Reinforcement Learning, Homework 1

Hannah Min: 11011580, Harm Manders: 10677186

September 2020

2.1 Dynamic Programming

1. Stochastic: $v(s) = \mathbb{E}_{\alpha\sim\pi}[q_\pi(s,a)]$
   Deterministic: $v(s) = q_\pi(s,a), a = \pi(a|s)$

2. Q-value Iteration updates the Q-values for a state-action pair independent of the policy.

$$q_{k+1}(s,a) = \sum_{s',r} p(s',r|s,a)[r + \gamma \max_{a'} q(s',a')]$$

3. The Q-values are weighted by the probabilities $\pi(a|s)$ given by the policy.

$$Q^\pi(a,s) \leftarrow \pi(a|s) \sum_{s',r} p(s',r|s,a)[r + \gamma \sum_{a'} \pi(a'|s')Q(s',a')]$$

4.

$$\pi(s) \leftarrow \arg\max_a \left(\sum_{s',r} p(s',r|s,a)[r + \gamma \cdot \max_{a'} Q(s',a')]) \right)$$

2.2 Coding Assignment - Dynamic Programming

1. Handed in on codegra.de

2. Because value iteration only takes the action into consideration that maximizes the value, it converges faster towards a stable policy. In the example given in the RLLab1, value iteration uses approximately 55 times less backup operations than policy iteration ($4 \cdot \#States$ compared to $220 \cdot \#States$).