Logarithmic Regret for Linear Markov Decision Processes with Adversarial Corruptions

Canzhe Zhao[1], Xiangcheng Zhang[2], Baoxiang Wang[3], Shuai Li [1*]

[1] John Hopcroft Center for Computer Science, Shanghai Jiao Tong University
[2] Weiyang College, Tsinghua University
[3] School of Data Science, The Chinese University of Hong Kong, Shenzhen
{canzhezhao,shuaili8}@sjtu.edu.cn, xc-zhang21@mails.tsinghua.edu.cn, bxiangwang@cuhk.edu.cn

Abstract

In this work, we study the logarithmic regret for reinforcement learning (RL) with linear function approximation and adversarial corruptions, in the formulation of linear Markov decision processes (MDPs). Specifically, we consider the case where there exist adversarial corruptions over the reward functions, and the total amount of the corruptions of each step h across all episodes K is bounded by a corruption level $C \geq 0$. We propose an algorithm, double-weighted least-squares value iteration with UCB (DW-LSVI-UCB), which leverages weighted linear regressions to learn the (corrupted) unknown reward parameters and unknown transition parameters simultaneously. We prove that DW-LSVI-UCB attains an $\widetilde{O}(\frac{d^2H^4\log^2(1+K/\delta)}{\mathrm{gap}_{\min}} + CdH^2)$ regret (omitting the dependence on lower order terms), where d is the ambient dimension of the feature mapping, H is the horizon length, $\mathrm{gap}_{\min}$ is the minimal sub-optimality gap, and K is the number of episodes. Additionally, when there are no adversarial corruptions over reward functions, the regret of our algorithm improves the previous best result by an $\widetilde{O}(dH/\log K)$ factor.

1 Introduction

Reinforcement learning (RL), which is typically modeled as Markov decision processes (MDPs) (Feinberg 1996), has garnered significant empirical success across diverse domains such as gaming, control systems, and robotics. For tabular MDPs with small finite state space and finite action space, nearly minimax optimal sample complexity has been attained in discounted MDPs utilizing a generative model (Azar, Munos, and Kappen 2013). In the absence of such a generative model, optimal sample complexity has been achieved in MDPs with finite and infinite horizons (Azar, Osband, and Munos 2017; He, Zhou, and Gu 2021b; Tossou, Basu, and Dimitrakakis 2019). Nevertheless, in practical RL applications, the state and action spaces can be immense, even approaching infinity, rendering tabular MDPs vulnerable to the curse of dimensionality. To address this challenge, recent research has pivoted towards examining MDPs under the lens of *function approximation*. Typically, this approach involves reparameterizing the values of state-action pairs by embedding them into a low-dimensional space using a specified feature mapping. RL with linear function approximation, in particular, has attracted significant research attention in recent years. Among these studies, linear mixture MDPs (Ayoub et al. 2020) and linear MDPs (Jin et al. 2020) are two of the most prominent MDP models that enable RL with linear function approximation. Notably, recent works have achieved the (nearly) minimax optimal regret guarantee of order $\widetilde{O}(dH\sqrt{KH})$ in both linear mixture MDPs (Zhou, Gu, and Szepesvári 2021) and linear MDPs (Hu, Chen, and Huang 2022; He et al. 2023) with stochastic reward functions, where d is the ambient dimension of the feature mapping, H is the horizon length, and K is the number of episodes.

Despite significant advances in establishing (nearly) minimax optimal regret for RL with linear function approximation, the previous algorithms are still not able to be shown to achieve better performance in "easier" environments and might not be able to work in the more "difficult" environments. Specifically, when the instance-dependent structure is relatively mild to be learned, the algorithms should be expected to perform better than in environments with instance-dependent structures that are hardest to learn. Indeed, for bandit problems, the more amenable special case of RL, and tabular MDP problems, instance-dependent regret of $O(\log K)$ has been established (Simchowitz and Jamieson 2019; Jin and Luo 2020; Lattimore and Szepesvári 2020; Yang, Yang, and Du 2021; Jin, Huang, and Luo 2021). One notable exception is the work of He, Zhou, and Gu (2021a), which establishes an instance-dependent regret of order $\widetilde{O}\left(d^3H^5\log K/\mathrm{gap}_{\min}\right)$ for linear MDPs and an instance-dependent regret of order $\widetilde{O}\left(d^2H^5\log^3 K/\mathrm{gap}_{\min}\right)$ for linear mixture MDPs, where $\mathrm{gap}_{\min}$ is the minimal sub-optimality gap. However, He, Zhou, and Gu (2021a) also provide a regret lower bound of order $\Omega\left(Hd/\mathrm{gap}_{\min}\right)$ for both problems, indicating that there still exist large gaps between the regret upper bounds and the regret lower bound. On the other hand, when non-stationarity arises in the environment, these algorithms will fail to work in such cases, except for the algorithms by Jin and Luo (2020); Jin, Huang, and Luo (2021), which adapt to the stochastic environment with $\widetilde{O}\left(\log K/\mathrm{gap}_{\min}\right)$ regret and also guarantee $\widetilde{O}(\log K + \sqrt{C})$ regret when there is a total amount of C

*Corresponding author.

corruptions over the reward functions in the environment. However, the algorithms by Jin and Luo (2020); Jin, Huang, and Luo (2021) are tailored to the cases of tabular MDPs, leaving the more challenging and practical cases with large-scale state-action spaces unexplored.

To address the aforementioned limitations of existing algorithms, in this work, we aim to establish an algorithm for RL with linear function approximation that can enjoy the $\widetilde{O}(\log K/\operatorname{gap}_{\min})$-type instance-dependent regret in stochastic environments, while being robust to adversarial corruptions over reward functions. Formally, we study linear MDPs where both the state transitions and the reward functions admit linear structures but the rewards can be corrupted by an *adaptive* adversary. At each step h in each episode k, the adversary may add an arbitrary corruption over the sampled reward $\hat{r}_{k,h}(s_{k,h}, a_{k,h})$ of state-action pair $(s_{k,h}, a_{k,h})$, which might potentially depend on all the information up to step h in episode k including the action $a_{k,h}$ taken by the learner. To tackle this problem, we propose an algorithm called DW-LSVI-UCB, which simultaneously leverages weighted regressions to learn the unknown reward parameters against the adversarial corruptions and the unknown transition parameters to adapt to the variances of the transition noises. Consequently, our algorithm achieves the regret of order $\widetilde{O}(\frac{d^2H^4\log^2(1+K/\delta)}{\operatorname{gap}_{\min}} + CdH^2)$ (omitting the dependence on lower order terms), where C is the total amount of the reward corruptions. Our regret upper bound also improves the previous best result by an $\widetilde{O}(dH/\log K)$ factor, when there are no adversarial corruptions over reward functions (*i.e.*, $C = 0$). The analysis for our algorithm is inspired by previous works studying gap-dependent regret for uncorrupted MDPs (Yang, Yang, and Du 2021; He, Zhou, and Gu 2021a) but with more refined parts to sharpen the dependence of $\log K$ (see Section 4 for details).

1.1 Additional Related Works

RL with Function Approximation To enable efficient learning in RL within vast state-action spaces, recent studies have gravitated towards RL algorithms employing function approximation. Broadly, these results can be grouped into three categories. The first line of research explores the linear MDP assumption (Yang and Wang 2019; Jin et al. 2020; Du et al. 2020; Zanette et al. 2020; Wang, Salakhutdinov, and Yang 2020; Wang et al. 2021; He, Zhou, and Gu 2021a; Luo, Wei, and Lee 2021; Dai et al. 2023; Sherman, Koren, and Mansour 2023; Zhong and Zhang 2024; Kong et al. 2024), wherein both the transition probabilities and the reward functions are modeled as linear functions of given state-action feature mappings. Notably, Jin et al. (2020) pioneer a statistically and computationally efficient algorithm with a regret guarantee of $\widetilde{O}(H^2\sqrt{d^3K})$. He et al. (2023) then enhance this achievement by leveraging weighted ridge regression and a Bernstein-type exploration bonus, achieving the (nearly) minimax optimal regret bound of $\widetilde{O}(dH\sqrt{KH})$. In this work, we likewise focus on linear MDPs. The second approach hinges on the linear mixture MDP assumption (Ayoub et al. 2020; Cai et al. 2020; Zhang et al. 2021; Zhou, Gu, and Szepesvári 2021; He, Zhou, and Gu 2021a; Zhou and Gu 2022; Wu, Zhou, and Gu 2022; Min et al. 2022; He, Zhou, and Gu 2022; Zhao et al. 2023; Li, Zhao, and Zhou 2024), where the transition probabilities are linear in an unknown parameter and a given feature mapping of state-action-next-state triplets. Among these studies, Zhou, Gu, and Szepesvári (2021) attain the (nearly) minimax optimal regret bound of $\widetilde{O}(dH\sqrt{KH})$. The final line of research examines the general function approximation in RL (Jiang et al. 2017; Dann et al. 2018; Sun et al. 2019; Du et al. 2019a, 2021; Jin, Liu, and Miryoosefi 2021; Uehara, Zhang, and Sun 2022).

Reward Corruptions in RL A substantial amount of literature has been dedicated to exploring the challenges in sequential decision-making problems with adversarially corrupted rewards, including bandits (Lykouris, Mirrokni, and Leme 2018; Gupta, Koren, and Talwar 2019; Yang et al. 2020; Garcelon et al. 2020; Lu, Wang, and Zhang 2021; Ito 2021a; Zimmert and Seldin 2021; Ito 2021b; Zhong, Cheung, and Tan 2021; He et al. 2022; Kong, Zhou, and Li 2022; Ito, Tsuchiya, and Honda 2022; Ye et al. 2023a; Ito and Takemura 2023; Kong, Zhao, and Li 2023; Dann, Wei, and Zimmert 2023; Ito, Tsuchiya, and Honda 2024) and RL (Lykouris et al. 2021; Wu et al. 2021; Zhang et al. 2020; Jin and Luo 2020; Jin, Huang, and Luo 2021; Wei, Dann, and Zimmert 2022; Ye et al. 2023a; Chen et al. 2023; Ye et al. 2023b, 2024). Amongst these works, some works study the best-of-both-worlds guarantee (Zimmert and Seldin 2021; Ito 2021b; Jin and Luo 2020; Jin, Huang, and Luo 2021; Kong, Zhou, and Li 2022; Ito, Tsuchiya, and Honda 2022; Ito and Takemura 2023; Kong, Zhao, and Li 2023; Dann, Wei, and Zimmert 2023; Ito, Tsuchiya, and Honda 2024), meaning that the algorithms in these works can simultaneously enjoy the $O(\log K/\operatorname{gap}_{\min})$ regret in stochastic environments and $O(\sqrt{K})$ regret in adversarial environments where an adversary may corrupt the rewards in each episode. In addition, we note that some works study RL problems where both the rewards and state transitions might be corrupted (Ye et al. 2023a,b, 2024).

2 Preliminaries

An episodic Markov decision process (MDP) is formally represented by the tuple $M(\mathcal{S}, \mathcal{A}, H, \{r_h\}_{h=1}^H, \{\mathbb{P}_h\}_{h=1}^H)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, H denotes the episode length, $r_h : \mathcal{S} \times \mathcal{A} \to [-1, 1]$ is the reward function at step h, and $\mathbb{P}_h : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the state transition probability with $\mathbb{P}_h(s'|s, a)$ being as the probability of transitioning from state s to state s' after taking action a at step h. Following Jin et al. (2020), we postulate that $\mathcal{S}$ is a measurable space potentially comprising an infinite number of states, while $\mathcal{A}$ is a finite set.

The interaction protocol between the learner and the environment is given as follows. At the commencement of each episode k, the learner specifies a policy set $\pi_k = \{\pi_{k,h}\}_{h=1}^H$ to adhere to during that episode. Meanwhile, an initial state $s_{k,1}$ is sample from $\mathbb{P}_0(\cdot)$. At each step h within episode k, the learner observes the current state $s_{k,h}$ and samples an action $a_{k,h} \sim \pi_{k,h}(\cdot|s_{k,h})$. Then the learner observes a reward $\hat{r}_{k,h}(s_{k,h}, a_{k,h}) = r_h(s_{k,h}, a_{k,h}) + \varepsilon_{k,h}$, where $\varepsilon_{k,h}$

is a conditionally 1-sub-Gaussian stochastic noise. Subsequently, the system transitions to the next state $s_{k,h+1} \sim \mathbb{P}_h(\cdot|s_{k,h}, a_{k,h})$, which will be observable to the learner at step $h+1$. Finally, episode k will end after the learner reaches the (fixed) terminal state s_{H+1}.

For a given policy set $\pi = \{\pi_h\}_{h=1}^H$, its state-action value $Q_h^\pi(s,a)$ and state value $V_h^\pi(s)$ for state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$ are defined as $Q_h^\pi(s,a) = \mathbb{E}\left[\sum_{h'=h}^H r_{h'}\left(s_{h'}, a_{h'}\right) \middle| (s_h, a_h) = (s,a), \pi, \mathbb{P}\right]$ and $V_h^\pi(s) = \mathbb{E}_{a\sim\pi_h(\cdot|s)}\left[Q_h^\pi\left(s,a\right)\right]$, where the expectation is taken over the randomness of the policy π and the environment state transition $\mathbb{P}$. Additionally, for $(s,a) \in \mathcal{S} \times \mathcal{A}$, we denote by $Q_h^*(s,a) = \max_\pi Q_h^\pi(s,a)$ the optimal state-action value and $V_h^*(s) = \max_\pi V_h^\pi(s)$ the optimal state value. It is known that there exists (at least one) optimal policy π^* such that $Q_h^{\pi^*}(s,a) = Q_h^*(s,a)$ and $V_h^{\pi^*}(s) = V_h^*(s)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $h \in [H]$.

We proceed to define the regret accumulated over the K episodes as

$$\mathrm{Reg}_K = \sum_{k=1}^K V_1^*(s_{k,1}) - V_1^{\pi_k}(s_{k,1}),$$

where π_k is the policy adopted by the learner in the k-th episode.

Following (Simchowitz and Jamieson 2019; Du et al. 2019b; Yang, Yang, and Du 2021; He, Zhou, and Gu 2021a), we formally define the minimal sub-optimality gap $\mathrm{gap}_{\min}$ for MDPs as detailed below, under the assumption that $\mathrm{gap}_{\min} > 0$.

Definition 2.1. *For each state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$ and step $h \in [H]$, let $\mathrm{gap}_h(s,a) = V_h^*(s) - Q_h^*(s,a)$ be the sub-optimality gap at (s,a). Then the minimal sub-optimality gap is defined as*

$$\mathrm{gap}_{\min} = \min_{h\in[H],(s,a)\in\mathcal{S}\times\mathcal{A}} \{\mathrm{gap}_h(s,a) : \mathrm{gap}_h(s,a) > 0\}.$$

Linear Markov Decision Process In this work, we study RL with linear function approximation, in the formulation of linear MDPs, detailed as follows.

Definition 2.2. *An MDP $\mathcal{M}(\mathcal{S}, \mathcal{A}, H, \{r_h\}_{h=1}^H, \{\mathbb{P}_h\}_{h=1}^H)$ is a linear MDP if for any step $h \in [H]$, there exist an unknown vector $\boldsymbol{\theta}_h \in \mathbb{R}^d$, an unknown (signed) measures $\boldsymbol{\mu}_h(\cdot) : \mathcal{S} \to \mathbb{R}^d$ and a known feature mapping $\boldsymbol{\phi} : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^d$, such that for each state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$ and state $s' \in \mathcal{S}$, it holds that*

$$r_h(s,a) = \left\langle \boldsymbol{\phi}(s,a), \boldsymbol{\theta}_h \right\rangle, \mathbb{P}_h(s'|s,a) = \left\langle \boldsymbol{\phi}(s,a), \boldsymbol{\mu}_h(s') \right\rangle.$$

Moreover, we further impose the regularity condition over the linear MDPs, assuming that $\|\boldsymbol{\theta}_h\|_2 \le 1$, $\left\|\boldsymbol{\mu}_h(\mathcal{S})\right\|_2 \le \sqrt{d}$ and $\|\boldsymbol{\phi}(s,a)\|_2 \le 1$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $h \in [H]$.

Corruptions over Reward Functions In this work, we consider the case where the rewards of linear MDPs might be adversarially corrupted, subject to the following conditions. At each step h in each episode k, *after* the learner takes action $a_{k,h}$ at state $s_{k,h}$ and the reward $\hat{r}_{k,h}(s_{k,h}, a_{k,h}) = r_h(s_{k,h}, a_{k,h}) + \varepsilon_{k,h}$ is sampled from the environment, the adversary observes the reward $\hat{r}_{k,h}$ and may impose an adversarial corruption $c_{k,h}$ onto the reward $\hat{r}_{k,h}$, which might depend on all the information up to step $h-1$ in episode k as well as $(s_{k,h}, a_{k,h}, \hat{r}_{k,h})$. Then the corrupted reward $\tilde{r}_{k,h}(s_{k,h}, a_{k,h}) = \hat{r}_{k,h}(s_{k,h}, a_{k,h}) + c_{k,h}$ is revealed to the learner. Note that we do not impose any structural assumptions over the corruption $c_{k,h}$, and $c_{k,h}$ is potentially dependent on the current action $a_{k,h}$ chosen by the learner (as well as the current state $s_{k,h}$ of the learner), while the corruption $c_{k,h}$ is determined by the adversary *before* it observes the current action $a_{k,h}$ of the learner in some prior works (Lykouris, Mirrokni, and Leme 2018; Gupta, Koren, and Talwar 2019; Jin and Luo 2020; Jin, Huang, and Luo 2021). In addition, we assume that the total amount of corruption at each step h is bounded by some corruption level $C \ge 0$. That is, $\max_{h\in[H]} \sum_{k=1}^K |c_{k,h}| \le C$.

Additional Notations For any state value function $V : \mathcal{S} \to \mathbb{R}$, we introduce the shorthand notations $[\mathbb{P}_h V](s,a) = \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s,a)}[V(s')]$ and $[\mathbb{V}_h V](s,a) = [\mathbb{P}_h V^2](s,a) - \left([\mathbb{P}_h V](s,a)\right)^2$ for notational convenience. Consequently, by the Bellman equation for MDPs, for each step h and policy π, the state-action value $Q_h^\pi(s,a)$ can be rewritten as $Q_h^\pi(s,a) = r_h(s,a) + [\mathbb{P}_h V_{h+1}^\pi](s,a)$, where we define $V_{H+1}^\pi(s) = 0$ for all $s \in \mathcal{S}$. We denote by $o_{i,j} = (s_{i,j}, a_{i,j}, \tilde{r}_{i,j}(s_{i,j}, a_{i,j}))$ be the observation of the learner at episode i and step j. Let $\mathcal{F}_{k,h}$ be the σ-algebra generated by $\{o_{1,1}, \ldots, o_{1,H}, o_{2,1}, \ldots, o_{k,1}, \ldots, o_{k,h}\}$. For simplicity, we sometimes abbreviate $\mathbb{E}[\cdot|\mathcal{F}_{k,h}]$, $\boldsymbol{\phi}(s_{k,h}, a_{k,h})$, $\tilde{r}_{k,h}(s_{k,h}, a_{k,h})$ as $\mathbb{E}_{k,h}[\cdot]$, $\boldsymbol{\phi}_{k,h}$ and $\tilde{r}_{k,h}$. We sometimes denote $\log_+(x) = \max\{1, \log x\}$.

3 Algorithm

This section introduces the proposed double-weighted least-squares value iteration with UCB (DW-LSVI-UCB) algorithm, detailed in Algorithm 1. Our algorithm is built upon the LSVI-UCB^{++}algorithm by He et al. (2023), but with several key alterations, illustrated in Section 3.1 and 3.2, respectively.

3.1 Transition Parameter Learning

To learn the unknown transition parameter $\boldsymbol{\mu}_h$, the LSVI-UCB^{++}algorithm (He et al. 2023) adopts a model-free manner and utilizes the fact the state-action value function of linear MDPs are linear with the feature mapping $\boldsymbol{\phi}(\cdot,\cdot)$ in the sense that $Q_h^\pi(s,a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{\mu}_h V_{h+1}^\pi + \boldsymbol{\theta}_h \rangle$. Nevertheless, in this way, the learning of the transition parameter $\boldsymbol{\mu}_h$ and the reward parameter $\boldsymbol{\theta}_h$ are coupled with each other, making it hard to handle the adversarial corruptions over rewards simultaneously. To this end, our algorithm adopts a model-based manner, enabling separate learning of the transition parameter $\boldsymbol{\mu}_h$ and the reward parameter $\boldsymbol{\theta}_h$. We detail the learning of $\boldsymbol{\mu}_h$ in this section and present the learning of $\boldsymbol{\theta}_h$ in the next section.

Weighted Regression for Learning $\boldsymbol{\mu}_h$ To enable efficient learning of the unknown transition parameter $\boldsymbol{\mu}_h$, we

utilize a weighted regression scheme to learn $\boldsymbol{\mu}_h$ by solving

$$\hat{\boldsymbol{\mu}}_{k,h} = \mathop{\arg\min}_{\boldsymbol{\mu}\in\mathbb{R}^{d\times S}} \sum_{i=1}^{k-1} \left\| \hat{\sigma}_{i,h}^{-1} \left[\boldsymbol{\mu}^\top \boldsymbol{\phi}_{i,h} - \boldsymbol{\delta}_{i,h+1}\right] \right\|_2^2 + \lambda^P \|\boldsymbol{\mu}\|_F^2 , \tag{1}$$

where $\hat{\sigma}_{i,h}$ is the estimated variance, $\boldsymbol{\delta}_{i,h+1} \in \{0,1\}^S$ is the Dirac measure centred on $s_{i,h+1}$ (*i.e.*, the one-hot vector with the one entry at $s_{i,h+1}$), $\lambda^P > 0$ is the transition regularization parameter and $\|\cdot\|_F$ denotes the Frobenius norm. The solution of Eq. (1) is given by

$$\hat{\boldsymbol{\mu}}_{k,h} = (\boldsymbol{\Sigma}_{k,h}^P)^{-1} \boldsymbol{b}_{k,h}^P , \tag{2}$$

where $\boldsymbol{\Sigma}_{k,h}^P = \sum_{i=1}^{k-1} \hat{\sigma}_{i,h}^{-2} \boldsymbol{\phi}_{i,h}\boldsymbol{\phi}_{i,h}^\top + \lambda^P \boldsymbol{I}$ is the weighted feature covariance matrix and $\boldsymbol{b}_{k,h}^P = \sum_{i=1}^{k-1} \hat{\sigma}_{i,h}^{-2} \boldsymbol{\phi}_{i,h}\boldsymbol{\delta}_{i,h+1}^\top$. Note that $\hat{\boldsymbol{\mu}}_{k,h}$ can be computed in an "implicit" manner with time and space complexities both of order $O(\mathrm{poly}(d))$ (see Appendix for details).

Construction of Variance Estimates The estimated variance $\hat{\sigma}_{i,h}$ of $[\mathbb{V}_h \hat{V}_{k,h+1}](s_{i,h}, a_{i,h})$ in Eq. (1) is used to construct a sharper confidence set that adapts to the variance of the transition noises so as to enable the use of the Bernstein-type self-normalized vector-valued martingale inequality, which is originally proposed by Zhou, Gu, and Szepesvári (2021) to obtain the (nearly) minimax optimal regret for learning stochastic linear mixture MDPs and subsequently used by Hu, Chen, and Huang (2022); He et al. (2023) to achieve the (nearly) minimax optimal regret for learning stochastic linear MDPs. Note that directly setting $\hat{\sigma}_{i,h}$ as the estimated variance of $[\mathbb{V}_h \hat{V}_{k,h+1}](s_{i,h}, a_{i,h})$ will induce the measurability issue (*i.e.*, the noise $\hat{V}_{k,h+1}(s_{i,h+1}) - [\mathbb{P}_h \hat{V}_{k,h+1}](s_{i,h}, a_{i,h})$ is not $\mathcal{F}_{i+1,h}$-measurable and $\mathbb{E}_{i,h}\left[\hat{V}_{k,h+1}(s_{i,h+1}) - [\mathbb{P}_h \hat{V}_{k,h+1}](s_{i,h}, a_{i,h})\right] \neq 0$), which requires a uniform covering argument to address and eventually leads to an additional $\widetilde{O}(\sqrt{d})$ dependence as discussed by Jin et al. (2020). To overcome this obstacle, we consider decomposing the Bellman error $[\hat{\mathbb{P}}_h - \mathbb{P}_h]\hat{V}_{k,h+1}$ into $[\hat{\mathbb{P}}_h - \mathbb{P}_h]V_{h+1}^*$ as well as $[\hat{\mathbb{P}}_h - \mathbb{P}_h](\hat{V}_{k,h+1} - V_{h+1}^*)$ and bound these two terms separately, which only requires to estimate the variances $[\mathbb{V}_h V_{h+1}^*](s_{k,h}, a_{k,h})$ and $[\mathbb{V}_h(\hat{V}_{k,h+1} - V_{h+1}^*)](s_{k,h}, a_{k,h})$, following Azar, Osband, and Munos (2017); Hu, Chen, and Huang (2022); He et al. (2023).

For the variance $[\mathbb{V}_h V_{h+1}^*](s_{k,h}, a_{k,h})$, with the observation that

$$\begin{aligned}
&[\mathbb{V}_h V_{h+1}^*](s_{k,h}, a_{k,h}) = [\hat{\mathbb{V}}_{k,h}\hat{V}_{k,h+1}](s_{k,h}, a_{k,h}) \\
&\quad + [\mathbb{V}_h V_{h+1}^*](s_{k,h}, a_{k,h}) - [\hat{\mathbb{V}}_{k,h}\hat{V}_{k,h+1}](s_{k,h}, a_{k,h}) \\
&\leq [\hat{\mathbb{V}}_{k,h}\hat{V}_{k,h+1}](s_{k,h}, a_{k,h}) \\
&\quad + \left|[\mathbb{V}_h \hat{V}_{k,h+1} - \mathbb{V}_h V_{h+1}^*](s_{k,h}, a_{k,h})\right| \\
&\quad + \left|[\hat{\mathbb{V}}_{k,h}\hat{V}_{k,h+1} - \mathbb{V}_h \hat{V}_{k,h+1}](s_{k,h}, a_{k,h})\right| ,
\end{aligned}$$

we upper bound $[\mathbb{V}_h V_{h+1}^*](s_{k,h}, a_{k,h})$ by $[\hat{\mathbb{V}}_{k,h}\hat{V}_{k,h+1}](s_{k,h}, a_{k,h}) + E_{k,h} + D_{k,h}$, where $E_{k,h}$ and $D_{k,h}$ serve as the upper bounds of $\left|[\hat{\mathbb{V}}_{k,h}\hat{V}_{k,h+1} - \mathbb{V}_h \hat{V}_{k,h+1}](s_{k,h}, a_{k,h})\right|$ and $\left|[\mathbb{V}_h \hat{V}_{k,h+1} - \mathbb{V}_h V_{h+1}^*](s_{k,h}, a_{k,h})\right|$, respectively (see Appendix for details). In particular, $E_{k,h}$ and $D_{k,h}$ satisfying

$$\begin{aligned}
E_{k,h} =& \min\left\{\tilde{\beta}^P \left\|\boldsymbol{\phi}_{k,h}\right\|_{(\boldsymbol{\Sigma}_{k,h}^P)^{-1}}, H^2\right\} \\
&+ \min\left\{2H\hat{\beta}^P \left\|\boldsymbol{\phi}_{k,h}\right\|_{(\boldsymbol{\Sigma}_{k,h}^P)^{-1}}, H^2\right\}, \\
D_{k,h} =& d^3H^2 \min\left\{4\left[\boldsymbol{\phi}_{k,h}^\top \hat{\boldsymbol{\mu}}_{k,h}\left(\hat{V}_{k,h+1} - \check{V}_{k,h+1}\right)\right.\right. \\
&\left.\left. + 2\hat{\beta}^P \|\boldsymbol{\phi}_{k,h}\|_{(\boldsymbol{\Sigma}_{k,h}^P)^{-1}}\right], H\right\} ,
\end{aligned}$$

where $\hat{V}_{k,h+1}$ and $\check{V}_{k,h+1}$ are the optimistic and pessimistic value function estimates, $\hat{\beta}^P = \widetilde{O}(Hd^{3/2})$ and $\tilde{\beta}^P = \widetilde{O}(H^2 d^{3/2})$ are the Hoeffding-type concentration coefficients for the self-normalized transition error sequences $\|(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_{k,h})\hat{V}_{k,h+1}\|_{\boldsymbol{\Sigma}_{k,h}^P}$ and $\|(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_{k,h})\hat{V}_{k,h+1}^2\|_{\boldsymbol{\Sigma}_{k,h}^P}$ (see Appendix for the detailed values of $\hat{\beta}^P$ and $\tilde{\beta}^P$), respectively.

On the other hand, for the variance $[\mathbb{V}_h(\hat{V}_{k,h+1} - V_{h+1}^*)](s_{k,h}, a_{k,h})$, it can be proven that $D_{k,h}$ above also serves as an upper bound of $[\mathbb{V}_h(\hat{V}_{k,h+1} - V_{h+1}^*)](s_{k,h}, a_{k,h})$. Therefore, the overall variance estimate is taken as

$$\sigma_{k,h} = \sqrt{[\hat{\mathbb{V}}_{k,h}\hat{V}_{k,h+1}](s_{k,h}, a_{k,h}) + E_{k,h} + D_{k,h} + H} . \tag{3}$$

To facilitate a slightly tighter confidence set, following Zhou and Gu (2022); He et al. (2023), our algorithm utilizes an additional uncertainty term $2d^3H^2\|\boldsymbol{\phi}_{k,h}\|_{(\boldsymbol{\Sigma}_{k,h}^P)^{-1}}^{1/2}$ to construct the final variance estimate $\hat{\sigma}_{k,h}$ used in the weighted regression for learning $\boldsymbol{\mu}_h$:

$$\hat{\sigma}_{k,h} = \max\left\{\sigma_{k,h}, H, 2d^3H^2\|\boldsymbol{\phi}_{k,h}\|_{(\boldsymbol{\Sigma}_{k,h}^P)^{-1}}^{1/2}\right\} . \tag{4}$$

Rare-updating of Value Function Estimates Besides, we also apply the determinant-based rare-updating scheme, originally proposed by Abbasi-Yadkori, Pál, and Szepesvári (2011) and also leveraged by He et al. (2023) to control the updates of the optimistic and pessimistic value functions (Line 4 - Line 11), which is critical in bounding the covering number of the value function classes. In particular, both the optimistic and pessimistic state-action value function estimates will be updated only when the determinants of the reward or transition feature covariance matrices are doubled. In this case, they will be updated as

$$\begin{aligned}
\hat{Q}_{k,h}(\cdot,\cdot) = \min\Big\{&\Big[\hat{r}_{k,h}(\cdot,\cdot) + \beta^r \|\boldsymbol{\phi}_{k,h}\|_{(\boldsymbol{\Sigma}_{k,h}^r)^{-1}} \\
&+ \langle \hat{\boldsymbol{\mu}}_{k,h}\hat{V}_{k,h+1}, \boldsymbol{\phi}(\cdot,\cdot)\rangle + \beta^P \|\boldsymbol{\phi}_{k,h}\|_{(\boldsymbol{\Sigma}_{k,h}^P)^{-1}}\Big]_{[-H,H]} , \\
&\hat{Q}_{g(k-1),h}(\cdot,\cdot)\Big\} ,
\end{aligned} \tag{5}$$

and

$$\check{Q}_{k,h}(\cdot,\cdot) = \max\Big\{\Big[\hat{r}_{k,h}(\cdot,\cdot) - \beta^r\|\boldsymbol{\phi}_{k,h}\|_{(\boldsymbol{\Sigma}^r_{k,h})^{-1}} + \langle\hat{\boldsymbol{\mu}}_{k,h}\check{V}_{k,h+1}, \boldsymbol{\phi}(\cdot,\cdot)\rangle - \hat{\beta}^P\|\boldsymbol{\phi}_{k,h}\|_{(\boldsymbol{\Sigma}^P_{k,h})^{-1}}\Big]_{[-H,H]}, \check{Q}_{g(k-1),h}(\cdot,\cdot)\Big\}, \quad (6)$$

where $\hat{V}_{k,h+1}(s) = \max_{a\in\mathcal{A}}\hat{Q}_{k,h+1}(s,a)$ and $\check{V}_{k,h+1}(s) = \max_{a\in\mathcal{A}}\check{Q}_{k,h+1}(s,a)$, and $g(k)$ denotes the index of the latest episode when updates of the value function occur up to episode k.

Algorithm 1: Algorithm for Linear MDPs with Corrupted Rewards

Input: λ^r, λ^P, β^r, $\hat{\beta}^P$, $\tilde{\beta}^P$, β^P, α.
for $k = 1,\ldots,K$ **do**
  Set $\hat{V}_{k,H+1}(s) = \check{V}_{k,H+1}(s) = 0$ for all $s\in\mathcal{S}$.
  if there exists some step $h' \in [H]$ s.t. $\det(\boldsymbol{\Sigma}_{k,h'}) \geq 2\det(\boldsymbol{\Sigma}_{g(k-1),h'})$ or $\det(\boldsymbol{\Sigma}_{k,h'}) \geq 2\det(\boldsymbol{\Sigma}_{g(k-1),h'})$ **then**
    Set $g(k) = k$.
    for $h = H,\ldots,1$ **do**
      Update optimistic and pessimistic value function estimate as in Eq. (5) and Eq. (6) respectively.
    end for
  else
    Set $g(k) = g(k-1)$.
    Set $\hat{Q}_{k,h}(\cdot,\cdot) = \hat{Q}_{k-1,h}(\cdot,\cdot)$ and $\check{Q}_{k,h}(\cdot,\cdot) = \check{Q}_{k-1,h}(\cdot,\cdot)$.
  end if
  for $h = 1,\ldots,H$ **do**
    Take action $a_{k,h} = \arg\max_{a\in\mathcal{A}}\hat{Q}_{k,h}(s_{k,h},a)$, observe reward $\tilde{r}_{k,h}$ and next-state $s_{k,h+1} \sim \mathbb{P}_h(\cdot|s_{k,h},a_{k,h})$.
    Compute empirical variance $\sigma_{k,h}$ as in Eq. (3).
    Set $\hat{\sigma}_{k,h}$ as in Eq. (4).
    Update transition statistics:

$$\boldsymbol{\Sigma}^P_{k+1,h} = \boldsymbol{\Sigma}^P_{k,h} + \hat{\sigma}^{-2}_{k,h}\boldsymbol{\phi}_{k,h}\boldsymbol{\phi}^\top_{k,h},$$
$$\boldsymbol{b}^P_{k+1,h} = \boldsymbol{b}^P_{k,h} + \hat{\sigma}^{-2}_{k,h}\boldsymbol{\phi}_{k,h}\boldsymbol{\delta}^\top_{k,h+1}.$$

    Compute $\hat{\boldsymbol{\mu}}_{k+1,h}$ as in Eq. (2).
    Set reward weight as in Eq. (7).
    Update reward statistics:

$$\boldsymbol{\Sigma}^r_{k+1,h} = \boldsymbol{\Sigma}^r_{k,h} + w_{k,h}\boldsymbol{\phi}_{k,h}\boldsymbol{\phi}^\top_{k,h},$$
$$\boldsymbol{b}^r_{k+1,h} = \boldsymbol{b}^r_{k,h} + w_{k,h}\boldsymbol{\phi}_{k,h}\tilde{r}_{k,h}.$$

    Compute $\hat{\boldsymbol{\theta}}_{k+1,h}$ as in Eq. (8).
  end for
end for

3.2 Reward Parameter Learning

In addition to the weighted regression used to learn the unknown transition parameter, our algorithm incorporates a second weighted regression scheme aimed at learning the unknown reward parameter $\boldsymbol{\theta}_h$, which gives rise to the name DW-LSVI-UCB for our algorithm. In particular, to deal with the potential adversarial corruptions over the rewards, the weight of each experienced state-action pair $(s_{k,h},a_{k,h})$ is inversely proportional to the width of the confidence set. Intuitively, a larger width of the confidence set of the experienced state-action pair $(s_{k,h},a_{k,h})$ indicates that the learning of the unknown reward parameter $\boldsymbol{\theta}_h$ is less accurate along with the direction of $\boldsymbol{\phi}_{k,h}$ and thus there might be larger adversarial corruptions over this state-action pair $(s_{k,h},a_{k,h})$. Consequently, our algorithm will assign a smaller weight in weighted regression to such $(s_{k,h},a_{k,h})$, which eventually minimizes the impacts of the experienced state-action pairs with large adversarial corruptions on the learning of $\boldsymbol{\theta}_h$. A similar design is first proposed by He et al. (2022) to achieve minimax optimal regret for linear contextual bandits with adversarial corruptions and is also subsequently used by Ye et al. (2023a) for learning corrupted contextual bandits and RL with bounded eluder dimension.

Formally, in each episode k, our DW-LSVI-UCB algorithm solves the following weighted regression to obtain an estimate of the unknown reward parameter $\boldsymbol{\theta}_h$:

$$\hat{\boldsymbol{\theta}}_{k,h} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d}\sum_{i=1}^{k-1} w_{i,h}\big(\boldsymbol{\theta}^\top\boldsymbol{\phi}_{i,h} - \tilde{r}_{i,h}\big)^2 + \lambda^r\|\boldsymbol{\theta}\|_2^2,$$

where $\lambda^r > 0$ is the reward regularization parameter and

$$w_{i,h} = \min\Big\{1, \alpha/\|\boldsymbol{\phi}_{i,h}\|_{(\boldsymbol{\Sigma}^r_{i,h})^{-1}}\Big\}, \quad (7)$$

is the regression weight with α as a parameter to be tuned in the analysis. The closed-form solution of $\hat{\boldsymbol{\theta}}_{k,h}$ is given as

$$\hat{\boldsymbol{\theta}}_{k,h} = (\boldsymbol{\Sigma}^r_{k,h})^{-1}\boldsymbol{b}^r_{k,h}, \quad (8)$$

where $\boldsymbol{\Sigma}^r_{k,h} = \sum_{i=1}^{k-1} w_{i,h}\boldsymbol{\phi}_{i,h}\boldsymbol{\phi}^\top_{i,h} + \lambda^r\boldsymbol{I}$ and $\boldsymbol{b}^r_{k,h} = \sum_{i=1}^{k-1} w_{i,h}\boldsymbol{\phi}_{i,h}\tilde{r}_{i,h}$.

4 Analysis

In this section, the theoretical guarantee of our algorithm will be introduced first, followed by the proof sketch and the technical challenges in the proof.

4.1 Main Results

The regret of Algorithm 1 is guaranteed by the following theorem.

Theorem 4.1. *For any linear MDP $\mathcal{M}$ satisfying Definition 2.2, if the total amount of corruption at each step h is bounded by $C \geq 0$ and C is known, then by setting $\beta^r = O(\sqrt{d\log((1+K/\lambda^r)/\delta)} + \sqrt{\lambda^r} + \alpha C)$, $\alpha = (\sqrt{d}+\sqrt{\lambda^r})/C$, $\hat{\beta}^P$, $\tilde{\beta}^P$, and $\beta^P = \beta^{P,(1)} + \beta^{P,(2)}$ specified in Appendix, $\lambda^r = 1$ and $\lambda^P = 1/H^2$, with probability $1-((K+6)H\lceil\log{}^H/_{\mathrm{gap}_{\min}}\rceil + \log K)\delta$, the gap-dependent regret of DW-LSVI-UCB is upper bounded as*

$$\widetilde{O}\left(\frac{d^2H^4\log^2_+(K/\delta)}{\mathrm{gap}_{\min}}\log^2\left(\frac{\log^2_+(K/\delta)}{\mathrm{gap}^2_{\min}}\right)\right.$$

$$+ CdH^2\sqrt{\log_+(K/\delta)}\log^2\left(\frac{\sqrt{\log_+(K/\delta)}}{\mathrm{gap}_{\min}}\right)$$
$$+ d^4H^9 + d^{7.5}H^6\log_+(K/\delta)\log^2\left(\frac{\log_+(K/\delta)}{\mathrm{gap}_{\min}}\right)$$
$$+\sqrt{(d^{13.5}H^{14}+Cd^6H^7)}\log_+(K/\delta)\log\frac{\log_+^2(K/\delta)}{\mathrm{gap}_{\min}^2}\Bigg),$$

where $\widetilde{O}(\cdot)$ suppresses all logarithmic terms expect for $\log_+(K/\delta)$ and $\log(1/\mathrm{gap}_{\min})$.

Remark 4.1. *Further ignoring all the log-logarithmic terms, the gap-dependent regret of Algorithm 1 is $\widetilde{O}(\frac{d^2H^4\log_+^2(K/\delta)}{\mathrm{gap}_{\min}} + CdH^2\sqrt{\log_+(K/\delta)} + (d^{7.5}H^6 + \sqrt{(d^{13.5}H^{14}+Cd^6H^7)})\log_+(K/\delta) + d^4H^9)$, where the corruption level C is not multiplicative to the leading term $\widetilde{O}(\frac{d^2H^4\log_+^2(K/\delta)}{\mathrm{gap}_{\min}})$; instead, it only appears in the lower-order terms $\widetilde{O}(CdH^2\sqrt{\log_+(K/\delta)} + \sqrt{(d^{13.5}H^{14}+Cd^6H^7)}\log_+(K/\delta)$. Besides, when there are no adversarial corruptions over rewards (i.e., corruption level $C = 0$), the regret of our DW-LSVI-UCB algorithm improves the previous best result of He, Zhou, and Gu (2021a) by an $\widetilde{O}(dH/\log K)$ factor. Please see Appendix for additional discussions on the results.*

Corollary 4.1. *For any linear MDP $\mathcal{M}$, if the total corruption level C is unknown and we have an estimated corruption level $\hat{C}$ such that $\hat{C} \geq C \geq 0$, by setting $\alpha = (\sqrt{d}+\sqrt{\lambda^r})/\hat{C}$ and $\beta^r = O(\sqrt{d\log\left((1+K/\lambda^r)/\delta\right)}+\sqrt{\lambda^r})$ and all other parameters the same as in Theorem 4.1, with probability $1-((K+6)H\lceil\log {}^H/_{\mathrm{gap}_{\min}}\rceil + \log K)\delta$, the gap-dependent regret of DW-LSVI-UCB is upper bounded as*

$$\widetilde{O}\Bigg(\frac{d^2H^4\log_+^2(K/\delta)}{\mathrm{gap}_{\min}}\log^2\left(\frac{\log_+^2(K/\delta)}{\mathrm{gap}_{\min}^2}\right)$$
$$+ \hat{C}dH^2\sqrt{\log_+(K/\delta)}\log^2\left(\frac{\sqrt{\log_+(K/\delta)}}{\mathrm{gap}_{\min}}\right)$$
$$+ d^4H^9 + d^{7.5}H^6\log(1+K/\delta)\log^2\left(\frac{\log_+(K/\delta)}{\mathrm{gap}_{\min}}\right)$$
$$+\sqrt{(d^{13.5}H^{14}+\hat{C}d^6H^7)}\log_+(K/\delta)\log\frac{\log_+^2(K/\delta)}{\mathrm{gap}_{\min}^2}\Bigg).$$

Remark 4.2. *When the upper bound of C is unknown and $C > \hat{C}$, our DW-LSVI-UCB algorithm fails to learn the corrupted linear MDPs and incurs a linear gap-independent regret. However, note that for a given class $\mathscr{A}$ of algorithms for learning uncorrupted linear MDPs with true regret Reg_K satisfying $\mathrm{Reg}_K \leq \mathrm{Reg}'_K \leq O(K)$ for some Reg'_K, DW-LSVI-UCB can also enjoy the Reg'_K regret by simply setting $\hat{C} = O(\mathrm{Reg}'_K/(dH))$. Further, we can show that when C is unknown and $C \geq \Omega(\mathrm{Reg}'_K/(dH))$, all the algorithms in this class suffer linear (expected) regret, following similar reasoning in He et al. (2022). This indicates that when C is unknown and $C \geq \Omega(\mathrm{Reg}'_K/(dH))$, the corrupted linear MDPs are not learnable for all the algorithms belong to class $\mathscr{A}$ including DW-LSVI-UCB .*

4.2 Proof Sketch

We now present the proof sketch of Theorem 4.1, along with the key technical challenges and discussions on how we overcome them.

To begin with, the following lemma guarantees that the estimation error of the true unknown reward parameter $\boldsymbol{\theta}_h$ is upper bounded by the reward exploration bonus.

Lemma 4.1 (Lemma 4.1, He et al. (2022))**.** *Fix $h \in [H]$. For any $0 < \delta < 1$ and known corruption budget $C \geq 0$, set the reward confidence radius $\beta^r = \sqrt{d\log\left((1+K/\lambda^r)/\delta\right)} + \sqrt{\lambda^r} + \alpha C$, then with probability at least $1-\delta$, for each episode k, the estimator $\hat{\boldsymbol{\theta}}_{k,h}$ satisfies that $\|\hat{\boldsymbol{\theta}}_{k,h} - \boldsymbol{\theta}_h\|_{\boldsymbol{\Sigma}^r_{k,h}} \leq \beta^r$.*

For the concentration of the transition parameter $\boldsymbol{\mu}_h$, we first introduce the following Hoeffding-type concentrations:

$$\hat{\mathcal{C}}_{k,h} = \left\{\boldsymbol{\mu} \in \mathbb{R}^{d\times S} : \|(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_{k,h})\hat{V}_{k,h+1}\|_{\boldsymbol{\Sigma}^P_{k,h}} \leq \hat{\beta}^P\right\},$$
$$\tilde{\mathcal{C}}_{k,h} = \left\{\boldsymbol{\mu} \in \mathbb{R}^{d\times S} : \|(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_{k,h})\hat{V}^2_{k,h+1}\|_{\boldsymbol{\Sigma}^P_{k,h}} \leqslant \tilde{\beta}^P\right\},$$
$$\check{\mathcal{C}}_{k,h} = \left\{\boldsymbol{\mu} \in \mathbb{R}^{d\times S} : \|(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_{k,h})\check{V}_{k,h+1}\|_{\boldsymbol{\Sigma}^P_{k,h}} \leqslant \hat{\beta}^P\right\},$$

where the values of $\hat{\beta}^P$ and $\tilde{\beta}^P$ are specified in Appendix.

With high probability, the following lemma guarantees that $\boldsymbol{\mu}_h$ lies in the above concentration sets. It will be useful for constructing the sharper Bernstein-type concentration sets and its proof is deferred to Appendix.

Lemma 4.2. *Define $\mathcal{E}$ as the event that $\boldsymbol{\mu}_h \in \hat{\mathcal{C}}_{k,h}\bigcap\tilde{\mathcal{C}}_{k,h}\bigcap\check{\mathcal{C}}_{k,h}$ for all $(k,h) \in [K]\times[H]$. Then $\mathcal{E}$ happens with probability $1-3\delta$.*

Based on the Hoeffding-type concentration above, we then construct the following Bernstein-type confidence set:

$$\mathcal{C}_{k,h} = \left\{\boldsymbol{\mu} \in \mathbb{R}^{d\times S} : \|(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_{k,h})\hat{V}_{k,h+1}\|_{\boldsymbol{\Sigma}^P_{k,h}} \leq \beta^P\right\},$$

where $\beta^P := \beta^{P,(1)} + \beta^{P,(1)}$ with $\beta^{P,(1)}$ and $\beta^{P,(1)}$ quantified in Appendix. Define $\tilde{\mathcal{E}}_h$ as the event that $\boldsymbol{\mu}_{h'} \in \mathcal{C}_{k,h'}$ for all $k \in [K]$ and $h' \in \{h, h+1, \ldots, H\}$. Let $\tilde{\mathcal{E}} = \tilde{\mathcal{E}}_1$. Then the following lemma guarantees that the true transition parameter $\boldsymbol{\mu}_h$ is contained in the constructed Bernstein-type confidence set with high probability. Its proof is postponed to Appendix.

Lemma 4.3. *Conditioned on event $\mathcal{E}$, event $\tilde{\mathcal{E}}$ holds with probability at least $1-2\delta$.*

Moreover, by the standard performance difference lemma (Kakade and Langford 2002), one can see that $\mathrm{Reg}_K = O(\sum_{k=1}^K\sum_{h=1}^H \mathrm{gap}_h(s_{k,h},a_{k,h}))$, where recall $\mathrm{gap}_h(s,a) = V_h^*(s) - Q_h^*(s,a)$. Then, to obtain the final gap-dependent regret, it suffices to bound the summation of the $\mathrm{gap}_h(s_{k,h},a_{k,h})$'s in the above display. To this end, we consider dividing the range $[\mathrm{gap}_{\min}, H]$

where all the $\mathrm{gap}_h(s_{k,h}, a_{k,h})$'s lie into several subintervals, and upper bounding the number of $\mathrm{gap}_h(s_{k,h}, a_{k,h})$'s that fall into each subinterval of $[\mathrm{gap}_{\min}, H]$ following Yang, Yang, and Du (2021); He, Zhou, and Gu (2021a). Specifically, we first divide the whole range $[\mathrm{gap}_{\min}, H]$ into $\lceil \log {}^{H}/_{\mathrm{gap}_{\min}} \rceil$ subintervals, with the i-th subinterval being as $\left[2^{i-1}\,\mathrm{gap}_{\min}, 2^i\,\mathrm{gap}_{\min}\right)$. Then with the observation that

$$\begin{aligned}\mathrm{gap}_h(s_{k,h}, a_{k,h}) =& V_h^*(s_{k,h}) - Q_h^*(s_{k,h}, a_{k,h})\\ =& Q_h^*(s_{k,h}, \pi^*(s_{k,h})) - Q_h^*(s_{k,h}, a_{k,h})\\ \le& \hat{Q}_{k,h}(s_{k,h}, \pi^*(s_{k,h})) - Q_h^{\pi_k}(s_{k,h}, a_{k,h})\\ \le& \hat{Q}_{k,h}(s_{k,h}, a_{k,h}) - Q_h^{\pi_k}(s_{k,h}, a_{k,h}),\end{aligned}$$

where the first inequality is due to the optimism of the constructed estimated state-action value function $\hat{Q}_{k,h}$ guaranteed by Lemma 4.1 and 4.3. Therefore, it turns out that the total number of $\mathrm{gap}_h(s_{k,h}, a_{k,h})$'s that fall into the i-th subinterval $\left[2^{i-1}\,\mathrm{gap}_{\min}, 2^i\,\mathrm{gap}_{\min}\right)$ is naturally upper bounded by the estimation errors of the state-action value functions in these episodes. This is demonstrated in the following key technical lemma (please see Appendix for its formal statement).

Lemma 4.4. *Let $N := \lceil \log {}^{H}/_{\mathrm{gap}_{\min}} \rceil$. Conditioned on the events in Lemma 4.1 and 4.3, for any fixed $h \in [H]$ and $i \in [N]$, with probability at least $1 - K\delta$, we have*

$$\begin{aligned}&\sum_{k=1}^{K} \mathbb{1}\{\mathrm{gap}_h(s_{k,h}, a_{k,h}) \ge 2^i\,\mathrm{gap}_{\min}\}\\ &= \widetilde{O}\left(\frac{d^2H^3\log_+^2(K/\delta)}{4^i\,\mathrm{gap}_{\min}^2}\log^2\left(\frac{d^2H^3\log_+^2(K/\delta)}{4^i\,\mathrm{gap}_{\min}^2}\right)\right.\\ &\left.+\frac{CdH\sqrt{\log_+(K/\delta)}}{2^i\,\mathrm{gap}_{\min}}\log^2\left(\frac{CdH\sqrt{\log_+(K/\delta)}}{2^i\,\mathrm{gap}_{\min}}\right)\right).\end{aligned}$$

The proof of Lemma 4.4 is inspired by Yang, Yang, and Du (2021), which study the gap-dependent regret of the canonical Q-learning for tabular MDPs. Similar reasoning has also been adopted by He, Zhou, and Gu (2021a) for achieving the gap-dependent regret for uncorrupted linear (mixture) MDPs. However, note that our proof adopts a refined analysis to further sharpen the dependence on $\log K$ in the final results. In particular, fix a step $h \in [H]$ and an index of the subinterval $i \in [N]$. Denote by $m = \sum_{k=1}^{K} \mathbb{1}\{\mathrm{gap}_h(s_{k,h}, a_{k,h}) \ge 2^i\,\mathrm{gap}_{\min}\}$. Then a straightforward analysis will deduce a quadratic equation regarding m as follows:

$$\begin{aligned}&m\cdot 2^i\,\mathrm{gap}_{\min} = \widetilde{O}\left(\left[CdH + d^4H^8 + d^{7.5}H^5\right.\right.\\ &\left.\left.+d\sqrt{H(mH^2 + d^{11.5}H^{11} + Cd^4H^4)}\right]\log_+^2(K/\delta)\right).\end{aligned}$$

Solving this quadratic equation can only lead to

$$\begin{aligned}m = \widetilde{O}\left(\frac{(CdH + d^4H^8 + d^{7.5}H^5)\log_+^2(K/\delta)}{2^i\,\mathrm{gap}_{\min}}\right.\\ \left.+\frac{d^2H^3\log_+^4(K/\delta)}{4^i\,\mathrm{gap}_{\min}^2} + \frac{\sqrt{d^{13.5}H^{12}\log_+^4(K/\delta)}}{2^i\,\mathrm{gap}_{\min}}\right.\\ \left.+\sqrt{Cd^6H^5\log_+^4(K/\delta)}/\left(2^i\,\mathrm{gap}_{\min}\right)\right).\end{aligned}$$

Multiplying such m with the maximum value $2^i\,\mathrm{gap}_{\min}$ of this subinterval and taking summation over i will inevitably lead to a $\log_+^4(K)/\,\mathrm{gap}_{\min}$ dependence of the leading term in the final regret bound.

To overcome this obstacle, we establish a more refined analysis to upper bound m. In specific, we first upper bound $m\cdot 2^i\,\mathrm{gap}_{\min}$ as

$$\begin{aligned}m =& \widetilde{O}\left(d^4H^8\log_+(m) + d^{7.5}H^5\log_+(K/\delta)\log_+(m)\right.\\ &+ d\sqrt{H(d^{11.5}H^{11} + Cd^4H^4)}\log_+(K/\delta)\log_+^{1/2}(m)\\ &+ d\sqrt{H(mH^2 + d^3H^{4.5}m^{1/2})}\log_+(K/\delta)\log_+^{1/2}(m)\\ &\left.+\beta^r H\sqrt{md\log_+(m)} + \beta^r Hd\log_+(m)/\alpha\right). \qquad (9)\end{aligned}$$

Further, we prove that for any two large enough positive numbers x and y satisfying $x/\log x \le y$, it holds that $x = O(y\log^2 y)$. With this observation, we delicately upper bound m using each term in the RHS of Eq. (9), which eventually shaves off a multiplicative $\log_+^2(K)$ factor in the leading term of the regret upper bound (please see Appendix for details).

Discussions on the Optimal Dependence on K As mentioned in Remark 4.1, the leading term in our regret upper bound currently has a $\log^2 K$ dependence. The extra $\log K$ dependence comes from the fact the confidence radius of the Bernstein-type concentration for vector-valued martingales (see, *e.g.*, Theorem 2 of Zhou, Gu, and Szepesvári (2021); Theorem 4.3 of Zhou and Gu (2022)) has an additional $\sqrt{\log K}$ dependence compared with the canonical Hoeffding-type concentrations (see, *e.g.*, Lemma 11 of Abbasi-Yadkori, Pál, and Szepesvári (2011)), though the former has the advantage of being adaptive to the variance of the noises. Currently, it is highly unclear whether it is feasible to establish a Bernstein-type concentration for vector-valued martingales with the confidence radius having the same $\log K$ dependence as that of the Hoeffding-type ones. We leave this interesting and challenging extension as our future study.

5 Conclusion

In this work, we propose the first algorithm for learning corrupted linear MDPs with provable $\widetilde{O}(\frac{d^2H^4\log_+^2(K/\delta)}{\mathrm{gap}_{\min}} + CdH^2)$ gap-dependent regret. When there are no corruptions over reward functions, this result also improves the previous best-known gap-dependent regret for linear MDPs by an $\widetilde{O}(dH/\log K)$ factor. We believe our results might shed light on better understandings of how to adapt to mild environments with fine-grained instance-dependent structures in RL with large-scale state-action spaces, while being robust to adversarial corruptions. One remaining question may be whether it is possible to obtain the optimal $\log K$ dependence for our problem. The other natural question might be extending the proposed algorithm and results to corrupted RL with general function approximation. We leave these extensions as our future directions.

Acknowledgments

The corresponding author Shuai Li is supported by National Science and Technology Major Project (2022ZD0114804) and is partly supported by the Guangdong Provincial Key Laboratory of Mathematical Foundations for Artificial Intelligence (2023B1212010001). Baoxiang Wang is partially supported by the National Natural Science Foundation of China (62106213, 72394361) and an extended support project from the Shenzhen Science and Technology Program.

References

Abbasi-Yadkori, Y.; Pál, D.; and Szepesvári, C. 2011. Improved Algorithms for Linear Stochastic Bandits. In *NeurIPS*.

Ayoub, A.; Jia, Z.; Szepesvári, C.; Wang, M.; and Yang, L. 2020. Model-Based Reinforcement Learning with Value-Targeted Regression. In *ICML*.

Azar, M. G.; Munos, R.; and Kappen, H. J. 2013. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Mach. Learn.*

Azar, M. G.; Osband, I.; and Munos, R. 2017. Minimax Regret Bounds for Reinforcement Learning. In *ICML*.

Cai, Q.; Yang, Z.; Jin, C.; and Wang, Z. 2020. Provably efficient exploration in policy optimization. In *ICML*. PMLR.

Chen, Y.; Zhang, X.; Zhang, K.; Wang, M.; and Zhu, X. 2023. Byzantine-Robust Online and Offline Distributed Reinforcement Learning. In *AISTATS*.

Dai, Y.; Luo, H.; Wei, C.-Y.; and Zimmert, J. 2023. Refined regret for adversarial mdps with linear function approximation. In *ICML*. PMLR.

Dann, C.; Jiang, N.; Krishnamurthy, A.; Agarwal, A.; Langford, J.; and Schapire, R. E. 2018. On Oracle-Efficient PAC RL with Rich Observations. In *NeurIPS*.

Dann, C.; Wei, C.-Y.; and Zimmert, J. 2023. A blackbox approach to best of both worlds in bandits and beyond. In *COLT*. PMLR.

Du, S. S.; Kakade, S. M.; Lee, J. D.; Lovett, S.; Mahajan, G.; Sun, W.; and Wang, R. 2021. Bilinear Classes: A Structural Framework for Provable Generalization in RL. In *ICML*.

Du, S. S.; Kakade, S. M.; Wang, R.; and Yang, L. F. 2020. Is a Good Representation Sufficient for Sample Efficient Reinforcement Learning? In *ICLR*.

Du, S. S.; Krishnamurthy, A.; Jiang, N.; Agarwal, A.; Dudík, M.; and Langford, J. 2019a. Provably efficient RL with Rich Observations via Latent State Decoding. In *ICML*.

Du, S. S.; Luo, Y.; Wang, R.; and Zhang, H. 2019b. Provably Efficient Q-learning with Function Approximation via Distribution Shift Error Checking Oracle. In *NeurIPS*.

Feinberg, A. 1996. Markov Decision Processes: Discrete Stochastic Dynamic Programming (Martin L. Puterman). *SIAM Rev.*

Garcelon, E.; Rozière, B.; Meunier, L.; Tarbouriech, J.; Teytaud, O.; Lazaric, A.; and Pirotta, M. 2020. Adversarial Attacks on Linear Contextual Bandits. In *NeurIPS*.

Gupta, A.; Koren, T.; and Talwar, K. 2019. Better Algorithms for Stochastic Bandits with Adversarial Corruptions. In *COLT*.

He, J.; Zhao, H.; Zhou, D.; and Gu, Q. 2023. Nearly Minimax Optimal Reinforcement Learning for Linear Markov Decision Processes. In *ICML*.

He, J.; Zhou, D.; and Gu, Q. 2021a. Logarithmic Regret for Reinforcement Learning with Linear Function Approximation. In *ICML*.

He, J.; Zhou, D.; and Gu, Q. 2021b. Nearly Minimax Optimal Reinforcement Learning for Discounted MDPs. In *NeurIPS*.

He, J.; Zhou, D.; and Gu, Q. 2022. Near-optimal policy optimization algorithms for learning adversarial linear mixture mdps. In *AISTATS*. PMLR.

He, J.; Zhou, D.; Zhang, T.; and Gu, Q. 2022. Nearly Optimal Algorithms for Linear Contextual Bandits with Adversarial Corruptions. In *NeurIPS*.

Hu, P.; Chen, Y.; and Huang, L. 2022. Nearly Minimax Optimal Reinforcement Learning with Linear Function Approximation. In *ICML*.

Ito, S. 2021a. On Optimal Robustness to Adversarial Corruption in Online Decision Problems. In *NeurIPS*.

Ito, S. 2021b. Parameter-Free Multi-Armed Bandit Algorithms with Hybrid Data-Dependent Regret Bounds. In *COLT*.

Ito, S.; and Takemura, K. 2023. Best-of-three-worlds linear bandit algorithm with variance-adaptive regret bounds. In *COLT*. PMLR.

Ito, S.; Tsuchiya, T.; and Honda, J. 2022. Nearly optimal best-of-both-worlds algorithms for online learning with feedback graphs. *NeurIPS*, 35.

Ito, S.; Tsuchiya, T.; and Honda, J. 2024. Adaptive learning rate for follow-the-regularized-leader: Competitive analysis and best-of-both-worlds. In *COLT*. PMLR.

Jiang, N.; Krishnamurthy, A.; Agarwal, A.; Langford, J.; and Schapire, R. E. 2017. Contextual Decision Processes with low Bellman rank are PAC-Learnable. In *ICML*.

Jin, C.; Liu, Q.; and Miryoosefi, S. 2021. Bellman Eluder Dimension: New Rich Classes of RL Problems, and Sample-Efficient Algorithms. In *NeurIPS*.

Jin, C.; Yang, Z.; Wang, Z.; and Jordan, M. I. 2020. Provably efficient reinforcement learning with linear function approximation. In *COLT*.

Jin, T.; Huang, L.; and Luo, H. 2021. The best of both worlds: stochastic and adversarial episodic MDPs with unknown transition. In *NeurIPS*.

Jin, T.; and Luo, H. 2020. Simultaneously Learning Stochastic and Adversarial Episodic MDPs with Known Transition. In *NeurIPS*.

Kakade, S. M.; and Langford, J. 2002. Approximately Optimal Approximate Reinforcement Learning. In *ICML*.

Kong, F.; Zhang, X.; Wang, B.; and Li, S. 2024. Improved Regret Bounds for Linear Adversarial MDPs via Linear Optimization. *Transactions on Machine Learning Research*.

Kong, F.; Zhao, C.; and Li, S. 2023. Best-of-three-worlds analysis for linear bandits with follow-the-regularized-leader algorithm. In *COLT*. PMLR.

Kong, F.; Zhou, Y.; and Li, S. 2022. Simultaneously learning stochastic and adversarial bandits with general graph feedback. In *ICML*, 11473–11482. PMLR.

Lattimore, T.; and Szepesvári, C. 2020. *Bandit algorithms*. Cambridge University Press.

Li, L.-F.; Zhao, P.; and Zhou, Z.-H. 2024. Improved algorithm for adversarial linear mixture MDPs with bandit feedback and unknown transition. In *AISTATS*. PMLR.

Lu, S.; Wang, G.; and Zhang, L. 2021. Stochastic Graphical Bandits with Adversarial Corruptions. In *AAAI*.

Luo, H.; Wei, C.-Y.; and Lee, C.-W. 2021. Policy optimization in adversarial mdps: Improved exploration via dilated bonuses. *NeurIPS*, 34.

Lykouris, T.; Mirrokni, V. S.; and Leme, R. P. 2018. Stochastic bandits robust to adversarial corruptions. In *STOC*.

Lykouris, T.; Simchowitz, M.; Slivkins, A.; and Sun, W. 2021. Corruption-robust exploration in episodic reinforcement learning. In *COLT*.

Min, Y.; He, J.; Wang, T.; and Gu, Q. 2022. Learning Stochastic Shortest Path with Linear Function Approximation. In *ICML*.

Sherman, U.; Koren, T.; and Mansour, Y. 2023. Improved regret for efficient online reinforcement learning with linear function approximation. In *ICML*. PMLR.

Simchowitz, M.; and Jamieson, K. G. 2019. Non-Asymptotic Gap-Dependent Regret Bounds for Tabular MDPs. In *NeurIPS*.

Sun, W.; Jiang, N.; Krishnamurthy, A.; Agarwal, A.; and Langford, J. 2019. Model-based RL in Contextual Decision Processes: PAC bounds and Exponential Improvements over Model-free Approaches. In *COLT*.

Tossou, A. C. Y.; Basu, D.; and Dimitrakakis, C. 2019. Near-optimal Optimistic Reinforcement Learning using Empirical Bernstein Inequalities. abs/1905.12425.

Uehara, M.; Zhang, X.; and Sun, W. 2022. Representation Learning for Online and Offline RL in Low-rank MDPs. In *ICLR*.

Wang, R.; Salakhutdinov, R.; and Yang, L. F. 2020. Reinforcement Learning with General Value Function Approximation: Provably Efficient Approach via Bounded Eluder Dimension. In *NeurIPS*.

Wang, Y.; Wang, R.; Du, S. S.; and Krishnamurthy, A. 2021. Optimism in Reinforcement Learning with Generalized Linear Function Approximation. In *ICLR*.

Wei, C.; Dann, C.; and Zimmert, J. 2022. A Model Selection Approach for Corruption Robust Reinforcement Learning. In *ALT*.

Wu, T.; Yang, Y.; Du, S. S.; and Wang, L. 2021. On Reinforcement Learning with Adversarial Corruption and Its Application to Block MDP. In *ICML*.

Wu, Y.; Zhou, D.; and Gu, Q. 2022. Nearly Minimax Optimal Regret for Learning Infinite-horizon Average-reward MDPs with Linear Function Approximation. In *AISTATS*.

Yang, K.; Yang, L. F.; and Du, S. S. 2021. Q-learning with Logarithmic Regret. In *AISTATS*.

Yang, L.; Hajiesmaili, M. H.; Talebi, M. S.; Lui, J. C. S.; and Wong, W. S. 2020. Adversarial Bandits with Corruptions: Regret Lower Bound and No-regret Algorithm. In *NeurIPS*.

Yang, L.; and Wang, M. 2019. Sample-Optimal Parametric Q-Learning Using Linearly Additive Features. In *ICML*.

Ye, C.; He, J.; Gu, Q.; and Zhang, T. 2024. Towards Robust Model-Based Reinforcement Learning Against Adversarial Corruption. abs/2402.08991.

Ye, C.; Xiong, W.; Gu, Q.; and Zhang, T. 2023a. Corruption-Robust Algorithms with Uncertainty Weighting for Nonlinear Contextual Bandits and Markov Decision Processes. In *ICML*.

Ye, C.; Yang, R.; Gu, Q.; and Zhang, T. 2023b. Corruption-Robust Offline Reinforcement Learning with General Function Approximation. In *NeurIPS*.

Zanette, A.; Brandfonbrener, D.; Brunskill, E.; Pirotta, M.; and Lazaric, A. 2020. Frequentist Regret Bounds for Randomized Least-Squares Value Iteration. In *AISTATS*.

Zhang, X.; Ma, Y.; Singla, A.; and Zhu, X. 2020. Adaptive Reward-Poisoning Attacks against Reinforcement Learning. In *ICML*.

Zhang, Z.; Yang, J.; Ji, X.; and Du, S. S. 2021. Improved Variance-Aware Confidence Sets for Linear Bandits and Linear Mixture MDP. In *NeurIPS*.

Zhao, C.; Yang, R.; Wang, B.; and Li, S. 2023. Learning adversarial linear mixture markov decision processes with bandit feedback and unknown transition. In *ICLR*.

Zhong, H.; and Zhang, T. 2024. A theoretical analysis of optimistic proximal policy optimization in linear markov decision processes. *NeurIPS*, 36.

Zhong, Z.; Cheung, W. C.; and Tan, V. Y. F. 2021. Probabilistic Sequential Shrinking: A Best Arm Identification Algorithm for Stochastic Bandits with Corruptions. In *ICML*.

Zhou, D.; and Gu, Q. 2022. Computationally efficient horizon-free reinforcement learning for linear mixture mdps. *NeurIPS*.

Zhou, D.; Gu, Q.; and Szepesvári, C. 2021. Nearly Minimax Optimal Reinforcement Learning for Linear Mixture Markov Decision Processes. In *COLT*.

Zimmert, J.; and Seldin, Y. 2021. Tsallis-INF: An Optimal Algorithm for Stochastic and Adversarial Bandits. *J. Mach. Learn. Res.*