	classmate
	Date Page
01	Monte Carlo ES (First visit)
	Initialize:
	Ti(s) & A(s) (arbitrarily), for all se S
	Q(s,a) E R (arbitrarily), for all s ∈ S, a ∈ A(s)
	Cs, a = 0, for all ses, a e A(s) // count of networks for each s,c
	Loop forever:
	choose So ∈ S, Ao ∈ A(So) grandomly s+ all pairs have psichobility>
	Generate episode from So, Ao following T: So, Ao, R1,, St., AT-1, RT
	C140
	Loop for each step of episode, t = T-1, T-2,, o
	9 < y9 + Re+1
	Unless the pair St, At appears in So, Ao, S1, A1,, St-1, At-1
	Cs, a += 1
	Q(St, At) + Q(St, At) + 1 (G - Q(St, At))
	C _S ,a
	T(St) ← original Q(Stra)
1 -	
	Instead of maintaining a list of netwins for each (SIA) pair
	we can simply keep totack of worent mean and total counts of
	neturns for each (s/a) pair.
	We can than simply use
4	0/2 / 42 0/2 / 1/2 0/2 / 7
TO ORE CO	$Q(s,a) \leftarrow Q(s,a) + Q(q-Q(s,a))$
	where $\alpha = 1$ since this corresponds to finding
	Cs,a new mean given the old mean,
Fred To	the step section and count of sections for pair (s,a).

Ω1

83 Q(s,a) = \(\sum_{\text{LET(S)}} \) \(\text{Per(the)} \) 83 Q(s,a) = \(\sum_{\text{LET(S)}} \) \(\text{Per(the)} \) 85 Set us assume we already have the estimate of V(s) for the time when we were in the old building. After moving to the new biding and soming back home, since highway is a common entry point for both the nowtees, we can hence use the quen of V(s) where s is the state when we are at highway to quickly estimate V(s') for the new moute change, where s comes before s. In other words, we are hootstrapping and using the guest estimate that we have for "nemaining thme" from highway to find values estimate of new states.	0.0	
Q(s,a) = \(\sum_{\text{term}} \) \(\text{ferm} \) \\ \[\text{Q(s,a)} = \(\sum_{\text{term}} \) \(\text{ferm} \) \(0.2	0
OS Q(s,a) = Secres fertities Q(s,a) = Secres fertities Steprospection Bet us assume we already have the estimate of V(s) for the time when we were in the old building. After moving to the new building and noming back home, since highway is a common entry point for both the nowice, we can hence use the quent of V(s) where s is the state when we are at highway to quickly estimate V(s') for the new nowice change, where s comes before s. In other words, we are bootstrapping and using the guest estimate that we have for nemaining time? from highway to find value estimate of new states.		
Q(s,a) = \(\sum_{\text{terms}} \) \(\text{Fetter} \) Q(s,a) = \(\sum_{\text{terms}} \) \(\text{Fetter} \) \[\text{Sterms} \) \(\text{ptotter} \) \[\text{Time when we were in the old building. After moving to the new building and noming back home. Since highway is a common entry point for both the nowtes, we can hence use the quest of V(s) where s is the state when we are at highway to quickly estimate \(\text{V}(s) \) where s is the state when we are at highway to quickly estimate \(\text{V}(s') \) \(\text{pte} \) the new shoute change, where s comes before s. In other words, we are bootstrapping and using the guest estimate that we have for "remaining time" from highway to find value estimate of mew states.		(Str Ht)
Q(s,a) = \(\sum_{\text{terms}} \) \(\text{Fetter} \) Q(s,a) = \(\sum_{\text{terms}} \) \(\text{Fetter} \) \[\text{Sterms} \) \(\text{PtotTcts-1} \) \[\text{Sterms} \) \(\text{Sterms} \) \(\text{PtotTcts-1} \) \[\text{Sterms} \) \(\text{Sterms} \) \(\text{PtotTcts-1} \) \[\text{Sterms} \) \(\text{Sterms} \) \(\text{PtotTcts-1} \) \[\text{Sterms} \) \(\text{Sterms} \) \(\text{PtotTcts-1} \) \[\text{Sterms} \) \(\text{Sterms} \) \(\text{PtotTcts-1} \) \[\text{Sterms} \) \(\text{Sterms} \) \(\text{PtotTcts-1} \) \[\text{Sterms} \) \(\text{Sterms} \) \(\text{PtotTcts-1} \) \[\text{Time when we were in the old building. After moving to the new building and using the point for both the nowtes, we can hence use the quextent of \(\text{PtotTcs} \) \(\text{PtotTcs} \		Ŏ
Q(s,a) = \(\sum_{\text{te}} \gamma(s) \) Ptu:\(T(t) - \text{Get} \) \[\text{Q(s,a)} = \(\sum_{\text{te}} \gamma(s) \) Ptu:\(T(t) - \text{Get} \) \[\text{Le T(s)} \) Ptu:\(T(t) - \text{Get} \) \[\text{Le T(s)} \] Ptu:\(T(t) - \text{Get} \) \[\text{Line T(s)} \] \[\text{Le T(s)} \] Ptu:\(T(t) - \text{Get} \) \[\text{Line T(s)} \] \[Line T(
Q(s,a) = \(\sum_{terres} \) \(\text{Fetter} \) \\ \[\text{Q(s,a)} = \sum_{terres} \) \(\text{Fetter} \) \\ \[\text{Ete T(s)} \) \(\text{Fetter} \) \\ \text{Ete T(s)} \) \\ \text{Fetter} \) \\ \text{Ete T(s)} \) \(\text{Fetter} \) \\ \text{Ete T(s)} \) \\\ \text{Ete T(s)} \) \\ \text{Ete T(s)} \) \\ \text{Ete T(s)} \)		· · ·
Q(s,a) = \(\sum_{terres} \) \(\text{Fetter} \) \\ \[\text{Q(s,a)} = \sum_{terres} \) \(\text{Fetter} \) \\ \[\text{Ete T(s)} \) \(\text{Fetter} \) \\ \text{Ete T(s)} \) \\ \text{Fetter} \) \\ \text{Ete T(s)} \) \(\text{Fetter} \) \\ \text{Ete T(s)} \) \\\ \text{Ete T(s)} \) \\ \text{Ete T(s)} \) \\ \text{Ete T(s)} \)		nav mi - of
Q(s,a) = \(\sum_{terres} \) \(\text{fert(t)} - \text{Get} \) \[\text{Q(s,a)} = \sum_{terres} \) \(\text{fert(t)} - \text{Get} \) \[\text{Left(s)} \) \(\text{PtwT(t)} - \text{Get} \) \[\text{Lime when we assume we already have the estimate of \(\text{V(s)} \) \\ \\ \text{ptw} \text{Left have} \\ \text{Lime have} \\	- 1	the state of the s
Q(s,a) = \(\sum_{\text{ET(S)}} \) \(\text{Fitter} \) \[\text{Q(s,a)} = \(\sum_{\text{ET(S)}} \) \(\text{Ft.T(t)-1} \) \(\text{Gt+1} \) \[\text{Ste} \text{T(s)} \) \(\text{Ft.T(t)-1} \) \(\text{Gt+1} \) \[\text{Ste} \text{T(s)} \) \(\text{Ft.T(t)-1} \) \(\text{Gt+1} \) \[\text{Ste} \text{T(s)} \) \(\text{Ft.T(t)-1} \) \(\text{Gt+1} \) \[\text{Ste} \text{T(s)} \) \(\text{Ft.T(t)-1} \) \(\text{Gt+1} \) \[\text{Time the men assume we already have the estimate of \(\text{V(s)} \) \\ \) \(for the men of the moving for the men of		THE CHARLES A RECT OF THE PROPERTY OF
Q(s,a) = \(\sum_{\text{ET(S)}} \) \(\text{Fitter} \) \[\text{Q(s,a)} = \(\sum_{\text{ET(S)}} \) \(\text{Ft.T(t)-1} \) \(\text{Gt+1} \) \[\text{Ste} \text{T(s)} \) \(\text{Ft.T(t)-1} \) \(\text{Gt+1} \) \[\text{Ste} \text{T(s)} \) \(\text{Ft.T(t)-1} \) \(\text{Gt+1} \) \[\text{Ste} \text{T(s)} \) \(\text{Ft.T(t)-1} \) \(\text{Gt+1} \) \[\text{Ste} \text{T(s)} \) \(\text{Ft.T(t)-1} \) \(\text{Gt+1} \) \[\text{Time the men assume we already have the estimate of \(\text{V(s)} \) \\ \) \(for the men of the moving for the men of		0.30
Q(s,a) = \(\sum_{eff}(s) \) \(\frac{\text{Ext(s)}}{\text{PtoT(t)-1}} \) \(\text{Ctot-1} \) \[\text{Ste J(s)} \) \(\text{PtoT(t)-1} \) \(\text{Ctot-1} \) \[\text{Ste J(s)} \) \(\text{PtoT(t)-1} \) \[Line when we were in the old building. After moving to the new building and roming back home, since highway is a common entry point for both the routes, we can hence use the query of \(V(s) \) where s is the state when we are at highway to quickly estimate \(V(s') \) for the new groute change, where s comes before s. In other words, we are bootstrapping and using the guery estimate that we have for "nemaining time" from highway to find values that we have for "nemaining time" from highway to find values that the property of the new states.	02	Oca in the state of the state o
Q(s,a) = \(\text{\text{E}}(s) \) \(\text{\text{Fth.T(t)-1}} \) \(\text{Gth.T(t)-1} \) \(Gth.	0.5	$\frac{-\sqrt{(s,\alpha)}}{-\sqrt{(s,\alpha)}} = \frac{1}{\sqrt{(s)}} + \frac{1}{\sqrt{(s)}} = \frac{1}{\sqrt{(s)}}$
DE Jet us assume we already have the estimate of V(s) for the time when we were in the old building. After moving to the new building and wring back home, since highway is a common entry point for both the noutes, we can hence use the guest of V(s) where s is the state when we are at highway to quickly estimate V(s') for the new noute change, where s comes before s. In other words, we are bootstrapping and using the guest estimate that we have for "nemaining time" from highway to find values estimate of new states.		
DE Set us assume we already have the estimate of V(s) for the time when we were in the old building. After moving to the new building and roming back home, since highway is a common entry point for both the rowtes, we can hence use the guest of V(s) where s is the state when we are at highway to quickly estimate V(s') for the new groute change, where s comes before s. In other words, we are bootstrapping and using the guest estimate that we have for "nemaining time" from highway to find values estimate of new states.		2007(s) [t+1: 1(t)-1 46+1
def us assume we already have the estimate of V(s) for the time when we were in the old building. After moving to the new building and roming back home, since highway is a common entry point for both the routes, we can hence use the quest of V(s) where s is the state when we are at highway to quickly estimate V(s') for the new moute change, where s comes before s. In other words, we are bootsterapping and using the guest estimate that we have for "nemaining time" from highway to find values estimate of new states.		Zte J(s) PtatCt1-1
def us assume we already have the estimate of V(s) for the time when we were in the old building. After moving to the new building and roming back home, since highway is a common entry point for both the routes, we can hence use the quest of V(s) where s is the state when we are at highway to quickly estimate V(s') for the new moute change, where s comes before s. In other words, we are bootsterapping and using the guest estimate that we have for "nemaining time" from highway to find values estimate of new states.		
new briding and soming back home, since highway is a common entry point for both the noutes, we can hence use the guest of V(s) where s is the state when we are at highway to quickly estimate V(s') for the new noute change, where s comes before s. In other words, we are bootstrapping and using the guest estimate that we have for "nemaining time" from highway to find values time of new states.	05	Let us are a discourse and the
common entry point for both the nowtes, we can hence use the guest of V(s) where s is the state when we are at highway to quickly estimate V(s') for the new nowte change, where s comes before s. In other words, we are bootstrapping and using the guest estimate that we have for "nemaining time" from highway to find values that we have for states.		the assume we acceedly have the estimate of V(s) for the
common entry point for both the moutes, we can hence use the guess of V(s) where s is the state when we are at highway to quickly estimate V(s') for the new moute change, where s comes before s. In other words, we are bootstrapping and using the guess estimate that we have for "nemaining time" from highway to find values that of new states.	t.	me were in the old building. After moving to the
to quickly estimate V(s') for the new monte change, where s comes before s. In other words, we are bootstrapping and using the guest estimate that we have for "nemaining time" from highway to find value estimate of new states.		state of the some pack home cine !!!
Comes before s. In other words, we are bootsterapping and using the guess estimate that we have for "nemaining time" from highway to find value estimate of new states.		arem of West of some both the soutes, we can hence use the
In other words, we are bootstrapping and using the guess estimate that we have for "nemaining time" from highway to find value estimate of new states.		to anickly activate 1/2/2 I li
In other words, we are bootstrapping and using the guess estimate that we have for "nemaining time" from highway to find value estimate of new states.		comes before &
estimate of new states.		30/10/2 3.
estimate of new states.		that me have be "remained and using the guest estimate
La ser restore got est		The transfer of the transfer to 1.
		STITUTE OF HEW STATES.
The state of the s	AND ASSESSMENT OF THE PARTY OF	

	This is because A is the only state in this coince of
	This is because A is the only state in this episode, herwould He was a will decrease. In terminal
	its value will decrease. In this episode, herwouly
	H changed his
	H charged by 0.05 and became 0.45 (since $\alpha = 0.1$)
	6-4 Ac 8000 7000 0
	= basis of the episodes (here the release) at the
	basis of the episodes (hence the returns) observed at each episode,
	a large value of & will bias the V() estimates towards the
V.	loss will never converge and keep assilts of any the
	loss will never converge and keep ascilla changing in response to
	A low value of alpha means that the loss or and
	a lot of time to converge. The best thing is to vary alpha on the
	bails of time Steps (high to low of as time increases)
	(noceses)
	65 At 12
	6.5 At large alphas, the value estimates are always biosed towards
	the latest episodes. Hence the loss may go up the as
	the no. of episodes are Inviersed.
	Heno the parameters need to carefully celected.
	This can also be a function of how the value functions were
	initialized first.
	6.12 SARSA S (gredy) S' (gredy) Empdate &}
	begin with (S', A')
	Q-leany SAIR S' supplate Q?
Mail	(gredy) begin with (s')
	In SARSA, we have already picked the new S' and A'
34	before updating a Ewith e=0, this is pro exploitation)
The state of the s	In O-leans, we have update of first and then selected
100	S'. This was , a new A' selected for S'might be different from
5	A' of SARSA. Hence there is a slight difference.
100,00	









