Reinforcement Learning Saturday, May 5, 2018 Menkov Decision fraces 15, A. R. P. O)
State action remand. I discourt faith.

action = 7/20000001 Pallay 71: TX = ZYTY Total = original E | X Yt | TU] Bolmon Equation. (2*(5,a) = Es' ~ E[Y+Ymex(2*(5),a)|S_a] Value iteration: Qiti(Soa)= E[Y+MaxQi (S,a')|S-a] Q(S,a,t) & Q*(S,a) forward approximentor $Li(\theta_i) = Ls, \alpha pc -) \left[ly_i - Q(s, \alpha) \theta_i \right]^2$. where : fi = [5 ~ r[x + ymax Q(5, a'x) -1) | s, a] berehverel. Toilibi) = Es, an pr.>:5'~ 6'[Y+Jmax Q(S,a'; 0:-1)-QS, aid) Toracs, aithi) parametrized pollules T= (The, OF RM)

for each policy: JOD = Et. p(T)() [FAT] = (tx) p(t; t) dt (tx) p(t; t) dt t = (50, 00, 50, 51) trick Toplito = P(tio) = P(tio) - Tolyp(tio) Tolo) = f (MI) Tolyp(I))) P(I) o) dr. = Etopicio, [rit) Rolog petro) I. Monte Carlo. PCTIB) = Top(St+1/St, at) To Cat/St) Thus: by P(T; 0) = Z by P(St+1 St, OL + by The (Ott St) Vo ly p(T) = } To log Tra (Cut 15e) To Jla) 2 Z r(t) To log to (and Se) baseline: Yollo) & Z Z t't -bst) To log Tolata) 6 function as buseline To Jun 2 Z (QT (St Olt) - VT (St)) To by The POH ISt) appor ximalim therom. Miverson