Review

* Markov decision process (MDP)
  $\{ S, A, P(s'|s,a), R(s), \gamma \}$
  states, actions, transitions, rewards, discount factor

* Policy: mapping $\pi : S \to A$ of states to actions

* Value functions

$$V^{\pi}(s) = E^{\pi}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t) \mid s_0 = s\right] \quad \text{state}$$

$$Q^{\pi}(s,a) = E^{\pi}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t) \mid s_0 = s, a_0 = a\right] \quad \text{action}$$

* Planning - given parameters of MDP what can we compute?

1. Policy evaluation - how to compute $V^{\pi}(s)$?
   Solve linear equations:

$$\sum_{s'} \left[I(s,s') - \gamma P(s'|s, \pi(s)) \right] V^{\pi}(s') = R(s)$$

2. Policy improvement - how to ~~compute~~ improve on $\pi(s)$?
   Greedy policy $\pi'(s) = \operatorname{argmax}_a Q^{\pi}(s,a)$

   Thm: $V^{\pi'}(s) \geq V^{\pi}(s)$ for all states s.

3. Policy iteration - how to compute $\pi^*(s)$?
   Algorithm [NOT the algorithm in HW]
   (1) initialize policy at random
   (2) repeat until convergence
   - compute state & action value function of current policy.
   - derive greedy policy from action value function.

$\pi_0 \xrightarrow{\text{evaluate}} V^{\pi_0}(s) \xrightarrow{\text{improve}} \pi_1 \xrightarrow{\text{evaluate}} V^{\pi_1}(s) \xrightarrow{\text{improve}} \dots$

$Q^{\pi_0}(s,a)$ $\qquad$ $Q^{\pi_1}(s,a)$

* Is policy iteration guaranteed to converge? yes
why? 1) # policies is finite $|A|^{|S|}$
2) cannot cycle b/c policy improvements at every iteration.
Typically converges in far less steps than $|A|^{|S|}$

* Does it always converge to an optimal policy $\pi^*(s)$?
yes.

* thm: suppose $\pi'(s) = \pi(s)$ for all states s, or even more generally, that $V^{\pi'}(s) = V^{\pi}(s)$
Then $V^{\pi}(s) = V^*(s)$

Note: optimal value functions are unique, even if there are many optimal policies.

* Proof strategy:
1) Derive "Bellman optimality eqn" satisfied by $V^{\pi}(s)$ at convergence.
2) Show that $V^{\pi}(s) \geq V^{\tilde{\pi}}(s)$ for all states s and other policies $\tilde{\pi}$. Hence $V^{\pi}(s) = V^*(s)$

Step 1
From Bellman eqn for $\pi'(s)$:
$$V^{\pi'}(s) = R(s) + \gamma \sum_{s'} P(s'|s, \pi'(s)) V^{\pi'}(s')$$
By assumption, $V^{\pi'}(s) = V^{\pi}(s)$ at convergence
$$V^{\pi}(s) = R(s) + \gamma \sum_{s'} P(s'|s, \pi'(s)) V^{\pi}(s')$$

By assumption, $\pi'(s)$ is greedy w.r.t. $V^{\pi}(s)$

Hence:

$$V^{\pi}(s) = R(s) + \gamma \max_a \sum_{s'} P(s'|s,a)\, V^{\pi}(s')$$

↑ different than linear Bellman equation.

"Bellman optimality equation"

(set of n ~~linear~~ non-linear eqns for $s = 1, 2, \ldots, n$ non-linear b/c of max operation)

Step 2

Iterate RHS

$$V^{\pi}(s) = R(s) + \gamma \max_a \sum_{s'} P(s'|s,a) \Big[\underbrace{R(s')}_{V^{\pi}(s')} + \underbrace{\gamma \max_{a'} P(s''|s',a')\, V^{\pi}(s'')}_{V^{\pi}(s')} \Big]$$

iterate again and again:

$$V^{\pi}(s) = R(s) + \gamma \max_a \sum_{s'} P(s'|s,a) \Big[R(s') + \gamma \max_{a'} \sum_{s''} P(s''|s',a') \Big[\quad \Big] \Big]$$

Now show that this iterated expression (taken out to an infinite # terms) implies optimality.

Let $\tilde{\pi}(s)$ be any other policy.

From Bellman eqn:

$$V^{\tilde{\pi}}(s) = R(s) + \gamma \sum_{s'} P(s'|s, \tilde{\pi}(s))\, V^{\tilde{\pi}}(s')$$

$\tilde{\pi}$ be greedy

$$\leq R(s) + \gamma \max_a \sum_{s'} P(s'|s,a)\, V^{\tilde{\pi}}(s')$$

$$= R(s) + \gamma \max_a \sum_{s'} P(s'|s,a) \Big[R(s') + \gamma \sum_{s''} P(s''|s', \tilde{\pi}(s'))\, V^{\tilde{\pi}}(s'') \Big]$$

$$\leq R(s) + \gamma \max_a \sum_{s'} P(s'|s,a) \Big[R(s') + \gamma \max_{a'} \sum_{s''} P(s''|s',a')\, V^{\tilde{\pi}}(s'') \Big]$$

$\tilde{\pi}$ be greedy

Consider upper bound on $V^{\tilde{\pi}}(s)$ from iterating above t times.

Compare to equality for $V^{\pi}(s)$ after iterating t times.

As $t \to \infty$, RHS on upper bound on $V^{\tilde{\pi}}(s)$ converges to RHS of equality for $V^{\pi}(s)$.

Thus as $t \to \infty$

$$V^{\tilde{\pi}}(s) \leq \lim_{t \to \infty} [\quad] = \lim_{t \to \infty} [\quad] = V^{\pi}(s)$$

Thus for all policies $\tilde{\pi}$ and states s, we have $V^{\tilde{\pi}}(s) \leq V^{\pi}(s)$

or: $V^{\pi}(s) = \max_{\tilde{\pi}} V^{\tilde{\pi}}(s) = V^*(s)$

Pros / cons of policy iteration:

(+) converges quickly (in handful of steps often)

(−) each iteration requires policy evaluation $O(n^3)$

4) Value iteration - how to compute $V^*(s)$ directly?

$$V^*(s) = \max_a Q^*(s,a)$$

$$= \max_a \left[R(s) + \gamma \sum_{s'} P(s'|s,a) V^*(s') \right]$$

$$\boxed{V^*(s) = R(s) + \gamma \max_a \sum_{s'} P(s'|s,a) V^*(s')}$$

* n non-linear eqns for $s = 1, 2, \ldots, n$

n unknowns $V^*(s)$

How to solve?

Algorithm : value iteration

(1) initialize $V_0(s) = 0$ for all states s

(2) iterate

$$V_{K+1}(s) = R(s) + \gamma \max_a \left[\sum_{s'} P(s'|s,a)\, V_K(s') \right]$$

for all $s = 1, 2, \ldots, n$

Note : this algorithm works directly on value functions not policies.

But incremental policies can be computed from :

$$\pi_{K+1}(s) = \text{greedy}\left[V_K(s) \right] = \arg\max_a \left[\sum_{s'} P(s'|s,a)\, V_K(s') \right]$$

(3) Suppose this converges :

$$\lim_{K \to \infty} V_K(s) \to V^*(s)$$

compute $$\pi^*(s) = \arg\max_a \left[\sum_{s'} P(s'|s,a)\, V^*(s') \right]$$

Does algorithm converge ? yes.