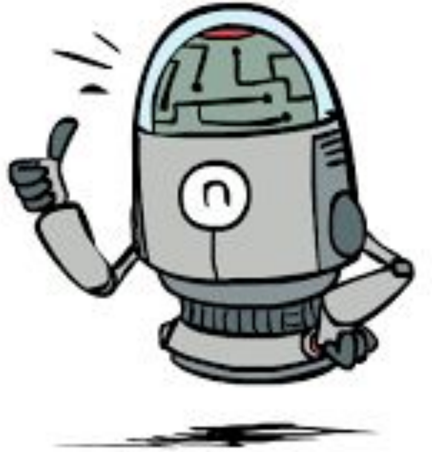
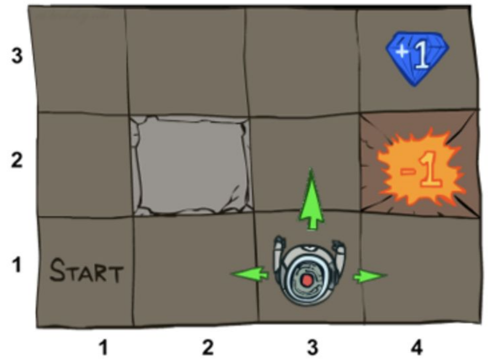
CS 188 Discussion 5b

Markov Decision Processes (MDPs)



Markov Decision Processes

- An MDP is defined by:
  - A set of states s ∈ S
  - A set of actions a ∈ A
  - A transition function T(s, a, s')
    - Probability that a from s leads to s', i.e., P(s'| s, a)
    - Also called the model or the dynamics
  - A reward function R(s, a, s')
    - Sometimes just R(s) or R(s')
  - A start state
  - Maybe a terminal state

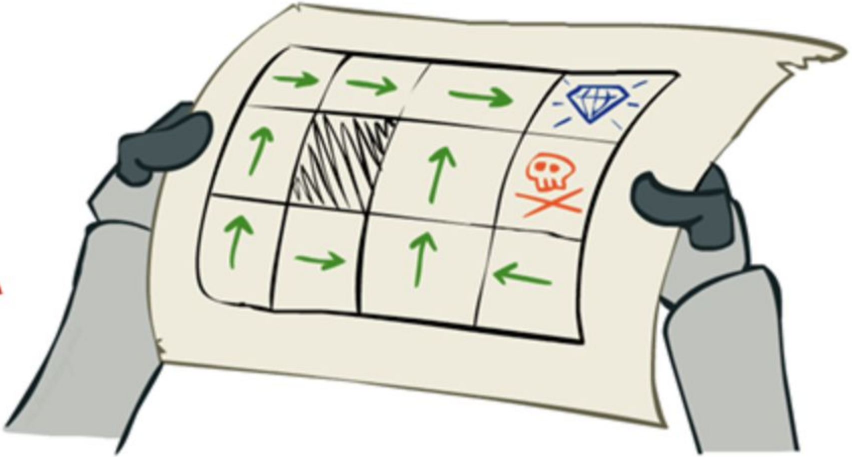


Policies

- In deterministic single-agent search problems, we wanted an optimal plan, or sequence of actions, from start to a goal

- For MDPs, we want an optimal policy $\pi^*: S \to A$
  - A policy π gives an action for each state
  - An optimal policy is one that maximizes expected utility if followed

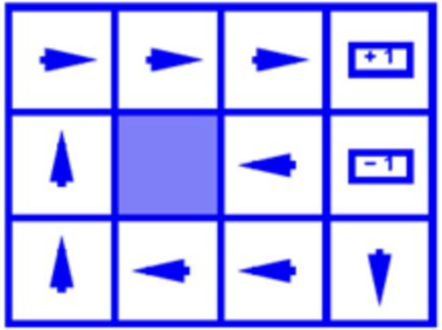
- Expectimax didn't compute entire policies
  - It computed the action for a single state only



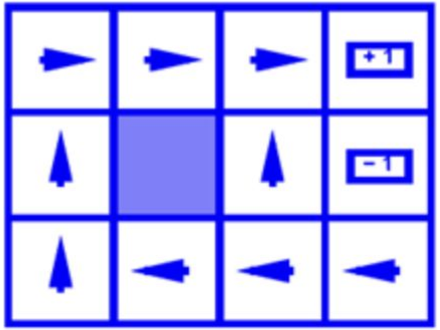
Optimal policy when R(s, a, s') = -0.03 for all non-terminals s

slide yoinked from sp23

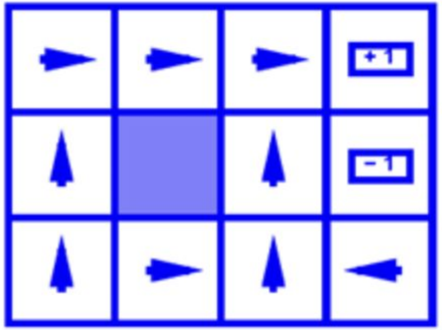
Optimal Policies



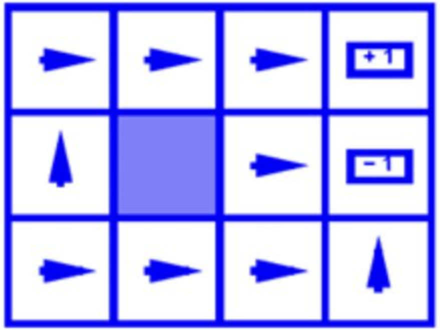
R(s) = -0.01



R(s) = -0.03



R(s) = -0.4



R(s) = -2.0

slide yoinked from sp23

Bellman Eqns & Value Iteration

- Bellman equations characterize the optimal values:
  - $$V^*(s) = \max_a Q^*(s,a)$$
    $$Q^*(s,a) = \sum_{s'} T(s,a,s')\left[R(s,a,s') + \gamma V^*(s')\right]$$
    $$V^*(s) = \max_a \sum_{s'} T(s,a,s')\left[R(s,a,s') + \gamma V^*(s')\right]$$

- Value iteration computes them:
  - $$V_{k+1}(s) \leftarrow \max_a \sum_{s'} T(s,a,s')\left[R(s,a,s') + \gamma V_k(s')\right]$$

Policy Evaluation, Extraction, & Iteration

- Policy Evaluation calculates the V's for a fixed policy
  - $$V_{k+1}^{\pi}(s) \leftarrow \sum_{s'} T(s, \pi(s), s')[R(s, \pi(s), s') + \gamma V_k^{\pi}(s')]$$
- Policy extraction determines optimal policy given optimal values V*(s)
  - $$\pi^*(s) = \arg\max_a \sum_{s'} T(s, a, s')[R(s, a, s') + \gamma V^*(s')]$$
- Policy iteration lets us find the optimal policy *faster* than value iteration
  - Evaluation:
    $$V_{k+1}^{\pi_i}(s) \leftarrow \sum_{s'} T(s, \pi_i(s), s') \left[R(s, \pi_i(s), s') + \gamma V_k^{\pi_i}(s')\right]$$
  - Improvement: $$\pi_{i+1}(s) = \arg\max_a \sum_{s'} T(s, a, s') \left[R(s, a, s') + \gamma V^{\pi_i}(s')\right]$$

tinyurl.com/cs188su23

tinyurl.com/188feedback