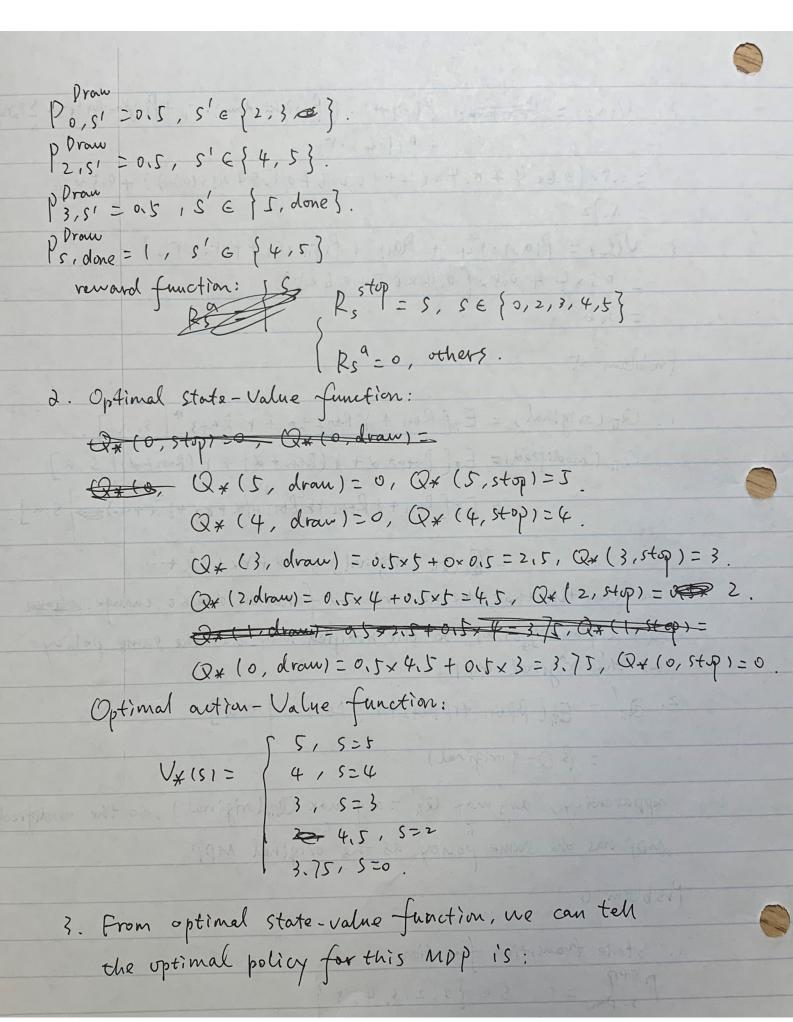
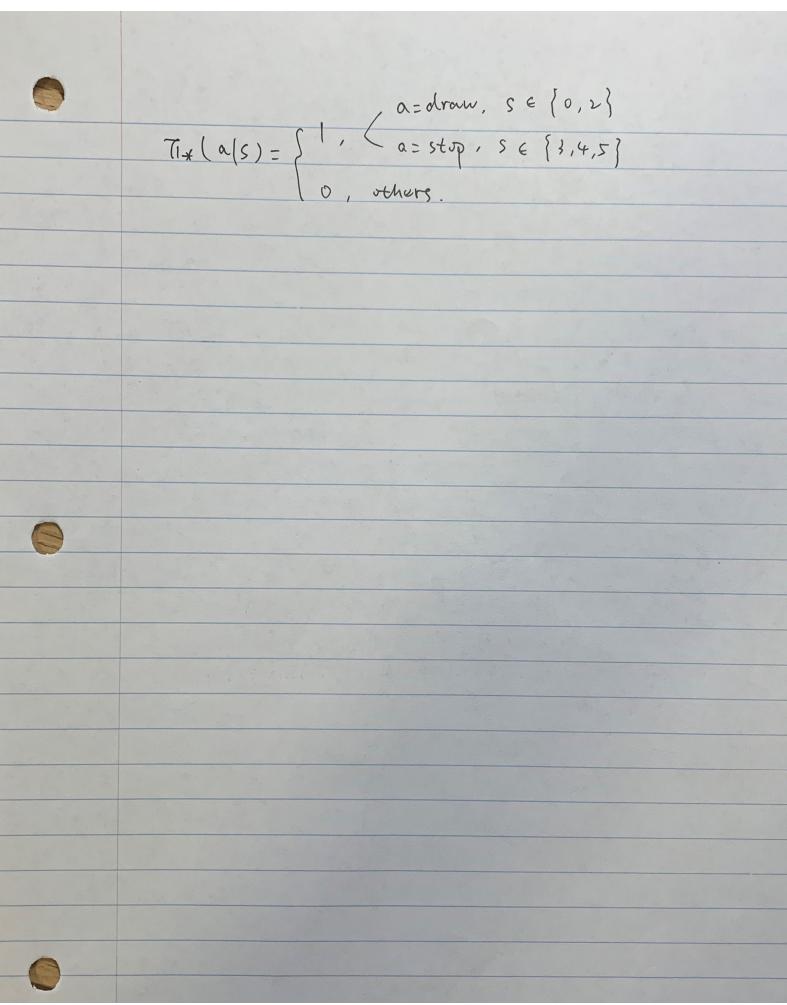
Homework 1 ELEN 6885 Reinforcement Learning t=2, 0, 0 1. t=1, 0, 0,3 t=}, D, 1 to 1 +24 ,000 of the state of the batter xameter $P_1 = \frac{0.3}{1} = 0.3, P_2 = \frac{0+1+0+0}{4} = \frac{1}{4} = 0.25.$ de PizPz, so arm, will be played if greedy method is used. 2. t=6, Polarmz) = = = =0.05. Polarmi) = 1-Plarmz) =0.95. t=7, -P(arm2)= P6(arm1) + 1 P6 (arm2) x if Darm=1 is played when t=6, fram P. >Pz, Pr(armz)=0.05.
if Darmz is played when t=6 and remard=0, P. 7Pz, Pr(armz)=0.05.
Darmz is played when t=6 and remard=1, P. <Pz, Pr(armz)=0.95. => P7 (armz) = P0 · P7 (armz) + P3 · P7"(armz) + P3 · P7"(armz) = 0.95 x 0.05 + 0.05 x 0.4 x 0.65 + 0.05 x 0.6 x 0.95 = 0.07] : t=6, the probability to play arm 2 is 0.05, t=7, the probability to play arms is 0.077. 3. Because the greedy method its stuck performing suboptimal actions. Problem 2 $\frac{Q_{+}(\alpha)}{P_{\alpha}(\tau \to 0)} = \frac{Q_{+}(\alpha)}{P_{\alpha}(\tau \to 0)} = \frac{Q_{+}(\alpha)}{Q_{+}(\alpha)} = \frac{Q_{+}(\alpha)}{Q_$ Problem 2 if an is the action with max Qtiin, Palton = 1.

if not, Pa(240)=0. => Softmax action selection becomes the same as greedy method. 2. if $\tau \to \infty$. Octob $\int_{\mathbb{R}^2} e^{\frac{Q+(i)}{\tau}} = \frac{1}{\sum_{i=1}^n e^{\frac{Q+(i)}{\tau}-Q+(i)}} = \frac{1}{\sum_{i=1}^n e^{0}} = \frac{1}{n}.$ => Softmax method yields equiprobable selection among all actions. which is the same form as to sigmoid function. Yoblem 3 m u-1 $V_{ne1} = \frac{\sum_{k=1}^{N} W_k G_k}{W_k G_k} = \frac{W_n G_n + \sum_{k=1}^{N} W_k}{W_n G_n + \sum_{k=1}^{N} W_k}$ $V_{ne1} = \frac{\sum_{k=1}^{N} W_k G_k}{W_n G_n + \sum_{k=1}^{N} W_k} = \frac{W_n G_n}{W_n G_n}$ $V_{ne1} = \frac{W_n G_n}{W_n G_n} + V_n = \frac{W_n G_n}{V_n G_n} + V_n = \frac{W_n G_n}{V_n G_n}$ $V_{ne1} = \frac{W_n G_n}{W_n G_n} + V_n = \frac{W_n G_n}{V_n G_n} + V_n = \frac{W_n G_n}{V_n G_n}$ $V_{ne1} = \frac{W_n G_n}{W_n G_n} + V_n = \frac{W_n G_n}{V_n G_n} + V_n = \frac{W_n G_n}{V_n G_n}$ $V_{ne1} = \frac{W_n G_n}{W_n G_n} + V_n = \frac{W_n G_n}{V_n G_n} + V$ = Un+ Wn Gn = Vn + Wn Gn - Vn Wn = Un (Gn - Vn)

Cn Cht1 = EWK+Wn+1 = Cn+Wn+1. Problem 4 1. V(St) = +(a1) + y (\(\sum_{a.r} \) 4+0,2x6+0-3x8+0,5x10=12,6

2. U(St) = Prantie P(right), (P(right-down), r, + Piright-up). (rz+ EPata) + P(left). r3 = 0.5x (0.6x 4 + 0.4x (4+ (0.2x6+0.3x8+0.5x10)))+0.5x4 3. U(Se) = P(az) · rz1 + P(a) · (P(r1) · r1 + P(rw) · r12) = 0.5x4 + 0.5x to,4x0,5+0.6x5) = 3.6 Problem 5 1. Q7 (original) = E7[Rt+1 + Y Pt+2 to + +2 Rt+3 S, a]. Q7, (modified) = E7[R+++ d+ & (Rten+d) + +2 (Re+3+d) | 5,a] = E7,[Rt+++Rt+2++Pt+3+(d++d++d+1-)= S,a]. = Q7 (original) + x+ rx+ rx+... we can tell that the new reward function down't change who we a = avg max 9 (s.a), so the modified MDP has the same policy as the original MDP. 2. Q7 = E7[BROOT +rfftotz+rfB Pet] ... S,a] = BQ7 (original) apparently, arg max Q7 = arg max Q7 (viginal), so the modified MOP has the same policy as the original MPP. Problem 6 1. State transition function:





This study source was downloaded by 100000850528670 from CourseHero.com on 09-18-2022 18:06:15 GMT -05:00