1. *Convert the equation of n-step off-policy TD (7.9) to semi-gradient form. Give accompanying definitions of the return for both the episodic and continuing cases.*

$$w_{t+n} = w_{t+n-1} + \alpha \rho_{t:t+n-1}(G_{t:t+n} - \hat{v}(S_t, w_{t+n-1}))\nabla \hat{v}(S_t, w_{t+n-1})$$

In the episodic case

$$G_{t:t+n} = R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{t+n-1} R_{t+n} + \hat{v}(S_{t+n}, w_{t+n-1}).$$

In the continuing case

$$G_{t:t+n} = R_{t+1} - \bar{R}_{t+n-1} + R_{t+2} - \bar{R}_{t+n-1} + \cdots + R_{t+n} - \bar{R}_{t+n-1} + \hat{v}(S_{t+n}, w_{t+n-1}).$$

2. *Convert the equations of n-step Q(σ) (7.11 and 7.17) to semi-gradient form. Give definitions that cover both the episodic and continuing cases.*

$$w_{t+n} = w_{t+n-1} + \alpha(G_{t:t+n} - \hat{q}(S_t, A_t, w_{t+n-1}))\nabla \hat{q}(S_t, A_t, w_{t+n-1})$$

In the episodic case

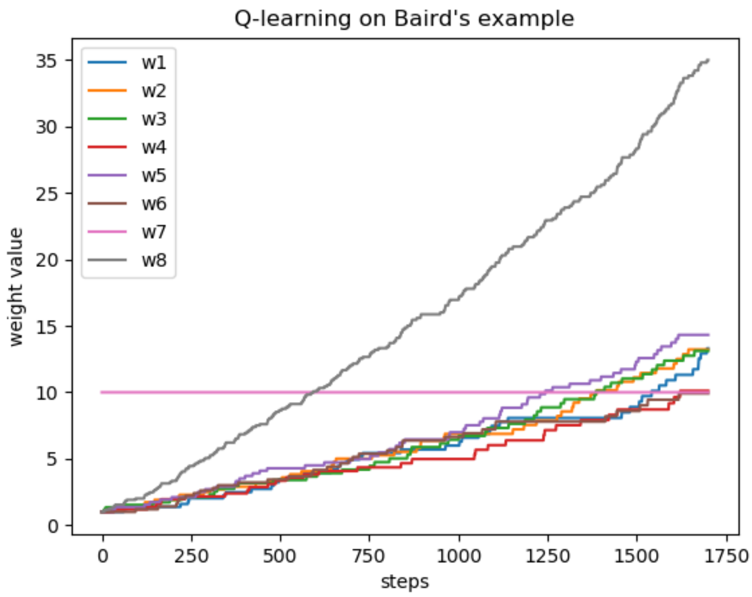
$$\begin{aligned} G_{t:h} = R_{t+1} &+ \gamma\Big(\sigma_{t+1}\rho_{t+1} + (1-\sigma)\pi(A_{t+1}|S_{t+1})\Big)\Big(G_{t+1:h} - \hat{q}(S_{t+1}, A_{t+1}, w_{h-1})\Big) \\ &+ \gamma \sum_a \pi(a|S_{t+1})\hat{q}(S_{t+1}, a, w_{h-1}) \end{aligned}$$

In the continuing case

$$\begin{aligned} G_{t:h} = R_{t+1} - \bar{R}_{h-1} &+ \gamma\Big(\sigma_{t+1}\rho_{t+1} + (1-\sigma)\pi(A_{t+1}|S_{t+1})\Big) \\ &\cdot \Big(G_{t+1:h} - \hat{q}(S_{t+1}, A_{t+1}, w_{h-1})\Big) + \gamma \sum_a \pi(a|S_{t+1})\hat{q}(S_{t+1}, a, w_{h-1}) \end{aligned}$$

3. *Apply one-step semi-gradient Q-learning to Baird's counterexample and show empirically that its weights diverge.*

   The code can be found at https://github.com/hannagabor/SBRL/blob/master/11.3/q_learning_bairds.py.



Q-learning on Baird's example
w1
w2
w3
w4
w5
w6
w7
w8
weight value
steps
0
5
10
15
20
25
30
35
0
250
500
750
1000
1250
1500
1750