A. LSVI-PHE with General Function Approximations

A.1. Noise

In the section, we specify how to choose σ in Algorithm 1. Note that we use $\xi_{h,k}^{\tau,m}$ for the noise added in episode k, timestep h, data from episode $\tau < k$ and sampling time m. Similarly, $\xi_{h,k}^{\prime i,m}$ is for episode k, timestep h, regularizer $p_i(\cdot)$ and sampling time m. We set $\lambda = 1$ in our algorithm. By Lemma A.6, there exists $\beta'(\mathcal{F}, \delta)$ such that with probability at least $1-\delta$, for all $(k,h) \in [K] \times [H]$, we have

$$f_h^k(\cdot,\cdot) := r(\cdot,\cdot) + P_h V_{h+1}^k(\cdot,\cdot) \in \mathcal{F}_h^k,$$

where $\mathcal{F}_h^k = \{f \in \mathcal{F} \,|\, \|f - \widehat{f}_h^k\|_{\mathcal{Z}_h^k}^2 + R(f - \widehat{f}_h^k) \le \beta'(\mathcal{F}, \delta)\}$. By Assumption C, for each $\mathcal{F}_h^k$, there exists a $\sigma_{h,k}$ such that

$$g_{\sigma_{h,k}}(s,a) \ge w(\mathcal{F}_h^k, s, a).$$

We define $\sigma = \max_{k\in[K], h\in[H]} \sigma_{h,k}$ to be the maximum standard deviation of the added noise.

A.2. Concentration

We first define few filtrations and good events that we will use in the proof of lemmas in this section.

Definition A.1 (Filtrations)**.** We denote the σ-algbera generated by the set $\mathcal{G}$ using $\sigma(\mathcal{G})$. We define the following filtrations

$$\begin{aligned}
\mathcal{G}^k &\stackrel{\text{def}}{=} \sigma\left(\{(s_t^i, a_t^i, r_t^i)\}_{\{i,t\}\in[k-1]\times[H]} \bigcup \{\xi_{t,l}^{i,j}\}_{i\in[l],\{t,j,l\}\in[H]\times[M]\times[k-1]} \bigcup \{\xi_{t,l}^{\prime i,j}\}_{\{i,t,j,l\}\in[D]\times[H]\times[M]\times[k-1]}\right),\\
\mathcal{G}_{h,1}^k &\stackrel{\text{def}}{=} \sigma\left(\mathcal{G}^k \bigcup \{(s_t^k, a_t^k, r_t^k)\}_{t\in[h]} \bigcup \{\xi_{t,k}^{i,j}\}_{i\in[k],t\ge h,j\in[M]} \bigcup \{\xi_{t,k}^{\prime i,j}\}_{i\in[D],t\ge h,j\in[M]}\right),\\
\mathcal{G}_{h,2}^K &\stackrel{\text{def}}{=} \sigma\left(\mathcal{G}^k \bigcup \{(s_t^k, a_t^k, r_t^k)\}_{t\in[h]}\right).
\end{aligned}$$

Definition A.2 (Good events)**.** For any $\delta > 0$, we define the following random events

$$\begin{aligned}
\mathcal{G}_h^k(\xi,\delta) &\stackrel{\text{def}}{=} \left\{\max_{i\in[k],j\in[M]} \left|\xi_{h,k}^{i,j}\right| \le \sqrt{\gamma_k(\delta)} \bigcap \max_{i\in[D],j\in[M]} |\xi_{h,k}^{\prime i,j}| \le \sqrt{\gamma_k(\delta)}\right\},\\
\mathcal{G}(K,H,\delta) &\stackrel{\text{def}}{=} \bigcap_{k\le K} \bigcap_{h\le H} \mathcal{G}_h^k(\xi,\delta),
\end{aligned}$$

where $\gamma_k(\delta)$ is some constant to be specified in Lemma A.3.

Notation: To simplify our presentation, in the remaining part of this section, we always denote $\sqrt{\gamma_k} := \sqrt{\gamma_k(\delta)}$.

The next lemma shows that the good event defined in Definition A.2 happens with high probability.

Lemma A.3. *For good event $\mathcal{G}(K,H,\delta)$ defined in Definition A.2, if we set $\sqrt{\gamma_k} = \widetilde{O}(\sigma)$, then it happens with probability at least $1-\delta$.*

Proof. Recall that $\xi_{t,l}^{i,j}$ is a zero-mean Gaussian noise with variance $\sigma_{t,l}^2$. By the concentration of Gaussian distribution (Lemma D.1), with probability $1-\delta'$, we have

$$|\xi_{t,l}^{i,j}| \le \sigma_{t,l}\sqrt{2\log(1/\delta')} \le \sigma\sqrt{2\log(1/\delta')}.$$

The same result holds for $\xi_{t,l}^{\prime i,j}$. We complete the proof by setting $\delta' = \delta/(K+D)MHK$ and using union bound. □

In Definition 3.1, for a regularizer $R(f) = \sum_{j=1}^D p_j(f)^2$, where $p_j(\cdot)$ are functionals, we defined the perturbed regularizer as $\widetilde{R}_\sigma(f) = \sum_{j=1}^D [p_j(f) + \xi_j']^2$ with ξ_j' being i.i.d. zero-mean Gaussian noise with variance σ^2. Note that in the algorithm, the variance of the noise for the regularizer is the same as the dataset, which is $\sigma_{h,k}^2$. Recall from Assumption D that for any $V : \mathcal{S} \to [0,H]$, our regularizer R satisfies $R(r+PV) \le B$ for some constant $B \in \mathbb{R}$.

Our next lemma establishes a bound on the perturbed estimate of a single backup.

Lemma A.4. *Consider a fixed $k \in [K]$ and a fixed $h \in [H]$. Let $\mathcal{Z}_h^k = \{(s_h^\tau, a_h^\tau)\}_{\tau\in[k-1]}$ and $\widetilde{\mathcal{D}}_{h,V}^k = \{(s_h^\tau, a_h^\tau, r_h^\tau + \xi_h^\tau + V(s_{h+1}^\tau))\}_{\tau\in[k-1]}$. Define $\widetilde{f}_{h,V}^k = \arg\min_{f\in\mathcal{F}} \|f\|^2_{\widetilde{\mathcal{D}}_{h,V}^k} + \widetilde{R}(f)$. Conditioned on the good event $\mathcal{G}(K,H,\delta)$, with probability at least $1-\delta$, for a fixed $V : \mathcal{S} \to [0,H]$ and any $V' : \mathcal{S} \to [0,H]$ with $\|V' - V\|_\infty \le 1/T$, we have*

$$\begin{aligned}
&\left\|\widetilde{f}_{h,V'}(\cdot,\cdot) - r_h(\cdot,\cdot) - P_hV'(\cdot,\cdot)\right\|^2_{\mathcal{Z}_h^k} + R\left(\widetilde{f}_{h,V'}(\cdot,\cdot) - r_h(\cdot,\cdot) - P_hV'(\cdot,\cdot)\right)\\
\le& c'\left[(H+1+\sqrt{\gamma_k})\sqrt{\log(2/\delta) + \log\mathcal{N}(\mathcal{F},1/T)} + \sqrt{B+\sqrt{\gamma_k BD}}\right]^2,
\end{aligned}$$

for some constant c'. Here B is the bound on the regularizer (Assumption D) and D is the number of regularizers (Definition 3.1). Define this event as $\mathcal{E}_{h,V}(\delta)$.

Proof. Recall that for notational simplicity, we denote $[\mathbb{P}_hV_{h+1}](s,a) = \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot\,|\,s,a)}V_{h+1}(s')$. Now consider a fixed $V : \mathcal{S} \to [0,H]$, and define

$$f_V(\cdot,\cdot) := r_h(\cdot,\cdot) + P_hV(\cdot,\cdot). \tag{6}$$

For any $f \in \mathcal{F}$, we consider $\sum_{\tau\in[k-1]} \chi_h^\tau(f)$ where

$$\chi_h^\tau(f) := 2(f(s_h^\tau,a_h^\tau) - f_V(s_h^\tau,a_h^\tau))(f_V(s_h^\tau,a_h^\tau) - r_h^\tau(s_h^\tau,a_h^\tau) - \xi_h^\tau - V(s_{h+1}^\tau)).$$

Recalling the definition of the filtration $\mathcal{G}_{h,1}^\tau$ from Definition A.1, we note

$$\begin{aligned}
\mathbb{E}[\chi_h^\tau(f)|\mathcal{G}_{h,1}^\tau] &= \mathbb{E}[2(f(s_h^\tau,a_h^\tau) - f_V(s_h^\tau,a_h^\tau))(f_V(s_h^\tau,a_h^\tau) - r_h^\tau(s_h^\tau,a_h^\tau) - \xi_h^\tau - V(s_{h+1}^\tau))|\mathcal{G}_{h,1}^\tau]\\
&= 2(f(s_h^\tau,a_h^\tau) - f_V(s_h^\tau,a_h^\tau))\mathbb{E}[(f_V(s_h^\tau,a_h^\tau) - r_h^\tau(s_h^\tau,a_h^\tau) - \xi_h^\tau - V(s_{h+1}^\tau))|\mathcal{G}_{h,1}^\tau]\\
&= 2(f(s_h^\tau,a_h^\tau) - f_V(s_h^\tau,a_h^\tau))(f_V(s_h^\tau,a_h^\tau) - r_h^\tau(s_h^\tau,a_h^\tau) - P_hV(s_h^\tau,a_h^\tau))\\
&= 0.
\end{aligned}$$

In addition, conditioning on the good event $\mathcal{G}(K,H,\delta)$, we have

$$|\chi_h^\tau(f)| \le 2(H+1+\sqrt{\gamma_\tau})|f(s_h^\tau,a_h^\tau) - f_V(s_h^\tau,a_h^\tau)|.$$

As $\chi_h^\tau(f)$ is a martingale difference sequence conditioned on the filtration $\mathcal{G}_{h,1}^\tau$, by Azuma-Hoeffding inequality, we have

$$\mathbb{P}\left[\left|\sum_{\tau\in[k-1]}\chi_h^\tau(f)\right| \ge \epsilon\right] \le 2\exp\left(-\frac{\epsilon^2}{8(H+1+\sqrt{\gamma_\tau})^2\|f-f_V\|^2_{\mathcal{Z}_h^k}}\right).$$

Now we set

$$\begin{aligned}
\epsilon &= \sqrt{8(H+1+\sqrt{\gamma_\tau})^2\log\left(\frac{2\mathcal{N}(\mathcal{F},1/T)}{\delta}\right)\|f-f_V\|^2_{\mathcal{Z}_h^k}}\\
&\le 4(H+1+\sqrt{\gamma_\tau})\|f-f_V\|_{\mathcal{Z}_h^k}\sqrt{\log(2/\delta)+\log\mathcal{N}(\mathcal{F},1/T)}.
\end{aligned}$$

With union bound, for all $g \in \mathcal{C}(\mathcal{F},1/T)$, with probability at least $1-\delta$ we have

$$\left|\sum_{(\tau)\in[k-1]}\xi_h^\tau(g)\right| \le 4(H+1+\sqrt{\gamma_\tau})\|f-f_V\|_{\mathcal{Z}_h^k}\sqrt{\log(2/\delta)+\log\mathcal{N}(\mathcal{F},1/T)}.$$

Thus, for all $f \in \mathcal{F}$, there exists $g \in \mathcal{C}(\mathcal{F}, 1/T)$ such that $\|f - g\|_\infty \le 1/T$ and

$$\begin{aligned}\left|\sum_{(\tau)\in[k-1]} \chi_h^\tau(f)\right| &\le \left|\sum_{(\tau)\in[k-1]} \chi_h^\tau(g)\right| + 2(H+1+\sqrt{\gamma_\tau}) \\ &\le 4(H+1+\sqrt{\gamma_\tau})\|g - f_V\|_{\mathcal{Z}_h^k}\sqrt{\log(2/\delta) + \log\mathcal{N}(\mathcal{F}, 1/T)} + 2(H+1+\sqrt{\gamma_\tau}) \\ &\le 4(H+1+\sqrt{\gamma_\tau})(\|f - f_V\|_{\mathcal{Z}_h^k} + 1)\sqrt{\log(2/\delta) + \log\mathcal{N}(\mathcal{F}, 1/T)} + 2(H+1+\sqrt{\gamma_\tau}).\end{aligned}$$

For $V' : \mathcal{S} \to [0, H]$ such that $\|V - V'\|_\infty \le 1/T$, we have $\|f_{V'} - f_V\|_\infty \le \|V' - V\|_\infty \le 1/T$.

For any $f \in \mathcal{F}$, we have

$$\begin{aligned}&\|f\|^2_{\widetilde{\mathcal{D}}^k_{h,V'}} - \|f_{V'}\|^2_{\widetilde{\mathcal{D}}^k_{h,V'}} \\ =&\|f - f_{V'}\|^2_{\mathcal{Z}_h^k} + 2\sum_{(s_h^\tau, a_h^\tau)\in\mathcal{Z}_h^k} (f(s_h^\tau, a_h^\tau) - f_{V'}(s_h^\tau, a_h^\tau))(f_{V'}(s_h^\tau, a_h^\tau) - r_h^\tau(s_h^\tau, a_h^\tau) - \xi_h^\tau - V'(s_{h+1}^\tau)) \\ \ge&\|f - f_{V'}\|^2_{\mathcal{Z}_h^k} + 2\sum_{(s_h^\tau, a_h^\tau)\in\mathcal{Z}_h^k} (f(s_h^\tau, a_h^\tau) - f_V(s_h^\tau, a_h^\tau))(f_V(s_h^\tau, a_h^\tau) - r_h^\tau(s_h^\tau, a_h^\tau) - \xi_h^\tau - V(s_{h+1}^\tau)) \\ &- 4(H+1+\sqrt{\gamma_k})\|V' - V\|_\infty|\mathcal{Z}_h^k| \\ \ge&\|f - f_{V'}\|^2_{\mathcal{Z}_h^k} + \sum_{(\tau,h)\in[k-1]\times[H]} \chi_h^\tau(f) - 4(H+1+\sqrt{\gamma_k}) \\ \ge&\|f - f_{V'}\|^2_{\mathcal{Z}_h^k} - 4(H+1+\sqrt{\gamma_k})(\|f - f_V\|_{\mathcal{Z}_h^k} + 1)\sqrt{\log(2/\delta) + \log\mathcal{N}(\mathcal{F}, 1/T)} - 6(H+1+\sqrt{\gamma_k}) \\ \ge&\|f - f_{V'}\|^2_{\mathcal{Z}_h^k} - 4(H+1+\sqrt{\gamma_k})(\|f - f_{V'}\|_{\mathcal{Z}_h^k} + 2)\sqrt{\log(2/\delta) + \log\mathcal{N}(\mathcal{F}, 1/T)} - 6(H+1+\sqrt{\gamma_k}).\end{aligned}$$

In addition, using Assumption D, we have the approximate triangle inequality for the perturbed regularizer:

$$\begin{aligned}&\widetilde{R}(f) - \widetilde{R}(f_{V'}) \\ =&\sum_i^D [p_i(f) + \xi_i']^2 - \sum_i^D [p_i(f_{V'}) + \xi_i']^2 \\ =&R(f) - R(f_{V'}) + 2\sum_i^D \xi_i'(p_i(f) - p_i(f_{V'})) \\ \ge&cR(f - f_{V'}) - 2R(f_{V'}) - 2\sum_i^D \sqrt{\gamma_k} p_i(f_{V'}) \\ \ge&cR(f - f_{V'}) - 2B - 2\sqrt{\gamma_k}\sqrt{BD}.\end{aligned}$$

Summing the above two inequalities we have

$$\|f\|^2_{\widetilde{\mathcal{D}}^k_{h,V'}} + \widetilde{R}(f) - \|f_{V'}\|^2_{\widetilde{\mathcal{D}}^k_{h,V'}} - \widetilde{R}(f_{V'}) \ge \|f - f_{V'}\|^2_{\mathcal{Z}_h^k} + cR(f - f_{V'}) - C,$$

where $C = 4(H+1+\sqrt{\gamma_k})(\|f - f_{V'}\|_{\mathcal{Z}_h^k} + 2)\sqrt{\log(2/\delta) + \log\mathcal{N}(\mathcal{F}, 1/T)} + 6(H+1+\sqrt{\gamma_k}) + 2B + 2\sqrt{\gamma_k}\sqrt{BD}$.

As $\widetilde{f}_{h,V'}$ is the minimizer of $\|f\|^2_{\widetilde{\mathcal{D}}^k_{h,V'}} + \widetilde{R}(f)$, we have

$$\|\widetilde{f}_{h,V'} - f_{V'}\|^2_{\mathcal{Z}_h^k} + cR(\widetilde{f}_{h,V'} - f_{V'}) \le c'\left[(H+1+\sqrt{\gamma_k})\sqrt{\log(2/\delta) + \log\mathcal{N}(\mathcal{F}, 1/T)} + \sqrt{B + \sqrt{\gamma_k BD}}\right]^2.$$

To prove the above argument, we use the inequality that if we have $x^2 + y \le ax + b$ for positive a, b, y, then $x \le a + \sqrt{b}$ and $x^2 + y \le (a + \sqrt{b})^2$. In addition, we can remove c by replacing c' with $c'/\min\{1, c\}$ and then we get our final bound.

□

Lemma A.5 (Confidence Region). *Let* $\mathcal{F}_h^{k,m} = \{f \in \mathcal{F} | \|f - \widetilde{f}_h^{k,m}\|_{\mathcal{Z}_h^k}^2 + R(f - \widetilde{f}_h^{k,m}) \le \beta(\mathcal{F}, \delta)\}$, *where*

$$\beta(\mathcal{F}, \delta) = c' \left[(H + 1 + \sqrt{\gamma_k})\sqrt{\log(2/\delta) + \log \mathcal{N}(\mathcal{F}, 1/T)} + \sqrt{B + \sqrt{\gamma_k BD}}\right]^2. \tag{7}$$

Conditioned on the event $\mathcal{G}(K, H, \delta)$, *with probability at least* $1 - \delta$, *for all* $(k, h, m) \in [K] \times [H] \times [M]$, *we have*

$$r_h(\cdot, \cdot) + P_h V_{h+1}^k(\cdot, \cdot) \in \mathcal{F}_h^{k,m}.$$

Proof. First note that for a fixed $(k, h, m) \in [K] \times [H] \times [M]$,

$$\mathcal{Q} = \{\min\{f(\cdot, \cdot), H\} \mid f \in \mathcal{C}(\mathcal{F}, 1/T)\} \cup \{0\}$$

is a $(1/T)$-cover of $Q_{h+1}^{k,m}(\cdot, \cdot)$. This implies $\mathcal{Q}$ is also a $(1/T)$-cover of $Q_{h+1}^k(\cdot, \cdot)$. This further implies

$$\mathcal{V} = \{\max_{a \in \mathcal{A}} q(\cdot, a) \mid q \in \mathcal{Q}\}$$

is a $1/T$ cover of $V_{h+1}^k(\cdot)$ where we have $\log(|\mathcal{V}|) = \log \mathcal{N}(\mathcal{F}, 1/T)$.

For the remaining part of the proof, we condition on $\bigcap_{V \in \mathcal{V}} \mathcal{E}_{h,V}(\delta/|\mathcal{V}|TM)$, where $\mathcal{E}_{h,V}(\delta)$ is the event defined in Lemma A.4. By Lemma A.4 and union bound, we have $\Pr\left[\bigcap_{V \in \mathcal{V}} \mathcal{E}_{h,V}(\delta/(8|\mathcal{V}|MT)\right] \ge 1 - \delta/(8MT)$.

Let $V \in \mathcal{V}$ such that $\|V - V_{h+1}^k\|_\infty \le 1/T$. By Lemma A.4 we have

$$\begin{aligned}
&\left\|\widetilde{f}_h^{k,m}(\cdot, \cdot) - r_h(\cdot, \cdot) - P_h V_{h+1}^k(\cdot, \cdot)\right\|_{\mathcal{Z}_h^k}^2 + R(\widetilde{f}_h^{k,m}(\cdot, \cdot) - r_h(\cdot, \cdot) - P_h V_{h+1}^k(\cdot, \cdot)) \\
&\qquad \le c' \left[(H + 1 + \sqrt{\gamma_k})\sqrt{\log(1/\delta) + \log \mathcal{N}(\mathcal{F}, 1/T)}\right]^2,
\end{aligned}$$

where c' is some absolute constant. By union bound, for all $(k, h, m) \in [K] \times [H] \times [M]$ we have $r_h(\cdot, \cdot) + P_h V_{h+1}^k(\cdot, \cdot) \in \mathcal{F}_h^{k,m}$ with probability $1 - \delta$. □

The last lemma guarantees that $r_h(\cdot, \cdot) + P_h V_{h+1}^k(\cdot, \cdot)$ lies in the confidence region $\mathcal{F}_h^{k,m}$ with high probability. Note that the confidence region $\mathcal{F}_h^{k,m}$ is centered at $\widetilde{f}_h^{k,m}$, which is the solution to the perturbed regression problem defined in (3). For the unperturbed regression problem and its solution as center of the confidence region, we get the following lemma as a direct consequence of Lemma A.5.

Lemma A.6. *Let* $\mathcal{F}_h^k = \{f \in \mathcal{F} | \|f - \widehat{f}_h^k\|_{\mathcal{Z}_h^k}^2 + R(f - \widehat{f}_h^k) \le \beta'(\mathcal{F}, \delta)\}$, *where*

$$\beta'(\mathcal{F}, \delta) \ge c' \left[(H + 1)\sqrt{\log(2/\delta) + \log \mathcal{N}(\mathcal{F}, 1/T)} + \sqrt{B}\right]^2. \tag{8}$$

With probability at least $1 - \delta$, *for all* $(k, h, m) \in [K] \times [H] \times [M]$, *we have*

$$r_h(\cdot, \cdot) + P_h V_{h+1}^k(\cdot, \cdot) \in \mathcal{F}_h^k.$$

Proof. This is a direct implication of Lemma A.5 with zero perturbance. □

A.3. Optimism

In this section, we will show that $\{Q_h^k\}_{(h,k)\in[H]\times[K]}$ is optimistic with high probability. Formally, we have the following lemma.

Lemma A.7. *Set $M = \ln(\frac{T|\mathcal{S}||\mathcal{A}|}{\delta})/\ln(\frac{1}{1-v})$ in Algorithm 1. Conditioned on the event $\mathcal{G}(K,H,\delta)$, with probability at least $1-\delta$, for all $s\in\mathcal{S}$, $a\in\mathcal{A}$, $h\in[H]$, $k\in[K]$, we have*

$$Q_h^*(s,a) \le Q_h^k(s,a).$$

Proof. For timestep $H+1$, we have $Q_{H+1}^k = Q_{H+1}^* = 0$. By Lemma A.6, there exists $\beta'(\mathcal{F},\delta)$ such that with probability at least $1-\delta$, for all $(k,h)\in[K]\times[H]$, we have

$$f_h^k(\cdot,\cdot) := r_h(\cdot,\cdot) + P_h V_{h+1}^k(\cdot,\cdot) \in \mathcal{F}_h^k,$$

where $\mathcal{F}_h^k = \{f\in\mathcal{F} \mid \|f-\widehat{f}_h^k\|^2_{\mathcal{Z}_h^k} + R(f-\widehat{f}_h^k) \le \beta'(\mathcal{F},\delta)\}$.

Using notations introduced in Definition 4.2, let $g_{h,\sigma}^k$ be a function such that $\widetilde{f}_h^{k,m}(s,a) \ge \widehat{f}(s,a) + g_{h,\sigma}^k(s,a)$ holds with probability at least v. We set $M = \ln(\frac{T|\mathcal{S}||\mathcal{A}|}{\delta})/\ln(\frac{1}{1-v})$ and then $\widetilde{f}_h^{k,m}(s,a) \ge \widehat{f}(s,a) + g_{h,\sigma}^k(s,a)$ with probability at least

$$1-(1-v)^M = 1 - \frac{\delta}{T|\mathcal{S}||\mathcal{A}|},$$

for any $(k,h)\in[K]\times[H]$ and $(s,a)\in\mathcal{S}\times\mathcal{A}$. By union bound, we have $\widetilde{f}_h^{k,m}(s,a) \ge \widehat{f}(s,a) + g_{h,\sigma}^k(s,a)$ for all $(k,h)\in[K]\times[H]$ and $(s,a)\in\mathcal{S}\times\mathcal{A}$ with probability at least $1-\delta$ and we have

$$\begin{aligned}
\widetilde{f}_h^k(s,a) &= \max_{m\in[M]} \widetilde{f}_h^{k,m}(s,a) \\
&\ge \widehat{f}_h^k(s,a) + g_{h,\sigma}^k(s,a) \\
&\ge \widehat{f}_h^k(s,a) + w(\mathcal{F}_h^k) \\
&\ge f_h^k(s,a),
\end{aligned}$$

where the second inequality is from Assumption C and the choise of σ as discussed in Appendix A.1. The last inequality follows from the definition of the width function and the previous observation that $f_h^k(\cdot,\cdot)\in\mathcal{F}_h^k$ with probability at least $1-\delta$. Now we induct on h from $h=H$ to 1.

$$\begin{aligned}
Q_h^*(s,a) &= \min\{r_h(s,a) + P_h V_{h+1}^*(s,a), H\} \\
&= \min\{f_h^k(s,a) + P_h(V_{h+1}^* - V_{h+1}^k)(s,a), H\} \\
&\le \min\{\widetilde{f}_h^k(s,a) + P_h(V_{h+1}^* - V_{h+1}^k)(s,a), H\} \\
&\le \min\{\widetilde{f}_h^k(s,a), H\} \\
&= Q_h^k(s,a).
\end{aligned}$$

Thus,

$$V_h^*(s) = \max_a Q_h^*(s,a) \le \max_a Q_h^k(s,a) = V_h^k(s).$$

where the second inequality is from $V_{h+1}^* \le V_{h+1}^k$, which is implied by induction. □

A.4. Regret Bound

We are now ready to provide the regret bound for Algorithm 1. The next lemma upper bounds the regret of the algorithm by the sum of the width functions.

Lemma A.8 (Regret decomposition)**.** *Denote $b_h^k(s,a) = w(\mathcal{F}_h^k,s,a)$. Conditioned on the event $\mathcal{G}(K,H,\delta)$, with probability at least $1-\delta$, we have*

$$\text{Regret}(K) \leq \sum_{k=1}^{K}\sum_{h=1}^{H} b_h^k(s_h^k, a_h^k) + \sum_{k=1}^{K}\sum_{h=1}^{H} \zeta_h^k,$$

where $\zeta_h^k = P(s_h^k, a_h^k)(V_{h+1}^k - V_{h+1}^{\pi^k}) - (V_{h+1}^k(s_{h+1}^k) - V_{h+1}^{\pi^k}(s_{h+1}^k))$ *is a martingale difference sequence with respect to the filtration* $\mathcal{G}_{h,2}^k$.

Proof. We condition on the good events in Lemma A.5. For all $(k, h, m) \in [K] \times [H] \times [M]$, we have

$$\left\| r_h(\cdot,\cdot) + P_h V_{h+1}^k(\cdot,\cdot) - \widetilde{f}_h^{k,m} \right\|_{\mathcal{Z}_h^k}^2 + R(r_h(\cdot,\cdot) + P_h V_{h+1}^k(\cdot,\cdot) - \widetilde{f}_h^{k,m}) \leq \beta(\mathcal{F}, \delta).$$

Recall that $\mathcal{F}_h^k = \{f \mid \left\| r_h(\cdot,\cdot) + P_h V_{h+1}^k(\cdot,\cdot) - f \right\|_{\mathcal{Z}_h^k}^2 + R(r_h(\cdot,\cdot) + P_h V_{h+1}^k(\cdot,\cdot) - \widetilde{f}_h^{k,m}) \leq \beta(\mathcal{F}, \delta)\}$ is the confidence region. Then for $(k, h, m) \in [K] \times [H] \times [M]$, $\widetilde{f}_h^{k,m} \in \mathcal{F}_h^k$. Defining $b_h^k(s, a) = w(\mathcal{F}_h^k, s, a)$, for all $(k, h, m) \in [K] \times [H] \times [M]$ we have,

$$b_h^k(s,a) \geq \left| r(s,a) + P(s,a)V_{h+1}^k - \widetilde{f}_h^{k,m}(s,a) \right|.$$

As $Q_h^k(s,a) = \min\{\max_{m\in[M]}\{\widetilde{f}_h^{k,m}(\cdot,\cdot)\}, H - h + 1\}$, we have

$$b_h^k(s,a) \geq \left| r(s,a) + P(s,a)V_{h+1}^k - Q_h^k(s,a) \right|.$$

By Lemma A.7 and standard telescoping argument, we have

$$\begin{aligned}
\text{Regret}(K) &\leq \sum_{k=1}^{K} V_1^k(s_1^k) - V_1^{\pi_k}(s_1^k) \\
&= \sum_{k=1}^{K} Q_1^k(s_1^k, a_1^k) - Q_1^{\pi^k}(s_1^k, a_1^k) \\
&= \sum_{k=1}^{K} Q_1^k(s_1^k, a_1^k) - (r(s_1^k, a_1^k) + P(s_1^k, a_1^k)V_2^k) + (r(s_1^k, a_1^k) + P(s_1^k, a_1^k)V_2^k) - Q_1^{\pi^k}(s_1^k, a_1^k) \\
&\leq \sum_{k=1}^{K} b_1^k(s_1^k, a_1^k) + P(s_1^k, a_1^k)(V_2^k - V_2^{\pi^k}) \\
&= \sum_{k=1}^{K} b_1^k(s_1^k, a_1^k) + (V_2^k(s_2^k) - V_2^{\pi^k}(s_2^k)) + \zeta_1^k \\
&\leq \sum_{k=1}^{K}\sum_{h=1}^{H} b_h^k(s_h^k, a_h^k) + \sum_{k=1}^{K}\sum_{h=1}^{H} \zeta_h^k.
\end{aligned}$$

□

Lemma A.9 (Time inhomogeneous version of Lemma 10 in (Wang et al., 2020))**.** *Let* $\mathcal{F}'$ *be a subset of function class* $\mathcal{F}$*, consisting of all* $f \in \mathcal{F}$ *such that*

$$\|f - \upsilon\|_{\mathcal{Z}}^2 + R(f - \upsilon) \leq \beta(\mathcal{F}, \delta),$$

where $\upsilon = r + PV$ *as in Assumption E and* $\beta(\mathcal{F}, \delta)$ *as defined in Lemma A.5. With probability at least* $1 - \delta$*, we have*

$$\sum_{k=1}^{K}\sum_{h=1}^{H} b_h^k(s_h^k, a_h^k) \leq H + 4H^3 \dim_{\mathcal{E}}(\mathcal{F}', 1/T) + H\sqrt{c\dim_{\mathcal{E}}(\mathcal{F}', 1/T)K\beta(\mathcal{F}, \delta)},$$

for some absolute constant $c > 0$.

Proof. Define

$$\mathcal{F}_h'^k = \{f \in \mathcal{F}' \,|\, \|f - \widehat{f}_h^k\|_{\mathcal{Z}_h^k}^2 \le \beta(\mathcal{F}, \delta)\} = \mathcal{F}' \bigcap \{f \in \mathcal{F} \,|\, \|f - \widehat{f}_h^k\|_{\mathcal{Z}_h^k}^2 \le \beta(\mathcal{F}, \delta)\}.$$

As $\mathcal{F}_h^k \subseteq \mathcal{F}'$ and $\mathcal{F}_h^k \subseteq \bigcap\{f \in \mathcal{F} \,|\, \|f - \widehat{f}_h^k\|_{\mathcal{Z}_h^k}^2 \le \beta(\mathcal{F}, \delta)\}$, we have $\mathcal{F}_h^k \subseteq \mathcal{F}_h'^k$ and $w(\mathcal{F}_h^k, s, a) \le w(\mathcal{F}_h'^k, s, a)$ for all s, a. By Assumption E, $\mathcal{F}'$ has bounded eluder dimension.

Similar to Lemma 10 in (Wang et al., 2020), we have for any h,

$$\sum_{k=1}^K b_h^k(s_h^k, a_h^k) \le \sum_{k=1}^K w(\mathcal{F}_h'^k, s, a) \le 1 + 4H^2 \dim_{\mathcal{E}}(\mathcal{F}', 1/T) + \sqrt{c \dim_{\mathcal{E}}(\mathcal{F}', 1/T) K \beta(\mathcal{F}, \delta)}.$$

Summing over all timestep h and we have the bound in the lemma.

□

Theorem A.10. *Under all the assumptions, with probability at least $1 - \delta$, Algorithm 1 achieves a regret bound of*

$$\text{Regret}(K) \le 4H^3 \dim_{\mathcal{E}}(\mathcal{F}, 1/T) + \sqrt{\dim_{\mathcal{E}}(\mathcal{F}, 1/T) \beta(\mathcal{F}, \delta) HT},$$

where

$$\beta(\mathcal{F}, \delta) = c' \left[(H + 1 + \sigma)\sqrt{\log(2/\delta) + \log \mathcal{N}(\mathcal{F}, 1/T)} + \sqrt{B + \sigma\sqrt{BD}}\right]^2,$$

for some constant c'.

Proof. By Assumption E, we can consider $\mathcal{F}' \subseteq \mathcal{F}$ as the whole function class in the analysis because it includes all the $\mathcal{F}_h^k, \forall h, k$. By Azuma-Hoeffding inequality and Lemma A.9, With probability at least $1 - \delta$, we have

$$\begin{aligned}
\text{Regret}(K) &\le \sum_{k=1}^K \sum_{h=1}^H b_h^k(s_h^k, a_h^k) + \sum_{k=1}^K \sum_{h=1}^H \zeta_h^k \\
&\le c' \left(H + 4H^3 \dim_{\mathcal{E}}(\mathcal{F}, 1/T) + H\sqrt{c \dim_{\mathcal{E}}(\mathcal{F}, 1/T) K \beta(\mathcal{F}, \delta)} + H\sqrt{KH \log(1/\delta)}\right),
\end{aligned}$$

for some constant c'. We plug in the definition of $\beta(\mathcal{F}, \delta)$ and $\sqrt{\gamma_k} = \widetilde{O}(\sigma)$, then we get the final bound. □

Remark A.11. *For linear MDP, as shown in Section 4.1.1, we have*

$$\sigma = 2\sqrt{\beta'(\mathcal{F}, \delta)} = c' \left[(H + 1)\sqrt{\log(2/\delta) + \log \mathcal{N}(\mathcal{F}, 1/T)} + \sqrt{B}\right]^2,$$

$B = 2Hd$ and $D = d$. In addition, we have $dim_{\mathcal{E}}(\mathcal{F}, 1/T) = \widetilde{O}(d)$ (Russo & Van Roy, 2013) and $\log \mathcal{N}(\mathcal{F}, 1/T) = \widetilde{O}(d)$. As a result, our bound implies a $\widetilde{O}(\sqrt{H^3 d^3 T})$ regret bound for linear MDP.

B. GFA With Model Misspecification

Assumption F. (Assumption 3 in (Wang et al., 2020)) For function class $\mathcal{F}$, there exists a real number ζ, such that for any $V : \mathcal{S} \to [0, H]$, there exists $g_V \in \mathcal{F}$ which satisfies

$$\max_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left| g_V(s, a) - r(s, a) - \sum_{s' \in \mathcal{S}} P(s'|s, a) V(s') \right| \le \zeta.$$

In addition, we assume g_V satisfies Assumption D, i.e. $R(g_V) \le B$.

Lemma B.1. *Consider a fixed $k \in [K]$ and a fixed $h \in [H]$. Let $\mathcal{Z}_h^k = \{(s_h^\tau, a_h^\tau)\}_{\tau\in[k-1]}$ and $\widetilde{\mathcal{D}}_{h,V}^k = \{(s_h^\tau, a_h^\tau, r_h^\tau + \xi_h^\tau + V(s_{h+1}^\tau))\}_{\tau\in[k-1]}$. Define $\widetilde{f}_{h,V}^k = \arg\min_{f\in\mathcal{F}} \|f\|^2_{\widetilde{\mathcal{D}}_{h,V}^k} + \widetilde{R}(f)$. Conditioned on the good event $\mathcal{G}(K,H,\delta)$, with probability at least $1-\delta$, for a fixed $V : \mathcal{S} \to [0,H]$ and any $V' : \mathcal{S} \to [0,H]$ with $\|V' - V\|_\infty \le 1/T$, we have*

$$\begin{aligned}
&\left\|\widetilde{f}_{h,V'}(\cdot,\cdot) - r_h(\cdot,\cdot) - P_hV'(\cdot,\cdot)\right\|^2_{\mathcal{Z}_h^k} + R(\widetilde{f}_{h,V'}(\cdot,\cdot) - r_h(\cdot,\cdot) - P_hV'(\cdot,\cdot)) \\
\le& c'\left[(H+1+\sqrt{\gamma_k})\sqrt{\log(2/\delta) + \log\mathcal{N}(\mathcal{F},1/T)} + \sqrt{B + \sqrt{\gamma_k BD} + \zeta K(H+\sqrt{\gamma_k})}\right]^2,
\end{aligned}$$

for some constant c'.

Proof. Recall that for notational simplicity, we denote $[\mathbb{P}_hV_{h+1}](s,a) = \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot\,|\,s,a)}V_{h+1}(s')$. Now consider a fixed $V : \mathcal{S} \to [0,H]$, and define

$$f_V(\cdot,\cdot) = r_h(\cdot,\cdot) + P_hV(\cdot,\cdot). \tag{9}$$

By Assumption F, there exists $g_V \in \mathcal{F}$ such that

$$\max_{(s,a)\in\mathcal{S}\times\mathcal{A}} |g_V(s,a) - f_V(s,a)| \le \zeta.$$

For any $f \in \mathcal{F}$, consider

$$\chi_h^\tau = 2(f(s_h^\tau,a_h^\tau) - f_V(s_h^\tau,a_h^\tau))(f_V(s_h^\tau,a_h^\tau) - r_h^\tau(s_h^\tau,a_h^\tau) - \xi_h^\tau - V(s_{h+1}^\tau)).$$

First we show that $\chi_h^\tau(f)$ is a martingale difference sequence with respect to the filtration $\mathcal{G}_{h,1}^\tau$.

$$\begin{aligned}
\mathbb{E}[\chi_h^\tau(f)|\mathcal{G}_{h,1}^\tau] &= \mathbb{E}[2(f(s_h^\tau,a_h^\tau) - f_V(s_h^\tau,a_h^\tau))(f_V(s_h^\tau,a_h^\tau) - r_h^\tau(s_h^\tau,a_h^\tau) - \xi_h^\tau - V(s_{h+1}^\tau))|\mathcal{G}_{h,1}^\tau] \\
&= 2(f(s_h^\tau,a_h^\tau) - f_V(s_h^\tau,a_h^\tau))\mathbb{E}[(f_V(s_h^\tau,a_h^\tau) - r_h^\tau(s_h^\tau,a_h^\tau) - \xi_h^\tau - V(s_{h+1}^\tau))|\mathcal{G}_{h,1}^\tau] \\
&= 2(f(s_h^\tau,a_h^\tau) - f_V(s_h^\tau,a_h^\tau))(f_V(s_h^\tau,a_h^\tau) - r_h^\tau(s_h^\tau,a_h^\tau) - P_hV(s_h^\tau,a_h^\tau)) \\
&= 0.
\end{aligned}$$

In addition, conditioning on good events $\mathcal{G}(K,H,\delta)$, we have

$$|\chi_h^\tau(f)| \le 2(H+1+\sqrt{\gamma_\tau})|f(s_h^\tau,a_h^\tau) - f_V(s_h^\tau,a_h^\tau)|.$$

As $\chi_h^\tau(f)$ is a martingale difference sequence conditioned on the filtration $\mathcal{G}_{h,1}^\tau$, by Azuma-Hoeffding inequality, we have

$$\mathbb{P}\left[\left|\sum_{\tau\in[k-1]}\chi_h^\tau(f)\right| \ge \epsilon\right] \le 2\exp\left(-\frac{\epsilon^2}{8(H+1+\sqrt{\gamma_\tau})^2\|f-f_V\|^2_{\mathcal{Z}_h^k}}\right).$$

Now we set

$$\begin{aligned}
\epsilon &= \sqrt{8(H+1+\sqrt{\gamma_\tau})^2\log\left(\frac{2N(\mathcal{F},1/T)}{\delta}\right)\|f-f_V\|^2_{\mathcal{Z}_h^k}} \\
&\le 4(H+1+\sqrt{\gamma_\tau})\|f-f_V\|_{\mathcal{Z}_h^k}\sqrt{\log(2/\delta)+\log\mathcal{N}(\mathcal{F},1/T)}.
\end{aligned}$$

With union bound, for all $g \in \mathcal{C}(\mathcal{F},1/T)$, with probability at least $1-\delta$ we have

$$\left|\sum_{(\tau)\in[k-1]}\xi_h^\tau(g)\right| \le 4(H+1+\sqrt{\gamma_\tau})\|f-f_V\|_{\mathcal{Z}_h^k}\sqrt{\log(2/\delta)+\log\mathcal{N}(\mathcal{F},1/T)}.$$

Thus, for all $f \in \mathcal{F}$, there exists $g \in \mathcal{C}(\mathcal{F}, 1/T)$ such that $\|f - g\|_\infty \leq 1/T$ and ,

$$\begin{aligned}\left|\sum_{(\tau)\in[k-1]} \chi_h^\tau(f)\right| &\leq \left|\sum_{(\tau)\in[k-1]} \chi_h^\tau(g)\right| + 2(H+1+\sqrt{\gamma_\tau}) \\ &\leq 4(H+1+\sqrt{\gamma_\tau})\|g - f_V\|_{\mathcal{Z}_h^k}\sqrt{\log(2/\delta) + \log\mathcal{N}(\mathcal{F}, 1/T)} + 2(H+1+\sqrt{\gamma_\tau}) \\ &\leq 4(H+1+\sqrt{\gamma_\tau})(\|f - f_V\|_{\mathcal{Z}_h^k} + 1)\sqrt{\log(2/\delta) + \log\mathcal{N}(\mathcal{F}, 1/T)} + 2(H+1+\sqrt{\gamma_\tau})\end{aligned}$$

For $V' : \mathcal{S} \to [0, H]$ such that $\|V - V'\|_\infty \leq 1/T$, we have $\|f_{V'} - f_V\|_\infty \leq \|V' - V\|_\infty \leq 1/T$.
For any $f \in \mathcal{F}$, we have

$$\begin{aligned}&\|f\|^2_{\widetilde{\mathcal{D}}^k_{h,V'}} - \|f_{V'}\|^2_{\widetilde{\mathcal{D}}^k_{h,V'}} \\ =&\|f - f_{V'}\|^2_{\mathcal{Z}_h^k} + 2\sum_{(s_h^\tau, a_h^\tau)\in\mathcal{Z}_h^k} (f(s_h^\tau, a_h^\tau) - f_{V'}(s_h^\tau, a_h^\tau))(f_{V'}(s_h^\tau, a_h^\tau) - r_h^\tau(s_h^\tau, a_h^\tau) - \xi_h^\tau - V'(s_{h+1}^\tau)) \\ \geq&\|f - f_{V'}\|^2_{\mathcal{Z}_h^k} + 2\sum_{(s_h^\tau, a_h^\tau)\in\mathcal{Z}_h^k} (f(s_h^\tau, a_h^\tau) - f_V(s_h^\tau, a_h^\tau))(f_V(s_h^\tau, a_h^\tau) - r_h^\tau(s_h^\tau, a_h^\tau) - \xi_h^\tau - V(s_{h+1}^\tau)) \\ &- 4(H+1+\sqrt{\gamma_k})\|V' - V\|_\infty|\mathcal{Z}_h^k| \\ \geq&\|f - f_{V'}\|^2_{\mathcal{Z}_h^k} + \sum_{(\tau,h)\in[k-1]\times[H]} \chi_h^\tau(f) - 4(H+1+\sqrt{\gamma_k}) \\ \geq&\|f - f_{V'}\|^2_{\mathcal{Z}_h^k} - 4(H+1+\sqrt{\gamma_k})(\|f - f_V\|_{\mathcal{Z}_h^k} + 1)\sqrt{\log(2/\delta) + \log\mathcal{N}(\mathcal{F}, 1/T)} - 6(H+1+\sqrt{\gamma_k}) \\ \geq&\|f - f_{V'}\|^2_{\mathcal{Z}_h^k} - 4(H+1+\sqrt{\gamma_k})(\|f - f_{V'}\|_{\mathcal{Z}_h^k} + 2)\sqrt{\log(2/\delta) + \log\mathcal{N}(\mathcal{F}, 1/T)} - 6(H+1+\sqrt{\gamma_k}).\end{aligned}$$

In addition, by Assumption D, we have

$$\begin{aligned}&\widetilde{R}(f) - \widetilde{R}(f_{V'}) \\ =&\sum_i [p_i(f) - \xi_i']^2 - \sum_i [p_i(f_{V'}) - \xi_i']^2 \\ =&R(f) - R(f_{V'}) - 2\sum_i \xi_i'(p_i(f) - p_i(f_{V'})) \\ \geq&cR(f - f_{V'}) - 2R(f_{V'}) - 2\sum_i \sqrt{\gamma_k} p_i(f_{V'}) \\ \geq&cR(f - f_{V'}) - 2B - 2\sqrt{\gamma_k}\sqrt{BD}.\end{aligned}$$

Summing the above two inequalities we have

$$\|f\|^2_{\widetilde{\mathcal{D}}^k_{h,V'}} + \widetilde{R}(f) - \|f_{V'}\|^2_{\widetilde{\mathcal{D}}^k_{h,V'}} - \widetilde{R}(f_{V'}) \geq \|f - f_{V'}\|^2_{\mathcal{Z}_h^k} + cR(f - f_{V'}) - C,$$

where $C = 4(H+1+\sqrt{\gamma_k})(\|f - f_{V'}\|_{\mathcal{Z}_h^k} + 2)\sqrt{\log(2/\delta) + \log\mathcal{N}(\mathcal{F}, 1/T)} + 6(H+1+\sqrt{\gamma_k}) + 2B + 2\sqrt{\gamma_k}\sqrt{DB}$.

Now we try to replace the $f_{V'}$ in the RHS with g'_V.

$$\begin{aligned}
&\|f_{V'}\|^2_{\widetilde{\mathcal{D}}^k_{h,V'}} - \|g_{V'}\|^2_{\widetilde{\mathcal{D}}^k_{h,V'}} \\
=& \sum_{\tau\in[k-1]} (f_{V'}(s^\tau_h,a^\tau_h) - (r^\tau_h + \xi^\tau_h + V(s^\tau_{h+1})))^2 - \sum_{\tau\in[k-1]} (g_{V'}(s^\tau_h,a^\tau_h) - (r^\tau_h + \xi^\tau_h + V(s^\tau_{h+1})))^2 \\
=& \sum_{\tau\in[k-1]} (f_{V'}(s^\tau_h,a^\tau_h) - g_{V'}(s^\tau_h,a^\tau_h))(f_{V'}(s^\tau_h,a^\tau_h) + g_{V'}(s^\tau_h,a^\tau_h) - 2(r^\tau_h + \xi^\tau_h + V(s^\tau_{h+1}))) \\
\geq& -\zeta K(4H + 2\sqrt{\gamma_k}).
\end{aligned}$$

By the boundedness of the regularizer (Assumption D), we have

$$\|f_{V'}\|^2_{\widetilde{\mathcal{D}}^k_{h,V'}} + \widetilde{R}(f_{V'}) - \|g_{V'}\|^2_{\widetilde{\mathcal{D}}^k_{h,V'}} - \widetilde{R}(g_{V'}) \geq -\zeta K(4H + 2\sqrt{\gamma_k}) - B.$$

Thus we have

$$\begin{aligned}
\|f\|^2_{\widetilde{\mathcal{D}}^k_{h,V'}} + \widetilde{R}(f) - \|g_{V'}\|^2_{\widetilde{\mathcal{D}}^k_{h,V'}} - \widetilde{R}(g_{V'}) &\geq \|f\|^2_{\widetilde{\mathcal{D}}^k_{h,V'}} + \widetilde{R}(f) - \|f_{V'}\|^2_{\widetilde{\mathcal{D}}^k_{h,V'}} - \widetilde{R}(f_{V'}) - \zeta K(4H + 2\sqrt{\gamma_k}) - B \\
&\geq \|f - f_{V'}\|^2_{\mathcal{Z}^k_h} + cR(f - f_{V'}) - C - \zeta K(4H + 2\sqrt{\gamma_k}) - B.
\end{aligned}$$

As $\widetilde{f}_{h,V'}$ is the minimizer of $\|f\|^2_{\widetilde{\mathcal{D}}^k_{h,V'}} + \widetilde{R}(f)$ for $f \in \mathcal{F}$ and note that $g_{V'} \in \mathcal{F}$, we have

$$\begin{aligned}
&\|\widetilde{f}_{h,V'} - f_{V'}\|^2_{\mathcal{Z}^k_h} + cR(\widetilde{f}_{h,V'} - f_{V'}) \\
\leq& c' \left[(H + 1 + \sqrt{\gamma_k})\sqrt{\log(2/\delta) + \log\mathcal{N}(\mathcal{F}, 1/T)} + \sqrt{B + \sqrt{\gamma_k BD} + \zeta K(H + \sqrt{\gamma_k})}\right]^2.
\end{aligned}$$

To prove the above argument, we use the inequality that if we have $x^2 + y \leq ax + b$ for positive a, b, y, then $x \leq a + \sqrt{b}$ and $x^2 + y \leq (a + \sqrt{b})^2$. In addition, we can remove c by replacing c' with $c'/\min\{1, c\}$ and then we get the final bound.

□

Lemma B.2. *(Misspecified Confidence Region) Let* $\mathcal{F}^{k,m}_h = \{f \in \mathcal{F} | \|f - \widetilde{f}^{k,m}_h\|^2_{\mathcal{Z}^k_h} + R(f - \widetilde{f}^{k,m}_h) \leq \beta(\mathcal{F}, \delta)\}$*, where*

$$\beta(\mathcal{F}, \delta) = c' \left[(H + 1 + \sqrt{\gamma_k})\sqrt{\log(2/\delta) + \log\mathcal{N}(\mathcal{F}, 1/T)} + \sqrt{B + \sqrt{\gamma_k BD} + \zeta K(H + \sqrt{\gamma_k})}\right]^2. \tag{10}$$

Conditioned on the event $\mathcal{G}(K, H, \delta)$*, with probability at least* $1 - \delta$*, for all* $(k, h, m) \in [K] \times [H] \times [M]$*, we have*

$$r_h(\cdot,\cdot) + P_h V^k_{h+1}(\cdot,\cdot) \in \mathcal{F}^{k,m}_h.$$

Proof. With Lemma B.1, the proof is same as Lemma A.5. □

Theorem B.3. *Under all the assumptions, with probability at least* $1 - \delta$*, Algorithm 1 achieves a regret bound of*

$$\mathrm{Regret}(K) \leq 4H^3 \dim_{\mathcal{E}}(\mathcal{F}, 1/T) + \sqrt{\dim_{\mathcal{E}}(\mathcal{F}, 1/T)\beta(\mathcal{F}, \delta)HT},$$

where

$$\beta(\mathcal{F}, \delta) = c' \left[(H + 1 + \sigma)\sqrt{\log(2/\delta) + \log\mathcal{N}(\mathcal{F}, 1/T)} + \sqrt{B + \sigma\sqrt{BD} + \zeta K(H + \sigma)}\right]^2,$$

for some constant c'*.*

Proof. With Lemma B.2, the proof is the same as Theorem A.10. □

C. LSVI-PHE with linear function approximation

In this section, we prove Theorem 4.7. Our analysis specilized to linear MDP setting is simpler and may provide additional insights. In addition, compared to GFA setting, we improve the bound for M and it no longer depends on $|\mathcal{S}|$ or $|\mathcal{A}|$. We first introduce the notation and few definitions that are used throughout this section. Upon presenting lemmas and their proofs, finally we combine the lemmas to prove Theorem 4.7.

Definition C.1 (Model prediction error)**.** For all $(k,h) \in [K] \times [H]$, we define the model prediction error associated with the reward r_h^k,

$$l_h^k(s,a) = r_h^k(s,a) + \mathbb{P}_h V_{h+1}^k(s,a) - Q_h^k(s,a).$$

This depicts the prediction error using V_{h+1}^k instead of $V_{h+1}^{\pi^k}$ in the Bellman equations (1).

Definition C.2 (Unperturbed estimated parameter)**.** For all $(k,h) \in [K] \times [H]$, we define the unperturbed estimated parameter as

$$\widehat{\theta}_h^k = (\Lambda_h^k)^{-1} \left(\sum_{\tau=1}^{k-1} [r_h^\tau + V_{h+1}^k(s_{h+1}^\tau)] \phi(s_h^\tau, a_h^\tau) \right).$$

Moreover, we denote the difference between the perturbed estimated parameter $\widetilde{\theta}_h^{k,j}$ and the unperturbed estimated parameter $\widehat{\theta}_h^k$ as

$$\zeta_h^{k,j} = \widetilde{\theta}_h^{k,j} - \widehat{\theta}_h^k.$$

C.1. Concentration

Our first lemma characterizes the difference between the perturbed estimated parameter $\widetilde{\theta}_h^{k,j}$ and the unperturbed estimated parameter $\widehat{\theta}_h^k$.

Proposition C.3 (restatement of Proposition 3.2)**.** *In step 9 of Algorithm 2, conditioned on all the randomness except* $\{\epsilon_h^{k,i,j}\}_{(i,j)\in[k-1]\times[M]}$ *and* $\{\xi_h^{k,j}\}_{j\in[M]}$*, the estimated parameter* $\widetilde{\theta}_h^{k,j}$ *satisfies*

$$\zeta_h^{k,j} = \widetilde{\theta}_h^{k,j} - \widehat{\theta}_h^k \sim N(0, \sigma^2(\Lambda_h^k)^{-1}),$$

where $\widehat{\theta}_h^k = (\Lambda_h^k)^{-1}(\sum_{\tau=1}^{k-1}[r_h^\tau + V_{h+1}^k(s_{h+1}^\tau)]\phi(s_h^\tau, a_h^\tau))$ *is the unperturbed estimated parameter from Definiton C.2.*

Proof. From Algorithm 2, note that

$$\begin{aligned}
\widetilde{\theta}_h^{k,j} &= (\Lambda_h^k)^{-1}(\rho_h^k + \xi_h^{k,j}) \\
&= (\Lambda_h^k)^{-1} \left(\sum_{\tau=1}^{k-1} \left([r_h^\tau + V_{h+1}^k(s_{h+1}^\tau) + \epsilon_h^{k,\tau,j}] \phi(s_h^\tau, a_h^\tau) \right) + \xi_h^{k,j} \right) \\
&= (\Lambda_h^k)^{-1} \left(\sum_{\tau=1}^{k-1} [r_h^\tau + V_{h+1}^k(s_{h+1}^\tau)] \phi(s_h^\tau, a_h^\tau) \right) + (\Lambda_h^k)^{-1} \left(\sum_{\tau=1}^{k-1} \epsilon_h^{k,\tau,j} \phi(s_h^\tau, a_h^\tau) + \xi_h^{k,j} \right) \\
&= \widehat{\theta}_h^k + (\Lambda_h^k)^{-1} \left(\sum_{\tau=1}^{k-1} \epsilon_h^{k,\tau,j} \phi(s_h^\tau, a_h^\tau) + \xi_h^{k,j} \right).
\end{aligned}$$

Since $\epsilon_h^{k,\tau,j} \sim N(0,\sigma^2)$, note that for $\tau \in [k-1]$,

$$\epsilon_h^{k,\tau,j} \phi(s_h^\tau, a_h^\tau) \sim N(0, \sigma^2 \phi(s_h^\tau, a_h^\tau) \phi(s_h^\tau, a_h^\tau)^\top).$$

Now, since $\xi_h^{k,j} \sim N(0, \sigma^2 \lambda I_d)$,

$$\begin{aligned}
(\Lambda_h^k)^{-1} \left(\sum_{\tau=1}^{k-1} \epsilon_h^{k,\tau,j} \phi(s_h^\tau, a_h^\tau) + \xi_h^{k,j} \right) &\sim (\Lambda_h^k)^{-1} \cdot N\left(0, \sigma^2 \left(\sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau) \phi(s_h^\tau, a_h^\tau)^\top + \lambda I_d \right) \right) \\
&\sim (\Lambda_h^k)^{-1} \cdot N\left(0, \sigma^2 \Lambda_h^k\right) \\
&\sim N(0, \sigma^2 (\Lambda_h^k)^{-1}).
\end{aligned}$$

Thus, we have

$$\zeta_h^{k,j} = \widetilde{\theta}_h^{k,j} - \widehat{\theta}_h^k \sim N(0, \sigma^2(\Lambda_h^k)^{-1}).$$

□

Lemma C.4 (Lemma B.1 in (Jin et al., 2020)). *Under Definition 4.3 of linear MDP, for any fixed policy π, let $\{\theta_h^\pi\}_{h\in[H]}$ be the corresponding weights such that $Q_h^\pi(s,a) = \langle \phi(s,a), \theta_h^\pi \rangle$ for all $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$. Then for all $h \in [H]$, we have*

$$\|\theta_h^\pi\| \le 2H\sqrt{d}.$$

Our next lemma states that the unperturbed estimated weight $\widehat{\theta}_h^k$ is bounded.

Lemma C.5. *For any $(k,h) \in [K]\times[H]$, the unperturbed estimated weight $\widehat{\theta}_h^k$ in Definition C.2 satisfies*

$$\|\widehat{\theta}_h^k\| \le 2H\sqrt{kd/\lambda}.$$

Proof. We have

$$\begin{aligned}
\|\widehat{\theta}_h^k\| &= \Big\|(\Lambda_h^k)^{-1}\sum_{\tau=1}^{k-1}[r_h^\tau(s_h^\tau,a_h^\tau) + V_{h+1}^k(s_{h+1}^\tau)]\cdot\phi(s_h^\tau,a_h^\tau)\Big\| \\
&= \Big\|(\Lambda_h^k)^{-1}\sum_{\tau=1}^{k-1}[r_h^\tau(s_h^\tau,a_h^\tau) + \max_{a\in\mathcal{A}} Q_{h+1}^k(s_{h+1}^\tau,a)]\cdot\phi(s_h^\tau,a_h^\tau)\Big\| \\
&\le \frac{1}{\sqrt{\lambda}}\sqrt{k-1}\Big(\sum_{\tau=1}^{k-1}\big\|[r_h^\tau(s_h^\tau,a_h^\tau) + \max_{a\in\mathcal{A}} Q_{h+1}^k(s_{h+1}^\tau,a)]\cdot\phi(s_h^\tau,a_h^\tau)\big\|^2_{(\Lambda_h^k)^{-1}}\Big)^{1/2} \\
&\le \frac{2H}{\sqrt{\lambda}}\sqrt{k-1}\Big(\sum_{\tau=1}^{k-1}\|\phi(s_h^\tau,a_h^\tau)\|^2_{(\Lambda_h^k)^{-1}}\Big)^{1/2} \\
&\le 2H\sqrt{kd/\lambda}.
\end{aligned}$$

Here, the first inequality follows from Lemma D.5. The second inequality follows from the truncation of Q_h^k to the range $[0, H-h+1]$ in Line 11 of Algorithm 2. The last inequality is due to Lemma D.3. □

For the ease of exposition, we now define the values $\beta_k(\delta)$, $\nu_k(\delta)$ and $\gamma_k(\delta)$ which we use to define our high confidence bounds.

Definition C.6 (Noise bounds). For any $\delta > 0$ and some large enough constants c_1,c_2 and c_3, let

$$\begin{aligned}
\sqrt{\beta_k(\delta)} &\stackrel{\text{def}}{=} c_1 H\sqrt{d\log(Hdk/\delta)}, \\
\sqrt{\nu_k(\delta)} &\stackrel{\text{def}}{=} c_2 H\sqrt{d\log(Hdk/\delta)}, \\
\sqrt{\gamma_k(\delta)} &\stackrel{\text{def}}{=} c_3\sqrt{d\nu_k(\delta)\log(d/\delta)}.
\end{aligned}$$

Definition C.7 (Noise distribution). In Algorithm 2, we set the following values for σ

$$\sigma_k = 2\sqrt{\nu_k(\delta)}.$$

Thus for all $j \in [M]$, we have,

$$\{\xi_h^{k,j}\} \sim \mathcal{N}\big(0, 4\nu_k(\delta)(\Lambda_h^k)^{-1}\big).$$

Now, we define some events based on the characterization of the random variable $\zeta_h^{k,j}$ as defined in Definition C.2.

Definition C.8 (Good events). For any $\delta > 0$, we define the following random events

$$\begin{aligned}
\mathcal{G}_h^k(\zeta,\delta) &\stackrel{\text{def}}{=} \Big\{\max_{j\in[M]}\|\zeta_h^{k,j}\|_{\Lambda_h^k} \le \sqrt{\gamma_k(\delta)}\Big\}, \\
\mathcal{G}(K,H,\delta) &\stackrel{\text{def}}{=} \bigcap_{k\le K}\bigcap_{h\le H}\mathcal{G}_h^k(\zeta,\delta).
\end{aligned}$$

Next, we present our main concentration lemma in this section.

Lemma C.9. *Let $\lambda = 1$ in Algorithm 2. For any fixed $\delta > 0$, conditioned on the event $\mathcal{G}(K, H, \delta)$, we have for all $(k, h) \in [K] \times [H]$,*

$$\Big\|\sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau)\big[\big(V_{h+1}^k - \mathbb{P}_h V_{h+1}^k\big)(s_h^\tau, a_h^\tau)\big]\Big\|_{(\Lambda_h^k)^{-1}} \le c_1 H\sqrt{d\log\left(Hdk/\delta\right)}, \tag{11}$$

with probability at least $1-\delta$ for some constant $c_1 > 0$.

Proof. From Lemma C.5, we know, for all $(k, h) \in [K] \times [H]$, we have $\|\widehat{\theta}_h^k\| \le 2H\sqrt{kd/\lambda}$. In addition, by construction of Λ_{h+1}^k, the minimum eigenvalue of Λ_{h+1}^k is lower bounded by λ. Thus we have $\sqrt{\lambda}\|\zeta_{h+1}^{k,j}\| \le \|\zeta_{h+1}^{k,j}\|_{\Lambda_{h+1}^k} \le \sqrt{\gamma_k(\delta)}$. Finally, triangle inequality implies, $\|\widetilde{\theta}_{h+1}^{k,j}\| = \|\widehat{\theta}_{h+1}^k + \zeta_{h+1}^{k,j}\| \le 2H\sqrt{kd/\lambda} + \sqrt{\gamma_k(\delta)/\lambda}$ for all $j \in [M]$. Combining Lemma D.6 and Lemma D.8, we have that, for any $\varepsilon > 0$ and $\delta > 0$, with probability at least $1-\delta$,

$$\begin{aligned}
&\Big\|\sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau)[\big(V_{h+1}^k - \mathbb{P}_h V_{h+1}^k\big)(s_h^\tau, a_h^\tau)]\Big\|_{(\Lambda_h^k)^{-1}} \\
&\le \Big(4H^2\Big[\frac{d}{2}\log\Big(\frac{k+\lambda}{\lambda}\Big) + d\log\Big(1 + \frac{4H\sqrt{kd/\lambda} + 2\sqrt{\gamma_k(\delta)/\lambda}}{\varepsilon}\Big) + \log\frac{1}{\delta}\Big] + \frac{8k^2\varepsilon^2}{\lambda}\Big)^{1/2} \\
&\le \Big(4H^2\Big[\frac{d}{2}\log\Big(\frac{k+\lambda}{\lambda}\Big) + d\log\Big(\frac{3(2H\sqrt{kd/\lambda} + \sqrt{\gamma_k(\delta)/\lambda})}{\varepsilon}\Big) + \log\frac{1}{\delta}\Big] + \frac{8k^2\varepsilon^2}{\lambda}\Big)^{1/2} \\
&\le 2H\Big[\frac{d}{2}\log\Big(\frac{k+\lambda}{\lambda}\Big) + d\log\Big(\frac{3(2H\sqrt{kd/\lambda} + \sqrt{\gamma_k(\delta)/\lambda})}{\varepsilon}\Big) + \log\frac{1}{\delta}\Big]^{1/2} + \frac{2\sqrt{2}k\varepsilon}{\sqrt{\lambda}} \\
&\le 2H\sqrt{d}\Big[\frac{1}{2}\log\Big(\frac{k+\lambda}{\lambda}\Big) + \log\Big(\frac{3(2H\sqrt{kd/\lambda} + \sqrt{\gamma_k(\delta)/\lambda})}{\varepsilon}\Big) + \log\frac{1}{\delta}\Big]^{1/2} + \frac{2\sqrt{2}k\varepsilon}{\sqrt{\lambda}}.
\end{aligned} \tag{12}$$

Setting $\lambda = 1$, $\varepsilon = H\sqrt{d}/k$ and substituting $\sqrt{\gamma_k(\delta)} = c_3\sqrt{d\nu_k(\delta)\log(d/\delta)} \le c_4 Hd\log\left(Hdk/\delta\right)$ for some constant $c_4 > 0$, we get

$$\begin{aligned}
&\Big\|\sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau)[\big(V_{h+1}^k - \mathbb{P}_h V_{h+1}^k\big)(s_h^\tau, a_h^\tau)]\Big\|_{(\Lambda_h^k)^{-1}} \\
&\le 2H\sqrt{d}\left[\frac{1}{2}\log(k+1) + \log(1/\delta) + \log\frac{3k[2H\sqrt{dk} + c_4 Hd\log\left(Hdk/\delta\right)]}{H\sqrt{d}}\right]^{1/2} + 2\sqrt{2}H\sqrt{d} \\
&\le c_1 H\sqrt{d\log\left(Hdk/\delta\right)},
\end{aligned} \tag{13}$$

for some constant $c_1 > 0$.

□

Lemma C.10. *Let $\lambda = 1$ in Algorithm 2. For any $\delta > 0$, conditioned on the event $\mathcal{G}(K, H, \delta)$, for any $(h, k) \in [H] \times [K]$ and $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have*

$$\big|\phi(s,a)^\top \widehat{\theta}_h^k - r_h^k(s,a) - \mathbb{P}_h V_{h+1}^k(s,a)\big| \le c_2 H\sqrt{d\log(Hdk/\delta)}\big\|\phi(s,a)\big\|_{(\Lambda_h^k)^{-1}},$$

with probability $1-\delta$, where $c_2 > 0$ is a constant.

Proof. Let us denote the inner product over $\mathcal{S}$ by $\langle\cdot,\cdot\rangle_{\mathcal{S}}$. Using linear MDP assumption for transition kernel from

Definition 4.3, we get

$$\begin{aligned}
\mathbb{P}_h V_{h+1}^k(s,a) &= \phi(s,a)^\top \langle \mu_h, V_{h+1}^k \rangle_{\mathcal{S}} \\
&= \phi(s,a)^\top (\Lambda_h^k)^{-1} \Lambda_h^k \langle \mu_h, V_{h+1}^k \rangle_{\mathcal{S}} \\
&= \phi(s,a)^\top (\Lambda_h^k)^{-1} \Big(\sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau) \phi(s_h^\tau, a_h^\tau)^\top + \lambda I \Big) \langle \mu_h, V_{h+1}^k \rangle_{\mathcal{S}} \\
&= \phi(s,a)^\top (\Lambda_h^k)^{-1} \Big(\sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau) \phi(s_h^\tau, a_h^\tau)^\top \langle \mu_h, V_{h+1}^k \rangle_{\mathcal{S}} + \lambda I \langle \mu_h, V_{h+1}^k \rangle_{\mathcal{S}} \Big) \\
&= \phi(s,a)^\top (\Lambda_h^k)^{-1} \Big(\sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau) (\mathbb{P}_h V_{h+1}^k)(s_h^\tau, a_h^\tau) + \lambda I \langle \mu_h, V_{h+1}^k \rangle_{\mathcal{S}} \Big), \qquad (14)
\end{aligned}$$

where in the last line we rely on the definition of $\mathbb{P}_h$.

Using (14) we obtain,

$$\begin{aligned}
\phi(s,a)^\top \widehat{\theta}_h^k - r_h^k(s,a) - (\mathbb{P}_h V_{h+1}^k)(s,a) &= \phi(s,a)^\top (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \big[r_h^\tau(s_h^\tau, a_h^\tau) + V_{h+1}^k(s_{h+1}^\tau) \big] \cdot \phi(s_h^\tau, a_h^\tau) - r_h^k(s,a) \\
&\quad - \phi(s,a)^\top (\Lambda_h^k)^{-1} \Big(\sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau) (\mathbb{P}_h V_{h+1}^k)(s_h^\tau, a_h^\tau) + \lambda I \langle \mu_h, V_{h+1}^k \rangle_{\mathcal{S}} \Big) \\
&= \underbrace{\phi(s,a)^\top (\Lambda_h^k)^{-1} \Big(\sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau) \big[\big(V_{h+1}^k - \mathbb{P}_h V_{h+1}^k \big)(s_h^\tau, a_h^\tau) \big] \Big)}_{\text{(i)}} \\
&\quad \underbrace{+ \phi(s,a)^\top (\Lambda_h^k)^{-1} \Big(\sum_{\tau=1}^{k-1} r_h^\tau(s_h^\tau, a_h^\tau) \phi(s_h^\tau, a_h^\tau) \Big) - r_h^k(s,a)}_{\text{(ii)}} \\
&\quad \underbrace{- \lambda \phi(s,a)^\top (\Lambda_h^k)^{-1} \langle \mu_h, V_{h+1}^k \rangle_{\mathcal{S}}}_{\text{(iii)}} . \qquad (15)
\end{aligned}$$

In the following we will analyze the each of the three terms in (15) separately and derive high probability bound for each of them.

Term (i). Since $(\Lambda_h^k)^{-1} \succ 0$, by Cauchy-Schwarz inequality and Lemma C.9, with probability at least $1 - \delta$, we have

$$\begin{aligned}
&\phi(s,a)^\top (\Lambda_h^k)^{-1} \Big(\sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau) \big[\big(V_{h+1}^k - \mathbb{P}_h V_{h+1}^k \big)(s_h^\tau, a_h^\tau) \big] \Big) \\
&\le \Big\| \sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau) \big[\big(V_{h+1}^k - \mathbb{P}_h V_{h+1}^k \big)(s_h^\tau, a_h^\tau) \big] \Big\|_{(\Lambda_h^k)^{-1}} \big\| \phi(s,a) \big\|_{(\Lambda_h^k)^{-1}} \\
&\le \sqrt{\beta_k(\delta)} \big\| \phi(s,a) \big\|_{(\Lambda_h^k)^{-1}} . \qquad (16)
\end{aligned}$$

Term (ii). Note that

$$\begin{aligned}
&\phi(s,a)^\top(\Lambda_h^k)^{-1}\Big(\sum_{\tau=1}^{k-1} r_h^\tau(s_h^\tau,a_h^\tau)\phi(s_h^\tau,a_h^\tau)\Big) - r_h^k(s,a)\\
&= \phi(s,a)^\top(\Lambda_h^k)^{-1}\Big(\sum_{\tau=1}^{k-1} r_h^\tau(s_h^\tau,a_h^\tau)\phi(s_h^\tau,a_h^\tau)\Big) - \phi(s,a)^\top w_h\\
&= \phi(s,a)^\top(\Lambda_h^k)^{-1}\Big(\sum_{\tau=1}^{k-1} r_h^\tau(s_h^\tau,a_h^\tau)\phi(s_h^\tau,a_h^\tau) - \Lambda_h^k w_h\Big)\\
&= \phi(s,a)^\top(\Lambda_h^k)^{-1}\Big(\sum_{\tau=1}^{k-1} r_h^\tau(s_h^\tau,a_h^\tau)\phi(s_h^\tau,a_h^\tau) - \sum_{\tau=1}^{k-1}\phi(s_h^\tau,a_h^\tau)\phi(s_h^\tau,a_h^\tau)^\top w_h - \lambda I w_h\Big)\\
&= \phi(s,a)^\top(\Lambda_h^k)^{-1}\Big(\sum_{\tau=1}^{k-1} r_h^\tau(s_h^\tau,a_h^\tau)\phi(s_h^\tau,a_h^\tau) - \sum_{\tau=1}^{k-1}\phi(s_h^\tau,a_h^\tau) r_h^\tau(s_h^\tau,a_h^\tau) - \lambda I w_h\Big)\\
&= -\lambda\phi(s,a)^\top(\Lambda_h^k)^{-1} w_h, \qquad (17)
\end{aligned}$$

where in the penultimate step, we used the fact $r_h(s,a) = \langle\phi(s,a), w_h\rangle$ from Definition 4.3. Applying Cauchy-Schwarz inequality we obtain,

$$\begin{aligned}
-\lambda\phi(s,a)^\top(\Lambda_h^k)^{-1} w_h &\le \lambda\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}\|w_h\|_{(\Lambda_h^k)^{-1}}\\
&\le \sqrt{\lambda}\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}\|w_h\|_2\\
&\le \sqrt{\lambda d}\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}. \qquad (18)
\end{aligned}$$

Here the second inequality follows by observing that the smallest eigenvalue of Λ_h^k is at least λ and thus the largest eigenvalue of $(\Lambda_h^k)^{-1}$ is at most $1/\lambda$. The last inequality follows from Definition 4.3. Combining (17) and (18) we get

$$\phi(s,a)^\top(\Lambda_h^k)^{-1}\Big(\sum_{\tau=1}^{k-1} r_h^\tau(s_h^\tau,a_h^\tau)\phi(s_h^\tau,a_h^\tau)\Big) - r_h^k(s,a) \le \sqrt{\lambda d}\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}. \qquad (19)$$

Term (iii). Similar to (18), applying Cauchy-Schwarz inequality, we get

$$\begin{aligned}
-\lambda\phi(s,a)^\top(\Lambda_h^k)^{-1}\langle\mu_h, V_{h+1}^k\rangle_{\mathcal{S}} &\le \lambda\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}\|\langle\mu_h, V_{h+1}^k\rangle_{\mathcal{S}}\|_{(\Lambda_h^k)^{-1}}\\
&\le \sqrt{\lambda}\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}\|\langle\mu_h, V_{h+1}^k\rangle_{\mathcal{S}}\|_2\\
&\le \sqrt{\lambda}\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}\Big(\sum_{\tau=1}^{d}\|\mu_h^\tau\|_1^2\Big)^{\frac{1}{2}}\|V_{h+1}^k\|_\infty\\
&\le H\sqrt{\lambda d}\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}. \qquad (20)
\end{aligned}$$

Here the second inequality follows using the same observation we did for **term (ii)**. The last inequality follows from $\sum_{\tau=1}^d \|\mu_h^\tau\|_1^2 \le d$ in Definition 4.3 and the clipping operation performed in Line 2 of Algorithm 2. Now combining (16), (19) and (20), and letting $\lambda = 1$, we get,

$$\begin{aligned}
\big|\phi(s,a)^\top\widehat{\theta}_h^k - r_h^k(s,a) - \mathbb{P}_h V_{h+1}^k(s,a)\big| &\le (\sqrt{\beta_k(\delta)} + H\sqrt{d} + \sqrt{d})\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}} \qquad (21)\\
&= (c_1 H\sqrt{d\log(Hdk/\delta)} + H\sqrt{d} + \sqrt{d})\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}} \qquad (22)\\
&\le c_2 H\sqrt{d\log(Hdk/\delta)}\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}, \qquad (23)
\end{aligned}$$

with probability $1-\delta$ for some constant $c_2 > 0$.

In addition, If we set $\theta_h^k : \phi(\cdot,\cdot)^\top \theta_h^k = r_h^k(\cdot,\cdot) + \mathbb{P}_h V_{h+1}^k(\cdot,\cdot)$ to be the true parameter and $\Delta\theta_h^k = \theta_h^k - \widehat{\theta}_h^k$ to be the regression error, then from the analysis above we can derive that $\|\Delta\theta_h^k\|_{\Lambda_h^k} \le \sqrt{\nu_k(\delta)} = c_2 H\sqrt{d\log(Hdk/\delta)}$. □

Lemma C.11 (stochastic upper confidence bound)**.** *Let $\lambda = 1$ in Algorithm 2. For any $\delta > 0$, conditioned on the event $\mathcal{G}(K,H,\delta)$, for any $(h,k) \in [H]\times[K]$ and $(s,a) \in \mathcal{S}\times\mathcal{A}$, with probability at least $1-(\delta + c_0^M)$, we have*

$$l_h^k(s,a) \le 0,$$

and

$$-l_h^k(s,a) \le \Big(\sqrt{\nu_k(\delta)} + \sqrt{\gamma_k(\delta)}\Big)\big\|\phi(s,a)\big\|_{(\Lambda_h^k)^{-1}},$$

where $c_0 = \Phi(1)$.

Proof. Applying Lemma C.10, for any $(h,k) \in [H]\times[K]$ and $(s,a) \in \mathcal{S}\times\mathcal{A}$, we have,

$$\begin{aligned}\big|r_h^k(s,a) + \mathbb{P}_h V_{h+1}^k(s,a) - \phi(s,a)^\top \widehat{\theta}_h^k\big| &\le c_2 H\sqrt{d\log(Hdk/\delta)} &(24)\\ &= \sqrt{\nu_k(\delta)}\big\|\phi(s,a)\big\|_{(\Lambda_h^k)^{-1}}, &(25)\end{aligned}$$

with probability at least $1-\delta$.

As we are conditioning on the event $\mathcal{G}(K,H,\delta)$, for any $(h,k) \in [H]\times[K]$ and $(s,a) \in \mathcal{S}\times\mathcal{A}$, we have

$$\max_{j\in[M]}\big|\phi(s,a)^\top \zeta_h^{k,j}\big| \le \sqrt{\gamma_k(\delta)}\big\|\phi(s,a)\big\|_{(\Lambda_h^k)^{-1}}. \tag{26}$$

Now from the definition of model prediction error, using (24) and (26), we get, with probability $1-\delta$,

$$\begin{aligned}-l_h^k(s,a) &= Q_h^k(s,a) - r_h^k(s,a) - \mathbb{P}_h V_{h+1}^k(s,a)\\ &= \min\{\max_{j\in[M]} \phi(s,a)^\top(\widehat{\theta}_h^k + \zeta_h^{k,j}), H\} - r_h^k(s,a) - \mathbb{P}_h V_{h+1}^k(s,a)\\ &\le \max_{j\in[M]} \phi(s,a)^\top(\widehat{\theta}_h^k + \zeta_h^{k,j}) - r_h^k(s,a) - \mathbb{P}_h V_{h+1}^k(s,a)\\ &= \max_{j\in[M]} \phi(s,a)^\top \zeta_h^{k,j} - \big(r_h^k(s,a) + \mathbb{P}_h V_{h+1}^k(s,a) - \phi(s,a)^\top\widehat{\theta}_h^k\big)\\ &\le \big|r_h^k(s,a) + \mathbb{P}_h V_{h+1}^k(s,a) - \phi(s,a)^\top\widehat{\theta}_h^k\big| + \max_{j\in[M]}\big|\phi(s,a)^\top\zeta_h^{k,j}\big|\\ &\le \Big(\sqrt{\nu_k(\delta)} + \sqrt{\gamma_k(\delta)}\Big)\big\|\phi(s,a)\big\|_{(\Lambda_h^k)^{-1}}, &(27)\end{aligned}$$

Set $\theta_h^k : \phi(\cdot,\cdot)^\top\theta_h^k = r_h^k(\cdot,\cdot) + \mathbb{P}_h V_{h+1}^k(\cdot,\cdot)$ to be the true parameter and $\Delta\theta_h^k = \theta_h^k - \widehat{\theta}_h^k$ to be the regression error. By the concentration part, conditioning on good events, we have $\|\Delta\theta_h^k\|_{\Lambda_h^k} \le \sqrt{\nu_k(\delta)}$ and $\|\xi_h^{k,j}\|_{\Lambda_h^k} \le \sqrt{\gamma_k(\delta)}$ for all $j \in [M]$.

For all $(h,k) \in [H]\times[K]$ and any $(s,a) \in \mathcal{S}\times\mathcal{A}$, we have

$$\begin{aligned}l_h^k(s,a) &= r_h^k(s,a) + \mathbb{P}_h V_{h+1}^k(s,a) - Q_h^k(s,a)\\ &= r_h^k(s,a) + \mathbb{P}_h V_{h+1}^k(s,a) - \min\big\{H, \max_{j\in[M]}\phi(s,a)^\top(\widehat{\theta}_h^k + \xi_h^{k,j})\big\}^+\\ &\le \max\{\phi(s,a)^\top\Delta\theta_h^k - \max_{j\in[M]}\phi(s,a)^\top\xi_h^{k,j}, 0\}\end{aligned}$$

Now we prove that with high probability, $\max_{j\in[M]}\phi(s,a)^\top\xi_h^{k,j} - \phi(s,a)^\top\Delta\theta_h^k \ge 0$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$. Note that the inequality still holds if we scale $\phi(s,a)$. Now we assume all $\phi(s,a)$ satisfy $\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}} = 1$. Define $\mathcal{C}(\epsilon)$ to be a

ϵ-cover of the ellipsoid $\{\phi | \|\phi\|_{(\Lambda_h^k)^{-1}} = 1\}$ with respect to norm $\|\cdot\|_{(\Lambda_h^k)^{-1}}$ and $\log|\mathcal{C}(\epsilon)| = \widetilde{O}(d\log(\frac{1}{\epsilon}))$. For all $j \in [M]$, we have,

$$\{\xi_h^{k,j}\} \sim N\big(0, 4\nu_k(\delta)(\Lambda_h^k)^{-1}\big).$$

Thus, for all $j \in [M]$ and for all $\phi \in \mathcal{C}(\epsilon)$, we have

$$\big\{\phi^\top \xi_h^{k,j}\big\} \sim N\Big(0, 4\nu_k(\delta)\|\phi\|^2_{(\Lambda_h^k)^{-1}}\Big).$$

Now, for all $j \in [M]$ and for all $\phi \in \mathcal{C}(\epsilon)$, we have

$$\mathbb{P}\Big(\phi^\top \xi_h^{k,j} - 2\sqrt{\nu_k(\delta)}\|\phi\|_{(\Lambda_h^k)^{-1}} \geq 0\Big) = \Phi(-1).$$

Now

$$\begin{aligned}\mathbb{P}\Big(\max_{j\in[M]} \phi^\top \xi_h^{k,j} - 2\sqrt{\nu_k(\delta)}\|\phi\|_{(\Lambda_h^k)^{-1}} \geq 0\Big) &\geq 1 - (1-\Phi(-1))^M \\ &= 1 - \Phi(1)^M \\ &= 1 - c_0^M, \end{aligned} \tag{28}$$

By union bound, with probability $1 - |\mathcal{C}(\epsilon)|c_0^M$, the above bound holds for all elements in $\mathcal{C}$ simultaneously.

Now condition on the previous event, for $\phi = \phi(s,a)$, we can find a $\phi' \in \mathcal{C}(\epsilon)$ such that $\|\phi - \phi'\|_{(\Lambda_h^k)^{-1}} \leq \epsilon$. Define $\Delta\phi = \phi - \phi'$.

$$\begin{aligned}\phi^\top \xi_h^{k,j} - \phi^\top \Delta\theta_h^k &= \phi'^\top \xi_h^{k,j} - \phi'^\top \Delta\theta_h^k + \Delta\phi^\top \xi_h^{k,j} + \Delta\phi^\top \Delta\theta_h^k \\ &\geq \phi'^\top \xi_h^{k,j} - 2\sqrt{\nu_k(\delta)}\|\phi'\|_{(\Lambda_h^k)^{-1}} + \sqrt{\nu_k(\delta)}\|\phi'\|_{(\Lambda_h^k)^{-1}} - \epsilon\|\xi_h^{k,j}\|_{\Lambda_h^k} - \epsilon\|\Delta\theta_h^k\|_{\Lambda_h^k} \\ &\geq \phi'^\top \xi_h^{k,j} - 2\sqrt{\nu_k(\delta)}\|\phi'\|_{(\Lambda_h^k)^{-1}} + \sqrt{\nu_k(\delta)}\|\phi'\|_{(\Lambda_h^k)^{-1}} - \epsilon\sqrt{\gamma_k(\delta)} - \epsilon\sqrt{\nu_k(\delta)}\end{aligned}$$

Set $\epsilon = \frac{\sqrt{\nu_k(\delta)}}{\sqrt{\gamma_k(\delta)}+\sqrt{\nu_k(\delta)}} = \widetilde{O}(\frac{1}{\sqrt{d}})$ and we have, with probability $1 - |\mathcal{C}(\epsilon)|c_0^M$,

$$\begin{aligned}\max_{j\in[M]} \phi^\top \xi_h^{k,j} - \phi^\top \Delta\theta_h^k &\geq \max_{j\in[M]} \phi'^\top \xi_h^{k,j} - 2\sqrt{\nu_k(\delta)}\|\phi'\|_{(\Lambda_h^k)^{-1}} \\ &\geq 0.\end{aligned}$$

Finally we have conditioning on good event $\mathcal{G}(K,H,\delta)$, with probability at least $1 - |\mathcal{C}(\epsilon)|c_0^M$, for all $(s,a) \in \mathcal{S}\times\mathcal{A}$, $l_h^k(s,a) \leq 0$. As $\log|\mathcal{C}(\epsilon)| = \widetilde{O}(d\log(\frac{1}{\epsilon}))$, we can set $M = \widetilde{O}(\frac{d\log(1/\epsilon\delta)}{\log(1/c_0)}) = \widetilde{O}(d)$ to have probability $1-\delta$.

□

C.2. Regret Bound

Definition C.12 (Filtrations)**.** We denote the σ-algbera generated by the set $\mathcal{G}$ using $\sigma(\mathcal{G})$. We define the following filtrations:

$$\begin{aligned}\mathcal{F}^k &\stackrel{\text{def}}{=} \sigma\left(\{(s_t^i, a_t^i, r_t^i)\}_{\{i,t\}\in[k-1]\times[H]} \bigcup \{\xi_t^{i,j}\}_{\{i,t,j\}\in[k-1]\times[H]\times[M]}\right), \\ \mathcal{F}_{h,1}^k &\stackrel{\text{def}}{=} \sigma\left(\mathcal{F}^k \bigcup \{(s_t^k, a_t^k, r_t^k)\}_{t\in[h]} \bigcup \{\xi_t^{k,j} : t \leq h,\ 1 \leq j \leq M\}\right), \\ \mathcal{F}_{h,2}^k &\stackrel{\text{def}}{=} \sigma\left(\mathcal{F}_{h,1}^k \bigcup \{x_{h+1}^k\}\right).\end{aligned}$$

Lemma C.13 (Lemma 4.2 in (Cai et al., 2019)). *It holds that*

$$\begin{aligned}\text{Regret}(T) &= \sum_{k=1}^{K}\left(V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k)\right) \\ &= \underbrace{\sum_{k=1}^{K}\sum_{t=1}^{H}\mathbb{E}_{\pi^*}\left[\langle Q_h^k(s_h,\cdot), \pi_h^*(\cdot \mid s_h) - \pi_h^k(\cdot \mid s_h)\rangle \,\middle|\, s_1 = s_1^k\right]}_{(i)} \\ &\quad + \underbrace{\sum_{k=1}^{K}\sum_{t=1}^{H}\mathcal{D}_h^k}_{(ii)} + \underbrace{\sum_{k=1}^{K}\sum_{t=1}^{H}\mathcal{M}_h^k}_{(iii)} \\ &\quad + \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\left(\mathbb{E}_{\pi^*}\left[l_h^k(s_h,a_h) \mid s_1 = s_1^k\right] - l_h^k(s_h^k,a_h^k)\right)}_{(iv)},\end{aligned} \tag{29}$$

where

$$\mathcal{D}_h^k := \langle (Q_h^k - Q_h^{\pi^k})(s_h^k,\cdot), \pi_h^k(\cdot, s_h^k)\rangle - (Q_h^k - Q_h^{\pi^k})(s_h^k,a_h^k), \tag{30}$$

$$\mathcal{M}_h^k := \mathbb{P}_h((V_{h+1}^k - V_{h+1}^{\pi^k}))(s_h^k,a_h^k) - (V_{h+1}^k - V_{h+1}^{\pi^k})(s_h^k). \tag{31}$$

Lemma C.14. *For the policy π_h^k at time-step k of episode h, it holds that*

$$\sum_{k=1}^{K}\sum_{t=1}^{H}\mathbb{E}_{\pi^*}\left[\langle Q_h^k(s_h,\cdot), \pi_h^*(\cdot \mid s_h) - \pi_h^k(\cdot \mid s_h)\rangle \mid s_1 = s_1^k\right] \le 0, \tag{32}$$

where $T = HK$.

Proof. Obvious from the observation that π_h^k acts greedily with respect to Q_h^k. Note that if $\pi_h^k = \pi_h^*$ then the difference is 0. Else the difference is negative since π_h^k is deterministic with respect to its action-values meaning it takes a value of 1 where π_h^* would take a value of 0 and Q_h^k would have the greatest value at the state-action pair that π_h^k equals one. □

Lemma C.15 (Bound on Martingale Difference Sequence). *For any $\delta > 0$, it holds with probability $1 - 2\delta/3$ that*

$$\sum_{k=1}^{K}\sum_{t=1}^{H}\mathcal{D}_h^k + \sum_{k=1}^{K}\sum_{t=1}^{H}\mathcal{M}_h^k \le 2\sqrt{2H^2T\log(3/\delta)}. \tag{33}$$

Proof. Recall that

$$\begin{aligned}\mathcal{D}_h^k &:= \langle (Q_h^k - Q_h^{\pi^k})(s_h^k,\cdot), \pi_h^k(\cdot, s_h^k)\rangle - (Q_h^k - Q_h^{\pi^k})(s_h^k,a_h^k), \\ \mathcal{M}_h^k &:= \mathbb{P}_h((V_{h+1}^k - V_{h+1}^{\pi^k}))(s_h^k,a_h^k) - (V_{h+1}^k - V_{h+1}^{\pi^k})(s_h^k).\end{aligned}$$

Note that in line 2 of Algorithm 2, we truncate Q_h^k to the range $[0, H-h]$. Thus for any $(k,t) \in [K] \times [H]$, we have, $|\mathcal{D}_h^k| \le 2H$. Moreover, since $\mathbb{E}[\mathcal{D}_h^k | \mathcal{F}_{h,1}^k] = 0$, $\mathcal{D}_h^k$ is a martingale difference sequence. So, applying Azuma-Hoeffding inequality we have with probability at least $1 - \delta/3$,

$$\sum_{k=1}^{K}\sum_{t=1}^{H}\mathcal{D}_h^k \le \sqrt{2H^2T\log(3/\delta)}, \tag{34}$$

where $T = KH$.

Similarly, $\mathcal{M}_h^k$ is a martingale difference sequence since for any $(k,t) \in [K] \times [H]$, $|\mathcal{M}_h^k| \leq 2H$ and $\mathbb{E}[\mathcal{M}_h^k | \mathcal{F}_{h,1}^k] = 0$. Applying Azuma-Hoeffding inequality we have with probability at least $1 - \delta/3$,

$$\sum_{k=1}^{K}\sum_{t=1}^{H} \mathcal{M}_h^k \leq \sqrt{2H^2T\log(3/\delta)}. \tag{35}$$

Applying union bound on (34) and (35) gives (33) and completes the proof.

□

Lemma C.16. *Let $\lambda = 1$ in Algorithm 2. For any $\delta > 0$, conditioned on the event $\mathcal{G}(K,H,\delta)$, we have,*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\big(\mathbb{E}_{\pi^*}\left[l_h^k(s_h,a_h)|s_1 = s_1^k\right] - l_h^k(s_h^k,a_h^k)\big) \leq \big(\sqrt{\nu_K(\delta)} + \sqrt{\gamma_K(\delta)}\big)\sqrt{2dHT\log(1+K)}, \tag{36}$$

with probability $1 - (\delta + c_0^M)$.

Proof. By Lemma C.11, with probability $1 - (\delta + c_0^M)$ it holds that

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\left[l_h^k(s_h,a_h)|s_1 = s_1^k\right] \leq 0, \tag{37}$$

and

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H} -l_h^k(s_h^k,a_h^k) &\leq \sum_{k=1}^{K}\sum_{h=1}^{H}\Big(\sqrt{\nu_k(\delta)} + \sqrt{\gamma_k(\delta)}\Big)\big\|\phi(s_h^k,a_h^k)\big\|_{(\Lambda_h^k)^{-1}} \\
&\leq \Big(\sqrt{\nu_K(\delta)} + \sqrt{\gamma_K(\delta)}\Big)\sum_{k=1}^{K}\sum_{h=1}^{H}\big\|\phi(s_h^k,a_h^k)\big\|_{(\Lambda_h^k)^{-1}} \\
&\leq \Big(\sqrt{\nu_K(\delta)} + \sqrt{\gamma_K(\delta)}\Big)\sum_{h=1}^{H}\sqrt{K}\Big(\sum_{k=1}^{K}\big\|\phi(s_h^k,a_h^k)\big\|^2_{(\Lambda_h^k)^{-1}}\Big)^{1/2} \\
&\leq \Big(\sqrt{\nu_K(\delta)} + \sqrt{\gamma_K(\delta)}\Big)H\sqrt{2dK\log(1+K)} \\
&= \Big(\sqrt{\nu_K(\delta)} + \sqrt{\gamma_K(\delta)}\Big)\sqrt{2dHT\log(1+K)}.
\end{aligned} \tag{38}$$

Here the second inequality follows from the fact that both $\nu_k(\delta)$ and $\gamma_k(\delta)$ are increasinig in k. The third and the fourth inequalities follow from Cauchy-Schwarz inequality and Lemma D.4. Combining (37) and (38) completes the proof. □

Lemma C.17 (Good event probability)**.** *For any $K \in \mathbb{N}$ and any $\delta > 0$, we would have the event $\mathcal{G}(K,H,\delta')$ with probability at least $1 - \delta$, where $\delta' = \delta/MT$.*

Proof. By Lemma D.2, we have, for any fixed t and k, the event $\mathcal{G}_h^k(\xi,\delta')$ occurs with probability at least $1 - M\delta'$. Recall from Definition C.8 that,

$$\mathcal{G}(K,H,\delta') = \bigcap_{k\leq K}\bigcap_{h\leq H}\mathcal{G}_h^k(\xi,\delta').$$

Now taking union bound over all $(t,k) \in [H] \times [K]$, we have

$$\mathbb{P}\Big(\bigcap_{k\leq K}\bigcap_{h\leq H}\mathcal{G}_h^k(\xi,\delta')\Big) \geq 1 - MT\delta' = 1 - \delta,$$

which completes the proof. □

Theorem C.18. *Let $\lambda = 1$, $\sigma = \widetilde{O}(H\sqrt{d})$ and $M = d\log(\delta/9)/\log c_0$, where $c_0 = \Phi(1)$ and $\delta \in (0, 1]$. Under Definition 4.3, the regret of Algorithm 2 satisfies*

$$\textit{Regret}(T) \leq \widetilde{\mathcal{O}}(d^{3/2}H^{3/2}\sqrt{T}),$$

with probability at least $1 - \delta$.

Proof of Theorem C.18. Let $\delta' = \delta/9$. From Lemma C.17, the event $\mathcal{G}(K, H, \delta')$ happens with probability $1 - \delta'$. Combining Lemma C.16 and Lemma C.17 we have that the event $\mathcal{G}(K, H, \delta')$ occurs and it holds that

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\left(\mathbb{E}_{\pi^*}\left[l_h^k(s_h, a_h)|s_1 = s_1^k\right] - l_h^k(s_h^k, a_h^k)\right) \leq \left(\sqrt{\nu_K(\delta')} + \sqrt{\gamma_K(\delta')}\right)\sqrt{2dHT\log(1+K)}, \tag{39}$$

with probability at least $(1-\delta')(1-(\delta'+c_0^M))$. Note that $c_0^M = \delta'$ and $(1-\delta')(1-(\delta'+c_0^M)) > 1 - 3\delta' = 1 - \delta/3$. The martingale inequalities from Lemma C.15 happens with probability $1 - 2\delta/3$.

Applying union bound on (32), (33) and (39) gives the final regret bound of $\widetilde{\mathcal{O}}(d^{3/2}H^{3/2}\sqrt{T})$ completes the proof. □

D. Auxiliary lemmas

This section presents several auxiliary lemmas and their proofs.

D.1. Gaussian Concentration

Lemma D.1 (Gaussian Concentration (Vershynin, 2018))**.** *Consider a d-dimensional multivariate normal distribution $\eta \sim N(0, A\Lambda^{-1})$ where A is a scalar. For any $\delta > 0$, with probability $1 - \delta$,*

$$\|\eta\|_\Lambda \leq c\sqrt{dA\log(d/\delta)},$$

where c is some absolute constant. For $d = 1$, we have $c = \sqrt{2}$.

Lemma D.2. *Consider a d-dimensional multivariate normal distribution $N(0, A\Lambda^{-1})$ where A is a scalar. Let $\eta_1, \eta_2, \ldots, \eta_M$ be M independent samples from the distribution. Then for any $\delta > 0$*

$$\mathbb{P}\left(\max_{j\in[M]}\|\eta_j\|_\Lambda \leq c\sqrt{dA\log(d/\delta)}\right) \geq 1 - M\delta,$$

where c is some absolute constant.

Proof. From Lemma D.1, for a fixed $j \in [M]$, with probability at least $1 - \delta$ we would have

$$\|\eta\|_\Lambda \leq c\sqrt{dA\log(d/\delta)}.$$

Applying union bound over all M samples completes the proof. □

D.2. Inequalities for summations

Lemma D.3 (Lemma D.1 in (Jin et al., 2020))**.** *Let $\Lambda_h = \lambda I + \sum_{i=1}^t \phi_i\phi_i^\top$, where $\phi_i \in \mathbb{R}^d$ and $\lambda > 0$. Then it holds that*

$$\sum_{i=1}^{t}\phi_i^\top(\Lambda_h)^{-1}\phi_i \leq d.$$

Lemma D.4 (Lemma 11 in (Abbasi-Yadkori et al., 2011))**.** *Using the same notation as defined in this paper*

$$\sum_{k=1}^{K}\left\|\phi(s_h^k, a_h^k)\right\|^2_{(\Lambda_h^k)^{-1}} \leq 2d\log\left(\frac{\lambda + K}{\lambda}\right).$$

Lemma D.5. *Let $A \in \mathbb{R}^{d\times d}$ be a positive definite matrix where its largest eigenvalue $\lambda_{max}(A) \le \lambda$. Let $x_1, \dots, x_k$ be k vectors in $\mathbb{R}^d$. Then it holds that*

$$\Big\|A\sum_{i=1}^{k} x_i\Big\| \le \sqrt{\lambda k}\Big(\sum_{i=1}^{k} \|x_i\|_A^2\Big)^{1/2}.$$

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

$$\begin{aligned}\|Av\| &= \|A^{1/2}A^{1/2}v\| \\ &\le \|A^{1/2}\|\|A^{1/2}v\| \\ &= \|A^{1/2}\|\|v\|_A.\end{aligned}$$

Here the inequality follows from the definition of the operator norm $\|A^{1/2}\|$. Moreover, $\|A^{1/2}\| \le \sqrt{\lambda}$ since $\lambda_{max}(A) \le \lambda$. Thus,

$$\Big\|A\sum_{i=1}^{k} x_i\Big\| \le \sqrt{\lambda}\Big\|\sum_{i=1}^{k} x_i\Big\|_A. \tag{40}$$

Now by Cauchy-Schwarz inequality,

$$\begin{aligned}\Big\|\sum_{i=1}^{k} x_i\Big\|_A^2 &= \sum_{i=1}^{k}\sum_{j=1}^{k} x_i^\top A x_j \\ &\le \sum_{i=1}^{k}\sum_{j=1}^{k} \|x_i\|_A\|x_j\|_A \\ &= \left(\sum_{i=1}^{k}\|x_i\|_A\right)^2 \\ &\le k\sum_{i=1}^{k}\|x_i\|_A^2.\end{aligned} \tag{41}$$

Combining (40) and (41), proves the lemma. □

D.3. Covering numbers and self-normalized processes

Lemma D.6 (Lemma D.4 in (Jin et al., 2020))**.** *Let $\{s_i\}_{i=1}^{\infty}$ be a stochastic process on state space $\mathcal{S}$ with corresponding filtration $\{\mathcal{F}_i\}_{i=1}^{\infty}$. Let $\{\phi_i\}_{i=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process where $\phi_i \in \mathcal{F}_{i-1}$, and $\|\phi_i\| \le 1$. Let $\Lambda_k = \lambda I + \sum_{i=1}^{k}\phi_i\phi_i^\top$. Then for any $\delta > 0$, with probability at least $1-\delta$, for all $k \ge 0$, and any $V \in \mathcal{V}$ with $\sup_{s\in\mathcal{S}}|V(s)| \le H$, we have*

$$\Big\|\sum_{i=1}^{k}\phi_i\big\{V(s_i) - \mathbb{E}[V(s_i)\,|\,\mathcal{F}_{i-1}]\big\}\Big\|_{\Lambda_k^{-1}}^2 \le 4H^2\Big[\frac{d}{2}\log\Big(\frac{k+\lambda}{\lambda}\Big) + \log\frac{N_\varepsilon}{\delta}\Big] + \frac{8k^2\epsilon^2}{\lambda},$$

where N_ε is the ε-covering number of $\mathcal{V}$ with respect to the distance dist$(V, V') = \sup_{s\in\mathcal{S}}|V(s) - V'(s)|$.

Lemma D.7 (Covering number of Euclidean ball, (Vershynin, 2018))**.** *For any $\varepsilon > 0$, the ε-covering number, N_ε, of the Euclidean ball of radius $B > 0$ in $\mathbb{R}^d$ satisfies*

$$N_\varepsilon \le \Big(1 + \frac{2B}{\varepsilon}\Big)^d \le \Big(\frac{3B}{\varepsilon}\Big)^d.$$

Lemma D.8. *Consider a class of functions $\mathcal{V} : \mathcal{S} \to \mathbb{R}$ which has the following parametric form*

$$V(\cdot) = \Big\langle \min\big\{\phi(\cdot,\cdot)^\top\theta, H\big\}^+, \pi(\cdot\,|\,\cdot)\Big\rangle_{\mathcal{A}},$$

where the parameter θ satisfies $\|\theta\| \le B$ and for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have $\|\phi(s,a)\| \le 1$. If $N_{\mathcal{V},\varepsilon}$ denotes the ε-covering number of $\mathcal{V}$ with respect to the distance dist$(V, V') = \sup_{s\in\mathcal{S}} |V(s) - V'(s)|$, then

$$\log N_{\mathcal{V},\varepsilon} \le d \log(1 + 2B/\varepsilon) \le d \log(3B/\varepsilon).$$

Proof. Consider any two functions $V_1, V_2 \in \mathcal{V}$ with parameters θ_1 and θ_2, respectively. Note that $\min\{\cdot, H\}$ is a contraction mapping. Thus we have

$$\begin{aligned}
\text{dist}(V_1, V_2) &\le \sup_s \big| \langle \phi(s,\cdot)^\top \theta_1 - \phi(s,\cdot)^\top \theta_2, \pi(\cdot \mid s) \rangle_{\mathcal{A}} \big| \\
&\le \sup_{\phi : \|\phi\| \le 1} \big| \phi^\top \theta_1 - \phi^\top \theta_2 \big| \\
&= \sup_{\phi : \|\phi\| \le 1} \big| \phi^\top (\theta_1 - \theta_2) \big| \\
&\le \sup_{\phi : \|\phi\| \le 1} \|\theta_1 - \theta_2\|_2 \|\phi\|_2 \\
&= \|\theta_1 - \theta_2\|, \qquad (42)
\end{aligned}$$

where the second inequality follows from the triangle inequality and the third inequality follows from the Cauchy-Schwarz inequality.

If $N_{\theta,\varepsilon}$ denotes the ε-covering number of $\{\theta \in \mathbb{R}^d \mid \|\theta\| \le B\}$, Lemma D.7 implies

$$N_{\theta,\varepsilon} \le \Big(1 + \frac{2B}{\varepsilon}\Big)^d \le \Big(\frac{3B}{\varepsilon}\Big)^d.$$

Let $\mathcal{C}_{\theta,\varepsilon}$ be an ε-cover of $\{\theta \in \mathbb{R}^d \mid \|\theta\| \le B\}$ with cardinality $N_{\theta,\varepsilon}$. Consider any $V_1 \in \mathcal{V}$. By (42), there exists $\theta_2 \in \mathcal{C}_{\theta,\varepsilon}$ such that V_2 parameterized by θ_2 satisfies dist$(V_1, V_2) \le \varepsilon$. Thus we have

$$\log N_{\mathcal{V},\varepsilon} \le \log N_{\theta,\varepsilon} \le d \log(1 + 2B/\varepsilon) \le d \log(3B/\varepsilon),$$

which concludes the proof. □

E. Experiment Details

In this section we include the figure for the RiverSwim environment from (Osband et al., 2013).

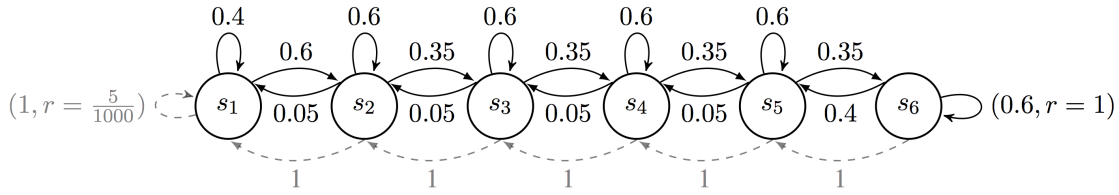

Figure 5: The 6 state RiverSwim environment (Osband et al., 2013). State s_1 has a small reward while state s_6 has a large reward. The action whose transition is denoted with a dashed arrow deterministically moves the agent left. The other action is stochastic, and with relative high probability moves the agent towards the goal state s_6. This action represents swimming against the current, hence the name RiverSwim.