Policy_gradient

~~$V(s)$, $Q(s,a)$ value functions~~ $s \to \boxed{\hat{Q}} = \hat{Q}(s,a)$

$s \to \boxed{\pi}(\bar{\theta}) = \pi(a|s,\theta) = Pr[A_t = a \mid S_t = s]$

$\pi(a|s,\theta)$, $V(s|\theta)$ } actor-critic methods

① M.S. пример модель π

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

$h(s,a,\bar{\theta}) = \bar{\theta}^T \bar{x}(s,a)$

② Можно сделать $\pi(a|s)$ – нейрон. сеть

③ $\bar{\theta}_{t+1} = \bar{\theta}_t + \alpha \nabla J(\bar{\theta}_t)$

$J(\bar{\theta}) = V_{\pi_{\bar{\theta}}}(s_0) \xrightarrow[\bar{\theta}]{} \max$

Policy_gradient theorem

$$\nabla_\theta V_\pi(s) = \nabla_\theta \left[\sum_a \pi(a|s) Q_\pi(s,a) \right] =$$

$$= \sum_a \left[(\nabla_\theta \pi(a|s)) \cdot Q_\pi(s,a) + \pi(a|s) \cdot \nabla_\theta Q_\pi(s,a) \right] =$$

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

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

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

$$= \sum_{s'} \sum_{k=0}^{\infty} \Pr[\text{за } k \text{ шагов из } s \to s' \text{ по ср. } \pi] \cdot \sum_a Q_\pi(s',a) \nabla_\theta \pi(a|s')$$

$$\nabla_\theta J(\theta) = \nabla_\theta V_\pi(s_0) = \sum_s \left(\sum_{k=0}^{\infty} \Pr[s_0 \to s \text{ за } k] \right) \left(\sum_a Q_\pi(s,a) \nabla_\theta \pi(a|s) \right) =$$

$$= \mathbb{E}[\#\text{ попаданий в } s]$$

$$\mu(s) = \Pr[s]$$

$$= \sum_s \mathbb{E}[\#\text{ попад. в } s] \sum_a (\dots)$$

$$\boxed{\nabla_\theta J(\theta) \propto \sum_s \mu(s) \sum_a Q_\pi(s,a) \nabla_\theta \pi(a|s)}$$

policy gradient theorem

① **All-actions method** (actor-critic)

$$\theta_{t+1} := \theta_t + \alpha \sum_a \hat{Q}(s,a|\bar{w}) \nabla_\theta \pi(a|s,\bar{\theta})$$

② **REINFORCE** (Williams, 1992)

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

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

$$= \mathbb{E}_\pi \left[Q_\pi(S_t,A_t) \cdot \frac{\nabla_\theta \pi(A_t|S_t)}{\pi(A_t|S_t)} \right] =$$

$$= \mathbb{E}_\pi[G_t|S_t,A_t]$$

$$= \mathbb{E}_\pi \left[G_t \cdot \frac{\nabla_\theta \pi(A_t|S_t)}{\pi(A_t|S_t)} \right]$$

$$\theta_{t+1} := \theta_t + \alpha G_t \frac{\nabla_\theta \pi(A_t|S_t,\theta_t)}{\pi(A_t|S_t,\theta_t)}$$

$$\boxed{\theta_{t+1} := \theta_t + \alpha G_t \cdot \nabla_\theta \log \pi(A_t|S_t,\theta_t)}$$

(3) REINFORCE with baseline

$$\nabla J(\theta) \propto \sum_s \mu(s) \sum_a \underbrace{Q_\pi(s,a)}_{-b(s)} \nabla \pi(a|s,\theta)$$

$$\sum_s \mu(s) \sum_a \left(Q_\pi(s,a) - b(s)\right) \nabla_\theta \pi(a|s,\theta) =$$

$$= \sum_s \mu(s) \sum_a Q_\pi(s,a) \nabla_\theta \pi - \sum_s \mu(s) b(s) \nabla_\theta \left[\sum_a \pi(a|s,\theta)\right]$$

($\sum_a \pi(a|s,\theta) = 1$, $\nabla_\theta[1] = 0$, $b(s) \approx V(s)$)

$$\nabla J(\theta) \propto \sum_s \mu(s) \sum_a \left(Q_\pi(s,a) - \underbrace{b(s)}_{\text{baseline}}\right) \nabla \pi(a|s,\theta)$$

REINFORCE w/baseline:

$$\theta_{t+1} = \theta_t + \alpha \left(G_t - \underbrace{b(S_t)}_{\hat{V}(S_t|\bar{w})}\right) \nabla_\theta \log \pi(A_t|S_t,\theta_t)$$

- $\alpha_\theta, \alpha_w, \bar{\theta}, \bar{w}$
- loop:
  - gen. episode $S_0, A_0, R_1, S_1, A_1 \dots S_{T-1}, A_{T-1}, R_T$ no π, $G=0$
  - for $t = T-1, \dots, 1, 0$:
    - $G := R_t + \gamma G$
    - $\bar{w} := \bar{w} + \alpha_w \cdot (G - \hat{V}(S_t|\bar{w})) \nabla_{\bar{w}} \hat{V}(S_t|\bar{w})$
    - $\bar{\theta} := \bar{\theta} + \alpha_\theta (G - \hat{V}(S_t|\bar{w})) \cdot \nabla_\theta \log \pi(A_t|S_t,\theta)$

One-step actor-critic

$$\bar{\theta}_{t+1} = \bar{\theta}_t + \alpha_\theta \left(R_{t+1} + \gamma \hat{v}(S_{t+1}|\bar{w}) - \hat{v}(S_t|\bar{w}) \right) \nabla_\theta \log_\pi (A_t | S_t, \bar{\theta})$$

④ Continuous actions

$a \in \mathbb{R}$

$$\pi(a|s,\theta) = \frac{1}{\sqrt{2\pi} \cdot \sigma(s|\bar{\theta})} \cdot e^{-\frac{1}{2\sigma(s|\theta)^2}(a - \mu(s|\theta))^2}$$

$\mu(s|\theta) = \bar{\theta}_\mu^T \bar{x}(s)$

$\sigma(s|\bar{\theta}) = e^{\bar{\theta}_\sigma^T \bar{x}(s)}$

$$\nabla_{\bar{\theta}_\mu, \bar{\theta}_\sigma} \log \pi(a|s,\theta) = \text{упр}$$