approximate policy iteration

goal: characterize performance of PI algorithms applied to function approximations

Bertsekas & Tsitsiklis Ch 6

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

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

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

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

- use Monte Carlo approximation
  if sum is onerous
- this policy is 1-stage greedy;
  generalizing to t-stage greediness
  yields a stochastic shortest-path
  problem ...

∘ assuming $\tilde{\mu}$ can be computed,
  alternately improving policy
  (i.e. computing $\tilde{\mu}$ given $\tilde{v}$ / $\tilde{q}$)
  and evaluating policy
  (i.e. computing $\tilde{v}$/$\tilde{q}$ given $\tilde{\mu}$)
  determines a policy iteration algo

* prop 6.1 might be the better
  one to discuss ...

prop: (6.2 in B&T 96)

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

values generated by policy iteration

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

$$\| \tilde{v}_k - v^{\tilde{\mu}_k} \|_\infty \le \varepsilon,$$

$$\| T_{\tilde{\mu}_{k+1}} \tilde{v}_k - T \tilde{v}_k \| \le \delta, \text{ then}$$

$$\limsup_{k \to \infty} \| v^{\mu_k} - v^* \| \le \frac{(\delta + 2 \cdot \varepsilon \cdot \gamma)}{(1-\gamma)^2}$$

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

- there are analogous generalizations of TD(λ), BUT:
  * these generalizations can fail to converge for nonlinear approximation architectures!

---

approximating policies

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

- we'll now assume it's impractical to solve for greedy μ given $\tilde{v}$ or $\tilde{q}$

solve for greedy μ given v or q
(too many states and/or P unknown)

- propose to approximate μ using
  $\tilde{\mu} : X \times \Psi \to \Delta(\mathcal{U})$

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

- solve the optimization problem

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

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

[terminology: $\tilde{\mu}$ is an <u>actor</u>,
$\tilde{v} / \tilde{q}$ is a <u>critic</u>]

- can solve for best $\tilde{\mu}$, eg offline, or can incrementally update toward best $\tilde{\mu}$, eg online:

B & T Eq. (6.51)

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

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

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

- there are TD(λ) variants of this as well; same non-convergence issues can arise as above

- here's the best we can hope for:

1°. suppose $\tilde{v}^\mu_k \to \tilde{v}^\mu$

and $\|\tilde{v}^\mu - v^\mu\| \leq \varepsilon$

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

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

$$T\tilde{v}^\mu = T_\mu \tilde{v}^\mu$$

(b/c PI converged to μ by (2°))

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

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

- so 1° & 2° together seem good, but: 2° is generally not true...

but: 2°. is generally not true...

(see Sec 6.4.2 in B&T96)