

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

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

In the continuing case

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

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

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

In the episodic case

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

In the continuing case

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