Multiagent Reinforcement Learning in Stochastic Games

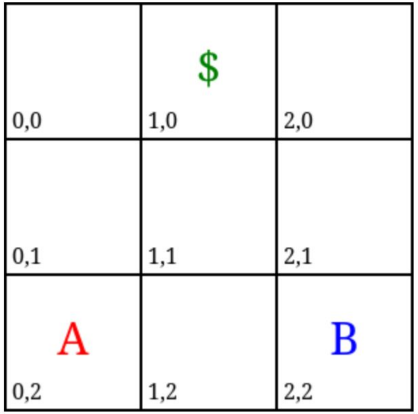
Lucas LUGAO GUIMARAES <@lucaslugao>
Victor VIANNA <@victorvianna>

Summary

- Motivation
- Background
  - Reinforcement Learning
  - Markov decision process
  - Policy and value function
  - Single-agent Q-learning
- Stochastic games
  - Definition
  - Nash equilibrium
  - Multi-agent Q-learning
  - Proof of convergence
- Demo!

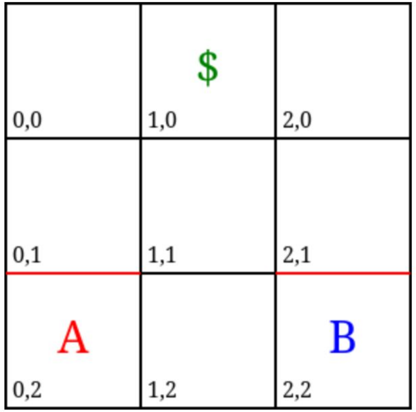
Motivation

- Goal: Get to $ cell ASAP
- A and B can't occupy the same cell (except for $)
- Game ends when someone reaches $



Motivation

- Red barrier with 50% probability of crossing

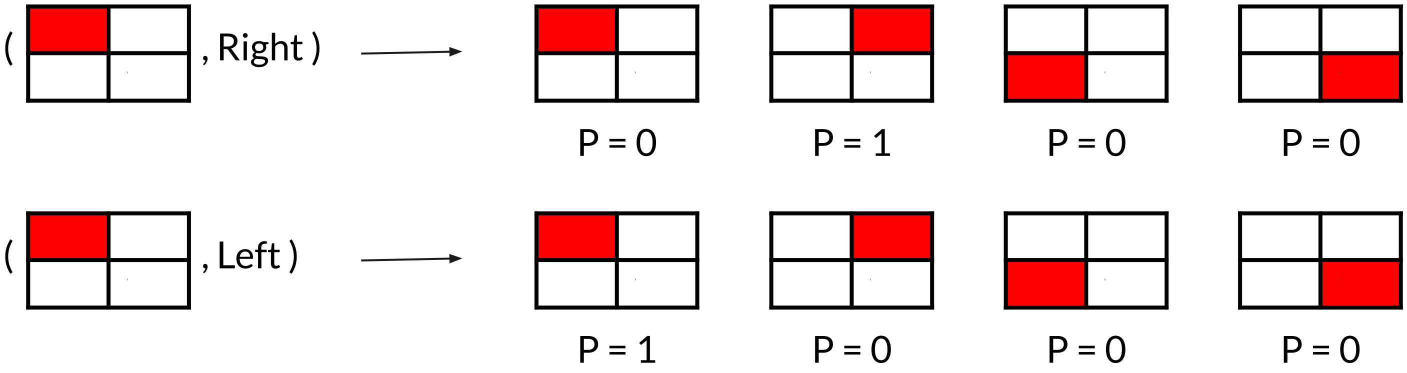


Background - Reinforcement Learning

- Agent(s) interacting with **environment** via **actions** (e.g. game)
- Only access to **very local** information
  - Observe current state
  - Perform specific action and observe reward
- May stop when some terminal state s_t is reached (e.g. win, lose)
- Aim to maximize “expected total reward”

Background - Markov decision process

- Space S (e.g. positions in the grid, pixel configurations in videogame)
- Actions A (e.g. {Up, Down, Left, Right})
- Stochastic transition s’(s, a) and reward r(s, s’, a)



Background - Policy and value function

- Policy Π, can be:
  - Stationary $\Pi = (\pi, \pi, \pi, \ldots)$
    - Deterministic π: S -> A
    - Stochastic π: S -> σ(A)
  - Non-stationary $\Pi = (\pi_0, \pi_1, \pi_2, \ldots)$
- Value function: Expected discounted sum of rewards ($0<\beta<1$)

$$v(s,\pi) = \sum_{t=0}^{\infty} \beta^t E(r_t | \pi, s_0 = s)$$

- Wish to find optimal policy π^*, i.e. $v(s_0, \pi^*) \geq v(s_0, \pi)$
- Theorem (Bellman): There is a stationary policy which is optimal for every state s

Background - Single-agent Q-learning

- Keep map Q(s, a) (arbitrary initialization)
- Starting at an arbitrary s, at each step t:
  - Choose random action a to perform
  - Observe reward r and new state s’
  - Update Q table for (s, a):

$$Q(s,a) \leftarrow (1-\alpha_t)\cdot \underbrace{Q(s,a)}_{\text{old value}} + \underbrace{\alpha_t}_{\text{learning rate}} \cdot \overbrace{\left(\underbrace{r}_{\text{reward}} + \underbrace{\beta}_{\text{discount factor}} \cdot \underbrace{\max_b Q(s',b)}_{\substack{\text{estimate of optimal future value}\\\text{following policy learned so far}}} \right)}^{\text{learned value}}$$

$t \leftarrow t + 1$
$s \leftarrow s'$

Background - Single-agent Q-learning

- Reset s if s_t reached
- Assumptions
  - Random actions s.t. all (s,a) visited ∞ times
  - $\sum \alpha_t = \infty$, $\sum \alpha_t^2 < \infty$ (locally in every (s,a))
- Optimal policy greedily recovered by :

$$\pi(s) = \arg\max_a Q(s,a)$$

Stochastic games - Definition

- S now encodes 2 players (e.g. S ={((0,0),(0,1)), ((0,0),(0,2)) , ...})
- Deterministic $r^1, r^2 : S \times A^1 \times A^2 \rightarrow \mathbb{R}$
  - Depend only on source state
  - Unknown in advance (“incomplete information game”)
- Bimatrix game at each fixed s
  - **UE19 course framework**
- Info available: observed rewards of everyone + new state
  - “perfect information game”

Stochastic games - Nash equilibrium

Definition 3 *In stochastic game* Γ, *a* Nash equilibrium point *is a pair of strategies* (π_*^1, π_*^2) *such that for all* $s \in S$

$$v^1(s, \pi_*^1, \pi_*^2) \geq v^1(s, \pi^1, \pi_*^2) \quad \forall \pi^1 \in \Pi^1$$

and

$$v^2(s, \pi_*^1, \pi_*^2) \geq v^2(s, \pi_*^1, \pi^2) \quad \forall \pi^2 \in \Pi^2$$

Stochastic games - Nash equilibrium

Theorem 2 *(Filar and Vrieze [4], Theorem 4.6.4) Every general-sum discounted stochastic game possesses at least one equilibrium point in stationary strategies.*

- Can we adapt Q-learning to find an equilibrium point in stationary strategies?
  - Yes! But we must give up on determinism.
  - (Intuition: There isn't always an eq. in simple strategies, but in mixed ones, yes.)

Stochastic games - Multi-agent Q-learning

- Before ...

$$Q(s,a) \leftarrow (1-\alpha_t)\cdot \underbrace{Q(s,a)}_{\text{old value}} + \underbrace{\alpha_t}_{\text{learning rate}} \cdot \overbrace{\left(\underbrace{r}_{\text{reward}} + \underbrace{\beta}_{\text{discount factor}} \cdot \underbrace{\boxed{\max_a Q(s',a)}}_{\substack{\text{estimate of optimal future value}\\ \text{following policy learned so far}}} \right)}^{\text{learned value}}$$

Stochastic games - Multi-agent Q-learning

- Now ...

$$Q^1_{t+1}(s,a^1,a^2) = (1-\alpha_t)Q^1_t(s,a^1,a^2) + \alpha_t[r^1_t + \beta\boxed{\pi^1(s')Q^1_t(s')\pi^2(s')}]$$
$$Q^2_{t+1}(s,a^1,a^2) = (1-\alpha_t)Q^2_t(s,a^1,a^2) + \alpha_t[r^2_t + \beta\boxed{\pi^1(s')Q^2_t(s')\pi^2(s')}]$$

Estimate of future value following equilibria learned so far

Stochastic games - Proof of convergence

- Assumptions:
  - All (s, a) visited ∞ times
  - Same condition on summability of α's
  - For all (s, a^1, a^2), ∃ local Nash equilibrium that **either**
    - Maximizes payoff for each player
    - Is a saddle point (deviating favours adversary)
- Three main ingredients for proof

Stochastic games - Proof of convergence

1. Local equilibrium => Global equilibrium

Theorem 3 *(Filar and Vrieze [4]) The following assertions are equivalent:*

1. *For each* $s \in S$, *the pair* $\boxed{(\pi^1(s), \pi^2(s))}$ *constitutes an equilibrium point in the static bimatrix game* $\left(Q^1(s), Q^2(s)\right)$ *with equilibrium payoffs* $\left(v^1(s, \pi^1, \pi^2), v^2(s, \pi^1, \pi^2)\right)$, *and for k=1,2 the entry* (a^1, a^2) *in* $Q^k(s)$ *equals*

$$Q^k(s, a^1, a^2) = r^k(s, a^1, a^2) + \beta \sum_{s'=1}^{N} p(s'|s, a^1, a^2) v^k(s', \pi^1, \pi^2).$$

2. $\boxed{(\pi^1, \pi^2)}$ *is an equilibrium point in the discounted stochastic game , with equilibrium payoff* $\left(\boldsymbol{v}^1(\pi^1, \pi^2), \boldsymbol{v}^2(\pi^1, \pi^2)\right)$, *where* $\boldsymbol{v}^k(\pi^1, \pi^2) = (v^k(s^1, \pi^1, \pi^2), \cdots, v^k(s^m, \pi^1, \pi^2))$, $k = 1, 2$.

Stochastic games - Proof of convergence

2. Understand averaged learning in terms of learned term

Lemma 1 *(Conditional Average Lemma) Under Assumptions 1-2, the process* $\boxed{Q_{t+1} = (1-\alpha_t)Q_t + \alpha_t w_t}$ *converges to* $E(w_t|h_t, \alpha_t)$*, where* h_t *is the history at time* t.

Lemma 2 *Under Assumptions 1-2, If the process defined by* $\boxed{U_{t+1}(x) = (1-\alpha_t(x))U_t(x) + \alpha_t(x)[P_t v^*](x)}$ *converges to* v^* *and* P_t *satisfies* $\| P_t V - P_t v^* \| \leq \gamma \| V - v^* \| + \lambda_t$ *for all* V*, where* $0 < \gamma < 1$ *and* $\lambda_t \geq 0$ *converges to 0, then the iteration defined by*

$$V_{t+1}(x) = (1-\alpha_t(x))V_t(x) + \alpha_t(x)[P_t V_t](x)$$

converges to v^*.

Stochastic games - Proof of convergence

3. Contraction mappings

Lemma 3 *Let* $P_t^k Q^k(s) = r_t^k + \beta \pi^1(s) Q^k(s) \pi^2(s)$, $k = 1, 2$, *where* $(\pi^1(s), \pi^2(s))$ *is a pair of mixed Nash equilibrium strategies for the bimatrix game* $(Q^1(s), Q^2(s))$. *Then* $P_t = (P_t^1, P_t^2)$ *is a contraction mapping.*

Demo!

"Never do a live demo."

Anyone

Thank you!

Lucas LUGAO GUIMARAES <@lucaslugao>
Victor VIANNA <@victorvianna>