Task 1.

State \ Action	up	down	left	right
E	0	0	0	0
T	0	0	0	0
S	0	0	0	0
B	0	0	0	0

Q-Table of Iteration „0"

$\alpha = 0.1$; $\gamma = 1$

- $r_E = -1$; $r_T = +100$; $r_B = -0.5$; $r_S = -1$
- $Q^{\pi}_{k+1}(x_t, u_t) = (1-\alpha) \cdot Q^{\pi}_{k}(x_t, u_t) + \alpha \left(r_{t+1} + \gamma \max_u Q(x_{t+1}, u) \right)$

By neglecting epsilon-greedy strategy, which action should be choosed? $\Rightarrow$ max Q or random u?

Step 2: Choose an action = e.g. $Q^{\pi}(S, up)$ with $x_1 = S$

Step 3: Perform an Action: $\rightarrow x_2 = E$

all Q-values are zero

Randomly select.

Step 4: Measure rewards: $r_E = -1$

Step 5: Update Q-Table: $Q^{\pi_v}_2(S, up) = (1-0.1) \times 0 + 0.1 \times (-1 + 1 \times 0)$

$= -0.1$

$\Rightarrow$ State E is now the current State.

State \ Action	up	down	left	right
E	0	0	0	0
T	0	0	0	0
S	-0.1	0	0	0
B	0	0	0	0

Q-Table of Iteration „1"

Task 2

State \ Action	up	down	left	right
E	0	-0.1	-0.1	0
T	0	0	0	0
S	-0.1	-0.1	-0.1	7.91
B	40.95	0	-0.1	0

Q-Table of Iteration „0"

$\alpha = 0.1$; $\gamma = 1$

- $r_E = -1$; $r_T = +100$; $r_B = -0.5$; $r_S = -1$
- $Q^{\pi}_{k+1}(x_t, u_t) = (1-\alpha) \cdot Q^{\pi}_{k}(x_t, u_t) + \alpha (r_{t+1} + \gamma \cdot \max_u Q(x_{t+1}, u))$

Step 2: Choose an action = e.g. $Q^{\pi}_1(S, \text{right})$ with $x_1 = S$

choose the largest Q-value in State S

Step 3: Perform an Action: $x_2 = B$

Step 4: Measure rewards: $r_B = -0.5$

Step 5: Update Q-Table: $Q^{\pi}_2(S, up) = (1-0.1) \times 7.91 + 0.1 \times (-0.5 + 1 \times 40.95)$

$= 11.764$

⇒ State B is now the current State.

State \ Action	up	down	left	right
⇒ E	0	-0.1	-0.1	0
T	0	0	0	0
S	-0.1	-0.1	-0.1	11.764
B	40.95	0	-0.1	0

Q-Table of Iteration „1"

Task 3 : ε-Greedy Strategy

$$Action = \begin{cases} \max_{u} Q(x,u), & R > \varepsilon \quad \Rightarrow R \text{ is uniform random value in } [0,1] \\ \text{Random } u, & R \leq \varepsilon \quad \Rightarrow \varepsilon(episode) = 0.9999^{episode} \end{cases}$$

- At the beginning, ε (epsilon) is close to 1, and the random number „R" is likely to be less than „ε", „Action" tends to be randomly selected.
- When the episode is large enough, the „ε" (epsilon) becomes very small, and the random number „R" is likely to be greater than „ε". The „Action" tends to be selected as the action with the largest Q-value in the current state.
- Q-Table: state S: Q(S, Left) will be the largest,
  Q(S up) will be larger than the rest two.
  State E: Q(E, right) will be the largest.
  State B: Q(B, up) will be the largest.
  State T = Target location, all Q-Value = 0.

State \ Action	up	down	left	right
E	-X	-X	-X	100
T	0	0	0	0
S	99	-X	-X	99.5
B	100	-X	-X	-X

Task 4

- When the episode is large enough, every Q-value update path will not pass through the square E, what means, the Q-value corresponding to state E will no longer be updated.