Finite Approximations and Q learning for Mean Field Type Multi Agent Control*

Erhan Bayraktar, Nicole Bäuerle, and Ali Devran Kara †

November 18, 2022

Abstract

We study a multi-agent mean field type control problem in discrete time where the agents aim to find a socially optimal strategy and where the state and action spaces for the agents are assumed to be continuous. We provide several approximations for this problem: (i) We first show that (building on Bauerle, N. 2021, arXiv:2106.08755) the problem can be seen as an empirical measure valued centralized Markov decision process (MDP), where the agents' identities are not relevant but only their state information is used for the optimal strategy without loss of optimality. (ii) Secondly, we construct a finite empirical measure valued MDP by discretizing the state spaces of the agents, and show that the strategies chosen based on this formulation, which only keep track of how many agents are in the discretized aggregate sets, are nearly optimal with increasing discretization rates. (iii) Finally, we look at large population problems, where the state space of the problem can be taken finite after discretization, but can be too large for computational efficiency. Different from the usual approach in the literature, instead of going to the infinite population limit, which results in a continuous problem and requires further discretizations again, we perform approximations by choosing a smaller sample of the large population and find strategies for this small sample and generalize them to the original large population.

Furthermore, we provide convergent Q learning iterations for the approximate models, and show that the learned policies are nearly op-

---

*E. Bayraktar is partially supported by the National Science Foundation under grant DMS-2106556 and by the Susan M. Smith chair.

†N. Bauerle is with the Department of Mathematics, Karlsruhe Institute of Technology, Karlsruhe, Germany. E. Bayraktar and A.D. Kara are with the Department of Mathematics, University of Michigan, Ann Arbor, MI, USA, Email: nicole.baueuerle@kit.edu, {erhan,alikara}@umich.edu
timal under standard regularity assumptions on the cost and dynamic functions of the problem.

1 Introduction

The goal of this paper is to develop a finite state and action approximation method for mean field type control in discrete time, and apply these findings to the Q learning of the mean field control problems with general state and action spaces.

The dynamics for the model are presented as follows: suppose $N$ agents (decision-makers or controllers) act in a cooperative way to minimize a cost function, and the agents share a common state and an action space denoted by $X \subset \mathbb{R}^l$ and $U \subset \mathbb{R}^m$ for some $l, m < \infty$, for any time step $t$, and agent $i \in \{1, \ldots, N\}$ we have

$$x_{t+1}^i = f(x_t^i, u_t^i, \mu_{X_t}, w_t^i, w_0^i)$$

for a measurable function $f$, where $w_t^i$ denotes the i.i.d. idiosyncratic noise process, and $w_0^i$ denotes the i.i.d. common noise process. Furthermore, $\mu_X$ denotes the empirical distribution of the agents on the state space $X$ such that for a given joint state $x := \{x^1, \ldots, x^N\} \in X^N$

$$\mu_X := \frac{1}{N} \sum_{i=1}^{N} \delta_{x^i}.$$  

At every time stage $t$, each agent receives a cost determined by a measurable stage-wise cost function $c : X \times U \times \mathcal{P}_N(X) \to \mathbb{R}$, where $\mathcal{P}_N(X)$ is the set of all empirical measures on $X$ constructed using $N$ dimensional state vectors.

For the remainder of the paper, by an abuse of notation, we will sometimes denote the dynamics in terms of the vector state and action variables, $\mathbf{x} = (x^1, \ldots, x^N)$, and $\mathbf{u} = (u^1, \ldots, u^N)$, and vector noise variables $\mathbf{w} = (w^0, w^1, \ldots, w^N)$ such that

$$\mathbf{x}_{t+1} = f(\mathbf{x}_t, \mathbf{u}_t, \mathbf{w}_t).$$

For the initial formulation, every agent is assumed to know the state and action variables of every other agent. We define an admissible policy for an agent $i$, as a sequence of functions $\gamma^i := \{\gamma^i_t\}_t$, where $\gamma^i_t$ is a $U$-valued (possibly randomized) function which is measurable with respect to the $\sigma$-algebra generated by

$$I_t = \{x_0, \ldots, x_t, u_0, \ldots, u_{t-1}\}. \quad (1)$$
Accordingly, an admissible team policy, is defined as $\gamma := \{\gamma^1, \ldots, \gamma^N\}$, where $\gamma^i$ is an admissible policy for the agent $i$. In other words, agents share full information.

The objective of the agents is to minimize the following cost function

$$J^N_\beta(x_0, \gamma) = \sum_{t=0}^{\infty} \beta^t E_{\gamma} [c(x_t, u_t)]$$

where $0 < \beta < 1$ is some discount factor and

$$c(x_t, u_t) := \frac{1}{N} \sum_{i=1}^{N} c(x^i_t, u^i_t, \mu_{x_t}).$$

The optimal cost is defined by

$$J^{N,*,\beta}_\gamma(x_0) := \inf_{\gamma \in \Gamma} J^{N}_\beta(x_0, \gamma)$$

where $\Gamma$ denotes the set of all admissible team policies.

We note that this information structure will be our benchmark to evaluate the performance of the approximate solutions that will be presented in the paper. In other words, the value function that is achieved when the agents share full information will be taken to be our reference point for simpler information structures.

1.1 Literature Review and Contributions

Mean-field setup is used to study multi agent problems where the population is homogeneous and weakly interacted such that the dynamics of each agent depends on the other agents only through the empirical distribution of agents. Depending on the cost structure of the problem, this setting is usually referred to as mean-field game theory if the agents are competitive or mean-field control if the agents are cooperative. The key observation for both settings is that for large populations (when the number of agents goes to infinity), the decentralized problem can be reduced to a centralized decision problem, which can be analyzed through the perspective of a representative agent.

Mean-field game theory has been introduced independently by [22, 26]. The area has gained a lot of attention after the first introduction. We refer the reader to [16, 8, 6, 38, 21, 1, 12, 14, 19, 30, 37, 33, 34] and references in them for mean-field game theory studies both in continuous and discrete
time, studying different models and cost structures including computational and learning related methods.

For multi-agent control problems, the solution, in general, is intractable except for special information structures between the agents. Mean-field type team problems, where the agents are only related through the mean field term, is one these special cases. For team control problems, where the agents are cooperative and work together to minimize (or maximize) a common cost (or reward) function, majority of the studies focus on the continuous time dynamics, for the limit problem, see [4, 11, 27, 31, 9, 15, 3, 5, 35, 36] and references therein for the study of dynamic programming principle, learning methods, and justification of the exchangeability of agents for large (possibly infinite) agent team settings. The papers [25, 13] provide the justification for studying the centralized limit problem by rigorously connecting the large population decentralized setting and the infinite population limit problem.

The closest papers to the formulation and setup of our paper are [29, 2, 18, 17, 28, 10] where the discrete time mean-field control problem is studied. [29, 28] study mean-field control for infinite population problems under so called open loop controls where the controller can use the realization of the noise process to decide on actions, furthermore the authors rigorously prove the connection between the limit problem and the finite population problem with a rate of convergence. [18, 17] study dynamic programming principle and solutions to the limit (infinite population) problem, and Q learning methods. Similarly, [10] studies different classes of policies that achieve optimal performance for the limit problem and focuses on Q learning of the problem after establishing the optimality of randomized feedback policies for the agents. Finally, [2] first shows that even without going to the infinite population problem, one can construct a measure valued MDP problem for the finite population problem where the new state space is the set of N-length empirical distributions defined on the state space of the agents, and hence the agents can choose their actions (in a randomized manner) by only looking at their own state and the empirical distribution of other agents. The paper also establishes the connection of the finite population problem to the limit problem. Furthermore, the author studies the implication of these findings to the infinite horizon average cost problems.

We work directly with finite population setup, inspired by the key observation in [2], that one does not have to go to the infinite population limit to reduce the problem into a measure valued MDP, and that the optimal performance can be achieved, if the agents have access to the (finite population) empirical distributions of the agents. We note that unlike the infinite-limit
measure valued MDP reduction, for the finite population measure valued reduction, the new state space becomes the set of all N-length (where N is the number of agents) empirical distributions rather than the set of all probability measures on the original state space. However, even after reducing the problem to the empirical measure valued counterpart, if the original state space is continuous, the state space of the measure valued problem will be a continuous state space. Hence, we then further approximate the problem by discretizing the state space and creating a measure valued MDP defined on a finite set, and we provide provable error bounds resulting from the discretization.

This measure valued finite MDP will be manageable for small populations, however, for large populations, we may encounter dimension problems even though the size of the new state space is finite. The common approach to deal with this issue has been to go to the infinite population limit and work with the measure valued MDP for the limit problem. However, in the limit we have an MDP problem with a continuous state space, since even when the state space of agents is finite, the set of probability measures on this finite set will be continuous. Hence, one usually needs to discretize the measure valued infinite population problem further for numerical and learning applications. Therefore, rather than going to the infinite population limit and discretizing it again, we directly discretize the finite population measure valued problem. For the discretization, we use two different methods: (i) we first discretize the set of N-length empirical measures directly by working on the set of measures, (ii) secondly, we pick a smaller sample of the original population and solve the problem for this smaller population, and generalize their policies to the large population, which indirectly discretizes the set of empirical measures for the original population. Note that the second method is opposed to the common approach in the literature, where one uses the infinite population solutions for finite (but large) populations, whereas, we analyze the case where one uses the small population policies for larger populations.

We now summarize our contributions in detail:

**Contributions:** Our main goal in this paper is to develop approximate solution methods and provide convergent learning algorithms for the approximate models. We will develop the approximation techniques in several steps. As introduced in the problem formulation, the original team problem can be modeled as a centralized MDP with state space $X^N$, where $N$ is the number of agents.

- In Section 2, we show that the team problem can be modeled as a
centralized measure valued MDP, and the state space can be reduced to \( \mathcal{P}_N(X) \), the set of empirical measures with length \( N \) on \( X \), without loss of optimality. We show that the optimal policies for agents can be realized as randomized policies for this formulation. This step can be thought of as removing the identities of the agents using the exchangibility of them and only focusing on which parts of the state space agents live, rather than keeping track of states of agents with agent identities.

- In Section 3, we show that by discretizing the state space \( X \), and separating \( X \) into \( M \) disjoint subsets, one can construct an MDP whose state space is finite with size \( \frac{(M+N-1)!}{(N-1)!M!} \). This step can be thought as only counting the number of agents in the chosen subsets of the state space \( X \), rather than looking at every point in the state space. We show that the policies constructed for this approximate model are near optimal for large \( M \), if the dynamics and the cost functions satisfy certain regularity properties.

- The approximation in Section 3 results in a finite model, however, for large number of agents, the problem can be intractable. In Section 4, we present further approximation methods, (i) first by choosing a further subset of the state space presented in Section 3, and using less refined counting methods, (ii) second, by choosing a smaller subsample of the large population.

- In Section 5, we show that a convergent Q learning algorithm can be constructed to estimate the policies for the models from Sections 3 and 4. The algorithm keeps track of the cost realizations of the agents and the number of agents in the discretization bins of the state space.

2 Measure Valued Centralized MDP Construction

In this section, we will define a Markov decision process, for the empirical distribution of the agents, where the control actions are the joint empirical distribution of the state and action vectors of the agents.

We let the state space to be \( Z = \mathcal{P}_N(X) \) which is the set of all empirical measures on \( X \) that can be constructed using the state vectors of the agents.

The admissible set of actions for some state \( \mu \in Z \), is denoted by \( U(\mu) \), where

\[
U(\mu) = \{ \Theta \in \mathcal{P}_N(X \times U) | \Theta(\cdot, U) = \mu(\cdot) \},
\]  

(3)
that is, the set of actions for a state $\mu$, is the set of all joint empirical measures on $X \times U$ whose marginal on $X$ coincides with $\mu$.

We equip the state space $\mathcal{Z}$, and the action sets $U(\mu)$, with the first order Wasserstein distance $W_1$.

In order to define the transition model for this centralized MDP, we note that the empirical distributions of the agents’ states of the original team problem induces a controlled Markov chain. In particular, for some set $B \in \mathcal{B}(\mathcal{Z})$, we can write

$$Pr(\mu_{t+1} \in B|\mu_t, \ldots, \mu_0, \Theta_t, \ldots, \Theta_0) = \int_{x_t, u_t \in X^N \times U^N} Pr(\mu_{t+1} \in B|x_t, u_t) Pr(dx_t, du_t|\mu_t, \ldots, \mu_0, \Theta_t, \ldots, \Theta_0).$$

For any $\mu_{x_t, u_t} = \Theta_t$, and $\mu_{x_t} = \mu_t$, the inside term can be written as

$$Pr(\mu_{t+1} \in B|x_t, u_t) = \int 1\{f(x_{t+1}|x_t, u_t, \mu_{x_t}, \mu_{x_{t+1}}) \in B\} P(dw_t)$$

where $P(\cdot)$ is the probability measure governing the idiosyncratic and the common noise processes, and $w_t$ is the noise vector with length $N + 1$.

If any two pairs $(x_t, u_t), (x'_t, u'_t)$ have the same empirical distribution $\Theta_t$, then they can be viewed as reordered versions of each other. Furthermore, since the dynamics are identical for every agent, i.e. since the agent are exchangeable, for some $w_t$, and $x_{t+1} = f(x_t, u_t, \mu_{x_t}, w_t)$, where $\mu_{x_{t+1}} = \mu_{t+1}$, by reordering $w_t$, one can construct some $w'_t$ such that $x'_{t+1} = f(x'_t, u'_t, \mu_{x_t}, w'_t)$, where $x'_{t+1}$ is just a reordered version of $x_{t+1}$, and in particular $\mu_{x_{t+1}} = \mu_{x_{t+1}}$.

Since the idiosyncratic noises are identically distributed for every agent, as a result of the above discussion, for any two pairs $(x_t, u_t), (x'_t, u'_t)$ with the same empirical distribution $\Theta_t$,

$$Pr(\mu_{t+1} \in B|x_t, u_t) = Pr(\mu_{t+1} \in B|x'_t, u'_t).$$

Therefore, the empirical distributions of the agents’ states $\mu_t$, and of the joint state and actions $\Theta_t$ define a controlled Markov chain such that

$$Pr(\mu_{t+1} \in B|\mu_t, \ldots, \mu_0, \Theta_t, \ldots, \Theta_0) = Pr(\mu_{t+1} \in B|\mu_t, \Theta_t)$$

$$:= \eta(B|\mu_t, Q_t)$$

$$= Pr(\mu_{t+1} \in B|x_t, u_t), \text{ for any } (x_t, u_t) : \mu_{(x_t, u_t)} = \Theta_t$$

where $\eta(\cdot|\mu, \Theta) \in \mathcal{P}(\mathcal{P}_N(X))$ is the transition kernel of the centralized measure valued MDP, which is induced by the dynamics of the team problem.
We define the stage-wise cost function $k(\mu, \Theta)$ by

$$k(\mu, \Theta) := \int c(x, u, \mu)\Theta(dx, du) = \frac{1}{N} \sum_{i=1}^{N} c(x^i, u^i, \mu).$$

Thus, we have an MDP with state space $\mathcal{Z}$, action space $\cup_{\mu \in \mathcal{Z}} U(\mu)$, transition kernel $\eta$ and the stage-wise cost function $k$.

We define the set of admissible policies for this measured valued MDP as a sequence of functions $g = \{g_0, g_1, g_2, \ldots\}$ such that at every time $t$, $g_t$ is measurable with respect to the $\sigma$-algebra generated by the information variables

$$I_M = \{\mu_0, \ldots, \mu_t, \Theta_0, \ldots, \Theta_{t-1}\}.$$ 

We denote the set of all admissible control policies by $G$ for the measure valued MDP.

In particular, we define the infinite horizon discounted expected cost function under a policy $g$ by

$$K^N_\beta(\mu_0, g) = E^\mu_0 \left[ \sum_{t=0}^{\infty} \beta^t k(\mu_t, \Theta_t) \right].$$

We also define the optimal cost by

$$K^N_\beta(\mu_0) = \inf_{g \in G} K^N_\beta(\mu_0, g).$$

2.1 Equivalence between the Team Problem and the Measure Valued MDP

In this section, we show that the optimal value function of the original team problem, $J^N_\beta$ (see (2)), and the optimal value function of the measure valued MDP, $K^N_\beta$ (see (6)) are equal. Furthermore, an optimal policy designed for the measure valued MDP can be realized as a randomized policy for the team problem, which achieves the optimal performance.

Before the main result of this section, we present the set of assumptions that will be used frequently in the paper and we present a key lemma:

**Assumption 1.** i. $X$ and $U$ are compact.

ii. $f$ is Lipschitz in $x, u, \mu_X$ such that

$$|f(x, u, \mu_X, w^i, w^0) - f(x', u', \mu_{X'}, w^i, w^0)| \leq K_f \left(|x - x'| + |u - u'| + W_1(\mu_X, \mu_{X'})\right)$$
for some $K_f < \infty$, uniformly in $w, w^0$ where $W_1$ is the first order Wasserstein distance.

iii $c$ is Lipschitz in $x, u, \mu$ such that

$$|c(x, u, \mu) - c(x', u', \mu')| \leq K_c \left( |x - x'| + |u - u'| + W_1(\mu, \mu') \right)$$

for some $K_c < \infty$.

**Lemma 1.** Under Assumption 1, we have that the optimal value function is constant over the states with the same empirical distribution, i.e. if $x, x' \in X^N$ have the same empirical distribution, $\mu = \mu'$, then

$$J_{\beta}^N(x) = J_{\beta}^N(x').$$

**Proof.** The proof can be found in Appendix A.1

**Theorem 1.** Under Assumption 1, for any $x_0$ that satisfies $x_0 = \mu_0$, we have that

i.

$$K_{\beta}^{N,*}(\mu_0) = J_{\beta}^{N,*}(x_0).$$

ii. For any stationary and Markov policy $g$ for the measure valued MDP, every agent can construct a randomized policy $\gamma : X \times \mathcal{P}_N(X) \to \mathcal{P}_N(U)$, that is each agent makes a randomized decision by looking at their own states and the empirical distribution of the other agents. If we denote the resulting team policy also by $\gamma$ with an abuse of notation, we have that

$$J_{\beta}^N(x_0, \gamma) = K_{\beta}^N(\mu_0, g).$$

iii. There exists a stationary and Markov optimal policy $g^*$ for the measure valued MDP, and as a particular instance of part (ii), using $g^*$, every agent can construct a randomized policy $\gamma : X \times \mathcal{P}_N(X) \to \mathcal{P}_N(U)$

$$J_{\beta}^N(x_0, \gamma) = J_{\beta}^{N,*}(x_0).$$

**Proof.** The proof can be found in Appendix A.2.
Remark 1. The second part of the previous result is significant for reducing the complexity of the problem. The result can be interpreted in different ways.

The first interpretation is that, a central planner who only keeps track of the empirical distribution of the agents, chooses an empirical measure on $\mathbb{X} \times \mathbb{U}$, as a control action, and transfers this empirical measure to agents as a ‘prescription’, and every agent constructs a randomized control function using the higher level policy prescribed by the central planner. Hence, if there exists a central planner, the optimal performance can be achieved if the central planner only keeps track of the empirical distribution of the agents, and the agents only keep track of their own states. This is a significant relaxation, compared to the original information structure where every agent keeps track of the states of all the other agents.

Another interpretation is that there does not need to be a central planner in the literal terms, if the agents have access to the empirical distribution of the other agents. In this case, any agent can be viewed as a central planner, as long as the agents have the same empirical distribution, and the resulting policies will be consistent since the dynamics are identical for agents and we are concerned with the team (or social) optimality.

In either case, our search space for optimality reduces from $\mathbb{X}^N$ to $\mathcal{P}_N(\mathbb{X})$ (set of all possible empirical measures on $\mathbb{X}$ with length $N$).

3 Finite Model Construction

3.1 Action Space Discretization

We start by reducing the action space $\mathbb{U}$ of agents to a finite set.

We let $\hat{\mathbb{U}} = \{u_1, \ldots, u_k\} \subset \mathbb{U}$ be a subset of the original action space and we define:

$$L_U := \sup_{u \in \mathbb{U}} \min_{\hat{u} \in \hat{\mathbb{U}}} |u - \hat{u}|$$

which is the worst error bound that results from the discretization of the action space.

One can define the team problem introduced in Section 1 in an identical way where the action space of the agents is replaced with $\hat{\mathbb{U}}$. We denote the optimal value function of the team problem with finite action spaces by $\hat{J}^N,\beta(x_0)$.

Recall that the team problem in its original formulation can be seen as a centralized MDP when every agent has access to the state and action.
information of every other agent. Hence the following is a direct implication of [32, Theorem 3.16]:

**Theorem 2.** Under Assumption 1, we have

\[ |\tilde{J}_{\beta}^N(x_0) - J_{\beta}^N(x_0)| \leq KL_{U} \]

for some constant \( K < \infty \) that depends on the discount factor \( \beta \) and the continuity parameters introduced in Assumption 1.

Since we can control the upperbound in the last result under the assumption that \( U \) is compact, in the sequel, we assume that \( U \) is finite.

### 3.2 State Space Discretization

For the finite state space approximation, we start by choosing a collection of disjoint sets \( \{B_i\}_{i=1}^M \) such that \( \cup_i B_i = X \), and \( B_i \cap B_j = \emptyset \) for any \( i \neq j \). Furthermore, we choose a representative state, \( \hat{x}_i \in B_i \), for each disjoint set. For this setting, we denote the new finite state space by \( \hat{X} := \{\hat{x}_1, \ldots, \hat{x}_M\} \).

The mapping from the original state space \( X \) to the finite set \( \hat{X} \) is done via

\[ \phi(x) = \hat{x}_i \quad \text{if} \quad x \in B_i. \]  

(8)

With an abuse of notation, we extend this definition to the state vector of the agents, such that \( \phi(x) \) denotes the mapping when \( \phi \) is applied to every element of \( x = \{x^1, \ldots, x^N\} \) separately.

Furthermore, we choose a weight measure

\[ \pi(\cdot) \in \mathcal{P}(X^N). \]  

(9)

such that \( \pi(B_{i_1} \times B_{i_2} \times \cdots \times B_{i_N}) > 0 \) for any \( B_{i_1}, B_{i_2}, \ldots, B_{i_N} \).

We also define the uniform quantization error that results from the discretiation of the state space

\[ L_X := \max_i \sup_{x, x' \in B_i} |x - x'|. \]  

(10)

### 3.3 Measure Valued Finite MDP Construction

In this section, building on the discretization in Section 3.2, we will construct a measure valued finite MDP where the state space is \( \mathcal{P}_N(\hat{X}) \), and the action space is \( \mathcal{P}_N(\hat{X} \times \hat{U}) \). In other words, the central controller observes the empirical distribution of the agents on the finite space \( \hat{X} \), say \( \mu_{\hat{X}} \), and
selects an empirical distribution on $\hat{X} \times \hat{U}$ as an action with the restriction that the marginal on $\hat{X}$ is $\mu_{\hat{X}}$.

Note that with a simple combinatorial analysis, it can be shown that the size of the state space $P_N(\hat{X})$ is $\frac{(M+N-1)!}{(N-1)!M!}$. Similarly, for a given state $\mu_{\hat{X}}$, the set of all admissible actions has the size $M \frac{(|U|+N-1)!}{(N-1)!|U|!}$. Hence, the new state and action spaces are finite.

We enumerate the states of the finite state space $P_N(\hat{X})$, and denote by $\mu_j$ the $j$th state. We now show that the discretization in the space $X$ induces a discretization in $P_N(X)$, which creates $P_N(\hat{X})$. We define the sets

$$A_j = \{\mu_x : \mu_{\phi(x)} = \mu_j\},$$

that is, $A_j$ is the set of empirical measures on $X$ for the vector states $x$, which give the finite empirical measure $\mu_j$ when they are discretized.

Note that, these sets define a map from $P_N(X)$ to $P_N(\hat{X})$, such that each $\mu \in P_N(X)$ is mapped to some $\mu_j \in P_N(\hat{X})$. We will denote this map by also $\phi$ with an abuse of notation to avoid notational clutter.

The next supporting result shows that the quantization error, $L_X$, for the space $X$, projects to the quantization error for the space $P_N(X)$, when it is metrized by the first order Wasserstein metric.

**Lemma 2.**

$$\sup_j \sup_{\mu, \nu \in A_j} W_1(\mu, \nu) \leq L_X.$$  

where $L_X$ is as defined in (10).

**Proof.** Let $x_1, x_2$ be such that $\mu_{\phi(x_1)} = \mu_{\phi(x_2)}$. Note that since the discretized versions of $x_1$ and $x_2$ have the same empirical distributions, then $x_1$ and $x_2$ have the exact same number of elements in every quantization bin of $X$. By reordering, and collecting the elements from the same quantization bin together:

$$W_1(\mu_{x_1}, \mu_{x_2}) = \sup_{Lip(h) \leq 1} \left| \int h d\mu_{x_1} - \int h d\mu_{x_2} \right| \leq \sup_{Lip(h) \leq 1} \frac{1}{N} \sum_{i=1}^{N} |h(x_1^i) - h(x_2^i)| \leq L_X$$

where the last step follows from the fact that $h$’s Lipschitz coefficient is smaller than 1, and $x_1^i, x_2^i$ belong to the same quantization bin of $X$ due to reordering. Furthermore, the quantization error on $X$ is upper bounded by $L_X$ by definition (see (10)).

12
We now construct the normalized weight measures for the subsets \( A_j \subset \mathcal{P}_N(X) \), using the previously chosen measures \( \pi \) (see (9)), such that given the finite empirical measure is \( \mu_j \), for \( A \subset A_j \)

\[
Pr(\mu \in A | \mu \in A_j) = \frac{Pr(x : \mu_x \in A)}{Pr(x : \mu_{\phi(x)} = \mu_j)} = \frac{\pi(\{x : \mu_x \in A\})}{\pi(\{x : \mu_{\phi(x)} = \mu_j\})} := \hat{\pi}_j(A).
\]

(11)

Using these normalized weight measures, we now define the cost function and the transition model for the finite measure valued MDP.

We denote the cost function by \( \hat{k} : \mathcal{P}_N(\hat{X}) \times \mathcal{P}_N(\hat{X} \times \hat{U}) \rightarrow \mathbb{R} \) such that for some \( \mu_j \in \mathcal{P}(\hat{X}) \) and \( \hat{\Theta} \in \mathcal{P}_N(\hat{X} \times \hat{U}) \)

\[
\hat{k}(\mu_j, \hat{\Theta}) := \int_{A_j} k(\mu, \Theta_{\mu})\hat{\pi}_j(d\mu)
\]

(12)

where \( \hat{\pi}_j \) is the normalized weight measure for the set \( A_j \subset \mathcal{P}_N(X) \) of \( \mu_j \), and \( k \) is the cost function of the original measure valued MDP as defined in (5). Furthermore, for \( \hat{\Theta}(dx, du) = \gamma(du|x)\mu_j(dx) \), we have \( \Theta_{\mu}(dx, du) = \gamma(du|\phi(x))\mu_j(dx) \).

We denote the transition law of the finite model by \( \hat{\eta} \) such that

\[
\hat{\eta}(\mu_i | \mu_j, \hat{\Theta}) = \int_{A_j} \eta(A_i | \mu, \Theta_{\mu})\hat{\pi}_j(d\mu)
\]

(13)

where \( \eta \) is the transition kernel of the original measure valued MDP as defined in (4).

In other words, to define the cost and the transitions of the finite model, we average over the cost and the transitions of the original model using the normalized weight measures.

For some \( \mu_j \in \mathcal{P}_N(\hat{X}) \) and an admissible policy \( g^N \), the infinite horizon discounted cost function is defined as

\[
\hat{K}^*_\beta(\mu_j, g^N) := \sum_{t=0}^{\infty} \beta^t E[\hat{k}(\mu_t, g^N(\mu_t))]
\]

where the expectation is defined with respect to the kernel \( \hat{\eta} \) and the initial point \( \mu_j \). We will denote the optimal value function of this finite measure valued MDP by

\[
\hat{K}^*_\beta(\mu_j) := \inf_{g^N \in G^N} \hat{K}^*_\beta(\mu_j, g^N)
\]
for every $j$, where $G^N$ is the set of all admissible policies for the finite measure valued MDP.

Note that we can extend the optimal value function $\hat{K}^{N,*}_{\beta}(\mu_j)$ which is defined on $\mathcal{P}_N(\hat{X})$, over the set $\mathcal{P}_N(X)$, by making it constant over the subsets $A_j$. Hence, we can compare the value functions of the original measure valued MDP problem, and the finite measure valued MDP problem. For the main result of this section, we need the following lemmas:

**Lemma 3.** Let $\mu, \mu' \in \mathcal{P}_N(X)$ and $u \in \hat{U}^N$. Under Assumption 1, there exist state vectors $x, x'$ such that $\mu_x = \mu$, $\mu_{x'} = \mu'$ and $\Theta = \mu(x, u), \Theta' = \mu(x', u) \in \mathcal{P}_N(X \times \hat{U})$ and we have

$$|k(\mu, \Theta) - k(\mu', \Theta')| \leq 2K_c W_1(\mu, \mu')$$

$$W_1(\eta(\cdot|\mu, \Theta), \eta(\cdot|\mu', \Theta')) \leq 2K_f W_1(\mu, \mu').$$

**Proof.** The proof can be found in Appendix A.3.

**Lemma 4.** Under Assumption 1, if $2K_f \beta < 1$ for any $\mu, \mu' \in \mathcal{P}_N X$

$$|K^{N,*}_{\beta}(\mu) - K^{N,*}_{\beta}(\mu')| \leq \frac{2K_c}{1 - 2K_f \beta} W_1(\mu, \mu').$$

**Proof.** The proof can be found in Appendix A.4. Note that the result also follows from Theorem 4.1. (c) in [20].

The following result shows that the difference between the value functions of the finite measure valued model and the original measure valued model can be bounded:

**Proposition 1.** Under Assumption 1, for any $\mu \in \mathcal{P}_N(X)$

$$\left|\hat{K}^{N,*}_{\beta}(\mu) - K^{N,*}_{\beta}(\mu)\right| \leq \frac{2K_c}{(1 - \beta)(1 - 2\beta K_f)} L_X.$$

**Proof.** The proof can be found in Appendix A.5.

3.4 Near Optimality of the Policy Constructed From the Finite Model

We first summarize the steps to construct the approximate policy.

- Construct the transition kernel and the cost function for the measure valued centralized MDP using (4) and (5).
• Choose a normalizing measure $\pi \in \mathcal{P}(X)$ and quantization bins $\{B_i\}_{i=1}^M \subset X$, such that $\bigcup_{i=1}^M B_i = X$. Choose representative states for the quantization bins, say $\{\hat{x}_i\}_{i=1}^M$ and denote the finite space by $\hat{X} = \{\hat{x}_1, \ldots, \hat{x}_M\}$.

• Construct a finite MDP based on the measure valued MDP using (10) and (9).

• Compute the optimal policy for the finite model, say $\hat{\gamma}^N$:

$$\hat{\gamma}^N : \mathcal{P}_N(\hat{X}) \rightarrow \mathcal{P}_N(\hat{X} \times \hat{U}).$$

• Compute the randomized policies for the agents using $\hat{\gamma}$, such that, for some $\hat{\mu} \in \mathcal{P}_N(\hat{X})$, disintegrate $\hat{\gamma}(\hat{\mu}) = \Theta(d\hat{x}, du) \in \mathcal{P}_N(\hat{X} \times \hat{U})$, to find $\hat{\gamma}_{\hat{\mu}}(du|\hat{x})\hat{\mu}(d\hat{x}) = \Theta(d\hat{x}, du)$.

• Randomized policies, $\hat{\gamma}_{\hat{\mu}}(du|\hat{x})$, are functions defined on the finite space $\hat{X}$. To use them in the original problem, simply extend them over the original space $X$, by making them constant over the quantization bins.

**Theorem 3.** Under Assumption 1 if $2\beta K_f < 1$, and if we denote the randomized policies induced by the measure valued finite model by $\hat{\gamma}^N$ and apply it in the original team problem, we will have

$$J^N_\beta (x_0, \hat{\gamma}^N) - J^{N,*}_\beta (x_0) \leq \frac{4K_\epsilon}{(1 - \beta)^2(1 - 2\beta K_f)} L_X$$

where $L_X$ is the uniform quantization error.

**Proof.** Note that the application of $\hat{\gamma}^N$ implies a policy for the measure valued MDP, say $\hat{\mu}$, which can be written for some empirical measure $\mu \in \mathcal{P}_N(X)$ as

$$\hat{\mu} = \hat{\gamma}_{\hat{\mu}}(du|\phi(\hat{x}))\mu(d\hat{x}).$$

Using Theorem 1 (ii), we have that

$$J^N_\beta (x_0, \hat{\gamma}^N) = K^N_\beta (\mu_0, \hat{\mu})$$

$$J^{N,*}_\beta (x_0) = K^{N,*}_\beta (\mu_0)$$

where $\mu_0 = \mu_{x_0}$.

We start with the following bound

$$K^N_\beta (\mu_0, \hat{\mu}) - K^{N,*}_\beta (\mu_0) \leq |K^N_\beta (\mu_0, \hat{\mu}) - \hat{K}^{N,*}_\beta (\mu_0)| - |\hat{K}^{N,*}_\beta (\mu_0) - K^{N,*}_\beta (\mu_0)|.$$  

(14)
The second term is bounded by Proposition 1, so we focus on the first term. Recall that, for \( \hat{K}^{N,*}_\beta (\mu_0) \), given that \( \mu_{\phi(x_0)} = \mu_j \), we have

\[
\hat{K}^{N,*}_\beta (\mu_0) = \int_{A_j} k(\mu', \Theta_{\mu'})\bar{\pi}_j(d\mu') + \beta \int_{A_j} \int_{\mu_1} \hat{K}^{N,*}_\beta (\mu_1)\eta(d\mu_1|\mu', \Theta_{\mu'})\bar{\pi}_j(d\mu'),
\]

where for \( \Theta_{\mu'} = \gamma^N(\mu') \), and thus, \( \Theta_{\mu'}(dx, du) = \gamma(dx)\mu'(dx) \).

The Bellman equation for \( K^{N}_{\beta}(\mu_0, \hat{g}) \) can be written as

\[
K^{N}_{\beta}(\mu_0, \hat{g}) = k(\mu_0, \hat{g}(\mu_0)) + \beta \int K^{N}_{\beta}(\mu_1, \hat{g})\eta(d\mu_1|\mu_0, \hat{g}(\mu_0)).
\]

For \( \mu_0 \in A_j \) and for any \( \mu' \in A_j \), since \( \hat{g} \) uses the discretized versions of the state variables, and since \( \mu \) and \( \mu_j \) belong to the same quantization bin, we can find state and action vectors \( x, x', u \) in accordance with Lemma 3 such that \( \mu(x, u) = \hat{g}(\mu_0) \) and \( \mu(x', u) = \hat{g}(\mu') \). Hence, by Lemma 3

\[
|k(\mu_0, \hat{g}(\mu_0)) - k(\mu', \hat{g}(\mu'))| \leq 2K_cW_1(\mu_0, \mu')
\]

\[
W_1(\eta(\cdot|\mu_0, \hat{g}(\mu_0)), \eta(\cdot|\mu', \hat{g}(\mu'))) \leq 2K_fW_1(\mu_0, \mu').
\]

We can then write the following:

\[
|K^{N}_{\beta}(\mu_0, \hat{g}) - \hat{K}^{N,*}_\beta (\mu_0)| \leq \int_{A_j} |k(\mu', \Theta_{\mu'}) - k(\mu_0, \hat{g}(\mu_0))| \bar{\pi}_j(d\mu')
\]

\[
+ \beta \int_{A_j} \int_{\mu_1} \hat{K}^{N,*}_\beta (\mu_1, \hat{g})\eta(d\mu_1|\mu', \Theta_{\mu'})\bar{\pi}_j(d\mu') - \beta \int_{A_j} \int_{\mu_1} K^{N,*}_\beta (\mu_1)\eta(d\mu_1|\mu', \Theta_{\mu'})\bar{\pi}_j(d\mu')
\]

\[
+ \beta \int_{A_j} \int_{\mu_1} K^{N,*}_\beta (\mu_1)\eta(d\mu_1|\mu_0, \hat{g}(\mu_0)) - \beta \int \hat{K}^{N,*}_\beta (\mu_1)\eta(d\mu_1|\mu_0, \hat{g}(\mu_0))
\]

\[
\leq 2K_cL_X + \beta \sup_\mu \left| \hat{K}^{N,*}_\beta (\mu) - K^{N,*}_\beta (\mu) \right| + \beta \|K^{N,*}_\beta \|_{Lip}2K_fL_X + \beta \sup_\mu \left| K^{N}_{\beta}(\mu, \hat{g}) - K^{N,*}_\beta (\mu) \right|.
\]

Combining this bound and (14), and with an application of Proposition 1 and Lemma 4, we get

\[
(1 - \beta) \sup_\mu \left| K^{N}_{\beta}(\mu, \hat{g}) - K^{N,*}_\beta (\mu) \right| \leq 2K_cL_X + \frac{2K_c(1 + \beta)}{(1 - \beta)(1 - 2K_f\beta)}L_X + \frac{2K_c2K_f\beta}{1 - 2K_f\beta}L_X.
\]

Combining the terms, we get

\[
\sup_\mu \left| K^{N}_{\beta}(\mu, \hat{g}) - K^{N,*}_\beta (\mu) \right| \leq \frac{4K_c}{(1 - \beta)^2(1 - 2K_f\beta)}L_X.
\]
4 Further Approximations for Large Populations

So far, we have focused on a finite approximate model construction for the team problem with $N$ agents by discretizing the state space, $X$, the agents live in. Recall that the state space of the finite measure valued model, $P_N(\hat{X})$, which is the set of all empirical measures on $\hat{X}$ with length $N$, will have the size $\frac{(M+N-1)!}{(N-1)!M!}$ where $M$ is the number of quantization bins. Hence, for large populations, i.e. when $N$ is large, the problem will still be complex although the model is finite.

In this section, we will introduce two different methods to further approximate the problem when $N$ is large and present error bounds. For the first method, we will directly work with the empirical measures on $P_N(\hat{X})$ and we will select a subset of these empirical measures to further approximate the model. For the second method, rather than working with probability measures directly, we will choose a small sample from the large population, and generalize the policy computed for the small population to the full population.

4.1 Approximation via Aggregation of Empirical Measures

Recall that discretizing the state space $X$, induces a discretization of the empirical measures, $P_N(\hat{X})$, as well, where

$$A_j = \{\mu_x : \mu_{\phi(x)} = \mu_j\},$$

are the set of empirical measures grouped together. However, there are $\frac{(M+N-1)!}{(N-1)!M!}$ many such sets, and thus, we might need to further group these sets in certain situations, e.g. when $N$ is large for computational efficiency.

Note that, each $A_j$ set corresponds to an empirical measure on the finite set $\hat{X}$ with length $N$. We then choose a subset $\{\hat{\mu}_i\}_{i=1}^K \subset P_N(\hat{X})$ by aggregating the sets $A_j$’s further such that $A_{j_1} \cup A_{j_2} \cup \ldots A_{j_k}$ is mapped to some $\hat{\mu}_i$ for some $j_1, \ldots, j_k$ and $i$. For this new finite model, a cost function and a transition probability can be defined similar to the construction in Section 3.3, and thus a policy can be computed that is optimal for the new finite model. We denote the randomized agent level policy by $\hat{\gamma}^N$.

Let the new aggregate sets are denoted by $\hat{A}_1, \ldots, \hat{A}_K$, and define

$$\bar{L} = \max_i \sup_{\mu, \mu' \in \hat{A}_i} W_1(\mu, \mu')$$

(15)

Corollary 1. Suppose Assumption 1 holds and let $\hat{\gamma}$ denotes the agent level randomized policies constructed using the subset $\{\hat{\mu}_i\}_{i=1}^K \subset P_N(\hat{X})$, we then
have

\[ J_\beta^N(x_0, \gamma) - J_\beta^{N,*}(x_0) \leq \frac{4K_c}{(1 - \beta)^2(1 - 2\beta K_f)} \bar{L}. \]

**Proof.** The proof follows from identical arguments made in the proof of Theorem 3.

Note that the above result holds for any form of aggregation of the sets \( A_j \)'s, but for every different aggregation method, the bound \( \bar{L} \) will be different. We now give an example for the further aggregation that can be made on the empirical measures on the finite set \( \hat{X} \).

**Example 4.1.** Let the original state space be \( X = [0, 3] \) and let \( N = 100 \), i.e. there are 100 agents. We first define the quantization bins as \( B_1 = [0, 1), B_2 = [1, 2), B_3 = [2, 3] \). Thus, after this quantization, we can construct an MDP whose state space is the set of empirical measures on \( \{B_1, B_2, B_3\} \) with length 100. Then, there are \( \frac{102!}{99!3!} = 171700 \) many possible empirical distributions.

To further approximate this problem, we can choose a subset of the empirical measures defined on \( \{B_1, B_2, B_3\} \). For example, we can consider the empirical distributions which only have multiples on 10 agents in the quantization bins. This can be viewed as distributing 10 agents over the sets \( \{B_1, B_2, B_3\} \) and multiplying the populations by 10. We will then have \( \frac{10^3!}{9^3!3!} = 220 \) possible states.

The mapping from the original empirical measures on \( \{B_1, B_2, B_3\} \) can be chosen in different ways, which will result in different \( \bar{L} \) values. For example, one might round the population numbers in the quantization bins, e.g. for a distribution \( \left[ \frac{27}{100}, \frac{42}{100}, \frac{31}{100} \right] \) that is when we have 27 agents in \( B_1 \), 42 agents in \( B_2 \), and 31 agents in \( B_3 \), we can use \( \left[ \frac{30}{100}, \frac{40}{100}, \frac{30}{100} \right] \) by rounding the populations.

The error bound \( \bar{L} \) for this particular example can be computed offline after the full mapping between the rounded distributions and the original distributions are determined.

4.2 Approximation via Population Sampling

We now perform the approximation by choosing a small sample, say \( n \) agents, from the large population. Note that these \( n \) agents are chosen uniformly from the total of \( N \) agents, that is any combination of \( n \) agents is equally likely to be chosen to get an unbiased sample of the large population.
For the error analysis of this approximation, one needs to define the control problem for the sampled small population. To construct the control problem for the small sample, we simply repeat the constructions from Section 2 and Section 3 for \( n \) agents. In what follows, we define the control problem for completeness.

Different from the previous constructions, we will define the cost and dynamics using the agent level randomized policies, \( \gamma : X \times P_n(X) \to P_n(\hat{U}) \). For a given sub-sample of \( n \) agents, the stage wise cost function, denoted by \( k^n(\mu^n, \gamma) \) for some empirical distribution \( \mu^n \) and a randomized policy \( \gamma \), is defined by

\[
\begin{align*}
k^n(\mu^n, \gamma) := & \frac{1}{N} \sum_{i=1}^{n} \int_{\hat{U}} c(x^i, u, \mu^n) \gamma_{\mu^n}(du|x^i) \\
= & \int_{X} \int_{\hat{U}} c(x, u, \mu^n) \gamma_{\mu^n}(du|x) \mu^n(dx).
\end{align*}
\]

(16) (17)

where \( \gamma_{\mu^n} \) is the randomized policy under the empirical distribution \( \mu^n \).

Similarly, for some empirical distribution \( \mu^n \) and a randomized policy \( \gamma \), the transition model is defined as

\[
\eta^n(d\mu^n_1|\mu^n, \gamma) := \int_{\hat{U}^n} Pr(d\mu^n_1|x, u) \gamma_{\mu^n}(du|x)
\]

(18)

for any \( x \) such that \( \mu_x = \mu^n \). Above, with an abuse of notation we use \( \gamma_{\mu^n}(du|x) := [\gamma_{\mu^n}(du_1|x^1), \ldots, \gamma_{\mu^n}(du_n|x^n)] \).

Similar to the original population construction, we define the expected cost function under a policy \( \gamma : X \times P_n(X) \to P_n(\hat{U}) \) for some initial empirical distribution \( \mu^n_0 \) by

\[
K^n_\beta(\mu^n_0, \gamma) = E_{\mu^n_0} \left[ \sum_{t=0}^{\infty} \beta^t k^n(\mu^n_t, \gamma) \right]
\]

We also define the optimal cost by

\[
K^n_{\beta, \ast}(\mu_0) = \inf_{\gamma} K^n_\beta(\mu^n_0, \gamma).
\]

For the discretized formulation, we use the same averaging we have used for the original population in Section 3.3. In particular, using the weight measures \( \hat{\pi}_j \), defined in (11), we denote the cost function by \( \hat{k}^n(\mu^n_j, \hat{\gamma}) \) such that for some \( \mu^n_j \in P_n(\hat{X}) \) and a randomized agent level policy \( \hat{\gamma} \) defined on
where $\hat{\pi}_j$ is the normalized weight measure for the set $A_j \subset P_n(\hat{X})$ of $\mu^n_j$, and $k^n$ is the cost function on continuous space as defined in (16).

We denote the transition model of the finite model by $\hat{\eta}^n$ such that

$$
\hat{\eta}^n(\mu^n_i | \mu^n_j, \hat{\gamma}) = \int_{A_j} \eta^n(A_i | \mu, \hat{\gamma}) \hat{\pi}_j(d\mu) \tag{20}
$$

where $\eta^n$ is the transition kernel on continuous space as defined in (16). The sets $A_i$ and $A_j$ are the discretization bins $\mu^n_i$ and $\mu^n_j$ belong to. That is

$$
A_j = \{ \mu \phi(x) \} = \mu^n_j \tag{21}
$$

such that, $A_j$ is the set of empirical measures on $X$ for the vector states $x$ with length $n$, which give the finite empirical measure $\mu_j$ when they are discretized.

We denote the infinite horizon discounted cost function and the optimal cost by $\hat{K}^n_{\beta}(\mu^n_j, \hat{\gamma}^n)$ and $\hat{K}^n_{\beta}^{\star}(\mu^n_j)$ for some finite $n$-sized empirical measure $\mu^n_j$ and some policy $\hat{\gamma}^n$.

### 4.2.1 Error Analysis of Sub-Sample Policies Generalized to the Full Population

We now present the main result of this section. Suppose we choose a sub sample with size $n$ of the population with size $N$, and construct a randomized policy say $\hat{\gamma}^n$ which maps the $n$-sized empirical distributions defined on the finite set $\hat{X}$ to $P(\hat{X})$ together with the states of the agents. To use these policies for the original problem, one simply looks at the empirical distribution of the sub-sampled population, decides on the randomized agent-level policy $\hat{\gamma}^n$, and every agent in the full population apply this agent level policy.

We introduce the following constant which will be used for the error upper bound:

$$
M_{N,n} := \sup_{\mu \in P(\hat{X})} E \left[ W_1(\mu^N, \mu^n) \right]
$$

which is the expectation of the first order Wasserstein distance of $N$ and $n$ sized empirical distributions defined on the discretized space $\hat{X}$ when they both come from the same probability measure $\mu$ and when we take supremum over such $\mu \in P(\hat{X})$.

The following result shows the near optimality of these policies
Theorem 4. Under Assumption 1, if we denote the policies induced by the measure valued finite model for the small population by $\hat{\gamma}^n$ and apply it in the original team problem, we will have

$$\sup_{\pi_0 \in \mathcal{P}(\mathcal{X})} E_{\pi_0} \left[ J^N_\beta(X^N_0, \hat{\gamma}^n) - J^N(N)^*(X^N_0) \right] \leq K (L_X + M_{N,n})$$

for some constant $K < \infty$, where $X^N_0$ is the vector of initial states of the full population with $N$ agents and $\pi_0$ is the initial probability measure governing the distribution of every agent.

Proof. We first fix a $\pi_0$ and a realization of $X^N_0$, say $x^N_0$. We denote the state vector of the selected $n$ agents by $x^N_0$. Note that as long as the selected agents are distributed initially with $\pi_0$, $\hat{\gamma}^n$ is going to be optimal for the discretized problem for $n$ agents. Let $\mu_{x^N_0} = \mu^N_0$ and let $\mu^N_0$ be the empirical distribution of the chosen $n$ agents. We denote by $\rho$ the map that chooses the sub-sample such that $\rho(X^N_0) = x^N_0$. We start with the following bound:

$$J^N_\beta(x^N_0, \hat{\gamma}^n) - J^N(N)^*(x^N_0) \leq \left| J^N_\beta(x^N_0, \hat{\gamma}^n) - J^N_\beta(\rho(x^N_0), \hat{\gamma}^n) \right|$$

$$+ \left| J^N_\beta(x^N_0, \hat{\gamma}^n) - J^N(N)^*(x^N_0) \right|$$

Note that the second term is bounded by Theorem 3.

We now focus on the first term. Since $\gamma^n$ is the optimal policy for the $N$ population defined on the discretized space, it induces the same randomized agent level policy for both terms, hence we can write:

$$| J^N_\beta(x^N_0, \hat{\gamma}^n) - J^N_\beta(\rho(x^N_0), \hat{\gamma}^n) | =$$

$$\int c(x, u, \mu_{x^N_0}^N) \hat{\gamma}^n(du|\phi(x))\mu_{x^N_0}^N(dx) - \int c(x, u, \mu_{x^N_0}^N) \hat{\gamma}^n(du|\phi(x))\mu_{x^N_0}^N(dx)$$

$$+ \beta \int J^N_\beta(x^N_1, \hat{\gamma}^n) Pr(dx^N_1|x^N_0, \hat{\gamma}^n) - \beta \int J^N_\beta(x^N_1, \hat{\gamma}^n) Pr(dx^N_1|x^N_0, \hat{\gamma}^n).$$

(22)

For the first term above, we write:

$$\int c(x, u, \mu_{x^N_0}^N) \hat{\gamma}^n(du|\phi(x))\mu_{x^N_0}^N(dx) - \int c(x, u, \mu_{x^N_0}^N) \hat{\gamma}^n(du|\phi(x))\mu_{x^N_0}^N(dx)$$

$$\leq KL_X + \int c(\phi(x), u, \mu_{x^N_0}^N) \hat{\gamma}^n(du|\phi(x))\mu_{x^N_0}^N(dx) - \int c(\phi(x), u, \mu_{x^N_0}^N) \hat{\gamma}^n(du|\phi(x))\mu_{x^N_0}^N(dx)$$

$$\leq KL_X + KW_1(\hat{\mu}_{x^N_0}^N, \hat{\mu}_{x^N_0}^N) + KW_1(\mu_{x^N_0}^N, \mu_{x^N_0}^N)$$

(23)

$$\leq KL_X + KW_1(\hat{\mu}_{x^N_0}^N, \hat{\mu}_{x^N_0}^N)$$

(24)
where we use a generic constant $K$. At the last step we used the inequality that $W_1(\mu_0^n, \mu_0^N) \leq KL_X + W_1(\hat{\mu}_0^n, \hat{\mu}_0^N)$, where $\hat{\mu}_0^n$ and $\hat{\mu}_0^N$ are the empirical measures on the finite set $\hat{X}$. Furthermore, above we used the fact that once we replace $x$ with the discretized version of it, $\phi(x)$, inside of the integral depends on $x$ only through the discretization and thus the integral is computed only on finitely many points and the corresponding empirical measures $\mu_0^n$ and $\mu_0^N$ can be replaced by their finite counterparts defined on $\hat{X}$ which we denote by $\hat{\mu}_0^n$ and $\hat{\mu}_0^N$. Furthermore, since the underlying space is finite, the function inside of the integral will be continuous with respect to an appropriate metric defined for the finite set, and the Lipschitz constant of the function inside of the integral will be bounded since we assume $X$ to be compact and hence bounded. Finally, the last term represents the difference between the correct empirical distribution of the full population, $\mu_0^N$, and the estimate of the empirical distribution of the full population based on the information coming from the sampled population.

Similarly, for the second term in (22), we write

$$|\beta \int J_\beta^N(x_1^N, \hat{\gamma}^n) Pr(dx_1^n|x_0^n, \hat{\gamma}^n) - \beta \int J_\beta^n(x_1^n, \hat{\gamma}^n) Pr(dx_1^n|x_0^n, \hat{\gamma}^n)|$$

$$\leq |\beta \int J_\beta^N(x_1^N, \hat{\gamma}^n) Pr(dx_1^n|x_0^n, \hat{\gamma}^n) - \beta \int J_\beta^n(\rho(x_1^n), \hat{\gamma}^n) Pr(dx_1^n|x_0^n, \hat{\gamma}^n)|$$

$$+ |\beta \int J_\beta^n(\rho(x_1^n), \hat{\gamma}^n) Pr(dx_1^n|x_0^n, \hat{\gamma}^n) - \beta \int J_\beta^n(x_1^n, \hat{\gamma}^n) Pr(dx_1^n|x_0^n, \hat{\gamma}^n)|.$$  

(25)

Note that the first term above is upper bounded by $\beta \times \sup_{X^n \in \mathcal{P}(X)} E_{\pi_0} \left[ J_\beta^N(x_0^n, \hat{\gamma}^n) - J_\beta^n(\rho(X_0^n), \hat{\gamma}^n) \right]$. We now focus on the second term:

$$|\beta \int J_\beta^n(\rho(x_1^n), \hat{\gamma}^n) Pr(dx_1^n|x_0^n, \hat{\gamma}^n) - \beta \int J_\beta^n(x_1^n, \hat{\gamma}^n) Pr(dx_1^n|x_0^n, \hat{\gamma}^n)|$$

$$\leq K \sup \left[ W_1(\mu_{\hat{X}_1^n}, \mu_{X_1^n}) \right]$$

where we used the fact the value function is Lipschitz in the empirical distributions (see Lemma 4). Furthermore, using the definition of the first order Wasserstein distance, the supremum is taken with respect to any $X_1^n \sim Pr(dx_1^n|x_0^n, \hat{\gamma}^n)$ and $\hat{X}_1^n = \rho(\hat{X}_1^n)$ such that $\hat{X}_1^n \sim Pr(dx_1^n|x_0^n, \hat{\gamma}^n)$. 

22
We can further upper-bound the last term by

\[
E \left[ W_1(\mu_{\hat{X}^n}, \mu_{X^n}) \right] \\
\leq \frac{1}{n} \sum_{i=1}^{n} \int |f(x_i^0, u, \mu_{\hat{X}^N}, w_i^0, w_0^0) - f(x_i^0, u, \mu_{X^n}, w_i^0, w_0^0)| \hat{\gamma}^n(du|\phi(x_i^0))P(dw_i^0, dw_0^0) \\
\leq KW_1(\mu_{\hat{X}^N}, \mu_{X^n}) \leq KL_{\hat{X}} + KW_1(\hat{\mu}_{\hat{X}^N}, \hat{\mu}_{X^n})
\]

(26)

where we used Lipschitz continuity of \( f \) (see Assumption 1).

Combining, (22), (23), (25) and (26) we can write

\[
\sup_{\pi_0 \in \mathcal{P}(X)} E_{\pi_0} \left[ |J_D^N(x_0^N, \gamma^n) - J_D^N(x_0^n, \hat{\gamma}^n)| \right] \leq KL_{X} + K \sup_{\pi_0 \in \mathcal{P}(X)} E_{\pi_0} \left[ W_1(\hat{\mu}_{\hat{X}^N}, \hat{\mu}_{X^n}) \right] \\
= K(L_{X} + M_{N,n})
\]

for a generic constant \( K < \infty \).

For the term \( |J_D^{\pi^*, n}(x_0^n) - J_D^N(x_0^n)| \), we can use identical arguments to find an upper bound, which completes the proof. \( \square \)

5 Q Learning for the Team Problem

In this section, we will introduce several coordinated learning algorithm for the agents to construct a near optimal team policy building on the approximate models presented earlier.

5.1 Q Learning for \( N \) Agents after Discretization of the Spaces

The MDP constructed in Section 3.3 has a finite state and a finite action space. Thus, one can apply a Q learning algorithm if the empirical distribution of the agents on the finite set \( \hat{X} \), and the cost realizations of the agents are available. Furthermore, for the learned policy, the bound presented in Theorem 3 will be valid.

We define the optimal Q values for the finite model introduced in Section 3.3, \( Q^*: \mathcal{P}_N(\hat{X}) \times \mathcal{P}_N(\hat{X} \times \hat{U}) \to \mathbb{R} \). They satisfy the following fixed point equation for every \( \mu_j \in \mathcal{P}_N(\hat{X}) \) and \( \Theta_i \in U(\mu_j) \), note that \( U(\mu_j) \) is the set of admissible actions for the empirical distribution \( \mu_j \) (see (3)):

\[
Q^*(\mu_j, U_i) = \hat{k}(\mu_j, \Theta_i) + \beta \sum_{\mu_1 \in \mathcal{P}_N(\hat{X})} \hat{\eta}(\mu_1|\mu_j, \Theta_i) \min_{V \in U(\mu_j)} Q^*(\mu_j, V). \tag{27}
\]
We note that given the optimal Q values, one can construct the optimal policy for the finite model. Hence, we aim to design a coordinated algorithm that will converge to these optimal Q values.

For the learning algorithm, during the exploration, each agent will use a randomized policy, say $\gamma^i$ for the agent $i$. We do not put any other assumption other than the condition that the exploration policies are stationary. Thus, the exploration policies can use the state vector $x$ or the empirical distribution of the agents, or they can be fully randomized and not use any information to choose the actions.

The following assumption will be used to show that the Q learning algorithm that will be presented in this section converges.

**Assumption 2.**

1. We let $\alpha_t(\mu_j, \Theta_i) = 0$ unless $(\phi(\mu_t), \Theta_t) = (\mu_j, \Theta_i)$. Otherwise, let
   \[
   \alpha_t(\mu_j, \Theta_i) = \frac{1}{1 + \sum_{k=0}^{t-1} 1\{\phi(\mu_k) = \mu_j, \Theta_k = \Theta_i\}}.
   \]

2. Under exploration policies, the state vector process $\{X_t\}_{t \geq 0}$ admits a unique invariant measure $\pi^*$.

3. During the exploration phase, every empirical distribution-action pair $(\mu_j, \Theta_i)$ is visited infinitely often.

**Remark 2.** One might notice that the assumptions are different from the classical conditions used to show the convergence of the Q learning algorithms (see [39, 23]). Note that, in our model, although the approximate model we have constructed is finite, the data we use to construct the algorithm comes from the original model which is continuous valued. Once we discretize this data, the Markov property no longer holds, which is a crucial assumption for the convergence of the Q learning algorithm. Hence, we need the extra stationarity assumption (2).

We claim that the below iterations converge to the optimal Q values:

\[
Q_{t+1}(\mu_j, \Theta_i) = (1 - \alpha_t(\mu_j, \Theta_i)) Q_t(\mu_j, \Theta_i) \\
+ \alpha_t(\mu_j, \Theta_i) \left( k(\mu_t, \Theta_t) + \beta \min_{V \in U(\phi(\mu_{t+1}))} Q_t(\phi(\mu_{t+1}), V) \right). \tag{28}
\]
Note that, by assumption, the learning rates, \( \alpha_t(\mu_j, \Theta_t) \)'s, are nonzero only when the realized or observed state and action variables are \((\mu_j, \Theta_t)\). For the remaining state and action pairs, the learning rates are zero, and thus the Q values remain unchanged. In other words, we only update the state and action pair hit by the given data, i.e. when \( \phi(\mu_t) = \mu_j \) and \( \Theta_t = \Theta_t \).

Before proving the main result of this section, we will summarize the learning algorithm step by step. The first part of the algorithm will be similar to the algorithm presented in Section 3.4. Recall that the algorithm presented in Section 3.4 assumes that the system model and the cost function are known and develops an approximate model. However, in this section, we will assume that the system model and the cost function are unknown, and the algorithm will estimate near optimal policies without these information.

- Choose the quantization bins \( \{B_i\}_{i=1}^M \subset \mathbb{X} \), such that \( \cup_{i=1}^M B_i = \mathbb{X} \).
- Assign a stationary and randomized policy \( \gamma^i \) to every agent for exploration.
- Collect the cost realizations from the agents, \( c(x^i_t, u^i_t, \mu_{x_t}) \), for every \( i \), at every time \( t \), and average them to find
  \[
  \frac{1}{N} \sum_{i=1}^N c(x^i_t, u^i_t, \mu_{x_t}) = k(\mu_t, \Theta_t).
  \]
  Note that we do not have the know the cost function but only the cost realizations to apply this step.
- At every time \( t \), count the number of agents in the quantization bins, \( B_i \)'s, which will imply the finite dimensional empirical measure \( \mu_j \in P_N(\hat{X}) \).
- Apply the iterations (28), by plugging in the values computed in the previous two steps, \( k(\mu_t, \Theta_t) \) and \( \mu_j \), and using the learning rates.
- From the limit Q values, compute the measure valued policies \( \hat{g} \). Then, compute the randomized policies for the agents using \( \hat{g} \), such that, for some \( \hat{\mu} \in P_N(\hat{X}) \), disintegrate \( \hat{g}(\hat{\mu}) = \Theta(\hat{d}\hat{x}, du) \in P_N(\hat{X} \times \hat{U}) \), to find \( \hat{\gamma}_{\hat{\mu}}(du|\hat{x})\hat{\mu}(d\hat{x}) = \Theta(\hat{d}\hat{x}, du) \).

**Theorem 5.** Under Assumption 2 and Assumption 1, the iterations in (28) converge to (27). The randomized policies for the agents constructed using
the limit point, say \( \hat{\gamma} \), satisfy the following bound when they are applied in the original problem:

\[
J_{N}^{\beta}(x_0, \hat{\gamma}) - J_{N}^{\ast}(x_0) \leq \frac{4K_c}{(1-\beta)^2(1-2\beta K_f)} \bar{L}.
\]

**Proof.** Note that the MDP model constructed in Section 2, has the stage-wise cost function \( \frac{1}{N} \sum_{i=1}^{N} c(x^i, u^i, \mu_x) \), and the state \( \mu_x \). Thus, the iterations (28), which uses the stage-wise cost function realizations of the original measure valued MDP, and the discrete versions of the state process of the measure valued MDP. Hence, it is a direct implication of [24, Theorem 3.1] that, under Assumption 2, the iterations converge to (27). Since, the limit Q values are the Q values of the model constructed in Section 3.3, Theorem 3 can be applied to conclude the proof. \( \square \)

### 5.2 Q learning via Aggregation of Empirical Measures

Recall that the algorithm presented in the last section will learn Q values of an MDP with a state space of size \( \frac{(M+N-1)!}{(N-1)!M!} \). Hence, one might need further approximations for an efficient learning algorithm. In this section we present an algorithm that will learn the policies of the model presented in Section 4.1. For this further approximation, we choose a subset \( \{\hat{\mu}_i\}_{i=1}^{K} \subset \mathcal{P}_N(\hat{X}) \) and assign a mapping, say \( h : \mathcal{P}_N(\hat{X}) \rightarrow \{\hat{\mu}_i\}_{i=1}^{K} \), which maps the empirical distributions of the agents on the finite set \( \hat{X} \) to the appropriate element of the chosen subset.

The following iterations are constructed for this subset:

\[
Q_{t+1}(h(\mu_j), \Theta_i) = (1 - \alpha_t(h(\mu_j), \Theta_i) Q_t(h(\mu_j), \Theta_i)
+ \alpha_t(h(\mu_j), \Theta_i) \left( k(\mu_t, \Theta_i) + \beta \min_{V \in U(h(\mu_{t+1}))} Q_t(h(\phi(\mu_{t+1})), V) \right),
\]

(29)

**Theorem 6.** Under Assumption 2 and Assumption 1, the iterations in (29) converge to the Q values of the model constructed in Section 4.1. Hence, the randomized policies for the agents constructed the limit Q values, say \( \hat{\gamma} \), satisfy the following bound when they are applied in the original problem:

\[
J_{N}^{\beta}(x_0, \hat{\gamma}) - J_{N}^{\ast}(x_0) \leq \frac{4K_c}{(1-\beta)^2(1-2\beta K_f)} \bar{L},
\]

where \( \bar{L} \) is as defined in (15).

**Proof.** The proof is identical to the proof of Theorem 5, because the model just involves a further discretization step. \( \square \)
5.3 Q Learning via Population Sampling

The approach presented in the last section, i.e. the iterations (29), reduces the number of Q values one needs to keep track of, however, to construct the iterations one still needs to keep track of the position and the cost realizations of the every agent in the population which might be costly for large populations. We now provide an algorithm that only collects the position and cost realization information from a sub-sample of the original population, say \(n\) agents. The algorithm will converge to the Q values of the \(n\)-agent team problem constructed in Section 4.2.

Since the algorithm will be identical to the one presented in Section 5.1 for \(n\)-agents, we will not present it again to avoid repetition.

**Theorem 7.** Under Assumption 2 and Assumption 1, the iterations in (28) constructed using the data of a subsample of \(n\) agents converge to the Q values of the model constructed in Section 4.2. Hence, the randomized policies for the agents constructed the limit Q values, say \(\hat{\gamma}^n\), satisfy the following bound when they are applied in the original problem:

\[
\sup_{\pi_0 \in \mathcal{P}(\mathcal{X})} E_{\pi_0} \left[ J^N_{\beta}(X^N_0, \hat{\gamma}^n) - J^{N,*}_{\beta}(X^N_0) \right] \leq K \left( L_X + M_{N,n} \right)
\]

for some constant \(K < \infty\), where \(X^N_0\) is the vector of initial states of the full population with \(N\) agents and \(\pi_0\) is the initial probability measure governing the distribution of every agent. Furthermore, we have

\[
M_{N,n} := \sup_{\mu \in \mathcal{P}(\hat{\mathcal{X}})} E \left[ W_1(\mu^N, \mu^n) \right]
\]

which is the expectation of the first order Wasserstein distance of \(N\) and \(n\) sized empirical distributions defined on the discretized space \(\hat{\mathcal{X}}\) when they both come from the same probability measure \(\mu\) and when we take supremum over such \(\mu \in \mathcal{P}(\hat{\mathcal{X}})\).

### A Proofs of Technical Results

#### A.1 Proof of Lemma 1

We will use value function approximation through value iterations. We define the sequence of functions \(v_k : \mathcal{X}^N \to \mathbb{R}\) such that

\[
v_{k+1}(x) = \inf_{u \in \tilde{U}^N} \left( c(x,u) + \beta E_{\nu_k}\left[ v_k(X_1) \big| x, u \right] \right)
\]
where \( v_0(x) = \inf_{u \in \tilde{U}} c(x, u) \).

We first note that if \( x = \{x^1, \ldots, x^N\} \) and \( x' = \{x'^1, \ldots, x'^N\} \), have the same empirical distribution, \( \mu_x \), then they can be viewed as different orderings of the same state vector. Furthermore, for an action vector \( u = \{u^1, \ldots, u^N\} \), one can construct another action vector \( u' = \{u'^1, \ldots, u'^N\} \), by reordering, such that the pairs \((x, u)\) and \((x', u')\) have the same empirical distribution. The immediate but key observation for the proof is that for the pairs \((x, u)\) and \((x', u')\), we have

\[
\begin{align*}
\mu(x, u) &= \frac{1}{N} \sum_{i=1}^{N} c(x^i, u^i, \mu_x) \\
&= \frac{1}{N} \sum_{i=1}^{N} c(x'^i, u'^i, \mu_{x'}) = \mu(x', u').
\end{align*}
\]

Also, for any noise vector \( w := \{w^0, w^1, \ldots, w^N\} \), the state vectors \( x_1 = f(x, u, \mu_x, w) \), and \( x'_1 = f(x', u', \mu_{x'}, w) \) will have the same empirical distribution since the dynamics governing every agent are identical and hence the agents are exchangeable. In other words,

\[
\mu(f(x, u, \mu_x, w)) = \mu(f(x', u', \mu_{x'}, w))
\]

if the pairs \((x, u)\) and \((x', u')\) have the same empirical distribution.

Using these observations, we now prove the result with induction. For \( v_0 \), if \( x, x' \in \mathbb{X}^N \) have the same empirical distribution, \( \mu_x = \mu_{x'} \), then we clearly have

\[
\inf_{u \in \tilde{U}} \frac{1}{N} \sum_{i=1}^{N} c(x^i, u^i, \mu_x) = \inf_{u \in \tilde{U}} \frac{1}{N} \sum_{i=1}^{N} c(x'^i, u'^i, \mu_{x'}). \]

We now assume that the claim is true for \( v_k \), i.e. if \( x, x' \in \mathbb{X}^N \) have the same empirical distribution, \( \mu_x = \mu_{x'} \), then \( v_k(x) = v_k(x') \). Note that for \( x, x' \in \mathbb{X}^N \) having the same empirical distribution, \( \mu_x = \mu_{x'} \)

\[
\begin{align*}
v_{k+1}(x) &= \inf_{u \in \tilde{U}} \left( c(x, u) + \beta E[v_k(X_1)|x, u] \right) \\
v_{k+1}(x') &= \inf_{u \in \tilde{U}} \left( c(x', u) + \beta E[v_k(X_1)|x', u] \right).
\end{align*}
\]

Suppose that \( v_{k+1}(x) < v_{k+1}(x') \), denoting the minimizer of the first equation by \( u \), (whose existence is guaranteed under Assumption 1), we construct
such that the pairs \((x, u)\) and \((x', u')\) have the same empirical distribution. We can then write

\[
v_{k+1}(x') \leq c(x', u') + \beta E[v_k(X_1)|x', u'] = c(x', u') + \beta \int v_k(x_1) P_r(dx_1|x', u')
\]

\[
= c(x', u') + \beta \int v_k(f(x', u', w)) P(dw)
\]

\[
= c(x, u) + \beta \int v_k(f(x, u, w)) P(dw)
\]

\[
= v_{k+1}(x)
\]

where \(P(dw)\) is the probability measure for the noise vector. We have used the symmetry of the cost function which implies that \(c(x', u) = c(x, u)\), and the induction step with the exchangeability of the agents to conclude that \(v_k(f(x', u', w)) = v_k(f(x, u, w))\). Note that this follows from the fact that \(f(x', u', w)\) and \(f(x, u, w)\) have the same empirical distribution, and the fact that \(v_k\) is constant over the states with same empirical distribution. Thus, we reach a contradiction. Using a symmetrical argument, we can also conclude that \(v_{k+1}(x) > v_{k+1}(x')\) is not possible either, which implies that \(v_{k+1}(x) = v_{k+1}(x')\).

Finally, we write

\[
|J_{\beta}^N(x) - J_{\beta}^N(x')| \leq |J_{\beta}^N(x) - v_k(x)| + |v_k(x) - v_k(x')| + |v_k(x') - J_{\beta}^N(x')|
\]

\[
\leq 2\|c\|_{\infty} \frac{\beta^k}{1 - \beta} \to 0,
\]

where we have used the fact that Bellman operator is a contraction under the uniform norm with modulus \(\beta\) and its fixed point is the optimal value function.

### A.2 Proof of Theorem A.1

This result has been proved in [2], however, we present a slightly different proof for completeness and because the proof method we use here will help us prove the other results in the paper.

First, note that by Lemma 1, for any \(x_0\) that satisfies \(x_0 = \mu_0\), \(J_{\beta}^*(x_0)\) has the same value. We now show that this value is also equal to the optimal value function of the measure valued MDP.

We prove the result using value iterations. In particular, we will approx-
imate, $J_{\beta}^{N,*}(x_0)$ and $K_{\beta}^{N,*}(\mu_0)$ using the iterations:

\[
v_{k+1}(x_0) = \inf_{u \in \hat{U}^N} \left\{ c(x_0, u) + \beta E[v_k(X_1)|x_0, u] \right\},
\]

\[
w_{k+1}(\mu_0) = \inf_{\Theta \in U(\mu_0)} \left\{ k(\mu_0, \Theta) + \beta E[w_k(\mu_1)|\mu_0, \Theta] \right\},
\]

where

\[
v_0(x_0) = \inf_{u \in \hat{U}^N} c(x_0, u)
\]

\[
w_0(\mu_0) = \inf_{\Theta \in U(\mu_0)} k(\mu_0, \Theta).
\]

We denote the minimizer of $v_0$ by $u^*$, and we construct $\Theta$ such that $\Theta = \mu(x_0, u^*)$. We then have that

\[
w_0(\mu_0) \leq k(\mu_0, \Theta) = \frac{1}{N} \sum_{i=1}^{N} c(x_0, u^*, \mu_0) = c(x_0, u^*) = v_0(x_0).
\]

Conversely, let the minimizer for $w_0$ be $\Theta^*$. We pick some $u$ such that $\mu(x_0, u) = \Theta^*$. We can then write

\[
v_0(x_0) \leq c(x_0, u) = \frac{1}{N} \sum_{i=1}^{N} c(x_0, u^*, \mu_0) = k(\mu_0, \Theta^*) = w_0(\mu_0).
\]

Hence, we have that $v_0(x_0) = w_0(\mu_0)$ for any $x_0$ that satisfies $x_0 = \mu_0$. We now assume that $v_k(x_0) = w_k(\mu_0)$, and study $v_{k+1}(x_0)$ and $w_{k+1}(\mu_0)$.

We denote the minimizer of $v_k$ by $u^*$, and we construct $\Theta$ such that $\Theta = \mu(x_0, u^*)$. We then have that

\[
w_{k+1}(\mu_0) \leq k(\mu_0, \Theta) + \beta E[w_k(\mu_1)|\mu_0, \Theta]
\]

\[
= k(\mu_0, \Theta) + \beta \int w_k(\mu_1) \eta(d\mu_1|\mu_0, \Theta)
\]

\[
= k(\mu_0, \Theta) + \beta \int_{\mu_1: \mu_1 = \mu_0} w_k(\mu_1) Pr(dx_1|x_0, u^*)
\]

\[
= c(x_0, u^*) + \beta \int v_k(x_1) Pr(dx_1|x_0, u^*)
\]

\[
= v_k(x_0)
\]

where for the second equality, we used the definition of $\eta$, and for the third equality we used the induction argument.

30
Conversely, we let the minimizer for \( w_k \) be \( \Theta^* \). We pick some \( u \) such that \( \mu_{x_0,u} = \Theta^* \). We can then write
\[
v_k(x_0) \leq c(x_0, u) + \beta E[v_k(X_1)|x_0, u] \\
= k(\mu_0, \Theta^*) + \beta E[w_k(\mu_1)|x_0, u] \\
= k(\mu_0, \Theta^*) + \beta E[w_k(\mu_1)|\mu_0, \Theta^*] \\
= w_k(\mu_0)
\]
where for the first equality, we used the induction argument, and for the second equality, we used the fact that the probability distribution of \( \mu_1 \) is fully determined by the empirical measure of the pair \((x_0, u)\). Hence, we have that \( v_k(x_0) = w_k(\mu_0) \) for any \( x_0 \) that satisfies \( x_0 = \mu_0 \). Then, the first part of the result is completed by noting that
\[
|J^{N,*}_\beta(x) - v_k(x)| \leq \|c\|_\infty \frac{\beta^k}{1-\beta}, \quad |K^{N,*}_\beta(\mu) - w_k(\mu)| \leq \|c\|_\infty \frac{\beta^k}{1-\beta}.
\]

For the second part of the result, we first show the construction of the randomized policy. For any given state vector \( x \in X \), by disintegrating \( \Theta = g(\mu_x) \), we can write
\[
\Theta(dx, du) = \gamma(dx|x)\mu_x(dx)
\]
where \( \gamma(\cdot|x) \in \mathcal{P}_{\mu_x}(\hat{U}) \). Agent \( i \), for any \( i \in \{1, \ldots, N\} \), with state \( x^i \), can choose an action according to \( \gamma(\cdot|x^i) \). We can then define a team policy \( \gamma(\cdot|x) = \{\gamma(\cdot|x^1), \ldots, \gamma(\cdot|x^N)\} \). By construction, the resulting team policy induces a empirical measure \( Q \) on the pair \((x, u)\).

We now define the following Bellman operators, \( T_{\gamma} \), and \( \hat{T}_g \) for admissible policies \( \gamma \) (possibly randomized) and \( g \), such that for some measurable and bounded \( v \in M_b(X^N) \) and \( w \in M_b(P_N(X)) \)
\[
T_{\gamma}(v)(x) := c(x, \gamma) + \beta E[v(X_1)|x, \gamma], \\
\hat{T}_g(w)(\mu) := k(\mu, g(\mu)) + \beta E[w(\mu_1)|\mu, g(\mu)].
\]
where \( c(x, \gamma) := \frac{1}{N} \sum_{i=1}^N \int_x c(x^i, u)\gamma(du|x^i) \) denotes the stage-wise cost function induced under the randomized policy, which by construction is equal to \( k(\mu_x, \Theta) \), and \( E[v(X_1)|x, \gamma] \) denotes the conditional expectation induced under the randomized policy \( \gamma \).

Note that these operators are contraction under the uniform bound, thus \( T_{\gamma}^k(v)(x) \to J^{N}_{\beta}(x, \gamma) \), and \( \hat{T}_g^k(w)(\mu) \to K^{N}_{\beta}(\mu, g) \), where \( T_{\gamma}^k \), \( \hat{T}_g^k \) denote the resulting operators from \( k \) consecutive application of the same operator.
Furthermore, if \( v(x) = c(x, \gamma) \) and \( w(\mu_x) = k(\mu_x, g(\mu_x)) \), by construction of \( \gamma \), using the identical arguments used for the proof of part (i), one can show that

\[
T^k_\gamma(v)(x) = \hat{T}^k_\gamma(w)(\mu_x),
\]

for all \( k \) which proves that \( J^N_\beta(x, \gamma) = K^N_\beta(\mu_x, f) \) for all \( x \in X^N \) with same empirical distribution.

For the proof of last part, we start by writing the Bellman equation for \( K^N_\beta^{*,*}(\mu_0) \):

\[
K^N_\beta^{*,*}(\mu_0) = \inf_{\Theta \in U(\mu_0)} \left\{ k(\mu_0, \Theta) + \beta E[K^N_\beta^{*,*}(\mu_1)|\mu_0, \Theta] \right\}.
\]

Under Assumption 1, the measurable selection conditions apply ([7]), and there exists some \( g : \mathcal{P}_N(X) \to \mathcal{P}_N(X \times \hat{U}) \), which achieves the minimum cost. Hence, the result follows from part (ii).

A.3 Proof of Lemma 3

Since \( \mu, \mu' \) are empirical distributions, there always exist state vectors \( x = [x^1, \ldots, x^N] \) and \( \hat{x} = [\hat{x}^1, \ldots, \hat{x}^N] \) such that \( \mu_x = \mu, \mu_{\hat{x}} = \mu' \). Furthermore, this holds for any permutation of the vectors \( x, \hat{x} \). If we denote by \( \rho(x) \) the possible permutations of \( x \), we then have \( \mu_{\rho(x)} = \mu \) for every different permutation \( \rho \).

The first order Wasserstein distance between \( \mu \) and \( \mu' \), can be written as

\[
W_1(\mu, \mu') = \inf_{\rho} \frac{1}{N} \sum_{i=1}^{N} |x^i - \hat{x}^i| \quad \text{where the infimum is over all couplings of } \mu \text{ and } \mu'.
\]

The last equality follows as \( \mu \) and \( \mu' \) are the empirical measures induced by the vectors \( x, x' \) (or permutations of them). Following the last term, the coupling that achieves the minimum must concentrate on the closest pairings. Hence, we have that

\[
W_1(\mu, \mu') = \min_{\rho} \frac{1}{N} \sum_{i=1}^{N} |x^i - \rho(\hat{x}^i)|
\]

where we fix the vector \( x \) and the minimum goes over all possible permutations, \( \rho \), of the vector \( \hat{x} \). Since, there are finitely many different permutations, the minimum can be achieved. For the rest of the proof, we will use the state vectors \( x, \hat{x} \) that achieve the minimum.
For the cost function, using the fact that the empirical distribution of the control actions are identical and the assumption that $c(x, u, \mu)$ is Lipschitz in $x$ and $\mu$, we can write

$$|k(\mu, \Theta) - k(\mu', \Theta')| = \left| \frac{1}{N} \sum_{i=1}^{N} c(x^i, u^i, \mu) - c(\hat{x}^i, u^i, \mu') \right|$$

$$\leq \frac{1}{N} \sum_{i=1}^{N} K_c (|x^i - \hat{x}^i| + W_1(\mu, \mu'))$$

$$= 2K_c W_1(\mu, \mu')$$

where $x^i$'s and $\hat{x}^i$'s are the elements of the state vectors that imply the empirical measures $\mu$ and $\mu'$ and that achieve the minimum in (30).

For the transition kernel, we will use a similar argument. We first note that by construction $\eta(\cdot|\mu, \Theta) = P r(\mu_1 \in \cdot | x, u)$ for any $\mu_x = \mu$ and $\mu_{x,u} = \Theta$. We can then write

$$W_1(\eta(\cdot|\mu, \Theta), \eta(\cdot|\mu', \Theta')) = \sup_{Lip(h) \leq 1} \left| \int h(\mu_1) \eta(d\mu_1|\mu, \Theta) - \int h(\mu_1) \eta(d\mu_1|\mu', \Theta') \right|$$

$$= \sup_{Lip(h) \leq 1} \left| \int h(\mu_1(x, u, \mu, w)) P(dw) - \int h(\mu_1(\hat{x}, u, \mu', w)) P(dw) \right|$$

$$\leq \int W_1(\mu_1(x, u, \mu, w), \mu_1(\hat{x}, u, \mu', w)) \ P(dw)$$

$$= \int \sup_{Lip(h) \leq 1} \left| \frac{1}{N} \sum_{i=1}^{N} h(f(x^i, u^i, \mu, w^i)) - h(f(\hat{x}^i, u^i, \mu', w^i)) \right| \ P(dw)$$

$$\leq \frac{1}{N} \sum_{i=1}^{N} K_f (|x^i - \hat{x}^i| + W_1(\mu, \mu'))$$

$$= 2K_f W_1(\mu, \mu').$$
A.4 Proof of Lemma 4

We start by writing the Bellman equation of the value functions

\[ K_{\beta}^{N,*}(\mu) = k(\mu, \Theta) + \beta \int_{\mu_1} K_{\beta}^{N,*}(\mu_1) \eta(d\mu_1 | \mu, \Theta), \]

\[ K_{\beta}^{N,*}(\mu') = k(\mu', \Theta') + \beta \int_{\mu_1} K_{\beta}^{N,*}(\mu_1) \eta(d\mu_1 | \mu', \Theta'), \]

where we denote the optimal actions achieving the minimum on the right side of the Bellman equation by \( \Theta \) and \( \Theta' \) respectively for \( \mu \) and \( \mu' \), note that the existence of the minimizers is guaranteed under Assumption 1.

Note that \( \mu, \mu', \Theta, \Theta' \) are empirical measures on \( \mathbb{X} \) and \( \mathbb{X} \times \hat{U} \), hence we can find \( x, x', u, u' \), vectors with length N, such that

\[ \mu(x) = \mu, \mu(x') = \mu', \]

\[ \mu(x, u) = \Theta, \mu(x', u') = \Theta' \] where \( x, x' \) are chosen in accordance with Lemma 3.

We will first assume that \( K_{\beta}^{N,*}(\mu) > K_{\beta}^{N,*}(\mu') \). Since, \( \Theta \) is the optimal action for the measure \( \mu \) if we use \( \theta = \mu(x, u') \), that is the empirical distribution of the pair \( x, u' \) where \( u' \) is the action vector resulting in \( \Theta' \), we then get a greater cost. Hence, we can write that

\[ K_{\beta}^{N,*}(\mu) - K_{\beta}^{N,*}(\mu') \leq k(\mu, \theta) + \beta \int_{\mu_1} K_{\beta}^{N,*}(\mu_1) \eta(d\mu_1 | \mu, \theta) \]

\[ - k(\mu', \Theta') - \beta \int_{\mu_1} K_{\beta}^{N,*}(\mu_1) \eta(d\mu_1 | \mu', \Theta') \]

\[ \leq 2K_c W_1(\mu, \mu') + 2K_f \beta \| K_{\beta}^{N,*} \|_{Lip} W_1(\mu, \mu'), \]

where the last inequality follows from Lemma 3. Hence, by rearranging the terms, we get

\[ K_{\beta}^{N,*}(\mu) - K_{\beta}^{N,*}(\mu') \leq \frac{2K_c}{1 - 2K_f \beta} W_1(\mu, \mu'). \]

The case \( K_{\beta}^{N,*}(\mu) < K_{\beta}^{N,*}(\mu') \) follows from identical steps.

A.5 Proof of Proposition 1

We first note that for some \( \mu_j \in \mathcal{P}(\hat{X}) \) the Bellman equation for the finite model value function can be written as

\[ \hat{K}_{\beta}^{N,*}(\mu_j) = \inf_{\hat{\Theta}} \left\{ \hat{k}(\mu_j, \hat{\Theta}) + \beta \sum_{\mu_1} \hat{K}_{\beta}^{N,*}(\mu_1) \hat{\eta}(\mu_1 | \mu_j, \hat{\Theta}) \right\}. \]
If we denote the minimizer for the right hand side by $\hat{\Theta}^*$, then when the value function is extended over $\mathcal{P}_N(X)$, by making it constant over the subsets $A_j$, we can write for any $\mu \in A_j$

$$\hat{K}^{N,*}_\beta(\mu) = \int_{A_j} k(\mu', \Theta_{\mu'}) \hat{\pi}_j(d\mu') + \beta \int_{A_j} \int_{\mu_1} \hat{K}^{N,*}_\beta(\mu_1) \eta(d\mu_1 | \mu', \Theta_{\mu'}) \hat{\pi}_j(d\mu'),$$

where for $\hat{\Theta}^*(dx, du) = \gamma(du|x)\mu_j(dx)$, $\Theta_{\mu'}(dx, du) = \gamma(du|\phi(x))\mu'(dx)$.

Furthermore, the Bellman equation for the value function of the original model, $K^{N,*}_\beta(\mu)$, is

$$K^{N,*}_\beta(\mu) = k(\mu, \Theta^*) + \beta \int_{\mu_1} K^{N,*}_\beta(\mu_1) \eta(d\mu_1 | \mu, \Theta^*),$$

where we used $\Theta^*$ for the minimizer of the right side.

We first assume that $\hat{K}^{N,*}_\beta(\mu) > K^{N,*}_\beta(\mu)$. Note that the minimizer for the finite model is $\hat{\Theta}^* \in \mathcal{P}_N(\bar{X} \times \bar{U})$, and we can find state and action vectors $\hat{x}, \hat{u}$ such that $\mu(\hat{x}, \hat{u}) = \hat{\Theta}^*$. The minimizer for the original model is $\Theta^* \in \mathcal{P}_N(X \times U)$, and we can find state and action vectors $(x, u)$ (in accordance with Lemma 3) such that $\mu(x, u) = \Theta^*$. Furthermore, if we use any other empirical measure on the space $\bar{X} \times \bar{U}$ we will get a greater cost for the finite model. In particular, we define $\hat{\Theta} \in \mathcal{P}_N(\bar{X} \times \bar{U})$, such that $\hat{\Theta} = \mu(\hat{x}, \hat{u})$, that is the marginal empirical distribution of the states on the finite set $\hat{X}$ stays the same, however, the marginal empirical distribution of the actions are chosen to be the same as the ones coming from the minimizer action of the original model. By construction of $\Theta_{\mu'}$, we can also find $x', u$ such that $\phi(x') = \hat{x}$, where $\phi$ is the discretization map, and $\mu(x', u) = \Theta_{\mu'}$. We can then write:

$$\hat{K}^{N,*}_\beta(\mu) - K^{N,*}_\beta(\mu) \leq \int_{A_j} k(\mu', \Theta_{\mu'}) \hat{\pi}_j(d\mu') + \beta \int_{A_j} \int_{\mu_1} \hat{K}^{N,*}_\beta(\mu_1) \eta(d\mu_1 | \mu', \Theta_{\mu'}) \hat{\pi}_j(d\mu') - k(\mu, \Theta^*) - \beta \int_{\mu_1} K^{N,*}_\beta(\mu_1) \eta(d\mu_1 | \mu, \Theta^*).$$

We now study the differences in the above term separately.

$$\int_{A_j} k(\mu', \Theta_{\mu'}) \hat{\pi}_j(d\mu') - k(\mu, \Theta^*)$$

$$\leq \int_{A_j} |k(\mu', \Theta_{\mu'}) - k(\mu, \Theta^*)| \hat{\pi}_j(d\mu')$$

$$\leq 2K_c \sup_{\mu' \in A_j} W_1(\mu, \mu').$$
where the last inequality follows from Lemma 3, as the empirical distributions of the control actions are identical.

For the second difference, we write

$$\beta \int_{A_j} \int_{\mu_1} \hat{K}^{N,*}_{\beta}(\mu_1) \eta(d\mu_1|\mu', \Theta_{\mu'}) \hat{\pi}_j(d\mu') - \beta \int_{\mu_1} K^{N,*}_{\beta}(\mu_1) \eta(d\mu_1|\mu, \Theta^*)$$

$$= \beta \int_{A_j} \int_{\mu_1} \hat{K}^{N,*}_{\beta}(\mu_1) \eta(d\mu_1|\mu', \Theta_{\mu'}) \hat{\pi}_j(d\mu') - \beta \int_{A_j} \int_{\mu_1} K^{N,*}_{\beta}(\mu_1) \eta(d\mu_1|\mu', \Theta_{\mu'}) \hat{\pi}_j(d\mu')$$

$$+ \beta \int_{A_j} \int_{\mu_1} \hat{K}^{N,*}_{\beta}(\mu_1) \eta(d\mu_1|\mu', \Theta_{\mu'}) \hat{\pi}_j(d\mu') - \beta \int_{\mu_1} K^{N,*}_{\beta}(\mu_1) \eta(d\mu_1|\mu, \Theta^*)$$

$$\leq \beta \sup_{\mu} \left| \hat{K}^{N,*}_{\beta}(\mu) - K^{N,*}_{\beta}(\mu) \right| + \|K^{N,*}_{\beta}\|_{\text{Lip}} \beta 2K_f \sup_{\mu' \in A_j} W_1(\mu, \mu'),$$

where the last line from Lemma 3. Thus, combining what we have so far:

$$\hat{K}^{N,*}_{\beta}(\mu) - K^{N,*}_{\beta}(\mu) \leq \frac{2K_c + 2K_f \beta \|K^{N,*}_{\beta}\|_{\text{Lip}}}{1 - \beta} \sup_{\mu' \in A_j} W_1(\mu, \mu')$$

we used Lemma 2 for the last step. The result then follows from Lemma 4.

The case $\hat{K}^{N,*}_{\beta}(\mu) < K^{N,*}_{\beta}(\mu)$ follows from almost identical steps. Hence, the proof is complete.

**References**

[1] Berkay Anahtarci, Can Deha Kariksiz, and Naci Saldi. Q-learning in regularized mean-field games. *Dynamic Games and Applications*, pages 1–29, 2022.

[2] Nicole Bäuerle. Mean field Markov decision processes. *arXiv preprint arXiv:2106.08755*, 2021.

[3] Erhan Bayraktar, Alekos Cecchin, and Prakash Chakraborty. Mean field control and finite dimensional approximation for regime-switching jump diffusions. *arXiv preprint arXiv:2109.09134*, 2021.

[4] Erhan Bayraktar, Andrea Cosso, and Huyën Pham. Randomized dynamic programming principle and Feynman-Kac representation for optimal control of Mckean-Vlasov dynamics. *Transactions of the American Mathematical Society*, 370(3):2115–2160, 2018.
[5] Erhan Bayraktar and Xin Zhang. Solvability of infinite horizon McKean-Vlasov FBSDEs in mean field control problems and games. arXiv preprint arXiv:2102.09619, 2021.

[6] Alain Bensoussan, Jens Frehse, Phillip Yam, et al. Mean field games and mean field type control theory, volume 101. Springer, 2013.

[7] Lawrence D Brown and Roger Purves. Measurable selections of extrema. The annals of statistics, pages 902–912, 1973.

[8] René Carmona and François Delarue. Probabilistic analysis of mean-field games. SIAM Journal on Control and Optimization, 51(4):2705–2734, 2013.

[9] René Carmona and Mathieu Laurière. Convergence analysis of machine learning algorithms for the numerical solution of mean field control and games i: the ergodic case. SIAM Journal on Numerical Analysis, 59(3):1455–1485, 2021.

[10] 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, 2019.

[11] Mao Fabrice Djetè, Dylan Possamaï, and Xiaolu Tan. McKean–Vlasov optimal control: the dynamic programming principle. The Annals of Probability, 50(2):791–833, 2022.

[12] 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.

[13] Massimo Fornasier, Stefano Lisini, Carlo Orrieri, and Giuseppe Savaré. Mean-field optimal control as gamma-limit of finite agent controls. European Journal of Applied Mathematics, 30(6):1153–1186, 2019.

[14] Zuyue Fu, Zhuoran Yang, Yongxin Chen, and Zhaoran Wang. Actor-critic provably finds Nash equilibria of linear-quadratic mean-field games. arXiv preprint arXiv:1910.07498, 2019.

[15] Maximilien Germain, Joseph Mikael, and Xavier Warin. Numerical resolution of McKean-Vlasov FBSDEs using neural networks. Methodology and Computing in Applied Probability, pages 1–30, 2022.
[16] Diogo A Gomes et al. Mean field games models—a brief survey. Dynamic Games and Applications, 4(2):110–154, 2014.

[17] Haotian Gu, Xin Guo, Xiaoli Wei, and Renyuan Xu. Dynamic programming principles for mean-field controls with learning. arXiv preprint arXiv:1911.07314, 2019.

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

[19] Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. Learning mean-field games. Advances in Neural Information Processing Systems, 32, 2019.

[20] Karl Hinderer. Lipschitz continuity of value functions in markovian decision processes. Mathematical Methods of Operations Research, 62(1):3–22, 2005.

[21] Minyi Huang, Peter E Caines, and Roland P Malhamé. Large-population cost-coupled LQG problems with nonuniform agents: individual-mass behavior and decentralized epsilon -Nash equilibria. IEEE transactions on automatic control, 52(9):1560–1571, 2007.

[22] Minyi Huang, Roland P Malhamé, and Peter E Caines. Large population stochastic dynamic games: closed-loop mckean-vlasov systems and the nash certainty equivalence principle. Communications in Information & Systems, 6(3):221–252, 2006.

[23] T. Jaakkola, M. I. Jordan, and S. P. Singh. On the convergence of stochastic iterative dynamic programming algorithms. Neural computation, 6(6):1185–1201, 1994.

[24] A. D. Kara, N. Saldi, and S. Yüksel. Q-learning for MDPs with general spaces: Convergence and near optimality via quantization under weak continuity. arXiv preprint arXiv:2111.06781, 2021.

[25] Daniel Lacker. Limit theory for controlled McKean–Vlasov dynamics. SIAM Journal on Control and Optimization, 55(3):1641–1672, 2017.

[26] Jean-Michel Lasry and Pierre-Louis Lions. Mean field games. Japanese journal of mathematics, 2(1):229–260, 2007.
[27] Mathieu Laurière and Olivier Pironneau. Dynamic programming for mean-field type control. *Comptes Rendus Mathematique*, 352(9):707–713, 2014.

[28] Médéric Motte and Huyên Pham. Mean-field Markov decision processes with common noise and open-loop controls. *The Annals of Applied Probability*, 32(2):1421–1458, 2022.

[29] Médéric Motte and Huyên Pham. Quantitative propagation of chaos for mean field Markov decision process with common noise. *arXiv preprint arXiv:2207.12738*, 2022.

[30] Sarah Perrin, Julien Pérolat, Mathieu Laurière, Matthieu Geist, Romuald Elie, and Olivier Pietquin. Fictitious play for mean field games: Continuous time analysis and applications. *Advances in Neural Information Processing Systems*, 33:13199–13213, 2020.

[31] Huyên Pham and Xiaoli Wei. Dynamic programming for optimal control of stochastic McKean–Vlasov dynamics. *SIAM Journal on Control and Optimization*, 55(2):1069–1101, 2017.

[32] N. Saldi, T. Linder, and S. Yüksel. *Finite Approximations in Discrete-Time Stochastic Control: Quantized Models and Asymptotic Optimality*. Springer, Cham, 2018.

[33] Naci Saldi, Tamer Basar, and Maxim Raginsky. Markov–nash equilibria in mean-field games with discounted cost. *SIAM Journal on Control and Optimization*, 56(6):4256–4287, 2018.

[34] Naci Saldi, Tamer Başar, and Maxim Raginsky. Approximate nash equilibria in partially observed stochastic games with mean-field interactions. *Mathematics of Operations Research*, 44(3):1006–1033, 2019.

[35] Sina Sanjari, Naci Saldi, and Serdar Yüksel. Optimality of independently randomized symmetric policies for exchangeable stochastic teams with infinitely many decision makers. *Mathematics of Operations Research*, 2022.

[36] Sina Sanjari and Serdar Yuksel. Optimal policies for convex symmetric stochastic dynamic teams and their mean-field limit. *SIAM Journal on Control and Optimization*, 59(2):777–804, 2021.
[37] Jayakumar Subramanian and Aditya Mahajan. Reinforcement learning in stationary mean-field games. In Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, pages 251–259, 2019.

[38] Hamidou Tembine, Quanyan Zhu, and Tamer Başar. Risk-sensitive mean-field games. IEEE Transactions on Automatic Control, 59(4):835–850, 2013.

[39] J. N. Tsitsiklis. Asynchronous stochastic approximation and q-learning. Machine Learning, 16:185–202, 1994.