1. Let $V_0(x) \equiv$ maximal expected return if gambler has a total amt. of x & is allowed N more gambling moves.

Then we have the DP algorithm as

(a) $$V_k(x) = \max_{0 \le \alpha \le 1} \left[p V_{k+1}(x + \alpha x) + q V_{k+1}(x - \alpha x) \right],$$

$$k = N-1, N-2, \dots, 0$$

with $V_N(x) = \log x$.

(b) When $p > q$, since $p + q = 1$,

$$1 > 2q \Rightarrow q < \frac{1}{2}$$

$$\Rightarrow p > 1/2$$

Then,

$$V_{N-1}(x) = \max_{0 \le \alpha \le 1} \left[p\, V_N(x+\alpha x) + q\, V_N(x-\alpha x) \right]$$

$$= \max_{0 \le \alpha \le 1} \left[p \log(x+\alpha x) + q \log(x-\alpha x) \right]$$

Let $f(\alpha) = p \log(x+\alpha x) + q \log(x-\alpha x)$

$$= p \log((1+\alpha)x) + q \log((1-\alpha)x)$$

$$= p \log(1+\alpha) + q \log(1-\alpha) + (p+q) \log x$$

$$\| \quad 1$$

Then $f'(\alpha) = \frac{p}{1+\alpha} - \frac{q}{1-\alpha} = 0$

$$\Rightarrow \quad p - q - (p+q)\alpha = 0$$

$\Rightarrow \alpha = p - q$

(Note: $0 \le \alpha \le 1$)

Now $f''(\alpha) = -p(1+\alpha)^{-2} - q(1-\alpha)^{-2}$

$< 0 \quad \forall \alpha \in [0,1]$

$\Rightarrow \alpha = p - q$ is maximizer for $f(\alpha)$

Now

$V_{N-1}(x) = p \log(1 + p - q)$

$+ q \log(1 - p + q)$

$+ \log x$

$= \underbrace{p \log(2p) + q \log(2q)}_{\text{//}} + \log x$

K

$\therefore V_{N-1}(x) = \log x + K$

Now

$$V_{N-2}(x) = \max_{0 \le \alpha \le 1} \left[p V_{N-1}(x+\alpha x) + q V_{N-1}(x-\alpha x) \right]$$

$$= \max_{0 \le \alpha \le 1} \left[p \log(x+\alpha x) + q \log(x-\alpha x) \right] + k$$

Again the max above is attained at

$$\alpha = p - q.$$

And $V_{N-2}(x) = \log x + 2k$

$\vdots$

Proceeding similarly,

$$V_0(x) = \log x + Nk.$$

Optimal action: Bet $(p-q)$ fraction of

the total amount at each instant.

(c) When $p < q$, again consider

$$f(\alpha) = p \log(1+\alpha) + q \log(1-\alpha)$$

$t \log x$,

$0 \le \alpha \le 1$.

So, $f'(\alpha) = \frac{p}{1+\alpha} - \frac{q}{1-\alpha}$.

For $0 \le \alpha \le 1$,

$$\frac{p}{1+\alpha} < \frac{q}{1-\alpha}$$

and $f'(\alpha) < 0 \quad \forall \alpha \in [0,1]$

$\Rightarrow$ f is a decreasing function on $[0,1]$

$\Rightarrow$ Max is attained at $\alpha = 0$ in interval $[0, 1]$.

& $V_{N-1}(x) = p\log(1) + q\log(1) + \log x$

$= \log x$.

Similarly, it can be seen that

$V_k(x) = \log x, \quad \forall\, k = 0, 1, \ldots, N-1$

& the optimal strategy is to bet 0 at any instant.

2. Note that

$$R(i, \mu(i)) + \gamma \sum_{j \in S} P_{ij}(\mu(i)) \max\{J^{\mu_1}(j), J^{\mu_2}(j)\}$$

$$= \max_a \left(R(i,a) + \gamma \sum_{j \in S} P_{ij}(a) \max\{J^{\mu_1}(j), J^{\mu_2}(j)\} \right)$$

$$\geq \max_a \left(R(i,a) + \gamma \sum_{j \in S} P_{ij}(a) J^{\mu_1}(j) \right)$$

$$\geq R(i, \mu_1(i)) + \gamma \sum_{j \in S} P_{ij}(\mu_1(i)) J^{\mu_1}(j)$$

$$= J^{\mu_1}(i). \longrightarrow (1)$$

Similarly, we obtain

$$R(i, \mu(i)) + \gamma \sum_{j \in S} p_{ij}(\mu(i)) \max\{J^{\mu_1}(j), J^{\mu_2}(j)\}$$

$$\geq R(i, \mu_2(i)) + \gamma \sum_{j \in S} p_{ij}(\mu_2(i)) J^{\mu_2}(j)$$

$$= J^{\mu_2}(i). \longrightarrow (2)$$

From (1) and (2), it follows that

$$R(i, \mu(i)) + \gamma \sum_{j \in S} p_{ij}(\mu(i)) \max\{J^{\mu_1}(j), J^{\mu_2}(j)\}$$

$$\geq \max\{J^{\mu_1}(i), J^{\mu_2}(i)\}$$

$\forall i \in S.$

In terms of the operator T_μ, the above

is analogous to

$$T_\mu \max\{J^{\mu_1}, J^{\mu_2}\} \geq \max\{J^{\mu_1}, J^{\mu_2}\}$$

Using monotonicity of T_μ, we have

$$J^\mu = \lim_{n\to\infty} T_\mu^n \max\{J^{\mu_1}, J^{\mu_2}\}$$

$$\geq \cdots \geq T_\mu^{k+1} \max\{J^{\mu_1}, J^{\mu_2}\}$$

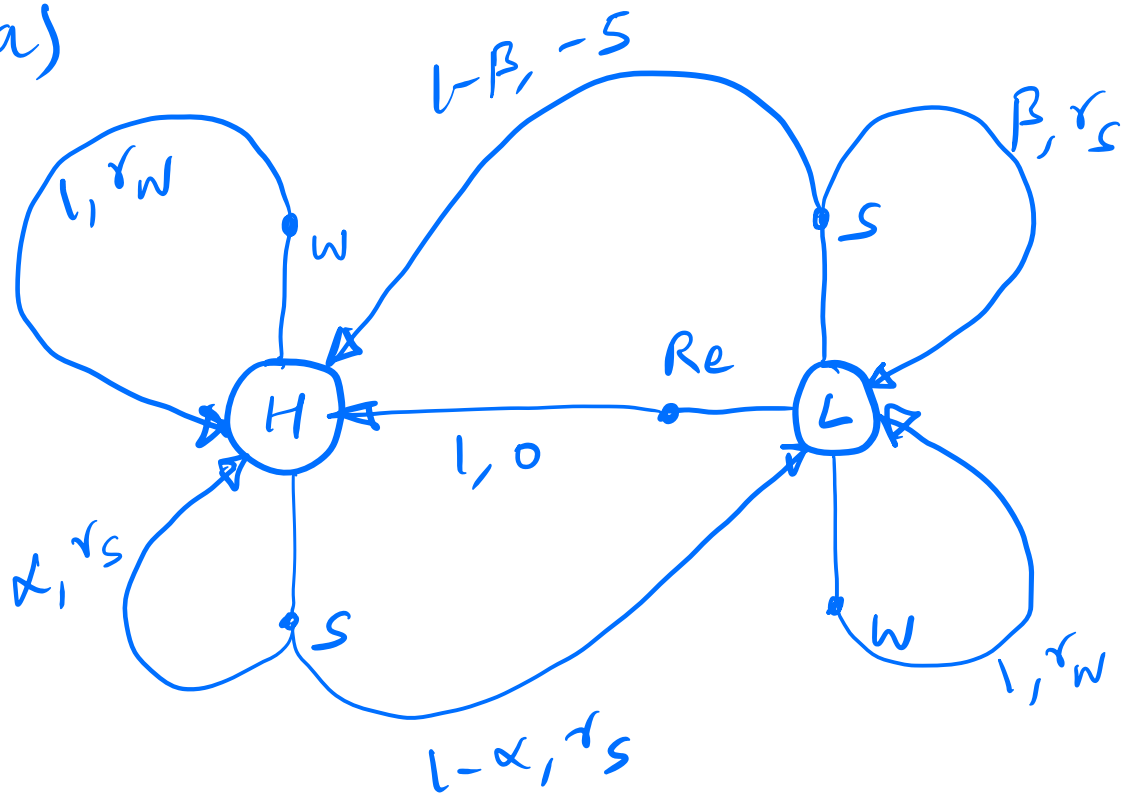
$$\geq T_\mu^k \max\{J^{\mu_1}, J^{\mu_2}\}$$

$$\geq \cdots \geq T_\mu \max\{J^{\mu_1}, J^{\mu_2}\}$$

$$\geq \max\{J^{\mu_1}, J^{\mu_2}\}$$

$$\Rightarrow J^\mu(i) \geq \max\{J^{\mu_1}(i), J^{\mu_2}(i)\}, \ \forall i \in S$$

3. (a)



State transition Diagram

(b) Sample trajectory:

final state

H, W, r_w, H, S, r_s, L, Re, 0, H, S, r_s, L, W, r_w, L, S, r_s, L, S, -5, H

- High state H occurs 3 times when an action is chosen & reward obtained and once as the final (terminal) state
- Low state L occurs 4 times

First visit Monte Carlo estimates for H & L

$\hat{V}(H) \leftarrow 1 + 2 + 0 + 2 + 1 + 2 - 5 = 3$

$\hat{V}(L) \leftarrow 0 + 2 + 1 + 2 - 5 = 0$

$\therefore$ First visit estimates

$$\hat{V}(H) = 3 \quad \& \quad \hat{V}(L) = 0$$

(c) Every visit Monte Carlo Estimates for H:

$\hat{V}_1(H) \leftarrow 1 + 2 + 0 + 2 + 1 + 2 - 5 = 3$

$\hat{V}_2(H) \leftarrow 2 + 0 + 2 + 1 + 2 - 5 = 2$

$\hat{V}_3(H) \leftarrow 2 + 1 + 2 - 5 = 0$

$\Rightarrow \hat{V}(H) = \frac{3+2+0}{3} = \frac{5}{3}$

Every visit Monte Carlo Estimates for L:

$$\hat{V}_1(L) \leftarrow 0 + 2 + 1 + 2 - 5 = 0$$

$$\hat{V}_2(L) \leftarrow 1 + 2 - 5 = -2$$

$$\hat{V}_3(L) \leftarrow 2 - 5 = -3$$

$$\hat{V}_4(L) \leftarrow -5$$

$$\Rightarrow \hat{V}(L) = \frac{0 - 2 - 3 - 5}{4} = -\frac{5}{2}.$$

$\therefore$ Every visit estimates

$$\hat{V}(H) = \frac{5}{3} \text{ and } \hat{V}(L) = -\frac{5}{2}.$$

(d) $V_0(H) = V_0(L) = 0$ (given initial values)

$\alpha = 1$ (learning rate)

$\gamma = 1$ (no discounting)

TD(0) update

$$V(s) \leftarrow V(s) + \zeta\left(r(s) + \gamma V(s') - V(s)\right)$$

steps of TD(0) update

$$V_1(H) = V_0(H) + 1.\left(r_w + \cancel{V_0(H)} - \cancel{V_0(H)}\right)$$

$$= 0 + 1 = 1$$

$$V_1(L) = V_0(L) = 0$$

$$V_2(H) = \cancel{V_1(H)} + 1.\left(r_s + V_1(L) - \cancel{V_1(H)}\right)$$

$$= 2 + 0 = 2$$

$$V_2(L) = V_1(L) = 0$$

$$V_3(L) = \cancel{V_2(L)} + 1.\left(r_{Re} + V_2(H) - \cancel{V_2(L)}\right)$$

$$= 0 + 2 = 2$$

$$V_3(H) = V_2(H) = 2$$

$$V_4(H) = V_3(H) + 1 \cdot (r_s + V_3(L) - V_3(H))$$

$$= 2 + 2 = 4$$

$$V_4(L) = V_3(L) = 2$$

$$V_5(L) = V_4(L) + 1 (r_w + V_4(L) - V_4(L))$$

$$= 2 + 1 = 3$$

$$V_5(H) = V_4(H) = 4$$

$$V_6(L) = V_5(L) + 1 (r_s + V_5(L) - V_5(L))$$

$$= 3 + 2 = 5$$

$$V_6(H) = V_5(H) = 4$$

$$V_7(L) = \cancel{V_6(L)} + 1(-5 + V_6(H) - \cancel{V_6(L)})$$

$$= -5 + 4 = -1$$

$$V_7(H) = V_6(H) = 4$$

Final values obtained from TD(0):

$$\hat{V}(H) = 4, \quad \hat{V}(L) = -1$$

(e) Application of TD(1):

$$V(H) = V(L) = 0 \quad \text{(initialization)}$$

$$V(H) := V(H) + 1\,[1 + 2 + 0 + 2 + 1 + 2 - 5 - V(H)]$$

$$= 3$$

$$V(H) := V(H) + 1(2+0+2+1+2-5 - V(H))$$

$$= 2$$

$$V(L) := V(L) + 1(0+2+1+2-5 - V(L))$$

$$= 0$$

$$V(H) := V(H) + 1(2+1+2-5 - V(H))$$

$$= 0$$

$$V(L) := V(L) + 1(1+2-5 - V(L))$$

$$= -2$$

$$V(L) := V(L) + 1(2-5 - V(L))$$

$$= -3$$

$$V(L) := V(L) + 1(-5 - V(L))$$

$$= -5$$

Av. of the value estimates for TD(1) gives

$$\hat{V}(H) = \frac{3+2+0}{3} = 5/3$$

$$\hat{V}(L) = \frac{0-2-3-5}{4} = -\frac{10}{4} = -\frac{5}{2}$$

Note: These are same as every visit Monte Carlo estimates.

4. (a) Note that $W^p : \mathbb{R}^{|S|} \to \mathbb{R}^{|S|}$ is a contraction.

$\Rightarrow W^p$ has a unique fixed pt. U^* such that $U^* = W^p U^*$. Applying W to both sides of the fixed pt. equation above gives

$$W U^* = W^{p+1} U^* = W^p W U^*.$$

$\Rightarrow$ WU^* is also a fixed pt. of W^p.

By uniqueness of fixed pt, we have

$$U^* = WU^*.$$

(b) Since W^p is a contraction,

$\exists\, \alpha \in (0,1)$ s.t.

$$\| W^p U - W^p \bar{U} \| \leq \alpha \| U - \bar{U} \|,$$

$\forall\, U, \bar{U} \in \mathbb{R}^{|S|}$.

Thus,

$$\| W^{2p} U - W^{2p} \bar{U} \| \leq \alpha \| W^p U - W^p \bar{U} \|$$

$$\leq \alpha^2 \| U - \bar{U} \|$$

$\vdots$

Proceeding in this manner, we obtain

$$\| W^{kp} U - W^{kp} \bar{U} \| \le \alpha^k \| U - \bar{U} \|, \quad k \ge 1.$$

Letting $\bar{U} = U^*$, since $W^p U^* = U^*$,

it follows that $W^{kp} U^* = U^*$ $\forall k \ge 1$.

Thus,

$$\| W^{kp} U - U^* \| \le \alpha^k \| U - U^* \|, \quad \forall k \ge 1.$$

$$\Rightarrow \lim_{k \to \infty} \| W^{kp} U - U^* \| = 0$$

Since U is arbitrary, we may replace U with $W^i U$ for $i = 0, 1, \cdots, p-1$.

Then, we have

$$\lim_{k \to \infty} \| W^{kp+i} U - U^* \| = 0,$$

$$\forall\, i = 0, 1, \ldots, p-1$$

Now any $n \in \{0, 1, 2, \ldots\}$ can be written as $n = kp + i$, for some $i \in \{0, 1, \ldots, p-1\}$, $k \geq 0$. Thus, we have that

$$\lim_{n \to \infty} W^n U = U^*. \qquad \square$$

Q5.

(a) Given a history dependent policy $\hat{\pi} \in \Pi_1$ and an initial state $i \in S$,

define a markovian policy $\pi \in \Pi_2$ such that

$$\pi = (\pi_0, \pi_1, \pi_2, \ldots)$$

where

$$\pi_t(a_t = a \mid s_t = s) = P^{\hat{\pi}}\{a_t = a \mid s_t = s, s_0 = i\}$$

Claim: $P^{\hat{\pi}}\{s_t = s, a_t = a \mid s_0 = i\} = P^{\pi}\{s_t = s, a_t = a \mid s_0 = i\}$

proof: (By induction)

Base Case: for $t = 0$,

$$LHS = P^{\hat{\pi}}\{S_0 = s, a_0 = a \mid S_0 = i\}$$

$$= P^{\hat{\pi}}\{a_0 = a \mid S_0 = i\}\, \mathbb{1}_{\{s=i\}}$$

$$= \pi_1(a_0 = a \mid S_0 = i)\mathbb{1}_{\{s=i\}} = P^{\pi}\{a_0 = a \mid S_0 = i\}\mathbb{1}_{\{s=i\}}$$

$$= P^{\pi}\{S_0 = s, a_0 = a \mid S_0 = i\}$$

Induction : Assume the relation holds $\forall\, t = 1, 2, \ldots, n-1$

Now, $P^{\hat{\pi}}\{ S_n = s \mid S_0 = i \}$

$$= \sum_{k \in S} \sum_{u \in A} P^{\hat{\pi}}\{ S_{n-1} = k,\ a_{n-1} = u \mid S_0 = i \}\ \underbrace{P(s \mid k, u)}$$

Transition prob. for the MDP.

$$= \sum_{k \in S} \sum_{u \in A} P^{\pi}\{ S_{n-1} = k, a_{n-1} = u \mid S_0 = i \}\ P(s, \mid k, u)$$

(from inductive step)

$$= P^{\pi}\{S_n = s \mid S_0 = i\}$$

Now, $P^{\hat{\pi}}\{S_n = s,\ a_n = a \mid S_0 = i\}$

$$= P^{\hat{\pi}}\{a_n = a \mid S_n = s,\ S_0 = i\}\ P^{\hat{\pi}}\{S_n = s \mid S_0 = i\}$$

$$= \pi_n(a_n = a \mid S_n = s)\ P^{\pi}\{S_n = s \mid S_0 = i\}$$

$$= P^{\pi}\{a_n = a \mid S_n = s,\ S_0 = i\}\ P^{\pi}\{S_n = s \mid S_0 = i\}$$

$$= P^{\pi}\{S_n = s,\ a_n = a \mid S_0 = i\}$$

This completes the proof for (a).

(b)

$$J_{\hat{\pi}}(i) = \mathbb{E}_{\hat{\pi}}\left(\sum_{t=0}^{T} c(s_t, a_t, s_{t+1}) \,\middle|\, s_0 = i \right)$$

$$= \sum_{t=0}^{T} \sum_{s_t, a_t, s_{t+1}} c(s_t, a_t, s_{t+1}) \; P^{\hat{\pi}}\{ s_t, a_t, s_{t+1} \mid s_0 = i \}$$

$$= \sum_{t=0}^{T} \sum_{s_t, a_t, s_{t+1}} c(s_t, a_t, s_{t+1}) \underbrace{P^{\hat{\pi}}\{ s_{t+1} \mid s_t, a_t, s_0 = i \}}_{\text{Depends only on the transition prob. of the MDP and not on the policy}} \underbrace{P^{\hat{\pi}}\{ s_t, a_t \mid s_0 = i \}}_{P^{\pi}\{ s_t, a_t \mid s_0 = i \} \text{ from part (a)}}$$

$$\mathbb{P}(s_{t+1} \mid s_t, a_t)$$

$$= \sum_{t=0}^{T} \sum_{s_t, a_t, s_{t+1}} c(s_t, a_t, s_{t+1}) \, p^{\pi}\{s_t, a_t, s_{t+1} \mid s_0 = i\}$$

$$= \mathbb{E}_{\pi}\left(\sum_{t=0}^{T} c(s_t, a_t, s_{t+1}) \mid s_0 = i \right)$$