Finite-sample Guarantees for Nash Q-learning with Linear Function Approximation (Supplementary Material)

We restate Theorem 3.1 below with a change of variables that moves the dependency on the number of agents from the upper bound of the regret to the probability expression. This is the version of the Theorem we will prove when referring to Theorem 3.1 from the main paper.

Theorem (Performance of the NQOVI algorithm)**.** *There exists an absolute constant $c_\beta > 0$ such that, for any fixed $\delta \in (0,1)$, if we set $\lambda = 1$ and $\beta = c_\beta dH\sqrt{\iota}$, with $\iota := \log(dKH/\delta)$, then, with probability at least $1-(n+2)\delta$,*

$$\text{Regret}(K) \leq \mathcal{O}\left(\sqrt{K}\sqrt{d^3H^5\iota^2}\right). \tag{.1}$$

Let $\mathbb{Z}_{\geq 0}$ ($\mathbb{Z}_{\geq 1}$) be the set of non-negative (positive) integers.

All results that are direct adaptations or restatements from existing results in the single RL literature from (Jin et al., 2020) will have their detailed proofs if necessary to understand the nuances of their adaptation to our setting. Such proofs will be deferred to the last Section C of the supplementary material.

A AUXILIARY RESULTS

The following proposition is an immediate adaptation of an existing one in (Jin et al., 2020) for MDPs.

Proposition A.1 (Bounded parameters for Q-functions – Proposition 2.3 and Lemma B.1 in (Jin et al., 2020))**.** *Consider a linear stochastic game $\mathcal{MG}$. Given a policy profile π, we have that for any $i \in [n]$, there exist paremeters $w_h^{i,\pi} \in \mathbb{R}^d$, $h \in [H]$, such that $Q_h^{i,\pi}(x,a) = \langle \phi(x,a), w_h^{i,\pi}\rangle$ for any $(x,a) \in \mathcal{S} \times \mathcal{A}$ and $\left\|w_h^{i,\pi}\right\| \leq 2H\sqrt{d}$.*

The following lemma is a restatement of another one in (Jin et al., 2020), though with some different notation.

Lemma A.1 (Concentration bound for self-normalized processes – Lemma D.4 in (Jin et al., 2020))**.** *Let $\{\mathcal{F}_\tau\}_{\tau=0}^\infty$ be a filtration. Let $\{x_\tau\}_{\tau=1}^\infty$ be a stochastic process on $\mathcal{S}$ such that $x_\tau \in \mathcal{F}_\tau$, and let $\{\phi_\tau\}_{\tau=1}^\infty$ be an $\mathbb{R}^d$-valued stochastic process such that $\phi_\tau \in \mathcal{F}_{\tau-1}$ and $\|\phi_\tau\| \leq 1$. Let $\mathcal{G}$ be a function class of real-valued functions such that $\sup_{x\in\mathcal{S}}|g(x)| \leq H$ for any $g \in \mathcal{G}$, and with ϵ-covering number $\mathcal{N}_\epsilon$ with respect to the distance $\text{dist}(g,g') = \sup_{x\in S}|g(x)-g'(x)|$. Let $\Lambda_A = \lambda I_d + \sum_{\tau=1}^A \phi_\tau\phi_\tau^\top$. Then for every $A \in \mathbb{Z}_{\geq 1}$, every $g \in \mathcal{G}$, and any $\delta \in (0,1]$, we have that with probability at least $1-\delta$,*

$$\left\|\sum_{\tau=1}^A \phi_\tau\{g(x_\tau) - \mathbb{E}[g(x_\tau)|\mathcal{F}_{\tau-1}]\}\right\|^2_{\Lambda_A^{-1}} \leq 4H^2\left[\frac{d}{2}\log\left(\frac{\lambda + A/d}{\lambda}\right) + \log\frac{\mathcal{N}_\epsilon}{\delta}\right] + \frac{8A^2\epsilon^2}{\lambda}.$$

The following lemma is a key result in the proof of Theorem 3.1, as seen in Section 4 from the main paper.

Lemma A.2 (Bounding the covering number)**.** *Let $i \in [n]$, and let $\bar{w}_i \in \mathbb{R}^d$ be such that $\|\bar{w}_i\| \leq L$, $\bar{\Lambda} \in \mathbb{R}^{d \times d}$ be such that its minimum eigenvalue is greater or equal than λ, and, for all $(x, a) \in \mathcal{S} \times \mathcal{A}$, let $\phi(x, a) \in \mathbb{R}^d$ be such that $\|\phi(x, a)\| \leq 1$, and let $\beta > 0$. Define the function class*

$$\mathcal{V}_i = \left\{ V : \mathcal{S} \to \mathbb{R} \;\middle|\; V(\cdot) = \max_{\nu \in \Delta(\mathcal{A}_i)} \mathop{\mathbb{E}}_{\substack{a_i \sim \nu \\ a_{-i} \sim \pi_{-i}(\cdot)}} \left[\min \left\{ \bar{w}_i^\top \phi(\cdot, a) + \beta \sqrt{\phi(\cdot, a)^\top \bar{\Lambda}^{-1} \phi(\cdot, a)}, H \right\} \right] \right\}, \tag{A.1}$$

where $\pi_{-i}(\cdot) \in \Delta(\mathcal{A}_{-i})$. Let $\mathcal{N}_{\epsilon_i}$ be the ϵ_i-covering number of $\mathcal{V}_i$ with respect to the distance $\operatorname{dist}(V, V') = \sup_{x \in \mathcal{S}} |V(x) - V'(x)|$. *Then,*

$$\log \mathcal{N}_{\epsilon_i} \leq d \log(1 + 4L/\epsilon) + d^2 \log\big[1 + 8d^{1/2}\beta^2/(\lambda\epsilon^2)\big].$$

Proof. Let $V, V' \in \mathcal{V}_i$. Let $\bar{u}(x, a) := \sqrt{\phi(x, a)^\top \bar{\Lambda}^{-1} \phi(x, a)}$ and $\bar{u}'(x, a) := \sqrt{\phi(x, a)^\top (\bar{\Lambda}')^{-1} \phi(x, a)}$, and let $\bar{g}(x, a) = \min\left\{\bar{w}_i^\top \phi(x, a) + \beta \bar{u}(x, a), H\right\}$ and $\bar{g}'(x, a) = \min\left\{\bar{w}_i^\top \phi(x, a) + \beta \bar{u}'(x, a), H\right\}$. Then,

$$\begin{aligned}
\operatorname{dist}(V, V') &= \sup_{x \in \mathcal{S}} \left| \max_{\nu \in \Delta(\mathcal{A}_i)} \mathop{\mathbb{E}}_{\substack{a_i \sim \nu \\ a_{-i} \sim \pi_{-i}(x)}} [\bar{g}(x, a)] - \max_{\nu \in \Delta(\mathcal{A}_i)} \mathop{\mathbb{E}}_{\substack{a_i \sim \nu \\ a_{-i} \sim \pi_{-i}(x)}} [\bar{g}'(x, a)] \right| \\
&\overset{(a)}{\leq} \sup_{\substack{x \in \mathcal{S} \\ \nu \in \Delta(\mathcal{A}_i)}} \left| \mathop{\mathbb{E}}_{\substack{a_i \sim \nu \\ a_{-i} \sim \pi_{-i}(x)}} [\bar{g}(x, a)] - \mathop{\mathbb{E}}_{\substack{a_i \sim \nu \\ a_{-i} \sim \pi_{-i}(x)}} [\bar{g}'(x, a)] \right| \\
&= \sup_{\substack{x \in \mathcal{S} \\ \nu \in \Delta(\mathcal{A}_i)}} \left| \mathop{\mathbb{E}}_{\substack{a_i \sim \nu \\ a_{-i} \sim \pi_{-i}(x)}} [\bar{g}(x, a) - \bar{g}'(x, a)] \right| \\
&\leq \sup_{\substack{x \in \mathcal{S} \\ \nu \in \Delta(\mathcal{A}_i)}} \mathop{\mathbb{E}}_{\substack{a_i \sim \nu \\ a_{-i} \sim \pi_{-i}(x)}} |\bar{g}(x, a) - \bar{g}'(x, a)| \\
&\leq \sup_{\substack{x \in \mathcal{S} \\ a \in \mathcal{A}}} \left| \bar{g}(x, a) - \bar{g}'(x, a) \right| \\
&\leq \sup_{\substack{x \in \mathcal{S} \\ a \in \mathcal{A}}} | \min\{\bar{w}_i^\top \phi(x, a) + \beta \bar{u}(x, a), H\} - \min\{(\bar{w}_i')^\top \phi(x, a) + \beta \bar{u}'(x, a), H\}| \\
&\overset{(b)}{\leq} \sup_{\substack{x \in \mathcal{S} \\ a \in \mathcal{A}}} |\bar{w}_i^\top \phi(x, a) + \beta \bar{u}(x, a) - ((\bar{w}_i')^\top \phi(x, a) + \beta \bar{u}'(x, a))| \\
&\leq \sup_{\substack{x \in \mathcal{S} \\ a \in \mathcal{A}}} |(\bar{w}_i - \bar{w}_i')^\top \phi(x, a)| + \beta \sup_{\substack{x \in \mathcal{S} \\ a \in \mathcal{A}}} |\bar{u}(x, a) - \bar{u}'(x, a)|
\end{aligned} \tag{A.2}$$

Inequality (a) follows from the property $|\max_{\nu \in \Delta(\mathcal{A}_i)} f(\nu) - \max_{\nu \in \Delta(\mathcal{A}_i)} h(\nu)| \leq \max_{\nu \in \Delta(\mathcal{A}_i)} |f(\nu) - h(\nu)|$ for any $f, h : \Delta(\mathcal{A}_i) \to \mathbb{R}$ since, letting $\bar{\nu} = \operatorname{argmax}_{\nu \in \Delta(\mathcal{A}_i)} f(\nu)$ and $\widetilde{\nu} = \operatorname{argmax}_{\nu \in \Delta(\mathcal{A}_i)} h(\nu)$, we observe that: (i) if $\max_{\nu \in \Delta(\mathcal{A}_i)} f(\nu) > \max_{\nu \in \Delta(\mathcal{A}_i)} h(\nu)$, then $\max_{\nu \in \Delta(\mathcal{A}_i)} f(\nu) - \max_{\nu \in \Delta(\mathcal{A}_i)} h(\nu) \leq f(\bar{\nu}) - h(\bar{\nu}) \leq \max_{\nu \in \Delta(\mathcal{A}_i)} |f(\nu) - h(\nu)|$; and (ii) if $\max_{\nu \in \Delta(\mathcal{A}_i)} f(\nu) \leq \max_{\nu \in \Delta(\mathcal{A}_i)} h(\nu)$, then $\max_{\nu \in \Delta(\mathcal{A}_i)} h(\nu) - \max_{\nu \in \Delta(\mathcal{A}_i)} f(\nu) \leq h(\widetilde{\nu}) - f(\widetilde{\nu}) \leq \max_{\nu \in \Delta(\mathcal{A}_i)} |f(\nu) - h(\nu)|$. Inequality (b) follows from $\min\{\cdot, H\}$ being a non-expansive operator.

We can continue bounding,

$$\begin{aligned}
\operatorname{dist}(V, V') &\overset{(a)}{\leq} \sup_{\phi : \|\phi\| \leq 1} \left| (\bar{w}_i - \bar{w}_i')^\top \phi \right| \\
&\qquad + \sup_{\phi : \|\phi\| \leq 1} \beta \left| \sqrt{\phi^\top \bar{\Lambda}^{-1} \phi} - \sqrt{\phi^\top (\bar{\Lambda}')^{-1} \phi} \right| \\
&\overset{(b)}{\leq} \|\bar{w}_i - \bar{w}_i'\| + \sup_{\phi : \|\phi\| \leq 1} \beta \sqrt{|\phi(x, a)^\top (\bar{\Lambda}^{-1} - (\bar{\Lambda}')^{-1}) \phi(x, a)|} \\
&= \|\bar{w}_i - \bar{w}_i'\| + \beta \sqrt{\|\bar{\Lambda}^{-1} - (\bar{\Lambda}')^{-1}\|} \\
&\leq \|\bar{w}_i - \bar{w}_i'\| + \beta \sqrt{\|\bar{\Lambda}^{-1} - (\bar{\Lambda}')^{-1}\|_F}
\end{aligned} \tag{A.3}$$

where (a) follows from the assumption $\sup_{x \in \mathcal{S}} \max_{a \in \mathcal{A}} \|\phi(x, a)\| \leq 1$, and (b) follows from the inequality $|\sqrt{p} - \sqrt{q}| \leq \sqrt{|p - q|}$ for any $p, q \geq 0$. Now, we notice that (A.3) is a bound of the same form of equation (28) from (Jin et al., 2020, Lemma D.6), and so we can use the proof of this lemma to obtain that the ϵ_i-covering number of $\mathcal{V}_i$, denoted by $\mathcal{N}_{\epsilon_i}$ can be upper bounded as $\log \mathcal{N}_{\epsilon_i} \leq d \log(1 + 4L/\epsilon) + d^2 \log\left[1 + 8d^{1/2}\beta^2/(\lambda\epsilon^2)\right]$. This finishes the proof. □

B PROVING THEOREM 3.1

For simplicity, we will use the following notation: at episode k, we denote $\pi^{i,k} = \{\pi_h^{i,k}\}_{h\in[H]}$ as the policy induced by $\{Q_h^{i,k}\}_{h=1}^H$ as performed by agent $i \in [n]$ (line 14 of Algorithm 1) across time steps $h \in [H]$, thus for a fixed step $h \in [H]$ we let $V_h^{i,k}(x_h^k) = \mathbb{E}_{a\sim\pi_h^k(x_h^k)}[Q_h^{i,k}(x_h^k, a)]$ with $\pi_h^k(x_h^k)$ being a Nash equilibrium from the stage game $(Q_h^{i,k}(x_h^k, \cdot))_{i\in[n]}$. With some abuse of notation, we similarly define $V_h^{i,k}(x) = \mathbb{E}_{a\sim\pi_h^k(x)}[Q_h^{i,k}(x, a)]$ with $\pi_h^k(x)$ being a Nash equilibrium from the game $(Q_h^{i,k}(x, \cdot))_{i\in[n]}$. Let $\phi_h^\tau := \phi(x_h^\tau, a_h^\tau)$.

B.1 PRELIMINARY TECHNICAL RESULTS

We now bound the parameters $\{w_h^{i,k}\}_{(i,h,k)\in[n]\times[H]\times[K]}$ from the NQOVI algorithm.

Lemma B.1 (Parameter bound – Lemma B.2 in (Jin et al., 2020))**.** *For any $(i, k, h) \in [n] \times [K] \times [H]$, the parameter $w_h^{i,k}$ in the NQOVI algorithm satisfies $\left\|w_h^{i,k}\right\| \leq (1 + H)\sqrt{\frac{d(k-1)}{\lambda}}$.*

Now we use Lemma A.2 and Lemma A.1 to prove a useful concentration bound for NQOVI.

Lemma B.2 (Concentration bound on value functions for NQOVI – Lemma B.3 in (Jin et al., 2020))**.** *Consider the setting of Theorem 3.1. There exists an absolute constant C independent of c_β such that for any fixed $\delta \in (0, 1)$, the following event $\mathcal{E}_i$ holds with probability at least $1 - \delta$ for a fixed $i \in [n]$: for every $(k, h) \in [K] \times [H]$,*

$$\left\|\sum_{\tau=1}^{k-1} \phi_h^\tau [V_{h+1}^{i,k}(x_{h+1}^\tau) - \mathbb{P}_h V_{h+1}^{i,k}(x_h^\tau, a_h^\tau)]\right\|_{(\Lambda_h^k)^{-1}} \leq CdH\sqrt{\log[(c_\beta + 1)dKH/\delta]}.$$

The following lemma crucially depends on the principle of optimism.

Lemma B.3 (Difference with an arbitrary Q-function – Lemma B.4 in (Jin et al., 2020))**.** *Consider the setting of Theorem 3.1. There exists an absolute constant c_β such that for $\beta = c_\beta dH\sqrt{\iota}$ with $\iota = \log(dKH/\delta)$ and any fixed joint policy $\bar{\pi}$, such that for any $i \in [n]$: given the event $\mathcal{E}_i$ defined in Lemma B.2, we have for all $(x, a, h, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$ that*

$$\langle \phi(x, a), w_h^{i,k}\rangle - Q_h^{i,\bar{\pi}}(x, a) = \mathbb{P}_h(V_{h+1}^{i,k} - V_{h+1}^{i,\bar{\pi}})(x, a) + \Delta_h^{i,k}(x, a),$$

for some $\Delta_h^{i,k}(x, a)$ such that $|\Delta_h^{i,k}(x, a)| \leq \beta\sqrt{\phi(x, a)^\top(\Lambda_h^k)^{-1}\phi(x, a)}$.

The following key lemma makes use of optimism by using Lemma B.3 and of the fact that we choose a Nash equilibrium at each stage game.

Lemma B.4 (Optimism bounds)**.** *Consider the setting of Theorem 3.1. Given the event $\mathcal{E}_i$ defined in Lemma B.2, we have that for all $(x, a, h, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$,*

$$Q_h^{i,\mathrm{br}(\pi_{-i}^k),\pi_{-i}^k}(x, a) \leq Q_h^{i,k}(x, a) \quad \textit{and} \quad V_h^{i,\mathrm{br}(\pi_{-i}^k),\pi_{-i}^k}(x) \leq V_h^{i,k}(x).$$

Proof. We prove the claims by induction in $h = H + 1, \ldots, 1$. The base case $H + 1$ is trivial, since $Q_{H+1}^{i,\mathrm{br}(\pi_{-i}^k),\pi_{-i}^k}(x, a) = Q_{H+1}^{i,k}(x, a) = 0$. Now, at step $h + 1$ we have the induction hypothesis $Q_{h+1}^{i,\mathrm{br}(\pi_{-i}^k),\pi_{-i}^k}(x, a) \leq Q_{h+1}^{i,k}(x, a)$. Then we have that

$$\begin{aligned} V_{h+1}^{i,\mathrm{br}(\pi_{-i}^k),\pi_{-i}^k}(x) &\overset{(a)}{=} \max_{\nu\in\Delta(\mathcal{A}_i)} \mathbb{E}_{\substack{a_i\sim\nu \\ a_{-i}\sim\pi_{-i,h+1}^k(x)}}[Q_{h+1}^{i,\mathrm{br}(\pi_{-i}^k),\pi_{-i}^k}(x, a)] \\ &\overset{(b)}{\leq} \max_{\nu\in\Delta(\mathcal{A}_i)} \mathbb{E}_{\substack{a_i\sim\nu \\ a_{-i}\sim\pi_{-i,h+1}^k(x)}}[Q_{h+1}^{i,k}(x, a)] \\ &\overset{(c)}{=} \mathbb{E}_{a\sim\pi_{h+1}^k(x)}[Q_{h+1}^{i,k}(x, a)] \\ &= V_{h+1}^{i,k}(x, a), \end{aligned} \tag{B.1}$$

where (a) follows by definition of best response, (b) from the induction hypothesis, and (c) from the fact that NQOVI chooses a Nash equilibrium at every stage game.

Now, we have

$$\begin{aligned}
Q_h^{i,\mathrm{br}(\pi^k_{-i}),\pi^k_{-i}}(x,a) &\overset{(a)}{\le} \langle \phi(x,a), w_h^{i,k}\rangle + \mathbb{P}_h(V_{h+1}^{i,k} - V_{h+1}^{i,\mathrm{br}(\pi^k_{-i}),\pi^k_{-i}})(x,a) + \beta\sqrt{\phi(x,a)^\top(\Lambda_h^k)^{-1}\phi(x,a)} \\
&\overset{(b)}{\le} \langle \phi(x,a), w_h^{i,k}\rangle + \beta\sqrt{\phi(x,a)^\top(\Lambda_h^k)^{-1}\phi(x,a)} \\
\overset{(c)}{\implies} Q_h^{i,\mathrm{br}(\pi^k_{-i}),\pi^k_{-i}}(x,a) &\le \min\{\langle \phi(x,a), w_h^{i,k}\rangle + \beta\sqrt{\phi(x,a)^\top(\Lambda_h^k)^{-1}\phi(x,a)}, H\} \\
&= Q_h^{i,k}(x,a),
\end{aligned} \tag{B.2}$$

where (a) follows from Lemma B.3, (b) from (B.1), and (c) from $Q_h^{i,\mathrm{br}(\pi^k_{-i}),\pi^k_{-i}} \le H$. From here we can repeat the steps in (B.1) to obtain $V_h^{i,\mathrm{br}(\pi^k_{-i}),\pi^k_{-i}}(x) \le V_h^{i,k}(x)$. This finishes the proof. □

B.2 PROOF OF THEOREM 3.1

Let us first condition on the event $\bigcap_{i=1}^n \mathcal{E}_i$ where $\mathcal{E}_i$ is defined in Lemma B.2. Since $\mathbb{P}[\text{not } \mathcal{E}_i] \le \delta$, applying union bound let us conclude that $\mathbb{P}[\bigcap_{i\in[n]} \mathcal{E}_i] \ge 1 - n\delta$.

For any $k \in [K]$, given the policy $\pi^k = \{\pi_i^k\}_{i\in[n]}$ defined by NQOVI, we define the functions $\widehat{Q}_h^k$ and $\widehat{V}_h^k$ recursively as: $\widehat{V}_{H+1}^k(x) = \widehat{Q}_{H+1}^k(x) = 0$ and

$$\begin{aligned}
\widehat{Q}_h^k(x,a) &= \mathbb{P}_h\widehat{V}_{h+1}^k(x,a) + 2\beta\sqrt{(\phi_h^k)^\top(\Lambda_h^k)^{-1}\phi_h^k}, \\
\widehat{V}_h^k(x) &= \mathbb{E}_{a\sim\pi_h^k}[\widehat{Q}_h^k(x,a)]
\end{aligned}$$

for any $h = H, \dots, 1$ and $(x,a) \in \mathcal{S}\times\mathcal{A}$. Notice that since $2\beta\sqrt{(\phi_h^k)^\top(\Lambda_h^k)^{-1}\phi_h^k} \le 2\beta\sqrt{(\phi_h^k)^\top\phi_h^k} = 2\beta\left\|\phi_h^k\right\| \le 2\beta$, we have that $\widehat{Q}_h^k$ and $\widehat{V}_h^k$ are nonnegative with maximum value $2\beta H$.

Let $k \in [K]$. We claim that for any $(h,x,a) \in [H]\times\mathcal{S}\times\mathcal{A}$,

$$\begin{aligned}
\max_{i\in[n]}(Q_h^{i,k}(x,a) - Q_h^{i,\pi^k}(x,a)) &\le \widehat{Q}_h^k(x,a), \text{ and} \\
\max_{i\in[n]}(V_h^{i,k}(x) - V_h^{i,\pi^k}(x)) &\le \widehat{V}_h^k(x).
\end{aligned} \tag{B.3}$$

We prove the claim by induction in $h = H+1, \dots, 1$. The base case $H+1$ is trivial, since $Q_{H+1}^{i,k}(x,a) = Q_{H+1}^{i,\pi^k}(x,a) = \widehat{Q}_{H+1}^k(x,a) = 0$ for every $i \in [n]$. Now, at step $h+1$ we have the induction hypothesis $\max_{i\in[h]}(Q_{h+1}^{i,k}(x,a) - Q_{h+1}^{i,\pi^k}(x,a)) \le \widehat{Q}_{h+1}^k(x,a)$. Taking expectations over $a \sim \pi_{h+1}^k(x)$ let us immediately obtain

$$\max_{i\in[h]}(V_{h+1}^{i,k}(x) - V_{h+1}^{i,\pi^k}(x)) \le \widehat{V}_{h+1}^k(x). \tag{B.4}$$

Now, for any $i \in [n]$,

$$\begin{aligned}
Q_h^{i,k}(x,a) - Q_h^{i,\pi^k}(x,a) &= \min\{(w_h^{i,k})^\top\phi(x,a) + \beta\sqrt{\phi(x,a)^\top(\Lambda_h^k)^{-1}\phi(x,a)}, H\} - Q_h^{i,\pi^k}(x,a) \\
&\overset{(a)}{\le} \mathbb{P}_h(V_{h+1}^{i,k} - V_{h+1}^{i,\pi^k})(x,a) + 2\beta\sqrt{\phi(x,a)^\top(\Lambda_h^k)^{-1}\phi(x,a)} \\
&\overset{(b)}{\le} \mathbb{P}_h\widehat{V}_{h+1}^k(x,a) + 2\beta\sqrt{\phi(x,a)^\top(\Lambda_h^k)^{-1}\phi(x,a)} \\
&= \widehat{Q}_h^k(x,a),
\end{aligned} \tag{B.5}$$

where (a) follows from Lemma B.3 and (b) from (B.4). Taking expectations let us obtain $V_h^{i,k}(x) - V_h^{i,\pi^k}(x) \leq \widehat{V}_h^k(x)$. This finishes the proof for the claim in (B.3).

We now introduce the following notation: $\delta_h^k := \mathbb{E}_{a \sim \pi_h^k(x_h^k)}[\widehat{Q}_h^k(x_h^k, a)] - \widehat{Q}_h^k(x_h^k, a_h^k)$, and $\xi_{h+1}^k := \mathbb{P}_h \widehat{V}_{h+1}^k(x_h^k, a_h^k) - \widehat{V}_{h+1}^k(x_{h+1}^k)$ with $\xi_1^k := 0$. Then, for any $(h,k) \in [H] \times [K]$,

$$\begin{aligned}
\widehat{V}_h^k(x_h^k) &= \mathbb{E}_{a \sim \pi_h^k(x_h^k)}[\widehat{Q}_h^k(x_h^k, a)] \\
&= \delta_h^k + \widehat{Q}_h^k(x_h^k, a_h^k) \\
&= \delta_h^k + \mathbb{P}_h \widehat{V}_{h+1}^k(x_h^k, a_h^k) + 2\beta\sqrt{(\phi_h^k)^\top (\Lambda_h^k)^{-1}\phi_h^k} \\
&= \delta_h^k + \xi_{h+1}^k + 2\beta\sqrt{(\phi_h^k)^\top (\Lambda_h^k)^{-1}\phi_h^k} + \widehat{V}_{h+1}^k(x_{h+1}^k).
\end{aligned}$$

Now, let us focus on the regret performance metric.

$$\begin{aligned}
\text{Regret}(K) &= \sum_{k=1}^K \max_{i \in [n]} (V_1^{i, \text{br}(\pi_{-i}^k), \pi_{-i}^k}(s_o) - V_1^{i,\pi^k}(s_o)) \\
&\overset{(a)}{\leq} \sum_{k=1}^K \max_{i \in [n]} (V_1^{i,k}(s_o) - V_1^{i,\pi^k}(s_o)) \\
&\overset{(b)}{\leq} \sum_{k=1}^K \widehat{V}_1^k(s_o) \\
&= \underbrace{\sum_{k=1}^K \sum_{h=1}^H \xi_h^k}_{\text{(I)}} + \underbrace{\sum_{k=1}^K \sum_{h=1}^H \delta_h^k}_{\text{(II)}} + \underbrace{2\beta \sum_{k=1}^K \sum_{h=1}^H \sqrt{(\phi_h^k)^\top (\Lambda_h^k)^{-1}\phi_h^k}}_{\text{(III)}},
\end{aligned} \tag{B.6}$$

where (a) follows from Lemma B.4 and the fact that we are conditioned on the event $\bigcap_{i=1}^n \mathcal{E}_i$; (b) follows from (B.3).

We first analyze the term (I) from (B.6). Let us define the filtration $\{\mathcal{F}_{(k,h)}\}_{(k,h) \in \mathcal{L}^\star}$ where $\mathcal{L}^\star$ is a sequence such that $\mathcal{L}^\star \subset \mathbb{Z}_{\geq 1} \times [H]$ and its elements are arranged as follows. Firstly, we let the second coordinate take values from 1 to H and repeat this periodically *ad infinitum*, so that each period has H elements of $\mathcal{L}^\star$. Finally, we let the first coordinate take the value corresponding to the current number of periods so far progressed in the second coordinate (and so its value is unbounded). Consider any element $(k,h) \in \mathcal{L}^\star$. We denote by $(k,h)^{-1}$ its previous element in $\mathcal{L}^\star$. We let $\mathcal{F}_{(k,h)}$ contain the information of the tuple $(x_{\bar{h}}^{\bar{k}}, a_{\bar{h}}^{\bar{k}})$ whose indexes $(\bar{k}, \bar{h})$ belong to the set $\mathcal{L}^\star$ up to the element $(k,h) \in \mathcal{L}^\star$.

We then can conclude that $\{\xi_h^k\}_{(k,h) \in \mathcal{L}^\star}$ is a martingale difference sequence due to the following two properties:

1. $\xi_h^k \in \mathcal{F}_{(k,h)^{-1}}$. For $h = 1$, $\mathbb{E}[\xi_h^k | \mathcal{F}_{(k,h)^{-1}}] = 0$ is trivial, so we focus on $h = 2, \ldots, H$. Then, since $x_h^k \sim \mathcal{P}_{h-1}(\cdot | x_{h-1}^k, a_{h-1}^k)$ (line 16 of NQOVI), we have $\mathbb{E}[\widehat{V}_h^k(x_h^k) | \mathcal{F}_{(k,h)^{-1}}] = \mathbb{E}_{x' \sim \mathcal{P}_{h-1}(\cdot | x_{h-1}^k, a_{h-1}^k)}[\widehat{V}_h^k(x')] = \mathbb{P}_{h-1}\widehat{V}_h^k(x_{h-1}^k, a_{h-1}^k)$, which immediately implies $\mathbb{E}[\xi_h^k | \mathcal{F}_{(k,h)^{-1}}] = 0$.
2. $|\xi_h^k| \leq |\mathbb{P}_{h-1}\widehat{V}_h^k(x_{h-1}^k, a_{h-1}^k)| + |\widehat{V}_h^k(x_h^k)| \leq 4\beta H < \infty$ since $\widehat{V}_h^k(x) \in [0, 2\beta H]$ for any $x \in \mathcal{S}$.

Therefore, we can use the Azuma-Hoeffding inequality to conclude that, for any $\epsilon > 0$,

$$\Pr\left(\sum_{k=1}^K \sum_{h=1}^H \xi_h^k > \epsilon\right) \leq \exp\left(\frac{-2\epsilon^2}{(KH)(16\beta^2 H^2)}\right).$$

We choose $\epsilon = \sqrt{8KH^3\beta^2 \log\left(\frac{1}{\delta}\right)}$. Then, with probability at least $1 - \delta$,

$$\text{(I)} = \sum_{k=1}^K \sum_{h=1}^H \xi_h^k \leq \sqrt{8KH^3\beta^2 \log\left(\frac{1}{\delta}\right)} \leq 8\beta H\sqrt{KH\iota}, \tag{B.7}$$

recalling that $\iota = \log\left(\frac{dKH}{\delta}\right)$. We call $\bar{\mathcal{E}}$ the event such that (B.7) holds.

The term (II) can be analyzed in a very similar way as in (I) to show that $\{\delta_h^k\}_{(k,h)\in\mathcal{L}^*}$ is a martingale difference sequence, and thus obtain that with probability at least $1-\delta$,

$$\text{(II)} = \sum_{k=1}^{K}\sum_{h=1}^{H} \delta_h^k \leq 8\beta H\sqrt{KH\iota}. \tag{B.8}$$

We call $\widetilde{\mathcal{E}}$ the event such that (B.8) holds.

We now analyze the term (III) from (B.6). Then for a fixed $h \in [H]$,

$$\sum_{k=1}^{K} (\phi_h^k)^\top (\Lambda_h^k)^{-1}\phi_h^k \leq 2\log\left[\frac{\det(\Lambda_h^{K+1})}{\det(\lambda I_d)}\right]$$

where the inequality follows from the so-called elliptical potential lemma (Abbasi-yadkori et al., 2011, Lemma 11), whose conditions are satisfied from our bounded sequence $\{\phi_h^k\}_{k=1}^K$ and the fact that the minimum eigenvalue of Λ_h^k is lower bounded by $\lambda = 1$ for every $(h,k) \in [H]\times[K]$. Now, we have that Λ_h^{K+1} is a positive definite matrix whose maximum eigenvalue can be bounded as $\|\Lambda_h^{K+1}\| \leq \left\|\sum_{k=1}^K \phi_h^k(\phi_h^k)^\top\right\| + \lambda \leq K+\lambda$, and so $\det(\Lambda_h^{K+1}) \leq \det((K+\lambda)I_d) = (K+\lambda)^d$. We also have that $\det(\lambda I_d) = \lambda^d$. Then, we obtain that

$$\sum_{k=1}^{K} (\phi_h^k)^\top (\Lambda_h^k)^{-1}\phi_h^k \leq 2\log\left[\frac{K+\lambda}{\lambda}\right]^d = 2d\log(K+1) \leq 2d\iota, \tag{B.9}$$

where the last inequality holds since $\log(K+1) \leq \log\left(\frac{dKH}{\delta}\right) = \iota$ for $d \geq 2$, $\delta > 0$.

Now, going back to term (III),

$$\text{(III)} = 2\beta\sum_{h=1}^{H}\sum_{k=1}^{K}\sqrt{(\phi_h^k)^\top(\Lambda_h^k)^{-1}\phi_h^k} \overset{\text{(a)}}{\leq} 2\beta\sum_{h=1}^{H}\sqrt{K}\sqrt{\sum_{k=1}^{K}(\phi_h^k)^\top(\Lambda_h^k)^{-1}\phi_h^k} \overset{(b)}{\leq} 2\beta H\sqrt{2dK\iota}, \tag{B.10}$$

where (a) follows from the Cauchy-Schwartz inequality, and (b) from (B.9).

Now, using the results in (B.7), (B.8), and (B.10) back in (B.6), we conclude that,

$$\begin{aligned}\text{Regret}(K) &\leq 8\beta H\sqrt{KH\iota} + 8\beta H\sqrt{KH\iota} + 2\beta H\sqrt{dK\iota} \\ &= 16c_\beta\sqrt{d^2KH^5\iota^2} + 2c_\beta\sqrt{d^3KH^4\iota^2} \overset{\text{(a)}}{\leq} 18c_\beta\sqrt{d^3KH^5\iota^2},\end{aligned} \tag{B.11}$$

where (a) follows from $\sqrt{\iota} \leq \iota$ which follows from equation (C.5).

Finally, applying union bound let us conclude that $\mathbb{P}[\bigcap_{i\in[n]}\mathcal{E}_i \cap \bar{\mathcal{E}} \cap \widetilde{\mathcal{E}}] \geq 1-(n+2)\delta$, i.e., our final result holds with probability at least $1-(n+2)\delta$. This finishes the proof of Theorem 3.1. □

C REMAINING PROOFS

Proof of Lemma A.1. First, from our assumptions, for any $g \in \mathcal{G}$, there exists a $\widetilde{g}$ in the ϵ-covering such that $g = \widetilde{g} + \Delta_g$ with $\sup_{x\in\mathcal{S}}|\Delta_g(x)| \leq \epsilon$. Then,

$$\begin{aligned}&\left\|\sum_{\tau=1}^{A}\phi_\tau\{g(x_\tau) - \mathbb{E}[g(x_\tau)|\mathcal{F}_{\tau-1}]\}\right\|_{\Lambda_A^{-1}}^2 \\ &\leq 2\underbrace{\left\|\sum_{\tau=1}^{A}\phi_\tau\{\widetilde{g}(x_\tau) - \mathbb{E}[\widetilde{g}(x_\tau)|\mathcal{F}_{\tau-1}]\}\right\|_{\Lambda_A^{-1}}^2}_{\text{(I)}} + 2\underbrace{\left\|\sum_{\tau=1}^{A}\phi_\tau\{\Delta_g(x_\tau) - \mathbb{E}[\Delta_g(x_\tau)|\mathcal{F}_{\tau-1}]\}\right\|_{\Lambda_A^{-1}}^2}_{\text{(II)}},\end{aligned} \tag{C.1}$$

where we used $\|a+b\| \le \|a\| + \|b\| \implies \|a+b\|^2 \le \|a\|^2 + \|b\|^2 + 2\|a\|\|b\| \le 2\|a\|^2 + 2\|b\|^2$ for any $a, b \in \mathbb{R}^d$, and which actually holds for any weighted Euclidean norm.

We start by analyzing the term (I) in equation (C.1). Let $\varepsilon_\tau := \widetilde{g}(x_\tau) - \mathbb{E}[\widetilde{g}(x_\tau)|\mathcal{F}_{\tau-1}]$. Now, we observe that 1) $\mathbb{E}[\varepsilon_\tau|\mathcal{F}_{\tau-1}] = 0$ and 2) $\varepsilon_\tau \in [-H, H]$ since $\widetilde{g}(x_\tau) \in [0, H]$. From these two facts we obtain that $\varepsilon_\tau|\mathcal{F}_{\tau-1}$ is H-sub-Gaussian. Therefore we can apply the concentration bound of self-normalized processes from Theorem 1 of (Abbasi-yadkori et al., 2011) along with a union bound over the ϵ-covering of $\mathcal{G}$ to conclude that, with probability at least $1-\delta$,

$$\text{(I)} = \left\| \sum_{\tau=1}^{A} \phi_\tau \varepsilon_\tau \right\|^2_{\Lambda_A^{-1}} \le \log\left(\frac{\det(\Lambda_A)^{1/2} \det(\lambda I_d)^{-1/2}}{\delta/\mathcal{N}_\epsilon} \right) \overset{\text{(a)}}{\le} 2H^2 \left(\frac{d}{2} \log\left(\frac{\lambda + AB/d}{\lambda} \right) + \log\left(\frac{\mathcal{N}_\epsilon}{\delta} \right) \right), \quad \text{(C.2)}$$

where (a) follows from $\det(\lambda I_d) = \lambda^d$ and from the determinant-trace inequality from Lemma 10 in (Abbasi-yadkori et al., 2011) which let us obtain $\det(\Lambda_A) \le (\lambda + AB/d)^d$.

Now we analyze the term (II) in equation (C.1). Let $\bar{\varepsilon}_\tau := \Delta_g(x_\tau) - \mathbb{E}[\Delta_g(x_\tau)|\mathcal{F}_{\tau-1}]$. Then,

$$\left\| \sum_{\tau=1}^{A} \phi_\tau \bar{\varepsilon}_\tau \right\| \le \sum_{\tau=1}^{A} \|\phi_\tau \bar{\varepsilon}_\tau\| \overset{\text{(a)}}{\le} \sum_{\tau=1}^{A} |\bar{\varepsilon}_\tau| \le \sum_{\tau=1}^{A} |\Delta_g(x_\tau)| + |\mathbb{E}[\Delta_g(x_\tau)|\mathcal{F}_{\tau-1}]| \le \sum_{\tau=1}^{A} 2\epsilon = 2A\epsilon,$$

where (a) follows from $\|\phi_\tau\| \le 1$. Thus, using this result, we obtain

$$\text{(II)} \le \frac{1}{\lambda} \left\| \sum_{\tau=1}^{A} \phi_\tau \bar{\varepsilon}_\tau \right\|^2 \le \frac{1}{\lambda} 4A^2\epsilon^2.$$

We finish the proof by multiplying by two the terms (I) and (II), and then adding them up to use them as an upper bound to (C.1) . □

Proof of Lemma B.1. For any vector $v \in \mathbb{R}^d$,

$$\begin{aligned}
|v^\top w_h^{i,k}| &= |v^\top (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi_h^\tau [r_h^i + \max_{\substack{a \sim \pi^* \\ \pi^* \text{ as in line 7 of Algorithm 1}}} Q_{h+1}^{i,k}(x_{h+1}^\tau, a)]| \\
&\overset{\text{(a)}}{\le} (1+H) \sum_{\tau=1}^{k-1} |v^\top (\Lambda_h^k)^{-1} \phi_h^\tau| \\
&\overset{(b)}{\le} (1+H) \sqrt{\left[\sum_{\tau=1}^{k-1} v^\top (\Lambda_h^k)^{-1} v \right] \left[\sum_{\tau=1}^{k-1} (\phi_h^\tau)^\top (\Lambda_h^k)^{-1} \phi_h^\tau \right]} \\
&\overset{\text{(c)}}{\le} (1+H)\sqrt{d} \sqrt{\sum_{\tau=1}^{k-1} v^\top (\Lambda_h^k)^{-1} v} \\
&\overset{\text{(d)}}{\le} (1+H) \sqrt{\frac{d(k-1)}{\lambda}} \|v\|,
\end{aligned}$$

where (a) follows from the bounded rewards and $Q_{h+1}^{i,k}(\cdot,\cdot) \le H$; (b) from applying Cauchy-Schwarz twice as in the following series of inequalities: given $q = (q_1, \dots, q_m)$ and $q = (p_1, \dots, p_m)$ where q_i and p_i are vectors of same arbitrary dimension we have $\sum_{i=1}^m |q_i^\top p_i| \le \sum_{i=1}^m \|q_i\| \|p_i\| \le \sqrt{\sum_{i=1}^m \|q_i\|} \sqrt{\sum_{i=1}^m \|p_i\|}$; (c) follows from (Jin et al., 2020, Lemma D.1); and (d) from $(\Lambda_h^k)^{-1} \preceq \lambda^{-1} I_d$. The proof concludes by considering that $\left\| w_h^{i,k} \right\| = \max_{v:\|v\|=1} |v^\top w_h^{i,k}|$. □

Proof of Lemma B.2. We obtain that, with probability at least $1-\delta$, $\delta \in (0,1)$,

$$\begin{aligned}
&\left\| \sum_{\tau=1}^{k-1} \phi_h^\tau [V_{h+1}^{i,k}(x_{h+1}^\tau) - \mathbb{P}_h V_{h+1}^{i,k}(x_h^\tau, a_h^\tau)] \right\|_{(\Lambda_h^k)^{-1}}^2 \\
&\overset{\text{(a)}}{\le} 4H^2 \left[\frac{d}{2} \log\left(\frac{\lambda + (k-1)/d}{\lambda} \right) + \log \mathcal{N}_{\epsilon_i} + \log \frac{1}{\delta} \right] + \frac{8(k-1)^2\epsilon^2}{\lambda} \\
&\overset{(b)}{\le} 4H^2 \left[\frac{d}{2} \log\left(\frac{\lambda + (k-1)/d}{\lambda} \right) + d \log\left(1 + \frac{4(1+H)\sqrt{d(k-1)}}{\epsilon\sqrt{\lambda}} \right) \right. \\
&\quad \left. + d^2 \log\left(1 + \frac{8d^{1/2}\beta^2}{\lambda\epsilon^2} \right) + \log\frac{1}{\delta} \right] + \frac{8(k-1)^2\epsilon^2}{\lambda}
\end{aligned} \tag{C.3}$$

where (a) is a direct application of Lemma A.1; and (b) follows from the realization that, from lines 9 and 10 in Algorithm 1, $V_{h+1}^{i,k}(\cdot) \in \mathcal{V}$ with $\mathcal{V}$ as in Lemma A.2 and so we can use the bound on the covering number derived in such lemma with $L = (1+H)\sqrt{\frac{d(k-1)}{\lambda}}$ by using the bound from Lemma B.1.

Recalling that $\lambda = 1$ and $\beta = c_\beta dH\iota$ with $\iota = \log(dKH/\delta)$ in the setting of Theorem 3.1, we claim that, after setting $\epsilon = \frac{dH}{K}$ in our previous equation, there exists an absolute constant $C > 0$ independent of c_β such that

$$\left\| \sum_{\tau=1}^{k-1} \phi_h^\tau [V_{h+1}^{i,k}(x_{h+1}^\tau) - \mathbb{P}_h V_{h+1}^{i,k}(x_h^\tau, a_h^\tau)] \right\|_{(\Lambda_h^k)^{-1}}^2 \le Cd^2H^2 \log((c_\beta+1)dKH/\delta). \tag{C.4}$$

Proving (C.4) would conclude the proof.

We first introduce a couple of useful results:

$$\iota^2 = \log\left(\frac{dKH}{\delta} \right) \ge \log(dKH) \ge \log(4) > 1, \tag{C.5}$$

$$\log\left(\frac{(c_\beta+1)dKH}{\delta} \right) = \log(c_\beta+1) + \iota \ge \iota > 1. \tag{C.6}$$

Replacing $\lambda = 1$ and $\epsilon = \frac{dH}{K}$ in the right-hand side of (C.3) and doing some algebraic calculations, let us conclude that

$$\begin{aligned}
\text{(C.3)} \le 4d^2H^2 &\left[\log\left(1 + \frac{K}{d} \right) + \log\left(1 + \frac{8K^{3/2}}{d^{1/2}} \right) + \log\left(\frac{1}{\delta}\left(1 + \frac{8\beta^2K^2}{d^{3/2}H^2} \right) \right) \right] \\
&+ 8d^2H^2.
\end{aligned} \tag{C.7}$$

Replacing $\beta = c_\beta dH\sqrt{\iota}$ in the previous expression and doing some algebraic work let us obtain

$$\begin{aligned}
\text{(C.7)} \le\ &\underbrace{8d^2H^2 \log\left(1 + \frac{8K^{3/2}}{d^{1/2}} \right)}_{\text{(I)}} + \underbrace{4d^2H^2 \log\left(\frac{1}{\delta}\left(1 + 8c_\beta^2 d^{1/2}\iota K^2 \right) \right)}_{\text{(II)}} \\
&+ 8d^2H^2 \log\left(\frac{(c_\beta+1)dKH}{\delta} \right)
\end{aligned} \tag{C.8}$$

where the inequality has made use of (C.6). We now upper bound the terms highlighted in (C.8). Then,

$$\begin{aligned}
\text{(I)} &\le 8d^2H^2 \log\left(1 + 8K^{3/2} \right) \\
&\overset{\text{(a)}}{\le} 8d^2H^2 \log\left(\frac{(1+c_\beta)^2(dKH)^2}{\delta^2} \right) + 8d^2H^2 \log(9) \log\left(\frac{(c_\beta+1)dKH}{\delta} \right) \\
&= (16 + 8\log(9))d^2H^2 \log\left(\frac{(c_\beta+1)dKH}{\delta} \right),
\end{aligned}$$

where (a) follows from (C.6) and $c_\beta > 0$. Similarly,

$$
\begin{aligned}
\text{(II)} &\overset{\text{(a)}}{\le} 4d^2H^2 \log\left(\frac{8(c_\beta+1)^2\iota(dKH)^2}{\delta}\right) \\
&\overset{(b)}{\le} 4d^2H^2 \log\left(\frac{(c_\beta+1)^2\iota(dKH)^2}{\delta^2}\right) + 4d^2H^2\log(8) \\
&= 4d^2H^2 \log\left(\frac{(c_\beta+1)^2(dKH)^2}{\delta^2}\right) + 4d^2H^2\log(\iota) + 4d^2H^2\log(8) \\
&\overset{\text{(c)}}{\le} 8d^2H^2 \log\left(\frac{(c_\beta+1)dKH}{\delta}\right) + 4d^2H^2\iota + 4d^2H^2\log(8) \\
&\overset{\text{(d)}}{\le} (12+4\log(8))d^2H^2 \log\left(\frac{(c_\beta+1)dKH}{\delta}\right)
\end{aligned}
$$

where (a) follows from $c_\beta > 0$, (b) from $\delta^2 < \delta$, (c) from $\log(\iota) < \iota$ (since $\iota > 1$ from (C.5)), and (d) from $\iota \le \log\left(\frac{(c_\beta+1)dKH}{\delta}\right)$ and from (C.6).

Now, joining the upper bounds for (I) and (II) in (C.8), we finally obtain

$$
\left\|\sum_{\tau=1}^{k-1} \phi_h^\tau [V_{h+1}^{i,k}(x_{h+1}^\tau) - \mathbb{P}_h V_{h+1}^{i,k}(x_h^\tau, a_h^\tau)]\right\|^2_{(\Lambda_h^k)^{-1}} \le (36 + 8\log(9) + 4\log(8))d^2H^2\log((c_\beta+1)dKH/\delta)
$$

which proves the claim and thus the proof. □

Proof of Lemma B.3. For any $(i,k) \in [n] \times [K]$,

$$
\begin{aligned}
w_h^{i,k} - w_h^{i,\bar\pi} &= (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi_h^\tau (r_h^\tau + V_{h+1}^{i,k}(x_{h+1}^\tau)) - w_h^{i,\bar\pi} \\
&\overset{\text{(a)}}{=} (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi_h^\tau ({\phi_h^\tau}^\top w_h^{i,\bar\pi} - \mathbb{P}_h V_{h+1}^{i,\bar\pi}(x_h^\tau, a_h^\tau) + V_{h+1}^{i,k}(x_{h+1}^\tau)) - w_h^{i,\bar\pi} \\
&= (\Lambda_h^k)^{-1} \left(\left(\sum_{\tau=1}^{k-1} \phi_h^\tau (\phi_h^\tau)^\top - \Lambda_h^k \right) w_h^{i,\bar\pi} \right. \\
&\quad \left. + \sum_{\tau=1}^{k-1} \phi_h^\tau \big(V_{h+1}^{i,k}(x_{h+1}^\tau) - \mathbb{P}_h V_{h+1}^{i,\bar\pi}(x_h^\tau, a_h^\tau)\big) \right) \\
&\overset{(b)}{=} (\Lambda_h^k)^{-1} \left(-\lambda w_h^{i,\bar\pi} + \sum_{\tau=1}^{k-1} \phi_h^\tau (V_{h+1}^{i,k}(x_{h+1}^\tau) - \mathbb{P}_h V_{h+1}^{i,\bar\pi}(x_h^\tau, a_h^\tau)) \right) \\
&= \underbrace{-\lambda(\Lambda_h^k)^{-1} w_h^{i,\bar\pi}}_{\text{(I)}} + \underbrace{(\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi_h^\tau (V_{h+1}^{i,k}(x_{h+1}^\tau) - \mathbb{P}_h V_{h+1}^{i,k}(x_h^\tau, a_h^\tau))}_{\text{(II)}} \\
&\quad + \underbrace{(\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi_h^\tau \mathbb{P}_h (V_{h+1}^{i,k} - V_{h+1}^{i,\bar\pi})(x_h^\tau, a_h^\tau)}_{\text{(III)}} .
\end{aligned}
$$

where (a) follows from the fact that, for any $(x,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, $Q_h^{i,\bar\pi}(x,a) := \langle \phi(x,a), w_h^{i,\bar\pi} \rangle = (r_h + \mathbb{P}_h V_{h+1}^{i,\bar\pi})(x,a)$ for some $w_h^{i,\bar\pi} \in \mathbb{R}^d$ (this follows from Proposition A.1 and the Bellman equation); and (b) follows from the definition of Λ_h^k. Since $\langle \phi(x,a), w_h^{i,k} \rangle - Q_h^{i,\bar\pi}(x,a) = \langle \phi(x,a), w_h^{i,k} - w_h^{i,\bar\pi} \rangle$ for any $(x,a) \in \mathcal{S} \times \mathcal{A}$, then we look to bound the inner product of each of the terms (I) – (III) with the term $\phi(x,a)$.

Regarding the term (I),

$$\begin{aligned}
|\langle \phi(x,a), \text{(I)} \rangle| &= |\langle \phi(x,a), \lambda (\Lambda_h^k)^{-1} w_h^{i,\bar{\pi}} \rangle| = |\lambda \langle (\Lambda_h^k)^{-1/2} \phi(x,a), (\Lambda_h^k)^{-1/2} w_h^{i,\bar{\pi}} \rangle| \\
&\leq \lambda \left\| w_h^{i,\bar{\pi}} \right\|_{(\Lambda_h^k)^{-1}} \sqrt{\phi(x,a)^\top (\Lambda_h^k)^{-1} \phi(x,a)} \leq \sqrt{\lambda} \left\| w_h^{i,\bar{\pi}} \right\| \sqrt{\phi(x,a)^\top (\Lambda_h^k)^{-1} \phi(x,a)}
\end{aligned}$$

where the last inequality follows from $\| \cdot \|_{(\Lambda_h^k)^{-1}} \leq \frac{1}{\sqrt{\lambda}} \| \cdot \|$.

For the term (II), since the event $\mathcal{E}_i$ from Lemma B.2 is given and $\lambda = 1$, we directly obtain

$$\begin{aligned}
|\langle \phi(x,a), \text{(II)} \rangle| &= \left| \left\langle \phi(x,a), (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi_h^\tau (V_{h+1}^{i,k}(x_{h+1}^\tau) - \mathbb{P}_h V_{h+1}^{i,k}(x_h^\tau, a_h^\tau)) \right\rangle \right| \\
&\leq \left\| \sum_{\tau=1}^{k-1} \phi_h^\tau (V_{h+1}^{i,k}(x_{h+1}^\tau) - \mathbb{P}_h V_{h+1}^{i,k}(x_h^\tau, a_h^\tau)) \right\|_{(\Lambda_h^k)^{-1}} \| \phi(x,a) \|_{(\Lambda_h^k)^{-1}} \\
&\leq C d H \sqrt{\log((c_\beta + 1) d K H / \delta)} \sqrt{\phi(x,a)^\top (\Lambda_h^k)^{-1} \phi(x,a)}
\end{aligned}$$

where C is an absolute constant independent of $c_\beta > 0$.

For the term (III),

$$\begin{aligned}
\langle \phi(x,a), \text{(III)} \rangle &= \left\langle \phi(x,a), (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi_h^\tau \mathbb{P}_h (V_{h+1}^{i,k} - V_{h+1}^{i,\bar{\pi}})(x_h^\tau, a_h^\tau) \right\rangle \\
&= \left\langle \phi(x,a), (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi_h^\tau (\phi_h^\tau)^\top \int_{\mathcal{S}} (V_{h+1}^{i,k} - V_{h+1}^{i,\bar{\pi}})(x') d\mu_h(x') \right\rangle \\
&\overset{\text{(a)}}{=} \underbrace{\left\langle \phi(x,a), \int_{\mathcal{S}} (V_{h+1}^{i,k} - V_{h+1}^{i,\bar{\pi}})(x') d\mu_h(x') \right\rangle}_{\text{(III.1)}} \\
&\quad \underbrace{- \lambda \left\langle \phi(x,a), (\Lambda_h^k)^{-1} \int_{\mathcal{S}} (V_{h+1}^{i,k} - V_{h+1}^{i,\bar{\pi}})(x') d\mu_h(x') \right\rangle}_{\text{(III.2)}}
\end{aligned}$$

where (a) follows from the definition of Λ_h^k. We immediately see from our assumption on linear stochastic game that (III.1) $= \mathbb{P}_h (V_{h+1}^{i,k} - V_{h+1}^{i,\bar{\pi}})(x,a)$ and

$$\begin{aligned}
|\text{(III.2)}| &\leq \lambda \left\| \int_{\mathcal{S}} (V_{h+1}^{i,k} - V_{h+1}^{i,\bar{\pi}})(x') d\mu_h(x') \right\|_{(\Lambda_h^k)^{-1}} \sqrt{\phi(x,a)^\top (\Lambda_h^k)^{-1} \phi(x,a)} \\
&\leq \sqrt{\lambda} \left\| \int_{\mathcal{S}} (V_{h+1}^{i,k} - V_{h+1}^{i,\bar{\pi}})(x') d\mu_h(x') \right\| \sqrt{\phi(x,a)^\top (\Lambda_h^k)^{-1} \phi(x,a)} \\
&\overset{\text{(a)}}{\leq} \sqrt{\lambda} 2H \int_{\mathcal{S}} \| \mu_h(x') \| \, dx' \sqrt{\phi(x,a)^\top (\Lambda_h^k)^{-1} \phi(x,a)} \overset{(b)}{\leq} 2H \sqrt{d\lambda} \sqrt{\phi(x,a)^\top (\Lambda_h^k)^{-1} \phi(x,a)}
\end{aligned}$$

where (a) follows from the value functions being bounded, and (b) from the definiton of the linear MDP.

Finally, putting it all together with $\lambda = 1$, we conclude that,

$$\begin{aligned}
&|\langle \phi(x,a), w_h^{i,k} \rangle - Q_h^{i,\bar{\pi}}(x,a) - \mathbb{P}_h (V_{h+1}^{i,k} - V_{h+1}^{i,\bar{\pi}})(x,a)| \\
&\qquad \leq \left(\left\| w_h^{i,\bar{\pi}} \right\| + C d H \sqrt{\log((c_\beta + 1) d K H / \delta)} + 2H\sqrt{d} \right) \sqrt{\phi(x,a)^\top (\Lambda_h^k)^{-1} \phi(x,a)} \\
&\qquad \leq \left(4H\sqrt{d} + C d H \sqrt{\log((c_\beta + 1) d K H / \delta)} \right) \sqrt{\phi(x,a)^\top (\Lambda_h^k)^{-1} \phi(x,a)}
\end{aligned}$$

where the last inequality follows from Proposition A.1.

Now, from equation (C.6) in Lemma B.2, we have $\sqrt{\log((c_\beta + 1)dKH/\delta)} > 1$ independently from $c_\beta > 0$, and thus

$$|\langle \phi(x,a), w_h^{i,k} \rangle - Q_h^{i,\bar{\pi}}(x,a) - \mathbb{P}_h(V_{h+1}^{i,k} - V_{h+1}^{i,\bar{\pi}})(x,a)| \leq \bar{C} dH \sqrt{\log((c_\beta + 1)dKH/\delta)} \sqrt{\phi(x,a)^\top (\Lambda_h^k)^{-1} \phi(x,a)},$$

for an absolute constant $\bar{C} = C + 4$ independent of c_β.

Finally, to prove this lemma, we only need to show that there exists a choice of the absolute positive constant c_β so that $\bar{C}\sqrt{\log((c_\beta + 1)dKH/\delta)} \leq c_\beta \sqrt{\iota}$, which is equivalent to

$$\bar{C}\sqrt{\iota + \log(c_\beta + 1)} \leq c_\beta \sqrt{\iota} \tag{C.9}$$

since $\sqrt{\log\left(\frac{(1+c_\beta)dKH}{\delta}\right)} = \sqrt{\log\left(\frac{dKH}{\delta}\right) + \log(1 + c_\beta)} = \sqrt{\iota + \log(1 + c_\beta)}$.

Two facts are known: 1) $\iota \in [\log(2), \infty)$ by its definition and $d \geq 2$; and 2) $\bar{C}$ is an absolute constant independent of c_β.

Since we know we are looking for $c_\beta > 0$ and using the bound $\log(x) \leq x - 1$ for any positive $x \in \mathbb{R}$, we conclude that proving the following equation implies (C.9),

$$\bar{C}\sqrt{\iota + c_\beta} \leq c_\beta \sqrt{\iota}. \tag{C.10}$$

After some algebraic calculations, we can show that

$$c_\beta \geq \frac{\bar{C}^2}{2\log(2)} + \frac{1}{2}\sqrt{\frac{\bar{C}^4}{(\log(2))^2} + 4\bar{C}^2} \tag{C.11}$$

suffices. This finishes the proof. □

References

Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2137–2143. PMLR, 09–12 Jul 2020.