EX1.

1) For $V^{\pi}(s)$ to be $-\infty$, the agent should not reach any terminal state.
Let the policy $\pi(S_t) =$ right, $S_t = 3$, and state $s = 3$.
left, $S_t \neq 3$

Hence, $V^{\pi}(3) = -\infty$ since the agent will oscillate between grid 3 and 4.

2) $Q^{\pi}(9, \text{right}) = -1 + V(10) = -1$, since 10 is the terminal state, $V(10)=0$, one step reward is -1.

$Q^{\pi}(12, \text{left}) = -1 + V(11) = -1 + (-13.39649575) \approx -14.396$

3) ① Initialize $V(s) = -1$ for $s \in S - \{1, 10\}$, and $V(s) = 0$ for $s \in \{1, 10\}$
$\pi(s)$ arbitrary for all $s \in S - \{1, 10\}$

② Loop:
$\Delta \leftarrow 0$
Loop for each $s \in S$:
If s is terminal state ($s = 1 / 10$)
continue
$v \leftarrow V[s]$
$V(s) \leftarrow -1 + \sum_{a \in A} \pi(a|s) * V(s')$ ($\pi(a|s) = \frac{1}{4}$ for all a in A and s' is the next step from s taking a)
$\Delta \leftarrow \max(\Delta, |v - V(s)|)$
Until $\Delta < \theta$ (a small positive number determining the accuracy of estimation)

③ policy_stable $\leftarrow$ true
For each state $s \in S$ except terminal states:
If s is terminal state ($s = 1 / 10$)
continue
old_action $\leftarrow \pi(s)$
$\pi(s) \leftarrow \text{argmax}_{a \in A} (-1 + \sum_{s' \in S} p(s'|s,a) * V(s'))$

If old_action $\neq \pi$, then policy_stable $\leftarrow$ false
If policy_stable, then stop and return $V \approx v_*$ and $\pi \approx \pi_*$
else go to ②

4) V^{π}:

0	-1	-2	-3
-1	-1	-2	-3
-1	0	-1	-2
-2	-1	-2	-3

π:

X	←	←	←
↑	↓	↓	←
→	X	←	←
↑	↑	↑	↑

"X": terminal state
the arrow points to the optimal move direction

EX3.

policy Iteration:

① Initialization

$Q(s,a) = 0$ for all $s \in S$ and $a \in A$
$\pi(s) \in A(s)$ arbitrarily for all $s \in S$

② Policy Evaluation

Loop:
$\Delta \leftarrow 0$
Loop for each $s \in S$:
Loop for each $a \in A$:

$\text{old\_q} \leftarrow Q(s,a)$
$Q(s,a) \leftarrow \sum_{s',r} p(s'|s,a)(r + \gamma \sum_{a'} \pi(a'|s') Q(s',a'))$
$\Delta \leftarrow \max(\Delta, |\text{old\_q} - Q(s,a)|)$
Until $\Delta < \theta$ (a small positive number determining the accuracy of estimation)

③ Policy Improvement

policy-stable ← true

For each $s \in S$:

old_action ← $\pi(s)$

$\pi(s) \leftarrow \underset{a \in A}{\text{argmax}}\ Q(s,a)$

If old_action ≠ $\pi(s)$, then policy-stable ← false

If policy_stable, then stop and return $Q \approx q_*$, and $\pi \approx \pi_*$

Else go to ②

Value Iteration:

Initialization

$Q(s,a) = 0$ for all $s \in S$ and $a \in A$

Loop:

$\Delta \leftarrow 0$

Loop for each $s \in S$:

Loop for each $a \in A$:

$old\_q \leftarrow Q(s,a)$

$Q(s,a) \leftarrow \sum_{s',r} p(s'|s,a)\left(r + \gamma \max_{a' \in A}(Q(s',a'))\right)$

$\Delta \leftarrow \max(\Delta, |old\_q - Q(s,a)|)$

Until $\Delta < \theta$ (a small positive number determining the accuracy of estimation)

For each $s \in S$:

$\pi(s) \leftarrow \underset{a \in A}{\text{argmax}}\ Q(s,a)$

Return Q and π