Softmax Policy Gradient

(1.1) $\pi_w(a|s) = \dfrac{e^{w_{s,a}/\tau}}{\sum_{a'} e^{w_{s,a'}/\tau}}$

$$\nabla \log(\pi(a|s)) = \nabla_w \left[w_{s,a}/\tau - \log\left(\textstyle\sum_{a'} e^{w_{s,a'}/\tau}\right)\right]$$

c1: $\nabla_{w_{s,b}} \left[w_{s,a}/\tau - \log\left(e^{w_{s,a}/\tau}\right)\right]$ $\quad b = a$

$= \frac{1}{\tau} - \nabla_{w_{s,b}} \log\left(\sum_{a'} e^{w_{s,a'}/\tau}\right)$

$= 1/\tau - \frac{1}{\sum_{a'} e^{w_{s,a'}/\tau}} \nabla_{w_{s,b}} \left(\sum_{a'} e^{w_{s,a'}/\tau}\right)$

$= 1/\tau - \frac{\nabla_{w_{s,b}} \left(e^{w_{s,a'}/\tau}\right)}{\sum_{a'} e^{w_{s,a'}/\tau}}$

$= \frac{1}{\tau} - \frac{e^{w_{s,a}/\tau}\left(\frac{1}{\tau}\right)}{\sum_{a'} e^{w_{s,a}/\tau}}$

for $\tau = 1$,

$= \boxed{1 - \frac{e^{w_{s,a}}}{\sum_{a'} e^{w_{s,a'}/\tau}}}$

c2: $\nabla_{w_{s,b}} \left[w_{s,a}/\tau - \log\left(\sum_{a'} e^{w_{s,a'}/\tau}\right)\right]$, $b \neq a$.

$= 0 - \nabla_{w_{s,b}} \log\left(\sum_{a'} e^{w_{s,a'}/\tau}\right)$

$= -\frac{1}{\sum_{a'} e^{w_{s,a'}/\tau}} \nabla_{w_{s,b}} \left(\sum_{a'} e^{w_{s,a'}/\tau}\right)$

$= \boxed{-\frac{e^{w_{s,b}/\tau}\left(\frac{1}{\tau}\right)}{\sum_{a'} e^{w_{s,a'}/\tau}}}$

c3: $\nabla_{w_{s',b}} \left[w_{s,a}/\tau - \log\left(\sum_{a'} e^{w_{s,a'}/\tau}\right)\right]$, $s' \neq s$, $b \neq a$.

$= \boxed{0}$.

Linear Policy Gradient

(2.1) $\pi_\theta(a|s) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(a-\theta s)^2}{2\sigma^2}}$

$$\log \pi_\theta(a|s) = \log\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(a-\theta s)^2}{2\sigma^2}}\right)$$

$$= \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) + \log\left(e^{-\frac{(a-\theta s)^2}{2\sigma^2}}\right)$$

$$\nabla_\theta \log \pi_\theta(a|s) = \nabla_\theta \left[\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) + \log\left(e^{-\frac{(a-\theta s)^2}{2\sigma^2}}\right)\right]$$

$$= \cancelto{0}{\nabla_\theta \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right)} + \nabla_\theta\left(-\frac{(a-\theta s)^2}{2\sigma^2}\right)$$

$$= -\frac{1}{2\sigma^2} \nabla_\theta (a-\theta s)^2$$

$$= -\frac{1}{\cancel{2}\sigma^2}(\cancel{2})(a-\theta s)(-s)$$

$$= \frac{(a-\theta s)}{\sigma^2} \cdot s$$, where θ, s are vectors.

Neural Network Gradient Derivation

(2-3) x - input

$z = Wx_i + b_1$

$h = ReLU(z)$

$\theta = Uh + b_2$

$\hat{y} = softmax(\theta)$

$J(\theta) = (CE(y, \hat{y}))(v+1)(x) \rightarrow$

alpha(x) - denotes learning rate

* Derivation does not include sampling from the normal distribution

- Here $J(\theta)$ would denote the error $u(5)$

Here, CE denotes the cross entropy loss function

$\Rightarrow$ We maximize the likelihood the estimator is close to the actual output of i

$CE = \sum_i y_i \log(\hat{y}_i)$

We assume that b_1, b_2 (the biases in our network = 0)

$\theta = Uh$ and $z = Wx_i$

We compute the following gradients:

$$\frac{\partial J}{\partial U}, \frac{\partial J}{\partial W}, \frac{\partial J}{\partial x}$$

$$\frac{\partial J}{\partial U} = \underbrace{\frac{\partial J}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial \theta}} \frac{\partial \theta}{\partial U}$$

$$\frac{\partial J}{\partial \theta} = (\hat{y} - y)(v+1)(x) = \frac{\partial J}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial \theta}$$

$$\frac{\partial \theta}{\partial U} = h$$

$$\boxed{\frac{\partial J}{\partial U} = (\hat{y} - y) h}$$

$ReLU(x) = \max(x, 0)$

$$\frac{\partial ReLU(x)}{\partial x} = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\Rightarrow \frac{\partial ReLU(x)}{\partial x} = sgn(ReLU(x))$$

$$\frac{\partial h}{\partial z} = \frac{\partial ReLU(z)}{\partial z} = sgn(ReLU(z))$$

z is a vector.

$$\frac{\partial J}{\partial W} = \left(\frac{\partial J}{\partial \theta}\right)\left(\frac{\partial \theta}{\partial h}\right)\left(\frac{\partial h}{\partial z}\right)\frac{\partial z}{\partial W}$$

$$\boxed{= (\hat{y} - y)(v+1)(x)(sgn(h))(x)}$$

$$\frac{\partial J}{\partial x} = \frac{\partial J}{\partial \theta} \cdot \frac{\partial \theta}{\partial h} \cdot \frac{\partial h}{\partial z} \cdot \frac{\partial z}{\partial x} = \boxed{(\hat{y} - y)(v+1)(x)(sgn(h))(W)}$$