① Policy-gradient

Direct RL / Approximate RL

$\pi_* = ?$

$s \to \pi(a|s)$

θ

$s, a \to h(a,s,\bar\theta) \to \pi(a|s,\bar\theta) = \dfrac{e^{h(a,s,\bar\theta)}}{\sum_{a'} e^{h(a',s,\bar\theta)}}$

1) параметризуем $\pi(a|s) \Rightarrow$ можем обучать стохаст. π

$Q(s,a)$ $\quad s: Q(s,a_1), Q(s,a_2), \dots, Q(s,a_k)$

2) π может состоять детерминир.

3) π – более простой объект, чем Q

s_0 – начальное сост. / $\gamma = 1$, $r = 0$

$$J(\bar\theta) = V_{\pi_{\bar\theta}}(s_0) = \mathbb{E}_{\pi_\theta}[G_t \mid S_t = s_0] \xrightarrow{\bar\theta} \max$$

$$= \mathbb{E}_{Traj(\pi_\theta)}[\dots]$$

$$\bar\theta := \bar\theta + \alpha \cdot \nabla_{\bar\theta} J(\bar\theta)$$

$$\nabla_{\bar\theta} J(\bar\theta) = \nabla_{\bar\theta} V_{\pi_\theta}(s) = \nabla_{\bar\theta}\left[\sum_a \pi_{\bar\theta}(a|s)\, Q_{\pi_{\bar\theta}}(s,a)\right] =$$

$$= \sum_a \left(\nabla_{\bar\theta}\pi_{\bar\theta}(a|s) \cdot Q_{\pi_\theta}(s,a) + \pi_\theta(a|s) \cdot \nabla_{\bar\theta} Q_{\pi_{\bar\theta}}(s,a) \right) =$$

$$= \sum_a \left[Q(s,a) \cdot \nabla_{\bar\theta}\pi(a|s) + \pi(a|s)\, \nabla_{\bar\theta}\left(\sum_{r,s'} p(r,s'|s,a)\left(\underbrace{r}_{=0} + V_{\pi_\theta}(s')\right)\right)\right] =$$

$$= \sum_a \left[Q(s,a)\nabla_{\bar\theta}\pi(a|s) + \pi(a|s) \sum_{r,s'} p(r,s'|s,a)\left(\nabla_{\bar\theta} V_{\pi_\theta}(s')\right)\right] =$$

$$\sum_{a'}\left(\nabla_{\bar\theta}\pi_{\bar\theta}(a'|s')\, Q_{\pi_\theta}(s',a') + \pi(a'|s')\sum_{r',s''} \dots\right)$$

$$= \sum_{s'}\left[\left(\sum_a Q(s,a)\cdot\nabla_{\bar\theta}\pi(a|s')\right) \cdot \sum_{t=0}^{\infty} P_\sigma\left[\begin{array}{c}\text{из } s \text{ в } s' \\ \text{ровно } t \\ \text{за } k \text{ шагов}\end{array}\right]\right]$$

$$\mathbb{E}[\#t : S_t = s' \mid s] \propto \mu(s')$$

$$\nabla_{\bar\theta} J(\bar\theta) \propto \sum_s \mu(s) \cdot \left(\sum_a Q_{\pi_{\bar\theta}}(s,a) \cdot \nabla_{\bar\theta} \pi_{\bar\theta}(a|s) \right)$$

policy gradient theorem

1) All-actions

$$\bar\theta := \bar\theta + \alpha \sum_a \hat{Q}(s,a|\bar w) \cdot \nabla_{\bar\theta} \pi_{\bar\theta}(a|s)$$

Actor-critic methods — $\pi_{\bar\theta}$ (actor), $\hat{V}_{\bar w}$, $\hat{Q}_{\bar w}$ (critic)

2) REINFORCE (Williams, 1992)

$$\nabla_{\bar\theta} J(\bar\theta) \propto \mathbb{E}_{\pi_{\bar\theta}} \left[\sum_a Q_{\pi_\theta}(s,a) \nabla_{\bar\theta} \pi(a|s) \right] =$$

$$= \mathbb{E}_{\pi_\theta} \left[\sum_a \pi_\theta(a|s) \cdot Q_{\pi_\theta}(s,a) \frac{\nabla_\theta \pi_\theta(a|s)}{\pi_\theta(a|s)} \right] =$$

$\mathbb{E}_{\pi_\theta}$

$$= \mathbb{E}_{\pi_\theta} \left[Q_{\pi_\theta}(s,a) \cdot \nabla_{\bar\theta} \log \pi_{\bar\theta}(a|s) \right] = \mathbb{E}_{\pi_{\bar\theta}} \left[G_t \cdot \nabla_{\bar\theta} \log \pi_\theta(A_t|S_t) \right]$$

$$= \mathbb{E}_{\pi_\theta} [G_t | S_t = s, A_t = a]$$

REINFORCE

$$\bar\theta := \bar\theta + \alpha \cdot G_t \cdot \nabla_{\bar\theta} \log \pi_\theta(A_t|S_t)$$

MC policy gradient

(2) Baselines & actor-critic algorithms

$$\nabla_{\bar\theta} J(\bar\theta) \propto \sum_s \mu(s) \sum_a \left(Q_{\pi_{\bar\theta}}(s,a) - b(s) \right) \nabla_\theta \pi_\theta(a|s)$$

baseline

$$\sum_a \left(Q_\pi(s,a) - b(s) \right) \nabla_\theta \pi_\theta(a|s) = \sum(\_) - b(s) \cdot \sum_a \nabla_\theta \pi_\theta(a|s)$$

$$= \sum(\_) - b(s) \cdot \nabla_\theta \underbrace{\left(\sum_a \pi_\theta(a|s) \right)}_{1} = \sum(\_)$$

$$\bar{\theta} := \bar{\theta} + \alpha \cdot \underbrace{\left(G_t - b(S_t)\right)}_{\text{advantage}} \nabla_{\bar{\theta}} \log \pi_{\bar{\theta}}(A_t \mid S_t)$$

$$\approx V(S_t) \approx \hat{V}_{\bar{w}}(S_t)$$

REINFORCE w/ baseline:

— gen. ep. from $\pi_{\bar{\theta}}$

— for t ...

- $G_t := \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$
- $\bar{w} := \bar{w} + \alpha_w \cdot \left(G_t - \hat{V}_{\bar{w}}(S_t)\right) \cdot \nabla_{\bar{w}} \hat{V}_{\bar{w}}(S_t)$
- $\bar{\theta} := \bar{\theta} + \alpha_\theta \cdot \gamma^t \left(G_t - \hat{V}_{\bar{w}}(S_t)\right) \cdot \nabla_{\bar{\theta}} \log \pi_{\bar{\theta}}(A_t \mid S_t)$

(δ)

$$\bar{\theta} := \bar{\theta} + \alpha_\theta \left(R_{t+1} + \gamma \cdot \hat{V}_{\bar{w}}(S_{t+1}) - \hat{V}_{\bar{w}}(S_t)\right) \nabla \log \pi$$

(δ)

One-step Actor-Critic:

S, A, R, S'

$$\delta := R + \gamma \hat{V}_{\bar{w}}(S') - \hat{V}_{\bar{w}}(S)$$