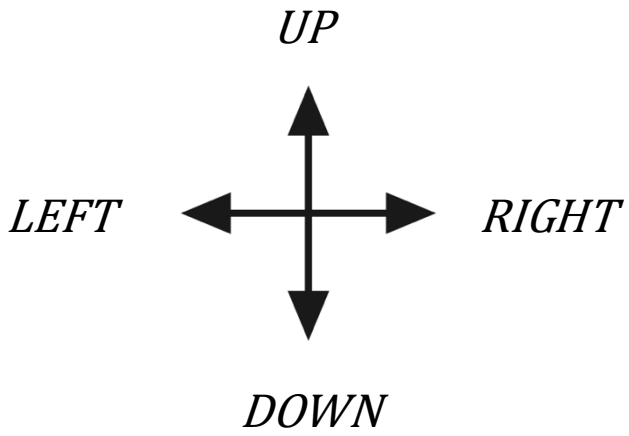
Exercise 4: 3x3 Gridworld

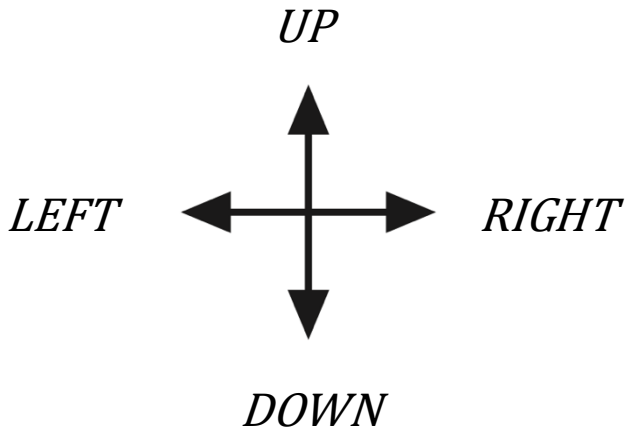
A	B	C
D	E	F
G	H	I



- **States** $\mathcal{S} = (A, B, C, D, E, F, G, H, I)$
- **Actions** $\mathcal{A} = (UP, DOWN, LEFT, RIGHT)$
- **Policy** $\mathcal{P} =$ From every state, choose each action with probability 0.25
- **Reward** $(\mathcal{R} = -1)\ per\ step$
- Discount Factor $(\gamma = 1)$

Exercise 4: 3x3 Gridworld

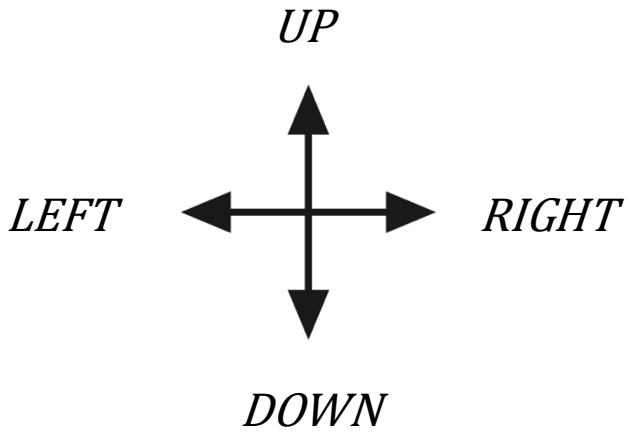
A	B	C
D	E	F
G	H	I



- Undiscounted MDP ($\gamma = 1$)
- Non-terminal states (A, B, D, E, F, H)
- Terminal State (C, G, I)
- Agent follows a uniform random policy

Exercise 4: 3x3 Gridworld

A	B	C
D	E	F
G	H	I

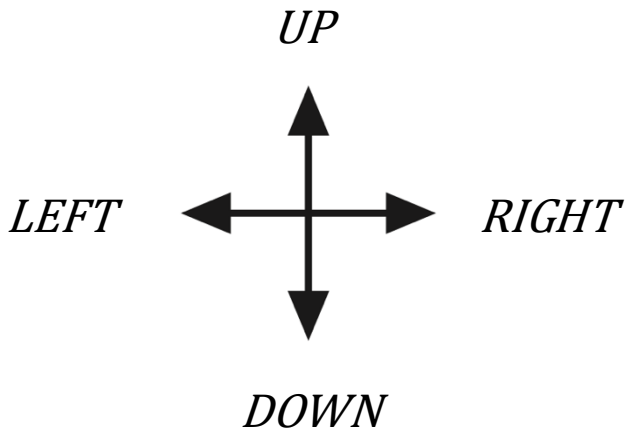


Rules:

- From each state, actions move you in that direction if possible, otherwise you stay in the same square.
- Reward is -1 until the terminal state is reached.

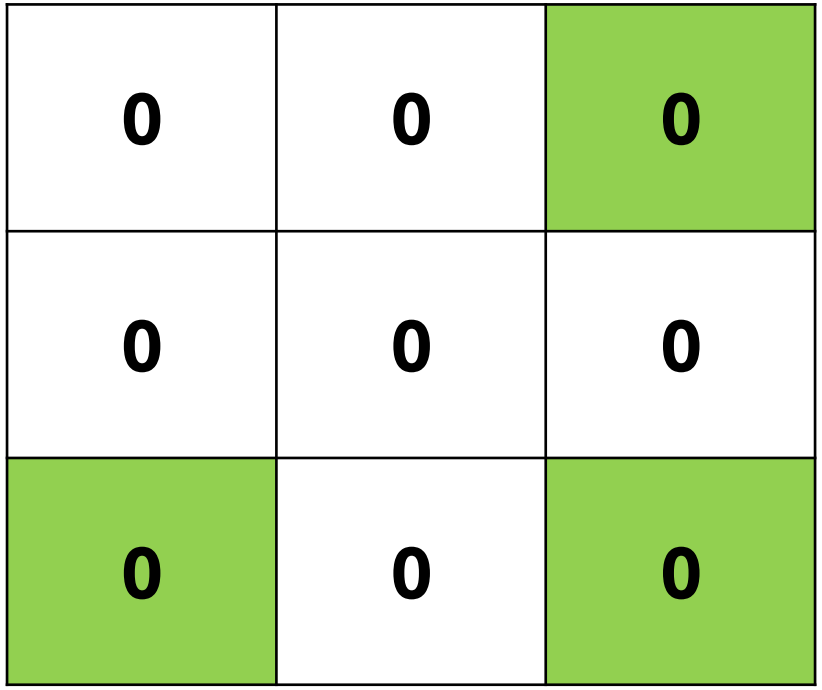
Exercise 4: 3x3 Gridworld

A	B	C
D	E	F
G	H	I

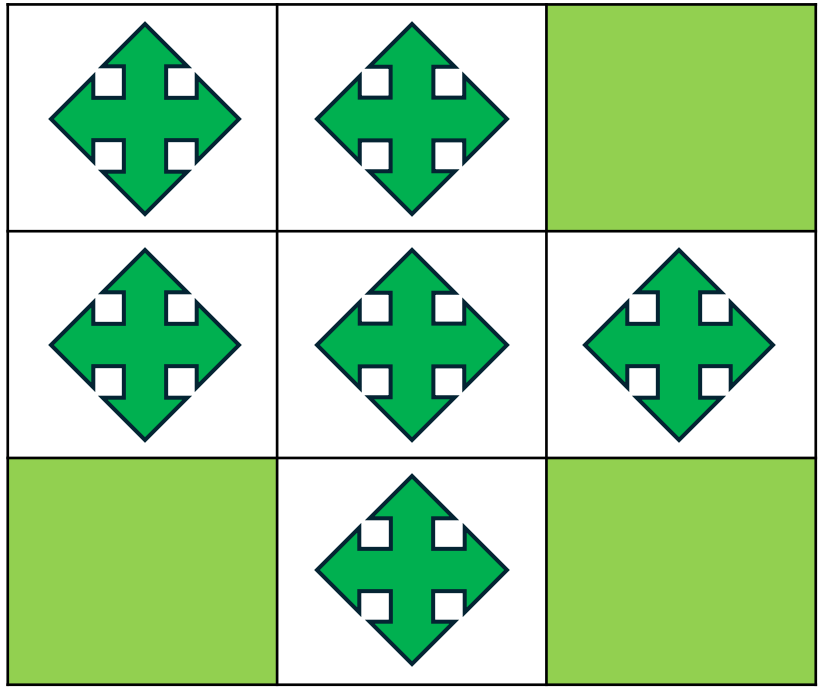


Goal

- The goal is to reach state C, G, I which gives **0 reward** and ends the episode.
- To reach the goal, we need to find the optimal policy π_*



value functions at $k = 0$



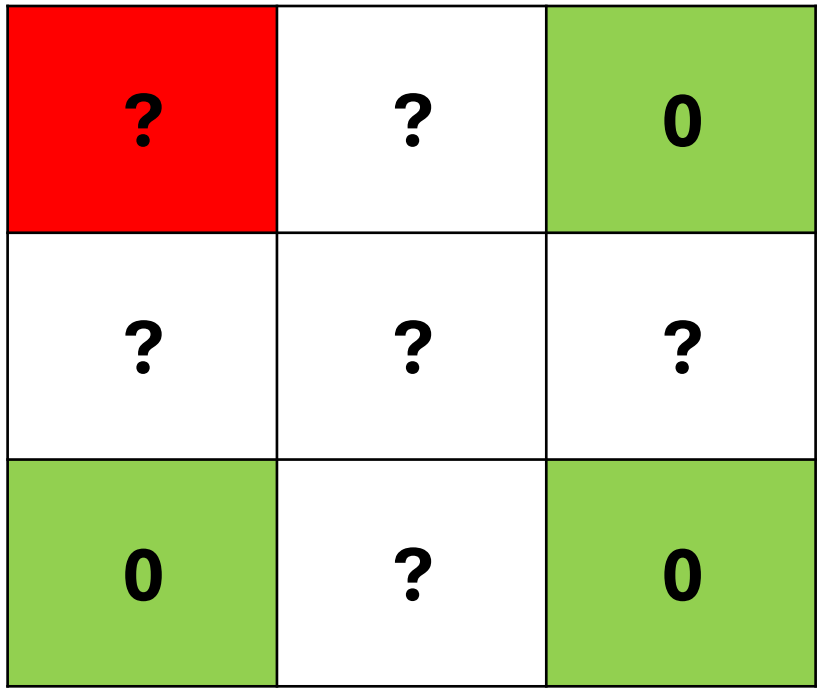
uniform random policy at $k = 0$

Step 1: Compute the value function of states A,B,D,E,F,H at $k = 1$

	$v_k(s)$	$v_{k+1}(s)$	$v_{k+2}(s)$
A	0	?	?
B	0	?	?
D	0	?	?
E	0	?	?
F	0	?	?
H	0	?	?

Step 1: Compute the value function of state A at $k = 1$

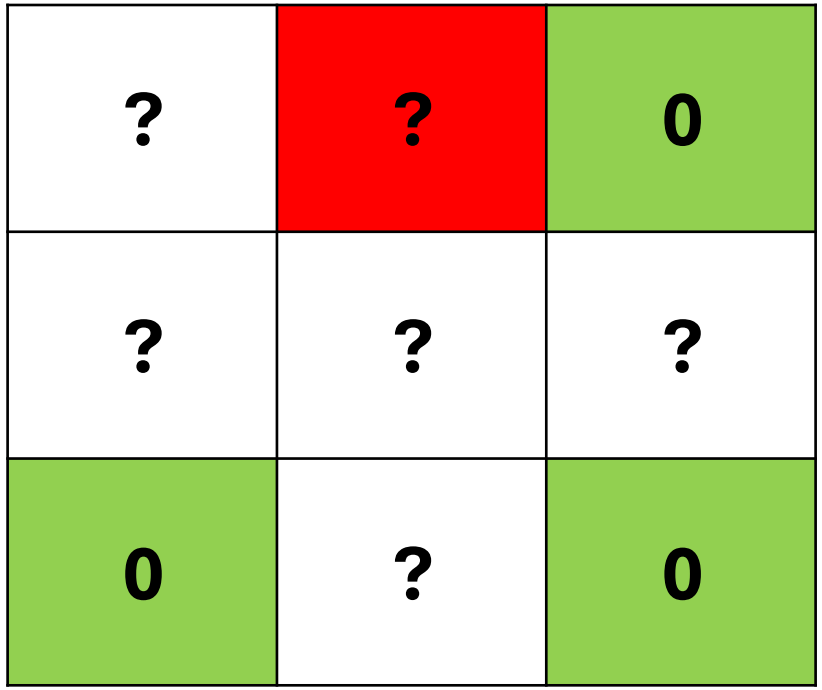
$1. v_{k+1}(A) = ?$



$k = 1$

Step 1: Compute the value function of state B at $k = 1$

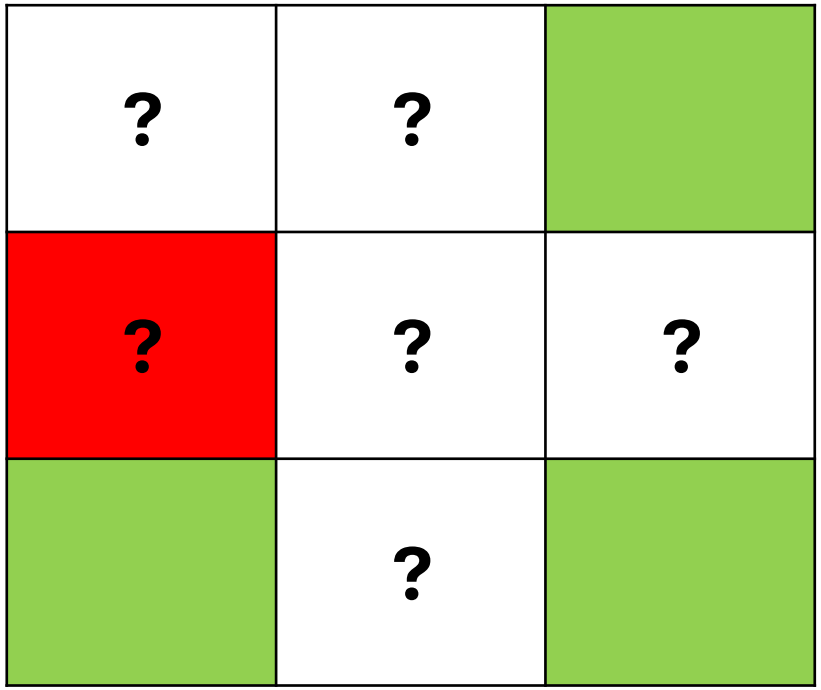
$2.\ v_{k+1}(B) = ?$



$k = 1$

Step 1: Compute the value function of state D at $k = 1$

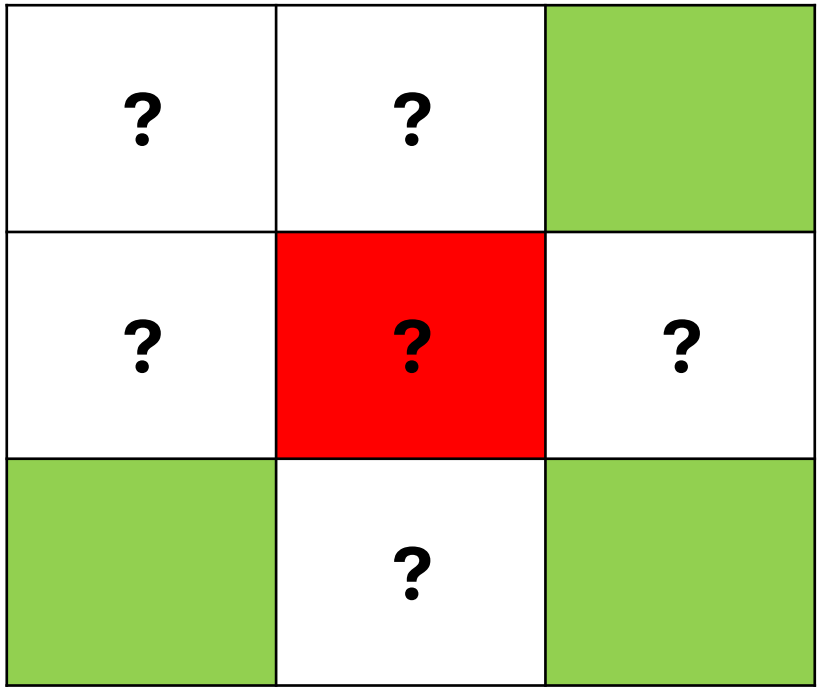
$3.\ v_{k+1}(D) = ?$



$k = 1$

Step 1: Compute the value function of state E at $k = 1$

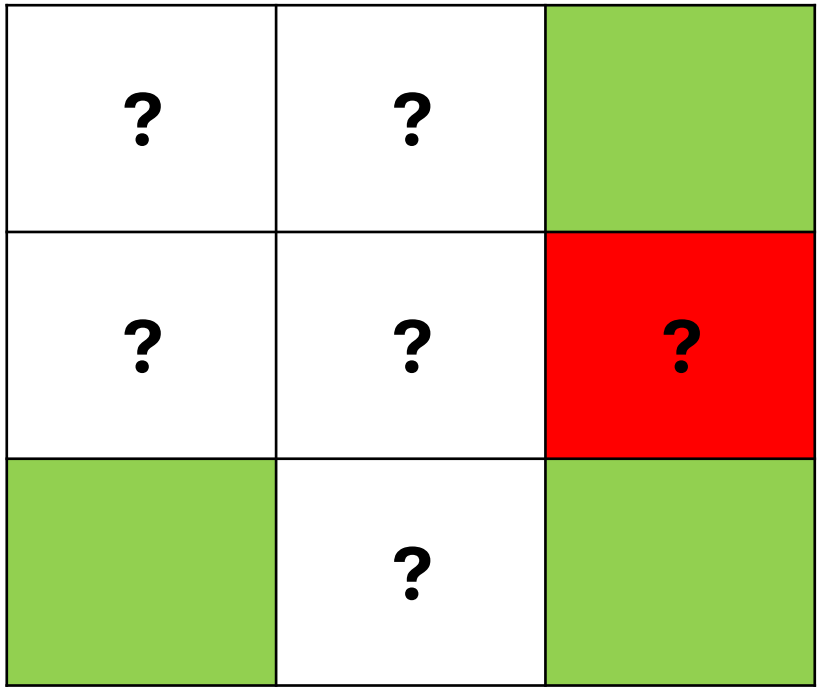
$4. v_{k+1}(E) = ?$



$k = 1$

Step 1: Compute the value function of state F at $k = 1$

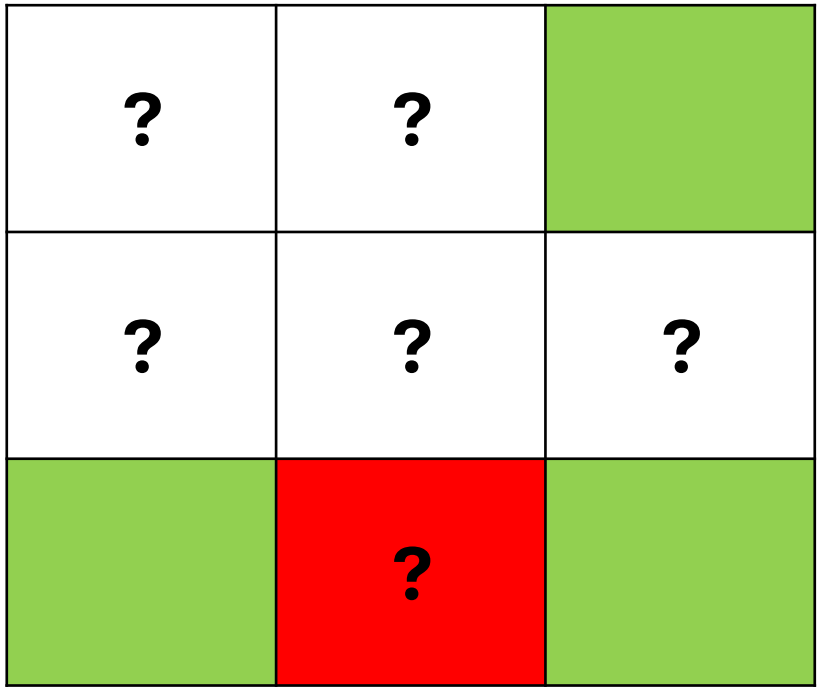
$5.\ v_{k+1}(F) = ?$



$k = 1$

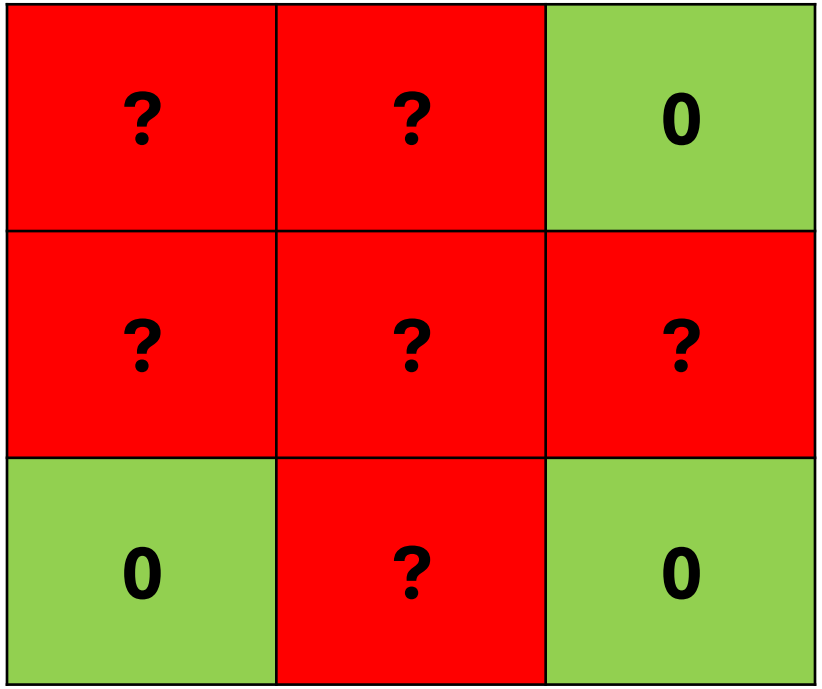
Step 1: Compute the value function of state H at $k = 1$

$6.\ v_{k+1}(H) = ?$



$k = 1$

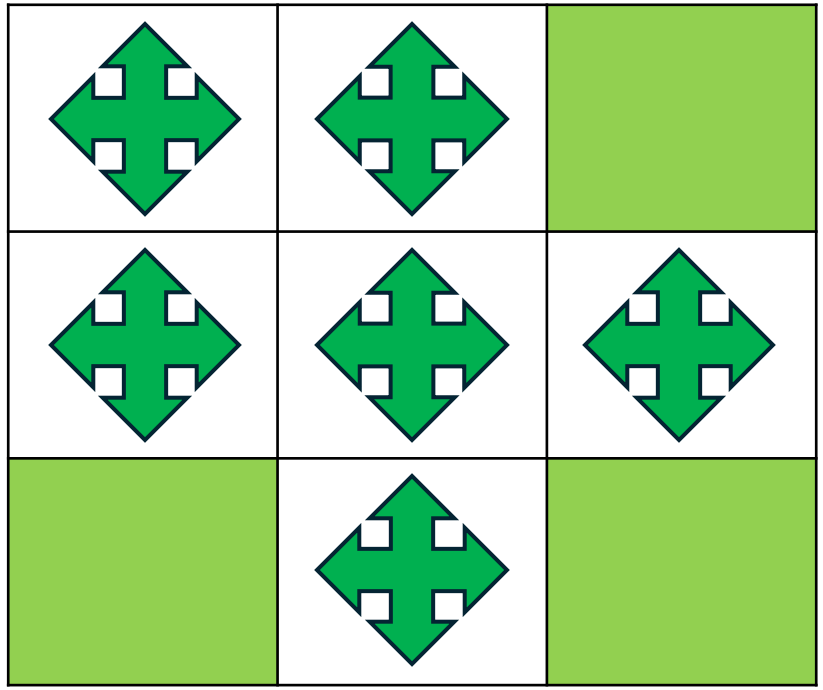
Step 1: Compute the value function of states A, B, D, E, F, H at $k = 1$



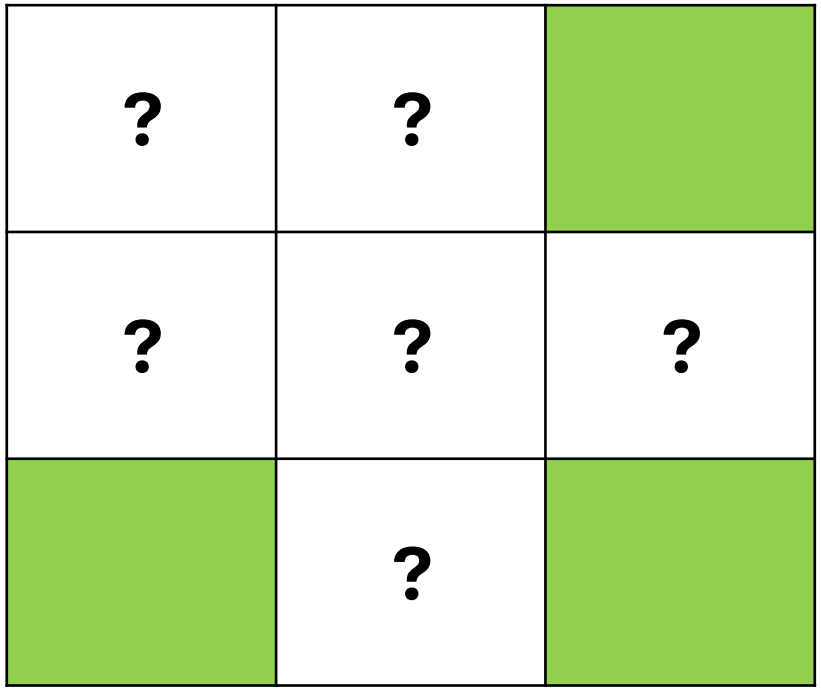
$k = 1$

7. *Put the new value functions in the 3x3 grid*

Step 2: Compute the action-value function and update the policy of states A, B, D, E, F, H at $k = 1$



$k = 0$



$k = 1$

Step 2: Compute the action-value function and update the policy of state A at $k = 1$

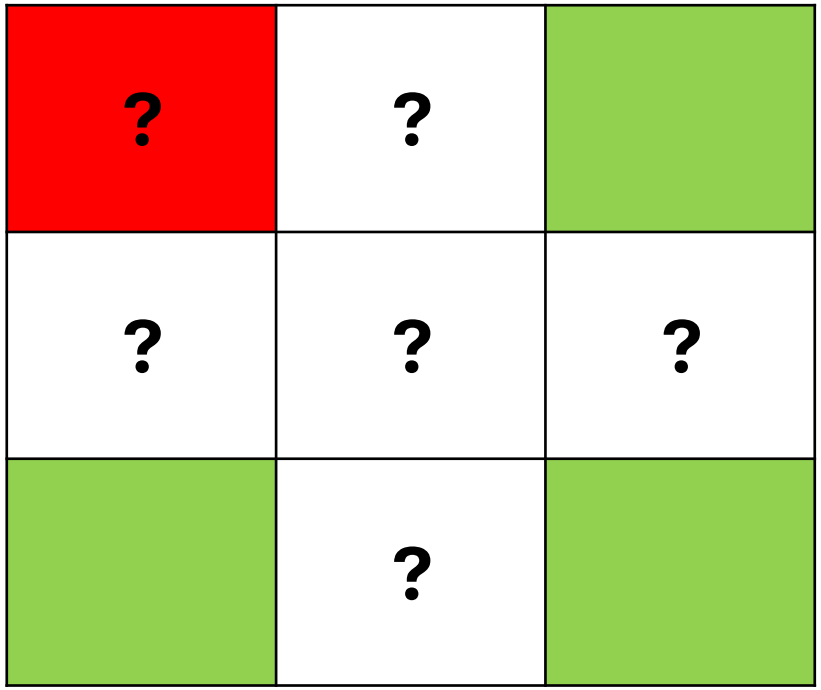
8. $q_{k+1}(A, LEFT) = ?$

9. $q_{k+1}(A, RIGHT) = ?$

10. $q_{k+1}(A, UP) = ?$

11. $q_{k+1}(A, DOWN) = ?$

12. $\pi_{k+1}(A) = ?$



$k = 1$

Step 2: Compute the action-value function and update the policy of state B at $k = 1$

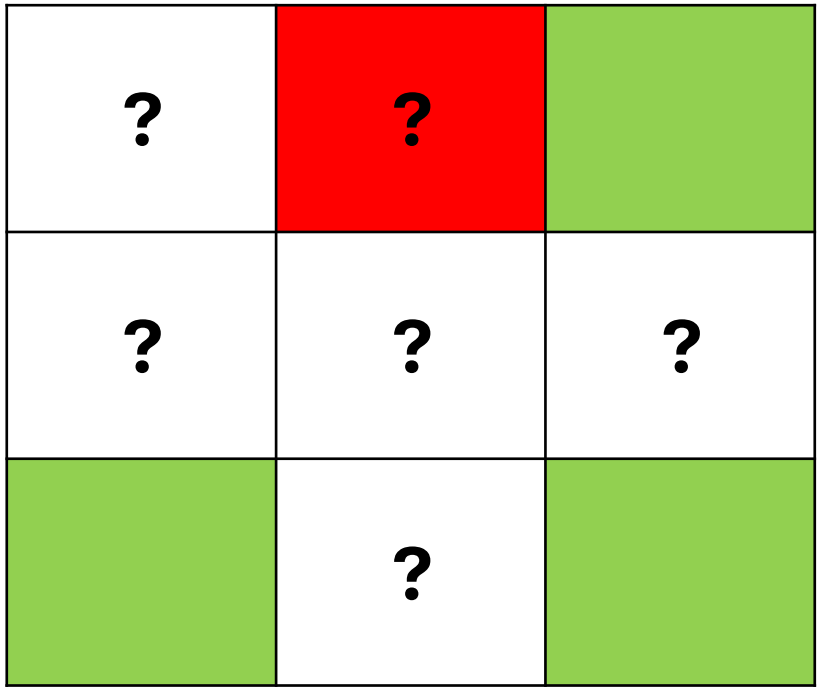
13. $q_{k+1}(B, LEFT) = ?$

14. $q_{k+1}(B, RIGHT) = ?$

15. $q_{k+1}(B, UP) = ?$

16. $q_{k+1}(B, DOWN) = ?$

17. $\pi_{k+1}(B) = ?$



$k = 1$

Step 2: Compute the action-value function and update the policy of state D at $k = 1$

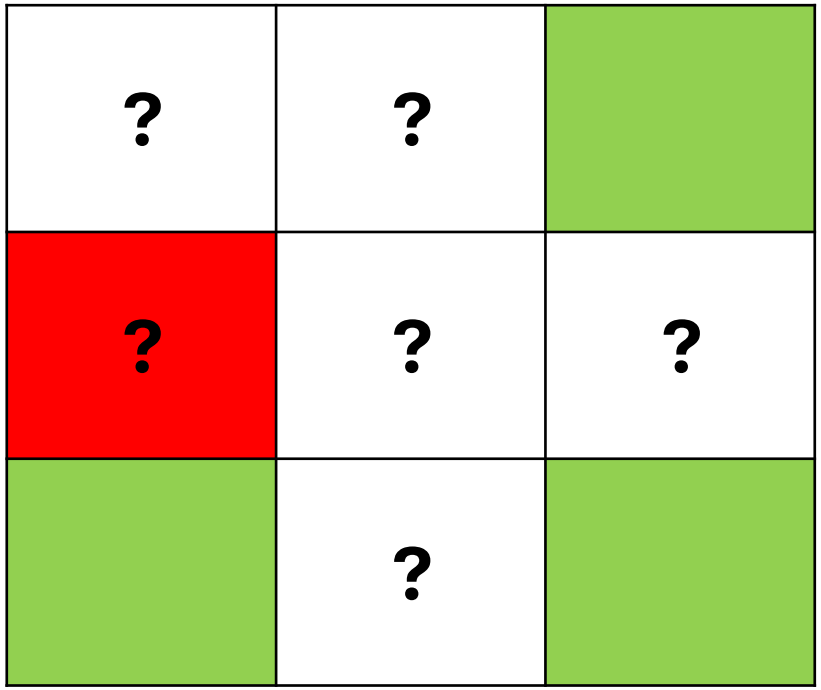
18. $q_{k+1}(D, LEFT) = ?$

19. $q_{k+1}(D, RIGHT) = ?$

20. $q_{k+1}(D, UP) = ?$

21. $q_{k+1}(D, DOWN) = ?$

22. $\pi_{k+1}(D) = ?$



$k = 1$

Step 2: Compute the action-value function and update the policy of state E at $k = 1$

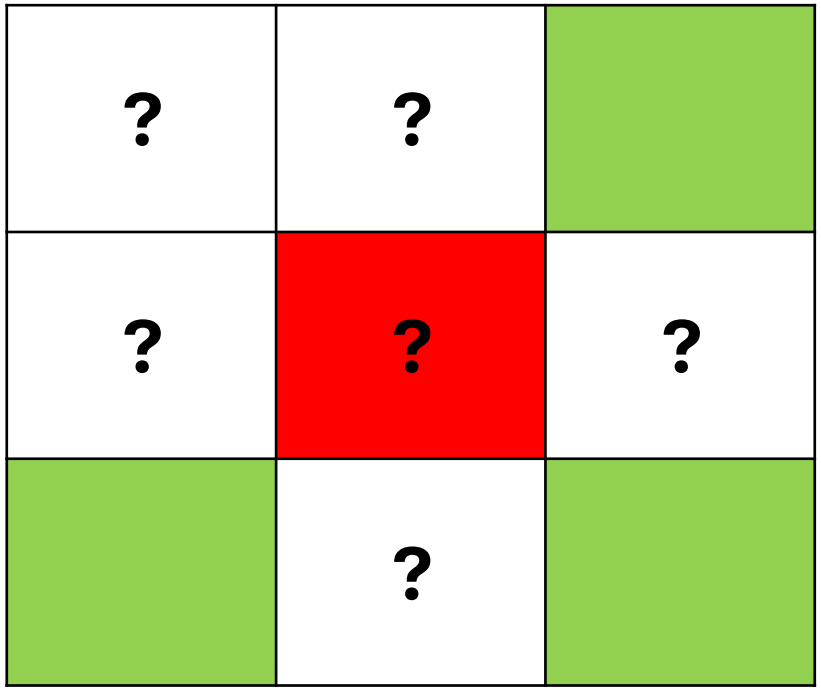
23. $q_{k+1}(E, LEFT) = ?$

24. $q_{k+1}(E, RIGHT) = ?$

25. $q_{k+1}(E, UP) = ?$

26. $q_{k+1}(E, DOWN) = ?$

27. $\pi_{k+1}(E) = ?$



$k = 1$

Step 2: Compute the action-value function and update the policy of state F at $k = 1$

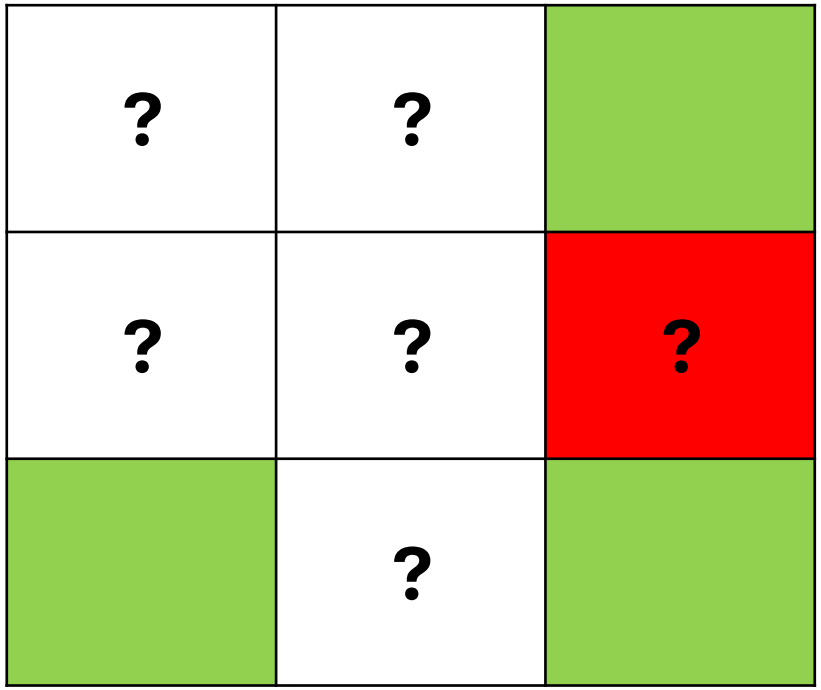
28. $q_{k+1}(F, LEFT) = ?$

29. $q_{k+1}(F, RIGHT) = ?$

30. $q_{k+1}(F, UP) = ?$

31. $q_{k+1}(F, DOWN) = ?$

32. $\pi_{k+1}(F) = ?$



$k = 1$

Step 2: Compute the action-value function and update the policy of state H at $k = 1$

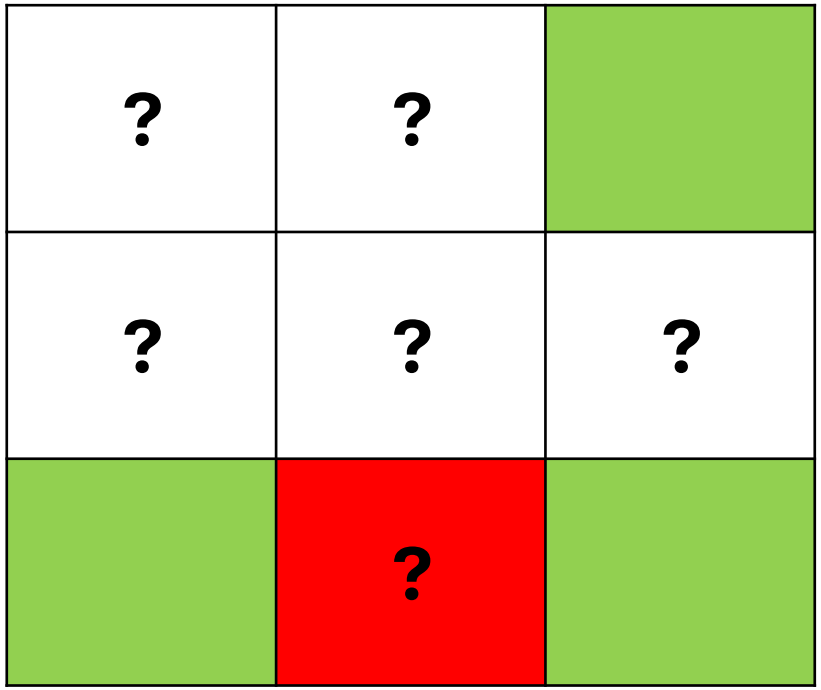
33. $q_{k+1}(H, LEFT) = ?$

34. $q_{k+1}(H, RIGHT) = ?$

35. $q_{k+1}(H, UP) = ?$

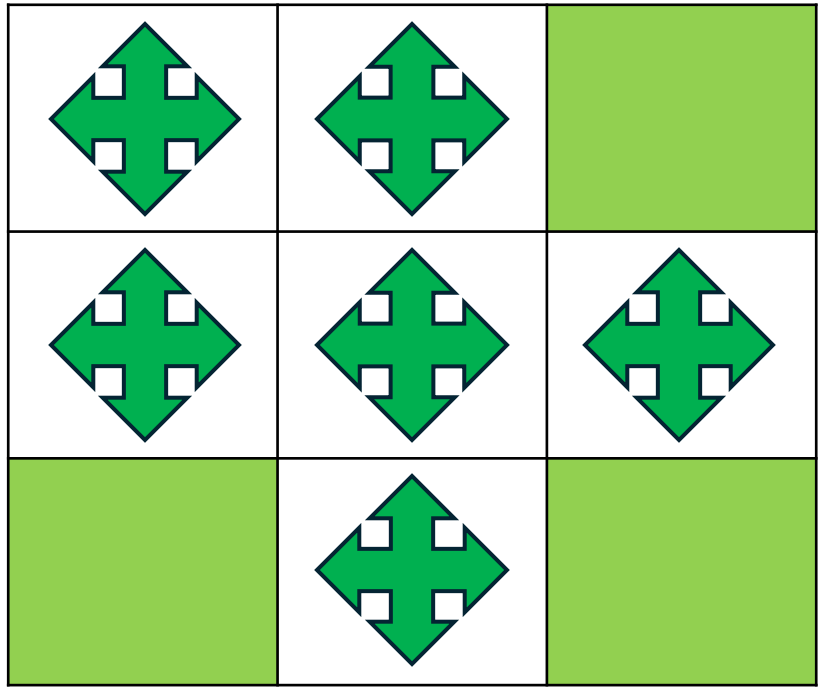
36. $q_{k+1}(H, DOWN) = ?$

37. $\pi_{k+1}(H) = ?$

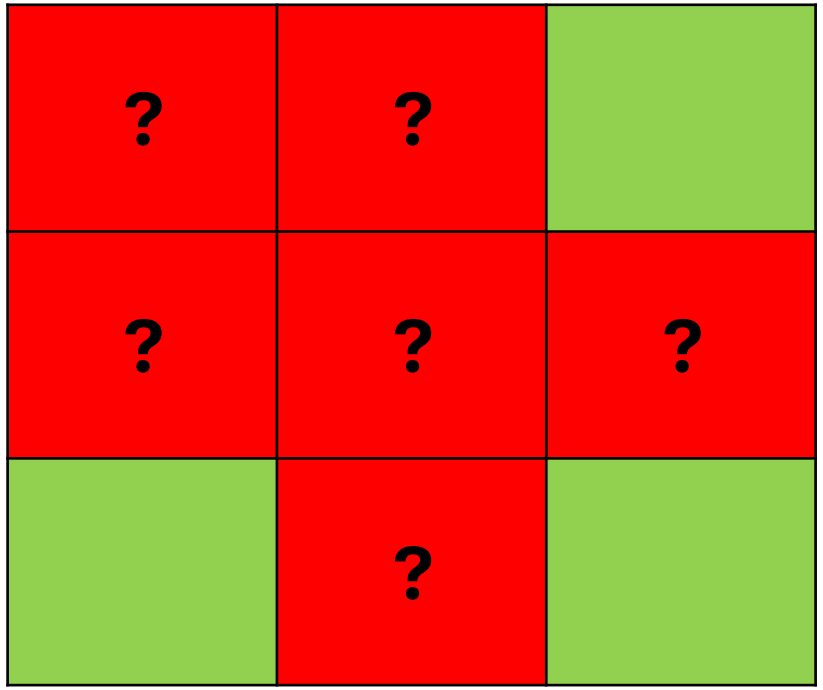


$k = 1$

Step 2: Compute the action-value function and update the policy of states A, B, D, E, F, H at $k = 1$



$k = 0$



$k = 0$

38. *Put the new policies in the 3x3 grid*

Step 3: Use DP to find the optimal value function $\boldsymbol{v}_*$ of states A, B, D, E, F, H

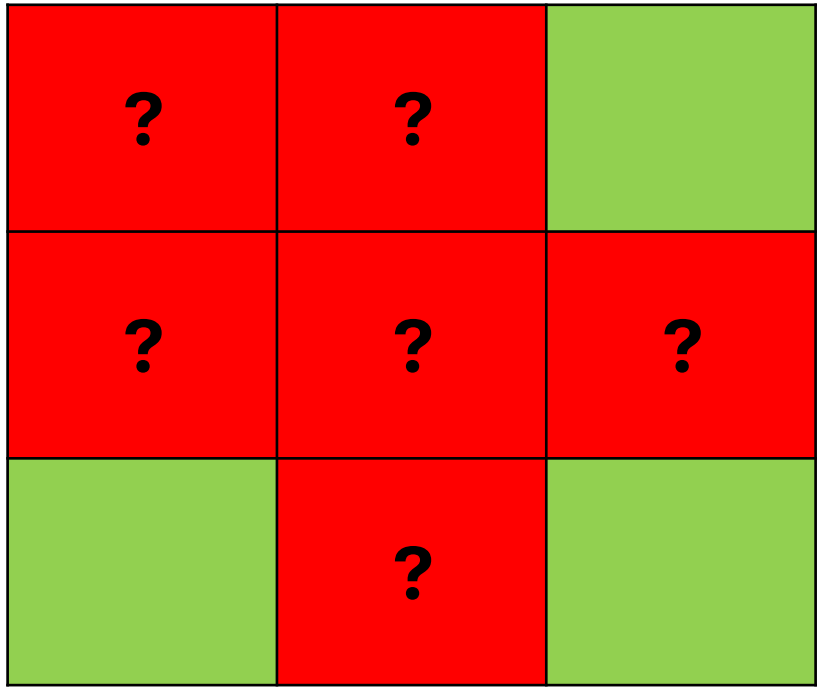
39. $v_*(A) = ?$

40. $v_*(B) = ?$

41. $v_*(D) = ?$

42. $v_*(F) = ?$

43. $v_*(H) = ?$



Step 3: Use DP to find the optimal action-value function $\boldsymbol{q}_*$ of states A, B, D, E, F, H

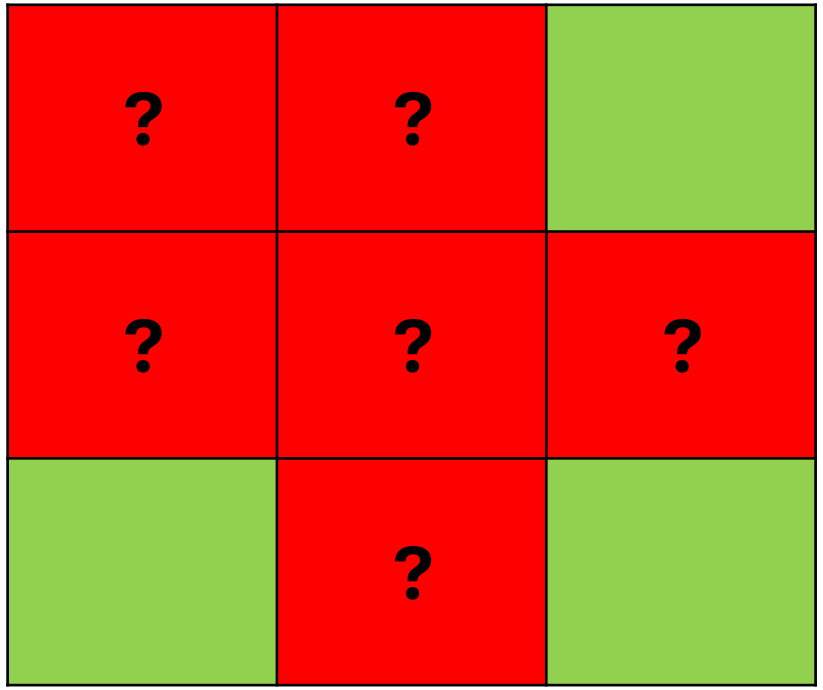
44. $q_*(A|\mathcal{A}) = ?$

45. $q_*(B|\mathcal{A}) = ?$

46. $q_*(D|\mathcal{A}) = ?$

47. $q_*(F|\mathcal{A}) = ?$

48. $q_*(H|\mathcal{A}) = ?$



Step 3: Use DP to find the optimal policy π_* of states A, B, D, E, F, H

49. $\pi_*(A) = ?$

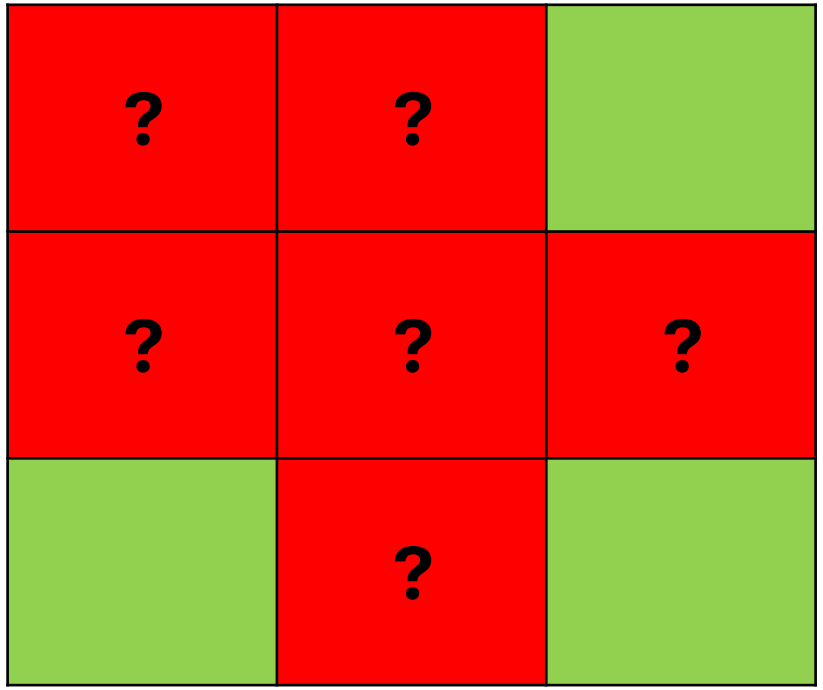
50. $\pi_*(B) = ?$

51. $\pi_*(D) = ?$

52. $\pi_*(E) = ?$

53. $\pi_*(F) = ?$

54. $\pi_*(H) = ?$



Final Step: Map the optimal value function v_* and the optimal policy π_*

?	?	0
?	?	?
0	?	0

55. *Put the optimal value functions in the 3x3 grid*

?	?	0
?	?	?
0	?	0

56. *Put the optimal policy in the 3x3 grid*