1.

There are 100 continuous blocks contain -1

$$R(UP) = 50 + (-1) * \gamma + (-1) * \gamma^2 + (-1) * \gamma^3 + \cdots + (-1) * \gamma^{100}$$
$$= 50 - (\gamma + \gamma^2 + \gamma^3 + \cdots + \cdots \gamma^{100})$$
$$= 50 - \sum_{n=1}^{100} \gamma^n$$

There are 100 continuous blocks contain 1

$$R(DOWN) = -50 + 1 * \gamma + 1 * \gamma^2 + 1 * \gamma^3 + \cdots + 1 * \gamma^{100}$$
$$= -50 + (\gamma + \gamma^2 + \gamma^3 + \cdots + \gamma^{100})$$
$$= -50 + \sum_{n=1}^{100} \gamma^n$$

If $R(UP) > R(DOWN)$

$$50 - \sum_{n=1}^{100} \gamma^n > -50 + \sum_{n=1}^{100} \gamma^n$$

$$2\sum_{n=1}^{100} \gamma^n < 100$$

$$\sum_{n=1}^{100} \gamma^n < 50$$

$$\gamma < 0.984398$$

If $\gamma < 0.984398$, it should take UP action, or $\gamma > 0.984398$ it should take DOWN action to get more reward.

2.

(a).

$$\gamma = 1$$

The initial policy has action UP, so the utilities for both gray and black state is:

If the initial state is gray:

$$U(gray) = R(gray) + \gamma\big(P(up) * U(white) + P(stay) * U(gray)\big)$$
$$U(gray) = -1 + 0.1 * U(white) + 0.9 * U(gray)$$

If the initial state is back:

$$U(black) = R(black) + \gamma\big(P(up) * U(white) + P(stay) * U(black)\big)$$
$$U(black) = -2 + 0.1 * U(white) + 0.9 * U(black)$$

If the initial state is white:

$$U(white) = R(white) = 0$$

We can find the value of $U(gray), U(black)$ by substitute $U(white) = 0$

$$U(gray) = -1 + 0.1 * 0 + 0.9 * U(gray)$$
$$U(gray) = -10$$

$$U(black) = -2 + 0.1 * 0 + 0.9 * U(black)$$
$$U(black) = -20$$

If the state is gray:

$$Q^*(gray, R) = T(gray, R, black) * U(black) + T(gray, S, gray) * U(gray)$$
$$= 0.8 * (-20) + 0.2 * (-10) = -18$$
$$Q^*(gray, U) = T(gray, U, white) * U(white) + T(gray, S, gray) * U(gray)$$
$$= 0.1 * 0 + 0.9 * (-10) = -9$$
$$\pi^*(gray) = \text{argmax}\, Q^*(gray, a) = \text{argmax}(-18, -9) = UP$$

The action should be UP when the state is gray.

If the state is black:

$$Q^*(black, L) = T(black, L, gray) * U(gray) + T(black, S, black) * U(black)$$
$$= 0.8 * (-10) + 0.2 * (-20) = -12$$
$$Q^*(black, U) = T(black, U, white) * U(white) + T(black, S, black) * U(black)$$
$$= 0.1 * 0 + 0.9 * (-20) = -18$$
$$\pi^*(black) = \text{argmax}\, Q^*(black, a) = \max(-18, -12) = LEFT$$

The action should be LEFT when the state is black. Refer to the action above, the agent should take UP action when the state is gray. So, the action remains same when the state is gray. Refer to the above iteration, follow the new path and do iteration again to check the utility is updated or not.

If the state is gray:

$$U(gray) = R(gray) + \gamma\big(P(up) * U(white) + P(stay) * U(gray)\big)$$
$$U(gray) = -1 + 0.1 * U(white) + 0.9 * U(gray)$$

If the state is back:

$$U(black) = R(black) + \gamma\big(P(left) * U(gray) + P(stay) * U(black)\big)$$
$$U(black) = -2 + 0.8 * U(gray) + 0.2 * U(black)$$

If the initial state is white:

$$U(white) = 0$$

We can find the value of $U(gray), U(black)$ by substitute $U(white) = 0$

$$U(gray) = -1 + 0.1 * 0 + 0.9 * U(gray)$$
$$U(gray) = -10$$
$$U(black) = -2 + 0.8 * U(gray) + 0.2 * U(black)$$
$$U(black) = -12.5$$

If the state is gray:

$$Q^*(gray, R) = T(gray, R, black) * U(black) + T(gray, S, gray) * U(gray)$$
$$= 0.8 * (-12.5) + 0.2 * (-10) = -12$$
$$Q^*(gray, U) = T(gray, U, white) * U(white) + T(gray, S, gray) * U(gray)$$
$$= 0.1 * 0 + 0.9 * (-10) = -9$$
$$\pi^*(gray) = \text{argmax}\, Q^*(gray, a) = \text{argmax}(-12, -9) = UP$$

The action should be UP when the state is gray.

If the state is black:

$$Q^*(black, L) = T(black, L, gray) * U(gray) + T(black, S, black) * U(black)$$
$$= 0.8 * (-10) + 0.2 * (-12.5) = -10.5$$
$$Q^*(black, U) = T(black, U, white) * U(white) + T(black, S, black) * U(black)$$
$$= 0.1 * 0 + 0.9 * (-12.5) = -11.25$$
$$\pi^*(black) = \text{argmax}\, Q^*(black, a) = \max(-10.5, -11.25) = LEFT$$

When the state is black the action should be LEFT and refer to the action of state gray, it should take UP.
The action remains same, and the iteration is terminated.

Initiative explanation:
the agent should reach to white as soon as possible because stay at gray or black or go left or right will not terminate walking and will cost more to reach to white.
It is not easy for the agent goes UP directly to white because the probability is small. The agent is unlikely going UP and it will remain at the same position.
To let agent reaches to terminal state white, the agent can act left or right first when the agent is in gray or black.
Based on the above iteration, the agent will take LEFT action when the state is in black and take UP action when it is in gray.

(b). when the initial policy is RIGHT and LEFT (for gray and black respectively) and $\gamma = 1$
If the initial state is gray:
$$U(gray) = R(gray) + \gamma\big(P(right) * U(black) + P(stay) * U(gray)\big)$$
$$U(gray) = -1 + 0.8 * U(black) + 0.2 * U(gray)$$
If the initial state is back:
$$U(black) = R(black) + \gamma\big(P(left) * U(gray) + P(stay) * U(black)\big)$$
$$U(black) = -2 + 0.8 * U(gray) + 0.2 * U(black)$$

If the initial state is white:
$$U(white) = R(white) = 0$$
Solve the equation:
$$0.8 * U(gray) = -1 + 0.8 * U(black)$$
$$0.8 * U(black) = -2 + 0.8 * U(gray)$$
$$-3 = 0$$

The equation is not consistent and the values of them are infinity.
When $\gamma = 1$ the initial action cannot be LEFT or RIGHT
The optimal policy depends on discount factor.
When discount factor is small which is less than 1, the factor will tend to be less influenced the result of utility, because γ^n tend to be 0 when n is large enough, and it will break ties when choosing the action.
Because the change of γ will lead to change of utility. When calculating Q-value, the value will change and choice of action will change when find the maximize value by action.

3.

(a).

It is passive reinforcement learning task, because it provides 15 samples to compute $V^{\pi}(G'), V^{\pi}(W'), V^{\pi}(G')$ and the goal is to learn the state values and given the fixed policies. The values we need to compute are the all $V^{\pi}(s')$, which is the value of next state of the current sample that was provided.

The input is each sample which (s, a, s', r), and output will be updated v-values after each sample.

(b).

$\alpha = 0.5$, and initial all Q-values are 0, assume $\gamma = 1$:

1.$(G, \rightarrow, B, -2)$: $sample = R(G, \rightarrow, B) = -2 + 1 * 0 = -2$

$$Q(G, \rightarrow) = (1 - \alpha)Q(G, \rightarrow) + \alpha * sample = 0.5 * 0 + 0.5 * -2 = -1$$

2. $(B, \uparrow, B, -2)$: $sample = R(G, \uparrow, B) + \gamma * 0 = -2 + 1 * 0 = -2$

$$Q(B, \uparrow) = (1 - \alpha)Q(B, \uparrow) + \alpha * sample = 0.5 * 0 + 0.5 * -2 = -1$$

3. $(B, \leftarrow, B, -2)$: $sample = R(B, \leftarrow, B) + \gamma * 0 = -2 + 1 * 0 = -2$

$$Q(B, \leftarrow) = (1 - \alpha)Q(B, \leftarrow) + \alpha * sample = 0.5 * 0 + 0.5 * -2 = -1$$

4. $(B, \leftarrow, G, -1)$: $sample = R(B, \leftarrow, G) + \gamma * (-1) = -1 + 1 * 0 = -1$

$$Q(B, \leftarrow) = (1 - \alpha)Q(B, \leftarrow) + \alpha * sample = 0.5 * -1 + 0.5 * -1 = -1$$

5. $(G, \rightarrow, B, -2)$: $sample = R(G, \rightarrow, B) + \gamma * -1 = -2 + 1 * -1 = -3$

$$Q(G, \rightarrow) = (1 - \alpha)Q(G, \rightarrow) + \alpha * sample = 0.5 * -1 + 0.5 * -3 = -2$$

6. $(B, \uparrow, B, -2)$: $sample = R(B, \uparrow, B) + \gamma * (-1) = -2 + 1 * -1 = -3$

$$Q(B, \uparrow) = (1 - \alpha)Q(B, \uparrow) + \alpha * sample = 0.5 * -1 + 0.5 * -3 = -2$$

7. $(G, \rightarrow, G, -1)$: $sample = R(G, \rightarrow, G) + \gamma * 0 = -1 + 1 * 0 = -1$

$$Q(G, \rightarrow) = (1 - \alpha)Q(G, \rightarrow) + \alpha * sample = 0.5 * -2 + 0.5 * -1 = -1.5$$

8. $(G, \uparrow, G, -1)$: $sample = R(G, \uparrow, G) + \gamma * 0 = -1 + 1 * 0 = -1$

$$Q(G, \uparrow) = (1 - \alpha)Q(G, \uparrow) + \alpha * sample = 0.5 * 0 + 0.5 * -1 = -0.5$$

9. $(G, \rightarrow, B, -2)$: $sample = R(G, \rightarrow, B) + \gamma * -1 = -2 + 1 * -1 = -3$

$$Q(G, \rightarrow) = (1 - \alpha)Q(G, \rightarrow) + \alpha * sample = 0.5 * -1.5 + 0.5 * -3 = -2.25$$

10. $(G, \rightarrow, B, -2)$: $sample = R(G, \rightarrow, B) + \gamma * -1 = -2 + 1 * -1 = -3$

$$Q(G, \rightarrow) = (1 - \alpha)Q(G, \rightarrow) + \alpha * sample = 0.5 * -2.25 + 0.5 * -3 = -2.625$$

11. $(B, \leftarrow, G, -1)$: $sample = R(B, \leftarrow, G) + \gamma * -0.5 = -1 + 1 * -0.5 = -1.5$

$$Q(B, \leftarrow) = (1 - \alpha)Q(B, \leftarrow) + \alpha * sample = 0.5 * -1 + 0.5 * -1.5 = -1.25$$

12. $(B, \uparrow, W, 0)$: $sample = R(B, \uparrow, W) + \gamma * 0 = 0 - 0 = 0$

$$Q(B, \uparrow) = (1 - \alpha)Q(B, \uparrow) + \alpha * sample = 0.5 * -2 + 0.5 * 0 = -1$$

13. $(B, \leftarrow, G, -1)$: $sample = R(B, \leftarrow, G) + \gamma * -0.5 = -1 + 1 * -0.5 = -1.5$

$$Q(B, \leftarrow) = (1 - \alpha)Q(B, \leftarrow) + \alpha * sample = 0.5 * -1.25 + 0.5 * -1.5 = -1.375$$

14. $(G, \uparrow, G, -1)$: $sample = R(G, \uparrow, G) + \gamma * (-0.5) = -1 + 1 * -0.5 = -1.5$

$$Q(G, \uparrow) = (1 - \alpha)Q(G, \uparrow) + \alpha * sample = 0.5 * -0.5 + 0.5 * -1.5 = -1$$

15. $(G, \uparrow, W, 0)$: $sample = R(G, \uparrow, W) + \gamma * 0 = 0 + 0 = 0$

$$Q(G, \uparrow) = (1 - \alpha)Q(G, \uparrow) + \alpha * sample = 0.5 * -1 + 0 = -0.5$$

The final values are:

$$Q(B, \uparrow) = -1$$
$$Q(B, \leftarrow) = -1.375$$
$$Q(G, \uparrow) = -0.5$$

$$Q(G,\rightarrow) = -2.625$$

from the Q values above, $Q(G,\uparrow) > Q(B,\uparrow) > Q(B,\leftarrow) > Q(G,\rightarrow)$

for the agent reach to white state, Q values will lead the agent reach to the gray state and turn UP. If the agent is in gray state, he will turn UP. If the agent is in black state, turning UP will be the best choice for him because this action has larger value than turn LEFT.