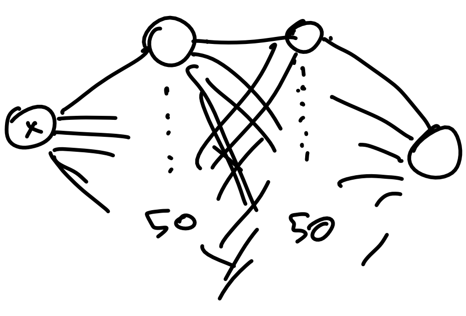
$$\theta \leftarrow \theta + \alpha \frac{\partial Q}{\partial \theta}\left(Q_\theta(s,a) - y\right)$$

$$\mathcal{L}(s,a,\theta) = \left(Q_\theta(s,a) - y\right)^2 \qquad y \leftarrow r + \gamma \max Q_{\theta'}(s',a')$$

$$\frac{\partial \mathcal{L}}{\partial \theta} = 2\left(Q_\theta(s,a) - y\right)\frac{\partial Q}{\partial \theta}$$

$$y = W_3\sigma\left(W_2\, \sigma\left(W_1 x + b_1\right) + b_2\right) + b_3$$

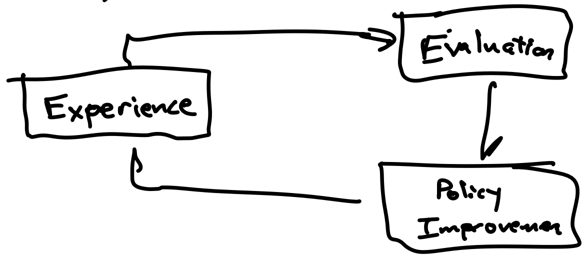
1×50 50×50 50×50 50 55 1



Policy Gradients

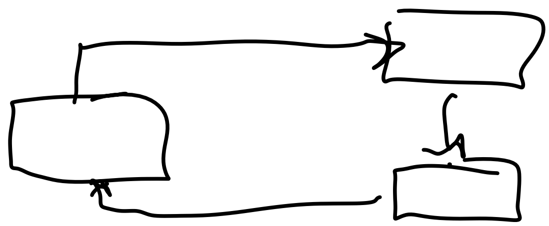
Q Learning

$$\theta \leftarrow \theta + \alpha \frac{\partial Q}{\partial \theta}\left(Q_\theta(s,a) - r - \gamma \max_{a'} Q_{\theta'}(s',a')\right)$$



$$\pi(a|s) \begin{cases} 1-\varepsilon & \text{if } a = \operatorname{argmax} Q_\theta(s,a) \\ \frac{\varepsilon}{|A|-1} & \text{o.w.} \end{cases}$$

Policy Grad



$$\hat{Q} = \sum_t r_t$$

$$\tilde{\pi}_\theta(a|s)$$

$$\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$$

$$\tau = (s_1, a_1, s_2, a_2 \ldots s_T, a_T)$$

$$J(\theta) = \mathop{E}_{\tau \sim p(\tau|\pi_\theta)}[r(\tau)]$$

$$J(\theta) = \int p(\tau|\pi_\theta)\, r(\tau)\, d\tau$$

$$\nabla_\theta J(\theta) = \int \nabla_\theta p(\tau|\pi_\theta)\, r(\tau)\, d\tau$$

$$\frac{d \log(f(x))}{dx} = \frac{df(x)}{dx}\frac{1}{f(x)}$$

$$= \int \frac{\nabla_\theta p(\tau|\pi_\theta)}{p(\tau|\pi_\theta)} p(\tau|\pi_\theta)\, r(\tau)\, d\tau$$

$$= \int \nabla_\theta \log p(\tau|\pi_\theta)\, p(\tau|\pi_\theta)\, r(\tau)\, d\tau$$

$\pi: S \rightarrow A$

$\pi: S \times A \rightarrow \mathbb{R}$

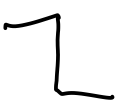
prob a ins

$$= \mathop{E}_{\tau \sim p(\tau|\pi_\theta)}\left[\nabla_\theta \underline{\log p(\tau|\pi_\theta)}\; r(\tau)\right]$$

$$\underbrace{p(\tau|\pi_\theta)} = p(s_1) \prod_t \pi_\theta(a_t|s_t) T(s_{t+1}|s_t,a_t)$$

$$\log p = \cancel{\log p(s_1)} + \underline{\sum_t \log \pi(a_t|s_t)} + \cancel{\log T(s_{t+1}|s_t,a_t)} \quad \log(xy) = \log x + \log y$$

$$\nabla_\theta \log p = \sum_t \nabla_\theta \log \pi_\theta(a_t|s_t)$$



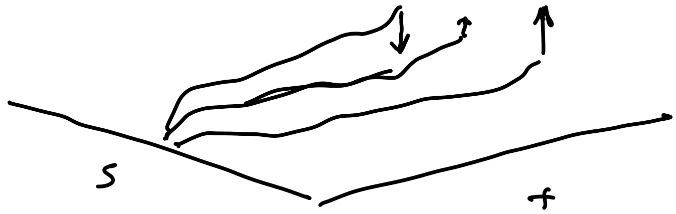
p(1) p(2)

$$\nabla_\theta J(\theta) = \mathop{E}_{\tau \sim p(\tau|\pi_\theta)}\left[\left(\sum_t \nabla_\theta \log \pi_\theta\right)\left(\underline{\sum R(s_t,a_t)}\right)\right]$$

simulation

MC Simulations

$$\approx \frac{1}{N}\sum_{i=1}^{N}\sum_t \nabla_\theta \log \pi_\theta(a_{t,i}, s_{t,i}) \underline{\sum_t r_{t,i}}$$



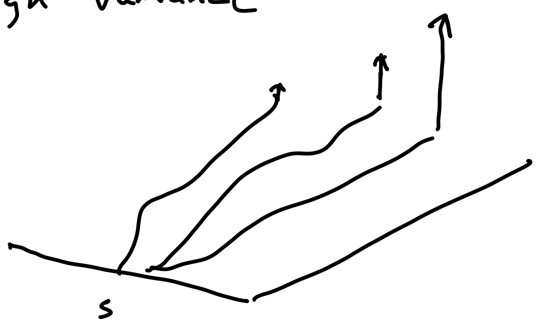
REINFORCE™

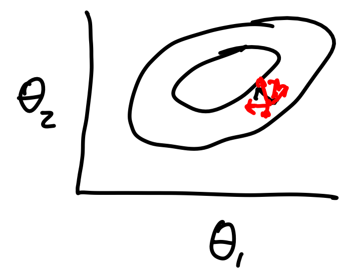
1. sample $\{\tau_i\}$ (run π_θ)
2. $\nabla_\theta J(\theta) = \frac{1}{N}\sum_i \sum_t \nabla_\theta \log \pi \;\; r(\tau)$

Evaluation

3. $\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$ improvement

Won't work: High Variance





1. Causality

$$\nabla_\theta J(\theta) \simeq \frac{1}{N} \sum_i \sum_t \left(\nabla_\theta \log \pi_\theta \sum_{t'=t}^{T} r_{i,t'} \right)$$

<u>Always do this</u>

$\hat{Q}(s_{i,t}, a_{i,t})$

2. Baselines

$$\nabla_\theta J(\theta) \simeq \frac{1}{N} \sum_i \sum_t \left(\nabla_\theta \log \pi_\theta \left[r(\tau_i) - \underline{b} \right] \right)$$

$$b = \frac{1}{N} \sum r(\tau)$$

$$\begin{aligned} \mathop{E}_{\tau}\left[\nabla_\theta \log \pi_\theta \, b\right] &= \int \pi_\theta \nabla_\theta \log \pi_\theta \, b \, d\tau \\ &= \int \nabla_\theta \, p(\tau|\pi_\theta) \, b \, d\tau \\ &= b \nabla_\theta \left(\underbrace{\int p(\tau|\pi_\theta) \, d\tau}_{1} \right) \\ &= 0 \end{aligned}$$

Adding a baseline does not bias gradien

$$\frac{dVar}{db} \doteq 0$$

$$b^* = \frac{E\left[g(\tau)^2 r(\tau)\right]}{E\left(g(\tau)^2\right)}$$

In practice use $b = \frac{1}{N} \sum r(\tau)$

REINFORCE$^+$ it works :)

REINFORCE is onpolicy :(

$\bar{\pi} \neq \pi_\theta$

Importance Sampling
$\mathop{E}_{x \sim p(x)}\left[f(x)\right] = \int p(x) f(x) \, dx$ $\int \frac{q(x)}{q(x)} p(x) f(x) \, dx$ $\int q(x) \frac{p(x)}{q(x)} f(x) \, dx$ $= \mathop{E}_{x \sim q(x)}\left[\frac{p(x)}{q(x)} f(x)\right]$

$$J(\theta) = \underset{\tau\sim p(\tau|\bar{\pi})}{E}\left[\frac{p(\tau|\pi_\theta)}{p(\tau|\bar{\pi})}\, r(\tau)\right]$$

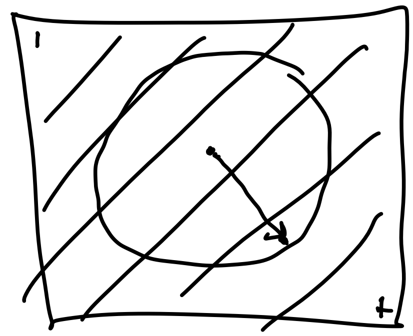
$$\nabla_{\theta'} J(\theta') = E\left[\prod_{t=1}^{T} \frac{\pi_{\theta'}(a_t|s_t)}{\bar{\pi}(a_t|s_t)} \left(\sum_{t=1}^{T} \nabla_\theta \log \pi_{\theta'}(a_t|s_t)\right)\left(r(\tau)\right)\right]$$

gets really small

~~Gradient Ascent~~ Policy Iteration

$$\theta' = \underset{\theta'}{\arg\max}\; \nabla_\theta J(\theta)^T (\theta' - \theta)$$

$$\text{s.t. } \|\theta - \theta'\|^2 \leq \varepsilon$$



Actually want something like

$$\theta' = \underset{\theta'}{\arg\max}\; \nabla_\theta J(\theta)^T (\theta' - \theta)$$

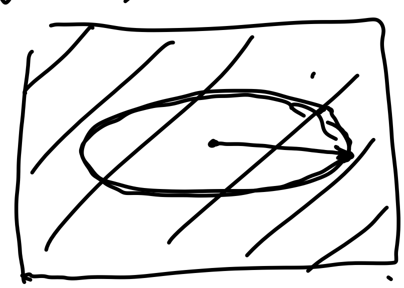
$$\text{s.t. } D_{KL}(\pi_{\theta'}(a|s)\,\|\,\pi_\theta(a|s)) \leq \varepsilon$$

$$D_{KL}(p\|q) = \underset{x\sim p}{E}\left[\log\frac{p(x)}{q(x)}\right]$$

Taylor Expansion

$$D_{KL}(\quad) \approx \frac{1}{2}(\theta' - \theta)^T F (\theta' - \theta)$$

Fischer info matrix



$$\theta' = \theta + \alpha F^{-1} \nabla_\theta J(\theta)$$

"Natural Gradient"

Trust Region Policy Opt TRPO

$$\alpha = \sqrt{\frac{2\varepsilon}{\nabla_\theta J(\theta)^T F \nabla_\theta J(\theta)}}$$

Proximal Policy Optimization PPO

Use regularization to stay close to old policy so that we can use importance sampling