DS 598
Introduction to RL

Xuezhou Zhang

Chapter 5: Policy-based RL (continued)

The REINFORCE algorithm

1. Initialize θ_0
2. For iteration t = 0,…,T
   1) Run π_{θ_t} and collect trajectories $\tau_1, \dots, \tau_n$
   2) Estimate the PG by

$$g_t = \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{h=0}^{\infty} \nabla_\theta \log \pi(a_{i;h}|s_{i;h}) R(\tau_i)\right]$$

   3) Do SGD update $\theta_{t+1} = \theta_t + \alpha_t g_t$

The REINFORCE algorithm

$$g_t = \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{h=0}^{\infty}\nabla_\theta \log \pi(a_{i;h}|s_{i;h})R(\tau_i)\right]$$

A couple of techniques to improve PG estimation:

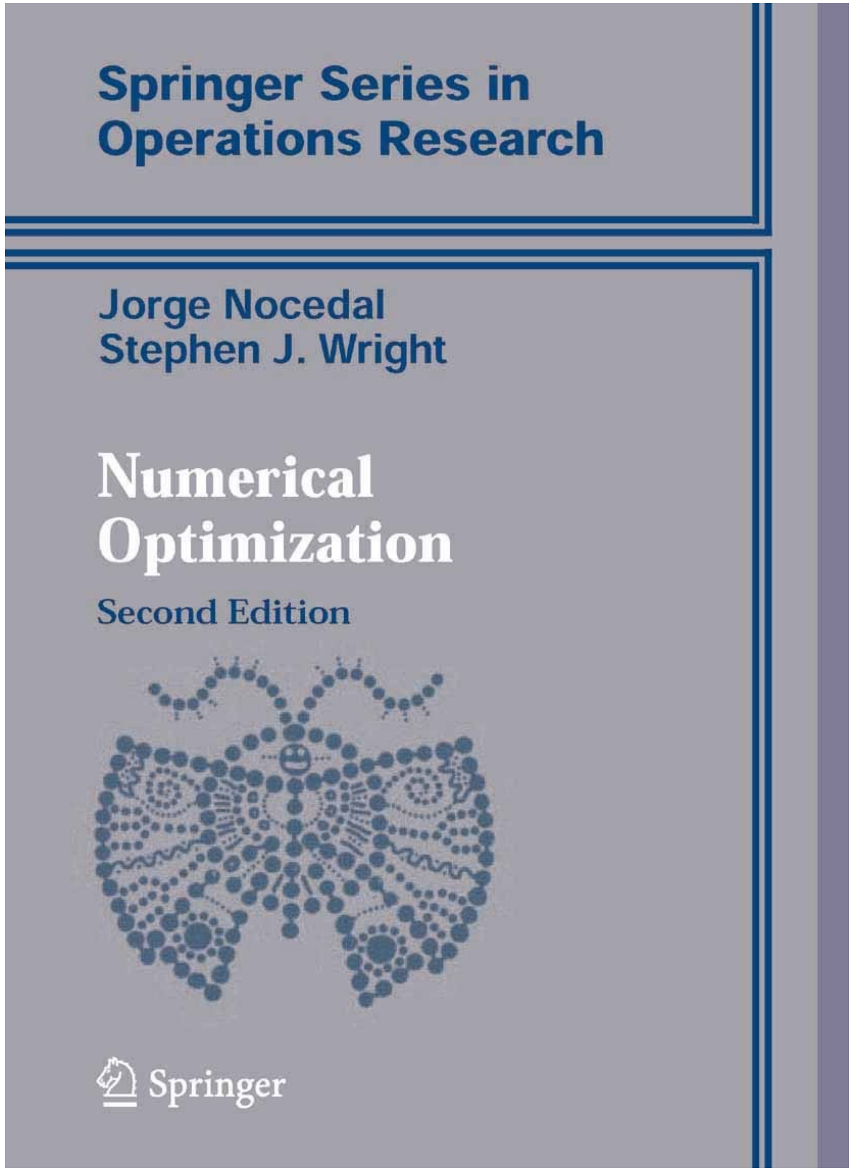
1. Baseline: variance reduction
2. Critic: off-policy learning of value function
3. Importance Sampling: off-policy estimation of PG
4. Deterministic PG: handles continuous and deterministic policy

An Optimization Viewpoint

- Numerical Optimization. Jorge Nocedal , Stephen J. Wright (2006)

- Highly recommended!

- Pillars of ML: statistics, calculus and linear algebra, numerical optimization.



An Optimization Viewpoint

Given a function $f(x)$, find $\mathrm{argmin}_x\, f(x)$.

- Approximation-based Optimization

1. Starting at some x_0.
2. For iteration k=0,2, …
   1) Find a local approximation $\hat{f}_k$ that can be minimized with less effort than f itself.
   2) Set $x_{k+1} = \mathrm{argmin}_{x \in \mathcal{X}}\, \hat{f}_k(x)$.

- Example 1:
- $\hat{f}_k(x) = f(x_k) + (x - x_k)^\top \cdot \nabla f(x_k) + \frac{1}{2t_k}||x - x_k||^2$
- $x_{k+1} = x_k - t_k \nabla f(x_k)$.
- This is gradient descent!

An Optimization Viewpoint

Given a function $f(x)$, find $\mathrm{argmin}_x\ f(x)$.

- Approximation-based Optimization

1. Starting at some x_0.
2. For iteration k=0,2, …
   1) Find a local approximation $\hat{f}_k$ that can be minimized with less effort than f itself.
   2) Set $x_{k+1} = \mathrm{argmin}_{x \in \mathcal{X}}\ \hat{f}_k(x)$.

- Example 2:
- $\hat{f}_k(x) = f(x_k) + (x - x_k)^\top \cdot \nabla f(x_k) + \frac{1}{2}(x - x_k)^\top H_k (x - x_k)$
- where H_k is the Hessian of f at x_k.
- $x_{k+1} = x_k - H_k^{-1} \nabla f(x_k)$.
- This is the Newton's method!

An Optimization Viewpoint

Given a function $f(x)$, find $\mathrm{argmin}_x\, f(x)$.

- Approximation-based Optimization

1. Starting at some x_0.
2. For iteration k=0,2, …
   1) Find a local approximation $\hat{f}_k$ that can be minimized with less effort than f itself.
   2) Set $x_{k+1} = \mathrm{argmin}_{x \in \mathcal{X}}\, \hat{f}_k(x)$.

- Problem?

- $\hat{f}_k$ will be a poor approximation of f far away from x_k.

- Solution: Sanity check on x_{k+1} before continue.

An Optimization Viewpoint

Given a function $f(x)$, find $\mathrm{argmin}_x\ f(x)$.

- Trust-region Method

1. Starting at some x_0.
2. For iteration k=0,2, …
   1) Find a local approximation $\hat{f}_k$.
   2) Choose a trust region U_k containing x_k, e.g.
      $U_k = \left\{x: ||x - x_k||_k \le \Delta_k\right\}$
   3) Set $x_{k+1} = \mathrm{argmin}_{x \in \mathrm{U_k}}\ \hat{f}_k(x)$.
   4) Sanity check: if $f(x_{k+1}) - f(x_k)$ is sufficiently large, continue; else, set $\Delta_k \leftarrow \epsilon_k \Delta_k$ and loop back to step 2.

- Example 1:
- $\hat{f}_k(x) = f(x_k) + (x - x_k)^\top \cdot \nabla f(x_k)$
- $U_k = \left\{x: \frac{1}{2}||x - x_k||_2^2 \le \delta^2\right\}$
- $x_{k+1} = x_k - \delta \frac{\nabla f(x_k)}{||\nabla f(x_k)||}$.
- Normalized gradient descent.
- Better distance metric?

An Optimization Viewpoint

Given a function $f(x)$, find $\mathrm{argmin}_x\, f(x)$.

- Trust-region Method

1. Starting at some x_0.
2. For iteration k=0,2, …
   1) Find a local approximation $\hat{f}_k$.
   2) Choose a trust region U_k containing x_k, e.g. $U_k = \left\{x: \left\|x - x_k\right\|_k \leq \Delta_k\right\}$
   3) Set $x_{k+1} = \mathrm{argmin}_{x\in \mathrm{U_k}}\, \hat{f}_k(x)$.
   4) Sanity check: if $f(x_{k+1}) - f(x_k)$ is sufficiently large, continue; else, set $\Delta_k \leftarrow \epsilon_k \Delta_k$ and loop back to step 2.

- Better distance metric?
- Linear model
- $U_k = \left\{x: \frac{1}{2}(x - x_k)^\top F_k (x - x_k) \leq \delta^2\right\}$
- $x_{k+1} = x_k - D_k \nabla f(x_k)$.
- where $D_k = \dfrac{\delta F^{-1}(x_k)}{\sqrt{\nabla f^\top (x_k) F^{-1}(x_k) \nabla \mathrm{f}(\mathrm{x_k})}}$.
- Damped Newton's Method ($F_k = H_k$)

An Optimization Viewpoint

Given a function $f(x)$, find $\text{argmin}_x\ f(x)$.

- Trust-region Method

1. Starting at some x_0.
2. For iteration k=0,2, …
   1) Find a local approximation $\hat{f}_k$.
   2) Choose a trust region U_k containing x_k, e.g.
      $$U_k = \left\{x: \left|\left|x - x_k\right|\right|_k \leq \Delta_k\right\}$$
   3) Set $x_{k+1} = \text{argmin}_{x \in \mathrm{U_k}}\ \hat{f}_k(x)$.
   4) Sanity check: if $f(x_{k+1}) - f(x_k)$ is sufficiently large, continue; else, set $\Delta_k \leftarrow \epsilon_k \Delta_k$ and loop back to step 2.

- Design Choices:

1. What is $\hat{f}_k$?
2. What is U_k?
3. How to do sanity check?

Back to RL

- $f(\pi_\theta) = \mathbb{E}_{\pi_\theta}[\sum_{h=0}^{\infty} \gamma^h \, r(s_h, a_h)]$

- Design Choices:

1. What is $\hat{f}_k$?
2. What is U_k?
3. How to do sanity check?

RL as Optimization

- $f(\pi_\theta) = \mathbb{E}_{\pi_\theta}[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h)]$

- Design Choices:

1. What is $\hat{f}_k$? How do we approximate $f(\pi_\theta)$ with data from π_k?

- Performance Difference Lemma:

$$f(\pi) - f(\pi') = \mathbb{E}_{s,a\sim d^\pi}\left[A^{\pi'}(s,a)\right]$$

RL as Optimization

- $f(\pi_\theta) = \mathbb{E}_{\pi_\theta}[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h)]$

- Design Choices:

1. What is $\hat{f}_k$? How do we approximate $f(\pi_\theta)$ with data from π_k?

$$
\begin{aligned}
f(\pi_\theta) &= f(\pi_k) + \mathbb{E}_{s,a\sim d^{\pi}}[A^{\pi_k}(s,a)] \\
&\approx \underbrace{f(\pi_k) + \mathbb{E}_{s,a\sim d^{\pi_k}}[\frac{\pi_\theta(a|s)}{\pi_k(a|s)} A^{\pi_k}(s,a)]}_{\hat{f}_k}
\end{aligned}
$$

RL as Optimization

- $f(\pi_\theta) = \mathbb{E}_{\pi_\theta}[\sum_{h=0}^{\infty} \gamma^h \, r(s_h, a_h)]$

- Design Choices:

1. What is $\hat{f}_k$? How do we approximate $f(\pi_\theta)$ with data from π_k?

$$\hat{f}(\pi_\theta) = f(\pi_k) + \mathbb{E}_{s,a\sim d^{\pi_k}}\left[\frac{\pi_\theta(a|s)}{\pi_k(a|s)} A^{\pi_k}(s,a)\right]$$

$\hat{f}(\pi_\theta)$ satisifies $\hat{f}(\pi_k) = \mathrm{f}(\theta_k)$ and

$$\nabla_\theta \hat{f}(\pi_k) = \nabla_\theta f(\pi_k) = \mathbb{E}_{s,a\sim d^{\pi_k}}[\nabla_\theta \log \pi_k(a|s) \cdot A^{\pi_k}(s,a)]$$

RL as Optimization

- $f(\pi_\theta) = \mathbb{E}_{\pi_\theta}[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h)]$

- Design Choices:

1. What is $\hat{f}_k$? How do we approximate $f(\pi_\theta)$ with data from π_k?

$$\hat{f}(\pi_\theta) = f(\pi_k) + \mathbb{E}_{s,a\sim d^{\pi_k}}\left[\frac{\pi_\theta(a|s)}{\pi_k(a|s)} A^{\pi_k}(s,a)\right]$$

First-order Taylor expansion at θ_k

$$\hat{f}_k \approx f(\pi_k) + (\theta - \theta_k)^\top \cdot \nabla_\theta f\left(\pi_{\theta_k}\right)$$

RL as Optimization

- $f(\pi_\theta) = \mathbb{E}_{\pi_\theta}[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h)]$

Can we make smarter choices?

- Design Choices:

1. What is $\hat{f}_k$? $\hat{f}_k = f(\pi_k) + (\theta - \theta_k)^\top \cdot \nabla_\theta f\left(\pi_{\theta_k}\right)$
2. What is U_k? $U_k = \left\{\theta : \frac{1}{2}\left|\left|\theta - \theta_k\right|\right|_2^2 \leq \delta^2\right\}$
3. How to do sanity check? No sanity check.

- Then, we get $\theta_{k+1} = \theta_k + \delta \frac{\nabla f(\theta_k)}{||\nabla f(\theta_k)||}$, which is exactly Vanilla PG!

RL as Optimization

- $f(\pi_\theta) = \mathbb{E}_{\pi_\theta}[\sum_{h=0}^{\infty} \gamma^h \, r(s_h, a_h)]$

- Design Choices:

2. What is a better U_k? Or rather, what metric should we use?

    - Policies $\pi_\theta(a|s)$ are probability distributions.
    - Different θ can map to the same policy.
    - A metric in the probability space?

Kullback–Leibler (KL) divergence

- $D_{KL}(p|q) = \mathbb{E}_{x\sim p}\log\left(\frac{p(x)}{q(x)}\right)$.

- In general, $D_{KL}(p|q) \neq D_{KL}(q|p)$, so it's not a metric.
- $D_{KL}(p|q) \geq 0$.
- $p = q$ iff $D_{KL}(p|q) = D_{KL}(q|p) = 0$.

- Example: If $p = \mathcal{N}(\mu_1, \sigma I)$, $q = \mathcal{N}(\mu_2, \sigma I)$,
- then $D_{KL}(p|q) = ||\mu_1 - \mu_2||_2^2/\sigma^2$.

Kullback–Leibler (KL) divergence

- $D_{KL}(\pi_k|\pi_\theta) = \mathbb{E}_{x\sim\pi_k}\log\left(\frac{x\sim\pi_k(x)}{x\sim\pi_\theta(x)}\right).$

- Fact:

The Fisher Information Matrix

- $\nabla_\theta D_{KL}(\pi_k|\pi_\theta)|_{\theta=\theta_k} = 0$
- $H_p D_{KL}(\pi_k|\pi_\theta)|_{\theta=\theta_k} = \mathbb{E}_{x\sim\pi_k}[\nabla_\theta \log \pi_k(a|s) \nabla_\theta \log \pi_k(a|s)^\top] := \mathrm{F_k}$

- Second-order Taylor expansion at θ_k:
- $D_{KL}(\pi_k|\pi_\theta) \approx (\theta - \theta_k)^\top F_k(\theta - \theta_k)$

Putting it together

- $\theta_{k+1} = \text{argmax}_{\theta \in U_k} f(\pi_k) + (\theta - \theta_k)^\top \cdot \nabla_\theta f\left(\pi_{\theta_k}\right)$,
- where $U_k = \left\{\theta: \frac{1}{2}(\theta - \theta_k)^\top F_k(\theta - \theta_k) \leq \delta^2\right\}$.
- This implies $\theta_{k+1} = \theta_k - D_k \nabla_\theta f(\theta_k)$,
- where $D_k = \frac{\delta F^{-1}(x_k)}{\sqrt{\nabla f^\top(x_k) F^{-1}(x_k) \nabla \mathrm{f}(\mathrm{x_k})}}$.

- Again, $F_k = \mathbb{E}_{x \sim \pi_k}[\nabla_\theta \log \pi_k(a|s) \nabla_\theta \log \pi_k(a|s)^\top]$.
- This is the Trusted-region Policy Optimization (TRPO) algorithm.

Natural Policy Gradient

- An earlier appearance of an update rule similar to TRPO is called Natural Policy Gradient (NPG).
- TRPO: $\theta_{k+1} = \theta_k - D_k \nabla_\theta f(\theta_k)$
- where $D_k = \frac{\delta F^{-1}(x_k)}{\sqrt{\nabla f^\top(x_k) F^{-1}(x_k) \nabla \mathrm{f}(\mathrm{x_k})}}$.
- NPG: $\theta_{k+1} = \theta_k - \alpha F_k^{-1} \nabla_\theta f(\theta_k)$

- NPG makes a less careful choice on the step-size of the update.

RL as Optimization

- $\hat{f}(\pi_\theta) = f(\pi_k) + \mathbb{E}_{s,a\sim d^{\pi_k}}\left[\frac{\pi_\theta(a|s)}{\pi_k(a|s)} A^{\pi_k}(s,a)\right]$

- An objective-specific U_k?
- Idea: we don't want to overfit too much on $\hat{f}$.

- Proximal Policy Optimization (PPO):

$$\mathrm{sign}\left(\left(\frac{\pi_\theta(a|s)}{\pi_k(a|s)} - 1\right) A^{\pi_k}(s,a)\right) \leq \epsilon$$

RL as Optimization

- $\hat{f}(\pi_\theta) = f(\pi_k) + \mathbb{E}_{s,a\sim d^{\pi_k}}\left[\frac{\pi_\theta(a|s)}{\pi_k(a|s)} A^{\pi_k}(s,a)\right]$

- Proximal Policy Optimization (PPO):

$$\left(\frac{\pi_\theta(a|s)}{\pi_k(a|s)} - 1\right)\mathrm{sign}(A^{\pi_k}(s,a)) \le \epsilon$$

- Instead of enforce it as a constraint, PPO modifies the objective as

$$\hat{f}(\pi_\theta) = f(\pi_k) + \mathbb{E}_{s,a\sim d^{\pi_k}}\left[\min\left(\frac{\pi_\theta(a|s)}{\pi_k(a|s)} A^{\pi_k}(s,a), \mathrm{clip}_\epsilon\left(\frac{\pi_\theta(a|s)}{\pi_k(a|s)}\right) A^{\pi_k}(s,a)\right)\right]$$

Summary

- REINFORCE:
  - 1st-order Taylor approximation of the objective.
  - Trusted region with Euclidean distance.

- TRPO/NPG:
  - 1st-order Taylor approximation of the objective.
  - Trusted region with KL divergence.

- PPO:
  - 1st-order Taylor approximation of the objective.
  - Trusted region with improvement constraints in $\hat{f}$.