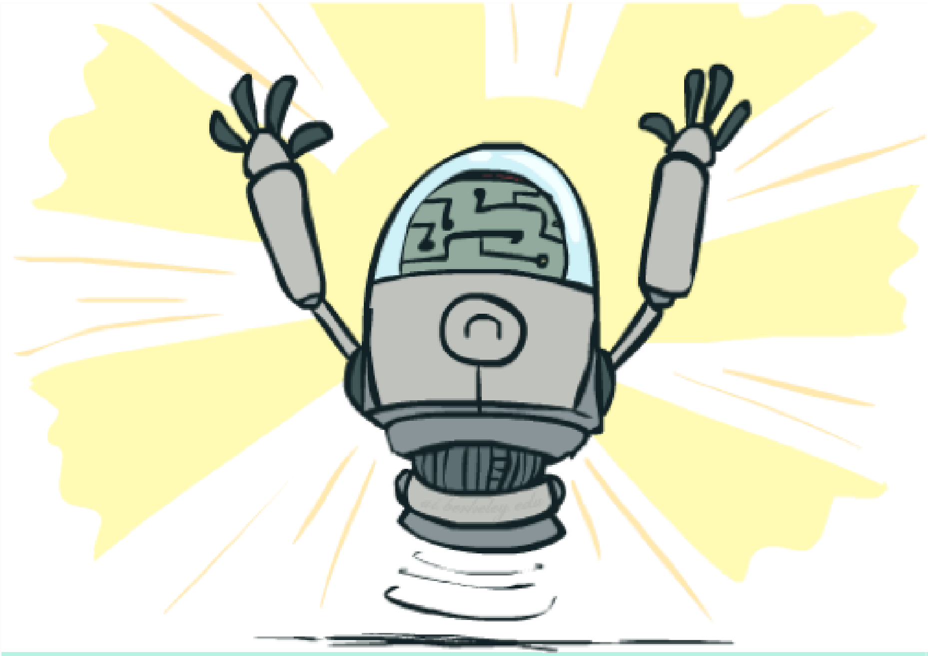
Advanced Topics in AI

Q-Learning



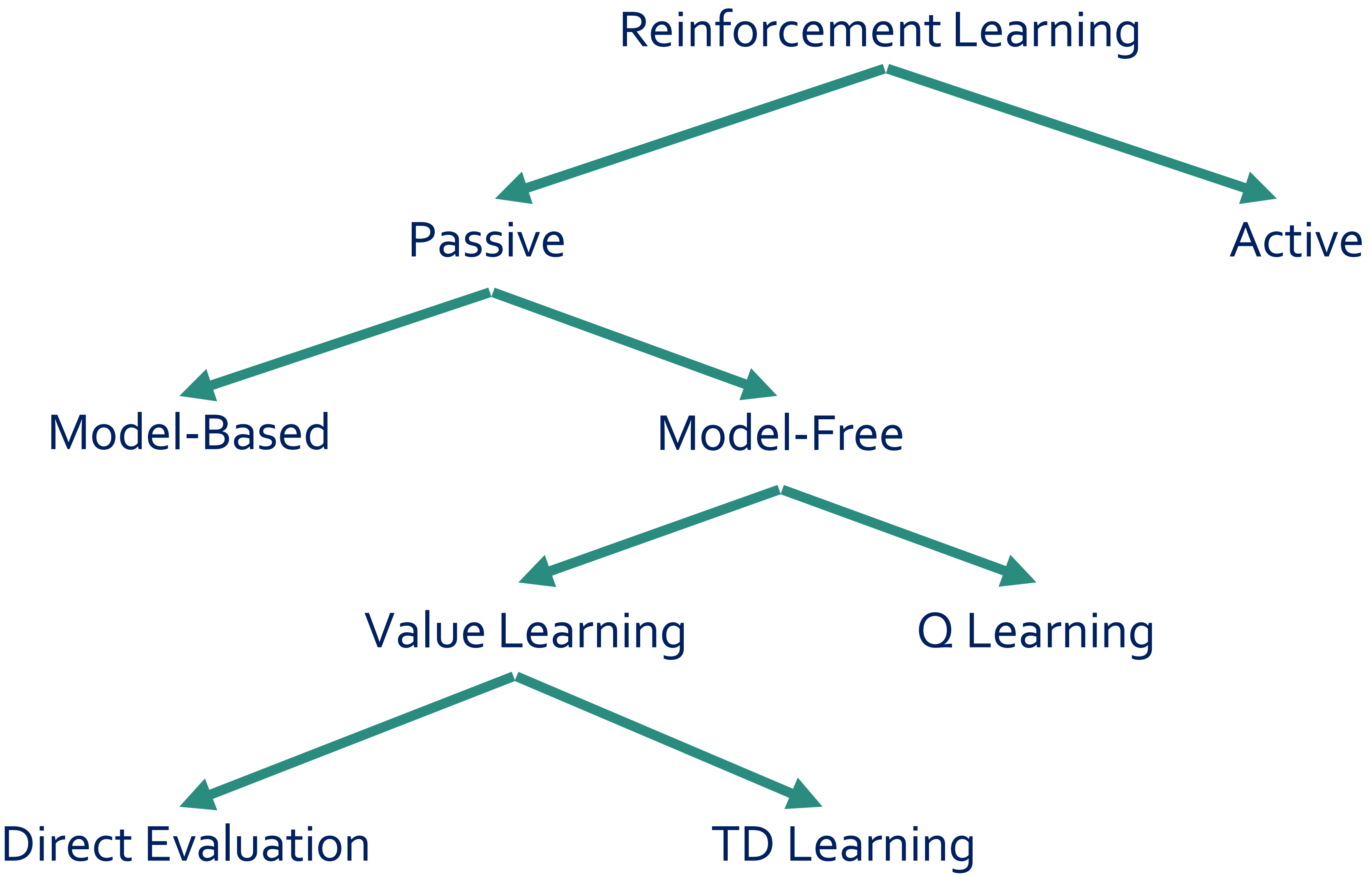
Instructor: Prof. Dr. techn. Wolfgang Nejdl

Leibniz University Hannover

[These slides were created by Dan Klein and Pieter Abbeel for CS188 Intro to AI at UC Berkeley. All materials are available at http://ai.berkeley.edu.]

Co-financed by the Connecting Europe Facility of the European Union

Reinforcement Learning Taxonomy



Co-financed by the Connecting Europe Facility of the European Union

Q-Value Iteration

- Value iteration: find successive (depth-limited) values
  - Start with $V_0(s) = 0$, which we know is right
  - Given V_k, calculate the depth $k+1$ values for all states:

$$V_{k+1}(s) \leftarrow \max_a \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V_k(s')]$$

- But Q-values are more useful, so compute them instead
  - Start with $Q_0(s,a) = 0$, which we know is right
  - Given Q_k, calculate the depth $k+1$ q-values for all q-states:

$$Q_{k+1}(s,a) \leftarrow \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma \max_{a'} Q_k(s',a')]$$

Co-financed by the Connecting Europe Facility of the European Union

Q-Learning

- Q-Learning: sample-based Q-value iteration

$$Q_{k+1}(s,a) \leftarrow \sum_{s'} T(s,a,s')\left[R(s,a,s') + \gamma \max_{a'} Q_k(s',a')\right]$$

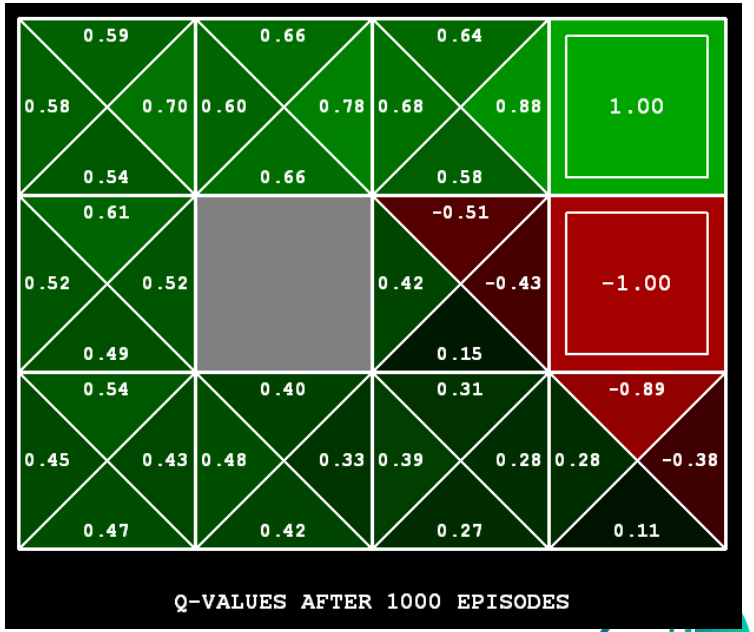
- Learn $Q(s,a)$ values as you go
  - Receive a sample (s, a, s', r)
  - Consider your old estimate: $Q(s,a)$
  - Consider your new sample estimate:

$$\text{sample} = R(s,a,s') + \gamma \max_{a'} Q(s',a')$$

no longer policy evaluation!

  - Incorporate the new estimate into a running average:

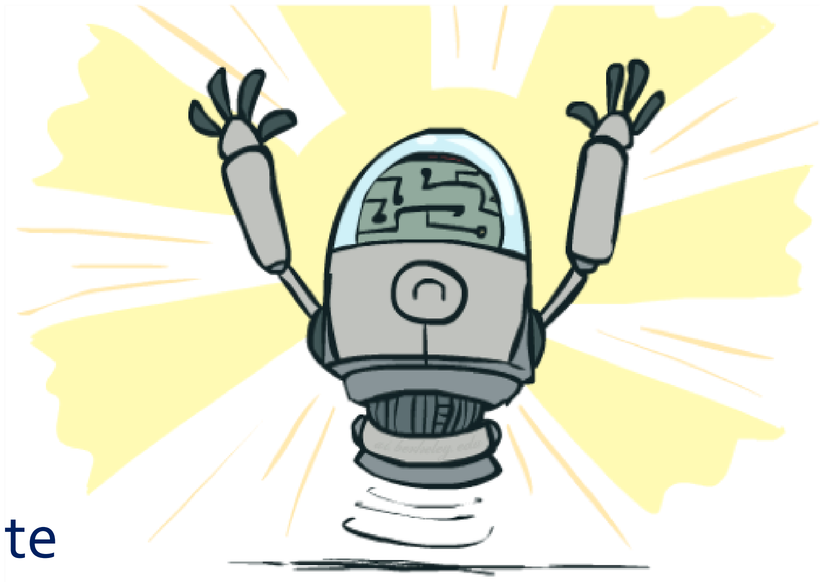
$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha \cdot \text{sample}$$



Co-financed by the Connecting Europe Facility of the European Union

Q-Learning Properties

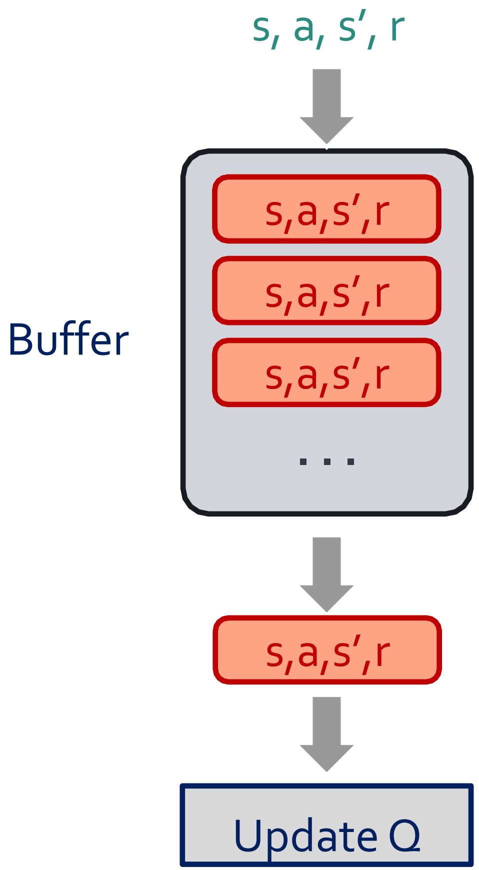
- Amazing result: Q-learning converges to optimal policy -- even if you're acting suboptimally!
- This is called off-policy learning
- Caveats:
  - You have to explore enough
  - You have to eventually make the learning rate small enough
  - … but not decrease it too quickly
  - Basically, in the limit, it doesn't matter how you select actions (!)



Co-financed by the Connecting Europe Facility of the European Union

Q-Learning with a Replay Buffer

- Problem:
  - Need to repeat same (s,a,s',r) transitions in environment many times to propagate values
- Solution:
  - Collect transitions in a memory buffer and "replay" them to update Q values
    - Uses memory of transitions only, no need to repeat them in environment
  - Evidence of such experience replay in the brain



Co-financed by the Connecting Europe Facility of the European Union

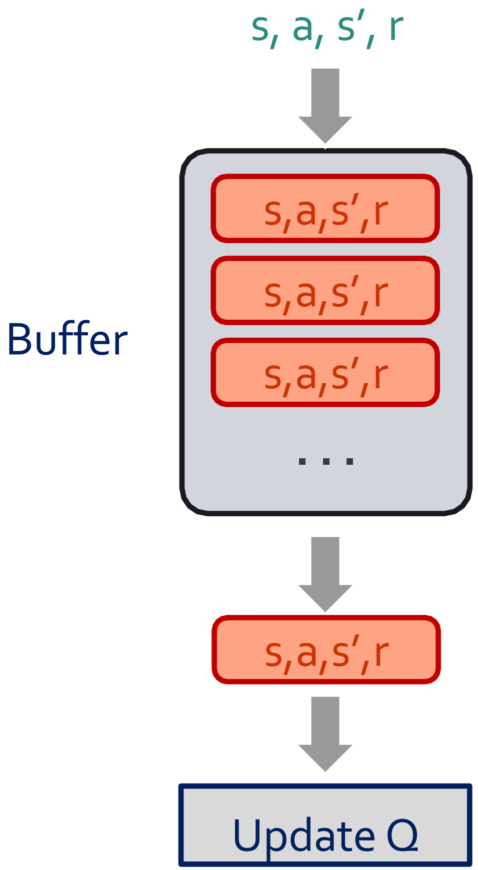
Q-Learning with a Replay Buffer

- At each step:
  - Receive a sample transition (s,a,s’,r)
  - Add (s,a,s’,r) to replay buffer
  - Repeat N times:
    - Randomly pick transition (s,a,s’,r) from replay buffer
    - Make sample based on (s,a,s’,r):

$$\text{sample} = R(s,a,s') + \gamma \max_{a'} Q(s',a')$$

    - Update Q based on picked sample:

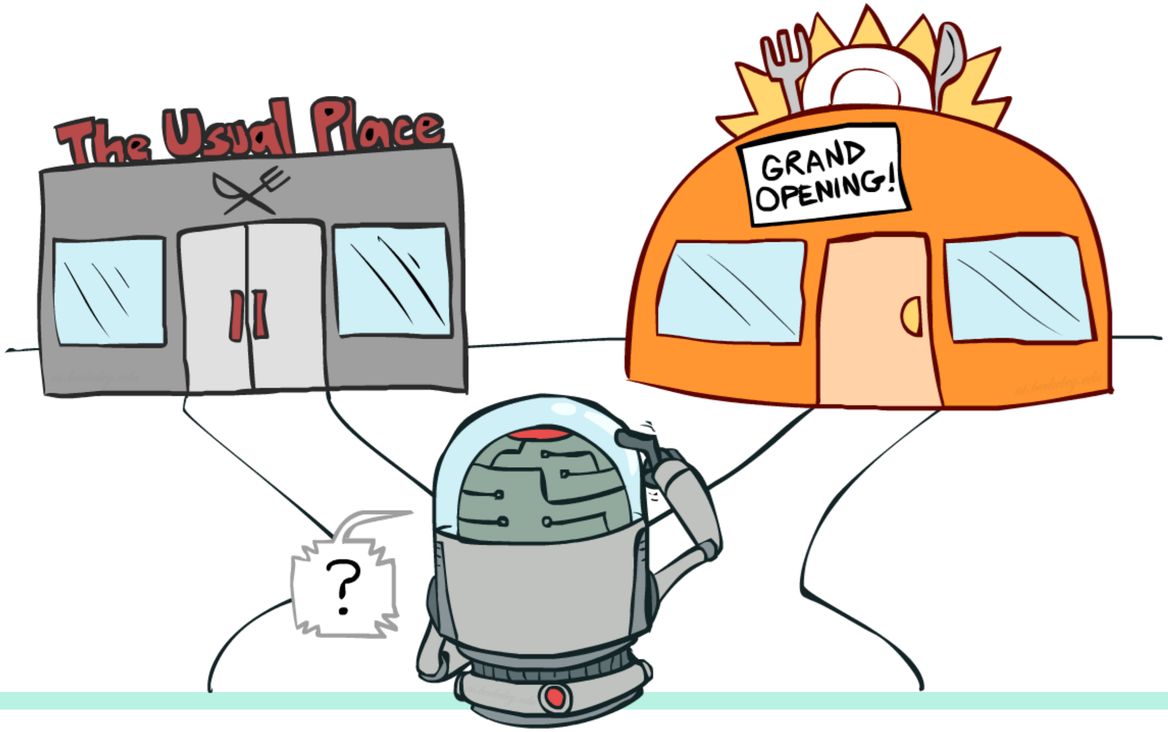
$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha \cdot \text{sample}$$



Co-financed by the Connecting Europe Facility of the European Union

Advanced Topics in AI

Next: Active RL & Exploration vs. Exploitation



Instructor: Prof. Dr. techn. Wolfgang Nejdl

Leibniz University Hannover

[These slides were created by Dan Klein and Pieter Abbeel for CS188 Intro to AI at UC Berkeley. All materials are available at http://ai.berkeley.edu.]

Co-financed by the Connecting Europe Facility of the European Union