goal: derive & analyze Bellman operators for MDP/SOCP

refs: *Neuro-Dynamic Programming*

Dimitri P. Bertsekas and John N. Tsitsiklis

chapter 2

- consider finite MDP/SOCP (X, U, P, c) — infinite-horizon exp.-discounted:

$$\min_u \mathbb{E}[c(x,u)] \quad \text{s.t. } x^+ \sim P(x,u) \qquad c(x,u) = \sum_{t=0}^{\infty} \gamma^t \cdot \ell(x_t, u_t)$$

↳ $\gamma \in (0,1)$, i.e. $\gamma < 1$

- given policy $\pi : X \to \Delta(U)$, know that value $v^\pi : X \to \mathbb{R}$ defined by $v^\pi(x) = \mathbb{E}[c(x,u) \mid x_0 \sim x,\ u_t \sim \pi(x_t)]$ satisfies Bellman equation

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

idea: use this equation to define operator $T_\pi : \mathbb{R}^X \to \mathbb{R}^X$, $\mathbb{R}^X = \{v : X \to \mathbb{R}\}$ "value-like" functions

$$: v \to T_\pi v = v^+$$

$: V \to T_\pi V = V'$ "value-like" functions

by $\forall x \in X: (T_\pi V)(x) = \sum_{u \in U} \pi(u|x) \sum_{x^+ \in X} P(x^+|u,x) \cdot \left(\ell(x_t, u_t) + \gamma \cdot V(x^+) \right)$

- given optimal policy $\pi^*: X \to \Delta(U)$, the optimal value $V^*: X \to \mathbb{R}$ satisfies $\forall x \in X: V^*(x) = \sum_{u \in U} \pi^*(u|x) \sum_{x^+ \in X} P(x^+|u,x) \cdot \left(\ell(x_t, u_t) + \gamma \cdot V^*(x^+) \right)$

optimal Bellman equation $\left\{ = \min_{u \in U} \sum_{x^+ \in X} P(x^+|u,x) \cdot \left(\ell(x_t, u_t) + \gamma \cdot V^*(x^+) \right) \right.$

so we can define $T: \mathbb{R}^X \to \mathbb{R}^X: V \mapsto T_V = V^+$

by $\forall x \in X: (TV)(x) = \min_{u \in U} \sum_{x^+ \in X} P(x^+|u,x) \cdot \left(\ell(x_t, u_t) + \gamma \cdot V(x^+) \right)$

→ what kind of operator is T_π? T? (give a simple expression for T_π)

- T_π is affine!
- letting $[A_\pi]_{x,x^+} = \sum_{u \in U} \pi(u|x) \cdot \sum_{x^+ \in X} P(x^+|x,u)$,

$$[b_\pi]_x = \sum_{u \in U} \pi(u|x) \cdot \sum_{x^+ \in X} P(x^+|x,u) \cdot \ell(x,u)$$

we have $T_\pi V = \gamma \cdot A_\pi \cdot V + b_\pi$ → Bellman equation is $T_\pi V^\pi = V^\pi$

- T is nonlinear (and non-affine)

- let $T_\pi^k = \underbrace{T_\pi \circ T_\pi \circ \cdots \circ T_\pi}_{k \text{ times}}$, $T^k = \underbrace{T \circ \cdots \circ T}_{k \text{ times}}$

and endow $\mathbb{R}^X$ with the max norm $\|V\|_\infty = \max_{x \in X} |V(x)|$:

thm: (Bellman operators are contractions — 2.5 in BT96)

$\forall V, W \in \mathbb{R}^X, \pi: X \to \Delta(U): \|TV - TW\|_\infty \leq \gamma \cdot \|V - W\|_\infty$

$\forall v, w \in \mathbb{R}^X$, $\pi : X \to \Delta(\mathcal{U})$: $\|Tv - Tw\|_\infty \leq \gamma \cdot \|v - w\|_\infty$

$\|T_\pi v - T_\pi w\|_\infty \leq \gamma \cdot \|v - w\|_\infty$

cor: (by Banach contraction mapping theorem) assuming $\gamma < 1$:

1°. $\forall v \in \mathbb{R}^X$: $\lim_{k \to \infty} T^k v = v^* = T v^*$ is the optimal value

2°. $\forall v \in \mathbb{R}^X$: $\lim_{k \to \infty} T_\pi^k v = v^\pi = T_\pi v^\pi$ is the value of π

converges at an exponential rate

3°. π is optimal $\iff T_\pi v^* = T v^* = v^*$

→ use these facts to derive algorithms to (approximately) solve MDP/SOCP