INFINITE HORIZON AVERAGE COST OPTIMALITY CRITERIA FOR MEAN-FIELD CONTROL*

ERHAN BAYRAKTAR AND ALI DEVRAN KARA [†]

Abstract. We study mean-field control problems in discrete-time under the infinite horizon average cost optimality criteria. We focus on both the finite population and the infinite population setups. We show the existence of a solution to the average cost optimality equation (ACOE) and the existence of optimal stationary Markov policies for finite population problems under (i) a minorization condition that provides geometric ergodicity on the collective state process of the agents, and (ii) under standard Lipschitz continuity assumptions on the stage-wise cost and transition function of the agents when the Lipschitz constant of the transition function satisfies a certain bound. For the infinite population problem, we establish the existence of a solution to the ACOE, and the existence of optimal policies under the continuity assumptions on the cost and the transition functions. Finally, we relate the finite population and infinite population control problems: (i) we prove that the optimal value of the finite population problem converges to the optimal value of the infinite population problem as the number of agents grows to infinity; (ii) we show that the accumulation points of the finite population optimal solution corresponds to an optimal solution for the infinite population problem, and finally (iii), we show that one can use the solutions of the infinite population problem for the finite population problem symmetrically across the agents to achieve near optimal performance when the population is sufficiently large.

1. Introduction. We focus on optimal control of a team problem where the agents are only correlated through the so-called mean-field term, i.e. the distribution of the states of all agents. We aim to analyze the optimal control problem under the infinite horizon average cost criteria in discrete time.

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

$$x_{t+1}^i = f(x_t^i, u_t^i, \mu_{\mathbf{x_t}}, w_t^i, w_t^0)$$

for a measurable function f, where $\{w_t^i\}$ denotes the i.i.d. idiosyncratic noise process, and $\{w_t^0\}$ denotes the i.i.d. common noise process. Furthermore, $\mu_{\mathbf{x}}$ denotes the distribution of the agents on the state space $\mathbb{X}$ such that for a given joint state $\mathbf{x} := \{x^1, \ldots, x^N\} \in \mathbb{X}^N$

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

Equivalently, for a given realization of the common noise, w_t^0, the next state of the agent i is determined by some stochastic kernel $\mathcal{T}^{w_t^0}(\cdot | x_t^i, u_t^i, \mu_{\mathbf{x_t}})$ such that

$$\Pr\left(X_{t+1}^i \in B \,\middle|\, (X^j, U^j)_{[0,t]} = (x^j, u^j)_{[0,t]}, \forall j = 1, \ldots, N\right)$$

$$= \int \int_B \mathcal{T}^{w_t^0}(dx_{t+1}^i | x_t^i, u_t^i, \mu_{\mathbf{x_t}}) Pr(dw_t^0), B \in \mathcal{B}(\mathbb{X}), t \in \mathbb{N}, \qquad (1.1)$$

that is, the conditional probability distribution of the state of some agent i, given the state and action history of all agents, depends only on the most recent state and action of agent i, and the state distribution of other agents by conditioning on the realizations of the common noise.

For the remainder of the paper, by an abuse of notation, we will sometimes also 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

$$\mathbf{x_{t+1}} = f(\mathbf{x_t}, \mathbf{u_t}, \mathbf{w_t}).$$

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

$$c(x_t^i, u_t^i, \mu_{\mathbf{x_t}}).$$

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 $\gamma^i := \{\gamma_t^i\}_t$, where γ_t^i is a $\mathbb{U}$-valued (possibly randomized) function which is measurable with respect to the σ-algebra generated by

$$I_t = \{\mathbf{x}_0, \ldots, \mathbf{x}_t, \mathbf{u}_0, \ldots, \mathbf{u}_{t-1}\}.$$

Accordingly, an admissible *team* policy, is defined as $\gamma := \{\gamma^1, \ldots, \gamma^N\}$, where γ^i is an admissible policy for the agent i. In other words, agents share the complete information.

The objective of the agents is to minimize the following infinite horizon average cost function

$$J_\infty^N(\mathbf{x}_0, \gamma) := \limsup_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} E_\gamma \left[\mathbf{c}(\mathbf{X}_t, \mathbf{U}_t) \right]$$

where

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

The optimal cost is defined by

$$J_\infty^{N,*}(\mathbf{x}_0) := \inf_{\gamma \in \Gamma} J_\infty^N(\mathbf{x}_0, \gamma) \tag{1.2}$$

where Γ denotes the set of all admissible team policies.

We note that the value function that is achieved when agents share full information, as presented here, will be taken to be our reference point for simpler information structures.

1.1. Literature Review. Multi-agent systems attract attention due to their application flexibility to real world problems. Many problems that can be modeled as a multi-agent system may be intractable to model as a single agent system. Although they offer a practically rich and relevant model, the mathematical analysis is much more challenging compared to the single agent systems in general. The dynamics we consider in this paper represents a class of multi-agent systems which we will refer to as mean-field type multi-agent problems. For the mean-field type multi-agent systems, the population is homogeneous and weakly interacting such that the dynamics of each agent depends on the other agents only through the state distribution of all the other agents. The motivation of this formulation comes from systems where the effect of a single agent on the whole system is minimal (which is usually the case for large population systems). The mathematical tractability of the mean-field systems mainly relies on the observation that this weak coupling between the agents disappears through *propagation of chaos*.

If agents are non-cooperative, the mean-field systems in general are studied under mean-filed game theory. If the agents are cooperative, these problems are referred to as mean-field control. Mean-field game theory has been introduced independently by [39, 42]. There has been a great progress and a significant number of publications since then. For some of the work that has been done since, we cite [31, 18, 14, 55, 38, 3, 25, 28, 34, 46, 54, 49, 50] and references therein for mean-field game theory studies both in continuous and discrete time.

For the cooperative case, i.e. for mean-field control, similarly, there have been a number of studies dealing with structural as well as computational and learning-related methods. See [12, 23, 22, 43, 47, 19, 29, 11, 13, 10, 52, 51, 53, 40, 27, 45, 44, 20, 32, 33, 4] and the references therein for the study of the dynamic programming principle, learning methods, and justification of the exchangeability of agents for large (possibly infinite) agent team settings.

We note that most of the papers we have cited so far deal with either the finite horizon or the infinite horizon discounted cost criteria. The standard tools used to study these problems mostly involve the dynamic programming principle or the contraction property of the Bellman operator via the discount factor. However, the analysis of the infinite horizon average cost optimality criteria requires significantly different tools than the ones used in the study of finite horizon or discounted cost criteria. For the average cost criteria, one usually needs certain stability (or ergodicity) properties of the controlled process, to ensure the existence of a solution.

There are several works studying the mean-field game problems, i.e. the non-cooperative case, in continuous time under infinite horizon average cost criteria, see [26, 17, 5, 8, 21, 24, 30]. We note that in the game problems, as the agents are self interested, it is relatively more direct to decouple the agents with respect to their cost. However, the same is not true for the mean-field control problems as the goal is the minimization of the social optima and thus the agents are coupled via their common cost functions as well. Furthermore, the state measure flow can be viewed as an exogenous flow for the game problems, and thus the stability and the ergodicity of the agent state processes can be analyzed separately. For the control problems, the stability of the state measure flow itself needs to be analyzed.

For continuous time mean-field control problems, we refer the reader to [1, 7] for the ergodic control of Mckean-Vlasov dynamics. [1] studies a class of team control problems where the dynamics of the agents does not depend on the mean-field term, however, the stage-wise cost function of the problem is dependent on the distribution of the states of the agents. In particular, the agent dynamics are coupled only through the control. The authors establish the precise relationship between the finite population team problem, and the infinite population control problem, where the agents are decoupled and can use feedback policies with their local states. [7] considers a general model where the mean-field term can affect the dynamics via drift and diffusion. The authors do not consider the finite population problem and work directly with the limiting McKean-Vlasov dynamics, and show the existence of the unique viscosity solution for the ergodic control problem under a class of deterministic control functions which are assumed to be Lipschitz continuous in state and the mean-field term.

In discrete-time, the number of studies that focus on the average-cost optimality criteria for mean-field problems is limited compared to the continuous-time studies. Furthermore, most of the existing work focuses on the mean-field game problems. [48, 2, 16, 57, 56] are among the papers which study ergodic mean-field game problems in discrete time. [57] considers mean-field game problems with discrete spaces. [16, 56] establish mean-field equilibria using the ergodicity properties of the agent state process under a mixing type condition on the agent dynamics. [16] also assumes that the dynamics do not depend on the mean-field of the agent states. [48, 2] provide value iteration algorithms for mean-field game problems where the convergence of these iterations are shown to the mean-field equilibria under certain minorization and mixing type conditions on the agent dynamics. We note that the papers that focus on the game setup aim to establish the mean field equilibria. Therefore, they are able

to use the ergodicity properties of the dynamics by using stationary policies for the agents (independent of the mean-field term) under minorization conditions on the transition kernel. However, for mean-field control problems, the focus is mainly on establishing the existence of the global optimal performance for the team problem. In particular, the resulting measure valued control problem requires the use of policies that depend on the mean-field term (empirical distribution of the states). Hence, the minorization condition by itself may not be sufficient to establish the convergence to the stationary distributions of the collective agent states. Hence, in this paper, we make use of weak continuity assumptions on the agent dynamics without needing the ergodicity or minorization assumptions on the state process to establish the existence of optimal stationary and Markov policies.

Finally, to our knowledge, the only paper that deals with the discrete-time mean-field control problems under average cost (reward) criteria is [9]. In [9], existence of optimal stationary Markov policies has been established through the average cost optimality inequality. It is assumed that the difference between the optimal value of any two initial points is uniformly bounded over all initial points and discount factors. This assumption is usually not easy to verify using conditions on the primitive system components. Furthermore, in [9], two special cases are considered, namely (i) when the cost function only depends on the mean-field term and the dynamics are independent of the mean-field term, and there is no common noise (ii) when the cost functions only depend on the mean-field term (transitions can depend on the mean field term) and the admissible policies are Lipschitz continuous in the mean-field term. It is shown that the dynamic optimization problem reduces to a static optimization problem under either of these special settings.

Contributions. Our contributors are itemized as follows:

- In Section 2, we focus on the finite population control problem.
  - In Section 2.2, we establish the existence of a solution to the average cost optimality equation (ACOE) and the existence of optimal stationary Markov policies under a minorization condition on the agent dynamics (Assumption 2.1). We also provide a relative value iteration algorithm and its convergence with a convergence rate. However, the convergence speed diminishes as the number of agents grows and the convergence may even fail for the infinite population problems.
  - In Section 2.3, we show the existence of a solution to the ACOE and the existence of optimal policies without the mixing (or minorization) condition. Instead, we work with the standard continuity conditions used to study mean-field control problems. Namely, we assume that the stage-wise cost function and the agent transition function are Lipschitz continuous, and the state and action spaces are compact (see Assumption 2.2). We note that avoiding the use of strong ergodicity assumptions (e.g. Assumption 2.1) on the state process is a significant relaxation.
- In Section 3, we focus on the infinite population control problem, which can be modeled as a measure valued single agent control problem. Even though the problem can be seen as a Markov decision process (MDP), the standard tools used to study the average cost optimality criteria fail for the measure valued mean-field MDP problem (see Remark 3.1 for more detail). Hence, we utilize different tools:
  - In Section 3.2, we show that Assumption 2.2, that is the continuity of the transitions and the stage-wise cost function, is also sufficient for the infinite population problem to establish the existence of ACOE and the existence of optimal policies.
- Finally, in Section 4, we focus on the relationship between the finite population and the infinite population control problems.
  - In Section 4.1 we show that the optimal average cost value of the finite population problem converges to the optimal average cost value of the infinite

4

1.2. Metrics for Probability Measures. For the analysis of the technical results, we will make use of different distance notions for probability measures; total variation distance and the Wasserstein metric.

For some Polish, a separable completely metrizable topological space $\mathbb{X}$ and for probability measures $\mu, \nu \in \mathcal{P}(\mathbb{X})$, the *total variation* metric is given by

$$\|\mu - \nu\|_{TV} = 2 \sup_{B \in \mathcal{B}(\mathbb{X})} |\mu(B) - \nu(B)| = \sup_{f:\|f\|_\infty \leq 1} \left| \int f(x)\mu(\mathrm{d}x) - \int f(x)\nu(\mathrm{d}x) \right|,$$

where the supremum is taken over all measurable real valued f such that $\|f\|_\infty = \sup_{x \in \mathbb{X}} |f(x)| \leq 1$.

The other distance notion we will use in the paper is the Kantorovich-Rubinstein distance. For two probability measures $\mu, \nu \in \mathcal{P}(\mathbb{X})$, the Kantorovich-Rubinstein distance between them can be written as

$$W_1(\mu, \nu) = \sup_{\|f\|_{Lip} \leq 1} \left| \int f(x)\mu(dx) - \int f(x)\nu(dx) \right|$$

where $\|f\|_{Lip}$ denotes the minimal Lipschitz constant of f. We note that this distance notion is equivalent to first order Wasserstein metric via the Kantorovich-Rubinstein duality. For the rest of the paper, we will sometimes refer to this metric as the Wasserstein distance.

2. Ergodic Control of Finite Populations. For the rest of the paper, we will analyze the problem by considering the controlled process to be the state distribution of the agents, rather than the state vector of the agents.

In this section, we will define an MDP for the distribution of the agents, where the control actions are the joint distribution of the state and action vectors of the agents.

We let the state space to be $\mathcal{Z} = \mathcal{P}_N(\mathbb{X})$ which is the set of all empirical measures on $\mathbb{X}$ that can be constructed using the state vectors of N-agents. In other words, for a given state vector $\mathbf{x} = \{x^1, \ldots, x^N\}$, we consider $\mu_{\mathbf{x}} \in \mathcal{P}_N(\mathbb{X})$ to be the new state variable of the team control problem.

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

$$U(\mu) = \{\Theta \in \mathcal{P}_N(\mathbb{U} \times \mathbb{X}) | \Theta(\mathbb{U}, \cdot) = \mu(\cdot)\}, \tag{2.1}$$

that is, the set of actions for a state μ, is the set of all joint empirical measures on $\mathbb{X} \times \mathbb{U}$ whose marginal on $\mathbb{X}$ coincides with μ.

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

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

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

$$Pr(\mu_{t+1} \in B | \mathbf{x_t}, \mathbf{u_t}) = \int \mathbb{1}_{\{\mu_{f(\mathbf{x_t}, \mathbf{u_t}, \mu_{\mathbf{x_t}}, \mathbf{w_t})} \in B\}} P(d\mathbf{w_t})$$

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

If any two pairs $(\mathbf{x_t}, \mathbf{u_t}), (\mathbf{x_t'}, \mathbf{u_t'})$ have the same empirical distribution Θ_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 $\mathbf{w_t}$, and $\mathbf{x_{t+1}} = f(\mathbf{x_t}, \mathbf{u_t}, \mu_{\mathbf{x_t}} \mathbf{w_t})$, where $\mu_{\mathbf{x_{t+1}}} = \mu_{t+1}$, by reordering $\mathbf{w_t}$, one can construct some $\mathbf{w_t'}$ such that $\mathbf{x_{t+1}'} = f(\mathbf{x_t'}, \mathbf{u_t'}, \mu_{\mathbf{x_t}}, \mathbf{w_t'})$, where $\mathbf{x_{t+1}'}$ is just a reordered version of $\mathbf{x_{t+1}}$, and in particular $\mu_{\mathbf{x_{t+1}}} = \mu_{\mathbf{x_{t+1}'}}$.

Since the idiosyncratic noises are identically distributed for every agent, as a result of the above discussion, for any two pairs $(\mathbf{x_t}, \mathbf{u_t}), (\mathbf{x_t'}, \mathbf{u_t'})$ with the same empirical distribution Θ_t,

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

Therefore, the empirical distributions of the agents' states μ_t, and of the joint state and actions Θ_t define a controlled Markov chain such that

$$Pr(\mu_{t+1} \in B | \mu_t, \ldots \mu_0, \Theta_t, \ldots, \Theta_0) = Pr(\mu_{t+1} \in B | \mu_t, \Theta_t)$$
$$:= \eta(B | \mu_t, \Theta_t) \tag{2.2}$$
$$= Pr(\mu_{t+1} \in B | \mathbf{x_t}, \mathbf{u_t}), \text{ for any } (\mathbf{x_t}, \mathbf{u_t}) : \mu_{(\mathbf{x_t}, \mathbf{u_t})} = \Theta_t$$

where $\eta(\cdot | \mu, \Theta) \in \mathcal{P}(\mathcal{P}_N(\mathbb{X}))$ 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

$$k(\mu, \Theta) := \int c(x, u, \mu) \Theta(du, dx) = \frac{1}{N} \sum_{i=1}^{N} c(x^i, u^i, \mu). \tag{2.3}$$

Thus, we have an MDP with state space $\mathcal{Z}$, action space $\cup_{\mu \in \mathcal{Z}} U(\mu)$, transition kernel η 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 σ-algebra generated by the information variables

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

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

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

$$K_\infty^N(\mu_0, g) = \limsup_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} E_{\mu_0}^\eta [k(\mu_t, \Theta_t)].$$

We also define the optimal cost by

$$K_\infty^{N,*}(\mu_0) = \inf_{g \in G} K_\infty^N(\mu_0, g). \tag{2.4}$$

REMARK 2.1. *We note that the measure valued control problem considered in this section, and the original team control problem are equivalent. Any admissible policy for the original team problem can be realized as an admissible policy for the measure valued MDP problem, vice versa, and they achieve the same accumulated expected cost. In particular, the optimal value of two problems coincide. For further discussion see [10, 9].*

2.1. Average Cost Optimality Equation for the Measure Valued MDP for Finite Populations. The average cost optimality equation (ACOE) is a common tool to analyze the optimality of the average cost criteria for MDPs. In particular for the model described in the last section, the ACOE is given by

$$j^* + h(\mu) = \inf_{\Theta \in U(\mu)} \left(k(\mu, \Theta) + \int_{\mathcal{P}_N(\mathbb{X})} h(\mu_1) \eta(d\mu_1 | \mu, \Theta) \right), \tag{2.5}$$

for some $h \in \mathcal{B}(\mathcal{P}_N(\mathbb{X}))$ where $B(\mathcal{P}_N(\mathbb{X}))$ denotes the set of bounded and measurable functions on $\mathcal{P}_N(\mathbb{X})$ and some constant j^* where $U(\mu)$ denotes the set of admissible actions for the empirical distribution μ. If a solution to the ACOE exists, that is if there exists some j^* and h that satisfy the ACOE, then the optimal value function is given by the constant j^*. We present the following result for completeness and for future references:

THEOREM 2.1. *[35] Suppose there exists $h \in \mathcal{B}(\mathcal{P}_N(\mathbb{X}))$ and some constant j^* such that*

$$j^* + h(\mu) = \inf_{\Theta \in U(\mu)} \left(k(\mu, \Theta) + \int_{\mathcal{P}_N(\mathbb{X})} h(\mu_1) \eta(d\mu_1 | \mu, \Theta) \right),$$

we then have

    (i) *for any initial point $\mu \in \mathcal{P}_N(\mathbb{X})$, the constant j^* is the optimal infinite horizon average cost, that is*

$$j^* = K_\infty^{N,*}(\mu) = \inf_{g \in G} K_\infty^N(\mu, g)$$

    for every $\mu \in \mathcal{P}_N(\mathbb{X})$.

    (ii) *If there exists a policy $g^* \in G$ achieving the minimum on the right hand side for every μ, then this stationary policy is an optimal policy for the average infinite horizon cost problem; that is, if g^* satisfies*

$$j^* + h(\mu) = k(\mu, g^*(\mu)) + \int_{\mathcal{P}_N(\mathbb{X})} h(\mu_1) \eta(d\mu_1 | \mu, g^*(\mu))$$

    then $K_\infty^N(\mu, g^*) = K_\infty^{N,*}(\mu)$.

Thus, if one can guarantee the existence of a solution to the ACOE (2.5), for the N-population measure valued MDP, existence of an optimal stationary policy for the distribution of the population can also be shown under appropriate measurable selection conditions. We will establish the existence of a solution to the ACOE under two different sets of assumptions. For the first case, we will assume a minorization condition on the dynamics of the agents, which will in turn give us a mixing type result for the state vector of all agents. For the second case, we will assume that the cost and transition functions of the agents are Lipschitz continuous without assuming the minorization condition.

REMARK 2.2. *We note that the solution to the ACOE may not be unique; it is clear that j^* is unique to the problem as it is equal to the optimal value; however, there may be several h that satisfy the ACOE. In fact any shifted version of h, e.g. $h' = h + M$ for some $M < \infty$ will also satisfy the ACOE.*

2.2. Existence of a Solution to ACOE under Minorization and Geometric Ergodicity. We introduce a relative value iteration approach that converges to the ACOE in (2.5). Define the operator $T : \mathcal{B}(\mathcal{P}_N(\mathbb{X})) \to \mathcal{B}(\mathcal{P}_N(\mathbb{X}))$ such that for $v \in \mathcal{B}(\mathcal{P}_N(\mathbb{X}))$

$$Tv(\mu) := \inf_{\Theta \in U(\mu)} \left(k(\mu, \Theta) + \int_{\mathcal{P}_N(\mathbb{X})} v(\mu_1) \eta(d\mu_1 | \mu, \Theta) \right), \tag{2.6}$$

Using this, we now define a relative operator, $T_0 : \mathcal{B}(\mathcal{P}_N(\mathbb{X})) \to \mathcal{B}(\mathcal{P}_N(\mathbb{X}))$, such that for some fixed $\mu_0 \in \mathcal{P}_N(\mathbb{X})$

$$T_0 v(\mu) := Tv(\mu) - Tv(\mu_0). \tag{2.7}$$

In what follows, we focus on the contraction property of the relative operator T_0.

Recall the stochastic kernel (see (1.1)), $\mathcal{T}^{w^0}(\cdot | x_t^i, u_t^i, \mu_{\mathbf{x_t}}) \in \mathcal{P}(\mathbb{X})$ for the agent dynamics.

ASSUMPTION 2.1. *There exists a non-trivial measure* $\pi(\cdot) \in \mathcal{B}(\mathbb{X})$ *with* $\pi(\mathbb{X}) > 0$, *and there exists measurable set* $B \in \mathcal{B}(\mathbb{W})$ *with positive probability,* $P(B) > 0$, *such that for any* $x_t^i, u_t^i, \mu_{\mathbf{x_t}}$ *we have*

$$\mathcal{T}^{w^0}(\cdot | x_t^i, u_t^i, \mu_{\mathbf{x_t}}) \geq \pi(\cdot)$$

for all $w^0 \in B$.

THEOREM 2.2. *Under Assumption 2.1, we have that*

$$\| T_0^k v_0 - h \|_\infty \to 0$$

for any $v_0 \in \mathcal{B}(\mathcal{P}_N(\mathbb{X}))$ *and for some* $h \in \mathcal{B}(\mathcal{P}_N(\mathbb{X}))$, *where* T_0^k *denotes the operator* T_0 *defined by (2.7) applied* k-*times. Furthermore,* h *satisfies the following ACOE*

$$j^* + h(\mu) = \inf_{\Theta \in U(\mu)} \left(k(\mu, \Theta) + \int_{\mathcal{P}_N(\mathbb{X})} h(\mu_1) \eta(d\mu_1 | \mu, \Theta) \right).$$

Proof. We define the *span* semi-norm for a function f such that

$$\| f \|_{sp} := \sup_x f(x) - \inf_x f(x).$$

We will show that the operator T_0 is a contraction under the span semi-norm.

Let f, g be measurable functions and $\mu, \mu' \in \mathcal{P}_N(\mathbb{X})$. Then by the definition of T_0, we can write:

$$| (T_0 f(\mu) - T_0 g(\mu)) - (T_0 f(\mu') - T_0 g(\mu')) |$$
$$= |(Tf(\mu) - Tf(\mu_0)) - (Tg(\mu) - Tg(\mu_0)) - (Tf(\mu') - Tf(\mu_0)) + (Tg(\mu') - Tg(\mu_0))|$$
$$= |(Tf(\mu) - Tg(\mu)) - (Tf(\mu') - Tg(\mu'))|$$
$$= \left| \inf_{\Theta \in U(\mu)} \left(k(\mu, \Theta) + \int_{\mathcal{P}_N(\mathbb{X})} f(\mu_1) \eta(d\mu_1 | \mu, \Theta) \right) \right.$$
$$- \inf_{\Theta \in U(\mu)} \left(k(\mu, \Theta) + \int_{\mathcal{P}_N(\mathbb{X})} g(\mu_1) \eta(d\mu_1 | \mu, \Theta) \right)$$
$$- \left(\inf_{\Theta \in U(\mu')} \left(k(\mu', \Theta) + \int_{\mathcal{P}_N(\mathbb{X})} f(\mu_1) \eta(d\mu_1 | \mu', \Theta) \right) \right.$$

8

$$- \inf_{\Theta \in U(\mu')} \left(k(\mu', \Theta) + \int_{\mathcal{P}_N(\mathbb{X})} g(\mu_1) \eta(d\mu_1 | \mu', \Theta) \right) \Bigg) \Bigg|.$$

Assume now without loss of generality that the first difference is greater than the second difference, and denote the ϵ-near achieving (since the existence of the minimizers is not guaranteed) action selections by $\Theta^f, \Theta^g, \Theta_0^f, \Theta_0^g$ respectively in the order of the terms in the equation. We can then find an upper-bound as:

$$|(Tf(\mu) - Tg(\mu)) - (Tf(\mu') - Tg(\mu'))|$$

$$\leq \left(k(\mu, \Theta_g) + \int_{\mathcal{P}_N(\mathbb{X})} f(\mu_1) \eta(d\mu_1 | \mu, \Theta_g) - k(\mu, \Theta_g) - \int_{\mathcal{P}_N(\mathbb{X})} g(\mu_1) \eta(d\mu_1 | \mu, \Theta_g) \right)$$

$$- \left(k(\mu', \Theta_f') + \int_{\mathcal{P}_N(\mathbb{X})} f(\mu_1) \eta(d\mu_1 | \mu', \Theta_f') - k(\mu', \Theta_f') - \int_{\mathcal{P}_N(\mathbb{X})} g(\mu_1) \eta(d\mu_1 | \mu', \Theta_f') \right) + \epsilon$$

$$\leq \int_{\mathcal{P}_N(\mathbb{X})} f(\mu_1) - g(\mu_1) \eta(d\mu_1 | \mu, \Theta_g) - \int_{\mathcal{P}_N(\mathbb{X})} f(\mu_1) - g(\mu_1) \eta(d\mu_1 | \mu', \Theta_f') + \epsilon.$$

$$(2.8)$$

For any $\mu, \mu' \in \mathcal{P}_N(\mathbb{X})$ and any $\Theta \in U(\mu), \Theta' \in U(\mu')$ there exists state and action vectors $\mathbf{x}, \mathbf{x}', \mathbf{u}, \mathbf{u}'$ such that $\mu_{\mathbf{x}} = \mu, \mu_{\mathbf{x}'} = \mu'$ and $\mu_{(\mathbf{x},\mathbf{u})} = \Theta, \mu_{(\mathbf{x}',\mathbf{u}')} = \Theta'$. Furthermore, following the definition of η (see (2.2)), we have that for any $A \in \mathcal{B}(\mathcal{P}_N(\mathbb{X}))$

$$\eta(A|\mu, \Theta) = Pr(\mu_1 \in A | \mathbf{x}, \mathbf{u}) = Pr(\{\mathbf{x_1} \in \mathbb{X} : \mu_{\mathbf{x_1}} \in A\} | \mathbf{x}, \mathbf{u})$$

$$\eta(A|\mu', \Theta') = Pr(\mu_1 \in A | \mathbf{x}', \mathbf{u}') = Pr(\{\mathbf{x_1} \in \mathbb{X} : \mu_{\mathbf{x_1}} \in A\} | \mathbf{x}', \mathbf{u}'). \quad (2.9)$$

Note that $\mathbb{X}^N$ and $\mathbb{U}^N$ are equipped with the product topology, and the vector state process $\mathbf{x}_t \in \mathbb{X}^N$ is also a controlled Markov chain, where the control variables are $\mathbf{u}_t \in \mathbb{U}^N$ are simply the action vector variables. We write the transition kernel of the vector state process as $P(\mathbf{x}_{t+1} \in \cdot | \mathbf{x_t}, \mathbf{u_t})$.

Consider $\pi(\cdot) \in \mathcal{P}(\mathbb{X})$ from Assumption 2.1, we define the product measure $\pi^N(\cdot) \in \mathcal{P}(\mathbb{X}^N)$ using $\pi(\cdot)$, that is, for any rectangular set $E = E_1 \times \cdots \times E_N$ where $E_1, \ldots, E_N \subset \mathbb{X}$, we have that

$$\pi^N(E) = \pi(E_1) \times \cdots \times \pi(E_N).$$

Furthermore, using the conditional independence of the agent states, x^j for $j = 1, \ldots, N$, given $\mathbf{x}, \mathbf{u}$ (or given $x^j, u^j, \mu_{\mathbf{x}}$), and given the common noise realization we can also write that

$$P(\mathbf{x_1} \in E | \mathbf{x}, \mathbf{u}) = \int \prod_{j=1}^{N} \mathcal{T}^{w^0}(x_1^j \in E_j | x^j, u^j, \mu_{\mathbf{x}}) P(dw^0)$$

Consider a general measurable set $A \in \mathcal{B}(\mathbb{X}^N)$, which can be written as a countable union of disjoint rectangular sets, say $\{E^i\}_i$, such that $\cup_i E^i = A$. We can then write

$$P(\mathbf{x_1} \in A | \mathbf{x}, \mathbf{u}) = \sum_i P(E^i | \mathbf{x}, \mathbf{u}) = \sum_i \int \prod_{j=1}^{N} \mathcal{T}^{w^0}(E_j^i | x^j, u^j, \mu_{\mathbf{x}}) P(dw^0)$$

$$\geq \sum_i \prod_{j=1}^{N} \pi(E_j^i) P(B) = P(B) \sum_i \pi^N(E^i) = P(B) \pi^N(A).$$

9

Note that by construction, $\pi^N(\cdot)$ is a non-trivial measure, and furthermore, $\pi^N(\mathbb{X}^N) = \pi(\mathbb{X})^N < 1$. We denote by $\hat{P}(\cdot|\mathbf{x}, \mathbf{u}) := P(\cdot|\mathbf{x}, \mathbf{u}) - P(B)\pi^N(\cdot)$. Going back to (2.8) and denoting $h(\mu) := f(\mu) - g(\mu)$, we write

$$\left(\int_{\mathcal{P}_N(\mathbb{X})} f(\mu_1) - g(\mu_1)\eta(d\mu_1|\mu, \Theta_g) - \int_{\mathcal{P}_N(\mathbb{X})} f(\mu_1) - g(\mu_1)\eta(d\mu_1|\mu', \Theta'_f) \right) + \epsilon$$

$$= \int h(\mu_{\mathbf{x_1}})P(d\mathbf{x_1}|\mathbf{x}, \mathbf{u_g}) - \int h(\mu_{\mathbf{x_1}})P(d\mathbf{x_1}|\mathbf{x}', \mathbf{u'_f}) + \epsilon$$

$$= \int h(\mu_{\mathbf{x_1}})\hat{P}(d\mathbf{x_1}|\mathbf{x}, \mathbf{u_g}) - \int h(\mu_{\mathbf{x_1}})\hat{P}(d\mathbf{x_1}|\mathbf{x}', \mathbf{u'_f}) + \epsilon$$

$$\leq \sup_\mu h(\mu)\hat{P}(\mathbb{X}^N|\mathbf{x}, \mathbf{u_g}) - \inf_\mu h(\mu)\hat{P}(\mathbb{X}^N|\mathbf{x}', \mathbf{u'_f}) + \epsilon$$

$$\leq \|f - g\|_{sp} \left(1 - P(B)\pi^N(\mathbb{X}^N) \right) + \epsilon.$$

Since μ, μ' and ϵ are arbitrary, taking $\epsilon \to 0$ shows that T_0 is a contraction under the span norm. Hence, using an extended version of Banach Fixed Point Theorem under the span semi-norm, one can show that there exists a h such that $\|T_0h - h\|_{sp} = 0$. Note that the fixed point theorem is used under the span semi-norm which is a pseudo metric. Hence, the fixed point is only unique under the quotient space that is defined by the equivalent classes of functions that are invariant under constant shifts. However, for the original space of functions, there might be several fixed points. In other words, any function v^* such that $T_0h(\mu) - h(\mu) = C$ for some constant $C < \infty$ for all μ, is a fixed point of the operator T_0 under the span semi-norm. We can also write this as

$$h(\mu) = \inf_{\Theta \in U(\mu)} \left(k(\mu, \Theta) + \int_{\mathcal{P}_N(\mathbb{X})} h(\mu_1)\eta(d\mu_1|\mu, \Theta) \right)$$

$$- \inf_{\Theta \in U(\mu_0)} \left(k(\mu_0, \Theta) + \int_{\mathcal{P}_N(\mathbb{X})} h(\mu_1)\eta(d\mu_1|\mu_0, \Theta) \right) + C.$$

Note that T_0 is constructed so that $T_0^k v_0(\mu_0) = 0$ for all k. Hence, the fixed point also satisfies $h(\mu_0) = 0$. For the above equation, we then must have that $C = 0$. Therefore, among the span-norm equivalent set of functions that T_0^k converges to under the span semi-norm, for the one with $h(\mu_0) = 0$, we have convergence under the uniform norm as well. That is $\|T_0^k v_0 - h\|_\infty \to 0$, for any $v_0 \in \mathcal{B}(\mathcal{P}_n(\mathbb{X}))$, such that $h(\mu_0) = 0$ and it satisfies the following

$$h(\mu) = \inf_{\Theta \in U(\mu)} \left(k(\mu, \Theta) + \int_{\mathcal{P}_N(\mathbb{X})} h(\mu_1)\eta(d\mu_1|\mu, \Theta) \right)$$

$$- \inf_{\Theta \in U(\mu_0)} \left(k(\mu_0, \Theta) + \int_{\mathcal{P}_N(\mathbb{X})} h(\mu_1)\eta(d\mu_1|\mu_0, \Theta) \right)$$

which is in the form of ACOE with

$$j^* = \inf_{\Theta \in U(\mu_0)} \left(k(\mu_0, \Theta) + \int_{\mathcal{P}_N(\mathbb{X})} h(\mu_1)\eta(d\mu_1|\mu_0, \Theta) \right).$$

□

REMARK 2.3. *Assumption 2.1 requires that under a set of positive probability on the common noise, an agent can go to a subset of the state space $\mathbb{X}$, with positive probability decided by the measure π no matter what the current state, action, and the empirical measure*

are. Note that π does not have to be a probability measure. However, if $\pi(\mathbb{X}) < 1$, then the contraction constant $1 - P(B)\pi(\mathbb{X})^N$ we use in the proof goes to 1 as N grows. This suggests that for large (or infinite) populations, Theorem 2.2 no longer has a useful conclusion. One way to avoid this problem is to require π to be a probability measure. In other words, if there exists a measurable set B with positive probability, $P(B) > 0$, such that for any $x_t^i, u_t^i, \mu_{\mathbf{x_t}}$ we have

$$\mathcal{T}^{w^0}(\cdot | x_t^i, u_t^i, \mu_{\mathbf{x_t}}) = \pi(\cdot)$$

for all $w^0 \in B$ where π is a probability measure, then Theorem 2.2 will still provide a contraction constant, uniform over the population N. This condition can be thought of as resetting the distribution of every agent to π under certain common noise realizations. This is clearly a strong assumption, and maybe an artificial one. Thus, in the following we will use a different set of assumptions for the analysis of ACOE which will be valid for large (and infinite) population problems as well.

REMARK 2.4. *Assumption 2.1, we use in this section, is also a sufficient condition for geometric ergodicity of the vector state process of the team under stationary policies. Consider two team state processes under the same policy γ where one starts from some $\hat{\pi}_0$ and the other starts from π_0. Denote by $\pi_t^\gamma \in \mathcal{P}(\mathbb{X}^N)$ and $\hat{\pi}_t^\gamma \in \mathcal{P}(\mathbb{X}^N)$ the marginal distributions of the state vectors at time t. One can show that under Assumption 2.1*

$$\|\pi_{t+1}^\gamma - \hat{\pi}_{t+1}^\gamma\|_{TV} \leq \alpha \|\pi_{t+1}^\gamma - \hat{\pi}_{t+1}^\gamma\|_{TV} \tag{2.10}$$

where $\alpha = (1 - P(B)\pi(\mathbb{X})^N) < 1$. In particular, one can use this relation to show that π_t^γ is a Cauchy sequence, and thus converges to some π^γ under the total variation metric, since $\mathcal{P}(\mathbb{X}^N)$ is complete under total variation. The limit point π^γ is the invariant measure of the vector state process under the stationary team policy γ. Furthermore, using (2.10), one can also show that for any starting distribution of agents

$$\|\pi_{t+1}^\gamma - \pi^\gamma\|_{TV} \leq 2\alpha^t$$

with $\alpha = (1 - P(B)\pi(\mathbb{X})^N) < 1$. Therefore, Assumption 2.1 implies the geometric ergodicity of the vector state process of the team.

2.3. Existence of a Solution to the ACOE and Optimal Policies under Continuity of Transition and Cost Functions. We have shown the existence of a solution to the ACOE and thus the existence of optimal stationary Markov policies under a minorization condition for the dynamics of the agents. However, we have also observed that this approach may fail when the number of agents is large.

In this section, we will show that a solution to the ACOE can be established using the vanishing discount approach under the assumption that the transition and cost functions of the agents are Lipschitz continuous. Furthermore, the approach presented in this section will be valid for large populations, and in particular for the infinite population problems as we will see later.

ASSUMPTION 2.2.
  i. $\mathbb{X}$ *and* $\mathbb{U}$ *are compact.*
  ii. f is Lipschitz in x, u, μ such that

$$|f(x, u, \mu, w^i, w^0) - f(x', u', \mu', w^i, w^0)| \leq K_f \left(|x - x'| + |u - u'| + W_1(\mu, \mu') \right)$$

  for some $2K_f < 1$, uniformly in w^i, w^0 where W_1 is the first order Wasserstein distance.

  iii. c is Lipschitz in x, u, μ 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 $K_c < \infty$.

11

THEOREM 2.3. *Under Assumption 2.2, there exists a constant j^*, and a function $h \in \mathcal{B}(\mathcal{P}_N(\mathbb{X}))$, such that*

$$j^* + h(\mu) = \inf_{\Theta \in U(\mu)} \left(k(\mu, \Theta) + \int_{\mathcal{P}_N(\mathbb{X})} h(\mu_1) \eta(d\mu_1 | \mu, \Theta) \right).$$

Furthermore, there exists a stationary and Markov policy g which achieves the infimum at the right hand side.

Proof. We start by analyzing the discounted cost criteria. Let

$$K_\beta^N(\mu, g) = \sum_{t=0}^\infty \beta^t E_\mu^g \left[k(\mu_t, \Theta_t) \right]$$

be the discounted cost under some policy g where $0 < \beta < 1$ is some discount factor. Furthermore, let $K_\beta^{N,*}(\mu)$ denote the optimal discounted cost function for the initial agent distribution $\mu \in \mathcal{P}_N(\mathbb{X})$.

We define the following relative discounted value function, for some fixed $\mu' \in \mathcal{P}_N(\mathbb{X})$:

$$h_\beta^N(\mu) := K_\beta^{N,*}(\mu) - K_\beta^{N,*}(\mu').$$

Using the Bellman equation for $K_\beta^{N,*}(\mu)$ and by rearranging some terms, one can write that

$$K_\beta^{N,*}(\mu) - K_\beta^{N,*}(\mu')$$
$$= \inf_{\Theta \in U(\mu)} \left(k(\mu, \Theta) + \beta \int \left(K_\beta^{N,*}(\mu_1) - K_\beta^{N,*}(\mu') \right) \eta(d\mu_1 | \mu, \Theta) - (1 - \beta) K_\beta^{N,*}(\mu') \right).$$

Using the definition h_β^N function, we can also write that

$$h_\beta^N(\mu) = \inf_{\Theta \in U(\mu)} \left(k(\mu, \Theta) + \beta \int h_\beta^N(\mu_1) \eta(d\mu_1 | \mu, \Theta) - (1 - \beta) K_\beta^{N,*}(\mu') \right) \quad (2.11)$$

Next, we will show that h_β^N is uniformly bounded and equicontinuous (over β) when we metrize $\mathcal{P}_N(\mathbb{X})$ with the first order Wasserstein distance. Thus we can invoke the Arzela-Ascoli Theorem since $\mathcal{P}_N(\mathbb{X})$ is complete and separable under the Wasserstein metric.

For the boundedness:

$$\left| h_\beta^N(\mu) \right| = \left| K_\beta^{N,*}(\mu) - K_\beta^{N,*}(\mu') \right| \leq \frac{2 K_c}{1 - 2 K_f \beta} W_1(\mu, \mu')$$

where we used [10, Lemma 4] for the last step. Furthermore, since we assume that $\mathbb{X}$ is compact (and bounded), $W_1(\mu, \mu')$ is uniformly bounded as well. Hence, $h_\beta^N(\mu)$ is uniformly bounded over β when $2 K_f < 1$.

For equicontinuity, similarly, for some $\mu, \hat{\mu}$ we write

$$\left| h_\beta^N(\mu) - h_\beta^N(\hat{\mu}) \right| = \left| K_\beta^{N,*}(\mu) - K_\beta^{N,*}(\mu') - K_\beta^{N,*}(\hat{\mu}) + K_\beta^{N,*}(\mu') \right| \leq \frac{2 K_c}{1 - 2 K_f \beta} W_1(\mu, \hat{\mu})$$
$$(2.12)$$

which proves that $h_\beta^N(\mu)$ is equicontinuous over β if $2 K_f < 1$, as the bound becomes uniform over β.

Furthermore, the stage-wise cost function is uniformly bounded as c is continuous and $\mathbb{X}$ is compact. Hence, for the fixed μ', $(1 - \beta) K_\beta^{N,*}(\mu')$ is a bounded sequence over β. Thus,

there exists a sequence of discount factors, say $\beta(k) \to 1$, such that, $(1 - \beta(k)) K_{\beta(k)}^{N,*}(\mu') \to j^*$ for some constant j^*. If we consider some other $\mu \in \mathcal{P}_N(\mathbb{X})$, we can also show that

$$
\begin{aligned}
&|(1 - \beta(k)) K_{\beta(k)}^{N,*}(\mu) - j^*| \\
&\leq |(1 - \beta(k)) K_{\beta(k)}^{N,*}(\mu) - (1 - \beta(k)) K_{\beta(k)}^{N,*}(\mu')| + |(1 - \beta(k)) K_{\beta(k)}^{N,*}(\mu') - j^*| \\
&\leq (1 - \beta(k)) \left| h_{\beta(k)}^N(\mu) \right| + |(1 - \beta(k)) K_{\beta(k)}^{N,*}(\mu') - j^*| \to 0
\end{aligned}
$$

where we used the boundedness of h_β^N and the fact that along the sequence $\beta(k)$, $|(1 - \beta(k)) K_{\beta(k)}^{N,*}(\mu') - j^*| \to 0$. Hence, not only for the fixed μ' but for all $\mu \in \mathcal{P}_N(\mathbb{X})$, we have that $|(1 - \beta(k)) K_{\beta(k)}^{N,*}(\mu) - j^*| \to 0$.

We have shown that h_β^N is equicontinuous, hence there exists a further subsequence, say $\beta(k')$, such that $h_{\beta(k')}^N \to h$ uniformly, for some continuous function h, using the Arzela-Ascoli Theorem. If we take the limit of both sides in (2.11) along $\beta(k')$, we can see

$$
h(\mu) = \lim_{k' \to \infty} \inf_{\Theta \in U(\mu)} \left(k(\mu, \Theta) + \beta(k') \int h_{\beta(k')}^N(\mu_1) \eta(d\mu_1 | \mu, \Theta) \right) - j^*
$$

One can also justify the exchange of infimum and the limit using the equicontinuity of h_β^N, compactness of $\mathbb{X}$, and thus the weakly pre-compactness of $\mathcal{P}_N(\mathbb{X})$, see [58, Theorem 7.3.3], once we exchange the limit and the infimum, we get

$$
h(\mu) = \inf_{\Theta \in U(\mu)} \left(k(\mu, \Theta) + \int h(\mu_1) \eta(d\mu_1 | \mu, \Theta) \right) - j^*
$$

which proves the first part. The second part follows as the measurable selection conditions are satisfied under Assumption 2.2, since the limit function h, and the cost function k are continuous, η is weakly continuous and the action set $U(\mu)$ is compact and lower continuous. □

The proof method used for the previous result reveals an important implication. Along any sequence (not just a subsequence) of vanishing discount factors, a normalized version of the discounted optimal value function converges to the average cost optimal value function.

PROPOSITION 2.4. *Under Assumption 2.2, we have that for any $\mu \in \mathcal{P}_N(\mathbb{X})$*

$$
\lim_{\beta \to 1} (1 - \beta) K_\beta^{N,*}(\mu) = K_\infty^{N,*}(\mu) = j^*.
$$

Proof. Note that we know from the proof of Theorem 2.3 that for any $\beta \to 1$, there exists a subsequence $\beta(k) \to 1$ such that

$$
\lim_{\beta(k) \to 1} (1 - \beta(k)) K_\beta^{N,*}(\mu) = j^*
$$

where j^* is the optimal value function for the average cost infinite horizon problem. The result then follows, since j^*, i.e. the optimal value function, is unique under Assumption 2.2 and thus any convergent subsequence should converge to the same value. □

3. Ergodic Control of Mean Field Limit Problem. In Section 2, we have focused on solutions to the problem of a finite population team, for which the resulting policy is a mapping $g : \mathcal{P}_N(\mathbb{X}) \to \mathcal{P}_N(\mathbb{U} \times \mathbb{X})$. In other words, the policy provides a distribution rule for the joint state-action of the agents, by looking at the state distribution of the agents. Note that this policy is a recipe for the team, however, to apply this policy at the agent level, one needs to coordinate the agents in order to realize the recipe state-action distribution. For

example, if the distribution of the agents is given by $\mu \in \mathcal{P}_N(\mathbb{X})$, and the optimal action is given by $\Theta(du, dx) \in \mathcal{P}_N(\mathbb{U} \times \mathbb{X})$, then $\Theta(du, dx)$ serves as a recipe for the team of agents. The agents need to coordinate to create this joint distribution among each other. In particular, every agent may be required to apply different, asymmetric policies, even if their states are the same, i.e. the agent level policies might depend on the identity of the agents. It might be the case for example, two agents $i \neq j$ with same state $x^i = x^j$, might need to apply different policies $\gamma^i(\cdot|x^i, \mu_{\mathbf{x}}) \neq \gamma^j(\cdot|x^j, \mu_{\mathbf{x}})$, see Example 4.1. Asymmetric policies, especially for the large number of agents, can be hard to coordinate.

To deal with this coordination challenge, the usual approach is to consider the infinite population problem. The solution of the infinite population problem which we will introduce in this section, provides a policy, say $g : \mathcal{P}(\mathbb{X}) \to \mathcal{P}(\mathbb{U} \times \mathbb{X})$, that maps the marginal distribution of a representative agent to a joint state-action measure. For example, if the state is $\mu \in \mathcal{P}(\mathbb{X})$, and the recipe action is some $\Theta(du, dx) \in \mathcal{P}(\mathbb{U} \times \mathbb{X})$, then by disintegrating, $\Theta(du, dx) = \gamma(du|x, \mu)\mu(dx)$, every agent can apply the agent level symmetric policy $\gamma(\cdot|x, \mu)$, which solves the coordination challenge.

In this section, we consider the control problem for the infinite population of agents, i.e. for $N \to \infty$. We will first provide a control problem for a single agent, and formulate it as a controlled measure valued Markov process.

3.1. Infinite Population Measure Valued Control Problem. We present a control problem for the measure process μ_t, by changing the control variables as well. We will let the control actions be the joint measures on $\mathbb{U} \times \mathbb{X}$. Let $\Theta_t \in \mathcal{P}(\mathbb{U} \times \mathbb{X})$ such that the first marginal of Θ_t agrees with the state distribution of the agent μ_t. We define the stage-wise cost function $k(\mu_t, \Theta_t)$ similar to the finite population measure valued MDP construction so that

$$k(\mu_t, \Theta_t) := \int c(x, u, \mu_t)\Theta(du, dx).$$

An admissible policy $g = \{g_t\}$ is a sequence of control functions such that each can use the information variables $\{\mu_0, \dots, \mu_t, \Theta_0, \dots, \Theta_{t-1}\}$. Let G be the set of all admissible policies for the measure valued process.

Recall the transition kernel $\mathcal{T}^{w^0}(\cdot|x, u, \mu)$ defined in (1.1), which defines a probability measure on $\mathbb{X}$ given the last state, action, and measure variables (x, u, μ) and the common noise w^0. We can define the dynamics of the measure process μ_t using $\mathcal{T}^{w^0}$ so that

$$\mu_{t+1}(\cdot) = F(\mu_t, \Theta_t, w^0) := \int \mathcal{T}^{w^0}(\cdot|x, u, \mu_t)\Theta(dx, du).$$

Note that the dynamics can also be represented using a Markov kernel, which we will denote by

$$\eta(d\mu_{t+1}|\mu_t, \Theta_t)$$

by an abuse of notation; recall that we have used the same notation to denote the transition kernel of the finite population measure valued MDP (see (2.2)). Even though these transition kernels are totally different, we use the same notation for notation simplicity.

We define the cost function to be minimized as

$$K_\infty(\mu, g) = \limsup_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} E_\mu^g [k(\mu_t, \Theta_t)]$$

where $\mu = \mu_0$.

14

We denote the optimal cost function by $K_\infty^*(\mu)$ such that

$$K_\infty^*(\mu) = \inf_{g \in G} K_\infty(\mu, g).$$

REMARK 3.1. *We note that even though the infinite population problem can be structured as a classical single agent MDP problem, the standard tools used to study the average cost optimality criteria for MDPs may not be directly applicable. The standard methods (see e.g. [36, 37, 6]) requires certain ergodicity conditions on the controlled state process, as well as mixing type conditions of the state process under total variation norm, e.g. conditions similar to Assumption 2.1. These conditions may be easy to verify for state variables living in general Euclidean state spaces, e.g. for the finite population problem as we have seen in Section 2.2. However, for the measure valued controlled process we have introduced in this section for the infinite population problem, these conditions are not applicable unless we assume very strong conditions on the system dynamics. For example, a slightly weaker version of Assumption 2.1 requires that*

$$\|\eta(\cdot|\mu, \Theta) - \eta(\cdot|\mu', \Theta')\|_{TV} \le 2\alpha \tag{3.1}$$

for some $\alpha < 1$. If we assume for simplicity that there is no common noise and consider the infinite population measure valued process, this inequality will always fail, since the process becomes deterministic, and in particular we get $\|\eta(\cdot|\mu, \Theta) - \eta(\cdot|\mu', \Theta')\|_{TV} = 2$ for any $\mu \ne \mu'$ and $\Theta \ne \Theta'$. In general, most of the tools developed to study MDPs under average cost criteria may not be applicable for deterministic systems.

With the presence of common noise, the dynamics becomes stochastic again, however, verifying the standard average cost criteria assumptions is not immediate. For example, as we have observed in Remark 2.3, in order to verify (3.1) for infinite population dynamics, a sufficient condition is that there exists a measurable set B with positive probability, $P(B) > 0$, such that for any x, u, μ we have

$$\mathcal{T}^{w^0}(\cdot|x, u, \mu) = \pi(\cdot)$$

for all $w^0 \in B$, where π is a probability measure. In words, there exists a set of common noise realizations which resets the measure flow to some $\pi(\cdot) \in \mathcal{P}(\mathbb{X})$. Obviously, this is a very restrictive assumption, and thus, in what follows, we use different tools to study the average cost optimality.

3.2. Existence of a Solution to the ACOE and Optimal Policies under Continuity of Transition and Cost Functions. We will first analyze the optimality problem, and establish the existence of stationary Markov optimal policies for the infinite population problem as in the finite population problem. We follow the same steps in Section 2.3 and establish a result parallel to Theorem 2.3. As before, we will use the vanishing discount approach. Let

$$K_\beta(\mu, g) = \sum_{t=0}^{\infty} \beta^t E_\mu^g [k(\mu_t, \Theta_t)]$$

where $0 < \beta < 1$ is some discount factor. Furthermore, let $K_\beta^*(\mu)$ denote the optimal discounted cost function.

We define the following relative discounted value function, for some fixed $\mu' \in \mathcal{P}(\mathbb{X})$:

$$h_\beta(\mu) := K_\beta^*(\mu) - K_\beta^*(\mu').$$

THEOREM 3.1. *Under Assumption 2.2, there exists a constant j^*, and a function $h \in \mathcal{B}(\mathcal{P}_N(\mathbb{X}))$, such that*

$$j^* + h(\mu) = \inf_{\Theta \in U(\mu)} \left(k(\mu, \Theta) + \int_{\mathcal{P}(\mathbb{X})} h(\mu_1) \eta(d\mu_1|\mu, \Theta) \right).$$

15

Furthermore, there exists a stationary and Markov policy g which achieves the infimum at the right hand side.

Proof. As in the proof of Theorem 2.3, we define the relative discounted value function for some fixed $\mu' \in \mathcal{P}(\mathbb{X})$:

$$h_\beta(\mu) = K_\beta^*(\mu) - K_\beta^*(\mu').$$

A careful look at the proof of Theorem 2.3 shows that it is sufficient to show h_β is uniformly bounded and equicontinuous over $\beta \in (0,1)$.

For h_β, we have that

$$
\begin{aligned}
|h_\beta(\mu)| = \left| K_\beta^*(\mu) - K_\beta^*(\mu') \right| \leq \lim_{N \to \infty} & \left| K_\beta^*(\mu) - K_\beta^{N,*}(\mu^N) \right| + \left| K_\beta^{N,*}(\mu^N) - K_\beta^{N,*}(\mu^{N'}) \right| \\
& + \left| K_\beta^{N,*}(\mu^{N'}) - K_\beta^*(\mu') \right| \\
& \leq \frac{2K_c}{1 - 2K_f\beta} W_1(\mu, \mu')
\end{aligned}
\tag{3.2}
$$

where μ^N and $\mu^{N'}$ are chosen such that $\mu^N \to \mu$ and $\mu^{N'} \to \mu'$ weakly. Taking the limit $N \to \infty$, the first and last terms go to 0 by [9, 44] under Assumption 2.2, and the middle term is bounded by $\frac{2K_c}{1-2K_f\beta} W_1(\mu^N, \mu^{N'})$ as we have shown in (2.12). Hence, we have uniform boundedness together with compactness of $\mathbb{X}$. For the equicontinuity one can follow the same steps.

For the second part of the result, we have that the measurable selections conditions holds under Assumption 2.2, as the cost function $k(\mu, \Theta)$ is continuous in both variables, and the transition kernel $\eta(\cdot|\mu, \Theta)$ is weakly continuous in both variables, furthermore $U(\mu)$ is compact and lower-semi continuous. Therefore, the proof is complete. □

We now present a result, similar to Proposition 2.4. The result is about the limit of normalized discounted optimal value functions. The proof is identical to the proof of Proposition 2.4 and follows from the uniqueness of the optimal value function for the average cost criteria.

PROPOSITION 3.2. *Under Assumption 2.2, we have that for any $\mu \in \mathcal{P}(\mathbb{X})$*

$$\lim_{\beta \to 1}(1 - \beta)K_\beta^*(\mu) = K_\infty^*(\mu) = j^*.$$

4. Limit Theorems for $N \to \infty$. Finally, in this section, we study the relation between the finite population control and the infinite population control. In particular, we will show that the value function of the N- population problem converges to the value function of the infinite population problem. We will show that the accumulation points of the optimal state-action distributions for the N-population problem are optimal state-action distribution for the infinite population problem. Furthermore, we will establish that one can symmetrically use policies designed for the infinite population problem for the finite population control with near optimality if the population is sufficiently large.

4.1. Convergence of Value Functions as $N \to \infty$. THEOREM 4.1. *Under Assumption 2.2*

$$\lim_{N \to \infty} K_\infty^{N,*}(\mu^N) = K_\infty^*(\mu)$$

Proof. We will prove the result using the ACOE for the finite population and the infinite population problem. Recall that we have proved the existence of a solution to the ACOE for the finite and infinite population problems under Assumption 2.2, in Theorem 3.1 and Theorem 2.3 such that:

$$h^N(\mu^N) = \inf_{\Theta^N} \left\{ k(\mu^N, \Theta^N) + \int h^N(\mu_1^N)\eta(d\mu_1^N|\mu^N, \Theta^N) \right\} - j^N$$

16

$$h(\mu) = \inf_{\Theta} \left\{ k(\mu, \Theta) + \int h(\mu_1) \eta(d\mu_1 | \mu, \Theta) \right\} - j$$

for some measurable relative value functions h^N, h, and where j^N, j are constants that are equal to the optimal value functions, i.e. $K_\infty^{N,*}(\mu^N) = j^N$ for all μ^N and $K_\infty^*(\mu) = j$ for all μ. In particular, under Assumption 2.2, there exist optimal selectors, say f and f^N such that

$$h^N(\mu^N) = k(\mu^N, f^N(\mu^N)) + \int h^N(\mu_1^N) \eta(d\mu_1^N | \mu^N, f^N(\mu^N)) - j^N$$

$$h(\mu) = k(\mu, f(\mu)) + \int h(\mu_1) \eta(d\mu_1 | \mu, f(\mu)) - j$$

Iterating these equalities T times we get

$$h^N(\mu^N) = \sum_{t=0}^{T-1} E_{\mu^N}^{f^N} \left[k(\mu_t^N, \Theta_t^N) \right] + E_{\mu^N}^{f^N} \left[h^N(\mu_T^N) \right] - Tj^N$$

$$h(\mu) = \sum_{t=0}^{T-1} E_\mu^f \left[k(\mu_t, \Theta_t) \right] + E_\mu^f \left[h(\mu_T) \right] - Tj \qquad (4.1)$$

If we denote the finite horizon costs under some policies g^N, g by

$$K_T^N(\mu^N, g^N) = E_{\mu^N}^{g^N} \left[\sum_{t=0}^{T-1} k(\mu_t^N, \Theta_t^N) \right]$$

$$K_T(\mu, g) = E_\mu^g \left[\sum_{t=0}^{T-1} k(\mu_t, \Theta_t) \right]$$

We can then write, using (4.1)

$$K_T^N(\mu^N, f^N) - Tj^N = h^N(\mu^N) - E_{\mu^N}^{f^N} \left[h^N(\mu_T^N) \right]$$

$$K_T(\mu, f) - Tj = h(\mu) - E_\mu^f \left[h(\mu_T) \right]$$

We will later show that $\|h\|_\infty \leq M$ and $\|h^N\|_\infty \leq M$, where M is uniform over N. Hence, we can write

$$\frac{K_T^N(\mu^N, f^N)}{T} - j^N \leq \frac{2M}{T}$$

$$\frac{K_T(\mu, f)}{T} - j \leq \frac{2M}{T}$$

If we iterate the ACOE's, over some policies g, g^N (including the finite horizon optimal policies), instead of the selectors f, f^N, the equations in (4.1), can be represented as the following inequalities:

$$h^N(\mu^N) \leq \sum_{t=0}^{T-1} E_{\mu^N}^{g^N} \left[k(\mu_t^N, \Theta_t^N) \right] + E_{\mu_0^N}^{g^N} \left[h^N(\mu_T^N) \right] - Tj^N$$

$$h(\mu) \leq \sum_{t=0}^{T-1} E_\mu^g \left[k(\mu_t, \Theta_t) \right] + E_\mu^g \left[h(\mu_T) \right] - Tj$$

Using the same bound on h, h^N

$$\frac{-2M}{T} \le \frac{K_T^N(\mu^N, g^N)}{T} - j^N$$
$$\frac{-2M}{T} \le \frac{K_T(\mu, g)}{T} - j$$

As we have noted, the above lower bounds are valid for the finite horizon optimal policies as well. Furthermore, because of the optimality, we have that $\frac{K_T^{N,*}(\mu^N)}{T} \le \frac{K_T^N(\mu^N, f^N)}{T}$ and $\frac{K_T^*(\mu)}{T} \le \frac{K_T(\mu, f)}{T}$ where f and f^N are the selectors for the ACOE, i.e. the optimal policies for the infinite horizon average cost problem. Hence, we can write

$$\frac{-2M}{T} \le \frac{K_T^{N,*}(\mu^N)}{T} - j^N \le \frac{K_T^N(\mu^N, f^N)}{T} - j^N \le \frac{2M}{T} \tag{4.2}$$
$$\frac{-2M}{T} \le \frac{K_T^*(\mu)}{T} - j \le \frac{K_T(\mu, f)}{T} - j \le \frac{2M}{T}$$

Hence, if M is uniform over N, taking $T \to \infty$, we get

$$\frac{K_T^{N,*}(\mu^N)}{T} - j^N \to 0$$
$$\frac{K_T^*(\mu)}{T} - j \to 0 \tag{4.3}$$

uniformly over N.

To show that h and h^N are uniformly bounded over N, we write using (2.12) and (3.2)

$$h^N(\mu^N) = \lim_{\beta \to 1} h_\beta^N(\mu^N) \le M$$
$$h(\mu) = \lim_{\beta \to 1} h_\beta(\mu) \le M$$

where M is independent of N.

Finally, we write

$$\left| K_\infty^{N,*}(\mu^N) - K_\infty^*(\mu) \right| = \left| K_\infty^{N,*}(\mu^N) - K_\infty^*(\mu) \pm \frac{K_T^*(\mu)}{T} \pm \frac{K_T^{N,*}(\mu^N)}{T} \right|$$

$$\le \left| K_\infty^{N,*}(\mu^N) - \frac{K_T^{N,*}(\mu^N)}{T} \right| + \left| \frac{K_T^{N,*}(\mu^N)}{T} - \frac{K_T^*(\mu)}{T} \right|$$

$$+ \left| \frac{K_T^*(\mu)}{T} - K_\infty^*(\mu) \right|$$

We have shown in (4.3) that we can choose a large enough T, independent of N such that the first and the last terms are less than $\epsilon/3$ for any given $\epsilon > 0$. For the chosen ϵ and T, it has been shown in [9, 44] that the middle term goes to 0, i.e. can be made less than $\epsilon/3$, for some large $N < \infty$, under Assumption 2.2. Hence the proof is completed.
□

4.2. Accumulation to Optimal Solutions for Infinite Populations. We have established the existence of an optimal policy for the N-population control problem, say $g^N : \mathcal{P}_N(\mathbb{X}) \to \mathcal{P}_N(\mathbb{U} \times \mathbb{X})$. Under this optimal policy, we have that

$$K_\infty^N(\mu^N, g^N) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} E^{g^N} \left[\int c(x, u, \mu_t^N) \Theta_t^N(du, dx) \right] = j^N$$

18

for all μ^N where j^N is the optimal value of the problem. The dynamics for the state-action and state distributions are given in (2.2). We can also rewrite the accumulated cost under the optimal policy as

$$K_\infty^N(\mu^N, g^N) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} E^{g^N} \left[\int c(x, u, \mu_t^N) \Theta_t^N(du, dx) \right]$$

$$= \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \int k(\mu, \Theta) P_t^N(d\Theta)$$

where $P_t^N(d\Theta) \in \mathcal{P}(\mathcal{P}_N(\mathbb{U} \times \mathbb{X})) \subset \mathcal{P}(\mathcal{P}(\mathbb{U} \times \mathbb{X}))$ is the marginal distribution for the state-action distribution of the agents at time t under the policy g^N. Note also that μ is determined by Θ as it is simply the second marginal of $\Theta(du, dx)$. The next result shows that the accumulation points of P_t^N coincides with the optimal flow of the infinite population control problem as N grows.

We first present a useful lemma.

LEMMA 4.2. *Assume Assumption 2.2 holds. Consider a sequence of functions $\{g^N\}_N$ defined on $\mathcal{P}_N(\mathbb{X})$ such that*

$$g^N(\mu^N) \to g(\mu)$$

for some function g and for any $W_1(\mu^N, \mu) \to 0$. Assume further that

$$\left| g^N(\mu^N) - g^N(\hat{\mu}^N) \right| \le K W_1(\mu^N, \hat{\mu}^N)$$

for all $\mu^N, \hat{\mu}^N \in \mathcal{P}_N(\mathbb{X})$ and for some $K < \infty$ that is independent of N.

We then have for any $\mu^N \in \mathcal{P}_N(\mathbb{X}), \Theta^N \in \mathcal{P}_N(\mathbb{U} \times \mathbb{X})$ such that $W_1(\mu^N, \mu) \to 0$ and $W_1(\Theta^N, \Theta) \to 0$ for some $\mu \in \mathcal{P}(\mathbb{X})$ and $\Theta \in \mathcal{P}(\mathbb{U} \times \mathbb{X})$

$$\lim_{N \to \infty} \int g^N(\mu_1^N) \eta^N(d\mu_1^N | \mu^N, \Theta^N) = \int g(\mu_1) \eta(d\mu_1 | \mu, \Theta).$$

Proof. The proof can be found in Appendix B. □

THEOREM 4.3. *Under Assumption 2.2, we can find a subsequence, say N' such that $P_t^{N'} \to P_t$ weakly for every t for some $P_t \in \mathcal{P}(\mathcal{P}(\mathbb{U} \times \mathbb{X}))$. Furthermore, the limit flow $\{P_t\}$ is the optimal flow for the infinite population control problem such that*

$$\lim_{T \to \infty} \frac{1}{T} \int k(\mu, \Theta) P_t(d\Theta) = K_\infty^*(\mu) = j$$

Proof. We first note that since $\mathbb{U} \times \mathbb{X}$ is assumed to be compact and Polish, $\mathcal{P}(\mathbb{U} \times \mathbb{X})$ is also a compact and Polish under the weak convergence topology. In turn, $\mathcal{P}(\mathcal{P}(\mathbb{U} \times \mathbb{X}))$ is also compact and Polish.

Hence, we can find a sebsequence N_0 such that $P_0^{N_0} \to P_0$ weakly for some $P_0 \in \mathcal{P}(\mathcal{P}(\mathbb{U} \times \mathbb{X}))$. In fact, for time $t = 0$, the common noise does not affect the distributions. Thus, if the initial state distribution $\mu^N \to \mu$, then one can show that $P_0(\cdot) = \delta_{\Theta_0}(\cdot)$ for some $\Theta_0 \in \mathcal{P}(\mathbb{U} \times \mathbb{X})$ such that $\Theta_0(\mathbb{U}, \cdot) = \mu(\cdot)$.

For $P_1^{N_0}$, one can find a further subsequence N_1, such that $P_1^{N_1} \to P_1$ for some $P_1 \in \mathcal{P}(\mathcal{P}(\mathbb{U} \times \mathbb{X}))$. Continuing in this manner, and using a standard diagonal argument, we can find a subsequence N' such that $P_t^{N'} \to P_t$ weakly for all $t \ge 0$.

We now show that the limit flow $\{P_t\}_t$ is consistent with the dynamics of the infinite population dynamics. Note that any $P_t \in \mathcal{P}(\mathcal{P}(\mathbb{U} \times \mathbb{X}))$ induces a probability measure in $\mathcal{P}(\mathcal{P}(\mathbb{X}))$, say $\bar{P}_t$ such that

$$\bar{P}_t(\mu \in A) = P_t(\{\Theta : \Theta(\mathbb{U}, \cdot) \in A\})$$

19

for any $A \in \mathcal{B}(\mathcal{P}(\mathbb{X}))$. The sequence $\{P_t\} \subset \mathcal{P}(\mathcal{P}(\mathbb{U} \times \mathbb{X}))$ is consistent with the infinite population dynamics, if we have

$$\bar{P}_{t+1}(\mu_{t+1} \in A) = \int \eta(\mu_{t+1} \in A|\mu, \Theta)P_t(d\Theta).$$

For the N- population dynamics, we have that

$$\int f(\mu_{t+1})\bar{P}_{t+1}^N(d\mu_{t+1}) = \int f(\mu_{t+1})\eta^N(d\mu_{t+1}|\mu, \Theta)P_t^N(d\Theta) \qquad (4.4)$$

for any continuous and bounded function f, where we use η^N to denote the one step transition kernel of the N-population dynamics. By Lemma 4.2, for any $\mu^N, \Theta^N \to \mu, \Theta$, we can write

$$\int f(\mu_{t+1})\eta^N(d\mu_{t+1}|\mu^N, \Theta^N) \to \int f(\mu_{t+1})\eta(d\mu_{t+1}|\mu, \Theta).$$

Hence, if we take the limit of both sides in (4.4) along the chosen subsquence N', we get

$$\int f(\mu_{t+1})\bar{P}_{t+1}(d\mu_{t+1}) = \int f(\mu_{t+1})\eta(d\mu_{t+1}|\mu, \Theta)P_t(d\Theta)$$

where we used the fact that $P_t^{N'} \to P_t$ weakly for all t and [41, Theorem 3.1]. Since the class of continuous and bounded functions are measure determining (see [15, Theorem 1.2]), we can conclude that the limit sequence $\{P_t\}_t$ is consistent with the infinite population dynamics.

We now prove the second part of the result. The result is an implication of Theorem 4.1. By (4.2), we have that

$$\frac{K_T^N(\mu^N, g^N)}{T} \to j^N$$

where g^N is the optimal stationary policy for the N-population. Furthermore, the convergence is uniform over N. Thus, if we take the limit of both sides along the subsequence N', we get

$$\lim_{N' \to \infty} \lim_{T \to \infty} \frac{K_T^{N'}(\mu^{N'}, g^{N'})}{T} = \lim_{N' \to \infty} j^{N'} = j$$

$$= \lim_{T \to \infty} \lim_{N' \to \infty} \frac{K_T^{N'}(\mu^{N'}, g^{N'})}{T} = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \lim_{N' \to \infty} \int k(\mu, \Theta)P_t^{N'}(d\Theta)$$

$$= \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \int k(\mu, \Theta)P_t(d\Theta)$$

where we used the weak convergence of $P_t^{N'} \to P_t$ and the continuity of the function $k(\mu, \Theta)$ on (μ, Θ) under Assumption 2.2. This shows that

$$K_\infty^*(\mu) = j = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \int k(\mu, \Theta)P_t(d\Theta).$$

□

20

4.3. Near Optimality of Symmetric Policies for Finite Population Control. In this section, we focus on the effect of using symmetric policies for finite population control problem. The following example shows that the symmetric policies may not achieve the optimal performance, and personalized policies have to be used for the optimality.

EXAMPLE 4.1. *Consider a team control problem with two agents, i.e. $N = 2$. We assume that $\mathbb{X} = \mathbb{U} = \{0, 1\}$. The stage wise cost function of the agents is defined as*

$$c(x, u, \mu_{\mathbf{x}}) = W_1(\mu_{\mathbf{x}}, \bar{\mu})$$

where

$$\bar{\mu} = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1.$$

In words, the state distribution should be distributed equally over the state space $\{0, 1\}$ for minimal stage-wise cost. For the dynamics we assume a deterministic model such that

$$x_{t+1} = u_t.$$

In words, the action of an agent purely determines the next state of the same agent. The goal of the agents is to minimize

$$K_\infty(x_0^1, x_0^2, g^1, g^2) = \limsup_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} E^{g^1, g^2} \left[\frac{c(x_t^1, u_t^1, \mu_{\mathbf{x_t}}) + c(x_t^2, u_t^2, \mu_{\mathbf{x_t}})}{2} \right]$$

for some initial state values $\mathbf{x}_0 = [x_0^1, x_0^2]$, by choosing policies g^1, g^2. The expectation is over the possible randomization of the policies. We assume full information sharing such that every agent has access to the state and action information of the other agent.

We let the initial states be $x_0^1 = x_0^2 = 0$. An optimal policy for the agents for the problem is given by

$$g^1(0, 0) = 0, \qquad g^2(0, 0) = 1$$
$$g^1(0, 1) = 0, \qquad g^2(0, 1) = 1$$
$$g^1(1, 0) = 1, \qquad g^2(1, 0) = 0$$
$$g^1(1, 1) = 1, \qquad g^2(1, 1) = 0$$

which always spreads the agents equally over the state space. One can realize that, when the agents are positioned at either $(0, 0)$ or $(1, 1)$, they have to use personalized policies to decide on which one to be placed at 0 or 1.

For any symmetric policy $g^1(x^1, x^2) = g^2(x^1, x^2) = g(x^1, x^2)$, including the randomized ones, there will always be cases with strict positive probability, where the agents are positioned at the same state, and thus the performance will be strictly worse than the optimal performance.

We now introduce the symmetric policies that will be used by the agents for the finite population setting. We will focus on the optimal policies for infinite population under the discounted cost criteria. Let $\Theta(du, dx)$ be an optimal state-action distribution for some measure $\mu \in \mathcal{P}(\mathbb{X})$. We then write

$$\Theta(du, dx) = \gamma(du|x, \mu)\mu(dx).$$

In the finite population setting, we let the agents use the symmetric randomized policy $\gamma(du|x, \mu)$. That is, an agent i, at time t, observes their local state x_t^i and the state distribution of the agents, say μ_t^N, and applies $\gamma(du|x_t^i, \mu_t^N)$. This agent level policy then defines a team policy for the state distribution μ^N such that

$$g_\beta(\mu^N) := \Theta^N(du, dx) = \gamma(du|x, \mu^N)\mu^N(dx).$$

We use the β dependence to emphasize that the policy is constructed using the discounted cost criteria.

Our next result shows that this policy, which is symmetric between the agents, will achieve near optimal performance for sufficiently large populations under the discounted cost criteria.

PROPOSITION 4.4. *Under Assumption 2.2, for any $\mu^N \to \mu$*

$$\lim_{N \to \infty} K_\beta^N(\mu^N, g_\beta) = K_\beta^*(\mu).$$

Furthermore,

$$\lim_{N \to \infty} K_\beta^N(\mu^N, g_\beta) - K_\beta^{N,*}(\mu^N) = 0.$$

Hence, g_β will achieve near optimal performance for finite populations.

Proof. The proof can be found in Appendix A. □

THEOREM 4.5. *Under Assumption 2.2, for any given $\epsilon > 0$ we can find a $\beta < 1$ sufficiently close to 1, and some $\bar{N}$ such that for all $N > \bar{N}$ we have that*

$$\left| K_\infty^N(\mu^N, g_\beta) - j^N \right| \le \epsilon.$$

In words, the symmetric feedback policy g_β is near optimal for N-population problem if N is sufficiently large.

Proof. Recall $h_\beta(\mu) := K_\beta^*(\mu) - K_\beta^*(\mu_0)$. Using the Bellman equation for $K_\beta^*(\mu)$ we can write

$$h_\beta(\mu) = k(\mu, \Theta) + \beta \int h_\beta(\mu_1) \eta(d\mu_1 | \mu, \Theta) - (1 - \beta) K_\beta^*(\mu_0)$$

$$= k(\mu, \Theta) + \int h_\beta(\mu_1) \eta(d\mu_1 | \mu, \Theta) - (1 - \beta) \int h_\beta(\mu_1) \eta(d\mu_1 | \mu, \Theta) - (1 - \beta) K_\beta^*(\mu_0)$$

(4.5)

where Θ is an optimal state-action distribution for μ. We also recall $h_\beta^N(\mu^N) := K_\beta^{N,*}(\mu^N) - K_\beta^{N,*}(\mu_0^N)$ where $\mu_0^N \in \mathcal{P}_N(\mathbb{X})$ is chosen such that $\mu_0^N \to \mu_0$. We can then rewrite (4.5) as

$$\pm h_\beta^N(\mu^N) \pm j^N + h_\beta(\mu) = k(\mu, \Theta) + \int h_\beta(\mu_1) \eta(d\mu_1 | \mu, \Theta)$$

$$- (1 - \beta) \int h_\beta(\mu_1) \eta(d\mu_1 | \mu, \Theta) - (1 - \beta) K_\beta^*(\mu_0)$$

$$\pm \left(k(\mu^N, g_\beta(\mu^N)) + \int h_\beta^N(\mu_1^N) \eta(d\mu_1^N | \mu^N, g_\beta(\mu^N)) \right).$$

By rearranging the terms we can write

$$k(\mu^N, g_\beta^\infty(\mu^N)) + \int h_\beta^N(\mu_1^N) \eta(d\mu_1^N | \mu^N, g_\beta^\infty(\mu^N))$$

$$= j^N + h_\beta^N(\mu^N)$$

$$+ h_\beta(\mu) - h_\beta^N(\mu^N) \qquad (4.6)$$

$$+ j - j^N \qquad (4.7)$$

$$- \left(j - (1 - \beta) K_\beta^*(\mu_0) \right) + (1 - \beta) \int h_\beta(\mu_1) \eta(d\mu_1 | \mu, \Theta) \qquad (4.8)$$

22

$$+ k(\mu^N, g_\beta^\infty(\mu^N)) - k(\mu, \Theta) \qquad\qquad (4.9)$$

$$+ \int h_\beta^N(\mu_1^N)\eta(d\mu_1^N|\mu^N, g_\beta^\infty(\mu^N)) - \int h_\beta(\mu_1)\eta(d\mu_1|\mu, \Theta) \qquad (4.10)$$

We first choose β large enough (independent of N) such that (4.8) is smaller than ϵ. This can be done using Proposition 3.2, and since h_β is uniformly bounded over β (see (3.2)).

For the chosen β, we can make (4.6) smaller than ϵ, since $K_\beta^{N,*}(\mu) \to K_\beta^*(\mu)$ and $K_\beta^{N,*}(\mu_0^N) \to K_\beta^*(\mu_0)$ as $N \to \infty$ for fixed $\beta < 1$. (4.7) can be made smaller than ϵ by Theorem 4.1.

Finally we focus on (4.9) and (4.10) :

$$k(\mu^N, g_\beta(\mu^N)) + \int h_\beta^N(\mu_1^N)\eta(d\mu_1^N|\mu^N, g_\beta(\mu^N)) - k(\mu, \Theta) - \int h_\beta(\mu_1)\eta(d\mu_1|\mu, \Theta)$$

$$= k(\mu^N, g_\beta(\mu^N)) + \beta \int h_\beta^N(\mu_1^N)\eta(d\mu_1^N|\mu^N, g_\beta(\mu^N)) + (1-\beta)\int h_\beta^N(\mu_1^N)\eta(d\mu_1^N, \mu^N, g_\beta(\mu^N))$$

$$- k(\mu, \Theta) - \beta\int h_\beta(\mu_1)\eta(d\mu_1|\mu, \Theta) - (1-\beta)\int h_\beta(\mu_1)\eta(d\mu_1|\mu, \Theta)$$

$$= k(\mu^N, g_\beta(\mu^N)) + \beta\int K_\beta^{N,*}(\mu_1^N)\eta(d\mu_1^N|\mu^N, g_\beta(\mu^N)) - K_\beta^*(\mu) - \beta K_\beta^{N,*}(\mu_0^N) + \beta K_\beta^*(\mu_0)$$

$$+ (1-\beta)\int h_\beta^N(\mu_1^N)\eta(d\mu_1^N|\mu^N, g_\beta(\mu^N)) - (1-\beta)\int h_\beta(\mu_1)\eta(d\mu_1|\mu, \Theta)$$

The last line can be made smaller than ϵ by choosing β sufficiently close to 1 independent of N since h_β^N and h_β are previously shown to be uniformly bounded (uniform over N as well). Furthermore, for the chosen β, we have that $\beta K_\beta^{N,*}(\mu_0^N) \to \beta K_\beta^*(\mu_0)$ and finally we have shown in the proof of Proposition 4.4 that

$$\lim_{N\to\infty} k(\mu^N, g_\beta(\mu^N)) + \beta \int K_\beta^{N,*}(\mu_1^N)\eta(d\mu_1^N|\mu^N, g_\beta(\mu^N)) = K_\beta^*(\mu).$$

We then can make the terms (4.9) and (4.10) smaller than ϵ by choosing β and N sufficiently large.

Hence we can write the following for the chosen β and N

$$k(\mu^N, g_\beta^\infty(\mu^N)) + \int h_\beta^N \eta(d\mu_1^N|\mu^N, g_\beta^\infty(\mu^N)) \le j^N + 5\epsilon + h_\beta^N(\mu^N).$$

We note the following inequality to conclude the result (see [58, Theorem 7.1.3]): if

$$g + h(\mu^N) \ge k(\mu^N, f(\mu^N)) + \int h(\mu_1^N)\eta(d\mu_1^N|\mu^N, f(\mu^N))$$

for some functions h, f and for some constant g and if h is bounded then

$$g \ge K_\infty^N(\mu^N, f).$$

We have established this inequality with $g = j^N + 5\epsilon$, $f \equiv g_\beta$ and $h \equiv h_\beta^N$. Therefore we can conclude that

$$K_\infty^N(\mu^N, g_\beta) \le j^N + 5\epsilon$$

as h_β^N is bounded uniformly. $\square$

Appendix A. Proof of Proposition 4.4.

Consider a sequence of measures $\{\mu_n\}_n \subset \mathcal{P}(\mathbb{X})$ such that $\mu_n \to \mu$ weakly for some $\mu \in \mathcal{P}(\mathbb{X})$. Furthermore, consider the corresponding optimal state-action distribution $\Theta_n(du, dx)$ for μ_n under the discounted cost criteria for the infinite population problem. Since $\mathbb{U} \times \mathbb{X}$ is assumed to be compact, there exists a convergent subsequence, say $\Theta_{n'} \to \Theta$. Furthermore, Θ is an admissible state-action distribution for μ, i.e. $\Theta(\mathbb{U}, \cdot) = \mu(\cdot)$ since $\mu_n \to \mu$. We claim that the limit Θ is also an optimal state-action distribution for μ under the discounted cost criteria. Consider the following Bellman equation:

$$K_\beta^*(\mu_{n'}) = k(\mu_{n'}, \Theta_{n'}) + \beta \int K_\beta^*(\mu_1)\eta(d\mu_1 | \mu_{n'}, \Theta_{n'}).$$

Note that under Assumption 2.2, K_β^* is continuous, and it can be shown that $k(\mu, \Theta)$ is continuous and $\eta(\cdot | \mu, \Theta)$ is weakly continuous in μ, Θ. Hence, taking the limit along $n' \to \infty$, we see that

$$K_\beta^*(\mu) = k(\mu, \Theta) + \beta \int K_\beta^*(\mu_1)\eta(d\mu_1 | \mu, \Theta)$$

which proves the claim that Θ is an optimal action for μ. In other words, any for some $\mu_n \to \mu$ and Θ_n such that Θ_n is optimal for μ_n, limit any convergent subsequence of Θ_n will be optimal for μ.

We now define the following operator for any function $h : \mathcal{P}_N(\mathbb{X}) \to \mathbb{R}$ for the finite population problem such that

$$(T^N h)(\mu^N) = k(\mu^N, g_\beta(\mu^N)) + \beta \int h(\mu_1^N)\eta(d\mu_1^N | \mu^N, g_\beta(\mu^N))$$

where $g_\beta(\mu^N) = \gamma(du | x, \mu^N)\mu^N(dx)$, i.e. g_β is the resulting team policy when the agents use symmetric policies from the infinite population problem. Note that T^N is a contraction and

$$\lim_{k \to \infty} (T_k^N h)(\mu^N) = K_\beta^N(\mu^N, g_\beta)$$

where T_k^N denotes T^N applied to the function h, k times. The convergence above is uniform over N since contraction modulus β is independent of N and since the stage-wise cost $c(c, u, \mu)$ is assumed to be uniformly bounded.

We claim that

$$\lim_{N \to \infty} T_k^N(K_\beta^{N,*})(\mu^N) = K_\beta^*(\mu)$$

for all $k < \infty$ for all $\mu^N \to \mu$. We prove the claim by induction. For $k = 1$:

$$T^N K_\beta^{N,*}(\mu^N) = k(\mu^N, g_\beta(\mu^N)) + \beta \int K_\beta^{N,*}(\mu_1^N)\eta(d\mu_1^N | \mu^N, g_\beta(\mu^N)).$$

We assume that there exists some $\epsilon > 0$ and a subsequence N' such that

$$\left| T^{N'} K_\beta^{N',*}(\mu^{N'}) - K_\beta^*(\mu) \right| > \epsilon.$$

For this subsequence, there exists a further subsequence, say N_m, such that $g_\beta(\mu^{N_m})$ converges to some Θ that is optimal for μ by the initial argument we had at the start of the proof.

24

Hence, taking the limit of both sides along $N^m \to \infty$ and using Lemma 4.2 with the fact that $K_\beta^{N,*}(\mu^N) \to K_\beta^*(\mu)$, we can have that

$$\lim_{N^m} T^{N^m} K_\beta^{N^m,*}(\mu^{N^m}) = k(\mu, \Theta) + \int K_\beta^*(\mu_1)\eta(d\mu_1|\mu, \Theta) = K_\beta^*(\mu)$$

where the last equality follows since Θ is an optimal action for μ. Finally, since the value function K_β^* is unique, we reach a contradiction, which implies that

$$\lim_{N \to \infty} (T^N K_\beta^{N,*})(\mu^N) = K_\beta^*(\mu)$$

For $k + 1$:

$$T_k^N K_\beta^{N,*}(\mu^N) = k(\mu^N, g_\beta(\mu^N)) + \beta \int T_k^N (K_\beta^{N,*})(\mu_1^N)\eta(d\mu_1^N|\mu^N, g_\beta(\mu^N)).$$

Using the same contradiction argument with the induction step that $\lim_{N \to \infty} T_k^N(K_\beta^{N,*})(\mu^N) = K_\beta^*(\mu)$ we can conclude that

$$\lim_{N \to \infty} (T_k^N K_\beta^{N,*})(\mu^N) = K_\beta^*(\mu)$$

for all $k < \infty$.

Finally, we write

$$\left| K_\beta^N(\mu^N, g_\beta) - K_\beta^*(\mu) \right| \leq \left| K_\beta^N(\mu^N, g_\beta) - (T_k^N K_\beta^{N,*})(\mu^N) \right| + \left| (T_k^N K_\beta^{N,*})(\mu^N) - K_\beta^*(\mu) \right|$$

where the first term can be made arbitrarily small by choosing k large enough for all N, and the the chosen k the second term is shown to converge to 0 as $N \to \infty$. This proves the first part of the result. The second part follows since we have $K_\beta^{N,*}(\mu^N) \to K_\beta^*(\mu)$ for any $\mu^N \to \mu$ weakly.

Appendix B. Proof of Lemma 4.2. For the given μ^N and Θ^N, one can find state and action vectors $\mathbf{x} = [x^1, x^2, \ldots, x^N]$ and $\mathbf{u} = [u^1, u^2, \ldots, u^N]$ such that $\mu_{\mathbf{x}} = \mu^N$ and $\mu_{(\mathbf{x},\mathbf{u})} = \Theta^N$ where $\mu_{\mathbf{x}}$ denotes the empirical distribution of the vector $\mathbf{x}$. Note that for any permutation $\sigma(\mathbf{x}, \mathbf{u})$, we will have that $\mu_{\sigma(\mathbf{x})} = \mu_{\mathbf{x}} = \mu^N$ and $\mu_{\sigma(\mathbf{x},\mathbf{u})} = \mu_{(\mathbf{x},\mathbf{u})} = \Theta^N$

We also consider some $\hat{\mathbf{x}}, \hat{\mathbf{u}} = [(x^1, u^1), \ldots, (x^N, u^N)]$ such that $(x^i, u^i) \sim \Theta(dx, du)$ for all $i = 1, \ldots, N$. We then have

$$W_1(\mu_{\hat{\mathbf{x}},\hat{\mathbf{u}}}, \mu_{\mathbf{x},\mathbf{u}}) = \min_\sigma \frac{1}{N} \sum_{i=1}^N \left| (\hat{x}^i, \hat{u}^i) - \sigma(x^i, u^i) \right|.$$

Note that the minimum above is achievable that is a particular permutation of the state-action vector $(\mathbf{x}, \mathbf{u})$ achieves the minimum. In what follows we consider this particular permutation.

We start by writing

$$\left| \int g^N(\mu_1^N)\eta(d\mu_1^N|\mu^N, \Theta^N) - \int g(\mu_1)\eta(d\mu_1|\mu, \Theta) \right|$$

$$\leq \left| \int g^N(\mu_1^N)\eta^N(d\mu_1^N|\mu_{\mathbf{x}}, \mu_{\mathbf{x},\mathbf{u}}) - \int g^N(\mu_1^N)\eta(d\mu_1^N|\mu_{\hat{\mathbf{x}}}, \mu_{\hat{\mathbf{x}},\hat{\mathbf{u}}}) \right|$$

$$+ \left| \int g^N(\mu_1^N)\eta(d\mu_1^N|\mu_{\hat{\mathbf{x}}}, \mu_{\hat{\mathbf{x}},\hat{\mathbf{u}}}) - \int g(\mu_1)\eta(d\mu_1|\mu, \Theta) \right|.$$

25

For the first term, we have that

$$\left| \int g^N(\mu_1^N) \eta(d\mu_1^N | \mu_{\mathbf{x}}, \mu_{\mathbf{x},\mathbf{u}}) - \int g^N(\mu_1^N) \eta(d\mu_1 | \mu_{\hat{\mathbf{x}}}, \mu_{\hat{\mathbf{x}},\hat{\mathbf{u}}}) \right|$$

$$= \left| \int g^N(\mu_{f(\mathbf{x},\mathbf{u},\mathbf{w})}^N) P(d\mathbf{w}) - \int g^N(\mu_{f(\hat{\mathbf{x}},\hat{\mathbf{u}},\mathbf{w})}^N) P(d\mathbf{w}) \right|$$

$$\leq K \int W_1(\mu_{f(\mathbf{x},\mathbf{u},\mathbf{w})}^N, \mu_{f(\hat{\mathbf{x}},\hat{\mathbf{u}},\mathbf{w})}^N) P(d\mathbf{w})$$

$$\leq K \int \frac{1}{N} \sum_{i=1}^N \left| f(x^i, u^i, \mu_{\mathbf{x}}, w^i, w^0) - f(\hat{x}^i, \hat{u}^i, \mu_{\hat{\mathbf{x}}}, w^i, w^0) \right| P(d\mathbf{w})$$

$$\leq K \frac{1}{N} \sum_{i=1}^N \left(|x^i - \hat{x}^i| + |u^i - \hat{u}^i| + W_1(\mu_{\mathbf{x}}, \mu_{\hat{\mathbf{x}}}) \right)$$

$$= K \left(W_1(\mu_{\mathbf{x},\mathbf{u}}, \mu_{\hat{\mathbf{x}},\hat{\mathbf{u}}}) + W_1(\mu_{\mathbf{x}}, \mu_{\hat{\mathbf{x}}}) \right) \to 0.$$

where $\mu_{f(\mathbf{x},\mathbf{u},\mathbf{w})}^N$ denotes the empirical distribution of the upcoming local state variables of the agents, and $P(d\mathbf{w})$ is simply the probability measure for the local noise variables of the agents $\{w_1, \ldots, w^N\}$ and the common noise w^0. The last inequality follows since we use a particular permutation of $f(\mathbf{x}, \mathbf{u}, \mathbf{w}) = \{f(x^i, u^i, \mu_{\mathbf{x}}, w^i, w^0)\}_{i=1}^N$. The last equality follows since the particular permutation fixed at the start is the one that achieves $W_1(\mu_{\mathbf{x},\mathbf{u}}, \mu_{\hat{\mathbf{x}},\hat{\mathbf{u}}})$. Finally, the convergence to 0 holds since $\mu_{\mathbf{x},\mathbf{u}} = \Theta^N \to \Theta$ by assumption, and $\mu_{\hat{\mathbf{x}},\hat{\mathbf{u}}} \to \Theta$ since they form an empirical measure for Θ.

We now focus on the second term:

$$\left| \int g^N(\mu_1^N) \eta(d\mu_1 | \mu_{\hat{\mathbf{x}}}, \mu_{\hat{\mathbf{x}},\hat{\mathbf{u}}}) - \int g(\mu_1) \eta(d\mu_1 | \mu, \Theta) \right|.$$

Note that the local noise of the agents, i.e. w^i, are i.i.d.. We consider an extended probability space for the infinite sequence of the local noise variables $\{w^1, w^2, \ldots, w^N, \ldots\}$, in which every $\omega \in \Omega$ gives rise to a possibly different sequence. Then we write the above term as

$$\int g^N(\mu_{f(\hat{\mathbf{x}},\hat{\mathbf{u}},\mathbf{w})}^N) \prod_{i=1}^N P(dw^i) P(dw^0) = \int g^N(\mu_{f(\hat{\mathbf{x}},\hat{\mathbf{u}},\mathbf{w}(\omega))}^N) P(d\omega) P(dw^0)$$

where $\mathbf{w}(\omega)$ is simply the first N terms of the local noise variables and the common noise variable.

Observe that for a given common noise w^0 and a sequence of local noise variables (or a given ω), $\mu_1^N = \mu_{f(\hat{\mathbf{x}},\hat{\mathbf{u}},\mathbf{w})}^N$ is the empirical distribution of $\{f(\hat{x}^i, \hat{u}^i, \mu_{\hat{\mathbf{x}}}, w^i, w^0)\}_i$ where $(\hat{x}^i, \hat{u}^i) \sim \Theta$, and the local noise w^i are i.i.d.. Hence, for a given common noise w^0, and for any continuous function h with unit Lipschitz constant, we have that

$$\int h(x_1) \mu_1^N(dx_1) = \frac{1}{N} \sum_{i=1}^N h(f(\hat{x}^i, \hat{u}^i, \mu_{\hat{\mathbf{x}}}, w_{(\omega)}^i, w^0)) \pm h(f(\hat{x}^i, \hat{u}^i, \mu, w_{(\omega)}^i, w^0))$$

$$\leq K_f W_1(\mu_{\mathbf{x}}, \mu) + \frac{1}{N} \sum_{i=1}^N h(f(\hat{x}^i, \hat{u}^i, \mu, w_{(\omega)}^i, w^0))$$

$$\to \int h(f(x, u, \mu, w, w^0)) \Theta(dx, du) P(dw) = \int h(x_1) \mathcal{T}^{w^0}(dx_1 | x, u, \mu) \Theta(dx, du)$$

for almost every $\omega \in \Omega$. For any given common noise realization $\mu_1^N(dx_1)$ converges to $\int \mathcal{T}^{w^0}(dx_1|x, u, \mu)\Theta(dx, du)$ almost surely. Hence by definition of the kernel η of the infnite population problem, and by the assumption that $g^N(\mu^N) \to g(\mu)$ for all $\mu^N \to \mu$ we can conclude

$$\int g^N(\mu_{f(\hat{\mathbf{x}}, \hat{\mathbf{u}}, \mathbf{w}(\omega))}^N)P(d\omega)P(dw^0) \to \int g(\mu_1(\mu, \Theta, w^0))P(dw^0) = \int g(\mu_1)\eta(d\mu_1|\mu, \Theta)$$

which conclude the proof.

REFERENCES

[1] S. Albeverio, F. C. De Vecchi, A. Romano, and S. Ugolini. Mean-field limit for a class of stochastic ergodic control problems. *SIAM Journal on Control and Optimization*, 60(1):479–504, 2022.

[2] B. Anahtarcı, C. D. Karıksız, and N. Saldi. Value iteration algorithm for mean-field games. *Systems & Control Letters*, 143:104744, 2020.

[3] B. Anahtarci, C. D. Kariksiz, and N. Saldi. Q-learning in regularized mean-field games. *Dynamic Games and Applications*, pages 1–29, 2022.

[4] A. Angiuli, J.-P. Fouque, and M. Laurière. Unified reinforcement q-learning for mean field game and control problems. *Mathematics of Control, Signals, and Systems*, 34(2):217–271, 2022.

[5] A. Arapostathis, A. Biswas, and J. Carroll. On solutions of mean field games with ergodic cost. *Journal de Mathématiques Pures et Appliquées*, 107(2):205–251, 2017.

[6] A. Arapostathis, V. S. Borkar, E. Fernandez-Gaucherand, M. K. Ghosh, and S. I. Marcus. Discrete-time controlled Markov processes with average cost criterion: A survey. *SIAM J. Control and Optimization*, 31:282–344, 1993.

[7] X. Bao and S. Tang. Ergodic control of mckean–vlasov sdes and associated bellman equation. *Journal of Mathematical Analysis and Applications*, 527(1):127404, 2023.

[8] M. Bardi and F. S. Priuli. Linear-quadratic n-person and mean-field games with ergodic cost. *SIAM Journal on Control and Optimization*, 52(5):3022–3052, 2014.

[9] N. Bäuerle. Mean field Markov decision processes. *Applied Mathematics & Optimization*, 88(1):12, 2023.

[10] E. Bayraktar, N. Bauerle, and A. D. Kara. Finite approximations for mean field type multi-agent control and their near optimality, 2023.

[11] E. Bayraktar, A. Cecchin, and P. Chakraborty. Mean field control and finite agent approximation for regime-switching jump diffusions. *Applied Mathematics & Optimization*, 88(2):36, 2023.

[12] E. Bayraktar, A. Cosso, and H. 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.

[13] E. Bayraktar and X. Zhang. Solvability of infinite horizon Mckean–Vlasov FBSDEs in mean field control problems and games. *Applied Mathematics & Optimization*, 87(1):13, 2023.

[14] A. Bensoussan, J. Frehse, and P. Yam. *Mean field games and mean field type control theory*, volume 101. Springer, 2013.

[15] P. Billingsley. *Convergence of Probability Measures*. Wiley, New York, 1968.

[16] A. Biswas. Mean field games with ergodic cost for discrete time markov processes. *arXiv preprint arXiv:1510.08968*, 2015.

[17] P. Cardaliaguet, J.-M. Lasry, P.-L. Lions, and A. Porretta. Long time average of mean field games. *Networks and heterogeneous media*, 7(2):279–301, 2012.

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

[19] R. Carmona and M. 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.

[20] R. Carmona, M. Laurière, and Z. Tan. Model-free mean-field reinforcement learning: mean-field MDP and mean-field Q-learning. *arXiv preprint arXiv:1910.12802*, 2019.

[21] M. Cirant. Stationary focusing mean-field games. *Communications in Partial Differential Equations*, 41(8):1324–1346, 2016.

[22] M. F. Djete, D. Possamaï, and X. Tan. Mckean–vlasov optimal control: limit theory and equivalence between different formulations. *Mathematics of Operations Research*, 47(4):2891–2930, 2022.

[23] M. F. Djete, D. Possamaï, and X. Tan. McKean–Vlasov optimal control: the dynamic programming principle. *The Annals of Probability*, 50(2):791–833, 2022.

[24] F. Dragoni and E. Feleqi. Ergodic mean field games with hörmander diffusions. *Calculus of Variations and Partial Differential Equations*, 57:1–22, 2018.

[25] R. Elie, J. Perolat, M. Laurière, M. Geist, and O. 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.

[26] E. Feleqi. The derivation of ergodic mean field game equations for several populations of players. *Dynamic Games and Applications*, 3:523–536, 2013.

[27] M. Fornasier, S. Lisini, C. Orrieri, and G. Savaré. Mean-field optimal control as gamma-limit of finite agent controls. *European Journal of Applied Mathematics*, 30(6):1153–1186, 2019.

[28] Z. Fu, Z. Yang, Y. Chen, and Z. Wang. Actor-critic provably finds Nash equilibria of linear-quadratic mean-field games. 2020.

[29] M. Germain, J. Mikael, and X. Warin. Numerical resolution of Mckean-Vlasov FBSDEs using neural networks. *Methodology and Computing in Applied Probability*, pages 1–30, 2022.

[30] D. A. Gomes, S. Patrizi, and V. Voskanyan. On the existence of classical solutions for stationary extended mean field games. *Nonlinear Analysis: Theory, Methods & Applications*, 99:49–79, 2014.

[31] D. A. Gomes and J. Saúde. Mean field games models—a brief survey. *Dynamic Games and Applications*, 4(2):110–154, 2014.

[32] H. Gu, X. Guo, X. Wei, and R. 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.

[33] H. Gu, X. Guo, X. Wei, and R. Xu. Dynamic programming principles for mean-field controls with learning. *Operations Research*, 2023.

[34] X. Guo, A. Hu, R. Xu, and J. Zhang. Learning mean-field games. *Advances in Neural Information Processing Systems*, 32, 2019.

[35] O. Hernández-Lerma. *Adaptive Markov Control Processes*. Springer-Verlag, 1989.

[36] O. Hernandez-Lerma and J. B. Lasserre. *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. Springer, 1996.

[37] O. Hernández-Lerma and J. B. Lasserre. *Further topics on discrete-time Markov control processes*. Springer, 1999.

[38] M. Huang, P. E. Caines, and R. 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.

[39] M. Huang, R. P. Malhamé, and P. 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.

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

[41] H. Langen. Convergence of dynamic programming models. *Math. Oper. Res.*, 6(4):493–512, Nov. 1981.

[42] J. M. Lasry and P. L. Lions. Mean field games. *Japanese journal of mathematics*, 2(1):229–260, 2007.

[43] M. Laurière and O. Pironneau. Dynamic programming for mean-field type control. *Comptes Rendus Mathematique*, 352(9):707–713, 2014.

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

[45] M. Motte and H. Pham. Quantitative propagation of chaos for mean field Markov decision process with common noise. *Electronic Journal of Probability*, 28:1–24, 2023.

[46] S. Perrin, J. Pérolat, M. Laurière, M. Geist, R. Elie, and O. Pietquin. Fictitious play for mean field games: Continuous time analysis and applications. *Advances in Neural Information Processing Systems*, 33:13199–13213, 2020.

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

[48] N. Saldi. Discrete-time average-cost mean-field games on polish spaces. *Turkish Journal of Mathematics*, 44(2):463–480, 2020.

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

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

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

[52] S. Sanjari, N. Saldi, and S. Yuksel. Optimality of decentralized symmetric policies for stochastic teams with mean-field information sharing. *arxiv preprint 2404.04957*, 2024.

[53] S. Sanjari and S. 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.

[54] J. Subramanian and A. 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.

[55] H. Tembine, Q. Zhu, and T. Başar. Risk-sensitive mean-field games. *IEEE Transactions on Automatic Control*, 59(4):835–850, 2013.

[56] P. Wiecek. Discrete-time ergodic mean-field games with average reward on compact spaces. *Dynamic Games and Applications*, 10(1):222–256, 2020.

[57] P. Wiecek and E. Altman. Stationary anonymous sequential games with undiscounted rewards. *Journal of optimization theory and applications*, 166:686–710, 2015.

[58] S. Yüksel. Optimization and control of stochastic systems. Queen's University, Lecture notes, available

online, URL: https://mast.queensu.ca/ yuksel/LectureNotesOnStochasticOptControl.pdf.