11/18 **Review**

* Markov decision process

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

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

* Value functions

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

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

* Policy evaluation

Solve linear equations: $\sum_{s'} [I(s,s') - \gamma P(s'|s, \pi(s))] V^\pi(s') = R(s)$

* Policy improvement

Greedy policy $\pi'(s) = \underset{a}{\text{argmax}}\ Q^\pi(s,a)$

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

* Policy iteration

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

Thm: if $\pi'(s) = \underset{a}{\text{argmax}}\ Q^\pi(s,a)$ and $V^{\pi'}(s) = V^\pi(s)$ for all s, then $V^\pi(s) = V^*(s)$.

Proof: 1) From last time:

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

2) Iterate RHS:

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

Imagine iterating above t times.

Let's show that this iterated expression implies optimality.

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

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

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

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

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

$= \cdots \leq \cdots = \cdots$: (A)

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

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

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

Thus, as $t \to \infty$: $V^{\tilde{\pi}}(s) \le \lim_{t\to\infty}[\text{(A)}] = \lim_{t\to\infty}[\text{(B)}] = V^{\pi}(s)$

Thus, for all policies $\tilde{\pi}(s)$ and states s: $V^{\pi}(s) \ge V^{\tilde{\pi}}(s)$

$$V^{\pi}(s) = \max_{\tilde{\pi}} V^{\tilde{\pi}}(s) \longrightarrow V^{\pi}(s) = V^*(s).$$

Finally, once you have $V^*(s)$,

$$\pi^*(s) = \arg\max_a Q^*(s,a)$$
$$= \arg\max_a \sum_{s'} P(s'|s,a)\, V^*(s')$$

* Pros & Cons of policy iteration.

(+) converges quickly in finite # steps

(-) requires policy evaluation $O(n^3)$ at each iteration.

Value iteration

* How to compute $V^*(s)$ directly?

* Bellman optimality equation.

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

Before we showed that if $V^{\pi}(s)$ satisfied (*), then π is optimal.

Here, we show that $V^*(s)$ must satisfy (*).

Q: How to solve this set of N nonlinear equations for N unknowns?

* Algorithm

1) initialize: $V_0(s) = 0$ for all 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 $s = 1, 2, \cdots, n$

($V_k(s')$: estimate at k^{th} iteration)

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

But, incremental policies can be computed as:

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

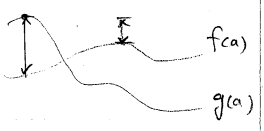
3) after convergence:

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

Does algorithm converge? $V^*(s)$ is obviously a fixed point.

But are there others? Does it always reach $V^*(s)$?

* Lemma: $\left| \max_a [f(a)] - \max_a [g(a)] \right| \le \max_a |f(a) - g(a)|$



Proof of lemma:

For all a: $f(a) - \max_{a'} g(a') \le f(a) - g(a)$

Max over a: $\max_a f(a) - \max_{a'} g(a') \le \max_a [f(a) - g(a)] \le \max_a |f(a) - g(a)|$

By symmetry: $\max_a g(a) - \max_{a'} f(a') \le \max_a |f(a) - g(a)|$.

Combining last two bounds proves lemma.

* Theorem : value iteration converges.

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

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

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

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

(from value iteration; from Bellman optimality)

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

Apply lemma

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

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

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

$$= \gamma \max_s \max_a \Big| \sum_{s'} P(s'|s,a) \Delta_k \Big|$$

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

Hence : $\Delta_{k+1} \le \gamma \Delta_k$ with $\gamma < 1$. so-called "contraction" mapping.

By iteration : $\Delta_k \le \gamma^k \Delta_0 \longrightarrow 0$ as $k \to \infty$.

Assume rewards are bounded : $\Delta_0 = \max_s |V_0(s) - V^*(s)|$

$$= \max_s |V^*(s)|$$

$$\le \max_s |R(s)| (1 + \gamma + \gamma^2 + \cdots)$$

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

More iterations required as $\gamma \to 1$.