NNs for control: Reinforcement Learning

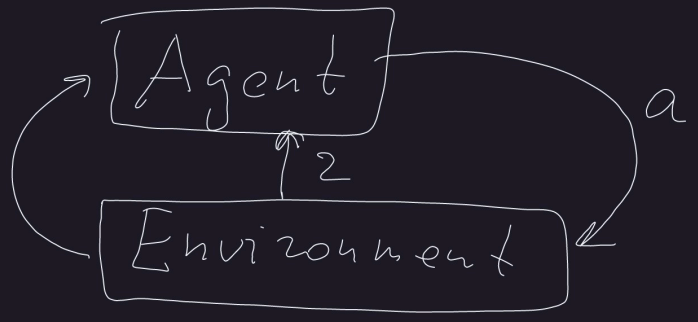
ML/DL

Dataset $\{x_i, y_i\}_{i=1}^{N}$ $\quad f(x, \theta)$

$$F(\theta) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(y_i, f(x_i, \theta)) + \lambda R(\theta) \to \min_{\theta}$$

$$\theta = \text{opt\_step}(\nabla_{\theta} F)$$

RL



$s \in S$ – state

$a \in A$ – action

$\pi(a|s)$ – policy

$p(s'|s,a)$ – transition prob.

$r(s,a)$ – reward

Trajectory

$\tau = \{ s_0, a_0, r_0, s_1, a_1, r_1, \dots \}$

$p(\tau|\pi) = p(s_0)\,\pi(a_0|s_0)\,p(s_1|s_0,a_0)\,\pi(a_1|s_1) \cdots$

MDP $p(s_t|a_{t-1}, s_{t-1}, a_{t-2}, s_{t-2}, \dots, a_0, s_0) = p(s_t|s_{t-1}, a_{t-2})$

$$R_t := r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots \qquad \gamma \in (0, 1)$$

↑
cumulative reward or return

$$J(\pi) = \mathop{\mathbb{E}}_{p(\tau|\pi)} R_0 \longrightarrow \max_{\pi}$$

$$V^{\pi}(s) := \mathbb{E}_{p(\tau|\pi)}[R_t | s_t = s]$$ — value function or state function

$$V^{*}(s) := \max_{\pi} V^{\pi}(s)$$

$$Q^{\pi}(s,a) := \mathbb{E}_{p(\tau|\pi)}[R_t | s_t = s, a_t = a]$$ — quality func. or state-action func.

$$Q^{*}(s,a) = \max_{\pi} Q^{\pi}(s,a)$$

$$R_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots = r_t + \gamma \left(r_{t+1} + \gamma r_{t+2} + \dots \right) =$$

$$= r_t + \gamma R_{t+1}$$

$$V^\pi(s_0) = \mathbb{E}\, R_0 = \mathbb{E}_{a_0, s_1, a_1, \dots} \left(r(s_0, a_0) + \gamma R_1 \right) =$$

$$= \mathbb{E}_{\pi(a_0|s_0)} \left(r(s_0, a_0) + \gamma \mathbb{E}_{p(s_1|s_0, a_0)} \left(\mathbb{E}_{a_1, s_2, a_2 \dots} R_1 \right) \right)$$

$$\Rightarrow \boxed{V^\pi(s) = \mathbb{E}_{\pi(a|s)} \left(r(s,a) + \gamma \mathbb{E}_{p(s'|s,a)} V^\pi(s') \right)} \quad V^\pi(s_1)$$

$$V^*(s_0) = \max_{\pi} V^{\pi}(s_0) =$$

$$= \max_{\pi_0(a_0|s_0), \pi_1(a_1|s_1), \ldots} \mathbb{E}_{\pi_0(a_0|s_0)} \left(r(s_0, a_0) + \gamma \mathbb{E}_{p(s_1|s_0, a_0)} V^{\pi}(s_1) \right) =$$

$$= \left\{ \max(a+b, a+c) = a + \max(b, c) \right\} = \quad \leftarrow V^*(s_1)$$

$$= \max_{\pi_0(a_0|s_0)} \mathbb{E}_{\pi_0(a_0|s_0)} \left(r(s_0, a_0) + \gamma \mathbb{E}_{p(s_1|s_0, a_0)} \left(\max_{\pi_1(a_1|s_1), \pi_2, \ldots} V^{\pi}(s_1) \right) \right) =$$

$$= \max_{\pi_0} \mathbb{E}_{\pi_0(a_0|s_0)} \left(\left(r(s_0, a_0) + \gamma \mathbb{E}_{p(s_1|s_0, a_0)} V^*(s_1) \right) \right) = y(s_0, a_0)$$

$$\mathbb{E}_{\pi_0(a_0|s_0)} y(s_0, a_0) = \begin{cases} \sum_{a_0} \pi_0(a_0|s_0)\, y(s_0, a_0) \to \max_{\pi_0} \\ \pi_0(a_0|s_0) \geq 0 \quad \forall a_0 \\ \sum_{a_0} \pi_0(a_0|s_0) = 1 \end{cases}$$

$$\pi^{opt}(a_0|s_0) = \mathbb{I}\left[a_0 = \arg\max_{a_0} y(s_0, a_0)\right]$$

$$\boxed{V^*(s) = \max_a \left(r(s,a) + \gamma \mathbb{E}_{p(s'|s,a)} V^*(s') \right)}$$

$$\pi^*(a|s) = \arg\max_a \left(\dots \right)$$

$$Q^{\pi}(s,a) = r(s,a) + \gamma \mathbb{E}_{p(s'|s,a)} \mathbb{E}_{\pi(a'|s')} Q^{\pi}(s',a')$$

$$Q^{*}(s,a) = r(s,a) + \gamma \mathbb{E}_{p(s'|s,a)} \max_{a'} Q^{*}(s',a')$$

$$V^{\pi}(s) = \mathbb{E}_{\pi(a|s)} Q^{\pi}(s,a)$$

$$Q^{\pi}(s,a) = r(s,a) + \gamma \mathbb{E}_{p(s'|s,a)} V^{\pi}(s')$$

$$V^{*}(s) = \max_{a} Q^{*}(s,a)$$

$$Q^{*}(s,a) = r(s,a) + \gamma \mathbb{E}_{p(s'|s,a)} V^{*}(s')$$

Value Iteration

$$V^*(s) = \max_a \left(r(s,a) + \gamma \mathbb{E}_{p(s'|s,a)} V^*(s') \right) \quad \forall s$$

$$V^* = F(V^*)$$

$$V_{new} = F(V_{old}) \quad \text{- fixed-point iteration}$$

$$|F(V) - F(W)| \leq \|V - W\| \quad \forall V, W$$

Init. V

For iterations:

$\Delta = 0$

For all s:

$v = V(s)$

$V(s) = \max_a \left(r(s,a) + \gamma \mathbb{E}_{p(s'|s,a)} V(s') \right)$

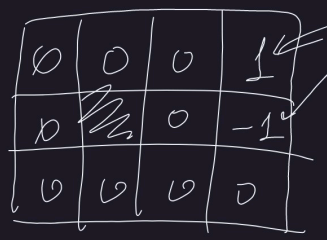
$\Delta = \max(\Delta, |v - V(s)|)$

if $\Delta < \varepsilon$, then break

Inference:

$\pi(a|s) = \arg\max_a \left(r(s,a) + \gamma \mathbb{E}_{p(s'|s,a)} V(s') \right)$

reward:

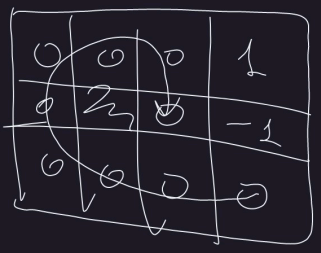


terminal state

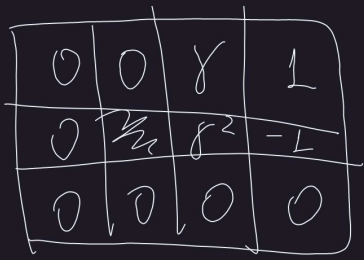
S – position

$a \in \{\leftarrow, \rightarrow, \uparrow, \downarrow\}$

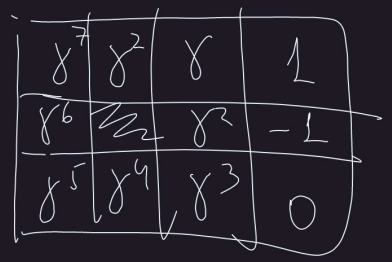
V_0:



$\rightarrow V_1$:

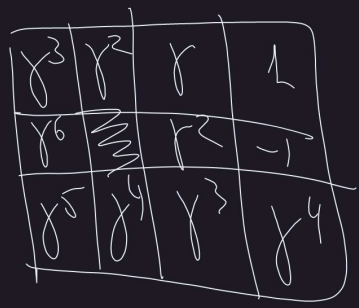


$\rightarrow V_2$:



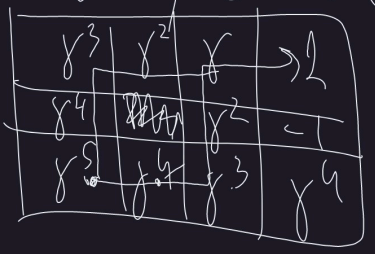
$\rightarrow$

$\rightarrow V_3$:



$\max(0 + \gamma \cdot 0,\ 0 + \gamma(-1)) = 0$

$\rightarrow V_4$:



$= V_3$

Q-learning

$$\forall s,a \quad Q^*(s,a) = r(s,a) + \gamma \mathbb{E}_{p(s'|s,a)} \max_{a'} Q^*(s',a')$$

$$\downarrow$$

$$F(Q^*) = \frac{1}{|S| \cdot |A|} \sum_{s,a} \frac{1}{2} \left(Q^*(s,a) - \left(r(s,a) + \gamma \mathbb{E}_{p(s'|s,a)} \max_{a'} Q^*(s',a') \right) \right)^2$$

$$s, a, r, s' \quad \left(Q_{new}(s,a) = Q_{old}(s,a) - \alpha \left(Q_{old}(s,a) - \right. \right. \quad \rightarrow \min_{Q^*}$$

$$\left. \left. y(s,a) = r + \gamma \max_{a'} Q_{old}(s',a') \right) - y(s,a) \right) = (1-\alpha) Q_{old}(s,a) + \alpha y(s,a)$$

Exploration – Exploitation Dilemma

$$a \sim \varepsilon\text{-greedy}(Q(s,a)): \begin{cases} \text{random action with } \frac{\varepsilon}{|A|} \\ \underset{a}{\operatorname{argmax}}\, Q(s,a) \text{ with prob. } 1-\varepsilon \end{cases}$$

$$a \sim \text{Softmax}\left(Q(s,a)/\tau\right) = \frac{\exp\left(\frac{Q(s,a)}{\tau}\right)}{\sum_a \exp\left(\frac{Q(s,a)}{\tau}\right)}$$

$\tau \to +\infty$ Softmax $\simeq$ Uniform

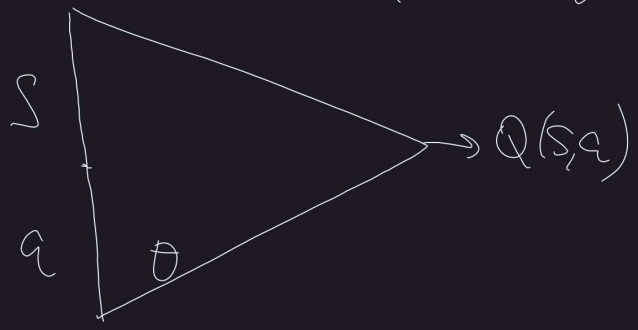
$\tau \to 0$ Softmax $\simeq \underset{a}{\operatorname{argmax}}\, Q(s,a)$

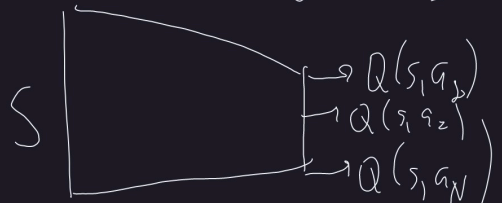
$$Q(s,a|\theta)$$

$$F(\theta) = \mathbb{E}_{s,a} \frac{1}{2}\left(Q(s,a|\theta) - \left(r(s,a) + \gamma \mathbb{E}_{p(s'|s,a)} \max_{a'} Q(s',a'|\theta)\right)\right)^2$$

$\sim s, a, r, s'$; $y(s,a) = r + \gamma \max_{a'} Q(s',a'|\theta)$ $\rightarrow \min_{\theta}$

$$\theta_{new} = \theta_{old} - \alpha\left(Q(s,a|\theta_{old}) - y(s,a)\right) \nabla_{\theta} Q(s,a|\theta)$$





DQN (Deep Q Network)

Inference:

$$\pi(a|s) = \arg\max_a Q(s,a|\theta)$$

Init. θ, buffer $M = \emptyset$ ← experience replay buffer

For episodes:

Init. s_0

For $t = 0, 1, 2, \ldots, T$:

$a_t \sim \varepsilon\text{-greedy}(Q(s_t, a|\theta))$

get r_t, s_{t+1}; $(s_t, a_t, r_t, s_{t+1}) \rightarrow M$

sample mini-batch $\{s_j, a_j, r_j, s'_j\}_{j=1}^{B}$ from M

$L_\theta := 0$

for $j = 1, \ldots, B$:

$y_j = r_j + \gamma \max_{a'} Q(s_j, a'|\bar{\theta})$ ← different from θ

$L_\theta = L_\theta + \frac{1}{B} \frac{1}{2} (Q(s_j, a_j|\theta) - y_j)^2$

$\theta = \text{opt\_step}(\nabla_\theta L_\theta)$

$\bar{\theta}$ - target network

$\bar{\theta} = \beta\bar{\theta} + (1-\beta)\theta$

Overestimation bias