Policy gradient

1 Problem

Control the cart to prevent the pole on it from from falling.

- Cart weight: $M = 1.0(kg)$
- Pole weight: $m = 0.1(kg)$
- Pole length: $2l = 1(m)$
- Gravity acceration: $g = 9.8(m/sec^2)$
- Force to cart: $a(N)$
- Fliction coefficient of cart: $\mu_c = 0.0005$
- Fliction coefficient of port: $\mu_p = 0.000002$

Cart pole dynamics is as follow:

$$\ddot{y} = \frac{g\sin(y) + \cos(y)(\mu_c \operatorname{sgn}(\dot{x}) - a - ml\dot{y}^2\sin(y))/(M+m) - \mu_p\dot{y}/(ml)}{l(4/3 - m\cos^2(y)/(M+m))}$$
$$\ddot{x} = \frac{a + ml(\dot{y}^2\sin(y) - \ddot{y}\cos(y)) - \mu_c \operatorname{sgn}(\dot{x})}{M+m},$$

where x is cart position, $\dot{x}$ is cart velocity, $\ddot{x}$ is cart acceleration, y is pole angle to verticle, $\dot{y}$ is pole angular velocity, and $\ddot{y}$ is pole angular acceleration.

Let $\tau = 1/60(sec^{-1})$ be time constant. Transition every $1/60(s)$ as follow:

$$x = x + \tau\dot{x}$$
$$\dot{x} = \dot{x} + \tau\ddot{x}$$
$$y = y + \tau\dot{y}$$
$$\dot{y} = \dot{y} + \tau\ddot{y}$$

When $|a| > 20(N)$, the environment ignores the excess.

2 Algorithm

The training episode consists of repeating the initial state selection, behavior selection, reward acquisition, policy gradient update, and state transition until the state does not satisfy the conditions $|x| > 2.4(m), |\dot{x}| > 2(m/sec), |y| > 12\pi/180(rad), |\dot{y}| > 1.5(rad/sec)$. By repeating the training episode, the curt can act so that the pole does not fall.

Let $\boldsymbol{\theta} = [\theta_1, \theta_2, \theta_3, \theta_4]^T, \eta$ be the policy parameter.

The initial state is

$$\begin{aligned} \boldsymbol{s} &= [x, \dot{x}, y, \dot{y}]^T \sim N(\mathbf{0}, \boldsymbol{\Sigma}) \\ \boldsymbol{\Sigma} &= \operatorname{diag}(0.01, 0.01, 0.01, 0.01) \\ [\ddot{x}, \ddot{y}]^T &= \mathbf{0}. \end{aligned}$$

The behavior selection is

$$\begin{aligned} a &\sim \pi(a|\boldsymbol{s}; \boldsymbol{\theta}, \eta) \\ \pi(a|s; \boldsymbol{\theta}, \eta) &= N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(a-\mu)^2}{2\sigma^2}\right), \end{aligned}$$

where

$$\begin{aligned} \mu &= \boldsymbol{\theta}^T \boldsymbol{C} \boldsymbol{s} \\ \boldsymbol{C} &= \operatorname{diag}(1/2.4, 1/2, 180/(12\pi), 1/1.5) \\ \sigma &= 0.1 + \frac{1}{1 + \exp(\eta)}. \end{aligned}$$

The reward r acquisition is

$$r = -\boldsymbol{s}^T \boldsymbol{Q} \boldsymbol{s} - aRa,$$

where

$$\begin{aligned} \boldsymbol{Q} &= \operatorname{diag}(1.25, 1, 12, 0.25) \\ R &= 0.01. \end{aligned}$$

The policy gradient update is

$$\begin{aligned} \boldsymbol{\nabla}_\theta \ln \pi(a|\boldsymbol{s}; \boldsymbol{\theta}, \eta) &= \frac{(a-\mu)\boldsymbol{C}\boldsymbol{s}}{\sigma^2} \\ \nabla_\eta \ln \pi(a|\boldsymbol{s}; \boldsymbol{\theta}, \eta) &= \frac{(\sigma^2 - (a - mu)^2)\exp(\eta)}{\sigma^3(1 + \exp(\eta))^2} \\ \boldsymbol{z} &= \boldsymbol{z} + [\boldsymbol{\nabla}_\theta \ln \pi(a|\boldsymbol{s}; \boldsymbol{\theta}, \eta)^T, \nabla_\eta \ln \pi(a|\boldsymbol{s}; \boldsymbol{\theta}, \eta)]^T \\ \boldsymbol{\delta} &= \boldsymbol{\delta} + \boldsymbol{z} r \gamma^t \\ t &= t + 1. \end{aligned}$$

At the beginning of each training episode, initialize as $\boldsymbol{z} = \mathbf{0}, \boldsymbol{\delta} = \mathbf{0}, t = 0$.

Let Δ be policy gradient. After each training episode,

$$\begin{aligned} \Delta_{past} &= \Delta \\ \Delta &= \frac{n-1}{n}\Delta + \frac{1}{n}\delta \\ n &= n+1 \end{aligned}$$

is calculated. If

$$\angle(\Delta_{past}, \Delta) < \epsilon = 3/1000, \tag{1}$$

the policy is improved as follow:

$$[\boldsymbol{\theta}^T, \eta]^T = [\boldsymbol{\theta}^T, \eta]^T + \alpha\Delta.$$

The initial state is $n = 1$.

The parameters are defined as $\gamma = 0.95, \alpha = 0.1$. The policy parameters $\theta_1, \theta_2, \theta_3, \theta_4$ are initialized by randomly selecting in the range $[-5, 5]$ and η is so in the range $[-1, 1]$.