Policy Gradient

VI & PI can yield π^t and π^{t+1} that are very different, leading to instabilities

Q & $\hat{Q}$ can be close but π & $\hat{\pi}$ can be far.

$\Rightarrow$ change policy incrementally ✓ by parametrizing π_θ & doing gradient ascent.

$\pi : S \to A$ can deal with continuous state spaces. ✓

$\to \pi_{\theta}(a|s)$

$$\min_\theta \; T\,E[V_{\pi^*}(S)] - \sum_{t=1}^{T} E[V_{\pi_{\theta_t}}(S)] =: J(\pi_\theta)$$

$$\begin{cases} \theta_0 \text{ initialize} \\ \theta_{t+1} \leftarrow \theta_t + \eta \nabla_\theta J(\pi_{\theta_t}) \end{cases}$$

Issue 1: π_θ needs to be differentiable ($\nexists$ deterministic policies)

stochastic policies:

e.g. softmax policies $\pi_\theta(a|s) \propto \exp(\theta_{s,a})$

loglinear " $\pi_\theta(a|s) \propto \exp(\theta \cdot \phi_{s,a})$

neural softmax policies $\pi_\theta(a|s) \propto \exp(f_\theta(s,a))$

Issue 2: $\nabla_\theta J(\pi_{\theta_t})$

$\hookrightarrow$ may require differentiation through entire dynamics of MDP.

Policy Gradient Theorem (REINFORCE version)

$$R(\tau) = \sum_{h=1}^{H} r_h$$

$$\tau = (s_1, a_1, r_1, s_2 \ldots s_H, a_H, r_H)$$

$$\nabla x = x \nabla \log x$$

$$\nabla_\theta E[V_{\pi_\theta}(S)] = \nabla_\theta E_{\tau \sim P^{\pi_\theta}_{d_0}}[R(\tau)]$$

$$= \nabla_\theta \sum_\tau R(\tau)\, Pr^{\pi_\theta}_{d_0}(\tau)$$

$$= \sum_\tau R(\tau)\, \nabla_\theta Pr^{\pi_\theta}_{d_0}(\tau)$$

$$= \sum_\tau R(\tau)\, Pr^{\pi_\theta}_{d_0}(\tau)\, \nabla \log Pr^{\pi_\theta}_{d_0}(\tau)$$

$$= \sum_\tau R(\tau)\, Pr^{\pi_\theta}_{d_0}(\tau)\, \nabla \log [d_0(s_1)\, \pi_\theta(a_1|s_1)\, P(s_2|s_1,a_1) \ldots \pi_\theta(a_H|s_H)\, P(s_{H+1}|s_H, a_H)]$$

$$= \sum_\tau R(\tau)\, \underset{\sim}{Pr^{\pi_\theta}_{d_0}}(\tau) \left(\sum_{h=1}^{H} \nabla \log \pi_\theta(a_h|s_h)\right)$$

$$= E_{\tau \sim p_{\mu_0}^{\pi_\theta}} \left[R(\tau) \sum_{h=1}^{H} \nabla \log \pi_\theta (a_h | s_h) \right]$$

REINFORCE

Compute stochastic gradient by 1) sampling trajectories using π_θ

2) Compute $R(\tau) \sum_{h=1}^{H} \nabla \log \pi_\theta (a_h | s_h)$

Note: estimate gradient with accuracy independent of size of state space.

Issue: Variance of gradient estimate can be high.

Per step importance sampling to lower variance

Q-version

$$\nabla_\theta J(\pi_\theta) = E_{(s,a) \sim d^{\pi_\theta}} \left[\nabla_\theta \log \pi_\theta (a|s) \; Q^{\pi_\theta}(s,a) \right]$$

1) generate trajectory using π_θ

2) pick random time step h.

3) compute $\nabla_\theta \log \pi_\theta (a_h | s_h) \underbrace{\sum_{h}^{H} r_h}$

Policy Gradient Theorem (Q-version)

$$\nabla_\theta V_{\pi_\theta}(s) = \nabla_\theta \; E_{a \sim \pi_\theta(s)} \left[Q_{\pi_\theta}(s,a) \right] = \nabla_\theta \sum_a \underbrace{\pi_\theta(a|s)} \; \underbrace{Q_{\pi_\theta}(s,a)}$$

$$= \sum_a \nabla_\theta \pi_\theta (a|s) \cdot Q_{\pi_\theta}(s,a) + \sum_a \pi_\theta(a|s) \; \nabla_\theta Q_{\pi_\theta}(s,a)$$

$$Q_{\pi_\theta}(s,a) = R(s,a) + \sum_{s'} P(s'|s,a) \underline{V_{\pi_\theta}(s')}$$

$$\nabla_\theta Q_{\pi_\theta}(s,a) = \sum_{s'} P(s'|s,a) \; \nabla_\theta V_{\pi_\theta}(s')$$

$$\Rightarrow \underbrace{\nabla_\theta V_{\pi_\theta}(s)} = \sum_a \nabla_\theta \pi_\theta(a|s) \cdot Q_{\pi_\theta}(s,a) + \sum_a \pi_\theta(a|s) \sum_{s'} P(s'|s,a) \underbrace{\nabla_\theta V_{\pi_\theta}(s')}$$

$$\nabla_\theta \pi_\theta = \pi_\theta \nabla_\theta \log \pi_\theta$$

Recursive equation in $\nabla_\theta V_{\pi_\theta}(s)$.

$$\nabla_\theta E_{s \sim d_h^{\pi_\theta}} \left[V_{\pi_\theta}(s) \right] = E_{s \sim d_h^{\pi_\theta}, a \sim \pi_\theta} \left[\nabla_\theta \log \pi_\theta(a|s) Q_{\pi_\theta}(s,a) \right]$$

$$+ \; \underbrace{E_{s \sim d_h^{\pi_\theta}, a \sim \pi_\theta, s' \sim P(\cdot|s,a)}}_{\star} \left[\nabla_\theta V_{\pi_\theta}(s') \right]$$

$\star$ $s' \sim d_{h+1}^{\pi_\theta}$

$$= \dots = \sum_{h'=h}^{H} E_{s \sim d_{h'}^{\pi_\theta}, a \sim \pi_\theta} \left[\nabla_\theta \log \pi_\theta(a|s) \, Q_{\pi_\theta}(s,a) \right]$$

Variance reduction in policy gradient

$$\Rightarrow \quad \mathbb{E}_{a \sim \pi_\theta(s)}[\underbrace{\nabla \log \pi_\theta(a|s)}_{f(s)}] = \sum_a \pi_\theta(a|s) \nabla \log \pi_\theta(a|s)$$

$$= \sum_a \nabla \pi_\theta(a|s) \qquad \nabla x = x \nabla \log x$$

$$= \nabla \cancelto{1}{\sum_a \pi_\theta(a|s)}$$

$$= 0$$

$\therefore$ can add any function $f : S \to \mathbb{R}$ to policy gradient as long as f doesn't depend on a / θ.

$\Rightarrow$ doesn't change bias of gradient estimate.

$$\underset{\text{vanilla}}{Q^{\pi_\theta}(s,a)} \longrightarrow Q^{\pi_\theta}(s,a) - f(s)$$

eg. $f = V^{\pi_\theta}(s)$

"critic"

Actor-critic approach:

$$\underset{\substack{\text{actor} \\ \text{(policy gradient)}}}{Q^{\pi_\theta}(s,a)} - \underset{\substack{\text{critic} \\ \text{(value iteration / policy iteration)}}}{V^{\pi_\theta}(s)} \equiv A_{\pi_\theta}(s,a) \qquad \text{best of both.}$$

effective update:

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{(s,a) \sim d^{\pi_\theta}}\left[\nabla_\theta \log \pi_\theta(a|s) \left(\underbrace{Q^{\pi_\theta}(s,a) - f(s)}_{A_{\pi_\theta}(s,a) \text{ if } f = V^{\pi_\theta}} \right) \right]$$

Analysis (sketch):

(stochastic) gradient descent guaranteed to find approx stationary point under mild conditions (even for non-convex) where size of gradient $\|\nabla J(\pi_{\hat\theta})\|$ is small.

Does small $\|\nabla J(\pi_{\hat\theta})\|$ imply optimality of $\hat\pi = \pi_{\hat\theta}$?

$$J(\pi^*) - J(\hat\pi) = E_{s\sim d^{\pi^*}}\left[Q^{\hat\pi}(s,\underset{=}{\pi^*}) - V^{\hat\pi}(s)\right]$$ Performance Diff^n lemma.

$$\le E_{s\sim d^{\pi^*}}\left[Q^{\hat\pi}(s,\pi_{Q^{\hat\pi}}) - V^{\hat\pi}(s)\right]$$

$$\le \left\|\frac{d^{\pi^*}}{d^{\hat\pi}}\right\|_\infty E_{s\sim d^{\hat\pi}}\left[Q^{\hat\pi}(s,\pi_{Q_{\hat\pi}}) - V^{\hat\pi}(s)\right]$$

$$= \underbrace{\left\|\frac{d^{\pi^*}}{d^{\hat\pi}}\right\|_\infty}_{1)} \underbrace{E_{s\sim d^{\hat\pi}}\left[\sum_a Q^{\hat\pi}(s,a)\left(\pi_{Q_{\hat\pi}}(a|s) - \hat\pi(a|s)\right)\right]}_{2)}$$

1) need to ensure $\hat\pi$ covers π_*

$\Rightarrow$ need to ensure initalization distribution covers π_*

$s_1 \sim d_0$

exploration $s_1 \sim d$

eg uniform

2) Does small $\|\nabla J(\hat\pi)\|$ imply small $\|\pi_{Q_{\hat\pi}} - \hat\pi\|$?

possible under smoothness

eg. tabular, linear.

$$\text{Regret} \sim \frac{d^{\pi^*}}{d^{\pi_0}}, |S|, |A|$$

Variance Reduction

- Q version of REINFORCE
- Actor, critic
- TRPO (Trust Region Policy optimization) - second order
- PPO (Proximal policy optimization) - first order.

PPO-clip $$\frac{\pi_\theta(a|s)}{\pi_{orig}(a|s)} = g_\theta$$

$$E\left[\log \pi_\theta(a|s)\ A_{\pi_\theta}(s,a)\right]$$