model-based solution of MDP / SOCP

goal: derive algorithms to solve MDP/SOCP when model is "known"

refs: *Neuro-Dynamic Programming*

Dimitri P. Bertsekas and John N. Tsitsiklis

chapter 2

- given finite MDP/SOCP $(X, \mathcal{U}, P, c)$ – infinite-horizon, exponentially-discounted

i.e. $\min_{u} E[c(x,u)]$ $\quad c(x,u) = \sum_{t=0}^{\infty} \gamma^t \cdot \mathcal{L}(x_t, u_t)$

s.t. $x^+ \sim P(x,u)$, $|X|, |\mathcal{U}| < \infty$ $\quad \gamma \in (0,1)$

- we can define Bellman operators associated with MDP, $\pi : X \to \Delta(\mathcal{U})$

$T : \mathbb{R}^X \to \mathbb{R}^X$ – nonlinear / piecewise-affine

$$: v \mapsto (Tv)(x) = \min_{u \in \mathcal{U}} \sum_{x^+ \in X} P(x^+ | x, u) \cdot (\mathcal{L}(x,u) + \gamma \cdot v(x^+))$$

$T_\pi : \mathbb{R}^X \to \mathbb{R}^X$ – affine

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

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

* recall that T & T_π are contractions: $\|Tv - Tw\|_\infty \le \gamma \|v - w\|_\infty$

$$\|T_\pi v - T_\pi w\|_\infty \le \gamma \|v - w\|_\infty$$

→ propose a value iteration algorithm that approximates $v^* \, (= Tv^*)$ (discuss computational complexity)

- starting from any $v \in \mathbb{R}^X$, iteratively evaluate (nonlinear) operator $T: v \mapsto Tv \mapsto T^2 v \mapsto \cdots \mapsto T^k v$

* $\|T^k v - v^*\| \le \gamma^k \|v - v^*\|$ so $\lim_{k \to \infty} T^k v = v^*$

- each iteration (evaluation of T) requires $O(|X| \cdot |\mathcal{U}|)$ operations

⇝ given v^*, can determine an optimal deterministic policy $\pi^* : X \to \mathcal{U}$ by evaluating T

→ propose a policy iteration algorithm that computes $v^\pi \, (= T_\pi v^\pi)$ and then approximates π^* (discuss computational complexity)

- given any $\pi : X \to \Delta(\mathcal{U})$, can compute v^π by solving affine equation $v^\pi = T_\pi v^\pi$ in $O(|X|^3)$ operations

actually $|X|^{2 \text{ to } 3}$

- can improve policy with greedy update:

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

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

* it turns out that $\forall x \in X: v^{\pi^+}(x) \leq v^\pi(x)$ and $\exists x^*: v^{\pi^+}(x^*) < v^\pi(x^*)$

<u>and</u> $\pi \longmapsto \pi^+ \longmapsto \pi^{++} \longmapsto \cdots \longmapsto \pi^* \rightsquigarrow v^{\pi^*} = v^*$

i.e. this iteration converges to an optimal (deterministic) policy!

$\rightsquigarrow$ in a <u>finite</u> number of iterations!

↳ follows from the fact that $|U^X| < \infty$, but $|U^X| = |U|^{|X|}$