16)

$$S = \{ \text{Fresh}, \text{Stale} \}$$

$$A = \{ \text{Query}, \text{Silent} \}$$

s	s'	a	$P[s' \mid s, a]$	r
Stale	Fresh	Query	0.8	-8
Stale	Stale	Query	0.2	-8
Fresh	Stale	Query	0.1	-4
Fresh	Fresh	Query	0.9	-4
Stale	Fresh	Silent	0	-
Stale	Stale	Silent	1	4
Fresh	Stale	Silent	0.5	4
Fresh	Fresh	Silent	0.5	4

16)) Gurmal self knowledge practice

$$U_\pi(x) = \sum_a \pi(a|x) \sum_{x', s'} P[x', s | x, a] \left(s + \gamma U_\pi(x')\right)$$

Stale	→ s	Query	→ a_1
Fresh	→ f	Silent	→ a_2

Notation

Let's say we start in stale state.

$$U_\pi(s) = \sum_a \pi(a|s) \sum_{s, x'} P[x', s | s, a] \left(s + \gamma U_\pi(x')\right)$$

x' can be 'f' or 's' depending on our action.

$\pi(a|s) = \frac{1}{2}$ always

Assuming this policy

Starting In stale

$$U_\pi(s) = \left(P[a_1 | s] \sum_{s, x'} P[x', s | s, a_1] \left(s + \gamma U_\pi(x')\right)\right) + \left(P[a_2 | s] \sum_{s, x'} P[x', s | s, a_2] \left(s + \gamma U_\pi(x')\right)\right)$$

If I choose query

$$U_\pi(s) = \frac{1}{2}\left(P[s, -8 | s, a_1]\left(-8 + \gamma U_\pi(s)\right) + P[f, -8 | s, a_1]\left(-8 + \gamma U_\pi(f)\right)\right)$$

If I choose to stay silent

$$+ \frac{1}{2}\left(P[s, 4 | s, a_2]\left(4 + \gamma U_\pi(s)\right) + P[f, -1 | s, a_2]\left(-1 + \gamma U_\pi(f)\right)\right)$$

$$u_{\pi}(s) = \frac{1}{2}\Big[(0.2(-8 + \gamma u_{\pi}(s)) + (0.8(-8 + \gamma u_{\pi}(f))\Big] + \frac{1}{2}\left(1(4 + \gamma u_{\pi}(s))\right)$$

$$u_{\pi}(s) = \frac{1}{2}\left(-1.6 + \frac{\gamma}{5} u_{\pi}(s) - 6.4 + \frac{4}{5}\gamma u_{\pi}(f)\right) + \frac{1}{2}\left(4 + \gamma u_{\pi}(s)\right)$$

$$u_{\pi}(s) = \frac{1}{2}\left(-8 + \frac{\gamma}{5} u_{\pi}(s) + \frac{4}{5}\gamma u_{\pi}(f) + 4 + \gamma u_{\pi}(s)\right)$$

$$u_{\pi}(s) = \frac{1}{2}\left(-4 + \frac{6\gamma}{5} u_{\pi}(s) + \frac{4}{5}\gamma u_{\pi}(f)\right)$$

$$u_{\pi}(s) = -2 + \frac{2}{5}\gamma u_{\pi}(f) + \frac{3\gamma}{5} u_{\pi}(s)$$

$$\boxed{u_{\pi}(s) - \frac{3}{5}\gamma u_{\pi}(s) = \frac{2}{5}\gamma u_{\pi}(f) - 2}$$

Now we start w/ $x = f$ & find $u_\pi(f)$

→ if I choose angry

$$u_\pi(f) = \frac{1}{2}\Big[P[f, -4 \mid f, a_1]\,(-4 + \gamma u_\pi(f)) + P[s, -4 \mid f, a_1]\,(-4 + \gamma u_\pi(s))\Big]$$

if I choose silent →

$$+ \frac{1}{2}\Big[P[f, 4 \mid f, a_2]\,(4 + \gamma u_\pi(f)) + P[s, 4 \mid f, a_2]\,(4 + \gamma u_\pi(s))\Big]$$

$$u_\pi(f) = \frac{1}{2}\Big(0.9\,(-4 + \gamma u_\pi(f)) + 0.1\,(-4 + \gamma u_\pi(s))\Big) + \frac{1}{2}\Big(0.5\,(4 + \gamma u_\pi(f)) + 0.5\,(4 + \gamma u_\pi(s))\Big)$$

$$u_\pi(f) = \frac{1}{2}\Big(-3.6 + \frac{9\gamma}{10} u_\pi(f) - 0.4 + \frac{\gamma u_\pi(s)}{10}\Big) + \frac{1}{2}\Big(4 + \frac{\gamma}{2} u_\pi(f) + \frac{\gamma}{2} u_\pi(s)\Big)$$

$$u_{\pi}(f) = \frac{1}{2}\left(-4 + \frac{9r}{10}u_{\pi}(f) + \frac{r}{10}u_{\pi}(s)\right)$$

$$+ \frac{1}{2}\left(4 + \frac{r}{2}u_{\pi}(f) + \frac{r}{2}u_{\pi}(s)\right)$$

$$u_{\pi}(f) = \cancel{-2+2} + \frac{9r}{20}u_{\pi}(f) + \frac{r}{20}u_{\pi}(s)$$

$$+ \frac{r}{4}u_{\pi}(f) + \frac{r}{4}u_{\pi}(s)$$

$$u_{\pi}(f) = \frac{14r}{20}u_{\pi}(f) + \frac{6r}{20}u_{\pi}(s)$$

$$\boxed{u_{\pi}(f) - \frac{14r}{20}u_{\pi}(f) = \frac{6r}{20}u_{\pi}(s)}$$

$$u_n(s) - \frac{3}{5} r\, u_n(s) = \frac{2}{5} r\, u_n(f) - 2$$

$u_n(s) \to i$

$u_n(f) \to j$

$r = 1/2$

$$u_n(f) - \frac{14r}{20} u_n(f) = \frac{6r}{20} u_n(s)$$

$$i - \frac{3}{10} i = \frac{2}{10} j - 2$$

$$\frac{7}{10} i = \frac{2}{10} j - 2$$

$$i = \frac{2}{7} j - \frac{20}{7}$$

$$j - \frac{14}{40} j = \frac{6}{40} i$$

$$\frac{6i}{40} = \frac{26}{40} j$$

$$i = \frac{13}{3} j$$

$$\frac{13}{3} j = \frac{2}{7} j - \frac{20}{7}$$

$$\frac{91j - 6j}{21} = -\frac{20}{7} \Rightarrow 85j = -60$$

$$j = \frac{-60}{85} = -\frac{12}{17}$$

$$i = \frac{13}{3}\left(-\frac{12}{17}\right) = -\frac{52}{17}$$

Ans

$u_n(s) = -52/17$

$u_n(f) = -12/17$

16)

Part 3)

$S = \{ \text{Fresh}, \text{Stale} \}$

$A = \{ \text{Query}, \text{Silent} \}$

s	s'	a	$P[s' \mid s, a]$	r
Stale	Fresh	Query	0.8	-8
Stale	Stale	Query	0.2	-8
Fresh	Stale	Query	0.1	-4
Fresh	Fresh	Query	0.9	-4
Stale	Fresh	Silent	0	-
Stale	Stale	Silent	1	4
Fresh	Stale	Silent	0.5	4
Fresh	Fresh	Silent	0.5	4

we found state values for policy π_0.

$$\pi_0(a|x) = \frac{1}{2}$$

Stale → s | Fresh → f | query → a_1 | silent → a_2

$$U_{\pi_0}(s) = -\frac{52}{17} \qquad U_{\pi_0}(f) = -\frac{12}{17}$$

now, we improve the policy :-

FOR STALE START

$$\pi_1(s) = \underset{a \in A(s)}{\operatorname{argmax}} \sum_{r, x'} P[r, x' | s, a] \left(r + \gamma U_{\pi_0}(x')\right)$$

for a_1

$$\sum_{r, x'} P[r, x' | s, a_1] \left(r + \gamma U_{\pi_0}(x')\right)$$

$$= (0.2)\left(-8 + \frac{1}{2}\left(-\frac{52}{17}\right)\right) + (0.8)\left(-8 + \frac{1}{2}\left(-\frac{12}{17}\right)\right)$$

$$= -1.9 - 6.688$$

$$= -8.585$$

for a_2

$$\sum_{r, x'} P[r, x' | s, a_2] \left(r + \gamma U_{\pi_0}(x')\right)$$

$$= 1\left(4 + \frac{1}{2}\left(-\frac{52}{17}\right)\right)$$

$$= 2.47$$

For stale start

∴ Choosing silent is the greedy policy

FOR FRESH START

$$\sum_{r, s'} P[r, s' | s, a_1] \left(r + \gamma v_{\pi_0}(s')\right)$$

for a_1

$$\sum_{r, s'} P[r, s' | s, a_1] \left(r + \gamma v_{\pi_0}(s')\right)$$

$$= (0.1)\left(-4 + \frac{1}{2}\left(-\frac{52}{17}\right)\right) + (0.9)\left(-4 + \left(\frac{-12}{34}\right)\right)$$

$$= -0.553 - 3.91$$

$$= -4.47$$

for a_2

$$\sum_{r, s'} P[r, s' | s, a_2] \left(r + \gamma v_{\pi_0}(s')\right)$$

$$= \frac{1}{2}\left(4 + \frac{1}{2}\left(-\frac{12}{17}\right)\right) + \frac{1}{2}\left(4 + \frac{1}{2}\left(-\frac{52}{17}\right)\right)$$

$$= 1.8235 + 1.2352$$

$$= 3.0587$$

Pick a_2 (Silent)

$\pi_1(x=s)$ = silent a_2

$\pi_1(x=y)$ = silent a_2

$$u_{\pi_1}(s) = 1\left(4 + \frac{1}{2} u_{\pi_1}(s)\right)$$

$$u_{\pi_1}(s) = 4 + \frac{u_{\pi_1}(s)}{2}$$

$$\boxed{u_{\pi_1}(s) = 8}$$

$$u_{\pi_1}(y) = \frac{1}{2}\left(4 + \frac{1}{2} u_{\pi_1}(s)\right) + \frac{1}{2}\left(4 + \frac{1}{2} u_{\pi_1}(y)\right)$$

$$u_{\pi_1}(y) = \frac{1}{2}(4+4) + 2 + \frac{u_{\pi_1}(y)}{4}$$

$$\frac{3}{4} u_{\pi_1}(y) = 6$$

$$\boxed{u_{\pi_1}(y) = 8}$$

Let's improve π_1 now.

$$\pi_2(s) = \underset{a \in A(s)}{\operatorname{argmax}} \left(\sum_{s', r} P[s', r \mid s, a] \left(r + \gamma v_{\pi_1}(s') \right) \right)$$

for a_1

$$\sum_{s', r} P[s', r \mid s, a_1] \left(r + \gamma v_{\pi_1}(s') \right)$$

for $s = s$

$$= (0.8)\left(-8 + \frac{8}{2}\right) + 0.2\left(-8 + \frac{8}{2}\right)$$

$$= -4$$

for $s = f$

$$0.1\left(-4 + \frac{8}{2}\right) + 0.9\left(-4 + \frac{8}{2}\right)$$

$$= 0$$

For a_2

for $s = s$

$$\sum_{s', r} P[s', r \mid s, a_1] \left(r + \gamma v_{\pi_1}(s') \right)$$

$$1\left(4 + \frac{4}{2}\right) = 6$$

for $s = f$

$$\frac{1}{2}\left(4 + \frac{4}{2}\right) + \frac{1}{2}\left(4 + \frac{4}{2}\right) = 6$$

∴ our policy is unchanged

$\pi_2(s) = a_2$

$\pi_2(f) = a_2$

$$U_{\pi_2}(s) = 1\left(4 + \frac{1}{2} U_{\pi_2}(s)\right)$$

$$U_{\pi_2}(s) = 4 + \frac{U_{\pi_2}(s)}{2}$$

$$\boxed{U_{\pi_2}(s) = 8}$$

$$U_{\pi_2}(f) = \frac{1}{2}\left(4 + \frac{1}{2} U_{\pi_2}(s)\right) + \frac{1}{2}\left(4 + \frac{1}{2} U_{\pi_2}(f)\right)$$

$$U_{\pi_2}(f) = \frac{1}{2}(4+4) + 2 + \frac{U_{\pi_2}(f)}{4}$$

$$\frac{3}{4} U_{\pi_2}(f) = 6$$

$$\boxed{U_{\pi_2}(f) = 8}$$

Always choose Silence !

Since $U_{\pi_1}(s) = U_{\pi_2}(s)$ & $U_{\pi_1}(f) = U_{\pi_2}(f)$

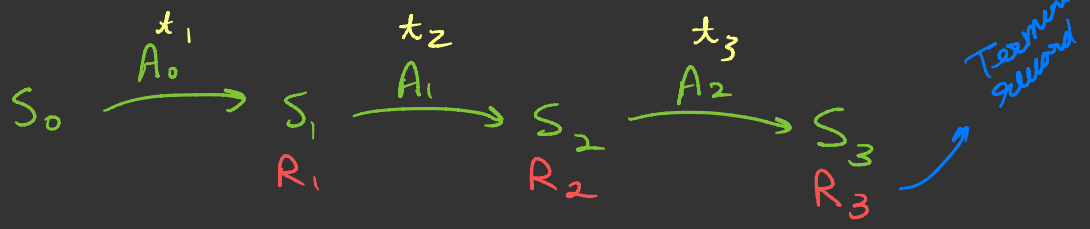
∴ We have converged : $\pi(s) = a_2$, $\pi(f) = a_2$

Part 2

16)

Question 16. A server requires information from a sensor. The server would like the information to be fresh. However, there is a cost to querying information from the sensor. Specifically, the state at the server can be either fresh or stale. The former indicates that the information at the server about the sensor is fresh and the latter indicates that it is stale. At any time, the server may choose to query or remain silent. A query makes stale information fresh with probability 0.8 and fresh information stale with probability 0.1. Staying silent keeps stale information stale with probability 1 and makes fresh information stale with probability 0.5. Also, a query when current information at the server is stale costs 8 dollars. Otherwise, it costs 4 dollars. Staying quiet has a reward of 4 dollars.

1. Draw and/or tabulate the MDP.
2. Consider a finite horizon problem in which the server gets to optimize costs over three time steps. Assume that if at the end of the three time steps the server has stale information it pays a cost of 10, else it receives a reward of 10. Also, assume that future costs are discounted with a factor 1/2. Calculate the optimal costs for every starting state. Calculate the optimal policy. Show all your steps. Guesses and intuition will bring you unbounded costs.
3. Suppose the server is to optimize for ever and is looking for a stationary policy. Write down the first four iterations of value iteration that will help the server get closer to a stationary policy. Do the same for policy iteration (where one iteration includes evaluation and improvement). Clearly identify each iteration.



$R_3 \in \{-10, 10\}$

$$E[G_t \mid S_t] = E[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots \mid S_{t=3}]$$

Bellman optimality states if we can find π_2^* then for π_1^* & π_0^*, we can use state values given by π_2^* as it will be the optimum.

If route 1 + route 4 is the optimum from place 1 to 3. Then route 4 is opt for 2-3.

At the 3rd time step, we take an action & we either get reward 10 or -10.

$$\pi_2^* = \underset{a \in A(s)}{\operatorname{argmax}} \; E[R_3 \mid X_2 = s, A_2 = a]$$

For Fresh ($x_2 = f$)

$$\pi_2^*(f) = \text{argmax} \begin{cases} 0.9(10) + 0.1(-10) & \rightarrow \text{query} \\ 0.5(10) + 0.5(-10) & \rightarrow \text{silent} \end{cases}$$

$$\boxed{\pi_2^*(f) = \text{query}}$$

For Stale ($x_2 = s$)

$$\pi_2^*(s) = \text{argmax} \begin{cases} 0.8(10) + 0.2(-10) & \rightarrow \text{query} \\ 1(10) + 1(-10) & \rightarrow \text{silent} \end{cases}$$

$$\therefore \boxed{\pi_2^*(s) = \text{query}}$$

Now, we have the optimum policy at third time step. we use this to find Expectation at t_2 & t_1.

at 2nd time step.

$\gamma = \frac{1}{2}$

$$E[R_2 + \gamma R_3 \mid X_1 = x]$$

we need to find the optimum policy.

$$\pi_1^*(x) = \underset{a \in A}{\operatorname{argmax}} \left(E[R_2 + \gamma R_3 \mid X_1 = x, A_1 = a] \right)$$

$$= \underset{a \in A}{\operatorname{argmax}} \left(E[R_2 \mid X_1 = x, A_1 = a] + \gamma E[R_3 \mid X_1 = x, A_1 = a] \right)$$

$$= \sum_{r'' \in R(s'')} r'' \sum_{s''} \sum_{a'} \sum_{s'} p[r'', s'' \mid s', a'] \pi[s', a' \mid s, a]$$

17)

$$U_{\pi_0}(s) = \sum_a \pi_0(a|s) \sum_{r,s'} (r + \gamma U_{\pi_0}(s')) P[r, s' | s, a]$$

$$\pi_1(s) = \underset{a \in A(s)}{\operatorname{argmax}} \left[\sum (r + \gamma U_{\pi_0}(s')) P[r, s' | s, a] \right]$$

↳ Since, this policy will get the action, which maximizes $U_{\pi_0}(s)$, This means that either $U_{\pi_1}(s)$ will be greater than $U_{\pi_0}(s)$ as $U_{\pi_0}(s)$ wasn't maximized w.r.t π_0 or they'll be equal if infact $\pi_0(s)$ picked the maximizing action.

18)

G, 1, 2, 3, ... n, F
(T) (T)

$i \in (1, n]$

$j \in [1, n)$

Actions: {D, U}

s	s'	a	r	P[r, s' | s, a]
i	i-1	D	2	0.5
i	i-1	D	0	0.5
1	G	D	1	1
j	j+1	U	-1	1
n	F	U	2n	1

$n = 2$

$\pi(a|s) = 0.5 \quad \forall \; a \in A(s)$

G, 1, 2, F

s	s'	a	r	$P[r, s' \mid s, a]$
1	2	U	-1	1
2	F	U	4	1
2	1	D	2	0.5
2	1	D	0	0.5
1	G	D	1	1

We have 4 states out of which two are terminal

Let's define the initial state value for each state as zero.

Since G, F are terminal states, their value will remain zero.

$V_\pi(G) = V_\pi(F) = 0 \qquad \gamma = 1$

first we find the state value for the given policy. $\gamma = 1$, Given

$$v_\pi(s) = E_\pi\left[R_{t+1} + \gamma v_\pi(s') \mid S_t = s, A_t = a\right]$$

$$v_\pi(s) = \sum_{a \in A(s)} \pi(a|s) \sum_{r, s'} \left(\gamma v_\pi(s') + r\right) p[r, s' | s, a]$$

$$v_\pi(1) = \pi(u|1)\left[\left(\gamma v_\pi(2) + (-1)\right) 1\right]$$

$$+ \pi(d|1)\left[\left(\gamma v_\pi(G) + (1)\right) 1\right]$$

$$v_\pi(1) = \frac{1}{2}\left(v_\pi(2) - 1\right) + \frac{1}{2}(0 + 1)$$

$$\boxed{v_\pi(1) = v_\pi(2)/2} \quad (1)$$

If we start in 2

$$v_\pi(2) = \frac{1}{2}\left(v_\pi(F) + 4\right) + \frac{1}{2}\left(\left(v_\pi(1) + 2\right)\frac{1}{2} + \left(v_\pi(1) + 0\right)\frac{1}{2}\right)$$

$$U_{\pi}(2) = 2 + \frac{1}{2}\left(U_{\pi}(1) + 1\right)$$

$$U_{\pi}(2) = 2.5 + \frac{U_{\pi}(1)}{2} \quad (2)$$

from (1) & (2)

$$2U_{\pi}(1) = 2.5 + \frac{U_{\pi}(1)}{2}$$

$$4U_{\pi}(1) = 5 + U_{\pi}(1)$$

$$U_{\pi}(1) = \frac{5}{3}$$

$$U_{\pi}(2) = \frac{10}{3}$$

$$U_{\pi_0}(1) = 5/3$$

$$U_{\pi_0}(2) = 10/3$$

now, we update our policy, for the first time!

$$\pi_1(s) = \underset{a \in A(s)}{\operatorname{argmax}}\ U_{\pi_0}(s)$$

$$\pi_1(s) = \underset{a \in A(s)}{\operatorname{argmax}} \sum_{r, s'} P[r, s' \mid s, a]\left(r + \gamma U_{\pi_0}(s')\right)$$

$$\pi_1(1) = \underset{a \in A(1)}{\operatorname{argmax}} \left[\underset{U}{(-1 + U_{\pi_0}(2))1}\ ,\ \underset{D}{(1+0)1}\right]$$

$$\pi_1(1) = \text{Up}$$

$$\pi_1(2) = \underset{a \in A(2)}{\operatorname{argmax}} \left[\overset{D}{(2 + U_{\pi_0}(1))\frac{1}{2}} + \overset{D}{(0 + U_{\pi_0}(1))\frac{1}{2}}, \overset{U}{(4+0)\,1} \right]$$

$$\pi_1(2) = \underset{a \in A(2)}{\operatorname{argmax}} \left[\overset{D}{\frac{8}{3}}, \overset{U}{4} \right]$$

$$\pi_1(2) = Up$$

$$\boxed{\begin{aligned} \pi_1(1) &= Up \\ \pi_1(2) &= Up \end{aligned}}$$

$$U_{\pi_1}(1) = (-1 + U_{\pi_1}(2))\,1$$

$$U_{\pi_1}(1) = U_{\pi_1}(2) - 1$$

$$U_{\pi_1}(2) = (4+0)\,1 = 4$$

$$\boxed{\begin{aligned} U_{\pi_1}(2) &= 4 \\ U_{\pi_1}(1) &= 3 \end{aligned}}$$

$$\pi_2(s) = \underset{a \in A(s)}{\operatorname{argmax}} \left[\sum_{r,s'} (r + U_{\pi_1}(s')) P[s', r | s, a] \right]$$

$$\pi_2(1) = \underset{a \in A(1)}{\operatorname{argmax}} \left[\overset{U}{(-1 + U_{\pi_1}(2))}, \overset{D}{(1 + U_{\pi_1}(3))} \right]$$

$$\pi_2(1) = \underset{a \in A(1)}{\operatorname{argmax}} \left[\overset{U}{3}, 1 \right]$$

$$\boxed{\pi_2(1) = \text{Up}}$$

$$\pi_2(2) = \underset{a \in A(2)}{\operatorname{argmax}} \left[(4 + U_{\pi_1}(F)), \frac{(2 + U_{\pi_1}(1))}{2} + \frac{(0 + U_{\pi_1}(1))}{2} \right]$$

$$\pi_2(2) = \underset{a \in A(2)}{\operatorname{argmax}} \left[\overset{U}{4}, \overset{D}{1 + 3} \right]$$

$$\boxed{\pi_2(2) = \text{Up or Down}}$$

$$\pi(U|2) = 1/2$$
$$\pi(D|2) = 1/2$$

$$U_{\pi_2}(1) = (-1 + U_{\pi_2}(2))\,1$$

$$\boxed{U_{\pi_2}(1) = U_{\pi_2}(2) - 1}$$

$$U_{\pi_2}(2) = \frac{1}{2}(4+0) + \frac{1}{2}\left[(2+U_{\pi_2}(1))\frac{1}{2} + (0+U_{\pi_2}(1))\frac{1}{2}\right]$$

$$U_{\pi_2}(2) = 2 + \frac{1}{2}\left[1 + U_{\pi_2}(1)\right]$$

$$\boxed{U_{\pi_2}(2) = \frac{5}{2} + \frac{U_{\pi_2}(1)}{2}}$$

$$\therefore U_{\pi_2}(2) = \frac{5}{2} + \frac{U_{\pi_2}(2)}{2} - \frac{1}{2}$$

$$\boxed{\begin{aligned} U_{\pi_2}(2) &= 4 \\ U_{\pi_2}(1) &= 3 \end{aligned}} \rightsquigarrow U^*(s)$$

$U_{\pi_{k+1}}(s) = U_{\pi_k}(s)\ \forall\ s$ ∴ we converged

$\pi^*(1) =$ Up $\pi^*(2) =$ {Up, Down}

If I have more states, I will have to iterate more, as there will be more paths to terminal states.

$n = 2$

$\pi(a|s) = 0.5 \quad \forall \; a \in A(s)$

G, 1, 2, F

s	s'	a	r	$P[r, s' \mid s, a]$
1	2	U	-1	1
2	F	U	4	1
2	1	D	2	0.5
2	1	D	0	0.5
1	G	D	1	1

We have 4 states out of which two are terminal

Let's define the initial state value for each state as zero.

Since G, F are terminal states, their value will remain zero.

$V_\pi(G) = V_\pi(F) = 0 \qquad \gamma = 1$