Chapter 10 Actor-Critic Method

Q actor-critic, Advantage actor-critic(A2C) / TD actor-critic, Off-policy actor-critic, Deterministic actor-critic

1. Explain the meaning of actor and critic.

Q actor-critic

2. What is the relationship between REINFORCE and QAC?
3. Explain the process of QAC.

Advantage actor-critic(A2C) / TD actor-critic

4. What is the idea of Advantage?
5. How to reduce the estimation variance?
6. Why and How to estimate q_t and v_t? Use 2 methods and explain the difference.
7. Explain the process of A2C.

Off-policy actor-critic

8. What is importance sampling?
9. Explain the process of off-policy actor-critic.

Deterministic actor-critic

10. What is the difference between μ and π?
11. The expression of the gradient of J(θ) in the deterministic situation.
12. The process of deterministic actor-critic.

1. Actor refers to the policy update step because actions are taken following the policy. Critic refers to the value update step because it criticizes the actor by evaluating the value.
2. Both of them are basically policy gradient methods and use gradient-ascent method to do policy update step. However, the way of value update is difference. If used MC method to do value update, then it is REINFORCE. If used TD method (Sarsa) to do value update, then it is QAC.
3. QAC is the simplist actor-critic method.

Algorithm 10.1: The simplest actor-critic algorithm (QAC)

Initialization: A policy function $\pi(a|s,\theta_0)$ where θ_0 is the initial parameter. A value function $q(s,a,w_0)$ where w_0 is the initial parameter. $\alpha_w, \alpha_\theta > 0$.
Goal: Learn an optimal policy to maximize $J(\theta)$.

At time step t in each episode, do

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

Actor (policy update):

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

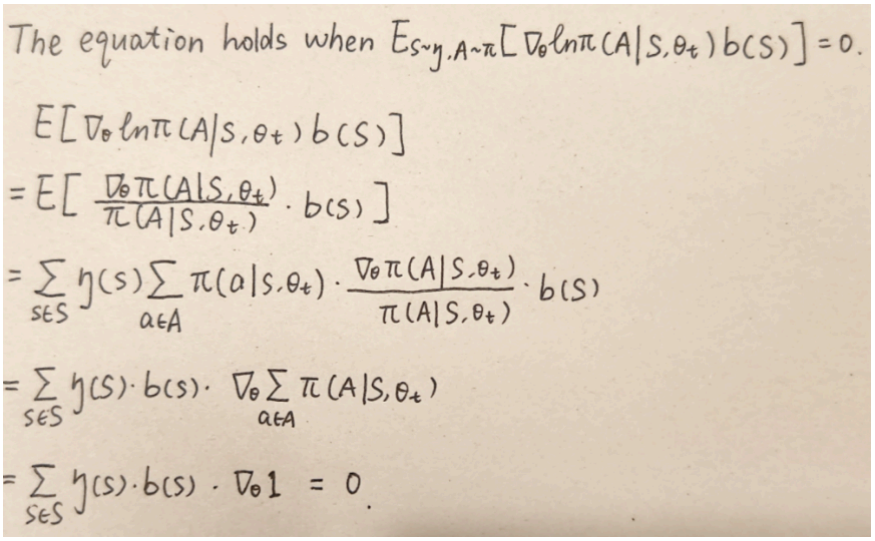
Critic (value update):

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

4. We add an additional baseline to the gradient of metric in policy gradient method to reduce the estimation variance. The lower it is, the more accurate it is to use samples to approximate the gradient.
5. When adding a baseline b(S) to the gradient, it remains unchanged.

$$\mathbb{E}_{S\sim\eta,A\sim\pi}\Big[\nabla_\theta \ln\pi(A|S,\theta_t)q_\pi(S,A)\Big] = \mathbb{E}_{S\sim\eta,A\sim\pi}\Big[\nabla_\theta \ln\pi(A|S,\theta_t)(q_\pi(S,A)-b(S))\Big]$$

The proof is as follows:



Although E[∇J(θ)] is invariant to the baseline, the estimation variance is not. In REINFORCE and QAC, the baseline is 0 and it is not a good baseline to reduce the estimation variance. The optimal baseline to reduce *var*(∇J(θ)) is:

$$b^*(s) = \frac{\mathbb{E}_{A\sim\pi}\left[\|\nabla_\theta \ln\pi(A|s,\theta_t)\|^2 q_\pi(s,A)\right]}{\mathbb{E}_{A\sim\pi}\left[\|\nabla_\theta \ln\pi(A|s,\theta_t)\|^2\right]}$$

which is too complex to be useful. Thus we remove the weight and get a suboptimal baseline:

$$b^\dagger(s) = \mathbb{E}_{A\sim\pi}[q_\pi(s,A)] = v_\pi(s)$$

which coincidentally, is the state value.

Then when b(S) = v_π, the gradient-ascent method becomes:

$$\begin{aligned}\theta_{t+1} &= \theta_t + \alpha\mathbb{E}\Big[\nabla_\theta \ln\pi(A|S,\theta_t)[q_\pi(S,A)-v_\pi(S)]\Big]\\ &\doteq \theta_t + \alpha\mathbb{E}\Big[\nabla_\theta \ln\pi(A|S,\theta_t)\delta_\pi(S,A)\Big].\end{aligned}$$

where $\delta_\pi(S,A)$ is the advantage function. The we use samples to approximate:

$$\begin{aligned}\theta_{t+1} &= \theta_t + \alpha\nabla_\theta \ln\pi(a_t|s_t,\theta_t)[q_t(s_t,a_t)-v_t(s_t)]\\ &= \theta_t + \alpha\nabla_\theta \ln\pi(a_t|s_t,\theta_t)\delta_t(s_t,a_t),\end{aligned}$$

where $\delta_t(s_t, a_t)$ reflects the advantage of a_t over others. The greater it is, the better the action is. This is also one of the differences between A2C and policy gradient in Chapter 9. We replace $q_t(s_t, a_t)$ with $\delta_t(s_t, a_t)$.

6. If we use MC method to estimate $q_t(s_t, a_t)$ and $v_t(s_t)$, then it is called REINFORCE with a baseline. If we use TD method to estimate, then it is called A2C, which is also called TD actor-critic:

$$q_t(s_t, a_t) - v_t(s_t) \approx r_{t+1} + \gamma v_t(s_{t+1}) - v_t(s_t)$$

The reason why we do the estimate is that we only need one neural network to represent $v_\pi(s)$ after doing the estimate. Otherwise, we need two neural networks to represent both $q_\pi(s, a)$ and $v_\pi(s)$.

7.

Algorithm 10.2: Advantage actor-critic (A2C) or TD actor-critic

Initialization: A policy function $\pi(a|s, \theta_0)$ where θ_0 is the initial parameter. A value function $v(s, w_0)$ where w_0 is the initial parameter. $\alpha_w, \alpha_\theta > 0$.
Goal: Learn an optimal policy to maximize $J(\theta)$.

At time step t in each episode, do

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

Advantage (TD error):

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

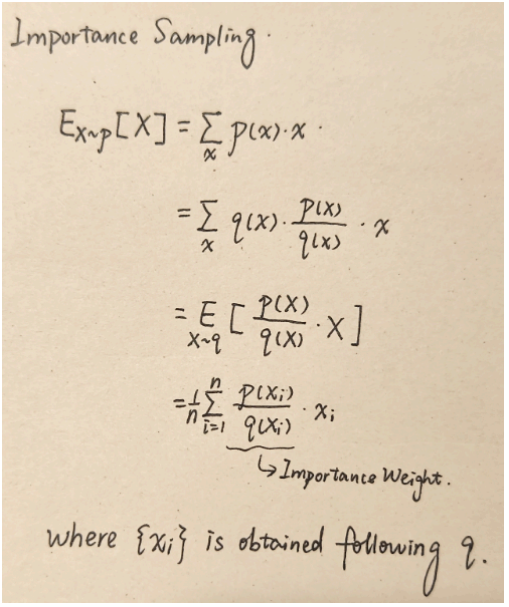
Actor (policy update):

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

Critic (value update):

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

8. Importance sampling can use sufficient samples following distribution q to approximate the expectation following distribution p when p is not easy to obtain enough samples and the expression of p is not available. The derivation is:



In off-policy actor-critic, we use samples of behavior policy to approximate the derivation expectation.

9. All the previous policy gradient methods (REINFORCE, QAC, Q2C) are on-policy. Using inportance sampling can help we utilize the samples generated by another policy β, the behavior policy. The expectation of the metric gradient can be written as:

$$\nabla_\theta J(\theta) = \mathbb{E}\left[\frac{\pi(A|S,\theta)}{\beta(A|S)}\nabla_\theta \ln \pi(A|S,\theta)\big(q_\pi(S,A) - v_\pi(S)\big)\right]$$

Then the gradient-ascent algorithm can be written as:

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

Algorithm 10.3: Off-policy actor-critic based on importance sampling

Initialization: A given behavior policy $\beta(a|s)$. A target policy $\pi(a|s,\theta_0)$ where θ_0 is the initial parameter. A value function $v(s,w_0)$ where w_0 is the initial parameter. $\alpha_w, \alpha_\theta > 0$.
Goal: Learn an optimal policy to maximize $J(\theta)$.

At time step t in each episode, do

Generate a_t following $\beta(s_t)$ and then observe r_{t+1}, s_{t+1}.

Advantage (TD error):

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

Actor (policy update):

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

Critic (value update):

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

10. μ is the deterministic policy, which indicates that at any state, one action is given a probablity of 1 and other actions are given 0. We denote a = μ(s) because μ directly gives an action at a state instead of a probablity space.

11.

$$\begin{aligned}\nabla_\theta J(\theta) &= \sum_{s\in\mathcal{S}} \eta(s)\nabla_\theta \mu(s)\big(\nabla_a q_\mu(s,a)\big)|_{a=\mu(s)} \\ &= \mathbb{E}_{S\sim\eta}\left[\nabla_\theta \mu(S)\big(\nabla_a q_\mu(S,a)\big)|_{a=\mu(S)}\right]\end{aligned}$$

12.

Algorithm 10.4: Deterministic policy gradient or deterministic actor-critic

Initialization: A given behavior policy $\beta(a|s)$. A deterministic target policy $\mu(s,\theta_0)$ where θ_0 is the initial parameter. A value function $q(s,a,w_0)$ where w_0 is the initial parameter. $\alpha_w, \alpha_\theta > 0$.
Goal: Learn an optimal policy to maximize $J(\theta)$.

At time step t in each episode, do

Generate a_t following β and then observe r_{t+1}, s_{t+1}.

TD error:

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

Actor (policy update):

$\theta_{t+1} = \theta_t + \alpha_\theta \nabla_\theta \mu(s_t,\theta_t)\big(\nabla_a q(s_t,a,w_t)\big)|_{a=\mu(s_t)}$

Critic (value update):

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