

5-minute quiz.

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

Policy Evaluation

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

Method: Solve Bellman Eqn

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

Dynamic Programming

Sequence of value funcs that approximate v^π :

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

↓

Choose arbitrarily, e.g.

0 in every state. (Must be 0 for terminal states.)

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

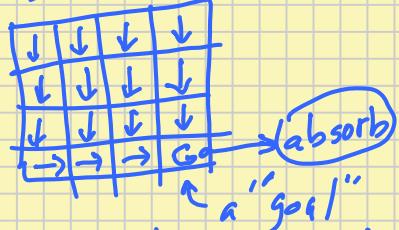
③ One pass over the state space a full backup.
A single state update

Properties

① $\hat{v}_k^\pi = v^\pi$ is a fixed point

② \hat{v}_k^π converges to v^π as $k \rightarrow \infty$ for finite MDPs with bounded rewards

Try it out!



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

$$\hat{V}_0^\pi = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\hat{V}_3^\pi = \begin{bmatrix} -3 & -3 & -3 & -3 \\ -3 & -3 & -3 & -2 \\ -3 & -3 & -2 & -1 \\ -3 & -2 & -1 & 0 \end{bmatrix}$$

$$\hat{V}_1^\pi = \begin{bmatrix} -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & 0 \end{bmatrix}$$

$$\hat{V}_4^\pi = \begin{bmatrix} -4 & -4 & -4 & -3 \\ -4 & -4 & -3 & -2 \\ -4 & -3 & -2 & -1 \\ -3 & -2 & -1 & 0 \end{bmatrix}$$

$$\hat{V}_2^\pi = \begin{bmatrix} -2 & -2 & -2 & -2 \\ -2 & -2 & -2 & -2 \\ -2 & -2 & -2 & -1 \\ -2 & -2 & -1 & 0 \end{bmatrix}$$

$$\hat{V}_5^\pi = \begin{bmatrix} -5 & -5 & -4 & -3 \\ -5 & -4 & -3 & -2 \\ -4 & -3 & -2 & -1 \\ -3 & -2 & -1 & 0 \end{bmatrix}$$

- 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.

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

$$\hat{V}_6^\pi = \begin{bmatrix} -6 & -5 & -4 & -3 \\ -5 & -4 & -3 & -2 \\ -4 & -3 & -2 & -1 \\ -3 & -2 & -1 & 0 \end{bmatrix}$$

Policy Improvement

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

If we have estimated \hat{q}^{π} , how

Can we improve π ? What if
we are just greedy?

Policy Improvement Thm:

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

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

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