Decentralized Value-Based Reinforcement Learning in Stochastic Potential Games

M.Sc. Thesis, Data Science and Engineering

Hafez Ghaemi
Supervisors: Prof. Fabio Fagnani, Prof. Giacomo Como

Politecnico di Torino

July 2022

Table of Contents

1. Motivation
   - Multi-Agent Systems and Stochastic Games
   - Decentralized Multi-Agent Reinforcement Learning
2. Theoretical Background
   - Game Theory
   - Stochastic Games and MARL
   - Stochastic Potential Game (SPG)
3. Learning Dynamics
4. Experimental Results
   - Single-State Potential Games
   - Network Routing Congestion Games
   - Distancing Game
5. Convergence Analysis
6. Discussion, Conclusion, and Future Works

Table of Contents

1 Motivation
- Multi-Agent Systems and Stochastic Games
- Decentralized Multi-Agent Reinforcement Learning

2 Theoretical Background
- Game Theory
- Stochastic Games and MARL
- Stochastic Potential Game (SPG)

3 Learning Dynamics

4 Experimental Results
- Single-State Potential Games
- Network Routing Congestion Games
- Distancing Game

5 Convergence Analysis

6 Discussion, Conclusion, and Future Works

]

Multi-Agent Systems

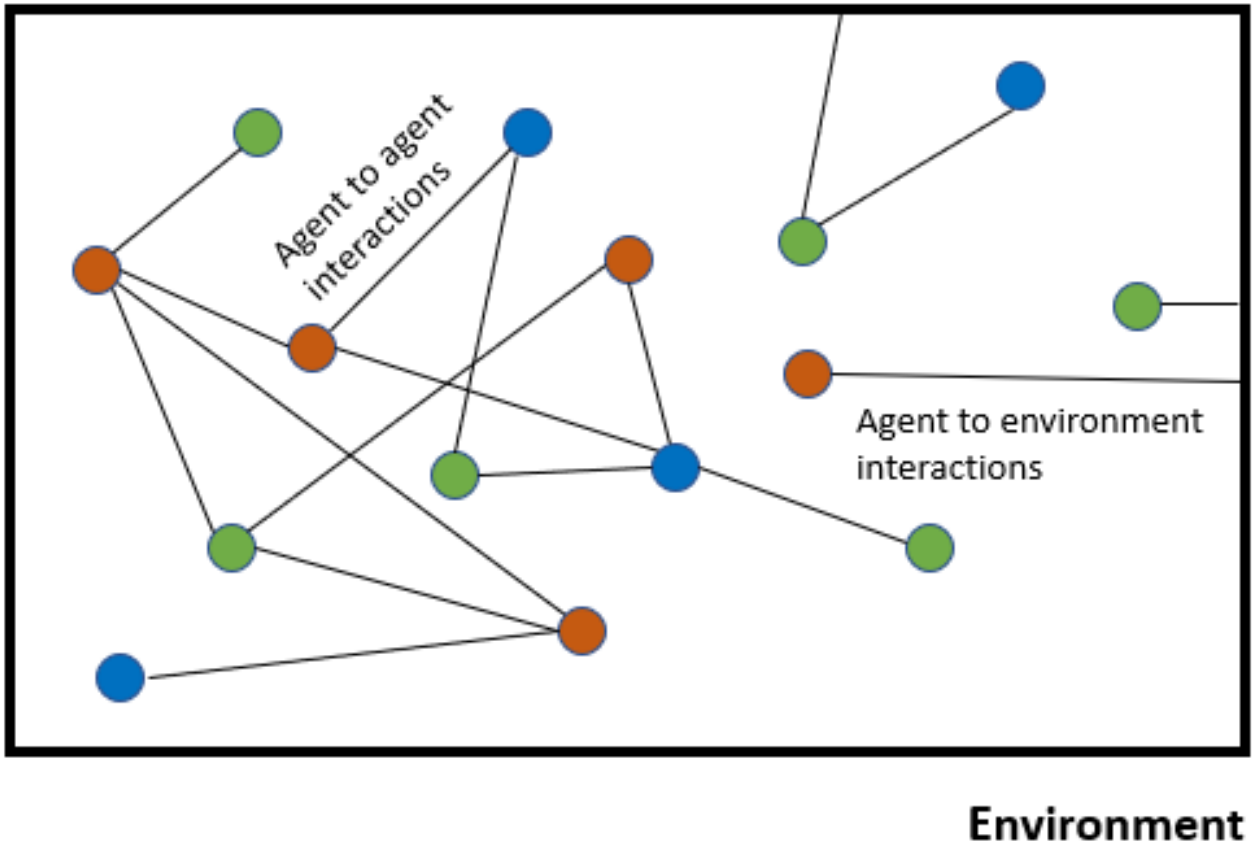


Figure: source: https://tinyurl.com/yeysd2ws

Multi-Agent Systems

Network of self-driving cars

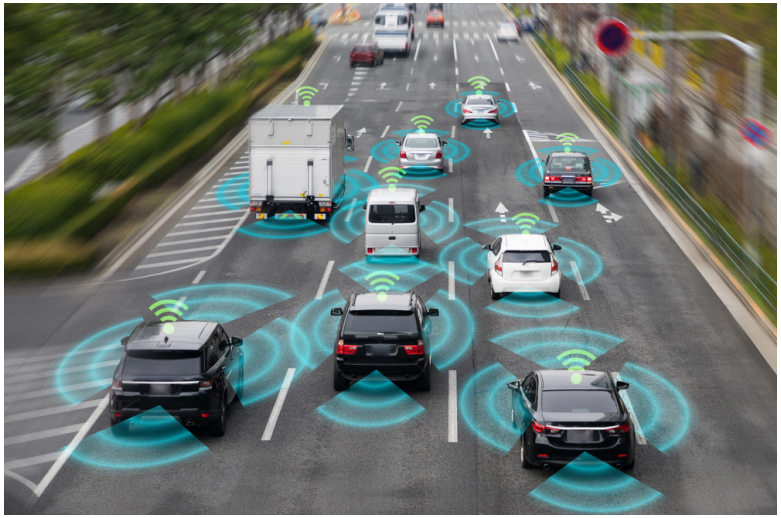


Figure: source: https://tinyurl.com/5c7mnt5x

Multi-robot systems

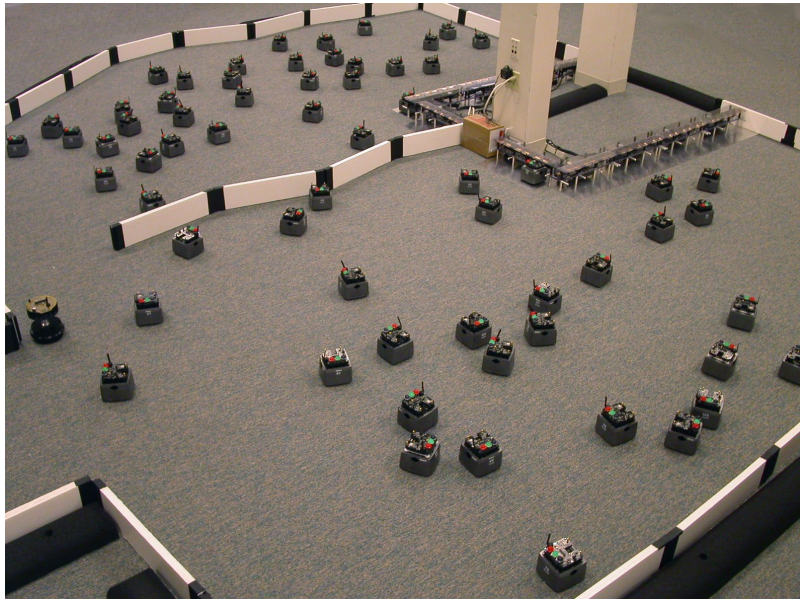
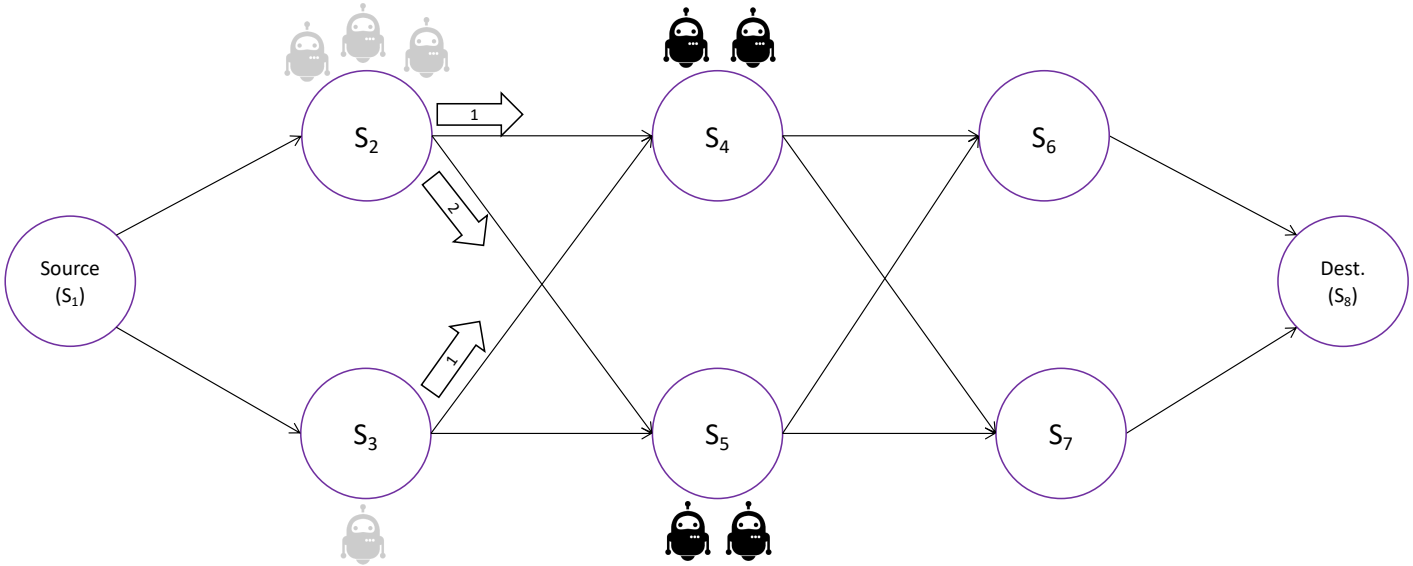


Figure: source: https://tinyurl.com/ymkupj9c

Network Congestion Games

Optimal routing in transportation or computer networks



The utility/reward functions are dynamic. Therefore, instead of a static game, we can model the problem by a stochastic game (SG).

Stochastic Potential Games

Level of competition in games:

- Fully-competitive games (zero-sum)

Stochastic Potential Games

Level of competition in games:

- Fully-competitive games (zero-sum)
- Fully-cooperative games (identical interests)

Stochastic Potential Games

Level of competition in games:

- Fully-competitive games (zero-sum)
- Fully-cooperative games (identical interests)
- Mixed-motive and general-sum games

Stochastic Potential Games

- Potential games are a class of mixed-motive games. The agents are self-interested, but because a global potential function can express the incentive of all players to change their strategy, a cooperative motive is hidden inside the game.

Stochastic Potential Games

- Potential games are a class of mixed-motive games. The agents are self-interested, but because a global potential function can express the incentive of all players to change their strategy, a cooperative motive is hidden inside the game.
- We focus on stochastic potential games (SPG) that can model many real-world scenarios, such as network congestion games.

MARL to Solve Stochastic Games

- Usually, in SGs, the environement dynamics are not a priori known and we cannot use optimization-based methods to find optimal strategies in each state.

MARL to Solve Stochastic Games

- Usually, in SGs, the environement dynamics are not a priori known and we cannot use optimization-based methods to find optimal strategies in each state.
- Furthermore, since the game is not static, we cannot use game-theoretic approaches, such as fictitious play, that rely on playing a single game repeatedly.

MARL to Solve Stochastic Games

- Usually, in SGs, the environement dynamics are not a priori known and we cannot use optimization-based methods to find optimal strategies in each state.
- Furthermore, since the game is not static, we cannot use game-theoretic approaches, such as fictitious play, that rely on playing a single game repeatedly.
- Therefore, adopting a multi-agent reinforcement learning (MARL) approach where agents learn the game's state and reward dynamics by trial and error is a feasible solution. In fact, SG is the main framework for formulating MARL.

Centralized vs. Decentralized MARL

Categorization of MARL algorithms in terms of centrality

- Centralized/coordinated MARL
  - Usually needs a simulator of the game.
  - Problems with privacy, scalibility, and computational cost

Centralized vs. Decentralized MARL

Categorization of MARL algorithms in terms of centrality

- Centralized/coordinated MARL
  - Usually needs a simulator of the game.
  - Problems with privacy, scalibility, and computational cost
- Decentralized/independent MARL
  - The main challenge is non-stationarity of the environement.
  - The evolution of the environment is dependent on the changing behavior of all agents. Therefore, the environment is highly non-stationary from an agent's prespective that only has access to local observations.
  - It can be addressed by either communication between agents, adaptive learning rates, or two-timescale learning.

Contribution and Gap Statements

Contribution and Gap Statements

Considering their real-world applications, we focus on solving SPGs using decentralized MARL. Up to now, only policy-based MARL methods that require gradient calculations have been applied to SPGs. In this work, we consider decentralized Q-learning as a value-based MARL algorithm, and show its convergence to Nash equilibrium in SPGs. Empirically, two SPG benchmarks, network routing games, and distancing games are used for performance evaluation. Theoretically, we prove the asymptotic convergence to a Nash equilibrium of the SPG under an initialization assumption. Based on our results, decentralized Q-learning can be considered the first game-agnostic value-based MARL algorithm.

Table of Contents

1 Motivation
- Multi-Agent Systems and Stochastic Games
- Decentralized Multi-Agent Reinforcement Learning

2 Theoretical Background
- Game Theory
- Stochastic Games and MARL
- Stochastic Potential Game (SPG)

3 Learning Dynamics

4 Experimental Results
- Single-State Potential Games
- Network Routing Congestion Games
- Distancing Game

5 Convergence Analysis

6 Discussion, Conclusion, and Future Works

]

Strategic Games

Definition. Strategic Game

A strategic game $\mathcal{G}$ can be characterized using a tuple $(\mathcal{N}, \mathcal{A} = \{\mathcal{A}_i | i \in N\}, U = \{u_i | i \in N\})$, where $\mathcal{N}$ is the set of players, $\mathcal{A}_i$ is the set of possible actions for player i, and $u_i(a^i, a^{-i})$ is the utility/reward function of player i.

Table: Two-player prisoner's dilemma

P2 P1	Confess	Don't confess
Confess	3, 3	0, 4
Don't confess	4, 0	1, 1

Mixed Strategies and Nash Equilibrium

Mixed utility:

$$U_i(\pi^i, \pi^{-i}) = (\pi^i)^T u^i \prod_{j \in N, j \neq i} \pi^j = \mathbb{E}_{(a^i, a^{-i}) \sim (\pi^i, \pi^{-i})}\{u_i(a^i, a^{-i})\} \quad (1)$$

Definition. Nash Equilibrium

A Nash equilibrium profile (π^i_*, π^{-i}_*) has the property

$$U_i(\pi^i_*, \pi^{-i}_*) \geq U_i(\pi^i, \pi^{-i}_*), \quad \forall i, \pi^i. \quad (2)$$

Every finite game has at least one Nash equilibrium. Deterministic $(\pi^i_*, \pi^{-i}_*) = (a^i_*, a^{-i}_*)$ corresponds to a deterministic Nash equilibrium.

Nash equilibrium

Every deterministic Nash equilibrium is a point where each player has played her best response with respect to others,

$$\mathcal{B}_i(a_{-i}) := \underset{a_i \in A_i}{\operatorname{argmax}}\, u_i(a_i, a_{-i}) \tag{3}$$

Table: Nash equilibria in two-player prisoner's dilemma

P2 P1	Confess	Don't confess
Confess	**3, 3**	0, 4
Don't confess	4, 0	**1, 1**

Stochastic Game as an MDP

Definition. Stochastic Game

A stochastic (Markov) game is a Markov decision process (MDP) with five key elements $(\mathcal{N}, \mathcal{S}, \mathcal{A} = \{\mathcal{A}_i | i \in N\}, \mathcal{P}, R = \{R_i | i \in \mathcal{N}\})$. $\mathcal{N}$ is the set of players, $\mathcal{S}$ is the set of states, $\mathcal{A}_i$ is the set of possible actions of player i, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition probabilities of agents, and $\mathcal{R} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to R$ is their reward functions.
Based on this definition, SPG is a set of states, each of which is a normal-form game, and the transition probabilities dictate which game state should be played at a given time step.

Nash Equilibrium in Stochastic Games

Utility function in a stochastic game that is played over the infinite-horizon:

$$U^i(\pi^i, \pi^{-i}) = \mathbb{E}_{(a_t^i, a_t^{-i}) \sim \pi} \left\{ \sum_{t=0}^{\infty} \gamma^t r_{s_t}^i(a^i, a^{-i}) \right\}, \tag{4}$$

where the policy $\pi^i \in \Delta(A^i)$ is a probability distribution over actions for each state and γ is the discount factor.

ϵ-Nash policy in SG:

$$U_i(\pi_*^i, \pi_*^{-i}) \geq U_i(\pi^i, \pi_*^{-i}) - \epsilon, \quad \forall i, \pi^i \in \Delta(A^i), \tag{5}$$

We can also define the Nash equilibrium using RL terminology.

MARL Terminology in Stochastic Games

Value function:

$$v_\pi^i(s) = \mathbb{E}_{(a^i, a^{-i}) \sim \pi} \left\{ r_s^i(a^i, a^{-i}) + \gamma \sum_{s' \in S} v_\pi^i(s') p(s' | s, a^i, a^{-i}) \right\}. \quad (6)$$

Therefore, $U_i(\pi^i, \pi^{-i}) = \mathbb{E}_{s_0 \sim p_{s_0}} \{ v(s_0) \}$.
Action-value (Q) function:

$$Q_\pi^i(s, a^i, a^{-i}) = r_s^i(a^i, a^{-i}) + \gamma \sum_{s' \in S} v_\pi^i(s') p(s' | s, a^i, a^{-i}). \quad (7)$$

ϵ-Nash policy formulation based on value and Q functions:

$$v^i_{\pi_*}(s) \geq v^i_{\pi}(s) - \epsilon \quad \forall \pi, i, s, \tag{8}$$

$$(\pi^{i,s}_*)^T Q^i_s \prod_{j \in N, j \neq i} \pi^{j,s}_* \geq (\pi^{i,s})^T Q^i_s \prod_{j \in N, j \neq i} \pi^{j,s}_* - \epsilon, \quad \forall \pi, i, s. \tag{9}$$

Value-based vs. Policy-based RL

Two Categories of Model-Free RL:

- Policy-based methods
  - Nowadays, they are mostly gradient-based and are also called policy-gradient methods.
  - Agent's policy is considered as a policy parameterized by θ and it is optimized using gradient descent,

$$\theta = \theta + \alpha \nabla_\theta \pi_\theta(s_t, a_t) g_t, \tag{10}$$

where g_t can be the current $Q^\pi(s_t, a_t)$ estimate.

Value-based vs. Policy-based RL

Two Categories of Model-Free RL:

- Policy-based methods
  - Nowadays, they are mostly gradient-based and are also called policy-gradient methods.
  - Agent's policy is considered as a policy parameterized by θ and it is optimized using gradient descent,

$$\theta = \theta + \alpha \nabla_\theta \pi_\theta(s_t, a_t) g_t, \tag{10}$$

  where g_t can be the current $Q^\pi(s_t, a_t)$ estimate.
- Value-based methods
  - They are based on iterative updates of the Q or value functions using temporal difference error.
  - Q-learning is a famous example,

$$Q(s,a) = Q(s,a) + \alpha[r + \gamma \max_{a' \in \mathcal{A}} Q(s',a') - Q(s,a)]. \tag{11}$$

Stochastic Potential Games

Definition. Stochastic potential game

A stochastic game is potential, if there exists a potential function $\Phi : \mathcal{S} \times \mathcal{A} \to R$ such that

$$r_i(s, a_i, a_{-i}) - r_i(s, a_i', a_{-i}) = \Phi(s, a_i, a_{-i}) - \Phi(s, a_i', a_{-i}), \quad \forall s, i, a^{-i}. \tag{12}$$

Furthermore, we say an SPG is state-transitive if

$$r_i(s, a_i, a_{-i}) - r_i(s', a_i, a_{-i}) = \Phi(s, a_i, a_{-i}) - \Phi(s', a_i, a_{-i}). \tag{13}$$

The state-transitivity is an intuitive property of many real-world SPGs, and we assume it is satisfied in our analysis.

Stochastic Potential Games

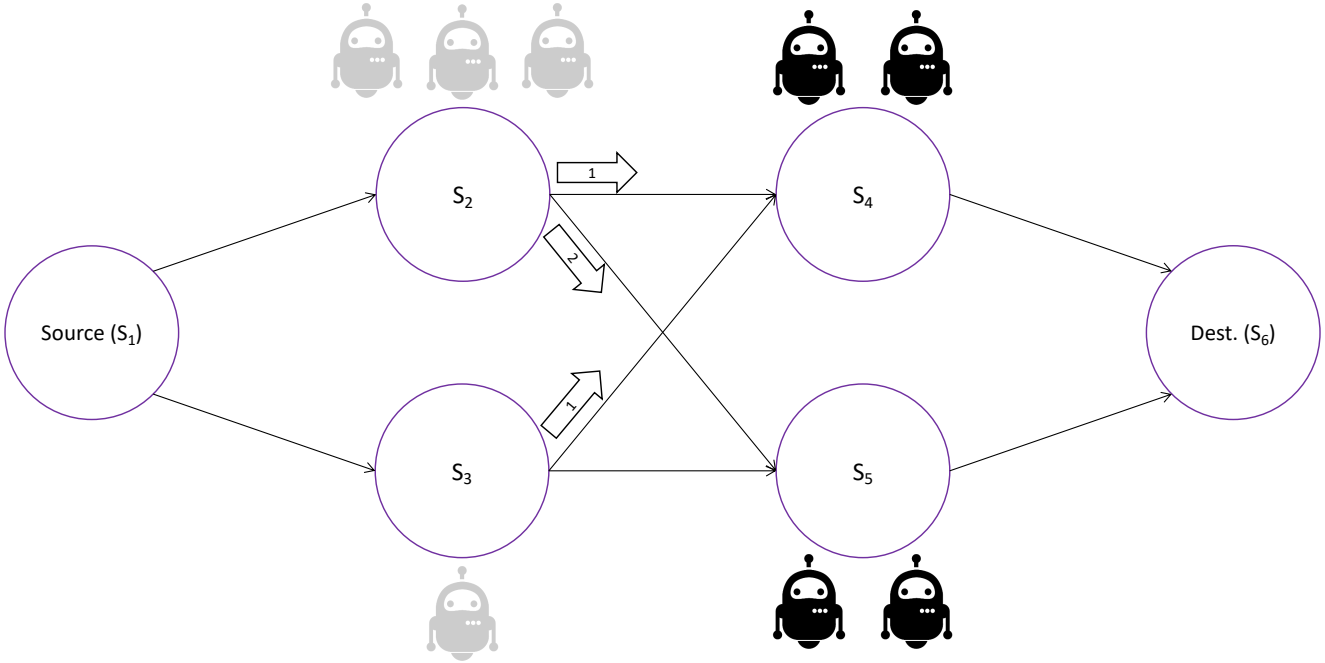
Lemma

If an SPG is state-transitive, Q and value functions of the game also admit potential functions [Mguni et al., 2021]:

$$v^i_\pi(s) - v^i_{\pi'}(s) = B(s)_\pi - B(s)_{\pi'}, \quad \forall i, s. \tag{14}$$

$$Q^i(s, a^{i'}, a^{-i}) - Q^i(s, a^i, a^{-i}) = G(s, a^{i'}, a^{-i}) - G(s, a^i, a^{-i}), \quad \forall i, s, a^{-i}. \tag{15}$$

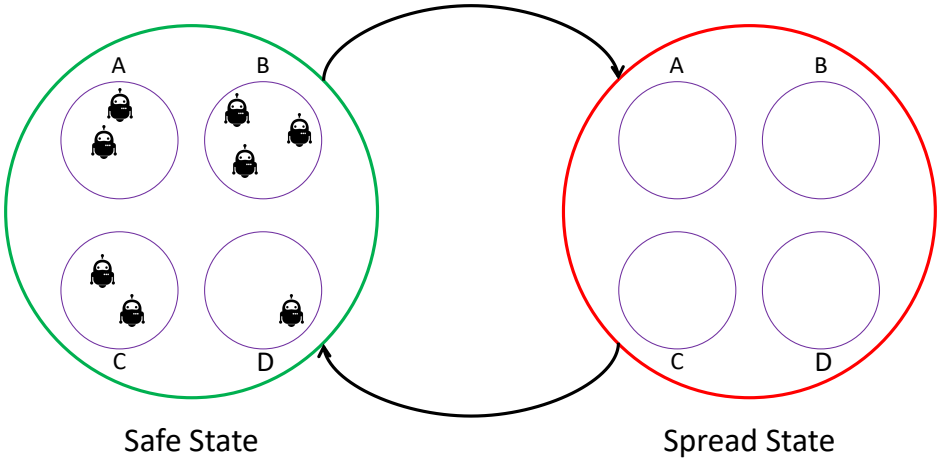
Benchmark: Network Routing Congestion Games



$$\Phi(s, a^i, a^{-i}) = \sum_e \sum_{k=1}^{n_e} c(k) \tag{16}$$

Benchmark: Distancing Game

$m = 4$ facilities (actions). A safe and a spread state. A positive reward proportional to the weight of their chosen facility ($w_D > w_C > w_B > w_A$) and the number of agents in that facility. With more than $N/2$ number of agents in a single facility, the agents will transit to the spread state, where they have to spread out and not be more than $N/4$ in the same facility to move back to the safe state, and avoid a negative large constant added to their reward.



Benchmark: Distancing Game

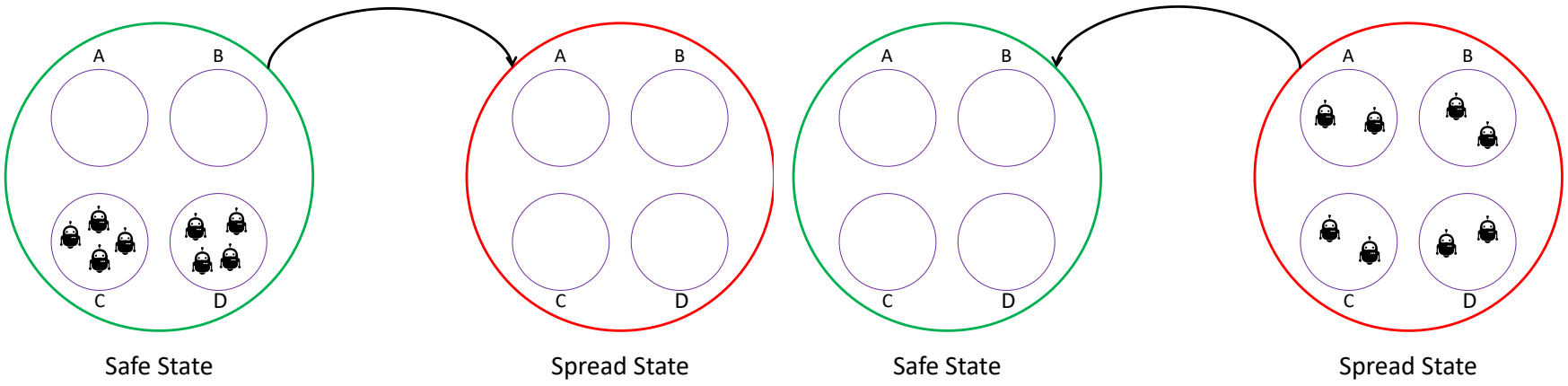


Figure: Optimum strategies in a distancing game

$$\Phi(s_{safe}, a^i, a^{-i}) = \sum_f \sum_{k=1}^{n_f} k w_f, \tag{17a}$$

$$\Phi(s_{spread}, a^i, a^{-i}) = \sum_f \sum_{k=1}^{n_f} k(w_f - c), \tag{17b}$$

Related Works

The SPG literature is relatively new and sparse, and the previous works have all adopted policy-based MARL techniques.

- Leonardos et al. [Leonardos et al., 2021] extended the policy gradient method [Agarwal et al., 2021] to a multi-player setting for SPGs.

Related Works

The SPG literature is relatively new and sparse, and the previous works have all adopted policy-based MARL techniques.

- Leonardos et al. [Leonardos et al., 2021] extended the policy gradient method [Agarwal et al., 2021] to a multi-player setting for SPGs.
- Zhang et al. [Zhang et al., 2021] proved the last iterate convergence of independent policy gradient (IPG) in SPGs.

Related Works

The SPG literature is relatively new and sparse, and the previous works have all adopted policy-based MARL techniques.

- Leonardos et al. [Leonardos et al., 2021] extended the policy gradient method [Agarwal et al., 2021] to a multi-player setting for SPGs.
- Zhang et al. [Zhang et al., 2021] proved the last iterate convergence of independent policy gradient (IPG) in SPGs.
- Fox et al. [Fox et al., 2022] showed the convergence of independent natural policy gradient (INPG), and empirically showed its faster convergence compared to IPG.

Related Works

The SPG literature is relatively new and sparse, and the previous works have all adopted policy-based MARL techniques.

- Leonardos et al. [Leonardos et al., 2021] extended the policy gradient method [Agarwal et al., 2021] to a multi-player setting for SPGs.
- Zhang et al. [Zhang et al., 2021] proved the last iterate convergence of independent policy gradient (IPG) in SPGs.
- Fox et al. [Fox et al., 2022] showed the convergence of independent natural policy gradient (INPG), and empirically showed its faster convergence compared to IPG.
- Mguni et al. [Mguni et al., 2021] proposed SPot-AC, an actor-critic algorithm that needs to first learn the potential function of SPG using minibatch data samples.

Table of Contents

1 Motivation
- Multi-Agent Systems and Stochastic Games
- Decentralized Multi-Agent Reinforcement Learning

2 Theoretical Background
- Game Theory
- Stochastic Games and MARL
- Stochastic Potential Game (SPG)

3 Learning Dynamics

4 Experimental Results
- Single-State Potential Games
- Network Routing Congestion Games
- Distancing Game

5 Convergence Analysis

6 Discussion, Conclusion, and Future Works

]

Decentralized Q-Learning

- A decentralized value-based MARL algorithm.

Decentralized Q-Learning

- A decentralized value-based MARL algorithm.
- First introduced for two-player zero-sum stochastic games [Sayin et al., 2021].

Decentralized Q-Learning

- A decentralized value-based MARL algorithm.
- First introduced for two-player zero-sum stochastic games [Sayin et al., 2021].
- Three important properties:
  - Two-timescale learning to solve the non-stationarity challenge in MARL; the slower update (with a smaller learning rate) of the value functions compared to Q-functions makes the game locally stationary for each agent.
  - The step sizes for Q-function updates are normalized to ensure that the entries corresponding to different actions in each state are updated with the same rate, even when they are not chosen with the same frequency.
  - Exploration is ensured by using a noisy best-response policy with a decaying temperature parameter (τ).

Decentralized Q-Learning

Because we do not have access to the global Q-function, we define a local Q-function that agents try to estimate:

$$q_\pi^i(s, a^i) = \mathbb{E}_{a^{-i} \sim \pi^{-i}} \left\{ Q_\pi^i(s, a^i, a^{-i}) \right\} \quad \forall i, s. \tag{18}$$

Agents follow a noisy best-response (BR) policy:

$$Br(q_s^i) = \underset{\mu \in \Delta(A)}{\operatorname{argmax}} \left\{ \mu q_s^i - \nu(\mu) \right\} \tag{19}$$

If ν is the negative of entropy function with noise level τ,

$$\nu(\mu, \tau) = \tau \sum_{a_j^i \in A_i} \mu_j \log(\mu_j) \tag{20}$$

Noisy BR will be the logit choice (softmax) function:

$$Br(q_s^i, \tau) = \frac{\exp(q_s^i/\tau)}{\sum_{a_j^i \in A_i} \exp(q_{j,s}^i/\tau)} \tag{21}$$

Decentralized Q-Learning Algorithm

$\#s$ is the number of visits to state s up to now. Used to decay step sizes, α and β, and temperature τ.

Require: Keep track of $\{q^i_{s,t},\ v^i_{s,t}, \#s\}$.

procedure SEPARATELY FOR EACH PLAYER i, RUN THE FOLLOWING UNTIL CONVERGENCE.

Choose a^i_{t-1} from current policy. Receive r^i_{t-1} and observe s_t.

$$q^i_{s_{t-1},t}[a^i_{t-1}] = q^i_{s_{t-1},t}[a^i_{t-1}] + \alpha_{\#s_{t-1}}(r^i_{t-1} + \gamma(v^i_{s_{t-1},t-1}) - q^i_{s_{t-1},t}[a^i_{t-1}]) \tag{22}$$

$$\pi^i_t = Br(q^i_{s_t}, \tau_{\#s_t}) = \frac{\exp(q^i/\tau)}{\sum_{a_j \in A} \exp(q^i_j/\tau)} \tag{23}$$

$$v^i_{s_t,t+1} = v^i_{s_t,t} + \beta_{\#s_t}(\pi^i_t \cdot q^i_{s_t,t} - v^i_{s_t,t}) \tag{24}$$

end procedure

Assumptions

Assumptions

a. $\{\alpha_t \in (0,1)\} and \{\beta_t \in (0,1)\}$ are non-increasing and satisfy $\lim_{t\to\infty} \beta_t = 0$ and $\sum_{t\geq 0} \alpha_t = \sum_{t\geq 0} \beta_t = \infty$

b. $\lim_{t\to\infty} \frac{\beta_{\lfloor mt \rfloor}}{\alpha_t} = 0$ for all $m \in (0,1)$.

c. Any state is reachable from others within a finite number of states with a positive probability (ergodicity of MDP)

d. The temperature sequence $\tau_{t>0}$ is non-increasing and satisfies

$$\lim_{t\to\infty} \frac{(\tau_{t+1} - \tau_t)}{\alpha_t} = 0, \quad and \quad \lim_{t\to\infty} \tau_t = 0 \tag{25}$$

Table of Contents

1 Motivation
- Multi-Agent Systems and Stochastic Games
- Decentralized Multi-Agent Reinforcement Learning

2 Theoretical Background
- Game Theory
- Stochastic Games and MARL
- Stochastic Potential Game (SPG)

3 Learning Dynamics

4 Experimental Results
- Single-State Potential Games
- Network Routing Congestion Games
- Distancing Game

5 Convergence Analysis

6 Discussion, Conclusion, and Future Works

Single-State Potential Games

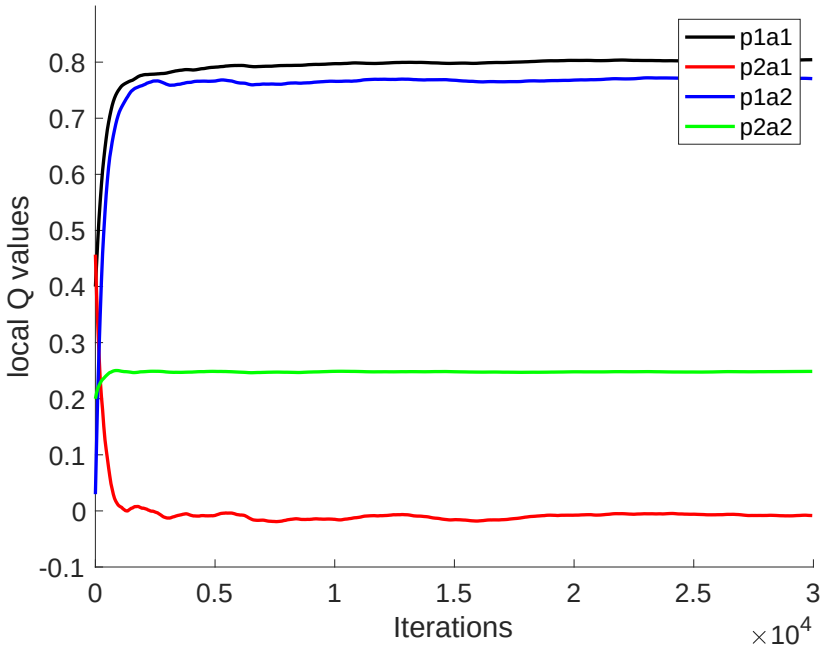
]

Two-player Single-State Potential Game with Two Actions

P2 / P1	a1	a2
a1	0.2, -0.3	**0.4, 0.1**
a2	**0.5, 0.2**	0.2, 0.1

Φ

0.1	**0.5**
0.4	0.3

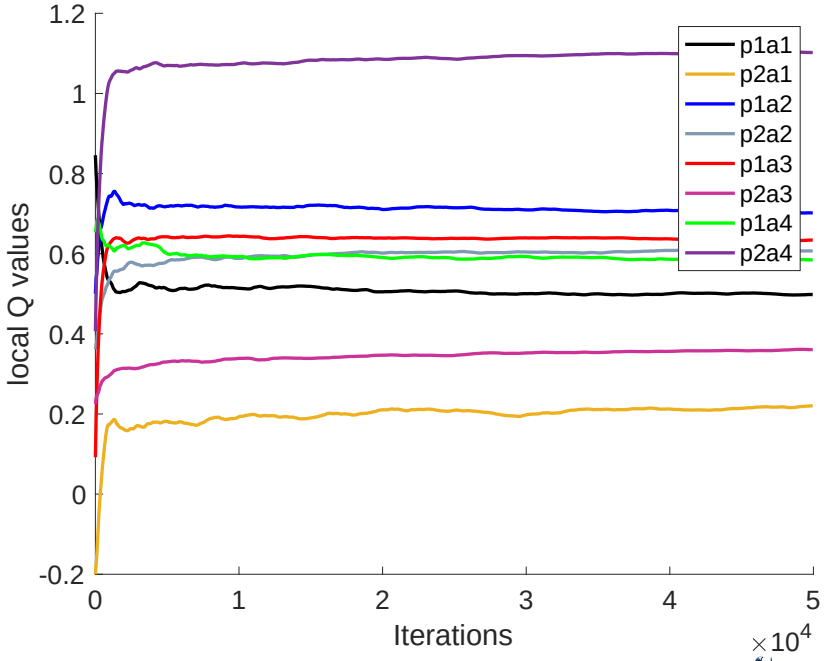


Two-player Single-State Potential Game with Four Actions

P2 / P1	a1	a2	a3	a4
a1	0.1,-0.4	0.3,-0.2	**0.4,0.3**	-0.1,-0.5
a2	0.5,0.2	**0.7, 0.4**	0.2,0.3	0.1,-0.1
a3	0.2,-0.2	0.3,-0.1	0.0,0.0	**0.4,0.1**
a4	**0.6, 0.9**	0.4,0.7	0.1,0.8	0.0, 0.4

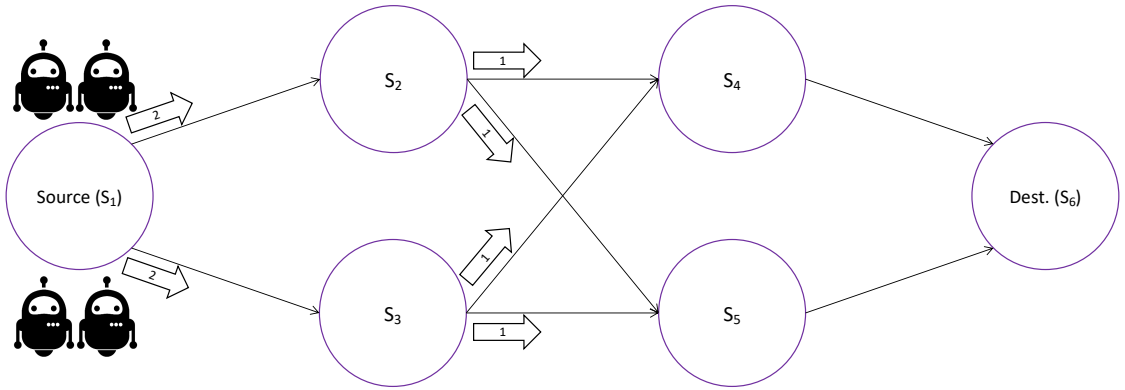
Φ

0.2	0.4	**0.9**	0.1
0.6	**0.8**	0.7	0.3
0.3	0.4	0.5	**0.6**
0.7	0.5	0.6	0.2

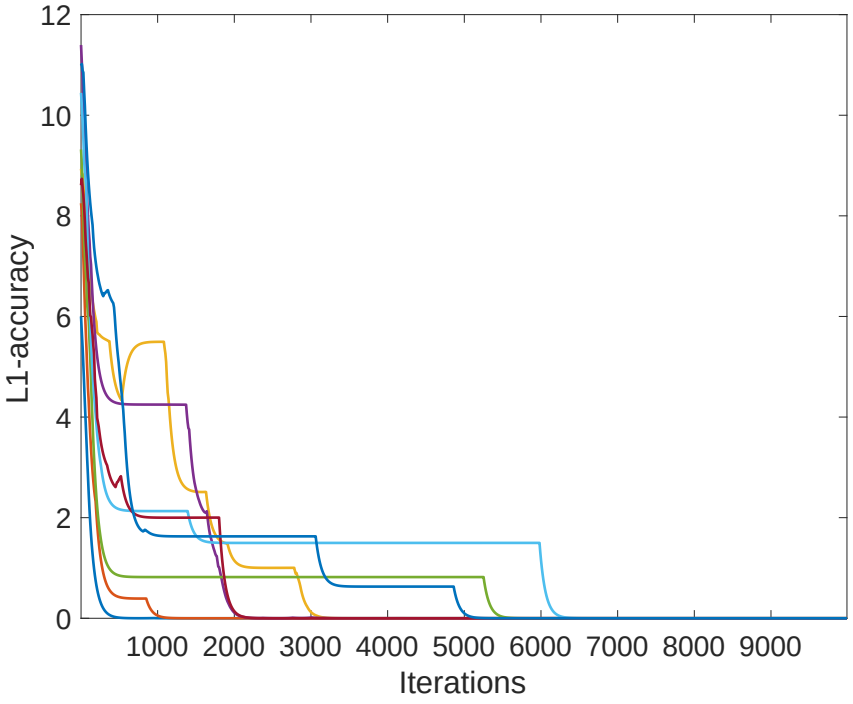


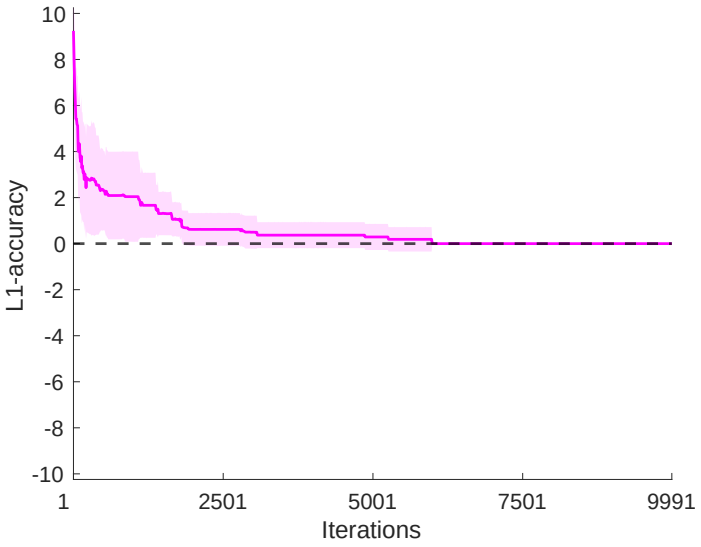
Optimal Strategy and Convergence Criterion

$$L1 - accuracy = \frac{1}{N}\sum_{i\in N}|BR^i - BR^i_{final}| = \frac{1}{N}\sum_{i\in N}\sum_{s\in\mathcal{S}}\sum_{a^i_j\in\mathcal{A}_i}|BR^i(s,a^i_j) - BR^i_{final}(s,a^i_j)|. \quad (26)$$

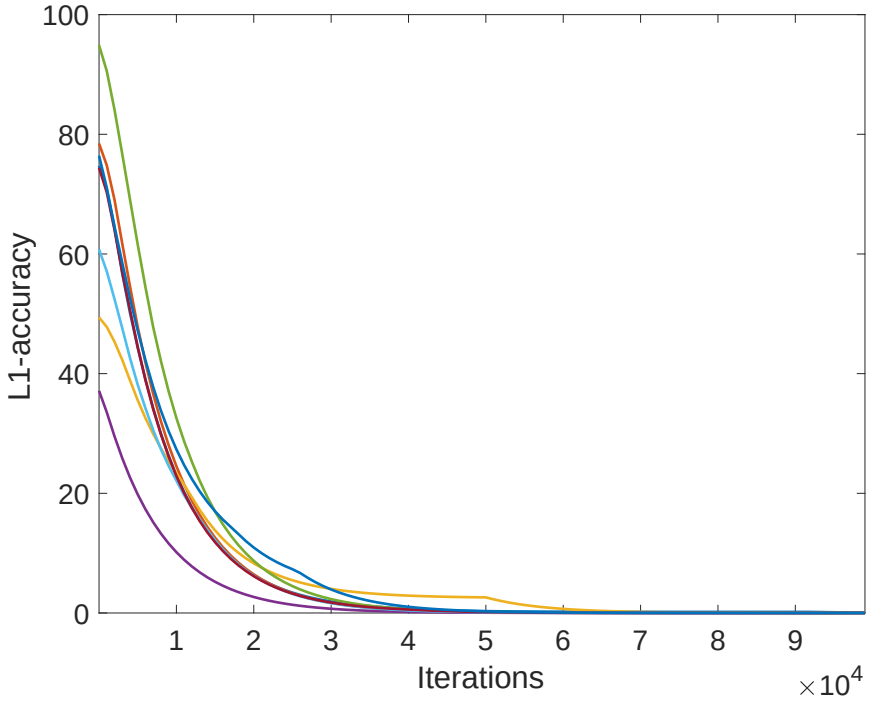


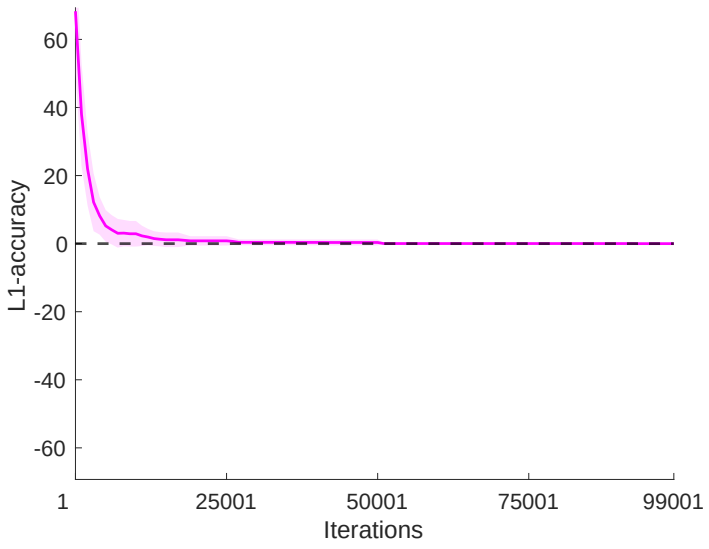
8 Independent Runs and Their Mean with Four Players





8 Independent Runs and Their Mean with Eight Players





Optimal Strategy

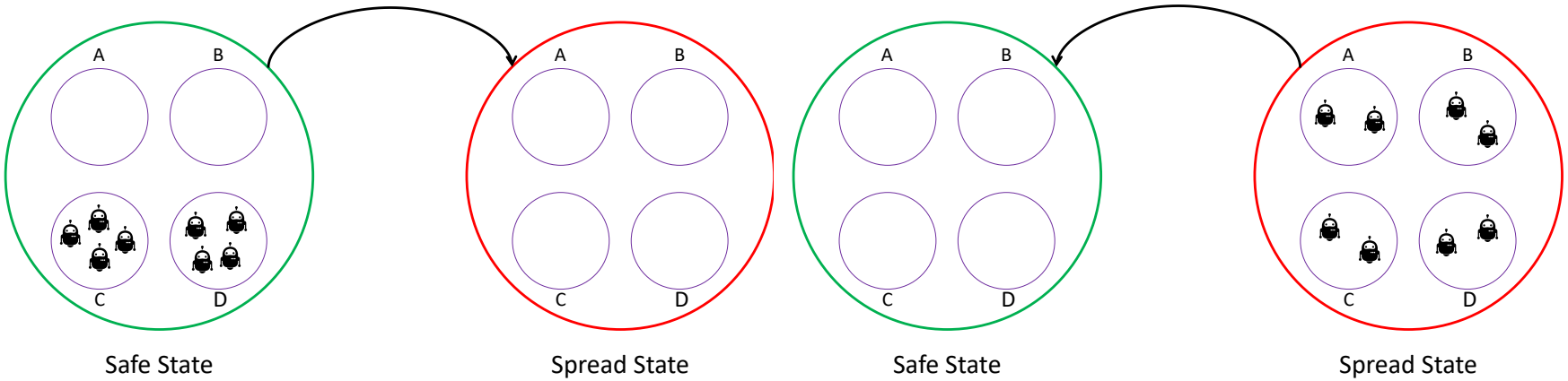
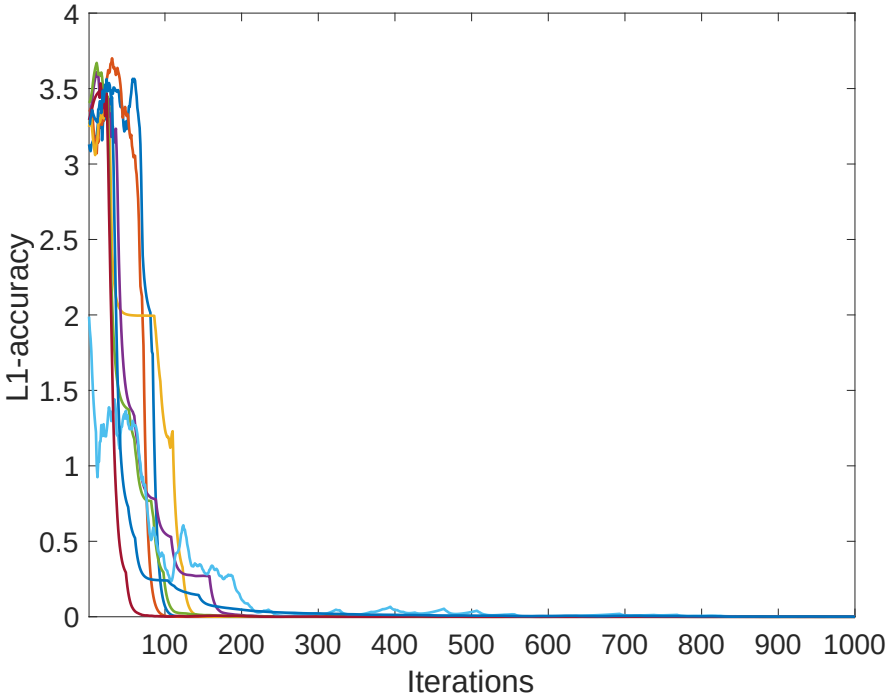


Figure: Optimum strategies in a distancing game

8 Independent Runs and Their Mean



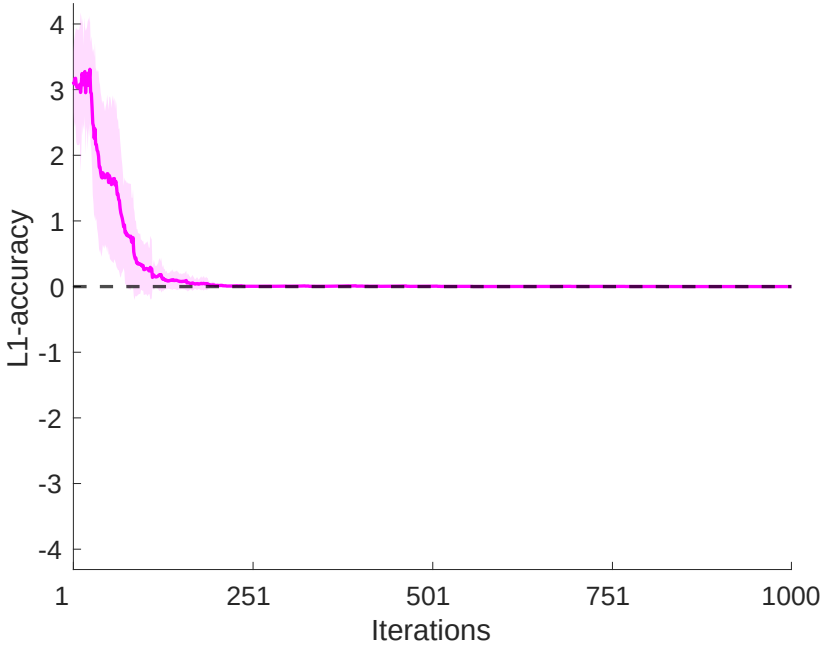


Table of Contents

1 Motivation
- Multi-Agent Systems and Stochastic Games
- Decentralized Multi-Agent Reinforcement Learning

2 Theoretical Background
- Game Theory
- Stochastic Games and MARL
- Stochastic Potential Game (SPG)

3 Learning Dynamics

4 Experimental Results
- Single-State Potential Games
- Network Routing Congestion Games
- Distancing Game

5 Convergence Analysis

6 Discussion, Conclusion, and Future Works

]

Proof Sketch

- We prove the convergence of the learning dynamics to a Nash equilibrium policy asymptotically, under an extra assumption for initializing local Q-functions. However, our conjecture and the focus of our future work is the prove of convergence under an arbitrary initialization.

Proof Sketch

- We prove the convergence of the learning dynamics to a Nash equilibrium policy asymptotically, under an extra assumption for initializing local Q-functions. However, our conjecture and the focus of our future work is the prove of convergence under an arbitrary initialization.
- In the first part of the analysis, similar to Sayin et al. [Sayin et al., 2021], the dynamics of a single state is decoupled from other states', but the formulation is given for a game with n players instead of two.

Proof Sketch

- We prove the convergence of the learning dynamics to a Nash equilibrium policy asymptotically, under an extra assumption for initializing local Q-functions. However, our conjecture and the focus of our future work is the prove of convergence under an arbitrary initialization.
- In the first part of the analysis, similar to Sayin et al. [Sayin et al., 2021], the dynamics of a single state is decoupled from other states', but the formulation is given for a game with n players instead of two.
- Afterwards, we associate this ODE with the one from stochastic fictitious play [Hofbauer and Sandholm, 2002] and using the initialization assumption related to belief-based Q-functions, we couple these two ODEs together to prove that they both follow the same trajectory and converge to the same solution that is a Nash equilibrium of the game in stochastic fictitious play.

Convergence Analysis

After decoupling the dynamics, the limiting ODE for a given state will be:

$$\frac{dq_s^i(t)}{dt} = Q_s^i Br^{-i}(q_s^{-i}(t), \tau(t)) - q^i(t) \quad \forall (i,s) \in N \times \mathcal{S}, \quad (27a)$$

$$\frac{dv_s^i(t)}{dt} = 0, \quad \forall (i,s) \in N \times \mathcal{S}, \quad (27b)$$

$$\frac{d\tau(t)}{dt} = 0, \quad \forall s \in \mathcal{S}. \quad (27c)$$

where

$$Q_s^i Br^{-i}(q_s^{-i}(t), \tau(t)) := Q_s^i \prod_{j \in N, j \neq i} Br^j(q_s^j(t), \tau(t)) \quad (28)$$

Convergence Analysis

Definition (Belief-based Q-function)

We call the local Q-function, q^i for player i, belief-based, if there exists opponent strategy profiles, where we could establish the following equality:

$$q^i(t) = Q^i \prod_{j \in N, j \neq i} \pi^j(t) \tag{29}$$

If we could establish that q^i are belief-based for every time step, then, the ODE would follow the same trajectory as the stochastic fictitious play [Hofbauer and Sandholm, 2002] with the following dynamics:

$$\frac{d\pi^i(t)}{dt} = Br^i(Q^i \pi^{-i}(t), \tau) - \pi^i(t), \tag{30}$$

Convergence Analysis

Without any extra assumptions, we can show that the original ODE is belief-based asymptotically by adding the following auxiliary ODEs:

$$\frac{d\pi_s^i(t)}{dt} = Br^i(q_s^i(t), \tau(t)) - \pi_s^i(t), \quad \forall i, s, \tag{31}$$

and using the following Lyapunov function:

$$H(t) = \sum_{i \in N} ||q^i - Q^i \pi^{-i}||^2 \tag{32}$$

for which:

$$\frac{dH(t)}{dt} = -nH(t) \tag{33}$$

Convergence Analysis

The convergence proof with an initialization assumption:
By establishing the belief-based property at the beginning of the trajectory using a communication-based or centralized initialization among the agents, based on Corollary 8.17 in Kelly and Peterson [Kelley and Peterson, 2010] and its trivial assumptions, the original system of ODE and the stochastic fictitious play ODE become coupled together by the initial condition and as a first-order vector differential equation, they will follow the same trajectory and have a unique solution that due to stochastic fictitious play is a Nash equilibrium [Hofbauer and Sandholm, 2002].

$$\frac{dq_s^i(t)}{dt} = Q_s^i Br^{-i}(q_s^{-i}(t), \tau(t)) - q^i(t) \quad \forall (i,s) \in N \times \mathcal{S}, \tag{34a}$$

$$\frac{d\pi_s^i(t)}{dt} = Br^i(Q_s^i \pi_s^{-i}(t), \tau) - \pi_s^i(t), \quad \forall (i,s) \in N \times \mathcal{S}. \tag{34b}$$

Table of Contents

1 Motivation
- Multi-Agent Systems and Stochastic Games
- Decentralized Multi-Agent Reinforcement Learning

2 Theoretical Background
- Game Theory
- Stochastic Games and MARL
- Stochastic Potential Game (SPG)

3 Learning Dynamics

4 Experimental Results
- Single-State Potential Games
- Network Routing Congestion Games
- Distancing Game

5 Convergence Analysis

6 Discussion, Conclusion, and Future Works

]

Game-Agnostic Convergence

- Game-agnostic convergence indicates that the learning dynamics converges regardless of the type of the game being played or the players' awareness of this type [Ding et al., 2022].

Game-Agnostic Convergence

- Game-agnostic convergence indicates that the learning dynamics converges regardless of the type of the game being played or the players' awareness of this type [Ding et al., 2022].
- Not yet explored in stochastic games and MARL. Has been only established in non-stochastic matrix games for zero-sum and potential games when players adopt fictitious play [Hofbauer and Sandholm, 2002] or Q-learning [Leonardos and Piliouras, 2022].

Game-Agnostic Convergence

- Game-agnostic convergence indicates that the learning dynamics converges regardless of the type of the game being played or the players' awareness of this type [Ding et al., 2022].
- Not yet explored in stochastic games and MARL. Has been only established in non-stochastic matrix games for zero-sum and potential games when players adopt fictitious play [Hofbauer and Sandholm, 2002] or Q-learning [Leonardos and Piliouras, 2022].
- Ding et al. [Ding et al., 2022], for the first time, show that a gradient-based algorithm converges to Nash equilibrium in both two-player stochastic zero-sum and cooperative games.

Game-Agnostic Convergence

- Game-agnostic convergence indicates that the learning dynamics converges regardless of the type of the game being played or the players' awareness of this type [Ding et al., 2022].
- Not yet explored in stochastic games and MARL. Has been only established in non-stochastic matrix games for zero-sum and potential games when players adopt fictitious play [Hofbauer and Sandholm, 2002] or Q-learning [Leonardos and Piliouras, 2022].
- Ding et al. [Ding et al., 2022], for the first time, show that a gradient-based algorithm converges to Nash equilibrium in both two-player stochastic zero-sum and cooperative games.
- Therefore, our results indicate that decentralized Q-learning is the first game-agnostic value-based MARL algorithm.

Conclusion

Conclusion

In this thesis, for the first time, we applied a value-based MARL approach, decentralized Q-learning, to stochastic potential games, a class of stochastic games that is capable of modeling many real-world scenarios. Following the previous established convergence in zero-sum games, our results show that decentralized Q-learning is a game-agnostic value-based MARL algorithm.

Future Works

- Analyzing the finite-time convergence properties of decentralized Q-learning.
- Extending the algorithm to continuous state spaces using function approximation.
- Delving deeper into studying game-agnostic convergence of the algorithm by considering other families of games.

Thank you!

hafez.ghaemi@studenti.polito.it

References I

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. (2021).
On the theory of policy gradient methods: Optimality, approximation, and distribution shift.
J. Mach. Learn. Res., 22(98):1–76.

Ding, D., Wei, C.-Y., Zhang, K., and Jovanovic, M. (2022).
Independent policy gradient for large-scale markov potential games: Sharper rates, function approximation, and game-agnostic convergence.
In International Conference on Machine Learning, pages 5166–5220. PMLR.

References II

Fox, R., Mcaleer, S. M., Overman, W., and Panageas, I. (2022).
Independent natural policy gradient always converges in markov potential games.
In International Conference on Artificial Intelligence and Statistics, pages 4414–4425. PMLR.

Hofbauer, J. and Sandholm, W. H. (2002).
On the global convergence of stochastic fictitious play.
Econometrica, 70(6):2265–2294.

Kelley, W. G. and Peterson, A. C. (2010).
The theory of differential equations: classical and qualitative.
Springer Science & Business Media.

References III

Leonardos, S., Overman, W., Panageas, I., and Piliouras, G. (2021).
Global convergence of multi-agent policy gradient in markov potential games.
arXiv preprint arXiv:2106.01969.

Leonardos, S. and Piliouras, G. (2022).
Exploration-exploitation in multi-agent learning: Catastrophe theory meets game theory.
Artificial Intelligence, 304:103653.

References IV

Mguni, D. H., Wu, Y., Du, Y., Yang, Y., Wang, Z., Li, M., Wen, Y., Jennings, J., and Wang, J. (2021).
Learning in nonzero-sum stochastic games with potentials.
In International Conference on Machine Learning, pages 7688–7699. PMLR.

Sayin, M., Zhang, K., Leslie, D., Basar, T., and Ozdaglar, A. (2021).
Decentralized q-learning in zero-sum markov games.
Advances in Neural Information Processing Systems, 34.

References V

Zhang, R., Ren, Z., and Li, N. (2021).
Gradient play in multi-agent markov stochastic games: Stationary points and convergence.
arXiv preprint arXiv:2106.00198.