Ch 4 Notes

* Remember Bellman optimality

$$v_*(s) = \max_a \mathbb{E}\left[R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a\right]$$

$$= \max_a \sum_{s', r} p(s', r \mid s, a)\left[r + \gamma v_*(s')\right]$$

or

$$q_*(s, a) = \mathbb{E}\left[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') \mid S_t = s, A_t = a\right]$$

$$= \sum_{s', r} p(s', r \mid s, a)\left[r + \gamma \max_{a'} q_*(s', a')\right]$$

DP algorithms are ~~turned into~~ obtained by turning ~~these~~ into assignments. (update rules).

* Policy Iteration

Policy Iteration = Policy Evaluation
→ Policy Improvement

* Policy Evaluation. (also called prediction)

Recall

$$v_\pi(s) \doteq \mathbb{E}_\pi [G_t \mid S_t = s]$$

$$= \mathbb{E}_\pi [R_{t+1} + \gamma G_{t+1} \mid S_t = s]$$

$$= \mathbb{E}_\pi [R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s]$$

$$= \sum_a \pi(a|s) \sum_{s',r} p(s', r | s, a) [r + \gamma v_\pi(s')]$$

In principle we can solve as a giant LP!

But we prefer iterative methods.

$$v_{k+1}(s) \doteq \mathbb{E}_\pi \left[R_{t+1} + \gamma v_k(S_{t+1}) \mid S_t = s \right] \quad (3)$$

$$= \sum_a \pi(a|s) \sum_{s', r} p(s', r | s, a) \left[r + \gamma v_k(s') \right]$$

$\forall\, s \in \mathcal{S}$.

Can be shown that $\{v_k\}$ converges as $k \to \infty$ (Iterative policy evaluation)

Policy Improvement

First let's look at Policy Improvement Theorem

Let π and π' be any pair of deterministic policies, s.t. $\forall s \in \mathcal{S}$.

$$q_\pi(s, \pi'(s)) \geq v_\pi(s) \quad (4.7)$$

Then π' must be $\geq \pi$.

$$\text{i.e. } v_{\pi'}(s) \geq v_\pi(s) \quad (4.8)$$

All they are saying is that if 4.7 is true at all states then it must be true that 4.8 is also true. This is because π' starting from any state will hit other states along the way and π' is better than π at all of those states so the sum must be better.

The proof of PIT proceeds by considering a policy π' that is identical to π everywhere except at a state s. So 4.7 holds everywhere except at s. So if $q_\pi(s,a) > v_\pi(s)$ then π' is indeed better.

The proof then shows by repeated applications of 4.7 that $v_\pi(s) \leq v_{\pi'}(s)$

So if you now have a $\pi'(s)$ better everywhere ...

$$\pi'(s) \doteq \arg\max_a q_{\pi}(s,a)$$

$$= \arg\max_a \mathbb{E}\left[R_{t+1} + \gamma v_{\pi}(S_{t+1}) \,\middle|\, S_t = s, A_t = a \right]$$

$$= \arg\max_a \sum_{s',r} p(s', r \mid s, a)\left[r + \gamma v_{\pi}(s') \right]$$

it is just greedy wrt. current q-values.

* Value Iteration

Drawback of PI,

→ Each iteration requires Policy Eval. which itself is a protracted iterative procedure requiring multiple sweeps through the state set.

Turns out we can stop early without losing convergence guarantees of policy iteration.

Important such variation:

Stop policy evaluation after 1-step.

Called as value iteration.

$$v_{k+1}(s) \doteq \max_a \mathbb{E}\left[R_{t+1} + \gamma v_k(S_{t+1}) \mid S_t = s, A_t = a\right]$$

$$= \max_a \sum_{s',r} p(s', r \mid s, a)\left[r + \gamma v_k(s')\right]$$

$$\forall s \in \mathcal{S}.$$

$\{v_k\}$ can be shown to converge.