A PROOF

A.1 PROOF OF THEOREM 4.2

Proof. We define a mapping $\Gamma : \Pi \to \Pi$ such that $\boldsymbol{\pi}_{n+1} = \Gamma(\boldsymbol{\pi}_n)$ satisfies that $P_T(x_T, \pi^i_{n+1}, \boldsymbol{\pi}^{-i}_n) \propto R^i(x_T; \pi^i_{n+1}, \boldsymbol{\pi}^{-i}_n)$, where Π is the joint policy space of $\boldsymbol{\pi}$. From Brouwer Fixed Points Theorem, there exists a profile $\boldsymbol{\pi}^\star = \Gamma(\boldsymbol{\pi}^\star)$. From the definition of FE, $\boldsymbol{\pi}^\star$ is an FE. □

A.2 PROOF OF THEOREM 4.3

Before the proof, we introduce a Lemma (Guo et al., 2019).

Lemma A.1. *Define the two distribution over vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ where $x_1 = x_2 = \ldots = x_m = x_{max} \geq \max_i x_i$: $\textbf{softmax}_c(\boldsymbol{x})_i = \frac{e^{cx_i}}{\sum_j e^{cx_j}}$ and $\textbf{argmax-e}(\boldsymbol{x})_i = \begin{cases} \frac{1}{m}, & i \leq m, \\ 0, & otherwise. \end{cases}$ The distance between $\textbf{softmax}_c$ and $\textbf{argmax-e}$ is bounded by*

$$\|\textbf{softmax}_1(\boldsymbol{x}) - \textbf{argmax-e}(\boldsymbol{x})\|_1 \leq 2n \exp(-c\delta), \tag{15}$$

where $\delta = x_{max} - \max_{x_j < x_{max}} x_j$ and $\delta := \infty$ when all x_j are equal.

Proof. Without loss of generality, assume that $\log R^i(x^1_T) = \log R^i(x^2_T) = \ldots = \log R^i(x^m_T) = \max_k \log R^i(x^k_T) := R_{\max}$. We denote $\mathbf{x}_T = (x^1_T, x^2_T, \ldots, x^{|\mathcal{X}|}_T)$. Given the opponent policy $\boldsymbol{\pi}^{-i}$, the expected cumulative reward of the best response for agent i equals to $\langle \textbf{argmax-e}(\log R^i(\mathbf{x}_T)), \log R^i(\mathbf{x}_T)\rangle$ and $\langle \textbf{softmax}_1(\log R^i(\mathbf{x}_T)), \log R^i(\mathbf{x}_T)\rangle$. The exploitability of agent i is

$$\begin{aligned} &|\langle \textbf{argmax-e}(\log R^i(\mathbf{x}_T)), \log R^i(\mathbf{x}_T)\rangle - \langle \textbf{softmax}_1(\log R^i(\mathbf{x}_T)), \beta \log R^i(\mathbf{x}_T)\rangle| \\ &\leq \text{Ret}_{\max} \|\textbf{softmax}_1(\log R^i(\mathbf{x}_T)) - \textbf{argmax-e}(\log R^i(\mathbf{x}_T))\|_1 = 2|\mathcal{X}|e^{-\delta} \end{aligned} \tag{16}$$

□

A.3 PROOF OF PROPOSITION 5.1

Proof. We first define the discounted state visitation distribution $d^{\boldsymbol{\pi}}(s) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t P(s_t = s)$. Denote $\theta^{(t)}_i$ is the parameter vector of policy $\pi^{i,(t)}$. And the element of $\theta^{(t)}$ is the approximation of action-value function, i.e. $\theta^{(t)}_{i,s,a,a^{-i}} = Q^i(s, a, a^{-i}; \boldsymbol{\pi}^{(t)})$. From (Cen et al., 2022), the update rule of θ is

$$\begin{aligned} \theta^{(t+1)}_{i,s,a,a^{-i}} =& \theta^{(t)}_{i,s,a,a^{-i}} + \eta \left[\mathcal{F}(\theta^{(t)}_i)^\dagger V^i(s; \boldsymbol{\pi}^{(t)})\right] \\ =& \theta^{(t)}_{i,s,a,a^{-i}} + \frac{\eta}{1-\gamma}\left(Q^i(s, a, a^{-i}; \boldsymbol{\pi}^{(t)}) - \log \pi(a|s, a^{-i})\rho(a^{-i}|s) - V^i(s; \boldsymbol{\pi}^{(t)})\right), \end{aligned}$$

where $\mathcal{F}(\theta^{(t)}_i)^\dagger$ is the pseudo-inverse of the Fischer information matrix

$$\mathcal{F}(\theta^{(t)}_i) = \mathbb{E}_{s\sim d^{\boldsymbol{\pi}^{(t)}}(\cdot), a^{-i}\sim\rho(\cdot|s), a^i\sim\pi^{i,(t)}(\cdot|a^{-i},s)\sim}\left[\nabla_{\theta^{(t)}_i} \log \pi^{i,(t)}(a^i|s, a^{-i}) \log \pi^{i,(t)}(a^i|s, a^{-i})^T\right].$$

Hence the update rule of policy is

$$\begin{aligned} &\pi^{i,(t+1)}(a|s, a^{-i}) \propto \exp(\theta^{(t+1)}_{i,s,a,a^{-i}}) \\ =& \exp\left(\theta^{(t)}_{i,s,a,a^{-i}} + \frac{\eta}{1-\gamma}\left(Q^i(s, a, a^{-i}; \boldsymbol{\pi}^{(t)}) - \log \pi(a|s, a^{-i})\rho(a^{-i}|s) - V^i(s; \boldsymbol{\pi}^{(t)})\right)\right) \\ &\propto \left(\pi^{i,(t)}(a|s, a^{-i})\right)^{1-\frac{\eta}{1-\gamma}} \exp\left(\frac{\eta}{1-\gamma} Q^i\left(s, a, a^{-i}; \boldsymbol{\pi}^{(t)}\right)\right). \end{aligned}$$

□

A.4 Proof of Proposition 4.6

Proof. From Pinsker's inequality, $D_{TV}(\rho(\cdot|s), \pi^{-i}(\cdot|s)) \leq \sqrt{\frac{1}{2}\mathrm{KL}(\rho(\cdot|s)||\pi^{-i}(\cdot|s))} \leq \sqrt{\frac{1}{2}\epsilon_\rho}$. Denote the value function derived using the opponent model as $\hat{V}^i(s;\boldsymbol{\pi})$. Define $P^{\boldsymbol{\pi}}_\rho(s'|s) := \mathbb{E}_{a^{-i}\sim\rho(\cdot|s)}[\sum_{a\in\mathcal{A}^i} P(s'|s,a,a^{-i})\pi(a|s,a^{-i})]$.

$$
\begin{aligned}
&|V^i(s;\boldsymbol{\pi}) - \hat{V}^i(s;\boldsymbol{\pi})| \\
\leq& 2(1+\log|\mathcal{A}_i|)D_{TV}(\rho(\cdot|s), \pi^{-i}(\cdot|s)) + \gamma|\mathbb{E}_{s'\sim P^{\boldsymbol{\pi}}_\rho(s'|s)}V^i(s';\boldsymbol{\pi}) - \mathbb{E}_{s'\sim P^{\boldsymbol{\pi}}_{\boldsymbol{\pi}^{-i}}(s'|s)}V^i(s';\boldsymbol{\pi})| \\
&+ \gamma|\mathbb{E}_{s'\sim P^{\boldsymbol{\pi}}_\rho(s'|s)}[V^i(s';\boldsymbol{\pi}) - \hat{V}^i(s';\boldsymbol{\pi})]| - \mathrm{KL}(\rho(\cdot|s)||\pi^{-i}(\cdot|s)) \\
\leq& 2(1+\log|\mathcal{A}_i|)D_{TV}(\rho(\cdot|s), \pi^{-i}(\cdot|s)) + 2\gamma(\max_{s'\in\mathcal{S}}V^i(s';\boldsymbol{\pi}))D_{TV}(P^{\boldsymbol{\pi}}_\rho(s'|s), P^{\boldsymbol{\pi}}_{\boldsymbol{\pi}^{-i}}(s'|s)) \\
&+ \gamma\max_{s'\sim\mathcal{S}}|V^i(s';\boldsymbol{\pi}) - \hat{V}^i(s';\boldsymbol{\pi})| - \mathrm{KL}(\rho(\cdot|s)||\pi^{-i}(\cdot|s)) \\
\leq& 2(1+\log|\mathcal{A}_i| + \frac{\gamma(1+\log|\mathcal{A}_i|)}{1-\gamma})\sqrt{\frac{1}{2}\epsilon_\rho} + \gamma\max_{s'\sim\mathcal{S}}|V^i(s';\boldsymbol{\pi}) - \hat{V}^i(s';\boldsymbol{\pi})| + \epsilon_\rho \\
=& \frac{2(1+\log|\mathcal{A}_i|)}{1-\gamma}\sqrt{\frac{1}{2}\epsilon_\rho} + \gamma\max_{s'\sim\mathcal{S}}|V^i(s';\boldsymbol{\pi}) - \hat{V}^i(s';\boldsymbol{\pi})| + \epsilon_\rho
\end{aligned}
$$

Then the estimated error of value function can be derived

$$
\max_{s\sim\mathcal{S}}|V^i(s;\boldsymbol{\pi}) - \hat{V}^i(s;\boldsymbol{\pi})| \leq \frac{2(1+\log|\mathcal{A}_i|)}{(1-\gamma)^2}\sqrt{\frac{1}{2}\epsilon_\rho} + \frac{\epsilon_\rho}{1-\gamma}
$$

Using soft Bellman equation, we have that

$$
\max_{s\in\mathcal{S},\boldsymbol{a}\in\mathcal{A}}|Q^i(s,\boldsymbol{a};\boldsymbol{\pi}) - \hat{Q}^i(s,\boldsymbol{a};\boldsymbol{\pi})| \leq \delta
$$

□

A.5 Proof of Proposition 5.3

Proof. We denote $\pi^i(a|s) = \sum_{a^{-i}\in\mathcal{A}^{-i}}\pi_i^{\theta_s}(a|a^{-i},s)\rho(a^{-i}|s)$. Then the joint policy $\boldsymbol{\pi}^{\theta_s}(\boldsymbol{a}|s) = \prod_{i\in\mathcal{N}}\pi_i^{\theta_s}(a^i|s)$. For all $i,j\in\mathcal{N}$, $i\neq j$.

$$
\mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}^{\theta_s}(\cdot|s)}\left[\nabla_{\theta_s}\pi_i^{\theta_s}(a^i|s)\nabla_{\theta_s}\pi_j^{\theta_s}(a^j|s)^T\right] = \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}^{\theta_s}(\cdot|s)}\left[\nabla_{\theta_s}\pi_i^{\theta_s}(a^i|s)\right]\left[\nabla_{\theta_s}\pi_j^{\theta_s}(a^j|s)^T\right] = 0
$$

Then the Fisher matrix

$$
\mathcal{F}(\theta_s) = \mathbb{E}[\nabla_{\theta_s}\log\boldsymbol{\pi}^{\theta}(\boldsymbol{a}|s)\nabla_{\theta_s}\log\boldsymbol{\pi}^{\theta_s}(\boldsymbol{a}|s)^T] = \sum_{i\in\mathcal{N}}\mathbb{E}_{a^i\sim\pi_i^{\theta_s}(a^i|s)}[\nabla_{\theta_s}\log\pi_i^{\theta_s}(a^i|s)\nabla_{\theta_s}\log\pi^{i,\theta_s}(a^i|s)^T].
$$

Therefore $\mathcal{F}(\theta_s)$ is a block-diagonal matrix, and each block is corresponding to the policy parameter of an agent. Since the pseudo-inverse of a block-diagonal matrix is block-diagonal with the pseudo-inverse of each block of the original matrix, VPG has the same dynamics as global NPG. □

A.6 Proof of Lemma 5.4

Proof. As the gradient of the value functions equals the potential function, we prove the smoothness of value functions. Define the $\tilde{\Phi}^i(s,\boldsymbol{\pi}) = \mathbb{E}[\sum_{t=0}^{\infty}\gamma^t r^i(s_t,\boldsymbol{a}_t)]$, where the expectation is taken with respect to $\boldsymbol{a}_t\sim\boldsymbol{\pi}_i(\cdot|s_t)$, $s_{t+1}\sim P(\cdot|s_t,\boldsymbol{a}_t)$. $n=2$ is the number of players. The value function can be decomposed

$$
\Phi(s,\boldsymbol{\pi}) = \tilde{\Phi}^i(s,\boldsymbol{\pi}) + \mathcal{H}(\boldsymbol{\pi}) \tag{17}
$$

where $\mathcal{H}(\boldsymbol{\pi}) = -\mathbb{E}[\sum_{t=0}^{\infty}\gamma^t\boldsymbol{\pi}(\boldsymbol{a}_t|s_t)\log\boldsymbol{\pi}(\boldsymbol{a}_t|s_t)]$. We first bound the smoothness of $\tilde{\Phi}^i(s,\boldsymbol{\pi})$. Let $\boldsymbol{\pi}^\alpha := \boldsymbol{\pi}^{\theta+\alpha u}$, where u is a unit vector.

$$
\left.\frac{d\boldsymbol{\pi}^\alpha(\boldsymbol{a}\mid s)}{d\alpha}\right|_{\alpha=0} = \boldsymbol{\pi}(\boldsymbol{a}|s)\sum_{i\in\mathcal{N}}\sum_{a'\in\mathcal{A}_i}u_{i,s,a',a^{-i}}(\mathbf{I}_{a'=a^i} - \pi(a'|s,a^{-i}))
$$

$$\left|\frac{d\boldsymbol{\pi}^{\alpha}(\boldsymbol{a}\mid s)}{d\alpha}\bigg|_{\alpha=0}\right| = \left|\boldsymbol{\pi}(\boldsymbol{a}|s)\sum_{i\in\mathcal{N}}\sum_{a'\in\mathcal{A}_i} u_{i,s,a',a^{-i}}(\mathbf{I}_{a'=a^i} - \pi^i(\cdot|s,a^{-i}))\right|$$
$$\leq \boldsymbol{\pi}(\boldsymbol{a}|s)n\left(|u_{i,s,a^i,a^{-i}}| + \sum_{i\in\mathcal{N}}\sum_{a'\in\mathcal{A}_i}|u_{i,s,a^i,a^{-i}}\pi^i(a'|s,a^{-i})|\right)$$
$$\leq (n+1)\boldsymbol{\pi}(\boldsymbol{a}|s)$$

$$\begin{aligned}\frac{d^2\boldsymbol{\pi}^{\alpha}(\boldsymbol{a}\mid s')}{(d\alpha)^2}\bigg|_{\alpha=0} =& \boldsymbol{\pi}(\boldsymbol{a}|s)(u_{i,s,a^i,a^{-i}}u_{j,s,a^j,a^{-j}} - \sum_{i,j\in\mathcal{N}}\sum_{a'\in\mathcal{A}_j} u_{i,s,a^i,a^{-i}}u_{j,s,a',a^{-j}}\pi_j(a'|s,a^{-j}) \\ &- \sum_{i,j\in\mathcal{N}}\sum_{a'\in\mathcal{A}_i} u_{j,s,a',a^{-i}}u_{i,s,a^j,a^{-h}}\pi^i(a'|s,a^{-i}) \\ &+ 2\sum_{i,j\in\mathcal{N}}\sum_{a'\in\mathcal{A}_i}\sum_{a''\in\mathcal{A}_j} u_{i,s,a',a^{-i}}u_{j,s,a'',a^{-j}}\pi^i(a'|s,a^{-i})\pi_j(a''|s,a^{-j}) \\ &+ \sum_{i\in\mathcal{N}}\sum_{a'\in\mathcal{A}_i} u^2_{i,s,a',a^{-i}}\pi^i(a'|s,a^{-i}))\end{aligned} \tag{18}$$

$$\begin{aligned}\left|\frac{d^2\boldsymbol{\pi}^{\alpha}(\boldsymbol{a}\mid s')}{(d\alpha)^2}\bigg|_{\alpha=0}\right| \leq& \boldsymbol{\pi}(\boldsymbol{a}|s)(u_{i,s,a^i,a^{-i}}u_{j,s,a^j,a^{-j}} + \left|\sum_{i,j\in\mathcal{N}}\sum_{a'\in\mathcal{A}_j} u_{i,s,a^i,a^{-i}}u_{j,s,a',a^{-j}}\pi_j(a'|s,a^{-j})\right| \\ &+ \left|\sum_{i,j\in\mathcal{N}}\sum_{a'\in\mathcal{A}_i} u_{j,s,a',a^{-i}}u_{i,s,a^j,a^{-h}}\pi^i(a'|s,a^{-i})\right| \\ &+ 2\left|\sum_{i,j\in\mathcal{N}}\sum_{a'\in\mathcal{A}_i}\sum_{a''\in\mathcal{A}_j} u_{i,s,a',a^{-i}}u_{j,s,a'',a^{-j}}\pi^i(a'|s,a^{-i})\pi_j(a''|s,a^{-j})\right| \\ &+ \left|\sum_{i\in\mathcal{N}}\sum_{a'\in\mathcal{A}_i} u^2_{i,s,a',a^{-i}}\pi^i(a'|s,a^{-i}))\right| \\ \leq& 2(1+n+n^2)\boldsymbol{\pi}(\boldsymbol{a}|s)\end{aligned} \tag{19}$$

Let $\tilde{P}(\alpha)$ be the state-action transition matrix under $\boldsymbol{\pi}$,

$$[\widetilde{P}(\alpha)]_{(s,\boldsymbol{a})\to(s',\boldsymbol{a}')} = \boldsymbol{\pi}^{\alpha}(\boldsymbol{a}'\mid s')\,P(s'\mid s,\boldsymbol{a}).$$

We can differentiate $\widetilde{P}(\alpha)$ w.r.t α to get

$$\left[\frac{d\widetilde{P}(\alpha)}{d\alpha}\bigg|_{\alpha=0}\right]_{(s,\boldsymbol{a})\to(s',\boldsymbol{a}')} = \frac{d\boldsymbol{\pi}^{\alpha}(\boldsymbol{a}'\mid s')}{d\alpha}\bigg|_{\alpha=0} P(s'\mid s,\boldsymbol{a}).$$

For an arbitrary vector x,

$$\left[\frac{d\widetilde{P}(\alpha)}{d\alpha}\bigg|_{\alpha=0} x\right]_{s,\boldsymbol{a}} = \sum_{\boldsymbol{a}',s'}\frac{d\boldsymbol{\pi}^{\alpha}(\boldsymbol{a}'\mid s')}{d\alpha}\bigg|_{\alpha=0} P(s'\mid s,\boldsymbol{a})\,x_{\boldsymbol{a}',s'}$$

$$\begin{aligned}
\max_{\|u\|_2=1}\left|\left[\left.\frac{d\widetilde{P}(\alpha)}{d\alpha}\right|_{\alpha=0} x\right]_{s,\boldsymbol{a}}\right| &= \max_{\|u\|_2=1}\left|\sum_{a',s'}\left.\frac{d\boldsymbol{\pi}^{\alpha}\left(\boldsymbol{a}' \mid s'\right)}{d\alpha}\right|_{\alpha=0} P\left(s' \mid s,\boldsymbol{a}\right) x_{\boldsymbol{a}',s'}\right| \\
&\leq \sum_{\boldsymbol{a}',s'}\left|\left.\frac{d\boldsymbol{\pi}^{\alpha}\left(\boldsymbol{a}' \mid s'\right)}{d\alpha}\right|_{\alpha=0}\left|P\left(s' \mid s,\boldsymbol{a}\right)\right| x_{\boldsymbol{a}',s'}\right| \\
&\leq \sum_{s'} P\left(s' \mid s,\boldsymbol{a}\right)\|x\|_{\infty}\sum_{\boldsymbol{a}'}\left|\left.\frac{d\boldsymbol{\pi}^{\alpha}\left(\boldsymbol{a}' \mid s'\right)}{d\alpha}\right|_{\alpha=0}\right| \\
&\leq \sum_{s'} P\left(s' \mid s,\boldsymbol{a}\right)\|x\|_{\infty}(n+1) \\
&\leq (n+1)\|x\|_{\infty}.
\end{aligned}$$

By definition of ℓ_∞ norm,

$$\max_{\|u\|_2=1}\left\|\frac{d\widetilde{P}(\alpha)}{d\alpha}x\right\|_{\infty} \leq (n+1)\|x\|_{\infty}$$

Similarly, we get

$$\left[\left.\frac{d^2\widetilde{P}(\alpha)}{(d\alpha)^2}\right|_{\alpha=0}\right]_{(s,\boldsymbol{a})\to(s',\boldsymbol{a}')} = \left.\frac{d^2\boldsymbol{\pi}^{\alpha}\left(\boldsymbol{a}' \mid s'\right)}{(d\alpha)^2}\right|_{\alpha=0} P\left(s' \mid s,\boldsymbol{a}\right).$$

An identical argument leads to that, for arbitrary x,

$$\max_{\|u\|_2=1}\left\|\left.\frac{d^2\tilde{P}(\alpha)}{(d\alpha)^2}\right|_{\alpha=0} x\right\|_{\infty} \leq 2(1+n+n^2)\|x\|_{\infty}$$

$$\max_{\|u\|_2=1}\|(I-\gamma\tilde{P}(\alpha))^{-1}x\|_{\infty} = \|\sum_{n=0}^{\infty}\gamma^n\tilde{P}(\alpha)^n x\|_{\infty} \leq \frac{1}{1-\gamma}\|x\|_{\infty} \tag{20}$$

Denote $Q^i(s,\boldsymbol{a};\boldsymbol{\pi}^{\alpha})$ as the action value function of $\boldsymbol{\pi}^{\alpha}$.

$$\max_{\|u\|_2=1}\left|\frac{dQ^i(s,\boldsymbol{a};\boldsymbol{\pi}^{\alpha})}{d\alpha}\right| = \max_{\|u\|_2=1}\gamma\left|e_{i,s,\boldsymbol{a}}^T(I-\gamma\tilde{P}(\alpha))^{-1}\frac{d\widetilde{P}(0)}{d\alpha}(I-\gamma\tilde{P}(\alpha))^{-1}r\right| \leq \frac{\gamma(n+1)}{(1-\gamma)^2} \tag{21}$$

where r is the reward function.

$$\begin{aligned}
&\max_{\|u\|_2=1}\left|\frac{d^2Q^i(s,\boldsymbol{a};\boldsymbol{\pi}^{\alpha})}{d\alpha^2}\right| \\
&=\max_{\|u\|_2=1}\left|2\gamma^2 e_{i,s,\boldsymbol{a}}^T(I-\gamma\tilde{P}(\alpha))^{-1}\frac{d\widetilde{P}(0)}{d\alpha}(I-\gamma\tilde{P}(\alpha))^{-1}\frac{d\widetilde{P}(0)}{d\alpha}(I-\gamma\tilde{P}(\alpha))^{-1}\right. \\
&\quad+\left|\gamma(I-\gamma\tilde{P}(\alpha))^{-1}\frac{d^2\widetilde{P}(0)}{d\alpha^2}(I-\gamma\tilde{P}(\alpha))^{-1}\right| \\
&\leq\max_{\|u\|_2=1}\left|2\gamma^2 e_{i,s,\boldsymbol{a}}^T(I-\gamma\tilde{P}(\alpha))^{-1}\frac{d\widetilde{P}(0)}{d\alpha}(I-\gamma\tilde{P}(\alpha))^{-1}\frac{d\widetilde{P}(0)}{d\alpha}(I-\gamma\tilde{P}(\alpha))^{-1}\right| \\
&\quad+\left|\gamma(I-\gamma\tilde{P}(\alpha))^{-1}\frac{d^2\widetilde{P}(0)}{d\alpha^2}(I-\gamma\tilde{P}(\alpha))^{-1}\right| \\
&\leq\frac{2\gamma^2(n+1)^2}{(1-\gamma)^3}+\frac{2\gamma(1+n+n^2)}{(1-\gamma)^2}
\end{aligned} \tag{22}$$

$$\tilde{\Phi}(s,\boldsymbol{\pi}^{\alpha}) = \sum_{\boldsymbol{a}\in\mathcal{A}} \boldsymbol{\pi}^{\alpha}(\boldsymbol{a}|s) Q^{i}(s,\boldsymbol{a};\boldsymbol{\pi}^{\alpha}) \tag{23}$$

$$\begin{aligned}\left.\frac{d^2\tilde{\Phi}(s,\boldsymbol{\pi}^{\alpha})}{d\alpha^2}\right|_{\alpha=0} &= \sum_{\boldsymbol{a}\in\mathcal{A}} \boldsymbol{\pi}^{\alpha}(\boldsymbol{a}|s) \left.\frac{d^2 Q^i(s,\boldsymbol{a};\boldsymbol{\pi}^{\alpha})}{d\alpha^2}\right|_{\alpha=0} + \sum_{\boldsymbol{a}\in\mathcal{A}} \left.\frac{d^2\boldsymbol{\pi}^{\alpha}(\boldsymbol{a}\mid s)}{d\alpha^2}\right|_{\alpha=0} Q^i(s,\boldsymbol{a};\boldsymbol{\pi}^{\alpha}) \\ &+ 2\sum_{\boldsymbol{a}\in\mathcal{A}} \left.\frac{d\boldsymbol{\pi}^{\alpha}(\boldsymbol{a}\mid s)}{d\alpha}\right|_{\alpha=0} \left.\frac{dQ^i(s,\boldsymbol{a};\boldsymbol{\pi}^{\alpha})}{d\alpha}\right|_{\alpha=0}\end{aligned} \tag{24}$$

$$\begin{aligned}\left|\left.\frac{d^2\tilde{\Phi}(s,\boldsymbol{\pi}^{\alpha})}{d\alpha^2}\right|_{\alpha=0}\right| \leq &\left|\sum_{\boldsymbol{a}\in\mathcal{A}} \boldsymbol{\pi}^{\alpha}(\boldsymbol{a}|s) \left.\frac{d^2 Q^i(s,\boldsymbol{a};\boldsymbol{\pi}^{\alpha})}{d\alpha^2}\right|_{\alpha=0}\right| + \left|\sum_{\boldsymbol{a}\in\mathcal{A}} \left.\frac{d^2\boldsymbol{\pi}^{\alpha}(\boldsymbol{a}\mid s)}{d\alpha^2}\right|_{\alpha=0} Q^i(s,\boldsymbol{a};\boldsymbol{\pi}^{\alpha})\right| \\ &+ 2\left|\sum_{\boldsymbol{a}\in\mathcal{A}} \left.\frac{d\boldsymbol{\pi}^{\alpha}(\boldsymbol{a}\mid s)}{d\alpha}\right|_{\alpha=0}\right| \left|\left.\frac{dQ^i(s,\boldsymbol{a};\boldsymbol{\pi}^{\alpha})}{d\alpha}\right|_{\alpha=0}\right| \\ \leq &\frac{2\gamma^2(n+1)^2}{(1-\gamma)^3} + \frac{2\gamma(1+n+n^2)}{(1-\gamma)^2} + \frac{2(1+n+n^2)}{1-\gamma} + \frac{2\gamma(n+1)^2}{(1-\gamma)^2} \\ < &\frac{2(n+1)^2}{(1-\gamma)^3}\end{aligned} \tag{25}$$

The second step is to bound the smoothness of $\mathcal{H}(\boldsymbol{\pi})$.

$$-(\boldsymbol{\pi}^{\alpha})^T \log \boldsymbol{\pi}^{\alpha} = -\,(\boldsymbol{\pi}^{\alpha})^T(\theta+\alpha u) + n\log\sum_{\boldsymbol{a}\in\mathcal{A}}\exp(\theta+\alpha u) \tag{26}$$

$$\begin{aligned}\left|-\left.\frac{d^2(\boldsymbol{\pi}^{\alpha})^T\log\boldsymbol{\pi}^{\alpha}}{d\alpha^2}\right|_{\alpha=0}\right| \leq &\sum_{\boldsymbol{a}\in\mathcal{A}}\left|\left.\frac{d^2\boldsymbol{\pi}^{\alpha}(\boldsymbol{a}\mid s)}{d\alpha^2}\right|_{\alpha=0}\theta\right| + 2\sum_{\boldsymbol{a}\in\mathcal{A}}\left|\left.\frac{d\boldsymbol{\pi}^{\alpha}(\boldsymbol{a}\mid s)}{d\alpha}\right|_{\alpha=0}\right||u| \\ &+ n\max_{i\in\mathcal{N}}\left|1^T\mathrm{diag}(\pi^i\odot u) - \pi^i(\pi^i\odot u)^T\right| \\ \leq &2(n^2+n+1)\frac{1+\log\max_{i\in\mathcal{N}}|A_i|}{1-\gamma} + 3n+2\end{aligned} \tag{27}$$

$$\begin{aligned}\left|\left.\frac{d^2\mathcal{H}(\boldsymbol{\pi}^{\alpha})}{d\alpha^2}\right|_{\alpha=0}\right| = &\left|-\left.\frac{d^2}{d\alpha^2}\mathbb{E}[\sum_{t=0}^{\infty}\gamma^t\boldsymbol{\pi}^{\alpha}(\boldsymbol{a}_t|s_t)\log\boldsymbol{\pi}^{\alpha}(\boldsymbol{a}_t|s_t)]\right|_{\alpha=0}\right| \\ \leq &\frac{1}{1-\gamma}\left\|\frac{d^2(\boldsymbol{\pi}^{\alpha})^T\log\boldsymbol{\pi}^{\alpha}}{d\alpha^2}\right\|_{\infty} \\ \leq &2(n^2+n+1)\frac{1+\log\max_{i\in\mathcal{N}}|A_i|}{(1-\gamma)^2} + \frac{3n+2}{1-\gamma} := c\end{aligned} \tag{28}$$

Therefore the potential function is $(\frac{2(n+1)^2}{(1-\gamma)^3}+c)$-smooth. □

A.7 Proof of Theorem 5.5

Proof. Since Φ is bounded, the monotone sequence $\{\Phi^{(t)}\}_{t=0}^{\infty}$ converges to fixed point. We denote $\boldsymbol{\pi}^* = \lim_{t\to\infty}\boldsymbol{\pi}^{(t)}$. Assume that $\tilde{\boldsymbol{\pi}}$ is a Nash equilibrium policy. We first derive the performance difference between $\tilde{\boldsymbol{\pi}}$ and $\boldsymbol{\pi}^*$.

$$\begin{aligned}\tilde{\Phi}(s;\tilde{\boldsymbol{\pi}}) - \tilde{\Phi}(s;\boldsymbol{\pi}^*) = &\tilde{\Phi}(s;\tilde{\boldsymbol{\pi}}) - \Phi(s;\tilde{\boldsymbol{\pi}}) + \Phi(s;\tilde{\boldsymbol{\pi}}) - \Phi(s;\boldsymbol{\pi}^*) + \Phi(s;\boldsymbol{\pi}^*) - \tilde{\Phi}(s;\boldsymbol{\pi}^*) \\ \leq &\Phi(s;\boldsymbol{\pi}^*) - \tilde{\Phi}(s;\boldsymbol{\pi}^*) \leq \frac{\log|A|}{1-\gamma}\end{aligned} \tag{29}$$

Note that opponent modelling will introduce extra estimation error. We denote the potential function derived by opponent modelling as $\hat{\Phi}$. And the policy derived using opponent modelling is $\hat{\boldsymbol{\pi}}^*$.

$$\begin{aligned}\|\Phi(\cdot;\boldsymbol{\pi}^*)-\Phi(\cdot;\hat{\boldsymbol{\pi}}^*)\|_\infty =&\|\Phi(\cdot;\boldsymbol{\pi}^*)-\hat{\Phi}(\cdot;\boldsymbol{\pi}^*)+\hat{\Phi}(\cdot;\boldsymbol{\pi}^*)-\hat{\Phi}(\cdot;\hat{\boldsymbol{\pi}}^*)+\hat{\Phi}(\cdot;\hat{\boldsymbol{\pi}}^*)-\Phi(\cdot;\hat{\boldsymbol{\pi}}^*)\|_\infty \\ \leq&\|\hat{Q}^i(s,\boldsymbol{a};\boldsymbol{\pi}^*)-Q^i(s,\boldsymbol{a};\boldsymbol{\pi}^*)\|_\infty\leq\delta\end{aligned}\tag{30}$$

Therefore the performance difference is $\delta+\frac{\log|A|}{1-\gamma}$. □

A.8 Proof of Proposition 4.7

Proof.

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{a}_{0:\infty},s_{0:\infty}\sim q}\left[\sum_{t=0}^{\infty}r^{-i}(s_t,\boldsymbol{a}_t)-\mathrm{KL}(\rho^{-i}(a_t^{-i}|s_t)||\hat{\boldsymbol{\pi}}^{-i}(\boldsymbol{a}_t^{-i}|s_t))-\mathrm{KL}(q(a_t^i|s_t)||\boldsymbol{\pi}^i(a_t^i|s_t))\middle|q\right]\\ &=\mathbb{E}_{\boldsymbol{a}_0,s_0\sim q}\left[Q_\rho^{-i}(s_0,\boldsymbol{a}_0;\rho)-\mathrm{KL}(\rho^{-i}(\boldsymbol{a}_0^{-i}|s_0)||\hat{\boldsymbol{\pi}}^{-i}(\boldsymbol{a}_0^{-i}|s_0))\middle|q\right]\\ &=\mathbb{E}_{s_0\sim q}\left[-\mathrm{KL}\left(\rho^{-i}(a_0^{-i}|s_0)||\frac{\hat{\boldsymbol{\pi}}^{-i}(\boldsymbol{a}_0^{-i}|s_0)\exp(\mathbb{E}_{a_0^i\sim q}[Q_\rho^{-i}(s_0,\boldsymbol{a}_0;\rho)])}{\mathbb{E}_{\boldsymbol{a}_0^{-i}\sim\hat{\boldsymbol{\pi}}^{-i}(\cdot|s_0)}\left[\exp(\mathbb{E}_{a_0^i\sim q}[Q_\rho^{-i}(s_0,\boldsymbol{a}_0;\rho)])\right]}\right)\middle|q\right]+\mathbb{E}_{s_0,\boldsymbol{a}_0\sim q}\left[Q_\rho^{-i}(s_0,\boldsymbol{a}_0;\rho)\right]\end{aligned}$$

From the non-negativity of KL divergence, the optimal opponent model of agent $-i$ is

$$\rho^{-i,*}(\boldsymbol{a}^{-i}|s)=\frac{\hat{\boldsymbol{\pi}}^{-i}(a^{-i}|s)\exp(\mathbb{E}_{a^i\sim q}[Q_\rho^{-i}(s,\boldsymbol{a};\rho)])}{\mathbb{E}_{\boldsymbol{a}^{-i}\sim\hat{\boldsymbol{\pi}}^{-i}(\cdot|s)}\left[\exp(\mathbb{E}_{a^i\sim q}[Q_\rho^{-i}(s,\boldsymbol{a};\rho)])\right]}$$

□

B Algorithm

B.1 Opponent modelling

Algorithm 3 Opponent modelling (OM)

input Initial the parameter of reward function ψ, trajectory replay buffer $\mathcal{D}$,
  for $i=1,2,\ldots$ **do**
    Sample trajectory τ from $\mathcal{D}$
    for $j=1$ to N **do**
      Update $\hat{r}^j(s_t,\boldsymbol{a}_t)$ using (10)
      Update $\rho_j(\tau_j)$ using (6)
    end for
  end for
output optimised opponent model ρ