Exercise 11.1 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. □

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

$$G_{t:t+n} = R_{t+1} - \bar{R}_t + \ldots + R_{t+n} - \bar{R}_{t+n-1} + \hat{v}(S_{t+n}, w_{t+n-1}) \ldots \underline{\text{continuing}}$$

$$G_{t:t+n} = R_{t+1} + \ldots + \gamma^{n-1} R_{t+n-1} + \gamma^{n} \hat{v}(S_{t+n}, w_{t+n-1}) \ldots \underline{\text{episodic}} \ (t+n < T)$$

$$G_{h:h} = R_T \text{ if } h = T \text{ or } \hat{v}(S_h) \text{ otherwise}$$

**Exercise 11.2* Convert the equations of n-step $Q(\sigma)$ (7.11 and 7.17) to semi-gradient form. Give definitions that cover both the episodic and continuing cases. □

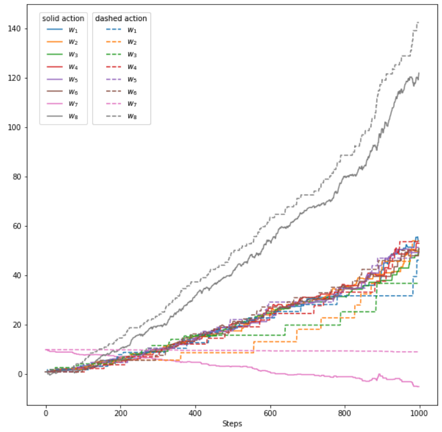
$$w_{t+n} = w_{t+n-1} + \alpha \rho_{t+1:t+n} \left[G_{t:t+n} - \hat{q}(S_t, A_t, w_{t+n-1}) \right] \nabla \hat{q}(S_t, A_t, w_{t+n-1})$$

$$G_{t:t+n} = R_{t+1} - \bar{R}_t + \gamma \left(\sigma_{t+1} \rho_{t+1} + (1 - \sigma_{t+1}) \pi(A_{t+1} | S_{t+1}) \right) \left(G_{t+1:t+n} - \hat{q}(S_{t+1}, A_{t+1}, w_{t+n-1}) \right) +$$
$$+ \sum_a \pi(a | S_{t+1}) \hat{q}(S_{t+1}, a, w_{t+n-1}) \ldots \underline{\text{continuing}}$$

$$G_{t:t+n} = R_{t+1} + \gamma \left(\sigma_{t+1} \rho_{t+1} + (1 - \sigma_{t+1}) \pi(A_{t+1} | S_{t+1}) \right) \left(G_{t+1:t+n} - \hat{q}(S_{t+1}, A_{t+1}, w_{t+n-1}) \right) +$$
$$+ \gamma \sum_a \pi(a | S_{t+1}) \hat{q}(S_{t+1} a, w_{t+n-1}) \ldots \underline{\text{episodic}} (t+n < T)$$

$$G_{h:h} = R_T \text{ if } h = T \text{ or } \hat{q}(S_h, A_h, w_{h-1}) \text{ otherwise}$$

Exercise 11.3 (programming) Apply one-step semi-gradient Q-learning to Baird's counterexample and show empirically that its weights diverge. □



**Exercise 11.4* Prove (11.24). Hint: Write the $\overline{\mathrm{RE}}$ 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). □

$$\overline{RE}(w) = \sum_s \mu(s)\left(G_t - \hat{v}(s,w)\right)^2 =$$

$$= \sum_s \mu(s)\left(\left(G_t - v_\pi(s)\right) + \left(v_\pi(s) - \hat{v}(s,w)\right)\right)^2 =$$

$$= \sum_s \mu(s)\left(\left(G_t - v_\pi(s)\right)^2 + \left(v_\pi(s) - \hat{v}(s,w)\right)^2 + 2\underbrace{\left(G_t - v_\pi(s)\right)}\left(v_\pi(s) - \hat{v}(s,w)\right)\right) =$$

$$= \sum_s \mu(s)\left(\underbrace{\left(G_t - v_\pi(s)\right)^2} + \underbrace{\left(v_\pi(s) - \hat{v}(s,w)\right)^2}\right)$$

$$= \mathbb{E}\left[\left(G_t - v_\pi(S_t)\right)^2\right] + \overline{VE}(w)$$

$\mathbb{E}\left[G_t - v_\pi(S_t)\right] = 0$

... in expectation true value function is equal to the return.