



AIL 7022: Reinforcement Learning

Lecture 6: Policy Evaluation

Instructor: Raunak Bhattacharyya



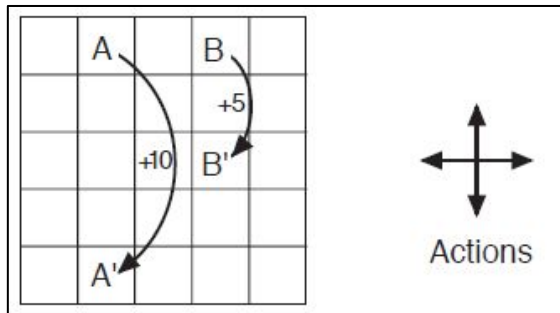
ScAI

YARDI SCHOOL OF ARTIFICIAL INTELLIGENCE
INDIAN INSTITUTE OF TECHNOLOGY DELHI

Recap

- History-dependent policies
- Markovian policies
- Stationary (dynamics, rewards, policies)
- FH and IFH MDP
- SSP MDP

Example of a Value Function



Off the grid actions result in reward -1; discount set to 0.9

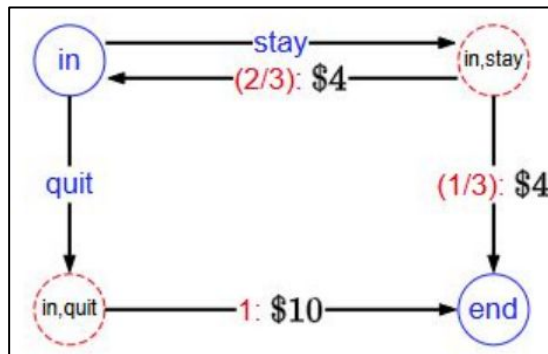
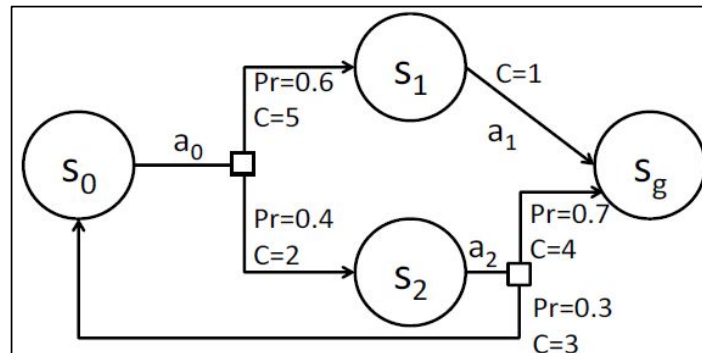
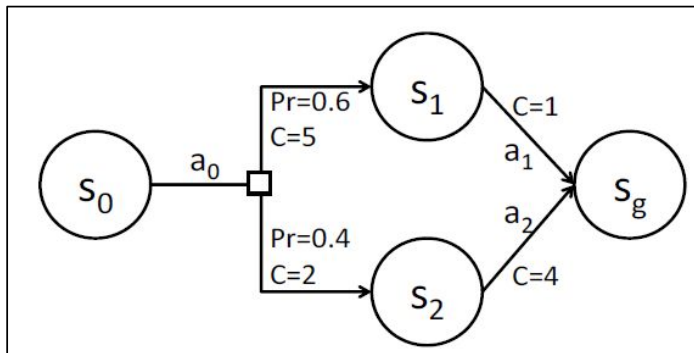
Find value of Equiprobable policy

						
	3.3	8.8	4.4	5.3	1.5	
	1.5	3.0	2.3	1.9	0.5	
	0.1	0.7	0.7	0.4	-0.4	
	-1.0	-0.4	-0.4	-0.6	-1.2	
	-1.9	-1.3	-1.2	-1.4	-2.0	

Plan: Policy Evaluation

- Models are known
 - Iterative approach
- Without models
 - Sampling based

Policy Evaluation



Value Function: Recursive Definition

$$\begin{aligned}v_{\pi}(s) &\doteq \mathbb{E}_{\pi}[G_t \mid S_t = s] \\&= \mathbb{E}_{\pi}[R_{t+1} + \gamma G_{t+1} \mid S_t = s] \\&= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r | s, a) \left[r + \gamma \mathbb{E}_{\pi}[G_{t+1} | S_{t+1} = s'] \right] \\&= \sum_a \pi(a|s) \sum_{s', r} p(s', r | s, a) \left[r + \gamma v_{\pi}(s') \right], \quad \text{for all } s \in \mathcal{S},\end{aligned}$$

Existence and uniqueness

Linear System

Side Problem: Expected number of Die Rolls

You roll a fair 6-sided die repeatedly.

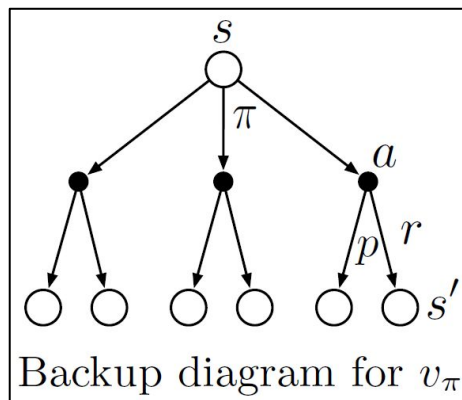
Each roll adds the outcome (1 to 6) to your current total.

You start at total = 0.

What is the expected number of rolls to exceed N ?

Value Function: Iterative Solution

$$\begin{aligned} v_{k+1}(s) &\doteq \mathbb{E}_{\pi}[R_{t+1} + \gamma v_k(S_{t+1}) \mid S_t = s] \\ &= \sum_a \pi(a|s) \sum_{s', r} p(s', r \mid s, a) [r + \gamma v_k(s')] \end{aligned}$$



Iterative Policy Evaluation: S&B

Iterative Policy Evaluation, for estimating $V \approx v_\pi$

Input π , the policy to be evaluated

Algorithm parameter: a small threshold $\theta > 0$ determining accuracy of estimation

Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(\text{terminal}) = 0$

Loop:

$\Delta \leftarrow 0$

Loop for each $s \in \mathcal{S}$:

$v \leftarrow V(s)$

$V(s) \leftarrow \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a) [r + \gamma V(s')]$

$\Delta \leftarrow \max(\Delta, |v - V(s)|)$

until $\Delta < \theta$

Iterative Policy Evaluation: M&K

Algorithm 3.1: Iterative Policy Evaluation

```
1  //Assumption:  $\pi$  is proper
2  initialize  $V_0^\pi$  arbitrarily for each state
3   $n \leftarrow 0$ 
4  repeat
5       $n \leftarrow n + 1$ 
6      foreach  $s \in \mathcal{S}$  do
7          compute  $V_n^\pi(s) \leftarrow \sum_{s' \in \mathcal{S}} \mathcal{T}(s, \pi(s), s') \left[ \mathcal{C}(s, \pi(s), s') + V_{n-1}^\pi(s') \right]$ 
8          compute  $\text{residual}_n(s) \leftarrow |V_n^\pi(s) - V_{n-1}^\pi(s)|$ 
9      end
10 until  $\max_{s \in \mathcal{S}} \text{residual}_n(s) < \epsilon$ ;
11 return  $V_n^\pi$ 
```

Example

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

$R_t = -1$
on all transitions

v_k for the
random policy

$k = 0$

0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0

$k = 1$

0.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	-1.0
-1.0	-1.0	-1.0	0.0

$k = 2$

0.0	-1.7	-2.0	-2.0
-1.7	-2.0	-2.0	-2.0
-2.0	-2.0	-2.0	-1.7
-2.0	-2.0	-1.7	0.0

$k = 3$

0.0	-2.4	-2.9	-3.0
-2.4	-2.9	-3.0	-2.9
-2.9	-3.0	-2.9	-2.4
-3.0	-2.9	-2.4	0.0

$k = 10$

0.0	-6.1	-8.4	-9.0
-6.1	-7.7	-8.4	-8.4
-8.4	-8.4	-7.7	-6.1
-9.0	-8.4	-6.1	0.0

$k = \infty$

0.0	-14.	-20.	-22.
-14.	-18.	-20.	-20.
-20.	-20.	-18.	-14.
-22.	-20.	-14.	0.0

Demo: Iterative Policy Evaluation

GridWorld: Dynamic Programming Demo

Policy Evaluation (one sweep)

Policy Update

Toggle Value Iteration

Reset

0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖
0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖
0.00 ↖					0.00 ↖				0.00 ↖
0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖ R -1.0		0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖
0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖		0.00 ↖ R -1.0	0.00 ↖ R -1.0	0.00 ↖	0.00 ↖	0.00 ↖
0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖		0.00 ↖ R 1.0	0.00 ↖ R -1.0	0.00 ↖	0.00 ↖ R -1.0	0.00 ↖
0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖		0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖ R -1.0	0.00 ↖
0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖ R -1.0		0.00 ↖ R -1.0	0.00 ↖ R -1.0	0.00 ↖	0.00 ↖	0.00 ↖
0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖	0.00 ↖

Theoretical Properties

- Convergence
- Optimality
- Termination
- Running time