지능 시스템
Intelligent Systems

Example Questions

1. Reward Computation 1

- Discounting factor is 0.9
- Rewards at each step from the start of the episode are as follows;
  - -1, 2, 6, 3, 2
- The episode starts from the initial state (state 0) and terminates at state 5.

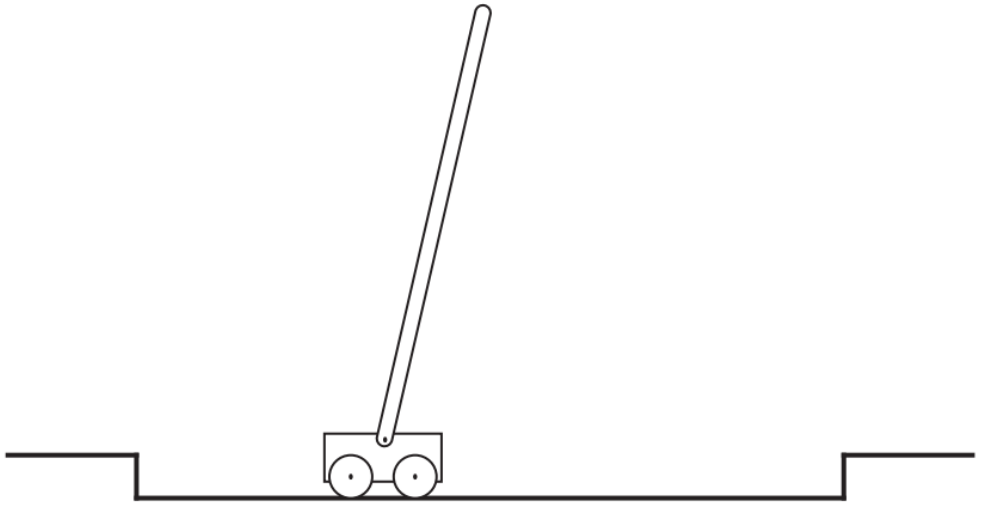
Q: What are $G_0, G_1, G_2, G_3, G_4, G_5$?

2. Reward Computation 2

Pole balancing Task

- Episodic, ends when the pole trips over.
- Reward = -1 for the pole tripping over and
  Reward = 0 otherwise
- Rewards are discounted.

Q: State the expression of the return at each timestep.



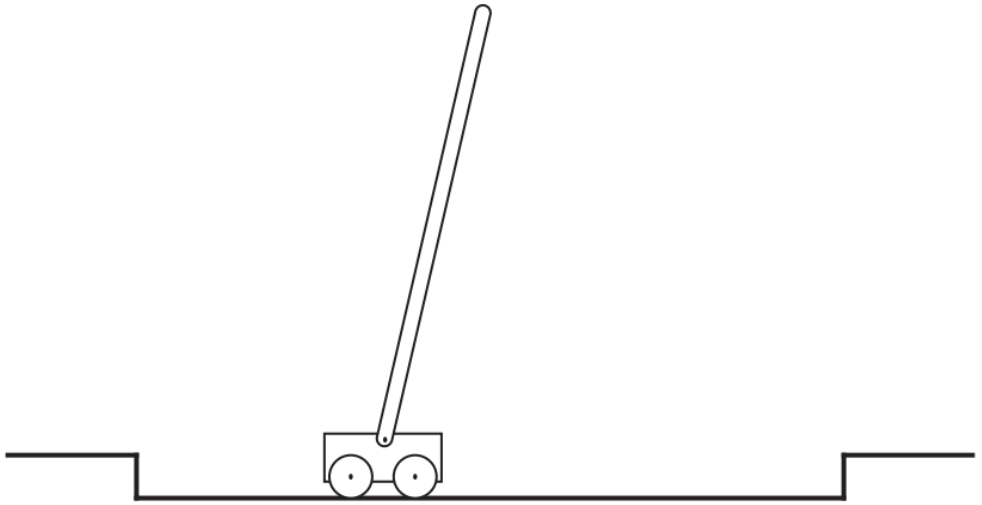
2. Reward Computation 2

Pole balancing Task

- Episodic, ends when the pole trips over.
- Reward = -1 for the pole tripping over and Reward = 0 otherwise.
- Rewards are discounted.

Q: State the expression of the return at each timestep.

A: $G_t = -\gamma^{H-t}$,
Where H is the time step at which the pole has been tripped over.

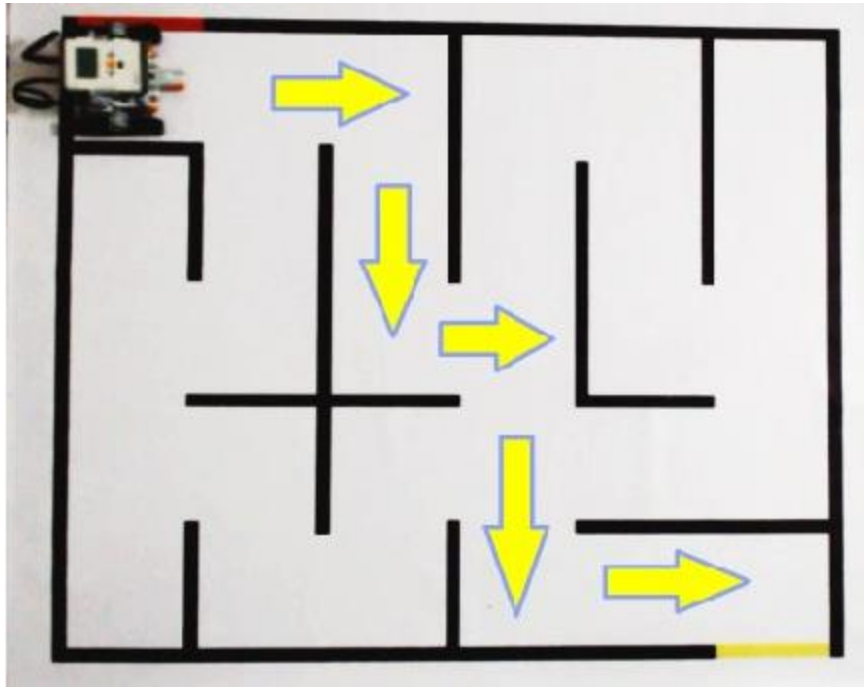


3. Reward Analysis

Maze Robot Task

- Goal: escape the maze ASAP
- Episodic, ends when the robot escapes.
- Reward = +1 for terminal state and
  Reward = 0 otherwise.
- Rewards are not discounted.

Q: After a significant amount of learning, the robot is not improving in escaping the maze. What is the problem? What should we change for the robot to learn properly?



3. Reward Analysis

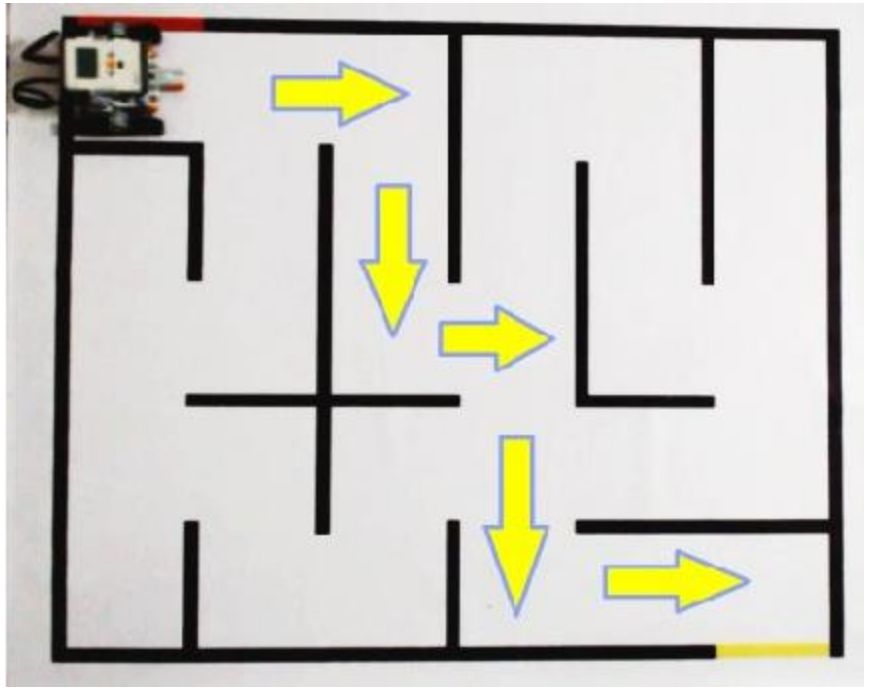
Maze Robot Task

- Goal: escape the maze ASAP
- Episodic, ends when the robot escapes.
- Reward = +1 for terminal state and
  Reward = 0 otherwise.
- Rewards are discounted.

Q: After a significant amount of learning, the robot is not improving in escaping the maze. What is the problem? What should we change for the robot to learn properly?

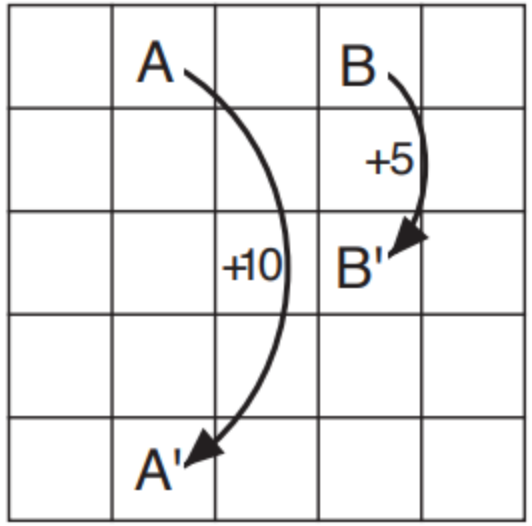
A:

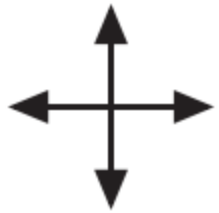
- No penalty for the duration until escape.
- Must add penalty reward for each step.
- No discount factor to indicate the time steps remaining until terminal state.



4. Grid World

- Possible Actions: move up/down/left/right
- When hitting the edge, the agent remains in the same cell but gets Reward = -1
- At state A, the agent directly moves to state A’ with any actions and gets Reward = +10
- At state B, the agent directly moves to state B’ with any actions and gets Reward = +5
- The agent gets Reward = 0 in any other cases.
- Grid on the right shows the values at each cell under the current policy of uniformly selecting the actions.





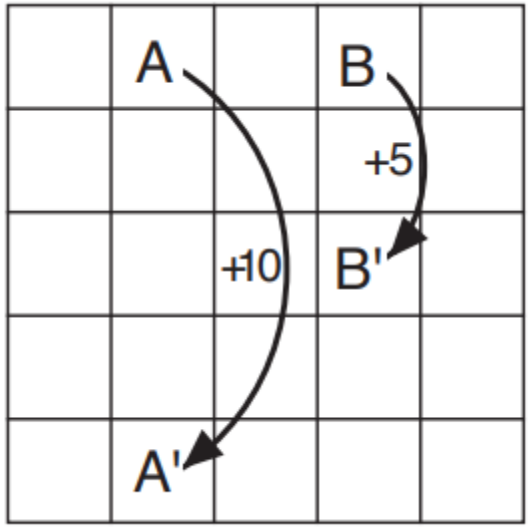
Actions

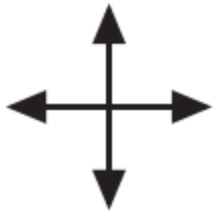
3.3	8.8	4.4	5.3	1.5
1.5	3.0	2.3	1.9	0.5
0.1	0.7	0.7	0.4	-0.4
-1.0	-0.4	-0.4	-0.6	-1.2
-1.9	-1.3	-1.2	-1.4	-2.0

4. Grid World

Q:

1. **Why is the value at A smaller than 10, which corresponds to the immediate reward?**
2. **Why is the value at B larger than 5, which corresponds to the immediate reward?**
3. **Show that the Bellman Equation holds for the centre cell with the value of 0.7 with respect to the neighbouring four states. (use 0.9 as the discount rate)**





Actions

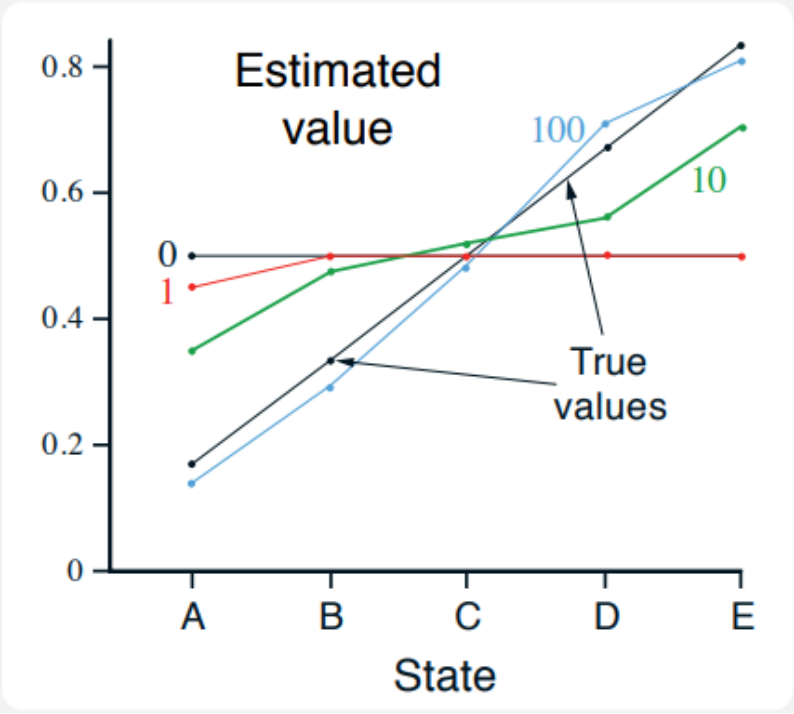
3.3	8.8	4.4	5.3	1.5
1.5	3.0	2.3	1.9	0.5
0.1	0.7	0.7	0.4	-0.4
-1.0	-0.4	-0.4	-0.6	-1.2
-1.9	-1.3	-1.2	-1.4	-2.0

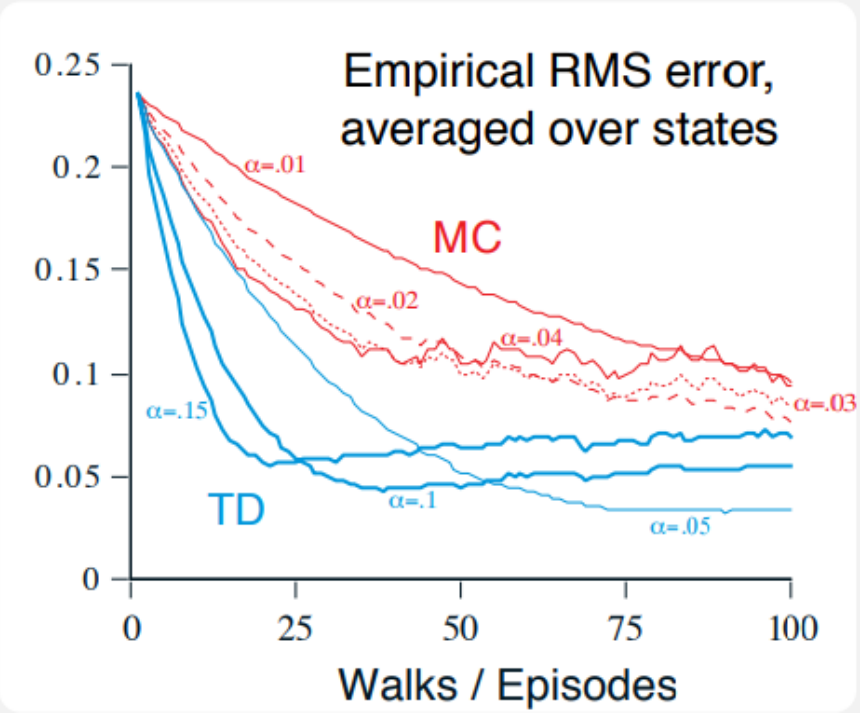
5. Random Walk

- Episodic
- Starts at state C
- Terminates either at the extreme left or the extreme right
- Reward = +1 when it ends on the extreme right, Reward = 0 otherwise
- No discount rate
- At each state, its Value is the probability of reaching to the extreme right.
- Hence, the value at state C, V(C) = 0.5

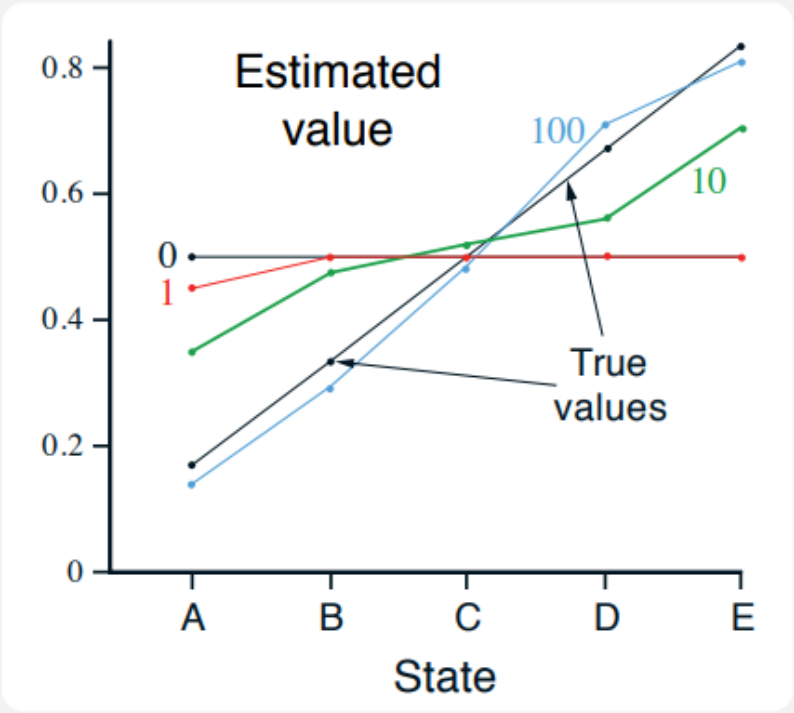
Q: Find the optimal values at states A, B, D, and E.

5. Random Walk





5. Random Walk

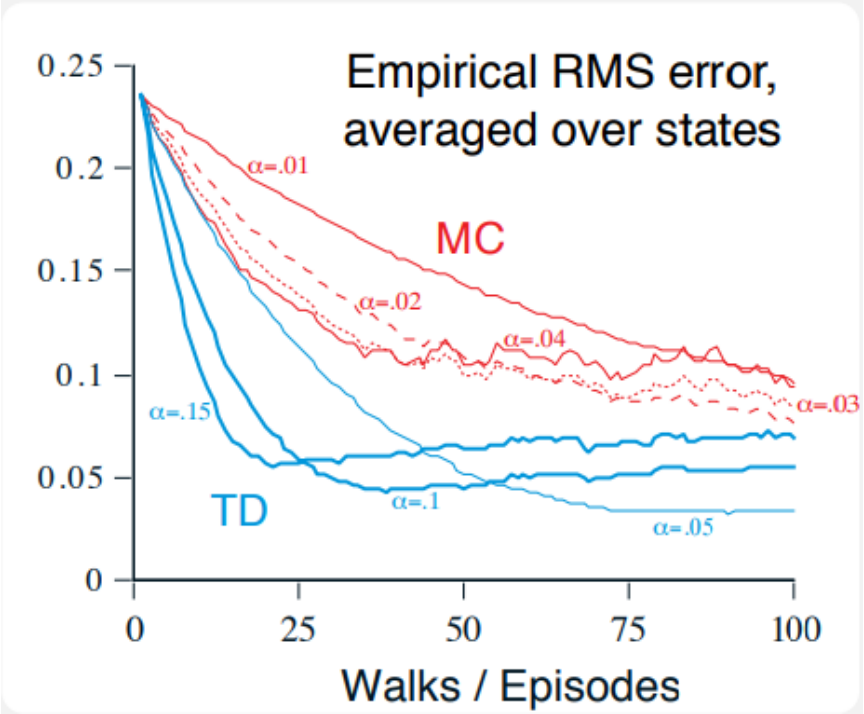


Q: Consider the case with the first episode result (red).
It only altered the value at state A. Describe what happened at the first episode and by how much the value at A changed. (learning rate, $\alpha = 0.1$)

5. Random Walk

Q:

1. **Describe the difference between the performances of MC and TD algorithm.**
2. **In TD, describe the learning curve for different values of learning rate.**



6. TD to MC

Bellman Equation:

$$V^{\pi}(s_t) = \mathbb{E}_{\tau_{a_t: s_{t+n}} \sim p_{\pi}(\tau|s_t,a_t)}[G_{t:t+n-1} + \gamma^n V^{\pi}(s_{t+n})|s_t]$$

For $n = 1$

$$V^{\pi}(s_t) = \mathbb{E}_{a_{t+1}\sim\pi(a_t|s_t),s_{t+1}\sim p(s_{t+1}|s_t,a_t)}[r_t + \gamma V^{\pi}(s_{t+1})|s_t]$$

Q:

1. **Rewrite the TD update equation for n=2.**
2. **Describe the change in the result compared to the case n=1.**
3. **What value of n would lead the TD update equation to MC?**

MC $$V(s_t) \leftarrow V(s_t) + \alpha(G_t - V(s_t))$$

TD $$V(s_t) \leftarrow V(s_t) + \alpha(r_t + \gamma V(s_{t+1}) - V(s_t))$$