RL: Lecture 16

2) Adjustments to syllabus

3) Office hours / logistics

2 - Now drop 2 hw/quiz assignments & drop 1 midterm
Off-policy learning with importance sampling: State value

Without control variate:

\[ \rho_t = \frac{\pi(a_t \mid s_t)}{\pi^b(a_t \mid s_t)} \]

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

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:

\[ E_{\pi}[Q_{t+1}(S_{t+1}, A_{t+1})] = E_{\pi}[Q_{h-1}(S_{t+1}, A_{t+1})] - \nabla E_{\pi}[V_{h-1}(S_{t+1})] \]

\[ \bar{G}_{t:h} = R_{t+1} + \gamma \rho_{t+1} \bar{G}_{t+1:h} \]

With control variate:

\[ E_b[\rho_{t+1} Q_{h-1}(S_{t+1}, A_{t+1})] - \rho_{t+1} Q_{h-1}(S_{t+1}, A_{t+1}) \]

\[ \bar{G}_{t:h} = R_{t+1} + \gamma (\rho_{t+1} \bar{G}_{t+1:h} + V_{h-1}(S_{t+1}) - \rho_{t+1} Q_{h-1}(S_{t+1}, A_{t+1})) \]

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

if \( \rho_{t+1} \approx 0 \)
\[ \tilde{V}(s) = \sum_a \pi(a|s) Q(s, a) \]

We are estimating action values \( Q(s, a) \)
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 $Q(\sigma)$

$$\sum_{a \in \text{Action}} \text{value} = \frac{\text{value}}{V(s)} - Q(s, A_{t+1})$$

**Tree-backup return:**

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

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

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

**Per-decision importance sampling return:**

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

**Combined:**

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

$\sigma_t \in \text{Binary } [0, 1]$  

$\sigma_t$ can be different
For one paper descending $\sigma$

$\sigma = 0$
$\sigma = 1$
$\sigma = 0.2$
$\sigma = 0.6$

change from start $\sigma @ 1$
over from reducing to $0.95$
saying $\sigma (0.95)^2$
to $\sigma$

$\text{RMS}$

19-state random walk tree backup $\sigma (0.95)^n \rightarrow 0$
importance sampling
ordinary importance sampling
weighted importance sampling
per-decision importance sampling

$V(s)$

1-step $\Rightarrow$

true value

tree backup

time

descending $\sigma$