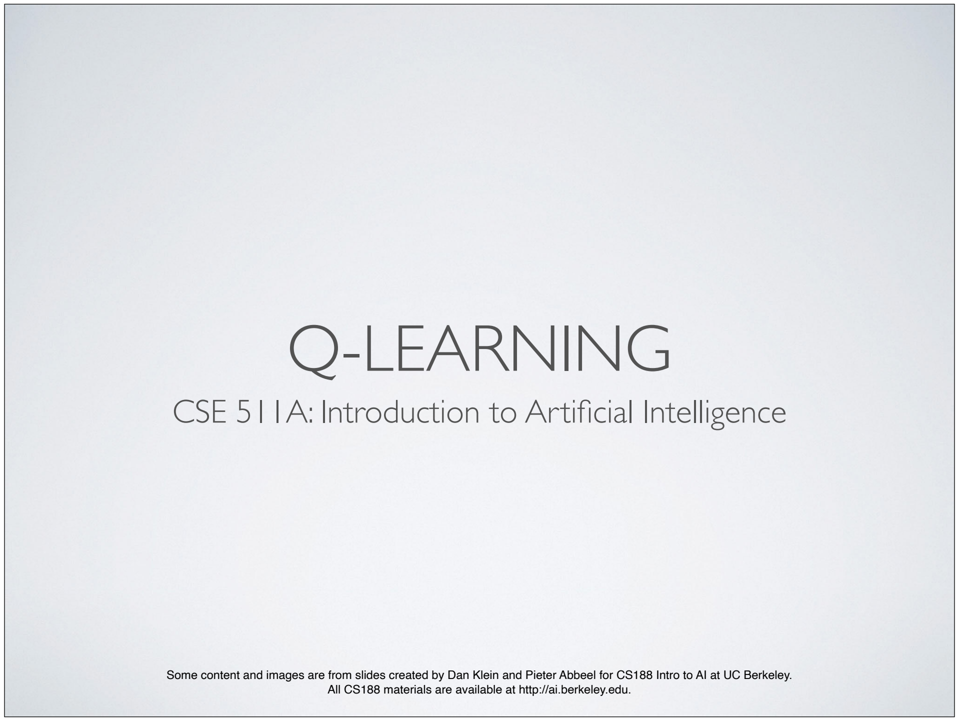
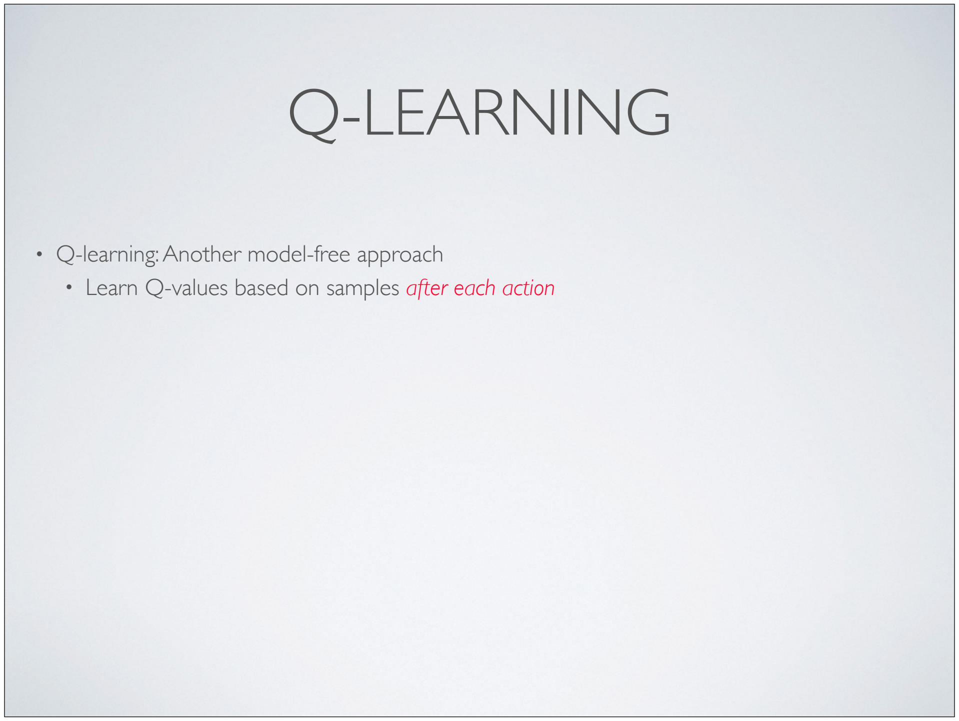


Q-LEARNING

CSE 511A: Introduction to Artificial Intelligence

Some content and images are from slides created by Dan Klein and Pieter Abbeel for CS188 Intro to AI at UC Berkeley.
All CS188 materials are available at http://ai.berkeley.edu.

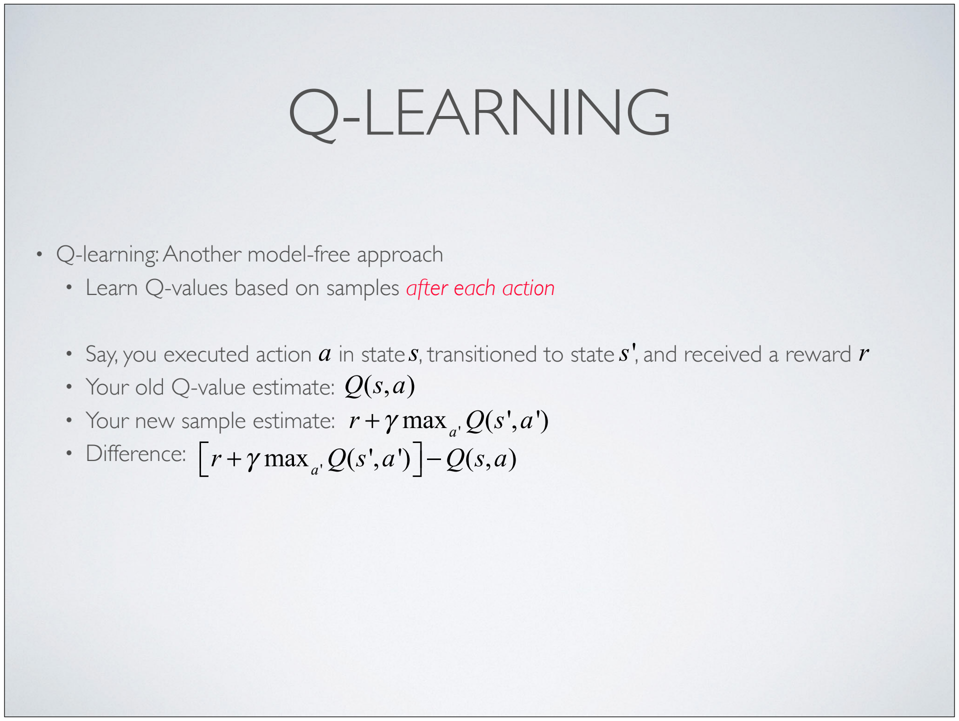
1



Q-LEARNING

- Q-learning: Another model-free approach
  - Learn Q-values based on samples *after each action*

2

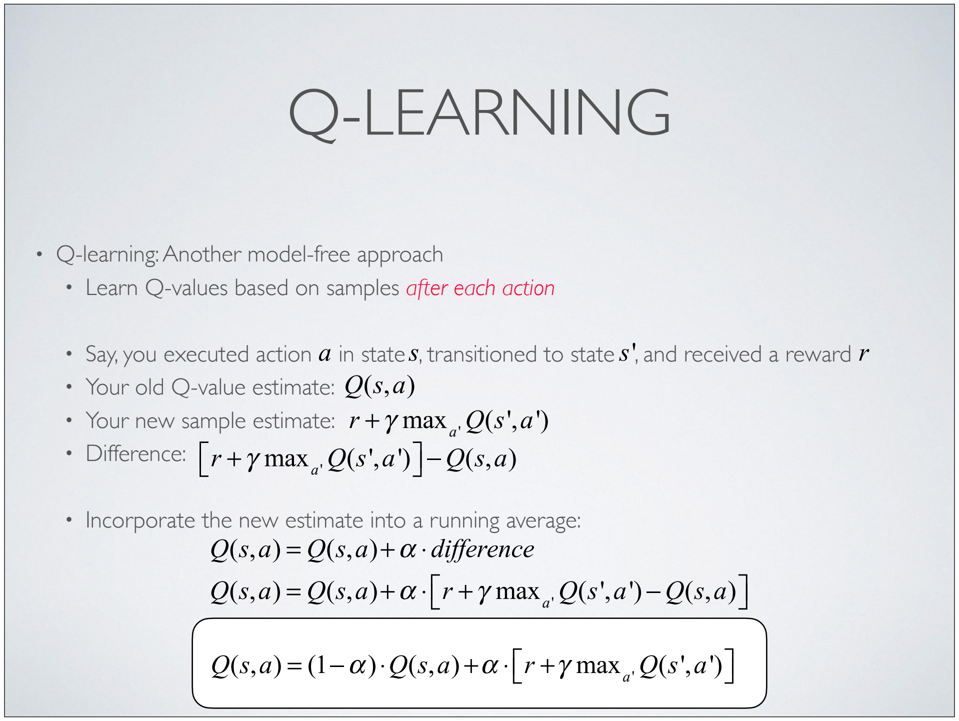


Q-LEARNING

- Q-learning: Another model-free approach
  - Learn Q-values based on samples *after each action*

  - Say, you executed action a in state s, transitioned to state s', and received a reward r
  - Your old Q-value estimate: $Q(s,a)$
  - Your new sample estimate: $r + \gamma \max_{a'} Q(s',a')$
  - Difference: $\left[r + \gamma \max_{a'} Q(s',a')\right] - Q(s,a)$

3



Q-LEARNING

- Q-learning: Another model-free approach
  - Learn Q-values based on samples *after each action*

  - Say, you executed action a in state s, transitioned to state s', and received a reward r
  - Your old Q-value estimate: $Q(s,a)$
  - Your new sample estimate: $r + \gamma \max_{a'} Q(s',a')$
  - Difference: $\left[r + \gamma \max_{a'} Q(s',a')\right] - Q(s,a)$

  - Incorporate the new estimate into a running average:

$$Q(s,a) = Q(s,a) + \alpha \cdot \mathit{difference}$$

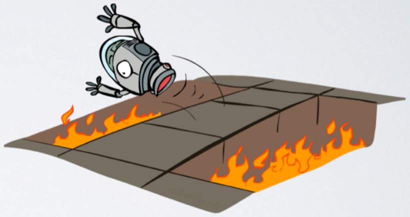
$$Q(s,a) = Q(s,a) + \alpha \cdot \left[r + \gamma \max_{a'} Q(s',a') - Q(s,a)\right]$$

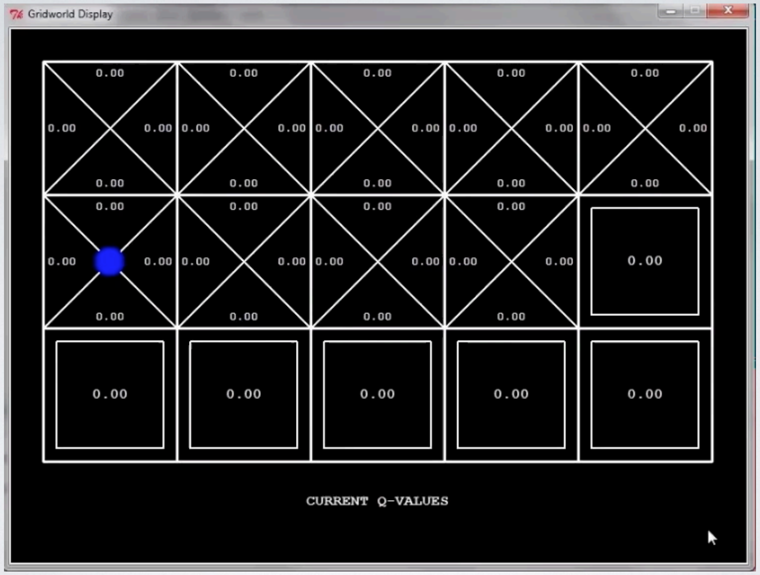
$$\boxed{Q(s,a) = (1-\alpha) \cdot Q(s,a) + \alpha \cdot \left[r + \gamma \max_{a'} Q(s',a')\right]}$$

4

EXAMPLE

- Start at one end of the cliff
- Reward of 10 if it gets to the other end
- Reward of -100 if it falls into the pit
- Learning rate (alpha) = 0.5
- Discount factor = 1
- Transitions are all deterministic





0.00 0.00 0.00 0.00	0.00 0.00 0.00 0.00	0.00 0.00 0.00 0.00	0.00 0.00 0.00 0.00	0.00 0.00 0.00 0.00
0.00 0.00 0.00 0.00	0.00 0.00 0.00 0.00	0.00 0.00 0.00 0.00	0.00 0.00 0.00 0.00	0.00
0.00	0.00	0.00	0.00	0.00

CURRENT Q-VALUES

Q-LEARNING

- Q-learning: Another model-free approach
  - Converges to an optimal policy — even if you are acting sub optimally!
- Caveats:
  - Your have to explore enough
  - You have to eventually make the learning rate small enough
    … but not decrease it too quickly
  - Basically, in the limit, it doesn't matter how you select actions!