# DS 598
# Introduction to RL

Xuezhou Zhang

# Reminder

- Sign up for your team by Jan 27th [link].

| Presentation Date | Team Name | Team Members | | |
|---|---|---|---|---|
| 03/19 | Team Zero | Mao Mao | Haotian Shangguan | |
| 03/21 | Team RL | Seunghwan Hyun | Zoey Yang | |
| 03/26 | Team Alpha | Ayush Sharma | Gauravdeep Singh Bindra | |
| 03/28 | | | | |
| 04/02 | Team S | Sahana Kowshik | | |
| 04/04 | Team Reward | Xinyu Zhang | Lilin Jin | Yan Si |
| 04/09 | | Xavier Thomas | Shiva Charan | |
| 04/11 | | | | |
| 04/16 | | | | |
| 04/18 | | | | |
| 04/23 | | | | |
| 04/25 | | | | |

# Reminder

- Sign up for your team by Jan 27th [link].

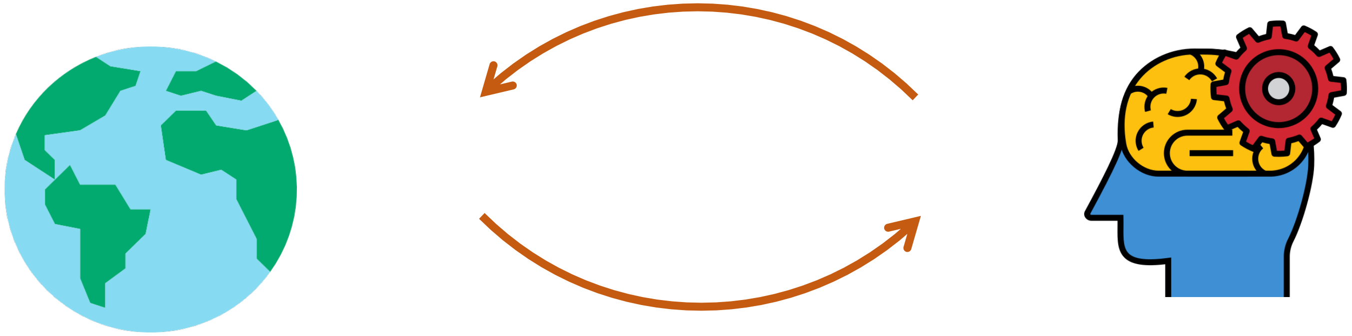
| | Team 1 | Team Members | | | Team 2 | Team Members | | |
|---|---|---|---|---|---|---|---|---|
| 03/19 | Team Zero | Mao Mao | Haotian Shangguan | | | | | |
| 03/21 | Team RL | Seunghwan Hyun | Zoey Yang | | | | | |
| 03/26 | Team Alpha | Ayush Sharma | Gauravdeep Singh Bindra | | | | | |
| 03/28 | Team WYA | Wylliam | Yichen | Andy Yang | | | | |
| 04/02 | Team S | Sahana Kowshik | | | | | | |
| 04/04 | Team Reward | Xinyu Zhang | Lilin Jin | Yan Si | | | | |
| 04/09 | Team Q | Xavier Thomas | Shiva Charan | | | | | |
| 04/11 | Team Gamma | | | | | | | |
| 04/16 | | | | | | | | |
| 04/18 | | | | | | | | |
| 04/23 | | | | | | | | |
| 04/25 | | | | | | | | |

# Reminder

- Course Announcements on the Blackboard site.
- Piazza created for any discussions.

# Recap: MDP

Perform action: $a_h \sim \pi(\cdot \mid s_h)$



Environment

RL Agent

Receive Reward: $r_h \sim r(s_h, a_h)$

Observe Next state: $s_{h+1} \sim P(\cdot \mid s_h, a_h)$

# Recap: Infinite Horizon MDP

- MDP $\mathcal{M} = \{S, A, P, r, \gamma\}$
  - $S$ is the state space.
  - $A$ is the action space.
  - $P: S \times A \to \Delta(S)$ is the transition probability function.
  - $r: S \times A \to [0,1]$ is the reward function.
  - $\gamma \in [0,1)$ is the discounting factor.

- A Markovian policy is defined as $\pi: S \to \Delta(A)$.

# Recall from last time

- Value function

$$V^\pi(s) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) | s_0 = s, a_h \sim \pi(s_h), s_{h+1} \sim P(\cdot|s_h, a_h)\right]$$

- Q function

$$Q^\pi(s,a) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) | (s_0, a_0) = (s,a), a_h \sim \pi(s_h), s_{h+1} \sim P(\cdot|s_h, a_h)\right]$$

$$\boxed{\text{Bellman Equation: } V^\pi(s) = \mathbb{E}[r(s, \pi(s))] + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^\pi(s')}$$

# Existence of an Optimal Policy

- $V^*(s) = \max_\pi V^\pi(s)$
- $Q^*(s,a) = \max_\pi Q^\pi(s,a)$
- So far, we know that there exists an optimal $\pi_s^*$ per state $s$.

🤔 Is there a single policy that achieves $V^*$ and $Q^*$ for all $s$?

Turns out the answer is Yes.

# Existence of an Optimal Policy

- [Claim] There exist a stationary and deterministic policy $\pi$, s.t.

$$\forall (s,a) \in S \times A, V^{\pi}(s) = V^{*}(s) \text{ and } Q^{\pi}(s,a) = Q^{*}(s,a)$$

- Proof by Construction:

$$\pi^{*}(s) = \arg\max_{a \in A} \left[ r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^{\star}(s') \right]$$

# Existence of an Optimal Policy

- Proof by Construction: $\pi^*(s) = \arg\max_{a \in A} \left[ r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^\star(s') \right]$

- Let's prove $V^{\pi^*}(s) = V^*(s) \text{ for all } s \in S.$

- We already know, by the definition of $V^*$, that $V^{\pi^*}(s) \leq V^*(s)$

- It remains to be shown that $V^{\pi^*}(s) \geq V^*(s)$

# Existence of an Optimal Policy

$$
\begin{aligned}
V^*(s_0) &= \max_\pi \mathbb{E}\left[r(s_0, a_0) + \gamma \mathbb{E}_{s_1 \sim P(\cdot|s_0, a_0)}[V^\pi(s_1)] | \pi\right] && \text{(Bellman Equation)} \\
&\leq \max_\pi \mathbb{E}\left[r(s_0, a_0) + \gamma \mathbb{E}_{s_1 \sim P(\cdot|s_0, a_0)}[\max_\pi V^\pi(s_1)] | \pi\right] && \text{(Jensen)} \\
&= \max_\pi \mathbb{E}\left[r(s_0, a_0) + \gamma \mathbb{E}_{s_1 \sim P(\cdot|s_0, a_0)}[V^*(s_1)] | \pi\right] && \text{(Definition of } V^*\text{)} \\
&= \mathbb{E}\left[r(s_0, a_0) + \gamma \mathbb{E}_{s_1 \sim P(\cdot|s_0, a_0)}[V^*(s_1)] | \pi^*\right] \\
&\leq \mathbb{E}\left[r(s_0, a_0) + \gamma r(s_1, a_1) + \gamma^2 \mathbb{E}_{s_2 \sim P(\cdot|s_1, a_1)}[V^*(s_2)] | \pi^*\right] && \text{(recursion)} \\
&\leq \mathbb{E}\left[r(s_0, a_0) + \gamma r(s_1, a_1) + \gamma^2 r(s_2, a_2) + ... | \pi^*\right] \\
&= V^{\pi^*}(s_0)
\end{aligned}
$$

# Bellman Optimality Equation

- We have shown that $V^* = V^{\pi^*}$ and thus

$$V^*(s) = \max_a \left[ r(s,a) + \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^*(s')] \right]$$

- Bellman Optimality Equation

$$\color{red}{f}(s) = \max_a \left[ r(s,a) + \mathbb{E}_{s' \sim P(\cdot|s,a)}[\color{red}{f}(s')] \right]$$

# Summary

| $V^{\pi}$ | $V^{*}$ |
|---|---|
| Bellman Equation: $V^{\pi}(s) = r\big(s, \pi(s)\big) + \gamma \mathbb{E}_{s' \sim P(\cdot \mid s, \pi(s))}[V^{\pi}(s')]$ | Bellman Optimality Equation: $V^{*}(s) = \max_{a} \left[ r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot \mid s, a)}[V^{*}(s')] \right]$ |

- $f$ satisfies Bellman Equation iff $f = V^{\pi}$ for some $\pi$.
- $f$ satisfies Bellman Optimality Equation iff $f = V^{*}$.

# Chapter 2: Planning

# What is planning?

- “Given” an MDP, find an optimal policy.
- It’s a pure “computational” problem.
- There is no “learning” involved.
- Still highly non-trivial!! e.g. AlphaGo.

# Approach 1: Solving the Bellman Optimality Equation

- It's easy! Simply solve the BOE:

$$Q^*(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[\max_{a' \in A} Q^*(s',a')]$$

# How do we solve BOE?

- Fixed-point iteration (FPI) method:

- To solve equation $x = f(x)$

1. Initialize $x^{(0)}$ arbitrarily.
2. For $t = 1, \ldots T$,
   - Compute $x^{(t)}(s, a) = f(x^{(t-1)})$
3. Return $x^{(T)}$

# When does FPI work?

- FPI doesn't necessarily converge.
- A sufficient condition for FPI to work is called the contraction property.
- Contraction: $\exists \gamma \in [0,1)$, s.t. $\forall\, x, x'$, $\left|\left|f(x) - f(x')\right|\right| \leq \gamma \cdot \left|\left|x - x'\right|\right|$
- Since $x^* = f(x^*)$, we have $\left|\left|f\left(x^{(t)}\right) - f(x^*)\right|\right| \leq \gamma \cdot \left|\left|x^{(t-1)} - x^*\right|\right|$.

# How do we solve BOE?

HW: Prove that BOE satisfies contraction.

- In RL, FPI is called Value Iteration

1. Initialize $Q^{(0)}$ arbitrarily.
2. For $t = 1, \ldots T$
   - $Q^{(i)}(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[ \max_{a'} Q^{(i-1)}(s', a') \right]$

- The implicit assumptions
  - 🥲 Finite S and A
  - 🥲 Can evaluate $\mathbb{E}_{s' \sim P(\cdot|s, a)}[\cdot]$

# Approach 1b: Policy Iteration

- Instead of updating the Q function, policy iteration updates the policy.

1. Initialize $\pi^{(0)}$ arbitrarily.
2. For $t = 1, \dots T$
    - **Policy Evaluation**: $Q^{\pi^{(t-1)}}$.
    - **Policy Improvement**: $\pi^{(t)}(s,a) = \text{argmax}_a Q^{\pi^{(t-1)}}(s,a)$.

- One can show that $\left\|Q^{\pi^{(t)}} - Q^*\right\|_\infty \leq \gamma \cdot \left\|Q^{\pi^{(t-1)}} - Q^*\right\|_\infty$.

# Approach 2: Linear Programming

- Occupancy Measure:

$$d_\mu^\pi(s,a) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathrm{Pr}^\pi(s_t = s, a_t = a | s_0 \sim \mu)$$

- Starting from $s_0 \sim \mu$, follow $\pi$,
- at every step, stop with prob. $(1-\gamma)$
- if stop, sample $(s,a)$ at that step.

# Approach 2: Linear Programming

- Connection to the Value Function:

$$V^\pi(\bar{s}) = r(s,a)^\top d_{\bar{s}}^\pi(s,a)$$

- Bellman-like Recursion:

$$\sum_a d_\mu^\pi(s,a) = (1-\gamma)\mu(s) + \gamma \sum_{s',a'} P(s|s',a')d_\mu^\pi(s',a') \qquad (1)$$

- $d$ satisfies (1) iff $d = d_\mu^\pi$ for some $\pi$, in particular $\pi_d(a|s) = \frac{d(s,a)}{\sum_a d(s,a)}$.

# Approach 2: Linear Programming

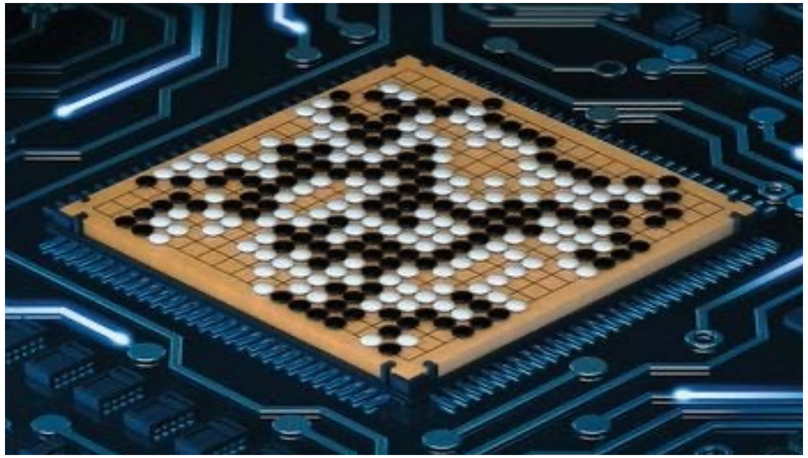
$$\max_{d \in \Delta(S \times A)} \sum_{s,a} r(s,a) d(s,a)$$

$$\text{s.t.} \sum_{a} d_\mu^\pi(s,a) = (1-\gamma)\mu(s) + \gamma \sum_{s',a'} P(s|s',a') d_\mu^\pi(s',a')$$

- It’s a linear program!
- Many efficient algorithms exist.

# “Given” an MDP, find an optimal policy.

- What does “given” mean?
- [Stronger] “Given” means you can sample any state, query any action, and observe the outcome.



- [Weaker] “Given” means you can play out a policy and observe the trajectory with no real-world cost.