- $w_i' = w_i$ for $i = 1, 2 \cdots d$

- For $t = 0, 1, \cdots T-1$:

$$G_{t:T} = \sum_{k=t}^{T} (r^k \cdot \gamma^{k-t}) \quad \text{// long term reward}$$

For $i = 1, 2, \cdots d$:

$$w_i' = w_i' + \alpha \cdot (G_{t:T}) \cdot \text{derivative}_i(s^t, a^t, w = (w_1 \cdots w_d))$$

- derivative$_i$(s, a, $w = (w_1, \cdots w_d)$):

// derivative of $\ln \pi_w(s,a)$ wrt w_i:

w_dot_phi = 0.

for $i = 1, 2, \cdots d$:

w_dot_phi += $w_i \cdot \phi_i(s)$

$$\text{grad} = \frac{+1}{(1 + e^{-\text{w\_dot\_phi}})^2} \cdot e^{-\text{w\_dot\_phi}} \cdot \phi_i(s)$$

// equivalently grad = $\pi_w(s,0) \cdot \pi_w(s,1) \cdot \phi_i(s)$

if $a == 0$:

return grad / $\pi_w(s,0)$

if $a == 1$:

return $-$grad / $\pi_w(s,1)$

Sunday 19