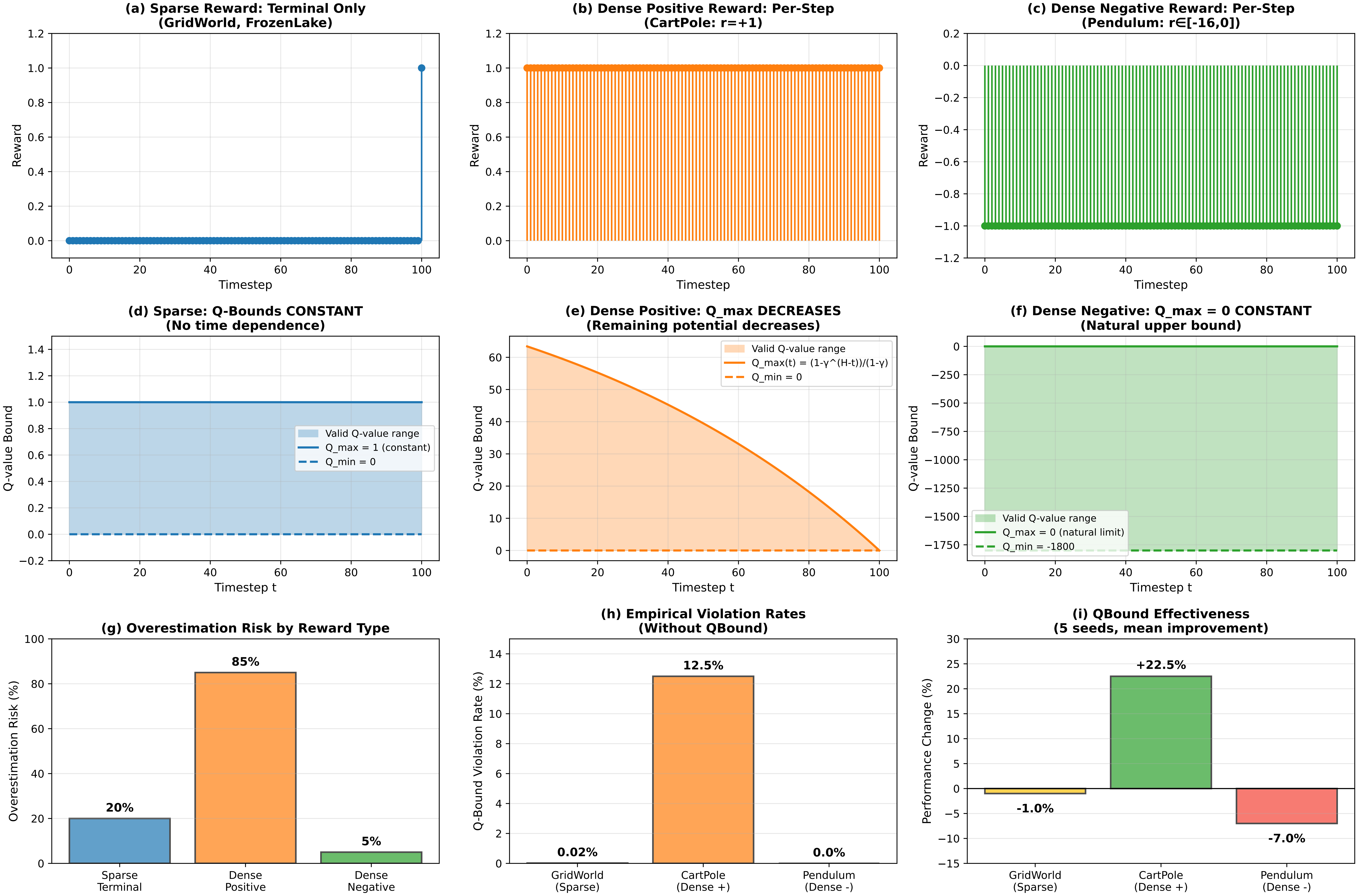
(a) Sparse Reward: Terminal Only
(GridWorld, FrozenLake)
Reward
Timestep
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
(b) Dense Positive Reward: Per-Step
(CartPole: r=+1)
Reward
Timestep
(c) Dense Negative Reward: Per-Step
(Pendulum: r∈[-16,0])
Reward
Timestep
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
−1.2
(d) Sparse: Q-Bounds CONSTANT
(No time dependence)
Q-value Bound
Timestep t
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
Valid Q-value range
Q_max = 1 (constant)
Q_min = 0
(e) Dense Positive: Q_max DECREASES
(Remaining potential decreases)
Q-value Bound
Timestep t
60
50
40
30
20
10
0
Valid Q-value range
Q_max(t) = (1-γ^(H-t))/(1-γ)
Q_min = 0
(f) Dense Negative: Q_max = 0 CONSTANT
(Natural upper bound)
Q-value Bound
Timestep t
0
−250
−500
−750
−1000
−1250
−1500
−1750
Valid Q-value range
Q_max = 0 (natural limit)
Q_min = -1800
(g) Overestimation Risk by Reward Type
Overestimation Risk (%)
100
80
60
40
20
0
20%
85%
5%
Sparse
Terminal
Dense
Positive
Dense
Negative
(h) Empirical Violation Rates
(Without QBound)
Q-Bound Violation Rate (%)
14
12
10
8
6
4
2
0
0.02%
12.5%
0.0%
GridWorld
(Sparse)
CartPole
(Dense +)
Pendulum
(Dense -)
(i) QBound Effectiveness
(5 seeds, mean improvement)
Performance Change (%)
30
25
20
15
10
5
0
−5
−10
−15
-1.0%
+22.5%
-7.0%
GridWorld
(Sparse)
CartPole
(Dense +)
Pendulum
(Dense -)