Policy Iterations

Ashwin

University of Colorado Boulder

ashwin.asokan@colorado.edu

April 15, 2017

Improve Policy by Evaluation

- Start with a policy π
- Evaluate the policy π
  $V_\pi(s) = E[R_{t+1} + \gamma R_{t+2} + ...|S_t = s]$
- Improve the policy by acting greedily with respect to V_π
  $\pi' = greedy(V_\pi)$
- Repeat the above steps till policy converges.

Policy Iteration Algorithm

Algorithm 1: Policy Iterations

```
Initialize a policy π, compute V_π;
for iteration=1,2... do
    Evaluate Policy V_π = π(S)
    Set New Policy to be the greedy policy for V_π
        π(s) = Max_a E_{s'|s,a}[r + γ V_π(s')]
end
```

Initialize a policy π, compute V_π;
for *iteration=1,2...* **do**
  Evaluate Policy $V_\pi = \pi(S)$
  Set New Policy to be the greedy policy for V_π
  $\pi(s) = Max_a E_{s'|s,a}[r + \gamma V_\pi(s')]$
end

GridWorld

G	1	2	3
4	5	6	7
8	9	10	11
12	13	14	G

- Two Terminal Goal States G
- 14 Non Terminal States 1,2...14
- Possible Action UP,DOWN,LEFT,RIGHT
- R = 0 for Goal State. R = -1 for every other possible state transition
- No Discounts ($\gamma = 1$)
- $V_{i+1}(s) = \sum_{a \in A} \pi(a|s) \langle R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a V_k(s')$
- $V_{i+1} = R_\pi + \gamma P_\pi V_i$

Iterative Policy Evaluation (1)

$V_{i,Random}$ $\qquad$ $\Pi_{Greedy}(V_i)$

I $= 0$

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

G	$\leftarrow\updownarrow\rightarrow$	$\leftarrow\updownarrow\rightarrow$	$\leftarrow\updownarrow\rightarrow$
$\leftarrow\updownarrow\rightarrow$	$\leftarrow\updownarrow\rightarrow$	$\leftarrow\updownarrow\rightarrow$	$\leftarrow\updownarrow\rightarrow$
$\leftarrow\updownarrow\rightarrow$	$\leftarrow\updownarrow\rightarrow$	$\leftarrow\updownarrow\rightarrow$	$\leftarrow\updownarrow\rightarrow$
$\leftarrow\updownarrow\rightarrow$	$\leftarrow\updownarrow\rightarrow$	$\leftarrow\updownarrow\rightarrow$	G

$V_1(i,j) = \frac{1}{4}(R(i,j-1) + (\gamma)(V_0(i,j-1))) + \frac{1}{4}(R(i-1,j) + (\gamma)(V_0(i-1,j))) + \frac{1}{4}(R(i,j+1) + (\gamma)(V_0(i,j+1))) + \frac{1}{4}(R(i+1,j) + (\gamma)(V_0(i+1,j)))$

I $= 1$

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

G	$\leftarrow$	$\leftarrow\updownarrow\rightarrow$	$\leftarrow\updownarrow\rightarrow$
$\uparrow$	$\leftarrow\updownarrow\rightarrow$	$\leftarrow\updownarrow\rightarrow$	$\leftarrow\updownarrow\rightarrow$
$\leftarrow\updownarrow\rightarrow$	$\leftarrow\updownarrow\rightarrow$	$\leftarrow\updownarrow\rightarrow$	$\downarrow$
$\leftarrow\updownarrow\rightarrow$	$\leftarrow\updownarrow\rightarrow$	$\rightarrow$	G

$V_1(1,2) =$
$\frac{1}{4}(-1 + (1)(0)) + \frac{1}{4}(-1 + (1)(0)) + \frac{1}{4}(-1 + (1)(0)) + \frac{1}{4}(-1 + (1)(0)) = -1$

Iterative Policy Evaluation (2)

$V_{i,Random}$ $\qquad\qquad$ $\Pi_{Greedy}(V_i)$

I = 2

0	-1.7	-2	-2
-1.7	-2	-2	-2
-2	-2	-2	-1.7
-2	-2	-1.7	0

G	$\leftarrow$	$\leftarrow$	$\leftarrow\updownarrow\rightarrow$
$\uparrow$	$\leftarrow\uparrow$	$\leftarrow\updownarrow\rightarrow$	$\downarrow$
$\uparrow$	$\leftarrow\updownarrow\rightarrow$	$\downarrow\rightarrow$	$\downarrow$
$\leftarrow\updownarrow\rightarrow$	$\rightarrow$	$\rightarrow$	G

$$V_3(i,j) = \frac{1}{4}(R(i,j-1) + (\gamma)(V_2(i,j-1))) + \frac{1}{4}(R(i-1,j) + (\gamma)(V_2(i-1,j))) + \frac{1}{4}(R(i,j+1) + (\gamma)(V_2(i,j+1))) + \frac{1}{4}(R(i+1,j) + (\gamma)(V_2(i+1,j)))$$

I = 3

0	-2.4		
		-2.9	
			0

G			
			G

$$V_3(1,2) = \frac{1}{4}(-1 + (1)(0)) + \frac{1}{4}(-1 + (1)(-1.7)) + \frac{1}{4}(-1 + (1)(-2)) + \frac{1}{4}(-1 + (1)(-2)) = -2.4$$

Iterative Policy Evaluation (3)

$V_{i,Random}$ | $\Pi_{Greedy}(V_i)$

I = 2

0	-1.7	-2	-2
-1.7	-2	-2	-2
-2	-2	-2	-1.7
-2	-2	-1.7	0

G	$\leftarrow$	$\leftarrow$	$\leftarrow\updownarrow\rightarrow$
$\uparrow$	$\leftarrow\uparrow$	$\leftarrow\updownarrow\rightarrow$	$\downarrow$
$\uparrow$	$\leftarrow\updownarrow\rightarrow$	$\downarrow\rightarrow$	$\downarrow$
$\leftarrow\updownarrow\rightarrow$	$\rightarrow$	$\rightarrow$	G

I = 3

0	-2.4	-2.9	-3
-2.4	-2.9	-3	-2.9
-2.9	-3	-2.9	-2.4
-3	-2.9	-2.4	0

G	$\leftarrow$	$\leftarrow$	$\leftarrow\downarrow$
$\uparrow$	$\leftarrow\uparrow$	$\leftarrow\downarrow$	$\downarrow$
$\uparrow$	$\uparrow\rightarrow$	$\downarrow\rightarrow$	$\downarrow$
$\uparrow\rightarrow$	$\rightarrow$	$\rightarrow$	G

Iterative Policy Evaluation (4)

$V_{i,Random}$ | $\Pi_{Greedy}(V_i)$

l = 3

0	-2.4	-2.9	-3
-2.4	-2.9	-3	-2.9
-2.9	-3	-2.9	-2.4
-3	-2.9	-2.4	0

G	$\leftarrow$	$\leftarrow$	$\leftarrow\downarrow$
$\uparrow$	$\leftarrow\uparrow$	$\leftarrow\downarrow$	$\downarrow$
$\uparrow$	$\uparrow\rightarrow$	$\downarrow\rightarrow$	$\downarrow$
$\uparrow\rightarrow$	$\rightarrow$	$\rightarrow$	G

l = 4

0		-3.8	
-3.9			0

G			
			G

$V_3(1,3) = \frac{1}{4}(-1+(1)(-2.4)) + \frac{1}{4}(-1+(1)(-2.9)) + \frac{1}{4}(-1+(1)(-3)) + \frac{1}{4}(-1+(1)(-3)) = -3.8$

$V_3(4,1) = \frac{1}{4}(-1+(1)(-3)) + \frac{1}{4}(-1+(1)(-2.9)) + \frac{1}{4}(-1+(1)(-2.9)) + \frac{1}{4}(-1+(1)(-3)) = -3.9$

Iterative Policy Evaluation (5)

$V_{i,Random}$

$\Pi_{Greedy}(V_i)$

I = 4

0	-3.1	-3.8	-3.9
-3.1	-3.7	-3.9	-3.8
-3.8	-3.9	-3.7	-3.1
-3.9	-3.8	-3.1	0

G	$\leftarrow$	$\leftarrow$	$\leftarrow\downarrow$
$\uparrow$	$\leftarrow\uparrow$	$\leftarrow\downarrow$	$\downarrow$
$\uparrow$	$\uparrow\rightarrow$	$\downarrow\rightarrow$	$\downarrow$
$\uparrow\rightarrow$	$\rightarrow$	$\rightarrow$	G

I = 10

0	-6.1	-8.4	-9
-6.1	-7.7	-8.4	-8.4
-8.4	-8.4	-7.7	-6.1
-9	-8.4	-6.1	0

G	$\leftarrow$	$\leftarrow$	$\leftarrow\downarrow$
$\uparrow$	$\leftarrow\uparrow$	$\leftarrow\downarrow$	$\downarrow$
$\uparrow$	$\uparrow\rightarrow$	$\downarrow\rightarrow$	$\downarrow$
$\uparrow\rightarrow$	$\rightarrow$	$\rightarrow$	G

Observations

- We could improve policy by iteratively spreading the information from goal towards the start.
- We have to visit all the states available at each pass to make sure it converges.
- Policy sequence $\pi_0, \pi_1, \pi_2 ....$is monotonically improving.
- $V_{\pi_1 \pi_0 \pi_0 ...} \geq V_{\pi_0 \pi_0 \pi_0 ...}$
- $V_{\pi_1 \pi_1 \pi_0 ...} \geq V_{\pi_1 \pi_0 \pi_0 ...}$
- $V_{\pi_1 \pi_1 \pi_1 ...} \geq V_{\pi_0 \pi_0 \pi_0 ...}$
- We did only one step look ahead here, why not look ahead few more steps? We are exactly going to do that next.

References

- Berkeley RL Course: Lecture 5 - Value Iteration, Policy Iteration
- UCL RL Course: Lecture 3: Planning by Dynamic Programming
- Value iteration and policy iteration algorithms for Markov decision problem