Value iteration

$$V(s) = \max a \in A(s) \sum_{s' \in S} P_a(s'|s)[r(s,a,s') + \gamma V(s')]$$

Algorithm

Value iteration:

1) Set V_0 to arbitrary value for each s in S (choose 0 as the value)
2) While diff is >= epsilon
   a. For each s in S do
      i. V_t+1(s) := $\max a \in A(s) \sum_{s' \in S} P_a(s'|s)[r(s,a,s') + \gamma V\_t(s')]$
3) Select policy

initialise V(s) with random value

close to optimal

given a value to terminate (small diff)

set a fixed time to iterate

between previous state & next state.

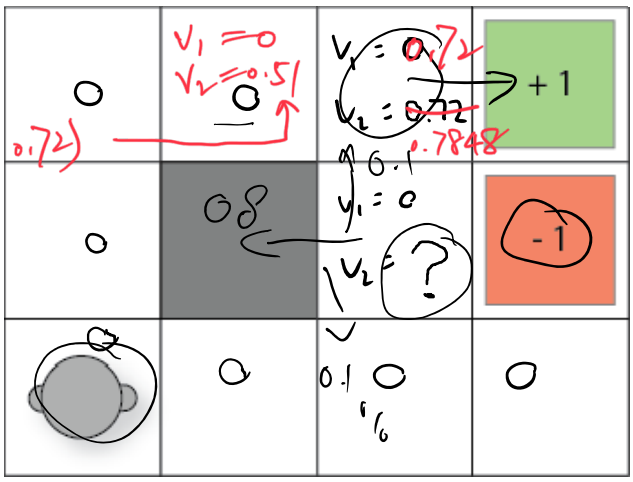
0 for all due to the initialisation

∴ we use previous time step

Value Iteration

V_t+1(s) :=

$$\max a \in A(s) \sum_{s' \in S} P_a(s'|s)[r(s,a,s') + \gamma V\_t(s')]$$



Assuming $\gamma = 0.9$

v_2=0.8*(0 + 0.9*1) +
0.1*(0 + 0.9*0) + 0.1*(0 + 0.9*0) = 0.72

0.8 × (0 + 0.9 × 0) + ... = 0

0.8 × (0 + 0.9 × 0) = V_0

0.8 × (0 + 0.9 × 1) + 0.1 × (0 + 0.9 × 0) + 0.1 × (0 + 0.9 × 0.72) = 0.7848

Deciding how to act

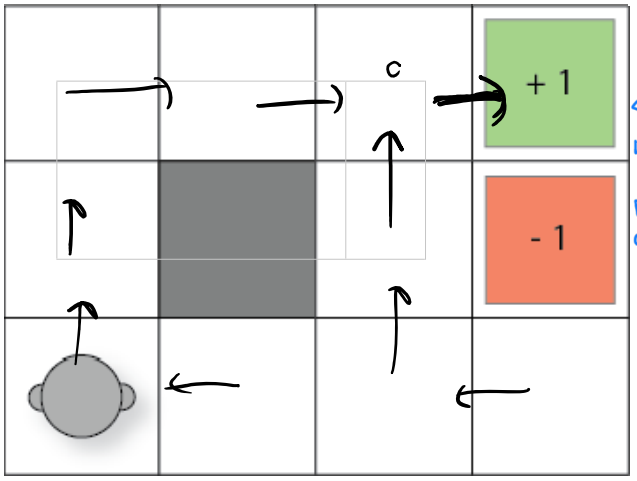
$$\text{Argmax}_{a \in A} \sum_{s' \in S} P_a(s'|s)\left[r(s,a,s') + \gamma V(s')\right]$$

$Q(s,a)$

Just intuition in Poosh

Policy iteration

0.6 * (2 + 0.9*10) = 6.6
0.4 * (0 + 0.9*12) = 4.32



π (random assign initially)

① Policy evaluation (How good is this state here if taking the given action)

② Policy update (choose the max (Best) action)

Stop when no policy change

VI $O(|S|^2|A|)$

PI $O(|S|^2|A| + |A|^3)$

$$V^{\pi}(s) = \sum_{s' \in S} P(s'|s)\,[r(s,a,s') + \gamma\, V^{\pi}(s')]$$

$a = \pi(s)$