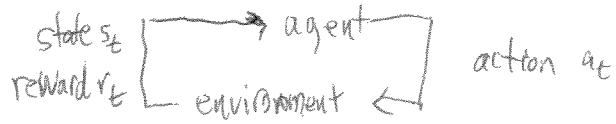
Review

- Reinforcement Learning



- Markov Decision Process (MDP): assume a complete model of fully observed, stochastic environment

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

S: states, A: actions, $P(s'|s,a)$: transitions, $R(s)$: rewards

- Policy: $\pi: S \to A$

- Long-term discounted return $= \sum_{t=0}^{\infty} \gamma^t r_t$

- State value function

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

$$V^{\pi}(s) = R(s) + \gamma \sum_{s'} P(s'|s, \pi(s)) V^{\pi}(s') \quad \star \text{ Bellman eq.}$$

- Action value function

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

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

- Value functions $\hookrightarrow$ policies

Given $\pi^*(s)$:

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

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

Given $V^*(s)$, $Q^*(s,a)$:

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

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

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

- Planning under uncertainty

How to compute $\pi^*(s)$, or equivalently, $V^*(s)$ or $Q^*(s,a)$?

1) Policy evaluation: compute $V^\pi(s)$
2) Policy improvement: compute π' s.t. $V^{\pi'}(s) \geq V^\pi(s)$ $\forall s$
3) Policy iteration: compute π^*
4) Value iteration: compute $V^*(s)$ directly

1) Policy evaluation

System of n linear equations for n unknowns

$$\sum_{s'} \left[I(s,s') - \gamma P(s'|s,\pi(s))\right] \cdot V^\pi(s') = R(s) \qquad s = 1 \ldots n$$

$\Rightarrow$ $(I - \gamma P)V = R$ ex $\left[\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \gamma \begin{pmatrix} P_{00} & P_{01} \\ P_{10} & P_{11} \end{pmatrix}\right] \begin{bmatrix} V_0 \\ V_1 \end{bmatrix} = \begin{bmatrix} r_0 \\ r_1 \end{bmatrix}$

$\Rightarrow$ $V^\pi = (I - \gamma P^\pi)^{-1} R$

2) Policy improvement

- How to compute π' such that $V^{\pi'}(s) \geq V^{\pi}(s)$ for all states s?
- Recall $Q^{\pi}(s,a)$ = "expected return from state s, follow action a, then follow policy π"

How to compute $Q^{\pi}(s,a)$?

- Evaluate policy to get $V^{\pi}(s)$
- $Q^{\pi}(s,a) = R(s) + \gamma \sum_{s'} P(s'|s,a) V^{\pi}(s')$

Define "greedy policy":

$$\pi'(s) = \underset{a}{\operatorname{argmax}} \; Q^{\pi}(s,a)$$

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

$$= \underset{a}{\operatorname{argmax}} \left[\sum_{s'} P(s'|s,a) V^{\pi}(s') \right]$$

Theorem: greedy policy π' everywhere performs better or equal to original policy π.

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

Intuition: ~~greedy policy π' everywhere performs better or equal to original policy π~~

if better to choose action a in state s, then follow π, it's always better to choose action a in state s.

Proof: $V^{\pi}(s) = Q^{\pi}(s, \pi(s))$

$\leq \max_a Q^{\pi}(s, a)$

$= Q^{\pi}(s, \pi'(s))$ by definition of greedy policy

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

So far, it is better to take one step under π', then revert to π, than to follow π.

"One-step inequality": $V^{\pi}(s) \leq R(s) + \gamma \sum_{s'} P(s' | s, \pi'(s)) V^{\pi}(s')$

Apply inequality to $V^{\pi}(s')$ on RHS: ↙ $V^{\pi}(s')$

$$V^{\pi}(s) \leq R(s) + \gamma \sum_{s'} P(s' | s, \pi'(s)) \cdot \left[R(s') + \gamma \sum_{s''} P(s'' | s', \pi'(s')) \cdot V^{\pi}(s'') \right]$$

So, better to take two steps under π', then revert to π, than to always follow π.

In general, apply "one-step inequality" t times, we will show:
better to take $t+1$ steps under π', then follow π,
than to always follow π.

Let $t \to \infty$, it's always better to follow $\pi'(s)$ than $\pi(s)$.

$\Rightarrow V^{\pi}(s) \leq V^{\pi'}(s)$ since RHS converges to $V^{\pi'}(s)$ for $\gamma = 1$.

3) Policy iteration: how to compute $\pi^*(s)$?

Algorithm:

(1) initialize policy π_0 at random

(2) repeat until convergence:

- compute state + action value function of current policy (policy evaluation)
- derive greedy policy from action value function (policy improvement)

$\pi_0 \xrightarrow{\text{evaluate}} V^{\pi_0}(s),\ Q^{\pi_0}(s,a) \xrightarrow{\text{improve}} \pi_1 \xrightarrow{\text{evaluate}} V^{\pi_1}(s),\ Q^{\pi_1}(s,a) \xrightarrow{\text{improve}} \cdots$

- Two questions: (i) does policy iteration converge at all?
  (ii) if so, what does it converge to?

(i) Is policy iteration guaranteed to converge? Yes.

why? 1) cannot cycle b/c policy improves at every iteration
2) # policies is finite: $|A|^{|S|}$

Typically converges in far fewer steps than $|A|^{|S|}$.

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

Thm: Suppose that we've converged at iteration k: $V^{\pi_{k+1}}(s) = V^{\pi_k}(s)$.
Then $V^{\pi_k}(s) = V^*(s)$.

(Note: optimal value function is unique, even if there are many optimal policies.)

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

Step 1

From Bellman eqn for $\pi_{k+1}(s)$:

$$V^{\pi_{k+1}}(s) = R(s) + \gamma \sum_{s'} P(s'|s, \pi_{k+1}(s)) V^{\pi_{k+1}}(s')$$

By assumption: $V^{\pi_{k+1}}(s) = V^{\pi_k}(s)$ at convergence.

$$V^{\pi_k}(s) = R(s) + \gamma \sum_{s'} P(s'|s, \pi_{k+1}(s)) V^{\pi_k}(s')$$

By assumption: $\pi_{k+1}(s)$ is greedy w.r.t. $V^{\pi_k}(s)$

Hence: $$V^{\pi_k}(s) = R(s) + \gamma \max_a \sum_{s'} P(s'|s,a) V^{\pi_k}(s')$$ "Bellman optimality eqn"

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

Step 2

Iterate RHS

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

($V^{\pi_k}(s')$ ↙)

denote @ for RHS

Iterate again + again.

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

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

From Bellman eqn:

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

↓ be greedy

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

↓ use Bellman eq.

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

↓ be greedy

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

denote [$] for RHS

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

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

As $t \to \infty$, [$] converges to [@].

↙ upper bound on $V^{\tilde{\pi}}(s)$

↘ RHS of equality for $V^{\pi_k}(s)$

Thus, as $t \to \infty$:

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

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

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

To compute π^*: $\pi^*(s) = \operatorname{argmax}_a Q^*(s,a) = \operatorname{argmax}_a \sum_{s'} P(s'|s,a) V^*(s)$.

Pros / cons of policy iteration:

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

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