Theory and Methods for Reinforcement Learning

Prof. Volkan Cevher
volkan.cevher@epfl.ch

Lecture 9: 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)

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

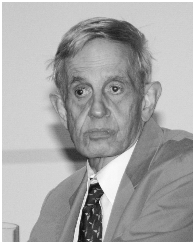


Figure: John Nash

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 [10]: each agent can choose to cooperate or defect

		Bob	
		cooperate	defect
Alex	cooperate	$1/1$	$-1/2$
	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	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 [9])

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

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

- **Bad news** Computing mixed NE in normal form games is intractable in general [2, 3].
- **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

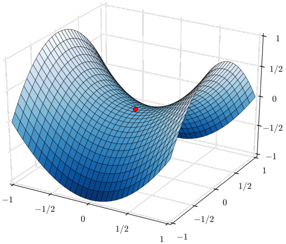
$$\max_{\pi_1 \in \Delta^{d_1}} \min_{\pi_2 \in \Delta^{d_2}} (\pi_1)^\top A \pi_2 \quad \text{(player 1)}$$

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

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

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

- 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 $\pi_i^{t+1}$ 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	
		L	R
Player X	T	$1/2$	$3/1$
	B	$2/1$	$4/3$

- ○ Starting from (A, B), two players update simultaneously.
- ○ (A,B) → (B,A) → (A,B)→...
- ○ It avoids NEs (A,A) and (B,B).

		Player Y	
		A	B
Player X	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

- **Required 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 [12])

		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: $(\mathrm{T},\mathrm{C}) \to (\mathrm{T},\mathrm{R}) \to (\mathrm{M},\mathrm{R}) \to (\mathrm{M},\mathrm{L}) \to (\mathrm{B},\mathrm{L}) \to (\mathrm{B},\mathrm{C}) \to (\mathrm{T},\mathrm{C}) \to \dots$
- 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 zero-sum games

Theorem ([11])

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 [4]

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

Remark: - Still an open problem

Gradient ascent

- 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

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

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

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

Gradient ascent in two-player zero-sum games

◦ The bilinear minimax optimization

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

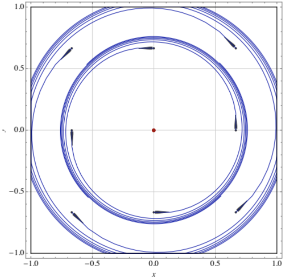
◦ Gradient ascent (also called gradient descent ascent or GDA in this case)

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

◦ 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 [6].

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

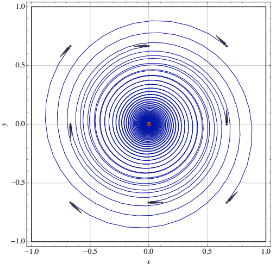
- Minimax optimization:

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

- 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 12 2021 (EE-556) [1]

Beyond normal form games / convex-concave

- ○ So far focused on normal form (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 $f(x,y) = x^\top Ay$ and $\mathcal{X} = \Delta$ and $\mathcal{X} = \Delta$ 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,

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

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

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

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

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

Remarks:

◦ NE ⇒ LNE ⇒ 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) [5]
- ◦ Large family of methods (including extra-gradient) may not converge to FOSP [8]
- ◦ **Example** [8]

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

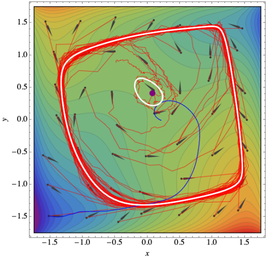


Figure: Neither last iterate (red) or time average (blue) of extra-gradient does converge to a FOSP.

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	No	No[1]	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 [7])

[1] The time average converges for an appropriate stepsize selection.

References I

[1] Volkan Cevher.
Lecture 12: Primal-dual optimization II: Extra-gradient method (Mathematics of Data 2021).
https://www.epfl.ch/labs/lions/wp-content/uploads/2022/01/lecture_12_2021.pdf.

[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.

[3] 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.

[4] 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.

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

[6] 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.

References II

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

[8] 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.

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

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

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

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