Can Mean Field Control (MFC) Approximate Cooperative Multi Agent Reinforcement Learning (MARL) with Non-Uniform Interaction? (Supplementary Material)

Washim Uddin Mondal[1, 2] **Vaneet Aggarwal**[1] **Satish V. Ukkusuri**[2]

[1]School of Industrial Engineering, Purdue University, West Lafayette, Indiana, USA 47907
[2]Lyles School of Civil Engineering, Purdue University, West Lafayette, Indiana, USA 47907

A PROOF OF COROLLARY 1

The following inequalities hold $\forall x \in \mathcal{X}$, $\forall u \in \mathcal{U}$, $\forall \boldsymbol{\mu}_1 \in \mathcal{P}(\mathcal{X})$, and $\forall \boldsymbol{\nu}_1 \in \mathcal{P}(\mathcal{U})$.

$$\begin{aligned}|r(x,u,\boldsymbol{\mu}_1,\boldsymbol{\nu}_1)| &\leq |\boldsymbol{a}^T\boldsymbol{\mu}_1| + |\boldsymbol{b}^T\boldsymbol{\nu}_1| + |f(x,u)| \\ &\leq |\boldsymbol{a}|_1|\boldsymbol{\mu}_1|_1 + |\boldsymbol{b}|_1|\boldsymbol{\nu}_1|_1 + |f(x,u)| \\ &\overset{(a)}{=} |\boldsymbol{a}|_1 + |\boldsymbol{b}|_1 + |f(x,u)|\end{aligned}$$

Equality (a) follows from the fact that both $\boldsymbol{\mu}_1$ and $\boldsymbol{\nu}_1$ are probability distributions. As the sets $\mathcal{X}, \mathcal{U}$ are finite, there must exist $M_F > 0$ such that, $|f(x,u)| \leq M_F$, $\forall x \in \mathcal{X}$, $\forall u \in \mathcal{U}$. Taking $M_R = |\boldsymbol{a}|_1 + |\boldsymbol{b}|_1 + M_F$, we can establish proposition (a).

Proposition (b) follows from the fact that $\forall x \in \mathcal{X}$, $\forall u \in \mathcal{U}$, $\forall \boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathcal{P}(\mathcal{X})$, $\forall \boldsymbol{\nu}_1, \boldsymbol{\nu}_2 \in \mathcal{P}(\mathcal{U})$, the following relations hold.

$$\begin{aligned}&|r(x,u,\boldsymbol{\mu}_1,\boldsymbol{\nu}_2) - r(x,u,\boldsymbol{\mu}_2,\boldsymbol{\nu}_2)| \\ &\quad\leq |\boldsymbol{a}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)| + |\boldsymbol{b}^T(\boldsymbol{\nu}_1 - \boldsymbol{\nu}_2)| \\ &\quad\leq |\boldsymbol{a}|_1|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2|_1 + |\boldsymbol{b}|_1|\boldsymbol{\nu}_1 - \boldsymbol{\nu}_2|_1\end{aligned}$$

Taking $L_R = \max\{|\boldsymbol{a}|_1, |\boldsymbol{b}|_1\}$, we conclude the result.

B PROOF OF THEOREM 1

The following results are necessary to establish the theorem.

B.1 LIPSCHITZ CONTINUITY

In the following three lemmas, we shall establish that the functions, ν^{MF}, P^{MF} and r^{MF} defined in (8), (9) and (10) are Lipschitz continuous. In all of these lemmas, the term Π denotes the set of policies that satisfies Assumption 3. The proofs of these lemmas are delegated to Appendix C, D, and E respectively.

Lemma B.1. *If $\nu^{\mathrm{MF}}(.,.)$ is defined by (8), then $\forall \boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathcal{P}(\mathcal{X})$, $\forall \pi \in \Pi$, the following inequality holds.*

$$|\nu^{\mathrm{MF}}(\boldsymbol{\mu}_1,\pi) - \nu^{\mathrm{MF}}(\boldsymbol{\mu}_2,\pi)|_1 \leq (1+L_Q)|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2|_1$$

where L_Q is defined in Assumption 3.

Lemma B.2. *If $P^{\mathrm{MF}}(.,.)$ is defined by (9), then $\forall \boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathcal{P}(\mathcal{X})$, $\forall \pi \in \Pi$, the following inequality holds.*

$$|P^{\mathrm{MF}}(\boldsymbol{\mu}_1,\pi) - P^{\mathrm{MF}}(\boldsymbol{\mu}_2,\pi)|_1 \leq S_P|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2|_1$$

where $S_P \triangleq (1+L_Q) + L_P(2+L_Q)$.

The terms L_P, and L_Q are defined in Assumption 1, and 3 respectively.

Lemma B.3. *If $r^{\mathrm{MF}}(.,.)$ is defined by (10), then $\forall \boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathcal{P}(\mathcal{X})$, $\forall \pi \in \Pi$, the following inequality holds.*

$$|r^{\mathrm{MF}}(\boldsymbol{\mu}_1,\pi) - r^{\mathrm{MF}}(\boldsymbol{\mu}_2,\pi)|_1 \leq S_R|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2|_1$$

where $S_R \triangleq M_R(1+L_Q) + L_R(2+L_Q)$.

The terms M_R, L_R, and L_Q are defined in Corollary 1 and Assumption 3 respectively.

B.2 APPROXIMATION RESULTS

The following Lemma B.5, B.6, B.7 establish that the state, action distributions and the average reward of an N-agent system closely approximate their mean-field counterparts when N is large. All of these results use Lemma B.4 as the key ingredient.

Lemma B.4. *[Mondal et al., 2022] Assume that $\forall m \in [M]$, $\{X_{m,n}\}_{n\in[N]}$ are independent random variables that lie in the interval $[0,1]$, and satisfy the following constraint: $\sum_{m\in[M]}\mathbb{E}[X_{m,n}] = 1$, $\forall n \in [N]$. If $\{C_{m,n}\}_{m\in[M],n\in[N]}$ are constants that obey $|C_{m,n}| \leq C$, $\forall m \in [M]$, $\forall n \in [N]$, then the following inequality holds.*

$$\sum_{m\in[M]}\mathbb{E}\Big|C_{m,n}(X_{m,n} - E[X_{m,n}])\Big| \leq C\sqrt{MN}$$

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

The proofs of Lemma B.5, B.6, and B.7 have been delegated to Appendix F, G, and H respectively.

Lemma B.5. *Assume $\{\boldsymbol{\mu}_t^N, \boldsymbol{\nu}_t^N\}_{t\in\mathbb{T}}$ are empirical state and action distributions of an N-agent system defined by (1), and (2) respectively. If these distributions are generated by a sequence of policies $\boldsymbol{\pi} = \{\pi_t\}_{t\in\mathbb{T}}$, then $\forall t \in \mathbb{T}$ the following inequality holds.*

$$\mathbb{E}|\boldsymbol{\nu}_t^N - \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^N, \pi_t)|_1 \leq \frac{\sqrt{|\mathcal{U}|}}{\sqrt{N}}$$

where ν^{MF} is defined in (8).

Lemma B.6. *Assume $\{\boldsymbol{\mu}_t^N, \boldsymbol{\nu}_t^N\}_{t\in\mathbb{T}}$ are empirical state and action distributions of an N-agent system defined by (1), and (2) respectively. If these distributions are generated by a sequence of policies $\boldsymbol{\pi} = \{\pi_t\}_{t\in\mathbb{T}}$, then $\forall t \in \mathbb{T}$ the following inequality holds.*

$$\mathbb{E}|\boldsymbol{\mu}_{t+1}^N - P^{\mathrm{MF}}(\boldsymbol{\mu}_t^N, \pi_t)|_1 \leq \frac{C_P}{\sqrt{N}}\left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|}\right]$$

where P^{MF} is defined in (9)*, $C_P \triangleq 2 + L_P$, and L_P is given in Assumption 1.*

Lemma B.7. *Assume $\{\boldsymbol{\mu}_t^N, \boldsymbol{\nu}_t^N\}_{t\in\mathbb{T}}$ are empirical state and action distributions of an N-agent system defined by (1), and (2) respectively. Also, $\forall i \in [N]$, let $\{\boldsymbol{\mu}_t^{i,N}, \boldsymbol{\nu}_t^{i,N}\}$ be weighted state and action distributions defined by* (3), (4)*. If these distributions are generated by a sequence of policies $\boldsymbol{\pi} = \{\pi_t\}_{t\in\mathbb{T}}$, then $\forall t \in \mathbb{T}$ the following inequality holds.*

$$\mathbb{E}\left|\frac{1}{N}\sum_{i=1}^N r(x_t^i, u_t^i, \boldsymbol{\mu}_t^{i,N}, \boldsymbol{\nu}_t^{i,N}) - r^{\mathrm{MF}}(\boldsymbol{\mu}_t^N, \pi_t)\right| \leq C_R\frac{\sqrt{|\mathcal{U}|}}{\sqrt{N}}$$

where r^{MF} is given in (10)*, $C_R \triangleq |\boldsymbol{b}|_1 + M_F$ and M_F is such that $|f(x,u)| \leq M_F$, $\forall x \in \mathcal{X}$, $\forall u \in \mathcal{U}$. The function $f(.,.)$ and the parameter $\boldsymbol{b}$ are defined in Assumption 2. We would like to mention that M_F always exists since $\mathcal{X}$,$\mathcal{U}$ are finite.*

B.3 PROOF OF THE THEOREM

Note that,

$$\begin{aligned}
&|v_{\mathrm{MARL}}(\boldsymbol{x}_0, \boldsymbol{\pi}) - v_{\mathrm{MF}}(\boldsymbol{\mu}_0, \boldsymbol{\pi})| \\
&\overset{(a)}{=} \Bigg|\sum_{t=0}^\infty \frac{1}{N}\sum_{i=1}^N \gamma^t \mathbb{E}[r(x_t^i, u_t^i, \boldsymbol{\mu}_t^{i,N}, \boldsymbol{\nu}_t^{i,N})] \\
&- \sum_{t=0}^\infty \gamma^t r^{\mathrm{MF}}(\boldsymbol{\mu}_t, \pi_t)\Bigg| \leq J_1 + J_2
\end{aligned}$$

Equality (a) directly follows from the definitions (7) and (10). The first term J_1 can be written as follows.

$$\begin{aligned}
J_1 &\triangleq \sum_{t=0}^\infty \gamma^t \mathbb{E}\left|\frac{1}{N}\sum_{i=1}^N [r(x_t^i, u_t^i, \boldsymbol{\mu}_t^{i,N}, \boldsymbol{\nu}_t^{i,N})] - r^{\mathrm{MF}}(\boldsymbol{\mu}_t^N, \pi_t)\right| \\
&\overset{(a)}{\leq} C_R \frac{\sqrt{|\mathcal{U}|}}{\sqrt{N}} \frac{1}{1-\gamma}
\end{aligned}$$

Equation (a) is a result of Lemma B.7. The second term can be expressed as follows.

$$\begin{aligned}
J_2 &\triangleq \sum_{t=0}^\infty \gamma^t \mathbb{E}|r^{\mathrm{MF}}(\boldsymbol{\mu}_t^N, \pi_t) - r^{\mathrm{MF}}(\boldsymbol{\mu}_t, \pi_t)| \\
&\overset{(a)}{\leq} S_R \sum_{t=0}^\infty \gamma^t |\boldsymbol{\mu}_t^N - \boldsymbol{\mu}_t|_1
\end{aligned}$$

Inequality (a) follows from Lemma B.3. Observe that, $\forall t \in \mathbb{T}$,

$$\begin{aligned}
&|\boldsymbol{\mu}_{t+1}^N - \boldsymbol{\mu}_{t+1}|_1 \\
&\leq |\boldsymbol{\mu}_{t+1}^N - P^{\mathrm{MF}}(\boldsymbol{\mu}_t^N, \pi_t)|_1 + |P^{\mathrm{MF}}(\boldsymbol{\mu}_t^N, \pi_t) - \boldsymbol{\mu}_{t+1}|_1 \\
&\overset{(a)}{\leq} \frac{C_P}{\sqrt{N}}\left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|}\right] \\
&\qquad + |P^{\mathrm{MF}}(\boldsymbol{\mu}_t^N, \pi_t) - P^{\mathrm{MF}}(\boldsymbol{\mu}_t, \pi_t)|_1 \\
&\overset{(b)}{\leq} \frac{C_P}{\sqrt{N}}\left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|}\right] + S_P|\boldsymbol{\mu}_t^N - \boldsymbol{\mu}_t|_1 \\
&\overset{(c)}{\leq} \frac{C_P}{\sqrt{N}}\left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|}\right] \frac{(S_P^{t+1} - 1)}{S_P - 1}
\end{aligned}$$

Inequality (a) follows from Lemma B.6 and Eq. (9) while (b) is a result of Lemma B.2. Finally, inequality (c) can be derived by recursively applying (b). Therefore, the term J_2 can be upper bounded as follows.

$$J_2 \leq \frac{1}{\sqrt{N}}\left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|}\right] \frac{S_R C_P}{S_P - 1}\left[\frac{1}{1-\gamma S_P} - \frac{1}{1-\gamma}\right]$$

This concludes the theorem.

C PROOF OF LEMMA B.1

The following inequalities hold true.

$$
\begin{aligned}
&|\nu^{\mathrm{MF}}(\boldsymbol{\mu}_1,\pi)-\nu^{\mathrm{MF}}(\boldsymbol{\mu}_2,\pi)|_1\\
&=\left|\sum_{x\in\mathcal{X}}\pi(x,\boldsymbol{\mu}_1)\boldsymbol{\mu}_1(x)-\sum_{x\in\mathcal{X}}\pi(x,\boldsymbol{\mu}_2)\boldsymbol{\mu}_2(x)\right|_1\\
&=\sum_{u\in\mathcal{U}}\left|\sum_{x\in\mathcal{X}}\pi(x,\boldsymbol{\mu}_1)(u)\boldsymbol{\mu}_1(x)-\sum_{x\in\mathcal{X}}\pi(x,\boldsymbol{\mu}_2)(u)\boldsymbol{\mu}_2(x)\right|\\
&\leq\sum_{u\in\mathcal{U}}\left|\sum_{x\in\mathcal{X}}\pi(x,\boldsymbol{\mu}_1)(u)\boldsymbol{\mu}_1(x)-\sum_{x\in\mathcal{X}}\pi(x,\boldsymbol{\mu}_2)(u)\boldsymbol{\mu}_1(x)\right|\\
&+\sum_{u\in\mathcal{U}}\left|\sum_{x\in\mathcal{X}}\pi(x,\boldsymbol{\mu}_2)(u)\boldsymbol{\mu}_1(x)-\sum_{x\in\mathcal{X}}\pi(x,\boldsymbol{\mu}_2)(u)\boldsymbol{\mu}_2(x)\right|\\
&\leq\sum_{x\in\mathcal{X}}\boldsymbol{\mu}_1(x)\sum_{u\in\mathcal{U}}|\pi(x,\boldsymbol{\mu}_1)(u)-\pi(x,\boldsymbol{\mu}_2)(u)|\\
&+\sum_{x\in\mathcal{X}}|\boldsymbol{\mu}_1(x)-\boldsymbol{\mu}_2(x)|\sum_{u\in\mathcal{U}}\pi(x,\boldsymbol{\mu}_2)(u)\\
&\overset{(a)}{\leq}L_Q|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2|_1\sum_{x\in\mathcal{X}}\boldsymbol{\mu}_1(x)+|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2|_1\\
&\overset{(b)}{=}(1+L_Q)|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2|_1
\end{aligned}
$$

Inequality (a) is a consequence of the fact that $\pi\in\Pi$ and $\pi(x,\boldsymbol{\mu}_2)$ is a distribution. Finally, the equality (b) follows because $\boldsymbol{\mu}_1$ is a distribution. This concludes the result.

D PROOF OF LEMMA B.2

Note the following inequalities.

$$
\begin{aligned}
&|P^{\mathrm{MF}}(\boldsymbol{\mu}_1,\pi)-P^{\mathrm{MF}}(\boldsymbol{\mu}_2,\pi)|_1\\
&=\Bigg|\sum_{x\in\mathcal{X}}\sum_{u\in\mathcal{U}}P(x,u,\boldsymbol{\mu}_1,\nu^{\mathrm{MF}}(\boldsymbol{\mu}_1,\pi))\pi(x,\boldsymbol{\mu}_1)(u)\boldsymbol{\mu}_1(x)\\
&-\sum_{x\in\mathcal{X}}\sum_{u\in\mathcal{U}}P(x,u,\boldsymbol{\mu}_2,\nu^{\mathrm{MF}}(\boldsymbol{\mu}_2,\pi))\pi(x,\boldsymbol{\mu}_2)(u)\boldsymbol{\mu}_2(x)\Bigg|_1\\
&\leq J_1+J_2
\end{aligned}
$$

where the term J_1 is as follows.

$$
\begin{aligned}
J_1&\triangleq\sum_{x\in\mathcal{X}}\sum_{u\in\mathcal{U}}\pi(x,\boldsymbol{\mu}_1)(u)\boldsymbol{\mu}_1(x)\\
&\times\left|P(x,u,\boldsymbol{\mu}_1,\nu^{\mathrm{MF}}(\boldsymbol{\mu}_1,\pi))-P(x,u,\boldsymbol{\mu}_2,\nu^{\mathrm{MF}}(\boldsymbol{\mu}_2,\pi))\right|_1\\
&\overset{(a)}{\leq}\sum_{x\in\mathcal{X}}\sum_{u\in\mathcal{U}}\pi(x,\boldsymbol{\mu}_1)(u)\boldsymbol{\mu}_1(x)\\
&\times L_P\Big\{|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2|_1+|\nu^{\mathrm{MF}}(\boldsymbol{\mu}_1,\pi)-\nu^{\mathrm{MF}}(\boldsymbol{\mu}_2,\pi)|_1\Big\}\\
&\overset{(b)}{\leq}L_P(2+L_Q)|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2|_1
\end{aligned}
$$

Inequality (a) follows from Assumption 1 whereas (b) uses Lemma B.1 and the fact that $\boldsymbol{\mu}_1$, $\pi(x,\boldsymbol{\mu}_1)$ are distributions. The term J_2 is given as follows.

$$
\begin{aligned}
J_2&\triangleq\sum_{x\in\mathcal{X}}\sum_{u\in\mathcal{U}}\left|P(x,u,\boldsymbol{\mu}_2,\nu^{\mathrm{MF}}(\boldsymbol{\mu}_2,\pi))\right|_1\\
&\times\left|\pi(x,\boldsymbol{\mu}_1)(u)\boldsymbol{\mu}_1(x)-\pi(x,\boldsymbol{\mu}_2)(u)\boldsymbol{\mu}_2(x)\right|\\
&\overset{(a)}{=}\sum_{x\in\mathcal{X}}\sum_{u\in\mathcal{U}}\left|\pi(x,\boldsymbol{\mu}_1)(u)\boldsymbol{\mu}_1(x)-\pi(x,\boldsymbol{\mu}_2)(u)\boldsymbol{\mu}_2(x)\right|\\
&\leq\sum_{x\in\mathcal{X}}\boldsymbol{\mu}_1(x)\sum_{u\in\mathcal{U}}|\pi(x,\boldsymbol{\mu}_1)(u)-\pi(x,\boldsymbol{\mu}_2)(u)|\\
&+\sum_{x\in\mathcal{X}}|\boldsymbol{\mu}_1(x)-\boldsymbol{\mu}_2(x)|\sum_{u\in\mathcal{U}}\pi(x,\boldsymbol{\mu}_2)(u)\\
&\overset{(b)}{\leq}L_Q|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2|_1\sum_{x\in\mathcal{X}}\boldsymbol{\mu}_1(x)+|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2|_1\\
&\overset{(c)}{=}(1+L_Q)|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2|_1
\end{aligned}
$$

Equality (a) uses the fact that $P(x,u,\boldsymbol{\mu}_2,\nu^{\mathrm{MF}}(\boldsymbol{\mu}_2,\pi))$ is a distribution. Inequality (b) follows from Assumption 3 while equation (c) holds because $\boldsymbol{\mu}_1$ is a distribution.

E PROOF OF LEMMA B.3

The following inequalities hold true.

$$
\begin{aligned}
&|r^{\mathrm{MF}}(\boldsymbol{\mu}_1,\pi)-r^{\mathrm{MF}}(\boldsymbol{\mu}_2,\pi)|_1\\
&=\Bigg|\sum_{x\in\mathcal{X}}\sum_{u\in\mathcal{U}}r(x,u,\boldsymbol{\mu}_1,\nu^{\mathrm{MF}}(\boldsymbol{\mu}_1,\pi))\pi(x,\boldsymbol{\mu}_1)(u)\boldsymbol{\mu}_1(x)\\
&-\sum_{x\in\mathcal{X}}\sum_{u\in\mathcal{U}}r(x,u,\boldsymbol{\mu}_2,\nu^{\mathrm{MF}}(\boldsymbol{\mu}_2,\pi))\pi(x,\boldsymbol{\mu}_2)(u)\boldsymbol{\mu}_2(x)\Bigg|_1\\
&\leq J_1+J_2
\end{aligned}
$$

where the term J_1 is given as follows.

$$
\begin{aligned}
J_1&\triangleq\sum_{x\in\mathcal{X}}\sum_{u\in\mathcal{U}}\pi(x,\boldsymbol{\mu}_1)(u)\boldsymbol{\mu}_1(x)\\
&\times\left|r(x,u,\boldsymbol{\mu}_1,\nu^{\mathrm{MF}}(\boldsymbol{\mu}_1,\pi))-r(x,u,\boldsymbol{\mu}_2,\nu^{\mathrm{MF}}(\boldsymbol{\mu}_2,\pi))\right|\\
&\overset{(a)}{\leq}\sum_{x\in\mathcal{X}}\sum_{u\in\mathcal{U}}\pi(x,\boldsymbol{\mu}_1)(u)\boldsymbol{\mu}_1(x)\\
&\times L_R\Big\{|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2|_1+|\nu^{\mathrm{MF}}(\boldsymbol{\mu}_1,\pi)-\nu^{\mathrm{MF}}(\boldsymbol{\mu}_2,\pi)|_1\Big\}\\
&\overset{(b)}{\leq}L_R(2+L_Q)|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2|_1
\end{aligned}
$$

Inequality (a) follows from Corollary 1(b) whereas (b) uses Lemma B.1 and the fact that $\boldsymbol{\mu}_1$, $\pi(x,\boldsymbol{\mu}_1)$ are distributions.

The term J_2 is given as follows.

$$\begin{aligned}
J_2 &\triangleq \sum_{x\in\mathcal{X}}\sum_{u\in\mathcal{U}} \left|r(x,u,\boldsymbol{\mu}_2,\nu^{\mathrm{MF}}(\boldsymbol{\mu}_2,\pi))\right| \\
&\times \left|\pi(x,\boldsymbol{\mu}_1)(u)\boldsymbol{\mu}_1(x)-\pi(x,\boldsymbol{\mu}_2)(u)\boldsymbol{\mu}_2(x)\right| \\
&\overset{(a)}{\leq} M_R\sum_{x\in\mathcal{X}}\sum_{u\in\mathcal{U}}\left|\pi(x,\boldsymbol{\mu}_1)(u)\boldsymbol{\mu}_1(x)-\pi(x,\boldsymbol{\mu}_2)(u)\boldsymbol{\mu}_2(x)\right| \\
&\leq M_R\sum_{x\in\mathcal{X}}\boldsymbol{\mu}_1(x)\sum_{u\in\mathcal{U}}|\pi(x,\boldsymbol{\mu}_1)(u)-\pi(x,\boldsymbol{\mu}_2)(u)| \\
&+ M_R\sum_{x\in\mathcal{X}}|\boldsymbol{\mu}_1(x)-\boldsymbol{\mu}_2(x)|\sum_{u\in\mathcal{U}}\pi(x,\boldsymbol{\mu}_2)(u) \\
&\overset{(b)}{\leq} M_RL_Q|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2|_1\sum_{x\in\mathcal{X}}\boldsymbol{\mu}_1(x)+M_R|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2|_1 \\
&\overset{(c)}{=} M_R(1+L_Q)|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2|_1
\end{aligned}$$

Inequality (a) uses Corollary 1(a). Inequality (b) follows from Assumption 3 while equation (c) holds because $\boldsymbol{\mu}_1$ is a distribution. This concludes the lemma.

F PROOF OF LEMMA B.5

Applying the definitions of $\boldsymbol{\nu}_t^N$ and ν^{MF}, we can write the following.

$$\begin{aligned}
&\mathbb{E}|\boldsymbol{\nu}_t^N-\nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^N,\pi_t)|_1 \\
&\quad=\sum_{u\in\mathcal{U}}\mathbb{E}|\boldsymbol{\nu}_t^N(u)-\nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^N,\pi_t)(u)| \\
&\quad=\sum_{u\in\mathcal{U}}\mathbb{E}\left|\frac{1}{N}\sum_{i=1}^N\delta(u_t^i=u)-\sum_{x\in\mathcal{X}}\pi_t(x,\boldsymbol{\mu}_t^N)(u)\boldsymbol{\mu}_t^N(x)\right|
\end{aligned} \tag{1}$$

Similarly, using the definition of $\boldsymbol{\mu}_t^N$, we get,

$$\begin{aligned}
&\sum_{x\in\mathcal{X}}\pi_t(x,\boldsymbol{\mu}_t^N)(u)\boldsymbol{\mu}_t^N(x) \\
&=\sum_{x\in\mathcal{X}}\pi_t(x,\boldsymbol{\mu}_t^N)(u)\frac{1}{N}\sum_{i=1}\delta(x_t^i=x) \\
&=\frac{1}{N}\sum_{i=1}^N\sum_{x\in\mathcal{X}}\pi_t(x,\boldsymbol{\mu}_t^N)(u)\delta(x_t^j=x) \\
&=\frac{1}{N}\sum_{i=1}^N\pi_t(x_t^j,\boldsymbol{\mu}_t^N)
\end{aligned} \tag{2}$$

Substituting into (1), we obtain the following.

$$\begin{aligned}
&\mathbb{E}|\boldsymbol{\nu}_t^N-\nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^N,\pi_t)|_1 \\
&\quad=\frac{1}{N}\sum_{u\in\mathcal{U}}\mathbb{E}\left|\sum_{i=1}^N\delta(u_t^j=u)-\pi_t(x_t^i,\boldsymbol{\mu}_t^N)(u)\right| \\
&\quad\overset{(a)}{\leq}\frac{\sqrt{|\mathcal{U}|}}{\sqrt{N}}
\end{aligned}$$

Inequality (a) is a consequence of Lemma B.4. Particularly, we use the fact that $\forall u\in\mathcal{U}$, the random variables $\{\delta(u_t^i=u)\}_{i\in[N]}$ lie in $[0,1]$, are conditionally independent given $\boldsymbol{x}_t\triangleq\{x_t^i\}_{i\in[N]}$ (thereby given $\boldsymbol{\mu}_t^N$), and satisfy the following constraints.

$$\begin{gathered}
\mathbb{E}\left[\delta(u_t^i=u)|\boldsymbol{x}_t\right]=\pi_t(x_t^i,\boldsymbol{\mu}_t^N) \\
\sum_{u\in\mathcal{U}}\mathbb{E}\left[\delta(u_t^i=u)|\boldsymbol{x}_t\right]=1,\ \forall i\in[N]
\end{gathered}$$

G PROOF OF LEMMA B.6

Using the definition of P^{MF}, we get the following.

$$\begin{aligned}
&P^{\mathrm{MF}}(\boldsymbol{\mu}_t^N,\pi_t) \\
&=\sum_{x\in\mathcal{X}}\sum_{u\in\mathcal{U}}P(x,u,\boldsymbol{\mu}_t^N,\nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^N,\pi_t))\pi_t(x,\boldsymbol{\mu}_t^N)(u)\boldsymbol{\mu}_t^N(x) \\
&=\sum_{x\in\mathcal{X}}\sum_{u\in\mathcal{U}}P(x,u,\boldsymbol{\mu}_t^N,\nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^N,\pi_t))\pi_t(x,\boldsymbol{\mu}_t^N)(u)\boldsymbol{\mu}_t^N(x) \\
&\qquad\times\frac{1}{N}\sum_{i=1}^N\delta(x_t^i=x) \\
&=\frac{1}{N}\sum_{i=1}^N\sum_{u\in\mathcal{U}}P(x_t^i,u,\boldsymbol{\mu}_t^N,\nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^N,\pi_t))\pi_t(x_t^i,\boldsymbol{\mu}_t^N)(u)
\end{aligned}$$

Using the definition of L_1 norm, we can write the following.

$$\begin{aligned}
&\mathbb{E}\left|\boldsymbol{\mu}_{t+1}^N-P^{\mathrm{MF}}(\boldsymbol{\mu}_t^N,\pi_t)\right|_1 \\
&=\sum_{x\in\mathcal{X}}\mathbb{E}\left|\boldsymbol{\mu}_{t+1}^N(x)-P^{\mathrm{MF}}(\boldsymbol{\mu}_t^N,\pi_t)(x)\right|_1 \\
&=\frac{1}{N}\sum_{x\in\mathcal{X}}\mathbb{E}\Bigg|\sum_{i=1}^N\delta(x_{t+1}^i=x) \\
&-\sum_{i=1}^N\sum_{u\in\mathcal{U}}P(x_t^i,u,\boldsymbol{\mu}_t^N,\nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^N,\pi_t))(x)\pi_t(x_t^i,\boldsymbol{\mu}_t^N)(u)\Bigg| \\
&\leq J_1+J_2+J_3
\end{aligned}$$

The first term, J_1 is given as follows.

$$\begin{aligned}
J_1&\triangleq\frac{1}{N}\sum_{x\in\mathcal{X}}\mathbb{E}\left|\sum_{i=1}^N\delta(x_{t+1}^j=x)-P(x_t^i,u_t^i,\boldsymbol{\mu}_t^N,\boldsymbol{\nu}_t^N)(x)\right| \\
&\overset{(a)}{\leq}\frac{\sqrt{|\mathcal{X}|}}{\sqrt{N}}
\end{aligned}$$

Inequality (a) follows from Lemma B.4. Specifically, we use the fact that, $\forall x \in \mathcal{X}$, the random variables $\{\delta(x_{t+1}^i = x)\}_{i\in[N]}$ lie in $[0,1]$, are conditionally independent given $\boldsymbol{x}_t \triangleq \{x_t^i\}_{i\in[N]}$, $\boldsymbol{u}_t \triangleq \{u_t^i\}_{i\in[N]}$, (thereby given $\boldsymbol{\mu}_t^N$, $\boldsymbol{\nu}_t^N$) and satisfy the following.

$$\mathbb{E}[\delta(x_{t+1}^i = x)|\boldsymbol{x}_t, \boldsymbol{u}_t] = P(x_t^i, u_t^i, \boldsymbol{\mu}_t^N, \boldsymbol{\nu}_t^N),$$
$$\sum_{x\in\mathcal{X}} \mathbb{E}[\delta(x_{t+1}^i = x)|\boldsymbol{x}_t, \boldsymbol{u}_t] = 1,\ \forall i \in [N]$$

The second term J_2 can be expressed as follows.

$$\begin{aligned}
J_2 &\triangleq \frac{1}{N}\sum_{x\in\mathcal{X}} \mathbb{E}\Big|\sum_{i=1}^N P(x_t^i, u_t^i, \boldsymbol{\mu}_t^N, \boldsymbol{\nu}_t^N)(x) \\
&- P(x_t^i, u_t^i, \boldsymbol{\mu}_t^N, \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^N, \pi_t))(x)\Big| \\
&\leq \frac{1}{N}\sum_{i=1}^N \mathbb{E}\Big|P(x_t^i, u_t^i, \boldsymbol{\mu}_t^N, \boldsymbol{\nu}_t^N) \\
&- P(x_t^i, u_t^i, \boldsymbol{\mu}_t^N, \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^N, \pi_t))\Big|_1 \\
&\overset{(a)}{\leq} L_P \mathbb{E}|\boldsymbol{\nu}_t^N - \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^N, \pi_t)| \overset{(b)}{\leq} L_P \frac{\sqrt{|\mathcal{U}|}}{\sqrt{N}}
\end{aligned}$$

Inequality (a) follows from Assumption 1 whereas (b) results from Lemma B.5. Finally, the term J_3 is defined as follows.

$$\begin{aligned}
J_3 &\triangleq \frac{1}{N}\sum_{x\in\mathcal{X}} \mathbb{E}\Bigg|\sum_{i=1}^N P(x_t^i, u_t^i, \boldsymbol{\mu}_t^N, \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^N, \pi_t))(x) \\
&- \sum_{i=1}^N \sum_{u\in\mathcal{U}} P(x_t^i, u, \boldsymbol{\mu}_t^N, \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^N, \pi_t))(x)\pi_t(x_t^i, \boldsymbol{\mu}_t^N)(u)\Bigg| \\
&\overset{(a)}{\leq} \frac{\sqrt{|\mathcal{X}|}}{\sqrt{N}}
\end{aligned}$$

Relation (a) results from Lemma B.4. Particularly we use the fact that $\forall x \in \mathcal{X}$, $\{P(x_t^i, u_t^i, \boldsymbol{\mu}_t^N, \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^N, \pi_t))(x)\}_{i\in[N]}$ lie in the interval $[0,1]$, are conditionally independent given $\boldsymbol{x}_t \triangleq \{x_t^i\}_{i\in[N]}$ (therefore, given $\boldsymbol{\mu}_t^N$), and satisfy the following constraints.

$$\begin{aligned}
&\mathbb{E}[P(x_t^i, u_t^i, \boldsymbol{\mu}_t^N, \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^N, \pi_t))(x)|\boldsymbol{x}_t] \\
&= \sum_{u\in\mathcal{U}} P(x_t^i, u, \boldsymbol{\mu}_t^N, \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^N, \pi_t))(x)\pi_t(x_t^i, \boldsymbol{\mu}_t^N)(u), \\
&\text{and} \sum_{x\in\mathcal{X}} \mathbb{E}[P(x_t^i, u_t^i, \boldsymbol{\mu}_t^N, \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^N, \pi_t))(x)|\boldsymbol{x}_t] = 1
\end{aligned}$$

This concludes the Lemma.

H PROOF OF LEMMA B.7

Note that,

$$\begin{aligned}
&r^{\mathrm{MF}}(\boldsymbol{\mu}_t^N, \pi_t) \\
&= \sum_{x\in\mathcal{X}}\sum_{u\in\mathcal{U}} r(x, u, \boldsymbol{\mu}_t^N, \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^N, \pi_t))\pi_t(x, \boldsymbol{\mu}_t^N)(u)\boldsymbol{\mu}_t^N(x) \\
&= \sum_{x\in\mathcal{X}}\sum_{u\in\mathcal{U}} r(x, u, \boldsymbol{\mu}_t^N, \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^N, \pi_t))\pi_t(x, \boldsymbol{\mu}_t^N)(u) \\
&\qquad \times \frac{1}{N}\sum_{i=1}^N \delta(x_t^i = x) \\
&= \frac{1}{N}\sum_{u\in\mathcal{U}}\sum_{i=1}^N r(x_t^i, u, \boldsymbol{\mu}_t^N, \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^N, \pi_t))\pi_t(x_t^i, \boldsymbol{\mu}_t^N)(u) \\
&\overset{(a)}{=} \frac{1}{N}\sum_{u\in\mathcal{U}}\sum_{i=1}^N \Big[\boldsymbol{a}^T\boldsymbol{\mu}_t^N + \boldsymbol{b}^T\nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^N, \pi_t) + f(x_t^i, u)\Big] \\
&\qquad \times \pi_t(x_t^i, \boldsymbol{\mu}_t^N)(u) \\
&\overset{(b)}{=} \boldsymbol{a}^T\boldsymbol{\mu}_t^N + \boldsymbol{b}^T\nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^N, \pi_t) \\
&\qquad + \frac{1}{N}\sum_{u\in\mathcal{U}}\sum_{i=1}^N f(x_t^i, u)\pi_t(x_t^i, \boldsymbol{\mu}_t^N)(u)
\end{aligned}$$

Equality (a) follows from Assumption 2 while (b) uses the fact that $\pi_t(x_t^i, \boldsymbol{\mu}_t^N)$ is a distribution. On the other hand,

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^N r(x_t^i, u_t^i, \boldsymbol{\mu}_t^{i,N}, \boldsymbol{\nu}_t^{i,N}) \\
&= \frac{1}{N}\sum_{i=1}^N \Big[\boldsymbol{a}^T\boldsymbol{\mu}_t^{i,N} + \boldsymbol{b}^T\boldsymbol{\nu}_t^{i,N} + f(x_t^i, u_t^i)\Big] \\
&= \frac{1}{N}\sum_{i=1}^N \Bigg[\sum_{x\in\mathcal{X}} a(x)\boldsymbol{\mu}_t^{i,N}(x) + \sum_{u\in\mathcal{U}} b(u)\boldsymbol{\nu}_t^{i,N}(u)\Bigg] \\
&+ \frac{1}{N}\sum_{i=1}^N f(x_t^i, u_t^i)
\end{aligned}$$

Now the first term can be simplified as follows.

$$\begin{aligned}
&\frac{1}{N}\sum_{x\in\mathcal{X}} a(x)\sum_{i=1}^N\sum_{j=1}^N W(i,j)\delta(x_t^j = x) \\
&= \frac{1}{N}\sum_{x\in\mathcal{X}} a(x)\sum_{j=1}^N \delta(x_t^j = x)\sum_{i=1}^N W(i,j) \\
&\overset{(a)}{=} \sum_{x\in\mathcal{X}} a(x)\frac{1}{N}\sum_{j=1}^N \delta(x_t^j = x) = \boldsymbol{a}^T\boldsymbol{\mu}_t^N
\end{aligned}$$

Equality (a) follows as W is doubly-stochastic (Assumption 4). Similarly, the second term can be simplified as shown below.

$$\frac{1}{N}\sum_{u\in\mathcal{U}} b(u)\sum_{i=1}^{N}\sum_{j=1}^{N} W(i,j)\delta(u_t^j = u)$$
$$= \frac{1}{N}\sum_{u\in\mathcal{U}} b(u)\sum_{j=1}^{N}\delta(u_t^j = u)\sum_{i=1}^{N} W(i,j)$$
$$\overset{(a)}{=} \sum_{u\in\mathcal{U}} b(u)\frac{1}{N}\sum_{j=1}^{N}\delta(u_t^j = u) = \boldsymbol{b}^T\boldsymbol{\nu}_t^N$$

Equality (a) follows from Assumption 4. Therefore, we get,

$$\mathbb{E}\left|\frac{1}{N}\sum_{i=1}^{N} r(x_t^i, u_t^i, \boldsymbol{\mu}_t^{i,N}, \boldsymbol{\nu}_t^{i,N}) - r^{\mathrm{MF}}(\boldsymbol{\mu}_t^N, \pi_t)\right|$$
$$\leq |\boldsymbol{b}|_1\mathbb{E}|\boldsymbol{\nu}_t^N - \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^N, \pi_t)|_1$$
$$+ \frac{1}{N}\mathbb{E}\left|\sum_{i=1}^{N} f(x_t^i, u_t^i) - \sum_{i=1}^{N}\sum_{u\in\mathcal{U}} f(x_t^i, u)\pi_t(x_t^i, \boldsymbol{\mu}_t^N)(u)\right|$$

Using Lemma B.5, the first term can be upper bounded by $|\boldsymbol{b}|_1\sqrt{|\mathcal{U}|/N}$. The second term can be bounded as follows.

$$\frac{1}{N}\mathbb{E}\left|\sum_{i=1}^{N} f(x_t^i, u_t^i) - \sum_{i=1}^{N}\sum_{u\in\mathcal{U}} f(x_t^i, u)\pi_t(x_t^i, \boldsymbol{\mu}_t^N)(u)\right|$$
$$\leq \frac{1}{N}\sum_{u\in\mathcal{U}}\mathbb{E}\left|\sum_{i=1}^{N} f(x_t^i, u)\left[\delta(u_t^i = u) - \pi_t(x_t^i, \boldsymbol{\mu}_t^N)(u)\right]\right|$$
$$\overset{(a)}{\leq} M_F\frac{\sqrt{|\mathcal{U}|}}{\sqrt{N}}$$

The term $M_F > 0$ is such that $|f(x,u)| \leq M_F, \forall x \in \mathcal{X}$, $\forall u \in \mathcal{U}$. Such M_F always exists since $\mathcal{X}$, and $\mathcal{U}$ are finite. Equality (a) is a result of Lemma $B.4$. In particular, we use the following facts to prove this result. The random variables $\{\delta(u_t^i = u)\}_{i\in[N]}$ are conditionally independent given $\boldsymbol{x}_t \triangleq \{x_t^i\}_{i\in[N]}$ (therefore, given $\boldsymbol{\mu}_t^N$), $\forall u \in \mathcal{U}$ and they lie in the interval $[0,1]$. Moreover,

$$|f(x_t^i, u)| \leq M_F, \forall i \in [N], \forall u \in \mathcal{U},$$
$$\mathbb{E}[\delta(u_t^i = u)|\boldsymbol{x}_t] = \pi_t(x_t^i, \boldsymbol{\mu}_t^N),$$
$$\sum_{u\in\mathcal{U}}\mathbb{E}[\delta(u_t^i = u)|\boldsymbol{x}_t] = 1$$

I SAMPLING PROCEDURE

Algorithm 1 Sampling Algorithm

Input: $\boldsymbol{\mu}_0, \boldsymbol{\pi}_{\Phi_j}, P, r$

Sample $x_0 \sim \boldsymbol{\mu}_0$.
Sample $u_0 \sim \pi_{\Phi_j}(x_0, \boldsymbol{\mu}_0)$
$\boldsymbol{\nu}_0 \leftarrow \nu^{\mathrm{MF}}(\boldsymbol{\mu}_0, \pi_{\Phi_j})$ where ν^{MF} is defined in (8).
$t \leftarrow 0$
$\mathrm{FLAG} \leftarrow \mathrm{FALSE}$
while FLAG is FALSE **do**
    $\mathrm{FLAG} \leftarrow \mathrm{TRUE}$ with probability $1-\gamma$.
    Execute Update
end while
$T \leftarrow t$
Accept $(x_T, \boldsymbol{\mu}_T, u_T)$ as a sample.
$\hat{V}_{\Phi_j} \leftarrow 0$, $\hat{Q}_{\Phi_j} \leftarrow 0$
$\mathrm{FLAG} \leftarrow \mathrm{FALSE}$
$\mathrm{SumRewards} \leftarrow 0$
while FLAG is FALSE **do**
    $\mathrm{FLAG} \leftarrow \mathrm{TRUE}$ with probability $1-\gamma$.
    Execute Update
    $\mathrm{SumRewards} \leftarrow \mathrm{SumRewards} + r(x_t, u_t, \boldsymbol{\mu}_t, \boldsymbol{\nu}_t)$
end while
With probability $\frac{1}{2}$, $\hat{V}_{\Phi_j} \leftarrow \mathrm{SumRewards}$. Otherwise $\hat{Q}_{\Phi_j} \leftarrow \mathrm{SumRewards}$.
$\hat{A}_{\Phi_j}(x_T, \boldsymbol{\mu}_T, u_T) \leftarrow 2(\hat{Q}_{\Phi_j} - \hat{V}_{\Phi_j})$.

Output: $(x_T, \boldsymbol{\mu}_T, u_T)$ and $\hat{A}_{\Phi_j}(x_T, \boldsymbol{\mu}_T, u_T)$

Procedure Update:
$x_{t+1} \sim P(x_t, u_t, \boldsymbol{\mu}_t, \boldsymbol{\nu}_t)$.
$\boldsymbol{\mu}_{t+1} \leftarrow P^{\mathrm{MF}}(\boldsymbol{\mu}_t, \pi_{\Phi_j})$ where P^{MF} is defined in (9).
$u_{t+1} \sim \pi_{\Phi_j}(x_{t+1}, \boldsymbol{\mu}_{t+1})$
$\boldsymbol{\nu}_{t+1} \leftarrow \nu^{\mathrm{MF}}(\boldsymbol{\mu}_{t+1}, \pi_{\Phi_j})$
$t \leftarrow t+1$

EndProcedure

References

Washim Uddin Mondal, Mridul Agarwal, Vaneet Aggarwal, and Satish V Ukkusuri. On the approximation of cooperative heterogeneous multi-agent reinforcement learning (marl) using mean field control (mfc). *Journal of Machine Learning Research*, 23(129):1–46, 2022.