\checkmark	Vishoal Udandame
1	20/6/19 (CLASSTIME Page No.
	THEORY ANSWERS Date 1
0.1)	We see from the plots that sample
	average action-value nothod is sub-
	Optimal to the constant step size
	parameter action value mothod
	Thus, for the non stationary setting,
	Constant step 812e pasanates nother
	outperforms sample average
	action -value nothed
0.2)	The generated figure 2.3 is stown
	in the code.
	Ex, 2,6
	The oscillations and spikes in the early
	part of the curve for optimistic method
	are due to the high exploratory nature
	of the notherd. In the initial esteps
	the nothods will choose all outions
	and eine the reward received from
	all there of the seward received from
	estimates of those actions, the
	method will be "disappointed" in the
	remards it receives.
	Now, even though it is "disappointed"
	of authorizing the nother will
	by all-the actions, the nother will charge the optimal action after trying out all the actions at least
	tend to another ent me mo grant
A market	

	CLASSTIME Page No. Date
	once sice the maximum occurand?
	albeit disappointing) in from the optimal
4	action. Therefore, all the agents will
della	almost smutaneously charge the optimal
1	action and hance this causes the
	spelve in the graph
	However, over after charging the optimal
	actions, the reward it receives is
, , , , , , , , , , , , , , , , , , ,	again bound to be disappointing" and
	here the nothed does were exploration.
Jan Kay	Thus, the oscillation in the plot can
	be explained by this
	Therefore, for the vitial stope in the
	method, the optimistic greedy initialization
	method, the aptimistic greedy initialization mothod is not a good methodology since it en courages exploration and is useful
to be seen	it en courages exploration and is useful
· · · · · · · · · · · · · · · · · · ·	only for stationary problems.
0.3)	Ex. 2.7 (who with a 12
L. Dary	
	As suggested in the question, when he
	use sample averages to estimate action
	values, the action values are biased by
Contract of the Contract of th	the initial estimate ise they are
	independent of a.
	Oct 1 de la constant etco
	But when we use constant etcp size parameter nothed to estimate
795	3/20 Parane

	CLASSTIME Page No. Date / /	
	action values, these actions values are	The state of the s
And the state of t	action values, these actions values are biased by the initial estimate Q,	_
the ground is the control of grown the deposits of the control of		_
	Morefore, we need to propose a method	<u>a</u>
क लालने द्वारा र स्टब्स्ट काम्या विश्वपादानामा निर्माण कर्ता के भाग कर कर मान वर्तनात कर कर कर कर कर कर कर कर -	which can be used to modify the	_
	the action is luce entireted and	
W BATS OCCUPANTAL THE PARTY TO SEE A BUT WOOD	The action values estimated are	
A second	a la trebragabri	
	Forthis, we can use the following	
The state of the s	as our stepsize in the action value	
data dikantan manakada sinya dimatahka digunas kasanya danama	up date step 3	
	Bn=vx / sontenen	
	in all the work of all the trees	
V Marine	THE THE PARTY OF THE PROPERTY OF THE PARTY O	
	Hose, By is the step size we will use and	
	& is the constant step. cize.	
	Op is defined as =)	
	On=On+ x.(1-0n-1) (08 n7,0	
		-
	$O_0 = 0$	T. C.
The second secon	Now, by this formula, we see s	1
	01=0+x(1-100)	1
	20 = 0 + x	-
	[O, = X]	+
	S = A = A = A = A	+
	一一一一次により、	+
192	- 1980년 1981년 - 1982년 - 1982년 - 1982년 - 1982	15,33

	CLASSTIME Page No. Date / /
	norsefore our action value estimate
	Prosefose our action value estimate for step 2 is =
	Q2=QTB1(R1-Q1)
No. 1	$Q_{2} = Q_{1} + \beta_{1}(R_{1} - Q_{1})$ $= Q_{1} + (R_{1} - Q_{1})$
	=> 02=0/+R1-Q1=R1
	$\Rightarrow \boxed{0_2 = R_1}$
	$\frac{1}{2} \left[\alpha_{2} = 1 \zeta_{1} \right]$
	Therefore Q is independent of the initial
	entinate a
	Similarly, we can orderd this =
300000000000000000000000000000000000000	Por Step 3 =)
	103=02+B2(R2-02)
	wese of the second
	0.20, + 2(1-0)
	= $d + d(1-d)$
	- du(2-d) -> 0, = x(2-d)
	- rusk-yach-
	1. B2 = x = 1 = B=1
	02 d(2-d) 2-d)
4 - 117	March Many Medical Commence
All sections	$\frac{3}{3}$
	$M_{AD} A = D - A - Abar = 1$
	0 R. + 1 (R-R)
	1 2 per 2 - 2 - 2 - 2 - 2 - 2 - 2 - 2 - 2 - 2
The state of the s	

CLASSTIME Page No. 82 = Ry (1-x)+Rz This is also independent of a. · In a similar fashion, we can show that they are independent of d . The step size that we should use for senoving the dependence on intral estimate Opis B. Malysis for Stationary The stationary plot shows that optimistic again at performs better than epular gready over the long for sine it charges timal value actions greedily over the Analysis for non stationary The por atationary plat shows that epsilon greedy should perform better in greating rom stationary environment, the optimal value Changes at every step. Here optimistic

CLASSTIME	Page No.
Date	

greedy performs poorer. Malyin for stationary In stationary setting, we see that the ophnistic extinate nothod performs bold followed by UCB and then epsilon goe This is because in stationary softing, the optimal action xenous same and here optimistic estimate nother is the best since in the long run it exploits the most UCB is a dightly more explor adoxy method and here performs worse than optiniste estimate. Epsilon-greedy is not an engraly exploitation method and here it performs worst prolysis for non stationary In non stationary setting, we see that UCB performs the worlt among all three. This is because in non stationery letting the optimal action heeps changing and here the exploratory aspect of a noted should be high. But we see that UCB is highly exploitation forwaring as time passes by. Hence, UCB perform The wast is non stationary setting