Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Decentralized Federated Mean-Field Learning\for Competitive Mean-Field Games

Wei Sun

wsun@ou.edu

University of Oklahoma

Theodore B. Trafalis

University of Oklahoma

Research Article

Keywords:

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11528578/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Decentralized Federated Mean-Field Learning for Competitive Mean-Field Games

Wei Sun[1*] and Theodore B. Trafalis[2]

[1*]School of Aerospace and Mechanical Engineering, University of Oklahoma, Norman, 73019, OK, USA.
[2]School of Industrial and Systems Engineering, University of Oklahoma, Norman, 73019, OK, USA.

*Corresponding author(s). E-mail(s): wsun@ou.edu;
Contributing authors: ttrafalis@ou.edu;

Abstract

Mean-field games (MFGs) make large-population strategic interaction tractable by replacing the joint behavior of agents with aggregate population measures. Many MFG solvers assume that agents observe the global mean field or receive it from a central coordinator, which may be impractical when data and computation are distributed across a communication network. We study a decentralized mean-field game (DMFG) between two teams in which agents estimate population behavior from locally visible actions. We propose the Decentralized Federated Mean-Field Learning (DF-MFL), which combines a hybrid team-aware mean-field estimator, local Double-DQN updates, and periodic parameter averaging within team-and-class federated groups, exchanging no raw data. We establish a finite-round bound on within-group parameter disagreement with explicit dependence on the communication topology, synchronization period, and local step sizes, and prove asymptotic parameter consensus under vanishing step sizes. We compare the proposed DF-MFL against Decentralized Mean Field Game Q-learning (DMFG-QL), the vanilla Q-learning algorithm for DMFG, and show improvements in policy while learning much fewer distinct parameters.

Keywords: Decentralized mean-field games, Networked multi-agent systems, Federated learning, Consensus over graphs, Reinforcement learning

1 Introduction

Mean-field games (MFGs) provide a rigorous framework for dynamic games among a large number of interacting agents by replacing the network of pairwise interactions with a macroscopic population measure, the *mean field*. This reduces an otherwise intractable N-player game in the large-population limit to a coupled system comprising a representative agent's optimal-control problem and a population-evolution equation [1, 2]. MFGs have been applied to a wide variety of research fields including economics [3], epidemiology [4], autonomous driving [5], energy markets [6], and crowd dynamics [7].

Many of the MFG solvers are model-based, iterating between a backward value sweep and a forward Kolmogorov equation, whose closed forms exist for the linear-quadratic-Gaussian class [2, 8], while finite-difference [9] and semi-Lagrangian [10] schemes handle nonlinear dynamics. To relax the explicit-model requirement, reinforcement learning (RL) has been applied to learn MFG equilibria from interaction data: mean-field Q-learning replaces each agent's dependence on the full joint action with the empirical mean action of its neighborhood in [11], which is extended to stationary and heterogeneous populations in [12]. These methods assume that each agent either observes the true population distribution or receives it from a central coordinator.

That assumption tends to fail in the networked settings such as connected-vehicle routing, smart-grid load balancing, multi-robot coordination, and distributed resource allocation, where the decision-making units are nodes of a communication graph. In the graph, the data are normally generated and stored locally, and centralized aggregation is precluded by bandwidth or privacy constraints. The *decentralized* MFG (DMFG) [13] is introduced to tackle these issues. In DMFG, agents communicate only with graph neighbors and need to *estimate* the mean field from local observations. Prior decentralized and network-structured MFG research has considered graph-coupled and graphon interactions [14, 15], noisy local observations [16], and approximation or finite-sample guarantees for networked learning [17, 18]. In [13], the authors proposed DMFG-Q-Learning (DMFG-QL), in which each agent learns independently and uses an auxiliary network to predict its next mean action.

On the other hand, federated learning enables distributed clients to train shared models through parameter exchange [19], with extensions addressing heterogeneity, communication efficiency, and privacy [20–22]. Federated reinforcement learning applies the same principle to value or policy learning [23–25]. By exchanging parameters rather than raw data, these methods reduce the direct exposure of raw observations and rewards.

In contrast to DMFG-QL in [13], which maintains an independent Q-network and learned mean-action predictor for each agent, we propose the Decentralized Federated Mean-Field Learning (DF-MFL) that uses an empirical mean field and periodically averages value-function parameters within each federated group. Our proposed work replaces per-agent mean-action prediction with local measurement and parameter sharing, trading agent-level specialization for substantially lower model complexity.

Contributions. The paper makes four contributions.

1. We introduce a hybrid team-aware mean-field estimator that combines the actions of neighbor agents with neighbor teammate actions, enabling mean-field conditioning from strictly local observations.
2. We develop DF-MFL, a decentralized learning architecture that combines local mean-field temporal-difference updates with periodic parameter mixing over intra-team communication graphs. The method requires no central parameter server and exchanges no raw data.
3. We establish a finite-round bound on within-group parameter disagreement with explicit dependence on the graph spectral gap, synchronization period, and local step sizes. We further prove asymptotic parameter consensus under vanishing step sizes and show that convergence of the network average transfers to every local parameter sequence.
4. We evaluate DF-MFL against DMFG-QL under matched environments and training budgets. On the heterogeneous benchmark, the strongest DF-MFL policies achieve higher field scores and win more games against the baseline policy while using approximately 42× fewer distinct learned parameters.

2 Decentralized Mean-Field Game Formulation

2.1 DMFG Model and Agent Dynamics

We represent the decentralized mean-field game (DMFG) by the tuple

$$\mathfrak{G} = \Big\langle \mathcal{I}, \{\mathcal{S}_i\}_{i\in\mathcal{I}}, \{\mathcal{A}_i\}_{i\in\mathcal{I}}, \{P_i\}_{i\in\mathcal{I}}, \{r_i\}_{i\in\mathcal{I}}, \boldsymbol{\mu}_0, \{\mathcal{G}^{\mathcal{T}}\}_{\mathcal{T}\in\{\mathcal{R},\mathcal{B}\}} \Big\rangle,$$

where $\mathcal{I} = \{1, \dots, N\}$ is the set of agents, $\mathcal{S}_i$ and $\mathcal{A}_i$ are the state and action spaces of agent i, P_i is its controlled transition kernel, r_i is its one-stage reward, $\boldsymbol{\mu}_0$ denotes the initial action mean fields, and $\mathcal{G}^{\mathcal{T}}$ is the communication graph associated with team $\mathcal{T}$. The joint state and action spaces are

$$\mathcal{S}^{(N)} := \prod_{i=1}^{N} \mathcal{S}_i, \qquad \mathcal{A}^{(N)} := \prod_{i=1}^{N} \mathcal{A}_i.$$

Agents are partitioned into two teams $\mathcal{R}$ and $\mathcal{B}$, where $|\mathcal{R}| = |\mathcal{B}| = M$ and $N = 2M$. Let $\mathcal{T}(i) \in \{\mathcal{R}, \mathcal{B}\}$ be the team containing agent i and $\mathcal{T}^{-}(i)$ be the opposing team. The single-population case is recovered by removing one of the two teams. As in standard mean-field-game formulations [1, 2], interactions in a large population are represented through aggregate probability measures rather than the complete joint action profile. The DMFG is defined in the limiting regime $N \to \infty$, while the computational experiments use finite-agent empirical approximations. For agent i, let $\mu_t^{\text{self}} \in \mathcal{P}(\mathcal{A}), \mu_t^{\text{opp}} \in \mathcal{P}(\mathcal{A})$ denote the action distributions of its own team and the opposing team, respectively. Their finite-population empirical counterparts are

denoted as

$$\mu_{i,t}^{\text{self},N} := \frac{1}{|\mathcal{T}(i)|} \sum_{j \in \mathcal{T}(i)} \delta_{\psi_{ji}(a_{j,t})}, \qquad \mu_{i,t}^{\text{opp},N} := \frac{1}{|\mathcal{T}^-(i)|} \sum_{j \in \mathcal{T}^-(i)} \delta_{\psi_{ji}(a_{j,t})}, \tag{1}$$

where δ_a is the Dirac measure concentrated at action a and ψ_{ji} maps the action of agent j into the action representation used by agent i. For homogeneous populations, ψ_{ji} is the identity map. For heterogeneous populations, it implements the action projection and zero-padding described in Section 2.2. Each state space $\mathcal{S}_i, i \in \mathcal{I}$ is assumed to be Polish. To accommodate heterogeneous populations, action spaces can vary across agents, where $\mathcal{A}_i$ is a compact subset of a finite-dimensional Euclidean space for each i. After applying the representation maps ψ_{ji}, all action measures are embedded in a common measurable action space $\mathcal{A}$. When the state spaces are shared, we write $\mathcal{S}_i = \mathcal{S} \subseteq \mathbb{R}^{d_x}, i \in \mathcal{I}$.

The space of action probability measures $\mathcal{P}(\mathcal{A})$ is equipped with the 2-Wasserstein metric W_2 whenever the measures have finite second moments. The transition kernel of agent i is

$$P_i : \mathcal{S}_i \times \mathcal{A}_i \times \mathcal{P}_2(\mathcal{A}) \times \mathcal{P}_2(\mathcal{A}) \longrightarrow \mathcal{P}_2(\mathcal{S}_i),$$

and its one-stage reward is

$$r_i : \mathcal{S}_i \times \mathcal{A}_i \times \mathcal{P}_2(\mathcal{A}) \times \mathcal{P}_2(\mathcal{A}) \longrightarrow \mathbb{R}.$$

in the DMFG, the transition and reward of an agent can depend on its local state and action, the aggregate action distribution of its teammates, and the aggregate action distribution of its opponents. On a complete filtered probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$, the corresponding state evolution is given by

$$x_{i,t+1} = f_i\big(x_{i,t}, a_{i,t}, \mu_t^{\text{self}}, \mu_t^{\text{opp}}\big) + \xi_{i,t}, \tag{2}$$

where f_i is the dynamics of agent i, and $\{\xi_{i,t}\}$ is an $\mathbb{F}$-adapted martingale-difference sequence satisfying

$$\mathbb{E}[\xi_{i,t} \mid \mathcal{F}_t] = 0, \qquad \mathbb{E}\big[\|\xi_{i,t}\|_2^2 \mid \mathcal{F}_t\big] \le \sigma_\xi^2.$$

The DMFG differs from a classical homogeneous MFG in two principal respects. First, agent indices are retained, since agents may have different action spaces, reward functions, transition laws, or policy parameters. Second, no agent is assumed to know the global action distributions μ_t^{self} and μ_t^{opp}. Instead, agent i acts using its local observation and the hybrid empirical action measure introduced in Section 2.2.

2.2 Decentralized Mean-Field Estimation

The population action measures μ_t^{self} and μ_t^{opp} describe the aggregate behavior of an agent's own team and the opposing team. These measures are useful for defining the population-level game, but are not directly available to individual agents in the DMFG setting. In particular, agent i neither observes the joint action profile nor receives the

current population measures from a central coordinator. Its decisions are instead based on its local state, its observation history, and the actions of agents visible within its observation region. Let $\mathcal{F}_{i,t}^{\mathrm{loc}} \subseteq \mathcal{F}_t$ denote the information available to agent i at time t. This local information contains the agent's current observation $o_{i,t}$, its previous actions and rewards, and the most recently observed actions of nearby agents. It does not contain μ_t^{self}, μ_t^{opp}, or the states and actions of agents outside the observation region. An admissible Markov policy is therefore a stochastic kernel of the form

$$\pi_{i,t} : \mathcal{O}_i \times \mathcal{P}_2(\mathcal{A}) \longrightarrow \mathcal{P}(\mathcal{A}_i),$$

so that

$$a_{i,t} \sim \pi_{i,t}(\,\cdot \mid o_{i,t}, \tilde{\mu}_{i,t}\,),$$

where $\tilde{\mu}_{i,t}$ is an $\mathcal{F}_{i,t}^{\mathrm{loc}}$-measurable approximation of the relevant population action distribution. Thus, the policy conditions on a locally computable mean-field feature rather than the unavailable global mean field. The observation neighborhood is determined by physical or sensing proximity. For an agent with position $p_{i,t}$ and observation radius ρ, define

$$\mathcal{N}_i(t) := \{j \in \mathcal{I} \setminus \{i\} : \|p_{j,t} - p_{i,t}\|_2 \le \rho\}\,. \tag{3}$$

This set may vary with time as the agents move. We further partition the visible agents into

$$\mathcal{N}_i^{\mathrm{team}}(t) := \mathcal{N}_i(t) \cap \mathcal{T}(i), \qquad \mathcal{N}_i^{\mathrm{opp}}(t) := \mathcal{N}_i(t) \cap \mathcal{T}^{-}(i).$$

Definition 1 (Local Empirical Action Measures) For agent i, define the empirical action measures over its visible teammates and all visible agents by

$$\hat{\mu}_{i,t}^{\mathrm{team}} := \frac{1}{|\mathcal{N}_i^{\mathrm{team}}(t)|} \sum_{j \in \mathcal{N}_i^{\mathrm{team}}(t)} \delta_{\psi_{ji}(a_{j,t-1})}, \tag{4}$$

$$\hat{\mu}_{i,t}^{\mathrm{all}} := \frac{1}{|\mathcal{N}_i(t)|} \sum_{j \in \mathcal{N}_i(t)} \delta_{\psi_{ji}(a_{j,t-1})}. \tag{5}$$

When the corresponding neighborhood is empty, its empirical measure is replaced by a fixed reference measure μ_i^0, taken in the experiments to be uniform over the valid actions of agent i. The map ψ_{ji} converts the action of agent j into the action representation used by agent i.

For homogeneous agents, ψ_{ji} is the identity map. For heterogeneous agents, the action spaces may have different dimensions or semantic meanings. In that case, ψ_{ji} projects a neighbor's action into the observer's action representation, after which the resulting vector is zero-padded to the common input width used by the mean-field encoder. This construction allows an agent to summarize locally observed behavior without requiring identical action spaces across the population. The teammate-only measure in (4) captures the recent behavior of agents with whom agent i has a shared team objective. The all-neighbor measure in (5) additionally reflects the competitive behavior visible in the local region. DF-MFL combines these two signals rather than requiring the agent to reconstruct separate global distributions for both populations.

Definition 2 (Hybrid Team-Aware Mean Field) The local mean-field input used by agent i is

$$\tilde{\mu}_{i,t} := \alpha \hat{\mu}_{i,t}^{\text{team}} + (1-\alpha)\hat{\mu}_{i,t}^{\text{all}}, \qquad \alpha \in [0,1]. \tag{6}$$

The weight α controls the relative emphasis placed on the team and overall local behavior. At $\alpha = 1$, the estimate contains only the actions of visible teammates. At $\alpha = 0$, it reduces to the empirical distribution over all visible agents. Intermediate values preserve information about team coordination while retaining sensitivity to locally observed opponents. Unlike predictive DMFG methods, such as DMFG-QL [13], that train an auxiliary network to estimate the current mean action, DF-MFL directly constructs $\tilde{\mu}_{i,t}$ from observed neighbor actions, which removes the auxiliary prediction model needed by [13] and its associated approximation error.

2.3 Policy, Objective, and Equilibrium

Given the locally observable process $\{(o_{i,t}, \tilde{\mu}_{i,t})\}_{t=0}^{T-1}$, agent i selects actions from a parameterized Markov policy

$$\pi_{\theta_i} : \mathcal{O}_i \times \mathcal{P}_2(\mathcal{A}) \longrightarrow \mathcal{P}(\mathcal{A}_i), \qquad \theta_i \in \Theta_i \subseteq \mathbb{R}^{d_i}.$$

At time t, the action is sampled according to

$$a_{i,t} \sim \pi_{\theta_i}\big(\cdot \mid o_{i,t}, \tilde{\mu}_{i,t}\big). \tag{7}$$

Since $\tilde{\mu}_{i,t}$ is measurable with respect to the local information $\mathcal{F}_{i,t}^{\text{loc}}$, the policy in (7) does not require access to the current global action profile or either population action measure. For a policy profile $\boldsymbol{\theta} := (\theta_1, \ldots, \theta_N)$, let

$$\boldsymbol{\mu}^{\boldsymbol{\theta}} := \left\{ \mu_t^{\mathcal{R},\boldsymbol{\theta}}, \mu_t^{\mathcal{B},\boldsymbol{\theta}} \right\}_{t=0}^{T-1}$$

denote the team action flows induced by the agents' policies and the state dynamics. For team $\mathcal{T} \in \{\mathcal{R}, \mathcal{B}\}$, the finite-population empirical action measure is

$$\mu_t^{\mathcal{T},N,\boldsymbol{\theta}} := \frac{1}{|\mathcal{T}|} \sum_{j \in \mathcal{T}} \delta_{\psi_j(a_{j,t})}, \tag{8}$$

and $\mu_t^{\mathcal{T},\boldsymbol{\theta}}$ denotes its mean-field limit, when that limit exists. Here, ψ_j embeds the action of agent j into the common action representation introduced in Section 2.2. The expected finite-horizon return of agent i is

$$J_i\big(\theta_i, \boldsymbol{\theta}_{-i}\big) := \mathbb{E}^{\boldsymbol{\pi}_{\boldsymbol{\theta}}} \left[\sum_{t=0}^{T-1} \gamma_d^t r_i \left(x_{i,t}, a_{i,t}, \mu_t^{\text{self},\boldsymbol{\theta}}, \mu_t^{\text{opp},\boldsymbol{\theta}} \right) \right], \tag{9}$$

where $\gamma_d \in [0,1]$ is the reward discount factor and $\boldsymbol{\theta}_{-i}$ denotes the parameters of all agents other than i.

We adapt from the standard equilibrium concept for mean-field games [1, 2].

Definition 3 (Decentralized Mean-Field Nash Equilibrium) A policy profile $\boldsymbol{\theta}^\star = (\theta_1^\star, \ldots, \theta_N^\star)$ and the associated population action flows $\boldsymbol{\mu}^\star = \boldsymbol{\mu}^{\boldsymbol{\theta}^\star}$ form a decentralized mean-field Nash equilibrium (D-MFNE) if the following conditions hold.

1. *Local-information optimality.* For every agent i and every admissible parameter $\theta_i \in \Theta_i$,
$$J_i\big(\theta_i^\star, \boldsymbol{\theta}_{-i}^\star\big) \geq J_i\big(\theta_i, \boldsymbol{\theta}_{-i}^\star\big), \tag{10}$$
where both the equilibrium policy and the deviating policy are restricted to the locally observable information $(o_{i,t}, \tilde{\mu}_{i,t})$.
2. *Population consistency.* For each team $\mathcal{T} \in \{\mathcal{R}, \mathcal{B}\}$, the candidate action flow coincides with the flow induced by the equilibrium policy profile:
$$\boldsymbol{\mu}^{\mathcal{T},\star} = \Psi^{\mathcal{T}}\big(\boldsymbol{\pi}_{\boldsymbol{\theta}^\star}, \boldsymbol{\mu}^\star\big), \tag{11}$$
where $\Psi^{\mathcal{T}}$ is the population-flow operator generated by the state dynamics and the policies of team $\mathcal{T}$.

2.4 Intra-Team Communication Network

The observation neighborhood introduced in Section 2.2 determines which agents' actions are visible to a learner. Federated parameter exchange is governed by a separate communication network, which allows sensing and communication to follow different topologies. For each team $\mathcal{T} \in \{\mathcal{R}, \mathcal{B}\}$, let

$$\mathcal{G}^{\mathcal{T}} = \big(\mathcal{V}^{\mathcal{T}}, \mathcal{E}^{\mathcal{T}}\big)$$

be an undirected communication graph, where $\mathcal{V}^{\mathcal{T}} = \mathcal{T}$ is the set of agents in the team and $\mathcal{E}^{\mathcal{T}}$ is the set of communication links. An edge $(i, j) \in \mathcal{E}^{\mathcal{T}}$ indicates that agents i and j may exchange value-function parameters during a synchronization round. It is assumed that no communication edge connects agents on opposing teams, and agents with incompatible value-function output dimensions cannot directly average all parameters for heterogeneous populations. Communication and averaging are therefore performed within federated groups. Let

$$\mathcal{H} := \{g = (\mathcal{T}, c) : \mathcal{T} \in \{\mathcal{R}, \mathcal{B}\},\ c \in \mathcal{C}_{\mathcal{T}}\}$$

denote the index set of federated groups, where $\mathcal{C}_{\mathcal{T}}$ is the set of agent classes present in team $\mathcal{T}$. For each $g = (\mathcal{T}, c) \in \mathcal{H}$, define

$$\mathcal{V}_g := \{i \in \mathcal{T} : c(i) = c\}, \qquad n_g := |\mathcal{V}_g|,$$

where $c(i)$ denotes the class of agent i.

Let $n_g := |\mathcal{V}_g|$ denote the number of agents in group g. For notational simplicity, the group subscript is omitted below when the relevant group is clear. Parameter mixing within a group is represented by a matrix

$$W = [W_{ij}] \in \mathbb{R}^{n \times n},$$

where $W_{ij} > 0$ only if $i = j$ or agents i and j are connected by a communication edge. At a synchronization round, agent i applies

$$\theta_i^+ = \sum_{j=1}^{n} W_{ij} \theta_j^-, \tag{12}$$

where θ_j^- and θ_i^+ denote the parameters immediately before and after communication, respectively. We impose the following standard conditions on the communication process:

1. The communication graph for each federated group is connected.
2. The matrix W is symmetric and doubly stochastic:

$$W = W^\top, \qquad W\mathbf{1} = \mathbf{1}, \qquad \mathbf{1}^\top W = \mathbf{1}^\top.$$

Double stochasticity ensures that parameter update through communication modifies the disagreement among local parameter vectors without changing the average of the parameter vector. To characterize the contraction of disagreement, define the averaging projection

$$J := \frac{1}{n} \mathbf{1}\mathbf{1}^\top$$

and the network contraction factor

$$q := \|W - J\|_2 \in [0, 1). \tag{13}$$

Since the communication graph is connected and W is symmetric and doubly stochastic, repeated application of W contracts vectors in the subspace orthogonal to $\mathbf{1}$. We define the corresponding spectral gap by

$$\gamma := 1 - q > 0. \tag{14}$$

Communication occurs once every τ_F episodes. Between synchronization rounds, agents update their parameters independently using locally generated experience, so their parameter vectors may drift apart. The communication graph is used only for model-parameter exchange. Federated averaging avoids the direct exchange and centralized collection of raw observations, rewards, and trajectories by communicating model parameters instead of local data.

3 Decentralized Federated Mean-Field Learning

3.1 Overview

DF-MFL combines local mean-field temporal-difference learning with periodic parameter averaging over intra-team communication graphs. Each agent selects actions from its own observation and the hybrid mean-field estimate in (6), stores locally generated transitions, and updates its value function independently. Every τ_F episodes, agents within the same federated group exchange and mix value-function parameters. Observations, rewards, trajectories, and replay-buffer contents remain local.

The method therefore alternates between two phases. During local learning, agents adapt to their own experiences and may develop different parameter vectors. During synchronization, parameter mixing reduces within-group disagreement without changing the group-average parameter vector. Exact group averaging eliminates disagreement at the synchronization instant, whereas graph-local mixing contracts it according to the communication topology. Section 4 quantifies this contraction.

3.2 Value-Function Architecture

Each agent maintains a mean-field-conditioned action-value function

$$Q_{\theta_i}(o_{i,t}, \tilde{\mu}_{i,t}) = F_{\text{out}}\left(\text{ReLU}\left(F_{\text{fuse}}\left[E_o(o_{i,t}); E_\mu(\tilde{\mu}_{i,t})\right]\right)\right), \tag{15}$$

where E_o encodes the local observation, E_μ encodes the hybrid mean-field input, F_{fuse} combines the two representations, and F_{out} produces one value for each admissible action. A target network with parameters $\bar{\theta}_i$ is maintained through Polyak averaging.

For heterogeneous populations, agents with different classes maintain class-specific architectures since their action spaces and output dimensions differ. Parameter averaging is therefore performed within the federated groups defined in Section 2.4.

3.3 Local Mean-Field Temporal-Difference Update

At each environment step, agent i constructs $\tilde{\mu}_{i,t}$ from the most recently observed actions of its visible neighbors and selects an action ε-greedily from $Q_{\theta_i}(o_{i,t}, \tilde{\mu}_{i,t})$. The resulting transition is stored in the agent's local replay buffer. Since observations, rewards, and neighborhoods differ across agents, the local updates generally cause their parameter vectors to drift apart between synchronization rounds.

The online network is updated by minimizing the Double-DQN loss

$$\mathcal{L}(\theta_i) = \mathbb{E}\Big[\big(r + \gamma_d(1-d)\, Q_{\bar{\theta}_i}(o', \tilde{\mu}', \arg\max_{a'} Q_{\theta_i}(o', \tilde{\mu}')) - Q_{\theta_i}(o, \tilde{\mu}, a)\big)^2\Big], \tag{16}$$

with discount γ_d (distinguished from the spectral gap γ).

3.4 Intra-Team Federated Averaging

Every τ_F episodes, agents within each federated group $g \in \mathcal{H}$ synchronize their value-function parameters. Under exact group averaging,

$$\theta_i \leftarrow \frac{1}{n_g} \sum_{j \in \mathcal{V}_g} \theta_j, \qquad i \in \mathcal{V}_g. \tag{17}$$

The target-network parameters are then reset to the synchronized online parameters. Equation (17) corresponds to $W_g = J_g$ and $q_g = 0$, so all parameter vectors in the group are identical immediately after synchronization.

For graph-local communication, the update is

$$\theta_i \leftarrow \sum_{j \in \mathcal{V}_g} (W_g)_{ij} \theta_j.$$

This update contracts parameter disagreement without necessarily eliminating it in one communication round. No parameter exchange occurs across teams or between agents with incompatible network architectures.

3.5 Training Stabilization

The competitive benchmarks exhibit two training pathologies that are separate from the local mean-field and federated-learning components of DF-MFL. First, the two starting positions may not be strategically equivalent, causing a policy to benefit from the side to which it is assigned. Second, self-play may converge to a passive regime in which both teams avoid engagement. We use side-mirroring and an intrinsic engagement bonus to reduce these effects during training. These mechanisms are training interventions rather than components of the consensus analysis in Section 4.

At the beginning of each episode, the assignment of team policies to starting positions is reversed with probability p_{swap}. Agent classes are preserved, and every transition is attributed to the policy that controlled the agent after the assignment. If $V^k(\pi^R, \pi^B)$ denotes the expected return of policy k when π^R and π^B occupy the original starting positions, then its effective training objective is

$$\bar{V}^k = (1 - p_{\text{swap}}) V^k(\pi^R, \pi^B) + p_{\text{swap}} V^k(\pi^B, \pi^R). \tag{18}$$

With $p_{\text{swap}} = \frac{1}{2}$, the objective assigns equal weight to both policy-to-position configurations and is therefore invariant to relabeling the two starting sides. This does not remove positional asymmetry from the environment, but it prevents either policy from being trained exclusively from one side.

To discourage mutually passive behavior, the environment reward is augmented during training by a two-tier intrinsic engagement bonus:

$$\tilde{r}_{i,t} = r^{\text{env}}_{i,t} + \beta_b \frac{K^{\text{team}}_{i,t}}{M_{\text{tot}}} + \beta_{\ell,t} \frac{K^{\text{team}}_{i,t}}{M_{\text{tot}}} \mathbf{1}\big[\mathcal{T}(i) = \ell_t\big], \tag{19}$$

Algorithm 1 Decentralized Federated Mean-Field Learning (DF-MFL)

Require: Teams $\mathcal{R}, \mathcal{B}$; hybrid weight α; federated period τ_F; radius ρ; swap probability p_s; bonuses; episodes E.

Initialize $\{Q_{\theta_i}\}, \{Q_{\bar{\theta}_i}\}$, replay buffers, and win history.
for $e = 1, \ldots, E$ **do**
  Sample swap flag b; set ε_e; compute imbalance ι and trailing team ℓ.
  for $t = 0, \ldots, T-1$ **do**
    for each agent i **do**
      Form $\mathcal{N}_i(t)$; compute $\tilde{\mu}_{i,t}$ by (6).
      Select $a_{i,t}$ ε-greedily w.r.t. $Q_{\theta_{\phi(i,b)}}$.
    end for
    Step environment; augment rewards by (19); store transitions.
    Update θ_i by (16); Polyak-update targets.
  end for
  Record winner; store federated snapshot.
  if $e \bmod \tau_F = 0$ **then** apply (17) within each team.
  end if
end for
return snapshot-averaged parameters $\{\theta_i\}$.

where $K_{i,t}^{\text{team}}$ measures the engagement achieved by agent i's team at time t, M_{tot} is the normalization constant, and $\beta_b \geq 0$ is a base bonus applied to both teams. The base component encourages active interaction regardless of the current balance of play.

The second component is applied only to the trailing team ℓ_t. Its coefficient is

$$\beta_{\ell,t} = \min\left\{\kappa \iota_t, \beta_{\max}\right\}, \qquad \iota_t = 2\left|\hat{w}_t^R - \frac{1}{2}\right|, \tag{20}$$

where $\hat{w}_t^R$ is the rolling win rate of the red team, $\kappa \geq 0$ controls the response to imbalance, and $\beta_{\max}$ caps the additional bonus. The adaptive component acts as a bounded restoring force, which increases with the observed win-rate imbalance and approaches zero as the rolling win rate approaches parity. The cap prevents a large temporary imbalance from producing an arbitrarily large reward modification.

3.6 DF-MFL Algorithm

The proposed DF-MFL algorithm is presented in Algorithm 1. Evaluation uses snapshot-averaged parameters: the last K federated snapshots per group are averaged, sampling the converged regime rather than a single late iterate.

4 Consensus, Communication and Parameter Complexity

4.1 Consensus Analysis

The following analysis isolates the effect of periodic parameter mixing from the convergence behavior of the underlying Double-DQN updates. We represent the local learning dynamics as a federated stochastic-approximation process, where each agent performs a noisy parameter update using locally generated experience before communicating with its federated group. Consistent with standard analyses of stochastic approximation and federated reinforcement learning [24], the following assumption requires the local update directions to have uniformly bounded second moments in order to bound the parameter drift accumulated between synchronization rounds.

Assumption 1 For each federated group, the local stochastic parameter updates can be written as

$$\Theta^{k+\frac{1}{2}} = \Theta^k - \eta_k G^k,$$

where the ith row of $\Theta^k \in \mathbb{R}^{n\times d}$ is $(\theta_i^k)^\top$ and the ith row of G^k is $(g_i^k)^\top$. $\theta_i^{k+\frac{1}{2}}$ denotes the parameter vector of agent i after its local learning update at the k-th iteration, and before the averaging step, g_i^k is the local stochastic update direction, and η_k is the local step size. There exists a finite constant $G > 0$ such that

$$\mathbb{E}\Big[\|G^k\|_F^2\Big] \le nG^2, \qquad k \ge 0.$$

While Assumption 1 bounds the disagreement introduced by local learning, the next assumption ensures that each communication round contracts the disagreement already present among agents in the federated group.

Assumption 2 For each federated group, the communication graph is connected and undirected, and its mixing matrix $W \in \mathbb{R}^{n\times n}$ is symmetric and doubly stochastic, that is, $W = W^\top, W\mathbf{1} = \mathbf{1}, \mathbf{1}^\top W = \mathbf{1}^\top$. Let the averaging matrix $J := \frac{1}{n}\mathbf{1}\mathbf{1}^\top$, the matrix $W - J$ measures how far graph-mixing step is from averaging. $q := \|W - J\|_2 \in [0, 1)$ denotes the worst-case contraction factor for parameter disagreement in one communication round. It is assumed that $\gamma := 1 - q > 0$, where γ is the spectral gap governing contraction on the disagreement subspace.

The evolution of parameter disagreement is governed by two effects. Between synchronization rounds, heterogeneous local stochastic updates generate drift among the agents' parameter vectors, while at synchronization, the mixing matrix contracts this disagreement and preserves the group average. The following theorem combines these two effects to provide a finite-round bound on the expected parameter disagreement after any finite number of communication rounds and to make explicit its dependence on the synchronization period, communication topology, and local step sizes.

Theorem 1 *Consider a federated group consisting of n agents, where a group may correspond to one team or a pair of teams. Let $\theta_i^k \in \mathbb{R}^d$ be the parameter vector of agent i at local iteration k, and let*

$$\Theta^k := \begin{bmatrix} (\theta_1^k)^\top \\ \vdots \\ (\theta_n^k)^\top \end{bmatrix} \in \mathbb{R}^{n\times d}.$$

The agents perform the local updates

$$\Theta^{k+\frac{1}{2}} = \Theta^k - \eta_k G^k \tag{21}$$

and communicate every τ_F local iterations according to

$$\Theta^{k+1} = \begin{cases} W\Theta^{k+\frac{1}{2}}, & k+1 \equiv 0 \pmod{\tau_F}, \\ \Theta^{k+\frac{1}{2}}, & \text{otherwise.} \end{cases} \tag{22}$$

At communication round r, let the post-communication iterate and their disagreement matrix be

$$X^r := \Theta^{r\tau_F}, \qquad E^r := (I-J)X^r.$$

Then under Assumptions 1 and 2, for every $r \geq 1$,

$$\mathbb{E}\Big[\|E^r\|_F^2\Big] \leq q^r\|E^0\|_F^2 + \frac{nG^2q^2\tau_F}{1-q}\sum_{s=0}^{r-1} q^{r-1-s}\sum_{k=s\tau_F}^{(s+1)\tau_F-1}\eta_k^2. \tag{23}$$

Moreover,

$$\|E^r\|_F^2 = \sum_{i=1}^n \|\theta_i^{r\tau_F} - \bar{\theta}^{r\tau_F}\|_2^2, \qquad \bar{\theta}^k := \frac{1}{n}\sum_{i=1}^n \theta_i^k,$$

so (23) bounds the total mean-square parameter disagreement. If $\{\eta_k\}$ is nonincreasing, then

$$\mathbb{E}\Big[\|E^r\|_F^2\Big] \leq q^r\|E^0\|_F^2 + \frac{nG^2q^2\tau_F^2}{1-q}\sum_{s=0}^{r-1} q^{r-1-s}\eta_{s\tau_F}^2. \tag{24}$$

Thus, initial disagreement contracts geometrically at rate $q = 1-\gamma$, while disagreement introduced by local updates is controlled by τ_F, the step sizes, and the graph spectral gap.

Proof For communication round r, let the accumulated local update be

$$B^r := \sum_{k=r\tau_F}^{(r+1)\tau_F-1} \eta_k G^k. \tag{25}$$

If $q = 0$, then $\|W-J\|_2 = 0$ and therefore $W = J$. Consequently, every communication step produces exact parameter consensus:

$$E^{r+1} = (W-J)(E^r - B^r) = 0.$$

Thus, the claimed bound holds. It remains to consider $q \in (0, 1)$.

Since W is doubly stochastic, it holds that $W\mathbf{1} = \mathbf{1}$ and $\mathbf{1}^\top W = \mathbf{1}^\top$. Thus

$$WJ = JW = J, \qquad (I-J)W = W-J, \qquad (W-J)J = 0.$$

Applying all local updates in the rth communication block and then one mixing step gives

$$X^{r+1} = W(X^r - B^r). \tag{26}$$

Projecting (26) onto the disagreement subspace yields

$$\begin{aligned} E^{r+1} &= (I-J)W(X^r - B^r) \\ &= (W-J)(X^r - B^r) \\ &= (W-J)(E^r + JX^r - B^r) \\ &= (W-J)(E^r - B^r) + (W-J)JX^r \\ &= (W-J)(E^r - B^r). \end{aligned} \tag{27}$$

The last equation holds since $(W-J)J = WJ - J^2 = J - J = 0$. Taking the Frobenius norm and applying the triangle inequality, we get

$$\|E^{r+1}\|_F \le q\big(\|E^r\|_F + \|B^r\|_F\big), \tag{28}$$

where $q = \|W - J\|_2$. From the weighted Young's inequality we have that for any $a, b \ge 0$ and $q \in (0,1)$,

$$q^2(a+b)^2 \le qa^2 + \frac{q^2}{1-q}b^2, \tag{29}$$

since subtracting the right-hand side with the left-hand side of (29) yields

$$\frac{q}{1-q}\big((1-q)a - qb\big)^2 \ge 0.$$

Applying (29) to (28) gives

$$\|E^{r+1}\|_F^2 \le q\|E^r\|_F^2 + \frac{q^2}{1-q}\|B^r\|_F^2. \tag{30}$$

To bound the second term in (30), we next estimate the norm of the accumulated local update B^r. By its definition in (25) and the Cauchy–Schwarz inequality,

$$\|B^r\|_F^2 = \left\|\sum_{k=r\tau_F}^{(r+1)\tau_F-1} \eta_k G^k\right\|_F^2 \le \tau_F \sum_{k=r\tau_F}^{(r+1)\tau_F-1} \eta_k^2 \|G^k\|_F^2. \tag{31}$$

Taking expectations and using Assumption 1,

$$\mathbb{E}\Big[\|B^r\|_F^2\Big] \le nG^2\tau_F \sum_{k=r\tau_F}^{(r+1)\tau_F-1} \eta_k^2. \tag{32}$$

Combining (30) and (32) yields

$$\mathbb{E}\Big[\|E^{r+1}\|_F^2\Big] \le q\mathbb{E}\Big[\|E^r\|_F^2\Big] + \frac{nG^2q^2\tau_F}{1-q} \sum_{k=r\tau_F}^{(r+1)\tau_F-1} \eta_k^2. \tag{33}$$

Notice that the inequality (33) holds for any r. For instance, when $r = 0$, we have $\mathbb{E}\Big[\|E^1\|_F^2\Big] \le q\|E^0\|_F^2 + \frac{nG^2q^2\tau_F}{1-q}\sum_{k=\tau_F}^{\tau_F-1}\eta_k^2$. By substituting the one-step inequality (33) into itself r times, we obtain the inequality in terms of the initial disagreement and all accumulated local-update terms

$$\mathbb{E}\Big[\|E^r\|_F^2\Big] \le q^r\|E^0\|_F^2 + \frac{nG^2q^2\tau_F}{1-q}\sum_{s=0}^{r-1} q^{r-1-s} \sum_{k=s\tau_F}^{(s+1)\tau_F-1} \eta_k^2. \tag{34}$$

which proves (23). If the step sizes are nonincreasing, then for every $k \in \{s\tau_F, \dots, (s+1)\tau_F - 1\}$, $\eta_k \le \eta_{s\tau_F}$. Since the sth communication block contains τ_F local iterations,

$$\sum_{k=s\tau_F}^{(s+1)\tau_F-1} \eta_k^2 \le \tau_F \eta_{s\tau_F}^2.$$

Substituting this bound into (23) yields (24). □

Theorem 1 provides a finite-round decomposition of parameter disagreement into the initial disagreement that converges to zero and the drift introduced by local updates between synchronization rounds. When the synchronization period is fixed and the local step sizes vanish, the latter contribution also vanishes. This yields the following consensus result.

Corollary 2 *Let the conditions of Theorem 1 hold, $\tau_F < \infty$ is fixed, and*

$$\eta_k \longrightarrow 0.$$

Then the post-communication iterates reach mean-square consensus

$$\lim_{r\to\infty} \mathbb{E}\left[\sum_{i=1}^{n} \|\theta_i^{r\tau_F} - \bar{\theta}^{r\tau_F}\|_2^2\right] = 0. \tag{35}$$

The consensus also holds at every local iteration

$$\lim_{k\to\infty} \mathbb{E}\left[\sum_{i=1}^{n} \|\theta_i^{k} - \bar{\theta}^{k}\|_2^2\right] = 0. \tag{36}$$

If, in addition, the network-average sequence converges in mean square to $\theta^\star$, that is,

$$\lim_{k\to\infty} \mathbb{E}\Big[\|\bar{\theta}^k - \theta^\star\|_2^2\Big] = 0, \tag{37}$$

then every local parameter sequence converges in mean square to the same limit

$$\lim_{k\to\infty} \mathbb{E}\Big[\|\theta_i^k - \theta^\star\|_2^2\Big] = 0, \qquad i = 1, \dots, n. \tag{38}$$

Proof When $q = 0$, every post-communication disagreement matrix is zero. Moreover, since τ_F is fixed and $\eta_k \to 0$, the parameter drift accumulated over any communication interval also vanishes asymptotically. The consensus conclusions follow directly in this case.

In the case where $q \in (0, 1)$, let $u_r := \mathbb{E}[\|E^r\|_F^2]$ and

$$c_r := \frac{nG^2q^2\tau_F}{1-q} \sum_{k=r\tau_F}^{(r+1)\tau_F-1} \eta_k^2.$$

Equation (33) can then be rewritten as

$$u_{r+1} \le qu_r + c_r, \qquad 0 < q < 1. \tag{39}$$

Since τ_F is fixed and $\eta_k \to 0$, $c_r \to 0$. Applying the one-step recursion (39) successively over r communication rounds yields

$$u_r \le q^r u_0 + \sum_{s=0}^{r-1} q^{r-1-s} c_s. \tag{40}$$

The first term in (40) goes to zero as $r \to \infty$. Let $\varepsilon > 0$, since $c_s \to 0$, there exists s_0 such that $c_s \le \varepsilon(1-q)$ for all $s \ge s_0$. Therefore

$$\sum_{s=0}^{r-1} q^{r-1-s} c_s \le \sum_{s=0}^{s_0-1} q^{r-1-s} c_s + \varepsilon(1-q) \sum_{s=s_0}^{r-1} q^{r-1-s}$$

$$\leq \sum_{s=0}^{s_0-1} q^{r-1-s} c_s + \varepsilon. \tag{41}$$

The finite sum in (41) goes to zero as $r \to \infty$. Since ε is arbitrary, $u_r \to 0$, proving (35).

For an iteration k between communication rounds, let $r := \lfloor k/\tau_F \rfloor$, the parameters have evolved from the last synchronized iterate X_r through local updates as $\Theta^k = X_r - \sum_{\ell=r\tau_F}^{k-1} \eta_\ell G_\ell$. Applying the disagreement projector $(I - J)$, we get

$$(I-J)\Theta^k = E_r - (I-J) \sum_{\ell=r\tau_F}^{k-1} \eta_\ell G_\ell, \tag{42}$$

since $E_r = (I - J)X_r$.

Using $(a+b)^2 \leq 2a^2 + 2b^2$, $\|I - J\|_2 = 1$, and the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned}
\mathbb{E}\Big[\|(I-J)\Theta^k\|_F^2\Big] &= \mathbb{E}\left[\left\| E_r - (I-J) \sum_{\ell=r\tau_F}^{k-1} \eta_\ell G_\ell \right\|_F^2\right] \\
&\leq 2\mathbb{E}\Big[\|E_r\|_F^2\Big] + 2\mathbb{E}\left[\left\| \sum_{\ell=r\tau_F}^{k-1} \eta_\ell G_\ell \right\|_F^2\right] \\
&\leq 2\mathbb{E}\Big[\|E_r\|_F^2\Big] + 2\tau_F \sum_{\ell=r\tau_F}^{k-1} \eta_\ell^2 \mathbb{E}\Big[\|G_\ell\|_F^2\Big] \\
&\leq 2\mathbb{E}\Big[\|E_r\|_F^2\Big] + 2nG^2\tau_F \sum_{\ell=r\tau_F}^{(r+1)\tau_F-1} \eta_\ell^2,
\end{aligned} \tag{43}$$

where the last inequality follows from Assumption 1.

Both terms on the right hand side of (43) go to zero as $r \to \infty$, and notice that $\|(I - J)\Theta^k\|_F^2 = \sum_{i=1}^n \|\theta_i^k - \bar{\theta}^k\|^2$, which proves (36).

Finally, for every agent i,

$$\|\theta_i^k - \theta^\star\|_2^2 \leq 2\|\theta_i^k - \bar{\theta}^k\|_2^2 + 2\|\bar{\theta}^k - \theta^\star\|_2^2.$$

Taking expectations and using (36) and (37) proves (38). □

Remark 1 Theorem 1 establishes convergence of parameter disagreement. It separates the effect of the communication mechanism from the underlying reinforcement-learning dynamics. If the network-average learning process converges to a stationary parameter $\theta^\star$, Corollary 2 guarantees that all members of the federated group converge to the same limit.

Remark 2 For graph-local mixing with $0 < q < 1$, the contraction factor in (23) is $q = 1 - \gamma$. For exact group averaging, $W = J$ and $q = 0$, so disagreement is eliminated at every synchronization. Better connected graphs have a larger spectral gap γ and therefore contract pre-existing disagreement more rapidly. The accumulation term also grows with the communication period τ_F, reflecting the larger parameter drift permitted between synchronization rounds. Thus, more frequent communication and better graph connectivity improve parameter consistency.

4.2 Communication and Parameter Complexity

Each averaging round transmits d parameters per agent over the team graph, giving a per-team volume of $O(Md/\tau_F)$ parameter values per episode; no raw observations or rewards are communicated, so τ_F directly trades communication cost against consensus speed as quantified by Theorem 1.

Beyond volume, consensus has a structural consequence that distinguishes DF-MFL sharply from independent-agent learners. Since (17) drives every agent in a (team, class) group to identical parameters at each synchronization, the system learns only *one distinct network per group*, not one per agent. Table 1 quantifies this against DMFG-QL, which maintains a Q-network *and* a mean-action predictor for each of the N agents independently.

Table 1: Learned parameter counts for the two benchmarks of Section 5. "Per agent" is the number of parameters held by a single agent. "Distinct (system)" is the number of independently learned parameters across the whole system: four (team, class) groups for DF-MFL under consensus, and N independent agents for DMFG-QL. The homogeneous benchmark has $N = 40$ single-class agents; the heterogeneous benchmark has $N = 40$ agents in two classes per team.

Benchmark	Method	Per agent	Distinct (system)
Homogeneous	DF-MFL	153,621	307,242
	DMFG-QL	810,364	32,414,560
Heterogeneous	DF-MFL	∼240,000	∼960,000
	DMFG-QL	∼1,016,000	∼40,700,000

Per agent, DF-MFL is smaller than DMFG-QL since it requires no mean-action predictor. At the system level the gap is much larger with approximately 42× on the heterogeneous benchmark and 106× on the homogeneous one.

5 Computational Experiments

5.1 Benchmarks and Baseline

We use two benchmarks built on the MAgent2 multi-agent simulator [26]. The *homogeneous* benchmark is a 25×25 grid with $M = 20$ agents per team, $|\mathcal{A}| = 21$ actions, and a local view flattened to $d_o = 845$. The *heterogeneous* benchmark fields 10 melee ($|\mathcal{A}| = 9$) and 10 ranged ($|\mathcal{A}| = 25$) agents per team with $d_o = 1521$, so DF-MFL maintains four (team, class) groups. Both use identical reward parameters: step -0.005, opponent elimination $+5.0$, own elimination -0.1, ineffective attack -0.1, horizon 500.

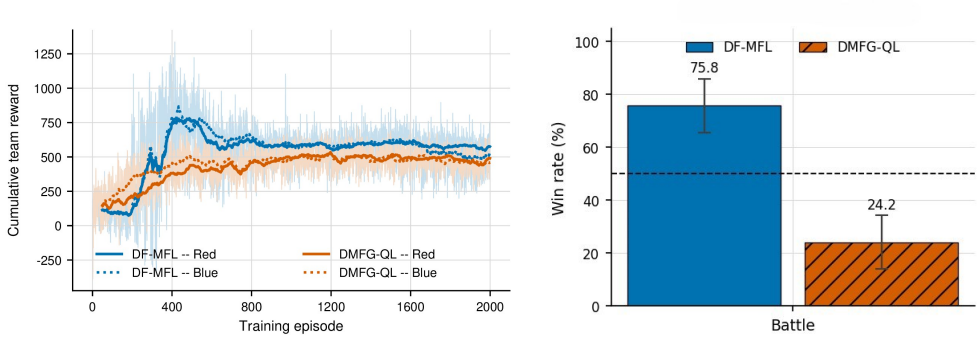
Our baseline is DMFG-QL [13], reimplemented from the authors' released code and ported to the same simulator. DMFG-QL maintains an independent Q-network and mean-action predictor for each agent, whereas DF-MFL uses a measured local mean field and periodic within-group parameter averaging. Both methods use the same environment configuration and the same training budget of 2000 episodes. Each retains its original mean-field convention and action-selection rule so that the comparison evaluates the complete algorithms rather than modified variants.

5.2 Evaluation Protocol

We evaluate on the game where all shaping is disabled, position assignment is randomized across paired games, and the unit of analysis is a *team policy*—one team's networks—since a training run yields two policies that can differ in strength. Policies are cross-played round-robin; the primary outcome is eliminations, with equal eliminations recorded as a genuine tie, and the mean elimination margin (denoted killΔ) reported as a continuous secondary statistic [27, 28].

5.3 Head-to-Head Comparison on the Homogeneous Benchmark

Cross-playing the four team policies (two per method, 30 games per ordered pair) yields a fully transitive ladder: DF-MFL red scores 74.4% against the field, DMFG-QL red 56.7%, DF-MFL blue 42.2%, and DMFG-QL blue 23.3%. Aggregated over the four cross-method pairings, DF-MFL wins 65% of 240 games, and the strongest single policy in the population is produced by DF-MFL training. The training reward per episode for our DF-MFL and DMFG-QL are shown in Fig. 1a and the head-to-head competition result between the trained red team policies of DF-MFL and DMFG-QL is shown in Fig. 1b.



(a) Training rewards

(b) Head-to-head competition

Fig. 1: Homogeneous battle results. In (a) average in lines, and standard deviation in shaded regions. In (b) average in bars and standard deviation in black lines.

Table 2: Heterogeneous benchmark at matched training budget (2000 episodes). "Field" is the mean score against all other policies in the cross-play, with ties counted as one half; "killΔ" is the mean per-game elimination margin, positive when the row policy eliminates more agents than it loses; the last two columns give the direct head-to-head win rate against each baseline policy.

Policy	Field	killΔ	vs. DMFG-QL red	vs. DMFG-QL blue
DF-MFL red	**56.9%**	+0.9	78%	85%
DMFG-QL red	50.6%	+0.1	—	72%
DF-MFL blue	48.3–49.2%	≈ 0	42–45%	42–48%
DMFG-QL blue	37.5%	−1.5	28%	—

5.4 Matched-Budget Comparison on the Heterogeneous Benchmark

This is our principal comparison. Both methods train for 2000 episodes on an identical configuration. Table 2 reports a ten-policy cross-play with 20 games per ordered pair.

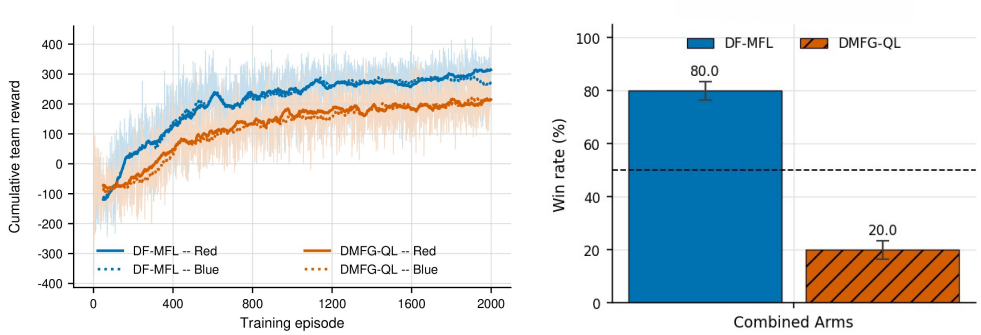
At matched budget, DF-MFL defeats the baseline DMFG-QL. At the matched training budget, the DF-MFL red policy beats DMFG-QL red in 78% of games and DMFG-QL blue in 85% of games. It also obtains a higher score against the common policy field, 56.9% compared with 50.6% for DMFG-QL red. The elimination margins provide a consistent secondary result: DF-MFL red obtains a mean margin of +0.9, compared with +0.1 for DMFG-QL red and −1.5 for DMFG-QL blue. Taken together with Section 4.2, DF-MFL achieves this while learning approximately 42× fewer distinct parameters.

The training reward coparision between the proposed DF-MFL and baseline DMFG-QL are shown in Fig. 2a and the head-to-head competition result between the trained red team policies of DF-MFL and DMFG-QL is shown in Fig. 2b.

6 Conclusion

We introduced DF-MFL, which combines a hybrid local mean-field estimator with periodic parameter averaging over intra-team communication graphs. The method keeps observations, rewards, and trajectories local while using parameter mixing to limit disagreement within each federated group. We established a finite-round consensus bound with explicit dependence on the graph spectral gap, synchronization period, and local step sizes. Under vanishing step sizes, the agents attain asymptotic parameter consensus, and if the network-average sequence converges, every local sequence converges to the same limit.

Compared against the baseline DMFG-QL [13] under an identical environment and an identical 2000-episode budget, DF-MFL policies win more head-to-head games on the heterogeneous benchmark and score higher against the common field (56.4–56.9% versus 50.6%), while learning approximately 42× fewer distinct parameters, since consensus shares one policy per (team, class) group instead of training every



(a) Training rewards

(b) Head-to-head competition

Fig. 2: Heterogeneous combined arms results. In (a) average in lines, and standard deviation in shaded regions. In (b) average in bars and standard deviation in black lines.

agent independently. On the homogeneous benchmark, DF-MFL yields the strongest single policy in the population.

Declarations

- **Funding.** The author declares that no funds, grants, or other support were received during the preparation of this manuscript.
- **Competing interests.** The author has no competing interests to declare that are relevant to the content of this article.
- **Ethics approval and consent to participate.** Not applicable. This study did not involve human participants or animals.
- **Consent for publication.** Not applicable.
- **Data availability.** No external datasets were used. All data reported in this article were generated by simulation using the open-source MAgent2 environment [26]; the environment configurations required to regenerate them are specified in full in Section 5.1.
- **Materials availability.** Not applicable.
- **Code availability.** The implementation of DF-MFL, the reimplementation of the DMFG-QL baseline, and the cross-play evaluation code will be made available upon request to the authors.
- **Author contribution.** W.S. and T.T. contributed to the study conception, design, and analysis. W.S. implemented the experiments and performed the data collection. W.S. wrote the first draft of the manuscript, and T.T. revised the manuscript. All authors read and approved the final manuscript.

References

[1] Lasry, J.-M., Lions, P.-L.: Mean field games. Japanese Journal of Mathematics

2(1), 229–260 (2007)

[2] Huang, M., Caines, P.E., Malhamé, R.P.: 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 (2007)

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

[4] Elie, R., Hubert, E., Turinici, G.: Contact rate epidemic control of COVID-19: an equilibrium view. Mathematical Modelling of Natural Phenomena **15**, 35 (2020)

[5] Huang, M., Zhou, M.: Linear quadratic mean field games: Asymptotic solvability and relation to the fixed point approach. IEEE Transactions on Automatic Control **65**(4), 1397–1412 (2020) https://doi.org/10.1109/TAC.2019.2926570

[6] Couillet, R., Perlaza, S.M., Tembine, H., Debbah, M.: Electrical vehicles in the smart grid: a mean field game analysis. IEEE Journal on Selected Areas in Communications **30**(6), 1086–1096 (2012)

[7] Lachapelle, A., Wolfram, M.-T.: On a mean field game approach modeling congestion and aversion in pedestrian crowds. Transportation Research Part B: Methodological **45**(10), 1572–1589 (2011)

[8] Bensoussan, A., Frehse, J., Yam, P.: Mean Field Games and Mean Field Type Control Theory. Springer, New York, NY (2013). https://doi.org/10.1007/978-1-4614-8508-7

[9] Achdou, Y., Capuzzo-Dolcetta, I.: Mean field games: numerical methods. SIAM Journal on Numerical Analysis **48**(3), 1136–1162 (2010)

[10] Carlini, E., Silva, F.J.: A semi-Lagrangian scheme for a degenerate second order mean field game system. Discrete and Continuous Dynamical Systems **35**(9), 4269–4292 (2015)

[11] Yang, Y., Luo, R., Li, M., Zhou, M., Zhang, W., Wang, J.: Mean field multi-agent reinforcement learning. In: Proc. International Conference on Machine Learning (ICML), pp. 5571–5580 (2018)

[12] Subramanian, S.G., Mahajan, A.: Reinforcement learning in stationary mean-field games. In: Proc. International Conference on Autonomous Agents and Multiagent Systems (AAMAS), pp. 251–259 (2019)

[13] Subramanian, S.G., Taylor, M.E., Crowley, M., Poupart, P.: Decentralized mean field games. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 9439–9447 (2022). https://doi.org/10.1609/aaai.v36i9.21176

[14] Carmona, R., Laurière, M.: Linear-quadratic mean field games of controls with absorption. ESAIM: Proceedings and Surveys **65**, 58–76 (2019)

[15] Aurell, A., Carmona, R., Dayanıklı, G., Laurière, M.: Finite state graphon games with applications to epidemics. Dynamic Games and Applications **12**, 49–81 (2022)

[16] Moon, J., Başar, T.: Linear quadratic risk-sensitive and robust mean field games. IEEE Transactions on Automatic Control **62**(3), 1062–1077 (2017) https://doi.org/10.1109/TAC.2016.2579264

[17] Cui, K., Koeppl, H.: Approximately solving mean field games via entropy-regularized deep reinforcement learning. In: Proc. International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 1909–1917 (2021)

[18] Zaman, M.A., Zhang, K., Miehling, E., Başar, T.: Reinforcement learning in non-stationary discrete-time linear-quadratic mean-field games. In: Proceedings of the IEEE Conference on Decision and Control (CDC), pp. 2278–2284 (2020)

[19] McMahan, H.B., Moore, E., Ramage, D., Hampson, S., Arcas, B.: Communication-efficient learning of deep networks from decentralized data. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 1273–1282 (2017)

[20] Li, T., Sahu, A.K., Talwalkar, A., Smith, V.: Federated learning: challenges, methods, and future directions. IEEE Signal Processing Magazine **37**(3), 50–60 (2020)

[21] Reisizadeh, A., Mokhtari, A., Hassani, H., Jadbabaie, A., Pedarsani, R.: FedPAQ: a communication-efficient federated learning method with periodic averaging and quantization. In: Proc. International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 2021–2031 (2020)

[22] Agarwal, N., Suresh, A.T., Yu, F., Kumar, S., McMahan, H.B.: cpSGD: communication-efficient and differentially-private distributed SGD. In: Proc. Advances in Neural Information Processing Systems (NeurIPS), pp. 7575–7586 (2018)

[23] Nadiger, C., Kumar, A., Abdelhak, S.: Federated reinforcement learning for fast personalization. In: Proc. IEEE International Conference on Consumer Electronics (ICCE), pp. 123–127 (2020)

[24] Khodadadian, S., Sharma, P., Joshi, G., Maguluri, S.T.: Federated reinforcement learning: linear speedup under Markovian sampling. In: Proc. International Conference on Machine Learning (ICML), pp. 10997–11057 (2022)

[25] Jin, H., Peng, Y., Yang, W., Wang, S., Zhang, Z.: Federated reinforcement

learning with environment heterogeneity. In: Proc. International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 18–37 (2022)

[26] Zheng, L., Yang, J., Cai, H., Zhou, M., Zhang, W., Wang, J., Yu, Y.: MAgent: a many-agent reinforcement learning platform for artificial collective intelligence. In: Proc. AAAI Conference on Artificial Intelligence, pp. 8222–8223 (2018)

[27] Balduzzi, D., Tuyls, K., Perolat, J., Graepel, T.: Re-evaluating evaluation. In: Proc. Advances in Neural Information Processing Systems (NeurIPS), pp. 3272–3283 (2018)

[28] Lanctot, M., Zambaldi, V., Gruslys, A., Lazaridou, A., Tuyls, K., Pérolat, J., Silver, D., Graepel, T.: A unified game-theoretic approach to multiagent reinforcement learning. Advances in Neural Information Processing Systems **30** (2017)