Data-driven intelligent systems July 8, 2020

Florian Vahl, Dominik Buchhardt

1 Markov Decision Process

1.1 Description

The Markov decision process is a process where the action of an agent only depends on the current state of the environment. There are is a set of possible actions and a set of possible states. There is also an agent policy which choses an action based on the given state. The agent receives information about the state, performs the action provided by its policy and may receives a reward for the action. The state is updated and executed another time. The MPD has fixed transition probabilities which depend as said only on the current state. The reward probability is also fixed.
The agents policy will be adapted to gain a maximum Return meaning a maximum future reward. The reward is estimated by estimating the Return for the next step and adding the reward for the transition.
The best estimates for this transition are stored as V values depending on the state or Q depending on both the state and the action.

1.2 Example for a MDP

A robot navigates through a maze, the states are the positions and orientations of the robot. The actions are the movement commands. The reward is given based on the proximity to the goal position.

1.3 Example for a POMPD

A real life application where the robot from above is only able to measure a few of the values which are defining its state. Therefore it is only able to approximate a distribution of possible states and apply a more general form of the MDP.

2 Exploration vs. exploitation

In reinforcement learning exists a tradeoff between exploration and exploitation. Exploitation means optimizing the currently best working policy towards a local maximum. Exploration on the other hand means trying out new or worse strategies to explore unknown states and skip local maxima in the best case towards a global maximum.

This tradeoff can be implemented using a ϵ-greedy action selection, where a random action is chosen with a probability of $1 - \epsilon$ for exploration purposes. If no random action is selected the currently best fitting action is chosen.

An agent may be failing or take a very long time to converge if it explores too much. On the other hand, exploiting too much results in a local maximum of the reward, meaning it maybe fails to choose an action that is suboptimal at the moment but leads to a much higher reward later on.