# Lecture 12: Introduction to Stochastic Games and Multi-Agent RL

Lecturer: Anas Barakat

October 23, 2025

**Abstract**

We introduce finite discounted infinite-time horizon Markov games as a mathematical framework for multi-agent RL and show existence of stationary Nash equilibrium policies. Then, we characterize Nash policies as first-order stationary points of the pseudo-gradient vector field. We subsequently introduce two fundamental classes of Markov games in which Nash equilibrium computation is tractable: zero-sum Markov games for which we prove the minmax theorem and Markov potential games for which any stationary point of the potential function is a Nash policy. We discuss independent learning in Markov games using policy gradient methods. Finally, we briefly comment on the regret analysis framework for Markov games.

*Disclaimer: These lecture notes are preliminary notes that will evolve and will be enriched over time, this is the first iteration of the course. The treatment is not comprehensive but focuses on some of the main ideas with accompanying proofs. Research on the topic is also still very active.*

## 1 Markov Games: A Theoretical Framework for Multi-Agent RL

In multi-agent reinforcement learning (RL), several agents interact within a shared dynamic and uncertain environment evolving over time depending on the individual strategic decisions of all the agents. Each agent aims to maximize their own individual reward which may however depend on all players' decisions. Markov games (a.k.a. stochastic games) offer a formal mathematical framework for modeling multi-agent reinforcement learning problems.

### 1.1 Definition of Markov Games

In Markov games, multiple agents interact with each other in a dynamically changing environment. The formal definition is as follows.

**Definition 1** (Markov Game, Shapley (1953))**.** *A discounted infinite time horizon Markov Game*[1] *is defined by a tuple:*

$$\Gamma := \langle \mathcal{N}, \mathcal{S}, (\mathcal{A}_i)_{i\in\mathcal{N}}, P, (r_i)_{i\in\mathcal{N}}, \rho, \gamma \rangle \tag{1}$$

*where*

- $\mathcal{N} \triangleq \{1, \cdots, N\}$ *with* $N \geq 2$ *is the set of players,*
- $\mathcal{S}$ *is a finite set of states*[2]*,*
- $\mathcal{A}_i$ *is a finite set of actions for each* $i \in \mathcal{N}$*. We denote by* $\mathcal{A} \triangleq \times_{i\in\mathcal{N}} \mathcal{A}_i$ *the set of joint actions,*
- $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ *is the Markov transition kernel, where we use throughout this paper the notation* $\Delta(\mathcal{S})$ *for the set of probability distributions over the set* $\mathcal{S}$*,*
- $r_i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ *is the reward function of agent* $i \in \mathcal{N}$ *,*
- $\rho \in \Delta(\mathcal{S})$ *is an initial distribution over states,*
- $\gamma \in (0, 1)$ *is a discount factor .*

[1] We focus on the discounted infinite horizon setting, variants for finite horizon and undiscounted/total cumulative sum rewards can also be defined similarly.

[2] Extension to infinite countable or even continuous state action spaces is possible under additional technical assumptions, we focus on the finite state action setting.

**Remark 2.** There are two useful interpretations of the definition:

- Markov games can be seen as a generalization of Markov Decision Processes (case $N = 1$) which have been extensively studied in the literature, see Puterman (2014); Howard (1960) for textbook references.
- Markov games extend normal-form games to a stateful sequential setting (with a shared state).

**Policies.** For each $i \in \mathcal{N}$, denote by $\pi_i : \mathcal{S} \to \Delta(\mathcal{A}_i)$ the policy of agent $i$ and $\Pi^i$ the set of all possible Markov stationary policies for agent $i \in \mathcal{N}$. We denote by $\pi = (\pi_i)_{i\in\mathcal{N}}$ the joint policy and by $\Pi \triangleq \times_{i\in\mathcal{N}}\Pi^i$ the set of joint Markov stationary policies.

**Interaction protocol.** In a Markov game, the agents' interaction with the environment unfolds as follows: At each time step $t \geq 0$, the agents observe a shared state $s_t \in \mathcal{S}$ and choose a joint action $a_t = (a_{t,i})_{i\in\mathcal{N}}$ according to their joint policy $\pi^{(t)}$, i.e. for every $i \in \mathcal{N}, a_{t,i}$ is sampled according to $\pi_i^{(t)}(\cdot|s_t)$. Each agent $i \in \mathcal{N}$ receives a reward $r_i(s_t, a_t)$. Then the game proceeds by transitioning to a state $s_{t+1}$ drawn from the distribution $P(s_t, a_t)$.

**Occupancy measure.** For any joint policy $\pi \in \Pi$, we define the state occupancy measure $d_\rho^\pi$ for every state $s \in \mathcal{S}$ by

$$d_\rho^\pi(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}_{\rho,\pi}(s_t = s)\,. \tag{2}$$

This quantity captures the visitation frequency of each state.

**Value functions.** For each joint policy $\pi \in \Pi$, and each $i \in \mathcal{N}$, define the state and state-action value functions $V_i^\pi : \mathcal{S} \to \mathbb{R}, Q_i^\pi : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ for each $s \in \mathcal{S}, a \in \mathcal{A}$ by

$$V_i^\pi(s) \triangleq \mathop{\mathbf{E}}_{\pi}\left[\sum_{t=0}^{\infty}\gamma^t r_i(s_t,a_t)|s_0 = s\right], \quad Q_i^\pi(s,a) \triangleq \mathop{\mathbf{E}}_{\pi}\left[\sum_{t=0}^{\infty}\gamma^t r_i(s_t,a_t)|s_0 = s, a_0 = a\right].$$

Observe then that for any joint policy $\pi \in \Pi$, any $i \in \mathcal{N}$ and any state $s \in \mathcal{S}$,

$$V_i^\pi(s) = \sum_{a\in\mathcal{A}}\pi(a|s)Q_i^\pi(s,a)\,. \tag{3}$$

We use the notation $V_i^\pi(\rho) \triangleq \mathbf{E}_{s\sim\rho}[V_i^\pi(s)]$ for any initial state distribution $\rho$ and any joint policy $\pi \in \Pi$.

**Definition 3** (Nash equilibria). *For any $\varepsilon \geq 0$, a joint policy $\pi = (\pi_i)_{i\in\mathcal{N}} \in \Pi$ is an $\varepsilon$-approximate Nash equilibrium for the game $\Gamma$ if for every $i \in \mathcal{N}$ and every $\pi_i' \in \Pi^i$, $V_i^{\pi_i,\pi_{-i}}(\rho) \geq V_i^{\pi_i',\pi_{-i}}(\rho) - \varepsilon$. When $\varepsilon = 0$, such a policy $\pi$ from which no agent has an incentive to deviate unilaterally, is called a Nash equilibrium policy.*

### 1.2 Existence of stationary Nash equilibria

**Theorem 4** (Existence of stationary Nash policy, Fink (1964)). *Every discounted infinite-horizon stochastic game with a finite number of states, actions, and players (as defined in Definition 1), has a stationary Markovian Nash equilibrium policy, i.e. there exist $\pi_i : \mathcal{S} \to (\Delta A_i)$ for all $i \in \mathcal{N}$ s.t. for all $i \in\in \mathcal{N}$ and any policy $\pi_i'$ (not necessarily Markovian), $V_i^{\pi_i',\pi_{-i}}(\rho) \leq V_i^{\pi_i,\pi_{-i}}(\rho)$.*

*Proof.* (Proof sketch) Similarly to the proofs of Nash's theorem, there are several proofs for this theorem based again on fixed point arguments. The original proof of Fink (1964) is based on

Kakutani's fixed point theorem. There is an alternative proof using a Nash mapping similar to the proof we provided in the case of finite normal-form games. We provide here a sketch of proof following this second strategy. A complete proof using this strategy can be found for instance in Deng et al. (2023). Consider the mapping $\varphi : \Pi \to \Pi$ defined for each player $i \in \mathcal{N}$, each action $a_i \in \mathcal{A}_i$ at each state $s \in \mathcal{S}$ as follows:

$$\varphi(\pi)_i(s, a_i) = \frac{\pi_i(a_i|s) + \max\left(0, Q_i^{\pi_i,\pi_{-i}}(s, a_i) - V_i^{\pi_i,\pi_{-i}}(s)\right)}{1 + \sum_{b_i \in \mathcal{A}_i} \max\left(0, Q_i^{\pi^i,\pi^{-i}}(s, b_i) - V_i^{\pi_i,\pi_{-i}}(s)\right)}. \tag{4}$$

The mapping $\varphi$ is continuous over the convex compact set of all stationary Markov policies (as value functions can be shown to be continuous w.r.t. policies). Hence, there exists a fixed point $\pi^\star$ of $\varphi$ by Brouwer's fixed point theorem. It remains to show that $\pi^\star$ is Nash policy.

□

### 1.3 Characterization of Nash equilibria as first-order stationary points

The value function of each agent is non-concave in their policy in general, as this is already the case in standard single-agent RL (see e.g. Lemma 1 in Agarwal et al. (2021)). Nevertheless, the following crucial result establishes a gradient domination property which holds per-agent. This structural property relates utility change under unilateral policy deviation to individual policy gradients.

**Proposition 5** (Agentwise Gradient Domination)**.** *Suppose that for all $\pi \in \Pi, s \in \mathcal{S}$, $d_\mu^\pi(s) > 0$. Then, for all $i \in \mathcal{N}$, all policies $\pi = (\pi_i, \pi_{-i}) \in \Pi$, $\pi_i' \in \Pi_i$, we have for any full-support distribution $\mu, \rho \in \Delta(\mathcal{S})$,*

$$V_i^{\pi_i',\pi_{-i}}(\mu) - V_i^{\pi_i,\pi_{-i}}(\mu) \leq C_\mathcal{G} \cdot \max_{\tilde{\pi}_i \in \Pi_i} \langle \nabla_{\pi_i} V_i^\pi(\rho), \tilde{\pi}_i - \pi_i \rangle$$

*where the minimax distribution mismatch coefficient:* $C_\mathcal{G} := \max_{i \in \mathcal{N}} \max_{\pi \in \Pi} \min_{\pi_i^\star \in \Pi_i^\star(\pi_{-i})} \left\| \frac{d_\mu^{\pi_i^\star,\pi_{-i}}}{d_\rho^{\pi_i,\pi_{-i}}} \right\|_\infty$,

*where $\Pi_i^\star(\pi_{-i})$ denotes the set of best response policies for agent $i$, is finite.*

The above result has been shown in Daskalakis et al. (2020) for 2-player zero-sum Markov games, in Leonardos et al. (2022) for Markov potential games and in Zhang et al. (2024); Giannou et al. (2022) for general-sum Markov games.

We define the *pseudo-gradient* vector field notation for every joint policy $\pi \in \Pi$:

$$v_i(\pi) := \nabla_{\pi_i} V_i^\pi(\mu), \quad v(\pi) = (v_i(\pi))_{i \in \mathcal{N}}. \tag{5}$$

Note that the vector field $v$ is not necessarily a gradient field in general, i.e. the existence of a function $f$ s.t. $v = \nabla f$ is not guaranteed in general. Indeed, each $v_i$ is the gradient of a different function.

**Proposition 6** (First-order Stationary policies are Nash policies)**.** *Suppose that for all $\pi \in \Pi, s \in \mathcal{S}$, $d_\mu^\pi(s) > 0$. A (joint) policy $\pi^\star \in \Pi$ is a Nash policy if and only if it is first-order stationary, i.e.*

$$\langle v(\pi^\star), \pi - \pi^\star \rangle \leq 0, \quad \forall \pi \in \Pi. \tag{FOS}$$

*Moreover, for any $\varepsilon > 0$, if a policy $\pi^\star$ is a $\varepsilon$-FOS policy, i.e. it satisfies: $\langle v(\pi^\star), \pi - \pi^\star \rangle \leq \varepsilon$, $\forall \pi \in \Pi$, then $\pi^\star$ is also a $(C_\mathcal{G}\, \varepsilon)$-Nash policy, where $C_\mathcal{G}$ is the distribution mismatch coefficient defined in Prop. 5.*

*Proof.* We prove both implications. We use the shorthand notation $u_i(\pi) = V_i^\pi(\rho)$ in this proof.

**(i) If $\pi^\star$ is a Nash policy then $\pi^\star$ is FOS.**

This implication is standard in optimization. We provide a proof here for completeness. Let $\pi^\star$ be a Nash policy. By definition of the differentiability of $u_i$ w.r.t $\pi_i$ at $\pi^\star$, we can write for any $\eta > 0$ and any policy $\pi_i' \in \Pi_i$ (using $\delta_i := \pi_i' - \pi_i^\star$) that

$$u_i(\pi_i^\star + \eta\delta_i, \pi_{-i}^\star) = u_i(\pi_i^\star, \pi_{-i}^\star) + \eta \left\langle \nabla_{\pi_i} u_i(\pi^\star), \delta_i \right\rangle + \eta \|\delta_i\| \cdot o_{\eta\to 0}(1)\,. \tag{6}$$

Therefore, we obtain by rearranging the identity

$$\left\langle \nabla_{\pi_i} u_i(\pi^\star), \pi_i' - \pi_i^\star \right\rangle = \frac{1}{\eta}(u_i(\pi_i^\star + \eta\,\delta_i, \pi_{-i}^\star) - u_i(\pi_i^\star, \pi_{-i}^\star)) - \|\delta_i\| \cdot o_{\eta\to 0}(1)\,. \tag{7}$$

By the definition of a Nash equilibrium policy, it follows that

$$\left\langle \nabla_{\pi_i} u_i(\pi^\star), \pi_i' - \pi_i^\star \right\rangle \leq -\|\delta_i\| \cdot o_{\eta\to 0}(1)\,. \tag{8}$$

Taking $\eta \to 0$ yields $\langle \nabla_{\pi_i} u_i(\pi^\star), \pi_i' - \pi_i^\star \rangle \leq 0$, i.e.$\langle v_i(\pi^\star), \pi_i' - \pi_i^\star \rangle \leq 0$. As $i$ is arbitrary in $\mathcal{N}$, the inequality holds for all $i \in \mathcal{N}$ and summing up these inequalities yields $\langle v(\pi^\star), \pi' - \pi^\star \rangle \leq 0$ for all $\pi' \in \Pi$ which means that $\pi^\star$ is a FOS policy.

**(ii) If $\pi^\star$ is a FOS policy then $\pi^\star$ is also a Nash policy.**

The result follows from our gradient domination result of Proposition 5.

Let $\pi^\star = (\pi_i^\star, \pi_{-i}^\star) \in \Pi$ be a FOS policy. By definition it satisfies for all $\pi \in \Pi$, $\langle v(\pi^\star), \pi - \pi^\star \rangle \leq 0$ for all $\pi \in \Pi$. Choose $\pi = (\pi_i', \pi_{-i}^\star)$ in this FOS inequality to obtain $\langle v_i(\pi^\star), \pi_i' - \pi_i^\star \rangle \leq 0$ for all $i \in \mathcal{N}$ and for all $\pi_i' \in \Pi_i$. By Proposition 5, for all $i \in \mathcal{N}$ and for all $\pi_i' \in \Pi_i$, we have

$$u_i(\pi_i', \pi_{-i}^\star) - u_i(\pi_i^\star, \pi_{-i}^\star) \leq C_{\mathcal{G}} \cdot \max_{\tilde{\pi}_i \in \Pi_i} \langle v_i(\pi^\star), \tilde{\pi}_i - \pi_i^\star \rangle\,. \tag{9}$$

This in turn implies that $u_i(\pi_i', \pi_{-i}^\star) - u_i(\pi_i^\star, \pi_{-i}^\star) \leq 0$ for all $i \in \mathcal{N}$ and for all $\pi_i' \in \Pi_i$ and hence $\pi^\star$ is a Nash policy. □

### 1.4 Hardness results for Nash equilibrium computation

Computing Nash equilibria in Markov games is at least as hard as in normal-form games as these are particular cases of Markov games when the state space consists of a single state (stateless setting) and $\gamma = 0$ (with the convention $0^0 = 1$). Complexity results for general-sum Markov games have been established recently in the literature (Daskalakis et al., 2023; Jin et al., 2023; Foster et al., 2023; Deng et al., 2023).

## 2 Subclasses of Markov Games

We focus here on two fundamental subclasses of Markov games in which Nash equilibrium computation is tractable, mirroring our presentation for (static) normal-form games. Other classes such as adversarial team Markov games have also been investigated in the literature.

### 2.1 Zero-Sum Markov Games

**Definition 7** (Two-Player Zero-Sum Markov Game)**.** *A Markov game $\Gamma$ is said to be zero-sum if the reward functions satisfy $r_1 + r_2 = 0$.*

Note that the definition immediately implies that $V_1^{\pi_1,\pi_2}(\rho) + V_2^{\pi_1,\pi_2}(\rho) = 0$ for all policies $\pi_1 \in \Pi^1, \pi_2 \in \Pi^2$.

Similarly to two-player zero-sum normal-form games, a minmax theorem holds.

**Theorem 8** (Minmax theorem, Shapley (1953)). *For any two-player zero-sum Markov game,*

$$\min_{\pi_1 \in \Pi^1} \max_{\pi_2 \in \Pi^2} V_2^{\pi_1,\pi_2}(\rho) = \max_{\pi_2 \in \Pi^2} \min_{\pi_1 \in \Pi^1} V_2^{\pi_1,\pi_2}(\rho)\,. \tag{10}$$

The duality gap is equal to zero. Note that the value function is not convex-concave in general.

*Proof.* (Proof sketch) The proof relies on a contraction argument based on Shapley's operator which we define in the following. Let $\mathrm{val}(\cdot)$ be the value operator which assigns to each payoff matrix the value of the corresponding finite zero-sum game which is well-defined by the (classical) min-max theorem. Then define the operator $\mathcal{T} : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ for any vector $V \in \mathbb{R}^{|\mathcal{S}|}$ and any $s \in \mathcal{S}$ by

$$\mathcal{T}V(s) := \mathrm{val}\left(\left[r_2(s,a,b) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a,b)V(s')\right]_{a \in \mathcal{A}_1, b \in \mathcal{A}_2}\right) \tag{11}$$

It can be easily shown using the non-expansiveness of the value operator (i.e. $|\mathrm{val}(A) - \mathrm{val}(B)| \leq \max_{i,j} |A_{i,j} - B_{i,j}|$ for two payoff matrices $A, B$) that $\mathcal{T}$ is a $\gamma$-contraction in the infinity norm, i.e. $\|\mathcal{T}V - \mathcal{T}V'\|_\infty \leq \gamma \|V - V'\|_\infty$ for any $V, V' \in \mathcal{R}^{|\mathcal{S}|}$. This implies in turn that $\mathcal{T}$ admits a unique fixed point by Banach fixed-point theorem (i.e. there exists a unique $V^\star \in \mathbb{R}^{|\mathcal{S}|}$ s.t. $\mathcal{T}V^\star = V^\star$). □

### 2.2 Markov Potential Games

The class of Markov Potential games defined below extends potential games in **static** normal-form games.

**Definition 9** (Markov Potential Game). *A Markov game $\Gamma$ is a Markov Potential Game (MPG) if $\forall s \in \mathcal{S}, \exists \Phi_s : \Pi \to \mathbb{R}$ (player independent) s.t. $\forall i \in \mathcal{N}$, $(\pi_i, \pi_{-i}) \in \Pi$, $\pi_i' \in \Delta(\mathcal{A}_i)^{\mathcal{S}}$,*

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

It follows from taking the expectation w.r.t. the inital state distribution $\rho$ in the above definition that we can also write:

$$V_i^{\pi_i',\pi_{-i}}(\rho) - V_i^{\pi_i,\pi_{-i}}(\rho) = \Phi^{\pi_i',\pi_{-i}}(\rho) - \Phi^{\pi_i,\pi_{-i}}(\rho)\,, \tag{12}$$

where we slightly abuse notation by using the overloaded notation $\Phi^\pi(\rho) \triangleq \mathbf{E}_{s\sim\rho}[\Phi^\pi(s)]$ for any policy $\pi \in \Pi$ (similarly as for value functions).

The identical-interest case is an important particular case of this definition. In this case in which all the reward functions are identical ($r_i = r_j, \forall i, j \in \mathcal{N}$), it can be easily seen that the total potential function is the value function of any of the players. Beyond this important case, conditions to obtain MPGs with non-identical rewards were identified in the literature.

## 3 Independent Policy Gradient Methods for Markov Games

### 3.1 Independent learning protocol

Independent learning has recently attracted increasing attention thanks to its versatility as a learning protocol. We refer the reader to a recent nice survey on the topic (Ozdaglar et al., 2021). In this protocol, agents can only observe the realized state and their own reward and action in each stage to individually optimize their return. In particular, each agent does not observe actions or policies from any other agent. This protocol offers several advantages including the following aspects:

(a) **Scaling:** independent learning dynamics do not scale exponentially with the number of players in the game (also known as the *curse of multi-agents*), guarantees will typically scale with $\sum_{i=1}^{N} |\mathcal{A}_i|$ rather than $\prod_{i=1}^{N} |\mathcal{A}_i|$;

(b) **Privacy protection:** agents may avoid sharing their local data and information to protect their privacy and autonomy;

(c) **Communication cost** a central node that can bidirectionally communicate with all agents may not exist or may be too expensive to afford.

Therefore, this protocol is particularly appealing in several applications where agents need to make decisions independently, in a decentralized manner.

### 3.2 Multi-agent policy gradient algorithm

In single agent RL, a natural algorithm for policy optimization consists in performing gradient ascent steps w.r.t. the policy optimization objective using gradients w.r.t. some policy parameterization. Policy gradient estimates can be computed using Monte-Carlo estimates only based on sampled trajectories (Sutton et al., 1999; Williams, 1992).

Theoretical results on the convergence of policy gradient methods in single-agent RL are recent in the literature. We refer the reader to Agarwal et al. (2021); Xiao (2022) for a more detailed exposition regarding these methods and their convergence rates.

In this section, we briefly show how this approach can be extended to our multi-agent setting. Note that Proposition 6 is an additional theoretical motivation for resorting to these policy gradient methods.

**Theorem 10** (Policy gradient theorem)**.** *For any player $i \in \mathcal{N}$ and any policy $\pi \in \Pi$,*

$$\nabla_{\pi_i} V_i^{\pi}(\mu) = \mathbb{E}_{\mu,\pi}\left[\sum_{t=0}^{+\infty} \gamma^t r_i(s_t, a_t) \sum_{t'=0}^{t} \nabla_{\pi_i} \log \pi_i(a_{i,t'}|s_{t'})\right]. \tag{13}$$

*Proof.* The result follows from applying the standard policy gradient theorem for single-agent RL (e.g. Sutton et al. (1999)) and selecting the $i$th coordinate. □

Using the above result, stochastic policy gradients $\hat{\nabla}_{\pi_i} V_i^{\pi}(\mu)$ can be computed by sampling finite-length trajectories using Monte-Carlo estimates. The multi-agent policy gradient algorithm is then run independently by each agent. Each agent $i \in \mathcal{N}$ seeks to maximize their own value function and updates their policy as follows:

$$\pi_i^{t+1} = \pi_i^{t+1} + \eta_i \hat{\nabla}_{\pi_i} V_i^{\pi^t}(\mu), \tag{14}$$

where $\eta_i$ is a positive step size.

## 4 Regret Analysis Framework for Markov Games

**Definition 11** (Nash regret)**.** *We define Nash regret for every time horizon $T \geq 1$ as follows:*

$$\textit{Nash-regret(T)} \triangleq \frac{1}{T}\sum_{t=1}^{T} \max_{i\in\mathcal{N}} \max_{\pi_i' \in \Pi^i} V_i^{\pi_i', \pi_{-i}^{(t)}}(\rho) - V_i^{\pi^{(t)}}(\rho),$$

*where $\pi^{(t)} = (\pi_i^{(t)}, \pi_{-i}^{(t)})$ is the joint policy of the $N$ players at time step $t \in \{1, \cdots, T\}$ and $\rho$ is the initial distribution over the state space $\mathcal{S}$.*

At each time step $t$ and for every player $i \in \mathcal{N}$, the joint policy $\pi^{(t)}$ is compared to the policy where player $i$ unilaterally deviates to its best response to policy $\pi_{-i}^{(t)}$. The difference in value functions quantifies player $i$'s Nash gap. The Nash regret computes the average over the worst player's Nash gap induced by the joint policy $\pi^{(t)}$ over the time horizon $T$.

**Proposition 12.** *The following statements hold true:*

*(i) For every $T \geq 1$, Nash-regret(T) $\geq 0$.*

*(ii) If Nash-regret(T) $\leq \varepsilon$ for some accuracy $\varepsilon > 0$, then there exists $t^* \in \{1, \cdots, T\}$ such that $\pi^{(t^*)}$ is an $\varepsilon$-Nash equilibrium.*

*Proof.* The result is immediate from the definition. □

Several Nash regret and convergence to approximate Nash guarantees can be found in the recent literature, using similar techniques to the ones we discussed in the previous lectures for normal-form games.

## References

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

Constantinos Daskalakis, Dylan J Foster, and Noah Golowich. Independent policy gradient methods for competitive reinforcement learning. *Advances in Neural Information Processing Systems*, 33:5527–5540, 2020.

Constantinos Daskalakis, Noah Golowich, and Kaiqing Zhang. The complexity of markov equilibrium in stochastic games. In *Annual Conference on Learning Theory*, pages 4180–4234. PMLR, 2023.

Xiaotie Deng, Ningyuan Li, David Mguni, Jun Wang, and Yaodong Yang. On the complexity of computing markov perfect equilibrium in general-sum stochastic games. *National Science Review*, 10(1):nwac256, 2023.

Arlington M Fink. Equilibrium in a stochastic $n$-person game. *Journal of science of the hiroshima university, series ai (mathematics)*, 28(1):89–93, 1964.

Dylan J Foster, Noah Golowich, and Sham M Kakade. Hardness of independent learning and sparse equilibrium computation in markov games. In *International Conference on Machine Learning*, pages 10188–10221. PMLR, 2023.

Angeliki Giannou, Kyriakos Lotidis, Panayotis Mertikopoulos, and Emmanouil-Vasileios Vlatakis-Gkaragkounis. On the convergence of policy gradient methods to nash equilibria in general stochastic games. *Advances in Neural Information Processing Systems*, 35:7128–7141, 2022.

Ronald A Howard. Dynamic programming and markov processes. 1960.

Yujia Jin, Vidya Muthukumar, and Aaron Sidford. The complexity of infinite-horizon general-sum stochastic games. *ITCS, arXiv preprint arXiv:2204.04186*, 2023.

Stefanos Leonardos, Will Overman, Ioannis Panageas, and Georgios Piliouras. Global convergence of multi-agent policy gradient in markov potential games. In *International Conference on Learning Representations*, 2022.

Asuman Ozdaglar, Muhammed O Sayin, and Kaiqing Zhang. Independent learning in stochastic games. *Invited chapter for the International Congress of Mathematicians 2022 (ICM 2022), arXiv preprint arXiv:2111.11743*, 2021.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Lloyd S Shapley. Stochastic games. *Proceedings of the National Academy of Sciences*, 39(10): 1095–1100, 1953.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in Neural Information Processing Systems*, 12, 1999.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Lin Xiao. On the convergence rates of policy gradient methods. *Journal of Machine Learning Research*, 23(282):1–36, 2022.

Runyu Zhang, Zhaolin Ren, and Na Li. Gradient play in stochastic games: stationary points, convergence, and sample complexity. *IEEE Transactions on Automatic Control*, 69(10):6499–6514, 2024.