Assignment 1

Given initial value of 0 and policy of [0.25, 0.25, 0.25, 0.25] for all states, except all zero for terminal states S1 and S35. In policy evaluation and policy iteration, both value and policy is given, while for value iteration, only value is given.

1. Policy Evaluation

$$v_{k+1}(s) = \sum_{a\in\mathcal{A}} \pi(a|s)\left(\mathcal{R}_s^a + \gamma \sum_{s'\in\mathcal{S}} \mathcal{P}_{ss'}^a v_k(s')\right)$$

In the given case, A includes moving north, south, west and east. For special cases, such as being on a boundary, any attempt to move out will end up as no movement.

In one iteration, each state will update its value using the formula mentioned above. And calculate the change of each state value. If the value is smaller than a certain threshold, the iteration would end.

When the threshold is set to 0.05, evaluation ends after 157 iteration, it’s 234 iterations for threshold of 0.01. The value and policy of each state are shown below.

−18.1→	0.0	−29←	−43.7←	−51.2←	−54.3←
−32.1↑	−29.9↑	−39.3↑	−47.1←	−51.5←	−53.4↓
−44.3↑	−44.4↑	−47.2↑	−49.7↑	−50.6↓	−50.4↓
−52.5→	−52.1↑	−51.5↑	−49.8→	−46.7↓	−43.3↓
−57.3→	−55.9↑	−53.1→	−47.6→	−39.1→	−28.9↓
−59.3→	−57.4→	−53.0→	−44.6→	−29.2→	0.0

It can be proved that each state would take the shortest path to destination.

2. Policy iteration

$$v_\pi(s) = \max_{a\in\mathcal{A}} q_\pi(s,a)$$

The case is quite similar for policy iteration. After each round of iteration, the policy would be updated according to the value of neighbor states. And the iteration would end after no more change in policy is made.

In this case, it stopped after 6 iterations. The value of each state are shown below, and the value is actually the distance to nearest neighbor.

−1	0	−1	−2	−3	−4
−2	−1	−2	−3	−4	−4
−3	−2	−3	−4	−4	−3
−4	−3	−4	−4	−3	−2
−5	−4	−4	−3	−2	−1
−5	−4	−3	−2	−1	0

3. Value iteration

$$v_{k+1}(s) = \max_{a \in \mathcal{A}} \left(\mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v_k(s') \right)$$

In value iteration, we don’t need to know what the initial policy is. In each iteration, value of state is updated using its best neighbor.

Similar to policy evaluation, iteration stop when the maximum change of state is under a certain threshold.

It stopped after 6 iterations. The value is the same as policy iteration.

−1	0	−1	−2	−3	−4
−2	−1	−2	−3	−4	−4
−3	−2	−3	−4	−4	−3
−4	−3	−4	−4	−3	−2
−5	−4	−4	−3	−2	−1
−5	−4	−3	−2	−1	0