Value Policy Iteration MDP. WS 7

$Q(s,a) = \sum_{s'} P_a(s'|s)\,[r(s,a,s') + \gamma V(s')]$.

$V(s,a) = \max_{a \in A} Q(s,a)$.

① $\gamma = 1.0$

$Q(\text{Messi}, \text{pass}) = P_{\text{pass}}(\text{Suarez}|\text{Messi})[r(\text{Messi}, \text{pass}, \text{Suarez}) + \gamma V(\text{Suarez})]$

$= 1 \times [0 - 1 + 1 \times (-1.2)] = -2.2$

$Q(\text{Messi}, \text{shoot}) = P_{\text{shoot}}(S|M)[r(M, \text{shoot}, S_{\text{uarez}}) + \gamma V(S_{\text{uarez}})]$

$+ P_{\text{shoot}}(S_{\text{coral}}|M)[r(\text{Mes}, \text{shoot}, S_{\text{coral}}) + \gamma V(S_{\text{coral}})]$

$= 0.8 \times [-2 + 1 \times (-1.2)] = 0.8 \times (-3.2) = -2.56$ (3.2 × 0.8 = 2.56)

$+ 0.2 \times (-2 + 1 \times 1) = -0.2$.

$= -2.76$.

$V(\text{Messi}, a) = \max(Q(\text{Messi}, \text{pass}), Q(\text{Messi}, \text{shoot}))$

$= \max(-2.2, -2.76) = -2.2$.

$\therefore$ pass

② $Q(S_u, \text{pass}) = P_{\text{pass}}(M|S_u)[r(S_u, \text{pass}, M) + \gamma V(M)] = -3$.

$Q(S_u, \text{shoot}) = P_{\text{shoot}}(M|S_u)[r(S_u, \text{shoot}, M_e) + \gamma V(M)]$

$+ P_{\text{shoot}}(S_{\text{coral}}|S_u)[r(S_u, \text{shoot}, S_{\text{coral}}) + \gamma V(S_{\text{coral}})]$

$= 0.4 \times [-2 + 1 \times (-2)] + 0.6 \times [-2 + 1 \times 1] = -2.2$

$V(S_{\text{coral}}) = Q(S_{\text{coral}}, \text{retire}) = P_{\text{retire}}(\text{Messi}, S_{\text{coral}})[r(S_{\text{coral}}, \text{retire}, M_e + \gamma V(M)]$

$= 1 \times [2 + 1 \times (-2)] = 0$.

③ $\gamma = 0.8$. policy -

$V^{\pi}(M) = Q^{\pi}(M, \text{Pass}) = P_{\text{pass}}(S_u|M)[r(M, \text{pass}, S_u) + \gamma V^{\pi}(S_u)] = 1 \times [-1 + \gamma V^{\pi}(S_u)]$

$= \gamma V^{\pi}(S_u) - 1$.

$V^{\pi}(S_u) = Q^{\pi}(S_u, \text{pass}) = 1 \times [-1 + \gamma V^{\pi}(M)] = \gamma V^{\pi}(M) - 1$.

$V^{\pi}(S_{\text{coral}}) = Q^{\pi}(S_{\text{coral}}, \text{return}) = 1 \times [2 + \gamma V^{\pi}(M)] = 2 + \gamma V^{\pi}(M)$

$\begin{cases} a = \gamma b - 1 \\ b = \gamma a - 1 \\ c = 2 + \gamma a \end{cases}$

$a = \gamma(\gamma a - 1) - 1 = \gamma^2 a - \gamma - 1 = a$

$(\gamma^2 - 1)a = \gamma + 1$

$a = \frac{\gamma + 1}{(\gamma + 1)(\gamma - 1)} = \frac{1}{\gamma - 1}$

$\therefore \begin{cases} a = \frac{1}{\gamma - 1} = -5 \\ b = \frac{1}{\gamma - 1} = -5 \\ c = 2 + 0.8 \times (-5) = -2 \end{cases}$

$Q^{\pi}(M,S) = 0.8 \times [-2 + 0.8 \cdot V(S_W)] + 0.2 \times [-2 + 0.8\, V(\text{scared})]$

$= 0.8 \times (-6) + 0.2 \times (-3.6) = -4.8 + (-0.72)$

$= -5.52.$

$Q^{\pi}(S,S) = 0.4 \times [-2 + 0.8\, V(M)] + 0.6 \times [-2 + 0.8\, V(\text{scared})]$

$= 0.4 \times (-6) + 0.6 \times (-3.6)$

$= -2.4 + (-2.16)$

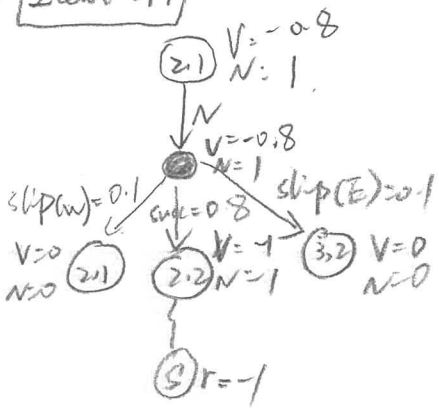
$= -4.56.$

3.6
× 0.6
2.16

WS 8 MCTS

$$V(s) = \max \sum P_a(s'|s)[r + \gamma V(s')]$$

① Iteration 1



sled (2,1)
expand N.
Sliva on slip(S).
Backup reveals -1.

② $Q((2,1),E) = -0.8$

$N = 0.08$

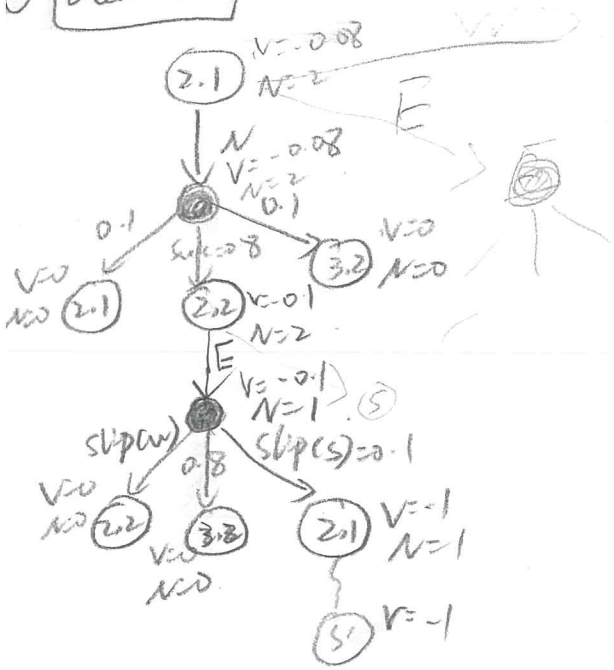
$Q((2,1),W) = 0.8$

$S = 0$

$UCT = \text{argmax}\, Q(s,a) + 2 \cdot c_p \sqrt{\frac{2\ln N(s)}{N(s,a)}}$

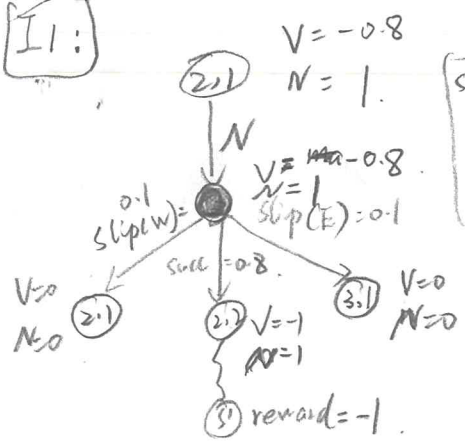
$$\pi(s) = \begin{cases} E: -0.8 + \sqrt{\frac{2\ln(5)}{1}} \\ N: 0.08 + \sqrt{\frac{2\ln(s)}{3}} \\ W: 0.8 + \sqrt{\frac{2\ln 5}{1}} \\ S: \infty \end{cases}$$

② Iteration 2



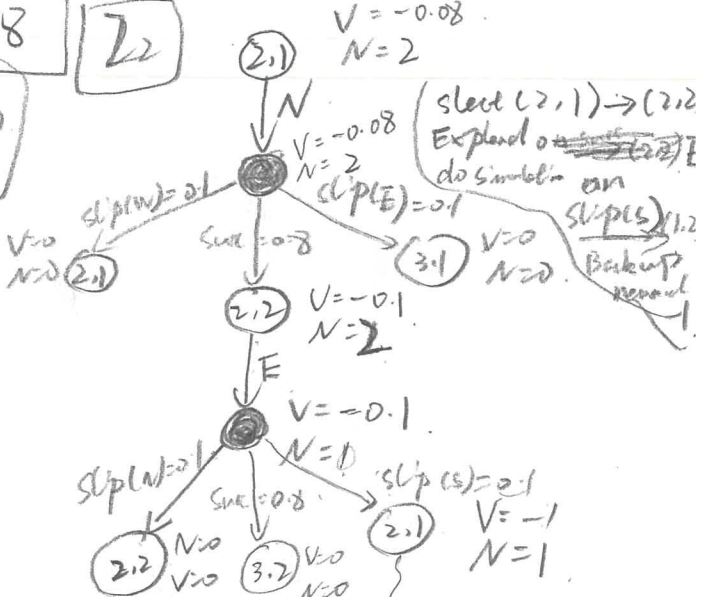
Week 8

I1:



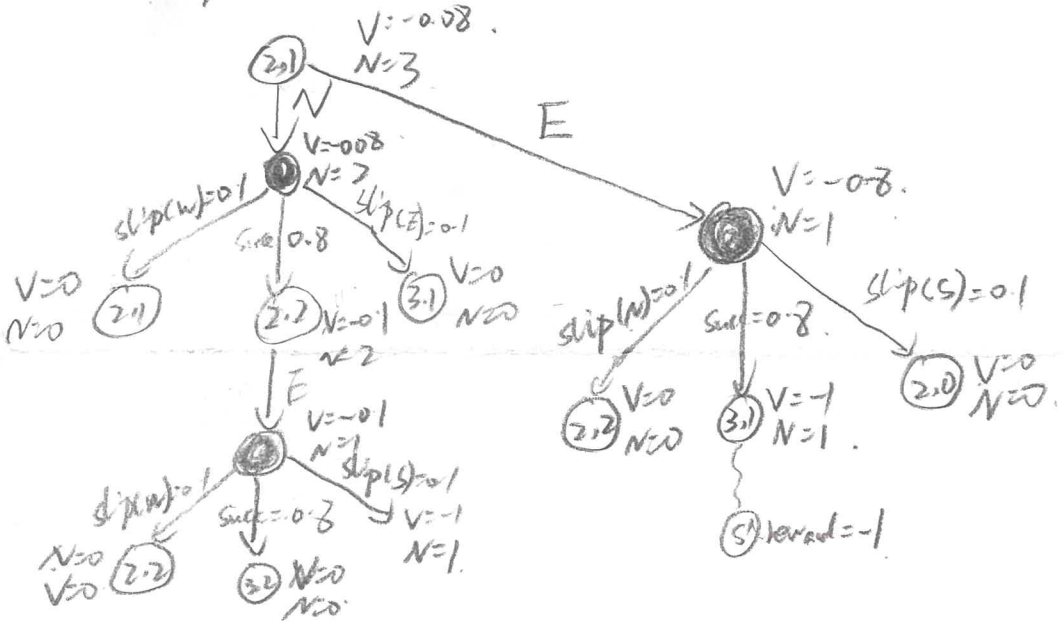
select (2,1), expand N,
do simulation on succ → (2,2)
Backup the reward -1

I2

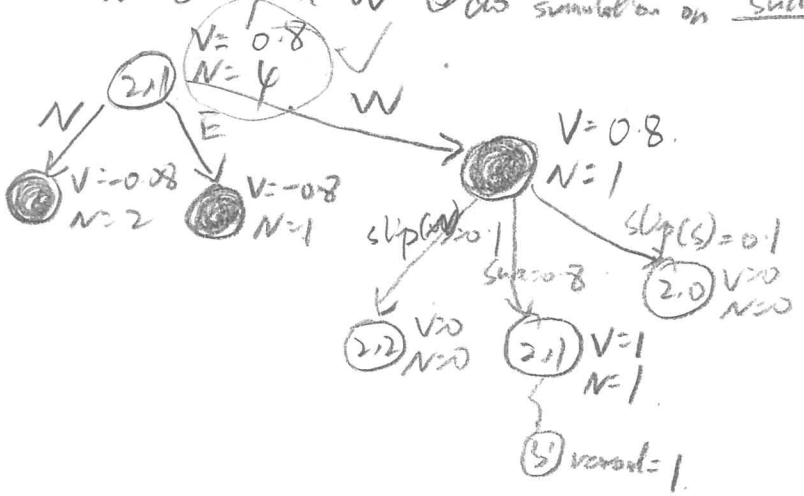


$V(s) = \max\{\ 0.1\times(0+\gamma 1\times 0)+0+0.8\times(-1+1\times 0)) = -0.8$

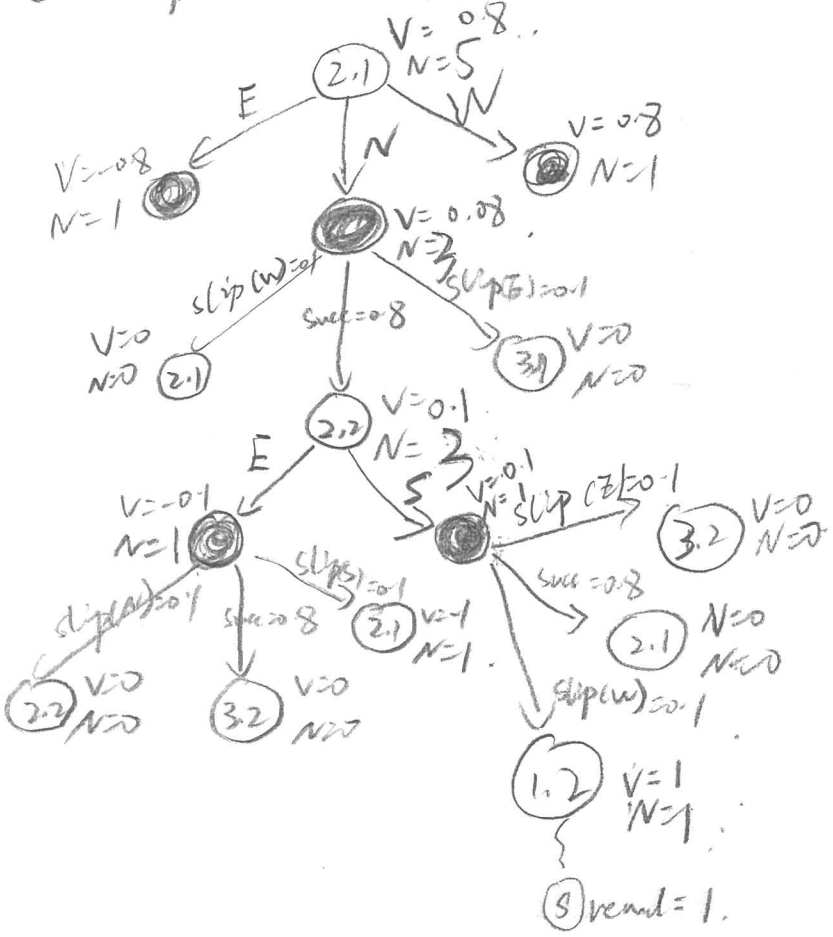
I3 ① select (2,1) ② expand E ③ do simulation on succ → (3,1)
④ Backup the rewards -1.



I4 ① Select (2,1) ② expand W ③ do simulation on succ → (2,1) ④ Backup the reward -1.



25: ① Select (2,1), $\xrightarrow{S_1}$ (2,2) ② expand S: ③ do simulation on $\xrightarrow{slip(W)}$ (2,2) ...
④ Backup the reward = 1



② $Q((2,1),W) = 0.8$ ✓ $\arg\max_{a \in A} Q(s,a) = W$

$Q((2,1),E) = -0.8$

$Q((2,1),N) = 0.08$

$Q((2,1),S) = 0$

α=0.4 γ=0.9

Week 9

2. $Q(S,pass) = Q(S,pass) + \alpha[r(S,pass,M) + \gamma \max_{a' \in A(M)} Q(M,a') - Q(S,pass)]$

$= -0.7 + 0.4[-1 + 0.9 \times (-0.4) - (-0.7)]$

$= -0.7 + 0.4(-1 - 0.36 + 0.7)$

$-0.66 \times 0.4 = 0.264$

$-0.7 - 0.264 = -0.964$

3 SARSA:

$Q(S,P) = Q(S,P) + \alpha[r(S,pass,M) + \gamma Q(M,S) - Q(S,P)]$

$= -0.7 + 0.4[-1 + 0.9 \times (-0.8) + 0.7]$

$= -0.7 + 0.4(-1 - 0.72 + 0.7)$

$-1.02 \times 0.4 = -0.408$

$= -1.108$

4 ☆ $Q(S,a) = Q(S,a) + \alpha[G_t^n - Q(S,a)]$ $G_t = \sum_{i=t+1}^{t+n} \gamma^{i-t-1} r_i$

$Q(S,a) + \alpha[G_t + \gamma^n Q(S_{t+n}', a_{t+n}') - Q(S,a)]$

question ① (S,P) ② (M,S) ③ (Scored, R)

$Q(S_{suarez}, Pass) = Q(S_{suarez}, Pass) + \alpha[G_{suarez}^3 + \gamma^3 Q(M,P) - Q(S,P)]$

$= -0.7 + 0.4[G_{suarez}^3 + 0.9^3 \times (-0.4) + 0.7]$

$= -0.7 + 0.4[-1.18 - 0.2916 + 0.7]$

$= -0.7 + 0.4 \times (-0.7716) = -1.00864$

$0.81 \times 0.9 = 0.729$

$G_{suarez}^3 = r(S,P) + \gamma r(M,S) + \gamma^2 r(S_{Scored}, R)$

$= -1 + 0.9 \times (-2) + 0.9^2 \times 2 = -1 - 1.8 + 1.62 = -1.18$

$0.81 \times 2 = 1.62$

-0.18

$0.81 \times 0.4 = 0.324$ $0.729 \times 0.4 = 0.2916$

$-0.7716 \times 0.4 = 0.30864$

$-0.7 - 0.324 = -0.376$

$-0.376 - 1.18 = -1.556$

$-0.7 + 0.30864 = -0.39136$

$1.18 + 0.324 = 1.504$

$1.18 + 0.2916 = 1.4716$

$1.4716 - 0.7 = 0.7716$

$-0.7 + 1.00864 = 1.00864$

Week 9

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

$Q(Su, pass) = -0.7 + 0.4[-1 + 0.9 \times (\max Q(M,)) + 0.7]$ (max = −0.4)

$= -0.7 + 0.4[-1 - 0.36 + 0.7]$ $= -0.7 - 0.264 = -0.964$

$-0.4 \times 0.9 = -0.36$

$-0.66 \times 0.4 = 0.264$

13. $Q(S, pass) = -0.7 + 0.4[-1 + 0.9 \times (-0.8) + 0.7]$

$= -0.7 + 0.4[-1 + (-0.72) + 0.7]$ $= -0.7 - 0.408 = -1.108$

$-0.8 \times 0.9 = -0.72$

$-1.02 \times 0.4 = 0.408$

14. $Q(Su, pass)$ γ_1

$Q(Mesu, shoot)$ γ_2

$Q(Scond, retire)$ γ_3

$Q(M, pass)$ γ_4

3-step SARSA

$Q(s,a) = Q(s,a) + \alpha[G_t + \gamma^3 Q(s',a') - Q(s,a)]$

$G(t) = r_1 + \gamma r_2 + \gamma^2 r_3 = -1 + 0.9 \times (-2) + 0.9^2 \times 2$

$= -1 - 1.8 + 1.62 = -1.18$

$0.81 \times 2 = 1.62$

$1.62 - 1.8 = -0.18$

$Q(Su, pass) = -0.7 + 0.4[-1.18 + 0.9^3 \times (-0.4)) + 0.7)$

$= -1.20864$

$-0.4 \times 0.9^3 = -0.4 \times 0.729 = -0.2916$

$-1.18 - 0.2916 = -1.4716$