RL: Lecture 16

Harvey Mudd College

March 30, 2020

Neil Rhodes

1) n-step TD

2) Adjustments to syllabus

3) Office hours / logistics

2 — Now drop 2 hws, quizzes, programming assignments

& drop 1 midterm

$E_b[(1-\rho_t)V_{h-1}(S_t)]$

$E_b[\rho_t] = 1$ & ρ_t independent of $V_{h-1}(S_t)$

behavior policy b

target policy π

Off-policy learning with importance sampling: State value

$\rho_t = \dfrac{\pi(a_t|s_t)}{b(a_t|s_t)}$

Without control variate:

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

curr state value X

target state value O

With control variate:

control variate

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

$\rho_t = 0$

$0(R_{t+1} + \gamma G_{t+1:h}) + 1 V_{h-1}(S_t)$

$= V_{h-1}(S_t)$

$E_b[\underbrace{\rho_{t+1}}Q_{h-1}(S_{t+1}, A_{t+1})] = E_\pi[Q_{h-1}(S_{t+1}, A_{t+1})]$

$= \boxed{V_{h-1}(S_{t+1})}$ $\quad E_b[V_{h-1}(S_{t+1})] = V_{h-1}(S_{t+1})$

Off-policy learning with importance sampling: Action value

$\bar{V}(s) = \sum_a \pi(a|s)\, Q(s,a)$

$Q_\pi(S_t, A_t)$

$E[5] = 5$

Without control variate:

$E[E[R] - R] = E[E[R]] - E[R] = 0$

ρ_t tells us the ratio of $\frac{\pi(A_t|S_t)}{b(A_t|S_t)}$

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

With control variate:

$E_b[\text{control variate}] = 0$

$E_b[E_b[\rho_{t+1} Q \cdots] - \rho_{t+1} Q \cdots]$

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

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

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

if $\rho_{t+1} = 0$

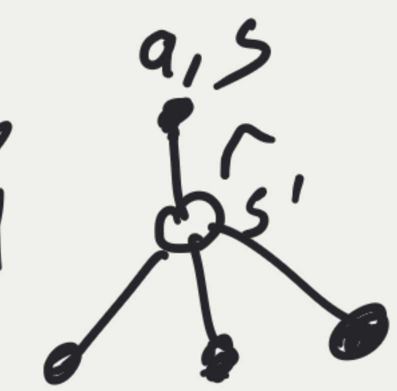
$$\bar{V}(s) = \sum_{a} \pi_\psi(a|s)\, Q(s,a)$$

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

If we don't know b, can we do importance sampling? No. requires ρ which requires b

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

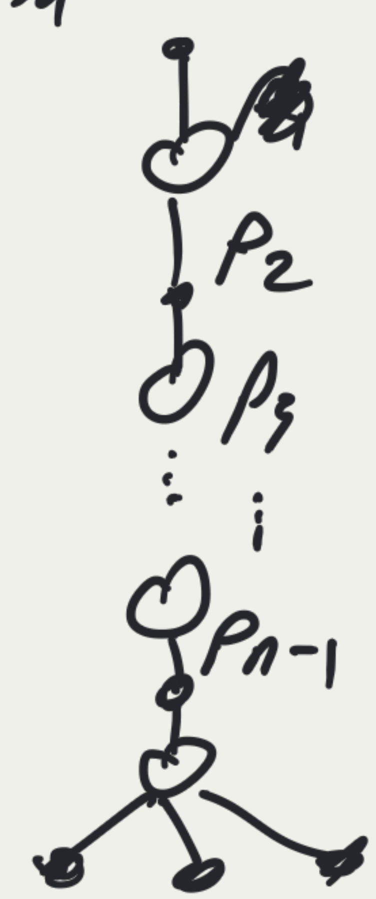
1-step expected Sarsa



one-step:

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

n-step expected Sarsa



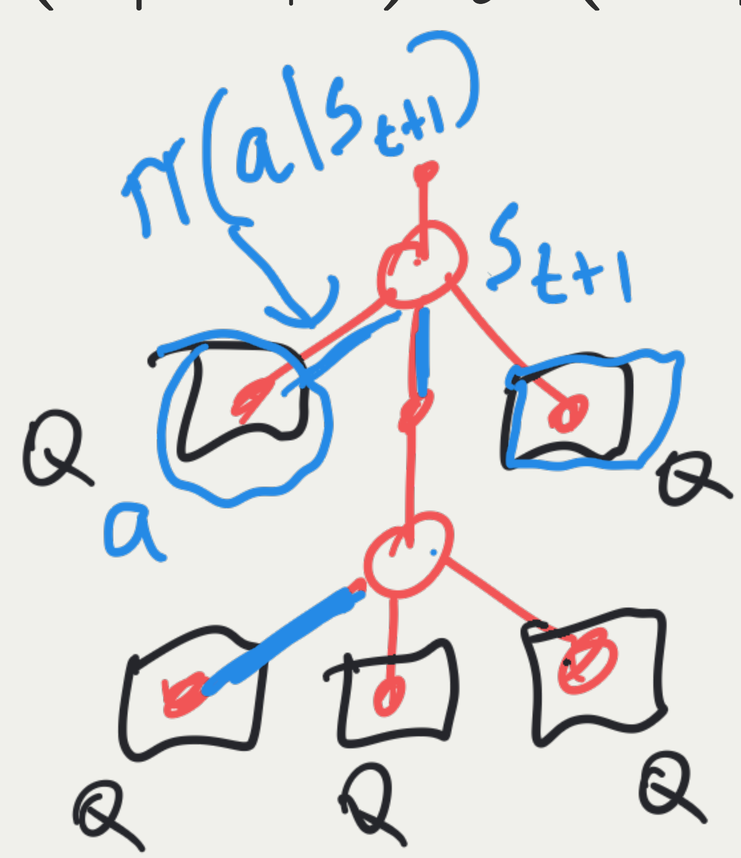
Off-Policy learning without importance sampling: n-step Tree Backup

two-step:

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

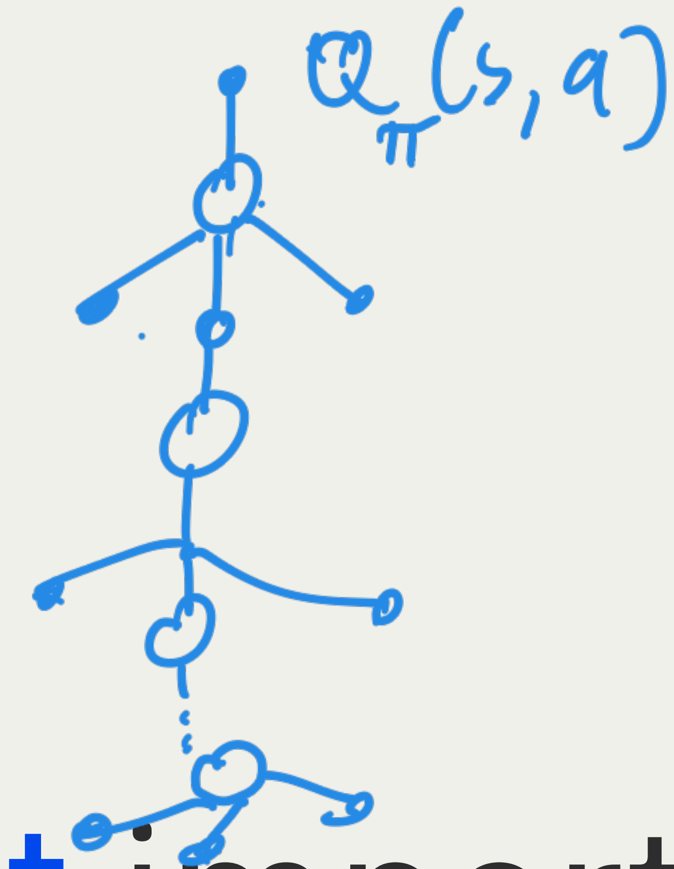
$$+\gamma \boxed{\pi(A_{t+1}|S_{t+1})} \left(\boxed{R_{t+2}} + \gamma \boxed{\sum_{a} \pi(a|S_{t+2}) Q_{t+1}(S_{t+2}, a)} \right)$$

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



	Bias	Variance
tree backup	higher	lower
importance sampling	lower	higher

goes down over time



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 $Q(\sigma)$

$\sum_{a \neq A_{t+1}} \text{value} = \underbrace{\sum_a \text{value}}_{= \bar{V}(s)} - \text{value@}a$

$\bar{V}(s) - Q(s, A_{t+1})$

tree-backup return:

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

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

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

per-decision importance sampling return:

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

combined:

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

$\sigma_t \in$ Range $[0,1]$

σ_i's can be different

For one paper

$\sigma = 0$

$\sigma = 1$

$\sigma = .2$

$\sigma = .6$

descending σ

change over from time

start σ @ 1

reduce to 0.95

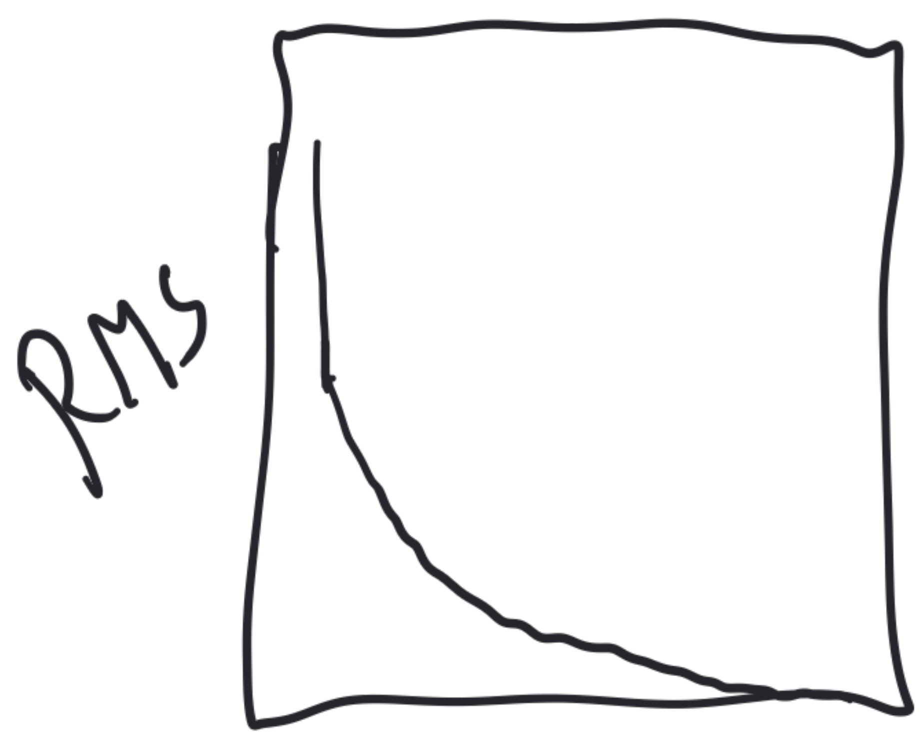
sampling " " $(0.95)^2$

⋮

$(0.95)^n \approx 0$

to

19-state random-walk tree backup



importance sampling

ordinary importance sampling

weighted importance sampling

n-step → per-decision importance sampling