A Doubly Robust Approach to Sparse Reinforcement Learning

Wonyoung Kim
Columbia University

Garud Iyengar
Columbia University

Assaf Zeevi
Columbia University

Abstract

We propose a new regret minimization algorithm for episodic sparse linear Markov decision process (SMDP) where the state-transition distribution is a linear function of observed features. The only previously known algorithm for SMDP requires the knowledge of the sparsity parameter and oracle access to an unknown policy. We overcome these limitations by combining the doubly robust method that allows one to use feature vectors of *all* actions with a novel analysis technique that enables the algorithm to use data from all periods in all episodes. The regret of the proposed algorithm is $\tilde{O}(\sigma_{\min}^{-1} s_{\star} H\sqrt{N})$, where $\sigma_{\min}$ denotes the restrictive the minimum eigenvalue of the average Gram matrix of feature vectors, $s_{\star}$ is the sparsity parameter, H is the length of an episode, and N is the number of rounds. We provide a lower regret bound that matches the upper bound to logarithmic factors on a newly identified subclass of SMDPs. Our numerical experiments support our theoretical results and demonstrate the superior performance of our algorithm.

1 INTRODUCTION

The goal of reinforcement learning (RL) is to maximize the cumulative expected reward while simultaneously learning the unknown transition structure of the underlying Markov decision process (MDP). RL has been applied to robotics (Kober et al., 2013), human-level game plays (Mnih et al., 2013; Silver et al., 2016), dialogue systems (Li et al., 2016), among others (Barto et al., 2017). Modern RL applications have exponentially large, possibly infinite state space, and therefore tabular RL (Auer et al., 2008; Osband et al., 2016; Azar et al., 2017; Dann et al., 2017; Jin et al., 2018; Strehl et al., 2006) is intractable, and value function approximation is essential.

RL with deep neural networks-based value function approximation had empirical success in a variety of settings with high dimensional state and action spaces (Mnih et al., 2013, 2016; Schulman et al., 2015). However, providing theoretical guarantees for these methods has been challenging because, in the high dimensional setting, most states are not visited even once during a set of learning episode (Sutton and Barto, 2018; Szepesvári, 2022) Consequently, there was an effort to establish convergence results when the true (unknown) value function is assumed to be a linear function of d features (Hu et al., 2022; Zhou et al., 2021; He et al., 2021a,b). Extensions to the setting where the value function is within a prescribed distance from a linear function (Cai et al., 2020; Jin et al., 2023; Ayoub et al., 2020; Zanette et al., 2020), and to settings the state transition, and therefore, the value function, is a sparse linear function of the features, i.e., a linear function of $s_{\star} \ll d$ features (Jiang et al., 2017; Sun et al., 2019; Agarwal et al., 2020; Hao et al., 2021b). The latter class of problems includes low-rank MDPs and sparse linear Markov decision processes (SMDPs). It is known that when humans play video games, their strategy depends on a few significant pixels. Once the set of informative features is identified, a policy linear in these features is often optimal (Barto et al., 2017). The SMDP approach provides flexibility over linear function approximation since one is now allowed to consider a much larger set $d \gg 1$ of features and select only the $s_{\star} \ll d$ informative features. In the SMDP setup, we jointly estimate the informative features and a linearly parameterized value function that is close to optimal.

Most of the prior work using sparse linear approximation is for offline RL, and the extension to online RL has remained challenging. Hao et al. (2021b) proposed the first online algorithm for RL with sparse linear approximation with a regret bound that is logarithmic in the number of features d. However, their algorithm

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

requires knowledge of the number of informative features $s_\star$, and needs oracle access to an exploratory policy, and, unfortunately, identifying such a policy is as hard as designing the learning algorithm. Therefore, the problem of designing a practical online algorithm for RL with sparse linear function approximation remains open.

Our main contributions are as follows.

(a) Online RL with bandit feedback is hard because only samples for the Q-value function of the selected actions are observed. We propose a novel algorithm that leverages a technique called doubly robust (DR) estimation to impute values for the Q-values for unselected actions (Section 4). The estimator proposed by Hao et al. (2021b) requires oracle access to an exploratory policy in order to guarantee that Q-value estimates converge to the true Q-values. In contrast, our estimator converges to the optimal Q-value function without the oracle access to an exploratory policy, which is possible using features of *all* actions.

(b) We develop a new analysis technique that carefully accounts for the dependence between Q-value estimates for different periods h, which allows us to use data from all H periods and all n episodes to estimate the Q-value function of a time-homogeneous MDP (Section 5.1). In contrast, previous methods partition the episodes into H groups and use the h-th group to estimate the Q-value function for period h. Thus, our estimation method increases the number of effective samples from n/H to n.

(c) We leverage our estimation to propose new algorithm `RDRLVI` for homogeneous SMDPs whose regret is $\tilde{O}(\sigma_{\min}^{-1} s_\star H\sqrt{N})$ (Theorem 5.5), where $\sigma_{\min}$ is the restrictive minimum eigenvalue defined in Definition 3.2 and $s_\star$ is the sparsity parameter defined in Definition 3.1. `RDRLVI` does not require knowledge of the number of informative features $s_\star$ and does not need oracle access to an exploratory policy, and yet the regret bound that is logarithm in the number of features d.

(d) We provide a novel lower bound on the regret for SMDPs (Theorem 3.2). We show that the lower bound critically depends on $\sigma_{\min}$. For SMDP instances with $\sigma_{\min}^2 \geq s^\star/d$ we show that the regret of our proposed algorithm is tight to within logarithmic factors; whereas when $\sigma_{\min}^2 \leq s^\star/d$ there is a gap that needs to addressed. This result is an improvement and an extension of the lower bound results for sparse linear bandits.

(e) The results of our numerical experiments demonstrate the superior performance of the proposed algorithm over the previously known algorithms. The results empirically verify the dependence of regret on $\sigma_{\min}$, and that the regret is almost independent of the dimension of the feature vector d.

2 RELATED WORK

Function approximation MDP is introduced by Sutton (1988); Tsitsiklis and Van Roy (1996) and Bradtke and Barto (1996). For inhomogeneous episodic MDP, Hu et al. (2022) and Zhou et al. (2021) proved an $\tilde{O}(dH^{3/2}\sqrt{N})$ regret bound with a nearly matching lower bound when the optimal value function is assumed to be a linear function of the features, and He et al. (2021a) established a logarithmic regret bound when there is a positive sub-optimality gap. Jin et al. (2023); Ayoub et al. (2020) and Zanette et al. (2020) established a regret bound when the true value function is within a prescribed distance from a linear function. For offline RL, Jiang et al. (2017) and Sun et al. (2019) considered a larger class of MDPs that have, respectively, low Bellman rank and witness rank. Agarwal et al. (2020) introduced the low-rank MDP setting where the algorithm chooses a low-dimensional feature from a certain function class. Kolter and Ng (2009); Geist and Scherrer (2011) and Painter-Wakefield and Parr (2012) studied the feature selection in offline RL using ℓ_1 regularization. Finite sample guarantees for offline RL were established by Ghavamzadeh et al. (2011); Geist et al. (2012) and Hao et al. (2021a).

Online SMDP reduces to contextual linear bandits with sparse parameters when the episode length $H = 1$. Abbasi-Yadkori et al. (2012) proposed an algorithm that achieves an $\tilde{O}(\sqrt{s_\star dN})$ regret bound and matches a lower bound established in Lattimore and Szepesvári (2020). Hao et al. (2020) and Jang et al. (2022) proposed an algorithm with a regret upper bound that does not have $\sqrt{d}$ and depends only on the minimum eigenvalue of the Gram matrix of contexts. (Oh et al., 2021) and (Kim and Paik, 2019; Bastani and Bayati, 2020) use results in high dimensional statistics (Bühlmann and Van De Geer, 2011; van de Geer and Bühlmann, 2009) to establish regret bounds that depend on the restrictive minimum eigenvalue and the compatibility condition, respectively. Even though the fact that (restrictive) minimum eigenvalue is the critical parameter determining the upper and lower bound of regret for linear bandits is known, extending these results to the SMDP is nontrivial and remains open.

Applying the DR method (Bang and Robins, 2005; Fleiss et al., 2013) to bandit literature was introduced by Kim and Paik (2019) and Dimakopoulou et al.

(2019). A line of linear bandit literature applied the DR method to design an algorithm with improved regret bound Kim et al. (2023b,c) and near-optimal regret bound Kim et al. (2021, 2023d) However, all preceding works are limited to (sparse) linear bandits, and extending these results to online sparse linear RL is nontrivial.

3 SPARSE LINEAR MARKOV DECISION PROCESS

In this section, we present the problem formulation of the SMDP and a lower bound that depends only on the restrictive minimum eigenvalue for the regret.

3.1 Problem Formulation

Let $\mathrm{MDP}(\mathcal{X}, \mathcal{A}, H, \mathbb{P}, r)$ denote an episodic homogeneous MDP where $\mathcal{X}$ and $\mathcal{A}$ are the sets of possible states and actions, $H \in \mathbb{N}$ is the length of each episode, $\mathbb{P}$ is the state transition probability measure, and r is the reward function. We allow the cardinality of the state space $\mathcal{X}$ to be infinite but require $\mathcal{A}$ to have finite cardinality $|\mathcal{A}|$. For $x \in \mathcal{X}$ and $a \in \mathcal{A}$, the probability measure $\mathbb{P}(\cdot|x, a)$ denotes over the state in the next time if action a is taken in state x, and $r : \mathcal{X} \times \mathcal{A} \to [0, 1]$ is the deterministic reward function. For simplicity of exposition, we assume that the reward function is known; all our results hold when the reward is unknown.

An agent interacts with this episodic MDP as follows. At the beginning of each episode, an initial state x_1 is sampled from the unknown distribution $\mathbb{P}_0$. Then, in each period $h \in [H]$, the agent observes the state $x_h \in \mathcal{X}$, picks an action $a_h \in \mathcal{A}$, and receives a reward $r(x_h, a_h)$. The MDP evolves into a new state x_{h+1} drawn from the probability measure $\mathbb{P}(\cdot|x, a)$. The episode terminates after H interactions, i.e. when x_{H+1} is observed. Note that the agent does take an action at x_{H+1} and hence receives no reward.

We focus on the sparse linear Markov decision process (SMDP) defined as follows.

Definition 3.1 (Sparse linear MDP, Hao et al. (2021b))**.** *The $\mathrm{MDP}(\mathcal{X}, \mathcal{A}, H, \mathbb{P}, r)$ is $s_\star$-sparse to a (known) feature map $\phi : \mathcal{X} \times \mathcal{A} \to [-1, 1]^d$ if there exists an (unknown) function $\psi = (\psi_1(x), \dots, \psi_d(x)) : \mathcal{X} \to \mathbb{R}^d$ and an (unknown) set $\mathcal{I}_\star \subseteq [d]$ with $|\mathcal{I}_\star| := s_\star \ll d$ such that $\psi_i(x) = 0$ for all $x \in \mathcal{X}$ and $i \notin \mathcal{I}_\star$, and*

$$\begin{aligned}\mathbb{P}\left(X_{h+1} = x | X_h = x', a_h = a'\right) &= \phi(x', a')^\top \psi(x), \\ &= \sum_{i \in \mathcal{I}_\star} \phi_i(x', a') \psi_i(x)\end{aligned}$$

for all $h \in [H]$ and $(x', a') \in \mathcal{X} \times \mathcal{A}$. We denote a sparse MDP by $\mathrm{SMDP}(\mathcal{X}, \mathcal{A}, H, \phi, \psi, r)$.

A policy $\pi := (\pi_1, \dots, \pi_H)$ where $\pi_h : \mathcal{X} \to \Delta_{\mathcal{A}}$, $h \in [H]$, is a function from the state $\mathcal{X}$ to the set $\Delta_{\mathcal{A}}$ of probability distributions over $\mathcal{A}$. Let

$$V_h^\pi(x) := \mathbb{E}^\pi \left[\sum_{h'=h}^{H} r\left(x_{h'}, a_{h'}\right) \middle| x_h = x \right], \quad \forall x \in \mathcal{X},$$

denote the expected reward of policy π over periods $h, \dots, H$ when the state in period $h \in [H]$ is x. For $(x, a) \in \mathcal{X} \times \mathcal{A}$, define the Q-value function

$$Q_h^\pi(x, a) := r(x, a) + \mathbb{E}_{x' \sim \mathbb{P}(\cdot | x_h = x, a_h = a)} \left[V_{h+1}^\pi(x') \right],$$

which is the expected value of cumulative rewards over $[h, H]$ when the agent takes action $a \in \mathcal{A}$ in period h, and follows policy π thereafter. Since $|\mathcal{A}|$ and H are both finite, there always exists an optimal policy $\pi^\star$ that achieves the optimal value $V_h^\star(x) = \sup_\pi V_h^\pi(x)$ for all $x \in \mathcal{X}$ and $h \in [H]$ (see e.g. Puterman (2014)). Let $\widehat{A}$ denote an algorithm that takes as input $(\mathcal{X}, \mathcal{A}, H, \phi, r)$ (ψ is *not* known to $\widehat{A}$) and a sequence of episodes and computes a sequence of policies $\widehat{\pi}^{(1)}, \dots, \widehat{\pi}^{(N)}$. The total regret $\mathrm{R}(N, \widehat{A})$ of $\widehat{A}$ over N episodes

$$\mathrm{R}(N, \widehat{A}) := \sum_{n=1}^{N} \left[V_1^\star(x_1^{(n)}) - V_1^{\widehat{\pi}^{(n)}}(x_1^{(n)}) \right],$$

where $V_1^\star(x_1^{(n)}) - V_1^{\widehat{\pi}^{(n)}}(x_1^{(n)})$ denotes the regret over episode $n \in [N]$.

For $f : \mathcal{X} \to \mathbb{R}$, let $[\mathbb{P}f](x, a) := \mathbb{E}_{x' \sim \mathbb{P}(\cdot|x,a)} f(x')$. Then the Q-value function $Q_h^\pi(x, a)$ and the value function $V_h^\pi(x)$ of the policy π is given by the Bellman equation: For all $(x, a) \in \mathcal{X} \times \mathcal{A}$,

$$\begin{aligned} Q_h^\pi(x, a) &= r(x, a) + [\mathbb{P} V_{h+1}^\pi](x, a), \\ V_h^\pi(x) &= \mathbb{E}_{a \sim \pi_h(x)} \left[Q_h^\pi(x, a) \right], \quad V_{H+1}^\pi(x) = 0. \end{aligned} \tag{1}$$

The optimal Q-value function $Q_h^\star(x, a)$ and the optimal value $V_h^\star(x)$ is given by the Bellman equations:

$$\begin{aligned} Q_h^\star(x, a) &= r(x, a) + \left[\mathbb{P} V_{h+1}^\star \right](x, a), \\ V_h^\star(x) &= \max_{a \in \mathcal{A}} Q_h^\star(x, a), \quad V_{H+1}^\star(x) = 0. \end{aligned} \tag{2}$$

The Bellman equation (2) implies that the optimal policy is the greedy policy with respect to the optimal Q-value function $\{Q_h^\star\}_{h \in [H]}$. Thus, to identify the optimal policy $\pi^\star$, it suffices to estimate the optimal Q-value functions.

We will extensively use a result that, for SMDPs, the Q-value function is linear in the feature map ϕ.

Proposition 3.1 (Sparse linearity of the expected value function)**.** *For an SMDP$(\mathcal{X}, \mathcal{A}, H, \phi, \psi, r)$ and for any policy π, there exists a set of vectors $\{w_h^\pi \in \mathbb{R}^d : h \in [H]\}$ such that*

$$[\mathbb{P}V_h^\pi](x,a) = \phi(x,a)^\top w_h^\pi.$$

for all $(x,a) \in \mathcal{X} \times \mathcal{A}$, and the i-th entry of $w_h^\pi = 0$ for all $h \in [H]$ and $i \notin \mathcal{I}_\star$.

3.2 A Regret Lower Bound

Proposition 3.1 implies that the learning task for SMDPs reduces to estimating weights $\{w_h^\pi\}_{h\in[H]}$. The following restricted minimum eigenvalue is critical in this estimation task.

Definition 3.2 (Restrictive minimum eigenvalue)**.** *The restricted minimum eigenvalue (RME) for a positive semi-definite matrix $M \in \mathbb{R}^{d\times d}$ as a function of a sparsity parameter $s \in [d]$ is defined as follows:*

$$\sigma_{\min}(M,s) := \min_{\mathcal{I}\subset[d],|\mathcal{I}|\le s} \left\{ \frac{\beta^\top M \beta}{\|\beta_{\mathcal{I}}\|_2^2} : \|\beta_{\mathcal{I}^c}\|_1 \le 3\|\beta_{\mathcal{I}}\|_1 \right\}.$$

The restricted minimum eigenvalue (RME) of the matrix M is the minimum value of the quadratic form $\frac{\beta^\top M \beta}{\beta_{\mathcal{I}}^\top \beta_{\mathcal{I}}}$, when β is restricted to the set $\{\beta \in \mathbb{R}^d : \|\beta_{\mathcal{I}^c}\|_1 \le 3\|\beta_{\mathcal{I}}\|_1, |\mathcal{I}| = s\}$. The RME controls the rate of convergence in sparse linear reinforcement learning (Hao et al., 2021a) and was introduced by Bühlmann and Van De Geer (2011) to understand the properties of the Lasso estimator. The RME is also essential for establishing finite sample bounds for the Lasso estimators (see Lemma B.4 for details), and Raskutti et al. (2011) established an RME-dependent lower bound on the ℓ_1 error for any estimator.

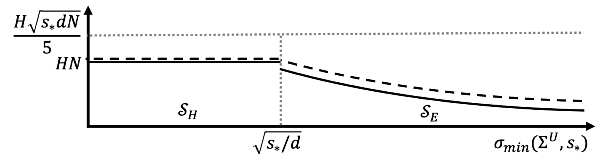
Let $\Sigma(\pi) := \mathbb{E}^\pi\big[\frac{1}{H}\sum_{h=1}^H \phi(x_h^{(n)}, a_h^{(n)})\phi(x_h^{(n)}, a_h^{(n)})^\top\big]$ denote the expected Gram matrix over the states and actions from a policy π. Let π^U denote the policy that chooses actions uniformly over the set $\mathcal{A}$ for all $h \in [H]$ and $x \in \mathcal{X}$, and let $\Sigma^U := \Sigma(\pi^U)$ denote the expected Gram matrix of the uniform policy π^U. Let $\mathcal{S}_{\mathrm{E}}$ (resp. $\mathcal{S}_{\mathrm{H}}$) denote a collection of SMDP instances of which satisfy $\sigma_{\min}(\Sigma^U, s_\star) \ge \sqrt{s_\star/d}$ (resp. $\sigma_{\min}(\Sigma^U, s_\star) < \sqrt{s_\star/d}$). We show that the lower bound on the regret of any algorithm on an SMDP instance depends on whether the instance is in $\mathcal{S}_{\mathrm{E}}$ or $\mathcal{S}_{\mathrm{H}}$.

Theorem 3.2 (SMDP Regret Lower Bound.)**.** *Suppose $N \ge s_\star \ge 5$ and $d \ge s_\star^2$. Then for any algorithm $\widehat{A}$,*

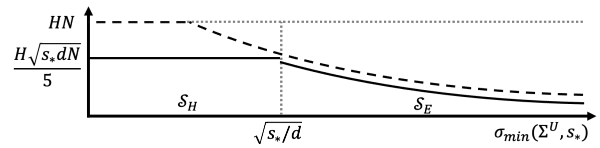
$$\sup_{SMDP\in\mathcal{S}_E} \mathbb{E}[\mathrm{R}(N,\widehat{A})] \ge \min\left\{ \frac{H s_\star \sqrt{N}}{20\sigma_{\min}(\Sigma^U, s_\star)},\ HN \right\}, \quad (3)$$

and

$$\sup_{SMDP\in\mathcal{S}_H} \mathbb{E}[\mathrm{R}(N,\widehat{A})] \ge \min\left\{ \frac{H\sqrt{s_\star d N}}{5},\ HN \right\}. \quad (4)$$



(a) Upper and lower regret bound when $ds_\star > 25N$.



(b) Upper and lower regret bound when $ds_\star \le 25N$.

Figure 1: Illustration of the regret lower bound (solid) and upper bound (dashed) proved in this paper (ignoring logarithmic terms). Our upper bound matches the lower bound for SMDPs in $\mathcal{S}_{\mathrm{E}}$ or when $ds_\star > 25N$.

The HN term in the bounds is from the rewards bounded by 1. For instances in $\mathcal{S}_{\mathrm{E}}$, i.e. when the RME $\sigma_{\min}(\Sigma^U, s_\star) \ge \sqrt{s_\star/d}$ the lower bound (3) does not depend on dimension d, which is because the RME provides enough variability on the $s_\star$ non-zero entries of the feature vectors to estimate the non-zero entries of $\{w_h^\star\}_{h\in[H]}$.

The RME $\sigma_{\min}(\Sigma^U, s_\star)$ is a measure of the variability of the $s_\star$ informative features: when RME is large (small), the reward provides a strong (weak) signal of the $s_\star$ non-zero features, and therefore, it is easier (harder) to select the $s_\star$ informative features; resulting in the faster (slower) convergence of the estimator. As the RME increases, the lower bound decreases because it is easier to identify $\mathcal{I}_\star$. However, when RME $\sigma_{\min}(\Sigma^U, s_\star) \le \sqrt{s_\star/d}$, the features do not have sufficient variability on $\mathcal{I}_\star$ and any algorithm must estimate all d entries of the weight $\{w_h^\star\}_{h\in[H]}$, resulting in d appearing in the regret bound (4).

We compare the regret upper and lower bounds (ignoring the trivial HN term and logarithmic terms) in Figure 1. The regret of our proposed algorithm is tight within a logarithmic term on $\mathcal{S}_{\mathrm{E}}$ or when $ds_\star/25 > N > \Omega(\log^7 d)$, i.e. d is large compared to N; however, there is a gap for SMDPs in $\mathcal{S}_{\mathrm{H}}$ when $d \le 25N/s_\star$.

The bound (4) generalizes $\Omega(\sqrt{s_\star dN})$ lower bound for sparse linear bandits established by Lattimore and Szepesvári (2020) to SMDPs. When $d \ge N$, Hao et al. (2020) proved $\Omega(\lambda_{\min}^{-1/3} s_\star^{1/3} N^{2/3})$ regret bound and Jang et al. (2022) improved the result to $\Omega(\lambda_{\min}^{-1/3} s_\star^{2/3} N^{2/3})$, where $\lambda_{\min}$ is the (unrestricted) minimum eigenvalue. Hao et al. (2021b) proved an $\Omega(dH)$ regret bound when $d \geq N$ for SMDPs. However, the impact of the restrictive minimum eigenvalue on the regret lower bound for general d and N has not been discovered. Theorem 3.2 establishes a lower bound for SMDPs with any d and N, and sparse linear bandits as a special case. Furthermore, we identify $\mathcal{S}_{\mathrm{E}}$ and a novel lower bound where the regret depends on the RME instead of d. When $\lambda_{\min} = \sigma_{\min}(\Sigma^U, s_\star)$, the lower bound (3) is an improvement over the best known lower bound $\Omega(\lambda_{\min}^{-1/3} s_\star^{2/3} N^{2/3})$ for sparse linear bandits. This is because $N^{2/3} \leq d^{\frac{1}{6}}\sqrt{N} \leq \lambda_{\min}^{1/3} s_\star^{1/3} \sqrt{N}$ is implied by the assumption that $d \geq N$ and $\sigma_{\min}(\Sigma^U, s_\star) \geq \sqrt{s_\star/d}$. The proof of the lower bound is in Appendix C.2.

4 PROPOSED METHOD

We propose a novel estimator and an algorithm that uses features from all actions, periods, and episodes.

4.1 Randomized Doubly Robust Q-Value Function

The Bellman equation (2) and Proposition 3.1 implies,

$$\begin{aligned} V_h^\star(x') &= \max_{a' \in \mathcal{A}} \left\{ r\left(x', a'\right) + \left[\mathbb{P} V_{h+1}^\star\right]\left(x', a'\right) \right\} \\ &= \max_{a' \in \mathcal{A}} \left\{ r\left(x', a'\right) + \phi(x', a')^\top w_{h+1}^\star \right\}. \end{aligned}$$

Thus, it follows that

$$\begin{aligned} &[\mathbb{P} V_h^\star](x, a) = \phi(x, a)^\top w_h^\star \\ &= \mathbb{E}_{x' \sim \mathbb{P}(\cdot | x, a)} \Big[\max_{a' \in \mathcal{A}} \{ r\left(x', a'\right) + \phi(x', a')^\top w_{h+1}^\star \} \Big]. \end{aligned}$$

Then,

$$\max_{a' \in \mathcal{A}} \widehat{Q}_{w_{h+1}^\star}(x', a') := \max_{a' \in \mathcal{A}} \{ r\left(x', a'\right) + \phi(x', a')^\top w_{h+1}^\star \}$$

is an unbiased estimator for $\phi(x, a)^\top w_h^\star$ for any $(x, a) \in \mathcal{X} \times \mathcal{A}$ when x' is generated from the distribution $\mathbb{P}(\cdot | x, a)$. Since $w_{h+1}^\star$ is unknown, we need an estimate $w_{h+1}^{(n)}$ for $w_{h+1}^\star$ using data from n episodes and H periods.

For period $k \in [H]$ and action $a \in \mathcal{A}$, let $X_{k+1}^{(\tau)}(a)$ denote a random sample of the state according to $\mathbb{P}(\cdot | x_k^{(\tau)}, a)$. Note that $X_{k+1}^{(\tau)}(a_k^{(\tau)}) = x_{k+1}^{(\tau)}$. Let $\Pi_{[0,H]}(x) := \min\{\max\{x, 0\}, H\}$ is a projection function onto $[0, H]$. Let

$$\widehat{Y}_{w_{h+1}^{(n)}}(x_k^{(\tau)}, a) := \Pi_{[0,H]} \Big(\max_{a' \in \mathcal{A}} \widehat{Q}_{w_{h+1}^{(n)}}(X_{k+1}^{(\tau)}(a), a') \Big) \quad (5)$$

denote an estimate for the Q-value function on $(x_k^{(\tau)}, a)$ for $a \in \mathcal{A}$. For $\tau \in [n]$ and $k \in [H]$, we only observe $x_{k+1}^{(\tau)} = X_{k+1}^{(\tau)}(a_k^{(\tau)})$, the estimate (5) is observable only when $a = a_k^{(\tau)}$. Therefore, the conventional least square value iteration for (inhomogeneous) MDP estimates $w_h^\star$ by minimizing the loss function,

$$\sum_{\tau=1}^{n} \left\{ \widehat{Y}_{w_{h+1}^{(n)}}(x_h^{(\tau)}, a_h^{(\tau)}) - w_h^\top \phi(x_h^{(\tau)}, a_h^{(\tau)}) \right\}^2. \quad (6)$$

Hao et al. (2021b) equally divide n episodes into H partitions $\{\mathcal{D}_h\}_{h \in [H]}$ and estimate $w_h^\star$ using the episodes in $\mathcal{D}_h$, i.e., by minimizing the loss function,

$$\sum_{\tau \in \mathcal{D}_h} \sum_{k=1}^{H} \left\{ \widehat{Y}_{w_{h+1}^{(n)}}(x_k^{(\tau)}, a_k^{(\tau)}) - w_h^\top \phi(x_k^{(\tau)}, a_k^{(\tau)}) \right\}^2. \quad (7)$$

The loss function (7) sums up over $k \in [H]$, enabling the estimation procedure to use a Gram matrix that sums up over all periods $k \in [H]$ in each episode. However, in order to ensure that the estimate for $w_h^{(n)}$ is independent of $w_k^{(n)}$ for $k > h$, the estimator of $w_h^{(n)}$ can only use episodes in D_h, since $w_{h+1}^{(n)}$ in (7) is estimated with $(x_k^{(\tau)}, a_k^{(\tau)})_{k \in [H], \tau \in \mathcal{D}_{h+1}}$. In order to use all n episodes in each estimation, the correlation between $w_{h+1}^{(n)}$ and $(x_k^{(\tau)}, a_k^{(\tau)})_{k \in [H], \tau \in \mathcal{D}_{h+1}}$ need to be analyzed carefully.

While the loss functions (6) and (7) used in previous work only utilize selected actions $a_k^{(\tau)}$, we consider the estimated Q-value function $\widehat{Y}_{w_{h+1}^{(n)}}(x_k^{(\tau)}, a)$ for unselected actions $a \neq a_k^{(\tau)}$ as missing data and apply the DR method to develop a novel estimator that uses *all* actions. Let $\tilde{a}_k^{(\tau)}$ denote a random variable sampled from the Uniform distribution on $\mathcal{A}$, i.e., $\mathbb{P}(\tilde{a}_k^{(\tau)} = a) = |\mathcal{A}|^{-1}$, independent of all other random variables. We define the pseudo-reward analogous to the DR method as follows:

$$\begin{aligned} \tilde{Y}_{w,k}^{(\tau)}(a) := &\frac{\mathbb{I}(\tilde{a}_k^{(\tau)} = a)}{|\mathcal{A}|^{-1}} \widehat{Y}_w(x_k^{(\tau)}, a) \\ &+ \left\{ 1 - \frac{\mathbb{I}(\tilde{a}_k^{(\tau)} = a)}{|\mathcal{A}|^{-1}} \right\} \phi(x_k^{(\tau)}, a)^\top \widehat{w}_h^{\mathrm{Im}(n)}, \end{aligned} \quad (8)$$

where the imputation estimator

$$\begin{aligned} \widehat{w}_h^{\mathrm{Im}(n)} = \arg\min_{w_h} &\Bigg\{ \lambda_{\mathrm{Im}}^{(n)} \left\| w_h \right\|_1 + \\ &\sum_{\tau=1}^{n} \sum_{k=1}^{H} \Big(\widehat{Y}_{\widehat{w}_{h+1}^{(n)}}(x_k^{(\tau)}, a_k^{(\tau)}) - w_h^\top \phi(x_k^{(\tau)}, a_k^{(\tau)}) \Big)^2 \Bigg\}. \end{aligned} \quad (9)$$

Taking expectation over $\tilde{a}_k^{(\tau)}$ on both sides of (8) gives $\mathbb{E}[\tilde{Y}_{h,k}^{(\tau)}(a)] = \widehat{Y}_h(x_k^{(\tau)}, a)$. Thus, the pseudo-reward $\tilde{Y}_{h,k}^{(\tau)}(a)$ is unbiased for all $a \in \mathcal{A}$.

Still, we observe $\widehat{Y}_h(x_k^{(\tau)}, a)$ only when $a = a_k^{(\tau)}$ and we resample $\tilde{a}_k^{(\tau)}$ until $\tilde{a}_k^{(\tau)} = a_k^{(\tau)}$. The resampling further randomizes the policy and connects it to the uniform policy. Let $\mathcal{M}_k^{(\tau)}$ denote the event of obtaining the matching $\tilde{a}_k^{(\tau)} = a_k^{(\tau)}$ with certain number of resamples. On the event $\mathcal{M}_k^{(\tau)}$, we use unbiased pseudo-rewards $\{\tilde{Y}_{h,k}^{(\tau)}(a)\}_{a\in\mathcal{A}}$, otherwise we do not use the data. Let $\widehat{w}_{H+1}^{(n)} = \mathbf{0}$ and we will construct our estimator $\widehat{w}_h^{(n)}$ recursively for $h = H, \ldots, 2$ by minimizing

$$\widehat{w}_h^{(n)} = \arg\min_{w_h} \Bigg\{ \lambda_{\text{Est}}^{(n)} \|w_h\|_1 + \sum_{\tau=1}^{n}\sum_{k=1}^{H} \mathbb{I}(\mathcal{M}_k^{(\tau)}) \sum_{a\in\mathcal{A}} \left(\tilde{Y}_{\widehat{w}_{h+1}^{(n)},k}^{(\tau)}(a) - w_h^\top \phi(x_k^{(\tau)}, a) \right)^2 \Bigg\}, \tag{10}$$

where $\lambda_{\text{Est}}^{(n)} > 0$ is another regularization parameter. Although $\widehat{w}_{h+1}^{(n)}$ is correlated with $x_k^{(\tau)}$, we develop a novel analysis technique to obtain finite sample guarantees (see Section 5.1 for details). Note that with a sufficiently large number of resamples, the event $\mathcal{M}_k^{(\tau)}$ happens with high probability. Then our estimator utilizes data that are not used by previous works in that (i) we use unbiased pseudo-rewards and feature vectors of all arms in $\mathcal{A}$ and (ii) we use all data points in $\tau \in [n]$ instead of splitting them into independent partitions as in (7). These two novel contributions enable us to design a practical and optimal algorithm for SMDP.

4.2 Proposed Algorithm

Our proposed algorithm, Randomized Doubly Robust Lasso Value Iteration (`RDRLVI`), is described in Algorithm 1. The `RDRLVI` samples $a_h^{(n)}$ as ϵ-greedy algorithm with $\epsilon = 1 - (1 - n^{-1/2})^{\frac{1}{H}}$ in order to induce exploration. Before taking the action $a_h^{(n)}$, `RDRLVI` resamples at most $M_h^{(n)} = \log(H(\tau+1)^2/\delta)/\log(1/(1 - |\mathcal{A}|^{-1}))$ times to ensure that the pseudo-action $\tilde{a}_h^{(n)} = a_h^{(n)}$. Since $\mathbb{P}(\tilde{a}_k^{(\tau)} = a_k^{(\tau)}) = |\mathcal{A}|^{-1}$, the matching event $\mathcal{M}_k(\tau)$ occurs with probability at least $1 - \delta H^{-1}(\tau+1)^{-2}$. The resampling couples the ϵ_n-greedy policy with the uniform policy. This coupling is crucial for employing the doubly robust estimator that is able to transfer information across actions to converge faster than conventional estimators (Theorem 5.2). In practice, resampling succeeds within a few trials; however, if there is no match after $M_h^{(\tau)}$ trials, the algorithm does not update the estimators.

The computational complexity of `RDRLVI` is higher than the previous algorithm because it needs to compute the imputation estimator and pseudo-rewards.

Algorithm 1 Randomized Doubly Robust Lasso Value Iteration (`RDRLVI`)

INPUT: Confidence parameter ($\delta > 0$).
Initialize $\widehat{w}_1^{(0)} = \cdots = \widehat{w}_H^{(0)} = 0$ and set $\widehat{w}_{H+1}^{(n)} = 0$.
for Episode $n = 1, \ldots, N$ **do**
  Receive the initial state $x_1^{(n)}$.
  Set $\epsilon_n = 1 - (1 - n^{-1/2})^{\frac{1}{H}}$
  Set $M_h^{(n)} = \ln(H(\tau+1)^2/\delta)/\ln(1/(1 - |\mathcal{A}|^{-1}))$
  for period $h = 1, \ldots, H$ **do**
    while ($\tilde{a}_h^{(n)} \neq a_h^{(n)}$) **and** (count $\leq M_h^{(n)}$) **do**
      Sample $\tilde{a}_h^{(n)} \sim \text{unif}(\mathcal{A})$
      Select $a_h^{(n)}$ using ϵ_n-greedy policy

$$a_h^{(n)} = \begin{cases} \underset{a\in\mathcal{A}}{\text{argmax}}\{\Pi_{[0,H]}\big(\widehat{Q}_{\widehat{w}_{h+1}^{(n-1)}}(x_h^{(n)}, a)\big) & \text{w.p. } 1 - \epsilon_n, \\ \sim \text{unif}(|\mathcal{A}| - 1) & \text{w.p. } \epsilon_n. \end{cases}$$

      count = count + 1
    end while
    Play $a_h^{(n)}$
  end for
  for period $h = H, \ldots, 1$ **do**
    Update $\widehat{w}_h^{\text{Im}(n)}$ by minimizing the loss (9).
    if $\tilde{a}_h^{(n)} \neq a_h^{(n)}$ **then**
      Set $\widehat{w}_h^{(n)} := \widehat{w}_h^{(n-1)}$
    else
      Compute pseudo-rewards $\tilde{Y}_{\widehat{w}_h^{\text{Im}(n)},k}^{(\tau)}(a)$ in (8).
      Compute $\widehat{w}_h^{(n)}$ by minimizing the loss (10).
    end if
  end for
end for

However, this additional cost is compensated by the benefits: `RDRLVI` uses all samples in estimating value function resulting in a faster convergence rate (Theorem 5.2) and a significantly superior regret bound (Theorem 5.5) – all without requiring oracle access to an exploratory policy whose expected Gram matrix has positive RME, or the knowledge of σ_U and $s_\star$. This relaxation is possible since the algorithm collects features from *all* actions in $\mathcal{A}$ to compose a Gram matrix with a larger RME than a Gram matrix generated by an exploratory policy.

5 REGRET ANALYSIS

Next, we present our novel analysis to establish an upper bound for the regret of `RDRLVI`.

5.1 Analysis for Tail Inequality

First, we bound the regret in terms of the ℓ_1 error of the estimator $\widehat{w}_h^{(n)}$.

Lemma 5.1 (Regret decomposition)**.** *Let $\widehat{A}_{\texttt{RDRLVI}}$ denote Algorithm* `RDRLVI`*, and for each $n \in [N]$, define $\widehat{w}_{H+1}^{(n)} = \mathbf{0}$ and*

$$\bar{w}_h^{(n)} := \int_{\mathcal{X}} \Pi_{[0,H]} \left(\max_{a' \in \mathcal{A}} \widehat{Q}_{\widehat{w}_{h+1}^{(n)}}(x, a') \right) \psi(x) dx, \quad (11)$$

Then, for any $N_1 \in [N]$,

$$\mathrm{R}(N, \widehat{A}_{\texttt{RDRLVI}}) \leq 2H(\sqrt{N} + N_1) + 2 \sum_{n=N_1}^{N-1} \sum_{h=2}^{H} \|\widehat{w}_h^{(n)} - \bar{w}_h^{(n)}\|_1.$$

In the bound, the first term comes from the $\epsilon_n = 1 - (1 - n^{-1/2})^{\frac{1}{H}}$-greedy policy in `RDRLVI` and the number of episodes N_1 required to obtain an effective ℓ_1- error bound of the estimator $\widehat{w}_h^{(n)}$.

The parameter $\bar{w}_h^{(n)}$ defined in (11) yields the expectation of the estimate $\hat{Q}$ of the true Q-value over the true state-transit distribution, i.e., $\phi(x_k^{(\tau)}, a_k^{(\tau)})^\top \bar{w}_h^{(n)} = \mathbb{E}[\Pi_{[0,H]}(\max_{a' \in \mathcal{A}} \widehat{Q}_{\widehat{w}_{h+1}^{(n)}}(X, a')) | x_k^{(\tau)}, a_k^{(\tau)}]$ and the estimator $\hat{w}_h^{(n)}$ is the finite sample approximate of $\bar{w}_n^{(n)}$. As $\hat{w}_h^{(n)}$ converges to $\bar{w}_h^{(n)}$ for all $h \in [H]$, the expected Q-value functions $\phi(x, a)^\top \bar{w}_h^{(n)}$ satisfy the Bellman equation for the optimal policy (2) and converges to $w_h^\star$.

Theorem 5.2 (Tail inequality for the estimator)**.** *For any given $\delta \in (0, 1)$, set $\lambda_{\mathrm{Im}}^{(n)} := 8H\sqrt{n \log \frac{2dHn^2}{\delta}}$ and $\lambda_{\mathrm{Est}}^{(n)} := 9|\mathcal{A}|H\sqrt{n \log \frac{2dHn^2}{\delta}}$. Then, there exists an absolute constant C such that for all $h \in [H] \setminus \{1\}$, and $n \geq C\sigma_U^{-4} s_\star^4 H^2 \log^5(dHn^2/\delta) \log^2(2d)$,*

$$\left\| \widehat{w}_h^{(n)} - \bar{w}_h^{(n)} \right\|_1 \leq \frac{8s_\star}{\sigma_U \sqrt{n}} \sqrt{\log \frac{dHn^2}{\delta}},$$

with probability at least $1 - 12\delta$, where $\bar{w}_h^{(n)}$ defined in (11) and $\sigma_U := \sigma_{\min}(\Sigma^U, s_\star)$.

Note that the episode length H appears only as $\sqrt{\log(H)}$ in the convergence rate of our estimator. This is a significant improvement compared to the rate $\tilde{O}(\sigma_U^{-1} s_\star (n/H)^{-\frac{1}{2}})$ of the estimator in Hao et al. (2021b) of which estimator only uses n/H episodes in each period h.

The main challenge here is to obtain a bound for the residual,

$$\eta_{\widehat{w}_{h+1}^{(n)}, k}^{(\tau)}(a) := \widehat{Y}_{\widehat{w}_{h+1}^{(k)}}(x_k^{(\tau)}, a) - \left[\mathbb{P}\Pi_{[0,H]}\left(\max_{a' \in \mathcal{A}} \widehat{Q}_{\widehat{w}_{h+1}^{(n)}}(x, a')\right)\right](x_k^{(\tau)}, a).$$

Here, $\widehat{w}_{h+1}^{(n)}$ is correlated with $(x_k^{(\tau)})_{k \in [H], \tau \in [n]}$ resulting in a bias. However, for sufficiently large n, the residual $\eta_{\widehat{w}_{h+1}^{(n)}, k}^{(\tau)}$ is close to $\eta_{w_{h+1}^\star, k}^{(\tau)}$ and the worst-case bias can be bounded. For $\rho > 0$, define $\mathcal{W}_{h+1}(\rho) := \{w \in \mathbb{R}^d : \|w - w_{h+1}^\star\|_1 \leq \rho\}$ and let $\phi_k^{(\tau)} := \phi(x_k^{(\tau)}, \tilde{a}_k^{(\tau)})$ and $\eta_{\widehat{w}_{h+1}^{(n)}, k}^{(\tau)} := \eta_{\widehat{w}_{h+1}^{(n)}, k}^{(\tau)}(\tilde{a}_k^{(\tau)})$. We decompose the residual vector as the worst case bound on the vicinity of $w_{h+1}^\star$ and on $w_{h+1}^\star$,

$$\begin{aligned} \left\| \sum_{\tau=1}^{n} \sum_{k=1}^{H} \eta_{\widehat{w}_{h+1}^{(n)}, k}^{(\tau)} \phi_k^{(\tau)} \right\|_\infty &\leq \left\| \sum_{\tau=1}^{n} \sum_{k=1}^{H} \eta_{w_{h+1}^\star, k}^{(\tau)} \phi_k^{(\tau)} \right\|_\infty \\ &+ \sup_{w \in \mathcal{W}_{h+1}(\rho)} \left\| \sum_{\tau=1}^{n} \sum_{k=1}^{H} \left(\eta_{w,k}^{(\tau)} - \eta_{w_{h+1}^\star, k}^{(\tau)}\right) \phi_k^{(\tau)} \right\|_\infty. \end{aligned} \quad (12)$$

The following lemma bounds the worst case bound on $\mathcal{W}_{h+1}(\rho)$.

Lemma 5.3 (Worst-case bound on the sum of residuals)**.** *Suppose $n^3 \geq 16e^2$ and let $a_1^{(\tau)}, \ldots, a_H^{(\tau)}$ denote the selected actions by policy $\pi^{(\tau)}$ and $\phi_k^{(\tau)} := \phi(x_k^{(\tau)}, a_k^{(\tau)})$. Then for any policy $\pi^{(\tau)}$,*

$$\begin{aligned} &\sup_{w \in \mathcal{W}_{h+1}(\rho)} \left\| \sum_{\tau=1}^{n} \sum_{k=1}^{H} \left\{ \eta_{w,k}^{(\tau)}(a_k^{(\tau)}) - \eta_{w_{h+1}^\star, k}^{(\tau)}(a_k^{(\tau)}) \right\} \phi_k^{(\tau)} \right\|_\infty \\ &\leq \rho \sqrt{2nH \log 2d} \Big(8 + \frac{256\sqrt{3}}{3} \log^{3/2} \frac{Hdn^2}{\delta} \Big), \end{aligned}$$

with probability at least $1 - \delta/(Hn^2)$.

The proof Lemma 5.3 involves nontrivial extensions using Rademacher complexity for sequential data developed by Rakhlin et al. (2015), and details are in Appendix B.4.

Another challenge arises in bounding the first term of (12). By definition of the optimal value function, it follows that $\eta_{w_{h+1}^\star, k}^{(\tau)} = V_h^\star(X_{k+1}^{(\tau)}(\tilde{a}_k^{(\tau)})) - [\mathbb{P}V_h^\star](x_k^{(\tau)}, \tilde{a}_k^{(\tau)})$. To bound the sum of the conditional variance of $V_h^\star$, previous inequalities (e.g., Lemma C.5 in Jin et al. (2018)) are not applicable because the actions are not from the optimal policy, and the summation is over the state $k \in [H]$ not the index of the value function $h \in [H]$. Hence, we develop the novel inequality in the following lemma.

Lemma 5.4 (Bound on sum of variance of the optimal value functions)**.** *Let $a_1^{(\tau)}, \ldots, a_H^{(\tau)}$ denote a sequence of actions selected by a policy $\pi^{(\tau)}$ and $\mathcal{H}^{(\tau)}$ denote the sigma algebra generated by $\{x_{h'}^{(u)}, a_{h'}^{(u)}\}_{u \in [\tau], h' \in [H]}$ Then, for any policy $\pi^{(\tau)}$ and $h \in [H]$, the sum of the variance of the optimal value function is bounded by*

$$\mathbb{E}\left[\sum_{k=1}^{H} \left\{ V_h^\star(x_{k+1}^{(\tau)}) - [\mathbb{P}V_h^\star](x_k^{(\tau)}, a_k^{(\tau)}) \right\}^2 \middle| \mathcal{H}^{(\tau-1)} \right] \leq 10H^2.$$

With two novel lemmas, we bound the sum of residuals of Q-value function in (12) by $\tilde{O}(H\sqrt{n})$. Lemma 5.3 and Lemma 5.4 can applied to handle correlation for a more general class of estimators. We defer the detailed derivation in Appendix C.3.

5.2 A Regret Bound of RDRLVI

Theorem 5.5 (A regret bound of RDRLVI). *Fix $\delta \in (0,1)$. Then, with probability at least $1-12\delta$,*

$$\mathrm{R}(N, \widehat{A}_{RDRLVI}) \leq \min\left\{ HN,\ \frac{16 s_\star H}{\sigma_U}\sqrt{N \log dHN^2} + 2H\left(\sqrt{N} + \frac{CH^2 s_\star^4}{\sigma_U^4}\log^5 \frac{dHN^2}{\delta}\log^2 2d\right)\right\}, \tag{13}$$

for all $N \geq \frac{C s_\star^4 H^2 \log^2(2d)}{\sigma_U^4}\log^5 \frac{2d s_\star^{4/5} H^{2/5}}{e\sigma^{4/5}\sqrt{\delta}}$, where $C>0$ is an absolute constant.

The first HN term represents the trivial bound resulting from the rewards bounded by 1. Thus, the leading order term is $\tilde{O}(\sigma_U^{-1} s_\star H\sqrt{N})$. For the SMDP such that $\sigma_U^2 \geq s_\star/d$, the upper regret bound matches the lower bound (3) up to logarithmic factors. As long as σ_U does not change, our regret bound increases in the logarithmic of the ambient dimension d.

In Section 6.1, we discuss how σ_U and d affect the regret RDRLVI in our numerical experiments.

With oracle access to an exploratory policy π^E such that $\sigma_E := \sigma_{\min}(\Sigma^{\pi^E}, s_\star)$ is a positive constant independent of d and N, Hao et al. (2021b) established an $\tilde{O}(\sigma_E^{-\frac{2}{3}} H^{\frac{4}{3}} s_\star^{\frac{2}{3}} N^{\frac{2}{3}})$ regret bound for SMDP. If $s_\star$ and $\sigma_{\min}$ are unknown, the regret bound increases to $\tilde{O}(\sigma_{\min}^{-1} H^{7/3} s_\star^{5/3} N^{2/3})$. Lemma B.1) establishes that the uniform policy π^U is also exploratory whenever the SMDP admits an exploratory policy. Therefore, one can design an algorithm that uses π^U as default choice for an exploratory policy; however, simply using π^U for pure exploration results in high regret. We employ π^U to introduce the random pseudo-actions $\tilde{a}_k^{(\tau)}$ for the DR method and use features from all actions. This approach yields $\tilde{O}(\sigma_U^{-1} H s_\star N^{\frac{1}{2}})$ regret bound, without the oracle access to π^E, $s_\star$ and σ_E.

SMDPs are a special case of low-rank MDPs with a function class Φ with cardinality $|\Phi| = O(s_\star^d)$; however, the low-rank MDP results imply a loose bound for the SMDP. Specifically, Lemma 9 in Uehara et al. (2021) implies $\tilde{O}(Hs_\star^2|\mathcal{A}|^2\sqrt{N\log|\Phi|}) = \tilde{O}(Hs_\star^2|\mathcal{A}|^2\sqrt{Nd})^*$ regret bound for the SMDP. By leveraging the linear structure, we establish a $\tilde{O}(\sigma_U^{-1} s_\star H\sqrt{N\log(dHN^2)})$ regret bound, independent of $|\mathcal{A}|$ and logarithmic in d. Thus, we are able to accommodate exponentially large action spaces. Note that our bound is tighter when $\sigma_U = \tilde{\Omega}(s_\star^{-1}|\mathcal{A}|^{-2}d^{-1/2})$, and, as Hao et al. (2021a) pointed out, σ_U is a constant in many applications of interest.

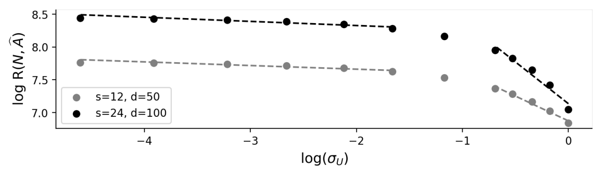


Figure 2: Logarithmic of cumulative regret of the proposed RDRLVI algorithm on RME σ_U. The dots are average regret based on ten experiments, and the slope of the right regression line is -0.76 ($s_\star = 12$) and -1.28 ($s_\star = 24$), respectively. The slopes of the flat regression lines are both -0.06. The figure supports our regret bound (13) which is proportional to σ_U^{-1} and converges to HN as σ_U decreases to 0.

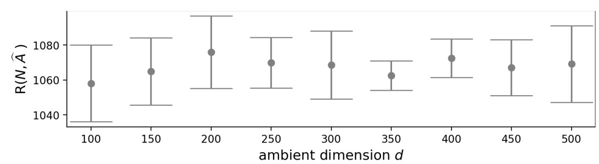


Figure 3: Cumulative regrets of the proposed RDRLVI algorithm on increasing ambient dimensions d with $\sigma_U = 1/6$. The dots and error bars represent the average and standard deviation based on ten experiments. As d increases, the regret remains flat since the algorithm selects $s_\star$ features among d features.

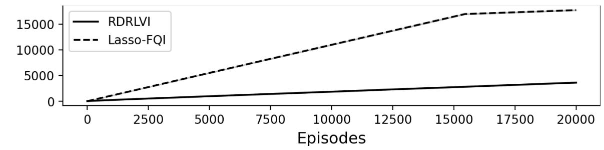
6 EXPERIMENTS

In this section, we discuss the results of numerical experiments that validate our theoretical results and superior performance of RDRLVI. We use an environment where the RME σ_U is explicitly computable. Details of the setting are in Appendix A.1.

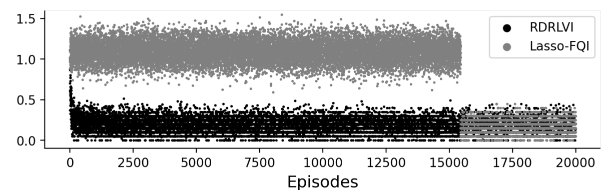
6.1 Empirical Analysis of the Regret Bound

Figure 2 shows the two-phase behavior of the cumulative regret as the RME σ_U decreases to 0. We set $H = 10$ and $N = 500$ for $s_\star = 12$ case and $N = 1000$ for $s_\star = 24$ case. For sufficiently large value of σ_U, $\log \mathrm{R}(N, \widehat{A})$ decreases linearly with $\log(\sigma_U)$ with the slope -0.76 for $s_\star = 12$ (-1.28 for $s_\star = 24$). The slope decreases for larger $s_\star$ because of the impact of the first term in (13). For sufficiently small σ_U, the regret reaches a plateau as it converges to the trivial bound HN. These results validate our regret bound (13).

In Figure 3, we plot the cumulative regret as a func-



(a) Cumulative regret comparison



(b) Average episodic regret comparison

Figure 4: Comparison of regrets of the proposed `RDRLVI` with `Lasso-FQI` (Hao et al., 2021b). The line and shade represent the average and standard deviation based on ten experiments. The figures show that `RDRLVI` finds a low-regret policy while exploiting the reward.

tion of d for a RME $\sigma_U = 1/6$, $H = 10$, $N = 500$ and $s_\star = 8$ (see Appendix A.2 for results for other values of $s_\star$). The results show that when RME is sufficiently large, `RDRLVI` quickly identifies the $s_\star$ non-zero features, and the dimension d does not impact the regret of the algorithm.

6.2 Comparison of `RDRLVI` and `Lasso-FQI`

We compare our `RDRLVI` with the Lasso fitted-Q-iteration algorithm (`Lasso-FQI`) proposed by Hao et al. (2021b). `Lasso-FQI` uses oracle access to the exploratory policy π^E, the size of the active entries $s_\star$ and RME σ_E. Hao et al. (2021b) proposed that the number of episodes N_1 for exploration bet set to $N_1 := (2048 s_\star^2 H^4 N^2 \sigma_E^{-2} \log(2dH/\delta))^{1/3}$. However, N_1 involves worst-case bounds, and the algorithm may over-explore. Hence, we reduce the number of episodes used for exploration to $N_1 := H^{4/3} N^{2/3} s_\star^{2/3} \sigma_E^{-1}$. Similarly, `RDRLVI` uses reduced the reduced value for $\lambda_{\mathrm{Im}} := H\sqrt{n \log(2dH/\delta)}$ with $\delta = 0.1$.

Figure 4 show cumulative and episodic regrets of the proposed `RDRLVI` and `Lasso-FQI` when $d = 200$, $H = 2$, $s_\star = 24$, and $\sigma_U = \sigma_E = 1$ (for results on other parameters, see Appendix A.2). Since `Lasso-FQI` chooses action according to the exploratory π_E without using the estimated value function when $n \leq N_1$, it causes high regret in most episodes. When $n \geq N_1$, `Lasso-FQI` finds the high-reward policy and takes greedy action until the end of the episodes. In contrast, the proposed `RDRLVI` finds the low-regret policy while selecting the best action at each episode. We see that `RDRLVI` balances the trade-off between exploration and exploitation by using unbiased pseudo-rewards and features of all actions and possible states.

Acknowledgements

We thank the anonymous referees for offering many helpful comments and valuable feedback. Wonyoung Kim was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. RS-2023-00240142) and Garud Iyengar was supported by NSF EFMA-2132142, ARPA-E PERFORM Program, and ONR N000142312374.

References

Yasin Abbasi-Yadkori, David Pal, and Csaba Szepesvari. Online-to-confidence-set conversions and application to sparse stochastic bandits. In *Artificial Intelligence and Statistics*, pages 1–9, 2012.

Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. Flambe: Structural complexity and representation learning of low rank mdps. *Advances in neural information processing systems*, 33:20095–20107, 2020.

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

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pages 463–474. PMLR, 2020.

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

Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

Andrew G Barto, Philip S Thomas, and Richard S Sutton. Some recent applications of reinforcement learning. In *Proceedings of the Eighteenth Yale Workshop on Adaptive and Learning Systems*, 2017.

Hamsa Bastani and Mohsen Bayati. Online decision making with high-dimensional covariates. *Operations Research*, 68(1):276–294, 2020.

Steven J Bradtke and Andrew G Barto. Linear least-squares algorithms for temporal difference learning. *Machine learning*, 22:33–57, 1996.

Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294. PMLR, 2020.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Maria Dimakopoulou, Zhengyuan Zhou, Susan Athey, and Guido Imbens. Balanced linear contextual bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3445–3453, 2019.

Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. *Statistical methods for rates and proportions*. john wiley & sons, 2013.

Matthieu Geist and Bruno Scherrer. l1-penalized projected bellman residual. In *European Workshop on Reinforcement Learning*, pages 89–101. Springer, 2011.

Matthieu Geist, Bruno Scherrer, Alessandro Lazaric, and Mohammad Ghavamzadeh. A dantzig selector approach to temporal difference learning. In *ICML-12*, pages 1399–1406. Omnipress, 2012.

Mohammad Ghavamzadeh, Alessandro Lazaric, Rémi Munos, and Matt Hoffman. Finite-sample analysis of lasso-td. In *International Conference on Machine Learning*, 2011.

Botao Hao, Tor Lattimore, and Mengdi Wang. High-dimensional sparse linear bandits. *Advances in Neural Information Processing Systems*, 33:10753–10763, 2020.

Botao Hao, Yaqi Duan, Tor Lattimore, Csaba Szepesvári, and Mengdi Wang. Sparse feature selection makes batch reinforcement learning more sample efficient. In *International Conference on Machine Learning*, pages 4063–4073. PMLR, 2021a.

Botao Hao, Tor Lattimore, Csaba Szepesvári, and Mengdi Wang. Online sparse reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 316–324. PMLR, 2021b.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Logarithmic regret for reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pages 4171–4180. PMLR, 2021a.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Uniform-pac bounds for reinforcement learning with linear function approximation. *Advances in Neural Information Processing Systems*, 34:14188–14199, 2021b.

Pihe Hu, Yu Chen, and Longbo Huang. Nearly minimax optimal reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pages 8971–9019. PMLR, 2022.

Kyoungseok Jang, Chicheng Zhang, and Kwang-Sung Jun. Popart: Efficient sparse regression and experimental design for optimal sparse linear bandits. *Advances in Neural Information Processing Systems*, 35:2102–2114, 2022.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2017.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. *Mathematics of Operations Research*, 2023.

Gisoo Kim and Myunghee Cho Paik. Doubly-robust lasso bandit. In *Advances in Neural Information Processing Systems*, pages 5869–5879, 2019.

Wonyoung Kim, Gi-Soo Kim, and Myunghee Cho Paik. Doubly robust thompson sampling with linear payoffs. In *Advances in Neural Information Processing Systems*, 2021.

Wonyoung Kim, Garud Iyengar, and Assaf Zeevi. Pareto front identification with regret minimization. *arXiv preprint arXiv:2306.00096*, 2023a.

Wonyoung Kim, Garud Iyengar, and Assaf Zeevi. Improved algorithms for multi-period multi-class packing problems with bandit feedback. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 16458–16501. PMLR, 23–29 Jul 2023b.

Wonyoung Kim, Kyungbok Lee, and Myunghee Cho Paik. Double doubly robust thompson sampling for generalized linear contextual bandits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37 (7):8300–8307, Jun. 2023c. doi: 10.1609/aaai.v37i7.26001.

Wonyoung Kim, Myunghee Cho Paik, and Min-Hwan Oh. Squeeze all: Novel estimator and self-normalized bound for linear contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3098–3124. PMLR, 2023d.

Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

J Zico Kolter and Andrew Y Ng. Regularization and feature selection in least-squares temporal difference learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 521–528, 2009.

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

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, 2016.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937. PMLR, 2016.

Min-hwan Oh, Garud Iyengar, and Assaf Zeevi. Sparsity-agnostic lasso bandit. In *International Conference on Machine Learning*, pages 8271–8280. PMLR, 2021.

Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*, pages 2377–2386. PMLR, 2016.

Christopher Painter-Wakefield and Ronald Parr. Greedy algorithms for sparse reinforcement learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 867–874, 2012.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Sequential complexities and uniform martingale laws of large numbers. *Probability theory and related fields*, 161:111–153, 2015.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over *ell_q* -balls. *IEEE transactions on information theory*, 57(10):6976–6994, 2011.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

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

Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Conference on learning theory*, pages 2898–2933. PMLR, 2019.

Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3:9–44, 1988.

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

Csaba Szepesvári. *Algorithms for reinforcement learning*. Springer Nature, 2022.

John Tsitsiklis and Benjamin Van Roy. Analysis of temporal-diffference learning with function approximation. *Advances in neural information processing systems*, 9, 1996.

Masatoshi Uehara, Xuezhou Zhang, and Wen Sun. Representation learning for online and offline rl in low-rank mdps. In *International Conference on Learning Representations*, 2021.

Sara A van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

Andrea Zanette, David Brandfonbrener, Emma Brunskill, Matteo Pirotta, and Alessandro Lazaric. Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pages 1954–1964. PMLR, 2020.

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pages 4532–4576. PMLR, 2021.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Doubly Robust Approach to Sparse Reinforcement Learning: Supplementary Materials

A SUPPLEMENTARY MATERIALS FOR EXPERIMENTS

A.1 Experiment Setting

In this section, we present the setting used in our numerical experiment. For given d, let $U_1^{(\tau)} \in [-1,1]^d$ denote a random variable whose entries are independent and have equal probability on $[-1,1]$ and we sample initial state $x_1^{(\tau)} := ((U^{(\tau)})^\top, 1)^\top$. For given $s_\star = 4, 8, 12, \ldots$, we set $\mathcal{A} := [s_\star]$. For each $a \in \mathcal{A}$, the reward is

$$r\left(x,a\right) := \mathbb{I}\left(x_{d+1}=1\right)\left(1-\frac{a-1}{s_\star}\right) + \mathbb{I}\left(x_{d+1}=-1\right)\frac{a}{2s_\star},$$

which heavily depends on the last entry of state x_{d+1}. When $x_{d+1} = -1$, the maximum reward is $1/2$ and increasing in a. In contrast, when $x_{d+1} = 1$, the maximum reward is 1 and decreasing in a. Let $\nu := a \mod (s_\star/4)$. For any $\sigma > 0$, we define feature,

$$\phi(x,a)^\top := \sigma(-x_1, \ldots, -x_{\nu-1}, x_\nu, \ldots, x_{s_\star/2}, -x_{s_\star/2+1}, \ldots, -x_{3s_\star/4-\nu+1}, x_{3s_\star/4-\nu+2} \ldots, x_{s_\star}, x_{s_\star+1}, \ldots, x_d)$$

The $\sigma > 0$ will control the (restrictive) minimum eigenvalue σ^U. To define transition distribution of states, let $x_{1:d} := (x_1, \ldots, x_d)$, and

$$\psi(x_{1:d},1)^\top := 2\sigma^{-1}s_\star^{-1}(x_1^{-1}, \ldots, x_{s_\star/2}^{-1}, 0, \ldots, 0),$$
$$\psi(x_{1:d},-1)^\top := 2\sigma^{-1}s_\star^{-1}(0, \ldots, 0, x_{s_\star/2+1}^{-1}, \ldots, x_{s_\star}^{-1}, 0, \ldots, 0).$$

Now we obtain the transition probability,

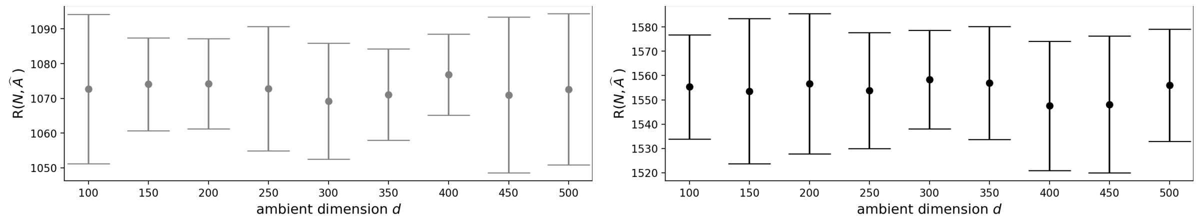
$$\mathbb{P}\left(\left(x_{1:d},1\right)\middle|\left(x_{1:d},1\right),a\right) = \phi((x_{1:d},\pm 1),a)^\top \psi(x_{1:d},1) = 1 - \frac{4(\nu-1)}{s_\star}$$
$$\mathbb{P}\left(\left(x_{1:d},-1\right)\middle|\left(x_{1:d},1\right),a\right) = \phi((x_{1:d},\pm 1),a)^\top \psi(x_{1:d},-1) = \frac{4(\nu-1)}{s_\star}.$$

Because $x_{1:d}^{(\tau)} = U_1^{(\tau)}$, we have $\sigma_U = \sigma/6$. The optimal policy is to choose $a = 1$, where the state stays $x_{d+1} = 1$ and reward $r\left(x,1\right) = 1$. Therefore, the optimal policy gains HN reward for N episodes.

A.2 Additional Experiment Results

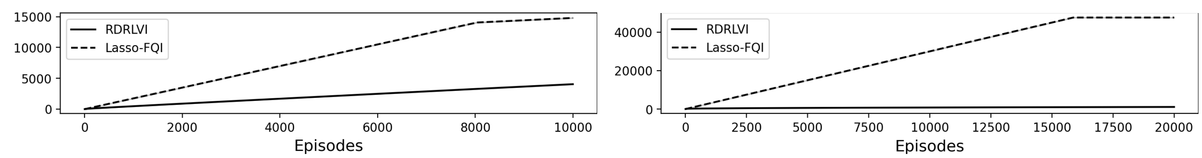
In this section, we present additional numerical results demonstrating the superior performance of `RDRLVI`. In Figure 5, we plot the cumulative regret as a function of d for a RME $\sigma_U = 1/6$, $H = 10$ and $N = 1000$ with two different values of $s_\star = 16, 24$. The results show that, for other environments than in Section 6.1, `RDRLVI`also quickly finds the $s_\star$ non-zero weights, and the dimension does not impact the regret.

As in Section 6.2, we present additional results of comparing `LASSO-FQI` (Hao et al., 2021b) with our `RDRLVI`. Figure 6 shows cumulative and episodic regrets of `LASSO-FQI` and our `RDRLVI`. The results show that, for other environments than in Section 6.2, `RDRLVI`also finds the low-regret policy while selecting the best action at each episode. We see that `RDRLVI`balances the trade-off between exploration and exploitation by using unbiased pseudo-rewards and features of all actions and possible states.

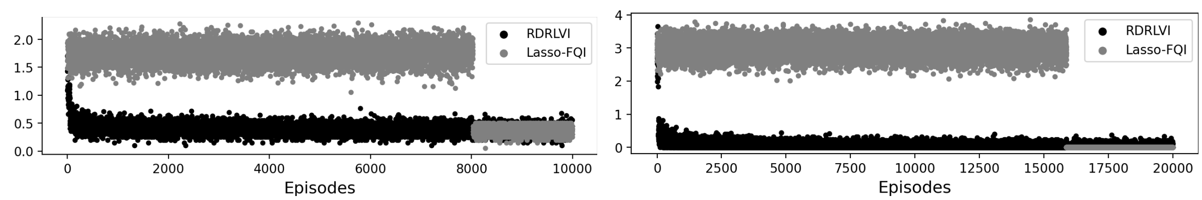


(a) Regret changes with increasing d when $s_\star = 16$. (b) Regret changes with increasing d when $s_\star = 24$.

Figure 5: Cumulative regrets of the proposed `RDRLVI` algorithm on increasing ambient dimensions d with $\sigma_U = 1/6$, $H = 10$, and $N = 1000$. The dots and error bars represent the average and standard deviation based on ten experiments. As d increases, the regret remains flat since the algorithm selects $s_\star$ features among d features.



(a) Cumulative regret comparison ($H = 3$, $s_\star = 8$). (b) Cumulative regret comparison ($H = 5$, $s_\star = 4$).



(c) Episodic regret comparison ($H = 3$, $s_\star = 8$). (d) Episodic regret comparison ($H = 5$, $s_\star = 4$)

Figure 6: Comparison of regrets of the proposed `RDRLVI` with `Lasso-FQI` (Hao et al., 2021b) when $d = 200$ and $\sigma_U = 1$. The line and shade represent the average and standard deviation based on ten experiments. The figures show that `RDRLVI` finds a low-regret policy while exploiting the reward and achieves lower regret than `Lasso-FQI`.

B TECHNICAL LEMMAS

In this section, we present technical lemmas used in our analysis. We provide proof after the novel lemmas.

B.1 Comparison of an Exploratory Policy and the Uniform Policy

Lemma B.1 (Comparison of σ_U and σ_E)**.** *Let π^E denote an exploratory policy such that $\sigma_E := \sigma_{\min}(\Sigma(\pi^E), s_\star) > 0$ and $\pi^E(x, a)$ denote the probability of selecting an action $a \in \mathcal{A}$ for the state $x \in \mathcal{X}$. Then,*

$$\sigma_U \geq \Big(\max_{(x,a)\in\mathcal{X}\times\mathcal{A}} \pi^E(x, a)\, |\mathcal{A}| \Big)^{-H} \sigma_E. \tag{14}$$

Remark B.2. The inequality (14) does not involve d and N. While the cost of the worst-case of replacing the π^E by π^U involves $|\mathcal{A}|$ and H, it does not involve d or N. We can redefine an exploratory policy π^E close to the uniform policy which gives $\max_{(x,a)\in\mathcal{X}\times\mathcal{A}} \pi^E(x, a) = O(|\mathcal{A}|^{-1})$.

Proof. Let $\Phi(x, a) := \phi(x, a)\phi(x, a)^\top$. Let $a_1, \ldots, a_H$ denote a sequence of actions selected by the policy π^E.

Note that

$$\begin{aligned}
H\Sigma^{\pi^E} :=& \mathbb{E}^{\pi^E}\left[\sum_{k=1}^{H}\Phi(x_k,a_k)\right] \\
=& \mathbb{E}^{\pi^E}\left[\sum_{k=2}^{H}\Phi(x_k,a_k)\right] + \mathbb{E}^{\pi^E}\left[\mathbb{E}\left[\Phi(x_1,a_1)\middle| x_1\right]\right] \\
=& \mathbb{E}^{\pi^E}\left[\sum_{k=2}^{H}\Phi(x_k,a_k)\right] + \mathbb{E}^{\pi^E}\left[\mathbb{E}\left[\sum_{a\in\mathcal{A}}\pi^E(x_1,a)\Phi(x_1,a)\middle| x_1\right]\right] \\
=& \mathbb{E}^{\pi^E}\left[\sum_{k=2}^{H}\Phi(x_k,a_k)\right] + \mathbb{E}^{\pi^E}\left[\sum_{a\in\mathcal{A}}\pi^E(x_1,a)\Phi(x_1,a)\right].
\end{aligned}$$

where $\pi^E(x,a)$ is the probability that the policy π^E selects an action a when the state is x. Recursively, we obtain,

$$H\Sigma^{\pi^E} = \sum_{k=1}^{H}\mathbb{E}^{\pi^E}\left[\sum_{a\in\mathcal{A}}\pi^E(x_k,a)\Phi(x_k,a)\right].$$

Because x_1 is sampled from $\mathbb{P}_0$,

$$\mathbb{E}^{\pi^E}\left[\sum_{a\in\mathcal{A}}\pi^E(x_1,a)\Phi(x_1,a)\right] = \int_{\mathcal{X}}\sum_{a\in\mathcal{A}}\pi^E(z_1,a)\Phi(z_1,a)d\mathbb{P}_0(z_1).$$

Using the SMDP setting, for each $k\geq 2$,

$$\begin{aligned}
&\mathbb{E}^{\pi^E}\left[\sum_{a\in\mathcal{A}}\pi^E(x_k,a)\Phi(x_k,a)\right] \\
&= \mathbb{E}^{\pi^E}\left[\mathbb{E}\left[\sum_{a\in\mathcal{A}}\pi^E(x_k,a)\Phi(x_k,a)\middle| x_{k-1},a_{k-1}\right]\right] \\
&= \mathbb{E}^{\pi^E}\left[\int_{\mathcal{X}}\sum_{a\in\mathcal{A}}\pi^E(z_k,a)\Phi(z_k,a)\phi(x_{k-1},a_{k-1})^\top\psi(z_k)dz_k\right] \\
&= \mathbb{E}^{\pi^E}\left[\mathbb{E}\left[\int_{\mathcal{X}}\sum_{a\in\mathcal{A}}\pi^E(z_k,a)\Phi(z_k,a)\phi(x_{k-1},a_{k-1})^\top\psi(z_k)dz_k\middle| x_{k-1}\right]\right] \\
&= \mathbb{E}^{\pi^E}\left[\int_{\mathcal{X}}\sum_{u_k\in\mathcal{A}}\sum_{u_{k-1}\in\mathcal{A}}\pi^E(z_k,u_k)\Phi(z_k,u_k)\pi^E(x_{k-1},u_{k-1})\phi(x_{k-1},u_{k-1})^\top\psi(z_k)dz_k\right].
\end{aligned}$$

Applying the equality recursively, we obtain,

$$\begin{aligned}
&\mathbb{E}^{\pi^E}\left[\sum_{a\in\mathcal{A}}\pi^E(x_k,a)\Phi(x_k,a)\right] \\
&= \int_{\mathcal{X}^k}\sum_{j=1}^{k}\sum_{u_j\in\mathcal{A}}\Phi(z_k,u_k)\left(\prod_{j=1}^{k}\pi^E(z_j,u_j)\right)\left(\prod_{j=2}^{k}\phi(z_{j-1},u_{j-1})^\top\psi(z_j)\right)dz_k\cdots dz_2 d\mathbb{P}(z_1) \\
&\preceq \left(\max_{(x,a)\in\mathcal{X}\times\mathcal{A}}\pi^E(x,a)\right)^k\int_{\mathcal{X}^k}\sum_{j=1}^{k}\sum_{u_j\in\mathcal{A}}\Phi(z_k,u_k)\left(\prod_{j=2}^{k}\phi(z_{j-1},u_{j-1})^\top\psi(z_j)\right)dz_k\cdots dz_2 d\mathbb{P}(z_1)
\end{aligned}$$

Because $\pi^U(x,a) = |\mathcal{A}|^{-1}$,

$$\begin{aligned}
&\mathbb{E}^{\pi^E}\left[\sum_{a\in\mathcal{A}} \pi^E(x_k,a)\Phi(x_k,a)\right] \\
&= \big(|\mathcal{A}| \max_{(x,a)\in\mathcal{X}\times\mathcal{A}} \pi^E(x,a)\big)^k \int_{\mathcal{X}^k} \sum_{j=1}^{k}\sum_{u_j\in\mathcal{A}} \Phi(z_k,u_k)\left(\prod_{j=1}^{k}\pi^U(z_j,u_j)\right)\left(\prod_{j=2}^{k}\phi(z_{j-1},u_{j-1})^\top\psi(z_j)\right) dz_k\cdots dz_2 d\mathbb{P}(z_1) \\
&\preceq \big(|\mathcal{A}| \max_{(x,a)\in\mathcal{X}\times\mathcal{A}} \pi^E(x,a)\big)^k \mathbb{E}^{\pi^U}\left[\sum_{a\in\mathcal{A}} \pi^U(x_k,a)\Phi(x_k,a)\right].
\end{aligned}$$

Thus,

$$\begin{aligned}
H\Sigma^{\pi^E} &\preceq \sum_{k=1}^{H}\left(|\mathcal{A}| \max_{(x,a)\in\mathcal{X}\times\mathcal{A}} \pi^E(x,a)\right)^k \mathbb{E}^{\pi^U}\left[\sum_{a\in\mathcal{A}} \pi^U(x_k,a)\Phi(x_k,a)\right] \\
&\preceq \left(|\mathcal{A}| \max_{(x,a)\in\mathcal{X}\times\mathcal{A}} \pi^E(x,a)\right)^H \sum_{k=1}^{H}\mathbb{E}^{\pi^U}\left[\sum_{a\in\mathcal{A}} \pi^U(x_k,a)\Phi(x_k,a)\right] \\
&= \left(|\mathcal{A}| \max_{(x,a)\in\mathcal{X}\times\mathcal{A}} \pi^E(x,a)\right)^H \mathbb{E}^{\pi^U}\left[\sum_{k=1}^{H} \Phi(H_k,a_k)\right].
\end{aligned}$$

This concludes the proof. □

B.2 Lower Bound for the Restrictive Minimum Eigenvalue

Lemma B.3. *(Corollary 6.8 in Bühlmann and Van De Geer (2011)) Let Σ_0 and Σ_1 be two positive semi-definite block diagonal matrices. Suppose that the restricted eigenvalue of Σ_0 satisfies $\sigma_{\min}(\Sigma_0, s) > 0$ and $\|\Sigma_1 - \Sigma_0\|_\infty \le \sigma_{\min}(\Sigma_0,s)/(32s)$. Then the restrictive eigenvalue of Σ_1 satisfies $\sigma_{\min}(\Sigma_1,s) > \sigma_{\min}(\Sigma_0,s)/2$.*

B.3 An Error bound for the Lasso Estimator

Lemma B.4. *(An ℓ_1-error bound for Lasso estimator) Let $\{x_\tau\}_{\tau\in[t]}$ denote the covariates in $[-1,1]^d$ and $y_\tau = x_\tau^\top \bar{w} + e_\tau$ for some $\bar{w}\in\mathbb{R}^d$ and $e_\tau\in\mathbb{R}$. For $\lambda > 0$, let*

$$\widehat{w}_t = \arg\min_w \sum_{\tau=1}^{t}\left(y_\tau - x_\tau^\top w\right)^2 + \lambda\|w\|_1 .$$

Let $\bar{\mathcal{S}} := \{i\in[d] : \bar{w}(i)\neq 0\}$ and $\Sigma_t := \sum_{\tau=1}^t x_\tau x_\tau^\top$. Suppose $\|\sum_{\tau=1}^t e_\tau x_\tau\| \le \frac{\lambda}{4}$, for some $\lambda > 0$ and $\|t^{-1}\Sigma_t - \bar{\Sigma}\|_\infty \le 32|\bar{\mathcal{S}}|^{-1}\sigma_{\min}(\bar{\Sigma}, |\bar{\mathcal{S}}|)$ for some $\bar{\Sigma}\in\mathbb{R}^{d\times d}$. Then the ℓ_1-error is bounded as

$$\|\widehat{w}_t - \bar{w}\|_1 \le \frac{8\lambda|\bar{\mathcal{S}}|}{t\sigma_{\min}\left(\bar{\Sigma}, |\bar{\mathcal{S}}|\right)}.$$

Proof. Let $X_t^\top := (x_1,\ldots,x_t)\in[-1,1]^{d\times t}$ and $\mathbf{e}_t^\top := (e_1,\ldots,e_t)\in\mathbb{R}^t$. We write $X_t(j)$ and $\widehat{w}_t(j)$ as the j-th column of X_t and j-th entry of $\widehat{w}_t$, respectively. By definition of $\widehat{w}_t$,

$$\|X_t(\bar{w}-\widehat{w}_t) + \mathbf{e}_t\|_2^2 + \lambda\|\widehat{w}_t\|_1 \le \left\|\mathbf{e}_t^{(j)}\right\|_2^2 + \lambda\|\bar{w}\|_1,$$

which implies

$$\begin{aligned}
\|X_t(\bar{w}-\widehat{w}_t)\|_2^2 + \lambda\|\widehat{w}_t\|_1 &\le 2\left(\widehat{w}_t - \bar{w}\right)^\top X_t^\top \mathbf{e}_t + \lambda\|\bar{w}\|_1 \\
&\le 2\|\widehat{w}_t - \bar{w}\|_1 \left\|X_t^\top \mathbf{e}_t\right\|_\infty + \lambda\|\bar{w}\|_1 \\
&\le \frac{\lambda}{2}\|\widehat{w}_t - \bar{w}\|_1 + \lambda\|\bar{w}\|_1,
\end{aligned}$$

where the last inequality uses the bound on λ. On the left hand side, by triangle inequality,

$$\begin{aligned}\|\widehat{w}_t\|_1 &= \sum_{i\in\bar{\mathcal{S}}} |\widehat{w}_t(i)| + \sum_{i\in[d]\setminus\bar{\mathcal{S}}} |\widehat{w}_t(i)| \\ &\geq \sum_{i\in\bar{\mathcal{S}}} |\widehat{w}_t(i)| - \sum_{i\in\mathcal{S}_\star} |\widehat{w}_t(i) - \bar{w}(i)| + \sum_{i\in[d]\setminus\bar{\mathcal{S}}} |\bar{w}(i)| \\ &= \|\bar{w}\|_1 - \sum_{i\in\bar{\mathcal{S}}} |\widehat{w}_t(i) - \bar{w}(i)| + \sum_{i\in[d]\setminus\bar{\mathcal{S}}} |\widehat{w}_t(i)|\end{aligned}$$

and for the right-hand side,

$$\|\widehat{w}_t - \bar{w}\|_1 = \sum_{i\in\bar{\mathcal{S}}} |\widehat{w}_t(i) - \bar{w}(i)| + \sum_{i\in[d]\setminus\bar{\mathcal{S}}} |\widehat{w}_t(i)| .$$

Plugging in both sides and rearranging the terms,

$$2\left\|X_t\left(\bar{w}-\widehat{w}_t\right)\right\|_2^2 + \lambda \sum_{i\in[d]\setminus\bar{\mathcal{S}}} |\widehat{w}_t(i)| \leq 3\lambda \sum_{i\in\bar{\mathcal{S}}} |\widehat{w}_t(i) - \bar{w}(i)| . \tag{15}$$

The inequality (15) implies $\sum_{i\in[d]\setminus\bar{\mathcal{S}}} |\widehat{w}_t(i) - \bar{w}(i)| \leq 3\sum_{i\in\bar{\mathcal{S}}} |\widehat{w}_t(i) - \bar{w}(i)|$ and

$$\begin{aligned}\left\|X_t\left(\bar{w}-\widehat{w}_t\right)\right\|_2^2 &\geq \sigma_{\min}\left(X_t^\top X_t, \left|\bar{\mathcal{S}}\right|\right) \sum_{i\in\bar{\mathcal{S}}} |\widehat{w}_t(i) - \bar{w}(i)|^2 \\ &\geq \frac{\sigma_{\min}\left(X_t^\top X_t, \left|\bar{\mathcal{S}}\right|\right)}{\left|\bar{\mathcal{S}}\right|} \left(\sum_{i\in\bar{\mathcal{S}}} |\widehat{w}_t(i) - \bar{w}(i)|\right)^2 \\ &\geq \frac{\sigma_{\min}\left(t\bar{\Sigma}, \left|\bar{\mathcal{S}}\right|\right)}{2\left|\bar{\mathcal{S}}\right|} \left(\sum_{i\in\bar{\mathcal{S}}} |\widehat{w}_t(i) - \bar{w}(i)|\right)^2 ,\end{aligned}$$

where the last inequality holds by assumption $\|t^{-1}\Sigma_t - \bar{\Sigma}\|_\infty \leq 32|\bar{\mathcal{S}}|^{-1}\sigma_{\min}(\bar{\Sigma}, \left|\bar{\mathcal{S}}\right|)$ and Lemma B.3. Plugging in (15) gives,

$$\begin{aligned}2\left\|X_t\left(\bar{w}-\widehat{w}_t\right)\right\|_2^2 + \lambda\left(\sum_{i\in[d]\setminus\bar{\mathcal{S}}} |\widehat{w}_t(i)| + \sum_{i\in\bar{\mathcal{S}}} |\widehat{w}_t(i) - \bar{w}(i)|\right) &\leq 4\lambda\sqrt{\frac{2\left|\bar{\mathcal{S}}\right|}{\sigma_{\min}\left(t\bar{\Sigma}, \left|\bar{\mathcal{S}}\right|\right)}} \left\|X_t\left(\bar{w}-\widehat{w}_t\right)\right\|_2 \\ &\leq \frac{8\lambda^2\left|\bar{\mathcal{S}}\right|}{\sigma_{\min}\left(t\bar{\Sigma}, \left|\bar{\mathcal{S}}\right|\right)} + \left\|X_t\left(\bar{w}-\widehat{w}_t\right)\right\|_2 ,\end{aligned}$$

where the last inequality uses $ab \leq a^2/4 + b^2$. Rearranging the terms,

$$\left\|X_t\left(\bar{w}-\widehat{w}_t\right)\right\|_2^2 + \lambda\left\|\widehat{w}_t - \bar{w}\right\|_1 \leq \frac{8\lambda^2\left|\bar{\mathcal{S}}\right|}{\sigma_{\min}\left(t\bar{\Sigma}, \left|\bar{\mathcal{S}}\right|\right)},$$

which proves the result. □

B.4 Sequential Rademacher Complexity for Martingales

The following lemma connects the sum of martingale differences to the sequential Rademacher complexity.

Lemma B.5. *(Lemma 4 in Rakhlin et al. (2015).) Let $Z_i \in \mathcal{Z}$ denote a stochastic process adapted to filtration $\mathcal{H}_i$ and $\mathcal{F}$ a class of functions $f : \mathcal{Z} \to [-1,1]$. Let $\mathbf{z} := (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ denote a sequence of binary trees $\mathbf{z}_i : \{\pm 1\}^{i-1} \to \mathcal{Z}$ and $\{\xi_i\}_{i\in[n]}$ denote independent Bernoulli random variables such that $\mathbb{P}(\xi_i = -1) = \mathbb{P}(\xi = 1) = 1/2$. Then for any $\alpha > 0$,*

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n f(Z_i) - \mathbb{E}\left[Z_i \middle| \mathcal{H}_{i-1}\right]\right| > \alpha\right) \leq 4\sup_{\mathbf{z}} \mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \xi_i f(\mathbf{z}_i(\xi_1,\ldots,\xi_{i-1}))\right| > \frac{\alpha}{4}\middle| \mathbf{z}\right).$$

We provide a novel lemma for a bound for sequential Rademacher complexity (Rakhlin et al., 2015). The following lemma is a generalization of Lemma 6 in Rakhlin et al. (2015).

Lemma B.6. *(Bound for sequential Rademacher complexity) Let $\{\xi_i\}_{i\in[n]}$ denote independent Bernoulli random variables such that $\mathbb{P}(\xi_i=-1)=\mathbb{P}(\xi=1)=1/2$ and $\mathbf{z}:=(\mathbf{z}_1,\ldots,\mathbf{z}_n)$ denote a sequence of binary trees $\mathbf{z}_i:\{\pm1\}^{i-1}\to\mathcal{Z}$. Let $\mathcal{F}$ denote a class of functions $f:\mathcal{Z}\to[-1,1]$. For a fixed tree $\mathbf{z}$ and $\epsilon>0$, let $N(\epsilon,\mathcal{F},\|\cdot\|_{\infty,\mathbf{z}})$ denote a covering number of $\mathcal{F}$ in the norm defined by $\|f\|_{\infty,\mathbf{z}}:=\max_{\{\xi_i\}_{i\in[n]}\in\{\pm1\}^n}|f(\mathbf{z}(\xi_1,\ldots,\xi_n)|$. Then with probability at least $1-\delta$,*

$$\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n\xi_i f(\mathbf{z}_i(\xi_{1:i-1}))\right|\le 2\inf_{\alpha>0}\left\{n\alpha+2\int_\alpha^{1/2}\sqrt{3n\log\frac{N(\epsilon,\mathcal{F},\|\cdot\|_{\infty,\mathbf{z}})}{\sqrt{\delta}}}d\epsilon\right\}.$$

Proof. For given $\epsilon>0$, define $\epsilon_j=2^{-j}$. For a fixed tree $\mathbf{z}$ of depth n, let V_j be an ϵ_j-cover with respect to ℓ_∞-norm, $\|\cdot\|_{\infty,\mathbf{z}}$. For any path $\xi:=\{\xi_i\}_{i\in[n]}\in\{\pm1\}^n$ and any $f\in\mathcal{F}$, let $\mathbf{v}^{(j)}(f,\xi)\in V_j$ denote a ϵ_j-close element of the cover in the $\|\cdot\|_{\infty,\mathbf{z}}$-norm. Now for any $f\in\mathcal{F}$ and $J\in\mathbb{N}$,

$$\begin{aligned}
&\left|\sum_{i=1}^n\xi_i f(\mathbf{z}_i(\xi_{1:i-1}))\right|\\
&\le\left|\sum_{i=1}^n\xi_i\left\{f(\mathbf{z}_i(\xi_{1:i-1}))-\mathbf{v}_i^{(J)}(f,\xi_{1:i-1})\right\}\right|+\sum_{j=1}^J\left|\sum_{i=1}^n\xi_i\left\{\mathbf{v}_i^{(j)}(f,\xi_{1:i-1})-\mathbf{v}_i^{(j-1)}(f,\xi_{1:i-1})\right\}\right|\\
&\le n\max_{i\in[n]}\left|f(\mathbf{z}_i(\xi_{1:i-1}))-\mathbf{v}_i^{(J)}(f,\xi_{1:i-1})\right|+\sum_{j=1}^J\left|\sum_{i=1}^n\xi_i\left\{\mathbf{v}_i^{(j)}(f,\xi_{1:i-1})-\mathbf{v}_i^{(j-1)}(f,\xi_{1:i-1})\right\}\right|\\
&\le n\max_{\xi'\in\{\pm1\}^n}\max_{i\in[n]}\left|f(\mathbf{z}_i(\xi'_{1:i-1}))-\mathbf{v}_i^{(J)}(f,\xi'_{1:i-1})\right|+\sum_{j=1}^J\left|\sum_{i=1}^n\xi_i\left\{\mathbf{v}_i^{(j)}(f,\xi_{1:i-1})-\mathbf{v}_i^{(j-1)}(f,\xi_{1:i-1})\right\}\right|\\
&=n\|f(\mathbf{z})-\mathbf{v}^{(J)}(f)\|_{\infty,\mathbf{z}}+\sum_{j=1}^J\left|\sum_{i=1}^n\xi_i\left\{\mathbf{v}_i^{(j)}(f,\xi_{1:i-1})-\mathbf{v}_i^{(j-1)}(f,\xi_{1:i-1})\right\}\right|
\end{aligned}$$

Because $\mathbf{v}^{(N)}(f,\xi)\in V_N$,

$$\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n\xi_i f(\mathbf{z}_i(\xi_{1:i-1}))\right|\le n\epsilon_J+\sup_{f\in\mathcal{F}}\left(\sum_{j=1}^J\left|\sum_{i=1}^n\xi_i\left\{\mathbf{v}_i^{(j)}(f,\xi_{1:i-1})-\mathbf{v}_i^{(j-1)}(f,\xi_{1:i-1})\right\}\right|\right).$$

To bound the second term, consider all possible pairs of $\mathbf{v}^r\in V_{j-1}$ and $\mathbf{v}^s\in V_j$, for $r\in[|V_{j-1}|]$ and $s\in[|V_j|]$. For each pair $(\mathbf{v}^r,\mathbf{v}^s)$, define a real-valued tree $\mathbf{w}^{(j|r,s)}$by

$$\mathbf{w}_i^{(j|r,s)}(\xi):=\begin{cases}\mathbf{v}_i^s(\xi)-\mathbf{v}_i^r(\xi) & \text{if there exists } f\in\mathcal{F} \text{ s.t. } \mathbf{v}^s=\mathbf{v}^{(j)}(f,\xi),\mathbf{v}^r=\mathbf{v}^{(j-1)}(f,\xi)\\ 0 & \text{otherwise}\end{cases},$$

for all $i\in[n]$ and $\xi\in\{\pm1\}^n$. Note that $\mathbf{w}^{(j|r,s)}$ is non-zero only on those ξ such that $\mathbf{v}^{(r)}$and $\mathbf{v}^{(s)}$ are the members of covers V_j and V_{j-1} close in the $\|\cdot\|_{\infty,\mathbf{z}}$-norm for some $f\in\mathcal{F}$. Define the set of trees W_j,

$$W_j:=\{\mathbf{w}^{(j|r,s)}:1\le r\le|V_{j-1}|,1\le s\le|V_j|\}.$$

Then we get

$$\begin{aligned}
\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n\xi_i f(\mathbf{z}_i(\xi_{1:i-1}))\right|\le& n\epsilon_J+\sup_{f\in\mathcal{F}}\left(\sum_{j=1}^J\left|\sum_{i=1}^n\xi_i\left\{\mathbf{v}_i^{(j)}(f,\xi_{1:i-1})-\mathbf{v}_i^{(j-1)}(f,\xi_{1:i-1})\right\}\right|\right).\\
\le& n\epsilon_J+\sum_{j=1}^J\sup_{\mathbf{w}^{(j)}\in W_j}\left|\sum_{i=1}^n\xi_i\mathbf{w}_i^{(j)}(\xi_{1:i-1})\right|.
\end{aligned}\tag{16}$$

Note that for $\mathbf{w}^{(j)} \in W_j$, there exists $f \in \mathcal{F}$ such that

$$
\begin{aligned}
\|\mathbf{w}^{(j)}\|_{\infty,\mathbf{z}} \leq & \sup_{\mathbf{v}^r \in V_{j-1}, \mathbf{v}^s \in V_j} \sup_{\xi \in \{\pm 1\}^n} \max_{i \in [n]} |\mathbf{v}_i^r(f, \xi_{1:i-1}) - \mathbf{v}_i^s(f, \xi_{1:i-1})| \\
= & \sup_{\mathbf{v}^r \in V_{j-1}, \mathbf{v}^s \in V_j} \sup_{\xi \in \{\pm 1\}^n} \max_{i \in [n]} |\mathbf{v}_i^r(f, \xi_{1:i-1}) - f(\mathbf{z}_i(\xi_{1:i-1})) + f(\mathbf{z}_i(\xi_{1:i-1})) - \mathbf{v}_i^s(f, \xi_{1:i-1})| \\
\leq & \epsilon_{j-1} + \epsilon_j = 3\epsilon_j.
\end{aligned}
$$

For any measurable set A and $\lambda \in \mathbb{R}$,

$$
\begin{aligned}
& \mathbb{E}\left[\sup_{\mathbf{w}^{(j)} \in W_j} \left|\sum_{i=1}^n \xi_i \mathbf{w}_i^{(j)}(\xi_{1:i-1})\right| \mathbb{I}(A)\right] \\
& = \mathbb{P}(A) \frac{\mathbb{E}\left[\sup_{\mathbf{w}^{(j)} \in W_j} \left|\sum_{i=1}^n \xi_i \mathbf{w}_i^{(j)}(\xi_{1:i-1})\right| \mathbb{I}(A)\right]}{\mathbb{P}(A)} \\
& = \mathbb{P}(A) \frac{\mathbb{E}\left[\log\left\{\exp\left(\lambda \sup_{\mathbf{w}^{(j)} \in W_j} \left|\sum_{i=1}^n \xi_i \mathbf{w}_i^{(j)}(\xi_{1:i-1})\right|\right)\right\} \mathbb{I}(A)\right]}{\lambda \mathbb{P}(A)} \\
& \leq \frac{\mathbb{P}(A)}{\lambda} \log\left(\frac{\mathbb{E}\left[\exp\left(\lambda \sup_{\mathbf{w}^{(j)} \in W_j} \left|\sum_{i=1}^n \xi_i \mathbf{w}_i^{(j)}(\xi_{1:i-1})\right|\right) \mathbb{I}(A)\right]}{\mathbb{P}(A)}\right) \\
& \leq \frac{\mathbb{P}(A)}{\lambda} \log\left(\frac{\mathbb{E}\left[\exp\left(\lambda \sup_{\mathbf{w}^{(j)} \in W_j} \left|\sum_{i=1}^n \xi_i \mathbf{w}_i^{(j)}(\xi_{1:i-1})\right|\right)\right]}{\mathbb{P}(A)}\right),
\end{aligned}
$$

where the first inequality holds by Jensen's inequality (Note that $\mathbb{E}[\cdot \mathbb{I}(A)]/\mathbb{P}(A) = \mathbb{E}[\cdot|A]$ defines a conditional distribution). Let us write the covering number $N(\epsilon_j) := |V_j|$. Because

$$
\begin{aligned}
& \mathbb{E}\left[\exp\left(\lambda \sup_{\mathbf{w}^{(j)} \in W_j} \left|\sum_{i=1}^n \xi_i \mathbf{w}_i^{(j)}(\xi_{1:i-1})\right|\right)\right] \\
& \leq \mathbb{E}\left[\sum_{\mathbf{w}^{(j)} \in W_j} \exp\left(\lambda \left|\sum_{i=1}^n \xi_i \mathbf{w}_i^{(j)}(\xi_{1:i-1})\right|\right)\right] \\
& \leq \mathbb{E}\left[\sum_{\mathbf{w}^{(j)} \in W_j} \exp\left(\lambda \sum_{i=1}^n \xi_i \mathbf{w}_i^{(j)}(\xi_{1:i-1})\right) + \exp\left(-\lambda \sum_{i=1}^n \xi_i \mathbf{w}_i^{(j)}(\xi_{1:i-1})\right)\right] \\
& \leq |W_j| \exp\left(\frac{3\lambda^2 \epsilon_j n}{2}\right) \\
& \leq N(\epsilon_j)^2 \exp\left(\frac{3\lambda^2 \epsilon_j^2 n}{2}\right),
\end{aligned}
$$

we obtain

$$
\mathbb{E}\left[\sup_{\mathbf{w}^{(j)} \in W_j} \left|\sum_{i=1}^n \xi_i \mathbf{w}_i^{(j)}(\xi_{1:i-1})\right| \mathbb{I}(A)\right] \leq \frac{\mathbb{P}(A)}{\lambda} \left(\frac{3\lambda^2 \epsilon_j^2 n}{2} + \log \frac{N(\epsilon_j)^2}{\mathbb{P}(A)}\right)
$$

Setting $\lambda = \epsilon_j^{-1} \sqrt{2 \log(N_j^2/\mathbb{P}(A))/(3n)}$,

$$
\mathbb{E}\left[\sup_{\mathbf{w}^{(j)} \in W_j} \left|\sum_{i=1}^n \xi_i \mathbf{w}_i^{(j)}(\xi_{1:i-1})\right| \mathbb{I}(A)\right] \leq 2\mathbb{P}(A) \epsilon_j \sqrt{3n \log \frac{N(\epsilon_j)}{\sqrt{\mathbb{P}(A)}}}
$$

Summing up over $j \in [J]$,

$$\begin{aligned}
\mathbb{E}\left[\sum_{j=1}^{J} \sup_{\mathbf{w}^{(j)} \in W_j} \left| \sum_{i=1}^{n} \xi_i \mathbf{w}_i^{(j)}(\xi_{1:i-1}) \right| \mathbb{I}(A)\right] \leq & 2\mathbb{P}(A) \sum_{j=1}^{J} \epsilon_j \sqrt{3n \log \frac{N(\epsilon_j)}{\sqrt{\mathbb{P}(A)}}} \\
= & 2\mathbb{P}(A) \sum_{j=1}^{J} \frac{\epsilon_j}{\epsilon_j - \epsilon_{j+1}} \int_{\epsilon_{j+1}}^{\epsilon_j} \sqrt{3n \log \frac{N(\epsilon_j)}{\sqrt{\mathbb{P}(A)}}} d\epsilon \\
= & 4\mathbb{P}(A) \sum_{j=1}^{J} \int_{\epsilon_{j+1}}^{\epsilon_j} \sqrt{3n \log \frac{N(\epsilon_j)}{\sqrt{\mathbb{P}(A)}}} d\epsilon.
\end{aligned}$$

Because $N(\epsilon)$ is nonincreasing in ϵ,

$$\begin{aligned}
\mathbb{E}\left[\sum_{j=1}^{J} \sup_{\mathbf{w}^{(j)} \in W_j} \left| \sum_{i=1}^{n} \xi_i \mathbf{w}_i^{(j)}(\xi_{1:i-1}) \right| \mathbb{I}(A)\right] \leq & 4\mathbb{P}(A) \sum_{j=1}^{J} \int_{\epsilon_{j+1}}^{\epsilon_j} \sqrt{3n \log \frac{N(\epsilon)}{\sqrt{\mathbb{P}(A)}}} d\epsilon \\
= & 4\mathbb{P}(A) \int_{\epsilon_{J+1}}^{1/2} \sqrt{3n \log \frac{N(\epsilon)}{\sqrt{\mathbb{P}(A)}}} d\epsilon.
\end{aligned}$$

From (16),

$$\begin{aligned}
\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} \xi_i f(\mathbf{z}_i(\xi_{1:i-1})) \right| \mathbb{I}(A)\right] \leq & \mathbb{P}(A) \left(n\epsilon_J + 4 \int_{\epsilon_{J+1}}^{1/2} \sqrt{3n \log \frac{N(\epsilon)}{\sqrt{\mathbb{P}(A)}}} d\epsilon \right) \\
= & 2\mathbb{P}(A) \left(n\epsilon_{J+1} + 2 \int_{\epsilon_{J+1}}^{1/2} \sqrt{3n \log \frac{N(\epsilon)}{\sqrt{\mathbb{P}(A)}}} d\epsilon \right).
\end{aligned}$$

For any $a > 0$, let $\mathcal{E}_t(a) := \{\sup_{f \in \mathcal{F}} |\sum_{i=1}^{n} \xi_i f(\mathbf{z}_i(\xi_{1:i-1}))| > a\}$. Then, by Markov inequality,

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_t(a)) \leq & \frac{1}{a} \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} \xi_i f(\mathbf{z}_i(\xi_{1:i-1})) \right| \mathbb{I}(\mathcal{E}_t(a))\right] \\
\leq & \frac{2}{a} \mathbb{P}(\mathcal{E}_t(a)) \left(n\epsilon_{J+1} + 2 \int_{\epsilon_{J+1}}^{1/2} \sqrt{3n \log \frac{N(\epsilon)}{\sqrt{\mathbb{P}(\mathcal{E}_t(a))}}} d\epsilon \right)
\end{aligned}$$

Canceling out the probability terms,

$$a \leq 2 \left(n\epsilon_{J+1} + 2 \int_{\epsilon_{J+1}}^{1/2} \sqrt{3n \log \frac{N(\epsilon)}{\sqrt{\mathbb{P}(\mathcal{E}_t(a))}}} d\epsilon \right)$$

Setting

$$a = 2 \left(n\epsilon_{J+1} + 2 \int_{\epsilon_{J+1}}^{1/2} \sqrt{3n \log \frac{N(\epsilon)}{\sqrt{\delta}}} d\epsilon \right)$$

gives

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} \xi_i f(\mathbf{z}_i(\xi_{1:i-1})) \right| > a \right) = \mathbb{P}(\mathcal{E}_t(a)) \leq \delta.$$

Setting suitable $J \in \mathbb{N}$ proves the result. $\square$

B.5 Probabilistic Inequalities

Lemma B.7. *(Exponential martingale inequality) If a martingale $(X_t; t \geq 0)$, adapted to filtration $\mathcal{F}_t$, satisfies $\mathbb{E}[\exp(\lambda X_t) | \mathcal{F}_{t-1}] \leq \exp(\lambda^2 \sigma_t^2 / 2)$ for some constant σ_t, for all t, then for any $a \geq 0$,*

$$\mathbb{P}(|X_T - X_0| \geq a) \leq 2 \exp\left(-\frac{a^2}{2 \sum_{t=1}^{T} \sigma_t^2} \right)$$

Thus, with probability at least $1-\delta$,

$$|X_T - X_0| \leq \sqrt{2\sum_{t=1}^{T}\sigma_t^2 \log\frac{2}{\delta}}.$$

Lemma B.8. *(Azuma-Bernstein inequality) Let $\{X_s\}_{s\geq 1}$ denote the martingale difference adapted to the filtration $\{\mathcal{F}_s\}_{s\geq 0}$ such that $\mathbb{E}\left[X_s|\,\mathcal{F}_{s-1}\right]=0$. Suppose that $|X_s|\leq M$ almost surely for $s\geq 1$. Then with probability at least $1-\delta$,*

$$\sum_{s=1}^{n} X_s \leq \frac{2}{3}M\log\frac{1}{\delta} + \sqrt{2\sum_{s=1}^{n}\mathbb{E}\left[X_s^2|\,\mathcal{F}_{s-1}\right]\log\frac{1}{\delta}}.$$

C MISSING PROOFS

In this section, we provide complete proofs omitted in the manuscript.

C.1 Proof of Proposition 3.1

Proof. For $h\in[H]$ and $(x,a)\in\mathcal{X}\times\mathcal{A}$,

$$\begin{aligned}[\mathbb{P}_h V_h^\pi](x,a) :=& \int_{\mathcal{X}} V_h^\pi(x')\phi(x,a)^\top\psi(x')dx' \\ =&\phi(x,a)^\top\left\{\int_{\mathcal{X}} V_h^\pi(x')\psi(x')dx'\right\}.\end{aligned}$$

Setting $w_h^\pi := \int_{\mathcal{X}} V_h^\pi(x')\psi(x')dx'$ proves the result. □

C.2 Proof of Theorem 3.2

Proof. Without loss of generality, suppose $s_\star$ and $s=s_\star/2$ are even. Let $e_i\in\mathbb{R}^{s_\star}$ denote the i-th Euclidean basis. We set the action space $\mathcal{A}:=[s_\star]^s$ and the state space $\mathcal{X}:=\{-1,1\}^d\times\{x_0,x_g,x_b\}$. Since $d>s_\star^2$, we can define $\psi(x_{1,1},\ldots,x_{|\mathcal{S}_\star|,2s},x_{|\mathcal{S}_\star|^2+1},\ldots,x_d,x_0):=\mathbf{0}$. Let us write $x_{1:d}:=(x_{1,1},\ldots,x_{|\mathcal{S}_\star|,2s},x_{|\mathcal{S}_\star|^2+1},\ldots,x_d)$. For $\sigma^2\in(0,1]$, given $(i_1,\ldots,i_s)\in[s_\star]^s$ set

$$\begin{aligned}\psi(x_{1:d},x_g|i_1,\ldots,i_s)^\top :=& \frac{1}{\sigma s}(\underbrace{x_{i_1,1}e_{i_1}^\top,\ldots,x_{i_s,s}e_{i_s}^\top}_{s\times s_\star},0,\ldots,0),\\ \psi(x_{1:d},x_b|i_1,\ldots,i_s)^\top :=& \frac{1}{\sigma s}(\underbrace{x_{j_1,1}e_{j_1}^\top,\ldots,x_{j_s,s}e_{j_s}^\top}_{s\times s_\star},0,\ldots,0),\end{aligned}$$

where $j_v\in[s_\star]\setminus\{i_v\}$ for each $v\in[s]$. Note that $\psi(x)_{\mathcal{S}_\star^c}=\mathbf{0}$ for $\mathcal{S}_\star:=\{i_1,j_1,s_\star+i_2,s_\star+j_2,\ldots,(s-1)s_\star+i_s,(s-1)s_\star+j_s\}$. Here, $x_{1:d}$ are sampled independently from d Bernouill distributions over $\{\pm 1\}$. For each $v\in[s]$, let $z_v:=\sum_{u\in[s_\star]\setminus\{i_v\}}x_{u,v}e_u$. Note that $z_v^\top e_{i_v}=0$ and $z_v^\top e_{j_v}=x_{j_v}$. Further, for each action $a=(a(1),\ldots,a(s))\in\mathcal{A}$, let $y_v(a):=\sum_{u\in[|\mathcal{S}_\star|]\setminus\{i_v,a(v)\}}x_{u,v}e_u+\mathbb{I}\left(a(v)\neq i_v\right)x_{a(v),v}e_{a(v)}$. we construct the feature vectors

$$\phi((x_{1:d},x_0),A)^\top := \begin{cases}\frac{\sigma}{2}(\underbrace{x_{i_1,1}e_{i_1}^\top+z_1^\top,\ldots,x_{i_s,s}e_{i_s}^\top+z_s^\top}_{s\times s_\star},x_{s_\star^2+1},\ldots,x_d) & a=(i_1,\ldots,i_s)\\ \sigma(\underbrace{v_1x_{i_1,1}e_{i_1}^\top+y_1(a)^\top,\ldots,v_sx_{i_s,s}e_{i_s}^\top+y_s(a)^\top}_{s\times s_\star},x_{s_\star^2+1},\ldots,x_d) & a\neq(i_1,\ldots,i_s)\end{cases}$$

$$\phi((x_{1:d},x_g),A)^\top = \sigma(\underbrace{x_{i_1,1}e_{i_1}^\top,\ldots,x_{i_s,s}e_{i_s}^\top}_{s\times s_\star},x_{s_\star^2+1},\ldots,x_d)$$

$$\phi((x_{1:d},x_b),A)^\top = \sigma\left(z_1^\top+v_1x_{i_1,1}e_{i_1}^\top,\ldots,z_s^\top+v_sx_{i_s,s}e_{i_s}^\top,x_{s_\star^2+1},\ldots,x_d\right),$$

where $v_1, \ldots, v_s \in \{\pm 1\}$ satisfies $v_1 + \cdots + v_s = 0$. The condition $\sigma^2 \leq 1$ ensures $\|\phi(x,a)\|_\infty \leq 1$. Under this construction, the transition probability is

$$
\begin{aligned}
\mathbb{P}\left((x_{1:d} x_g) \middle| (x_{1:d}, x_0), a\right) &= \phi((x_{1:d}, x_0), a)^\top \psi(x_{1:d}, x_g; i_1, \ldots, i_s) = \begin{cases} \frac{1}{2} & a = (i_1, \ldots, i_s) \\ 0 & a \neq (i_1, \ldots, i_s) \end{cases}, \\
\mathbb{P}\left((x_{1:d}, x_b) \middle| (x_{1:d}, x_0), a\right) &= \phi((x_{1:d}, x_0), a)^\top \psi(x_{1:d}, x_b; i_1, \ldots, i_s) = \begin{cases} \frac{1}{2} & a = (i_1, \ldots, i_s) \\ 1 & a \neq (i_1, \ldots, i_s) \end{cases}, \\
\mathbb{P}\left((x_{1:d}, x_b) \middle| (x_{1:d}, x_b), a\right) &= \phi((x_{1:d}, x_b), a)^\top \psi(x_{1:d}, x_b; i_1, \ldots, i_s) = 1, \\
\mathbb{P}\left((x_{1:d}, x_g) \middle| (x_{1:d}, x_g), a\right) &= \phi((x_{1:d}, x_g), a)^\top \psi(x_{1:d}, x_g; i_1, \ldots, i_s) = 1.
\end{aligned}
$$

The construction fixes $x_{1:d}$ and the good (x_g) or bad state (x_b) after the choice of the first step. To evaluate the restrictive minimum eigenvalue,

$$
\begin{aligned}
\Sigma^{\pi^U} &= \mathbb{E}^{\pi^U}\left[\frac{1}{H|\mathcal{A}|}\sum_{h=1}^{H} \phi(X_h, a_h)\phi(X_h, a_h)^\top\right] \\
&= \mathbb{E}^{\pi^U}\left[\frac{1}{H|\mathcal{A}|}\sum_{h=1}^{H}\sum_{a\in\mathcal{A}} \phi(X_h, a)\phi(X_h, a)^\top\right] \\
&\succeq \mathbb{E}^{\pi^U}\left[\frac{1}{H|\mathcal{A}|}\sum_{h=2}^{H}\sum_{a\in\mathcal{A}} \phi(X_h, a)\phi(X_h, a)^\top\right] \\
&= \mathbb{E}^{\pi^U}\left[\mathbb{E}\left[\frac{1}{H|\mathcal{A}|}\sum_{h=2}^{H}\sum_{a\in\mathcal{A}} \phi(X_h, a)\phi(X_h, a)^\top \middle| a_1 \neq (i_1, \ldots, i_s)\right] \mathbb{P}\left(a_1 \neq (i_1, \ldots, i_s)\right)\right] \\
&= \mathbb{E}^{\pi^U}\left[\mathbb{E}\left[\frac{1}{H|\mathcal{A}|}\sum_{h=2}^{H}\sum_{a\in\mathcal{A}} \phi(X_h, a)\phi(X_h, a)^\top \middle| X_2(d+1) = \cdots = X_H(d+1) = x_b\right] \frac{|\mathcal{A}|-1}{|\mathcal{A}|}\right]
\end{aligned}
$$

Because $x_{1:d}$ are independent random variables such that $\mathbb{E}[x_i x_j] = 0$ and $\mathbb{E}[x_i^2] = 1$, we obtain $\mathbb{E}\phi(X_h, a)\phi(X_h, a)^\top = \sigma I_d$ for $X_h(d+1) = x_b$. Thus,

$$
\begin{aligned}
\Sigma^{\pi^U} &\succeq \frac{|\mathcal{A}|-1}{H|\mathcal{A}|^2}\sum_{h=2}^{H} |\mathcal{A}|\,\sigma^2 I_d \\
&\succeq \frac{H-1}{H}\frac{|\mathcal{A}|-1}{|\mathcal{A}|}\sigma^2 I_d \\
&\succeq \frac{\sigma^2}{4} I_d.
\end{aligned}
$$

Thus we obtain $\sigma_{\min}(i_1, \ldots, i_s) := \sigma_{\min(i_1,\ldots,i_s)}(\Sigma^{\pi_U}, s_\star) \geq \sigma^2/4$. Define $y := \min\{5/s_\star, \sqrt{s_\star d/N}, 1/(\sigma^2\sqrt{N})\} \in [0,1]$ and set the rewards for good state $r\left((x_{1:d}, x_g), a\right) = y$ and for the bad states $r\left((x_{1:d}, x_b), a\right) = 0$ for all $a \in \mathcal{A}$. For the initial state we set $r\left((x_{1:d}, x_0), (i_1, \ldots, i_s)\right) = y/2$ and $r\left((x_{1:d}, x_0), (j_1, \ldots, j_s)\right) = y/2$. For $a \neq (i_1, \ldots, i_s)$ and $a \neq (j_1, \ldots, j_s)$ we set, $r\left((x_{1:d}, x_0), a\right) = 0$. Because the optimal policy gains expected reward $HyN/2$, for any $(i_1, \ldots, i_s) \in [s_\star]^s$ and any algorithm $\widehat{A}$ which generates the policy $\widehat{\pi}^{(n)}$ that selects $\widehat{a}_1^{(n)}, \ldots \widehat{a}_H^{(n)}$, the expected regret is

$$
\mathbb{E}_{(i_1,\ldots,i_s)}\left[\mathrm{R}(N, \widehat{A})\right] = \frac{Hy}{2}N - \sum_{n=1}^{N} \mathbb{E}_{(i_1,\ldots,i_s)}\left[V_1^{\widehat{\pi}^{(n)}}(X_1)\right].
$$

By construction of the SMDP,

$$
\begin{aligned}
&\mathbb{E}_{(i_1,\ldots,i_s)}\left[V_1^{\widehat{\pi}^{(n)}}(X_1)\right] \\
&= \frac{Hy}{2}\mathbb{E}_{(i_1,\ldots,i_s)}\left[\mathbb{I}\left(\widehat{a}_1^{(n)} = (i_1, \ldots, i_s)\right)\right] + \frac{y}{2}\mathbb{E}_{(i_1,\ldots,i_s)}\left[\mathbb{I}\left(\widehat{a}_1^{(n)} = (j_1, \ldots, j_s)\right)\right] \\
&\leq \frac{Hy}{2}\mathbb{E}_{(i_1,\ldots,i_s)}\left[\mathbb{I}\left(\widehat{a}_1^{(n)} = (i_1, \ldots, i_s) \cup \widehat{a}_1^{(n)} = (j_1, \ldots, j_s)\right)\right].
\end{aligned}
$$

Let $M^{(N)}_{(i_1,\dots,i_s)} = \sum_{n=1}^{N} \mathbb{I}(\widehat{a}^{(n)}_1 = (i_1,\dots,i_s) \cup \widehat{a}^{(n)}_1 = (j_1,\dots,j_s))$. Note that $0 \le M^{(N)}_{(i_1,\dots,i_s)} \le N$, almost surely. For each $v \in [s]$, by Pinsker's inequality,

$$
\begin{aligned}
&\mathbb{E}_{(i_1,\dots,i_s)}\left[M^{(N)}_{(i_1,\dots,i_s)}\right] \\
&\le \mathbb{E}_{(i_1,\dots,i_{v-1}0,i_{v+1}\dots,i_s)}\left[M^{(N)}_{(i_1,\dots,i_s)}\right] + N\sqrt{\frac{1}{2}D(\mathbb{P}_{(i_1,\dots,i_{v-1}0,i_{v+1}\dots,i_p)},\mathbb{P}_{(i_1,\dots,i_p)})},
\end{aligned} \tag{17}
$$

where the distribution of $\mathbb{E}_{(i_1,\dots,i_{v-1}0,i_{v+1}\dots,i_s)}$ is constructed by modifying

$$
\begin{aligned}
\psi(x_{1:d},x_g|i_1,\dots,i_{v-1}0,i_{v+1}\dots,i_s)^\top &:= \frac{1}{(s-1)}(e^\top_{i_1},\dots,e^\top_{i_{v-1}},0^\top,e^\top_{i_{v+1}},\dots,e^\top_{i_s},0^\top,\dots,0^\top)\cdot\frac{1}{2^d} \\
\psi(x_{1:d},x_b|i_1,\dots,i_{v-1}0,i_{v+1}\dots,i_s)^\top &:= \frac{1}{(s-1)}(e^\top_{j_1},\dots,e^\top_{j_{v-1}},0^\top,e^\top_{j_{v+1}}\dots,e^\top_{j_s},0^\top,\dots,0^\top)\cdot\frac{1}{2^d}
\end{aligned}
$$

and for $\Delta \in (0,1/4)$ to be determined later

$$
\begin{aligned}
&\phi((x_{1:d},x_0),a)^\top \\
&:= \begin{cases}
\left(\frac{1}{2}+\Delta\right)\left(e^\top_{i_1},\dots,e^\top_{i_{v-1}},0^\top,e^\top_{i_{v+1}},\dots,e^\top_{i_s},0^\top,\dots,0^\top\right) & \\
+\left(\frac{1}{2}-\Delta\right)\left(e^\top_{j_1},\dots,e^\top_{j_{v-1}},0^\top,e^\top_{j_{v+1}}\dots,x_{j_s}e^\top_{j_s},0^\top,\dots,0^\top\right) & a=(i_1,\dots,i_s) \\
\left(e^\top_{j_1},\dots,e^\top_{j_{v-1}},0^\top,e^\top_{j_{v+1}}\dots,x_{j_s}e^\top_{j_s},0^\top,\dots,0^\top\right) & a\neq(i_1,\dots,i_s)
\end{cases}.
\end{aligned}
$$

This construction modifies the distribution when $a = (i_1,\dots,i_s)$

$$
\begin{aligned}
\mathbb{P}\left(X_2(d+1)=x_g|\,x_0,a\right) &= \frac{1}{2}+\Delta, \\
\mathbb{P}\left(X_2(d+1)=x_b|\,x_0,a\right) &= \frac{1}{2}-\Delta.
\end{aligned}
$$

Other feature vectors are constructed as:

$$
\begin{aligned}
\phi((x_{1:d},x_g),a)^\top &= \left(e^\top_{i_1},\dots,e^\top_{i_{v-1}},0^\top,e^\top_{i_{v+1}},\dots,e^\top_{i_s},0^\top,\dots,0^\top\right), \\
\phi((x_{1:d},x_b),a)^\top &= \left(e^\top_{j_1},\dots,e^\top_{j_{v-1}},0^\top,e^\top_{j_{v+1}}\dots,x_{j_s}e^\top_{j_s},0^\top,\dots,0^\top\right),
\end{aligned}
$$

for all $a \in \mathcal{A}$. Then the distribution of $\mathbb{P}_{(i_1,\dots,i_s)}$ and $\mathbb{P}_{(i_1,\dots,i_{v-1}0,i_{v+1}\dots,i_s)}$ only differs when the action of the first step $\widehat{a}^{(n)}_1 = (i_1,\dots,i_s)$ (Note that this problem does not count in hard instances and its RME can be zero). Let $D(\mathbb{P}_1,\mathbb{P}_2)$ denote the relative entropy between probability measures $\mathbb{P}_1$ and $\mathbb{P}_2$ and $\mathbb{P}_{(i_1,\dots,i_s)}(a)$ denote the distribution of states when $a_1 = a$. By the divergence decomposition (Lemma 15.1 in Lattimore and Szepesvári (2020)),

$$
\begin{aligned}
&D(\mathbb{P}_{(i_1,\dots,i_{v-1}0,i_{v+1}\dots,i_s)},\mathbb{P}_{(i_1,\dots,i_s)}) \\
&= \sum_{a\in\mathcal{A}}\mathbb{E}_{(i_1,\dots,i_{v-1}0,i_{v+1}\dots,i_s)}\left[\sum_{n=1}^{N}\mathbb{I}\left(\widehat{a}^{(n)}_1=a\right)\right]D(\mathbb{P}_{(i_1,\dots,i_{v-1}0,i_{v+1}\dots,i_s)}(a),\mathbb{P}_{(i_1,\dots,i_s)}(a)) \\
&= \mathbb{E}_{(i_1,\dots,i_{v-1}0,i_{v+1}\dots,i_s)}\left[\sum_{n=1}^{N}\mathbb{I}\left(\widehat{a}^{(n)}_1=(i_1,\dots,i_s)\right)\right]D(\mathbb{P}_{(i_1,\dots,i_{v-1}0,i_{v+1}\dots,i_s)}((i_1,\dots,i_s)),\mathbb{P}_{(i_1,\dots,i_s)}((i_1,\dots,i_s))) \\
&= \mathbb{E}_{(i_1,\dots,i_{v-1}0,i_{v+1}\dots,i_s)}\left[\sum_{n=1}^{N}\mathbb{I}(\widehat{a}^{(n)}_1=(i_1,\dots,i_s))\right]\left(\left(\frac{1}{2}+\Delta\right)\log(1+2\Delta)+\left(\frac{1}{2}-\Delta\right)\log(1-2\Delta)\right) \\
&= \mathbb{E}_{(i_1,\dots,i_{v-1}0,i_{v+1}\dots,i_s)}\left[\sum_{n=1}^{N}\mathbb{I}(\widehat{a}^{(n)}_1=(i_1,\dots,i_s))\right]\left(\frac{\log(1-4\Delta^2)}{2}+\Delta\log\frac{1+2\Delta}{1-2\Delta}\right) \\
&\le \mathbb{E}_{(i_1,\dots,i_{v-1}0,i_{v+1}\dots,i_s)}\left[M^{(N)}_{(i_1,\dots,i_s)}\right]\left(\frac{\log(1-4\Delta^2)}{2}+\Delta\log\frac{1+2\Delta}{1-2\Delta}\right).
\end{aligned}
$$

Because $\Delta \leq 1/4$, we have $\left(\frac{\log\left(1-4\Delta^2\right)}{2} + \Delta \log \frac{1+2\Delta}{1-2\Delta}\right) \leq 4\Delta^2$ and

$$D(\mathbb{P}_{(i_1,\ldots,i_{v-1}0,i_{v+1}\ldots,i_s)}, \mathbb{P}_{(i_1,\ldots,i_s)}) \leq \mathbb{E}_{(i_1,\ldots,i_{v-1}0,i_{v+1}\ldots,i_s)}\left[M^{(N)}_{(i_1,\ldots,i_s)}\right]\left(4\Delta^2\right).$$

From (17)

$$\begin{aligned}
&\mathbb{E}_{(i_1,\ldots,i_s)}\left[M^{(N)}_{(i_1,\ldots,i_s)}\right] \\
&\leq \mathbb{E}_{(i_1,\ldots,i_{v-1}0,i_{v+1}\ldots,i_s)}\left[M^{(N)}_{(i_1,\ldots,i_s)}\right] + N\sqrt{\frac{1}{2}D(\mathbb{P}_{(i_1,\ldots,i_{v-1}0,i_{v+1}\ldots,i_s)}, \mathbb{P}_{(i_1,\ldots,i_s)})} \\
&\leq \mathbb{E}_{(i_1,\ldots,i_{v-1}0,i_{v+1}\ldots,i_s)}\left[M^{(N)}_{(i_1,\ldots,i_s)}\right] + N\Delta\sqrt{2\mathbb{E}_{(i_1,\ldots,i_{v-1}0,i_{v+1}\ldots,i_s)}\left[M^{(N)}_{(i_1,\ldots,i_s)}\right]}
\end{aligned}$$

From the regret, we obtain

$$\begin{aligned}
\mathbb{E}_{(i_1,\ldots,i_s)}\left[\mathrm{R}(N,\widehat{A})\right] =& \frac{Hy}{2}N - \sum_{n=1}^{N}\mathbb{E}_{(i_1,\ldots,i_s)}\left[V_1^{\widehat{\pi}^{(n)}}(X_1)\right] \\
\geq& \frac{Hy}{2}N - \frac{Hy}{2}\mathbb{E}_{(i_1,\ldots,i_s)}\left[M^{(N)}_{(i_1,\ldots,i_s)}\right] \\
\geq& \frac{Hy}{2}\left(N - \mathbb{E}_{(i_1,\ldots,i_{v-1}0,i_{v+1}\ldots,i_s)}\left[M^{(N)}_{(i_1,\ldots,i_s)}\right] - N\Delta\sqrt{2\mathbb{E}_{(i_1,\ldots,i_{v-1}0,i_{v+1}\ldots,i_s)}\left[M^{(N)}_{(i_1,\ldots,i_s)}\right]}\right).
\end{aligned}$$

Taking supremum over $(i_1,\ldots,i_s) \in [s_\star]^s$,

$$\begin{aligned}
\sup_{(i_1,\ldots,i_s)}\mathbb{E}_{(i_1,\ldots,i_s)}\left[\mathrm{R}(N,\widehat{A})\right] \geq& \frac{1}{s_\star^s}\sum_{(i_1,\ldots,i_s)}\mathbb{E}_{(i_1,\ldots,i_s)}\left[\mathrm{R}(N,\widehat{A})\right] \\
=& \frac{1}{s_\star^{s-1}}\sum_{v=1}^{s}\sum_{(i_1,\ldots,i_{v-1},i_{v+1},\ldots,i_s)}\frac{1}{s_\star}\sum_{i_v=1}^{s_\star}\mathbb{E}_{(i_1,\ldots,i_s)}\left[\mathrm{R}(N,\widehat{A})\right].
\end{aligned}$$

Taking the average over $i_v \in [s_\star]$,

$$\begin{aligned}
&\frac{1}{s_\star}\sum_{i_v=1}^{s_\star}\mathbb{E}_{(i_1,\ldots,i_s)}\left[\mathrm{R}(N,\widehat{A})\right] \\
&\geq \frac{Hy}{2}\frac{1}{s_\star}\sum_{i_v=1}^{s_\star}\left(N - \mathbb{E}_{(i_1,\ldots,i_{v-1}0,i_{v+1}\ldots,i_s)}\left[M^{(N)}_{(i_1,\ldots,i_s)}\right] - N\Delta\sqrt{2\mathbb{E}_{(i_1,\ldots,i_{v-1}0,i_{v+1}\ldots,i_s)}\left[M^{(N)}_{(i_1,\ldots,i_s)}\right]}\right) \\
&\geq \frac{Hy}{2}\left(N - \frac{1}{s_\star}\mathbb{E}_{(i_1,\ldots,i_{v-1}0,i_{v+1}\ldots,i_s)}\left[\sum_{i_v=1}^{s_\star}M^{(N)}_{(i_1,\ldots,i_s)}\right] - \frac{N\Delta}{s_\star}\sqrt{2s_\star\mathbb{E}_{(i_1,\ldots,i_{v-1}0,i_{v+1}\ldots,i_s)}\left[\sum_{i_v=1}^{s_\star}M^{(N)}_{(i_1,\ldots,i_s)}\right]}\right) \\
&\geq \frac{Hy}{2}\left(N - \frac{N}{s_\star} - N\Delta\sqrt{\frac{2N}{s_\star}}\right) \\
&\geq \frac{HyN}{2}\left(\frac{4}{5} - \Delta\sqrt{\frac{2N}{s_\star}}\right),
\end{aligned}$$

where the third inequality holds by $s_\star \geq 5$. Setting $\Delta = (1/5)\sqrt{s_\star/N} \in (0, 1/4)$ gives

$$\frac{1}{s_\star}\sum_{i_v=1}^{s_\star}\mathbb{E}_{(i_1,\ldots,i_s)}\left[\mathrm{R}(N,\widehat{A})\right] \geq \frac{HyN}{5}.$$

Thus,

$$
\begin{aligned}
\sup_{(i_1,\ldots,i_s)} \mathbb{E}_{(i_1,\ldots,i_s)}\left[\mathrm{R}(N,\widehat{A})\right] \geq & \frac{1}{s_\star^{s-1}} \sum_{v=1}^{s} \sum_{(i_1,\ldots,i_{v-1},i_{v+1},i_s)} \frac{HyN}{5} \\
= & \frac{s_\star HyN}{5} \\
= & H \min\{N, \frac{\sqrt{s_\star dN}}{5}, \frac{s_\star \sqrt{N}}{5\sigma^2}\}
\end{aligned}
$$

We conclude that there exists $(\tilde{i}_1,\ldots,\tilde{i}_s) \in [s_\star]^s$ such that

$$
\begin{aligned}
\mathbb{E}_{(\tilde{i}_1,\ldots,\tilde{i}_s)}\left[\mathrm{R}(N,\widehat{A})\right] \geq & H \min\{N, \frac{\sqrt{s_\star dN}}{5}, \frac{s_\star \sqrt{N}}{5\sigma^2}\} \\
\geq & H \min\{N, \frac{\sqrt{s_\star dN}}{5}, \frac{s_\star \sqrt{N}}{20\sigma_{\min}(\tilde{i}_1,\ldots,\tilde{i}_s)}\},
\end{aligned}
$$

where the last inequality holds by $\sigma_{\min}(\tilde{i}_1,\ldots,\tilde{i}_s) \geq \sigma^2/4$. □

C.3 Proof of Theorem 5.2

Proof. By Lemma B.4, it is sufficient to prove the following inequalities

$$
\left\| \frac{1}{nH|\mathcal{A}|} \sum_{\tau=1}^{n} \sum_{k=1}^{H} \sum_{a\in\mathcal{A}} \phi(x_k^{(\tau)},a)\phi(x_k^{(\tau)},a)^\top - \Sigma^{\pi_U} \right\|_\infty \leq \frac{\sigma_U}{32|\mathcal{S}_\star|} \tag{18}
$$

$$
\left\| \sum_{\tau=1}^{n} \sum_{k=1}^{H} \sum_{a\in\mathcal{A}} \left\{ \tilde{Y}^{(\tau)}_{\widehat{w}^{(n)}_{h+1}}(a) - \phi(x_k^{(\tau)},a)^\top \bar{w}_h^{(n)} \right\} \phi(x_k^{(\tau)},a) \right\|_\infty \leq \lambda_{\mathrm{Est}}^{(n)} \tag{19}
$$

To prove (18), let $A := (a_1^{(1)},\ldots,a_H^{(n)})$ and $\tilde{A} := (\tilde{a}_1^{(1)},\ldots,\tilde{a}_H^{(n)})$ denote a collection of actions selected by the policy of RDRLVI and pseudo-actions selected by Uniform policy π^U, respectively. Let $X := (x_1^{(1)},\ldots,x_H^{(N)})$ and $\tilde{X}_h := (\tilde{x}_1^{(1)},\ldots \tilde{x}_H^{(N)})$ denote a sample path for states under algorithm policy and the uniform policy π^U. Let $n_1 := \min\{n \in \mathbb{N} : n \geq 1024\sigma_U^{-4} s_\star^4 H^2 \log \frac{2d^2Hn^2}{\delta}\}$ and

$$
\begin{aligned}
\mathcal{B}(X,A) &:= \bigcup_{n=n_1}^{N} \left\{ \left\| \frac{1}{nH|\mathcal{A}|} \sum_{\tau=1}^{n} \sum_{k=1}^{H} \sum_{a\in\mathcal{A}} \phi(x_k^{(\tau)},a)\phi(x_k^{(\tau)},a)^\top - \Sigma^{\pi_U} \right\|_\infty > \frac{\sigma_U}{32 s_\star} \right\}, \\
B(\tilde{X},\tilde{A}) &:= \bigcup_{n=n_1}^{N} \left\{ \left\| \frac{1}{nH} \sum_{\tau=1}^{n} \sum_{k=1}^{H} \phi(\tilde{x}_k^{(\tau)},\tilde{a}_k^{(\tau)})\phi(\tilde{x}_k^{(\tau)},\tilde{a}_k^{(\tau)})^\top - \Sigma^{\pi_U} \right\|_\infty > \frac{\sigma_U}{32 s_\star} \right\}.
\end{aligned}
$$

Note the algorithm restricts the event on $A = \tilde{A}$. For $Z := (z_1^{(1)},\ldots,z_H^{(n)}) \in \mathcal{X}^{HN}$,

$$
\begin{aligned}
&\mathbb{P}\left(\mathcal{B}(X,A) \cap \left\{A = \tilde{A}\right\}\right) \\
&= \int_{\mathcal{X}^{NH}\times\mathcal{A}^{NH}\times\mathcal{A}^{NH}} \mathbb{I}\left(\mathcal{B}(X,A)\right) \mathbb{I}\left(A = \tilde{A}\right) d\mathbb{P}(X,A,\tilde{A}) \\
&= \int_{\mathcal{X}^N} \int_{\mathcal{X}^{(H-1)N}} \int_{\mathcal{A}^{NH}} \int_{\mathcal{A}^{NH}} \mathbb{I}\left(\mathcal{B}(X,A)\right) \mathbb{I}\left(A = \tilde{A}\right) \prod_{n=1}^{N} \prod_{h=2}^{H+1} \phi(z_{h-1}^{(n)}, a_{h-1}^{(n)})^\top \psi\left(x_h^{(n)}\right) d\mathbb{P}(A) d\mathbb{P}(\tilde{A}) d\mathbb{P}(Z) \\
&= \int_{\mathcal{X}^N} \int_{\mathcal{X}^{(H-1)N}} \int_{\mathcal{A}^{NH}} \int_{\mathcal{A}^{NH}} \mathbb{I}\left(\mathcal{B}(X,\tilde{A})\right) \mathbb{I}\left(A = \tilde{A}\right) \prod_{n=1}^{N} \prod_{h=2}^{H+1} \phi(z_{h-1}^{(n)}, \tilde{a}_{h-1}^{(\tau)})^\top \psi\left(x_h^{(n)}\right) d\mathbb{P}(A) d\mathbb{P}(\tilde{A}) d\mathbb{P}(Z).
\end{aligned}
$$

Because the term

$$
\prod_{n=1}^{N} \prod_{h=2}^{H+1} \phi(z_{h-1}^{(n)}, \tilde{a}_{h-1}^{(\tau)})^\top \psi\left(x_h^{(n)}\right)
$$

is the density function for $\tilde{X}$. we obtain,

$$
\begin{aligned}
&\mathbb{P}\left(\mathcal{B}(X,A)\cap\left\{A=\tilde{A}\right\}\right)\\
&=\int_{\mathcal{X}^N}\int_{\mathcal{X}^{(H-1)N}}\int_{\mathcal{A}^{NH}}\int_{\mathcal{A}^{NH}}\mathbb{I}\left(\mathcal{B}(X,\tilde{A})\right)\mathbb{I}\left(A=\tilde{A}\right)\prod_{n=1}^{N}\prod_{h=2}^{H+1}\phi(z_{h-1}^{(n)},\tilde{a}_{h-1}^{(\tau)})^\top\psi\left(x_h^{(n)}\right)d\mathbb{P}(A)d\mathbb{P}(\tilde{A})d\mathbb{P}(Z)\\
&\le\int_{\mathcal{X}^N}\int_{\mathcal{X}^{(H-1)N}}\int_{\mathcal{A}^{NH}}\int_{\mathcal{A}^{NH}}\mathbb{I}\left(\mathcal{B}(X,\tilde{A})\right)\prod_{n=1}^{N}\prod_{h=2}^{H+1}\phi(z_{h-1}^{(n)},\tilde{a}_{h-1}^{(\tau)})^\top\psi\left(x_h^{(n)}\right)d\mathbb{P}(A)d\mathbb{P}(\tilde{A})d\mathbb{P}(Z)\\
&=\mathbb{P}\left(\mathcal{B}(\tilde{X},\tilde{A})\right).
\end{aligned}
$$

For $i,j\in[d]$, let

$$
v_{ij}^{(\tau)}:=\frac{1}{H}\sum_{k=1}^{H}\phi(\tilde{x}_k^{(\tau)},\tilde{a}_k^{(\tau)})(i)\phi(\tilde{x}_k^{(\tau)},\tilde{a}_k^{(\tau)})^\top(j)-\Sigma_{ij}^{\pi_U}.
$$

Then $\mathbb{E}\left[v_{ij}^{(\tau)}\right]=0$ and $|v_{ij}^{(\tau)}|\le 1$. Applying Lemma B.7, with probability at least $1-2\delta/(dn)^2$,

$$
\left|\sum_{\tau=1}^{n}v_{ij}^{(\tau)}\right|\le\sqrt{2n\log\frac{d^2n^2}{\delta}}.
$$

Thus, with probability at least $1-2\delta/n^2$,

$$
\left\|\frac{1}{nH}\sum_{\tau=1}^{n}\sum_{k=1}^{H}\phi(\tilde{x}_k^{(\tau)},\tilde{a}_k^{(\tau)})\phi(\tilde{x}_k^{(\tau)},\tilde{a}_k^{(\tau)})^\top-\Sigma^{\pi_U}\right\|_\infty\le\sqrt{\frac{2}{n}\log\frac{d^2n^2}{\delta}}.
$$

For all $n\ge n_1$, we have $n\ge 2^{11}\sigma_U^{-2}s_\star^2\log\frac{d^2n^2}{\delta}$ and

$$
\left\|\frac{1}{nH}\sum_{\tau=1}^{n}\sum_{k=1}^{H}\phi(\tilde{x}_k^{(\tau)},\tilde{a}_k^{(\tau)})\phi(\tilde{x}_k^{(\tau)},\tilde{a}_k^{(\tau)})^\top-\Sigma^{\pi_U}\right\|_\infty\le\frac{\sigma_U}{32s_\star}.
$$

Therefore, we obtain,

$$
\mathbb{P}\left(\mathcal{B}(X,A)\cap\left\{A=\tilde{A}\right\}\right)\le\mathbb{P}\left(\mathcal{B}(\tilde{X},\tilde{A})\right)\le\sum_{n=n_1}^{N}\frac{\delta}{n^2}\le\delta,
$$

which proves the inequality (18) for all $n\ge n_1$. Similarly, we can prove

$$
\mathbb{P}\left(\bigcup_{n=n_1}^{N}\left\{\left\|\frac{1}{nH}\sum_{\tau=1}^{n}\sum_{k=1}^{H}\phi(x_k^{(\tau)},a_k^{(\tau)})\phi(x_k^{(\tau)},a_k^{(\tau)})^\top-\Sigma^{\pi_U}\right\|_\infty\ge\frac{\sigma_U}{32s_\star}\right\}\cap\left\{A=\tilde{A}\right\}\right)\le\mathbb{P}\left(\mathcal{B}(\tilde{X},\tilde{A})\right)\le\delta,
$$

and with probability at least $1-2\delta$,

$$
\left\|\frac{1}{nH}\sum_{\tau=1}^{n}\sum_{k=1}^{H}\phi(x_k^{(\tau)},a_k^{(\tau)})\phi(x_k^{(\tau)},a_k^{(\tau)})^\top-\Sigma^{\pi_U}\right\|_\infty\le\frac{\sigma_U}{32s_\star},\tag{20}
$$

for all $n\ge n_1$.

To prove the inequality (19), recall that for $h\in[H]$ and $n\in[N]$,

$$
\bar{w}_h^{(n)}:=\int_{\mathcal{X}}\Pi_{[0,H]}\left(\max_{a'\in\mathcal{A}}\widehat{Q}_{\widehat{w}_{h+1}^{(n)}}(x,a')\right)\psi(x)dx.
$$

Define

$$
\eta_{w,k}^{(\tau)}(a):=\Pi_{[0,H]}\left(\max_{a'\in\mathcal{A}}\widehat{Q}_w(X_{k+1}^{(\tau)}(a),a')\right)-\mathbb{E}_{X\sim\mathbb{P}\left(\cdot|x_k^{(\tau)},a\right)}\left[\Pi_{[0,H]}\left(\max_{a'\in\mathcal{A}}\widehat{Q}_w(X,a')\right)\right].\tag{21}
$$

Note that

$$\begin{aligned}
&\widehat{Y}^{(\tau)}_{\widehat{w}^{(n)}_{h+1},k}(x^{(\tau)}_k, a) - \phi(x^{(\tau)}_k, a)^\top \bar{w}^{(n)}_h \\
&= \Pi_{[0,H]}\left(\max_{a'\in\mathcal{A}} \widehat{Q}_{\widehat{w}^{(n)}_{h+1}}(X^{(\tau)}_{k+1}(a), a')\right) - \phi(x^{(\tau)}_k, a)^\top \bar{w}^{(n)}_h \\
&= \Pi_{[0,H]}\left(\max_{a'\in\mathcal{A}} \widehat{Q}_{\widehat{w}^{(n)}_{h+1}}(X^{(\tau)}_{k+1}(a), a')\right) - \phi(x^{(\tau)}_k, a)^\top \left\{\int_{\mathcal{X}} \Pi_{[0,H]}\left(\max_{a'\in\mathcal{A}} \widehat{Q}_{\widehat{w}^{(n)}_{h+1}}(x, a')\right)\psi(x)dx\right\} \\
&= \Pi_{[0,H]}\left(\max_{a'\in\mathcal{A}} \widehat{Q}_{\widehat{w}^{(n)}_{h+1}}(X^{(\tau)}_{k+1}(a), a')\right) - \mathbb{E}_{X\sim\mathbb{P}\left(\cdot|x^{(\tau)}_k, a\right)}\left[\Pi_{[0,H]}\left(\max_{a'\in\mathcal{A}} \widehat{Q}_{\widehat{w}^{(n)}_{h+1}}(X, a')\right)\right] \\
&:= \eta^{(\tau)}_{\widehat{w}^{(n)}_{h+1},k}(\tilde{a}^{(\tau)}_k)
\end{aligned}$$

and

$$\begin{aligned}
\tilde{\eta}^{(\tau)}_{\widehat{w}^{(n)}_{h+1},k}(a) :=& \tilde{Y}^{(\tau)}_{\widehat{w}^{(n)}_{h+1}}(a) - \phi(x^{(\tau)}_k, a)^\top \bar{w}^{(n)}_h \\
=& \frac{\mathbb{I}\left(\tilde{a}^{(\tau)}_k = a\right)}{|\mathcal{A}|^{-1}} \eta^{(\tau)}_{\widehat{w}^{(n)}_{h+1},k}(a) + \left(1 - \frac{\mathbb{I}\left(\tilde{a}^{(\tau)}_k = a\right)}{|\mathcal{A}|^{-1}}\right)\phi(x^{(\tau)}_k, a)^\top\left(\widehat{w}^{\mathrm{Im}(n)}_h - \bar{w}^{(n)}_h\right) \\
=& |\mathcal{A}|\,\eta^{(\tau)}_{\widehat{w}^{(n)}_{h+1},k}(\tilde{a}^{(\tau)}_k) + \left(1 - \frac{\mathbb{I}\left(\tilde{a}^{(\tau)}_k = a\right)}{|\mathcal{A}|^{-1}}\right)\phi(x^{(\tau)}_k, a)^\top\left(\widehat{w}^{\mathrm{Im}(n)}_h - \bar{w}^{(n)}_h\right)
\end{aligned}$$

Then the inequality (19) becomes

$$\left\|\sum_{\tau=1}^{n}\sum_{k=1}^{H}\sum_{a\in\mathcal{A}} \tilde{\eta}^{(\tau)}_{\widehat{w}^{(n)}_{h+1},k}(a)\phi(x^{(\tau)}_k, a)\right\|_\infty \le \lambda^{(n)}_{\mathrm{Est}}.$$

The left-hand side is decomposed as

$$\begin{aligned}
&\left\|\sum_{\tau=1}^{n}\sum_{k=1}^{H}\sum_{a\in\mathcal{A}} \tilde{\eta}^{(\tau)}_{\widehat{w}^{(n)}_{h+1},k}(a)\phi(x^{(\tau)}_k, a)\right\|_\infty \\
&\le |\mathcal{A}|\left\|\sum_{\tau=1}^{n}\sum_{k=1}^{H} \eta^{(\tau)}_{\widehat{w}^{(\tau)}_{h+1},k}(\tilde{a}^{(\tau)}_k)\phi(x^{(\tau)}_k, \tilde{a}^{(\tau)}_k)\right\|_\infty \\
&\quad + \left\|\sum_{\tau=1}^{n}\sum_{k=1}^{H}\sum_{a\in\mathcal{A}}\left(1 - \frac{\mathbb{I}\left(\tilde{a}^{(\tau)}_k = a\right)}{|\mathcal{A}|^{-1}}\right)\phi(x^{(\tau)}_k, a)^\top\left(\widehat{w}^{\mathrm{Im}(n)}_h - \bar{w}^{(n)}_h\right)\phi(x^{(\tau)}_k, a)\right\|_\infty \\
&\le |\mathcal{A}|\left\|\sum_{\tau=1}^{n}\sum_{k=1}^{H} \eta^{(\tau)}_{\widehat{w}^{(\tau)}_{h+1},k}(\tilde{a}^{(\tau)}_k)\phi(x^{(\tau)}_k, \tilde{a}^{(\tau)}_k)\right\|_\infty + \left\|\sum_{\tau=1}^{n}\sum_{k=1}^{H}\sum_{a\in\mathcal{A}}\left(1 - \frac{\mathbb{I}\left(\tilde{a}^{(\tau)}_k = a\right)}{|\mathcal{A}|^{-1}}\right)\phi(x^{(\tau)}_k, a)\right\|_\infty \left\|\widehat{w}^{\mathrm{Im}(n)}_h - \bar{w}^{(n)}_h\right\|_1,
\end{aligned}$$

where the last inequality involves $\|\phi(x,a)\|_\infty \le 1$. Since

$$\mathbb{E}\left[\sum_{a\in\mathcal{A}}\left(1 - \frac{\mathbb{I}\left(\tilde{a}^{(\tau)}_k = a\right)}{|\mathcal{A}|^{-1}}\right)\phi(x^{(\tau)}_k, a)\,\middle|\, x^{(\tau)}_k\right] = 0,$$

we can use Lemma B.7 to obtain with probability at least $1 - 2\delta/(Hn^2)$

$$\left\|\sum_{\tau=1}^{n}\sum_{k=1}^{H}\sum_{a\in\mathcal{A}}\left(1 - \frac{\mathbb{I}\left(\tilde{a}^{(\tau)}_k = a\right)}{|\mathcal{A}|^{-1}}\right)\phi(x^{(\tau)}_k, a)\right\|_\infty \le |\mathcal{A}|\sqrt{nH\log\frac{dHn^2}{\delta}}.$$

Thus it is sufficient to prove

$$|\mathcal{A}| \left\| \sum_{\tau=1}^{n} \sum_{k=1}^{H} \eta^{(\tau)}_{\widehat{w}^{(n)}_{h+1},k}(\tilde{a}^{(\tau)}_k)\phi(x^{(\tau)}_k, \tilde{a}^{(\tau)}_k) \right\|_\infty + |\mathcal{A}| \sqrt{nH \log \frac{dHn^2}{\delta}} \left\| \widehat{w}^{\mathrm{Im}(n)}_h - \bar{w}^{(n)}_h \right\|_1 \leq \lambda^{(n)}_{\mathrm{Est}}. \tag{22}$$

We prove (22) by inductive arguments. For step H, we have $\widehat{w}^{(n)}_{H+1} = \mathbf{0}$ and $\widehat{Q}^{\widehat{w}^{(n)}_{H+1}}_{H+1}(x,a) = 0$ for all $(x,a) \in \mathcal{X} \times \mathcal{A}$. This implies $\eta^{(\tau)}_{\widehat{w}^{(n)}_{H+1},k}(a) = 0$ for all $a \in \mathcal{A}$, and the inequality

$$\begin{aligned}
&\left\| \sum_{\tau=1}^{n} \sum_{k=1}^{H} \left\{ \widehat{Y}^{(\tau)}_{\widehat{w}^{(n)}_{H+1}}(x^{(\tau)}_k, a^{(\tau)}_k) - \phi(x^{(\tau)}_k, a^{(\tau)}_k)^\top \bar{w}^{(n)}_h \right\} \phi(x^{(\tau)}_k, a^{(\tau)}_k) \right\|_\infty \\
&= \left\| \sum_{\tau=1}^{n} \sum_{k=1}^{H} \eta_{\widehat{w}^{(n)}_{H+1}}(a^{(\tau)}_k)\phi(x^{(\tau)}_k, a^{(\tau)}_k) \right\|_\infty \leq \lambda^{(n)}_{\mathrm{Im}}
\end{aligned}$$

holds. By Lemma B.4 and (20), we obtain

$$\left\| \widehat{w}^{\mathrm{Im}(n)}_H - \bar{w}^{(n)}_H \right\|_1 \leq \frac{8\lambda^{(n)}_{\mathrm{Im}} s_\star}{Hn\sigma_U} = \frac{64 s_\star}{\sigma_U \sqrt{n}} \sqrt{\log \frac{dHn^2}{\delta}},$$

which implies

$$\begin{aligned}
&|\mathcal{A}| \left\| \sum_{\tau=1}^{n} \sum_{k=1}^{H} \eta^{(\tau)}_{\widehat{w}^{(n)}_{H+1},k}(\tilde{a}^{(\tau)}_k)\phi(x^{(\tau)}_k, \tilde{a}^{(\tau)}_k) \right\|_\infty + |\mathcal{A}| \sqrt{nH \log \frac{dHn^2}{\delta}} \left\| \widehat{w}^{\mathrm{Im}(n)}_h - \bar{w}^{(n)}_h \right\|_1 \\
&= \frac{64\,|\mathcal{A}|\, s_\star \sqrt{H}}{\sigma_U} \log \frac{dHn^2}{\delta} \\
&\leq 9\,|\mathcal{A}|\, H \sqrt{n \log \frac{dHn^2}{\delta}} \\
&= \lambda^{(n)}_{\mathrm{Est}},
\end{aligned}$$

where the last inequality holds because $n \geq 2^{12} s_\star^2 \sigma_U^{-2} H^{-1} \log(dHn^2/\delta)$ for $n \geq n_1$. Suppose (22) holds for steps $H, \ldots, h+1$. Then by Lemma B.4 and (18),

$$\max_{h' \geq h+1} \left\| \widehat{w}^{(n)}_{h'} - \bar{w}^{(n)}_{h'} \right\|_1 \leq \frac{8\lambda^{(n)}_{\mathrm{Est}} s_\star}{nH\,|\mathcal{A}|\,\sigma_U} = \frac{72 s_\star}{\sqrt{n}\sigma_U} \sqrt{\log \frac{dHn^2}{\delta}}.$$

We decompose

$$\begin{aligned}
&\left\| \sum_{\tau=1}^{n} \sum_{k=1}^{H} \eta^{(\tau)}_{\widehat{w}^{(n)}_{h+1},k}(\tilde{a}^{(\tau)}_k)\phi(x^{(\tau)}_k, \tilde{a}^{(\tau)}_k) \right\|_\infty \\
&\leq \left\| \sum_{\tau=1}^{n} \sum_{k=1}^{H} \left\{ \eta^{(\tau)}_{\widehat{w}^{(n)}_{h+1},k}(\tilde{a}^{(\tau)}_k) - \eta^{(\tau)}_{w^\star_{h+1},k}(\tilde{a}^{(\tau)}_k) \right\} \phi(x^{(\tau)}_k, \tilde{a}^{(\tau)}_k) \right\|_\infty + \left\| \sum_{\tau=1}^{n} \sum_{k=1}^{H} \eta^{(\tau)}_{w^\star_{h+1},k}(\tilde{a}^{(\tau)}_k)\phi(x^{(\tau)}_k, \tilde{a}^{(\tau)}_k) \right\|_\infty
\end{aligned}$$

Because $\widehat{w}^{(n)}_{h+1}$ depends on the data, we take supremum over $\mathcal{W}_{h+1} := \{ w \in \mathbb{R}^d : \|w - w^\star_{h+1}\|_1 \leq \sqrt{H}/(52 \log \frac{Hdn^2}{\delta} \sqrt{\log 2d}) \}$. To prove $\widehat{w}^{(n)}_{h+1} \in \mathcal{W}_{h+1}$, we observe

$$\begin{aligned}
\left\| \widehat{w}^{(n)}_{h+1} - w^\star_{h+1} \right\|_1 &\leq \left\| \widehat{w}^{(n)}_{h+1} - \bar{w}^{(n)}_{h+1} \right\|_1 + \left\| \bar{w}^{(n)}_{h+1} - w^\star_{h+1} \right\|_1 \\
&\leq \left\| \widehat{w}^{(n)}_{h+1} - \bar{w}^{(n)}_{h+1} \right\|_1 + \sqrt{s_\star} \left\| \bar{w}^{(n)}_{h+1} - w^\star_{h+1} \right\|_2 \\
&\leq \left\| \widehat{w}^{(n)}_{h+1} - \bar{w}^{(n)}_{h+1} \right\|_1 + \sqrt{\frac{2s_\star}{\sigma_U}} \left\| \bar{w}^{(n)}_{h+1} - w^\star_{h+1} \right\|_{\frac{1}{nH} \sum_{\tau=1}^{n} \sum_{k=1}^{H} \phi(\tilde{x}^{(\tau)}_k, \tilde{a}^{(\tau)}_k)\phi(\tilde{x}^{(\tau)}_k, \tilde{a}^{(\tau)}_k)^\top},
\end{aligned}$$

where the last inequality holds by Lemma B.3 and (20). By definition of $\bar{w}_{h+1}^{(n)}$,

$$
\begin{aligned}
&\left\|\bar{w}_{h+1}^{(n)} - w_{h+1}^{\star}\right\|^2_{\frac{1}{nH}\sum_{\tau=1}^{n}\sum_{k=1}^{H}\phi(\tilde{x}_k^{(\tau)},\tilde{a}_k^{(\tau)})\phi(\tilde{x}_k^{(\tau)},\tilde{a}_k^{(\tau)})^\top} \\
&\le \max_{(x,a)\in\mathcal{X}\times\mathcal{A}} \left|\phi(x,a)^\top(\bar{w}_{h+1}^{(n)} - w_{h+1}^{\star})\right| \\
&\le \max_{(x,a)\in\mathcal{X}\times\mathcal{A}} \left|\int \left\{\Pi_{[0,H]}\left(\max_{a'\in\mathcal{A}}\widehat{Q}_{\widehat{w}_{h+2}^{(n)}}(x',a)\right) - V_{h+1}^{\star}(x')\right\}\phi(x,a)^\top\psi(x')dx'\right| \\
&\le \max_{x'\in\mathcal{X}}\left|\Pi_{[0,H]}\left(\max_{a'\in\mathcal{A}}\widehat{Q}_{\widehat{w}_{h+2}^{(n)}}(x',a)\right) - \max_{a'\in\mathcal{A}} Q_{h+1}^{\star}(x',a)\right|\int \phi(x,a)^\top\psi(x')dx' \\
&= \max_{x'\in\mathcal{X}}\left|\Pi_{[0,H]}\left(\max_{a'\in\mathcal{A}}\widehat{Q}_{\widehat{w}_{h+2}^{(n)}}(x',a)\right) - \max_{a'\in\mathcal{A}} Q_{h+1}^{\star}(x',a)\right|.
\end{aligned}
$$

By definition of $Q_{h+1}^{\star}(x',a) = r(x',a) + \phi(x',a)^\top w_{h+2}^{\star} = \widehat{Q}_{w_{h+2}^{\star}}(x',a) \in [0,H]$, we obtain

$$
\begin{aligned}
&\left\|\bar{w}_{h+1}^{(n)} - w_{h+1}^{\star}\right\|^2_{\frac{1}{nH}\sum_{\tau=1}^{n}\sum_{k=1}^{H}\phi(\tilde{x}_k^{(\tau)},\tilde{a}_k^{(\tau)})\phi(\tilde{x}_k^{(\tau)},\tilde{a}_k^{(\tau)})^\top} \\
&\le \max_{x'\in\mathcal{X}}\left|\Pi_{[0,H]}\left(\max_{a'\in\mathcal{A}}\widehat{Q}_{\widehat{w}_{h+2}^{(n)}}(x',a)\right) - \max_{a'\in\mathcal{A}}\widehat{Q}_{w_{h+2}^{\star}}(x',a)\right| \\
&= \max_{x'\in\mathcal{X}}\left|\Pi_{[0,H]}\left(\max_{a'\in\mathcal{A}}\widehat{Q}_{\widehat{w}_{h+2}^{(n)}}(x',a)\right) - \Pi_{[0,H]}\left(\max_{a'\in\mathcal{A}}\widehat{Q}_{w_{h+2}^{\star}}(x',a)\right)\right| \\
&\le \max_{x'\in\mathcal{X}}\max_{a'\in\mathcal{A}}\left|\widehat{Q}_{\widehat{w}_{h+2}^{(n)}}(x',a) - \max_{a'\in\mathcal{A}}\widehat{Q}_{w_{h+2}^{\star}}(x',a)\right| \\
&\le \max_{(x,a)\in\mathcal{X}\times\mathcal{A}}\left|\phi(x,a)^\top\left(\widehat{w}_{h+2}^{(n)} - w_{h+2}^{\star}\right)\right|.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\left\|\widehat{w}_{h+1}^{(n)} - w_{h+1}^{\star}\right\|_1 &\le \left\|\widehat{w}_{h+1}^{(n)} - \bar{w}_{h+1}^{(n)}\right\|_1 + \sqrt{\frac{2s_\star}{\sigma_U}\max_{(x,a)\in\mathcal{X}\times\mathcal{A}}\left|\phi(x,a)^\top\left(\widehat{w}_{h+2}^{(n)} - w_{h+2}^{\star}\right)\right|} \\
&\le \left\|\widehat{w}_{h+1}^{(n)} - \bar{w}_{h+1}^{(n)}\right\|_1 + \sqrt{\frac{2s_\star}{\sigma_U}\left(\left\|\widehat{w}_{h+2}^{(n)} - \bar{w}_{h+2}^{(n)}\right\|_1 + \max_{(x,a)\in\mathcal{X}\times\mathcal{A}}\left|\phi(x,a)^\top\left(\bar{w}_{h+2}^{(n)} - w_{h+2}^{\star}\right)\right|\right)}
\end{aligned}
$$

Applying the inequality recursively,

$$
\left\|\widehat{w}_{h+1}^{(n)} - w_{h+1}^{\star}\right\|_1 \le \left\|\widehat{w}_{h+1}^{(n)} - \bar{w}_{h+1}^{(n)}\right\|_1 + \sqrt{\frac{2s_\star}{\sigma_U}\sum_{h'=h+2}^{H}\left\|\widehat{w}_{h'}^{(n)} - \bar{w}_{h'}^{(n)}\right\|_1}.
$$

By inductive assumption,

$$
\begin{aligned}
\left\|\widehat{w}_{h+1}^{(n)} - w_{h+1}^{\star}\right\|_1 \le& \frac{72 s_\star}{\sqrt{n}\sigma_U}\sqrt{\log\frac{dHn^2}{\delta}} + \sqrt{\frac{144 H s_\star^2}{\sqrt{n}\sigma_U^2}\sqrt{\log\frac{dHn^2}{\delta}}} \\
\le& \frac{\sqrt{H}}{52\log\frac{Hdn^2}{\delta}\sqrt{\log 2d}}
\end{aligned}
$$

where the last inequality holds by $n \ge C s_\star^4 \sigma_U^{-4} H^{-2}\log^5(dHn^2/\delta)\log^2(2d)$ for some absolute constant $C := (144)^2\cdot(52)^4 + 8\cdot(72)^2(52)^2$ for $n \ge n_1$. Thus, we obtain $\widehat{w}_{h+1}^{(n)} \in \mathcal{W}_{h+1}$, and

$$
\begin{aligned}
&\left\|\sum_{\tau=1}^{n}\sum_{k=1}^{H}\eta_{\widehat{w}_{h+1}^{(n)},k}^{(\tau)}(\tilde{a}_k^{(\tau)})\phi(x_k^{(\tau)},\tilde{a}_k^{(\tau)})\right\|_\infty \\
&\le \sup_{w\in\mathcal{W}_{h+1}}\left\|\sum_{\tau=1}^{n}\sum_{k=1}^{H}\left\{\eta_{w,k}^{(\tau)}(\tilde{a}_k^{(\tau)}) - \eta_{w_{h+1}^{\star},k}^{(\tau)}(\tilde{a}_k^{(\tau)})\right\}\phi(x_k^{(\tau)},\tilde{a}_k^{(\tau)})\right\|_\infty + \left\|\sum_{\tau=1}^{n}\sum_{k=1}^{H}\eta_{w_{h+1}^{\star},k}^{(\tau)}(\tilde{a}_k^{(\tau)})\phi(x_k^{(\tau)},\tilde{a}_k^{(\tau)})\right\|_\infty
\end{aligned}
$$

By Lemma 5.3, with probability at least $1-\delta/(Hn^2)$

$$\sup_{w\in\mathcal{W}_{h+1}}\left\|\sum_{\tau=1}^{n}\sum_{k=1}^{H}\left\{\eta_{w,k}^{(\tau)}(\tilde{a}_k^{(\tau)})-\eta_{w_{h+1}^\star,k}^{(\tau)}(\tilde{a}_k^{(\tau)})\right\}\phi(x_k^{(\tau)},\tilde{a}_k^{(\tau)})\right\|_\infty\leq 3H\sqrt{n\log\frac{dHn^2}{\delta}} \tag{23}$$

Note that $\|\eta_h^w(x_u^{(\tau)},a_u^{(\tau)})\phi(x_u^{(\tau)},a_u^{(\tau)})\|_\infty\leq H$. By Lemma B.8, with probability at least $1-2\delta/(Hn^2)$,

$$\begin{aligned}&\left\|\sum_{\tau=1}^{n}\sum_{k=1}^{H}\eta_{w_{h+1}^\star,k}^{(\tau)}(\tilde{a}_k^{(\tau)})\phi(x_k^{(\tau)},\tilde{a}_k^{(\tau)})\right\|_\infty\\&\leq\frac{2H}{3}\log\frac{dHn^2}{\delta}+\sqrt{2\sum_{\tau=1}^{n}\sum_{k=1}^{H}\mathbb{E}\left[\left(\eta_{w_{h+1}^\star,k}^{(\tau)}(\tilde{a}_k^{(\tau)})\right)^2\middle|\mathcal{H}_k^{(\tau)}\right]\log\frac{dHn^2}{\delta}},\end{aligned}$$

where $\mathcal{H}_k^{(\tau)}$ is a sigma algebra generated by $\{x_{h'}^{(u)},a_{h'}^{(u)}\}_{u\in[\tau-1],h'\in[H]}\cup\{x_{h'}^{(\tau)},a_{h'}^{(\tau)}\}_{h'\in[k]}$. Note that

$$\begin{aligned}\Pi_{[0,H]}\left(\max_{a'\in\mathcal{A}}\widehat{Q}_{w_{h+1}^\star}(x,a')\right)=&\Pi_{[0,H]}\left(\max_{a\in\mathcal{A}}\left\{r(x,a)+\phi(x,a)^\top w_{h+1}^\star\right\}\right)\\=&\Pi_{[0,H]}\left(\max_{a\in\mathcal{A}}\left\{r(x,a)+\left[\mathbb{P}V_{h+1}^\star\right](x,a)\right\}\right)\\=&\Pi_{[0,H]}\left(\max_{a\in\mathcal{A}}Q_h^\star(x,a)\right)\\=&\Pi_{[0,H]}\left(V_h^\star(x)\right)\\=&V_h^\star(x).\end{aligned}$$

By definition (21),

$$\begin{aligned}\eta_{w_{h+1}^\star,k}^{(\tau)}(\tilde{a}_k^{(\tau)})=&\Pi_{[0,H]}\left(\max_{a'\in\mathcal{A}}\widehat{Q}_{w_{h+1}^\star}(X_{k+1}^{(\tau)}(\tilde{a}_k^{(\tau)}),a')\right)-\mathbb{E}_{X\sim\mathbb{P}\left(\cdot|x_k^{(\tau)},\tilde{a}_k^{(\tau)}\right)}\left[\Pi_{[0,H]}\left(\max_{a'\in\mathcal{A}}\widehat{Q}_{w_{h+1}^\star}(X,a')\right)\right]\\=&V_h^\star(X_{k+1}^{(\tau)}(\tilde{a}_k^{(\tau)}))-[\mathbb{P}V_h^\star](x_k^{(\tau)},\tilde{a}_k^{(\tau)}).\end{aligned}$$

Applying Lemma B.7, with probability at least $1-2\delta/(Hn^2)$,

$$\sum_{\tau=1}^{n}\sum_{k=1}^{H}\mathbb{E}\left[\eta_{w_{h+1}^\star,k}^{(\tau)}(\tilde{a}_k^{(\tau)})^2\middle|\mathcal{H}_k^{(\tau)}\right]\leq\sum_{\tau=1}^{n}\mathbb{E}\left[\sum_{k=1}^{H}\mathbb{E}\left[\eta_{w_{h+1}^\star,k}^{(\tau)}(\tilde{a}_k^{(\tau)})^2\middle|\mathcal{H}_k^{(\tau)}\right]\right]+H^2\sqrt{2n\log\frac{Hn^2}{\delta}}.$$

Using the variance bound (Lemma 5.4) we get

$$\begin{aligned}\sum_{\tau=1}^{n}\sum_{k=1}^{H}\mathbb{E}\left[\eta_{w_{h+1}^\star,k}^{(\tau)}(\tilde{a}_k^{(\tau)})^2\middle|\mathcal{H}_k^{(\tau)}\right]\leq&5n(H^2+H)+H^3\sqrt{2n\log\frac{Hn^2}{\delta}}\\\leq&11nH^2,\end{aligned}$$

where the last inequality holds by $n\geq n_1\geq 2H^2\log(Hn^2/\delta)$. Thus,

$$\left\|\sum_{\tau=1}^{n}\sum_{k=1}^{H}\eta_{w_{h+1}^\star,k}^{(\tau)}(\tilde{a}_k^{(\tau)})\phi(x_k^{(\tau)},\tilde{a}_k^{(\tau)})\right\|_\infty\leq\frac{2H}{3}\log\frac{dHn^2}{\delta}+H\sqrt{22n\log\frac{dHn^2}{\delta}}\leq 5H\sqrt{n\log\frac{dHn^2}{\delta}},$$

where the last inequality holds by $n\geq n_1\geq(100/9)\log(dHn^2/\delta)$. Thus, we obtain

$$\left\|\sum_{\tau=1}^{n}\sum_{k=1}^{H}\eta_{\widehat{w}_{h+1}^{(n)},k}^{(\tau)}(\tilde{a}_k^{(\tau)})\phi_k^{(\tau)}\right\|_\infty\leq 8H\sqrt{n\log\frac{dHn^2}{\delta}}$$

which implies

$$\begin{aligned}
&|\mathcal{A}| \left\| \sum_{\tau=1}^{n} \sum_{k=1}^{H} \eta^{(\tau)}_{\widehat{w}^{(n)}_{h+1},k} (\tilde{a}^{(\tau)}_k) \phi(x^{(\tau)}_k, \tilde{a}^{(\tau)}_k) \right\|_\infty + |\mathcal{A}| \sqrt{nH \log \frac{dHn^2}{\delta}} \left\| \widehat{w}^{\mathrm{Im}(n)}_h - \bar{w}^{(n)}_h \right\|_1 \\
&\le 8 \, |\mathcal{A}| \, H \sqrt{n \log \frac{dHn^2}{\delta}} + |\mathcal{A}| \sqrt{nH \log \frac{dHn^2}{\delta}} \left\| \widehat{w}^{\mathrm{Im}(n)}_h - \bar{w}^{(n)}_h \right\|_1 .
\end{aligned}$$

By using a similar argument, we obtain with probability at least $1 - 5\delta/(Hn^2)$,

$$\left\| \sum_{\tau=1}^{n} \sum_{k=1}^{H} \eta^{(\tau)}_{\widehat{w}^{(n)}_{h+1},k} (a^{(\tau)}_k) \phi^{(\tau)}_k \right\|_\infty \le \lambda^{(n)}_{\mathrm{Im}}, \tag{24}$$

Using Lemma B.4,

$$\left\| \widehat{w}^{\mathrm{Im}(n)}_h - \bar{w}^{(n)}_h \right\|_1 \le \frac{8\lambda^{(n)}_{\mathrm{Im}} s_\star}{\sigma_U H n} = \frac{64 s_\star}{\sigma_U \sqrt{n}} \sqrt{\log \frac{dHn^2}{\delta}},$$

which implies

$$\begin{aligned}
&|\mathcal{A}| \left\| \sum_{\tau=1}^{n} \sum_{k=1}^{H} \eta^{(\tau)}_{\widehat{w}^{(n)}_{h+1},k} (\tilde{a}^{(\tau)}_k) \phi(x^{(\tau)}_k, \tilde{a}^{(\tau)}_k) \right\|_\infty + |\mathcal{A}| \sqrt{nH \log \frac{dHn^2}{\delta}} \left\| \widehat{w}^{\mathrm{Im}(n)}_h - \bar{w}^{(n)}_h \right\|_1 \\
&\le 8 \, |\mathcal{A}| \, H \sqrt{n \log \frac{dHn^2}{\delta}} + \frac{64 s_\star \sqrt{H}}{\sigma_U} \log \frac{dHn^2}{\delta} \\
&\le 9 \, |\mathcal{A}| \, H \sqrt{n \log \frac{dHn^2}{\delta}},
\end{aligned}$$

where the last inequality holds by $n \ge n_1$. Therefore we conclude

$$\left\| \sum_{\tau=1}^{n} \sum_{k=1}^{H} \sum_{a \in \mathcal{A}} \tilde{\eta}^{(\tau)}_{\widehat{w}^{(n)}_{h+1},k} (a) \phi(x^{(\tau)}_k, a) \right\|_\infty \le 9 \, |\mathcal{A}| \, H \sqrt{n \log \frac{dHn^2}{\delta}} = \lambda^{(n)}_{\mathrm{Est}}$$

□

C.4 Proof of Lemma 5.3

Proof. Fix the policies $\pi^{(1)}, \ldots, \pi^{(n)}$ and set $Z^{(\tau)}_k := (x^{(\tau)}_k, a^{(\tau)}_k, x^{(\tau)}_{k+1}, u^{(\tau)}_k)$ and $\mathcal{Z} := (\mathcal{X} \times \mathcal{A})^2$, where $\{u^{(\tau)}_k\}_{k \in [H], \tau \in [n]}$ are the IID Uniform random variables over $\mathcal{A}$. Let $\mathcal{H}^{(\tau)}_k$ denote the sigma-algebra generated by $\{x^{(u)}_v, a^{(u)}_v\}_{v \in [H], u \in [\tau-1]} \cup \{x^{(\tau)}_v, a^{(\tau)}_v\}_{v \in [k]}$ with $\mathcal{H}^{(\tau)}_0 := \mathcal{H}^{(\tau-1)}_H$. For $i \in [d]$,

$$f_{w,i}(x_1, a, x_2, u) := \rho^{-1} \left\{ \Pi_{[0,H]} \left(\max_{a' \in \mathcal{A}} \widehat{Q}_w(x_2, a') \right) - \Pi_{[0,H]} \left(\max_{a' \in \mathcal{A}} \widehat{Q}_{w^\star_{h+1}}(x_2, a') \right) \right\} \phi(x_1, a)(i)$$

Then with the function class $\mathcal{F}_i := \{f_{w,i} : w \in \mathcal{W}_{h+1}\}$,

$$\sup_{w \in \mathcal{W}_{h+1}} \left\| \sum_{\tau=1}^{n} \sum_{k=1}^{H} \left\{ \eta^{(\tau)}_{w,k} - \eta^{(\tau)}_{w^\star_{h+1},k} \right\} \phi^{(\tau)}_k \right\|_\infty = \rho \max_{i \in [d]} \sup_{f_i \in \mathcal{F}_i} \left| \sum_{\tau=1}^{n} \sum_{k=1}^{H} f_i(x^{(\tau)}_k, a^{(\tau)}_k, x^{(\tau)}_{k+1}, u^{(\tau)}_k) \right|.$$

Note that for any $f \in \mathcal{F}_i$, there exists $w \in \mathcal{W}_{h+1}$ such that

$$\begin{aligned}
\max_{(x_1,a,x_2,u) \in \mathcal{Z}} |f(x_1, a, x_2, u)| &\le \rho^{-1} \left| \Pi_{[0,H]} \left(\max_{a' \in \mathcal{A}} \widehat{Q}_w(x_2, a') \right) - \Pi_{[0,H]} \left(\max_{a' \in \mathcal{A}} \widehat{Q}_{w^\star_{h+1}}(x_2, a') \right) \right| \\
&\le \rho^{-1} \max_{a' \in \mathcal{A}} \left| \widehat{Q}^w_h(x, a') - \widehat{Q}^{w^\star_{h+1}}_h(x, a') \right| \\
&\le \rho^{-1} \left\| w - w^\star_{h+1} \right\|_1 \\
&\le 1.
\end{aligned}$$

Let $\mathbf{z} := (\mathbf{z}_1^{(1)}, \dots, \mathbf{z}_H^{(n)})$ denote a sequence of binary tree such that $\mathbf{z}_k^{(\tau)} : \{\pm 1\}^{\tau H + k} \to \mathcal{Z}$ and $\xi := (\xi_1^{(1)}, \dots, \xi_H^{(n)})$ denote a sequence of IID Bernoulli random variables with $\mathbb{P}(\xi_1^{(1)} = -1) = \mathbb{P}(\xi_1^{(1)} = 1) = 1/2$. By Lemma B.5, for any $x > 0$,

$$\mathbb{P}\left(\max_{i \in [d]} \sup_{f \in \mathcal{F}_i} \left| \sum_{\tau=1}^{n} \sum_{k=1}^{H} f(x_k^{(\tau)}, a_k^{(\tau)}, x_{k+1}^{(\tau)}, u_k^{(\tau)}) \right| > x \right) \le 4 \sum_{i=1}^{d} \sup_{\mathbf{z}} \mathbb{P}\left(\sup_{f \in \mathcal{F}_i} \left| \sum_{\tau=1}^{n} \sum_{k=1}^{H} \xi_k^{(\tau)} f(\mathbf{z}_k^{(\tau)}(\xi)) \right| > \frac{x}{4} \middle| \mathbf{z} \right).$$

By Lemma B.6, setting

$$x = 8 \sup_{\mathbf{z}} \inf_{\alpha > 0} \left\{ nH\alpha + 2 \int_{\alpha}^{1/2} \sqrt{3nH \log \frac{N(\epsilon, \mathcal{F}, \|\cdot\|_{\infty, \mathbf{z}})\sqrt{n^2 H d}}{\sqrt{\delta}}} d\epsilon \right\}$$

we obtain

$$\mathbb{P}\left(\max_{i \in [d]} \sup_{f \in \mathcal{F}_i} \left| \sum_{\tau=1}^{n} \sum_{k=1}^{H} f(x_k^{(\tau)}, a_k^{(\tau)}, x_{k+1}^{(\tau)}, u_k^{(\tau)}) \right| > x \right) \le \frac{\delta}{n^2 H d}.$$

To find an upper bound for x, define a function $g_{w,i} : \mathcal{Z} \to \mathbb{R}$ by

$$g_{w,i}(x_1, a, x_2, u) := L^{-1} \phi(x_2, u)^\top w \phi(x_1, a)(i),$$

and a function class $\mathcal{G}_i := \{ g_{w,i} - g_{w_{h+1}^\star, i} : w \in \mathcal{W}_{h+1} \}$. Given $\epsilon > 0$ and a binary tree $\mathbf{z} := (\mathbf{z}_1^{(1)}, \dots, \mathbf{z}_H^{(n)}) = ((x_1^{(1)}, a_1^{(1)}, x_2^{(1)}, u_1^{(1)}), \dots (x_H^{(n)}, a_H^{(n)}, x_{H+1}^{(n)}, u_H^{(n)}))$, for any $f \in \mathcal{F}_i$, there exists $g_{\tilde{w},i}$ in the ϵ-cover of $\mathcal{G}_i$ such that

$$\begin{aligned}
&\max_{\tau,k} \Bigg| f_{w,i}(\mathbf{z}_k^{(\tau)}) - \Pi_{[0,H]} \left(\max_{a' \in \mathcal{A}} \widehat{Q}_h^{w_{h+1}^\star}(x_{k+1}^{(\tau)}, a') \right) \phi(x_k^{(\tau)}, a_k^{(\tau)})(i) \\
&\quad - \Pi_{[0,H]} \left(\max_{u \in \mathcal{A}} r(u, x_{k+1}^{(\tau)}) + g_{\tilde{w},i}(x_k^{(\tau)}, a_k^{(\tau)}, x_{k+1}^{(\tau)}, u) \right) \phi(x_k^{(\tau)}, a_k^{(\tau)})(i) \Bigg| \\
&\le \max_{\tau,k} \left| \Pi_{[0,H]} \left(\max_{u \in \mathcal{A}} r(u, x_{k+1}^{(\tau)}) + g_{w,i}(x_k^{(\tau)}, a_k^{(\tau)}, x_{k+1}^{(\tau)}, u) \right) - \Pi_{[0,H]} \left(\max_{u \in \mathcal{A}} r(u, x_{k+1}^{(\tau)}) + g_{\tilde{w},i}(x_k^{(\tau)}, a_k^{(\tau)}, x_{k+1}^{(\tau)}, u) \right) \right| \\
&\le \max_{\tau,k} \max_{u \in \mathcal{A}} \left| g_{w,i}(x_k^{(\tau)}, a_k^{(\tau)}, x_{k+1}^{(\tau)}, u) - g_{\tilde{w},i}(x_k^{(\tau)}, a_k^{(\tau)}, x_{k+1}^{(\tau)}, u) \right| \\
&\le \max_{\tau,k} \left| g_{w,i}(\mathbf{z}_k^{(\tau)}) - g_{w_{h+1}^\star, i}(\mathbf{z}_k^{(\tau)}) + g_{w_{h+1}^\star, i}(\mathbf{z}_k^{(\tau)}) - g_{\tilde{w},i}(\mathbf{z}_k^{(\tau)}) \right| \\
&\le \epsilon.
\end{aligned}$$

Thus ,$N(\epsilon, \mathcal{F}_i, \|\cdot\|_{\infty,\mathbf{z}}) \le N(\epsilon, \mathcal{G}_i, \|\cdot\|_{\infty,\mathbf{z}})$ and

$$2 \int_{\alpha}^{1/2} \sqrt{3nH \log \frac{N(\epsilon, \mathcal{F}_i, \|\cdot\|_{\infty,\mathbf{z}})\sqrt{n^2 H d}}{\sqrt{\delta}}} d\epsilon \le 2 \int_{\alpha}^{1/2} \sqrt{3nH \log \frac{N(\epsilon, \mathcal{G}_i, \|\cdot\|_{\infty,\mathbf{z}})\sqrt{n^2 H d}}{\sqrt{\delta}}} d\epsilon.$$

Define the sequential Rademacher complexity,

$$R_H^{(n)}(\mathcal{G}_i) := \sup_{\mathbf{z}} \mathbb{E}\left[\sup_{g \in \mathcal{G}_i} \sum_{\tau=1}^{n} \sum_{k=1}^{H} \xi_k^{(\tau)} g(\mathbf{z}_k^{(\tau)}(\xi)) \right].$$

Note that the Rachmechar complexity is bounded as

$$\begin{aligned}
R_H^{(n)}(\mathcal{G}_i) =& \sup_{\mathbf{z}} \mathbb{E}\left[\sup_{g \in \mathcal{G}_i} \sum_{\tau=1}^{n} \sum_{k=1}^{H} \xi_k^{(\tau)} g(\mathbf{z}_k^{(\tau)}(\xi)) \right] \\
=& \rho^{-1} \sup_{\mathbf{z}} \mathbb{E}\left[\sup_{w \in \mathcal{W}_{h+1}} \sum_{\tau=1}^{n} \sum_{k=1}^{H} \xi_k^{(\tau)} \phi(x_k^{(\tau)}, u_k^{(\tau)})^\top \left(w - w_{h+1}^\star \right) \phi(x_k^{(\tau)}, a_k^{(\tau)})(i) \right] \\
\le& \rho^{-1} \sup_{\mathbf{z}} \sup_{w \in \mathcal{W}_{h+1}} \left\| w - w_{h+1}^\star \right\|_1 \mathbb{E}\left[\left\| \sum_{\tau=1}^{n} \sum_{k=1}^{H} \xi_k^{(\tau)} \phi(x_k^{(\tau)}, u_k^{(\tau)}) \phi(x_k^{(\tau)}, a_k^{(\tau)})(i) \right\|_\infty \right] \\
\le& \sup_{\mathbf{z}} \mathbb{E}\left[\left\| \sum_{\tau=1}^{n} \sum_{k=1}^{H} \xi_k^{(\tau)} \phi(x_k^{(\tau)}, u_k^{(\tau)}) \phi(x_k^{(\tau)}, a_k^{(\tau)})(i) \right\|_\infty \right].
\end{aligned}$$

By Jensen's inequality, for any $\lambda > 0$,

$$\begin{aligned}
&\mathbb{E}\left[\left\|\sum_{\tau=1}^{n}\sum_{k=1}^{H}\xi_k^{(\tau)}\phi(x_k^{(\tau)},u_k^{(\tau)})\phi(x_k^{(\tau)},a_k^{(\tau)})(i)\right\|_\infty\right]\\
&\leq \frac{1}{\lambda}\log\mathbb{E}\exp\left(\lambda\left\|\sum_{\tau=1}^{n}\sum_{k=1}^{H}\xi_k^{(\tau)}\phi(x_k^{(\tau)},u_k^{(\tau)})\phi(x_k^{(\tau)},a_k^{(\tau)})(i)\right\|_\infty\right)\\
&\leq \frac{1}{\lambda}\log\sum_{j\in[d]}\mathbb{E}\exp\left(\lambda\left|\sum_{\tau=1}^{n}\sum_{k=1}^{H}\xi_k^{(\tau)}\phi(x_k^{(\tau)},u_k^{(\tau)})(j)\phi(x_k^{(\tau)},a_k^{(\tau)})(i)\right|\right)\\
&\leq \frac{1}{\lambda}\log\left\{\sum_{j\in[d]}\mathbb{E}\exp\left(\lambda\sum_{\tau=1}^{n}\sum_{k=1}^{H}\xi_k^{(\tau)}\phi(x_k^{(\tau)},u_k^{(\tau)})(j)\phi(x_k^{(\tau)},a_k^{(\tau)})(i)\right)\right.\\
&\left.+\mathbb{E}\exp\left(-\lambda\sum_{\tau=1}^{n}\sum_{k=1}^{H}\xi_k^{(\tau)}\phi(x_k^{(\tau)},u_k^{(\tau)})(j)\phi(x_k^{(\tau)},a_k^{(\tau)})(i)\right)\right\}\\
&\leq \frac{1}{\lambda}\log 2d\exp\left(\frac{\lambda^2 nH}{2}\right)\\
&= \frac{\log 2d}{\lambda}+\frac{\lambda nH}{2}.
\end{aligned}$$

where the last inequality uses $\|\phi(x,a)\|_\infty \leq 1$. Minimizing over $\lambda > 0$ gives

$$\mathbb{E}\left[\left\|\sum_{\tau=1}^{n}\sum_{k=1}^{H}\xi_k^{(\tau)}\phi(x_k^{(\tau)},u_k^{(\tau)})\phi(x_k^{(\tau)},a_k^{(\tau)})(i)\right\|_\infty\right] \leq \sqrt{\frac{nH\log 2d}{2}}.$$

Therefore we obtain,

$$R_H^{(n)}(\mathcal{G}_i) \leq \sqrt{\frac{nH\log 2d}{2}}.$$

Setting $\alpha = \sqrt{\frac{2\log 2d}{nH}}$,

$$\begin{aligned}
&\sup_{\mathbf{z}}\inf_{\alpha>0}\left\{nH\alpha + 2\int_{\alpha}^{1/2}\sqrt{3nH\log\frac{N(\epsilon,\mathcal{G}_i,\|\cdot\|_{\infty,\mathbf{z}})\sqrt{n^2Hd}}{\sqrt{\delta}}}d\epsilon\right\}\\
&\leq \sqrt{2nH\log 2d} + 2\sup_{\mathbf{z}}\int_{\sqrt{\frac{2\log 2d}{nH}}}^{1/2}\sqrt{3nH\log\frac{N(\epsilon,\mathcal{G}_i,\|\cdot\|_{\infty,\mathbf{z}})\sqrt{n^2Hd}}{\sqrt{\delta}}}d\epsilon.
\end{aligned}$$

By Corollary 1 and Lemma 2 in Rakhlin et al. (2015), whenever $\epsilon \geq \sqrt{\frac{2\log 2d}{nH}} \geq 2n^{-1}H^{-1}R_H^{(n)}(\mathcal{G}_i)$,

$$\log N(\epsilon,\mathcal{G},\|\cdot\|_{\infty,\mathbf{z}}) \leq \frac{32}{nH\epsilon^2}R_H^{(n)}(\mathcal{G}_i)^2\log\frac{2enH}{\epsilon}.$$

Thus,

$$\begin{aligned}
&\int_{\sqrt{\frac{2\log 2d}{nH}}}^{1/2}\sqrt{3nH\log\frac{N(\epsilon,\mathcal{G}_i,\|\cdot\|_{\infty,\mathbf{z}})\sqrt{n^2Hd}}{\sqrt{\delta}}}d\epsilon\\
&\leq 4\sqrt{6}R_H^{(n)}(\mathcal{G}_i)\int_{\sqrt{\frac{2\log 2d}{nH}}}^{1/2}\frac{1}{\epsilon}\sqrt{-\log\epsilon+\log\frac{2enH\sqrt{n^2Hd}}{\sqrt{\delta}}}d\epsilon\\
&= 4\sqrt{6}R_H^{(n)}(\mathcal{G}_i)\left[-\frac{2}{3}\left(-\log\epsilon+\log\frac{2enH\sqrt{n^2Hd}}{\sqrt{\delta}}\right)^{3/2}\right]_{\sqrt{\frac{2\log 2d}{nH}}}^{1/2}\\
&\leq \frac{8\sqrt{6}}{3}R_H^{(n)}(\mathcal{G}_i)\left(\log\sqrt{\frac{nH}{2\log 2d}}+\log\frac{2enH\sqrt{n^2Hd}}{\sqrt{\delta}}\right)^{3/2}.
\end{aligned}$$

Now we obtain

$$
\begin{aligned}
x \leq & 8\sqrt{2nH\log 2d} + 16\frac{8\sqrt{6}}{3} R_H^{(n)}(\mathcal{G}_i) \log^{3/2} \frac{2enH^2\sqrt{n^3 d}}{\sqrt{2\delta \log 2d}} \\
\leq & 8\sqrt{2nH\log 2d} + \frac{128\sqrt{3}}{3}\sqrt{nH\log 2d} \log^{3/2} \frac{4enH^2\sqrt{n^3 d}}{\sqrt{2\delta \log 2d}} \\
\leq & 8\sqrt{2nH\log 2d} + \frac{256\sqrt{3}}{3}\sqrt{2nH\log 2d} \log^{3/2} \frac{2\sqrt{e}Hdn^{\frac{5}{4}}}{\delta} \\
\leq & 8\sqrt{2nH\log 2d} + \frac{256\sqrt{3}}{3}\sqrt{2nH\log 2d} \log^{3/2} \frac{Hdn^2}{\delta},
\end{aligned}
$$

the last inequality holds by $n^3 \geq 16e^2$. Thus we conclude with probability at least $1-\delta/(Hn^2)$

$$
\begin{aligned}
\sup_{w \in \mathcal{W}_{h+1}(\rho)} \left\| \sum_{\tau=1}^{n}\sum_{k=1}^{H} \left\{ \eta_{w,k}^{(\tau)} - \eta_{w_{h+1}^{\star},k}^{(\tau)} \right\} \phi_k^{(\tau)} \right\|_\infty \leq & \rho \left(8\sqrt{2nH\log 2d} + \frac{256\sqrt{3}}{3}\sqrt{2nH\log 2d} \log^{3/2} \frac{Hdn^2}{\delta} \right) \\
\leq & \rho\sqrt{2nH\log 2d} \left(8 + \frac{256\sqrt{3}}{3} \log^{3/2} \frac{Hdn^2}{\delta} \right)
\end{aligned}
$$

□

C.5 Proof of Lemma 5.4

Proof. For each $k \in [H]$, the definition of action value function $Q_{h-1}^{\star}(x,a)$ gives,

$$
\begin{aligned}
\left\{ V_h^{\star}(x_{k+1}^{(\tau)}) - [\mathbb{P}V_h^{\star}](x_k^{(\tau)}, a_k^{(\tau)}) \right\}^2 = & \left\{ V_h^{\star}(x_{k+1}^{(\tau)}) - Q_{h-1}^{\star}(x_k^{(\tau)}, a_k^{(\tau)}) + r\left(x_k^{(\tau)}, a_k^{(\tau)}\right) \right\}^2 \\
\leq & \frac{5}{4}\left\{ V_h^{\star}(x_{k+1}^{(\tau)}) - Q_{h-1}^{\star}(x_k^{(\tau)}, a_k^{(\tau)}) \right\}^2 + 5r\left(x_{h'-1}^{(\tau)}, a_{h'-1}^{(\tau)}\right)^2
\end{aligned}
$$

where the second inequality holds by $(a+b)^2 \leq \frac{5}{4}a^2 + 5b^2$ for $a, b \in \mathbb{R}$. Because the reward function is bounded by 1,

$$
\left\{ V_h^{\star}(x_{k+1}^{(\tau)}) - [\mathbb{P}V_h^{\star}](x_k^{(\tau)}, a_k^{(\tau)}) \right\}^2 \leq \frac{5}{4}\left\{ V_h^{\star}(x_{k+1}^{(\tau)}) - Q_{h-1}^{\star}(x_k^{(\tau)}, a_k^{(\tau)}) \right\}^2 + 5.
$$

For $k \in [H]$, let $\mathcal{H}_k^{(\tau)}$denote the sigma algebra generated by $\{x_u^{(\tau)}, a_u^{(\tau)}\}_{u=1,\ldots,k} \cup \{x_u^{(s)}, a_u^{(s)}\}_{s\in[\tau-1], u\in[H]}$. Taking conditional expectations on both sides,

$$
\mathbb{E}\left[\left\{ V_h^{\star}(x_{k+1}^{(\tau)}) - [\mathbb{P}V_h^{\star}](x_k^{(\tau)}, a_k^{(\tau)}) \right\}^2 \middle| \mathcal{H}_k^{(\tau)} \right] \leq \frac{5}{4}\mathbb{E}\left[\left\{ V_h^{\star}(x_{k+1}^{(\tau)}) - Q_{h-1}^{\star}(x_k^{(\tau)}, a_k^{(\tau)}) \right\}^2 \middle| \mathcal{H}_k^{(\tau)} \right] + 5.
$$

In the first term,

$$
\begin{aligned}
& \mathbb{E}\left[\left\{ V_h^{\star}(x_{k+1}^{(\tau)}) - Q_{h-1}^{\star}(x_k^{(\tau)}, a_k^{(\tau)}) \right\}^2 \middle| \mathcal{H}_k^{(\tau)} \right] \\
& = \mathbb{E}\left[V_h^{\star}(x_{k+1}^{(\tau)})^2 \middle| \mathcal{H}_k^{(\tau)} \right] - 2Q_{h-1}^{\star}(x_k^{(\tau)}, a_k^{(\tau)}) [\mathbb{P}V_h^{\star}](x_k^{(\tau)}, a_k^{(\tau)}) + Q_{h-1}^{\star}(x_k^{(\tau)}, a_k^{(\tau)})^2.
\end{aligned}
$$

Note that for any $a_k^{(\tau)} \in \mathcal{A}$, we have

$$
\begin{aligned}
[\mathbb{P}V_h^{\star}](x_k^{(\tau)}, a_k^{(\tau)}) = & Q_{h-1}^{\star}(x_k^{(\tau)}, a_k^{(\tau)}) - r\left(x_k^{(\tau)}, a_k^{(\tau)}\right) \\
\leq & Q_{h-1}^{\star}(x_k^{(\tau)}, a_k^{(\tau)}).
\end{aligned}
$$

Because the function $f(x) = -2xb + x^2$ is non-decreasing for $x \geq b$,

$$
\begin{aligned}
&\mathbb{E}\left[V_h^\star(x_{k+1}^{(\tau)})^2 \middle| \mathcal{H}_k^{(\tau)} \right] - 2Q_{h-1}^\star(x_k^{(\tau)}, a_k^{(\tau)}) \left[\mathbb{P}V_h^\star\right](x_k^{(\tau)}, a_k^{(\tau)}) + Q_{h-1}^\star(x_k^{(\tau)}, a_k^{(\tau)})^2 \\
&\leq \mathbb{E}\left[V_h^\star(x_{k+1}^{(\tau)})^2 \middle| \mathcal{H}_k^{(\tau)} \right] - 2\max_{a\in\mathcal{A}} Q_{h-1}^\star(x_k^{(\tau)}, a) \left[\mathbb{P}V_h^\star\right](x_k^{(\tau)}, a_k^{(\tau)}) + \max_{a\in\mathcal{A}} Q_{h-1}^\star(x_k^{(\tau)}, a)^2 \\
&= \mathbb{E}\left[V_h^\star(x_{k+1}^{(\tau)})^2 \middle| \mathcal{H}_k^{(\tau)} \right] - 2V_{h-1}^\star(x_k^{(\tau)}) \left[\mathbb{P}V_h^\star\right](x_k^{(\tau)}, a_k^{(\tau)}) + V_{h-1}^\star(x_k^{(\tau)})^2 \\
&= \mathbb{E}\left[\left\{ V_h^\star(x_{k+1}^{(\tau)}) - V_{h-1}^\star(x_k^{(\tau)}) \right\}^2 \middle| \mathcal{H}_k^{(\tau)} \right],
\end{aligned}
$$

Summing up over $k \in [H]$,

$$
\sum_{k=1}^{H} \mathbb{E}\left[\left\{ V_h^\star(x_{k+1}^{(\tau)}) - \left[\mathbb{P}V_h^\star\right](x_k^{(\tau)}, a_k^{(\tau)}) \right\}^2 \middle| \mathcal{H}_k^{(\tau)} \right] \leq \frac{5}{4} \sum_{k=1}^{H} \mathbb{E}\left[\left\{ V_h^\star(x_{k+1}^{(\tau)}) - V_{h-1}^\star(x_k^{(\tau)}) \right\}^2 \middle| \mathcal{H}_k^{(\tau)} \right] + 5H.
$$

Note that $\left[\mathbb{P}V_h^\star\right](x_k^{(\tau)}, a_k^{(\tau)}) \leq Q_{h-1}^\star(x_k^{(\tau)}, a_k^{(\tau)}) \leq \max_{a'\in\mathcal{A}} Q_{h-1}^\star(x_k^{(\tau)}, a') = V_{h-1}^\star(x_k^{(\tau)})$ for any $k \in [H]$. Thus, for any $k_1 \neq k_2$, the cross-product terms,

$$
\mathbb{E}\left[V_h^\star(x_{k_1+1}^{(\tau)}) - V_{h-1}^\star(x_{k_1}^{(\tau)}) \middle| \mathcal{H}_{k_1}^{(\tau)} \right] \mathbb{E}\left[V_h^\star(x_{k_2+1}^{(\tau)}) - V_{h-1}^\star(x_{k_2}^{(\tau)}) \middle| \mathcal{H}_{k_2}^{(\tau)} \right] \geq 0,
$$

which implies

$$
\sum_{k=1}^{H} \mathbb{E}\left[\left\{ V_h^\star(x_{k+1}^{(\tau)}) - V_{h-1}^\star(x_k^{(\tau)}) \right\}^2 \middle| \mathcal{H}_k^{(\tau)} \right] \leq \left\{ \sum_{k=1}^{H} \mathbb{E}\left[V_h^\star(x_{k+1}^{(\tau)}) - V_{h-1}^\star(x_k^{(\tau)}) \middle| \mathcal{H}_k^{(\tau)} \right] \right\}^2
$$

Taking conditional expectations on both sides,

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^{H} \mathbb{E}\left[\left\{ V_h^\star(x_{k+1}^{(\tau)}) - V_{h-1}^\star(x_k^{(\tau)}) \right\}^2 \middle| \mathcal{H}_k^{(\tau)} \right] \middle| \mathcal{H}^{(\tau)} \right] \\
&\leq \mathbb{E}\left[\left\{ \sum_{k=1}^{H} \mathbb{E}\left[V_h^\star(x_{k+1}^{(\tau)}) - V_{h-1}^\star(x_k^{(\tau)}) \middle| \mathcal{H}_k^{(\tau)} \right] \right\}^2 \middle| \mathcal{H}^{(\tau)} \right] \\
&= \mathbb{E}\left[\left\{ \sum_{k=1}^{H-1} \mathbb{E}\left[V_h^\star(x_{k+1}^{(\tau)}) - V_{h-1}^\star(x_k^{(\tau)}) \middle| \mathcal{H}_k^{(\tau)} \right] + \mathbb{E}\left[V_h^\star(x_{H+1}^{(\tau)}) - V_{h-1}^\star(x_H^{(\tau)}) \middle| \mathcal{H}_H^{(\tau)} \right] \right\}^2 \middle| \mathcal{H}^{(\tau)} \right] \\
&\leq \mathbb{E}\left[\mathbb{E}\left[\left\{ \sum_{k=1}^{H-1} \mathbb{E}\left[V_h^\star(x_{k+1}^{(\tau)}) - V_{h-1}^\star(x_k^{(\tau)}) \middle| \mathcal{H}_k^{(\tau)} \right] + V_h^\star(x_{H+1}^{(\tau)}) - V_{h-1}^\star(x_H^{(\tau)}) \right\}^2 \middle| \mathcal{H}_H^{(\tau)} \right] \middle| \mathcal{H}^{(\tau)} \right] \\
&= \mathbb{E}\left[\left\{ \sum_{k=1}^{H-1} \mathbb{E}\left[V_h^\star(x_{k+1}^{(\tau)}) - V_{h-1}^\star(x_k^{(\tau)}) \middle| \mathcal{H}_k^{(\tau)} \right] + V_h^\star(x_{H+1}^{(\tau)}) - V_{h-1}^\star(x_H^{(\tau)}) \right\}^2 \middle| \mathcal{H}^{(\tau)} \right],
\end{aligned}
$$

where the second inequality holds by Jensen's inequality. Applying the inequality recursively,

$$
\mathbb{E}\left[\sum_{k=1}^{H} \mathbb{E}\left[\left\{ V_h^\star(x_{k+1}^{(\tau)}) - V_{h-1}^\star(x_k^{(\tau)}) \right\}^2 \middle| \mathcal{H}_k^{(\tau)} \right] \middle| \mathcal{H}^{(\tau)} \right] \leq \mathbb{E}\left[\left\{ \sum_{k=1}^{H} V_h^\star(x_{k+1}^{(\tau)}) - V_{h-1}^\star(x_k^{(\tau)}) \right\}^2 \middle| \mathcal{H}^{(\tau)} \right].
$$

There we obtain

$$
\begin{aligned}
\left\{ \sum_{k=1}^{H} V_h^\star(x_{k+1}^{(\tau)}) - V_{h-1}^\star(x_k^{(\tau)}) \right\}^2 &= \left\{ \sum_{k=1}^{H} V_h^\star(x_{k+1}^{(\tau)}) - V_h^\star(x_k^{(\tau)}) + V_h^\star(x_k^{(\tau)}) - V_{h-1}^\star(x_k^{(\tau)}) \right\}^2 \\
&= \left\{ V_h^\star(x_{H+1}^{(\tau)}) - V_h^\star(x_1^{(\tau)}) + \sum_{k=1}^{H} V_h^\star(x_k^{(\tau)}) - V_{h-1}^\star(x_k^{(\tau)}) \right\}^2.
\end{aligned}
$$

Note that $\left|V_h^\star(x_{H+1}^{(\tau)}) - V_h^\star(x_1^{(\tau)})\right| \le H$. Because $V_h^\star(x) := \sup_\pi V_h^\star(x)$ for all $x \in \mathcal{X}$ and $h \in [H]$,

$$
\begin{aligned}
\left|\sum_{k=1}^{H} V_h^\star(x_k^{(\tau)}) - V_{h-1}^\star(x_k^{(\tau)})\right| \le& \sum_{k=1}^{H} \left|V_h^\star(x_k^{(\tau)}) - V_{h-1}^\star(x_k^{(\tau)})\right| \\
\le& \sum_{k=1}^{H} \sup_{x\in\mathcal{X}} \left|V_h^\star(x) - V_{h-1}^\star(x)\right| \\
=& \sum_{k=1}^{H} \sup_{x\in\mathcal{X}} \left|\sup_\pi V_h^\pi(x) - \sup_\pi V_{h-1}^\pi(x)\right| \\
\le& \sum_{k=1}^{H} \sup_{x\in\mathcal{X}} \sup_\pi \left|V_h^\pi(x) - V_{h-1}^\pi(x)\right| \\
=& \sum_{k=1}^{H} \sup_{x\in\mathcal{X}} \max_{a\in\mathcal{A}} \left|r\left(x,a\right)\right| \\
\le& H.
\end{aligned}
$$

Thus, we obtain

$$
\left\{\sum_{k=1}^{H} V_h^\star(x_{k+1}^{(\tau)}) - V_{h-1}^\star(x_k^{(\tau)})\right\}^2 \le 4H^2.
$$

Gathering the inequalities proves the variance bound. □

C.6 Proof of Lemma 5.1

Proof. By definition of the regret,

$$
\mathbb{E}\left[\mathrm{R}(N,\widehat{A})\right] := \mathbb{E}\left[\sum_{\tau=1}^{N} V_1^\star(x_1^{(\tau)}) - V_1^{\widehat{\pi}^{(\tau)}}(x_1^{(\tau)})\right]
$$

For any $\tau \in [N]$ and $h \in [H]$, define $\mathcal{S}_h^{(\tau)} := \{a_h^{(\tau)} = \arg\max_{a\in\mathcal{A}} r\left(x_h^{(\tau)},a\right) + \phi(x_h^{(\tau)},a)^\top \widehat{w}_{h+1}^{(\tau-1)})\}$. By construction of the algorithm we have $\mathbb{P}(\mathcal{S}_h^{(\tau)}) = (1-\tau^{-1/2})^{1/H}$.For $\tau \in [N]$ and $h \in [H]$, define $\widehat{a}_H^{(\tau)}(x) := \arg\max_{a\in\mathcal{A}} r\left(x,a\right) + \phi(x,a)^\top \widehat{w}_{h+1}^{(\tau-1)})$. Because $\widehat{w}_{H+1}^{(\tau-1)} = \mathbf{0}$, for any $x \in \mathcal{X}$,

$$
\begin{aligned}
V_H^{\widehat{\pi}^{(\tau)}}(x) &= \mathbb{E}^{\widehat{\pi}}\left[r\left(x,a_H\right)\right] \\
&\ge \left(1-\tau^{-1/2}\right)^{1/H} r\left(x,\widehat{a}_H^{(\tau)}\right) \\
&= \left(1-\tau^{-1/2}\right)^{1/H} \Pi_{[0,H]}(r\left(x,\widehat{a}_H^{(\tau)}\right) + \phi(x,\widehat{a}_H^{(\tau)})^\top \widehat{w}_{H+1}^{(\tau-1)}) \\
&= \left(1-\tau^{-1/2}\right)^{1/H} \max_{a\in\mathcal{A}} \Pi_{[0,H]}(r\left(x,a\right) + \phi(x,a)^\top \widehat{w}_{H+1}^{(\tau-1)}) \\
&= \left(1-\tau^{-1/2}\right)^{1/H} \max_{a\in\mathcal{A}} \Pi_{[0,H]}\left(\widehat{Q}_{\widehat{w}_{H+1}^{(\tau-1)}}(x,a)\right),
\end{aligned}
$$

where the first inequality holds because the reward function is nonnegative. For step $H-1$,

$$
\begin{aligned}
V_{H-1}^{\widehat{\pi}^{(\tau)}}(x) =& \mathbb{E}\left[Q_{H-1}^{\widehat{\pi}}(x,a_{H-1})\right] \\
\ge& \left(1-\tau^{-1/2}\right)^{1/H} \Pi_{[0,H]}\left(Q_{H-1}^{\widehat{\pi}}(x,\widehat{a}_{H-1}^{(\tau)}(x))\right) \\
=& \left(1-\tau^{-1/2}\right)^{1/H} \Pi_{[0,H]}\left(r\left(x,\widehat{a}_{H-1}^{(\tau)}(x)\right) + [\mathbb{P}V_H^{\widehat{\pi}}](x,\widehat{a}_{H-1}^{(\tau)}(x))\right) \\
\ge& \left(1-\tau^{-1/2}\right)^{2/H} \Pi_{[0,H]}\left(r\left(x,\widehat{a}_{H-1}^{(\tau)}(x)\right) + \int_{\mathcal{X}} \max_{a\in\mathcal{A}} \Pi_{[0,H]}\left(\widehat{Q}_{\widehat{w}_{H+1}^{(\tau-1)}}(x,a)\right) \phi(x,\widehat{a}_{H-1}^{(\tau)}(x))^\top \psi(x')dx'\right)
\end{aligned}
$$

By definition of $\bar{w}_H^{(\tau-1)}$,

$$\begin{aligned}
V_{H-1}^{\widehat{\pi}^{(\tau)}}(x) &\geq \left(1-\tau^{-1/2}\right)^{2/H} \Pi_{[0,H]}\left(r\left(x,\widehat{a}_{H-1}^{(\tau)}(x)\right)+\phi(x,\widehat{a}_{H-1}^{(\tau)}(x))^\top \bar{w}_H^{(\tau-1)}\right) \\
&\geq \left(1-\tau^{-1/2}\right)^{2/H} \Pi_{[0,H]}\left(r\left(x,\widehat{a}_{H-1}^{(\tau)}(x)\right)+\phi(x,\widehat{a}_{H-1}^{(\tau)}(x))^\top \widehat{w}_H^{(\tau-1)}\right) - \left\|\bar{w}_H^{(\tau-1)}-\widehat{w}_H^{(\tau-1)}\right\|_1 \\
&= \left(1-\tau^{-1/2}\right)^{2/H} \max_{a\in\mathcal{A}} \Pi_{[0,H]}\left(\widehat{Q}_{\widehat{w}_H^{(\tau-1)}}(x,a)\right) - \left\|\bar{w}_H^{(\tau-1)}-\widehat{w}_H^{(\tau-1)}\right\|_1 .
\end{aligned}$$

Recursively, for step 1,

$$V_1^{\widehat{\pi}^{(\tau)}}(x) \geq \left(1-\tau^{-1/2}\right) \max_{a\in\mathcal{A}} \Pi_{[0,H]}\left(\widehat{Q}_{\widehat{w}_2^{(\tau-1)}}(x,a)\right) - \sum_{h=2}^{H}\left\|\widehat{w}_2^{(\tau-1)}-\bar{w}_2^{(\tau-1)}\right\|_1 .$$

For the optimal value function,

$$\begin{aligned}
V_1^\star(x) &= \max_{a\in\mathcal{A}} Q_1^\star(x,a) \\
&= \max_{a\in\mathcal{A}} \left\{r\left(x,a\right)+[\mathbb{P}V_2^\star](x,a)\right\} \\
&= \max_{a\in\mathcal{A}} \Pi_{[0,H]}\left(r\left(x,a\right)+\phi(x,a)^\top w_2^\star\right) \\
&\leq \max_{a\in\mathcal{A}} \Pi_{[0,H]}\left\{r\left(x,a\right)+\phi(x,a)^\top \widehat{w}_2^{(\tau-1)}\right\} + \max_{(x,a)\in\mathcal{X}\times\mathcal{A}}\left|\phi(x,a)^\top\left(w_2^\star-\widehat{w}_2^{(\tau-1)}\right)\right| \\
&= \max_{a\in\mathcal{A}} \Pi_{[0,H]}\left(\widehat{Q}_{\widehat{w}_2^{(\tau-1)}}(x,a)\right) + \max_{(x,a)\in\mathcal{X}\times\mathcal{A}}\left|\phi(x,a)^\top\left(w_2^\star-\widehat{w}_2^{(\tau-1)}\right)\right| .
\end{aligned}$$

By definition of $\bar{w}_2^{(\tau-1)}$,

$$\begin{aligned}
&\max_{(x,a)\in\mathcal{X}\times\mathcal{A}}\left|\phi(x,a)^\top\left(w_2^\star-\widehat{w}_2^{(\tau-1)}\right)\right| \\
&\leq \max_{(x,a)\in\mathcal{X}\times\mathcal{A}}\left|\phi(x,a)^\top\left(w_2^\star-\bar{w}_2^{(\tau-1)}\right)\right| + \left\|\widehat{w}_2^{(\tau-1)}-\bar{w}_2^{(\tau-1)}\right\|_1 \\
&= \max_{(x,a)\in\mathcal{X}\times\mathcal{A}}\left|\phi(x,a)^\top\left\{\int_{\mathcal{X}}\left\{\max_{a\in\mathcal{A}} Q_2^\star(x',a)-\max_{a\in\mathcal{A}}\Pi_{[0,H]}\left(\widehat{Q}_{\widehat{w}_3^{(\tau-1)}}(x',a)\right)\right\}\psi(x')dx'\right\}\right| + \left\|\widehat{w}_2^{(\tau-1)}-\bar{w}_2^{(\tau-1)}\right\|_1
\end{aligned}$$

Because $\int\phi(x,a)^\top\psi(x')dx' = \int\mathbb{P}(x'|x,a)dx' = 1$ for all $(x,a)\in\mathcal{X}\times\mathcal{A}$,

$$\begin{aligned}
&\max_{(x,a)\in\mathcal{X}\times\mathcal{A}}\left|\phi(x,a)^\top\left(w_2^\star-\widehat{w}_2^{(\tau-1)}\right)\right| \\
&\leq \max_{x\in\mathcal{X}}\left|\max_{a\in\mathcal{A}} Q_2^\star(x',a)-\max_{a\in\mathcal{A}}\Pi_{[0,H]}\left(\widehat{Q}_{\widehat{w}_3^{(\tau-1)}}(x',a)\right)\right| + \left\|\widehat{w}_2^{(\tau-1)}-\bar{w}_2^{(\tau-1)}\right\|_1 \\
&= \max_{x\in\mathcal{X}}\left|\max_{a\in\mathcal{A}} \Pi_{[0,H]}\left(Q_2^\star(x,a)\right)-\max_{a\in\mathcal{A}}\Pi_{[0,H]}\left(\widehat{Q}_{\widehat{w}_3^{(\tau-1)}}(x,a)\right)\right| + \left\|\widehat{w}_2^{(\tau-1)}-\bar{w}_2^{(\tau-1)}\right\|_1 \\
&\leq \max_{(x,a)\in\mathcal{X}\times\mathcal{A}}\left|\phi(x,a)^\top\left(w_3^\star-\widehat{w}_3^{(\tau-1)}\right)\right| + \left\|\widehat{w}_2^{(\tau-1)}-\bar{w}_2^{(\tau-1)}\right\|_1 .
\end{aligned}$$

Recursively, we obtain

$$\begin{aligned}
V_1^\star(x) &\leq \max_{a\in\mathcal{A}} \Pi_{[0,H]}\left(\widehat{Q}_{\widehat{w}_2^{(\tau-1)}}(x,a)\right) + \max_{(x,a)\in\mathcal{X}\times\mathcal{A}}\left|\phi(x,a)^\top\left(w_{H+1}^\star-\widehat{w}_{H+1}^{(\tau-1)}\right)\right| + \sum_{h=2}^{H}\left\|\widehat{w}_h^{(\tau-1)}-\bar{w}_h^{(\tau-1)}\right\|_1 \\
&= \max_{a\in\mathcal{A}} \Pi_{[0,H]}\left(\widehat{Q}_{\widehat{w}_2^{(\tau-1)}}(x,a)\right) + \sum_{h=2}^{H}\left\|\widehat{w}_h^{(\tau-1)}-\bar{w}_h^{(\tau-1)}\right\|_1 ,
\end{aligned}$$

where the inequality holds by $w^{\star}_{H+1} = \widehat{w}^{(\tau-1)}_{H+1} = \mathbf{0}$. Therefore

$$\begin{aligned} V_1^{\star}(x_1^{(\tau)}) - V_1^{\widehat{\pi}^{(\tau)}}(x_1^{(\tau)}) \leq & \frac{1}{\sqrt{\tau}} \max_{a \in \mathcal{A}} \Pi_{[0,H]} \left(\widehat{Q}_{\widehat{w}_2^{(\tau-1)}}(x_1^{(\tau)}, a) \right) + 2 \sum_{h=2}^{H} \left\| \widehat{w}_h^{(\tau-1)} - \bar{w}_h^{(\tau-1)} \right\|_1 \\ \leq & \frac{H}{\sqrt{\tau}} + 2 \sum_{h=2}^{H} \left\| \widehat{w}_h^{(\tau-1)} - \bar{w}_h^{(\tau-1)} \right\|_1 . \end{aligned}$$

Summing over $\tau \in [N]$ proves the result. □

C.7 Proof of Theorem 5.5

By Lemma B.4 in Kim et al. (2023a), $N \geq \frac{C s_\star^4 H^2 \log^2(2d)}{\sigma_U^4} \log^5 \frac{2d s_\star^{4/5} H^{2/5}}{e \sigma^{4/5} \sqrt{\delta}}$, implies

$$N \geq C \sigma_U^{-4} s_\star^4 H^2 \log^5(dHN^2/\delta) \log^2(2d).$$

By the regret decomposition (Lemma 5.1), with $N_1 = C \sigma_U^{-4} s_\star^4 H^2 \log^5(dHN^2/\delta) \log^2(2d)$,

$$\mathrm{R}(N, \widehat{A}_{\texttt{RDRLVI}}) \leq 2H(\sqrt{N} + C \sigma_U^{-4} s_\star^4 H^2 \log^5(dHN^2/\delta) \log^2(2d)) + 2 \sum_{n=N_1}^{N-1} \sum_{h=2}^{H} \| \widehat{w}_h^{(n)} - \bar{w}_h^{(n)} \|_1.$$

Applying the tail inequality (Theorem 5.2) for the estimator proves the regret bound.