policy and value iteration

Bertsekas & Tsitsiklis 1996
Ch 2

goal: intuition & facts about
policy & value iteration

- given MDP $(X, \mathcal{U}, P, c)$, i.e.

  $\min_u E[c]$ s.t. $x^+ \sim P(x,u)$,

  with exponentially-discounted
  infinite-horizon cost

  $c(x,u) = \sum_{s=0}^{\infty} \gamma^s \mathcal{L}(x_s, u_s)$,

  consider the corresponding
  Bellman equation:

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

  - given (non-optimal) value $v : X \to \mathbb{R}$,
    could interpret the right-hand side
    of this equation as determining an
    operator $T : \mathbb{R}^X \to \mathbb{R}^X$ on values:

$$\forall x \in X: (Tv)(x) =$$

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

- similarly, given (non-optimal) policy $\mu : X \to \Delta(\mathcal{U})$, define operator $T_\mu : \mathbb{R}^X \to \mathbb{R}^X$:

$$\forall x \in X: (T_\mu v)(x) =$$

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

→ what kind of operator is T? T_μ?
(provide a simpler expression for T_μ)

- T_μ is affine in $v \in \mathbb{R}^X$
(recall that $\mathbb{R}^X = \{f : X \to \mathbb{R}\}$ is a vector space, so affine is defined)

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

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

we see that $T_\mu v = g_\mu + \alpha \cdot P_\mu \cdot v$

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

- the operators T^k, T_μ^k have nice properties:

<u>lem</u>: (monotinicity; 2.3 in BT96)

for any $v, \bar{v} \in \mathbb{R}^X$ s.t. $v(x) \le \bar{v}(x)$, $x \in X$

we have $(T^k v)(x) \le (T^k \bar{v})(x)$,

$$(T_\mu^k v)(x) \le (T_\mu^k \bar{v})(x)$$

<u>lem</u>: (2.4 in BT96)

$$(T^k(v + \nu \cdot \underline{1}))(x) = (T^k v)(x) + \gamma^k \cdot \nu$$

$$(T_\mu^k(v + \nu \cdot \underline{1}))(x) = (T_\mu^k v)(x) + \gamma^k \cdot \nu$$

→ prove these lem's

- both follow from the fact that P_μ is row-stochastic

- these two properties together give the T's a strong contraction property with respect to max norm $\|v\|_\infty = \max_{x \in X} |v(x)|$

<u>thm</u>: (2.5 in BT96)

given $v, \bar{v} \in \mathbb{R}^X$ and policy $\mu : X \to \Delta(\mathcal{U})$, given $v, \bar{v} \in \mathbb{R}^{\mathcal{X}}$ and policy $\mu : \mathcal{X} \to \Delta(\mathcal{U})$,

$$\|Tv - T\bar{v}\|_\infty \leq \gamma \|v - \bar{v}\|_\infty$$

$$\|T_\mu v - T_\mu \bar{v}\|_\infty \leq \gamma \|v - \bar{v}\|_\infty$$

→ prove this contraction result for T

- letting $m = \max_{x \in \mathcal{X}} |v(x) - \bar{v}(x)|$,

  we have $v(x) - m \leq \bar{v}(x) \leq v(x) + m$

- applying T using lemis above,

  $(Tv)(x) - \gamma \cdot m \leq (T\bar{v})(x) \leq (Tv)(x) + \gamma \cdot m$,

  hence $|(Tv)(x) - (T\bar{v})(x)| \leq \gamma \cdot m$

  $$= \gamma \cdot \|v - \bar{v}\|_\infty$$

(the proof for T_μ is similar, just requires marginalizing over $u \sim \mu$)

- since T, T_μ are contractions, their asymptotic behavior is nice:

<u>prop</u>: (2.6 in BT 96) if $\gamma < 1$:

prop: (2.6 in BT 96) if $\gamma < 1$:

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

2° for any $v \in \mathbb{R}^X$: $\lim_{k\to\infty} T_\mu^k v = v^\mu$
is the unique value satisfying $v^\mu = T_\mu v^\mu$

3° a policy $\mu : X \to \Delta(\mathcal{U})$ is optimal
if and only if $T_\mu v^* = Tv^* \; (= v^*)$
(in which case we'll write $\mu = \mu^*$)

- these facts suggest some straightforward algorithms to compute (or approximate) v^*

→ propose a "value iteration" algorithm that uses the operator T to approximate v^*, and discuss its properties

- starting from any $v \in \mathbb{R}^X$, it's straightforward to evaluate the (nonlinear) operator T on v, yielding $Tv \in \mathbb{R}^X$ that's closer to v^* by yielding $Tv \in \mathbb{R}^X$ that's closer to v^* by a factor α: $\|Tv - v^*\|_\infty \leq \alpha \cdot \|v - v^*\|_\infty$

- guaranteed to converge at an exponential rate; each evaluation of T is $O(|X| \cdot |\mathcal{U}|)$

→ propose a "policy iteration" algorithm that uses T_μ to approximate μ^*, and discuss its properties

- given $\mu : X \to \Delta(X)$, can compute v^μ by solving linear equation:

$$v^\mu = T_\mu v^\mu = g_\mu + P_\mu v^\mu \left(= \lim_{k \to \infty} T_\mu^k v, \text{ any } v \right)$$

- now that we know the value of μ, we can improve the policy:

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

- it turns out this will converge to optimal policy <u>in a finite number of steps</u>!

(but requires solving $|X|$ linear equations, which takes $O(|X|^2)$ to $O(|X|^3)$ ...)

* there are many elaborations / variations on

* there are many elaborations / variations on these simple schemes, but they all rely on contraction properties of Bellman-inspired operators:

- Gauss-Seidel (i.e. asynchronous) value iteration
- multistage look-ahead policy iteration
- modified policy iteration (i.e. combine the two - use a couple of value iterations to approximate v^μ, then improve policy)
- asynchronous modified policy iteration
- linear (i.e. convex) programming

• overall, we're looking at an actor-critic setup: