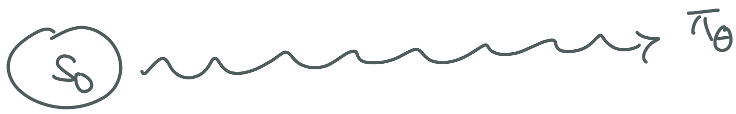
LET'S SAY WE HAVE A POLICY $\pi_\theta(a|s)$

ROLLOUT π_θ FROM START STATE S_0

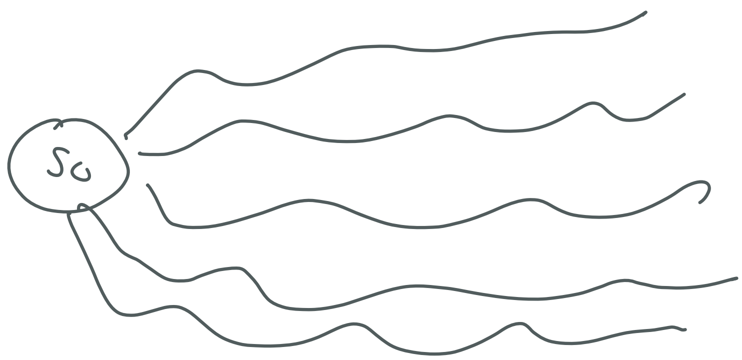


$a_0 \quad s_1 \quad a_1 \quad s_2 \quad a_2 \quad - - - - - - - \quad s_{T-1} \quad a_{T-1}$

$$\xi = (s_0, a_0, s_1, a_1, - - - - - - s_{T-1}, a_{T-1})$$

$$\boxed{P_\theta(\xi) = P(s_0)\,\pi_\theta(a_0|s_0)\,P(s_1|s_0, a_0)\,\pi_\theta(a_1|s_1) - - - - - -}$$

EXPECTED TOTAL REWARD



$$J(\theta) = \mathop{E}_{\xi \sim P_\theta(\xi)} R(\xi)$$

$$\nabla_\theta J(\theta) = \nabla_\theta \mathop{E}_{\xi \sim P_\theta(\xi)} R(\xi) = \nabla_\theta \sum_\xi P_\theta(\xi)\, R(\xi)$$

NAIVE APPROACH:

$$\nabla_\theta J(\theta) = \sum_\xi \boxed{\nabla_\theta P_\theta(\xi)}\, R(\xi)$$

APPLY CHAIN RULE.

$$\nabla_\theta P_\theta(\xi) = P(s_0)\,\nabla_\theta \pi_\theta(a_0|s_0)\,\underbrace{\boxed{P(s_1|s_0, a_0)}}_{\text{UNKNOWN}} - - - - - - -$$

$$+ P(s_0)\,\pi_\theta(a_0|s_0)\,\underbrace{\boxed{P(s_1|s_0, a_0)}}_{\text{UNKNOWN.}}\,\nabla \pi_\theta(a_1|s_1) - - - - -$$

$$+ \vdots$$

$$P_\theta(\xi) = P(s_0)\,\pi_\theta(a_0|s_0)\,P(s_1|s_0,a_0)\,\pi_\theta(a_1|s_1)\cdots\cdots$$

$$\log P_\theta(\xi) = \log P(s_0) + \log \pi_\theta(a_0|s_0) + \log P(s_1|s_0,a_0) + \cdots\cdots$$

$$\nabla_\theta \log P_\theta(\xi) = 0 + \nabla_\theta \log \pi_\theta(a_0|s_0) + 0 + \nabla_\theta \log \pi_\theta(a_1|s_1) + \cdots$$

$$\boxed{\underbrace{\nabla_\theta \log P_\theta(\xi)}_{\frac{1}{P_\theta(\xi)}\nabla_\theta P_\theta(\xi)} = \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t)} \rightarrow (1)$$

$$\nabla_\theta J(\theta) = \sum_{\xi} \boxed{\nabla_\theta P_\theta(\xi)}\, R(\xi)$$

$$= \sum_{\xi} P_\theta(\xi)\, \nabla_\theta \log P_\theta(\xi)\, R(\xi)$$

$$= \mathop{E}_{\xi \sim P_\theta(\xi)} \left[\nabla_\theta \log P_\theta(\xi)\, R(\xi) \right]$$

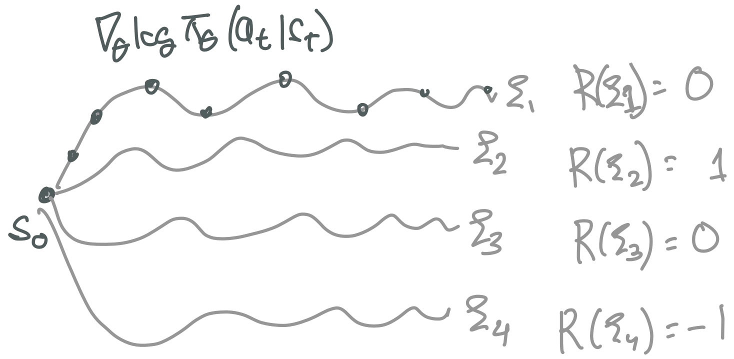
APPROXIMATE THE EXPECTATION BY SAMPLING TRAJECTORIES

$\{\xi_i\}_{i=1}^{N}$ BY ROLLING OUT POLICY π_θ IN REAL WORLD

$$\tilde{\nabla}_\theta J(\theta) = \frac{1}{N}\sum_{i=1}^{N} \left[\nabla_\theta \log P_\theta(\xi_i)\, R(\xi_i) \right]$$

PLUGGING IN (1)

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



$$\mathbb{E}_{\substack{s \sim d^{\pi_\theta} \\ a \sim \pi_\theta}} \left[\nabla_\theta \log \pi_\theta (a | s) \cdot Q^{\pi_\theta}(s, a) \right]$$