# Section B

Wednesday, April 5, 2023 12:30 PM

The goal of RL is to find an <u>optimal policy</u>:

$$\pi^* = \underset{\pi}{\text{argmax}}\ V^{\pi}(s) \quad \forall s \in \mathcal{S}$$

<u>Value function</u>:

$$V^{\pi}(s_0) = \mathbb{E}\left[\text{total discounted reward for starting in state } s_0 \text{ and executing } \pi\right]$$

Given $\pi$, $p(s_{t+1}|s_t, a_t)$ there exists a distribution over <u>state trajectories</u>

$s_0 \xrightarrow{a_0} s_1 \xrightarrow{a_1} s_2 \xrightarrow{a_2} s_3 \xrightarrow{a_3} \cdots$ (because $p(s'|s,a)$ is stochastic)

$$= \mathbb{E}_{p(s'|s,a)}\left[R(s_0,a_0) + \gamma R(s_1,a_1) + \gamma^2 R(s_2,a_2) + \cdots \mid s_0\right]$$

$$r_0 + \gamma r_1 + \gamma^2 r_2 + \cdots$$

$$= R(s_0,a_0) + \gamma\, \mathbb{E}_{p(s'|s,a)}\Big[\underbrace{R(s_1,a_1)}_{f(s_1)} + \underbrace{\gamma R(s_2,a_2) + \gamma^2 R(s_3,a_3) + \cdots}_{g(s_2, s_3, \cdots)} \mid s_0\Big]$$

$$= R(s_0,a_0) + \gamma \sum_{s_1 \in \mathcal{S}} p(s_1|s_0,a_0)\underbrace{\Big(R(s_1,a_1) + \gamma\, \mathbb{E}_{p(s'|s,a)}\left[R(s_2,a_2) + \gamma R(s_3,a_3) + \cdots \mid s_0\right]\Big)}_{V^{\pi}(s_1)}$$

$$= R(s_0,a_0) + \gamma \sum_{s_1 \in \mathcal{S}} p(s_1|s_0,a_0)\, V^{\pi}(s_1)$$

Aside: $\mathbb{E}_{x,y \sim p(x,y)}\left[f(x) + g(y)\right] =$

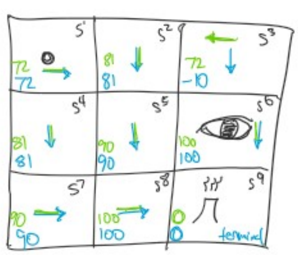
$$= \sum_x \sum_y p(x)p(y|x)\left[f(x) + g(y)\right]$$

$$= \sum_x p(x)\left(f(x) + \mathbb{E}_{y \sim p(y|x)}\left[g(y)\right]\right)$$

<u>Bellman Equations</u>: $a$

$$V^{\pi}(s) = R(s, \pi(s)) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, \pi(s))\, V^{\pi}(s')$$

**Ex: Value functions and policies**




Assume $\gamma = 0.9$

$A = \{\leftarrow, \downarrow, \uparrow, \rightarrow\}$

$R(s,a) = +100$ if entering $\Pi$

$R(s,a) = -100$ if entering 👁

$R(s,a) = 0$

Transitions are deterministic, $\delta(s,a)$

$$V^{\pi}(s) = R(s,\pi(s)) + \gamma V^{\pi}(\delta(s,\pi(s)))$$

$\leftarrow$ some policy $\pi'$

$$V^{\pi'}(s^3) = -100 + (0.9)100 = -10$$

$$V^*(s) = \max_{a \in A} R(s,a) + \gamma V^*(\delta(s,a))$$

$\leftarrow$ optimal policy $\pi^*$

$$V^*(s^3) = \max\left(R(s^3,\downarrow) + 0.9\, V^*(s^6),\ R(s^3,\leftarrow) + 0.9\, V^*(s^2)\right) = 72$$

- <u>Optimal value function</u> $V^* = V^{\pi^*}$

- Given $V^*, R(s,a), p(s'|s,a), \gamma$ we can compute $\pi^*$!

$$\pi^*(s) = \underset{a \in A}{\text{argmax}}\ \underbrace{R(s,a)}_{\text{immediate reward}} + \underbrace{\gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) V^*(s')}_{\text{future discounted reward}}$$

☹ Problem: our definitions of $\pi^*$ and $V^*$ are cyclic ☹

- Can compute $V^*$ without $\pi^*$

$$V^*(s) = \max_{a \in A} R(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) V^*(s')$$

$\uparrow$ ... definition of optimal value function

$$V^*(s) = \max_{a \in A} R(s,a) + \gamma \sum_{s' \in S} p(s' \mid s,a) V^*(s')$$

★ recursive definition of optimal value function

system of $|S|$ equations and $|S|$ variables

each variable is $V^*(s)$ for some $s \in S$

**Key Idea of value iteration:**

- Apply dynamic programming, i.e. fixed point iteration, to the recursive def. of $V^*$