

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:

$$\min_u E[c(x, u)]$$

$$\text{s.t. } x^+ \sim P(x, u)$$

$$c(x, u) = \sum_{t=0}^{\infty} \gamma^t \cdot \mathcal{L}(x_t, u_t)$$

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

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

Bellman equation

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

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

$$: V \rightarrow T_{\pi} V = V' \quad \text{"value-like" functions}$$

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

• given optimal policy $\pi^*: X \rightarrow \Delta(\mathcal{U})$, the optimal value $V^*: X \rightarrow \mathbb{R}$

$$\text{satisfies } \forall x \in X: V^*(x) = \sum_{u \in \mathcal{U}} \pi^*(u|x) \sum_{x^+ \in X} P(x^+|u, x) \cdot (\mathcal{L}(x, u) + \gamma \cdot V^*(x^+))$$

$$\text{optimal Bellman equation } \left\{ \begin{array}{l} = \min_{u \in \mathcal{U}} \sum_{x^+ \in X} P(x^+|u, x) \cdot (\mathcal{L}(x, u) + \gamma \cdot V^*(x^+)) \end{array} \right.$$

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

$$\text{by } \forall x \in X: (T V)(x) = \min_{u \in \mathcal{U}} \sum_{x^+ \in X} P(x^+|u, x) \cdot (\mathcal{L}(x, u) + \gamma \cdot V(x^+))$$

→ what kind of operator is T_{π} ? T ? (give a simple expression for T_{π})

- T_{π} is affine!

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

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

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

- T is nonlinear (and non-affine)

$$\text{• let } T_{\pi}^k = \underbrace{T_{\pi} \circ T_{\pi} \circ \dots \circ T_{\pi}}_{k \text{ times}}, \quad T^k = \underbrace{T \circ \dots \circ T}_{k \text{ times}}$$

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)

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

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

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

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

$$1^\circ: \forall v \in \mathbb{R}^X: \lim_{k \rightarrow \infty} T^k v = v^* = Tv^* \text{ is the optimal value}$$

$$2^\circ: \forall v \in \mathbb{R}^X: \lim_{k \rightarrow \infty} T_\pi^k v = v^\pi = T_\pi v^\pi \text{ is the value of } \pi$$

converges
at an
exponential
rate

$$3^\circ: \pi \text{ is optimal} \iff T_\pi v^* = Tv^* = v^*$$

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