Policy Evaluation

Dynamic Programming

Marius Lindauer

Leibniz Universität Hannover

Winter Term 2021

Dynamic Programming for Evaluating Value of Policy π

- Initialize $V_0^\pi(s) = 0$ for all $s \in S$
- For $k = 1$ until convergence
  - For all s in S

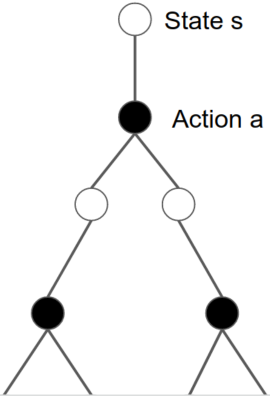
$$V_k^\pi(s) = r(s, \pi(s)) + \gamma \sum_{s' \in S} p(s' \mid s, \pi(s)) V_{k-1}^\pi(s')$$

- $V_k^\pi(s)$ is exact value of k-horizon value of state s under policy π
- $V_k^\pi(s)$ is an estimate of infinite horizon value of state s under policy π

$$V^\pi(s) = \mathbb{E}_\pi[G_t \mid s_t = s] \approx \mathbb{E}[r_t + \gamma V_{k-1} \mid s_t = s]$$

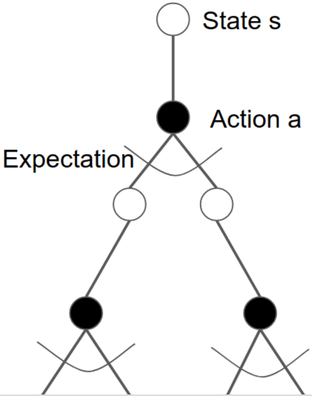
Dynamic Programming for Evaluating Value of Policy π

$$V^\pi(s) = \mathbb{E}_\pi[G_t \mid s_t = s] \approx \mathbb{E}[r_t + \gamma V_{k-1} \mid s_t = s]$$



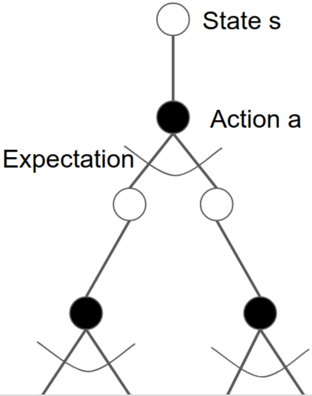
Dynamic Programming for Evaluating Value of Policy π

$$V^{\pi}(s) = \mathbb{E}_{\pi}[G_t \mid s_t = s] \approx \mathbb{E}[r_t + \gamma V_{k-1} \mid s_t = s]$$



Dynamic Programming for Evaluating Value of Policy π

$$V^{\pi}(s) = \mathbb{E}_{\pi}[G_t \mid s_t = s] \approx \mathbb{E}[r_t + \gamma V_{k-1} \mid s_t = s]$$



Policy Evaluation: $V^\pi = \mathbb{E}[G_t \mid s_t = s]$

- $G_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots$ in MDP M under policy π
- Dynamic Programming
  - $V^\pi(s) = \mathbb{E}_\pi[r_t + \gamma V_{k-1} \mid s_t = s]$
  - Requires model of MDP M
  - Bootstraps future return using value estimate
  - Requires Markov assumption: bootstrapping regardless of history

Policy Evaluation: $V^\pi = \mathbb{E}[G_t \mid s_t = s]$

- $G_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots$ in MDP M under policy π
- Dynamic Programming
  - $V^\pi(s) = \mathbb{E}_\pi[r_t + \gamma V_{k-1} \mid s_t = s]$
  - Requires model of MDP M
  - Bootstraps future return using value estimate
  - Requires Markov assumption: bootstrapping regardless of history

- What if we don't know the dynamic model P and/or reward model R? ($\leadsto$ see next videos)