6.S890: Topics in Multi-Agent Learning

Lecture 9: Stochastic Games Supplement

↳ Existence of Nash Equilibrium in infinite horizon, discounted, stochastic games w/ finite states and actions

Theorem: Every infinite-horizon discounted stochastic game w/ a finite #players, #states, #actions has a Nash equilibrium in stationary Markov policies.

That is there $\exists$ a collection of policies $\pi_1, \dots, \pi_m$ where $\pi_i : \underset{\uparrow \text{states}}{S} \to \Delta(\underset{\uparrow \text{actions of player } i}{A_i})$,
such that:

$$u_i(\pi_i ; \pi_{-i}) \geq u_i(\pi_i' ; \pi_{-i}), \ \forall i, \forall \pi_i'$$

↳ even if allowed to be history-dependent

Proof: For policy profile $\pi = (\pi_1, \pi_2, \dots, \pi_m)$
player $i \in \{1, \dots, n\}$

we'll define $\left(v_i^{\pi}(s)\right)_{s \in S}$

and $\left(q_i^{\pi}(s, a_i)\right)_{s \in S, a_i \in A_i}$

as follows:

- $v_i^{\pi}(s)$: infinite discounted utility of player i if game started at state s & players used policies $\pi_1, \dots, \pi_m$

$$\hookrightarrow v_i^{\pi}(s) = \underbrace{\sum_a r_i(s,a)\cdot\pi(a|s)}_{\text{II}\triangleright\ r_i^{\pi}(s)} + \sum_{s'} v_i^{\pi}(s')\cdot\underset{\swarrow\text{discount factor}}{\gamma}\cdot\underbrace{\sum_a \pi(a|s)\cdot P(s'|s,a)}_{\text{II}\triangleright\ \Gamma^{\pi}(s,s')}$$

$$\Rightarrow (I-\gamma\cdot\Gamma^{\pi})\, v_i^{\pi} = r_i^{\pi} \overset{(*)}{\Rightarrow} v_i^{\pi} = (I-\gamma\cdot\Gamma^{\pi})^{-1}\cdot r_i^{\pi}$$

$\boxed{\begin{array}{l} (*) \text{ why is } I-\gamma\cdot\Gamma^{\pi} \text{ invertible?} \\ \text{Row sums of } \Gamma^{\pi}: \quad \sum_{s'}\Gamma^{\pi}(s,s') = \sum_a \pi(a|s) = 1 \\ \overset{\gamma<1}{\Rightarrow} I-\gamma\cdot\Gamma^{\pi} \text{ is strictly diagonally dominant} \\ \Rightarrow I-\gamma\cdot\Gamma^{\pi} \text{ is non-singular} \end{array}}$

by above note that, as a f'n of π, $v_i^{\pi}(s)$ is continuous

- $q_i^{\pi}(s,a_i)$: infinite discounted utility of player i if game started at s, players used policies $\pi_1, \dots, \pi_m$ throughout, except at the very first step player i plays a_i

$$q_i^\pi(s,a_i) = \sum_{a_{-i}} r_i(s,a)\cdot\pi_{-i}(a_{-i}|s) + \sum_{s'} v_i^\pi(s')\cdot\gamma\cdot\sum_{a_{-i}}\pi(a_{-i}|s)\cdot P(s'|s,a)$$

↳ as a function of π $q_i^\pi(s,a_i)$ is continuous b.c. everything on th RHS is cont. wrt. π including $v_i^\pi(s')$

- Now define Nash-type function mapping $\pi \overset{f}{\longmapsto} \pi'$ as follows:

$\forall i, \forall s, \forall a_i$:

$$\pi_i'(a_i|s) \longleftarrow \frac{\pi_i(a_i|s) + \overbrace{\max(0, q_i^\pi(s,a_i) - v_i^\pi(s))}^{\varphi_i^\pi(s,a_i)}}{1 + \sum_{a_i'} \max\left(0, q_i^\pi(s,a_i') - v_i^\pi(s)\right)}$$

f: continuous over convex, compact set $\times_i \big(\underbrace{\Delta(A_i)}_{\text{set of all possible stationary Markov policies of player } i}\big)^S$

$\underset{\text{Brouwer}}{\Longrightarrow}$ $\exists$ fixed point $\pi = f(\pi)$

Claim: π is Nash Equilibrium

Proof: suffices to prove π_i is best response to π_{-i} among all stationary Markovian policies

[why? b.c. fixing π_{-i}, player i faces Markov Decision process and it's known that MDPs have optimal policies that are stationary & Markovian]

Interesting to note: $v_i^{\pi}(s) = \sum_{a_i} \pi_i(a_i)\, q_i^{\pi}(s,a_i)$ (1)

↳ thus by doing the same logic we did in Nash's proof for each s separately it follows from $\pi = f(\pi)$ that:

$$\forall s, \forall i, \forall a_i : \ \varphi_i^{\pi}(s,a_i) = 0$$

$$\iff \forall s, \forall i, \forall a_i : \ v_i^{\pi}(s) \geq q_i^{\pi}(s,a_i)$$

$$\overset{(*)(*)}{\iff} \pi \text{ is a Nash Equilibrium.}$$

(*)(*) $\forall i, \forall s, \forall a_i : \ \underbrace{v_i^{\pi}(s)}_{\text{player i's expected infinite discounted payoff from } \pi_i} \geq \underbrace{q_i^{\pi}(s,a_i)}_{\text{player i's expected infinite discounted payoff at s from single-round deviation from } \pi_i}$ (2)

- clearly implied by π being a Nash Eq (so $\Leftarrow$ easy)
- why does it imply Nash eq?

fixing π_{-i}, player i faces an MDP where:

$\widetilde{\text{MDP}}$

$$\tilde{r}_i(s,a_i) = \sum_{a_{-i}} r(s,a) \cdot \pi_{-i}(a_{-i}|s)$$

$$\tilde{P}(s'|s,a_i) = \sum_{a_{-i}} P(s'|s,a)\, \pi_{-i}(a_{-i}|s)$$

given a policy π_i, denote by $\tilde{v}_i^{\pi_i}(s)$ the expected infinite discounted payoff starting at s & using π_i in $\widetilde{\text{MDP}}$

(2) + (1) $\Rightarrow \pi_i$ satisfies what are called "Bellman equations"

namely:

$$\forall s: \tilde{v}_i^{\pi_i}(s) = \max_{a_i} \left(\tilde{r}_i(s,a_i) + \gamma \cdot \sum_{s'} \tilde{P}(s'|s,a_i)\, \tilde{v}_i^{\pi_i}(s') \right)$$

Bellman Equations $\Rightarrow \pi_i$ is optimal in $\widetilde{MDP}$

so π_i is best response to π_{-i}

$\boxtimes$