Policies for the 6 cases:

1)

a)

- - R -
T L R T
T T - T
T L - R

b)

- - R -
R R R T
R T - T
T T - T

2)

a)

- - L -
B R L B
R T - T
T L - R

b)

- - R -
R R R T
T T - T
T T - T

c)

- - R -
T R R T
T T - T
T R - T

d)

- - R -
T L R T
T B - T
R R - L

Observations:

All the above policies except 2(a) plans a ‘move_right’ in cell (0, 2). The reason being that this is the only case with a high positive step cost. Hence, one would prefer to make more moves than to reach a goal state and stop. It plans a left instead of bottom move such that the probability of reaching the goal state stays 0. In this case, one avoids hitting an end state in every scenario. Right beside a goal state, a move is chosen such that the probability of reaching the goal state stays 0. In (1, 2) and (1, 3), a circular movement is planned. We try to keep moving and not hit an end state ever.

1(b) seems a very rational plan. At every cell, we take the step which will allow us to reach the end state with the highest reward using a minimum number of steps (if we assume that we get to take the move we want i.e the perpendicular moves with probabilities 0.1 do not take place). This is sane and understandable.

1(a) is similar to 1(b), however, the iterations end soon as the discount factor is low. Hence, immediate goals are prioritized more, over what is

best for the agent in the long run. Hence, we try to pick the closest end state with a positive reward and try to avoid the end state with a negative reward (in cell (3, 3), the plan is to go Right).

The only difference between 1(b) and 2(b) is in the plan in cell (0, 2). 2(b) chooses Top while 1(b) chooses right. 2(b) has a more negative step cost. Both are almost equally good in terms of reaching the best possible end state although in 2(b), we reduce the chance of getting closer to the negative reward end state by choosing to go to the top and hence, it seems to be slightly better than 1(b).

In 1(c), the step cost becomes more negative. The agent prefers to end up in the end state with negative reward when it is in (3, 1) than travelling ahead. Again in (1, 0), it prefers the next immediate end state with a lesser reward than travelling ahead to reach the best end state.

In 1(d), a very high negative step cost is given. This is almost similar to saying, reach the nearest end state and prevent moving as much as possible. When the end state with the negative step cost is the closest, we try to end up there and stop. When we have the option of 2 equidistant end states, we pick the one with the higher reward. This can be understood from the plan in cell (2, 3) where going up is chosen.