On the Approximation of Cooperative Heterogeneous Multi-Agent Reinforcement Learning (MARL) using Mean Field Control (MFC)

Washim Uddin Mondal   WMONDAL@PURDUE.EDU
Lyles School of Civil Engineering,
School of Industrial Engineering,
Purdue University,
West Lafayette, IN, 47907, USA

Mridul Agarwal   AGARW180@PURDUE.EDU
School of Electrical and Computer Engineering,
Purdue University,
West Lafayette, IN, 47907, USA

Vaneet Aggarwal   VANEET@PURDUE.EDU
School of Industrial Engineering,
School of Electrical and Computer Engineering,
Purdue University,
West Lafayette, IN, 47907, USA

Satish V. Ukkusuri   SUKKUSUR@PURDUE.EDU
Lyles School of Civil Engineering,
Purdue University,
West Lafayette, IN, 47907, USA *

Editor: Animashree Anandkumar

Abstract

Mean field control (MFC) is an effective way to mitigate the curse of dimensionality of cooperative multi-agent reinforcement learning (MARL) problems. This work considers a collection of N_{pop} heterogeneous agents that can be segregated into K classes such that the k-th class contains N_k homogeneous agents. We aim to prove approximation guarantees of the MARL problem for this heterogeneous system by its corresponding MFC problem. We consider three scenarios where the reward and transition dynamics of all agents are respectively taken to be functions of (1) joint state and action distributions across all classes, (2) individual distributions of each class, and (3) marginal distributions of the entire population. We show that, in these cases, the K-class MARL problem can be approximated by MFC with errors given as $e_1 = \mathcal{O}(\frac{\sqrt{|\mathcal{X}|}+\sqrt{|\mathcal{U}|}}{N_{\mathrm{pop}}} \sum_k \sqrt{N_k})$, $e_2 = \mathcal{O}([\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|}] \sum_k \frac{1}{\sqrt{N_k}})$ and $e_3 = \mathcal{O}\left(\left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|}\right] \left[\frac{A}{N_{\mathrm{pop}}} \sum_{k\in[K]} \sqrt{N_k} + \frac{B}{\sqrt{N_{\mathrm{pop}}}}\right] \right)$, respectively, where A, B are some constants and $|\mathcal{X}|, |\mathcal{U}|$ are the sizes of state and action spaces of each agent. Finally, we design a Natural Policy Gradient (NPG) based algorithm that, in the three cases stated above, can converge to an optimal MARL policy within $\mathcal{O}(e_j)$ error with a sample complexity of $\mathcal{O}(e_j^{-3})$, $j \in \{1, 2, 3\}$, respectively.

---

1. Introduction

The control of a large number of interacting agents is a common problem in social science and engineering with applications in finance, smart grids, transportation, wireless networks, epidemic control, etc. (Schwartz, 2014; Zhang et al., 2021). A common approach for decision making in such environments is multi-agent reinforcement learning (MARL). In *cooperative* MARL, the target is to design a sequence of *decision rules* or a *policy* that instructs the agents how to select *actions* based on their observed *state* of the environment such that the long-term *collective* reward is maximized. The joint state and action spaces of the agents, however, increase exponentially with the size of the population. This makes the computation of reward maximizing policy an incredibly challenging pursuit, especially when the number of agents is large.

To overcome the exponential blow-up of joint state and action spaces in collaborative MARL, several computationally efficient approaches have been proposed, including Independent Q-learning (IQL) (Tan, 1993), centralized training with decentralized execution (CTDE) (Rashid et al., 2020; Sunehag et al., 2018; Son et al., 2019; Rashid et al., 2018), and mean-field control (MFC) (Angiuli et al., 2020). IQL forces the environment to be non-stationary and thus its global convergence cannot be shown in general (Zhu et al., 2019). Global convergence for CTDE-type algorithms is also not known. On the other hand, the core idea of MFC is that, if the population size is infinite and the agents are *homogeneous*, then one can draw accurate inferences about the population by studying only one representative agent (Bensoussan et al., 2018). The assumption of homogeneity, however, does not go hand-in-hand with many scenarios of practical interest. For example, ride-hailing services typically offer multiple types of vehicles and drivers, each with different accommodation capacity, driving behavior, searching behavior and preferred travel range. If the profit earned per unit time is considered as reward, then each type of vehicle/driver will possess a distinct reward function and thus the system as a whole cannot be homogeneous.

It is evident from the above discussion that there are no scalable approaches in the literature to solve the problem of heterogeneous MARL with global convergence guarantees. The goal of our paper is to bridge this gap. In particular, we consider a population of N_{pop} heterogeneous agents that can be partitioned into K classes such that k-th class consists of N_k homogeneous agents. In other words, the agents in each class are assumed to have identical reward function and state transition dynamics. However, those functions are different in different classes. In this framework, we prove that MARL can be approximated as a K-class MFC problem and obtain the approximation error as a function of different class sizes. We further develop an algorithm to solve the K-class MFC problem and with the help of our approximation result, show that it efficiently converges to a provably near-optimal policy of heterogeneous MARL.

K-class MFC can be depicted as a generalization of traditional MFC-based approach which as stated before, assumes all agents to be identical. Homogeneity enforces the impact of the population on any agent to be summarized by the state and action distributions of the entire population. In contrast, K-class MFC does not allow such simplification.

The agents in such a case, not only influence other agents from the same class but their influence extends to agents from other classes as well. Due to the inter-class interaction, the influence of the whole population must be summarized either via joint state and action distributions over all classes or via the collection of distributions of each individual classes. The analysis of a K-class MFC, as a result, turns out to be very different from that of a single class/traditional MFC.

1.1 Key Contributions:

We analyse the above heterogeneous system under two generic setups. In the first case, the reward and transition functions of all agents are assumed to be functions of joint state and action distributions across all classes while in the second scenario, those are taken to be functions of state and action distributions of each individual classes. We prove that, in the first case, the N_{pop}-agent RL problem can be approximated by the K-class MFC problem within an error of $e_1 = \mathcal{O}(\left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|}\right] \frac{1}{N_{\text{pop}}} \sum_{k \in [K]} \sqrt{N_k})$ where N_k is the population size of k-th class, $k \in \{1, \cdots, K\} \triangleq [K]$ and $|\mathcal{X}|, |\mathcal{U}|$ denote the size of state and action spaces of individual agents, respectively. In the second case, the approximation error is proven to be $e_2 = \mathcal{O}(\left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|}\right] \sum_{k \in [K]} \frac{1}{\sqrt{N_k}})$.

For single class of agents, the approximation error reduces to $\mathcal{O}\left(\left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|}\right] \frac{1}{\sqrt{N_{\text{pop}}}}\right)$ which matches a recent result of (Gu et al., 2020). It is worthwhile to point out that our proof methods are distinct from that used in (Gu et al., 2020). In particular, at the heart of our approximation, lies a novel inequality on independent random variables bounded in $[0, 1]$ with constrained parameters (Lemma 11 of Appendix A). This, in conjugation with two important observations about state and action evolution of the agents, establishes our preliminary results. In contrast, (Gu et al., 2020) utilises a well-known property of sub-Gaussian variables. Although for $K = 1$, both our bound and that suggested in (Gu et al., 2020) are of the same order, our bounds possess smaller leading constant terms[1].

We also consider a special case where the reward and transition dynamics are functions of aggregate state and action distributions of the entire population. In this case, the approximation error reduces to $e_3 = \mathcal{O}\left(\left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|}\right] \left[\frac{A}{N_{\text{pop}}} \sum_{k \in [K]} \sqrt{N_k} + \frac{B}{\sqrt{N_{\text{pop}}}}\right]\right)$ where A, B are some constants.

Finally, extending the approach in (Liu et al., 2020), we develop a natural policy-gradient (NPG) based algorithm for MFC, which, combined with the approximation results between MARL and MFC, shows that the proposed NPG algorithm converges to the optimal MARL policy within $\mathcal{O}(e_j)$ error with a sample complexity of $\mathcal{O}(e_j^{-3})$, $j \in \{1, 2, 3\}$ for the three cases, respectively.

---

1. We note that the authors of (Gu et al., 2020) had an incorrect result when we first posted our version on arXiv in Sept 2021 (https://arxiv.org/pdf/2109.04024.pdf), and the error was detailed in our arXiv version. The authors of (Gu et al., 2020) fixed the error in the final version, acknowledging our manuscript.

2. Related Work

Approaches for RL: Tabular algorithms such as Q-learning (Watkins and Dayan, 1992) and SARSA (Rummery and Niranjan, 1994) were the earliest approaches to solve RL problems. However, they are not suitable for large state-action space due to their huge memory requirement. Recently, Deep Q-network (DQN) (Mnih et al., 2015) and policy-gradient based algorithms (Mnih et al., 2016) have shown promising results in terms of scalability. Although these algorithms can handle large state-space due to neural network (NN) based architecture, the approach is not scalable to multiple agents. Further, the guarantees of these algorithms either require the underlying Markov Decision Processes to be linear (Jin et al., 2020), of low Bellman rank (Jiang et al., 2017), or the scaling of parameters of NNs to be increasing with time (Wang et al., 2019) - all of which are restrictive assumptions and may not hold for general MARL.

   Use of MFC for MARL problems: MFC has found its application in various MARL setups. For example, it has been used in traffic signal control (Wang et al., 2020), management of power grids (Chen et al., 2016), ride-sharing (Al-Abbasi et al., 2019), and epidemic control (Watkins et al., 2016), among others.

   Learning Algorithms for MFC: To solve homogeneous MFC problems, several learning algorithms have been proposed. For example, model-free Q-learning algorithms have been suggested in (Angiuli et al., 2020; Gu et al., 2020; Carmona et al., 2019b) while (Carmona et al., 2019a) designed a policy-gradient based method. Recently, (Pasztor et al., 2021) proposed a model-based algorithm for MFC. All of these works are appropriate only for homogeneous MFC.

   Theoretical Relation between MARL and MFC: It is well known that when the number of agents approaches infinity, the limiting behaviour of homogeneous MARL is described by MFC (Lacker, 2017). However, it was proven only recently (Gu et al., 2020) that for a finite N_{pop} number of agents, MARL is approximated by MFC within $\mathcal{O}(1/\sqrt{N_{\text{pop}}})$ error margin. Our work is the first to provide such approximation bound for the heterogeneous MARL.

   Mean Field Games: Alongside MFC, mean field games (MFG) has garnered attention in the mean-field community. MFG analyses an infinite population of *non-cooperative* homogeneous agents. The target is to identify the Nash equilibrium (NE) of the game and design learning algorithms that converge to such an equilibrium (Guo et al., 2019; Elie et al., 2020; Yang et al., 2018; Agarwal et al., 2022).

3. Model for Heterogeneous Cooperative MARL

We consider K classes of agents where the agents belonging to each class are identical and interchangeable. The population size of k-th class, where $k \in \{1, \cdots, K\} \triangleq [K]$ is N_k, while the total population size is $N_{\text{pop}} \triangleq \sum_{k \in [K]} N_k$. Also, $\mathbf{N} \triangleq \{N_k\}_{k \in [K]}$. Let $\mathcal{X}, \mathcal{U}$ be (finite) state and action spaces of each agent. At time $t \in \{0, 1, \cdots\}$, j-th agent belonging to k-th class possesses a state $x_{j,k}^{t,\mathbf{N}} \in \mathcal{X}$ and takes an action $u_{j,k}^{t,\mathbf{N}} \in \mathcal{U}$. As a consequence, it receives a reward $r_{j,k}^{t,\mathbf{N}}$ and its state changes to $x_{j,k}^{t+1,\mathbf{N}}$ following some transition probability law. In general $r_{j,k}^{t,\mathbf{N}}$ is a function of $(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}})$, i.e, the state and action of the concerned agent at time t, as well as the joint states and actions of all the agents at time t which are denoted

by $\mathbf{x}_t^{\mathbf{N}}$ and $\mathbf{u}_t^{\mathbf{N}}$, respectively. Mathematically,

$$r_{j,k}^{t,\mathbf{N}} = \tilde{r}_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \mathbf{x}_t^{\mathbf{N}}, \mathbf{u}_t^{\mathbf{N}}) \tag{1}$$

Note that the function $\tilde{r}_k(\cdot, \cdot, \cdot, \cdot)$ is identical for all agents of k-th class. This is due to the fact that the agents of a certain class are homogeneous. Recall that the agents belonging to a given class are interchangeable as well. Thus if $\boldsymbol{\mu}_t^{\mathbf{N}}$, $\boldsymbol{\nu}_t^{\mathbf{N}}$ are empirical joint distributions of states and actions of all agents at time t, i.e, $\forall x \in \mathcal{X}$, $\forall u \in \mathcal{U}$, $\forall k \in [K]$,

$$\boldsymbol{\mu}_t^{\mathbf{N}}(x, k) \triangleq \frac{1}{N_{\text{pop}}} \sum_{j=1}^{N_k} \delta(x_{j,k}^{t,\mathbf{N}} = x), \tag{2}$$

$$\boldsymbol{\nu}_t^{\mathbf{N}}(u, k) \triangleq \frac{1}{N_{\text{pop}}} \sum_{j=1}^{N_k} \delta(u_{j,k}^{t,\mathbf{N}} = u) \tag{3}$$

where $\delta(\cdot)$ is an indicator function, then, for some function r_k, we can rewrite (1) as

$$r_{j,k}^{t,\mathbf{N}} = r_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\nu}_t^{\mathbf{N}}, N_{\text{pop}}) \tag{4}$$

Note that the output of r_k, in general, is dependent on the total number of agents, N_{pop}. Moreover, if, for an arbitrary set $\mathcal{A}$, the collection of all distributions over $\mathcal{A}$ is denoted as $\mathcal{P}(\mathcal{A})$, then $\boldsymbol{\mu}_t^{\mathbf{N}} \in \mathcal{P}(\mathcal{X} \times [K])$, and $\boldsymbol{\nu}_t^{\mathbf{N}} \in \mathcal{P}(\mathcal{U} \times [K])$.

We shall now show that (1) can be also written in an alternate form. Let, $\bar{\boldsymbol{\mu}}_t^{\mathbf{N}}$, $\bar{\boldsymbol{\nu}}_t^{\mathbf{N}}$ be such that $\bar{\boldsymbol{\mu}}_t^{\mathbf{N}}(., k)$ and $\bar{\boldsymbol{\nu}}_t^{\mathbf{N}}(., k)$ are state and action distributions of the agents of k-th class, i.e., $\bar{\boldsymbol{\mu}}_t^{\mathbf{N}} \in \mathcal{P}^K(\mathcal{X}) \triangleq \mathcal{P}(\mathcal{X}) \times \cdots \times \mathcal{P}(\mathcal{X})$, $\bar{\boldsymbol{\nu}}_t^{\mathbf{N}} \in \mathcal{P}^K(\mathcal{U})$, and $\forall x \in \mathcal{X}$, $\forall u \in \mathcal{U}$, $\forall k \in [K]$

$$\bar{\boldsymbol{\mu}}_t^{\mathbf{N}}(x, k) \triangleq \frac{1}{N_k} \sum_{j=1}^{N_k} \delta(x_{j,k}^{t,\mathbf{N}} = x), \tag{5}$$

$$\bar{\boldsymbol{\nu}}_t^{\mathbf{N}}(u, k) \triangleq \frac{1}{N_k} \sum_{j=1}^{N_k} \delta(u_{j,k}^{t,\mathbf{N}} = u) \tag{6}$$

With this notation, for some $\bar{r}_k$, we can rewrite (1) as

$$r_{j,k}^{t,\mathbf{N}} = \bar{r}_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\boldsymbol{\nu}}_t^{\mathbf{N}}, \mathbf{N}) \tag{7}$$

Note that the output of $\bar{r}_k$ is, in general, dependent on $\mathbf{N}$, i.e., the population size of each of the classes. Similar to (1), the state transition law in general can be written as

$$x_{j,k}^{t+1,\mathbf{N}} \sim \tilde{P}_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \mathbf{x}_t^{\mathbf{N}}, \mathbf{u}_t^{\mathbf{N}}), \tag{8}$$

for some function $\tilde{P}_k$. Using the same argument as used in (4) and (7), we can express (8) in the following two equivalent forms for some functions P_k and $\bar{P}_k$.

$$x_{j,k}^{t+1,\mathbf{N}} \sim P_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\nu}_t^{\mathbf{N}}, N_{\text{pop}}), \tag{9}$$

$$x_{j,k}^{t+1,\mathbf{N}} \sim \bar{P}_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\boldsymbol{\nu}}_t^{\mathbf{N}}, \mathbf{N}) \tag{10}$$

To proceed with the analysis, we need to assume one of the following assumptions to be true.

Assumption 1 *(a)* $\forall k \in [K]$, *the outputs of* r_k, P_k *are independent of the last argument* N_{pop}. *To simplify notations,* N_{pop} *can be dropped as argument from both the functions.*

$$(b) |r_k(x, u, \boldsymbol{\mu}_1, \boldsymbol{\nu}_1)| \leq M_R$$
$$(c) |r_k(x, u, \boldsymbol{\mu}_1, \boldsymbol{\nu}_1) - r_k(x, u, \boldsymbol{\mu}_2, \boldsymbol{\nu}_2)| \leq L_R \left[|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2|_1 + |\boldsymbol{\nu}_1 - \boldsymbol{\nu}_2|_1 \right]$$
$$(d) |P_k(x, u, \boldsymbol{\mu}_1, \boldsymbol{\nu}_1) - P_k(x, u, \boldsymbol{\mu}_2, \boldsymbol{\nu}_2)|_1 \leq L_P \left[|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2|_1 + |\boldsymbol{\nu}_1 - \boldsymbol{\nu}_2|_1 \right]$$

$\forall x \in \mathcal{X}, \forall u \in \mathcal{U}, \forall \boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathcal{P}(\mathcal{X} \times [K]), \forall \boldsymbol{\nu}_1, \boldsymbol{\nu}_2 \in \mathcal{P}(\mathcal{U} \times [K])$, $\forall k \in [K]$. *The terms* M_R, L_R, L_P *denote some positive constants. The function* $|.|_1$ *indicates* L_1-*norm.*

Assumption 2 *(a)* $\forall k \in [K]$, *the outputs of* $\bar{r}_k, \bar{P}_k$ *are independent of the last argument* $\mathbf{N}$. *For simplifying notations,* $\mathbf{N}$ *can be dropped as argument from both the functions.*

$$(b) |\bar{r}_k(x, u, \bar{\boldsymbol{\mu}}_1, \bar{\boldsymbol{\nu}}_1)| \leq \bar{M}_R$$
$$(c) |\bar{r}_k(x, u, \bar{\boldsymbol{\mu}}_1, \bar{\boldsymbol{\nu}}_1) - \bar{r}_k(x, u, \bar{\boldsymbol{\mu}}_2, \bar{\boldsymbol{\nu}}_2)| \leq \bar{L}_R \left[|\bar{\boldsymbol{\mu}}_1 - \bar{\boldsymbol{\mu}}_2|_1 + |\bar{\boldsymbol{\nu}}_1 - \bar{\boldsymbol{\nu}}_2|_1 \right]$$
$$(d) |\bar{P}_k(x, u, \bar{\boldsymbol{\mu}}_1, \bar{\boldsymbol{\nu}}_1) - \bar{P}_k(x, u, \bar{\boldsymbol{\mu}}_2, \bar{\boldsymbol{\nu}}_2)|_1 \leq \bar{L}_P \left[|\bar{\boldsymbol{\mu}}_1 - \bar{\boldsymbol{\mu}}_2|_1 + |\bar{\boldsymbol{\nu}}_1 - \bar{\boldsymbol{\nu}}_2|_1 \right]$$

$\forall x \in \mathcal{X}, \forall u \in \mathcal{U}, \forall \bar{\boldsymbol{\mu}}_1, \bar{\boldsymbol{\mu}}_2 \in \mathcal{P}^K(\mathcal{X}), \forall \bar{\boldsymbol{\nu}}_1, \bar{\boldsymbol{\nu}}_2 \in \mathcal{P}^K(\mathcal{U})$, $\forall k \in [K]$. *The terms* $\bar{M}_R, \bar{L}_R$ *and* $\bar{L}_P$ *are constants.*

Assumptions $1(a)$, $2(a)$ state that the influence of the population on individual agents is summarized by the state and action distributions only and it does not vary with the scale of the population. In particular, Assumption $1(a)$ dictates that such influence is conveyed through joint state and action distributions across all classes which makes the reward and transition functions invariant to N_{pop}. In contrast, Assumption $2(a)$ presumes that the joint influence of the whole population can be segregated based on the class it originated from. This makes the reward and transition law invariant to the population size of each individual class. For single class of agents (i.e., when $K = 1$), both assumptions are identical. Scale invariance is one of the fundamental assumptions in the mean-field literature (Carmona and Delarue, 2018; Gu et al., 2020; Angiuli et al., 2020).

Assumptions $1(b)$, $2(b)$ state that the reward functions are bounded while Assumptions $1(c)$, $2(c)$ and $1(d)$, $2(d)$ dictate that the reward functions and the transition probabilities are Lipschitz continuous w. r. t. their respective state and action distribution arguments. These assumptions are common in the literature (Carmona and Delarue, 2018; Gu et al., 2020; Angiuli et al., 2020).

It is worthwhile to mention that for given r_k's and P_k's satisfying Assumption 1, one can define equivalent $\bar{r}_k$'s and $\bar{P}_k$'s that satisfy Assumption 2 and vice versa. For example, in appendix P, we exhibit that if r_k's and P_k's satisfy Assumption 1 with Lipschitz constants L_R, L_P respectively, then we can define equivalent $\bar{r}_k$'s and $\bar{P}_k$'s that satisfy Assumption 2 with constants $L_P \boldsymbol{\theta}_M^{-1}, L_Q \boldsymbol{\theta}_M^{-1}$ respectively where $\boldsymbol{\theta}_M^{-1} \triangleq \max_{k \in [K]} \{N_{\text{pop}}/N_k\}$. Note that the modified 'constants' are dependent on the population sizes of different classes. Therefore, if we have an approximation bound for Assumption 2, by injecting the values of the modified constants into the expression of that bound, we can obtain a bound for Assumption 1. In appendix P, however, we demonstrate that such translated bounds are, in general, loose. This is primarily because, in the derivation of the bound for Assumption 2, the Lipschitz constants are not treated as functions of the population sizes. Therefore, it cannot account

for any stringent inequality that might be applicable due to the special structure of the translated functions. We can similarly argue why a translation from Assumption 2 to Assumption 1 may not produce a tight result. In summary, although the approximation result derived for one of the above assumptions can be cast, with slight modifications, as an approximation result for the other assumption, in general, such translated results are loose. To derive tighter bounds, it is therefore necessary to produce analysis for each of these assumptions separately. We shall establish our approximation result first with Assumption 1 and then with Assumption 2.

4. Policy, Value Function and Mean Field Limit under Assumption 1

4.1 Policy and Value Function

Recall that the distributions $\boldsymbol{\mu}_t^{\mathbf{N}}$ and $\boldsymbol{\nu}_t^{\mathbf{N}}$ defined by $(2), (3)$ are elements of $\mathcal{P}(\mathcal{X} \times [K])$ and $\mathcal{P}(\mathcal{U} \times [K])$ respectively. Therefore, presuming Assumption 1 to be true, the reward function r_k for k-th class of agents can be described as a map of the following form, $r_k : \mathcal{X} \times \mathcal{U} \times \mathcal{P}(\mathcal{X} \times [K]) \times \mathcal{P}(\mathcal{U} \times [K]) \to \mathbb{R}$. Similarly, the transition law P_k can be described as, $P_k : \mathcal{X} \times \mathcal{U} \times \mathcal{P}(\mathcal{X} \times [K]) \times \mathcal{P}(\mathcal{U} \times [K]) \to \mathcal{P}(\mathcal{X})$.

A time-dependent decision rule π_k^t for k-th class of agents is a map, $\pi_k^t : \mathcal{X} \times \mathcal{P}(\mathcal{X} \times [K]) \to \mathcal{P}(\mathcal{U})$. In simple words, a decision rule π_k^t states with what probability a certain action $u \in \mathcal{U}$ should be selected by any agent of k-th class at time t, given its own state and the state distribution across all classes at time t. A policy $\boldsymbol{\pi} \triangleq \{(\pi_k^t)_{k \in [K]}\}_{t \in \{0,1,\cdots\}}$ is defined as a sequence of decision rules over all classes of agents. For a policy $\boldsymbol{\pi}$ and given initial states $\mathbf{x}_0^{\mathbf{N}}$, the infinite-horizon $\gamma \in [0,1)$-discounted value of the policy $\boldsymbol{\pi}$ for j-th agent of k-th class is defined as

$$v_{j,k}^{\mathbf{N}}(\mathbf{x}_0^{\mathbf{N}}, \boldsymbol{\pi}) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\nu}_t^{\mathbf{N}})\right], \tag{11}$$

where the expectation is taken over $u_{j,k}^{t,\mathbf{N}} \sim \pi_k^t(x_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^{\mathbf{N}})$, $x_{j,k}^{t+1,\mathbf{N}} \sim P_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\nu}_t^{\mathbf{N}})$. Also, $\boldsymbol{\mu}_t^{\mathbf{N}}$, and $\boldsymbol{\nu}_t^{\mathbf{N}}$ are obtained from $\mathbf{x}_t^{\mathbf{N}}$ and $\mathbf{u}_t^{\mathbf{N}}$ respectively. The average infinite-horizon discounted value of policy $\boldsymbol{\pi}$ is defined as

$$v^{\mathbf{N}}(\mathbf{x}_0^{\mathbf{N}}, \boldsymbol{\pi}) \triangleq \frac{1}{N_{\text{pop}}} \sum_{k \in [K]} \sum_{j=1}^{N_k} v_{j,k}^{\mathbf{N}}(\mathbf{x}_0^{\mathbf{N}}, \boldsymbol{\pi}) \tag{12}$$

In the next subsection, we discuss how to compute the mean-field limit of the empirical value function $v^{\mathbf{N}}$. The following two observations will be useful in many of our forthcoming analyses.

Observation 1 $\{u_{j,k}^{t,\mathbf{N}}\}_{j \in [N_k], k \in [K]}$ *are independent conditioned on* $\mathbf{x}_t^{\mathbf{N}}$, $\forall t \in \{0, 1, \cdots\}$. *Specifically, for a given policy* $\boldsymbol{\pi}$, *and* $\forall j \in [N_k], \forall j' \in [N_{k'}]$, $\forall k, k' \in [K]$,

$$\mathbb{P}(u_{j,k}^{t,\mathbf{N}}, u_{j',k'}^{t,\mathbf{N}} | \mathbf{x}_t^{\mathbf{N}}) = \mathbb{P}(u_{j,k}^{t,\mathbf{N}} | \mathbf{x}_t^{\mathbf{N}}) \mathbb{P}(u_{j',k'}^{t,\mathbf{N}} | \mathbf{x}_t^{\mathbf{N}})$$

Observation 2 $\{x_{j,k}^{t+1,\mathbf{N}}\}_{j \in [N_k], k \in [K]}$ *are independent conditioned on* $\mathbf{x}_t^{\mathbf{N}}, \mathbf{u}_t^{\mathbf{N}}$, $\forall t \in \{0, 1, \cdots\}$.

4.2 Mean Field Limit for K Classes

In the mean-field limit, i.e., when $N_k \to \infty$, $\forall k \in [K]$, it is enough to consider a representative for each of the classes. The state and action of the representative of k-th class at time t are indicated as $x_k^t \in \mathcal{X}$ and $u_k^t \in \mathcal{U}$ respectively. The joint distribution of states and actions of all classes of agents are symbolized as $\boldsymbol{\mu}_t \in \mathcal{P}(\mathcal{X} \times [K])$ and $\boldsymbol{\nu}_t \in \mathcal{P}(\mathcal{U} \times [K])$. If Assumption 1 holds, then the reward and the transition probability law of the representative of k-th class at time t can be expressed as, $r_k(x_k^t, u_k^t, \boldsymbol{\mu}_t, \boldsymbol{\nu}_t)$ and $P_k(x_k^t, u_k^t, \boldsymbol{\mu}_t, \boldsymbol{\nu}_t)$ respectively. For a given policy, $\boldsymbol{\pi} \triangleq \{\boldsymbol{\pi}_t\}_{t \in \{0,1,\dots\}}$, $\boldsymbol{\pi}_t \triangleq \{(\pi_k^t)_{k \in [K]}\}$, where $\{\pi_k^t\}_{t \in \{0,1,\dots\}}$ is a sequence of decision rules for k-th class, the action distribution at time t can be obtained as follows.

$$\boldsymbol{\nu}_t = \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t, \boldsymbol{\pi}_t) \triangleq \{\nu_k^{\mathrm{MF}}(\boldsymbol{\mu}_t, \boldsymbol{\pi}_t)\}_{k \in [K]},$$
$$\nu_k^{\mathrm{MF}}(\boldsymbol{\mu}_t, \boldsymbol{\pi}_t) \triangleq \sum_{x \in \mathcal{X}} \pi_k^t(x, \boldsymbol{\mu}_t) \boldsymbol{\mu}_t(x, k) \tag{13}$$

Using the definition of ν^{MF}, the evolution of the state distribution can be written as

$$\boldsymbol{\mu}_{t+1} = P^{\mathrm{MF}}(\boldsymbol{\mu}_t, \boldsymbol{\pi}_t) \triangleq \{P_k^{\mathrm{MF}}(\boldsymbol{\mu}_t, \boldsymbol{\pi}_t)\}_{k \in [K]},$$
$$P_k^{\mathrm{MF}}(\boldsymbol{\mu}_t, \boldsymbol{\pi}_t) \triangleq \sum_{x \in \mathcal{X}} \sum_{u \in \mathcal{U}} \boldsymbol{\mu}_t(x, k) \pi_k^t(x, \boldsymbol{\mu}_t)(u) \times P_k\left(x, u, \boldsymbol{\mu}_t, \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t, \boldsymbol{\pi}_t)\right) \tag{14}$$

Finally, the average reward of k-th class is computed as

$$r_k^{\mathrm{MF}}(\boldsymbol{\mu}_t, \boldsymbol{\pi}_t) \triangleq \sum_{x \in \mathcal{X}} \sum_{u \in \mathcal{U}} \boldsymbol{\mu}_t(x, k) \pi_k^t(x, \boldsymbol{\mu}_t)(u) \times r_k(x, u, \boldsymbol{\mu}_t, \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t, \boldsymbol{\pi}_t)) \tag{15}$$

For a given initial state distribution $\boldsymbol{\mu}_0$, and a policy $\boldsymbol{\pi}$, the infinite-horizon γ-discounted average reward is

$$v^{\mathrm{MF}}(\boldsymbol{\mu}_0, \boldsymbol{\pi}) = \sum_{k \in [K]} \sum_{t=0}^{\infty} \gamma^t r_k^{\mathrm{MF}}(\boldsymbol{\mu}_t, \boldsymbol{\pi}_t) \tag{16}$$

In the following section, we show how well the function $v^{\mathbf{N}}$, given by (12) can be approximated by v^{MF} as the population sizes, $\mathbf{N}$ and the cardinality of state and action spaces, indicated by $|\mathcal{X}|$ and $|\mathcal{U}|$ respectively, become large.

5. MFC as an Approximation of MARL with Assumption 1

To establish the approximation result, we need to restrict the policies to a set Π such that the following assumption holds.

Assumption 3 *Every policy $\boldsymbol{\pi} \triangleq \{(\pi_k^t)_{k \in [K]}\}_{t \in \{0,1,\dots\}}$ in Π is such that, $\forall x \in \mathcal{X}, \forall \boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathcal{P}(\mathcal{X} \times [K]), \forall k \in [K]$*

$$|\pi_k^t(x, \boldsymbol{\mu}_1) - \pi_k^t(x, \boldsymbol{\mu}_2)|_1 \le L_Q |\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2|_1$$

for some positive real L_Q.

Assumption 3 states that the decision rules π_k^t, associated with any policy in Π are Lipschitz continuous w. r. t. the state distribution argument. Such assumption holds in practice because the decision rules are commonly realised by Neural Networks possessing bounded weights (Pasztor et al., 2021). Below we state our first result.

Theorem 1 *Let $\mathbf{x}_0^{\mathbf{N}}$ be the initial states and $\boldsymbol{\mu}_0$ be their corresponding distribution. If $v^{\mathbf{N}}$ is the empirical value function given by (12) and v^{MF} is its mean-field limit defined in (16), then for any policy, $\boldsymbol{\pi} \in \Pi$, the following inequality holds if $\gamma S_P < 1$ and Assumptions 1, 3 are true.*

$$
\left| v^{\mathbf{N}}(\mathbf{x}_0^{\mathbf{N}}, \boldsymbol{\pi}) - v^{\mathrm{MF}}(\boldsymbol{\mu}_0, \boldsymbol{\pi}) \right| \leq \frac{C_R}{1-\gamma} \sqrt{|\mathcal{U}|} \frac{1}{N_{\mathrm{pop}}} \left(\sum_{k \in [K]} \sqrt{N_k} \right)
$$

$$
+ C_P \left(\frac{S_R}{S_P - 1} \right) \left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|} \right] \frac{1}{N_{\mathrm{pop}}} \left(\sum_{k \in [K]} \sqrt{N_k} \right) \times \left[\frac{1}{1-\gamma S_P} - \frac{1}{1-\gamma} \right]
$$

(17)

where $S_R \triangleq M_R(1 + L_Q) + L_R(2 + L_Q)$, $S_P \triangleq (1 + L_Q) + L_P(2 + L_Q)$, $C_R \triangleq M_R + L_R$, $C_P \triangleq 2 + L_P$.

Theorem 1 dictates that the empirical value function, $v^{\mathbf{N}}$, can be approximated by its mean-field limit, v^{MF}, within an error margin of $\mathcal{O}\left(\frac{1}{N_{\mathrm{pop}}} \sum_{k \in [K]} \sqrt{N_k} \right)$. In a special case, where the number of agents in each classes are equal, the error is $\mathcal{O}(\sqrt{K/N_{\mathrm{pop}}})$. Additionally, Theorem 1 also dictates how the error varies as a function of the state-action cardinality. For example, given other things as constant, the error is $\mathcal{O}\left(\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|} \right)$.

The implication of this result is profound. It essentially assures that, if we can come up with an algorithm to compute the optimal MFC policy, then the obtained policy is guaranteed to be close to the optimal MARL policy. In practice, an MFC problem is much easier to solve than a MARL problem, primarily because in MFC, we are needed to keep track of only one representative agent from each class. Therefore, if the number of agents is large and individual state-action spaces are relatively small, MFC can be utilized as an easier route to obtain an approximate MARL solution. However, Theorem 1 also suggests that the error of approximation increases with the number of classes, K. As a consequence, if the level of heterogeneity in the population is too high, then MFC may not be a good approximation of MARL.

5.1 Proof Outline

A detailed proof of Theorem 1 is provided in Appendix A. Here we present a brief outline.

Step 0: $v^{\mathbf{N}}$ and v^{MF} respectively are time-discounted average rewards for a finite agent system and that of an infinite agent system. To estimate their difference, we need to evaluate the difference between mean rewards of these systems at a given time t.

Step 1: To achieve that, we introduce an intermediate system X whose state-action evolutions are identical to the $\mathbf{N}$-agent system upto time t, but after that, it follows the

update process of an infinite agent system. Our first task is to bound the difference between the average reward of system X and that of the **N**-agent system at time t (Lemma 13).

Step 2: The next task is to estimate the difference between the average reward of system X and the mean-field reward at time t. Using the continuity of mean-field reward function (Lemma 9), this difference can be bounded by a multiple of the difference between the empirical state distribution, $\boldsymbol{\mu}_t^{\mathbf{N}}$ of the **N**-agent system and the mean-field distribution, $\boldsymbol{\mu}_t$.

Step 3: The difference between $\boldsymbol{\mu}_t^{\mathbf{N}}$ and $\boldsymbol{\mu}_t$ can be obtained in a recursive manner. To achieve this, we introduce another intermediate system Y whose state, action distributions upto time $t - 1$ are same as the **N**-agent system, but after that, those evolve following mean-field updates. First, we evaluate the difference between $\boldsymbol{\mu}_t^{\mathbf{N}}$ and the state distribution of the system Y at time t (Lemma 14).

Step 4: Using the continuity of mean-field state-transition function (Lemma 10), the difference between $\boldsymbol{\mu}_t$ and the state distribution of system Y at t is upper bounded by a multiple of the difference between $\boldsymbol{\mu}_{t-1}^{\mathbf{N}}$ and $\boldsymbol{\mu}_{t-1}$.

Step 5: Combining the above results, the difference between the average finite-agent reward and the mean-field reward at time t can be written as a function of t.

Step 6: Taking a γ-discounted sum of the these estimate errors over t, we arrive at the desired result.

6. MFC as an Approximation of MARL with Assumption 2

We shall now discuss how well the empirical value function is approximated by its mean-field counterpart if Assumption 2 is true. The empirical state and action distributions of k-th class at time t are denoted as $\bar{\boldsymbol{\mu}}_t^{\mathbf{N}}(\cdot, k)$, $\bar{\boldsymbol{\nu}}_t^{\mathbf{N}}(\cdot, k)$ and defined by (5), (6) respectively. Clearly, $\bar{\boldsymbol{\mu}}_t^{\boldsymbol{N}} \in \mathcal{P}^K(\mathcal{X})$ and $\bar{\boldsymbol{\nu}}_t^{\mathbf{N}} \in \mathcal{P}^K(\mathcal{U})$ where $\mathcal{P}^K(\cdot) \triangleq \mathcal{P}(\cdot) \times \cdots \times \mathcal{P}(\cdot)$.

The reward function, $\bar{r}_k$ and the transition probability law, $\bar{P}_k$ of k-th class of agents are defined to be functions of the following forms, $\bar{r}_k : \mathcal{X} \times \mathcal{U} \times \mathcal{P}^K(\mathcal{X}) \times \mathcal{P}^K(\mathcal{U}) \to \mathbb{R}$ and $\bar{P}_k : \mathcal{X} \times \mathcal{U} \times \mathcal{P}^K(\mathcal{X}) \times \mathcal{P}^K(\mathcal{U}) \to \mathcal{P}(\mathcal{X})$. Similarly, a policy $\bar{\boldsymbol{\pi}} \triangleq \{(\bar{\pi}_k^t)_{k \in [K]}\}_{t \in \{0,1,\cdots\}}$ is defined as a sequence of collection of decision rules, $\bar{\pi}_k^t$ where $\bar{\pi}_k^t : \mathcal{X} \times \mathcal{P}^K(\mathcal{X}) \to \mathcal{P}(\mathcal{U})$. Similar to Assumption 3, we restrict the policies to a set $\bar{\Pi}$ such that the decision rules associated with each elements of $\bar{\Pi}$ are Lipschitz continuous. This is formally expressed as follows.

Assumption 4 *Every policy* $\bar{\boldsymbol{\pi}} \triangleq \{(\bar{\pi}_k^t)_{k \in [K]}\}_{t \in \{0,1,\cdots\}}$ *in* $\bar{\Pi}$ *is such that,* $\forall x \in \mathcal{X}, \forall \bar{\boldsymbol{\mu}}_1, \bar{\boldsymbol{\mu}}_2 \in \mathcal{P}^K(\mathcal{X})$, $\forall k \in [K]$

$$|\bar{\pi}_k^t(x, \bar{\boldsymbol{\mu}}_1) - \bar{\pi}_k^t(x, \bar{\boldsymbol{\mu}}_2)|_1 \leq \bar{L}_Q |\bar{\boldsymbol{\mu}}_1 - \bar{\boldsymbol{\mu}}_2|_1$$

for some positive real $\bar{L}_Q$.

For initial states $\mathbf{x}_0^{\mathbf{N}}$, the empirical value of a given policy $\bar{\boldsymbol{\pi}}$ is defined as follows.

$$\bar{v}^{\mathbf{N}}(\mathbf{x}_0^{\mathbf{N}}, \bar{\boldsymbol{\pi}}) = \frac{1}{N_{\text{pop}}} \sum_{k \in [K]} \sum_{j=1}^{N_k} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \bar{r}_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\boldsymbol{\nu}}_t^{\mathbf{N}})\right] \qquad (18)$$

where the expectation is taken over $u_{j,k}^{t,\mathbf{N}} \sim \bar{\pi}_k^t(x_{j,k}^{t,\mathbf{N}}, \bar{\boldsymbol{\mu}}_t^{\mathbf{N}})$, $x_{j,k}^{t+1,\mathbf{N}} \sim \bar{P}_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\boldsymbol{\nu}}_t^{\mathbf{N}})$. Also, $\bar{\boldsymbol{\mu}}_t^{\mathbf{N}}$, $\bar{\boldsymbol{\nu}}_t^{\mathbf{N}}$ are obtained from $\mathbf{x}_t^{\mathbf{N}}$, $\mathbf{u}_t^{\mathbf{N}}$. If $\bar{v}^{\text{MF}}$ denotes the mean-field limit of $\bar{v}^{\mathbf{N}}$, then the following approximation result holds.

Theorem 2 *If $\mathbf{x}_0^{\mathbf{N}}$ are initial states and $\bar{\boldsymbol{\mu}}_0 \in \mathcal{P}^K(\mathcal{X})$ is the resulting distribution, then under Assumptions 2, 4, $\forall \bar{\boldsymbol{\pi}} \in \bar{\Pi}$,*

$$
\begin{aligned}
\left| \bar{v}^{\mathbf{N}}(\mathbf{x}_0^{\mathbf{N}}, \bar{\boldsymbol{\pi}}) - \bar{v}^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}_0, \bar{\boldsymbol{\pi}}) \right| &\leq \frac{\bar{C}_R}{1-\gamma} \sqrt{|\mathcal{U}|} \left(\sum_{k \in [K]} \frac{1}{\sqrt{N_k}} \right) \\
&+ \bar{C}_P \left(\frac{\bar{S}_R}{\bar{S}_P - 1} \right) \left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|} \right] \left(\sum_{k \in [K]} \frac{1}{\sqrt{N_k}} \right) \times \left[\frac{1}{1 - \gamma \bar{S}_P} - \frac{1}{1-\gamma} \right]
\end{aligned}
\tag{19}
$$

whenever $\gamma \bar{S}_P < 1$ where $\bar{v}^{\mathbf{N}}(\cdot, \cdot)$ denotes the empirical value function defined in (18) and $\bar{v}^{\mathrm{MF}}(\cdot, \cdot)$ is its mean-field limit. The other terms are given as follows: $\bar{C}_R \triangleq \bar{M}_R + \bar{L}_R$, $\bar{C}_P \triangleq 2 + K \bar{L}_P$, $\bar{S}_R \triangleq \bar{M}_R(1 + \bar{L}_Q) + \bar{L}_R(2 + K \bar{L}_Q)$, and $\bar{S}_P \triangleq (1 + K \bar{L}_Q) + K \bar{L}_P(2 + K \bar{L}_Q)$.

Therefore, Theorem 2 asserts that the error in approximating the value function $\bar{v}^{\mathbf{N}}$ by its mean-field limit, $\bar{v}^{\mathrm{MF}}$, is $\mathcal{O}(\left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|} \right] \sum_{k \in [K]} \frac{1}{\sqrt{N_k}})$. Note that, Theorem 1 gives a tighter bound than Theorem 2 (if Lipschitz constants are same in both the cases). This can be attributed to the fact that the difference between two joint distributions $\boldsymbol{\mu}, \boldsymbol{\mu}'$ (which is used to bound the approximation error in Theorem 1) is, in general, less than the difference between the resulting collection of distributions of all classes $\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\mu}}'$ (which is used to bound the approximation error in Theorem 2).

7. Improved Results when Transition and Reward Functions Depend on Aggregate Distributions

In this section, the transition and reward functions (and thus, the decision rules associated with the policies) are assumed to be Lipschitz continuous functions of aggregate/marginal state and action distributions of the entire population. It is easy to see that, this assumption is stronger than Assumption 1 and 2 as any Lipschitz continuous function of the marginal distributions is necessarily a Lipschitz continuous function of both the joint distributions and the collection of distributions of each classes, with the same Lipschitz parameter. We shall demonstrate that such stronger assumption leads to improved approximation result. Mathematically, if the reward and state transition functions are indicated as r_k's, and P_k's, a generic policy is denoted as $\boldsymbol{\pi} \triangleq \{(\pi_k^t)_{k \in [K]}\}_{t \in \{0, 1, \cdots\}}$, and the class of policies is defined by Π, then our assumption can be stated as follows.

Assumption 5 *(a) The reward functions, transition dynamics and the decision rules are of the following form.*

$$
\begin{aligned}
r_k &: \mathcal{X} \times \mathcal{U} \times \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{U}) \to \mathbb{R} \\
P_k &: \mathcal{X} \times \mathcal{U} \times \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{U}) \to \mathcal{P}(\mathcal{X}) \\
\pi_k^t &: \mathcal{X} \times \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{U})
\end{aligned}
$$

$\forall \boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathcal{P}(\mathcal{X} \times [K]),\ \forall \boldsymbol{\nu}_1, \boldsymbol{\nu}_2 \in \mathcal{P}(\mathcal{U} \times [K]),\ \forall x \in \mathcal{X},\ \forall u \in \mathcal{U},\ \forall k \in [K],\ \forall \boldsymbol{\pi} = \{(\pi_k^t)_{k \in [K]}\}_{t \in \{0,1,\cdots\}} \in \Pi,$

$(b) |r_k(x, u, \boldsymbol{\mu}_1[\mathcal{X}], \boldsymbol{\nu}_1[\mathcal{U}])| \leq M_R$

$(c) |r_k(x, u, \boldsymbol{\mu}_1[\mathcal{X}], \boldsymbol{\nu}_1[\mathcal{U}]) - r_k(x, u, \boldsymbol{\mu}_2[\mathcal{X}], \boldsymbol{\nu}_2[\mathcal{U}])| \leq L_R [|\boldsymbol{\mu}_1[\mathcal{X}] - \boldsymbol{\mu}_2[\mathcal{X}]|_1 + |\boldsymbol{\nu}_1[\mathcal{U}] - \boldsymbol{\nu}_2[\mathcal{U}]|_1]$

$(d) |P_k(x, u, \boldsymbol{\mu}_1[\mathcal{X}], \boldsymbol{\nu}_1[\mathcal{U}]) - P_k(x, u, \boldsymbol{\mu}_2[\mathcal{X}], \boldsymbol{\nu}_2[\mathcal{U}])|_1 \leq L_P [|\boldsymbol{\mu}_1[\mathcal{X}] - \boldsymbol{\mu}_2[\mathcal{X}]|_1 + |\boldsymbol{\nu}_1[\mathcal{U}] - \boldsymbol{\nu}_2[\mathcal{U}]|_1]$

$(e) |\pi_k^t(x, \boldsymbol{\mu}_1[\mathcal{X}]) - \pi_k^t(x, \boldsymbol{\mu}_2[\mathcal{X}])|_1 \leq L_Q |\boldsymbol{\mu}_1[\mathcal{X}] - \boldsymbol{\mu}_2[\mathcal{X}]|_1$

where $\boldsymbol{\mu}[\mathcal{X}], \boldsymbol{\nu}[\mathcal{U}]$ are marginal distributions on $\mathcal{X}, \mathcal{U}$ resulting from $\boldsymbol{\mu}, \boldsymbol{\nu}$ and M_R, L_R, L_P, L_Q are some constants.

Theorem 3 *Assume $\mathbf{x}_0^{\mathbf{N}}$ to be the initial states and $\boldsymbol{\mu}_0$ their corresponding joint distribution. If $\gamma S_P < 1$, and Assumption 5 holds, then for any arbitrary policy, $\boldsymbol{\pi} \in \Pi$,*

$$\left| v^{\mathbf{N}}(\mathbf{x}_0^{\mathbf{N}}, \boldsymbol{\pi}) - v^{\mathrm{MF}}(\boldsymbol{\mu}_0, \boldsymbol{\pi}) \right| \leq \frac{C_R}{1 - \gamma} \sqrt{|\mathcal{U}|} \frac{1}{\sqrt{N_{\mathrm{pop}}}}$$

$$+ \left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|} \right] \left(\frac{\gamma C_P}{1 - \gamma} \right) \left[\frac{S_R'}{N_{\mathrm{pop}}} \left(\sum_{k \in [K]} \sqrt{N_k} \right) + \frac{S_R''}{\sqrt{N_{\mathrm{pop}}}} \right]$$

$$+ C_P \left(\frac{S_R}{S_P - 1} \right) \left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|} \right] \left(\frac{\gamma}{1 - \gamma S_P} - \frac{\gamma}{1 - \gamma} \right) \times \left[\frac{S_P'}{N_{\mathrm{pop}}} \left(\sum_{k \in [K]} \sqrt{N_k} \right) + \frac{S_P''}{\sqrt{N_{\mathrm{pop}}}} \right]$$

$$(20)$$

where $v^{\mathbf{N}}$ denotes the empirical value function and v^{MF} is its mean-field limit. Also, $S_R' \triangleq M_R + L_R$, $S_R'' \triangleq M_R L_Q + L_R(1 + L_Q)$, $S_P' \triangleq 1 + L_P$, and $S_P'' \triangleq L_Q + L_P(1 + L_Q)$. The terms S_R, S_P, C_R, C_P are defined in Theorem 1.

Theorem 3 states that the error in approximating the empirical value function, $v^{\mathbf{N}}$, by its mean-field limit, v^{MF}, can be written as $\mathcal{O}\left(\left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|} \right] \left[\frac{A}{N_{\mathrm{pop}}} \sum_{k=1}^{K} \sqrt{N_k} + \frac{B}{\sqrt{N_{\mathrm{pop}}}} \right] \right)$ where A, B are some constants. It is easy to show that the approximation error suggested by Theorem 3 is strictly better than the errors given by Theorem 1, 2. Intuitively, if the reward and transition functions only depend on the marginal distributions, not on the joint distributions, then those functions overlook the heterogeneity of the agents and treat the whole population holistically. This leads to the $\frac{1}{\sqrt{N_{\mathrm{pop}}}}$ component of the error which matches the error of a single class system. However, the reward and transition functions (and hence, the decision rules) themselves are different for different classes. This variation enforces the other part of the error to align towards a general heterogeneous system.

8. Global Convergence of MARL using Natural Policy Gradient Algorithm

The previous sections showed that a K-class heterogeneous MARL can be approximated as a K-class MFC. This section develops a Natural Policy Gradient (NPG) based algorithm for

K-class MFC that can obtain policies with guaranteed optimality gaps for heterogeneous MARL. We limit our discussion to the category of systems that satisfy the same set of assumptions as used in Theorem 1. For other assumptions, one can replicate similar result and the processes have been briefly described in sections 8.1 and 8.2.

Let the policies in the set Π be parametrized by Φ. Without loss of generality, we can restrict the set Π to comprise of only stationary policies (Puterman, 2014). To simplify notations, we denote a stationary policy with the parameter Φ as $\boldsymbol{\pi}_\Phi \triangleq \{\pi_\Phi^k\}_{k\in[K]}$, where π_Φ^k's are stationary decision rules for each class. In a K-class MFC, we need to track only one representative agent from each class. The k-th representative takes its action u_k by observing its own state x_k and the joint distribution $\boldsymbol{\mu}$. If $\mathbf{x} \triangleq \{x_k\}_{k\in[K]}$ and $\mathbf{u} \triangleq \{u_k\}_{k\in[K]}$, then K-class MFC can effectively be described as a single agent RL problem with $(\mathbf{x}, \boldsymbol{\mu}) \in \mathcal{X}^K \times \mathcal{P}(\mathcal{X}\times[K])$ and $\mathbf{u} \in \mathcal{U}^K$ as its state and action respectively. However, such a system comes with the additional advantage that the actions u_k's are conditionally independent given $\mathbf{x}$. It will be clear from our later result (Theorem 5) that, this prevents the complexity of the problem from being an exponential function of K.

For arbitrary $\boldsymbol{\mu} \in \mathcal{P}(\mathcal{X}\times[K])$, $\mathbf{x} \in \mathcal{X}^K$, and $\mathbf{u} \in \mathcal{U}^K$, denote the Q-value and the advantage value associated with the policy $\boldsymbol{\pi}_\Phi$ as $Q_\Phi(\mathbf{x}, \boldsymbol{\mu}, \mathbf{u})$ and $A_\Phi(\mathbf{x}, \boldsymbol{\mu}, \mathbf{u})$ respectively. The precise definition of Q-function is as follows.

$$Q_\Phi(\mathbf{x}, \boldsymbol{\mu}, \mathbf{u}) = \mathbb{E}_\Phi\left[\sum_{k\in[K]}\sum_{t=0}^\infty \gamma^t r_k\left(x_k^t, u_k^t, \boldsymbol{\mu}_t, \boldsymbol{\nu}_t\right)\middle| \mathbf{x}_0 = \mathbf{x}, \boldsymbol{\mu}_0 = \boldsymbol{\mu}, \mathbf{u}_0 = \mathbf{u}\right] \qquad (21)$$

where the expectation is computed over $u_k^t \sim \pi_\Phi^k(x_k^t, \boldsymbol{\mu}_t)$, $x_k^t \sim P_k(x_k^{t-1}, u_k^{t-1}, \boldsymbol{\mu}_{t-1}, \boldsymbol{\nu}_{t-1})$, $\forall t \in \{1, 2, \cdots\}$, $\forall k \in [K]$. Moreover, $\forall t \in \{1, 2, \cdots\}$, $\boldsymbol{\mu}_t = P^{\mathrm{MF}}(\boldsymbol{\mu}_{t-1}, \boldsymbol{\pi}_\Phi)$, $\boldsymbol{\nu}_t = \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t, \boldsymbol{\pi}_\Phi)$ where $P^{\mathrm{MF}}(\cdot, \cdot)$, $\nu^{\mathrm{MF}}(\cdot, \cdot)$ are given by (14), (13) respectively. Incorporating (21), we now define the advantage function as follows

$$A_\Phi(\mathbf{x}, \boldsymbol{\mu}, \mathbf{u}) = Q_\Phi(\mathbf{x}, \boldsymbol{\mu}, \mathbf{u}) - \mathbb{E}\left[Q_\Phi(\mathbf{x}, \boldsymbol{\mu}, \mathbf{u})\right],$$

where the expectation is over $u_k \sim \pi_\Phi^k(x_k, \boldsymbol{\mu})$, $\forall k \in [K]$.

Define $v_{\mathrm{MF}}^*(\boldsymbol{\mu}_0) \triangleq \sup_{\Phi\in\mathbb{R}^{\mathrm{d}}} v^{\mathrm{MF}}(\boldsymbol{\mu}_0, \boldsymbol{\pi}_\Phi)$, $\boldsymbol{\mu}_0 \in \mathcal{P}(\mathcal{X}\times[K])$, where $v^{\mathrm{MF}}(\cdot, \cdot)$ is the mean-field value function given by (16) and $\mathbb{R}^{\mathrm{d}}$ denotes the space of Φ. Consider a sequence of parameters $\{\Phi_j\}_{j=1}^J$ that is recursively calculated by following the natural policy gradient (NPG) (Kakade, 2001; Liu et al., 2020; Agarwal et al., 2021) update as described below.

$$\Phi_{j+1} = \Phi_j + \eta\mathbf{w}_j, \mathbf{w}_j \triangleq \arg\min_{\mathbf{w}\in\mathbb{R}^{\mathrm{d}}} L_{\zeta_{\boldsymbol{\mu}_0}^{\Phi_j}}(\mathbf{w}, \Phi_j) \qquad (22)$$

where η is the learning rate. The definitions of the function $L_{\zeta_{\boldsymbol{\mu}_0}^{\Phi_j}}$ and the distribution $\zeta_{\boldsymbol{\mu}_0}^{\Phi_j}$ are provided below. Define the following function $\forall \Phi, \Phi' \in \mathbb{R}^{\mathrm{d}}$,

$$L_{\zeta_{\boldsymbol{\mu}_0}^{\Phi'}}(\mathbf{w}, \Phi) \triangleq \mathbb{E}_{(\mathbf{x}, \boldsymbol{\mu}, \mathbf{u})\sim\zeta_{\boldsymbol{\mu}_0}^{\Phi'}}\left[\left(A_\Phi(\mathbf{x}, \boldsymbol{\mu}, \mathbf{u}) - (1-\gamma)\mathbf{w}^{\mathrm{T}}\nabla_\Phi\log\prod_{k\in[K]}\pi_\Phi^k(x_k, \boldsymbol{\mu})(u_k)\right)^2\right|$$

$$(23)$$

and $\zeta_{\boldsymbol{\mu}_0}^{\Phi'}(\mathbf{x}, \boldsymbol{\mu}, \mathbf{u}) \triangleq (1-\gamma)\sum_{\tau=0}^\infty \gamma^\tau \mathbb{P}(\mathbf{x}_\tau = \mathbf{x}, \boldsymbol{\mu}_\tau = \boldsymbol{\mu}, \mathbf{u}_\tau = \mathbf{u}|\mathbf{x}_0 = \mathbf{x}, \boldsymbol{\mu}_0 = \boldsymbol{\mu}, \mathbf{u}_0 = \mathbf{u}, \boldsymbol{\pi}_{\Phi'})$$

$$(24)$$

It is evident from (22) that at each NPG update, one needs to solve a stochastic minimization problem to find the update direction. This sub-problem can be solved by another stochastic gradient descent algorithm with the update equation $\mathbf{w}_{j,l+1} = \mathbf{w}_{j,l} - \alpha\mathbf{h}_{j,l}$ (Liu et al., 2020), where α is the learning rate and the update direction $\mathbf{h}_{j,l}$ is defined as:

$$\mathbf{h}_{j,l} \triangleq \left(\mathbf{w}_{j,l}^{\mathrm{T}}\nabla_{\Phi_j}\log\prod_{k\in[K]}\pi_{\Phi_j}^k(x_k,\boldsymbol{\mu})(u_k) - \frac{1}{1-\gamma}\hat{A}_{\Phi_j}(\mathbf{x},\boldsymbol{\mu},\mathbf{u})\right)\nabla_{\Phi_j}\log\prod_{k\in[K]}\pi_{\Phi_j}^k(x_k,\boldsymbol{\mu})(u_k)$$

where $\mathbf{x} = \{x_k\}_{k\in[K]}, \mathbf{u} = \{u_k\}_{k\in[K]}$, $(\mathbf{x},\boldsymbol{\mu},\mathbf{u})$ are sampled from $\zeta_{\boldsymbol{\mu}_0}^{\Phi_j}$ and $\hat{A}_{\Phi_j}$ is an unbiased estimator of A_Φ. The details of procuring the samples and the unbiased estimate is provided in Algorithm 2 which is based on Algorithm 3 of (Agarwal et al., 2021). In Algorithm 1, we summarize the NPG-based procedure to obtain the optimal MFC policy.

Algorithm 1 Natural Policy Gradient for K-class MFC

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: $\Phi \leftarrow \Phi_0$

1: **for** $j \in \{0, 1, \cdots, J-1\}$ **do**
2:     $\mathbf{w}_{j,0} \leftarrow \mathbf{w}_0$
3:     **for** $l \in \{0, 1, \cdots, L-1\}$ **do**
4:         Sample $(\mathbf{x},\boldsymbol{\mu},\mathbf{u}) \sim \zeta_{\boldsymbol{\mu}_0}^{\Phi_j}$ and $\hat{A}_{\Phi_j}(\mathbf{x},\boldsymbol{\mu},\mathbf{u})$ using Algorithm 2
5:         Compute $\mathbf{h}_{j,l}$ using (25)
        $\mathbf{w}_{j,l+1} \leftarrow \mathbf{w}_{j,l} - \alpha\mathbf{h}_{j,l}$
6:     **end for**
7:     $\mathbf{w}_j \leftarrow \dfrac{1}{L}\sum_{l=1}^{L}\mathbf{w}_{j,l}$
8:     $\Phi_{j+1} \leftarrow \Phi_j + \eta\mathbf{w}_j$
9: **end for**
Output: $\{\Phi_1, \cdots, \Phi_J\}$: Policy parameters

Following Theorem 4.9 of (Liu et al., 2020), we can now state the global convergence result of NPG as given below. For the result to hold, the following Assumptions need to be satisfied. These assumptions are similar to Assumptions 2.1, 4.2, 4.4 respectively in (Liu et al., 2020).

Assumption 6 $\forall\Phi \in \mathbb{R}^{\mathrm{d}}$, $\forall\boldsymbol{\mu}_0 \in \mathcal{P}(\mathcal{X} \times [K])$, the matrix $F_{\boldsymbol{\mu}_0}(\Phi) - \chi I_{\mathrm{d}}$ is positive semi-definite for some $\chi > 0$ where $F_{\boldsymbol{\mu}_0}(\Phi)$ is defined as,

$$F_{\boldsymbol{\mu}_0}(\Phi) \triangleq \mathbb{E}_{(\mathbf{x},\boldsymbol{\mu},\mathbf{u})\sim\zeta_{\boldsymbol{\mu}_0}^{\Phi}}\left[\left\{\nabla_\Phi\log\prod_{k\in[K]}\pi_\Phi^k(x_k,\boldsymbol{\mu})(u_k)\right\}\left\{\nabla_\Phi\log\prod_{k\in[K]}\pi_\Phi^k(x_k,\boldsymbol{\mu})(u_k)\right\}^{\mathrm{T}}\right]$$

Assumption 7 $\forall\Phi \in \mathbb{R}^{\mathrm{d}}$, $\forall\boldsymbol{\mu} \in \mathcal{P}(\mathcal{X} \times [K])$, $\forall x_k \in \mathcal{X}$, $\forall u_k \in \mathcal{U}$, $\forall k \in [K]$,

$$\left|\nabla_\Phi\log\prod_{k\in[K]}\pi_\Phi^k(x_k,\boldsymbol{\mu})(u_k)\right|_1 \leq G$$

Algorithm 2 Algorithm to sample $(\mathbf{x}, \boldsymbol{\mu}, \mathbf{u}) \sim \zeta_{\boldsymbol{\mu}_0}^{\Phi_j}$ and $\hat{A}_{\Phi_j}(\mathbf{x}, \boldsymbol{\mu}, \mathbf{u})$

Input: $\boldsymbol{\mu}_0$: Initial joint state distribution, $\boldsymbol{\pi}_{\Phi_j} \triangleq \{\pi_{\Phi_j}^k\}_{k \in [K]}$: Policy,
$\{P_k(.,.,.,.)\}_{k \in [K]}$: Transition laws, $\{r_k(.,.,.,.)\}_{k \in [K]}$: Reward functions,
$\boldsymbol{\theta} \triangleq \{\theta_k\}_{k \in [K]}$: Prior probabilities of different classes.

1: Sample $\mathbf{x}_0 \triangleq \{x_k^0\}_{k \in [K]} \sim \boldsymbol{\mu}_0$.
2: Sample $\mathbf{u}_0 \triangleq \{u_k^0\}_{k \in [K]} \sim \boldsymbol{\pi}_{\Phi_j}(\mathbf{x}_0, \boldsymbol{\mu}_0)$ i.e., sample $u_k^0 \sim \pi_{\Phi_j}^k(x_k^0, \boldsymbol{\mu}_0), \forall k \in [K]$.
3: $\boldsymbol{\nu}_0 \leftarrow \nu^{\mathrm{MF}}(\boldsymbol{\mu}_0, \boldsymbol{\pi}_{\Phi_j})$ where ν^{MF} is defined in (13).

4: $t \leftarrow 0$
5: FLAG $\leftarrow$ FALSE
6: **while** FLAG is FALSE **do**
7:     FLAG $\leftarrow$ TRUE with probability $1 - \gamma$.
8:     Execute SystemUpdate
9: **end while**
10: $T \leftarrow t$
11: Accept $(\mathbf{x}_T, \boldsymbol{\mu}_T, \mathbf{u}_T)$ as a sample.

12: $\hat{V}_{\Phi_j} \leftarrow 0, \hat{Q}_{\Phi_j} \leftarrow 0$
13: FLAG $\leftarrow$ FALSE
14: SumRewards $\leftarrow 0$
15: **while** FLAG is FALSE **do**
16:     FLAG $\leftarrow$ TRUE with probability $1 - \gamma$.
17:     Execute SystemUpdate
18:     SumRewards $\leftarrow$ SumRewards $+ \sum_{k \in [K]} \theta_k r_k(x_k^t, u_k^t, \boldsymbol{\mu}_t, \boldsymbol{\nu}_t)$
19: **end while**

20: With probability $\frac{1}{2}$, $\hat{V}_{\Phi_j} \leftarrow$ SumRewards. Otherwise $\hat{Q}_{\Phi_j} \leftarrow$ SumRewards.
21: $\hat{A}_{\Phi_j}(\mathbf{x}_T, \boldsymbol{\mu}_T, \mathbf{u}_T) \leftarrow 2(\hat{Q}_{\Phi_j} - \hat{V}_{\Phi_j})$.
Output: $(\mathbf{x}_T, \boldsymbol{\mu}_T, \mathbf{u}_T)$ and $\hat{A}_{\Phi_j}(\mathbf{x}_T, \boldsymbol{\mu}_T, \mathbf{u}_T)$

Procedure SystemUpdate:
1: Execute the actions $\mathbf{u}_t \triangleq \{u_k^t\}_{k \in [K]}$.
2: Transition to $\mathbf{x}_{t+1} \triangleq \{x_k^{t+1}\}_{k \in [K]}$ following $x_k^{t+1} \sim P_k(x_k^t, u_k^t, \boldsymbol{\mu}_t, \boldsymbol{\nu}_t), \forall k \in [K]$.
3: $\boldsymbol{\mu}_{t+1} \leftarrow P^{\mathrm{MF}}(\boldsymbol{\mu}_t, \boldsymbol{\pi}_{\Phi_j})$ where P^{MF} is defined in (14).
4: Sample $\mathbf{u}_{t+1} \triangleq \{u_k^{t+1}\}_{k \in [K]} \sim \boldsymbol{\pi}_{\Phi_j}(\mathbf{x}_{t+1}, \boldsymbol{\mu}_{t+1})$ i.e., sample $u_k^{t+1} \sim \pi_{\Phi_j}^k(x_k^{t+1}, \boldsymbol{\mu}_{t+1})$,
    $\forall k \in [K]$.
5: $\boldsymbol{\nu}_{t+1} \leftarrow \nu^{\mathrm{MF}}(\boldsymbol{\mu}_{t+1}, \boldsymbol{\pi}_{\Phi_j})$ where ν^{MF} is defined in (13).
6: $t \leftarrow t + 1$
EndProcedure

---

for some positive constant G.

Assumption 8 $\forall \Phi_1, \Phi_2 \in \mathbb{R}^d, \forall \boldsymbol{\mu} \in \mathcal{P}(\mathcal{X} \times [K]), \forall x_k \in \mathcal{X}, \forall u_k \in \mathcal{U}, \forall k \in [K],$

$$\left| \nabla_{\Phi_1} \log \prod_{k \in [K]} \pi_{\Phi_1}^k (x_k, \boldsymbol{\mu})(u_k) - \nabla_{\Phi_2} \log \prod_{k \in [K]} \pi_{\Phi_2}^k (x_k, \boldsymbol{\mu})(u_k) \right|_1 \leq M |\Phi_1 - \Phi_2|_1$$

for some positive constant M.

Assumption 9 $\forall \Phi \in \mathbb{R}^d, \forall \boldsymbol{\mu}_0 \in \mathcal{P}(\mathcal{X} \times [K])$, *the following holds true*

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

where Φ^* is the parameter associated with an optimal policy.

Lemma 4 *If $\{\Phi_j\}_{j=1}^J$ are computed following Algorithm 1, and Assumptions 6−9 are satisfied, then for appropriate choices of η, α, J, L,*

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

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

The parameter ϵ_{bias} measures the capacity of parametrization. For rich neural network based policies, we can assume ϵ_{bias} to be small (Liu et al., 2020).

Lemma 4 states that, with a sample complexity of $\mathcal{O}(\epsilon^{-3})$, Algorithm 1 can approximate the optimal mean-field value function with an error margin of ϵ. Combining this with Theorem 1, we obtain the following result.

Theorem 5 *Let $\mathbf{x}_0^{\mathbf{N}}$ be the initial states and $\boldsymbol{\mu}_0$ their associated distribution. If the parameters $\{\Phi_j\}_{j=1}^J$ are obtained by following Algorithm 1, then under Assumptions 1, 3, and the set of assumptions used in Lemma 4, the following inequality holds for appropriate choices of η, α, J, L if $\gamma S_P < 1$*

$$\left| \sup_{\Phi \in \mathbb{R}^d} v^{\mathbf{N}}(\boldsymbol{\mu}_0, \pi_\Phi) - \frac{1}{J} \sum_{j=1}^J v^{\text{MF}}(\boldsymbol{\mu}_0, \boldsymbol{\pi}_{\Phi_j}) \right| \leq \frac{\sqrt{\epsilon_{\text{bias}}}}{1 - \gamma} + C e_1$$

$$\text{where } e_1 \triangleq \left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|} \right] \frac{1}{N_{\text{pop}}} \sum_{k \in [K]} \sqrt{N_k} \tag{25}$$

where S_P is defined in Theorem 1, C is a constant and the parameter ϵ_{bias} is defined in Lemma 4. The sample complexity of the process is $\mathcal{O}(e_1^{-3})$.

Theorem 5 states that, with a sample complexity of $\mathcal{O}(e_1^{-3})$, Algorithm 1 generates a policy which is within $\mathcal{O}(e_1)$ error of the optimal heterogeneous MARL policy.

Note that both time and space complexity of the sampling step in Algorithm 1 is $\mathcal{O}(K)$. In contrast, if NPG is directly applied to MARL, those complexities increase to $\mathcal{O}(N_{\text{pop}})$. Therefore, MFC based NPG provides an advantage of the order of N_{pop}/K in comparison to MARL based NPG.

In the following subsections, we shall establish results similar to Theorem 5 for the set of assumptions used in Theorem 2 and 3.

8.1 NPG with Assumption 2 and 4

If a multi-agent system satisfies Assumption 2, 4, and the set of stationary policies, $\bar{\Pi}$ is parametrized by $\Phi \in \mathbb{R}^d$, then similar to Algorithm 1, an NPG-based algorithm can be made to obtain its global optimal policy within $\bar{\Pi}$. Let this algorithm be denoted as $\text{NPG}_{2,4}$. Algorithm $\text{NPG}_{2,4}$ is identical to Algorithm 1 except the joint distribution $\boldsymbol{\mu} \in \mathcal{P}(\mathcal{X} \times [K])$ in Algorithm 1 is replaced by $\bar{\boldsymbol{\mu}} \in \mathcal{P}^K(\mathcal{X})$, in $\text{NPG}_{2,4}$. To show its global convergence, we need to assume a set of assumptions that are identical to those used in Lemma 4, except the joint distributions in all those assumptions must be replaced by the collection of distributions over all classes. Let this set of assumptions be denoted as $\text{ASMP}_{2,4}$.

Following the same line of argument as is used in Theorem 5, we can derive the result stated below.

Theorem 6 *Let* $\mathbf{x}_0^{\mathbf{N}}$ *be the initial states and* $\bar{\boldsymbol{\mu}}_0 \in \mathcal{P}^K(\mathcal{X})$ *their associated distribution. If the parameters* $\{\Phi_j\}_{j=1}^J$ *are obtained by following* $\text{NPG}_{2,4}$*, then under Assumptions 2, 4, and* $\text{ASMP}_{2,4}$*, the following inequality holds for appropriate choices of the Algorithm parameters,* η, α, J, L *if* $\gamma \bar{S}_P < 1$*.*

$$\left| \sup_{\Phi \in \mathbb{R}^d} \bar{v}^{\mathbf{N}}(\mathbf{x}_0^{\mathbf{N}}, \bar{\pi}_\Phi) - \frac{1}{J} \sum_{j=1}^J \bar{v}^{\text{MF}}(\bar{\boldsymbol{\mu}}_0, \bar{\boldsymbol{\pi}}_{\Phi_j}) \right| \leq \frac{\sqrt{\epsilon_{\text{bias}}}}{1 - \gamma} + \bar{C} e_2$$

$$\text{where } e_2 \triangleq \left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|} \right] \sum_{k \in [K]} \frac{1}{\sqrt{N_k}}$$

(26)

where $\bar{v}^{\mathbf{N}}$ *is the empirical value function of the* $\mathbf{N}$*-agent system,* $\bar{v}^{\text{MF}}$ *is its mean-field limit,* $\bar{\pi}_\Phi$ *is the stationary decision rules associated with the policy* $\bar{\boldsymbol{\pi}}_\Phi$*,* $\bar{S}_P$ *is defined in Theorem 2,* $\bar{C}$ *is a constant and the parameter* ϵ_{bias} *is a measure of the capacity of parametrization. The sample complexity of the process is* $\mathcal{O}(e_2^{-3})$*.*

Theorem 6 states that, with a sample complexity of $\mathcal{O}(e_2^{-3})$, Algorithm $\text{NPG}_{2,4}$ can approximate the empirical value function of MARL within an error margin of $\mathcal{O}(e_2)$.

8.2 NPG with Assumption 5

If a multi-agent system satisfies Assumption 5, and the set of stationary policies, Π is parametrized by $\Phi \in \mathbb{R}^d$, then similar to Algorithm 1, an NPG-based algorithm can be made to obtain its global optimal policy within Π. Let this algorithm be denoted as NPG_5. Algorithm NPG_5 is identical to Algorithm 1 except the joint distribution $\boldsymbol{\mu} \in \mathcal{P}(\mathcal{X} \times [K])$ in Algorithm 1 must be replaced by $\boldsymbol{\mu}[\mathcal{X}] \in \mathcal{P}(\mathcal{X})$, in NPG_5. To show its global convergence, we need to assume a set of assumptions that are same as those used in Lemma 4, except the joint distributions in those assumptions must be replaced by marginal distributions. Let this set of assumptions be denoted as ASMP_5. Following the same line of argument as is used in Theorem 5, we can derive the result stated below.

Theorem 7 *Let* $\mathbf{x}_0^{\mathbf{N}}$ *be the initial states and* $\boldsymbol{\mu}_0 \in \mathcal{P}(\mathcal{X} \times [K])$ *their associated joint distribution. If the parameters* $\{\Phi_j\}_{j=1}^J$ *are obtained by following* NPG_5*, then under Assumptions*

5, and ASMP$_5$*, the following inequality holds for appropriate choices of the Algorithm parameters,* η, α, J, L *if* $\gamma S_P < 1$.

$$\left| \sup_{\Phi \in \mathbb{R}^d} v^{\mathbf{N}}(\mathbf{x}_0^{\mathbf{N}}, \pi_\Phi) - \frac{1}{J} \sum_{j=1}^{J} v^{\mathrm{MF}}(\boldsymbol{\mu}_0, \boldsymbol{\pi}_{\Phi_j}) \right| \leq \frac{\sqrt{\epsilon_{\mathrm{bias}}}}{1 - \gamma} + e_3$$

$$\textit{where } e_3 \triangleq \left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|} \right] \left[\frac{A}{N_{\mathrm{pop}}} \sum_{k=1}^{K} \sqrt{N_k} + \frac{B}{\sqrt{N_{\mathrm{pop}}}} \right]$$

(27)

where $v^{\mathbf{N}}$ *is the empirical value function of the* $\mathbf{N}$*-agent system,* v^{MF} *is its mean-field limit,* π_Φ *is the stationary decision rules associated with policy* $\boldsymbol{\pi}_\Phi$*,* S_P *is defined in Theorem 1,* A, B *are constants and the parameter* ϵ_{bias} *is a measure of the capacity of parametrization. The sample complexity of the process is* $\mathcal{O}(e_3^{-3})$*.*

Theorem 7 states that, with $\mathcal{O}(e_3^{-3})$ sample complexity, Algorithm NPG$_5$ can approximate the empirical value function of MARL within an error margin of $\mathcal{O}(e_3)$.

9. Conclusions

In this paper, we prove that a K-class heterogeneous cooperative MARL problem can be approximated by its associated MFC problem. We also provide estimates of the approximation error as a function of class sizes for various set of assumptions. Finally, we propose a natural policy gradient based algorithm that approximates the optimal MARL policy in a sample efficient manner. Exchangeability among agents is one of the most important assumptions in MFC-type analyses. It allows the influence of the whole population to be summarized by the state-action distribution. In many scenarios of practical interest, however, agents interact only with certain number of neighbouring agents. As a result, the presumption of exchangeability may only hold locally. Establishing MFC-type approximation for system with limited agent exchangeability is an important direction to pursue in the future.

Appendix A. Proof of Theorem 1

The following results are needed to prove the theorem. The proofs of Lemma 8-14 are relegated to Appendix D-J respectively.

A.1 Continuity Lemmas

Lemma 8 *If $\nu^{\mathrm{MF}}(\cdot, \cdot)$ is defined by (13), then $\forall \boldsymbol{\mu}, \boldsymbol{\mu}' \in \mathcal{P}(\mathcal{X} \times [K])$ and $\forall \boldsymbol{\pi} = \{\pi_k\}_{k \in [K]}$ where π_k's are decision rules satisfying Assumption 3, the following inequality holds.*

$$|\nu^{\mathrm{MF}}(\boldsymbol{\mu}, \boldsymbol{\pi}) - \nu^{\mathrm{MF}}(\boldsymbol{\mu}', \boldsymbol{\pi})|_1 \leq (1 + L_Q)|\boldsymbol{\mu} - \boldsymbol{\mu}'|_1 \tag{28}$$

Lemma 9 *If $r_k^{\mathrm{MF}}(\cdot, \cdot)$ satisfies (15), then $\forall \boldsymbol{\mu}, \boldsymbol{\mu}' \in \mathcal{P}(\mathcal{X} \times [K])$ and $\forall \boldsymbol{\pi} = \{\pi_k\}_{k \in [K]}$ where π_k's are decision rules satisfying Assumption 3, the following inequality holds.*

$$\sum_{k \in [K]} |r_k^{\mathrm{MF}}(\boldsymbol{\mu}, \boldsymbol{\pi}) - r_k^{\mathrm{MF}}(\boldsymbol{\mu}', \boldsymbol{\pi})| \leq S_R |\boldsymbol{\mu} - \boldsymbol{\mu}'|_1$$
$$\text{where } S_R \triangleq M_R(1 + L_Q) + L_R [2 + L_Q] \tag{29}$$

Lemma 10 *If $P^{\mathrm{MF}}(\cdot, \cdot)$ is defined by (14), then $\forall \boldsymbol{\mu}, \boldsymbol{\mu}' \in \mathcal{P}(\mathcal{X} \times [K])$ and $\forall \boldsymbol{\pi} = \{\pi_k\}_{k \in [K]}$ where π_k's denote decision rules satisfying Assumption 3, the following inequality holds.*

$$|P^{\mathrm{MF}}(\boldsymbol{\mu}, \boldsymbol{\pi}) - P^{\mathrm{MF}}(\boldsymbol{\mu}', \boldsymbol{\pi})|_1 \leq S_P |\boldsymbol{\mu} - \boldsymbol{\mu}'|_1$$
$$\text{where } S_P \triangleq (1 + L_Q) + L_P [2 + L_Q] \tag{30}$$

Lemma 8-10 essentially state that the average reward function, $r_k^{\mathrm{MF}}(\cdot, \cdot)$ defined by (15) and the state and action evolution operators $P^{\mathrm{MF}}(\cdot, \cdot)$, $\nu^{\mathrm{MF}}(\cdot, \cdot)$ defined by (14), (13) respectively are Lipschitz continuous. These lemmas will be important in deriving the main result.

A.2 Approximation Lemmas

Recall that our primary goal is to prove that the value functions generated by a certain policy in a finite agent system can be well approximated by those generated by the same policy in the mean-field limit. As a precursor to this grand target, in this section, we discuss how various components of the value functions themselves behave when the population sizes become large. Lemma 11 serves as a key ingredient in many of the forthcoming lemmas.

Lemma 11 *If $\forall m \in [M]$, $\{X_{m,n}\}_{n \in [N]}$ are independent random variables bounded within $[0,1]$ with $\sum_{m \in [M]} \mathbb{E}[X_{m,n}] = 1$, $\forall n \in [N]$ and $\{C_{m,n}\}_{m \in [M], n \in [N]} \in \mathbb{R}$ are constants obeying $|C_{m,n}| \leq C$, $\forall m \in [M], \forall n \in [N]$, then the following holds.*

$$\sum_{m=1}^{M} \mathbb{E} \left| \sum_{n=1}^{N} C_{m,n} \left(X_{m,n} - \mathbb{E}[X_{m,n}] \right) \right| \leq C \sqrt{MN} \tag{31}$$

Below we state our first approximation result. Essentially, Lemma 12 provides an estimate of the difference between the empirical action distributions, $\boldsymbol{\nu}_t^{\mathbf{N}}$ and the action distribution that would have been obtained by following the mean-field action evolution operator $\nu(\cdot, \cdot)$, defined by (13), in a finite agent system.

Lemma 12 *If* $\{\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\nu}_t^{\mathbf{N}}\}_{t \in \{0,1,\cdots\}}$ *are empirical joint state and action distributions induced by the policy* $\boldsymbol{\pi} = \{\boldsymbol{\pi}_t\}_{t \in \{0,1,\cdots\}}$, *then the following inequality holds* $\forall t \in \{0, 1, \cdots\}$.

$$\mathbb{E}|\boldsymbol{\nu}_t^{\mathbf{N}} - \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t)|_1 \leq \frac{1}{N_{\mathrm{pop}}} \left(\sum_{k \in [K]} \sqrt{N_k} \right) \sqrt{|\mathcal{U}|} \tag{32}$$

Lemma 13 (stated below) bounds the error between the empirical average reward and the reward obtained by following the mean-field averaging process quantified by (15).

Lemma 13 *If* $\{\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\nu}_t^{\mathbf{N}}\}_{t \in \{0,1,\cdots\}}$ *are empirical joint state and action distributions induced by the policy* $\boldsymbol{\pi} = \{\boldsymbol{\pi}_t\}_{t \in \{0,1,\cdots\}}$, *then the following holds* $\forall t \in \{0, 1, \cdots\}$.

$$\mathbb{E} \left| \frac{1}{N_{\mathrm{pop}}} \sum_{k \in [K]} \sum_{j=1}^{N_k} r_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\nu}_t^{\mathbf{N}}) - \sum_{k \in [K]} r_k^{\mathrm{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t) \right| \leq C_R \sqrt{|\mathcal{U}|} \frac{1}{N_{\mathrm{pop}}} \left(\sum_{k \in [K]} \sqrt{N_k} \right) \tag{33}$$

where $C_R = M_R + L_R$.

Finally, Lemma 14 computes an upper bound on the error between the empirical state distribution, $\boldsymbol{\mu}_{t+1}^{\mathbf{N}}$ and the distribution that would have been obtained by following the mean-field state distribution evolution operator $P(\cdot, \cdot)$, defined by (14) in a finite agent system.

Lemma 14 *If* $\{\boldsymbol{\mu}_t^{\mathbf{N}}\}_{t \in \{0,1,\cdots\}}$ *are empirical joint state distributions induced by the policy* $\boldsymbol{\pi} = \{\boldsymbol{\pi}_t\}_{t \in \{0,1,\cdots\}}$, *then the following inequality holds* $\forall t \in \{0, 1, \cdots\}$.

$$\mathbb{E} \left| \boldsymbol{\mu}_{t+1}^{\mathbf{N}} - P^{\mathrm{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t) \right|_1 \leq C_P \left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|} \right] \frac{1}{N_{\mathrm{pop}}} \left(\sum_{k \in [K]} \sqrt{N_k} \right) \tag{34}$$

where $C_P = 2 + L_P$.

A.3 Proof of the Theorem

We are now ready to prove the theorem. Using $(11), (12)$, and (16), we can write,

$$\left| v^{\mathbf{N}}(\mathbf{x}_0^{\mathbf{N}}, \boldsymbol{\pi}) - v^{\mathrm{MF}}(\boldsymbol{\mu}_0, \boldsymbol{\pi}) \right| \leq J_1 + J_2 \tag{35}$$

where the first term J_1 is defined as follows:

$$J_1 \triangleq \sum_{t=0}^{\infty} \gamma^t \mathbb{E} \left| \frac{1}{N_{\mathrm{pop}}} \sum_{k \in [K]} \sum_{j=1}^{N_k} r_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\nu}_t^{\mathbf{N}}) - \sum_{k \in [K]} r_k^{\mathrm{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t) \right|$$

$$\overset{(a)}{\leq} \frac{C_R}{1 - \gamma} \sqrt{|\mathcal{U}|} \frac{1}{N_{\mathrm{pop}}} \left(\sum_{k \in [K]} \sqrt{N_k} \right)$$

The inequality (a) follows from Lemma 13. The second term, J_2 is given as follows:

$$
\begin{aligned}
J_2 &\triangleq \sum_{t=0}^{\infty} \gamma^t \left| \sum_{k \in [K]} r_k^{\mathrm{MF}}(\boldsymbol{\mu}_t, \boldsymbol{\pi}_t) - \sum_{k \in [K]} \mathbb{E} \left[r_k^{\mathrm{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t) \right] \right| \\
&\leq \sum_{t=0}^{\infty} \gamma^t \sum_{k \in [K]} \left| r_k^{\mathrm{MF}}(\boldsymbol{\mu}_t, \boldsymbol{\pi}_t) - \mathbb{E} \left[r_k^{\mathrm{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t) \right] \right| \\
&\overset{(a)}{=} \sum_{t=0}^{\infty} \gamma^t \sum_{k \in [K]} \left| \mathbb{E} \left[r_k^{\mathrm{MF}}(\boldsymbol{\mu}_t, \boldsymbol{\pi}_t) - r_k^{\mathrm{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t) \right] \right| \\
&\leq \sum_{t=0}^{\infty} \gamma^t \sum_{k \in [K]} \mathbb{E} \left| r_k^{\mathrm{MF}}(\boldsymbol{\mu}_t, \boldsymbol{\pi}_t) - r_k^{\mathrm{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t) \right| \overset{(b)}{\leq} S_R \left(\sum_{t=0}^{\infty} \gamma^t \mathbb{E} \left| \boldsymbol{\mu}_t^{\mathbf{N}} - \boldsymbol{\mu}_t \right|_1 \right)
\end{aligned}
\tag{36}
$$

Equation (a) holds because the sequence $\{\boldsymbol{\mu}_t\}_{t \in \{0,1,\cdots\}}$ is deterministic. Inequality (b) is due to Lemma 9. Observe that, $\forall t \geq 0$ the following holds,

$$
\mathbb{E} \left| \boldsymbol{\mu}_{t+1}^{\mathbf{N}} - \boldsymbol{\mu}_{t+1} \right|_1 \leq \mathbb{E} \left| \boldsymbol{\mu}_{t+1}^{\mathbf{N}} - P^{\mathrm{MF}}\left(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t \right) \right|_1 + \mathbb{E} \left| P^{\mathrm{MF}}\left(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t \right) - \boldsymbol{\mu}_{t+1} \right|_1 \tag{37}
$$

The first term can be upper bounded by invoking Lemma 14. Using Lemma 10, the second term can be upper bounded as follows:

$$
\mathbb{E} \left| P^{\mathrm{MF}}\left(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t \right) - \boldsymbol{\mu}_{t+1} \right|_1 = \mathbb{E} \left| P^{\mathrm{MF}}\left(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t \right) - P^{\mathrm{MF}}\left(\boldsymbol{\mu}_t, \boldsymbol{\pi}_t \right) \right|_1 \leq S_P(\mathbb{E} | \boldsymbol{\mu}_t^{\mathbf{N}} - \boldsymbol{\mu}_t |_1) \tag{38}
$$

Recall that, $\boldsymbol{\mu}^{0,\mathbf{N}} = \boldsymbol{\mu}^0$. Therefore,

$$
\begin{aligned}
\mathbb{E} \left| \boldsymbol{\mu}_{t+1}^{\mathbf{N}} - \boldsymbol{\mu}_{t+1} \right|_1 &\leq C_P \left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|} \right] \frac{1}{N_{\mathrm{pop}}} \left(\sum_{k \in [K]} \sqrt{N_k} \right) + S_P \left(\mathbb{E} \left| \boldsymbol{\mu}_t^{\mathbf{N}} - \boldsymbol{\mu}_t \right|_1 \right) \\
&\leq C_P \left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|} \right] \frac{1}{N_{\mathrm{pop}}} \left(\sum_{k \in [K]} \sqrt{N_k} \right) \left(\frac{S_P^{t+1} - 1}{S_P - 1} \right)
\end{aligned}
\tag{39}
$$

Clearly, J_2 is upper bounded as follows,

$$
J_2 \leq C_P \left(\frac{S_R}{S_P - 1} \right) \left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|} \right] \frac{1}{N_{\mathrm{pop}}} \left(\sum_{k \in [K]} \sqrt{N_k} \right) \left(\frac{1}{1 - \gamma S_P} - \frac{1}{1 - \gamma} \right)
$$

This completes the proof of (17).

Appendix B. Proof of Theorem 2

The collection of empirical state and action distributions of all classes at time t are denoted as $\bar{\boldsymbol{\mu}}_t^{\mathbf{N}} \in \mathcal{P}^K(\mathcal{X})$ and $\bar{\boldsymbol{\nu}}_t^{\mathbf{N}} \in \mathcal{P}^K(\mathcal{U})$ respectively and their mean-field counterparts are $\bar{\boldsymbol{\mu}}_t, \bar{\boldsymbol{\nu}}_t$. The prior probability of k-th class, $k \in [K]$ will be denoted as $\theta_k = N_k/N_{\mathrm{pop}}$ and $\boldsymbol{\theta} \triangleq \{\theta_k\}_{k \in [K]}$.

B.1 Mean-field equations

For a policy $\bar{\boldsymbol{\pi}} \triangleq \{\bar{\boldsymbol{\pi}}_t\}_{t \in \{0,1,\cdots\}} \triangleq \{(\bar{\pi}_k^t)_{k \in [K]}\}_{t \in \{0,1,\cdots\}}$, the mean-field action distribution is updated as,

$$\begin{aligned}
\bar{\boldsymbol{\nu}}_t = \bar{\nu}^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}_t, \bar{\boldsymbol{\pi}}_t) &\triangleq \{\bar{\nu}_k^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}_t, \bar{\boldsymbol{\pi}}_t)\}_{k \in [K]} \\
\bar{\nu}_k^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}_t, \bar{\boldsymbol{\pi}}_t) &\triangleq \sum_{x \in \mathcal{X}} \bar{\pi}_k^t(x, \bar{\boldsymbol{\mu}}_t) \bar{\boldsymbol{\mu}}_t(x, k)
\end{aligned} \tag{40}$$

Similarly, the state distribution is updated as,

$$\begin{aligned}
\bar{\boldsymbol{\mu}}_{t+1} = \bar{P}^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}_t, \bar{\boldsymbol{\pi}}_t) &\triangleq \{\bar{P}_k^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}_t, \bar{\boldsymbol{\pi}}_t)\}_{k \in [K]} \\
\bar{P}_k^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}_t, \bar{\boldsymbol{\pi}}_t) &\triangleq \sum_{x \in \mathcal{X}} \sum_{u \in \mathcal{U}} \bar{P}_k\left(x, u, \bar{\boldsymbol{\mu}}_t, \bar{\nu}^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}_t, \bar{\boldsymbol{\pi}}_t)\right) \times \bar{\boldsymbol{\mu}}_t(x, k) \bar{\pi}_k^t(x, \bar{\boldsymbol{\mu}}_t)(u)
\end{aligned} \tag{41}$$

Finally, the average reward of k-th class are computed as,

$$\bar{r}_k^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}_t, \bar{\boldsymbol{\pi}}_t) \triangleq \sum_{x \in \mathcal{X}} \sum_{u \in \mathcal{U}} \bar{r}_k(x, u, \bar{\boldsymbol{\mu}}_t, \bar{\nu}^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}_t, \bar{\boldsymbol{\pi}}_t)) \times \bar{\boldsymbol{\mu}}_t(x, k) \bar{\pi}_k^t(x, \bar{\boldsymbol{\mu}}_t)(u) \tag{42}$$

For an initial state distribution $\bar{\boldsymbol{\mu}}_0$, and a policy $\bar{\boldsymbol{\pi}}$, the infinite-horizon γ-discounted average reward in the mean-field limit is,

$$\bar{v}^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}_0, \bar{\boldsymbol{\pi}}) = \sum_{k \in [K]} \theta_k \sum_{t=0}^{\infty} \gamma^t \bar{r}_k^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}_t, \bar{\boldsymbol{\pi}}_t) \tag{43}$$

B.2 Helper Lemmas

The following results are necessary to prove the theorem. The proofs of Lemma 15, and 16 have been relegated to Appendix K, and L respectively.

Lemma 15 *The following inequalities hold $\forall \bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\mu}}' \in \mathcal{P}^K(\mathcal{X})$ and $\forall \bar{\boldsymbol{\pi}} = \{\bar{\pi}_k\}_{k \in [K]}$ where $\bar{\pi}_k$'s are decision rules satisfying Assumption 4.*

$$\begin{aligned}
&(a) \ |\bar{\nu}^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\pi}}) - \bar{\nu}^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}', \bar{\boldsymbol{\pi}})|_1 \le (1 + K\bar{L}_Q)|\bar{\boldsymbol{\mu}} - \bar{\boldsymbol{\mu}}'|_1 \\
&(b) \ \sum_{k \in [K]} \theta_k |\bar{r}_k^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\pi}}) - \bar{r}_k^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}', \bar{\boldsymbol{\pi}})| \le \bar{S}_R |\bar{\boldsymbol{\mu}} - \bar{\boldsymbol{\mu}}'|_1 \\
&(c) \ |\bar{P}^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\pi}}) - \bar{P}^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}', \bar{\boldsymbol{\pi}})|_1 \le \bar{S}_P |\bar{\boldsymbol{\mu}} - \bar{\boldsymbol{\mu}}'|_1
\end{aligned}$$

where $\bar{S}_R \triangleq \bar{M}_R(1 + \bar{L}_Q) + \bar{L}_R(2 + K\bar{L}_Q)$ and $\bar{S}_P \triangleq (1 + K\bar{L}_Q) + K\bar{L}_P(2 + K\bar{L}_Q)$.

Lemma 16 *If $\{\bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\boldsymbol{\nu}}_t^{\mathbf{N}}\}_{t \in \{0,1,\cdots\}}$ are the collections of empirical state and action distributions of each classes induced by policy $\bar{\boldsymbol{\pi}} = \{\bar{\boldsymbol{\pi}}_t\}_{t \in \{0,1,\cdots\}}$, then the following inequalities*

22

hold true $\forall t \in \{0, 1, \cdots\}$.

$$(a) \ \mathbb{E}|\bar{\boldsymbol{\nu}}_t^{\mathbf{N}} - \bar{\nu}^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\boldsymbol{\pi}}_t)|_1 \leq \left(\sum_{k \in [K]} \frac{1}{\sqrt{N_k}} \right) \sqrt{|\mathcal{U}|} \tag{44}$$

$$(b) \ \mathbb{E} \left| \frac{1}{N_{\mathrm{pop}}} \sum_{k \in [K]} \sum_{j=1}^{N_k} \bar{r}_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\boldsymbol{\nu}}_t^{\mathbf{N}}) - \sum_{k \in [K]} \theta_k \bar{r}_k^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\boldsymbol{\pi}}_t) \right|$$

$$\leq \bar{C}_R \left(\sum_{k \in [K]} \frac{1}{\sqrt{N_k}} \right) \sqrt{|\mathcal{U}|} \tag{45}$$

$$(c) \ \mathbb{E} \left| \bar{\boldsymbol{\mu}}_{t+1}^{\mathbf{N}} - \bar{P}^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\boldsymbol{\pi}}_t) \right|_1 \leq \bar{C}_P \left(\sum_{k \in [K]} \frac{1}{\sqrt{N_k}} \right) \left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|} \right] \tag{46}$$

where $\bar{C}_R \triangleq \bar{M}_R + \bar{L}_R$, *and* $\bar{C}_P \triangleq 2 + K\bar{L}_P$.

B.3 Proof of the Theorem

We are now ready to prove the theorem. Using (18) and (43), we can write,

$$\left| \bar{v}^{\mathbf{N}}(\mathbf{x}_0^{\mathbf{N}}, \bar{\boldsymbol{\pi}}) - \bar{v}^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}_0, \bar{\boldsymbol{\pi}}) \right| \leq J_1 + J_2 \tag{47}$$

where the first term J_1 is defined as follows:

$$J_1 \triangleq \sum_{t=0}^{\infty} \gamma^t \mathbb{E} \left| \frac{1}{N_{\mathrm{pop}}} \sum_{k \in [K]} \sum_{j=1}^{N_k} \bar{r}_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\boldsymbol{\nu}}_t^{\mathbf{N}}) - \sum_{k \in [K]} \theta_k \bar{r}_k^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\boldsymbol{\pi}}_t) \right|$$

$$\overset{(a)}{\leq} \frac{\bar{C}_R}{1 - \gamma} \left(\sum_{k \in [K]} \frac{1}{\sqrt{N_k}} \right) \sqrt{|\mathcal{U}|}$$

The inequality (a) follows from Lemma 16. The second term, J_2 is given as follows:

$$J_2 \triangleq \sum_{t=0}^{\infty} \gamma^t \left| \sum_{k \in [K]} \theta_k \bar{r}_k^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}_t, \bar{\boldsymbol{\pi}}_t) - \sum_{k \in [K]} \theta_k \mathbb{E} \left[\bar{r}_k^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\boldsymbol{\pi}}_t) \right] \right|$$

$$\leq \sum_{t=0}^{\infty} \gamma^t \sum_{k \in [K]} \theta_k \mathbb{E} \left| \bar{r}_k^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}_t, \bar{\boldsymbol{\pi}}_t) - \bar{r}_k^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\boldsymbol{\pi}}_t) \right| \tag{48}$$

$$\overset{(a)}{\leq} \bar{S}_R \left(\sum_{t=0}^{\infty} \gamma^t \mathbb{E} \left| \bar{\boldsymbol{\mu}}_t^{\mathbf{N}} - \bar{\boldsymbol{\mu}}_t \right|_1 \right)$$

Inequality (a) is due to Lemma 15. Observe that, $\forall t \geq 0$ the following holds,

$$\mathbb{E} \left| \bar{\boldsymbol{\mu}}_{t+1}^{\mathbf{N}} - \bar{\boldsymbol{\mu}}_{t+1} \right|_1 \leq \mathbb{E} \left| \bar{\boldsymbol{\mu}}_{t+1}^{\mathbf{N}} - \bar{P}^{\mathrm{MF}} \left(\bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\boldsymbol{\pi}}_t \right) \right|_1 + \mathbb{E} \left| \bar{P}^{\mathrm{MF}} \left(\bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\boldsymbol{\pi}}_t \right) - \bar{\boldsymbol{\mu}}_{t+1} \right|_1 \tag{49}$$

23

The first term can be upper bounded by invoking Lemma 16. Using Lemma 15, the second term can be upper bounded as follows:

$$\mathbb{E}\left|\bar{P}^{\mathrm{MF}}\left(\bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\boldsymbol{\pi}}_t\right) - \bar{\boldsymbol{\mu}}_{t+1}\right|_1 = \mathbb{E}\left|\bar{P}^{\mathrm{MF}}\left(\bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\boldsymbol{\pi}}_t\right) - \bar{P}^{\mathrm{MF}}\left(\bar{\boldsymbol{\mu}}_t, \bar{\boldsymbol{\pi}}_t\right)\right|_1 \leq \bar{S}_P(\mathbb{E}|\bar{\boldsymbol{\mu}}_t^{\mathbf{N}} - \bar{\boldsymbol{\mu}}_t|_1) \quad (50)$$

Recall that, $\bar{\boldsymbol{\mu}}^{0,\mathbf{N}} = \bar{\boldsymbol{\mu}}^0$. Therefore,

$$
\begin{aligned}
\mathbb{E}\left|\bar{\boldsymbol{\mu}}_{t+1}^{\mathbf{N}} - \bar{\boldsymbol{\mu}}_{t+1}\right|_1 &\leq \bar{C}_P\left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|}\right]\left(\sum_{k\in[K]}\frac{1}{\sqrt{N_k}}\right) + \bar{S}_P\left(\mathbb{E}\left|\bar{\boldsymbol{\mu}}_t^{\mathbf{N}} - \bar{\boldsymbol{\mu}}_t\right|_1\right) \\
&\leq \bar{C}_P\left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|}\right]\left(\sum_{k\in[K]}\frac{1}{\sqrt{N_k}}\right)\left(\frac{\bar{S}_P^{t+1} - 1}{\bar{S}_P - 1}\right)
\end{aligned}
\tag{51}
$$

Clearly, J_2 is upper bounded as follows,

$$J_2 \leq \bar{C}_P\left(\frac{\bar{S}_R}{\bar{S}_P - 1}\right)\left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|}\right]\left(\sum_{k\in[K]}\frac{1}{\sqrt{N_k}}\right)\left(\frac{1}{1 - \gamma\bar{S}_P} - \frac{1}{1 - \gamma}\right)$$

Appendix C. Proof of Theorem 3

The following results are required to prove the theorem. The proofs of Lemma 17, 18 are given in Appendix M, N respectively. We define mean-field state, action distribution evolution functions $P^{\mathrm{MF}}(\cdot,\cdot)$, $\nu^{\mathrm{MF}}(\cdot,\cdot)$ and the class-average reward functions $r_k^{\mathrm{MF}}(\cdot,\cdot)$'s by $(14), (13), (15)$, respectively.

C.1 Helper Lemmas

Lemma 17 *The following inequalities hold $\forall\boldsymbol{\mu}, \boldsymbol{\mu}' \in \mathcal{P}(\mathcal{X} \times [K])$ and $\forall\boldsymbol{\pi} = \{\pi_k\}_{k\in[K]}$ where π_k's denote Lipschitz continuous decision rules with parameter L_Q.*

$$
\begin{aligned}
(a) \quad &|\nu^{\mathrm{MF}}(\boldsymbol{\mu}, \boldsymbol{\pi})[\mathcal{U}] - \nu^{\mathrm{MF}}(\boldsymbol{\mu}', \boldsymbol{\pi})[\mathcal{U}]|_1 \leq |\nu^{\mathrm{MF}}(\boldsymbol{\mu}, \boldsymbol{\pi}) - \nu^{\mathrm{MF}}(\boldsymbol{\mu}', \boldsymbol{\pi})|_1 \\
&\hspace{8.5cm} \leq |\boldsymbol{\mu} - \boldsymbol{\mu}'|_1 + L_Q|\boldsymbol{\mu}[\mathcal{X}] - \boldsymbol{\mu}'[\mathcal{X}]|_1 \\
(b) \quad &\sum_{k\in[K]}|r_k^{\mathrm{MF}}(\boldsymbol{\mu}, \boldsymbol{\pi}) - r_k^{\mathrm{MF}}(\boldsymbol{\mu}', \boldsymbol{\pi})| \leq S_R'|\boldsymbol{\mu} - \boldsymbol{\mu}'|_1 + S_R''|\boldsymbol{\mu}[\mathcal{X}] - \boldsymbol{\mu}'[\mathcal{X}]|_1 \\
(c) \quad &|P^{\mathrm{MF}}(\boldsymbol{\mu}, \boldsymbol{\pi})[\mathcal{X}] - P^{\mathrm{MF}}(\boldsymbol{\mu}', \boldsymbol{\pi})[\mathcal{X}]|_1 \leq |P^{\mathrm{MF}}(\boldsymbol{\mu}, \boldsymbol{\pi}) - P^{\mathrm{MF}}(\boldsymbol{\mu}', \boldsymbol{\pi})|_1 \\
&\hspace{8.5cm} \leq S_P'|\boldsymbol{\mu} - \boldsymbol{\mu}'|_1 + S_P''|\boldsymbol{\mu}[\mathcal{X}] - \boldsymbol{\mu}'[\mathcal{X}]|_1
\end{aligned}
$$

where $S_R' \triangleq M_R + L_R$, $S_R'' \triangleq M_R L_Q + L_R(1 + L_Q)$, $S_P' \triangleq 1 + L_P$, and $S_P'' \triangleq L_Q + L_P(1 + L_Q)$. Note that, $S_R' + S_R'' = S_R$ and $S_P' + S_P'' = S_P$ where S_R, S_P are defined in $(29), (30)$ respectively.

Similar to Lemma 12, 13, and 14, we can derive the approximation results as follows.

Lemma 18 *If* $\{\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\nu}_t^{\mathbf{N}}\}_{t \in \{0,1,\cdots\}}$ *are the empirical joint state and action distributions induced by the policy* $\boldsymbol{\pi} = \{\boldsymbol{\pi}_t\}_{t \in \{0,1,\cdots\}}$, *then the following inequalities hold true* $\forall t \in \{0, 1, \cdots\}$.

(a) $\mathbb{E}|\boldsymbol{\nu}_t^{\mathbf{N}}[\mathcal{U}] - \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t)[\mathcal{U}]|_1 \leq \dfrac{1}{\sqrt{N_{\mathrm{pop}}}} \sqrt{|\mathcal{U}|}$ (52)

(b) $\mathbb{E} \left| \dfrac{1}{N_{\mathrm{pop}}} \displaystyle\sum_{k \in [K]} \sum_{j=1}^{N_k} r_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^{\mathbf{N}}[\mathcal{X}], \boldsymbol{\nu}_t^{\mathbf{N}}[\mathcal{U}]) - \displaystyle\sum_{k \in [K]} r_k^{\mathrm{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t) \right| \leq \dfrac{C_R}{\sqrt{N_{\mathrm{pop}}}} \sqrt{|\mathcal{U}|}$ (53)

(c) $\mathbb{E} \left| \boldsymbol{\mu}_{t+1}^{\mathbf{N}}[\mathcal{X}] - P^{\mathrm{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t)[\mathcal{X}] \right|_1 \leq \dfrac{C_P}{\sqrt{N_{\mathrm{pop}}}} \left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|} \right]$ (54)

where C_R, C_P *are same as defined in Lemma 13, 14 respectively.*

C.2 Proof of the Theorem

Following the proof of Theorem 1, we can write,

$$\left| v^{\mathbf{N}}(\mathbf{x}_0^{\mathbf{N}}, \boldsymbol{\pi}) - v^{\mathrm{MF}}(\boldsymbol{\mu}_0, \boldsymbol{\pi}) \right| \leq J_1 + J_2 \tag{55}$$

where the first term J_1 is defined as follows:

$$
\begin{aligned}
J_1 &\triangleq \sum_{t=0}^{\infty} \gamma^t \mathbb{E} \left| \frac{1}{N_{\mathrm{pop}}} \sum_{k \in [K]} \sum_{j=1}^{N_k} r_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^{\mathbf{N}}[\mathcal{X}], \boldsymbol{\nu}_t^{\mathbf{N}}[\mathcal{U}]) - \sum_{k \in [K]} r_k^{\mathrm{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t) \right| \\
&\overset{(a)}{\leq} \frac{C_R}{1 - \gamma} \sqrt{|\mathcal{U}|} \frac{1}{\sqrt{N_{\mathrm{pop}}}}
\end{aligned}
\tag{56}
$$

The inequality (a) follows from Lemma 18. The second term, J_2 is given as follows:

$$
\begin{aligned}
J_2 &\triangleq \sum_{t=0}^{\infty} \gamma^t \left| \sum_{k \in [K]} r_k^{\mathrm{MF}}(\boldsymbol{\mu}_t, \boldsymbol{\pi}_t) - \sum_{k \in [K]} \mathbb{E} \left[r_k^{\mathrm{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t) \right] \right| \\
&\leq \sum_{t=0}^{\infty} \gamma^t \sum_{k \in [K]} \mathbb{E} \left| r_k^{\mathrm{MF}}(\boldsymbol{\mu}_t, \boldsymbol{\pi}_t) - r_k^{\mathrm{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t) \right| \\
&\overset{(a)}{\leq} \sum_{t=0}^{\infty} \gamma^t \left\{ S_R' \mathbb{E} \left| \boldsymbol{\mu}_t^{\mathbf{N}} - \boldsymbol{\mu}_t \right|_1 + S_R'' \mathbb{E} \left| \boldsymbol{\mu}_t^{\mathbf{N}}[\mathcal{X}] - \boldsymbol{\mu}_t[\mathcal{X}] \right|_1 \right\}
\end{aligned}
\tag{57}
$$

Inequality (a) is due to Lemma 17. Observe that, $\forall t \geq 0$ the following holds,

$$
\begin{aligned}
&S_R' \mathbb{E} \left| \boldsymbol{\mu}_{t+1}^{\mathbf{N}} - \boldsymbol{\mu}_{t+1} \right|_1 + S_R'' \mathbb{E} \left| \boldsymbol{\mu}_{t+1}^{\mathbf{N}}[\mathcal{X}] - \boldsymbol{\mu}_{t+1}[\mathcal{X}] \right|_1 \\
&\leq S_R' \mathbb{E} \left| \boldsymbol{\mu}_{t+1}^{\mathbf{N}} - P^{\mathrm{MF}}\left(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t\right) \right|_1 + S_R'' \mathbb{E} \left| \boldsymbol{\mu}_{t+1}^{\mathbf{N}}[\mathcal{X}] - P^{\mathrm{MF}}\left(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t\right)[\mathcal{X}] \right|_1 \\
&\quad + S_R' \mathbb{E} \left| P^{\mathrm{MF}}\left(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t\right) - \boldsymbol{\mu}_{t+1} \right|_1 + S_R'' \mathbb{E} \left| P^{\mathrm{MF}}\left(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t\right)[\mathcal{X}] - \boldsymbol{\mu}_{t+1}[\mathcal{X}] \right|_1
\end{aligned}
\tag{58}
$$

25

The first two terms can be upper bounded by invoking Lemma 14 and 18 respectively. Utilising Lemma 17, the last two term can be upper bounded as follows:

$$
\begin{aligned}
\mathbb{E}\left|P^{\mathrm{MF}}\left(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t\right)[\mathcal{X}]-\boldsymbol{\mu}_{t+1}[\mathcal{X}]\right|_1 & \leq \mathbb{E}\left|P^{\mathrm{MF}}\left(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t\right)-\boldsymbol{\mu}_{t+1}\right|_1 \\
& =\mathbb{E}\left|P^{\mathrm{MF}}\left(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t\right)-P^{\mathrm{MF}}\left(\boldsymbol{\mu}_t, \boldsymbol{\pi}_t\right)\right|_1 \\
& \leq S_P^{\prime}(\mathbb{E}|\boldsymbol{\mu}_t^{\mathbf{N}}-\boldsymbol{\mu}_t|_1)+S_P^{\prime\prime}(\mathbb{E}|\boldsymbol{\mu}_t^{\mathbf{N}}[\mathcal{X}]-\boldsymbol{\mu}_t[\mathcal{X}]|_1)
\end{aligned}
\tag{59}
$$

Therefore, (58) can be rewritten as,

$$
\begin{aligned}
& S_R^{\prime} \mathbb{E}\left|\boldsymbol{\mu}_{t+1}^{\mathbf{N}}-\boldsymbol{\mu}_{t+1}\right|_1+S_R^{\prime\prime} \mathbb{E}\left|\boldsymbol{\mu}_{t+1}^{\mathbf{N}}[\mathcal{X}]-\boldsymbol{\mu}_{t+1}[\mathcal{X}]\right|_1 \\
& \leq C_P\left[\sqrt{|\mathcal{X}|}+\sqrt{|\mathcal{U}|}\right]\left[\frac{S_R^{\prime}}{N_{\mathrm{pop}}}\left(\sum_{k \in[K]} \sqrt{N_k}\right)+\frac{S_R^{\prime\prime}}{\sqrt{N_{\mathrm{pop}}}}\right] \\
& \quad+S_R\left(S_P^{\prime} \mathbb{E}\left|\boldsymbol{\mu}_t^{\mathbf{N}}-\boldsymbol{\mu}_t\right|_1+S_P^{\prime} \mathbb{E}\left|\boldsymbol{\mu}_t^{\mathbf{N}}[\mathcal{X}]-\boldsymbol{\mu}_t[\mathcal{X}]\right|_1\right)
\end{aligned}
\tag{60}
$$

where $S_R=S_R^{\prime}+S_R^{\prime\prime}$. Similarly, one can show that,

$$
\begin{aligned}
& S_P^{\prime} \mathbb{E}\left|\boldsymbol{\mu}_{t+1}^{\mathbf{N}}-\boldsymbol{\mu}_{t+1}\right|_1+S_P^{\prime\prime} \mathbb{E}\left|\boldsymbol{\mu}_{t+1}^{\mathbf{N}}[\mathcal{X}]-\boldsymbol{\mu}_{t+1}[\mathcal{X}]\right|_1 \\
& \leq C_P\left[\sqrt{|\mathcal{X}|}+\sqrt{|\mathcal{U}|}\right]\left[\frac{S_P^{\prime}}{N_{\mathrm{pop}}}\left(\sum_{k \in[K]} \sqrt{N_k}\right)+\frac{S_P^{\prime\prime}}{\sqrt{N_{\mathrm{pop}}}}\right] \\
& \quad+S_P\left(S_P^{\prime} \mathbb{E}\left|\boldsymbol{\mu}_t^{\mathbf{N}}-\boldsymbol{\mu}_t\right|_1+S_P^{\prime} \mathbb{E}\left|\boldsymbol{\mu}_t^{\mathbf{N}}[\mathcal{X}]-\boldsymbol{\mu}_t[\mathcal{X}]\right|_1\right)
\end{aligned}
\tag{61}
$$

Recall that $\boldsymbol{\mu}_0^{\mathbf{N}}=\boldsymbol{\mu}_0$. Combining the above results, we therefore obtain,

$$
\begin{aligned}
& S_R^{\prime} \mathbb{E}\left|\boldsymbol{\mu}_{t+1}^{\mathbf{N}}-\boldsymbol{\mu}_{t+1}\right|_1+S_R^{\prime\prime} \mathbb{E}\left|\boldsymbol{\mu}_{t+1}^{\mathbf{N}}[\mathcal{X}]-\boldsymbol{\mu}_{t+1}[\mathcal{X}]\right|_1 \\
& \leq C_P\left[\sqrt{|\mathcal{X}|}+\sqrt{|\mathcal{U}|}\right]\left\{\left[\frac{S_R^{\prime}}{N_{\mathrm{pop}}}\left(\sum_{k \in[K]} \sqrt{N_k}\right)+\frac{S_R^{\prime\prime}}{\sqrt{N_{\mathrm{pop}}}}\right]\right. \\
& \quad\left.+S_R\left[\frac{S_P^{\prime}}{N_{\mathrm{pop}}}\left(\sum_{k \in[K]} \sqrt{N_k}\right)+\frac{S_P^{\prime\prime}}{\sqrt{N_{\mathrm{pop}}}}\right]\left(\frac{S_P^t-1}{S_P-1}\right)\right\}
\end{aligned}
\tag{62}
$$

Clearly, J_2 is upper bounded as follows,

$$
\begin{aligned}
J_2 \leq & C_P\left[\sqrt{|\mathcal{X}|}+\sqrt{|\mathcal{U}|}\right]\left(\frac{\gamma}{1-\gamma}\right)\left[\frac{S_R^{\prime}}{N_{\mathrm{pop}}}\left(\sum_{k \in[K]} \sqrt{N_k}\right)+\frac{S_R^{\prime\prime}}{\sqrt{N_{\mathrm{pop}}}}\right] \\
& +C_P\left(\frac{S_R}{S_P-1}\right)\left[\sqrt{|\mathcal{X}|}+\sqrt{|\mathcal{U}|}\right]\left[\frac{S_P^{\prime}}{N_{\mathrm{pop}}}\left(\sum_{k \in[K]} \sqrt{N_k}\right)+\frac{S_P^{\prime\prime}}{\sqrt{N_{\mathrm{pop}}}}\right]\left(\frac{\gamma}{1-\gamma S_P}-\frac{\gamma}{1-\gamma}\right)
\end{aligned}
$$

This completes the proof of the Theorem.

Appendix D. Proof of Lemma 8

The following chain of inequalities hold true.

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

Inequality (a) follows from Assumption 3 and the fact that $\pi_k(x, \boldsymbol{\mu})$ is a distribution. Finally, equality (b) uses the fact that $\boldsymbol{\mu}'$ is a distribution. This concludes the result.

Appendix E. Proof of Lemma 9

Note that,

$$
\begin{aligned}
\sum_{k \in [K]} |r_k^{\mathrm{MF}}&(\boldsymbol{\mu}, \boldsymbol{\pi}) - r_k^{\mathrm{MF}}(\boldsymbol{\mu}', \boldsymbol{\pi})| \\
&\leq \sum_{k \in [K]} \sum_{x \in \mathcal{X}} \sum_{u \in \mathcal{U}} |r_k(x, u, \boldsymbol{\mu}, \nu^{\mathrm{MF}}(\boldsymbol{\mu}, \boldsymbol{\pi})) - r_k(x, u, \boldsymbol{\mu}', \nu^{\mathrm{MF}}(\boldsymbol{\mu}', \boldsymbol{\pi}))| \times \boldsymbol{\mu}(x, k) \pi_k(x, \boldsymbol{\mu})(u) \\
&+ \sum_{k \in [K]} \sum_{x \in \mathcal{X}} \sum_{u \in \mathcal{U}} |r_k(x, u, \boldsymbol{\mu}', \nu^{\mathrm{MF}}(\boldsymbol{\mu}', \boldsymbol{\pi}))| \times |\boldsymbol{\mu}(x, k) \pi_k(x, \boldsymbol{\mu})(u) - \boldsymbol{\mu}'(x, k) \pi_k(x, \boldsymbol{\mu}')(u)|
\end{aligned}
$$

Utilising Assumption 1(c), and the facts that $\boldsymbol{\mu}$, $\pi_k(x, \boldsymbol{\mu})$ are probability distributions, the first term can be upper bounded by the following expression,

$$
L_R \left(|\boldsymbol{\mu} - \boldsymbol{\mu}'|_1 + |\nu^{\mathrm{MF}}(\boldsymbol{\mu}, \boldsymbol{\pi}) - \nu^{\mathrm{MF}}(\boldsymbol{\mu}', \boldsymbol{\pi})|_1 \right) \leq L_R [1 + (1 + L_Q)] |\boldsymbol{\mu} - \boldsymbol{\mu}'|_1
$$

Lemma 8 is applied to derive the above inequality. Utilising Assumption 1(b), the second term can be upper bounded by the following quantity:

$$M_R \sum_{k\in[K]} \sum_{x\in\mathcal{X}} \sum_{u\in\mathcal{U}} |\boldsymbol{\mu}(x,k)\pi_k(x,\boldsymbol{\mu})(u) - \boldsymbol{\mu}'(x,k)\pi_k(x,\boldsymbol{\mu}')(u)| \overset{(a)}{\leq} M_R(1+L_Q)|\boldsymbol{\mu}-\boldsymbol{\mu}'|_1$$

Inequality (a) can be proved using identical arguments as used in Lemma 8. This concludes the result.

Appendix F. Proof of Lemma 10

Note that,

$$|P^{\mathrm{MF}}(\boldsymbol{\mu},\boldsymbol{\pi}) - P^{\mathrm{MF}}(\boldsymbol{\mu}',\boldsymbol{\pi})|_1 = \sum_{k\in[K]} |P_k^{\mathrm{MF}}(\boldsymbol{\mu},\boldsymbol{\pi}) - P_k^{\mathrm{MF}}(\boldsymbol{\mu}',\boldsymbol{\pi})|_1$$

$$\leq \sum_{k\in[K]} \sum_{x\in\mathcal{X}} \sum_{u\in\mathcal{U}} |P_k(x,u,\boldsymbol{\mu},\nu^{\mathrm{MF}}(\boldsymbol{\mu},\boldsymbol{\pi})) - P_k(x,u,\boldsymbol{\mu}',\nu^{\mathrm{MF}}(\boldsymbol{\mu}',\boldsymbol{\pi}))|_1 \times \boldsymbol{\mu}(x,k)\pi_k(x,\boldsymbol{\mu})(u)$$

$$+ \sum_{k\in[K]} \sum_{x\in\mathcal{X}} \sum_{u\in\mathcal{U}} |P_k(x,u,\boldsymbol{\mu}',\nu^{\mathrm{MF}}(\boldsymbol{\mu}',\boldsymbol{\pi}))|_1 \times |\boldsymbol{\mu}(x,k)\pi_k(x,\boldsymbol{\mu})(u) - \boldsymbol{\mu}'(x,k)\pi_k(x,\boldsymbol{\mu}')(u)|$$

Utilising Assumption 1(d), and the facts that $\boldsymbol{\mu}$, $\pi_k(x,\boldsymbol{\mu})$ are probability distributions, the first term can be upper bounded by the following expression,

$$L_P\left(|\boldsymbol{\mu}-\boldsymbol{\mu}'|_1 + |\nu^{\mathrm{MF}}(\boldsymbol{\mu},\boldsymbol{\pi}) - \nu^{\mathrm{MF}}(\boldsymbol{\mu}',\boldsymbol{\pi})|_1\right) \leq L_P\left[1+(1+L_Q)\right]|\boldsymbol{\mu}-\boldsymbol{\mu}'|_1$$

Lemma 8 is applied to derive the above inequality. Note that, $|P_k(x,u,\boldsymbol{\mu}',\nu(\boldsymbol{\mu}',\boldsymbol{\pi}))|_1 = 1$. Therefore, the second term can be bounded by the following quantity.

$$\sum_{k\in[K]} \sum_{x\in\mathcal{X}} \sum_{u\in\mathcal{U}} |\boldsymbol{\mu}(x,k)\pi_k(x,\boldsymbol{\mu})(u) - \boldsymbol{\mu}'(x,k)\pi_k(x,\boldsymbol{\mu}')(u)| \overset{(a)}{\leq} (1+L_Q)|\boldsymbol{\mu}-\boldsymbol{\mu}'|_1$$

Inequality (a) can be proved using identical arguments as used in Lemma 8. This concludes the result.

Appendix G. Proof of Lemma 11

Let, $Y_{m,n} \triangleq X_{m,n} - \mathbb{E}[X_{m,n}]$, $\forall m \in [M]$, $\forall m \in [N]$. We need the following results to prove Lemma 11.

Proposition 19 $\forall m \in [M]$, $\forall n \in [N]$, $\mathbb{E}[Y_{m,n}^2] \leq \mathbb{E}[X_{m,n}]$.

Proof For random variables $X_{m,n} \in [0,1]$, note that,

$$\mathbb{E}[Y_{m,n}^2] = E[X_{m,n}^2] - (\mathbb{E}[X_{m,n}])^2$$
$$\leq E[X_{m,n}] - (\mathbb{E}[X_{m,n}])^2 \leq \mathbb{E}[X_{m,n}]$$

∎

28

Proposition 20 $\forall m \in [M]$, $\mathbb{E}[\sum_{n=1}^{N} C_{m,n} Y_{m,n}]^2 \leq C^2 \sum_{n=1}^{N} \mathbb{E}[Y_{m,n}^2]$.

Proof Using the independence of $Y_{m,n}$'s, we deduce, $\forall m \in [M]$,

$$
\mathbb{E}\left[\sum_{n=1}^{N} C_{m,n} Y_{m,n}\right]^2
$$

$$
= \mathbb{E}\left[\sum_{n_1=1}^{N} \sum_{n_2=1}^{N} C_{m,n_1} C_{m,n_2} Y_{m,n_1} Y_{m,n_2}\right]
$$

$$
= \sum_{n=1}^{N} C_{m,n}^2 \mathbb{E}\left[Y_{m,n}^2\right] + 2 \sum_{n_1=1}^{N} \sum_{n_2 > n_1}^{N} C_{m,n_1} C_{m,n_2} \mathbb{E}[Y_{m,n_1}] \mathbb{E}[Y_{m,n_2}]
$$

$$
\overset{(a)}{=} \sum_{n=1}^{N} C_{m,n}^2 \mathbb{E}\left[Y_{m,n}^2\right]
$$

$$
\leq C^2 \sum_{n=1}^{N} \mathbb{E}[Y_{m,n}^2]
$$

Equality (a) uses the fact that $\mathbb{E}[Y_{m,n}] = 0$, $\forall m \in [M], \forall n \in [N]$. ∎

We are now ready to prove Lemma 11. Note that,

$$
\sum_{m=1}^{M} \mathbb{E}\left|\sum_{n=1}^{N} C_{m,n} Y_{m,n}\right|
$$

$$
\overset{(a)}{\leq} \sqrt{M} \left\{\sum_{m=1}^{M} \mathbb{E}\left[\sum_{n=1}^{N} C_{m,n} Y_{m,n}\right]^2\right\}^{\frac{1}{2}}
$$

$$
\overset{(b)}{\leq} C\sqrt{M} \left\{\sum_{m=1}^{M} \sum_{n=1}^{N} \mathbb{E}\left[Y_{m,n}^2\right]\right\}^{\frac{1}{2}}
$$

$$
\overset{(c)}{\leq} C\sqrt{M} \left\{\sum_{n=1}^{N} \sum_{m=1}^{M} \mathbb{E}\left[X_{m,n}\right]\right\}^{\frac{1}{2}}
$$

$$
= C\sqrt{MN}
$$

Result (a) is a consequence of Cauchy-Schwarz inequality, and (b), (c) follow from Proposition 20, and 19 respectively. This concludes the result.

Appendix H. Proof of Lemma 12

Using the definition of L_1-norm, we get:

$$\mathbb{E}\left|\boldsymbol{\nu}_t^{\mathbf{N}} - \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t)\right|_1 = \sum_{k \in [K]} \sum_{u \in \mathcal{U}} \mathbb{E}\left|\boldsymbol{\nu}_t^{\mathbf{N}}(u, k) - \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t)(u, k)\right|$$

$$= \frac{1}{N_{\mathrm{pop}}} \sum_{k \in [K]} \sum_{u \in \mathcal{U}} \mathbb{E}\left|\sum_{j=1}^{N_k} \delta(u_{j,k}^{t,\mathbf{N}} = u) - \sum_{j=1}^{N_k} \pi_k^t(x_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^{\mathbf{N}})(u)\right|$$

Recall from Observation 1 that, the random variables $u_{j,k}^{t,\mathbf{N}}$'s are independent conditioned on $\mathbf{x}_t^{\mathbf{N}}$. Also, it is easy to check the following relations,

$$\mathbb{E}\left[\delta(u_{j,k}^{t,\mathbf{N}} = u)\Big|\mathbf{x}_t^{\mathbf{N}}\right] = \pi_k^t(x_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^{\mathbf{N}})(u) \text{ and } \sum_{u \in \mathcal{U}} \mathbb{E}\left[\delta(u_{j,k}^{t,\mathbf{N}} = u)\Big|\mathbf{x}_t^{\mathbf{N}}\right] = 1$$

Using Lemma 11, we therefore conclude:

$$\mathbb{E}\left|\boldsymbol{\nu}_t^{\mathbf{N}} - \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t)\right|_1 \leq \frac{1}{N_{\mathrm{pop}}}\left(\sum_{k \in [K]} \sqrt{N_k}\right)\sqrt{|\mathcal{U}|}$$

Appendix I. Proof of Lemma 13

Note that,

$$r_k^{\mathrm{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t) = \frac{1}{N_{\mathrm{pop}}} \sum_{j=1}^{N_k} \sum_{x \in \mathcal{X}} \sum_{u \in \mathcal{U}} r_k(x, u, \boldsymbol{\mu}_t^{\mathbf{N}}, \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t)) \times \pi_k^t(x, \boldsymbol{\mu}_t^{\mathbf{N}})(u)\delta(x_{j,k}^{t,\mathbf{N}} = x)$$

$$= \frac{1}{N_{\mathrm{pop}}} \sum_{j=1}^{N_k} \sum_{u \in \mathcal{U}} r_k(x_{j,k}^{t,\mathbf{N}}, u, \boldsymbol{\mu}_t^{\mathbf{N}}, \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t)) \times \pi_k^t(x_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^{\mathbf{N}})(u)$$

We can upper bound the LHS of (33) by $J_1 + J_2$ where J_1 is defined as follows.

$$J_1 \triangleq \frac{1}{N_{\mathrm{pop}}}\mathbb{E}\left|\sum_{k \in [K]} \sum_{j=1}^{N_k} r_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\nu}_t^{\mathbf{N}}) - r_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^{\mathbf{N}}, \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t))\right|$$

$$\leq \frac{1}{N_{\mathrm{pop}}}\mathbb{E} \sum_{k \in [K]} \sum_{j=1}^{N_k}\left|r_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\nu}_t^{\mathbf{N}}) - r_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^{\mathbf{N}}, \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t))\right|$$

$$\overset{(a)}{\leq} \frac{1}{N_{\mathrm{pop}}} \sum_{k \in [K]} \sum_{j=1}^{N_k} L_R\mathbb{E}\left|\boldsymbol{\nu}_t^{\mathbf{N}} - \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t)\right|_1 \overset{(b)}{\leq} \frac{L_R}{N_{\mathrm{pop}}}\left(\sum_{k \in [K]} \sqrt{N_k}\right)\sqrt{|\mathcal{U}|}$$

Inequality (a) follows from Assumption 1(c) whereas inequality (b) follows from Lemma 12. The term J_2 is defined below.

$$J_2 \triangleq \mathbb{E} \left| \frac{1}{N_{\text{pop}}} \sum_{k \in [K]} \sum_{j=1}^{N_k} \sum_{u \in \mathcal{U}} r_k(x_{j,k}^{t,\mathbf{N}}, u, \boldsymbol{\mu}_t^{\mathbf{N}}, \nu^{\text{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t)) \times \left[\delta(u_{j,k}^{t,\mathbf{N}} = u) - \pi_k^t(x_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^{\mathbf{N}})(u) \right] \right|$$

$$\leq \frac{1}{N_{\text{pop}}} \sum_{u \in \mathcal{U}} \mathbb{E} \left| \sum_{k \in [K]} \sum_{j=1}^{N_k} r_k(x_{j,k}^{t,\mathbf{N}}, u, \boldsymbol{\mu}_t^{\mathbf{N}}, \nu^{\text{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t)) \times \left[\delta(u_{j,k}^{t,\mathbf{N}} = u) - \pi_k^t(x_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^{\mathbf{N}})(u) \right] \right|$$

Recall from Observation 1 that $u_{j,k}^{t,\mathbf{N}}$'s are independent conditioned on $\mathbf{x}_t^{\mathbf{N}}$. Therefore, $\forall u \in \mathcal{U}$, $\delta(u_{j,k}^{t,\mathbf{N}} = u)$'s are independent, conditioned on $\mathbf{x}_t^{\mathbf{N}}$. Moreover,

$$\mathbb{E} \left[\delta(u_{j,k}^{t,\mathbf{N}} = u) \Big| \mathbf{x}_t^{\mathbf{N}} \right] = \pi_k^t(x_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^{\mathbf{N}})(u), \ \forall u \in \mathcal{U},$$

$$\sum_{u \in \mathcal{U}} \mathbb{E} \left[\delta(u_{j,k}^{t,\mathbf{N}} = u) \Big| \mathbf{x}_t^{\mathbf{N}} \right] = 1$$

$$\text{and } |r_k(x, u, \boldsymbol{\mu}_t^{\mathbf{N}}, \nu^{\text{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t))| \leq M_R, \ \forall x \in \mathcal{X}, \forall u \in \mathcal{U}$$

Using Lemma 11, we therefore get,

$$J_2 \leq \frac{M_R}{\sqrt{N}} \sqrt{|\mathcal{U}|} \leq \frac{M_R}{N} \left(\sum_{k \in [K]} \sqrt{N_k} \right) \sqrt{|\mathcal{U}|}$$

This concludes the result.

Appendix J. Proof of Lemma 14

Note that the LHS of (34) can be upper bounded as follows.

$$\text{LHS} = \sum_{k \in [K]} \sum_{x \in \mathcal{X}} \mathbb{E} \left| \boldsymbol{\mu}_{t+1}^{\mathbf{N}}(x, k) - P_k^{\text{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t)(x) \right|$$

$$= \sum_{k \in [K]} \sum_{x \in \mathcal{X}} \frac{1}{N_{\text{pop}}} \mathbb{E} \left| \sum_{j=1}^{N_k} \delta(x_{j,k}^{t+1,\mathbf{N}} = x) \right.$$

$$\left. - \sum_{j=1}^{N_k} \sum_{u \in \mathcal{U}} \pi_k^t(x_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^{\mathbf{N}})(u) P_k(x_{j,k}^{t,\mathbf{N}}, u, \boldsymbol{\mu}_t^{\mathbf{N}}, \nu^{\text{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t))(x) \right|$$

$$\leq J_1 + J_2 + J_3$$

The first term, J_1 is defined as follows:

$$J_1 \triangleq \sum_{k \in [K]} \sum_{x \in \mathcal{X}} \frac{1}{N_{\text{pop}}} \mathbb{E} \left| \sum_{j=1}^{N_k} \delta(x_{j,k}^{t+1,\mathbf{N}} = x) - \sum_{j=1}^{N_k} P_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\nu}_t^{\mathbf{N}})(x) \right|$$

Recall from observation 2 that $x_{j,k}^{t+1,\mathbf{N}}$'s are independent conditional on $\mathbf{x}_t^\mathbf{N}, \mathbf{u}_t^\mathbf{N}$. Also,

$$\mathbb{E}\left[\delta\left(x_{j,k}^{t+1,\mathbf{N}} = x\right)\middle|\mathbf{x}_t^\mathbf{N}, \mathbf{u}_t^\mathbf{N}\right] = P_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^\mathbf{N}, \boldsymbol{\nu}_t^\mathbf{N})(x), \ \forall x \in \mathcal{X}$$

$$\text{and } \sum_{x\in\mathcal{X}} \mathbb{E}\left[\delta\left(x_{j,k}^{t+1,\mathbf{N}} = x\right)\middle|\mathbf{x}_t^\mathbf{N}, \mathbf{u}_t^\mathbf{N}\right] = 1$$

Applying Lemma 11, we can conclude that,

$$J_1 \le \frac{1}{N_{\text{pop}}}\left(\sum_{k\in[K]}\sqrt{N_k}\right)\sqrt{|\mathcal{X}|} \le \frac{1}{N_{\text{pop}}}\left(\sum_{k\in[K]}\sqrt{N_k}\right)\left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|}\right]$$

The second term, J_2 is defined as follows,

$$J_2 \triangleq \sum_{k\in[K]}\sum_{x\in\mathcal{X}}\frac{1}{N_{\text{pop}}}\mathbb{E}\left|\sum_{j=1}^{N_k} P_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^\mathbf{N}, \boldsymbol{\nu}_t^\mathbf{N})(x) - P_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^\mathbf{N}, \nu^{\text{MF}}(\boldsymbol{\mu}_t^\mathbf{N}, \boldsymbol{\pi}_t))(x)\right|$$

$$\le \frac{1}{N_{\text{pop}}}\sum_{k\in[K]}\sum_{j=1}^{N_k}\sum_{x\in\mathcal{X}}\mathbb{E}\left|P_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^\mathbf{N}, \boldsymbol{\nu}_t^\mathbf{N})(x) - P_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^\mathbf{N}, \nu^{\text{MF}}(\boldsymbol{\mu}_t^\mathbf{N}, \boldsymbol{\pi}_t))(x)\right|$$

$$\overset{(a)}{\le} L_P\left|\boldsymbol{\nu}_t^\mathbf{N} - \nu^{\text{MF}}(\boldsymbol{\mu}_t^\mathbf{N}, \boldsymbol{\pi}_t)\right|_1$$

$$\overset{(b)}{\le} \frac{L_P}{N_{\text{pop}}}\left(\sum_{k\in[K]}\sqrt{N_k}\right)\sqrt{|\mathcal{U}|} \le \frac{L_P}{N_{\text{pop}}}\left(\sum_{k\in[K]}\sqrt{N_k}\right)\left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|}\right]$$

Relation (a) is a consequence of Assumption 1(d) and the inequality (b) follows from Lemma 12. Finally,

$$J_3 = \sum_{k\in[K]}\sum_{x\in\mathcal{X}}\mathbb{E}\left|\frac{1}{N_{\text{pop}}}\sum_{j=1}^{N_k}\left[P_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^\mathbf{N}, \nu^{\text{MF}}(\boldsymbol{\mu}_t^\mathbf{N}, \boldsymbol{\pi}_t))(x)\right.\right.$$

$$\left.\left. - \sum_{u\in\mathcal{U}}\pi_k^t(x_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^\mathbf{N})(u) P_k(x_{j,k}^{t,\mathbf{N}}, u, \boldsymbol{\mu}_t^\mathbf{N}, \nu^{\text{MF}}(\boldsymbol{\mu}_t^\mathbf{N}, \boldsymbol{\pi}_t))(x)\right]\right|$$

$$\overset{(a)}{\le} \frac{1}{N_{\text{pop}}}\left(\sum_{k\in[K]}\sqrt{N_k}\right)\sqrt{|\mathcal{X}|} \le \frac{1}{N_{\text{pop}}}\left(\sum_{k\in[K]}\sqrt{N_k}\right)\left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|}\right]$$

Inequality (a) is a result of Lemma 11 and the facts that $\{\mathbf{u}_{j,k}^{t,\mathbf{N}}\}_{j,k}$'s are independent conditioned on $\mathbf{x}_t^\mathbf{N}$ and

$$\mathbb{E}\left[P_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^\mathbf{N}, \nu^{\text{MF}}(\boldsymbol{\mu}_t^\mathbf{N}, \boldsymbol{\pi}_t))(x)\middle|\mathbf{x}_t^\mathbf{N}\right]$$

$$= \sum_{u\in\mathcal{U}}\pi_k^t(x_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^\mathbf{N})(u) P_k(x_{j,k}^{t,\mathbf{N}}, u, \boldsymbol{\mu}_t^\mathbf{N}, \nu^{\text{MF}}(\boldsymbol{\mu}_t^\mathbf{N}, \boldsymbol{\pi}_t))(x), \ \forall x \in \mathcal{X},$$

$$\sum_{x\in\mathcal{X}}\mathbb{E}\left[P_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^\mathbf{N}, \nu^{\text{MF}}(\boldsymbol{\mu}_t^\mathbf{N}, \boldsymbol{\pi}_t))(x)\middle|\mathbf{x}_t^\mathbf{N}\right] = 1$$

This concludes the result.

Appendix K. Proof of Lemma 15

K.1 Proof of Proposition (a)

Following similar line of argument as used in the proof of Lemma 8, we obtain,

$$
\begin{aligned}
|\bar{\nu}^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\pi}}) &- \bar{\nu}^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}', \bar{\boldsymbol{\pi}})|_1 \\
&= \sum_{k \in [K]} |\bar{\nu}_k^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\pi}}) - \bar{\nu}_k^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}', \bar{\boldsymbol{\pi}})|_1 \\
&= \sum_{k \in [K]} \sum_{u \in \mathcal{U}} \left| \sum_{x \in \mathcal{X}} \bar{\boldsymbol{\mu}}(x, k)\bar{\pi}_k(x, \bar{\boldsymbol{\mu}})(u) - \sum_{x \in \mathcal{X}} \bar{\boldsymbol{\mu}}'(x, k)\bar{\pi}_k(x, \bar{\boldsymbol{\mu}}')(u) \right| \\
&\leq \sum_{k \in [K]} \sum_{u \in \mathcal{U}} \sum_{x \in \mathcal{X}} \left| \bar{\boldsymbol{\mu}}(x, k)\bar{\pi}_k(x, \bar{\boldsymbol{\mu}})(u) - \bar{\boldsymbol{\mu}}'(x, k)\bar{\pi}_k(x, \bar{\boldsymbol{\mu}}')(u) \right| \\
&\leq \sum_{k \in [K]} \sum_{x \in \mathcal{X}} |\bar{\boldsymbol{\mu}}(x, k) - \bar{\boldsymbol{\mu}}'(x, k)| \sum_{u \in \mathcal{U}} \bar{\pi}_k(x, \bar{\boldsymbol{\mu}})(u) \\
&\qquad + \sum_{k \in [K]} \sum_{x \in \mathcal{X}} \bar{\boldsymbol{\mu}}'(x, k) \sum_{u \in \mathcal{U}} \left| \bar{\pi}_k(x, \bar{\boldsymbol{\mu}})(u) - \bar{\pi}_k(x, \bar{\boldsymbol{\mu}}')(u) \right| \\
&\overset{(a)}{\leq} \sum_{k \in [K]} \sum_{x \in \mathcal{X}} |\bar{\boldsymbol{\mu}}(x, k) - \bar{\boldsymbol{\mu}}'(x, k)| + \bar{L}_Q |\bar{\boldsymbol{\mu}} - \bar{\boldsymbol{\mu}}'|_1 \sum_{k \in [K]} \sum_{x \in \mathcal{X}} \bar{\boldsymbol{\mu}}'(x, k) \\
&\overset{(b)}{=} \left(1 + K\bar{L}_Q \right) |\bar{\boldsymbol{\mu}} - \bar{\boldsymbol{\mu}}'|_1
\end{aligned}
$$

Inequality (a) follows from Assumption 4 and the fact that $\bar{\pi}_k(x, \bar{\boldsymbol{\mu}})$ is a distribution $\forall x \in \mathcal{X}$, $\forall k \in [K]$. Equality (b) uses the fact that $\bar{\boldsymbol{\mu}}'(., k)$ is a distribution $\forall k \in [K]$.

K.2 Proof of Proposition (b)

Note that,

$$
\begin{aligned}
\sum_{k \in [K]} \theta_k &|\bar{r}_k^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\pi}}) - \bar{r}_k^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}', \bar{\boldsymbol{\pi}})| \\
&\leq \sum_{k \in [K]} \sum_{x \in \mathcal{X}} \sum_{u \in \mathcal{U}} |\bar{r}_k(x, u, \bar{\boldsymbol{\mu}}, \bar{\nu}^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\pi}})) - \bar{r}_k(x, u, \bar{\boldsymbol{\mu}}', \bar{\nu}^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}', \bar{\boldsymbol{\pi}}))| \times \theta_k \bar{\boldsymbol{\mu}}(x, k)\bar{\pi}_k(x, \bar{\boldsymbol{\mu}})(u) \\
&+ \sum_{k \in [K]} \sum_{x \in \mathcal{X}} \sum_{u \in \mathcal{U}} |\bar{r}_k(x, u, \bar{\boldsymbol{\mu}}', \bar{\nu}^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}', \bar{\boldsymbol{\pi}}))| \times \theta_k |\bar{\boldsymbol{\mu}}(x, k)\bar{\pi}_k(x, \bar{\boldsymbol{\mu}})(u) - \bar{\boldsymbol{\mu}}'(x, k)\bar{\pi}_k(x, \bar{\boldsymbol{\mu}}')(u)|
\end{aligned}
$$

Utilising Assumption 2(c), and the facts that $\boldsymbol{\theta}$, $\bar{\boldsymbol{\mu}}(\cdot, k)$, $\pi_k(x, \bar{\boldsymbol{\mu}})$ are probability distributions $\forall k \in [K]$, $\forall x \in \mathcal{X}$, the first term can be upper bounded by the following expression,

$$
\bar{L}_R \left(|\bar{\boldsymbol{\mu}} - \bar{\boldsymbol{\mu}}'|_1 + |\bar{\nu}^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\pi}}) - \bar{\nu}^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}', \bar{\boldsymbol{\pi}})|_1 \right) \leq \bar{L}_R \left[1 + (1 + K\bar{L}_Q) \right] |\bar{\boldsymbol{\mu}} - \bar{\boldsymbol{\mu}}'|_1
$$

Proposition (a) is used to derive the above inequality. Applying assumption 2(b), the second term can be upper bounded by the following quantity.

$$\bar{M}_R \sum_{k\in[K]} \sum_{x\in\mathcal{X}} \sum_{u\in\mathcal{U}} \theta_k |\bar{\boldsymbol{\mu}}(x,k)\bar{\pi}_k(x,\bar{\boldsymbol{\mu}})(u) - \bar{\boldsymbol{\mu}}'(x,k)\bar{\pi}_k(x,\bar{\boldsymbol{\mu}}')(u)|$$

$$\leq \bar{M}_R \sum_{k\in[K]} \sum_{x\in\mathcal{X}} \theta_k |\bar{\boldsymbol{\mu}}(x,k) - \bar{\boldsymbol{\mu}}'(x,k)| \sum_{u\in\mathcal{U}} \bar{\pi}_k(x,\bar{\boldsymbol{\mu}})(u)$$

$$+ \bar{M}_R \sum_{k\in[K]} \sum_{x\in\mathcal{X}} \theta_k \bar{\boldsymbol{\mu}}'(x,k) \sum_{u\in\mathcal{U}} \left| \bar{\pi}_k(x,\bar{\boldsymbol{\mu}})(u) - \bar{\pi}_k(x,\bar{\boldsymbol{\mu}}')(u) \right|$$

$$\overset{(a)}{\leq} \bar{M}_R \sum_{k\in[K]} \sum_{x\in\mathcal{X}} \theta_k |\bar{\boldsymbol{\mu}}(x,k) - \bar{\boldsymbol{\mu}}'(x,k)| + \bar{M}_R \bar{L}_Q |\bar{\boldsymbol{\mu}} - \bar{\boldsymbol{\mu}}'|_1 \sum_{k\in[K]} \sum_{x\in\mathcal{X}} \theta_k \bar{\boldsymbol{\mu}}'(x,k)$$

$$\overset{(b)}{\leq} \bar{M}_R \sum_{k\in[K]} \sum_{x\in\mathcal{X}} |\bar{\boldsymbol{\mu}}(x,k) - \bar{\boldsymbol{\mu}}'(x,k)| + \bar{M}_R \bar{L}_Q |\bar{\boldsymbol{\mu}} - \bar{\boldsymbol{\mu}}'|_1 \sum_{k\in[K]} \theta_k$$

$$\overset{(c)}{=} \bar{M}_R (1 + \bar{L}_Q) |\bar{\boldsymbol{\mu}} - \bar{\boldsymbol{\mu}}'|_1$$

Inequality (a) follows from Assumption 4 and the fact that $\bar{\pi}_k(x,\bar{\boldsymbol{\mu}})$ is a distribution $\forall x \in \mathcal{X}, \forall k \in [K]$ while result (b) is derived from the fact that $\bar{\boldsymbol{\mu}}'(\cdot,k)$ is a distribution and $\theta_k \leq 1, \forall k \in [K]$. Finally, equality (c) holds because $\boldsymbol{\theta}$ is a distribution. This proves the proposition.

K.3 Proof of Proposition (c)

Note that,

$$|\bar{P}^{\mathrm{MF}}(\bar{\boldsymbol{\mu}},\bar{\boldsymbol{\pi}}) - \bar{P}^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}',\bar{\boldsymbol{\pi}})|_1 = \sum_{k\in[K]} |\bar{P}_k^{\mathrm{MF}}(\bar{\boldsymbol{\mu}},\bar{\boldsymbol{\pi}}) - \bar{P}_k^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}',\bar{\boldsymbol{\pi}})|_1$$

$$\leq \sum_{k\in[K]} \sum_{x\in\mathcal{X}} \sum_{u\in\mathcal{U}} |\bar{P}_k(x,u,\bar{\boldsymbol{\mu}},\bar{\nu}^{\mathrm{MF}}(\bar{\boldsymbol{\mu}},\bar{\boldsymbol{\pi}})) - \bar{P}_k(x,u,\bar{\boldsymbol{\mu}}',\bar{\nu}^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}',\bar{\boldsymbol{\pi}}))|_1 \times \bar{\boldsymbol{\mu}}(x,k)\bar{\pi}_k(x,\bar{\boldsymbol{\mu}})(u)$$

$$+ \sum_{k\in[K]} \sum_{x\in\mathcal{X}} \sum_{u\in\mathcal{U}} |\bar{P}_k(x,u,\bar{\boldsymbol{\mu}}',\bar{\nu}^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}',\bar{\boldsymbol{\pi}}))|_1 \times |\bar{\boldsymbol{\mu}}(x,k)\bar{\pi}_k(x,\bar{\boldsymbol{\mu}})(u) - \bar{\boldsymbol{\mu}}'(x,k)\bar{\pi}_k(x,\bar{\boldsymbol{\mu}}')(u)|$$

Using Assumption 2(d) and the facts that $\bar{\boldsymbol{\mu}}(\cdot,k)$, $\bar{\pi}_k(x,\bar{\boldsymbol{\mu}})$ are probability distributions $\forall x \in \mathcal{X}, \forall k \in [K]$, the first term can be upper bounded by the following expression,

$$K\bar{L}_P \left(|\bar{\boldsymbol{\mu}} - \bar{\boldsymbol{\mu}}'|_1 + |\bar{\nu}^{\mathrm{MF}}(\bar{\boldsymbol{\mu}},\bar{\boldsymbol{\pi}}) - \bar{\nu}^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}',\bar{\boldsymbol{\pi}})|_1 \right) \leq K\bar{L}_P \left[1 + (1 + K\bar{L}_Q) \right] |\bar{\boldsymbol{\mu}} - \bar{\boldsymbol{\mu}}'|_1$$

Proposition (a) is applied to derive the above inequality. Note that, $|\bar{P}_k(x,u,\bar{\boldsymbol{\mu}}',\bar{\nu}(\bar{\boldsymbol{\mu}}',\bar{\boldsymbol{\pi}}))|_1 = 1$. Therefore, the second term can be upper bounded by the following quantity.

$$\sum_{k\in[K]} \sum_{x\in\mathcal{X}} \sum_{u\in\mathcal{U}} |\bar{\boldsymbol{\mu}}(x,k)\bar{\pi}_k(x,\bar{\boldsymbol{\mu}})(u) - \bar{\boldsymbol{\mu}}'(x,k)\bar{\pi}_k(x,\bar{\boldsymbol{\mu}}')(u)| \overset{(a)}{\leq} (1 + K\bar{L}_Q)|\bar{\boldsymbol{\mu}} - \bar{\boldsymbol{\mu}}'|_1$$

Inequality (a) can be established by following identical arguments as used in Proposition (a). This concludes the result.

Appendix L. Proof of Lemma 16

L.1 Proof of Proposition (a)

Using the definition of L_1-norm, we get:

$$
\mathbb{E}\left|\bar{\boldsymbol{\nu}}_t^{\mathbf{N}} - \bar{\nu}^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\boldsymbol{\pi}}_t)\right|_1 = \sum_{k\in[K]}\sum_{u\in\mathcal{U}}\mathbb{E}\left|\bar{\boldsymbol{\nu}}_t^{\mathbf{N}}(u,k) - \bar{\nu}^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\boldsymbol{\pi}}_t)(u,k)\right|
$$

$$
= \sum_{k\in[K]}\sum_{u\in\mathcal{U}}\frac{1}{N_k}\mathbb{E}\left|\sum_{j=1}^{N_k}\delta(u_{j,k}^{t,\mathbf{N}}=u) - \sum_{j=1}^{N_k}\pi_k^t(x_{j,k}^{t,\mathbf{N}}, \bar{\boldsymbol{\mu}}_t^{\mathbf{N}})(u)\right|
$$

$$
\overset{(a)}{\leq} \left(\sum_{k\in[K]}\frac{1}{\sqrt{N_k}}\right)\sqrt{|\mathcal{U}|}
$$

Inequality (a) follows from Lemma 11. This concludes the proposition.

L.2 Proof of Proposition (b)

Note that,

$$
\theta_k \bar{r}_k^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\boldsymbol{\pi}}_t)
$$

$$
= \left(\frac{N_k}{N_{\mathrm{pop}}}\right)\frac{1}{N_k}\sum_{j=1}^{N_k}\sum_{x\in\mathcal{X}}\sum_{u\in\mathcal{U}}\bar{r}_k(x,u,\bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\nu}^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\boldsymbol{\pi}}_t))\times\bar{\pi}_k^t(x,\bar{\boldsymbol{\mu}}_t^{\mathbf{N}})(u)\delta(x_{j,k}^{t,\mathbf{N}}=x)
$$

$$
= \frac{1}{N_{\mathrm{pop}}}\sum_{j=1}^{N_k}\sum_{u\in\mathcal{U}}\bar{r}_k(x_{j,k}^{t,\mathbf{N}}, u, \bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\nu}^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\boldsymbol{\pi}}_t))\times\bar{\pi}_k^t(x_{j,k}^{t,\mathbf{N}}, \bar{\boldsymbol{\mu}}_t^{\mathbf{N}})(u)
$$

We can upper bound the LHS of (45) by $J_1 + J_2$ where J_1 is defined as follows.

$$
J_1 \triangleq \frac{1}{N_{\mathrm{pop}}}\mathbb{E}\left|\sum_{k\in[K]}\sum_{j=1}^{N_k}\bar{r}_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\boldsymbol{\nu}}_t^{\mathbf{N}}) - \bar{r}_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\nu}^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\boldsymbol{\pi}}_t))\right|
$$

$$
\leq \frac{1}{N_{\mathrm{pop}}}\sum_{k\in[K]}\sum_{j=1}^{N_k}\mathbb{E}\left|\bar{r}_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\boldsymbol{\nu}}_t^{\mathbf{N}}) - \bar{r}_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\nu}^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\boldsymbol{\pi}}_t))\right|
$$

$$
\overset{(a)}{\leq} \frac{1}{N_{\mathrm{pop}}}\sum_{k\in[K]}\sum_{j=1}^{N_k}\bar{L}_R\mathbb{E}\left|\bar{\boldsymbol{\nu}}_t^{\mathbf{N}} - \bar{\nu}^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\boldsymbol{\pi}}_t)\right|_1
$$

$$
\overset{(b)}{\leq} \bar{L}_R\left(\sum_{k\in[K]}\frac{1}{\sqrt{N_k}}\right)\sqrt{|\mathcal{U}|}
$$

Inequality (a) follows from Assumption 2(c) while inequality (b) follows from Proposition (a). The term J_2 is defined as

$$J_2 \triangleq \mathbb{E} \left| \frac{1}{N_{\text{pop}}} \sum_{k \in [K]} \sum_{j=1}^{N_k} \sum_{u \in \mathcal{U}} \bar{r}_k(x_{j,k}^{t,\mathbf{N}}, u, \bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\nu}^{\text{MF}}(\bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\boldsymbol{\pi}}_t)) \times \left[\delta(u_{j,k}^{t,\mathbf{N}} = u) - \bar{\pi}_k^t(x_{j,k}^{t,\mathbf{N}}, \bar{\boldsymbol{\mu}}_t^{\mathbf{N}})(u) \right] \right|$$

$$\leq \frac{1}{N_{\text{pop}}} \sum_{u \in \mathcal{U}} \mathbb{E} \left| \sum_{k \in [K]} \sum_{j=1}^{N_k} \bar{r}_k(x_{j,k}^{t,\mathbf{N}}, u, \bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\nu}^{\text{MF}}(\bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\boldsymbol{\pi}}_t)) \times \left[\delta(u_{j,k}^{t,\mathbf{N}} = u) - \bar{\pi}_k^t(x_{j,k}^{t,\mathbf{N}}, \bar{\boldsymbol{\mu}}_t^{\mathbf{N}})(u) \right] \right|$$

Using similar argument as used in Lemma 13, we therefore get,

$$J_2 \leq \frac{\bar{M}_R}{\sqrt{N}} \sqrt{|\mathcal{U}|} \leq \bar{M}_R \left(\sum_{k \in [K]} \frac{1}{\sqrt{N_k}} \right) \sqrt{|\mathcal{U}|}$$

This concludes the result.

L.3 Proof of Proposition (c)

Note that the LHS of (46) can be upper bounded by the following quantity..

$$\sum_{k \in [K]} \sum_{x \in \mathcal{X}} \mathbb{E} \left| \bar{\boldsymbol{\mu}}_{t+1}^{\mathbf{N}}(x, k) - \bar{P}_k^{\text{MF}}(\bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\boldsymbol{\pi}}_t)(x) \right|$$

$$= \sum_{k \in [K]} \sum_{x \in \mathcal{X}} \frac{1}{N_k} \mathbb{E} \left| \sum_{j=1}^{N_k} \delta(x_{j,k}^{t+1,\mathbf{N}} = x) \right.$$

$$\left. - \sum_{j=1}^{N_k} \sum_{u \in \mathcal{U}} \bar{\pi}_k^t(x_{j,k}^{t,\mathbf{N}}, \bar{\boldsymbol{\mu}}_t^{\mathbf{N}})(u) \bar{P}_k(x_{j,k}^{t,\mathbf{N}}, u, \bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\nu}^{\text{MF}}(\bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\boldsymbol{\pi}}_t))(x) \right| \leq J_1 + J_2 + J_3$$

The first term, J_1 is defined as follows:

$$J_1 \triangleq \sum_{k \in [K]} \sum_{x \in \mathcal{X}} \frac{1}{N_k} \mathbb{E} \left| \sum_{j=1}^{N_k} \delta(x_{j,k}^{t+1,\mathbf{N}} = x) - \sum_{j=1}^{N_k} \bar{P}_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\boldsymbol{\nu}}_t^{\mathbf{N}})(x) \right|$$

Using similar argument as used in Lemma 14 to bound J_1, we get,

$$J_1 \leq \left(\sum_{k \in [K]} \frac{1}{\sqrt{N_k}} \right) \sqrt{|\mathcal{X}|} \leq \left(\sum_{k \in [K]} \frac{1}{\sqrt{N_k}} \right) \left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|} \right]$$

The second term, J_2 is defined as follows,

$$J_2 \triangleq \sum_{k \in [K]} \sum_{x \in \mathcal{X}} \frac{1}{N_k} \mathbb{E} \left| \sum_{j=1}^{N_k} \bar{P}_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\boldsymbol{\nu}}_t^{\mathbf{N}})(x) - \sum_{j=1}^{N_k} \bar{P}_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\nu}^{\text{MF}}(\bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\boldsymbol{\pi}}_t))(x) \right|$$

$$\overset{(a)}{\leq} K \bar{L}_P \left| \bar{\boldsymbol{\nu}}_t^{\mathbf{N}} - \bar{\nu}^{\text{MF}}(\bar{\boldsymbol{\mu}}_t^{\mathbf{N}}, \bar{\boldsymbol{\pi}}_t) \right|_1$$

$$\overset{(b)}{\leq} K \bar{L}_P \left(\sum_{k \in [K]} \frac{1}{\sqrt{N_k}} \right) \sqrt{|\mathcal{U}|} \leq K \bar{L}_P \left(\sum_{k \in [K]} \frac{1}{\sqrt{N_k}} \right) \left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|} \right]$$

36

Relation (a) is a result of Assumption 2(d) and the inequality (b) follows from Proposition (a). Finally,

$$
\begin{aligned}
J_3 &= \sum_{k\in[K]}\sum_{x\in\mathcal{X}}\mathbb{E}\left|\frac{1}{N_k}\sum_{j=1}^{N_k}\left[\bar{P}_k(x_{j,k}^{t,\mathbf{N}},u_{j,k}^{t,\mathbf{N}},\bar{\boldsymbol{\mu}}_t^{\mathbf{N}},\bar{\nu}^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}_t^{\mathbf{N}},\bar{\boldsymbol{\pi}}_t))(x)\right.\right. \\
&\qquad\qquad\left.\left. - \sum_{u\in\mathcal{U}}\pi_k^t(x_{j,k}^{t,\mathbf{N}},\bar{\boldsymbol{\mu}}_t^{\mathbf{N}})(u)\bar{P}_k(x_{j,k}^{t,\mathbf{N}},u,\bar{\boldsymbol{\mu}}_t^{\mathbf{N}},\bar{\nu}^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}_t^{\mathbf{N}},\bar{\boldsymbol{\pi}}_t))(x)\right]\right| \\
&\overset{(a)}{\leq} \left(\sum_{k\in[K]}\frac{1}{\sqrt{N_k}}\right)\sqrt{|\mathcal{X}|} \leq \left(\sum_{k\in[K]}\frac{1}{\sqrt{N_k}}\right)\left[\sqrt{|\mathcal{X}|}+\sqrt{|\mathcal{U}|}\right]
\end{aligned}
$$

Inequality (a) is a result of Lemma 11. This concludes the result.

Appendix M. Proof of Lemma 17

M.1 Proof of Proposition (a)

The following chain of inequalities hold true.

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

Result (a) follows from Lipschitz continuity of π_k^t and the fact that $\pi_k(x,\boldsymbol{\mu})$ is a probability distribution. Finally, inequality (b) uses the fact that $\boldsymbol{\mu}'$ is a distribution. This concludes the result.

M.2 Proof of Proposition (b)

Note that,

$$\sum_{k\in[K]} |r_k^{\mathrm{MF}}(\boldsymbol{\mu}, \boldsymbol{\pi}) - r_k^{\mathrm{MF}}(\boldsymbol{\mu}', \boldsymbol{\pi})|$$

$$\leq \sum_{k\in[K]} \sum_{x\in\mathcal{X}} \sum_{u\in\mathcal{U}} |r_k(x, u, \boldsymbol{\mu}[\mathcal{X}], \nu(\boldsymbol{\mu}, \boldsymbol{\pi})[\mathcal{U}]) - r_k(x, u, \boldsymbol{\mu}'[\mathcal{X}], \nu(\boldsymbol{\mu}', \boldsymbol{\pi})[\mathcal{U}])| \times \boldsymbol{\mu}(x, k)\pi_k(x, \boldsymbol{\mu}[\mathcal{X}])(u)$$

$$+ \sum_{k\in[K]} \sum_{x\in\mathcal{X}} \sum_{u\in\mathcal{U}} |r_k(x, u, \boldsymbol{\mu}'[\mathcal{X}], \nu(\boldsymbol{\mu}', \boldsymbol{\pi})[\mathcal{U}])| \times |\boldsymbol{\mu}(x, k)\pi_k(x, \boldsymbol{\mu}[\mathcal{X}])(u) - \boldsymbol{\mu}'(x, k)\pi_k(x, \boldsymbol{\mu}'[\mathcal{X}])(u)|$$

Using the Lipschitz continuity of r_k, and the facts that $\boldsymbol{\mu}$, $\pi_k(x, \boldsymbol{\mu})$ are distributions, the first term can be upper bounded by the following expression,

$$L_R \left(|\boldsymbol{\mu}[\mathcal{X}] - \boldsymbol{\mu}'[\mathcal{X}]|_1 + |\nu^{\mathrm{MF}}(\boldsymbol{\mu}, \boldsymbol{\pi})[\mathcal{U}] - \nu^{\mathrm{MF}}(\boldsymbol{\mu}', \boldsymbol{\pi})[\mathcal{U}]|_1 \right)$$
$$\leq L_R |\boldsymbol{\mu} - \boldsymbol{\mu}'|_1 + L_R(1 + L_Q)|\boldsymbol{\mu}[\mathcal{X}] - \boldsymbol{\mu}'[\mathcal{X}]|_1$$

Proposition (a) is used to derive the above inequality. Utilising similar logic as used in Proposition (a), we can upper bound the second term by the following quantity:

$$M_R |\boldsymbol{\mu} - \boldsymbol{\mu}'|_1 + M_R L_Q |\boldsymbol{\mu}[\mathcal{X}] - \boldsymbol{\mu}'[\mathcal{X}]|_1$$

M.3 Proof of Proposition (c)

The proof is similar to that of Proposition (b). Note that,

$$|P^{\mathrm{MF}}(\boldsymbol{\mu}, \boldsymbol{\pi}) - P^{\mathrm{MF}}(\boldsymbol{\mu}', \boldsymbol{\pi})|_1 = \sum_{k\in[K]} |P_k^{\mathrm{MF}}(\boldsymbol{\mu}, \boldsymbol{\pi}) - P_k^{\mathrm{MF}}(\boldsymbol{\mu}', \boldsymbol{\pi})|_1$$

$$\leq \sum_{k\in[K]} \sum_{x\in\mathcal{X}} \sum_{u\in\mathcal{U}} |P_k(x, u, \boldsymbol{\mu}[\mathcal{X}], \nu^{\mathrm{MF}}(\boldsymbol{\mu}, \boldsymbol{\pi})[\mathcal{U}]) - P_k(x, u, \boldsymbol{\mu}'[\mathcal{X}], \nu^{\mathrm{MF}}(\boldsymbol{\mu}', \boldsymbol{\pi})[\mathcal{U}])|_1 \boldsymbol{\mu}(x, k)\pi_k(x, \boldsymbol{\mu}[\mathcal{X}])(u)$$

$$+ \sum_{k\in[K]} \sum_{x\in\mathcal{X}} \sum_{u\in\mathcal{U}} |P_k(x, u, \boldsymbol{\mu}'[\mathcal{X}], \nu^{\mathrm{MF}}(\boldsymbol{\mu}', \boldsymbol{\pi})[\mathcal{U}])|_1 \times |\boldsymbol{\mu}(x, k)\pi_k(x, \boldsymbol{\mu}[\mathcal{X}])(u) - \boldsymbol{\mu}'(x, k)\pi_k(x, \boldsymbol{\mu}'[\mathcal{X}])(u)|$$

Using the Lipschitz continuity of P_k and the facts that $\boldsymbol{\mu}$, $\pi_k(x, \boldsymbol{\mu})$ are distributions, the first term can be upper bounded by the following expression,

$$L_P \left(|\boldsymbol{\mu}[\mathcal{X}] - \boldsymbol{\mu}'[\mathcal{X}]|_1 + |\nu^{\mathrm{MF}}(\boldsymbol{\mu}, \boldsymbol{\pi})[\mathcal{U}] - \nu^{\mathrm{MF}}(\boldsymbol{\mu}', \boldsymbol{\pi})[\mathcal{U}]|_1 \right)$$
$$\leq L_P |\boldsymbol{\mu} - \boldsymbol{\mu}'|_1 + L_P(1 + L_Q)|\boldsymbol{\mu}[\mathcal{X}] - \boldsymbol{\mu}'[\mathcal{X}]|_1$$

Proposition (a) is used to derive the above inequality. Utilising similar logic as used in Proposition (a), and the fact that $|P_k(x, u, \boldsymbol{\mu}'[\mathcal{X}], \nu^{\mathrm{MF}}(\boldsymbol{\mu}', \boldsymbol{\pi})[\mathcal{U}])|_1 = 1, \forall x \in \mathcal{X}, \forall u \in \mathcal{U}$, we can bound the second term by the following quantity:

$$|\boldsymbol{\mu} - \boldsymbol{\mu}'|_1 + L_Q |\boldsymbol{\mu}[\mathcal{X}] - \boldsymbol{\mu}'[\mathcal{X}]|_1$$

This concludes the result.

Appendix N. Proof of Lemma 18

N.1 Proof of Proposition (a)

Using the definition of L_1-norm, we get:

$$\mathbb{E}\left|\boldsymbol{\nu}_t^{\mathbf{N}}[\mathcal{U}] - \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t)[\mathcal{U}]\right|_1$$

$$= \sum_{u\in\mathcal{U}}\mathbb{E}\left|\sum_{k\in[K]}\boldsymbol{\nu}_t^{\mathbf{N}}(u,k) - \sum_{k\in[K]}\nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t)(u,k)\right|$$

$$= \frac{1}{N_{\mathrm{pop}}}\sum_{u\in\mathcal{U}}\mathbb{E}\left|\sum_{k\in[K]}\sum_{j=1}^{N_k}\delta(u_{j,k}^{t,\mathbf{N}} = u) - \sum_{k\in[K]}\sum_{j=1}^{N_k}\pi_k^t(x_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^{\mathbf{N}})(u)\right|$$

Using Lemma 11, we conclude the proposition.

N.2 Proof of Proposition (b)

Using similar argument as used in Lemma 13, we can bound the LHS of (53) by $J_1 + J_2$ where

$$J_1 \triangleq \frac{1}{N_{\mathrm{pop}}}\mathbb{E}\left|\sum_{k\in[K]}\sum_{j=1}^{N_k} r_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^{\mathbf{N}}[\mathcal{X}], \boldsymbol{\nu}_t^{\mathbf{N}}[\mathcal{U}]) - r_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^{\mathbf{N}}[\mathcal{X}], \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t)[\mathcal{U}])\right|$$

$$\leq \frac{1}{N_{\mathrm{pop}}}\sum_{k\in[K]}\sum_{j=1}^{N_k}\mathbb{E}\left|r_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^{\mathbf{N}}[\mathcal{X}], \boldsymbol{\nu}_t^{\mathbf{N}}[\mathcal{U}]) - r_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^{\mathbf{N}}[\mathcal{X}], \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t)[\mathcal{U}])\right|$$

$$\overset{(a)}{\leq} L_R\mathbb{E}\left|\boldsymbol{\nu}_t^{\mathbf{N}}[\mathcal{U}] - \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t)[\mathcal{U}]\right|_1$$

$$\overset{(b)}{\leq} \frac{L_R}{\sqrt{N_{\mathrm{pop}}}}\sqrt{|\mathcal{U}|}$$

Inequality (a) follows from the Lipschitz continuity of r_k while (b) is a consequence of proposition (a). The second term, J_2 is as follows,

$$J_2 \triangleq \frac{1}{N_{\mathrm{pop}}}\mathbb{E}\left|\sum_{k\in[K]}\sum_{j=1}^{N_k}\sum_{u\in\mathcal{U}} r_k(x_{j,k}^{t,\mathbf{N}}, u, \boldsymbol{\mu}_t^{\mathbf{N}}[\mathcal{X}], \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t)[\mathcal{U}])\right.$$

$$\left.\times \left[\delta(u_{j,k}^{t,\mathbf{N}} = u) - \pi_k^t(x_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^{\mathbf{N}}[\mathcal{X}])(u)\right]\right|$$

$$\overset{(a)}{\leq} \frac{M_R}{\sqrt{N_{\mathrm{pop}}}}\sqrt{|\mathcal{U}|}$$

Inequality (a) can be proved using Lemma 11.

N.3 Proof of Proposition (c)

Note that the LHS of (54) can be upper bounded as follows,

$$
\mathbb{E}\left|\boldsymbol{\mu}_{t+1}^{\mathbf{N}}[\mathcal{X}] - P^{\mathrm{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t)[\mathcal{X}]\right|_1
$$

$$
= \sum_{x \in \mathcal{X}} \mathbb{E}\left|\sum_{k \in [K]} \boldsymbol{\mu}_{t+1}^{\mathbf{N}}(x, k) - \sum_{k \in [K]} P_k^{\mathrm{MF}}\left(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t\right)(x)\right|
$$

$$
= \frac{1}{N_{\mathrm{pop}}} \sum_{x \in \mathcal{X}} \mathbb{E}\left|\sum_{k \in [K]} \sum_{j=1}^{N_k}\left\{\delta(x_{j,k}^{t+1,\mathbf{N}} = x)\right.\right.
$$

$$
\left.\left. - \sum_{u \in \mathcal{U}} \pi_k^t(x_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^{\mathbf{N}}[\mathcal{X}])(u) P_k(x_{j,k}^{t,\mathbf{N}}, u, \boldsymbol{\mu}_t^{\mathbf{N}}[\mathcal{X}], \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t)[\mathcal{U}])(x)\right\}\right|
$$

$$
\leq J_1 + J_2 + J_3
$$

The first term is defined as:

$$
J_1 \triangleq \frac{1}{N_{\mathrm{pop}}} \sum_{x \in \mathcal{X}} \mathbb{E}\left|\sum_{k \in [K]} \sum_{j=1}^{N_k} \delta(x_{j,k}^{t+1,\mathbf{N}} = x) - \sum_{k \in [K]} \sum_{j=1}^{N_k} P_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}^{t,\mathbf{N}}[\mathcal{X}], \boldsymbol{\nu}^{t,\mathbf{N}}[\mathcal{U}])(x)\right|
$$

Applying Lemma 11, we can conclude that,

$$
J_1 \leq \frac{1}{\sqrt{N_{\mathrm{pop}}}} \sqrt{|\mathcal{X}|} \leq \frac{1}{\sqrt{N_{\mathrm{pop}}}}\left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|}\right]
$$

The second term, J_2 is as follows,

$$
J_2 \triangleq \frac{1}{N_{\mathrm{pop}}} \sum_{x \in \mathcal{X}} \mathbb{E}\left|\sum_{k \in [K]} \sum_{j=1}^{N_k}\left\{\tilde{P}_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^{\mathbf{N}}[\mathcal{X}], \boldsymbol{\nu}_t^{\mathbf{N}}[\mathcal{U}])(x)\right.\right.
$$

$$
\left.\left. - \tilde{P}_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^{\mathbf{N}}[\mathcal{X}], \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t)[\mathcal{U}])(x)\right\}\right|
$$

$$
\leq \frac{1}{N_{\mathrm{pop}}} \sum_{k \in [K]} \sum_{j=1}^{N_k} \sum_{x \in \mathcal{X}} \mathbb{E}\left|\tilde{P}_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^{\mathbf{N}}[\mathcal{X}], \boldsymbol{\nu}_t^{\mathbf{N}}[\mathcal{U}])(x)\right.
$$

$$
\left. - \tilde{P}_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^{\mathbf{N}}[\mathcal{X}], \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t)[\mathcal{U}])(x)\right|
$$

$$
\overset{(a)}{\leq} L_P\left|\boldsymbol{\nu}_t^{\mathbf{N}}[\mathcal{U}] - \nu^{\mathrm{MF}}(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t)[\mathcal{U}]\right|_1
$$

$$
\overset{(b)}{\leq} \frac{L_P}{\sqrt{N_{\mathrm{pop}}}} \sqrt{|\mathcal{U}|}
$$

$$
\leq \frac{L_P}{\sqrt{N_{\mathrm{pop}}}}\left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|}\right]
$$

Inequality (a) is due to Lipschitz continuity of P_k and (b) follows from Proposition (a). Finally,

$$J_3 = \sum_{x \in \mathcal{X}} \mathbb{E} \left| \frac{1}{N_{\text{pop}}} \sum_{k \in [K]} \sum_{j=1}^{N_k} \left[P_k(x_{j,k}^{t,\mathbf{N}}, u_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^{\mathbf{N}}[\mathcal{X}], \nu(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t)[\mathcal{U}])(x) - \right. \right.$$

$$\left. \left. - \sum_{u \in \mathcal{U}} \pi_k^t(x_{j,k}^{t,\mathbf{N}}, \boldsymbol{\mu}_t^{\mathbf{N}}[\mathcal{X}])(u) P_k(x_{j,k}^{t,\mathbf{N}}, u, \boldsymbol{\mu}_t^{\mathbf{N}}[\mathcal{X}], \nu(\boldsymbol{\mu}_t^{\mathbf{N}}, \boldsymbol{\pi}_t)[\mathcal{U}])(x) \right] \right|$$

Applying Lemma 11, we finally obtain, $J_3 \leq \frac{1}{\sqrt{N_{\text{pop}}}} \sqrt{|\mathcal{X}|} \leq \frac{1}{\sqrt{N_{\text{pop}}}} \left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|} \right]$.

Appendix O. Proof of Theorem 5

Note that the LHS of (25) can be upper bounded as,

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

Using Theorem 1, the first term can be bounded by $C' \left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|} \right] \frac{1}{N_{\text{pop}}} \sum_{k \in [K]} \sqrt{N_k}$ for some constant C'. Using Lemma 4, the second term be bounded by $\sqrt{\epsilon_{\text{bias}}}/(1-\gamma) + \epsilon$ with a sample complexity $\mathcal{O}(\epsilon^{-3})$. Choosing $\epsilon = C' \left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|} \right] \frac{1}{N_{\text{pop}}} \sum_{k \in [K]} \sqrt{N_k}$, we obtain the result as in the statement of the theorem.

Appendix P. Loose Bounds

In this section, we shall demonstrate that one can derive loose bounds for multi-agent systems satisfying Assumption 1, 3 using Theorem 2. Similarly, loose bounds for systems satisfying Assumption 2 and 4 can be derived using Theorem 1.

P.1 Loose Bound Using Theorem 1

Consider a multi-agent system satisfying Assumptions 2 and 4. We shall use the notations of Theorem 2. Let, $\boldsymbol{\theta} \triangleq \{\theta_k\}_{k \in [K]}$ be prior probabilities of different classes. If $\bar{r}_k$'s and $\bar{P}_k$'s are given reward and transition functions of the system, then one can define r_k's and P_k's such that, $\forall x \in \mathcal{X}, \forall u \in \mathcal{U}, \forall \bar{\boldsymbol{\mu}} \in \mathcal{P}^K(\mathcal{X}), \forall \bar{\boldsymbol{\nu}} \in \mathcal{P}^K(\mathcal{U})$ and $\forall k \in [K]$,

$$\bar{r}_k(x, u, \bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\nu}}) = r_k(x, u, \boldsymbol{\mu}, \boldsymbol{\nu}),$$
$$\bar{P}_k(x, u, \bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\nu}}) = P_k(x, u, \boldsymbol{\mu}, \boldsymbol{\nu})$$

where $\boldsymbol{\mu}, \boldsymbol{\nu}$ are uniquely defined as, $\boldsymbol{\mu} \triangleq \{\theta_k \bar{\boldsymbol{\mu}}(., k)\}_{k \in [K]}$ and $\boldsymbol{\nu} \triangleq \{\theta_k \bar{\boldsymbol{\nu}}(., k)\}_{k \in [K]}$. Clearly, $\boldsymbol{\mu} \in \mathcal{P}_{\boldsymbol{\theta}}(\mathcal{X} \times [K])$ where $\mathcal{P}_{\boldsymbol{\theta}}(\mathcal{X} \times [K])$ is the collection of distributions over $\mathcal{X} \times [K]$ such that the marginal distribution over $[K]$ derived from each of its elements is $\boldsymbol{\theta}$. Similarly, $\boldsymbol{\nu} \in \mathcal{P}_{\boldsymbol{\theta}}(\mathcal{U} \times [K])$. Also, for every policy $\bar{\boldsymbol{\pi}} \triangleq \{(\bar{\pi}_k^t)_{k \in [K]}\}_{t \in \{0,1,\cdots\}}$, one can define $\boldsymbol{\pi} \triangleq \{(\pi_k^t)_{k \in [K]}\}_{t \in \{0,1,\cdots\}}$ such that, $\forall x \in \mathcal{X}, \forall \bar{\boldsymbol{\mu}} \in \mathcal{P}^K(\mathcal{X})$ and $\forall k \in [K]$,

$$\bar{\pi}_k^t(x, \bar{\boldsymbol{\mu}}) = \pi_k^t(x, \boldsymbol{\mu})$$

Note that, the following inequality holds $\forall \boldsymbol{\mu}, \boldsymbol{\mu}' \in \mathcal{P}_{\boldsymbol{\theta}}(\mathcal{X} \times [K])$, $\forall \boldsymbol{\nu}, \boldsymbol{\nu}' \in \mathcal{P}_{\boldsymbol{\theta}}(\mathcal{U} \times [K])$, $\forall x \in \mathcal{X}$, $\forall u \in \mathcal{U}$, and $\forall k \in [K]$

$$
\begin{aligned}
|r_k(x, u, \boldsymbol{\mu}, \boldsymbol{\nu}) - r_k(x, u, \boldsymbol{\mu}', \boldsymbol{\nu}')| &= |\bar{r}_k(x, u, \bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\nu}}) - \bar{r}_k(x, u, \bar{\boldsymbol{\mu}}', \bar{\boldsymbol{\nu}}')| \\
&\leq \bar{L}_R \left[|\bar{\boldsymbol{\mu}} - \bar{\boldsymbol{\mu}}'|_1 + |\bar{\boldsymbol{\nu}} - \bar{\boldsymbol{\nu}}'|_1 \right] \\
&= \bar{L}_R \sum_{k \in [K]} \theta_k^{-1} \left[|\boldsymbol{\mu}(., k) - \boldsymbol{\mu}'(., k)|_1 + |\boldsymbol{\nu}(., k) - \boldsymbol{\nu}'(., k)|_1 \right] \quad (63) \\
&\leq \bar{L}_R \boldsymbol{\theta}_M^{-1} \left[|\boldsymbol{\mu} - \boldsymbol{\mu}'|_1 + |\boldsymbol{\nu} - \boldsymbol{\nu}'|_1 \right]
\end{aligned}
$$

where we have, $\boldsymbol{\theta}_M^{-1} \triangleq \max\{\theta_k^{-1}\}_{k \in [K]}$, $\bar{\boldsymbol{\mu}} \triangleq \{\theta_k^{-1} \boldsymbol{\mu}(., k)\}_{k \in [K]}$, $\bar{\boldsymbol{\mu}}' \triangleq \{\theta_k^{-1} \boldsymbol{\mu}'(., k)\}_{k \in [K]}$, $\bar{\boldsymbol{\nu}} \triangleq \{\theta_k^{-1} \boldsymbol{\nu}(., k)\}_{k \in [K]}$, and $\bar{\boldsymbol{\nu}}' \triangleq \{\theta_k^{-1} \boldsymbol{\nu}'(., k)\}_{k \in [K]}$. Similarly, $\forall \boldsymbol{\mu}, \boldsymbol{\mu}' \in \mathcal{P}_{\boldsymbol{\theta}}(\mathcal{X} \times [K])$, $\forall \boldsymbol{\nu}, \boldsymbol{\nu}' \in \mathcal{P}_{\boldsymbol{\theta}}(\mathcal{U} \times [K])$, $\forall x \in \mathcal{X}$, $\forall u \in \mathcal{U}$, $\forall k \in [K]$, $\forall t \in \{0, 1, \cdots\}$,

$$
|P_k(x, u, \boldsymbol{\mu}, \boldsymbol{\nu}) - P_k(x, u, \boldsymbol{\mu}', \boldsymbol{\nu}')|_1 \leq \bar{L}_P \boldsymbol{\theta}_M^{-1} \left[|\boldsymbol{\mu} - \boldsymbol{\mu}'|_1 + |\boldsymbol{\nu} - \boldsymbol{\nu}'|_1 \right] \quad (64)
$$

$$
|\pi_k^t(x, \boldsymbol{\mu}) - \pi_k^t(x, \boldsymbol{\mu}')|_1 \leq \bar{L}_Q \boldsymbol{\theta}_M^{-1} |\boldsymbol{\mu} - \boldsymbol{\mu}'|_1 \quad (65)
$$

Hence, the given system can equivalently be thought as a multi-agent system satisfying Assumptions 1 and 3 with parameters $\bar{M}_R$, $\bar{L}_R \boldsymbol{\theta}_M^{-1}$, $\bar{L}_P \boldsymbol{\theta}_M^{-1}$ and $\bar{L}_Q \boldsymbol{\theta}_M^{-1}$. Using Theorem 1, the approximation error bound for this translated system can be expressed as follows.

Theorem 21 *Let $\mathbf{x}_0^{\mathbf{N}}$ be the initial states and $\bar{\boldsymbol{\mu}}_0 \in \mathcal{P}^K(\mathcal{X})$ their corresponding distribution. If $\bar{v}^{\mathbf{N}}$ denotes the empirical value function and $\bar{v}^{\mathrm{MF}}$ is its mean-field limit, then for any policy, $\bar{\boldsymbol{\pi}} \in \bar{\Pi}$, the following inequality holds*

$$
\left| \bar{v}^{\mathbf{N}}(\mathbf{x}_0^{\mathbf{N}}, \bar{\boldsymbol{\pi}}) - \bar{v}^{\mathrm{MF}}(\bar{\boldsymbol{\mu}}_0, \bar{\boldsymbol{\pi}}) \right| \leq \frac{\bar{C}_R(\boldsymbol{\theta})}{1 - \gamma} \sqrt{|\mathcal{U}|} \frac{1}{N_{\mathrm{pop}}} \left(\sum_{k \in [K]} \sqrt{N_k} \right)
$$

$$
+ \bar{C}_P(\boldsymbol{\theta}) \left(\frac{\bar{S}_R(\boldsymbol{\theta})}{\bar{S}_P(\boldsymbol{\theta}) - 1} \right) \left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|} \right] \frac{1}{N_{\mathrm{pop}}} \left(\sum_{k \in [K]} \sqrt{N_k} \right) \times \left[\frac{1}{1 - \gamma \bar{S}_P(\boldsymbol{\theta})} - \frac{1}{1 - \gamma} \right] \quad (66)
$$

whenever $\gamma \bar{S}_P(\boldsymbol{\theta}) < 1$ where the parameters are defined as follows,

$$
\begin{aligned}
\bar{C}_R(\boldsymbol{\theta}) &\triangleq \bar{M}_R + \bar{L}_R \boldsymbol{\theta}_M^{-1} \\
\bar{C}_P(\boldsymbol{\theta}) &\triangleq 2 + \bar{L}_P \boldsymbol{\theta}_M^{-1} \\
\bar{S}_R(\boldsymbol{\theta}) &\triangleq \bar{M}_R(1 + \bar{L}_Q \boldsymbol{\theta}_M^{-1}) + \bar{L}_R \boldsymbol{\theta}_M^{-1}(2 + \bar{L}_Q \boldsymbol{\theta}_M^{-1}) \\
\bar{S}_P(\boldsymbol{\theta}) &\triangleq (1 + \bar{L}_Q \boldsymbol{\theta}_M^{-1}) + \bar{L}_P \boldsymbol{\theta}_M^{-1}(2 + \bar{L}_Q \boldsymbol{\theta}_M^{-1})
\end{aligned}
$$

One can verify that the bound (66) is weaker than the bound provided by Theorem 2.

P.2 Loose Bound Using Theorem 2

Consider a multi-agent system satisfying Assumptions 1 and 3. We shall use the notations of Theorem 1. Let, $\boldsymbol{\theta} \triangleq \{\theta_k\}_{k \in [K]}$ be prior probabilities of different classes. If r_k's and P_k's

are given reward and transition functions of the system, then one can define $\bar{r}_k$'s and $\bar{P}_k$'s such that, $\forall x \in \mathcal{X}$, $\forall u \in \mathcal{U}$, $\forall \boldsymbol{\mu} \in \mathcal{P}(\mathcal{X} \times [K])$, $\forall \boldsymbol{\nu} \in \mathcal{P}(\mathcal{U} \times [K])$ and $\forall k \in [K]$,

$$r_k(x, u, \boldsymbol{\mu}, \boldsymbol{\nu}) = \bar{r}_k(x, u, \bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\nu}}),$$
$$P_k(x, u, \boldsymbol{\mu}, \boldsymbol{\nu}) = \bar{P}_k(x, u, \bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\nu}})$$

where $\bar{\boldsymbol{\mu}}$, $\bar{\boldsymbol{\nu}}$ are uniquely defined as, $\bar{\boldsymbol{\mu}} \triangleq \{\theta_k^{-1} \boldsymbol{\mu}(., k)\}_{k \in [K]}$ and $\bar{\boldsymbol{\nu}} \triangleq \{\theta_k^{-1} \boldsymbol{\nu}(., k)\}_{k \in [K]}$. Clearly, $\bar{\boldsymbol{\mu}} \in \mathcal{P}^K(\mathcal{X})$, $\bar{\boldsymbol{\nu}} \in \mathcal{P}^K(\mathcal{U})$. Also, for every policy $\boldsymbol{\pi} \triangleq \{(\pi_k^t)_{k \in [K]}\}_{t \in \{0,1,\cdots\}}$, one can define $\bar{\boldsymbol{\pi}} \triangleq \{(\bar{\pi}_k^t)_{k \in [K]}\}_{t \in \{0,1,\cdots\}}$ such that, $\forall x \in \mathcal{X}$, $\forall k \in [K]$, and $\forall \boldsymbol{\mu} \in \mathcal{P}(\mathcal{X} \times [K])$,

$$\pi_k^t(x, \boldsymbol{\mu}) = \bar{\pi}_k^t(x, \bar{\boldsymbol{\mu}})$$

Note that, the following inequality holds $\forall \bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\mu}}' \in \mathcal{P}^K(\mathcal{X})$, $\forall \bar{\boldsymbol{\nu}}, \bar{\boldsymbol{\nu}}' \in \mathcal{P}^K(\mathcal{U})$, $\forall x \in \mathcal{X}$, $\forall u \in \mathcal{U}$, and $\forall k \in [K]$

$$
\begin{aligned}
|\bar{r}_k(x, u, \bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\nu}}) - \bar{r}_k(x, u, \bar{\boldsymbol{\mu}}', \bar{\boldsymbol{\nu}}')| &= |r_k(x, u, \boldsymbol{\mu}, \boldsymbol{\nu}) - r_k(x, u, \boldsymbol{\mu}', \boldsymbol{\nu}')| \\
&\leq L_R \left[|\boldsymbol{\mu} - \boldsymbol{\mu}'|_1 + |\boldsymbol{\nu} - \boldsymbol{\nu}'|_1 \right] \\
&= L_R \sum_{k \in [K]} \theta_k \left[|\bar{\boldsymbol{\mu}}(., k) - \bar{\boldsymbol{\mu}}'(., k)|_1 + |\bar{\boldsymbol{\nu}}(., k) - \bar{\boldsymbol{\nu}}'(., k)|_1 \right] \quad (67) \\
&\leq L_R \left[|\bar{\boldsymbol{\mu}} - \bar{\boldsymbol{\mu}}'|_1 + |\bar{\boldsymbol{\nu}} - \bar{\boldsymbol{\nu}}'|_1 \right]
\end{aligned}
$$

where $\boldsymbol{\mu} \triangleq \{\theta_k \bar{\boldsymbol{\mu}}(., k)\}_{k \in [K]}$, $\boldsymbol{\mu}' \triangleq \{\theta_k \bar{\boldsymbol{\mu}}'(., k)\}_{k \in [K]}$, $\boldsymbol{\nu} \triangleq \{\theta_k \bar{\boldsymbol{\nu}}(., k)\}_{k \in [K]}$, $\boldsymbol{\nu}' \triangleq \{\theta_k \bar{\boldsymbol{\nu}}'(., k)\}_{k \in [K]}$. Similarly, $\forall \bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\mu}}' \in \mathcal{P}^K(\mathcal{X})$, $\forall \bar{\boldsymbol{\nu}}, \bar{\boldsymbol{\nu}}' \in \mathcal{P}^K(\mathcal{U})$, $\forall x \in \mathcal{X}$, $\forall u \in \mathcal{U}$, $\forall k \in [K]$, $\forall t \in \{0, 1, \cdots\}$,

$$|\bar{P}_k(x, u, \bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\nu}}) - \bar{P}_k(x, u, \bar{\boldsymbol{\mu}}', \bar{\boldsymbol{\nu}}')|_1 \leq L_P \left[|\bar{\boldsymbol{\mu}} - \bar{\boldsymbol{\mu}}'|_1 + |\bar{\boldsymbol{\nu}} - \bar{\boldsymbol{\nu}}'|_1 \right] \qquad (68)$$
$$|\bar{\pi}_k^t(x, \bar{\boldsymbol{\mu}}) - \bar{\pi}_k^t(x, \bar{\boldsymbol{\mu}}')|_1 \leq L_Q |\bar{\boldsymbol{\mu}} - \bar{\boldsymbol{\mu}}'|_1 \qquad (69)$$

Hence, the given system can equivalently be thought as a multi-agent system satisfying Assumptions 2 and 4 with parameters M_R, L_R, L_P and L_Q. Using Theorem 2, the approximation error bound for this translated system can be expressed as follows.

Theorem 22 *If $\mathbf{x}_0^{\mathbf{N}}$ be initial states and $\boldsymbol{\mu}_0 \in \mathcal{P}(\mathcal{X} \times [K])$ its resulting distribution, then $\forall \boldsymbol{\pi} \in \Pi$,*

$$
\begin{aligned}
\left| v^{\mathbf{N}}(\mathbf{x}_0^{\mathbf{N}}, \boldsymbol{\pi}) - v^{\mathrm{MF}}(\boldsymbol{\mu}_0, \boldsymbol{\pi}) \right| &\leq \frac{C_R}{1 - \gamma} \sqrt{|\mathcal{U}|} \left(\sum_{k \in [K]} \frac{1}{\sqrt{N_k}} \right) \\
&+ C_P \left(\frac{S_R}{S_P - 1} \right) \left[\sqrt{|\mathcal{X}|} + \sqrt{|\mathcal{U}|} \right] \left(\sum_{k \in [K]} \frac{1}{\sqrt{N_k}} \right) \times \left[\frac{1}{1 - \gamma S_P} - \frac{1}{1 - \gamma} \right]
\end{aligned}
\qquad (70)
$$

whenever $\gamma S_P < 1$ where $v^{\mathbf{N}}(\cdot, \cdot)$ denotes the empirical value function and $v^{\mathrm{MF}}(\cdot, \cdot)$ is its mean-field limit. The other terms are given as follows: $C_R \triangleq M_R + L_R$, $C_P \triangleq 2 + K L_P$, $S_R \triangleq M_R(1 + L_Q) + L_R(2 + K L_Q)$, and $S_P \triangleq (1 + K L_Q) + K L_P(2 + K L_Q)$.

Clearly, the bound provided by (70) is weaker than the bound suggested in Theorem 1.

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.

Mridul Agarwal, Vaneet Aggarwal, Arnob Ghosh, and Nilay Tiwari. Reinforcement learning for mean-field game. *Algorithms*, 15(3):73, 2022.

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.

Andrea Angiuli, Jean-Pierre Fouque, and Mathieu Laurière. Unified reinforcement q-learning for mean field game and control problems. *arXiv preprint arXiv:2006.13912*, 2020.

Alain Bensoussan, Tao Huang, and Mathieu Laurière. Mean field control and mean field game models with several populations. *Minimax Theory and its Applications*, 3(2):173–209, 2018.

René Carmona and François Delarue. *Probabilistic Theory of Mean Field Games with Applications II: Mean Field Games with Common Noise and Master Equations*, volume 84. Springer, 2018.

René Carmona, Mathieu Laurière, and Zongjun Tan. Linear-quadratic mean-field reinforcement learning: convergence of policy gradient methods. *arXiv preprint arXiv:1910.04295*, 2019a.

René Carmona, Mathieu Laurière, and Zongjun Tan. Model-free mean-field reinforcement learning: mean-field MDP and mean-field Q-learning. *arXiv preprint arXiv:1910.12802*, 2019b.

Yue Chen, Ana Bušić, and Sean P Meyn. State estimation for the individual and the population in mean field control with application to demand dispatch. *IEEE Transactions on Automatic Control*, 62(3):1138–1149, 2016.

Romuald Elie, Julien Perolat, Mathieu Laurière, Matthieu Geist, and Olivier Pietquin. On the convergence of model free learning in mean field games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7143–7150, 2020.

Haotian Gu, Xin Guo, Xiaoli Wei, and Renyuan Xu. Mean-Field Controls with Q-learning for Cooperative MARL: Convergence and Complexity Analysis. *arXiv:2002.04131 [cs, math, stat]*, October 2020. URL http://arxiv.org/abs/2002.04131. arXiv: 2002.04131.

Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. Learning mean-field games. *Advances in Neural Information Processing Systems*, 32:4966–4976, 2019.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2017.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020.

Sham M Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

Daniel Lacker. Limit theory for controlled mckean–vlasov dynamics. *SIAM Journal on Control and Optimization*, 55(3):1641–1672, 2017.

Yanli Liu, Kaiqing Zhang, Tamer Basar, and Wotao Yin. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. In *NeurIPS*, 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.

Barna Pasztor, Ilija Bogunovic, and Andreas Krause. Efficient Model-Based Multi-Agent Mean-Field Reinforcement Learning. *arXiv:2107.04050 [cs, stat]*, July 2021. URL `http://arxiv.org/abs/2107.04050`. arXiv: 2107.04050.

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, 2020.

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

Howard M Schwartz. *Multi-agent machine learning: A reinforcement approach*. John Wiley & Sons, 2014.

Kyunghwan Son, Daewoo Kim, Wan Ju Kang, David Earl Hostallero, and Yung Yi. Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 5887–5896. PMLR, 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 based on team reward. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'18)*, volume 3, pages 2085–2087, 2018.

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

Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. In *International Conference on Learning Representations*, 2019.

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(3-4):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.

Jiachen Yang, Xiaojing Ye, Rakshit Trivedi, Huan Xu, and Hongyuan Zha. Learning deep mean field games for modeling large population behavior. In *International Conference on Learning Representations (ICLR)*, 2018.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, pages 321–384, 2021.

Changxi Zhu, Ho-fung Leung, Shuyue Hu, and Yi Cai. A Q-values sharing framework for multiple independent Q-learners. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 2324–2326, 2019.