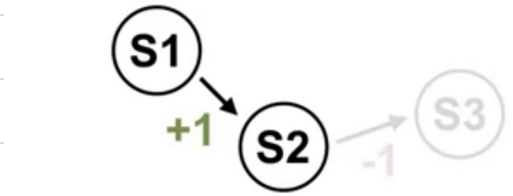
Dealing with inaccurate models

Objectives

- Explain the effects of planning with an inaccurate model
- Explain how model inaccuracies produce another exploration-exploitation trade-off
- Compare Dyna-Q and Dyna-Q+

How models can be inaccurate

- Models are inaccurate when transitions they store are different from transitions that happen in the environment.
- Incomplete models.
  - The agent hasn't tried most of the actions in almost all of the states. The transitions associated with trying those actions in those states are simply missing from the model.



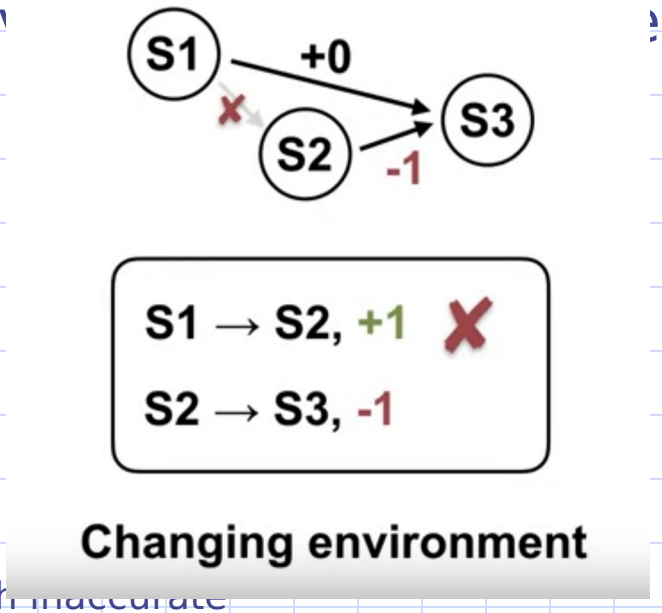
S1 → S2, +1

S2 → ??

Incomplete model

How models can be inaccurate

- The model could also be an accurate if the environment changes. Taking an action in a state could result in a different next state and reward than what the agent obser[illegible].
- Inaccurate model:
  - what actually happens is different from what the model says.

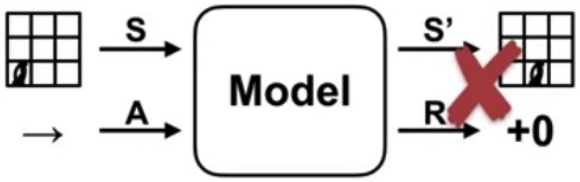


How models can be inaccurate

- The effect of planning with inaccurate models depends on how the model is inaccurate.
  - Since the model can't produce a next step or a reward, it can't be used for planning.
  - However, as the agent interacts with the environment, the model stores more and more transitions.
  - Then, the agent can perform updates by simulating transitions it's seen before.
  - That means that as long as the agent has seen some transitions, it can plan with the model.

How models can be inaccurate

- What will happen when the agent tries to plan using its inaccurate model?
  - If the agent tries to perform a planning update with one of the incorrect transitions in the model, the value function or policy that the agent updates might change in the wrong di



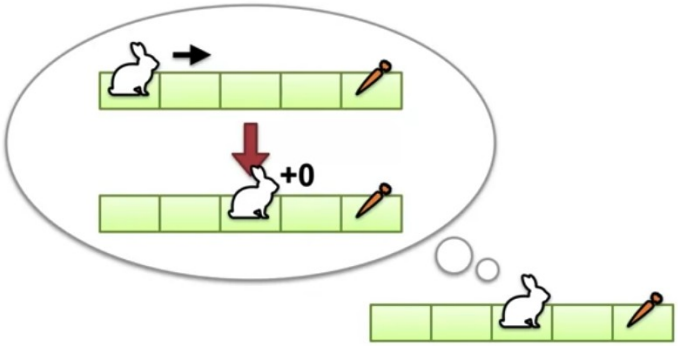
$$Q(S,A) \leftarrow Q(S,A) + \alpha(R + \gamma \max_a Q(S',a) - Q(S,A))$$

How models can be inaccurate

- When the agents model is inaccurate, planning will likely make the policy or value function worse with respect to the environment.
- That means the agent has a vested interest in making sure it's model stays accurate.

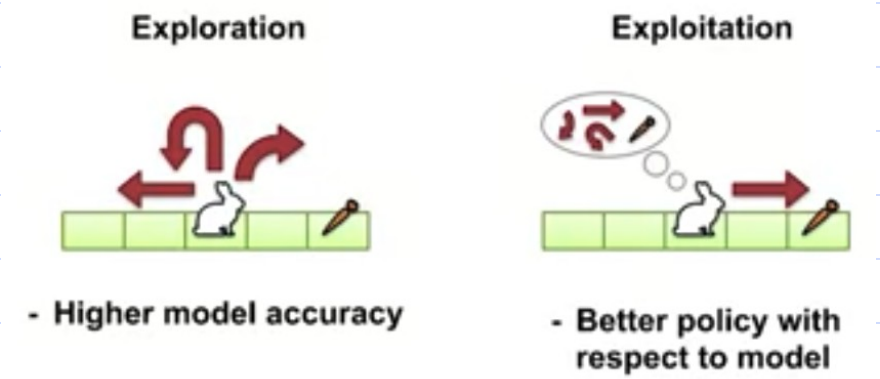
How models can be inaccurate

- Example:
  - The robust model initially says that the rabbit will go directly to the carrot if it chooses the move right.
  - However, after moving right, the rabbit in the middle square.
  - After experiencing this transition, rabbit update its model of the world.
  - The update reflects that choosing the move right action, in the leftmost squ middle square.



How models can be inaccurate

- In changing environments, the agent's model might become inaccurate at any time. So the agent has to make a choice;
  - explorer to make sure it's model is accurate
  - exploit the model to compute the optimal policy, assuming that the model is correct



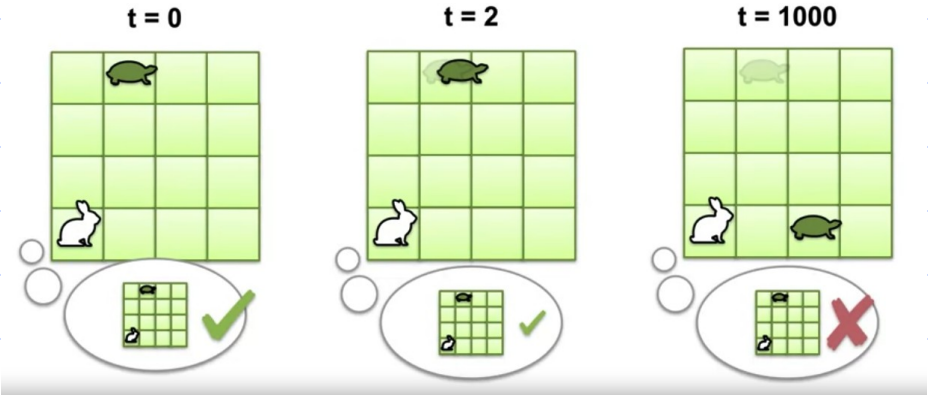
FPT Education
FPT UNIVERSITY

How models can be inaccurate

- When the environment changes, the model will be incorrect.
- It will remain incorrect until the agent revisits that part of the environment that changed and updates the model.
- This suggests that the agent should explore places it has not been to in a while.

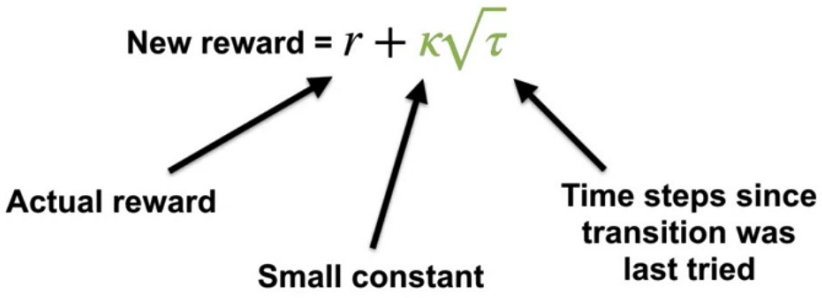
How models can be inaccurate

- Example:
  - The rabbit knows the turtle starts in the following cell.
  - Since the turtle moves slowly, it can't move very far in the first few time-steps.
  - However, after a long time, the turtle might be in a totally different cell than it started in, making the rabbits model incorrect.



How models can be inaccurate

- Bonus rewards for exploration:
  - To encourage the agent to revisit its state periodically, we can add a bonus to the reward used in planning.
  - This bonus is simply Kappa, times the square root of Tau, where r, is the reward from the model and Tau is the amount of time it's been since the state action pair was last visited in

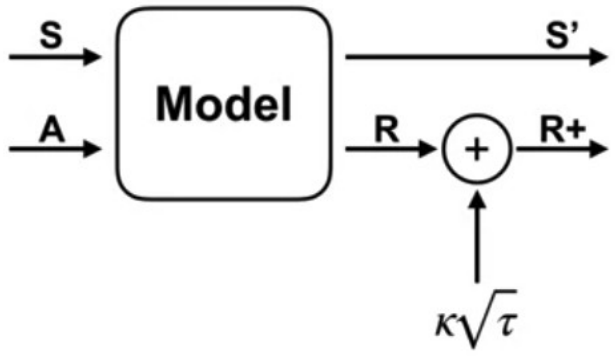


How models can be inaccurate

- Tau is not updated in the planning loop, that would not be a real visit.
- Kappa is a small constant that controls the influence of the bonus on the planning update.
- If Kappa was zero, we would ignore the bonus completely.

How models can be inaccurate

- The Dyna-Q+ algorithm
  - Adding this exploration bonus to the planning updates results in the Dyna-Q+ algorithm.
  - By artificially increasing the rewards used in planning, we increase the value of state action pairs it haven't been visited recent



Dealing with inaccurate models

Dyna-Q vs Dyna-Q+

- Dyna-Q and Dyna-Q+ both use model-based planning to improve value estimates in reinforcement learning tasks:
  - Dyna-Q uses a fixed exploration bonus and updates all state-action pairs during planning
  - Dyna-Q+ adapts the exploration bonus over time and focuses planning efforts on state-action pairs with the highest potential for improvement.

Dyna-Q vs Dyna-Q+

- Exploration
  - In Dyna-Q, the agent uses a fixed exploration bonus during planning to encourage exploration of less-visited state-action pairs.
  - In Dyna-Q+, the exploration bonus used during planning decays over time as the agent gains more experience..

Dyna-Q vs Dyna-Q+

- Planning Strategy
  - Dyna-Q follows a straightforward planning strategy where the agent performs updates for all state-action pairs visited during planning, regardless of their importance or relevance.
  - Dyna-Q+ prioritizes planning efforts by focusing on state-action pairs with the largest uncertainty or potential for improvement.

Dyna-Q vs Dyna-Q+

- Additional Reward
  - Dyna-Q does not incorporate additional rewards during planning. It relies solely on the exploration bonus to drive exploration and improve value estimates.
  - Dyna-Q+ incorporates an additional reward signal during planning based on the novelty of experiences.

Summary

- Explain the effects of planning with an inaccurate model
- Describe how Dyna can plan successfully with a partially inaccurate model
- Explain how model inaccuracies produce another exploration-exploitation trade-off
- Compare Dyna-Q and Dyna-Q+

Q & A