Reinforcement Learning – Mid Semester Assignment

Nadav Shaked 312494925, Michael Glustein 203929500

Grid World Problem

Goal: Reach top left or bottom right corner.

Actions: U – up, D – down, R – right, L – left.

Rewards: -1 for each step.

Transition Model: Deterministic

Discount Factor: $\gamma = 1$

Random Policy: 0.25 probability of any direction

Values (partially given for a specific state, by current policy):

0	S1	-20	-22
-14	-18	-20	S2
-20	-20	-18	-14
-22	-20	-14	**0**

Solution

Since $\gamma \leq 1$, we know the policy evaluation will always converge.

Let's recall the equation:

$$V_\pi(s) = \sum_a \pi(a|s) \cdot \sum_{s',r} p(s',r|s,a)\,[r + \gamma V_\pi(s')]$$

$$V(S1) = 1 \cdot 0.25 \cdot (-1 + 1 \cdot 0) + 1 \cdot 0.25 \cdot \big(-1 + 1 \cdot (-20)\big) + 1 \cdot 0.25 \cdot \big(-1 + 1 \cdot (-18)\big) + 1 \cdot 0.25 \cdot \big(-1 + 1 \cdot V(S1)\big)$$

$$V(S1) = -0.25 - 5.25 - 4.75 - 0.25 + 0.25\big(V(S1)\big) = -\frac{10.5}{0.75} = \boxed{-14}$$

$$V(S2) = 1 \cdot 0.25 \cdot (-1 + 1 \cdot (-22)) + 1 \cdot 0.25 \cdot \big(-1 + 1 \cdot (-20)\big) + 1 \cdot 0.25 \cdot \big(-1 + 1 \cdot (-14)\big) + 1 \cdot 0.25 \cdot \big(-1 + 1 \cdot V(S2)\big)$$

$$V(S2) = -5.75 - 5.25 - 3.75 - 0.25 + 0.25\big(V(S2)\big) = -\frac{15}{0.75} = \boxed{-20}$$