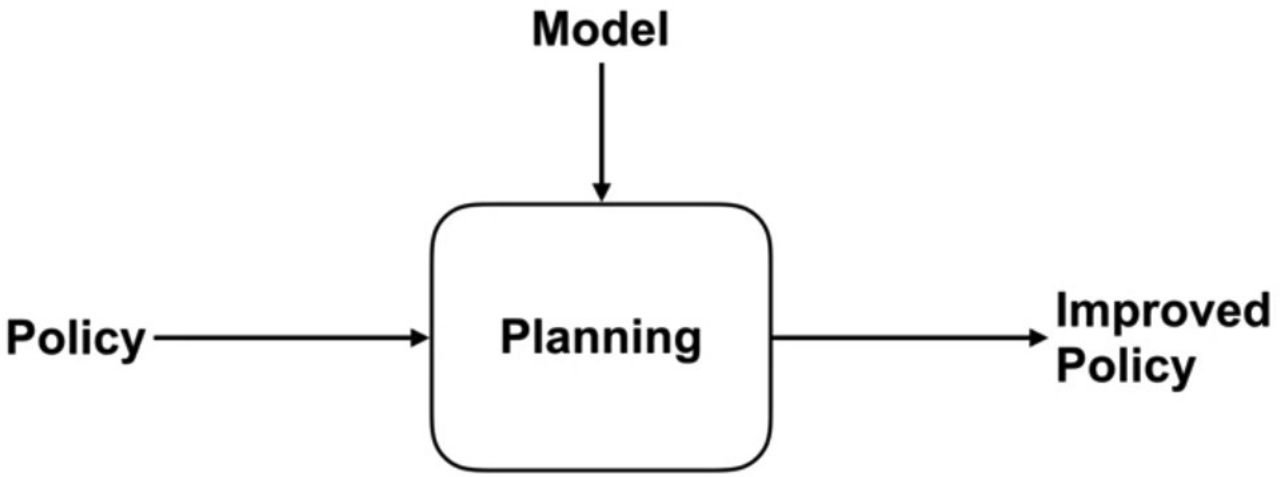
Define Planning in Reinforcement Learning

Objectives

- Explain how planning is used to improve policies
- Describe random-sample one-step tabular Q-planning
- How planning typically works in RL

Planning

- ❑ Planning is a process which takes a model as input and produces unimproved policy.

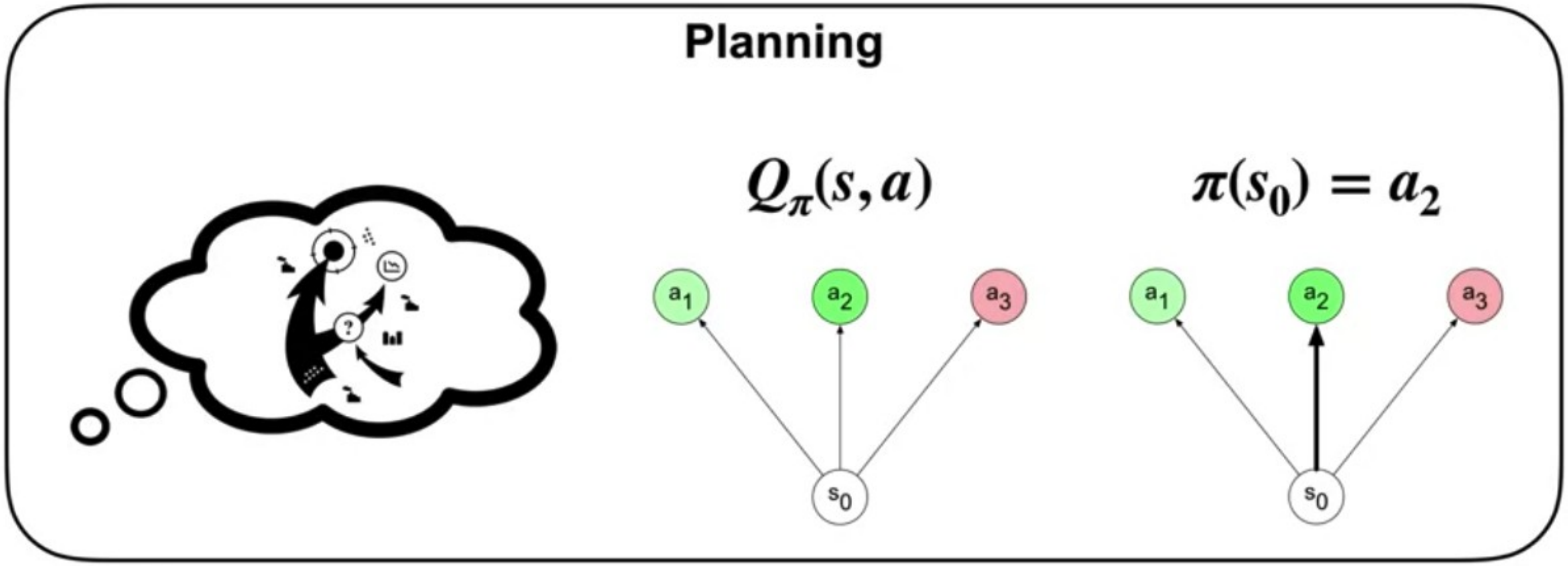


Planning

- One possible approach to planning is to first sample experience from the model.
- This generated experience can then be use to perform updates to the value function as if these interactions actually occurred in the world.
- Behaving greedily with respect to these improved values results in improved policy

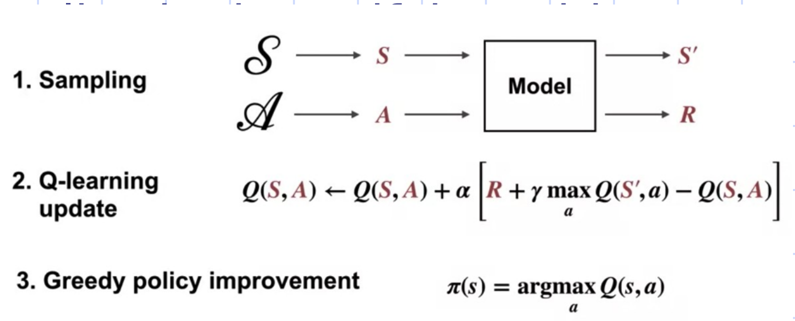
Planning

- □ Planning improve policies



Planning

- ❑ This algorithm first chooses a state action pair at random from the set of all states and actions.
  - ❑ It then queries the sample model with this state action pair to produce a sample of the next state and reward.
  - ❑ It then performs a Q-Learning Update on this model transition.
  - ❑ Finally, it improves the t to the updated action v


1. Sampling

$$\mathcal{S} \longrightarrow S \longrightarrow \boxed{\text{Model}} \longrightarrow S'$$
$$\mathcal{A} \longrightarrow A \longrightarrow \boxed{\text{Model}} \longrightarrow R$$

2. Q-learning update

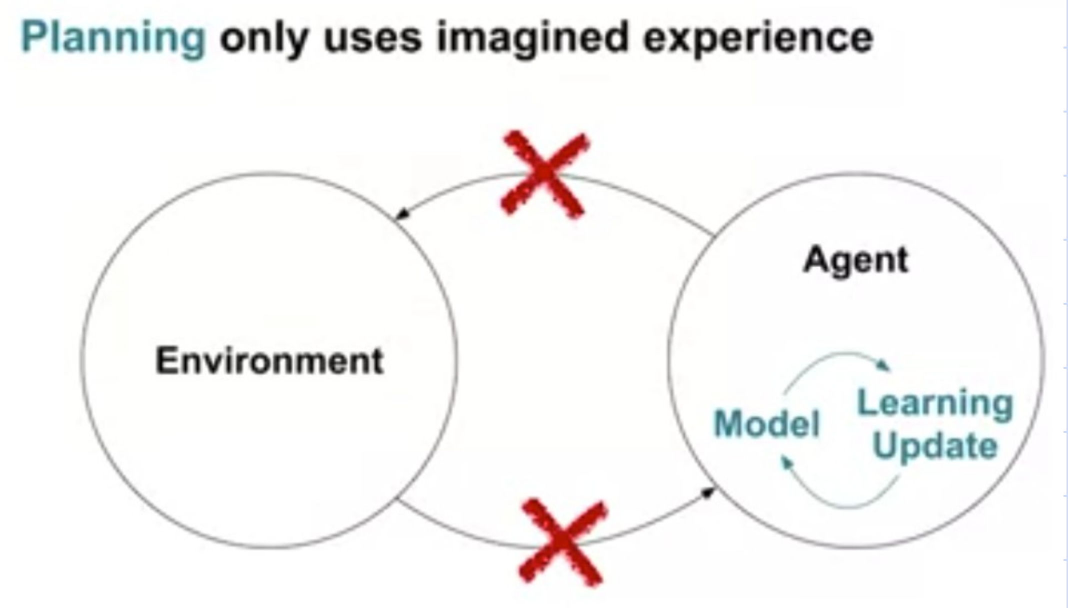
$$Q(S,A) \leftarrow Q(S,A) + \alpha\left[R + \gamma \max_a Q(S',a) - Q(S,A)\right]$$

3. Greedy policy improvement

$$\pi(s) = \operatorname{argmax}_a Q(s,a)$$

Planning

- ❑ Planning method only uses imagined or simulated experience.
- ❑ All of these updates can be done without behaving in the wo loop.



Planning only uses imagined experience

Planning

- In reinforcement learning (RL), planning refers to the process of making decisions about what actions to take in an environment in order to achieve a specific goal.
- Unlike learning from direct interaction with the environment (as in model-free RL algorithms like Q-learning or SARSA), planning involves simulating possible future states and rewards to determine the best course of action.

How planning works in RL

- ❑ Model Representation:
  - ❑ Planning often requires a model of the environment, which describes how the environment transitions from one state to another and the rewards associated with each state-action pair. This model can be learned from experience (model-based RL) or provided by the environment (model-based planning).
- ❑ Simulation:
  - ❑ Using the model, the agent simulates possible sequences of states and actions to predict the consequences of different action choices. This simulation allows the agent to explore potential future trajectories without actually interacting with

How planning works in RL

- Action Selection:
  - Based on the simulated trajectories, the agent selects actions that are expected to lead to desirable outcomes. This could involve selecting actions that maximize expected rewards, minimize expected costs, or achieve specific objectives.
- Policy Improvement:
  - After selecting actions, the agent may update its policy based on the simulated outcomes. For example, it may learn from the simulated trajectories to improve its decision-making strategy in similar situations in the future.

How planning works in RL

- Execution:
  - Once a plan has been determined, the agent executes the chosen actions in the real environment and observes the actual outcomes. These outcomes can then be used to evaluate the effectiveness of the plan and inform future planning decisions.

Summary

- Explain how planning is used to improve policies
- Describe random-sample one-step tabular Q-planning
- How planning typically works in RL

Q & A