Zero-Sum Games Between Large-Population Teams under Mean-Field Sharing

Panagiotis Tsiotras

School of Aerospace Engineering
Institute for Robotics and Intelligent Machines
Georgia Institute of Technology

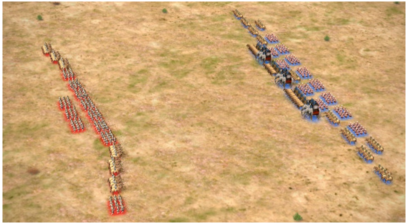
Workshop on Large Population Teams: Control, Equilibria & Learning
63rd IEEE Conference on Decision and Control
Milan, Italy, Dec. 15, 2024

Outline

Motivation

Large-Population Multi-Agent Interactions

- Mixed collaborative-competitive setting with large number of agents
  - Team level: competition
  - Within each team: collaboration
- Battlefield offense-defense, swarm robotics, sports



Key Challenges

- **Scalability:** Complexity increases exponentially as the number of agents increase
- Solution must respect the underlying **information structure**
- Information about the **opponents** is often unknown

Mean-Field Team Games

Problem Setting

- Zero-sum finite horizon problem with simultaneous moves
- Finite state and action spaces
- Each team (Blue and Red) consists of N_i **homogeneous** agents ($i =$ Blue, Red)
- Agents interact via weak coupled dynamics (transitions and rewards only depend on the state distributions)

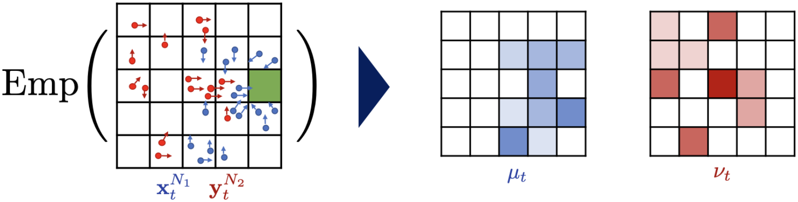


Figure: State distribution at time t is the empirical distribution μ_t and ν_t

Information Structure

Mean-Field Information Sharing Structure

- Each agent observes its own state (local information) and the empirical distribution (common information) of both its team and the opponent team
- We consider mixed Markov policies for each agent:

$$\phi_{i,t} : \mathcal{U} \times \mathcal{X} \times \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}) \to [0,1],$$
$$\psi_{j,t} : \mathcal{V} \times \mathcal{Y} \times \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}) \to [0,1],$$

where $\phi_{i,t}(u|x_{i,t}, \mu_t, \nu_t)$ is the probability that Blue agent i selects action u given its local state $x_{i,t}$ and the team EDs μ_t and ν_t; similarly for the Red agent $\psi_{j,t}(v|y_{j,t}, \mu_t, \nu_t)$

Notation

Individual Blue agent strategy $\phi_i = \{\phi_{i,t}\}_{t=0}^T$

Blue team strategy $\phi^{N_1} = \{\phi_i\}_{i=1}^{N_1}$

Individual Red agent strategy $\psi_j = \{\psi_{j,t}\}_{t=0}^T$

Red team strategy $\psi^{N_2} = \{\psi_j\}_{j=1}^{N_2}$

Information Structure

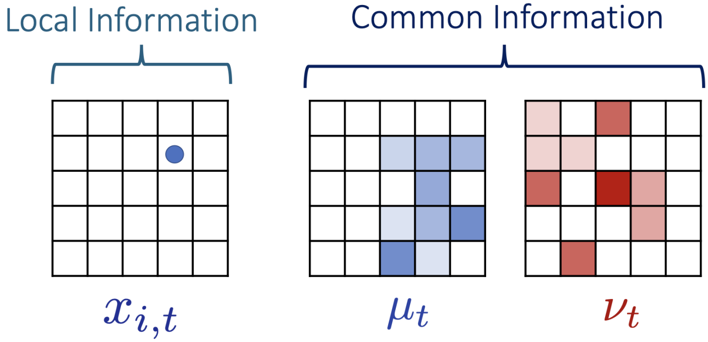


Figure: Information structure in a 2D grid world

Optimization

- We let the Blue team **maximize** and the Red team **minimize** (general zero-sum game)
- Performance of team strategy pair (ϕ^{N_1}, ψ^{N_2}) is given by the expected cumulative reward

$$J^{N,\phi^{N_1},\psi^{N_2}}(\mathbf{x}_0^{N_1}, \mathbf{y}_0^{N_2}) = \mathbb{E}_{\phi^{N_1},\psi^{N_2}}\left[\sum_{t=0}^{T} r_t(\mu_t, \nu_t) \mid \mu_0, \nu_0\right]$$

Objective

When Blue team considers its worst-case performance, we have the max-min optimization:

$$\underline{J}^{N*}(\mathbf{x}_0^{N_1}, \mathbf{y}_0^{N_2}) = \max_{\phi^{N_1}\in\Phi^{N_1}} \min_{\psi^{N_2}\in\Psi^{N_2}} J^{N,\phi^{N_1},\psi^{N_2}}(\mathbf{x}_0^{N_1}, \mathbf{y}_0^{N_2})$$

ϕ^{N_1}, ψ^{N_2} are the team strategies and $\underline{J}^{N*}$ is lower game value for the *finite-population game.*[a]

[a] Allows agents apply different strategies, especially the opponent red agents

Identical Team Strategies

The set of identical team strategies $\phi_{i,t} = \phi_t \ \forall i = 1, \dots, N_1$ is rich enough to approximate team behaviors induced by non-identical team strategies when team size is large

Approximation Lemma (Informal)

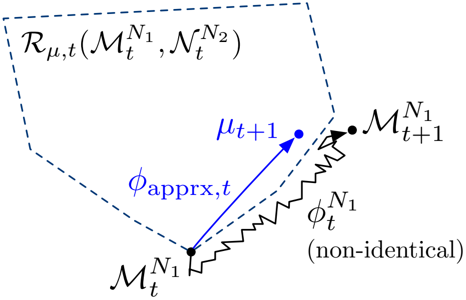
Given a non-identical team strategy $\phi_t^{N_1}$ there exists an identical team strategy ϕ_t such that the distribution μ_{t+1} induced by ϕ_t is close to the empirical distribution $\mu_{t+1}^{N_1}$ induced by $\phi_t^{N_1}$

$$\mathbb{E}_{\phi_t}\left[\mathrm{d}_{\mathrm{TV}}\left(\mu_{t+1}^{N_1}, \mu_{t+1}\right)\right] \leq \mathcal{O}\left(\sqrt{\frac{1}{N_1}}\right)$$

For a finite set E, the total variation between two probability measures $\mu, \mu' \in \mathcal{P}(E)$ is given by

$$\mathrm{d}_{\mathrm{TV}}(\mu, \mu') = \frac{1}{2}\sum_{e \in E} \left|\mu(e) - \mu'(e)\right| = \frac{1}{2}\left\|\mu - \mu'\right\|_1$$

Intuition: A Reachability Result



- It suffices for the Blue team to approximate all possible future ED outcomes using the mean-fields within the reachable set
- There exists a MF within the reachable set that is ϵ-close to the (*finite-population*) ED induced by that team policy

$$\mathcal{R}_{\mu,t}(\mu_t, \nu_t) = \{\mu_{t+1} \mid \exists \phi_t \in \Phi_t \text{ s.t. } \mu_{t+1} = \mu_t F_t(\mu_t, \nu_t, \phi_t)\}$$

Zero-Sum Coordinator Game

- Construct equivalent system, where the **state distributions** act as the **common information** to generate Blue (α) and Red (β) coordination strategies
- These coordination strategies **select local policies** π_t and σ_t that only depend on the agent's individual state
- One-to-one correspondence between identical team and coordination strategies

$$\phi_t(u|x,\mu,\nu) = \underbrace{\alpha_t(\mu,\nu)}_{\pi_t}(u|x)$$

$$\psi_t(v|y,\mu,\nu) = \underbrace{\beta_t(\mu,\nu)}_{\sigma_t}(v|y)$$

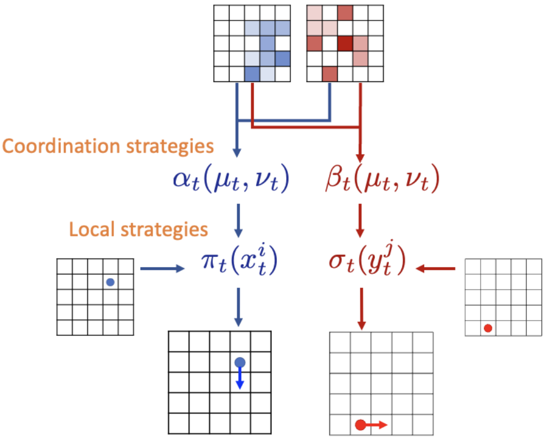
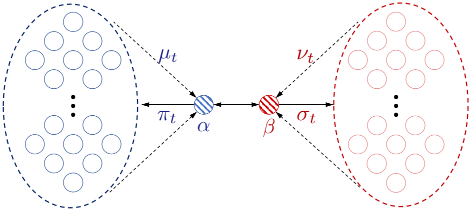


Figure: Illustration of the coordinator game

Zero-Sum Coordinator Game

Problem Setting

- Zero-sum infinite population game
- Players: Blue and Red coordinators
- States: Distributions μ_t and ν_t
- Actions: Local policies π_t and σ_t
- Strategies: Coordination strategies α and β
- Dynamics: Deterministic (Law of Large Numbers)



$$\mu_{t+1} = \mu_t F_t(\mu_t, \nu_t, \alpha_t)$$
$$\nu_{t+1} = \nu_t G_t(\mu_t, \nu_t, \beta_t)$$

Can be solved using:

- Dynamics Programming
- Reinforcement Learning

$$\underline{J}_{\text{coord}}(\mu_0, \nu_0) = \max_{\alpha \in \mathcal{A}} \min_{\beta \in \mathcal{B}} \sum_{t=0}^{T} r_t(\mu_t, \nu_t)$$

Performance Guarantees

Main Result

The optimal Blue coordination strategy α^* obtained from the infinite-population coordinator game induces an ϵ-optimal Blue team strategy for the finite-population game

$$\underline{J}^{N*}(\mathbf{x}^{N_1}, \mathbf{y}^{N_2}) \geq \min_{\psi^{N_2} \in \Psi^{N_2}} J^{N, \alpha^*, \psi^{N_2}}(\mathbf{x}^{N_1}, \mathbf{y}^{N_2}) \geq \underline{J}^{N*}(\mathbf{x}^{N_1}, \mathbf{y}^{N_2}) - \mathcal{O}\left(\sqrt{1/\underline{N}}\right)$$

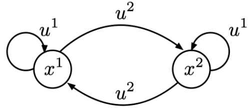
for all $\mathbf{x}^{N_1} \in \mathcal{X}^{N_1}$ and $\mathbf{y}^{N_2} \in \mathcal{Y}^{N_2}$ where $\underline{N} = \min\{N_1, N_2\}$

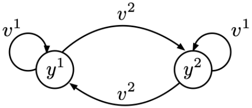
Key Takeaways:

- We can solve the mean-field team game assuming identical team strategies
- Even if opponent employs a non-identical strategy to exploit our identical strategy, the performance degradation is within a bound from the best attainable performance
- The error diminishes as the size of the team population increases

- Two-state example (T=2)   At $t=0$, all Red agents are frozen, at $t=1$, y^2 is absorbing

$$r_0(\mu,\nu) = r_1(\mu,\nu) = 0 \quad \forall \mu \in \Delta_{|\mathcal{X}|}, \nu \in \Delta_{|\mathcal{Y}|},$$
$$r_2(\mu,\nu) = -\nu(y^2)$$





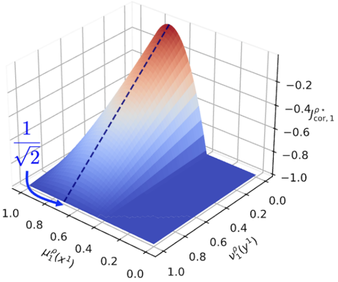
Incentivizes Red agents to move to y^2 using v^2 with transition probability

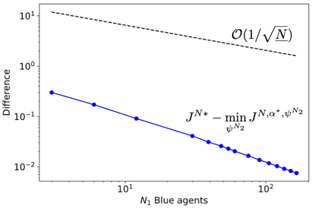
$$\min\left\{5\Big((\mu(x^1)-\frac{1}{\sqrt{2}})^2 + (\mu(x^2)-(1-\frac{1}{\sqrt{2}}))^2\Big), 1\right\}$$

- Optimal Blue team strategy is to match distribution $[1/\sqrt{2}, 1-1/\sqrt{2}]$
  - Feasible only in infinite-population case

Finite-population Blue optimal strategy is **non-identical**

Transition probability from y^1 to y^2 is 0.016 (non-identical) vs 0.518 (identical)

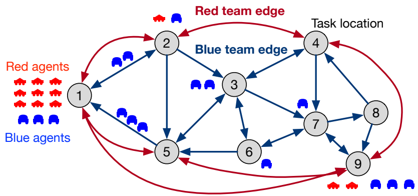




Blue agent either deterministically stays at its current state or deterministically moves to the other state

Application to MARL

- State-of-the-art multi-agent reinforcement learning (MARL) algorithms like MADDPG (Lowe et al., 2017) fail to scale in situations where the number of agents becomes large
- Complexities due to the training of individual policies for each agent
- Parameter sharing (Li et al. 2024) can help, but NN uses all states and actions of all agents
- MF theory approximations show promising results (Cui et al. 2023; Yardim and He 2024)



Salient Features of MF-MAPPO

MF-MAPPO: Mean-Field Multi-Agent Proximal Policy Optimization

- Only requires common information in order to learn the value function (minimally informed critic network)
  - Network complexity does not scale with the number of agents
  - Input to the network is much smaller in size compared to Q-function based methods (do not require the actions as an input)
- Simultaneous training and updates of both competing teams instead of an alternating optimization
- Train for N agents, deploy for $M \neq N$

Training

Common Information-Based Critic Network

- For $j \in \{\text{Blue}, \text{Red}\}$, critic $V_j(\mu, \nu)$ is parametrized by ζ_j
- Objective:

$$L_{\text{critic}}(\zeta_j) = \frac{1}{|B|} \Sigma_{\tau \in B} \Big(V_j(\mu, \nu; \zeta_j) - \hat{R}^j \Big)^2,$$

  where τ sampled from the mini-batch of size B and $\hat{R}^j$ is the discounted reward-to-go

$$\hat{R}^j_t = \sum_{t'=t}^{\infty} \gamma^{t'-t} r^j_t(\mu_t, \nu_t), \quad j \in \{\text{Blue}, \text{Red}\}$$

- Recall that $r^{\text{Red}}_t = -r^{\text{Blue}}_t$

Mean-Field Multi-Agent Proximal Policy Optimization

Actor Network

- Blue team with identical team strategy $\phi_{\theta_{\text{Blue}}}$ parametrized by θ_{Blue}
- Objective:

$$L(\theta_{\text{Blue}}) = \frac{1}{|B|}\sum_{k=1}^{B}\left[\min\Big(g_k A_k,\ \text{clip}(g_k, 1-\epsilon, 1+\epsilon)A_k\Big) + \omega S(\phi_{\theta_{\text{Blue}}}(x_k, \mu_k, \nu_k))\right],$$

$$\text{where } g = g(\theta) = \frac{\phi_\theta(u|x,\mu,\nu)}{\phi_{\theta^{\text{old}}}(u|x,\mu,\nu)}$$

A_k is the GAE function sampled at time t from a trajectory with a T-step rollout $A_k = \delta_t + (\gamma\lambda)\delta_{t+1} + \cdots + \cdots + (\gamma\lambda)^{T-t+1}\delta_{T-1}$[a] and ω weighs the contribution of the entropy term S and decays during training

[a] $\delta_t = r_{t,\text{Blue}} + \gamma V_{\text{Blue}}(\mu_{t+1}, \nu_{t+1}) - V_{\text{Blue}}(\mu_t, \nu_t)$

Constrained Rock-Paper-Scissors (cRPS)

For both teams we have:

- State space $\mathcal{S} = \{s_0, s_1, s_2\}$
- Action space $\mathcal{A} = \{a_0, a_1\}$
- Deterministic transitions
- Reward at each time step $r_t^{\text{Blue}} = -r_t^{\text{Red}} = \boldsymbol{\mu}_t^T A \boldsymbol{\nu}_t$, where

$$A = \begin{bmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{bmatrix}$$

Equilibrium Distribution

$$\boldsymbol{\mu}^* = \boldsymbol{\nu}^* = \left[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right]$$

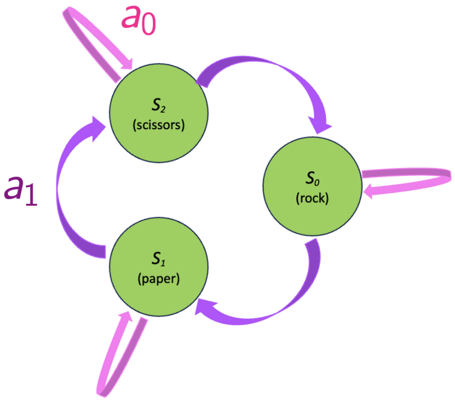
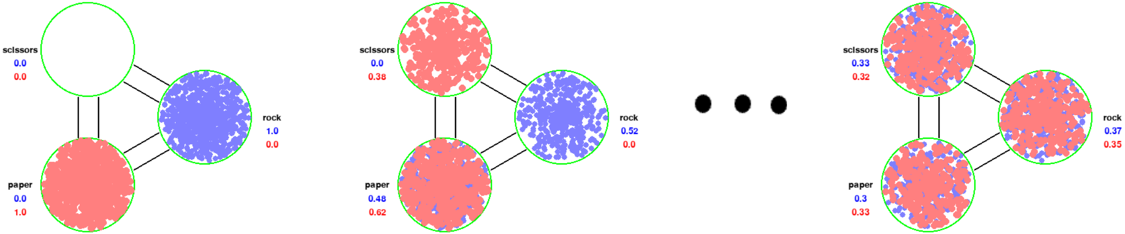


Figure: States and Actions for cRPS

Constrained Rock-Paper-Scissors (cRPS)



Observations

- For this initial condition, the equilibrium distribution is not reachable at $t = 1$
- Algorithm effectively explores and learns to achieve the distribution $[1/3, 1/3, 1/3]$ (Note: $N_1 = N_2 = 1,000$)

Mixed Collaborative-Competitive Battlefield with Target Capture

- Battlefield on a 2D grid world
  - Blue team: reach the target location(s)
  - Red team: defend target(s)
- Zero-sum game where $r_t^{\text{Blue}} = -r_t^{\text{Red}} \propto$ fraction of blue population at target
- Agent state is its position and status
- Teams must learn to remain alive, not be deactivated by the opponent team's agents (based on relative numerical advantage at each cell) and circumnavigate obstacles
- Agent observation: local position and state distributions of both teams
- Here, $N_1 = N_2 = 100$

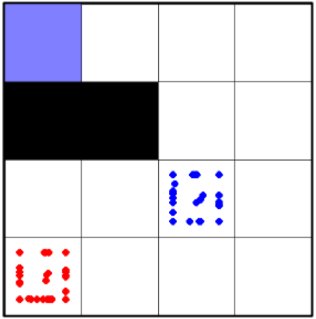
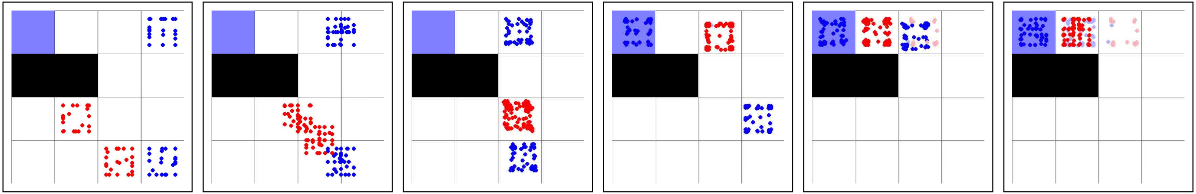


Figure: Example of a 4×4 Map with 1 Target (Lilac Square) and 2 obstacles (Black)

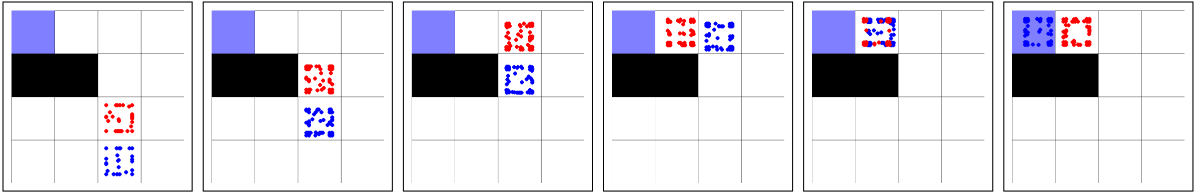
Mixed Collaborative-Competitive Battlefield with Target Capture



Observations

- The Blue agents learn to reach the target
- The Red team learns to assemble and position itself at the target and deactivate remaining Blue agents

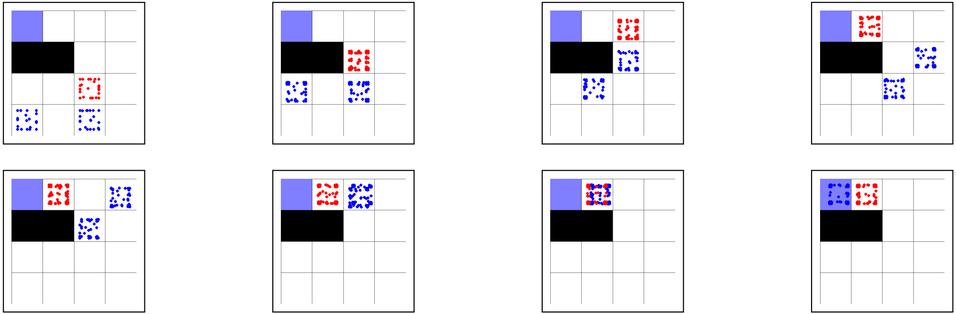
Mixed Collaborative-Competitive Battlefield with Target Capture



Observations

- The Blue agents move as a blob so that not be deactivated by the Red team (since numerical advantage $= 0$)

Mixed Collaborative-Competitive Battlefield with Target Capture



Observations

- The Blue agents, upon seeing the state distribution, learn to change their path and combine with the remaining agents in order to ”push through” to the target

Animations

Conclusions and Future Work

Summary

- **Zero-sum mean-field team games** with weakly coupled dynamics: mixed collaborative and competitive
- **Common information approach** with **mean field sharing** $\Rightarrow$ equivalent coordinator game
- Identical team strategies with **theoretical performance guarantees** (performance improves as the number of agents increase)
- Novel **common information critic** based MARL algorithm for solving large scale real-world complex team games

Future Work

- More realistic and complex scenarios
- Limited information/partial observability
- Heterogeneous agents and sub-team roles/behaviors