POLICY, VALUE, AND BELLMAN EXERCISES

RECAP

SOLVING MDPS

- Defined problem space, i.e., an MDP $M = (\mathcal{S}, \mathcal{A}, \mathcal{R}, p, d_0, \gamma)$
- Defined solution space, i.e., $\pi \in \Pi$

  $$\pi(a \mid s) \doteq \Pr(A_t = a \mid S_t = s)$$

- Defined ways to evaluate policies

  $$v_\pi(s) \doteq \mathbb{E}[G_t \mid S_t = s], \; q_\pi(s, a) \doteq \mathbb{E}[G_t \mid S_t = s, A_t = a]$$

- Define Optimal

  $$\pi \in \Pi_* \text{ if } \forall s, \; v_\pi(s) = v_*(s) \doteq \max_{\pi' \in \Pi} v_{\pi'}(s)$$

RECAP

VALUE FUNCTIONS AND BELLMAN

- $v_\pi(s) \doteq \mathbb{E}[G_t \,|\, S_t = s]$ Requires considering all possible futures (intractable)

Bellman Equation — Express value using only immediate reward and value from the next state

$$v_\pi(s) = \sum_a \pi(a \,|\, s) \left(r(s,a) + \gamma \sum_{s'} p(s,a,s') v_\pi(s') \right)$$

$$q_\pi(s,a) = r(s,a) + \gamma \sum_{s'} p(s,a,s') v_\pi(s')$$

RECAP

BELLMAN OPTIMALITY

- Express Optimality without considering the policy

$$v_*(s) = \max_a r(s,a) + \gamma \sum_{s'} p(s,a,s') v_*(s') = \max_a q_*(s,a)$$

if $\pi(s) \in \arg\max_a r(s,a) + \gamma \sum_{s'} p(s,a,s') v_*(s')$ or

$$\pi(s) \in \arg\max_a q_*(s,a)$$

then $\pi \in \Pi_*$

NEXT

COMPUTING VALUES AND FINDING OPTIMAL POLICIES

- Compute value function iteratively
- Methods for searching for π_* and v_* if we know p

OPTIMALITY AND CHANGES IN REWARD

SCALING BY A CONSTANT

For some $\alpha > 0$ let $R'_t = \alpha R_t$ with $G'_t = \sum_{k=0}^{\infty} \gamma^k R'_{t+1+k}$, and $q'_\pi(s,a) = \mathbb{E}[G'_t \,|\, S_t = s, A_t = a]$

Prove that $\forall \alpha > 0$ the optimal policy remains unchanged, i.e.,

$$\Pi_* \cap \Pi'_* = \Pi_* = \Pi'_*$$

Hint: show

$$\arg\max_a q'_*(s,a) = \arg\max_a q_*(s,a)$$

OPTIMALITY AND CHANGES IN REWARD

ADDING A CONSTANT TO CONTINUING MDPS

For some constant $c \in \mathbb{R}$ let $R'_t = R_t + c$

Prove that for continuing problems (infinite-length episodes) and $\forall c \in \mathbb{R}$ the optimal policies are unchanged.

OPTIMALITY AND CHANGES IN REWARD

ADDING A CONSTANT TO EPISODIC MDPS

For some constant $c \in \mathbb{R}$ let $R'_t = R_t + c$

Prove that for episodic problems (infinite-length episodes) adding a constant $c \in \mathbb{R}$ **can** change the optimal policy.

Hint: construct an MDP to show that the optimal policy changes when adding c

NEXT CLASS

WHAT YOU SHOULD DO

1. Quiz due tonight: Value Functions and Bellman Equations 2
2. Watch the next week’s videos before Friday’s class

Friday: Dynamic Programming