Markov Decision Processes

Andreas Orthey

February 2020

A Markov decision process (MDP) is a tuple (S, A, R, P) with S being the states, A being the actions, rewards $R = R(s, a)$ which you get when in state s and applying action a and a transition probability $P = P(s' \mid s, a)$ which gives the probability of ending in state s' if we are at state s and apply action a. Additional information can be an initial state probability distribution $P(s_0)$, which tells us where we likely will start. If rewards are non-deterministic (same action and state transition might lead to different rewards), you need to replace $R(s, a)$ by a probability distribution $P(r \mid s, a)$. We consider here only *stationary* MDPs, which are independent of time. To define a *non-stationary* MDP, every state, action and reward becomes time dependent, i.e. $s \to s_t$, $a \to a_t$ and $r \to r_t$.

Given an MDP, our goal is to find an optimal policy $\pi^\star(s)$. A policy tells us, in each state, which action to take. An optimal policy gives us the action which will optimize a cost function over rewards. A commonly used cost function is the expected discounted reward over an infinite horizon, which we define as

$$\pi^\star(s) = \underset{\pi}{\operatorname{argmax}}\ E_\pi \left\{ \sum_{i=0}^{\infty} \gamma^i r_i \mid s \right\}. \tag{1}$$

There are two main motivations for this formulation. First, it makes intuitively sense to maximize the cumulative reward over all future actions. Rewards close in time should have a bigger influence because we should try to avoid disastrous outcomes (airplane crashing, humanoid robot falling down). Actions further away in time could potentially be fixed at a later time and are therefore less important. Second, this formulation allows us to write algorithms which can be shown to converge to the optimal policy. This would, however, not work if we use e.g. $\gamma = 1$.

1 Terminal states

Let us assume that we want the MDP to end at some point. This can be modeled by adding a Nirvana state s_N to our MDP, i.e. we increase our state space with $S = S \cup s_N$. This state has a transition probability of

$$P(s'|s = \{s_N, s_T\}, a) = \begin{cases} 1 & \text{if } s' = s_N \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

and returns a reward of $R(s_N, a) = 0$. The total return for the agent is then given by

$$\sum_{t=0}^{\infty} \gamma^t r_t = \sum_{t=0}^{T} \gamma^t r_t + \sum_{t=T}^{\infty} \gamma^t R(s_N, a) = \sum_{t=0}^{T} \gamma^t r_t. \tag{3}$$