Sample Complexity Reduction via Policy Difference Estimation in Tabular Reinforcement Learning

Adhyyan Narang
University of Washington
adhyyan@uw.edu

Andrew Wagenmaker
University of California, Berkeley
ajwagen@berkeley.edu

Lillian J. Ratliff
University of Washington
ratliffl@uw.edu

Kevin Jamieson
University of Washington
jamieson@cs.washington.edu

Abstract

In this paper, we study the non-asymptotic sample complexity for the pure exploration problem in contextual bandits and tabular reinforcement learning (RL): identifying an ϵ-optimal policy from a set of policies Π with high probability. Existing work in bandits has shown that it is possible to identify the best policy by estimating only the *difference* between the behaviors of individual policies– which can be substantially cheaper than estimating the behavior of each policy directly —yet the best-known complexities in RL fail to take advantage of this, and instead estimate the behavior of each policy directly. Does it suffice to estimate only the differences in the behaviors of policies in RL? We answer this question positively for contextual bandits, but in the negative for tabular RL, showing a separation between contextual bandits and RL. However, inspired by this, we show that it *almost* suffices to estimate only the differences in RL: if we can estimate the behavior of a *single* reference policy, it suffices to only estimate how any other policy deviates from this reference policy. We develop an algorithm which instantiates this principle and obtains, to the best of our knowledge, the tightest known bound on the sample complexity of tabular RL.

1 Introduction

Online platforms, such as AirBnB, often try to improve their services by A/B testing different marketing strategies. Based on the inventory, their strategy could include emphasizing local listings versus tourist destinations, providing discounts for longer stays, or de-prioritizing homes that have low ratings. In order to choose the best strategy, the standard approach would be to apply each strategy sequentially and measure outcomes. However, recognize that the choice of strategy (policy) affects the future inventory (state) of the platform. This complex interaction between different strategies makes it difficult to estimate the impact of any strategy, if it were to be applied independently. To address this, we can model the platform as an Markov Decision Process (MDP) with an observed state [17, 15] and a finite set of policies Π corresponding to possible strategies. We wish to collect data by playing *exploratory* actions which will enable us to estimate the true value of each policy $\pi \in \Pi$, and identify the best policy from Π as quickly as possible.

In addition to A/B testing, similar challenges arise in complex medical trials, learning robot policies to pack totes, and autonomous navigation in unfamiliar environments. All of these problems can be formally modeled as the PAC (Probably Approximately Correct) policy identification problem in reinforcement learning (RL). An algorithm is said to be (ϵ, δ)-PAC if, given a set of policies Π, it returns a policy $\pi \in \Pi$ that performs within ϵ of the optimal policy in Π, with probability $1 - \delta$. The

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-0717

goal is to satisfy this condition whilst minimizing the number of interactions with the environment (the *sample complexity*).

Traditionally, prior work has aimed to obtain *minimax* or *worst-case* guarantees for this problem—guarantees that hold across *all* environments within a problem class. Such worst-case guarantees typically scale with the "size" of the environment, for example, scaling as $\mathcal{O}(\mathrm{poly}(S, A, H)/\epsilon^2)$, for environments with S states, A actions, horizon H. While guarantees of this form quantify which classes of problems are efficiently learnable, they fail to characterize the difficulty of particular problem instances—producing the same complexity on both "easy" and "hard" problems that share the same "size". This is not simply a failure of analysis—recent work has shown that algorithms that achieve the minimax-optimal rate could be very suboptimal on particular problem instances [46]. Motivated by this, a variety of recent work has sought to obtain *instance-dependent* complexity measures that capture the hardness of learning each particular problem instance. However, despite progress in this direction, the question of the *optimal* instance-dependent complexity has remained elusive, even in tabular settings.

Towards achieving instance-optimality in RL, the key question is: *what aspects* of a given environment must be learned, in order to choose a near-optimal policy? In the simpler bandit setting, this question has been settled by showing that it is sufficient to learn the *differences* between values of actions rather than learning the value of each individual action: it is only important whether a given action's value is greater or lesser than that of other actions. This observation can yield significant improvements in sample efficiency [37, 16, 13, 30]. Precisely, the best-known complexity measures in the bandit setting scale as:

$$\inf_{\pi_{\mathrm{exp}}} \max_{\pi \in \Pi} \frac{\|\phi^{\pi} - \phi^{\star}\|^2_{\Lambda(\pi_{\mathrm{exp}})^{-1}}}{\Delta(\pi)^2}, \tag{1.1}$$

where ϕ^{π} is the feature vector of action π, $\phi^{\star}$ the feature vector of the optimal action, $\Delta(\pi)$ is the suboptimality of action π. Here, $\Lambda(\pi_{\mathrm{exp}})$ are the covariates induced by π_{exp}, our distribution of exploratory actions. The denominator of this expression measures the performance gap between action π and the optimal action. The numerator measures the variance of the estimated (from data collected by π_{exp}) difference in values between $(\pi, \pi^{\star})$. The max over actions follows because to choose the best action, we have to rule out every sub-optimal action from the set of candidates Π; the infimum optimizes over data collection strategies.

In contrast, in RL, instead of estimating the difference between policy values *directly*, the best known algorithms simply estimate the value of each individual policy *separately* and then take the difference. This obtains instance-dependent complexities which scale as follows [42]:

$$\sum_{h=1}^{H} \inf_{\pi_{\mathrm{exp}}} \max_{\pi \in \Pi} \frac{\|\phi^{\pi}_h\|^2_{\Lambda_h(\pi_{\mathrm{exp}})^{-1}} + \|\phi^{\star}_h\|^2_{\Lambda_h(\pi_{\mathrm{exp}})^{-1}}}{\Delta(\pi)^2} \tag{1.2}$$

where ϕ^{π}_h is the state-action visitation of policy π at step h. Since now the difference is calculated *after* estimation, the variance of the difference is the sum of the individual variances of the estimates of each policy, captured in the numerator of (1.2). Comparing the numerator of (1.2) to that of (1.1) begs the question: in RL can we estimate the *difference* of policies directly to reduce the sample complexity of RL?

To motivate why this distinction is important, consider the tabular MDP example of Figure 1. In this example, the agent starts in state s_1, takes one of three actions, and then transitions to one of states s_2, s_3, s_4. Consider the policy set $\Pi = \{\pi_1, \pi_2\}$, where π_1 always plays action a_1, and π_2 is identical, except plays actions a_2 in the red states. If $\phi^{\pi_i}_h \in \triangle_{\mathcal{S}\times\mathcal{A}}$ denotes the state-action visitations of policy π_i at time $h = 1, 2$, then we see that $\phi^{\pi_1}_1 = \phi^{\pi_2}_1$ since π_1 and π_2 agree on the action in s_1. But $\phi^{\pi_1}_2 \neq \phi^{\pi_2}_2$ as their actions differ on the red states.

Since these red states will be reached with probability at most 3ϵ, the norm of the *difference*

$$\|\phi^{\pi}_2 - \phi^{\star}_2\|^2_{\Lambda_2(\pi_{\mathrm{exp}})^{-1}} = \sum_{s,a} \frac{(\phi^{\pi}_2(s,a) - \phi^{\star}_2(s,a))^2}{\phi^{\pi_{\mathrm{exp}}}_2(s,a)}$$

is significantly less than the sum of the individual norms

$$\|\phi^{\pi}_2\|^2_{\Lambda_2(\pi_{\mathrm{exp}})^{-1}} + \|\phi^{\star}_2\|^2_{\Lambda_2(\pi_{\mathrm{exp}})^{-1}} = \sum_{s,a} \frac{\phi^{\pi}_2(s,a)^2 + \phi^{\star}(s,a)^2}{\phi^{\pi_{\mathrm{exp}}}_2(s,a)}.$$

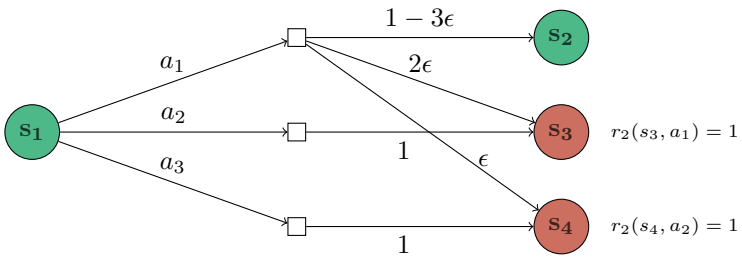

Figure 1: A motivating example for differences. The rewards for all actions other than the ones specified in the figure are 0. Define policy set $\Pi = \{\pi_1, \pi_2\}$ so that π_1 always plays a_1, whereas π_2 plays a_1 on green states but a_2 on red states. The difference of their state-action visitation probabilities is only non-zero in states s_3, s_4 and are just $O(\epsilon)$ apart.

Intuitively, to minimize differences π_{exp} can explore just states s_3, s_4 where the policies differ, whereas minimizing the individual norms requires wasting lots of energy in state s_2 where the two policies and the difference is zero. Formally:

Proposition 1. *On the MDP and policy set* Π *from Figure 1, we have that*

$$\inf_{\pi_{\text{exp}}} \max_{\pi\in\Pi} \|\phi_2^\pi\|^2_{\Lambda_2(\pi_{\text{exp}})^{-1}} \geq 1 \quad \textit{and} \quad \inf_{\pi_{\text{exp}}} \max_{\pi\in\Pi} \|\phi_2^\star - \phi_2^\pi\|^2_{\Lambda_2(\pi_{\text{exp}})^{-1}} \leq 15\epsilon^2.$$

Proposition 1 shows that indeed, the complexity of the form Equation (1.1) (generalized to RL) in terms of differences could be significantly tighter than Equation (1.2); in this case, it is a factor of ϵ^2 better. But achieving a sample complexity that depends on the differences requires more than just a better analysis: it requires a new estimator and an algorithm to exploit it.

Contributions. In this work, we aim to understand whether such a complexity is achievable in RL. Letting ρ_Π denote the generalization of (1.1) to the RL case—that is, (1.2) but with $\|\phi_h^\pi\|^2_{\Lambda_h(\pi_{\text{exp}})^{-1}}$ replaced by $\|\phi_h^\pi - \phi_h^{\pi^\star}\|^2_{\Lambda_h(\pi_{\text{exp}})^{-1}}$, our contributions are as follows:

1. In the Tabular RL case, [2] recently showed that ρ_Π is a lower bound on the sample complexity of RL by characterizing the difficulty of learning the unknown reward function; however, they did not resolve whether it is achievable when the state-transitions are unknown as well. We provide a lower bound which demonstrates that $\mathcal{O}(\rho_\Pi)$ is *not* sufficient for learning with state transitions.
2. We provide an algorithm PERP, which first learns the behavior a particular reference policy $\bar{\pi}$, and then estimates the difference in behavior between $\bar{\pi}$ and every other policy π, rather than estimating the behavior of each π directly.
3. In the case of tabular RL, we show that PERP obtains a complexity that scales with $\mathcal{O}(\rho_\Pi)$, in addition to an extra term which measures the cost of learning the behavior of the reference policy $\bar{\pi}$. We argue that this additional term is critical to achieving instance-optimal guarantees in RL, and that PERP leads to improved complexities over existing work.
4. In the contextual bandit setting, we provide an upper bound that scales (up to lower order terms) as $\mathcal{O}(\rho_\Pi)$ for the *unknown-context* distribution case. This matches the lower bound from [30] for the known context distribution case, thus showing that ρ_Π is necessary and sufficient in contextual bandits even when the context distribution is unknown. Hence, we observe a qualitative information-theoretic separation between contextual bandits and RL.

The key insight from our work is that it does not suffice to *only* learn the differences between policy values in RL, but it *almost* suffices to—if we can learn how a single policy behaves, it suffices to learn the difference between this policy and every other policy.

2 Related Work

The reinforcement learning literature is vast, and here we focus on results in tabular RL and instance-dependent guarantees in RL.

https://doi.org/10.52202/079017-0717

Minimax Guarantees Tabular RL. Finite-time minimax-style results on policy identification in tabular MDPs go back to at least the late 90s and early 2000s [24, 26, 25, 8, 21]. This early work was built upon and refined by a variety of other works over the following decade [38, 4, 34, 39], leading up to works such as [28, 9], which establish sample complexity bounds of $\mathcal{O}(S^2A \cdot \text{poly}(H)/\epsilon^2)$. More recently, [10, 11, 33] have proposed algorithms which achieve the optimal dependence of $\mathcal{O}(SA \cdot \text{poly}(H)/\epsilon^2)$, with [11, 33] also achieving the optimal H dependence. The question of regret minimization is intimately related to that of policy identification—any low-regret algorithm can be used to obtain a near-optimal policy via an online-to-batch conversion [19]. Early examples of low-regret algorithms in tabular MDPs are [3, 4, 5, 48], with more recent works removing the horizon dependence or achieving the optimal lower-order terms as well [50, 51]. Recently, [6, 7] provide minimax guarantees in the multi-task RL setting as well.

Instance-Dependence in RL. While the problem of obtaining worst-case optimal guarantees in tabular RL is nearly closed, we are only beginning to understand what types of instance-dependent guarantees are possible. In the setting of regret minimization, [35, 14] achieve instance-optimal regret for tabular RL asymptotically. Simchowitz and Jamieson [36] show that standard optimistic algorithms achieve regret bounded as $\mathcal{O}(\sum_{s,a,h} \frac{\log K}{\Delta_h(s,a)})$, a result later refined by [47, 12]. In settings of RL with linear function approximation, several works achieve instance-dependent regret guarantees [18, 44]. Recently, Wagenmaker and Foster [45] achieved finite-time guarantees on instance-optimal regret in general decision-making settings, a setting encompassing much of RL.

On the policy identification side, early works obtaining instance-dependent guarantees for tabular MDPs include [49, 20, 31, 32], but they all exhibit shortcomings such as requiring access to a generative model or lacking finite-time results. The work of Wagenmaker et al. [46] achieves a finite-time instance-dependent guarantee for tabular RL, introducing a new notion of complexity, the *gap-visitation complexity*. In the special case of deterministic, tabular MDPs, Tirinzoni et al. [41] show matching finite-time instance-dependent upper and lower bounds. For RL with linear function approximation, [42, 43] achieve instance-dependent guarantees on policy identification, in particular, the complexity given in (1.2), and propose an algorithm, PEDEL, which directly inspires our algorithmic approach. On the lower bound side, Al-Marjani et al. [2] show that ρ_Π is necessary for tabular RL, but fail to close the aforementioned gap between ρ_Π and (1.2). We will show instead that this gap is real and both the lower bound of Al-Marjani et al. [2] and upper bound of Wagenmaker and Jamieson [42] are loose.

Several works on linear and contextual bandits are also relevant. In the seminal work, [37] posed the best-arm identification problem for linear bandits and beautifully argued—without proof—that estimating differences were crucial and that (1.1) ought to be the true sample complexity of the problem. Over time, this conjecture was affirmed and generalized [16, 13, 22]. This improved understanding of pure-exploration directly led to instance-dependent optimal linear bandit algorithms for regret [29, 27]. More recently, contextual bandits have also been given a similar treatment [40, 30].

3 Preliminaries and Problem Setting

Let $\|x\|_\Lambda^2 = x^\top \Lambda x$ for any (x, Λ). We let $\mathbb{E}_\pi$ denote the probability measure induced by playing policy π in our MDP.

Tabular Markov Decision Processes. We study episodic, finite-horizon, time inhomogenous and tabular Markov Decision Processes (MDPs), denoted by the tuple $(\mathcal{S}, \mathcal{A}, H, \{P_h\}_{h=1}^H, \{\nu_h\}_{h=1}^H)$ where the state space $\mathcal{S}$ and action space $\mathcal{A}$ are finite, H is the horizon, $P_h \in \mathbb{R}^{S \times SA}$ denote the transition matrix at stage h where $[P_h]_{s',sa} = \mathbb{P}(s_{h+1} = s' | s_h = s, a_h = a)$, and $\nu_h(s,a) \in \triangle_{[0,1]}$ denote the distribution over reward at stage h when the state of the system is s and action a is chosen. Let $r_h(s,a)$ be the expectation of a reward drawn from $\nu_h(s,a)$. We assume that every episode starts in state s_1, and that ν_h and P_h are initially unknown and must be estimated over time.

Let $\pi = \{\pi_h\}_{h=1}^H$ denote a policy mapping states to actions, so that $\pi_h(s) \in \triangle_{\mathcal{A}}$ denotes the distribution over actions for the policy at (s,h); when the policy is deterministic, $\pi_h(s) \in \mathcal{A}$ outputs a single action. An episode begins in state s_1, the agent takes action $a_1 \sim \pi_1(s_1)$ and receives reward $R_1 \sim \nu_1(s_1, a_1)$ with expectation $r_1(s_1, a_1)$; the environment transitions to state $s_2 \sim P_h(s_1, a_1)$. The process repeats until timestep H, at which point the episode ends and the agent returns to state s_1. Let $V_h^\pi(s) = \mathbb{E}_\pi[\sum_{h'=h}^H r_{h'}(s_{h'}, a_{h'}) | s_h = s]$, V_0^π the total expected reward, $V_0^\pi := V_1^\pi(s_0)$,

and $Q_h^\pi(s,a) = \mathbb{E}_\pi[\sum_{h'=h}^H r_{h'}(s_{h'}, a_{h'})|s_h = s, a_h = a]$ the amount of reward we expect to collect if we are in state s at step h, play action a and then play policy π for the remainder of the episode. Note that we can understand these functions as S and SA-dimensional vectors respectively. We use $V^\pi = V_0^\pi$ when clear from context.

We call $w_h^\pi \in \triangle_S$ the *state visitation vector* at step h for policy π, so that $w_h^\pi(s)$ captures the probability that policy π would land in state s at step h during an episode. Let $\boldsymbol{\pi}_h \in \mathbb{R}^{SA \times S}$ denote the policy matrix for policy π, that maps states to state-actions as follows

$$[\boldsymbol{\pi}_h]_{(s,a),s'} = \mathbb{I}(s = s')[\pi_h(s)]_a.$$

Denote $\phi_h^\pi \in \triangle_{SA}$ as $\phi_h^\pi := \boldsymbol{\pi}_h w_h^\pi$ as the *state-action visitation vector*: $\phi_h^\pi(s,a)$ measures the the probability that policy π would land in state s and play action a at step h during an episode. From these definitions, it follows that $[P_h \phi_h^\pi]_s = [P_h \boldsymbol{\pi}_h w_h^\pi]_s = w_{h+1}^\pi(s)$. For policy π, denote the covariance matrix at timestep h as $\Lambda_h(\pi) = \sum_{s,a} \phi_h^\pi(s,a) \mathbf{e}_{(s,a)} \mathbf{e}_{(s,a)}^\top$.

(ϵ, δ)-PAC Best Policy Identification. For a collection of policies Π, define $\pi^\star := \arg\max_{\pi \in \Pi} V^\pi$ as the optimal policy, $V^\star$ its value, and $\phi_h^\star$ as its state-action visitation vector. Let $\Delta_{\min} := \min_{\pi \in \Pi \setminus \{\pi^\star\}} V^\star - V^\pi$ in the case when $\pi^\star$ is unique, and otherwise $\Delta_{\min} := 0$. Define $\Delta(\pi) := \max\{V^\star - V^\pi, \Delta_{\min}\}$. Given $\epsilon \geq 0$, $\delta \in (0,1)$ an algorithm is said to be (ϵ, δ)-PAC if at a stopping time τ of its choosing, it returns a policy $\widehat{\pi}$ which satisfies $\Delta(\pi) \leq \epsilon$ with probability $1 - \delta$. Our goal is to obtain an (ϵ, δ)-PAC algorithm that minimizes τ. A fundamental complexity measure used throughout this work is defined as

$$\rho_\Pi := \sum_{h=1}^H \inf_{\pi_{\exp}} \max_{\pi \in \Pi} \frac{\|\phi_h^\star - \phi_h^\pi\|^2_{\Lambda_h(\pi_{\exp})^{-1}}}{\max\{\epsilon^2, \Delta(\pi)^2\}} \quad \text{for} \quad \|\phi_h^\star - \phi_h^\pi\|^2_{\Lambda_h(\pi_{\exp})^{-1}} := \sum_{s,a} \frac{(\phi_h^\star(s,a) - \phi_h^\pi(s,a))^2}{\phi_h^{\pi_{\exp}}(s,a)}$$

where the infimum is over all exploration policies $\pi_{\exp}$ (not necessarily just those in Π). Recall that for $\epsilon = 0$, [2] showed any (ϵ, δ)-PAC algorithm satisfies $\mathbb{E}[\tau] \geq \rho_\Pi \log(\frac{1}{2.4\delta})$.

4 What is the Sample Complexity of Tabular RL?

In this section, we seek to understand the complexity of tabular RL. We start by showing that ρ_Π is not sufficient. We have the following result.

Lemma 1. *For the MDP $\mathcal{M}$ and policy set Π from Figure 1,*

1. $\sum_{h=1}^H \inf_{\pi_{\exp}} \max_{\pi \in \Pi} \frac{\|\phi_h^\star - \phi_h^\pi\|^2_{\Lambda_h(\pi_{\exp})^{-1}}}{\max\{\epsilon^2, \Delta(\pi)^2\}} \leq 15$,
2. *Any (ϵ, δ)-PAC algorithm must collect at least $\mathbb{E}^{\mathcal{M}}[\tau] \geq \frac{1}{\epsilon} \cdot \log \frac{1}{2.4\delta}$. samples.*

Where does the additional complexity arise on the instance of Figure 1? As described in the introduction, π_1 and π_2 differ only on the red states, and a complexity scaling as ρ_Π quantifies only the difficulty of distinguishing $\{\pi_1, \pi_2\}$ on these states. Note that on this example π_1 plays the optimal action in state s_3 and a suboptimal action in state s_4, and π_2 plays a suboptimal action in s_3 and the optimal action in s_4. The total reward of policy π_1 is therefore equal to the reward achieved at state s_3 times the probability it reaches state s_3, and the total reward of policy π_2 is the reward achieved at state s_4 times the probability it reaches state s_4. Here, ρ_Π would quantify the difficulty of learning the reward achieved at each state. However, it fails to quantify *the probability of reaching each state*, since this depends on the behavior at step 1, not step 2.

Thus, on this example, to determine whether π_1 or π_2 is optimal, we must pay some additional complexity to learn the outgoing transitions from the initial state, giving rise to the lower bound in Lemma 1. Inspecting the lower bound of [2], one realizes that the construction of this lower bound only quantifies the cost of learning the reward distributions $\{\nu_h\}_h$ and *not* the state transition matrices $\{P_h\}_h$. On examples such as Figure 1, this lower bound then does not quantify the cost of learning the probability of visiting each state, which we've argued is necessary. We therefore conclude that, while ρ_Π may be enough for learning the rewards, it is *not* sufficient for solving the full tabular RL problem. Our main algorithm builds on this intuition, and, in addition to estimating the rewards, aims to estimate where policies visit as efficiently as possible.

https://doi.org/10.52202/079017-0717

4.1 Main Result

If ρ_Π is not achievable as the sample complexity for Tabular RL, what is the best that we can do? In this section, we answer this question with our sample complexity bound; we later describe the algorithmic insights that enable us to achieve this result in the following section. First, for any $\pi, \bar{\pi} \in \Pi$, we define

$$U(\pi, \bar{\pi}) := \sum_{h=1}^{H} \mathbb{E}_{s_h \sim w_h^{\bar{\pi}}}[(Q_h^{\pi}(s_h, \pi_h(s_h)) - Q_h^{\pi}(s_h, \bar{\pi}_h(s_h)))^2]. \tag{4.1}$$

Now, we state our main result.

Theorem 1. *There exists an algorithm (Algorithm 1) which, with probability at least $1 - 2\delta$, finds an ϵ-optimal policy and terminates after collecting at most*

$$\sum_{h=1}^{H} \inf_{\pi_{\exp}} \max_{\pi \in \Pi} \frac{H^4 \|\phi_h^\star - \phi_h^\pi\|^2_{\Lambda_h(\pi_{\exp})^{-1}}}{\max\{\epsilon^2, \Delta(\pi)^2\}} \cdot \iota \beta^2 + \max_{\pi \in \Pi} \frac{HU(\pi, \pi^\star)}{\max\{\epsilon^2, \Delta(\pi)^2\}} \log \tfrac{H|\Pi|\iota}{\delta} + \frac{C_{\mathrm{poly}}}{\max\{\epsilon^{\frac{5}{3}}, \Delta_{\min}^{\frac{5}{3}}\}}$$

episodes, for $C_{\mathrm{poly}} := \mathrm{poly}(S, A, H, \log 1/\delta, \iota, \log|\Pi|)$, $\beta := C\sqrt{\log(\frac{SH|\Pi|}{\delta} \cdot \frac{1}{\Delta_{\min} \vee \epsilon})}$ and $\iota := \log \frac{1}{\Delta_{\min} \vee \epsilon}$.

Theorem 1 shows that, up to terms lower-order in ϵ and $\Delta_{\min}$, ρ_Π is almost sufficient, if we are willing to pay for an additional term scaling as $U(\pi, \pi^\star)/\Delta(\pi)^2$. Recognize the similarity of this term to the that from the performance difference lemma: if there were no square inside the expectation, the quantity $U(\pi, \pi^\star)$ would be equal to $\Delta(\pi)$. However, the square may change the scaling in some instances. Below, Lemma 2 shows that there exist settings where the complexity of Theorem 1 could be significantly tighter than Equation (1.2), the complexity achieved by the PEDEL algorithm of [42]. We revisit the instance from Figure 1 to show this; recall from Lemma 1 that the first term from Theorem 1 is a universal constant for this instance.

Lemma 2. *On MDP $\mathcal{M}$ and policy set Π from Figure 1, we have:*

1. $\max_{\pi \in \Pi} \frac{HU(\pi, \pi^\star)}{\max\{\epsilon^2, \Delta(\pi)^2\}} = \frac{3H}{\epsilon}$,
2. $\sum_{h=1}^{H} \inf_{\pi_{\exp}} \max_{\pi \in \Pi} \frac{\|\phi_h^\pi\|^2_{\Lambda_h(\pi_{\exp})^{-1}}}{\max\{\epsilon^2, \Delta(\pi)^2\}} \geq \frac{H}{\epsilon^2}$.

Furthermore, the complexity of Theorem 1 is never worse than Equation (1.2).

Lemma 3. *For any MDP instance and policy set Π, we have that*

$$\max\left\{\sum_{h=1}^{H} \inf_{\pi_{\exp}} \max_{\pi \in \Pi} \frac{\|\phi_h^\star - \phi_h^\pi\|^2_{\Lambda_h(\pi_{\exp})^{-1}}}{\max\{\epsilon^2, \Delta(\pi)^2\}}, \frac{HU(\pi, \pi^\star)}{\max\{\epsilon^2, \Delta(\pi)^2\}}\right\} \leq \sum_{h=1}^{H} \inf_{\pi_{\exp}} \max_{\pi \in \Pi} \frac{\|\phi_h^\pi\|^2_{\Lambda_h(\pi_{\exp})^{-1}}}{\max\{\epsilon^2, \Delta(\pi)^2\}}.$$

We briefly remark on the lower-order term for Theorem 1, $\frac{C_{\mathrm{poly}}}{\max\{\epsilon^{5/3}, \Delta_{\min}^{5/3}\}}$. Note that for small ϵ or $\Delta_{\min}$, this term will be dominated by the leading-order terms, which scale with $\min\{\epsilon^{-2}, \Delta_{\min}^{-2}\}$. While we make no claims on the tightness of this term, we note that recent work has shown that some lower-order terms are necessary for achieving instance-optimality [45].

4.2 The Main Algorithmic Insight: The Reduced-Variance Difference Estimator

In this section, we describe how we can estimate the difference between the values of policies directly, and provide intuition for why this results in the two main terms in Theorem 1. Fix any reference policy $\bar{\pi}$ and logging policy μ (neither are necessarily in Π). Here μ can be thought of as playing the role of $\pi_{\exp}$. Or, we can consider the A/B testing scenario from the introduction, where a policy μ is taking random actions and one wishes to perform off-policy estimation over some set of policies Π [17, 15]. For any $s \in \mathcal{S}$, we define

$$\delta_h^\pi(s) := w_h^\pi(s) - w_h^{\bar{\pi}}(s)$$

as the difference in state-visitations of policy π from reference policy $\bar{\pi}$, and $\delta_h^\pi \in \mathbb{R}^S$ as the vectorization of $\delta_h^\pi(s')$.

Policy selection rule. First, we describe our procedure of data collection and estimation. We collect $K_{\bar{\pi}}$ trajectories from $\bar{\pi}$ and K_μ trajectories from μ, and let $\{\widehat{w}_h^{\bar{\pi}}(s)\}_{s,h}$ denote the empirical state visitations from playing $\bar{\pi}$. From the data collected by playing μ, we construct estimates $\{\widehat{P}_h(s'|s,a)\}_{s,a,s',h}$ of the transition matrices. Note that $\widehat{w}_h^{\bar{\pi}}(s)$ simply counts visitations, so that $\mathbb{E}[(\widehat{w}_h^{\bar{\pi}}(s) - w_h^{\bar{\pi}}(s))^2] \le \frac{w_h^{\bar{\pi}}(s)}{K_{\bar{\pi}}}$ for all h, s. Define estimated state visitations for policy π in terms of deviations from $\bar{\pi}$ as $\widehat{w}_h^{\pi} := \widehat{w}_h^{\bar{\pi}} + \widehat{\delta}_h^{\pi}$. Here, $\widehat{\delta}_h^{\pi}$ is defined recursively as:

$$\widehat{\delta}_{h+1}^{\pi} := \widehat{P}_h \boldsymbol{\pi}_h \widehat{\delta}_h^{\pi} + \widehat{P}_h(\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_h)\widehat{w}_h^{\bar{\pi}}$$

Then, assuming, for simplicity, that rewards are known, we recommend the following policy:

$$\widehat{\pi} = \arg\max_{\pi \in \Pi} \widehat{D}^{\pi} \qquad \text{where} \qquad \widehat{D}^{\pi} := \textstyle\sum_{h=1}^{H} \langle r_h, \boldsymbol{\pi}_h \widehat{\delta}_h^{\pi} \rangle - \langle r_h, (\bar{\boldsymbol{\pi}}_h - \boldsymbol{\pi}_h)\widehat{w}_h^{\bar{\pi}} \rangle$$

Sufficient condition for ϵ-optimality. Here, we show that if

$$\forall \pi \in \Pi, \qquad |\widehat{D}^{\pi} - D^{\pi}| \le \frac{1}{3}\max\{\epsilon, \Delta(\pi)\} \tag{4.2}$$

then $\widehat{\pi}$ is ϵ-optimal. First, write the difference between values of policies π and $\bar{\pi}$ as:

$$\begin{aligned} D^{\pi} := V_0^{\pi} - V_0^{\bar{\pi}} &= \textstyle\sum_{h=1}^{H} \langle r_h, \boldsymbol{\pi}_h w_h^{\pi} \rangle - \sum_{h=1}^{H} \langle r_h, \bar{\boldsymbol{\pi}}_h w_h^{\bar{\pi}} \rangle \\ &= \textstyle\sum_{h=1}^{H} \langle r_h, \boldsymbol{\pi}_h \delta_h^{\pi} \rangle - \langle r_h, (\bar{\boldsymbol{\pi}}_h - \boldsymbol{\pi}_h) w_h^{\bar{\pi}} \rangle. \end{aligned} \tag{4.3}$$

Then, it is easy to verify that if $|\widehat{D}^{\pi} - D^{\pi}| \le 1/3\ \Delta(\pi)$, then $\widehat{D}^{\pi^\star} - \widehat{D}^{\pi} \ge 0$; hence, $\widehat{\pi} \ne \pi$. Hence, under Condition (4.2), either $\widehat{\pi} = \pi^\star$ or or $|\widehat{D}^{\pi} - D^{\pi}| \le \epsilon$. In the first case, clearly $\widehat{\pi}$ is ϵ-optimal. In the second case, we can add and subtract terms to write

$$V^\star - V^{\widehat{\pi}} \le |D^{\pi^\star} - \widehat{D}^{\pi^\star}| + \widehat{D}^{\pi^\star} - \widehat{D}^{\widehat{\pi}} + |\widehat{D}^{\widehat{\pi}} - D^{\widehat{\pi}}| \le \frac{2\epsilon}{3} + \hat{D}^{\pi^\star} - \hat{D}^{\widehat{\pi}} \le \frac{2\epsilon}{3}.$$

The last inequality follows since $\widehat{\pi}$ maximizes $\widehat{D}^{\pi}$. Hence, $\widehat{\pi}$ would be ϵ-optimal in this case as well.

Sample complexity. Now, we characterize how many samples must be collected from μ and $\bar{\pi}$ in order to meet Condition (4.2). After dropping some lower-order terms and unrolling the recursion (see Section A for details), we observe that

$$\begin{aligned} \widehat{\delta}_{h+1}^{\pi} - \delta_{h+1}^{\pi} &\approx (\widehat{P}_h - P_h)(\phi_h^{\pi} - \phi_h^{\bar{\pi}}) + P_h(\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_h)(\widehat{w}_h^{\bar{\pi}} - w_h^{\bar{\pi}}) + P_h \boldsymbol{\pi}_h (\widehat{\delta}_h^{\pi} - \delta_h^{\pi}) \\ &= \textstyle\sum_{k=0}^{h} \big(\prod_{j=k+1}^{h} P_j \boldsymbol{\pi}_j \big) \big((\widehat{P}_k - P_k)(\phi_k^{\pi} - \phi_k^{\bar{\pi}}) + P_k(\boldsymbol{\pi}_k - \bar{\boldsymbol{\pi}}_k)(\widehat{w}_k^{\bar{\pi}} - w_k^{\bar{\pi}}) \big). \end{aligned}$$

After manipulating this expression a bit more, we observe that

$$\sum_{h=1}^{H} \langle r_h, \boldsymbol{\pi}_h (\widehat{\delta}_h^{\pi} - \delta_h^{\pi}) \rangle = \sum_{k=0}^{H-1} \langle V_{k+1}^{\pi}, (\widehat{P}_k - P_k)(\phi_k^{\pi} - \phi_k^{\bar{\pi}}) + P_k(\boldsymbol{\pi}_k - \bar{\boldsymbol{\pi}}_k)(\widehat{w}_k^{\bar{\pi}} - w_k^{\bar{\pi}}) \rangle$$

Recognizing $Q_h^{\pi} = r_h + P_h^\top V_{h+1}^{\pi}$,

$$\begin{aligned} |\widehat{D}^{\pi} - D^{\pi}| &= \left| \sum_{h=1}^{H} \langle r_h, \boldsymbol{\pi}_h (\widehat{\delta}_h^{\pi} - \delta_h^{\pi}) \rangle + \langle r_h, (\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_h)(\widehat{w}_h^{\bar{\pi}} - w_h^{\bar{\pi}}) \rangle \right| \\ &= \left| \sum_{h=0}^{H-1} \langle V_{h+1}^{\pi}, (\widehat{P}_h - P_h)(\phi_h^{\pi} - \phi_h^{\bar{\pi}}) \rangle + \langle r_h + P_h^\top V_{h+1}^{\pi}, (\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_h)(\widehat{w}_h^{\bar{\pi}} - w_h^{\bar{\pi}}) \rangle \right| \end{aligned}$$

We can bound this as:

$$\begin{aligned} &\lesssim \sqrt{H^2 \sum_{h=0}^{H-1} \sum_{s,a} \frac{(\phi_h^{\pi}(s,a) - \phi_h^{\bar{\pi}}(s,a))^2}{K_\mu \mu_h(s,a)}} + \sqrt{\sum_{h=0}^{H-1} \sum_{s} \big(Q_h^{\pi}(s, \pi_h(s)) - Q_h^{\pi}(s, \bar{\pi}_h(s)) \big)^2 \frac{w_h^{\bar{\pi}}(s)}{K_{\bar{\pi}}}} \\ &= \sqrt{H^2 \sum_{h=0}^{H-1} \frac{\|\phi_h^{\pi} - \phi_h^{\bar{\pi}}\|^2_{\Lambda_h(\mu)^{-1}}}{K_\mu}} + \sqrt{\frac{U(\pi, \bar{\pi})}{K_{\bar{\pi}}}}. \end{aligned}$$

https://doi.org/10.52202/079017-0717

Here, we applied Bernstein's inequality and observed that $\sum_{s'} V_{h+1}^{\pi}(s')^2 P_h(s'|s,a) \leq H^2$. Now, we have that if

$$K_\mu \gtrsim \max_{\pi \in \Pi} \sum_{h=0}^{H-1} \frac{H^2 \|\phi_h^\pi - \phi_h^{\bar{\pi}}\|^2_{\Lambda_h(\mu)^{-1}}}{\max\{\epsilon^2, \Delta(\pi)^2\}} \qquad \text{and} \qquad K_{\bar{\pi}} \gtrsim \max_{\pi \in \Pi} \frac{U(\pi, \bar{\pi})}{\max\{\epsilon^2, \Delta(\pi)^2\}} \tag{4.4}$$

then Condition (4.2) holds. Notice that up to H and $\log(\cdot)$ factors, this is precisely the sample complexity of Theorem 1 if we set $\bar{\pi} = \pi^\star$ and minimize over all logging/exploration policies μ/π_{exp}. Note that, if $\bar{V}$ denotes the average reward collected from rolling out $\bar{\pi}$ $K_{\bar{\pi}}$ times, then $|\bar{V} - V_0^{\bar{\pi}}| \leq \sqrt{\frac{H^2}{K_{\bar{\pi}}}}$ by Hoeffding's inequality. Thus, one could use $\widehat{V}^\pi = \widehat{D}^\pi + \bar{V}$ as an effective off-policy estimator. Likewise, $\widehat{D}^\pi - \widehat{D}^{\pi'}$ is an effective estimator for $V_0^\pi - V_0^{\pi'}$.

This calculation (elaborated on in Appendix A) suggests that our analysis is tight, and clearly illustrates that the $U(\pi, \bar{\pi})$ term arises due to estimating the behavior of the reference policy $w_h^{\bar{\pi}}$. The $U(\pi, \bar{\pi})$ term is, to the best of our knowledge, novel in the literature. More precisely, this term corresponds to the cost of estimating where $\bar{\pi}$ visits, if our goal is to estimate the difference in value between policy π and $\bar{\pi}$. If, for a given state, the actions taken by π and $\bar{\pi}$ achieve the same long-term reward, then it is not critical that the frequency with which $\bar{\pi}$ visits this state is estimated, as it does not affect the difference in values between π and $\bar{\pi}$; if the actions take by π and $\bar{\pi}$ do achieve different long-term reward at s, then we must estimate the behavior of each policy at this state. This is reflected by the term inside the expectation of $U(\pi, \bar{\pi})$; this will be 0 in the former case, and scale with the difference between long-term action reward in the latter case.

Additionally, note that if we had offline data from *some* policy $\bar{\pi}$, that had been played for a long time, so that $K_{\bar{\pi}} \approx \infty$, then we would only incur the K_μ term; this is precisely ρ_Π, but with $\pi^\star$ replaced with our reference policy $\bar{\pi}$ in the numerator.

5 Achieving Theorem 1: PERP Algorithm

While the above section provides intuition for where the terms in Theorem 1 come from, it does not lead to a practical algorithm. This is because the desired number of samples in Equation (4.4) are in terms of unknown quantities: $\{\|\phi_h^\pi - \phi_h^{\bar{\pi}}\|^2_{\Lambda_h(\mu)^{-1}}, \Delta(\pi), U(\pi, \bar{\pi})\}$, which depend on our unknown environment variables ν_h, P_h; hence, we would not know how many samples to collect. In this section, we propose an algorithm that will proceed in rounds, successively improving our estimates of these quantities. Define

$$\widehat{U}_{\ell,h}(\pi, \pi') := \widehat{\mathbb{E}}_{\pi',\ell}[(\widehat{Q}^\pi_{\ell,h}(s_h, \pi_h(s)) - \widehat{Q}^\pi_{\ell,h}(s_h, \pi'_h(s)))^2], \tag{5.1}$$

where $\widehat{\mathbb{E}}_{\pi',\ell}$ denotes the expectation induced playing policy π' on the MDP with transitions $\widehat{P}_{\ell,h}$, and $\widehat{Q}^\pi_{\ell,h}$ denotes the Q-function of policy π on this same MDP. To compute $\widehat{P}_{\ell,h}$, we use the standard estimator: $\widehat{P}_{\ell,h}(s' \mid s, a) = \frac{N_{\ell,h}(s,a,s')}{N_{\ell,h}(s,a)}$ for $N_{\ell,h}(s,a)$ and $N_{\ell,h}(s,a,s')$ the visitation counts in $\mathfrak{D}^{\text{ED}}_{\ell,h}$. We set $\widehat{P}_{\ell,h}(s' \mid s, a) = \text{unif}(\mathcal{S})$ if $N_{\ell,h}(s,a) = 0$. The analogous estimator is used to estimate $\widehat{r}_{\ell,h}$. The quantity $\phi_h^\pi - \phi_h^{\bar{\pi}}$ is estimated as in the previous section: $(\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h})\widehat{w}^{\bar{\pi}}_{\ell,h} + \boldsymbol{\pi}_h \widehat{\delta}^\pi_{\ell,h}$.

Algorithm 1 proceeds in epochs. It begins with a policy set Π_1, which contains all policies of interest, Π. It then gradually begins to refine this policy set, seeking to estimate the *difference* in values between policies in the set up to tolerance $\epsilon_\ell = 2^{-\ell}$. To achieve this, it instantiates the intuition above. First, it chooses a reference policy $\bar{\pi}_\ell$, then running this estimate a sufficient number of times to estimate $w_h^{\bar{\pi}_\ell}$. Given this estimate, it then seeks to estimate δ_h^π for each π in the active set of policies, Π_ℓ, by collecting data covering the directions $(\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h})\widehat{w}^{\bar{\pi}}_{\ell,h} + \boldsymbol{\pi}_h \widehat{\delta}^\pi_{\ell,h}$ for all $\pi \in \Pi_\ell$. To efficiently collect this covering data, on line 12, we run a data collection procedure first developed in [42]. Finally, after estimating each δ_h^π, it estimates the differences between policy values as in (4.3), and eliminates suboptimal policies.

The computational complexity of PERP is poly $(S, A, H, 1/\epsilon, |\Pi|, \log(1/\delta))$. The primary contributor to the computational complexity is the the use of the Franke-Wolfe algorithm for experiment design in the OPTCOV subroutine. Lemma 37 from Wagenmaker and Pacchiano [43] shows that the number of iterations of the Franke-Wolfe algorithm is bounded polynomially in the problem parameters,

Algorithm 1 PERP: Policy Elimination with Reference Policy (informal)

Require: tolerance ϵ, confidence δ, policies Π

$\Pi_1 \leftarrow \Pi$, $\widehat{P}_0 \leftarrow$ arbitrary transition matrix
for $\ell = 1, 2, 3, \ldots, \lceil \log_2 \frac{16}{\epsilon} \rceil$ **do**
  Set $\epsilon_\ell \leftarrow 2^{-\ell}$
  `// Compute new reference policy`
  Compute $\widehat{U}_{\ell-1,h}(\pi, \pi')$ as in (5.1) for all $(\pi, \pi') \in \Pi_\ell$
  Choose $\bar{\pi}_\ell \leftarrow \min_{\bar{\pi} \in \Pi_\ell} \max_{\pi \in \Pi_\ell} \sum_{h=1}^{H} \widehat{U}_{\ell-1,h}(\pi, \bar{\pi})$
  Collect the following number of episodes from $\bar{\pi}_\ell$ and store in dataset $\mathfrak{D}_\ell^{\mathrm{ref}}$

$$\bar{n}_\ell = \mathcal{O}\Big(\max_{\pi \in \Pi_\ell} c \cdot \tfrac{H \widehat{U}_{\ell-1}(\pi, \bar{\pi}_\ell)}{\epsilon_\ell^2} \cdot \log \tfrac{H \ell^2 |\Pi_\ell|}{\delta} \Big)$$

  Compute $\{\widehat{w}^{\bar{\pi}}_{\ell,h}(s)\}_{h=1}^{H}$ using empirical state visitation frequencies in $\mathfrak{D}_\ell^{\mathrm{ref}}$
  `// Estimate Policy Differences`
  Initialize $\widehat{\delta}_1^{\pi} \leftarrow 0$
  for $h = 1, \ldots, H$ **do**
    Run OPTCOV (Algorithm 3) to collect dataset $\mathfrak{D}_{\ell,h}^{\mathrm{ED}}$ such that:

$$\sup_{\pi \in \Pi_\ell} \|(\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h}) \widehat{w}^{\bar{\pi}}_{\ell,h} + \boldsymbol{\pi}_h \widehat{\delta}^{\pi}_{\ell,h}\|^2_{\Lambda_{\ell,h}^{-1}} \le \epsilon_\ell^2 / H^4 \beta_\ell^2 \quad \text{for} \quad \Lambda_{\ell,h} = \textstyle\sum_{(s,a) \in \mathfrak{D}_{\ell,h}^{\mathrm{ED}}} e_{sa} e_{sa}^\top$$

    and $\beta_\ell \leftarrow \mathcal{O}(\sqrt{\log SH\ell^2|\Pi_\ell|/\delta})$
    Use $\mathfrak{D}_{\ell,h}^{\mathrm{ED}}$ to compute $\widehat{P}_{\ell,h}(s'|s,a)$ and $\widehat{r}_{\ell,h}$
    Compute $\widehat{\delta}^{\pi}_{\ell,h+1} \leftarrow \widehat{P}_{\ell,h}(\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h}) \widehat{w}^{\bar{\pi}}_{\ell,h} + \widehat{P}_{\ell,h} \boldsymbol{\pi}_h \widehat{\delta}^{\pi}_{\ell,h})$
  end for
  `// Eliminate suboptimal policies`
  Compute $\widehat{D}_{\bar{\pi}_\ell}(\pi) \leftarrow \sum_h \langle \widehat{r}_{\ell,h}, \boldsymbol{\pi}_h \widehat{\delta}_{\ell,h} \rangle + \sum_h \langle \widehat{r}_{\ell,h}, (\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h}) \widehat{w}^{\bar{\pi}}_{\ell,h} \rangle$
  Update $\Pi_{\ell+1} = \Pi_\ell \backslash \{\pi \in \Pi_\ell : \max_{\pi'} \widehat{D}_{\bar{\pi}_\ell}(\pi') - \widehat{D}_{\bar{\pi}_\ell}(\pi) > 8\epsilon_\ell \}$
  if $|\Pi_{\ell+1}| = 1$ **then return** $\pi \in \Pi_{\ell+1}$
end for
return any $\pi \in \Pi_{\ell+1}$

and from the definition of this procedure given in Wagenmaker and Pacchiano [43], we see that each iteration of Franke-Wolfe has computational complexity polynomial in problem parameters. We omit several technical details from Algorithm 1 for simplicity, but present the full definition in Algorithm 2.

6 When is ρ_Π Sufficient?

Our results so far show that ρ_Π is not in general sufficient for tabular RL. In this section, we consider several special cases where it *is* sufficient.

Tabular Contextual Bandits. The tabular contextual bandit setting is the special case of the RL setting with $H = 1$ and where the initial action does not affect the next-state transition. Theorem 2.2 of Li et al. [30] show that if the rewards distributions $\nu(s, a)$ are Gaussian for each (s, a), where here s denotes the context, any $(0, \delta)$-PAC algorithm requires at least ρ_Π samples. Crucially, however, they assume that the context distribution—in this case corresponding to the initial transition P_1—is known. Their algorithm makes explicit use of this fact, using this to estimate the value of ϕ^π. The following result shows that knowing the context distribution is not critical—we can achieve a complexity of $\mathcal{O}(\rho_\Pi)$ without this prior knowledge.

Corollary 1. *For the setting of tabular contextual bandits, there exists an algorithm such that with probability at least* $1 - 2\delta$*, as long as* Π *contains only deterministic policies, it finds an* ϵ*-optimal*

https://doi.org/10.52202/079017-0717

policy and terminates after collecting at most the following number of samples:

$$\inf_{\pi_{\text{exp}}} \max_{\pi \in \Pi} \frac{\|\phi^\star - \phi^\pi\|^2_{\Lambda(\pi_{\text{exp}})^{-1}}}{\max\{\epsilon^2, \Delta(\pi)^2\}} \cdot \beta^2 \log \frac{1}{\Delta_{\min} \vee \epsilon} + \frac{C_{\text{poly}}}{\max\{\epsilon^{5/3}, \Delta_{\min}^{5/3}\}},$$

for $C_{\text{poly}} = \text{poly}(|\mathcal{S}|, A, \log 1/\delta, \log 1/(\Delta_{\min} \vee \epsilon), \log |\Pi|)$ *and* $\beta = C\sqrt{\log(\frac{S|\Pi|}{\delta} \cdot \frac{1}{\Delta_{\min} \vee \epsilon})}$.

The theorem is proved in Appendix D, and follows from the application of our algorithm PERP to the contextual bandit problem. The key intuition behind this result is that, in the contextual case:

$$U(\pi, \bar{\pi}) = \mathbb{E}_{s \sim P_1}[(r_1(s, \pi_1(s)) - r_1(s, \bar{\pi}_1(s))^2] \leq \mathbb{E}_{s \sim P_1}[\mathbb{I}\{\pi_1(s) \neq \bar{\pi}_1(s)\}].$$

It is then possible to show that, since π_{exp} only has choices of which actions are taken (and cannot affect the context distribution), this can be further bounded by $\inf_{\pi_{\text{exp}}} \|\phi^\pi - \phi^{\bar{\pi}}\|^2_{\Lambda(\pi_{\text{exp}})^{-1}}$. This is not true in the full MDP case, where our choice of exploration policy in π_{exp} could make $\inf_{\pi_{\text{exp}}} \|\phi^\pi - \phi^{\bar{\pi}}\|^2_{\Lambda(\pi_{\text{exp}})^{-1}}$ significantly smaller than $U(\pi, \bar{\pi})$ (as is the case in Lemma 2). Hence, we observe that the cost of learning the contexts is dominated by that of learning the rewards in the case of contextual bandits. This is the opposite of tabular RL, where our complexity from Theorem 1 is unchanged (as seen in Section 4.2) even if we knew the reward distribution. This shows that there is a distinct separation between instance-optimal learning in tabular RL vs contextual bandits.

MDPs with Action-Independent Transitions. In the special case of MDPs where the transitions do not depend on the actions selected, the complexity simplifies to $\mathcal{O}(\rho_\Pi)$. Note that this exactly matches (up to lower order terms) the lower bound from [2].

Corollary 2. *Assume that all* P_h *are such that* $P_h(s'|s, a) = P_h(s'|s, a')$ *for all* $(a, a') \in \mathcal{A}$. *Then, with probability at least* $1 - 2\delta$, PERP *(Algorithm 2) finds an* ϵ*-optimal policy and terminates after collecting at most the following number of episodes:*

$$\sum_{h=1}^{H} \inf_{\pi_{\text{exp}}} \max_{\pi \in \Pi} \frac{\|\phi_h^\star - \phi_h^\pi\|^2_{\Lambda_h(\pi_{\text{exp}})^{-1}}}{\max\{\epsilon^2, \Delta(\pi)^2\}} \cdot \iota H^4 \beta^2 + \frac{C_{\text{poly}}}{\max\{\epsilon^{5/3}, \Delta_{\min}^{5/3}\}}$$

for C_{poly}, β *as defined in Theorem 1.*

The intuition for Corollary 2 is similar to that of Corollary 1, and proved in Appendix E.

7 Discussion

In this paper, we performed a fine-grained study of the instance-dependent complexity of tabular RL. We proposed a new off-policy estimator that estimates the value relative to a reference policy. We leveraged this insight to close the instance-dependent contextual bandits problem and obtained the tightest known upper bound for tabular MDPs.

Limitations and Future work One limitation of the present work is that PERP, in it's current form, would be too computationally expensive to run for most practical applications; enumerating the policy set Π is often intractable, but works in contextual bandits have avoided this issue by only relying on argmax oracles over this set [1, 30]; an interesting direction of future work would be to extend this technique to tabular RL. Extending the results from this paper to obtain refined instance-dependent bounds for linear MDPs and general function approximation is an exciting direction as well.

The new estimator and its improved sample complexity raise additional theoretical questions. Our upper bound has unfortunate low order terms; can these be removed? Can one show that $\frac{U(\pi, \bar{\pi})}{\max\{\Delta(\pi)^2, \epsilon^2\}}$ is unavoidable for all MDPs in general, thereby matching our upper bound? As discussed above, a few works have proven gap-dependent regret upper bounds, but we are unaware of any matching lower bounds besides over restricted classes of MDPs; can our estimator involving the differences result in even tighter instance-dependent regret bounds for MDPs?

Acknowledgments

AN and LJR are supported in part by ONR YIP N000142012571 and NSF CAREER 1844729. AN was supported, in part by the Amazon Hub Fellowship at the University of Washington. KJ and AW were funded in part by NSF CAREER 2141511 and NSF TRIPODS 2023239.

References

[1] Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pages 1638–1646. PMLR, 2014.

[2] Aymen Al-Marjani, Andrea Tirinzoni, and Emilie Kaufmann. Towards instance-optimality in online pac reinforcement learning. *arXiv preprint arXiv:2311.05638*, 2023.

[3] Peter Auer and Ronald Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. *Advances in neural information processing systems*, 19, 2006.

[4] Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

[5] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

[6] Avinandan Bose, Mihaela Curmei, Daniel L. Jiang, Jamie Morgenstern, Sarah Dean, Lillian J. Ratliff, and Maryam Fazel. Initializing Services in Interactive ML Systems for Diverse Users. *arXiv preprint arXiv:2312.11846*, 2023.

[7] Avinandan Bose, Simon Shaolei Du, and Maryam Fazel. Offline Multi-task Transfer RL with Representational Penalization. *arXiv preprint arXiv:2402.12570*, 2024.

[8] Ronen I Brafman and Moshe Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct):213–231, 2002.

[9] Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. *Advances in Neural Information Processing Systems*, 2015.

[10] Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 2017.

[11] Christoph Dann, Lihong Li, Wei Wei, and Emma Brunskill. Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning*, pages 1507–1516. PMLR, 2019.

[12] Christoph Dann, Teodor Vanislavov Marinov, Mehryar Mohri, and Julian Zimmert. Beyond value-function gaps: Improved instance-dependent regret bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[13] Rémy Degenne, Pierre Ménard, Xuedong Shang, and Michal Valko. Gamification of pure exploration for linear bandits. In *International Conference on Machine Learning*, pages 2432–2442. PMLR, 2020.

[14] Kefan Dong and Tengyu Ma. Asymptotic instance-optimal algorithms for interactive decision making. *arXiv preprint arXiv:2206.02326*, 2022.

[15] Vivek Farias, Andrew Li, Tianyi Peng, and Andrew Zheng. Markovian interference in experiments. *Advances in Neural Information Processing Systems*, 35:535–549, 2022.

[16] Tanner Fiez, Lalit Jain, Kevin G Jamieson, and Lillian Ratliff. Sequential experimental design for transductive linear bandits. *Advances in neural information processing systems*, 32, 2019.

[17] Peter W Glynn, Ramesh Johari, and Mohammad Rasouli. Adaptive experimental design with temporal interference: A maximum likelihood approach. *Advances in Neural Information Processing Systems*, 33:15054–15064, 2020.

[18] Jiafan He, Dongruo Zhou, and Quanquan Gu. Logarithmic regret for reinforcement learning with linear function approximation. *International Conference on Machine Learning*, 2021.

 https://doi.org/10.52202/079017-0717

[19] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 4868–4878, 2018.

[20] Anders Jonsson, Emilie Kaufmann, Pierre Ménard, Omar Darwiche Domingues, Edouard Leurent, and Michal Valko. Planning in markov decision processes with gap-dependent sample complexity. *Advances in Neural Information Processing Systems*, 2020.

[21] Sham Machandranath Kakade. *On the sample complexity of reinforcement learning*. PhD thesis, UCL (University College London), 2003.

[22] Julian Katz-Samuels, Lalit Jain, and Kevin G Jamieson. An empirical process approach to the union bound: Practical algorithms for combinatorial and linear bandits. *Advances in Neural Information Processing Systems*, 33:10371–10382, 2020.

[23] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1): 1–42, 2016.

[24] Michael Kearns and Satinder Singh. Finite-sample convergence rates for q-learning and indirect algorithms. *Advances in neural information processing systems*, 11, 1998.

[25] Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2):209–232, 2002.

[26] Michael Kearns, Yishay Mansour, and Andrew Ng. Approximate planning in large pomdps via reusable trajectories. *Advances in Neural Information Processing Systems*, 12, 1999.

[27] Johannes Kirschner, Tor Lattimore, Claire Vernade, and Csaba Szepesvári. Asymptotically optimal information-directed sampling. In *Conference on Learning Theory*, pages 2777–2821. PMLR, 2021.

[28] Tor Lattimore and Marcus Hutter. Pac bounds for discounted mdps. In *International Conference on Algorithmic Learning Theory*, pages 320–334. Springer, 2012.

[29] Tor Lattimore and Csaba Szepesvari. The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *Artificial Intelligence and Statistics*, pages 728–737. PMLR, 2017.

[30] Zhaoqi Li, Lillian Ratliff, Houssam Nassif, Kevin Jamieson, and Lalit Jain. Instance-optimal PAC algorithms for contextual bandits. *Advances in Neural Information Processing Systems*, 2022.

[31] Aymen Al Marjani and Alexandre Proutiere. Adaptive sampling for best policy identification in markov decision processes. *International Conference on Machine Learning*, 2021.

[32] Aymen Al Marjani, Aurélien Garivier, and Alexandre Proutiere. Navigating to the best policy in markov decision processes. *Neural Information Processing Systems*, 2021.

[33] Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Emilie Kaufmann, Edouard Leurent, and Michal Valko. Fast active learning for pure exploration in reinforcement learning. *International Conference on Machine Learning*, 2021.

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

[35] Jungseul Ok, Alexandre Proutiere, and Damianos Tranos. Exploration in structured reinforcement learning. *arXiv preprint arXiv:1806.00775*, 2018.

[36] Max Simchowitz and Kevin G Jamieson. Non-asymptotic gap-dependent regret bounds for tabular mdps. *Advances in Neural Information Processing Systems*, 32, 2019.

[37] Marta Soare, Alessandro Lazaric, and Rémi Munos. Best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 27, 2014.

[38] Alexander L Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michael L Littman. Pac model-free reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 881–888, 2006.

[39] Alexander L Strehl, Lihong Li, and Michael L Littman. Reinforcement learning in finite mdps: Pac analysis. *Journal of Machine Learning Research*, 10(11), 2009.

[40] Andrea Tirinzoni, Matteo Pirotta, Marcello Restelli, and Alessandro Lazaric. An asymptotically optimal primal-dual incremental algorithm for contextual linear bandits. *Neural Information Processing Systems*, 2020.

[41] Andrea Tirinzoni, Aymen Al-Marjani, and Emilie Kaufmann. Near instance-optimal pac reinforcement learning for deterministic mdps. *Neural Information Processing Systems*, 2022.

[42] Andrew Wagenmaker and Kevin G Jamieson. Instance-dependent near-optimal policy identification in linear mdps via online experiment design. *Advances in Neural Information Processing Systems*, 35:5968–5981, 2022.

[43] Andrew Wagenmaker and Aldo Pacchiano. Leveraging offline data in online reinforcement learning. *International Conference of Machine Learning*, 2023.

[44] Andrew Wagenmaker, Yifang Chen, Max Simchowitz, Simon S Du, and Kevin Jamieson. First-order regret in reinforcement learning with linear function approximation: A robust estimation approach. *International Conference of Machine Learning*, 2022.

[45] Andrew J Wagenmaker and Dylan J Foster. Instance-optimality in interactive decision making: Toward a non-asymptotic theory. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1322–1472. PMLR, 2023.

[46] Andrew J Wagenmaker, Max Simchowitz, and Kevin Jamieson. Beyond no regret: Instance-dependent pac reinforcement learning. In *Conference on Learning Theory*, pages 358–418. PMLR, 2022.

[47] Haike Xu, Tengyu Ma, and Simon S Du. Fine-grained gap-dependent bounds for tabular mdps via adaptive multi-step bootstrap. *Conference on Learning Theory*, 2021.

[48] Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pages 7304–7312. PMLR, 2019.

[49] Andrea Zanette, Mykel J Kochenderfer, and Emma Brunskill. Almost horizon-free structure-aware best policy identification with a generative model. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[50] Zihan Zhang, Xiangyang Ji, and Simon S Du. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. *Conference on Learning Theory*, 2021.

[51] Zihan Zhang, Yuxin Chen, Jason D Lee, and Simon S Du. Settling the sample complexity of online reinforcement learning. *arXiv preprint arXiv:2307.13586*, 2023.

 https://doi.org/10.52202/079017-0717

Contents

Notation	Description
$\mathcal{S}$	State space
$\mathcal{A}$	Action space
H	Horizon
P_h	Transition matrix at stage h
ν_h	Distribution over reward at stage h
$r_h(s,a)$	Expected reward at stage h for state s and action a
π	Policy
Π	Set of candidate policies
$\pi_h(s)$	Distribution over actions for policy π at state s and stage h
w_h^π	State visitation vector at step h for policy π
$\boldsymbol{\pi}_h$	Policy matrix for policy π at step h
ϕ_h^π	State-action visitation vector for policy π at step h
$\Lambda_h(\pi)$	Expected covariance matrix at timestep h for policy π
$Q_h^\pi(s,a)$	Q-value function for policy π at state s, action a, and step h
$V_h^\pi(s)$	Value function for policy π at state s and step h
V^π	Value of policy π
$\pi^\star$	Optimal policy within Π
$\Delta(\pi)$	Suboptimality of policy π
$W_h^\star(s)$	Maximum probability of reaching state s at step h over all policies
$\mathcal{C}$	Context space (for contextual bandits)
$\mu^\star$	Context distribution (for contextual bandits)
$\theta^\star$	Reward parameters (for contextual bandits)
ρ_Π	Complexity measure based on feature differences
$\bar{\pi}_\ell$	Reference policy
δ_h^π	Difference in state visitation between policy π and reference policy at step h
$D_{\bar{\pi}_\ell}(\pi)$	Difference in value between policy π and reference policy
$U_h(\pi,\pi')$	Expected squared difference in Q-values between policies π and π' at step h
$\mathcal{S}_\ell^{\text{keep}}$	Set of reachable states at epoch ℓ
$\epsilon_{\text{unif}}^\ell$	Minimum reachability threshold at epoch ℓ
$\epsilon_{\text{exp}}^\ell$	Tolerance for experiment design at epoch ℓ
β_ℓ	Confidence parameter at epoch ℓ
$n_\ell, K_{\text{unif}}^\ell$	Number of samples and minimum exploration at epoch ℓ
$\mathfrak{D}_{\ell,h}^{\text{ED}}$	Dataset collected during exploration in PERP
$\mathfrak{D}_\ell^{\text{ref}}$	Dataset collected from reference policy

Table 1: Table of notation used in the paper

A Understanding the origins of $U(\pi,\bar{\pi})$

This section is inspired by the exposition of Soare et al. [37] for justifying the sample complexity of linear bandits. Fix a reference policy $\bar{\pi}$ and some (stochastic) logging policy μ. For $K \in \mathbb{N}$ to be determined later, roll out $\bar{\pi}$ K times and compute the empirical state visitations $\widehat{w}_h^{\bar{\pi}}(s) = \frac{1}{K}\sum_{k=1}^K \sum_{s,h} \mathbf{1}\{s_h^k = s\}$. Also roll out μ K times and compute the empirical transition probabilities $\widehat{P}_h(s'|s,a) = \frac{\sum_{k=1}^K \mathbf{1}\{(s_h^k,a_h^k,s_{h+1}^k)=(s,a,s')\}}{\sum_{k=1}^K \mathbf{1}\{(s_h^k,a_h^k)=(s,a)\}}$. For any $\pi \neq \bar{\pi}$, use $\{\widehat{P}_h(s'|s,a)\}_{s,a,s',h}$ to compute $\widehat{w}_h^\pi(s)$. With $\delta_{h+1}^\pi := w_{h+1}^\pi - w_{h+1}^{\bar{\pi}} = P_h\boldsymbol{\pi}_h w_h^\pi - P_h\bar{\boldsymbol{\pi}}_h w_h^{\bar{\pi}} = P_h\boldsymbol{\pi}_h\delta_h^\pi + P_h(\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_h)w_h^{\bar{\pi}}$ set

$$D(\pi) = V_0^\pi - V_0^{\bar{\pi}} = \sum_{h=1}^H \langle r_h, \boldsymbol{\pi}_h w_h^\pi - \bar{\boldsymbol{\pi}}_h w_h^{\bar{\pi}}\rangle = \sum_{h=1}^H \langle r_h, \boldsymbol{\pi}_h\delta_h^\pi\rangle + \langle r_h, (\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_h)w_h^{\bar{\pi}}\rangle$$

and also define the empirical counterparts $\widehat{\delta}_{h+1}^\pi := \widehat{P}_h\boldsymbol{\pi}_h\widehat{\delta}_h^\pi + \widehat{P}_h(\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_h)\widehat{w}_h^{\bar{\pi}}$ with

$$\widehat{D}(\pi) = \sum_{h=1}^H \langle r_h, \boldsymbol{\pi}_h\widehat{\delta}_h^\pi\rangle + \langle r_h, (\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_h)\widehat{w}_h^{\bar{\pi}}\rangle.$$

https://doi.org/10.52202/079017-0717

If $\widehat{\pi} = \arg\max_{\pi\in\Pi} \widehat{D}(\pi)$, how large must K be to ensure that $\widehat{\pi} = \pi^\star := \arg\max_{\pi\in\Pi} D(\pi) = \arg\max_{\pi\in\Pi} V_0^\pi$?

Assume at time $h = 0$ all policies are initialized arbitrarily in some state s_0 so that $\widehat{P}_0(s'|s_0, a)$ simply defines the initial empirical state distribution at time $h = 1$. Let $\widehat{w}_0^\pi(s_0) = w_0^\pi(s_0) = 1$ We can then unroll the recursion for $h = 0, \ldots, H-1$

$$
\begin{aligned}
&\widehat{\delta}_{h+1}^\pi - \delta_{h+1}^\pi = \widehat{P}_h \boldsymbol{\pi}_h \widehat{\delta}_h^\pi + \widehat{P}_h(\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_h)\widehat{w}_h^{\bar{\pi}} - \delta_{h+1}^\pi \\
&= (\widehat{P}_h - P_h)\boldsymbol{\pi}_h \delta_h^\pi + (\widehat{P}_h - P_h)(\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_h) w_h^{\bar{\pi}} + P_h(\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_h)(\widehat{w}_h^{\bar{\pi}} - w_h^{\bar{\pi}}) + P_h \boldsymbol{\pi}_h(\widehat{\delta}_h^\pi - \delta_h^\pi) \\
&\qquad + \underbrace{(\widehat{P}_h - P_h)\boldsymbol{\pi}_h(\widehat{\delta}_h^\pi - \delta_h^\pi) + (\widehat{P}_h - P_h)(\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_h)(\widehat{w}_h^{\bar{\pi}} - w_h^{\bar{\pi}})}_{\text{Low order terms} \approx 0} \\
&\approx (\widehat{P}_h - P_h)(\phi_k^\pi - \phi_k^{\bar{\pi}}) + P_h(\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_h)(\widehat{w}_h^{\bar{\pi}} - w_h^{\bar{\pi}}) + P_h \boldsymbol{\pi}_h(\widehat{\delta}_h^\pi - \delta_h^\pi) \\
&\approx \sum_{i=0}^{h} \Big(\prod_{j=h-i+1}^{h} P_j \boldsymbol{\pi}_j \Big) \big((\widehat{P}_{h-i} - P_{h-i})(\phi_{h-i}^\pi - \phi_{h-i}^{\bar{\pi}}) + P_{h-i}(\boldsymbol{\pi}_{h-i} - \bar{\boldsymbol{\pi}}_{h-i})(\widehat{w}_{h-i}^{\bar{\pi}} - w_{h-i}^{\bar{\pi}}) \big) \\
&= \sum_{k=0}^{h} \Big(\prod_{j=k+1}^{h} P_j \boldsymbol{\pi}_j \Big) \big((\widehat{P}_k - P_k)(\phi_k^\pi - \phi_k^{\bar{\pi}}) + P_k(\boldsymbol{\pi}_k - \bar{\boldsymbol{\pi}}_k)(\widehat{w}_k^{\bar{\pi}} - w_k^{\bar{\pi}}) \big)
\end{aligned}
$$

where we recall $\phi_k^\pi = \boldsymbol{\pi}_k w_k^\pi$. If $\epsilon_{k+1} := (\widehat{P}_k - P_k)(\boldsymbol{\pi}_h w_k^\pi - \bar{\boldsymbol{\pi}} w_k^{\bar{\pi}}) + P_k(\boldsymbol{\pi}_k - \bar{\boldsymbol{\pi}}_k)(\widehat{w}_k^{\bar{\pi}} - w_k^{\bar{\pi}})$ then

$$
\begin{aligned}
\sum_{h=1}^{H} \langle r_h, \boldsymbol{\pi}_h(\widehat{\delta}_h^\pi - \delta_h^\pi) \rangle &= \sum_{h=1}^{H} \sum_{k=0}^{h-1} \langle r_h, \boldsymbol{\pi}_h \Big(\prod_{j=k+1}^{h-1} P_j \boldsymbol{\pi}_j \Big) \epsilon_{k+1} \rangle \\
&= \sum_{k=0}^{H-1} \sum_{h=k+1}^{H} \langle r_h, \boldsymbol{\pi}_h \Big(\prod_{j=k+1}^{h-1} P_j \boldsymbol{\pi}_j \Big) \epsilon_{k+1} \rangle = \sum_{k=0}^{H-1} \langle V_{k+1}^\pi, \epsilon_{k+1} \rangle \\
&= \sum_{k=0}^{H-1} \langle V_{k+1}^\pi, (\widehat{P}_k - P_k)(\phi_k^\pi - \phi_k^{\bar{\pi}}) + P_k(\boldsymbol{\pi}_k - \bar{\boldsymbol{\pi}}_k)(\widehat{w}_k^{\bar{\pi}} - w_k^{\bar{\pi}}) \rangle.
\end{aligned}
$$

Finally, we use these calculations to compute the deviation

$$
\begin{aligned}
\widehat{D}(\pi) - D(\pi) &= \sum_{h=1}^{H} \langle r_h, \boldsymbol{\pi}_h(\widehat{\delta}_h^\pi - \delta_h^\pi) \rangle + \langle r_h, (\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_h)(\widehat{w}_h^{\bar{\pi}} - w_h^{\bar{\pi}}) \rangle \\
&= \sum_{h=0}^{H-1} \langle V_{h+1}^\pi, (\widehat{P}_h - P_h)(\phi_h^\pi - \phi_h^{\bar{\pi}}) \rangle + \langle r_h + P_h^\top V_{h+1}^\pi, (\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_h)(\widehat{w}_h^{\bar{\pi}} - w_h^{\bar{\pi}}) \rangle \\
&= \sum_{h=0}^{H-1} \langle V_{h+1}^\pi, (\widehat{P}_h - P_h)(\phi_h^\pi - \phi_h^{\bar{\pi}}) \rangle + \langle Q_h^\pi, (\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_h)(\widehat{w}_h^{\bar{\pi}} - w_h^{\bar{\pi}}) \rangle \\
&= \sum_{h=0}^{H-1} \sum_{s,a,s'} V_{h+1}^\pi(s')(\widehat{P}_h(s'|s,a) - P_h(s'|s,a))(\phi_h^\pi(s,a) - \phi_h^{\bar{\pi}}(s,a)) \\
&\qquad + \sum_{h=0}^{H-1} \sum_{s} \big(Q_h^\pi(s, \pi_h(s)) - Q_h^\pi(s, \bar{\pi}_h(s)) \big)(\widehat{w}_h^{\bar{\pi}}(s) - w_h^{\bar{\pi}}(s)) \\
&\lesssim \sqrt{\sum_{h=0}^{H-1} \sum_{s,a,s'} V_{h+1}^\pi(s')^2 \frac{P_h(s'|s,a)}{K \mu_h(s,a)} (\phi_h^\pi(s,a) - \phi_h^{\bar{\pi}}(s,a))^2} \\
&\qquad + \sqrt{\sum_{h=0}^{H-1} \sum_{s} \big(Q_h^\pi(s, \pi_h(s)) - Q_h^\pi(s, \bar{\pi}_h(s)) \big)^2 \frac{w_h^{\bar{\pi}}(s)}{K}}.
\end{aligned}
$$

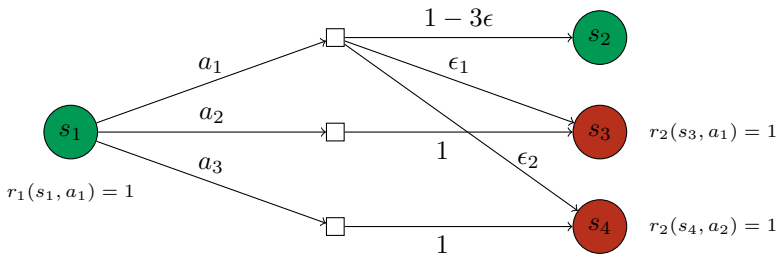

Figure 2: A motivating example for differences. All rewards other than the ones specified in the figure are 0.

Applying $\sum_{s'} V_{h+1}^{\pi}(s')^2 P_h(s'|s,a) \le H^2$, we observe that if

$$K \ge \min_{\mu,\bar{\pi}} \max_{\pi} H^2 \sum_{h=1}^{H-1} \frac{\sum_{s,a}(\phi_h^{\pi}(s,a) - \phi_h^{\bar{\pi}}(s,a))^2/\mu_h(s,a)}{\Delta(\pi)^2} + \sum_{h=1}^{H-1} \frac{\sum_s \left(Q_h^{\pi}(s,\pi_h(s)) - Q_h^{\pi}(s,\bar{\pi}_h(s))\right)^2 w_h^{\bar{\pi}}(s)}{\Delta(\pi)^2}$$

and we employ the minimizers $\mu, \bar{\pi}$ to collect data, then $\widehat{D}(\pi) - D(\pi) < \Delta(\pi)$ and $\widehat{\pi} = \arg\max_{\pi\in\Pi} \widehat{D}(\pi) = \arg\max_{\pi\in\Pi} D(\pi)$. Notice that up to H and log factors, this is precisely the sample complexity of our algorithm. A natural candidate for $\bar{\pi}$ is $\pi^\star$ so that the first term matches the lower bound of [2].

On the other hand, suppose we used the data from the logging policy μ to compute the empirical state visitations $\widehat{w}_h^{\pi}$ for all $\pi \in \Pi$ and set $\widehat{\pi} = \arg\max_{\pi\in\Pi} \sum_{h=1}^{H} \langle r_h, \boldsymbol{\pi} \widehat{w}_h^{\pi} \rangle =: \widehat{V}_0^{\pi}$. Using the same techniques as above, it is straightforward to show that if

$$\begin{aligned}
\widehat{w}_{h+1}^{\pi} - w_{h+1}^{\pi} &= \widehat{P}_h \boldsymbol{\pi}_h \widehat{w}_h^{\pi} - P_h \boldsymbol{\pi}_h w_h^{\pi} \\
&= (\widehat{P}_h - P_h + P_h)\boldsymbol{\pi}_h(\widehat{w}_h^{\pi} - w_h^{\pi} + w_h^{\pi}) - P_h \boldsymbol{\pi}_h w_h^{\pi} \\
&= (\widehat{P}_h - P_h)\boldsymbol{\pi}_h w_h^{\pi} + P_h \boldsymbol{\pi}_h(\widehat{w}_h^{\pi} - w_h^{\pi}) + \underbrace{(\widehat{P}_h - P_h)\boldsymbol{\pi}_h(\widehat{w}_h^{\pi} - w_h^{\pi})}_{\text{Low order terms} \approx 0} \\
&\approx \sum_{i=0}^{h} \Big(\prod_{j=h-i+1}^{h} P_j \boldsymbol{\pi}_j \Big)(\widehat{P}_{h-i} - P_{h-i})\boldsymbol{\pi}_{h-i} w_{h-i}^{\pi} \\
&= \sum_{k=0}^{h} \Big(\prod_{j=k+1}^{h} P_j \boldsymbol{\pi}_j \Big)(\widehat{P}_k - P_k)\boldsymbol{\pi}_k w_k^{\pi}
\end{aligned}$$

and we employ the minimizer μ to collect data, then $\widehat{V}_0^{\pi} - V_0^{\pi} \le \Delta(\pi)$ and $\widehat{\pi} = \arg\max_{\pi\in\Pi} \widehat{V}_0^{\pi} = \arg\max_{\pi\in\Pi} V_0^{\pi}$.

B Tabular MDPs: Comparison with Prior Work and Lower Bounds

Illustrative Family of MDP Instances Recall the family of MDP instances in the introduction (visualized in Figure 2 for ease of reference). The family of MDPs is parameterized by $\epsilon, \epsilon_1, \epsilon_2 > 0$, with $H = 2$, $\mathcal{S} = \{s_1, s_2, s_3, s_4\}$, and $\mathcal{A} = \{a_1, a_2, a_3\}$, which start in state s_0 and are defined as:

$$P_1(s_2 \mid s_1, a_1) = 1 - 3\epsilon, \quad P_1(s_3 \mid s_1, a_1) = \epsilon_1, \quad P_1(s_4 \mid s_1, a_1) = \epsilon_2$$
$$P_1(s_3 \mid s_1, a_2) = P_1(s_4 \mid s_1, a_3) = 1.$$

We define the reward function so that all rewards are 0 except $r_1(s_1, a_1) = r_2(s_3, a_1) = r_2(s_4, a_2) = 1$ for all a.

Let $\mathcal{M}$ denote the MDP above with $\epsilon_1 = 2\epsilon, \epsilon_2 = \epsilon$, and $\mathcal{M}'$ the MDP above with $\epsilon_1 = \epsilon, \epsilon_2 = 2\epsilon$.

https://doi.org/10.52202/079017-0717

Let $\Pi = \{\pi_1, \pi_2\}$ denote some set of policies. Let π_1 denote the policy which always plays a_1, and π_2 the policy which plays a_1 at green states and a_2 at red states i.e $\pi_2(s_1) = \pi_2(s_2) = a_1$ and $\pi_2(s_3) = \pi_2(s_4) = a_2$.

Now note that $V_0^{\mathcal{M},\pi_1} = 1 + 2\epsilon$, $V_0^{\mathcal{M},\pi_2} = 1 + \epsilon$, $V_0^{\mathcal{M}',\pi_1} = 1 + \epsilon$, and $V_0^{\mathcal{M}',\pi_2} = 1 + 2\epsilon$.

B.1 Comparison with complexities from prior work

The lemma below shows that the upper bound presented in Theorem 1 is smaller than that of PEDEL from Theorem 1 of [42] for all MDP instances.

Lemma 4. *For any MDP instance and policy set* Π*, we have that*

1. $\inf_{\pi_{\exp}} \max_{\pi\in\Pi} \frac{\|\phi_h^\pi\|^2_{\Lambda_h(\pi_{\exp})^{-1}}}{\max\{\epsilon^2,\Delta(\pi)^2,\Delta^2_{\min}\}} \geq \frac{1}{\max\{\epsilon^2,\Delta(\pi)^2,\Delta^2_{\min}\}}$
2. $$H^4 \sum_{h=1}^{H} \inf_{\pi_{\exp}} \max_{\pi\in\Pi} \frac{\|\phi_h^\star - \phi_h^\pi\|^2_{\Lambda_h(\pi_{\exp})^{-1}}}{\max\{\epsilon^2,\Delta(\pi)^2,\Delta^2_{\min}\}} \leq 4H^4 \sum_{h=1}^{H} \inf_{\pi_{\exp}} \max_{\pi\in\Pi} \frac{\|\phi_h^\pi\|^2_{\Lambda_h(\pi_{\exp})^{-1}}}{\max\{\epsilon^2,\Delta(\pi)^2,\Delta^2_{\min}\}}$$
3. $\frac{HU(\pi,\pi^\star)}{\max\{\epsilon^2,\Delta(\pi)^2,\Delta^2_{\min}\}} \leq H^4 \sum_{h=1}^{H} \inf_{\pi_{\exp}} \max_{\pi\in\Pi} \frac{\|\phi_h^\pi\|^2_{\Lambda_h(\pi_{\exp})^{-1}}}{\max\{\epsilon^2,\Delta(\pi)^2,\Delta^2_{\min}\}}$

Proof. **Proof of Claim 1.** Note that

$$\|\phi_h^\pi\|^2_{\Lambda_h(\pi_{\exp})^{-1}} = \sum_{s,a} \frac{\phi_h^\pi(s,a)^2}{\phi_h^{\pi_{\exp}}(s,a)} \geq \inf_{\lambda\in\Delta_{SA}} \sum_{s,a} \frac{\phi_h^\pi(s,a)^2}{\lambda_{s,a}}$$

In order to solve this optimization problem, we can consider the KKT conditions. We can verify from stationarity that at optimality, $\lambda_{s,a} = \frac{\phi_h^\pi(s,a)}{\sqrt{\beta}}$ for some constant $\beta > 0$. But since $\lambda_{s,a}$ must live in the simplex Δ_{SA}, and since $\phi_h^\pi(s,a)$ is itself a distribution over $\mathcal{S}\times\mathcal{A}$, it follows that $\beta = 1$ must be true. Plugging this optimal value into the above, we obtain that

$$\|\phi_h^\pi\|^2_{\Lambda_h(\pi_{\exp})^{-1}} \geq \inf_{\lambda\in\Delta_{SA}} \sum_{s,a} \frac{\phi_h^\pi(s,a)^2}{\lambda_{s,a}} = 1$$

Then,

$$\inf_{\pi_{\exp}} \max_{\pi\in\Pi} \frac{\|\phi_h^\pi\|^2_{\Lambda_h(\pi_{\exp})^{-1}}}{\max\{\epsilon^2,\Delta(\pi)^2,\Delta^2_{\min}\}} \geq \frac{1}{\max\{\epsilon^2,\Delta(\pi)^2,\Delta^2_{\min}\}}$$

directly follows from the above.

Proof of Claim 2. From the triangle inequality,

$$\begin{aligned}
&\inf_{\pi_{\exp}} \max_{\pi\in\Pi} \frac{\|\phi_h^\star - \phi_h^\pi\|^2_{\Lambda_h(\pi_{\exp})^{-1}}}{\max\{\epsilon^2,\Delta(\pi)^2,\Delta^2_{\min}\}} \\
&\leq 2 \inf_{\pi_{\exp}} \max_{\pi\in\Pi} \left(\frac{\|\phi_h^\star\|^2_{\Lambda_h(\pi_{\exp})^{-1}}}{\max\{\epsilon^2,\Delta(\pi)^2,\Delta^2_{\min}\}} + \frac{\|\phi_h^\pi\|^2_{\Lambda_h(\pi_{\exp})^{-1}}}{\max\{\epsilon^2,\Delta(\pi)^2,\Delta^2_{\min}\}} \right) \\
&\leq 2 \inf_{\pi_{\exp}} \max_{\pi\in\Pi} \left(\frac{\|\phi_h^\star\|^2_{\Lambda_h(\pi_{\exp})^{-1}}}{\max\{\epsilon^2,\Delta(\pi^\star)^2,\Delta^2_{\min}\}} + \frac{\|\phi_h^\pi\|^2_{\Lambda_h(\pi_{\exp})^{-1}}}{\max\{\epsilon^2,\Delta(\pi)^2,\Delta^2_{\min}\}} \right) \\
&\leq 4 \inf_{\pi_{\exp}} \max_{\pi\in\Pi} \frac{\|\phi_h^\pi\|^2_{\Lambda_h(\pi_{\exp})^{-1}}}{\max\{\epsilon^2,\Delta(\pi)^2,\Delta^2_{\min}\}}
\end{aligned}$$

where we have used that $\Delta(\pi) \geq \Delta(\pi^\star)$ for all π. Plugging this bound into the expression from (2) from the Lemma statement completes the proof.

Proof of Claim 3. We have that

$$HU(\pi,\pi^\star) = H\sum_{h=1}^{H}\mathbb{E}_{s_h\sim w_h^{\pi^\star}}[(Q_h^\pi(s_h,\pi_h(s)) - Q_h^\pi(s_h,\pi_h^\star(s)))^2] \le H\sum_{h=1}^{H} H^2 \le H^4$$

Then,

$$\begin{aligned}\frac{HU(\pi,\pi^\star)}{\max\{\epsilon^2,\Delta(\pi)^2,\Delta_{\min}^2\}} &\le \frac{H^4}{\max\{\epsilon^2,\Delta(\pi)^2,\Delta_{\min}^2\}}\\ &\le H^4\sum_{h=1}^{H}\inf_{\pi_{\exp}}\max_{\pi\in\Pi}\frac{\|\phi_h^\pi\|^2_{\Lambda_h(\pi_{\exp})^{-1}}}{\max\{\epsilon^2,\Delta(\pi)^2,\Delta_{\min}^2\}}\end{aligned}$$

Where the final inequality follows from Claim 1 above. □

The lemma below shows that there are some instances where the complexity from Theorem 1 is strictly smaller in terms of ϵ dependence than that from Theorem 1 from [42] for PEDEL.

Lemma 5. *On MDP $\mathcal{M}$ defined above, we have:*

1. $\sum_{h=1}^{H}\inf_{\pi_{\exp}}\max_{\pi\in\Pi}\frac{\|\phi_h^\star-\phi_h^\pi\|^2_{\Lambda_h(\pi_{\exp})^{-1}}}{\max\{\epsilon^2,\Delta(\pi)^2,\Delta_{\min}^2\}} \le 15$
2. $\max_{\pi\in\Pi}\frac{HU(\pi,\pi^\star)}{\max\{\epsilon^2,\Delta(\pi)^2,\Delta_{\min}^2\}} = \frac{3H}{\epsilon}$
3. $\sum_{h=1}^{H}\inf_{\pi_{\exp}}\max_{\pi\in\Pi}\frac{\|\phi_h^\pi\|^2_{\Lambda_h(\pi_{\exp})^{-1}}}{\max\{\epsilon^2,\Delta(\pi)^2,\Delta_{\min}^2\}} \ge \frac{H}{\epsilon^2}$

Proof. **Proof of 1.** In this case we have that $\pi^* = \pi_1$, and the only other π of interest is π_2. Note that π_1 and π_2 differ only at state s_3 and s_4 at $h = 2$. Let $\pi_{\exp}$ be the policy that plays actions uniformly at random. Then, we have

$$\begin{aligned}\sum_{h=1}^{H}\inf_{\pi_{\exp}}\max_{\pi\in\Pi}\frac{\|\phi_h^\star-\phi_h^\pi\|^2_{\Lambda_h(\pi_{\exp})^{-1}}}{\max\{\epsilon^2,\Delta(\pi)^2,\Delta_{\min}^2\}} &\le \inf_{\pi_{\exp}}\frac{\|\phi_2^{\pi_1}-\phi_2^{\pi_2}\|^2_{\Lambda_h(\pi_{\exp})^{-1}}}{\epsilon^2}\\ &= \frac{1}{\epsilon^2}\left(\frac{w_2^{\pi_1}(s_3)^2}{w_2^{\pi_{\exp}}(s_3)} + \frac{w_2^{\pi_1}(s_4)^2}{w_2^{\pi_{\exp}}(s_4)}\right)\\ &\le \frac{1}{\epsilon^2}\left(\frac{4\epsilon^2}{1/3} + \frac{\epsilon^2}{1/3}\right)\\ &= 15.\end{aligned}$$

Proof of 2. Note that

$$\max_{\pi\in\Pi}\frac{HU(\pi,\pi^\star)}{\max\{\epsilon^2,\Delta(\pi)^2,\Delta_{\min}^2\}} = \frac{HU(\pi_2,\pi_1)}{\epsilon^2}.$$

Then,

$$\begin{aligned}U(\pi_2,\pi_1) &= \sum_{h=1}^{H}\mathbb{E}_{s\sim w_h^{\pi_1}}[(Q_h^{\pi_1}(s,\pi_{1,h}(s)) - Q_h^{\pi_1}(s,\pi_{2,h}(s)))^2]\\ &= \mathbb{E}_{s\sim w_2^{\pi_1}}[(Q_2^{\pi_1}(s,\pi_{1,2}(s)) - Q_2^{\pi_1}(s,\pi_{2,2}(s)))^2]\\ &= 2\epsilon + \epsilon = 3\epsilon.\end{aligned}$$

Combining these proves the result.

Proof of 3. By Claim 1 in Lemma 4, the stated result then follows by recognizing that $\max\{\epsilon^2,\Delta(\pi)^2,\Delta_{\min}^2\} \le \epsilon^2$.

□

https://doi.org/10.52202/079017-0717

B.2 Lower bound

Lemma 6. *On MDP $\mathcal{M}$ defined above, any (ϵ, δ)-PAC algorithm must collect*

$$\mathbb{E}^{\mathcal{M}}[\tau] \geq \frac{1}{\epsilon} \cdot \log \frac{1}{2.4\delta}.$$

samples.

Proof. Consider $\Pi, \mathcal{M}$, and $\mathcal{M}'$ defined above. Let $\mathcal{E}$ denote the event $\{\widehat{\pi} = \pi_1\}$. By the above observations, we have that π_1 is ϵ-optimal on $\mathcal{M}$ while π_2 is not, and that π_2 is ϵ-optimal on $\mathcal{M}'$ while π_1 is not. Then by the definition of an (ϵ, δ)-PAC algorithm, $\mathbb{P}^{\mathcal{M}}[\mathcal{E}] \geq 1 - \delta$ and $\mathbb{P}^{\mathcal{M}'}[\mathcal{E}] \leq \delta$.

Let $\gamma_h(s, a)$ denote the distribution of (r_h, s_{h+1}) given (s, a, h) on $\mathcal{M}$, and $\gamma'_h(s, a)$ is the same on $\mathcal{M}'$. Then, letting $\nu_h \leftarrow \gamma_h, \nu'_h \leftarrow \gamma'_h$ and otherwise adopting the same notation as in Lemma F.1 of [46], we have from Lemma F.1 of [46] that:

$$\begin{aligned}
\sum_{s,a,h} \mathbb{E}^{\mathcal{M}}[N_h^\tau(s, a)]\mathrm{KL}(\gamma_h(s, a), \gamma'_h(s, a)) &\geq \sup_{\mathcal{E}' \in \mathcal{F}_\tau} d(\mathbb{P}^{\mathcal{M}}[\mathcal{E}'], \mathbb{P}^{\mathcal{M}'}[\mathcal{E}']) \\
&\geq d(\mathbb{P}^{\mathcal{M}}[\mathcal{E}], \mathbb{P}^{\mathcal{M}'}[\mathcal{E}]) \\
&\geq \log \frac{1}{2.4\delta}
\end{aligned}$$

where the last inequality follows from [23].

Note that $\mathcal{M}$ and $\mathcal{M}'$ differ only at (s_1, a_1), so

$$\sum_{s,a,h} \mathbb{E}^{\mathcal{M}}[N_h^\tau(s, a)]\mathrm{KL}(\gamma_h(s, a), \gamma'_h(s, a)) = \mathbb{E}^{\mathcal{M}}[N_1^\tau(s_1, a_1)]\mathrm{KL}(\gamma_1(s_1, a_1), \gamma'_1(s_1, a_1)).$$

Furthermore, we see that

$$\mathrm{KL}(\gamma_1(s_1, a_1), \gamma'_1(s_1, a_1)) = 2\epsilon \log \frac{2\epsilon}{\epsilon} + \epsilon \log \frac{\epsilon}{2\epsilon} \leq \epsilon.$$

So it follows that we must have

$$\mathbb{E}^{\mathcal{M}}[N_1^\tau(s_1, a_1)] \geq \frac{1}{\epsilon} \cdot \log \frac{1}{2.4\delta}.$$

Noting that $\mathbb{E}^{\mathcal{M}}[N_1^\tau(s_1, a_1)] \leq \mathbb{E}^{\mathcal{M}}[\tau]$ completes the proof. □

C Tabular MDP Upper Bound

C.1 Notation

Covariance matrices. We use

$$\Lambda_h(\pi_{\mathrm{exp}}) = \mathbb{E}_{\pi_{\mathrm{exp}}}[e_{s_h a_h} e_{s_h a_h}^\top]$$

to denote the expected covariance matrix and $\widehat{\Lambda}_{\ell,h}$ to denote the empirical covariance matrix collected from $\mathfrak{D}_{\ell,h}^{\mathrm{ED}}$.

State visitations. Let $\delta_{\ell,h}^\pi(s') := w_h^\pi(s') - w_h^{\bar{\pi}_\ell}(s')$, for $\bar{\pi}_\ell$ the reference policy, $\delta_{\ell,h}^\pi$ the vectorization of $\delta_{\ell,h}^\pi(s')$, and $w_h^\pi(s) = \mathbb{P}_\pi[s_h = s]$ the visitation probability, and $W_h^\star(s) = \sup_\pi w_h^\pi(s)$. Then, we can recursively define

$$\delta_{\ell,h+1}^\pi = P_h(\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h}) w_{\ell,h}^{\bar{\pi}} + P_h \boldsymbol{\pi}_h \delta_{\ell,h}^\pi. \tag{C.1}$$

Similarly,

$$\widetilde{\delta}_{\ell,h+1}^\pi = M_h \left(P_h(\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h}) \widehat{w}_{\ell,h}^{\bar{\pi}} + P_h \boldsymbol{\pi}_h \widetilde{\delta}_{\ell,h}^\pi \right). \tag{C.2}$$

And

$$\widehat{\delta}_{\ell,h+1}^\pi = M_h \left(\widehat{P}_{\ell,h}(\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h}) \widehat{w}_{\ell,h}^{\bar{\pi}} + \widehat{P}_{\ell,h} \boldsymbol{\pi}_h \widehat{\delta}_{\ell,h}^\pi \right). \tag{C.3}$$

Algorithm 2 PERP: Policy Elimination with Reference Policy

Require: tolerance ϵ, confidence δ, policies Π

$\Pi_1 \leftarrow \Pi$, $\widehat{P}_0 \leftarrow$ arbitrary transition matrix

for $\ell = 1, 2, 3, \ldots, \lceil \log_2 \frac{16}{\epsilon} \rceil$ **do**

  Set $\epsilon_\ell \leftarrow 2^{-\ell}$, $\epsilon_{\text{unif}}^\ell \leftarrow \frac{\epsilon_\ell}{64 S^{3/2} H^2}$, $K_{\text{unif}}^\ell \leftarrow \frac{\epsilon_\ell^{-2/3}}{\epsilon_{\text{unif}}^\ell}$

  $\mathcal{S}_\ell^{\text{keep}} = \textsc{Prune}(\epsilon_{\text{unif}}^\ell, \delta/3\ell^2)$ (Algorithm 5) `// Prune states that are hard to reach`

  Use $\{\widehat{P}_{\ell-1,h}\}_{h=1}^H$ to compute $\widehat{U}_{\ell-1,h}(\pi, \pi')$ for all $(\pi, \pi') \in \Pi_\ell$ `// Compute new reference policy`

  Choose $\bar{\pi}_\ell \leftarrow \min_{\bar{\pi} \in \Pi_\ell} \max_{\pi \in \Pi_\ell} \sum_{h=1}^H \widehat{U}_{\ell-1,h}(\pi, \bar{\pi})$

  Collect the following number of episodes from $\bar{\pi}_\ell$ and store in dataset $\mathfrak{D}_\ell^{\text{ref}}$

$$\bar{n}_\ell = \max_{\pi \in \Pi_\ell} c \cdot \frac{H \widehat{U}_{\ell-1}(\pi, \bar{\pi}_\ell) + H^4 S^{3/2} \sqrt{A} \log \frac{SAH\ell^2}{\delta} \cdot \epsilon_\ell^{1/3} + S^2 H^4 \epsilon_{\text{unif}}^\ell}{\epsilon_\ell^2} \cdot \log \frac{60 H \ell^2 |\Pi_\ell|}{\delta}$$

  Compute $\{\widehat{w}_{\ell,h}^{\bar{\pi}}(s)\}_{h=1}^H$ using empirical state visitation frequencies in $\mathfrak{D}_\ell^{\text{ref}}$

  Initialize $\widehat{\delta}_1^\pi \leftarrow 0$ `// Exploration via experiment design`

  for $h = 1, \ldots, H$ **do**

    Define $M_{\ell,h} \in \mathbb{R}^{SA \times SA}$ as $M_{\ell,h} \leftarrow \text{diag}(\alpha_{s_1,a_1} \ldots \alpha_{s_S,a_A})$, where $\alpha_{s,a} = \mathbf{1}(s \in \mathcal{S}_{\ell,h}^{\text{keep}})$.

    $\Phi^\ell \leftarrow \left\{ M_{\ell,h} \left((\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h}) \widehat{w}_{\ell,h}^{\bar{\pi}} + \boldsymbol{\pi}_h \widehat{\delta}_{\ell,h}^\pi \right) \ : \ \pi \in \Pi_\ell \right\}$

    $\epsilon_{\text{exp}}^\ell \leftarrow \epsilon_\ell^2 / H^4 \beta_\ell^2$ for $\beta_\ell \leftarrow \left(\sqrt{2 \log \left(\frac{60 S H^2 \ell^2 |\Pi_\ell|}{\delta} \right)} + \frac{4}{3} \sqrt{\frac{SA}{\epsilon_{\text{unif}}^\ell K_{\text{unif}}^\ell}} \log \left(\frac{60 H^2 \ell^2 |\Pi_\ell|}{\delta} \right) \right)$

    Run $\mathfrak{D}_{\ell,h}^{\text{ED}} \leftarrow \textsc{OptCov}\left(\Phi^\ell, \epsilon_{\text{exp}}^\ell, \frac{\delta}{6H\ell^2}, \epsilon_{\text{unif}}^\ell, K_{\text{unif}}^\ell, \mathcal{S}_{\ell,h}^{\text{keep}}, h \right)$ (Algorithm 3)

    Use $\mathfrak{D}_{\ell,h}^{\text{ED}}$ to compute $\widehat{P}_{\ell,h}(s'|s,a) \leftarrow \frac{N_{\ell,h}(s',s,a)}{N_{\ell,h}(s,a)}$ if $N_{\ell,h}(s,a) > 0$, $\text{unif}(\mathcal{S})$ otherwise, and $\widehat{r}_{\ell,h}(s,a) = \frac{1}{N_{\ell,h}(s,a)} \sum_{(s',a',r',s'') \in \mathfrak{D}_{\ell,h}^{\text{ED}}} r' \cdot \mathbb{I}\{(s,a) = (s',a')\}$ if $N_{\ell,h}(s,a) > 0$, 0 otherwise

    Compute $\widehat{\delta}_{\ell,h+1}^\pi \leftarrow M_{\ell,h}(\widehat{P}_{\ell,h}(\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h}) \widehat{w}_{\ell,h}^{\bar{\pi}} + \widehat{P}_{\ell,h} \boldsymbol{\pi}_h \widehat{\delta}_{\ell,h}^\pi)$

  end for

  Compute $\widehat{D}_{\bar{\pi}_\ell}(\pi) \leftarrow \sum_h \langle \widehat{r}_{\ell,h}, \boldsymbol{\pi}_h \widehat{\delta}_{\ell,h} \rangle + \sum_h \langle \widehat{r}_{\ell,h}, (\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h}) \widehat{w}_{\ell,h}^{\bar{\pi}} \rangle$

  Update $\Pi_{\ell+1} = \Pi_\ell \backslash \{ \pi \in \Pi_\ell : \max_{\pi'} \widehat{D}_{\bar{\pi}_\ell}(\pi') - \widehat{D}_{\bar{\pi}_\ell}(\pi) > 8\epsilon_\ell \ \}$

  if $|\Pi_{\ell+1}| = 1$ **then return** $\pi \in \Pi_{\ell+1}$

end for

return any $\pi \in \Pi_{\ell+1}$

Value functions. Note that we can express the value function as:

$$V_h^\pi = \sum_{k=h}^H \left(\prod_{j=h+1}^k P_j \boldsymbol{\pi}_j \right)^\top \boldsymbol{\pi}_k^\top r_k$$

On the "pruned" MDP, define

$$\widetilde{r}_{\ell,h} = M_{\ell,h} r_h,$$

and

$$\widetilde{V}_{\ell,h} := \sum_{k=h}^H \left(\prod_{j=h+1}^k M_{\ell,j+1} P_j \boldsymbol{\pi}_j \right)^\top \boldsymbol{\pi}_k^\top \widetilde{r}_{\ell,k}.$$

Reward difference term. Define

$$U_h(\pi, \pi') := \mathbb{E}_{\pi'}[(Q_h^\pi(s_h, \pi_h(s)) - Q_h^\pi(s_h, \pi_h'(s)))^2]$$

https://doi.org/10.52202/079017-0717

and $U(\pi, \pi') := \sum_{h=1}^{H} U_h(\pi, \pi')$. Additionally, define

$$\widehat{U}_{\ell,h}(\pi, \pi') := \mathbb{E}_{\pi',\ell}[(\widehat{Q}^{\pi}_{\ell,h}(s_h, \pi_h(s)) - \widehat{Q}^{\pi}_{\ell,h}(s_h, \pi'_h(s)))^2]$$

where $\mathbb{E}_{\pi',\ell}$ denotes the expectation induced playing π' on the MDP with transitions $\widehat{P}_\ell$, and $\widehat{Q}^{\pi}_{\ell,h}$ denotes the Q-function for policy π on this same MDP. Let $\widehat{U}_\ell(\pi, \pi') := \sum_{h=1}^{H} \widehat{U}_{\ell,h}(\pi, \pi')$.

C.2 Technical Results

Lemma 7. *Let $\mathfrak{D} = \{(s_1, a_1, s'_1), \dots (s_T, a_T, s'_T)\}$ be any dataset of transitions collected from level h. Let $\widehat{P} \in \mathbb{R}^{S \times SA}$ denote the empirical transition matrix with $[\widehat{P}]_{s',sa} = \frac{N(s'|s,a)}{N(s,a)}$ if $N(s,a) > 0$, and 0 otherwise, for $N(s' \mid s, a) = \sum_t \mathbb{I}\{(s_t, a_t, s'_t) = (s, a, s')\}$ and $N(s,a) = \sum_t \mathbb{I}\{(s_t, a_t) = (s, a)\}$. Consider any $v \in [0,1]^S$ and $u \in \mathbb{R}^{SA}$ and assume that $N(s,a) > \underline{\lambda} > 0$ for all $(s,a) \in \mathrm{support}(u)$. Then, for P the true transition matrix, we have that with probability at least $1 - \delta$:*

$$\left| v^\top (P - \widehat{P}) u \right| \le \sqrt{\sum_{s,a} \frac{[u]^2_{s,a}}{N(s,a)}} \cdot \left(\sqrt{2 \log\left(\frac{1}{\delta}\right)} + \frac{4}{3\sqrt{\underline{\lambda}}} \log\left(\frac{1}{\delta}\right) \right).$$

Proof. First write

$$\begin{aligned} v^\top (P - \widehat{P}) u &= \sum_{s'} \sum_{s,a} v_{s'} \left(P(s' \mid s, a) - \frac{N(s' \mid s, a)}{N(s,a)} \right) u_{sa} \\ &= \sum_t \sum_{s'} \frac{v_{s'} \left(P(s' \mid s_t, a_t) - \mathbb{I}\{s'_t = s'\} \right) u_{s_t a_t}}{N(s_t, a_t)} \end{aligned}$$

where the second equality follows from some simple manipulations. Note that, for any t, we have

$$\mathbb{E}\left[\frac{v_{s'} \left(P(s' \mid s_t, a_t) - \mathbb{I}\{s'_t = s'\} \right) u_{s_t a_t}}{N(s_t, a_t)} \mid s_t, a_t \right] = 0$$

and can bound

$$\begin{aligned} \left| \sum_{s'} \frac{v_{s'} \left(P(s' \mid s_t, a_t) - \mathbb{I}\{s'_t = s'\} \right) u_{s_t a_t}}{N(s_t, a_t)} \right| &\le \frac{2 u_{s_t a_t}}{N(s_t, a_t)} \le \frac{2}{\sqrt{\underline{\lambda}}} \cdot \frac{u_{s_t a_t}}{\sqrt{N(s_t, a_t)}} \\ &\le \frac{2}{\sqrt{\underline{\lambda}}} \cdot \sqrt{\sum_{s,a} \frac{u^2_{sa}}{N(s,a)}} \end{aligned}$$

where we have used the fact that $N(s,a) \ge \underline{\lambda}$ for $(s,a) \in \mathrm{support}(u)$, and since v has entries in $[0,1]$ and $P(s' \mid s_t, a_t)$ and $\mathbb{I}\{s'_t = s'\}$ are valid distributions, so $\sum_{s'} v_{s'}(P(s' \mid s_t, a_t) - \mathbb{I}\{s'_t = s'\}) \in [-1, 1]$. Furthermore, we have that

$$\mathbb{E}_{s'_t}\left[\left(\sum_{s'} \frac{v_{s'} \left(P(s' \mid s_t, a_t) - \mathbb{I}\{s'_t = s'\} \right) u_{s_t a_t}}{N(s_t, a_t)} \right)^2 \right] \le \mathbb{E}_{s'_t}\left[\left(\frac{u_{s_t a_t}}{N(s_t, a_t)} \right)^2 \right] = \left(\frac{u_{s_t a_t}}{N(s_t, a_t)} \right)^2$$

where we have again used that $\sum_{s'} v_{s'}(P(s' \mid s_t, a_t) - \mathbb{I}\{s'_t = s'\}) \in [-1, 1]$.

By Bernstein's inequality, we therefore have that with probability at least $1 - \delta$:

$$\begin{aligned} \left| v^\top (P - \widehat{P}) u \right| &\le \sqrt{2 \sum_t \left(\frac{u_{s_t a_t}}{N(s_t, a_t)} \right)^2 \cdot \log \frac{2}{\delta}} + \frac{4}{3\sqrt{\underline{\lambda}}} \cdot \sqrt{\sum_t \frac{u^2_{s_t a_t}}{N(s_t, a_t)}} \cdot \log \frac{2}{\delta} \\ &= \left(\sqrt{2 \log \frac{2}{\delta}} + \frac{4}{3\sqrt{\underline{\lambda}}} \log \frac{2}{\delta} \right) \cdot \sqrt{\sum_{s,a} \frac{u^2_{sa}}{N(s,a)}}. \end{aligned}$$

□

Lemma 8. *Let $\mathfrak{D} = \{(s_1, a_1, r_1), \ldots (s_T, a_T, r_T)\}$ be any dataset of state-action-reward tuples collected from level h. Let $\widehat{r} \in \mathbb{R}^{SA}$ denote the empirical reward estimation with $[\widehat{r}]_{sa} = \frac{1}{N(s,a)} \cdot \sum_{t=1}^{T} r_t \cdot \mathbb{I}\{(s_t, a_t) = (s, a)\}$ if $N(s, a) > 0$, and 0 otherwise, for $N(s, a) = \sum_t \mathbb{I}\{(s_t, a_t) = (s, a)\}$. Consider any $u \in \mathbb{R}^{SA}$ and assume that $N(s, a) > \underline{\lambda} > 0$ for all $(s, a) \in \mathrm{support}(u)$. Then, for r the true reward mean, we have that with probability at least $1 - \delta$:*

$$\left|(r - \widehat{r})^\top u\right| \leq \sqrt{\sum_{s,a} \frac{[u]_{s,a}^2}{N(s,a)}} \cdot \left(\sqrt{2 \log\left(\frac{1}{\delta}\right)} + \frac{4}{3\sqrt{\underline{\lambda}}} \log\left(\frac{1}{\delta}\right)\right).$$

Proof. First write

$$(r - \widehat{r})^\top u = \sum_t \frac{(r(s_t, a_t) - r_t)\, u_{s_t a_t}}{N(s_t, a_t)}.$$

Note that, for any t, we have

$$\mathbb{E}\left[\frac{(r(s_t, a_t) - r_t)\, u_{s_t a_t}}{N(s_t, a_t)} \mid s_t, a_t\right] = 0$$

and can bound

$$\left|\frac{(r(s_t, a_t) - r_t)\, u_{s_t a_t}}{N(s_t, a_t)}\right| \leq \frac{u_{s_t a_t}}{N(s_t, a_t)} \leq \frac{1}{\sqrt{\underline{\lambda}}} \cdot \frac{u_{s_t a_t}}{\sqrt{N(s_t, a_t)}} \leq \frac{1}{\sqrt{\underline{\lambda}}} \cdot \sqrt{\sum_{s,a} \frac{u_{sa}^2}{N(s,a)}}$$

where we have used the fact that $N(s, a) \geq \underline{\lambda}$ for $(s, a) \in \mathrm{support}(u)$, and since we assume our rewards are in $[0, 1]$. Furthermore, we have that

$$\mathbb{E}_{r_t}\left[\left(\frac{(r(s_t, a_t) - r_t)\, u_{s_t a_t}}{N(s_t, a_t)}\right)^2\right] \leq \mathbb{E}_{r_t}\left[\left(\frac{u_{s_t a_t}}{N(s_t, a_t)}\right)^2\right] = \left(\frac{u_{s_t a_t}}{N(s_t, a_t)}\right)^2.$$

By Bernstein's inequality, we therefore have that with probability at least $1 - \delta$:

$$\begin{aligned}\left|(r - \widehat{r})^\top u\right| &\leq \sqrt{2 \sum_t \left(\frac{u_{s_t a_t}}{N(s_t, a_t)}\right)^2 \cdot \log\frac{2}{\delta}} + \frac{4}{3\sqrt{\underline{\lambda}}} \cdot \sqrt{\sum_t \frac{u_{s_t a_t}^2}{N(s_t, a_t)}} \cdot \log\frac{2}{\delta} \\ &= \left(\sqrt{2 \log\frac{2}{\delta}} + \frac{4}{3\sqrt{\underline{\lambda}}} \log\frac{2}{\delta}\right) \cdot \sqrt{\sum_{s,a} \frac{u_{sa}^2}{N(s,a)}}.\end{aligned}$$

□

Lemma 9. *Let $u \in \mathbb{R}^S$ be any vector such that $\forall s, |u_s| \leq M$. Then, for any (ℓ, h), the following holds with probability $(1 - \delta)$:*

$$\left|\mathbb{E}_{s \sim w_{\ell,h}^{\bar{\pi}}}[u_s] - \mathbb{E}_{s \sim \widehat{w}_{\ell,h}^{\bar{\pi}}}[u_s]\right| \leq \sqrt{\frac{2\mathbb{E}_{s \sim w_{\ell,h}^{\bar{\pi}}}[u_s^2]}{\bar{n}_\ell} \log\left(\frac{2}{\delta}\right)} + \frac{2M}{3\bar{n}_\ell} \log\left(\frac{2}{\delta}\right)$$

Proof. The left side of the inequality above takes the form of the deviation between an empirical and true mean of the random variable u_s. Hence, the result follows directly from Bernstein's inequality since we know $|u_s| \leq M$ is bounded. □

Lemma 10. *Assume that A and B are matrices with entries in $[0, 1]$ and whose rows sum to a value ≤ 1. Then AB also satisfies this.*

Proof. To see this, consider the ith row of AB, and note that the sum of the elements in this row can be written as, for $a_i^\top$ the ith row of A, and b_j the jth column of B:

$$\sum_j a_i^\top b_j = \sum_k \sum_j a_{ik} b_{jk} = \sum_k a_{ik} (\sum_j b_{jk}).$$

Now note that $\sum_j b_{jk}$ is the sum across the kth row of B, so this is ≤ 1 by assumption. Furthermore, $\sum_k a_{ik} \leq 1$ for the same reason. Thus, the ith row of AB sums to a value ≤ 1. Furthermore, it is easy to see $a_i^\top b_j \leq 1$ for each j. Thus, AB has values in $[0, 1]$ and rows that sum to a value ≤ 1. □

https://doi.org/10.52202/079017-0717

Lemma 11. *We have that* $\|\Pi_{h=i}^{j} M_{h+1} P_h \boldsymbol{\pi}_h\|_2, \|\Pi_{h=i}^{j} P_h \boldsymbol{\pi}_h\|_2 \le \sqrt{S}$ *for any* i, j, h.

Proof. By definition $P_h \boldsymbol{\pi}_h$ is a transition matrix—each row has values in $[0,1]$ and sums to 1—and M_{h+1} is diagonal with diagonal elements either 0 or 1. Thus, each matrix $M_h P_h \boldsymbol{\pi}_h$ has values in $[0,1]$ and rows that sum to a value ≤ 1, so Lemma 10 implies that $\Pi_{h=i}^{j} M_{h+1} P_h \boldsymbol{\pi}_h$ does as well. Denote $A := \|\Pi_{h=i}^{j} M_h P_h \boldsymbol{\pi}_h\|_2$. We can then bound

$$\|\Pi_{h=i}^{j} M_{h+1} P_h \boldsymbol{\pi}_h\|_2^2 = \|A\|_2^2 \le \|A\|_{\mathrm{F}}^2 = \sum_i \sum_j A_{ij}^2 \le \sum_i 1 \le S,$$

which proves the result. The bound on $\|\Pi_{h=i}^{j} P_h \boldsymbol{\pi}_h\|_2$ follows from the same argument. □

Lemma 12. *We have*

$$\widetilde{\delta}_{\ell,h+1}^{\pi} - \widehat{\delta}_{\ell,h+1}^{\pi} = \sum_{i=0}^{h-2} \left(\prod_{j=h-i+1}^{h} M_{\ell,j+1} P_j \boldsymbol{\pi}_j \right) (P_{h-i} - \widehat{P}_{\ell,h-i}) M_{\ell,h-i} \Big[(\boldsymbol{\pi}_{h-i} - \bar{\boldsymbol{\pi}}_{\ell,h-i}) \widehat{w}_{\ell,h-i}^{\bar{\pi}} + \boldsymbol{\pi}_{h-i} \widehat{\delta}_{\ell,h-i}^{\pi} \Big].$$

Proof. This follows immediately from the definition of $\widetilde{\delta}_{\ell,h+1}^{\pi}, \widehat{\delta}_{\ell,h+1}^{\pi}$, and simple manipulations. □

C.3 Concentration Arguments and Good Events

Lemma 13. *Let* $\mathcal{E}_{\mathrm{prune}}^{\ell}$ *be the event for which the call to* PRUNE *in epoch* ℓ *in Algorithm 2 will terminate after running for at most*

$$\mathrm{poly}(S, A, H, \log \frac{SAH\ell}{\delta \epsilon_\ell}) \cdot \frac{1}{\epsilon_{\mathrm{unif}}^{\ell}}$$

episodes and will return a set $\mathcal{S}_{\ell}^{\mathrm{keep}}$ *such that, for every* $(s,h) \in \mathcal{S}_{\ell}^{\mathrm{keep}}$, *we have* $W_h^{\star}(s) \ge \epsilon_{\mathrm{unif}}^{\ell}$, *and, if* $(s,h) \notin \mathcal{S}_{\ell}^{\mathrm{keep}}$, *then* $W_h^{\star}(s) \le 32\epsilon_{\mathrm{unif}}^{\ell}$. *Then* $\mathbb{P}(\mathcal{E}_{\mathrm{prune}}^{\ell}) \ge 1 - \frac{\delta}{3\ell^2}$.

Proof. From Lemma 38, this event follows directly with probability $(1 - \frac{\delta}{3\ell^2})$. □

Lemma 14. *Let* $\mathcal{E}_{\mathrm{exp}}^{\ell,h}$ *be the event for which:*

1. *The exploration procedure in Algorithm 3 will produce* $\mathfrak{D}_{\ell,h}^{\mathrm{ED}}$ *such that*

$$\max_{\pi \in \Pi_\ell} \|M_{\ell,h}((\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h}) \widehat{w}_{\ell,h}^{\bar{\pi}} + \boldsymbol{\pi}_h \widehat{\delta}_{\ell,h}^{\pi})\|_{\widehat{\Lambda}_{\ell,h}^{-1}}^2 \le \epsilon_{\mathrm{exp}}^{\ell} \quad \textit{for} \quad \widehat{\Lambda}_{\ell,h} = \sum_{(s,a) \in \mathfrak{D}_{\ell,h}^{\mathrm{ED}}} e_{sa} e_{sa}^{\top}, \tag{C.4}$$

   and will collect at most

$$C \cdot \frac{\inf_{\pi_{\mathrm{exp}}} \max_{\pi \in \Pi_\ell} \|M_{\ell,h}((\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h}) \widehat{w}_{\ell,h}^{\bar{\pi}} + \boldsymbol{\pi}_h \widehat{\delta}_{\ell,h}^{\pi})\|_{\Lambda_h(\pi_{\mathrm{exp}})^{-1}}^2}{\epsilon_{\mathrm{exp}}^{\ell}} + \frac{C_{\mathrm{fw}}^{\ell}}{(\epsilon_{\mathrm{exp}}^{\ell})^{4/5}} + \frac{C_{\mathrm{fw}}^{\ell}}{\epsilon_{\mathrm{unif}}^{\ell}} + \log(C_{\mathrm{fw}}^{\ell}) \cdot K_{\mathrm{unif}}^{\ell}$$

   episodes.

2. *For each* $s \in \mathcal{S}_{\ell}^{\mathrm{keep}}$, *we have that* $\sum_{(s',a') \in \mathfrak{D}_{\ell,h}^{\mathrm{ED}}} \mathbb{I}\{(s',a') = (s,a)\} \ge \frac{K_{\mathrm{unif}}^{\ell} \epsilon_{\mathrm{unif}}^{\ell}}{SA}$ *for any* $a \in \mathcal{A}$.

Above, C *is a universal constant and* $C_{\mathrm{fw}}^{\ell} = \mathrm{poly}(S, A, H, \log \ell/\delta, \log 1/\epsilon, \log |\Pi|)$. *Then* $\mathbb{P}[(\mathcal{E}_{\mathrm{exp}}^{\ell,h})^c \cap \mathcal{E}_{\mathrm{prune}}^{\ell} \cap \bar{\mathcal{E}}_{\mathrm{est}}^{\ell} \cap (\cap_{h' \le h-1} \mathcal{E}_{\mathrm{est}}^{\ell,h'}) \cap (\cap_{h' \le h-1} \mathcal{E}_{\mathrm{exp}}^{\ell,h'})] \le \frac{\delta}{6H\ell^2}$.

Proof. Since the event $\mathcal{E}^\ell_{\text{prune}}$ holds, for each $s \in \mathcal{S}^{\text{keep}}_\ell$ we have $W^\star_h(s) \geq \epsilon^\ell_{\text{unif}}$. Now, observe that, for $s \in \mathcal{S}^{\text{keep}}_\ell$ and any a:

$$\begin{aligned}&|[(\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h'})\widehat{w}^{\bar{\pi}}_{\ell,h} + \boldsymbol{\pi}_h \widehat{\delta}^{\pi}_{\ell,h}]_{(s,a)}| \\ &\leq [\widehat{w}^{\bar{\pi}}_{\ell,h}]_s + |[\widehat{\delta}^{\pi}_{\ell,h}]_s| \leq [w^{\bar{\pi}}_{\ell,h}]_s + |[\delta^{\pi}_{\ell,h}]_s| + |[\widehat{w}^{\bar{\pi}}_{\ell,h} - w^{\bar{\pi}}_{\ell,h}]_{(s)}| + |[\delta^{\pi}_{\ell,h}]_s - |[\widehat{\delta}^{\pi}_{\ell,h}]_s||.\end{aligned}$$

By construction, we have $[w^{\bar{\pi}}_{\ell,h}]_s, |[\delta^{\pi}_{\ell,h}]_s| \leq W^\star_h(s)$. By Lemma 19, on $\bar{\mathcal{E}}^\ell_{\text{est}}$, we can bound $|[\widehat{w}^{\bar{\pi}}_{\ell,h} - w^{\bar{\pi}}_{\ell,h}]_{(s)}| \leq \sqrt{8S\epsilon_\ell^{5/3}}$. By Lemma 18, on $\mathcal{E}^\ell_{\text{prune}} \cap (\cap_{h' \leq h-1} \mathcal{E}^{\ell,h'}_{\text{est}}) \cap (\cap_{h' \leq h-1} \mathcal{E}^{\ell,h'}_{\text{exp}})$, we can bound

$$|[\delta^{\pi}_{\ell,h}]_s - |[\widehat{\delta}^{\pi}_{\ell,h}]_s|| \leq \sqrt{SH\beta_\ell \epsilon^\ell_{\text{exp}}} + SH(\sqrt{8\epsilon_\ell^{5/3}} + 32\epsilon^\ell_{\text{unif}}).$$

Altogether then, we have

$$\begin{aligned}&|[(\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h'})\widehat{w}^{\bar{\pi}}_{\ell,h} + \boldsymbol{\pi}_h \widehat{\delta}^{\pi}_{\ell,h}]_{(s,a)}| \\ &\leq 2W^\star_h(s) + \sqrt{SH\beta_\ell \epsilon^\ell_{\text{exp}}} + SH(\sqrt{8\epsilon_\ell^{5/3}} + 32\epsilon^\ell_{\text{unif}}) + \sqrt{8S\epsilon_\ell^{5/3}}.\end{aligned}$$

By our choice of $\epsilon^\ell_{\text{exp}}$ and $\epsilon^\ell_{\text{unif}}$, we can bound all of this as

$$\leq C_{\boldsymbol{\phi}} \cdot (W^\star_h(s) + \sqrt{K^\ell_{\text{unif}} \epsilon^\ell_{\text{unif}}} \epsilon^\ell_{\text{exp}})$$

for $C_{\boldsymbol{\phi}} = cSH\beta_\ell$. This is the condition required by Theorem 2, so the result follows from Theorem 2. □

Lemma 15. *Let $\mathcal{E}^{\ell,h}_{\text{est}}$ be the event at epoch ℓ for step h on which:*

(1) For all $\pi \in \Pi_\ell$, $h' \leq h$:

$$\begin{aligned}&\left|\left\langle \boldsymbol{\pi}_h^\top \widetilde{r}_{\ell,h}, \left(\prod_{i=h'+1}^{h} M_{\ell,i+1} P_i \boldsymbol{\pi}_i\right)(P_{h'} - \widehat{P}_{\ell,h'}) M_{\ell,h'}\left[(\boldsymbol{\pi}_{h'} - \bar{\boldsymbol{\pi}}_{\ell,h'})\widehat{w}^{\bar{\pi}}_{\ell,h'} + \boldsymbol{\pi}_{h'}\widehat{\delta}^{\pi}_{\ell,h'}\right]\right\rangle\right| \\ &\leq \beta_\ell \sqrt{\sum_{s,a} \frac{\left[M_{\ell,h'}\left((\boldsymbol{\pi}_{h'} - \bar{\boldsymbol{\pi}}_{\ell,h'})\widehat{w}^{\bar{\pi}}_{\ell,h'} + \boldsymbol{\pi}_{h'}\widehat{\delta}^{\pi}_{\ell,h'}\right)\right]^2_{(s,a)}}{N_{\ell,h'}(s,a)}}.\end{aligned}$$

(2) For all canonical vectors $e_{s'}$ in $\mathbb{R}^S$, $\pi \in \Pi_\ell$, and $h' \leq h$,

$$\begin{aligned}&\left|\left\langle e_{s'}, \left(\prod_{i=h'+1}^{h} M_{\ell,i+1} P_i \boldsymbol{\pi}_i\right)(P_{h'} - \widehat{P}_{\ell,h'}) M_{\ell,h'}\left[(\boldsymbol{\pi}_{h'} - \bar{\boldsymbol{\pi}}_{\ell,h'})\widehat{w}^{\bar{\pi}}_{\ell,h'} + \boldsymbol{\pi}_{h'}\widehat{\delta}^{\pi}_{\ell,h'}\right]\right\rangle\right| \\ &\leq \beta_\ell \sqrt{\sum_{s,a} \frac{[M_{\ell,h'}((\boldsymbol{\pi}_{h'} - \bar{\boldsymbol{\pi}}_{\ell,h'})\widehat{w}^{\bar{\pi}}_{\ell,h'} + \boldsymbol{\pi}_{h'}\widehat{\delta}^{\pi}_{\ell,h'})]^2_{s,a}}{N_{\ell,h'}(s,a)}}.\end{aligned}$$

(3) For each (s,a), we have

$$\sum_{s'} |\widehat{P}_{\ell,h}(s' \mid s,a) - P_h(s' \mid s,a)| \leq S\sqrt{\frac{\log \frac{48S^2AH\ell^2}{\delta}}{N_{\ell,h}(s,a)}}.$$

(4) For each $\pi \in \Pi_\ell$,

$$\begin{aligned}&|\langle \widehat{r}_{\ell,h} - \widetilde{r}_{\ell,h}, \boldsymbol{\pi}_h \widehat{\delta}^{\pi}_{\ell,h} + (\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h})\widehat{w}^{\bar{\pi}}_{\ell,h}\rangle| \\ &\leq \beta_\ell \sqrt{\sum_{s,a} \frac{\left[M_{\ell,h}\left((\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h'})\widehat{w}^{\bar{\pi}}_{\ell,h} + \boldsymbol{\pi}_h \widehat{\delta}^{\pi}_{\ell,h}\right)\right]^2_{(s,a)}}{N_{\ell,h}(s,a)}}.\end{aligned}$$

Then $\mathbb{P}[(\mathcal{E}^{\ell,h}_{\text{est}})^c \cap \mathcal{E}^\ell_{\text{prune}} \cap (\cap_{h' \leq h} \mathcal{E}^{\ell,h}_{\text{exp}})] \leq \frac{\delta}{6H\ell^2}$.

Proof. We prove each of the events sequentially.

https://doi.org/10.52202/079017-0717

Proof of Event (1). Consider any fixed choice of (π, h'). By Lemma 10 and since our rewards are in $[0,1]$, we have that $\left(\prod_{i=h'+1}^{h} M_{\ell,i+1} P_i \boldsymbol{\pi}_i\right)^\top \boldsymbol{\pi}_h^\top \widetilde{r}_{\ell,h}$ is a vector in $[0,1]$. Let $v \leftarrow \left(\prod_{i=h'+1}^{h} M_{\ell,i+1} P_i \boldsymbol{\pi}_i\right)^\top \boldsymbol{\pi}_h^\top \widetilde{r}_{\ell,h}$ and $u \leftarrow M_{\ell,h'}\left[(\boldsymbol{\pi}_{h'} - \bar{\boldsymbol{\pi}}_{\ell,h'})\widehat{w}^{\bar{\pi}}_{\ell,h'} + \boldsymbol{\pi}_{h'}\widehat{\delta}^{\pi}_{\ell,h'}\right]$. Note that by construction we have that $u_{sa} = 0$ for $s \notin \mathcal{S}^{\text{keep}}_{\ell,h'}$, and so on $\mathcal{E}^{\ell,h'}_{\text{exp}}$, we have $N_{\ell,h'}(s,a) \geq \frac{K^\ell_{\text{unif}}\epsilon^\ell_{\text{unif}}}{2SA}$ for all $(s,a) \in \text{support}(u)$. On $\mathcal{E}^\ell_{\text{prune}} \cap \mathcal{E}^{\ell,h'}_{\text{exp}}$, we can then apply Lemma 7 with u and v as defined above to get that the bound fails with probability at most $\frac{\delta}{30H^2\ell^2|\Pi_\ell|}$. Union bounding over h' and π we get that the stated result fails with probability at most $\frac{\delta}{30H\ell^2}$.

Proof of Event (2). Choose

$$v = e_i^\top \left(\prod_{i=h'+1}^{h} M_{\ell,i} P_i \boldsymbol{\pi}_i\right) \quad \text{and} \quad u = M_{h',\ell}\left((\boldsymbol{\pi}_{h'} - \bar{\boldsymbol{\pi}}_{\ell,h'})w^{\bar{\pi}}_{\ell,h'} + \boldsymbol{\pi}_{h'}\widehat{\delta}^{\pi}_{\ell,h'}\right).$$

Note that by construction of $w^{\bar{\pi}}_{\ell,h'}$ and $\widehat{\delta}^{\pi}_{\ell,h'}$ we have that $u_{sa} = 0$ for $s \notin \mathcal{S}^{\text{keep}}_{\ell,h'}$, and so on $\mathcal{E}^{\ell,h'}_{\text{exp}}$, we have $N_{\ell,h'}(s,a) \geq \frac{K^\ell_{\text{unif}}\epsilon^\ell_{\text{unif}}}{2SA}$ for all $(s,a) \in \text{support}(u)$. Furthermore, we have that $v \in [0,1]^S$ by Lemma 10. Then, the event follows by invoking Lemma 7.

Proof of Event (3). By Hoeffding's inequality, for any (s,a), we have, with probability at least $1 - \frac{\delta}{24S^2AH\ell^2}$:

$$|\widehat{P}_{\ell,h}(s' \mid s,a) - P_h(s' \mid s,a)| \leq \sqrt{\frac{\log\frac{24S^2AH\ell^2}{\delta}}{N_{\ell,h}(s,a)}}.$$

Thus, we have that with probability at least $1 - \frac{\delta}{24SAH\ell^2}$:

$$\sum_{s'} |\widehat{P}_{\ell,h}(s' \mid s,a) - P_h(s' \mid s,a)| \leq S\sqrt{\frac{\log\frac{24S^2AH\ell^2}{\delta}}{N_{\ell,h}(s,a)}}.$$

Union bounding over all (s,a), we obtain that this holds with probability at least $1 - \frac{\delta}{24H\ell^2}$.

Proof of Event (4). Note first that $\langle \widehat{r}_{\ell,h} - \widetilde{r}_{\ell,h}, \boldsymbol{\pi}_h\widehat{\delta}^{\pi}_{\ell,h} + (\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h})\widehat{w}^{\bar{\pi}}_{\ell,h}\rangle = \langle \widehat{r}_{\ell,h} - \widetilde{r}_{\ell,h}, M_{\ell,h}(\boldsymbol{\pi}_h\widehat{\delta}^{\pi}_{\ell,h} + (\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h})\widehat{w}^{\bar{\pi}}_{\ell,h})\rangle$. The result then follows on $\mathcal{E}^\ell_{\text{prune}}$ by a direct application of Lemma 8.

The final result then holds by a union bound. □

Lemma 16. *Let $\bar{\mathcal{E}}^\ell_{\text{est}}$ denote the event that at epoch ℓ and for each h:*

(1) For all $\pi \in \Pi_\ell$ and $h \in [H]$, we have

$$\left|\langle P_h^\top M_{\ell,h+1}\widetilde{V}_{\ell,h+1} + r_h, (\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h})(w^{\bar{\pi}}_{\ell,h} - \widehat{w}^{\bar{\pi}}_{\ell,h})\rangle\right| \leq \frac{2H}{3\bar{n}_\ell}\log\frac{60H\ell^2|\Pi_\ell|}{\delta}$$
$$+ \sqrt{\frac{2\mathbb{E}_{s\sim w^{\bar{\pi}_\ell}_{\ell,h}}[\langle P_h^\top M_{\ell,h+1}\widetilde{V}^\pi_{\ell,h+1} + r_h, (\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h})e_s\rangle^2]}{\bar{n}_\ell}\cdot\log\frac{60H\ell^2|\Pi_\ell|}{\delta}}.$$

(2) For all canonical vectors $e_s \in \mathbb{R}^S$,

$$|\langle e_s, \widehat{w}^{\bar{\pi}}_{\ell,h} - w^{\bar{\pi}}_{\ell,h}\rangle| \leq \sqrt{\frac{2\log\left(\frac{30H\ell^2S}{\delta}\right)}{\bar{n}_\ell}} + \frac{2\log\left(\frac{30H\ell^2S}{\delta}\right)}{\bar{n}_\ell}.$$

Then $\mathbb{P}[(\bar{\mathcal{E}}^\ell_{\text{est}})^c] \leq \frac{\delta}{15\ell^2}$.

Proof. **Proof of Event (1).** Consider a fixed choice of π, and let $u_s^\pi = \left\langle P_h^\top \widetilde{V}_{\ell,h+1}^\pi + r_h, (\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h}) e_s \right\rangle$, and note that $|u_s^\pi| \le H$ for all s. Lemma 9 then gives that with probability at least $1 - \frac{\delta}{30H\ell^2|\Pi_\ell|}$ we have

$$\left|\langle P_h^\top M_{\ell,h+1} \widetilde{V}_{\ell,h+1} + r_h, (\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h})(w_{\ell,h}^{\bar{\pi}} - \widehat{w}_{\ell,h}^{\bar{\pi}})\rangle\right|$$

$$\le \sqrt{\frac{2\mathbb{E}_{s\sim w_{\ell,h}^{\bar{\pi}_\ell}}[\langle P_h^\top M_{\ell,h+1}\widetilde{V}_{\ell,h+1}^\pi + r_h, (\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h}) e_s\rangle^2]}{\bar{n}_\ell} \cdot \log\frac{60H\ell^2|\Pi_\ell|}{\delta}} + \frac{2H}{3\bar{n}_\ell}\log\frac{60H\ell^2|\Pi_\ell|}{\delta}.$$

Proof of Event (2). For a fixed choice of $s \in [S]$, the event follows from Lemma 9 with $u = e_s$ with probability $1-\delta$, where $\delta = \frac{\delta}{30H\ell^2 S}$. Once we take the union bound over all $s \in [S]$, then the event follows with probability $1 - \frac{\delta}{30H\ell^2}$.

The result then holds by union bounding over each of these for all h. □

Lemma 17. *On $\mathcal{E}_{\text{prune}}^\ell$, for all h and π we have*

$$\delta_{\ell,h+1}^\pi - \widetilde{\delta}_{\ell,h+1}^\pi = \sum_{i=0}^{h-2}\left(\prod_{j=h-i+1}^{h} M_{\ell,j+1}P_j\boldsymbol{\pi}_j\right) M_{\ell,h-i+1}P_{h-i}(\boldsymbol{\pi}_{h-i} - \bar{\boldsymbol{\pi}}_{h-i})(w_{\ell,h-i}^{\bar{\pi}_\ell} - \widehat{w}_{\ell,h-i}^{\bar{\pi}_\ell}) + \Delta_{\ell,h+1}^\pi$$

for some $\Delta_{\ell,h}^\pi \in \mathbb{R}^S$ with $\|\Delta_{\ell,h}^\pi\|_2 \le 32SH\epsilon_{\text{unif}}^\ell$. Furthermore, for any π and any i,k satisfying $0 \le i \le k \le H$, we have

$$\left\|\left(\prod_{j=i}^{k} M_{\ell,j+1}P_j\boldsymbol{\pi}_j - \prod_{j=i}^{k} P_j\boldsymbol{\pi}_j\right) w_i^\pi\right\|_2 \le 32SH\epsilon_{\text{unif}}^\ell.$$

Proof. By definition, we have that

$$\begin{aligned}
&\delta_{\ell,h+1}^\pi - \widetilde{\delta}_{\ell,h+1}^\pi \\
&= P_h(\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h})w_{\ell,h}^{\bar{\pi}_\ell} + P_h\boldsymbol{\pi}_h\delta_{\ell,h}^\pi - M_{\ell,h+1}P_h(\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h})\widehat{w}_{\ell,h}^{\bar{\pi}} - M_{\ell,h+1}P_h\boldsymbol{\pi}_h\widetilde{\delta}_{\ell,h}^\pi \\
&= (I - M_{\ell,h+1})P_h(\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h})w_{\ell,h}^{\bar{\pi}_\ell} + M_{\ell,h+1}P_h(\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h})(w_{\ell,h}^{\bar{\pi}_\ell} - \widehat{w}_{\ell,h}^{\bar{\pi}}) \\
&\quad + (I - M_{\ell,h+1})P_h\boldsymbol{\pi}_h\delta_{\ell,h}^\pi + M_{\ell,h+1}P_h\boldsymbol{\pi}_h(\delta_{\ell,h}^\pi - \widetilde{\delta}_{\ell,h}^\pi) \\
&\vdots \\
&= \sum_{i=0}^{h-2}\left(\prod_{j=h-i+1}^{h} M_{\ell,j+1}P_j\boldsymbol{\pi}_j\right)\Big[(I - M_{\ell,h-i+1})P_{h-i}(\boldsymbol{\pi}_{h-i} - \bar{\boldsymbol{\pi}}_{h-i})w_{\ell,h-i}^{\bar{\pi}_\ell} \\
&\quad + M_{\ell,h-i+1}P_{h-i}(\boldsymbol{\pi}_{h-i} - \bar{\boldsymbol{\pi}}_{h-i})(w_{\ell,h-i}^{\bar{\pi}_\ell} - \widehat{w}_{\ell,h-i}^{\bar{\pi}_\ell}) + (I - M_{\ell,h-i+1})P_{h-i}\boldsymbol{\pi}_{h-i}\delta_{\ell,h-i}^\pi\Big].
\end{aligned}$$

Note that $[P_{h-i}(\boldsymbol{\pi}_{h-i} - \bar{\boldsymbol{\pi}}_{h-i})w_{\ell,h'}^{\bar{\pi}_\ell}]_s \le W_{h-i+1}^\star(s)$, and similarly $[P_{h-i}\boldsymbol{\pi}_{h-i}\delta_{\ell,h-i}^\pi]_s \le W_{h-i+1}^\star(s)$. On the event $\mathcal{E}_{\text{prune}}^\ell$, we have that if $[M_{\ell,h-i+1}]_{s,s} = 0$, then $W_{h-i+1}^\star(s) \le 32\epsilon_{\text{unif}}^\ell$. It follows from this that every non-zero element in $(I - M_{\ell,h-i+1})P_{h-i}(\boldsymbol{\pi}_{h-i} - \bar{\boldsymbol{\pi}}_{h-i})w_{\ell,h-i}^{\bar{\pi}_\ell}$ and $(I - M_{\ell,h-i+1})P_{h-i}\boldsymbol{\pi}_{h-i}\delta_{\ell,h-i}^\pi$ is bounded by $32\epsilon_{\text{unif}}^\ell$, so:

$$\|(I - M_{\ell,h-i+1})P_{h-i}(\boldsymbol{\pi}_{h-i} - \bar{\boldsymbol{\pi}}_{h-i})w_{\ell,h-i}^{\bar{\pi}_\ell}\|_2 \le 32\sqrt{S}\epsilon_{\text{unif}}^\ell \text{ and}$$
$$\|(I - M_{\ell,h-i+1})P_{h-i}\boldsymbol{\pi}_{h-i}\delta_{\ell,h-i}^\pi\|_2 \le 32\sqrt{S}\epsilon_{\text{unif}}^\ell.$$

By Lemma 11, we can bound

$$\|\prod_{j=h-i+1}^{h} M_{\ell,j+1}P_j\boldsymbol{\pi}_j\|_2 \le \sqrt{S}.$$

https://doi.org/10.52202/079017-0717

Combining these gives the result.

We now prove the second part of the result. Denote $A_j := M_{\ell,j+1}P_j\boldsymbol{\pi}_j$ and $B_j := P_j\boldsymbol{\pi}_j$. Then

$$\begin{aligned}\prod_{j=i}^{k} M_{\ell,j+1}P_j\boldsymbol{\pi}_j - \prod_{j=i}^{k} P_j\boldsymbol{\pi}_j &= \prod_{j=i}^{k} A_j - \prod_{j=i}^{k} B_j \\ &= A_k\left(\prod_{j=i}^{k-1} A_j - \prod_{j=i}^{k-1} B_j\right) + (A_k - B_k)\prod_{j=i}^{k-1} B_j \\ &\vdots \\ &= \sum_{s=i}^{k}\left(\prod_{j=s+1}^{k} A_j\right)(A_s - B_s)\left(\prod_{j'=i}^{s-1} B_{j'}\right).\end{aligned}$$

By Lemma 11 we have $\|\prod_{j=s+1}^{k} A_j\|_2 \le \sqrt{S}$. Furthermore, note that $\prod_{j'=i}^{s-1} B_{j'} w_i^\pi = w_s^\pi$. So it follows that

$$\left\|(\prod_{j=i}^{k} M_{\ell,j+1}P_j\boldsymbol{\pi}_j - \prod_{j=i}^{k} P_j\boldsymbol{\pi}_j)w_i^\pi\right\|_2 \le \sum_{s=i}^{k}\sqrt{S}\|(A_s - B_s)w_s^\pi\|_2.$$

By the same argument as above, we can bound $\|(A_s - B_s)w_s^\pi\|_2 \le 32\sqrt{S}\epsilon_{\text{unif}}^\ell$. □

Lemma 18. *On the event $\mathcal{E}_{\text{prune}}^\ell \cap (\cap_{h'\le h}\mathcal{E}_{\text{est}}^{\ell,h'}) \cap (\cap_{h'\le h}\mathcal{E}_{\text{exp}}^{\ell,h'})$, we have, for all $\pi \in \Pi_\ell$:*

$$\|\widehat{\delta}_{\ell,h+1}^\pi - \delta_{\ell,h+1}^\pi\|_2 \le \sqrt{SH\beta_\ell\epsilon_{\text{exp}}^\ell} + SH(\sqrt{8\epsilon_\ell^{5/3}} + 32\epsilon_{\text{unif}}^\ell).$$

Proof. We can write

$$\|\widehat{\delta}_{\ell,h+1}^\pi - \delta_{\ell,h+1}^\pi\|_2 \le \|\widehat{\delta}_{\ell,h+1}^\pi - \widetilde{\delta}_{\ell,h+1}^\pi\|_2 + \|\widetilde{\delta}_{\ell,h+1}^\pi - \delta_{\ell,h+1}^\pi\|_2.$$

From Lemma 12 we have

$$\begin{aligned}&\widetilde{\delta}_{\ell,h+1}^\pi - \widehat{\delta}_{\ell,h+1}^\pi \\ &= \sum_{i=0}^{h-2}\left(\prod_{j=h-i+1}^{h} M_{\ell,j+1}P_j\boldsymbol{\pi}_j\right)(P_{h-i} - \widehat{P}_{\ell,h-i})M_{\ell,h-i}\Big[(\boldsymbol{\pi}_{h-i} - \bar{\boldsymbol{\pi}}_{\ell,h-i})\widehat{w}_{\ell,h-i}^{\bar{\pi}} + \boldsymbol{\pi}_{h-i}\widehat{\delta}_{\ell,h-i}^\pi\Big].\end{aligned}$$

From Event (2) of $\mathcal{E}_{\text{est}}^{\ell,h}$ in Lemma 15, we have that for all canonical vectors e_s and $\pi \in \Pi_\ell$:

$$\begin{aligned}&\left\langle e_s, \left(\prod_{j=h-i+1}^{h} M_{\ell,j+1}P_j\boldsymbol{\pi}_j\right)(P_{h-i} - \widehat{P}_{\ell,h-i})M_{\ell,h-i}\Big[(\boldsymbol{\pi}_{h-i} - \bar{\boldsymbol{\pi}}_{h-i})\widehat{w}_{\ell,h-i}^{\bar{\pi}} + \boldsymbol{\pi}_{h-i}\widehat{\delta}_{\ell,h-i}^\pi\Big]\right\rangle \\ &\le \beta_\ell\sqrt{\sum_{s,a}\frac{[M_{\ell,h-i}((\boldsymbol{\pi}_{h-i} - \bar{\boldsymbol{\pi}}_{\ell,h-i})\widehat{w}_{\ell,h-i}^{\bar{\pi}} + \boldsymbol{\pi}_{h-i}\widehat{\delta}_{\ell,h-i}^\pi)]_{s,a}^2}{N_{\ell,h-i}(s,a)}}.\end{aligned}$$

Now, summing over the bound above for all canonical vectors, and applying this for each i, it follows that

$$\|\widehat{\delta}_{\ell,h+1}^\pi - \widetilde{\delta}_{\ell,h+1}^\pi\|_2^2 \le S\beta_\ell^2\sum_{h'=1}^{h}\sum_{s,a}\frac{[M_{\ell,h'}((\boldsymbol{\pi}_{h'} - \bar{\boldsymbol{\pi}}_{\ell,h'})\widehat{w}_{\ell,h'}^{\bar{\pi}} + \boldsymbol{\pi}_{h'}\widehat{\delta}_{\ell,h'}^\pi)]_{s,a}^2}{N_{\ell,h'}(s,a)} \le SH\beta_\ell\epsilon_{\text{exp}}^\ell$$

where the last inequality holds on $\cap_{h'\le h}\mathcal{E}_{\text{exp}}^{\ell,h'}$.

We now turn to bounding $\|\widetilde{\delta}^{\pi}_{\ell,h+1} - \delta^{\pi}_{\ell,h+1}\|_2$. By Lemma 17 we have

$$\delta^{\pi}_{\ell,h+1} - \widetilde{\delta}^{\pi}_{\ell,h+1}$$
$$= \sum_{i=0}^{h-2} \left(\prod_{j=h-i+1}^{h} M_{\ell,j+1} P_j \boldsymbol{\pi}_j \right) M_{\ell,h-i+1} P_{h-i} (\boldsymbol{\pi}_{h-i} - \bar{\boldsymbol{\pi}}_{h-i}) (w^{\bar{\pi}_\ell}_{\ell,h-i} - \widehat{w}^{\bar{\pi}_\ell}_{\ell,h-i}) + \Delta^{\pi}_{\ell,h+1}$$

for some $\Delta^{\pi}_{\ell,h} \in \mathbb{R}^S$ with $\|\Delta^{\pi}_{\ell,h}\|_2 \le 32SH\epsilon^{\ell}_{\text{unif}}$. Furthermore, on $\mathcal{E}^{\ell,h-i}_{\text{est}}$, by Lemma 19 we can bound

$$\|w^{\bar{\pi}_\ell}_{\ell,h-i} - \widehat{w}^{\bar{\pi}_\ell}_{\ell,h-i}\|_2 \le \sqrt{8S\epsilon^{5/3}_{\ell}}.$$

Combining this with Lemma 11 gives the result.

□

Lemma 19. *On event $\bar{\mathcal{E}}^{\ell}_{\text{est}}$ we have:*

$$\|\widehat{w}^{\bar{\pi}}_{\ell,h} - w^{\bar{\pi}}_{\ell,h}\|_2^2 \le 8S\epsilon^{5/3}_{\ell}.$$

Proof. From Event (2) of Lemma 16, we have that for all canonical vectors $e_i \in \mathbb{R}^S$:

$$|\langle e_i, \widehat{w}^{\bar{\pi}}_{\ell,h} - w^{\bar{\pi}}_{\ell,h}\rangle| \le \sqrt{\frac{2\log\left(\frac{30H\ell^2 S}{\delta}\right)}{\bar{n}_\ell}} + \frac{2\log\left(\frac{30H\ell^2 S}{\delta}\right)}{\bar{n}_\ell}.$$

Then, combining these bounds together for all s:

$$\|\widehat{w}^{\bar{\pi}}_{\ell,h} - w^{\bar{\pi}}_{\ell,h}\|_2^2 \le \frac{4S\log\left(\frac{30H\ell^2 S}{\delta}\right)}{\bar{n}_\ell} + \frac{4S\log^2\left(\frac{30H\ell^2 S}{\delta}\right)}{\bar{n}^2_\ell} \le 4S\epsilon^{5/3}_{\ell} + 4S\epsilon^{10/3}_{\ell} \le 8S\epsilon^{5/3}_{\ell},$$

where the last inequality follows from our choice of $\bar{n}_\ell$ in Algorithm 2.

□

Lemma 20. *Let* $\mathcal{E}_{\text{good}} := (\cap_{\ell=1}^{\infty}\mathcal{E}^{\ell}_{\text{prune}}) \cap (\cap_{\ell=1}^{\infty}\bar{\mathcal{E}}^{\ell}_{\text{est}}) \cap (\cap_{\ell=1}^{\infty} \cap_{h\in[H]} \mathcal{E}^{\ell,h}_{\text{est}}) \cap (\cap_{\ell=1}^{\infty} \cap_{h\in[H]} \mathcal{E}^{\ell,h}_{\text{exp}})$. *Then* $\mathbb{P}[\mathcal{E}_{\text{good}}] \ge 1 - 2\delta$.

Proof. By a union bound and basic set manipulations, we have:

$$\mathbb{P}[\mathcal{E}^c_{\text{good}}] \le \sum_{\ell=1}^{\infty}\mathbb{P}[(\mathcal{E}^{\ell}_{\text{prune}})^c] + \sum_{\ell=1}^{\infty}\mathbb{P}[(\bar{\mathcal{E}}^{\ell}_{\text{est}})^c]$$
$$+ \sum_{\ell=1}^{\infty}\sum_{h=1}^{H}\mathbb{P}[(\mathcal{E}^{\ell,h}_{\text{exp}})^c \cap \mathcal{E}^{\ell}_{\text{prune}} \cap \bar{\mathcal{E}}^{\ell}_{\text{est}} \cap (\cap_{h'\le h-1}\mathcal{E}^{\ell,h'}_{\text{est}}) \cap (\cap_{h'\le h-1}\mathcal{E}^{\ell,h'}_{\text{exp}})]$$
$$+ \sum_{\ell=1}^{\infty}\sum_{h=1}^{H}\mathbb{P}[(\mathcal{E}^{\ell,h}_{\text{est}})^c \cap \mathcal{E}^{\ell}_{\text{prune}} \cap (\cap_{h'\le h}\mathcal{E}^{\ell,h}_{\text{exp}})].$$

By Lemma 13, we have $\mathbb{P}[(\mathcal{E}^{\ell}_{\text{prune}})^c] \le \delta/3\ell^2$. By By Lemma 16, we have $\mathbb{P}[(\bar{\mathcal{E}}^{\ell}_{\text{est}})^c] \le \frac{\delta}{15\ell^2}$. By Lemma 14, we have $\mathbb{P}[(\mathcal{E}^{\ell,h}_{\text{exp}})^c \cap \mathcal{E}^{\ell}_{\text{prune}} \cap \bar{\mathcal{E}}^{\ell}_{\text{est}} \cap (\cap_{h'\le h-1}\mathcal{E}^{\ell,h'}_{\text{est}}) \cap (\cap_{h'\le h-1}\mathcal{E}^{\ell,h'}_{\text{exp}})] \le \frac{\delta}{6H\ell^2}$. By Lemma 15 we have $\mathbb{P}[(\mathcal{E}^{\ell,h}_{\text{est}})^c \cap \mathcal{E}^{\ell}_{\text{prune}} \cap (\cap_{h'\le h}\mathcal{E}^{\ell,h}_{\text{exp}})] \le \frac{\delta}{6H\ell^2}$. Putting this together we can bound the above as

$$\le \sum_{\ell=1}^{\infty}\left(\frac{\delta}{3\ell^2} + \frac{\delta}{15\ell^2}\right) + \sum_{\ell=1}^{\infty}\sum_{h=1}^{H}\frac{2\delta}{6H\ell^2} \le 2\delta.$$

□

https://doi.org/10.52202/079017-0717

C.4 Estimation of Reference Policy and Values

Lemma 21. *On $\mathcal{E}_{\text{good}}$ we have that:*

$$\left|\sum_{h=1}^{H}\langle\widetilde{r}_{\ell,h},\boldsymbol{\pi}_h(\widetilde{\delta}^{\pi}_{\ell,h}-\widehat{\delta}^{\pi}_{\ell,h})\rangle\right|\le\epsilon_\ell \quad \textit{and} \quad \sum_{h=1}^{H}|\langle\widehat{r}_{\ell,h}-\widetilde{r}_{\ell,h},\boldsymbol{\pi}_h\widehat{\delta}^{\pi}_{\ell,h}+(\boldsymbol{\pi}_h-\bar{\boldsymbol{\pi}}_{\ell,h})\widehat{w}^{\bar{\pi}}_{\ell,h}\rangle|\le\epsilon_\ell. \tag{C.5}$$

Proof. From Lemma 12 we have:

$$\begin{aligned}&\widetilde{\delta}^{\pi}_{\ell,h+1}-\widehat{\delta}^{\pi}_{\ell,h+1}\\&=\sum_{i=0}^{h-2}\left(\prod_{j=h-i+1}^{h}M_{\ell,j+1}P_j\boldsymbol{\pi}_j\right)(P_{h-i}-\widehat{P}_{\ell,h-i})M_{\ell,h-i}\Big[(\boldsymbol{\pi}_{h-i}-\bar{\boldsymbol{\pi}}_{\ell,h-i})\widehat{w}^{\bar{\pi}}_{\ell,h-i}+\boldsymbol{\pi}_{h-i}\widehat{\delta}^{\pi}_{\ell,h-i}\Big].\end{aligned}$$

A sufficient condition for (C.5) is that, for each i:

$$\begin{aligned}\left|\left\langle\boldsymbol{\pi}_h^{\top}\widetilde{r}_{\ell,h},\left(\prod_{j=h-i+1}^{h}M_{\ell,j+1}P_j\boldsymbol{\pi}_j\right)(P_{h-i}-\widehat{P}_{\ell,h-i})\right.\right.\\\left.\left.M_{\ell,h-i}\Big[(\boldsymbol{\pi}_{h-i}-\bar{\boldsymbol{\pi}}_{\ell,h-i})\widehat{w}^{\bar{\pi}}_{\ell,h-i}+\boldsymbol{\pi}_{h-i}\widehat{\delta}^{\pi}_{\ell,h-i}\Big]\right\rangle\right|\le\epsilon_\ell.\end{aligned}$$

On $\mathcal{E}_{\text{good}}$, and in particular $\mathcal{E}^{\ell,h}_{\text{est}}$ (Lemma 15), we can bound the left-hand side of this as:

$$\begin{aligned}&\le\beta_\ell\sqrt{\sum_{s,a}\frac{\left[M_{\ell,h-i}\left((\boldsymbol{\pi}_{h-i}-\bar{\boldsymbol{\pi}}_{\ell,h-i})\widehat{w}^{\bar{\pi}}_{\ell,h-i}+\boldsymbol{\pi}_{h-i}\widehat{\delta}^{\pi}_{\ell,h-i}\right)\right]^2_{(s,a)}}{N_{\ell,h-i}(s,a)}}\\&\le\beta_\ell\sqrt{\epsilon_\ell^2/H^4\beta_\ell^2}\\&\le\epsilon_\ell/H^2\end{aligned}$$

where the second inequality holds on $\mathcal{E}_{\text{good}}$ (in particular $\mathcal{E}^{\ell,h-i}_{\text{exp}}$). This proves the first inequality.

On $\mathcal{E}^{\ell,h}_{\text{est}}$ we can also bound

$$\begin{aligned}&|\langle\widehat{r}_{\ell,h}-\widetilde{r}_{\ell,h},\boldsymbol{\pi}_h\widehat{\delta}^{\pi}_{\ell,h}+(\boldsymbol{\pi}_h-\bar{\boldsymbol{\pi}}_{\ell,h})\widehat{w}^{\bar{\pi}}_{\ell,h}\rangle|\\&\le\beta_\ell\sqrt{\sum_{s,a}\frac{\left[M_{\ell,h}\left((\boldsymbol{\pi}_h-\bar{\boldsymbol{\pi}}_{\ell,h'})\widehat{w}^{\bar{\pi}}_{\ell,h}+\boldsymbol{\pi}_h\widehat{\delta}^{\pi}_{\ell,h}\right)\right]^2_{(s,a)}}{N_{\ell,h}(s,a)}}\\&\le\epsilon_\ell/H^2.\end{aligned}$$

This proves the second inequality.

□

Lemma 22. *On event $\mathcal{E}_{\text{good}}$, for any timestep h, policies π,π', and action a, we have:*

$$\mathbb{E}_{\pi'}[|\widehat{Q}^{\pi}_{\ell,h}(s_h,a)-Q^{\pi}_h(s_h,a)|]\le H^2S^{3/2}\sqrt{A\log\frac{24S^2AH\ell^2}{\delta}}\cdot\epsilon_\ell^{1/3}+64H^2S\epsilon^{\ell}_{\text{unif}}. \tag{C.6}$$

Proof. By Lemma E.15 of [10], we have that:

$$\begin{aligned}&\widehat{Q}^{\pi}_{\ell,h}(s,a)-Q^{\pi}_h(s,a)\\&=\mathbb{E}_\pi\left[\sum_{h'=h}^{H}\sum_{s'}(\widehat{P}_{\ell,h'}(s'\mid s_{h'},a_{h'})-P_h(s'\mid s_{h'},a_{h'}))\widehat{V}^{\pi}_{\ell,h'+1}(s_{h'})\mid s_h=s,a_h=a\right].\end{aligned}$$

On $\mathcal{E}_{\mathrm{good}}$, in particular $\mathcal{E}_{\mathrm{est}}^{\ell,h'}$, we can bound, for $s \in \mathcal{S}_{\ell,h'}^{\mathrm{keep}}$ and any a:

$$\left|\sum_{s'}(\widehat{P}_{\ell,h'}(s' \mid s,a) - P_h(s' \mid s,a))\widehat{V}^{\pi}_{\ell,h'+1}(s')\right|$$
$$\leq SH\sqrt{\frac{\log\frac{24S^2AH\ell^2}{\delta}}{N_{\ell,h'}(s,a)}} \leq SH\sqrt{\frac{SA\log\frac{24S^2AH\ell^2}{\delta}}{K_{\mathrm{unif}}^{\ell}\epsilon_{\mathrm{unif}}^{\ell}}}$$

and where the last inequality follows on $\mathcal{E}_{\mathrm{exp}}^{\ell,h'}$. By our choice of K_{unif}^{ℓ} and $\epsilon_{\mathrm{unif}}^{\ell}$, we can further bound this as

$$\leq SH\sqrt{SA\log\frac{24S^2AH\ell^2}{\delta}}\cdot\epsilon_{\ell}^{1/3}.$$

For $s \notin \mathcal{S}_{\ell,h'}^{\mathrm{keep}}$, we can bound $|\sum_{s'}(\widehat{P}_{\ell,h'}(s' \mid s,a) - P_h(s' \mid s,a))\widehat{V}^{\pi}_{\ell,h'}(s_{h'})| \leq 2H$. We therefore have that

$$\begin{aligned}
&\mathbb{E}_{\pi'}[|\widehat{Q}^{\pi}_{\ell,h}(s_h,a) - Q^{\pi}_h(s_h,a)|] \\
&\leq \mathbb{E}_{\pi'}\Bigg[\mathbb{E}_{\pi}\Bigg[\sum_{h'=h}^{H} SH\sqrt{SA\log\frac{24S^2AH\ell^2}{\delta}}\cdot\epsilon_{\ell}^{1/3}\cdot\mathbb{I}\{s_{h'}\in\mathcal{S}_{\ell,h'}^{\mathrm{keep}}\} \\
&\qquad + 2H\mathbb{I}\{s_{h'}\notin\mathcal{S}_{\ell,h'}^{\mathrm{keep}}\} \mid s_h = s, a_h = a\Bigg]\Bigg] \\
&= \sum_{h'=h}^{H}\mathbb{E}_{\widetilde{\pi}}\left[SH\sqrt{SA\log\frac{24S^2AH\ell^2}{\delta}}\cdot\epsilon_{\ell}^{1/3}\cdot\mathbb{I}\{s_{h'}\in\mathcal{S}_{\ell,h'}^{\mathrm{keep}}\} + 2H\mathbb{I}\{s_{h'}\notin\mathcal{S}_{\ell,h'}^{\mathrm{keep}}\}\right] \\
&\leq H^2S^{3/2}\sqrt{A\log\frac{24S^2AH\ell^2}{\delta}}\cdot\epsilon_{\ell}^{1/3} + 64H^2S\epsilon_{\mathrm{unif}}^{\ell},
\end{aligned}$$

where the last inequality follows by definition of $\mathcal{S}_{\ell,h'}^{\mathrm{keep}}$, and π' is the policy which plays $\bar{\pi}_{\ell}$ for the first h steps and then plays π. This proves the result. □

Lemma 23. *On event $\mathcal{E}_{\mathrm{good}}$, for all h and any π and π', we have that*

$$|\widehat{U}_{\ell,h}(\pi,\pi') - U_h(\pi,\pi')| \leq 9H^3S^{3/2}\sqrt{A\log\frac{24S^2AH\ell^2}{\delta}}\cdot\epsilon_{\ell}^{1/3} + 576H^3S\epsilon_{\mathrm{unif}}^{\ell}.$$

Proof. We have

$$\widehat{U}_{\ell,h}(\pi,\pi') = \mathbb{E}_{\pi',\ell}\left[\left(\widehat{Q}^{\pi}_{\ell,h}(s_h,\pi_h(s_h)) - \widehat{Q}^{\pi}_{\ell,h}(s_h,\pi'_h(s_h))\right)^2\right]$$

where $\mathbb{E}_{\pi',\ell}$ denotes the expectation induced playing policy π' on the MDP with transition $\widehat{P}_{\ell}$. We can think of this as simply a value function for policy π on the reward $\check{r}_h(s,a) = \left(\widehat{Q}^{\pi}_{\ell,h}(s,\pi_h(s)) - \widehat{Q}^{\pi}_{\ell,h}(s,a)\right)^2$. Let $\check{V}$ denote the value function on this reward on $\widehat{P}_{\ell}$, and note that $\check{V}_h(s) \in [0,H^2]$ for all (s,h). By Lemma E.15 of [10], we then have that

$$\begin{aligned}
&\left|\widehat{U}_{\ell,h}(\pi,\pi') - \mathbb{E}_{\pi'}\left[\left(\widehat{Q}^{\pi}_{\ell,h}(s_h,\pi_h(s_h)) - \widehat{Q}^{\pi}_{\ell,h}(s_h,\pi'_h(s_h))\right)^2\right]\right| \\
&= \mathbb{E}_{\pi'}\left[\sum_{h=1}^{H}\sum_{s'}(\widehat{P}_{\ell,h}(s' \mid s_h,a_h) - P_h(s' \mid s_h,a_h))\check{V}_{h+1}(s')\right] \\
&\leq H^2\sum_{h=1}^{H}\mathbb{E}_{\pi'}\left[\sum_{s'}|\widehat{P}_{\ell,h}(s' \mid s_h,a_h) - P_h(s' \mid s_h,a_h)|\right].
\end{aligned}$$

https://doi.org/10.52202/079017-0717

Note that we always have $\sum_{s'} |\widehat{P}_{\ell,h}(s' \mid s_h, a_h) - P_h(s' \mid s_h, a_h)| \le 2$. Furthermore, on $\mathcal{E}_{\mathrm{good}}$ we also have $\sum_{s'} |\widehat{P}_{\ell,h}(s' \mid s_h, a_h) - P_h(s' \mid s_h, a_h)| \le S\sqrt{\frac{\log \frac{24S^2AH\ell^2}{\delta}}{N_{\ell,h}(s_h,a_h)}}$. We can therefore bound the above as

$$\le H^2 \sum_{h=1}^{H} \mathbb{E}_{\pi'} \left[\min\left\{ 2, S\sqrt{\frac{\log \frac{24S^2AH\ell^2}{\delta}}{N_{\ell,h}(s_h,a_h)}} \right\} \right]$$
$$\le H^2 \sum_{h=1}^{H} \mathbb{E}_{\pi'} \left[2 \cdot \mathbb{I}\{s_h \notin \mathcal{S}^{\mathrm{keep}}_{\ell,h}\} + S\sqrt{\frac{\log \frac{24S^2AH\ell^2}{\delta}}{N_{\ell,h}(s_h,a_h)}} \cdot \mathbb{I}\{s_h \in \mathcal{S}^{\mathrm{keep}}_{\ell,h}\} \right].$$

For $s \in \mathcal{S}^{\mathrm{keep}}_{\ell,h}$, on $\mathcal{E}_{\mathrm{good}}$ we have $N_{\ell,h}(s_h,a_h) \ge \frac{K^{\ell}_{\mathrm{unif}}\epsilon^{\ell}_{\mathrm{unif}}}{SA} = \epsilon_\ell^{2/3}/SA$, and we also have for $s_h \notin \mathcal{S}^{\mathrm{keep}}_{\ell,h}$ that $W^\star_h(s) \le 32\epsilon^{\ell}_{\mathrm{unif}}$. Putting this together we can bound the above as

$$\le H^2 \sum_{h=1}^{H} \left[64S\epsilon^{\ell}_{\mathrm{unif}} + S\sqrt{SA \log \frac{24S^2AH\ell^2}{\delta}} \cdot \epsilon_\ell^{1/3} \right]$$
$$\le 64SH^3\epsilon^{\ell}_{\mathrm{unif}} + H^3S^{3/2}\sqrt{A \log \frac{24S^2AH\ell^2}{\delta}} \cdot \epsilon_\ell^{1/3}.$$

Furthermore,

$$\begin{aligned}
&\left| \mathbb{E}_{\pi'} \left[\left(\widehat{Q}^{\pi}_{\ell,h}(s,\pi_h(s)) - \widehat{Q}^{\pi}_{\ell,h}(s,\pi'_h(s)) \right)^2 \right] - \mathbb{E}_{\pi'} \left[(Q^{\pi}_h(s,\pi_h(s)) - Q^{\pi}_h(s,\pi'_h(s)))^2 \right] \right| \\
&= \Bigg| \mathbb{E}_{\pi'} \left[\left(\widehat{Q}^{\pi}_{\ell,h}(s,\pi_h(s)) - Q^{\pi}_h(s,\pi_h(s)) + Q^{\pi}_h(s,\pi'_h(s)) - \widehat{Q}^{\pi}_{\ell,h}(s,\pi'_h(s)) \right)^2 \right] \\
&\quad + \mathbb{E}_{\pi'} \Bigg[\left(\widehat{Q}^{\pi}_{\ell,h}(s,\pi_h(s)) - Q^{\pi}_h(s,\pi_h(s)) + Q^{\pi}_h(s,\pi'_h(s)) - \widehat{Q}^{\pi}_{\ell,h}(s,\pi'_h(s)) \right) \\
&\qquad (Q^{\pi}_h(s,\pi_h(s)) - Q^{\pi}_h(s,\pi'_h(s))) \Bigg] \Bigg| \\
&\le 4H\mathbb{E}_{\pi'}[|\widehat{Q}^{\pi}_{\ell,h}(s,\pi_h(s)) - Q^{\pi}_h(s,\pi_h(s))|] + 4H\mathbb{E}_{\pi'}[|Q^{\pi}_h(s,\pi'_h(s)) - \widehat{Q}^{\pi}_{\ell,h}(s,\pi'_h(s))|] \\
&\le 8H^3S^{3/2}\sqrt{A \log \frac{24S^2AH\ell^2}{\delta}} \cdot \epsilon_\ell^{1/3} + 512H^3S\epsilon^{\ell}_{\mathrm{unif}}
\end{aligned}$$

where the final inequality follows from Lemma 22. Combining this with the above bound completes the argument. □

Lemma 24. *On event $\mathcal{E}_{\mathrm{good}}$, for all epochs ℓ, we have that*

$$\left| \sum_{h=1}^{H} \langle \widetilde{r}_{\ell,h}, \boldsymbol{\pi}_h(\delta^{\pi}_{\ell,h} - \widetilde{\delta}^{\pi}_{\ell,h}) \rangle + \langle \widetilde{r}_{\ell,h}, (\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h})(w^{\bar{\pi}}_{\ell,h} - \widehat{w}^{\bar{\pi}}_{\ell,h}) \rangle \right| \le \epsilon_\ell. \tag{C.7}$$

Proof. We first bound $|\langle M_{\ell,h} r_h, \boldsymbol{\pi}_h(\delta^{\pi}_{\ell,h} - \widetilde{\delta}^{\pi}_{\ell,h}) \rangle|$. By Lemma 17 we have that

$$\begin{aligned}
&\delta^{\pi}_{\ell,h+1} - \widetilde{\delta}^{\pi}_{\ell,h+1} \\
&= \sum_{i=0}^{h-2} \left(\prod_{j=h-i+1}^{h} M_{\ell,j+1} P_j \boldsymbol{\pi}_j \right) M_{\ell,h-i+1} P_{h-i} (\boldsymbol{\pi}_{h-i} - \bar{\boldsymbol{\pi}}_{\ell,h-i})(w^{\bar{\pi}}_{h-i} - \widehat{w}^{\bar{\pi}}_{\ell,h-i}) + \Delta^{\pi}_{\ell,h+1}
\end{aligned}$$

for some $\Delta^{\pi}_{\ell,h} \in \mathbb{R}^S$ with $\|\Delta^{\pi}_{\ell,h}\|_2 \le 32SH\epsilon^{\ell}_{\mathrm{unif}}$. Furthermore, note that

$$\begin{aligned}
&\sum_{h=1}^{H}\sum_{i=0}^{h-2}\left\langle \widetilde{r}_{\ell,h}, \boldsymbol{\pi}_h \left(\prod_{j=h-i+1}^{h} M_{\ell,j+1}P_j\boldsymbol{\pi}_j\right) M_{\ell,h-i+1}P_{h-i}(\boldsymbol{\pi}_{h-i}-\bar{\boldsymbol{\pi}}_{\ell,h-i})(w^{\bar{\pi}}_{h-i}-\widehat{w}^{\bar{\pi}}_{\ell,h-i})\right\rangle \\
&= \sum_{h=1}^{H}\sum_{k=2}^{h}\left\langle \widetilde{r}_{\ell,h}, \boldsymbol{\pi}_h \left(\prod_{j=k+1}^{h} M_{\ell,j+1}P_j\boldsymbol{\pi}_j\right) M_{\ell,k+1}P_k(\boldsymbol{\pi}_{k}-\bar{\boldsymbol{\pi}}_{\ell,k})(w^{\bar{\pi}}_{k}-\widehat{w}^{\bar{\pi}}_{\ell,k})\right\rangle \\
&= \sum_{k=2}^{H}\sum_{h=k}^{H}\left\langle \widetilde{r}_{\ell,h}, \boldsymbol{\pi}_h \left(\prod_{j=k+1}^{h} M_{\ell,j+1}P_j\boldsymbol{\pi}_j\right) M_{\ell,k+1}P_k(\boldsymbol{\pi}_{k}-\bar{\boldsymbol{\pi}}_{\ell,k})(w^{\bar{\pi}}_{k}-\widehat{w}^{\bar{\pi}}_{\ell,k})\right\rangle \\
&= \sum_{k=2}^{H}\langle P_k^\top M_{\ell,k+1}\widetilde{V}_{\ell,k+1}, (\boldsymbol{\pi}_k-\bar{\boldsymbol{\pi}}_{\ell,k})(w^{\bar{\pi}}_{k}-\widehat{w}^{\bar{\pi}}_{\ell,k})\rangle.
\end{aligned}$$

It follows that

$$\begin{aligned}
&\sum_{h=1}^{H}\langle \widetilde{r}_{\ell,h}, \boldsymbol{\pi}_h(\delta^{\pi}_{\ell,h}-\widetilde{\delta}^{\pi}_{\ell,h})\rangle + \langle \widetilde{r}_{\ell,h}, (\boldsymbol{\pi}_h-\bar{\boldsymbol{\pi}}_{\ell,h})(w^{\bar{\pi}}_{h}-\widehat{w}^{\bar{\pi}}_{\ell,h})\rangle \\
&= \sum_{h=2}^{H}\langle P_h^\top M_{\ell,h+1}\widetilde{V}_{\ell,h+1} + \widetilde{r}_{\ell,h}, (\boldsymbol{\pi}_h-\bar{\boldsymbol{\pi}}_{\ell,h})(w^{\bar{\pi}}_{\ell,h}-\widehat{w}^{\bar{\pi}}_{\ell,h})\rangle + \Delta
\end{aligned}$$

for some Δ satisfying $|\Delta| \le 32S^{3/2}H^2\epsilon^{\ell}_{\mathrm{unif}}$. On $\mathcal{E}_{\mathrm{good}}$ (specifically $\bar{\mathcal{E}}^{\ell}_{\mathrm{est}}$), we can bound

$$\begin{aligned}
&\sum_{h=2}^{H}|\langle P_h^\top M_{\ell,h+1}\widetilde{V}_{\ell,h+1} + \widetilde{r}_{\ell,h}, (\boldsymbol{\pi}_h-\bar{\boldsymbol{\pi}}_{\ell,h})(w^{\bar{\pi}}_{\ell,h}-\widehat{w}^{\bar{\pi}}_{\ell,h})\rangle| \\
&\le \sum_{h=2}^{H}\sqrt{\frac{2\mathbb{E}_{s\sim w^{\bar{\pi}}_{\ell,h}}[\langle P_h^\top M_{\ell,h+1}\widetilde{V}^{\pi}_{\ell,h+1} + \widetilde{r}_{\ell,h}, (\boldsymbol{\pi}_h-\bar{\boldsymbol{\pi}}_{\ell,h})e_s\rangle^2]}{\bar{n}_\ell}\cdot\log\frac{60H\ell^2|\Pi_\ell|}{\delta}} \\
&\quad + \frac{2H}{3\bar{n}_\ell}\log\frac{60H^2\ell^2|\Pi_\ell|}{\delta}
\end{aligned}$$

We can also bound

$$\begin{aligned}
&\mathbb{E}_{s\sim w^{\bar{\pi}}_{\ell,h}}[\langle P_h^\top M_{\ell,h+1}\widetilde{V}^{\pi}_{\ell,h+1} + \widetilde{r}_{\ell,h}, (\boldsymbol{\pi}_h-\bar{\boldsymbol{\pi}}_{\ell,h})e_s\rangle^2] \\
&\le 2\mathbb{E}_{s\sim w^{\bar{\pi}}_{\ell,h}}[\langle P_h^\top V^{\pi}_{h+1} + r_h, (\boldsymbol{\pi}_h-\bar{\boldsymbol{\pi}}_{\ell,h})e_s\rangle^2] + 2H\mathbb{E}_{s\sim w^{\bar{\pi}}_{\ell,h}}[|[\boldsymbol{\pi}_h^\top P_h^\top(M_{\ell,h+1}\widetilde{V}^{\pi}_{\ell,h+1} - V^{\pi}_{h+1})]_s|] \\
&\quad + 2H\mathbb{E}_{s\sim w^{\bar{\pi}}_{\ell,h}}[|[\bar{\boldsymbol{\pi}}_{\ell,h}^\top P_h^\top(M_{\ell,h+1}\widetilde{V}^{\pi}_{\ell,h+1} - V^{\pi}_{h+1})]_s|] + 4\mathbb{E}_{s\sim w^{\bar{\pi}}_{\ell,h}}[\sup_a |r_h(s,a)-\widetilde{r}_{\ell,h}(s,a)|]
\end{aligned}$$

Furthermore,

$$\begin{aligned}
&\mathbb{E}_{s\sim w^{\bar{\pi}}_{\ell,h}}[|[\boldsymbol{\pi}_h^\top P_h^\top(M_{\ell,h+1}\widetilde{V}^{\pi}_{\ell,h+1} - V^{\pi}_{h+1})]_s|] \\
&= \sum_s |[\boldsymbol{\pi}_h^\top P_h^\top(M_{\ell,h+1}\widetilde{V}^{\pi}_{\ell,h+1} - V^{\pi}_{h+1})]_s| w^{\bar{\pi}}_{\ell,h}(s) \\
&\le \sqrt{S}\|(M_{\ell,h+1}\widetilde{V}^{\pi}_{\ell,h+1} - V^{\pi}_{h+1})^\top P_h\boldsymbol{\pi}_h w^{\bar{\pi}}_{\ell,h}\|_2 \\
&\le \sqrt{S}\|(\widetilde{V}^{\pi}_{\ell,h+1} - V^{\pi}_{h+1})^\top P_h\boldsymbol{\pi}_h w^{\bar{\pi}}_{\ell,h}\|_2 + \sqrt{S}\|(M_{\ell,h+1}\widetilde{V}^{\pi}_{\ell,h+1} - \widetilde{V}^{\pi}_{\ell,h+1})^\top P_h\boldsymbol{\pi}_h w^{\bar{\pi}}_{\ell,h}\|_2 \\
&\le 64S^2H^2\epsilon^{\ell}_{\mathrm{unif}}
\end{aligned}$$

where the last inequality follows from the definition of $\widetilde{V}$ and Lemma 17. A similar bound can be shown for $\mathbb{E}_{s\sim w^{\bar{\pi}}_{\ell,h}}[|[\bar{\boldsymbol{\pi}}_{\ell,h}^\top P_h^\top(M_{\ell,h+1}\widetilde{V}^{\pi}_{\ell,h+1} - V^{\pi}_{h+1})]_s|]$. In addition, by definition of $\widetilde{r}_{\ell,h}$ we have

$$\mathbb{E}_{s\sim w^{\bar{\pi}}_{\ell,h}}[\sup_a |r_h(s,a)-\widetilde{r}_{\ell,h}(s,a)|] \le \mathbb{E}_{s\sim w^{\bar{\pi}}_{\ell,h}}[\mathbb{I}\{s\notin\mathcal{S}^{\mathrm{keep}}_{\ell,h}\}] \le 32S\epsilon^{\ell}_{\mathrm{unif}}.$$

https://doi.org/10.52202/079017-0717

Thus, we have

$$\begin{aligned}
&\sum_{h=2}^{H} |\langle P_h^\top M_{\ell,h+1} \widetilde{V}_{\ell,h+1} + r_h, (\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h})(w_{\ell,h}^{\bar{\pi}} - \widehat{w}_{\ell,h}^{\bar{\pi}})\rangle| \\
&\le \sum_{h=2}^{H} \sqrt{\frac{2\mathbb{E}_{s\sim w_{\ell,h}^{\bar{\pi}}}[\langle P_h^\top M_{\ell,h+1} \widetilde{V}_{\ell,h+1}^{\pi} + r_h, (\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h}) e_s\rangle^2]}{\bar{n}_\ell} \cdot \log \frac{60H\ell^2|\Pi_\ell|}{\delta}} \\
&\quad + \frac{2H}{3\bar{n}_\ell} \log \frac{60H^2\ell^2|\Pi_\ell|}{\delta} \\
&\le \sum_{h=2}^{H} \sqrt{\frac{4U_h(\pi, \bar{\pi}_\ell) + 384S^2H^3\epsilon_{\mathrm{unif}}^\ell}{\bar{n}_\ell} \cdot \log \frac{60H\ell^2|\Pi_\ell|}{\delta}} + \frac{2H}{3\bar{n}_\ell} \log \frac{60H^2\ell^2|\Pi_\ell|}{\delta}.
\end{aligned}$$

By Lemma 23 and Jensen's inequality, this can be further bounded as

$$\begin{aligned}
&\le \sum_{h=2}^{H} c\sqrt{\frac{\widehat{U}_{\ell-1,h}(\pi, \bar{\pi}_\ell) + S^{3/2}H^3\sqrt{A\log\frac{24S^2AH\ell^2}{\delta}} \cdot \epsilon_\ell^{1/3} + S^2H^3\epsilon_{\mathrm{unif}}^\ell}{\bar{n}_\ell} \cdot \log \frac{60H\ell^2|\Pi_\ell|}{\delta}} \\
&\quad + \frac{2H}{3\bar{n}_\ell} \log \frac{60H^2\ell^2|\Pi_\ell|}{\delta} \\
&\le c\sqrt{\frac{H\widehat{U}_{\ell-1}(\pi, \bar{\pi}_\ell) + S^{3/2}H^4\sqrt{A\log\frac{24S^2AH\ell^2}{\delta}} \cdot \epsilon_\ell^{1/3} + S^2H^4\epsilon_{\mathrm{unif}}^\ell}{\bar{n}_\ell} \cdot \log \frac{60H\ell^2|\Pi_\ell|}{\delta}} \\
&\quad + \frac{2H}{3\bar{n}_\ell} \log \frac{60H^2\ell^2|\Pi_\ell|}{\delta}.
\end{aligned}$$

The result then follows from this, our choice of $\bar{n}_\ell$ and $\epsilon_{\mathrm{unif}}^\ell$, and the bound on Δ above.

□

Lemma 25. *On $\mathcal{E}_{\mathrm{good}}$, we can bound*

$$\begin{aligned}
&\frac{\inf_{\pi_{\mathrm{exp}}} \max_{\pi\in\Pi_\ell} \|M_{\ell,h}((\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h})\widehat{w}_{\ell,h}^{\bar{\pi}} + \boldsymbol{\pi}_h\widehat{\delta}_{\ell,h}^{\pi})\|_{\Lambda_h(\pi_{\mathrm{exp}})^{-1}}^2}{\epsilon_{\mathrm{exp}}^\ell} \\
&\le \frac{\inf_{\pi_{\mathrm{exp}}} \max_{\pi\in\Pi_\ell} 4\|\bar{\boldsymbol{\pi}}_{\ell,h} w_{\ell,h}^{\bar{\pi}} - \boldsymbol{\pi}_h w_h^{\pi}\|_{\Lambda_h(\pi_{\mathrm{exp}})^{-1}}^2}{\epsilon_{\mathrm{exp}}^\ell} \\
&\quad + \frac{(8S^2A + 32S^3AH^2)\epsilon_\ell^{5/3} + 2S^2AH\beta_\ell\epsilon_{\mathrm{exp}}^\ell + 4096S^3AH^2(\epsilon_{\mathrm{unif}}^\ell)^2}{\epsilon_{\mathrm{unif}}^\ell\epsilon_{\mathrm{exp}}^\ell}.
\end{aligned}$$

Proof. We can bound:

$$\begin{aligned}
&\inf_{\pi_{\mathrm{exp}}} \max_{\pi\in\Pi_\ell} \|M_{\ell,h}((\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h})\widehat{w}_{\ell,h}^{\bar{\pi}} + \boldsymbol{\pi}_h\widehat{\delta}_{\ell,h}^{\pi})\|_{\Lambda_h(\pi_{\mathrm{exp}})^{-1}}^2 \\
&\le \inf_{\pi_{\mathrm{exp}}} \max_{\pi\in\Pi_\ell} 4\|M_{\ell,h}((\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h})w_{\ell,h}^{\bar{\pi}} + \boldsymbol{\pi}_h\delta_{\ell,h}^{\pi})\|_{\Lambda_h(\pi_{\mathrm{exp}})^{-1}}^2 \\
&\quad + \inf_{\pi'_{\mathrm{exp}}} \max_{\pi\in\Pi_\ell} \Big[8\|M_{\ell,h}(\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h})(w_{\ell,h}^{\bar{\pi}} - \widehat{w}_{\ell,h}^{\bar{\pi}})\|_{\Lambda_h(\pi'_{\mathrm{exp}})^{-1}}^2 \\
&\quad + 8\|M_{\ell,h}\boldsymbol{\pi}_h(\delta_{\ell,h}^{\pi} - \widehat{\delta}_{\ell,h}^{\pi})\|_{\Lambda_h(\pi'_{\mathrm{exp}})^{-1}}^2\Big].
\end{aligned}$$

We can write

$$\begin{aligned}
&\|M_{\ell,h}(\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h})(w^{\bar{\pi}}_{\ell,h} - \widehat{w}^{\bar{\pi}}_{\ell,h})\|^2_{\Lambda_h(\pi'_{\mathrm{exp}})^{-1}} \\
&= \sum_{s,a} \frac{(\boldsymbol{\pi}_h(a \mid s) - \bar{\boldsymbol{\pi}}_{\ell,h}(a \mid s))^2 (w^{\bar{\pi}}_{\ell,h}(s) - \widehat{w}^{\bar{\pi}}_{\ell,h}(s))^2}{[\Lambda_h(\pi'_{\mathrm{exp}})]_{sa,sa}} \cdot \mathbb{I}\{(s,a) \in \mathcal{S}^{\mathrm{keep}}_{\ell,h}\} \\
&\le \sum_{s,a} \frac{(w^{\bar{\pi}}_{\ell,h}(s) - \widehat{w}^{\bar{\pi}}_{\ell,h}(s))^2}{[\Lambda_h(\pi'_{\mathrm{exp}})]_{sa,sa}} \cdot \mathbb{I}\{(s,a) \in \mathcal{S}^{\mathrm{keep}}_{\ell,h}\}.
\end{aligned}$$

On $\mathcal{E}_{\mathrm{good}}$, for each $(s,a) \in \mathcal{S}^{\mathrm{keep}}_{\ell,h}$ we have $W^\star_h(s) \ge \epsilon^\ell_{\mathrm{unif}}$. Let π^{sh} denote the policy which achieves $w^{\pi^{sh}}_h(s) = W^\star_h(s)$, and then plays actions uniformly at random at (s,h). Let $\pi'_{\mathrm{exp}} = \mathrm{unif}(\{\pi^{sh}\}_s)$. Then we have $[\Lambda_h(\pi'_{\mathrm{exp}})]_{sa,sa} \ge W^\star_h(s)/SA \ge \epsilon^\ell_{\mathrm{unif}}/SA$ for each $(s,a) \in \mathcal{S}^{\mathrm{keep}}_{\ell,h}$, so we can bound the above as

$$\le \frac{SA}{\epsilon^\ell_{\mathrm{unif}}} \sum_{s,a} (w^{\bar{\pi}}_{\ell,h}(s) - \widehat{w}^{\bar{\pi}}_{\ell,h}(s))^2 = \frac{SA}{\epsilon^\ell_{\mathrm{unif}}} \|w^{\bar{\pi}}_{\ell,h} - \widehat{w}^{\bar{\pi}}_{\ell,h}\|_2^2 \le \frac{8S^2A\epsilon_\ell^{5/3}}{\epsilon^\ell_{\mathrm{unif}}},$$

where the last inequality follows from Lemma 19.

We can obtain a bound on $\|M_{\ell,h}\boldsymbol{\pi}_h(\delta^\pi_{\ell,h} - \widehat{\delta}^\pi_{\ell,h})\|^2_{\Lambda_h(\pi'_{\mathrm{exp}})^{-1}}$ using a similar argument but now applying Lemma 18 to get that:

$$\|M_{\ell,h}\boldsymbol{\pi}_h(\delta^\pi_{\ell,h} - \widehat{\delta}^\pi_{\ell,h})\|^2_{\Lambda_h(\pi'_{\mathrm{exp}})^{-1}} \le \frac{2S^2AH\beta_\ell\epsilon^\ell_{\mathrm{exp}}}{\epsilon^\ell_{\mathrm{unif}}} + \frac{32S^3AH^2\epsilon_\ell^{5/3}}{\epsilon^\ell_{\mathrm{unif}}} + 4096S^3AH^2\epsilon^\ell_{\mathrm{unif}}.$$

Finally, note that

$$\begin{aligned}
\|M_{\ell,h}((\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h})w^{\bar{\pi}}_{\ell,h} + \boldsymbol{\pi}_h\delta^\pi_{\ell,h})\|^2_{\Lambda_h(\pi_{\mathrm{exp}})^{-1}} &= \|M_{\ell,h}(\bar{\boldsymbol{\pi}}_{\ell,h}w^{\bar{\pi}}_{\ell,h} + \boldsymbol{\pi}_h w^\pi_h)\|^2_{\Lambda_h(\pi_{\mathrm{exp}})^{-1}} \\
&\le \|\bar{\boldsymbol{\pi}}_{\ell,h}w^{\bar{\pi}}_{\ell,h} - \boldsymbol{\pi}_h w^\pi_h\|^2_{\Lambda_h(\pi_{\mathrm{exp}})^{-1}}
\end{aligned}$$

where the equality holds by definition, and the inequality by simply manipulations. Combining these bounds gives the result. □

C.5 Correctness and Sample Complexity

Lemma 26. *On the event $\mathcal{E}_{\mathrm{good}}$, for all $\pi \in \Pi_{\ell+1}$, we have $V^\star_0(\Pi) - V^\pi_0 \le 16\epsilon_\ell$, and $\pi^\star \in \Pi_\ell$.*

Proof. Recall $D_{\bar{\pi}_\ell}(\pi) = V^\pi_0 - V^{\bar{\pi}_\ell}_0$. For $\pi \in \Pi_\ell$, we have

$$\begin{aligned}
&|\widehat{D}_{\bar{\pi}_\ell}(\pi) - D_{\bar{\pi}_\ell}(\pi)| \\
&= \left|\sum_{h=1}^H \left[\langle \widehat{r}_{\ell,h}, \boldsymbol{\pi}_h\widehat{\delta}^\pi_{\ell,h} + (\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h})\widehat{w}^{\bar{\pi}}_{\ell,h}\rangle - \langle r_h, \boldsymbol{\pi}_h\delta^\pi_{\ell,h} + (\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h})w^{\bar{\pi}}_{\ell,h}\rangle\right]\right| \\
&\le \underbrace{\sum_{h=1}^H |\langle \widehat{r}_{\ell,h} - \widetilde{r}_{\ell,h}, \boldsymbol{\pi}_h\widehat{\delta}^\pi_{\ell,h} + (\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h})\widehat{w}^{\bar{\pi}}_{\ell,h}\rangle|}_{(a)} + \underbrace{\sum_{h=1}^H |\langle \widetilde{r}_{\ell,h}, \boldsymbol{\pi}_h(\widetilde{\delta}^\pi_{\ell,h} - \widehat{\delta}^\pi_{\ell,h})\rangle|}_{(b)} \\
&\quad + \underbrace{\left|\sum_{h=1}^H \langle \widetilde{r}_{\ell,h}, \boldsymbol{\pi}_h(\delta^\pi_{\ell,h} - \widetilde{\delta}^\pi_{\ell,h})\rangle + \langle r_h, (\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h})(w^{\bar{\pi}}_{\ell,h} - \widehat{w}^{\bar{\pi}}_{\ell,h})\rangle\right|}_{(c)} \\
&\quad + \underbrace{\sum_{h=1}^H |\langle \widetilde{r}_{\ell,h} - r_h, \boldsymbol{\pi}_h\delta^\pi_{\ell,h} + (\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h})w^{\bar{\pi}}_{\ell,h}\rangle|}_{(d)}.
\end{aligned}$$

https://doi.org/10.52202/079017-0717

By Lemma 21, on $\mathcal{E}_{\mathrm{good}}$ we have $(a) \le \epsilon_\ell$ and $(b) \le \epsilon_\ell$, and by Lemma 24, $(c) \le \epsilon_\ell$. To bound (d), we note that $\boldsymbol{\pi}_h \delta^\pi_{\ell,h} + (\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h}) w^{\bar{\pi}}_{\ell,h} = \boldsymbol{\pi}_h w^\pi_h - \bar{\boldsymbol{\pi}}_{\ell,h} w^{\bar{\pi}}_{\ell,h}$, and so, on $\mathcal{E}_{\mathrm{good}}$ and by definition of $\widetilde{r}_{\ell,h}$,

$$(d) \le \sum_{h=1}^{H} \sum_{s \notin \mathcal{S}^{\mathrm{keep}}_{\ell,h}} (w^\pi_h(s) + w^{\bar{\pi}}_{\ell,h}(s)) \le 64 H S \epsilon^\ell_{\mathrm{unif}} \le \epsilon_\ell.$$

Note that we only eliminate policy $\pi \in \Pi_\ell$ at round ℓ if $\max_{\pi'} \widehat{D}_{\bar{\pi}_\ell}(\pi') - \widehat{D}_{\bar{\pi}_\ell}(\pi) > 8\epsilon_\ell$. Assume that $\pi^\star \in \Pi_\ell$. By what we have just shown, if policy π is eliminated, we then have

$$8\epsilon_\ell < \max_{\pi' \in \Pi_\ell} D_{\bar{\pi}_\ell}(\pi') - D_{\bar{\pi}_\ell}(\pi) + 8\epsilon_\ell = V^\star_0 - V^\pi_0 + 8\epsilon_\ell \implies V^\pi_0 < V^\star_0.$$

It follows that $\pi^\star$ will not be eliminated at round ℓ, as long as $\pi^\star \in \Pi_\ell$. By a simple inductive argument, since $\pi^\star \in \Pi_0$, it follows that on $\mathcal{E}_{\mathrm{good}}$, $\pi^\star \in \Pi_\ell$ for all ℓ.

Furthermore, for each $\pi \in \Pi_{\ell+1}$, we have $\max_{\pi'} \widehat{D}_{\bar{\pi}_\ell}(\pi') - \widehat{D}_{\bar{\pi}_\ell}(\pi) \le 8\epsilon_\ell$. Which, again by what we have just shown, implies that

$$8\epsilon_\ell \ge \max_{\pi' \in \Pi_\ell} D_{\bar{\pi}_\ell}(\pi') - D_{\bar{\pi}_\ell}(\pi) - 8\epsilon_\ell = V^\star_0 - V^\pi_0 - 8\epsilon_\ell \implies V^\star_0 - V^\pi_0 \le 16\epsilon_\ell.$$

□

Theorem 1. *There exists an algorithm (Algorithm 1) which, with probability at least* $1 - 2\delta$*, finds an* ϵ*-optimal policy and terminates after collecting at most*

$$\sum_{h=1}^{H} \inf_{\pi_{\mathrm{exp}}} \max_{\pi \in \Pi} \frac{H^4 \|\phi^\star_h - \phi^\pi_h\|^2_{\Lambda_h(\pi_{\mathrm{exp}})^{-1}}}{\max\{\epsilon^2, \Delta(\pi)^2\}} \cdot \iota\beta^2 + \max_{\pi \in \Pi} \frac{HU(\pi, \pi^\star)}{\max\{\epsilon^2, \Delta(\pi)^2\}} \log \tfrac{H|\Pi|\iota}{\delta} + \frac{C_{\mathrm{poly}}}{\max\{\epsilon^{\frac{5}{3}}, \Delta^{\frac{5}{3}}_{\min}\}}$$

episodes, for $C_{\mathrm{poly}} := \mathrm{poly}(S, A, H, \log 1/\delta, \iota, \log|\Pi|)$, $\beta := C\sqrt{\log(\frac{SH|\Pi|}{\delta} \cdot \frac{1}{\Delta_{\min} \vee \epsilon})}$ *and* $\iota := \log \frac{1}{\Delta_{\min} \vee \epsilon}$.

Proof. First, by Lemma 20, we have that $\mathbb{P}[\mathcal{E}_{\mathrm{good}}] \ge 1 - 2\delta$. For the remainder of the proof we assume we are on $\mathcal{E}_{\mathrm{good}}$.

By Lemma 26, we have that on $\mathcal{E}_{\mathrm{good}}$, for every $\pi \in \Pi_{\ell+1}$, $V^\star_0 - V^\pi_0 \le 16\epsilon_\ell$, and that $\pi^\star \in \Pi_\ell$ for all ℓ. It follows that, since we run for $\ell_\epsilon = \lceil \log_2 16/\epsilon \rceil$ epochs, when we terminate each policy $\pi \in \Pi_{\ell_\epsilon}$ satisfies $V^\star_0 - V^\pi_0 \le 16\epsilon_{\ell_\epsilon} = 16 \cdot 2^{-\ell_\epsilon} \le \epsilon$. Furthermore, if we terminate early on Line 20, then we know that $|\Pi_{\ell+1}| = 1$, and since $\pi^\star \in \Pi_{\ell+1}$, we have that the algorithm returns $\pi^\star$. Thus, the policy returned by Algorithm 2 is always ϵ-optimal.

It therefore remains to bound the sample complexity of Algorithm 2. At round ℓ of Algorithm 2, we collect $\bar{n}_\ell$ samples plus the number of samples collected from OptCov. On $\mathcal{E}_{\mathrm{good}}$, we have that the

number of samples collected by OptCov at round ℓ step h is bounded by

$$
\begin{aligned}
&C \cdot \frac{\inf_{\pi_{\mathrm{exp}}} \max_{\pi \in \Pi_\ell} \|M_h^\ell((\boldsymbol{\pi}_h - \bar{\boldsymbol{\pi}}_{\ell,h})\widehat{w}_{\ell,h}^{\bar{\pi}} + \boldsymbol{\pi}_h \widehat{\delta}_{\ell,h}^{\pi})\|^2_{\Lambda_h(\pi_{\mathrm{exp}})^{-1}}}{\epsilon_{\mathrm{exp}}^\ell} \\
&\quad + \frac{C_{\mathrm{fw}}^\ell}{(\epsilon_{\mathrm{exp}}^\ell)^{4/5}} + \frac{C_{\mathrm{fw}}^\ell}{\epsilon_{\mathrm{unif}}^\ell} + \log(C_{\mathrm{fw}}^\ell) \cdot K_{\mathrm{unif}}^\ell \\
&\overset{(a)}{\leq} C \cdot \frac{\inf_{\pi_{\mathrm{exp}}} \max_{\pi \in \Pi_\ell} \|\bar{\boldsymbol{\pi}}_{\ell,h} w_{\ell,h}^{\bar{\pi}} - \boldsymbol{\pi}_h w_h^\pi\|^2_{\Lambda_h(\pi_{\mathrm{exp}})^{-1}}}{\epsilon_{\mathrm{exp}}^\ell} + \frac{C_{\mathrm{fw}}^\ell}{(\epsilon_{\mathrm{exp}}^\ell)^{4/5}} + \frac{C_{\mathrm{fw}}^\ell}{\epsilon_{\mathrm{unif}}^\ell} + \log(C_{\mathrm{fw}}^\ell) \cdot K_{\mathrm{unif}}^\ell \\
&\qquad + \frac{(8S^2A + 32S^3AH^2)\epsilon_\ell^{5/3} + 2S^2AH\beta_\ell\epsilon_{\mathrm{exp}}^\ell + 4096S^3AH^2(\epsilon_{\mathrm{unif}}^\ell)^2}{\epsilon_{\mathrm{unif}}^\ell \epsilon_{\mathrm{exp}}^\ell} \\
&\overset{(b)}{\leq} C \cdot \frac{\inf_{\pi_{\mathrm{exp}}} \max_{\pi \in \Pi_\ell} \|\bar{\boldsymbol{\pi}}_{\ell,h} w_{\ell,h}^{\bar{\pi}} - \boldsymbol{\pi}_h w_h^\pi\|^2_{\Lambda_h(\pi_{\mathrm{exp}})^{-1}}}{\epsilon_\ell^2} \cdot H^4\beta_\ell^2 + \frac{C_{\mathrm{poly}}^\ell}{\epsilon_\ell^{5/3}} \\
&\overset{(c)}{\leq} C \cdot \frac{\inf_{\pi_{\mathrm{exp}}} \max_{\pi \in \Pi_\ell} \|\boldsymbol{\pi}_h^\star w_h^{\pi^\star} - \boldsymbol{\pi}_h w_h^\pi\|^2_{\Lambda_h(\pi_{\mathrm{exp}})^{-1}}}{\epsilon_\ell^2} \cdot H^4\beta_\ell^2 + \frac{C_{\mathrm{poly}}^\ell}{\epsilon_\ell^{5/3}} \\
&\overset{(d)}{\leq} C \cdot \inf_{\pi_{\mathrm{exp}}} \max_{\pi \in \Pi} \frac{\|\boldsymbol{\pi}_h^\star w_h^{\pi^\star} - \boldsymbol{\pi}_h w_h^\pi\|^2_{\Lambda_h(\pi_{\mathrm{exp}})^{-1}}}{\max\{\epsilon_\ell^2, \Delta(\pi)^2\}} \cdot H^4\beta_\ell^2 + \frac{C_{\mathrm{poly}}^\ell}{\epsilon_\ell^{5/3}}
\end{aligned}
$$

where the initial bound holds from Lemma 14, the (a) follows from Lemma 25, and (b) follows plugging in our choice of $\epsilon_{\mathrm{unif}}^\ell$ and $\epsilon_{\mathrm{exp}}^\ell$, and with $C_{\mathrm{poly}}^\ell = \mathrm{poly}(S, A, H, \log \ell/\delta, \log 1/\epsilon, \log|\Pi|)$, (c) holds by the triangle inequality and since $\bar{\pi}_\ell \in \Pi_\ell$, and (d) holds because, for all $\pi \in \Pi_\ell$, we have $\Delta(\pi) \leq 32\epsilon_\ell$. Furthermore, we can bound $\bar{n}_\ell$ as

$$
\begin{aligned}
\bar{n}_\ell &= \min_{\bar{\pi} \in \Pi_\ell} \max_{\pi \in \Pi_\ell} c \cdot \frac{H\widehat{U}_{\ell-1}(\pi, \bar{\pi}) + H^4S^{3/2}\sqrt{A}\log\frac{SAH\ell^2}{\delta} \cdot \epsilon_\ell^{1/3} + S^2H^4\epsilon_{\mathrm{unif}}^\ell}{\epsilon_\ell^2} \cdot \log\frac{60H\ell^2|\Pi_\ell|}{\delta} \\
&\overset{(a)}{\leq} \min_{\bar{\pi} \in \Pi_\ell} \max_{\pi \in \Pi_\ell} c \cdot \frac{HU(\pi, \bar{\pi}) + H^4S^{3/2}\sqrt{A}\log\frac{SAH\ell^2}{\delta} \cdot \epsilon_\ell^{1/3} + S^2H^4\epsilon_{\mathrm{unif}}^\ell}{\epsilon_\ell^2} \cdot \log\frac{60H\ell^2|\Pi_\ell|}{\delta} \\
&\overset{(b)}{\leq} \max_{\pi \in \Pi} c \cdot \frac{HU(\pi, \pi^\star)}{\max\{\epsilon_\ell^2, \Delta(\pi)^2\}} \cdot \log\frac{60H\ell^2|\Pi_\ell|}{\delta} + \frac{C_{\mathrm{poly}}^\ell}{\epsilon_\ell^{5/3}}
\end{aligned}
$$

where (a) follows from Lemma 23, and (b) since $\pi^\star \in \Pi_\ell$, and by a similar argument as above.

Thus, if we run for a total of L rounds, the sample complexity is bounded as

$$
\begin{aligned}
\sum_{\ell=1}^{L} \Bigg(C \cdot \sum_{h=1}^{H} \inf_{\pi_{\mathrm{exp}}} \max_{\pi \in \Pi} \frac{\|\boldsymbol{\pi}_h^\star w_h^{\pi^\star} - \boldsymbol{\pi}_h w_h^\pi\|^2_{\Lambda_h(\pi_{\mathrm{exp}})^{-1}}}{\max\{\epsilon_\ell^2, \Delta(\pi)^2\}} \cdot H^4\beta_\ell^2 \\
+ \max_{\pi \in \Pi} c \cdot \frac{HU(\pi, \pi^\star)}{\max\{\epsilon_\ell^2, \Delta(\pi)^2\}} \cdot \log\frac{60H\ell^2|\Pi_\ell|}{\delta} \Bigg) + \frac{LC_{\mathrm{poly}}^L}{\epsilon_L^{5/3}}.
\end{aligned}
$$

By construction, we have that $L \leq \lceil \log_2 16/\epsilon \rceil$. However, we terminate early if $|\Pi_{\ell+1}| = 1$, and since each $\pi \in \Pi_{\ell+1}$ satisfies $\Delta(\pi) \leq \epsilon_\ell$, it follows that we will have $|\Pi_{\ell+1}| = 1$ once $\epsilon_\ell < \Delta_{\min}$, which will occur for $\ell \geq \lceil \log_2 \frac{1}{\Delta_{\min}} \rceil + 1$. Thus, we can bound

$$
L \leq \min\{\lceil \log_2 16/\epsilon \rceil, \lceil \log_2 1/\Delta_{\min} \rceil + 1\},
$$

and so for all $\epsilon_\ell, \ell \leq L$, we have $\epsilon_\ell \geq c \cdot \max\{\epsilon, \Delta_{\min}\}$. Plugging this into the above gives the final complexity.

□

D Tabular Contextual Bandits: Upper Bound

Setting and notation. We study stochastic tabular contextual bandits, denoted by the tuple $(\mathcal{C}, \mathcal{A}, \mu^\star, \nu)$. At each episode, a context $c \sim \mu^\star$ arrives, the agent chooses an action $a \in \mathcal{A}$,

https://doi.org/10.52202/079017-0717

and receives reward $r(c,a) \sim \nu(c,a)$ in $\mathbb{R}$. Note that this is a special case of the Tabular MDP when $H = 1$. In this setting, we use the terminology "contexts" instead of "states" to highlight that the agent has no impact on these. The vector $\mu^\star$ plays the same role as the state visitation vectors w_h^π previously, except this is now policy-independent. The notation for policy matrix $\boldsymbol{\pi}$, values V^π, features $\phi^\pi(c,a)$ are inherited directly from the general case.

Define $\theta^\star \in R^{|\mathcal{C}|A}$ as the vector of reward means, so that $[\theta^\star]_{(c,a)} = \mathbb{E}_\nu[r(c,a)]$. Then, we can write the value of π as:

$$\mathbb{E}_{\nu,\mu^\star}[r(c,\pi(c))] = \sum_{c,a} \theta^\star_{c,a}[\mu^\star]_c[\pi(c)]_a = (\theta^\star)^\top \boldsymbol{\pi}\mu^\star$$

For any (θ,μ) define $\mathsf{OPT}(\theta,\mu) := \arg\max_{\pi\in\Pi} \theta^\top \boldsymbol{\pi}\mu$, where θ is any hypothetical vector of reward-means and $\mu \in \Delta_{|\mathcal{C}|}$ is a hypothetical context distribution.

Recall that we use $\boldsymbol{\pi} \in \mathbb{R}^{|\mathcal{C}|A\times|\mathcal{C}|}$ to refer to the policy matrix. The vector $\boldsymbol{\pi}\mu \in \mathbb{R}^{|\mathcal{C}|A}$ contains context-action visitations for policy π under context distribution μ. Define function $G(\mu,\pi) = \mathbb{E}_{\mu,\pi}[(\boldsymbol{\pi}\mu)(\boldsymbol{\pi}\mu)^\top]$ which returns the expected covariance matrix of policy π under context distribution μ. For shorthand, we refer to $\hat{A}(\pi) = G(\widehat{\mu}_\ell, \pi_{\exp})$ and $A(\pi) = G(\mu^\star, \pi_{\exp})$ for any π.

Lemma 27. *Define the experimental design objective*

$$F(\pi_{\exp},\mu,\pi,\pi') = \|(\boldsymbol{\pi}' - \boldsymbol{\pi})\mu\|^2_{G(\mu,\pi_{\exp})^{-1}}.$$

Then, for any $\mu \in \Delta_{\mathcal{C}}$,

$$\min_{\pi_{\exp}} \max_{\pi,\pi'\in\Pi_\ell} F(\pi_{\exp},\mu,\pi,\pi') = \max_{\pi,\pi'\in\Pi_\ell} \min_{\pi_{\exp}} F(\pi_{\exp},\mu,\pi,\pi')$$

Proof. We can rewrite the maximization problem to be over the simplex $\Delta_{\Pi_\ell\times\Pi_\ell}$ instead:

$$\min_{\pi_{\exp}} \max_{\lambda\in\Delta_{\Pi_\ell\times\Pi_\ell}} \sum_{\pi,\pi'\in\Pi_\ell\times\Pi_\ell} \lambda_{\pi,\pi'} F(\pi_{\exp},\mu,\pi,\pi') \tag{D.1}$$

This does not change the objective value. To see this, note that for any selection (π_1,π_2) in the original problem, the same objective value can be obtained by setting $\lambda = e_{\pi_1,\pi_2}$; hence, the modification to the optimization cannot reduce the value. Further if $F(\pi_{\exp},\mu,\pi,\pi')$ is maximized by (π_1,π_2), setting λ as anything other than e_{π_1,π_2} cannot increase the objective value.

Now, note that both the minimization and maximization problems are over simplices, which are compact and convex sets. The objective is linear in the maximization variable, and hence concave. The objective can be rewritten as

$$\sum_{c\in\mathcal{S}}\sum_{a\in\mathcal{A}} \frac{(\boldsymbol{\pi}-\boldsymbol{\pi}')^\top e_{a,c} e_{a,c}^\top (\boldsymbol{\pi}-\boldsymbol{\pi}')}{p_{c,a}}.$$

Here, $p_{c,a}$ as the probability that $\pi_{\exp}$ plays action a, given that we are in context c. From this representation, we can clearly see that the objective is convex in each $p_{c,a}$. Hence, since we are optimizing over finite-dimensional spaces ($|\mathcal{A}|$ and $|\mathcal{C}|$ are finite), Von Neumann's minimax theorem applies and the proof is complete. □

Lemma 28. *For the contextual bandit problem, define the experimental design objective*

$$F(\pi_{\exp},\mu,\pi,\pi') = \|(\boldsymbol{\pi}' - \boldsymbol{\pi})\mu\|^2_{G(\mu,\pi_{\exp})^{-1}}.$$

Then, for any μ and assuming that all policies in Π_ℓ are deterministic, we have:

$$\min_{\pi_{\exp}} \max_{\pi,\pi'\in\Pi_\ell} F(\pi_{\exp},\mu,\pi,\pi') = \max_{\pi,\pi'\in\Pi_\ell} \mathbb{E}_{c\sim\mu}[4\mathbb{I}[\pi(c)\neq\pi'(c)]], \tag{D.2}$$

Proof. Below, we refer to $p_{c,a}$ as the probability that $\pi_{\exp}$ plays action a, given that we are in context c. We have:

$$\begin{aligned}
&\min_{\pi_{\exp}} \max_{\pi,\pi' \in \Pi_\ell} \|(\boldsymbol{\pi}' - \boldsymbol{\pi})\mu\|^2_{G(\mu,\pi_{\exp})^{-1}} \\
&= \max_{\pi,\pi' \in \Pi_\ell} \min_{\pi_{\exp}} \|(\boldsymbol{\pi}' - \boldsymbol{\pi})\mu\|^2_{G(\mu,\pi_{\exp})^{-1}} \\
&= \max_{\pi,\pi' \in \Pi_\ell} \min_{p_1 \ldots p_C \in \Delta_{\mathcal{A}}} \sum_{a,c} \mu_c^2 \frac{(\boldsymbol{\pi} - \boldsymbol{\pi}')^\top e_{a,c} e_{a,c}^\top (\boldsymbol{\pi} - \boldsymbol{\pi}')}{\mu_c p_{c,a}} \\
&= \max_{\pi,\pi' \in \Pi_\ell} \sum_c \mu_c \min_{p_c} \sum_{a \in \mathcal{A}} \frac{(\boldsymbol{\pi} - \boldsymbol{\pi}')^\top e_{a,c} e_{a,c}^\top (\boldsymbol{\pi} - \boldsymbol{\pi}')}{p_{c,a}} \\
&= \max_{\pi,\pi' \in \Pi_\ell} \sum_c \mu_c \left(\sum_{a \in \mathcal{A}} \sqrt{(\boldsymbol{\pi} - \boldsymbol{\pi}')^\top e_{a,c} e_{a,c}^\top (\boldsymbol{\pi} - \boldsymbol{\pi}')} \right)^2 .
\end{aligned}$$

Here the first equality follows from Lemma 27, and the last from Lemma D.6 of [30].

We have assumed that the policies in Π_ℓ are deterministic. Hence, the only two actions in the summation over $\mathcal{A}$ above that are relevant are $\pi(c)$ and $\pi'(c)$. For all other $a \in \mathcal{A}$, the term in the square root evaluates to 0. If $\pi(c) = \pi'(c)$, then the entire summation over $\mathcal{A}$ evaluates to 0; else, the terms indexed by $\pi(c)$ and $\pi'(c)$ are both 1, and the summation evalutes to 2. Hence, we can simplify the expression to exactly the form of Equation (D.2) from the lemma statement, and the proof is complete. □

Lemma 29. *For the contextual bandits problem, we have that*

$$\max_{\pi \in \Pi} \mathbb{E}_{c \sim \mu^\star}[\mathbb{E}_{\nu^\star}[(r(c,\pi(c)) - r(c,\pi^\star(c)))^2 | c]] \leq \inf_{\pi_{\exp}} \max_{\pi \in \Pi} \|\phi^\star - \phi^\pi\|^2_{\Lambda(\pi_{\exp})^{-1}}$$

Proof. Observe that $r(c,\pi(c)) - r(c,\pi^\star(c)) = 0$ if $\pi(c) = \pi^\star(c)$; else, $|r(c,\pi(c)) - r(c,\pi^\star(c))| \leq 2$. Then, it follows that

$$\begin{aligned}
&\max_{\pi \in \Pi} \mathbb{E}_{c \sim \mu^\star}[\mathbb{E}_{\nu^\star}[(r(c,\pi(c)) - r(c,\pi^\star(c)))^2 | c]] \\
&\leq \max_{\pi \in \Pi} 4\mathbb{E}_{c \sim \mu^\star} \mathbb{I}(\pi(c) \neq \pi^\star(c)) \\
&= \inf_{\pi_{\exp}} \max_{\pi \in \Pi} \|\phi^\star - \phi^\pi\|^2_{\Lambda(\pi_{\exp})^{-1}},
\end{aligned}$$

where the equality follows from Lemma 28. □

Now, we state our main upper bound for contextual bandits.

Corollary 1. *For the setting of tabular contextual bandits, there exists an algorithm such that with probability at least $1 - 2\delta$, as long as Π contains only deterministic policies, it finds an ϵ-optimal policy and terminates after collecting at most the following number of samples:*

$$\inf_{\pi_{\exp}} \max_{\pi \in \Pi} \frac{\|\phi^\star - \phi^\pi\|^2_{\Lambda(\pi_{\exp})^{-1}}}{\max\{\epsilon^2, \Delta(\pi)^2\}} \cdot \beta^2 \log \frac{1}{\Delta_{\min} \vee \epsilon} + \frac{C_{\text{poly}}}{\max\{\epsilon^{5/3}, \Delta_{\min}^{5/3}\}},$$

for $C_{\text{poly}} = \text{poly}(|\mathcal{S}|, A, \log 1/\delta, \log 1/(\Delta_{\min} \vee \epsilon), \log |\Pi|)$ *and* $\beta = C\sqrt{\log(\frac{S|\Pi|}{\delta} \cdot \frac{1}{\Delta_{\min} \vee \epsilon})}$.

Proof. In the special case of contextual bandits, $U(\pi, \pi^\star)$ defined in Theorem 1 can be written more simply as $\mathbb{E}_{c \sim \mu^\star}[\mathbb{E}_{\nu^\star}[(r(c,\pi(c)) - r(c,\pi^\star(c)))^2 | c]]$. Then, by Lemma 29, we have that:

$$\frac{U(\pi, \pi^\star)}{\max\{\epsilon^2, \Delta(\pi)^2, \Delta_{\min}^2\}} \leq \inf_{\pi_{\exp}} \max_{\pi \in \Pi} \frac{\|\phi^\star - \phi^\pi\|^2_{\Lambda(\pi_{\exp})^{-1}}}{\max\{\epsilon^2, \Delta(\pi)^2, \Delta_{\min}^2\}}$$

Plugging this into Theorem 1 completes the proof. □

https://doi.org/10.52202/079017-0717

E MDPs with Action-Independent Transitions

We consider here a special class of MDPs where the transitions only depend on the states and are independent of the actions selected i.e all P_h are such that $P_h(s,a) = P_h(s,a')$ for all $(a,a') \in \mathcal{A}$. In this special case, we prove in this subsection that the (leading order) complexity of PERP reduces to $O(\rho_\Pi)$.

Lemma 30. *For the ergodic MDP problem,*

$$\min_{\pi_{\exp}} \max_{\pi \in \Pi} \|\phi_h^\pi - \phi_h^\star\|^2_{\Lambda_h(\pi_{\exp})^{-1}} = \max_{\pi \in \Pi} \min_{\pi_{\exp}} \|\phi_h^\pi - \phi_h^\star\|^2_{\Lambda_h(\pi_{\exp})^{-1}}$$

Proof. We can rewrite the maximization problem to be over the simplex Δ_Π instead:

$$\min_{\pi_{\exp}} \max_{\lambda \in \Delta_\Pi} \sum_{\pi \in \Pi} \lambda_\pi \|\phi_h^\pi - \phi_h^\star\|^2_{\Lambda_h(\pi_{\exp})^{-1}} \tag{E.1}$$

This does not change the objective value. To see this, note that for any selection $\pi \in \Pi$ in the original problem, the same objective value can be obtained by setting $\lambda = e_\pi$ in Equation (E.1); hence, the modification to the optimization cannot reduce the value. Further if $\|\phi_h^\pi - \phi_h^\star\|^2_{\Lambda_h(\pi_{\exp})^{-1}}$ is maximized by π for any fixed $\pi_{\exp}$, setting λ as anything other than e_π cannot increase the objective value.

Now, note that both the minimization and maximization problems are over simplices, which are compact and convex sets. The objective is linear in the maximization variable, and hence concave. The objective can be rewritten as

$$\sum_a \frac{(\boldsymbol{\pi}_h - \boldsymbol{\pi}_h^\star)^\top e_{s,a} e_{s,a}^\top (\boldsymbol{\pi}_h - \boldsymbol{\pi}_h^\star)}{p_{s,a}}$$

Here, $p_{s,a}$ is the probability that $\pi_{\exp}$ plays action a, given that it is in context s. From this representation, we can clearly see that the objective is convex in each $p_{s,a}$. Hence, Von Neumann's minimax theorem applies and the proof is complete. □

Lemma 31. *For the setting of ergodic MDPs,*

$$\min_{\pi_{\exp}} \max_{\pi \in \Pi} \|\phi_h^\pi - \phi_h^\star\|^2_{\Lambda_h(\pi_{\exp})^{-1}} = \max_{\pi \in \Pi} 2\mathbb{E}_{s \sim w_h^\star} \mathbb{I}[\pi_h(s) \neq \pi'_h(s)], \tag{E.2}$$

Proof. Below, we refer to $p_{s,a}$ as the probability that $\pi_{\exp}$ plays action a, given that it is in context s. The second equality follows from Lemma 30.

$$\begin{aligned}
&\min_{\pi_{\exp}} \max_{\pi \in \Pi} \|\phi_h^\pi - \phi_h^\star\|^2_{\Lambda_h(\pi_{\exp})^{-1}} \\
&= \min_{\pi_{\exp}} \max_{\pi \in \Pi} \|(\boldsymbol{\pi}_h - \boldsymbol{\pi}_h^\star) w_h^\star\|^2_{\Lambda_h(\pi_{\exp})^{-1}} \\
&= \max_{\pi \in \Pi} \min_{\pi_{\exp}} \|(\boldsymbol{\pi}_h - \boldsymbol{\pi}_h^\star) w_h^\star\|^2_{\Lambda_h(\pi_{\exp})^{-1}} \\
&= \max_{\pi \in \Pi} \min_{p_1 \ldots p_S \in \Delta_\mathcal{A}} \sum_{s,a} (w_h^\star(s))^2 \frac{(\boldsymbol{\pi}_h - \boldsymbol{\pi}_h^\star)^\top e_{s,a} e_{s,a}^\top (\boldsymbol{\pi}_h - \boldsymbol{\pi}_h^\star)}{w_h^\star(s) p_{s,a}} \\
&= \max_{\pi \in \Pi} \sum_s w_h^\star(s) \min_{p_s \in \Delta_\mathcal{A}} \sum_a \frac{(\boldsymbol{\pi}_h - \boldsymbol{\pi}_h^\star)^\top e_{s,a} e_{s,a}^\top (\boldsymbol{\pi}_h - \boldsymbol{\pi}_h^\star)}{p_{s,a}} \\
&= \max_{\pi \in \Pi} \sum_s w_h^\star(s) \left(\sum_a \sqrt{(\boldsymbol{\pi}_h - \boldsymbol{\pi}_h^\star)^\top e_{s,a} e_{s,a}^\top (\boldsymbol{\pi}_h - \boldsymbol{\pi}_h^\star)} \right)^2
\end{aligned}$$

The optimization problems in the final line were solved using KKT conditions. We assume that the two policies are deterministic. Hence, the only two actions in the summation over $\mathcal{A}$ above that are relevant are $\pi_h(s)$ and $\pi'_h(s)$. For all other $a \in \mathcal{A}$, the term in the square root evaluates to 0. If $\pi_h(s) = \pi'_h(s)$, then the entire summation over $\mathcal{A}$ evaluates to 0; else, the terms indexed by $\pi(c)$ and $\pi'(c)$ are both 1, and the summation evalutes to 2. Hence, we can simplify the expression to exactly the form of Equation (E.2) from the lemma statement, and the proof is complete. □

Lemma 32. *For the ergodic MDP problem, we have that*

$$\max_{\pi\in\Pi}\frac{HU(\pi,\pi^\star)}{\max\{\epsilon^2,\Delta(\pi)^2,\Delta_{\min}^2\}}\le 2H^4\sum_{h=1}^{H}\inf_{\pi_{\exp}}\max_{\pi\in\Pi}\frac{\|\phi_h^\star-\phi_h^\pi\|^2_{\Lambda_h(\pi_{\exp})^{-1}}}{\max\{\epsilon^2,\Delta(\pi)^2,\Delta_{\min}^2\}}$$

Proof. Recall the definition of $U(\pi,\pi^\star)$

$$U(\pi,\pi^\star)=\sum_{h=1}^{H}\mathbb{E}_{s_h\sim w_h^{\pi^\star}}[(Q_h^\pi(s_h,\pi_h(s))-Q_h^\pi(s_h,\pi_h^\star(s)))^2].$$

Then, we have that

$$\begin{aligned}
&\max_{\pi\in\Pi}\frac{HU(\pi,\pi^\star)}{\max\{\epsilon^2,\Delta(\pi)^2,\Delta_{\min}^2\}}\\
&=\max_{\pi\in\Pi}\frac{H\sum_{h=1}^{H}\mathbb{E}_{s_h\sim w_h^{\pi^\star}}[(Q_h^\pi(s_h,\pi_h(s))-Q_h^\pi(s_h,\pi_h^\star(s)))^2]}{\max\{\epsilon^2,\Delta(\pi)^2,\Delta_{\min}^2\}}\\
&\le H\sum_{h=1}^{H}\max_{\pi\in\Pi}\frac{\mathbb{E}_{s_h\sim w_h^{\pi^\star}}[(Q_h^\pi(s_h,\pi_h(s))-Q_h^\pi(s_h,\pi_h^\star(s)))^2]}{\max\{\epsilon^2,\Delta(\pi)^2,\Delta_{\min}^2\}}\\
&\le H\sum_{h=1}^{H}\max_{\pi\in\Pi}\frac{2H^2\mathbb{E}_{s\sim w_h^\star}\mathbb{I}[\pi_h(s)\neq\pi_h'(s)]}{\max\{\epsilon^2,\Delta(\pi)^2,\Delta_{\min}^2\}}\\
&=H^4\sum_{h=1}^{H}\inf_{\pi_{\exp}}\max_{\pi\in\Pi}\frac{\|\phi_h^\star-\phi_h^\pi\|^2_{\Lambda_h(\pi_{\exp})^{-1}}}{\max\{\epsilon^2,\Delta(\pi)^2,\Delta_{\min}^2\}}.
\end{aligned}$$

The final equality follows from Lemma 31. □

Corollary 2. *Assume that all P_h are such that $P_h(s'|s,a)=P_h(s'|s,a')$ for all $(a,a')\in\mathcal{A}$. Then, with probability at least $1-2\delta$,* PERP *(Algorithm 2) finds an ϵ-optimal policy and terminates after collecting at most the following number of episodes:*

$$\sum_{h=1}^{H}\inf_{\pi_{\exp}}\max_{\pi\in\Pi}\frac{\|\phi_h^\star-\phi_h^\pi\|^2_{\Lambda_h(\pi_{\exp})^{-1}}}{\max\{\epsilon^2,\Delta(\pi)^2\}}\cdot\iota H^4\beta^2+\frac{C_{\mathrm{poly}}}{\max\{\epsilon^{5/3},\Delta_{\min}^{5/3}\}}$$

for C_{poly},β as defined in Theorem 1.

Proof. The proof follows directly from Theorem 1 and Lemma 32. □

F Tabular Franke Wolfe

Theorem 2. *Fix parameters $K_{\mathrm{unif}}>0$, $\epsilon_{\exp}>0$, and consider some $\Phi\subseteq\mathbb{R}^{SA}$ and set $\mathcal{S}_0\subseteq\mathcal{S}$. Let $\epsilon_{\mathrm{unif}}>0$ be some value satisfying*

$$W_h^\star(s)>\epsilon_{\mathrm{unif}},\forall s\in\mathcal{S}_0,\quad \text{and}\quad K_{\mathrm{unif}}\ge\epsilon_{\mathrm{unif}}^{-1}.$$

Assume that $|[\boldsymbol{\phi}]_{(s,a)}|\le C_{\boldsymbol{\phi}}\cdot(W_h^\star(s)+\sqrt{\epsilon_\phi})$ for all $s\in\mathcal{S}_0$, $\boldsymbol{\phi}\in\Phi$, and some $C_{\boldsymbol{\phi}}>0$, and that $[\boldsymbol{\phi}]_{(s,a)}=0$ for $s\notin\mathcal{S}_0$. Additionally, let the parameters be such that $\epsilon_\phi/(K_{\mathrm{unif}}\epsilon_{\mathrm{unif}})\le\epsilon_{\exp}$. Then with probability at least $1-\delta$, algorithm Algorithm 3 run with these parameters will collect at most

$$\min\left\{C\cdot\frac{\inf_{\boldsymbol{\Lambda}\in\boldsymbol{\Omega}_h}\max_{\boldsymbol{\phi}\in\Phi}\|\boldsymbol{\phi}\|^2_{\boldsymbol{\Lambda}^{-1}}}{\epsilon_{\exp}}+\frac{C_{\mathrm{fw}}}{\epsilon_{\exp}^{4/5}},C_{\mathrm{fw}}(\frac{1}{\epsilon_{\exp}}+K_{\mathrm{unif}})\right\}+\frac{C_{\mathrm{fw}}}{\epsilon_{\mathrm{unif}}}+\log(C_{\mathrm{fw}})\cdot K_{\mathrm{unif}}$$

episodes, for C a universal constant and $C_{\mathrm{fw}}=\mathrm{poly}(S,A,H,C_{\boldsymbol{\phi}},\log 1/\delta,\log 1/\epsilon_{\exp},\log|\Phi|)$, and will produce covariates $\widehat{\boldsymbol{\Sigma}}$ such that

$$\max_{\boldsymbol{\phi}\in\Phi}\|\boldsymbol{\phi}\|^2_{\widehat{\boldsymbol{\Sigma}}^{-1}}\le\epsilon_{\exp}\tag{F.1}$$

and, for all $s\in\mathcal{S}_0$,

$$[\widehat{\boldsymbol{\Sigma}}]_{(s,a)}\ge\frac{\epsilon_{\mathrm{unif}}}{2SA}\cdot K_{\mathrm{unif}}.\tag{F.2}$$

https://doi.org/10.52202/079017-0717

Algorithm 3 Online Experiment Design (OPTCOV)

input: directions Φ, tolerance ϵ_{exp}, confidence δ, minimum reachability ϵ_{unif}, minimum exploration K_{unif}, pruned states $\mathcal{S}_0$, step h
$i \leftarrow 1$
while $T_i K_i \leq \mathrm{poly}(S, A, H, C_{\boldsymbol{\phi}}, \log 1/\delta, \log 1/\epsilon_{\mathrm{exp}}, \log|\Phi|) \cdot \epsilon_{\mathrm{exp}}^{-1}$ **do**
  $\mathfrak{D}^i_{\mathrm{unif}} \leftarrow \textsc{UnifExp}(\epsilon_{\mathrm{unif}}, K_i T_i + K_{\mathrm{unif}}, \delta/8i^2)$
  $\boldsymbol{\Lambda}^i_0 \leftarrow \frac{1}{T_i K_i}\mathrm{diag}(v^i)$ where $[v^i]_{sa} = \sum_{(s',a')\in\mathfrak{D}^i_{\mathrm{unif}}} \mathbb{I}\{(s',a') = (s,a)\}$ for $s \in \mathcal{S}_0$, and $T_i K_i$ otherwise
  Run iteration i of Algorithm 4 of [43] on objective

$$f_i(\boldsymbol{\Lambda}) \leftarrow \frac{1}{\eta_i}\log\left(\sum_{\boldsymbol{\phi}\in\Phi} e^{\eta_i \|\boldsymbol{\phi}\|^2_{\mathbf{A}(\boldsymbol{\Lambda})^{-1}}}\right) \quad \text{for} \quad \mathbf{A}(\boldsymbol{\Lambda}) = \boldsymbol{\Lambda} + \boldsymbol{\Lambda}^i_0, \eta_i = 2^{2i/5}$$

  to obtain data $\mathfrak{D}^i$
  if Algorithm 4 reaches termination condition **then**
    return $\mathfrak{D}^i \cup \mathfrak{D}^i_{\mathrm{unif}}$
  end if
  $i \leftarrow i + 1$
end while
$\mathfrak{D} \leftarrow \textsc{UnifExp}(\epsilon_{\mathrm{unif}}, \frac{8S^2A^2C^2_{\boldsymbol{\phi}}}{\epsilon_{\mathrm{exp}}} + (8S^2A^2C^2_{\boldsymbol{\phi}} + 1)K_{\mathrm{unif}}, \delta/4)$
return $\mathfrak{D}$

Proof. To prove this result, we apply Lemma 37 combined with Lemma 36.

Let $\mathcal{E}^i_{\mathrm{exp}}$ denote the success event of running Algorithm 4 at epoch i, as defined in Lemma 36. On this event, and under the assumption that $W^\star_h(s) > \epsilon_{\mathrm{unif}}$ for each $s \in \mathcal{S}_0$, we have that $[\boldsymbol{\Sigma}_i]_{(s,a)} \geq \frac{W^\star_h(s)}{2SA} \cdot (T_i K_i + K_{\mathrm{unif}})$ for each (s,a) with $s \in \mathcal{S}_0$ and $\boldsymbol{\Sigma}_i$ the covariates induced by $\mathfrak{D}^i_{\mathrm{unif}}$, which implies that

$$[\boldsymbol{\Lambda}^i_0]_{(s,a)} \geq \frac{1}{T_i K_i}\frac{W^\star_h(s)}{2SA} \cdot (T_i K_i + K_{\mathrm{unif}}) \geq \frac{W^\star_h(s)}{2SA}$$

for each (s,a) with $s \in \mathcal{S}_0$, and, furthermore, Algorithm 4 collects at most

$$T_i K_i + K_{\mathrm{unif}} + \mathrm{poly}(S, A, H, \log\frac{T_i K_i i^2}{\delta\epsilon_{\mathrm{unif}}}) \cdot \frac{1}{\epsilon_{\mathrm{unif}}} \tag{F.3}$$

episodes. Furthermore, by Lemma 36, we have $\mathbb{P}[\mathcal{E}^i_{\mathrm{exp}}] \geq \delta/2i^2$, so it follows that

$$\mathbb{P}[\cup_{i\geq 1}(\mathcal{E}^i_{\mathrm{exp}})^c] \leq \sum_{i=1}^{\infty}\frac{\delta}{8i^2} \leq \delta/4.$$

Henceforth, we therefore assume that $\mathcal{E}^i_{\mathrm{exp}}$ holds for each i. This immediately implies that (F.2) holds.

It remains to show that (F.1) is satisfied, and that our sample complexity guarantee is met. To this end we apply Lemma 37 with $\boldsymbol{\Lambda}_0$ a diagonal matrix, with $[\boldsymbol{\Lambda}_0]_{(s,a)} = \frac{W^\star_h(s)}{2SA}$ for $s \in \mathcal{S}_0$, and otherwise $[\boldsymbol{\Lambda}_0]_{(s,a)} = 1$. Note that with this choice of $\boldsymbol{\Lambda}_0$, by what we just showed above, we have $\boldsymbol{\Lambda}^i_0 \succeq \boldsymbol{\Lambda}_0$, as required by Lemma 37.

We next turn to bounding the smoothness constants, β and M. First, note that by Lemma 34, at epoch i we have that all iterates of FWREGRET live in the set $\widehat{\boldsymbol{\Omega}}_{h,T_i K_i}(\delta/8i^2)$ with probability $1 - \delta/8i^2$. Union bounding over this event for all i, with probability at least $1 - \delta/4$, we have that for each i all iterates of FWREGRET live in the set $\widehat{\boldsymbol{\Omega}}_{h,T_i K_i}(\delta/8i^2)$. By Lemma 35, since we have assumed that $|[\boldsymbol{\phi}]_{(s,a)}| \leq C_{\boldsymbol{\phi}} \cdot (W^\star_h(s) + \sqrt{\epsilon_{\boldsymbol{\phi}}})$ for all (s,a) with $s \in \mathcal{S}_0$ and otherwise $[\boldsymbol{\phi}]_{(s,a)} = 0$ for all

$\boldsymbol{\phi} \in \Phi$, we can then bound

$$M_i \le \max_{s \in \mathcal{S}_0} \left(\frac{2SAC_{\boldsymbol{\phi}}^2}{C'} + \frac{2SAC_{\boldsymbol{\phi}}^2 \epsilon_\phi}{C' \cdot W_h^\star(s)} \right) \cdot \left(\frac{2}{C'} + \frac{2}{C' T_i K_i W_h^\star(s)} \cdot \log \frac{SAH}{\delta} \right)$$

$$\beta_i \le \max_{s \in \mathcal{S}_0} (2\eta_i + 2) \left(\frac{2SAC_{\boldsymbol{\phi}}^2}{C'} + \frac{2SAC_{\boldsymbol{\phi}}^2 \epsilon_\phi}{C' \cdot W_h^\star(s)} \right)^2 \cdot \left(\frac{2}{C'} + \frac{2}{C' T_i K_i W_h^\star(s)} \cdot \log \frac{SAH}{\delta} \right)^2$$

On the event $\mathcal{E}_{\text{exp}}^i$, as noted above we have $[\boldsymbol{\Lambda}_0^i]_{(s,a)} \ge \frac{W_h^\star(s)}{2SA}(1 + \frac{K_{\text{unif}}}{T_i K_i})$ for $s \in \mathcal{S}_0$, so we can take $C' = \frac{1}{2SA}(1 + \frac{K_{\text{unif}}}{T_i K_i})$. We can then bound

$$\begin{aligned}
&\max_{s \in \mathcal{S}_0} \left(\frac{2SAC_{\boldsymbol{\phi}}^2}{C'} + \frac{2SAC_{\boldsymbol{\phi}}^2 \epsilon_\phi}{C' \cdot W_h^\star(s)} \right) \cdot \left(\frac{2}{C'} + \frac{2}{C' T_i K_i W_h^\star(s)} \cdot \log \frac{SAH}{\delta} \right) \\
&\le \left(4S^2 A^2 C_{\boldsymbol{\phi}} + \frac{4S^2 A^2 C_{\boldsymbol{\phi}}^2 \epsilon_\phi \cdot T_i K_i}{K_{\text{unif}} \epsilon_{\text{unif}}} \right) \cdot \left(4SA + \frac{4SA}{K_{\text{unif}} \epsilon_{\text{unif}}} \log \frac{SAH}{\delta} \right)
\end{aligned}$$

where we have used that $W_h^\star(s) \ge \epsilon_{\text{unif}}$ for all $s \in \mathcal{S}_0$, by assumption. By assumption we have $\frac{\epsilon_\phi}{K_{\text{unif}} \epsilon_{\text{unif}}} \le \epsilon_{\text{exp}}$. Note that by construction, the while statement on Line 3 will ensure that we always have $T_i K_i \le \text{poly}(S, A, H, C_{\boldsymbol{\phi}}, \log 1/\delta, \log 1/\epsilon_{\text{exp}}, \log |\Phi|) \cdot \epsilon_{\text{exp}}^{-1}$, so we can bound

$$\epsilon_{\text{exp}} \cdot T_i K_i \le \text{poly}(S, A, H, C_{\boldsymbol{\phi}}, \log 1/\delta, \log 1/\epsilon_{\text{exp}}, \log |\Phi|).$$

It follows that it suffices to take

$$\beta, M \le \text{poly}(S, A, H, C_{\boldsymbol{\phi}}, \log 1/\delta, \log 1/\epsilon_{\text{exp}}, \log |\Phi|).$$

We now consider two cases. In the first case, when the termination criteria on Line 7 is met, we can apply Lemma 37, to get that with probability at least $1 - \delta/4$ we have that the procedure terminates after running for at most

$$\begin{aligned}
&\max \left\{ \min_N 16N \quad \text{s.t.} \quad \inf_{\boldsymbol{\Lambda} \in \boldsymbol{\Omega}} \max_{\boldsymbol{\phi} \in \Phi} \boldsymbol{\phi}^\top (N\boldsymbol{\Lambda} + \boldsymbol{\Lambda}_0)^{-1} \boldsymbol{\phi} \le \frac{\epsilon_{\text{exp}}}{6}, \right. \\
&\qquad \left. \frac{\text{poly}(\beta, R, d, H, M, \log 1/\delta, \log 1/\epsilon_{\text{exp}}, \log |\Phi|)}{\epsilon_{\text{exp}}^{4/5}} \right\} \\
&\le \max \left\{ \min_N 16N \quad \text{s.t.} \quad \inf_{\boldsymbol{\Lambda} \in \boldsymbol{\Omega}} \max_{\boldsymbol{\phi} \in \Phi} \boldsymbol{\phi}^\top (N\boldsymbol{\Lambda} + \boldsymbol{\Lambda}_0)^{-1} \boldsymbol{\phi} \le \frac{\epsilon_{\text{exp}}}{6}, \right. \\
&\qquad \left. \frac{\text{poly}(S, A, H, C_{\boldsymbol{\phi}}, \log 1/\delta, \log 1/\epsilon_{\text{exp}}, \log |\Phi|)}{\epsilon_{\text{exp}}^{4/5}} \right\}
\end{aligned}$$

episodes, and returns data $\widehat{\boldsymbol{\Sigma}}_N$ such that

$$f_{\widehat{i}}(N^{-1} \widehat{\boldsymbol{\Sigma}}_N) \le N \epsilon_{\text{exp}},$$

where $\widehat{i}$ is the index of the epoch on which it terminates. By Lemma D.1 of [42], we have

$$\max_{\boldsymbol{\phi} \in \Phi} \|\boldsymbol{\phi}\|^2_{\mathbf{A}(N^{-1} \widehat{\boldsymbol{\Sigma}}_N)^{-1}} \le f_{\widehat{i}}(N^{-1} \widehat{\boldsymbol{\Sigma}}_N) \le N \epsilon_{\text{exp}}$$

which implies

$$\max_{\boldsymbol{\phi} \in \Phi} \|\boldsymbol{\phi}\|^2_{(\widehat{\boldsymbol{\Sigma}}_N + \boldsymbol{\Sigma}_{\widehat{i}})^{-1}} \le \epsilon_{\text{exp}},$$

which proves (F.1). Furthermore, (F.2) holds since as noted $[\boldsymbol{\Sigma}_i]_{(s,a)} \ge \frac{W_h^\star(s)}{2SA} \cdot (T_i K_i + K_{\text{unif}})$ for each (s, a) with $s \in \mathcal{S}_0$, and since $W_h^\star(s) \ge \epsilon_{\text{unif}}$ for all $s \in \mathcal{S}_0$.

In the second case, when the while loop on Line 3 terminates since $T_i K_i \le \text{poly}(S, A, H, C_{\boldsymbol{\phi}}, \log 1/\delta, \log 1/\epsilon_{\text{exp}}, \log |\Phi|) \cdot \epsilon_{\text{exp}}^{-1}$, we can bound the total number of episodes collected within the calls to Algorithm 4 of [43] within the while loop by

https://doi.org/10.52202/079017-0717

$\mathrm{poly}(S, A, H, C_{\boldsymbol{\phi}}, \log 1/\delta, \log 1/\epsilon_{\mathrm{exp}}, \log|\Phi|) \cdot \epsilon_{\mathrm{exp}}^{-1}$. Furthermore, by Lemma 36, with probability at least $1-\delta/4$, we have that the call to UNIFEXP on Line 12 terminates after running for at most

$$\frac{8S^2A^2C_{\boldsymbol{\phi}}^2}{\epsilon_{\mathrm{exp}}} + (8S^2A^2C_{\boldsymbol{\phi}}^2+1)K_{\mathrm{unif}} + \mathrm{poly}(S, A, H, \log\frac{T_iK_ii^2}{\delta\epsilon_{\mathrm{unif}}}) \cdot \frac{1}{\epsilon_{\mathrm{unif}}}$$

episodes, and that the returned data satisfies $N_h(s,a) \geq \frac{W_h^\star(s)}{2SA} \cdot (\frac{8S^2A^2C_{\boldsymbol{\phi}}^2}{\epsilon_{\mathrm{exp}}} + 8S^2A^2C_{\boldsymbol{\phi}}^2K_{\mathrm{unif}} + K_{\mathrm{unif}})$. Since $|[\boldsymbol{\phi}]_{(s,a)}| \leq C_{\boldsymbol{\phi}} \cdot (W_h^\star(s) + \sqrt{\epsilon_\phi})$ and $\epsilon_\phi/(K_{\mathrm{unif}}\epsilon_{\mathrm{unif}}) \leq \epsilon_{\mathrm{exp}}$ by assumption, some manipulation shows that

$$\frac{[\boldsymbol{\phi}]_{(s,a)}^2}{N_h(s,a)} \leq \frac{C_{\boldsymbol{\phi}}^2 \cdot (W_h^\star(s)+\sqrt{\epsilon_\phi})^2}{\frac{W_h^\star(s)}{2SA} \cdot (\frac{8S^2A^2C_{\boldsymbol{\phi}}^2}{\epsilon_{\mathrm{exp}}} + 8S^2A^2C_{\boldsymbol{\phi}}^2K_{\mathrm{unif}} + K_{\mathrm{unif}})} \leq \frac{\epsilon_{\mathrm{exp}}}{SA}.$$

It follows then that, letting $\widehat{\boldsymbol{\Sigma}}$ denote the covariance obtained by the call to UNIFEXP on Line 12,

$$\max_{\boldsymbol{\phi}\in\Phi} \|\boldsymbol{\phi}\|_{\widehat{\boldsymbol{\Sigma}}^{-1}}^2 \leq \epsilon_{\mathrm{exp}}$$

as desired. Furthermore, it is straightforward to see that $[\widehat{\boldsymbol{\Sigma}}]_{(s,a)} \geq \frac{\epsilon_{\mathrm{unif}}}{2SA} \cdot K_{\mathrm{unif}}$ for $s \in \mathcal{S}_0$ as well.

To complete the proof, we union bound over these events holding, and take the minimum of the sample complexity bounds from either case.

□

F.1 Data Conditioning

Lemma 33. *Consider running any algorithm for K episodes. Let $K_h(s,a)$ denote the number of visits to (s,a,h). Then with probability at least $1-\delta$, for all (s,a,h) simultaneously, we have*

$$K_h(s,a) \leq W_h^\star(s)K + \sqrt{2W_h^\star(s)K \cdot \log\frac{SAH}{\delta}} + \log\frac{SAH}{\delta}.$$

Proof. By definition, we have

$$\sup_\pi w_h^\pi(s) = W_h^\star(s).$$

This implies that any policy will reach (s,h) with probability at most $W_h^\star(s)$. We can therefore think of this as the sum of Bernoullis with parameter at most $W_h^\star(s)$, so the bound follows by applying Bernstein's inequality and a union bound. □

Lemma 34. *Consider the set*

$$\widehat{\boldsymbol{\Omega}}_{h,K}(\delta) := \left\{ \mathrm{diag}(\boldsymbol{v}) \ : \ \boldsymbol{v} \in \mathbb{R}_+^{SA}, [\boldsymbol{v}]_{(s,a)} \leq W_h^\star(s) + \sqrt{\frac{2W_h^\star(s)}{K} \cdot \log\frac{SAH}{\delta}} + \frac{1}{K}\log\frac{SAH}{\delta} \right\}.$$

Consider running some set of policies for K episodes, and let $\widehat{\boldsymbol{\Lambda}}$ be defined as

$$\widehat{\boldsymbol{\Lambda}}_h = \mathrm{diag}(\widehat{\boldsymbol{v}}), \quad [\boldsymbol{v}]_{(s,a)} = \frac{K_h(s,a)}{K}.$$

Then with probability at least $1-\delta$, we have that $\widehat{\boldsymbol{\Lambda}}_h \in \widehat{\boldsymbol{\Omega}}_{h,K}(\delta)$ for all $h \in [H]$ simultaneously.

Proof. This is an immediate consequence of Lemma 33. □

We will denote $\widehat{\boldsymbol{\Omega}}_{h,K} := \widehat{\boldsymbol{\Omega}}_{h,K}(\delta)$ when the choice of δ is clear from context.

Lemma 35. *Consider the function*

$$f(\boldsymbol{\Lambda}) = \frac{1}{\eta}\log\left(\sum_{\boldsymbol{\phi}\in\Phi} e^{\eta\|\boldsymbol{\phi}\|^2_{\mathbf{A}(\boldsymbol{\Lambda})^{-1}}}\right) \quad \textit{for} \quad \mathbf{A}(\boldsymbol{\Lambda}) = \boldsymbol{\Lambda} + \boldsymbol{\Lambda}_0$$

Assume that for all $\phi \in \Phi$ we have

$$\max_{\phi \in \Phi} |[\phi]_{(s,a)}| \leq C_\phi \cdot (W_h^\star(s) + \epsilon), \quad \forall s \in \mathcal{S}_0$$

for some $\mathcal{S}_0$ and some $C_\phi, \epsilon > 0$, and otherwise $[\phi]_{(s,a)} = 0$. Assume that $\mathbf{\Lambda}_0 = \mathrm{diag}(\boldsymbol{v})$ for some $\boldsymbol{v}$ satisfying

$$[\boldsymbol{v}]_{(s,a)} \geq C' \cdot W_h^\star(s), \quad \forall s \in \mathcal{S}_0$$

and otherwise $[\boldsymbol{v}]_{(s,a)} \geq \lambda$, for some $C', \lambda > 0$. Then we can bound

$$\begin{aligned}
&\sup_{\widehat{\mathbf{\Lambda}}, \widehat{\mathbf{\Lambda}}' \in \widehat{\mathbf{\Omega}}_{h,K}} |\nabla_{\mathbf{\Lambda}} f(\mathbf{\Lambda})|_{\mathbf{\Lambda}=\widehat{\mathbf{\Lambda}}}[\widehat{\mathbf{\Lambda}}']| \\
&\leq \max_{s \in \mathcal{S}_0} \left(\frac{2SAC_\phi^2}{C'} + \frac{2SAC_\phi^2 \epsilon^2}{C' \cdot W_h^\star(s)} \right) \cdot \left(\frac{2}{C'} + \frac{2}{C'KW_h^\star(s)} \cdot \log \frac{SAH}{\delta} \right)
\end{aligned}$$

and

$$\begin{aligned}
&\sup_{\widehat{\mathbf{\Lambda}}, \widehat{\mathbf{\Lambda}}', \widehat{\mathbf{\Lambda}}'' \in \widehat{\mathbf{\Omega}}_{h,K}} |\nabla^2_{\mathbf{\Lambda}} f(\mathbf{\Lambda})|_{\mathbf{\Lambda}=\widehat{\mathbf{\Lambda}}}[\widehat{\mathbf{\Lambda}}', \widehat{\mathbf{\Lambda}}'']| \\
&\leq \max_{s \in \mathcal{S}_0} (2 + 2\eta) \left(\frac{2SAC_\phi^2}{C'} + \frac{2SAC_\phi^2 \epsilon^2}{C' \cdot W_h^\star(s)} \right)^2 \cdot \left(\frac{2}{C'} + \frac{2}{C'KW_h^\star(s)} \cdot \log \frac{SAH}{\delta} \right)^2.
\end{aligned}$$

Proof. By Lemma D.5 of [42], we have that

$$\nabla_{\mathbf{\Lambda}} f(\mathbf{\Lambda})|_{\mathbf{\Lambda}=\widehat{\mathbf{\Lambda}}}[\widehat{\mathbf{\Lambda}}'] = - \left(\sum_{\phi \in \Phi} e^{\eta \|\phi\|^2_{\mathbf{A}(\widehat{\mathbf{\Lambda}})^{-1}}} \right) \cdot \sum_{\phi \in \Phi} e^{\eta \|\phi\|^2_{\mathbf{A}(\widehat{\mathbf{\Lambda}})^{-1}}} \phi^\top \mathbf{A}(\widehat{\mathbf{\Lambda}})^{-1} \widehat{\mathbf{\Lambda}}' \mathbf{A}(\widehat{\mathbf{\Lambda}})^{-1} \phi.$$

We have

$$\phi^\top \mathbf{A}(\widehat{\mathbf{\Lambda}})^{-1} \widehat{\mathbf{\Lambda}}' \mathbf{A}(\widehat{\mathbf{\Lambda}})^{-1} \phi = \sum_{s,a} \frac{[\phi]^2_{(s,a)} \cdot [\widehat{\mathbf{\Lambda}}']_{(s,a)}}{[\mathbf{A}(\widehat{\mathbf{\Lambda}})]^2_{(s,a)}} = \sum_{s \in \mathcal{S}_0} \sum_a \frac{[\phi]^2_{(s,a)} \cdot [\widehat{\mathbf{\Lambda}}']_{(s,a)}}{[\mathbf{A}(\widehat{\mathbf{\Lambda}})]^2_{(s,a)}}$$

where the last equality follows since, for $s \notin \mathcal{S}_0$, we have assumed $[\phi]_{(s,a)} = 0$.

Now consider some $s \in \mathcal{S}_0$. By assumption we have $[\phi]^2_{(s,a)} \leq 2C_\phi^2 \cdot (W_h^\star(s)^2 + \epsilon^2)$ and by our assumption on $\mathbf{\Lambda}_0$ we can lower bound $[\mathbf{A}(\widehat{\mathbf{\Lambda}})]_{(s,a)} \geq C' \cdot W_h^\star(s)$. Furthermore, since $\widehat{\mathbf{\Lambda}}' \in \widehat{\mathbf{\Omega}}_{h,K}$, we have

$$\begin{aligned}
[\widehat{\mathbf{\Lambda}}']_{(s,a)} &\leq W_h^\star(s) + \sqrt{\frac{2W_h^\star(s)}{K} \cdot \log \frac{SAH}{\delta}} + \frac{1}{K} \log \frac{SAH}{\delta} \\
&\leq 2W_h^\star(s) + \frac{2}{K} \log \frac{SAH}{\delta}.
\end{aligned}$$

Putting this together, we have

$$\begin{aligned}
\frac{[\phi]^2_{(s,a)} \cdot [\widehat{\mathbf{\Lambda}}']_{(s,a)}}{[\mathbf{A}(\widehat{\mathbf{\Lambda}})]^2_{(s,a)}} &\leq \frac{4C_\phi^2 \cdot (W_h^\star(s)^2 + \epsilon^2) \cdot (W_h^\star(s) + \frac{1}{K} \log \frac{SAH}{\delta})}{(C' \cdot W_h^\star(s))^2} \\
&\leq \left(\frac{2C_\phi^2}{C'} + \frac{2C_\phi^2 \epsilon^2}{C' W_h^\star(s)} \right) \cdot \left(\frac{2}{C'} + \frac{2}{C'KW_h^\star(s)} \log \frac{SAH}{\delta} \right).
\end{aligned}$$

It follows that

$$\sum_{s \in \mathcal{S}_0} \sum_a \frac{[\phi]^2_{(s,a)} \cdot [\widehat{\mathbf{\Lambda}}']_{(s,a)}}{[\mathbf{A}(\widehat{\mathbf{\Lambda}})]^2_{(s,a)}} \leq \max_{s \in \mathcal{S}_0} \left(\frac{2SAC_\phi^2}{C'} + \frac{2SAC_\phi^2 \epsilon^2}{C' W_h^\star(s)} \right) \cdot \left(\frac{2}{C'} + \frac{2}{C'KW_h^\star(s)} \log \frac{SAH}{\delta} \right).$$

The second bound follows in an analogous fashion, using the expression for the second derivative given in Lemma D.5 of [42].

□

https://doi.org/10.52202/079017-0717

Algorithm 4 Uniform Exploration (UNIFEXP)

input: tolerance ϵ_{unif}, reruns K, confidence δ, step h
$\mathfrak{D} \leftarrow \emptyset$
for $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**
  // LEARN2EXPLORE is as defined in [46]
  $\{(\mathcal{X}_j, \Pi_j, N_j)\}_{j=1}^{\lceil \log_2 1/\epsilon_{\text{unif}} \rceil} \leftarrow$ LEARN2EXPLORE$(\{(s,a)\}, h, \frac{\delta}{2SA}, \frac{\delta}{2KSA}, \epsilon_{\text{unif}})$
  if $\exists j_{sa}$ such that $(s,a) \in \mathcal{X}_{j_{sa}}$ **then**
    Rerun every policy in $\Pi_{j_{sa}}$ $K_{sa} := \lceil \frac{K}{SA|\Pi_{j_{sa}}|} \rceil$ times, store observed transitions in $\mathfrak{D}$
  end if
end for
return $\mathfrak{D}$

Lemma 36. *With probability at least $1-\delta$, Algorithm 4 will terminate after running for at most*

$$K + \text{poly}(S, A, H, \log \frac{K}{\delta \epsilon_{\text{unif}}}) \cdot \frac{1}{\epsilon_{\text{unif}}}$$

episodes and will collect at least $\frac{W_h^\star(s)K}{2SA}$ samples from each (s,a) such that $W_h^\star(s) > \epsilon_{\text{unif}}$.

Proof. By Theorem 13 of [46], with probability at least $1 - \delta/2SA$, for any (s,a):

- LEARN2EXPLORE will run for at most $\text{poly}(S, A, H, \log \frac{K}{\delta \epsilon_{\text{unif}}}) \cdot \frac{1}{\epsilon_{\text{unif}}}$ episodes.
- Rerunning every policy in $\Pi_{j_{sa}}$ once, with probability at least $1 - \delta/K$ we will collect $N = 2^{-j_{sa}}|\Pi_{j_{sa}}|$ samples from (s,a), for $|\Pi_{j_{sa}}| = \mathcal{O}(2^{j_{sa}} \cdot S^3A^2H^4 \log^3 1/\delta)$.
- We have that $W_h^\star(s) \le 2^{-j_{sa}+1}$.
- IF $(s,a) \notin \mathcal{X}_j$ for all $j = 1, 2, \ldots, \lceil \log 1/\epsilon_{\text{unif}} \rceil$, then $W_h^\star(s) \le \epsilon_{\text{unif}}$.

By the above conclusions, rerunning policies in $\Pi_{j_{sa}}$ on Line 7, with probability at least $1 - \delta/2SA$ we will collect

$$N \cdot K_{sa} \ge N \cdot \frac{K}{SA|\Pi_{j_{sa}}|} = \frac{2^{-j_{sa}}K}{SA}$$

samples from (s,a). As noted, $W_h^\star(s) \le 2^{-j_{sa}+1}$, so this implies that we will collect at least $\frac{W_h^\star(s)K}{2SA}$ samples from (s,a). Union bounding over this holding for all (s,a), and noting that we only fail to collect this many samples if $W_h^\star(s) \le \epsilon_{\text{unif}}$ gives the collection guarantee.

To bound the total number of episodes, we note that the procedure on Line 7 will, in total collect at most

$$\sum_{s,a:j_{sa}\text{ exists}} |\Pi_{j_{sa}}|\lceil K_{sa} \rceil \le \sum_{s,a:j_{sa}\text{ exists}} |\Pi_{j_{sa}}| + \sum_{s,a} \frac{K}{SA} = \sum_{s,a} |\Pi_{j_{sa}}| + K$$

episodes. IF j_{sa} exists, this implies that $|\Pi_{j_{sa}}| \le \mathcal{O}(2^{j_{sa}} \cdot S^3A^2H^4 \log^3 1/\delta)$, and since $j_{sa} \in \{1, 2, \ldots, \lceil \log 1/\epsilon_{\text{unif}} \rceil\}$, this implies that the above is bounded by

$$K + \mathcal{O}(\epsilon_{\text{unif}}^{-1} \cdot S^3A^2H^4 \log^3 1/\delta).$$

Combining this with our bound on the total number of episodes collected by LEARN2EXPLORE, we have that the number of episodes collected by Algorithm 4 is bounded by

$$K + \text{poly}(S, A, H, \log \frac{K}{\delta \epsilon_{\text{unif}}}) \cdot \frac{1}{\epsilon_{\text{unif}}}.$$

□

F.2 Online Frank-Wolfe

Lemma 37. *Let*

$$f_i(\mathbf{\Lambda}) = \frac{1}{\eta_i}\log\left(\sum_{\boldsymbol{\phi}\in\Phi} e^{\eta_i\|\boldsymbol{\phi}\|^2_{\mathbf{A}_i(\mathbf{\Lambda})^{-1}}}\right), \quad \mathbf{A}_i(\mathbf{\Lambda}) = \mathbf{\Lambda} + \frac{1}{T_iK_i}\mathbf{\Lambda}_{0,i}$$

for some $\mathbf{\Lambda}_{0,i}$ satisfying $\mathbf{\Lambda}_{0,i} \succeq \mathbf{\Lambda}_0$ for all i, and $\eta_i = 2^{2i/5}$. Let (β_i, M_i) denote the smoothness and magnitude constants for f_i. Let (β, M) be some values such that $\beta_i \le \eta_i\beta, M_i \le M$ for all i, and R the diameter of the domain of possible values of $\mathbf{\Lambda}$.

Then, if we run Algorithm 4 of [43] on $(f_i)_i$ with constraint tolerance ϵ and confidence δ and $K_i = T_i = 2^i$, we have that with probability at least $1-\delta$, it will run for at most

$$\max\left\{\min_N 16N \text{ s.t. } \inf_{\mathbf{\Lambda}\in\mathbf{\Omega}}\max_{\boldsymbol{\phi}\in\Phi}\boldsymbol{\phi}^\top(N\mathbf{\Lambda}+\mathbf{\Lambda}_0)^{-1}\boldsymbol{\phi} \le \frac{\epsilon}{6}, \frac{\mathrm{poly}(\beta,R,d,H,M,\log 1/\delta,\log|\Phi|)}{\epsilon^{4/5}}\right\}.$$

episodes, and will return data $\{\boldsymbol{\phi}_\tau\}_{\tau=1}^N$ with covariance $\widehat{\mathbf{\Sigma}}_N = \sum_{\tau=1}^N \boldsymbol{\phi}_\tau\boldsymbol{\phi}_\tau^\top$ such that

$$f_{\widehat{i}}(N^{-1}\widehat{\mathbf{\Sigma}}_N) \le N\epsilon,$$

where $\widehat{i}$ is the iteration on which OptCov *terminates.*

Proof. Our goal is to simply find a setting of i that is sufficiently large to guarantee the condition $f_i(\widehat{\mathbf{\Lambda}}_i) \le K_iT_i\epsilon$ is met. By Lemma C.1 of [43], we have with probability at least $1-\delta/(2i^2)$:

$$\begin{aligned}
f_i(\widehat{\mathbf{\Lambda}}_i) &\le \inf_{\mathbf{\Lambda}\in\mathbf{\Omega}} f_i(\mathbf{\Lambda}) + \frac{\beta_iR^2(\log T_i+3)}{2T_i} + \sqrt{\frac{4M^2\log(8i^2T_i/\delta)}{K_i}} \\
&\quad + \sqrt{\frac{c_1M^2d^4H^4\log^3(8i^2HK_iT_i/\delta)}{K_i}} + \frac{c_2Md^4H^3\log^{7/2}(4i^2HK_iT_i/\delta)}{K_i} \\
&\le 3\max\left\{\inf_{\mathbf{\Lambda}\in\mathbf{\Omega}} f_i(\mathbf{\Lambda}), \frac{\beta_iR^2(\log T_i+3)}{2T_i}, \sqrt{\frac{4M^2\log(8i^2T_i/\delta)}{K_i}}\right. \\
&\quad \left. + \sqrt{\frac{c_1M^2d^4H^4\log^3(8i^2HK_iT_i/\delta)}{K_i}} + \frac{c_2Md^4H^3\log^{7/2}(4i^2HK_iT_i/\delta)}{K_i}\right\}.
\end{aligned}$$

So a sufficient condition for $f_i(\widehat{\mathbf{\Lambda}}_i) \le K_iT_i\epsilon$ is that

$$\begin{aligned}
K_iT_i \ge \frac{3}{\epsilon}\max&\left\{\inf_{\mathbf{\Lambda}\in\mathbf{\Omega}} f_i(\mathbf{\Lambda}), \frac{\beta_iR^2(\log T_i+3)}{2T_i}, \sqrt{\frac{4M^2\log(8i^2T_i/\delta)}{K_i}}\right. \\
&\left. + \sqrt{\frac{c_1M^2d^4H^4\log^3(8i^2HK_iT_i/\delta)}{K_i}} + \frac{c_2Md^4H^3\log^{7/2}(4i^2HK_iT_i/\delta)}{K_i}\right\}.
\end{aligned} \tag{F.4}$$

Recall that

$$f_i(\mathbf{\Lambda}) = \frac{1}{\eta_i}\log\left(\sum_{\boldsymbol{\phi}\in\Phi} e^{\eta_i\|\boldsymbol{\phi}\|^2_{\mathbf{A}_i(\mathbf{\Lambda})^{-1}}}\right), \quad \mathbf{A}_i(\mathbf{\Lambda}) = \mathbf{\Lambda} + \frac{1}{T_iK_i}\mathbf{\Lambda}_{0,i}.$$

By Lemma D.1 of [42], we can bound

$$\max_{\boldsymbol{\phi}\in\Phi}\|\boldsymbol{\phi}\|^2_{\mathbf{A}_i(\mathbf{\Lambda})^{-1}} \le f_i(\mathbf{\Lambda}) \le \max_{\boldsymbol{\phi}\in\Phi}\|\boldsymbol{\phi}\|^2_{\mathbf{A}_i(\mathbf{\Lambda})^{-1}} + \frac{\log|\Phi|}{\eta_i}.$$

Thus,

$$\begin{aligned}
\inf_{\mathbf{\Lambda}\in\mathbf{\Omega}} f_i(\mathbf{\Lambda}) &\le \inf_{\mathbf{\Lambda}\in\mathbf{\Omega}}\max_{\boldsymbol{\phi}\in\Phi}\|\boldsymbol{\phi}\|^2_{\mathbf{A}_i(\mathbf{\Lambda})^{-1}} + \frac{\log|\Phi|}{\eta_i} \\
&= \inf_{\mathbf{\Lambda}\in\mathbf{\Omega}}\max_{\boldsymbol{\phi}\in\Phi} T_iK_i\boldsymbol{\phi}^\top(T_iK_i\mathbf{\Lambda} + \mathbf{\Lambda}_{0,i} + \mathbf{\Lambda}_{\mathrm{off}})^{-1}\boldsymbol{\phi} + \frac{\log|\Phi|}{\eta_i}
\end{aligned}$$

https://doi.org/10.52202/079017-0717

By our choice of $\eta_i = 2^{2i/5}$, and $K_i = 2^i$, $T_i = 2^i$, we can ensure that

$$K_i T_i \geq \frac{6}{\epsilon} \frac{\log |\Phi|}{\eta_i}$$

as long as $i \geq \frac{2}{5} \log_2[\frac{6 \log |\Phi|}{\epsilon}]$. To ensure that

$$T_i K_i \geq \frac{6}{\epsilon} \inf_{\boldsymbol{\Lambda} \in \boldsymbol{\Omega}} \max_{\boldsymbol{\phi} \in \Phi} T_i K_i \boldsymbol{\phi}^\top (T_i K_i \boldsymbol{\Lambda} + \boldsymbol{\Lambda}_{0,i})^{-1} \boldsymbol{\phi}$$

it suffices to take

$$i \geq \arg\min_i i \quad \text{s.t.} \quad \inf_{\boldsymbol{\Lambda} \in \boldsymbol{\Omega}} \max_{\boldsymbol{\phi} \in \Phi} \boldsymbol{\phi}^\top (2^{3i} \boldsymbol{\Lambda} + \boldsymbol{\Lambda}_{0,i})^{-1} \boldsymbol{\phi} \leq \frac{\epsilon}{6}.$$

Since we assume that we can lower bound $\boldsymbol{\Lambda}_{0,i} \succeq \boldsymbol{\Lambda}_0$ for each i, so this can be further simplified to

$$i \geq \arg\min_i i \quad \text{s.t.} \quad \inf_{\boldsymbol{\Lambda} \in \boldsymbol{\Omega}} \max_{\boldsymbol{\phi} \in \Phi} \boldsymbol{\phi}^\top (2^{3i} \boldsymbol{\Lambda} + \boldsymbol{\Lambda}_0)^{-1} \boldsymbol{\phi} \leq \frac{\epsilon}{6}. \tag{F.5}$$

We next want to show that

$$T_i K_i \geq \frac{3}{\epsilon} \cdot \frac{\beta_i R^2 (\log T_i + 3)}{2 T_i}.$$

Bounding $\beta_i \leq \eta_i \beta$, a sufficient condition for this is that

$$i \geq \frac{2}{5} \left(\log_2(12 \beta R^2 i) + \log_2 \frac{1}{\epsilon} \right).$$

By Lemma A.1 of [43], it suffices to take

$$i \geq \frac{6}{5} \log_2(9 \beta R^2 \log_2 \frac{1}{\epsilon}) + \frac{2}{5} \log_2 \frac{1}{\epsilon} \tag{F.6}$$

to meet this condition (this assumes that $12 \beta R^2 \geq 1$ and $\frac{2}{5} \log_2 \frac{1}{\epsilon} \geq 1$—if either of these is not the case we can just replace them with 1 without changing the validity of the final result).

Finally, we want to ensure that

$$T_i K_i \geq \frac{3}{\epsilon} \Bigg(\sqrt{\frac{4 M^2 \log(8 i^2 T_i / \delta)}{K_i}} + \sqrt{\frac{c_1 M^2 d^4 H^4 \log^3(8 i^2 H K_i T_i / \delta)}{K_i}} + \frac{c_2 M d^4 H^3 \log^{7/2}(4 i^2 H K_i T_i / \delta)}{K_i} \Bigg).$$

To guarantee this, it suffices that

$$2^{5i/2} \geq \frac{c}{\epsilon} \sqrt{M^2 d^4 H^4 i^3 \log^3(iH/\delta)}, \quad 2^{3i} \geq \frac{c}{\epsilon} \cdot M d^4 H^3 i^{7/2} \log^{7/2}(iH/\delta).$$

or

$$i \geq \frac{4}{5} \log_2(cMdHi \log(H/\delta)) + \frac{2}{5} \log_2 \frac{1}{\epsilon}, \quad i \geq \frac{4}{3} \log_2(cMdH \log(H/\delta)) + \frac{1}{3} \log_2 \frac{1}{\epsilon}.$$

By Lemma A.1 of [43], it then suffices to take

$$\begin{aligned} i &\geq \frac{12}{5} \log(cMdH \log(H/\delta) \log_2 1/\epsilon) + \frac{2}{5} \log_2 \frac{1}{\epsilon}, \\ i &\geq 4 \log_2(cMdH \log(H/\delta) \log_2 1/\epsilon) + \frac{1}{3} \log_2 \frac{1}{\epsilon} \end{aligned} \tag{F.7}$$

Thus, a sufficient condition to guarantee (F.4) is that i is large enough to satisfy (F.5), (F.6), and (F.7) and $i \geq \frac{2}{5} \log_2[\frac{6 \log |\Phi|}{\epsilon}]$.

If $\widehat{i}$ is the final round, the total complexity scales as

$$\sum_{i=1}^{\widehat{i}} T_i K_i = \sum_{i=1}^{\widehat{i}} 2^{2i} \leq 2 \cdot 2^{2\widehat{i}}.$$

Using the sufficient condition on i given above, we can bound the total complexity as

$$\max \left\{ \min_N 16N \text{ s.t. } \inf_{\boldsymbol{\Lambda} \in \boldsymbol{\Omega}} \max_{\boldsymbol{\phi} \in \Phi} \boldsymbol{\phi}^\top (N \boldsymbol{\Lambda} + \boldsymbol{\Lambda}_0)^{-1} \boldsymbol{\phi} \leq \frac{\epsilon}{6}, \frac{\text{poly}(\beta, R, d, H, M, \log 1/\delta, \log |\Phi|)}{\epsilon^{4/5}} \right\}.$$

□

F.3 Pruning Hard-to-Reach States

Algorithm 5 PRUNE: Prune Hard-to-Reach States

input: tolerance ϵ_{unif}, confidence δ
$\mathcal{S}^{\text{keep}} \leftarrow \emptyset$
for $h \in [H]$ **do**
  for $s \in \mathcal{S}$ **do**
    // LEARN2EXPLORE is as defined in [46]
    $\{(\mathcal{X}_j, \Pi_j, N_j)\}_{j=1}^{\lceil \log_2 \frac{1}{32\epsilon_{\text{unif}}} \rceil} \leftarrow$ LEARN2EXPLORE$(\{(s,a)\}, h, \frac{\delta}{SH}, \frac{1}{2}, 32\epsilon_{\text{unif}})$ for any $a \in \mathcal{A}$
    if $\exists j_s$ such that $(s,a) \in \mathcal{X}_{j_s}$ **then**
      $\mathcal{S}^{\text{keep}} = \mathcal{S}^{\text{keep}} \cup \{(s,h)\}$
    end if
  end for
end for
return $\mathcal{S}^{\text{keep}}$

Lemma 38. *With probability at least $1-\delta$, Algorithm 5 will terminate after running for at most*

$$\text{poly}(S, A, H, \log \frac{1}{\delta\epsilon_{\text{unif}}}) \cdot \frac{1}{\epsilon_{\text{unif}}}$$

episodes and will return a set $\mathcal{S}^{\text{keep}}$ such that, for every $(s,h) \in \mathcal{S}^{\text{keep}}$, we have $W_h^\star(s) \geq \epsilon_{\text{unif}}$, and, if $(s,h) \notin \mathcal{S}^{\text{keep}}$, then $W_h^\star(s) \leq 32\epsilon_{\text{unif}}$.

Proof. As in Lemma 36, by Theorem 13 of [46], with probability at least $1 - \delta/SH$, for any (s,h):

- LEARN2EXPLORE will run for at most $\text{poly}(S, A, H, \log \frac{1}{\delta\epsilon_{\text{unif}}}) \cdot \frac{1}{\epsilon_{\text{unif}}}$ episodes.
- Rerunning every policy in Π_{j_s} once, with probability at least $1/2$ we will collect $N = 2^{-j_s}|\Pi_{j_s}|$ samples from (s,a,h).
- If $(s,a) \notin \mathcal{X}_j$ for all $j = 1, 2, \ldots, \lceil \log 1/\epsilon_{\text{unif}} \rceil$, then $W_h^\star(s) \leq 32\epsilon_{\text{unif}}$.

We union bound over this event holding for all (s,h), which occurs with probability at least $1-\delta$.

It is immediate by the last property that, if $(s,h) \notin \mathcal{S}^{\text{keep}}$ then $W_h^\star(s) \leq 32\epsilon_{\text{unif}}$.

We next show that if $(s,h) \in \mathcal{S}^{\text{keep}}$, then this implies that $W_h^\star(s) \geq \epsilon_{\text{unif}}$. Let X be a random variable denoting the total number of samples we collect from (s,a,h) when rerunning all policies in Π_{j_s}. Then by Markov's Inequality, by the above properties we have

$$\frac{1}{2} \leq \mathbb{P}[X \geq N_{j_s}/2] \leq \frac{2\mathbb{E}[X]}{N_{j_s}} \leq \frac{2|\Pi_{j_s}|W_h^\star(s)}{N_{j_s}} = 8 \cdot 2^{j_s} W_h^\star(s).$$

It follows that

$$W_h^\star(s) \geq \frac{1}{16 \cdot 2^{j_s}} \geq \frac{1}{16 \cdot 2^{\lceil \log_2 \frac{1}{32\epsilon_{\text{unif}}} \rceil}} \geq \frac{1}{32 \cdot 2^{\log_2 \frac{1}{32\epsilon_{\text{unif}}}}} = \epsilon_{\text{unif}}.$$

This completes the proof.

□

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

https://doi.org/10.52202/079017-0717

Justification: These can be found in the abstract and the contributions section of the introduction.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: These can be found in the Discussion section of the main body of the paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All proofs are found in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.

- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: The contributions of this paper are entirely theoretical.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA]

   Justification: The contributions of this paper are entirely theoretical.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.

https://doi.org/10.52202/079017-0717

- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: The contributions of this paper are entirely theoretical.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: The contributions of this paper are entirely theoretical.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: The contributions of this paper are entirely theoretical.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: There are no human subjects and we discuss the ethical consequences in the "Broader Impact" section of the discussion.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: This theoretical paper poses minimal public concerns but holds significant potential to inspire advancements in algorithm development, contributing positively to the field of machine learning.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

https://doi.org/10.52202/079017-0717

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The contributions are theoretical.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: The contributions are theoretical, and we do not use any such assets. For prior theoretical work, we have credited the authors appropriately.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The contributions are theoretical.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: No human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: No human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-0717