Policy Gradient Theorem

$$J(\theta) = \mathop{E}_{\substack{a_t \sim \pi_\theta(\cdot|s_t) \\ s_{t+1} \sim P(\cdot|s_t, a_t)}} \left[\sum_{t=0}^{T-1} r(s_t, a_t) \right]$$

$$\nabla_\theta J(\theta) = \nabla_\theta \mathop{E}_{\substack{a_t \sim \pi_\theta(\cdot|s_t) \\ s_{t+1} \sim P(\cdot|s_t, a_t)}} \left[\sum_{t=0}^{T-1} r(s_t, a_t) \right]$$

$$= \nabla_\theta \sum_{s_0, a_0, s_1, a_1, \ldots} P(s_0)\, \pi_\theta(a_0|s_0)\, P(s_1|s_0, a_0)\, \pi_\theta(a_1|s_1) \ldots \left[\sum_{t=0}^{T-1} r(s_t, a_t) \right]$$

All trajectories

BAD IDEA: TRY CHAIN RULE.

$$\nabla_\theta \sum_{s_0, a_0, s_1, a_1, \ldots} P(s_0)\, \nabla_\theta \pi_\theta(a_0|s_0)\, P(s_1|s_0, a_0) \ldots +$$

$$\sum_{s_0, a_0, s_1, a_1, \ldots} P(s_0)\, \pi_\theta(a_0|s_0)\, P(s_1|s_0, a_0)\, \nabla_\theta \pi_\theta(a_1|s_1) \ldots$$

Much Better Idea

$\tau = (s_0, a_0, s_1, a_1, \ldots, s_T)$

$$\nabla_\theta J(\theta) = \nabla_\theta \underset{\substack{a_t \sim \pi_\theta(\cdot|s_t) \\ s_{t+1} \sim \mathbb{P}(\cdot|s_t, a_t)}}{E} \left[\sum_{t=0}^{T-1} r(s_t, a_t) \right]$$

$$= \nabla_\theta \left(\sum_\tau P(\tau) \frac{P_\theta(\tau)}{P_\theta(\tau)} \right) \left[\sum_{t=0}^{T-1} r(s_t, a_t) \right]$$

$$= \underset{\tau \sim P_\theta(\tau)}{E} \left[\frac{\nabla_\theta P_\theta(\tau)}{P_\theta(\tau)} \left[\sum_{t=0}^{T-1} r(s_t, a_t) \right] \right]$$

Just rollout using current policies

$$= \underset{\tau \sim P_\theta(\tau)}{E} \left[\nabla_\theta \log P_\theta(\tau) \left[\sum_{t=0}^{T-1} r(s_t, a_t) \right] \right]$$

$$\nabla_\theta \left[\log \left(P(s_0) \; \pi_\theta(a_0|s_0) \; P(s_1|s_0, a_0) \; \pi_\theta(a_1|s_1) \; \ldots \right) \right]$$

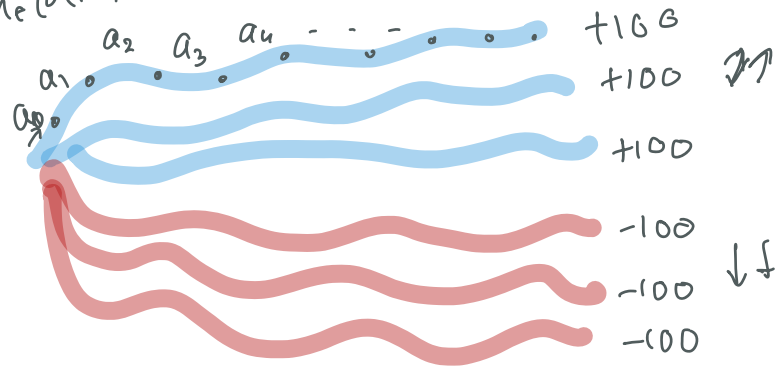
$$\nabla_\theta \left[\log P(s_0) + \log \pi_\theta(a_0|s_0) + \log P(s_1|s_0, a_0) + \log \pi_\theta(a_1|s_1) + \ldots \right]$$

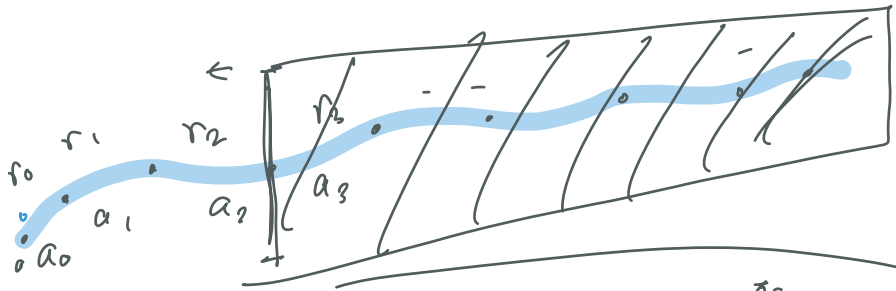
$$\nabla_\theta J(\theta) = \mathop{E}_{\tau \sim p_\theta(\tau)} \left[\nabla_\theta \log P_\theta(\tau) \left[\sum_{t=0}^{T-1} r(s_t, a_t) \right] \right]$$

$$= \mathop{E}_{\tau \sim p_\theta(\tau)} \left[\left(\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t) \right) \left(\sum_{t=0}^{T-1} r(s_t, a_t) \right) \right]$$

$r_1 + r_2 + r_3 + \cdots$

① Roll out $\pi_\theta(a_t|s_t)$





$$\nabla_\theta \log \pi_\theta(a_3|s_3) \; Q^{\pi_\theta}(s_3, a_3)$$