	CLASSTIME Page No. Date
	RL HW 3
	Vishaal V dardarao
	2016119.1
)	Ex. S. 4
X 1, 3	Assume that we touch the returns for
	Assume that we touch the roturns for each State-action pairs (5,0) by (i)
	For updating the action value function, he need to askerage out over Gis.
	hat need to alkerage out over Gis.
	Koneidoring first-vicit MC∋
	N .
	Qn+1 (s,a) = 1 (\(\frac{\xi}{2} \) \(\frac{\xi}{1} \)
10 (to 2)	
	where a is initialized arbitrarily for
	every state - action pairs.
	Now, we know =)
	$a_{n+1}(s,a) = 1 \left(\sum_{i=1}^{n-1} G_i + G_n \right)$
	$= \int_{N} \left(\frac{N-1}{N-1} \sum_{i=1}^{N-1} G_{i} + G_{i} \right)$
	$= \frac{1}{n} \left((n-1) \cdot Q_n \left(s, q \right) + G_n \right)$
	n \
6	

Ne clearly see from this modified update rule that the east of Number of tries the particular stateation pair has been encountered and the previous action value function suffice to from the updated action The updated psoudo code will be => Initialize: T(c) & A(s) (arbitrarily) for all s & S. Q(s,a) & B (arbitrarily) for all s & S, a & A(s) Returno (s,a) = 0 for all s & S, a & A(s) Loop Forever: (for each episode) Choop So & S, A & A(s) randomly st all pairs have of probability Cuevrate on episode from so, The following T: So, A= R, S, A-1, A-1, R, Q(s,a) & T(s,a) + = 1 Unloss poir s, a appears in So, Ao,, St-1, Atri Q(s,a) & Q(s,a) + Q(s,a) & Q(s,a) +	CLASSTIME Page No. Date	Communication of the state of t
update rule that the eart of rumber of this the particular stateation pair has been encountered and the previous action value function suffice to form the updated action The updated psoudo code will be =) Initialize: T(c) & A(s) (arbitarily) for all s & S. Q(s,a) & B. (arbitarily) for all s & S. Q(s,a) & B. (arbitarily) for all s & S. Returno (s,a) = 0 for all s & S. Loop Forevers: (for each episode) Choose So & S. A. & A(s) rendomly s t all pairs have 0 probability Cenerate on episode from so, Ao following T: So, Az, R., S., A., A., R., Q(s,a) & Q(s,a) + = 1 Unlus poir S, a appears in So, Ao,; St., At., Q(s,a) & Q(s,a) + (q - Q(s,a))	=) $Q_{n+1}(S,\alpha) = Q_n(S,\alpha) + L(G_n - Q_n(S,\alpha))$	
action value function suffice to from the updated action The updated psoudo code will be => Initialize: T(s) & A(s) (assistavily) for all s & S. Q(s,a) & B. (assistavily) for all s & S, 9 & A(s) Returns (s,a) = 0 for all s & S, 9 & A(s) Loop Forever: (for each episode) Choose So & S, A & A(s) randomly s t all pairs have D probability Cenerate an episode from so, No following TI: So, Ao, R, 1 ST-1, AT-1, RT. Q(S,a) & Q(S,a) + 1	update rule that the east of humber of times the particular stateaction pair	
Initialize: TI(s) E A(s) (axbitrarily) for all s ES. Q(s,a) EB (axbitrarily) for all s ES, 9 EA(s) Returns (s,a) = 0 for all s ES, a EA(s) LOOP Fosewer: (for each episode) Choose So ES, A of EA(s) randomly s t all pairs have 0 probability. Generate an episode from so, Ao following TT: So, Ap, R, 15-1, A-1, RT. GEO Loop for each step of episode, t=T-1, T-2, .0: (E (4+Rt+1) Rotums (s,a) += 1 Unloss pair S, a appears in So, Ao,; St-1, At+1 Q(s,a) EQ(s,a) (4-Q(s,a)	action value function suffice to from the updated action	
T(s) EA(s) (asbitarily) for all s ES. Q(s,a) EB (asbitarily) for all s ES, a EA(s) Raturno (s,a) = 0 for all S ES, a EA(s) Loop Forever: (for each episode) Choose So Es, A of EA(s) randomly s t all paire have of probability. Generate an episode from so, the following TI: So, As, R., ST-1, AT-1, RT. Q(S,a) += 1 Unless pair S, a appears in So, tho,; St-1, At+1 Q(S,a) & Q(S,a) + L x (9-Q(S,a))	The updated pseudo code will be =>	-
Q(s,a) & B (axbit sarry) for all SES, 9 EA(s) Returns (s,a) = 0 for all SES, a EA(s) Loop Forever: (for each episode) Choop So ES, Ag EA(s) rondomly st all paixe have of probability. Cunerate an episode from so, Ao following IT: So, As, R, ST., AT., RT., GEO Loop for each step of episode, t= T-1, T-2, 0: (E M+R++ Returns (s,a)+= Unloss pair S, a appears in So, Ao,, St-1, A++1 Q(s,a) & Q(s,a) + (9 - Q(s,a))	TICOEACS) (assistavily) for oils ES.	
Choose So ES, A & EA(Co) randomly st all paixe have of probability. Generate an episode from Co, Ao fellowing IT: So, Ao, R., ST., AT., RT. GEO Loop for each step of episode, t=T-1, T-2, -0: GE M+R++ Rotums (S, a) += 1 Unless pair S, a appears in So, Ao,, St-1, Atr) Q(S, a) E Q(S, a) + L x (4-Q(S, a))	Q(C, a) EB (asbitsarry) for all SES, 9 = HO	<u>) </u>
Concrate an episode from So, Hotellowing TT: So, Ao, R,, ST-1, AT-1, RT, QEO Loop for each step of episode, t=T-1, T-2,O: (E M+R++) Returns (s, a) += Unless pair s, a appears in So, Ao,, St-1, A++1 Q(s, a) E Q(s, a) + [X (9-Q(s, a))	Charte So ES, A CA(s) randomly st	
Loop for each step of spisode, t=T-1, T-2, -0: GE - (U+R+1) Returns (s, a) += Unloss pair s, a appears in So, Ao,; St-1, At+1 Q(s,a) E Q(s,a) + Lx (9-Q(s,a))	Generate an episode from so, Ho tolow.	¥ —
Returns $(S, \alpha) + = 1$ Unless pair S, α appears in $S_0, A_0,;$ S_{t-1}, A_{t+1} $Q(S, \alpha) \leftarrow Q(S, \alpha) + \dots$ $Q(S, \alpha) \leftarrow Q(S, \alpha)$	GEO Lon for each step of episode, t=T-1, T-2,	 <u>]:</u>
1 x (9 - Q(S,9))	Roturns (5, a) += 1	ナ デ
Returns (S,a)	Q(S,a) = Q(S,a) + 1 x (9 - Q(S,a) Returns (S,a)	2)

	CLASSTIME Page No. Date
100000	Tr(s) targ max a(s, d)
	The above pseudor code is the same as the MC is cordsol ES pseudo code as we have previously shown that both the update order are equivalent.
2)	The backup digs san for 9,7 (s=)
	(S, a) $O(S')$
	$\sqrt{(s',a')}$
	(termin) state
	The topmast node shows the initial State—action pour of a particular episode, the rest of the rodes are state nodes and state—action pains
	Black dot deput state - action pair nodes White Godes represent state nodes.

CLASSTIME	Page No	
Date /	1	· .
- P	1.	

3)	Assume the set of all time - steps in which
	The state-action pair (S,a) is visited to
	be denoted as >) T(s,a)
	and the soult of the work.
	Then, we can propresent a (s,a) as =)
	real survival survival and it
	0(ca)= 5 Prom G
	$Q(s,a) = \underbrace{Z}_{t \in T(s,a)} f_{t:T(t)-1} f_{t}$
1.4.	€ T(s,a) (E:T(t)-1
	. The west was a war as a second
	where Gt is the return at steps t to
	and (t:T(t)-1 are the 6 importance
	sampling ratios.
	TO BE THE DESCRIPTION OF THE PROPERTY OF THE P
5)	In this particular case TD updates would
	perform bother due to =)
	i) AR TD bootstrops, the original state
	0 0 00 redained
	is I be only initial source or the proper
	I and shore will be many common
	he so pottation and ald pott brem
	and out - value tunon of 11) for on since
	and hand
	La luck line line
	of piox experience while driving home
	form week.

Manager at the second distribution and the second s	CLASSTIME Page No. Date / /
	to) There state value function can
	to used in the Nows and I. then
	Jone state values of the new problem.
	Hence, There will be faster convergence
	to true state values.
	Even, in the asignal scenario if
	ne intrat value estimates are
	close to true values, then those would
/ #1 I	Le fastor convergence.
3)	If action-selection procedure is
	greedy, a leaving is not the
	Same as the sarsa algo. This is
	because sassa pichs the rext action
	from a given state next state
	to the Q-values is then based
11/2/2	on this next action.
	However, in Q-learning, fix+ the
	a trainer are updated and then
	the next action is piched according
10 10 16 9 16 9 16 16 16 16 16 16 16 16 16 16 16 16 16	to the updated d-value. This is
75	from sassa.
	side of the many and the

7 I	CLASSTIME Page No. Date / /
6)	Exercise 6.3
	We are given the following hyposparansters
	$\alpha = 0.1$ $\alpha = 1$ (undiscourted)
P	We know, update to state-value is =)
	V(St) = V(St) + & (n+++ x x(s+1) - v(s+))
	V(St) = V(St) + & (Pt+1+ YV(Str)) - V(St)) V(St)) Since all rewards are 0 except for transitions to terminal state =>
	- V(St)= (Ct)+0.1(0+V(Ct1)-V(Ct))
	Now, all V(St) initially were O.
	=> V(St)=V(St)+0.1(0.5-0.5)
San Arthur	>> V(St) = V(St)
. AS #	For tempinal states=)
2	In the first episode, when we so to
	A that is the left terminal state,
	our reward is 0

CLASSIEME	Page NA
Oute 1	1

=> V(A) = V(A) + 0:1 (0+0 - V(A))

2 (A) = V(A) - 0 1 (A) = 0.9 (A) = 0.9 (A)

Thorefore, we do conclude that only stak value of A charged in first apisode in it decreased to origin

Therefores the amount of charge was

GLONGISC 6 4

Yes, the conclusions about to the better performing also will be enfected in the performing also will be enfected in the performing a widow large of alpha values.

Suppose we take lasge values of construct of them even TD(0) built alpha, then even TD(0) built extends performing pointly share oscillations are library to hopean oscillations are library to hopean on the sty cize of being causes as the sty cize of being causes land for the convergence to be achieved.

Sinjan is the care with M cappers in the pick a widow sed of applies.

	CLASSTIME Page No. Date / /
	yes, There will be a specific alpha at which the algos might perform bother. This can be found out by doing an alpha-tuning to find out the best alpha at which print is reduced.
attacking Change	Exercise 6.5
	For high & values, those will be oscillation, in the convergence (with which can be seen as a gradient descent update sule). Therefore, for high with themse & values, when convergence is achieved, of all ations will occur.
	No, these is not a function of how the value function estimates are inflicted of all high or values, or illations will occur.
201	
por plos	