Temporary Difference: Q(s,a)

Sunday, May 8, 2022 4:05 PM

Temporary Difference Method

b) For $Q(s,a)$: $Q(s,a) \leftarrow Q(s,a) + \alpha \Delta Q(s,a)$

where $\Delta Q(s,a) = \left[f_R(s,a,s_f) + \gamma \max_{a_f} [Q(s_f,a_f)] \right] - Q(s,a)$

the state from which we move and the action we chose

$\max$ (Q cols: Q in s_f with all actions available at that s_f)

reward after transition

initial state s_1

	sF1	s1	a2=→	s2	s3	sF2	
		a1=←					
fR(s)=	-10	0		-0.04	-0.04	10	→ $f_R(s, a, s_f)$
	sf	s = ←					
Q(s,a1=←)=	0.2	0.6		0.14	0.27	0.34	a1=←
Q(s,a2=→)=	0.67	0.15		0.8	0.535	0.89	a2=→
		$Q(s_f,a_2)$	$Q(s_f,a_1)$				
γ=	0.9						
α=	0.8						

mtx Q: s ↓

Start with random $\in [0,1]$

↳ s=← since $Q(s_1,a)$ tells us what action to take, and we apply max(0.6, 0.15) = 0.6 which will also be the cell to modify

cell = 0.6 + α(f·10 + γ max(0.2, 0.67)) − 0.6)

= −8.997 instead of 0.6 now

Next

	sF1	s1	a2=→	s2	s3	sF2	
		a1=←					
fR(s)=	-10	0		-0.04	-0.04	10	✓ modified after some iterations
	s = →	s_f					new iteration!
Q(s,a1=←)=	0.2	−8.997 ✓		0.14	0.27	0.34	a1=←
Q(s,a2=→)=	0.67 ✓	0.15 ✓		0.8 ✓	0.535 ✓	0.89 ✓	a2=→

MAX →

chooses the action (→) and the one we update

this is updated to 8.7478 since it is near to 10: the agent needs to experience the goal in order to back propagate and guide the politic

¿when do we start again in s(0)?

cell = 0.67 + α((0 + γ MAX(8.997, 0.15)) − 0.67)

= 0.292 instead of 0.67

and s = s_1 again with $s_f = s_2$, r = −0.04

and a will be max(−8.997, 0.15) = →

and 0.15 will be updated

α: tells us at which speed we achieve the learning

```
In [1]: runfile('C:/Users/José A
wdir='C:/Users/José Abdón/Downlo
i= 321  r_prom= 0.0 -> 0.2
i= 334  r_prom= 0.2 -> 0.25
i= 432  r_prom= 0.25 -> 0.3
i= 442  r_prom= 0.3 -> 0.35
i= 1785  r_prom= 0.35 -> 0.45
i= 2385  r_prom= 0.45 -> 0.5
i= 3066  r_prom= 0.5 -> 0.55
```

i = 321 transitions it took to r_prom to another 0.00

at the end it will tell us in how many transitions it solved Q to a good approximation to the optimal politic

→ The transition model is being known as Temporary Diff.

Q is being solved: that is why we say P_{MT} is not needed, since it gets discovered on the way.

At the end we can see the learned Q

s ↓ a →

	0	1	2	3
0	0.0376883	0.0352659	0.0351618	0.0337093
1	0.0195259	0.0200393	0.0202732	0.0413555
2	0.0568662	0.0472828	0.073121	0.0295862
3	0.0292544	0.0219538	0.00795138	0.0330317
4	0.0464733	0.0304132	0.0280761	0.0241096
5	0	0	0	0
6	0.0655563	0.0758191	0.124899	0.0258071
7	0	0	0	0
8	0.0286933	0.0501687	0.0459952	0.0507164
9	0.147413	0.149075	0.1327	0.0364522
10	0.216103	0.269461	0.124771	0.0342354
11	0	0	0	0
12	0	0	0	0

s ↓

	0	1	2	3
0	0.0376883	0.0352659	0.0351618	0.0337093
1	0.0195259	0.0200393	0.0202732	0.0413555
2	0.0568662	0.0472828	0.073121	0.0295862
3	0.0292544	0.0219538	0.00795138	0.0330317
4	0.0464733	0.0304132	0.0280761	0.0241096
5	0	0	0	0
6	0.0655563	0.0758191	0.124899	0.0258071
7	0	0	0	0
8	0.0286933	0.0501687	0.0459952	0.0507164
9	0.147413	0.149075	0.1327	0.0364522
10	0.216103	0.269461	0.124771	0.0342354
11	0	0	0	0
12	0	0	0	0
13	0.0352794	0.0225662	0.234773	0.17616
14	0.0248262	0.553156	0.14272	0.197373
15	0	0	0	0