Theory and Methods for Reinforcement Learning

Prof. Volkan Cevher
volkan.cevher@epfl.ch

Lecture 10: Solving Markov Games

Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)

EE-618 (Spring 2022)

License Information for Theory and Methods for Reinforcement Learning (EE-618)

Markov games

- What is Markov game?
- Value functions and Nash equilibrium
- Algorithms for Markov games
  - Nonlinear programming
  - Fictitious play
  - Policy gradient
  - Nash Q-learning

Markov games

- A **Markov game** (MG) can be viewed as a MDP involving multiple **agents** with their own rewards
- Introduced by L.S.Shapley [5] as stochastic games, referred to with a tuple $(\mathcal{S}, \boldsymbol{\mathcal{A}}, P, \boldsymbol{r}, \gamma)$
- A Markov game is an extension of normal form game with multiple stages and a **shared state** $s \in \mathcal{S}$
- **Joint action**: $\boldsymbol{a} = (a_i)_i$, where $a_i \in \mathcal{A}_i$ is the action of agent $i \in \mathcal{I}$
- **Transition function**: $\mathrm{P}\left(s' \mid s, \boldsymbol{a}\right)$ is the likelihood of transitioning from a state s to s' under an action $\boldsymbol{a}$
- **Reward function**: $r_i(s, \boldsymbol{a})$ is the reward received by agent i at state s with a joint action $\boldsymbol{a}$
- **Discount factor**: γ
- **Stationary policy**: $\pi_i(a_i \mid s)$ is the probability that agent i selects action a_i at state s

An example

◦ Consider the interaction between drivers in the traffic as a markov game.

© eyetronic, Adobe Stock

- agents: commuters/drivers in the traffic
- states: locations of all cars
- action: which road to drive for each car
- reward: negative of time spent on the road

Normal form games and Markov games

	action	state	transition	reward	policy	multi-stage
Normal form game	$a_i \in \mathcal{A}_i$	no	no	$r_i(\boldsymbol{a})$	$\pi_i(a)$	no
Markov game	$a_i \in \mathcal{A}_i$	$s \in \mathcal{S}$	$\mathsf{P}\left(s' \mid s, \boldsymbol{a}\right)$	$r_i(s, \boldsymbol{a})$	$\pi_i(a_i \mid s)$	yes

- We focus on infinite horizon Markov games
- Compared to a normal form game, agents in MG consider not only the current reward of the action...
  ...but also its effect in the long run!
- Compared to an MDP, MG has multiple agents and the reward also depends on other agents' action.

Markov games

- What is Markov game?
- Value functions and Nash equilibrium
- Algorithms for Markov games
  - Nonlinear programming
  - Fictitious play
  - Policy gradient
  - Nash Q-learning

Value function

- **Value function**: the expected γ discounted sum of rewards for a player i starting from state s, when all players play their part of the joint policy $(\pi_i)_{i\in\mathcal{I}}$:

$$V_i^{\boldsymbol{\pi}}(s) = \mathbb{E}\left[\sum_{t=0}^{+\infty} \gamma^t r_i\left(s^t, \boldsymbol{a}^t\right) \mid s^0 = s, \boldsymbol{a}^t \sim \boldsymbol{\pi}\left(\cdot \mid s^t\right), s^{t+1} \sim \mathsf{P}\left(\cdot \mid s^t, \boldsymbol{a}^t\right)\right].$$

- **Action-value function**:

$$Q_i^{\boldsymbol{\pi}}(s, \boldsymbol{a}) = \mathbb{E}\left[\sum_{t=0}^{+\infty} \gamma^t r_i\left(s^t, \boldsymbol{a}^t\right) \mid s^0 = s, \boldsymbol{a}^0 = \boldsymbol{a}, \boldsymbol{a}^t \sim \boldsymbol{\pi}\left(\cdot \mid s^t\right), s^{t+1} \sim \mathsf{P}\left(\cdot \mid s^t, \boldsymbol{a}^t\right)\right].$$

Remarks:

- Relation between $Q_i^{\boldsymbol{\pi}}(s, \boldsymbol{a})$ and $V_i^{\boldsymbol{\pi}}(s)$

$$Q_i^{\boldsymbol{\pi}}(s, \boldsymbol{a}) = r_i(s, \boldsymbol{a}) + \gamma \sum_{s' \in S} \mathsf{P}\left(s' \mid s, \boldsymbol{a}\right) V_i^{\boldsymbol{\pi}}\left(s'\right).$$

- Each agent wants to maximize its value.

Response model – best response

- The expected reward to agent i from state s when following joint policy $\boldsymbol{\pi}$ is

$$r_i(s, \boldsymbol{\pi}(\cdot|s)) = \sum_{\boldsymbol{a}} r_i(s, \boldsymbol{a}) \prod_{j \in \mathcal{I}} \pi_j \left(a_j \mid s\right).$$

- The probability of transitioning from state s to s' when following $\boldsymbol{\pi}$ is

$$\mathsf{P}\left(s' \mid s, \boldsymbol{\pi}(\cdot|s)\right) = \sum_{\boldsymbol{a}} \mathsf{P}\left(s' \mid s, \boldsymbol{a}\right) \prod_{j \in \mathcal{I}} \pi_j \left(a_j \mid s\right).$$

- **Best response policy** for agent i is a policy π_i that maximizes expected utility given the fixed policies of other agents $\boldsymbol{\pi}_{-i}$. This best response can be computed by solving the MDP with

$$\mathsf{P}'\left(s' \mid s, a_i\right) = \mathsf{P}\left(s' \mid s, a_i, \boldsymbol{\pi}_{-i}(s)\right)$$
$$r'\left(s, a_i\right) = r_i\left(s, a_i, \boldsymbol{\pi}_{-i}(s)\right).$$

Nash equilibrium

- In a **Nash equilibrium** (NE) $\boldsymbol{\pi}^\star$, no player can improve its value by changing its policy if the other players stick to their policy.

- Or we can say, $\pi_i^\star$ is the best policy for agent i if other agents stick to $\boldsymbol{\pi}^\star_{-i}$.

- In NE, we can write for each agent i

$$V_i^{\boldsymbol{\pi}^\star}(s) \geq V_i^{\pi_i, \boldsymbol{\pi}^\star_{-i}}(s), \quad \forall \pi_i, \forall s \in \mathcal{S}.$$

- ϵ-Nash equilibrium:

$$V_i^{\boldsymbol{\pi}}(s) + \epsilon \geq \max_{\pi_i} V_i^{\boldsymbol{\pi}}(s), \quad \forall i, \forall s \in \mathcal{S}.$$

Existence of Nash equilibrium

Theorem (Existence of Nash equilibrium [3])

All finite Markov games with a discounted infinite horizon have a Nash equilibrium.

Exercise: ◦ Show this with the theorem of the existence of Nash equilibrium in the normal form games.

Hint: ◦ Construct a new normal form game with each player and state pair in the original Markov game, i.e. (i, s), as an agent in the new game.

Markov games

- What is Markov game?
- Value functions and Nash equilibrium
- Algorithms for Markov games
  - Nonlinear programming
  - Fictitious play
  - Policy gradient
  - Nash Q-learning

Nonlinear optimization to find NE [2]

- Minimizes the sum of the lookahead utility deviations
- Constrains the policies to be valid distributions
- Assume we know reward and transition functions

$$\begin{aligned} \underset{\boldsymbol{\pi}, V}{\operatorname{minimize}} \quad & \sum_{i \in \mathcal{I}} \sum_{s} \left(V_i(s) - Q_i(s, \boldsymbol{\pi}(\cdot|s))\right) \\ \text{subject to} \quad & V_i(s) \geq Q_i\left(s, a_i, \boldsymbol{\pi}_{-i}(\cdot|s)\right) \text{ for all } i, s, a_i \\ & \sum_{a_i} \pi_i\left(a_i \mid s\right) = 1 \text{ for all } i, s \\ & \pi_i\left(a_i \mid s\right) \geq 0 \text{ for all } i, s, a_i, \end{aligned}$$

where $Q_i(s, \boldsymbol{\pi}(\cdot|s)) = r_i(s, \boldsymbol{\pi}(\cdot|s)) + \gamma \sum_{s'} \mathsf{P}\left(s' \mid s, \boldsymbol{\pi}(\cdot|s)\right) V_i\left(s'\right)$.

Nonlinear optimization: Equivalence between the optimal solution and NE

Theorem (Equivalence between optimal solution and NE[2])

A joint policy $\boldsymbol{\pi}^\star$ is a NE with value $V^\star$ if and only if $(\boldsymbol{\pi}^\star, V^\star)$ is a global minimum to this nonlinear programming.

Remarks:

- The nonlinearity arises in $r_i(s, \boldsymbol{\pi}(\cdot|s))$ and $\mathsf{P}\left(s' \mid s, \boldsymbol{\pi}(\cdot|s)\right)$.
- The proof of the theorem uses the following lemma.

Lemma

In an MDP, $V^\star$ is the optimal value with the optimal policy $\pi^\star$ if and only if

$$V^\star(s) = r(s, \pi^\star(\cdot|s)) + \sum_{s' \in \mathcal{S}} \mathsf{P}\left(s' \mid s, \pi^\star(\cdot|s)\right) V^\star(s'), \quad \forall s \in \mathcal{S}$$

$$V^\star(s) \geq r(s, a) + \sum_{s' \in \mathcal{S}} \mathsf{P}\left(s' \mid s, a\right) V^\star(s'), \quad \forall s \in \mathcal{S}, a \in \mathcal{A}.$$

Nonlinear optimization: Equivalence between the optimal solution and NE

◦ We are ready to prove the theorem.

Proof.

◦ ($\Longrightarrow$) Assume $\pi^\star$ is a NE with value $V^\star$

1. The second and third constraints hold trivially.
2. The first constraint makes the optimum at least 0.
3. The lemma implies the first constraint is feasible and the objective value at $(\pi^\star, V^\star)$ is 0.

◦ ($\Longleftarrow$) Assume $(\pi^\star, V^\star)$ is a global minimum to the nonlinear programming

1. The optimum is 0 and is achievable by the reasoning above.
2. By the lemma, three constraints and the objective at $(\pi^\star, V^\star)$ being 0 implies that $\pi^\star$ is a NE with value $V^\star$.

□

Fictitious play in Markov games

- **Required feedback** Each agent i counts opponent's actions at state s: $N_t(j, a_j, s)$ for $j \neq i, s \in \mathcal{S}$.

- **Behavioural assumption** Each agent i assumes its opponents use the empirical distribution as the same stationary mixed strategy

$$\widetilde{\pi}_j^t(a_j \mid s) = \frac{N_t(j, a_j, s)}{\sum_{\bar{a}_j \in \mathcal{A}_j} N_t(j, \bar{a}_j, s)}.$$

- Each agent i considers the following MDP,

$$\mathsf{P}^t\left(s' \mid s, a_i\right) = \mathsf{P}\left(s' \mid s, a_i, \widetilde{\pi}_{-i}^t(s)\right)$$
$$r^t\left(s, a_i\right) = r_i\left(s, a_i, \widetilde{\pi}_{-i}^t(s)\right),$$

  and computes

$$Q_i^t(s, a_i, \widetilde{\boldsymbol{\pi}}_{-i}^t(\cdot|s)).$$

- Each agent i updates their policy as follows

$$\pi_i^{t+1}(s) = \arg\max_{a_i} Q_i^t(s, a_i, \widetilde{\boldsymbol{\pi}}_{-i}^t(\cdot|s)) \quad \forall s \in \mathcal{S}.$$

Policy gradient methods

- Also referred to as gradient ascent.
- Take the gradient of value function at $\boldsymbol{\pi}^t$: $\left.\frac{\partial V_i^{\pi}(s)}{\partial \pi_i(a_i|s)}\right|_{\boldsymbol{\pi}=\boldsymbol{\pi}^t}$.
- Apply gradient ascent to each agent

$$\pi_i^{t+1}(a_i \mid s) = \pi_i^t(a_i \mid s) + \alpha_i^t \left.\frac{\partial V_i^{\pi}(s)}{\partial \pi_i(a_i \mid s)}\right|_{\boldsymbol{\pi}=\boldsymbol{\pi}^t}.$$

- Project π_i^{t+1} to a valid probability distribution.

Policy gradient algorithms in linear quadratic (LQ) games

- Generalization of LQR to multiple agents setting
- Continuous, vector valued state $s \in \mathbb{R}^m$ and action space $a_i \in \mathbb{R}^{d_i}$ for agent i.
- Linear dynamics for state transition: with matrices $A \in \mathbb{R}^{m \times m}$ and $B_i \in \mathbb{R}^{d_i \times m}$

$$s^{t+1} = As^t + \sum_{i=1}^{n} B_i a_i^t.$$

- Consider the linear feedback policy $a_i = \pi_i(s) = -K_i s$ with $K_i \in \mathbb{R}^{m \times d_i}$.
- Player i's loss function is quadratic function: with $Q_i \in \mathbb{R}^{m \times m}$, $R_i \in \mathbb{R}^{d_i \times d_i}$ and initial state distribution $\mathcal{D}_0$

$$f_i(K_1, ..., K_n) = \mathbb{E}_{s^0 \sim \mathcal{D}_0} \left[\sum_{t=0}^{\infty} (s^t)^T Q_i s^t + (a_i^t)^T R_i a_i^t \right].$$

Non-convergence of policy gradient algorithms in linear quadratic games

- Each player wants to minimize its loss $f_i(K_1, \ldots, K_i, ..., K_n)$
- $(K_1^\star, ..., K_n^\star)$ is a Nash equilibrium if for each agent i

$$f_i\left(K_1^\star, \ldots, K_i^\star, \ldots, K_N^\star\right) \le f_i\left(K_1^\star, \ldots, K_i, \ldots, K_N^\star\right), \forall K_i \in \mathbb{R}^{d_i \times m}.$$

- Policy gradient algorithms

$$K_i^{t+1} = K_i^t - \alpha_i \frac{\partial f}{\partial K_i}(K_1^t, ..., K_n^t).$$

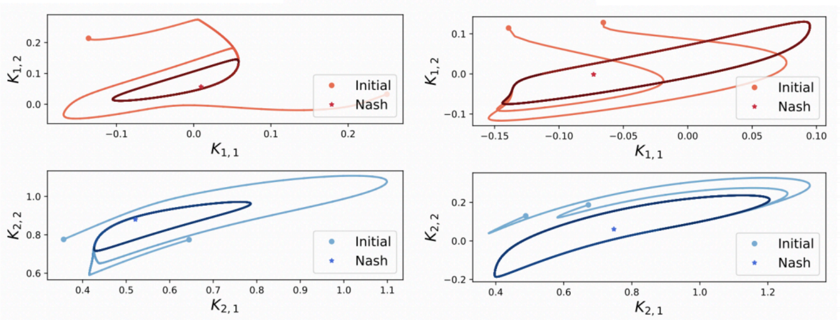
Theorem (Non-convergence of policy gradient in LQ games [4])

There is a LQ game that the set of initial conditions in a neighborhood of the Nash equilibrium from which gradient converges to the Nash equilibrium is of measure zero.

- **Remark**: When the initial policy is close enough to NE and stepsize is small enough, it still may not converge.

Non-convergence of policy gradient algorithms in linear quadratic games

- Implement policy gradient on two LQ games with two players with dimension $d_1 = d_2 = 1$ and $m = 2$.
- Nash equilibrium is avoided by the gradient dynamics.
- Players converge to the same cycle from different initializations.



Two-player zero-sum Markov games

- What is two-player zero-sum Markov games?
- Bellman operators in two-player zero-sum Markov games
- Algorithms for two-player zero-sum games
  - Value iteration
  - Policy iteration and its variants

Two-player zero-sum Markov games

◦ Markov games with two agents

◦ Sum of two agents' rewards is 0, i.e. $r_1(s, a_1, a_2) = -r_2(s, a_1, a_2) = r(s, a_1, a_2)$ for any $s \in \mathcal{S}$.

◦ Value function:

$$V^{\pi_1, \pi_2}(s) = E\left[\sum_{t=0}^{+\infty} \gamma^t r\left(s_t, a_1^t, a_2^t\right) \mid s_0 = s, a_1^t \sim \pi_1\left(\cdot \mid s_t\right), a_2^t \sim \pi_2\left(\cdot \mid s_t\right), s_{t+1} \sim \mathsf{P}\left(\cdot \mid s_t, a_1^t, a_2^t\right)\right].$$

◦ Agent 1 wants to maximize the value function and agent 2 wants to minimize it.

◦ There exists a unique value for all Nash equilibrium

$$V^\star(s) = \min_{\pi_1} \max_{\pi_2} V^{\pi_1, \pi_2}(s) = \max_{\pi_2} \min_{\pi_1} V^{\pi_1, \pi_2}(s).$$

Applications of two-player zero-sum Markov games

- Includes many sequential games. When one wins, the other loses.
- Poker.
- Tennis.
- Go
  - agents: players
  - states: the states of the board
  - action: move in each turn
  - reward: zero for all non-terminal steps; the terminal reward at the end of the game: +1 for winning and -1 for losing.

Two-player zero-sum Markov games

- What is two-player zero-sum Markov games?
- Bellman operators in two-player zero-sum Markov games
- Algorithms for two-player zero-sum games

Bellman operators in two-player zero-sum Markov games

- Let $r(s,\pi_1(s),\pi_2(s))$ the expected immediate reward/cost (player 1/player 2) at state s under policies π_1,π_2.
- Define the operator $\mathcal{T}_{\pi_1}$ as follows,

$$[\mathcal{T}_{\pi_1}\boldsymbol{V}](s) = \max_{\pi_1}\min_{\pi_2}\left[r(s,\pi_1(s),\pi_2(s)) + \gamma\sum_{s'}\mathsf{P}(s' \mid s,\pi_1(s),\pi_2(s))\cdot V(s')\right]$$

- Define the operator $\mathcal{T}_{\pi_2}$ as follows,

$$[\mathcal{T}_{\pi_2}\boldsymbol{V}](s) = \min_{\pi_2}\max_{\pi_1}\left[r(s,\pi_1(s),\pi_2(s)) + \gamma\sum_{s'}\mathsf{P}(s' \mid s,\pi_1(s),\pi_2(s))\cdot V(s')\right]$$

- $\mathcal{T}_{\pi_1}$ and $\mathcal{T}_{\pi_2}$ are equivalent. Let $\mathcal{T}\equiv\mathcal{T}_{\pi_1}\equiv\mathcal{T}_{\pi_2}$
- The fixed point of $\mathcal{T}$ is $\boldsymbol{V}^\star$.

Two-player zero-sum Markov games

- What is two-player zero-sum Markov games?
- Bellman operators in two-player zero-sum Markov games
- Algorithms for two-player zero-sum games

Value iteration for two-player zero-sum Markov games

Value iteration for two-player zero-sum Markov games [5]

for each stage t **do**

Apply the Bellman operator $\mathcal{T}$ at each iteration

$$V^{t+1} = \mathcal{T}V^t.$$

end for

Theorem (Convergence of value iteration)

$$\left\|\mathbf{V}^t - \mathbf{V}^\star\right\|_\infty \leq \gamma^t \left\|\mathbf{V}^0 - \mathbf{V}^\star\right\|_\infty.$$

Policy iteration for two-player zero-sum Markov games

◦ π_1 is said to be **greedy**, denoted as $\pi_1 \in \mathcal{G}(V)$ if and only if **for each state** $s \in S$,

$$\pi_1(\cdot|s) := \underset{\pi_1(\cdot|s)}{\arg\max} \; \min_{\pi_2(\cdot|s)} \left[r(s, \pi_1(s), \pi_2(s)) + \gamma \sum_{s'} \mathsf{P}(s' \mid s, \pi_1(s), \pi_2(s)) \cdot V(s') \right]$$

Policy iteration for two-player zero-sum Markov games

for each stage t **do**
  find $\pi_1^t \in \mathcal{G}(V^{t-1})$
  compute $V^t = \min_{\pi_2} V^{\pi_1^t, \pi_2}$
end for

Remarks:

- The first step requires the solution of $|\mathcal{S}|$ **linear programs**.
- The second step to compute $V^t = \min_{\pi_2} V^{\pi_1, \pi_2}$ requires solving the MDP with transition $\mathbb{E}_{a_1 \sim \pi_1^t(\cdot|s)}[P(\cdot \mid s, a_1, a_2)]$ and reward $-\mathbb{E}_{a_1 \sim \pi_1^t(\cdot|s)}[r(s, a_1, a_2)]$.

Value and Policy Iteration in zero-sum Markov games

Pros

- Compute Nash Equilibrium.
- Simple to implement.

Cons

- Computationally expensive.
- Model-based (they need the exact description of the Markov game).

Model-free methods for NE

- Policy gradient [1]
- Optimistic mirror decent + actor-critic [6]
- Natural policy gradient + actor-critic [Alacaoglu et al.]

Policy gradient in two-player zero-sum Markov games

Policy gradient in two-player zero-sum Markov games [1]

for each stage $i = 1$ to ... **do**

A trajectory $\{(s^t, \alpha_1^t, \alpha_2^t)\}_{t=0}^{H-1}$ is sampled according to policies π_1^i, π_2^i.

- Player 1 updates π_1^{i+1} as follows,

$$\pi_1^{i+1} \leftarrow \Pi_{\text{eucl}}\left[\pi_1^i + \left(\sum_{t=0}^{H-1} r(s^t, \alpha_1^t, \alpha_2^t)\right) \cdot \sum_{t=0}^{H-1} \nabla \log(\pi_1^i(a_1^t|s^t)\right]$$

- Player 2 updates π_2^{i+1} as follows,

$$\pi_2^{i+1} \leftarrow \Pi_{\text{eucl}}\left[\pi_2^i - \left(\sum_{t=0}^{H-1} r(s^t, \alpha_1^t, \alpha_2^t)\right) \cdot \sum_{t=0}^{H-1} \nabla \log(\pi_2^i(a_2^t|s^t)\right]$$

where $\Pi_{\text{eucl}}[\cdot]$ is the euclidean projection to the set of policies.

end for

Policy gradient in two-player zero-sum Markov games

Theorem (Informal, [1])

Policy-gradient in two-player zero-sum games requires $O(1/\epsilon^{12.5})$ *stages to converge to an* ϵ*-Nash Equilibrium.*

Policy gradient in two-player zero-sum Markov games

- Model-free
- Each player needs to learn only her individual experienced payoffs.
- Efficient and simple to implement.

Cons

Huge sample-complexity, PL needs to sample $O(1/\epsilon^{12.5})$ trajectories to find an ϵ-NE.

Other model-free methods for two-player zero-sum Markov games

○ Recent methods model-free drastically improve on the sample complexity.

Optimistic gradient decent/ascent with actor-critic [6]

- At each stage i a trajectory $\{(s^t, \alpha_1^t, \alpha_2^t)\}_{t=0}^{H-1}$ is sampled according to π_1^i, π_2^i.
- Agent 1 (resp. agent 2) estimates the $\hat{Q}^i(s, a_1)$ as follows,

$$\hat{Q}^i(s, a_1) \leftarrow \frac{\sum_{t=0}^{H-1} \mathbf{1}[s^t = s, a_1^t = a_1] \cdot \left(r(a_1^t, a_2^t, s^t) + \gamma V^{i-1}(s^{t+1})\right)}{\sum_{t=0}^{H-1} \mathbf{1}[s^t = s, a_1^t = a_1]} \quad \leftarrow \text{Critic}$$

- At each state s, optimistic gradient ascent (descent for player 2) uses $\hat{Q}^i(s, a)$ to update $\pi^i(\cdot|s)$.

Convergence [6]

Optimistic gradient decent/ascent with actor-critic in two-player zero-sum games requires $O(1/\epsilon^4)$ stages to converge to an ϵ-Nash Equilibrium.

State of the art [Alacaoglu et. al.]

Natural policy gradient with actor-critic in two-player zero-sum games requires $O(1/\epsilon^2)$ stages to converge to an ϵ-Nash Equilibrium.

Summary

- Markov games
  - What is Markov game?
  - Value functions and Nash equilibria
  - Algorithms for Markov games

- Two-player zero-sum Markov games
  - What is two-player zero-sum Markov games?
  - Bellman operators in two-player zero-sum Markov games
  - Algorithms for two-player zero-sum games

References

[1] Constantinos Daskalakis, Dylan J. Foster, and Noah Golowich.
Independent policy gradient methods for competitive reinforcement learning, 2021.

[2] Jerzy A Filar, Todd A Schultz, Frank Thuijsman, and OJ Vrieze.
Nonlinear programming and stationary equilibria in stochastic games.
Mathematical Programming, 50(1):227–237, 1991.

[3] A. M. Fink.
Equilibrium in a stochastic n-person game.
Journal of Science of the Hiroshima University, Series A-I (Mathematics), 28(1):89 – 93, 1964.

[4] Eric Mazumdar, Lillian J Ratliff, Michael I Jordan, and S Shankar Sastry.
Policy-gradient algorithms have no guarantees of convergence in linear quadratic games.
In *AAMAS Conference proceedings*, 2020.

[5] Lloyd S Shapley.
Stochastic games.
Proceedings of the national academy of sciences, 39(10):1095–1100, 1953.

[6] Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo.
Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive markov games, 2021.