Data-Driven Network LQG Mean Field Games with Heterogeneous Populations via Integral Reinforcement Learning ⋆

Jean Zhu * **Shuang Gao** **

* *Department of Mechanical Engineering, Polytechnique Montréal & GERAD, Montreal, Canada (email: jean.zhu@etud.polymtl.ca)*
** *Department of Electrical Engineering, Polytechnique Montréal & GERAD, Montreal, Canada (email: shuang.gao@polymtl.ca)*

Abstract: This paper establishes a data-driven solution for infinite horizon linear quadratic Gaussian Mean Field Games with network-coupled heterogeneous agent populations where the dynamics of the agents are unknown. The solution technique relies on Integral Reinforcement Learning and Kleinman's iteration for solving algebraic Riccati equations (ARE). The resulting algorithm uses trajectory data to generate network-coupled MFG strategies for agents and does not require parameters of agents' dynamics. Under technical conditions on the persistency of excitation and on the existence of unique stabilizing solution to the corresponding AREs, the learned network-coupled MFG strategies are shown to converge to their true values.

Keywords: Mean field games, data-driven control, integral reinforcement learning, adaptive dynamic programming, multi-agent systems

1. INTRODUCTION

Large-scale systems composed of heterogeneous agent populations, such as renewable energy grids with different types of renewable sources and various types of users, and autonomous vehicle networks with different types of vehicles, naturally arise in the transition to sustainable energy systems. Such systems motivate the modelling and the strategy design in this paper for heterogeneous populations of strategically competitive agents.

To achieve tractable strategy design for large populations of competitive agents, Mean Field Game (MFG) theory was proposed by Huang et al. (2006, 2007) and independently by Lasry and Lions (2006, 2007). In the linear-quadratic-Gaussian (LQG) case, the MFG solution involves solving coupled Riccati equations in (Huang et al., 2007; Huang, 2010). MFG problems with multiple classes have been investigated in (Huang et al., 2006, 2007; Huang, 2010) and MFGs with network interactions in e.g. (Huang et al., 2010; Gao et al., 2023).

For MFGs where the dynamics of agents are unknown, several data-driven solutions have been established. In continuous-time settings, adaptive control techniques with system identification have been applied to MFGs with heterogeneous populations with mean field cost couplings by Kizilkale and Caines (2012), and integral reinforcement learning (IRL) has been used by Xu et al. (2023, 2025a); Li et al. (2025) to generate the data-driven strategies for LQG mean field game problems without system identification, where a homogeneous population of agents with mean field coupling is assumed. For discrete-time MFGs, standard reinforcement learning techniques have been employed in (Subramanian and Mahajan, 2019; Guo et al., 2019; Fu et al., 2019; Zaman et al., 2020, 2023; Angiuli et al., 2022).

IRL was developed to generate data-driven optimal control solutions for continuous-time systems with unknown dynamics (e.g. partially unknown nonlinear dynamics (Vrabie and Lewis, 2009), completely unknown linear dynamics (Jiang and Jiang, 2012) and stochastic dynamics (Li et al., 2022)). Such a data-driven technique also applies to multi-agent systems with unknown dynamics in (Vamvoudakis and Lewis, 2011) and cases with heterogeneous agents in (Modares et al., 2016). In the context of mean field control with agent populations, IRL was used for continuous-time problems in (Xu et al., 2025b) whereas standard reinforcement learning has been empolyed in (Subramanian and Mahajan, 2019; Carmona et al., 2019; Angiuli et al., 2022) for discrete-time problems.

Contribution: This current paper establishes (a) the LQG-MFG strategies with multi-class heterogeneous populations with inter-class network couplings and (b) an IRL algorithm for learning the agent strategies from trajectory data, extending the IRL algorithm for homogeneous LQG-MFG by Xu et al. (2023, 2025a). The heterogeneity of agent classes and the network interaction lead to a new set of algebraic Riccati equations (ARE), one for each class, and one capturing cross-class interactions, all of which can be learned simultaneously using the proposed algorithm from trajectory data without knowing the parameters of the underlying system dynamics.

Notation and Definition: $\mathbb{R}$ and $\mathbb{N}$ denote the set of real and nonzero natural numbers respectively. For a random variable x, $\bar{x}$ denotes its expectation $\bar{x} = \mathbb{E}[x]$. For a matrix A, $A^\top$ denotes its transpose. For any $n, m \in \mathbb{N}$, let $I_n \in \mathbb{R}^{n\times n}$ denote the identity matrix,

⋆ This work is supported in part by FRQNT and NSERC (Canada).

$\mathbf{0}_{n\times m} \in \mathbb{R}^{n\times m}$ the zero matrix, and $\mathbf{1}_{n\times m} \in \mathbb{R}^{n\times m}$ the matrix of ones. For a matrix $A \in \mathbb{R}^{m\times n}$, $\mathrm{vec}(A)$ corresponds to the mn-dimensional vector formed by stacking the columns of A on top of each other. $\otimes$ denotes the Kronecker product between two matrices. $A \succeq 0$ (resp. $A \succ 0$) denotes that matrix A is positive semidefinite (resp. positive definite). $\mathrm{diag}(M_1, \ldots, M_N)$ denotes the matrix with diagonal blocks $M_1, \cdots, M_N$ and zero elsewhere. For $K, n_k, m_l, N, M \in \mathbb{N}$, $\mathcal{M} = [\![M_{kl}]\!] \in \mathbb{R}^{N\times M}$, where $N = \sum_{k=1}^{K} n_k$ and $M = \sum_{l=1}^{K} m_l$, denotes a block matrix with blocks $M_{kl} \in \mathbb{R}^{n_k\times m_l}$, with $k, l \in \{1, ..., K\}$.

Consider a matrix $A \in \mathbb{R}^{k\times k}$. A is called *strong* (k_1, k_2) *c-splitting* if, of its $k = k_1 + k_2$ eigenvalues, it has k_1 eigenvalues in the open left half plane, and k_2 in the open right half plane, with none on the imaginary axis. An l-dimensional subspace (with $l \le k$) of A denoted by $\mathcal{V}$ of $\mathbb{R}^k$ is an *invariant subspace* if $A\mathcal{V} \subset \mathcal{V}$; in this case, $AV = VA_0$ for some $A_0 \in \mathbb{R}^{l\times l}$ where $V \in \mathbb{R}^{k\times l}$ and $\mathrm{span}(V) = \mathcal{V}$. If A_0 is Hurwitz, $\mathcal{V}$ is called a *stable invariant subspace.* An l-dimensional subspace (with $l \le k$) of $A \in \mathbb{R}^{k\times k}$ is called a *graph subspace* if it's spanned by the columns of a $k \times l$ matrix whose first l rows form an invertible submatrix (see Huang and Zhou (2020)).

2. PROBLEM FORMULATION

Consider a population of agents grouped into $\mathcal{K}$ classes, where each class $k \in \{1, \ldots, \mathcal{K}\}$ contains a large number of identical agents. Let c_k denotes the number of agents in class k. An agent $a_{k,i}$ of class k has dynamics given by

$$dx_{k,i}(t) = (A_k x_{k,i}(t) + B_k u_{k,i}(t))dt + D_k dw_{k,i}(t) \quad (1)$$

where $x_{k,i}(t) \in \mathbb{R}^{n_k}$ and $u_{k,i}(t) \in \mathbb{R}^{m_k}$ are the states and control inputs respectively, and $w_{k,i}(t) \in \mathbb{R}^{d_k}$ is a standard Wiener process. A_k, B_k and D_k are system matrices for class k, which are all assumed to be unknown. For simplicity of presentation, we omit the time index hereafter. An agent $a_{k,i}$ seeks to find the control $u_{k,i}$ that minimizes the class-specific discounted cost functional

$$J_k(u_{k,i}) = \mathbb{E}\int_0^{\infty} e^{-\rho t}\left(\tilde{x}_{k,i}^\top Q_k \tilde{x}_{k,i} + u_{k,i}^\top R_k u_{k,i}\right) dt \quad (2)$$

where $\tilde{x}_{k,i} = x_{k,i} - \sum_{m=1}^{\mathcal{K}} H_{km}\bar{x}_m$ is the tracking error with $\bar{x}_m = \frac{1}{c_m}\sum_{i=1}^{c_m} x_{m,i}$, $Q_k = Q_k^\top \succeq 0$ and $R_k = R_k^\top \succ 0$ the class-specific state cost and control cost matrices respectively, $\rho > 0$ is the discount factor, and $H_{km} \in \mathbb{R}^{n_k\times n_m}$ is a class-mean coupling matrix representing how the states of an agent of class k depends on the means of the states of class m. The resulting $\mathcal{H} = [\![H_{km}]\!] \in \mathbb{R}^{N\times N}$ with $N = \sum_{k=1}^{\mathcal{K}} n_k$ is the network coupling matrix.

Remark 1. In the special case where the agents have homogeneous state dimensions $n_k = n$ for all k and are to track the global mean field $\bar{x}(t)$, defined as the convex combination of the mean of all classes

$$\bar{x}(t) = \sum_{k=1}^{\mathcal{K}} \pi_k \bar{x}_k(t), \qquad \pi_k > 0, \qquad \sum_{k=1}^{\mathcal{K}} \pi_k = 1 \quad (3)$$

where $\bar{x}_k(t)$ is the mean of all agents belonging to class k and π_k are scalar weights, then the coupling matrices H_{km} are to be chosen such that

$$H_{km} = \pi_m I_n \ \text{ for all } k, m \in \{1, \ldots, \mathcal{K}\}.$$

Assumption 1. The pair $(A_k - \frac{1}{2}\rho I_n, B_k)$ is stabilizable and the pair $(A_k - \frac{1}{2}\rho I_n, Q_k^{\frac{1}{2}})$ is observable for all class $k \in \{1, ..., \mathcal{K}\}$.

The following block diagonal matrices are then defined.

$$\mathcal{A} = \mathrm{diag}(A_1, \ldots, A_{\mathcal{K}}), \quad \mathcal{B} = \mathrm{diag}(B_1, \ldots, B_{\mathcal{K}}),$$
$$\mathcal{Q} = \mathrm{diag}(Q_1, \ldots, Q_{\mathcal{K}}), \quad \mathcal{R} = \mathrm{diag}(R_1, \ldots, R_{\mathcal{K}}),$$

where $\mathcal{A}, \mathcal{Q} \in \mathbb{R}^{N\times N}$, $\mathcal{B} \in \mathbb{R}^{N\times M}$, and $\mathcal{R} \in \mathbb{R}^{M\times M}$ with

$$M = \sum_{k=1}^{\mathcal{K}} m_k \quad \text{and} \quad N = \sum_{k=1}^{\mathcal{K}} n_k.$$

Remark 2. Clearly $\mathcal{R} \succ 0$. In addition, under Assumption 1, the pair $(\mathcal{A} - \frac{1}{2}\rho I_N, \mathcal{B})$ is stabilizable and the pair $(\mathcal{A} - \frac{1}{2}\rho I_N, \mathcal{Q})$ is observable as an immediate consequence.

Assumption 2. The eigenvalues of the Hamiltonian matrix $H_\Omega \in \mathbb{R}^{2N\times 2N}$ defined by

$$H_\Omega = \begin{bmatrix} \mathcal{A} - \dfrac{1}{2}\rho I_N & -\mathcal{B}\mathcal{R}^{-1}\mathcal{B}^\top \\ -\mathcal{Q}(I_N - \mathcal{H}) & -\mathcal{A}^\top + \dfrac{1}{2}\rho I_N \end{bmatrix} \in \mathbb{R}^{2N\times 2N} \quad (4)$$

are strong (N, N) c-splitting, and the associated N-dimensional stable invariant subspace is a graph subspace.

Following the fixed point approach in (Huang et al., 2007; Huang and Zhou, 2020), we let the population size in all classes go to infinity and solve the corresponding limit problem, which generates the following MFG strategy.

Proposition 1. Under Assumptions 1 and 2, the MFG strategy of a generic agent $a_{k,i}$ in class $k \in \{1, ..., \mathcal{K}\}$ exists and is uniquely given by

$$u_{k,i}^*(t) = -R_k^{-1}B_k^\top(P_k x_{k,i} + s_k) \quad (5)$$

where $P_k \succ 0$ and $s_k = \sum_{m=1}^{\mathcal{K}} \Pi_{km}\bar{x}_m$ follow from the solutions to the following algebraic equations

$$\rho\mathcal{P} = \mathcal{Q} + \mathcal{P}\mathcal{A} + \mathcal{A}^\top\mathcal{P} - \mathcal{P}\mathcal{B}\mathcal{R}^{-1}\mathcal{B}^\top\mathcal{P} \quad (6a)$$

$$\begin{aligned} \rho\Pi = &-\mathcal{Q}\mathcal{H} + \Pi(\mathcal{A} - \mathcal{B}\mathcal{R}^{-1}\mathcal{B}^\top(\mathcal{P} + \Pi)) \\ &+ (\mathcal{A}^\top - \mathcal{P}\mathcal{B}\mathcal{R}^{-1}\mathcal{B}^\top)\Pi \end{aligned} \quad (6b)$$

and the mean field dynamics

$$\dot{\bar{\mathcal{X}}} = (\mathcal{A} - \mathcal{B}\mathcal{R}^{-1}\mathcal{B}^\top(\mathcal{P} + \Pi))\bar{\mathcal{X}} \quad (7)$$

with $\mathcal{P} \in \mathbb{R}^{N\times N}$, $\bar{X} \in \mathbb{R}^N$ given by

$$\mathcal{P} = \mathrm{diag}(P_1, \ldots, P_{\mathcal{K}}), \quad \bar{\mathcal{X}} = [\bar{x}_1^\top, \ldots, \bar{x}_{\mathcal{K}}^\top]^\top,$$

and $\Pi = [\![\Pi_{km}]\!] \in \mathbb{R}^{N\times N}$.

See Appendix A for the detailed proof.

By defining $\Omega = \mathcal{P} + \Pi$ and summing the two equations in (6), the coupled equations are reduced to a set of decoupled Algebraic Riccati Equations (AREs) in $\mathcal{P}$ and Ω:

$$\rho\mathcal{P} = \mathcal{Q} + \mathcal{P}\mathcal{A} + \mathcal{A}^\top\mathcal{P} - \mathcal{P}\mathcal{B}\mathcal{R}^{-1}\mathcal{B}^\top\mathcal{P} \quad (8a)$$

$$\rho\Omega = \mathcal{Q}(I_N - \mathcal{H}) + \Omega\mathcal{A} + \mathcal{A}^\top\Omega - \Omega\mathcal{B}\mathcal{R}^{-1}\mathcal{B}^\top\Omega \quad (8b)$$

where $\mathcal{P}$ and Ω can be solved simultaneously and Π obtained by $\Pi = \Omega - \mathcal{P}$.

A solution to (8b) is called *stabilizing* if the closed-loop matrix $\mathcal{A} - \frac{1}{2}\rho I_N - \mathcal{B}\mathcal{R}^{-1}\mathcal{B}^\top\Omega$ is Hurwitz. Assumption 2 ensures the existence of a unique stabilizing solution Ω to (8b) following (Huang and Zhou, 2020, Thm. 18).

Since all involved matrices in (8a) are block diagonal, $\mathcal{P}$ can be solved class by class by simultaneously solving for

P_k for each class k, where P_k is the solution to the following class-specific ARE

$$\rho P_k = Q_k + P_k A_k + A_k^\top P_k - P_k B_k R_k^{-1} B_k^\top P_k. \quad (9)$$

3. DATA-DRIVEN MULTI-CLASS LQG-MFG WITH COMPLETELY UNKNOWN DYNAMICS

To establish the multi-class LQG-MFG strategy purely based on trajectory data, we adapt the data-driven approach developed by Jiang and Jiang (2012) for linear quadratic control problems and later applied to single class LQG-MFGs by Xu et al. (2023, 2025a). This approach combines the Kleinman algorithm (Kleinman, 1968) for iteratively solving symmetric ARE and the IRL technique for updating strategies based on trajectory data.

Since the data-driven method based on (Jiang and Jiang, 2012; Xu et al., 2023) requires the ARE to be symmetric, we introduce the following assumption.

Assumption 3. The matrix $\mathcal{Q}(I_N - \mathcal{H})$ is symmetric.

Kleinman's Iteration Procedure: Under Assumptions 1,2 and 3, initially stabilizing gains $L_{P,k}^{(0)}$ and $\mathcal{L}_\Omega^{(0)}$ can be selected. An iterative procedure is then formed to solve equation (9) for P_k, for all k, and (8b) for Ω. Equations (9) and (8b) can thus be solved iteratively by

$$\begin{aligned}\rho P_k^{(\ell)} &= P_k^{(\ell)}(A_k - B_k L_{P,k}^{(\ell-1)}) + (A_k - B_k L_{P,k}^{(\ell-1)})^\top P_k^{(\ell)} \\ &\quad + (L_{P,k}^{(\ell-1)})^\top R_k L_{P,k}^{(\ell-1)} + Q_k \quad (10a)\\ \rho \Omega^{(\ell)} &= \Omega^{(\ell)}(\mathcal{A} - \mathcal{B}\mathcal{L}_\Omega^{(\ell-1)}) + (\mathcal{A} - \mathcal{B}\mathcal{L}_\Omega^{(\ell-1)})^\top \Omega^{(\ell)} \\ &\quad + (\mathcal{L}_\Omega^{(\ell-1)})^\top \mathcal{R} \mathcal{L}_\Omega^{(\ell-1)} + \mathcal{Q}(I_N - \mathcal{H}) \quad (10b)\end{aligned}$$

where $L_{P,k}^{(\ell)} = R_k^{-1} B_k^\top P_k^{(\ell)}$ and $\mathcal{L}_\Omega^{(\ell)} = \mathcal{R}^{-1}\mathcal{B}^\top \Omega^{(\ell)}$. $\mathcal{P}$ can then be reconstructed by $\mathcal{P} = \mathrm{diag}(P_1, \ldots, P_\mathcal{K})$.

Proposition 2. Under Assumptions 1,2 and 3, the resulting matrices $P_k^{(\ell)}$, $\Omega_k^{(\ell)}$, $L_{P,k}^{(\ell)}$ and $\mathcal{L}_\Omega^{(\ell)}$ from the Kleinman's iteration procedure above satisfy the following properties:

(1) $A_k - \frac{1}{2}\rho I_{n_k} - B_k L_{P,k}^{(\ell)}$ and $\mathcal{A} - \frac{1}{2}\rho I_N - \mathcal{B}\mathcal{L}_\Omega^{(\ell)}$ are Hurwitz,
(2) $P_k^* \preceq P_k^{(\ell+1)} \preceq P_k^{(\ell)}$ and $\Omega^* \preceq \Omega^{(\ell+1)} \preceq \Omega^{(\ell)}$,
(3) $\lim_{\ell\to\infty} L_{P,k}^{(\ell)} = L_{P,k}^*$, $\lim_{\ell\to\infty} P_k^{(\ell)} = P_k^*$, $\lim_{\ell\to\infty} \mathcal{L}_\Omega^{(\ell)} = \mathcal{L}_\Omega^*$, and $\lim_{\ell\to\infty} \Omega^{(\ell)} = \Omega^*$, □

where P_k^* is the unique positive definite solution to the ARE (9) for class $k \in \{1, 2, ..., \mathcal{K}\}$ and Ω^* the unique stablizing solution to (8b).

Proof. Since the class-specific AREs (9) form $\mathcal{K}$ independent standard Riccati equations associated with LQR problems, the convergence of the iteration for P_k follows directly from (Kleinman, 1968). The ARE for Ω in (8b) corresponds to a global LQR problem with a state penalty matrix $\mathcal{Q}(I_N - \mathcal{H})$, which may be indefinite. Under Assumptions 1, 2, and 3, the convergence of the policy iteration for Ω follows from (Xu et al., 2025a, Lemma 4.1), which extends Kleinman's iteration to symmetric indefinite AREs.

In order to solve the AREs without the knowledge of system dynamics, a representative agent $a_{k,1}$ is selected for each class k, and $\mathcal{X} = [x_{1,1}^\top, x_{2,1}^\top, \ldots, x_{\mathcal{K},1}^\top]^\top \in \mathbb{R}^N$ is defined as the augmented states vector, and $\mathcal{U} = [u_{1,1}^\top, u_{2,1}^\top, \ldots, u_{\mathcal{K},1}^\top]^\top \in \mathbb{R}^M$, the augmented control inputs vector. The value function ansatzs for each P_k and another for Ω are respectively defined as follows:

$$\Psi_{1,k}(t, x_{k,1}) = e^{-\rho t} x_{k,1}^\top P_k^{(\ell)} x_{k,1} \quad (11a)$$

$$\Psi_2(t, \mathcal{X}) = e^{-\rho t} \mathcal{X}^\top \Omega^{(\ell)} \mathcal{X}. \quad (11b)$$

Under Assumption 1, a control input

$$\begin{aligned}\mathcal{U}(t) = \alpha(t) &= [\alpha_{1,1}^\top(t), \ldots, \alpha_{\mathcal{K},1}^\top(t)]^\top \\ &= -\mathcal{L}_\Omega^{(0)} \mathcal{X}(t) + l(t) \quad (12)\end{aligned}$$

can be selected, where $\alpha(t)$ denotes the control input during the learning phase, composed of an initially stabilizing global gain $\mathcal{L}_\Omega^{(0)}$ such that $\mathcal{A} - \frac{1}{2}\rho I_N - \mathcal{B}\mathcal{L}_\Omega^{(0)}$ is Hurwitz, and a global exploratory noise $l(t) = [l_{1,1,1}(t), \ldots, l_{\mathcal{K},1,m_k}(t)]^\top \in \mathbb{R}^M$, composed of independent noise inputs for all elements of $\mathcal{U}(t)$. The associated $L_k^{(0)}$ included in $\mathcal{L}_\Omega^{(0)}$ are also selected to be initially stabilizing gain for each class k such that $A_k - \frac{1}{2}\rho I_{n_k} - B_k L_k^{(0)}$ are Hurwitz. The dynamics of the representative agents and the augmented states are given by

$$dx_{k,1}(t) = (A_k x_{k,1}(t) + B_k \alpha_{k,1}(t))dt + D_k dw_{k,1}(t) \quad (13a)$$

$$d\mathcal{X}(t) = (\mathcal{A}\mathcal{X}(t) + \mathcal{B}\alpha(t))\, dt + \mathcal{D} d\mathcal{W} \quad (13b)$$

where

$$\mathcal{D} = \mathrm{diag}(D_1, \ldots, D_\mathcal{K}), \quad \mathcal{W}(t) = [w_{1,1}^\top(t), \ldots, w_{\mathcal{K},1}^\top(t)]^\top.$$

Applying Itô's formula to (11), and using (10) and (13), we obtain

$$\begin{aligned}d\Psi_{1,k} = e^{-\rho t}\Big[&- x_{k,1}^\top (L_{P,k}^{(\ell-1)})^\top R_k L_{P,k}^{(\ell-1)} x_{k,1} \quad (14a)\\ &- x_{k,1}^\top Q_k x_{k,1} + 2(\alpha_{k,1} + L_{P,k}^{(\ell-1)} x_{k,1})^\top R_k L_{P,k}^{(\ell)} x_{k,1} \\ &+ \mathrm{Tr}(D_k D_k^\top P_k^{(\ell)})\Big] dt + 2e^{-\rho t} x_{k,1}^\top P_k^{(\ell)} D_k dw_{k,1}\end{aligned}$$

$$\begin{aligned}d\Psi_2 = e^{-\rho t}\Big[&- \mathcal{X}^\top (\mathcal{L}_\Omega^{(\ell-1)})^\top \mathcal{R} \mathcal{L}_\Omega^{(\ell-1)} \mathcal{X} \quad (14b)\\ &- \mathcal{X}^\top \mathcal{Q}(I_N - \mathcal{H})\mathcal{X} + 2(\alpha + \mathcal{L}_\Omega^{(\ell-1)}\mathcal{X})^\top \mathcal{R} \mathcal{L}_\Omega^{(\ell)} \mathcal{X} \\ &+ \mathrm{Tr}(\mathcal{D}\mathcal{D}^\top \Omega^{(\ell)})\Big] dt + 2e^{-\rho t} \mathcal{X}^\top \Omega^{(\ell)} \mathcal{D} d\mathcal{W}.\end{aligned}$$

Let Δt denote the length of integration chosen for transforming the trajectory data. Integrating both sides for a time interval $[t, t + \Delta t]$ and taking its expectation, the expectation of the integral forms of (14) is given by

$$\Delta\Psi_{1,k}^t = -\mathcal{I}_{q,k}^t + 2\mathcal{I}_{1,P,k}^t + \mathcal{I}_{2,P,k}^t \quad (15a)$$

$$\Delta\Psi_2^t = -\mathcal{I}_K^t + 2\mathcal{I}_{1,\Omega}^t + \mathcal{I}_{2,\Omega}^t \quad (15b)$$

where the terms for (15a) are defined by

$$\begin{aligned}\Delta\Psi_{1,k}^t = \mathbb{E}\big[&e^{-\rho(t+\Delta t)} x_{k,1}^\top(t+\Delta t) P_k^{(\ell)} x_{k,1}(t+\Delta t) \\ &- e^{-\rho t} x_{k,1}^\top(t) P_k^{(\ell)} x_{k,1}(t)\big]\end{aligned}$$

$$\mathcal{I}_{q,k}^t = \mathbb{E}\int_t^{t+\Delta t} e^{-\rho\tau}\Big[x_{k,1}^\top\big((L_{P,k}^{(\ell-1)})^\top R_k L_{P,k}^{(\ell-1)} + Q_k\big)x_{k,1}\Big] d\tau$$

$$\mathcal{I}_{1,P,k}^t = \mathbb{E}\int_t^{t+\Delta t} e^{-\rho\tau}(\alpha_{k,1} + L_{P,k}^{(\ell-1)} x_{k,1})^\top R_k L_{P,k}^{(\ell)} x_{k,1} d\tau$$

$$\mathcal{I}_{2,P,k}^t = \tfrac{1}{\rho}(e^{-\rho t} - e^{-\rho(t+\Delta t)})\,\mathrm{Tr}(D_k D_k^\top P_k^{(\ell)}),$$

and the terms for (15b) are defined by

$$\Delta\Psi_2^t = \mathbb{E}\big[e^{-\rho(t+\Delta t)} \mathcal{X}^\top(t+\Delta t) \Omega^{(\ell)} \mathcal{X}(t+\Delta t)$$

$$- e^{-\rho t}\mathcal{X}^\top(t)\Omega^{(\ell)}\mathcal{X}(t)]$$

$$\mathcal{I}_K^t = \mathbb{E}\int_t^{t+\Delta t} e^{-\rho\tau}\Big[\mathcal{X}^\top\big((\mathcal{L}_\Omega^{(\ell-1)})^\top\mathcal{R}\mathcal{L}_\Omega^{(\ell-1)} + \mathcal{Q}(I_N-\mathcal{H})\big)\mathcal{X}\Big]d\tau$$

$$\mathcal{I}_{1,\Omega}^t = \mathbb{E}\int_t^{t+\Delta t} e^{-\rho\tau}(\alpha+\mathcal{L}_\Omega^{(\ell-1)}\mathcal{X})^\top\mathcal{R}\mathcal{L}_\Omega^{(\ell)}\mathcal{X}d\tau$$

$$\mathcal{I}_{2,\Omega}^t = \tfrac{1}{\rho}(e^{-\rho t}-e^{-\rho(t+\Delta t)})\operatorname{Tr}(\mathcal{D}\mathcal{D}^\top\Omega^{(\ell)}).$$

Equations (15) can be expressed using the Kronecker product representation, which yields

$$\begin{aligned}0 &= (\delta_{P,k}^t)^\top\hat{P}_k^{(\ell)} + \delta_\rho^t\theta_{P,k}^{(\ell)} \\ &+ (\mathcal{I}_{xx,k}^t)^\top\operatorname{vec}((L_{P,k}^{(\ell-1)})^\top R_k L_{P,k}^{(\ell-1)} + Q_k) \\ &- 2\Big[(\mathcal{I}_{x\alpha,k}^t)^\top(I_{n_k}\otimes R_k) \\ &+ (\mathcal{I}_{xx,k}^t)^\top(I_{n_k}\otimes(L_{P,k}^{(\ell-1)})^\top R_k)\Big]\operatorname{vec}(K_{P,k}^{(\ell)})\end{aligned} \tag{16a}$$

$$\begin{aligned}0 &= (\delta_{\Omega,k}^t)^\top\hat{\Omega}^{(\ell)} + \delta_\rho^t\theta_\Omega^{(\ell)} \\ &+ (\mathcal{I}_{XX}^t)^\top\operatorname{vec}((\mathcal{L}_\Omega^{(\ell-1)})^\top\mathcal{R}\mathcal{L}_\Omega^{(\ell-1)} + \mathcal{Q}(I_N-\mathcal{H})) \\ &- 2\Big[(\mathcal{I}_{X\alpha}^t)^\top(I_N\otimes\mathcal{R}) \\ &+ (\mathcal{I}_{XX}^t)^\top(I_N\otimes(\mathcal{L}_\Omega^{(\ell-1)})^\top\mathcal{R})\Big]\operatorname{vec}(\mathcal{L}_\Omega^{(\ell)})\end{aligned} \tag{16b}$$

where the terms for the equation for (16a) are

$$\hat{P}_k^{(\ell)} = \big[P_{k,11}^{(\ell)}, 2P_{k,12}^{(\ell)}, \ldots, 2P_{k,1n_k}^{(\ell)}, P_{k,22}^{(\ell)}, \ldots, P_{k,n_kn_k}^{(\ell)}\big]^\top \in \mathbb{R}^{\frac{n_k(n_k+1)}{2}}$$

$$\hat{x}_{k,1} = \big[x_1^2, 2x_1x_2, \ldots, 2x_1x_{n_k}, x_2^2, \ldots, x_{n_k}^2\big]^\top \in \mathbb{R}^{\frac{n_k(n_k+1)}{2}}, \text{ where } x_j = x_{k,1,j}$$

$$\delta_{P,k}^t = \mathbb{E}\big[e^{-\rho(t+\Delta t)}\hat{x}_{k,1}(t+\Delta t) - e^{-\rho t}\hat{x}_{k,1}(t)\big] \in \mathbb{R}^{\frac{n_k(n_k+1)}{2}}$$

$$\mathcal{I}_{xx,k}^t = \mathbb{E}\left[\int_t^{t+\Delta t} e^{-\rho\tau}(x_{k,1}\otimes x_{k,1})d\tau\right] \in \mathbb{R}^{n_k^2}$$

$$\mathcal{I}_{x\alpha,k}^t = \mathbb{E}\left[\int_t^{t+\Delta t} e^{-\rho\tau}(x_{k,1}\otimes \alpha_{k,1})d\tau\right] \in \mathbb{R}^{n_km_k}$$

$$\delta_\rho^t = e^{-\rho(t+\Delta t)} - e^{-\rho t} \in \mathbb{R}$$

$$\theta_{P,k}^{(\ell)} = \frac{1}{\rho}\operatorname{Tr}(D_kD_k^\top P_k^{(\ell)}) \in \mathbb{R}$$

and the terms for (16b) are

$$\hat{\Omega}^{(\ell)} = \big[\Omega_{11}^{(\ell)}, 2\Omega_{12}^{(\ell)}, \ldots, 2\Omega_{1N}^{(\ell)}, \Omega_{22}^{(\ell)}, \ldots, \Omega_{NN}^{(\ell)}\big]^\top \in \mathbb{R}^{\frac{N(N+1)}{2}}$$

$$\hat{\mathcal{X}} = \big[\mathcal{X}_1^2, 2\mathcal{X}_1\mathcal{X}_2, \ldots, 2\mathcal{X}_1\mathcal{X}_N, \mathcal{X}_2^2, \ldots, \mathcal{X}_N^2\big]^\top \in \mathbb{R}^{\frac{N(N+1)}{2}}$$

$$\delta_\Omega^t = \mathbb{E}\big[e^{-\rho(t+\Delta t)}\hat{\mathcal{X}}(t+\Delta t) - e^{-\rho t}\hat{\mathcal{X}}(t)\big] \in \mathbb{R}^{\frac{N(N+1)}{2}}$$

$$\mathcal{I}_{XX}^t = \mathbb{E}\left[\int_t^{t+\Delta t} e^{-\rho\tau}(\mathcal{X}\otimes\mathcal{X})d\tau\right] \in \mathbb{R}^{N^2}$$

$$\mathcal{I}_{X\alpha}^t = \mathbb{E}\left[\int_t^{t+\Delta t} e^{-\rho\tau}(\mathcal{X}\otimes\alpha)d\tau\right] \in \mathbb{R}^{NM}$$

$$\delta_\rho^t = e^{-\rho(t+\Delta t)} - e^{-\rho t} \in \mathbb{R}$$

$$\theta_{\Omega,k}^{(\ell)} = \tfrac{1}{\rho}\operatorname{Tr}(\mathcal{D}\mathcal{D}^\top\Omega^{(\ell)}) \in \mathbb{R}.$$

Using $l \in \mathbb{N}$ time steps associated with real-time data, the following matrices are defined

$$\Delta_{1k} = [\delta_{P,k}^{t_1}, \ldots, \delta_{P,k}^{t_l}]^\top \in \mathbb{R}^{l\times\frac{n_k(n_k+1)}{2}}$$

$$\Delta_1 = [\delta_\Omega^{t_1}, \ldots, \delta_\Omega^{t_l}]^\top \in \mathbb{R}^{l\times\frac{N(N+1)}{2}}$$

$$\begin{aligned}\Delta_{2k} &= -2[\mathcal{I}_{xx,k}^{t_1}, \ldots, \mathcal{I}_{xx,k}^{t_l}]^\top(I_{n_k}\otimes(L_{P,k}^{(k-1)})^\top R_k) \\ &\quad - 2[\mathcal{I}_{x\alpha,k}^{t_1}, \ldots, \mathcal{I}_{x\alpha,k}^{t_l}]^\top(I_{n_k}\otimes R_k) \in \mathbb{R}^{l\times m_kn_k}\end{aligned}$$

$$\begin{aligned}\Delta_3 &= -2[\mathcal{I}_{XX}^{t_1}, \ldots, \mathcal{I}_{XX}^{t_l}]^\top(I_N\otimes(\mathcal{L}_\Omega^{(\ell-1)})^\top\mathcal{R}) \\ &\quad - 2[\mathcal{I}_{X\alpha}^{t_1}, \ldots, \mathcal{I}_{X\alpha}^{t_l}]^\top(I_N\otimes\mathcal{R}) \in \mathbb{R}^{l\times NM}\end{aligned}$$

$$\begin{aligned}\Delta_{4k} &= -[\mathcal{I}_{xx,k}^{t_1}, \ldots, \mathcal{I}_{xx,k}^{t_l}]^\top \\ &\quad \cdot \operatorname{vec}((L_{P,k}^{(\ell-1)})^\top R_k L_{P,k}^{(\ell-1)} + Q_k) \in \mathbb{R}^l\end{aligned}$$

$$\begin{aligned}\Delta_5 &= -[\mathcal{I}_{XX}^{t_1}, \ldots, \mathcal{I}_{XX}^{t_l}]^\top \\ &\quad \cdot \operatorname{vec}((\mathcal{L}_\Omega^{(\ell-1)})^\top\mathcal{R}\mathcal{L}_\Omega^{(\ell-1)} + \mathcal{Q}(I_N-\mathcal{H})) \in \mathbb{R}^l\end{aligned}$$

$$\Delta_6 = [\delta_\rho^{t_1}, \ldots, \delta_\rho^{t_l}]^\top \in \mathbb{R}^l$$

and equations (16), using the data from l time steps, can be expressed as

$$0 = \Delta_{1k}\hat{P}_k^{(\ell)} + \Delta_{2k}\operatorname{vec}(L_{P,k}^{(\ell)}) + \Delta_6\theta_{P,k}^{(\ell)} - \Delta_{4k} \tag{17a}$$

$$0 = \Delta_1\hat{\Omega}_k^{(\ell)} + \Delta_3\operatorname{vec}(\mathcal{L}_\Omega^{(\ell)}) + \Delta_6\theta_\Omega^{(\ell)} - \Delta_5. \tag{17b}$$

Equation (17) can then be reformulated as the matrix form

$$\underbrace{[\Delta_{1k}\ \Delta_{2k}\ \Delta_6]}_{\Xi_{1,k}}\begin{bmatrix}\hat{P}_k^{(\ell)} \\ \operatorname{vec}(L_{P,k}^{(\ell)}) \\ \theta_{P,k}^{(\ell)}\end{bmatrix} = \Delta_{4k} \tag{18a}$$

$$\underbrace{[\Delta_1\ \Delta_3\ \Delta_6]}_{\Xi_2}\begin{bmatrix}\hat{\Omega}^{(\ell)} \\ \operatorname{vec}(\mathcal{L}_\Omega^{(\ell)}) \\ \theta_\Omega^{(\ell)}\end{bmatrix} = \Delta_5. \tag{18b}$$

We introduce the following assumption regarding the requirement for the trajectory data.

Assumption 4. There exists an integer $L > 0$ such that for $l \geq L$, the matrices

$$\begin{bmatrix}\mathcal{I}_{xx,k}^{t_1} & \mathcal{I}_{xx,k}^{t_2} & \cdots & \mathcal{I}_{xx,k}^{t_l} \\ \mathcal{I}_{x\alpha,k}^{t_1} & \mathcal{I}_{x\alpha,k}^{t_2} & \cdots & \mathcal{I}_{x\alpha,k}^{t_l} \\ \delta_\rho^{t_1} & \delta_\rho^{t_2} & \cdots & \delta_\rho^{t_l}\end{bmatrix}$$

are of rank $\frac{n_k(n_k+1)}{2} + m_kn_k + 1$ for all class k and the matrix

$$\begin{bmatrix}\mathcal{I}_{XX}^{t_1} & \mathcal{I}_{XX}^{t_2} & \cdots & \mathcal{I}_{XX}^{t_l} \\ \mathcal{I}_{X\alpha,k}^{t_1} & \mathcal{I}_{X\alpha,k}^{t_2} & \cdots & \mathcal{I}_{X\alpha,k}^{t_l} \\ \delta_\rho^{t_1} & \delta_\rho^{t_2} & \cdots & \delta_\rho^{t_l}\end{bmatrix}$$

is of rank $\frac{N(N+1)}{2} + NM + 1$.

Remark 3. Assumption 4 ensures that (18) have unique solutions. In practice, it can be satisfied by injecting exploration noise into the control input during the learning phase, such as Gaussian noise, or a sum of sinusoids, as shown in (Jiang and Jiang, 2012) and (Xu et al., 2023).

The multi-class LQG-MFG integral reinforcement learning can thus be formulated in Algorithm 1.

Remark 4. (Trajectory Data Required by the Algorithm). The data-driven algorithm requires multiple trajectories of the states and control inputs of the representative agent $a_{k,1}$ of class k for all $k \in \{1, \cdots, \mathcal{K}\}$. These trajectories

Algorithm 1 Multi-class Mean Field Games Integral Reinforcement Learning

Initialization:
Choose $\mathcal{L}_\Omega^{(0)}$ s.t. $\mathcal{A} - \frac{1}{2}\rho I_N - \mathcal{B}\mathcal{L}_\Omega^{(0)}$ is Hurwitz and s.t. the associated class-specific $L_{P,k}^{(0)}$ ensure $A_k - \frac{1}{2}\rho I_{n_k} - B_k L_{P,k}^{(0)}$ are Hurwitz for all k.
Select representative agents $a_{k,1}$ for all class k, set threshold ε, and set iteration counter $\ell = 1$.

Data Collection:
Apply $\mathcal{U}(t) = -\mathcal{L}_\Omega^{(0)}\mathcal{X}(t) + l(t)$ to global system composed of selected agents.
Compute data vectors $\delta_{P,k}, \mathcal{I}_{xx,k}, \mathcal{I}_{x\alpha,k}, \mathcal{I}_{XX}, \mathcal{I}_{X\alpha}, \delta_t$ for t_j, $j \in \{1, \dots, l\}$ until rank$(\Xi_{1,k})$ and rank(Ξ_2) satisfy persistence of excitation. (Assumption 4)

Policy Iteration:
repeat
Solve for class-specific parameters:

$$\begin{bmatrix} \hat{P}_k^{(\ell)} \\ \text{vec}(L_{P,k}^{(\ell)}) \\ \theta_{P,k}^{(\ell)} \end{bmatrix} = (\Xi_{1,k}^\top \Xi_{1,k})^{-1}\Xi_{1,k}^\top \Delta_{4k}, \quad \text{for all } k. \quad (19)$$

Solve for global parameters:

$$\begin{bmatrix} \hat{\Omega}^{(\ell)} \\ \text{vec}(\mathcal{L}_\Omega^{(\ell)}) \\ \theta_\Omega^{(\ell)} \end{bmatrix} = (\Xi_2^\top \Xi_2)^{-1}\Xi_2^\top \Delta_5. \quad (20)$$

Update $\ell \leftarrow \ell + 1$
until $\|P_k^{(\ell)} - P_k^{(\ell-1)}\| \le \varepsilon$ for all k **&** $\|\Omega^{(\ell)} - \Omega^{(\ell-1)}\| \le \varepsilon$

are defined for the interval $[t_1, t_l + \Delta t]$ and satisfy Assumption 4. Using a large finite number of trajectories, the mean of data integrals in $\delta_{P,k}, \mathcal{I}_{xx,k}, \mathcal{I}_{x\alpha,k}, \delta_\Omega, \mathcal{I}_{XX}$ and $\mathcal{I}_{x\alpha}$ can approximate their expectations.

4. NUMERICAL EXAMPLE

A numerical simulation for learning multi-class LQG-MFG gain matrices using Algorithm 1 is carried out. The parameters of the problems are given in Table 1.

Table 1. Class-Specific Parameters

Parameter	Class 1	Class 2	Class 3
A_k	$\begin{bmatrix} 0 & 10 \\ -10 & -3 \end{bmatrix}$	$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -2 & -3 & -5 \end{bmatrix}$	$\begin{bmatrix} 0 & -4 \\ 3 & -6 \end{bmatrix}$
B_k	$\begin{bmatrix} 1.0 \\ 1.0 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 0 & 1.0 \\ 1.0 & 0.5 \end{bmatrix}$	$\begin{bmatrix} 0.8 \\ 3.0 \end{bmatrix}$
D_k	$0.1I_2$	$0.1I_3$	$0.1I_2$
Q_k	diag(20, 10)	diag(10, 15, 20)	diag(30, 20)
R_k	0.8	diag(0.5, 0.7)	0.6
ρ_k	0.1	0.1	0.1

The global interaction matrix $\mathcal{H}$ is chosen to ensure Assumptions 2 and 3 are respected. By eigen decomposition of matrix $\mathcal{Q} = U\Lambda U^\top$, where U is the matrix where the columns are the eigenvectors of $\mathcal{Q}$, and Λ is a diagonal matrix where the diagonal elements are the corresponding eigenvalues, then

$$\mathcal{H} = (U\Lambda^{-\frac{1}{2}}U^\top)\tilde{\mathcal{H}}(U\Lambda^{\frac{1}{2}}U^\top). \quad (21)$$

The normalized interaction matrix $\tilde{\mathcal{H}}$ is constructed as a block matrix

$$\tilde{\mathcal{H}} = \frac{1}{\lambda_{\max}}\begin{bmatrix} H_{11} & H_{12} & H_{13} \\ H_{21} & H_{22} & H_{23} \\ H_{31} & H_{32} & H_{33} \end{bmatrix} = \frac{1}{\lambda_{\max}}\tilde{H} \quad (22)$$

where $\lambda_{\max}$ is the largest eigenvalue of $\tilde{H}$. Diagonal blocks are defined such that $H_{km} = \mathbf{0}_{n_k \times n_m}$ if $k = m$, indicating no self-coupling within each class. The remaining blocks, representing inter-class coupling between the first two states of each class, are defined as

$$H_{12} = \frac{1}{2}\left[I_2 \ \mathbf{0}_{2\times1}\right], \ H_{21} = H_{12}^\top,$$
$$H_{32} = \frac{1}{2}\left[I_2 \ \mathbf{0}_{2\times1}\right], \ H_{23} = H_{32}^\top,$$
$$H_{13} = H_{31} = \frac{1}{2}I_2.$$

Under Assumptions 1-3, the AREs (9) and (8b) admit unique positive-definite solutions P_k and Ω, respectively.

To run the model-free Algorithm 1, a representative agent $a_{k,1}$ is selected for all k classes. The initial states and initial gains are the same for each class

$$L_k^{(0)} = \mathbf{0}_{m_k \times n_k} \qquad x_{k,1}(0) = \mathbf{1}_{n_k \times 1}$$

and $\mathcal{L}_\Omega^{(0)} = \text{diag}(L_1^{(0)}, \dots, L_k^{(0)})$.

Using an exploration noise $\mathcal{U}(t) = \alpha(t) = -\mathcal{L}_\Omega^{(0)}\mathcal{X}(t) + l(t)$, where each channel $l_{k,1,i}(t)$ for each class $k \in \{1, \dots, \mathcal{K}\}$ and each control input channel $i \in \{1, \dots, m_k\}$ is composed by a sum of sinusoids, such that $l_{k,1,i}(t) = \sum_{j=1}^{500} A_e \sin(\omega_j t)$ where ω_j is selected uniformly randomly in $[-100, 100]$ and independently across agents and channels, and $A_e = 25$. The global system composed of agents $a_{k,1}$ runs 100 times for 20 s. Algorithm 1 is then carried out, with $\varepsilon = 10^{-9}$. The results are presented in Tables 2 and 3. The convergence of the matrices $(L_{P,k}, P_k)$ and $(\mathcal{L}_\Omega, \Omega)$ to their ground truth values is plotted in Fig. 1.

Table 2. Comparison of Learned Parameters vs. Ground Truth for Individual Classes $(L_{P,k}, P_k)$

Parameter	Learned (IRL)	Ground Truth
$L_{P,1}$	$[3.9969, 2.7750]$	$[3.9608, 2.7376]$
P_1	$\begin{bmatrix} 2.8624 & 0.3584 \\ 0.3584 & 1.8810 \end{bmatrix}$	$\begin{bmatrix} 2.8102 & 0.3584 \\ 0.3584 & 1.8316 \end{bmatrix}$
$L_{P,2}$	$\begin{bmatrix} -0.4359 & -0.6928 & 2.9514 \\ 3.6698 & 5.5736 & 0.5743 \end{bmatrix}$	$\begin{bmatrix} -0.4233 & -0.6794 & 2.9314 \\ 3.6677 & 5.5738 & 0.5616 \end{bmatrix}$
P_2	$\begin{bmatrix} 13.4680 & 2.6812 & -0.2104 \\ 2.6812 & 4.0818 & -0.3260 \\ -0.2104 & -0.3260 & 1.5024 \end{bmatrix}$	$\begin{bmatrix} 13.4070 & 2.6732 & -0.2117 \\ 2.6732 & 4.0715 & -0.3397 \\ -0.2117 & -0.3397 & 1.4657 \end{bmatrix}$
$L_{P,3}$	$[-3.6928, 6.0612]$	$[-3.6908, 6.0446]$
P_3	$\begin{bmatrix} 10.2930 & -3.4970 \\ -3.4970 & 2.1593 \end{bmatrix}$	$\begin{bmatrix} 10.2340 & -3.4672 \\ -3.4672 & 2.1335 \end{bmatrix}$

By iteration 11, all matrices have converged to the fixed threshold ε, and the Frobenius norm error of the learned matrices to their ground truth obtained using MATLAB's `care()` function is small. Individual elements, a sample of

Table 3. Comparison of Selected Parameters for $(\mathcal{L}_\Omega, \Omega)$

Parameter	Learned (IRL)	Ground Truth	Error
$\Omega_{1,1}$	2.7679	2.7111	0.0568
$2\Omega_{1,2}$	0.6943	0.7514	0.0571
$2\Omega_{1,3}$	-1.2242	-1.2599	0.0357
$\Omega_{6,6}$	9.6314	9.5677	0.0637
$\mathcal{L}_{\Omega,1,1}$	3.9079	3.8585	0.0494
$\mathcal{L}_{\Omega,1,2}$	2.7259	2.7032	0.0227
$\mathcal{L}_{\Omega,2,3}$	-0.5188	-0.4515	0.0673
$\mathcal{L}_{\Omega,3,6}$	-1.3534	-1.3572	0.0038

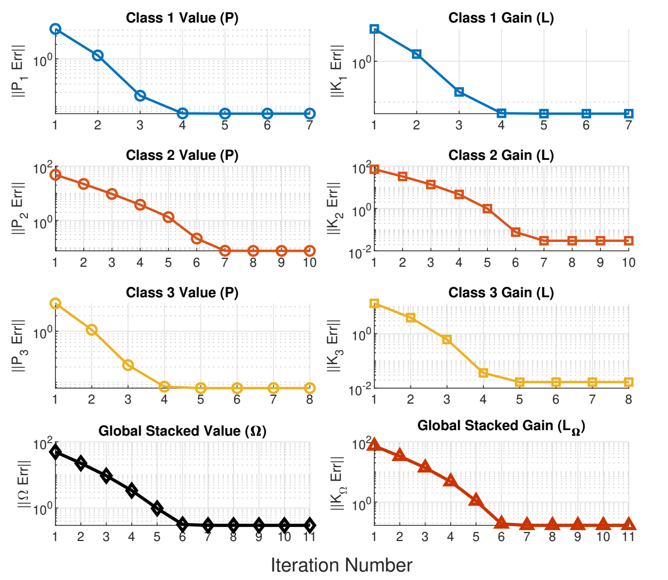

Fig. 1. Convergence analysis of the IRL algorithm for local class-level results and global system-level results

which is presented in Tables 2 and 3, are also close to their ground truth value.

Similar to the methodology of Xu et al. (2023), the mean field trajectories are computed offline using the learned control gains. A representative augmented state $\mathcal{X} = [x_{1,1}^\top, x_{2,1}^\top, x_{3,1}^\top]^\top \in \mathbb{R}^N$ is simulated $N_s = 100$ times using the learned global feedback law $\alpha = -\mathcal{L}_\Omega \mathcal{X}$ with initial states $\mathcal{X}(0) = \mathbf{1}_{7\times 1}$. The empirical mean field is then constructed as $\mathcal{X}^{(N_s)}(t) = \frac{1}{N_s}\sum_{j=1}^{N_s} \mathcal{X}_j(t)$, where $\mathcal{X}_j(t)$ is the trajectory from the j-th run. By the law of large numbers $\mathcal{X}^{(N_s)}(t) \to \bar{\mathcal{X}}(t)$ as $N_s \to \infty$. The class-specific mean fields $\bar{x}_k(t)$ are approximated as the corresponding subvectors of $\mathcal{X}^{(N_s)}(t)$ and the global mean field corresponds to their average.

For validation, a finite population of 50 agents per class is simulated, with all initial states located uniformly in [0.5, 1.5]. The agents are simulated using control input

$$u_{k,i}(t) = -L_{P,k} x_{k,i}(t) - L_{\Pi,k} \mathcal{X}^{(50)}(t), \quad (23)$$

where $L_{\Pi,k} \in \mathcal{R}^{m_k \times N}$ is the k-th block row of the matrix $\mathcal{L}_\Pi = \mathcal{L}_\Omega - \mathcal{L}_\mathcal{P} \in \mathbb{R}^{M\times N}$, with $\mathcal{L}_\mathcal{P} = \mathrm{diag}(L_{P,1}, \ldots, L_{P,\mathcal{K}}) \in \mathbb{R}^{M\times N}$.

The results comparing trajectories using learned gains to those using gains computed by `care()` serving as ground truth, are presented in Fig. 2. Class-specific plots show the resulting empirical class mean field $x_k^{(N_s)}$, and the shaded area shows the spread of the agents around the mean (± 2 standard deviations).

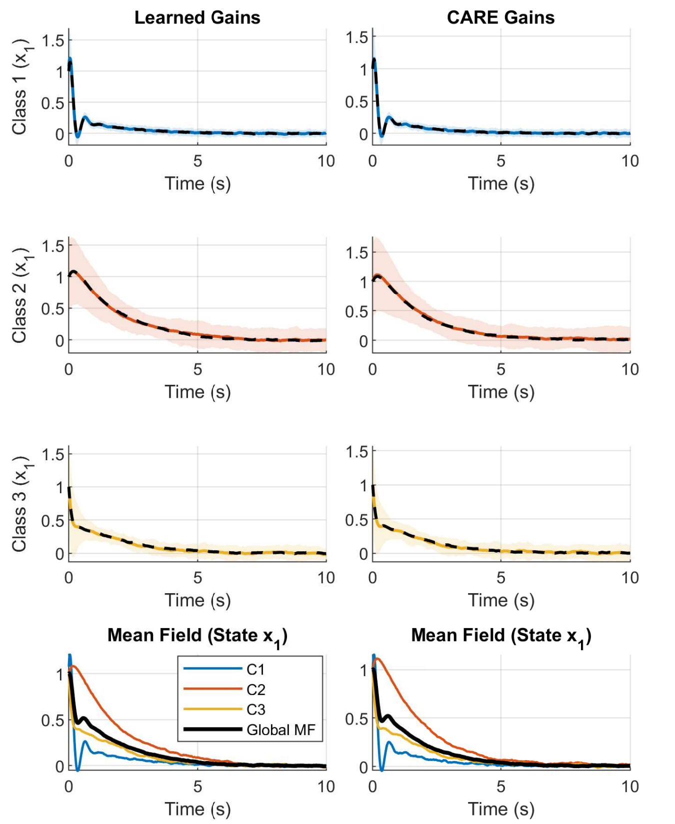

Fig. 2. Mean field trajectories under data-driven learned controls (left) and standard `care()` gains computed assuming known system and cost matrices (right)

5. CONCLUSION

This paper established a data-driven algorithm for computing strategies for continuous-time infinite horizon LQG MFGs with heterogeneous network-coupled populations that contain completely unknown dynamics. During the data collection and learning phase, a global system is formed for a generic agent in each class, and an exploration noise is applied to an initially stabilizing control to ensure persistency of excitation. Under conditions on the persistency of excitation and on the existence of unique stabilizing solution for the corresponding AREs, the algorithm converges to the MFG strategies that depend on the classes and network couplings.

Future investigations should extend the solutions to the cases with network-coupled dynamics, directed network couplings, finite time horizons, and nonlinear agent dynamics, and apply the algorithm in applications including renewable energy systems with user populations.

REFERENCES

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

Carmona, R., Laurière, M., and Tan, Z. (2019). Linear-quadratic mean-field reinforcement learning: convergence of policy gradient methods. *arXiv preprint arXiv:1910.04295*.

Fu, Z., Yang, Z., Chen, Y., and Wang, Z. (2019). Actor-critic provably finds Nash equilibria of linear-quadratic mean-field games. *arXiv preprint arXiv:1910.07498*.

Gao, S., Caines, P.E., and Huang, M. (2023). LQG graphon mean field games: Analysis via graphon-invariant subspaces. *IEEE Transactions on Automatic Control*, 68(12), 7482–7497.

Guo, X., Hu, A., Xu, R., and Zhang, J. (2019). Learning mean-field games. *Advances in neural information processing systems*, 32.

Huang, M., Caines, P., and Malhamé, R. (2007). Large-population cost-coupled LQG problems with nonuniform agents: Individual-mass behavior and decentralized ε-Nash equilibria. *IEEE Transactions on Automatic Control*, 52(9), 1560–1571.

Huang, M., Malhamé, R., and Caines, P. (2006). Large population stochastic dynamic games: Closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle. *Communications in Information & Systems*, 6(3), 221–252.

Huang, M. (2010). Large-population LQG games involving a major player: the Nash certainty equivalence principle. *SIAM Journal on Control and Optimization*, 48(5), 3318–3353.

Huang, M., Caines, P.E., and Malhamé, R.P. (2010). The NCE (mean field) principle with locality dependent cost interactions. *IEEE Transactions on Automatic Control*, 55(12), 2799–2805.

Huang, M. and Zhou, M. (2020). Linear quadratic mean field games: Asymptotic solvability and relation to the fixed point approach. *IEEE Transactions on Automatic Control*, 65(4), 1397–1412.

Jiang, Y. and Jiang, Z.P. (2012). Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics. *Automatica*, 48(10), 2699–2704.

Kizilkale, A.C. and Caines, P.E. (2012). Mean field stochastic adaptive control. *IEEE Transactions on Automatic Control*, 58(4), 905–920.

Kleinman, D. (1968). On an iterative technique for riccati equation computations. *IEEE Transactions on Automatic Control*, 13(1), 114–115.

Lasry, J.M. and Lions, P.L. (2006). Jeux à champ moyen. I-Le cas stationnaire. *Comptes Rendus Mathématique*, 343(9), 619–625.

Lasry, J.M. and Lions, P.L. (2007). Mean field games. *Japanese Journal of Mathematics*, 2(1), 229–260.

Li, N., Li, X., Peng, J., and Xu, Z.Q. (2022). Stochastic linear quadratic optimal control problem: A reinforcement learning method. *IEEE Transactions on Automatic Control*, 67(9), 5009–5016.

Li, X., Wang, G., Wang, Y., Xiong, J., and Zhang, H. (2025). Two system transformation data-driven algorithms for linear quadratic mean-field games. *European Journal of Control*, 83, 101226.

Modares, H., Nageshrao, S.P., Lopes, G.A.D., Babuška, R., and Lewis, F.L. (2016). Optimal model-free output synchronization of heterogeneous systems using off-policy reinforcement learning. *Automatica*, 71, 334–341.

Subramanian, J. and Mahajan, A. (2019). Reinforcement learning in stationary mean-field games. In *Proceedings of the 18th international conference on autonomous agents and multiagent systems*, 251–259.

Vamvoudakis, K.G. and Lewis, F.L. (2011). Multi-player non-zero-sum games: Online adaptive learning solution of coupled Hamilton–Jacobi equations. *Automatica*, 47(8), 1556–1569.

Vrabie, D. and Lewis, F. (2009). Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems. *Neural Networks*, 22(3), 237–246.

Wonham, W.M. (1968). On a matrix Riccati equation of stochastic control. *SIAM Journal on Control*, 6(4), 681–697.

Xu, Z., Chen, J., Wang, B.C., and Shen, T. (2025a). Data-driven mean field equilibrium computation in large-population LQG games. *IEEE Transactions on Control of Network Systems*, 12(4), 2713–2725.

Xu, Z., Shen, T., and Huang, M. (2023). Model-free policy iteration approach to NCE-based strategy design for linear quadratic Gaussian games. *Automatica*, 155, 111162.

Xu, Z., Wang, B.C., and Shen, T. (2025b). Mean field LQG social optimization: A reinforcement learning approach. *Automatica*, 172, 111924.

Zaman, M.A.u., Miehling, E., and Başar, T. (2023). Reinforcement learning for non-stationary discrete-time linear–quadratic mean-field games in multiple populations. *Dynamic Games and Applications*, 13(1), 118–164.

Zaman, M.A.u., Zhang, K., Miehling, E., and Başar, T. (2020). Approximate equilibrium computation for discrete-time linear-quadratic mean-field games. In *Proceedings of the American Control Conference (ACC)*, 333–339.

Appendix A. PROOF OF PROPOSITION 1

Let the population size in all classes go to infinity. Then, in the limit problem, an individual agent is negligible in the mean field of every class. Thus, the mean fields of all classes can be treated as deterministic trajectories in the limit problem. For an agent $\alpha_{k,i}$, consider the augmented state $z_{k,i} = [x_{k,i}^\top, \bar{\mathcal{X}}^\top]^\top$ where $\bar{\mathcal{X}} = [\bar{x}_1^\top, \ldots, \bar{x}_k^\top]^\top$. Then the dynamics of the augmented states $z_{k,i}$ satisfy

$$dz_{k,i} = \left(\tilde{A}_k z_{k,i} + \tilde{B}_k u_{k,i}\right) dt + \tilde{D}_k dw_{k,i} \tag{A.1}$$

where

$$\tilde{A}_k = \begin{bmatrix} A_k & \mathbf{0}_{n_k \times N} \\ \mathbf{0}_{N \times n_k} & \bar{G} \end{bmatrix}, \tilde{B}_k = \begin{bmatrix} B_k \\ \mathbf{0}_{N \times m_k} \end{bmatrix}, \tilde{D}_k = \begin{bmatrix} D_k \\ \mathbf{0}_{N \times d_k} \end{bmatrix},$$

$\bar{G}$ is the drift of the mean field, assumed to be known. The linear mean-field dynamics with the drift $\bar{G}$ will be identified later from the consistency condition required by the fixed-point approach for MFGs (Huang et al., 2007).

The problem can then be framed as a standard LQR problem with cost functional

$$J_k = \mathbb{E} \int_0^\infty e^{-\rho t} \left(\tilde{x}_{k,i}^\top Q_k \tilde{x}_{k,i} + u_{k,i}^\top R_k u_{k,i}\right) dt. \tag{A.2}$$

where $\tilde{x}_{k,i} = x_{k,i} - \sum_{m=1}^{\mathcal{K}} H_{km} \bar{x}_m$ and with a slight abuse of notation $\bar{x}_m = \lim_{c_m \to \infty} \frac{1}{c_m} \sum_{i=1}^{c_m} x_{m,i}$. Then, assuming a standard value ansatz for an LQR problem, the class-specific value ansatz is

$$V_k(z_{k,i}) = \begin{bmatrix} x_{k,i} \\ \bar{\mathcal{X}} \end{bmatrix}^\top \begin{bmatrix} P_{11,k} & P_{12,k} \\ P_{21,k} & P_{22,k} \end{bmatrix} \begin{bmatrix} x_{k,i} \\ \bar{\mathcal{X}} \end{bmatrix} \triangleq z_{k,i}^\top \tilde{P}_k z_{k,i}$$

where $P_{11,k} \in \mathbb{R}^{n_k \times n_k}, P_{12,k} = P_{21,k}^\top \in \mathbb{R}^{n_k \times N}$, and $P_{22,k} \in \mathbb{R}^{N \times N}$. The infinite-time Hamilton-Jacobi-Bellman (HJB) equation is thus

$$\rho V_k(z_{k,i}) = \inf_{u_{k,i}} \{L(\tilde{x}_{k,i}, u_{k,i}) + \nabla_{z_{k,i}} V_k(z_{k,i})^\top f(z_{k,i}, u_{k,i})\} + \mathrm{Tr}(\frac{1}{2}\tilde{D}_k\tilde{D}_k^\top V_k(z_{k,i}))$$

where

$$L(\tilde{x}_{k,i}, u_{k,i}) = \tilde{x}_{k,i}^\top Q_k \tilde{x}_{k,i} + u_{k,i}^\top R_k u_{k,i}$$
$$f(z_{k,i}, u_{k,i}) = \tilde{A}_k z_{k,i} + \tilde{B}_k u_{k,i}.$$

Expanding and replacing the terms in the HJB, it becomes

$$\rho z_{k,i}^\top \tilde{P}_k z_{k,i} = \inf_{u(\cdot)} \{z_{k,i}^\top \tilde{Q}_k z_{k,i} + u_{k,i} R_k u_{k,i} + 2 z_{k,i}^\top \tilde{P}_k \tilde{A}_k z_{k,i} + 2 z_{k,i}^\top \tilde{P}_k \tilde{B}_k u_{k,i}\} + \mathrm{Tr}(\frac{1}{2}\tilde{D}_k\tilde{D}_k^\top V_k(z_{k,i}))$$

where

$$\tilde{Q}_k = \begin{bmatrix} Q_k & -Q_k H_k \\ -H_k^\top Q_k & H_k^\top Q_k H_k \end{bmatrix}$$
$$H_k = [H_{k1}, \dots, H_{k\mathcal{K}}] \in \mathbb{R}^{n_k \times N}.$$

Setting the gradient with respect to u to zero yields the optimal control law (i.e. the best response for the limit MFG problem)

$$u_{k,i}^* = -R_k^{-1} \tilde{B}_k^\top \tilde{P}_k z_{k,i} \quad \text{(A.3)}$$
$$= -R_k^{-1} \tilde{B}_k^\top (P_{11,k} x_{k,i} + P_{12,k} \bar{\mathcal{X}}). \quad \text{(A.4)}$$

The dynamics of the class-specific mean-field associated with class k are obtained by

$$\dot{\bar{x}}_k = A_k \bar{x}_k - B_k R_k^{-1} B_k^\top (P_{11,k} \bar{x}_k + P_{12,k} \bar{\mathcal{X}}) \quad \text{(A.5)}$$

and the dynamics of $\bar{\mathcal{X}}$ are given by

$$\dot{\bar{\mathcal{X}}} = (\mathcal{A} - \mathcal{B}\mathcal{R}^{-1}\mathcal{B}^\top(\mathcal{P}_{11} + \mathcal{P}_{12}))\bar{\mathcal{X}} \quad \text{(A.6)}$$

where $\mathcal{A}, \mathcal{B}, \mathcal{R}$ and $\mathcal{P}_{11}$ are block diagonal matrices composed of matrices A_k, B_k, R_k and $P_{11,k}$ for each class k, and $\mathcal{P}_{12} = [P_{12,1}^\top, \dots, P_{12,k}^\top]^\top$. Thus, the consistency condition for the mean field in the fixed-point approach (Huang et al., 2007) is then equivalently given by

$$\bar{G} = \mathcal{A} - \mathcal{B}\mathcal{R}^{-1}\mathcal{B}^\top(\mathcal{P}_{11} + \mathcal{P}_{12}).$$

Plugging in the optimal control back into the HJB and then matching the coefficients with the value function ansatz yield the infinite-time discounted cost Riccati equation

$$\rho \tilde{P}_k = \tilde{Q}_k - \tilde{P}_k \tilde{B}_k R_k^{-1} \tilde{B}_k^\top \tilde{P}_k + \tilde{A}_k^\top \tilde{P}_k + \tilde{P}_k \tilde{A}_k. \quad \text{(A.7)}$$

Developing the terms, the following equations for $P_{11,k}$ and $P_{12,k}$ are obtained

$$\rho P_{11,k} = Q_k - P_{11,k} B_k R_k^{-1} B_k^\top P_{11,k} + A_k^\top P_{11,k} + P_{11,k} A_k$$

$$\rho P_{12,k} = -Q_k H_k - P_{11,k} B_k R_k^{-1} B_k^\top P_{12,k} + A_k^\top P_{12,k} + P_{12,k}(\mathcal{A} - \mathcal{B}\mathcal{R}^{-1}\mathcal{B}^\top(\mathcal{P}_{11} + \mathcal{P}_{12}))$$

$$\rho P_{22,k} = H_k^\top Q_k H_k - P_{12,k}^\top B_k R_k^{-1} B_k^\top P_{12,k} + \bar{G}^\top P_{22,k} + P_{22,k} \bar{G}.$$

Stacking the equations for $P_{11,k}$ for all $\mathcal{K}$ classes yields

$$\rho \mathcal{P}_{11} = \mathcal{Q} + \mathcal{P}_{11}\mathcal{A} + \mathcal{A}^\top \mathcal{P}_{11} - \mathcal{P}_{11}\mathcal{B}\mathcal{R}^{-1}\mathcal{B}^\top \mathcal{P}_{11} \quad \text{(A.8)}$$

and stacking the equations for $P_{12,k}$ yields

$$\rho \mathcal{P}_{12} = -\mathcal{Q}\mathcal{H} + (\mathcal{A}^\top - \mathcal{P}_{11}\mathcal{B}\mathcal{R}^{-1}\mathcal{B}^\top)\mathcal{P}_{12} + \mathcal{P}_{12}(\mathcal{A} - \mathcal{B}\mathcal{R}^{-1}\mathcal{B}^\top\mathcal{P}_{11}) - \mathcal{P}_{12}\mathcal{B}\mathcal{R}^{-1}\mathcal{B}^\top\mathcal{P}_{12} \quad \text{(A.9)}$$

which are analogous to the equations in (6), with $\mathcal{P}_{11} = \mathcal{P}$ and $\mathcal{P}_{12} = \Pi$. Summing (A.8) with (A.9) and defining $\Omega = \mathcal{P}_{11} + \mathcal{P}_{12}$, an ARE for Ω is obtained

$$\rho\Omega = \mathcal{Q}(I_N - \mathcal{H}) + \Omega\mathcal{A} + \mathcal{A}^\top\Omega - \Omega\mathcal{B}\mathcal{R}^{-1}\mathcal{B}^\top\Omega. \quad \text{(A.10)}$$

Under Assumptions 1 and 2, the pair $(\tilde{A} - \frac{1}{2}\rho I_N, \tilde{B})$ is stabilizable, and both (A.8) and (A.10), and thus also (A.9), all admit unique stabilizing solutions $\mathcal{P}_{11}, \Omega$ and $\mathcal{P}_{12}$ respectively, following (Huang and Zhou, 2020, Thm. 18) and Lemma 1. Therefore, the MFG strategy is uniquely given by (5)-(7). This completes the proof.

Appendix B. LEMMA USE IN THE PROOF OF PROP. 1

Let $\bar{\mathcal{A}} \triangleq \mathcal{A} - \frac{1}{2}\rho I_N$ and $\mathcal{M} \triangleq \mathcal{B}\mathcal{R}^{-1}\mathcal{B}^\top$. Consider the Hamiltonian matrix associated with the ARE in (6b):

$$H_\Pi = \begin{bmatrix} \bar{\mathcal{A}} - \mathcal{M}\mathcal{P} & -\mathcal{M} \\ \mathcal{Q}\mathcal{H} & -\bar{\mathcal{A}}^\top + \mathcal{P}\mathcal{M}^\top \end{bmatrix}. \quad \text{(B.1)}$$

Lemma 1. Assume $(\bar{\mathcal{A}}, \mathcal{B})$ is stabilizable and the pair $(\bar{\mathcal{A}}, \mathcal{Q})$ is observable. Then the following hold:

(1) H_Ω in (4) is strong (N, N) c-splitting if and only if H_Π is strong (N, N) c-splitting;
(2) the N-dimensional stable invariant subspace of H_Ω in (4) is a graph subspace if and only if the N-dimensional stable invariant subspace of H_Π is a graph subspace.

Proof.

The Hamiltonian matrix in (4) is equivalently given by

$$H_\Omega = \begin{bmatrix} \bar{\mathcal{A}} & -\mathcal{M} \\ -\mathcal{Q}(I_N - \mathcal{H}) & -\bar{\mathcal{A}}^\top \end{bmatrix}. \quad \text{(B.2)}$$

Then

$$\begin{bmatrix} I_N & 0 \\ -\mathcal{P} & I_N \end{bmatrix} H_\Omega \begin{bmatrix} I_N & 0 \\ \mathcal{P} & I_N \end{bmatrix} = \begin{bmatrix} \bar{\mathcal{A}} - \mathcal{M}\mathcal{P} & -\mathcal{M} \\ Z & -\bar{\mathcal{A}}^\top + \mathcal{P}\mathcal{M} \end{bmatrix}, \quad \text{(B.3)}$$

with $Z \triangleq -\mathcal{P}\bar{\mathcal{A}} + \mathcal{P}\mathcal{M}\mathcal{P} - \mathcal{Q}(I_N - \mathcal{H}) - \bar{\mathcal{A}}^\top\mathcal{P}$. The stabilizability of $(\bar{\mathcal{A}}, \mathcal{B})$ and the observability of $(\bar{\mathcal{A}}, \mathcal{Q})$ ensure that (8a) has a unique positive definite solution (Wonham, 1968, Thm. 4.1). Simplifying Z with (8a) yields $Z = \mathcal{Q}\mathcal{H}$. The right hand side of (B.3) becomes H_Π in (B.1). In addtion, we note that

$$\begin{bmatrix} I_N & 0 \\ -\mathcal{P} & I_N \end{bmatrix} \begin{bmatrix} I_N & 0 \\ \mathcal{P} & I_N \end{bmatrix} = I_{2N}. \quad \text{(B.4)}$$

Hence, H_Ω and H_Π are similar matrices. As a consequence, H_Ω and H_Π share the same eigenvalues, and hence H_Ω is strong (N, N) c-splitting if and only if H_Π is strong (N, N) c-splitting. By the definition of stable graph subspace, it is easy to verify that the N-dimensional stable invariant subspace of H_Ω is a graph subspace if and only if the N-dimensional stable invariant subspace of H_Π is a graph subspace.