goal: characterize solution of finite MDP/SOCP

refs:

Neuro-Dynamic Programming

Dimitri P. Bertsekas and John N. Tsitsiklis

chapter 2

Reinforcement Learning

An Introduction

Richard S. Sutton and Andrew G. Barto

chapter 3

- consider MDP/SOCP $(X, \mathcal{U}, P, c)$ with infinite-horizon exponentially-discounted cost $\min_{u} E[c(x,u)]$ s.t. $x^+ \sim P(x,u)$ where $c(x,u) = \sum_{t=0}^{\infty} \gamma^t \cdot \mathcal{L}(x_t, u_t)$, $\gamma \in (0,1)$

- given policy $\pi : X \to \Delta(\mathcal{U})$, define <u>value function</u> $v^\pi : X \to \mathbb{R}$

$$\forall x \in X: \; v^\pi(x) = E[c(x,u) \mid x_0 = x, \; u_t \sim \pi(x_t)]$$

→ assuming $|X|, |\mathcal{U}| < \infty$, show v^π satisfies Bellman equation

$$\forall x \in X: \; v^\pi(x) = \sum_{u \in \mathcal{U}} \pi(u|x) \sum_{x^+ \in X} P(x^+ | x, u) \cdot \left(\mathcal{L}(x,u) + \gamma \cdot v^\pi(x^+) \right)$$

$$- v^\pi(x) = E\left[\sum_{t=0}^{\infty} \gamma^t \cdot \mathcal{L}(x_t, u_t) \mid x_0 = x, \; u_t \sim \pi(x_t) \right]$$

$$= \mathbb{E}\left[\gamma^0 \cdot \mathcal{L}(x_0, u_0) + \sum_{t=1}^{\infty} \gamma^t \cdot \mathcal{L}(x_t, u_t) \,\middle|\, x_0 = x,\ u_t \sim \pi(x_t),\ x_1 \sim P(x^+ | x_0, u_0) \right]$$

$$= \sum_{u_0 \in \mathcal{U}} \pi(u_0 | x_0) \cdot \mathcal{L}(x_0, u_0) \Big\} = \sum_{u \in \mathcal{U}} \pi(u | x) \sum_{x^+ \in X} P(x^+ | x, u) \cdot \mathcal{L}(x, u)$$

$$+ \sum_{u_0 \in \mathcal{U}} \pi(u_0 | x_0) \cdot \sum_{x^+ \in X} P(x^+ | x_0, u_0) \cdot \underbrace{\mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^t \cdot \mathcal{L}(x_t, u_t) \,\middle|\, x_1 = x^+,\ u_t \sim \pi(x_t) \right]}$$

$$= \gamma \cdot \mathbb{E}\left[\sum_{\tau=0}^{\infty} \gamma^\tau \cdot \mathcal{L}(x_\tau, u_\tau) \,\middle|\, x_0 = x^+,\ u_\tau \sim \pi(x_\tau) \right]$$

$\tau = t - 1 \Rightarrow (t = 1 \Rightarrow \tau = 0)$

$$= \gamma \cdot v^\pi(x^+)$$

$$* \quad \forall x \in X: \ v^\pi(x) = \sum_{u \in \mathcal{U}} \pi(u | x) \sum_{x^+ \in X} P(x^+ | x, u) \cdot \left(\mathcal{L}(x, u) + \gamma \cdot v^\pi(x^+) \right)$$

* NOTE: value $v^\pi : X \to \mathbb{R} \in \mathbb{R}^{|X|}$ appears <u>linearly</u>!

→ determine L^π, b^π so that $L^\pi \cdot v^\pi = b^\pi$ is the Bellman equation

⇝ see Python notebook provided after homework completed

⇒ the value of <u>any</u> policy can be computed by solving a linear equation