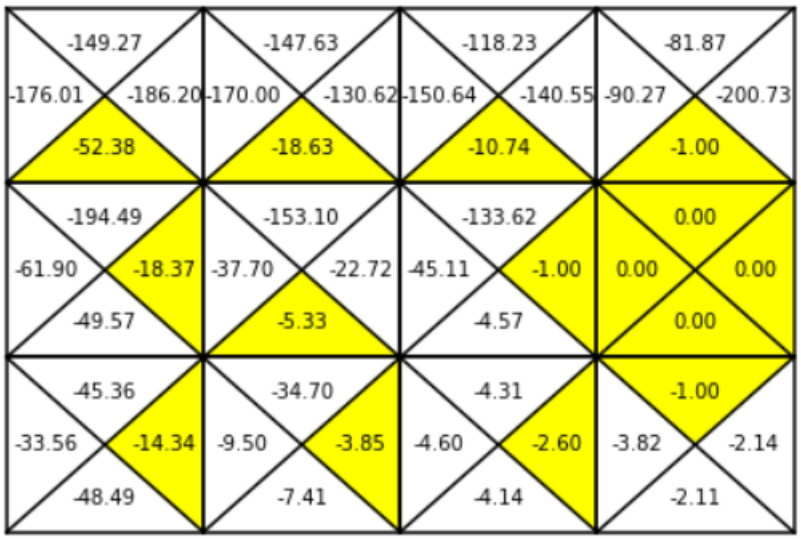
108061217 鍾永桓

1.



2. q_values 大致上是合理的，由這張圖可以看出只要踩進 swamp，也就是最上面一行時，向下的 action 的 q_value 最大，因為踩在 swamp 會使 reward 大幅下降，所以應該學習的結果應該會偏向離開 swamp，故向下的 q_value 最大，至於中間那行，q_value 最大的可能是向右或向下，仍算合理，會選擇向右的原因在於希望可以快速達到 terminal state，因為每多一個 action，reward 便會下降，而 q_value 會偏向向下的原因在於，policy 在選擇行動時有一定機率隨機選擇，所以導致在中間那行有非常高的踩進 swamp 的風險，所以為了躲避 swamp，也可能會學習到向下來遠離 swamp，而最下面一行除了最右邊的一格外 q_value 都是向右最大，主要是因為學習到的結果會偏好於向 terminal state 接近但又同時盡可能遠離 swamp，因此在向右有最高的 q_value。

3.

	1	2	3	4
1	-78.95	-48.06	-39.03	-28.74
2	-37.19	-20.14	-14.52	0.0
3	-20.67	-6.86	-3.0	-1.38

由此圖來看 state_value 由收斂至合理的值，在最上面一行因為位於 swamp，做出任何行動有很高機率仍在 swamp，因此 state_value 最低，而第二行雖然不在 swamp 上，而且離 terminal state 最為接近，但因為相比於最下面一行離 swamp 過於接近，很容易因為策略的隨機選擇而落入 swamp，所以 state_value 仍較第三行更為低，而這三行當中都是越往右 state_value 越高，因為越來越接近 terminal state。