687 2017-10-03

5-minute quiz.

Today: Solving the Bellman Optimality Equ more efficiently
(will use Dynamic Programming)

Policy Evaluation
- Given a policy π, find v^π
- Assume $P + R$ are known

Method: Solve Bellman Equ

$$v^\pi(s) = \sum_a \pi(s,a) \sum_{s'} P(s,a,s')\left[R(s,a,s') + \gamma v^\pi(s')\right]$$

Dynamic Programming

Sequence of value funcs that approximate v^π:

$\hat{v}_0^\pi, \hat{v}_1^\pi, \hat{v}_2^\pi, \hat{v}_3^\pi, \ldots$

$\downarrow$

Choose arbitrarily, e.g. 0 in every state. (Must be 0 for terminal states.)

Do: $\hat{v}_{k+1}^\pi(s) \leftarrow \sum_a \pi(s,a) \sum_{s'} P(s,a,s')\left[R(s,a,s') + \gamma \hat{v}_k^\pi(s')\right]$

Properties

1. $\hat{v}_k^\pi = v^\pi$ is a fixed point
2. $\hat{v}_k^\pi$ converges to v^π as $k \to \infty$ for finite MDPs with bounded rewards
3. One pass over the state space a full backup. A single state update is a backup.

Try it out!

↓	↓	↓	↓
↓	↓	↓	↓
↓	↓	↓	↓
→	→	→	G → absorb

a "goal"

- $R = -1$ always
- $\gamma = 1$
- Actions succeed (except at a wall)
- π as in the arrows above

$\hat{V}_0^\pi =$

0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0

$\hat{V}_1^\pi =$

-1	-1	-1	-1
-1	-1	-1	-1
-1	-1	-1	-1
-1	-1	-1	0

$\hat{V}_2^\pi =$

-2	-2	-2	-2
-2	-2	-2	-2
-2	-2	-2	-1
-2	-2	-1	0

$\hat{V}_3^\pi =$

-3	-3	-3	-3
-3	-3	-3	-2
-3	-3	-2	-1
-3	-2	-1	0

$\hat{V}_4^\pi$

-4	-4	-4	-3
-4	-4	-3	-2
-4	-3	-2	-1
-3	-2	-1	0

$\hat{V}_5^\pi =$

-5	-5	-4	-3
-5	-4	-3	-2
-4	-3	-2	-1
-3	-2	-1	0

$\hat{V}_6^\pi =$

-6	-5	-4	-3
-5	-4	-3	-2
-4	-3	-2	-1
-3	-2	-1	0

Information flows backwards from the "goal", hence the term backup.

- To speed this up, can do an in-place state update.
- Can update in any order.
- Can do updates asynchronously
  - ↪ Can update one state multiple times before updating some other state.

⇓

Guaranteed to converge to v^π if no state is starved for updates.

Can we do the same thing for q? Yes.

Policy Improvement

Q-update: $\hat{q}^{\pi}_{k+1}(s,a) = \sum_{s'} P(s,a,s')\left[R(s,a,s') + \gamma \sum_{a'} \pi(s',a')\hat{q}^{\pi}_{k}(s',a')\right]$

If we have estimated q^{π}, how can we improve π? What if we are just greedy?

Policy Improvement Thm:

Let π + π' be deterministic policies s.t. $\forall s$: $q^{\pi}(s,\pi'(s)) \geq v^{\pi}(s)$.

$= q^{\pi}(s,\pi(s))$ for a deterministic π.

Then $\pi' \geq \pi$. (Recall: $\pi' \geq \pi \triangleq \forall s.\ v^{\pi'}(s) \geq v^{\pi}(s)$)