Chapter 13

Policy gradient methods :- Action without estimating value function

Parameterized policy :

$$\pi(a|s,\theta) = \Pr\{A_t = a \mid S_t = s, \theta_t = \theta\}$$

Notation value function $\hat{v}(s, w)$

Gradient Ascent, $\theta_{t+1} = \theta_t + \alpha \widehat{\nabla J(\theta_t)}$ → policy gradient method.

method learns both value & policy → Actor critic Algorithm.

$$\pi(a|s,\theta) = \frac{e^{h(s,a,\theta)}}{\sum_a e^{h(s,a,\theta)}}$$ // $h(s,a,\theta)$: can be function of $(s,a),\theta)$

maybe $\approx \theta^T x(s,a)$

Two advantages :- ① Each deterministic policy ② Stochastic value consideration.

Policy gradient theorem :

$$J(\theta) = v_{\pi_\theta}(s_0)$$

$$\nabla J(\theta) \propto \sum_s \mu(s) \sum_a q_\pi(s,a) \nabla \pi(a|s,\theta)$$

($\mu(s)$: distribution of staying at state s.)

$$= E_\pi \left[\sum_a q_\pi(S_t, a) \nabla \pi(a|S_t,\theta) \right]$$ (look)

Reinforce :

$$\theta_{t+1} = \theta_t + \alpha \sum_a q_\pi(S_t, a, w) \nabla \pi(a|S_t,\theta)$$

$$\nabla J(\theta) = E_\pi \left[\sum_a \pi(a|S_t,\theta)\, q_\pi(S_t,a) \frac{\nabla \pi(a|S_t,\theta)}{\pi(a|S_t,\theta)} \right]$$

$$= E_\pi \left[q_\pi(S_t, A_t) \frac{\nabla \pi(A_t|S_t,\theta)}{\pi(A_t|S_t,\theta)} \right]$$

$$= E_\pi \left[G_t \frac{\nabla \pi(A_t|S_t,\theta)}{\pi(A_t|S_t,\theta)} \right]$$

$A_t \sim \pi$ (sampling)

$$\theta_{t+1} = \theta_t + \alpha G_t \frac{\nabla \pi(a_t|s_t,\theta)}{\pi(a_t|s_t,\theta)}$$

$$= \theta_t + \alpha G_t \nabla \ln [\pi(a_t|s_t,\theta)]$$

Reinforce : monte carlo Policy - Gradient Control ~~policy~~ (episodic) π_*

Input : $\pi(a|s,\theta)$

$\alpha > 0$

$\theta \in \mathbb{R}^d$

Loop forever :

Generate $S_0, A_0, R_1, \ldots S_{T-1}, A_{T-1}, R_T$ following $\pi(\cdot|\cdot,\theta)$

Loop for each episode $t = 0, 1 \ldots T-1$

$$G \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$$

$$\theta \leftarrow \theta + \alpha \gamma^t G \nabla \ln \pi(A_t|S_t,\theta)$$

$$\nabla \ln \pi(a|s,\theta) = X(s,a) - \sum_b \pi(b|s,\theta) X(s,b)$$

Reinforce with baseline :

$$\theta_{t+1} = \theta_t + \alpha (G_t - b(S_t)) \frac{\nabla \pi(A_t|S_t,\theta_t)}{\pi(A_t|S_t,\theta_t)}$$

Algorithm :

Input : $\pi(a|s,\theta)$, $\hat{v}(s,w)$, $\alpha^\theta > 0$, $\alpha^w > 0$, $\theta \in \mathbb{R}^d$, $w \in \mathbb{R}^d$

Loop forever (for each episode) :

Generate $S_0, A_0, R_1, \ldots S_{T-1}, A_{T-1}, R_T$ following $\pi(\cdot|\cdot,\theta)$

Loop for each step of episode : $t = 0, 1 \ldots T-1$

$$G \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$$

$$\delta \leftarrow G - \hat{v}(S_t, w)$$

$$w \leftarrow w + \alpha^w \delta \nabla \hat{v}(S_t, w)$$

$$\theta \leftarrow \theta + \alpha^\theta \gamma^t \delta \nabla \ln \pi(A_t|S_t,\theta)$$

Actor-critique method:

$$\theta_{t+1} = \theta_t + \alpha \left(G_{t:t+1} - \hat{v}(s_t, \underline{w}) \right) \frac{\nabla \pi(A_t | s_t, \underline{\theta})}{\pi(A_t | s_t, \underline{\theta})}$$

$$= \theta_t + \alpha \left(R_{t+1} + \gamma \hat{v}(s_{t+1}, \underline{w}) - \hat{v}(s_t, \underline{w}) \right) \frac{\nabla \pi(A_t | s_t, \underline{\theta})}{\pi(A_t | s_t, \underline{\theta})}$$

$$= \theta_t + \alpha \delta_t \frac{\nabla \pi(A_t | s_t, \underline{\theta})}{\pi(A_t | s_t, \underline{\theta})}$$

One step Acton-critic:

input: $\pi(a|s,\underline{\theta})$, $\hat{v}(s, \underline{w})$, $\alpha^{\theta} > 0$, $\alpha^{w} > 0$

param: $\underline{\theta} \in \mathbb{R}^d$, $\underline{w} \in \mathbb{R}^d$

Loop forever:

init s

$I \leftarrow 1$

Loop while s is not terminal:

$A \sim \pi(\cdot|s,\underline{\theta})$

Take action A, observe s', R

$\delta \leftarrow R + \gamma \hat{v}(s', \underline{w}) - \hat{v}(s, \underline{w})$ // if terminal then $\hat{v}(s', \underline{w}) = 0$

$\underline{w} \leftarrow \underline{w} + \alpha^{w} \delta \nabla \hat{v}(s, \underline{w})$

$\underline{\theta} \leftarrow \underline{\theta} + \alpha^{\theta} I \delta \nabla \ln \pi(A|s,\underline{\theta})$

$I \leftarrow \gamma I$

$s \leftarrow s'$

eligibility trace (episode)

Loop forever:

init s, $z^{\theta} \leftarrow 0$, $z^{w} \leftarrow 0$, $I \leftarrow 1$

Loop while s not terminal

$A \sim \pi(\cdot|s,\underline{\theta})$ → observe s', R, find δ

$z^{w} \leftarrow \gamma \lambda^{w} z^{w} + \nabla \hat{v}(s, w)$

$z^{\theta} \leftarrow \gamma \lambda^{\theta} z^{\theta} + I \nabla \ln \pi(A|s,\theta)$

$w \leftarrow w + \alpha^{w} \delta z^{\theta}$

$\theta \leftarrow \theta + \alpha^{\theta} \delta z^{\theta}$

$I \leftarrow \gamma I$

$s \leftarrow s'$

actor critique eligibility trace continuing

input : $\pi(a|s,\theta)$

input $\hat{v}(s,\underline{w})$

$\lambda^w \in [0,1]$, $\lambda^\theta \in [0,1)$, $\alpha^w > 0$, $\alpha^\theta > 0$ $\alpha^{\bar{R}} > 0$

init $\bar{R} \in \mathbb{R}$

init $\underline{w} \in \mathbb{R}^d$, $\theta \in \mathbb{R}^{d'}$, $S \in \mathcal{S}$

$\underline{z}^w \leftarrow 0$ // eligibility trace vector

$\underline{z}^\theta \leftarrow 0$

Loop forever:

$A \sim \pi(\cdot|S,\theta)$: take action

observe S', R

$\delta \leftarrow R - \bar{R} + \hat{v}(S',\underline{w}) - \hat{v}(S,\underline{w})$

$\bar{R} \leftarrow \bar{R} + \alpha^{\bar{R}}\delta$

$\underline{z}^w \leftarrow \lambda^w \underline{z}^w + \nabla \hat{v}(S,\underline{w})$

$\underline{z}^\theta \leftarrow \lambda^\theta \underline{z}^\theta + \nabla \ln \pi(A|S,\theta)$

$\theta \leftarrow \theta + \alpha^\theta \delta \underline{z}^\theta$

$\underline{w} \leftarrow \underline{w} + \alpha^w \delta \underline{z}^w$

$S \leftarrow S'$