Sol^n of Chapter 3

3.1 i) Stock Trader :-
State : Current price, Chart of previous days. Amount to be invested
Action :- Buy, Sell, hold.
Reward : Total profit = Reward.

ii) Rubix Cube Solver :-
State :- Current Cube State.
Action :- All Valid Cube moves.
Reward :- +100 on correct solve

iii) QWOP game agent :
State :- Players Stance
Action :- Q, W, O, P
Reward :- Distance Covered

3.2 Nope, MDP framework doesn't represent all tasks.
ex :- Chat Bot
Next responce depend on entire chat history & not just most recent

3.3 There is no fundamental reason to prefer one location over other, its a free choice.

3.4

Page No.
Date

3.5 $\sum_{s' \in S} \sum_{r \in R} p(s', r | s, a) = 1 \quad \forall s \in S, a \in A(s)$ ... eq (3.3)

Modified eq^n

$\sum_{s' \in S^+} \sum_{r \in R} p(s', r | s, a) = 1 \quad \forall s \in S \quad a \in A(s)$

3.6 Return $= -\gamma^{T-1}$

3.7 $G_t = 1$ for every episode irrespective of Time taken for each episode
Hence Discounting should be used

3.8
$G_4 = R_5 + 0.5 G_5 = 2$
$G_3 = R_4 + 0.5 G_4 = 4$
$G_2 = R_3 + 0.5 G_3 = 8$
$G_1 = R_2 + 0.5 G_2 = 6$
$G_0 = R_1 + 0.5 G_1 = 2$

3.9
$G_0 = 2 + (0.9)7 + (0.9)^2 7 + \ldots$
$= 2 + 7 \times 0.9 [1 + 0.9 + 0.9^2 + \ldots]$
$= 2 + 7 \times 0.9 \left[\frac{1}{1 - 0.9}\right]$

$G_0 = 2 + \frac{7 \times 0.9}{0.1} = 65$

$G_1 = 7 + (0.9)7 + \ldots$
$= 7[1 + 0.9 + (0.9)^2 + \ldots]$
$G_1 = 7 \left[\frac{1}{0.1}\right] = 70$

3.10 Summation of G.P.

3.11 $E[R_{t+1} | S_t] = \sum_{a \in A} \pi(a|S_t) \cdot r(a, S_t)$.

$= \sum_{a \in A} \pi(a, S_t) \cdot \sum_{r \in R} \sum_{S' \in S} r \cdot p(S', r | S_t,$

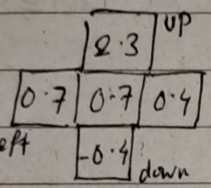
3.12 $V_\pi(S) = \sum_{a \in A} q_\pi(a, S) \cdot \pi(a|S)$

3.13 $q_\pi(S, a) = E[R_{t+1} + \gamma G_{t+1} | S_t = S, A_t = a]$

$= \sum_{S', r} p(S', r | S, a) \cdot E_\pi[R_{t+1} + \gamma G_{t+1} | S_{t+1} = S', R_{t+1} =$

$q_\pi(S, a) = \sum_{S', r} p(S', r | S, a) [r + \gamma V_\pi(S')]$

3.14



$V(up) =$

$V_\pi(S) = \sum_a \pi(a|S) \sum_{S', r} p(S', r | S, a) [r + \gamma V_\pi(S')]$

$= 0.9 [2.3 + 0.7 + 0.4 + 0 - 0.4]$

$= 0.9 \times 0.75 = 0.675 \approx 0.7.$

3.15 $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+}$

$G_{t_{new}} = G_t + \sum_{k=0}^{\infty} \gamma^k C = G_t + \frac{C}{1-\gamma}$

$\therefore E[G_{t_{new}} | S_t = S] = E\left[G_t + \frac{C}{1-\gamma} \middle| S_t = S\right]$

$\boxed{V_{\pi_{new}}(S) = V_\pi(S) + \frac{C}{1-\gamma}}$

3.25 $v_*(s) = \max_a q_*(s,a)$

3.26 $q_*(a,s) = \sum_{s',r} p(s',r|s,a)[r + \gamma v_*(s')]$

3.27 $\pi^*(s) = \arg\max_a q_*(s,a)$

3.28 $\pi^*(s) = \arg\max_a \left[\sum_{s',r} p(s',r|s,a)[r + \gamma v_*(s')]\right]$

3.29