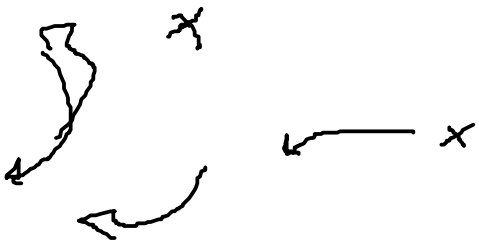
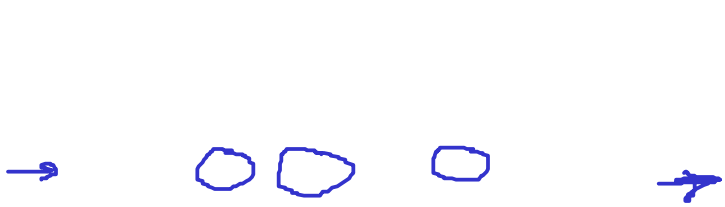




$$Q(state, action) \leftarrow (1-\alpha)Q(state, action) + \alpha\left(reward + \gamma \max_{a} Q(next\ state, all\ actions)\right)$$

* assume $\alpha = 1$ (Learning Rate) $\gamma$ - (discount Rate)

$$Q(S,a) = R(s,a) + \gamma . \max[Q(ns, At)]$$

$R =$

| State \ Action | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| 0 | -1 | -1 | -1 | -1 | 0 | -1 |
| 1 | -1 | -1 | -1 | 0 | -1 | 100 |
| 2 | -1 | -1 | -1 | 0 | -1 | -1 |
| 3 | -1 | 0 | 0 | -1 | 0 | -1 |
| 4 | 0 | -1 | -1 | 0 | -1 | 100 |
| 5 | -1 | 0 | -1 | -1 | 0 | 100 |

$Q =$

| state \ Action | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 80 | 0 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 100 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 |

$\gamma = 0.8$ $Q^{epoch=1}$

$Q(1,3) = R(1,3) + \gamma \max(Q(3,1/2/5))$

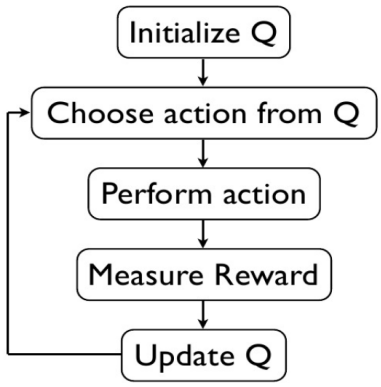
$Q(1,3) = 0 + 0.8 \max(0,0,0)$

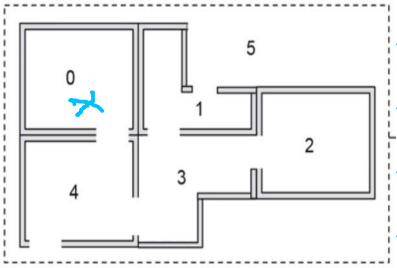
$Q(2,3) = 0$

$Q(3,4) = R(3,4) + \gamma \max(Q(4,0/3/5))$

$Q(3,4) = 0 + 0.8 \max(0,100,0)$

$Q(3,5) = 0$




$Q(4,5) = 100 + 0.8 \max(Q(5,1), Q(5,4), Q(5,5))$

$Q(4,5) = 100 + 0 = 100$

epoch = 2 $Q(0,4) = R(0,4) + \gamma \max(Q(4,0), Q(4,3), Q(4,5))$

$Q(0,4) = 0 + 0.8(0,0,100) = 80$

| actions \ states | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| 0 | 0. | 0. | 0. | 0. | 80. | 0. |
| 1 | 0. | 0. | 0. | 64. | 0. | 100. |
| 2 | 0. | 0. | 0. | 64. | 0. | 0. |
| 3 | 0. | 80. | 51.2 | 0. | 80. | 0. |
| 4 | 64. | 0. | 0. | 64. | 0. | 100. |
| 5 | 0. | 80. | 0. | 0. | 80. | 100. |