RL: Lecture 16

Harvey Mudd College

March 30, 2020

Neil Rhodes

Off-policy learning with importance sampling: State value

Without control variate:

$$\tilde{G}_{t:h} = \rho_t (R_{t+1} + \gamma \tilde{G}_{t+1:h})$$

With control variate:

$$\hat{G}_{t:h} = \rho_t (R_{t+1} + \gamma \hat{G}_{t+1:h}) + (1 - \rho_t) V_{h-1}(S_t)$$

Off-policy learning with importance sampling: Action value

Without control variate:

$$\tilde{G}_{t:h} = R_{t+1} + \gamma \rho_{t+1} \tilde{G}_{t+1:h}$$

With control variate:

$$\begin{aligned}
\hat{G}_{t:h} &= R_{t+1} + \gamma(\rho_{t+1}\hat{G}_{t+1:h} + \mathbb{E}_b[\rho_{t+1}Q_{h-1}(S_{t+1}, A_{t+1})] - \rho_{t+1}Q_{h-1}(S_{t+1}, A_{t+1})) \\
&= R_{t+1} + \gamma(\rho_{t+1}\hat{G}_{t+1:h} + \bar{V}_{h-1}(S_{t+1}) - \rho_{t+1}Q_{h-1}(S_{t+1}, A_{t+1})) \\
&= R_{t+1} + \gamma\rho_{t+1}(\hat{G}_{t+1:h} - Q_{h-1}(S_{t+1}, A_{t+1})) + \gamma\bar{V}_{h-1}(S_{t+1})
\end{aligned}$$

Off-Policy learning without importance sampling: n-step Tree Backup

one-step:

$$G_{t:t+1} = R_t + \gamma \sum_a \pi(a|S_{t+1}) Q_t(S_{t+1}, a)$$

Off-Policy learning without importance sampling: n-step Tree Backup

two-step:

$$G_{t:t+2} = R_{t+1} + \gamma \sum_{a \neq A_{t+1}} \pi(a|S_{t+1}) Q_t(S_{t+1}, a)$$
$$+ \gamma \pi(A_{t+1}|S_{t+1}) \left(R_{t+2} + \gamma \sum_a \pi(a|S_{t+2}) Q_{t+1}(S_{t+2}, a) \right)$$
$$= R_{t+1} + \gamma \sum_{a \neq A_{t+1}} \pi(a|S_{t+1}) Q_t(S_{t+1}, a) + \gamma \pi(A_{t+1}, S_{t+1}) G_{t+1:t+2}$$

Off-Policy learning without importance sampling: n-step Tree Backup

n-step:

$$G_{t:t+n} = R_{t+1} + \gamma \sum_{a \neq A_{t+1}} \pi(a|S_{t+1}) Q_t(S_{t+1}, a) + \gamma \pi(A_{t+1}, S_{t+1}) G_{t+1:t+n}$$

n-step $Q(\sigma)$

tree-backup return:

$$G_{t:t+n} = R_{t+1} + \gamma \sum_{a \neq A_{t+1}} \pi(a|S_{t+1}) Q_t(S_{t+1}, a) + \gamma \pi(A_{t+1}, S_{t+1}) G_{t+1:t+n}$$

$$= R_{t+1} + \gamma \bar{V}_{h-1}(S_{t+1}) - \gamma \pi(A_{t+1}|S_{t+1}) Q_{h-1}(S_{t+1}, A_{t+1}) + \gamma \pi(A_{t+1}, S_{t+1}) G_{t+1:t+n}$$

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

per-decision importance sampling return:

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

combined:

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