1) PG with value functions

claim: $s,a \sim d_{\mu_0}^{\pi_\theta}$

$$g = \frac{1}{1-\gamma} \nabla_\theta \log(\pi_\theta(a|s)) \left[Q^{\pi_\theta}(s,a) - b(s) \right]$$

(action-independent baseline)

is an unbiased estimate of $\nabla J(\theta)$

proof:

$$\nabla J(\theta) = \nabla_\theta \mathop{\mathbb{E}}_{s_0 \sim \mu_0} [V^{\pi_\theta}(s_0)]$$

skipping steps (lecture notes)

$$= \mathop{\mathbb{E}}_{\substack{s_0 \sim \mu_0 \\ a_0 \sim \pi_\theta(s_0)}} [\nabla_\theta \log \pi_\theta(a_0|s_0) Q^{\pi_\theta}(s_0,a_0)] + \gamma \mathop{\mathbb{E}}_{s_1 \sim P_1^{\pi_\theta}} [\nabla_\theta V^{\pi_\theta}(s_1)]$$

↓ iteration

marginalized over states

$$\nabla J(\theta) = \sum_{t=0}^{\infty} \gamma^t \mathop{\mathbb{E}}_{s_t, a_t \sim P_t^{\pi_\theta}} [\nabla_\theta \log \pi_\theta(a_t|s_t) Q^{\pi_\theta}(s_t,a_t)]$$

$$= \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s,a \sim d_{\mu_0}^{\pi_\theta}} [\nabla_\theta [\log \pi_\theta(a|s)] Q^{\pi_\theta}(s,a)]$$

"Actor Critic" Algorithm

1) $s,a \sim d_{\mu_0}^{\pi_{\theta_t}}$

2) $g = \frac{1}{1-\gamma} \nabla_\theta \log(\pi_\theta(a|s)) \underbrace{(Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s))}_{A^{\pi_\theta}(s,a)}$

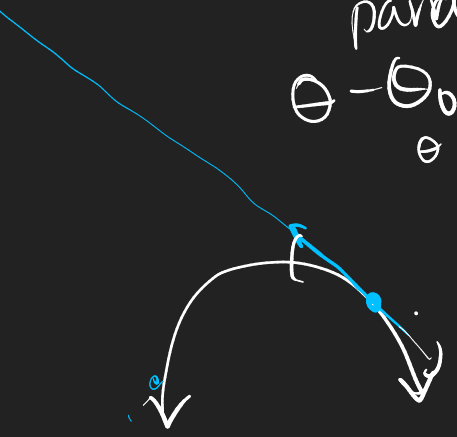
2) Trust Regions & KL-Divergence

$$\max_\theta J(\theta) \approx \left[\max_\theta J(\theta_0) + \underbrace{\nabla J(\theta_0)^T(\theta-\theta_0)}_{\text{first order approx}}\right]$$

The maximum occurs when $\theta-\theta_0$ is parallel to $\nabla J(\theta_0)$

$$\theta-\theta_0 = \alpha \nabla J(\theta_0) \quad \alpha>0$$

$$\theta = \theta_0 + \alpha \nabla J(\theta_0)$$



Choose small enough α so that the linearization is good enough

A <u>trust region</u> approach makes this intuition formal:

$$\left[\begin{array}{l}\max_\theta J(\theta) \\ \text{s.t. } d(\theta,\theta_0) < \delta\end{array}\right.$$

← bounded "distance" from θ_0

Another motivation: if $J(\theta)$ is <u>estimated</u> it might only be a good approx. when θ is near θ_0

e.g. conservative policy iteration

$$\pi'(s) = \text{argmax}_a \hat{Q}(s,a)$$

$$\pi^{t+1}(\cdot|s) = (1-\alpha)\pi^t(\cdot|s) + \alpha\pi'(s)$$

<u>KL Divergence</u>

way to measure "distance" between θ_0 and θ

<u>KL Divergence</u>: for $P \in \Delta(X)$ and $Q \in \Delta(X)$

$$KL(P|Q) = \mathop{\mathbb{E}}_{x\sim P}\left(\log\left(\frac{P(x)}{Q(x)}\right)\right) = \sum_{x\in X} P(x)\log\left(\frac{P(x)}{Q(x)}\right)$$

example: $P = \mathcal{N}(\mu_1, \sigma^2 I)$ $\quad Q = \mathcal{N}(\mu_2, \sigma^2 I)$

$$KL(P|Q) = \frac{\|\mu_1 - \mu_2\|_2^2}{\sigma}$$

Define our "distance" in terms of the KL divergence between $\pi_{\theta_0}(\cdot|s) \in \Delta(\mathcal{A})$ and $\pi_\theta(\cdot|s) \in \Delta(\mathcal{A})$ on average over $d_{\mu_0}^{\pi_{\theta_0}}$

$$d_{KL}(\theta_0, \theta) = \mathbb{E}_{s \sim d_{\mu_0}^{\pi_{\theta_0}}(s)} \left[KL\left(\pi_{\theta_0}(\cdot|s) \,\middle|\, \pi_\theta(\cdot|s)\right)\right]$$

marginalized over actions

$$= \mathbb{E}_{s,a \sim d_{\mu_0}^{\pi_{\theta_0}}} \left[\log\left(\frac{\pi_{\theta_0}(a|s)}{\pi_\theta(a|s)}\right)\right]$$

3) Natural Policy Gradient

Alg: Natural PG

Initiate θ_0

for $t = 0, 1, \dots$

Estimating $\nabla J(\theta_t)$ with g_t ← trajectory, Q, A functions based

Estimate Fisher Information matrix by

$$F_t = \nabla_\theta \log\left(\pi_{\theta_t}(a|s)\right) \nabla_\theta \log\left(\pi_{\theta_t}(a|s)\right)^T$$

$$s,a \sim d_\mu^{\pi_{\theta_t}}$$

Natural gradient step ← pseudoinverse

$$\theta_{t+1} = \theta_t + \alpha F_t^{+} g_t$$

The gradient is <u>preconditioned</u> by the Fisher Info. matrix

Derive as an approximation to constrained optimization:

$$\begin{cases} \max_\theta J(\theta) & \longrightarrow \text{first order approx} \\ \text{s.t. } d_{KL}(\theta_0, \theta) \le \delta & \longrightarrow \text{second order approx} \end{cases}$$

Second order approx

$$\ell(\theta) = d_{KL}(\theta_0, \theta) = \mathop{\mathbb{E}}_{s,a \sim d_{\mu_0}^{\pi_{\theta_0}}} \left(\log \left(\frac{\pi_{\theta_0}(a|s)}{\pi_\theta(a|s)} \right) \right)$$

$$\ell(\theta) \approx \ell(\theta_0) + \nabla \ell(\theta_0)^T (\theta - \theta_0) + (\theta - \theta_0)^T \nabla^2 J(\theta_0)(\theta - \theta_0)$$

Claim: $\ell(\theta_0) = 0$! $\quad \nabla \ell(\theta_0) = 0$

$$\nabla^2 \ell(\theta_0) = \mathop{\mathbb{E}}_{s,a \sim d_{\mu_0}^{\pi_{\theta_0}}} \left[\nabla_\theta \log(\pi_\theta(a|s)) \nabla_\theta \log(\pi_\theta(a|s))^T \Big|_{\theta=\theta_0} \right]$$

"Fisher Information matrix" F_{θ_0}

Proof:

$$\nabla_\theta \ell(\theta) = \mathop{\mathbb{E}}_{s,a \sim d^{\pi_\theta}} \left[\nabla_\theta \left(\log \pi_{\theta_0}(a|s) - \log(\pi_\theta(a|s)) \right) \right]$$

$$\nabla_\theta \ell(\theta)\Big|_{\theta=\theta_0} = -\mathop{\mathbb{E}}_{s \sim d^{\pi_{\theta_0}}(s)} \left[\sum_{a \in A} \pi_{\theta_0}(a|s) \frac{\nabla_\theta \pi_\theta(a|s)}{\pi_\theta(a|s)} \right] = 0 \quad \checkmark$$

Trust Region approximate maximization

$$\theta^* = \arg\max_\theta \nabla J(\theta_0)^T (\theta - \theta_0) + J(\theta_0)$$
$$\text{s.t. } (\theta - \theta_0)^T F_{\theta_0} (\theta - \theta_0) \le \delta$$

Claim: Solve in closed form:

$$\theta^* = \theta_0 + \alpha F_{\theta_0}^{-1} \nabla J(\theta_0)$$

$$\alpha = \left(\frac{\delta}{\nabla J(\theta_0)^T F_{\theta_0}^{-1} \nabla J(\theta_0)} \right)^{1/2}$$

Hint: $v = F_{\theta_0}^{1/2}(\theta - \theta_0)$ and $c = F_{\theta_0}^{1/2} \nabla J(\theta_0)$

$$\begin{bmatrix} \max & c^T v \\ \text{s.t.} & \|v\|_2^2 \leq \delta \end{bmatrix}$$

$$v = c \cdot \frac{\delta}{\|c\|_2}$$