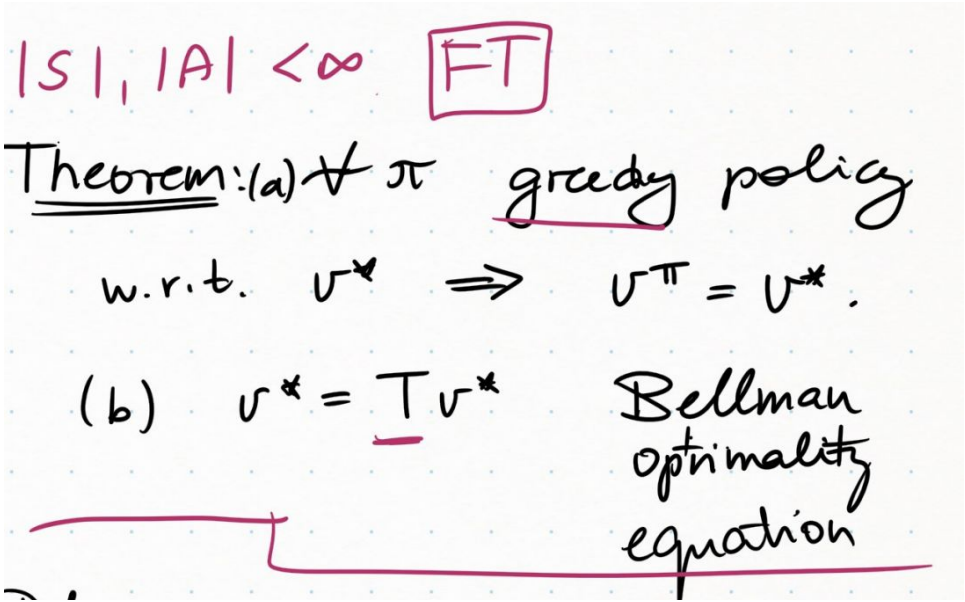
Lecture 2 flipped

Review



Why should we care?
Does it hold more generally?
Is not this too brittle?

Questions

Prabhat Nagarajan 15 hours ago

Typically when we write the definition of an MDP we use (S, A, T gamma, R).

- Should we (always) include the start state distribution (or at least the existence of it) in the definition of an MDP for completeness?

While the optimal policy/value function does not change with the start-state distribution (assuming ergodicity and things), learning those optimal policies/values should be affected by the start state distribution.

- Moreover, if an MDP specifies how the world works, it is incomplete without a start-state distribution.

9

Gábor Mihucz 14 hours ago

Do you find the average reward setup promising from a theoretical perspective?

6

Kushagra Chandak 13 hours ago

The end of the lecture talks about a bound for value iteration - something like log(1/\epsilon)? How do we obtain that?

3

Yongchang Hao 16 hours ago

In lec 2, there is a sentence saying "there are non-memoryless policies whose value function cannot be reproducec by a memoryless policy at every state". But I cannot easily get the intuition. Could you give a hint or an example?

2

Prabhat Nagarajan 11 hours ago

In Sutton and Barto, they define reward functions as a function of s,a,s'. However, I have seen works (including our own lecture notes for this course) using either r(s') or r(s,a) to specify the reward. Does this make things okay mathematically (i.e., do all standard theorems/statements hold under these different formulations)?From personal experience, it seems that all of these are "okay", but there are changes that need to be made in implementation for algorithms to work. E.g., for episodic tasks with a "final state", value iteration converges straightforwardly in the s,a,s' formulation. However, in other scenarios, you might have to subsequently transition to a special cost absorbing state.

2

Jiamin He 2 hours ago

I may have missed this somewhere. But I just want to ask why we didn't define the policy as the memoryless policy in the first place? Since the fundamental theorem of MDPs indicates that the optimal value can be achieved by a memoryless policy (which is greedy with respect to v*), there should be no loss if we define the policy space as the memoryless policy space. Are there any intuitions or benefits of considering non-memoryless policies over memoryless policies? I'm asking because all of the formulations of MDPs I've seen start with memoryless policies.

2

Jiayi Dai 14 hours ago

For a goal, is it possible to prove that some epsilon is the minimum one of all epsilon-optimal policies?

1

Kushagra Chandak 13 hours ago

Also, if I'm not wrong, occupancy measure is something like the expected discounted number of times we visit a state-action pair?

1

Prabhat Nagarajan 12 hours ago

Maybe this is part of the exercise of proving the second theorem, but is there a policy-state-occupancy operator that acts on a state-occupancy vector/matrix that can converge to the state-occupancy measure for a given policy?

Rohini Das 16 minutes ago

What is the difference between denoting policies as a probability simplex over the actions and denoting them using $M_1(A)$? Is it because we want to assign the distributions over subsets of actions? (edited)

Discussion points

Simulate or not?

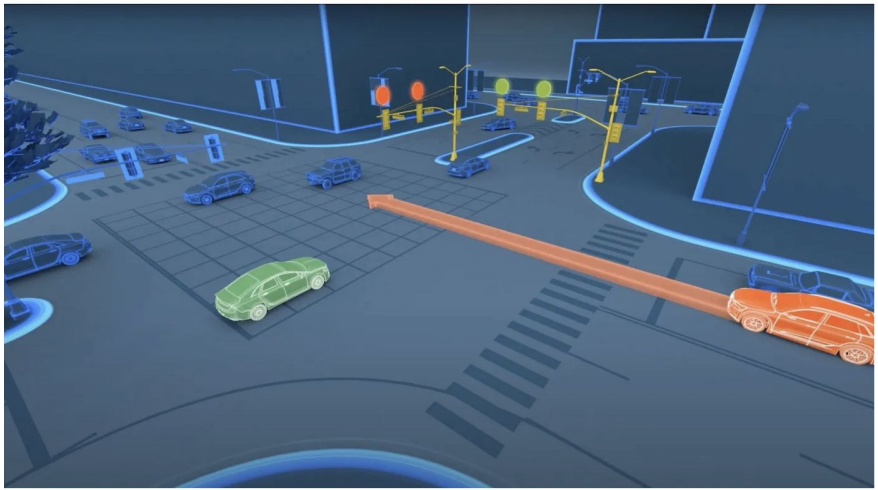
https://bdtechtalks.com/2022/01/06/real-world-reinforcement-learning/

"Basically, it comes down to this question: is it easier to create a brain, or is it easier to create the universe? I think it's easier to create a brain, because it is part of the universe,"

Sergey Levine

Do you agree?

No more simulations



One of the great benefits of offline and self-supervised RL is learning from real-world data instead of simulated environments.