Goal: Maximise objective function

$$U(\theta) = E_{\pi_\theta}(R(\tau))$$
$$= \sum_{\tau} P(\tau;\theta) R(\tau)$$

Use gradient ascent to optimise parameters.

$$\theta \leftarrow \theta + \alpha \cdot \nabla_\theta [U(\theta)]$$

$$\nabla_\theta [U(\theta)] = \nabla_\theta [E_{\pi_\theta}(R(\tau))]$$
$$= \nabla_\theta \left[\sum_{\tau} P(\tau;\theta) R(\tau)\right]$$
$$= \sum_{\tau} \nabla_\theta [P(\tau;\theta)] R(\tau)$$

note we have introduced these as they cancel.

$$= \sum_{\tau} P(\tau;\theta) \cdot \frac{\nabla_\theta [P(\tau;\theta)]}{P(\tau;\theta)} R(\tau)$$

$$= \sum_{\tau} P(\tau;\theta) \nabla_\theta [\log(P(\tau;\theta))] R(\tau)$$

This is an expectation! i.e.

$$= E_{\pi_\theta}\left(\nabla_\theta [\log(P(\tau;\theta))] R(\tau)\right)$$

We can use Monte Carlo sampling to approximate this:

$$\simeq \frac{1}{m} \sum_{i=1}^{m} \left(\nabla_\theta [\log(P(\tau^{(i)};\theta))] R(\tau^{(i)})\right)$$

But for each trajectory $\tau^{(i)}$ we sample, how do we calculate

$$\nabla_\theta \left[\log(P(\tau^{(i)};\theta)) \right] R(\tau^{(i)}) \; ?$$

$R(\tau^{(i)})$ is easy - it's just the sum of rewards over ~~m~~ trajectory i.

$$\nabla_\theta \left[\log(P(\tau;\theta)) \right] = \nabla_\theta \Big[\log\big(\underbrace{\pi_\theta(a_0|s_0)}_{\text{prob. of selecting action } a_0 \text{ from state } s_0} \cdot \underbrace{P(s_1|s_0,a_0)}_{\text{Prob. of ending up in state } s_1 \text{ having chosen action } a_0 \text{ from state } s_0.} \times \pi_\theta(a_1|s_1) \cdot P(s_2|s_1,a_1) \times \ldots \Big]$$

$$= \nabla_\theta \left[\log\left(\prod_{t=0}^{T-1} \pi_\theta(a_t|s_t) \cdot P(s_{t+1}|s_t,a_t) \right) \right]$$

$$= \nabla_\theta \left[\sum_{t=0}^{T-1} \left(\log(\pi_\theta(a_t|s_t) \cdot P(s_{t+1}|s_t,a_t)) \right) \right]$$

$$= \nabla_\theta \left[\sum_{t=0}^{T-1} \left(\log(\pi_\theta(a_t|s_t)) \right) + \sum_{t=0}^{T-1} \left(P(s_{t+1}|s_t,a_t) \right) \right]$$

$$= \sum_{t=0}^{T-1} \left(\nabla_\theta \left[\log(\pi_\theta(a_t|s_t)) \right] \right) + \sum_{t=0}^{T-1} \left(\nabla_\theta \Big[\underbrace{P(s_{t+1}|s_t,a_t)}_{\text{no dependence on } \theta \text{ so derivative is } 0} \Big] \right)$$

$$= \sum_{t=0}^{T-1} \left(\nabla_\theta \left[\log(\pi_\theta(a_t|s_t)) \right] \right)$$

So

$$\nabla_\theta \left[U(\theta) \right] \simeq \frac{1}{m} \sum_{i=1}^{m} \left(\sum_{t=0}^{T-1} \left(\nabla_\theta \left[\log(\pi_\theta(a_t^{(i)}|s_t^{(i)})) \right] \right) R(\tau^{(i)}) \right)$$