Theory and Methods for Reinforcement Learning

Prof. Volkan Cevher
volkan.cevher@epfl.ch

Lecture 7: Markov Games

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

EE-618 (Spring 2023)

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

Games

- The mathematical discussion of games can be traced back to 16th century by Gerolamo Cardano.
- From 17th-19th century, many different games are analyzed, such as the card game le Her and chess game.
- John von Neumann published the paper *On the Theory of Games of Strategy* in 1928.
- John Nash formalized Nash equilibrium in broad classes of games.

Figure: John von Neumann

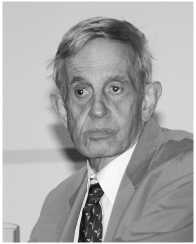


Figure: John Nash

Normal form games

- What is normal form game?
- Equilibria
- Dynamics for games
  - Iterated best response
  - Fictitious play
  - Gradient ascent

Normal form games

- What is normal form game?
- Equilibria
- Dynamics for games
  - Iterated best response
  - Fictitious play
  - Gradient ascent

Normal form games

- There is a set of **players/agents**: $\mathcal{I}$
- **Joint action**: $\boldsymbol{a} = (a_i)_i$, where $a_i \in \mathcal{A}_i$ is the action of agent $i \in \mathcal{I}$
- **Reward/Payoff**: $r_i(\boldsymbol{a})$ is the reward received by agent i with a joint action $\boldsymbol{a}$
- The game can be represented as above is called **normal form game**
- Other types of games:
  - Extensive form games
  - Markov games
  - Continuous action games
  - Cournot oligopolies

Strategies

- **Strategy/Policy**: $\pi_i \in \Delta(\mathcal{A}_i)$: $\pi_i(a_i)$ is the probability that agent i selects action a_i
  - pure strategy (deterministic policy): only play one action
  - mixed strategy (stochastic policy): a distribution over the set of actions
- **Strategy profile**: one strategy of each player $\boldsymbol{\pi} = (\pi_i)_i$
- Each player wants to maximize its payoff
- The expected payoff of player i when a strategy profile $\boldsymbol{\pi}$ is used

$$\underbrace{r_i(\boldsymbol{\pi}) = \sum_{\boldsymbol{a}} r_i(\boldsymbol{a}) \prod_{j \in \mathcal{I}} \pi_j(a_j)\,.}_{\text{expected payoff}}$$

Remark: We will see why mixed strategies can be necessary to consider.

A special case: Two-player games

- The game with two players
- The payoffs of two player normal form games can be represent with matrix forms
- Prisoners dilemma [14]: each agent can choose to cooperate or defect

		Bob: cooperate	Bob: defect
Alex	cooperate	$1/1$	$-1/2$
Alex	defect	$2/-1$	$0/0$

- Example: if Alex plays defect and Bob plays cooperate they receive 2 and -1 respectively.

A special case: Two-player zero-sum games

- The sum of two players' payoffs are zero, i.e., $r_1(a_1, a_2) = -r_2(a_1, a_2)$
- The payoff of a two-player zero-sum normal form game can be represented with a matrix A
- $A(i, j)$ is the payoff of player 1 (loss of player 2) when choosing i-th action and player 2 chooses its j-th action
- The expected payoff of player 1 / loss of player 2:

$$r_1(\pi_1, \pi_2) = (\pi_1)^\top A\pi_2$$

- Player 1 wants to maximize $(\pi_1)^\top A\pi_2$ and player 2 wants to minimize it

Response models

∘ What will a player do if other players' strategies are fixed at $\boldsymbol{\pi}_{-i} \triangleq (\pi_1, \ldots, \pi_{i-1}, \pi_{i+1}, \ldots, \pi_n)$?

∘ A **best response** of agent i to the policies of the other agents $\boldsymbol{\pi}_{-i}$ is a policy π_i such that

$$r_i\left(\pi_i, \boldsymbol{\pi}_{-i}\right) \geq r_i\left(\widetilde{\pi}_i, \boldsymbol{\pi}_{-i}\right), \quad \forall \widetilde{\pi}_i$$

∘ A **softmax response** of agent i to the policies of the other agents $\boldsymbol{\pi}_{-i}$ is a policy π_i such that

$$\pi_i\left(a_i\right) \propto \exp\left(\lambda r_i\left(a_i, \boldsymbol{\pi}_{-i}\right)\right)$$

Remarks:

∘ A best response can be either deterministic or mixed.

∘ when $\lambda \to \infty$ coincides softmax response with best response.

Normal form games

- What is normal form game?
- Equilibria
  - Dominant Strategy Equilibrium
  - Nash Equilibrium
- Dynamics for games
  - Iterated best response
  - Fictitious play
  - Gradient ascent

Dominant strategy equilibrium

- A **dominant strategy** π_i for player i is a strategy that is a best response against all $\boldsymbol{\pi}_{-i}$

$$r_i\left(\pi_i, \boldsymbol{\pi}_{-i}\right) \geq r_i\left(\widetilde{\pi}_i, \boldsymbol{\pi}_{-i}\right), \quad \forall \widetilde{\pi}_i, \boldsymbol{\pi}_{-i}$$

- In a **dominant strategy equilibrium**, every player adopts a dominant strategy.
- Dominant strategy and dominant strategy equilibrium may not exist.
- (defect, defect) is a dominant strategy equilibrium in prisoner dilemma game

		Bob: cooperate	Bob: defect
Alex	cooperate	$1/1$	$-1/2$
	defect	$2/-1$	$0/0$

- Bob can always improve his payoff by defecting (irrespectable of Alex's strategy)

Nash equilibrium

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

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

◦ In NE, we can write for each agent i

$$r_i(\boldsymbol{\pi}^\star) \geq r_i(\pi_i, \boldsymbol{\pi}^\star_{-i}), \quad \forall \pi_i.$$

◦ All dominant strategy equilibria are Nash equilibria (the reverse does not hold).

Nash equilibrium - good news

- Rock-paper-scissor game

		Bob		
		rock	paper	scissor
Alex	rock	$0/0$	$-1/1$	$1/-1$
	paper	$1/-1$	$0/0$	$-1/1$
	scissor	$-1/1$	$1/-1$	$0/0$

- No dominant strategy equilibrium. No pure NE.

- Each player playing a mixed strategy $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ is a NE.

Theorem (Existence of Nash equilibrium [13])

In a normal form game with finite players and actions, there exists a Nash equilibrium in mixed strategies.

Computing Nash equilibrium

- Consider a game with different payoff matrices

$$r_1(\pi_1, \pi_2) = (\pi_1)^\top A \pi_2 \quad \text{(player 1)}$$
$$r_2(\pi_1, \pi_2) = (\pi_1)^\top B \pi_2 \quad \text{(player 2)}$$

- **Bad news** Computing mixed NE in normal form games is intractable in general [2, 4].
- **Good news** However, NE of zero-sum games ($A = -B^\top$) can be efficiently computed as we will see.

Nash equilibria in two-player zero-sum games

- We can find a Nash equilibrium by solving a minimax formulation
- Consider the following bilinear minimax optimization problems

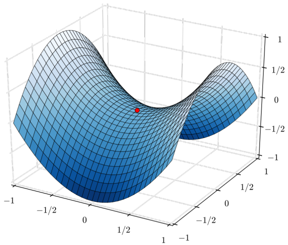
$$\max_{\pi_1 \in \Delta^{d_1}} \min_{\pi_2 \in \Delta^{d_2}} (\pi_1)^\top A \pi_2 \quad \text{(player 1)}$$

$$\min_{\pi_2 \in \Delta^{d_2}} \max_{\pi_1 \in \Delta^{d_1}} (\pi_1)^\top A \pi_2 \quad \text{(player 2)}$$

- NE corresponds to $(\pi_1^\star, \pi_2^\star)$ such that

$$(\pi_1)^\top A \pi_2^\star \le (\pi_1^\star)^\top A \pi_2^\star \le (\pi_1^\star)^\top A \pi_2, \quad \forall \pi_1, \pi_2$$

- It is also called a saddle point for the function $f(\pi_1, \pi_2) = (\pi_1)^\top A \pi_2$.



Connection with minimax optimization

- More generally $(x^\star, y^\star)$ is called a saddle point for f if

$$f(x^\star, y) \leq f(x^\star, y^\star) \leq f(x, y^\star) \tag{1}$$

Theorem (Minimax theorem)

Let $X \in \mathbb{R}^{d_1}$ and $Y \in \mathbb{R}^{d_2}$ be compact convex sets. If $f : X \times Y \to \mathbb{R}$ is a continous function such that $f(\cdot, y)$ is convex for any y and $f(x, \cdot)$ is concave for any x then

$$\max_{x \in X} \min_{y \in Y} f(x, y) = \min_{y \in Y} \max_{x \in X} f(x, y). \qquad \text{(minimax equality)}$$

Proposition: - $(x^\star, y^\star)$ is a saddle point for f if and only if the minimax equality holds and

$$x^\star \in \arg\min_{x \in X} \max_{y \in Y} f(x, y), \quad y^\star \in \arg\max_{y \in Y} \min_{x \in X} f(x, y).$$

Normal form games

- What is normal form game?
- Equilibria
  - Dominant Strategy Equilibrium
  - Nash Equilibrium
  - Correlated Equilibrium
- Dynamics for games
  - Iterated best response
  - Fictitious play
  - Gradient ascent

Iterated best response

◦ Each player iteratively find the best response to other player's strategies

Iterated best response (IBR)

for $t = 1, \ldots$ **do**

Each player i updates its strategy π_i^{t+1} such that

$$r_i\left(\pi_i^{t+1}, \boldsymbol{\pi}_{-i}^t\right) \geq r_i\left(\pi_i, \boldsymbol{\pi}_{-i}^t\right), \quad \forall \pi_i$$

end for

Remark: ◦ Players can update simultaneously or sequentially.

Non-convergence of iterated best response - bad news

- Starting from (T,L), two players update simultaneously.
- After 2 iterations, it arrives NE (B,R).

Player Y

Player X		L	R
	T	$1/2$	$3/1$
	B	$2/1$	$4/3$

- Starting from (A, B), two players update simultaneously.
- (A,B) $\rightarrow$ (B,A) $\rightarrow$ (A,B)$\rightarrow$...
- It avoids NEs (A,A) and (B,B).

Player Y

Player X		A	B
	A	$1/1$	$0/0$
	B	$0/0$	$1/1$

Convergence of IBR in potential games - good news

- The potential function for a game is a function $\Phi : \mathcal{A} \to \mathbb{R}$ such that

$$r_i\left(a_i, a_{-i}\right) - r_i\left(\widetilde{a}_i, a_{-i}\right) = \Phi\left(a_i, a_{-i}\right) - \Phi\left(\widetilde{a}_i, a_{-i}\right), \quad \forall a_i, \widetilde{a}_i \in \mathcal{A}_i, a^{-i} \in \mathcal{A}_{-i}.$$

- A game with a potential function is called potential game.

		Player Y	
		cooperate	defect
Player X	cooperate	$1/1$	$-1/2$
	defect	$2/-1$	$0/0$

Table: Prisoner's dilemma

		Player Y	
		cooperate	defect
Player X	cooperate	$\Phi = 0$	$\Phi = 1$
	defect	$\Phi = 1$	$\Phi = 2$

Table: Potential function

Proposition

If a potential game is finite, it has at least one pure Nash equilibrium. If players use iterated best response sequentially (or one at a time), the dynamic will terminate at a NE after finite step.

Fictitious play

- **Feedback** In fictitious play each agent i counts opponent's actions $N_t(j, a_j)$ for $j \neq i$. The initial counts $N_0(j, a_j)$ can be based on agents' initial guess.

- **Behavioural assumption** Each agent i assumes its opponents are using a stationary mixed strategy the same as empirical distribution of their actions

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

- Each agent i maximizes their reward assuming other agents are playing $\widetilde{\boldsymbol{\pi}}_{-i}^t$.

$$a_i^{t+1} = \max_{a_i} r_i(a_i, \widetilde{\boldsymbol{\pi}}_{-i}^t).$$

Non-convergence of fictitious play - bad news

◦ Fictitious play is not guaranteed to converge.

◦ Consider the following game (also known as the Shapley game [16])

		Player Y		
		Left	**C**enter	**R**ight
Player X	**T**op	$0/0$	$1/0$	$0/1$
	Middle	$0/1$	$0/0$	$1/0$
	Bottom	$1/0$	$0/1$	$0/0$

Table: Sharpley's dilemma

◦ The policy cycles: $(\mathrm{T},\mathrm{C}) \to (\mathrm{T},\mathrm{R}) \to (\mathrm{M},\mathrm{R}) \to (\mathrm{M},\mathrm{L}) \to (\mathrm{B},\mathrm{L}) \to (\mathrm{B},\mathrm{C}) \to (\mathrm{T},\mathrm{C}) \to \ldots$

◦ After one play stays on a wining position long enough, the other player will change its action

◦ Empirical distributions do not converge.

Convergence of fictitious play in some games - good news

◦ Fictitious play converges for two-player zero-sum games

Theorem ([15])

For two-player zero-sum games the empirical distribution of fictitious play converges to a NE, i.e. $(\widetilde{\pi}_1^t, \widetilde{\pi}_2^t) \to (\pi_1^\star, \pi_2^\star)$ *where* $(\pi_1^\star, \pi_2^\star)$ *is a NE.*

Karlin's conjecture [5]

The convergence rate of fictitious play for zero-sum games is $\mathcal{O}(1/\sqrt{T})$.

Remark: ◦ Still an open problem

Gradient ascent

- **Feedback** Assume agent i has access to all other mixed strategies π_j for $j \neq i$.
- Take the gradient of value function at $\boldsymbol{\pi}^t$: $\frac{\partial r_i(\boldsymbol{\pi})}{\partial \pi_i(a_i)}\Big|_{\boldsymbol{\pi}=\boldsymbol{\pi}^t}$.
- Apply gradient ascent to each agent

$$\pi_i^{t+1}(a_i) = \pi_i^t(a_i) + \alpha_i^t \frac{\partial r_i(\boldsymbol{\pi})}{\partial \pi_i(a_i)}\Big|_{\boldsymbol{\pi}=\boldsymbol{\pi}^t}.$$

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

$$\frac{\partial r_i(\boldsymbol{\pi})}{\partial \pi_i(a_i)}\Big|_{\boldsymbol{\pi}=\boldsymbol{\pi}^t} = \frac{\partial}{\partial \pi_i(a_i)} \left(\sum_{\boldsymbol{a}} r_i(\boldsymbol{a}) \prod_j \pi_j(a_j) \right)\Bigg|_{\boldsymbol{\pi}=\boldsymbol{\pi}_t} = \sum_{\boldsymbol{a}_{-i}} r_i(a_i, \boldsymbol{a}_{-i}) \prod_{j \neq i} \pi_j^t(a_j).$$

Gradient ascent in two-player zero-sum games

◦ The bilinear minimax optimization

$$\min_{\pi_2 \in \Delta^{d_2}} \max_{\pi_1 \in \Delta^{d_1}} (\pi_1)^\top A \pi_2$$

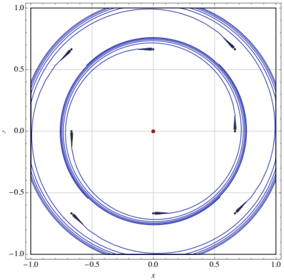
◦ Gradient ascent (also called gradient descent ascent or GDA in this case)

$$\begin{aligned} \pi_1^{t+1} &= \mathcal{P}_{\Delta^{d_1}} \left(\pi_1^t + \alpha_1^t A \pi_2^t \right), \\ \pi_2^{t+1} &= \mathcal{P}_{\Delta^{d_2}} \left(\pi_2^t - \alpha_2^t A^\top \pi_1^t \right). \end{aligned}$$

◦ Gradient descent ascent with constant stepsizes (i.e. $\alpha_1^t = \alpha_1$ and $\alpha_2^t = \alpha_2$) does not always converge for bilinear minimax optimization [9].

Gradient ascent in two-player zero-sum games - non-convergence

- The function $f(x, y) = xy$ has saddle point $(0, 0)$.
- GDA update $x_{t+1} = x_t - \alpha y_t$, $y_{t+1} = y_t + \alpha x_t$
- Since $x_{t+1}^2 + y_{t+1}^2 = (1 + \alpha^2)(x_t^2 + y_t^2)$, it does not converge to the saddle point.



- GDA with constant stepsize may not converge even if $f(x, y)$ is convex-concave!

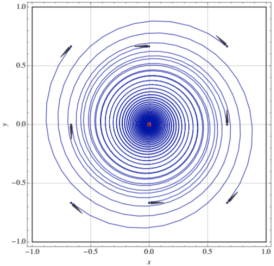
Extra-gradient - a simple fix to GDA

- Minimax optimization:

$$\min_{x\in X}\max_{y\in Y} f(x,y).$$

- Extra-gradient (EG) update:

$$x_{t+\frac{1}{2}} = \mathcal{P}_X\Big(x_t - \alpha\nabla_x f(x_t,y_t)\Big), \qquad y_{t+\frac{1}{2}} = \mathcal{P}_Y\Big(y_t + \alpha\nabla_y f(x_t,y_t)\Big)$$
$$x_{t+1} = \mathcal{P}_X\left(x_t - \alpha\nabla_x f(x_{t+\frac{1}{2}},y_{t+\frac{1}{2}})\right), \quad y_{t+1} = \mathcal{P}_Y\left(y_t + \alpha\nabla_y f(x_{t+\frac{1}{2}},y_{t+\frac{1}{2}})\right)$$



Convergence of extra-gradient

- Assumption 1: $f(x, y)$ is convex-concave,
- Assumption 2: $f(x, y)$ is L-smooth,
- Assumption 3: $D_X^2 = \frac{1}{2}\max_{x,x'} \|x - x'\|^2$ and $D_Y^2 = \frac{1}{2}\max_{y,y'} \|y - y'\|^2$ are finite.

Theorem

If the assumptions above holds, then EG with stepsize $\alpha = \frac{1}{2L}$ satisfies

$$f(\bar{x}_T, y) - f(x, \bar{y}_T) \leq \frac{2L(D_X^2 + D_Y^2)}{T}.$$

for any $x \in X$ and $y \in Y$ where $\bar{x}_T = \frac{1}{T}\sum_{t=1}^T x_t$ and $\bar{x}_T = \frac{1}{T}\sum_{t=1}^T y_t$.

Remarks:
- The time average $(\bar{x}_T, \bar{y}_T)$ produced by EG converges to a saddle point.
- For strongly-convex strongly-concave see Mathematics of Data lecture 14 2022 (EE-556) [1]

Beyond normal form games / convex-concave

◦ So far focused on normal form games (contained in convex-concave)

General zero-sum games

Consider

$$\min_{x \in X} \max_{y \in Y} f(x, y) \tag{2}$$

where $f(\cdot, y)$ is nonconvex and $f(x, \cdot)$ is nonconcave.

Remarks:

◦ If $f(x, y) = x^\top A y$ and $\mathcal{X} = \Delta$ and $\mathcal{Y} = \Delta$ this reduces to a normal form game.

◦ x,y can be the parameters of deep neural networks (e.g., generative adversarial networks)

Beyond normal form games / convex-concave

◦ A **Nash equilibrium** (NE) is a pair $(x^\star, y^\star) \in \mathcal{X} \times \mathcal{Y}$ for which,

$$f(x^\star, y) \leq f(x^\star, y^\star) \leq f(x, y^\star) \quad \forall x \in \mathcal{X}, y \in \mathcal{Y} \tag{3}$$

◦ A **local Nash equilibrium** (LNE) is a pair $(x^\star, y^\star) \in \mathcal{X} \times \mathcal{Y}$ for which,

$$f(x^\star, y) \leq f(x^\star, y^\star) \leq f(x, y^\star) \quad \text{for all } (x, y) \text{ in a neighborhood } \mathcal{U} \text{ of } (x^\star, y^\star) \text{ in } \mathcal{X} \times \mathcal{Y} \tag{4}$$

◦ A **first order stationary point** (FOSP) is a pair $(x^\star, y^\star) \in \mathcal{X} \times \mathcal{Y}$ for which,

$$\begin{aligned} \nabla_x f(x^\star, y^\star)^\top (x - x^\star) \geq 0 \quad \forall x \in \mathcal{X} \\ \nabla_y f(x^\star, y^\star)^\top (y - y^\star) \leq 0 \quad \forall y \in \mathcal{Y} \end{aligned} \tag{5}$$

Remarks:

◦ NE $\Rightarrow$ LNE $\Rightarrow$ FOSP

◦ In case f is not convex-concave Nash equilibrium may not exist

Nonconvex-nonconcave - bad news

- Computing FOSP is PPAD-complete (similar to NP-completeness) [6]
- Large family of methods (including extra-gradient) may not converge to FOSP [11]
- **Example** [11]

$$f(x, y) = y(x - 0.5) + \phi(y) - \phi(x) \quad \text{where} \quad \phi(u) = \frac{1}{4}u^2 - \frac{1}{2}u^4 + \frac{1}{6}u^6 \tag{6}$$

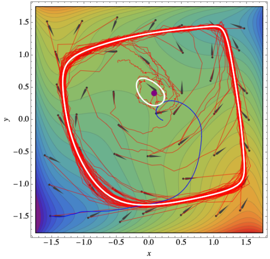


Figure: Neither last iterate (red) or time average (blue) of extra-gradient does converge to a FOSP.

Summary

- Normal form games:
  - What is normal form game?
  - Equilibrium
  - Algorithms for games

Table: Does the algorithm converge?

Setting (solution concept)	Best response	Fictitious play	GDA	Extra-gradient
Potential games (NE)	Yes	Yes	Yes	Yes
Normal form games (NE)	No	No	No	No
Zero-sum games (NE)	No	Yes[1]	Yes[2]	Yes
general zero-sum games (FOSP)	No	No	No	No

Remarks:

- All require full access on the payoff vector (**oracle based**)
- Weaker feedback model (**loss based**):
  - only access to randomly sampled pure strategy of opponents (e.g. Exp3 [10])

[1] Rates for fictitious play is still open.

[2] The time average of GDA converges for an appropriate stepsize selection. However, fixed stepsize does not.

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 [17] 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**: $\mathsf{P}\left(s' \mid s, \boldsymbol{a}\right)$ 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	$a_i \in \mathcal{A}_i$	$s \in \mathcal{S}$	$\mathsf{P}(s' \mid s, \boldsymbol{a})$	$r_i(s, \boldsymbol{a})$	$\pi_i(a_i \mid s)$	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 [8])

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 [7]

- 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[7])

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 $\mathcal{D}_0$

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

Non-convergence of policy gradient algorithms in linear quadratic games

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

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

- Policy gradient algorithms

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

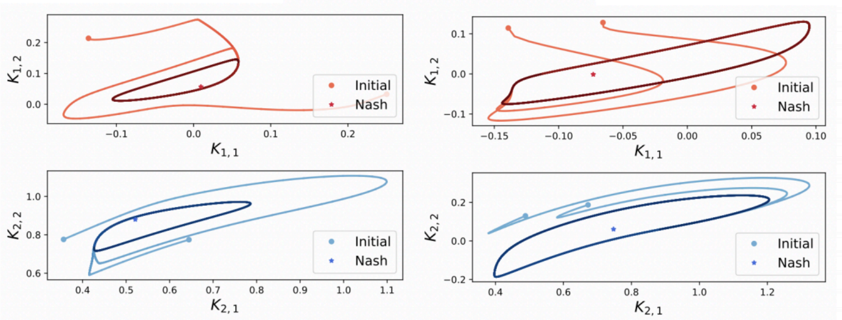
Theorem (Non-convergence of policy gradient in LQ games [12])

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 [17]

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** $s \in S$,

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

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 [3]
- Optimistic mirror decent + actor-critic [18]
- 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 [3]

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, [3])

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 [18]

- 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 [18]

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 I

[1] Volkan Cevher.
Lecture 14: Primal-dual optimization ii: The extra-gradient method (Mathematics of Data 2022).
https://www.epfl.ch/labs/lions/wp-content/uploads/2023/02/lecture_14_2022.pdf.
29

[2] Xi Chen, Xiaotie Deng, and Shang-Hua Teng.
Settling the complexity of computing two-player nash equilibria.
Journal of the ACM (JACM), 56(3):1–57, 2009.
15

[3] Constantinos Daskalakis, Dylan J. Foster, and Noah Golowich.
Independent policy gradient methods for competitive reinforcement learning, 2021.
60, 61, 62

[4] Constantinos Daskalakis, Paul W Goldberg, and Christos H Papadimitriou.
The complexity of computing a nash equilibrium.
SIAM Journal on Computing, 39(1):195–259, 2009.
15

[5] Constantinos Daskalakis and Qinxuan Pan.
A counter-example to karlin's strong conjecture for fictitious play.
In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 11–20. IEEE, 2014.
24

References II

[6] Constantinos Daskalakis, Stratis Skoulakis, and Manolis Zampetakis.
The complexity of constrained min-max optimization.
arXiv preprint arXiv:2009.09623, 2020.
32

[7] 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.
44, 45

[8] 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.
42

[9] Gauthier Gidel, Reyhane Askari Hemmat, Mohammad Pezeshki, Rémi Le Priol, Gabriel Huang, Simon Lacoste-Julien, and Ioannis Mitliagkas.
Negative momentum for improved game dynamics.
In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1802–1811, 2019.
26

[10] Elad Hazan.
Introduction to online convex optimization.
Foundations and Trends® in Optimization, 2(3-4):157–325, 2016.
33

References III

[11] Ya-Ping Hsieh, Panayotis Mertikopoulos, and Volkan Cevher.
The limits of min-max optimization algorithms: Convergence to spurious non-critical sets.
arXiv preprint arXiv:2006.09065, 2020.
32

[12] 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.
50

[13] John F Nash Jr.
Equilibrium points in n-person games.
Proceedings of the national academy of sciences, 36(1):48–49, 1950.
14

[14] William Poundstone.
Prisoner's Dilemma/John Von Neumann, game theory and the puzzle of the bomb.
Anchor, 1993.
8

[15] Julia Robinson.
An iterative method of solving a game.
Annals of mathematics, pages 296–301, 1951.
24

References IV

[16] Lloyd Shapley.
Some topics in two-person games.
Advances in game theory, 52:1–29, 1964.
23

[17] Lloyd S Shapley.
Stochastic games.
Proceedings of the national academy of sciences, 39(10):1095–1100, 1953.
35, 58

[18] 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.
60, 63