Last Time

Policy Gradient π_θ

$$\theta \leftarrow \theta + \alpha \nabla U(\theta)$$

↑ estimate with 1 sim

Likelihood Ratio
Causality
Baseline subtraction

Model-Based ⟷ **Model Free**

Learn T R
solve

Learn π
Learn Q ← Value Learning / TD Learning

Today

Incremental Mean Estimation

$$\hat{x}_m = \frac{1}{m} \sum_{i=1}^{m} x^{(i)}$$

$$= \frac{1}{m}\left(x^{(m)} + \sum_{i=1}^{m-1} x^{(i)}\right)$$

$$= \frac{1}{m}\left(x^{(m)} + (m-1)\hat{x}_{m-1}\right)$$

$$\rightarrow = \hat{x}_{m-1} + \frac{1}{m}\left(x^{(m)} - \hat{x}_{m-1}\right)$$

$q = r + \gamma$ simulate!

$$Q(s,a) = Q(s,a) + \frac{1}{N(s,a)}\left(q - Q(s,a)\right)$$

$$\hat{x} \leftarrow \hat{x} + \alpha\left(x - \hat{x}\right)$$

↑ learning rate

Q-learning (Conceptual)

$$Q(s,a) = R(s,a) + \gamma \sum_{s'} T(s'|s,a)\, V(s')$$

$$\doteq R(s,a) + \gamma \sum_{s'} T(s'|s,a) \max_{a'} Q(s',a')$$

$$Q(s,a) = \mathop{E}_{r,s'}\left[r + \gamma \max_{a'} Q(s',a') \right]$$

Q-learning (s,a,r,s')

$Q(s,a) \leftarrow 0 \quad \forall s,a$

$s \leftarrow s_0$

loop

$a \leftarrow$ exploration_policy (Q) ← e.g. ε-greedy

$s', r \leftarrow$ step! (env, a)

$Q(s,a) \leftarrow Q(s,a) + \alpha \left(r + \gamma \max_{a'} Q(s',a') - Q(s,a) \right)$

TD

book version

SARSA (s,a,r,s',a')

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left(r + \gamma Q(s',a') - Q(s,a) \right)$$

Off-Policy ⟷ **On-Policy**

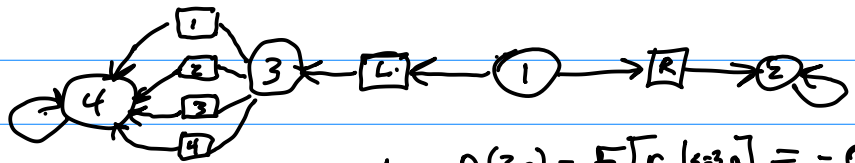
Learn only from actions that are in the policy they are executing

Off-Policy: Q-learning; Fewer Samples

On-Policy: SARSA; More stable; Policy Gradient

Breakout Rooms

$A(3) = \{1,2,3,4\}$ $A(1) = \{L, R\}$



same reward $R(3,a) = E[r \mid s=3, a] = -0.1$

$r \sim N(-0.1, 1)$

After a few episodes

$Q(3,1)$

$Q(3,2)$

Will Q-learning work well?

Right side stays at 0

Max throws off estimate on Left

Maximization Bias

$Q_2(3,1) = -0.1$ $\quad Q_1(3,1) = 0.2$

$\quad\quad\quad\quad\quad\quad Q_1(3,2) = 0.1$

$Q_2(3,3) = 0.1$ $\quad Q_1(3,3) = -0.3$

$\quad\quad\quad\quad\quad\quad Q_1(3,4) = -0.5$

$\max \rightarrow Q(1,L)$

Solution

Double Q-Learning

$Q_1 \quad Q_2$

$$Q_1(s,a) \leftarrow Q_1(s,a) + \alpha\left(R(s,a) + \gamma\, Q_2\left(s', \underset{a'}{\text{argmax}}\, Q_1(s',a')\right) - Q_1(s,a)\right)$$

S+B section 6.7

Credit Assignment

Problem: takes a long time to propagate backwards

Solution: Eligibility Traces (look at S+B)

Sarsa(λ) $\quad$ Q(λ) (Tricky) $\quad$ TD(λ) $\quad$ λ: decay param $\in [0,1]$

s,a,r,s',a'

$Q(s,a), N(s,a) \leftarrow 0 \quad \forall s,a$

loop

$a_{t+1} \leftarrow$ explore (Q)

$N(s_t,a_t) \leftarrow N(s_t,a_t) + 1$

$\delta \leftarrow r_t + \gamma\, Q(s_{t+1}, a_{t+1}) - Q(s_t,a_t)$

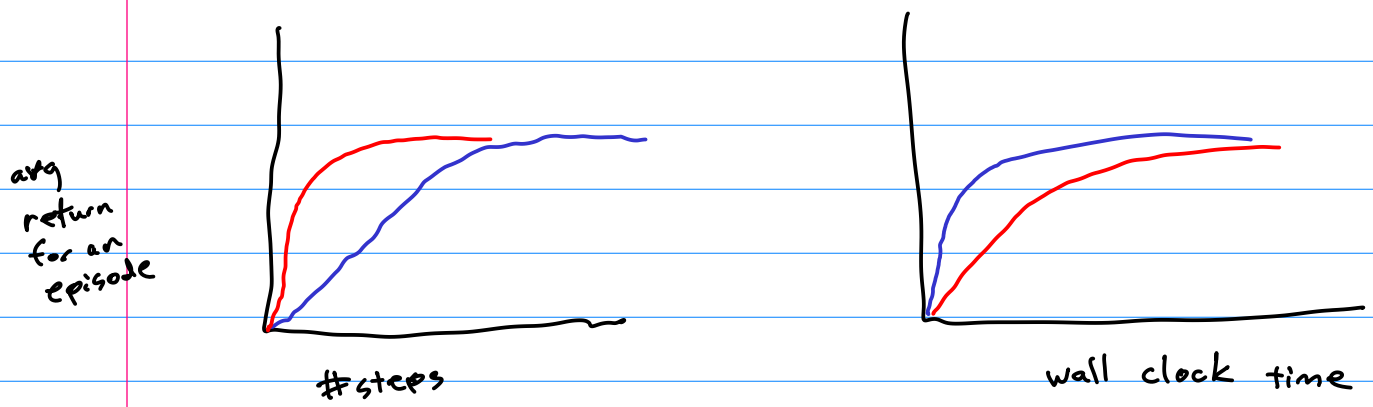
$Q(s,a) \leftarrow Q(s,a) + \alpha \delta N(s,a) \quad \forall s,a$

$N(s,a) \leftarrow \gamma \lambda N(s,a) \quad \forall s,a$

Convergence

Q-learning converges to optimal Q-values w.p. 1 S+B p.131

Sarsa converges to optimal Q-values w.p. 1 provided that $\pi \rightarrow$ greedy(Q) S+B p.129


avg return for an episode
#steps
wall clock time