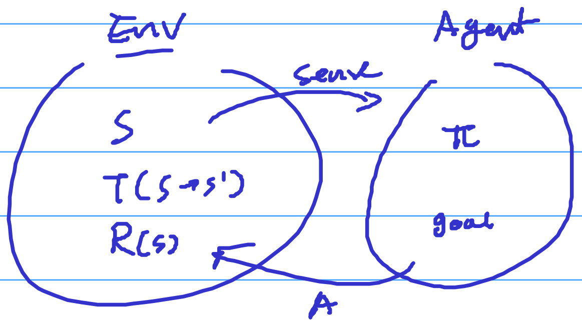
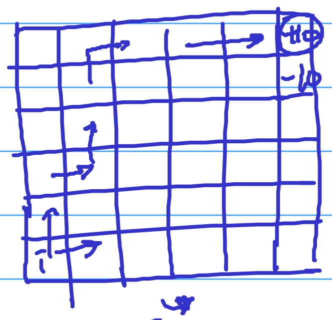
AI



finite, discrete state-action space

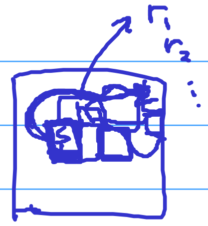


MDP

$\left\{ \begin{array}{l} T(s,a,s') \\ R(s) \end{array} \right\} \to$ DP

RL

$\left. \begin{array}{l} T \times \\ R \times \end{array} \right\} \to$



RL
- model-based : $\hat{T}, \hat{R}$ } → DP { VI, PI
- model-free :
  - Passive RL : evaluate π → V { MC (Direct sampling); TD (temporal-difference) = Moving average
  - active RL π

$$V_{k+1}(s) \leftarrow V_k(s) + \alpha\left[R_k + \gamma V_k(s') - V_k(s)\right]$$

- MC control : Q(s,a)
- TD control : Q

- onpolicy : ε-greedy { ε: π ; 1-ε : explore
- off-policy : π, b

Q-learning : off policy , TD-control

SARSA : on-policy "

SARSA vs Q-Learning

- A from s using ε-greedy Q ;
- A' from s' " "

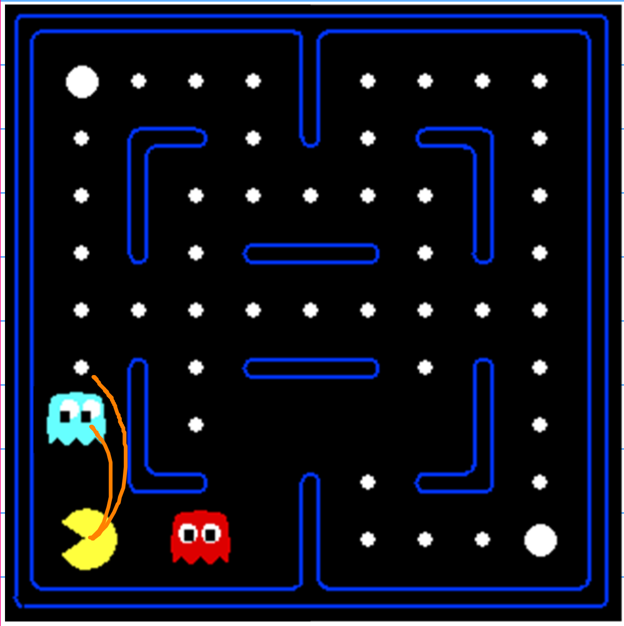
b: A from s using ε-greedy

$$Q(S,A) \leftarrow Q(S,A) + \alpha\left[R + \gamma Q(S',A') - Q(S,A)\right]$$

π

$$Q(S,A) \leftarrow Q(S,A)\ \alpha\left[R + \gamma \max_a Q(S',a) - Q(S,A)\right]$$

Approximation Method



$$\hat{V} = \sum_j w_j \cdot f_j(s)$$

$$\hat{Q} = \sum_j w_j \cdot f_j(s,a)$$

$$\boxed{Q^{new} \leftarrow Q^{old} + \alpha(\text{difference})}$$

$$Q_{target} - \hat{Q}$$

$$w_j \leftarrow w_j + \alpha \frac{\partial \ell}{\partial w_j}$$

$$\ell = \frac{1}{2}(Q_t - \hat{Q})^2 \qquad \boxed{(Q_t - \hat{Q}) \cdot \frac{\partial \hat{Q}}{\partial w_j}}$$

$$\boxed{w_j \leftarrow w_j + \alpha \cdot (Q_t - \hat{Q}) \cdot f_j(s,a)}$$

1. non-linear model → ?
2. $\boxed{Q_{true}}$
3. Deep RL

ML

Supervised learning $\ell(y, \hat{y})$

- parametric model $\hat{y} = f_\theta(x)$ $\{\theta\}$, gradient desc
- non-parametric, $\hat{y} = f(x)$ dis ent

•) MC → $E(r)$ → $Q(s,a) = G_t$

•) Bootstrapping $[r + \gamma \max_{a'} Q_\theta(s', a')]$ → semi-gradient method

Q_t

$$(Q_t - \hat{Q}) \frac{\partial \hat{Q}}{\partial \theta}$$

$$w_j \leftarrow w_j + \alpha \frac{\partial \hat{\phi}_w}{\partial w}(s_i, a_i)\,(Q_t - \hat{Q}_\theta(s_i, a_i))$$

$$Q_t = r(s,a) + \gamma \max_{a'} \hat{\phi}_w(s', a')$$

Q-learning /w $\boxed{\text{F.A}}$

1) sample (a) $s, a \rightarrow s', r$ # 1 Samples are correlate ✓

2) $Q_t = r(s,a) + \gamma \max_{a'} \hat{Q}_{w}(s', a')$ # 2 off-policy, bootstrapped ✓

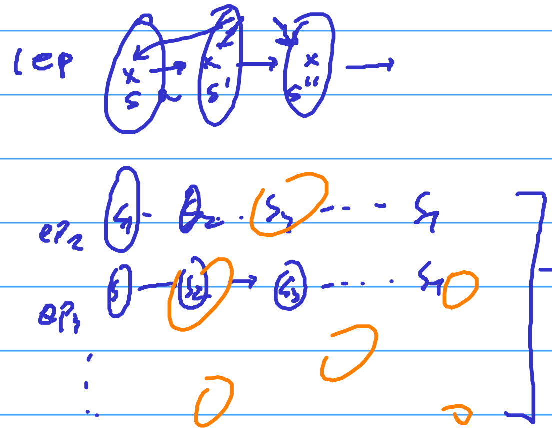
3) $w \leftarrow w + \alpha \frac{\partial \hat{Q}(s,a)}{\partial w} \left(Q_t - \hat{Q}(s,a) \right)$ # Semi-gradient descent

Deep Q-Network 2015

1. F.A → DNN
2. Off-policy
3. Bootstrapping

→ never converge!

P1. Correlated samples. → replay buffer



→ No sample correlation

P2, 3: Regression target not stable

$w' \leftarrow w$

$\{s_i, a_i, s'_i, r_i\} \rightarrow$ RB

N, K: $w \leftarrow w + \alpha \sum_i \frac{\partial \hat{Q}_w}{\partial w} \left(r(s_i, a_i) + \gamma \max_{a'} \hat{Q}_{w'}(s'_i, a'_i) - \hat{Q}_w(s_i, a_i) \right)$

target network

$w' = w$

→ Stable regression

N = 1000 ~ 10,000

$w' \leftarrow \tau w' + (1-\tau) w$