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

Abstract

Mean-Field Control (MFC) is a powerful tool to solve Multi-Agent Reinforcement Learning (MARL) problems. Recent studies have shown that MFC can well-approximate MARL when the population size is large and the agents are exchangeable. Unfortunately, the presumption of exchangeability implies that all agents uniformly interact with one another which is not true in many practical scenarios. In this article, we relax the assumption of exchangeability and model the interaction between agents via an arbitrary doubly stochastic matrix. As a result, in our framework, the mean-field 'seen' by different agents are different. We prove that, if the reward of each agent is an affine function of the mean-field seen by that agent, then one can approximate such a non-uniform MARL problem via its associated MFC problem within an error of $e = \mathcal{O}(\frac{1}{\sqrt{N}}[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|}])$ where N is the population size and $|\mathcal{X}|$, $|\mathcal{U}|$ are the sizes of state and action spaces respectively. Finally, we develop a Natural Policy Gradient (NPG) algorithm that can provide a solution to the non-uniform MARL with an error $\mathcal{O}(\max\{e, \epsilon\})$ and a sample complexity of $\mathcal{O}(\epsilon^{-3})$ for any $\epsilon > 0$.

1 INTRODUCTION

Multi-Agent Systems (MAS) are ubiquitous in the modern world. Many engineered systems such as transportation networks, power distribution and wireless communication systems can be modeled as MAS. Modeling, analysis and control of such systems to improve the overall performance is a central goal of research across multiple disciplines. Multi-Agent Reinforcement Learning (MARL) is a popular approach to achieve that target. In this article, we primarily focus on *cooperative* MARL where the goal is to determine *policies* for each individual agent such that the aggregate cumulative *reward* of the entire population is maximized. However, the sizes of joint state, and action spaces of the population grows exponentially with the number of agents. This makes the computation of the solution prohibitively hard for large MAS.

Two major computationally efficient approaches have been developed to tackle this problem. The first approach restricts its attention to local policies. In other words, it is assumed that each individual agent makes its decision solely based on its local state/observation. Algorithms that fall into this category are independent Q-learning (IQL) [Tan, 1993], centralised training and decentralised execution (CTDE) based algorithms such as VDN [Sunehag et al., 2017], QMIX [Rashid et al., 2018], WQMIX [Rashid et al., 2020], etc. Unfortunately, none of these algorithms can provide theoretical convergence guarantees. The other approach is called mean-field control (MFC) [Angiuli et al., 2022]. It is grounded on the idea that in an infinite population of homogeneous agents, it is sufficient to study the behaviour of only one representative agent in order to draw accurate conclusions about the whole population. Recent studies have shown that, if the agents are exchangeable, then MFC can be proven to be a good approximation of MARL [Gu et al., 2021].

Unfortunately, the idea of exchangeability essentially states that all agents in a population uniformly interact with each other (uniform means that all pairwise interactions are the same). This is not true in many practical scenarios. For example, in a traffic control network, the congestion at an intersection is highly influenced by the control policies adopted at its immediate neighbouring intersections. Moreover, the influence of an intersection on another intersection rapidly diminishes with increase of their separation distance. Non-uniform interaction is a hallmark characteristic of many other MASs such as social networks, wireless networks etc. In the absence of uniformity of the interaction between the agents, the framework of MFC no longer applies, and the problem becomes challenging. In this paper, we come up a new result which assures that even with non-uniform interac-

Submitted to the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022). **To be used for reviewing only**.

tions, MFC is a good choice for approximating MARL if the reward of each agent is an affine function of the mean-field distributions 'seen' by that agent. We note that the behaviour of agents in multitude of social and economic networks can be modeled via affine rewards (refer the examples given in [Chen et al., 2021]), and thus for many cases of practical interest, MFC can approximate MARL with non-uniform interactions.

1.1 CONTRIBUTIONS

We consider a non-uniform MARL setup where the pairwise interaction between the agents is described by an arbitrary doubly stochastic matrix (DSM). As a result of non-uniform interaction, the so-called mean-field effect of the population on an agent is determined by the identity of the agent. This is in stark contrast with other existing works [Gu et al., 2021, Mondal et al., 2021] where the presumption of exchangeability washes away the dependence on identity. We demonstrate that, if the reward of each agent is an affine function of the mean-field distribution 'seen' by that agent, then the standard MFC approach can approximate the non-uniform MARL with an error bound of $e \triangleq \mathcal{O}(\frac{1}{\sqrt{N}}[\sqrt{|\mathcal{X}|}+\sqrt{|\mathcal{U}|}])$, where N is the number of agents and $|\mathcal{X}|, |\mathcal{U}|$ indicate the sizes of state and action spaces of individual agent.

We would like to emphasize the importance of this result. MFC is traditionally seen as an approximation method of MARL when the agents are exchangeable and hence their interactions are uniform. Uniformity allows us to solve MFC problems by tracking only one representative agent. In this paper, we show that, under certain conditions, a non-uniform MARL can also be approximated by the MFC approach. Thus, although the non-uniform interaction is a major part of the original MARL problem, the assumed affine structure of the reward function allows us to evade non-uniformity while obtaining an approximate solution. The key result is established in Lemma 8 (Appendix B.2) where, using the affine structure of the reward function, we show that the instantaneous reward generated from non-uniform MARL can be closely approximated by MFC-generated instantaneous reward.

Finally, using the results of [Liu et al., 2020], in section 5, we design a natural policy-gradient based algorithm that can solve MFC within an error of $\mathcal{O}(\epsilon)$ for any $\epsilon > 0$, with a sample complexity of $\mathcal{O}(\epsilon^{-3})$. Invoking our approximation result, we prove that the devised algorithm can yield a solution that is $\mathcal{O}(\max\{e, \epsilon\})$ error away from the optimal MARL solution, with a sample complexity of $\mathcal{O}(\epsilon^{-3})$ for any $\epsilon > 0$.

1.2 RELATED WORKS

Single Agent RL: The classical algorithms in single agent learning include tabular Q-learning Watkins and Dayan [1992], SARSA Rummery and Niranjan [1994], etc. Although they provide theoretical guarantees, these algorithms can only be applied to small state-action space based systems due to their large memory requirements. Recently Neural Network (NN) based Q-iteration Mnih et al. [2015], and policy gradient Mnih et al. [2016] algorithms have becomes popular due to the large expressive power of NN. However, they cannot be applied to large MAS due to the exponential blow-up of joint state-space.

MFC as an Approximation to Uniform MARL: Recently, MFC is gaining traction as a scalable approximate solution to uniform MARL. On the theory side, recently it has been proven that MFC can approximate uniform MARL within an error of $\mathcal{O}(1/\sqrt{N})$ [Gu et al., 2021]. However, the result relies on the assumption that all agents are homogeneous. Later, this approximation result was extended to heterogeneous agents [Mondal et al., 2021]. We would like to clarify that the idea of heterogeneity is different from the idea of non-uniformity. In the first case, the agents are divided into multiple classes. However, the identities of different agents within a given class are irrelevant. In contrast, non-uniform interaction takes the identity of each agent into account.

Graphon Approximation: One possible approach to consider non-uniform agent interaction is the notion of Graphon mean-field, which is recently gaining popularity in the *non-cooperative* MARL setup [Caines and Huang, 2019, Cui and Koeppl, 2021]. The main idea is to approximate the finite indices of the agents as a continuum of real numbers and the discrete interaction graph between agents as a continuous, symmetric, measurable function, called graphon, in the asymptotic limit of infinite population. The unfortunate consequence of this approximation is that one is left to deal with an infinite dimensional mean-field distribution. In order to obtain practical solution from graphon-approximation, one must therefore discretise the continuum of agent indices [Cui and Koeppl, 2021], which limits the use of this approximation. Our paper establishes that for affine reward functions, we do not need to go to the complexity of Graphon approximation.

Applications of MFC: Alongside the theory, MFC has also become popular as an application tool. It has been used in ride-sharing [Al-Abbasi et al., 2019], epidemic management [Watkins et al., 2016], congestion control in road network [Wang et al., 2020] etc.

Learning Algorithms for MFC: Both model-free [Angiuli et al., 2022, Gu et al., 2021] and model-based [Pasztor et al., 2021] Q-learning algorithms have been proposed in the literature to solve uniform MARL via MFC with homogeneous agents. Recently, [Mondal et al., 2021] proposed a policy-gradient algorithm for heterogeneous-MFC.

2 COOPERATIVE MARL WITH NON-UNIFORM INTERACTION

We consider a system comprising of N interacting agents. The (finite) state and action spaces of each agent are denoted as $\mathcal{X}$, and $\mathcal{U}$ respectively. Time is assumed to belong to the discrete set, $\mathbb{T} \triangleq \{0, 1, 2, \cdots\}$. The state and action of i-th agent at time t are symbolized as x_t^i and u_t^i. The empirical state and action distributions of the population of agents at time t are denoted by $\boldsymbol{\mu}_t^N$, and $\boldsymbol{\nu}_t^N$ respectively, and defined as follows.

$$\boldsymbol{\mu}_t^N(x) \triangleq \frac{1}{N}\sum_{i=1}^{N}\delta(x_t^i = x),\ \forall x \in \mathcal{X}, \forall t \in \mathbb{T} \quad (1)$$

$$\boldsymbol{\nu}_t^N(u) \triangleq \frac{1}{N}\sum_{i=1}^{N}\delta(u_t^i = u),\ \forall u \in \mathcal{U}, \forall t \in \mathbb{T} \quad (2)$$

where $\delta(\cdot)$ is the indicator function.

Each agent, $i \in [N] \triangleq \{1, \cdots, N\}$ is endowed with a reward function r and a state transition function P that are of the following forms: $r : \mathcal{X}\times\mathcal{U}\times\mathcal{P}(\mathcal{X})\times\mathcal{P}(\mathcal{U}) \to \mathbb{R}$ and $P : \mathcal{X}\times\mathcal{U}\times\mathcal{P}(\mathcal{X})\times\mathcal{P}(\mathcal{U}) \to \mathcal{P}(\mathcal{X})$ where $\mathcal{P}(\cdot)$ is the set of all Borel probability measures over its argument. In particular, r, P take the followings as arguments: (a) the state, x_t^i and action, a_t^i of the corresponding agent and (b) the weighted state distribution $\boldsymbol{\mu}_t^{i,N}$ and the weighted action distribution $\boldsymbol{\nu}_t^{i,N}$ of the population as seen from the perspective of the agent. The terms $\boldsymbol{\mu}_t^{i,N}$ and $\boldsymbol{\nu}_t^{i,N}$ are defined as follows.

$$\boldsymbol{\mu}_t^{i,N}(x) \triangleq \sum_{j=1}^{N} W(i,j)\delta(x_t^j = x),\ \forall x \in \mathcal{X}, \forall t \in \mathbb{T} \quad (3)$$

$$\boldsymbol{\nu}_t^{i,N}(u) \triangleq \sum_{j=1}^{N} W(i,j)\delta(u_t^j = u),\ \forall u \in \mathcal{U}, \forall t \in \mathbb{T} \quad (4)$$

The function $W : [N] \times [N] \to [0, 1]$ dictates the influence of one agent on another. In particular, $W(i, j)$ specifies how j-th agent influences i-th agent's reward and transition functions. Observe that, for $\boldsymbol{\mu}_t^{i,N}$, and $\boldsymbol{\nu}_t^{i,N}$ to be probability distributions, W must be right-stochastic i.e.,

$$\sum_{j=1}^{N} W(i,j) = 1,\ \forall i \in \{1, \cdots, N\} \quad (5)$$

In summary, the reward received by the i-th agent at time t can be expressed as $r(x_t^i, u_t^i, \boldsymbol{\mu}_t^{i,N}, \boldsymbol{\nu}_t^{i,N})$. Moreover, the state of the agent at time $t + 1$ is decided by the following probability law: $x_{t+1}^i \sim P(x_t^i, u_t^i, \boldsymbol{\mu}_t^{i,N}, \boldsymbol{\nu}_t^{i,N})$. We would like to point out that, in contrast to our framework, existing works assume reward and state transition to be functions of $\boldsymbol{\mu}_t^N, \boldsymbol{\nu}_t^N$, thereby making the influence of population to be identical for every agent [Mondal et al., 2021, Gu et al., 2021]. If we take the influence function W to be uniform i.e., $W(i, j) = 1/N, \forall i, j \in [N]$, then $\forall i \in [N]$, $\boldsymbol{\mu}_t^{i,N} = \boldsymbol{\mu}_t^N$, and $\boldsymbol{\nu}_t^{i,N} = \boldsymbol{\nu}_t^N$, which forces our framework to collapse onto that described in the above mentioned papers.

At time $t \in \mathbb{T}$, each agent is also presumed to have a policy function $\pi_t : \mathcal{X} \times \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{U})$ that maps $(x_t^i, \boldsymbol{\mu}_t^{i,N})$ to a distribution over the action space, $\mathcal{U}$. In simple words, a policy function π_t is a rule that (probabilistically) dictates what action must be chosen by an agent given its current state and the mean-distribution of the population as observed by the agent. Note that the policy function is presumed to be the same for all the agents as the reward function, r and the transition function, P is taken to homogeneous across the population. Homogeneity of r, P is a common assumption in the mean-field literature [Gu et al., 2021, Vasal et al., 2021].

For a given set of initial states $\boldsymbol{x}_0 \triangleq \{x_0^i\}_{i\in\mathbb{N}}$, the value of the sequence of policies, $\boldsymbol{\pi} \triangleq \{\pi_t\}_{t\in\mathbb{T}}$, for the i-th agent is defined as follows.

$$v_i(\boldsymbol{x}_0, \boldsymbol{\pi}) \triangleq \sum_{t\in\mathbb{T}} \gamma^t \mathbb{E}\left[r\left(x_t^i, u_t^i, \boldsymbol{\mu}_t^{i,N}, \boldsymbol{\nu}_t^{i,N}\right)\right] \quad (6)$$

where $\boldsymbol{\mu}_t^{i,N}, \boldsymbol{\nu}_t^{i,N}$ are defined by $(3), (4)$, respectively, and the expectation is computed over all the state-action trajectories generated by the transition function P and the sequence of policy functions, $\boldsymbol{\pi}$. The term, $\gamma \in [0, 1]$ is called the time discount factor. We would like to emphasize that the value function v_i is dependent on the interaction matrix W (because so are $\boldsymbol{\mu}_t^{i,N}$, and $\boldsymbol{\nu}_t^{i,N}$). However, such dependence is not explicitly shown to keep the notation uncluttered. The average value function of the entire population is expressed as below.

$$v_{\mathrm{MARL}}(\boldsymbol{x}_0, \boldsymbol{\pi}) = \frac{1}{N}\sum_{i=1}^{N} v_i(\boldsymbol{x}_0, \boldsymbol{\pi}) \quad (7)$$

The goal of MARL is to maximize $v_{\mathrm{MARL}}(\boldsymbol{x}_0, .)$ over all policy sequences $\boldsymbol{\pi}$. Such an optimization is hard to solve in general, especially for large N.

Before concluding this section, we would like to point out two important observations that will be extensively used in many of our forthcoming results.

Remark 1. *$\forall t \in \mathbb{T}$, the random variables $\{u_t^i\}_{i\in[N]}$ are conditionally independent given $\{x_t^i\}_{i\in[N]}$. In other words, given current states, each agent chooses its action independent of each other.*

Remark 2. *$\forall t \in \mathbb{T}$, the random variables $\{x_{t+1}^i\}_{i\in[N]}$ are conditionally independent given $\{x_t^i\}_{i\in[N]}$, and $\{u_t^i\}_{i\in[N]}$. In other words, given current states and actions, the next state of each agent evolves independent of each other.*

3 MEAN-FIELD CONTROL

MFC is an approximation method of $N-$agent MARL that takes away many of the complexities of the later. The main

idea of MFC is to consider an infinite population of homogeneous agents, instead of a finite population as considered in MARL. The advantage of such presumption is that it allows us to draw accurate inferences about the whole population by tracking only a single representative agent. Unfortunately, as stated before, such approximation method is known to work [Gu et al., 2021] when the interactions between different agents are uniform, i.e., $W(i,j) = 1/N, \forall i,j \in [N]$. In this article, we shall show that, under certain conditions, we can show MFC as an approximation of MARL, even with non-uniform W. Below we describe the MFC method.

As explained above, in MFC, we only need to track a single representative agent. Let the state and action of the agent at time t be denoted as x_t, and u_t respectively. Also, let $\boldsymbol{\mu}_t$, $\boldsymbol{\nu}_t$ be the state, and action distributions of the infinite population at time t. The reward and state transition laws of the representative at time t are denoted as $r(x_t, u_t, \boldsymbol{\mu}_t, \boldsymbol{\nu}_t)$ and $P(x_t, u_t, \boldsymbol{\mu}_t, \boldsymbol{\nu}_t)$, respectively. For a given policy sequence $\boldsymbol{\pi} \triangleq \{\pi_t\}_{t\in\mathbb{T}}$, the action distribution $\boldsymbol{\nu}_t$ can be expressed as a deterministic function of the state distribution, $\boldsymbol{\mu}_t$ as follows.

$$\boldsymbol{\nu}_t = \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t, \pi_t) \triangleq \sum_{x\in\mathcal{X}} \pi_t(x, \boldsymbol{\mu}_t)\boldsymbol{\mu}_t(x) \tag{8}$$

In a similar fashion, the state distribution at time $t+1$ can be written as a deterministic function of $\boldsymbol{\mu}_t$ as follows.

$$\begin{aligned}\boldsymbol{\mu}_{t+1} &= P^{\mathrm{MF}}(\boldsymbol{\mu}_t, \pi_t) \\ &\triangleq \sum_{x\in\mathcal{X}}\sum_{u\in\mathcal{U}} P(x, u, \boldsymbol{\mu}_t, \nu^{\mathrm{MF}}(x, \boldsymbol{\mu}_t)) \\ &\qquad \times \pi_t(x, \boldsymbol{\mu}_t)(u)\boldsymbol{\mu}_t(x)\end{aligned} \tag{9}$$

For an initial state distribution $\boldsymbol{\mu}_0$, the value of a sequence of policies $\boldsymbol{\pi} \triangleq \{\pi_t\}_{t\in\mathbb{T}}$, is defined as written below.

$$\begin{aligned}v_{\mathrm{MF}}(\boldsymbol{\mu}_0, \boldsymbol{\pi}) &\triangleq \sum_{t\in\mathbb{T}} \gamma^t r^{\mathrm{MF}}(\boldsymbol{\mu}_t, \pi_t), \\ \text{where } r^{\mathrm{MF}}(\boldsymbol{\mu}_t, \pi_t) &\triangleq \sum_{x\in\mathcal{X}}\sum_{u\in\mathcal{U}} r(x, u, \boldsymbol{\mu}_t, \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t, \pi_t)) \\ &\qquad \times \pi_t(x, \boldsymbol{\mu}_t)(u)\boldsymbol{\mu}_t(x)\end{aligned} \tag{10}$$

The term $r^{\mathrm{MF}}(\boldsymbol{\mu}_t, \pi_t)$ indicates the average reward of the population. Alternatively, it can also be expressed as the ensemble average of the reward of the representative agent i.e., $r^{\mathrm{MF}}(\boldsymbol{\mu}_t, \pi_t) = \mathbb{E}[r(x_t, u_t, \boldsymbol{\mu}_t, \boldsymbol{\nu}_t)]$ where the expectation is computed over all possible states $x_t \sim \boldsymbol{\mu}_t$, and actions $u_t \sim \pi_t(x_t, \boldsymbol{\mu}_t)$ at time t. The mean distributions $\boldsymbol{\mu}_t, \boldsymbol{\nu}_t$ are sequentially determined by $(8), (9)$ from a given initial state distribution, $\boldsymbol{\mu}_0$.

The goal of MFC is to maximize $v_{\mathrm{MF}}(\boldsymbol{\mu}_0, \cdot)$ over all policy sequences. In the next section, we shall demonstrate that, under certain conditions, v_{MARL} is well-approximated by v_{MF}. Therefore, in order to solve MARL, it is sufficient to solve its associated MFC.

It is worthwhile to point out that $\boldsymbol{\mu}_t, \boldsymbol{\nu}_t$ can be thought of as limiting values of the empirical distributions $\boldsymbol{\mu}_t^N, \boldsymbol{\nu}_t^N$ in the asymptotic limit of infinite population. Note that, $\boldsymbol{\mu}_t^N, \boldsymbol{\nu}_t^N$ and thereby, $\boldsymbol{\mu}_t, \boldsymbol{\nu}_t$ are NOT dependent on W. This makes the MFC problem agnostic of W. In contrary, agents in the $N-$agent MARL problem are influenced by the weighted mean-field distribution $\{\boldsymbol{\mu}_t^{i,N}, \boldsymbol{\nu}_t^{i,N}\}_{i\in[N]}$ which do depend on W via $(3), (4)$. Therefore, unlike in the existing works, the mean-field representative in our case cannot be described as a randomly chosen typical agent in the limit $N \to \infty$. The concept of mean-field representative, in our work, is a useful construct that, under certain conditions, can provide well-approximated solution to MARL.

In the next section, we describe how these seemingly incompatible frameworks, namely non-uniform MARL where the behaviour of agents are dependent on W, and the framework of W-agnostic MFC, can be merged together.

4 MFC AS AN APPROXIMATION TO NON-UNIFORM MARL

Before formally stating our main result, we would like to describe the assumptions that the result is grounded upon. Our first assumption is on the structure of state-transition function.

Assumption 1. *The state-transition function P is Lipschitz continuous with parameter L_P with respect to the mean-distribution arguments. Mathematically, the inequality,*

$$\begin{aligned}&|P(x, u, \boldsymbol{\mu}_1, \boldsymbol{\nu}_1) - P(x, u, \boldsymbol{\mu}_2, \boldsymbol{\nu}_2)|_1 \\ &\qquad \leq L_P[|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2|_1 + |\boldsymbol{\nu}_1 - \boldsymbol{\nu}_2|_1]\end{aligned}$$

holds $\forall x \in \mathcal{X}, \forall u \in \mathcal{U}$, $\forall \boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathcal{P}(\mathcal{X})$ and $\forall \boldsymbol{\nu}_1, \boldsymbol{\nu}_2 \in \mathcal{P}(\mathcal{U})$. The symbol $|\cdot|_1$ denotes L_1 norm.

Assumption 1 states that the transition function, P, is Lipschitz continuous with respect to its mean-field arguments. Essentially, this implies that if the state-distribution changes from $\boldsymbol{\mu}$ to $\boldsymbol{\mu} + \Delta\boldsymbol{\mu}$, then the corresponding change in the transition-function can be bounded by a term proportional to $|\Delta\boldsymbol{\mu}|_1$. Similar property holds for the change in the action-distribution. This useful assumption commonly appears in the mean-field literature [Gu et al., 2021, Mondal et al., 2021, Carmona et al., 2018].

The second assumption is on the structure of r, the reward function.

Assumption 2. *The reward function, r is affine with respect to mean-distribution arguments. Mathematically, for some $\boldsymbol{a} \in \mathbb{R}^{|\mathcal{X}|}, \boldsymbol{b} \in \mathbb{R}^{|\mathcal{U}|}$, and $f : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$, the equality,*

$$r(x, u, \boldsymbol{\mu}, \boldsymbol{\nu}) = \boldsymbol{a}^T\boldsymbol{\mu} + \boldsymbol{b}^T\boldsymbol{\nu} + f(x, u)$$

holds $\forall x \in \mathcal{X}$, $\forall u \in \mathcal{U}$, $\forall \boldsymbol{\mu} \in \mathcal{P}(\mathcal{X})$, *and* $\forall \boldsymbol{\nu} \in \mathcal{P}(\mathcal{U})$.

Assumption 2 dictates that the reward is an affine function of the mean-field distributions. Although this assumption does not allow us to encapsulate a large variety of reward functions, we would like to point out that the behaviour of agents in multitude of social and economic networks can be modeled via affine rewards (refer the examples given in [Chen et al., 2021]). We shall provide one explicit example at the end of this section. We would also like to reiterate that the benefit of this seemingly restrictive assumption of affine reward is it allows us to apply the principles of MFC to an arbitrarily interacting N-agent system which is notoriously complex to solve in general.

The immediate corollary of Assumption 2 is that the reward function is bounded and Lipschitz continuous. The formal proposition is given below.

Corollary 1. *If the reward function,* r *satisfies Assumption* 2*, then for some* $M_R, L_R > 0$*, the following holds*

$$\begin{aligned}&(a)|r(x,u,\boldsymbol{\mu}_1,\boldsymbol{\nu}_1)| \leq M_R,\\ &(b)|r(x,u,\boldsymbol{\mu}_1,\boldsymbol{\nu}_1) - r(x,u,\boldsymbol{\mu}_2,\boldsymbol{\nu}_2)|\\ &\qquad\qquad \leq L_R\left[|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2|_1 + |\boldsymbol{\nu}_1-\boldsymbol{\nu}_2|_1\right]\end{aligned}$$

$\forall x \in \mathcal{X}$, $\forall u \in \mathcal{U}$, $\forall \boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathcal{P}(\mathcal{X})$, *and* $\forall \boldsymbol{\nu}_1, \boldsymbol{\nu}_2 \in \mathcal{P}(\mathcal{U})$.

The third assumption concerns the set of allowable policy functions.

Assumption 3. *The set of allowable policy functions,* Π*, is such that each of its element is Lipschitz continuous with respect to its mean-state distribution argument. Mathematically,* $\forall \pi \in \Pi$*, the following inequality holds*

$$|\pi(x,\boldsymbol{\mu}_1) - \pi(x,\boldsymbol{\mu}_2)|_1 \leq L_Q|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2|_1$$

for some $L_Q > 0$ *and* $\forall x \in \mathcal{X}$, $\forall \boldsymbol{\mu} \in \mathcal{P}(\mathcal{X})$.

Assumption 3 states that the allowable policy functions must be Lipschitz continuous with respect to its state-distribution argument. Such requirement typically holds for neural network based policies and are commonly presumed to be true in the literature [Gu et al., 2021, Cui and Koeppl, 2021, Pasztor et al., 2021].

The final assumption imposes some constraints on the interaction function, W.

Assumption 4. *The interaction function,* W *is such that,*

$$\sum_{i=1}^{N} W(i,j) = 1,\ \forall j \in \{1,\cdots,N\} \tag{11}$$

In conjunction with (5)*, this assumption implies that* W *is doubly-stochastic.*

Assumption 4 requires W to be an $N \times N$ doubly stochastic matrix (DSM). Such presumption is commonly applied in many multi-agent tasks, e.g., distributed consensus [Alaviani and Elia, 2019a], distributed optimization [Alaviani and Elia, 2019b], and multi-agent learning [Wai et al., 2018].

We now state our main result.

Theorem 1. *Let,* $\boldsymbol{x}_0 \triangleq \{x_0^i\}_{i\in[N]}$ *be the initial states in an* N*-agent non-uniform MARL problem and* $\boldsymbol{\mu}_0$ *be its associated empirical distribution defined by* (1)*. Assume* Π *to be a set of policies that obeys Assumption 3, and* $\boldsymbol{\pi} \triangleq \{\pi_t\}_{t\in\mathbb{T}}$ *is a sequence of policies such that* $\pi_t \in \Pi$, $\forall t \in \mathbb{T}$*. If Assumption 1, 2 and* 4 *hold, then*

$$\begin{aligned}&|v_{\mathrm{MARL}}(\boldsymbol{x}_0,\boldsymbol{\pi}) - v_{\mathrm{MF}}(\boldsymbol{\mu}_0,\boldsymbol{\pi})| \leq C_R\frac{\sqrt{|\mathcal{U}|}}{\sqrt{N}}\frac{1}{1-\gamma}\\ &+\frac{1}{\sqrt{N}}\left[\sqrt{|\mathcal{X}|}+\sqrt{|\mathcal{U}|}\right]\frac{S_RC_P}{S_P-1}\left[\frac{1}{1-\gamma S_P}-\frac{1}{1-\gamma}\right]\end{aligned} \tag{12}$$

whenever $\gamma S_P < 1$ *where* $S_P \triangleq (1+L_Q)+L_P(2+L_Q)$, $S_R \triangleq M_R(1+L_Q)+L_R(2+L_Q)$, $C_P \triangleq 2+L_P$, *and* $C_R \triangleq |\boldsymbol{b}|_1 + M_F$*. The parameters* $L_P, \boldsymbol{b}, L_Q, L_R, M_R$ *have been defined in Assumption 1, 2, 3, and Corollary 1, respectively. The term* M_F *is such that* $|f(x,u)| \leq M_F$, $\forall x \in \mathcal{X}$, $\forall u \in \mathcal{U}$ *where* f *is stated in Assumption 2. The functions* v_{MARL}*, and* v_{MF} *are defined in* (7), (10) *respectively.*

Theorem 1 has an important implication. Specifically, it states that, if reward and transition functions respectively are affine and Lipschitz continuous functions of the mean-distributions, and the interaction between the agents is described by a DSM, then the solution of MFC is at most $\mathcal{O}(1/\sqrt{N})$ error away from the solution of the non-uniform MARL problem. Therefore, the larger the number of agents, the better is the MFC-based approximation. It also described how the approximation error changes with the size of the state, and action spaces. Specifically, if all other parameters are kept fixed, then the error increases as $\mathcal{O}(\sqrt{|\mathcal{X}|}+\sqrt{|\mathcal{U}|})$. In other words, if individual state and action spaces are large, then MFC may not be a good approximation to non-uniform MARL.

Now we shall discuss one example where the reward, transition function and the interaction function satisfy Assumption 1, 2, and 4 respectively.

Example 1. *A version of this model has been adapted in [Subramanian and Mahajan, 2019] and [Chen et al., 2021]. Consider a network of N firms operated by a single operator. All of the firms produce the same product but with varying quality. A discrete set* $\mathcal{X} \triangleq \{1,2,\cdots,Q\}$ *(state-space) describes the possible levels of quality of the product. At each time instant, each firm decides whether to invest to improve the quality of its product which leads to the following action set:* $\mathcal{U} = \{0,1\}$*. If at time t, the* i*-th firm decides to invest,*

i.e., $u_t^i = 1$*, its current quality,* x_t^i*, improves according to the following transition law.*

$$x_{t+1}^i = \begin{cases} x_t^i + \left\lfloor \chi \left(1 - \frac{\bar{\boldsymbol{\mu}}_t^{i,N}}{Q}\right)(Q - x_t^i) \right\rfloor & \text{if } u_t^i = 1, \\ x_t^i & \text{otherwise} \end{cases}$$

where χ *is a uniform random variable between* $[0, 1]$*, and* $\bar{\boldsymbol{\mu}}_t^{i,N}$ *is average product quality of its* $K < N$ *neighbouring firms. The intuition is that improving product quality might be difficult if the quality maintained in the local economy is high. Formally, we assume that each firm equally influences and is influenced by* K *other firms. Hence,* $W(i, j) = 1/K$ *for all* $i, j \in [N]$ *that influence each other and* $W(i, j) = 0$ *otherwise. The local average product quality is computed as,* $\bar{\boldsymbol{\mu}}_t^{i,N} \triangleq \sum_{x \in \mathcal{X}} x \boldsymbol{\mu}_t^{i,N}(x)$ *where* $\boldsymbol{\mu}_t^{i,N}$ *is given in* (3)*. At time* t*, the* i*-th firm earns a positive reward,* αx_t^i *due to its revenue, a negative reward,* $\beta \bar{\boldsymbol{\mu}}_t^{i,N}$ *due to the average local quality, and a cost* λu_t^i *due to investment. Hence, the total reward can be expressed as follows.*

$$r(x_t^i, u_t^i, \boldsymbol{\mu}_t^{i,N}, \boldsymbol{\nu}_t^{i,N}) = \alpha x_t^i - \beta \bar{\boldsymbol{\mu}}_t^{i,N} - \lambda u_t^i$$

Clearly, in this example, Assumption 1, 2, and 4 are satisfied.

4.1 PROOF OUTLINE

In this subsection, we shall provide a brief sketch of the proof of Theorem 1.

Step 0: The difference between v_{MARL} and v_{MF} is essentially the time-discounted sum of differences between the average N-agent reward and average mean-field (MF) reward at time t. Our first goal, therefore, is to estimate the difference between these rewards.

Step 1: Average N-agent reward at t depends on weighted empirical distributions $\{\boldsymbol{\mu}_t^{i,N}\}_{i \in [N]}$, $\{\boldsymbol{\nu}_t^{i,N}\}_{i \in [N]}$ whereas average MF reward depends on the distributions $\boldsymbol{\mu}_t, \boldsymbol{\nu}_t$. To estimate their difference, we first compute the difference between average N-agent reward at t and average MF reward at the same instant generated from the distribution $\boldsymbol{\mu}_t^N$. This estimate is provided by Lemma 8 in the Appendix. Assumption 2 is invoked to establish this result.

Step 2: Next we estimate the difference between the average MF reward generated by $\boldsymbol{\mu}_t^N$ and that generated by $\boldsymbol{\mu}_t$. Lemma 4 in the Appendix bounds this difference by a term proportional to $|\boldsymbol{\mu}_t^N - \boldsymbol{\mu}_t|$.

Step 3: Using Lemma 3 and 7, we now establish a recursive relation on $|\boldsymbol{\mu}_t^N - \boldsymbol{\mu}_t|$. Via induction, we can now write this difference as a function of t.

Step 4: Finally, by computing a time-discounted sum of all the upper bounds described above, we arrive at the desired result.

5 SOLUTION OF MFC VIA NATURAL POLICY GRADIENT ALGORITHM

In this section, we develop a Natural Policy Gradient (NPG) algorithm to solve the MFC problem. By virtue of Theorem 1, it provides an approximate solution to the non-uniform MARL problem. Recall from section 3 that, in MFC, it is sufficient to track only one representative agent. At time t, that agent takes its decision u_t based on its own state x_t, and the mean-field state distribution $\boldsymbol{\mu}_t$. Thus, MFC essentially reduces to a single-agent Markov Decision Problem (MDP) with extended state space $\mathcal{X} \times \mathcal{P}(\mathcal{X})$ and action space $\mathcal{U}$. To solve MFC, it is therefore sufficient to consider only stationary policies [Puterman, 2014].

Let the set of stationary policies be denoted by Π and its elements be parameterized by $\Phi \in \mathbb{R}^{\text{d}}$. For a given policy $\pi_\Phi \in \Pi$, we shall define its sequence as $\boldsymbol{\pi}_\Phi \triangleq \{\pi_\Phi, \pi_\Phi, \cdots\}$. Let, Q_Φ be the Q-function associated with policy π_Φ. We define $Q_\Phi(x, \boldsymbol{\mu}, u)$ for arbitrary $x \in \mathcal{X}$, $\boldsymbol{\mu} \in \mathcal{P}(\mathcal{X})$, and $u \in \mathcal{U}$, as follows.

$$Q_\Phi(x, \boldsymbol{\mu}, u) \triangleq \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(x_t, u_t, \boldsymbol{\mu}_t, \boldsymbol{\nu}_t) \Big| x_0 = x, \boldsymbol{\mu}_0 = \boldsymbol{\mu}, u_0 = u\right] \tag{13}$$

where the expectation is over $u_{t+1} \sim \pi_\Phi(x_{t+1}, \boldsymbol{\mu}_{t+1})$, and $x_{t+1} \sim P(x_t, u_t, \boldsymbol{\mu}_t, \boldsymbol{\nu}_t)$, $\forall t \in \mathbb{T}$. The mean-field distributions $\{\boldsymbol{\mu}_{t+1}, \boldsymbol{\nu}_t\}_{t \in \mathbb{T}}$ are updated via deterministic update equations (8), and (9). We now define the advantage function as follows.

$$A_\Phi(x, \boldsymbol{\mu}, u) \triangleq Q_\Phi(x, \boldsymbol{\mu}, u) - \mathbb{E}[Q_\Phi(x, \boldsymbol{\mu}, u)] \tag{14}$$

where the expectation is over $u \sim \pi_\Phi(x, \boldsymbol{\mu})$.

Let, $v_{\text{MF}}^*(\boldsymbol{\mu}_0) = \sup_{\Phi \in \mathbb{R}^{\text{d}}} v_{\text{MF}}(\boldsymbol{\mu}_0, \boldsymbol{\pi}_\Phi)$ where v_{MF} is the value function of MFC problem and is defined in (10). Let, $\{\Phi_j\}_{j=1}^J$ be a sequence of parameters that are generated by the NPG algorithm [Liu et al., 2020, Agarwal et al., 2021] as follows.

$$\Phi_{j+1} = \Phi_j + \eta \mathbf{w}_j, \mathbf{w}_j \triangleq \arg\min_{\mathbf{w} \in \mathbb{R}^{\text{d}}} L_{\zeta_{\boldsymbol{\mu}_0}^{\Phi_j}}(\mathbf{w}, \Phi_j) \tag{15}$$

The term η is defined as the learning parameter. The function $L_{\zeta_{\boldsymbol{\mu}_0}^{\Phi_j}}$ and the distribution $\zeta_{\boldsymbol{\mu}_0}^{\Phi_j}$ are defined below.

$$L_{\zeta_{\boldsymbol{\mu}_0}^{\Phi'}}(\mathbf{w}, \Phi) \triangleq \mathbb{E}_{(x, \boldsymbol{\mu}, u) \sim \zeta_{\boldsymbol{\mu}_0}^{\Phi'}}\Big[\Big(A_\Phi(x, \boldsymbol{\mu}, u) - (1 - \gamma)\mathbf{w}^{\text{T}} \nabla_\Phi \log \pi_\Phi(x, \boldsymbol{\mu})(u)\Big)^2\Big], \tag{16}$$

$$\zeta_{\boldsymbol{\mu}_0}^{\Phi'}(x, \boldsymbol{\mu}, u) \triangleq \sum_{\tau=0}^{\infty} \gamma^\tau \mathbb{P}(x_\tau = x, \boldsymbol{\mu}_\tau = \boldsymbol{\mu}, u_\tau = u | x_0 = x, \boldsymbol{\mu}_0 = \boldsymbol{\mu}, u_0 = u, \boldsymbol{\pi}_{\Phi'})(1 - \gamma) \tag{17}$$

NPG update (15) indicates that, at each iteration, one must solve another minimization problem to obtain the gradient direction. It can be solved by applying a stochastic gradient descent (SGD) approach. In particular, the update equation, in this case, turns out to be the following: $\mathbf{w}_{j,l+1} = \mathbf{w}_{j,l} - \alpha \mathbf{h}_{j,l}$ [Liu et al., 2020]. The term α is the learning rate for this sub-problem. The update direction $\mathbf{h}_{j,l}$ can be defined as follows.

$$\mathbf{h}_{j,l} \triangleq \Bigg(\mathbf{w}_{j,l}^{\mathrm{T}} \nabla_{\Phi_j} \log \pi_{\Phi_j}(x, \boldsymbol{\mu})(u) - \frac{1}{1-\gamma} \hat{A}_{\Phi_j}(x, \boldsymbol{\mu}, u) \Bigg) \nabla_{\Phi_j} \log \pi_{\Phi_j}(x, \boldsymbol{\mu})(u) \tag{18}$$

where $(x, \boldsymbol{\mu}, u) \sim \zeta_{\boldsymbol{\mu}_0}^{\Phi_j}$, and $\hat{A}_{\Phi_j}$ is a unbiased estimator of A_{Φ_j}. The process to obtain the samples and the estimator has been detailed in Algorithm 2 in the Appendix I. We would like to point out that Algorithm 2 is based on Algorithm 3 of [Agarwal et al., 2021]. We summarize the whole NPG process in Algorithm 1.

Algorithm 1 Natural Policy Gradient

Input: η, α: Learning rates, J, L: Number of execution steps
$\mathbf{w}_0, \Phi_0$: Initial parameters, $\boldsymbol{\mu}_0$: Initial state distribution
Initialization: $\Phi \leftarrow \Phi_0$

```
for j ∈ {0, 1, ··· , J − 1} do
    w_{j,0} ← w_0
    for l ∈ {0, 1, ··· , L − 1} do
        Sample (x, μ, u) ~ ζ^{Φ_j}_{μ_0} and Â_{Φ_j}(x, μ, u) using Algorithm 2
        Compute h_{j,l} using (18)
        w_{j,l+1} ← w_{j,l} − α h_{j,l}
    end for
    w_j ← (1/L) Σ_{l=1}^{L} w_{j,l}
    Φ_{j+1} ← Φ_j + η w_j
end for
```

Output: $\{\Phi_1, \cdots, \Phi_J\}$: Policy parameters

The global converge of NPG is stated in Lemma 1 which is a direct consequence of Theorem 4.9 of [Liu et al., 2020]. However, the following assumptions are needed to establish the Lemma. These are similar to Assumptions 2.1, 4.2, 4.4 respectively in [Liu et al., 2020].

Assumption 5. *$\forall \Phi \in \mathbb{R}^{\mathrm{d}}$, $\forall \boldsymbol{\mu}_0 \in \mathcal{P}(\mathcal{X})$, for some $\chi > 0$, $F_{\boldsymbol{\mu}_0}(\Phi) - \chi I_{\mathrm{d}}$ is positive semi-definite where $F_{\boldsymbol{\mu}_0}(\Phi)$ can be expressed as follows.*

$$F_{\boldsymbol{\mu}_0}(\Phi) \triangleq \mathbb{E}_{(x,\boldsymbol{\mu},u)\sim\zeta_{\boldsymbol{\mu}_0}^{\Phi}} \Big[\{\nabla_\Phi \pi_\Phi(x, \boldsymbol{\mu})(u)\} \times \{\nabla_\Phi \log \pi_\Phi(x, \boldsymbol{\mu})(u)\}^{\mathrm{T}} \Big]$$

Assumption 6. *$\forall \Phi \in \mathbb{R}^{\mathrm{d}}$, $\forall \boldsymbol{\mu} \in \mathcal{P}(\mathcal{X})$, $\forall x \in \mathcal{X}$, $\forall u \in \mathcal{U}$,*

$$|\nabla_\Phi \log \pi_\Phi(x, \boldsymbol{\mu})(u)|_1 \leq G$$

for some positive constant G.

Assumption 7. *$\forall \Phi_1, \Phi_2 \in \mathbb{R}^{\mathrm{d}}$, $\forall \boldsymbol{\mu} \in \mathcal{P}(\mathcal{X})$, $\forall x \in \mathcal{X}$, $\forall u \in \mathcal{U}$,*

$$|\nabla_{\Phi_1} \log \pi_{\Phi_1}(x, \boldsymbol{\mu})(u) - \nabla_{\Phi_2} \log \pi_{\Phi_2}(x, \boldsymbol{\mu})(u)|_1 \leq M|\Phi_1 - \Phi_2|_1$$

for some positive constant M.

Assumption 8. *$\forall \Phi \in \mathbb{R}^{\mathrm{d}}$, $\forall \boldsymbol{\mu}_0 \in \mathcal{P}(\mathcal{X})$,*

$$L_{\zeta_{\boldsymbol{\mu}_0}^{\Phi^*}}(\mathbf{w}_\Phi^*, \Phi) \leq \epsilon_{\mathrm{bias}}, \quad \mathbf{w}_\Phi^* \triangleq \arg\min_{\mathbf{w}\in\mathbb{R}^{\mathrm{d}}} L_{\zeta_{\boldsymbol{\mu}_0}^{\Phi}}(\mathbf{w}, \Phi)$$

where Φ^* is the parameter of the optimal policy.

Lemma 1. *Let $\{\Phi_j\}_{j=1}^J$ be the sequence of policy parameters obtained from Algorithm 1. If Assumptions 5−8 hold, then the following inequality holds for some η, α, J, L,*

$$v_{\mathrm{MF}}^*(\boldsymbol{\mu}_0) - \frac{1}{J}\sum_{j=1}^{J} v_{\mathrm{MF}}(\boldsymbol{\mu}_0, \pi_{\Phi_j}) \leq \frac{\sqrt{\epsilon_{\mathrm{bias}}}}{1-\gamma} + \epsilon,$$

for arbitrary initial parameter Φ_0 and initial state distribution $\boldsymbol{\mu}_0 \in \mathcal{P}(\mathcal{X})$. The parameter ϵ_{bias} is a constant. The sample complexity of Algorithm 1 is $\mathcal{O}(\epsilon^{-3})$.

The bias ϵ_{bias} turns out to be small for rich neural network based policies [Liu et al., 2020]. Intuitively, it indicates the expressive power of the policy class, Π.

Lemma 1 establishes that Algorithm 1 can approximate the optimal mean-field value function with an error bound of ϵ, and a sample complexity of $\mathcal{O}(\epsilon^{-3})$. Using Theorem 1, we can now state the following result.

Theorem 2. *Let $\boldsymbol{x}_0 \triangleq \{x_0^i\}_{i\in[N]}$ be the initial states in an N-agent system and $\boldsymbol{\mu}_0$ their associated empirical distribution. Assume that $\{\Phi_j\}_{j=1}^J$ are the policy parameters generated from Algorithm 1, and the set of policies, Π satisfies Assumption 3. If Assumptions 1, 2, 4, 5 - 8 are satisfied, then, any $\epsilon > 0$, the following inequality holds for certain choices of η, α, J, L*

$$\left| \sup_{\Phi\in\mathbb{R}^{\mathrm{d}}} v_{\mathrm{MARL}}(\boldsymbol{x}_0, \pi_\Phi) - \frac{1}{J}\sum_{j=1}^{J} v_{\mathrm{MF}}(\boldsymbol{\mu}_0, \pi_{\Phi_j}) \right| \leq \frac{\sqrt{\epsilon_{\mathrm{bias}}}}{1-\gamma} + C\max\{e, \epsilon\}$$
$$\text{where } e \triangleq \frac{1}{\sqrt{N}}\left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|}\right] \tag{19}$$

whenever $\gamma S_P < 1$ where S_P is given in Theorem 1. The term, C is a constant and the parameter ϵ_{bias} is defined in Lemma 1. The sample complexity of the process is $\mathcal{O}(\epsilon^{-3})$.

Proof. Note that following inequality,

$$\left|\sup_{\Phi\in\mathbb{R}^{\mathrm{d}}} v_{\mathrm{MARL}}(\boldsymbol{x}_0,\pi_\Phi) - \frac{1}{J}\sum_{j=1}^{J} v_{\mathrm{MF}}(\boldsymbol{\mu}_0,\pi_{\Phi_j})\right|$$
$$\leq \left|\sup_{\Phi\in\mathbb{R}^{\mathrm{d}}} v_{\mathrm{MARL}}(\boldsymbol{x}_0,\pi_\Phi) - v^*_{\mathrm{MF}}(\boldsymbol{\mu}_0)\right|$$
$$+ \left|v^*_{\mathrm{MF}}(\boldsymbol{\mu}_0) - \frac{1}{J}\sum_{j=1}^{J} v_{\mathrm{MF}}(\boldsymbol{\mu}_0,\pi_{\Phi_j})\right|$$

Using Theorem 1, the first term can be bounded by $C'e$ for some constant C'. The second term can be bounded by $\sqrt{\epsilon_{\mathrm{bias}}}/(1-\gamma)+\epsilon$ with a sample complexity of $\mathcal{O}(\epsilon^{-3})$ (Lemma 1). Assigning $C = 2\max\{C', 1\}$, we conclude the result. □

Theorem 2 guarantees that Algorithm 1 can yield a policy such that its associated value is $\mathcal{O}(\max\{e,\epsilon\})$ error away from the optimal value of the non-uniform MARL problem. Moreover, it also dictates such a policy can be obtained with a sample complexity of $\mathcal{O}(\epsilon^{-3})$.

6 CONCLUSION

In this article, we consider a multi-agent reinforcement learning (MARL) problem where the interaction between agents is described by a doubly stochastic matrix. We prove that, if the reward function is affine, one can well-approximate this non-uniform MARL problem via an associated Mean-Field Control (MFC) problem. We obtain an upper bound of the approximation error as a function of the number of agents, and also propose a natural policy gradient (NPG) algorithm to solve the MFC problem with polynomial sample complexity. The obvious drawback of our approach is the restriction on the structure of the reward function. Therefore, extension of our techniques to non-affine reward functions is an important future goal.

References

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

Abubakr O Al-Abbasi, Arnob Ghosh, and Vaneet Aggarwal. Deeppool: Distributed model-free algorithm for ride-sharing using deep reinforcement learning. *IEEE Transactions on Intelligent Transportation Systems*, 20 (12):4714–4727, 2019.

S Sh Alaviani and Nicola Elia. Distributed average consensus over random networks. In *2019 American Control Conference (ACC)*, pages 1854–1859. IEEE, 2019a.

Seyyed Shaho Alaviani and Nicola Elia. Distributed multiagent convex optimization over random digraphs. *IEEE Transactions on Automatic Control*, 65(3):986–998, 2019b.

Andrea Angiuli, Jean-Pierre Fouque, and Mathieu Laurière. Unified reinforcement q-learning for mean field game and control problems. *Mathematics of Control, Signals, and Systems*, pages 1–55, 2022.

Peter E Caines and Minyi Huang. Graphon mean field games and the gmfg equations: ε-nash equilibria. In *2019 IEEE 58th conference on decision and control (CDC)*, pages 286–292. IEEE, 2019.

René Carmona, François Delarue, et al. *Probabilistic Theory of Mean Field Games with Applications I-II*. Springer, 2018.

Yang Chen, Jiamou Liu, and Bakhadyr Khoussainov. Agent-level maximum entropy inverse reinforcement learning for mean field games. *arXiv preprint arXiv:2104.14654*, 2021.

Kai Cui and Heinz Koeppl. Learning graphon mean field games and approximate nash equilibria. *arXiv preprint arXiv:2112.01280*, 2021.

Haotian Gu, Xin Guo, Xiaoli Wei, and Renyuan Xu. Mean-field controls with Q-learning for cooperative MARL: convergence and complexity analysis. *SIAM Journal on Mathematics of Data Science*, 3(4):1168–1196, 2021.

Yanli Liu, Kaiqing Zhang, Tamer Basar, and Wotao Yin. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. *Advances in Neural Information Processing Systems*, 33:7624–7636, 2020.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937. PMLR, 2016.

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). *arXiv preprint arXiv:2109.04024*, 2021.

Barna Pasztor, Ilija Bogunovic, and Andreas Krause. Efficient model-based multi-agent mean-field reinforcement learning. *arXiv preprint arXiv:2107.04050*, 2021.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 4295–4304. PMLR, 2018.

Tabish Rashid, Gregory Farquhar, Bei Peng, and Shimon Whiteson. Weighted qmix: Expanding monotonic value function factorisation for deep multi-agent reinforcement learning. *Advances in neural information processing systems*, 33:10199–10210, 2020.

Gavin A Rummery and Mahesan Niranjan. *On-line Q-learning using connectionist systems*, volume 37. Citeseer, 1994.

Jayakumar Subramanian and Aditya Mahajan. Reinforcement learning in stationary mean-field games. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 251–259, 2019.

Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, et al. Value-decomposition networks for cooperative multi-agent learning. *arXiv preprint arXiv:1706.05296*, 2017.

Ming Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the tenth international conference on machine learning*, pages 330–337, 1993.

Deepanshu Vasal, Rajesh Mishra, and Sriram Vishwanath. Sequential decomposition of graphon mean field games. In *2021 American Control Conference (ACC)*, pages 730–736. IEEE, 2021.

Hoi-To Wai, Zhuoran Yang, Zhaoran Wang, and Mingyi Hong. Multi-agent reinforcement learning via double averaging primal-dual optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

Xiaoqiang Wang, Liangjun Ke, Zhimin Qiao, and Xinghua Chai. Large-scale traffic signal control using a novel multiagent reinforcement learning. *IEEE transactions on cybernetics*, 51(1):174–187, 2020.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3):279–292, 1992.

Nicholas J Watkins, Cameron Nowzari, Victor M Preciado, and George J Pappas. Optimal resource allocation for competitive spreading processes on bilayer networks. *IEEE Transactions on Control of Network Systems*, 5(1): 298–307, 2016.

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 2. *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 3. *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 4. *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 6, 7, 8 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 5 as the key ingredient.

Lemma 5. *[Mondal et al., 2021] 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.*

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

The proofs of Lemma 6, 7, and 8 have been delegated to Appendix F, G, and H respectively.

Lemma 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.*

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

where ν^{MF} is defined in (8).

Lemma 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. 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 8. *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,

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

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 8. 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 4. 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 7 and Eq. (9) while (b) is a result of Lemma 3. 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 2

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 3

Note the following inequalities.

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

where the term J_1 is as follows.

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

Inequality (a) follows from Assumption 1 whereas (b) uses Lemma 2 and the fact that $\boldsymbol{\mu}_1$, $\pi(x,\boldsymbol{\mu}_1)$ are distributions. The term J_2 is given as follows.

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

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 4

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.

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

Inequality (a) follows from Corollary 1(b) whereas (b) uses Lemma 2 and the fact that $\boldsymbol{\mu}_1$, $\pi(x,\boldsymbol{\mu}_1)$ are distributions. The term J_2 is given as follows.

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

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 6

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

$$
\begin{aligned}
&\mathbb{E}|\boldsymbol{\nu}_t^N - \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^N,\pi_t)|_1 \\
&= \sum_{u\in\mathcal{U}} \mathbb{E}|\boldsymbol{\nu}_t^N(u) - \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^N,\pi_t)(u)| \\
&= \sum_{u\in\mathcal{U}} \mathbb{E}\left|\frac{1}{N}\sum_{i=1}^N \delta(u_t^i = u) - \sum_{x\in\mathcal{X}} \pi_t(x,\boldsymbol{\mu}_t^N)(u)\boldsymbol{\mu}_t^N(x)\right|
\end{aligned} \tag{20}
$$

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

$$\begin{aligned}
&\sum_{x\in\mathcal{X}} \pi_t(x,\boldsymbol{\mu}_t^N)(u)\boldsymbol{\mu}_t^N(x)\\
&= \sum_{x\in\mathcal{X}} \pi_t(x,\boldsymbol{\mu}_t^N)(u)\frac{1}{N}\sum_{i=1}\delta(x_t^i=x)\\
&= \frac{1}{N}\sum_{i=1}^{N}\sum_{x\in\mathcal{X}} \pi_t(x,\boldsymbol{\mu}_t^N)(u)\delta(x_t^j=x)\\
&= \frac{1}{N}\sum_{i=1}^{N}\pi_t(x_t^j,\boldsymbol{\mu}_t^N)
\end{aligned} \tag{21}$$

Substituting into (20), we obtain the following.

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

Inequality (a) is a consequence of Lemma 5. 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.

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

G PROOF OF LEMMA 7

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

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

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 5. 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.

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

The second term J_2 can be expressed as follows.

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

Inequality (a) follows from Assumption 1 whereas (b) results from Lemma 6. 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 5. In particular, 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 8

Note that,

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

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,

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

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.

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

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

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

Using Lemma 6, the first term can be upper bounded by $|\boldsymbol{b}|_1\sqrt{|\mathcal{U}|/N}$. The second term can be bounded as follows.

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

The term $M_F > 0$ is such that $|f(x, u)| \le M_F$, $\forall x \in \mathcal{X}$, $\forall u \in \mathcal{U}$. Such M_F always exists since $\mathcal{X}$, and $\mathcal{U}$ are finite. Equality (a) is a result of Lemma 5. 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,

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

This concludes the Lemma.

I SAMPLING PROCEDURE

Algorithm 2 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).

$t \leftarrow 0$
FLAG $\leftarrow$ FALSE
while FLAG is FALSE **do**
  FLAG $\leftarrow$ TRUE with probability $1 - \gamma$.
  Execute Update
end while
$T \leftarrow t$
Accept $(x_T, \boldsymbol{\mu}_T, u_T)$ as a sample.

$\hat{V}_{\Phi_j} \leftarrow 0$, $\hat{Q}_{\Phi_j} \leftarrow 0$
FLAG $\leftarrow$ FALSE
SumRewards $\leftarrow 0$
while FLAG is FALSE **do**
  FLAG $\leftarrow$ TRUE with probability $1 - \gamma$.
  Execute Update
  SumRewards $\leftarrow$ SumRewards $+ r(x_t, u_t, \boldsymbol{\mu}_t, \boldsymbol{\nu}_t)$
end while

With probability $\frac{1}{2}$, $\hat{V}_{\Phi_j} \leftarrow$ SumRewards. Otherwise $\hat{Q}_{\Phi_j} \leftarrow$ 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