Exercises - Chapter 11

Carl Fredriksson, c@msp.se

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.

My answer:

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

If episodic: $\rho_k = 1$ for all $k \geq T$ (where T is the last step of the episode).

Episodic return:

$$G_{t:t+n} \doteq R_{t+1} + \cdots + \gamma^{n-1}R_{t+n} + \gamma^n\hat{v}(S_{t+n}, \mathbf{w}_{t+n-1})$$

with $G_{t:t+n} = G_t$ if $t + n \geq T$.

Continuing return:

$$G_{t:t+n} \doteq R_{t+1} - \bar{R}_t + \cdots + R_{t+n} - \bar{R}_{t+n-1} + \hat{v}(S_{t+n}, \mathbf{w}_{t+n-1})$$

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.

My answer:

From section 7.6:

> Then we use the earlier update for n-step Sarsa without importance-sampling ratios (7.5) instead of (7.11), because now the ratios are incorporated in the n-step return.

I believe the exercise description is wrong and it should be (7.5 and 7.17) rather than (7.11 and 7.17). Let $h = t + n$, we have

$$\mathbf{w}_h \doteq \mathbf{w}_{h-1} + \alpha\big[G_{t:h} - \hat{q}(S_t, A_t, \mathbf{w}_{h-1})\big]\nabla\hat{q}(S_t, A_t, \mathbf{w}_{h-1})$$

If episodic: $\rho_k = 1$ for all $k \geq T$ (where T is the last step of the episode).

Episodic return:

$$G_{t:h} \doteq R_{t+1} + \gamma\Big(\sigma_{t+1}\rho_{t+1} + (1 - \sigma_{t+1})\pi(A_{t+1}|S_{t+1})\Big)\Big(G_{t+1:h} - \hat{q}(S_{t+1}, A_{t+1}, \mathbf{w}_{h-1})\Big) + \gamma\sum_a \pi(a|S_{t+1})\hat{q}(S_{t+1}, a, \mathbf{w}_{h-1})$$

for $t < h \leq T$. The recursion ends with $G_{h:h} \doteq \hat{q}(S_h, A_h, \mathbf{w}_{h-1})$ if $h < T$, or with $G_{T-1:T} \doteq R_T$ if $h = T$.

Continuing return:

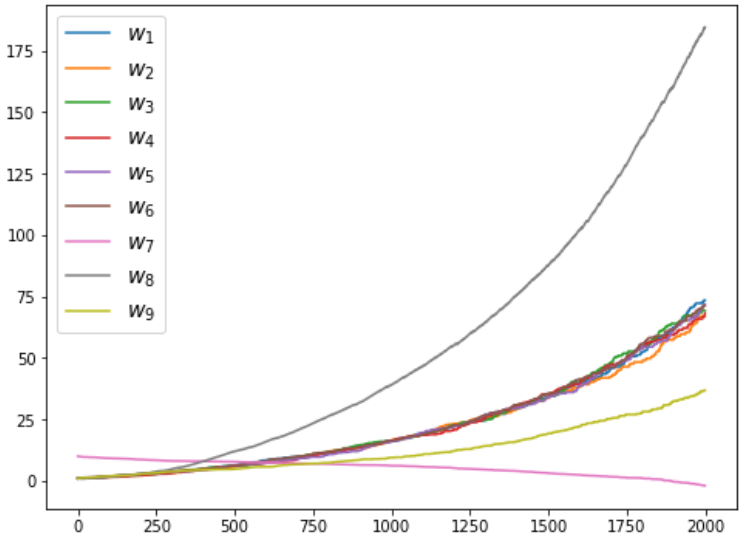
$$G_{t:h} \doteq R_{t+1} - \bar{R}_t + \Big(\sigma_{t+1}\rho_{t+1} + (1 - \sigma_{t+1})\pi(A_{t+1}|S_{t+1})\Big)\Big(G_{t+1:h} - \hat{q}(S_{t+1}, A_{t+1}, \mathbf{w}_{h-1})\Big) + \sum_a \pi(a|S_{t+1})\hat{q}(S_{t+1}, a, \mathbf{w}_{h-1})$$

The recursion ends with $G_{h:h} \doteq \hat{q}(S_h, A_h, \mathbf{w}_{h-1})$.

Exercise 11.3 (programming)

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

My answer:



Exercise 11.4

Prove (11.24). Hint: Write the $\overline{RE}$ 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).

My answer:

$$
\begin{aligned}
\overline{RE}(\mathbf{w}) &= \mathbb{E}\Big[\big(G_t - \hat{v}(S_t, \mathbf{w})\big)^2\Big] \\
&= \mathbb{E}\Big[\mathbb{E}\Big[\big(G_t - \hat{v}(s, \mathbf{w})\big)^2 \mid S_t = s\Big]\Big] \\
&= \sum_{s \in \mathcal{S}} \mu(s) \mathbb{E}\Big[\big([G_t - v_\pi(s)] + [v_\pi(s) - \hat{v}(s, \mathbf{w})]\big)^2 \mid S_t = s\Big] \\
&= \sum_{s \in \mathcal{S}} \mu(s) \mathbb{E}\Big[\big(G_t - v_\pi(s)\big)^2 + 2\big(G_t - v_\pi(s)\big)\big(v_\pi(s) - \hat{v}(s, \mathbf{w})\big) + \big(v_\pi(s) - \hat{v}(s, \mathbf{w})\big)^2 \mid S_t = s\Big] \\
&= \overline{VE}(\mathbf{w}) + \mathbb{E}\Big[\big(G_t - v_\pi(S_t)\big)^2\Big] + 2\sum_{s \in \mathcal{S}} \mu(s) \mathbb{E}\Big[G_t v_\pi(s) - G_t \hat{v}(s, \mathbf{w}) - v_\pi(s)^2 + v_\pi(s)\hat{v}(s, \mathbf{w}) \mid S_t = s\Big] \\
&= \overline{VE}(\mathbf{w}) + \mathbb{E}\Big[\big(G_t - v_\pi(S_t)\big)^2\Big] + 2\sum_{s \in \mathcal{S}} \mu(s) \Big[v_\pi(s)^2 - v_\pi(s)\hat{v}(s, \mathbf{w}) - v_\pi(s)^2 + v_\pi(s)\hat{v}(s, \mathbf{w})\Big] \\
&= \overline{VE}(\mathbf{w}) + \mathbb{E}\Big[\big(G_t - v_\pi(S_t)\big)^2\Big]
\end{aligned}
$$