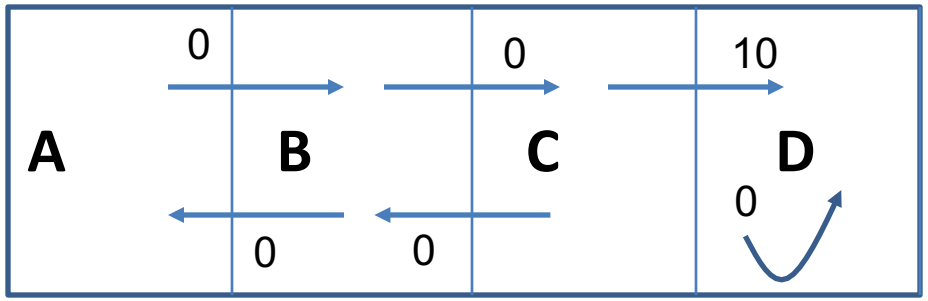
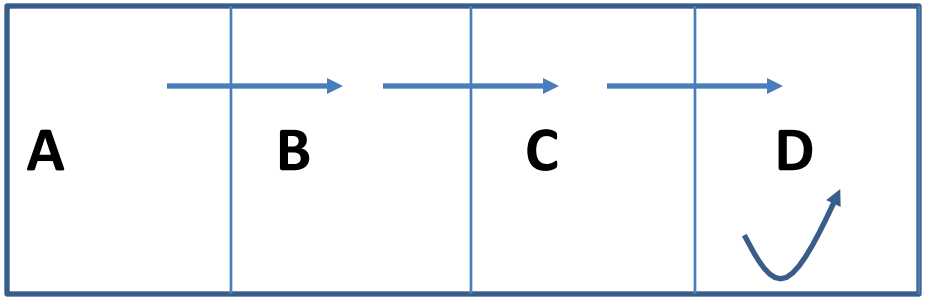
- Given below is a robotic world with immediate rewards as specified



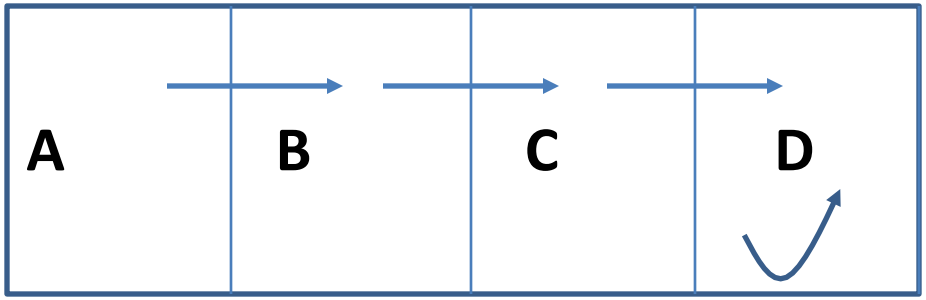
- Given the below policy, what is the total reward that we can expect starting from A assuming a discount factor of 0.5



- Ans: 0+0.5*0+0.5*0.5*10+0+0+…..

- What is the optimal policy

- Ans:



- What values would Q-learning converge to assuming a discount factor of 0.5 ?
- Ans: For this example, you don't really need to run Q-Learning algorithm, the optimal policy is evident from the example. Upon convergence, the value of Q for each state-action will be the optimal reward that can be received for an action if we follow the optimal policy