AIL 722: Reinforcement Learning

Lecture 7: Value Functions

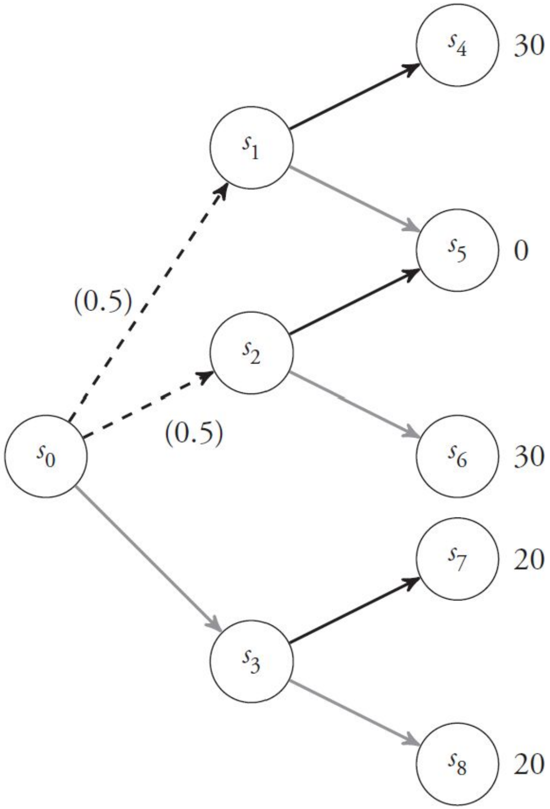
Raunak Bhattacharyya

ScAI YARDI SCHOOL OF ARTIFICIAL INTELLIGENCE
INDIAN INSTITUTE OF TECHNOLOGY DELHI

Outline

- Policy: Recap
- Goal of the agent: Objective Function
- Dealing with Expectations
- Value Functions

Open Loop Plan



$$
\begin{aligned}
U(\text{up}, \text{up}) &= 0.5 \times 30 + 0.5 \times 0 = 15 \\
U(\text{up}, \text{down}) &= 0.5 \times 0 + 0.5 \times 30 = 15 \\
U(\text{down}, \text{up}) &= 20 \\
U(\text{down}, \text{down}) &= 20
\end{aligned}
$$

Open loop plan chooses a down action from s_0

From Decision Making under Uncertainty, Mykel Kochenderfer

Policy

Search problem: path (sequence of actions)

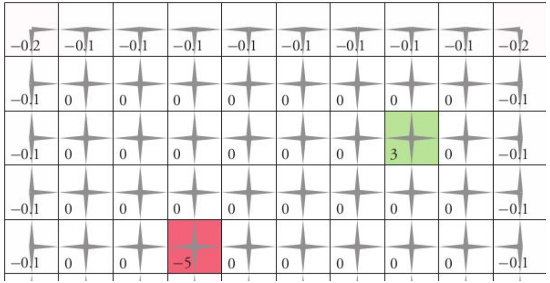
MDP:

Definition: policy

A **policy** π is a mapping from each state $s \in \text{States}$ to an action $a \in \text{Actions}(s)$.

An example policy:

- (1,1): right
- (4,7): left
- (6,2): up
- … (have to map every state to an action)



MDP

$$\text{MDP : Tuple}\langle \mathcal{S}, \mathcal{A}, T, R, \rho \rangle$$

$$\mathcal{S} : \text{State Space}$$

$$\mathcal{A} : \text{Action Space}$$

$$T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1] : \text{Probabilistic Transition Function}$$

$$R : \mathcal{S} \times \mathcal{A} \to \mathbb{R} : \text{Reward Function}$$

$$\rho : \text{Initial State Distribution}$$

$$\boxed{\pi : \mathcal{S} \to \mathcal{A}}$$

The Objective

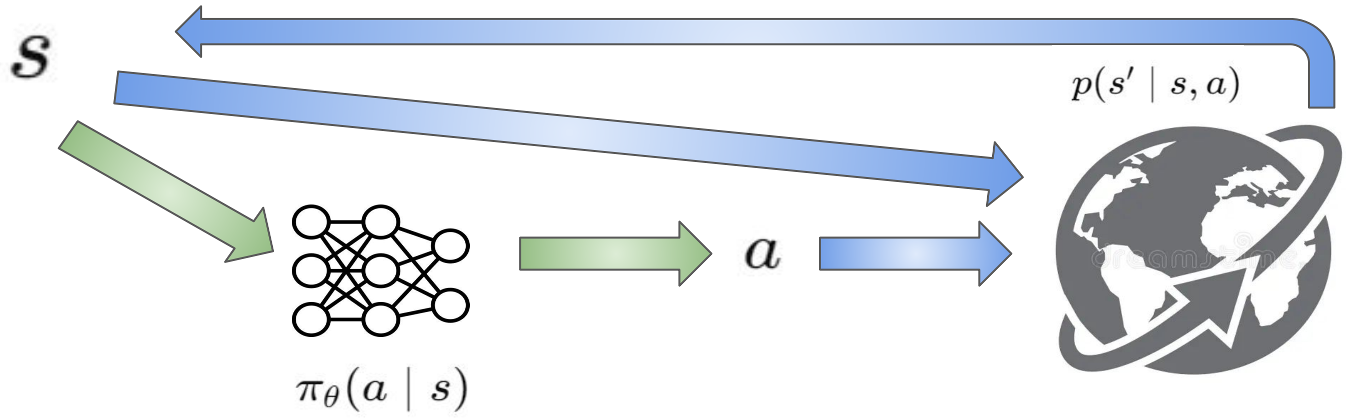
- The reward hypothesis

  That all of what we mean by goals and purposes can be well thought of as the maximization of the expected value of the cumulative sum of a received scalar signal (called reward).

  - Any problem of learning goal-directed behavior can be reduced to three signals passing back and forth between an agent and its environment:
    - Represent choices made by the agent (the actions)
    - Represent basis on which choices are made (the states)
    - Define the agent's goal (the rewards)

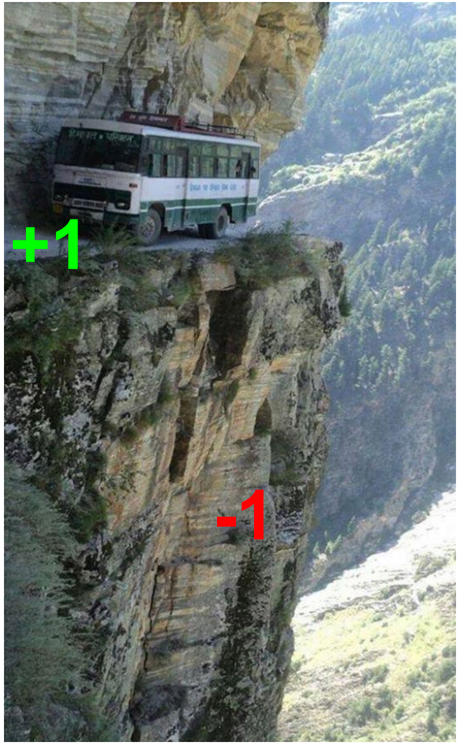
From Reinforcement Learning: An Introduction, Sutton & Barto

Objective



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

Expectations



Source: Pinterest

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

RL is really about optimising expectations

$r(s_t, a_t)$: not smooth

Suppose policy $\pi_\theta(a_t = \text{fall}) = \theta$

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

Why RL can use smooth optimisation techniques even though rewards are highly discontinuous

Adapted from Berkeley CS 285, Instructor: Sergey Levine

Working with Expectations

Expectations in the Objective

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

Expanding it out for clarity

$$J(\theta) = \mathbb{E}_{(s_1, a_1, s_2, a_2, \ldots, s_T, a_T) \sim p_\theta(s_1, a_1, \ldots, s_T, a_T)} \left[\sum_{t=1}^{T} r(s_t, a_t) \right]$$

Factorising the Trajectory Distribution

$$p_\theta(s_1, a_1, \ldots, s_T, a_T) = p(s_1) \prod_{t=1}^{T} \pi_\theta(a_t \mid s_t) \, p(s_{t+1} \mid s_t, a_t)$$

$$\begin{aligned} p(s_1, a_1, s_2, a_2, s_3) &= p(s_1) \cdot p(a_1, s_2, a_2, s_3 \mid s_1) \\ &= p(s_1) \cdot p(a_1 \mid s_1) \cdot p(s_2, a_2, s_3 \mid s_1, a_1) \\ &= p(s_1) \cdot p(a_1 \mid s_1) \cdot p(s_2 \mid s_1, a_1) \cdot p(a_2, s_3 \mid s_1, a_1, s_2) \\ &= p(s_1) \cdot p(a_1 \mid s_1) \cdot p(s_2 \mid s_1, a_1) \cdot p(a_2 \mid s_2) \cdot p(s_3 \mid s_2, a_2) \end{aligned}$$

Can we use this factorization in the objective function?

Conditional Expectations

$$J(\theta) = \mathbb{E}_{(s1,a1,s2,a2,\ldots,sT,aT)\sim p_\theta(s_1,a_1,\ldots,s_T,a_T)}\left[\sum_{t=1}^{T} r(s_t, a_t)\right]$$

$$J(\theta) = \mathbb{E}_{s_1\sim p(s_1)}\left[\mathbb{E}_{a_1\sim\pi_\theta(a_1|s_1)}\left[r(s_1, a_1) + \mathbb{E}_{s_2\sim p(s_2|s_1,a_1)}\left[\mathbb{E}_{a_2\sim\pi_\theta(a_2|s_2)}\left[r(s_2, a_2) + \cdots \mid s_2\right] \mid s_1, a_1\right] \mid s_1\right]\right]$$

Introducing the Q-function

$$J(\theta) = E_{s_1 \sim p(s_1)}\left[E_{a_1 \sim \pi_\theta(a_1|s_1)}\left[r(s_1, a_1) + E_{s_2 \sim p(s_2|s_1,a_1)}\left[E_{a_2 \sim \pi_\theta(a_2|s_2)}\left[r(s_2, a_2) + \cdots \mid s_2\right] \mid s_1, a_1\right] \mid s_1\right]\right]$$

Suppose we knew this part

Value Functions

Definition: Q-function

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

Expected cumulative reward obtained by taking a_t in s_t and then following the policy

What is the expectation over?

What is the objective in terms of Q?

$$J(\theta) = \mathbb{E}_{s_1 \sim p(s_1)}\left[\mathbb{E}_{a_1 \sim \pi_\theta(a_1|s_1)}\left[Q(s_1, a_1) \middle| s_1\right]\right]$$

Definition: Value Function (V)

$$J(\theta) = \mathbb{E}_{s_1 \sim p(s_1)}\left[\mathbb{E}_{a_1 \sim \pi_\theta(a_1|s_1)}\left[Q(s_1, a_1) \,\middle|\, s_1\right]\right]$$

$$V^\pi(s_t) = \mathbb{E}_{a_t \sim \pi_\theta(a_t|s_t)}\left[Q^\pi(s_t, a_t) \middle| s_t\right]$$

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

Expected cumulative reward obtained by taking by following the policy starting from s_t

What is the RL objective in terms of V?

$$J(\theta) = \mathbb{E}_{s_1 \sim p(s_1)}\left[V^\pi(s_1)\right]$$

Approaches Using Value Functions

If we have a policy, and we know its corresponding Q-function, we can improve the policy

Compute the gradient and do gradient ascent to increase the probability of good actions

$$\text{Set } \pi'(a \mid s) = 1$$

$$\text{if } a = \arg\max_a Q^\pi(s, a)$$

$$\text{If } Q^\pi(s, a) > V^\pi(s)$$

$$\text{modify } \pi(a \mid s) \text{ to increase probability of } a$$