CS 4/5789 16 March 2022 Prof Sarah Dean

Lecture 15: Policy Optimization with Trust Regions

1) Policy Gradient with value functions

PG w/ trajectories often has high variance. An alternative commonly used in practice uses an alternative estimate using Q functions.

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

$$g = \frac{1}{1-\gamma} \underbrace{\nabla_\theta \log(\pi_\theta(a|s))}_{\text{the gradient of the log-likelihood is called the score}} Q^{\pi_\theta}(s,a)$$

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

Proof:

$$\nabla J(\theta) = \nabla_\theta \mathbb{E}_{s_0 \sim \mu}\left[V^{\pi_\theta}(s_0)\right]$$ value fn. def.

$$= \mathbb{E}_{s_0 \sim \mu}\left[\nabla_\theta \mathbb{E}_{a_0 \sim \pi_\theta(s_0)}\left[Q^{\pi_\theta}(s_0, a_0)\right]\right]$$ s_0 doesn't depend on θ & Q fn. def.

$$\nabla_\theta \mathbb{E}_{a_0 \sim \pi_\theta(s_0)}\left[Q^{\pi_\theta}(s_0,a_0)\right] = \sum_{a_0 \in \mathcal{A}} \nabla_\theta \left[\pi(a_0|s_0) Q^{\pi_\theta}(s_0,a_0)\right]$$ defn. of expectation

product rule

$$= \sum_{a_0 \in \mathcal{A}} (\nabla_\theta \pi(a_0|s_0)) Q^{\pi_\theta}(s_0,a_0) + \pi(a_0|s_0) \nabla_\theta Q^{\pi_\theta}(s_0,a_0)$$

importance weighting trick; $r(s_0,a_0)$ doesn't depend on θ

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

$$\nabla J(\theta) = \mathop{\mathbb{E}}_{\substack{s_0 \sim \mu_0 \\ a_0 \sim \pi_\theta(s_0)}}\left[\nabla_\theta \log \pi_\theta(a_0|s_0)\, Q^{\pi_\theta}(s_0,a_0)\right] + \gamma \mathop{\mathbb{E}}_{s_1 \sim \mathbb{P}_1^{\pi_\theta}}\left[\nabla_\theta V^{\pi_\theta}(s_1)\right]$$

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

We can iterate!

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

expanding expectation

$$= \sum_{t=0}^{\infty} \sum_{\substack{s \in \mathcal{S} \\ a \in \mathcal{A}}} \mathbb{P}_t^{\pi_\theta}(s,a;\mu_0)\, \gamma^t \cdot \nabla_\theta \log \pi_\theta(a|s) \cdot Q^{\pi_\theta}(s,a)$$

definition of $d_{\mu_0}^{\pi_\theta}$

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

One final gradient estimate: $s, a \sim d_{\mu_0}^{\pi_\theta}$

$$g = \frac{1}{1-\gamma} \underbrace{\nabla_\theta \underbrace{\log \pi_\theta(a|s)}_{\text{log likelihood}}}_{\text{score}} \cdot \left(Q^{\pi_\theta}(s,a) - b(s)\right)$$

Baseline function $b(s)$ further helps in variance reduction. Most common $b(s) = V^{\pi_\theta}(s)$ resuts in <u>advantage function</u>-<u>based PG</u>

$$A^{\pi_\theta}(s,a) = Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s).$$

Policy gradients that use estimate value functions (Q or A) are called

"Actor critic"

policy ↗ ↖ value fn.

To show that g with a baseline is unbiased, we show that

$$\mathbb{E}_{a \sim \pi_\theta(s)}\left[\nabla_\theta \log \pi_\theta(a|s) \cdot b(s)\right] = 0$$

for any action-independent baseline.

$$\sum_a \cancel{\pi_\theta(a|s)} \cdot \frac{\nabla_\theta \pi_\theta(a|s)}{\cancel{\pi_\theta(a|s)}} \cdot b(s)$$ (expanding exp & grad)

$$= \nabla_\theta \sum_a \pi_\theta(a|s) \cdot b(s)$$ (linearity of grad)

$$= \nabla_\theta [\underbrace{1 \cdot b(s)}_{\text{doesn't depend on } \theta}] = 0$$ ($\pi(\cdot|s)$ is probability distribution)

2) Trust Regions & KL-Divergence

Recall: motivation of GA by first order approximate maximization

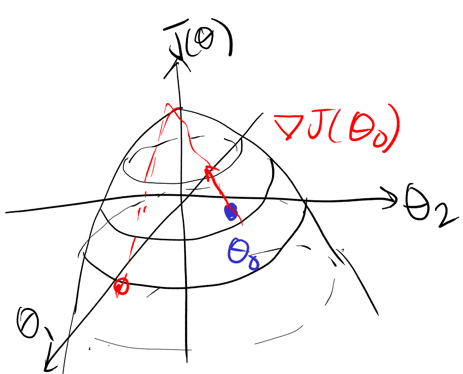
$$\max_{\theta} J(\theta) \approx \max_{\theta} J(\theta_0) + \nabla J(\theta_0)^T (\theta - \theta_0)$$

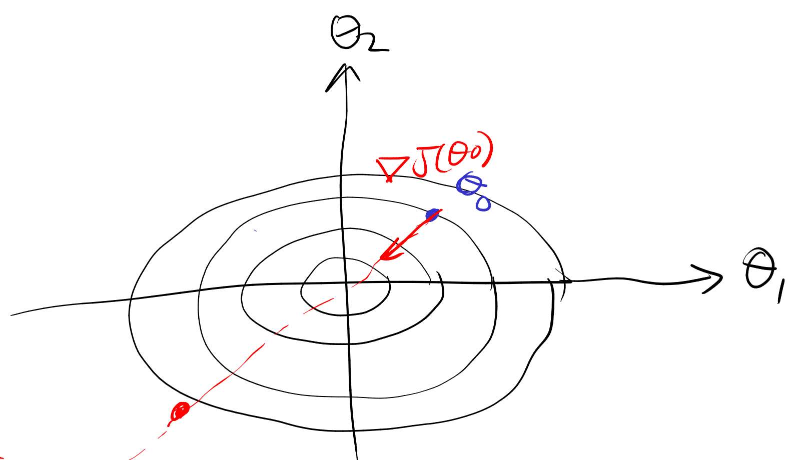
The maximum occurs when $\theta - \theta_0$ is parrallel to $\nabla J(\theta_0)$

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

Question: why do we normally use a small step size α? wouldn't as big α as possible acheive a higher maximum value?

Answer: The linear approximation is only locally valid, so by choosing small step size α, we ensure that θ is close to θ_0.





Following the gradient too far might even lead to decreasing $J(\theta)$

A trust region approach makes the intuition about the step size more precise:

$$\max_{\theta} J(\theta)$$
$$\text{s.t. } d(\theta, \theta_0) < \delta$$

trust region is described by bounded "distance" from θ_0

Another motivation for trust regions when it comes to RL: we might estimate $J(\theta)$ using data collected with θ_0 (i.e. a policy π_{θ_0}). So our estimate might only be good close to θ_0.

E.g. in Conservative policy iteration, incremental update:

$$\pi'(s) = \arg\max_a \hat{Q}(s,a)$$
$$\pi^{t+1}(\cdot|s) = (1-\alpha)\,\pi^t(\cdot|s) + \alpha\,\pi'(\cdot|s)$$

K-L Divergence:

In order to formulate a trust region problem for policy optimization, we need to decide how to measure the "distance" between θ_t and θ_{t+1}.

The K-L Divergence measures the "distance" between two distributions. Given $P \in \Delta(\mathcal{X})$ and $Q \in \Delta(\mathcal{X})$

Define (K-L Divergence)

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

Ex: if $P = \mathcal{N}(\mu_1, \sigma^2 I)$ and $Q = \mathcal{N}(\mu_2, \sigma^2 I)$ then

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

Fact: $KL(P|Q) \geq 0$ and $KL(P|Q) = 0 \iff P = Q$.

KL divergence is a natural way to constrain policy updates because it directly considers the difference in the distributions.

we define a measure of "distance" between $\pi_{\theta_0}(\cdot|s)$ and $\pi_\theta(\cdot|s)$ averaged over states s from the discounted-steady-state distribution of π_{θ_0}

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

marginalized over a

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

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

3) Natural Policy Gradient

Alg: Natural PG

initialize θ_0

for $t = 0, 1, \ldots$

Estimate $\nabla J(\theta_t)$ with g_t ← either trajectory, Q, or A function

Estimate Fisher information matrix by

$$F_t = \underbrace{\nabla \log(\pi_{\theta_t}(a|s))}\underbrace{\nabla \log(\pi_{\theta_t}(a|s))^T}_{\text{score}} \text{ for } s,a \sim d_\mu^{\pi_{\theta_t}}$$

Natural Gradient step:

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

The gradient is <u>preconditioned</u> by the Fischer information matrix.

Derive as approximating constrained optimization

$$\max_\theta J(\theta)$$ → Gradient Ascent: first order approx

$$\text{s.t. } d_{KL}(\theta_0, \theta) \leq \delta$$ → idea: second order approx!

A second order approximation to the divergence

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

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

Claim: $\ell(\theta_0) = 0$, $\nabla \ell(\theta_0) = 0$, and

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

fischer information matrix F_{θ_0}

Proof: $\ell(\theta_0) = KL(\rho_{\theta_0} | \rho_{\theta_0}) = 0$ ✓

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

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

$$\nabla \ell(\theta_0) = \mathbb{E}_{s \sim d_{\mu_0}^{\pi_{\theta_0}}} \left[\sum_a \cancel{\pi_{\theta_0}(a|s)} \cdot \frac{-\nabla \pi_{\theta_0}(a|s)}{\cancel{\pi_{\theta_0}(a|s)}} \right]$$

$$= - \mathbb{E}_{s} \left[\nabla_\theta \underbrace{\sum_a \pi_\theta(a|s)}_{=1} \Big|_{\theta=\theta_0} \right]$$

$$= - \mathbb{E}_{s} [\nabla_\theta (1)] = 0$$

✓

$$\nabla^2 \ell(\theta) = \mathop{\mathbb{E}}_{s, a \sim d_\mu^{\pi_\theta}} \left[\frac{-\nabla_\theta^2 \pi_\theta(a|s)}{\pi_\theta(a|s)} + \frac{\nabla_\theta \pi_\theta(a|s) \nabla_\theta \pi_\theta(a|s)^T}{\pi_\theta(a|s)^2} \right]$$

$$\nabla^2 \ell(\theta_0) = \underbrace{\mathop{\mathbb{E}}_{s} \sum_a \cancel{\pi_{\theta_0}(a|s)} \frac{-\nabla_{\theta_0}^2 \pi_{\theta_0}(a|s)}{\cancel{\pi_{\theta_0}(a|s)}}}_{\text{0 by same logic as above}} + \mathop{\mathbb{E}}_{s, a \sim d} \left[\nabla \log(\pi_{\theta_0}(a|s)) \nabla \log(\pi_{\theta_0}(a|s))^T \right]$$

$\square$

Therefore, the Trust Region constrained approximate maximization:

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

<u>Claim</u>: This maximization can be solved in closed form:

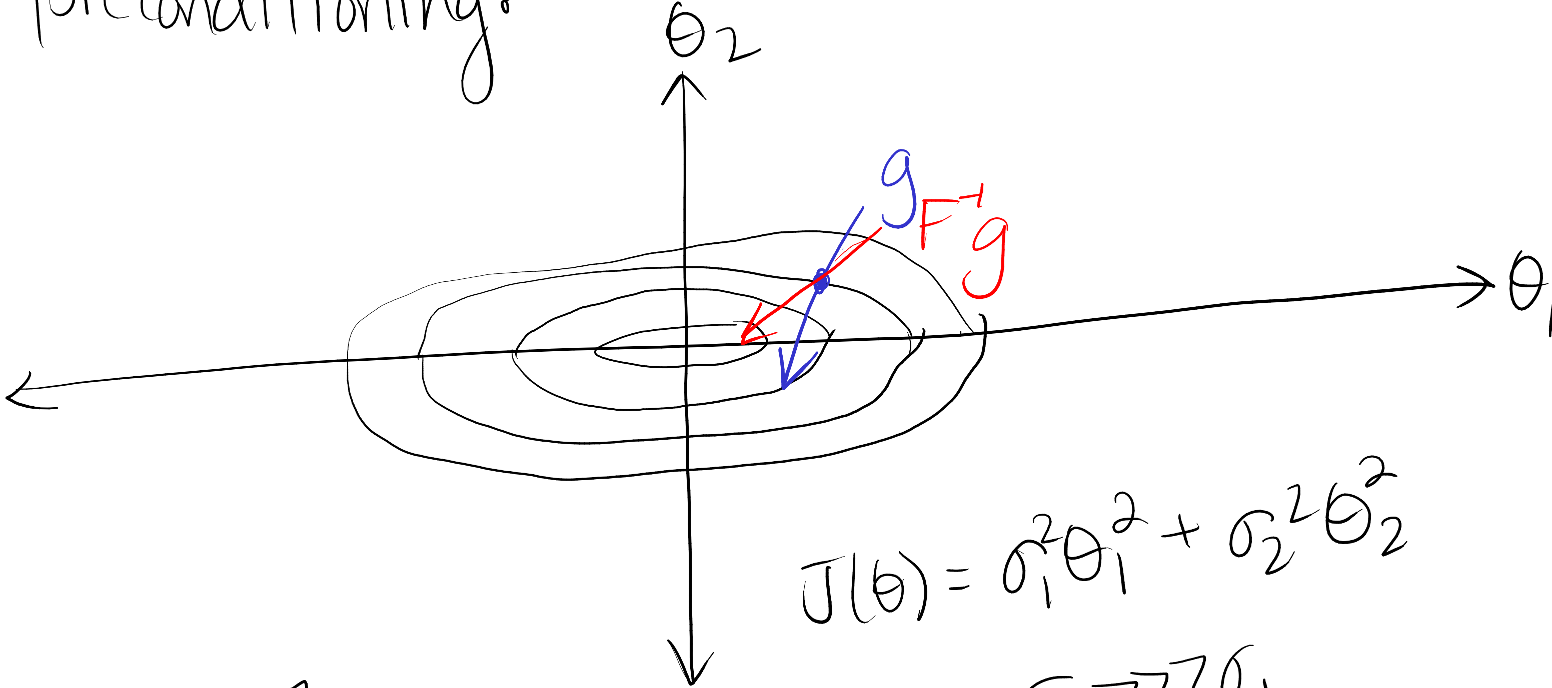
$$\theta = \theta_0 + \alpha F_{\theta_0}^{-1} \nabla J(\theta_0)$$

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

Exercise: show that this is true.

Hint: let $v = F_{\theta_0}^{1/2}(\theta - \theta_0)$ and $c = F_{\theta_0}^{-1/2} \nabla J(\theta_0)$ and consider $\max c^T v$ s.t. $\|v\|_2^2 \le \delta$

Intuitive explaination of the benefit preconditioning:



$$J(\theta) = \sigma_1^2\theta_1^2 + \sigma_2^2\theta_2^2$$

$$\sigma_2 \ggg \sigma_1$$

Steep along vertical axis. Preconditioning by $F = \begin{bmatrix} \sigma_1 & \\ & \sigma_2 \end{bmatrix}$ accounts for this and adjusts the stepsize on θ_1 vs. θ_2.