Info

CS260R Discussion
Jan 16, 2026

No participation will be recorded for any discussion.

You can go whatever discussion session.

Discussion 1A Friday 1200 - 1350: Matthew Leng (matthewleng@cs.ucla.edu)

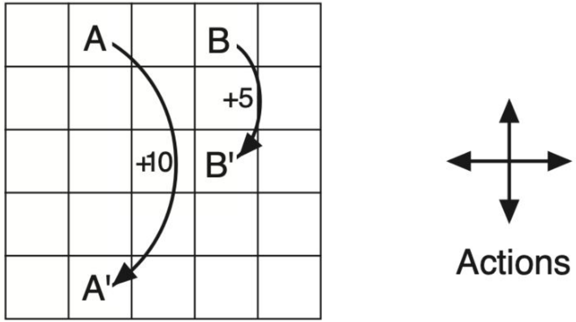
Discussion 1B Friday 1400 - 1550: Haoyuan Cai (haoyuan@cs.ucla.edu)

Agenda Today

1. policy evaluation
2. policy iteration = policy evaluation + policy improvement
3. Q&A

Policy Evaluation

Example 3.5: Gridworld Figure 3.2 (left) shows a rectangular gridworld representation of a simple finite MDP. The cells of the grid correspond to the states of the environment. At each cell, four actions are possible: **`north`**, **`south`**, **`east`**, and **`west`**, which deterministically cause the agent to move one cell in the respective direction on the grid. Actions that would take the agent off the grid leave its location unchanged, but also result in a reward of -1. Other actions result in a reward of 0, except those that move the agent out of the special states A and B. From state A, all four actions yield a reward of $+10$ and take the agent to A$'$. From state B, all actions yield a reward of $+5$ and take the agent to B$'$.



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

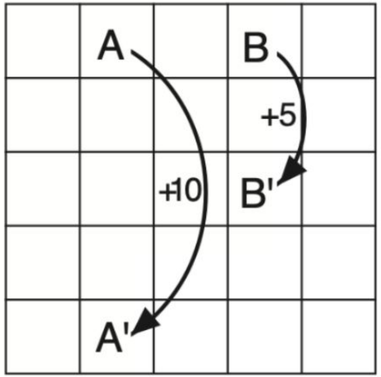
Figure 3.2: Gridworld example: exceptional reward dynamics (left) and state-value function for the equiprobable random policy (right).

Policy Evaluation

Initialize all values to 0

MDP Summary:

- Four actions: Up, Down, Left, Right
- Assume uniform random action
- R(A, A') = +10
- R(B, B') = +5
- Action takes off-grid is invalid, R(S, S) = -1



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

Policy Evaluation

Update value table with Empirical Bellman backup

$$\hat{V}^{\pi}(s) = \mathbb{E}_{a\sim\pi(\cdot|s),s'\sim\mathbb{P}(\cdot|s,a)}[r(s,a)+\gamma\hat{V}^{\pi}(s')] = \Sigma_{s'\in\mathcal{S}}\mathbb{P}(s'|s,\pi)[r(s,a)+\gamma\hat{V}^{\pi}(s')]$$

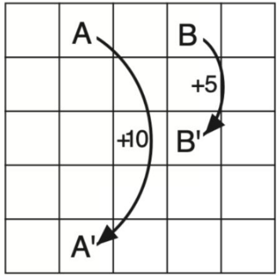
```
While value table is not stationary:
    For all states:
        Compute new values with ⬆️ equation and store in a temporary table
    Update all entries of value table with new table simultaneously
```

Policy Evaluation Step 1:

Update value table with Empirical Bellman backup

MDP Summary:

- Four actions: Up, Down, Left, Right
- Assume uniform random action
- R(A, A') = +10
- R(B, B') = +5
- Action takes off-grid is invalid, R(S, S) = -1
- Gamma = 1



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

Table at iteration=0

-0.5	+10	-0.25	+5	-0.5
-0.25	0	0	0	-0.25
-0.25	0	0	0	-0.25
-0.25	0	0	0	-0.25
-0.5	-0.25	-0.25	-0.25	-0.5

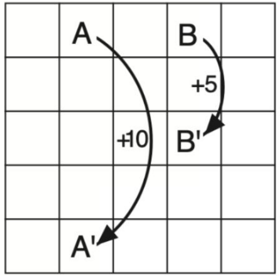
Table at iteration=1

$$\hat{V}^{\pi}(s) = \Sigma_{s'\in\mathcal{S}}\mathbb{P}(s'|s,\pi)[r(s,a)+\gamma\hat{V}^{\pi}(s')]$$

Policy Evaluation Step 2:

Update value table with Empirical Bellman backup

R(A, A') = +10
R(B, B') = +5
Action takes off-grid is invalid,
R(S, S) = -1
Gamma = 1



-0.5	+10	-0.25	+5	-0.5
-0.25	0	0	0	-0.25
-0.25	0	0	0	-0.25
-0.25	0	0	0	-0.25
-0.5	-0.25	-0.25	-0.25	-0.5

Table at iteration=1

	2.4375			

Table at iteration=2

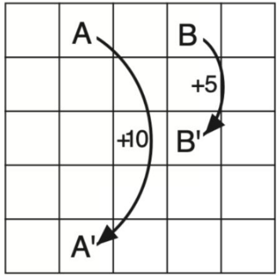
- Tips: 0.25*0.25 = 0.0625

$$\hat{V}^{\pi}(s) = \Sigma_{s' \in \mathcal{S}} \mathbb{P}(s'|s,\pi)[r(s,a) + \gamma \hat{V}^{\pi}(s')]$$

Policy Evaluation Step 2:

Update value table with Empirical Bellman backup

R(A, A') = +10
R(B, B') = +5
Action takes off-grid is invalid,
R(S, S) = -1
Gamma = 1



-0.5	+10	-0.25	+5	-0.5
-0.25	0	0	0	-0.25
-0.25	0	0	0	-0.25
-0.25	0	0	0	-0.25
-0.5	-0.25	-0.25	-0.25	-0.5

Table at iteration=1

	2.4375			
-0.5				

Table at iteration=2

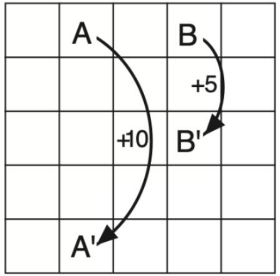
- Tips: 0.25*0.25 = 0.0625

$$\hat{V}^\pi(s) = \Sigma_{s'\in\mathcal{S}}\mathbb{P}(s'|s,\pi)[r(s,a)+\gamma\hat{V}^\pi(s')]$$

Policy Evaluation Step 2:

Update value table with Empirical Bellman backup

R(A, A') = +10
R(B, B') = +5
Action takes off-grid is invalid,
R(S, S) = -1
Gamma = 1



-0.5	+10	-0.25	+5	-0.5
-0.25	0	0	0	-0.25
-0.25	0	0	0	-0.25
-0.25	0	0	0	-0.25
-0.5	-0.25	-0.25	-0.25	-0.5

Table at iteration=1

			+5	
	2.4375			
-0.5				

Table at iteration=2

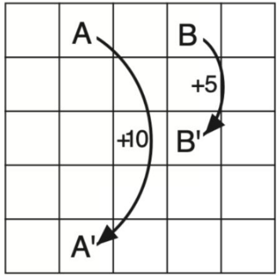
- Tips: 0.25*0.25 = 0.0625

$$\hat{V}^{\pi}(s) = \Sigma_{s' \in \mathcal{S}} \mathbb{P}(s'|s,\pi)[r(s,a) + \gamma \hat{V}^{\pi}(s')]$$

Policy Evaluation Step 2:

Update value table with Empirical Bellman backup

R(A, A') = +10
R(B, B') = +5
Action takes off-grid is invalid,
R(S, S) = -1
Gamma = 1



-0.5	+10	-0.25	+5	-0.5
-0.25	0	0	0	-0.25
-0.25	0	0	0	-0.25
-0.25	0	0	0	-0.25
-0.5	-0.25	-0.25	-0.25	-0.5

Table at iteration=1

			+5	
	2.4375			
-0.5				
				-0.875

Table at iteration=2

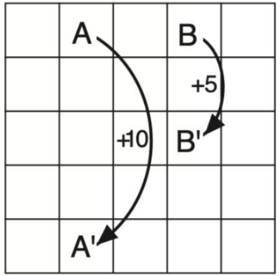
- Tips: 0.25*0.25 = 0.0625

$$\hat{V}^{\pi}(s) = \Sigma_{s' \in \mathcal{S}} \mathbb{P}(s'|s, \pi)[r(s, a) + \gamma \hat{V}^{\pi}(s')]$$

Policy Evaluation Step 2:

Update value table with Empirical Bellman backup

R(A, A') = +10
R(B, B') = +5
Action takes off-grid is invalid,
R(S, S) = -1
Gamma = 1



-0.5	+10	-0.25	+5	-0.5
-0.25	0	0	0	-0.25
-0.25	0	0	0	-0.25
-0.25	0	0	0	-0.25
-0.5	-0.25	-0.25	-0.25	-0.5

Table at iteration=1

1.6875	9.75	3.4375	+5	0.4375
-0.5	2.4375	0.0625	1.1875	-0.5
-0.4375	0.0625	0	0.0625	-0.4375
-0.5	-0.125	0.0625	-0.125	-0.5
-0.875	-0.5	-0.4375	-0.5	-0.875

Table at iteration=2

- Tips: 0.25*0.25 = 0.0625

$$\hat{V}^{\pi}(s) = \Sigma_{s' \in \mathcal{S}} \mathbb{P}(s'|s,\pi)[r(s,a) + \gamma \hat{V}^{\pi}(s')]$$

Policy Iteration

```
While the policy is still changing:
    While the value is still changing:
        For all states:
            Compute new values with ⬆️ equation and store in a temporary table
        Update all entries of value table with new table simultaneously
    For all states s:
        For all actions a: Compute Q(s, a) = $r(s, a) + \gamma \hat{V}^{\pi}(s')$
        Set the action π(a|s) = argmax Q(s, a)
```

1.6875	9.75	3.4375	5	0.4375
-0.5	2.4375	0.0625	1.1875	-0.5
-0.4375	0.0625	0	0.0625	-0.4375
-0.5	-0.125	0.0625	-0.125	-0.5
-0.875	-0.5	-0.4375	-0.5	-0.875

Q&A

Discussion 1A Friday 1200 - 1350: Matthew Leng (matthewleng@cs.ucla.edu)

Discussion 1B Friday 1400 - 1550: Haoyuan Cai (haoyuan@cs.ucla.edu)