Damoun Ayman
Github repository

Reinforcement Learning
November 14, 2020

Dynamic Programming
(Value Iteration, Policy Iteration, Q-Learning)

0.1 Value Iteration

- For each state s , we need to figure out the expected reward of strating in s an acting optimaly.
-

Algorithm 1: Policy Improvement

```
Input: MDP, value function $V$
Output: policy $\pi'$
for $s \in \mathcal{S}$ do
    for $a \in \mathcal{A}(s)$ do
        $Q(s,a) \leftarrow \sum_{s' \in \mathcal{S}, r \in \mathcal{R}} p(s', r|s, a)(r + \gamma V(s'))$
    end
    $\pi'(s) \leftarrow \arg\max_{a \in \mathcal{A}(s)} Q(s,a)$
end
return $\pi'$
```

Algorithm 2: Value Iteration

```
Input: MDP, small positive number $\theta$
Output: policy $\pi \approx \pi_*$
Initialize $V$ arbitrarily (e.g., $V(s) = 0$ for all $s \in \mathcal{S}^+$)
repeat
    $\Delta \leftarrow 0$
    for $s \in \mathcal{S}$ do
        $v \leftarrow V(s)$
        $V(s) \leftarrow \max_{a \in \mathcal{A}(s)} \sum_{s' \in \mathcal{S}, r \in \mathcal{R}} p(s', r|s, a)(r + \gamma V(s'))$
        $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
    end
until $\Delta < \theta$;
$\pi \leftarrow$ Policy_Improvement(MDP, $V$)
return $\pi$
```

- $p(s', r|s, a)$: probability of next state s' and reward r, given current state s and current action a ($\mathbb{P}(S_{t+1} = s', R_{t+1} = r | S_t = s, A_t = a)$)

0.2 Policy Iteration