MIT 6.S890 — Topics in Multiagent Learning

Lecture 16

Markov (aka stochastic) games

Instructor: Prof. Gabriele Farina

In this last lecture for Part III, we take a quick peek at one last model of games that is especially popular in (multi-agent) reinforcement learning: *Markov games*.

1 The model

The model of Markov games, also known as stochastic games, was introduced by Shapley, L. S. [Sha53] as a generalization of the Markov decision process to the multi-agent setting. In this model, the agents interact with each other and with the environment, and the environment is affected by the joint actions of the agents. In this lecture, we will draw a distinction between *infinite-horizon* games, and *finite-horizon* games (also known as *episodic*). We start from the former class.

Definition 1.1 (Infinite-horizon stochastic game). An m-player, infinite-horizon, finite state/action space, stochastic (aka Markov) game is a tuple $G = (S, A, \mathbb{P}, r, \gamma, \mu)$ where

- S is a finite set of states,
- $A = A_1 \times A_2 \times ... \times A_m$ is the set of joint actions,
- $\mathbb{P}(s'|s,a)$, for $s, s' \in S, a \in A$ is the transition matrix of the environment,
- $r = (r_1, ..., r_m)$ is a tuple of reward functions, wjere $r_{i(s,a)}$ specifies the reward of player i for taking action a in state s,
- $\gamma \in (0,1)$ is a discount factor, and
- $\mu \in \Delta(S)$ is the initial state distribution.

Given an infinite state-action sequence $\left(s^{(t)}, a^{(t)}\right)_t$, each player derives a *discounted utility*

$$u_i\Big(\left(s^{(t)}, a^{(t)}\right)_t\Big) \coloneqq \sum_{t \geq 0} \gamma^t \cdot r_i\big(s^{(t)}, a^{(t)}\big).$$

As the name suggests, an infinite-horizon stochastic game is played over an infinite number of steps. The goal of each player is to maximize their discounted utility. By fixing the number of steps, we instead obtain a finite-horizon stochastic game, as defined next.

Definition 1.2 (Finite-horizon stochastic game). In the finite-horizon case, the game is endowed with an addition parameter $H \in \mathbb{N}$, called the *horizon*, which indicates for how many steps the game will unfold. For these games, a value of $\gamma = 1$ (*i.e.*, no discounting) is acceptable since the utility is a finite sum and therefore it cannot diverge.

⋆These notes are class material that has not undergone formal peer review. The TA and I are grateful for any reports of typos.

2 Strategies: Stationarity vs Markovianity

In general, when we talk about a *strategy*, or *policy*, we allow for the possibility that the players remember their past actions and transitions. In other words, when left unqualified, the term *policy* allows for history-dependence and we think of it as a mapping

$$\pi_i : S \times (S \times A)^* \to \Delta(A_i),$$

where the asterisk denotes a tuple of arbitrary length representing the history of play up to any point.

When further restrictions are imposed on how the policy can depend on the history, we arrive at two important distinctions.

Definition 2.1 (Markovian policy). A policy is *history-independent*, or *Markovian*, if it only depends on the current state and time. This means that given any two histories of the same length, the policy is the same. In other words, the policy is a function

$$\pi_i : S \times \mathbb{N} \to \Delta(A_i).$$

Definition 2.2 (Stationary policy). A policy is *stationary* if it only depends on the current state and not even time. In other words, the policy is just a function of the current state

$$\pi_i : S \to \Delta(A_i).$$

(Note that stationarity implies Markovianity).

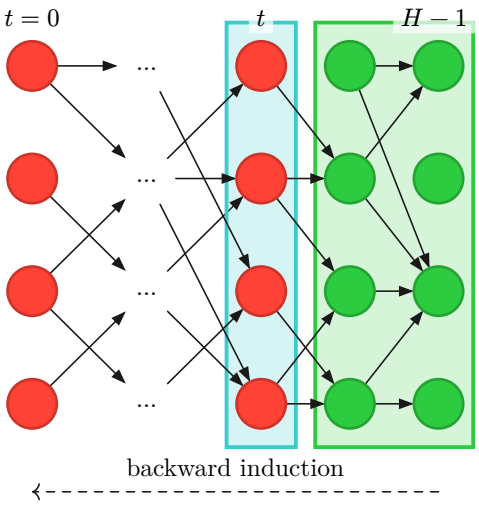
3 Equilibria in Markov games

3.1 The finite-horizon case

If we are seeking non-Markovian strategies, a finite-horizon stochastic game can just be "unrolled" and converted into a perfect-recall extensive-form game.

When Markovian strategies are sought, then the process is not as straightforward. However, the game can still be solved efficiently via *backward induction*. We illustrate this with an example.

The idea is to solve right-to-left the game, but inductively picking strategies that maximize the immediate reward plus the continuation value $V_{i,t}(s)$ at every state s.



- **Initialization:**
  - Set $V_{i,H} := 0$ for all states $s \in S$ and players i
- **Inductive step** (at time t)
  - Assume given $V_{i,t+1} : S \to \mathbb{R}$.
  - For each $s \in S$, construct a game where i's utility is defined as

$$F_{i,s}(a) := r_i(s,a) + \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)}\left[V_{i,t+1}(s')\right].$$

  - Compute a Nash equilibrium of the *normal-form game* given by the utilities F, and let that be

$$\pi(\cdot\,|s,t) \in \Delta(A).$$

  - Let $V_{i,t}(s) := \mathbb{E}_{a \sim \pi(\cdot|s,t)}\left[F_{i,s}(a)\right]$.

Finally, we remark that in finite-horizon games, stationary policies are typically not optimal, because optimal behavior usually depends on the amount of time left. In infinite-horizon games, however, equilibria in stationary policies exist, as we show next.

3.2 The infinite-horizon case

Theorem 3.1 ([Fin64; Tak64]). Every infinite-horizon discounted, stochastic game with a finite number of states, actions, and players, has a Nash equilibrium in stationary, Markovian strategies. More formally, there exists a collection of policies $\pi_1, ..., \pi_m$ where $\pi_i : S \to \Delta(A_i)$ such that

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

where π_i' is *any* policy for player i, *not necessarily Markovian.*

Proof. Given a policy profile $\pi = (\pi_1, .., ., \pi_m)$ and a player $i \in [m]$, we will introduce the following notation:

- $\boxed{v_i^\pi(s)}$, for $s \in S$, is the infinite discounted utility of player i if the game started at state s and all players used policies $\pi_1, ..., \pi_m$. In symbols,

$$v_i^\pi(s) = \underbrace{\sum_a r_i(s,a) \cdot \pi(a|s)}_{=: r_i^\pi(s)} + \gamma \sum_{s'} v_i^\pi(s') \underbrace{\sum_a \pi(a|s) \cdot \mathbb{P}(s'|s,a)}_{=: \Gamma^\pi(s,s')}.$$

  This is a linear system of equations in the variables $v_i^\pi(s)$, which we can rewrite more compactly as

$$(I - \gamma\Gamma^\pi) v_i^\pi = r_i^\pi.$$

  We now argue that the matrix $I - \gamma\Gamma^\pi$ is invertible. To see this, note that the matrix Γ^π is a stochastic matrix:

$$\sum_{s'} \Gamma^\pi(s,s') = \sum_{s'} \sum_a \pi(a|s)\mathbb{P}(s'|s,a) = \sum_a \pi(a|s) \left(\sum_{s'} \mathbb{P}(s'|s,a) \right) = \sum_a \pi(a|s) = 1.$$

  Since $\gamma < 1$ by Definition 1.1, we have that $I - \gamma\Gamma^\pi$ is *strictly* diagonally dominant, which implies that $I - \gamma\Gamma^\pi$ cannot be singular. Therefore, the system of equations has a unique solution, which corresponds to

$$v_i^\pi = (I - \gamma\Gamma^\pi)^{-1} r_i^\pi.$$

  Furthermore, the values v_i^π are *continuous* in the policies π.

- $\boxed{q_i^\pi(s, a_i)}$, for $s \in S$ and $a_i \in A_i$, is the infinite discounted utility of player i if the game started at state s, and players used policies $\pi_1, ..., \pi_m$, with the only exception that the very first action of player i is set to a_i. In symbols,

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

  Like before, the function q_i^π is continuous in the policies π, since everything on the right-hand side is continuous, including the v_i^π as discussed above. Furthermore,

$$v_i^\pi(s) = \sum_{a_i} \pi_i(a_i|s) \cdot q_i^\pi(s, a_i). \tag{1}$$

We now define a Nash-type function φ, similar to what we used in Lecture 2, mapping policy profiles to improved policy profiles as follows:

$$\forall i, s, a_i : \qquad \pi_i'(a_i|s) \leftarrow \frac{\pi_{i(a_i|s)} + [q_i^\pi(s,a_i) - v_i^\pi(s)]^+}{1 + \sum_{a_i'} [q_i^\pi(s,a_i') - v_i^\pi(s)]^+}.$$

This mapping is continuous over the convex compact set of all stationary Markov policy profiles. Hence, by Brouwer's fixed-point theorem, there exists a fixed point $\pi^* = \varphi(\pi^*)$.

To complete the proof, we then only have to argue that the fixed point π^* is a Nash equilibrium. Using the same logic as Lecture 2, we infer that

$$\forall i \in [m], s \in S, \text{and } a_i \in A_i, \qquad v_i^{\pi^*}(s) \geq q_i^{\pi^*}(s, a_i). \tag{2}$$

We need to show that, fixing π_{-i}^*, π_i^* is a best response for each player i. From the point of view of player i, computing a best response amounts to solving a Markov decision process (MDP) supported on S with rewards and transitions given by

$$\tilde{r}_i(s, a_i) \coloneqq \sum_{a_{-i}} r(s,a) \cdot \pi_{-i}(a_{-i}|s),$$
$$\tilde{\mathbb{P}}(s'|s, a_i) \coloneqq \sum_{a_{-i}} P(s'|s,a) \cdot \pi_{-i}(a_{-i}|s).$$

Equations (1) and (2) together imply that the expected discounted payoff $\tilde{v}_i^{\pi_i^*}(s)$ starting at s in MDP satisfies

$$\forall s \in S, \qquad \tilde{v}_i^{\pi_i^*}(s) = \max_{a_i} \left\{ \tilde{r}_i(s, a_i) + \gamma \sum_{s'} \tilde{v}_i^{\pi_i^*}(s') \tilde{\mathbb{P}}(s'|s, a_i) \right\}.$$

The previous condition is the Bellman equation for the MDP. From the theory of MDPs, we conclude that π_i^* is an optimal policy for the MDP, and therefore π_i^* is a best response to π_{-i}^*. □

From a computational point of view, finding equilibria in Markov games is a challenging task. In general, the problem is open already in the two-player zero-sum case. However, this becomes tractable if the discount factor γ is bounded away from 1, and the goal is approximate equilibria. The computation of correlated and coarse correlated equilibria is also mostly open, with some hardness results depending on the type of strategies under consideration.

3.3 Shapley's theorem for two-player zero-sum Markov games

Finally, we mention a result by Shapley, L. S. [Sha53] that characterizes the value of two-player zero-sum Markov games. The result is a generalization of the minimax theorem for two-player zero-sum games.

Theorem 3.2 (Shapley's minimax theorem). The value of a two-player zero-sum Markov game is given by

$$v = \max_{\pi_1} \min_{\pi_2} u_1(\pi_1, \pi_2) = \min_{\pi_2} \max_{\pi_1} u_1(\pi_1, \pi_2),$$

where the max and min can be taken over all policies, Markov policies, and for infinite-horizon games, even stationary Markov policies.

The proof of the theorem is based on showing contraction of a certain Bellman iterator.

Bibliography

[Sha53] L. S. Shapley, "Stochastic games," *Proceedings of the national academy of sciences*, vol. 39, no. 10, pp. 1095–1100, 1953.

[Tak64] M. Takahashi, "Equilibrium points of stochastic non-cooperative n-person games", *Journal of Science of the Hiroshima University, Series AI (Mathematics)*, vol. 28, no. 1, pp. 95–99, 1964.

[Fin64] A. M. Fink, "Equilibrium in a stochastic n-person game", *Journal of science of the hiroshima university, series ai (mathematics)*, vol. 28, no. 1, pp. 89–93, 1964.

6.S890 Topics in Multiagent Learning

Fall 2024