Causality Principle

Remember our gradient:

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{i=1}^{N} \left(\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta (a_{i,t} | s_{i,t}) \right) \left(\sum_{t=1}^{T} r(s_{i,t}, a_{i,t}) \right)$$

Remember the example we solved in class with two states and two rewards: s_1, s_2, r_1, r_2 ; actions: a_1, a_2

Now the gradient: (For one sample to write easily)

$$\nabla_\theta J(\theta) \approx \underbrace{\left(\nabla_\theta \log \pi_\theta(a_1|s_1) + \nabla_\theta \log \pi_\theta(a_2|s_2) \right)}_{\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)} \underbrace{\left(r(s_1,a_1) + r(s_2,a_2) \right)}_{\sum_{t=1}^{T} r(s_t,a_t)}$$

Lets distribute:

$$\nabla_\theta J(\theta) \approx \nabla_\theta \log \pi_\theta(a_1|s_1) \cdot \left(r(s_1,a_1) + r(s_2,a_2) \right) + \nabla_\theta \log \pi_\theta(a_2|s_2) \left(r(s_1,a_1) + r(s_2,a_2) \right)$$

Here we should realize that reward from the 1. timestep has no relation with the action we take at the 2. timestep.

So the equation becomes:

$$\nabla_\theta J(\theta) \approx \nabla_\theta \log \pi_\theta(a_1|s_1) \left(r(s_1,a_1) + r(s_2,a_2) \right) + \nabla_\theta \log \pi_\theta(a_2|s_2) \cdot r(s_2,a_2)$$

You can also add the discount factor γ to determine the emphasis of future rewards on the action we take:

$$\nabla_\theta J(\theta) \approx \nabla_\theta \log \pi_\theta(a_1|s_1) \left(r(s_1,a_1) + \gamma\, r(s_2,a_2) \right) + \nabla_\theta \log \pi_\theta(a_2|s_2) \cdot r(s_2,a_2)$$

The general form will become:

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{i=1}^{N} \left[\sum_{t=1}^{T} \left(\nabla_\theta \log \pi_\theta(a_{i,t}|s_{i,t}) \cdot \left(\sum_{k=t}^{T} \gamma^{k-t} r(s_{i,k}, a_{i,k}) \right) \right) \right]$$

Remember to use baseline as well in homework.