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 QA functions.

Claim: for $s, a \sim \pi^\theta_{y_0}$,

$$g = \frac{1}{1 - \chi} \nabla_\theta \log(\pi^\theta(a|s)) \ Q^\pi_{\theta}(s, a)$$

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

Proof: $\nabla J(\theta) = \nabla_\theta \mathbb{E}_{S_0 \sim \mathcal{Y}} [V^{\pi_\theta}(S_0)]$

$$= \mathbb{E}_{S_0 \sim \mathcal{Y}} \nabla_\theta \mathbb{E}_{a_0 \sim \pi^\theta_{y_0}(S_0)} [Q^{\pi_\theta}(S_0, a_0)]$$

$$\nabla_\theta \mathbb{E}_{a_0 \sim \pi^\theta_{y_0}(S_0)} [Q^{\pi_\theta}(S_0, a_0)] = \sum_{a_0 \in \mathcal{A}} \nabla_\theta \mathbb{E}_{S_0 \sim \mathcal{Y}} [\pi^\theta(a_0|S_0) Q^{\pi_\theta}(S_0, a_0)]$$

Product rule

$$= \sum_{a_0 \in \mathcal{A}} \mathbb{E}_{S_0 \sim \mathcal{Y}} [\nabla_\theta \pi^\theta(a_0|S_0)] Q^{\pi_\theta}(S_0, a_0) + \pi^\theta(S_0, a_0) \nabla_\theta Q^{\pi_\theta}(S_0, a_0)$$

$$= \mathbb{E}_{a_0 \sim \pi^\theta_{y_0}(S_0)} [\nabla_\theta \log(\pi^\theta(a_0|S_0)) Q^{\pi_\theta}(S_0, a_0)] + \chi \mathbb{E}_{S_1 \sim P(S_0, a_0)} \mathbb{E}_{a_0 \sim \pi^\theta_{y_0}(S_0)} [
abla_\theta V^{\pi_\theta}(S_1)]$$
\[ \nabla J(\theta) = \mathbb{E}\left[ \nabla_\theta \log \pi_\theta(a_t|s_t) \cdot Q_{\pi_\theta}(s_t, a_t) \right] + \gamma \mathbb{E}\left[ \nabla_\theta V_{\pi_\theta}(s_{t+1}) \right] \]

We can iterate.

\[ \nabla J(\theta) = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}\left[ \nabla_\theta \log \pi_\theta(a_t|s_t) \cdot Q_{\pi_\theta}(s_t, a_t) \right] \]

Expanding expectation

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

Definition of \( \Delta_{\mu_0} \)

\[ \frac{1}{\gamma} \mathbb{E}\left[ \nabla_\theta \log \pi_\theta(a|s) \cdot Q_{\pi_\theta}(s, a) \right] \]

One final gradient estimate: \( s, a \sim \Delta_{\mu_0} \)

\[ g = \frac{1}{1-\gamma} \frac{\nabla_\theta \log \pi_\theta(a|s) \cdot (Q_{\pi_\theta}(s, a) - b(s))}{\text{score}} \]

Baseline function \( b(s) \) further helps in variance reduction. Most common \( b(s) = V_{\pi_\theta}(s) \) results in advantage function-based PG

\[ A_{\pi_\theta}(s, a) = Q_{\pi_\theta}(s, a) - V_{\pi_\theta}(s) \]

Policy gradients that use estimate value \( Q_{\pi_\theta}(A) \) functions are called "Actor critic" policy value fn.
To show that $q$ with a baseline $\bar{b}(s)$ is unbiased, we show that

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

for any action-independent baseline.

$$\sum_a \pi^\theta(a|s) \cdot \frac{\nabla_\theta \pi^\theta(a|s)}{\pi^\theta(a|s)} \cdot \bar{b}(s)$$

(expanding exp & grad)

$$= \nabla_\theta \sum_a \pi^\theta(a|s) \cdot \bar{b}(s)$$

(linearity of grad)

$$= \nabla_\theta \left[ 1 \cdot \bar{b}(s) \right] = 0$$

($\pi^\theta(s)$ is probability distribution)

doesn't depend on $\theta$
2) Trust Regions & KL-Divergence

Recall: motivation of GGA 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 parallel to $\nabla J(\theta_0)$

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

Question: why do we normally use a small step size $\alpha$? Wouldn't a big $\alpha$ lead to possible achieving a higher maximum value?

Answer: the linear approximation is only locally valid, so by choosing small step size $\alpha$, we ensure that $\theta$ is close to $\theta_0$.

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:

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

trust region is described by bounded "distance" from $\theta_0$
Another motivation for trust regions when it comes to RL: we might estimate \( J(\theta) \) using data collected with \( \Theta_0 \) (i.e. a policy \( \pi_{\theta_0} \)). So our estimate might only be good close to \( \Theta_0 \).

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

\[
\pi'_{t+1}(s) = \frac{\pi_{\theta_t}(s)}{\sum_a \pi_{\theta_t}(s,a)} \\
\pi_{\theta_{t+1}}(s) = (1-\alpha) \pi_{\theta_t}(s) + \alpha \pi'_{t+1}(s)
\]

**K-L Divergence**: 
In order to formulate a trust region problem for policy optimization, we need to decide how to measure the "distance" between \( \Theta_t \) and \( \Theta_{t+1} \).

The **K-L Divergence** measures the "distance" between two distributions. Given \( P \in \Delta(X) \) and \( Q \in \Delta(X) \) define (K-L Divergence)

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

Ex: if \( P = N(y_1, \sigma^2 I) \) and \( Q = N(y_2, \sigma^2 I) \) then

\[
KL(P \| Q) = \frac{1}{2} \log \left( \frac{(\sigma^2)^2}{\sigma^2} \right)
\]

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 $\Pi_{\theta_0}(s)$ and $\Pi_0(s)$ averaged over states $s$ from the discounted-steady-state distribution of $\Pi_{\theta_0}$.

$$d_{KL}(\theta_0, \theta) = \mathbb{E} \left[ KL(\Pi_{\theta_0}(s) \| \Pi_0(s)) \right]$$

$$\sum_{s \sim \Pi_{\theta_0}(s)} \uparrow$$

Marginalized over $a$.

$$= \mathbb{E} \left[ \mathbb{E} \left[ \log \left( \frac{\Pi_{\theta_0}(a(s))}{\Pi_0(a(s))} \right) \right] \right]$$

$$\sum_{s \sim \Pi_{\theta_0}(s)} \sum_{a \sim \Pi_0(s)}$$

$$= \mathbb{E} \log \left( \frac{\Pi_{\theta_0}(a(s))}{\Pi_0(a(s))} \right)$$
Algorithm: Natural PG

1. Initialize $\theta_0$
2. for $t = 0, 1, \ldots$
   - Estimate $\nabla J(\theta_t)$ with $g_t$
   - Estimate Fisher information matrix by
     \[ F_t = \nabla \log(T_{\theta_{t-1}}(a|s)) \cdot \nabla \log(T_{\theta_{t-1}}(a|s)) \] for $s, a, c_{\theta_{t-1}}$
   - Natural Gradient step:
     \[ \theta_{t+1} = \theta_t + \alpha F_t^{-1} g_t \]

The gradient is **preconditioned** by the Fisher information matrix.

---

Derive as approximating constrained optimization

\[
\max_{\theta} J(\theta) \quad \text{Gradient ascent: first order approx}
\]

subject to \[ d_{KL}(\theta_0, \theta) \leq \delta \quad \text{Idea: second order approx!} \]
A second order approximation to the divergence

\[
e(\Theta) = \mathbb{E} \left[ \log \left( \frac{\prod_{\Theta_0} (a|s)}{\prod_{\Theta} (a|s)} \right) \right]_{s,a \sim \prod_{\Theta_0}}
\]

\[
e(\Theta) \approx e(\Theta_0) + \nabla e(\Theta_0)^T (\Theta - \Theta_0) + (\Theta - \Theta_0)^T \nabla^2 e(\Theta_0) (\Theta - \Theta_0)
\]

Claim: \( e(\Theta_0) = 0, \ \nabla e(\Theta_0) = 0, \) and

\[
\nabla^2 e(\Theta_0) = \mathbb{E} \left[ \nabla_{\Theta} \log (\prod_{\Theta_0} (a|s)) \nabla_{\Theta} \log (\prod_{\Theta_0} (a|s))^T \right]_{\Theta = \Theta_0}
\]

Fischer information matrix

\( F_{\Theta_0} \)

Proof:

\[
e(\Theta_0) = KL(\rho_{\Theta_0} \| \rho_{\Theta_0}) = 0
\]

\[
\nabla e(\Theta) = \mathbb{E} \left[ \nabla_{\Theta} \left( \log \prod_{\Theta_0} (a|s) - \log \prod_{\Theta_0} (a|s) \right) \right]_{s,a \sim \prod_{\Theta_0}, \prod_{\Theta_0}}
\]

\[
= \mathbb{E} \left( - \frac{\nabla_{\Theta} \prod_{\Theta_0} (a|s)}{\prod_{\Theta_0} (a|s)} \right)
\]

\[
\nabla e(\Theta_0) = \mathbb{E} \left[ \sum_{a} \frac{\prod_{\Theta_0} (a|s)}{\prod_{\Theta_0} (a|s)} \right.\left. - \frac{\nabla \prod_{\Theta_0} (a|s)}{\prod_{\Theta_0} (a|s)} \right]_{s \sim \prod_{\Theta_0}, \prod_{\Theta_0}}
\]

\[
= - \mathbb{E} \left[ \nabla_{\Theta} \sum_{a} \prod_{\Theta_0} (a|s) \right]_{\Theta = \Theta_0} \bigg|_{\Theta = \Theta_0}
\]

\[
= - \mathbb{E} \left[ \nabla_{\Theta} (1) \right] = 0
\]
\[ \mathcal{D}^2 l (\theta) = \mathbb{E} \left[ \frac{\nabla^2 \log \pi_\theta (a | s)}{\pi_\theta (a | s)} + \frac{\nabla \log \pi_\theta (a | s) \nabla \log \pi_\theta (a | s)^T}{\pi_\theta (a | s)^2} \right] \]

\[ \mathcal{D}^2 l (\theta_0) = \mathbb{E} \sum_{s,a} \pi_\theta (a | s) \nabla^2 \log \pi_\theta (a | s) + \mathbb{E} \left[ \nabla \log \pi_\theta (a | s) \nabla \log \pi_\theta (a | s)^T \right] \]

Therefore, the Trust Region constrained approximate maximization:

\[
\max_{\theta} \mathcal{D} J (\theta_0)^T (\theta - \theta_0) \\
\text{s.t. } (\theta - \theta_0)^T F_{\theta_0} (\theta - \theta_0) \leq \delta 
\]

Claim: This maximization can be solved in closed form:

\[
\theta = \theta_0 + \alpha F_{\theta_0}^{-1} \nabla J (\theta_0) \\
\text{where } \alpha = \left( \frac{\delta}{\nabla J (\theta_0)^T F_{\theta_0}^{-1} \nabla J (\theta_0)} \right)^{1/2}
\]

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 \( \max_{CV} \text{ s.t. } \| CV \|_2 \leq \delta \)
Intuitive explanation of the benefit of preconditioning:

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

\[ \sigma_2 \rightarrow 17 \sigma_1 \]

Steep along vertical axis. Preconditioning by

\[ F = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \]

accounts for this and adjusts the stepsizes on \( \theta_1 \) vs. \( \theta_2 \).