Course 1, Module 4
Policies and Value Functions

CMPUT 397
Fall 2019

Reminders: Sept 25, 2019

- Make sure you are in the Private Session!
  - To check this, just look at the deadlines. If they match the schedule on github, then you are in the private session
- Graded Assessment for Course 1, Module 4 due **this Friday**

A few comments about clarity

- University is about becoming more knowledgeable and more precise
- Writing clarity is extremely important in almost all jobs
- Examples of clarity issues (not about grammar, but about being clear in what you are saying):
  - “How to set the reward value in program?” —> Alternative: “We have to specify the rewards for the agent, to get the desired behaviour. How do we go about specifying the rewards?”
  - “Can it always be known from an agent that used Bellman optimality policy, why this policy is the best (reason the policy in ways that we understand)?”

Discussion Questions

- Many questions “guessing” how we are going to use these definitions
  - e.g., focusing on relationships to epsilon-greedy bandit algorithms
  - e.g., wondering how to improve on solutions using Bellman equations
- Right now, we are still just defining terms

Clarifications on policy optimality

- Why can there be many optimal policies? How does the agent know which optimal policy to choose?

- What if the reward is stochastic and in this case how to deal with the optimal policies?

Clarifications on Access to a model

- For getting the optimal policy from the optimal state value function, How do we have the access to the one-step dynamics of the MDP?

Connections to policies from Bandits section

- Would optimal policy has the save disadvantage as the greedy algorithm we discussed before? That is, we may neglect global max return and stuck in local max?

Bellman equations

- The Bellman optimality equation for Vpi has a unique solution that is independent of policy in finite MDPs. Why it is independent?

- In Policy Iteration, in the Policy Evaluation part, does the algorithm choose next action based on dynamics of the state? How come it will converge if it may choose different action every time?

- For the grid world example, how are we supposed to build up the V* map since it's more like a recursive problem? (related question: When we calculate the value function in this state, how do we know the value function in other states?)

Conditions on solution

- Why does a policy depend only on the current state? Why is this assumption important?
- Under what circumstances we can use Bellman equations to solve a MDP problem?

Problem specification

- With different Gama values, the optimal policy will be different, so how do we know which policy is the optimal policy we want

- Can you think of real world examples where there would be multiple optimal states?

Two related questions with an answer, though its not that obvious

- “If trying to simulate a human being using reinforcement learning, would the optimal policy be a simulation of a human (who makes mistakes) or a "perfect" person who always makes the correct choice?”

- “Would there ever be a case in which we would want to create an agent that does not use an optimal policy? ie. Would there ever be a situation where running a sub optimal policy of not maximizing the long term rewards be good for the designer? ”

Discussion topics for today

1. Debugging agents: Occasionally, mistakes can be made when creating complex things like reinforcement learning programs. Can we use the Bellman equations to verify the correctness of our agent's values and policy? (Xutong, your TA)

2. How can we decide what the optimal reward to assign if we're defining our own MDP? For example, in what cases should we give only negative or only positive rewards, and how do we decide how large/small they should actually be? (Sungsu, your TA)

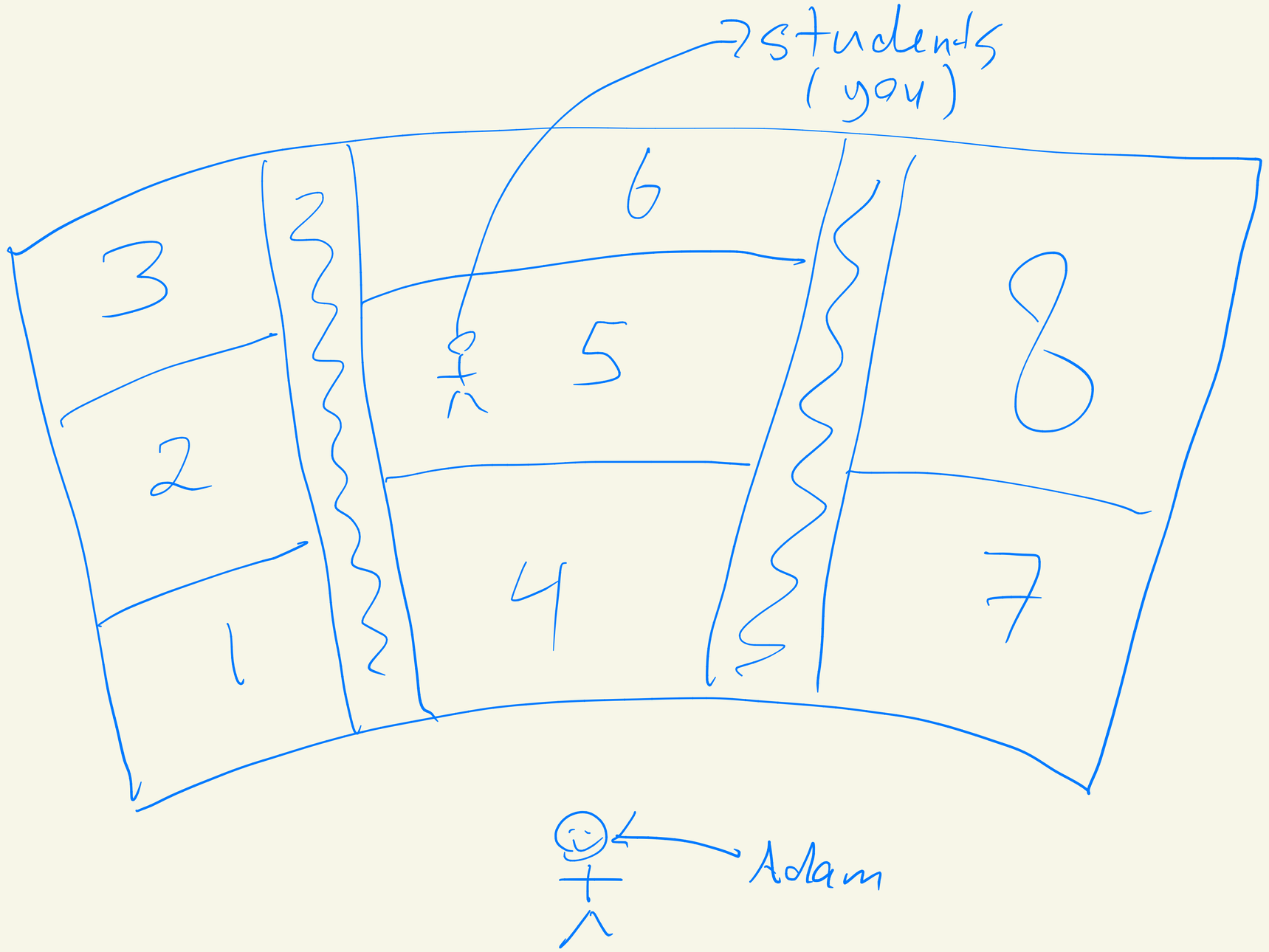
3. How do we know there is always an optimal policy?

4. Is it possible to generalize to infinite MDPs? Do we need to?

5. Why are we learning about Bellman equations? And why does this seem so impractical? (Adam)

6. Give examples where two policies can be optimal but different. (Derek, your TA)

7. Math Formulation Discussion: There were several different formulations for the dynamics of the MDP including the state transition probabilities and specifying the reward function as r(s,a). Is it possible to rewrite the Bellman equations using either 3.4 or 3.5 in the textbook? (Ryan, your TA)


7 students
(you)
3
2
6
5
8
2
1
4
7
Adam