L17 – Week 9
Introduction to Multi-agent RL

CS 295 Optimization for Machine Learning

Ioannis Panageas

The framework

A finite Markov Game or Stochastic Game is defined as follows:

- N agents
- A finite state space S.
- A finite action space $A := A_1 \times ... \times A_n$.
- A transition model P where $P(s'|s, a_1, ..., a_n)$ is the probability of transitioning into state s' upon agent i taking action a_i in state s. P is a matrix of size $(S \cdot A) \times S$.
- Reward function $r_i : S \times A \to [0, 1]$ for each i.
- A discounted factor $\gamma \in [0, 1)$.

The framework

Goal: Find a Nash policy $\pi^* = (\pi_1^*, ..., \pi_n^*)$ with $\pi^* : S \to \Delta(A_1) \times ... \times \Delta(A_n)$, that is

$$V_i^{\pi_i, \pi^*_{-i}}(s) = (1-\gamma)\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_i(s_t, a_{1,t}, ..., a_{n,t}) | (\pi_i, \pi^*_{-i}), s_0 = s\right]$$

is maximized for $\pi_i = \pi_i^*$. This is the Infinite Time Horizon case.

The framework

Goal: Find a Nash policy $\pi^* = (\pi_1^*, ..., \pi_n^*)$ with $\pi^* : S \to \Delta(A_1) \times ... \times \Delta(A_n)$, that is

$$V_i^{\pi_i, \pi^*_{-i}}(s) = (1-\gamma)\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_i(s_t, a_{1,t}, ..., a_{n,t}) \big| (\pi_i, \pi^*_{-i}), s_0 = s\right]$$

is maximized for $\pi_i = \pi_i^*$. This is the Infinite Time Horizon case.

Remarks

- Fixing all agents but i, induces a classic MDP. Every agent plays best response.
- Generalizes notion of Nash Equilibrium.
- Nash policy always exist!

Two player zero sum

p

$$\text{state } s_0 = \begin{pmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix} \qquad \text{state } s_1 = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

q

Two agents, it holds $r_2(s, a_1, a_2) = -r_1(s, a_1, a_2)$!

Two player zero sum

p

$$\text{state } s_0 = \begin{pmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix} \qquad \text{state } s_1 = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

q

Two agents, it holds $r_2(s, a_1, a_2) = -r_1(s, a_1, a_2)!$

Remarks

- Shapley showed that such games have value, i.e., minmax = maxmin.
- This fact was used recently in Daskalakis et al 2020 to show policy gradient converges to the Nash policy!
- The general problem is hard!

Gradient Policy Iteration

Definition (Direct Parametrization). *Every agent uses the following:*

$$\pi_i(a \mid s) = x_{i,s,a}$$

with $x_{i,s,a} \geq 0$ *and* $\sum_{a \in A_i} x_{i,s,a} = 1$.

Definition (Policy Gradient Ascent). *PGA is defined iteratively:*

$$\pi_i^{(t+1)} := P_{\Delta(A_i)^S}(\pi_i^{(t)} + \eta \nabla_{\pi_i} V^i(\pi^{(t)})),$$

where P denotes projection on simplex.

Gradient Policy Iteration

Definition (Policy Gradient Ascent). *PGA is defined iteratively:*

$$\pi_i^{(t+1)} := P_{\Delta(A_i)^S}(\pi_i^{(t)} + \eta \nabla_{\pi_i} V^i(\pi^{(t)})),$$

where P denotes projection on simplex.

Theorem (Policy Gradient Ascent). *It can be shown for one agent that after $O(1/\epsilon^2)$ iterations, an ϵ-optimal policy can be reached. With some modifications, it works for two-player zero sum games too.*

Remarks

- No guarantees for more than two players (only very specific settings).
- Can we find other classes of stochastic games that PGA converges?

Beyond two player zero sum: Markov Potential Games

Definition (Markov Potential Game). *A Markov Decision Process (MDP), G, is called a* Markov Potential Game (MPG) *if there exists a (state-dependent) function* $\Phi_s : \Pi \to \mathbb{R}$ *for* $s \in S$ *so that*

$$\Phi_s(\pi_i, \pi_{-i}) - \Phi_s(\pi_i', \pi_{-i}) = V_s^i(\pi_i, \pi_{-i}) - V_s^i(\pi_i', \pi_{-i}),$$

for all agents $i \in N$*, all states* $s \in S$ *and all policies* $\pi_i, \pi_i' \in \Pi_i, \pi_{-i} \in \Pi_{-i}$.

Remarks

- This notion generalizes the Potential Games in Game Theory.
- Potential Games capture routing (congestion games), important class.
- Deterministic Nash policies always exist!
- Each state a potential game does not imply MPG. Might have also zero sum game states!

Not a Markov Potential Game

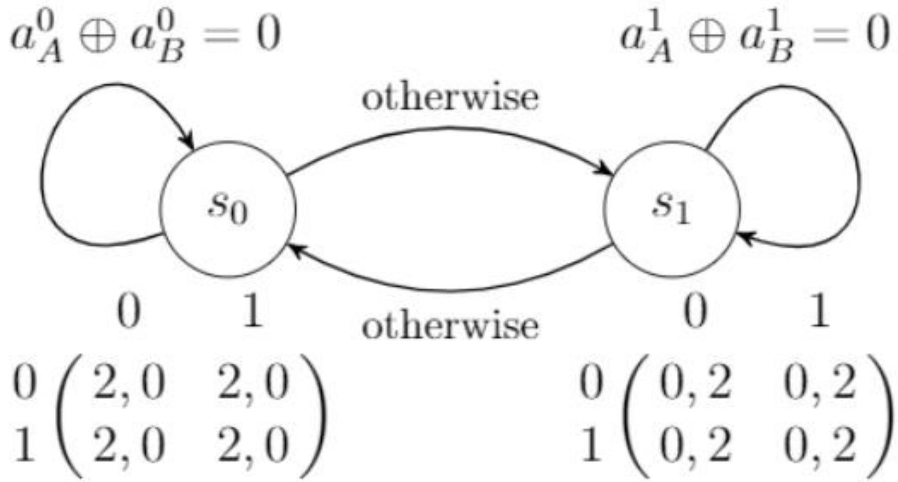

Figure 1: A MDP which is potential at every state but which not a MPG due to conflicting preferences over states. The agents' instantaneous rewards, $(R_A(s, \mathbf{a}), R_B(s, \mathbf{a}))$, are shown in matrix form below each state $s = 0, 1$.

Gradient Policy Iteration

Theorem (PGA for Markov Potential Games). *Suppose all agents run policy gradient iteration independently and update simultaneously. It can be shown that after $O(1/\epsilon^2)$ iterations, an ϵ-Nash policy can be reached.*

Remarks

- This result can be generalized if agents do not have access to exact gradients.
- It matches the result for single-agent.
- Proof steps on the board!

Conclusion

- Introduction to Markov Games.
  - Stochastic Games
  - Potential Games
  - Policy Gradient

- I hope you enjoyed the Lectures!