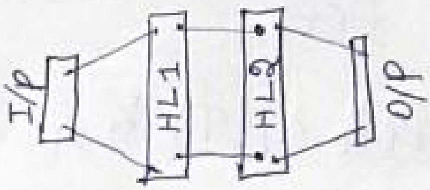
Policy Gradient



Policy Net - Q Network - NN

→ Q-learning → Minimizing the F1 smooth loss. means, it is working for maximizing two consecutive action cumulative reward.

→ Policy Gradient → Sampling the episodes and optimizing the policy net by gradient ascent.

objective Function: $$J(\theta) = E\left[\sum_{t=0}^{T-1} r_{t+1}\right] \quad \longrightarrow ①$$

here, Nomenclature : $\{S_t, a_t, S_{t+1}, r_{t+1}\}$

Gradient Ascent : $$\theta \leftarrow \theta + \frac{\partial}{\partial\theta}\left[J(\theta)\right] \quad ----②$$

As we all know that, $E[f(x)] = \sum_x P(x) f(x)$.

$$J(\theta) = E\left[\sum_{t=0}^{T-1} r_{t+1} \mid \pi_\theta\right]$$

$$J(\theta) = \sum_{t=0}^{T-1} P(s_t, a_t \mid \tau)\, r_{t+1} \quad -----③$$

$i \longrightarrow$ Arbitary Starting Point

$\tau \longrightarrow$ Given Trajectory

Differentiating both sides with respect to policy Parameter θ

using $\frac{d}{dx}[\log f(x)] = \frac{f'(x)}{f(x)}$ ---- ④

$$\nabla_\theta J(\theta) = \sum_{t=i}^{T-1} \nabla_\theta P(S_t, a_t | \tau)\, r_{t+1}$$

$$= \sum_{t=i}^{T-1} P(S_t, a_t | \tau) \cdot \frac{\nabla_\theta P(S_t, a_t | \tau)\, r_{t+1}}{P(S_t, a_t | \tau)}$$

$$= \sum_{t=i}^{T-1} P(S_t, a_t | \tau) \cdot \nabla_\theta \log P(S_t, a_t | \tau) \cdot r_{t+1}$$

so, substituting in the equation ①,

$$\nabla_\theta J(\theta) = E\left[\sum_{t=i}^{T-1} \nabla_\theta \log P(S_t, a_t | \tau)\, r_{t+1} \right]$$

This approximate can be done by re-writing the equation,

$$\nabla_\theta J(\theta) \sim \sum_{t=i}^{T-1} \nabla_\theta \log P(S_t, a_t | \tau)\, r_{t+1} \text{ ---- ⑤}$$

looking expression for $\nabla_\theta \log P(S_t, a_t | \tau)$,

$$P(S_t, a_t | \tau) = P(S_0, a_0, S_1, a_1, \ldots\ldots, S_{t-1}, a_{t-1}, S_t, a_t \mid \pi_\theta)$$

$$= P(S_0)\, \pi_\theta(a_1 | S_0)\, P(S_1 | S_0, a_0)\, \pi_\theta(a_2 | S_1) P(S_2 | S_1, a$$

$$\pi_\theta(a_3 | S_2)\, P(S_3 | S_2, a_2) \ldots\ldots$$

$$\ldots\ldots P(S_{t-1} | S_{t-2}, a_{t-2})\, \pi_\theta(a_{t-1} | S_{t-2})$$

$$\ldots\ldots\ldots P(S_t | S_{t-1}, a_{t-1})\, \pi_\theta(a_t, S$$

here, the function term, $\nabla_\theta \log P(s_t, a_t|\tau)$,

$$\nabla_\theta \log P(s_t, a_t|\tau) = \nabla_\theta \log P(s_0) + \nabla_\theta \log \pi_\theta(a_0|s_0) + \nabla_\theta \log P(s_1|s_0, a_0) + \nabla_\theta \log \pi_\theta(a_1|s_1) + \ldots \ldots + \nabla_\theta \log P(s_{t-1}|s_{t-2}, a_{t-2}) + \nabla_\theta \log \pi_\theta(a_t|s_{t-1}) + \ldots \ldots + \nabla_\theta \log P(s_t, s_{t-1}, a_{t-1}) + \nabla_\theta \log \pi_\theta(a_t|s_{t-1})$$

Here, it is important to note that $P(s_t|s_{t-1}, a_{t-1})$ is not dependant on the policy parameter θ and solely dependant on environment reinforcement learning.

$$\nabla_\theta \log P(s_t, a_t|\tau) = \sum_{t'=0}^{t} \nabla_\theta \log \pi_\theta(a_{t'}|s_{t'})$$

$$\text{from, } \nabla_\theta \log P(s_t, a_t|\tau) = 0 + \nabla_\theta \log \pi_\theta(a_1|s_0) + 0 + \nabla_\theta \log \pi_\theta(a_2|s_1) + 0 + \ldots + \ldots + 0 + \nabla_\theta \log \pi_\theta(a_{t-1}|s_{t-2}) + \ldots + \nabla_\theta \log \pi_\theta(a_t|s_{t-1})$$

so,

$$\nabla_\theta \log P(s_t, a_t|\tau) = \sum_{t'=0}^{t} \nabla_\theta \log \pi_\theta(a_{t'}|s_{t'}) \quad \text{---} (6)$$

from equation (5) & (6),

$$\nabla_\theta J(\theta) = \sum_{t=0}^{T-1} r_{t+1} \left\{ \sum_{t'=0}^{t} \nabla_\theta \log \pi_\theta(a_{t'}|s_{t'}) \right\} \quad \text{---} (7)$$

$$= r_1\left[\sum_{t'=0}^{0} \nabla_\theta \log \pi_\theta(a_{t'}|s_{t'})\right] + r_2\left[\sum_{t'=0}^{1} \nabla_\theta \log \pi_\theta(a_{t'}|s_t\right.$$

$$+ r_3\left[\sum_{t'=0}^{2} \nabla_\theta \log \pi_\theta(a_{t'}|s_{t'})\right] + \cdots\cdots$$

$$\cdots\cdots + r_{T-1}\left[\sum_{t'=0}^{T-1} \nabla_\theta \log \pi_\theta(a_{t'}|s_{t'})\right]$$

$$= r_1 \nabla_\theta \log \pi_\theta(a_0|s_0) + r_2\left[\nabla_\theta \log \pi_\theta(a_0|s_0) + \nabla_\theta \log \pi_\theta(a_1|s_1)\right]$$

$$+ r_3\left[\nabla_\theta \log \pi_\theta(a_0|s_0) + \nabla_\theta \log \pi_\theta(a_1|s_1) + \nabla_\theta \log \pi_\theta(a_2|s_2)\right]$$

$$\cdots\cdots + \nabla_\theta \log \pi_\theta(a_{T-1}|s_{T-1}) \cdot r_T$$

$$= \nabla_\theta \log \pi_\theta(a_0|s_0)\left[r_1 + r_2 + r_3 + \cdots + r_T\right]$$

$$+ \nabla_\theta \log \pi_\theta(a_1|s_1)\left[r_2 + r_3 + \cdots + r_T\right]$$

$$+ \nabla_\theta \log \pi_\theta(a_2|s_2)\left[r_3 + r_4 + \cdots + r_T\right] + \cdots$$

$$\cdots\cdots + \nabla_\theta \log \pi_\theta(a_{T-1}|s_{T-1})\, r_T$$

$$\nabla_\theta J(\theta) = \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t)\left[\sum_{t'=t+1}^{T} r_{t'}\right] \quad \text{---- (8)}$$

Simplifying the term $\sum_{t'=t+1}^{T} r_t'$ to $G_{t'}$.

$$\nabla_\theta J(\theta) = \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t)\, G_t \quad \text{------ (9)}$$

Incorporating the discount factor $\gamma \in [0, 1]$ into our objective.

$$J(\theta) = E\left[\gamma^0 r_1 + \gamma^1 r_2 + \gamma^2 r_3 + \cdots + \gamma^{T-1} r_{T-1} | \pi_\theta\right] \quad \text{------- (10)}$$

We can perform a similar derivation to obtain,

$$\nabla_\theta J(\theta) = \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|S_t) \left[\sum_{t'=t+1}^{T} \gamma^{t'-t-1} r_{t'} \right]$$

and simplifying $\sum_{t'=t+1}^{T} \gamma^{t'-t-1} r_{t'}$ to G_t

$$\boxed{\nabla_\theta J(\theta) = \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|S_t) G_t}$$

Policy update equation

Pseudo Code:

Initialise θ arbitarily

for each episode $\{S_1, a_1, r_2, \ldots, S_{T-1}, a_{T-1}, r_T) - \pi_\theta$ do

for $t = 1$ to $T-1$ do

$\theta \leftarrow \theta + \alpha \nabla_\theta \log \pi_\theta(S_t, a_t) V_t$

end for

end for

return θ