Juan Miguel Valverde Martinez

Sarsa Algorithm

$(s_t, a_t, r_t, s_{t+1}, a_{t+1})$

Characteristics

Monte Carlo

Dynamic Programming

Time Difference (TD)

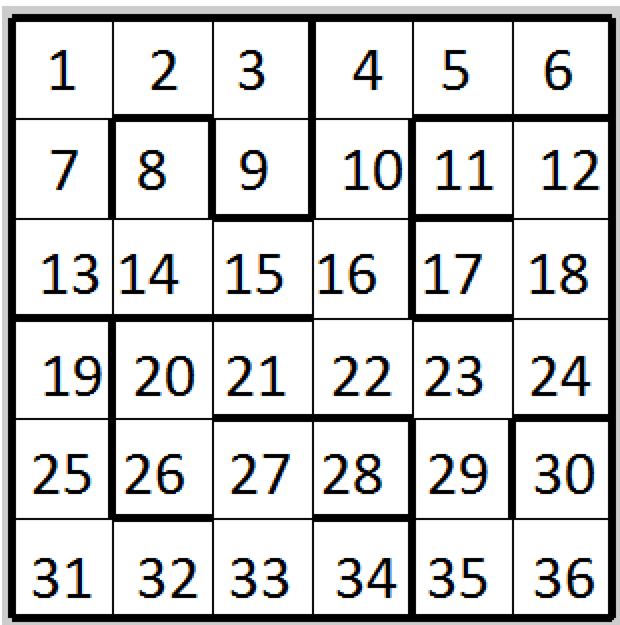
On Policy

Algorithm

Initialize $Q(s,a)$ arbitrarily
Repeat (for each episode):
  Initialize s
  Choose a from s using policy derived from Q (e.g., ε-greedy)
  Repeat (for each step of episode):
    Take action a, observe r, s'
    Choose a' from s' using policy derived from Q (e.g., ε-greedy)
    $Q(s,a) \leftarrow Q(s,a) + \alpha\big[r + \gamma Q(s',a') - Q(s,a)\big]$
    $s \leftarrow s'$; $a \leftarrow a'$;
  until s is terminal

Algorithm

Model



Rewards

30 → 36:	2
35 → 36:	2
Else:	0

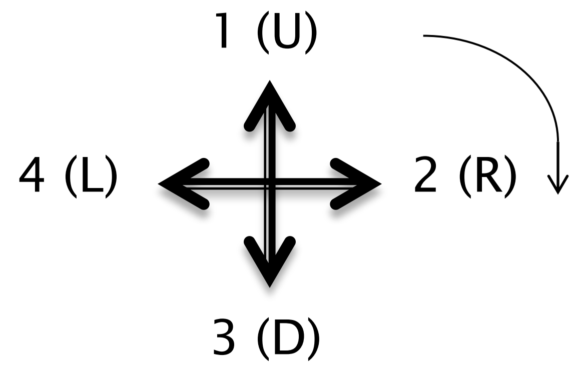
Why 0?

Because the algorithm does not know when and in which direction it has to go.

Algorithm

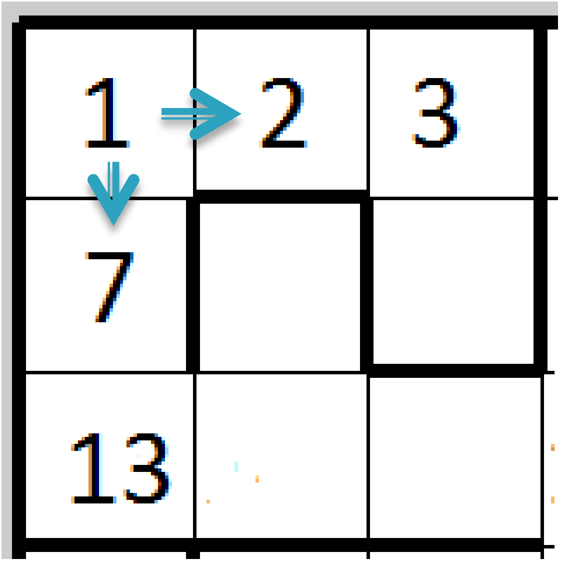
Structure Q(s,a)

X_1: U_1 R_1 D_1 L_1
X_2: U_3 R_2 D_2 L_2
X_3: U_2 R_3 D_3 L_3
…
…
X_{36}: U_{36} R_{36} D_{36} L_{36}



Algorithm

First iteration



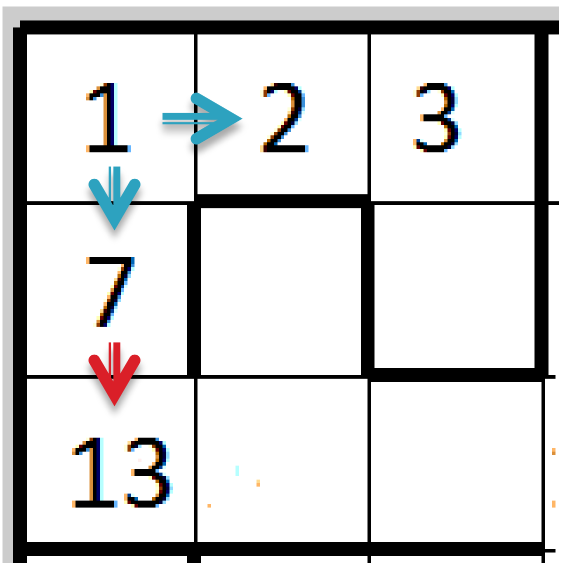
S: state 1
A: best action → · ε prob. to choose a random action.
· if same reward, random action
R:
S:
B:

↓

Both have a reward of 0, so it will choose a random action.

Algorithm

First iteration



S: state **1**
A: action **3 (down)**
R: reward **0**
S: state **7**
A: best action →

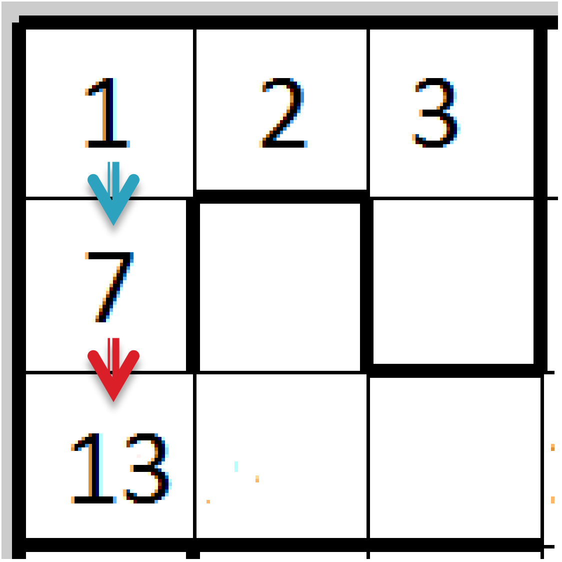
- ε prob. to choose a random action.
- if same reward, random action

↓

Both have a reward of 0, so it will choose a random action.

Algorithm

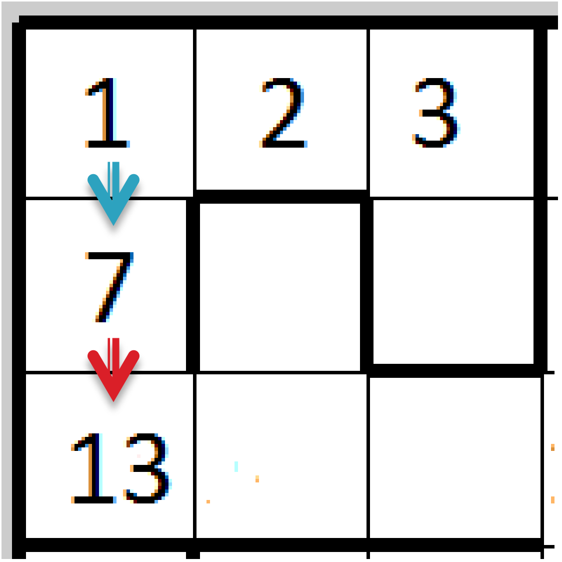
First iteration



S: state **1**
A: action **3 (down)**
R: reward **0**
S: state **7**
A: action **3 (down)**

Algorithm

First iteration



Algorithm

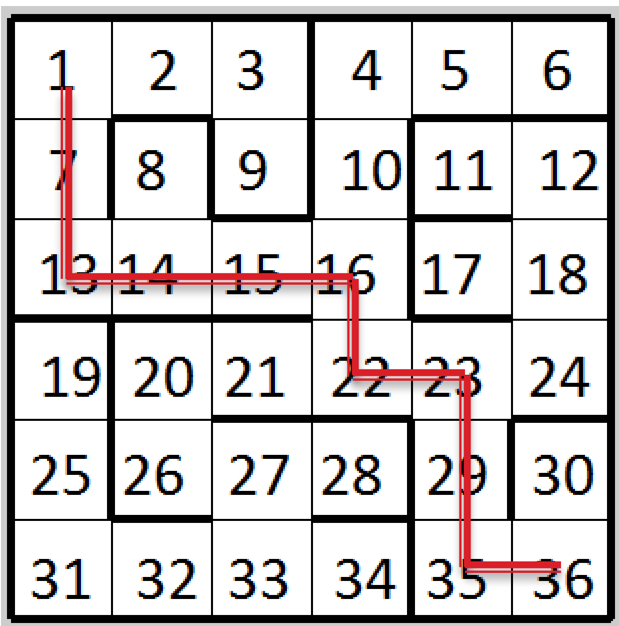
$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma Q(s',a') - Q(s,a)]$$

$$Q(1,3) \leftarrow Q(1,3) + \alpha[0 + \gamma Q(7,3) - Q(1,3)]$$

α: learning rate
r: reward
γ: discount factor

Algorithm

After many iterations



Final Q(s,a)

1:	0	0.0000	**0.0007**	0
7:	0.0002	0	**0.0033**	0
13:	0.0006	**0.0143**	0	0
14:	0.0007	**0.0276**	0	0.0001
15:	0	**0.0758**	0	0.0055
16:	0.0043	0	**0.3093**	0.0239
22:	0.0814	**0.4906**	0	0.0546
23:	0	0.0063	**0.9059**	0.1223
29:	0.2403	0	**1.3936**	0
35:	0.4974	**1.9444**	0	0

Juan Miguel Valverde Martinez

Explanation

Complementary explanation

http://laid.delanover.com/reinforcement-learning-sarsa-algorithm-a-practical-case

Video

https://www.youtube.com/watch?v=MZSK-Ho2-NA