$$Q(s,a) = Q(s,a) + \alpha[r + \gamma \max Q(s',a') - Q(s,a)]$$

Initial

$\alpha = 0.5 \quad \gamma = 0.5$

	a1	a2
S1	0	0
S2	0	0
S3	0	0

s: S1

a: a1

r: 0.5

s': S2

$Q(S1,a1) = 0 + 0.5[0.5 + 0.5(0) - 0] = 0.25$

	a1	a2
S1	0.25	0
S2	0	0
S3	0	0

s: S2

a: a2

r: -1

s': S3

$Q(S2,a2) = 0 + 0.5[-1 + 0.5(0) - 0] = -0.5$

	a1	a2
S1	0.25	0
S2	0	-0.5
S3	0	0

Optimal policy

- from state 1, best action is a1 → S2 because it yields an reward of 0.5 or 10 from S2 → S1 by a1
- from S2, best action is a1 → S1 because it has a reward of 10 from S1 → S2 through a loop. So you get rewards of 10 and 0.5 infinitely.
- from S3, best action is a2, taking you to S2. Yields a reward of 10. It also has a loop from S2 → S1 resulting in a self loop which yields a constant reward of 0.5 and 10