Chapter 10. Actor-Critic Methods.

actor: policy update.

critic: policy evaluation.

① simple AC.

② A2C

③ off-policy AC.

④ DPG.

1. The simplest A-C

- reinforce (MC)
- Temporal-difference learning

At time step t:

generate a_t following $\pi(a|s_t, \theta_t)$ observe r_{t+1}, s_{t+1}

and then generate a_{t+1} following $\pi(a|s_{t+1}, \theta_t)$

critics:

$$w_{t+1} = w_t + \alpha_w [r_{t+1} + \gamma q(s_{t+1}, a_{t+1}, w_t) - q(s_t, a_t, w_t)] \nabla_w q(s_t, a_t, w_t)$$

Actor:

$$\theta_{t+1} = \theta_t + \alpha_\theta \nabla_\theta \ln \pi(a_t|s_t, \theta_t)\, q(s_t, a_t, w_{t+1})$$

policy gradient is on policy

2. Advantage A2C

The gradient is $\nabla_\theta J(\theta) = E[X]$

where: $X(S,A) = \nabla_\theta \ln \pi(A|S,\theta_t)[q(S,A) - b(S)]$

$E[X]$ is invariant to $b(S)$

$\triangle$ $\mathrm{var}[X]$ is not invariant to $b(S)$

$tr[\mathrm{var}(X)] = E[X^T X] - \bar{x}^T \bar{x}$

$\bar{x} = E(X)$

we hope $X(S,A)$ $\mathrm{var}(X)$ smallest.

the optimal baseline:

$$b^*(S) = \frac{E_{A\sim\pi}[\|\nabla_\theta \ln \pi(A|S,\theta_t)\|^2 q(S,A)]}{E_{A\sim\pi}[\|\nabla_\theta \ln \pi(A|S,\theta_t)\|^2]}$$

$$b^*(S) \approx b(S) = E_{A\sim\pi}[q(S,A)] = V_\pi(S)$$

the gradient-ascent algorithm is:

$$\theta_{t+1} = \theta_t + \alpha E[\nabla_\theta \ln \pi(A|S,\theta_t)\, \delta_\pi(S,A)]$$

advantage function: $\delta_\pi(S,A) = q_\pi(S,A) - V_\pi(S)$

Stochastic version:

$$\theta_{t+1} = \theta_t + \alpha \nabla_\theta \ln \pi(a_t | s_t, \theta_t)\, \delta_t(s_t, a_t)$$

$$\doteq \theta_t + \alpha \underbrace{\frac{\delta_t(s_t, a_t)}{\pi(a_t | s_t, \theta_t)}}_{\text{Step size}} \nabla_\theta \pi(a_t | s_t, \theta_t)$$

$$\delta_t = q_t(s_t, a_t) - V_t(s_t) \longrightarrow = r_{t+1} + \gamma V_t(s_{t+1}) - V_t(s_t)$$

q_t, V_t: 2 networks

V_t: one network.

A2C / TD actor-critic:

At step time t:

generate a_t following $\pi(a | s_t, \theta_t)$ and then observe r_{t+1}, s_{t+1}

TD error:

$$\delta_t = r_{t+1} + \gamma V(s_{t+1}, w_t) - V(s_t, w_t)$$

Critic:

$$w_{t+1} = w_t + \alpha_w \delta_t \nabla_w V(s_t, w_t)$$

Actor:

$$\theta_{t+1} = \theta_t + \alpha_\theta \delta_t \nabla_\theta \ln \pi(a_t | s_t, \theta_t)$$

3. off policy actor-critic

Introduction: importance sampling

$$E_{x\sim p_0}[X] = \sum_x p_0(x)\,x = \sum_x p_1(x)\,\underbrace{\frac{p_0(x)}{p_1(x)}}_{f(x)}\,x = E_{x\sim p_1}[f(x)]$$

$$\bar{f} = \frac{1}{n}\sum_{i=1}^{n} f(x_i) \qquad \frac{p_0(x_i)}{p_1(x_i)} \text{ is called importance weight}$$

The theorem of off-policy policy gradient.

β is the behavior policy generate episodes

$$J(\theta) = \sum_{s\in S} d_\beta(s)\, V_\pi(s) =$$

d_β — stationary distribution under policy β.

$$\nabla_\theta J(\theta) = E_{s\sim\rho,A\sim\beta}\left[\frac{\pi(A|S,\theta)}{\beta(A|S)}\nabla_\theta \ln \pi(A|S,\theta)\, \underbrace{q_\pi(S,A)}_{\Downarrow\ q_\pi(S,A)-V_\pi(S)}\right]$$

$$\theta_{t+1} = \theta_t + \alpha_\theta \frac{\pi(a_t|s_t,\theta_t)}{\beta(a_t|s_t)} \nabla_\theta \ln \pi(a_t|s_t,\theta_t)\,(q_t(s_t,a_t) - V_t(s_t))$$

$$\theta_{t+1} = \theta_t + \alpha_\theta \left(\frac{\delta_t(s_t, a_t)}{\beta(a_t|s_t)} \right) \nabla_\theta \pi(a_t|s_t, \theta_t)$$

off-policy sod

given policy $\beta(a|s)$ target policy: $\pi(a|s, \theta_0)$

A value function $V(s, w_0)$ where w_0 is the initial parameter vector

① generate a_t following $\beta(s_t)$

observe r_{t+1}, s_{t+1}

② TD error

$$\delta_t = r_{t+1} + \gamma V(s_{t+1}, w_t) - V(s_t, w_t)$$

③ Critic:

$$w_{t+1} = w_t + \alpha_w \frac{\pi(a_t|s_t, \theta_t)}{\beta(a_t|s_t)} \delta_t \nabla_w V(s_t, w_t)$$

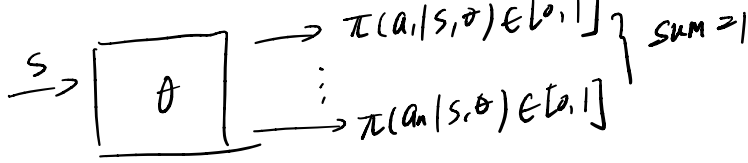
④ Actor:

$$\theta_{t+1} = \theta_t + \alpha_\theta \frac{\pi(a_t|s_t, \theta_t)}{\beta(a_t|s_t)} \delta_t \nabla_\theta \ln \pi(a_t|s_t, \theta_t)$$

4. Deterministic actor-critic (DDPG)

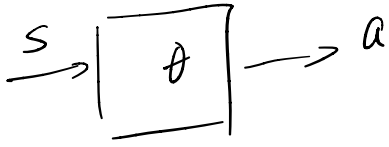
when action is infinite. — deterministic

finite — stochastic



deterministic policy is $a = \mu(s,\theta) \doteq \mu(s)$

$\mu : S \to A$.



$\mu(s,\theta)$ in short $\mu(s)$

① policy gradient:

$$J(\theta) = \sum_{s \in S} d_0(s)\, V_\mu(s)$$

$$\nabla_\theta J(\theta) = E_{s\sim\rho_\mu}\left[\nabla_\theta \mu(s) \left(\nabla_a q_\mu(s,a)\right) \Big|_{a=\mu(s)} \right]$$

$$\theta_{t+1} = \theta_t + \alpha_\theta E_{s\sim\rho_\mu}\left[\nabla_\theta \mu(s)\left(\nabla_a q_w(s,a)\right)\Big|_{a=\mu(s)}\right]$$

$$\theta_{t+1} = \theta_t + \alpha_\theta \nabla_\theta \mu(s_t)\left(\nabla_a q_w(s_t,a)\right)\Big|_{a=\mu(s_t)}$$

behavior policy: $\beta(a|s)$ A deterministic target policy: $\mu(s,\theta_0)$ A value function: $V(s,w_0)$

TD error:

$$\delta_t = r_{t+1} + \gamma q(s_{t+1}, \mu(s_{t+1},\theta_t), w_t) - q(s_t,a_t,w_t)$$

critic:

$$w_{t+1} = w_t + \alpha_w \delta_t \nabla_w q(s_t,a_t,w_0)$$

Actor:

$$\theta_{t+1} = \theta_t + \alpha_\theta \nabla_\theta \mu(s_t,\theta_t)\left(\nabla_a q(s_t,a|w_{t+1})\right)\Big|_{a=\mu(s_t)}$$