Data: $(s_1, a_1, \cdots, s_H, a_H)$ $a_h \sim \pi_b$.

Goal: $J(\pi) := \mathbb{E}_\pi\left[\sum_{h=1}^{H} r_h\right]$.

Estimator: let $\rho_h := \frac{\pi(a_h|s_h)}{\pi_b(a_h|s_h)}$ $\Rightarrow$ $\rho_{1:h} := \prod_{h'=1}^{h} \rho_{h'}$.

(step-wise) IS: $\sum_{h=1}^{H} \rho_{1:h} r_h$.

$\rho_{1:H}\left(\sum_h r_h\right)$.

WIS: $\dfrac{\sum_{h=1}^{H} \rho_{1:h}^i r_h^i}{\sum \rho_{1:h}^i}$

$\sum_{h=1}^{H} \rho_{1:h} r_h$ $\quad \rho_1 r_1 + \rho_{1:2} r_2 + \cdots + \rho_{1:H} r_H$

$= \rho_1(r_1 + \rho_2(r_2 + \rho_3(r_3 + \cdots + \rho_H r_H)))$

Define $v_0 := 0$.

$v_{H-h+1} = \rho_h(r_h + v_{H-h})$.

$\Rightarrow v_H$.

$V^\pi(s_h)$.

$r_h + V^\pi(s_{h+1})$.

ρ_h: $\dfrac{\pi(a_h|s_h)}{\pi_b(a_h|s_h)}$.

$V^\pi(s_{h+1})$.

$s_h, a_h \sim \pi_b$.

$Q^\pi(s_h, a_h)$.

$\downarrow \rho_h$

$V^\pi(s_h)$

s_h, $a_h \sim \pi_b$.

$Q^\pi(s_h, a_h)$.

DR: $V_{H-h+1} = \rho_h \left(r_h + V_{H-h} - \hat{Q}^\pi(s_h, a_h) \right) + \hat{Q}^\pi(s_h, \pi)$.

Policy Gradient $\Pi = \{\pi_\theta : \theta \in \Theta\}$.

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

Idea: $\nabla J(\pi) = \nabla_\theta J(\pi_\theta)$.

Goal: Given $\tau = (s_1, a_1, s_2, a_2 \dots, s_H, a_H)$ w/ $a_h \sim \pi$. estimate $\nabla J(\pi)$ from τ.

$\nabla_\theta J(\pi_\theta) \stackrel{?}{=} \mathbb{E}_{\tau \sim \pi}[f(\tau)] = \sum_\tau P^\pi(\tau) f(\tau)$

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

$= \sum R(\tau) \nabla P^\pi(\tau)$.

SGD: $\theta \leftarrow \theta + \alpha \cdot f(\tau)$ ✓.

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

$$= \sum_\tau R(\tau) P^\pi(\tau) \nabla \log \left(d_0(s_1) \cdot \pi(a_1|s_1) \cdot P(s_2|s_1,a_1) \cdot \pi(a_2|s_2) \cdots \right)$$

$$= \sum_\tau R(\tau) P^\pi(\tau) \sum_{h=1}^{H} \left(\nabla \log P(s_h|s_{h-1},a_{h-1}) + \nabla \log \pi(a_h|s_h) \right).$$

$$\propto \mathbb{E}_{\tau\sim\pi} \left[R(\tau) \sum_{h=1}^{H} \nabla \log \pi(a_h|s_h) \right]$$ "REINFORCE"

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

$\tau \sim \pi_\theta$. $\mathbb{E}_\pi[R(\tau)] = J(\pi_\theta)$.

$$\left(\prod_{h=1}^{H} \frac{\pi_{\theta+\Delta\theta}(a_h|s_h)}{\pi_\theta(a_h|s_h)} \cdot \right) R(\tau).$$

$$\left(\sum_{h=1}^{H} r_h\right)\left(\sum_{h=1}^{H} \nabla \log \pi(a_h|s_h)\right)$$

discounted

$$\nabla J(\pi) = \frac{1}{1-\gamma}\, \mathbb{E}_{(s,a)\sim d^{\pi}}\left[\nabla \log \pi(a|s)\cdot Q^{\pi}(s,a)\right]$$

$$\sum_{h=1}^{H} \nabla \log \pi_{\underline{h}}(a_h|s_h)\left(\sum_{h'=h}^{H} r_{h'}\right)$$

"Actor-critic"

$$\mathbb{E}_{(s,a)\sim\pi_\theta}\left[\nabla \log \pi_{\underline{\theta}}(a|s)\cdot \hat{Q}^{\pi}(s,a)\right] \quad > \sum r_h$$

FQE

$$\nabla J(\pi) = \frac{1}{1-\gamma}\, \mathbb{E}_{(s,a)\sim d^{\pi}}\left[\nabla \log \pi(a|s)\cdot\left(Q^{\pi}(s,a) - g(s)\right)\right]$$

$\forall g: S \to \mathbb{R}$

$$\mathbb{E}_{a\sim\pi}\left[\nabla \log \pi(a|s)\cdot\right] g(s)$$

$$= \sum_a \pi(a|s) \cdot \frac{\nabla \pi(a|s)}{\pi(a|s)} = \vec{0}$$

PG: find $\hat{\pi}$, $\|\nabla J(\hat{\pi})\| = \varepsilon$.

$(\Theta = \mathbb{R}^d)$.

$$\left|(\nabla J(\hat{\pi}))^T (\hat{\theta} - \theta)\right| \le \varepsilon \quad \forall \theta$$

$$J(\pi^*) - J(\hat{\pi})$$

$$= \frac{1}{1-\gamma} \mathbb{E}_{d^{\pi^*}}\left[Q^{\hat{\pi}}(s, \pi^*) - Q^{\hat{\pi}}(s, \hat{\pi}) \right].$$

$$\le \frac{1}{1-\gamma} \mathbb{E}_{d^{\pi^*}}\left[Q^{\hat{\pi}}(s, \pi_{Q^{\hat{\pi}}}) - Q^{\hat{\pi}}(s, \hat{\pi}) \right]$$

$$\mathbb{E}_{(s,a) \sim d^{\hat{\pi}}}\left[Q^{\hat{\pi}}(s,a) \cdot \nabla \log \hat{\pi}(a|s) \right]$$

$$\le \frac{1}{1-\gamma} \left\| \frac{d^{\pi^*}}{d^{\hat{\pi}}} \right\|_\infty \cdot \mathbb{E}_{d^{\hat{\pi}}}\left[Q^{\hat{\pi}}(s, \pi_{Q^{\hat{\pi}}}) - Q^{\hat{\pi}}(s, \hat{\pi}) \right].$$

I

II

Policy parameters

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