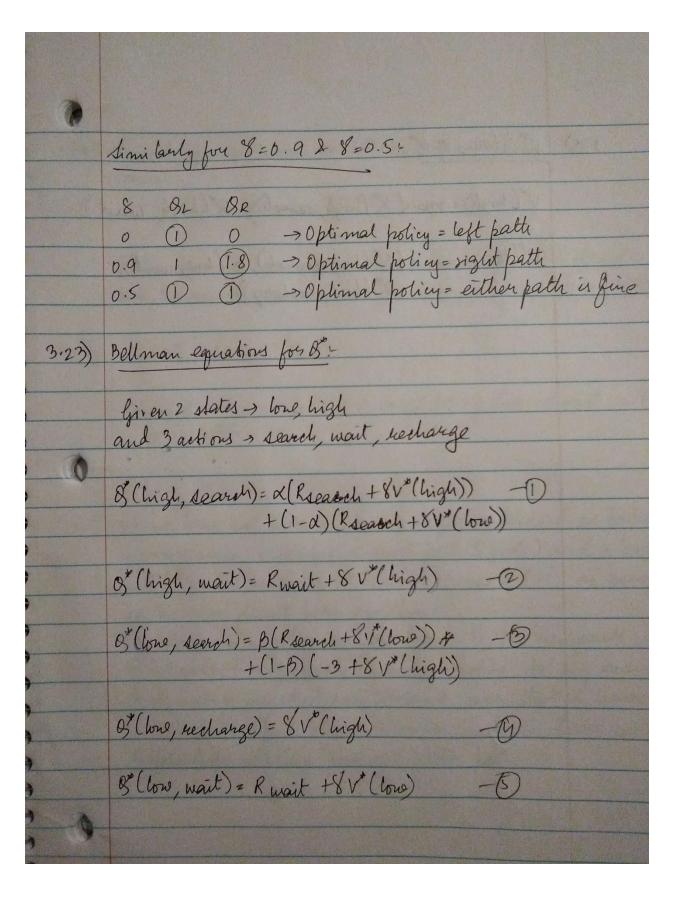
3.6) Reward for a continuous task ER) = T+1+87+2+876+3. In this case since we also in troduce discourts in the episodie task, the reward neturn will be the same as a continuous test! K=8++88++2+82++2++... Return at each sime will be 0, encept for a failure volvereit will be = -8t-1 3.7) No, we have not introduced any negative remord on talking a wrong more, meaning a short path (5 grids) and a longer path(8 grids) will give at the same neword (+1) on escape, so the agent doesn't learn amything. by introducing negative neward (say - 1), the newards will he -4 and -7 respectively in sted, according to my example. So it receives negative reinforcement, such that it leaves from the policy / path it took to escape, and can dearn and differentiate between good and bether patty.

3.14) Figure 3.2 Bedman equation: Va(s) = = 2 x(a/s) = p(s', x | s, a) [x+ y x (s')], for all ses Hore, given + $b(s', \sigma | s, a) = 1$ (transition probability) Y=0 1(a/s) = 1/4 = 0.25 (4 choices) Therefore: Va(0.7) = 0.75 ×1 (0+(0.9 × 0.23)) to.25 x 1 (0+(0.9 x 0.4)) +0.25 ×1 (0+(0.9 x-0.4)) to.75+1(0+(0.9×0.7)) 2.07 =0.15 (08207+0.36-0.36+0.63) = 0.675 ≈ 0.7

3.22) 2mm Bollman equation Va(s) = £ 7(a|s) £ p(s', 8 |s, 9) [+ 8 4 (3')] In this case; n(a(3)=0.5 (two choices) b(s', 8 | s, a) = 1 for 8=0 : Vx(3) = 0.5 x1-Va(S) = man (On, OD), for state s, L= left policy f= right policy where : $0_{1} = \frac{2}{5} |(3',8|5,a)[s+8 \frac{1}{5}(3')]$ $= \frac{1}{5} [1+0 \times 0]$ 8x= 10+0x2] =0 :.v*(3) for 8=0 = 91



v"interns of 8" > 3.25) VC (high) = max (8 (high, search), 8 (high, wait). V*(lone) = max (8" (lone, search), B*(lone, want), B" (love, necharge)).