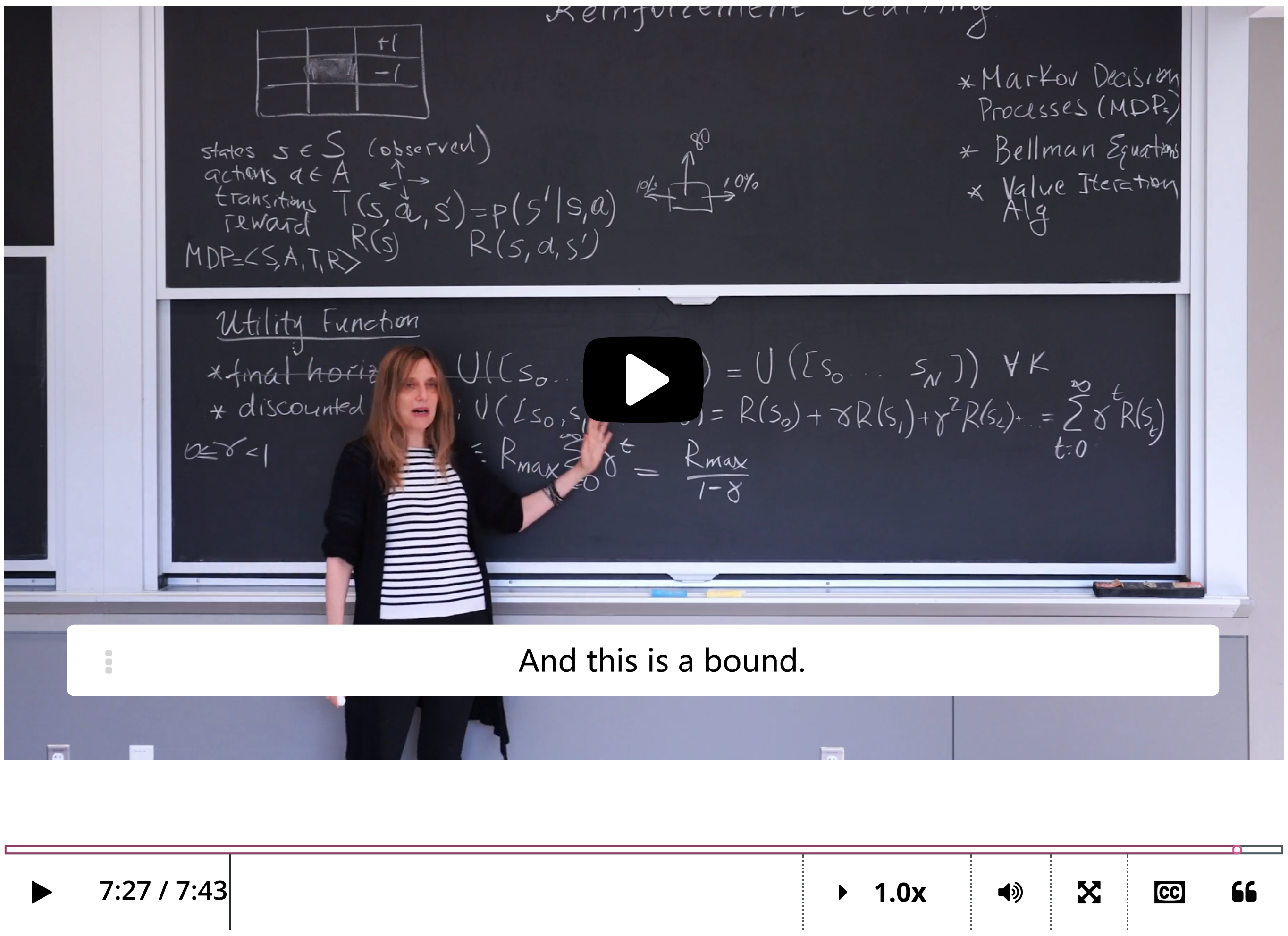
4. Utility Function

Video note: In the video below at 1:25, Prof Barzilay miswrote “final horizon" on the board. The correct term should be meant **finite horizon**.

Utility Function



is actually equal to 1 divided by 1 minus gamma.

So what we will get here is our max divided by 1 minus gamma.

And this is a bound.

And what you will see is that actually,

the fact that we're using the discount and rewards

would be essential for us to make our algorithms converge.

So this is a discounted reward that we

will be using for the duration.

End of transcript. Skip to the start.

Video

Download video file

Transcripts

Download SubRip (.srt) file

Download Text (.txt) file

Finite horizon vs Discounted reward

1/1 point (graded)

The main problem for MDPs is to optimize the agent's behavior. To do so, we first need to specify the criterion that we are trying to maximize in terms of accumulated rewards. We will define a **utility function** and maximize its expectation.

We consider two different types of utility functions:

1. **Finite horizon based utility** : The utility function is the sum of rewards after acting for a fixed number n steps. For example, in the case when the rewards depend only on the states, the utility function is

$$U\left[s_0, s_1, \ldots, s_n\right] = \sum_{i=0}^{n} R\left(s_i\right) \quad \text{for some fixed number of steps} n.$$

In particular $U\left[s_0, s_1, \ldots, s_{n+m}\right] = U\left[s_0, s_1, \ldots, s_n\right]$for any positive integer m.

2. **(Infinite horizon) discounted reward based utility** : In this setting, the reward one step into the future is discounted by a factor γ , the reward two steps ahead by γ^2, and so on. The goal is to continue acting (without an end) while maximizing the expected discounted reward. The discounting allows us to focus on near term rewards, and control this focus by changing γ. For example, if the rewards depend only on the states, the utility function is

$$U\left[s_0, s_1, \ldots\right] = \sum_{k=0}^{\infty} \gamma^k R\left(s_k\right).$$

How do these two types of utility function depend on the time steps?
(Choose all that apply.)

☑ The action at state s that maximizes a finite horizon based utility can depend on how many steps have been taken. ✓

☐ The action at state s that maximizes a finite horizon based utility does **not** depend on how many steps have been taken.

☑ The action at state s that maximizes a discount reward based utility does **not** depend on how many steps have been taken. ✓

☐ The action at state s that maximizes a discount reward based utility can depend on how many steps have been taken.

✓

Solution:

When using a finite horizon utility function, when computing the action at time step i, it can depend on the amount of steps that we have left until we reach N. For example, if we are at state s at time N, the agent will want to act greedily and take the action that leads to the immediate highest reward. However, if we are at time 0, the agent can allow to move towards areas with higher rewards while getting an immediate lower reward.

Discounted reward based utility under a markovian setting would lead to an optimal policy that only depends on the state and is independent of the step where the state occurs.

Submit

You have used 1 of 3 attempts

Answers are displayed within the problem

Discounted reward

1/1 point (graded)

Recall that the discounted reward in the case when $R\left(s, a, s'\right) = R\left(s\right)$is given by:

$$R\left(s_0\right) + \gamma R\left(s_1\right) + \gamma^2 R\left(s_2\right) \ldots = \sum_{t=0}^{\infty} \gamma^t R\left(s_t\right) \text{ where } 0 \leq \gamma < 1.$$

Which of the following is true about discounted reward?
(Choose all that apply.)

☑ For $\gamma = 0$, maximizing for discounted reward boils down to greedily maximizing for the immediate reward. ✓

☑ Discounted reward is guaranteed to be finite when the maximum reward is finite. ✓

☐ Discounted reward can be unbounded when the maximum reward is finite.

☐ Discounted reward converges to $R_{\text{min}} / \left(1 - \gamma\right)$ where R_{min} is the minimum reward possible from any state.

✓

Solution:

For $\gamma = 0$, the discounted reward is given by

$$R(s) + (0)\,R(s_1) + \ldots = R(s)\,,$$

which depends only on the reward for the current step.
When the maximum reward is finite, the discounted reward as derived in the lecture is bounded above by $R_{\max}/\,(1-\gamma)$.

Submit You have used 2 of 3 attempts

Answers are displayed within the problem

Discussion

Show Discussion

Topic: Unit 5 Reinforcement Learning (2 weeks) :Lecture 17. Reinforcement Learning 1 / 4. Utility Function