1) Performance Difference Lemma

Goal: understand V^{π} vs. $V^{\pi'}$ in terms of π vs. π'

Lemma (PDL):

$$V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_{s_0}^{\pi}} \left[\underbrace{\mathbb{E}_{a \sim \pi(s)}[Q^{\pi'}(s,a)] - V^{\pi'}(s)} \right]$$

for $r(s,a) \in [0,1] \ \forall s,a$

$$|V^{\pi}(s_0) - V^{\pi'}(s_0)| \le \frac{1}{(1-\gamma)^2} \mathbb{E}_{s \sim d_{s_0}^{\pi}} \left[\underbrace{\sum_a |\pi(a|s) - \pi'(a|s)|}_{\|\pi(\cdot|s) - \pi'(\cdot|s)\|_1} \right]$$

Def (Advantage Function)

$$A^{\pi}(s,a) = Q^{\pi}(s,a) - V^{\pi}(s) \qquad A^{\pi}(s, \pi(s)) = 0$$

"advantage" of taking action a in state s rather than following π

$$A^{\pi^*}(s,a) \le 0 \qquad \pi^*(s) = \arg\max_a Q^*(s,a)$$

$$Q^{\pi^*}(s,a) - V^{\pi^*}(s)$$

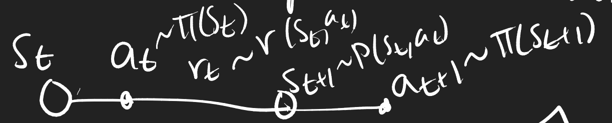
$$\arg\max_a Q^{\pi}(s,a) = \arg\max_a A^{\pi}(s,a)$$

2) Supervision via Bellman Equation

The Bellman Expection Equation

$$Q^{\pi}(s,a) = r(s,a) + \underbrace{\gamma \mathbb{E}_{\substack{s' \sim P(s,a) \\ a' \sim \pi(s)}} [Q^{\pi}(s', a')]}$$

IDEA: Bootstrap a label with one timestep



$$y_t = r_t + \gamma \hat{Q}(s_{t+1}, a_{t+1}) \approx Q^{\pi}(s_t, a_t)$$

"Temporal Difference" target

TD error : $r_t + \gamma \hat{Q}(s_{t+1}, a_{t+1}) - \hat{Q}(s_t, a_t)$

Alg SARSA subroutine ("state-action-reward-state-action")

initialize $\hat{Q}^0$, $s_0 \sim \mu_0$, $a_0 \sim \pi(s_0)$

for $t = 0, 1, \ldots$

Take a_t, observe $s_{t+1} \sim P(s_t, a_t)$ & $r_t \sim r(s_t, a_t)$

sample $a_{t+1} \sim \pi(s_{t+1})$

$$\hat{Q}^{t+1}(s_t, a_t) = (1-\alpha)\hat{Q}^t(s_t, a_t) + \alpha \big(\underbrace{r_t + \gamma \hat{Q}^t(s_{t+1}, a_{t+1})}_{y_t}\big)$$

ε-Greedy Policy improvement

$$\pi(s) = \begin{cases} \arg\max_a \hat{Q}(s,a) & \text{w.p. } 1-\varepsilon \\ \begin{bmatrix} a_0 & \text{w.p. } \varepsilon/A \\ a_1 & \text{w.p. } \varepsilon/A \\ \vdots & \vdots \end{bmatrix} \end{cases}$$

$$\pi(a|s) = \begin{cases} \varepsilon/A & a \neq \arg\max \hat{Q}(s,a) \quad (A-1) \\ 1 - \varepsilon + \frac{\varepsilon}{A} & \text{o.w.} \end{cases}$$

compare Rollout (MC) vs. Bellman Exp. (T.D.) Supervision

1) TD can update $\hat{Q}$ online @ every step
   MC waits until end of rollout

2) TD is biased when $\hat{Q} \neq Q^{\pi}$
   $r_t + \gamma Q^{\pi}(s_{t+1}, a_{t+1})$ is unbiased

   MC is unbiased

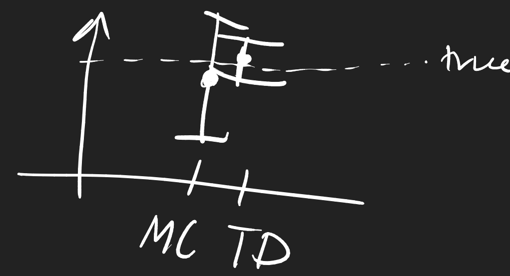
3) TD has smaller variance (informally)

$a_t \sim \pi(s_t)$

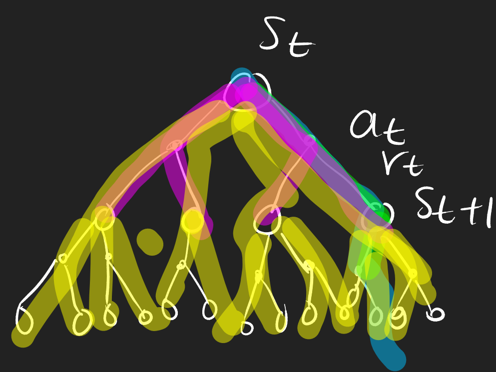
$s_{t+1} \sim P(s_t, a_t)$

$a_{t+1} \sim \pi(s_{t+1})$

MC has higher variance



4) Both methods use data collected with π – "on policy"



MC TD DP/BE $Q^\pi(s,a) = \mathbb{E}\left[\sum_0^t \gamma^t r_t \,\middle|\, \substack{\pi, P, r \\ s_0, a_0 = s, a}\right]$

3) Supervision w/ Bellman Optimality

Can we estimate Q^*?

Recall: Value Iteration

$$\boxed{\begin{array}{l} \text{Initialize } Q^0 \\ \text{for } t = 0, 1, \ldots, \\ \quad Q^{t+1} = \text{BellmanOperator}(Q^t) \end{array}}$$

BellmanOp(Q^t):

$$Q^{t+1}(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)}\left[\max_{a'} Q^t(s',a')\right]$$

Recall: Bellman Optimality

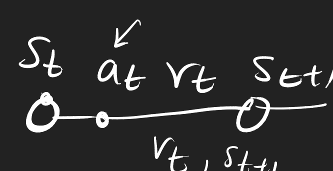
$$\to Q^*(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)}\Big[\underbrace{\max_{a'} Q^*(s',a')}_{V^*(s) = Q^*(s, \pi^*(s))}\Big]$$

$\pi^*(s) = \arg\max_a Q^*$

Bellman-Opt. Supervision:

$$y_t = r_t + \gamma \max_{a \in \mathcal{A}} \hat{Q}(s_{t+1}, a)$$

$$y_t \approx Q^*(s_t, a_t)$$



(s_t, a_t, y_t)

Q-learning

initilize Q^0

for $t = 0, 1, \cdots$

(e.g. εgreedy Q^t)

take action a_t & observe $s_{t+1} \sim P(s_t, a_t)$, $r_t \sim r(s_t, a_t)$

$$Q^{t+1}(s_t, a_t) \leftarrow (1-\alpha)\, Q^t(s_t, a_t) + \alpha \left(r_t + \gamma \max_{a'} Q(s_{t+1}, a') \right)$$