Pursuit-Evasion Games II

Tomáš Kroupa

Department of Computer Science
Faculty of Electrical Engineering
Czech Technical University in Prague

2022

What we have learned

- Matrix games capture the strategic interaction between two players
- Their solutions are found by LP
- Efficient algorithms
- Single stage games

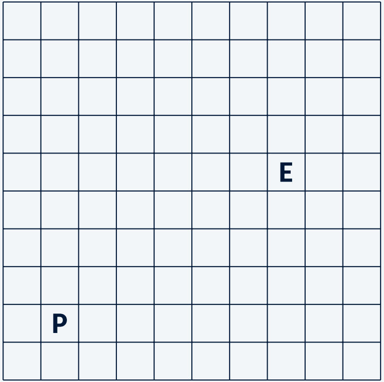
Where we will go

- Dynamic games
- Iterative interaction between players in the changing environment
- Scalability issues

Stochastic Games

Repeating Zero-Sum Games

Or making MDPs competitive?



- Pursuer **P** tries to capture evader **E**
- Stochastic policy describes the mixed strategy of each player in every state
- We are seeking a common generalization of
  - repeated TPZS games and
  - Markov decision processes (MDPs)

1. Players are the planner and the adversary
2. Finite action sets M and N for the planner and the adversary, respectively
3. Finite set S of states
4. Transition function

$$T\colon S \times M \times N \to \Delta_S$$

where $T(s, i, j)$ is a probability distribution on S

5. Reward function

$$R\colon S \times M \times N \to \mathbb{R}$$

where $R(s, i, j)$ is a reward to the planner

How to Play a Stochastic Game?

An initial state is $s_0 \in S$. At each stage $\tau = 0, 1, 2, \ldots$:

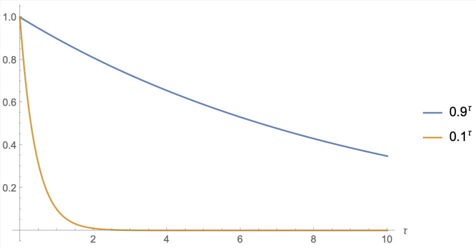
1. The players choose $(i_\tau, j_\tau) \in M \times N$ observing the history of past states/actions
2. The planner receives $R(s_\tau, i_\tau, j_\tau)$ from the adversary
3. A new state $s_{\tau+1} \in S$ is determined randomly according to $T(s_\tau, i_\tau, j_\tau)$

How to aggregate the rewards over an infinite history?

Discounting

- A discount factor is a number $\gamma \in (0, 1)$
- The discounted reward of planner over an infinite history $(s_0, i_0, j_0, \ldots)$ is

$$\sum_{\tau=0}^{\infty} \gamma^{\tau} R(s_\tau, i_\tau, j_\tau) = R(s_0, i_0, j_0) + \gamma R(s_1, i_1, j_1) + \gamma^2 R(s_2, i_2, j_2) + \cdots$$



At stage τ the players observe an entire history

$$h_\tau := (s_0, i_0, j_0, \ldots, s_\tau, i_\tau, j_\tau).$$

A strategy of the planner is

1. behavioral if it is a mapping $h_\tau \mapsto p \in \Delta_M$ from histories to mixed actions
2. Markov if it is a behavioral strategy depending only on the current state s_τ
3. stationary if it is a Markov strategy not depending on time τ

Stationary Strategies

A stationary strategy for the planner/adversary is a mapping

$$\pi : S \to \Delta_M \quad \text{and} \quad \sigma : S \to \Delta_N,$$

respectively.

Special cases

- If $|S| = 1$, then we obtain a two-person zero-sum game and the concept of mixed strategy
- If $|N| = 1$, then we get an MDP with the concept of stochastic policy

How to evaluate stationary strategies?

Value/Quality Function

👉 $(s_\tau), (i_\tau), (j_\tau)$ are stochastic processes with respect to T, π, σ

- The value function is the expected reward starting from state s and then following strategies π and σ,

$$\mathcal{V}_{\pi,\sigma}(s) = \mathbb{E}\left(\sum_{\tau=0}^{\infty} \gamma^\tau R(s_\tau, i_\tau, j_\tau)\right), \qquad s_0 = s.$$

- The Q-function is the expected reward starting from state s, taking actions i and j, and then following strategies π and σ,

$$\mathcal{Q}_{\pi,\sigma}(s, i, j) = \mathbb{E}\left(\sum_{\tau=0}^{\infty} \gamma^\tau R(s_\tau, i_\tau, j_\tau)\right), \qquad s_0 = s, i_0 = i, j_0 = j.$$

$$\mathcal{V}_{\pi,\sigma}(s) = \sum_{i \in M} \sum_{j \in N} \pi_s(i) \cdot \sigma_s(j) \cdot \mathcal{Q}_{\pi,\sigma}(s, i, j)$$

$$\mathcal{Q}_{\pi,\sigma}(s, i, j) = R(s, i, j) + \gamma \sum_{s' \in S} T(s, i, j)(s') \cdot \mathcal{V}_{\pi,\sigma}(s')$$

Equation for the Value Function

$$\mathcal{V}_{\pi,\sigma}(s) = \sum_{i \in M} \sum_{j \in N} \pi_s(i) \cdot \sigma_s(j) \cdot \left(R(s,i,j) + \gamma \sum_{s' \in S} T(s,i,j)(s') \cdot \mathcal{V}_{\pi,\sigma}(s') \right)$$

Equation for the Q-Function

$$\mathcal{Q}_{\pi,\sigma}(s,i,j) = R(s,i,j) + \gamma \sum_{s' \in S} T(s,i,j)(s') \sum_{i' \in M} \sum_{j' \in N} \pi_{s'}(i') \cdot \sigma_{s'}(j') \cdot \mathcal{Q}_{\pi,\sigma}(s',i',j')$$

Equilibrium of a Stochastic Game

☞ Maximin/minimax

- A stationary strategy π^* is an equilibrium strategy of the planner if

$$\max_{\pi} \min_{\sigma} \mathcal{V}_{\pi,\sigma}(s) = \min_{\sigma} \mathcal{V}_{\pi^*,\sigma}(s) \qquad \text{for every } s \in S,$$

  and analogously for the adversary
- Every stochastic game has an equilibrium (π^*, σ^*) in stationary strategies

Quality/Value Functions

👉 For equilibrium strategies (π^*, σ^*)

Define:

$$\mathcal{Q}_*(s, i, j) = R(s, i, j) + \gamma \sum_{s' \in S} T(s, i, j)(s') \cdot \mathcal{V}_*(s')$$

$$\mathcal{V}_*(s) = \mathcal{V}_{\pi^*, \sigma^*}(s)$$

How to obtain the estimate of $\mathcal{V}_*(s)$?

Value Iteration

☞ A variant for stochastic games

$$\mathcal{Q}(s,i,j) := R(s,i,j) + \gamma \sum_{s' \in S} T(s,i,j)(s') \cdot \mathcal{V}(s') \tag{1}$$

$$\mathcal{V}(s) := \max_{\pi_s \in \Delta_M} \min_{j \in N} \sum_{i \in M} \pi_s(i) \cdot \mathcal{Q}(s,i,j) \tag{2}$$

- Initialize $\mathcal{V}$ arbitrarily
- Update (1)–(2) iteratively
- This procedure converges to $\mathcal{V}_*$ (Shapley, 1953)

- Set a learning rate $\alpha \in (0,1)$

$$\mathcal{Q}(s,i,j) := (1-\alpha)\mathcal{Q}(s,i,j) + \alpha(R(s,i,j) + \gamma\mathcal{V}(s)) \tag{3}$$

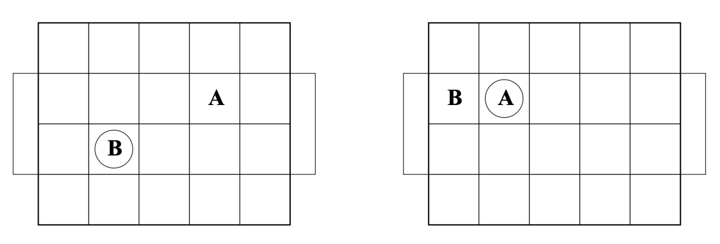
$$\mathcal{V}(s) := \max_{\pi_s \in \Delta_M} \min_{j \in N} \sum_{i \in M} \pi_s(i) \cdot \mathcal{Q}(s,i,j) \tag{4}$$

- Initialize $\mathcal{V}$ and $\mathcal{Q}$ arbitrarily
- Update (3)–(4) iteratively

Examples

Soccer (Littman, 1994) with $\gamma = 0.9$

- 780 states (players' position and ball possession)
- 5 actions (N, S, E, W, stand)
- Random change of ball possession and the action selection



Examples

Goofspiel (Flood, 1930s)

- Each player and the deck have one suit of cards numbered $1, \ldots, n$
- A card from the deck is bid on secretly
- The player with the highest card gets the corresponding reward
- Each player discards the card bid
- Repeat for n rounds

n	$\|S\|$	$\|S \times A\|$	SIZEOF(π or Q)	V(det)	V(random)
4	692	15150	$\sim$ 59KB	-2	-2.5
8	3×10^6	1×10^7	$\sim$ 47MB	-20	-10.5
13	1×10^{11}	7×10^{11}	$\sim$ 2.5TB	-65	-28

References

- Bowling, Michael, and Manuela Veloso. ”Scalable learning in stochastic games.” AAAI Workshop on Game Theoretic and Decision Theoretic Agents. 2002.
- Littman, Michael L. 1994. Markov Games as a Framework for Multi-Agent Reinforcement Learning. *Machine Learning Proceedings*, 1994. https://doi.org/10.1016/b978-1-55860-335-6.50027-1.