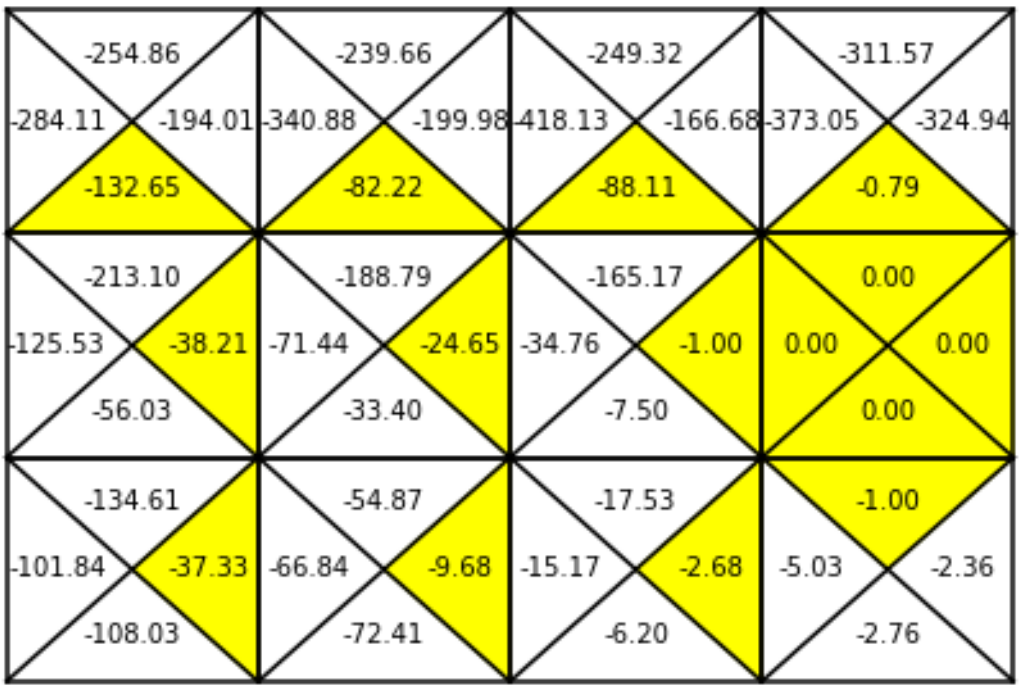
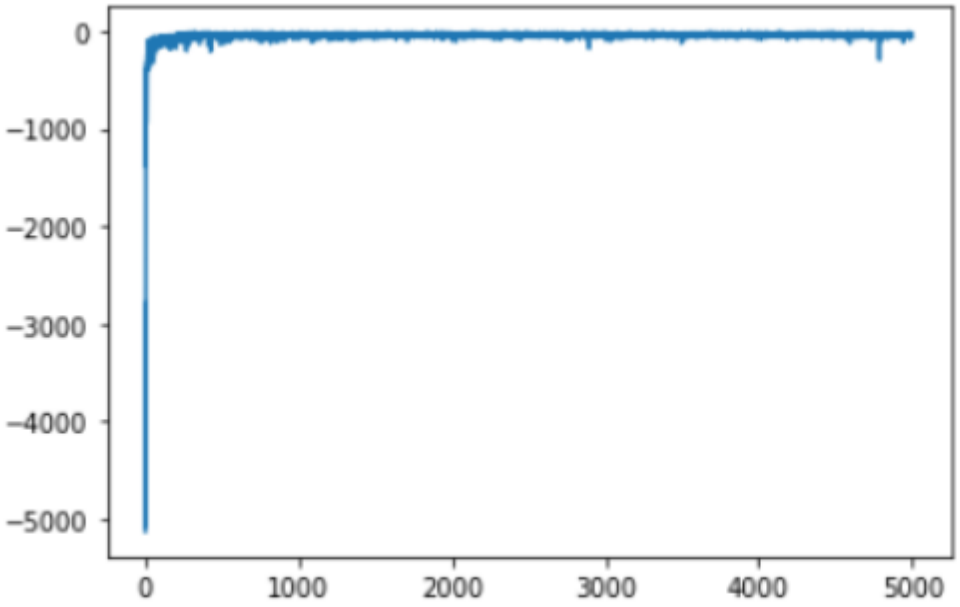
1. State action value

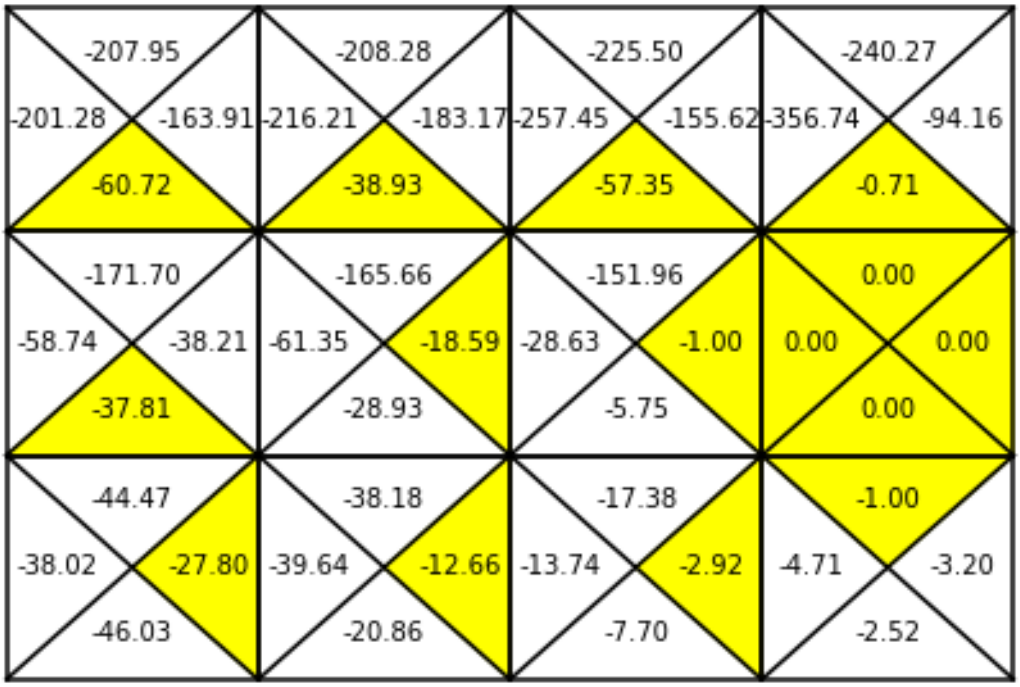
Every-visit MC control



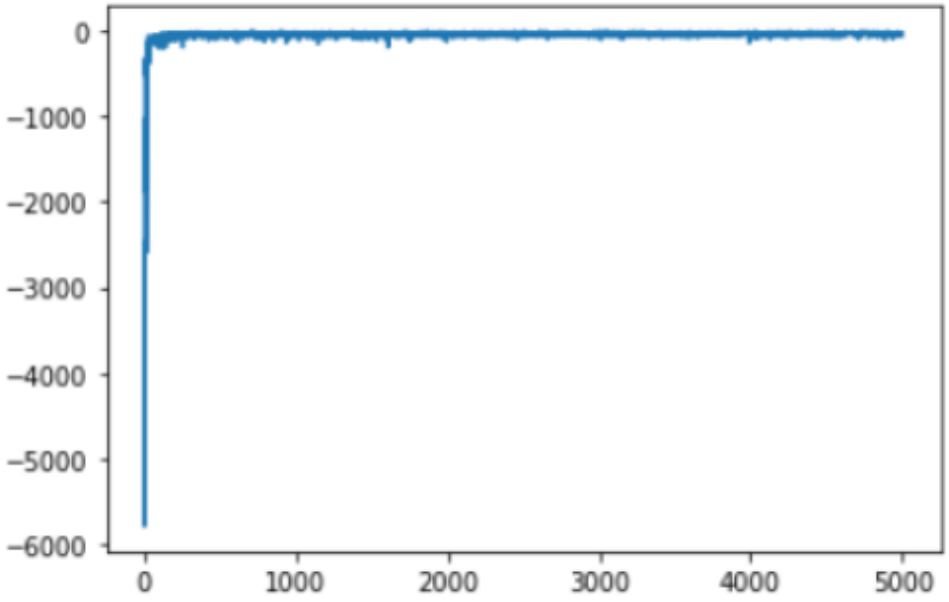
Learning curve



First-visit MC control



Learning curve



2.

是合理的，因為從 state action value 圖中可以觀察到，基本在 swamp 中，往下的動作和往右的動作價值最大，因為可以脫離沼澤或是更接近終點。

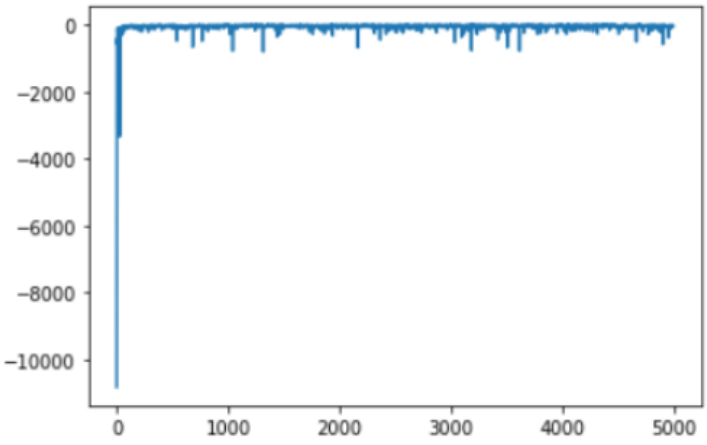
而終點的動作價值都是 0 因為是在終點。而圖中可以看到起點的動作價值最大是往下動作可以離沼澤遠一點，因為 epsilon-greedy action selection，所以有機會往上走到沼澤，因此往下的 Reward 比往右的 Reward 平均會高一點。

3. State value

Every-visit MC control

	1	2	3	4
1	-176.16	-310.78	-76.38	-243.41
2	-33.57	-30.34	-14.85	0.0
3	-43.16	-26.44	-8.31	-2.2

Learning curve



First-visit MC control

	1	2	3	4
1	-91.24	-131.57	-71.53	-502.71
2	-28.07	-31.78	-18.38	0.0
3	-20.07	-8.49	-6.17	-1.74

Learning curve