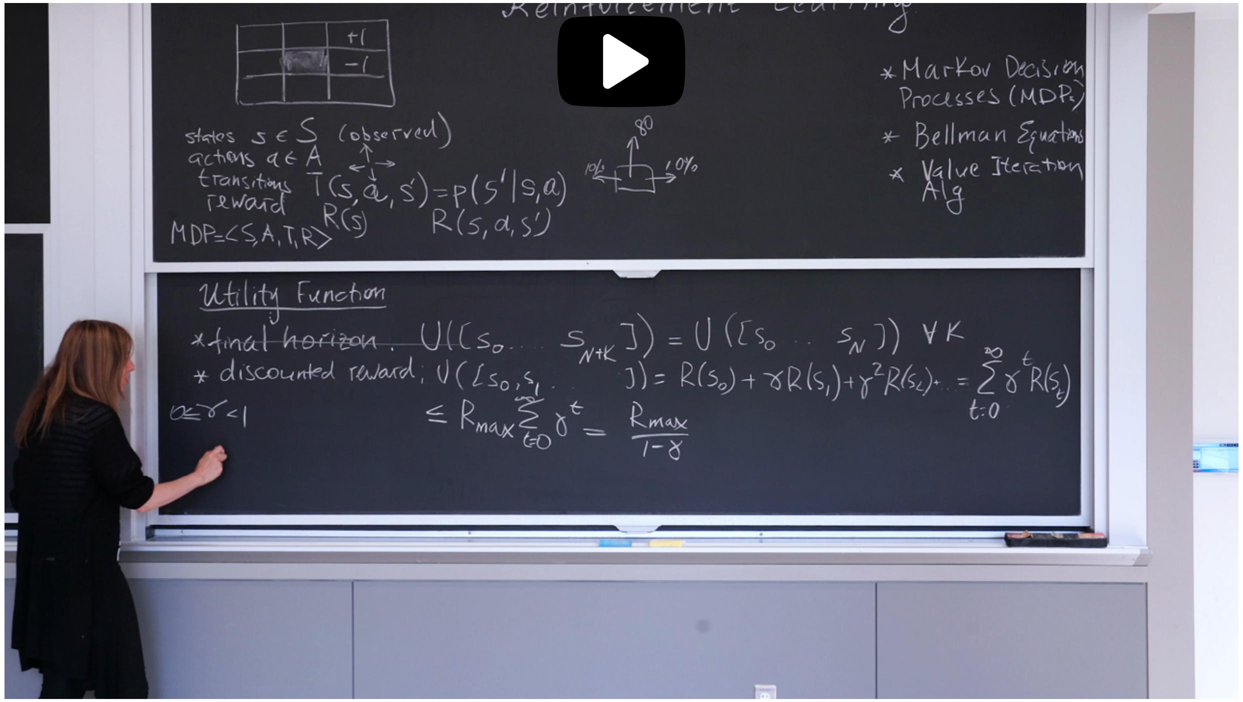
Course › Unit 5 Reinforcement Learning (2 weeks) › Lecture 17. Reinforcement Learning 1 › 5. Policy and Value Functions

5. Policy and Value Functions

Policy and Value Functions

Start of transcript. Skip to the end.

The next thing that I need to introduce to you before we can


+1
-1
states s ∈ S (observed)
actions a ∈ A
transitions T(s, a, s') = p(s'|s, a)
reward R(s) R(s, a, s')
MDP=<S,A,T,R>
80
10%
10%
* Markov Decision Processes (MDPs)
* Bellman Equations
* Value Iteration Alg
Utility Function
* ~~final horizon~~ U([s0 ... sN+K]) = U([s0 ... sN]) ∀K
* discounted reward; U([s0, s1 ...]) = R(s0) + γR(s1) + γ²R(s2) + ... = Σ γ^t R(st)
0 ≤ γ < 1
≤ Rmax Σ γ^t = Rmax / (1-γ)

move to the algorithm-- you can see we're doing a lot

of introduction, but this, I promise, is the last one--

we need to talk about policy, OK?

So what is policy?

Policy is an optimal action that you can take in a state.

OK?

0:00 / 0:00 1.50x

Video

Download video file

Transcripts

Download SubRip (.srt) file

Download Text (.txt) file

Definition of Optimal Policy

1/1 point (graded)

Which of the following options are correct about the optimal policy function:

○ Optimal Policy function would only depend on the state and action space but is independent of the reward structure

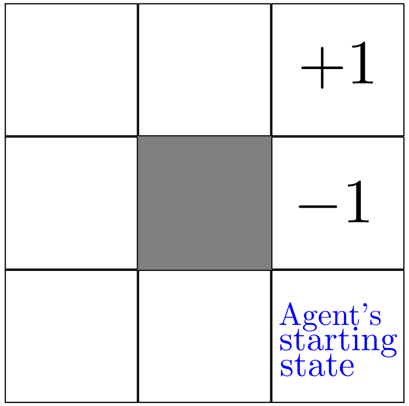
● Optimal Policy assigns an action at every state that maximizes the expected utility ✓

○ For any given state, the optimal Policy function should always take an action that results in the best expected immediate reward for that state

○ Policy function specifies an action for every different sequence of states starting from the initial state until a current state

Submit You have used 1 of 2 attempts

Consider the MDP example presented in the lecture: An AI agent is trying to navigate a 3x3 grid. It receives a reward of +1 for ending up in the top right corner and a reward of -1 for ending up in the cell immediately below it. It doesn't receive any reward at the other states as illustrated in the following figure.



Besides the reward for reaching one of the two shaded states in the figure from the text above, the agent also gets a reward of -10 for every action that it takes.

For the following set of problems, assume that all the transitions are deterministic. That is, there is only one initial state and all the actions are deterministic: All the future states are completely predictable.

For intance, taking the action “Left" from the centermost cell in the grid would always take the agent to the cell adjoining it towards the left. Any action pointing off the grid would lead it to remain in its current cell.

Also, assume that the agent always starts off from the bottom right corner of the grid, and it comes to a complete halt if and when it reaches the top right corner.

Optimal Policy - Numerical Example

1/2 points (graded)

What is the best discounted reward that an agent could accumulate starting from the bottom right corner of the grid for different values of the discount factor γ? Remember that the agent should keep on making actions until it reaches the top right corner.

If $\gamma = 0$:

-11

✗ **Answer:** -10

If $\gamma = 0.5$:

-15.5

✓ **Answer:** -15.5

Solution:

If $\gamma = 0$, then the discounted reward is the same as the reward received after the first step. The agent could receive a reward of -10 if it choose any one of the actions “Left", “Down", “Right". It would receive an additional reward of -1 if it selects to go up. So, the best discounted reward under this condition would be -10.

If $\gamma = 0.5$, then the best discounted reward would occur for the following sequence of actions starting from the initial state: “Up", “Up". It would recieve rewards of $-11, -9$ for these two actions respectively. The agent would reach the top right state and come to a complete halt after taking the above sequence of actions. The discounted reward amounts to $-11 + 0.5 * -9 = -15.5$

Submit You have used 3 of 3 attempts

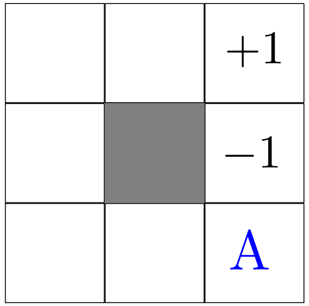
Answers are displayed within the problem

Value Function

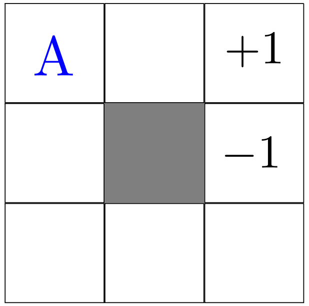
1/1 point (graded)

As above, we are working with the 3x3 grid with $+1$ reward at the top right corner and -1 at the cell below it. The agent also gets a reward of -10 for every action that it takes. We would like our agent to reach the $+1$ cell with minimum steps.

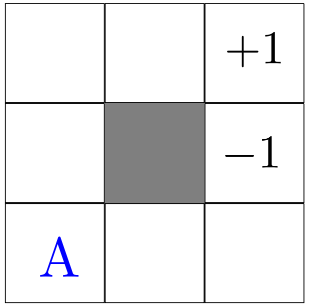
The following figures shows states s_1, s_2, s_3 for our grid example (A marks the cell of the current location of the agent).



s_1



s_2



s_3

Which of the following should hold true for a good value function $V(s)$ under the above reward structure?

○ $V(s_1) < V(s_1) < V(s_3)$

○ $V(s_3) << V(s_2) < V(s_1)$

● $V(s_3) < V(s_1) < V(s_2)$ ✓

Submit You have used 1 of 2 attempts

Discussion

Hide Discussion

Topic: Unit 5 Reinforcement Learning (2 weeks) :Lecture 17. Reinforcement Learning 1 / 5. Policy and Value Functions

Add a Post

Show all posts ▼ by recent activity ▼

? [Staff] Optimal Policy - Numerical Example 1 new_ 4

? Optimal Policy - Numerical Example: the accepted answer is inconsistent with the instructions
We are explicitly told "Remember that the agent should keep on making actions until it reaches the top right corner." 4

? Optimal Policy - Numerical Example: what are the two shaded states mentioned in the text? 2

? Value Function: s1 is repeated in one of the answers
. 1