Policy Gradient

find π to maximize $J(\pi)$

$= \mathbb{E}_{s \sim d_0}[V^{\pi}(s)]$

$\max_\theta J(\pi_\theta)$.

G.D. $\theta \leftarrow \theta + \alpha \cdot \boxed{\nabla_\theta J(\pi_\theta)}$ $\rightarrow \nabla_\theta J(\pi_\theta)|_\theta$

Claim: given an on-policy traj. $\tau = (s_1, a_1, r_1, \cdots, s_H, a_H, r_H)$ $a_t \sim \pi_\theta$.

$\rightarrow$ can construct PG estimator s.t.

$$\boxed{\mathbb{E}_{\pi_\theta}[(\cdot)] = \nabla_\theta J(\pi_\theta) = \nabla J(\pi)}$$

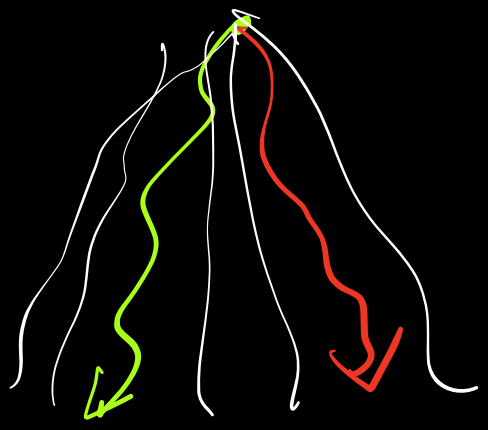
$$\nabla J(\pi) = \nabla \mathbb{E}_{\tau \sim \pi}[R(\tau)] \quad \rightarrow \sum_{t=1}^{H} \gamma^{t-1} r_t.$$

$\frac{d}{dx} \log x = \frac{1}{x}$,

$$= \nabla \sum_\tau P^{\pi}(\tau) R(\tau).$$

$$= \sum_\tau \nabla P^{\pi}(\tau) R(\tau).$$

$$= \sum_\tau P^{\pi}(\tau) \nabla \log P^{\pi}(\tau) \cdot R(\tau).$$



$$= \mathbb{E}_{\tau \sim \pi}[\nabla \log P^{\pi}(\tau) \cdot R(\tau)].$$

$$\nabla \log P^{\pi}(s_1, a_1, r_1, s_2, a_2, r_2, \cdots, s_H, a_H, r_H)$$

$$= \nabla \log \Big(d_0(s_1) \cdot \pi(a_1|s_1) \cdot R(r_1|s_1,a_1) \cdot P(s_2|s_1,a_1) \cdot \pi(a_2|s_2) \cdots \Big).$$

$$= \nabla \Big(\log d_0(s_1) + \log \pi(a_1|s_1) + \log R(r_1|s_1,a_1) + \log P(s_2|s_1,a_1) + \log \pi(a_2|s_2) + \cdots \Big).$$

$$= \sum_{t=1}^{H} \nabla \log \pi(a_t|s_t).$$

$$\Rightarrow \nabla J(\pi) = \mathbb{E}_{\pi}\left[\left(\sum_{t=1}^{H} \nabla \log \pi(a_t|s_t) \right) \left(\sum_{t=1}^{H} \gamma^{t-1} r_t \right) \right]$$

Prompt x LLM: $\begin{cases} y_{1:L} & \text{text,} \\ y_{L+1:2L} & \text{code output} \\ y_{2L+1:3L} & \text{text} \end{cases}$

$$\sum_{t=1}^{3L} \nabla \log \pi(y_t | x, y_{<t}).$$

"REINFORCE": $\left(\sum_{t=1}^{H} \gamma^{t-1} r_t \right) \left(\sum_{t=1}^{H} \nabla \log \pi(a_t|s_t) \right)$.

$$= \sum_{t=1}^{H} \left(\nabla \log \pi_\theta(a_t|s_t) \cdot \left(\sum_{t'=1}^{H} \gamma^{t'-1} r_{t'} \right) \right).$$

$$\mathbb{E}_{\pi}[\ \ ] = \mathbb{E}_{\pi}\left[\sum_{t=1}^{H} \nabla \log \pi_\theta(a_t|s_t) \cdot \sum_{t'=t}^{H} \gamma^{t'-1} r_{t'} \right].$$

$$\mathbb{E}_{\pi}\left[\nabla \log \pi_\theta(a_t|s_t) \sum_{t'=1}^{t-1} \gamma^{t'-1} r_{t'}\right].$$

$$= \mathbb{E}_{(s_1,a_1,r_1,\cdots,s_t)\sim\pi}\left[\cdots \;\middle|\; s_1,a_1,r_1,\cdots,s_t\right].$$

$$= \mathbb{E}\left[\mathbb{E}_{a_t\sim\pi(\cdot|s_t)}\left[\nabla \log \pi_\theta(a_t|s_t)\right]\left(\sum_{t'=1}^{t-1}\gamma^{t'-1} r_{t'}\right)\right].$$

$$\sum_a \pi_\theta(a|s_t)\cdot\frac{\nabla\pi_\theta(a|s_t)}{\pi_\theta(a|s_t)} = \nabla\sum_a \pi_\theta(a|s_t) = \nabla 1 = 0$$

$$\mathbb{E}_{\pi}\left[\sum_{t=1}^{H}\nabla \log \pi(a_t|s_t)\cdot\sum_{t'=t}^{H}\gamma^{t'-1} r_{t'}\right].$$

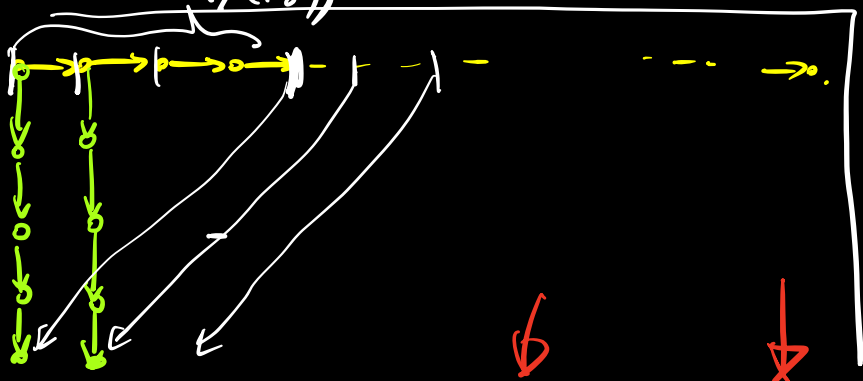
$$= \mathbb{E}_{\pi}\left[\gamma^{t-1}\sum_{t=1}^{H}\nabla \log \pi(a_t|s_t)\cdot Q^{\pi}(s_t,a_t)\right]$$

$$= \mathbb{E}_{\pi}\left[\sum_{t=1}^{H}\gamma^{t-1}\nabla \log \pi(a_t|s_t)\cdot\left(Q^{\pi}(s_t,a_t) - f(s_t)\right)\right]$$

$\gamma^{t-1} \le O(1/(1-\gamma))$

$f(s_t) = V^{\pi}(s_t)$

$A^{\pi}(s_t,a_t)$



$$\frac{d}{d\theta} J(\pi_\theta) = \lim_{\Delta\theta \to 0} \frac{J(\pi_{\theta+\Delta\theta}) - J(\pi_\theta)}{\Delta\theta}$$

IS

$$\tau \sim \pi_\theta \qquad R(\tau) = \sum_{t=1}^{H} \gamma^{t-1} r_t$$