A. Discussion of Function Space Complexity

To characterize the function space complexity, we first introduce the notions for the eigenvalues of the RKHS. Define $\mathcal{L}^2(\mathcal{Z})$ as the space of square-integrable functions on $\mathcal{Z}$ w.r.t. Lebesgue measure and define $\langle\cdot,\cdot\rangle_{\mathcal{L}^2}$ as the inner product on the space $\mathcal{L}^2(\mathcal{Z})$. According to Mercer's Theorem (Steinwart & Christmann, 2008), the kernel function $\ker(z,z')$ has a spectral expansion as $\ker(z,z') = \sum_{i=1}^{\infty}\sigma_i\varrho_i(z)\varrho_i(z')$ where $\{\varrho_i\}_{i\geq 1}$ are a set of orthonormal basis on $\mathcal{L}^2(\mathcal{Z})$ and $\{\sigma_i\}_{i\geq 1}$ are positive eigenvalues. In this paper, we consider two types of eigenvalues' properties and make the following assumptions.

Assumption A.1. *Assume $\{\sigma_i\}_{i\geq 1}$ satisfies one of the following eigenvalue decay conditions for some constant $\gamma > 0$:*

(a) γ-finite spectrum: we have $\sigma_i = 0$ for all $i > \gamma$;

(b) γ-exponential spectral decay: there exist constants $C_1 > 0$ and $C_2 > 0$ such that $\sigma_i \leq C_1\exp(-C_2\cdot i^{\gamma})$ for all $i \geq 1$.

Covering Numbers. Next, we characterize the upper bound of the covering numbers of the Q-function sets $\overline{\mathcal{Q}}(c,R,B)$ and $\underline{\mathcal{Q}}(c,R,B)$. For any $Q_1, Q_2 \in \overline{\mathcal{Q}}(c,R,B)$, we have

$$Q_1(z) = \min\Big\{c(z) + \Pi_{[0,H]}[\langle\mathbf{w}_1,\phi(z)\rangle] + B\cdot\max\{\|\phi(z)\|_{\Lambda_{\mathcal{D}_1}^{-1}}, H/\beta\}^+, H\Big\}^+,$$
$$Q_2(z) = \min\Big\{c(z) + \Pi_{[0,H]}[\langle\mathbf{w}_2,\phi(z)\rangle] + B\cdot\max\{\|\phi(z)\|_{\Lambda_{\mathcal{D}_2}^{-1}}, H/\beta\}^+, H\Big\}^+,$$

for some $\mathbf{w}_1, \mathbf{w}_2$ satisfying $\|\mathbf{w}_1\|_{\mathcal{H}} \leq R$ and $\|\mathbf{w}_2\|_{\mathcal{H}} \leq R$. Then, due to the fact that the truncation operator is non-expansive, we have

$$\|Q_1(\cdot) - Q_2(\cdot)\|_{\infty} \leq \sup_z|\langle\mathbf{w}_1 - \mathbf{w}_2, \phi(z)\rangle_{\mathcal{H}}| + B\sup_z\Big|\|\phi(z)\|_{\Lambda_{\mathcal{D}_1}^{-1}} - \|\phi(z)\|_{\Lambda_{\mathcal{D}_2}^{-1}}\Big|.$$

The above inequality shows that it suffices to bound the covering numbers of of the RKHS norm ball of radius R and the set of functions of the form $\|\phi(z)\|_{\Lambda_{\mathcal{D}}^{-1}}$. Thus, we define the function class $\mathcal{F}_{\lambda} := \{\|\phi(\cdot)\|_{\Upsilon} : \|\Upsilon\|_{\text{op}} \leq 1/\lambda\}$ since $\|\Lambda_{\mathcal{D}}^{-1}\|_{\text{op}} \leq 1/\lambda$ according to the definition of $\Lambda_{\mathcal{D}}$. Let $\overline{\mathcal{N}}_{\infty}(\epsilon;R,B)$ be the ϵ-covering number of $\overline{\mathcal{Q}}$ w.r.t. $\|\cdot\|_{\infty}$, $\mathcal{N}_{\infty}(\epsilon,\mathcal{H},R)$ be the ϵ-covering number of RKHS norm ball of radius R w.r.t. $\|\cdot\|_{\infty}$, and $\mathcal{N}_{\infty}(\epsilon,\mathcal{F},1/\lambda)$ be the ϵ-covering number of $\mathcal{F}_{\lambda}$ w.r.t. $\|\cdot\|_{\infty}$. Thus, we have

$$\overline{\mathcal{N}}_{\infty}(\epsilon;R,B) \leq \mathcal{N}_{\infty}(\epsilon/2,\mathcal{H},R)\cdot\mathcal{N}_{\infty}(\epsilon/(2B),\mathcal{F},1/\lambda).$$

We define the upper bound

$$\mathcal{N}_{\infty}(\epsilon;R,B) := \mathcal{N}_{\infty}(\epsilon/2,\mathcal{H},R)\cdot\mathcal{N}_{\infty}(\epsilon/(2B),\mathcal{F},1/\lambda),$$

in the main text of this paper. Then, we know

$$\log\mathcal{N}_{\infty}(\epsilon;R,B) = \log\mathcal{N}_{\infty}(\epsilon/2,\mathcal{H},R) + \log\mathcal{N}_{\infty}(\epsilon/(2B),\mathcal{F},1/\lambda).$$

Moreover, for any $Q_1, Q_2 \in \underline{\mathcal{Q}}(c,R,B)$, we have

$$Q_1(z) = \min\Big\{c(z) + \Pi_{[0,H]}[\langle\mathbf{w}_1,\phi(z)\rangle] - B\cdot\max\{\|\phi(z)\|_{\Lambda_{\mathcal{D}_1}^{-1}}, H/\beta\}^+, H\Big\}^+,$$
$$Q_2(z) = \min\Big\{c(z) + \Pi_{[0,H]}[\langle\mathbf{w}_2,\phi(z)\rangle] - B\cdot\max\{\|\phi(z)\|_{\Lambda_{\mathcal{D}_2}^{-1}}, H/\beta\}^+, H\Big\}^+,$$

which also implies

$$\|Q_1(\cdot) - Q_2(\cdot)\|_{\infty} \leq \sup_z|\langle\mathbf{w}_1 - \mathbf{w}_2, \phi(z)\rangle_{\mathcal{H}}| + B\sup_z\Big|\|\phi(z)\|_{\Lambda_{\mathcal{D}_1}^{-1}} - \|\phi(z)\|_{\Lambda_{\mathcal{D}_2}^{-1}}\Big|.$$

Thus, we can bound the covering number $\underline{\mathcal{N}}_{\infty}(\epsilon;R,B)$ of $\underline{\mathcal{Q}}(c,R,B)$ in the same way, i.e., $\underline{\mathcal{N}}_{\infty}(\epsilon;R,B) \leq \mathcal{N}_{\infty}(\epsilon;R,B)$.

According to Yang et al. (2020), we have the following covering number upper bounds

(a) γ-finite spectrum:

$$\log \mathcal{N}_\infty(\epsilon/2, \mathcal{H}, R) \le C_3\gamma[\log(2R/\epsilon) + C_4], \quad \log \mathcal{N}_\infty(\epsilon/(2B), \mathcal{F}, 1/\lambda) \le C_5\gamma^2[\log(2B/\epsilon) + C_6];$$

(b) γ-exponential spectral decay:

$$\log \mathcal{N}_\infty(\epsilon/2, \mathcal{H}, R) \le C_3[\log(2R/\epsilon) + C_4]^{1+1/\gamma}, \quad \log \mathcal{N}_\infty(\epsilon/(2B), \mathcal{F}, 1/\lambda) \le C_5[\log(2B/\epsilon) + C_6]^{1+2/\gamma}.$$

Maximal Information Gain. Here we give the definition of maximal information gain and discuss its upper bounds based on different kernels.

Definition A.2 (Maximal Information Gain (Srinivas et al., 2009))**.** *For any fixed integer $\mathfrak{C}$ and any $\sigma > 0$, we define the maximal information gain associated with the RKHS $\mathcal{H}$ as*

$$\Gamma(\mathfrak{C}, \lambda; \ker) = \sup_{\mathcal{D} \subseteq \mathcal{Z}} \frac{1}{2} \log\det(I + \mathcal{K}_\mathcal{D}/\lambda),$$

where the supremum is taken over all discrete subsets of $\mathcal{Z}$ with cardinality no more than $\mathfrak{C}$, and $\mathcal{K}_\mathcal{D}$ is the Gram matrix induced by $\mathcal{D} \subseteq \mathcal{Z}$ based on the kernel ker.

According to Theorem 5 in Srinivas et al. (2009), we have the maximal information gain characterized as follows

(a) γ-finite spectrum:

$$\Gamma(K, \lambda; \ker) \le C_7\gamma \log K;$$

(b) γ-exponential spectral decay:

$$\Gamma(K, \lambda; \ker) \le C_7(\log K)^{1+1/\gamma}.$$

Sample Complexity. Given the above results, for the kernel approximation setting, according to the discussion in the proof of Corollary 4.4 in Yang et al. (2020), under the parameter settings in Theorem 3.3 or Theorem 4.1, we have that for γ-finite spectrum setting,

$$\beta = \mathcal{O}(\gamma H\sqrt{\log(\gamma KH)}), \quad \log \mathcal{N}_\infty(\varsigma^*; R_K, 2\beta) = \mathcal{O}(\gamma^2 \log(\gamma KH)), \quad \Gamma(K, \lambda; \ker) = \mathcal{O}(\gamma \log K),$$

which implies after K episodes of exploration, the upper bound in Theorem 3.3 or Theorem 4.1 is

$$\mathcal{O}\left(\sqrt{H^6\gamma^3 \log^2(\gamma KH)/K}\right).$$

This result further implies that to obtain an ε-suboptimal policy or ε-approximate NE, it requires $\widetilde{\mathcal{O}}(H^6\gamma^3/\varepsilon^2)$ rounds of exploration. In addition, for the γ-exponential spectral decay setting, we have

$$\beta = \mathcal{O}(H\sqrt{\log(KH)}(\log K)^{1/\gamma}), \quad \log \mathcal{N}_\infty(\varsigma^*; R_K, 2\beta) = \mathcal{O}((\log K)^{1+2/\gamma} + (\log\log H)^{1+2/\gamma}),$$
$$\Gamma(K, \lambda; \ker) = \mathcal{O}((\log K)^{1+1/\gamma}),$$

which implies that after K episodes of exploration, the upper bound in Theorem 3.3 or Theorem 4.1 is

$$\mathcal{O}\left(\sqrt{H^6 \log^{2+3/\gamma}(KH)/K}\right).$$

Then, to obtain an ε-suboptimal policy or ε-approximate NE, it requires $\mathcal{O}(H^6 C_\gamma \log^{4+6/\gamma}(\varepsilon^{-1})/\varepsilon^2) = \widetilde{\mathcal{O}}(H^6 C_\gamma/\varepsilon^2)$ episodes of exploration, where C_γ is some constant depending on $1/\gamma$.

The above results also hold for the neural function approximation under both single-agent MDP and Markov game setting if the kernel $\ker_m$ satisfies the γ-finite spectrum or γ-exponential spectral decay and the network width m is sufficiently large such that the error term $H^2\beta\iota \le \varepsilon$. Then, we can similarly obtain the upper bounds in Theorems 3.5 and 4.2.

Linear and Tabular Cases. For the linear function approximation case, we have a feature map $\phi(s) \in \mathbb{R}^{\mathfrak{d}}$, where $\mathfrak{d}$ is the feature dimension. Therefore, the associated kernel can be represented as $\ker(s, s') = \phi(s)^\top \phi(s') = \sum_{i=1}^{\mathfrak{d}} \phi_i(s)\phi_i(s')$. Thus, we know that under the linear setting, the kernel $\ker$ has $\mathfrak{d}$-finite spectrum. Thus, letting $\gamma = \mathfrak{d}$ in the γ-finite spectrum case, we have

$$\beta = \mathcal{O}(\mathfrak{d}H\sqrt{\log(\mathfrak{d}KH)}), \quad \log \mathcal{N}_\infty(\varsigma^*; R_K, 2\beta) = \mathcal{O}(\mathfrak{d}^2 \log(\mathfrak{d}KH)), \quad \Gamma(K, \lambda; \ker) = \mathcal{O}(\mathfrak{d} \log K),$$

which further implies that to achieve $V_1^*(s_1, r) - V_1^\pi(s_1, r) \le \varepsilon$, it requires $\widetilde{\mathcal{O}}(H^6\mathfrak{d}^3/\varepsilon^2)$ rounds of exploration. This is consistent with the result in Wang et al. (2020a) for the single-agent MDP. This result also hold for the Markov game setting.

For the tabular case, since $\phi(z) = \boldsymbol{e}_z$ is the canonical basis in $\mathbb{R}^{|\mathcal{Z}|}$, we have $\gamma = |\mathcal{Z}|$ for the above γ-finite spectrum case. Therefore, for the single-agent MDP setting, we have $|\mathcal{Z}| = |\mathcal{S}||\mathcal{A}|$, which implies

$$\begin{aligned}&\beta = \mathcal{O}(H|\mathcal{S}||\mathcal{A}|\sqrt{\log(|\mathcal{S}||\mathcal{A}|KH)}), \quad \log \mathcal{N}_\infty(\varsigma^*; R_K, 2\beta) = \mathcal{O}(|\mathcal{S}|^2|\mathcal{A}|^2 \log(|\mathcal{S}||\mathcal{A}|KH)),\\ &\Gamma(K, \lambda; \ker) = \mathcal{O}(|\mathcal{S}||\mathcal{A}| \log K).\end{aligned}$$

Then, the sample complexity becomes $\widetilde{\mathcal{O}}(H^6|\mathcal{S}|^3|\mathcal{A}|^3/\varepsilon^2)$ to obtain an ε-suboptimal policy. For the two-player Markov game setting, we have $|\mathcal{Z}| = |\mathcal{S}||\mathcal{A}||\mathcal{B}|$, which implies

$$\begin{aligned}&\beta = \mathcal{O}(H|\mathcal{S}||\mathcal{A}||\mathcal{B}|\sqrt{\log(|\mathcal{S}||\mathcal{A}||\mathcal{B}|KH)}), \quad \log \mathcal{N}_\infty(\varsigma^*; R_K, 2\beta) = \mathcal{O}(|\mathcal{S}|^2|\mathcal{A}|^2|\mathcal{B}|^2 \log(|\mathcal{S}||\mathcal{A}||\mathcal{B}|KH)),\\ &\Gamma(K, \lambda; \ker) = \mathcal{O}(|\mathcal{S}||\mathcal{A}||\mathcal{B}| \log K).\end{aligned}$$

Then, the sample complexity becomes $\widetilde{\mathcal{O}}(H^6|\mathcal{S}|^3|\mathcal{A}|^3|\mathcal{B}|^3/\varepsilon^2)$ to obtain an ε-approximate NE.

B. Proofs for Single-Agent MDP Setting with Kernel Function Approximation

B.1. Lemmas

Lemma B.1 (Solution of Kernel Ridge Regression)**.** *The approximation vector $\widehat{f}_h^k \in \mathcal{H}$ is obtained by solving the following kernel ridge regression problem*

$$\underset{f \in \mathcal{H}}{\text{minimize}} \sum_{\tau=1}^{k-1} [V_{h+1}^k(s_{h+1}^\tau) - f(z_h^\tau)\rangle_{\mathcal{H}}]^2 + \lambda \|f\|_{\mathcal{H}}^2,$$

such that we have

$$\langle \phi(z), \widehat{f}_h^k(z)\rangle_{\mathcal{H}} = \psi_h^k(z)^\top (\lambda \cdot I + \mathcal{K}_h^k)^{-1} \mathbf{y}_h^k,$$

where we define

$$\begin{aligned}&\psi_h^k(z) := \Phi_h^k \phi(z) = [\ker(z, z_h^1), \cdots, \ker(z, z_h^{k-1})]^\top,\\ &\Phi_h^k = [\phi(z_h^1), \phi(z_h^2), \cdots, \phi(z_h^{k-1})]^\top,\\ &\mathbf{y}_h^k = [V_{h+1}^k(s_{h+1}^1), V_{h+1}^k(s_{h+1}^2), \cdots, V_{h+1}^k(s_{h+1}^{k-1})]^\top,\\ &\mathcal{K}_h^k := \Phi_h^k (\Phi_h^k)^\top = \begin{bmatrix} \ker(z_h^1, z_h^1) & \dots & \ker(z_h^1, z_h^{k-1}) \\ \vdots & \ddots & \vdots \\ \ker(z_h^{k-1}, z_h^1) & \dots & \ker(z_h^{k-1}, z_h^{k-1}) \end{bmatrix},\end{aligned} \tag{13}$$

with denoting $z = (s, a)$ and $z_h^\tau = (s_h^\tau, a_h^\tau)$, and $\ker(x, y) = \langle \phi(z), \phi(z')\rangle_{\mathcal{H}}, \forall z, z' \in \mathcal{Z} = \mathcal{S} \times \mathcal{A}$.

Proof. We seek to solve the following kernel ridge regression problem in the RKHS

$$\widehat{f}_h^k = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \sum_{\tau=1}^{k-1} [V_{h+1}^k(s_{h+1}^\tau) - f(s_h^\tau, a_h^\tau)\rangle_{\mathcal{H}}]^2 + \lambda \|f\|_{\mathcal{H}}^2,$$

which is equivalent to

$$\widehat{f}_h^k = \operatorname*{argmin}_{f\in\mathcal{H}} \sum_{\tau=1}^{k-1} [V_{h+1}^k(s_{h+1}^\tau) - \langle f, \phi(s_h^\tau, a_h^\tau)\rangle_{\mathcal{H}}]^2 + \lambda\langle f, f\rangle_{\mathcal{H}}.$$

By the first-order optimality condition, the above kernel ridge regression problem admits the following closed-form solution

$$\widehat{f}_h^k = (\Lambda_h^k)^{-1}(\Phi_h^k)^\top \mathbf{y}_h^k, \tag{14}$$

where we define

$$\Lambda_h^k = \sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top + \lambda \cdot I_{\mathcal{H}} = \lambda \cdot I_{\mathcal{H}} + (\Phi_h^k)^\top \Phi_h^k,$$

with $I_{\mathcal{H}}$ being the identity mapping in RKHS. Thus, by (14), we have

$$\langle \widehat{f}_h^k, \phi(z)\rangle_{\mathcal{H}} = \langle (\Lambda_h^k)^{-1}(\Phi_h^k)^\top \mathbf{y}_h^k, \phi(s,a)\rangle_{\mathcal{H}}, \quad \forall (z) \in \mathcal{S}\times\mathcal{A}, \tag{15}$$

which can be further rewritten in terms of kernel $\ker$ as follows

$$\begin{aligned}
\langle \widehat{f}_h^k, \phi(z)\rangle_{\mathcal{H}} &= \langle (\Lambda_h^k)^{-1}(\Phi_h^k)^\top \mathbf{y}_h^k, \phi(z)\rangle_{\mathcal{H}} \\
&= \phi(z)^\top [\lambda \cdot I_{\mathcal{H}} + (\Phi_h^k)^\top \Phi_h^k]^{-1} (\Phi_h^k)^\top \mathbf{y}_h^k \\
&= \phi(z)^\top (\Phi_h^k)^\top [\lambda \cdot I + \Phi_h^k (\Phi_h^k)^\top]^{-1} \mathbf{y}_h^k \\
&= \psi_h^k(z)^\top (\lambda \cdot I + \mathcal{K}_h^k)^{-1} \mathbf{y}_h^k.
\end{aligned} \tag{16}$$

The third equality is by

$$(\Phi_h^k)^\top [\lambda \cdot I + \Phi_h^k (\Phi_h^k)^\top] = [\lambda \cdot I_{\mathcal{H}} + (\Phi_h^k)^\top \Phi_h^k](\Phi_h^k)^\top,$$

such that

$$[\lambda \cdot I_{\mathcal{H}} + (\Phi_h^k)^\top \Phi_h^k]^{-1} (\Phi_h^k)^\top = (\Phi_h^k)^\top [\lambda \cdot I + \Phi_h^k (\Phi_h^k)^\top]^{-1}, \tag{17}$$

where I is an identity matrix in $\mathbb{R}^{(k-1)\times(k-1)}$. The last equality in (16) is by the definitions of $\psi_h^k(z)$ and $\mathcal{K}_h^k$ in (13). This completes the proof. □

Lemma B.2 (Boundedness of Solution)**.** *When $\lambda \geq 1$, for any $(k,h) \in [K]\times[H]$, $\widehat{f}_h^k$ defined in (14) satisfies*

$$\|\widehat{f}_h^k\|_{\mathcal{H}} \leq H\sqrt{2/\lambda \cdot \log\det(I + \mathcal{K}_h^k/\lambda)} \leq 2H\sqrt{\Gamma(K,\lambda;\ker)},$$

where $\mathcal{K}_h^k$ is defined in (13) and $\Gamma(K,\lambda;\ker)$ is defined in Definition A.2.

Proof. For any vector $f \in \mathcal{H}$, we have

$$|\langle f, \widehat{f}_h^k\rangle_{\mathcal{H}}| = |f^\top (\Lambda_h^k)^{-1}(\Phi_h^k)^\top \mathbf{y}_h^k| = \left| f^\top (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau) V_{h+1}^k(s_{h+1}^\tau) \right| \leq H \sum_{\tau=1}^{k-1} \left| f^\top (\Lambda_h^k)^{-1} \phi(s_h^\tau, a_h^\tau) \right|,$$

where the last inequality is due to $|V_{h+1}^k(s_{h+1}^\tau)| \leq H$. Then, with Lemma F.2, the rest of the proof is the same as the proof of Lemma C.5 in Yang et al. (2020), which finishes the proof. □

Lemma B.3. *With probability at least $1-\delta'$, we have $\forall (h,k) \in [H]\times[K]$,*

$$\begin{aligned}
&\left\| \sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau)[V_{h+1}^k(s_{h+1}^\tau) - \mathbb{P}_h V_{h+1}^k(s_h^\tau, a_h^\tau)] \right\|_{(\Lambda_h^k)^{-1}}^2 \\
&\qquad \leq 4H^2\Gamma(K,\lambda;\ker) + 10H^2 + 4H^2 \log \mathcal{N}_\infty(\varsigma^*; R_K, B_K) + 4H^2 \log(K/\delta'),
\end{aligned}$$

where we set $\varsigma^* = H/K$ and $\lambda = 1 + 1/K$.

Proof. We first define a value function class as follows

$$\overline{\mathcal{V}}(\mathbf{0}, R, B) = \{V : V(\cdot) = \max_{a\in\mathcal{A}} Q(\cdot, a) \text{ with } Q \in \overline{\mathcal{Q}}(\mathbf{0}, R, B)\},$$

where $\overline{\mathcal{Q}}$ is defined in (10). We denote the covering number of $\overline{\mathcal{V}}(\mathbf{0}, R, B)$ w.r.t. the distance dist as $\mathcal{N}^{\overline{\mathcal{V}}}_{\text{dist}}(\epsilon; R, B)$, where the distance dist is defined by $\text{dist}(V_1, V_2) = \sup_{s\in\mathcal{S}} |V_1(s) - V_2(s)|$. Specifically, for any $k \times h \in [K] \times [H]$, we assume that there exist constants R_K and B_K that depend on the number of episodes K such that any $V^k_h \in \overline{\mathcal{V}}(\mathbf{0}, R_K, B_K)$ with $R_K = 2H\sqrt{\Gamma(K, \lambda; \ker)}$ and $B_K = (1 + 1/H)\beta$ since $V^k_h(s) = (r^k_h + u^k_h + f^k_h)(z) = \Pi_{[0,H]}[\langle \widehat{f}^k_h, \phi(z)\rangle_{\mathcal{H}}] + (1 + 1/H)\beta \cdot \min\{\|\phi(z)\|_{(\Lambda^k_h)^{-1}}, H\}^+$ (See the next lemma for the reformulation of the bonus term). By Lemma F.1 with δ'/K, we have

$$\begin{aligned}
&\left\|\sum_{\tau=1}^{k-1} \phi(s^\tau_h, a^\tau_h)[V^k_{h+1}(s^\tau_{h+1}) - \mathbb{P}_h V^k_{h+1}(s^\tau_h, a^\tau_h)]\right\|^2_{(\Lambda^k_h)^{-1}} \\
&\quad\le \sup_{V\in\overline{\mathcal{V}}(\mathbf{0}, R_K, B_K)} \left\|\sum_{\tau=1}^{k-1} \phi(s^\tau_h, a^\tau_h)[V(s^\tau_{h+1}) - \mathbb{P}_h V(s^\tau_h, a^\tau_h)]\right\|^2_{(\Lambda^k_h)^{-1}} \\
&\quad\le 2H^2 \log\det(I + \mathcal{K}_k/\lambda) + 2H^2 k(\lambda - 1) + 4H^2 \log(K\mathcal{N}^{\overline{\mathcal{V}}}_{\text{dist}}(\epsilon; R_K, B_K)/\delta') + 8k^2\epsilon^2/\lambda \\
&\quad\le 4H^2\Gamma(K, \lambda; \ker) + 10H^2 + 4H^2 \log\mathcal{N}_\infty(\varsigma^*; R_K, B_K) + 4H^2 \log(K/\delta'),
\end{aligned}$$

where the last inequality is by setting $\lambda = 1 + 1/K$ and $\epsilon = \varsigma^* = H/K$. Moreover, the last inequality is also due to

$$\begin{aligned}
\text{dist}(V_1, V_2) &= \sup_{s\in\mathcal{S}} |V_1(s) - V_2(s)| = \sup_{s\in\mathcal{S}} \left|\max_{a\in\mathcal{A}} Q_1(s, a) - \max_{a\in\mathcal{A}} Q_2(s, a)\right| \\
&\le \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} |Q_1(s, a) - Q_2(s, a)| = \|Q_1 - Q_2\|_\infty,
\end{aligned}$$

which indicates that $\mathcal{N}^{\overline{\mathcal{V}}}_{\text{dist}}(\varsigma^*; R_K, B_K)$ upper bounded by the covering number of the class $\overline{\mathcal{Q}}$ w.r.t. $\|\cdot\|_\infty$, such that

$$\mathcal{N}^{\overline{\mathcal{V}}}_{\text{dist}}(\varsigma^*; R_K, B_K) \le \mathcal{N}_\infty(\varsigma^*; R_K, B_K).$$

Here $\mathcal{N}_\infty(\epsilon; R, B)$ denotes the upper bound of the covering number of $\overline{\mathcal{Q}}(h, R, B)$ w.r.t. ℓ_∞-norm, which is characterized in Section A. Further by union bound, we know that the above inequality holds for all $k \in [K]$ with probability at least $1 - \delta'$. This completes the proof. □

Lemma B.4. *We define the event $\mathcal{E}$ as that the following inequality holds $\forall z = (s, a) \in \mathcal{S} \times \mathcal{A}, \forall (h, k) \in [H] \times [K]$,*

$$|\mathbb{P}_h V^k_{h+1}(z) - f^k_h(z)| \le u^k_h(z),$$

where $f^k_h(z) = \min\{\widehat{f}^k_h(z), H\}^+$ and $u^k_h(z) = \min\{w^k_h(z), H\}^+$ with $w^k_h(z) = \beta\lambda^{-1/2}[\ker(z, z) - \psi^k_h(z)^\top(\lambda I + \mathcal{K}^k_h)^{-1}\psi^k_h(z)]^{1/2}$. Thus, setting $\beta = B_K/(1 + 1/H)$, if B_K satisfies

$$16H^2\big[R^2_Q + 2\Gamma(K, \lambda; \ker) + 5 + \log\mathcal{N}_\infty(\varsigma^*; R_K, B_K) + 2\log(K/\delta')\big] \le B^2_K, \forall h \in [H],$$

then we have that with probability at least $1 - \delta'$, the event $\mathcal{E}$ happens, i.e.,

$$\Pr(\mathcal{E}) \ge 1 - \delta'.$$

Proof. We assume that $\mathbb{P}_h V^k_{h+1}(s, a) = \langle \widetilde{f}^k_h, \phi(s, a)\rangle_{\mathcal{H}}$ for some $\widetilde{f}^k_h \in \mathcal{H}$. Then, we bound the difference between $f^k_h(z)$ and $\mathbb{P}_h V^k_{h+1}(s, a)$ in the following way

$$\begin{aligned}
&|\mathbb{P}_h V^k_{h+1}(s, a) - f^k_h(s, a)| \\
&\quad\le |\langle \widetilde{f}^k_h, \phi(s, a)\rangle_{\mathcal{H}} - \psi^k_h(s, a)^\top(\lambda \cdot I + \mathcal{K}^k_h)^{-1}\mathbf{y}^k_h| \\
&\quad= |\lambda\phi(s, a)^\top(\Lambda^k_h)^{-1}\widetilde{f}^k_h + \psi^k_h(s, a)^\top(\lambda \cdot I + \mathcal{K}^k_h)^{-1}\Phi^k_h\widetilde{f}^k_h - \psi^k_h(s, a)^\top(\lambda \cdot I + \mathcal{K}^k_h)^{-1}\mathbf{y}^k_h| \\
&\quad= |\lambda\phi(s, a)^\top(\Lambda^k_h)^{-1}\widetilde{f}^k_h + \psi^k_h(s, a)^\top(\lambda \cdot I + \mathcal{K}^k_h)^{-1}(\Phi^k_h\overline{f}^k_h - \mathbf{y}^k_h)|,
\end{aligned}$$

where the first inequality is due to non-expansiveness of the operator $\min\{\cdot, H\}^+$ and the definition of $\widehat{f}_h^k(z)$, and the first equality is due to

$$\begin{aligned}\phi(s,a) &= (\Lambda_h^k)^{-1}\Lambda_h^k\phi(s,a) = (\Lambda_h^k)^{-1}(\lambda\cdot I + (\Phi_h^k)^\top\Phi_h^k)\phi(s,a)\\ &= \lambda(\Lambda_h^k)^{-1}\phi(s,a) + (\Lambda_h^k)^{-1}(\Phi_h^k)^\top\Phi_h^k\phi(s,a)\\ &= \lambda(\Lambda_h^k)^{-1}\phi(s,a) + (\Phi_h^k)^\top(\lambda\cdot I + \mathcal{K}_h^k)^{-1}\Phi_h^k\phi(s,a)\\ &= \lambda(\Lambda_h^k)^{-1}\phi(s,a) + (\Phi_h^k)^\top(\lambda\cdot I + \mathcal{K}_h^k)^{-1}\psi_h^k(s,a).\end{aligned} \tag{18}$$

Thus, we have

$$|\mathbb{P}_h V_{h+1}^k(s,a,r^k) - f_h^k(s,a)| \le \underbrace{\lambda\|\phi(s,a)^\top(\Lambda_h^k)^{-1}\|_{\mathcal{H}}\cdot\|\widetilde{f}_h^k\|_{\mathcal{H}}}_{\text{Term(I)}} + \underbrace{|\psi_h^k(s,a)^\top(\lambda\cdot I + \mathcal{K}_h^k)^{-1}(\Phi_h^k\widetilde{f}_h^k - \mathbf{y}_h^k)|}_{\text{Term(II)}}. \tag{19}$$

For Term(I), we have

$$\begin{aligned}\text{Term(I)} &\le \sqrt{\lambda}R_Q H\sqrt{\phi(s,a)^\top(\Lambda_h^k)^{-1}\cdot\lambda I\cdot(\Lambda_h^k)^{-1}\phi(s,a)}\\ &\le \sqrt{\lambda}R_Q H\sqrt{\phi(s,a)^\top(\Lambda_h^k)^{-1}\cdot\Lambda_h^k\cdot(\Lambda_h^k)^{-1}\phi(s,a)}\\ &\le \sqrt{\lambda}R_Q H\sqrt{\phi(s,a)^\top(\Lambda_h^k)^{-1}\phi(s,a)} = \sqrt{\lambda}R_Q H\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}},\end{aligned} \tag{20}$$

where the first inequality is due to Assumption 3.2 and the second inequality is by $\theta^\top(\Phi_h^k)^\top\Phi_h^k\theta = \|\Phi_h^k\theta\|_{\mathcal{H}} \ge 0$ for any $\theta\in\mathcal{H}$.

For Term(II), we have

$$\begin{aligned}\text{Term(II)} &= \left|\phi(s,a)^\top(\Lambda_h^k)^{-1}\left\{\sum_{\tau=1}^{k-1}\phi(s_h^\tau,a_h^\tau)[V_{h+1}^k(s_{h+1}^\tau) - \mathbb{P}_h V_{h+1}^k(s_h^\tau,a_h^\tau)]\right\}\right|\\ &= \left|\phi(s,a)^\top(\Lambda_h^k)^{-1/2}(\Lambda_h^k)^{-1/2}\left\{\sum_{\tau=1}^{k-1}\phi(s_h^\tau,a_h^\tau)[V_{h+1}^k(s_{h+1}^\tau) - \mathbb{P}_h V_{h+1}^k(s_h^\tau,a_h^\tau)]\right\}\right|\\ &\le \|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}\left\|\sum_{\tau=1}^{k-1}\phi(s_h^\tau,a_h^\tau)[V_{h+1}^k(s_{h+1}^\tau) - \mathbb{P}_h V_{h+1}^k(s_h^\tau,a_h^\tau)]\right\|_{(\Lambda_h^k)^{-1}}\end{aligned} \tag{21}$$

By Lemma B.3, we have that with probability at least $1-\delta'$, the following inequality holds for all $k\in[K]$

$$\begin{aligned}&\left\|\sum_{\tau=1}^{k-1}\phi(s_h^\tau,a_h^\tau)[V_{h+1}^k(s_{h+1}^\tau) - \mathbb{P}_h V_{h+1}^k(s_h^\tau,a_h^\tau)]\right\|_{(\Lambda_h^k)^{-1}}\\ &\quad\le [4H^2\Gamma(K,\lambda;\ker) + 10H^2 + 4H^2\log\mathcal{N}_\infty(\varsigma^*;R_K,B_K) + 4H^2\log(K/\delta')]^{1/2}.\end{aligned}$$

Thus, Term(II) can be further bounded as

$$\text{Term(II)} \le \left[4H^2\Gamma(K,\lambda;\ker) + 10H^2 + 4H^2\log\mathcal{N}_\infty(\varsigma^*;R_K,B_K) + 4H^2\log(K/\delta')\right]^{1/2}\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}.$$

Plugging the upper bounds of Term(I) and Term(II) into (19), we obtain

$$\begin{aligned}&|\mathbb{P}_h V_{h+1}^k(s,a,r^k) - f_h^k(s,a)|\\ &\quad\le \left[\sqrt{\lambda}R_Q H + [4H^2\Gamma(K,\lambda;\ker) + 10H^2 + 4H^2\log\mathcal{N}_\infty(\varsigma^*;R_K,B_K) + 4H^2\log(K/\delta')]^{1/2}\right]\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}\\ &\quad\le \left[2\lambda R_Q^2H^2 + 8H^2\Gamma(K,\lambda;\ker) + 20H^2 + 4H^2\log\mathcal{N}_\infty(\varsigma^*;R_K,B_K) + 8H^2\log(K/\delta')\right]^{1/2}\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}\\ &\quad\le \beta\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}} = \beta\lambda^{-1/2}[\ker(z,z) - \psi_h^k(s,a)^\top(\lambda I + \mathcal{K}_h^k)^{-1}\psi_h^k(s,a)]^{1/2},\end{aligned}$$

where $\varsigma^* = H/K$, and $\lambda = 1 + 1/K$ as in Lemma B.3. In the last equality, we also use the identity that

$$\begin{aligned}\|\phi(s,a)\|^2_{(\Lambda_h^k)^{-1}} &= \lambda^{-1}\phi(s,a)^\top\phi(s,a) - \lambda^{-1}\psi_h^k(s,a)^\top(\lambda\cdot I + \mathcal{K}_h^k)^{-1}\psi_h^k(s,a) \\ &= \lambda^{-1}\ker(z,z) - \lambda^{-1}\psi_h^k(s,a)^\top(\lambda I + \mathcal{K}_h^k)^{-1}\psi_h^k(s,a).\end{aligned} \tag{22}$$

This is proved by

$$\begin{aligned}\|\phi(s,a)\|^2_{\mathcal{H}} &= \phi(s,a)^\top[\lambda(\Lambda_h^k)^{-1}\phi(s,a) + (\Phi_h^k)^\top(\lambda\cdot I + \mathcal{K}_h^k)^{-1}\Phi_h^k\phi(s,a)] \\ &= \lambda\phi(s,a)^\top(\Lambda_h^k)^{-1}\phi(s,a) + \psi_h^k(s,a)^\top(\lambda\cdot I + \mathcal{K}_h^k)^{-1}\psi_h^k(s,a),\end{aligned}$$

where the first equality is by (18).

According to Lemma B.2, we know that $\widehat{f}_h^k$ satisfies $\|\widehat{f}_h^k\|_{\mathcal{H}} \le H\sqrt{2/\lambda\cdot\log\det(I+\mathcal{K}_h^k/\lambda)} \le 2H\sqrt{\Gamma(K,\lambda;\ker)}$. Then, one can set $R_K = 2H\sqrt{\Gamma(K,\lambda;\ker)}$. Moreover, as we set $(1+1/H)\beta = B_K$, then $\beta = B_K/(1+1/H)$. Thus, we let

$$\big[2\lambda R_Q^2H^2 + 8H^2\Gamma(K,\lambda;\ker) + 20H^2 + 4H^2\log\mathcal{N}_\infty(\varsigma^*;R_K,B_K) + 8H^2\log(K/\delta')\big]^{1/2} \le \beta = B_K/(1+1/H),$$

which can be further guaranteed by

$$16H^2\big[R_Q^2 + 2\Gamma(K,\lambda;\ker) + 5 + \log\mathcal{N}_\infty(\varsigma^*;R_K,B_K) + 2\log(K/\delta')\big] \le B_K^2$$

as $(1+1/H)\le 2$ and $\lambda = 1+1/K \le 2$.

According to the above result, letting $w_h^k = \beta\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}} = \beta\lambda^{-1/2}[\ker(z,z) - \psi_h^k(s,a)^\top(\lambda I + \mathcal{K}_h^k)^{-1}\psi_h^k(s,a)]^{1/2}$, we have $-w_h^k \le \mathbb{P}_hV_{h+1}^k(s,a,r^k) - f_h^k(s,a) \le w_h^k$. Note that we also have $|\mathbb{P}_hV_{h+1}^k(s,a,r^k) - f_h^k(s,a)| \le H$ due to $0 \le f_h^k(s,a) \le H$ and $0 \le \mathbb{P}_hV_{h+1}^k(s,a,r^k) \le H$. Thus, there is $|\mathbb{P}_hV_{h+1}^k(s,a,r^k) - f_h^k(s,a)| \le \min\{w_h^k, H\}$. This completes the proof. □

Lemma B.5. *Conditioned on the event $\mathcal{E}$ defined in Lemma B.4, with probability at least $1-\delta'$, we have*

$$\sum_{k=1}^K V_1^*(s_1,r^k) \le \sum_{k=1}^K V_1^k(s_1) \le \mathcal{O}\left(\sqrt{H^3K\log(1/\delta')} + \beta\sqrt{H^2K\cdot\Gamma(K,\lambda;\ker)}\right).$$

Proof. We first show the first inequality in this lemma, i.e., $\sum_{k=1}^K V_1^*(s_1,r^k) \le \sum_{k=1}^K V_1^k(s_1)$. To show this inequality holds, it suffices to show $V_h^*(s,r^k) \le V_h^k(s)$ for all $s\in\mathcal{S}, h\in[H]$. We prove it by induction.

When $h = H+1$, we know $V_{H+1}^*(s,r^k) = 0$ and $V_{H+1}^k(s) = 0$ such that $V_{H+1}^*(s,r^k) = V_{H+1}^k(s_1)$. Now we assume that $V_{h+1}^*(s,r^k) \le V_{h+1}^k(s)$. Then, conditioned on the event $\mathcal{E}$ defined in Lemma B.4, for all $s\in\mathcal{S}$, $(h,k)\in[H]\times[K]$, we further have

$$\begin{aligned}&Q_h^*(s,a,r^k) - Q_h^k(s,a) \\ &\quad= r_h^k(s,a) + \mathbb{P}_hV_{h+1}^*(s,a,r^k) - \min\{[r_h^k(s,a) + f_h^k(s,a) + u_h^k(s,a)], H\}^+ \\ &\quad\le \min\{[\mathbb{P}_hV_{h+1}^*(s,a,r^k) - f_h^k(s,a) - u_h^k(s,a)], 0\} \\ &\quad\le \min\{[\mathbb{P}_hV_{h+1}^k(s,a) - f_h^k(s,a) - u_h^k(s,a)], 0\} \\ &\quad\le 0\end{aligned} \tag{23}$$

where the first inequality is due to $0 \le r_h^k(s,a) + \mathbb{P}_hV_{h+1}^*(s,a,r^k) \le H$ and $\min\{x,y\}^+ \ge \min\{x,y\}$, the second inequality is by the assumption that $V_{h+1}^*(s,r^k) \le V_{h+1}^k(s)$, the last inequality is by Lemma B.4 such that $\mathbb{P}_hV_{h+1}^k(s,a) - f_h^k(s,a) \le u_h^k(s,a)$ holds for any $(s,a)\in\mathcal{S}\times\mathcal{A}$ and $(k,h)\in[K]\times[H]$. The above inequality (23) further leads to

$$V_h^*(s,r^k) = \max_{a\in\mathcal{A}} Q_h^*(s,a,r^k) \le \max_{a\in\mathcal{A}} Q_h^k(s,a) = V_h^k(s).$$

Therefore, we obtain that conditioned on event $\mathcal{E}$, we have

$$\sum_{k=1}^K V_1^*(s,r^k) \le \sum_{k=1}^K V_1^k(s).$$

Next, we prove the second inequality in this lemma, namely the upper bound of $\sum_{k=1}^{K} V_1^k(s_1)$. Specifically, conditioned on $\mathcal{E}$ defined in Lemma B.4, we have

$$\begin{aligned}
V_h^k(s_h^k) = Q_h^k(s_h^k, a_h^k) &\leq f_h^k(s_h^k, a_h^k) + r_h^k(s_h^k, a_h^k) + u_h^k(s_h^k, a_h^k) \\
&\leq \mathbb{P}_h V_{h+1}^k(s_h^k, a_h^k) + u_h^k(s_h^k, a_h^k) + r_h^k(s_h^k, a_h^k) + u_h^k(s_h^k, a_h^k) \\
&\leq \mathbb{P}_h V_{h+1}^k(s_h^k, a_h^k) + (2 + 1/H) w_h^k \\
&= \zeta_h^k + V_{h+1}^k(s_{h+1}^k) + (2 + 1/H)\beta \|\phi(s_h^k, a_h^k)\|_{(\Lambda_h^k)^{-1}},
\end{aligned}$$

where the second inequality is due to Lemma B.4 and in the last equality, we define

$$\zeta_h^k := \mathbb{P}_h V_{h+1}^k(s_h^k, a_h^k) - V_{h+1}^k(s_{h+1}^k).$$

Recursively applying the above inequality gives

$$V_1^k(s_1) \leq \sum_{h=1}^{H} \zeta_h^k + (2 + 1/H)\beta \sum_{h=1}^{H} \|\phi(s_h^k, a_h^k)\|_{(\Lambda_h^k)^{-1}},$$

where we use the fact that $V_{H+1}^k(\cdot) = 0$. Taking summation on both sides of the above inequality, we have

$$\sum_{k=1}^{K} V_1^k(s_1) = \sum_{k=1}^{K} \sum_{h=1}^{H} \zeta_h^k + (2 + 1/H)\beta \sum_{k=1}^{K} \sum_{h=1}^{H} \|\phi(s_h^k, a_h^k)\|_{(\Lambda_h^k)^{-1}}.$$

By Azuma-Hoeffding inequality, with probability at least $1 - \delta'$, the following inequalities hold

$$\sum_{k=1}^{K} \sum_{h=1}^{H} \zeta_h^k \leq \mathcal{O}\left(\sqrt{H^3 K \log \frac{1}{\delta'}}\right).$$

On the other hand, by Lemma F.2, we have

$$\begin{aligned}
\sum_{k=1}^{K} \sum_{h=1}^{H} \|\phi(s_h^k, a_h^k)\|_{(\Lambda_h^k)^{-1}} &= \sum_{k=1}^{K} \sum_{h=1}^{H} \sqrt{\phi(s_h^k, a_h^k)^\top (\Lambda_h^k)^{-1} \phi(s_h^k, a_h^k)} \\
&\leq \sum_{h=1}^{H} \sqrt{K \sum_{k=1}^{K} \phi(s_h^k, a_h^k)^\top (\Lambda_h^k)^{-1} \phi(s_h^k, a_h^k)} \\
&\leq \sum_{h=1}^{H} \sqrt{2K \log \det(I + \lambda \mathcal{K}_h^K)} = 2H\sqrt{K \cdot \Gamma(K, \lambda; \ker)}.
\end{aligned}$$

where the first inequality is by Jensen's inequality. Thus, conditioned on event $\mathcal{E}$, we obtain that with probability at least $1 - \delta'$, there is

$$\sum_{k=1}^{K} V_1^k(s_1) \leq \mathcal{O}\left(\sqrt{H^3 K \log(1/\delta')} + \beta\sqrt{H^2 K \cdot \Gamma(K, \lambda; \ker)}\right),$$

which completes the proof. □

Lemma B.6. *We define the event $\widetilde{\mathcal{E}}$ as that the following inequality holds $\forall z = (s, a) \in \mathcal{S} \times \mathcal{A}, \forall h \in [H]$,*

$$|\mathbb{P}_h V_{h+1}(z) - f_h(z)| \leq u_h(z),$$

where $u_h(z) = \min\{w_h(z), H\}^+$ with $w_h(z) = \beta\lambda^{-1/2}[\ker(z, z) - \psi_h(z)^\top (\lambda I + \mathcal{K}_h)^{-1} \psi_h(z)]^{1/2}$. Thus, setting $\beta = \widetilde{B}_K$, if $\widetilde{B}_K$ satisfies

$$4H^2\big[R_Q^2 + 2\Gamma(K, \lambda; \ker) + 5 + \log \mathcal{N}_\infty(\varsigma^*; \widetilde{R}_K, \widetilde{B}_K) + 2\log(K/\delta')\big] \leq \widetilde{B}_K^2, \forall h \in [H],$$

then we have that with probability at least $1 - \delta'$, the event $\mathcal{E}$ happens, i.e.,

$$\Pr(\widetilde{\mathcal{E}}) \geq 1 - \delta'.$$

Proof. The proof of this lemma is nearly the same as the proof of Lemma B.4. We provide the sketch of this proof below. We assume that the true transition is formulated as $\mathbb{P}_h V_{h+1}(z) = \langle \widetilde{f}_h, \phi(z) \rangle_{\mathcal{H}} =: \widetilde{f}_h(z)$. We have the following definitions

$$\Phi_h = [\phi(s_h^1, a_h^1), \phi(s_h^2, a_h^2), \cdots, \phi(s_h^K, a_h^K)]^\top, \quad \Lambda_h = \sum_{\tau=1}^{K} \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top + \lambda \cdot I_{\mathcal{H}} = \lambda \cdot I_{\mathcal{H}} + (\Phi_h)^\top \Phi_h,$$
$$\mathbf{y}_h = [V_{h+1}(s_{h+1}^1), V_{h+1}(s_{h+1}^2), \cdots, V_{h+1}(s_{h+1}^K)]^\top, \quad \mathcal{K}_h = \Phi_h \Phi_h^\top, \quad \psi_h(s,a) = \Phi_h \phi(s,a).$$

Then, we bound the following term

$$\begin{aligned}
&\mathbb{P}_h V_{h+1}(s,a) - f_h(s,a) \\
&\quad \le |\langle \widetilde{f}_h, \phi(s,a) \rangle_{\mathcal{H}} - \psi_h(s,a)^\top (\lambda \cdot I + \mathcal{K}_h)^{-1} \mathbf{y}_h| \\
&\quad = |\lambda \phi(s,a)^\top \Lambda_h^{-1} \widetilde{f}_h + \psi_h(s,a)^\top (\lambda \cdot I + \mathcal{K}_h)^{-1} \Phi_h \widetilde{f}_h - \psi_h(s,a)^\top (\lambda \cdot I + \mathcal{K}_h)^{-1} \mathbf{y}_h| \\
&\quad = |\lambda \phi(s,a)^\top \Lambda_h^{-1} \widetilde{f}_h + \psi_h^k(s,a)^\top (\lambda \cdot I + \mathcal{K}_h)^{-1} (\Phi_h \widetilde{f}_h - \mathbf{y}_h)|,
\end{aligned}$$

where the first equality is by the same reformulation as (18) such that

$$\phi(s,a) = \lambda \Lambda_h^{-1} \phi(s,a) + (\Phi_h)^\top (\lambda \cdot I + \mathcal{K}_h)^{-1} \psi_h(s,a).$$

Thus, we have

$$|\mathbb{P}_h V_{h+1}(s,a) - f_h(s,a)| \le \underbrace{\lambda \|\phi(s,a)^\top \Lambda_h^{-1}\|_{\mathcal{H}} \cdot \|\widetilde{f}_h\|_{\mathcal{H}}}_{\text{Term(I)}} + \underbrace{|\psi_h(s,a)^\top (\lambda \cdot I + \mathcal{K}_h)^{-1} (\Phi_h \widetilde{f}_h - \mathbf{y}_h)|}_{\text{Term(II)}}. \tag{24}$$

Analogous to (20), for Term(I) here, we have

$$\text{Term(I)} \le \sqrt{\lambda} R_Q H \|\phi(s,a)\|_{\Lambda_h^{-1}}.$$

Similar to (21), for Term(II), we have

$$\text{Term(II)} \le \|\phi(s,a)\|_{\Lambda_h^{-1}} \left\| \sum_{\tau=1}^{K} \phi(s_h^\tau, a_h^\tau)[V_{h+1}(s_{h+1}^\tau) - \mathbb{P}_h V_{h+1}(s_h^\tau, a_h^\tau)] \right\|_{\Lambda_h^{-1}}.$$

Then, we need to bound the last factor in the above inequality. Here we apply the similar argument as Lemma B.3. We have the function class for V_h is

$$\overline{\mathcal{V}}(r_h, \widetilde{R}_K, \widetilde{B}_K) = \{V : V(\cdot) = \max_{a \in \mathcal{A}} Q(\cdot, a) \text{ with } Q \in \overline{\mathcal{Q}}(r_h, \widetilde{R}_K, \widetilde{B}_K)\}.$$

By Lemma F.1 with δ', we have

$$\begin{aligned}
&\left\| \sum_{\tau=1}^{K} \phi(s_h^\tau, a_h^\tau)[V_{h+1}(s_{h+1}^\tau) - \mathbb{P}_h V_{h+1}(s_h^\tau, a_h^\tau)] \right\|_{(\Lambda_h)^{-1}}^2 \\
&\quad \le \sup_{V \in \overline{\mathcal{V}}(r_h, \widetilde{R}_K, \widetilde{B}_K)} \left\| \sum_{\tau=1}^{K} \phi(s_h^\tau, a_h^\tau)[V(s_{h+1}^\tau) - \mathbb{P}_h V(s_h^\tau, a_h^\tau)] \right\|_{(\Lambda_h)^{-1}}^2 \\
&\quad \le 2H^2 \log\det(I + \mathcal{K}/\lambda) + 2H^2 K(\lambda - 1) + 4H^2 \log(\mathcal{N}_{\text{dist}}^{\overline{\mathcal{V}}}(\epsilon; \widetilde{R}_K, \widetilde{B}_K)/\delta') + 8K^2\epsilon^2/\lambda \\
&\quad \le 4H^2 \Gamma(K, \lambda; \ker) + 10H^2 + 4H^2 \log \mathcal{N}_\infty(\varsigma^*; \widetilde{R}_K, \widetilde{B}_K) + 4H^2 \log(1/\delta'),
\end{aligned}$$

where the last inequality is by setting $\lambda = 1 + 1/K$ and $\epsilon = \varsigma^* = H/K$, and also due to

$$\mathcal{N}_{\text{dist}}^{\overline{\mathcal{V}}}(\varsigma^*; \widetilde{R}_K, \widetilde{B}_K) \le \mathcal{N}_\infty(\varsigma^*; \widetilde{R}_K, \widetilde{B}_K).$$

We have that with probability at least $1-\delta'$, the following inequality holds for all $k \in [K]$

$$\begin{aligned}&\bigg\|\sum_{\tau=1}^{K}\phi(s_h^\tau,a_h^\tau)[V_{h+1}(s_{h+1}^\tau)-\mathbb{P}_hV_{h+1}^k(s_h^\tau,a_h^\tau)]\bigg\|_{\Lambda_h^{-1}}\\&\quad\le[4H^2\Gamma(K,\lambda;\ker)+10H^2+4H^2\log\mathcal{N}_\infty(\varsigma^*;\widetilde{R}_K,\widetilde{B}_K)+4H^2\log(K/\delta')]^{1/2}.\end{aligned}$$

Thus, Term(II) can be further bounded as

$$\text{Term(II)}\le\big[4H^2\Gamma(K,\lambda;\ker)+10H^2+4H^2\log\mathcal{N}_\infty(\varsigma^*;\widetilde{R}_K,\widetilde{B}_K)+4H^2\log(K/\delta')\big]^{1/2}\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}.$$

Plugging the upper bounds of Term(I) and Term(II) into (24), we obtain

$$\begin{aligned}&|\mathbb{P}_hV_{h+1}(s,a)-f_h(s,a)|\\&\quad\le u_h(s,a)\le\beta\|\phi(s,a)\|_{\Lambda_h^{-1}}=\beta\lambda^{-1/2}[\ker(z,z)-\psi_h(s,a)^\top(\lambda I+\mathcal{K}_h)^{-1}\psi_h(s,a)]^{1/2},\end{aligned}$$

where we let $z=(s,a)$, $\varsigma^*=H/K$, and $\lambda=1+1/K$. In the last equality, similar to (22), we have

$$\begin{aligned}\|\phi(s,a)\|_{\Lambda_h^{-1}}^2&=\lambda^{-1}\phi(s,a)^\top\phi(s,a)-\lambda^{-1}\phi(s,a)^\top(\Phi_h)^\top[\lambda I+\Phi_h(\Phi_h)^\top]^{-1}\Phi_h\phi(s,a)\\&=\lambda^{-1}\ker(z,z)-\lambda^{-1}\psi_h(s,a)^\top(\lambda I+\mathcal{K}_h)^{-1}\psi_h(s,a).\end{aligned}\tag{25}$$

Similar to Lemma B.2, we know that the estimated function f_h satisfies $\|f_h\|_{\mathcal{H}}\le H\sqrt{2/\lambda\cdot\log\det(I+\mathcal{K}_h^k/\lambda)}\le 2H\sqrt{\Gamma(K,\lambda;\ker)}$. Then, one can set $\widetilde{R}_K=2H\sqrt{\Gamma(K,\lambda;\ker)}$. Moreover, as we set $\beta=\widetilde{B}_K$. Thus, we let

$$\big[2\lambda R_Q^2H^2+8H^2\Gamma(K,\lambda;\ker)+20H^2+4H^2\log\mathcal{N}_\infty(\varsigma^*;\widetilde{R}_K,\widetilde{B}_K)+8H^2\log(K/\delta')\big]^{1/2}\le\beta=\widetilde{B}_K,$$

which can be further guaranteed by

$$4H^2\big[R_Q^2+2\Gamma(K,\lambda;\ker)+5+\log\mathcal{N}_\infty(\varsigma^*;\widetilde{R}_K,\widetilde{B}_K)+2\log(K/\delta')\big]\le\widetilde{B}_K^2$$

as $(1+1/H)\le 2$ and $\lambda=1+1/K\le 2$. This completes the proof. □

Lemma B.7. *Conditioned on the event $\widetilde{\mathcal{E}}$ as defined in Lemma B.6, we have*

$$V_h^*(s,r)\le V_h(s)\le r_h(s,\pi_h(s))+\mathbb{P}_hV_{h+1}(s,\pi_h(s))+2u_h(s,\pi_h(s)),\forall s\in\mathcal{S},\forall h\in[H],$$

where $\pi_h(s)=\operatorname{argmax}_{a\in\mathcal{A}}Q_h(s,a)$.

Proof. We first prove the first inequality in this lemma. We prove it by induction. For $h=H+1$, by the planning algorithm, we have $V_{H+1}^*(s,r)=V_{H+1}(s)=0$ for any $s\in\mathcal{S}$. Then, we assume that $V_{h+1}^*(s,r)\le V_{h+1}(s)$. Thus, conditioned on the event $\widetilde{\mathcal{E}}$ as defined in Lemma B.6, we have

$$\begin{aligned}&Q_h^*(s,a,r)-Q_h(s,a)\\&\quad=r_h(s,a)+\mathbb{P}_hV_{h+1}^*(s,a,r)-\min\{[r_h(s,a)+f_h(s,a)+u_h(s,a)],H\}^+\\&\quad\le\min\{[\mathbb{P}_hV_{h+1}^*(s,a,r)-f_h(s,a)-u_h(s,a)],0\}\\&\quad\le\min\{[\mathbb{P}_hV_{h+1}(s,a)-f_h(s,a)-u_h(s,a)],0\}\\&\quad\le 0\end{aligned}$$

where the first inequality is due to $0\le r_h(s,a)+\mathbb{P}_hV_{h+1}^*(s,a,r)\le H$ and $\min\{x,y\}^+\ge\min\{x,y\}$, the second inequality is by the assumption that $V_{h+1}^*(s,a,r)\le V_{h+1}(s,a)$, the last inequality is by Lemma B.6 such that $|\mathbb{P}_hV_{h+1}(s,a)-f_h(s,a)|\le u_h(s,a)$ holds for any $(s,a)\in\mathcal{S}\times\mathcal{A}$ and $(k,h)\in[K]\times[H]$. The above inequality further leads to

$$V_h^*(s,r)=\max_{a\in\mathcal{A}}Q_h^*(s,a,r)\le\max_{a\in\mathcal{A}}Q_h(s,a)=V_h(s).$$

Therefore, we have

$$V_h^*(s,r) \le V_h(s), \forall h \in [H], \forall s \in \mathcal{S}.$$

In addition, we prove the second inequality in this lemma. We have

$$\begin{aligned} Q_h(s,a) &= \min\{[r_h(s,a) + f_h(s,a) + u_h(s,a)], H\}^+ \\ &\le \min\{[r_h(s,a) + \mathbb{P}_h V_{h+1}(s,a) + 2u_h(s,a)], H\}^+ \\ &\le r_h(s,a) + \mathbb{P}_h V_{h+1}(s,a) + 2u_h(s,a), \end{aligned}$$

where the first inequality is also by Lemma B.6 such that $|\mathbb{P}_h V_{h+1}(s,a) - f_h(s,a)| \le u_h(s,a)$, and the last inequality is because of the non-negativity of $r_h(s,a) + \mathbb{P}_h V_{h+1}(s,a) + 2u_h(s,a)$. Therefore, we have

$$V_h(s) = \max_{a \in \mathcal{A}} Q_h(s,a) = Q_h(s, \pi_h(s)) \le r_h(s, \pi_h(s)) + \mathbb{P}_h V_{h+1}(s, \pi_h(s)) + 2u_h(s, \pi_h(s)).$$

This completes the proof. □

Lemma B.8. *With the exploration and planning phases, we have the following inequality*

$$K \cdot V_1^*(s_1, u/H) \le \sum_{k=1}^K V_1^*(s_1, r^k).$$

Proof. As shown in (25), we know that

$$w_h(s,a) = \beta \|\phi(s,a)\|_{\Lambda_h^{-1}} = \beta \sqrt{\phi(s,a)^\top \left[\lambda I_{\mathcal{H}} + \sum_{\tau=1}^K \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top\right]^{-1} \phi(s,a)}.$$

On the other hand, by (22), we similarly have

$$w_h^k(s,a) = \beta \|\phi(s,a)\|_{(\Lambda_h^k)^{-1}} = \beta \sqrt{\phi(s,a)^\top \left[\lambda I_{\mathcal{H}} + \sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top\right]^{-1} \phi(s,a)}.$$

Since $k-1 \le K$ and $f^\top \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top f = [f^\top \phi(s_h^\tau, a_h^\tau)]^2 \ge 0$, then we know that

$$\Lambda_h = \lambda I_{\mathcal{H}} + \sum_{\tau=1}^K \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top \succcurlyeq \lambda I_{\mathcal{H}} + \sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top = \Lambda_h^k.$$

The above relation further implies that $\Lambda_h^{-1} \preccurlyeq (\Lambda_h^k)^{-1}$ such that $\phi(s,a)^\top \Lambda_h^{-1} \phi(s,a) \le \phi(s,a)^\top (\Lambda_h^k)^{-1} \phi(s,a)$. This can be proved by extending the standard matrix case to the self-adjoint operator here. Thus, we have

$$w_h(s,a) \le w_h^k(s,a).$$

Since $r_h^k = H \cdot u_h^k(s,a) = H \cdot \min\{w_h^k(s,a), H\}^+$ and $u_h(s,a) = \min\{w_h(s,a), H\}^+$, then we have

$$u_h(s,a)/H \le r_h^k(s,a),$$

such that

$$V_1^*(s_1, u/H) \le V_1^*(s_1, r^k),$$

and thus

$$K \cdot V_1^*(s_1, u/H) \le \sum_{k=1}^K V_1^*(s_1, r^k).$$

This completes the proof. □

B.2. Proof of Theorem 3.3

Proof. Conditioned on the event $\mathcal{E}$ defined in Lemma B.4 and the event $\widetilde{\mathcal{E}}$ defined in Lemma B.6, we have

$$V_1^*(s_1, r) - V_1^\pi(s_1, r) \le V_1(s_1) - V_1^\pi(s_1, r), \tag{26}$$

where the inequality is by Lemma B.7. Further by this lemma, we have

$$\begin{aligned} V_h(s) - V_h^\pi(s, r) &\le r_h(s, \pi_h(s)) + \mathbb{P}_h V_{h+1}(s, \pi_h(s)) + 2u_h(s, \pi_h(s)) - Q_h^\pi(s, \pi_h(s), r) \\ &= r_h(s, \pi_h(s)) + \mathbb{P}_h V_{h+1}(s, \pi_h(s)) + 2u_h(s, \pi_h(s)) - r_h(s, \pi_h(s)) - \mathbb{P}_h V_{h+1}^\pi(s, \pi_h(s), r) \\ &= \mathbb{P}_h V_{h+1}(s, \pi_h(s)) - \mathbb{P}_h V_{h+1}^\pi(s, \pi_h(s), r) + 2u_h(s, \pi_h(s)). \end{aligned}$$

Recursively applying the above inequality and making use of $V_{H+1}^\pi(s, r) = V_{H+1}(s) = 0$ gives

$$\begin{aligned} V_1(s_1) - V_1^\pi(s_1, r) &\le \mathbb{E}_{\forall h \in [H]:\, s_{h+1} \sim \mathbb{P}_h(\cdot | s_h, \pi_h(s_h))} \left[\sum_{h=1}^H 2u_h(s_h, \pi_h(s_h)) \middle| s_1 \right] \\ &= 2H \cdot V_1^\pi(s_1, u/H). \end{aligned}$$

Combining this inequality with (26) gives

$$\begin{aligned} V_1^*(s_1, r) - V_1^\pi(s_1, r) \le 2H \cdot V_1^\pi(s_1, u/H) &\le \frac{2H}{K} \sum_{k=1}^K V_1^*(s_1, r^k) \\ &\le \frac{2H}{K} \mathcal{O}\left(\sqrt{H^3 K \log(1/\delta')} + \beta \sqrt{H^2 K \cdot \Gamma(K, \lambda; \ker)} \right) \\ &= \mathcal{O}\left([\sqrt{H^5 \log(1/\delta')} + \beta \sqrt{H^4 \cdot \Gamma(K, \lambda; \ker)}] / \sqrt{K} \right), \end{aligned}$$

where the second inequality is due to Lemma B.8 and the last inequality is by Lemma B.5.

By union bound, we have $P(\mathcal{E} \wedge \widetilde{\mathcal{E}}) \ge 1 - 2\delta'$. Therefore, by setting $\delta' = \delta/2$, we obtain that with probability at least $1 - \delta$

$$V_1^*(s_1, r) - V_1^\pi(s_1, r) \le \mathcal{O}\left([\sqrt{H^5 \log(2/\delta)} + \beta \sqrt{H^4 \cdot \Gamma(K, \lambda; \ker)}] / \sqrt{K} \right).$$

Note that $\mathcal{E} \wedge \widetilde{\mathcal{E}}$ happens when the following two conditions are satisfied, i.e.,

$$\begin{aligned} &4H^2 \big[R_Q^2 + 2\Gamma(K, \lambda; \ker) + 5 + \log \mathcal{N}_\infty(\varsigma^*; \widetilde{R}_K, \widetilde{B}_K) + 2\log(2K/\delta) \big] \le \widetilde{B}_K^2, \\ &16H^2 \big[R_Q^2 + 2\Gamma(K, \lambda; \ker) + 5 + \log \mathcal{N}_\infty(\varsigma^*; R_K, B_K) + 2\log(2K/\delta) \big] \le B_K^2, \forall h \in [H], \end{aligned}$$

where $\beta = \widetilde{B}_K$, $(1 + 1/H)\beta = B_K$, $\lambda = 1 + 1/K$, $\widetilde{R}_K = R_K = 2H\sqrt{\Gamma(K, \lambda; \ker)}$, and $\varsigma^* = H/K$. The above inequalities hold if we further let β satisfy

$$16H^2 \big[R_Q^2 + 2\Gamma(K, \lambda; \ker) + 5 + \log \mathcal{N}_\infty(\varsigma^*; R_K, 2\beta) + 2\log(2K/\delta) \big] \le \beta^2, \forall h \in [H],$$

since $2\beta \ge (1 + 1/H)\beta \ge \beta$ such that $\mathcal{N}_\infty(\varsigma^*; R_K, 2\beta) \ge \mathcal{N}_\infty(\varsigma^*; R_K, B_K) \ge \mathcal{N}_\infty(\varsigma^*; \widetilde{R}_K, \widetilde{B}_K)$. Since the above conditions imply that $\beta \ge H$, further setting $\delta = 1/(2K^2H^2)$, we obtain that

$$V_1^*(s_1, r) - V_1^\pi(s_1, r) \le \mathcal{O}\left(\beta \sqrt{H^4 [\Gamma(K, \lambda; \ker) + \log(KH)]} / \sqrt{K} \right),$$

with further letting

$$16H^2 \big[R_Q^2 + 2\Gamma(K, \lambda; \ker) + 5 + \log \mathcal{N}_\infty(\varsigma^*; R_K, 2\beta) + 6\log(2KH) \big] \le \beta^2, \forall h \in [H].$$

This completes the proof. □

C. Proofs for Single-Agent MDP Setting with Neural Function Approximation

C.1. Lemmas

Lemma C.1 (Lemma C.7 of Yang et al. (2020)). *With $TH^2 = \mathcal{O}(m\log^{-6} m)$, then there exists an constant $F \geq 1$ such that the following inequalities hold with probability at least $1 - 1/m^2$ for any $z \in \mathcal{S} \times \mathcal{A}$ and any $W \in \{W : \|W - W^{(0)}\| \leq H\sqrt{K/\lambda}\}$,*

$$|f(z;W) - \varphi(z;W^{(0)})^\top (W - W^{(0)})| \leq F K^{2/3} H^{4/3} m^{-1/6}\sqrt{\log m},$$
$$\|\varphi(z;W) - \varphi(z;W^{(0)})\| \leq F(KH^2/m)^{1/6}\sqrt{\log m}, \qquad \|\varphi(z;W)\| \leq F.$$

Lemma C.2. *We define the event $\mathcal{E}$ as that the following inequality holds $\forall (s,a) \in \mathcal{S}\times\mathcal{A}, \forall (h,k) \in [H]\times[K]$,*

$$|\mathbb{P}_h V_{h+1}^k(s,a) - f_h^k(s,a)| \leq u_h^k(s,a) + \beta\iota,$$
$$\left|\|\varphi(z;W_h^k)\|_{(\Lambda_h^k)^{-1}} - \|\varphi(z;W^{(0)})\|_{(\widetilde{\Lambda}_h^k)^{-1}}\right| \leq \iota,$$

where $\iota = 5K^{7/12}H^{1/6}m^{-1/12}\log^{1/4} m$ and we define

$$\Lambda_h^k = \sum_{\tau=1}^{k-1} \varphi(s_h^\tau, a_h^\tau; W_h^k)\varphi(s_h^\tau, a_h^\tau; W_h^k)^\top + \lambda\cdot I, \quad \widetilde{\Lambda}_h^k = \sum_{\tau=1}^{k-1} \varphi(s_h^\tau, a_h^\tau; W^{(0)})\varphi(s_h^\tau, a_h^\tau; W^{(0)})^\top + \lambda\cdot I.$$

Setting $(1+1/H)\beta = B_K$, $R_K = H\sqrt{K}$, $\varsigma^* = H/K$, and $\lambda = F^2(1+1/K)$, $\varsigma^* = H/K$, if we set

$$\beta^2 \geq 8R_Q^2 H^2(1+\sqrt{\lambda/d})^2 + 32H^2\Gamma(K,\lambda;\ker_m) + 80H^2 + 32H^2\log\mathcal{N}_\infty(\varsigma^*; R_K, B_K) + 32H^2\log(K/\delta'),$$

and also

$$m = \Omega(K^{19}H^{14}\log^3 m),$$

then we have that with probability at least $1 - 2/m^2 - \delta'$, the event $\mathcal{E}$ happens, i.e.,

$$\Pr(\mathcal{E}) \geq 1 - 2/m^2 - \delta'.$$

Proof. Recall that we assume $\mathbb{P}_h V_{h+1}$ for any V can be expressed as

$$\mathbb{P}_h V_{h+1}(z) = \int_{\mathbb{R}^d} \mathtt{act}'(\boldsymbol{\omega}^\top z)\cdot z^\top \boldsymbol{\alpha}(\boldsymbol{\omega})\mathrm{d}p_0(\boldsymbol{\omega}),$$

which thus implies that we have

$$\mathbb{P}_h V_{h+1}^k(z) = \int_{\mathbb{R}^d} \mathtt{act}'(\boldsymbol{\omega}^\top z)\cdot z^\top \boldsymbol{\alpha}_h^k(\boldsymbol{\omega})\mathrm{d}p_0(\boldsymbol{\omega}),$$

for some $\boldsymbol{\alpha}_h^k(\boldsymbol{\omega})$. Our algorithm suggests to estimate $\mathbb{P}_h V_{h+1}^k(s,a)$ via learning the parameters W_h^k by solving

$$W_h^k = \operatorname*{argmin}_W \sum_{\tau=1}^{k-1} [V_{h+1}^k(s_{h+1}^\tau) - f(s_h^\tau, a_h^\tau; W)]^2 + \lambda\|W - W^{(0)}\|^2, \tag{27}$$

such that we have the estimation of $\mathbb{P}_h V_{h+1}^k(s,a)$ as $f_h^k(z) = \Pi_{[0,H]}[f_h^k(z)]$ with

$$f(z;W_h^k) = \frac{1}{\sqrt{2m}}\sum_{i=1}^{2m} v_i\cdot \mathtt{act}([W_h^k]_i^\top z).$$

Furthermore, we have

$$\begin{aligned}\|W_h^k - W^{(0)}\|^2 &\leq \frac{1}{\lambda}\left(\sum_{\tau=1}^{k-1}[V_{h+1}^k(s_{h+1}^\tau) - f(s_h^\tau, a_h^\tau; W_h^k)]^2 + \lambda\|W_h^k - W^{(0)}\|^2\right)\\ &\leq \frac{1}{\lambda}\left(\sum_{\tau=1}^{k-1}[V_{h+1}^k(s_{h+1}^\tau) - f(s_h^\tau, a_h^\tau; W^{(0)})]^2 + \lambda\|W^{(0)} - W^{(0)}\|^2\right)\\ &= \frac{1}{\lambda}\sum_{\tau=1}^{k-1}[V_{h+1}^k(s_{h+1}^\tau)]^2 \leq H^2K/\lambda,\end{aligned}$$

where the second inequality is due to W_h^k is the minimizer of the objective function.

We also define a linearization of the function $f(z; W)$ at the point $W^{(0)}$, which is

$$f_{\texttt{lin}}(z; W) = f(z; W^{(0)}) + \langle\varphi(z; W^{(0)}), W - W^{(0)}\rangle = \langle\varphi(z; W^{(0)}), W - W^{(0)}\rangle, \tag{28}$$

where

$$\varphi(z; W) = \nabla_W f(z; W) = [\nabla_{W_1} f(z; W), \cdots, \nabla_{W_{2m}} f(z; W)].$$

Based on this linearization formulation, we similarly define a parameter matrix $W_{\texttt{lin},h}^k$ that is generated by solving an optimization problem with the linearied function $f_{\texttt{lin}}$, such that

$$W_{\texttt{lin},h}^k = \operatorname*{argmin}_W \sum_{\tau=1}^{k-1}[V_{h+1}^k(s_{h+1}^\tau) - f_{\texttt{lin}}(s_h^\tau, a_h^\tau; W)]^2 + \lambda\|W - W^{(0)}\|^2. \tag{29}$$

Due to the linear structure of $f_{\texttt{lin}}(z; W)$, one can easily solve the above optimization problem and obtain the closed form of the solution $W_{\texttt{lin},h}^k$, which is

$$W_{\texttt{lin},h}^k = W^{(0)} + (\widetilde{\Lambda}_h^t)^{-1}(\widetilde{\Phi}_h^k)^\top \mathbf{y}_h^k, \tag{30}$$

where we define Λ_h^t, Φ_h^k, and $\mathbf{y}_h^k$ as

$$\begin{aligned}\widetilde{\Phi}_h^k &= [\varphi(s_h^1, a_h^1; W^{(0)}), \cdots, \varphi(s_h^{k-1}, a_h^{k-1}; W^{(0)})]^\top,\\ \widetilde{\Lambda}_h^k &= \sum_{\tau=1}^{k-1}\varphi(s_h^\tau, a_h^\tau; W^{(0)})\varphi(s_h^\tau, a_h^\tau; W^{(0)})^\top + \lambda\cdot I = \lambda\cdot I + (\widetilde{\Phi}_h^k)^\top\widetilde{\Phi}_h^k,\\ \mathbf{y}_h^k &= [V_{h+1}^k(s_{h+1}^1), V_{h+1}^k(s_{h+1}^2), \cdots, V_{h+1}^k(s_{h+1}^{k-1})]^\top.\end{aligned}$$

Here we also have the upper bound of $\|W_{\texttt{lin},h}^k - W^{(0)}\|$ as

$$\begin{aligned}\|W_{\texttt{lin},h}^k - W^{(0)}\|^2 &\leq \frac{1}{\lambda}\left(\sum_{\tau=1}^{k-1}[V_{h+1}^k(s_{h+1}^\tau) - f_{\texttt{lin}}(s_h^\tau, a_h^\tau; W_{\texttt{lin},h}^k)]^2 + \lambda\|W_{\texttt{lin},h}^k - W^{(0)}\|^2\right)\\ &\leq \frac{1}{\lambda}\left(\sum_{\tau=1}^{k-1}[V_{h+1}^k(s_{h+1}^\tau) - f_{\texttt{lin}}(s_h^\tau, a_h^\tau; W^{(0)})]^2 + \lambda\|W^{(0)} - W^{(0)}\|^2\right)\\ &= \frac{1}{\lambda}\sum_{\tau=1}^{k-1}[V_{h+1}^k(s_{h+1}^\tau)]^2 \leq H^2K/\lambda,\end{aligned}$$

where the second inequality is due to $W_{\texttt{lin},h}^k$ is the minimizer of the objective function. Based on the matrix $W_{\texttt{lin},h}^k$, we define the function

$$f_{\texttt{lin},h}^k(z) := \Pi_{[0,H]}[f_{\texttt{lin}}(z; W_{\texttt{lin},h}^k)],$$

where $\Pi_{[0,H]}[\cdot]$ is short for $\min\{\cdot, H\}^+$.

Moreover, we further define an approximation of $\mathbb{P}_h V_{h+1}^h$ as

$$\widetilde{f}(z) = \Pi_{[0,H]}\left[\frac{1}{\sqrt{m}}\sum_{i=1}^{m} \mathtt{act}'(W_i^{(0)\top} z) z^\top \boldsymbol{\alpha}_i\right],$$

where $\|\boldsymbol{\alpha}_i\| \le R_Q H/\sqrt{dm}$. According to Gao et al. (2019), we have that with probability at least $1 - 1/m^2$ over the randomness of initialization, for any $(h,k) \in [H] \times [K]$, there exists a constant $C_{\mathtt{act}}$ such that

$$\left|\mathbb{P}_h V_{h+1}^k(z) - \frac{1}{\sqrt{m}}\sum_{i=1}^{m} \mathtt{act}'(W_i^{(0)\top} z) z^\top \boldsymbol{\alpha}_i\right| \le 10 C_{\mathtt{act}} R_Q H \sqrt{\log(mKH)/m}, \ \ \forall z = (s,a) \in \mathcal{S} \times \mathcal{A}.$$

which further implies that

$$|\mathbb{P}_h V_{h+1}^k(z) - \widetilde{f}(z)| \le 10 C_{\mathtt{act}} R_Q H \sqrt{\log(mKH)/m}, \ \ \forall z = (s,a) \in \mathcal{S} \times \mathcal{A}. \tag{31}$$

This indicates that $\widetilde{f}(z)$ can be regarded as a good surrogate of $P_h V_{h+1}^h(z)$ particularly when m is large, i.e., their difference $10 C_{\mathtt{act}} R_Q H \sqrt{\log(mKH)/m}$ is small.

Now, based on the above definitions and descriptions, we are ready to present our proof of this lemma. Overall, the basic idea of proving the upper bound of $|P_h V_{h+1}^k(z) - f_h^k(z)|$ is to bound the following difference terms, i.e.,

$$|f_h^k(z) - f_{\mathtt{lin},h}^k(z)| \quad \text{and} \quad |f_{\mathtt{lin},h}^k(z) - \widetilde{f}(z)|. \tag{32}$$

As we already have known the upper bound of the term $|\mathbb{P}_h V_{h+1}^h(z) - \widetilde{f}(z)|$ in (31), one can immediately obtain the upper bound of $|\mathbb{P}_h V_{h+1}^k(z) - f_h^k(z)|$ by decomposing it into the two aforementioned terms and bounding them separately.

We first bound the first term in (32), i.e., $|f_h^k(z) - f_{\mathtt{lin}}(z; W_{\mathtt{lin},h}^k)|$, in the following way

$$\begin{aligned}
|f_h^k(z) - f_{\mathtt{lin},h}^k(z)| &\le |f(z; W_h^k) - \langle \varphi(z; W^{(0)}), W_{\mathtt{lin},h}^k - W^{(0)} \rangle| \\
&\le |f(z; W_h^k) - \langle \varphi(z; W^{(0)}), W_h^k - W^{(0)} \rangle| + |\langle \varphi(z; W^{(0)}), W_h^k - W_{\mathtt{lin},h}^k \rangle| \\
&\le \digamma K^{2/3} H^{4/3} m^{-1/6} \sqrt{\log m} + \digamma \underbrace{\|W_h^k - W_{\mathtt{lin},h}^k\|}_{\text{Term(I)}},
\end{aligned} \tag{33}$$

where the first inequality is due to the non-expansiveness of projection operation $\Pi_{[0,H]}$, the third inequality is by Lemma C.1 that holds with probability at least $1 - m^{-2}$. Then, we need to bound Term(I) in the above inequality. Specifically, by the first order optimality condition for the objectives in (27) and (29), we have

$$\begin{aligned}
\lambda(W_h^k - W^{(0)}) &= \sum_{\tau=1}^{k-1} [V_{h+1}^k(s_{h+1}^\tau) - f(z_h^\tau; W_h^k)] \varphi(z_h^\tau; W_h^k) = (\Phi_h^k)^\top (\mathbf{y}_h^k - \mathbf{f}_h^k), \\
\lambda(W_{\mathtt{lin},h}^k - W^{(0)}) &= \sum_{\tau=1}^{k-1} [V_{h+1}^k(s_{h+1}^\tau) - \langle \varphi(z_h^\tau; W^{(0)}), W_{\mathtt{lin},h}^k - W^{(0)} \rangle] \varphi(z_h^\tau; W^{(0)}) \\
&= (\widetilde{\Phi}_h^k)^\top \mathbf{y}_h^k - (\widetilde{\Phi}_h^k)^\top \widetilde{\Phi}_h^k (W_{\mathtt{lin},h}^k - W^{(0)}),
\end{aligned}$$

where we define

$$\begin{aligned}
\Phi_h^k &= [\varphi(s_h^1, a_h^1; W_h^k), \cdots, \varphi(s_h^{k-1}, a_h^{k-1}; W_h^k)]^\top, \\
\Lambda_h^k &= \sum_{\tau=1}^{k-1} \varphi(s_h^\tau, a_h^\tau; W_h^k) \varphi(s_h^\tau, a_h^\tau; W_h^k)^\top + \lambda \cdot I = \lambda \cdot I + (\Phi_h^k)^\top \Phi_h^k, \\
\mathbf{f}_h^k &= [f(z_h^1; W_h^k), f(z_h^2; W_h^k), \cdots, f(z_h^{k-1}; W_h^k)]^\top.
\end{aligned}$$

Thus, we have

$$\begin{aligned}\text{Term(I)} &= \lambda^{-1}\|(\Phi_h^k)^\top(\mathbf{y}_h^k - \mathbf{f}_h^k) - (\widetilde{\Phi}_h^k)^\top \mathbf{y}_h^k + (\widetilde{\Phi}_h^k)^\top\widetilde{\Phi}_h^k(W_{\texttt{lin},h}^k - W^{(0)})\| \\ &= \lambda^{-1}\|(\Phi_h^k)^\top(\mathbf{y}_h^k - \mathbf{f}_h^k) - (\widetilde{\Phi}_h^k)^\top \mathbf{y}_h^k + (\widetilde{\Phi}_h^k)^\top\widetilde{\Phi}_h^k(W_{\texttt{lin},h}^k - W^{(0)})\| \\ &\leq \lambda^{-1}\|((\Phi_h^k)^\top - (\widetilde{\Phi}_h^k)^\top)\mathbf{y}_h^k\| + \lambda^{-1}\|(\Phi_h^k)^\top[\mathbf{f}_h^k - \widetilde{\Phi}_h^k(W_{\texttt{lin},h}^k - W^{(0)})]\| \\ &\quad + \lambda^{-1}\|((\Phi_h^k)^\top - (\widetilde{\Phi}_h^k)^\top)\widetilde{\Phi}_h^k(W_{\texttt{lin},h}^k - W^{(0)})\|.\end{aligned}$$

According to Lemma C.1, we can bound the last three terms in the above inequality separately as follows

$$\lambda^{-1}\|((\Phi_h^k)^\top - (\widetilde{\Phi}_h^k)^\top)\mathbf{y}_h^k\| \leq \lambda^{-1}K\max_{\tau\in[k-1]}|[\varphi(z_h^\tau; W_h^k) - \varphi(z_h^\tau; W^{(0)})]\cdot[\mathbf{y}_h^k]_\tau| \leq F\lambda^{-1}K^{7/6}H^{4/3}m^{-1/6}\sqrt{\log m},$$

and similarly,

$$\begin{aligned}&\lambda^{-1}\|(\Phi_h^k)^\top[\mathbf{f}_h^k - \widetilde{\Phi}_h^k(W_{\texttt{lin},h}^k - W^{(0)})]\| \leq \lambda^{-1}F^2K^{5/3}H^{4/3}m^{-1/6}\sqrt{\log m}, \\ &\lambda^{-1}\|((\Phi_h^k)^\top - (\widetilde{\Phi}_h^k)^\top)\widetilde{\Phi}_h^k(W_{\texttt{lin},h}^k - W^{(0)})\| \leq \lambda^{-3/2}F^2K^{5/3}H^{4/3}m^{-1/6}\sqrt{\log m}.\end{aligned}$$

Thus, we have

$$\text{Term(I)} \leq \lambda^{-1}(FK^{7/6} + 2F^2K^{5/3})H^{4/3}m^{-1/6}\sqrt{\log m} \leq 3K^{5/3}H^{4/3}m^{-1/6}\sqrt{\log m}.$$

where we set $\lambda = F^2(1+1/K)$, and use the fact that $\lambda \geq 1$ as $F \geq 1$ as well as $F^2/\lambda \in [1/2, 1]$ and $F/\lambda \in [1/2, 1]$. Combining the above upper bound of Term(I) with (33), we obtain

$$|f_h^k(z) - f_{\texttt{lin},h}^k(z)| \leq 4FK^{5/3}H^{4/3}m^{-1/6}\sqrt{\log m}. \tag{34}$$

Next, we bound the second term in (32), namely $|f_{\texttt{lin}}(z; W_{\texttt{lin},h}^k) - \widetilde{f}(z)|$. We further have

$$\begin{aligned}&\frac{1}{\sqrt{m}}\sum_{i=1}^m \texttt{act}'(W_i^{(0)\top}z)z^\top\boldsymbol{\alpha}_i \\ &\quad= \frac{1}{\sqrt{2m}}\sum_{i=1}^m \frac{(v_i^{(0)})^2}{\sqrt{2}}\texttt{act}'(W_i^{(0)\top}z)z^\top\boldsymbol{\alpha}_i + \frac{1}{\sqrt{2m}}\sum_{i=1}^m \frac{(v_i^{(0)})^2}{\sqrt{2}}\texttt{act}'(W_i^{(0)\top}z)z^\top\boldsymbol{\alpha}_i \\ &\quad= \frac{1}{\sqrt{2m}}\sum_{i=1}^m \frac{(v_i^{(0)})^2}{\sqrt{2}}\texttt{act}'(W_i^{(0)\top}z)z^\top\boldsymbol{\alpha}_i + \frac{1}{\sqrt{2m}}\sum_{i=m+1}^{2m} \frac{(v_{i-m}^{(0)})^2}{\sqrt{2}}\texttt{act}'(W_i^{(0)\top}z)z^\top\boldsymbol{\alpha}_{i-m} \\ &\quad= \frac{1}{\sqrt{2m}}\sum_{i=1}^m \frac{(v_i^{(0)})^2}{\sqrt{2}}\texttt{act}'(W_i^{(0)\top}z)z^\top\boldsymbol{\alpha}_i + \frac{1}{\sqrt{2m}}\sum_{i=m+1}^{2m} \frac{(v_i^{(0)})^2}{\sqrt{2}}\texttt{act}'(W_i^{(0)\top}z)z^\top\boldsymbol{\alpha}_i \\ &\quad= \frac{1}{\sqrt{2m}}\sum_{i=1}^{2m} v_i^{(0)}\texttt{act}'(W_i^{(0)\top}z)z^\top(\widetilde{W}_i - W_i^{(0)}) \\ &\quad= \langle\varphi(z; W^{(0)}), \widetilde{W} - W^{(0)}\rangle,\end{aligned}$$

where we define

$$\widetilde{W}_i = \begin{cases} W_i^{(0)} + \frac{v_i^{(0)}}{\sqrt{2}}\boldsymbol{\alpha}_i, \text{ if } 1 \leq i \leq m, \\ W_i^{(0)} + \frac{v_i^{(0)}}{\sqrt{2}}\boldsymbol{\alpha}_{i-m}, \text{ if } m+1 \leq i \leq 2m. \end{cases}$$

Then, we can reformulate $\widetilde{f}(z)$ as follows

$$\widetilde{f}(z) = \Pi_{[0,H]}[\langle\varphi(z; W^{(0)}), \widetilde{W} - W^{(0)}\rangle].$$

Since $\|\boldsymbol{\alpha}_i\| \leq R_Q H/\sqrt{d}$, then there is $\|\widetilde{W} - W^{(0)}\| \leq R_Q H/\sqrt{d}$. Equivalently, we further have

$$\begin{aligned}\langle \varphi(z;W^{(0)}), \widetilde{W} - W^{(0)}\rangle &= \langle \varphi(z;W^{(0)}), (\widetilde{\Lambda}_h^k)^{-1}\widetilde{\Lambda}_h^k(\widetilde{W} - W^{(0)})\rangle \\ &= \langle \varphi(z;W^{(0)}), \lambda(\widetilde{\Lambda}_h^k)^{-1}(\widetilde{W} - W^{(0)})\rangle + \langle \varphi(z;W^{(0)}), (\widetilde{\Lambda}_h^k)^{-1}(\widetilde{\Phi}_h^k)^\top \widetilde{\Phi}_h^k(\widetilde{W} - W^{(0)})\rangle,\end{aligned} \tag{35}$$

since $\Lambda_h^k = \lambda I + (\widetilde{\Phi}_h^k)^\top \widetilde{\Phi}_h^k$. Thus, by the above equivalent form of $\widetilde{f}(z)$ in (35), and further with the formulation of $f_{\mathtt{lin},h}^k(z)$ according to (28) and (30), we have

$$\begin{aligned}&|f_{\mathtt{lin},h}^k(z) - \widetilde{f}(z)| \\ &\quad\leq |\langle \varphi(z;W^{(0)}), W_{\mathtt{lin},h}^k - \widetilde{W}\rangle| \\ &\quad\leq \underbrace{|\langle \varphi(z;W^{(0)}), \lambda(\widetilde{\Lambda}_h^k)^{-1}(\widetilde{W} - W^{(0)})\rangle|}_{\text{Term(II)}} + \underbrace{|\langle \varphi(z;W^{(0)}), (\widetilde{\Lambda}_h^t)^{-1}(\widetilde{\Phi}_h^k)^\top[\mathbf{y}_h^k - \widetilde{\Phi}_h^k(\widetilde{W} - W^{(0)})]\rangle|}_{\text{Term(III)}}.\end{aligned}$$

The first term Term(II) can be bounded as

$$\begin{aligned}\text{Term(II)} &= |\langle \varphi(z;W^{(0)}), \lambda(\widetilde{\Lambda}_h^k)^{-1}(\widetilde{W} - W^{(0)})\rangle| \\ &\leq \lambda\|\varphi(z;W^{(0)})\|_{(\widetilde{\Lambda}_h^k)^{-1}}\|\widetilde{W} - W^{(0)}\|_{(\widetilde{\Lambda}_h^k)^{-1}} \\ &\leq \sqrt{\lambda}\|\varphi(z;W^{(0)})\|_{(\widetilde{\Lambda}_h^k)^{-1}}\|\widetilde{W} - W^{(0)}\| \\ &\leq \sqrt{\lambda}R_Q H/\sqrt{d}\cdot\|\varphi(z;W^{(0)})\|_{(\widetilde{\Lambda}_h^k)^{-1}},\end{aligned}$$

where the first inequality is by $\|\widetilde{W} - W^{(0)}\|_{(\widetilde{\Lambda}_h^k)^{-1}} = \sqrt{(\widetilde{W} - W^{(0)})^\top(\widetilde{\Lambda}_h^k)^{-1}(\widetilde{W} - W^{(0)})} \leq 1/\sqrt{\lambda}\|\widetilde{W} - W^{(0)}\|_2$ since $(\widetilde{\Lambda}_h^k)^{-1} \preccurlyeq 1/\lambda \cdot I$ and the last inequality is due to $\|\widetilde{W} - W^{(0)}\|_2 \leq R_Q H/\sqrt{d}$.

Next, we prove the bound of Term(III) in the following way

$$\begin{aligned}\text{Term(III)} &= |\langle \varphi(z;W^{(0)}), (\widetilde{\Lambda}_h^t)^{-1}(\widetilde{\Phi}_h^k)^\top[\mathbf{y}_h^k - \widetilde{\Phi}_h^k(\widetilde{W} - W^{(0)})]\rangle| \\ &\leq |\langle \varphi(z;W^{(0)}), (\widetilde{\Lambda}_h^t)^{-1}(\widetilde{\Phi}_h^k)^\top[\widetilde{\mathbf{y}}_h^k - \widetilde{\Phi}_h^k(\widetilde{W} - W^{(0)})]\rangle| + |\langle \varphi(z;W^{(0)}), (\widetilde{\Lambda}_h^t)^{-1}(\widetilde{\Phi}_h^k)^\top[\mathbf{y}_h^k - \widetilde{\mathbf{y}}_h^k]\rangle| \\ &\leq \|\varphi(z;W^{(0)})\|_{(\widetilde{\Lambda}_h^k)^{-1}}\cdot\|(\widetilde{\Phi}_h^k)^\top[\widetilde{\mathbf{y}}_h^k - \widetilde{\Phi}_h^k(\widetilde{W} - W^{(0)})]\|_{(\widetilde{\Lambda}_h^k)^{-1}} + \|\varphi(z;W^{(0)})\|_{(\widetilde{\Lambda}_h^k)^{-1}}\cdot\|(\Phi_h^k)^\top[\mathbf{y}_h^k - \widetilde{\mathbf{y}}_h^k]\|_{(\widetilde{\Lambda}_h^k)^{-1}} \\ &\leq 10C_{\mathtt{act}}R_Q H\sqrt{K\log(mKH)/m}\|\varphi(z;W^{(0)})\|_{(\widetilde{\Lambda}_h^k)^{-1}} + \|\varphi(z;W^{(0)})\|_{(\widetilde{\Lambda}_h^k)^{-1}}\cdot\underbrace{\|(\widetilde{\Phi}_h^k)^\top[\mathbf{y}_h^k - \widetilde{\mathbf{y}}_h^k]\|_{(\widetilde{\Lambda}_h^k)^{-1}}}_{\text{Term(IV)}},\end{aligned}$$

where we define $\widetilde{\mathbf{y}}_h^k = [\mathbb{P}_h V_{h+1}^k(s_{h+1}^1), \mathbb{P}_h V_{h+1}^k(s_{h+1}^2), \cdots, \mathbb{P}_h V_{h+1}^k(s_{h+1}^{k-1})]^\top$. Here, the last inequality is by

$$\begin{aligned}&\|(\Phi_h^k)^\top[\widetilde{\mathbf{y}}_h^k - \widetilde{\Phi}_h^k(\widetilde{W} - W^{(0)})]\|_{(\widetilde{\Lambda}_h^k)^{-1}} \\ &\quad= \sqrt{[\widetilde{\mathbf{y}}_h^k - \widetilde{\Phi}_h^k(\widetilde{W} - W^{(0)})]^\top\widetilde{\Phi}_h^k[\lambda I + (\widetilde{\Phi}_h^k)^\top\widetilde{\Phi}_h^k]^{-1}(\widetilde{\Phi}_h^k)^\top[\widetilde{\mathbf{y}}_h^k - \widetilde{\Phi}_h^k(\widetilde{W} - W^{(0)})]} \\ &\quad= \sqrt{[\widetilde{\mathbf{y}}_h^k - \widetilde{\Phi}_h^k(\widetilde{W} - W^{(0)})]^\top\widetilde{\Phi}_h^k(\widetilde{\Phi}_h^k)^\top[\lambda I + \widetilde{\Phi}_h^k(\widetilde{\Phi}_h^k)^\top]^{-1}[\widetilde{\mathbf{y}}_h^k - \widetilde{\Phi}_h^k(\widetilde{W} - W^{(0)})]} \\ &\quad\leq \sqrt{[\widetilde{\mathbf{y}}_h^k - \widetilde{\Phi}_h^k(\widetilde{W} - W^{(0)})]^\top[\lambda I + \widetilde{\Phi}_h^k(\widetilde{\Phi}_h^k)^\top][\lambda I + \widetilde{\Phi}_h^k(\widetilde{\Phi}_h^k)^\top]^{-1}[\widetilde{\mathbf{y}}_h^k - \widetilde{\Phi}_h^k(\widetilde{W} - W^{(0)})]} \\ &\quad= \|\widetilde{\mathbf{y}}_h^k - \widetilde{\Phi}_h^k(\widetilde{W} - W^{(0)})\| \leq 10C_{\mathtt{act}}R_Q H\sqrt{K\log(mKH)/m},\end{aligned}$$

where the second equality is by Woodbury matrix identity, the first inequality is due to $[\lambda I + \widetilde{\Phi}_h^k(\widetilde{\Phi}_h^k)^\top]^{-1} \succ 0$, and the second inequality is by (31) such that

$$\begin{aligned}\|\widetilde{\mathbf{y}}_h^k - \widetilde{\Phi}_h^k(\widetilde{W} - W^{(0)})\| &\leq \sqrt{k-1}\|\widetilde{\mathbf{y}}_h^k - \widetilde{\Phi}_h^k(\widetilde{W} - W^{(0)})\|_\infty \\ &= \sqrt{k-1}\sup_{\tau\in[k-1]}|\mathbb{P}_h V_{h+1}^k(s_h^\tau, a_h^\tau) - \widetilde{f}(s_h^\tau, a_h^\tau)| \\ &\leq 10C_{\mathtt{act}}R_Q H\sqrt{K\log(mKH)/m}.\end{aligned}$$

In order to further bound Term(IV), we define a new Q function based on $W^k_{\texttt{lin},h}$, which is

$$Q^k_{\texttt{lin},h}(z) := \min\{r^k_{\texttt{lin},h}(z) + f^k_{\texttt{lin},h}(z) + u^k_{\texttt{lin},h}(z), H\}^+,$$

where $r_{\texttt{lin},h}(s,a) = u^k_{\texttt{lin},h}(z)/H$, and $u^k_{\texttt{lin},h}(z) = \min\{\beta\|\varphi(z;W^{(0)})\|_{(\widetilde{\Lambda}^k_h)^{-1}}, H\}$. This Q function can be equivalently reformulated with a normalized feature representation $\vartheta = \varphi/\digamma$ as follows

$$Q^k_{\texttt{lin},h}(z) = \min\{\Pi_{[0,H]}[\langle\vartheta(z;W^{(0)}), \digamma\cdot(W^k_{\texttt{lin},h} - W^{(0)})\rangle] + (1+1/H)\cdot\min\{\beta\|\vartheta(z;W^{(0)})\|_{(\Xi^k_h)^{-1}}\}, H\}^+, \quad (36)$$

where we have

$$\Xi^k_h := \lambda/\digamma^2\cdot I + (\Theta^k_h)^\top\Theta^k_h, \qquad \Theta^k_h := \Phi^k_h/\digamma.$$

Note that $\digamma\|W^k_{\texttt{lin},h} - W^{(0)}\| \le \digamma H\sqrt{K/\lambda} \le H\sqrt{K}$ since $\lambda = \digamma^2(1+1/K)$. Thus, we can see that this new Q function lies in the space $\overline{\mathcal{Q}}(\mathbf{0}, R_K, B_K)$ as in (10), with $R_K = H\sqrt{K}$ and $B_K = (1+1/H)\beta$ with the kernel function defined as $\widetilde{\texttt{ker}}_m(z,z') := \langle\vartheta(z),\vartheta(z')\rangle$.

Now we try to bound the difference between the Q function $Q^k_h(z)$ in the exploration algorithm and the one $Q^k_{\texttt{lin},h}(z)$, which is

$$|Q^k_h(z) - Q^k_{\texttt{lin},h}(z)| \le |f^k_h(z) - f^k_{\texttt{lin},h}(z)| + (1+1/H)\beta\left|\|\varphi(z;W^k_h)\|_{(\Lambda^k_h)^{-1}} - \|\varphi(z;W^{(0)})\|_{(\widetilde{\Lambda}^k_h)^{-1}}\right|,$$

where the inequality is by the contraction of the operator $\min\{\cdot, H\}^+$. The upper bound of the term $|f^k_h(z) - f^k_{\texttt{lin},h}(z)|$ has already been studied in (34). Then, we focus on bounding the last term. Thus, we have

$$\begin{aligned}
&\left|\|\varphi(z;W^k_h)\|_{(\Lambda^k_h)^{-1}} - \|\varphi(z;W^{(0)})\|_{(\widetilde{\Lambda}^k_h)^{-1}}\right| \\
&\quad\le \sqrt{\left|\varphi(z;W^k_h)^\top(\Lambda^k_h)^{-1}\varphi(z;W^k_h) - \varphi(z;W^{(0)})^\top(\widetilde{\Lambda}^k_h)^{-1}\varphi(z;W^{(0)})\right|} \\
&\quad\le \sqrt{\left|[\varphi(z;W^k_h) - \varphi(z;W^{(0)})]^\top(\Lambda^k_h)^{-1}\varphi(z;W^k_h)\right|} + \sqrt{\left|\varphi(z;W^{(0)})^\top((\Lambda^k_h)^{-1} - (\widetilde{\Lambda}^k_h)^{-1})\varphi(z;W^k_h)\right|} \\
&\qquad + \sqrt{\left|\varphi(z;W^{(0)})^\top(\widetilde{\Lambda}^k_h)^{-1}[\varphi(z;W^k_h) - \varphi(z;W^{(0)})]\right|}.
\end{aligned}$$

Conditioned on the event that all the inequalities in Lemma C.1 hold, we can bound the last three terms above as follows

$$\begin{aligned}
&\left|[\varphi(z;W^k_h) - \varphi(z;W^{(0)})]^\top(\Lambda^k_h)^{-1}\varphi(z;W^k_h)\right| \\
&\qquad\le \|\varphi(z;W^k_h) - \varphi(z;W^{(0)})\|\|(\Lambda^k_h)^{-1}\|\|\varphi(z;W^k_h)\| \le \lambda^{-1}\digamma^2(KH^2/m)^{1/6}\sqrt{\log m}, \\
&\left|\varphi(z;W^{(0)})^\top(\widetilde{\Lambda}^k_h)^{-1}[\varphi(z;W^k_h) - \varphi(z;W^{(0)})]\right| \le \lambda^{-1}\digamma^2(KH^2/m)^{1/6}\sqrt{\log m}, \\
&\left|\varphi(z;W^{(0)})^\top((\Lambda^k_h)^{-1} - (\widetilde{\Lambda}^k_h)^{-1})\varphi(z;W^k_h)\right| \\
&\qquad\le \|\varphi(z;W^{(0)})\|\|(\Lambda^k_h)^{-1}(\Lambda^k_h - \widetilde{\Lambda}^k_h)(\widetilde{\Lambda}^k_h)^{-1}\|\|\varphi(z;W^k_h)\| \\
&\qquad\le \lambda^{-2}\digamma^2\|(\Phi^k_h)^\top\Phi^k_h - (\widetilde{\Phi}^k_h)^\top\widetilde{\Phi}^k_h\|_{\text{fro}} \le \lambda^{-2}\digamma^2(\|(\Phi^k_h - \widetilde{\Phi}^k_h)^\top\Phi^k_h\|_{\text{fro}} + \|(\widetilde{\Phi}^k_h)^\top(\Phi^k_h - \widetilde{\Phi}^k_h)\|_{\text{fro}}) \\
&\qquad\le \lambda^{-2}\digamma^4K^{7/6}H^{1/3}m^{-1/6}\sqrt{\log m},
\end{aligned}$$

which thus lead to

$$\left|\|\varphi(z;W^k_h)\|_{(\Lambda^k_h)^{-1}} - \|\varphi(z;W^{(0)})\|_{(\widetilde{\Lambda}^k_h)^{-1}}\right| \le 3K^{7/12}H^{1/6}m^{-1/12}\log^{1/4}m, \quad (37)$$

and thus

$$|Q^k_h(z) - Q^k_{\texttt{lin},h}(z)| \le 4\digamma K^{5/3}H^{4/3}m^{-1/6}\sqrt{\log m} + 3(1+1/H)\beta K^{7/12}H^{1/6}m^{-1/12}\log^{1/4}m,$$

where we use the fact that $\lambda = \digamma^2(1+1/K) \in [\digamma^2, 2\digamma^2]$. This further implies that we have the same bound for $|V_h^k(s) - V_{\mathtt{lin},h}^k(s)|$, .i.e.,

$$\begin{aligned}|V_h^k(s) - V_{\mathtt{lin},h}^k(s)| &\le \max_{a\in\mathcal{A}} |Q_h^k(s,a) - Q_{\mathtt{lin},h}^k(s,a)| \\ &\le 4\digamma K^{5/3}H^{4/3}m^{-1/6}\sqrt{\log m} + 3(1+1/H)\beta K^{7/12}H^{1/6}m^{-1/12}\log^{1/4} m,\end{aligned} \tag{38}$$

where we define $V_{\mathtt{lin},h}^k(s) = \max_{a\in\mathcal{A}} Q_{\mathtt{lin},h}^k(s,a)$.

Now, we are ready to give the upper bound of Term(IV). With probability at least $1-\delta'$, we have

$$\begin{aligned}\text{Term(IV)} &= \bigg\|\sum_{\tau=1}^{k-1}[V_{h+1}^k(s_{h+1}^\tau) - \mathbb{P}_h V_{h+1}^k(z_h^\tau)]\varphi(z_h^\tau; W^{(0)})\bigg\|_{(\widetilde{\Lambda}_h^k)^{-1}} \\ &\le \sqrt{\bigg\|\sum_{\tau=1}^{k-1}[V_{\mathtt{lin},h+1}^k(s_{h+1}^\tau) - \mathbb{P}_h V_{\mathtt{lin},h+1}^k(z_h^\tau)]\varphi(z_h^\tau; W^{(0)})\bigg\|_{(\widetilde{\Lambda}_h^k)^{-1}}^2} \\ &\quad + \bigg\|\sum_{\tau=1}^{k-1}\{[V_{h+1}^k(s_{h+1}^\tau) - V_{\mathtt{lin},h+1}^k(s_{h+1}^\tau)] - \mathbb{P}_h[V_{h+1}^k - V_{\mathtt{lin},h+1}^k(s_{h+1}^\tau)]\}\varphi(z_h^\tau; W^{(0)})\bigg\|_{(\widetilde{\Lambda}_h^k)^{-1}} \\ &\le [4H^2\Gamma(K,\lambda';\widetilde{\ker}_m) + 10H^2 + 4H^2\log\mathcal{N}_\infty(\varsigma^*; R_K, B_K) + 4H^2\log(K/\delta')]^{1/2} \\ &\quad + 8\digamma K^{8/3}H^{4/3}m^{-1/6}\sqrt{\log m} + 12\beta K^{19/12}H^{1/6}m^{-1/12}\log^{1/4} m.\end{aligned}$$

Here we set $\lambda' = \lambda/\digamma^2 = (1+1/K)$, $\varsigma^* = H/K$, $R_K = H\sqrt{K}$, $B_K = (1+1/H)\beta$, and $\widetilde{\ker}_m(z,z') = \langle\vartheta(z),\vartheta(z')\rangle$. Here the second inequality is by (36), and also follows the similar proof of Lemma B.3. The last inequality is by (38) and Lemma C.1, which lead to

$$\begin{aligned}&\bigg\|\sum_{\tau=1}^{k-1}\{[V_{h+1}^k(s_{h+1}^\tau) - V_{\mathtt{lin},h+1}^k(s_{h+1}^\tau)] - \mathbb{P}_h[V_{h+1}^k - V_{\mathtt{lin},h+1}^k(s_{h+1}^\tau)]\}\varphi(z_h^\tau; W^{(0)})\bigg\|_{(\widetilde{\Lambda}_h^k)^{-1}} \\ &\quad\le \sum_{\tau=1}^{k-1} 2[4\digamma K^{5/3}H^{4/3}m^{-1/6}\sqrt{\log m} + 3(1+1/H)\beta K^{7/12}H^{1/6}m^{-1/12}\log^{1/4} m]\|\varphi(z_h^\tau; W^{(0)})\|_{(\widetilde{\Lambda}_h^k)^{-1}} \\ &\quad\le K\digamma/\sqrt{\lambda}[8\digamma K^{5/3}H^{4/3}m^{-1/6}\sqrt{\log m} + 6(1+1/H)\beta K^{7/12}H^{1/6}m^{-1/12}\log^{1/4} m] \\ &\quad\le 8\digamma K^{8/3}H^{4/3}m^{-1/6}\sqrt{\log m} + 12\beta K^{19/12}H^{1/6}m^{-1/12}\log^{1/4} m.\end{aligned}$$

Now we let β satisfy

$$\begin{aligned}&\sqrt{\lambda}R_Q H/\sqrt{d} + 10C_{\mathtt{act}}R_Q H\sqrt{K\log(mKH)/m} + [4H^2\Gamma(K,\lambda';\widetilde{\ker}_m) + 10H^2 + 4H^2\log\mathcal{N}_\infty(\varsigma^*; R_K, B_K) \\ &+ 4H^2\log(K/\delta')]^{1/2} + 8\digamma K^{8/3}H^{4/3}m^{-1/6}\sqrt{\log m} + 12\beta K^{19/12}H^{1/6}m^{-1/12}\log^{1/4} m \le \beta.\end{aligned}$$

To obtain the above relation, it suffices to set

$$m = \Omega(K^{19}H^{14}\log^3 m)$$

such that m is sufficient large which results in

$$10C_{\mathtt{act}}R_Q H\sqrt{K\log(mKH)/m} + 8\digamma K^{8/3}H^{4/3}m^{-1/6}\sqrt{\log m} + 12\beta K^{19/12}H^{1/6}m^{-1/12}\log^{1/4} m \le R_Q H + \beta/2.$$

Then, there is

$$\sqrt{\lambda}R_Q H/\sqrt{d} + R_Q H + \beta/2 + [4H^2\Gamma(K,\lambda;\ker_m) + 10H^2 + 4H^2\log\mathcal{N}_\infty(\varsigma^*; R_K, B_K) + 4H^2\log(K/\delta')]^{1/2} \le \beta,$$

where $\Gamma(K,\lambda;\ker_m) = \Gamma(K,\lambda';\widetilde{\ker}_m)$ with $\ker_m := \langle\varphi(z;W^{(0)}),\varphi(z';W^{(0)})\rangle$. This inequality can be satisfied if we set β as

$$\beta^2 \ge 8R_Q^2H^2(1+\sqrt{\lambda/d})^2 + 32H^2\Gamma(K,\lambda;\ker_m) + 80H^2 + 32H^2\log\mathcal{N}_\infty(\varsigma^*; R_K, B_K) + 32H^2\log(K/\delta').$$

If the above conditions hold, we have

$$|f^k_{\mathtt{lin},h}(z) - \widetilde{f}(z)| \leq \beta\|\varphi(z;W^{(0)})\|_{(\widetilde{\Lambda}^k_h)^{-1}} \leq w^k_h + \beta(3K^{7/12}H^{1/6}m^{-1/12}\log^{1/4} m),$$

where the inequality is due to (37). Since $f^k_{\mathtt{lin},h}(z) \in [0,H]$ and $\widetilde{f}(z) \in [0,H]$, thus we have $|f^k_{\mathtt{lin},h}(z) - \widetilde{f}(z)| \leq H$, which further gives

$$\begin{aligned}|f^k_{\mathtt{lin},h}(z) - \widetilde{f}(z)| &\leq \min\{w^k_h, H\} + \beta(3K^{7/12}H^{1/6}m^{-1/12}\log^{1/4} m)\\ &= u^k_h + \beta(3K^{7/12}H^{1/6}m^{-1/12}\log^{1/4} m).\end{aligned} \tag{39}$$

Now we combine (34) and (39) as well as (31) and obtain

$$\begin{aligned}&|\mathbb{P}_h V^k_{h+1}(z) - f^k_h(z)|\\ &\quad\leq |\mathbb{P}_h V^k_{h+1}(z) - \widetilde{f}(z)| + |f^k_h(z) - f^k_{\mathtt{lin},h}(z)| + |f^k_{\mathtt{lin},h}(z) - \widetilde{f}(z)|\\ &\quad\leq 10C_{\mathtt{act}}R_Q H\sqrt{\log(mKH)/m} + 4\digamma K^{5/3}H^{4/3}m^{-1/6}\sqrt{\log m} + u^k_h + \beta(3K^{7/12}H^{1/6}m^{-1/12}\log^{1/4} m)\\ &\quad\leq u^k_h + \beta(5K^{7/12}H^{1/6}m^{-1/12}\log^{1/4} m),\end{aligned}$$

with m are sufficiently. We also have $\Big|\|\varphi(z;W^k_h)\|_{(\Lambda^k_h)^{-1}} - \|\varphi(z;W^{(0)})\|_{(\widetilde{\Lambda}^k_h)^{-1}}\Big| \leq \iota$ according to (37). The above inequalities hold with probability at least $1 - 2/m^2 - \delta'$ by union bound. This completes the proof. □

Lemma C.3. *Conditioned on the event $\mathcal{E}$ defined in Lemma C.2, with probability at least $1-\delta'$, we have*

$$\begin{aligned}&\sum_{k=1}^K V^*_1(s_1, r^k) \leq \sum_{k=1}^K V^k_1(s_1) + \beta HK\iota,\\ &\sum_{k=1}^K V^k_1(s_1) \leq \mathcal{O}\Big(\sqrt{H^3K\log(1/\delta')} + \beta\sqrt{H^2K\cdot\Gamma(K,\lambda;\ker_m)}\Big) + \beta HK\iota,\end{aligned}$$

where $\iota = 5K^{7/12}H^{1/6}m^{-1/12}\log^{1/4} m$.

Proof. We first show the first inequality in this lemma. We prove $V^*_h(s,r^k) \leq V^k_h(s) + (H+1-h)\iota$ for all $s\in\mathcal{S}, h\in[H]$ by induction. When $h = H+1$, we know $V^*_{H+1}(s,r^k) = 0$ and $V^k_{H+1}(s) = 0$ such that $V^*_{H+1}(s,r^k) \leq V^k_{H+1}(s_1)$. Now we assume that $V^*_{h+1}(s,r^k) \leq V^k_{h+1}(s) + (H-h)\beta\iota$. Then, conditioned on the event $\mathcal{E}$ defined in Lemma B.4, for all $s\in\mathcal{S}$, $(h,k)\in[H]\times[K]$, we further have

$$\begin{aligned}&Q^*_h(s,a,r^k) - Q^k_h(s,a)\\ &\quad= r^k_h(s,a) + \mathbb{P}_h V^*_{h+1}(s,a,r^k) - \min\{[r^k_h(s,a) + f^k_h(s,a) + u^k_h(s,a)], H\}^+\\ &\quad\leq \min\{[\mathbb{P}_h V^*_{h+1}(s,a,r^k) - f^k_h(s,a) - u^k_h(s,a)], 0\}\\ &\quad\leq \min\{[\mathbb{P}_h V^k_{h+1}(s,a) + \beta(H-h)\iota - f^k_h(s,a) - u^k_h(s,a)], 0\}\\ &\quad\leq \beta(H+1-h)\iota,\end{aligned} \tag{40}$$

where the first inequality is due to $0 \leq r^k_h(s,a) + \mathbb{P}_h V^*_{h+1}(s,a,r^k) \leq H$ and $\min\{x,y\}^+ \geq \min\{x,y\}$, the second inequality is by the assumption that $V^*_{h+1}(s,r^k) \leq V^k_{h+1}(s) + (H-h)\beta\iota$, the last inequality is by Lemma C.2 such that $|\mathbb{P}_h V^k_{h+1}(s,a) - f^k_h(s,a)| \leq u^k_h(s,a) + \beta\iota$ holds for any $(s,a)\in\mathcal{S}\times\mathcal{A}$ and $(k,h)\in[K]\times[H]$. The above inequality (40) further leads to

$$V^*_h(s,r^k) = \max_{a\in\mathcal{A}} Q^*_h(s,a,r^k) \leq \max_{a\in\mathcal{A}} Q^k_h(s,a) = V^k_h(s) + \beta(H+1-h)\iota.$$

Therefore, we obtain that conditioned on event $\mathcal{E}$, we have

$$\sum_{k=1}^K V^*_1(s,r^k) \leq \sum_{k=1}^K V^k_1(s) + \beta HK\iota.$$

Next, we prove the second inequality in this lemma. Conditioned on $\mathcal{E}$ defined in Lemma C.2, we have

$$\begin{aligned}V_h^k(s_h^k) = Q_h^k(s_h^k, a_h^k) &\leq \max\{0, f_h^k(s_h^k, a_h^k) + r_h^k(s_h^k, a_h^k) + u_h^k(s_h^k, a_h^k)\}\\ &\leq \mathbb{P}_h V_{h+1}^k(s_h^k, a_h^k) + u_h^k(s_h^k, a_h^k) + r_h^k(s_h^k, a_h^k) + u_h^k(s_h^k, a_h^k)\\ &\leq \zeta_h^k + V_{h+1}^k(s_{h+1}^k) + (2 + 1/H)\beta\|\varphi(s_h^k, a_h^k; W_h^k)\|_{(\Lambda_h^k)^{-1}},\end{aligned}$$

where we define

$$\zeta_h^k := \mathbb{P}_h V_{h+1}^k(s_h^k, a_h^k) - V_{h+1}^k(s_{h+1}^k).$$

Recursively applying the above inequality gives

$$V_1^k(s_1) \leq \sum_{h=1}^H \zeta_h^k + (2 + 1/H)\beta \sum_{h=1}^H \|\varphi(s_h^k, a_h^k; W_h^k)\|_{(\Lambda_h^k)^{-1}},$$

where we use the fact that $V_{H+1}^k(\cdot) = 0$. Taking summation on both sides of the above inequality, we have

$$\sum_{k=1}^K V_1^k(s_1) = \sum_{k=1}^K \sum_{h=1}^H \zeta_h^k + (2 + 1/H)\beta \sum_{k=1}^K \sum_{h=1}^H \|\varphi(s_h^k, a_h^k; W_h^k)\|_{(\Lambda_h^k)^{-1}}.$$

By Azuma-Hoeffding inequality, with probability at least $1 - \delta'$, the following inequalities hold

$$\sum_{k=1}^K \sum_{h=1}^H \zeta_h^k \leq \mathcal{O}\left(\sqrt{H^3 K \log \frac{1}{\delta'}}\right).$$

On the other hand, by Lemma F.2, we have

$$\begin{aligned}\sum_{k=1}^K \sum_{h=1}^H \|\phi(s_h^k, a_h^k)\|_{(\Lambda_h^k)^{-1}} &= \sum_{k=1}^K \sum_{h=1}^H \sqrt{\varphi(s_h^k, a_h^k; W_h^k)^\top (\Lambda_h^k)^{-1} \phi(s_h^k, a_h^k; W_h^k)}\\ &\leq \sum_{k=1}^K \sum_{h=1}^H \sqrt{\varphi(s_h^k, a_h^k; W^{(0)})^\top (\widetilde{\Lambda}_h^k)^{-1} \varphi(s_h^k, a_h^k; W^{(0)})} + HK\iota\\ &\leq \sum_{h=1}^H \sqrt{K \sum_{k=1}^K \varphi(s_h^k, a_h^k; W^{(0)})^\top (\widetilde{\Lambda}_h^k)^{-1} \varphi(s_h^k, a_h^k; W^{(0)}))} + HK\iota\\ &= 2H\sqrt{K \cdot \Gamma(K, \lambda; \ker_m)} + HK\iota.\end{aligned}$$

where the first inequality is due to Lemma C.2, the second inequality is by Jensen's inequality. Thus, conditioned on event $\mathcal{E}$, we obtain that with probability at least $1 - \delta'$, there is

$$\sum_{k=1}^K V_1^k(s_1) \leq \mathcal{O}\left(\sqrt{H^3 K \log(1/\delta')} + \beta\sqrt{H^2 K \cdot \Gamma(K, \lambda; \ker)}\right) + \beta HK\iota,$$

which completes the proof. □

Lemma C.4. *We define the event $\widetilde{\mathcal{E}}$ as that the following inequality holds $\forall(s, a) \in \mathcal{S} \times \mathcal{A}, \forall h \in [H]$,*

$$\begin{aligned}&|\mathbb{P}_h V_{h+1}(s, a) - f_h(s, a)| \leq u_h(s, a) + \beta\iota,\\ &\left|\|\varphi(z; W_h)\|_{(\Lambda_h)^{-1}} - \|\varphi(z; W^{(0)})\|_{(\widetilde{\Lambda}_h)^{-1}}\right| \leq \iota,\end{aligned}$$

where $\iota = 5K^{7/12}H^{1/6}m^{-1/12}\log^{1/4} m$ and we define

$$\Lambda_h = \sum_{\tau=1}^K \varphi(z_h^\tau; W_h^k)\varphi(z_h^\tau; W_h^k)^\top + \lambda \cdot I, \quad \widetilde{\Lambda}_h = \sum_{\tau=1}^K \varphi(z_h^\tau; W^{(0)})\varphi(z_h^\tau; W^{(0)})^\top + \lambda \cdot I.$$

Setting $\beta = \widetilde{B}_K$, $\widetilde{R}_K = H\sqrt{K}$, $\varsigma^* = H/K$, *and* $\lambda = \digamma^2(1+1/K)$, $\varsigma^* = H/K$, *if we set*

$$\beta^2 \geq 8R_Q^2H^2(1+\sqrt{\lambda/d})^2 + 32H^2\Gamma(K,\lambda;\ker_m) + 80H^2 + 32H^2\log\mathcal{N}_\infty(\varsigma^*;\widetilde{R}_K,\widetilde{B}_K) + 32H^2\log(K/\delta'),$$

and also

$$m = \Omega(K^{19}H^{14}\log^3 m),$$

then we have that with probability at least $1-2/m^2-\delta'$, *the event* $\widetilde{\mathcal{E}}$ *happens, i.e.,*

$$\Pr(\widetilde{\mathcal{E}}) \geq 1-2/m^2-\delta'.$$

Proof. The proof of this lemma exactly follows our proof of Lemma C.2. There are several minor differences here. In the proof of this lemma, we set $\widetilde{B}_K = \beta$ instead of $(1+1/H)\beta$ due to the structure of the planning phase. Moreover, we use $\mathcal{N}_\infty(\epsilon; R_K, B_K)$ to denote covering number of the Q function class $\overline{\mathcal{Q}}(r_h, R_K, B_K)$. Since the covering numbers of $\overline{\mathcal{Q}}(r_h, R_K, B_K)$ and $\overline{\mathcal{Q}}(\mathbf{0}, R_K, B_K)$ are the same where the former one only has an extra bias r_h, we use the same notation $\mathcal{N}_\infty(\epsilon; R_K, B_K)$ to denote their covering number. Then, the rest of this proof can be completed by using the same argument as the proof of Lemma C.2. □

Lemma C.5. *Conditioned on the event* $\widetilde{\mathcal{E}}$ *as defined in Lemma C.4, we have*

$$V_h^*(s,r) \leq V_h(s) + (H+1-h)\beta\iota, \forall s\in\mathcal{S}, \forall h\in[H],$$
$$V_h(s) \leq r_h(s,\pi_h(s)) + \mathbb{P}_hV_{h+1}(s,\pi_h(s)) + 2u_h(s,\pi_h(s)) + \beta\iota, \forall s\in\mathcal{S}, \forall h\in[H],$$

where $\pi_h(s) = \operatorname{argmax}_{a\in\mathcal{A}} Q_h(s,a)$.

Proof. We first prove the first inequality in this lemma by induction. For $h = H+1$, we have $V_{H+1}^*(s,r) = V_{H+1}(s) = 0$ for any $s\in\mathcal{S}$. Then, we assume that $V_{h+1}^*(s,r) \leq V_{h+1}(s) + (H-h)\beta\iota$. Thus, conditioned on the event $\widetilde{\mathcal{E}}$ as defined in Lemma C.4, we have

$$\begin{aligned}
&Q_h^*(s,a,r) - Q_h(s,a)\\
&\quad= r_h(s,a) + \mathbb{P}_hV_{h+1}^*(s,a,r) - \min\{[r_h(s,a)+f_h(s,a)+u_h(s,a)], H\}^+\\
&\quad\leq \min\{[\mathbb{P}_hV_{h+1}^*(s,a,r) - f_h(s,a) - u_h(s,a)], 0\}\\
&\quad\leq \min\{[\mathbb{P}_hV_{h+1}(s,a) + (H-h)\beta\iota - f_h(s,a) - u_h(s,a)], 0\}\\
&\quad\leq (H+1-h)\beta\iota,
\end{aligned}$$

where the first inequality is due to $0 \leq r_h(s,a) + \mathbb{P}_hV_{h+1}^*(s,a,r) \leq H$ and $\min\{x,y\}^+ \geq \min\{x,y\}$, the second inequality is by the assumption that $V_{h+1}^*(s,a,r) \leq V_{h+1}(s,a) + (H-h)\beta\iota$, the last inequality is by Lemma C.4 such that $|\mathbb{P}_hV_{h+1}(s,a) - f_h(s,a)| \leq u_h(s,a) + \beta\iota$ holds for any $(s,a)\in\mathcal{S}\times\mathcal{A}$ and $(k,h)\in[K]\times[H]$. The above inequality further leads to

$$V_h^*(s,r) = \max_{a\in\mathcal{A}} Q_h^*(s,a,r) \leq \max_{a\in\mathcal{A}} Q_h(s,a) + (H+1-h)\beta\iota = V_h(s) + (H+1-h)\beta\iota.$$

Therefore, we have

$$V_h^*(s,r) \leq V_h(s) + (H+1-h)\beta\iota, \forall h\in[H], \forall s\in\mathcal{S}.$$

We further prove the second inequality in this lemma. We have

$$\begin{aligned}
Q_h(s,a) &= \min\{[r_h(s,a)+f_h(s,a)+u_h(s,a)], H\}^+\\
&\leq \min\{[r_h(s,a)+\mathbb{P}_hV_{h+1}(s,a)+2u_h(s,a)+\beta\iota], H\}^+\\
&\leq r_h(s,a)+\mathbb{P}_hV_{h+1}(s,a)+2u_h(s,a)+\beta\iota,
\end{aligned}$$

where the first inequality is also by Lemma C.4 such that $|\mathbb{P}_hV_{h+1}(s,a) - f_h(s,a)| \leq u_h(s,a) + \beta\iota$, and the last inequality is because of the non-negativity of $r_h(s,a) + \mathbb{P}_hV_{h+1}(s,a) + 2u_h(s,a) + \beta\iota$. Therefore, we have

$$V_h(s) = \max_{a\in\mathcal{A}} Q_h(s,a) = Q_h(s,\pi_h(s)) \leq r_h(s,\pi_h(s)) + \mathbb{P}_hV_{h+1}(s,\pi_h(s)) + 2u_h(s,\pi_h(s)) + \beta\iota.$$

This completes the proof. □

Lemma C.6. *With the exploration and planning phases, conditioned on the event $\mathcal{E}$ and $\widetilde{\mathcal{E}}$, we have the following inequality*

$$K \cdot V_1^*(s_1, u/H) \leq \sum_{k=1}^{K} V_1^*(s_1, r^k) + 2K\beta\iota,$$

where $\iota = 5K^{7/12}H^{1/6}m^{-1/12}\log^{1/4} m$.

Proof. The bonus for the planning phase is $u_h(s,a) = \beta\|\varphi(s,a;W_h)\|_{\Lambda_h^{-1}}$. We also have $H \cdot r_h^k(s,a) = u_h^k(s,a) = \beta\|\varphi(s,a;W_h^k)\|_{(\Lambda_h^k)^{-1}}$. Conditioned on the event $\mathcal{E}$ and $\widetilde{\mathcal{E}}$, according to Lemmas C.2 and C.4, we have

$$\begin{aligned}
\Big|\|\varphi(s,a;W_h^k)\|_{(\Lambda_h^k)^{-1}} - \|\varphi(s,a;W^{(0)})\|_{(\widetilde{\Lambda}_h^k)^{-1}}\Big| &\leq \iota,\\
\Big|\|\varphi(s,a;W_h)\|_{(\Lambda_h)^{-1}} - \|\varphi(s,a;W^{(0)})\|_{(\widetilde{\Lambda}_h)^{-1}}\Big| &\leq \iota,
\end{aligned}$$

such that

$$\begin{aligned}
u_h(s,a) - \beta\iota &\leq \beta\|\varphi(s,a;W^{(0)})\|_{(\widetilde{\Lambda}_h)^{-1}},\\
\beta\iota + u_h^k(s,a) = \beta\iota + H \cdot r_h^k(s,a) &\geq \beta\|\varphi(s,a;W^{(0)})\|_{(\widetilde{\Lambda}_h^k)^{-1}}.
\end{aligned}$$

Moreover, since

$$\|\varphi(s,a;W^{(0)})\|_{(\widetilde{\Lambda}_h)^{-1}} = \sqrt{\varphi(s,a;W^{(0)})^\top \left[\lambda I + \sum_{\tau=1}^{K} \varphi(s_h^\tau, a_h^\tau; W^{(0)})\varphi(s_h^\tau, a_h^\tau; W^{(0)})^\top\right]^{-1} \varphi(s,a;W^{(0)})},$$

and also

$$\|\varphi(s,a;W^{(0)})\|_{(\widetilde{\Lambda}_h^k)^{-1}} = \sqrt{\varphi(s,a;W^{(0)})^\top \left[\lambda I + \sum_{\tau=1}^{k-1} \varphi(s_h^\tau, a_h^\tau; W^{(0)})\varphi(s_h^\tau, a_h^\tau; W^{(0)})^\top\right]^{-1} \varphi(s,a;W^{(0)})}.$$

Since $k-1 \leq K$ and $x^\top\phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top x = [x^\top\phi(s_h^\tau, a_h^\tau)]^2 \geq 0, \forall x$, then we know that

$$\widetilde{\Lambda}_h = \lambda I + \sum_{\tau=1}^{K} \varphi(s_h^\tau, a_h^\tau; W^{(0)})\varphi(s_h^\tau, a_h^\tau; W^{(0)})^\top \succcurlyeq \lambda I + \sum_{\tau=1}^{k-1} \varphi(s_h^\tau, a_h^\tau; W^{(0)})\varphi(s_h^\tau, a_h^\tau; W^{(0)})^\top = \widetilde{\Lambda}_h^k.$$

The above relation further implies that $\widetilde{\Lambda}_h^{-1} \preccurlyeq (\widetilde{\Lambda}_h^k)^{-1}$ such that

$$\varphi(s,a;W^{(0)})^\top\widetilde{\Lambda}_h^{-1}\varphi(s,a;W^{(0)}) \leq \varphi(s,a;W^{(0)})^\top(\widetilde{\Lambda}_h^k)^{-1}\varphi(s,a;W^{(0)}).$$

Thus, we have

$$u_h(s,a) - \beta\iota \leq H \cdot r_h^k(s,a) + \beta\iota,$$

such that

$$V_1^*(s_1, u/H) \leq V_1^*(s_1, r^k) + 2\beta\iota,$$

and thus

$$K \cdot V_1^*(s_1, u/H) \leq \sum_{k=1}^{K} V_1^*(s_1, r^k) + 2K\beta\iota.$$

This completes the proof. □

C.2. Proof of Theorem 3.5

Proof. Conditioned on the event $\mathcal{E}$ defined in Lemma C.2 and the event $\widetilde{\mathcal{E}}$ defined in Lemma C.4, we have

$$V_1^*(s_1, r) - V_1^\pi(s_1, r) \le V_1(s_1) - V_1^\pi(s_1, r) + H\beta\iota, \tag{41}$$

where the inequality is by Lemma C.5. Further by this lemma, we have

$$\begin{aligned} V_h(s) - V_h^\pi(s, r) &\le r_h(s, \pi_h(s)) + \mathbb{P}_h V_{h+1}(s, \pi_h(s)) + 2u_h(s, \pi_h(s)) - Q_h^\pi(s, \pi_h(s), r) + \beta\iota \\ &= r_h(s, \pi_h(s)) + \mathbb{P}_h V_{h+1}(s, \pi_h(s)) + 2u_h(s, \pi_h(s)) - r_h(s, \pi_h(s)) - \mathbb{P}_h V_{h+1}^\pi(s, \pi_h(s), r) + \beta\iota \\ &= \mathbb{P}_h V_{h+1}(s, \pi_h(s)) - \mathbb{P}_h V_{h+1}^\pi(s, \pi_h(s), r) + 2u_h(s, \pi_h(s)) + \beta\iota. \end{aligned}$$

Recursively applying the above inequality and making use of $V_{H+1}^\pi(s, r) = V_{H+1}(s) = 0$ gives

$$\begin{aligned} V_1(s_1) - V_1^\pi(s_1, r) &\le \mathbb{E}_{\forall h\in[H]:\, s_{h+1}\sim\mathbb{P}_h(\cdot|s_h,\pi_h(s_h))} \left[\sum_{h=1}^H 2u_h(s_h, \pi_h(s_h))\middle| s_1\right] + H\beta\iota \\ &= 2H \cdot V_1^\pi(s_1, u/H) + H\beta\iota. \end{aligned}$$

Combining with (41) gives

$$\begin{aligned} V_1^*(s_1, r) - V_1^\pi(s_1, r) &\le 2H \cdot V_1^\pi(s_1, u/H) + 2H\beta\iota \le \frac{2H}{K}\sum_{k=1}^K V_1^*(s_1, r^k) + 4H\beta\iota \\ &\le \frac{2H}{K}\mathcal{O}\left(\sqrt{H^3K\log(1/\delta')} + \beta\sqrt{H^2K\cdot\Gamma(K,\lambda;\ker_m)}\right) + H^2\beta\iota + 4H\beta\iota \\ &\le \mathcal{O}\left([\sqrt{H^5\log(1/\delta')} + \beta\sqrt{H^4\cdot\Gamma(K,\lambda;\ker_m)}]/\sqrt{K} + H^2\beta\iota\right), \end{aligned}$$

where the second inequality is due to Lemma C.6 and the last inequality is by Lemma C.3.

By union bound, we have $P(\mathcal{E}\wedge\widetilde{\mathcal{E}}) \ge 1 - 2\delta' - 4/m^2$. Therefore, by setting $\delta' = 1/(4K^2H^2)$, we obtain that with probability at least $1 - 1/(2K^2H^2) - 4/m^2$

$$\begin{aligned} V_1^*(s_1, r) - V_1^\pi(s_1, r) &\le \mathcal{O}\left([\sqrt{H^5\log(1/\delta')} + \beta\sqrt{H^4\cdot\Gamma(K,\lambda;\ker_m)}]/\sqrt{K} + H^2\beta\iota\right) \\ &\le \mathcal{O}\left(\beta\sqrt{H^4[\Gamma(K,\lambda;\ker_m) + \log(KH)]}/\sqrt{K} + H^2\beta\iota\right). \end{aligned}$$

where the last inequality is due to $\beta \ge H$. Note that $\mathcal{E}\wedge\widetilde{\mathcal{E}}$ happens when the following two conditions are satisfied, i.e.,

$$\begin{aligned} \beta^2 &\ge 8R_Q^2H^2(1+\sqrt{\lambda/d})^2 + 32H^2\Gamma(K,\lambda;\ker_m) + 80H^2 + 32H^2\log\mathcal{N}_\infty(\varsigma^*; R_K, B_K) + 32H^2\log(K/\delta'), \\ \beta^2 &\ge 8R_Q^2H^2(1+\sqrt{\lambda/d})^2 + 32H^2\Gamma(K,\lambda;\ker_m) + 80H^2 + 32H^2\log\mathcal{N}_\infty(\varsigma^*; \widetilde{R}_K, \widetilde{B}_K) + 32H^2\log(K/\delta'), \end{aligned}$$

where $\beta = \widetilde{B}_K$,$(1+1/H)\beta = B_K$, $\lambda = \mathit{\Gamma}(1+1/K)$, $\widetilde{R}_K = R_K = H\sqrt{K}$, and $\varsigma^* = H/K$. The above inequalities hold if we further let β satisfy

$$\beta^2 \ge 8R_Q^2H^2(1+\sqrt{\lambda/d})^2 + 32H^2\Gamma(K,\lambda;\ker_m) + 80H^2 + 32H^2\log\mathcal{N}_\infty(\varsigma^*; R_K, 2\beta) + 96H^2\log(2KH),$$

since $2\beta \ge (1+1/H)\beta \ge \beta$ such that $\mathcal{N}_\infty(\varsigma^*; R_K, 2\beta) \ge \mathcal{N}_\infty(\varsigma^*; R_K, B_K) \ge \mathcal{N}_\infty(\varsigma^*; \widetilde{R}_K, \widetilde{B}_K)$. This completes the proof. □

D. Proofs for Markov Game Setting with Kernel Function Approximation

D.1. Lemmas

Lemma D.1. *We define the event $\mathcal{E}$ as that the following inequality holds $\forall(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}, \forall(h,k)\in[H]\times[K]$,*

$$|\mathbb{P}_h V_{h+1}^k(s,a,b) - f_h^k(s,a,b)| \le u_h^k(s,a,b),$$

where $u_h^k(s,a,b) = \min\{w_h^k(s,a,b), H\}^+$ *with* $w_h^k(s,a,b) = \beta\lambda^{-1/2}[\ker(z,z) - \psi_h^k(s,a,b)^\top(\lambda I + \mathcal{K}_h^k)^{-1}\psi_h^k(s,a,b)]^{1/2}$ *with* $z = (s,a,b)$*, and*

$$\begin{aligned}
&\psi_h^k(z) = \Phi_h^k\phi(z) = [\ker(z, z_h^1), \cdots, \ker(z, z_h^{k-1})]^\top,\\
&\Phi_h^k = [\phi(z_h^1), \phi(z_h^2), \cdots, \phi(z_h^{k-1})]^\top,\\
&\mathbf{y}_h^k = [V_{h+1}^k(s_{h+1}^1), V_{h+1}^k(s_{h+1}^2), \cdots, V_{h+1}^k(s_{h+1}^{k-1})]^\top,\\
&\mathcal{K}_h^k = \Phi_h^k(\Phi_h^k)^\top = \begin{bmatrix} \ker(z_h^1, z_h^1) & \dots & \ker(z_h^1, z_h^{k-1}) \\ \vdots & \ddots & \vdots \\ \ker(z_h^{k-1}, z_h^1) & \dots & \ker(z_h^{k-1}, z_h^{k-1}) \end{bmatrix},
\end{aligned}$$

Thus, setting $\beta = B_K/(1+1/H)$*, if* B_K *satisfies*

$$16H^2\big[R_Q^2 + 2\Gamma(K,\lambda;\ker) + 5 + \log\mathcal{N}_\infty(\varsigma^*; R_K, B_K) + 2\log(K/\delta')\big] \le B_K^2, \forall h \in [H],$$

then we have that with probability at least $1-\delta'$*, the event* $\mathcal{E}$ *happens, i.e.,*

$$\Pr(\mathcal{E}) \ge 1 - \delta'.$$

Proof. According to the exploration algorithm for the game, we can see that by letting $\boldsymbol{a} = (a,b)$ be an action in the space $\mathcal{A}\times\mathcal{B}$, Algorithm 3 reduces to Algorithm 1 with the action space $\mathcal{A}\times\mathcal{B}$ and state space $\mathcal{S}$. Now, we also have a transition in the form of $P_h(s|\boldsymbol{a})$ and a product policy $(\pi_h^k \otimes \nu_h^k)(s)$ such that $\boldsymbol{a} \sim (\pi_h^k \otimes \nu_h^k)(s)$ at state $s \in \mathcal{S}$ for all $(h,k) \in [H]\times[K]$. Similarly, we have $Q_h^k(s,a,b) = Q_h^k(s,\boldsymbol{a})$ and $V_h^k(s,a,b) = V_h^k(s,\boldsymbol{a})$ as well as $u_h^k(s,a,b) = u_h^k(s,\boldsymbol{a})$ and $u_h^k(s,a,b) = u_h^k(s,\boldsymbol{a})$ and $r_h^k(s,a,b) = r_h^k(s,\boldsymbol{a})$. Thus, we can simply apply the proof of Lemma B.4 and obtain the proof for this lemma. This completes the proof. □

Lemma D.2. *Conditioned on the event* $\mathcal{E}$ *defined in Lemma D.1, with probability at least* $1-\delta'$*, we have*

$$\sum_{k=1}^K V_1^*(s_1, r^k) \le \sum_{k=1}^K V_1^k(s_1) \le \mathcal{O}\left(\sqrt{H^3K\log(1/\delta')} + \beta\sqrt{H^2K\cdot\Gamma(K,\lambda;\ker)}\right).$$

Proof. By the reduction of Algorithm 3 to Algorithm 1, we can apply the same proof as the one for Lemma B.5, which completes the proof. □

Lemma D.3. *We define the event* $\widetilde{\mathcal{E}}$ *as that the following inequality holds* $\forall(s,a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}, \forall h \in [H]$*,*

$$|\mathbb{P}_h\overline{V}_{h+1}(s,a,b) - \overline{f}_h(s,a,b)| \le u_h(s,a,b), \tag{42}$$
$$|\mathbb{P}_h\underline{V}_{h+1}(s,a,b) - \underline{f}_h(s,a,b)| \le u_h(s,a,b), \tag{43}$$

where $u_h(s,a,b) = \min\{w_h(s,a,b), H\}^+$ *with* $z = (s,a,b)$*,* $w_h(s,a,b) = \beta\lambda^{-1/2}[\ker(z,z) - \psi_h(s,a,b)^\top(\lambda I + \mathcal{K}_h)^{-1}\psi_h(s,a,b)]^{1/2}$*,* $\mathcal{K}_h = \Phi_h\Phi_h^\top$*, and* $\psi_h(s,a) = \Phi_h\phi(s,a)$ *with* $\Phi_h = [\phi(z_h^1), \phi(z_h^2), \cdots, \phi(z_h^K)]^\top$*.*

Thus, setting $\beta = \widetilde{B}_K$*, if* $\widetilde{B}_K$ *satisfies*

$$4H^2\big[R_Q^2 + 2\Gamma(K,\lambda;\ker) + 5 + \log\mathcal{N}_\infty(\varsigma^*; \widetilde{R}_K, \widetilde{B}_K) + 2\log(2K/\delta')\big] \le \widetilde{B}_K^2, \forall h \in [H],$$

then we have that with probability at least $1-\delta'$*, the event* $\mathcal{E}$ *happens, i.e.,*

$$\Pr(\widetilde{\mathcal{E}}) \ge 1 - \delta'.$$

Proof. According to the construction of u_h and $\overline{f}_h$, the proof for the the first inequality in this lemma is nearly the same as the proof of Lemma B.6 but one difference for computing the covering number of the value function space. Specifically, we have the function class for $\overline{V}_h$ which is

$$\overline{\mathcal{V}}(r_h, \widetilde{R}_K, \widetilde{B}_K) = \{V : V(\cdot) = \max_{a\sim\pi'}\min_{b\sim\nu'}\mathbb{E}_{\pi',\nu'}Q(\cdot,a,b) \text{ with } Q \in \overline{\mathcal{Q}}(r_h, \widetilde{R}_K, \widetilde{B}_K)\}.$$

By Lemma F.1 with $\delta'/2$, we have

$$\begin{aligned}
&\left\|\sum_{\tau=1}^{K}\phi(s_h^\tau,a_h^\tau,b_h^\tau)[\overline{V}_{h+1}(s_{h+1}^\tau)-\mathbb{P}_h\overline{V}_{h+1}(s_h^\tau,a_h^\tau,b_h^\tau)]\right\|^2_{(\Lambda_h)^{-1}}\\
&\quad\le \sup_{V\in\overline{\mathcal{V}}(r_h,\widetilde{R}_K,\widetilde{B}_K)}\left\|\sum_{\tau=1}^{K}\phi(s_h^\tau,a_h^\tau,b_h^\tau)[\overline{V}(s_{h+1}^\tau)-\mathbb{P}_h\overline{V}(s_h^\tau,a_h^\tau,b_h^\tau)]\right\|^2_{(\Lambda_h)^{-1}}\\
&\quad\le 2H^2\log\det(I+\mathcal{K}/\lambda)+2H^2K(\lambda-1)+4H^2\log(\mathcal{N}^{\overline{\mathcal{V}}}_{\mathrm{dist}}(\epsilon;\widetilde{R}_K,\widetilde{B}_K)/\delta')+8K^2\epsilon^2/\lambda\\
&\quad\le 4H^2\Gamma(K,\lambda;\ker)+10H^2+4H^2\log\mathcal{N}_\infty(\varsigma^*;\widetilde{R}_K,\widetilde{B}_K)+4H^2\log(2/\delta'),
\end{aligned}$$

where the last inequality is by setting $\lambda=1+1/K$ and $\epsilon=\varsigma^*=H/K$. Here $\mathcal{N}^{\overline{\mathcal{V}}}_{\mathrm{dist}}$ is the covering number of the function space $\overline{\mathcal{V}}$ w.r.t. the distance $\mathrm{dist}(V_1,V_2)=\sup_s|V_1(s)-V_2(s)|$, and $\mathcal{N}_\infty$ is the covering number for the function space $\overline{\mathcal{Q}}$ w.r.t. the infinity norm. In the last inequality, we also use

$$\mathcal{N}^{\overline{\mathcal{V}}}_{\mathrm{dist}}(\varsigma^*;\widetilde{R}_K,\widetilde{B}_K)\le\mathcal{N}_\infty(\varsigma^*;\widetilde{R}_K,\widetilde{B}_K),$$

which is in particular due to

$$\begin{aligned}
\mathrm{dist}(V_1,V_2)&=\sup_{s\in\mathcal{S}}|V_1(s)-V_2(s)|\\
&=\sup_{s\in\mathcal{S}}|\max_{\pi'}\min_{\nu'}\mathbb{E}_{a\sim\pi',b\sim\nu'}[Q_1(s,a,b)]-\max_{\pi''}\min_{\nu''}\mathbb{E}_{a\sim\pi'',b\sim\nu''}[Q_2(s,a,b)]|\\
&\le\sup_{s\in\mathcal{S}}\sup_{a\in\mathcal{A}}\sup_{b\in\mathcal{B}}|Q_1(s,a,b)-Q_2(s,a,b)|\\
&=\|Q_1(\cdot,\cdot,\cdot)-Q_2(\cdot,\cdot,\cdot)\|_\infty,
\end{aligned}\tag{44}$$

where we use the fact that maximin operator is the non-expansive. Thus, we have that with probability at least $1-\delta'/2$, the following inequality holds for all $k\in[K]$

$$\begin{aligned}
&\left\|\sum_{\tau=1}^{K}\phi(s_h^\tau,a_h^\tau,b_h^\tau)[\overline{V}_{h+1}(s_{h+1}^\tau)-\mathbb{P}_h\overline{V}_{h+1}(s_h^\tau,a_h^\tau,b_h^\tau)]\right\|_{\Lambda_h^{-1}}\\
&\quad\le[4H^2\Gamma(K,\lambda;\ker)+10H^2+4H^2\log\mathcal{N}_\infty(\varsigma^*;\widetilde{R}_K,\widetilde{B}_K)+4H^2\log(2K/\delta')]^{1/2}.
\end{aligned}$$

Then, the rest of the proof for (42) follows the proof of Lemma B.6.

Next, we give the proof of (43). We define another function class for $\underline{V}_h$ as

$$\underline{\mathcal{V}}(r_h,\widetilde{R}_K,\widetilde{B}_K)=\{V:V(\cdot)=\max_{a\sim\pi'}\min_{b\sim\nu'}\mathbb{E}_{\pi',\nu'}Q(\cdot,a,b)\text{ with }Q\in\underline{\mathcal{Q}}(r_h,\widetilde{R}_K,\widetilde{B}_K)\}.$$

Note that as we can show in the covering number for the function spaces $\underline{\mathcal{Q}}$ and $\overline{\mathcal{Q}}$ have the same covering number upper bound. Therefore, we use the same notation $\mathcal{N}_\infty$ for their upper bound. Thus, by the similar argument as (44), we have that with probability at least $1-\delta'/2$, the following inequality holds for all $k\in[K]$

$$\begin{aligned}
&\left\|\sum_{\tau=1}^{K}\phi(s_h^\tau,a_h^\tau,b_h^\tau)[\underline{V}_{h+1}(s_{h+1}^\tau)-\mathbb{P}_h\underline{V}_{h+1}(s_h^\tau,a_h^\tau,b_h^\tau)]\right\|_{\Lambda_h^{-1}}\\
&\quad\le[4H^2\Gamma(K,\lambda;\ker)+10H^2+4H^2\log\mathcal{N}_\infty(\varsigma^*;\widetilde{R}_K,\widetilde{B}_K)+4H^2\log(2K/\delta')]^{1/2},
\end{aligned}$$

where we use the fact that

$$\mathcal{N}^{\underline{\mathcal{V}}}_{\mathrm{dist}}(\varsigma^*;\widetilde{R}_K,\widetilde{B}_K)\le\mathcal{N}_\infty(\varsigma^*;\widetilde{R}_K,\widetilde{B}_K).$$

The rest of the proof are exactly the same as the proof of Lemma B.6.

In this lemma, we let

$$\big[2\lambda R_Q^2 H^2 + 8H^2\Gamma(K,\lambda;\ker) + 20H^2 + 4H^2\log\mathcal{N}_\infty(\varsigma^*;\widetilde{R}_K,\widetilde{B}_K) + 8H^2\log(2K/\delta')\big]^{1/2} \le \beta = \widetilde{B}_K,$$

which can be further guaranteed by

$$4H^2\big[R_Q^2 + 2\Gamma(K,\lambda;\ker) + 5 + \log\mathcal{N}_\infty(\varsigma^*;\widetilde{R}_K,\widetilde{B}_K) + 2\log(2K/\delta')\big] \le \widetilde{B}_K^2$$

as $(1+1/H)\le 2$ and $\lambda = 1+1/K \le 2$. This completes the proof. □

Lemma D.4. *Conditioned on the event $\widetilde{\mathcal{E}}$ as defined in Lemma D.3, we have*

$$V_h^\dagger(s,r) \le \overline{V}_h(s) \le \mathbb{E}_{a\sim\pi_h,b\sim\mathrm{br}(\pi)_h}[(\mathbb{P}_h\overline{V}_{h+1} + r_h + 2u_h)(s,a,b)], \forall s\in\mathcal{S}, \forall h\in[H], \tag{45}$$

$$V_h^\dagger(s,r) \ge \underline{V}_h(s) \ge \mathbb{E}_{a\sim\mathrm{br}(\nu)_h,b\sim\nu_h}[(\mathbb{P}_h\underline{V}_{h+1} - r_h - 2u_h)(s,a,b)], \forall s\in\mathcal{S}, \forall h\in[H]. \tag{46}$$

Proof. For the first inequality of (45), we can prove it by induction. We first prove the first inequality in this lemma. We prove it by induction. For $h = H+1$, by the planning algorithm, we have $V_{H+1}^\dagger(s,r) = V_{H+1}(s) = 0$ for any $s\in\mathcal{S}$. Then, we assume that $V_{h+1}^\dagger(s,r) \le \overline{V}_{h+1}(s)$. Thus, conditioned on the event $\widetilde{\mathcal{E}}$ as defined in Lemma D.3, we have

$$\begin{aligned}
&Q_h^\dagger(s,a,b,r) - \overline{Q}_h(s,a,b)\\
&\quad= r_h(s,a,b) + \mathbb{P}_hV_{h+1}^\dagger(s,a,b,r) - \min\{[r_h(s,a,b) + \overline{f}_h(s,a,b) + u_h(s,a,b)], H\}^+\\
&\quad\le \min\{[\mathbb{P}_hV_{h+1}^\dagger(s,a,b,r) - \overline{f}_h(s,a,b) - u_h(s,a,b)], 0\}\\
&\quad\le \min\{[\mathbb{P}_h\overline{V}_{h+1}(s,a,b) - \overline{f}_h(s,a,b) - u_h(s,a,b)], 0\}\\
&\quad\le 0,
\end{aligned}$$

where the first inequality is due to $0 \le r_h(s,a,b) + \mathbb{P}_hV_{h+1}^\dagger(s,a,b,r) \le H$ and $\min\{x,y\}^+ \ge \min\{x,y\}$, the second inequality is by the assumption that $V_{h+1}^\dagger(s,a,b,r) \le \overline{V}_{h+1}(s,a,b)$, the last inequality is by Lemma D.3 such that $|\mathbb{P}_h\overline{V}_{h+1}(s,a,b) - \overline{f}_h(s,a,b)| \le u_h(s,a,b)$ holds for any $(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}$ and $(k,h)\in[K]\times[H]$. Thus, the above inequality leads to

$$V_h^\dagger(s,r) = \max_{\pi_h'}\min_{\nu_h'}\mathbb{E}_{a\sim\pi_h',b\sim\nu_h'}[Q_h^\dagger(s,a,b,r)] \le \max_{\pi_h'}\min_{\nu_h'}\mathbb{E}_{a\sim\pi_h',b\sim\nu_h'}[\overline{Q}_h(s,a,b)] = \overline{V}_h(s),$$

which eventually gives

$$V_h^*(s,r) \le V_h(s), \forall h\in[H], \forall s\in\mathcal{S}.$$

To prove the second inequality of (45), we have

$$\begin{aligned}
\overline{V}_h(s) &= \min_{\nu'}\mathbb{E}_{a\sim\pi_h,b\sim\nu'}\overline{Q}_h(s,a,b)\\
&\le \mathbb{E}_{a\sim\pi_h,b\sim\mathrm{br}(\pi)_h}\overline{Q}_h(s,a,b)\\
&= \mathbb{E}_{a\sim\pi_h,b\sim\mathrm{br}(\pi)_h}\min\{(\overline{f}_h + r_h + u_h)(s,a,b), H\}^+\\
&\le \mathbb{E}_{a\sim\pi_h,b\sim\mathrm{br}(\pi)_h}\min\{(\mathbb{P}_h\overline{V}_{h+1} + r_h + 2u_h)(s,a,b), H\}^+\\
&\le \mathbb{E}_{a\sim\pi_h,b\sim\mathrm{br}(\pi)_h}[(\mathbb{P}_h\overline{V}_{h+1} + r_h + 2u_h)(s,a,b)],
\end{aligned}$$

where the first and the second equality is by the iterations in Algorithm 4, the second inequality is by Lemma D.3, and the last inequality is due to the non-negativity of $(\mathbb{P}_h\overline{V}_{h+1} + r_h + 2u_h)(s,a,b)$.

For the inequalities in (46), one can similarly adopt the argument above to give the proof. In fact, from the perspective of Player 2, this player is trying to find a policy to maximize the cumulative rewards w.r.t. a reward function $\{-r_h(s,a,b)\}_{h\in[H]}$, while the opponent (Player 1) is trying to minimize the cumulative reward w.r.t. $\{-r_h(s,a,b)\}_{h\in[H]}$. Thus, the proof of (46) exactly follows the proof of (45). This completes the proof. □

Lemma D.5. *With the exploration and planning phases, we have the following inequalities*

$$K \cdot V_1^{\pi,\mathrm{br}(\pi)}(s_1, u/H) \le \sum_{k=1}^{K} V_1^*(s_1, r^k), \quad K \cdot V_1^{\mathrm{br}(\nu),\nu}(s_1, u/H) \le \sum_{k=1}^{K} V_1^*(s_1, r^k).$$

Proof. First, we have $K \cdot V_1^{\pi,\mathrm{br}(\pi)}(s_1, u/H) \le K \cdot V_1^*(s_1, u/H)$, as well as $K \cdot V_1^{\mathrm{br}(\nu),\nu}(s_1, u/H) \le K \cdot V_1^*(s_1, u/H)$ due to the definition of $V_1^*(\cdot, u/H)$. Thus, to prove this lemma, we only need to show

$$K \cdot V_1^*(s_1, u/H) \le \sum_{k=1}^{K} V_1^*(s_1, r^k).$$

Since the constructions of u_h and r_h^k are the same as the ones for the single-agent case, similar to the proof of Lemma B.8, we have

$$u_h(s,a)/H \le r_h^k(s,a),$$

such that

$$V_1^*(s_1, u/H) \le V_1^*(s_1, r^k),$$

and thus

$$K \cdot V_1^*(s_1, u/H) \le \sum_{k=1}^{K} V_1^*(s_1, r^k).$$

Therefore, we eventually obtain

$$K \cdot V_1^{\pi,\mathrm{br}(\pi)}(s_1, u/H) \le K \cdot V_1^*(s_1, u/H) \le \sum_{k=1}^{K} V_1^*(s_1, r^k),$$
$$K \cdot V_1^{\mathrm{br}(\nu),\nu}(s_1, u/H) \le K \cdot V_1^*(s_1, u/H) \le \sum_{k=1}^{K} V_1^*(s_1, r^k).$$

This completes the proof. □

D.2. Proof of Theorem 4.1

Proof. Conditioned on the event $\mathcal{E}$ defined in Lemma D.1 and the event $\widetilde{\mathcal{E}}$ defined in Lemma D.3, we have

$$V_1^\dagger(s_1, r) - V_1^{\pi,\mathrm{br}(\pi)}(s_1, r) \le \overline{V}_1(s_1) - V_1^{\pi,\mathrm{br}(\pi)}(s_1, r), \tag{47}$$

where the inequality is by Lemma D.4. Further by this lemma, we have

$$\begin{aligned}
&\overline{V}_h(s_h) - V_h^{\pi,\mathrm{br}(\pi)}(s_h, r) \\
&\quad \le \mathbb{E}_{a_h \sim \pi_h, b_h \sim \mathrm{br}(\pi)_h}[(\mathbb{P}_h \overline{V}_{h+1} + r_h + 2u_h)(s_h, a_h, b_h)] - V_h^{\pi,\mathrm{br}(\pi)}(s_h, r) \\
&\quad = \mathbb{E}_{a_h \sim \pi_h, b_h \sim \mathrm{br}(\pi)_h}[(r_h + \mathbb{P}_h \overline{V}_{h+1} + 2u_h)(s_h, a_h, b_h) - r_h(s_h, a_h, b_h) - \mathbb{P}_h V_{h+1}^{\pi,\mathrm{br}(\pi)}(s_h, a_h, b_h, r)] \\
&\quad = \mathbb{E}_{a_h \sim \pi_h, b_h \sim \mathrm{br}(\pi)_h}[\mathbb{P}_h \overline{V}_{h+1}(s_h, a_h, b_h) - \mathbb{P}_h V_{h+1}^{\pi,\mathrm{br}(\pi)}(s_h, a_h, b_h, r) + 2u_h(s_h, a_h, b_h)] \\
&\quad = \mathbb{E}_{a_h \sim \pi_h, b_h \sim \mathrm{br}(\pi)_h, s_{h+1} \sim \mathbb{P}_h(\cdot|s_h, a_h, b_h)}[\overline{V}_{h+1}(s_{h+1}) - V_{h+1}^{\pi,\mathrm{br}(\pi)}(s_{h+1}, r) + 2u_h(s_h, a_h, b_h)].
\end{aligned}$$

Recursively applying the above inequality and making use of $\overline{V}_{H+1}(s) = V_{H+1}^{\pi,\mathrm{br}(\pi)}(s, r) = 0$ yield

$$\begin{aligned}
\overline{V}_1(s_1) - V_1^{\pi,\mathrm{br}(\pi)}(s_1, r) &\le \mathbb{E}_{\forall h \in [H]:\, a_h \sim \pi_h, b_h \sim \mathrm{br}(\pi)_h, s_{h+1} \sim \mathbb{P}_h(\cdot|s_h, a_h, b_h)} \left[\sum_{h=1}^{H} 2u_h(s_h, a_h, b_h) \middle| s_1 \right] \\
&= 2H \cdot V_1^{\pi,\mathrm{br}(\pi)}(s_1, u/H).
\end{aligned}$$

Combining this inequality with (47) gives

$$\begin{aligned} V_1^{\dagger}(s_1, r) - V_1^{\pi, \mathrm{br}(\pi)}(s_1, r) &\leq 2H \cdot V_1^{\pi, \mathrm{br}(\pi)}(s_1, u/H) \leq \frac{2H}{K} \sum_{k=1}^{K} V_1^*(s_1, r^k) \\ &\leq \frac{2H}{K} \mathcal{O}\left(\sqrt{H^3 K \log(1/\delta')} + \beta \sqrt{H^2 K \cdot \Gamma(K, \lambda; \ker)}\right) \\ &\leq \mathcal{O}\left([\sqrt{H^5 \log(1/\delta')} + \beta \sqrt{H^4 \cdot \Gamma(K, \lambda; \ker)}]/\sqrt{K}\right), \end{aligned}$$

where the second inequality is due to Lemma D.5 and the last inequality is by Lemma D.2.

Next, we prove the upper bound of the term $V_1^{\mathrm{br}(\nu), \nu}(s_1, r) - V_1^{\dagger}(s_1, r)$. Conditioned on the event $\mathcal{E}$ defined in Lemma D.1 and the event $\widetilde{\mathcal{E}}$ defined in Lemma D.3, we have

$$V_1^{\mathrm{br}(\nu), \nu}(s_1, r) - V_1^{\dagger}(s_1, r) \leq V_1^{\mathrm{br}(\nu), \nu}(s_1, r) - \underline{V}_1(s_1, r), \tag{48}$$

where the inequality is by Lemma D.4. Further by Lemma D.4, we have

$$\begin{aligned} &V_h^{\mathrm{br}(\nu), \nu}(s_h, r) - \underline{V}_h(s_h) \\ &\quad \leq V_h^{\mathrm{br}(\nu), \nu}(s_h, r) - \mathbb{E}_{a \sim \mathrm{br}(\nu)_h, b \sim \nu_h}[(\mathbb{P}_h \underline{V}_{h+1} - r_h - 2u_h)(s_h, a_h, b_h)] \\ &\quad = \mathbb{E}_{a_h \sim \mathrm{br}(\nu)_h, b_h \sim \nu_h}[\mathbb{P}_h V_{h+1}^{\mathrm{br}(\nu), \nu}(s_h, a_h, b_h, r) - \mathbb{P}_h \underline{V}_{h+1}(s_h, a_h, b_h) + 2u_h(s_h, a_h, b_h)] \\ &\quad = \mathbb{E}_{a_h \sim \mathrm{br}(\nu)_h, b_h \sim \nu_h, s_{h+1} \sim \mathbb{P}_h(\cdot | s_h, a_h, b_h)}[V_{h+1}^{\mathrm{br}(\nu), \nu}(s_{h+1}, r) - \mathbb{P}_h \underline{V}_{h+1}(s_{h+1}) + 2u_h(s_h, a_h, b_h)]. \end{aligned}$$

Recursively applying the above inequality yields

$$V_1^{\mathrm{br}(\nu), \nu}(s_1, r) - \underline{V}_1(s_h, r) \leq 2H \cdot V_1^{\mathrm{br}(\nu), \nu}(s_1, u/H).$$

Combining this inequality with (48) gives

$$\begin{aligned} V_1^{\mathrm{br}(\nu), \nu}(s_1, r) - V_1^{\dagger}(s_1, r) &\leq 2H \cdot V_1^{\mathrm{br}(\nu), \nu}(s_1, u/H) \leq \frac{2H}{K} \sum_{k=1}^{K} V_1^*(s_1, r^k) \\ &\leq \frac{2H}{K} \mathcal{O}\left(\sqrt{H^3 K \log(1/\delta')} + \beta \sqrt{H^2 K \cdot \Gamma(K, \lambda; \ker)}\right) \\ &\leq \mathcal{O}\left([\sqrt{H^5 \log(1/\delta')} + \beta \sqrt{H^4 \cdot \Gamma(K, \lambda; \ker)}]/\sqrt{K}\right), \end{aligned}$$

where the second inequality is due to Lemma D.5 and the third inequality is by Lemma D.2.

Since $\Pr(\mathcal{E} \wedge \widetilde{\mathcal{E}}) \geq 1 - 2\delta'$ by union bound, by setting $\delta' = 1/(4H^2K^2)$, we obtain that with probability at least $1 - 1/(2H^2K^2)$

$$\begin{aligned} V_1^{\dagger}(s_1, r) - V_1^{\pi, \mathrm{br}(\pi)}(s_1, r) &\leq \mathcal{O}\left([\sqrt{2H^5 \log(2HK)} + \beta \sqrt{H^4 \cdot \Gamma(K, \lambda; \ker)}]/\sqrt{K}\right), \\ V_1^{\mathrm{br}(\nu), \nu}(s_1, r) - V_1^{\dagger}(s_1, r) &\leq \mathcal{O}\left([\sqrt{2H^5 \log(2HK)} + \beta \sqrt{H^4 \cdot \Gamma(K, \lambda; \ker)}]/\sqrt{K}\right), \end{aligned}$$

such that

$$\begin{aligned} V_1^{\mathrm{br}(\nu), \nu}(s_1, r) - V_1^{\pi, \mathrm{br}(\pi)}(s_1, r) &\leq \mathcal{O}\left([\sqrt{2H^5 \log(2HK)} + \beta \sqrt{H^4 \cdot \Gamma(K, \lambda; \ker)}]/\sqrt{K}\right) \\ &\leq \mathcal{O}\left(\beta \sqrt{H^4 [\Gamma(K, \lambda; \ker) + \log(HK)]}/\sqrt{K}\right), \end{aligned}$$

where the last inequality is due to $\beta \geq H$. The event $\mathcal{E} \wedge \widetilde{\mathcal{E}}$ happens if we further let β satisfy

$$16H^2 \big[R_Q^2 + 2\Gamma(K, \lambda; \ker) + 5 + \log \mathcal{N}_\infty(\varsigma^*; R_K, 2\beta) + 6 \log(2HK)\big] \leq \beta^2, \forall h \in [H],$$

where $\lambda = 1 + 1/K$, $\widetilde{R}_K = R_K = 2H\sqrt{\Gamma(K, \lambda; \ker)}$, and $\varsigma^* = H/K$. This completes the proof. □

E. Proofs for Markov Game Setting with Neural Function Approximation

E.1. Lemmas

Lemma E.1 (Lemma C.7 of Yang et al. (2020)). *With $TH^2 = \mathcal{O}(m\log^{-6} m)$, then there exists an constant F such that the following inequalities hold with probability at least $1-1/m^2$ for any $z \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}$ and any $W \in \{W : \|W - W^{(0)}\| \le H\sqrt{K/\lambda}\}$,*

$$|f(z;W) - \varphi(z;W^{(0)})^\top (W - W^{(0)})| \le F K^{2/3} H^{4/3} m^{-1/6}\sqrt{\log m},$$
$$\|\varphi(z;W) - \varphi(z;W^{(0)})\| \le F (KH^2/m)^{1/6}\sqrt{\log m}, \qquad \|\varphi(z;W)\| \le F,$$

with $F \ge 1$.

Lemma E.2. *We define the event $\mathcal{E}$ as that the following inequality holds $\forall z = (s,a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}, \forall (h,k) \in [H]\times[K]$,*

$$|\mathbb{P}_h V_{h+1}^k(s,a,b) - f_h^k(s,a,b)| \le u_h^k(s,a,b) + \beta\iota,$$
$$\left|\|\varphi(z;W_h^k)\|_{(\Lambda_h^k)^{-1}} - \|\varphi(z;W^{(0)})\|_{(\widetilde{\Lambda}_h^k)^{-1}}\right| \le \iota,$$

where $\iota = 5K^{7/12}H^{1/6}m^{-1/12}\log^{1/4} m$ and we define

$$\Lambda_h^k = \sum_{\tau=1}^{k-1} \varphi(z_h^\tau; W_h^k)\varphi(z_h^\tau; W_h^k)^\top + \lambda\cdot I, \quad \widetilde{\Lambda}_h^k = \sum_{\tau=1}^{k-1} \varphi(z_h^\tau; W^{(0)})\varphi(z_h^\tau; W^{(0)})^\top + \lambda\cdot I.$$

Setting $(1+1/H)\beta = B_K$, $R_K = H\sqrt{K}$, $\varsigma^* = H/K$, and $\lambda = F^2(1+1/K)$, $\varsigma^* = H/K$, if we set

$$\beta^2 \ge 8R_Q^2 H^2(1+\sqrt{\lambda/d})^2 + 32H^2\Gamma(K,\lambda;\ker_m) + 80H^2 + 32H^2\log\mathcal{N}_\infty(\varsigma^*; R_K, B_K) + 32H^2\log(K/\delta'),$$

and also

$$m = \Omega(K^{19}H^{14}\log^3 m),$$

then we have that with probability at least $1-2/m^2-\delta'$, the event $\mathcal{E}$ happens, i.e.,

$$\Pr(\mathcal{E}) \ge 1 - 2/m^2 - \delta'.$$

Proof. By letting $\boldsymbol{a} = (a,b)$ be an action in the space $\mathcal{A}\times\mathcal{B}$, Algorithm 3 reduces to Algorithm 1 with the action space $\mathcal{A}\times\mathcal{B}$ and state space $\mathcal{S}$. We have $Q_h^k(s,a,b) = Q_h^k(s,\boldsymbol{a})$, $V_h^k(s,a,b) = V_h^k(s,\boldsymbol{a})$, $u_h^k(s,a,b) = u_h^k(s,\boldsymbol{a})$, $u_h^k(s,a,b) = u_h^k(s,\boldsymbol{a})$ and $r_h^k(s,a,b) = r_h^k(s,\boldsymbol{a})$. Simply applying the proof of Lemma C.2 yields the proof for this lemma. □

Lemma E.3. *Conditioned on the event $\mathcal{E}$ defined in Lemma E.2, with probability at least $1-\delta'$, we have*

$$\sum_{k=1}^K V_1^*(s_1, r^k) \le \sum_{k=1}^K V_1^k(s_1) + \beta HK\iota,$$
$$\sum_{k=1}^K V_1^k(s_1) \le \mathcal{O}\left(\sqrt{H^3K\log(1/\delta')} + \beta\sqrt{H^2K\cdot\Gamma(K,\lambda;\ker_m)}\right) + \beta HK\iota,$$

where $\iota = 5K^{7/12}H^{1/6}m^{-1/12}\log^{1/4} m$.

Proof. By the reduction of Algorithm 3 to Algorithm 1, we can apply the same proof for Lemma C.3, which completes the proof. □

Lemma E.4. *We define the event $\widetilde{\mathcal{E}}$ as that the following inequality holds $\forall (s,a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}, \forall h \in [H]$,*

$$|\mathbb{P}_h \overline{V}_{h+1}(s,a,b) - \overline{f}_h(s,a,b)| \le \overline{u}_h(s,a) + \beta\iota,$$
$$|\mathbb{P}_h \underline{V}_{h+1}(s,a,b) - \underline{f}_h(s,a,b)| \le \underline{u}_h(s,a) + \beta\iota,$$
$$\left|\|\varphi(z;\overline{W}_h)\|_{(\overline{\Lambda}_h)^{-1}} - \|\varphi(z;W^{(0)})\|_{(\widetilde{\Lambda}_h)^{-1}}\right| \le \iota,$$
$$\left|\|\varphi(z;\underline{W}_h)\|_{(\underline{\Lambda}_h)^{-1}} - \|\varphi(z;W^{(0)})\|_{(\widetilde{\Lambda}_h)^{-1}}\right| \le \iota.$$

where $\iota = 5K^{7/12}H^{1/6}m^{-1/12}\log^{1/4} m$*, and we define* $\overline{f}_h(z) = \Pi_{[0,H]}[f(z;\overline{W}_h)]$ *and* $\underline{f}_h(z) = \Pi_{[0,H]}[f(z;\underline{W}_h)]$ *as well as*

$$\overline{\Lambda}_h = \sum_{\tau=1}^{K} \varphi(z_h^\tau;\overline{W}_h)\varphi(z_h^\tau;\overline{W}_h)^\top + \lambda\cdot I, \quad \underline{\Lambda}_h = \sum_{\tau=1}^{K} \varphi(z_h^\tau;\underline{W}_h)\varphi(z_h^\tau;\underline{W}_h)^\top + \lambda\cdot I,$$
$$\widetilde{\Lambda}_h = \sum_{\tau=1}^{K} \varphi(z_h^\tau;W^{(0)})\varphi(z_h^\tau;W^{(0)})^\top + \lambda\cdot I.$$

Setting $\beta = \widetilde{B}_K$, $\widetilde{R}_K = H\sqrt{K}$, $\varsigma^* = H/K$, *and* $\lambda = \digamma^2(1+1/K)$, $\varsigma^* = H/K$, *if we set*

$$\beta^2 \geq 8R_Q^2H^2(1+\sqrt{\lambda/d})^2 + 32H^2\Gamma(K,\lambda;\ker_m) + 80H^2 + 32H^2\log\mathcal{N}_\infty(\varsigma^*;\widetilde{R}_K,\widetilde{B}_K) + 32H^2\log(2K/\delta'),$$

and also

$$m = \Omega(K^{19}H^{14}\log^3 m),$$

then we have that with probability at least $1-2/m^2-\delta'$*, the event* $\widetilde{\mathcal{E}}$ *happens, i.e.,*

$$\Pr(\widetilde{\mathcal{E}}) \geq 1-2/m^2-\delta'.$$

Proof. The proof of this lemma follows our proof of Lemmas C.2 and C.4 and apply some similar ideas from the proof of Lemma D.3. Particularly, to deal with the upper bounds of the estimation errors of $\mathbb{P}_h\overline{V}_{h+1}$ and $\mathbb{P}_h\underline{V}_{h+1}$, we define the two value function space $\overline{\mathcal{V}}$ and $\underline{\mathcal{V}}$ and show their covering numbers similar to the proof of Lemma D.3. Then, we further use the proof of Lemma C.4, which is derived from the proof of Lemma C.2, to show the eventual results in this lemma. In the proof of this lemma, we set $\widetilde{B}_K = \beta$ instead of $(1+1/H)\beta$ due to the structure of the planning phase. This completes the proof. □

Lemma E.5. *Conditioned on the event* $\widetilde{\mathcal{E}}$ *as defined in Lemma E.4, we have*

$$\begin{aligned} &V_h^\dagger(s,r) \leq \overline{V}_h(s) + (H+1-h)\beta\iota, \forall s\in\mathcal{S}, \forall h\in[H],\\ &\overline{V}_h(s) \leq \mathbb{E}_{a\sim\pi_h, b\sim \mathrm{br}(\pi)_h}[(\mathbb{P}_h\overline{V}_{h+1} + r_h + 2\overline{u}_h)(s,a,b)] + \beta\iota, \forall s\in\mathcal{S}, \forall h\in[H], \end{aligned} \tag{49}$$

$$\begin{aligned} &V_h^\dagger(s,r) \geq \underline{V}_h(s) - (H+1-h)\beta\iota, \forall s\in\mathcal{S}, \forall h\in[H],\\ &\underline{V}_h(s) \geq \mathbb{E}_{a\sim \mathrm{br}(\nu)_h, b\sim\nu_h}[(\mathbb{P}_h\underline{V}_{h+1} - r_h - 2\underline{u}_h)(s,a,b)] - \beta\iota, \forall s\in\mathcal{S}, \forall h\in[H]. \end{aligned} \tag{50}$$

Proof. We prove the first inequality in (49) by induction. For $h = H+1$, we have $V_{H+1}^\dagger(s,r) = \overline{V}_{H+1}(s) = 0$ for any $s\in\mathcal{S}$. Then, we assume that $V_{h+1}^\dagger(s,r) \leq \overline{V}_{h+1}(s) + (H-h)\beta\iota$. Thus, conditioned on the event $\widetilde{\mathcal{E}}$ as defined in Lemma E.4, we have

$$\begin{aligned} &Q_h^\dagger(s,a,b,r) - \overline{Q}_h(s,a,b)\\ &\quad= r_h(s,a,b) + \mathbb{P}_hV_{h+1}^\dagger(s,a,b,r) - \min\{[r_h(s,a,b) + \overline{f}_h(s,a,b) + u_h(s,a,b)], H\}^+\\ &\quad\leq \min\{[\mathbb{P}_hV_{h+1}^\dagger(s,a,b,r) - \overline{f}_h(s,a,b) - \overline{u}_h(s,a,b)], 0\}\\ &\quad\leq \min\{[\mathbb{P}_hV_{h+1}(s,a,b) + (H-h)\beta\iota - f_h(s,a,b) - \overline{u}_h(s,a,b)], 0\}\\ &\quad\leq (H+1-h)\beta\iota, \end{aligned}$$

where the first inequality is due to $0 \leq r_h(s,a,b) + \mathbb{P}_hV_{h+1}^\dagger(s,a,b,r) \leq H$ and $\min\{x,y\}^+ \geq \min\{x,y\}$, the second inequality is by the assumption that $V_{h+1}^\dagger(s,a,b,r) \leq \overline{V}_{h+1}(s,a,b) + (H-h)\beta\iota$, the last inequality is by Lemma E.4 such that $|\mathbb{P}_h\overline{V}_{h+1}(s,a,b) - \overline{f}_h(s,a,b)| \leq \overline{u}_h(s,a,b) + \beta\iota$ holds for any $(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}$ and $(k,h)\in[K]\times[H]$. The above inequality further leads to

$$\begin{aligned} V_h^\dagger(s,r) &= \max_{\pi_h'}\min_{\nu_h'}\mathbb{E}_{a\sim\pi_h', b\sim\nu_h'}[Q_h^\dagger(s,a,b,r)] \leq \max_{\pi_h'}\min_{\nu_h'}\mathbb{E}_{a\sim\pi_h', b\sim\nu_h'}[\overline{Q}_h(s,a,b)] + (H+1-h)\beta\iota\\ &= \overline{V}_h(s) + (H+1-h)\beta\iota. \end{aligned}$$

Therefore, we have

$$V_h^{\dagger}(s,r) \le \overline{V}_h(s) + (H+1-h)\beta\iota, \forall h \in [H], \forall s \in \mathcal{S}.$$

We further prove the second inequality in (49). We have

$$\begin{aligned}
\overline{Q}_h(s,a,b) &= \min\{[r_h(s,a,b) + \overline{f}_h(s,a,b) + \overline{u}_h(s,a,b)], H\}^+ \\
&\le \min\{[r_h(s,a,b) + \mathbb{P}_h\overline{V}_{h+1}(s,a,b) + 2\overline{u}_h(s,a,b) + \beta\iota], H\}^+ \\
&\le r_h(s,a,b) + \mathbb{P}_h\overline{V}_{h+1}(s,a,b) + 2\overline{u}_h(s,a,b) + \beta\iota,
\end{aligned}$$

where the first inequality is also by Lemma E.4 such that $|\mathbb{P}_h\overline{V}_{h+1}(s,a,b) - \overline{f}_h(s,a,b)| \le \overline{u}_h(s,a,b) + \beta\iota$, and the last inequality is because of the non-negativity of $r_h(s,a,b) + \mathbb{P}_h V_{h+1}(s,a,b) + 2\overline{u}_h(s,a,b) + \beta\iota$. Therefore, we have

$$\begin{aligned}
\overline{V}_h(s) &= \min_{\nu'} \mathbb{E}_{a\sim\pi_h, b\sim\nu'}\overline{Q}_h(s,a,b) \\
&\le \mathbb{E}_{a\sim\pi_h, b\sim \mathrm{br}(\pi)_h}\overline{Q}_h(s,a,b) \\
&\le \mathbb{E}_{a\sim\pi_h, b\sim \mathrm{br}(\pi)_h}[r_h(s,a,b) + \mathbb{P}_h\overline{V}_{h+1}(s,a,b) + 2\overline{u}_h(s,a,b)] + \beta\iota.
\end{aligned}$$

For the inequalities in (50), we can prove them in the same way to proving (49). In fact, Player 2 is trying to find a policy to maximize the cumulative rewards w.r.t. a reward function $\{-r_h(s,a,b)\}_{h\in[H]}$, while the opponent (Player 1) is trying to minimize the cumulative reward w.r.t. $\{-r_h(s,a,b)\}_{h\in[H]}$. Thus, one can also convert the results in (49) into (50) by this trick. This completes the proof. □

Lemma E.6. *With the exploration and planning phases, conditioned on the event $\mathcal{E}$ defined in Lemma E.2 and the event $\widetilde{\mathcal{E}}$ defined in Lemma E.4, we have the following inequalities*

$$K \cdot V_1^{\pi,\mathrm{br}(\pi)}(s_1, \overline{u}/H) \le \sum_{k=1}^K V_1^*(s_1, r^k) + 2K\beta\iota, \qquad K \cdot V_1^{\mathrm{br}(\nu),\nu}(s_1, \underline{u}/H) \le \sum_{k=1}^K V_1^*(s_1, r^k) + 2K\beta\iota.$$

Proof. First, we have $K \cdot V_1^{\pi,\mathrm{br}(\pi)}(s_1, \overline{u}/H) \le K \cdot V_1^*(s_1, \overline{u}/H)$ as well as $K \cdot V_1^{\mathrm{br}(\nu),\nu}(s_1, \underline{u}/H) \le K \cdot V_1^*(s_1, \underline{u}/H)$ according to the definition of V_1^*. Thus, to prove this lemma, we only need to show

$$K \cdot V_1^*(s_1, \overline{u}/H) \le \sum_{k=1}^K V_1^*(s_1, r^k) + 2K\beta\iota, \quad K \cdot V_1^*(s_1, \underline{u}/H) \le \sum_{k=1}^K V_1^*(s_1, r^k) + 2K\beta\iota.$$

Because the constructions of $\overline{u}_h$ and r_h^k are the same as the ones for the single-agent case, similar to the proof of Lemma C.6, and according to Lemmas E.2 and E.4, we have

$$\overline{u}_h(s,a,b) - \beta\iota \le H \cdot r_h^k(s,a,b) + \beta\iota, \quad \underline{u}_h(s,a,b) - \beta\iota \le H \cdot r_h^k(s,a,b) + \beta\iota$$

such that

$$V_1^*(s_1, \overline{u}/H) \le V_1^*(s_1, r^k) + 2\beta\iota, \quad V_1^*(s_1, \underline{u}/H) \le V_1^*(s_1, r^k) + 2\beta\iota,$$

Therefore, we eventually obtain

$$\begin{aligned}
K \cdot V_1^{\pi,\mathrm{br}(\pi)}(s_1, \overline{u}/H) \le K \cdot V_1^*(s_1, \overline{u}/H) \le \sum_{k=1}^K V_1^*(s_1, r^k) + 2\beta\iota, \\
K \cdot V_1^{\mathrm{br}(\nu),\nu}(s_1, \underline{u}/H) \le K \cdot V_1^*(s_1, \underline{u}/H) \le \sum_{k=1}^K V_1^*(s_1, r^k) + 2\beta\iota.
\end{aligned}$$

This completes the proof. □

E.2. Proof of Theorem 4.2

Proof. Conditioned on the event $\mathcal{E}$ defined in Lemma E.2 and the event $\widetilde{\mathcal{E}}$ defined in Lemma E.4, we have

$$V_1^\dagger(s_1, r) - V_1^{\pi, \mathrm{br}(\pi)}(s_1, r) \le \overline{V}_1(s_1) - V_1^{\pi, \mathrm{br}(\pi)}(s_1, r) + H\beta\iota, \tag{51}$$

where the inequality is by Lemma E.5. Further by this lemma, we have

$$\begin{aligned}
&\overline{V}_h(s_h) - V_h^{\pi, \mathrm{br}(\pi)}(s_h, r) \\
&\quad \le \mathbb{E}_{a_h \sim \pi_h, b_h \sim \mathrm{br}(\pi)_h}[(\mathbb{P}_h \overline{V}_{h+1} + r_h + 2\overline{u}_h)(s_h, a_h, b_h)] - V_h^{\pi, \mathrm{br}(\pi)}(s_h, r) + \beta\iota \\
&\quad = \mathbb{E}_{a_h \sim \pi_h, b_h \sim \mathrm{br}(\pi)_h}[(r_h + \mathbb{P}_h \overline{V}_{h+1} + 2\overline{u}_h)(s_h, a_h, b_h) - r_h(s_h, a_h, b_h) - \mathbb{P}_h V_{h+1}^{\pi, \mathrm{br}(\pi)}(s_h, a_h, b_h, r)] + \beta\iota \\
&\quad = \mathbb{E}_{a_h \sim \pi_h, b_h \sim \mathrm{br}(\pi)_h}[\mathbb{P}_h \overline{V}_{h+1}(s_h, a_h, b_h) - \mathbb{P}_h V_{h+1}^{\pi, \mathrm{br}(\pi)}(s_h, a_h, b_h, r) + 2\overline{u}_h(s_h, a_h, b_h)] + \beta\iota \\
&\quad = \mathbb{E}_{a_h \sim \pi_h, b_h \sim \mathrm{br}(\pi)_h, s_{h+1} \sim \mathbb{P}_h(\cdot | s_h, a_h, b_h)}[\overline{V}_{h+1}(s_{h+1}) - V_{h+1}^{\pi, \mathrm{br}(\pi)}(s_{h+1}, r) + 2\overline{u}_h(s_h, a_h, b_h)] + \beta\iota.
\end{aligned}$$

Recursively applying the above inequality and making use of $\overline{V}_{H+1}(s, r) = V_{H+1}^{\pi, \mathrm{br}(\pi)}(s) = 0$ gives

$$\begin{aligned}
\overline{V}_1(s_1) - V_1^{\pi, \mathrm{br}(\pi)}(s_1, r) &\le \mathbb{E}_{\forall h \in [H]:\, a_h \sim \pi_h, b_h \sim \mathrm{br}(\pi)_h, s_{h+1} \sim \mathbb{P}_h(\cdot | s_h, a_h, b_h)} \left[\sum_{h=1}^H 2\overline{u}_h(s_h, a_h, b_h) \middle| s_1 \right] \\
&= 2H \cdot V_1^{\pi, \mathrm{br}(\pi)}(s_1, \overline{u}/H) + H\beta\iota.
\end{aligned}$$

Combining with (51) gives

$$\begin{aligned}
V_1^\dagger(s_1, r) - V_1^{\pi, \mathrm{br}(\pi)}(s_1, r) &\le 2H \cdot V_1^{\pi, \mathrm{br}(\pi)}(s_1, \overline{u}/H) + 2H\beta\iota \le \frac{2H}{K} \sum_{k=1}^K V_1^*(s_1, r^k) + 4H\beta\iota \\
&\le \frac{2H}{K} \mathcal{O}\left(\sqrt{H^3 K \log(1/\delta')} + \beta \sqrt{H^2 K \cdot \Gamma(K, \lambda; \ker_m)} \right) + H^2\beta\iota + 4H\beta\iota \\
&\le \mathcal{O}\left([\sqrt{H^5 \log(1/\delta')} + \beta \sqrt{H^4 \cdot \Gamma(K, \lambda; \ker_m)}] / \sqrt{K} + H^2\beta\iota \right),
\end{aligned}$$

where the second inequality is due to Lemma E.6 and the last inequality is by Lemma E.3.

Next, we give the upper bound of $V_1^{\mathrm{br}(\nu), \nu}(s_1, r) - V_1^\dagger(s_1, r)$. Conditioned on the event $\mathcal{E}$ defined in Lemma E.2 and the event $\widetilde{\mathcal{E}}$ defined in Lemma E.4, we have

$$V_1^{\mathrm{br}(\nu), \nu}(s_1, r) - V_1^\dagger(s_1, r) \le V_1^{\mathrm{br}(\nu), \nu}(s_1, r) - \underline{V}_1(s_1) + H\beta\iota, \tag{52}$$

where the inequality is by Lemma E.5. Further by this lemma, we have

$$\begin{aligned}
&V_h^{\mathrm{br}(\nu), \nu}(s_h, r) - \underline{V}_h(s_h) \\
&\quad \le V_h^{\mathrm{br}(\nu), \nu}(s_h, r) - \mathbb{E}_{a \sim \mathrm{br}(\nu)_h, b \sim \nu_h}[(\mathbb{P}_h \underline{V}_{h+1} - r_h - 2\underline{u}_h)(s_h, a_h, b_h)] + \beta\iota \\
&\quad = \mathbb{E}_{a_h \sim \mathrm{br}(\nu)_h, b_h \sim \nu_h}[\mathbb{P}_h V_{h+1}^{\mathrm{br}(\nu), \nu}(s_h, a_h, b_h, r) - \mathbb{P}_h \underline{V}_{h+1}(s_h, a_h, b_h) + 2\underline{u}_h(s_h, a_h, b_h)] + \beta\iota \\
&\quad = \mathbb{E}_{a_h \sim \mathrm{br}(\nu)_h, b_h \sim \nu_h, s_{h+1} \sim \mathbb{P}_h(\cdot | s_h, a_h, b_h)}[V_{h+1}^{\mathrm{br}(\nu), \nu}(s_{h+1}, r) - \mathbb{P}_h \underline{V}_{h+1}(s_{h+1}) + 2\underline{u}_h(s_h, a_h, b_h)] + \beta\iota.
\end{aligned}$$

Recursively applying the above inequality gives

$$V_1^{\mathrm{br}(\nu), \nu}(s_1, r) - \underline{V}_1(s_1) \le 2H \cdot V_1^{\mathrm{br}(\nu), \nu}(s_1, \underline{u}/H) + H\beta\iota.$$

Combining with (52) gives

$$\begin{aligned}
V_1^{\mathrm{br}(\nu), \nu}(s_1, r) - V_1^\dagger(s_1, r) &\le 2H \cdot V_1^\pi(s_1, \underline{u}/H) + 2H\beta\iota \le \frac{2H}{K} \sum_{k=1}^K V_1^*(s_1, r^k) + 4H\beta\iota \\
&\le \mathcal{O}\left([\sqrt{H^5 \log(1/\delta')} + \beta \sqrt{H^4 \cdot \Gamma(K, \lambda; \ker_m)}] / \sqrt{K} + H^2\beta\iota \right),
\end{aligned}$$

where the second inequality is due to Lemma E.6 and the last inequality is by Lemma E.3. Thus, we eventually have

$$V_1^{\mathrm{br}(\nu),\nu}(s_1,r) - V_1^{\pi,\mathrm{br}(\pi)}(s_1,r) \le \mathcal{O}\left(\big[\sqrt{H^5\log(1/\delta')} + \beta\sqrt{H^4\cdot\Gamma(K,\lambda;\ker_m)}\big]/\sqrt{K} + H^2\beta\iota\right).$$

Moreover, we also have $P(\mathcal{E}\wedge\widetilde{\mathcal{E}}) \ge 1-2\delta'-4/m^2$ by union bound. Therefore, since $\beta \ge H$ as shown in Lemmas E.2 and E.4, setting $\delta' = 1/(4K^2H^2)$, we obtain that with probability at least $1-1/(2K^2H^2)-4/m^2$,

$$V_1^*(s_1,r) - V_1^{\pi}(s_1,r) \le \mathcal{O}\left(\beta\sqrt{H^4[\Gamma(K,\lambda;\ker_m)+\log(KH)]}/\sqrt{K} + H^2\beta\iota\right).$$

The event $\mathcal{E}\wedge\widetilde{\mathcal{E}}$ happens if we further let β satisfy

$$\beta^2 \ge 8R_Q^2H^2(1+\sqrt{\lambda/d})^2 + 32H^2\Gamma(K,\lambda;\ker_m) + 80H^2 + 32H^2\log\mathcal{N}_\infty(\varsigma^*;R_K,2\beta) + 96H^2\log(2KH).$$

where guarantees the conditions in Lemmas E.2 and E.4 hold. This completes the proof. □

F. Other Supporting Lemmas

Lemma F.1 (Lemma E.2 of Yang et al. (2020)). *Let $\{s_\tau\}_{\tau=1}^\infty$ and $\{\phi_\tau\}_{\tau=1}^\infty$ be $\mathcal{S}$-valued and $\mathcal{H}$-valued stochastic processes adapted to filtration $\{\mathcal{F}_\tau\}_{\tau=0}^\infty$, respectively, where we assume that $\|\phi_\tau\|\le 1$ for all $\tau\ge 1$. Moreover, for any $t\ge 1$, we let $\mathcal{K}_t\in\mathbb{R}^{t\times t}$ be the Gram matrix of $\{\phi_\tau\}_{\tau\in[t]}$ and define an operator $\Lambda_t:\mathcal{H}\mapsto\mathcal{H}$ as $\Lambda_t = \lambda I + \sum_{\tau=1}^t\phi_\tau\phi_\tau^\top$ with $\lambda>1$. Let $\mathcal{V}\subseteq\{V:\mathcal{S}\mapsto[0,H]\}$ be a class of bounded functions on $\mathcal{S}$. Then for any $\delta\in(0,1)$, with probability at least $1-\delta$, we have simultaneously for all $t\ge 1$ that*

$$\sup_{V\in\mathcal{V}}\left\|\sum_{\tau=1}^t\phi_\tau\{V(s_\tau)-\mathbb{E}[V(s_\tau)|\mathcal{F}_{\tau-1}]\}\right\|_{\Lambda_t^{-1}}^2 \le 2H^2\log\det(I+\mathcal{K}_t/\lambda) + 2H^2t(\lambda-1) + 4H^2\log(\mathcal{N}_\epsilon/\delta) + 8t^2\epsilon^2/\lambda,$$

where $\mathcal{N}_\epsilon$ is the ϵ-covering number of $\mathcal{V}$ with respect to the distance $\mathrm{dist}(\cdot,\cdot) := \sup_{\mathcal{S}}|V_1(s)-V_2(s)|$.

Lemma F.2 (Lemma E.3 of Yang et al. (2020)). *Let $\{\phi_t\}_{t\ge1}$ be a sequence in the RKHS $\mathcal{H}$. Let $\Lambda_0:\mathcal{H}\mapsto\mathcal{H}$ be defined as λI where $\lambda\ge1$ and I is the identity mapping on $\mathcal{H}$. For any $t\ge1$, we define a self-adjoint and positive-definite operator Λ_t by letting $\Lambda_t = \Lambda_0 + \sum_{j=1}^t\phi_j\phi_j^\top$. Then, for any $t\ge1$, we have*

$$\sum_{j=1}^t\min\{1,\phi_j\Lambda_{j-1}^{-1}\phi_j^\top\} \le 2\log\det(I+\mathcal{K}_t/\lambda),$$

where $\mathcal{K}_t\in\mathbb{R}^{t\times t}$ is the Gram matrix obtained from $\{\phi_j\}_{j\in[t]}$, i.e., for any $j,j'\in[t]$, the (j,j')-th entry of $\mathcal{K}_t$ is $\langle\phi_j,\phi_j\rangle_{\mathcal{H}}$. Moreover, if we further have $\sup_{t\ge0}\{\|\phi_t\|_{\mathcal{H}}\}\le1$, then it holds that

$$\log\det(I+\mathcal{K}_t/\lambda) \le \sum_{j=1}^t\phi_j^\top\Lambda_{j-1}^{-1}\phi_j \le 2\log\det(I+\mathcal{K}_t/\lambda).$$