\left\{\frac{1}{8H}, \frac{1}{32}\sqrt{\frac{KA}{N}}\right\}.$$

Then $0 < \alpha \le 1/(8H) < p$ and $p + \alpha \le 3/(8H) \le 1/(2H)$. Thus, equation 5 holds, and the resulting instances belong to $\mathcal{M}_{K,A,H}$.

Since the parameters do not depend on h, neither does κ_h; denote its common value by κ. For $u, v \in (0,1)$, the response-divergence formula and $\log x \le x - 1$ give

$$\begin{aligned}
D_{\mathrm{KL}}(B_u \| B_v) &= u\log\frac{u}{v} + (1-u)\log\frac{1-u}{1-v} \\
&\le u\left(\frac{u}{v} - 1\right) + (1-u)\left(\frac{1-u}{1-v} - 1\right) \\
&= \frac{(u-v)^2}{v(1-v)}.
\end{aligned}$$

Applying this inequality with $u = p - \alpha$ and $v = p + \alpha$ yields

$$\kappa \le \frac{4\alpha^2}{(p+\alpha)(1-p-\alpha)} \le 32H\alpha^2,$$

where the last inequality uses $p + \alpha \ge 1/(4H)$ and $1 - p - \alpha \ge 1/2$. The choice of α therefore implies

$$\frac{N\kappa}{AK} \le 32H\,\frac{N\alpha^2}{AK} \le \frac{H}{32}.$$

Define the horizon sums

$$\Lambda_1 := \sum_{h=1}^{H-1}(H-h) = \frac{H(H-1)}{2},$$
$$\Lambda_2 := \sum_{h=1}^{H-1}(H-h)^2 = \frac{H(H-1)(2H-1)}{6}.$$

For $H \ge 2$,

$$\frac{\Lambda_2}{\Lambda_1^2} = \frac{2(2H-1)}{3H(H-1)} \le \frac{2}{H}.$$

Consequently, the square-root term in equation 26 satisfies

$$\sqrt{\frac{N}{AK}\sum_{h=1}^{H-1}\kappa\alpha^2(H-h)^2} = \alpha\sqrt{\frac{N\kappa}{AK}\Lambda_2} \le \alpha\sqrt{\frac{H}{32}\Lambda_2} \le \frac{\alpha\Lambda_1}{4}.$$

Using $1 - 1/A \ge 1/2$ in Lemma 5, we obtain

$$\mathbb{E}_\nu\big[V^{\star,1}(\rho) - V^{\pi_{\mathrm{out}},1}(\rho)\big] \ge \frac{3}{8}\left[\left(1-\frac{1}{A}\right)\alpha\Lambda_1 - \frac{\alpha\Lambda_1}{4}\right] \ge \frac{3}{32}\alpha\Lambda_1 \ge \frac{3}{128}\alpha H^2,$$

where the last inequality uses $H - 1 \ge H/2$. Substituting the choice of α gives

$$\frac{3}{128}\alpha H^2 = \frac{3}{128}\min\left\{\frac{H}{8},\, \frac{H^2}{32}\sqrt{\frac{KA}{N}}\right\} \ge \frac{3}{4096}\min\left\{H,\, H^2\sqrt{\frac{KA}{N}}\right\}.$$

By Lemma 5, some instance $M \in \mathcal{M}_{K,A,H}$ has expected optimality gap at least this quantity. This proves equation 8 with $c = 3/4096$.

Finally, suppose the agent satisfies equation 9 for some $0 < \varepsilon < cH$. Then

$$c\min\left\{H,\, H^2\sqrt{\frac{KA}{N}}\right\} \le \varepsilon < cH.$$

The minimum must equal its second argument, so $cH^2\sqrt{KA/N} \le \varepsilon$. Squaring and rearranging yields

$$N \ge \frac{c^2 KAH^4}{\varepsilon^2},$$

which proves the claim. □

12 Analysis of NPG with one-step policy evaluation

Fix an MDP M satisfying Section 2.1, an initial distribution ρ, and an optimal policy $\pi^\star$. Consider Algorithm 1 with integers $T, m \ge 1$ and constant step size $\eta > 0$. Throughout this appendix, expectations are over the oracle observations and the algorithm's internal randomness on M; we suppress the instance subscript. We use $L_h := H - h + 1$ and the sigma-fields $\mathcal{F}_t$ defined in Section 5.1.

12.1 Properties of policy evaluation

For $t \in \{0, \dots, T-1\}$, $h \in [H]$, and $s \in \mathcal{S}$, define the estimation error

$$\zeta_t^h(s,\cdot) := \widehat{Q}^{\pi_t,h}(s,\cdot) - Q^{\pi_t,h}(s,\cdot).$$

Lemma 6. *For every $t \in \{0, \dots, T-1\}$, $h \in [H]$, $s \in \mathcal{S}$, and $a \in \mathcal{A}$, almost surely,*

$$0 \le \widehat{Q}^{\pi_t,h}(s,a) \le L_h, \qquad 0 \le \widehat{V}^{\pi_t,h}(s) \le L_h,$$

and

$$\mathbb{E}\Big[\widehat{Q}^{\pi_t,h}(s,a) \mid \mathcal{F}_t\Big] = Q^{\pi_t,h}(s,a),$$
$$\mathbb{E}\Big[\widehat{V}^{\pi_t,h}(s) \mid \mathcal{F}_t\Big] = V^{\pi_t,h}(s). \tag{27}$$

Moreover, almost surely,

$$\mathbb{E}\big[\|\zeta_t^h(s,\cdot)\|_\infty^2 \,\big|\, \mathcal{F}_t\big] \le L_h^2. \tag{28}$$

Proof. Fix $t \in \{0,\dots,T-1\}$. We first prove the range bounds by backward induction. At horizon H, $\widehat{Q}^{\pi_t,H}(s,a) = R^H(s,a) \in [0,1] = [0,L_H]$. Since equation 11 is a convex combination of the action-value estimates, $\widehat{V}^{\pi_t,H}(s)$ lies in the same interval. If $\widehat{V}^{\pi_t,h+1}(s') \in [0,L_{h+1}]$ for every s', then equation 10 gives

$$0 \le \widehat{Q}^{\pi_t,h}(s,a) \le 1 + L_{h+1} = L_h.$$

Averaging over actions gives the corresponding bound for $\widehat{V}^{\pi_t,h}(s)$, completing the induction.

We next prove conditional unbiasedness. At horizon H, both estimates equal their corresponding value functions. Fix $h \in [H-1]$ and suppose

$$\mathbb{E}\Big[\widehat{V}^{\pi_t,h+1}(s') \mid \mathcal{F}_t\Big] = V^{\pi_t,h+1}(s')$$

for every $s' \in \mathcal{S}$. Let $\mathcal{G}_{t,h+1}$ be the sigma-field generated by $\mathcal{F}_t$ and all samples collected at horizons $h+1,\dots,H-1$ during iteration t. In particular, $\mathcal{G}_{t,H} = \mathcal{F}_t$. The table $\widehat{V}^{\pi_t,h+1}$ is $\mathcal{G}_{t,h+1}$-measurable. By the conditional independence specified in Section 5.1, each $S_{t,\ell}^{h+1}(s,a)$ has distribution $P^h(\cdot \mid s,a)$ given $\mathcal{G}_{t,h+1}$. Therefore,

$$\mathbb{E}\Big[\widehat{Q}^{\pi_t,h}(s,a) \,\Big|\, \mathcal{G}_{t,h+1}\Big] = R^h(s,a) + \sum_{s'\in\mathcal{S}} P^h(s' \mid s,a)\widehat{V}^{\pi_t,h+1}(s').$$

Applying the tower property and the induction hypothesis gives

$$\begin{aligned}\mathbb{E}\Big[\widehat{Q}^{\pi_t,h}(s,a) \mid \mathcal{F}_t\Big] &= R^h(s,a) + \sum_{s'\in\mathcal{S}} P^h(s' \mid s,a)V^{\pi_t,h+1}(s') \\ &= Q^{\pi_t,h}(s,a).\end{aligned}$$

Since π_t^h is $\mathcal{F}_t$-measurable,

$$\begin{aligned}\mathbb{E}\Big[\widehat{V}^{\pi_t,h}(s) \mid \mathcal{F}_t\Big] &= \sum_{a\in\mathcal{A}} \pi_t^h(a \mid s)Q^{\pi_t,h}(s,a) \\ &= V^{\pi_t,h}(s).\end{aligned}$$

This completes the induction.

Finally, both $\widehat{Q}^{\pi_t,h}(s,a)$ and $Q^{\pi_t,h}(s,a)$ lie in $[0,L_h]$. Hence $\|\zeta_t^h(s,\cdot)\|_\infty \le L_h$ almost surely. Squaring and taking conditional expectation proves equation 28. □

12.2 The NPG update with estimated action values

Uniform initialization, the bounded estimates, and equation 12 ensure that $\pi_t^h(a \mid s) > 0$ for every $t \in \{0,\dots,T\}$, $h \in [H]$, $s \in \mathcal{S}$, and $a \in \mathcal{A}$. Define the local KL potential

$$\Psi_t^h(s) := D_{\mathrm{KL}}\big(\pi^{\star,h}(\cdot \mid s) \,\big\|\, \pi_t^h(\cdot \mid s)\big).$$

For $t \in \{0,\dots,T-1\}$, define the estimation-error contribution

$$\chi_t^h(s) := \big\langle \zeta_t^h(s,\cdot), \pi_{t+1}^h(\cdot \mid s) - \pi^{\star,h}(\cdot \mid s)\big\rangle,$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product on $\mathbb{R}^{\mathcal{A}}$.

Lemma 7. *For every $t \in \{0,\dots,T-1\}$, $h \in [H]$, and $s \in \mathcal{S}$, almost surely,*

$$\big\langle Q^{\pi_t,h}(s,\cdot), \pi^{\star,h}(\cdot \mid s) - \pi_t^h(\cdot \mid s)\big\rangle \le \frac{\Psi_t^h(s) - \Psi_{t+1}^h(s)}{\eta} + \frac{\eta L_h^2}{8} + \chi_t^h(s). \tag{29}$$

Proof. We first define

$$p := \pi_t^h(\cdot \mid s), \qquad p^+ := \pi_{t+1}^h(\cdot \mid s), \qquad p^\star := \pi^{\star,h}(\cdot \mid s),$$

and

$$q := Q^{\pi_t,h}(s,\cdot), \qquad \widehat{q} := \widehat{Q}^{\pi_t,h}(s,\cdot).$$

The multiplicative update gives

$$\log \frac{p^+(a)}{p(a)} = \eta \widehat{q}(a) - \log\left(\sum_{a' \in \mathcal{A}} p(a') e^{\eta \widehat{q}(a')}\right).$$

Substituting into the KL divergences yields

$$\eta \langle \widehat{q}, p^\star - p^+ \rangle = D_{\mathrm{KL}}(p^\star \| p) - D_{\mathrm{KL}}(p^\star \| p^+) - D_{\mathrm{KL}}(p^+ \| p).$$

Using $q = \widehat{q} - \zeta_t^h(s,\cdot)$ and the definition of $\chi_t^h(s)$, we obtain

$$\langle q, p^\star - p \rangle = \frac{\Psi_t^h(s) - \Psi_{t+1}^h(s)}{\eta} + \langle q, p^+ - p \rangle - \frac{D_{\mathrm{KL}}(p^+ \| p)}{\eta} + \chi_t^h(s).$$

Since $q(a) \in [0, L_h]$ and $\sum_a (p^+(a) - p(a)) = 0$,

$$\begin{aligned} \langle q, p^+ - p \rangle &= \sum_{a \in \mathcal{A}} \left(q(a) - \frac{L_h}{2}\right)\big(p^+(a) - p(a)\big) \\ &\le \frac{L_h}{2} \|p^+ - p\|_1. \end{aligned}$$

Pinsker's inequality gives $D_{\mathrm{KL}}(p^+ \| p) \ge \frac{1}{2}\|p^+ - p\|_1^2$. Thus, with $x := \|p^+ - p\|_1$,

$$\begin{aligned} \langle q, p^+ - p \rangle - \frac{D_{\mathrm{KL}}(p^+ \| p)}{\eta} &\le \frac{L_h x}{2} - \frac{x^2}{2\eta} \\ &= -\frac{1}{2\eta}\left(x - \frac{\eta L_h}{2}\right)^2 + \frac{\eta L_h^2}{8} \\ &\le \frac{\eta L_h^2}{8}. \end{aligned}$$

Substituting this bound into the preceding decomposition proves equation 29. □

Although $\zeta_t^h(s,\cdot)$ has zero conditional mean, $\chi_t^h(s)$ need not, because the updated decision rule depends on the same estimates. The next lemma controls this dependence.

Lemma 8. *For every $t \in \{0, \dots, T-1\}$, $h \in [H]$, and $s \in \mathcal{S}$, almost surely,*

$$\mathbb{E}\big[\chi_t^h(s) \mid \mathcal{F}_t\big] \le \eta L_h^2. \tag{30}$$

Proof. Fix t, h, s, and write $q := Q^{\pi_t,h}(s,\cdot)$ and $\zeta := \zeta_t^h(s,\cdot)$. For $u \in [0,1]$, define

$$p_u(a) := \frac{\pi_t^h(a \mid s) \exp(\eta[q(a) + u\zeta(a)])}{\sum_{a' \in \mathcal{A}} \pi_t^h(a' \mid s) \exp(\eta[q(a') + u\zeta(a')])}.$$

Then $p_1 = \pi_{t+1}^h(\cdot \mid s)$, whereas p_0 is the NPG update using the true action values q. In particular, p_0 is $\mathcal{F}_t$-measurable. By equation 27, $\mathbb{E}[\zeta \mid \mathcal{F}_t] = 0$, so

$$\mathbb{E}\big[\langle \zeta, p_0 - \pi^{\star,h}(\cdot \mid s)\rangle \,\big|\, \mathcal{F}_t\big] = 0.$$

Consequently,

$$\mathbb{E}[\chi_t^h(s) \mid \mathcal{F}_t] = \mathbb{E}[\langle \zeta, p_1 - p_0 \rangle \,|\, \mathcal{F}_t].$$

For each realization of the estimates, differentiation gives

$$\frac{d}{du} p_u(a) = \eta p_u(a)\big(\zeta(a) - \langle p_u, \zeta \rangle\big).$$

Integrating over $u \in [0,1]$ therefore yields

$$\begin{aligned}\langle \zeta, p_1 - p_0 \rangle &= \eta \int_0^1 \left[\sum_{a \in \mathcal{A}} p_u(a)\zeta(a)^2 - \langle p_u, \zeta \rangle^2 \right] du \\ &\le \eta \int_0^1 \sum_{a \in \mathcal{A}} p_u(a)\zeta(a)^2 \, du \\ &\le \eta \|\zeta\|_\infty^2.\end{aligned}$$

Taking conditional expectations and applying equation 28 gives

$$\mathbb{E}[\chi_t^h(s) \mid \mathcal{F}_t] \le \eta \, \mathbb{E}\big[\|\zeta\|_\infty^2 \mid \mathcal{F}_t\big] \le \eta L_h^2.$$

□

12.3 Performance difference and the NPG guarantees

Extend the visitation notation of Section 2.1 to the initial distribution by defining

$$d_\rho^{1 \to h, \pi^\star}(s) := \sum_{s_0 \in \mathcal{S}} \rho(s_0) \, d_{s_0}^{1 \to h, \pi^\star}(s), \qquad h \in [H], \; s \in \mathcal{S}.$$

This is the state distribution at horizon h when the initial state is drawn from ρ and the policy is $\pi^\star$.

Lemma 9 (Performance difference lemma)**.** *For every policy π,*

$$\begin{aligned}&V^{\star,1}(\rho) - V^{\pi,1}(\rho) \\ &\quad = \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} d_\rho^{1 \to h, \pi^\star}(s) \left\langle Q^{\pi,h}(s, \cdot), \pi^{\star,h}(\cdot \mid s) - \pi^h(\cdot \mid s) \right\rangle.\end{aligned} \tag{31}$$

Proof. For $h \in [H]$, write $d_h(s) := d_\rho^{1 \to h, \pi^\star}(s)$. For $h \in [H-1]$, the Bellman equation and the visitation recursion give

$$\begin{aligned}&\sum_{s \in \mathcal{S}} d_h(s) \left\langle Q^{\pi,h}(s, \cdot), \pi^{\star,h}(\cdot \mid s) \right\rangle \\ &\quad = \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} d_h(s) \pi^{\star,h}(a \mid s) R^h(s,a) + \sum_{s' \in \mathcal{S}} d_{h+1}(s') V^{\pi,h+1}(s').\end{aligned}$$

At horizon H, $Q^{\pi,H} = R^H$, so only the reward term remains. Moreover,

$$\left\langle Q^{\pi,h}(s, \cdot), \pi^h(\cdot \mid s) \right\rangle = V^{\pi,h}(s), \qquad h \in [H].$$

Subtracting these value terms and summing over h telescopes the continuation values. Since $d_1 = \rho$, we obtain

$$\begin{aligned}&\sum_{h=1}^{H} \sum_{s \in \mathcal{S}} d_h(s) \left\langle Q^{\pi,h}(s, \cdot), \pi^{\star,h}(\cdot \mid s) - \pi^h(\cdot \mid s) \right\rangle \\ &\quad = \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} d_h(s) \pi^{\star,h}(a \mid s) R^h(s,a) - \sum_{s \in \mathcal{S}} \rho(s) V^{\pi,1}(s) \\ &\quad = V^{\star,1}(\rho) - V^{\pi,1}(\rho).\end{aligned}$$

□

Proof of Theorem 2. For $t \in \{0, \dots, T\}$, define the aggregate KL potential

$$\Phi_t := \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} d_\rho^{1 \to h, \pi^\star}(s) \Psi_t^h(s).$$

Apply Lemma 9 with $\pi = \pi_t$, bound each inner product using Lemma 7, and take conditional expectations given $\mathcal{F}_t$. The visitation weights are deterministic and sum to one at each horizon, while the optimality gap and Φ_t are $\mathcal{F}_t$-measurable. Lemma 8 therefore gives

$$V^{\star,1}(\rho) - V^{\pi_t,1}(\rho) \le \frac{\Phi_t - \mathbb{E}[\Phi_{t+1} \mid \mathcal{F}_t]}{\eta} + \frac{9\eta}{8} \sum_{h=1}^{H} L_h^2.$$

Taking expectations and summing over $t = 0, \dots, T-1$ telescopes the potential differences:

$$\begin{aligned} \sum_{t=0}^{T-1} \mathbb{E}\big[V^{\star,1}(\rho) - V^{\pi_t,1}(\rho)\big] &\le \frac{\Phi_0 - \mathbb{E}[\Phi_T]}{\eta} + \frac{9\eta T}{8} \sum_{h=1}^{H} L_h^2 \\ &\le \frac{\Phi_0}{\eta} + \frac{9\eta T}{8} \sum_{h=1}^{H} L_h^2, \end{aligned}$$

where the last inequality uses nonnegativity of KL divergence. Under uniform initialization,

$$\Psi_0^h(s) = \log A + \sum_{a \in \mathcal{A}} \pi^{\star,h}(a \mid s) \log \pi^{\star,h}(a \mid s) \le \log A,$$

with the convention $0 \log 0 = 0$. Since each visitation distribution has total mass one, $\Phi_0 \le H \log A$.

The output index $\widehat{T}$ is uniform on $\{0, \dots, T-1\}$ and independent of the randomness generating the iterates. Hence,

$$\mathbb{E}\big[V^{\star,1}(\rho) - V^{\pi_{\widehat{T}},1}(\rho)\big] = \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\big[V^{\star,1}(\rho) - V^{\pi_t,1}(\rho)\big].$$

Dividing the preceding bound by T and using

$$\sum_{h=1}^{H} L_h^2 = \frac{H(H+1)(2H+1)}{6}$$

gives

$$\mathbb{E}\big[V^{\star,1}(\rho) - V^{\pi_{\widehat{T}},1}(\rho)\big] \le \frac{H \log A}{\eta T} + \frac{3\eta H(H+1)(2H+1)}{16},$$

which proves equation 13. □

Proof of Corollary 1. The step size in equation 14 makes the two terms in equation 13 equal. Substitution gives

$$\mathbb{E}\big[V^{\star,1}(\rho) - V^{\pi_{\widehat{T}},1}(\rho)\big] \le \sqrt{\frac{3H^2(H+1)(2H+1)\log A}{4T}}.$$

Define

$$x := \frac{3H^2(H+1)(2H+1)\log A}{4\varepsilon^2}.$$

Choosing $T = \lceil x \rceil$ therefore guarantees an expected optimality gap of at most ε.

With $m = 1$, the algorithm uses $N = SA(H-1)T$ transition queries. Since $H \ge 2$, $A \ge 2$, and $0 < \varepsilon < H$, we have $x > 1$ and hence $\lceil x \rceil \le 2x$. Therefore,

$$N = SA(H-1)\lceil x \rceil \le 2SA(H-1)x = O\left(\frac{SAH^5 \log A}{\varepsilon^2}\right),$$

proving the claimed transition-query bound. □

13 Analysis of backward induction with fresh Monte Carlo evaluation

Consider the instance $M \in \mathcal{M}_{K,A,H}$ from Section 4.2 with

$$\theta_i^h = 1, \qquad p^h = p := \frac{1}{4H}, \qquad \alpha^h = \alpha \in \left(0, \frac{1}{8H}\right], \qquad h \in [H-1],\ i \in [K].$$

Since $0 < \alpha < p$ and $p + \alpha \le 3/(8H) \le 1/(2H)$, these parameters satisfy equation 5. Thus, at every decision state and horizon $h < H$, action a_1 enters g with probability $p + \alpha$, whereas every other action enters g with probability $p - \alpha$.

We use the rollout notation and sigma-fields $\mathcal{F}_h$ defined in Section 6.1. Throughout this section, probabilities, expectations, and variances are evaluated on M; we suppress the instance subscript.

Lemma 10. *Fix $h \in [H-1]$, $i \in [K]$, and $a \in \mathcal{A}$, and write $L := H - h$. For every $\ell \in [m]$, almost surely, $0 \le G_\ell^h(s_i, a) \le L$,*

$$\mathbb{E}\big[G_\ell^h(s_i, a) \mid \mathcal{F}_h\big] = Q^{\pi_{0,h},h}(s_i, a), \tag{32}$$

and

$$\mathrm{Var}\big(G_\ell^h(s_i, a) \mid \mathcal{F}_h\big) \ge \frac{L^3}{96H}. \tag{33}$$

These statements hold uniformly over the continuation policy.

Proof. Condition on $\mathcal{F}_h$, so the continuation rules are fixed. The reward at horizon h is zero, and at most L subsequent unit rewards can be received. This proves the range bound. The rollout takes the prescribed initial action and then follows the continuation rules, so its conditional mean is $Q^{\pi_{0,h},h}(s_i, a)$.

For the variance bound, write $U := G_\ell^h(s_i, a)$. Number the remaining transitions by $j = 1, \dots, L$, with transition j occurring at horizon $h + j - 1$. If the first entry into g occurs at transition j, then $U = L - j + 1$; if no entry occurs, then $U = 0$.

Define

$$q_{\min} := \frac{1}{8H}, \qquad q_{\max} := \frac{3}{8H}.$$

Let β_j be the conditional probability of entering g at transition j, given $\mathcal{F}_h$ and no earlier entry. Before entry, the process is at a decision state. Each available action has entry probability between $p - \alpha$ and $p + \alpha$, so

$$q_{\min} \le \beta_j \le q_{\max}.$$

The probability chain rule therefore gives

$$\mathbb{P}(U = L - j + 1 \mid \mathcal{F}_h) = \beta_j \prod_{r=1}^{j-1}(1 - \beta_r),$$

$$\mathbb{P}(U = 0 \mid \mathcal{F}_h) = \prod_{r=1}^{L}(1 - \beta_r).$$

Since $L \le H - 1$, Bernoulli's inequality yields

$$(1 - q_{\max})^L \ge 1 - L q_{\max} \ge 1 - \frac{3(H-1)}{8H} \ge \frac{5}{8}.$$

Consequently,

$$\mathbb{P}(U = 0 \mid \mathcal{F}_h) \ge \frac{5}{8}, \tag{34}$$

and, for every $k \in [L]$,

$$\mathbb{P}(U = k \mid \mathcal{F}_h) \ge q_{\min}(1 - q_{\max})^L \ge \frac{5}{64H} \ge \frac{1}{16H}.$$

It follows that

$$\mathbb{E}[U^2 \mid \mathcal{F}_h] \geq \frac{1}{16H} \sum_{k=1}^{L} k^2.$$

By conditional Cauchy–Schwarz,

$$\begin{aligned}\big(\mathbb{E}[U \mid \mathcal{F}_h]\big)^2 &= \big(\mathbb{E}[U\mathbf{1}\{U>0\} \mid \mathcal{F}_h]\big)^2 \\ &\leq \mathbb{E}[U^2 \mid \mathcal{F}_h]\,\mathbb{P}(U>0 \mid \mathcal{F}_h).\end{aligned}$$

Combining the preceding bounds gives

$$\begin{aligned}\mathrm{Var}(U \mid \mathcal{F}_h) &\geq \mathbb{E}[U^2 \mid \mathcal{F}_h]\,\mathbb{P}(U=0 \mid \mathcal{F}_h) \\ &\geq \frac{1}{2}\mathbb{E}[U^2 \mid \mathcal{F}_h] \\ &\geq \frac{1}{32H}\sum_{k=1}^{L} k^2 \geq \frac{L^3}{96H},\end{aligned}$$

where the last inequality uses $\sum_{k=1}^{L} k^2 = L(L+1)(2L+1)/6 \geq L^3/3$. The argument uses only the entry-probability bounds and therefore holds for every continuation policy. □

Lemma 11. *There exist universal constants $c_0, c_1 > 0$ with the following property. Fix $i \in [K]$ and $h \in [H-1]$ such that $L := H - h \geq H/2$, and define*

$$D^h(i) := \widehat{Q}^{\pi_{0,h},h}(s_i, a_1) - \widehat{Q}^{\pi_{0,h},h}(s_i, a_2),$$

where the estimates use independent batches of m rollouts. If $\alpha \leq c_1/\sqrt{m}$, then, almost surely,

$$\mathbb{P}\big(D^h(i) < 0 \mid \mathcal{F}_h\big) \geq c_0. \tag{35}$$

The bound holds uniformly over the continuation policy.

Proof. Condition on $\mathcal{F}_h$ and suppress the conditioning throughout the proof. For $\ell \in [m]$, denote the rollout returns for a_1 and a_2 by $G_{\ell,1}$ and $G_{\ell,2}$, respectively, and define

$$Y_\ell := (G_{\ell,1} - \mathbb{E}G_{\ell,1}) - (G_{\ell,2} - \mathbb{E}G_{\ell,2}), \qquad Z := \frac{1}{m}\sum_{\ell=1}^{m} Y_\ell.$$

The variables $Y_1, \ldots, Y_m$ are independent and identically distributed, with mean zero. By equation 32 and equation 20, $D^h(i) = Z + \delta$, where

$$\delta := \mathbb{E}[D^h(i)] = 2\alpha\left[L - \overline{V}^{\pi_{0,h},h+1}\right], \qquad 0 < \delta \leq 2\alpha L.$$

Define the standard deviations

$$\sigma_Y := \big(\mathbb{E}[Y_1^2]\big)^{1/2}, \qquad \sigma_Z := \big(\mathbb{E}[Z^2]\big)^{1/2} = \frac{\sigma_Y}{\sqrt{m}}.$$

Independence of the two batches and Lemma 10 imply

$$\sigma_Y^2 = \mathrm{Var}(G_{1,1}) + \mathrm{Var}(G_{1,2}) \geq \frac{L^3}{48H}.$$

Since $L \geq H/2$,

$$\sigma_Z^2 \geq \frac{L^3}{48Hm} \geq \frac{L^2}{96m}.$$

In particular, both standard deviations are positive.

Each return lies in $[0, L]$, so $|Y_\ell| \leq 2L$ and $\mathbb{E}[Y_1^4] \leq 4L^2\sigma_Y^2$. Expanding the fourth moment of the sum gives

$$\mathbb{E}[Z^4] = \frac{\mathbb{E}[Y_1^4]}{m^3} + \frac{3(m-1)}{m^3}\sigma_Y^4.$$

Therefore,

$$\begin{aligned}\frac{\mathbb{E}[Z^4]}{\sigma_Z^4} &= \frac{\mathbb{E}[Y_1^4]}{m\sigma_Y^4} + \frac{3(m-1)}{m} \\ &\le 3 + \frac{4L^2}{m\sigma_Y^2} \\ &\le 3 + \frac{192H}{mL} \le 3 + \frac{384}{m} \le 387.\end{aligned}$$

Set $C := 387$, so $\mathbb{E}[Z^4] \le C\sigma_Z^4$.

Hölder's inequality gives

$$\mathbb{E}[Z^2] \le \big(\mathbb{E}|Z|\big)^{2/3}\big(\mathbb{E}[Z^4]\big)^{1/3}.$$

Rearranging and applying the fourth-moment bound yields

$$\mathbb{E}|Z| \ge \frac{(\mathbb{E}[Z^2])^{3/2}}{(\mathbb{E}[Z^4])^{1/2}} \ge \frac{\sigma_Z}{\sqrt{C}}.$$

Define $Z^- := \max\{-Z, 0\}$ and $\mu_- := \mathbb{E}[Z^-]$. Since $\mathbb{E}Z = 0$,

$$\mu_- = \frac{1}{2}\mathbb{E}|Z| \ge \frac{\sigma_Z}{2\sqrt{C}}.$$

Let $E := \{Z^- > \mu_-/2\}$. By Cauchy–Schwarz,

$$\begin{aligned}\mu_- &= \mathbb{E}[Z^-\mathbf{1}_{E^c}] + \mathbb{E}[Z^-\mathbf{1}_E] \\ &\le \frac{\mu_-}{2} + \sqrt{\mathbb{E}[(Z^-)^2]\,\mathbb{P}(E)}.\end{aligned}$$

Using $\mathbb{E}[(Z^-)^2] \le \sigma_Z^2$, we obtain

$$\mathbb{P}(E) \ge \frac{\mu_-^2}{4\mathbb{E}[(Z^-)^2]} \ge \frac{1}{16C}.$$

Moreover, on E,

$$Z < -\frac{\mu_-}{2} \le -\frac{\sigma_Z}{4\sqrt{C}}.$$

Choose

$$c_0 := \frac{1}{16C}, \qquad c_1 := \frac{1}{8\sqrt{96C}}.$$

If $\alpha \le c_1/\sqrt{m}$, then

$$\delta \le 2\alpha L \le \frac{2c_1 L}{\sqrt{m}} = \frac{L}{4\sqrt{96Cm}} \le \frac{\sigma_Z}{4\sqrt{C}}.$$

Thus, $D^h(i) = Z + \delta < 0$ on E, and $\mathbb{P}(E) \ge c_0$. Restoring the conditioning proves equation 35. All bounds are independent of the continuation policy. □

Proof of Theorem 3. For the instance specified at the beginning of this section, choose

$$\alpha := \min\left\{\frac{1}{8H}, \frac{c_1}{\sqrt{m}}\right\},$$

where c_1 is the constant constructed in the proof of Lemma 11.

Fix $i \in [K]$ and $h \in \{1, \dots, \lfloor H/2 \rfloor\}$. Then $H - h \ge H/2$, so Lemma 11 applies. On $\{D^h(i) < 0\}$, action a_2 has a strictly larger estimate than a_1. Hence, a_1 is excluded from the maximizing set under every tie-breaking ordering. The decision rule chosen at horizon h is retained in the returned policy, so

$$w_{\pi_1}^h(i) = 1 - \pi_1^h(a_1 \mid s_i) = 1 \qquad \text{on } \{D^h(i) < 0\}.$$

Consequently, almost surely,

$$\mathbb{E}\big[w_{\pi_1}^h(i) \mid \mathcal{F}_h\big] \ge \mathbb{P}\big(D^h(i) < 0 \mid \mathcal{F}_h\big) \ge c_0.$$

Taking expectations and averaging over i gives

$$\mathbb{E}[\overline{w}^h_{\pi_1}] \geq c_0, \qquad h \in \{1, \ldots, \lfloor H/2 \rfloor\}.$$

Apply equation 22 to each realization of π_1, with $\alpha^h = \alpha$. Since all summands are nonnegative,

$$\begin{aligned}\mathbb{E}\big[V^{\star,1}(\rho) - V^{\pi_1,1}(\rho)\big] &\geq \frac{3\alpha}{4} \sum_{h=1}^{\lfloor H/2 \rfloor} (H-h)\, \mathbb{E}[\overline{w}^h_{\pi_1}] \\ &\geq \frac{3\alpha c_0}{4} \sum_{h=1}^{\lfloor H/2 \rfloor} (H-h).\end{aligned}$$

For $H \geq 2$, we have $\lfloor H/2 \rfloor \geq H/4$ and $H - h \geq H/2$ at every retained horizon. Therefore,

$$\sum_{h=1}^{\lfloor H/2 \rfloor} (H-h) \geq \frac{H^2}{8},$$

and hence

$$\mathbb{E}\big[V^{\star,1}(\rho) - V^{\pi_1,1}(\rho)\big] \geq \frac{3c_0}{32}\alpha H^2.$$

Since $c_1 \leq 1/8$, the choice of α satisfies

$$\alpha \geq c_1 \min\left\{\frac{1}{H}, \frac{1}{\sqrt{m}}\right\}.$$

Define $c_{\mathrm{BI}} := 3c_0c_1/32 > 0$. It follows that

$$\mathbb{E}\big[V^{\star,1}(\rho) - V^{\pi_1,1}(\rho)\big] \geq c_{\mathrm{BI}} \min\left\{H, \frac{H^2}{\sqrt{m}}\right\},$$

which proves equation 17.

Finally, suppose Algorithm 2 satisfies equation 9 with $\pi_{\mathrm{out}} = \pi_1$ for some $0 < \varepsilon < c_{\mathrm{BI}}H$. Applying the guarantee to the constructed instance gives

$$c_{\mathrm{BI}} \min\left\{H, \frac{H^2}{\sqrt{m}}\right\} \leq \varepsilon < c_{\mathrm{BI}}H.$$

The minimum must equal its second argument, so $c_{\mathrm{BI}}H^2/\sqrt{m} \leq \varepsilon$. Squaring and rearranging yields

$$m \geq \frac{c_{\mathrm{BI}}^2 H^4}{\varepsilon^2}.$$

If every rollout transition is queried, then equation 16 and $H(H-1)/2 \geq H^2/4$ imply

$$N = KAm\,\frac{H(H-1)}{2} \geq \frac{KAmH^2}{4} \geq \frac{c_{\mathrm{BI}}^2}{4}\frac{KAH^6}{\varepsilon^2}.$$

This proves the transition-query lower bound. □

14 Additional Simulations

14.1 Variability across sampling runs

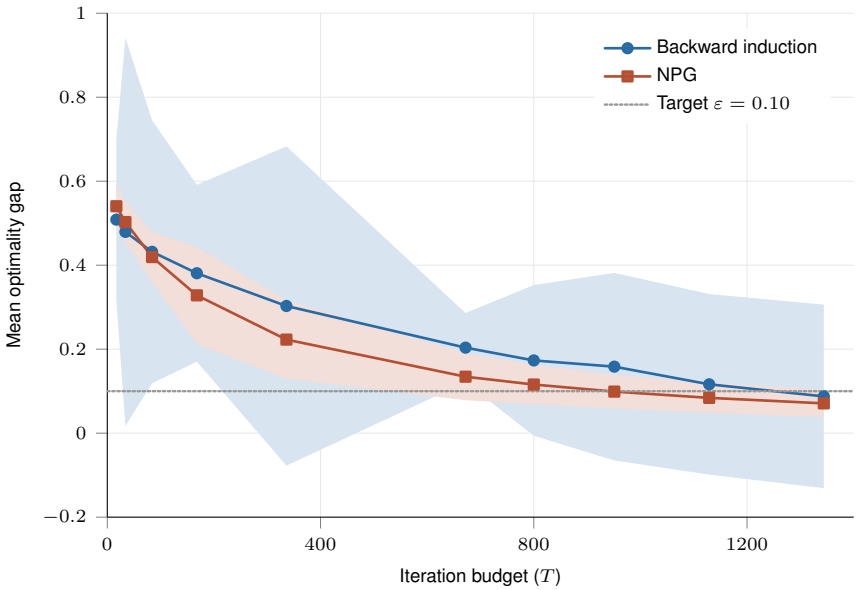


Figure 2: Sampling variability on the fixed MDP with $S = 25$, $A = 8$, and $H = 7$. Curves show five-run mean gaps for backward induction with fresh Monte Carlo evaluation and NPG with one-step evaluation, using $\eta = (\log A)/H$ for NPG. Shading shows the mean ± 10 sample standard deviations. The factor 10 magnifies variability for display; these bands are not confidence intervals. Both axes are linear, and the complete bands are shown. Here, T counts NPG updates; each backward induction point represents one complete sweep at the paired query budget.

We retain the MDP, procedures, step size, budget grid, and performance criteria of Section 7. Figure 2 uses the first five of the 500 independently seeded runs. At each budget, we recompute the mean and sample standard deviation from these five runs, using denominator 4 for the sample variance.

NPG has the smaller sample standard deviation at every displayed budget, consistent with the standard deviations across all 500 runs reported with the main figure. The backward induction band extends below zero at five of the ten budgets. These negative endpoints arise from the symmetric band construction; every observed optimality gap is nonnegative.