FPT Education

FPT UNIVERSITY

Understand Average Reward

FPT Education
FPT UNIVERSITY

Objectives

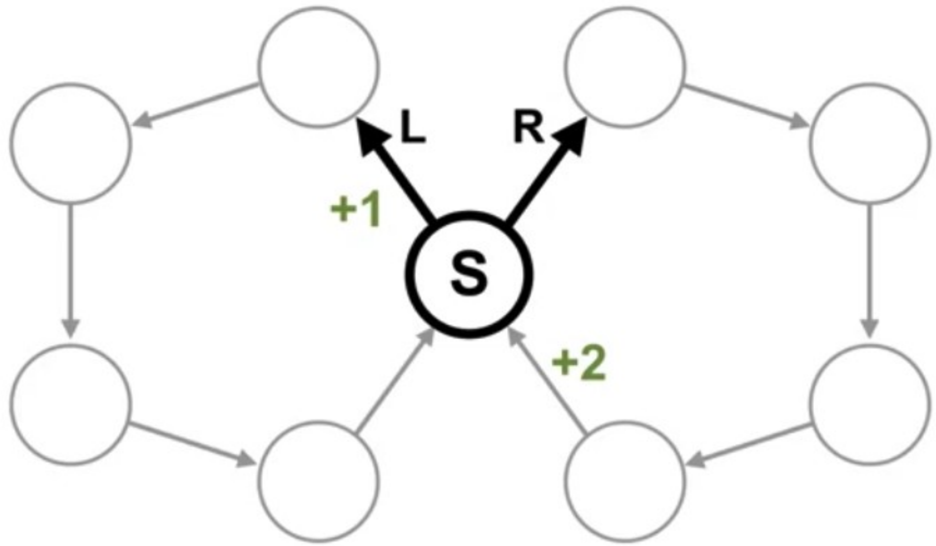
- ❑ Describe the average reward
- ❑ Understand differential value functions.

Average Reward

- Example:
  - In most states, there's a single action available, leaving no room for decisions.
  - However, in one specific state, the agent can choose between two actions: traversing the left or right ring.
  - The reward structure entails zero rewards except for specific transitions; for instance, in the left ring, there's a +1 reward immediately after state S, while in the right ring, there's a +2 reward just before state S, implying the intuitive choice of the right action for a higher reward.

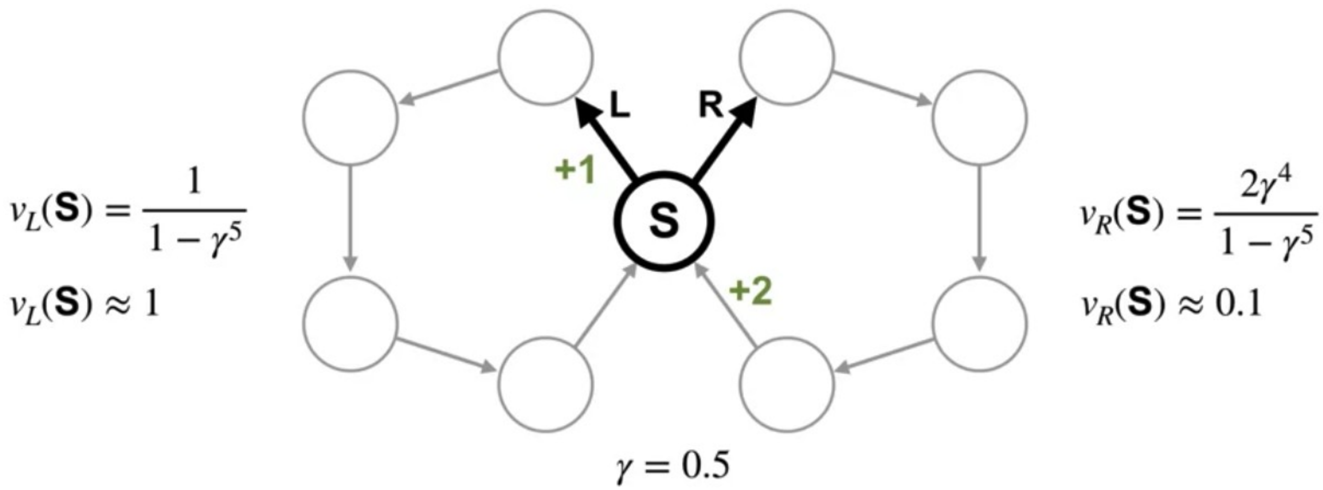
Average Reward

- Example:
  - What would you pick? Left or Right?



Average Reward

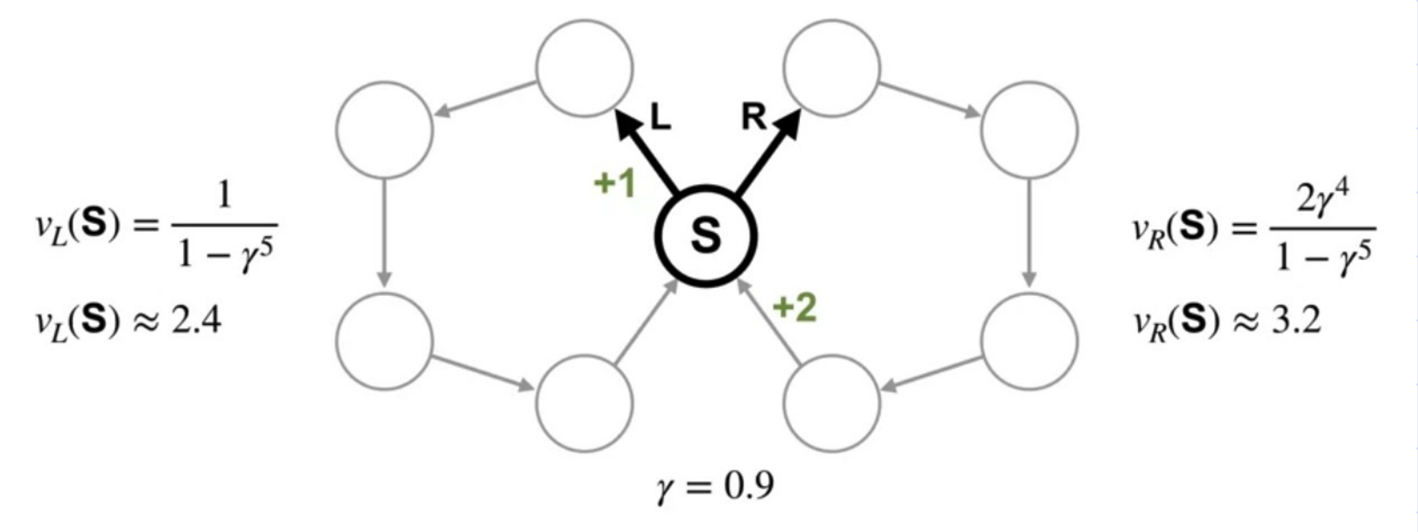
- Example:



- This means the policy that takes the left action is preferable under this more myopic discount.

Average Reward

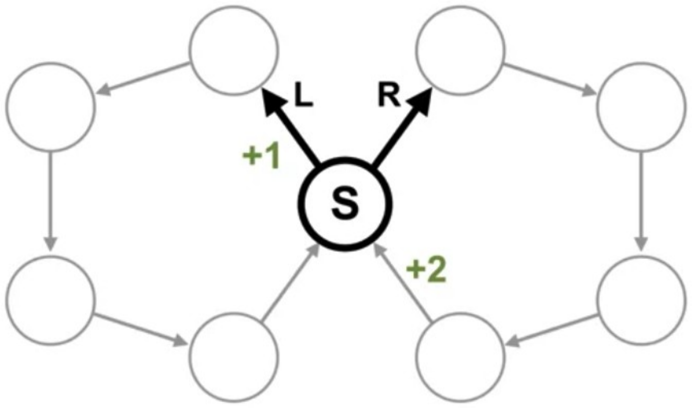
- Example:



- The agent prefers the policy that goes right

Average Reward

- Example:
  - we can figure out the minimum value of gamma so that the agent prefers the policy that goes right. Gamma needs to be at least 0.841.



$$v_L(\mathbf{S}) = \frac{1}{1-\gamma^5}$$

$$v_R(\mathbf{S}) = \frac{2\gamma^4}{1-\gamma^5}$$

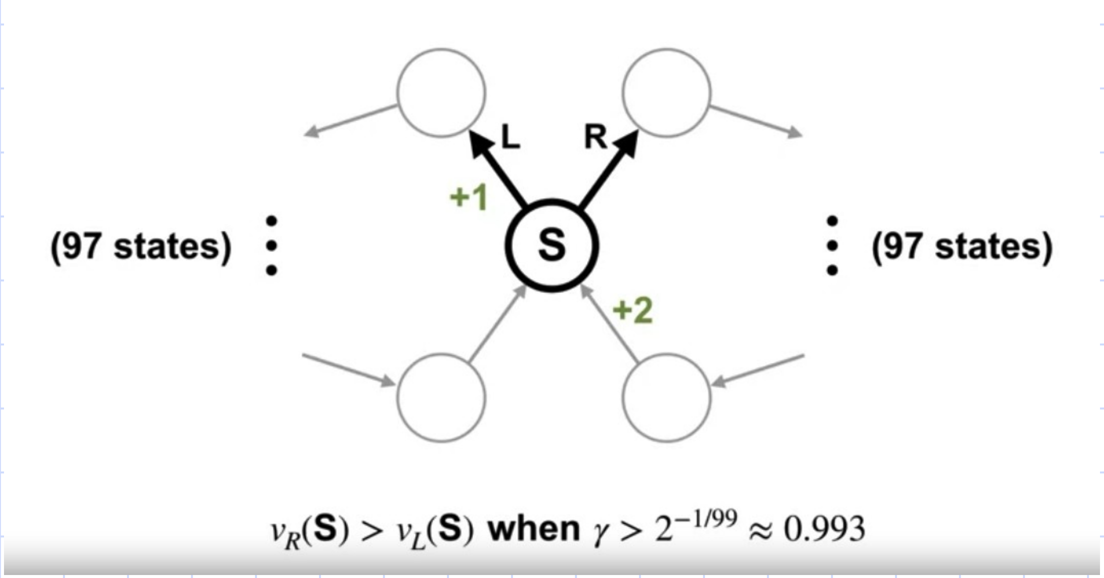
$$v_R(\mathbf{S}) > v_L(\mathbf{S}) \text{ when } \gamma > 2^{-1/4} \approx 0.841$$

Average Reward

- Example:
  - The only way to ensure that the agents actions maximize reward over time is to keep increasing the discount factor towards 1.
  - Depending on the problem, we might need gamma to be quite large.
  - We can't set it to 1 in a continuing setting because then the return might be infinite.

Average Reward

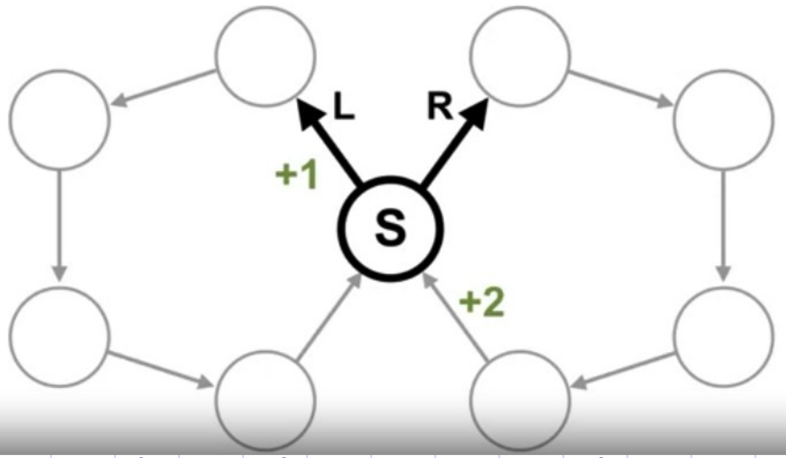
- Example:
  - What's wrong with having larger gamma? -->Larger values of gamma can also result in larger and more variables sums, which might be difficult to learn.



Average Reward

- Example
  - Imagine the agent has interacted with the world for H steps.
  - This is the reward it has received on average across those H steps.
  - It's rate of re

$$r(\pi) \doteq \lim_{h\to\infty} \frac{1}{h}\sum_{t=1}^{h} \mathbb{E}[R_t \mid S_0, A_{0:t-1} \sim \pi]$$



Average Reward

- Example:
  - If the agents goal is to maximize this average reward, then it cares equally about nearby and distant rewards.
  - We denote the average reward of a policy with R pi.

$$r(\pi) = \sum_{s} \mu_{\pi}(s) \sum_{a} \pi(a \mid s) \sum_{s',r} p(s', r \mid s, a) r$$

Average Reward

- Example
  - We can write the average reward using the state visitation, mu.
  - This inner term is the expected reward in a state under policy pi.
  - The outer sum takes the expectation over how frequently the po

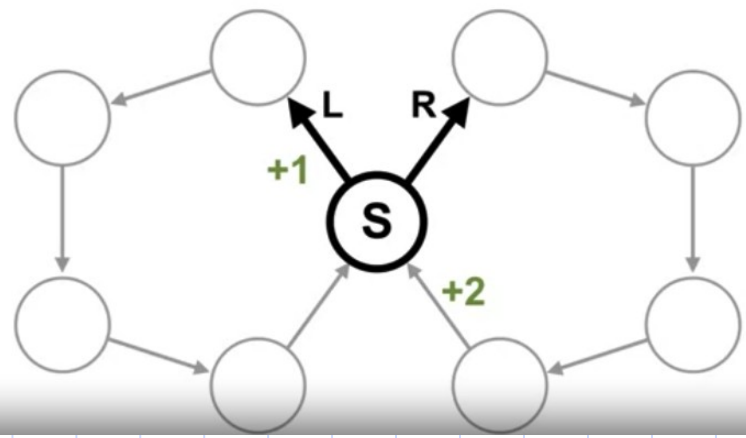
$$r(\pi) = \sum_{s} \mu_{\pi}(s) \sum_{a} \pi(a \mid s) \sum_{s',r} p(s', r \mid s, a) r$$

Average Reward

- Example:
  - The average reward puts preference on the policy that receives more reward in total without having to consider larger and larger discounts.

$$r(\pi) = \sum_{s} \mu_{\pi}(s) \sum_{a} \pi(a \mid s) \sum_{s',r} p(s', r \mid s, a) r$$

$$r(\pi_L) = 1/5 = 0.2$$



$$r(\pi_R) = 2/5 = 0.4$$

Average Reward

- Average reward is a formulation in reinforcement learning that focuses on maximizing the long-term average reward obtained by an agent interacting with an environment.
  - In traditional reinforcement learning formulations: the objective is to maximize the expected cumulative discounted reward over time.
  - Particularly those involving continuous or episodic tasks with unknown episode lengths, the average reward formulation offers advantages.

FPT Education
FPT UNIVERSITY

Average Reward

- Objective:
  - In the average reward formulation, the objective is to maximize the long-term average reward per time step rather than maximizing the discounted cumulative reward.
  - The agent aims to find a policy that maximizes the average reward obtained over an infinite horizon or a sufficiently long time period.

Average Reward

- Mathematical Formulation:
  - The average reward formulation is typically expressed as maximizing the following objective:

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[R_t]$$

where:

T is the time horizon or the number of time steps.

Rt is the reward obtained at time step

E[·] denotes the expectation operator.

  - The objective is to find a policy that maximizes the long-term average of rewards obtained per time step.

Average Reward

- Advantages:
  - Simplicity: simplify the analysis and computation of optimal policies
  - Stationarity: the assumption of stationarity is often implicit, making it suitable for environments with unknown episode lengths or varying dynamics.
  - Performance Metrics: clear performance metric which can be directly optimized by reinforcement learning algorithms.

Average Reward

- Disadvantages:
  - Discounting Effects: not explicitly consider the time value of rewards or future consequences, potentially leading to suboptimal behavior in tasks where immediate rewards are prioritized over delayed rewards.
  - Convergence Challenges: Finding optimal policies in average reward formulations may present convergence challenges, particularly in environments with complex dynamics or non-stationarity.

Average Reward

- Value functions for Average Reward:
  - We define value functions, as the expected return

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

  - we define differential value functions as the expected differential return under a policy from a given state or stat

$$G_t = R_{t+1} - r(\pi) + R_{t+2} - r(\pi) + R_{t+3} - r(\pi) + \ldots$$

  - Dif ıan eq

$$q_\pi(s,a) = \sum_{s',r} p(s',r \mid s,a)\Big(r - r(\pi) + \sum_{a'} \pi(a' \mid s') q_\pi(s',a')\Big)$$

Differential Sarsa

- Differential semi-gradient Sarsa for estimating q hat

Differential semi-gradient Sarsa for estimating $\hat{q} \approx q_*$

Input: a differentiable action-value function parameterization $\hat{q} : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}$
Algorithm parameters: step sizes $\alpha, \beta > 0$
Initialize value-function weights $\mathbf{w} \in \mathbb{R}^d$ arbitrarily (e.g., $\mathbf{w} = \mathbf{0}$)
Initialize average reward estimate $\bar{R} \in \mathbb{R}$ arbitrarily (e.g., $\bar{R} = 0$)

Initialize state S, and action A
Loop for each step:
  Take action A, observe R, S'
  Choose A' as a function of $\hat{q}(S', \cdot, \mathbf{w})$ (e.g., ε-greedy)
  $\delta \leftarrow R \quad + \hat{q}(S', A', \mathbf{w}) - \hat{q}(S, A, \mathbf{w})$
  $\bar{R} \leftarrow \bar{R} + \beta(R - \bar{R})$
  $\mathbf{w} \leftarrow \mathbf{w} + \alpha\delta\nabla\hat{q}(S, A, \mathbf{w})$
  $S \leftarrow S'$
  $A \leftarrow A'$

Summary

- ❑ Describe the average reward
- ❑ Understand differential value functions.

FPT Education
FPT UNIVERSITY

Q & A