AIL 722: Reinforcement Learning

Lecture 8: Value Functions (Part 2)

Raunak Bhattacharyya

ScAI YARDI SCHOOL OF ARTIFICIAL INTELLIGENCE
INDIAN INSTITUTE OF TECHNOLOGY DELHI

Outline

- Value functions

- The advantage function

- Policy Iteration

- Policy Evaluation

Questions & Clarifications

- What do trajectories mean, what does distribution over trajectories mean, and what is their role in the objective function
- Example of Q and V
- Factorising the trajectory distribution
- Tree view of Q and V

About Trajectories

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

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

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

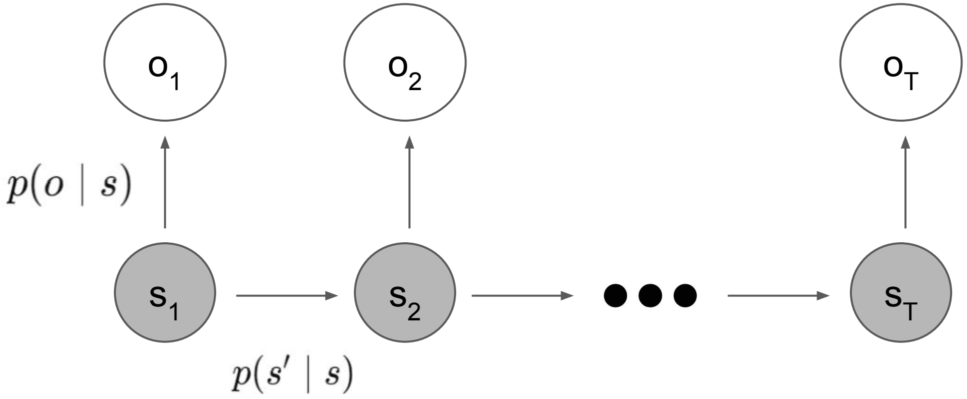
Can we visualise this?

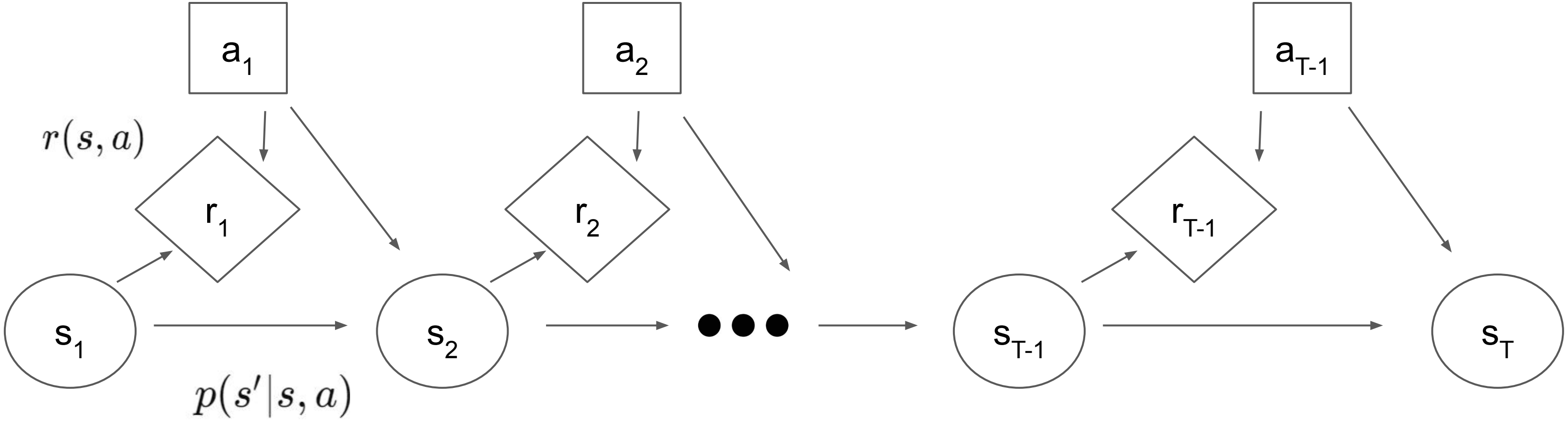
Factorising the Trajectory Distribution

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

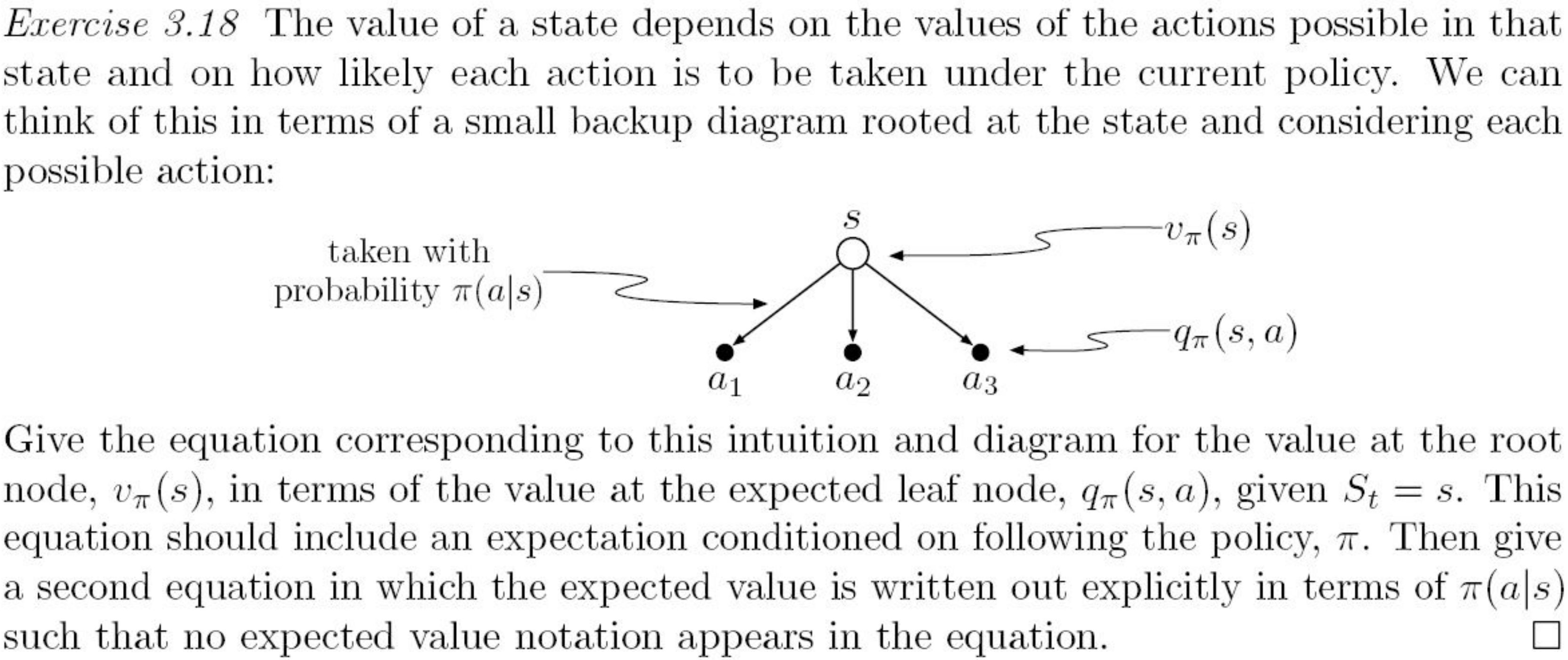
$$\begin{aligned} p(s_1, a_1, s_2, a_2, s_3) &= p(s_1) \cdot p(a_1, s_2, a_2, s_3 \mid s_1) \\ &= p(s_1) \cdot p(a_1 \mid s_1) \cdot p(s_2, a_2, s_3 \mid s_1, a_1) \\ &= p(s_1) \cdot p(a_1 \mid s_1) \cdot p(s_2 \mid s_1, a_1) \cdot p(a_2, s_3 \mid s_1, a_1, s_2) \\ &= p(s_1) \cdot p(a_1 \mid s_1) \cdot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

MDP: State Evolution





The Tree View



Exercise 3.18 The value of a state depends on the values of the actions possible in that state and on how likely each action is to be taken under the current policy. We can think of this in terms of a small backup diagram rooted at the state and considering each possible action:

s
$v_\pi(s)$
taken with
probability $\pi(a|s)$
$q_\pi(s, a)$
a_1
a_2
a_3

Give the equation corresponding to this intuition and diagram for the value at the root node, $v_\pi(s)$, in terms of the value at the expected leaf node, $q_\pi(s, a)$, given $S_t = s$. This equation should include an expectation conditioned on following the policy, π. Then give a second equation in which the expected value is written out explicitly in terms of $\pi(a|s)$ such that no expected value notation appears in the equation. □

From Reinforcement Learning: An Introduction, Sutton & Barto

Conditional Expectations and Q-function

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

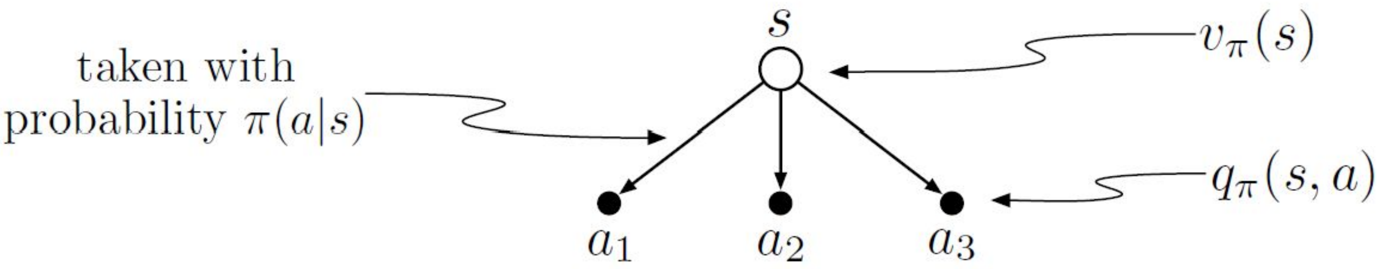
$$J(\theta) = \mathbb{E}_{(s_1,a_1,s_2,a_2,\ldots,s_T,a_T) \sim p_\theta(s_1,a_1,\ldots,s_T,a_T)} \left[\sum_{t=1}^{T} r(s_t, a_t) \right]$$

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

Suppose we knew this part

The Tree View

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

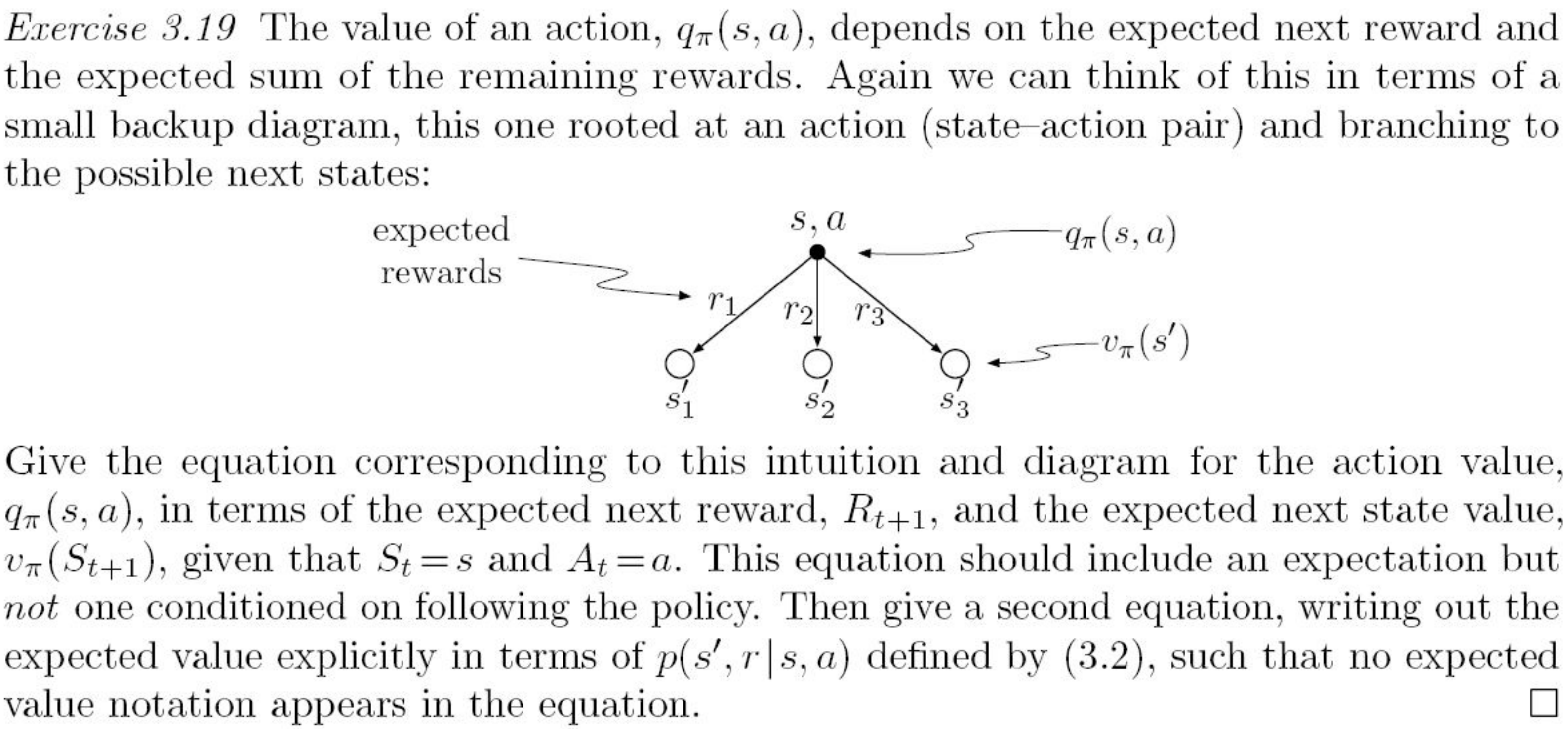


$$V^\pi(s_t) = \mathbb{E}_{a_t \sim \pi_\theta(a_t|s_t)}\left[Q^\pi(s_t, a_t) \middle| s_t\right]$$

Expand the expectation?

From Reinforcement Learning: An Introduction, Sutton & Barto

The Tree View: With Q



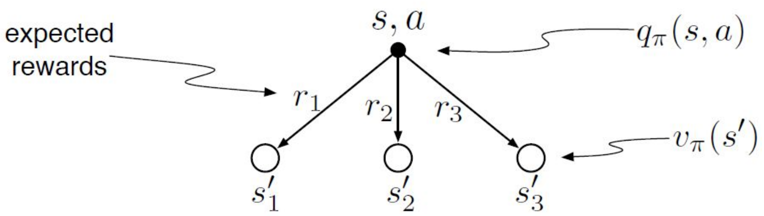
Exercise 3.19 The value of an action, $q_\pi(s,a)$, depends on the expected next reward and the expected sum of the remaining rewards. Again we can think of this in terms of a small backup diagram, this one rooted at an action (state–action pair) and branching to the possible next states:

expected
rewards
s, a
$q_\pi(s,a)$
r_1
r_2
r_3
$v_\pi(s')$
s'_1
s'_2
s'_3

Give the equation corresponding to this intuition and diagram for the action value, $q_\pi(s,a)$, in terms of the expected next reward, R_{t+1}, and the expected next state value, $v_\pi(S_{t+1})$, given that $S_t = s$ and $A_t = a$. This equation should include an expectation but *not* one conditioned on following the policy. Then give a second equation, writing out the expected value explicitly in terms of $p(s', r \,|\, s, a)$ defined by (3.2), such that no expected value notation appears in the equation. □

From Reinforcement Learning: An Introduction, Sutton & Barto

The Tree View: With Q



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

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

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

Why is this true?

From Reinforcement Learning: An Introduction, Sutton & Barto

Reminder: Value function

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

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

Which Action to Pick at t=1?

Suppose I gave you $Q^{\pi}(s_1, a_1)$.

How would you find the action to take at $t = 1$?

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

Algorithms

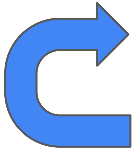
Definition: Advantage Function

$$A^\pi(s_t, a_t) = Q^\pi(s_t, a_t) - V^\pi(s_t)$$

On average, how much better is the action a_t when taken at state s_t as compared to the average over all actions possible at s_t?

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

Policy Iteration



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

How do we do this?

$$A^{\pi}(s_t, a_t) = Q^{\pi}(s_t, a_t) - V^{\pi}(s_t)$$

Can we simplify this?

Advantage Function

$$A^\pi(s_t, a_t) = Q^\pi(s_t, a_t) - V^\pi(s_t)$$

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

What is s_{t+1} sampled from?

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

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

Thus, finding V is a route to find A

Policy Iteration

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

How do we do this?

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

Policy Evaluation

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

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

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

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

The Bellman equation