MIT 6.7980 — Topics in Multiagent Learning Fall 2026

Supplementary reading S5

Markov (aka stochastic) games

Instructor: Prof. Constantinos Daskalakis (costis@mit.edu)

In this lecture, we turn our attention to *Markov games*, also known as *stochastic games*. These are an expressive family of games which has become especially popular recently as a mathematical model underlying multi-agent reinforcement learning. Markov games capture strategic interactions that take place over a number of rounds or perhaps an infinite number of rounds, in some environment whose state is influenced by the actions taken by players, and which in turn influences the players' rewards.

S5.1 The model

The model of Markov games was introduced in the seminal work of Shapley, L. S. [Sha53] as a generalization of Markov decision process from the single-agent 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 with the former. While the definition is a mouthful, the model is very natural in its examination.

Definition S5.1 (Infinite-horizon stochastic game)**.** An m-player, infinite-horizon, finite state and action *stochastic game*, also called *Markov game*, is a tuple $G = (S, A, \mathbb{P}, r, \gamma, \mu)$ where

- S is a finite set of states that the environment can be in;
- $A = A_1 \times A_2 \times ... \times A_m$ is the set of action profiles, where A_i are the actions available to player i;
- $\mathbb{P}(s' \mid s, a)$, for $s, s' \in S$ and $a \in A$ are the transition probabilities of the environment; in particular, $\mathbb{P}(s' \mid s, a)$ is the probability that the state of the environment becomes s' if action profile $a \in A$ is taken by the players in some state s;
- $r = (r_1, ..., r_m)$ is a tuple of reward functions, where $r_i(s, a)$ specifies the immediate reward received by player i when the action profile $a \in A$ is taken by the players in some state s;
- $\gamma \in [0, 1)$ is the discount factor; and
- $\mu \in \Delta(S)$ is the initialization distribution, sampling the state $s^{(0)}$ of the environment at the beginning of the interaction.

Given an infinite state-action sequence $\left(s^{(t)}, a^{(t)}\right)_{t=0}^{\infty}$, each player derives a *discounted utility* of

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

We will use the convention that $\gamma^0 = 1$, when $\gamma = 0$.

As the name suggests, an infinite-horizon stochastic game is played over an infinite number of steps, and there is discounting of future rewards. The goal of each player is to maximize their discounted utility. If the interaction takes place over a finite number of steps, we have a finite-horizon stochastic game, as defined next.

Definition S5.2 (Finite-Horizon Stochastic Game)**.** A *finite-horizon stochastic game* is defined in terms of the same primitives used to defined an infinite-horizon stochastic game, together with an additional parameter $H \in \mathbb{N}$, called the *horizon*, which indicates that the interaction takes place over H steps, indexed $t = 0, ..., H - 1$. Because of the finiteness of the number of steps, the discount factor can now take any value in $[0, 1]$, i.e. the value of $\gamma = 1$ is acceptable since the discounted utility is a finite sum and therefore cannot diverge. If $\gamma = 1$, we say that there is *no discounting* of future rewards.

S5.2 Strategies and Nash Equilibrium

In general, when we think about *strategies*, a.k.a. *policies*, for players in a stochastic game, we may allow for the possibility that the actions taken at some state s at some time t may depend on the entire trajectory $(s^{(\tau)}, a^{(\tau)})_{\tau < t}$ so far. 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 S5.3 (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 particular, the policy is a function

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

Definition S5.4 (Stationary and Markovian policy)**.** A policy is *stationary and Markovian* if it only depends on the current state. In particular, the policy is just a function of the current state

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

Given a collection of policies $\pi_1, ..., \pi_m$ for the players of a stochastic game, the expected utility of each player is naturally defined as follows:

$$u_i(\pi_1, ..., \pi_m) = \mathop{\mathbb{E}}_{\substack{s_0 \sim \mu; \\ \forall t > 0 : s^{(t)} \sim \mathbb{P}(\cdot \mid s^{(t-1)}, a^{(t-1)}) \\ \forall t \geq 0, i : a_i^{(t)} \sim \pi_i(s^{(t)}, (s^{(\tau)}, a^{(\tau)})_{\tau < t})}} \left[\sum_{t \geq 0} \gamma^t r_i(s^{(t)}, a^{(t)}) \right].$$

In particular, $u_i(\pi_1, ..., \pi_m)$ is the expected discounted utility of player i under the random trajectory which starts at $s^{(0)} \sim \mu$ and is sampled by having each player sampling an action from their policy at each state, and having the environment transition according to its dynamics. In terms of these utilities, Nash equilibrium is defined in the natural way as follows. Notice that this definition generalizes the concept of Nash equilibrium in normal-form games.

Definition S5.5 (Nash equilibrium)**.** A collection of policies $\pi = (\pi_1, ..., \pi_m)$ is a *Nash equilibrium* of a stochastic game iff for all players i, for all policies $\pi' : S \times (S \times A)^* \to \Delta(A_i)$ it holds that

$$u_i(\pi_i ; \pi_{-i}) \geq u_i(\pi'_i ; \pi_{-i}).$$

If all strategies $\pi_1, ..., \pi_m$ are Markovian, the Nash equilibrium is called *Nash equilibrium in Markovian strategies.* Similarly, if all strategies $\pi_1, ..., \pi_m$ are stationary and Markovian, the Nash equilibrium is called *Nash equilibrium in stationary and Markovian strategies.*

S5.3 Nash Equilibrium Existence

S5.3.1 The finite-horizon case

If we are content with non-Markovian strategies, a finite-horizon stochastic game can just be "unrolled" and converted into a perfect-recall extensive-form game, whose Nash equilibrium strategies can be converted to a Nash equilibrium of the stochastic game. In general, this Nash equilibrium will not be in Markovian strategies. However, finite-horizon stochastic games do have Nash equilibria in Markovian strategies, as can be seen by a *backward induction* argument.

Theorem S5.6. Every finite-horizon stochastic game with a finite number of states, actions, and players, has a Nash equilibrium in Markovian strategies. More formally, in the setting of Definition S5.2, there exists a collection of policies $\pi_1, ..., \pi_m$ where $\pi_i : S \times \{0, ..., H-1\} \to \Delta(A_i)$ 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, not necessarily Markovian,* policy for player i.

Proof Sketch. The idea is to solve the game backwards, starting at the end of the horizon and proceeding backwards, down to the first step of the interaction, inductively picking Nash equilibrium strategies for hypothetical games that would start at all possible interaction steps t and all possible states s. To compute these Nash equilibrium strategies inductively, we need, for all t and all s, to find Nash equilibrium strategies for a game whose payoff, $U_{i,t,s}(a)$, for each player i, is the immediate reward $r_i(s, a)$ plus the continuation value expected for this player under the inductively computed strategies and the transitions of the environment.

Below is a Nash equilibrium computation algorithm, whose correctness establishes the existence of Nash equilibrium in Markov policies.

1. Initialization ($t = H$):
    1.1. $V_{i,H}(s) = 0$, for all i, s; in particular, the expected continuation value for each player i at each state s at time $t = H$ is 0, as the game has ended at $t = H$.
2. Inductive Step (from $t = H - 1$ down to $t = 0$):
    2.1. Assume already computed expected continuation values $V_{i,t+1} : S \to \mathbb{R}$ for each player i.
    2.2. For each state s:
        2.2.1. define a game wherein player i's utility is
        $$U_{i,t,s}(a) = r_i(s, a) + \gamma \mathop{\mathbb{E}}_{s' \sim \mathbb{P}(\cdot \mid s,a)}[V_{i,t+1}(s')];$$
        2.2.2. pick an arbitrary Nash equilibrium $\pi(\cdot \mid s, t) \in \Delta(A)$ of the normal-form game with the above utility functions;
        2.2.3. set $V_{i,t}(s) = \mathbb{E}_{a \sim \pi(\cdot \mid s,t)}[U_{i,t,s}(a)]$.

To argue the correctness of the above algorithm, one proceeds as follows. Suppose that $\pi(\cdot \mid s, t) = (\pi_1(\cdot \mid s, t), ..., \pi_m(\cdot \mid s, t))$ are the Nash equilibrium strategies picked in Step 2.2.2 of the algorithm for all s, t. For each player i, define a Markovian policy $\pi_i : S \times \{0, ..., H-1\} \to \Delta(A_i)$ as follows:

$$\pi_i(s,t)(\cdot) := \pi_i(\cdot \mid s,t).$$

We claim that the collection of policies $(\pi_1, ..., \pi_m)$ is a Nash equilibrium of the game in Markovian policies. The Markovianity of the policies is clear from the definition of this policies in a backwards induction manner.

So all we need to prove is that π_i is a best-response to π_{-i}. This can be shown inductively. The base case is arguing that, for each s, the distribution $\pi_i(s, H-1)$ is optimal for player i to use, if he finds himself at state s at time $H-1$, given the policies of the other players. This follows immediately by the Nash equilibrium conditions satisfied by $\pi(\cdot \mid s, H-1)$. The inductive hypothesis is that policy

$$\pi_i^{\geq t+1} := (\pi_i(s,\tau)(\cdot))_{s\in S, \tau \geq t+1},$$

is optimal for player i to continue the game with against the policies of the other players if the player finds himself at some state s at time $t+1$. The induction step is showing that under the induction hypothesis, $\pi_i^{\geq t}$ is optimal for continuing the game with against the policies of the other players if the player finds himself at some state s at time t. To show the inductive step we will use the definition of the game in Step 2.2.1 of the algorithm and the Nash equilibrium properties of the strategies picked in Step 2.2.2. We leave the complete details to the reader. □

Finally, we remark that in finite-horizon games, there typically do not exist Nash equilibria in stationary and Markovian policies, because the best response of a player to the policies of the other players, even if all other players use stationary and Markovian policies, typically depends on the number of interaction steps that remain. In infinite-horizon games, however, equilibria in stationary and Markovian policies do exist, as we show in the next section.

S5.3.2 The infinite-horizon case

Theorem S5.7 ([Fin64; Tak64])**.** Every infinite-horizon stochastic game with a finite number of states, actions, and players, has a Nash equilibrium in stationary, Markovian strategies. In particular, in the setting of Definition S5.1, 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, not necessarily stationary and Markovian,* policy for player i.

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

- $v_i^{\pi}(s)$, 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,

$$\forall s : v_i^\pi(s) = \underbrace{\sum_a r_i(s,a)\pi(a \mid s)}_{=:r_i^\pi(s)} + \gamma \sum_{s'} v_i^\pi(s') \underbrace{\sum_a \pi(a \mid s)\mathbb{P}(s' \mid s,a)}_{=:\Gamma^\pi(s,s')} \tag{1}$$

Notice that 1 is a linear system of equations in the variables $(v_i^\pi(s))_s$, 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 row-stochastic matrix:

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

Since $\gamma < 1$ by Definition S5.1, 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, because the inverse matrix of the non-singular matrix $I - \gamma\Gamma^\pi$ is continuous in the values of the entries, and these are continuous in π.

- $q_i^\pi(s, a_i)$, 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,

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

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,

$$v_i^\pi(s) = \sum_{a_i} \pi_i(a_i \mid s) \cdot q_i^\pi(s, a_i). \tag{2}$$

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:

$$\forall i, s, a_i : \qquad \pi_i'(a_i \mid s) \leftarrow \frac{\pi_i(a_i \mid s) + [q_i^\pi(s,a_i) - v_i^\pi(s)]^+}{1 + \sum_{a_i'} [q_i^\pi(s,a_i') - v_i^\pi(s)]^+}. \tag{3}$$

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 need to argue that the fixed point π^* is a Nash equilibrium, that is, for all i, π_i^* is a best response to π_{-i}^*, even if the best response is computed with respect to arbitrary policies $\pi_i' : S \times (S \times A)^* \to \Delta(A_i)$.

Pick an arbitrary player i and state s. Using the same logic in the Nash equilibrium existence proof in Lecture 2, we infer that

$$\forall a_i \in A_i, \qquad v_i^{\pi^*}(s) \geq q_i^{\pi^*}(s, a_i). \tag{4}$$

Indeed, all that is needed to repeat the argument from Nash's proof is the linearity $v_i^{\pi}(s) = \sum_{a_i} \pi_i(a_i \mid s) \cdot q_i^{\pi}(s, a_i)$ and form of 3.

With this in hand, let us now show that π_i^* is a best response to π_{-i}^* for player i. From the point of view of player i, computing a best response to π_{-i}^* amounts to solving a Markov decision process (MDP) with states S and actions A_i and rewards, transitions given by

$$\tilde{r}(s, a_i) \coloneqq \sum_{a_{-i}} r_i(s, a) \cdot \pi_{-i}^*(a_{-i} \mid s),$$
$$\tilde{\mathbb{P}}(s' \mid s, a_i) \coloneqq \sum_{a_{-i}} \mathbb{P}(s' \mid s, a) \cdot \pi_{-i}^*(a_{-i} \mid s).$$

Notice that the V-value function induced by policy π_i^* in this MDP, denoted $\tilde{V}^{\pi_i^*}(s)$, coincides with $v_i^{\pi^*}(s)$, as defined above in the stochastic game, i.e.

$$\tilde{V}^{\pi_i^*}(s) \equiv v_i^{\pi^*}(s), \forall s.$$

Moreover, 2 and 4 together imply that

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

The previous condition is the Bellman equation for the MDP. From the theory of MDPs, we conclude that π_i^* is an optimal policy in the MDP, even among history-dependent policies. Therefore π_i^* is a best response to π_{-i}^*, even among history-dependent policies.□

S5.3.3 Shapley's theorem for two-player zero-sum Markov games

In this section, we turn our attention to *two-player zero-sum* stochastic games. These are games where the interests of the two players are strictly opposed, i.e., $r_1(s, a) = -r_2(s, a)$ for all states s and action profiles a. We typically denote $r(s, a) \coloneqq r_1(s, a)$ as the reward for Player 1 (the maximizer) and the cost for Player 2 (the minimizer).

For the remainder of this section, we will focus specifically on the more interesting case of *infinite-horizon* games. As we discussed in the previous section, finite-horizon games generally do not admit Nash equilibria in stationary Markovian strategies (strategies that depend only on the state, not the time step). In contrast, infinite-horizon games do admit stationary equilibria.

Shapley, L. S. [Sha53] established a sharp existence result for this setting, showing that these games have a unique *value*. While we have already established that Nash equilibria exist in general-sum Markov games (using Brouwer's fixed-point theorem), the zero-sum setting admits a "simpler" reason for existence. This simplicity translates into better computational guarantees. The existence of equilibrium here is guaranteed not merely by the existence of a

fixed point of a continuous map (as in Brouwer), but specifically by the existence of a fixed point of a contraction map.

S5.3.3.1 Contraction Mappings and Banach's Theorem

In the general-sum case, Brouwer's theorem guarantees a fixed point exists but provides no recipe for finding it. In contrast, a contraction mapping guarantees that simply *iterating* the function will converge to the unique fixed point.

Definition S5.8 (Contraction Mapping)**.** Let (X, d) be a complete metric space. A function $T : X \to X$ is called a λ-*contraction* if there exists a constant $\lambda \in [0, 1)$ such that for all $u, v \in X$:

$$d(T(u), T(v)) \leq \lambda \cdot d(u, v).$$

The mathematical engine behind Shapley's result is the following fundamental theorem from analysis.

Theorem S5.9 (Banach Fixed-Point Theorem)**.** Let (X, d) be a non-empty complete metric space and $T : X \to X$ be a λ-contraction. Then:

1. T admits a *unique* fixed point $x^* \in X$ (i.e., $T(x^*) = x^*$).
2. For any initial guess $x^{(0)} \in X$, the sequence defined by $x^{(t+1)} = T(x^{(t)})$ converges to x^*.

Proof Sketch. The reason contraction maps have fixed points is intuitive: applying the map strictly shrinks the distance between points. Consider the distance between two successive iterates:

$$d(x^{(t+1)}, x^{(t)}) = d(T(x^{(t)}), T(x^{(t-1)})) \leq \lambda \cdot d(x^{(t)}, x^{(t-1)}).$$

By induction, the steps become exponentially smaller: $d(x^{(t+1)}, x^{(t)}) \leq \lambda^t d(x^{(1)}, x^{(0)})$. The sequence is therefore Cauchy and must converge to a unique limit. □

S5.3.3.2 Shapley's Operator

Shapley used this machinery to prove that zero-sum stochastic games have a value. He constructed a contraction mapping over the space of *value functions* (not strategies). Let $V \in \mathbb{R}^{|S|}$ be a vector representing the value of the game to Player 1 at each state. We use the infinity norm $\|V\|_\infty = \max_s |V(s)|$.

We define the *Shapley Operator* (or Bellman Operator) $\mathcal{T} : \mathbb{R}^{|S|} \to \mathbb{R}^{|S|}$ as follows. For a given estimate of future values V, we construct a "local" matrix game at each state s where the payoff for joint action a is the immediate reward plus the discounted future value:

$$Q_{s,V}(a) := r(s, a) + \gamma \sum_{s'} \mathbb{P}(s' \mid s, a) V(s').$$

The operator updates the value of state s to be the minimax value of this local game:

$$(\mathcal{T}V)(s) := \max_{\pi_1 \in \Delta(A_1)} \min_{\pi_2 \in \Delta(A_2)} \mathop{\mathbb{E}}_{a \sim (\pi_1, \pi_2)} [Q_{s,V}(a)].$$

Theorem S5.10 (Shapley's Minimax Theorem)**.** The operator $\mathcal{T}$ is a contraction mapping with modulus γ. That is, $\|\mathcal{T}U - \mathcal{T}V\|_\infty \leq \gamma \|U - V\|_\infty$. Consequently, there exists a unique value vector V^* such that $\mathcal{T}V^* = V^*$. This V^* is the value of the stochastic game.

Proof Sketch. The proof relies on the fact that the value of a zero-sum matrix game is *non-expansive* with respect to its payoffs (if payoffs change by δ, the value changes by at most δ). Here, if future values U and V differ by ϵ, the payoffs in the local matrix games differ by at most $\gamma\epsilon$. Thus, the values of these local games differ by at most $\gamma\epsilon$. □

S5.3.3.3 Computation: Value Iteration

The constructive nature of the Banach Fixed-Point Theorem yields a natural algorithm for computing the Nash equilibrium, known as *Value Iteration*. Conceptually, this algorithm is "decentralized." We do not need to solve one giant optimization problem for the whole game. Instead, in each round, we update the value of every state by solving a local normal-form game based on the values from the previous round.

1. **Initialization:** Set $V^{(0)}(s) = 0$ for all $s \in S$.
2. **Iterative Step:** For $t = 0, 1, 2, ...$:
    2.1. For each state $s \in S$, construct the matrix game M_s with payoffs:
    $$A_{ij} = r(s, a_i, a_j) + \gamma \sum_{s'} \mathbb{P}(s' \mid s, a_i, a_j) V^{(t)}(s').$$
    2.2. Compute the value v_s of the matrix game M_s (this can be done efficiently using Linear Programming).
    2.3. Update the estimate: $V^{(t+1)}(s) \leftarrow v_s$.
3. **Termination:** Stop when $\left\|V^{(t+1)} - V^{(t)}\right\|_\infty \leq \epsilon \frac{1-\gamma}{2\gamma}$.

Once the value function V^* (or a sufficient approximation) is computed, the stationary Nash equilibrium strategy for each player at state s is simply the optimal (minimax) strategy in the local matrix game defined by payoffs Q_{s,V^*}.

Note that the convergence rate of the algorithm above depends heavily on the contraction modulus γ. Specifically, the number of iterations required to reach precision ϵ scales with $\frac{1}{1-\gamma}$. This raises a difficulty if the discount factor is extremely close to 1 (the "high-precision" regime), or if we are investigating the *undiscounted* setting where $\gamma \to 1$. In such cases, the term $\frac{1}{1-\gamma}$ blows up, potentially rendering the basic algorithm inefficient.

However, if our goal is merely to find an ϵ-*approximate* Nash equilibrium, we can circumvent this issue by *artificially contracting* the map. Even if the true game has $\gamma \approx 1$, we can substitute it with a surrogate game having a discount factor $\tilde{\gamma} = 1 - \epsilon$. It can be shown that the value of this surrogate game differs from the true value by at most $O(\epsilon)$. Crucially, running Value Iteration on this surrogate game requires a number of iterations proportional

to $\frac{1}{1-\tilde{\gamma}} = \frac{1}{\epsilon}$. This standard trick effectively converts the dependence on the discount factor into a polynomial dependence on $1/\epsilon$.

S5.3.3.4 The Complexity Landscape

We can now place the difficulty of solving two-player zero-sum Markov games in a rigorous context. By analyzing how the computational cost scales with the precision ϵ, we observe a striking hierarchy that mirrors the mathematical tools used to prove existence:

- **Two-player Zero-sum Normal-form Games ($\log(1/\epsilon)$):** These are the "easiest" case. Existence is guaranteed by **Linear Programming Duality**. Consequently, we can use linear programming algorithms (like Interior Point methods) to find ϵ-approximate equilibria in time proportional to $\log(1/\epsilon)$.
- **Two-player Zero-sum Markov Games ($\text{poly}\ (1/\epsilon)$):** This setting occupies the *middle ground*. Existence is guaranteed by **Banach's Fixed Point Theorem** (Contraction Mappings). As discussed above, even in the worst case where $\gamma \to 1$, we can use the artificial contraction trick to find ϵ-approximate equilibria in time proportional to $\text{poly}\ (1/\epsilon)$.
- **General-sum Normal-form Games ($\exp(1/\epsilon)$):** These are the "hardest" case. Existence is guaranteed only by **Brouwer's Fixed Point Theorem** (Topology), which provides no constructive recipe. Computing a Nash equilibrium is PPAD-complete. The best known algorithms run in time $(n^{O(\log n/\epsilon^2)})$, quasi-polynomial in the number of actions and exponential in $1/\epsilon$.

This spectrum ($\log(1/\epsilon)$ vs. $\text{poly}\ (1/\epsilon)$ vs. $\exp(1/\epsilon)$) highlights the unique position of Markov games. While the $\text{poly}\ (1/\epsilon)$ result ensures tractability, a major open problem is whether two-player zero-sum Markov games can be pushed into the "logarithmic" category. Specifically, can they be solved in *strongly polynomial time* (time independent of the transitions and discount factor)? This is known as the **Simplest Stochastic Games Conjecture** and it is one of the major open questions in algorithmic game theory.

S5.4 Bibliography for this lecture

[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.