A PROOF OF COROLLARY 1

The following inequalities hold \( \forall x \in X, \forall u \in U, \forall \mu_1 \in \mathcal{P}(X), \) and \( \forall \nu_1 \in \mathcal{P}(U). \)

\[
|r(x, u, \mu_1, \nu_1)| \leq |a^T \mu_1| + |b^T \nu_1| + |f(x, u)| \\
\leq |a|_1 |\mu_1|_1 + |b|_1 |\nu_1|_1 + |f(x, u)| \\
\overset{(a)}{=} |a|_1 + |b|_1 + |f(x, u)|
\]

Equality (a) follows from the fact that both \( \mu_1 \) and \( \nu_1 \) are probability distributions. As the sets \( X, U \) are finite, there must exist \( M_F > 0 \) such that \( |f(x, u)| \leq M_F, \forall x \in X, \) \( \forall u \in U. \) Taking \( M_R = |a|_1 + |b|_1 + M_F, \) we can establish proposition (a).

Proposition (b) follows from the fact that \( \forall x \in X, \forall u \in U, \forall \mu_1, \mu_2 \in \mathcal{P}(X), \forall \nu_1, \nu_2 \in \mathcal{P}(U), \) the following relations hold.

\[
|r(x, u, \mu_1, \nu_2) - r(x, u, \mu_2, \nu_2)| \\
\leq |a^T (\mu_1 - \mu_2)| + |b^T (\nu_1 - \nu_2)| \\
\leq |a|_1 |\mu_1 - \mu_2|_1 + |b|_1 |\nu_1 - \nu_2|_1
\]

Taking \( L_R = \max\{|a|_1, |b|_1\}, \) 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, \( \nu^{MF}, P^{MF} \) and \( \tau^{MF} \) defined in [8], [9] and [10] are Lipschitz continuous. In all of these lemmas, the term \( \Pi \) denotes the set of policies that satisfies Assumption 3. The proofs of these lemmas are delegated to Appendix B.4 and B.5 respectively.

\[
\text{Lemma B.1. If } \nu^{MF}(\cdot, \cdot) \text{ is defined by [8], then } \forall \mu_1, \mu_2 \in \mathcal{P}(X), \forall \pi \in \Pi, \text{ the following inequality holds.}
\]

\[
|\nu^{MF}(\mu_1, \pi) - \nu^{MF}(\mu_2, \pi)|_1 \leq L_Q |\mu_1 - \mu_2|_1
\]

where \( L_Q \) is defined in Assumption 3.

\[
\text{Lemma B.2. If } P^{MF}(\cdot, \cdot) \text{ is defined by [9], then } \forall \mu_1, \mu_2 \in \mathcal{P}(X), \forall \pi \in \Pi, \text{ the following inequality holds.}
\]

\[
|P^{MF}(\mu_1, \pi) - P^{MF}(\mu_2, \pi)|_1 \leq \sum_{T} |\mu_1 - \mu_2|_1
\]

where \( \sum_{T} \) is defined in Assumption 3 and Assumption 4 respectively.

\[
\text{Lemma B.3. If } \tau^{MF}(\cdot, \cdot) \text{ is defined by [10], then } \forall \mu_1, \mu_2 \in \mathcal{P}(X), \forall \pi \in \Pi, \text{ the following inequality holds.}
\]

\[
|\tau^{MF}(\mu_1, \pi) - \tau^{MF}(\mu_2, \pi)|_1 \leq \sum_{T} |\mu_1 - \mu_2|_1
\]

where \( \sum_{T} \) is defined in Assumption 3 respectively.

B.2 APPROXIMATION RESULTS

The following Lemma B.4, 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.

\[
\text{Lemma B.4.[of Mondal et al., 2022]} \text{ Assume that } \forall m \in [M], \{X_{m,n}\}_{n \in [N]} \text{ are independent random variables that lie in the interval } [0, 1], \text{ and satisfy the following constraint:}
\]

\[
\sum_{m \in [M]} E[X_{m,n}] = 1, \forall n \in [N]. \text{ If } \{C_{m,n}\}_{m \in [M], n \in [N]} \text{ are constants that obey } |C_{m,n}| \leq C, \forall m \in [M], \forall n \in [N], \text{ then the following inequality holds.}
\]

\[
\sum_{m \in [M]} E\left[C_{m,n}(X_{m,n} - E[X_{m,n}])\right] \leq C(\sqrt{MN})
\]
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 \( \{\mu_i^N, \nu_i^N\}_{i \in \mathbb{T}} \) 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 \( \pi = \{\pi_t\}_{t \in \mathbb{T}} \), then \( \forall t \in \mathbb{T} \) the following inequality holds.

\[
\mathbb{E}[\nu_t^N - \nu_{\text{MF}}(\mu_t^N, \pi_t)] \leq \frac{\sqrt{|U|}}{\sqrt{N}}
\]

where \( \nu_{\text{MF}} \) is defined in (8).

**Lemma B.6.** Assume \( \{\mu_i^N, \nu_i^N\}_{i \in \mathbb{T}} \) 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 \( \pi = \{\pi_t\}_{t \in \mathbb{T}} \), then \( \forall t \in \mathbb{T} \) the following inequality holds.

\[
\mathbb{E}[\mu_{t+1}^N - P_{\text{MF}}(\mu_t^N, \pi_t)] \leq \frac{C_P}{\sqrt{N}} \left[ \sqrt{|X|} + \sqrt{|U|} \right]
\]

where \( P_{\text{MF}} \) is defined in (9), \( C_P \triangleq 2 + L_P \), and \( L_P \) is given in Assumption 7.

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

\[
\mathbb{E}\left[ \frac{1}{N} \sum_{i=1}^{N} r(x_i^t, u_i^t, \mu_i^{1,N}, \nu_i^{1,N}) - r_{\text{MF}}(\mu_t^N, \pi_t) \right] \leq C_R \frac{\sqrt{|U|}}{\sqrt{N}}
\]

where \( r_{\text{MF}} \) is given in (10), \( C_R \triangleq |b|_1 + M_F \) 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 \( b \) 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,

\[
|v_{\text{MARL}}(x_0, \pi) - v_{\text{MF}}(\mu_0, \pi)|
\]

\[
\overset{(a)}{=} \left| \sum_{t=0}^{\infty} \frac{1}{N} \sum_{i=1}^{N} \gamma^t \mathbb{E}[r(x_i^t, u_i^t, \mu_i^N, \nu_i^N)] \right|
\]

\[
- \sum_{t=0}^{\infty} \gamma^t r_{\text{MF}}(\mu_t, \pi_t) \leq J_1 + J_2
\]

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

\[
J_1 \overset{(a)}{=} \sum_{t=0}^{\infty} \gamma^t \mathbb{E}\left[ \frac{1}{N} \sum_{i=1}^{N} \left[ r(x_i^t, u_i^t, \mu_i^N, \nu_i^N) - r_{\text{MF}}(\mu_i^N, \pi_t) \right] \right]
\]

\[
\overset{(a)}{\leq} C_R \gamma \frac{\sqrt{|U|}}{\sqrt{N}} \frac{1}{1-\gamma}
\]

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

\[
J_2 \overset{(a)}{=} \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[r_{\text{MF}}(\mu_t^N, \pi_t) - r_{\text{MF}}(\mu_t, \pi_t)]
\]

\[
\overset{(a)}{\leq} S_R \sum_{t=0}^{\infty} \gamma^t |\mu_t^N - \mu_t|_1
\]

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.

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

This concludes the theorem.
C PROOF OF LEMMA B.1

The following inequalities hold true.

\[ |\nu^{MF}(\mu_1, \pi) - \nu^{MF}(\mu_2, \pi)|_1 \]

\[ = \left| \sum_{x \in \mathcal{X}} \pi(x, \mu_1)\mu_1(x) - \sum_{x \in \mathcal{X}} \pi(x, \mu_2)\mu_2(x) \right|_1 \]

\[ = \sum_{u \in \mathcal{U}} \left| \sum_{x \in \mathcal{X}} \pi(x, \mu_1)(u)\mu_1(x) - \sum_{x \in \mathcal{X}} \pi(x, \mu_2)(u)\mu_2(x) \right| \]

\[ \leq \sum_{u \in \mathcal{U}} \left| \sum_{x \in \mathcal{X}} \pi(x, \mu_1)(u)\mu_1(x) - \sum_{x \in \mathcal{X}} \pi(x, \mu_2)(u)\mu_2(x) \right| \]

\[ + \sum_{u \in \mathcal{U}} \left| \pi(x, \mu_2)(u)\mu_1(x) - \sum_{x \in \mathcal{X}} \pi(x, \mu_2)(u)\mu_2(x) \right| \]

\[ \leq \sum_{x \in \mathcal{X}} \mu_1(x) \sum_{u \in \mathcal{U}} |\pi(x, \mu_1)(u) - \pi(x, \mu_2)(u)| \]

\[ + \sum_{x \in \mathcal{X}} |\mu_1(x) - \mu_2(x)| \sum_{u \in \mathcal{U}} \pi(x, \mu_2)(u) \]

\[ \leq L_Q |\mu_1 - \mu_2|_1 \sum_{x \in \mathcal{X}} \mu_1(x) + |\mu_1 - \mu_2|_1 \]

\[ \leq (1 + L_Q)|\mu_1 - \mu_2|_1 \]

Inequality (a) is a consequence of the fact that \( \pi \in \Pi \) and \( \pi(x, \mu_2) \) is a distribution. Finally, the equality (b) follows because \( \mu_1 \) is a distribution. This concludes the result.

D PROOF OF LEMMA B.2

Note the following inequalities.

\[ |P^{MF}(\mu_1, \pi) - P^{MF}(\mu_2, \pi)|_1 \]

\[ = \left| \sum_{x \in \mathcal{X}} \sum_{u \in \mathcal{U}} P(x, u, \mu_1, \nu^{MF}(\mu_1, \pi))\pi(x, \mu_1)(u)\mu_1(x) \right|_1 \]

\[ - \sum_{x \in \mathcal{X}} \sum_{u \in \mathcal{U}} P(x, u, \mu_2, \nu^{MF}(\mu_2, \pi))\pi(x, \mu_2)(u)\mu_2(x) \]

\[ \leq J_1 + J_2 \]

where the term \( J_1 \) is as follows.

\[ J_1 \triangleq \sum_{x \in \mathcal{X}} \sum_{u \in \mathcal{U}} |\pi(x, \mu_1)(u)\mu_1(x) \]

\[ \times \left| P(x, u, \mu_1, \nu^{MF}(\mu_1, \pi)) - P(x, u, \mu_2, \nu^{MF}(\mu_2, \pi)) \right|_1 \]

\[ \leq \sum_{x \in \mathcal{X}} \sum_{u \in \mathcal{U}} |\pi(x, \mu_1)(u)\mu_1(x) \]

\[ \times \left| \nu^{MF}(\mu_1, \pi) - \nu^{MF}(\mu_2, \pi) \right|_1 \]

\[ \leq L_P(2 + L_Q)|\mu_1 - \mu_2|_1 \]

Inequality (a) follows from Assumption [A] whereas (b) uses Lemma B.1 and the fact that \( \mu_1, \pi(x, \mu_1) \) are distributions.

E PROOF OF LEMMA B.3

The following inequalities hold true.

\[ |r^{MF}(\mu_1, \pi) - r^{MF}(\mu_2, \pi)|_1 \]

\[ = \left| \sum_{x \in \mathcal{X}} \sum_{u \in \mathcal{U}} r(x, u, \mu_1, \nu^{MF}(\mu_1, \pi))\pi(x, \mu_1)(u)\mu_1(x) \right|_1 \]

\[ - \sum_{x \in \mathcal{X}} \sum_{u \in \mathcal{U}} r(x, u, \mu_2, \nu^{MF}(\mu_2, \pi))\pi(x, \mu_2)(u)\mu_2(x) \]

\[ \leq J_1 + J_2 \]

where the term \( J_1 \) is given as follows.

\[ J_1 \triangleq \sum_{x \in \mathcal{X}} \sum_{u \in \mathcal{U}} |\pi(x, \mu_1)(u)\mu_1(x) \]

\[ \times \left| r(x, u, \mu_1, \nu^{MF}(\mu_1, \pi)) - r(x, u, \mu_2, \nu^{MF}(\mu_2, \pi)) \right|_1 \]

\[ \leq \sum_{x \in \mathcal{X}} \sum_{u \in \mathcal{U}} |\pi(x, \mu_1)(u)\mu_1(x) \]

\[ \times \left| \nu^{MF}(\mu_1, \pi) - \nu^{MF}(\mu_2, \pi) \right|_1 \]

\[ \leq L_R(2 + L_Q)|\mu_1 - \mu_2|_1 \]

Inequality (a) follows from Corollary [A] whereas (b) uses Lemma B.1 and the fact that \( \mu_1, \pi(x, \mu_1) \) are distributions.
The term $J_2$ is given as follows.

$$J_2 \triangleq \sum_{x \in X} \sum_{u \in \mathcal{U}} \left| r(x, u, \mu_2, \nu^{MF}(\mu_2, \pi)) \right| \times \left| \pi(x, \mu_1)(u) \mu_1(x) - \pi(x, \mu_2)(u) \mu_2(x) \right|$$
$$\leq M_R \sum_{x \in X} \sum_{u \in \mathcal{U}} \left| \pi(x, \mu_1)(u) \mu_1(x) - \pi(x, \mu_2)(u) \mu_2(x) \right|$$
$$\leq M_R \sum_{x \in X} |\mu_1(x) - \mu_2(x)| \sum_{u \in \mathcal{U}} \pi(x, \mu_2)(u)$$

Substituting into (1), we obtain the following.

$$\mathbb{E}[\mu_i^N - \nu^{MF}(\mu_i^N, \pi_i)|_1]$$

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

$$\mathbb{E}[\delta(x_i = u) | x_i] = \pi(x_i, \mu_i^N)$$
$$\sum_{u \in \mathcal{U}} \mathbb{E}[\delta(x_i = u) | x_i] = 1, \forall i \in [N]$$

**PROOF OF LEMMA B.5**

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

$$\mathbb{E}[\nu_i^N(\mu_i^N, \pi_i)]_1$$

Similarly, using the definition of $\mu_i^N$, we get,

$$\sum_{x \in X} \pi_t(x, \mu_i^N)(u) \mu_i^N(x)$$

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

$$\mathbb{E}[\mu_{i+1}^N - P^{MF}(\mu_i^N, \pi_i)|_1]$$

Similarly, the first term, $J_1$, is given as follows.

$$J_1 \triangleq \sum_{x \in X} \pi_t(x, \mu_i^N)(u) \delta(x_i = x)$$

$$\leq J_1 + J_2 + J_3$$

The first term, $J_1$, is given as follows.
Inequality (a) follows from Lemma [B.4]. Specifically, we use the fact that, \( \forall x \in X \), the random variables \( \{ \delta(x_i^t = x) \}_{i \in [N]} \) lie in \([0, 1]\), are conditionally independent given \( x_t \triangleq \{ x_i^t \}_{i \in [N]} \), \( u_t \triangleq \{ u_i^t \}_{i \in [N]} \) (thereby given \( \mu_t^N, \nu_t^N \)) and satisfy the following.

\[
E[\delta(x_i^{t+1} = x)|x_t, u_t] = P(x_i^t, u_i^t, \mu_t^N, \nu_t^N), \\
\sum_{x \in X} E[\delta(x_i^{t+1} = x)|x_t, u_t] = 1, \ \forall i \in [N]
\]

The second term \( J_2 \) can be expressed as follows.

\[
J_2 = \frac{1}{N} \sum_{x \in X} \sum_{i=1}^{N} \left| \sum_{i=1}^{N} P(x_i^t, u_i^t, \mu_t^N, \nu_t^N)(x) - P(x_i^t, u_i^t, \mu_t^N, \nu_t^N)(x) - P(x_i^t, u_i^t, \mu_t^N, \nu_t^N)(x) \right| \\
\leq \frac{1}{N} \sum_{i=1}^{N} \left| \sum_{x \in X} \right| \leq \sum_{x \in X} \sum_{i=1}^{N} P(x_i^t, u_i^t, \mu_t^N, \nu_t^N)(x) \\
\leq L_P \| \nu_t^N - \nu \|_{\infty, \pi_t} \leq \frac{L_P \sqrt{|U|}}{\sqrt{N}}
\]

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

\[
J_3 = \frac{1}{N} \sum_{x \in X} \left| \sum_{i=1}^{N} P(x_i^t, u_i^t, \mu_t^N, \nu_t^N)(x) \right| \\
\leq \frac{1}{N} \sum_{i=1}^{N} \left| \sum_{x \in X} \right| \leq \sum_{x \in X} \sum_{i=1}^{N} P(x_i^t, u_i^t, \mu_t^N, \nu_t^N)(x) \\
\leq \frac{1}{N} \sum_{x \in X} \sum_{i=1}^{N} f(x_i^t, u_i^t)
\]

Relation (a) results from Lemma [B.4]. Particularly we use the fact that \( \forall x \in X \), \( \{ P(x_i^t, u_i^t, \mu_t^N, \nu_t^N)(x) \}_{i \in [N]} \) lie in the interval \([0, 1]\), are conditionally independent given \( x_t \triangleq \{ x_i^t \}_{i \in [N]} \) (therefore, given \( \mu_t^N \)), and satisfy the following constraints.

\[
E[P(x_i^t, u_i^t, \mu_t^N, \nu_t^N)(x)|x_t] = \sum_{u \in U} P(x_i^t, u, \mu_t^N, \nu_t^N)(x) \pi_t(x_i^t, \mu_t^N)(u), \\
\sum_{x \in X} E[P(x_i^t, u_i^t, \mu_t^N, \nu_t^N)(x)|x_t] = 1
\]

This concludes the Lemma.

### H PROOF OF LEMMA [B.7]

Note that,

\[
r_{MF}(\mu_t^N, \nu_t^N) = \sum_{x \in X} \sum_{u \in U} r(x, u, \mu_t^N, \nu_t^N(\mu_t^N, \nu_t^N)) \pi_t(x, \mu_t^N(u) \mu_t^N(x) \\
= \sum_{x \in X} \sum_{u \in U} r(x, u, \mu_t^N, \nu_t^N(\mu_t^N, \nu_t^N)) \pi_t(x, \mu_t^N(u) \\
\times \frac{1}{N} \sum_{i=1}^{N} \delta(x_i^t = x)
\]

\[
= \frac{1}{N} \sum_{u \in U} \sum_{i=1}^{N} r(x_i^t, u, \mu_t^N, \nu_t^N(\mu_t^N, \nu_t^N)) \pi_t(x_i^t, \mu_t^N(u) \\
= \frac{1}{N} \sum_{i=1}^{N} \left[ a^T \mu_t^N + b^T \nu_t^N(\mu_t^N, \nu_t^N) + f(x_i^t, u) \right] \\
\times \pi_t(x_i^t, \mu_t^N(u) \\
= a^T \mu_t^N + b^T \nu_t^N(\mu_t^N, \nu_t^N) \\
+ \frac{1}{N} \sum_{u \in U} \sum_{i=1}^{N} f(x_i^t, u) \pi_t(x_i^t, \mu_t^N(u))
\]

Equality (a) follows from Assumption [2] while (b) uses the fact that \( \pi_t(x_i^t, \mu_t^N) \) is a distribution. On the other hand,

\[
\frac{1}{N} \sum_{i=1}^{N} r(x_i^t, u_i^t, \mu_t^N, \nu_t^N) \\
= \frac{1}{N} \sum_{i=1}^{N} \left[ a^T \mu_t^N + b^T \nu_t^N + f(x_i^t, u_i^t) \right] \\
= \frac{1}{N} \sum_{i=1}^{N} \left[ \sum_{x \in X} a(x) \pi_t(x_i^t, \mu_t^N) + \sum_{u \in U} b(u) \nu_t^N(\mu_t^N, \nu_t^N) \right] \\
+ \frac{1}{N} \sum_{i=1}^{N} f(x_i^t, u_i^t)
\]

Now the first term can be simplified as follows.

\[
\frac{1}{N} \sum_{x \in X} \sum_{i=1}^{N} \sum_{j=1}^{N} W(i, j) \delta(x_i^t = x) \\
= \frac{1}{N} \sum_{x \in X} \sum_{j=1}^{N} \delta(x_j^t = x) \sum_{i=1}^{N} W(i, j) \\
= (a) \sum_{x \in X} a(x) \frac{1}{N} \sum_{j=1}^{N} \delta(x_i^t = x) = a^T \mu_t^N
\]

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

\[ \frac{1}{N} \sum_{u \in \mathcal{U}} b(u) \sum_{i=1}^{N} \sum_{j=1}^{N} W(i,j) \delta(u_{i}^{j} = u) \]

\[ = \frac{1}{N} \sum_{u \in \mathcal{U}} b(u) \sum_{j=1}^{N} \delta(u_{j}^{j} = u) \sum_{i=1}^{N} W(i,j) \]

\[ \overset{(a)}{=} \sum_{u \in \mathcal{U}} b(u) \left( \frac{1}{N} \sum_{j=1}^{N} \delta(u_{j}^{j} = u) = b^T \nu_i \right) \]

Equality (a) follows from Assumption \[\Box\]. Therefore, we get,

\[ \mathbb{E} \left| \frac{1}{N} \sum_{i=1}^{N} f(x_{i}^{j}, u_{i}^{j}) - \sum_{i=1}^{N} \sum_{u \in \mathcal{U}} f(x_{i}^{j}, u) \pi_i(x_{i}^{j}, \mu_{i}^{N})(u) \right| \]

\[ \leq \frac{1}{N} \mathbb{E} \left| \sum_{i=1}^{N} f(x_{i}^{j}, u_{i}^{j}) - \sum_{i=1}^{N} \sum_{u \in \mathcal{U}} f(x_{i}^{j}, u) \pi_i(x_{i}^{j}, \mu_{i}^{N})(u) \right| \]

\[ \leq \frac{1}{N} \sum_{u \in \mathcal{U}} \mathbb{E} \left| \sum_{i=1}^{N} f(x_{i}^{j}, u) \left[ \delta(u_{i}^{j} = u) - \pi_i(x_{i}^{j}, \mu_{i}^{N})(u) \right] \right| \]

\[ \overset{(a)}{\leq} M_F \sqrt{\frac{\mathbb{E}[f]}{N}} \]

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 \[\Box\]. In particular, we use the following facts to prove this result. The random variables \( \{\delta(u_{i}^{j} = u)\}_{i \in [N]} \) are conditionally independent given \( x_i \overset{\text{i.i.d.}}{\approx} \mathcal{X} \) (therefore, given \( \mu_{i}^{N} \), \( \forall u \in \mathcal{U} \) and they lie in the interval \([0, 1]\). Moreover,

\[ |f(x_{i}^{j}, u)| \leq M_F, \forall i \in [N], \forall u \in \mathcal{U}, \]

\[ \mathbb{E}[\delta(u_{i}^{j} = u) | x_i] = \pi_i(x_{i}^{j}, \mu_{i}^{N}), \]

\[ \sum_{u \in \mathcal{U}} \mathbb{E}[\delta(u_{i}^{j} = u) | x_i] = 1 \]

I \quad SAMPLING PROCEDURE

Algorithm 1 Sampling Algorithm

\begin{enumerate}
\item \textbf{Input:} \( \mu_0, \pi_{\Phi_j}, P, r \)
\item Sample \( x_0 \sim \mu_0 \).
\item Sample \( u_0 \sim \pi_{\Phi_j}(x_0, \mu_0) \)
\item \( \nu_0 \sim \nuMF(\mu_0, \pi_{\Phi_j}) \) where \( \nuMF \) is defined in \[8\].
\item \( t \leftarrow 0 \)
\item \( \text{FLAG} \leftarrow \text{FALSE} \)
\item \text{while} \( \text{FLAG} \) is \text{FALSE} \text{ do}
\item \( \text{FLAG} \leftarrow \text{TRUE} \) with probability \( 1 - \gamma \).
\item Execute Update
\item \text{end while}
\item \( T \leftarrow t \)
\item Accept \( (x_T, \mu_T, u_T) \) as a sample.
\item \( \hat{V}_{\Phi_j} \leftarrow 0, \hat{Q}_{\Phi_j} \leftarrow 0 \)
\item \( \text{FLAG} \leftarrow \text{FALSE} \)
\item \( \text{SUMRewards} \leftarrow 0 \)
\item \text{while} \( \text{FLAG} \) is \text{FALSE} \text{ do}
\item \( \text{SUMRewards} \leftarrow \text{SUMRewards} + r(x_T, u_T, \mu_T, \nu_T) \)
\item \text{end while}
\item \( \hat{Q}_{\Phi_j} \leftarrow \text{SUMRewards}. \) Otherwise \( \hat{Q}_{\Phi_j} \leftarrow \text{SUMRewards} \).
\item With probability \( \frac{1}{2} \), \( \hat{V}_{\Phi_j} \leftarrow \text{SUMRewards}. \) Otherwise \( \hat{V}_{\Phi_j} \leftarrow \hat{Q}_{\Phi_j} \).
\end{enumerate}

Output: \( (x_T, \mu_T, u_T) \) and \( \hat{A}_{\Phi_j}(x_T, \mu_T, u_T) \)

Procedure Update:

\begin{enumerate}
\item \( x_{t+1} \sim P(x_t, u_t, \mu_t, \nu_t) \).
\item \( \mu_{t+1} \leftarrow \nuMF(\mu_t, \pi_{\Phi_j}) \) where \( \nuMF \) is defined in \[9\].
\item \( u_{t+1} \sim \pi_{\Phi_j}(x_{t+1}, \mu_{t+1}) \)
\item \( \nu_{t+1} \leftarrow \nuMF(\mu_{t+1}, \pi_{\Phi_j}) \)
\item \( t \leftarrow t + 1 \)
\end{enumerate}

EndProcedure

References

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