JUSTINE JOHN A SERDONCILLO
IE ST 71 REINFORCEMENT LEARNING & DYNAMIC PROGRAMMING

NOV 21, 2023
HOMEWORK #3

EXERCISE 7.4 PROVE THAT THE N-STEP RETURN OF SARSA (7.4) CAN BE WRITTEN EXACTLY IN TERMS OF A NOVEL TD ERROR, AS

(7.4)

$$G_{t:t+n} \doteq R_{t+1} + \gamma R_{t+2} + \dots + \gamma^{n-1} R_{t+n} + \gamma^n Q_{t+n-1}(S_{t+n}, A_{t+n})$$

$$n \geq 1, 0 \leq t < T-n$$

(7.6)

$$G_{t:t+n} = Q_{t-1}(S_t, A_t) + \sum_{k=t}^{\min(t+n,T)-1} \gamma^{k-t}\left[R_{k+1} + \gamma Q_k(S_{k+1}, A_{k+1}) - Q_{k-1}(S_k, A_k)\right]$$

Start with (7.6)

$$G_{t:t+n} = Q_{t-1}(S_t, A_t) + \sum_{k=t}^{\min(t+n-1,T-1)} \gamma^{k-t}\left[R_{k+1} + \gamma Q_k(S_{k+1}, A_{k+1}) - Q_{k-1}(S_k, A_k)\right]$$

expand a lot to get

$$= Q_{t-1}(S_t, A_t) + \left[R_{t+1} + \gamma Q_t(S_{t+1}, A_{t+1}) - Q_{t-1}(S_t, A_t)\right]$$

$$+ \gamma R_{t+2} + \gamma^2 Q_{t+1}(S_{t+2}, A_{t+2}) - \gamma Q_t(S_{t+1}, A_{t+1})$$

$$+ \gamma^2 R_{t+3} - \gamma^2 Q_{t+1}(S_{t+2}, A_{t+2}) + \gamma^3 Q_{t+2}(S_{t+3}, A_{t+3})$$

lots of terms will cancel out then

$$\boxed{G_{t:t+n} = R_{t+1} + \gamma R_{t+2} + \dots + \gamma^{n-1} R_{t+n} + \gamma^n Q_{t+n-1}(S_{t+n}, A_{t+n})}$$

EXERCISE 7.6 PROVE THAT THE CONTROL VARIATE IN THE ABOVE EQUATIONS DOES NOT CHANGE THE EXPECTED OF THE RETURN.

---

off policy

$$G_{t:h} = R_{t+1} + \gamma\left(\rho_{t+1} G_{t+1:h} + \bar{V}_{h-1}(S_{t+1})\right) - \rho_{t+1} Q_{h-1}(S_{t+1}, A_{t+1})$$

$$= R_{t+1} + \gamma\rho_{t+1}\left(G_{t+1:h} - Q_{h-1}(S_{t+1}, A_{t+1})\right) + \gamma \bar{V}_{h-1}(S_{t+1}) \quad t < h < T.$$

combine this w/ algorithm of n-step SARSA

$$E[G_{t:h}] = E\left[R_{t+1} + \gamma\rho_{t+1}\left(G_{t+1:h} - Q_{h-1}(S_{t+1}, A_{t+1})\right) + \gamma \bar{V}_{h-1}(S_{t+1})\right]$$

$$= E[R_{t+1}] + E\left[\gamma\rho_{t+1}\left(G_{t+1:h} - Q_{h-1}(S_{t+1}, A_{t+1})\right)\right] + E\left[\gamma \bar{V}_{h-1}(S_{t+1})\right]$$

$E[\rho] = 1$

$$= R_{t+1} + E\left[\gamma G_{t+1:h} - \gamma Q_{h-1}(S_{t+1}, A_{t+1})\right] + \gamma \bar{V}_{h-1}(S_{t+1})$$

$$= R_{t+1} + E\left[G_{t:h} - R_{t+1} - \gamma Q_{h-1}(S_{t+1}, A_{t+1})\right] + \gamma \bar{V}_{h-1}(S_{t+1})$$

$$= E[G_{t:h}] + \gamma\left[E\left[-Q_{h-1}(S_{t+1}, A_{t+1})\right] + \bar{V}_{h-1}(S_{t+1})\right]$$

$$= E[G_{t:h}]$$