$$\nabla_\theta J(\theta) = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)\, A^\pi(s_t, a_t)$$

$\theta \longrightarrow \theta'$ How to show improvement

$*$: $J(\theta') - J(\theta)$

$J(\theta') - J(\theta) \geq +$ $\Rightarrow$ policy improvement

$$J(\theta') - J(\theta) = E_{\tau \sim p_{\theta'}(\tau)}\left[\sum_t \gamma^t r(s_t, a_t)\right] - E_{\tau \sim p_\theta(\tau)}\left[\sum_t \gamma^t r(s_t, a_t)\right]$$

$$\boxed{J(\theta) = E_{s_0 \sim p(s_0)}\left[V^{\pi_\theta}(s_0)\right]}$$

$$= E_{\tau \sim p_{\theta'}(\tau)}\left[\sum_t \gamma^t r(s_t, a_t)\right] - \underbrace{E_{s_0 \sim p(s_0)}\left[V^{\pi_\theta}(s_0)\right]}_{= E_{\tau \sim p_{\theta'}(\tau)}\left[V^{\pi_\theta}(s_0)\right]}$$

$$= E_{\tau \sim p_{\theta'}(\tau)}\left[\sum_t \gamma^t r(s_t, a_t) + \sum_{t=1}^{\infty} \gamma^t V^{\pi_\theta}(s_t) - \sum_t^{\infty} \gamma^t V^{\pi_\theta}(s_t)\right]$$

$$= E_{\tau \sim p_{\theta'}(\tau)}\left[\sum_t \gamma^t \left(r(s_t, a_t) + \gamma V^{\pi_\theta}(s_{t+1}) - V^{\pi}(s_t)\right)\right]$$

$$= \mathbb{E}_{\tau \sim p_{\theta'}(\tau)} \left[\sum_t \gamma^t A^{\pi_\theta}(s_t, a_t) \right]$$

$$P_{\theta'}(\tau) = P(s_0) \prod \pi_{\theta'}(a_t | s_t) \, \mathbb{P}(s_t | s_{t-1}, a_{t-1})$$

$$= \sum_t \gamma^t \mathbb{E}_{\tau \sim p_{\theta'}(\tau)} \left[A^{\pi_\theta}(s_t, a_t) \right]$$

$$= \sum_t \gamma^t \mathbb{E}_{(s_t, a_t) \sim \pi^{\theta'}(a_t | s_t) p^{\theta'}(s_t)} \left[A^{\pi_\theta}(s_t, a_t) \right]$$

$$\Rightarrow \sum_t \mathbb{E}_{s_t \sim p^{\theta'}(s_t)} \left[\mathbb{E}_{a_t \sim \pi^{\theta'}(a_t|s_t)} \left[A^{\pi_\theta}(s_t, a_t) \right] \right]$$

$$\Rightarrow \sum_t \mathbb{E}_{s_t \sim p^{\theta'}(s_t)} \left[\mathbb{E}_{a_t \sim \pi^{\theta}(a_t|s_t)} \left[\gamma^t \frac{\pi^{\theta'}(a_t|s_t)}{\pi^{\theta}(a_t|s_t)} A^{\pi_\theta}(s_t, a_t) \right] \right]$$

$$\approx s_t \sim p^{\theta}(s_t)$$

$$G$$

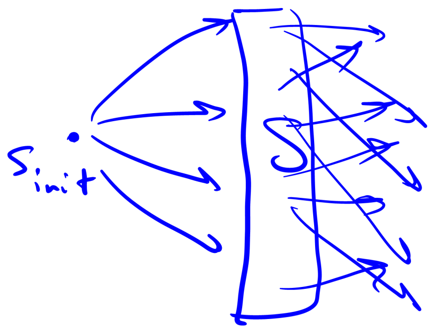
$$\underset{\theta'}{\operatorname{argmax}}\ G$$

Claim: $P_\theta(s_t)$ is close to $P_{\theta'}(s_t)$ if π_θ and $\pi_{\theta'}$ are close enough.

π_θ deterministic.

$$\pi_{\theta'}(a_t \neq \pi_\theta(s_t) \mid s_t) = \varepsilon$$

$$|P_{\theta'}(s_t) - P_\theta(s_t)| \leq \, ?$$



s_t

$$P_{\theta'}(s_t) = (1-\varepsilon)^t P_\theta(s_t) + \left(1-(1-\varepsilon)^t\right) P_{\text{بدردنخور}}(s_t)$$

$$\left|P_{\theta'}(s_t) - P_\theta(s_t)\right| = \left|\left(1-(1-\varepsilon)^t\right)\right| \left|P_{\text{بدردنخور}}(s_t) - P_\theta(s_t)\right|$$

$$\leq 2\left|\left(1-(1-\varepsilon)^t\right)\right|$$

$$\leq 2\left|\left(1-(1-\varepsilon t)\right)\right|$$

$$= 2\varepsilon t$$

General Case

$|\pi_{\theta'}(\cdot|s_t) - \pi_\theta(\cdot|s_t)| \leq \varepsilon$

if $|p_X(x) - p_Y(x)| = \epsilon$, exists $p(x, y)$ such that $p(x) = p_X(x)$ and $p(y) = p_Y(y)$ and $p(x = y) = 1 - \epsilon$

$\Rightarrow p_X(x)$ "agrees" with $p_Y(y)$ with probability ϵ

$\Rightarrow \pi_{\theta'}(\mathbf{a}_t|\mathbf{s}_t)$ takes a different action than $\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)$ with probability at most ϵ

$\leq 2\varepsilon t$

$$\sum_t \mathbb{E}_{s_t \sim p^{\theta'}(s_t)} [X] = \sum_t \sum_{s_t} p^{\theta'}(s_t) \, X(s_t)$$

$$= \sum_t \left[\sum_{s_t} \left(p^{\theta}(s_t) + p^{\theta'}(s_t) - p^{\theta}(s_t) \right) X(s_t) \right]$$

$$\geq \underbrace{\sum_t \sum_{s_t} p^{\theta}(s_t) X(s_t)}_{①} - \sum_t \sum_{s_t} \left| p^{\theta'}(s_t) - p^{\theta}(s_t) \right| X(s_t)$$

$$\geq ① - \underbrace{\sum_t \sum_{s_t} \left| p^{\theta'}(s_t) - p^{\theta}(s_t) \right|}_{} \max \left| X(s_t) \right|$$

$$\left| p^{\theta'}(\cdot) - p^{\theta}(\cdot) \right|$$

$$TV\left(p(x), q(x)\right) = \sum_x \left| p(x) - q(x) \right|$$

$$\overset{*}{\geq} ① - \sum_t 2\epsilon t \max |\chi(s_t)|$$

$$\Rightarrow \sum_t \mathbb{E}_{s_t \sim p^{\theta'}(s_t)} \mathbb{E}_{a_t \sim \pi^{\theta}(a_t|s_t)} \left[\gamma^t \frac{\pi^{\theta'}(a_t|s_t)}{\pi^{\theta}(a_t|s_t)} A^{\pi^{\theta}}(a_t, s_t) \right]$$

$$\overset{*}{\geq} \sum_t \mathbb{E}_{s_t \sim p^{\theta}(s_t)} \mathbb{E}_{a_t \sim \pi^{\theta}(a_t|s_t)} \left[\gamma^t \frac{\pi^{\theta'}(a_t|s_t)}{\pi^{\theta}(a_t|s_t)} A^{\pi^{\theta}}(a_t, s_t) \right] - O\left(\frac{\epsilon t R_{max}}{1-\gamma} \right)$$

The End.