Efficient Exploration in Linear MDPs with Nonlinear Confounding Rewards

Anonymous Author
Anonymous Institution

Abstract

Recent theoretical work establishes provably efficient algorithms using linear function approximation when the rewards are linear. However, nonlinear reward signals are common in practice. We therefore seek to construct similarly provably efficient algorithms using linear function approximation—even when the underlying value function is nonlinear. We extend the linear Markov decision process setting to consider a reward structure that has a linear component, confounded by a nonlinear adversarial baseline reward that is out of the agent's control. We show that knowing the linear component of the reward is sufficient to learn an optimal policy, and therefore propose to isolate the linear component by averaging out the nonlinear confounding reward via importance sampling. Our algorithm combines this with optimism under uncertainty exploration to provably achieve a regret bound of $\widetilde{\mathcal{O}}(T^{3/4})$, where T is the total number of steps. This contrasts recent work, which suffers from linear regret due to the additional nonlinear rewards.

1 Introduction

Many existing reinforcement learning (RL) algorithms [Mnih et al., 2015; Gu et al., 2017] require huge amounts of data to explore the environment and learn good decisions. This amount of data may be infeasible or too costly to support in domains that involve interacting with people, like an online shopping session or an application providing healthy living nudges. Thus we seek RL algorithms that learn with provable efficiency guarantees in complex domains. In some important cases, the complexity of domains may come from aspects *outside the control of the decision-maker*: e.g., a customer's shopping experience may be influenced by the number of other household members using internet bandwidth, the amount of sleep they had last night, and whether they are worried about an upcoming election. Ideally, an RL algorithm's learning efficiency would only depend on the complexity of the aspects of the reward impacted by the RL algorithm's choices, but we don't typically have access a priori to a model of these additional complexities.

Function approximation is a popular approach to learning in such complex spaces. However, such approaches have long been known to have potential shortcomings: results in the mid-1990s illustrated how popular algorithms could fail to even converge in certain cases [Baird, 1995]. In recent years there has been significant empirical success by new methods that largely avoid such instabilities, but the theory of RL in complex, large state or action spaces remains largely open.

Recent progress on theoretical RL with function approximation provides promising regret guarantees under strong restrictive assumptions, such as linear dynamics and reward models [Yang and Wang, 2020; Zanette et al., 2020a; Jin et al., 2020b], or requiring that the value function can be represented as a linear structure with zero or low inherent Bellman error [Zanette et al., 2020b]. Unfortunately, when such strong assumptions are violated, existing approaches typically provide no performance guarantees or incur additional linear regret if the algorithm is provided a bound on the amount of model misspecification. Closer to our previously described setting with decision-independent aspects is algorithmic work [Dietterich et al., 2018; Chitnis and Lozano-Pérez, 2020] that assumes some part of the state space may be (approximately) independent of the agent's actions, but it does not consider strategic exploration or offer performance guarantees. Interestingly, in the 1-step

Preliminary work. Under review by AISTATS 2021. Do not distribute.

decision-making literature in econometrics and healthcare research, it is quite common to focus on directly modeling the treatment effect; that is, the difference in expected outcomes between taking an action (e.g., a medication, a job training program) vs. taking no action. Recent work on contextual bandits [Greenewald et al., 2017; Krishnamurthy et al., 2018] takes a similar approach to avoid modeling a potentially complicated action-independent influence on the reward signal, and obtains sublinear theoretical regret bounds.

Inspired by these recent bandit results, we seek similar strong theoretical guarantees in the multi-step RL setting where rewards may be influenced by a complex episode-dependent baseline. We assume that the reward at each state-action in an episode consists of two parts: a simpler reward that depends on the agent's actions, and a potentially much more complicated *confounding reward* that is independent of the agent's actions and may adversarially vary across episodes. These assumptions can model scenarios like a shopping browsing session, a customer support phone call, or health nudges over a day, where we expect that an external baseline might systematically elevate or depress all rewards within an episode. Critically, this external baseline is unknown. Note that unlike in the contextual bandit setting, a confounding reward signal that changes the reward at each state, (even if the change is independent of the actions), can impact the optimal policy in RL settings. In contrast, as we show in Section 4, the confounding reward we consider still preserves the optimal policy but allows for a richer set of more realistic modeling assumptions that enable better robustness.

We present an algorithm, called Action-Centered Value Iteration (ACVI), that strategically explores and provably achieves sublinear regret in such spaces given a linear MDP structure [Jin et al., 2020b] with episodic confounding (Section 5). Our algorithm builds on optimistic Least-Squares Value Iteration (LSVI) and uses an action-centering approach from the contextual bandits setting [Greenewald et al., 2017; Krishnamurthy et al., 2018] to isolate the action-dependent rewards from the confounding rewards. Unlike the bandits setting, we must carefully control how the increased variability of our underlying reward estimator impacts the resulting exploration.

The key ingredients of our algorithm are: 1) an optimism bonus based on a bi-level upper confidence bound (UCB) on reward and value function, and 2) a schedule of the sampling probabilities that eventually ensures the greedy action is always chosen with high probability but still provides sufficient uncertainty for importance sampling as well. Our proposed algorithm achieves $\widetilde{\mathcal{O}}(d^2H^{3/2}T^{3/4})$ expected regret with high probability, where d is the dimension of the feature space, H is the length of the horizon, and T is the overall number of steps across all episodes. Our algorithm empirically achieves low regret on synthetic experiments, while existing (misspecified) optimistic LSVI algorithms for the linear MDP setting [Jin et al., 2020b] empirically are not robust to the confounding reward and achieve higher regret. Our results provide a useful step towards methods that can handle more complicated settings while preserving strong theoretical guarantees.

2 Related Work

Efficient exploration with linear value function approximation. Function approximation is necessary for reinforcement learning in large state-action spaces. In recent years, provably efficient exploration with linear value function approximation has attracted significant attention. A popular set of frameworks that yield formal guarantees for exploration are linear/low-rank MDPs [Yang and Wang, 2020; Zanette et al., 2020a; Jin et al., 2020b] and their extensions [Wang et al., 2019, 2020], which essentially assume that any function will be mapped to a linear function of features by the Bellman operator. Our work is based on similar assumptions on transition dynamics and further studies the case when the reward consists of a linear component and a nonlinear baseline component. Several recent works do not make the linear MDP assumption, and instead assume policy-value functions are linear [Lattimore and Szepesvari, 2020] or have low inherent Bellman error with the linear function class [Zanette et al., 2020b,c].

Nonlinear, confounding rewards in contextual bandits. Greenewald et al. [2017] and Krishnamurthy et al. [2018] propose two different approaches to learn linear functions to solve contextual bandit from nonlinear feedback signals. Both assume the nonlinear component only depends on the state in order to make sublinear regret possible. Though the algorithms and regret guarantees are different, the main idea of both algorithms is to sample the actions with additional stochasticity to average out the nonlinear component. This work extends this idea to tackle sequential decision-making problems.

Adversarial MDPs. Previous works for adversarial MDPs mainly consider general adversarial noise in finite state action spaces (e.g. most recently Jin et al. [2020a]), and achieves $\sqrt{T}$-regret by leveraging the ideas from adversarial bandits and online learning algorithms. Lykouris et al. [2019] consider the linear function approximation settings but suffer from a

linear dependency on the total amount of adversarial corruptions. Instead, this work assumes a structure on the adversarial noise such that it will corrupt the observed rewards but will not change the optimal policies, thus making sublinear regret with linear function approximation possible.

3 Problem Setting

In this section, we formally define the problem setting that we study. We first discuss the preliminaries of the episodic RL setting (Section 3.1). Then, we formalize additional assumptions on linear transition dynamics and nonlinear confounding rewards (Section 3.2).

3.1 Episodic reinforcement learning

We consider a sequential decision process over K episodes. In the k-th episode, the agent interacts with an environment defined by $\langle \mathcal{S}, \mathcal{A}, \Pr, H, r^k \rangle$, where the initial state can be chosen adversarially. All episodes share the same states $\mathcal{S}$; actions $\mathcal{A} = \{0, \dots, N\}$; time-dependent Markov transition dynamics $\Pr = \{\Pr_h\}_{h=1}^H$, where $\Pr_h(\cdot|s,a)$ is the distribution over next states after taking action a at state s on time-step h; and episode length H. The time-dependent rewards are $r^k = \{r_h^k\}_{h=1}^H$, where $r_h^k(s,a)$ is the deterministic reward achieved by taking action a at state s and time-step h. Notice here that we allow additional flexibility in reward as $r_h^k(s,a)$ can vary across the episodes. In the next section, we will introduce a type of reward structure we study in this paper so that learning a Markovian policy is still sufficient. In particular, a component of the reward (described in Section 3.2) is allowed to adversarially depend on the initial state of the episode and the history of states and actions over the previous $k-1$ episodes, $\bar{\mathcal{H}}^{k-1} = \mathcal{H}_1, \dots, \mathcal{H}_{k-1}$, where $\mathcal{H}_\tau = \{s_h^\tau, a_h^\tau\}_{h=1}^H$.

We denote the state and action at timestep h on the k-th episode as s_h^k and a_h^k, which yields reward $r_h^k(s_h^k, a_h^k)$ and transitions to $s_{h+1}^k \sim \Pr_h(\cdot \mid s_h^k, a_h^k)$. The agent's goal is to learn a policy $\pi_k : \mathcal{S} \times [H] \to A$, that maximizes the expected returns given the r_k in each episode. Formally, we define the policy value function as the expected returns of a policy π_k at episode k starting from state s at time-step h:

$$V_{h,r^k}^{\pi_k}(s) \coloneqq \mathbb{E}_{\pi_k}\left[\sum_{h'=h}^{H} r_{h'}^k(s_{h'}^k, \pi_k(s_{h'}^k, h'))\Big| s_h^k = s\right].$$

Note that we define the value function at the k-th episodes given that the reward specification r_k is fixed. Then, we define the cumulative regret over K episodes of a policy $\{\pi_1, \dots \pi_k\}$ with respect to the optimal policy as:

$$\text{Regret}(K) \coloneqq \sum_{k=1}^{K} \sup_{\pi} V_{1,r^k}^{\pi}(s) - V_{1,r^k}^{\pi_k}(s_1^k). \quad (1)$$

3.2 Linear MDP with confounding rewards

A common assumption [Jin et al., 2020b] is that the rewards and transition dynamics are linear in a feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, which further implies a linear value function. We similarly assume that the dynamics take the following linear form:

$$\Pr_h(\cdot \mid s,a) = \phi(s,a)^\top \mu_h(\cdot),$$

where $\mu_h(\cdot) \in \mathbb{R}^d$ is an unknown measure such that $\phi(s,a)^\top \sum_{s' \in \mathcal{S}} \mu_h(s') = 1$ for all s, a, h.

However, as rewards are often non-linear in practice, we seek a more complex and adversarial reward structure. To do this, we consider rewards that consist of: (1) a *linear reward* component r^{lin}, that is still linear in $\phi(s,a)$; (2) a non-linear *confounding reward* component r^{conf}, which, on episode k, may adversarially depend on the history of states and actions $\bar{\mathcal{H}}^{k-1}$, and the current episode's initial state s_1^k; and (3) independent sub-Gaussian noise. The confounding reward component represents a baseline signal that is independent of the interaction with the agent, but can often be more complex due to the complexity of environment.

Following Greenewald et al. [2017], we also assume there is a *default action* $a_{0,h} \in \mathcal{A}$ (often represented as 0) corresponding to "doing nothing" at timestep h. Only the non-linear confounding reward is observed when this default action is taken, and we only incur the baseline reward caused by factors out of the agent's control. More precisely, we assume the following reward structure for all s, a, h, k:

$$r_h^k(s,a) = \underbrace{\langle \phi(s,a), \theta_h \rangle \mathbb{1}\{a \neq a_{h,0}\}}_{\text{linear reward } r_h^{\text{lin}}} + \underbrace{f_h^k\left(s_1^k; \bar{\mathcal{H}}^{k-1}\right)}_{r_{h,k}^{\text{conf}}} + \underbrace{\varepsilon_h^k}_{\text{noise}},$$

where f_h^k is a bounded function and the agent can only observe a real-valued reward $r_h^k(s,a)$ rather than the decomposition or parameters θ_h and f_h^k. Note that the linear term r^{lin} is the same across all episodes, while the confounding reward r^{conf} may adversarially change. For our analysis, we assume that $\|\phi(s,a)\| \leq 1$; $\|\theta_h\|, \|\mu_h(\mathcal{S})\| \leq \sqrt{d}$; $|r_h^{\text{lin}}| \leq 1$, $|r_h^{\text{lin}} + r_{h,k}^{\text{conf}}| \leq r_{\max} < \infty$; and each ε_h^k is independent and σ^2 sub-Gaussian.

4 Equivalence in Optimality to Linear MDPs

Learning the value function in this setting can be very challenging due to the confounding rewards r^{conf}, which can be arbitrarily complex. However, we show below that the optimal policy for our setting depends only on the much simpler linear reward component r^{lin}, and not on the nonlinear confounding component r^{conf}. It is thus possible to learn the optimal policy by estimating the linear reward component (Section 5).

To see that the optimal policy only depends on the linear reward component r^{lin}, we decompose the value-function into the returns due to the linear reward component and the returns due to the confounding rewards and noise, similar to Dietterich et al. [2018]:

$$V^{\pi_k}_{h,r^k}(s) = \mathbb{E}_{\pi_k}\left[\sum_{h'=h}^{H} r^{\text{lin}}_h(s^k_h, a^k_h)\Big| s^k_h = s\right] + \sum_{h'=h}^{H} f^k_h\left(s^k_1; \bar{\mathcal{H}}^{k-1}\right)$$

Only the first term, which is just the returns due to the linear reward component, depends on the actions a^k_h and the policy in the current episode. Therefore, the optimal policy $\pi^\star_k := \arg\max_{\pi_k} V^{\pi_k}_{h,r^k}$ only depends on the linear reward component and is independent of the confounding rewards and noise. In addition, r^{lin}_h is constant across different episodes, so there is an optimal policy $\pi^\star$ that is stationary across episodes, (i.e., $\pi^\star_k = \pi^\star_{k'}$ for all $k, k' \in [H]$).

In particular, we notice that $\pi^\star$ is optimal for the linear MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, H, \Pr, r^{\text{lin}}\rangle$, which can be viewed as setting the confounding rewards and noise to be 0. So we define the policy value function in this MDP as

$$V^{\pi}_{h,\mathcal{M}} = \mathbb{E}_{\pi}\left[\sum_{h'=h}^{H} r^{\text{lin}}_{h'}(s_{h'}, \pi(s_{h'}, h'))\big| s_h = s\right].$$

For the ease of notation we will use V^π_h in the rest of this paper to refer to the value function $V^\pi_{h,\mathcal{M}}$ in the linear MDP. Notably, the regret with respect to the overall reward as we defined in Equation 1 is the same as the regret in the linear MDP $\mathcal{M}$:

$$\text{Regret}(K) = \sum_{k=1}^{K} V^{\pi^\star}_1(s^k_1) - V^{\pi_k}_1(s^k_1), \tag{2}$$

We denote the value function of the optimal policy as $V^\star_h(s) \coloneqq V^{\pi^\star}_h(s)$ and $Q^\star_h(s,a)$ as the corresponding state-action value function which is defined similarly. Notice that the value function and state-action value function satisfy the Bellman equations:

$$Q^\star_h(s,a) = r^{\text{lin}}_h(s,a) + \mathbb{E}\left[V^\star_{h+1}(s_{h+1})|s_h = s, a_h = a\right],$$

where $V^\star_h(s) = \max_{a\in A} Q^\star_h(s,a)$ and $V^\star_{h+1}(s) = 0$. Plugging the linearity of reward and transition dynamics in, it is easy to verify that $Q^\star_h(s,a)$ is a linear function of the feature map $\phi(s,a)$. Furthermore, it can be verified that the value function of any policy is a linear function of $\phi(s,a)$ by the Bellman equation for policy evaluation.

5 Action-Centered Value Iteration

We now present a provably efficient exploration algorithm, which rests on our key insight that the optimal policy is not impacted by confounding rewards. Since only the overall reward $r^k_h(s,a)$ can be observed rather than r^{lin}, we seek to estimate r^{lin} via the action-centering trick [Greenewald et al., 2017]: importance sampling between the default action and greedy action (section 5.1). This yields our algorithm, Action-Centered Value Iteration (ACVI), which combines least-squares value iteration, used to learn the linear value function $Q^\star_h(s,a)$, with action-centered reward estimates in Section 5.2.

5.1 Estimating the linear reward component

Given the reward structure, it would be trivial to estimate the linear reward part if we could take two actions at each timestep. We could take the difference between the reward of the default action, which only consists of r^{conf}, and any other action, which gives us a value that is linear in expectation in the feature map:

$$\mathbb{E}_{\varepsilon^k_h}\left[r^k_h(s,a) - r^k_h(s,a_{0,h})\right] = \langle \phi(s,a), \theta_h\rangle \tag{3}$$

Though only one action can be taken at each timestep, this intuition inspires a randomization mechanism to estimate the reward difference. We randomly sample between the default action $a_{0,h}$ and the greedy action at that timestep, denoted as $\bar{a}^k_h$, to determine a^k_h. (The greedy action can be the default action.) Let $p_k = \Pr_k(a_{0,h})$ be the probability of choosing the default action for any h during episode k, which we refer to as the *sampling schedule*. Then, define the estimator

$$\hat{r}^k_h(s,a) = (\mathbb{1}\left\{a = \bar{a}^k_h\right\} - (1-p_k))r^k_h(s,a). \tag{4}$$

Its expectation over the sub-Gaussian noise and the randomization between the default and greedy action is

$$\begin{aligned} &\mathbb{E}_{\varepsilon^k_h, a^k_h}\left[\hat{r}^k_h(s, a^k_h)\right] \\ =&\mathbb{E}_{\varepsilon^k_h}\left[(1-p_k)p_k r^k_h(s,\bar{a}^k_h) - p_k(1-p_k)r^k_h(s,a_{0,h})\right] \\ =&p_k(1-p_k)\langle\phi(s,\bar{a}^k_h), \theta_h\rangle \end{aligned} \tag{5}$$

Therefore, we can learn an estimate $\hat{\theta}^k_h$ based on the previous $k-1$ episodes for each θ_h using $\hat{r}^k_h(s^k_h, a^k_h)$,

Algorithm 1 Action-Centered LSVI (ACVI)

1: **Input:** $K, H, \alpha, \lambda, p_k, \beta_k \; \forall \; k \in [K]$
2: **for** $k = 1, \ldots, K$: **do**
3: Receive s_1^k.
4: **for** $h = H, \ldots, 1$ **do**
5: Estimate $\hat{\theta}_h^k = (B_h^k)^{-1}\hat{b}_h^k$ for $\hat{b}_h^k = \sum_{\tau=1}^{k-1} \hat{r}_h^\tau \phi_h^\tau$, $B_h^k = \alpha I + \sum_{\tau=1}^{k-1} \eta_\tau \phi_h^\tau \phi_h^{\tau\top}$.
6: Set $\Lambda_h^k = \lambda I + \sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top$.
7: Update weight estimate of value function $w_h^k = (\Lambda_h^k)^{-1}\big(\sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau)[\langle \phi(s_h^\tau, a_h^\tau), \hat{\theta}_h^k \rangle \mathbb{1}\{a_h^\tau \neq a_{h,0}\} + \max_a Q_{h+1}^k(s_{h+1}^\tau, a)]\big)$.
8: Set value function $Q_h^k(\cdot,\cdot) = \min\{\langle w_h^k, \phi(\cdot,\cdot)\rangle + \beta_k \sqrt{\phi(\cdot,\cdot)^\top (\Lambda_h^k)^{-1}\phi(\cdot,\cdot)}, H\}$.
9: **end for**
10: **for** $h = 1, \ldots, H$ **do**
11: $a_h^k = \begin{cases} a_{h,0} & \text{w.p. } p_k \\ \bar{a}_h^k = \text{argmax}_a Q_h^k(s_h^k, a) & \text{otherwise} \end{cases}$
12: **end for**
13: **end for**

which is observable. However, we do not want p_k to be constant across all k since taking the default action with a fixed probability as k increases incurs linear regret. This implies that the variance of each $\hat{r}_h^k$ is different across k. We use weighted ridge regression to estimate each θ_h, where the objective function to minimize over θ is:

$$\sum_{\tau=1}^{k-1} \eta_\tau \left(\frac{\hat{r}_h^\tau}{\eta_\tau} - \langle \phi_h^\tau, \theta \rangle \right)^2 + \alpha \|\theta\|_2^2 \tag{6}$$

for some $\alpha \geq 0$, where we've used notation $\eta_\tau = p_\tau(1 - p_\tau), \hat{r}_h^\tau = \hat{r}_h^\tau(s_h^\tau, a_h^\tau)$, $\phi_h^\tau = \phi(s_h^\tau, \bar{a}_h^\tau)$. Solving (6) gives us the closed-form solution $\hat{\theta}_h^k = (B_h^k)^{-1}\hat{b}_h^k$, where

$$B_h^k = \alpha I + \sum_{\tau=1}^{k-1} \eta_\tau \phi_h^k \phi_h^{k\top}, \quad \hat{b}_h^k = \sum_{\tau=1}^{k-1} \hat{r}_h^\tau \phi_h^\tau. \tag{7}$$

5.2 Our Algorithm

Now that we have discussed how the optimal policy is not impacted by the confounding term and how to compute $\hat{\theta}_h^k$, we explain how these ideas are used in ACVI, described in Algorithm 1. First, after each episode, we calculate $\hat{\theta}_h^k$ for each h and then use this to estimate the linear component of the reward as $\hat{r^{\text{lin}}}_h(s,a) := \langle \phi(s,a), \hat{\theta}_h^k \rangle$ in line 5. This reward estimate is then used in the standard LSVI framework [Bradtke and Barto, 1996; Osband et al., 2016], which aims to learn the optimal policy by iteratively fitting a linear function $\langle w_h^\star, \phi(s,a) \rangle$ to the Bellman update of $Q_h^\star(s,a)$; that is, our estimate w_h^k in line 7 is the closed form solution to

$$\text{argmin}_w \sum_{\tau=1}^{k-1} \big(\hat{r^{\text{lin}}}_h(s_h^\tau, a_h^\tau) + \max_{a \in \mathcal{A}} Q_{h+1}^k(s_{h+1}^\tau, a) - \langle w, \phi(s_h^\tau, a_h^\tau) \rangle \big)^2 + \lambda \|w\|^2.$$

After w_h^k is estimated, in line 8 we construct our estimate of the value function, Q_h^k, which contains an additional UCB bonus term to encourage exploration. This UCB term has a β_k constant that is used to capture the uncertainty from both estimating the reward and the value function, such that with high probability $Q_h^k(s,a) \geq Q_h^\star(s,a)$ for all (s,a). We also control the magnitude of our approximation Q_h^k by noting that the action-value function should always be less than H under our assumption $|r_h^{\text{lin}}| \leq 1$. Lastly, we use this updated value function to select a greedy action $\bar{a}_h^k$, and we randomize between this action and the default action according to the sampling schedule p_k in lines 10-11. This action centering is necessary for us to produce accurate estimates of $\hat{\theta}_h^k$.

6 Regret Analysis

We present our main theoretical result, the sublinear expected regret bound for ACVI, and provide a proof sketch. We then compare our results to prior regret bounds.

Theorem 1. *Set $\lambda = 1, \alpha = 1$, $p_k = \frac{1}{(k+1)^{1/4}}$, and $\delta \in [0,1]$. There exists an absolute constant $c_\beta > 0$ such that if $\beta_k = c_\beta d\sqrt{\iota}\left(H + \frac{(r_{\max}+2\sigma^2)\sqrt{d}}{\eta_{k-1}} \mathbb{1}\{k > 1\}\right)$, where $\iota = \log \frac{dT}{\delta}$, then the expected regret of Algorithm 1 is $\mathcal{O}((r_{\max} + 2\sigma^2)d^2 \iota H^{3/2} T^{3/4})$ with probability at least $1 - \delta$.*

This theorem shows we can achieve sublinear regret by carefully setting a sampling schedule p_k and particular UCB bonus structure β_k. This bound scales with the number of steps, the horizon, and the dimension of the feature map, which are all standard parameters in linear MDP's regret analysis, and is independent of confounding reward structure beyond magnitude.

6.1 Proof Sketch

The proof of theorem 1 involves three main parts: 1) bounding the error of the estimated linear reward component, 2) bounding the error of the value function approximator, and 3) setting a sampling schedule that balances exploration, exploitation, and learning the linear reward. We discuss our results for each and how they impact the cumulative regret bound.

Bounding the linear reward estimation error. It is clear that the ability to estimate θ_h accurately will affect the regret of the algorithm since this impacts the approximation of the value function. We apply the result of Theorem 1 of Abbasi-Yadkori et al. [2011] for self-normalized processes to produce a concentration inequality on $\hat{b}_h^k$ first and then propagate it to produce a bound on the accuracy of the linear reward estimate.

Lemma 1. *Fix $\delta' > 0$ and pick any s, a, h, k. For $\hat{\theta}_h^k$ defined in Algorithm 1 and $k > 1$, with probability at least $1 - \frac{\delta'}{T}$,*

$$|\langle \phi(s,a), \theta_h - \hat{\theta}_h^k \rangle| \leq \sqrt{d\phi(s,a)^\top (B_h^k)^{-1} \phi(s,a)} \quad (8)$$
$$\times \Big(\frac{r_{\max} + 2\sigma^2}{2\sqrt{\eta_{k-1}}} \cdot \sqrt{\log \frac{Tk}{\delta'}} + 1 \Big).$$

When $k = 1$, $|\langle \phi(s,a), \theta_h - \hat{\theta}_h^k \rangle|$ is trivially at most $\sqrt{d}$, since $\hat{\theta}_h^1 = 0$.

As k increases, the term $\sqrt{d\phi(s,a)^\top (B_h^k)^{-1} \phi(s,a)}$ decreases. However, note that $\frac{1}{\sqrt{\eta_{k-1}}} = \frac{1}{\sqrt{p_{k-1}(1-p_{k-1})}} \approx \frac{1}{\sqrt{p_k}}$, showing how p_k significantly impacts the estimation error.

Bounding the value function estimation error. The next challenge is to understand how this reward estimation error impacts exploration. To determine the UCB bonus, we consider the difference between the true value function and our linear approximation: the gap between $\langle w_h^k, \phi(s,a) \rangle$ and $Q_h^\star(s,a)$. Recall that in the linear MDP setting, the value function regardless of policy is linear. In particular, $Q_h^\pi(s,a) = \langle w_h^\pi, \phi(s,a) \rangle$, where $w_h^\pi := \theta_h + \int V_{h+1}^\pi(s) d\mu_h(s)$ for any policy π. If the linear reward were observable, both w_h^π and w_h^k would be defined in terms of θ_h; however, our analysis must account for fact that w_h^k is instead constructed using $\hat{\theta}_h^k$ estimated via action-centering. We define an upper confidence bonus that ensures optimism by incorporating both the reward estimation error of Lemma 1 and the standard value function estimation error (under no confounding).

Lemma 2. *For any s, a, h, k, with probability at least $1 - \delta$ there exists a constant c_β such that for $\beta_k = c_\beta d \sqrt{\iota} \Big(H\sqrt{d} + \frac{(r_{\max} + 2\sigma^2)\sqrt{d}}{\eta_{k-1}} \Big)$, $\iota = \log \frac{dT}{\delta}$, and any fixed policy π_k, we have that*

$$\langle w_h^k, \phi(s,a) \rangle - Q_h^{\pi_k}(s,a) \leq$$
$$\mathbb{E}\left[V_{h+1}^k(s_{h+1}^k) - V_{h+1}^{\pi_k}(s_{h+1}^k) | s_h^k = s, a_h^k = a \right]$$
$$+ \beta_k \sqrt{\phi(s,a)^T (\Lambda_h^k)^{-1} \phi(s,a)},$$

where $V_h^k(s) = \max_{a \in \mathcal{A}} Q_h^k(s,a)$.

Notice that β_k has two parts, and the second part depends on our reward estimation error. This is because there is an additional term in the decomposition of $\langle w_h^k, \phi(s,a) \rangle - Q_h^{\pi_k}(s,a)$, namely $\phi(s,a)^\top (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau) \langle \phi(s_h^\tau, a_h^\tau), \hat{\theta}_h^k - \theta_h \rangle$, which captures the role of using $\hat{\theta}_h^k$ instead of θ_h in w_h^k and explains how reward error is propagated.

As illustrated in line 8 of Algorithm 1, this bound is used to construct Q_h^k. Via an induction argument on this bound, we can show that $Q_h^*(s,a) \leq Q_h^k(s,a)$ for all s, a with probability at least $1 - \delta$.

Minimizing regret with choice of p_k. We have shown so far that the estimation error of r^{lin} and consequently the UCB bonus both depend on p_k. We now sketch how we derive our regret bound, which captures a tradeoff that the sampling schedule presents.

We first discuss the challenge in choosing p_k. First, if p_k is very small, the default action is sampled infrequently, and the estimation error of the linear reward is high. This results in a large UCB bonus that dominates the linear approximation $\langle w_h^k, \phi(s,a) \rangle$, and intuitively our agent may explore too much. However, sampling the default action less results in the agent being able to follow the learned policy Q_h^k more often. On the other hand, when p_k is large or clipped to some range in $(0, 1)$ as in Greenewald et al. [2017], action-centering yields a better estimate of the reward that also leads to a better linear approximation of the value function intuitively. However, having a constant clipped probability of taking a default action across all k will certainly result in linear regret.

Our approach to bounding the regret is as follows. By definition of the value-action function, our expected regret over the action-centering is bounded by

$$\begin{aligned} \mathbb{E}\left[\text{Regret}(K)\right] &= \sum_{k=1}^{K} \mathbb{E}\left[Q_1^\star(s_1^k, \widetilde{a}_1^k) - Q_1^{\pi_k}(s_1^k, a_1^k) \right] \\ &\leq \sum_{k=1}^{K} \mathbb{E}\left[Q_1^k(s_1^k, \bar{a}_1^k) - Q_1^{\pi_k}(s_1^k, a_1^k) \right]. \end{aligned}$$

The first line follows from the definition of the state-action value function, where $\widetilde{a}_h^k = \text{argmax}_{a \in \mathcal{A}} Q_h^\star(s_h^k, a)$. The second line uses the fact that $Q_1^\star(s_1^k, \widetilde{a}_1^k) \leq Q_1^k(s_1^k, \widetilde{a}_1^k)$ w.p $\leq 1 - \delta$ (UCB), and then the fact that the the greedy action $\bar{a}_1^k$ maximizes $Q_1^k(s_1^k, \cdot)$. This expression, however, presents a challenge since a_1^k and $\bar{a}_1^k$ are not always the same.

We address this by considering two cases for each episode depending on if the agent selects any default action throughout the kth episode. For each k, define the event E_k that $a_h^k = a_{0,h}$ for at least one $h \in [H]$.

Then our regret bound becomes

$$\sum_{k=1}^{K} \mathbb{E}\left[Q_1^k(s_1^k, \bar{a}_1^k) - Q_1^{\pi_k}(s_1^k, a_1^k) | E_k\right] \Pr(E_k) \quad (9)$$

$$+ \sum_{k=1}^{K} \mathbb{E}\left[Q_1^k(s_1^k, \bar{a}_1^k) - Q_1^{\pi_k}(s_1^k, a_1^k) | E_k^C\right] \Pr(E_k^C) \quad (10)$$

In the first line, $\Pr(E_k) = 1 - (1-p_k)^H$, and expected regret of the episode conditioned on at least one default action being taken is trivially bounded by $2H$. Then regret due to choosing default actions in (9) is bounded by

$$2H \sum_{k=1}^{K} (1 - (1-p_k)^H) \leq 2H^2 \sum_{k=1}^{K} p_k \quad (11)$$

after using the fact that $(1-p_k)^H \geq 1 - Hp_k$. In the second line, with probability $\Pr(E_k^C) = (1-p_k)^H$, the agent will follow the greedy trajectory throughout the kth episode, in which case our expected regret per episode, $\mathbb{E}\left[Q_1^k(s_1^k, \bar{a}_1^k) - Q_1^{\pi_k}(s_1^k, \bar{a}_1^k) | E_k^C\right]$, is now defined on the same state-action inputs. For this case, we write Lemma 2 recursively to decompose the regret as a summation over a martingale difference sequence for the value function estimate and a summation of the UCB bonuses accumulated across all h and k. The former can be bounded using standard concentration inequalities to get a regret less than $c_1 H\sqrt{T\iota}$ for some constant c_1 independent of p_k. The summation over UCB bonuses is of the form $c_2 \iota H d^{3/2}\Big(H\sqrt{K} + (r_{\max} + 2\sigma^2)\Big(\sum_{k=1}^{K} \frac{d}{p_k^2(1-p_k)^2}\Big)^{1/2}\Big)$ for some independent constant c_2. Therefore, our regret bound in terms of the sampling schedule can be written as

$$\mathbb{E}\left[\text{Regret}(K)\right] \leq c_1 H\sqrt{T\iota} + c_2 \iota H^2 d^{3/2}\sqrt{K}$$
$$+ c_2 \iota H d^{3/2}(r_{\max} + 2\sigma^2)\Big(\sum_{k=1}^{K} \frac{d}{p_k^2(1-p_k)^2}\Big)^{1/2}$$
$$+ 2H^2 \sum_{k=1}^{K} p_k$$

This expression formalizes the tradeoff previously discussed: large p_k incurs regret from taking the default action too frequently, while small p_k makes the UCB bonus large. Taking p_k to be of order $\frac{1}{k^{1/4}}$ minimizes the above expression and yields a regret of order $T^{3/4}$.

6.2 Comparison to other approaches and bounds

We now discuss our analysis and bounds with respect to two algorithms from Jin et al. [2020b] that provide guarantees under the linear MDP and "ζ-approximate" linear MDP.

Algorithm 1 of Jin et al. [2020b] (LSVI-UCB) assumes that true reward is exactly r^{lin}, and under such a setting their algorithm attains regret $\mathcal{O}(\sqrt{d^3H^3T\iota^2})$ according to their Theorem 3.1. They are able to directly use the reward in estimating w_h^k, and there is no need to do action-centering or learning of θ_h. As a result, their UCB bonus β is of order $dH\sqrt{\iota}$ whereas our β_k requires an extra term of order $\frac{(r_{\max}+2\sigma^2)d^{3/2}\sqrt{\iota}}{\eta_{k-1}}$ to factor in the estimation error for the linear reward. Therefore, the additional action-centering and learning of the linear component of the reward in our nonlinear setting increases regret by $T^{1/4}$ but is still sublinear.

Theorem 3.2 of Jin et al. [2020b] considers a misspecified setting that encapsulates our confounding reward assumptions: they assume that 1) rewards are "close" to linear, i.e. $|r_h(s,a) - \langle \phi(s,a), \theta_h \rangle| \leq \zeta$ and 2) transition dynamics are approximately linear, i.e. $\|\Pr_h(\cdot|s,a) - \langle \phi(s,a), \mu_h(\cdot)\rangle\|_{\text{TV}} \leq \zeta$. Under this ζ-approximate linear MDP, they use knowledge of the magnitude of misspecification ζ to construct a UCB bonus with β_k of order $H(d\sqrt{\iota} + \zeta\sqrt{kd})$. Using this β_k in their Algorithm 1, which we refer to here as LSVI-M, rather than β defined previously, yields regret $\mathcal{O}(\sqrt{d^3H^3T\iota^2} + \zeta dHT\sqrt{\iota})$. We note that keeping the linear dynamics assumption does not change their linear regret bound. We observe that their additional term in β_k, $\zeta\sqrt{kd}$, is a factor of $k^{1/4}$ larger than ours. This suggests that only using knowledge of the magnitude of misspecification (i.e. $|f_h^k|$) without learning the linear reward requires more optimism, which explains the additional linear regret term in their case.

7 Experiments

We evaluate our algorithm on synthetic experiments for the linear MDP framework confounded by nonlinear baseline rewards. We find that ACVI achieves lower average regret than alternatives that do not exploit confounding structure.

Experimental setup. We construct a linear MDP and then add confounding rewards to it. For the linear MDP, we follow an example from Jin et al. [2020b], where each state $s \in \mathbb{R}^{d'}$ satisfies $\|s\|_1 = 1$ with nonnegative entries. We construct a feature map $\phi(s,a) = s \cdot [\mathbb{1}\{a=0\} \ldots \mathbb{1}\{a=N\}]^\top$ by slotting the vector s into the a-th position of a zero vector of length $d = (N+1)d'$. We further require that each $e_i^\top \mu_h$ is a measure and $\|\theta_h\|_1 = 1$. We set $|\mathcal{S}| = 100$, $d' = 8$, $N = 5$ actions, $H = 3$, and $K = 100000$. The con-

founding reward has an episode-dependent structure:

$$f_h^k(\cdot) = \frac{1}{10}(z_h^\top s_1^k + s_1^k[h])(-1)^h \tag{12}$$

$$+ \frac{1}{10c}\Big(k - \Big(\Big\lfloor \frac{k}{c} - \frac{1}{2} \Big\rceil c + \frac{c}{2}\Big)\Big), \tag{13}$$

where $z_h \sim \text{Unif}(-1,1)^{d'}$, $s_1^k[h]$ represents the hth-indexed element of s_1^k, $c = 50000$ and $\lfloor \cdot \rceil$ represents rounding. This confounding reward structure has a component based on initial state and h, and a component that oscillates between -0.05 and 0.05 every 50000 episodes, which roughly models periodic external influences. This perturbs the linear reward significantly in magnitude and also makes fitting a linear approximation to the value function difficult.

Evaluation. We compare the average regret up to the current episode k, Regret(k), of ACVI with the two algorithms discussed in Section 6.2, restated below:

1. LSVI-UCB [Jin et al., 2020b]: An algorithm that ensures provably efficient exploration in linear MDPs without confounding rewards. It attains sublinear regret when transition dynamics and rewards are exactly linear. It explores with a UCB reward bonus of the form $\beta = cdH\sqrt{\iota}$.

2. LSVI-M [Jin et al., 2020b]: A variant of LSVI-UCB with provable guarantees when transition dynamics and rewards are ζ-approximately linear. It yields regret linear in the magnitude of misspecification ζ. The UCB reward bonus takes the form $\beta_k = cH(d\sqrt{\iota} + \zeta\sqrt{kd})$ on episode k.

To determine the UCB constant $c > 0$ in each case, we tune the minimum value of c such that $Q_h^*(s,a) \leq Q_h^k(s,a)$ for all s, a, h, k.

Results. The regret curves for each approach are in Figure 1.

The average regret of ACVI consistently decreases, without being impacted by the periodic confounding reward, and eventually ACVI achieves 25.1% lower average regret than LSVI-M and 13.3% lower than LSVI-UCB. Initially, ACVI incurs higher average regret than LSVI-UCB and LSVI-M from frequently taking the default action to estimate the linear reward component. However, for LSVI-M, the average regret increases after a while. We hypothesize that the value function approximation step attempts to fit a w_h^k according to the initial periodic increase in f_h^k, however, this linear approximation is worsened as f_h^k oscillates, changing the policy. For LSVI-UCB, the UCB constant was tuned to be very high to counter the fact that the UCB bonus does not account for misspecification. As a result, the algorithm focuses on exploring and does not obtain a regret curve as low as ACVI.

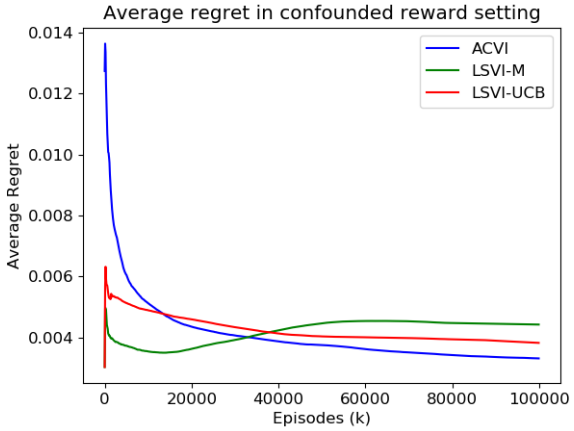

Figure 1: Average regret for ACVI, LSVI-M, and LSVI-UCB. ACVI initially incurs higher regret from action-centering, but achieves the lowest final regret.

We evaluate on other confounding rewards $f_h^k(\cdot)$ in the Appendix to show that confounding structure does not significantly impact our regret. We also empirically test different values of the sampling schedule p_k and find that our choice of order $\frac{1}{k^{1/4}}$ performs best.

8 Discussion

We present a provably efficient exploration algorithm for an episodic RL setting with linear dynamics and a nonlinear reward function, composed of a linear component and a nonlinear confounding component. We leverage the fact that the optimal policy only depends on the linear reward component and estimate this component via importance sampling. This enables our algorithm, ACVI, to achieve sublinear regret with high probability, and our synthetic experiments support our theoretical findings.

Our work opens several directions for future work. First, the action centering approach could be applicable to more general reward structures consisting of some complex latent reward and confounding reward; the action-centering would simply be done along with a more complex regression procedure. Second, the default action assumption could possibly be relaxed via methods similar to Krishnamurthy et al. [2018]. Lastly, misspecification in both rewards and dynamics contribute the same order terms in the linear MDP regret bound [Jin et al., 2020b]. It is interesting to consider perturbed nonlinear dynamics models analogous to the nonlinear confounded rewards we consider.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 24*, pages 2312–2320. Curran Associates, Inc.

Baird, L. (1995). Residual algorithms: Reinforcement learning with function approximation. In *Machine Learning Proceedings 1995*, pages 30–37. Elsevier.

Bradtke, S. J. and Barto, A. G. (1996). Linear least-squares algorithms for temporal difference learning. *Mach. Learn.*, 22(1–3):33–57.

Chitnis, R. and Lozano-Pérez, T. (2020). Learning compact models for planning with exogenous processes. volume 100 of *Proceedings of Machine Learning Research*, pages 813–822. PMLR.

Dietterich, T., Trimponias, G., and Chen, Z. (2018). Discovering and removing exogenous state variables and rewards for reinforcement learning. volume 80 of *Proceedings of Machine Learning Research*, pages 1262–1270, Stockholmsmässan, Stockholm Sweden. PMLR.

Greenewald, K., Tewari, A., Murphy, S., and Klasnja, P. (2017). Action centered contextual bandits. In *Advances in neural information processing systems*, pages 5977–5985.

Gu, S., Holly, E., Lillicrap, T., and Levine, S. (2017). Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 3389–3396. IEEE.

Jin, C., Jin, T., Luo, H., Sra, S., and Yu, T. (2020a). Learning adversarial markov decision processes with bandit feedback and unknown transition. *International Conference on Machine Learning (ICML)*.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. (2020b). Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143.

Krishnamurthy, A., Wu, Z. S., and Syrgkanis, V. (2018). Semiparametric contextual bandits. In *International Conference on Machine Learning*, pages 2776–2785.

Lattimore, T. and Szepesvari, C. (2020). Learning with good feature representations in bandits and in rl with a generative model. *International Conference on Machine Learning (ICML)*.

Lykouris, T., Simchowitz, M., Slivkins, A., and Sun, W. (2019). Corruption robust exploration in episodic reinforcement learning. *arXiv preprint arXiv:1911.08689*.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533.

Osband, I., Roy, B. V., and Wen, Z. (2016). Generalization and exploration via randomized value functions.

Wang, R., Salakhutdinov, R., and Yang, L. F. (2020). Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *Advances in neural information processing systems*.

Wang, Y., Wang, R., Du, S. S., and Krishnamurthy, A. (2019). Optimism in reinforcement learning with generalized linear function approximation. *arXiv preprint arXiv:1912.04136*.

Yang, L. F. and Wang, M. (2020). Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. *International Conference on Machine Learning (ICML)*.

Zanette, A., Brandfonbrener, D., Brunskill, E., Pirotta, M., and Lazaric, A. (2020a). Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pages 1954–1964.

Zanette, A., Lazaric, A., Kochenderfer, M., and Brunskill, E. (2020b). Learning near optimal policies with low inherent bellman error. *International Conference on Machine Learning (ICML)*.

Zanette, A., Lazaric, A., Kochenderfer, M. J., and Brunskill, E. (2020c). Provably efficient reward-agnostic navigation with linear value iteration. *Advances in neural information processing systems*.

Supplementary Materials

A Glossary

The glossary is given in Table 1 below.

Symbol	Used for
K	Number of episodes.
H	Length of the horizon.
T	Total number of steps HK.
$\mathcal{S}$	State space
$\mathcal{A}$	Action space $\mathcal{A} = \{0, \dots N\}$ with N actions.
Pr	Transition dynamics $\{\Pr_h\}_{h=1}^H$, defined as $\Pr(\cdot\|s,a)$ for each s, a.
r^k	Reward function $\{r_h^k\}_{h=1}^H$ defined as $r_h^k(s,a)$ for each s, a.
s_h^k	State the agent is at during the hth timestep of the kth episode.
a_h^k	Action the agent takes during the hth timestep of the kth episode.
$\mathcal{H}_k$	The history $\{s_h^k, a_h^k\}_{h=1}^H$ of states visited and actions taken in the kth episode.
$\bar{\mathcal{H}}^{k-1}$	Total history $\mathcal{H}_1, \dots \mathcal{H}_{k-1}$ so far at the kth episode.
π_k	A policy $\pi_k : \mathcal{S} \times [H] \to \mathcal{A}$ to learn in episode k.
$V_{h,r^k}^{\pi_k}$	Policy value function for policy π_k at episode k, starting timestep h.
d	Dimension of feature map ϕ.
$\phi(s,a)$	Feature map $\phi : \mathcal{S} \times A \to \mathbb{R}^d$.
$\mu_h(\cdot)$	Measure $\mu_h(\cdot) \in \mathbb{R}^d$ for linear dynamics.
r_h^{lin}	Linear component of the reward, which is constant across episodes.
$r_{h,k}^{\text{conf}}$	Nonlinear confounding component of the reward, which can vary across episodes.
$a_{0,h}$	Default action. When taken, the agent receives only the confounding reward $r_{h,k}^{\text{conf}}$ plus noise.
θ_h	Unknown weight vector in $\mathbb{R}^d$ for r_h^{lin}.
$f_h^k(\cdot)$	Confounding reward function, with input s_1^k and $\bar{H}^{k-1}$.
ε_h^k	Random σ^2 sub-Gaussian noise.
$r_{\max}$	Upper bound on $\lvert r_h^{\text{lin}} + r_{h,k}^{\text{conf}}\rvert$.
$\pi^\star$	Optimal policy for the agent, which is equal to $\pi_k^\star$, the optimal policy at each episode.
$\mathcal{M}$	A linear MDP $\langle \mathcal{S}, \mathcal{A}, H, \Pr, r^{\text{lin}} \rangle$ with linear rewards and dynamics.
$V_h^\pi(s)$	Policy value function for policy π of linear MDP $\mathcal{M}$, starting from timestep h.
$V_h^\star(s), Q_h^\star(s,a)$	Optimal value function and action-value function for $\mathcal{M}$ corresponding to $\pi^\star$.
p_k	The sampling schedule, i.e. $\Pr(a_h^k = a_{0,h})$ for any h.
$\bar{a}_h^k$	The algorithm's greedy action at timestep h episode k.
$\hat{r}_h^k(s,a)$	The reward estimator $(\mathbb{1}\{a = \bar{a}_h^k\} - (1-p_k))r_h^k(s,a)$. $\hat{r}_h^k$ is equal to $\hat{r}_h^k(s_h^k, a_h^k)$.
η_k	Equal to $p_k(1-p_k)$ for the kth episode.
ϕ_h^k	Equal to $\phi(s_h^k, \bar{a}_h^k)$.
$w_h^\star$	Weight vector that satisfies $Q_h^\star(s,a) = \langle \phi(s,a), w_h^\star \rangle$.
w_h^k	Weight vector estimate at episode k, computed using LSVI.
β_k	Parameter for the UCB bonus in constructing Q_h^k at the kth episode.
Q_h^k	Function used in kth episode to choose greedy action at timestep h. $V_h^k(s) = \max_a Q_h^k(s,a)$.
ι	Equal to $\log \frac{dT}{\delta}$, such that results hold w.p. at least $1-\delta$ for $\delta \in (0,1)$.
E_k	Event that $a_h^k = a_{0,h}$ for at least one h during episode k.
ζ	Magnitude of misspecification, i.e. $\lvert r_h(s,a) - \langle \phi(s,a), \theta_h \rangle\rvert \le \zeta$.

Table 1: Glossary of variables and symbols used in this paper.

B Assumptions and properties of the linear value function

We review some assumptions on the linear MDP with confounding rewards. Since optimal policy is not impacted by confounding, we examine the value function for the standard linear MDP, and show that it is also linear.

Assumption 1. *Define a feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$. For each timestep h and any s, a, we have a confounded reward at episode k:*

$$r_h^k(s,a) = \langle \phi(s,a), \theta_h \rangle \mathbb{1}\{a \neq a_{h,0}\} + f_h^k\left(s_1^k; \bar{\mathcal{H}}^{k-1}\right) + \varepsilon_h^k, \tag{14}$$

Furthermore, the linear transition dynamics are

$$\Pr_h(\cdot \mid s,a) = \phi(s,a)^\top \mu_h(\cdot),$$

where $\mu_h(\cdot) \in \mathbb{R}^d$ is some measure measure.

For all s, a, h, k, assume that $\mu_h(\mathcal{S}), \|\theta_h\| \leq \sqrt{d}, \|\phi(s,a)\| \leq 1$, $|\mathbb{E}_{\varepsilon_h^k}[r_h^k(s,a)]| \leq r_{\max}$, $\langle \phi(s,a), \theta_h \rangle \leq 1$ and ε_h^k is σ^2 sub-Gaussian.

Next, we restate the Bellman optimality equations for the value function and action-value function for an arbitrary policy π and the optimal policy. For any s, a, h:

$$Q_h^\pi(s,a) = \langle \phi(s,a), \theta_h \rangle + \mathbb{E}\left[V_{h+1}^\pi(s_{h+1}) | s_h = s, a_h = a\right] \quad V_h^\pi(s) = Q_h^\pi(s, \pi_k(s)) \quad V_{H+1}^{\pi_k}(s) = 0 \tag{15}$$

$$Q_h^\star(s,a) = \langle \phi(s,a), \theta_h \rangle + \mathbb{E}\left[V_{h+1}^\star(s_{h+1}) | s_h = s, a_h = a)\right] \quad V_h^\star(s) = \max_{a \in \mathcal{A}} Q_h^\star(s,a) \quad V_{H+1}^\star(s) = 0 \tag{16}$$

(Note that we drop the episode-dependency for s_{h+1}, s_h, a_h, since the linear dynamics are constant across episodes). Recall that π_k is the policy that the agent takes at the kth episode. Furthermore, let $\max_{s,a,h,\pi}\{|Q_h^\pi(s,a)|, |V_h^\pi(s,a)|\} \leq V_{\max}$. Using our assumptions, $V_{\max} = H$, but we use $V_{\max}$ throughout our results to understand how regret depends on the magnitude of reward.

Finally, recall the parameters for which our regret bound holds:

$$p_k = \frac{1}{(k+1)^{1/4}} \tag{17}$$

$$\beta_k = c_\beta d\iota \left(V_{\max} + \frac{(r_{\max} + 2\sigma^2)\sqrt{d}}{\eta_{k-1}} \mathbb{1}\{k > 1\} \right) \tag{18}$$

Lemma 3. *Define*

$$w_h^{\pi_k} = \theta_h + \int V_{h+1}^{\pi_k}(s') d\mu_h(s'), \qquad w_h^\star = \theta_h + \int V_{h+1}^\star(s') d\mu_h(s') \tag{19}$$

Then

$$Q_h^{\pi_k}(s,a) = \langle \phi(s,a), w_h^{\pi_k} \rangle, \qquad Q_h^\star(s,a) = \langle \phi(s,a), w_h^\star \rangle \tag{20}$$

Proof. This is equivalent to Proposition 2.3 of Jin et al. [2020b]. By definition of $w_h^{\pi_k}$, we have

$$\langle \phi(s,a), w_h^{\pi_k} \rangle = \langle \phi(s,a), \theta_h \rangle + \int V_{h+1}^{\pi_k}(s') \langle \phi(s,a), d\mu_h(s') \rangle \tag{21}$$

$$= \langle \phi(s,a), \theta_h \rangle + \mathbb{E}_{s' \sim \mathbb{P}_h(\cdot|s,a)}\left[V_{h+1}^{\pi_k}(s')\right] \tag{22}$$

due to the linear transition model. The same approach holds for $w_h^\star$. □

Lemma 4. *For any h, k*

$$\|w_h^\star\| \leq 2V_{max}\sqrt{d} \tag{23}$$
$$\|w_h^{\pi_k}\| \leq 2V_{max}\sqrt{d} \tag{24}$$

Proof. We have that $w_h^{\pi_k} = \theta_h + \int V_{h+1}^{\pi_k}(x')d\mu_h(x')$. $\| \int V_{h+1}^{\pi_k}(x')d\mu_h(x')\| \leq V_{max}\sqrt{d}$ and $\|\theta_h\| \leq \sqrt{d}$. The same approach holds for $w_h^\star$. □

C Proof of Lemma 1

The goal of this section is to bound $|\langle \phi(s,a), \hat{\theta}_h^k \rangle - \langle \phi(s,a), \theta_h \rangle|$ for any s, h, a, k with high probability.

C.1 Setup for estimator

Recall our reward estimator:

$$\hat{r}_h^k(s,a) = (\mathbb{1}\left\{a = \bar{a}_h^k\right\} - (1-p_k))r_h^k(s,a). \tag{25}$$

For $\tau = 1, \ldots, k-1$ and a fixed h (which we omit in indexing), define the following for convenience:

$$\phi_\tau = \phi(s_h^\tau, \bar{a}_h^\tau) \tag{26}$$
$$\eta_\tau = p_\tau(1-p_\tau) \tag{27}$$
$$y_\tau = \frac{\hat{r}_h^\tau(s_h^\tau, a_h^\tau)}{\eta_\tau} \tag{28}$$
$$\Sigma_k = \sum_{\tau=1}^{k-1} \eta_\tau \phi_\tau \phi_\tau^T \tag{29}$$

Using weighted ridge regression with penalty $\lambda = 1$, our estimator $\hat{\theta}_h^k$ minimizes $\sum_{\tau=1}^{k-1} \eta_\tau (y_\tau - \theta^\top \phi_\tau)^2 + \|\theta\|_2^2$. This gives us the closed form:

$$\hat{\theta}_h^k = (I + \Sigma_k)^{-1} \sum_{\tau=1}^{k-1} \eta_\tau y_\tau \phi_\tau = (B_h^k)^{-1} \hat{b}_h^k. \tag{30}$$

$$B_h^k = I + \sum_{\tau=1}^{k-1} p_\tau(1-p_\tau)\phi(s_h^\tau, \bar{a}_h^\tau)\phi(s_h^\tau, \bar{a}_h^\tau)^\top = I + \Sigma_k \tag{31}$$

$$\hat{b}_h^k = \sum_{\tau=1}^{k-1} \hat{r}_h^\tau(s_h^\tau, a_h^\tau)\phi(s_h^\tau, \bar{a}_h^\tau) = \sum_{\tau=1}^{k-1} \eta_\tau y_\tau \phi_\tau \tag{32}$$

To write θ_h in this form, we need to define a filtration $\mathcal{F}_k = \{\bar{a}_h^{k+1}, s_h^{\tau+1}, a_h^\tau \; \forall \; \tau \leq k\}$ (e.g. all past actions, all states up to the current state, and what the greedy action at the current state is). It is true that

$$\mathbb{E}\left[y_\tau | \mathcal{F}_{\tau-1}\right] = \langle \phi_\tau, \theta_h \rangle \tag{33}$$

Then, θ_h must satisfy

$$\theta_h = \text{argmin}_\theta \sum_{\tau=1}^{k-1} \eta_\tau (\mathbb{E}\left[y_\tau | \mathcal{F}_{\tau=1}\right] - \langle \phi_\tau, \theta \rangle)^2, \tag{34}$$

which is equivalent to

$$\Sigma_k \theta_h = \sum_{\tau=1}^{k-1} \eta_\tau \mathbb{E}\left[y_\tau | \mathcal{F}_{\tau-1}\right] \phi_\tau. \tag{35}$$

For ease of notation, let $\bar{b}_h^k = \sum_{\tau=1}^{k-1} \eta_\tau \mathbb{E}\left[y_\tau | \mathcal{F}_{\tau-1}\right] \phi_\tau$.

C.2 Computing the reward bound

We define the following:

$$m_\tau = \sqrt{\eta_\tau}\phi_\tau \tag{36}$$

$$\alpha_\tau = \frac{\hat{r}_h^\tau(s_h^\tau, a_h^\tau)}{\sqrt{\eta_\tau}} - \sqrt{\eta_\tau}\phi_\tau^T \theta_h \tag{37}$$

Then m_τ is a $\mathbb{R}^d$-valued stochastic process such that m_τ is $\mathcal{F}_{\tau-1}$-measurable, and α_τ is a martingale difference process that is $\mathcal{F}_\tau$-measurable (it is simple to verify that $\mathbb{E}\left[\alpha_\tau | \mathcal{F}_{\tau-1}\right] = 0$). Next, we expand $\alpha_\tau | \mathcal{F}_{\tau-1}$ and combine constant terms to get

$$\frac{1}{\sqrt{\eta_\tau}}\left(\mathbb{1}\left\{a_h^\tau = \bar{a}_h^\tau\right\} - (1-p_\tau)\right)\left(\langle \phi(s_h^\tau, a_h^\tau), \theta_h\rangle \mathbb{1}\left\{a_h^\tau = \bar{a}_h^\tau\right\} + f_h^\tau(s_1^\tau; \bar{\mathcal{H}}^{\tau-1}) + \varepsilon_h^\tau\right)\right) + \text{constant} \tag{38}$$

$$= \frac{1}{\sqrt{\eta_\tau}}\left(\mathbb{1}\left\{a_h^\tau = \bar{a}_h^\tau\right\}\left(p_\tau \langle \phi(s_h^\tau, a_h^\tau), \theta_h\rangle + f_h^\tau(s_1^\tau; \bar{\mathcal{H}}^{\tau-1})\right) + \varepsilon_h^\tau(\mathbb{1}\left\{a_h^\tau = \bar{a}_h^\tau\right\} - (1-p_\tau))\right) + \text{constant} \tag{39}$$

We use the fact that Bernoulli random variables (i.e. the indicator variable above) are $\frac{1}{4}$ sub-Gaussian and ε_h^k is σ^2 sub-Gaussian. Then, $\alpha_\tau | \mathcal{F}_{\tau-1}$ is $\frac{1}{\sqrt{\eta_\tau}}\left(\frac{r_{\max}}{4} + \sigma^2\right) \leq \frac{r_{\max}+2\sigma^2}{2\sqrt{\eta_\tau}}$ sub-Gaussian. Next, define ξ_k as

$$\xi_k = \sum_{\tau=1}^{k-1} m_\tau \alpha_\tau = \sum_{\tau=1}^{k-1} \sqrt{\eta_\tau}\phi_\tau \left(\frac{\hat{r}_h^\tau(s_h^\tau, a_h^\tau)}{\sqrt{\eta_\tau}} - \sqrt{\eta_\tau}\phi_\tau^\top \theta_h\right) \tag{40}$$

$$= \sum_{\tau=1}^{k-1} \hat{r}_h^\tau(s_h^\tau, a_h^\tau)\phi_\tau - \eta_\tau \phi_\tau \phi_\tau^\top \theta_h = \sum_{\tau=1}^{k-1} \eta_\tau y_\tau \phi_\tau - \eta_\tau \mathbb{E}\left[y_\tau | \mathcal{F}_{\tau-1}\right]\phi_\tau \tag{41}$$

$$= \hat{b}_h^k - \bar{b}_h^k \tag{42}$$

We can express $\hat{\theta}_h^k - \theta_h$ in terms of B_h^k and ξ_k.

Lemma 5. *For any h, k,*

$$\hat{\theta}_h^k - \theta_h = (B_h^k)^{-1}(\xi_k - \theta_h). \tag{43}$$

Proof. We can write $\bar{b}_h^k$ as $\Sigma_k \theta_h = (B_h^k - I)\theta_h$. Then $(B_h^k)^{-1}(\xi_k - \theta_h)$ is equivalent to

$$(B_h^k)^{-1}(\hat{b}_h^k - \bar{b}_h^k - \theta_h) = (B_h^k)^{-1}(\hat{b}_h^k - B_h^k\theta_h + \theta_h - \theta_h) \tag{44}$$

$$= (B_h^k)^{-1}\hat{b}_h^k - \theta_h = \hat{\theta}_h^k - \theta_h. \tag{45}$$

□

Then, our desired expression $|\langle \phi(s,a), \theta_h - \hat{\theta}_h^k\rangle|$ (when $k > 1$) can be written as follows:

$$|\langle \phi(s,a), \theta_h - \hat{\theta}_h^k\rangle| = |\langle \phi(s,a), (B_h^k)^{-1}(\xi_k - \theta_h)\rangle| \tag{46}$$

$$= |\langle (B_h^k)^{-1/2}\phi(s,a), (B_h^k)^{-1/2}(\xi_k - \theta_h)\rangle| \tag{47}$$

$$\leq \|(B_h^k)^{-1/2}\phi(s,a)\| \|(B_h^k)^{-1/2}(\xi_k - \theta_h)\| \tag{48}$$

$$= \sqrt{\phi(s,a)^\top (B_h^k)^{-1}\phi(s,a)} \cdot \|\xi_k - \theta_h\|_{(B_h^k)^{-1}} \tag{49}$$

We can bound $\|\xi_k - \theta_h\|_{(B_h^k)^{-1}} \leq \|\xi_k\|_{(B_h^k)^{-1}} + \|\theta_h\|_{(B_h^k)^{-1}}$, and $\|\theta_h\|_{(B_h^k)^{-1}}$ is less than $\|(B_h^k)^{-1/2}\theta_h\|_2 \leq \sqrt{d}$, since $\|(B_h^k)^{-1/2}\|_2 \leq 1$. Therefore,

$$|\langle \phi(s,a), \theta_h - \hat{\theta}_h^k \rangle| \leq \sqrt{\phi(s,a)^\top (B_h^k)^{-1} \phi(s,a)} \cdot \left(\|\xi_k\|_{(B_h^k)^{-1}} + \sqrt{d} \right) \tag{50}$$

We use now use Theorem 1 of Abbasi-Yadkori et al. [2011]. We have that with probability $1 - \delta'$,

$$\|\xi_k\|_{(B_h^k)^{-1}} \leq \frac{r_{\max} + 2\sigma^2}{2\sqrt{\eta_{k-1}}} \cdot \sqrt{d \log \frac{k}{\delta'}}. \tag{51}$$

Let $\delta' = \frac{\delta}{T(K-1)}$ for some $\delta > 0$. Fix $\delta > 0$ and pick any s, a, h, k. For $\hat{\theta}_h^k$ defined in Algorithm 1 and $k > 1$, then with probability at least $1 - \frac{\delta}{T(K-1)}$,

$$|\langle \phi(s,a), \theta_h - \hat{\theta}_h^k \rangle| \leq \sqrt{d\phi(s,a)^\top (B_h^k)^{-1} \phi(s,a)} \cdot \left(\frac{r_{\max}}{2\sqrt{\eta_{k-1}}} \cdot \sqrt{\log \frac{Tk(K-1)}{\delta}} + 1 \right). \tag{52}$$

When $k = 1$, $|\phi(s,a), \theta_h - \hat{\theta}_h^k|$ is trivially at most $\sqrt{d}$, since $\hat{\theta}_h^1 = 0$.

For the rest of this section, we will need to condition on the event that (52) holds across various s, a, h, k. Denote $E(S)$ where S is a set containing tuples of the form (s, a, h, k) to be the event that for each element of S, we have that (52) holds. In addition, define the event $Z = E(\bigcap_{h,k} \{s_h^\tau, a_h^\tau, h, k\}_{\tau=1}^{k-1})$.

D Proof of Lemma 2

First, we bound the norm of w_h^k using our result from Lemma 1.

Lemma 6. *For any k, h, the weight w_h^k in Algorithm 1 satisfies*

$$\|w_h^k\| \leq (M_k + V_{max})\sqrt{dk} \tag{53}$$

conditioned on the event $E(\{s_h^\tau, a_h^\tau, h, k\}_{\tau=1}^{k-1})$, where $M_k = \sqrt{d}\left(\frac{r_{\max} + 2\sigma^2}{2\sqrt{w_{k-1}}} \cdot \sqrt{\log \frac{Tk(K-1)}{\delta}} + 2 \right)$.

Proof. For any vector $v \in \mathbb{R}^d$,

$$|v^\top w_h^k| = \left| v^\top (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau) [\langle \phi(s_h^\tau, a_h^\tau), \hat{\theta}_h^k \rangle + \max_a Q_{h+1}^k(s_{h+1}^\tau, a)] \right| \tag{54}$$

$$\leq \sum_{\tau=1}^{k-1} |v^\top (\Lambda_h^k)^{-1} \phi(s_h^\tau, a_h^\tau)| \cdot |\langle \phi(s_h^\tau, a_h^\tau), \hat{\theta}_h^k \rangle + \max_a Q_{h+1}^k(s_{h+1}^\tau, a)| \tag{55}$$

By definition, $\max_a Q_{h+1}^k(s_{h+1}^\tau, a) \leq V_{max}$. We now bound $\langle \phi(s_h^\tau, a_h^\tau), \hat{\theta}_h^k \rangle$ across all s_h^τ, a_h^τ for $\tau = 1, \ldots, k-1$ using Lemma 1. This requires the event $E(\{s_h^\tau, a_h^\tau, h, k\}_{\tau=1}^{k-1})$.

$$\langle \phi(s_h^\tau, a_h^\tau), \hat{\theta}_h^k \rangle \leq \langle \phi(s_h^\tau, a_h^\tau), \theta_h \rangle + \sqrt{d\phi(s_h^\tau, a_h^\tau)^T (B_h^k)^{-1} \phi(s_h^\tau, a_h^\tau)} \cdot \left(\frac{r_{\max} + 2\sigma^2}{2\sqrt{\eta_{k-1}}} \cdot \sqrt{\log \frac{Tk(K-1)}{\delta}} + 1 \right) \tag{56}$$

$$\leq 1 + \sqrt{d}\left(\frac{r_{\max} + 2\sigma^2}{2\sqrt{\eta_{k-1}}} \cdot \sqrt{\log \frac{Tk(K-1)}{\delta}} + 1 \right) \leq \sqrt{d}\left(\frac{r_{\max} + 2\sigma^2}{2\sqrt{\eta_{k-1}}} \cdot \sqrt{\log \frac{Tk(K-1)}{\delta}} + 2 \right) \tag{57}$$

We have used the fact that $\sqrt{\phi(s,a)^\top (B_h^k)^{-1}\phi(s,a)} \leq 1$, since the minimum eigenvalue of B_h^k is at least 1. Denote M_k as this upper bound in (57). We use the Cauchy-Schwarz inequality to get

$$\begin{aligned}|v^\top w_h^k| &\leq \sum_{\tau=1}^{k-1} |v^\top (\Lambda_h^k)^{-1}\phi(s_h^\tau, a_h^\tau)| \cdot (M_k + V_{max}) \\ &\leq \sqrt{\Big(\sum_{\tau=1}^{k-1} v^\top (\Lambda_h^k)^{-1} v\Big) \cdot \Big(\sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau)^\top (\Lambda_h^k)^{-1}\phi(s_h^\tau, a_h^\tau)\Big)} \cdot (M_k + V_{max}) \end{aligned} \tag{58}$$

By Lemma D.1 of Jin et al. [2020b], this becomes

$$|v^\top w_h^k| \leq (M_k + V_{max})\sqrt{d} \cdot \sqrt{\sum_{\tau=1}^{k-1} v^\top (\Lambda_h^k)^{-1} v} \tag{59}$$

Note that $\Lambda_h^k - I \succeq 0$, which means that $I^{-1} - (\Lambda_h^k)^{-1} \succeq 0$, so

$$|v^\top w_h^k| \leq (M_k + V_{max})\sqrt{d} \cdot \sqrt{\sum_{\tau=1}^{k-1} v^\top v} \leq (M_k + V_{max})\|v\|\sqrt{dk} \tag{60}$$

Since $\|w_h^k\| = \max_{\|v\|=1} v^\top w_h^k$, then the above expression implies that $\|w_h^k\| \leq (M_k + V_{max})\sqrt{dk}$ conditioned on the event $E(\{s_h^\tau, a_h^\tau, h, k\}_{\tau=1}^{k-1})$. □

Next, we present a concentration inequality used for the analysis of LSVI.

Lemma 7. *If we let the event $\mathfrak{E}$ be, for some fixed constant C,*

$$\forall (k,h): \quad \Big\| \sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau)[V_{h+1}^k(s_{h+1}^\tau) - \mathbb{E}\left[V_{h+1}^k(s_{h+1}^\tau)|s_h^\tau, a_h^\tau\right]] \Big\|_{(\Lambda_h^k)^{-1}} \leq C \cdot dV_{max}\sqrt{\chi} \tag{61}$$

where $V_h^k(s) = \max_a Q_h^k(s,a)$ and $\chi = \log[c_\beta r_{\max} dT/\delta]$, then $\mathbb{P}(\mathfrak{E}|Z) \geq 1 - \delta/2$.

Proof. Applying Lemma D.4 of Jin et al. [2020b], note that V^k is bounded and thus $V_{\max}$-subgaussian, so we get

$$\Big\| \sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau)[V_{h+1}^k(s_{h+1}^\tau) - \mathbb{E}\left[V_{h+1}^k(s_{h+1}^\tau)|s_h^\tau, a_h^\tau\right]] \Big\|_{(\Lambda_h^k)^{-1}}^2 \leq 4V_{\max}^2 \Big[\frac{d}{2}\log(k+1) + \log\frac{2\mathcal{N}_\varepsilon}{\delta}\Big] + 8k^2\varepsilon^2 \tag{62}$$

with probability $1 - \frac{\delta}{2}$, where $\mathcal{N}_\varepsilon$ is the ε-covering number of $\mathcal{V}$, the set of all V_h^k. By Lemma D.6 of Jin et al. [2020b], $\mathcal{N}_\varepsilon$ satisfies

$$\log \mathcal{N}_\varepsilon \leq d\log(1 + 4L/\varepsilon) + d^2 \log[1 + 8d^{1/2}B^2/\varepsilon^2], \tag{63}$$

where L is an upper bound on w_h^k, which holds for all w_h^k given the event Z and B is an upper bound on β_k. Using knowledge of p_k and $V_{\max}$, we have the following values for L and B when conditioned on Z:

$$L = (M_k + V_{max})\sqrt{dk} = \left(\sqrt{d}\left(\frac{r_{\max} + 2\sigma^2}{2\sqrt{\eta_{k-1}}} \cdot \sqrt{\log\frac{Tk(K-1)}{\delta}} + 2\right) + V_{max}\right)\sqrt{dk} \tag{64}$$

$$= \left(\frac{r_{\max} + 2\sigma^2}{2\sqrt{\eta_{k-1}}} \cdot \sqrt{\log\frac{Tk(K-1)}{\delta}} + 2 + \frac{V_{max}}{\sqrt{d}}\right) d\sqrt{k} \tag{65}$$

$$B = c_\beta d\sqrt{\log\frac{dT}{\delta}} \cdot \left(\frac{(r_{\max} + 2\sigma^2)\sqrt{d}}{\eta_{k-1}} + V_{\max}\right) \tag{66}$$

We use these values and let $\varepsilon = dV_{\max}/k$ to get

$$\log(1+4L/\varepsilon) = \log\left(1+4\left(\frac{r_{\max}+2\sigma^2}{2V_{\max}\sqrt{\eta_{k-1}}}\cdot\sqrt{\log\frac{Tk(K-1)}{\delta}}+2+\frac{1}{\sqrt{d}}\right)k\sqrt{k}\right) \tag{67}$$

We know there exists some constant C_1 such that this expression can be bounded by $C_1\log\frac{(r_{\max}+2\sigma^2)T}{\delta}$, where we use the fact that $V_{\max}$ scales in H and the polynomial dependence on T inside the log only impacts C_1. For $\log(1+8d^{1/2}B^2/\varepsilon^2)$, we use the fact that $(a+b)^2 \le 2a^2+2b^2$ to get

$$\log(1+8\sqrt{d}B^2/\varepsilon^2) = \log\left(1+16d^{5/2}c_\beta^2\log\left(\frac{dT}{\delta}\right)\left(\frac{(r_{\max}+2\sigma^2)^2 d}{\eta_{k-1}^2}+V_{\max}^2\right)\cdot\frac{k^2}{d^2V_{\max}^2}\right) \tag{68}$$

$$\le \log\left(1+16\sqrt{d}c_\beta^2\log\left(\frac{dT}{\delta}\right)\left(\frac{(r_{\max}+2\sigma^2)^2 d}{\eta_{k-1}^2V_{\max}^2}+1\right)k^2\right) \tag{69}$$

There exists some different constant C_2 such that this expression can be bounded by $C_2\log\frac{c_\beta(r_{\max}+2\sigma^2)dT}{\delta}$. Putting this all together,

$$\Big\|\sum_{\tau=1}^{k-1}\phi(s_h^\tau,a_h^\tau)[V_{h+1}^k(s_{h+1}^\tau)-\mathbb{P}_hV_{h+1}^k(s_h^\tau,a_h^\tau)]\Big\|^2_{(\Lambda_h^k)^{-1}} \tag{70}$$

$$\le 4V_{\max}^2\Big[\frac{d}{2}\log(k+1)+\log(2/\delta)+d\log(1+4L/\varepsilon)+d^2\log(1+8\sqrt{d}B^2/\varepsilon^2)\Big]+8d^2V_{\max}^2 \tag{71}$$

$$\le C^2V_{\max}^2d^2\log\left(\frac{c_\beta(r_{\max}+2\sigma^2)dT}{\delta}\right) \tag{72}$$

Taking the square root, we get our desired bound.

□

We are now ready to prove Lemma 2. Part of this proof is similar to Lemma B.4 of Jin et al. [2020b]), except w_h^k is defined with $\hat{\theta}_h^k$ and w_h^π uses θ_h, which results in an additional error term. By Lemma 3, we have that

$$\langle\phi(s,a),w_h^k\rangle - Q_h^\pi(s,a) = \langle\phi(s,a),w_h^k-w_h^\pi\rangle \tag{73}$$

We can write $w_h^k - w_h^\pi$ as

$$\begin{aligned}
w_h^k - w_h^\pi &= (\Lambda_h^k)^{-1}\Big(\sum_{\tau=1}^{k-1}\phi(s_h^\tau,a_h^\tau)[\langle\phi(s_h^\tau,a_h^\tau),\hat{\theta}_h^k\rangle+V_{h+1}^k(s_{h+1}^\tau)]\Big)-w_h^\pi \\
&= (\Lambda_h^k)^{-1}\Big(\sum_{\tau=1}^{k-1}\phi(s_h^\tau,a_h^\tau)[\langle\phi(s_h^\tau,a_h^\tau),\hat{\theta}_h^k\rangle+V_{h+1}^k(s_{h+1}^\tau)]-\Lambda_h^kw_h^\pi\Big) \\
&= (\Lambda_h^k)^{-1}\Big(\sum_{\tau=1}^{k-1}\phi(s_h^\tau,a_h^\tau)[\langle\phi(s_h^\tau,a_h^\tau),\hat{\theta}_h^k\rangle+V_{h+1}^k(s_{h+1}^\tau)-\phi(s_h^\tau,a_h^\tau)^\top w_h^\pi]\Big) \\
&\quad-(\Lambda_h^k)^{-1}w_h^\pi
\end{aligned} \tag{74}$$

The last line comes from the definition of $\Lambda_h^k = I+\sum_{\tau=1}^{k-1}\phi(s_h^\tau,a_h^\tau)\phi(s_h^\tau,a_h^\tau)^\top$. Furthermore, recall that w_h^π is defined as $\theta_h+\int V_{h+1}^\pi(s')d\mu_h(s')$. Then $\phi(s_h^\tau,a_h^\tau)^\top w_h^\pi = \langle\phi(s_h^\tau,a_h^\tau),\theta_h\rangle+\langle\phi(s_h^\tau,a_h^\tau),\int V_{h+1}^\pi(s')d\mu_h(s')\rangle =$

$\langle \phi(s_h^\tau, a_h^\tau), \theta_h \rangle + \mathbb{E}\left[V_{h+1}^\pi(s_{h+1}^\tau)|s_h^\tau, a_h^\tau\right]$. Plugging this back into (75), we have

$$\begin{aligned} w_h^k - w_h^\pi = \underbrace{(-(\Lambda_h^k)^{-1} w_h^\pi)}_{q_1} &+ \underbrace{(\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau)\left[V_{h+1}^k(s_{h+1}^\tau) - \mathbb{E}\left[V_{h+1}^k(s_{h+1}^\tau)|s_h^\tau, a_h^\tau\right]\right]}_{q_2} \\ &+ \underbrace{(\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau)\mathbb{E}\left[V_{h+1}^k(s_{h+1}^\tau) - V_{h+1}^\pi(s_{h+1}^\tau)|s_h^\tau, a_h^\tau\right]}_{q_3} \\ &+ \underbrace{(\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau)\langle \phi(s_h^\tau, a_h^\tau), \hat{\theta}_h^k - \theta_h \rangle}_{q_4} \end{aligned} \tag{75}$$

Note that q_4 is an additional term that arises due to estimating the linear reward component. Substituting back into (73), we have

$$\langle \phi(s,a), w_h^k \rangle - Q_h^{\pi_k}(s,a) = \langle \phi(s,a), q_1 + q_2 + q_3 + q_4 \rangle \tag{76}$$

We bound $\langle \phi(s,a), q_1 \rangle, \langle \phi(s,a), q_2 \rangle$, $\langle \phi(s,a), q_3 \rangle$, and $\langle \phi(s,a), q_4 \rangle$ separately.

- q_1:

$$\begin{aligned} \langle \phi(s,a), q_1 \rangle = \langle \phi(s,a), -(\Lambda_h^k)^{-1} w_h^\pi \rangle &= \langle (\Lambda_h^k)^{-1}\phi(s,a), -w_h^\pi \rangle && (77) \\ &\leq \|w_h^\pi\| \sqrt{\phi(s,a)^\top (\Lambda_h^k)^{-1}(\Lambda_h^k)^{-1}\phi(s,a)} && (78) \\ &\leq \|w_h^\pi\| \sqrt{\phi(s,a)^\top (\Lambda_h^k)^{-1}\phi(s,a)} && (79) \\ &\leq 2V_{max}\sqrt{d} \cdot \sqrt{\phi(s,a)^\top (\Lambda_h^k)^{-1}\phi(s,a)} && (80) \end{aligned}$$

  In the second inequality, we've used the fact that $(\Lambda_h^k)^{-1} - (\Lambda_h^k)^{-1}(\Lambda_h^k)^{-1} \succeq 0$ (this can be checked using Weyl's ineqality on $I - (\Lambda_h^k)^{-1}$). In the third inequality, we apply Lemma 4.

- q_2: Since $(\Lambda_h^k)^{-1}$ is positive definite and symmetric, there exists a symmetric matrix $(\Lambda_h^k)^{-1/2}$ such that

$$\begin{aligned} \langle \phi(s,a), q_2 \rangle &= \phi(s,a)^\top (\Lambda_h^k)^{-1/2}(\Lambda_h^k)^{-1/2} \sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau)\left[V_{h+1}^k(s_{h+1}^\tau) - \mathbb{E}\left[V_{h+1}^k(s_{h+1}^\tau)|s_h^\tau, a_h^\tau\right]\right] && (81) \\ &= \left\langle (\Lambda_h^k)^{-1/2}\phi(s,a), (\Lambda_h^k)^{-1/2} \sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau)\left[V_{h+1}^k(s_{h+1}^\tau) - \mathbb{E}\left[V_{h+1}^k(s_{h+1}^\tau)|s_h^\tau, a_h^\tau\right]\right] \right\rangle && (82) \\ &\leq \|(\Lambda_h^k)^{-1/2}\phi(s,a)\| \cdot \|(\Lambda_h^k)^{-1/2} \sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau)\left[V_{h+1}^k(s_{h+1}^\tau) - \mathbb{E}\left[V_{h+1}^k(s_{h+1}^\tau)|s_h^\tau, a_h^\tau\right]\right]\| && (83) \\ &\leq C \cdot dV_{max}\sqrt{\chi} \cdot \sqrt{\phi(s,a)^\top (\Lambda_h^k)^{-1}\phi(s,a)} && (84) \end{aligned}$$

  We get the last line from Lemma 7, conditioned on $\mathfrak{E}$ and Z.

- q_3: We first write the expectation as an integral using the linear dynamics assumption:

$$\begin{aligned} \langle \phi(s,a), q_3 \rangle &= \left\langle \phi(s,a), (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau)\mathbb{E}\left[V_{h+1}^k(s_{h+1}^\tau) - V_{h+1}^\pi(s_{h+1}^\tau)|s_h^\tau, a_h^\tau\right] \right\rangle && (85) \\ &\leq \left\langle \phi(s,a), (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top \int V_{h+1}^k(s') - V_{h+1}^\pi(s') d\mu_h(s') \right\rangle. && (86) \end{aligned}$$

Note that $(\Lambda_h^k)^{-1}\sum_{\tau=1}^{k-1}\phi(s_h^\tau,a_h^\tau)\phi(s_h^\tau,a_h^\tau)^\top = I-(\Lambda_h^k)^{-1}$. Then the dot product is written as

$$\left\langle \phi(s,a), \int V_{h+1}^k(s') - V_{h+1}^\pi(s')d\mu_h(s') \right\rangle - \left\langle \phi(s,a), (\Lambda_h^k)^{-1}\int V_{h+1}^k(s') - V_{h+1}^\pi(s')d\mu_h(s') \right\rangle \tag{87}$$

By definition of expectation, the first dot product above is equal to $\mathbb{E}\left[V_{h+1}^k(s_{h+1}^k) - V_{h+1}^\pi(s_{h+1}^k)|s_h^k=s,a_h^k=a\right]$. Using Cauchy-Schwarz, the second one is less than $\sqrt{\phi(s,a)^\top(\Lambda_h^k)^{-1}\phi(s,a)}\cdot\|(\Lambda_h^k)^{-1/2}\int V_{h+1}^k(s') - V_{h+1}^\pi(s')d\mu_h(s')\|$. Using the fact that $\|(\Lambda_h^k)^{-1/2}\|\leq 1$, $\|\mu_h(\mathcal{S})\|\leq\sqrt{d}$, and $|V_{h+1}^k(s')|, |V_{h+1}^\pi(s')|\leq V_{\max}$, the second dot product in (87) is at most $2V_{\max}\sqrt{d}\cdot\sqrt{\phi(s,a)^\top(\Lambda_h^k)^{-1}\phi(s,a)}$. Putting this together, we get

$$\langle\phi(s,a),q_3\rangle \leq \mathbb{E}\left[V_{h+1}^k(s_{h+1}^k) - V_{h+1}^\pi(s_{h+1}^k)|s_h^k=s,a_h^k=a\right] + 2V_{max}\sqrt{d}\cdot\sqrt{\phi(s,a)^\top(\Lambda_h^k)^{-1}\phi(s,a)}. \tag{88}$$

- q_4:

$$\langle\phi(s,a),q_4\rangle = \langle\phi(s,a),(\Lambda_h^k)^{-1}\sum_{\tau=1}^{k-1}\phi(s_h^\tau,a_h^\tau)\phi(s_h^\tau,a_h^\tau)^\top(\hat{\theta}_h^k-\theta_h) \tag{89}$$

$$= \langle(\Lambda_h^k)^{-1/2}\phi(s,a),(\Lambda_h^k)^{-1/2}\sum_{\tau=1}^{k-1}\phi(s_h^\tau,a_h^\tau)\phi(s_h^\tau,a_h^\tau)^\top(\hat{\theta}_h^k-\theta_h)\rangle \tag{90}$$

$$\leq \|(\Lambda_h^k)^{-1/2}\phi(s,a)\|\cdot\|(\Lambda_h^k)^{-1/2}\sum_{\tau=1}^{k-1}\phi(s_h^\tau,a_h^\tau)\phi(s_h^\tau,a_h^\tau)^\top(\hat{\theta}_h^k-\theta_h)\| \tag{91}$$

$$\leq \sqrt{\phi(s,a)^\top(\Lambda_h^k)^{-1}\phi(s,a)}\cdot\sum_{\tau=1}^{k-1}\|(\Lambda_h^k)^{-1/2}\phi(s_h^\tau,a_h^\tau)\|\cdot\phi(s_h^\tau,a_h^\tau)^\top(\hat{\theta}_h^k-\theta_h) \tag{92}$$

$$\leq \sqrt{\phi(s,a)^\top(\Lambda_h^k)^{-1}\phi(s,a)}\cdot\sqrt{\sum_{\tau=1}^{k-1}\phi(s_h^\tau,a_h^\tau)^\top(\Lambda_h^k)^{-1}\phi(s_h^\tau,a_h^\tau)} \tag{93}$$

$$\cdot\sqrt{\sum_{\tau=1}^{k-1}(\phi(s_h^\tau,a_h^\tau)^\top(\hat{\theta}_h^k-\theta_h))^2} \tag{94}$$

We can use Lemma D.1 of Jin et al. [2020b] to bound $\sqrt{\sum_{\tau=1}^{k-1}\phi(s_h^\tau,a_h^\tau)^\top(\Lambda_h^k)^{-1}\phi(s_h^\tau,a_h^\tau)}\leq\sqrt{d}$. Next, we bound $\sqrt{\sum_{\tau=1}^{k-1}(\phi(s_h^\tau,a_h^\tau)^\top(\hat{\theta}_h^k-\theta_h))^2}$ using Lemma 1, which requires the event $E(\{s_h^\tau,a_h^\tau,h,k\}_{\tau=1}^{k-1})$.

$$\sqrt{\sum_{\tau=1}^{k-1}(\phi(s_h^\tau,a_h^\tau)^T(\hat{\theta}_h^k-\theta_h))^2} \leq \left(\frac{r_{\max}+2\sigma^2}{2\sqrt{\eta_{k-1}}}\cdot\sqrt{\log\frac{Tk(K-1)}{\delta}}+1\right) \tag{95}$$

$$\cdot\sqrt{d\sum_{\tau=1}^{k-1}\phi(s_h^\tau,a_h^\tau)^\top(B_h^k)^{-1}\phi(s_h^\tau,a_h^\tau)} \tag{96}$$

We can apply Lemma D.1 of Jin et al. [2020b] again, but since B_h^k is a weighted covariance matrix, we have that

$$\sum_{\tau=1}^{k-1}\eta_\tau\phi(s_h^\tau,a_h^\tau)^\top(B_h^k)^{-1}\phi(s_h^\tau,a_h^\tau)\leq d, \tag{97}$$

and hence

$$\sqrt{\sum_{\tau=1}^{k-1}(\phi(s_h^\tau, a_h^\tau)^\top(\hat{\theta}_h^k - \theta_h))^2} \leq \left(\frac{r_{\max} + 2\sigma^2}{2\sqrt{\eta_{k-1}}} \cdot \sqrt{\log\frac{Tk(K-1)}{\delta}} + 1\right) \cdot \frac{d}{\sqrt{\eta_{k-1}}} \tag{98}$$

$$\leq \frac{C_r d(r_{\max} + 2\sigma)^2}{\eta_{k-1}} \sqrt{\log\frac{T}{\delta}} \tag{99}$$

for some constant C_r. Note that when $k = 1$, this term is equal to 0.

Putting everything together in (73), we get that there exists a constant c' such that $\langle \phi(s,a), w_h^k \rangle - Q_h^\pi(s,a)$ is at most

$$\mathbb{E}\left[V_{h+1}^k(s_{h+1}^k) - V_{h+1}^{\pi_k}(s_{h+1}^k) | s_h^k = s, a_h^k = a\right] + d\sqrt{\chi}\left(c' \cdot V_{max} + \frac{C_r(r_{\max} + 2\sigma^2)\sqrt{d}}{\eta_{k-1}}\right)\sqrt{\phi(s,a)^T(\Lambda_h^k)^{-1}\phi(s,a)} \tag{100}$$

$$\leq \mathbb{E}\left[V_{h+1}^k(s_{h+1}^k) - V_{h+1}^{\pi_k}(s_{h+1}^k) | s_h^k = s, a_h^k = a\right] + c_\beta d\sqrt{\iota}\left(V_{max} + \frac{(r_{\max} + 2\sigma^2)\sqrt{d}}{\eta_{k-1}}\right)\sqrt{\phi(s,a)^\top(\Lambda_h^k)^{-1}\phi(s,a)} \tag{101}$$

where the term $\frac{(r_{\max}+2\sigma^2)\sqrt{d}}{\eta_{k-1}}$ is absent if $k = 1$. We know there exists some constant c_β that bounds $\sqrt{\chi}$ in terms of $\sqrt{\iota}$. Therefore, by setting $\beta_k = c_\beta d\sqrt{\iota}\left(V_{max} + \frac{(r_{\max}+2\sigma^2)\sqrt{d}}{\eta_{k-1}}\right)$ when $k > 1$ and $\beta_1 = c_\beta d\sqrt{\iota}V_{\max}$, we get our desired result. Note that the result is conditional on $\mathfrak{E} \cap Z$, but the reward estimate bound doesn't need to hold for the s, a in the lemma statement.

E Proof of Theorem 1

Lemma 2 is used to show that Q_h^k is an upper confidence bound on $Q_h^\star$ and to bound the gap between Q_h^k and $Q_h^{\pi_k}$, which make up two intermediate lemmas used in the regret bound analysis.

Lemma 8. *(UCB) For any s, a, h, k, conditioned on the event $\mathfrak{E} \cap Z$ we have that*

$$Q_h^\star(s,a) \leq Q_h^k(s,a) \tag{102}$$

Proof. We prove this by induction and use the result from Lemma 2 that

$$\left|\langle \phi(s,a), w_h^k \rangle - Q_h^\star(s,a) - \mathbb{E}\left[V_{h+1}^k(s_{h+1}^k) - V_{h+1}^\star(s_{h+1}^k) | s_h^k = s, a_h^k = a\right]\right| \leq \beta_k\sqrt{\phi(s,a)^\top(\Lambda_h^k)^{-1}\phi(s,a)}. \tag{103}$$

Base case We show $Q_H^\star(s,a) \leq Q_H^k(s,a)$ conditioned on $\mathfrak{E} \cap Z$. $\mathbb{E}\left[V_{H+1}^k(s_{h+1}^k) - V_{H+1}^\star(s_{h+1}^k) | s_h^k = s, a_h^k = a\right] = 0$ since the value of a state at time $H+1$ is 0, and thus (103) gives us

$$Q_H^\star(s,a) \leq \langle \phi(s,a), w_H^k \rangle + \beta_k\sqrt{\phi(s,a)^\top(\Lambda_H^k)^{-1}\phi(s,a)} \tag{104}$$

We also know that $Q_H^\star(s,a) \leq V_{\max}$, so by definition of Q_H^k we have $Q_H^\star(s,a) \leq Q_H^k(s,a)$.

Inductive hypothesis Suppose that $Q_{h+1}^\star(s,a) \leq Q_{h+1}^k(s,a)$ conditioned on $\mathfrak{E} \cap Z$.

Inductive step Applying the inductive hypothesis,

$$\mathbb{E}\left[V_{h+1}^k(s_{h+1}^k) - V_{h+1}^\star(s_{h+1}^k) | s_h^k = s, a_h^k = a\right] \geq \mathbb{E}\left[Q_{h+1}^k(s_{h+1}^k, a') - Q_{h+1}^\star(s_{h+1}^k, a') | s_h^k = s, a_h^k = a\right] \geq 0 \tag{105}$$

where $a' = \text{argmax}_a Q^\star_{h+1}(s^k_{h+1}, a)$. Therefore, using (103) again, we get that

$$\mathbb{E}\left[V^k_{h+1}(s^k_{h+1}) - V^\star_{h+1}(s^k_{h+1})|s^k_h = s, a^k_h = a\right] + Q^\star_h(s,a) - \langle \phi(s,a), w^k_h\rangle \leq \beta_k \sqrt{\phi(s,a)^\top (\Lambda^k_h)^{-1}\phi(s,a)} \quad (106)$$

$$\Rightarrow Q^\star_h(s,a) \leq \langle \phi(s,a), w^k_h\rangle + \beta_k \sqrt{\phi(s,a)^\top (\Lambda^k_h)^{-1}\phi(s,a)} \quad (107)$$

and therefore $Q^\star_h(s,a) \leq Q^k_h(s,a)$. We complete this proof by induction.

□

We look at expected regret over the importance sampling done with p_k. Recall that $\bar{a}^k_h$ is the greedy action $\text{argmax}_a Q^k_h(s^k_h, a)$, and define $a^{k\star}_h = \text{argmax}_a Q^\star_h(a^k_h, a)$ as the optimal action. Expected regret is defined as

$$\mathbb{E}\left[\text{Regret}(K)\right] = \mathbb{E}\left[\sum_{k=1}^K V^\star_1(s^k_1) - V^{\pi_k}_1(s^k_1)\right] = \mathbb{E}\left[\sum_{k=1}^K Q^\star_1(s^k_1, a^{k\star}_1) - Q^{\pi_k}_1(s^k_1, a^k_1)\right] \quad (108)$$

By Lemma 8, we know that $Q^\star_1(s^k_1, a^{k\star}_1) \leq Q^k_1(s^k_1, \tilde{a}^k_1)$ conditioned on $\mathfrak{E} \cap Z$ for each k, to get that

$$\begin{aligned}\mathbb{E}\left[\text{Regret}(K)\right] &\leq \mathbb{E}\left[\sum_{k=1}^K Q^k_1(s^k_1, a^{k\star}_1) - Q^{\pi_k}_1(s^k_1, a^k_1)\right] \\ &\leq \sum_{k=1}^K \mathbb{E}\left[Q^k_1(s^k_1, \bar{a}^k_1) - Q^{\pi_k}_1(s^k_1, a^k_1)\right], \end{aligned} \quad (109)$$

where the second inequality follows by definition of $\bar{a}^k_h$. Note that this expression is challenging because $\bar{a}^k_h$ and a^k_h are not the same. For each k, define the event E_k that $a^k_h = a_{0,h}$ for at least one $h \in [H]$. We can condition $\mathbb{E}\left[Q^{\pi_k}_1(s^k_1, a^k_1)\right]$ based on following the trajectory of E_k or not:

$$\begin{aligned}&\sum_{k=1}^K \mathbb{E}\left[Q^k_1(s^k_1, \bar{a}^k_1) - Q^{\pi_k}_1(s^k_1, a^k_1)\right] \\ =&\sum_{k=1}^K \Big(\mathbb{E}\left[Q^k_1(s^k_1, \bar{a}^k_1) - Q^{\pi_k}_1(s^k_1, a^k_1)|E^C_k\right]\Pr(E^C_k) + \mathbb{E}\left[Q^k_1(s^k_1, \bar{a}^k_1) - Q^{\pi_k}_1(s^k_1, a^k_1)|E_k\right]\Pr(E_k)\Big)\end{aligned}$$

Under the event E^C_k, the agent selects the greedy action $\bar{a}^k_h$ for the entire episode k, and then the action taken at each timestep is equal to the greedy action, making for easier comparison using Lemma **??**. The probability of E^C_k is $(1-p_k)^H$. On the other hand, under the event E_k there is at least one default action taken in episode k. However, we know that $Q^k_1(s^k_1, \bar{a}^k_1) - Q^{\pi_k}_1(s^k_1, a^k_1)$ is trivially less than $2V_{\max}$. The probability of E_k is $1-(1-p_k)^H$, which is at most $1-(1-Hp_k) = Hp_k$. Putting this back together, our regret bound is now

$$\mathbb{E}\left[\text{Regret}(K)\right] \leq \sum_{k=1}^K (1-p_k)^H \mathbb{E}\left[Q^k_1(s^k_1, \bar{a}^k_1) - Q^{\pi_k}_1(s^k_1, \bar{a}^k_1)|E^C_k\right] + 2HV_{\max}\sum_{k=1}^K p_k \quad (110)$$

We now focus on bounding the regret with each episode conditioned on E^C_k. We need the following intermediate lemma.

Lemma 9. *Define* $\delta^k_h = \mathbb{E}\left[Q^{k\star}_h(s^k_h, \bar{a}^k_h) - Q^{\pi_k}_h(s^k_h, \bar{a}^k_h)|E^C_k\right]$ *and* $\zeta^k_{h+1} = \mathbb{E}\left[\delta^k_{h+1}|s^k_h, \bar{a}^k_h\right] - \delta^k_{h+1}$. *Then conditioned on the event* $\mathfrak{E} \cap Z$ *and* E^C_k, *for any* h, k,

$$\delta^k_h \leq \zeta^k_{h+1} + \delta^k_{h+1} + 2\beta^k_h \sqrt{\phi(s^k_h, \bar{a}^k_h)^\top (\Lambda^k_h)^{-1}\phi(s^k_h, \bar{a}^k_h)} \quad (111)$$

Proof. The policy π_k conditioned on E_k^C can be redefined as a new policy, which we denote as $\bar{\pi}_k$—in this case, this policy is equivalent to discarding the action-centering step completely. By Lemma **??**, we have that

$$\begin{aligned}
&\langle \phi(s,a), w_h^k \rangle + \beta_k \sqrt{\phi(s,a)^\top (\Lambda_h^k)^{-1} \phi(s,a)} - \beta_k \sqrt{\phi(s,a)^\top (\Lambda_h^k)^{-1} \phi(s,a)} - Q_h^{\bar{\pi}_k}(s,a) \\
&\leq \mathbb{E}\left[V_{h+1}^k(s_{h+1}^k) - V_{h+1}^{\bar{\pi}_k}(s_{h+1}^k) | s_h^k = s, a_h^k = a \right] + \beta_k \sqrt{\phi(s,a)^\top (\Lambda_h^k)^{-1} \phi(s,a)} \\
\Rightarrow Q_h^k(s,a) - Q_h^{\bar{\pi}_k}(s,a) &\leq \mathbb{E}\left[V_{h+1}^k(s_{h+1}^k) - V_{h+1}^{\bar{\pi}_k}(s_{h+1}^k) | s_h^k = s, a_h^k = a \right] + 2\beta_k \sqrt{\phi(s,a)^\top (\Lambda_h^k)^{-1} \phi(s,a)}
\end{aligned} \tag{112}$$

If we let $s = s_h^k$, $a = \bar{a}_h^k$, this becomes:

$$\delta_h^k \leq \mathbb{E}\left[V_{h+1}^k(s_{h+1}^k) - V_{h+1}^{\bar{\pi}_k}(s_{h+1}^k) | s_h^k, \bar{a}_h^k \right] + 2\beta_k \sqrt{\phi(s_h^k, \bar{a}_h^k)^\top (\Lambda_h^k)^{-1} \phi(s_h^k, \bar{a}_h^k)} \tag{113}$$

Note that $\mathbb{E}\left[\delta_{h+1}^k | s_h^k, \bar{a}_h^k\right]$ is equal to $\mathbb{E}\left[Q_{h+1}^k(s_{h+1}^k, \bar{a}_{h+1}^k) - Q_{h+1}^{\bar{\pi}_k}(s_{h+1}^k, \bar{a}_{h+1}^k) | s_h^k, \bar{a}_h^k\right]$ (we implicitly condition on E_k^C by using $\bar{\pi}_k$). Since we are taking the greedy action under policy $\bar{\pi}_k$, this is equal to $\mathbb{E}\left[V_{h+1}^k(s_{h+1}^k) - V_{h+1}^{\bar{\pi}_k}(s_{h+1}^k) | s_h^k, \bar{a}_h^k\right]$. Therefore, we have that

$$\delta_h^k \leq \zeta_{h+1}^k + \delta_{h+1}^k + 2\beta_k \sqrt{\phi(s_h^k, a_h^k)^\top (\Lambda_h^k)^{-1} \phi(s_h^k, a_h^k)} \tag{114}$$

□

By Lemma 9, conditioned on $\mathfrak{E} \cap Z$ and E_k^C,

$$\begin{aligned}
&\sum_{k=1}^K (1-p_k)^H \mathbb{E}\left[Q_1^k(s_1^k, \bar{a}_1^k) - Q_1^{\pi_k}(s_1^k, \bar{a}_1^k) | E_k^C \right] \leq \sum_{k=1}^K \delta_1^k \\
&= \sum_{k=1}^K \left(\sum_{h=1}^H \zeta_h^k + 2 \sum_{h=1}^H \beta_k \sqrt{\phi(s_h^k, \bar{a}_h^k)^\top (\Lambda_h^k)^{-1} \phi(s_h^k, \bar{a}_h^k)} \right) \\
&= \underbrace{\sum_{k=1}^K \sum_{h=1}^H \zeta_h^k}_{a_1} + \underbrace{2 \sum_{k=1}^K \sum_{h=1}^H \beta_k \sqrt{\phi(s_h^k, \bar{a}_h^k)^\top (\Lambda_h^k)^{-1} \phi(s_h^k, \bar{a}_h^k)}}_{a_2}
\end{aligned} \tag{115}$$

We bound a_1 and a_2 separately:

- a_1: Note that $\{\zeta_h^k\}_{h,k}$ is a martingale difference sequence. Furthermore, $|\zeta_h^k| \leq 4V_{max}$ since $|\delta_h^k| \leq 2V_{max}$. Therefore the difference between the largest possible and smallest possible value of ζ_h^k is $c_h^k = 8V_{max}$, and we apply the Azuma-Hoeffding inequality:

$$P\left(a_1 \geq \epsilon\right) \leq \exp\left(-\frac{2\epsilon^2}{\sum_{h,k} 64 V_{max}^2} \right) \tag{116}$$

$$= \exp\left(-\frac{\epsilon^2}{32 V_{max}^2 T} \right) \tag{117}$$

  This can be written as

$$a_1 \leq \left(32 V_{max}^2 T \log(2/\delta) \right)^{1/2} \qquad w.p.\ 1 - \frac{\delta}{2} \tag{118}$$

- a_2: We use Cauchy-Schwarz to get

$$a_2 \leq 2\sum_{h=1}^{H}\Big(\sum_{k=1}^{K}\beta_k^2\Big)^{1/2}\cdot\Big(\sum_{k=1}^{K}\phi(s_h^k,\bar{a}_h^k)^\top(\Lambda_h^k)^{-1}\phi(s_h^k,\bar{a}_h^k)\Big)^{1/2} \leq 2\sum_{h=1}^{H}\Big(\sum_{k=1}^{K}\beta_k^2\Big)^{1/2}\sqrt{2d\iota}$$

$$\leq 2\sum_{h=1}^{H}\Big(c_\beta^2 d^2\iota V_{\max}^2 + \sum_{k=2}^{K} 2c_\beta^2 d^2\iota(V_{\max}^2 + \frac{(r_{\max}+2\sigma^2)^2 d}{p_{k-1}^2(1-p_{k-1})^2})\Big)^{1/2}\sqrt{2d\iota} \tag{119}$$

$$\leq 2\sum_{h=1}^{H}\Big(2Kc_\beta^2 d^2\iota V_{\max}^2 + \sum_{k=2}^{K} 2c_\beta^2 d^2\iota\frac{(r_{\max}+2\sigma^2)^2 d}{p_{k-1}^2(1-p_{k-1})^2}\Big)^{1/2}\sqrt{2d\iota} \tag{120}$$

$$\leq 2H\sqrt{2d\iota}\cdot c_\beta d\sqrt{\iota}\left(\sqrt{2V_{\max}^2 K} + \sqrt{2(r_{\max}+2\sigma^2)^2 d\sum_{k=2}^{K}\frac{1}{p_{k-1}^2(1-p_{k-1})^2}}\right) \tag{121}$$

$$\leq 4Hc_\beta\iota d^{3/2}\left(V_{\max}\sqrt{K} + (r_{\max}+2\sigma^2)\sqrt{\sum_{k=2}^{K}\frac{d}{p_{k-1}^2(1-p_{k-1})^2}}\right) \tag{122}$$

Where the second inequality comes from Lemma D.2 of Jin et al. [2020b].

Putting this together using (110), our regret bound is now

$$\mathbb{E}\left[\text{Regret}(K)\right] \leq \Big(32V_{max}^2 T\log(2/\delta)\Big)^{1/2} + 4Hc_\beta\iota d^{3/2}\left(V_{\max}\sqrt{K} + (r_{\max}+2\sigma^2)\sqrt{\sum_{k=2}^{K}\frac{d}{p_{k-1}^2(1-p_{k-1})^2}}\right)$$
$$+ 2HV_{max}\sum_{k=1}^{K}p_k \tag{123}$$

This inequality illustrates a tradeoff depending on the value of p_k. We thus must set a p_k as to minimize the order of T in the bound. Let p_k be of the form $\frac{1}{(k+1)^q}$ (since p_k cannot be 0 or 1). Then note that

$$\sum_{k=1}^{K}p_k = \sum_{k=1}^{K}\frac{1}{(k+1)^q} \leq \int_1^{K+1}k^{-q}dk = \frac{1}{1-q}k^{1-q}\Big|_1^{K+1} \leq \frac{1}{1-q}(K+1)^{1-q} \tag{124}$$

And

$$\sqrt{\sum_{k=2}^{K}\frac{1}{p_{k-1}^2(1-p_{k-1})^2}} = \sqrt{\sum_{k=1}^{K-1}\frac{1}{p_k^2(1-p_k)^2}} \leq \sqrt{\frac{1}{(1-\frac{1}{2^q})^2}\sum_{k=1}^{K-1}\frac{1}{p_k^2}} \tag{125}$$

$$= \sqrt{\frac{1}{(1-\frac{1}{2^q})^2}\sum_{k=1}^{K-1}(k+1)^{2q}} \tag{126}$$

$$\leq \sqrt{\frac{1}{(1-\frac{1}{2^q})^2}}K^{\frac{2q+1}{2}} \tag{127}$$

This justifies why we want $q=\frac{1}{4}$, since it satisfies $1-q=\frac{2q+1}{2}$. Substituting this in along with $V_{\max}=H$ gives an overall regret bound of

$$\mathbb{E}\left[Regret(K)\right] \leq \Big(32H^2 T\log(2/\delta)\Big)^{1/2} + 4Hc_\beta\iota d^{3/2}\left(H\sqrt{K} + 7(r_{\max}+2\sigma^2)K^{3/4}\sqrt{d}\right)$$
$$+ \frac{8}{3}H^2(K+1)^{3/4} \tag{128}$$

This is $\mathcal{O}((r_{\max}+2\sigma^2)d^2\iota H^{3/2}T^{3/4})$.

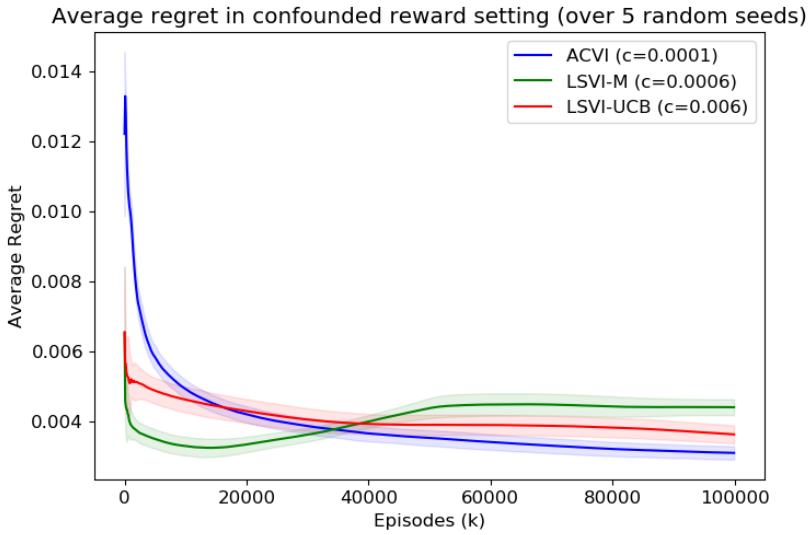

Figure 2: Average regret for ACVI ($c = 0.0001$), LSVI-M ($c = 0.0006$), and LSVI-UCB ($c = 0.006$) over 5 random seeds.

Probability of regret bound Note that the above regret bound is conditioned on the event $\mathfrak{E} \cap Z$. This probability is

$$\Pr(\mathfrak{E} \cap Z) = \Pr(\mathfrak{E}|Z) \Pr(Z) \geq \left(1 - \frac{\delta}{2}\right) \Pr(Z) \tag{129}$$

Next, Z consists of $H \sum_{k=1}^{K-1} k = H\binom{K}{2}$ tuples. Lemma 1 holds for an individual (s, a, h, k) with probability at least $1 - \frac{\delta}{(T(K-1)}$. Then, using a union bound we get that $\Pr(Z) \geq 1 - \frac{\delta}{T(K-1)} \cdot H\binom{K}{2} = 1 - \frac{\delta}{2}$. Therefore,

$$\Pr(\mathfrak{E} \cap Z) \geq \left(1 - \frac{\delta}{2}\right)^2 \geq 1 - \delta. \tag{130}$$

F Additional Experimental Results

We discuss more details about our experimental setup and provide additional results. First, we describe our approach to select a reasonable UCB constant c. Define $c_0 = \Delta$, and increment this by Δ until $Q_h^\star(s, a) \leq Q_h^k(s, a)$ holds for all s, a, h, k. For ACVI and LSVI-M, we use $\Delta = 0.0001$, and for LSVI-UCB we use a larger increment $\Delta = 0.001$ since β for LSVI-UCB is smaller. Our choice of c using this approach is $0.0001, 0.0006, 0.006$ for ACVI, LSVI-M, and LSVI-UCB, respectively. We rerun our procedure in section 7 over 5 random seeds, plotting the average and standard error in Figure 2 to verify that ACVI attains lower average regret than LSVI-M and LSVI-UCB empirically.

F.1 Setting the sampling schedule p_k

We examine the effect of changing p_k on the average regret and empirically verify that $p_k \sim \frac{1}{k^{1/4}}$ outperforms other rates for p_k. We run ACVI with $c = 0.0001$ using our proposed p_k, and compare against p_k scaling with $k^{-1}, k^{-1/2}, k^{-1/8}$. Under our protocol for choosing the UCB constant, we select $c = 0.0001$ for the p_k alternatives as well. The average regret and standard error across the same 5 random seeds are plotted in Figure 3. We see that having p_k decrease at a faster rate leads to the average regret converging to a nonzero constant, which matches our theoretical results that smaller p_k leads to poorer estimates of θ_h. p_k decreasing at a slower rate also has a higher average regret, matching our theoretical results that frequently choosing the default action worsens regret.

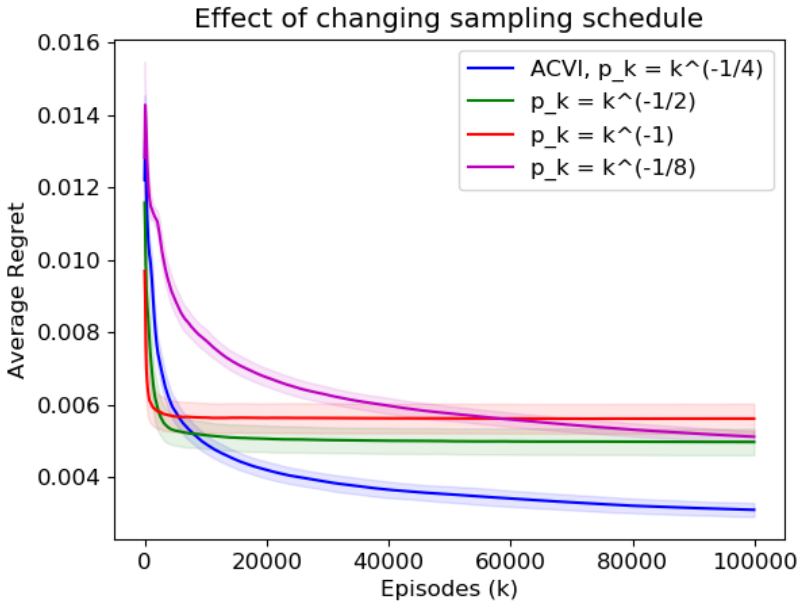

Figure 3: Average regret for ACVI with sampling schedule $p_k \approx k^{-1/4}, k^{-1/2}, k^{-1}, k^{-1/8}$ over 5 random seeds, $c = 0.0001$. The regret curve corresponding to ACVI ($p_k \approx k^{-1/4}$) attains the lowest average regret.

F.2 Robustness to confounding reward structure

We examine the effect of changing $f_h^k(\cdot)$ on the average regret and verify that, if $|f_h^k|$ is fixed, the performance of ACVI is largely unaffected by further reward structure. In addition to the confounding reward defined in (13), which we refer to as "oscillating", we consider two simple f_h^k, "index" and "dot product", scaled to the same approximate magnitude as that of (13):

$$\begin{aligned} f_{h,\text{index}}^k(s_1^k) &= 0.43 \cdot s_1^k[h] \\ f_{h,\text{dot product}}^k(s_1^k) &= 0.15 z_h^\top s_1^k. \end{aligned} \tag{131}$$

Recall that $z_h \sim \text{Unif}(-1,1)^{d'}$ and $s_1^k[h]$ represents the hth-indexed element of s_1^k. These confounding reward components differ from the one defined in (13) because they no longer oscillate as a function of k. Since the magnitude of these confounding reward components are the same, we use $c = 0.0001$ for these reward models. The average regret and standard error across the same 5 random seeds are plotted in Figure 4. This empirically confirms that the structure of the reward beyond the magnitude does not impact ACVI significantly.

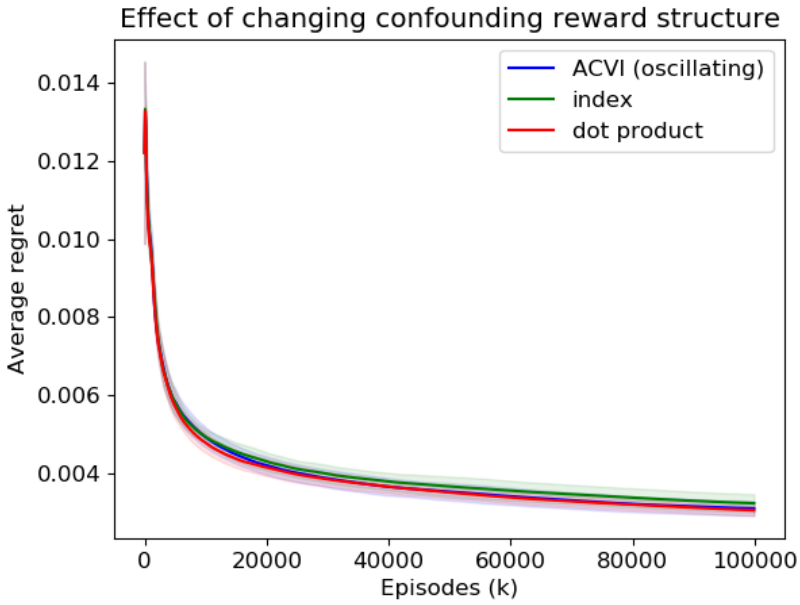

Figure 4: Average regret for ACVI, using the confounding rewards defined in (13) and (131) over 5 random seeds, $c = 0.0001$. All three curves decrease at roughly the same rate.