$\nabla_\theta J(\theta)$ $\quad \tau = (s_0, a_0, s_1, a_1, s_2, \dots\dots)$

TRAJECTORY

$$J(\theta) = \underset{\tau \sim P_\theta(\tau)}{E} R(\tau) = \sum_\tau P_\theta(\tau) R(\tau)$$

$$\nabla_\theta J(\theta) = \sum_\tau \nabla_\theta P_\theta(\tau) R(\tau)$$

$$\boxed{P_\theta(\tau) = P(s_0)\, \pi_\theta(a_0|s_0)\, P(s_1|s_0,a_0)\, \pi_\theta(a_1|s_1)\, P(s_2|s_1,a_1) \dots}$$

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

$$\nabla_\theta \log P_\theta(\tau) = \cancel{\nabla_\theta \log P(s_0)} + \nabla_\theta \log \pi_\theta(a_0|s_0) + \cancel{0} + \dots\dots$$

($\nabla_\theta \log P(s_0) = 0$)

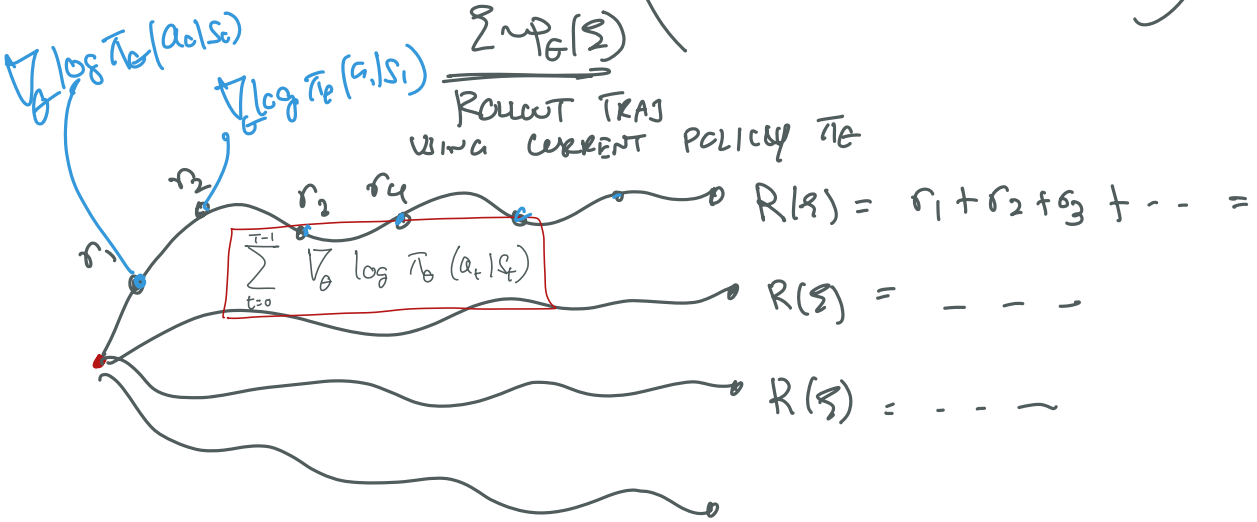
$$= \boxed{\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t)}$$

$$= \nabla_\theta \log P_\theta(\tau)$$

$$\nabla_\theta J(\theta) = \sum_\tau P_\theta(\tau) \boxed{\frac{\nabla_\theta P_\theta(\tau)}{P_\theta(\tau)}} R(\tau)$$

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

$$= \mathbb{E}\left(\boxed{\nabla_\theta \log P_\theta(\xi)} \, R(\xi) \right)$$



PERFORMANCE DIFFERENCE LEMMA

$$J(\pi_{\theta'}) - J(\pi_\theta) \geq 0$$

"NEW POLICY" — "OLD POLICY"

$$= \sum_{t \geq 0}^{T-1} \mathbb{E}_{s_t \sim d_{\pi_{\theta'}}} \Big[\underbrace{Q^{\tilde{\pi}_\theta}(s_t, \pi_{\theta'}(s_t)) - Q^{\tilde{\pi}_\theta}(s_t, \pi_\theta(s_t))}_{\geq 0} \Big]$$

$$V^{\pi_\theta}(s_t)$$

$$= \sum_{t=0}^{T} \mathbb{E}_{s_t \sim d_{\pi_{\theta'}}} \Big[\underbrace{Q^{\pi_\theta}(\cdot,\cdot) - V^{\pi_\theta}(\cdot)}_{A^{\pi_\theta}(s, \pi_{\theta'}(s))} \Big]$$