Dynamic Games and Applications
https://doi.org/10.1007/s13235-022-00448-w

Reinforcement Learning for Non-stationary Discrete-Time Linear–Quadratic Mean-Field Games in Multiple Populations

Muhammad Aneeq uz Zaman[1] · **Erik Miehling**[1] · **Tamer Başar**[1]

Accepted: 6 April 2022

Abstract
Scalability of reinforcement learning algorithms to multi-agent systems is a significant bottleneck to their practical use. In this paper, we approach multi-agent reinforcement learning from a mean-field game perspective, where the number of agents tends to infinity. Our analysis focuses on the structured setting of systems with linear dynamics and quadratic costs, named linear–quadratic mean-field games, evolving over a discrete-time infinite horizon where agents are assumed to be partitioned into finitely many populations connected by a network of known structure. The functional forms of the agents' costs and dynamics are assumed to be the same within populations, but differ between populations. We first characterize the equilibrium of the mean-field game which further prescribes an ϵ-Nash equilibrium for the finite population game. Our main focus is on the design of a learning algorithm, based on zero-order stochastic optimization, for computing mean-field equilibria. The algorithm exploits the affine structure of both the equilibrium controller and equilibrium mean-field trajectory by decomposing the learning task into first learning the linear terms and then learning the affine terms. We present a convergence proof and a finite-sample bound quantifying the estimation error as a function of the number of samples

Keywords Mean-field games · Large population games on networks · Multi-agent reinforcement learning · Zero-order stochastic optimization

1 Introduction

The successes of reinforcement learning (RL) in recent years are significant. However, to effectively scale to real-world systems, RL must be made efficient in systems with a large number of interacting agents. The primary issue with the generalization to multiple agents, i.e., multi-agent RL, is that the environment is no longer stationary from each agent's per-

Research leading to this work was supported in part by AFOSR Grant FA9550-19-1-0353. This article is part of the topical collection "Multi-agent Dynamic Decision Making and Learning" edited by Konstantin Avrachenkov, Vivek S. Borkar and U. Jayakrishnan Nair.

✉ Muhammad Aneeq uz Zaman
mazaman2@illinois.edu

[1] Coordinated Science Laboratory, University of Illinois at Urbana–Champaign, Urbana, IL 61801, USA

Published online: 10 May 2022

spective. Agents adapt their behavior as they gather new information and thus, since each agent takes actions (as a function of their own information) that influence the dynamics, the description of the world from each agent's perspective changes as other agents' understanding of the world changes. This non-stationarity leads to significant scalability challenges since each agent must reason about every other agent's internal state of information. The issues are even more pronounced in the case of agents with competing objectives as one must also account for the strategic behavior of other agents, e.g., agents taking actions so as to manipulate the beliefs of others.

An alternate paradigm, and one we explore in this paper, is the *mean-field* approach. The fundamental idea is to compute strategies in an approximate game, one which consists of infinitely many agents, and analyze how these computed strategies perform in the original finite-population game. This idea of *scaling down* from infinity, rather than scaling up from one, leads to a much more tractable analysis while allowing for one to explicitly quantify the performance loss of the approximation. In fact, fundamental results from the literature, e.g., [23], establish that the suboptimality gap evolves on the order of $1/\sqrt{N}$ where N is the number of agents in the original finite-population game. In other words, the approximation error is still satisfactory even for modestly sized systems.

The key simplification afforded by the mean-field approach is that, in the limiting approximate game, the influence of individual agents becomes negligible. As a result, any strategic or deceptive behavior by any single agent does not have a noticeable effect on the population. Rather, equilibrium behavior can be computed by analyzing how a representative agent interacts with the average population state. This concept is described formally by the *Nash-certainty equivalence* (NCE) principle [19].

Our analysis focuses on the structured setting of systems with (uncoupled) linear dynamics and (coupled) quadratic costs, referred to as *linear–quadratic mean-field games* (LQ-MFGs) [23], evolving over a discrete-time infinite horizon. Agents are assumed to be spread across multiple populations, where the form of the costs and dynamics are homogenous within populations but differ between populations. Agents' costs are coupled via a network structure where agents within each population are assumed to be fully connected, but agents between populations are connected via a sparse (but known) graph. Agents are assumed to have access to a centralized simulator which can simulate the behavior of agents in each population; however, they are assumed to be unaware the structural connections between populations and of the parameters of their own dynamics and cost functions. As such, we seek a reinforcement learning solution.

Establishing structure of the *mean-field equilibrium* (MFE), comprising both the equilibrium controller and equilibrium mean-field trajectory, facilitates application of reinforcement learning. We show that in the LQ-MFG setting, both the controller and the mean-field trajectory possess affine forms. The design of our learning algorithm takes advantage of this structure by decomposing the learning problem into two parts: (i) learning the linear (multiplicative) term of both the controller and the mean-field trajectory, and (ii) learning the affine (additive) term of the controller and the mean-field trajectory. The learning algorithm is built upon a derivative-free learning algorithm [25] termed zero-order stochastic optimization [12]. Of note is that the affine structure of mean-field trajectory allows for the development of RL algorithms that learn *non-stationary* mean-field equilibria, as we can describe how the trajectory evolves in time, as opposed to existing algorithms that learn stationary mean-field equilibria [13]. We present a convergence proof as well as a finite-sample bound quantifying the estimation error as a function of the number of samples.

1.1 Contribution

The contributions of this paper are as follows:

1. *Formulation of a multi-population LQ-MFG problem*: We present a novel mean-field game setting in which agents are partitioned into multiple populations which are connected via a sparse network. Heterogeneous costs and dynamics are assumed between populations.
2. *Existence, uniqueness, and structure of the MFE*: We prove existence and uniqueness of the MFE under general conditions. Furthermore, we show that this this equilibrium admits a controller and mean-field trajectory that are affine in their respective arguments.
3. *An ϵ-Nash bound for the multi-population LQ-MFG*: We present bounds on how "close" (in terms of cost) the mean-field equilibrium strategies are to the Nash equilibrium strategies. Our result generalizes existing bounds for the single-population setting [23, 29], demonstrating that the error for each population l grows on the order of $1/\sqrt{\min_{k\in\mathcal{L}_l} N_k}$, where the minimum is the size of the smallest population immediately neighboring (and including) l.
4. *Development of a provably convergent zero-order stochastic optimization algorithm for learning an MFE*: We develop an efficient derivative-free learning algorithm, based on zero-order stochastic optimization. We present convergence and finite-sample bounds for the algorithm and demonstrate the performance numerically.

1.2 Related Work

Since the seminal works on mean-field games [18, 20], there has been burgeoning interest on linear–quadratic mean-field games for both continuous-time [5, 17, 19] and discrete-time settings [13, 23, 28]. In addition to characterizing and analyzing the equilibria of these mean-field games, many existing works have focused on developing RL algorithms for LQ-MFGs for both stationary settings [13] and (the more general) non-stationary setting [29]. Many of these works rely on policy gradient methods. In contrast, the proposed algorithm of our paper is built upon the "*truly model-free*" [22] zero-order stochastic optimization (ZSO) algorithm where only access to the long run cost is required, rather than access to the state and action sequences. In addition to the LQ setting, several works have considered estimation of mean-field equilibria in the nonlinear setting using fictitious play [11] or fitted-Q learning [2], among others.

Our work is distinct from the body of literature titled *mean-field games on networks*. In these works, e.g., Camilli and Marchi, [1, 8], the underlying state space of agents consists of a network of nodes. In contrast, our work can be considered as a type of *networked* mean-field game in which the network structure describes the (cost) coupling among agents, rather than comprising the state of the problem. Of particular note is that the network in our setting may not be fully connected. In this sense, our work is similar to MFGs on Erdös-Renyi graphs [10] and, more recently, graphon MFGs [7, 14], both of which consider sparsity in the interaction among agents; however, these works differ in that they do not model clustering of agents into multiple populations. Some other related works are as follows. The work of Bauso et al. [4] considers a network of networks setting where the game at the lower level is a two player game and the game at the higher level is a mean-field game on a network. The work of Zhu and Başar [31] introduces a multi-resolution mean-field game but does not consider sparsity in interaction of agents. To the best of our knowledge, this paper is the first one

to propose an RL algorithm to approximate the non-stationary mean-field equilibrium of a multi-population MFG.

1.3 Outline

The remainder of the paper is organized as follows. Section 2 presents background on reinforcement learning in linear–quadratic mean-field games. Section 3 presents our multi-population model and corresponding ϵ-Nash bound. Section 4 presents the development of our RL algorithm, convergence and finite-sample results, and some numerical results. Section 5 provides concluding remarks. Most of the proofs have been relegated to an Appendix, which follows the references section.

1.4 Notation

All vectors are assumed to be column vectors, e.g., $x \in \mathbb{R}^n$ is interpreted as $x \in \mathbb{R}^{n\times 1}$ and is written inline as $x = (x_1, \ldots, x_n)$. A similar convention applies to matrices, e.g., for $A, B \in \mathbb{R}^{n\times n}$, $C = (A, B)$ denotes the $\mathbb{R}^{2n\times n}$ matrix. Mathematical spaces are denoted by blackboard characters, e.g., the real numbers $\mathbb{R}$. Superscript indices index agents and populations, e.g., $Z_t^{n,l}$, whereas superscripts in parentheses are used to denote iteration indices. A variable enclosed by square brackets, such as $[N]$, represents sets of the form $\{1, \ldots, N\}$, with all other sets either written out explicitly $\{1, 2, \ldots\}$ or denoted by a calligraphic symbol, such as, $\mathcal{N}$. Upright sans-serif notation is used to denote collections across populations, such as $N = (N_1, \ldots, N_L)$ and $\phi = (\phi^1, \ldots, \phi^L)$.

2 Preliminaries: Learning in Non-stationary Linear–Quadratic Mean-Field Games

Consider a dynamic game consisting of $N < \infty$ agents interacting over an infinite horizon. Associated with each agent $n \in [N]$ is a state $Z_t^n \in \mathcal{Z}{=}\mathbb{R}^m$ which obeys controlled linear time-invariant (LTI) stochastic dynamics of the form,

$$Z_{t+1}^n = AZ_t^n + BU_t^n + W_t^n, \tag{1}$$

where $A \in \mathbb{R}^{m\times m}$ is the state matrix, $B \in \mathbb{R}^{m\times p}$ is the input matrix, $U_t^n \in \mathcal{U} = \mathbb{R}^p$ is agent n's control actions at time t, and $W_t^n \in \mathbb{R}^m$, $t = 0, 1, \ldots$, are noise terms drawn independently from a Gaussian distribution, $W_t^n \sim \mathcal{N}(0, \Sigma_w)$. Each agent's initial state Z_0^n is drawn from a Gaussian distribution, $Z_0^n \sim \mathcal{N}(\nu_0, \Sigma_0)$, independently of all other initial states and noise terms. Note that dynamics are uncoupled, i.e., the evolution of agent n's state does not depend on the states or controls of other agents. Lastly, the pair (A, B) is assumed to be controllable.

Agents are assumed to possess local information. Agent n's information at time t is denoted by $I_t^n := (I_{t-1}^n, U_{t-1}^n, Z_t^n) \in \mathcal{I}_t := (\mathcal{Z} \times \mathcal{U})^t \times \mathcal{Z}$ initialized by $I_0^n = Z_0^n$. A control policy for agent n at time t, denoted by π_t^n, is a function that maps information I_t^n to control $U_t^n \in \mathbb{R}^p$. Let the sequence $\pi^n := (\pi_0^n, \pi_1^n, \ldots) \in \Pi := \{\varpi = (\varpi_0, \varpi_1, \ldots) \mid \varpi_t : \mathcal{I}_t \to \mathcal{U}, t = 0, 1, \ldots\}$ denote a control law, where Π is the set of all admissible control laws. A joint control law is given by $\pi = (\pi^1, \ldots, \pi^N)$. Agent n's expected cost under a given joint

control law $\pi = (\pi^n, \pi^{-n})$, where $\pi^{-n} = (\pi^1, \ldots, \pi^{n-1}, \pi^{n+1}, \ldots, \pi^N)$, is defined as

$$J_n^{(N)}(\pi^n, \pi^{-n}) := \limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\Big[\|Z_t^n\|_Q^2 + \|U_t^n\|_{C_U}^2 + \Big\| Z_t^n - \frac{1}{N-1} \sum_{n' \neq n} Z_t^{n'} \Big\|_{C_Z}^2 \Big], \tag{2}$$

which consists of penalties for state magnitude, control magnitude, and deviation of each agent's state from the average state. Norms are taken with respect to symmetric matrices $Q \geq 0$, $C_U > 0$, and $C_Z \geq 0$ of appropriate dimensions, where the pair $(A, Q^{1/2})$ is assumed to be observable. The expectation is defined over the probability measure on state-action trajectories induced by the joint control law π, the initial state distribution, and the noise distribution.

The above dynamics and costs form a non-cooperative stochastic dynamic game among agents. Due to the local information of the agents, these stochastic games are nearly impossible to solve for their Nash equilibria due to second guessing and infinite recursion. As such, approximate Nash equilibria of the game, termed ϵ-Nash equilibria, are defined as follows.

Definition 1 A joint control law $\tilde{\pi}$ is said to be in ϵ-Nash equilibrium if

$$J_n^{(N)}(\tilde{\pi}^n, \tilde{\pi}^{-n}) \leq J_n^{(N)}(\pi^n, \tilde{\pi}^{-n}) + \epsilon \tag{3}$$

for all $\pi^n \in \Pi$, $n \in [N]$.

The standard approach for finding an ϵ-Nash equilibrium is to apply the *Nash-certainty equivalence* NCE principle [18]. The general idea is to study the interaction between a representative agent, termed a *generic* agent, and the state distribution of the population. The generic agent faces a special type of optimal control problem in which the state distribution over time is constrained by the effect of the aggregate population.

Denoting the state of the generic agent by $Z_t \in \mathbb{R}^m$, its dynamics are identical to those in (1),

$$Z_{t+1} = AZ_t + BU_t + W_t, \tag{4}$$

where $U_t \in \mathcal{U}$ is the control action and $Z_0 \sim \mathcal{N}(\nu_0, \Sigma_0)$, $W_t \sim \mathcal{N}(0, \Sigma_w)$ are drawn independently. The generic agent's control policy at time t, denoted by ϕ_t, maps its information $I_t = (I_{t-1}, U_{t-1}, Z_t) \in \mathcal{I}_t$, $I_0 = Z_0$, to a control action U_t. The control policy is dependent in an implicit manner on the *mean-field trajectory* $\bar{Z} := (\bar{Z}_0, \bar{Z}_1, \ldots) \in \bar{\mathcal{Z}}$, which represents the average state trajectory of the agents in the population. Given a control law $\phi = (\phi_1, \phi_2, \ldots) \in \Phi$, the generic agent's expected cost is defined as

$$J(\phi, \bar{Z}) = \limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\Big[\|Z_t\|_Q^2 + \|U_t\|_{C_U}^2 + \|Z_t - \bar{Z}_t\|_{C_Z}^2 \Big], \tag{5}$$

It is assumed that mean-field trajectories lie in the space of deterministic bounded sequences, that is, $\bar{\mathcal{Z}} := \{x = (x_0, x_1, \ldots) \mid \sup_{t\geq 0} |x_t| < \infty\}$.[1] As before, the mean-field trajectory serves as a reference trajectory in the generic agent's LQT problem.

[1] See Moon and Başar, [23] for a justification of this assumption.

A mean-field equilibrium (MFE) is defined as a control law-trajectory pair $(\phi, \bar{Z})$ such that ϕ is optimal for $\bar{Z}$ and $\bar{Z}$ is consistent with ϕ. Formally, let $\Lambda : \Phi \to \bar{\mathcal{Z}}$ be defined recursively as follows

$$\bar{Z}_{t+1} = A\bar{Z}_t + B\phi_t(\bar{Z}_t), \tag{6}$$

with $\bar{Z}_0 = \nu_0$. Note that since a mean-field trajectory $\bar{Z}$ is computed as a function of the controlled state dynamics (4), the control law implicitly depends on $\bar{Z}$. A mean-field trajectory $\bar{Z}$ is termed *consistent with policy* ϕ if $\bar{Z} = \Lambda(\phi)$. Conversely, let $\Psi : \bar{\mathcal{Z}} \to \Phi$ be defined as

$$\Psi(\bar{Z}) = \operatorname*{argmin}_{\phi \in \Phi} J(\phi, \bar{Z}) \tag{7}$$

where $J(\phi, \bar{Z})$ is given by (5). A policy ϕ^* is termed a *cost-minimizing controller for* $\bar{Z}$ if $\phi^* \in \Psi(\bar{Z})$. An MFE is defined as follows.

Definition 2 ([24]) The tuple $(\phi^*, \bar{Z}^*) \in \Phi \times \mathcal{Z}$ is an MFE if $\phi^* \in \Psi(\bar{Z}^*)$ and $\bar{Z}^* = \Lambda(\phi^*)$.

The associated mean-field trajectory $\bar{Z}^*$ is termed an *equilibrium mean-field trajectory*, whereas the associated control law ϕ^* is termed an *equilibrium controller*. Note that the above definition corresponds to a non-stationary MFE, in that both the mean-field trajectory and the controller are allowed to vary in time, in contrast to stationary MFE [13, 16, 26].[2] The corresponding game is referred to as a non-stationary LQ-MFG.

A primary result of mean-field games is a quantification of the error when the computed strategy (from the infinite-population approximation) is used in the original, finite-population game. The following result, now standard, characterizes this error.

Theorem 1 *(ϵ-Nash bound [23, 29]) Let $\tilde{\phi}$ denote the N-tuple collection of controllers where each agent n employs the controller ϕ^* prescribed by the MFE, i.e., $\tilde{\phi} = \underbrace{(\phi^*, \ldots, \phi^*)}_{N \text{ times}}$. Then, for all $n \in [N]$,*

$$J_n^{(N)}(\tilde{\phi}) \leq \inf_{\pi^n \in \Pi} J_n^{(N)}(\pi^n, \tilde{\phi}^{-n}) + \mathcal{O}(1/\sqrt{N}) \tag{8}$$

where $J_n^{(N)}(\cdot)$ is the cost function of agent n in the finite population game and $\tilde{\phi}^{-n}$ denotes the collection of $N-1$ controllers where all agents except agent n employ controller ϕ^*.

The above bound demonstrates that, when the model is perfectly known, the error of the computed strategy decreases on the order of $1/\sqrt{N}$ where N is the population size. The proof proceeds by establishing the bound in an expanded information structure where each agent additionally possesses state information of all other agents. If the bound holds under this expanded information structure, then it also holds under the information structure where each agent only possesses local information. Since this holds for all agents, the MFE $\tilde{\phi}$ is an ϵ-Nash equilibrium for the finite population game. When the model parameters are unknown, and a learning algorithm is employed to compute an equilibrium, an additional error is introduced due to the stopping tolerance of the algorithm.

For the class of LQ-MFGs of this paper, a convenient parameterization of MFE exists which enables a direct application of RL. In particular, the equilibrium mean-field can be shown to possess linear dynamics, i.e., $\bar{Z}^*_{t+1} = F^* \bar{Z}^*_t$ [9, 29]. Combined with the parameterization of the control law by a gain matrix $K \in \mathbb{R}^{p \times n}$ (due to the linear–quadratic structure

[2] In other words, equilibrium mean-field trajectories are allowed to vary in time; see definition (A3) in Subramanian and Mahajan, [26].

of the problem), the generic agent's LQT problem can be reformulated as a state-feedback LQG problem with an extended state $(Z_t, \bar{Z}_t)$ and cost function in terms of the pair (K, F). Our previous work [29] had proposed a provably convergent learning algorithm that consists of two parts: i) an actor-critic process for learning an approximate cost-minimizing controller for a given mean-field trajectory, and ii) an aggregation step, carried out by a central *simulator* agent, for updating the mean-field state matrix. This yields a data-driven process for learning the MFE. The associated error bound demonstrates that the algorithm yields an approximate MFE that converges to an ϵ-Nash equilibrium of the finite-population game with probability at least $1 - \epsilon^5$ if the algorithm iterates are on the order of $\log(1/\epsilon)$ and the number of agents are on the order of $1/\epsilon^2$ [29].

The discussion up to this point has reviewed the case of a single population of homogeneous agents. The objective of the remainder of the paper is to generalize the above results to the multiple population case where agents in different populations exhibit heterogeneity in their dynamics and costs, and interact over a network.

3 Non-stationary Linear–Quadratic Mean-Field Games in Multiple Populations

The setting described in the aforementioned section, while theoretically interesting, has limited application in real-world settings. The limitation arises primarily from the assumption that agents are homogeneous, i.e., they possess the same dynamics and functional form of the cost, and that agents are fully connected, i.e., every agent can "talk" with every other agent. These two assumptions are often violated in practice. To model such environments, we generalize the setting of Sect. 2 to the case where agents are subdivided into populations, with agents within a given population possessing homogeneous dynamics and costs, but allowing for heterogeneity between populations.

In this section, we prove the existence and uniqueness of an affine mean-field equilibrium and present an ϵ-Nash bound for the multi-population case, analogous to Theorem 1 of Sect. 2.

Consider a collection of $L < \infty$ populations connected via a network with undirected edge set $\mathcal{E} \subseteq [L] \times [L]$. Each population $l \in [L]$ consists of $N_l < \infty$ agents, where each agent $n \in [N_l]$ possesses a state $Z_t^{n,l} \in \mathcal{Z}^l = \mathbb{R}^m$ which obeys uncoupled controlled LTI stochastic dynamics of the form

$$Z_{t+1}^{n,l} = A^l Z_t^{n,l} + B^l U_t^{n,l} + W_t^{n,l} \tag{9}$$

where $A^l \in \mathbb{R}^{m\times m}$ is the state matrix, $B^l \in \mathbb{R}^{m\times p}$ is the input matrix, $U_t^{n,l} \in \mathcal{U}^l = \mathbb{R}^p$ is the control at time t, and $W_t^{n,l} \in \mathbb{R}^m$, $t = 0, 1, \ldots$. Initial states and noise terms are assumed to be drawn independently as $Z_0^{n,l} \sim \mathcal{N}(\nu_0^l, \Sigma_0^l)$, $W_t^{n,l} \sim \mathcal{N}(0, \Sigma_w^l)$, $t = 0, 1, \ldots$. Each pair (A^l, B^l), $l \in [L]$, is assumed to be controllable.

As in the single population setting of Sect. 2, agents possess local information. The information available to agent n in population l at time t is given by $I_t^{n,l} := (I_{t-1}^{n,l}, U_{t-1}^{n,l}, Z_t^l) \in \mathcal{I}_t^l := (\mathcal{Z}^l \times \mathcal{U}^l)^t \times \mathcal{Z}^l$, $I_0^{n,l} = Z_0^l$. A control policy for agent n in population l at time t is given by $\pi_t^{n,l} : \mathcal{I}_t^l \to \mathcal{U}^l$ with a control law denoted by $\pi^{n,l} = (\pi_0^{n,l}, \pi_1^{n,l}, \ldots) \in \Pi^l := \{\varpi = (\varpi_0, \varpi_1, \ldots) \mid \varpi_t : \mathcal{I}_t^l \to \mathcal{U}^l, t = 0, 1, \ldots\}$. Define $\pi^l := (\pi^{1,l}, \ldots, \pi^{N_l,l})$ as population l's control law, equivalently written as $\pi^l = (\pi^{n,l}, \pi^{-n,l})$, where $\pi^{-n,l} = (\pi^{1,l}, \ldots, \pi^{n-1,l}, \pi^{n+1,l}, \ldots, \pi^{N_l,l})$ is the set of policies of all agents in population l excluding agent n. Similarly, let $\pi = (\pi^l, \pi^{-l})$ denote the joint control law across populations

where $\pi^{-l} = (\pi^1, \ldots, \pi^{l-1}, \pi^{l+1}, \ldots, \pi^L)$.[3] Denote by $\mathcal{L}_l \subseteq [L]$ the neighborhood of population l, defined as $\mathcal{L}_l := \{l\} \cup \{k \in [L] \mid e_{kl} = 1\}$, and $N = (N_1, \ldots, N_L)$ the set of finite population sizes. The expected cost of agent n in population l under joint control law $\pi = (\pi^l, \pi^{-l}) = \big((\pi^{n,l}, \pi^{-n,l}), \pi^{-l}\big)$ is given by,

$$J_{n,l}^{(N)}\big((\pi^{n,l}, \pi^{-n,l}), \pi^{-l}\big) := \limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\Bigg[\big\| Z_t^{n,l} \big\|_{Q^l}^2 + \big\| U_t^{n,l} \big\|_{C_U^l}^2 + \Big\| Z_t^{n,l} - \frac{1}{N_l - 1} \sum_{\substack{n' \in [N_l] \\ n' \neq n}} Z_t^{n',l} \Big\|_{C_Z^{ll}}^2 + \sum_{\substack{k \in \mathcal{L}_l \\ k \neq l}} \Big\| Z_t^{n,l} - \Big(\beta^{lk} + \frac{1}{N_k} \sum_{n' \in [N_k]} Z_t^{n',k} \Big) \Big\|_{C_Z^{lk}}^2 \Bigg], \quad (10)$$

where the first three terms inside the expectation share the same meaning to those in (2) (now with population dependence) and the last term penalizes deviation from a potentially offset average of neighboring populations' states, where $\beta^{lk} \geq 0$. The offset term models scenarios in which neighboring populations wish to maintain a fixed deviation from agents in other populations, e.g., a formation control application where each swarm (population) of agents maintains a fixed distance to neighboring swarms. For all $l, k \in [L]$, norms are taken with respect to matrices $Q^l \geq 0$, $C_U^l > 0$, $C_Z^{ll} \geq 0$, and $C_Z^{lk} \geq 0$ of appropriate dimensions,[4] where each pair $(A^l, (Q^l)^{1/2})$ is assumed to be observable. The expectation is taken with respect to the measure on state-action trajectories induced by the control, initial state distributions, and the noise distributions.

Note that our investigation is motivated by problems that only possess cost coupling and not state coupling, namely those of consensus [3] and formation control [15], just to cite a few. As a result, the network structure, encoded by the edge set $\mathcal{E}$, only influences agents' costs and not their dynamics. This coupling structure enables us to establish certain properties of the mean-field equilibrium (i.e., affineness) which in turn facilitate the design of an RL algorithm. Other papers that develop learning algorithms for MFGs with coupled dynamics either consider stationary MFE [13, 16] or impose stricter conditions on the mean-field trajectory simulator [11].

As in the single population setting of Sect. 2, the above described dynamics and costs give rise to a non-cooperative dynamic game among agents, but now across multiple populations. An ϵ-Nash equilibrium of this game is defined as follows.

Definition 3 A joint control law $\tilde{\pi}$ is said to be an ϵ-Nash equilibrium if

$$J_{n,l}^{(N)}\big((\tilde{\pi}^{n,l}, \tilde{\pi}^{-n,l}), \tilde{\pi}^{-l}\big) \leq J_{n,l}^{(N)}\big((\pi^{n,l}, \tilde{\pi}^{-n,l}), \tilde{\pi}^{-l}\big) + \epsilon \quad (11)$$

for all $\pi^{n,l} \in \Pi^l$, $n \in [N_l]$, $l \in [L]$.

Similar to the single population model of Sect. 2, an ϵ-Nash equilibrium is obtained by taking an infinite population limit of each population, i.e., $N_l \to \infty$, $l \in [L]$. The game is solved by applying the NCE; however, unlike in the single population setting, a generic agent is associated with *each* population $l \in [L]$. The state of the generic agent of population l, denoted by $Z_t^l \in \mathbb{R}^m$, obeys dynamics identical to (9),

$$Z_{t+1}^l = A^l Z_t^l + B^l U_t^l + W_t^l, \quad (12)$$

[3] The notation for π has been overwritten from Sect. 2 as the network case will be the topic of the remainder of the paper.

[4] If populations l and k are disconnected then $C_Z^{lk} = 0$.

with control $U_t^l \in \mathcal{U}^l$ and independently drawn terms $Z_0^l \sim \mathcal{N}(\nu_0^l, \Sigma_0^l)$, $W_t^l \sim \mathcal{N}(0, \Sigma_w^l)$, $t = 0, 1, \ldots$. The l'th generic agent's control policy at time t is given by $\phi_t^l : \mathcal{I}_t^l \to \mathcal{U}^l$, with its control law denoted by $\phi^l = (\phi_0^l, \phi_1^l, \ldots) \in \Phi^l$. The cost for generic agent l under control law ϕ^l and mean-field trajectory $\bar{Z} := (\bar{Z}_1, \bar{Z}_2, \ldots) \in \bar{\mathscr{Z}}$, $\bar{Z}_t := (\bar{Z}_t^1, \ldots, \bar{Z}_t^L)$, is given by,

$$J_l(\phi^l, \bar{Z}) := \limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\Big[\big\| Z_t^l \big\|_{Q^l}^2 + \big\| U_t^l \big\|_{C_U^l}^2 + \Big\| Z_t^l - \bar{Z}_t^l \Big\|_{C_Z^{ll}}^2 + \sum_{\substack{k\in\mathcal{L}_l \\ k\neq l}} \Big\| Z_t^l - \Big(\beta^{lk} + \bar{Z}_t^k \Big) \Big\|_{C_Z^{lk}}^2 \Big]. \tag{13}$$

where $\bar{Z}^l := (\bar{Z}_0^l, \ldots)$ represents the average state trajectory of agents in population $l \in [L]$. The mean-field trajectory lies in the space $\bar{\mathscr{Z}} := \prod_{l\in[L]} \bar{\mathscr{Z}}^l$, where each $\bar{\mathscr{Z}}^l$ is a space of deterministic bounded sequences.

3.1 Mean-Field Equilibrium of the Multi-population Game

Let $\phi := (\phi^1, \ldots, \phi^L) \in \Phi$ denote the joint control law across populations. Let $\Lambda : \Phi \to \bar{\mathscr{Z}}$ be defined as,

$$\bar{Z}_{t+1}^l = A^l \bar{Z}_t^l + B^l \phi_t^l(\bar{Z}_t^l), \;\; \bar{Z}_0^l = \nu_0^l$$

for all $l \in [L]$. Conversely, let $\Gamma : \bar{\mathscr{Z}} \to \Phi$ be defined as $(\Gamma(\bar{Z})) = \big(\psi^1(\bar{Z}), \ldots, \psi^L(\bar{Z})\big)$, where

$$\psi^l(\bar{Z}) = \operatorname*{argmin}_{\phi^l \in \Phi^l} J_l(\phi^l, \bar{Z})$$

for all $l \in [L]$. An MFE of the multi-population game is defined as follows.

Definition 4 The tuple $(\phi^*, \bar{Z}^*)$ is an MFE of the multi-population LQ-MFG if $\phi^* \in \mathrm{fl}(\bar{Z}^*)$ and $\bar{Z}^* = (\phi^*)$.

Denote by $Z_t := (Z_t^1, \ldots, Z_t^L)$ the joint mean-field at time t, i.e., the mean-field across all populations. The following proposition establishes the form of the equilibrium controllers and demonstrates that the joint mean-field evolves according to affine dynamics.

Proposition 1 *An MFE $(\phi^*, \bar{Z}^*)$ exists and is unique. Furthermore, the equilibrium controller $\phi^* = (\phi^{1*}, \ldots, \phi^{L*})$ and the equilibrium mean-field $\bar{Z}^* = (\bar{Z}_1^*, \bar{Z}_2^*, \ldots)$ take the following forms:*

1. $\phi^{l*}(Z_t^l, \bar{Z}_t^*) = -K_{l,1}^* \begin{bmatrix} Z_t^l \\ \bar{Z}_t^* \end{bmatrix} - K_{l,2}^* = -K_{l,1}^{1*} Z_t^l - K_{l,1}^{2*} \bar{Z}_t^* - K_{l,2}^*$ *for all* $l \in [L]$
2. $\bar{Z}_{t+1}^* = F^* \bar{Z}_t^* + C^*$

for some matrices $K_{l,1}^* = (K_{l,1}^{1*}, K_{l,1}^{2*})$, $K_{l,2}^*$, F^* and C^*.

Proof Proof of this result can be found in the Appendix, where also the expressions for the matrices in the affine structure are given. □

Note that since the mean-field trajectory $\bar{Z}^*$ has deterministic dynamics, it can be computed offline; hence, the generic agent in each population does not need to observe the mean-field trajectory. Compared to the linear dynamics of the mean-field trajectory in the single

population case of Sect. 2, the joint mean-field trajectory in the multi-population case evolves according to *affine* dynamics, rather than linear, due to the bias term β_{lk} between populations (see (10)).

Under the mean-field approach, the above MFE inherently assumes that each population contains an infinite number of agents. The following bound is one of the main results of the paper, serving as an analogous result to the ϵ-Nash bound of Theorem 1 in Sect. 2, and quantifies how the equilibrium controller of the MFE performs in the original finite-population game. Proof of this theorem can be found in the Appendix.

Theorem 2 *Let $\tilde{\phi}$ denote the collection of controllers where each agent n in each population l employs the equilibrium controller of the corresponding generic agent, that is,*

$$\tilde{\phi} = \big(\underbrace{\phi^{1*}, \ldots, \phi^{1*}}_{N_1 \text{ times}}, \underbrace{\phi^{2*}, \ldots, \phi^{2*}}_{N_2 \text{ times}}, \ldots, \underbrace{\phi^{L*}, \ldots, \phi^{L*}}_{N_L \text{ times}}\big)$$

where ϕ^{l*} is given by Proposition 1. Then, for all $n \in [N_l]$, $l \in [L]$,

$$J_{n,l}^{(N)}(\tilde{\phi}) \leq \inf_{\pi^{n,l} \in \Pi^l} J_{n,l}^{(N)}\big((\pi^{n,l}, \tilde{\phi}^{-n,l}), \tilde{\phi}^{-l}\big) + \mathcal{O}\Big(1/\sqrt{\min_{k \in \mathcal{L}_l} N_k}\Big) \tag{14}$$

where $J_{n,l}^{(N)}(\cdot)$ is the cost function of (10), $\tilde{\phi}^{-n,l}$ denotes the elements of $\tilde{\phi}$ corresponding to population l excluding agent n, and $\tilde{\phi}^{-l}$ denotes the elements of $\tilde{\phi}$ excluding population l.

The above result states that the performance loss of the agents using their respective population's equilibrium controller is bounded by a term that grows on the order of $\mathcal{O}\Big(1/\sqrt{\min_{k \in \mathcal{L}_l} N_k}\Big)$. Similar to the discussion of the bound following Theorem 1 in Sect. 2, the proof of the above result proceeds by assuming an expanded information structure where each agent is assumed to have access to the states of neighboring agents in their own population. Demonstrating the bound under this information structure implies that the bound holds in the information structure where each agent possesses local information. Notice that the bound of Theorem 2 reduces to that of Theorem 1 when there is a single population.

4 Zero-Order Stochastic Optimization for Non-stationary LQ-MFGs in Multiple Populations

Every agent l is assumed to be unaware of both their specific dynamics, e.g., A^l, B^l, as well as the parameters of their cost function, e.g., Q^l, C_U^l, C_Z^{ll}, C_Z^{lk}. As such, a learning algorithm must be employed in order for agents to compute an MFE. In this section we develop a reinforcement learning algorithm, based on the zero-order stochastic optimization (ZSO) algorithm [12], to obtain the MFE in a setting where the model parameters of the generic agents are unknown. The ZSO algorithm is a "truly model-free" algorithm [22] in the sense that agents are assumed to have access to only the cost sequence generated by a given policy. In our setting, agents are further assumed to have access to the current mean-field trajectory. Our main result is a finite-sample bound quantifying the estimation error as a function of the number of samples.

4.1 Zero-Order Stochastic Optimization for Learning Mean-Field Equilibria

As a consequence of Proposition 1, the search for an MFE is restricted to affine control laws and affine mean-field dynamics. For a given mean-field trajectory, each generic agent computes his control law and reports it to a centralized entity termed a *simulator*. The role of the simulator is to update the estimate of the mean-field as a function of the policies of the (generic) agents.[5] The assumption of the presence of such a centralized entity is common in the literature [11, 13, 16, 26] and is practically motivated in many real-world environments, e.g., an aerial base station in a 5G network scenario [30].

For the purposes of designing the learning algorithm, it is useful to define an extended state for each population. For each $l \in [L]$, define the *extended state* as $X_t^l := (Z_t^l, \bar{Z}_t) \in \mathbb{R}^{m(L+1)}$. Given Proposition 1, the extended state follows stochastic affine dynamics of the form,

$$X_{t+1}^l = \bar{A}^l X_t^l + \bar{B}^l U_t^l + \bar{C} + \bar{W}_t^l \tag{15}$$

with

$$\bar{A}^l = \begin{bmatrix} A^l & 0 \\ 0 & F \end{bmatrix}, \bar{B}^l = \begin{bmatrix} B^l \\ 0 \end{bmatrix}, \bar{C} = \begin{bmatrix} 0 \\ C \end{bmatrix}, \bar{W}_t^l = \begin{bmatrix} W_t^l \\ \omega_t \end{bmatrix},$$

where $\bar{W}_t^l \sim \mathcal{N}(0, \bar{\Sigma}^l)$, $\bar{\Sigma}^l := \mathrm{diag}(\Sigma_w^l, \Sigma)$, and the term $\omega_t \sim \mathcal{N}(0, \Sigma)$, $\Sigma \in \mathbb{R}^{M \times M}$, is an additive noise term included to allow for inherent stochasticity in the centralized simulator's update.

For a given mean-field trajectory $\bar{Z}$, the cost of generic agent l under control law ϕ^l can be expressed in terms of the extended state X_t^l as

$$J_l(\phi^l, \bar{Z}) := \limsup_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\big[\big\| X_t^l - \bar{\beta}^l \big\|_{\bar{Q}^l}^2 + \big\| U_t^l \big\|_{C_U^l}^2\big], \tag{16}$$

where

$$\bar{Q}^l = \begin{bmatrix} Q^l + \sum_{k \in \mathcal{L}} C_Z^{lk} & -C_Z^{l1} & -C_Z^{l2} & \cdots & -C_Z^{lL} \\ -C_Z^{l1} & C_Z^{l1} & 0 & \cdots & 0 \\ -C_Z^{l2} & 0 & C_Z^{l2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \\ -C_Z^{lL} & 0 & 0 & & C_Z^{lL} \end{bmatrix} \text{ and } \bar{\beta}^l = \begin{bmatrix} 0 \\ -\beta^{l1} \\ \vdots \\ -\beta^{lL} \end{bmatrix}. \tag{17}$$

The problem defined by finding the minimizing $\phi^l \in \Phi^l$ of (16) subject to the dynamics (15) is a stochastic LQ problem with drift and a constant tracking signal. The optimal control law for a given mean-field trajectory can thus be written as Yang et al. [27]

$$\phi_t^l(X_t^l) = -K_{l,1} X_t^l - K_{l,2}. \tag{18}$$

where $K_{l,1} \in \mathbb{R}^{p \times m(L+1)}$ is the control gain and $K_{l,2} \in \mathbb{R}^p$ is the control offset. As shown in Fu et al. [13], the cost of control law $K_l = (K_{l,1}, K_{l,2})$, with slight abuse of notation, decomposes as follows,

[5] The zero-order stochastic optimization algorithm requires access to a finite length mean-field trajectory. This rollout length (also called the truncation length) is tied to the accuracy of the ZSO algorithm. For a stable system (which is the case in this paper) the rollout length is $\mathcal{O}(\log(1/\bar{\delta}))$, where $\bar{\delta}$ is the accuracy of cost estimate J_l. The reader is referred to [22], Section 2.2.2, for a detailed explanation of this ZSO-based truncation scheme.

$$J_l((K_{l,1}, K_{l,2}), \bar{Z}) = J_l^1(K_{l,1}, \bar{Z}) + J_l^2((K_{l,1}, K_{l,2}), \bar{Z}) + \bar{\beta}^{l\top} \bar{Q}^l \bar{\beta}^l. \tag{19}$$

The intuition for the design of the algorithm is as follows. Disregarding the last term in (19), total cost J_l is comprised of costs J_l^1 and J_l^2. The cost J_l^1 is the standard LQR cost and depends on the control gain $K_{l,1}$. The cost J_2^1 depends on the mean-field drift C and the control offset $K_{l,2}$ in addition to the control gain $K_{l,1}$. If the control offset is ignored, i.e., $K_{l,2} = 0$, and the mean-field trajectory has no drift, $C = 0$, then, $J_l^2 = 0$ and hence optimizing the total cost J_l is equivalent to optimizing the cost J_l^1. Moreover, as shown in Fu et al. [13] for a fixed mean-field offset C, the cost J_l^2 for the optimal control offset $K_{l,2}$ is independent of the control gain $K_{l,1}$. These two pieces of information inspire the two-part RL algorithm (Algorithm 1).

For a given set of controllers $\phi = (\phi^1, \ldots, \phi^L)$, where each ϕ^l is given by (18), the centralized simulator *approximates* the mean-field trajectory $\bar{Z}$ consistent with ϕ. This is obtained through simulating the state dynamics $\bar{Z}_t^l$ of a single agent for each population l, with initial state $Z_0^l = \nu_0^l$, under its corresponding control law ϕ^l (18). This simulator is similar to the one in Elie et al. [11] where a single agent is used to approximate the mean-field trajectory. Since a single agent is being used to approximate the behavior of a large number of agents, i.e., mean-field trajectory, this introduces an approximation error (justifying the need for stochasticity in the simulated mean-field trajectory). For a given mean-field trajectory $\bar{Z}$, the state of agent l under control law ϕ^l follows the dynamics

$$Z_{t+1}^l = A^l Z_t^l + B^l \phi^l \left(\begin{bmatrix} Z_t^l \\ \bar{Z}_t \end{bmatrix} \right) + \omega_t^l$$

where $\omega_t^l \sim \mathcal{N}(0, \Sigma_w^l)$, $\Sigma_w^l \in \mathbb{R}^{m \times m}$, are drawn independently. Since the simulator uses a single agent to approximate the mean-field trajectory, the mean-field trajectory of population l is taken to be the state of the agent l, $\bar{Z}_t^l = Z_t^l$. Hence, the mean-field trajectory $\bar{Z}^l$ is generated by the dynamics

$$\bar{Z}_{t+1}^l = A^l \bar{Z}_t^l + B^l \phi^l \left(\begin{bmatrix} \bar{Z}_t^l \\ \bar{Z}_t \end{bmatrix} \right) + \omega_t^l. \tag{20}$$

Using this iterative process, the mean-field trajectory $\bar{Z}$ can be simulated under the controller ϕ. In the next section, we show that due to the affine (or linear) nature of ϕ^l, (20) will lead to $\bar{Z}$ having an affine (or linear) dynamics.

Having detailed the working of the simulator, we now outline the working of the two part RL algorithm. The first part of the algorithm approximates the equilibrium control gains $(K_{l,1}^*)_{l \in [L]}$ and the dynamics matrix of the equilibrium mean-field trajectory F^*. This is achieved by fixing the control offsets $K_{l,2} = 0$ for all l, which results in mean-field drift $C = 0$ as shown in the next subsection. During each iteration of the algorithm, the update of the control gains $K_{l,1}$ is achieved through ZSO (presented in Algorithm 2) and the mean-field dynamics matrix F is updated by the simulator. The second part of the algorithm approximates the equilibrium control offsets $K_{l,2}^*$ and the equilibrium mean-field drift C^*. The control gains $K_{l,1}$ are retained from the first part of the algorithm, which results in the mean-field dynamics matrix F also being fixed. Then, the control offsets $K_{l,2}$ are updated using ZSO, and the mean-field drift C is updated using the mean-field simulator.

4.1.1 Approximating Linear Terms

In the first part of Algorithm 1 (lines 2-6), each generic agent l keeps its control offset $K_{l,2} = 0$. This causes the mean-field drift term C in the mean-field trajectory dynamics to

be 0. Hence, the mean-field trajectory has stochastic linear dynamics. As a result, the cost of each generic agent l for a given $\bar{Z}$ can be reduced to the standard LQR cost plus a constant term:

$$J_l((K_{l,1}, 0), \bar{Z}) = J_l^1(K_{l,1}, \bar{Z}) + \bar{\beta}^{l\top} \bar{Q}^l \bar{\beta}^l. \tag{21}$$

Hence, the generic agent's problem for a given linear mean-field trajectory $\bar{Z}$ is an LQR problem.

Each generic agent uses ZSO (line 4) to approximate the optimal policy for this LQR problem. We briefly describe ZSO here; the reader can refer to Fazel et al. [12]; Malik et al. [22] for details. The ZSO algorithm (Algorithm 2) approximates the optimal policy of the LQR problem by stochastic gradient descent. The stochastic gradient in the ZSO algorithm is computed by using k_1 (mini-batch size) noisy estimates of the cost function (line 4 in Algorithm 2). Each noisy estimate is generated by long run cost of control law $K_{l,1}$ perturbed by a random perturbation from sphere $\mathbb{S}^1$ with radius r_1 (smoothing radius). The stochastic gradient descent is performed for R_1 iterations with constant learning rate η_1. We denote the policy obtained by the ZSO by $K'_{l,1}$.

Proving convergence of the ZSO algorithm is complicated by several factors among which are stability considerations, non-convexity of feasible set of controllers, and non-convexity of the cost itself. By proving local smoothness, Lipschitzness, and gradient domination properties of the cost function J_l^1, and ensuring that the ZSO algorithm is contained inside a feasibility region, the ZSO algorithm has been shown to converge to the optimal controller [12, 22].

Using the set of control gains $(K'_{l,1})_{l\in[L]}$, the simulator generates mean-field trajectory for each population $l \in [L]$ as shown in (20),

$$\begin{aligned}\bar{Z}_{t+1}^l &= A^l \bar{Z}_t^l - B^l K'_{l,1} \begin{bmatrix} \bar{Z}_t^l \\ \bar{Z}_t \end{bmatrix} + \omega_t^l \\ &= A^l \bar{Z}_t^l - B^l [K_{l,1}^{\prime,1},\ K_{l,1}^{\prime,2}] \begin{bmatrix} \bar{Z}_t^l \\ \bar{Z}_t \end{bmatrix} + \omega_t^l\end{aligned}$$

where $K'_{l,1} = [K_{l,1}^{\prime,1},\ K_{l,1}^{\prime,2}]$ such that $K_{l,1}^{\prime,1} \in \mathbb{R}^{p\times m}$ and $K_{l,1}^{\prime,2} \in \mathbb{R}^{p\times mL}$. By concatenating the mean-field trajectories of all populations, we obtain $\bar{Z} = (\bar{Z}^1, \ldots, \bar{Z}^L) \in \mathbb{R}^{mL}$, which follows linear dynamics of the form,

$$\bar{Z}_{t+1} = F\bar{Z}_t + \omega_t \text{ where } F := A - B(K_1^{\prime,1} + K_1^{\prime,2}), \tag{22}$$

where

$$\begin{gathered} A = \operatorname{diag}(A^1, \ldots, A^L),\ B = \operatorname{diag}(B^1, \ldots, B^L), \\ K_1^{\prime,1} = \operatorname{diag}(K_{1,1}^{\prime,1}, \ldots, K_{L,1}^{\prime,1}),\quad K_1^{\prime,2} = \begin{bmatrix} K_{1,1}^{\prime,2} \\ \vdots \\ K_{L,1}^{\prime,2} \end{bmatrix},\quad \omega_t = \begin{bmatrix} \omega_t^1 \\ \vdots \\ \omega_t^L \end{bmatrix}, \end{gathered} \tag{23}$$

where $\omega_t \sim \mathcal{N}(0, \Sigma)$, $\Sigma = \operatorname{diag}(\Sigma_w^1, \ldots, \Sigma_w^L)$. Hence, we have shown that if the control offsets $K_{l,2} = 0$, the mean-field drift $C = 0$ and the mean-field trajectory follows linear dynamics.

4.1.2 Approximating Offset Terms

In the second part of Algorithm 1 (lines 7-11), each generic agent l keeps the control gain $K_{l,1}$ fixed and iterates its control offset $K_{l,2}$. As a result, the mean-field dynamics matrix F remains constant but the drift C changes which will be shown later in the subsection. The problem faced by each generic agent l (15)-(16) is that of drifted LQ with constant tracking signal .

The generic agent l approximates the control offset by utilizing ZSO (line 9). The ZSO algorithm (Algorithm 2) performs R_2 iterations of stochastic gradient descent with learning rate η_2. The stochastic gradient is computed with mini-batch size k_2 and smoothing radius r_1. Let us denote the control offset obtained by ZSO as $K'_{l,2}$.

Proving convergence of the ZSO algorithm for a nonzero mean-field drift term C is further complicated (as compared with the drift-free case mentioned earlier) by several factors among which are stability considerations, non-convexity of feasible set of controllers, and non-convexity of the cost itself (as in Sect. 4.1.1). An additional challenge is introduced here by the fact that (unlike the previous section) the cost for a nonzero mean-field drift term C does satisfy the gradient domination condition. Rather, for this case we leverage local smoothness, Lipschitzness, and strong convexity properties of the cost function J_l^2, and ensure that the iterates of the ZSO algorithm are contained within a feasibility region. Lemma 2 presents the sample complexity guarantees of the ZSO algorithm for the nonzero drift term C.

The simulator uses the set of control laws $((K_{l,1}, K'_{l,2}))_{l\in[L]}$ to generate the mean-field trajectory for each population $l \in [L]$ as follows:

$$\bar{Z}^l_{t+1} = A^l \bar{Z}^l_t - B_i[K^1_{l,1},\ K^2_{l,1}] \begin{bmatrix} \bar{Z}^l_t \\ \bar{Z}_t \end{bmatrix} - B_i K'_{l,2} + \omega^l_t.$$

Hence, the joint mean-field trajectory follows affine dynamics,

$$\bar{Z}_{t+1} = F\bar{Z}_t + C' + \omega_t \text{ where } F := A - B(K^1_1 + K^2_1),\ C' = -BK'_2 \tag{24}$$

such that

$$K^1_1 = \text{diag}(K^1_{1,1}, \ldots, K^1_{L,1}),\ K^2_1 = \begin{bmatrix} K^2_{1,1} \\ \vdots \\ K^2_{L,1} \end{bmatrix},\ K'_2 = \begin{bmatrix} K'_{1,2} \\ \vdots \\ K'_{L,2} \end{bmatrix},\ \omega_t = \begin{bmatrix} \omega^1_t \\ \vdots \\ \omega^L_t \end{bmatrix}$$

where $\omega_t \sim \mathcal{N}(0, \Sigma)$, $\Sigma = \text{diag}(\Sigma^1_w, \ldots, \Sigma^L_w)$.

4.2 Finite Sample Bounds for Convergence of the RL Algorithm

We now present finite sample convergence analysis of Algorithm 1. We start by presenting convergence guarantees of ZSO for linear terms (Lemma 1) and for offset terms (Lemma 2). In Lemma 1, we prove finite sample bounds for ZSO on cost J^1_l of agent l using the control gain $K_{l,1}$ for a given linear mean-field trajectory $\bar{Z}$. The bounds depend on Lipschitz constant φ^l_1, smoothness constant λ^l_1, and gradient domination constant μ^l of J^1_l and its local radius ρ^l_1. The following Lemma provides a high confidence bound on the estimation error ϵ_1, given that the smoothing radius r_1, learning rate η_1, mini-batch size k_1, and iterations R_1 are chosen as specified.

Algorithm 1: RL for Multi-Population LQ-MFGs

Input: Number of iterations: S_1, S_2

Initialize: $(K_{l,1}^{(1)})_{l\in[L]}$ with stabilizing $K_{l,1}^{(1,1)}$ and $K_{l,1}^{(1,2)} = 0$, $K_{l,2}^{(0)} = 0$, $\bar{Z}^{(1)} = 0$

for $s \in \{1, \ldots, S_1 - 1\}$ **do**

▷ `Each generic agent performs ZSO to update` $K_{l,1}^{(s+1)}$

$$K_{l,1}^{(s+1)} = ZSO\big((K_{l,1}^{(1)}, K_{l,2}^{(0)}), 1, \bar{Z}^{(s)}, R_1, r_1, \eta_1, k_1\big)$$

Simulator uses $(K_{l,1}^{(s+1)}, K_{l,2}^{(0)})$ to obtain $\bar{Z}^{(s+1)}$.

end for

Initialize: $K_{l,2}^{(1)}$

for $s \in \{1, \ldots, S_2\}$ **do**

▷ `Each generic agent performs ZSO to obtain` $K_{l,2}^{(s+1)}$

$$K_{l,2}^{(s+1)} = ZSO((K_{l,1}^{(S_1)}, K_{l,2}^{(1)}), 2, \bar{Z}^{(s+S_1)}, R_2, r_2, \eta_2, k_2)$$

Simulator uses $(K_{l,1}^{(S_1)}, K_{l,2}^{(s+1)})$ to obtain $\bar{Z}^{(S_1+s+1)}$.

end for

Output: $\big(K_{l,1}^{(S_1)}, K_{l,2}^{(S_2)}\big)_{l\in[L]}, \bar{Z}^{(S_1+S_2)}$

Algorithm 2: Zero-order Stochastic Optimization (ZSO)

Input: Control gain $K_l^{(1)}$, controller number j, mean-field trajectory $\bar{Z}$, number of iterations R_j, smoothing radius r_j, step size η_j, mini-batch size k_j

for $r \in \{1, \ldots, R_j\}$ **do**

Generate $\tilde{K}_{l,j}^{(i)} \in \mathbb{S}^1(r_j)$ for all $i \in \{1, \ldots, k_j\}$.

Compute $\tilde{\nabla}_{K_{i,j}^{(r)}} J_l(K_l^{(r)}, \bar{Z})$

$$\tilde{\nabla}_{K_{i,j}^{(r)}} J_l(K_l^{(r)}, \bar{Z}) = \frac{1}{k_j}\sum_{l=1}^{k_j} \frac{mL}{r_j^2} J_l\big(K_{i,j}^{(r)} + r_j \tilde{K}_{l,j}^{(i)}, \bar{Z}\big)\tilde{K}_{l,j}^{(i)}$$

$K_{i,j}^{(r+1)} = K_{i,j}^{(r)} - \eta\tilde{\nabla}_{K_{i,j}^{(r)}} J_l(\cdot,\cdot)$

end for

Return: $K_{i,j}^{(R)}$

Lemma 1 *For a given linear mean-field trajectory $\bar{Z}$ and $\epsilon_1, \delta_1 > 0$, if the smoothing radius r_1, the learning rate η_1 and the mini-batch size k_1 are chosen such that*

$$r_1 = \frac{1}{8\varphi_1^l}\min\left(\theta_1^l\mu^l\sqrt{\frac{\epsilon_1}{240}}, \frac{1}{\varphi_1^l}\sqrt{\frac{\epsilon_1\mu^l}{30}}\right), \eta_1 = \min\left(1, \frac{1}{8\varphi_1^l}, \frac{\rho_1^l}{\sqrt{\mu^l}/32 + \varphi_1^l + \lambda_1^l}\right)$$

$$k_1 = 1024\frac{(mL)^2}{r_1^2}\left(J_l(K_i^{(0)}) + \frac{\lambda_1^l}{\rho_1}\right)^2 \log\left(\frac{2mL}{\delta}\right)\frac{1}{\mu^l\epsilon_1}$$

and the number of iterations is $R_1 = \frac{8}{\eta_1 \mu^l} \log\left(\frac{2(J_l^1(K_{l,1}^{(1)}) - J_l^1(K_{l,1}^*))}{\epsilon_1}\right)$, *then*

$$J_l^1(K_{l,1}^{(R_1)}) - J_l^1(\bar{K}_{l,1}^*) \leq \frac{\epsilon_1}{2},$$
$$\|K_{l,1}^{(R_1)} - \bar{K}_{l,1}^*\|_F \leq \sigma_{\min}^{-1}(\bar{\Sigma}^l)\sigma_{\min}^{-1}(C_U^l)\frac{\epsilon_1}{2}$$

with probability at least $1 - \delta_1 R_1$, *and the control gain* $\bar{K}_{l,1}^* = \operatorname{argmin}_{K_{l,1}} J_l^1(K_{l,1}, \bar{Z})$.

The first inequality in Lemma 1 can be obtained by combining Lemmas 1–3 and Theorem 2 in [22] and the second inequality is a consequence of Lemma D.4 of Fu et al. [13]. Next, we prove finite sample bounds for ZSO on J_l^2 for each $l \in [L]$ using control offset $K_{l,2}$ given an affine mean-field trajectory $\bar{Z}$ and control gain $K_{l,1}$. The bounds depend on the Lipschitz constant φ_2^l, smoothness constant λ_2^l, strong convexity constant ν^l of J_l^2, and local radius ρ_2^l. The Lemma provides a high confidence bound on the estimation error ϵ_2, given that the smoothing radius r_2, learning rate η_2, mini-batch size k_2, and inner loop iterations R_2 are chosen as specified.

Lemma 2 *For a given affine mean-field trajectory* $\bar{Z}$, *control gain* $K_{l,1}$ *and* $\epsilon_2, \delta_2 > 0$, *if the smoothing radius* r_2, *the learning rate* η_2 *and the mini-batch size* k_2 *are chosen such that*

$$r_2 = \min\left(1, \rho_2^l, \frac{\nu^l \epsilon_2}{32\varphi_2^l \lambda_2^l}\right), \qquad \eta_2 = \min\left(\frac{1}{\varphi_2^l}, \rho_2^l\left(\frac{\nu^l}{32} + \varphi_2^l + \lambda_2^l\right)^{-1}\right)$$
$$k_2 = 1024\frac{m^2}{r_2^2}\left(J_l(K_{l,2}^{(0)}) + \frac{\lambda_2^l}{\rho_2^l}\right)^2 \log\left(\frac{2m}{\delta}\right)\max\left(\frac{1}{\nu^l\epsilon_2}, \left(\frac{\lambda_2^l}{\nu^l\epsilon_2}\right)^2\right)$$

and the number of inner loop iterations is

$$R_2 = \frac{1}{\nu^l \eta_2}\log\left(\frac{4(J_l^2((K_{l,1}, K_{l,2}^{(1)}), \bar{Z}) - J_l^2((K_{l,1}, \bar{K}_{l,2}^*), \bar{Z}))}{\epsilon_2}\right)$$

then the difference between the output cost $J_l^2((K_{l,1}, K_{l,2}^{(R_2)}), \bar{Z})$ *and the optimal cost* $J_l^2((K_{l,1}, \bar{K}_{l,2}^*), \bar{Z})$ *is*

$$J_l^2((K_{l,1}, K_{l,2}^{(R_2)}), \bar{Z}) - J_l^2((K_{l,1}, \bar{K}_{l,2}^*), \bar{Z}) \leq \epsilon_2/2,$$
$$\|K_{l,2}^{(R_2)} - \bar{K}_{l,2}^*\|_2 \leq \sqrt{\frac{\epsilon_2}{\nu^l}}$$

with probability at least $1 - \delta_2 R_2$, *and* $\bar{K}_{l,2}^* = \operatorname{argmin}_{K_{l,2}} J_l^2((K_{l,1}, K_{l,2}), \bar{Z})$.

Proof of Lemma 2 can be found in the Appendix. The proof starts by computing the global Lipschitz, smoothness, and convexity constants for the cost function J_l^2. Using these constants, we prove finite sample bounds for the stochastic gradient descent algorithm.

Finally, we present the culmination of our analysis, as Theorem 3, which provides high confidence error bounds between the estimated MFE and the exact MFE given that the parameters are chosen appropriately. That is, for sufficiently small (given) $\epsilon, \delta > 0$, if we choose the parameters of ZSO as follows,

$$\epsilon_1 = \frac{(1 - T_1)\epsilon}{\|B\|_F \sum_{l\in[L]} \sigma_{\min}^{-1}(\bar{\Sigma}^l)\sigma_{\min}^{-1}(C_U^l)}, \delta_1 = \frac{\delta}{S_1 R_1}, \epsilon_2 = \epsilon^2, \delta_2 = \frac{\delta}{S_2 R_2} \tag{25}$$

then the RL algorithm will provide an arbitrarily accurate estimate of the MFE with high confidence.

Theorem 3 *Assume that*

$$T_1 := \left\| E^{-1} B R^{-1} B^T (I - H^k)^{-1} C_Z \right\|_2 < 1,$$

$$T_2 := \left\| E^{-1} B R^{-1} B^T \sum_{k=0}^{\infty} H^k C_Z (I - (F^*)^k)(I - F^*)^{-1} \right\|_2 < 1,$$

where B is defined in (23), $R := \mathrm{diag}(C_U^1, \ldots, C_U^L)$, and the remaining quantities, E, C_Z, H, are defined in the proof. If the outer loop iterations S_1 and S_2 are defined such that

$$S_1 = \frac{1}{1 - T_1} \log\left(\frac{2\|F^{(1)} - F^*\|_F}{\epsilon} \right), \quad S_2 = \frac{1}{1 - T_2} \log\left(\frac{2\|\bar{C}^{(1)} - \bar{C}^*\|_2}{\epsilon} \right), \tag{26}$$

$\epsilon_1, \delta_1, \epsilon_2, \delta_2$ are defined as in (25), and the parameters $r_1, r_2, \eta_1, \eta_2, k_1, k_2, R_1, R_2$ are defined as in the statements of Lemmas 1 and 2, then the error between the approximate MFE $\left(\left(K_{l,1}^{(S_1)}, K_{l,2}^{(S_2)} \right)_{l \in [L]}, \bar{Z}^{(S_1+S_2)} \right)$ and the MFE $\left(\left(K_l^* \right)_{l \in [L]}, \bar{Z}^* \right)$ is

$$\|F^{(S_1)} - F^*\|_F \leq \epsilon, \quad \|K_{l,1}^{(S_1)} - K_{l,1}^*\|_F \leq D_l^2 \epsilon \tag{27}$$

and $F^{(s)}$ are stable $\forall s \in [S_1]$ with probability at least $1 - \delta$. Furthermore if $\epsilon \leq \min\left(1, \frac{1-T_2}{2D^3\|B\|_2}\right)$, then

$$\|C^{(S_2)} - C^*\|_2 \leq D^4 \epsilon, \quad \|K_{l,2}^{(S_2+1)} - K_{l,2}^*\|_2 \leq D_l^5 \epsilon, \forall l \in [L] \tag{28}$$

with probability at least $1 - 2\delta$.

Proof Proof of this result can be found in the Appendix, where also the expressions for the scalar constants D_l^2, D^3, D^4, and D_l^5 are given. □

Note that the assumption on the quantities T_1 and T_2 in Theorem 3 ensures that the mean-field update is contractive; it is similar to standard assumptions found in Fu et al. [13], Zaman et al. [29] and Saldi et al. [24]. As a result, iterative application of the joint operator $\circ$ will lead to the MFE of the multi-population MFG. Algorithm 1 can thus be viewed as a data-driven version of this operator $\circ$.

The proof of Theorem 3 is presented in two parts. The first part of the proof proves finite sample bounds for the linear terms in the MFE. Provided that the estimation errors of the linear terms are sufficiently small (as characterized in the proof), the second part of the proof establishes finite sample bounds on the affine terms of the MFE (and thus the complete expressions). Theorem 3 requires that the inner loop errors (ϵ_1 and ϵ_2) be bounded by functions of ϵ. These bounds along with the contractive property of the mean-field update $\circ$ ensure that the approximate MFE computed by Algorithm 1 falls within the final error bound ϵ.

4.3 Experiments

The performance of the learning algorithm (Algorithm 1) is evaluated in the context of two small sample networks. The first network, termed a *chain network*, consists of three populations connected in a line, e.g., $1 - 2 - 3$. The second network, termed a *ring network*, consists of three populations that are fully connected. The state and action spaces of the agents are assumed to be scalar in both networks. Furthermore, the underlying dynamics, costs, and model parameters are chosen so as to satisfy the assumptions of Theorem 3. The

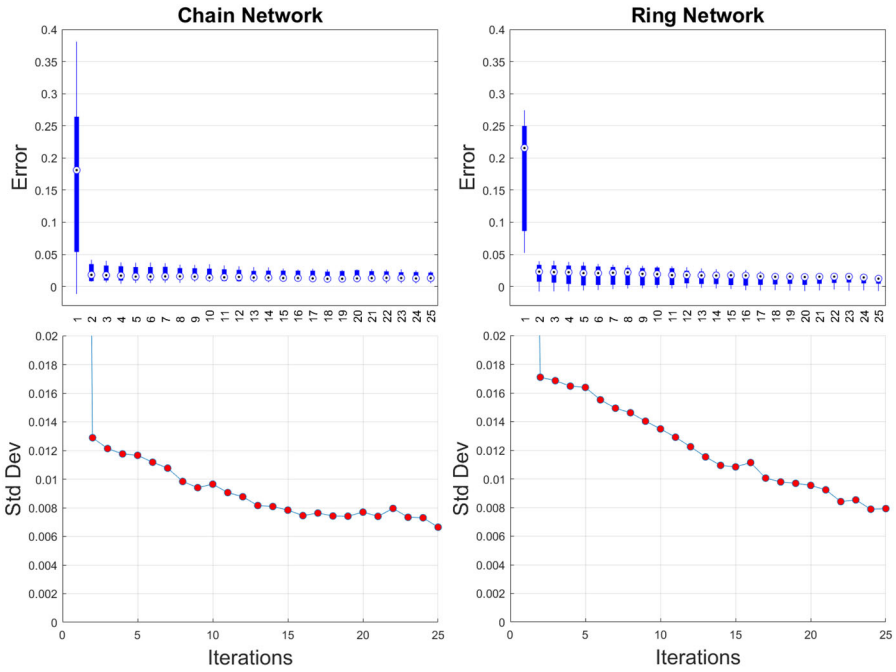

Fig. 1 Approximation error convergence (top) and standard deviation (bottom) for the chain and ring networks. Plots were averaged over 10 runs, 25 iterations each, of Algorithm 1. Markers in the boxplots show the median, with the box indicating the 25'th and 75'th percentile (whiskers cover the full range of the data). Both the number of iterations and number of rollouts in the ZSO algorithm 2 were set to 1500

top two plots in Fig. 1 illustrate the approximation error between the approximate mean-field trajectory $\bar{Z}^s$ and the equilibrium mean-field trajectory $\bar{Z}^*$ for the two networks. The bottom two plots depict the standard deviation of the approximation for both networks.

The plots show a fast initial drop which slows to a gradual decrease in error and standard deviation. This behavior is aligned with the behavior of the *exact* mean-field update $\circ$. The ZSO algorithm yields very good control gain and offset estimates, which enables the data-driven RL algorithm to closely mimic the exact algorithm.

Even though the observed effect is small in the sample networks, the degree of connectivity of the populations has an influence on the variance of the learning algorithm. The more densely connected the network (among populations) is, the higher is the variance of the learning algorithm. This is intuitive given that generic agents must *negotiate* with higher number of neighbors in dense networks as compared to sparse networks.

5 Concluding Remarks and Future Directions

This paper has approached the problem of multi-agent reinforcement learning from the perspective of mean-field games. We proposed a model that consists of multiple heterogeneous populations, each containing homogeneous agents, that are connected via a network. Our results establish existence, uniqueness, and the affine structure of the mean-field equilibrium. Furthermore, we quantified how the mean-field equilibrium performs in the original

finite-population game as a function of the population sizes. The proposed reinforcement learning algorithm, based on zero-order stochastic optimization, leverages the affine form of the equilibrium by individually estimating the linear and affine terms.[6] Finite sample convergence results of the algorithm are presented. While the current paper establishes ϵ-Nash bounds for the mean-field equilibrium (recall the third listed contribution), we do not establish here bounds for the approximate mean-field equilibrium, i.e., the equilibrium computed by the learning algorithm. This is beyond the scope of the current work and is left for future investigation.

Additional future work includes generalization of both the model and the learning algorithm. One modeling extension would be to consider settings where the populations are not taken as given (as assumed in the present paper), but rather learned by clustering agents based on similarity of dynamics and costs. Another direction is to investigate how the degree of heterogeneity among the populations, i.e., in their dynamics and costs, influences the convergence properties of the algorithm. Lastly, as with any algorithm applied to scenarios with strategic agents, a natural question that arises is whether the agents would abide by the rules of the algorithm. An interesting (and challenging) question is to evaluate how disobedience by a group of agents impacts the learned equilibrium, and investigating how agents can be incentivized to abide by the algorithm (by leveraging tools from mechanism design).

Acknowledgements We thank the anonymous reviewers for their useful suggestions. We also thank Dr. Kaiqing Zhang for his technical expertise and useful discussions.

Data Availability Statement: The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Appendix

Proof of Proposition 1

Proof We organize the proof of Proposition 1 in two parts. In the first part we prove the existence and uniqueness of the equilibrium mean-field $\bar{Z}$. In the second part we prove the form of the equilibrium controller.

Part I: The Hamiltonian function for generic agent in population l given a mean-field trajectory $\bar{Z} = (\bar{Z}_0, \bar{Z}_1, \ldots)$ is

$$H_t^l(Z_t^l, U_t^l, \bar{Z}_t, \zeta_t^l) = \frac{1}{2}\Big(\big\|Z_t^l\big\|_{Q^l}^2 + \big\|U_t^l\big\|_{C_U^l}^2 + \big\|Z_t^l - \bar{Z}_t^l\big\|_{C_Z^{ll}}^2 + \sum_{k\in\mathcal{L}_l} \big\|Z_t^l - \bar{Z}_t^k - \beta^{lk}\big\|_{C_Z^{lk}}^2\Big) + (\zeta_{t+1}^l)^\top (A^l Z_t^l + B^l U_t^l + W_t^l) \tag{29}$$

for $l \in \mathcal{L}$ where ζ_t^l is the co-state variable. Notice that we have scaled the cost function by $1/2$ to simplify expressions. Recall the dynamics of generic agent in population l,

$$Z_{t+1}^l = A^l Z_t^l + B^l U_t^l + W_t^l. \tag{30}$$

[6] This is in contrast to earlier work [13] where only the affine terms needed update due to the assumed stationarity of the MFE.

The co-state dynamics and the optimal control are obtained using the Hamiltonian and the conditions for optimality,

$$\begin{aligned}\frac{\partial H_t^l}{\partial Z_t^l} &= A^{l\top}\zeta_{t+1}^l + Q^l Z_t^l + \sum_{k\in\mathcal{L}_l} C_Z^{lk}(Z_t^l - \bar{Z}_t^k - \beta^{lk}) - \Delta M_t^l = \zeta_t^l,\\ \frac{\partial H_t^l}{\partial U_t^l} &= C_U^l U_t^l + B^{l\top}\zeta_{t+1}^l = 0 \implies U_t^l = -(C_U^l)^{-1}B^{l\top}\zeta_{t+1}^l,\end{aligned} \tag{31}$$

where ΔM_t^l is a Martingale difference sequence,

$$\Delta M_t^l = A^{l\top}\zeta_{t+1}^l - A^{l\top}\mathbb{E}[\zeta_{t+1}^l \mid \mathcal{F}_t^l]$$

where $\mathcal{F}_t^l$ is the σ-algebra generated by process Z_t^l up to time t and $\beta^{ll} = 0$. Aggregating the dynamics of the generic agent and co-state (with the assumption that the tracking signal $\bar{Z}^l$ is also the expected trajectory of Z_t^l) we obtain the dynamics of equilibrium mean-field of population l denoted by $\bar{Z}^{l*}$ and aggregated co-state $\bar{\zeta}^{l*}$,

$$\begin{aligned}\bar{Z}_{t+1}^{l*} &= A^l\bar{Z}_t^{l*} - B^l(C_U^l)^{-1}B^{l\top}\bar{\zeta}_{t+1}^{l*},\\ \bar{\zeta}_t^{l*} &= A^{l\top}\bar{\zeta}_{t+1}^{l*} + Q^l\bar{Z}_t^{l*} + \sum_{k\in\mathcal{L}_l} C_Z^{lk}\big(\bar{Z}_t^{l*} - \bar{Z}_t^{k*} - \beta^{lk}\big).\end{aligned} \tag{32}$$

Defining the mean-field trajectory of the system as $\bar{Z}_t^* = (\bar{Z}_t^{1*}, \ldots, \bar{Z}_t^{L*}) \in \mathbb{R}^{mL}$ and the co-state trajectory of the system as $\bar{\zeta}_t^* = (\bar{\zeta}_t^{1*}, \ldots, \bar{\zeta}_t^{L*}) \in \mathbb{R}^{mL}$, their dynamics are

$$\bar{Z}_{t+1}^* = A\bar{Z}_t^* - BR^{-1}B^\top\bar{\zeta}_{t+1}^*, \tag{33}$$

$$\bar{\zeta}_t^* = A^\top\bar{\zeta}_{t+1}^* + Q\bar{Z}_t^* - \hat{\beta}, \tag{34}$$

where A and B are defined in (23) and

$$\begin{aligned}&R = \operatorname{diag}(C_U^1, \ldots, C_U^L),\\ &\hat{\beta} = \begin{bmatrix}\sum_{k\in\mathcal{L}_1} C_Z^{1k}\beta^{1k}\\ \vdots\\ \sum_{k\in\mathcal{L}_L} C_Z^{Lk}\beta^{Lk}\end{bmatrix},\ Q = \begin{bmatrix}Q^1 + \sum_{k\in\mathcal{L}_1} C_Z^{1k} & \cdots & -C_Z^{1L}\\ -C_z^{21} & \cdots & -C_Z^{2L}\\ \vdots & \ddots & \vdots\\ -C_Z^{L1} & \cdots & Q^L + \sum_{k\in\mathcal{L}_L} C_Z^{Lk}\end{bmatrix}\end{aligned} \tag{35}$$

To solve the set of Eqs. (33)-(34), we use the sweep method [6] and assume $\bar{\zeta}_t^*$ has the form $\bar{\zeta}_t^* = S_t\bar{Z}_t^* + L_t$. Under this assumption Eq. (33) yields,

$$\begin{aligned}\bar{Z}_{t+1}^* &= A\bar{Z}_t^* - BR^{-1}B^\top(S_{t+1}\bar{Z}_{t+1}^* + L_{t+1})\\ \bar{Z}_{t+1}^* &= (I + BR^{-1}B^\top S_{t+1})^{-1}\big(A\bar{Z}_t^* - BR^{-1}B^\top L_{t+1}\big),\end{aligned} \tag{36}$$

and Eq. (34) results in

$$S_t\bar{Z}_t^* + L_t = A^\top(S_{t+1}\bar{Z}_{t+1}^* + L_{t+1}) + Q\bar{Z}_t^* - \hat{\beta}. \tag{37}$$

Substituting Eq. (36) into (37) yields

$$\begin{aligned}S_t\bar{Z}_t^* + L_t = {}& A^\top L_{t+1} + Q\bar{Z}_t^* - \hat{\beta} +\\ & A^\top S_{t+1}(I + BR^{-1}B^\top S_{t+1})^{-1}\big(A\bar{Z}_t^* - BR^{-1}B^\top L_{t+1}\big).\end{aligned} \tag{38}$$

Comparing coefficients of $\bar{Z}_t^*$ yields

$$\begin{aligned} S_t &= Q + A^\top S_{t+1}(I + BR^{-1}B^\top S_{t+1})^{-1}A \\ S_t &= A^\top S_{t+1}A + Q - A^\top S_{t+1}B(R + B^\top S_{t+1}B)^{-1}B^\top S_{t+1}A \end{aligned} \tag{39}$$

where the last equation is obtained using the Woodbury Matrix Identity. Comparing the remaining terms, we obtain the independent backwards process

$$\begin{aligned} L_t &= A^\top(I - S_{t+1}(I + BR^{-1}B^\top S_{t+1})^{-1}BR^{-1}B^\top)L_{t+1} - \hat{\beta} \\ &= A^\top(I + S_{t+1}BR^{-1}B^\top)^{-1}L_{t+1} - \hat{\beta}. \end{aligned} \tag{40}$$

For the infinite horizon case, consider the limiting Algebraic Riccati Equation,

$$S = A^\top SA - A^\top SB(R + B^\top SB)^{-1}B^\top SA + Q. \tag{41}$$

If a unique solution to the above ARE exists and the matrix $A^\top(I + SBR^{-1}B^\top)$ is stable, then the limiting backwards Eq. (40) becomes

$$L = A^\top(I + SBR^{-1}B^\top)^{-1}L - \beta. \tag{42}$$

If the Riccati Eq. (41) admits a unique positive definite solution S, then the MFE will be unique, linear and follow dynamics,

$$\begin{aligned} \bar{Z}_{t+1}^* &= (I + BR^{-1}B^\top S)^{-1}\left(A\bar{Z}_t^* - BR^{-1}B^\top L\right) \\ &= F^*\bar{Z}_t^* + C^* \end{aligned} \tag{43}$$

where $F^* = (I + BR^{-1}B^\top S)^{-1}A$ and $C^* = -(I + BR^{-1}B^\top S)^{-1}BR^{-1}B^\top L$.

Now we prove that the ARE (41) has a unique positive definite solution. We split up the matrix $Q = Q_1 + Q_2$ such that,

$$Q_1 = \operatorname{diag}(Q^1, \ldots, Q^L), \tag{44}$$

$$Q_2 = \begin{bmatrix} \sum_{k\in\mathcal{L}_1} C_Z^{1k} & \cdots & -C_Z^{1L} \\ \vdots & \ddots & \vdots \\ -C_Z^{L1} & \cdots & \sum_{k\in\mathcal{L}_L} C_Z^{Lk} \end{bmatrix}. \tag{45}$$

Both matrices Q_1 and Q_2 are symmetric positive semi-definite, since Q^l and $C_Z^{lk} = C_Z^{kl}$ are symmetric positive semi-definite matrices.

As a consequence of the observability of the pair $(A^l, Q_l^{1/2})$, the pair $(A, Q_1^{1/2})$ is also observable. Hence, for any vector x in the eigenspace of A, $x^\top Q_1 x > 0$. For such a vector,

$$x^\top(Q_1 + Q_2)x > 0 \tag{46}$$

since $Q_2 \geq 0$. This in turn implies that pair

$$\left(A, (Q_1 + Q_2)^{1/2}\right) = \left(A, Q^{1/2}\right) \tag{47}$$

is also observable. Finally for all $l \in [L], C_U^l > 0 \implies R > 0$ and (A^l, B^l) being controllable for all $l \in [L]$ implies (A, B) is controllable. This is a sufficient condition for the existence and uniqueness of the solution to Riccati Eq. (41). Moreover, F^* is also stable. Due to this the matrix in Eq. (42), $A^\top(I + SBR^{-1}B^\top)^{-1} = F^{*\top}$ is also stable; hence, a unique L satisfies the limiting Eq. (42). And hence due to Theorem 3.34 in Carmona and Delarue, [9] there exists a unique equilibrium mean-field trajectory given by (43).

Part II: To prove existence and uniqueness (and characterization) of the equilibrium controller, we formulate the problem (12)-(13) as a tracking control problem with the mean-field trajectory $\bar{Z}^*$ given. Restating the conditions of optimality (31),

$$\begin{aligned}
\frac{\partial H_t^l}{\partial Z_t^l} &= {A^l}^\top \zeta_{t+1}^l + Q^l Z_t^l + \sum_{k\in\mathcal{L}_l} C_Z^{lk}(Z_t^l - \bar{Z}_t^k - \beta^{lk}) - \Delta M_t^l = \zeta_t^l,\\
\frac{\partial H_t^l}{\partial U_t^l} &= C_U^l U_t^l + {B^l}^\top \zeta_{t+1}^l = 0 \implies U_t^l = -(C_U^l)^{-1}{B^l}^\top \zeta_{t+1}^l
\end{aligned} \tag{48}$$

where ΔM_t^l is a Martingale difference sequence,

$$\Delta M_t^l = {A^l}^\top \zeta_{t+1}^l - \mathbb{E}[{A^l}^\top \zeta_{t+1}^l \mid \mathcal{F}_t^l]$$

where $\mathcal{F}_t^l$ is the σ-algebra generated by process Z_t^l up to time t and $\beta^{ll} = 0$. Assuming the form of the co-state $\zeta_t^l = P_t^l Z_t^l + s_t^l$ and substituting into the equations we obtain

$$\begin{aligned}
Z_{t+1}^l &= (I + B^l(C_U^l)^{-1}{B^l}^\top P_{t+1}^l)^{-1}(A^l Z_t^l - B^l(C_U^l)^{-1}{B^l}^\top s_{t+1}^l),\\
P_t^l Z_t^l + s_t^l &= {A^l}^\top(P_{t+1}^l \mathbb{E}[Z_{t+1}^l \mid \mathcal{F}_t^l] + s_{t+1}^l) + Q^l Z_t^l + \sum_{k\in\mathcal{L}_l} C_Z^{lk}(Z_t^l - \bar{Z}_t^{k*}).
\end{aligned}$$

Substituting the first equation into the second yields

$$\begin{aligned}
P_t^l Z_t^l + s_t^l =& {A^l}^\top(P_{t+1}^l(I + B^l(C_U^l)^{-1}{B^l}^\top P_{t+1}^l)^{-1}(A^l Z_t^l - B^l(C_U^l)^{-1}{B^l}^\top s_{t+1}^l)\\
&+ s_{t+1}^l) + Q^l Z_t^l + \sum_{k\in\mathcal{L}_l} C_Z^{lk}(Z_t^l - \bar{Z}_t^{k*} - \beta^{lk}).
\end{aligned}$$

Comparing coefficients of Z_t^l yields the Riccati equation

$$P_t^l = {A^l}^\top P_{t+1}^l(I + B^l(C_U^l)^{-1}{B^l}^\top P_{t+1}^l)^{-1}A^l + Q^l + \sum_{k\in\mathcal{L}_l} C_Z^{lk} \tag{49}$$

and comparing the remaining terms yields a backwards recursive expression for s_t^l,

$$\begin{aligned}
s_t^l =& -{A^l}^\top\big(P_{t+1}^l(I + B^l(C_U^l)^{-1}{B^l}^\top P_{t+1}^l)^{-1}B^l(C_U^l)^{-1}{B^l}^\top + I\big)s_{t+1}^l\\
&- \sum_{k\in\mathcal{L}_l} C_Z^{lk}(\bar{Z}_t^{k*} + \beta^{lk}).
\end{aligned} \tag{50}$$

The infinite horizon Riccati equation is

$$P^l = {A^l}^\top P^l(I + B^l(C_U^l)^{-1}{B^l}^\top P^l)^{-1}A^l + Q^l + \sum_{k\in\mathcal{L}_l} C_Z^{lk}. \tag{51}$$

Given the above, if the pair $(A^l, (Q^l + \sum C_Z^{lk})^{1/2})$ is observable, then the Riccati equation will have unique positive definite solution P^l. Since the pair $(A^l, (Q^l)^{1/2})$ is observable, the pair $(A^l, (Q^l + \sum C_Z^{lk})^{1/2})$ is also observable. The reason for that is that for any vector x in the eigenspace of A^l, $x^\top Q^l x > 0$ due to the observability of the pair $(A^l, (Q^l)^{1/2})$. This implies that $x^\top(Q^l + \sum C_Z^{lk})x > 0$ as $C_Z^{lk} \geq 0$. This implies that the pair $(A^l, (Q^l + \sum C_Z^{lk})^{1/2})$ is also observable. Hence, there exists a unique positive definite P^l that satisfies the Riccati equation.

Now we characterize the form of equilibrium control law. Using (50) we obtain

$$s_t^l = -\sum_{k\in[L]} C_Z^{lk}(\bar{Z}_t^{k*} + \beta^{lk}) + H^l s_{t+1}^l \tag{52}$$

where

$$H^l = ((E^l)^{-1}A^l)^\top \text{ and } E^l = I + B^l(C_U^l)^{-1}B^{l\top}P^l \tag{53}$$

where P^l is the solution to the Riccati Eq. (51) and $\bar{Z}^{k*}$ represents the k'th population's mean-field trajectory from $\bar{Z}^*$. The stability of the sequence s_t^l is dependent on the matrix H^l being stable. We know that the matrix $(H^l)^\top = (E^l)^{-1}A^l$ is the closed-loop gain matrix of the LQR system $(A^l, B^l, Q^l + \sum_k C_Z^{lk}, C_U^l)$. This matrix is bound to be stable since its corresponding Riccati equation of the LQR system (51) has a unique positive definite solution. As a result, the matrix H^l is also stable and hence the sequence s_t^l is bounded. This yields the existence and uniqueness of the equilibrium controller. The closed-loop dynamics of generic agent l are thus

$$\begin{aligned}
Z_{t+1}^l &= H^{l\top}Z_t^l - (E^l)^{-1}B^l(C_U^l)^{-1}B^{l\top}s_{t+1}^l,\\
&= H^{l\top}Z_t^l - (E^l)^{-1}B^l(C_U^l)^{-1}B^{l\top}\sum_{i=0}^{\infty}(H^l)^i\sum_{k\in[L]}C_Z^{lk}(\bar{Z}_{t+i+1}^{k*} + \beta^{lk}),\\
&= H^{l\top}Z_t^l - (E^l)^{-1}B^l(C_U^l)^{-1}B^{l\top}\sum_{i=0}^{\infty}(H^l)^i C_Z^l(\bar{Z}_{t+i+1}^{*} + \beta^{l})
\end{aligned}$$

where $C_Z^l = (C_Z^{l1}, C_Z^{l2}, \ldots)$ and $\beta^l = (\beta^{l1}, \ldots, \beta^{lL}) \in \mathbb{R}^{mL}$. Since $\bar{Z}^{(s)}$ is assumed to follow affine dynamics $\bar{Z}_{t+1}^* = F^*\bar{Z}_t^* + C^*$, the above can further be simplified to

$$\begin{aligned}
Z_{t+1}^l &= H^{l\top}Z_t^l\\
&- (E^l)^{-1}B^l(C_U^l)^{-1}B^{l\top}\sum_{i=0}^{\infty}(H^l)^i C_Z^l((F^*)^{i+1}\bar{Z}_t^* + (I-(F^*)^i)(I-(F^*))^{-1}C^* + \beta^l).
\end{aligned}$$

Rewriting in terms of the controller $(K_{l,1}^*, K_{l,2}^*)$,

$$Z_{t+1}^l = A^l Z_t^l - B^l K_{l,1}^* \begin{bmatrix} Z_t^l \\ \bar{Z}_t^* \end{bmatrix} - B^l K_{l,2}^* \tag{54}$$

where

$$K_{l,1}^* = \begin{bmatrix} G^l A^l & (I - G^l B^l)(C_U^l)^{-1}B^{l\top}\sum_{i=0}^{\infty}(H^l)^i C_Z^l (F^*)^{i+1} \end{bmatrix} \tag{55}$$

where $G^l = (C_U^l + B^{l\top}P^l B^l)^{-1}B^{l\top}P^l$ and

$$K_{l,2}^* = (I - G^l B^l)(C_U^l)^{-1}B^l\sum_{i=0}^{\infty}(H^l)^i C_Z^l((I-(F^*)^i)(I-(F^*))^{-1}C^{(s)} + \beta^l) \tag{56}$$

which completes the proof. □

Proof of Theorem 2

Proof The following investigates the dependence of the ϵ-Nash bound $J_{n,l}^{(N)}(\tilde{\phi}) - \inf_{\pi^n \in \Pi^n} J_{n,l}^{(N)}((\pi^{n,l}, \tilde{\phi}^{-n,l}), \tilde{\phi}^{-l})$ on $N = (N_l)_{l \in L}$. We begin by writing the above quantity as

$$\begin{aligned} &J_{n,l}^{(N)}(\tilde{\phi}) - \inf_{\pi^n \in \Pi^n} J_{n,l}^{(N)}((\pi^{n,l}, \tilde{\phi}^{-n,l}), \tilde{\phi}^{-l}) = \\ &\quad J_{n,l}^{(N)}(\tilde{\phi}) - J_l(\phi^{l*}, \bar{Z}^*) + J_l(\phi^{l*}, \bar{Z}^*) \\ &\quad - \inf_{\pi^n \in \Pi^n} J_{n,l}^{(N)}((\pi^{n,l}, \tilde{\phi}^{-n,l}), \tilde{\phi}^{-l}). \end{aligned} \tag{57}$$

The first expression on the RHS of equation (57) can be bounded as follows

$$\begin{aligned} J_{n,l}^{(N)}(\tilde{\phi}) &\le \limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\Big[\|Z_t^{l*}\|_{Q^l}^2 + \|U_t^{l*}\|_{C_U^l}^2 + \sum_{k \in \mathcal{L}_l} \big(\|Z_t^{l*} - \bar{Z}_t^{k*} - \beta^{lk}\|_{C_Z^{lk}}^2 \\ &\quad + \|\bar{Z}_t^{k*} - Y_t^{k*}\|_{C_Z^{lk}}^2 + 2(Z_t^{l*} - \bar{Z}_t^{k*} - \beta^{lk})^\top C_Z^{lk} (\bar{Z}_t^{k*} - Y_t^{k*}) \big) \Big] \\ &= J_l(\phi^{l*}, \bar{Z}^*) + \limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{k \in \mathcal{L}_l} \mathbb{E}\big[\|\bar{Z}_t^{k*} - Y_t^{k*}\|_{C_Z^{lk}}^2 \\ &\quad + 2(Z_t^{l*} - \bar{Z}_t^{k*} - \beta^{lk})^\top C_Z^{lk} (\bar{Z}_t^{k*} - Y_t^{k*}) \big] \\ &\le J_l(\phi^{l*}, \bar{Z}^*) + \mathcal{O}\Big(\sum_{k \in \mathcal{L}_l} \sqrt{\limsup_{T\to\infty} \varepsilon_T^k} \Big) \end{aligned} \tag{58}$$

where $\beta^{ll} = 0$ and Y_t^{l*}, Y_t^{k*} are the empirical mean-field trajectories

$$Y_t^{l*} = \frac{1}{N_l - 1} \sum_{\substack{n' \in [N_l] \\ n' \neq n}} Z_t^{n',l}, \qquad Y_t^{k*} = \frac{1}{N_k} \sum_{n' \in [N_k]} Z_t^{n',k} \tag{59}$$

of populations l and $k \in \mathcal{L}_l \setminus \{l\}$, respectively, under equilibrium controller $\tilde{\phi}$, and

$$\varepsilon_T^k = \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\|\bar{Z}_t^{k*} - Y_t^{k*}\|_{C_Z^{lk}}^2 \tag{60}$$

for $k \in \mathcal{L}_l$. The last inequality in (58) is due to the fact that $(\phi^{l*})_{l \in [L]}$ are stabilizing control laws and $\bar{Z}^*$ is also stable. Using techniques similar to proof of Theorem 2 in Zaman et al. [29], we will now bound $\sum_{k \in \mathcal{L}_l} \sqrt{\limsup_{T\to\infty} \varepsilon_T^k}$ for $k \in \mathcal{L}_l$. The dynamics of the empirical mean-field trajectory Y_t^{k*} can be expressed using (59), (9), and the form of the equilibrium control law from Proposition 1,

$$Y_{t+1}^{k*} = (A^k - B^k K_{k,1}^{1*}) Y_t^{k*} - B^k K_{k,1}^{2*} \bar{Z}^* - K_{k,2}^* + \hat{\omega}_t^k$$

where $\hat{\omega}_t^k := \sum_{n' \in [N_k]} W_t^{n',k} / N_k$ is a Gaussian random variable with zero mean and covariance Σ_w^k / N_k. The covariance matrix for the stationary distribution of Y_t^{k*}, denoted by $\hat{\sigma^k}$, is the solution to the Lyapunov equation

$$\hat{\sigma^k} = \Sigma_w^k / N_k + (A^k - B^k K_{k,1}^{1*}) \hat{\sigma^k} (A^k - B^k K_{k,1}^{1*})^\top$$

hence

$$\mathrm{Tr}(\hat{\sigma^k}) = \mathcal{O}(1/N_k). \tag{61}$$

Next, we define $Y^* := (Y_t^{1*}, \ldots, Y_t^{L*}) \in \mathbb{R}^{mL}$ as the joint empirical mean-field trajectory under equilibrium controller with dynamics

$$Y_{t+1}^* = (A - BK_1^{1*})Y_t^* - BK_1^{2*}\bar{Z}_t^* - BK_2^* + \hat{\omega}_t$$

where A, B are defined in (35) ,

$$K_1^{1*} = \mathrm{diag}(K_{1,1}^{1*}, \ldots, K_{L,1}^{1*}), \quad K_1^{2*} = \begin{bmatrix} K_{1,1}^{2*} \\ \vdots \\ K_{L,1}^{2*} \end{bmatrix}, K_2^* = \begin{bmatrix} K_{1,2}^{*} \\ \vdots \\ K_{L,2}^{*} \end{bmatrix}$$

and $\hat{\omega}_t = (\hat{\omega}_t^1, \ldots, \hat{\omega_t^L}) \in \mathbb{R}^{mL}$. Note that $F^* = A - B(K_1^{1*} + K_1^{2*})$ and $C^* = -BK_2^*$ since the equilibrium mean-field trajectory is generated by the equilibrium mean-field controller. Consequently, the stationary distribution of Y_t^* is $\bar{Z}_\infty^*$ where $\bar{Z}_\infty^* := \lim_{t\to\infty} \bar{Z}_t^*$. As a result, $\mathbb{E}[Y_t^*] - \bar{Z}_t^* \to 0$ as $t \to \infty$ which implies

$$\mathbb{E}[Y_t^{k*}] - \bar{Z}_t^{k*} \to 0, \tag{62}$$

for all $l \in [L]$. Using (62) and (61),

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\|\bar{Z}_t^{k*} - Y_t^{k*}\|^2_{C_Z^{lk}} = \mathbb{E}_{Y_t^{k*}\sim\mathcal{N}(\bar{Z}_\infty^{k*}, \hat{\sigma^k})} \|\bar{Z}_t^{k*} - Y_t^{k*}\|^2_{C_Z^{lk}} = \mathcal{O}\Big(\frac{1}{N_k}\Big)$$

and so

$$\sum_{k\in\mathcal{L}_l} \sqrt{\limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\|\bar{Z}_t^{k*} - Y_t^{k*}\|^2_{C_Z^{lk}}} = \mathcal{O}\Big(1/\sqrt{\min_{k\in\mathcal{L}_l} N_k}\Big).$$

Hence, we have the first inequality,

$$J_{n,l}^{(N)}(\tilde{\phi}) - J_l(\phi^{l*}, \bar{Z}^*) = \mathcal{O}\left(1/\sqrt{\min_{k\in\mathcal{L}_l} N_k}\right)$$

Next, for the second term in (57), we denote the trajectory of agent n in population l which minimizes the following cost by $Z_t^{n,l}$,

$$\begin{aligned}
&\inf_{\pi^n\in\Pi^n} J_{n,l}^{(N)}((\pi^{n,l}, \tilde{\phi}^{-n,l}), \tilde{\phi}^{-l}) \\
&= \limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\big[\|Z_t^{n,l}\|^2_{Q^l} + \|U_t^{n,l}\|^2_{C_U^l} + \sum_{k\in\mathcal{L}_l} \|Z_t^{n,l} - Y_t^k - \beta^{lk}\|^2_{C_Z^{lk}}\big] \\
&\geq J_l(\phi^{l*}, \bar{Z}^*) + \limsup_{T\to\infty} \frac{2}{T} \sum_{t=0}^{T-1} (Z_t^{i*} - \bar{Z}_t^{k*} - \beta^{lk})^\top C_Z^{lk} (\bar{Z}_t^{k*} - Y_t^{k*}).
\end{aligned}$$

Using the same process as in (58) we arrive at,

$$J_l(\phi^{l*}, \bar{Z}^*) - \inf_{\pi^n\in\Pi^n} J_{n,l}^{(N)}((\pi^{n,l}, \tilde{\phi}^{-n,l}), \tilde{\phi}^{-l}) = \mathcal{O}\left(1/\sqrt{\min_{k\in\mathcal{L}_l} N_k}\right)$$

which concludes the proof. □

Proof of Lemma 2

Proof Let $J_l^2(K_{l,2}^{(r)})$ denote the cost of the control offset $K_{l,2}^{(r)}$ in the controller. This is the abridged version of the real cost $J_l^2((K_{l,1}, K_{l,2}^{(r)}), \bar{Z})$ but since none of the other parameters are changing we can disregard them for this proof. Let $\bar{K}_{l,2}^*$ denote the control offset which minimizes cost $J_l^2(K_{l,2})$. First we study some properties of cost J_l^2.

Let us define two sublevel sets based on the initial cost,

$$\mathcal{G}_l^0 := \{K_{l,2} \mid J_l^2(K_{l,2}) \leq 4J_l^2(K_{l,2}^{(1)})\},$$
$$\mathcal{G}_l^1 := \{K_{l,2} \mid J_l^2(K_{l,2}) \leq 10J_l^2(K_{l,2}^{(1)})\}.$$

We will show that $K_{l,2}^{(r)} \in \mathcal{G}_l^0$ and $K_{l,2}^{(r)} + r_2 D \in \mathcal{G}_l^1$ for $r \in [R_2]$, where r_2 is the smoothing radius and D is a random $p \times m$ matrix generated on a unit sphere. We start by proving some properties of J_l^2 over these sets.

Lemma 3 *The cost $J_l^2(K_{l,2})$ satisfies the following properties.*

1. *The cost $J_l^2(K_{l,2})$ can be written down as*

$$J_l^2(K_{l,2}) = (K_{l,2} - \bar{K}_{l,2}^*)^\top \boldsymbol{A}_l (K_{l,2} - \bar{K}_{l,2}^*) - \frac{1}{4}\boldsymbol{c}_l^\top \boldsymbol{A}_l^{-1}\boldsymbol{c}_l + \boldsymbol{d}_l$$

   where constants $\boldsymbol{A}_l$, $\boldsymbol{c}_l$ and $\boldsymbol{d}_l$ are given in the proof of the Lemma.
2. *The cost $J_l^2(K_{l,2})$ is continuously differentiable with respect to $K_{l,2}$. In addition, any non-empty sublevel set $\mathcal{G}_l(c) := \{K_{l,2} \mid J_l^2(K_{l,2}) \leq c\}$ for $c > 0$ is compact.*
3. *The cost $J_l^2(K_{l,2})$ is smooth and strongly convex, with coefficients φ_2^l and v^l.*
4. *Given that $K_{l,2} \in \mathcal{G}_l^0$, there exists a $\rho_2^l > 0$ and $\lambda_2^l > 0$ s.t. for any $K'_{l,2}$ where $\|K'_{l,2} - K_{l,2}\| \leq \rho_2^l$, we have $K'_{l,2} \in \mathcal{G}_l^1$ and $|J_l^2(K'_{l,2}) - J_l^2(K_{l,2})| \leq \lambda_2^l \|K'_{l,2} - K_{l,2}\|$.*

Proof The proof of this Lemma and the associated constants are provided in Sect. 5. □

We denote the exact gradient of J_l^2 with respect to $K_{l,2}$ by ∇J_l^2, the smoothed (with radius r_2) gradient by $\nabla_{r_2} J_l^2$ and the stochastic gradient by $\tilde{\nabla} J_l^2$. Now we prove the counterpart of Lemma 6 in Malik et al. [22].

Lemma 4 *For $K_{l,2} \in \mathcal{G}_l^0$ the gradients ∇J_l^2, $\nabla_{r_2} J_l^2$ and $\tilde{\nabla} J_l^2$ satisfy*

1. $\mathbb{E}[\tilde{\nabla} J_l^2(K_{l,2})] = \nabla_{r_2} J_l^2(K_{l,2})$
2. $\|\nabla_{r_2} J_l^2(K_{l,2}) - \nabla J_l^2(K_{l,2})\|_2 \leq \varphi_2^l r_2$

Proof Proof of the first part follows from the proof of Lemma 6 in Malik et al. [22]. For the second part for any $K_{l,2} \in \mathcal{G}_l^0$ and $\hat{K}_{l,2}$ sampled uniformly from a unit sphere,

$$\begin{aligned}
\|\nabla_{r_2} J_l^2(K_{l,2}) - \nabla J_l^2(K_{l,2})\|_2 &= \|\nabla \mathbb{E}[J_l^2(K_{l,2} + r_2 \hat{K}_{l,2})] - \nabla J_l^2(K_{l,2})\|_2 \\
&= \|\mathbb{E}[\nabla J_l^2(K_{l,2} + r_2 \hat{K}_{l,2}) - \nabla J_l^2(K_{l,2})]\|_2 \\
&\leq \mathbb{E}[\|\nabla J_l^2(K_{l,2} + r_2 \hat{K}_{l,2}) - \nabla J_l^2(K_{l,2})\|_2] \\
&\leq \varphi_2^l r_2
\end{aligned}$$

The second to last step follows from Jensen's inequality, and the last step is due to the fact that $r_2 < \rho_2^l$ and J_l^2 is smooth with parameter φ_2^l. □

The first part of the Lemma proves that the stochastic gradient is an unbiased estimate of the smoothed gradient and the second part bounds the difference between the exact gradient and the smoothed gradient. We also present a Lemma from Malik et al. [22] which bounds the estimation error between the smoothed gradient and the stochastic gradient with high probability. This Lemma also presents a method to decrease the variance of the stochastic gradient by increasing the minibatch-size k_2.

Lemma 5 *([22]) For any $r_2 \in (0, \rho_2^l)$, the k_2-sample minibatch gradient estimate satisfies the bound*

$$\|\tilde{\nabla} J_l^2(K_{l,2}^{(r)}) - \nabla_{r_2} J_l^2(K_{l,2}^{(r)})\|_2 \leq \frac{1}{\sqrt{k_2}} \frac{m(L+1)}{r_2} \left(J_l^2(K_{l,2}^{(r)}) + \frac{\lambda_2^l}{\rho_2^l} \right) \sqrt{\log\left(\frac{2m(L+1)}{\delta_2} \right)}$$

with probability at least $1 - \delta_2$.

Now we need to ensure that the stepsize is less than ρ_2^l to ensure Lipschitzness. Towards that end, let us first define the optimality gap by $\Delta_r := J_l^2(K_{l,2}^{(r)}) - J_l^2(\bar{K}_{l,2}^*)$ and assume that $K_{l,2}^{(r)} \in \mathcal{G}_l^0$. If we use a minibatch size of $k = 1024 \frac{m^2(L+1)^2}{r^2} \left(J_l^2(K_{l,2}^{(r)}) + \frac{\lambda_2^l}{\rho_2^l} \right)^2 \log\left(\frac{2m(L+1)}{\delta_2} \right) \max\left(\frac{1}{\nu^l \epsilon_2}, (\frac{\lambda_2^l}{\nu^l \epsilon_2})^2 \right)$ then using Lemma 5 we get,

$$\|\tilde{\nabla} J_l^2(K_{l,2}^{(r)}) - \nabla_{r_2} J_l^2(K_{l,2}^{(r)})\|_2 \leq \frac{\sqrt{\nu^l \epsilon_2}}{32}$$

with probability at least $1 - \delta_2$. Conditioned on this event

$$\begin{aligned}
&\|\eta_2 \tilde{\nabla} J_l^2(K_{l,2}^{(r)})\|_2 \\
&= \eta_2 \|\tilde{\nabla} J_l^2(K_{l,2}^{(r)}) - \nabla_{r_2} J_l^2(K_{l,2}^{(r)}) + \nabla_{r_2} J_l^2(K_{l,2}^{(r)}) - \nabla J_l^2(K_{l,2}^{(r)}) + \nabla J_l^2(K_{l,2}^{(r)})\|_2 \\
&\leq \eta_2 \|\tilde{\nabla} J_l^2(K_{l,2}^{(r)}) - \nabla_{r_2} J_l^2(K_{l,2}^{(r)})\|_2 + \eta_2 \|\nabla_{r_2} J_l^2(K_{l,2}^{(r)}) - \nabla J_l^2(K_{l,2}^{(r)})\|_2 + \eta_2 \|\nabla J_l^2(K_{l,2}^{(r)})\|_2 \\
&\leq \eta_2 \left(\frac{\sqrt{\nu^l \epsilon_2}}{32} + \varphi_2^l r_2 + \lambda_2^l \right)
\end{aligned}$$

where the last inequality is obtained by using Lemmas 4 and 5. Since $\epsilon_2, r < 1$

$$\|\eta_2 \tilde{\nabla} J_l^2(K_{l,2}^{(r)})\|_2 \leq \eta_2 \left(\frac{\sqrt{\nu^l}}{32} + \varphi_2^l + \lambda_2^l \right);$$

hence, by choosing $\eta_2 \leq \rho_2^l \left(\frac{\sqrt{\nu^l}}{32} + \varphi_2^l + \lambda_2^l \right)^{-1}$ we ensure,

$$\|\eta_2 \tilde{\nabla} J_l^2(K_{l,2}^{(r)})\|_2 \leq \rho_2^l \tag{63}$$

with probability at least $1 - \delta_2$. Thus, the size of the step $\|\eta_2 \tilde{\nabla} J_l^2(K_{l,2}^{(r)})\|_2$ has been shown to be bounded by ρ_2^l; hence, the Lipschitzness properties of J_l^2 are satisfied with the corresponding coefficients λ_2^l. Notice that using the method shown above $\|\eta_2 \nabla J_l^2(K_{l,2}^{(r)})\|_2$ can also be shown to be bounded by ρ_2^l.

Next we will show that with high probability $K_{l,2}^{(r)} \in \mathcal{G}_l^0$ for any $r \in [R_2]$. Let us trivially assume that $\epsilon_2/2 < \Delta_0$. Now we prove that if for $r \in [R_2]$, $\Delta_r > \epsilon_2/2$ then,

$$J_l^2(K_{l,2}^{(r+1)}) \leq J_l^2(K_{l,2}^{(r)}). \tag{64}$$

Recall that

$$K_{l,2}^{(r+1)} = K^{(r)} - \eta_2 \tilde{\nabla} J_l^2(K_{l,2}^{(r)});$$

similarly we define $\bar{K}_{l,2}^{(r+1)}$ as one step in the direction of the exact gradient:

$$\bar{K}_{l,2}^{(r+1)} = K^{(r)} - \eta_2 \nabla J_l^2(K_{l,2}^{(r)})$$

Due to the smoothness property of J_l^2,

$$\begin{aligned} J_l^2(\bar{K}_{l,2}^{(r+1)}) - J_l^2(K_{l,2}^{(r)}) \\ &\leq \eta_2 \langle \nabla J_l^2(K_{l,2}^{(r)}), \nabla J_l^2(K_{l,2}^{(r)}) \rangle + \frac{\varphi_2^l}{2} \eta_2^2 \|\nabla J_l^2(K_{l,2}^{(r)})\|_2^2 \\ &= \big(\frac{\varphi_2^l}{2} \eta_2^2 - \eta_2\big) \|\nabla J_l^2(K_{l,2}^{(r)})\|_2^2 \end{aligned}$$

Since $\eta_2 \leq 1/\varphi_2^l$,

$$\begin{aligned} J_l^2(\bar{K}_{l,2}^{(r+1)}) - J_l^2(K_{l,2}^{(r)}) &\leq -\frac{\eta_2}{2} \|\nabla J_l^2(K_{l,2}^{(r)})\|_2^2 \\ &\leq -\eta_2 \nu^l \Delta_r < -\eta_2 \nu^l \epsilon_2/2 < 0 \end{aligned} \tag{65}$$

The following Lemma upper bounds the cost gap $|J_l^2(\bar{K}_{l,2}^{(r+1)}) - J_l^2(K_{l,2}^{(r+1)})|$.

Lemma 6 *It holds with probability at least* $1 - \delta_2$, *that,*

$$|J_l^2(\bar{K}_{l,2}^{(r+1)}) - J_l^2(K_{l,2}^{(r+1)})| \leq \eta_2 \nu^l \epsilon_2/16$$

Proof Due to the cost being Lipschitz we have,

$$\begin{aligned} &|J_l^2(\bar{K}_{l,2}^{(r+1)}) - J_l^2(K_{l,2}^{(r+1)})| \\ &\leq \lambda_2^l \|\bar{K}_{l,2}^{(r+1)} - K_{l,2}^{(r+1)}\|_2 \\ &= \eta_2 \lambda_2^l \|\tilde{\nabla} J_l^2(K_{l,2}^{(r)}) - \nabla J_l^2(K_{l,2}^{(r)})\|_2 \\ &\leq \eta_2 \lambda_2^l \|\tilde{\nabla} J_l^2(K_{l,2}^{(r)}) - \nabla_{r_2} J_l^2(K_{l,2}^{(r)})\|_2 + \eta_2 \lambda_2^l \|\nabla_{r_2} J_l^2(K_{l,2}^{(r)}) - \nabla J_l^2(K_{l,2}^{(r)})\|_2 \\ &\leq \eta_2 \lambda_2^l \Big(\frac{\nu^l \epsilon_2}{32 \lambda_2^l} + \varphi_2^l r \Big) \end{aligned}$$

The last step is due to the fact that

$$k_2 = 1024 \frac{m^2 (L+1)^2}{r^2} \big(J_l^2(K_{l,2}^{(r)}) + \frac{\lambda_2^l}{\rho_2^l} \big)^2 \log \big(\frac{2m(L+1)}{\delta_2} \big) \max \big(\frac{1}{\nu^l \epsilon_2}, (\frac{\lambda_2^l}{\nu^l \epsilon_2})^2 \big)$$

which can be used along with Lemma 5 to arrive at the inequality $\|\tilde{\nabla} J_l^2(K_{l,2}^{(r)}) - \nabla_{r_2} J_l^2(K_{l,2}^{(r)})\|_2 \leq \nu^l \epsilon_2 / 32 \lambda_2^l$ with probability at least $1 - \delta_2$. Furthermore by having $r \leq \frac{\nu^l \epsilon_2}{32 \varphi_2^l \lambda_2^l}$ we arrive at

$$|J_l^2(\bar{K}_{l,2}^{(r+1)}) - J_l^2(K_{l,2}^{(r+1)})| \leq \eta_2 \nu^l \epsilon_2/16 \tag{66}$$

with probability at least $1 - \delta_2$. □

Combining Eq. (65) and Lemma 6 we get,

$$\begin{aligned} J_l^2(K_{l,2}^{(r+1)}) - J_l^2(K_{l,2}^{(r)}) &= J_l^2(K_{l,2}^{(r+1)}) - J_l^2(\bar{K}_{l,2}^{(r+1)}) + J_l^2(\bar{K}_{l,2}^{(r+1)}) - J_l^2(K_{l,2}^{(r)}) \\ &< 7\eta_2 \nu^l \epsilon_2/16 < 0 \end{aligned}$$

Hence, if at any iteration r, $\Delta_r > \epsilon_2/2$, then $\Delta_{r+1} < \Delta_r$ with probability at least $1-\delta$. Now we prove that if $\Delta_r \leq \epsilon_2/2$ then $K_{l,2}^{(r+1)} \in \mathcal{G}_l^0$. Using the expression for J_l^2 as in Lemma 3,

$$\begin{aligned} J_l^2(K_{l,2}^{(r+1)}) &= (K_{l,2}^{(r+1)} - \bar{K}_{l,2}^*)^\top \mathbf{A}_l (K_{l,2}^{(r+1)} - \bar{K}_{l,2}^*) - \frac{1}{4}\mathbf{c}_l^\top \mathbf{A}_l^{-1}\mathbf{c}_l + \mathbf{d}_l \\ &= (K_{l,2}^{(r+1)} - K_{l,2}^{(r)} + K_{l,2}^{(r)} - \bar{K}_{l,2}^*)^\top \mathbf{A}_l (K_{l,2}^{(r+1)} - K_{l,2}^{(r)} + K_{l,2}^{(r)} - \bar{K}_{l,2}^*) \\ &\quad - \frac{1}{4}\mathbf{c}_l^\top \mathbf{A}_l^{-1}\mathbf{c}_l + \mathbf{d}_l \\ &\leq 2\|K_{l,2}^{(r+1)} - K_{l,2}^{(r)}\|_{\mathbf{A}_l}^2 + 2\|K_{l,2}^{(r)} - \bar{K}_{l,2}^*\|_{\mathbf{A}_l}^2 - \frac{1}{2}\mathbf{c}_l^\top \mathbf{A}_l^{-1}\mathbf{c}_l + 2\mathbf{d}_l \\ &\leq 2\|K_{l,2}^{(r+1)} - K_{l,2}^{(r)}\|_{\mathbf{A}_l}^2 + 2\Delta_r - \frac{1}{2}\mathbf{c}_l^\top \mathbf{A}_l^{-1}\mathbf{c}_l + 2\mathbf{d}_l \\ &\leq 2\|K_{l,2}^{(r+1)} - K_{l,2}^{(r)}\|_{\mathbf{A}_l}^2 + \epsilon_2 + 2J_l^2(\bar{K}_{l,2}^*) \\ &\leq 2\|\mathbf{A}_l\|_2(\rho_2^l)^2 + \epsilon_2 + 2J_l^2(\bar{K}_{l,2}^*) \\ &= 2J_l^2(K_{l,2}^{(0)}) + \epsilon_2 + 2J_l^2(\bar{K}_{l,2}^*) \\ &\leq 4J_l^2(K_{l,2}^{(0)}) \end{aligned}$$

where the above quantities are defined in (97). Hence, $K_{l,2}^{(r+1)} \in \mathcal{G}_l^0$ with probability $1-\delta_2$. The second inequality is due to the fact that $\Delta_r = J_l^2(K_{l,2}^{(r)}) - J_l^2(\bar{K}_{l,2}^*) = \|K_{l,2}^{(r)} - \bar{K}_{l,2}^*\|_{\mathbf{A}_l}^2$. The third inequality follows from $J_l^2(\bar{K}_{l,2}^*) = -\frac{1}{4}\mathbf{c}_l^\top \mathbf{A}_l^{-1}\mathbf{c}_l + \mathbf{d}_l$ and the fact that $\Delta_r \leq \epsilon_2/2$. The second last inequality follows from the definition of ρ_2^l (99), and the last one follows from the trivial assumption that $\Delta_0 = J_l^2(K_{l,2}^{(r)}) - J_l^2(\bar{K}_{l,2}^*) \geq \epsilon_2/2$.

Now we will show that $J_l^2(K_{l,2}^{(R_2)}) - J_l^2(\bar{K}_{l,2}^*) \leq \epsilon_2/2$ with high probability.

$$\begin{aligned} \Delta_{r+1} - \Delta_r &= J_l^2(K_{l,2}^{(r+1)}) - J_l^2(K_{l,2}^{(r)}) \\ &= J_l^2(K_{l,2}^{(r+1)}) - J_l^2(\bar{K}_{l,2}^{(r+1)}) + J_l^2(\bar{K}_{l,2}^{(r+1)}) - J_l^2(K_{l,2}^{(r)}) \\ &\leq -\eta_2 \nu^l \Delta_r + \nu^l \eta_2 \epsilon_2/16 \end{aligned}$$

with probability at least $1-\delta_2$. The last inequality is due to Eq. (65) and Lemma 6. Hence, we get,

$$\Delta_{r+1} \leq (1-\eta_2\nu^l)\Delta_r + \nu^l \eta_2 \epsilon_2/16$$

with probability at least $1-\delta_2$. Using a union bound type argument and strong recursion, we get

$$\begin{aligned} \Delta_{R_2} &\leq (1-\eta_2\nu^l)^{R_2}\Delta_0 + \sum_{i=0}^{\infty}(1-\nu^l\eta_2)^i \nu^l \eta_2 \frac{\epsilon_2}{16} \\ &= (1-\eta_2\nu^l)^{R_2}\Delta_0 + \frac{\epsilon_2}{16} \end{aligned}$$

with probability at least $1-\delta_2 R_2$. Since $R_2 = \frac{1}{\eta_2 \nu^l}\log(\frac{4\Delta_0}{\epsilon_2})$, $\Delta_{R_2} \leq \frac{\epsilon_2}{2}$ with probability at least $1-\delta_2 R_2$. Furthermore since the cost J_l^2 is strongly convex,

$$\|K_{l,2}^{(R_2)} - \bar{K}_{l,2}^*\|_2 \leq \sqrt{\frac{\epsilon_2}{\nu^l}} \tag{67}$$

with probability at least $1-\delta_2 R_2$. This concludes the proof. □

Proof of Theorem 3

Proof This proof provides finite sample bounds on the estimation error of the MFE computed by the RL algorithm. Due to the stochastic nature of the RL algorithm, the learned policy of each generic agent has some error which causes an asymmetry in the joint learned policy. This results in an error in the mean-field trajectory computed by the centralized simulator. However, since the errors in the learned policy are restricted to be within carefully crafted bounds ($\mathcal{O}(\epsilon_1)$ and $\mathcal{O}(\epsilon_2)$ for the linear and offset terms, respectively), the accumulated error in the mean-field trajectory is shown to be bounded. Using the (corrective) contraction property of the mean-field update operator (by the assumption in Theorem 3), we prove convergence of the RL algorithm to an ϵ neighborhood of the MFE taking into account the bounded errors introduced by the stochastic nature of the RL algorithm.

The proof is organized in two parts. The first part deals with providing finite sample bounds for linear terms in the MFE and the second part deals with providing finite sample bounds for the affine terms in the MFE.

Part I: We will start by proving the bound on $\|F^{(S_1)} - F^*\|_2$ and $\|K_{l,1}^{(S_1)} - K_{l,1}^*\|_2$. From (22) we know that for $s \in [S_1]$, under controllers $(K_{l,1}^{(s)})_{l\in[L]} = \left([K_{l,1}^{(1,s)}, K_{l,1}^{(2,s)}]\right)_{l\in[L]}$, the mean-field trajectory $\bar{Z}^{(s)}$ follows stochastic linear dynamics

$$\bar{Z}_{t+1}^{(s)} = F^{(s)}\bar{Z}_t^{(s)} + \omega_t,\ F^{(s)} = A - B(K_1^{(1,s)} + K_1^{(2,s)}) \tag{68}$$

where A and B are defined in (35) and

$$K_1^{(1,s)} = \text{diag}(K_{1,1}^{(1,s)}, \ldots, K_{L,1}^{(1,s)}),\quad K_1^{(2,s)} = \begin{bmatrix} K_{1,1}^{(2,s)} \\ \vdots \\ K_{L,1}^{(2,s)} \end{bmatrix},\quad \omega_t = \begin{bmatrix} \omega_t^1 \\ \vdots \\ \omega_t^L \end{bmatrix}$$

Let us similarly define

$$\begin{aligned} \bar{F}^{(s)} &:= A - B(\bar{K}_1^{(1,s)} + \bar{K}_1^{(2,s)}), \\ \bar{K}_1^{(1,s)} &= \text{diag}(\bar{K}_{1,1}^{(1,s)}, \ldots, \bar{K}_{L,1}^{(1,s)}),\quad \bar{K}_1^{(2,s)} = \begin{bmatrix} \bar{K}_{1,1}^{(2,s)} \\ \vdots \\ \bar{K}_{L,1}^{(2,s)} \end{bmatrix} \end{aligned} \tag{69}$$

where $\bar{K}_{l,1}^{(s+1)} = \text{argmin}_{K_{l,1}} J_l^1(K_{l,1}, \bar{Z}^{(s)})$. Essentially $\bar{F}^{(s)}$ represents the mean-field trajectory dynamics consistent with the set of controllers $(K_{l,1}^{(s)})_{l\in[L]}$. The following Lemma characterizes $\bar{K}_{l,1}^{(s+1)}$ and $\bar{F}^{(s)}$.

Lemma 7 *The optimal controller for agent l at iteration $s \in S_1$, $\bar{K}_{l,1}^{(s)}$ for the stochastic control problem, with dynamics*

$$X_{t+1}^l = \bar{A}^{l,(s)} X_t^l + \bar{B}^l U_t^l + \bar{W}_t^l, \quad \bar{A}^{l,(s)} = \begin{bmatrix} A^l & 0 \\ 0 & F^{(s)} \end{bmatrix}, \bar{B}^l = \begin{bmatrix} B^l \\ 0 \end{bmatrix}, \bar{W}_t^l = \begin{bmatrix} W_t^l \\ \omega_t \end{bmatrix},$$

and cost

$$J_l(\phi^l, \bar{Z}^{(s)}) := \sum_{t=0}^{\infty} [\| X_t^l \|_{\bar{Q}_l}^2 + \| U_t^l \|_{C_U^l}^2],$$

is given by (106) and mean-field trajectory consistent with $(\bar{K}_{l,1}^{(s)})_{l\in[L]}$ has dynamics matrix $\bar{F}^{(s+1)}$ where $\bar{F}^{(s+1)} = \mathbb{T}(F^{(s)})$ and $\mathbb{T}$ is defined as

$$\mathbb{T}(M) = H^\top + E^{-1} B R^{-1} B^\top \sum_{i=0}^{\infty} H^i C_Z M^{i+1}. \tag{70}$$

This operator is also called the mean-field dynamics update operator.

Proof The proof of this Lemma is provided in Sect. 6. □

The following Lemma introduces some properties of the mean-field dynamics update operator $\mathbb{T}$.

Lemma 8 *Assume that*

$$T_1 := \left\| E^{-1} B R^{-1} B^T (I - H^k)^{-1} C_Z \right\|_2 < 1,$$

$$T_2 := \left\| E^{-1} B R^{-1} B^T \sum_{k=0}^{\infty} H^k C_Z (I - (F^*)^k)(I - F^*)^{-1} \right\|_2 < 1.$$

Then, the operator $\mathbb{T}$ is contractive with coefficient T, and F^* is its fixed point.

This Lemma can be proved following the proof of Proposition 1 in Zaman et al. [29]. Having characterized $\bar{K}_{l,1}^{(s)}$ we now prove the bound on $\| F^{(S_1)} - F^* \|_2$. For $s \in [S_1 - 1]$,

$$\begin{aligned} \| F^{(s+1)} - F^* \|_F &\leq \| F^{(s+1)} - \bar{F}^{(s+1)} \|_F + \| \bar{F}^{(s+1)} - F^* \|_F \\ &\leq \| B \|_F (\| K_1^{(1,s+1)} - \bar{K}_1^{(1,s+1)} \|_F + \| K_1^{(2,s+1)} - \bar{K}_1^{(2,s+1)} \|_F) + T_1 \| F^{(s)} - F^* \|_F \\ &\leq \| B \|_F \sum_{l\in[L]} \sigma_{\min}^{-1}(\bar{\Sigma}^l) \sigma_{\min}^{-1}(C_U^l) \epsilon_1 + T_1 \| F^{(s)} - F^* \|_F \end{aligned} \tag{71}$$

with probability at least $1 - \delta_1 R_1$. The second inequality is due to the definitions of $F^{(s+1)}$ and $\bar{F}^{(s+1)}$ ((68)-(69), respectively), the fact that $\bar{F}^{(s+1)} = \mathbb{T}(F^{(s)})$ and the contractive property of $\mathbb{T}$. The third inequality is obtained by using Lemma 1. Using a union bound type argument, we get

$$\| F^{(S_1)} - F^* \|_F \leq T_1^{S_1} \| F^{(1)} - F^* \|_F + \sum_{j=0}^{S_1 - 1} T_1^j \| B \|_F \frac{\epsilon_1}{2} \sum_{l\in[L]} \sigma_{\min}^{-1}(\bar{\Sigma}^l) \sigma_{\min}^{-1}(C_U^l)$$

with probability at least $1 - \delta_1 S_1 R_1$. Since,

$$\epsilon_1 \leq \frac{(1 - T_1)\epsilon}{\| B \|_F \sum_{l\in[L]} \sigma_{\min}^{-1}(\bar{\Sigma}^l) \sigma_{\min}^{-1}(C_U^l)}, l \in [L] \text{ and } \delta_1 = \frac{\delta}{S_1 R_1} \tag{72}$$

we arrive at

$$\|F^{(S_1)} - F^*\|_F \leq T_1^{S_1}\|F^{(1)} - F^*\|_F + \sum_{j=0}^{S_1-1} T_1^j(1-T)\frac{\epsilon}{2}$$
$$\leq T_1^{S_1}\|F^{(1)} - F^*\|_F + \frac{\epsilon}{2}$$

with probability at least $1-\delta$. Since $S_1 = \frac{1}{1-T_1}\log(\frac{2\|F^{(1)}-F^*\|_F}{\epsilon})$,

$$\|F^{(S_1)} - F^*\|_F \leq \epsilon \tag{73}$$

with probability at least $1-\delta$. Now we prove that $F^{(s)}$ are stable for $s \in [S_1]$. Using reasoning similar to (71), we arrive at

$$\|F^{(s+1)}\|_F \leq \|B\|_F \sum_{l\in[L]} \sigma_{\min}^{-1}(\bar{\Sigma}^l)\sigma_{\min}^{-1}(C_U^l)\epsilon_1 + T\|F^{(s)}\|_F \tag{74}$$

We know that $\epsilon_1 \leq \frac{(1-T_1)\epsilon}{\|B\|_F \sum_{l\in[L]} \sigma_{\min}^{-1}(\bar{\Sigma}^l)\sigma_{\min}^{-1}(C_U^l)}$, $l \in [L]$ and $\epsilon < 1$. Moreover, if we assume $\|F^{(s)}\|_F < 1$ then $\|F^{(s+1)}\|_F < 1$. Now we know that $\|F^{(1)}\|_F = 0$ because $\bar{Z}^{(1)} = 0 < 1$, and using recursion we can show that $\|F^{(s)}\|_F < 1$, and hence, $F^{(s)}$ is stable $s \in [S_1]$.

Now we move on to upper bounding $\|K_{l,1}^{(S_1+1)} - K_{l,1}^*\|_F$. First we consider for $s \in [S_1-1]$,

$$\|K_{l,1}^{(s+1)} - K_{l,1}^*\|_F \leq \|K_{l,1}^{(s+1)} - \bar{K}_{l,1}^{(s+1)}\|_F + \|\bar{K}_{l,1}^{(s+1)} - K_{l,1}^*\|_F$$

From Lemma 1 we know that

$$\|K_{l,1}^{(s+1)} - K_{l,1}^*\|_F \leq \sigma_{\min}^{-1}(\bar{\Sigma}^l)\sigma_{\min}^{-1}(C_U^l)\frac{\epsilon_1}{2} + \|\bar{K}_{l,1}^{(s+1)} - K_{l,1}^*\|_F$$

Recalling the definitions of $\bar{K}_{l,1}^{(s+1)}$ and $K_{l,1}^*$ from Lemma 7,

$$\bar{K}_{l,1}^{(s+1)} = \begin{bmatrix} G^l A^l & (I - G^l B^l)(C_U^l)^{-1} {B^l}^\top \sum_{i=0}^{\infty}(H^l)^i C_Z^l (F^{(s)})^{i+1} \end{bmatrix}$$
$$K_{l,1}^* = \begin{bmatrix} G^l A^l & (I - G^l B^l)(C_U^l)^{-1} {B^l}^\top \sum_{i=0}^{\infty}(H^l)^i C_Z^l (F^*)^{i+1}. \end{bmatrix}$$

Using these expressions $\|K_{l,1}^{(s+1)} - K_{l,1}^*\|_F$ can be upper bounded by

$$\|K_{l,1}^{(s+1)} - K_{l,1}^*\|_F \leq \sigma_{\min}^{-1}(\bar{\Sigma}^l)\sigma_{\min}^{-1}(C_U^l)\frac{\epsilon_1}{2} + D_l^1\|F^{(s)} - F^*\|_F$$

where

$$D_l^1 = \|(I - G^l B^l)(C_U^l)^{-1}{B^l}^\top\|_F \|C_Z^l\|_F/(1 - \|H^l\|_F)^2.$$

Using the value of ϵ_1 from (72) we get

$$\|K_{l,1}^{(S_1+1)} - K_{l,1}^*\|_F \leq D_l^2\epsilon \tag{75}$$

with probability at least $1-\delta$, where

$$D_l^2 = \frac{(1-T_1)\sigma_{\min}^{-1}(\bar{\Sigma}^l)\sigma_{\min}^{-1}(C_U^l)}{2\|B\|_F \sum_{k\in[L]} \sigma_{\min}^{-1}(\bar{\Sigma}^j)\sigma_{\min}^{-1}(R_j)} \tag{76}$$

Part II: Now we focus on the bounds for $\|C^{(S_2)} - C^*\|_2$ and $\|K_{l,2}^{(S_2)} - K_{l,2}^*\|_F$. We know that in the second part of algorithm, $s \in [S_2]$, $\bar{Z}^{(s)}$ follows stochastic affine dynamics,

$$\bar{Z}_{t+1}^{(s)} = F^{(S_1)}\bar{Z}_t^{(s)} + C^{(s)} + \omega_t,$$

where

$$F^{(S_1)} = A - B(K_1^{(1,S_1)} + K_1^{(2,S_1)}), \quad C^{(s)} = -BK_2^{(s)}, \quad K_2^{(s)} = \begin{bmatrix} K_{1,2}^{(s)} \\ \vdots \\ K_{L,2}^{(s)} \end{bmatrix} \tag{77}$$

Let us define $\bar{K}_{l,2}^{(s+1)} = \text{argmin}_{K_{l,2}} J_l^2((K_{l,1}^{(S_1)}, K_{l,2}), \bar{Z}^{(s)})$. Control offset $\bar{K}_{l,2}^{(s+1)}$ can be characterized using the following Lemma.

Lemma 9 *The optimal control offset for agent l at iteration $s \in S_2$, $\bar{K}_{l,2}^{(s)}$ for the stochastic control problem, with drifted dynamics*

$$X_{t+1}^l = \bar{A}^l X_t^l + \bar{B}^l U_t^l + \bar{C}^{(s)} + \bar{W}_t^l,$$

where

$$\bar{A}^l = \begin{bmatrix} A^l & 0 \\ 0 & F \end{bmatrix}, \bar{B}^l = \begin{bmatrix} B^l \\ 0 \end{bmatrix}, \bar{C}^{(s)} = \begin{bmatrix} 0 \\ C^{(s)} \end{bmatrix} \bar{W}_t^l = \begin{bmatrix} W_t^l \\ \omega_t \end{bmatrix},$$

and cost with constant tracking

$$J_l(\phi^l, \bar{Z}^{(s)}) := \sum_{t=0}^{\infty} \left[\left\| X_t^l - \bar{\beta}^l \right\|_{\bar{Q}_l}^2 + \left\| U_t^l \right\|_{C_U^l}^2 \right],$$

is given as follows:

$$\bar{K}_{l,2}^{(s)} = (I - G^l B^l)(C_U^l)^{-1} B^l \sum_{i=0}^{\infty} (H^l)^i C_Z^l ((I - F^i)(I - F)^{-1} C^{(s)} + \beta^l) \tag{78}$$

where $\beta^l = (\beta^{l1}, \ldots, \beta^{lL}) \in \mathbb{R}^{mL}$ and mean-field trajectory consistent with $(\bar{K}_{l,2}^{(s)})_{l \in [L]}$ has offset $\Lambda(C^{(s)})$ and operator Λ is defined as

$$\Lambda(C^{(s)}) = E^{-1} B R^{-1} B^\top \sum_{i=0}^{\infty} H^i [C_Z (I - F^i)(I - F)^{-1} C^{(s)}] + \text{diag}(C_Z^1, \ldots, C_Z^L)\beta \tag{79}$$

where $\beta = (\beta^1, \ldots, \beta^L) \in \mathbb{R}^{mLL}$. This operator is also called the mean-field offset update operator.

Proof The proof of this Lemma is provided in Sect. 7. □

As the operator Λ is defined for a fixed matrix F, we define two operators for specific matrices. We define $\bar{\Lambda}$ and Λ^* for the dynamics matrices $F^{(S_1)}$ and F^*, respectively.

$$\bar{\Lambda}(C^{(s)}) = E^{-1}BR^{-1}B^\top \sum_{i=0}^{\infty} H^i [C_Z(I - (F^{(S_1)})^i)(I - F^{(S_1)})^{-1}C^{(s)}] - \text{diag}(C_Z^1, \ldots, C_Z^L)\beta \tag{80}$$

$$\Lambda^*(C^{(s)}) = E^{-1}BR^{-1}B^\top \sum_{i=0}^{\infty} H_p^i [C_Z(I - (F^*)^i)(I - F^*)^{-1}C^{(s)}] - \text{diag}(C_Z^1, \ldots, C_Z^L)\beta \tag{81}$$

We require Λ^* to be contractive, and hence, we require the Lipschitz constant T_2 of Λ^* to be less than one,

$$T_2 := \left\| E^{-1}BR^{-1}B^\top \sum_{i=0}^{\infty} H^i C_Z(I - (F^*)^i)(I - F^*)^{-1} \right\|_2 < 1 \tag{82}$$

Since $T_2 < 1$, then Λ^* is contractive. Also C^* is the fixed point of operator Λ^*, that is $C^* = \Lambda^*(C^*)$. Let us analyze the convergence of $C^{(s)}$ to C^*. Toward that end, let us consider the following inequality:

$$\begin{aligned} &\|C^{(s+1)} - C^*\|_2 \\ &\quad \le \|C^{(s+1)} - \bar{\Lambda}(C^{(s)})\|_2 + \|\bar{\Lambda}(C^{(s)}) - \Lambda^*(C^{(s)})\|_2 + \|\Lambda^*(C^{(s)}) - C^*\|_2 \end{aligned} \tag{83}$$

We now bound the three terms in (83) separately. Using (77) and (80), the first term can be bounded as follows:

$$\begin{aligned} \|C^{(s+1)} - \bar{\Lambda}(C^{(s)})\|_2 &\le \|B\|_2 \|\text{diag}(K_{1,2}^{(s+1)} - \bar{K}_{1,2}^{(s+1)}, \ldots, K_{L,2}^{(s+1)} - \bar{K}_{L,2}^{(s+1)})\|_2, \\ &\le \frac{\|B\|_2}{\min_{l\in[L]} \sqrt{\nu^l}} \epsilon \end{aligned} \tag{84}$$

with probability at least $1 - \delta_2 R_2$, where the last inequality is obtained using Lemma 2 and the fact that $\epsilon_2 \le \epsilon^2$. The last term in (83) can be similarly bounded

$$\|\Lambda^*(C^{(s)}) - C^*\|_2 \le T_2 \|C^{(s)} - C^*\|_2 \tag{85}$$

To bound the second term in (83) we must first bound the following quantity

$$\begin{aligned} &\left\|(I - (F^{(s)})^k)(I - F^{(s)})^{-1} - (I - (F^*)^k)(I - F^*)^{-1}\right\|_2 = \left\| \sum_{i=0}^{k-1} (F^{(S_1)})^i - (F^*)^i \right\|_2 \\ &= \left\| \sum_{i=0}^{k-1} \sum_{j=0}^{i-1} (F^{(S_1)})^{i-1-j} (F^{(S_1)} - F^*)(F^*)^j \right\|_2 \le \sum_{i=1}^{k-1} i \bar{F}^{i-1} \|F^{(S_1)} - F^*\|_2 \\ &\le \frac{\|F^{(S_1)} - F^*\|_2}{(1 - \bar{F})^2} \end{aligned} \tag{86}$$

Now let us look at the second term in (83),

$$\begin{aligned}
&\|\bar{\Lambda}(C^{(s)}) - \Lambda^*(C^{(s)})\|_2 \\
&= \left\| E^{-1}BR^{-1}B^\top \sum_{i=0}^{\infty} H^i C_Z [(I - (F^{(s)})^i)(I - F^{(s)})^{-1} - (I - (F^*)^i)(I - F^*)^{-1}]C^{(s)} \right\|_2 \\
&\leq \left\| E^{-1}BR^{-1}B^\top \sum_{i=0}^{\infty} H^i C_Z \right\|_2 \frac{\|F^{(S_1)} - F^*\|_2}{(1-\bar{F})^2} \|C^{(s)}\|_2 \\
&\leq \frac{\|F^{(S_1)} - F^*\|_2}{(1-\bar{F})^2} \|C^{(s)}\|_2 = D^3 \|C^{(s)}\|_2 \epsilon
\end{aligned} \tag{87}$$

where the first inequality is obtained using (86), the second inequality using the assumptions of Theorem 3 and

$$D^3 := (1 - \bar{F})^{-2}. \tag{88}$$

Using (84), (85), (87) and a union bound type argument, we obtain

$$\|C^{(s+1)} - C^*\|_2 \leq D^3 \|C^{(s)}\|_2 \epsilon + \frac{\|B\|_2}{\min_{l\in[L]} \sqrt{\nu^l}} \epsilon + T_2 \|C^{(s)} - C^*\|_2 \tag{89}$$

with probability at least $1 - \delta - \delta_2 R_2$. Due to the $C^{(s)}$ term in the right hand side of (89) we first find an upper bound for $C^{(s)}$. Toward that end we use (89) to obtain:

$$\begin{aligned}
&\|C^{(s+1)} - C^*\|_2 \\
&\leq D^3 \|C^{(s)} - C^*\|_2 \epsilon + T_2 \|C^{(s)} - C^*\|_2 + \left(D^3 \|B\|_2 \|C^*\|_2 + \frac{\|B\|_2}{\min_{l\in[L]} \sqrt{\nu^l}} \right) \epsilon \\
&\leq \frac{1 + T_2}{2} \|C^{(s)} - C^*\|_2 + \left(\frac{1 - T_2}{2} \|C^*\|_2 + \frac{\|B\|_2}{\min_{l\in[L]} \sqrt{\nu^l}} \right).
\end{aligned}$$

The last inequality is due to the fact $\epsilon \leq \min(1, \frac{1-T_2}{2D^3\|B\|_2})$. Now $\|C^{(s)} - C^*\|_2$ can be bounded as follows,

$$\begin{aligned}
&\|C^{(s)} - C^*\|_2 \\
&\leq \left(\frac{1 + T_2}{2} \right)^{s-1} \|C^{(1)} - C^*\|_2 + \sum_{i=0}^{s-2} \left(\frac{1 + T_2}{2} \right)^i \left(\frac{1 - T_2}{2} \|C^*\|_2 + \frac{\|B\|_2}{\min_{l\in[L]} \sqrt{\nu^l}} \right) \\
&\leq \|C^{(1)} - C^*\|_2 + \|C^*\|_2 + \frac{2\|B\|_2}{(1 - T_2) \min_{l\in[L]} \sqrt{\nu^l}},
\end{aligned}$$

and hence,

$$\|C^{(s)}\|_2 \leq 2\|C^*\|_2 + \|C^{(1)} - C^*\|_2 + \frac{2\|B\|_2}{(1 - T_2) \min_{l\in[L]} \sqrt{\nu^l}} =: \bar{C}. \tag{90}$$

Now we can write (89) as

$$\|C^{(s+1)} - C^*\|_2 \leq T_2 \|C^{(s)} - C^*\|_2 + \left(D^3 \|B\|_2 \bar{C} + \frac{\|B\|_2}{\sqrt{\nu^l}} \right) \epsilon \tag{91}$$

with probability at least $1-\delta-\delta_2 R_2$, which using a union bound type argument leads to

$$\begin{aligned}\|C^{(s)}-C^*\|_2 &\le (T_2)^{s-1}\|C^{(1)}-C^*\|_2+\sum_{i=0}^{s-2}(T_2)^i\left(D^3\|B\|_2\bar{C}+\frac{\|B\|_2}{\min_{l\in[L]}\sqrt{\nu^l}}\right)\epsilon\\ &\le (T_2)^{s-1}\|C^{(1)}-C^*\|_2+\frac{1}{1-T_2}\left(D^3\|B\|_2\bar{C}+\frac{\|B\|_2}{\min_{l\in[L]}\sqrt{\nu^l}}\right)\epsilon\end{aligned}$$

with probability at least $1-\delta-s\delta_2 R_2$. Plugging in the values $S_2=\frac{1}{1-T_2}\log(\frac{2\|C^{(1)}-C^*\|_2}{\epsilon})$ and $\delta_2=\frac{\delta}{S_2R_2}$

$$\|C^{(S_2)}-C^*\|_2\le D^4\epsilon$$

with probability at least $1-2\delta$, where

$$D^4=\left(\frac{1}{2}+\frac{1}{1-T_2}\left(D^3\|B\|_2\bar{C}+\frac{\|B\|_2}{\min_{l\in[L]}\sqrt{\nu^l}}\right)\right) \tag{92}$$

Now we bound the quantity $\|K_{l,2}^{(S_2+1)}-K_{l,2}^*\|_2$.

$$\begin{aligned}&\|K_{l,2}^{(S_2+1)}-K_{l,2}^*\|_2\\ &\le \|K_{l,2}^{(S_2+1)}-\tilde{K}_{l,2}^{(S_2+1)}\|_2+\|\tilde{K}_{l,2}^{(S_2+1)}-\lambda_l^*C^{(S_2)}\|_2+\|\lambda_l^*C^{(S_2)}+K_{l,2}^*\|_2\\ &= \|K_{l,2}^{(S_2+1)}-\tilde{K}_{l,2}^{(S_2+1)}\|_2+\|\tilde{\lambda}_l C^{(S_2)}-\lambda_l^*C^{(S_2)}\|_2+\|\lambda_l^*C^{(S_2)}+\lambda_l^*C^*\|_2\\ &\le \sqrt{\frac{1}{\nu^l}}\epsilon+\bar{C}D_l^3\epsilon+\|\lambda_l^*\|_2D^4\epsilon\\ &\le D_l^5\epsilon\end{aligned}$$

with probability at least $1-2\delta$, where

$$D_l^5=\sqrt{\frac{1}{\nu^l}}+\bar{C}D_l^3+\|\lambda_l^*\|_2D^4 \tag{93}$$

Now we prove that global constants ρ_1^l, φ_1^l, λ_1^l, ν^l, φ_2^l, ρ_2^l and λ_2^l for each $l\in[L]$ do exist and characterize them.

Lemma 10 *If* $\epsilon\le\frac{1}{\sqrt{m(L+1)}}\min(1,\min_{l\in[L]}c_{16}^l,\min_{l\in[L]}\frac{1}{D_l^2})$ *where* c_{16}^l *and* D_l^2 *are defined in (121) and (76), respectively, then global constants* μ^l, ρ_1^l, φ_1^l, λ_1^l, ν^l, φ_2^l, ρ_2^l *and* λ_2^l *for each* $l\in[L]$ *are defined in (125) and (138).*

Proof The proof of this Lemma is provided in Sect. 8. □

Hence, we have completed the proof of Theorem 3. □

Proof of Lemma 3

Proof We know from Proposition B2 in Fu et al. [13] that the cost J_l^2 is quadratic in $K_{l,2}$,

$$\begin{aligned}&J_l^2((K_{l,1},K_{l,2}),\bar{Z})\\ &=\begin{pmatrix}\mu_K\\K_{l,2}\end{pmatrix}^\top\begin{pmatrix}\bar{Q}_l+K_{l,1}^\top C_U^l K_{l,1} & -K_{l,1}^\top C_U^l\\ -C_U^l K_{l,1} & C_U^l\end{pmatrix}\begin{pmatrix}\mu_K\\K_{l,2}\end{pmatrix}-2(\bar{\beta}^l)^\top\bar{Q}_l\mu_K\end{aligned} \tag{94}$$

where

$$\mu_K = \left(I - \bar{A}^l + \bar{B}^l K_{l,1}\right)^{-1}\left(\bar{B}^l K_{l,2} + \bar{C}\right) \tag{95}$$

As J_l^2 is quadratic in $K_{l,2}$, it is continuously differentiable with respect to $K_{l,2}$. Moreover, as the Hessian of J_l^2 is positive definite (Proposition 3.3 [13]), the non-empty level sets of J_l^2 are ellipsoids and hence the non-empty sublevel sets are compact.

Now we aim to derive the values of the Lipschitz constant φ_2^l and radius ρ_2^l. First we notice that the cost J_l^2 can be written in the form,

$$J_l^2(K_{l,2}) = K_{l,2}^\top \mathbf{A}_l K_{l,2} + \mathbf{c}_l^\top K_{l,2} + \mathbf{d}_l \tag{96}$$

where

$$\begin{aligned}
\mathbf{A}_l &= \left\| \begin{matrix} (I - \bar{A}^l + \bar{B}^l K_{l,1})^{-1}\bar{B}^l \\ I \end{matrix} \right\|^2_{\begin{pmatrix} \bar{Q}_l + (K_{l,1})^\top C_U^l K_{l,1} & -(K_{l,1})^\top C_U^l \\ -C_U^l K_{l,1} & C_U^l \end{pmatrix}} \\
\mathbf{c}_l &= 2\left((I - \bar{A}^l + \bar{B}^l K_{l,1})^{-1}\bar{C}\right)^\top \left(\bar{Q}_l + (K_{l,1})^\top C_U^l K_{l,1}\right)\left((I - \bar{A}^l + \bar{B}^l K_{l,1})^{-1}\bar{B}^l\right) \\
&\quad - 2\left((I - \bar{A}^l + \bar{B}^l K_{l,1})^{-1}\bar{C}\right)^\top (K_{l,1})^\top C_U^l - 2(\bar{\beta})^l{}^\top \bar{Q}_l (I - \bar{A}^l + \bar{B}^l K_{l,1})^{-1}\bar{B}^l, \\
\mathbf{d}_l &= \left((I - \bar{A}^l + \bar{B}^l K_{l,1})^{-1}\bar{C}\right)^\top \left(\bar{Q}_l + (K_{l,1})^\top C_U^l K_{l,1}\right)\left((I - \bar{A}^l + \bar{B}^l K_{l,1})^{-1}\bar{C}\right) \\
&\quad - 2(\bar{\beta}^l)^\top \bar{Q}_l (I - \bar{A}^l + \bar{B}^l K_{l,1})^{-1}\bar{C}.
\end{aligned} \tag{97}$$

The matrix $\mathbf{A}_l$ is symmetric positive definite. Proposition 3.3 of Fu et al. [13] proves smoothness and strong convexity of J_l^2 with coefficients φ_2^l and ν^l such that

$$\nu^l = \sigma_{\min}(\mathbf{A}_l), \qquad \varphi_2^l = \|\mathbf{A}_l\|_2. \tag{98}$$

Recall that the $K_{l,2}$ which minimizes J_l^2 is denoted by $\bar{K}_{l,2}^*$ and is given by

$$\bar{K}_{l,2}^* = -\frac{1}{2}\mathbf{A}_l^{-1}\mathbf{c}_l$$

which exists since $\mathbf{A}_l > 0$. By completing the square we can write the cost as

$$J_l^2(K_{l,2}) = (K_{l,2} - \bar{K}_{l,2}^*)^\top \mathbf{A}_l (K_{l,2} - \bar{K}_{l,2}^*) - \frac{1}{4}\mathbf{c}_l^\top \mathbf{A}_l^{-1}\mathbf{c}_l + \mathbf{d}_l$$

As in the statement of the Lemma assume $K_{l,2} \in \mathcal{G}_l^0$ and $\|K'_{l,2} - K_{l,2}\|_2 \leq \rho_2^l$ where ρ_2^l satisfies

$$\rho_2^l = \sqrt{\frac{J_l^2(K_i, \bar{Z})}{\|\mathbf{A}_l\|_2}} \tag{99}$$

Then the cost of controller $K'_{l,2}$ is

$$\begin{aligned}
&J_l^2(K'_{l,2}) = (K'_{l,2} - \bar{K}^*_{l,2})^\top \mathbf{A}_l (K'_{l,2} - \bar{K}^*_{l,2}) - \frac{1}{4}\mathbf{c}_l^\top \mathbf{A}_l^{-1}\mathbf{c}_l + \mathbf{d}_l \\
&= (K'_{l,2} - K_{l,2} + K_{l,2} - \bar{K}^*_{l,2})^\top \mathbf{A}_l (K'_{l,2} - K_{l,2} + K_{l,2} - \bar{K}^*_{l,2}) - \frac{1}{4}\mathbf{c}_l^\top \mathbf{A}_l^{-1}\mathbf{c}_l + \mathbf{d}_l \\
&\le 2(K'_{l,2} - K_{l,2})^\top \mathbf{A}_l (K'_{l,2} - K_{l,2}) + 2(K_{l,2} - \bar{K}^*_{l,2})^\top \mathbf{A}_l (K_{l,2} - \bar{K}^*_{l,2}) \\
&\quad - \frac{1}{2}\mathbf{c}_l^\top \mathbf{A}_l^{-1}\mathbf{c}_l + 2\mathbf{d}_l \\
&= 2\|K'_{l,2} - K_{l,2}\|^2_{\mathbf{A}_l} + 2J_l^2(K_{l,2}) \\
&\le 2\|\mathbf{A}_l\|_2(\rho_2^l)^2 + 8J_l^2(K_i, \bar{Z}) \le 10J_l^2(K_i, \bar{Z})
\end{aligned}$$

Hence, $K'_{l,2} \in \mathcal{G}_l^1$. Since J_l^2 is smooth with coefficient φ_2^l, for any $K'_{l,2} \in \mathcal{G}_l^1$ we have

$$\|\nabla J_l^2(K'_{l,2})\|_2^2 \le 2\varphi_2^l(J_l^2(K'_{l,2}) - J_l^2(\bar{K}^*_{l,2})) \le 20\varphi_2^l J_l^2(K_i, \bar{Z})$$

Hence, for $K'_{l,2} \in \mathcal{G}_l^1$, $J_l^2(K'_{l,2})$ is Lipschitz with coefficient,

$$\lambda_2^l = \sqrt{20\varphi_2^l J_l^2(K_i, \bar{Z})} \tag{100}$$

This concludes the proof. □

Proof of Lemma 7

Proof Due to certainty equivalence we instead consider the deterministic LQR problem with dynamics

$$X_{t+1}^l = \bar{A}_l^{l,(s)} X_t^l + \bar{B}^l U_t^l, \text{ where } \bar{A}^{l,(s)} = \begin{bmatrix} A^l & 0 \\ 0 & F^{(s)} \end{bmatrix}, \bar{B}^l = \begin{bmatrix} B^l \\ 0 \end{bmatrix} \tag{101}$$

and cost

$$J_l(\phi^l, \bar{Z}^{(s)}) := \sum_{t=0}^{\infty} \left[\left\| X_t^l \right\|^2_{\bar{Q}_l} + \left\| U_t^l \right\|^2_{C_U^l} \right] \tag{102}$$

Since for $s \in [S_1]$ the control offset $K_{l,2}^{(0)}$ and mean-field drift are 0 and the class of controllers is restricted to linear controllers $\phi^l(X_t^l) = K_{l,1}X_t^l$, then optimal controller $K_{l,1}$ for the stochastic drifted-LQR problem (15)-(16) will also be optimal for the deterministic LQR problem shown above. This deterministic problem can be rewritten as a Linear Quadratic Tracking (LQT) problem with dynamics

$$Z_{t+1}^l = A^l Z_t^l + B^l U_t^i$$

and cost

$$J_l(\phi^l, \bar{Z}^{(s)}) := \sum_{t=0}^{\infty} [\|Z_t^l\|^2_{Q^l} + \|U_t^l\|^2_{C_U^l} + \sum_{k\in[L]} \|Z_t^l - \bar{Z}_t^{(k,s)}\|^2_{C_Z^{lk}}],$$

where $\bar{Z}^{(s,i)}$ is the mean-field trajectory of population l in the joint mean-field trajectory $\bar{Z}^{(s)}$. This problem can be solved by using the maximum principle approach as shown in Proof of

Proposition 1. From Eq. (31) we surmise

$$s_t^l = -\sum_{k\in[L]} C_Z^{lk}\bar{Z}_t^{(k,s)} + H^l s_{t+1}^l \tag{103}$$

where H^l is defined in (53), P^l is the solution to the Riccati Eq. (51), and $\bar{Z}^{(j,s)}$ represents the jth population's mean-field trajectory in the joint mean-field trajectory $\bar{Z}^{(s)}$. Using (101) we write down the closed-loop dynamics of generic agent l.

$$\begin{aligned} Z_{t+1}^l &= {H^l}^\top Z_t^l - (E^l)^{-1}B^l(C_U^l)^{-1}{B^l}^\top s_{t+1}^l, \\ &= {H^l}^\top Z_t^l - (E^l)^{-1}B^l(C_U^l)^{-1}{B^l}^\top \sum_{i=0}^{\infty}(H^l)^i \sum_{k\in[L]} C_Z^{lk}\bar{Z}_{t+i+1}^{(j,s)}, \\ &= (A^l - B^l(C_U^l + {B^l}^\top P^l B^l)^{-1}{B^l}^\top P^l A^l)Z_t^l - (E^l)^{-1}B^l(C_U^l)^{-1}{B^l}^\top \sum_{i=0}^{\infty}(H^l)^i C_Z^l \bar{Z}_{t+i+1}^{(s)} \end{aligned} \tag{104}$$

where $C_Z^l = (C_Z^{l1}, C_Z^{l2}, \ldots)$. Since $\bar{Z}^{(s)}$ is assumed to follow linear dynamics $\bar{Z}_{t+1}^{(s)} = F^{(s)}\bar{Z}_t^{(s)}$, this can be further simplified into,

$$\begin{aligned} Z_{t+1}^l = (A^l - B^l(C_U^l + {B^l}^\top P^l B^l)^{-1}{B^l}^\top P^l A^l)Z_t^l \\ - (E^l)^{-1}B^l(C_U^l)^{-1}{B^l}^\top \sum_{i=0}^{\infty}(H^l)^i C_Z^l (F^{(s)})^{i+1}\bar{Z}_t^{(s)} \end{aligned}$$

This can be rewritten in terms of the controller $\bar{K}_{l,1}^{(s+1)}$,

$$Z_{t+1}^l = A^l Z_t^l - B^l \bar{K}_{l,1}^{(s+1)} \begin{bmatrix} Z_t^l \\ \bar{Z}_t^{(s)} \end{bmatrix} \tag{105}$$

where

$$\bar{K}_{l,1}^{(s+1)} = \begin{bmatrix} G^l A^l & (I - G^l B^l)(C_U^l)^{-1}{B^l}^\top \sum_{i=0}^{\infty}(H^l)^i C_Z^l (F^{(s)})^{i+1} \end{bmatrix} \tag{106}$$

where $G^l = (C_U^l + {B^l}^\top P^l B^l)^{-1}{B^l}^\top P^l$. We know that $\bar{K}_{l,1}^{(s+1)}$ exists since H^l is Hurwitz. Now we simulate the behavior of infinitely many agents in population l under controller $\bar{K}_{l,1}^{(s+1)}$ using (104),

$$\bar{Z}_{t+1}^l = {H^l}^\top \bar{Z}_t^l - (E^l)^{-1}B^l(C_U^l)^{-1}{B^l}^\top \sum_{i=0}^{\infty}(H^l)^i C_Z^l (F^{(s)})^{i+1}\bar{Z}_t \tag{107}$$

Writing down the closed-loop dynamics for the joint mean-field trajectory we get,

$$\bar{Z}_{t+1} = \left(H^\top - E^{-1}BR^{-1}B^\top \sum_{i=0}^{\infty} H^k C_Z (F^{(s)})^{k+1}\right)\bar{Z}_t \tag{108}$$

where,

$$H = \text{diag}(H^1, H^2, \ldots), \quad E = \text{diag}(E^1, E^2, \ldots), \quad C_Z = ({C_Z^1}^\top, {C_Z^2}^\top, \ldots)^\top \tag{109}$$

Now we define a mean-field dynamics update operator $\mathbb{T}$ as follows

$$\mathbb{T}(M) = H^\top + E^{-1}BR^{-1}B^\top \sum_{i=0}^{\infty} H^i C_Z M^{i+1} \tag{110}$$

and $\bar{F}^{(s+1)} = \mathbb{T}(F^{(s)})$. □

Proof of Lemma 9

Proof Using logic similar to proof of Lemma 7, we arrive at the deterministic tracking control problem for agent l where the dynamics of agent l are

$$Z_{t+1}^l = A^l Z_t^l + B^l U_t^i$$

and cost has constant tracking terms,

$$J_l(\phi^l, \bar{Z}^{(s)}) := \sum_{t=0}^{\infty} [\|Z_t^l\|_{Q^l}^2 + \|U_t^l\|_{C_U^l}^2 + \sum_{k\in[L]} \|Z_t^l - \bar{Z}_t^{(k,s)} - \beta^{lk}\|_{C_Z^{lk}}^2].$$

where $\bar{Z}^{(l,s)}$ is the mean-field trajectory of population l in the joint mean-field trajectory $\bar{Z}^{(s)}$. This problem can be solved by using the maximum principle approach as shown in Proof of Proposition 1. From Eq. (31) we surmise,

$$s_t^l = -\sum_{k\in[L]} C_Z^{lk}(\bar{Z}_t^{(k,s)} + \beta^{lk}) + H^l s_{t+1}^l \tag{111}$$

where H^l is defined in (53). As in proof of Lemma 7 we write down the closed-loop dynamics of generic agent l.

$$\begin{aligned}
Z_{t+1}^l &= {H^l}^\top Z_t^l - (E^l)^{-1}B^l(C_U^l)^{-1}{B^l}^\top s_{t+1}^l, \\
&= {H^l}^\top Z_t^l - (E^l)^{-1}B^l(C_U^l)^{-1}{B^l}^\top \sum_{i=0}^{\infty}(H^l)^i \sum_{k\in[L]} C_Z^{lk}(\bar{Z}_{t+i+1}^{(j,s)} + \beta^{lk}), \\
&= {H^l}^\top Z_t^l - (E^l)^{-1}B^l(C_U^l)^{-1}{B^l}^\top \sum_{i=0}^{\infty}(H^l)^i C_Z^l(\bar{Z}_{t+i+1}^{(s)} + \beta^l)
\end{aligned}$$

where $C_Z^l = (C_Z^{l1}, C_Z^{l2}, \ldots)$ and $\beta^l = (\beta^{l1}, \ldots, \beta^{lL}) \in \mathbb{R}^{mL}$. Since $\bar{Z}^{(s)}$ is assumed to follow affine dynamics $\bar{Z}_{t+1}^{(s)} = F\bar{Z}_t^{(s)} + C^{(s)}$, this can be further simplified into,

$$\begin{aligned}
Z_{t+1}^l =& {H^l}^\top Z_t^l \\
&- (E^l)^{-1}B^l(C_U^l)^{-1}{B^l}^\top \sum_{i=0}^{\infty}(H^l)^i C_Z^l(F^{i+1}\bar{Z}_t^{(s)} + (I-F^i)(I-F)^{-1}C^{(s)} + \beta^l)
\end{aligned}$$

This can be rewritten in terms of the controller $\bar{K}_{l,1}^{(s+1)}$ and $\bar{K}_{l,2}^{(s+1)}$,

$$Z_{t+1}^l = A^l Z_t^l - B^l \bar{K}_{l,1}^{(s+1)} \begin{bmatrix} Z_t^l \\ \bar{Z}_t^{(s)} \end{bmatrix} - B^l \bar{K}_{l,2}^{(s+1)} \tag{112}$$

where,

$$\bar{K}_{l,2}^{(s)} = (I - G^l B^l)(C_U^l)^{-1} B^l \sum_{i=0}^{\infty} (H^l)^i C_Z^l ((I - F^i)(I - F)^{-1} C^{(s)} + \beta^l) \tag{113}$$

Simulating the behavior of infinitely many agents as in Lemma 7, we get the mean-field offset update operator Λ defined as

$$\Lambda(C^{(s)}) = E^{-1} B R^{-1} B^\top \sum_{i=0}^{\infty} H^i [C_Z (I - F^i)(I - F)^{-1} C^{(s)}] + \text{diag}(C_Z^1, \ldots, C_Z^l)\beta$$

where $\beta = (\beta^1, \ldots, \beta^L) \in \mathbb{R}^{mLL}$. □

Proof of Lemma 10

Proof We define the global constants μ^l, ρ_1^l, φ_1^l and λ_1^l for Lemma 1. We observe from Section A in Malik et al. [22] that these constants depend on norms of matrices $\|\bar{A}^l\|_2$, which depend on the norm of mean-field trajectory dynamics matrix $\|F^{(s)}\|_2$. Furthermore the constants also depend on the initial cost $J_l((K_{l,1}^{(1)}, K_{l,2}^{(0)}), \bar{Z}^{(s)})$. We start by obtaining a bound for $\|F^{(s)}\|_2$. From (71) we observe

$$\|F^{(s+1)} - F^*\|_F \leq \|B\|_F \sum_{l \in [L]} \sigma_{\min}^{-1}(\bar{\Sigma}^l) \sigma_{\min}^{-1}(C_U^l)\epsilon_1 + T_1 \|F^{(s)} - F^*\|_F.$$

This implies

$$\|F^{(s)} - F^*\|_F \leq \|F^{(1)} - F^*\|_F + \frac{1}{1 - T_1} \|B\|_F \sum_{l \in [L]} \sigma_{\min}^{-1}(\bar{\Sigma}^l) \sigma_{\min}^{-1}(C_U^l)\epsilon_1.$$

Hence,

$$\begin{aligned} \|F^{(s)}\|_2 \leq & \|F^*\|_2 + m(L+1)\|F^{(1)} - F^*\|_F \\ & + \frac{m(L+1)}{1 - T_1} \|B\|_F \sum_{l \in [L]} \sigma_{\min}^{-1}(\bar{\Sigma}^l) \sigma_{\min}^{-1}(C_U^l)\epsilon_1 =: \bar{F} \end{aligned} \tag{114}$$

Now that an upper bound on $\|F^{(s)}\|_2$ has been defined in (114), we compute $\mathbf{J}_l^1$ which is the upper bound on $J_l((K_{l,1}^{(1)}, K_{l,2}^{(0)}), \bar{Z}^{(s)})$ for $s \in [S_1]$. Under controller $(K_{l,1}^{(1)}, K_{l,2}^{(0)})$ the dynamics of generic agent l and the mean-field trajectory dynamics are decoupled.

$$\begin{aligned} Z_{t+1}^l &= (A^l - B^l K_{l,1}^{1,1}) Z_t^l + W_t^l \\ \bar{Z}_{t+1}^{(s)} &= F^{(s)} \bar{Z}_t^{(s)} + \omega_t \end{aligned}$$

The cost function for the generic agent l is

$$J_l((K_{l,1}^{(1)}, K_{l,2}^{(0)}), \bar{Z}^{(s)}) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\left\| \begin{bmatrix} Z_t^l \\ \bar{Z}_t \end{bmatrix} - \bar{\beta}^l \right\|_{\bar{Q}_l}^2 + \|K_{l,1}^{1,1} Z_t^l\|_{C_U^l}^2 \right]$$

$$= \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[\|Z_t^l\|_{Q^l}^2 + \|K_{l,1}^{1,1} Z_t^l\|_{C_U^l}^2 + \sum_{k\in[L]} \|Z_t^l - (\bar{Z}_t^k + \beta^{lk})\|^2]$$

$$\leq \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[2\|Z_t^l\|_{Q^l+\sum_{k\in[L]}}^2 + \|K_{l,1}^{1,1} Z_t^l\|_{C_U^l}^2 + 2\sum_{k\in[L]} \|\bar{Z}_t^k\|_{C_Z^{lk}}^2] + 2\sum_{k\in[L]} \|\beta^{lk}\|_{C_Z^{lk}}^2$$

$$= \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[2\|Z_t^l\|_{Q^l+\sum_{k\in[L]}}^2 + \|K_{l,1}^{1,1} Z_t^l\|_{C_U^l}^2 + 2\|\bar{Z}_t\|_{\bar{C}_l}^2] + 2\sum_{k\in[L]} \|\beta^{lk}\|_{C_Z^{lk}}^2$$

where $\bar{C}_l = \text{diag}(C_Z^{l1}, \ldots, C_Z^{lL})$. Using results in standard LQR analysis [21], this cost is given by

$$J_l((K_{l,1}^{(1)}, K_{l,2}^{(0)}), \bar{Z}^{(s)}) = \text{Tr}(\bar{P}_l \Sigma_w^{(i)}) + 2\,\text{Tr}(\bar{P}_l^{(s)}\sigma) + \sum_{k\in[L]} \|\beta^{lk}\|_{C_Z^{lk}}^2,$$

where the matrices $\bar{P}_l$ and $\bar{P}_l^{(s)}$ are solutions to the Lyapunov equations,

$$\bar{P}_l = 2(Q^l + \sum_{k\in[L]} C_Z^{lk}) + (K_{l,1}^{(1,1)})^\top C_U^l K_{l,1}^{(1,1)} + (A^l - B^l K_{l,1}^{(1,1)})^\top \bar{P}_l (A^l - B^l K_{l,1}^{(1,1)}),$$

$$\bar{P}_l^{(s)} = \bar{C}_l + (F^{(s)})^\top \bar{P}_l^{(s)} F^{(s)}. \tag{115}$$

We upper bound $\text{Tr}(\bar{P}_l^{(s)}\sigma)$ using Lemma 20 in Fazel et al. [12]. Toward that end, we first define matrix $\bar{P}_l^*$ as the solution to the Lyapunov equation

$$\bar{P}_l^* = \bar{C}_l + (F^*)^\top \bar{P}_l^* F^* \tag{116}$$

We also introduce the following operators,

$$\mathcal{T}^{(s)}(X) = \sum_{t=0}^{\infty} ((F^{(s)})^\top)^t X (F^{(s)})^t, \qquad \mathcal{T}^*(X) = \sum_{t=0}^{\infty} ((F^*)^\top)^t X (F^*))^t,$$
$$\mathcal{F}^{(s)}(X) = (F^{(s)})^\top X F^{(s)}, \qquad \mathcal{F}^*(X) = (F^*)^\top X F^* \tag{117}$$

where $\mathcal{T}^{(s)}(\bar{C}_l) = \bar{P}_l^{(s)}$ and $\mathcal{T}^*(\bar{C}_l) = \bar{P}_l^*$. Towards upper bounding $\text{Tr}(\bar{P}_l^{(s)}\sigma)$ we first recognize

$$\text{Tr}(\bar{P}_l^{(s)}\sigma) = \text{Tr}((\bar{P}_l^{(s)} - \bar{P}_l^*)\sigma) + \text{Tr}(\bar{P}_l^*\sigma) \leq \|\bar{P}_l^{(s)} - \bar{P}_l^*\|_2 \|\sigma\|_2 + \text{Tr}(\bar{P}_l^*\sigma) \tag{118}$$

So we need to bound $\|\bar{P}_l^{(s)} - \bar{P}_l^*\|_2$ using Lemma 20 in Fazel et al. [12]. First we obtain a bound on $\|\mathcal{F}^{(s)} - \mathcal{F}^*\|_2$ which is similar to Lemma 19 of Fazel et al. [12], where $\|\cdot\|_2$ is the operator norm $\|\mathcal{F}\|_2 = \sup_X \frac{\|\mathcal{F}(X)\|_2}{\|X\|_2}$. Let us first define $\tilde{F} = F^{(s)} - F^*$; then, for any matrix X,

$$\mathcal{F}^{(s)}(X) - \mathcal{F}^*(X) = (F^*)^\top X \tilde{F} + (\tilde{F})^\top X F^* - (\tilde{F})^\top X \tilde{F}$$

Then, using the definition of operator norm $\|\cdot\|_2$ we get

$$\|\mathcal{F}^{(s)} - \mathcal{F}^*\|_2 \leq 2\|F^{(s)} - F^*\|_2\|F^*\|_2 + \|F^{(s)} - F^*\|_2^2 \tag{119}$$

Now we obtain a bound on $\|\mathcal{T}^*\|$ using techniques similar to Lemma 17 of Fazel et al. [12]. Consider a unit norm vector v and unit spectral norm matrix X.

$$\begin{aligned}
v^\top \mathcal{T}^*(X)v &= \sum_{t=0}^{\infty} v^\top ((F^*)^\top)^t X (F^*)^t v = \sum_{t=0}^{\infty} \mathrm{Tr}((F^*)^t v v^\top ((F^*)^\top)^t X) \\
&= \sum_{t=0}^{\infty} \mathrm{Tr}(\sigma^{1/2}(F^*)^t v v^\top ((F^*)^\top)^t \sigma^{1/2}\sigma^{-1/2} X \sigma^{-1/2}) \\
&\leq \sum_{t=0}^{\infty} \mathrm{Tr}(\sigma^{1/2}(F^*)^t v v^\top ((F^*)^\top)^t \sigma^{1/2})\|\sigma^{-1/2} X \sigma^{-1/2}\|_2 \\
&= \|\sigma^{-1/2} X \sigma^{-1/2}\|_2 (v^\top \mathcal{T}^*(\sigma)v) \leq \frac{\|\mathcal{T}^*(\sigma)\|_2}{\sigma_{\min}(\sigma)} \leq \frac{\mathrm{Tr}(\bar{P}_l^*\sigma)}{\sigma_{\min}(\sigma)\sigma_{\min}(\bar{C}_l)}
\end{aligned} \tag{120}$$

Hence, $\|\mathcal{T}^*\|_2 \leq \frac{\mathrm{Tr}(\bar{P}_l^*\sigma)}{\sigma_{\min}(\sigma)\sigma_{\min}(\bar{C}_l)}$. Using (119)-(120), we get

$$\|\mathcal{T}^*\|_2\|\mathcal{F}^{(s)} - \mathcal{F}^*\|_2 \leq \frac{\mathrm{Tr}(\bar{P}_l^*\sigma)}{\sigma_{\min}(\sigma)\sigma_{\min}(\bar{C}_l)}(2\|F^*\|_2 + \|F^{(s)} - F^*\|_2)\|F^{(s)} - F^*\|_2$$

Since $\epsilon \leq \frac{1}{\sqrt{m(L+1)}} \min_{l\in[L]}(1, c_{16}^l)$, $\|F^{(s)} - F^*\|_2 \leq \min_{l\in[L]}(1, c_{16}^l)$, where

$$c_{16}^l = \frac{\sigma_{\min}(\sigma)\sigma_{\min}(\bar{C}_l)}{2\,\mathrm{Tr}(\bar{P}_l^*\sigma)(2\|F^*\|_2 + 1)}, \tag{121}$$

then

$$\|\mathcal{T}^*\|_2\|\mathcal{F}^{(s)} - \mathcal{F}^*\|_2 \leq 1/2$$

This satisfies the conditions for Lemma 20 in Fazel et al. [12], so we obtain,

$$\|\bar{P}_l^{(s)} - \bar{P}_l^*\|_2 = \|\mathcal{T}^{(s)}(\bar{C}_l) - \mathcal{T}^*(\bar{C}_l)\|_2 \leq c_{15}^l\|F^{(s)} - F^*\|_2 \tag{122}$$

where

$$c_{15}^l = \left(\frac{\mathrm{Tr}(\bar{P}_l^*\sigma)}{\sigma_{\min}(\sigma)\sigma_{\min}(\bar{C}_l)}\right)^2 (2\|F^*\|_2 + 1)\|\bar{C}_l\|_2 \tag{123}$$

Hence, using (118) and (122),

$$\mathrm{Tr}(\bar{P}_l^{(s)}\sigma) \leq c_{15}^l\|\sigma\|_2\|F^{(s)} - F^*\|_2 + \mathrm{Tr}(\bar{P}_l^*\sigma)$$

Now we can bound the cost $J_l((K_{l,1}^{(1)}, K_{l,2}^{(0)}), \bar{Z}^{(s)})$,

$$\begin{aligned}
J_l((K_{l,1}^{(1)}, K_{l,2}^{(0)}), \bar{Z}^{(s)}) \leq \mathrm{Tr}(\bar{P}_l\Sigma_w^{(i)}) + 2c_{15}^l\|\sigma\|_2\|F^{(s)} - F^*\|_2 + 2\,\mathrm{Tr}(\bar{P}_l^*\sigma) \\
+ \sum_{k\in[L]} \|\beta^{lk}\|_{C_Z^{lk}}^2 =: \mathbf{J}_l^1,
\end{aligned} \tag{124}$$

Firstly we can bound μ^l. Using Corollary 5 from [12],

$$\mu^l \leq \frac{\|\Sigma_{\bar{K}_l^{(s)}}\|_2}{\sigma_{\min}^2(\bar{\Sigma}^l)\sigma_{\min}(C_U^l)} \leq \frac{J_l^1(\bar{K}^{(s)}, \bar{Z}^{(s)})}{\sigma_{\min}^2(\bar{\Sigma}^l)\sigma_{\min}(C_U^l)\sigma_{\min}(\bar{Q}_l)} \leq \frac{\mathbf{J}_l^1 \sigma_{\min}^{-1}(\bar{Q}_l)}{\sigma_{\min}^2(\bar{\Sigma}^l)\sigma_{\min}(C_U^l)},$$

where we use Lemma 5.1 from [27] for the second inequality. Now using (114), (124) and Lemma 9 from Malik et al. [22] we define

$$\begin{aligned}
c_0^l &= \frac{\sqrt{\|C_U^l\|_2 + 10\|B\|_2^2\mathbf{J}_l^1} + 10\|B^l\|_2(\|A^l\|_2 + \bar{F})\mathbf{J}_l^1}{\sigma_{\min}(C_U^l)},\\
c_1^l &= \max\left(\frac{10\mathbf{J}_l^1}{\sigma_{\min}(\bar{Q}_l)}\sqrt{\|C_U^l\|_2 + \|B^l\|_2^2(10\mathbf{J}_l^1)^2}, c_0^l\right),\\
c_2^l &= 4\left(\frac{10\mathbf{J}_l^1}{\sigma_{\min}(\bar{Q}_l)}\right)^2\|\bar{Q}_l\|_2\|B^l\|_2(\|A^l\|_2 + \bar{F} + \|B^l\|_2 c_1^l + 1),\\
c_3^l &= 8\left(\frac{10\mathbf{J}_l^1}{\sigma_{\min}(\bar{Q}_l)}\right)^2(c_1^l)^2\|C_U^l\|_2\|B^l\|_2(\|A^l\|_2 + \bar{F} + \|B^l\|_2 c_1^l + 1),\\
c_4^l &= 2\left(\frac{10\mathbf{J}_l^1}{\sigma_{\min}(\bar{Q}_l)}\right)^2(c_1^l + 1)\|C_U^l\|_2, c_5^l = \sqrt{\|C_U^l\|_2 + \|B^l\|_2^2(10\mathbf{J}_l^1)^2},\\
c_6^l &= \|C_U^l\|_F + \|B^l\|_F^2(c_1^l + 1)(c_2^l + c_3^l + c_4^l) + 10\|B^l\|_F^2\mathbf{J}_l^1\\
&\qquad + \|B^l\|_F(\|A^l\|_2 + \bar{F})(c_2^l + c_3^l + c_4^l),\\
c_7^l &= 50c_6^l\frac{\mathbf{J}_l^1}{\sigma_{\min}(\bar{Q}_l)} + 4c_5^l\left(\frac{10\mathbf{J}_l^1}{\sigma_{\min}(\bar{Q}_l)}\right)^2\|B^l\|_2(\|A^l\|_2 + \bar{F} + \|B^l\|_2 c_1^l) + c_1^l,\\
c_8^l &= \|\bar{\Sigma}^l\|_2(c_2^l + c_3^l + c_4^l),\\
c_9^l &= \min\left(\frac{\sigma_{\min}(\bar{Q}_l)}{40\mathbf{J}_l^1\|B^l\|_2(\|A^l\|_2 + \bar{F} + \|B^l\|_2 c_1^l + 1)}, 1\right)
\end{aligned}$$

The global constants for Lemma 1 can now be defined as

$$\mu^l \leq \frac{\mathbf{J}_l^1}{\sigma_{\min}^2(\bar{\Sigma}^l)\sigma_{\min}(C_U^l)\sigma_{\min}(\bar{Q}_l)}, \rho_1^l = c_9^l, \varphi_1^l = c_7^l, \lambda_1^l = c_8^l. \tag{125}$$

Now we move on to defining the constants ν^l, φ_2^l, ρ_2^l and λ_2^l for Lemma 2. First we find the upper bound for $\|\mathbf{A}_l\|_2$ in the definition of cost J_l^2 (96). From the definition (97),

$$\begin{aligned}
\|\mathbf{A}_l\|_2 \leq \big(&\|(I - \bar{A}^l + \bar{B}^l K_{l,1}^{(S_1)})^{-1}\|_2^2\|B^l\|_2^2 + 1\big)\\
&\big(\|\bar{Q}_l\|_2 + \|C_U^l\|_2\|K_{l,1}^{(S_1)}\|_2^2 + \|C_U^l\|_2\|K_{l,1}^{(S_1)}\|_2 + \|C_U^l\|_2\big)
\end{aligned} \tag{126}$$

First let us observe that

$$\begin{aligned}
I - \bar{A}^l + \bar{B}^l K_{l,1}^{(S_1)} &= \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} - \begin{pmatrix} A^l & 0 \\ 0 & F^{(S_1)} \end{pmatrix} + \begin{pmatrix} B^l \\ 0 \end{pmatrix}\begin{pmatrix} K_{l,1}^{(1,S_1)} & K_{l,1}^{(2,S_1)} \end{pmatrix},\\
&= \begin{pmatrix} I - A^l + B^l K_{l,1}^{(1,S_1)} & B^l K_{l,1}^{(2,S_1)} \\ 0 & I - F^{(S_1)} \end{pmatrix}
\end{aligned}$$

and thus

$$(I - \bar{A}^l + \bar{B}^l K_{l,1}^{(S_1)})^{-1} = \begin{pmatrix} (I - A^l + B^l K_{l,1}^{(1,S_1)})^{-1} & -(I - A^l + B^l K_{l,1}^{(1,S_1)})^{-1} B^l K_{l,1}^{(2,S_1)} (I - F^{(S_1)})^{-1} \\ 0 & (I - F^{(S_1)})^{-1} \end{pmatrix}$$

As a result,

$$\begin{aligned} \|(I - \bar{A}^l + \bar{B}^l K_{l,1}^{(S_1)})^{-1}\|_2 \leq \|(I - A^l + B^l K_{l,1}^{(1,S_1)})^{-1}\|_2 + \|(I - F^{(S_1)})^{-1}\|_2 \\ + \|(I - A^l + B^l K_{l,1}^{(1,S_1)})^{-1}\|_2 \|B^l K_{l,1}^{(2,S_1)}\|_2 \|(I - F^{(S_1)})^{-1}\|_2 \end{aligned} \tag{127}$$

So we bound the quantities $\|(I - A^l + B^l K_{l,1}^{(1,S_1)})^{-1}\|_2$ and $\|(I - F^{(S_1)})^{-1}\|_2$.

$$\begin{aligned} &\|(I - A^l + B^l K_{l,1}^{(1,S_1)})^{-1}\|_2 \\ &= \|(I - A^l + B^l K_{l,1}^{*,1} + B^l (K_{l,1}^{(1,S_1)} - K_{l,1}^{*,1}))^{-1})\|_2 \\ &= \|(I - A^l + B^l K^{*,1})^{-1} (I + (I - A^l + B^l K^{*,1})^{-1} B^l (K_{l,1}^{(1,S_1)} - K_{l,1}^{*,1}))^{-1}\|_2 \\ &\leq (1 - \rho(A^l - B^l K^{*,1}))^{-1} \|(I + (I - A^l + B^l K^{*,1})^{-1} B^l (K_{l,1}^{(1,S_1)} - K_{l,1}^{*,1}))^{-1}\|_2 \\ &\leq (1 - \rho(A^l - B^l K^{*,1}))^{-1} (1 - \|(I - A^l + B^l K^{*,1})^{-1} B^l\|_2)^{-1} =: c_{10}^l \end{aligned} \tag{128}$$

where the last inequality is due to (75) and the fact that $\epsilon \leq \frac{1}{\sqrt{m(L+1)}} \min_{l \in [L]} \left(1, \frac{1}{D_l^2}\right)$ which implies $\|F^{(S_1)} - F^*\|_2 \leq 1$ and $\|K_{l,1}^{(S_1)} - K_{l,1}^*\|_2 \leq 1$. Similarly,

$$\begin{aligned} \|(I - F^{(S_1)})^{-1}\|_2 &= \|(I - F^* + (F^* - F^{(S_1)}))^{-1}\|_2 \\ &= \|(I - F^*)^{-1} (I - (I - F^*)^{-1} (F^{(S_1)} - F^*))^{-1}\|_2 \\ &\leq (1 - \rho(F^*))^{-1} \|(I - (I - F^*)^{-1} (F^{(S_1)} - F^*))^{-1}\|_2 \\ &\leq (1 - \rho(F^*))^{-1} (1 - \|I - F^*\|_2)^{-1} =: c_{11}^l \end{aligned} \tag{129}$$

Similarly the following terms can be upper bounded,

$$\|B^l K_{l,1}^{(2,S_1)}\|_2 \leq \|B^l K_{l,1}^{*,2}\|_2 + \|B^l\|_2 =: c_{12}^l, \tag{130}$$

$$\|K_{l,1}^{(S_1)}\|_2 \leq \|K_{l,1}^*\|_2 + 1 =: c_{13}^l. \tag{131}$$

Using (126)-(131), we can bound

$$\|\mathbf{A}_l\|_2 \leq \bar{\mathbf{A}}_l \tag{132}$$

where

$$\bar{\mathbf{A}}_l := (c_{14}^l \|B^l\|_2^2 + 1)(\|\bar{Q}_l\|_2 + \|C_U^l\|_2 (1 + c_{13}^l + (c_{13}^l)^2)), \tag{133}$$

and thus

$$c_{14}^l = c_{10}^l + c_{10}^l c_{11}^l c_{12}^l + c_{12}^l \geq \|(I - \bar{A}^l + \bar{B}^l K_{l,1}^{(S_1)})^{-1}\|_2. \tag{134}$$

Now we move on to the bound on $J_l((K_{l,1}^{(S_1)}, K_{l,2}^{(1)}), \bar{Z}^{(s)})$. From the definition of cost J_l,

$$\begin{aligned} J_l((K_{l,1}^{(S_1)}, K_{l,2}^{(1)}), \bar{Z}^{(s+S_1)}) &= J_l^1(K_{l,1}^{(S_1)}, \bar{Z}^{(s+S_1)}) \\ &\quad + J_l^2((K_{l,1}^{(S_1)}, K_{l,2}^{(1)}), \bar{Z}^{(s+S_1)}) + (\bar{\beta}^l)^\top \bar{Q}_l \bar{\beta}^l \\ &\leq \mathbf{J}_l^1 + J_l^2((K_{l,1}^{(S_1)}, K_{l,2}^{(1)}), \bar{Z}^{(s+S_1)}) + (\bar{\beta}^l)^\top \bar{Q}_l \bar{\beta}^l \end{aligned}$$

Hence, we need to bound $J_l^2((K_{l,1}^{(S_1)}, K_{l,2}^{(1)}), \bar{Z}^{(s+S_1)})$. Recall the definition (96)

$$J_l^2(K_{l,2}) = K_{l,2}^\top \mathbf{A}_l K_{l,2} + \mathbf{c}_l^\top K_{l,2} + \mathbf{d}_l$$

where $\mathbf{A}_l$, $\mathbf{c}_l$ and $\mathbf{d}_l$ are defined in (97). As shown in (132) $\|\mathbf{A}_l\|_2 \leq \bar{\mathbf{A}}_l$. Using definition of $\mathbf{c}_l$ as defined in (97),

$$\begin{aligned} \|\mathbf{c}_l\|_2 &\leq 2\|(I - \bar{A}^l + \bar{B} K_{l,1}^{(S_1)})^{-1}\|_2^2 (\|\bar{Q}_l\|_2 + \|C_U^l\|_2 \|K_{l,1}^{(S_1)}\|_2^2) \|\bar{B}^l\|_2 \|\bar{C}^{(s)}\|_2 \\ &\quad + 2\|(I - \bar{A}^l + \bar{B} K_{l,1}^{(S_1)})^{-1}\|_2 (\|C_U^l\|_2 \|K_{l,1}^{(S_1)}\|_2 + 2\|(\bar{\beta}^l)^\top \bar{Q}_l\|_2 \|\bar{B}^l\|_2) \\ &\leq 2(c_{14}^l)^2 (\|\bar{Q}_l\|_2 + \|C_U^l\|_2 (2\|K_{l,1}^*\|_2^2 + 2)) \|\bar{B}^l\|_2 \bar{C} \\ &\quad + c_{14}^l (\|C_U^l\|_2 (\|K_{l,1}^*\|_2 + 1) + 2\|(\bar{\beta}^l)^\top \bar{Q}_l\|_2 \|\bar{B}^l\|_2) =: \bar{\mathbf{c}}_l \end{aligned} \tag{135}$$

The last inequality is obtained using (134), (90) and the fact that $\|K_{l,1}^* - K_{l,1}^{(S_1)}\|_2 \leq 1$. Similarly using the definition of $\mathbf{d}_l$ we obtain

$$\|\mathbf{d}_l\|_2 \leq (c_{14}^l)^2 (\|\bar{Q}_l\|_2 + \|C_U^l\|_2 (2\|K_{l,1}^*\|_2^2 + 2)) \bar{C}^2 + 2c_{14}^l \|\bar{\beta}^l \bar{Q}_l\|_2 \bar{C}^2 =: \bar{\mathbf{d}}_l \tag{136}$$

Now we can bound $J_l^2((K_{l,1}^{(S_1)}, K_{l,2}^{(1)}), \bar{Z}^{(s+S_1)})$ as follows:

$$J_l^2((K_{l,1}^{(S_1)}, K_{l,2}^{(1)}), \bar{Z}^{(s+S_1)}) \leq \bar{\mathbf{A}}_l \|K_{l,2}^{(1)}\|_2^2 + \bar{\mathbf{c}}_l \|K_{l,2}^{(1)}\|_2 + \bar{\mathbf{d}}_l =: \mathbf{J}_l^2 \tag{137}$$

Hence,

$$\nu^l = \bar{\mathbf{A}}_l, \quad \varphi_2^l = \bar{\mathbf{A}}_l, \quad \rho_2^l = \sqrt{\frac{4\mathbf{J}_l^2}{\|\mathbf{A}_l\|_2}}, \lambda_2^l = \sqrt{80\varphi_2^l \mathbf{J}_l^2} \tag{138}$$

concluding the proof. □

References

1. Achdou Y, Dao M-K, Ley O, Tchou N (2020) Finite horizon mean field games on networks. Calc Var Partial Differ Equ 59(5):1–34
2. Anahtarcı B, Karıksız CD, Saldi N (2019) Fitted Q-learning in mean-field games. arXiv preprint arXiv:1912.13309
3. Bauso D (2017) Consensus via multi-population robust mean-field games. Syst Control Lett 107:76–83
4. Bauso D, Tembine H, Başar T (2016) Opinion dynamics in social networks through mean-field games. SIAM J Control Optim 54(6):3225–3257
5. Bensoussan A, Sung K, Yam SCP, Yung S-P (2016) Linear-quadratic mean field games. J Optim Theory Appl 169(2):496–529
6. Bryson AE, Ho Y-C (1975) Applied optimal control, revised printing. Hemisphere, New York
7. Caines, PE, Huang M (2019) Graphon mean field games and the GMFG equations: ε-Nash equilibria. In: 2019 IEEE 58th conference on decision and control (CDC), pp 286–292. IEEE
8. Camilli F, Marchi C (2016) Stationary mean field games systems defined on networks. SIAM J Control Optim 54(2):1085–1103

9. Carmona R, Delarue F (2018) Probabilistic theory of mean field games with applications I. Springer, Cham
10. Delarue F (2017) Mean field games: a toy model on an Erdös–Renyi graph. ESAIM Proc Surv 60:1–26
11. Elie R, Pérolat J, Laurière M, Geist M, Pietquin O (2019) Approximate fictitious play for mean field games. arXiv preprint arXiv:1907.02633
12. Fazel M, Ge R, Kakade SM, Mesbahi M (2018) Global convergence of policy gradient methods for the linear quadratic regulator. In: International conference on machine learning, pp 1467–1476
13. Fu Z, Yang Z, Chen Y, Wang Z (2020) Actor-critic provably finds Nash equilibria of linear-quadratic mean-field games. In: International conference on learning representation
14. Gao S, Caines PE, Huang M (2020) LQG graphon mean field games. arXiv preprint arXiv:2004.00679
15. Gu D (2007) A differential game approach to formation control. IEEE Trans Control Syst Technol 16(1):85–93
16. Guo X, Hu A, Xu R, Zhang J (2019) Learning mean-field games. In: Advances in neural information processing systems
17. Huang M, Zhou M (2018) Linear quadratic mean field games–Part I: the asymptotic solvability problem. arXiv preprint arXiv:1811.00522
18. Huang M, Malhamé RP, Caines PE et al (2006) Large population stochastic dynamic games: Closed-loop Mckean-Vlasov systems and the Nash certainty equivalence principle. Commun Inf Syst 6(3):221–252
19. Huang M, Caines PE, Malhamé RP (2007) Large-population cost-coupled LQG problems with nonuniform agents: individual-mass behavior and decentralized ε-Nash equilibria. IEEE Trans Autom Control 52(9):1560–1571
20. Lasry J-M, Lions P-L (2007) Mean field games. Jpn J Math 2(1):229–260
21. Lewis FL, Zhang H, Hengster-Movric K, Das A (2013) Cooperative control of multi-agent systems: optimal and adaptive design approaches. Springer, Berlin
22. Malik D, Pananjady A, Bhatia K, Khamaru K, Bartlett P, Wainwright M (2019) Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. In: The 22nd international conference on artificial intelligence and statistics, pp 2916–2925. PMLR
23. Moon J, Başar T (2014) Discrete-time LQG mean field games with unreliable communication. In: 53rd IEEE conference on decision and control, pp 2697–2702. IEEE
24. Saldi N, Başar T, Raginsky M (2018) Markov-Nash equilibria in mean-field games with discounted cost. SIAM J Control Optim 56(6):4256–4287
25. Spall JC (2005) Introduction to stochastic search and optimization: estimation, simulation, and control, vol 65. Wiley, Hoboken
26. Subramanian J, Mahajan A (2019) Reinforcement learning in stationary mean-field games. In: International conference on autonomous agents and multiagent systems, pp 251–259
27. Yang Z, Chen Y, Hong M, Wang Z (2019) Provably global convergence of actor-critic: A case for linear quadratic regulator with ergodic cost. In: Advances in neural information processing systems, pp 8351–8363
28. Zaman MAu, Zhang K, Miehling E, Başar T (2020a) Approximate equilibrium computation for discrete-time linear-quadratic mean-field games. In: 2020 American control conference (ACC), pp 333–339. IEEE
29. Zaman MAu, Zhang K, Miehling E, Başar T (2020b) Reinforcement learning in non-stationary discrete-time linear-quadratic mean-field games. In: 2020 59th IEEE conference on decision and control (CDC), pp 2278–2284. IEEE
30. Zeng Y, Wu Q, Zhang R (2019) Accessing from the sky: a tutorial on UAV communications for 5G and beyond. Proc IEEE 107(12):2327–2375
31. Zhu Q, Başar T (2011) A multi-resolution large population game framework for smart grid demand response management. In: International conference on network games, control and optimization (NetGCooP 2011), pp 1–8. IEEE