CSE 150 Lecture 15. 3.8.2010

$\boxed{\text{Review}}$

* Markov decision process (MDP)

  $MDP = \{ \mathcal{S}, A, P(s'|s,a), R(s) \}$

  states, actions, transactions, rewards

* Policy = deterministic mapping $\pi : \mathcal{S} \to A$

* Value functions

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

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

* Policy evaluation

  Solve linear equations :

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

* Policy improvement

  greedy policy $\pi'(s) = \arg\max_a Q^\pi(s,a)$

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

* Policy iteration

  $\pi_0 \xrightarrow{\text{evaluate}} \begin{matrix} V^{\pi_0}(s) \\ Q^{\pi_0}(s,a) \end{matrix} \xrightarrow{\text{improve}} \pi_1 \xrightarrow{\text{evaluate}} \begin{matrix} V^{\pi_1}(s) \\ Q^{\pi_1}(s,a) \end{matrix} \xrightarrow{\text{improve}} \pi_2 \to \cdots$

* Is policy iteration guaranteed to converge? yes
* Does it always converge to an optimal policy π^*? yes.
* Theorem: 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 function $V^*(s)$ is unique, even if there are many optimal policies.

* proof strategy:

1) Derive "Bellman optimality equation" satisfied by $V^{\pi}(s)$ when $V^{\pi'}(s) = V^{\pi}(s)$.

2) Show that $V^{\pi}(s) \geq V^{\tilde{\pi}}(s)$ for all policies $\tilde{\pi}$ and states s in MDP.
   Hence: $V^{\pi}(s) = V^{*}(s)$

<u>Step 1.</u>

From Bellman equation 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

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

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

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

"Bellman optimality equation"

(set of n non-linear equations for $s = 1, 2, \ldots, n$)
non-linear b/c max operation is not linear.

* different than linear Bellman equation

<u>Step 2</u>

Iterate right hand side:

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

Iterate again and again

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

Let $\tilde{\pi}(s)$ be any other policy with Bellman equation:

$$\begin{aligned} V^{\tilde{\pi}}(s) &= R(s) + \gamma \sum_{s'} P(s' \mid s, \tilde{\pi}(s))\, V^{\tilde{\pi}}(s') \\ &\le R(s) + \gamma \max_a \sum_{s'} P(s' \mid s, a)\, V^{\tilde{\pi}}(s') \\ &= R(s) + \gamma \max_a \sum_{s'} P(s' \mid s, a) \left[R(s') + \gamma \sum_{s''} P(s'' \mid s', \tilde{\pi}(s'))\, V^{\tilde{\pi}}(s'') \right] \\ &\le R(s) + \gamma \max_a \sum_{s'} P(s' \mid s, a) \left[R(s') + \gamma \max_{a'} \sum_{s''} P(s'' \mid s', \tilde{\pi}(s'))\, V^{\tilde{\pi}}(s'') \right] \end{aligned}$$

"be greedy"

"use Bellman equation"

greedy

Consider upper bound on $V^{\tilde{\pi}}(s)$ from iterating above t times (being greedy, then applying Bellman equation)

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

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

Thus as $t \to \infty$:

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

Thus for all policies $\tilde{\pi}$ and states s, we have

$$V^{\pi}(s) \ge V^{\tilde{\pi}}(s)$$

$$V^{\pi}(s) = \max_{\tilde{\pi}} V^{\tilde{\pi}}(s) \quad \text{or} \quad V^{\pi}(s) = V^{*}(s).$$

To compute π^*:

$$\begin{aligned} \pi^*(s) &= \arg\max_a Q^*(s,a) \\ &= \arg\max_a \sum_{s'} P(s' \mid s, a)\, V^*(s') \end{aligned}$$

pros/cons of policy evaluation:

(+) Converges quickly (in few steps)

(−) each step requires policy evaluations $O(n^3)$

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 nonlinear equations for n unknowns $V^*(s)$ for $s=1,\ldots,n$
How to solve?

* Algorithm: Value iteration.

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

(2) iterate

current estimate of $V^*(s')$ at k-th iteration

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

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

Note: this algorithm works directly on value functions, no 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: $\lim_{k\to\infty} V_k(s) = V^*(s)$

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

Does algorithm converge?

Clearly, $V^*(s)$ is fixed point of iteration. But are there other fixed points? No. Does it always reach $V^*(s)$? yes.

* Lemma:

for any functions $f(a)$ and $g(a)$:

$$\left|\max_a f(a) - \max_a g(a)\right| \leq \max_a \left|f(a) - g(a)\right|$$

Proof of lemma:

for all a: $f(a) - \max_{a'} g(a') \leq f(a) - g(a)$

max over a: $\max_a f(a) - \max_a g(a) \leq \max_a \left[f(a) - g(a)\right]$

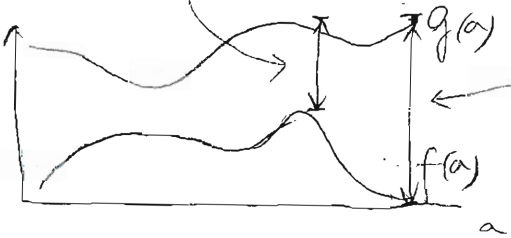
$$\leq \max_a \left|f(a) - g(a)\right|$$

By symmetry, exchanging $f \longleftrightarrow g$ everywhere:

$$\max_a g(a) - \max_a f(a) \leq \max_a \left|g(a) - f(a)\right|$$

Combining last two inequalities:

$$\left|\max_a f(a) - \max_a g(a)\right| \leq \max_a \left|g(a) - f(a)\right|$$



Thm: Value iteration converges.

$$\lim_{k\to\infty}\left[V_k(s)\right] \to V^*(s) \text{ for all states } s$$

Proof: let $\Delta_k = \max_s \left|V_k(s) - V^*(s)\right|$ error at k-th iteration

$$\Delta_{k+1} = \max_s \left|V_{k+1}(s) - V^*(s)\right|$$

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

plug in def for value iteration and Bellman optimality equation for $V^*(s)$.

(Cont.)

$$\Delta_{k+1} = \gamma \max_s \Big| \max_a \underbrace{\Big(\sum_{s'} P(s'|s,a) V_k(s')\Big)}_{f(a)} - \max_a \underbrace{\Big(\sum_{s'} P(s'|s,a) V^*(s')\Big)}_{g(a)} \Big|$$

apply lemma: $\left|\max_a f(a) - \max_a g(a)\right| \le \max_a \left|f(a) - g(a)\right|$

$$\Delta_{k+1} \le \gamma \max_s \max_a \Big| \sum_{s'} P(s'|s,a)\big[V_k(s') - V^*(s')\big] \Big|$$

$$\le \gamma \max_s \max_a \Big| \sum_{s'} P(s'|s,a) \Big(\max_{s''} \big|V_k(s'') - V^*(s'')\big|\Big) \Big|$$

$$= \gamma \max_s \max_a \Big| \underbrace{\Big\{ \sum_{s'} P(s'|s,a) \Big\}}_{1} \Delta_k \Big|$$ Worst case bound on difference.

$$= \gamma \Delta_k \max_s \max_a (1)$$

$$= \gamma \Delta_k$$

Hence: $\Delta_{k+1} \le \gamma \Delta_k$

By iteration: $\Delta_k \le \gamma^k \Delta_0 \to 0$ as $k \to \infty$ for $\gamma < 1$

Assume rewards are bounded:

$$\Delta_0 = \max_s \left|V_0(s) - V^*(s)\right| = \max_s \left|V^*(s)\right|$$

$$\le \Big[\max_s |R(s)|\Big] (1 + \gamma + \gamma^2 + \gamma^3 + \cdots)$$

$$= \max_s |R(s)| \frac{1}{1-\gamma}$$

Thm: $\Delta_k \le \left(\frac{\gamma^k}{1-\gamma}\right) \max_s |R(s)| \to 0$ as $k \to \infty$

Convergence rate depends on γ.

Suggests that more iterations are required as $\gamma \to 1$.