approximate solution of MDP

goal: characterize performance of
PI algorithms applied to
function approximations

Bertsekas & Tsitsiklis Ch 6

- consider MDP

$$\min_{u} E\left[\sum_{t=0}^{\infty} \gamma^t \cdot c(x_t, u_t)\right]$$

$$\text{s.t. } x^+ \sim P(x,u), \quad x \in X = \mathbb{R}^n, \; u \in \mathcal{U} = \mathbb{R}^m$$

---

greedy policies

- suppose given an approximation
  $\tilde{v}$ of v^* or $\tilde{q}$ of q^*

→ how would you choose policy?
(what information do you need,
and what is computational complexity?)

- greedy policy choice is

$$\tilde{\pi}(x) = \arg\min_{u \in \mathcal{U}} \tilde{q}(x, u; \theta)$$

$$= \operatorname*{argmin}_{u \in \mathcal{U}} \sum_{x^+ \in X} P(x^+ | x, u) \cdot \left(c(x,u) + \gamma \cdot \tilde{v}(x^+; \theta) \right)$$

* note that $\tilde{q}$ allows $\tilde{\pi}$ to be determined without model (P)

* equivalent def of greedy policy in terms of Bellman operators: $T\tilde{v} = T_{\tilde{\pi}}\tilde{v}$

<u>note</u>: if $\mathcal{U} = \mathbb{R}^m$ then $\tilde{\pi}(x)$ obtained by solving NLP
↳ we'll consider role of function approximator for π later on...

- assuming $\tilde{\pi}$ can be computed, alternately improving policy (i.e. computing $\tilde{\pi}$ given $\tilde{v} / \tilde{q}$) and evaluating policy (i.e. computing $\tilde{v}/\tilde{q}$ given $\tilde{\pi}$) determines a <u>policy iteration</u> algo w/ "actor / critic" architecture

* natural questions:

1°. does a given algo converge?

1°. does a given algo converge?

2°. if so, is limit close to v^*/π^*?

- partial answers to (2°) are the following:

<u>prop</u>: (6.1 in B&T 96)

suppose $\|v - v^*\|_\infty = \varepsilon$

if π is greedy policy wrt v

then $\|v^\pi - v^*\|_\infty \leq \frac{2\gamma\varepsilon}{1-\gamma}$

* this bound is tight: ex 6.2 in B&T 96

- Suppose $\{(\tilde{\pi}_k, \tilde{v}_k)\}_{k=1}^\infty$ is a sequence of policies and (approximate) values generated by policy iteration

prop: (6.2 in B&T 96)

- if $\exists\, \delta, \varepsilon > 0$ s.t.

$$\|\tilde{v}_k - v^{\tilde{\pi}_k}\|_\infty \leq \varepsilon,$$

$$\|T_{\tilde{\pi}_{k+1}} \tilde{v}_k - T\tilde{v}_k\|_\infty \leq \delta, \text{ then}$$

$$\limsup_{k\to\infty} \|v^{\tilde{\pi}_k} - v^*\|_\infty \leq \frac{(\delta + 2\cdot\varepsilon\cdot\gamma)}{(1-\gamma)^2}$$

– this bound is tight; see ex 6.4 in B&T 96

approximating values

- suppose π given and we seek to approximate $v^\pi(x)$ by $\tilde{v}^\pi(x;\theta)$
- can use gradient-like "TD(1)" method:

  suppose $x:[0,T] \to X$, $u:[0,T] \to \mathcal{U}$ given
- then
$$\theta^+ = \theta - \alpha \sum_{t=0}^{T} \underbrace{D\tilde{v}^\pi}_{\text{derivative wrt } \theta}(x_t;\theta) \cdot \left(\tilde{v}^\pi(x;\theta) - \sum_{\tau=t}^{T} \gamma^{\tau-t} c(x_\tau, u_\tau) \right)$$

  can be rewritten
$$\theta^+ = \theta + \alpha \sum_{t=0}^{T} D\tilde{v}^\pi(x_t;\theta) \cdot (d_t + d_{t+1} + \cdots + d_{T-1})$$
$$= \theta + \alpha \sum_{t=0}^{T} D\tilde{v}^\pi(x_t;\theta) \sum_{\tau=t}^{T} d_\tau$$
  where $d_t = c(x_t, u_t) + \gamma \cdot \tilde{v}^\pi(x_{t+1};\theta) - \tilde{v}^\pi(x_t;\theta)$

- "TD(λ)" variant:
$$\theta^+ = \theta + \alpha \sum_{t=0}^{T} D\tilde{v}^\pi(x_t;\theta) \sum_{\tau=t}^{T} d_\tau (\gamma \cdot \lambda)^{\tau-t}$$

- performance with $\lambda < 1$ unreliable:

  can fail to converge! but helpful in practice?

can fail to converge! but helpful in practice:
cf ex 6.6 in B&T 96

---

approximating policies

B&T 96: Ch. 6.4 optimistic PI
Ch. 6.2

- we'll now assume it's impractical to solve for greedy π given $\tilde{v}$ or $\tilde{q}$ (too many states and/or P unknown)
- propose to approximate π using $\tilde{\pi} : X \times \underline{\Psi} \to \Delta(U)$

→ how would you determine $\tilde{\pi}$? (what information do you need, and what is computational complexity?)

- solve the optimization problem

$$\min_{\psi \in \Psi} \| \tilde{\pi}_\psi - \pi \|^2$$

- though it seems like we'd need π to solve this problem, can approximate solution (online) using data (stochastic descent)

can solve for best $\tilde{\ }$ ... offline

– can solve for best $\tilde{\pi}$, eg offline, or can incrementally update toward best $\tilde{\pi}$, eg online: B&T Eq. (6.51)

$$\psi^+ = \psi - \alpha \cdot D_\psi \tilde{\pi}(x;\psi) \cdot D_u \sum_{x^+ \in X} P(x^+ | x, u)\left(c(x,u) + \gamma \cdot \tilde{v}(x^+;\theta)\right)$$

$$\theta^+ = \theta - \alpha \cdot D_\theta \tilde{v}(x,\theta) \cdot \left(\tilde{v}(x;\theta) - \gamma \cdot \tilde{v}(x^+;\theta)\right)$$

* $D_u E\left[c(x,u) + \gamma \cdot \tilde{v}(x^+;\theta)\right]$ can be approximated using "log likelihood trick" from 1st lecture!

∘ here's the best we can hope for:

1°. suppose $\tilde{v}^\pi_k \to \tilde{v}^\pi$

and $\|\tilde{v}^\pi - v^\pi\| \le \varepsilon$

2°. suppose $\tilde{\pi}_k \to \pi$ (so $\tilde{v}^\pi_k \to \tilde{v}^\pi$ as in)

then: $\tilde{\pi}$ greedy wrt $\tilde{v}^\pi$

$$\left(T\tilde{v}^\pi = T_\pi \tilde{v}^\pi\right)$$

(b/c PI converged to μ by (2°))

and $v^\pi \le v^* + \dfrac{2 \cdot \varepsilon \cdot \gamma}{1-\gamma}$

(since we know $T^k v^\pi \to v^*$)

- so 1°. & 2°. together seem good,
  but: 2°. is generally not true...
  (see Sec 6.4.2 in B&T 96)