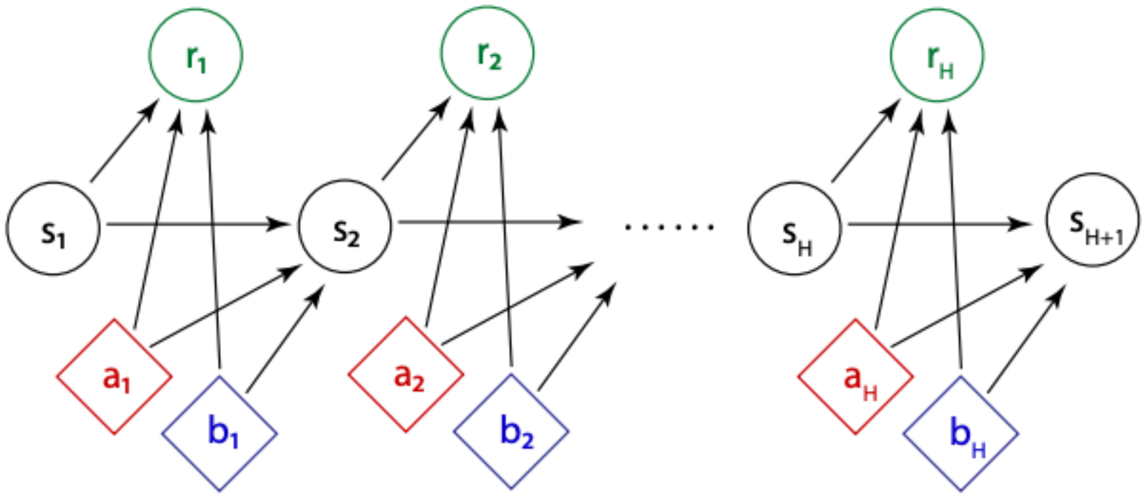
Chapter 10: Multi-agent RL (Continued)

Reminder: Course Project due next Tuesday

#	Team	Members	Score	Agents	Last	Join
1	Team GO		3000.8	2	1d	
2	Team Q		2523.0	2	1d	
3	Team S		2439.3	2	17d	
4	Team S_1		2347.4	2	22d	
5	MilesLiiii		2019.9	2	6h	
6	Team Lux		1464.8	2	1d	

Stochastic/Markov Games



Two-player zero-sum Markov Game $(\mathcal{S}, \mathcal{A}, \mathcal{B}, \mathbb{P}, r, H)$ [Shapley 1953].

- $\mathcal{S}$: set of states; $\mathcal{A}, \mathcal{B}$: set of actions for the max-player/the min-player.
- $\mathbb{P}_h(s_{h+1}|s_h, a_h, b_h)$: transition probability.
- $r_h(s_h, a_h, b_h) \in [0, 1]$: reward for the max-player (loss for the min-player).
- H: horizon/the length of the game.

Planning in Markov Games

A dynamical programming approach to find a Nash equilibrium.

Nash Value Iteration (Nash VI)

Initialize $V^{\star}_{H+1}(s) = 0$ for all s.

for $h = H, \ldots, 1$,

  for all (s, a, b),

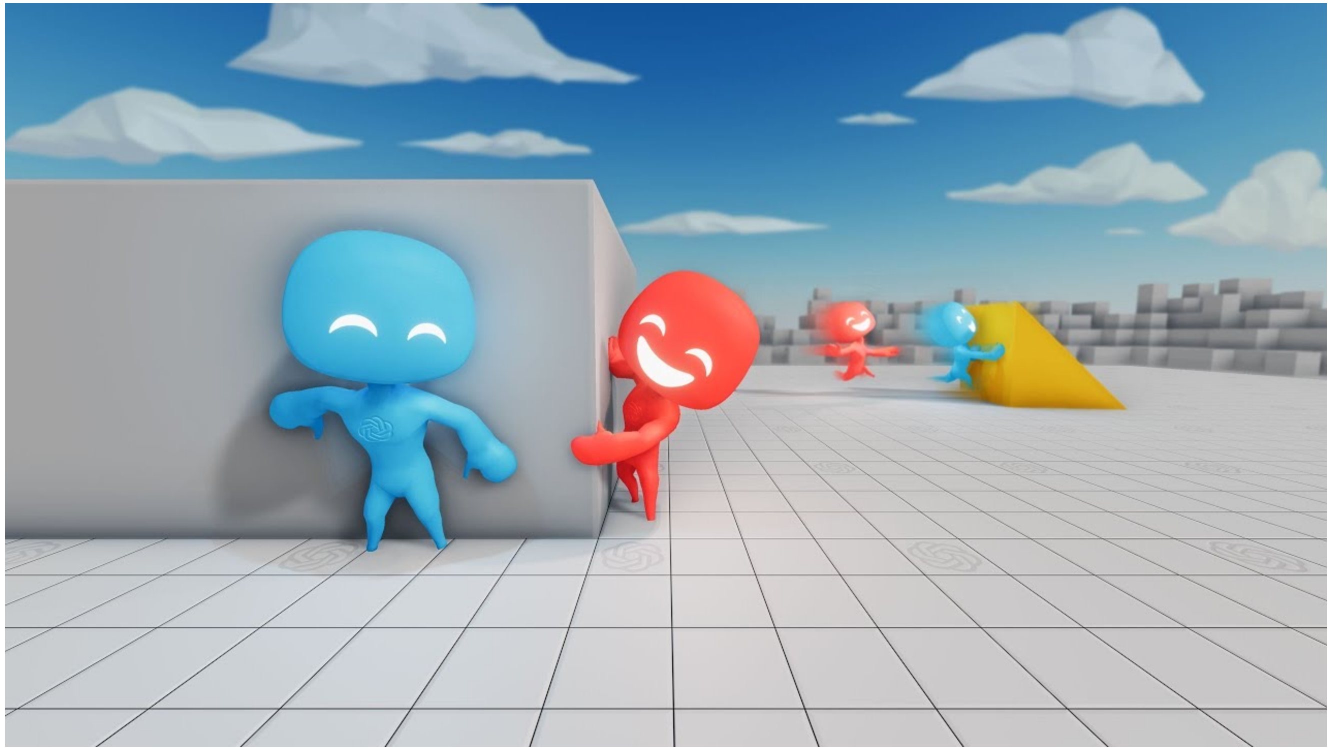
    $Q^{\star}_{h}(s, a, b) \leftarrow r_h(s, a, b) + \mathbb{E}_{s' \sim \mathbb{P}_h(\cdot|s,a,b)} V^{\star}_{h+1}(s')$

  for all s

    $\boxed{(\pi^{\star}_{1,h}(\cdot|s), \pi^{\star}_{2,h}(\cdot|s)) \leftarrow \mathrm{Nash}(Q^{\star}_{h}(s, \cdot, \cdot))}$ **NE for Normal-form Game**

    $V^{\star}_{h}(s) \leftarrow \langle \pi^{\star}_{1,h}(\cdot|s) \times \pi^{\star}_{2,h}(\cdot|s), Q^{\star}_{h}(s, \cdot, \cdot) \rangle$

Today: Online Learning in Unknown MGs



How do we explore in an unknown Markov Game to learn an ϵ-Nash strategy?

Online Learning in Unknown MGs

Interaction protocol

Environment samples initial state s_1.

for step $h = 1, \ldots, H$,

two agents take their own actions (a_h, b_h) simultaneously.

both agents receive their own immediate reward $\pm r_h(s_h, a_h, b_h)$.

environment transitions to the next state $s_{h+1} \sim \mathbb{P}_h(\cdot | s_h, a_h, b_h)$.

Recall UCBVI for Single-agent RL

Inside iteration n :

Use all previous data to estimate transitions $\widehat{P}^n$

Design reward bonus $b_h^n(s,a), \forall s, a, h$

Optimism

Optimistic planning with learned model: $\pi^n = \text{Value-Iter}\left(\widehat{P}^n, \{r_h + b_h^n\}_{h=1}^{H-1}\right)$

Collect a new trajectory by executing π^n in the real world P starting from s_0

🤔 How do we achieve optimism in Two-Player Zero-sum MG?

How do we modify Nash-VI?

A dynamical programming approach to find a Nash equilibrium.

Nash Value Iteration (Nash VI)
Initialize $V^{\star}_{H+1}(s) = 0$ for all s.
for $h = H, \ldots, 1$,
for all (s, a, b),
$Q^{\star}_h(s, a, b) \leftarrow r_h(s, a, b) + \mathbb{E}_{s' \sim \mathbb{P}_h(\cdot\|s,a,b)} V^{\star}_{h+1}(s')$
for all s
$(\pi^{\star}_{1,h}(\cdot\|s), \pi^{\star}_{2,h}(\cdot\|s)) \leftarrow \mathrm{Nash}(Q^{\star}_h(s, \cdot, \cdot))$
$V^{\star}_h(s) \leftarrow \langle \pi^{\star}_{1,h}(\cdot\|s) \times \pi^{\star}_{2,h}(\cdot\|s), Q^{\star}_h(s, \cdot, \cdot) \rangle$

Optimistic Nash-VI

Optimistic Nash VI

Reward bonus:

$$\beta = O\left(\sqrt{\frac{1}{N_k(s,a,b)}}\right)$$

for $k = 1, \ldots, K,$

  for $h = H, \ldots, 1,$

    for all $(s, a, b),$

      $\overline{Q}_h(s,a,b) \leftarrow r_h(s,a,b) + \mathbb{E}_{s' \sim \hat{\mathbb{P}}_h(\cdot|s,a,b)} \overline{V}_{h+1}(s') + \beta$

      $\underline{Q}_h(s,a,b) \leftarrow r_h(s,a,b) + \mathbb{E}_{s' \sim \hat{\mathbb{P}}_h(\cdot|s,a,b)} \underline{V}_{h+1}(s') - \beta$

    for all s

      $\boxed{\pi_h(\cdot,\cdot|s) \leftarrow \mathrm{CCE}(\overline{Q}_h(s,\cdot,\cdot), \underline{Q}_h(s,\cdot,\cdot))}$ **CCE instead of NE**

      $\overline{V}_h(s) \leftarrow \langle \pi_h(\cdot,\cdot|s), \overline{Q}_h(s,\cdot,\cdot) \rangle$

      $\underline{V}_h(s) \leftarrow \langle \pi_h(\cdot,\cdot|s), \underline{Q}_h(s,\cdot,\cdot) \rangle$

  execute policy π, collect samples, and update estimation $\hat{\mathbb{P}}$.

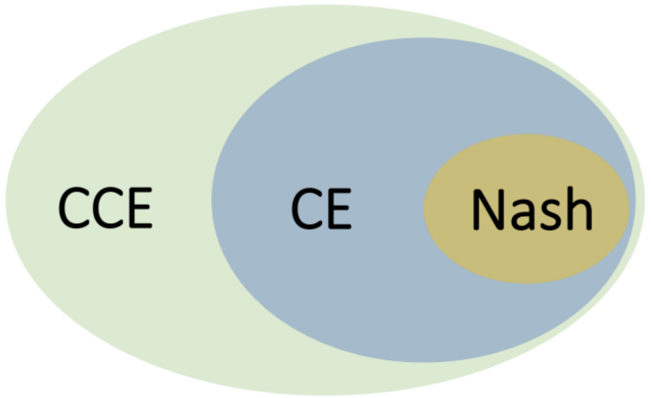
Coarse Correlated Equilibria

- **Coarse Correlated Equilibria (CCE)**: A joint policy $\boldsymbol{\pi}: \boldsymbol{S} \to \boldsymbol{A} \times \boldsymbol{B}$ is a CCE if

$$\max_{\pi':S\to A} V^{\pi',\pi_{-1}} \leq V^{\pi} \text{ and } \max_{\pi':S\to B} V^{\pi_{-2},\pi'} \geq V^{\pi}$$

- **CCE v.s. NE**:
  - CCE allows correlated polices, e.g. traffic light.

	STOP	GO
STOP	(0,0)	(0,1)
GO	(1,0)	(-100,-100)



  - CCE is efficiently computable for general-sum games, while NE isn't.

Theoretical Guarantee of Nash-VI

Theorem [Liu, Yu, Bai, Jin 2020]

With high probability, optimistic Nash VI finds an ϵ-Nash equilibrium in $\tilde{O}(H^3SAB/\epsilon^2)$ episodes.

H: horizon; S: number of states; A, B: number of actions for each player.

Optimistic Nash VI finds ϵ-Nash in polynomial time and samples!

Drawbacks of Nash-VI

- Centralized learning: Requires keeping track of $Q(s, a, b)$.
- The algorithm can be generalized to the multi-agent setting:
- Nash-VI finds an **ϵ-CCE** with $\boldsymbol{O\left(\mathrm{poly}(S \prod_{i=1}^{n} A_i)\right)}$ sample and computational complexity.
- "The Curse of Multi-agent": $\boldsymbol{\prod_{i=1}^{n} A_i}$ scaling

The Curse of Multi-agent

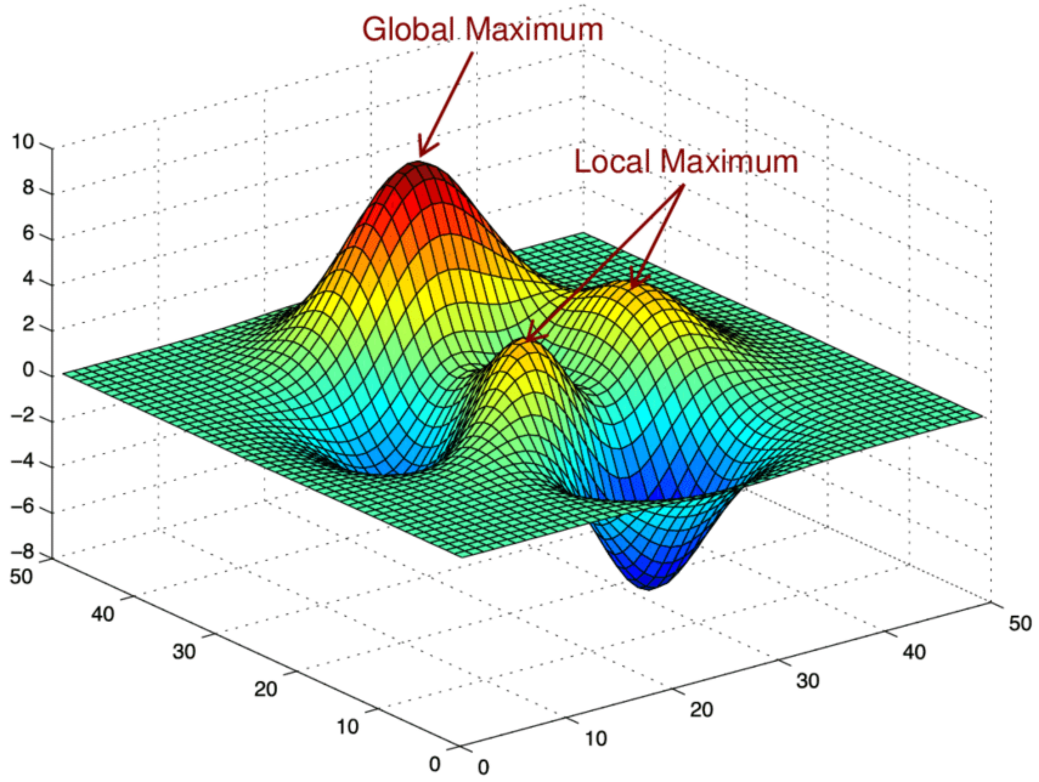
- Can we avoid the $O(AB)$ scaling?

Information theoretical lower bound: $\Omega(H^3 S \max\{A, B\}/\epsilon^2)$

- Observation: Nash-VI requires estimating the Q function with SAB entries, naturally resulting in the scaling with $O(SAB)$.

The Curse of Multi-agent

- But why can we avoid trying each (s, a, b) tuple at least once?



Simpler Setting: Normal-form Game

Each agent runs no-regret algorithm for adversarial bandit (e.g. EXP3) independently.

$$\sum_{t=1}^{T}\langle \mu_t, \ell_t\rangle - \min_{a\in\mathcal{A}}\sum_{t=1}^{T}\langle a, \ell_t\rangle \le \mathrm{poly}(A)\, T^{1-\alpha}.$$

- two-player zero-sum games: $(\mathbb{E}_{t\sim \mathrm{Unif}(T)}\mu_t^{(1)}) \times (\mathbb{E}_{t\sim \mathrm{Unif}(T)}\mu_t^{(2)}) \to$ Nash.
- sample complexity scales with $\tilde{O}(A+B)$.

Unfortunately, cannot run no-regret algorithm in MGs (recall from last lecture).

V-Learning

V-learning [Bai, Jin, Yu, 2020] [Jin, Liu, Wang, Yu, 2021]

for $k = 1, \ldots, K$, receive s_1,

for step $h = 1, \ldots, H$,

take action $a_h \sim \pi_h(\cdot|s_h)$, observe reward r_h and next state s_{h+1}.

$t = N_h(s_h) \leftarrow N_h(s_h) + 1$.

$V_h(s_h) \leftarrow (1 - \alpha_t)V_h(s_h) + \alpha_t(r_h + V_{h+1}(s_{h+1}) + \beta_t)$.

$\pi_h(\cdot|s_h) \leftarrow$ Adv_Bandit_Update$(a_h, r_h + V_{h+1}(s_{h+1}))$

on the $(s_h, h)^{\text{th}}$ adversarial bandit.

- Incremental updates of V instead of Q!
- Is a single-agent algorithm.

Theoretical Guarantee

- Multiagent setting: both agents run V-learning independently.
- Adversarial bandit subroutine: FTRL.

Theorem [Bai, Jin, Yu, 2020]

In two-player zero-sum Markov games, V-learning with FTRL finds ϵ-Nash in $\tilde{O}(H^5 S \max\{A, B\}/\epsilon^2)$ episodes.

V-learning is a decentralized algorithm that achieves optimal $O(\max\{A, B\})$ sample complexity!

Readily Generalize to Multi-agent MGs

Theorem (CCE & CE) [Song et al. 2021][Jin, Liu, Wang, Yu, 2021]

In general-sum Markov games,
(1) V-learning with FTRL finds ϵ-CCE in $\tilde{\mathcal{O}}(H^5 S(\max_{i\in[m]} A_i)/\epsilon^2)$ episodes;
(2) V-learning with FTRL_swap finds ϵ-CE in $\tilde{\mathcal{O}}(H^5 S(\max_{i\in[m]} A_i)^2/\epsilon^2)$ episodes.

Summary of Algorithms

Algorithm	Training	Main estimand	Sample complexity
Nash-VI	centralized	$\mathbb{P}_h(s'\|s, a, b)$	$\tilde{\mathcal{O}}(H^3 SAB/\epsilon^2)$
Nash Q-Learning	centralized	$Q_h^\star(s, a, b)$	$\tilde{\mathcal{O}}(H^5 SAB/\epsilon^2)$
V-Learning	decentralized	$V_h^\star(s)$	$\tilde{\mathcal{O}}(H^5 S \max\{A, B\}/\epsilon^2)$
Lower bound	-	-	$\Omega(H^3 S \max\{A, B\}/\epsilon^2)$

Lots of Future Work to be done

- Behavior of Decentralized Algorithms.
- Policy Gradient for Markov Games?
- Scalable algorithms? (closing theory-practice gap)
- Imperfect Information Markov Games.