Exploration-Free Reinforcement Learning with Linear Function Approximation

Anonymous authors
Paper under double-blind review

Keywords: exploration-free, linear function approximation, no-regret.

Summary

In the context of Markov Decision Processes (MDPs) with linear Bellman completeness, a generalization of linear MDPs, we reconsider the learning capabilities of a *greedy* algorithm. The motivation is that, when exploration is costly or dangerous, an exploration-free approach may be preferable to optimistic or randomized solutions. We show that, under a condition of sufficient diversity in the feature distribution, Least-Squares Value Iteration (LSVI) can achieve sublinear regret. Specifically, we show that the expected cumulative regret is at most $\tilde{\mathcal{O}}(H^3\sqrt{dK/\lambda_0})$, where K is the number of episodes, H is the task horizon, d is the dimension of the feature map and λ_0 is a measure of feature diversity. We empirically validate our theoretical findings on synthetic linear MDPs. Our analysis is a first step towards exploration-free reinforcement learning in MDPs with large state spaces.

Contribution(s)

1. The definition of a new diversity condition for linear MDPs.
   Context: Inspired from prior work of Bastani et al. (2021) and Kannan et al. (2018).
2. Proved that a greedy algorithm (LSVI) achieves sublinear cumulative regret with high probability when the here defined diversity condition is satisfied.
   Context: Proof built upon the related work on linear contextual bandit of Bastani et al. (2021).
3. Proved that a greedy algorithm (LSVI) achieves sublinear cumulative regret with high probability when the here defined diversity condition is satisfied, under a misspecified setting.
   Context: Proof built upon the related work on approximately linear MDPs of Zanette et al. (2020).

Exploration-Free Reinforcement Learning with Linear Function Approximation

Anonymous authors
Paper under double-blind review

Abstract

In the context of Markov Decision Processes (MDPs) with linear Bellman completeness, a generalization of linear MDPs, we reconsider the learning capabilities of a *greedy* algorithm. The motivation is that, when exploration is costly or dangerous, an exploration-free approach may be preferable to optimistic or randomized solutions. We show that, under a condition of sufficient diversity in the feature distribution, Least-Squares Value Iteration (LSVI) can achieve sublinear regret. Specifically, we show that the expected cumulative regret is at most $\tilde{\mathcal{O}}(H^3\sqrt{dK/\lambda_0})$, where K is the number of episodes, H is the task horizon, d is the dimension of the feature map and λ_0 is a measure of feature diversity. We empirically validate our theoretical findings on synthetic linear MDPs. Our analysis is a first step towards exploration-free reinforcement learning in MDPs with large state spaces.

1 INTRODUCTION

Reinforcement Learning (RL) is one of the most popular approaches to sequential decision making under uncertainty. In the last few years, RL in large state spaces has received a lot of attention both in theory (Long & Han, 2023) and practice, with applications ranging from robotics (Singh et al., 2022) to LLM finetuning (Ahmadian et al., 2024). One great potential of RL solutions, still largely untapped, is their intrinsically *adaptive* nature: RL agents, once deployed, can improve over time from interaction data. This requires a careful balancing of *exploitation* (taking decisions that are known to be good) and *exploration* (taking decisions that may be even better, but of which little is known).

This exploration-exploitation dilemma is well known in the RL literature since its beginnings (Sutton & Barto, 2018) and is the main subject of study of the bandit literature (Lattimore & Szepesvári, 2020) and of a good part of RL theory (Agarwal et al., 2019). All agree on this basic principle: that some form of exploration is necessary. A purely *greedy* agent can easily get stuck on a promising course of action, without ever discovering better but neglected alternatives. Some of the most popular exploration strategies are based on the *optimism in the face of uncertainty* principle (Lai & Robbins, 1985), of which (Azar et al., 2017) and (Jaksch et al., 2010) are notable applications to RL, *posterior sampling* (Thompson, 1933), like (Osband et al., 2013), or simple noise injection (Haarnoja et al., 2018).

In practice, however, there are several reasons to avoid exploration in favor of a greedy approach. In safety-critical applications, such as robotic (Brunke et al., 2022), explorative actions may be dangerous. In many cases, exploration for the sake of learning can also be considered unethical (Bird et al., 2016), some prominent examples being drug trials, predictive policing, lending, resume screening, and social media personalization. It is not hard to imagine that chatbots will incur in similar ethical issues (Følstad et al., 2021). Furthermore, explorative solutions are more expensive to implement, their behavior is less predictable, and their decisions less interpretable. Greedy approaches are not only favored for the aforementioned reasons, but often are also surprisingly effective in practice (e.g.,

Li et al., 2024). Hence, even if theory clearly shows the necessity of exploration, common sense may suggest otherwise in many real-world scenarios.

To reconcile theory and practice, Bastani et al. (2021), closely followed by Kannan et al. (2018), proposed to study special conditions under which exploration-free learning *is* possible. They did so within the framework of *linear contextual bandits* (Lattimore & Szepesvári, 2020, Chapter 19). In this model, at each timestep t, the agent observes a context X_t (e.g., data about the current user) and selects an action A_t (e.g., an item to recommend). The agent receives a reward that is *linear* in some context-action features. Clearly, some *structure* in the rewards (such as linearity) is necessary for exploration-free learning. If rewards of different actions are completely uncorrelated, active exploration is the only way to compare the value of different actions. On the other hand, if some structure is present, an action may reveal something about other actions, reducing or even removing the need for exploration. Indeed, Bastani et al. (2021) show that under sufficient *diversity* of contexts, exploration-free learning is possible in linear contextual bandits. In particular, they introduce a covariate-diversity assumption and prove that the regret of a simple greedy algorithm is sublinear. This does not mean that exploration is in general unnecessary for linear contextual bandits, but provides a possible characterization of tasks for which pure exploitation suffices.

Our purpose is to provide a similar characterization for Markov Decision Processes with structure, showing *when* exploration-free RL is possible. To leverage results from the linear contextual bandit literature, we examine MDPs with some kind of linear structure. These are commonly studied in the context of no-regret RL with linear function approximation. This line of work was pioneered by Jin et al. (2023), who first designed a no-regret algorithm for finite-horizon MDPs with linear rewards and transition probabilities, also known as low-rank MDPs (Yang & Wang, 2019). The algorithm is called LSVI-UCB and is based on the optimism principle. A follow-up work by Zanette et al. (2020) considers a more general class of "linear" MDPs where the class of linear action-value functions is closed under the Bellman optimality operator. This is the framework that we will adopt for our analysis, although we will use low-rank MDPs as numerical examples.[1] Nonlinear function approximation is also an active area of research (e.g., Jin et al., 2021). This is beyond the scope of this paper, but we believe that our analysis of linear MDPs is a necessary step in the study of exploration-free reinforcement learning in complex environments requiring general function approximation.

Our main contributions are as follows: we define a novel diversity condition, inspired by Bastani et al. (2021) and Kannan et al. (2018), for Markov Decision Processes with linear function approximation, and present new insights into how feature coverage affects the performance of exploration-free reinforcement learning algorithms. We prove that a greedy algorithm (LSVI) achieves sublinear cumulative regret with high probability when the diversity condition is satisfied. We also establish an *any-time* bound on the expected cumulative regret. Finally, we empirically validate our theoretical findings on synthetic linear MDPs.

The paper is structured as follows. In Section 2 we present all the necessary preliminaries for understanding and developing the concepts discussed in this work. We begin by introducing Markov Decision Processes (MDPs), followed by the specific case of MDPs that satisfy the linear Bellman completeness condition, which is the setting of this work. We also consider the special case of low-rank MDPs. Section 3 describes the analyzed algorithm, outlines the assumptions required for our analysis, and presents the theoretical results. Section 4 provides more details on the theoretical analysis, where we state the key lemmas used in the proof of the main theorem, followed by a detailed proof of the latter. Other proofs can be found in the Appendix. In Section 5, we discuss related works, while Section 6 focuses on the experiments conducted to empirically validate our theoretical results.

[1] A more intuitive generalization of low-rank MDPs is linear realizability of action-value functions. However, this has so far proven to be much more challenging to analyze (Weisz et al., 2023).

Table 1: Notation

SYMBOL	DESCRIPTION
$[n]$	$\{1, \ldots, n\}$
$\mathbb{I}$	Indicator function
$\lambda_{min}(A)$	Minimum eigenvalue of matrix A
$\langle x, y \rangle$	Inner product, $\langle x, y \rangle = \sum_i x_i y_i$
$\|\|x\|\|_p$	p-norm of vector x
$\|\|x\|\|_A^2$	$x^\top A x$

2 PRELIMINARIES

In this section, we provide the necessary background on Markov decision processes and the linearity assumption under which our work is conducted. Our notation is summarized in Table 1.

2.1 Markov Decision Processes

A finite-horizon Markov Decision Process (MDP, Puterman, 1994) is denoted by the tuple $M = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r, \mu)$, where $\mathcal{S}$ is the space of states, $\mathcal{A}$ is the space of actions, $H \in \mathbb{N}$ is the length of each episode, $\mathbb{P} = \{\mathbb{P}_h\}_{h=1}^H$ and $r = \{r_h\}_{h=1}^H$ are, respectively, the state transition probabilities and the reward functions. We assume that $\mathcal{S}$ is a measurable space and $\mathcal{A}$ has finite cardinality. For each step $h \in [H]$, $\mathbb{P}_h(\cdot|s, a)$ denotes the transition kernel over the next states if we choose action a in state s, and $r_h : \mathcal{S} \times \mathcal{A} \to [-1, 1]$ is the deterministic reward function. Finally μ is the starting-state probability distribution over $\mathcal{S}$.

An agent interacts with the MDP as follows: an initial state s_1 is drawn from μ, then at each step $h \in [H]$ the agent observes the state s_h, picks an action a_h and receives a reward $r_h(s_h, a_h)$. The MDP evolves into a new state s_{h+1} that is drawn from the transition kernel $\mathbb{P}_h(\cdot|s_h, a_h)$. The episode ends when state s_{H+1} is reached. A (deterministic) policy π of an agent is a function $\pi : \mathcal{S} \times [H] \to \mathcal{A}$, where $\pi(s, h)$ is the action that the agent takes in state s at the h-th step of the episode. We will abbreviate $\pi(s, h)$ as $\pi_h(s)$ in the following. For a policy π, for each $h \in [H]$, we can define the value function $V_h^\pi : \mathcal{S} \to \mathbb{R}$, which, given the current state at step h, returns the cumulative expected reward following policy π:

$$V_h^\pi(s) \coloneqq \mathbb{E}_\pi \left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}) \middle| s_h = s \right],$$

where $\mathbb{E}_\pi$ is short for $a_h \sim \pi(\cdot|s_h), s_{h+1} \sim \mathbb{P}(\cdot|s_h, a_h), \ldots, a_H \sim \pi(\cdot|s_H)$ conditional on π. We also define the action-value function $Q_h^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, which gives the expected value of cumulative rewards when the agent starts from a given state-action pair at the h-th step and follows policy π afterwards. We have:

$$Q_h^\pi(s, a) \coloneqq r_h(s, a) + \mathbb{E}_\pi \left[\sum_{h'=h+1}^{H} r_{h'}(s_{h'}, \pi_{h'}(s_{h'})) \middle| s_h = s, a_h = a \right],$$

for all $(s, a) \in \mathcal{S} \times \mathcal{A}, h \in [H]$.

Finally, we can define the occupancy measure of the policy π:

$$\rho_h^\pi(s) \coloneqq \mathbb{E}_{\pi, s_0 \sim \mu}[\mathbb{I}\{s_h = s\}].$$

There always exists an optimal deterministic policy π^* which gives the optimal value $V_h^*(s) = \sup_\pi V_h^\pi(s)$ for all $s \in \mathcal{S}$ and $h \in [H]$ (Puterman, 1994). Similarly, $Q_h^*(s, a) = Q_h^{\pi^*}(s, a)$.

In an episodic MDP, the agent aims to learn the optimal policy by interacting with the environment over a series of K episodes. For each $k \geq 1$, an initial state s_1^k is drawn from μ and the agent chooses policy π_k. The difference in values between $V_1^{\pi_k}(s_k)$ and $V_1^*(s_k)$ is the instantaneous regret, or suboptimality, of the agent at the k-th episode. Thus, after playing for K episodes, the total regret is

$$R(K) \coloneqq \sum_{k=1}^{K} \mathbb{E}_{s_k \sim \mu} \left[V_1^*(s_k) - V_1^{\pi_k}(s_k) \right].$$

We can also rewrite the total regret, by using a performance difference lemma (e.g., Proposition 29 from Papini et al. (2021a)), as follows:

$$R(K) \coloneqq \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{s_h \sim \rho_h^{\pi_k}} \left[\Delta_h(s_h, \pi_k(s_h)) \right], \tag{1}$$

where $\Delta_h(s,a) \coloneqq V_h^*(s) - Q_h^*(s,a)$ is the suboptimality gap.

2.2 Linear Bellman Completeness

We will consider a setting in which we have a set of features that satisfy the linear Bellman completeness condition, which we will refer to as linear MDPs for brevity. In this scenario we work with a feature map $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$. Let us first define the set of admissible parameters as:

$$\mathcal{W} = \{\mathbf{w} \in \mathbb{R}^d \text{ s.t. } |\langle \boldsymbol{\phi}(s,a), \mathbf{w} \rangle| \leq H \ \forall s \in \mathcal{S}, \forall a \in \mathcal{A}\}.$$

We restrict our analysis to MDPs equipped with a feature map that satisfies the following:

Assumption 2.1 (Linear Bellman completeness, Agarwal et al. (2021a))**.** *We say that the feature map $\boldsymbol{\phi}$ satisfies the linear Bellman completeness property if, for all $\boldsymbol{\theta} \in \mathcal{W}$ and $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, there exists $\mathbf{w} \in \mathcal{W}$ such that:*

$$\mathbf{w}^\top \boldsymbol{\phi}(s,a) = r(s,a) + \mathbb{E}_{s' \sim \mathbb{P}_h(s,a)} \max_{a'} \boldsymbol{\theta}^\top \boldsymbol{\phi}(s',a').$$

This condition implies that $Q_h^*(s,a)$ is linear in $\boldsymbol{\phi}$, i.e., there exists $\boldsymbol{\theta}_h^*$ such that $Q_h^*(s,a) = (\boldsymbol{\theta}_h^*)^\top \boldsymbol{\phi}(s,a)$ (Zanette et al., 2020, Lemma 6). This justifies the use of linear function approximation.

2.3 Low-Rank Markov Decision Processes

Although our theoretical results apply to general linear-Bellman-complete MDPs, we mention a particular case in which Assumption 2.1 holds, *low-rank* Markov Decision Processes (Jin et al., 2023). In this scenario, the transition kernel and the reward function are assumed to be linear w.r.t. known state-action features.

Formally, a Markov Decision Process defined as $M = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$, with a feature map $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, is considered a *low-rank MDP* (Yang & Wang, 2019; Jin et al., 2023) if, for each time step $h \in [H]$, there exist d signed measures $\boldsymbol{\rho}_h = (\rho^{(1)}, \ldots, \rho^{(d)})$ over the state space $\mathcal{S}$, and a vector $\boldsymbol{\theta}_h \in \mathbb{R}^d$, such that, for any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, the following holds:

$$\mathbb{P}_h(\cdot \mid s,a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{\rho}_h(\cdot) \rangle, \quad r_h(s,a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{\theta}_h \rangle.$$

A key characteristic of a low-rank MDP is that the action-value functions of *all* policies are linear with respect to the same feature map $\boldsymbol{\phi}$ (Jin et al., 2023, Proposition 2.3). It is easy to show that all low-rank MDPs are linear-Bellman-complete. The opposite is not true (Zanette et al., 2020).

3 GREEDY LEARNING

In this section, after reviewing the LSVI algorithm, we present our feature-diversity assumption and show how this is sufficient to achieve sublinear regret in an exploration-free manner.

3.1 Algorithm

The algorithm we consider in our work is Least-Square Value Iteration (LSVI, Bradtke & Barto, 1996), a simple *greedy* algorithm, based on value-iteration, which finds the optimal Q-function by iterative application of Bellman's optimality equation:

$$Q_h^*(s,a) = r_h(s,a) + \mathbb{E}_{s' \sim \mathbb{P}_h(\cdot|s,a)} \max_{a' \in \mathcal{A}} Q_{h+1}^*(s', a').$$

LSVI parametrizes $Q_h^*(s,a)$ by a linear form and approximates the optimality equation with a regularized least-squares problem in which we solve for $\mathbf{w}_h$. The algorithm solves the following program at each stage of each episode:

$$\mathbf{w}_h \leftarrow \operatorname*{argmin}_{\mathbf{w} \in \mathcal{W}} \sum_{\tau=1}^{k-1} [r_h(s_h^\tau, a_h^\tau) + \max_{a \in \mathcal{A}} Q_{h+1}(s_{h+1}^\tau, a) - \mathbf{w}^\top \boldsymbol{\phi}(s_h^\tau, a_h^\tau)]^2 + \lambda ||\mathbf{w}||^2.$$

Algorithm 1 LSVI

1: **for** episode $k = 1, \ldots, K$ **do**
2: Observe the initial state $s_1^k \sim \mu$
3: **for** step $h = H, \ldots, 1$ **do**
4: $\hat{\Sigma}_{k,h} = \sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau) \boldsymbol{\phi}(s_h^\tau, a_h^\tau)^\top + \lambda \cdot \mathbf{I}$
5: $\widetilde{\mathbf{w}}_h^k = \hat{\Sigma}_{k,h}^{-1} \sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau)[r_h(s_h^\tau, a_h^\tau) + \max_a Q_{h+1}^k(s_{h+1}^\tau, a)]$
6: $\widehat{\mathbf{w}}_h^k = \arg\min_{\mathbf{w} \in \mathcal{W}} \left\| \mathbf{w} - \widetilde{\mathbf{w}}_h^k \right\|_{\hat{\Sigma}_{k,h}}$
7: $Q_h^k = \langle \widehat{\mathbf{w}}_h^k, \boldsymbol{\phi}(\cdot, \cdot) \rangle$
8: **end for**
9: **for** step $h = 1, \ldots, H$ **do**
10: take action $a_h^k = \operatorname{argmax}_{a \in \mathcal{A}} Q_h^k(s_h^k, a)$ and observe s_{h+1}^k
11: **end for**
12: **end for**

At a high level, each episode involves two main passes through all time-steps. The first *backward* pass (lines 3-8) updates $\widehat{\mathbf{w}}_h^k$ and $\hat{\Sigma}_{k,h}$, that are, respectively, the parameters we are trying to estimate and the covariance matrix, which are used to construct the action-value function Q_h^k. In the second pass (lines 9-11), the greedy policy is executed: $a_h^k = \operatorname{argmax}_{a \in \mathcal{A}} Q_h^k(s_h^k, a)$, using the Q_h^k computed in the first pass. It's important to note that $Q_{H+1} \equiv 0$ since no reward is given after the H-th step. In the first episode ($k = 1$), the summations in lines 4 and 5 run from $\tau = 1$ to 0, meaning $\hat{\Sigma}_{1,h} = \lambda \cdot \mathbf{I}$ and $\widehat{\mathbf{w}}_h^1 = 0$. The inverse covariance matrix can be updated directly using Sherman-Morrison's formula for improved computational complexity. Line 6 is a projection step ensuring $\widehat{\mathbf{w}}_h^k \in \mathcal{W}$.[2]

3.2 Assumptions

We will now outline the assumptions necessary for our regret analysis. The first is a technical one on the parameter set:

Assumption 3.1. *$\mathcal{W}$ is a convex set. Moreover, there exists a constant ϕ_{max} such that $\|\boldsymbol{\phi}(s,a)\|_2 \leq \phi_{max}$ for all s, a, and a constant $w_{\max}$ such that $\|\mathbf{w}\|_2 \leq w_{\max}$ for all $\mathbf{w} \in \mathcal{W}$.*

[2] It is more common to directly clip the Q-function estimate in $[-H, H]$. However, for technical reasons, we need to preserve the linearity of the estimator.

The most important assumption is the following, inspired by conceptually similar conditions proposed by Bastani et al. (2021) and Kannan et al. (2018) for linear contextual bandits:

Assumption 3.2. ***(Covariate Diversity).*** *There exists a positive constant λ_0 such that, for each policy $\pi, \mathbf{w} \in \mathcal{W}$, and for each $h \in [H]$,*

$$\lambda_{\min}\left(\mathbb{E}_{s\sim\rho_h^\pi(s)}\left[\boldsymbol{\phi}(s,\pi(s))\boldsymbol{\phi}(s,\pi(s))^\top \mathbb{I}\{\langle\boldsymbol{\phi}(s,\pi(s)),\mathbf{w}\rangle \geq \max_{a\in\mathcal{A}}\langle\boldsymbol{\phi}(s,a),\mathbf{w}\rangle\}\right]\right) \geq \lambda_0.$$

Intuitively, the feature vectors witnessed by the agent in "sensible" rounds must cover the whole feature space. Fix a linear Q-function estimator. A round is "sensible" if the agent plays an action that would appear optimal according to the Q-function estimate. It must hold true for all deterministic policies the agent may play, all linear Q-function estimators, and separately for each episode's timestep. This is a joint property of the MDP and of the feature map. It is encouraged by feature maps showing great diversity across states, but also by strongly connected MDPs and starting-states distributions with a large support. The constant λ_0 is a measure of diversity. We expect exploration-free learning to be easier when λ_0 is larger.

A simple example where Assumption 3.2 holds is the following. For simplicity we consider two actions and $d = 1$, but similar constructions can be made for a generic number of actions and a larger feature dimension.

Proposition 3.3 (Noisy features)**.** *Let $|\mathcal{A}| = 2$ and $\boldsymbol{\phi}(s,a) = f(s,a) + \eta(a)$ for some function $f : \mathcal{S}\times\mathcal{A} \to [0, \sqrt{2}\sigma]$ and independent Gaussian noises $\eta(a) \sim \mathcal{N}(0,\sigma^2)$. Then Assumption 3.2 holds with $\lambda_0 \geq 0.2\sigma^2$.*

3.3 Regret of LSVI with Covariate Diversity

We now establish an upper bound on the cumulative regret of LSVI in the case of an MDP whose representation satisfies both the Assumption 2.1 and Assumption 3.2.

Theorem 3.4. *Under Assumptions 2.1, 3.1, and 3.2, with probability $1-\delta$, the cumulative regret of LSVI is at most:*

$$R(K) = \mathcal{O}\left(H^3\sqrt{\frac{dK}{\lambda_0}}\log(K/\delta)\right).$$

Notice that Algorithm 1 is not parametric in the failure probability δ. By setting this free parameter to $\delta = 1/\sqrt{K}$, by a standard argument, we obtain an upper bound on the *expected* regret, where the extra expectation is over the random sequence of (deterministic) policies played by LSVI.

Corollary 3.5. *Under the same assumptions as Theorem 3.4, the expected cumulative regret of LSVI is at most:*

$$\mathbb{E}\left[R(K)\right] = \mathcal{O}\left(H^3\sqrt{\frac{dK}{\lambda_0}}\log(K)\right).$$

The result is still *any-time*, that is, the algorithm does not need to know the number of episodes K in advance.

Our regret upper bounds, scaling with $\sqrt{d}$, seem to contradict existing $\Omega(d\sqrt{K})$ lower bounds (cf. Zanette et al. (2020), Theorem 2). This may actually be possible under the non-standard Assumption 3.2. Anyway, notice that $\lambda_0 \leq 1/d$, the minimum eigenvalue of the covariance matrix of a D-optimal design (Lattimore et al., 2020). Hence, linear dependence on the dimension of the feature map is not avoided. If $\lambda_0 \simeq 1/d$, LSVI with covariate diversity has a better dependence than LSVI-UCB ($d\sqrt{d}$) and matches that of the computationally inefficient ELEANOR (Zanette et al., 2020). This is possible thanks to the linearity of the Q-function estimates, while LSVI-UCB incurs an extra $\sqrt{d}$ factor due to its nonlinear exploration bonuses.

3.4 Misspecification

Our results extend to the case where the MDP is only *approximately* linear. In particular, we consider the notion of *low inherent Bellman error* introduced by Zanette et al. (2020):

Assumption 3.6.

$$\sup_{\mathbf{w}' \in \mathcal{W}} \inf_{\mathbf{w} \in \mathcal{W}} \sup_{s \in \mathcal{S}, a \in \mathcal{A}} \left| \langle \boldsymbol{\phi}(s,a), \mathbf{w} \rangle - r_h(s,a) - \mathbb{E}_{s' \sim \mathbb{P}_h(\cdot|s,a)} \left[\max_{a'} \langle \boldsymbol{\phi}(s',a'), \mathbf{w}' \rangle \right] \right| \leq \zeta.$$

The constant ζ measures the level of misspecification, and the linear Bellman completeness case we considered so far corresponds to $\zeta = 0$, no misspecification. The optimal action-value function is no longer linear, but is well approximated by a linear function (Zanette et al., 2020, Lemma 6). Our results generalize well to this misspecified setting.

Theorem 3.7. *If Assumptions 3.6 and 3.2 are satisfied, with probability $1-\delta$, the cumulative regret of LSVI is at most:*

$$R(K) = \widetilde{\mathcal{O}}\left(H^3 \sqrt{\frac{dK}{\lambda_0}} + H^2 \zeta \frac{K}{\sqrt{\lambda_0}} + H^2 \zeta K \right).$$

With misspecification, the linear term in K is inevitable (Zanette et al., 2020), but is controlled by ζ, which is supposed to be very small. In fact, our result seems to violate a fundamental $\Omega(\zeta\sqrt{d}K)$ lower bound (Lattimore et al., 2020). Again, this is not the case since $\lambda_0 \leq 1/d$, making the second term in the regret $\widetilde{O}(H^2\zeta\sqrt{d}K)$.

4 ANALYSIS

In this section, we prove our main result, Theorem 3.4. We first provide two fundamental lemmas, whose proofs are deferred to Appendix A and B.

The first lemma provides an upper bound on the difference between the estimated Q-function at episode k and step h, and the actual optimal Q-function.

Lemma 4.1. *Assume $\lambda_{\min}(\hat{\Sigma}_{k,h}) \geq \lambda_k$ for all $k \geq 1$ and $h \in [H]$. Under Assumptions 2.1 and 3.1, with probability $1-\delta/2$, for all $k \geq 1, h \in [H], s \in \mathcal{S}, a \in \mathcal{A}$:*

$$|\widehat{Q}_h^k(s,a) - Q_h^*(s,a)| \leq (H-h) \left(\left(\sqrt{\beta_k(\delta)} \right) \frac{\phi_{\max}}{\sqrt{\lambda_k}} \right),$$

where

$$\sqrt{\beta_k(\delta)} \coloneqq H\sqrt{A+B+C} + 1 + w_{\max},$$

and $A \coloneqq d \ln\left(1 + \frac{\phi_{\max}^2 k}{d}\right)$, $B \coloneqq d\ln(w_{\max}^2 \phi_{\max}^2 k)$, $C \coloneqq \ln(2H\delta^{-1})$.

Next, we show that the minimum eigenvalue of the sample covariance matrix at time step h until episode k, $\lambda_{\min}(\hat{\Sigma}_{k,h})$, grows linearly with k. This will guarantee the convergence of our regression estimate.

Lemma 4.2. *Given Assumptions 3.1 and 3.2, the following holds for the minimum eigenvalue of the empirical covariance matrix for each $h \in [H]$ and for each $k \geq 1$:*

$$\mathbb{P}\left[\lambda_{min}(\hat{\Sigma}_{k,h}) \geq \lambda + \lambda_0 k - 8\phi_{max}^2 \sqrt{k \log(4dk/\delta)} \right] \geq 1 - \frac{\delta}{2}$$

Proof of Theorem 3.4. Given the representation of regret from Equation (1):

$$
\begin{aligned}
R(K) &= \sum_{k=1}^{K} \mathbb{E}_{s\sim\rho_h^{\pi_k}}\left[\sum_{h=1}^{H} \Delta_h(s,\pi_k(s))\right] \\
&= \sum_{k=1}^{K} \mathbb{E}_{s\sim\rho_h^{\pi_k}}\left[\sum_{h=1}^{H} Q_h^*(s,\pi_h^*(s)) - Q_h^*(s,\pi_h^k(s))\right] \\
&= \sum_{k=1}^{K} \mathbb{E}_{s\sim\rho_h^{\pi_k}}\left[\sum_{h=1}^{H} Q_h^*(s,\pi_h^*(s)) - \widehat{Q}_h^k(s,\pi_h^k(s)) + \widehat{Q}_h^k(s,\pi_h^k(s)) - Q_h^*(s,\pi_h^k(s))\right] \\
&\le \sum_{k=1}^{K} \mathbb{E}_{s\sim\rho_h^{\pi_k}}\left[\sum_{h=1}^{H} Q_h^*(s,\pi_h^*(s)) - \widehat{Q}_h^k(s,\pi_h^*(s)) + \widehat{Q}_h^k(s,\pi_h^k(s)) - Q_h^*(s,\pi_h^k(s))\right] \\
&\le \sum_{k=1}^{K} \mathbb{E}_{s\sim\rho_h^{\pi_k}}\left[\sum_{h=1}^{H} 2\sup_{a\in\mathcal{A}}\left|\widehat{Q}_h^k(s,a) - Q_h^*(s,a)\right|\right] \\
&\le \sum_{k=1}^{K}\sum_{h=1}^{H} 2(H-h)\sqrt{\beta_k(\delta)}\frac{\phi_{\max}}{\sqrt{\lambda_k}} \qquad \left(\text{w.p. } 1-\frac{\delta}{2}\right) \\
&\le \sum_{k=1}^{K} H(H+1)\sqrt{\beta_k(\delta)}\frac{\phi_{\max}}{\sqrt{\lambda_k}} \\
&= \mathcal{O}\left(H^3\sqrt{d}\log(K/\delta)\sum_{k=1}^{K}\frac{1}{\sqrt{\lambda_0 k}}\right) \qquad \left(\text{w.p. } 1-\frac{\delta}{2}\right) \\
&= \mathcal{O}\left(H^3\sqrt{\frac{dK}{\lambda_0}}\log(K/\delta)\right),
\end{aligned}
$$

The first inequality is by definition of the greedy algorithm (Alg. 1, line 10). The third inequality follows from Lemma 4.1 with a choice of λ_k provided by Lemma 4.2, and the final inequality is an elementary upper bound on the sum $\sum_{h=1}^{H} 2(H-h)$. The last two equalities result from bounding $\beta_k(\delta)$ by $\beta_K(\delta)$, and then substituting $\sqrt{\beta_k(\delta)} = \mathcal{O}(H\sqrt{d\log(k/\delta)})$ as defined in Lemma 4.1. Additionally, we use the fact that $1/(\sqrt{k} - \mathcal{O}(\sqrt{k\log(k)})) = \mathcal{O}(\sqrt{\log(k)/k})$ and this derives from the definition of λ_k in Lemma 4.2. □

5 RELATED WORKS

Our work is mainly inspired by the one of Bastani et al. (2021) on linear contextual bandits. They first introduced a covariate-diversity assumption that allows a greedy algorithm to achieve sublinear regret. In the same paper, they also proposed a *greedy-first* algorithm that operates greedily until it detects that convergence to the optimal policy is unlikely, at which point it begins exploring. This was shown to outperform existing exploration-based bandit algorithms. A similar analysis of greedy algorithms, using a slightly different diversity condition, was carried out by Kannan et al. (2018). To the best of our knowledge, we are the first to extend this kind of analysis to MDPs.

With similar motivations, but a radically different approach, Saleh et al. (2022) studied noise-free reinforcement learning in MDPs with Lipschitz-continuous transition models. They proposed a regularized policy gradient approach called truly deterministic policy optimization.[3] They proved

[3]The name is a reference to the more popular *deterministic policy gradient* algorithms (Silver et al., 2014; Lillicrap et al., 2016; Fujimoto et al., 2018). These optimize a deterministic parametric policy using data collected by a stochastic counterpart obtained by noise injection. For this reason, despite the name, they cannot be considered exploration-free. The reasons for deploying deterministic policies are similar to ours, but are only applied to the final product of learning and not to the learning process itself. See Montenegro et al. (2024) for a recent investigation of this approach.

monotonic improvement guarantees, but did not study sample complexity nor regret. Their experiments show promising results in robotics applications, but are essentially incomparable to ours. Also, because of the (deterministic) policy regularization, their algorithm may not be considered greedy.

5.1 Diversity Conditions

Besides (Bastani et al., 2021) and (Kannan et al., 2018), which serve as the foundation for our work, several papers have adopted covariate-diversity assumptions in linear contextual bandits for diverse reasons, often as mere technical assumptions (Foster et al., 2019; Chatterji et al., 2020; Ghosh et al., 2021; Hao et al., 2020; Wu et al., 2020; Tirinzoni et al., 2022), sometimes, with a representation learning perspective, as a characterization of "good" feature maps (Papini et al., 2021b; Tirinzoni et al., 2022; 2023). See (Papini et al., 2021b) for a discussion and comparison of the different conditions.

For linear MDPs, Papini et al. (2021a) proposed a diversity condition, called UNISOFT, under which LSVI-UCB and other optimistic algorithms achieve *constant* regret (under a minimum-gap assumption). In our notation, they require:

$$\lambda_{\min}\left(\mathbb{E}_{s\sim\rho_h^{\pi^*}(s)}\left[\boldsymbol{\phi}(s,\pi^*(s))\boldsymbol{\phi}(s,\pi^*(s))^\top\right]\right) \geq \lambda^*. \tag{2}$$

It is sufficient to observe that $\langle\boldsymbol{\phi}(s,\pi^*(s)),\mathbf{w}^*\rangle \geq \max_a\langle\boldsymbol{\phi}(s,a),\mathbf{w}^*\rangle$, where $\mathbf{w}^*$ is the linear parameter of Q^*, to see that Assumption 3.2 implies UNISOFT and $\lambda_0 \leq \lambda^*$. Intuitively, UNISOFT requires optimal trajectories to be informative, while we ask the same of all trajectories that are optimal *according to some linear estimate*. In fact, our design of Assumption 3.2 was partly inspired by UNISOFT.

In the theory of policy gradient algorithms, *concentrability coefficients* (Agarwal et al., 2021b) play a similar role than our covariate-diversity assumption: by assuming that the starting-state distribution covers well the subset of the state space visited by the optimal policy, they remove part of the challenges of exploration. This allows to study policy gradient algorithms with simple exploration strategies (e.g., noise injection) from the perspective of stochastic optimization. The algorithms considered in this line of work always employ stochastic policies, at least for data collection. Our covariate diversity assumption is also reminiscent of some *coverage ratios* used in offline RL (Uehara & Sun, 2022) with linear function approximation and of some notions of coverability (Xie et al., 2023).

5.2 Safe Exploration

A related body of literature focuses on developing reinforcement learning algorithms that prioritize controlled exploration for ethical and safety reasons, moving away from conventional exploratory methods. This challenge is well outlined in Amodei et al. (2016), which identifies several AI safety problems. These include issues like "avoiding side effects" and "reward hacking," where agents can inadvertently perform harmful actions due to poorly designed objective functions, but also concerns regarding undesirable behavior during the learning process, a problem known as *safe exploration*.

The concept of safe exploration was first introduced by Moldovan & Abbeel (2012), who presents an algorithm for safe but potentially suboptimal exploration in Markov Decision Processes (MDPs). A key contribution of their work is a formal definition of safety that focuses on maintaining ergodicity with a controlled probability. Although the problem is NP-hard, the authors propose an approximation scheme that balances safety and performance.

A natural way to ensure safe exploration is by adding constraints to the MDP (Altman, 2021) and enforcing them during the learning process. A notable example is the Constrained Policy Optimization (CPO) algorithm by Achiam et al. (2017), that ensures near-satisfaction of safety constraints at each iteration.

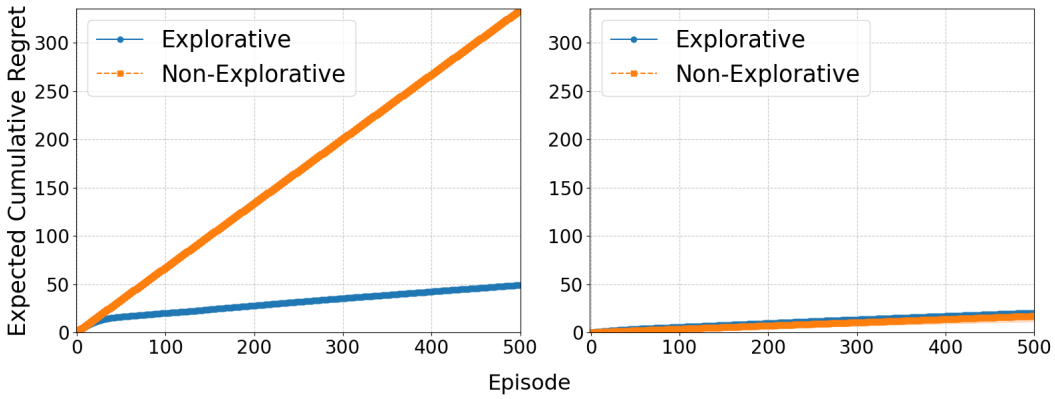

Figure 1: Expected cumulative regret of LSVI (without exploration) and LSVI-UCB (with exploration), with (*right*) and without (*left*) covariate diversity.

Safe exploration can also be defined in terms of stability, as in (Berkenkamp et al., 2017). A model-based predictive approach allows the agent to avoid exploratory actions that may lead to irrecoverable states.

Finally, with an appropriately designed reward function, safety during the learning process can be ensured by enforcing monotonic performance improvement (Papini et al., 2022).

6 EXPERIMENTS

In this section, we present the experimental results obtained from synthetic low-rank MDPs (as defined in Section 2.3). For each scenario, we evaluate the performance of LSVI and of an optimistic algorithm, LSVI-UCB (Jin et al., 2023).

Synthetic Problems. We define a randomly generated low-rank MDP with two distinct realizable linear representations, both obtained following Example 2.2 in Jin et al. (2023), each exhibiting different characteristics. The first representation satisfies covariate diversity by generating random features (cf. Proposition 3.3), while the second is specifically constructed to violate the diversity assumption, using orthogonal features to simulate an MDP without any correlation between actions (a tabular MDP). The purpose of these experiments is to demonstrate the varying behavior of LSVI-UCB and LSVI across different MDPs and to highlight the impact of covariate diversity, comparing the cases where it is satisfied and where it is not. We conduct our experiments in a setting where $H = 3$, $d = 10$ and $K = 500$. Each experiment is replicated 30 times under these same parameters. We plot cumulative regret normalized by V^*, averaged over the independent runs, with shaded areas corresponding to one standard deviation.

Covariate vs Non-Covariate Diversity. We construct two different environments with randomly generated parameters and compare, normalizing both with respect to their V^*, the expected cumulative regret obtained when using a representation that satisfies covariate diversity against one that does not. In the left image of Fig. 1, it is evident that the absence of covariate diversity causes the expected cumulative regret to increase linearly with LSVI (exploration-free). Conversely, the right image, depicting a setting satisfying covariate diversity, shows sub-linear curves for both LSVI (exploration-free) and LSVI-UCB (optimistic). Fig. 2 shows a close-up of the results under covariate diversity. Additionally, in Fig. 3, we show the average over 100 different MDPs with covariate diversity for both LSVI (non-explorative) and LSVI-UCB (explorative).

As shown by the experimental results, our findings align with the theory. Specifically, in both Fig. 1 and Fig. 3, we observe that the presence of covariate diversity makes exploration-free learning feasible. In Fig. 3, we average the (normalized) performance across 100 randomly generated MDPs that satisfy covariate diversity, with each parameter sampled from a uniform distribution between 0

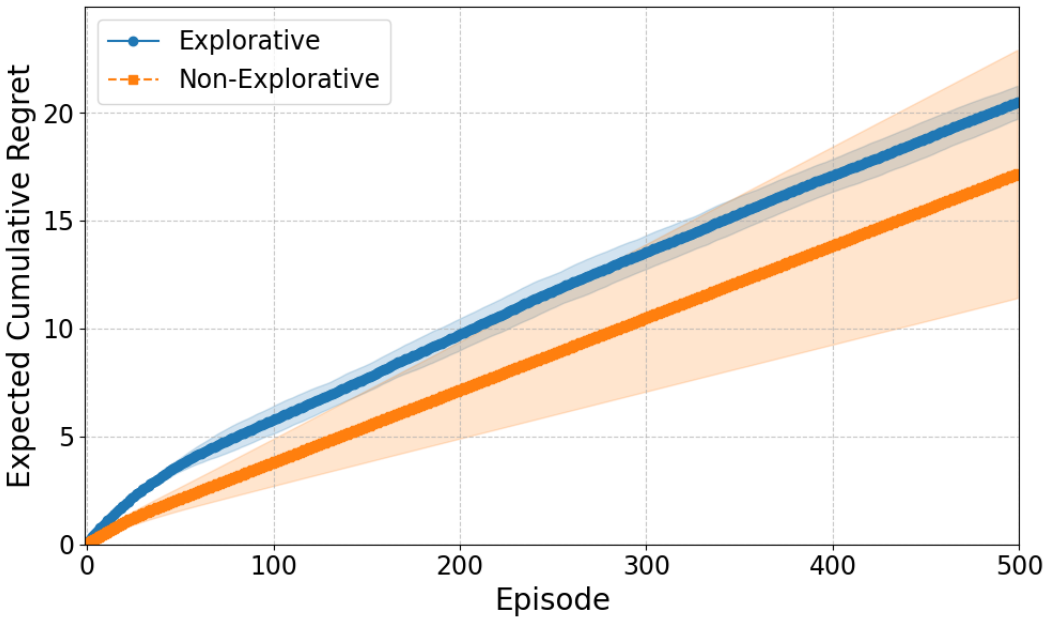

Figure 2: Expected cumulative regret of LSVI (without exploration) and LSVI-UCB (with exploration), with covariate diversity.

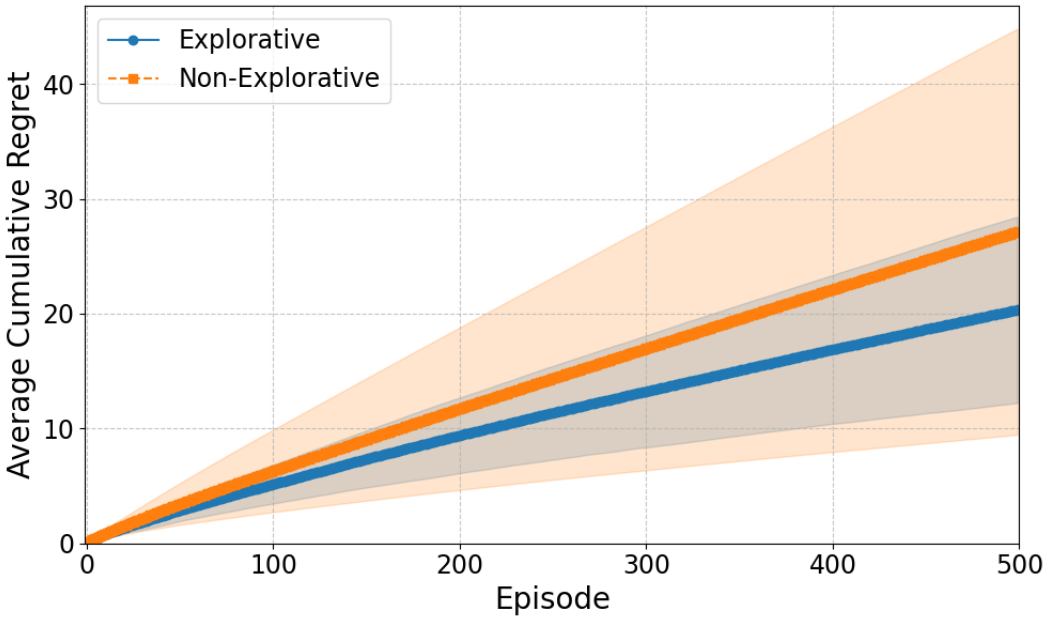

Figure 3: Average cumulative regret of LSVI (without exploration) and LSVI-UCB (with exploration), with covariate diversity over 100 MDP.

and 1, using parameters within this range can result in some MDPs having very small suboptimality gaps, which makes the learning process harder. This variability is reflected in the large standard deviation of Figure 3. Additionally, some of them may have a very small λ_0, resulting in weak covariate diversity. Unfortunately, this condition is difficult to measure, as λ_0 is defined for every parameter w and policy π.

7 CONCLUSION

In this paper, we have proven that in the setting of Markov Decision Processes, under the assumption of *linear Bellman completeness*, LSVI, a greedy algorithm, can achieve *sub-linear regret* if there is sufficient *diversity* in the feature distribution, as defined by our proposed diversity condition. This eliminates the need for explicit exploration for the agent to learn the optimal policy. Experimental results are coherent with the theory.

Our results on linear function approximation pave the way for exploration-free RL in MDPs with structure. Future work should focus on nonlinear function approximation in order to scale to complex control problems. Some questions remain even in the linear realm. Is exploration-free learning possible in Q^π-realizable MDPs (Weisz et al., 2023)? Since our covariate diversity condition is a special case of UNISOFT (Papini et al., 2021a), it might be possible for LSVI to achieve *constant* regret like LSVI-UCB under a minimum-gap assumption. From the perspective of representation learning, our notion of feature diversity could be encouraged by some form of spectral regularization as proposed by Tirinzoni et al. (2022) for UNISOFT. Our approach is inherently value-based, but similar ideas could be applied to policy-based RL to reconcile theory with practical "deterministic policy gradient" algorithms such as the one proposed by (Saleh et al., 2022). Finally, we may try to develop a greedy-first exploration algorithm for MDPs following the example of Bastani et al. (2021) on linear contextual bandits.

References

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 22–31. PMLR, 2017.

Alekh Agarwal, Nan Jiang, Sham M Kakade, and Wen Sun. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 32:96, 2019.

Alekh Agarwal, Nan Jiang, Sham M. Kakade, and Wen Sun. *Reinforcement Learning: Theory and Algorithms*. 2021a.

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *J. Mach. Learn. Res.*, 22: 98:1–98:76, 2021b.

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce-style optimization for learning from human feedback in llms. In *ACL (1)*, pp. 12248–12267. Association for Computational Linguistics, 2024.

Eitan Altman. *Constrained Markov decision processes*. Routledge, 2021.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul F. Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *CoRR*, abs/1606.06565, 2016.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pp. 263–272. PMLR, 2017.

Hamsa Bastani, Mohsen Bayati, and Khashayar Khosravi. Mostly exploration-free algorithms for contextual bandits. *Manag. Sci.*, 67(3):1329–1349, 2021.

Felix Berkenkamp, Matteo Turchetta, Angela P. Schoellig, and Andreas Krause. Safe model-based reinforcement learning with stability guarantees. In *NIPS*, pp. 908–918, 2017.

Sarah Bird, Solon Barocas, Kate Crawford, Fernando Diaz, and Hanna Wallach. Exploring or exploiting? social and ethical implications of autonomous experimentation in ai. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 2016.

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

Lukas Brunke, Melissa Greeff, Adam W. Hall, Zhaocong Yuan, Siqi Zhou, Jacopo Panerati, and Angela P. Schoellig. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annu. Rev. Control. Robotics Auton. Syst.*, 5:411–444, 2022.

Niladri S. Chatterji, Vidya Muthukumar, and Peter L. Bartlett. OSOM: A simultaneously optimal algorithm for multi-armed and linear contextual bandits. In *AISTATS*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1844–1854. PMLR, 2020.

Asbjørn Følstad, Theo Araujo, Effie Lai-Chong Law, Petter Bae Brandtzaeg, Symeon Papadopoulos, Lea Reis, Marcos Baez, Guy Laban, Patrick McAllister, Carolin Ischen, et al. Future directions for chatbot research: an interdisciplinary research agenda. *Computing*, 103(12):2915–2942, 2021.

Dylan J. Foster, Akshay Krishnamurthy, and Haipeng Luo. Model selection for contextual bandits. In *NeurIPS*, pp. 14714–14725, 2019.

Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1582–1591. PMLR, 2018.

Avishek Ghosh, Abishek Sankararaman, and Kannan Ramchandran. Problem-complexity adaptive model selection for stochastic linear bandits. In *AISTATS*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1396–1404. PMLR, 2021.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1856–1865. PMLR, 2018.

Botao Hao, Tor Lattimore, and Csaba Szepesvári. Adaptive exploration in linear contextual bandit. In *AISTATS*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3536–3545. PMLR, 2020.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *J. Mach. Learn. Res.*, 11:1563–1600, 2010.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of RL problems, and sample-efficient algorithms. In *NeurIPS*, pp. 13406–13418, 2021.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I. Jordan. Provably efficient reinforcement learning with linear function approximation. *Math. Oper. Res.*, 48(3):1496–1521, 2023.

Sampath Kannan, Jamie Morgenstern, Aaron Roth, Bo Waggoner, and Zhiwei Steven Wu. A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. In *NeurIPS*, pp. 2231–2241, 2018.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Tor Lattimore, Csaba Szepesvári, and Gellért Weisz. Learning with good feature representations in bandits and in RL with a generative model. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5662–5670. PMLR, 2020.

Zaile Li, Weiwei Fan, and L Jeff Hong. The (surprising) sample optimality of greedy procedures for large-scale ranking and selection. *Management Science*, 2024.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *ICLR (Poster)*, 2016.

Zhishuai Liu and Pan Xu. Distributionally robust off-dynamics reinforcement learning: Provable efficiency with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, 2024.

Jihao Long and Jiequn Han. Reinforcement learning with function approximation: From linear to nonlinear. *CoRR*, abs/2302.09703, 2023.

Teodor Mihai Moldovan and Pieter Abbeel. Safe exploration in markov decision processes. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress, 2012.

Alessandro Montenegro, Marco Mussi, Alberto Maria Metelli, and Matteo Papini. Learning optimal deterministic policies with stochastic policy gradients. In *ICML*. OpenReview.net, 2024.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. In *NIPS*, pp. 3003–3011, 2013.

Matteo Papini, Andrea Tirinzoni, Aldo Pacchiano, Marcello Restelli, Alessandro Lazaric, and Matteo Pirotta. Reinforcement learning in linear mdps: Constant regret and representation selection. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 16371–16383, 2021a.

Matteo Papini, Andrea Tirinzoni, Marcello Restelli, Alessandro Lazaric, and Matteo Pirotta. Leveraging good representations in linear contextual bandits. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8371–8380. PMLR, 2021b.

Matteo Papini, Matteo Pirotta, and Marcello Restelli. Smoothing policies and safe policy gradients. *Mach. Learn.*, 111(11):4081–4137, 2022.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Statistics. Wiley, 1994.

Ehsan Saleh, Saba Ghaffari, Timothy Bretl, and Matthew West. Truly deterministic policy optimization. In *NeurIPS*, 2022.

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin A. Riedmiller. Deterministic policy gradient algorithms. In *ICML*, volume 32 of *JMLR Workshop and Conference Proceedings*, pp. 387–395. JMLR.org, 2014.

Bharat Singh, Rajesh Kumar, and Vinay Pratap Singh. Reinforcement learning in robotic applications: a comprehensive survey. *Artif. Intell. Rev.*, 55(2):945–990, 2022.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

Andrea Tirinzoni, Matteo Papini, Ahmed Touati, Alessandro Lazaric, and Matteo Pirotta. Scalable representation learning in linear contextual bandits with constant regret guarantees. In *NeurIPS*, 2022.

Andrea Tirinzoni, Matteo Pirotta, and Alessandro Lazaric. On the complexity of representation learning in contextual linear bandits. In *AISTATS*, volume 206 of *Proceedings of Machine Learning Research*, pp. 7871–7896. PMLR, 2023.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, August 2011. ISSN 1615-3383. DOI: 10.1007/s10208-011-9099-z.

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. In *ICLR*. OpenReview.net, 2022.

Gellért Weisz, András György, and Csaba Szepesvári. Online RL in linearly q^π-realizable mdps is as easy as in linear mdps if you learn what to ignore. In *NeurIPS*, 2023.

Weiqiang Wu, Jing Yang, and Cong Shen. Stochastic linear contextual bandits with diverse contexts. In *AISTATS*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2392–2401. PMLR, 2020.

Tengyang Xie, Dylan J. Foster, Yu Bai, Nan Jiang, and Sham M. Kakade. The role of coverage in online reinforcement learning. In *ICLR*. OpenReview.net, 2023.

Lin Yang and Mengdi Wang. Sample-optimal parametric q-learning using linearly additive features. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6995–7004. PMLR, 2019.

Andrea Zanette, Alessandro Lazaric, Mykel J. Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10978–10989. PMLR, 2020.

Supplementary Materials

The following content was not necessarily subject to peer review.

A Proof of Lemma 4.1

Lemma 4.1. *Assume $\lambda_{\min}(\hat{\Sigma}_{k,h}) \geq \lambda_k$ for all $k \geq 1$ and $h \in [H]$. Under Assumptions 2.1 and 3.1, with probability $1-\delta/2$, for all $k \geq 1, h \in [H], s \in \mathcal{S}, a \in \mathcal{A}$:*

$$|\widehat{Q}_h^k(s,a) - Q_h^*(s,a)| \leq (H-h)\left(\left(\sqrt{\beta_k(\delta)}\right)\frac{\phi_{\max}}{\sqrt{\lambda_k}}\right),$$

where

$$\sqrt{\beta_k(\delta)} \coloneqq H\sqrt{A+B+C} + 1 + w_{\max},$$

and $A \coloneqq d\ln\left(1+\frac{\phi_{\max}^2 k}{d}\right)$, $B \coloneqq d\ln(w_{\max}^2\phi_{\max}^2 k)$, $C \coloneqq \ln(2H\delta^{-1})$.

We shall prove a more general version of Lemma 4.1 that takes misspecification into account.

Let the Bellman Error be defined as:

$$\mathcal{L}_h(\mathbf{w}; s, a, \mathbf{w}') \coloneqq \left|\langle \boldsymbol{\phi}(s,a), \mathbf{w}\rangle - r_h(s,a) - \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s,a)}\left[\max_{a'}\langle \boldsymbol{\phi}(s',a'), \mathbf{w}'\rangle\right]\right|. \tag{3}$$

Assumption 3.6 can be rephrased as follows.

Assumption A.1 (ζ-Approximate Linear Bellman Completeness, Zanette et al. (2020))**.** *For all $h \in [H]$ and $\mathbf{w}' \in \mathcal{W}$, there exists a $\mathbf{w} \in \mathcal{W}$ such that $\sup_{s\in\mathcal{S},a\in\mathcal{A}} \mathcal{L}_h(\mathbf{w}; s, a, \mathbf{w}') \leq \zeta$.*

We will denote:

$$\mathcal{T}_h(\mathbf{w}') \coloneqq \arg\min_{\mathbf{w}\in\mathcal{W}} \sup_{s\in\mathcal{S},a\in\mathcal{A}} \mathcal{L}_h(\mathbf{w}; s, a, \mathbf{w}'). \tag{4}$$

Clearly, by definition, for all $s \in \mathcal{S}, a \in \mathcal{A}, h \in [H]$:

$$\mathcal{L}_h(\mathcal{T}_h(\mathbf{w}'); s, a, \mathbf{w}') \leq \zeta. \tag{5}$$

Let $\mathcal{Q} = \{\langle \boldsymbol{\phi}(\cdot,\cdot), \mathbf{w}\rangle \text{ s.t. } \mathbf{w} \in \mathcal{W}\}$ and $\mathcal{V} = \{\max_{a\in\mathcal{A}} Q(\cdot, a) \text{ s.t. } Q \in \mathcal{Q}\}$. Let $\widehat{Q}_h^k(s,a) = \langle \boldsymbol{\phi}(s,a), \widehat{\mathbf{w}}_h^k\rangle$ and $\widehat{V}_h^k(s) = \max_{a\in\mathcal{A}}\langle \boldsymbol{\phi}(s,a), \widehat{\mathbf{w}}_h^k\rangle$. Clearly $\widehat{Q}_h^k \in \mathcal{Q}$ and $\widehat{V}_h^k \in \mathcal{V}$ for all k, h.

Proposition A.2 (Lemma 3 by Zanette et al. (2020)[4])**.** *Fix $h \in [H]$. If $\lambda = 1$, with probability $1-\delta$, for all $V \in \mathcal{V}$:*

$$\left\|\sum_{t=1}^{k-1} \boldsymbol{\phi}(s_h^t, a_h^t)\left(V(s_{h+1}^t) - \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_h^t,a_h^t)}\left[V(s')\right]\right)\right\|_{\Sigma_{t,h}^{-1}} \leq H\sqrt{d\ln\left(1+\frac{\phi_{\max}^2 k}{d}\right) + d\ln(w_{\max}^2\phi_{\max}^2 k) + \ln(\delta^{-1}) + 1}.$$

Proposition A.3 (Lemma 8 by Zanette et al. (2020))**.** *Let $a_1, \ldots, a_k \in \mathbb{R}^d$ and $b_1, \ldots, b_k \in \mathbb{R}$ such that $|b_t| \leq \epsilon$ for all $t \in [k]$. Let $\Sigma = \sum_{t=1}^k a_t a_t^\top + \lambda I$. Then, for any $k > 1$ and $\lambda \geq 0$:*

$$\left\|\sum_{t=1}^k a_t b_t\right\|_{\Sigma^{-1}}^2 \leq k\epsilon^2. \tag{6}$$

[4]The extra H factor is due to the fact that we assume value functions to be in $[0, H]$ rather than in $[0, 1]$.

Lemma A.4. *Under Assumption A.1, if $\lambda = 1$, with probability $1-\delta$, for all $k \geq 1$ and $h \in [H]$:*

$$\left\|\widehat{\mathbf{w}}_h^k - \mathcal{T}_h(\widehat{\mathbf{w}}_{h+1}^k)\right\|_{\hat{\Sigma}_{k,h}} \leq \sqrt{\beta_k(\delta)} + \sqrt{k}\zeta,$$

where $\sqrt{\beta_k(\delta)} \coloneqq H\sqrt{d\ln\left(1+\frac{\phi_{\max}^2 k}{d}\right) + d\ln(w_{\max}^2\phi_{\max}^2 k) + \ln(H\delta^{-1})} + 1 + w_{\max}$.

Proof. First, we show that $\left\|\widehat{\mathbf{w}}_h^k - \mathcal{T}_h(\widehat{\mathbf{w}}_{h+1}^k)\right\|_{\hat{\Sigma}_{k,h}} \leq \left\|\widetilde{\mathbf{w}}_h^k - \mathcal{T}_h(\widehat{\mathbf{w}}_{h+1}^k)\right\|_{\hat{\Sigma}_{k,h}}$, that is, the projection step can only bring the estimated parameter closer to its target. Fix k, h and let $\mathrm{Proj}(\mathbf{w}) \coloneqq \arg\max_{\mathbf{w}' \in \mathcal{W}} \|\mathbf{w} - \mathbf{w}'\|_{\hat{\Sigma}_{k,h}}$. Since $\mathcal{T}_h(\widehat{\mathbf{w}}_{h+1}^k) \in \mathcal{W}$ by definition, $\mathrm{Proj}(\mathcal{T}_h(\widehat{\mathbf{w}}_{h+1}^k)) = \mathcal{T}_h(\widehat{\mathbf{w}}_{h+1}^k)$. Then:

$$\left\|\widehat{\mathbf{w}}_h^k - \mathcal{T}_h(\widehat{\mathbf{w}}_{h+1}^k)\right\|_{\hat{\Sigma}_{k,h}} = \left\|\mathrm{Proj}(\widetilde{\mathbf{w}}_h^k) - \mathrm{Proj}(\mathcal{T}_h(\widehat{\mathbf{w}}_{h+1}^k))\right\| \leq \left\|\widetilde{\mathbf{w}}_h^k - \mathcal{T}_h(\widehat{\mathbf{w}}_{h+1}^k)\right\|_{\hat{\Sigma}_{k,h}}, \tag{7}$$

where the last inequality is by contractivity of metric projections onto convex sets ($\mathcal{W}$ is convex by Asm. 3.1). We then proceed to upper bound $\left\|\widetilde{\mathbf{w}}_h^k - \mathcal{T}_h(\widehat{\mathbf{w}}_{h+1}^k)\right\|_{\hat{\Sigma}_{k,h}}$. First notice that:

$$\begin{aligned}
\widetilde{\mathbf{w}}_h^k &= \hat{\Sigma}_{k,h}^{-1}\sum_{t=1}^{k-1}\boldsymbol{\phi}(s_h^t,a_h^t)\left(r_h(s_h^t,a_h^t) + \max_{a'\in\mathcal{A}}\langle\boldsymbol{\phi}(s_{h+1}^t,a'),\widehat{\mathbf{w}}_{h+1}^k\rangle\right) && (8)\\
&= \hat{\Sigma}_{k,h}^{-1}\sum_{t=1}^{k-1}\boldsymbol{\phi}(s_h^t,a_h^t)\left(r_h(s_h^t,a_h^t) + \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_h^t,a_h^t)}\left[\max_{a'\in\mathcal{A}}\langle\boldsymbol{\phi}(s',a'),\widehat{\mathbf{w}}_{h+1}^k\rangle\right]\right) + (A) && (9)\\
&= \hat{\Sigma}_{k,h}^{-1}\sum_{t=1}^{k-1}\boldsymbol{\phi}(s_h^t,a_h^t)\langle\boldsymbol{\phi}(s_h^t,a_t^h),\mathcal{T}_h(\widehat{\mathbf{w}}_{h+1}^k)\rangle + (A) + (B) && (10)\\
&= \mathcal{T}_h(\widehat{\mathbf{w}}_{h+1}^k) + (A) + (B) + (C), && (11)
\end{aligned}$$

where

$$\begin{aligned}
(A) &\coloneqq \hat{\Sigma}_{k,h}^{-1}\sum_{t=1}^{k-1}\boldsymbol{\phi}(s_h^t,a_h^t)\left(\max_{a'\in\mathcal{A}}\langle\boldsymbol{\phi}(s_{h+1}^t,a'),\widehat{\mathbf{w}}_{h+1}^k\rangle - \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_h^t,a_h^t)}\left[\max_{a'\in\mathcal{A}}\langle\boldsymbol{\phi}(s',a'),\widehat{\mathbf{w}}_{h+1}^k\rangle\right]\right)\\
&= \hat{\Sigma}_{k,h}^{-1}\sum_{t=1}^{k-1}\boldsymbol{\phi}(s_h^t,a_h^t)\left(\widehat{V}_{h+1}^k(s_{h+1}^t) - \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_h^t,a_h^t)}\left[\widehat{V}_{h+1}^k(s')\right]\right), && (12)
\end{aligned}$$

and

$$(B) \coloneqq \hat{\Sigma}_{k,h}^{-1}\sum_{t=1}^{k-1}\boldsymbol{\phi}(s_h^t,a_h^t)\left(r_h(s_h^t,a_h^t) + \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_h^t,a_h^t)}\left[\max_{a'\in\mathcal{A}}\langle\boldsymbol{\phi}(s',a'),\widehat{\mathbf{w}}_{h+1}^k\rangle\right] - \langle\boldsymbol{\phi}(s_h^t,a_h^t),\mathcal{T}_h(\widehat{\mathbf{w}}_{h+1}^k)\rangle\right), \tag{13}$$

and

$$(C) \coloneqq \hat{\Sigma}_{k,h}^{-1}(-\lambda\mathbf{I}\mathcal{T}_h(\widehat{\mathbf{w}}_{h+1}^k)). \tag{14}$$

By the triangular inequality, $\left\|\widetilde{\mathbf{w}}_h^k - \mathcal{T}_h(\widehat{\mathbf{w}}_{h+1}^k)\right\|_{\hat{\Sigma}_{k,h}} \leq \|(A)\|_{\hat{\Sigma}_{k,h}} + \|(B)\|_{\hat{\Sigma}_{k,h}} + \|(C)\|_{\hat{\Sigma}_{k,h}}$.

By Proposition A.2, if $\lambda = 1$, since $\widehat{V}_{h+1}^k \in \mathcal{V}$, the following holds with probability $1-\delta$ for all $h \in [H]$:

$$\begin{aligned}
\|(A)\|_{\hat{\Sigma}_{k,h}} &= \left\|\sum_{t=1}^{k-1}\boldsymbol{\phi}(s_h^t,a_h^t)\left(\widehat{V}_{h+1}^k(s_{h+1}^t) - \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_h^t,a_h^t)}\left[\widehat{V}_{h+1}^k(s')\right]\right)\right\|_{\hat{\Sigma}_{k,h}^{-1}} && (15)\\
&\leq H\sqrt{d\ln\left(1+\frac{\phi_{\max}^2 k}{d}\right) + d\ln(w_{\max}^2\phi_{\max}^2 k) + \ln(H\delta^{-1})} + 1. && (16)
\end{aligned}$$

By Proposition A.3 and Equation (5):

$$\begin{aligned}
\|(B)\|_{\hat{\Sigma}_{k,h}} &= \left\| \sum_{t=1}^{k-1} \boldsymbol{\phi}(s_h^t, a_h^t) \left(r_h(s_h^t, a_h^t) + \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s_h^t, a_h^t)} \left[\max_{a' \in \mathcal{A}} \langle \boldsymbol{\phi}(s', a'), \widehat{\mathbf{w}}_{h+1}^k \rangle \right] - \langle \boldsymbol{\phi}(s_h^t, a_h^t), \mathcal{T}_h(\widehat{\mathbf{w}}_{h+1}^k) \rangle \right) \right\|_{\hat{\Sigma}_{k,h}^{-1}} \\
&\leq \left\| \sum_{t=1}^{k-1} \boldsymbol{\phi}(s_h^t, a_h^t) \mathcal{L}(\mathcal{T}_h(\widehat{\mathbf{w}}_{h+1}^k); s_h^t, a_h^t, \widehat{\mathbf{w}}_{h+1}^k) \right\|_{\hat{\Sigma}_{k,h}^{-1}} && (17) \\
&\leq \sqrt{k}\zeta. && (18)
\end{aligned}$$

Finally:

$$\begin{aligned}
\|(C)\|_{\hat{\Sigma}_{k,h}} &= \lambda \left\| \Sigma_{k,h}^{-1} \mathcal{T}_h(\widehat{\mathbf{w}}_{h+1}^k) \right\|_{\hat{\Sigma}_{k,h}} && (19) \\
&= \lambda \left\| \mathcal{T}_h(\widehat{\mathbf{w}}_{h+1}^k) \right\|_{\hat{\Sigma}_{k,h}^{-1}} && (20) \\
&\leq \sqrt{\lambda} \left\| \mathcal{T}_h(\widehat{\mathbf{w}}_{h+1}^k) \right\| && (21) \\
&\leq \sqrt{\lambda} w_{\max}. && (22)
\end{aligned}$$

□

Lemma A.5. *Assume $\lambda_{\min}(\hat{\Sigma}_{k,h}) \geq \lambda_k$ for all $k \geq 1$ and $h \in [H]$. Under Assumption A.1, with probability $1-\delta$, for all $k \geq 1, h \in [H], s \in \mathcal{S}, a \in \mathcal{A}$:*

$$|\widehat{Q}_h^k(s,a) - Q_h^*(s,a)| \leq (H-h) \left(\left(\sqrt{\beta_k(\delta)} + \sqrt{k}\zeta \right) \frac{\phi_{\max}}{\sqrt{\lambda_k}} + \zeta \right),$$

where $\sqrt{\beta_k}$ is defined in Lemma A.4.

Proof. First:

$$\begin{aligned}
|\widehat{Q}_h^k(s,a) - \mathcal{T}_h \widehat{Q}_{h+1}^k(s,a)| &= \left| \langle \boldsymbol{\phi}(s,a), \widehat{\mathbf{w}}_h^k \rangle - r_h(s,a) - \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)} \left[\max_{a' \in \mathcal{A}} \langle \boldsymbol{\phi}(s', a'), \widehat{\mathbf{w}}_{h+1}^k \rangle \right] \right| && (23) \\
&= \left| \langle \boldsymbol{\phi}(s,a), \widehat{\mathbf{w}}_h^k - \mathcal{T}_h(\widehat{\mathbf{w}}_{h+1}^k) \rangle + \langle \boldsymbol{\phi}(s,a), \mathcal{T}_h(\widehat{\mathbf{w}}_{h+1}^k) \rangle - r_h(s,a) - \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)} \left[\max_{a' \in \mathcal{A}} \langle \boldsymbol{\phi}(s', a'), \widehat{\mathbf{w}}_{h+1}^k \rangle \right] \right| && (24) \\
&\leq \left| \langle \boldsymbol{\phi}(s,a), \widehat{\mathbf{w}}_h^k - \mathcal{T}_h(\widehat{\mathbf{w}}_{h+1}^k) \rangle \right| + \zeta \qquad \text{(by Equation 5)} && (25) \\
&\leq \|\boldsymbol{\phi}(s,a)\|_{\hat{\Sigma}_{k,h}^{-1}} \left\| \widehat{\mathbf{w}}_h^k - \mathcal{T}_h(\widehat{\mathbf{w}}_{h+1}^k) \right\|_{\hat{\Sigma}_{k,h}} + \zeta && (26) \\
&\leq \frac{\phi_{\max}}{\sqrt{\lambda_{\min}(\hat{\Sigma}_{k,h})}} \left\| \widehat{\mathbf{w}}_h^k - \mathcal{T}_h(\widehat{\mathbf{w}}_{h+1}^k) \right\|_{\hat{\Sigma}_{k,h}} + \zeta && (27) \\
&\leq \frac{\phi_{\max}}{\sqrt{\lambda_{\min}(\hat{\Sigma}_{k,h})}} \left(\sqrt{\beta_k(\delta)} + \sqrt{k}\zeta \right) + \zeta && (28) \\
&\leq \frac{\phi_{\max}}{\sqrt{\lambda_k}} \left(\sqrt{\beta_k(\delta)} + \sqrt{k}\zeta \right) + \zeta := \varepsilon_k. && (29)
\end{aligned}$$

The rest of the proof is by (backward) induction. Note that $Q_{H+1}^* = \widehat{Q}_{H+1}^k = 0$. Then, $\mathcal{T}_H \widehat{Q}_{H+1}^k = \mathcal{T}_H Q_{H+1}^*(s,a) = Q_H^*(s,a)$. By Equation (29):

$$|\widehat{Q}_H^k(s,a) - Q_H^*(s,a)| = |\widehat{Q}_H^k(s,a) - \mathcal{T}_H \widehat{Q}_{H+1}^k| \leq \varepsilon_k. \tag{30}$$

This is our base case ($h = H$). The inductive hypothesis is:

$$|\widehat{Q}^k_{h+1}(s,a) - Q^*_{h+1}(s,a)| \leq (H-h-1)\varepsilon_k. \tag{31}$$

Then:

$$\begin{aligned}
|\widehat{Q}^k_h(s,a) - Q^*_h(s,a)| &= |\widehat{Q}^k_h(s,a) - \mathcal{T}_h\widehat{Q}^k_{h+1}(s,a) + \mathcal{T}_h\widehat{Q}^k_{h+1}(s,a) - Q^*_h(s,a)| && (32)\\
&\leq |\widehat{Q}^k_h(s,a) - \mathcal{T}_h\widehat{Q}^k_{h+1}(s,a)| + |\mathcal{T}_h\widehat{Q}^k_{h+1}(s,a) - Q^*_h(s,a)| && (33)\\
&\leq \varepsilon_k + |\mathcal{T}_h\widehat{Q}^k_{h+1}(s,a) - Q^*_h(s,a)| && (34)\\
&= \varepsilon_k + |\mathcal{T}_h\widehat{Q}^k_{h+1}(s,a) - \mathcal{T}_hQ^*_{h+1}(s,a)| && (35)\\
&\leq \varepsilon_k + |\widehat{Q}^k_{h+1}(s,a) - Q^*_{h+1}(s,a)| && (36)\\
&\leq \varepsilon_k + (H-h-1)\varepsilon_k && (37)\\
&= (H-h)\varepsilon_k, && (38)
\end{aligned}$$

where the inequalities are, in order: by triangular inequality, by Equation (29), by the contraction property of Bellman's operator, by the induction hypothesis. □

Lemma 4.1 is just a special case of Lemma A.5 when $\zeta = 0$.

B Proof of Lemma 4.2

Lemma 4.2. *Given Assumptions 3.1 and 3.2, the following holds for the minimum eigenvalue of the empirical covariance matrix for each $h \in [H]$ and for each $k \geq 1$:*

$$\mathbb{P}\left[\lambda_{min}(\hat{\Sigma}_{k,h}) \geq \lambda + \lambda_0 k - 8\phi^2_{max}\sqrt{k\log(4dk/\delta)}\right] \geq 1 - \frac{\delta}{2}$$

Proof. Let π_τ be the policy played by Algorithm 1 in the τ-th episode. We can rewrite the design matrix as:

$$\begin{aligned}
\hat{\Sigma}_{k,h} &= \sum_{\tau=1}^{k}\phi(s^\tau_h,a^\tau_h)\phi(s^\tau_h,a^\tau_h)^\top + \lambda\mathbf{I} && (39)\\
&= \lambda\mathbf{I} + \sum_{\tau=1}^{k}\mathbb{E}_{s\sim\rho^{\pi_\tau}_h}[\phi(s^\tau_h,a^\tau_h)\phi(s^\tau_h,a^\tau_h)^\top] - \sum_{\tau=1}^{k}\left(\mathbb{E}_{s\sim\rho^{\pi_\tau}_h}[\phi(s^\tau_h,a^\tau_h)\phi(s^\tau_h,a^\tau_h)^\top] - \phi(s^\tau_h,a^\tau_h)\phi(s^\tau_h,a^\tau_h)^\top\right) && (40)
\end{aligned}$$

If $X_\tau = \mathbb{E}_{s\sim\rho^{\pi_\tau}_h}[\phi(s^\tau_h,a^\tau_h)\phi(s^\tau_h,a^\tau_h)^\top] - \phi(s^\tau_h,a^\tau_h)\phi(s^\tau_h,a^\tau_h)^\top$, then we have that $\mathbb{E}_\tau[X_\tau] = 0$. Also since X_τ is symmetric:

$$X^2_\tau \leq \lambda_{\max}(X^2_\tau)\mathbf{I} \leq ||X_\tau||^2\mathbf{I} \leq 4\phi^4_{\max}\mathbf{I} \tag{41}$$

Hence from the matrix Azuma inequality by Tropp (2011), with probability $1-\delta_k$, for all $k \geq 1$:

$$\lambda_{\max}\left(\sum_{\tau=1}^{k}X_\tau\right) \leq 4\phi^2_{\max}\sqrt{2k\log d/\delta_k} \tag{42}$$

We set $\delta_k = \delta/(2k^2)$ and perform a union bound over time. Finally with probability at least $1-\delta$ for all k≥ 1:

$$\lambda_{\max}\left(\sum_{\tau=1}^{k}X_\tau\right) \leq 4\phi^2_{\max}\sqrt{2k\log(4dk^2/\delta)} \leq 8\phi^2_{\max}\sqrt{k\log(4dk/\delta)}. \tag{43}$$

Now let $\pi = \pi_k$. By definition of Algorithm 1:

$$\sum_{\tau=1}^{k} \mathbb{E}_{s\sim\rho_h^{\pi_\tau}}[\phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top] = \sum_{\tau=1}^{k} \mathbb{E}_{s\sim\rho_h^{\pi_\tau}}[\phi(s_h^\tau, \pi_h^\tau(s))\phi(s_h^\tau, \pi_h^\tau(s))^\top] \tag{44}$$

$$= \sum_{\tau=1}^{k} \mathbb{E}_{s\sim\rho_h^{\pi_\tau}}[\phi(s_h^\tau, \pi_h^\tau(s))\phi(s_h^\tau, \pi_h^\tau(s))^\top \mathbb{I}\{\langle \boldsymbol{\phi}(s, \pi_h^\tau(s)), \widehat{\mathbf{w}}_h^k\rangle \geq \max_{a\in\mathcal{A}} \langle \boldsymbol{\phi}(s,a), \widehat{\mathbf{w}}_h^k\rangle\}]. \tag{45}$$

So, by Assumption 3.2:

$$\lambda_{\min}(\hat{\Sigma}_{k,h}) \geq \lambda \mathbf{I} + \lambda_0 k - 8\phi_{\max}^2 \sqrt{k \log(4dk/\delta)}. \tag{46}$$

□

C Other Proofs

Proposition 3.3 (Noisy features)**.** *Let $|\mathcal{A}| = 2$ and $\boldsymbol{\phi}(s,a) = f(s,a) + \eta(a)$ for some function $f : \mathcal{S}\times\mathcal{A} \to [0, \sqrt{2}\sigma]$ and independent Gaussian noises $\eta(a) \sim \mathcal{N}(0, \sigma^2)$. Then Assumption 3.2 holds with $\lambda_0 \geq 0.2\sigma^2$.*

Proof. Let $\mathcal{A} = \{a_1, a_2\}$, $w \in \mathcal{W}$, and π be any deterministic policy:

$$\lambda_0 = \mathbb{E}_{s\sim\rho_h^\pi(s)}\big[\phi(s,\pi(s))^2 \mathbb{I}\{\phi(s,\pi(s))w \geq \max_{a\in\mathcal{A}}\{\phi(s,a)w\}\}\big] \tag{47}$$

$$= \mathbb{E}_{s\sim\rho_h^\pi(s)}\big[(f(s,a) + \eta(a))^2 \mathbb{I}\{\phi(s,\pi(s))w \geq \max_{a\in\mathcal{A}}\{\phi(s,a)w\}\}\big] \tag{48}$$

$$\geq \mathbb{E}_{s\sim\rho_h^\pi(s)}\big[\eta(a)^2 \mathbb{I}\{\phi(s,\pi(s))w \geq \max_{a\in\mathcal{A}}\{\phi(s,a)w\}\}\big] \tag{49}$$

$$= \sigma^2 \mathbb{P}\big(\phi(s,\pi(s))w \geq \max_{a\in\mathcal{A}}\{\phi(s,a)w\}\big) \tag{50}$$

$$= \sigma^2 \mathbb{P}\big((f(s,\pi(s)) + \eta(\pi(s)))w \geq \max_{a\in\mathcal{A}}\{(f(s,a) + \eta(a))w\}\big). \tag{51}$$

Fix a state s and let $\pi(s) = a_1$ and $w > 0$ without loss of generality. Then

$$\mathbb{P}\big((f(s,a_1) + \eta(a_1))w \geq \max_{a\in\mathcal{A}}\{(f(s,a) + \eta(a))w\}\big) \geq \mathbb{P}\big((f(s,a_1) + \eta(a_1))w \geq (f(s,a_2) + \eta(a_2))w\big) \tag{52}$$

$$= \mathbb{P}\big(f(s,a_1) + \eta(a_1) \geq f(s,a_2) + \eta(a_2)\big) \tag{53}$$

$$\geq \mathbb{P}\big(\eta(a_1) - \eta(a_2) > \sqrt{2}\sigma\big) \tag{54}$$

$$= \mathbb{P}(X > \sqrt{2}\sigma) > 0.2, \tag{55}$$

where $X \sim \mathcal{N}(0, 2\sigma^2)$. □

Corollary 3.5. *Under the same assumptions as Theorem 3.4, the expected cumulative regret of LSVI is at most:*

$$\mathbb{E}[R(K)] = \mathcal{O}\left(H^3 \sqrt{\frac{dK}{\lambda_0}} \log(K)\right).$$

Proof. We define an event P, under which the cumulative regret is bounded as described in Theorem 3.4, that occurs with probability $1-\delta$:

$$\begin{aligned}\mathbb{E}[R(K)] &= \mathbb{E}[R(K)\mathbb{I}(P)] + \mathbb{E}[R(K)\mathbb{I}(\overline{P})] \\ &\leq \mathcal{O}\left(H^3\sqrt{\frac{dK}{\lambda_0}}\log(K/\delta)\right)(1-\delta) + 2HK\delta \\ &\leq \mathcal{O}\left(H^3\sqrt{\frac{dK}{\lambda_0}}\log(K^{3/2})\right) + 2H\sqrt{K} \\ &= \mathcal{O}\left(H^3\sqrt{\frac{dK}{\lambda_0}}\log(K)\right)\end{aligned}$$

The first inequality is by Theorem 3.4, and by upper bounding the regret with the trivial $2HK$ when P does not hold. The second inequality follows by $1-\delta \leq 1$ and by setting $\delta = 1/\sqrt{K}$. □

Theorem 3.7. *If Assumptions 3.6 and 3.2 are satisfied, with probability $1-\delta$, the cumulative regret of LSVI is at most:*

$$R(K) = \widetilde{\mathcal{O}}\left(H^3\sqrt{\frac{dK}{\lambda_0}} + H^2\zeta\frac{K}{\sqrt{\lambda_0}} + H^2\zeta K\right).$$

Proof. This is a simple variant of Theorem 3.4 that uses the more general Lemma A.5 in place of Lemma 4.1. Logarithmic terms are omitted for brevity. □

D Code and Computing Infrastructure

The experimental results presented in Section 6 were obtained using Python. The code, a modified version of the official implementation from Liu & Xu (2024), is attached to the paper and was executed on Kaggle.com, a cloud computing platform.