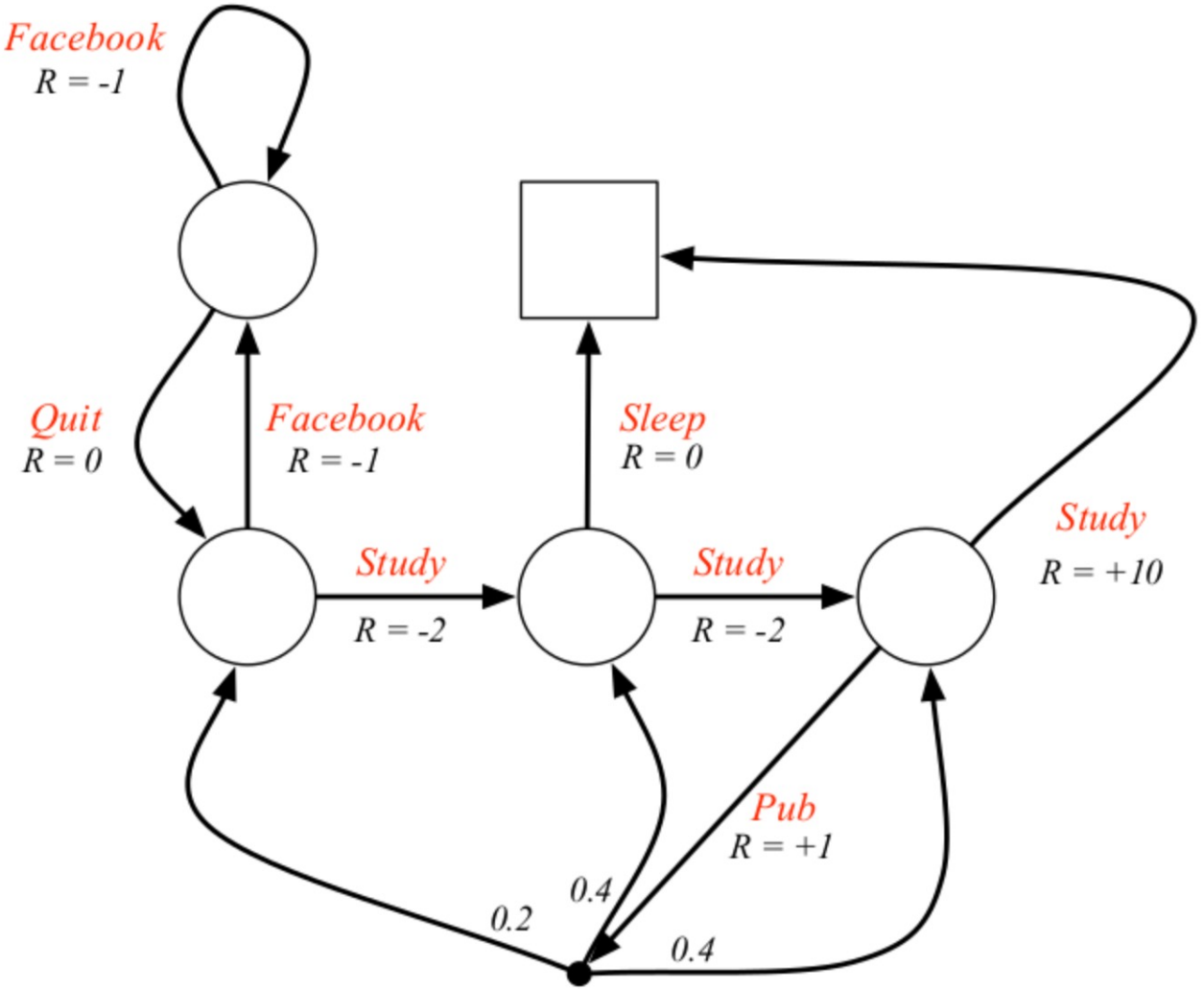
CSCI 3202: Intro to Artificial Intelligence

Lecture 33: MDP– Value iteration, Policy iteration

Rhonda Hoenigman

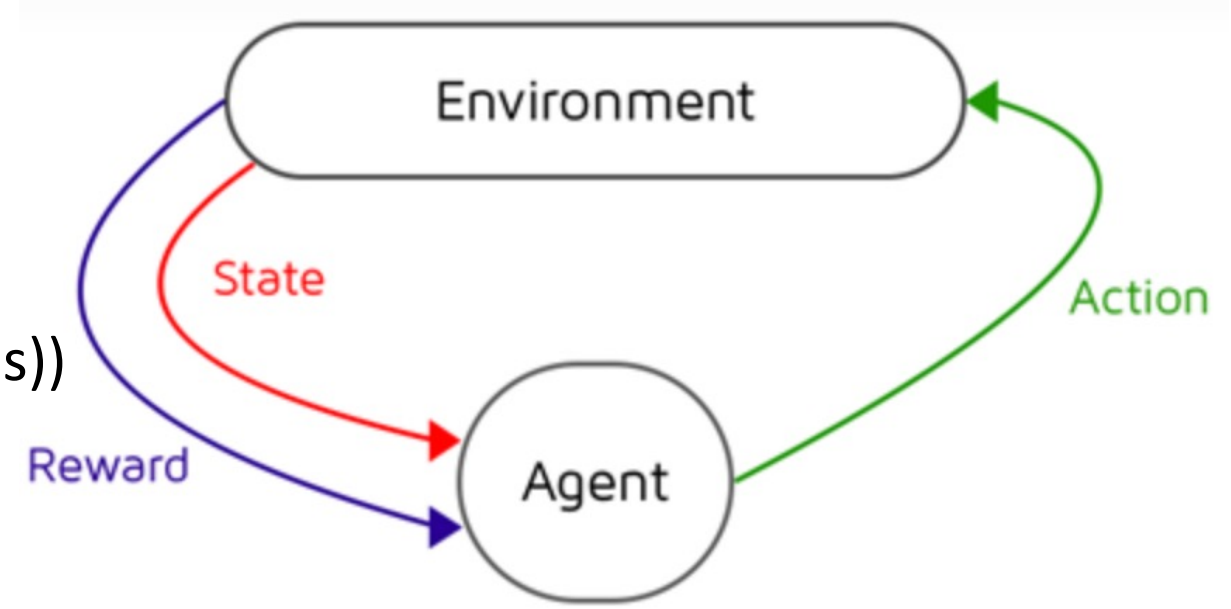
Department of Computer Science



Student MDP with actions

Markov Decision Process – Overview

A Markov Decision Process (MDP): Markov decision processes (MDP) are a framework for sequential decision making. At each step, the agent chooses an action from a set of possible actions.



Requires:

- States (call them s, with initial s_0)
- Actions available in each state (Actions (s))
- Transition model (P(s' | s, a))
- Reward function R(s) (or R(s, a, s'))

Markov Decision Process – Overview

A Markov Decision Process (MDP) objective: In an MDP, we are looking for an action in each state that solves the problem and maximize the agent's expected utility.

Policy:

Utility of state:

- Real number that captures how well the state represents agent's goals

Markov Decision Process

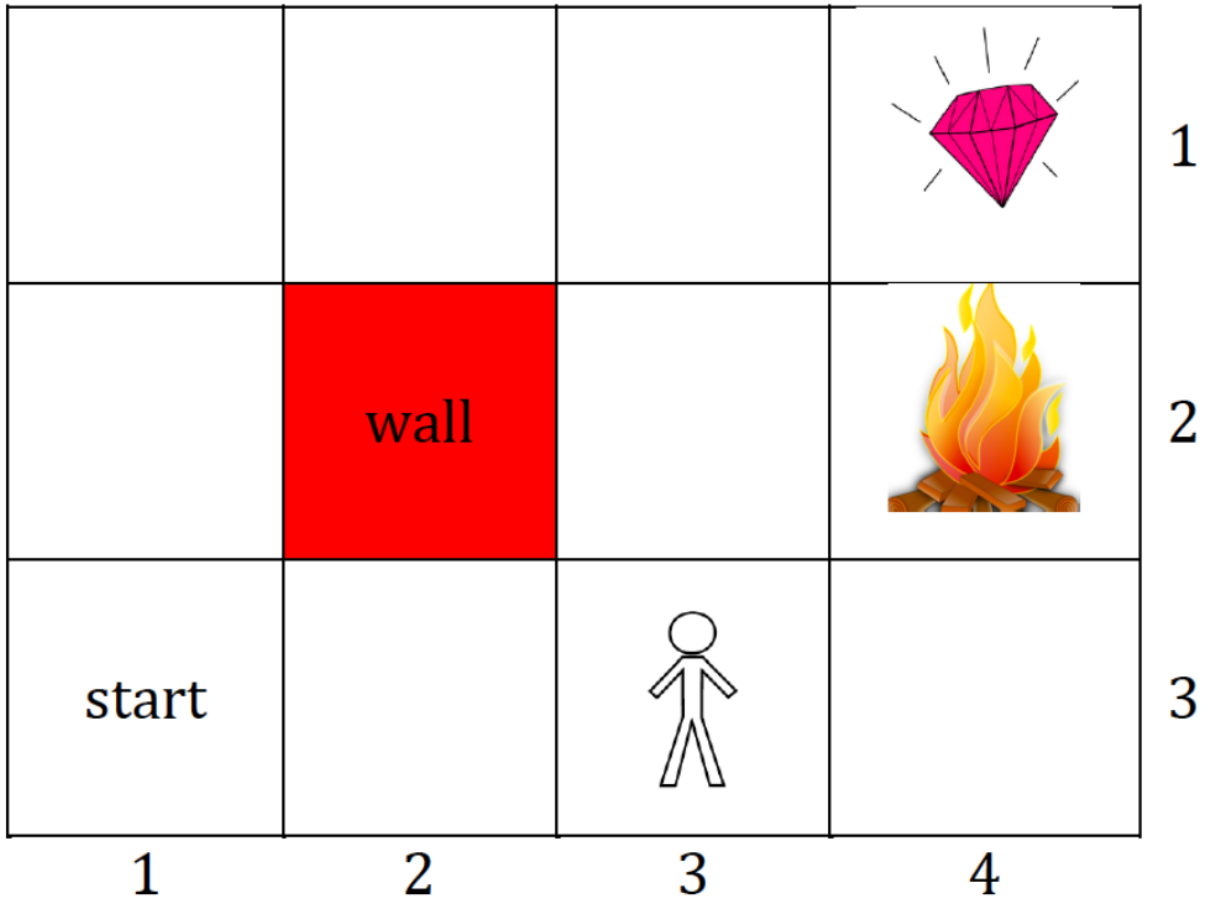
Example: Move agent from "start" to the diamond without falling into the fire pit.

- Diamond has reward of +1.
- Firepit has reward of -1.

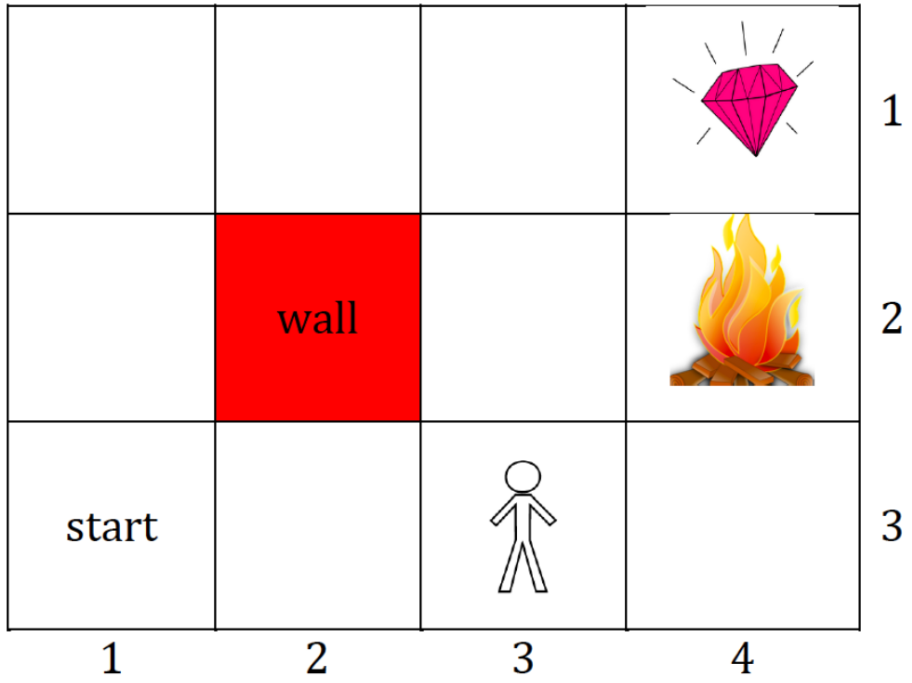
- With certainty in actions, the solution is trivial.

- Agent successful 80% of time. 10% of time goes left, 10% of time goes right.

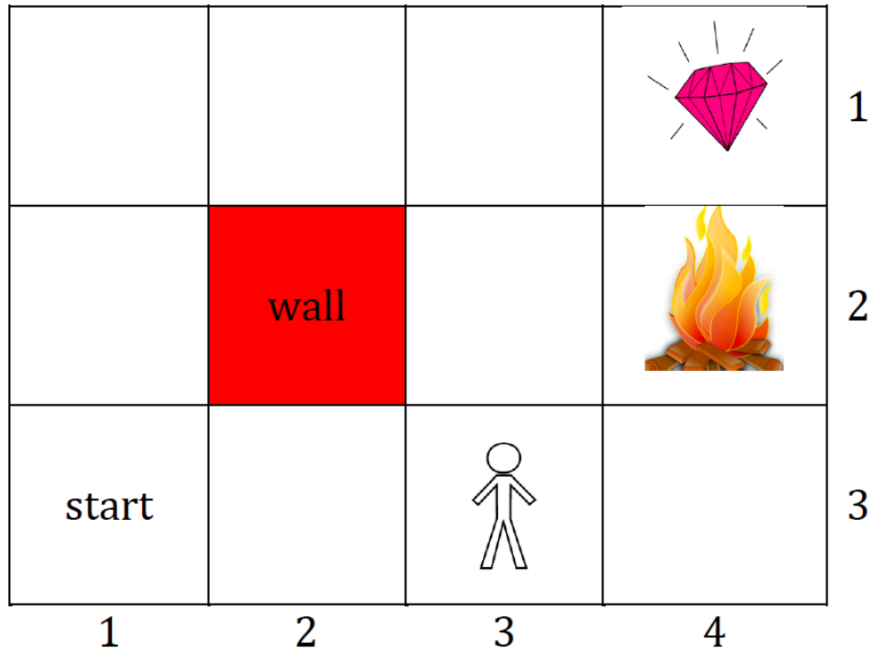
Eg. Assume the agent wants to go north. 80% of the time the action is successful, 10% of the time the agent goes east and 10% of the time the agent goes west. If agent goes into a wall, the agent just bounces off and stays in the same location.



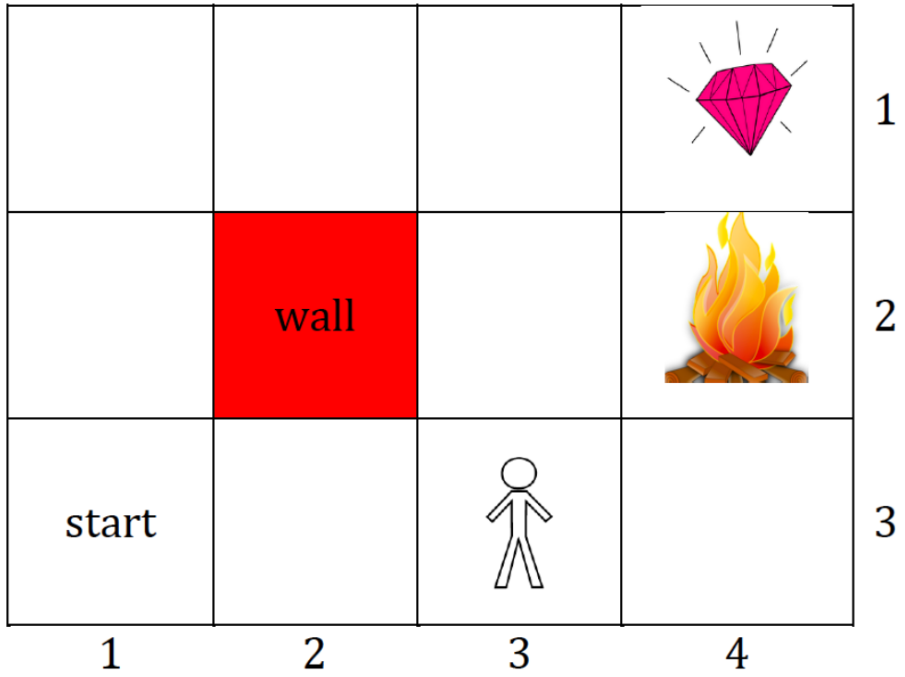
Markov Decision Process – Bellman equations



Markov Decision Process – Bellman equations

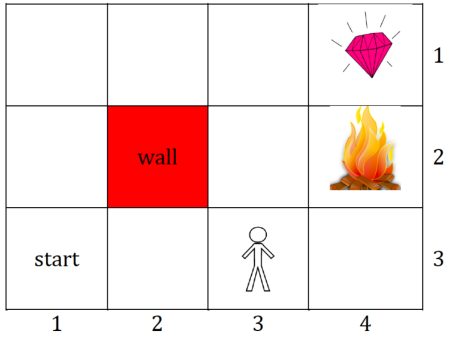


Markov Decision Process – Bellman equations



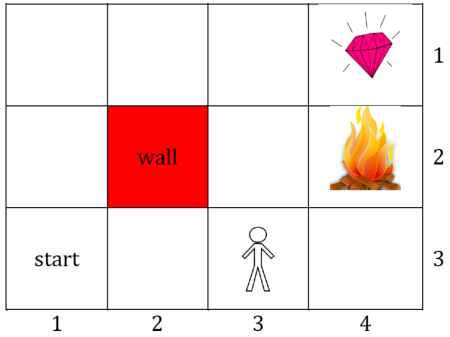
Markov Decision Process – value iteration

Iterative algorithm using Bellman equations to find utility of each state.

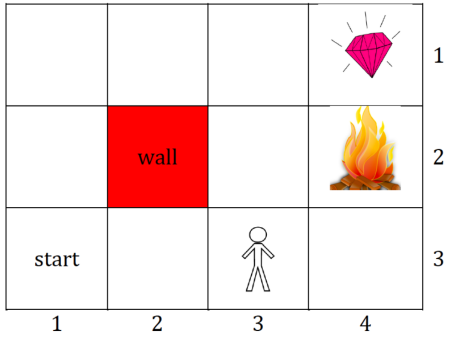


Markov Decision Process – policy iteration

Iterate through policies rather than values.



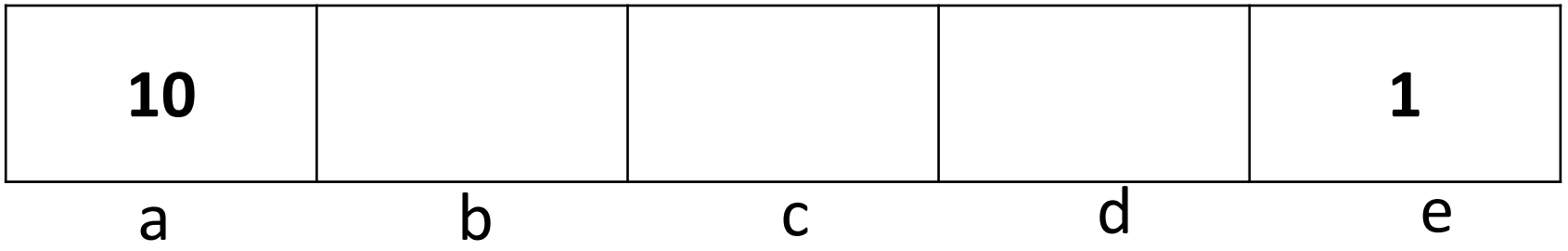
Markov Decision Process – policy iteration



Policy Iteration

Example: Given the following grid, find the optimal policy of each state using the policy iteration algorithm. The PolicyEvaluation() function sets U for all states using the current policy, U, and the MDP.

- The terminal states are a and e, and those states have the rewards shown. Let $\gamma = 0.9$
- The actions are move left and move right.
- $P(s'|a, s) = 0.80$ success and 0.20 that the agent stays in the same state.
- $R(s) = -0.04$



Policy Iteration

Example: (continued)

Policy Iteration

Example: (continued)

Policy Iteration

Example: (continued)

Policy Iteration

Example: (continued)