A. Proofs of propositions and lemmas

In this section, we give the proofs of propositions and lemmas in §4. To simplify the notation, we write $\mathrm{KL}(\pi^1 \,\|\, \pi^2) = \mathrm{KL}(\pi^1(\cdot \mid s) \,\|\, \pi^2(\cdot \mid s))$, where KL computes the Kullback–Leibler divergence.

A.1. Proof of Lemma 3.1

The maximization problem that we are supposed to solve is formulated as

$$\max_{\pi^1} \mathbb{E}_{\nu_k}\Big[\big\langle \widehat{Q}^{\pi_k^1}(s,\cdot), \pi^1(\cdot \mid s)\big\rangle - \beta \cdot \mathrm{KL}\big(\pi^1(\cdot \mid s) \,\|\, \pi_k^1(\cdot \mid s)\big)\Big] \tag{A.1}$$

$$\text{subject to} \quad \sum_{a \in \mathcal{A}^1} \pi^1(a^1 \mid s) = 1, \quad \forall s \in \mathcal{S}. \tag{A.2}$$

The Lagrangian of the above problem is given by

$$\int_{s \in \mathcal{S}} \Big(\big\langle \widehat{Q}^{\pi_k^1}(s,\cdot), \pi^1(\cdot \mid s)\big\rangle - \beta \cdot \mathrm{KL}\big(\pi^1(\cdot \mid s) \,\|\, \pi_k^1(\cdot \mid s)\big)\Big)\mathrm{d}\nu_k + \int_{s \in \mathcal{S}} \Big(\sum_{a^1 \in \mathcal{A}^1} \pi^1(a^1 \mid s) - 1\Big)\mathrm{d}\lambda(s), \tag{A.3}$$

where $\lambda(\cdot)$ is the introduced dual parameter. Plugging in (3.1), we obtain the following optimality condition,

$$\pi^1(a^1 \mid s) = \frac{\exp\big\{\beta^{-1} \cdot Q_k^1(s,a^1) + \tau_k^{-1} \cdot \phi^1(s,a^1)^\top \theta_k^1\big\}}{\exp\big\{1 + \sum_{a' \in A^1} \exp(\tau_k^{-1} \cdot \phi^1(s,a')^\top \theta_k^1) + \beta^{-1} \cdot \frac{\mathrm{d}\lambda(s)}{\mathrm{d}\nu_k(s)}\big\}} \tag{A.4}$$

Hence, we have $\pi^1(a^1 \mid s) \propto \exp\{\beta^{-1} \cdot Q_k^1(s,a^1) + \tau_k^{-1} \cdot \phi^1(s,a^1)^\top \theta_k^1\}$. By plugging in the definition of Q_k^1 in (3.3), it follows that

$$\theta_{k+1}^1 = \tau_{k+1} \cdot (\beta^{-1}\omega_k^1 + \tau_k^{-1}\theta_k^1), \tag{A.5}$$

which concludes the proof of Lemma 3.1.

A.2. Proof of Proposition 4.4

Proof. Let

$$\widehat{\Phi} = \Big(\frac{1}{N}\sum_{i=1}^N \phi^1(s_{i,0}, a_{i,0}^1)\phi^1(s_{i,0}, a_{i,0}^1)^\top\Big)^{-1}, \tag{A.6}$$

$$\Phi = \Big(\mathbb{E}_{\rho_{k+1}}\big[\phi^1(s,a^1)\phi^1(s,a^1)^\top\big]\Big)^{-1}, \tag{A.7}$$

$$\widehat{v} = \frac{1}{N}\sum_{i=1}^N \big((1-\gamma)\cdot r_{i,1}^1 + \gamma \cdot \widehat{Q}^{\pi_k^1}(s_{i,2}, a_{i,2}^1)\big) \cdot \phi^1(s_{i,1}, a_{i,1}^1), \tag{A.8}$$

$$v = \mathbb{E}_{\rho_{k+1}}\big[\mathbb{T}^{\pi_{k+1}^1, \pi_{k+1}^2}\widehat{Q}^{\pi_k^1}(s,a^1,a^2) \cdot \phi^1(s,a^1)\big]. \tag{A.9}$$

Then, we can write

$$\begin{aligned}\mathbb{T}^{\pi_{k+1}^1, \widetilde{\pi}_{k+1}^2}\widehat{Q}^{\pi_k^1} &= \phi^1(s,a^1)^\top \widetilde{\omega}_{k+1} = \phi^1(s,a^1)^\top(\Phi v),\\ \widehat{Q}^{\pi_{k+1}^1} &= \phi^1(s,a^1)^\top \omega_{k+1} = \phi^1(s,a^1)^\top \Pi_R(\widehat{\Phi}\widehat{v}).\end{aligned}$$

Here, recall that $\widetilde{\omega}_{k+1}$ is the solution to the critic update problem in (3.8). The closed form of $\widetilde{\omega}_{k+1}$ is given in (3.7). By Assumption 4.2, it holds that $\Phi v = \Pi_R(\Phi v)$. Then, we have

$$\begin{aligned}&|\mathbb{T}^{\pi_{k+1}^1, \widetilde{\pi}_{k+1}^2}\widehat{Q}^{\pi_k^1} - \widehat{Q}^{\pi_{k+1}^1}|\\ &\quad= |\phi^1(s,a^1)^\top \Pi_R(\Phi v - \widehat{\Phi}\widehat{v})|\\ &\quad\le \|\phi^1(s,a^1)\|_2 \cdot \|\Pi_R(\Phi v - \widehat{\Phi}\widehat{v})\|_2\\ &\quad\le \|\Phi v - \widehat{\Phi}\widehat{v}\|_2\\ &\quad\le \|\Phi - \widehat{\Phi}\|_2 \cdot \|v\|_2 + \|\widehat{\Phi}\|_2 \cdot \|\widehat{v} - v\|_2,\end{aligned}$$

where the second inequality follows from the fact that $\|\phi^1\|_2 \leq 1$ uniformly, and Π_R is a contraction, and the last inequality follows from the triangle inequality. Since $|r^1| \leq r_{\max}$, $\|\omega\|_2 \leq R$, and $\|\phi^1\|_2 \leq 1$, it follows that $\|v\|_2 \leq r_{\max} + R$. The following lemmas provide upper bound for $\|\widehat{v} - v\|$ and high-probability upper bounds for $\|\Phi - \widehat{\Phi}\|$and $\|\widehat{\Phi}\|_2$.

Lemma A.1. For $\widehat{v}$ and v defined in (A.8) and (A.9), it holds with probability at least $1 - \delta$ that

$$\|\widehat{v} - v\|_2 \leq \frac{4(r_{\max} + R)}{\sqrt{N}} \log\Big(\frac{2d}{\delta}\Big),$$

where the uncertainty comes from $\widehat{v}$.

Proof. See §B.2 for a detailed proof. □

Lemma A.2. For Φ and $\widehat{\Phi}$ defined in (A.7) and (A.6), it holds with probability at least $1 - \delta$ that

$$\|\widehat{\Phi} - \Phi\|_2 \leq \frac{1}{(\sqrt{N}\lambda_*/4) \cdot \log^{-1}(2d/\delta) - 1}, \text{ and } \|\widehat{\Phi}\|_2 \leq \frac{\sqrt{N}}{\sqrt{N}\lambda_* - 4\log(2d/\delta)},$$

where the uncertainty comes from $\widehat{\Phi}$.

Proof. See §B.3 for a detailed proof. □

Then, it holds with probability at least $1 - 2\delta$ that

$$\begin{aligned}
&\mathbb{E}_\rho\big[|\mathbb{T}^{\pi^1_{k+1},\widetilde{\pi}^2_{k+1}}\widehat{Q}^{\pi^1_k} - \widehat{Q}^{\pi^1_{k+1}}|\big] \\
&\quad\leq \mathbb{E}_\rho\big[\|\Phi - \widehat{\Phi}\|_2 \cdot \|v\|_2 + \|\widehat{\Phi}\|_2 \cdot \|\widehat{v} - v\|_2\big] \\
&\quad\leq \frac{r_{\max} + R}{(\sqrt{N}\lambda_*/4) \cdot \log^{-1}(2d/\delta) - 1} + \frac{4(r_{\max} + R)\log(2d/\delta)}{\sqrt{N}\lambda_* - 4\log(2d/\delta)} \\
&\quad\leq \frac{2(r_{\max} + R)}{(\sqrt{N}\lambda_*/4) \cdot \log^{-1}(2d/\delta) - 1} \\
&\quad\leq \frac{16(r_{\max} + R)}{\sqrt{N}\lambda_*} \cdot \log\Big(\frac{2d}{\delta}\Big).
\end{aligned}$$

Here, the expectation $\mathbb{E}$ is taken with respect to $(s, a^1, a^2) \sim \rho$, the uncertainty comes from $\widehat{\Phi}$ and $\widehat{v}$, and the last inequality holds whenever $\sqrt{N}\lambda_*/(4\log(2d/\delta)) \geq 2$, which is satisfied by our choose of N. Thus, we conclude the proof of Proposition 4.4. □

A.3. Proof of Lemma 5.1

Proof. We first give the following lemma which holds for π^1_{k+1}, π^1_k and $\widehat{Q}^{\pi^1_k}$ that comes from Algorithm 3.

Lemma A.3. Let π^1_k, π^1_{k+1}, and $\widehat{Q}^{\pi^1_k}$ be given by Algorithm 3, and π^1 be any policy of player 1. Then we have

$$\big\langle \log(\pi^1_{k+1}/\pi^1_k), \pi^1 - \pi^1_k \big\rangle = \beta^{-1} \cdot \big\langle \widehat{Q}^{\pi^1_k}, \pi^1 - \pi^1_k \big\rangle.$$

Proof. See §B.1 for a detailed proof. □

Then, it holds that

$$
\begin{aligned}
f_k &= \big[\gamma(\mathbb{P}_1^{\pi_*^1} - \mathbb{P}_1^{\pi_{k+1}^1})\widehat{Q}^{\pi_k^1}\big] \\
&= \big[\gamma\mathbb{P}\langle \widehat{Q}^{\pi_k^1}, \pi_*^1 - \pi_{k+1}^1\rangle\big] \\
&= \gamma\beta\big[\mathbb{P}\langle \log(\pi_{k+1}^1/\pi_k^1), \pi_*^1 - \pi_{k+1}^1\rangle\big] \\
&= \gamma\beta\Big[\mathbb{P}\big\langle \log(\pi_{k+1}^1/\pi_*^1) + \log(\pi_*^1/\pi_k^1), \pi_*^1\big\rangle - \mathbb{P}\big\langle \log(\pi_{k+1}^1/\pi_*^1), \pi_k^1\big\rangle\Big] \\
&= \gamma\beta\Big[\mathbb{P}\big(\mathrm{KL}(\pi_*^1 \,\|\, \pi_k^1) - \mathrm{KL}(\pi_*^1 \,\|\, \pi_{k+1}^1) - \mathrm{KL}(\pi_{k+1}^1 \,\|\, \pi_k^1)\big)\Big] \\
&\le \gamma\beta\Big[\mathbb{P}\big(\mathrm{KL}(\pi_*^1 \,\|\, \pi_k^1) - \mathrm{KL}(\pi_*^1 \,\|\, \pi_{k+1}^1)\big)\Big],
\end{aligned}
$$

where the third equality follows from Lemma A.3. Upon telescoping with respect to k, it holds that

$$
\begin{aligned}
\sum_{k=0}^{K} f_k &\le \sum_{k=0}^{K} \beta \cdot \Big[\mathbb{P}\big(\mathrm{KL}(\pi_*^1 \,\|\, \pi_k^1) - \mathrm{KL}(\pi_*^1 \,\|\, \pi_{k+1}^1)\big)\Big] \\
&= \beta \cdot \Big[\mathbb{P}\big(\mathrm{KL}(\pi_*^1 \,\|\, \pi_0^1) - \mathrm{KL}(\pi_*^1 \,\|\, \pi_{K+1}^1)\big)\Big] \\
&\le \beta \cdot \big[\mathbb{P}\mathrm{KL}(\pi_*^1 \,\|\, \pi_0^1)\big].
\end{aligned} \tag{A.10}
$$

By the initialization of policies in (3.2), π_0^1 is a uniform policy. Thus, it holds for any state $s \in \mathcal{S}$ that

$$
\mathrm{KL}\big(\pi_*^1 \,\|\, \pi_0^1\big) = \sum_{a^1 \in \mathcal{A}^1} \pi_*^1(a^1 \mid s) \log \frac{\pi_*^1(a^1 \mid s)}{\pi_0^1(a^1 \mid s)} = \sum_{a^1 \in \mathcal{A}^1} \pi_*^1(a^1 \mid s) \log \frac{\pi_*^1(a^1 \mid s)}{1/|\mathcal{A}^1|} = \log(|\mathcal{A}^1|) - \mathcal{H}(\pi_*^1). \tag{A.11}
$$

Here, $\mathcal{H}(\pi_*^1)$ is the Shannon entropy of π_*^1, and thus is always positive. Then, we obtain that

$$
\sum_{k=0}^{K} f_k \le \beta \log(|\mathcal{A}^1|),
$$

which concludes the proof of Lemma 5.1. □

A.4. Proof of Lemma 5.2

Proof. It holds that

$$
\begin{aligned}
h_k &= \mathbb{T}^{\pi_{k+1}^1, \widetilde{\pi}_{k+1}^2} \widehat{Q}^{\pi_k^1} - Q^{\pi_{k+1}^1, \widetilde{\pi}_{k+1}^2} \\
&= (1-\gamma) \cdot r + \gamma \mathbb{P}^{\pi_{k+1}^1, \widetilde{\pi}_{k+1}^2} \widehat{Q}^{\pi_k^1} - Q^{\pi_{k+1}^1, \widetilde{\pi}_{k+1}^2} \\
&= (1-\gamma) \cdot r + \sum_{t=1}^{\infty} \big[(\gamma \mathbb{P}^{\pi_{k+1}^1, \widetilde{\pi}_{k+1}^2})^t - (\gamma \mathbb{P}^{\pi_{k+1}^1, \widetilde{\pi}_{k+1}^2})^{t+1}\big] \widehat{Q}^{\pi_k^1} - \sum_{t=0}^{\infty} (1-\gamma)(\gamma \mathbb{P}^{\pi_{k+1}^1, \widetilde{\pi}_{k+1}^2})^t r \\
&= \sum_{t=1}^{\infty} (\gamma \mathbb{P}^{\pi_{k+1}^1, \widetilde{\pi}_{k+1}^2})^t \big(\widehat{Q}^{\pi_k^1} - \gamma \mathbb{P}^{\pi_{k+1}^1, \widetilde{\pi}_{k+1}^2} \widehat{Q}^{\pi_k^1} - (1-\gamma) r\big) \\
&= \sum_{t=1}^{\infty} (\gamma \mathbb{P}^{\pi_{k+1}^1, \widetilde{\pi}_{k+1}^2})^t \big(\underbrace{\widehat{Q}^{\pi_k^1} - \mathbb{T}^{\pi_{k+1}^1, \widetilde{\pi}_{k+1}^2} \widehat{Q}^{\pi_k^1}}_{\triangleq e_k}\big).
\end{aligned} \tag{A.12}
$$

Here, the second equality follows from the definition of Q^{π^1, π^2} in (2.1), and we define

$$
\begin{aligned}
e_k &= \widehat{Q}^{\pi_k^1} - \mathbb{T}^{\pi_{k+1}^1, \widetilde{\pi}_{k+1}^2} \widehat{Q}^{\pi_k^1} \\
&= \widehat{Q}^{\pi_k^1} - \mathbb{T}^{\pi_k^1, \widetilde{\pi}_k^2} \widehat{Q}^{\pi_k^1} + \gamma(\mathbb{P}^{\pi_k^1, \widetilde{\pi}_k^2} - \mathbb{P}^{\pi_{k+1}^1, \widetilde{\pi}_{k+1}^2}) \widehat{Q}^{\pi_k^1} \\
&= \widehat{Q}^{\pi_k^1} - \mathbb{T}^{\pi_k^1, \widetilde{\pi}_k^2} \widehat{Q}^{\pi_k^1} + \gamma(\mathbb{P}^{\pi_k^1} - \mathbb{P}^{\pi_k^2}) \widehat{Q}^{\pi_k^1}
\end{aligned} \tag{A.13}
$$

Since

$$
\begin{aligned}
(\mathbb{P}^{\pi_k^1} - \mathbb{P}^{\pi_k^2})\widehat{Q}^{\pi_k^1} &= \mathbb{P}\big[\langle \pi_k^1 - \pi_{k+1}^1, \widehat{Q}^{\pi_k^1}\rangle\big] \\
&= \beta \cdot \mathbb{P}\Big[\big\langle \log(\pi_{k+1}^1/\pi_k^1), \pi_k^1 - \pi_{k+1}^1\big\rangle\Big] \\
&= \beta \cdot \mathbb{P}\big[-\mathrm{KL}(\pi_k^1 \,\|\, \pi_{k+1}^1) - \mathrm{KL}(\pi_{k+1}^1 \,\|\, \pi_k^1)\big] \\
&\leq 0,
\end{aligned}
$$

where the second equality follows from Lemma A.3. We then have

$$
e_k = \widehat{Q}^{\pi_k^1} - \mathbb{T}^{\pi_{k+1}^1, \widetilde{\pi}_{k+1}^2}\widehat{Q}^{\pi_k^1} \leq \widehat{Q}^{\pi_k^1} - \mathbb{T}^{\pi_k^1, \widetilde{\pi}_k^2}\widehat{Q}^{\pi_k^1}.
$$

Then, it holds that

$$
\begin{aligned}
e_{k+1} &\leq \widehat{Q}^{\pi_{k+1}^1} - \mathbb{T}^{\pi_{k+1}^1, \widetilde{\pi}_{k+1}^2}\widehat{Q}^{\pi_{k+1}^1} \\
&= \underbrace{\widehat{Q}^{\pi_{k+1}^1} - \mathbb{T}^{\pi_{k+1}^1, \widetilde{\pi}_{k+1}^2}\widehat{Q}^{\pi_k^1}}_{=-\epsilon_k} + \underbrace{\mathbb{T}^{\pi_{k+1}^1, \widetilde{\pi}_{k+1}^2}\widehat{Q}^{\pi_k^1} - (\mathbb{T}^{\pi_{k+1}^1, \widetilde{\pi}_{k+1}^2})^2\widehat{Q}^{\pi_k^1}}_{=\gamma\mathbb{P}^{\pi_{k+1}^1, \widetilde{\pi}_{k+1}^2} e_k} \\
&\quad + \underbrace{(\mathbb{T}^{\pi_{k+1}^1, \widetilde{\pi}_{k+1}^2})^2\widehat{Q}^{\pi_k^1} - \mathbb{T}^{\pi_{k+1}^1, \widetilde{\pi}_{k+1}^2}\widehat{Q}^{\pi_{k+1}^1}}_{=\gamma\mathbb{P}^{\pi_{k+1}^1, \widetilde{\pi}_{k+1}^2}\epsilon_k} \\
&= \gamma\mathbb{P}^{\pi_{k+1}^1, \widetilde{\pi}_{k+1}^2} e_k - (I - \gamma\mathbb{P}^{\pi_{k+1}^1, \widetilde{\pi}_{k+1}^2})\epsilon_k.
\end{aligned} \tag{A.14}
$$

By applying (A.14) k times, we obtain that

$$
e_k \leq \gamma^k\Big(\prod_{i=1}^{k}\mathbb{P}^{\pi_i^1, \pi_i^2}\Big)e_0 - \sum_{i=0}^{k-1}\gamma^i\Big(\prod_{j=k-i+1}^{k}\mathbb{P}^{\pi_j^1, \pi_j^2}\Big)(I - \gamma\mathbb{P}^{\pi_{k-i}^1, \widetilde{\pi}_{k-i}^2})\epsilon_{k-i-1}. \tag{A.15}
$$

Plugging (A.15) into (A.12), we obtain that

$$
h_k \leq \sum_{t=1}^{\infty}(\gamma\mathbb{P}^{\pi_{k+1}^1, \widetilde{\pi}_{k+1}^2})^t\Big(\gamma^k\Big(\prod_{i=1}^{k}\mathbb{P}^{\pi_i^1, \pi_i^2}\Big)e_0 - \sum_{i=0}^{k-1}\gamma^i\Big(\prod_{j=k-i+1}^{k}\mathbb{P}^{\pi_j^1, \pi_j^2}\Big)(I - \gamma\mathbb{P}^{\pi_{k-i}^1, \widetilde{\pi}_{k-i}^2})\epsilon_{k-i-1}\Big). \tag{A.16}
$$

Since $|r^1| \leq r_{\max}$, $\|\omega\|_2 \leq R$, and $\|\phi\|_2 \leq 1$, it holds that

$$
|e_0| = |\widehat{Q}^{\pi_0^1} - \mathbb{T}^{\pi_1^1, \pi_1^2}\widehat{Q}^{\pi_0^1}| = |\widehat{Q}^{\pi_0^1} - (1-\gamma)r - \gamma\mathbb{P}^{\pi_1^1, \pi_1^2}\widehat{Q}^{\pi_0^1}| \leq r_{\max} + R.
$$

Upon telescoping with respect to k and taking expectation with respect to ρ, it holds that

$$
\mathbb{E}_\rho\Big[\sum_{k=0}^{K}\sum_{t=1}^{\infty}(\gamma\mathbb{P}^{\pi_{k+1}^1, \widetilde{\pi}_{k+1}^2})^t\gamma^k\Big(\prod_{i=1}^{k}\mathbb{P}^{\pi_i^1, \pi_i^2}\Big)e_0\Big] \leq \sum_{k=0}^{K}\sum_{t=1}^{\infty}\gamma^{k+t}(r_{\max} + R) \leq \frac{r_{\max} + R}{(1-\gamma)^2}, \tag{A.17}
$$

and

$$
\begin{aligned}
&\mathbb{E}_\rho\Big[\sum_{k=0}^{K}\sum_{t=1}^{\infty}\sum_{i=0}^{k-1}(\gamma\mathbb{P}^{\pi^1_{k+1},\widetilde{\pi}^2_{k+1}})^t\gamma^i\Big(\prod_{j=k+1-i}^{k}\mathbb{P}^{\pi^1_j,\pi^2_j}\Big)(I-\gamma\mathbb{P}^{\pi^1_{k-i},\widetilde{\pi}^2_{k-i}})\epsilon_{k-i-1}\Big]\\
&\le\sum_{k=1}^{K}\sum_{t=1}^{\infty}\sum_{i=0}^{k-1}\gamma^{t+i}\mathbb{E}_\rho\Big[(\mathbb{P}^{\pi^1_{k+1},\widetilde{\pi}^2_{k+1}})^t\Big(\prod_{j=k+1-i}^{k}\mathbb{P}^{\pi^1_j,\pi^2_j}\Big)|\epsilon_{k-i-1}|\Big]\\
&\le\sum_{k=1}^{K}\sum_{t=1}^{\infty}\sum_{i=0}^{k-1}\gamma^{t+i}c(t+i)\mathbb{E}_\rho\big[|\epsilon_{k-i-1}|\big]\\
&=\sum_{k=1}^{K}\sum_{t=1}^{\infty}\sum_{i=t}^{k+t-1}\gamma^{i}c(i)\mathbb{E}_\rho\big[|\epsilon_{k-i+t-1}|\big]\\
&=\sum_{k=1}^{K}\sum_{i=1}^{\infty}\sum_{t=\max\{1,i-k+1\}}^{i}\gamma^{i}c(i)\epsilon_Q\\
&\le\sum_{k=1}^{K}\sum_{i=1}^{\infty}i\gamma^{i}c(i)\epsilon_Q\\
&\le\frac{KC_\rho\epsilon_Q}{(1-\gamma)^2}.
\end{aligned}
\tag{A.18}
$$

Here, the second and the last inequality follows from Assumption 4.1, where $c(\cdot)$ and C_ρ are defined, and the second equality follows from the definition of ϵ_Q in (4.2). Combining (A.16), (A.17), and (A.18) together, we obtain that

$$
\mathbb{E}_\rho[\sum_{k=0}^{K}h_k]\le\frac{r_{\max}+R+KC_\rho\epsilon_Q}{(1-\gamma)^2},
$$

which concludes the proof of Lemma 5.2. □

B. Proof of Auxiliary Lemmas

In this section, we give proofs of the lemmas in §A.

B.1. Proof of Lemma A.3

Proof. Let π^1_k,π^1_{k+1}, and $\widehat{Q}^{\pi^1_k}$ be generated by Algorithm 3. Then it follows that

$$
\begin{aligned}
\big\langle\log(\pi^1_{k+1}/\pi^1_k),\pi^1-\pi^1_k\big\rangle&=\sum_{a\in\mathcal{A}^1}\Big[\log\big(\pi^1_{k+1}(s,a)/\pi^1_k(s,a)\big)\Big]\big(\pi^1(a\,|\,s)-\pi^1_k(a\,|\,s)\big)\\
&=\sum_{a\in\mathcal{A}^1}\Big[\log\Big(\frac{\exp\big(\tau_{k+1}^{-1}\cdot\phi^1(s,a)^\top\theta^1_{k+1}\big)}{\exp\big(\tau_k^{-1}\cdot\phi^1(s,a)^\top\theta^1_k\big)}\Big)\Big]\big(\pi^1(a\,|\,s)-\pi^1_k(a\,|\,s)\big)\\
&\quad-\underbrace{\sum_{a\in\mathcal{A}^1}\Big[\log\Big(\frac{\sum_{a'\in\mathcal{A}^1}\exp\big(\tau_{k+1}^{-1}\cdot\phi^1(s,a')^\top\theta^1_{k+1}\big)}{\sum_{a'\in\mathcal{A}^1}\exp\big(\tau_k^{-1}\cdot\phi^1(s,a')^\top\theta^1_k\big)}\Big)\Big]\big(\pi^1(a\,|\,s)-\pi^1_k(a\,|\,s)\big)}_{=\left(\sum_{a\in\mathcal{A}^1}\left(\pi^1(a\,|\,s)-\pi^1_k(a\,|\,s)\right)\right)\log\left(\frac{\sum_{a'\in\mathcal{A}^1}\exp\left(\tau_{k+1}^{-1}\cdot\phi^1(s,a')^\top\theta^1_{k+1}\right)}{\sum_{a'\in\mathcal{A}^1}\exp\left(\tau_k^{-1}\cdot\phi^1(s,a')^\top\theta^1_k\right)}\right)=0}\\
&=\sum_{a\in\mathcal{A}^1}\Big[\log\Big(\frac{\exp\big(\phi^1(s,a)^\top(\beta^{-1}\omega^1_k+\tau_k^{-1}\theta^1_k)\big)}{\exp\big(\tau_k^{-1}\cdot\phi^1(s,a)^\top\theta^1_k\big)}\Big)\Big]\big(\pi^1(a\,|\,s)-\pi^1_k(a\,|\,s)\big)\\
&=\sum_{a\in\mathcal{A}^1}\big(\phi^1(s,a)^\top\omega^1_k\big)\big(\pi^1(a\,|\,s)-\pi^1_k(a\,|\,s)\big)\\
&=\beta^{-1}\cdot\big\langle\widehat{Q}^{\pi^1_k},\pi^1-\pi^1_k\big\rangle,
\end{aligned}
$$

which concludes the proof of Lemma A.3. □

B.2. Proof of Lemma A.1

Proof. We denote

$$\Lambda_i = \frac{1}{N}\Big(\big((1-\gamma)\cdot r^1_{i,1} + \gamma\cdot \widehat{Q}^{\pi^1_k}(s_{i,2}, a^1_{i,2})\big)\cdot \phi^1(s_{i,1}, a^1_{i,1}) - \mathbb{E}_{\rho_{k+1}}\big[\mathbb{T}^{\pi^1_{k+1},\pi^2_{k+1}}\widehat{Q}^{\pi^1_k}(s, a^1, a^2)\cdot\phi^1(s, a^1)\big]\Big).$$

By the fact that $(s_{i,1}, a^1_{i,1}, a^2_{i,1}) \sim \rho_{k+1}$, $s_{i,2} \sim \mathcal{P}(\cdot \mid s, a^1_{i,1}, a^2_{i,1})$, and $a^1_{i,2} \sim \pi^1_{k+1}(\cdot \mid s_{i,2})$, we have $\mathbb{E}[\Lambda_i] = 0$. Since $|r^1| \le r_{\max}$, $\|\omega\| \le R$, and $\|\phi^1\|_2 \le 1$, it holds that $\|\Lambda_i\|_2 \le 2(r_{\max} + R)/N$. Also, the following matrix variance statistic of the sum $\sum \Lambda_i$ is upper bounded,

$$V(\sum_{j=1}^N \Lambda_j) = \mathbb{E}\Big[\Big\|(\sum_{j=1}^N \Lambda_i)(\sum_{j=1}^N \Lambda_i)^\top\Big\|_2\Big] = \mathbb{E}\Big[\Big\|\sum_{j=1}^N \Lambda_i\Lambda_i^\top\Big\|_2\Big] \le \frac{4(r_{\max}+R)^2}{N}.$$

Here, the second equality follows from $\mathbb{E}[\Lambda_i] = 0$ and the fact that $\Lambda_i, i = 1, 2, \ldots, N$ are independent. Thus, following from the matrix Bernstein inequality, it holds with probability at least $1-\delta$ that

$$\|\widehat{v} - v\|_2 \le \frac{4(r_{\max}+R)}{\sqrt{N}}\log\Big(\frac{2d}{\delta}\Big)$$

which concludes the proof of Lemma A.1. □

B.3. Proof of Lemma A.2

Proof. We denote

$$\Delta_i = \frac{1}{N}\Big(\phi^1(s_{i,0}, a^1_{i,0})\phi^1(s_{i,0}, a^1_{i,0})^\top - \mathbb{E}_{\rho_{k+1}}\big[\phi^1(s, a^1)\phi^1(s, a^1)^\top\big]\Big).$$

By the fact that $(s_{i,0}, a^1_{i,0}, a^2_{i,0}) \sim \rho_{k+1}$ and $\|\phi^1\|_2 \le 1$, we have $\mathbb{E}[\Delta_i] = 0$ and $\|\Delta_i\|_2 \le 2/N$. Note that the following matrix variance statistic of the sum $\sum \Delta_i$ is upper bounded,

$$V\Big(\sum_{i=1}^N \Delta_i\Big) = \mathbb{E}\Big[\Big\|(\sum_{i=1}^N \Delta_i)(\sum_{i=1}^N \Delta_i)^\top\Big\|_2\Big] = \mathbb{E}\Big[\Big\|\sum_{i=1}^N \Delta_i\Delta_i^\top\Big\|_2\Big] \le \frac{4}{N},$$

Here, the expectation is taken with respect to $(s_{i,0}, a^1_{i,0}, a^2_{i,0}) \sim \rho_{k+1}$ and the second equality follows from $\mathbb{E}[\Delta_i] = 0$ and the fact that $\Delta_i, i = 1, 2, \ldots, N$ are independent. Following from the matrix Bernstein inequality, it holds with probability at least $1-\delta$ that

$$\|\widehat{\Phi}^{-1} - \Phi^{-1}\|_2 \le \frac{2}{3N}\log\Big(\frac{2d}{\delta}\Big) + \sqrt{\Big(\frac{2}{3N}\log\Big(\frac{2d}{\delta}\Big)\Big)^2 + \frac{8}{N}\log\Big(\frac{2d}{\delta}\Big)} \le \frac{4}{\sqrt{N}}\log\Big(\frac{2d}{\delta}\Big),$$

where $\widehat{\Phi}$ and Φ are defined in (A.6) and (A.7), respectively. Following from Assumption 4.3, it holds that $\|\Phi^{-1}\|_2 \ge \lambda_*$. Thus, we have $\|\Phi\|_2 \le 1/\lambda_*$. In what follows, we upper bound $\|\widehat{\Phi}\|_2$, which is the last part of upper bounding $\|\widehat{\Phi} - \Phi\|_2$. It holds that

$$\|\widehat{\Phi}\|_2 = \big\|(\Phi^{-1} + \widehat{\Phi}^{-1} - \Phi^{-1})^{-1}\big\|_2 \le \|\Phi\|_2 + \|\Phi\|_2\cdot\|\widehat{\Phi}^{-1} - \Phi^{-1}\|_2\cdot\|\widehat{\Phi}\|_2.$$

Here, the first inequality follows from the fact that $(A+B)^{-1} = A^{-1} - A^{-1}B(A+B)^{-1}$ for non-singular matrices A and $A+B$ (Henderson & Searle, 1981), and the triangle inequality. Thus, it follows that

$$\|\widehat{\Phi}\|_2 \le \frac{\|\Phi\|_2}{1 - \|\Phi\|_2\cdot\|\widehat{\Phi}^{-1} - \Phi^{-1}\|_2}.$$

Then, we obtain that

$$\begin{aligned}\|\widehat{\Phi}-\Phi\|_2 &= \|\widehat{\Phi}\Phi^{-1}\Phi-\widehat{\Phi}\widehat{\Phi}^{-1}\Phi\|_2\\ &= \|\widehat{\Phi}(\Phi^{-1}-\widehat{\Phi}^{-1})\Phi\|_2\\ &\leq \|\widehat{\Phi}\|_2\cdot\|\widehat{\Phi}^{-1}-\Phi^{-1}\|_2\cdot\|\Phi\|_2\\ &\leq \frac{\|\Phi\|_2^2}{\|\widehat{\Phi}^{-1}-\Phi^{-1}\|_2^{-1}-\|\Phi\|_2}.\end{aligned}$$

Hence, it holds with probability at least $1-\delta$ that

$$\|\widehat{\Phi}-\Phi\|_2 \leq \frac{1/\lambda_*}{(\sqrt{N}/4)\cdot\log^{-1}(2d/\delta)-1/\lambda_*} = \frac{1}{(\sqrt{N}\lambda_*/4)\cdot\log^{-1}(2d/\delta)-1},$$

and

$$\|\widehat{\Phi}\|_2 \leq \frac{\sqrt{N}}{\sqrt{N}\lambda_*-4\log(2d/\delta)},$$

which concludes the proof of Lemam A.2. □