Lecture Notes on

Reinforcement Learning

Dhilan Teeluckdharry
teeluckn@mcmaster.ca

Contents

1 Dynamic Programming

1.1 Policy Improvement Theorem

Consider a policies π, π' s.t. $q_\pi(s, \pi'(s)) \geq v_\pi(s)\ \forall\ s \in \mathbb{S}$

$$v_\pi(s) \leq q_\pi(s, \pi'(s))$$

$$= \mathbb{E}[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s, A_t = \pi'(s)]$$

$$= \mathbb{E}_{\pi'}[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s]$$

$$\leq \mathbb{E}_{\pi'}[R_{t+1} + \gamma q_\pi(S_{t+1}, \pi'(S_{t+1})) \mid S_t = s]$$

$$= \mathbb{E}_{\pi'}[R_{t+1} + \gamma\, \mathbb{E}_{\pi'}[R_{t+2} + \gamma v_\pi(S_{t+2}) \mid S_{t+1}, A_{t+1} = \pi'(S_{t+1})] \mid S_t = s]$$

$$= \mathbb{E}_{\pi'}[R_{t+1} + \gamma R_{t+2} + \gamma^2 v_\pi(S_{t+2}) \mid S_t = s]$$

$$\leq \mathbb{E}_{\pi'}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 v_\pi(S_{t+3}) \mid S_t = s]$$

$$\cdots \leq \mathbb{E}_\pi[\sum_{k=0}^{N} \gamma^k R_{t+k+1} \mid S_t = s] = v_{\pi'}(s) \quad \square$$