AIL 722: Reinforcement Learning

Lecture 9: Discounting and Policy Evaluation

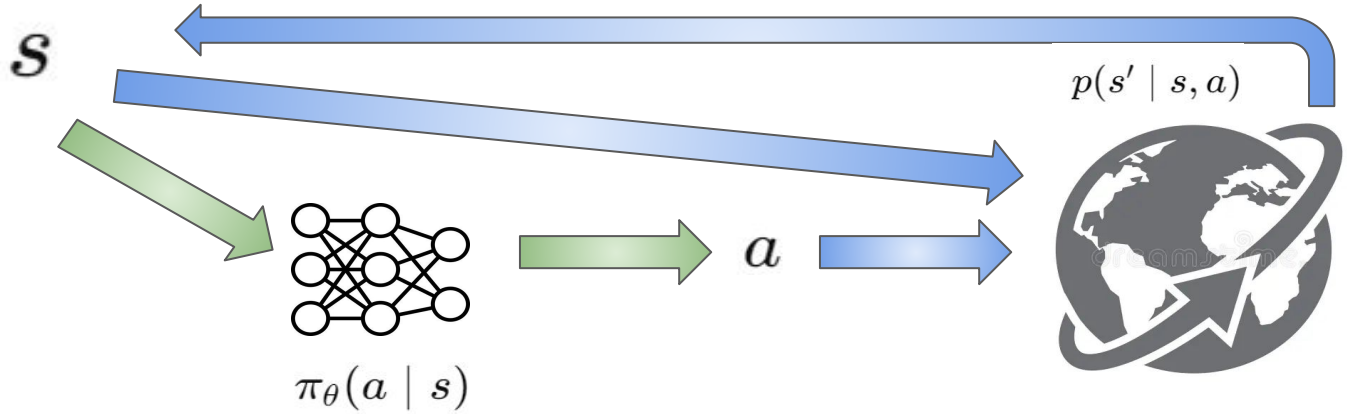
Raunak Bhattacharyya

ScAI YARDI SCHOOL OF ARTIFICIAL INTELLIGENCE
INDIAN INSTITUTE OF TECHNOLOGY DELHI

Outline

- Discounting
- Policy Evaluation
- Policy Iteration
- Assignment 1 Overview

Story So Far



$$\theta^* = \arg\max_\theta \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\sum_{t=1}^{T} r(s_t, a_t) \right]$$

Perspective

$$\theta^* = \arg\max_\theta \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\sum_{t=1}^{T} r(s_t, a_t) \right]$$

$$J(\theta) = \mathbb{E}_{p_\theta(\tau)} \left[\sum_{t=1}^{T} r(s_t, a_t) \right]$$

$$J(\theta) = \mathbb{E}_{p(s_1)} \left[V^\pi(s_1) \right] \qquad V^\pi(s_t) = \mathbb{E} \left[\sum_{t'=t}^{T} r(s_{t'}, a_{t'}) \middle| s_t \right]$$

Policy Iteration

$$\theta^* = \arg\max_\theta \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\sum_{t=1}^{T} r(s_t, a_t) \right]$$

1. Evaluate $V^\pi(s_t)$
2. Set $\pi \longleftarrow \pi_{\text{new}}$

$$\pi_{\text{new}} = \begin{cases} 1 & \text{if } a_t = \arg\max_{a_t} A^\pi(s_t, a_t) \\ 0 & \text{otherwise} \end{cases}$$

Value Function Recurrence

$$V^{\pi}(s_t) = \mathbb{E}\left[\sum_{t'=t}^{T} r(s_{t'}, a_{t'}) \middle| s_t\right]$$

$$V^{\pi}(s_t) = \mathbb{E}\left[r(s_t, a_t) + \sum_{t'=t+1}^{T} r(s_{t'}, a_{t'}) \middle| s_t\right]$$

$$V^{\pi}(s_t) = \mathbb{E}_{a_t \sim \pi_\theta(a_t|s_t)}\left[r(s_t, a_t)\right] + \mathbb{E}\left[\sum_{t'=t+1}^{T} r(s_{t'}, a_{t'})\right]$$

$$V^{\pi}(s_t) = \mathbb{E}_{a_t \sim \pi_\theta(a_t|s_t)}\left[r(s_t, a_t)\right] + \mathbb{E}_{s_{t+1} \sim p(s_{t+1}|s_t, a_t)}\left[V^{\pi}(s_{t+1})\right]$$

The Bellman equation

Deterministic Policy

$$\pi_{\text{new}} = \begin{cases} 1 & \text{if } a_t = \arg\max_{a_t} A^{\pi}(s_t, a_t) \\ 0 & \text{otherwise} \end{cases}$$

1. Evaluate $V^{\pi}(s_t)$
2. Set $\pi \longleftarrow \pi_{\text{new}}$

We are working with deterministic policies

$$V^{\pi}(s_t) = \mathbb{E}_{a_t \sim \pi_\theta(a_t|s_t)}\Big[r(s_t, a_t)\Big] + \mathbb{E}_{s_{t+1} \sim p(s_{t+1}|s_t, a_t)}\Big[V^{\pi}(s_{t+1})\Big]$$

$$V^{\pi}(s_t) = r(s_t, \pi(s_t)) + \mathbb{E}_{p(s_{t+1}|s_t, a_t)}\Big[V^{\pi}(s_{t+1})\Big]$$

Discounting

Episodic

Source: Youtube

Infinite horizon

Source: Youtube

How do we write the objective including discounting?

Discount factor: $\gamma \in [0, 1]$

Infinite horizon: $\gamma \in [0, 1)$

Value Function Recurrence: With Discount

Discount factor determines the present value of future rewards

Better to get rewards sooner rather than later

$$V^{\pi}(s_t) = r(s_t, \pi(s_t)) + \gamma \cdot \mathbb{E}_{p(s_{t+1}|s_t,a_t)}\left[V^{\pi}(s_{t+1})\right]$$

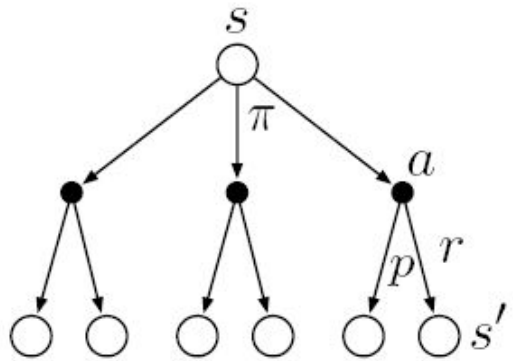
Policy Evaluation

$$V^{\pi}(s_t) = r(s_t, \pi(s_t)) + \gamma \cdot \mathbb{E}_{p(s_{t+1}|s_t,a_t)}\left[V^{\pi}(s_{t+1})\right]$$

Can we view this as a system of equations?

Policy Evaluation: Iterative Approach

$$V^{\pi}(s_t) = r(s_t, \pi(s_t)) + \gamma \cdot \mathbb{E}_{p(s_{t+1}|s_t,a_t)}\left[V^{\pi}(s_{t+1})\right]$$



Backup diagram for v_π

A state may be its own successor

From Reinforcement Learning: An Introduction, Sutton & Barto

Iterative Policy Evaluation

Iterative Policy Evaluation, for estimating $V \approx v_\pi$

Input π, the policy to be evaluated
Algorithm parameter: a small threshold $\theta > 0$ determining accuracy of estimation
Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(terminal) = 0$

Loop:
  $\Delta \leftarrow 0$
  Loop for each $s \in \mathcal{S}$:
    $v \leftarrow V(s)$
    $V(s) \leftarrow \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)\left[r + \gamma V(s')\right]$
    $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
until $\Delta < \theta$

From Reinforcement Learning: An Introduction, Sutton & Barto