Question 1

UCB : $\rho_a + c\sqrt{\frac{\log N}{N_a}}$

a) $\rho_A = 0.8$ UCB: $0.8 + 1\sqrt{\frac{\log 14}{10}} = 1.314$

$\rho_B = 0.75$ UCB: $0.75 + 1\sqrt{\frac{\log 14}{4}} = 1.56$

$\boxed{\text{choose } B}$

b) $\rho_A = 0.9$ UCB: $0.9 + 1\sqrt{\frac{\log 20}{10}} = 1.44$

$\rho_B = 0.8$ UCB: $0.8 + 1\sqrt{\frac{\log 20}{10}} = 1.35$

$\boxed{\text{choose } A}$

c) A greedy policy chooses the action that maximizes ρ_a.

Greedy for part (a): A

Greedy for part (b): A

In $\boxed{\text{part (b)}}$ the greedy action was chosen by UCB because both arms had been tried the same number of times.

Question 2

a) $$T(s'|s,a) = \frac{N(s,a,s')}{N(s,a)}$$

$$\boxed{\begin{aligned} T(s'=1|s=1,a=0) &= \frac{2}{2} = 1 \\ T(s'=2|s=1,a=0) &= \frac{0}{2} = 0 \\ T(s'=1|s=2,a=0) &= 0 \\ T(s'=2|s=2,a=0) &= 0 \end{aligned}}$$

(assumed because there is no data)
" " "

b) Use $Q(s,a) \leftarrow Q(s,a) + \alpha\left(r + \gamma \max_{a'} Q(s',a') - Q(s,a)\right)$

step 1: $Q(1,0) \leftarrow \cancelto{0}{Q(1,0)} + \alpha\left(1 + \gamma \max_{a'} \cancelto{0}{Q(1,a')} - \cancelto{0}{Q(1,0)}\right)$

$Q(1,0) \leftarrow 0.1 \cdot 1 = 0.1$

step 2: $Q(1,0) \leftarrow Q(1,0) + \alpha\left(1 + \gamma \max_{a'} Q(1,a') - Q(1,0)\right)$

$\leftarrow 0.1 + 0.1\,(1 + 0.9 \cdot 0.1 - 0.1)$

$\leftarrow 0.1 + 0.1\,(1.09 - 0.1)$

$Q(1,0) \leftarrow 0.199$

step 3: $Q(1,1) \leftarrow 0 + \alpha\left(1 + \gamma \max_{a'} \cancelto{0}{Q(2,a')} - \cancelto{0}{Q(1,1)}\right)$

$Q(1,1) \leftarrow 0.1$

Final Q values:

$$\boxed{\begin{array}{c} \quad\quad a \\ Q = \begin{array}{cc} 0 & 1 \\ \end{array} \\ \begin{bmatrix} 0.199 & 0.1 \\ 0 & 0 \end{bmatrix} \begin{array}{c} 1 \\ 2 \end{array}\ s \end{array}}$$

c) $$\nabla_\theta(U) = \sum_{k=1}^{3} \nabla_\theta \log \pi_\theta(a|s)\, \gamma^{k-1} r_{to\text{-}go}$$

step 1: $\nabla_\theta \log \pi_\theta = \begin{bmatrix} \frac{\partial}{\partial\theta_1} \log(1-\theta_1) \\ \frac{\partial}{\partial\theta_2} \log(1-\theta_1) \end{bmatrix} = \begin{bmatrix} -\frac{1}{1-0.5} \\ 0 \end{bmatrix} = \begin{bmatrix} -2 \\ 0 \end{bmatrix}$

$r_{to\text{-}go} = 1 + 0.9 + 0.81 = 2.71$

step 2: $\nabla_\theta \log \pi_\theta = \begin{bmatrix} -2 \\ 0 \end{bmatrix}$ (same as step 1)

$r_{to\text{-}go} = 1 + 0.9 = 1.9$

step 3: $\nabla_\theta \log \pi_\theta = \begin{bmatrix} \frac{\partial}{\partial\theta_1} \log\theta_1 \\ \frac{\partial}{\partial\theta_2} \log\theta_1 \end{bmatrix} = \begin{bmatrix} \frac{1}{0.5} \\ 0 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$

$r_{to\text{-}go} = 1$

(continued on next page)

Question 3(c) (continued)

$$\nabla_\theta(U) = \begin{bmatrix}-2\\0\end{bmatrix}\cdot 2.71 + \begin{bmatrix}-2\\0\end{bmatrix}\cdot 0.9\cdot 1.9 + \begin{bmatrix}2\\0\end{bmatrix}\cdot 0.9 = \boxed{\begin{bmatrix}-7.04\\0\end{bmatrix}}$$

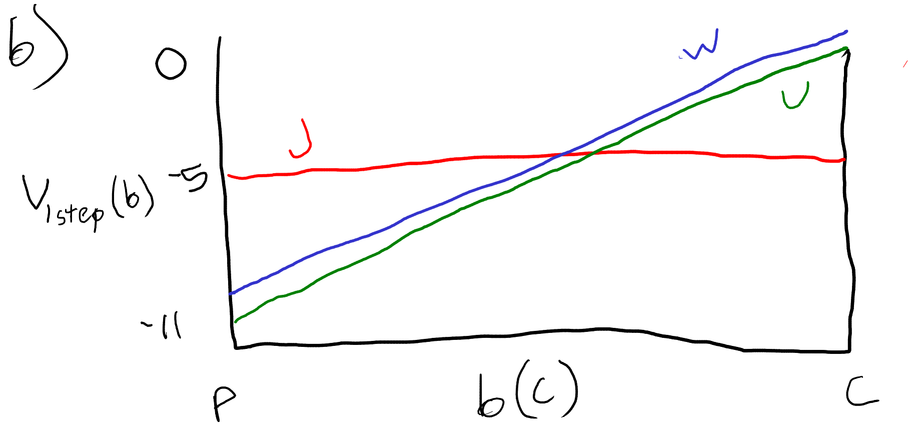
Question 4

a) one step α vectors

$$\alpha_a = \begin{bmatrix}R(P,a)\\R(C,a)\end{bmatrix}$$

$$\boxed{\alpha_U = \begin{bmatrix}-11\\-1\end{bmatrix} \quad \alpha_J = \begin{bmatrix}-5\\-5\end{bmatrix} \quad \alpha_W = \begin{bmatrix}-10\\0\end{bmatrix}}$$

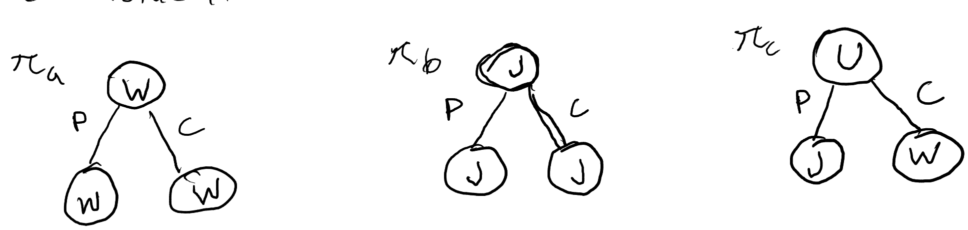
b)



c) You would never take action U because it is completely dominated by action W.

d) certainty-equivalence would avoid the U action because the best action to take in state P is J and the best action to take in state C is W. U is an info-gathering action and CE assumes that there is no state uncertainty, so U will be avoided.

e)



f) Use the conditional plan backup equation:

$$U^{\pi}(s) = R(s,\pi()) + \gamma\left[\sum_{s'} T(s' \mid s,\pi()) \sum_{o} Z(o \mid \pi(), s')\, U^{\pi(o)}(s')\right]$$

since $T(s'|s,a) \in \{1, 0\}$, calculations are simplified.

For action W, $s' = s$.

$$\alpha_a = \begin{bmatrix} U^{\pi_a}(P) \\ U^{\pi_a}(C) \end{bmatrix} = \begin{bmatrix} R(P,W) + R(P,W) \\ R(C,W) + R(C,W) \end{bmatrix} = \begin{bmatrix} -20 \\ 0 \end{bmatrix}$$

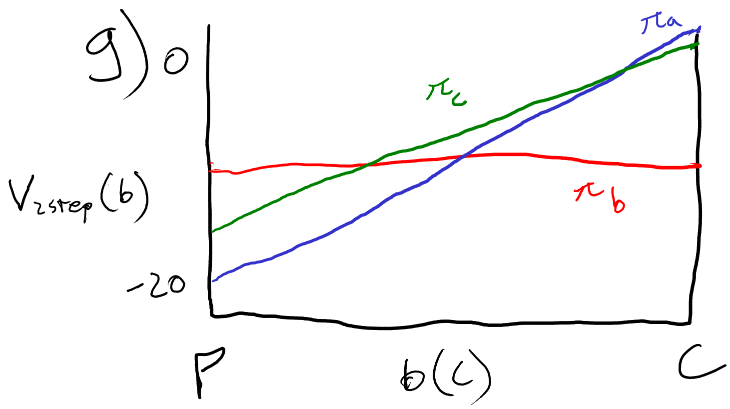
For action J, $s' = C$

$$\alpha_b = \begin{bmatrix} U^{\pi_b}(P) \\ U^{\pi_b}(C) \end{bmatrix} = \begin{bmatrix} R(P,J) + R(C,J) \\ R(C,J) + R(C,J) \end{bmatrix} = \begin{bmatrix} -10 \\ -10 \end{bmatrix}$$

For action U, $s' = s$, and $o = s$

$$\alpha_c = \begin{bmatrix} U^{\pi_c}(P) \\ U^{\pi_c}(C) \end{bmatrix} = \begin{bmatrix} R(P,U) + R(P,J) \\ R(C,U) + R(C,W) \end{bmatrix} = \begin{bmatrix} -16 \\ -1 \end{bmatrix}$$

g)



h) $b = [0.5, 0.5]$

$b \cdot \alpha_a = -10$

$b \cdot \alpha_b = -10$

$b \cdot \alpha_c = -8.5$

α_c is dominant at b, so π_c will be chosen.

$$\boxed{\pi_c() = U}$$