1 *Off-policy Methods with Approximation

1.1 Exercise 11.1

Q

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.

A

Tabular case is

$$V_{t+n}(S_t) = V_{t+n-1} + \alpha\rho_{t:t+n-1}[G_{t:t+n} - V_{t+n-1}(S_t)].$$

The semi-gradient weight update is

$$\mathbf{w}_{t+n} = \mathbf{w}_{t+n-1} + \alpha\rho_{t:t+n-1}[G_{t:t+n} - \hat{v}(S_t, \mathbf{w}_{t+n-1})]\nabla_{\mathbf{w}}\hat{v}(S_t, \mathbf{w}_{t+n-1}),$$

noting the occurrence of the $n$step TD Error

$$\delta_t^n = G_{t:t+n} - \hat{v}(S_t, \mathbf{w}_{t+n-1}).$$

We define the returns in the two cases

episodic $G_{t:t+n} = \sum_{i=t}^{t+n-1} \gamma_{i-t}R_{i+1} + \gamma^n\hat{v}(S_{t+n}, \mathbf{w}_{t+n-1})$

continuing $G_{t:t+n} = \sum_{i=t}^{t+n-1}(R_{i+1} - \bar{R}_i) + \hat{v}(S_{t+n}, \mathbf{w}_{t+n-1})$

where in each case $G_{t:h} = G_t$ if $h \geq T$.

1.2 *Exercise 11.2

Q

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.

A

The update is

$$\mathbf{w}_{t+n} = \mathbf{w}_{t+n-1} + \alpha[G_{t:t+n} - \hat{q}(S_t, A_t, \mathbf{w}_{t+n-1})]\nabla_{\mathbf{w}}\hat{q}(S_t, A_t, \mathbf{w}_{t+n-1}$$

with the following definitions of returns targets

Episodic

$$G_{t:h} = R_{t+1} + \gamma\left[\sigma_{t+1}\rho_{t+1} + (1 - \sigma_{t+1})\pi(A_{t+1}|S_{t+1})\right]\left[G_{t:h} - \hat{q}(S_t, A_t, \mathbf{w}_{h-1})\right] + \gamma\bar{V}_{h-1}(S_{t+1})$$

Continuing

$$G_{t:h} = R_{t+1} - \bar{R}_t + \left[\sigma_{t+1}\rho_{t+1} + (1 - \sigma_{t+1})\pi(A_{t+1}|S_{t+1})\right]\left[G_{t:h} - \hat{q}(S_t, A_t, \mathbf{w}_{h-1})\right] + \bar{V}_{h-1}(S_{t+1})$$

where

$$\bar{V}_i(s) = \sum_a \pi(a|s)\hat{q}(s, \mathbf{w}_i)$$

and $G_{h:h} = \hat{q}_(S_h, A_h, \mathbf{w}_{h-1})$ if $h < T$ while if $h = T$ we have $G_{T-1:T} = R_T$ in the episodic case and $G_{T-1:T} = R_T - \bar{R}_{T-1}$ in the continuing case.

Note that in each case the value functions are defined with respect to the relevant episodic discounted or continuing average excess return.

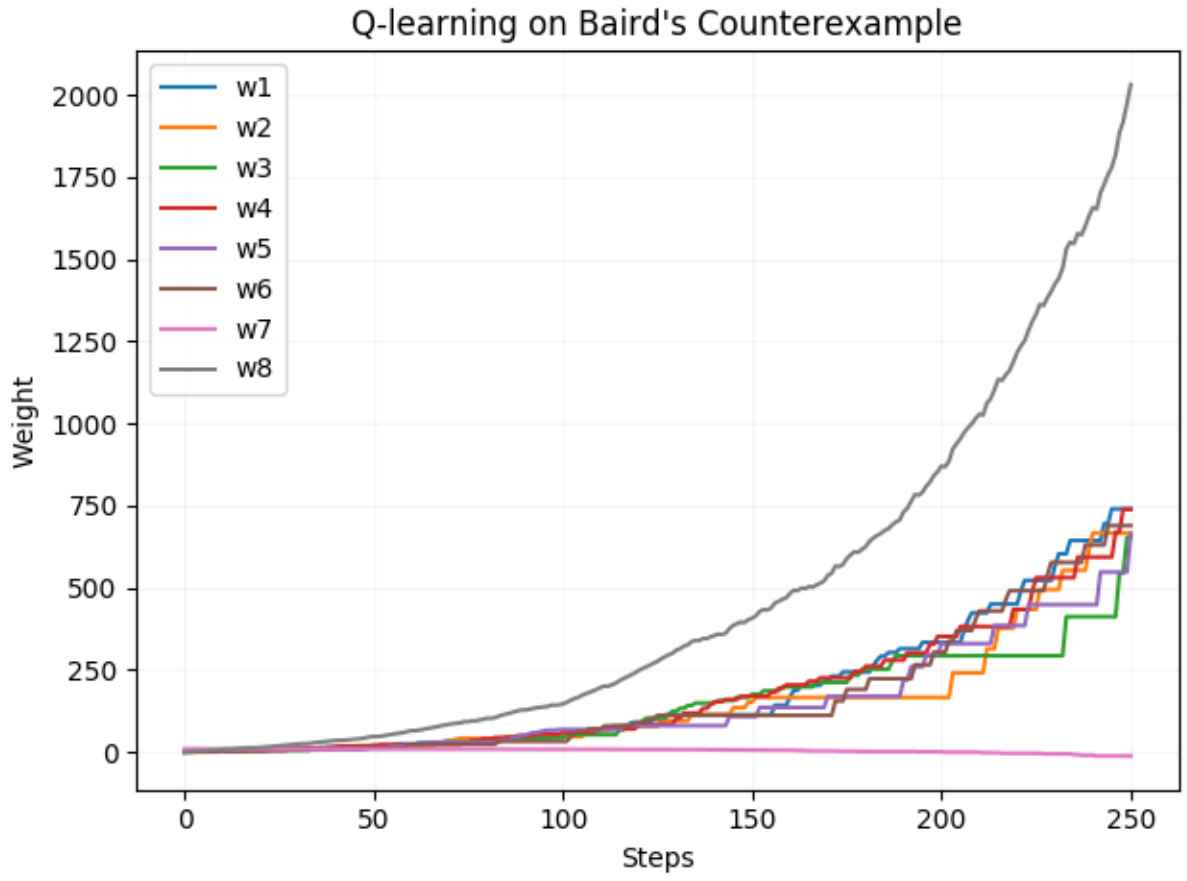
1.3 Exercise 11.3 (programming)

Q

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

A

This is a programming exercise. For the relevant code please see the repo.



1.4 Exercise 11.4

Q

Prove (11.24). Hint: Write the $\bar{\text{RE}}$ as an expectation over possible states s of the expectation of the squared error given that $S_t = s$. Then add and subtract the true value of state s from the error (before squaring), grouping the subtracted true value with the return and the added true value with

the estimated value. Then, if you expand the square, the most complex term will end up being zero, leaving you with (11.24).

A

Define

$$\overline{\mathrm{VE}}(\mathbf{w}) = \mathbb{E}_{s\sim\mu}[v_\pi(s) - \hat{v}(s,\mathbf{w})]$$

Now have the return error

$$\begin{aligned}\overline{\mathrm{RE}} &\doteq \mathbb{E}\left[(G_t - \hat{v}(S_t,\mathbf{w}))^2\right] &(1)\\ &= \overline{\mathrm{VE}}(\mathbf{w}) + \mathbb{E}\left[(G_t - v_\pi(S_t))^2\right] + 2\mathbb{E}\left[(G_t - v_\pi(S_t))[v_\pi(S_t) - \hat{v}(S_t,\mathbf{w})]\right]. &(2)\end{aligned}$$

The final term is

$$\begin{aligned}\mathbb{E}\left[(G_t - v_\pi(S_t))[v_\pi(S_t) - \hat{v}(S_t,\mathbf{w})]\right] &= \mathbb{E}_{s\sim\mu}\left\{\mathbb{E}\left[(G_t - v_\pi(s))[v_\pi(s) - \hat{v}(s,\mathbf{w})]\right]|s\right\} &(3)\\ &= 0 &(4)\end{aligned}$$