Quasi-Newton Methods for Saddle Point Problems

Chengchang Liu Luo Luo

September, 2022

Introduction

We focus on the minimax optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^{n_x}} \max_{\mathbf{y}\in\mathbb{R}^{n_y}} f(\mathbf{x},\mathbf{y}).$$

The applications

- Adversarial Learning
- Domain Adaption
- Fairness-Aware Machine Learning
- Few-Shot Learning
- Reinforcement Learning
- Robust Optimization
- Two-Player Games

First-Order Optimization

The assumptions

- $f(\mathbf{x}, \mathbf{y})$ is μ-strongly-convex in $\mathbf{x}$ and μ-strongly-concave in $\mathbf{y}$
- $\nabla f(\mathbf{x}, \mathbf{y})$ is L-Lipschitz continuous

The update of extragradient method is

$$\begin{cases} \mathbf{x}_{k+1/2} = \mathbf{x}_k - \eta \nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k), \\ \mathbf{y}_{k+1/2} = \mathbf{y}_k + \eta \nabla_{\mathbf{y}} f(\mathbf{x}_k, \mathbf{y}_k), \\ \mathbf{x}_{k+1} = \mathbf{x}_k - \eta \nabla_{\mathbf{x}} f(\mathbf{x}_{k+1/2}, \mathbf{x}_{k+1/2}), \\ \mathbf{y}_{k+1} = \mathbf{y}_k + \eta \nabla_{\mathbf{y}} f(\mathbf{x}_{k+1/2}, \mathbf{x}_{k+1/2}), \end{cases}$$

which has optimal linear convergence rate

$$\|\mathbf{x}_k - \mathbf{x}^*\|_2^2 + \|\mathbf{y}_k - \mathbf{y}^*\|_2^2 \leq \left(1 - \frac{\mu}{4L}\right)^k \left(\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 + \|\mathbf{y}_0 - \mathbf{y}^*\|_2^2 \right).$$

Second-Order Optimization

The assumptions

- $f(\mathbf{x}, \mathbf{y})$ is μ-strongly-convex in $\mathbf{x}$ and μ-strongly-concave in $\mathbf{y}$
- $\nabla f(\mathbf{x}, \mathbf{y})$ is L-Lipschitz continuous
- $\nabla^2 f(\mathbf{x}, \mathbf{y})$ is L_2-Lipschitz continuous

The update of Newton method:

$$\mathbf{z}_{k+1} = \mathbf{z}_k - \left(\nabla^2 f(\mathbf{z}_k)\right)^{-1} \nabla f(\mathbf{z}_k).$$

It has local quadratic convergence rate, where $\mathbf{z}_k = (\mathbf{x}_k, \mathbf{y}_k)$.

The weakness

1. computing $\left(\nabla^2 f(\mathbf{z}_k)\right)^{-1}$ requires $\mathcal{O}(n_x^3 + n_y^3)$
2. construction $\nabla^2 f(\mathbf{z}_k)$ is also expensive

Classical Quasi-Newton Methods

Classical quasi-Newton methods (BFGS, SR1, DFP...)

1. strongly convex minimization
2. avoiding construct or inverse the Hessian
3. superlinear local convergence rate

Quasi-Newton Methods for Convex Optimization

Quasi-Newton methods for

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}).$$

- Broyden family update (1970s)

$$\mathrm{Broyd}_\tau(\mathbf{G},\mathbf{H},\mathbf{u}) \overset{\text{def}}{=} \tau\left[\mathbf{G} - \frac{\mathbf{H}\mathbf{u}\mathbf{u}^\top\mathbf{G} + \mathbf{G}\mathbf{u}\mathbf{u}^\top\mathbf{H}}{\mathbf{u}^\top\mathbf{H}\mathbf{u}} + \left(\frac{\mathbf{u}^\top\mathbf{G}\mathbf{u}}{\mathbf{u}^\top\mathbf{H}\mathbf{u}} + 1\right)\frac{\mathbf{H}\mathbf{u}\mathbf{u}^\top\mathbf{H}}{\mathbf{u}^\top\mathbf{H}\mathbf{u}}\right] + (1-\tau)\left[\mathbf{G} - \frac{(\mathbf{G}-\mathbf{H})\mathbf{u}\mathbf{u}^\top(\mathbf{G}-\mathbf{H})}{\mathbf{u}^\top(\mathbf{G}-\mathbf{H})\mathbf{u}}\right]$$

- $\nabla^2 f(\mathbf{x}_{k+1}) \approx \mathbf{G}_{k+1} = \mathrm{Broyd}_\tau(\mathbf{G}_k, \mathbf{H}_{k+1}, \mathbf{u}_k)$
- $\mathbf{H}_{k+1}\mathbf{u}_k$ and $(\mathbf{G}_{k+1})^{-1}$ only require $\mathcal{O}(n^2)$

We require $\nabla^2 f(\mathbf{x})$ is positive-definite!

Quasi-Newton Methods for Convex Optimization

The different choice of τ

- SR1 method: $\tau_k = 0$
- BFGS method: $\tau_k = \dfrac{\mathbf{u}_k^\top \mathbf{H}_{k+1} \mathbf{u}_k}{\mathbf{u}_k^\top \mathbf{G}_k \mathbf{u}_k} \in [0, 1]$
- DFP method: $\tau_k = 1$

The different choice of $\mathbf{u}_k$ and $\mathbf{H}_{k+1}$:

- The classical quasi-Newton methods (1970s):

$$\mathbf{u}_k = \mathbf{x}_{k+1} - \mathbf{x}_k, \quad \mathbf{H}_{k+1} = \int_0^1 \nabla^2 f(\mathbf{x}_k + t\mathbf{u}_k)\,\mathrm{d}t$$

$$\Longrightarrow \mathbf{H}_{k+1}\mathbf{u}_k = \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k).$$

- The greedy quasi-Newton methods (2020):

$$\mathbf{u}_k = \mathop{\arg\max}_{\mathbf{u} \in \{\mathbf{e}_1, \dots, \mathbf{e}_n\}} \frac{\mathbf{u}^\top \mathbf{G}_k \mathbf{u}}{\mathbf{u}^\top \mathbf{H}_{k+1} \mathbf{u}}, \quad \mathbf{H}_{k+1} = \nabla^2 f(\mathbf{x}_{k+1})$$

- The random quasi-Newton methods (2020)

$$\mathbf{u}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \mathbf{H}_{k+1} = \nabla^2 f(\mathbf{x}_{k+1}).$$

Quasi-Newton Methods for Convex Optimization

We denote $\hat{\lambda}_k = \langle \nabla f(\mathbf{x}), (\nabla^2 f(\mathbf{x}))^{-1} \nabla f(\mathbf{x}) \rangle$.

Local superlinear convergence:

- The classical quasi-Newton methods
  - 1970s: $\lim\limits_{k\to\infty} \dfrac{\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2}{\|\mathbf{x}_k - \mathbf{x}^*\|_2} = 0$
  - 2020 : $\hat{\lambda}_k \leq \mathcal{O}\Big(\left(\frac{L^2}{\mu^2 k}\right)^{k/2} \Big)$
- The greedy/random quasi-Newton methods
  - 2020: $\hat{\lambda}_{k+k_0} \leq \mathcal{O}\left(\left(1 - \frac{\mu}{L}\right)^{k_0} \left(1 - \frac{\mu}{Ln}\right)^{k(k-1)/2} \right)$
  - 2021: $\hat{\lambda}_{k+k_0} \leq \mathcal{O}\left(\left(1 - \frac{\mu}{L}\right)^{k_0} \left(1 - \frac{1}{n}\right)^{k(k-1)/2} \right)$

We require $\nabla^2 f(\mathbf{x})$ is positive-definite!

Quasi-Newton Methods for Minimax Optimization

In 1970s, the asymptotic rates is based on Dennis–Moré Theorem:

$$\lim_{k\to\infty} \frac{\left\|(\mathbf{G}_k - \nabla^2 f(\mathbf{x}^*))(\mathbf{x}_{k+1} - \mathbf{x}_k)\right\|_2}{\left\|\mathbf{x}_{k+1} - \mathbf{x}_k\right\|_2} = 0.$$

In 2020s, the non-asymptotic rates is based on

1. (BFGS) $\sigma_k = \operatorname{tr}\big((\mathbf{G}_k - \nabla^2 f(\mathbf{x}_k))((\nabla^2 f(\mathbf{x}_k))^{-1}\big)$,
2. (SR1) $\tau_k = \operatorname{tr}(\mathbf{G}_k - \nabla^2 f(\mathbf{x}_k))$

converge to zero, which means $\mathbf{G}_k$ converges to $\nabla^2 f(\mathbf{x}_k)$.

Quasi-Newton Methods for Minimax Optimization

Come back to minimax optimization

$$\min_{\mathbf{x}\in\mathbb{R}^{n_x}} \max_{\mathbf{y}\in\mathbb{R}^{n_y}} f(\mathbf{x},\mathbf{y}).$$

The convexity on $\mathbf{x}$ and concavity on $\mathbf{x}$ means

$$\nabla^2 f_{\mathbf{xx}}(\mathbf{x},\mathbf{y}) \succ \mathbf{0} \quad \text{and} \quad \nabla^2 f_{\mathbf{yy}}(\mathbf{x},\mathbf{y}) \prec \mathbf{0}.$$

Then the Hessian

$$\nabla^2 f(\mathbf{x},\mathbf{y}) = \begin{bmatrix} \nabla^2 f_{\mathbf{xx}}(\mathbf{x},\mathbf{y}) & \nabla^2 f_{\mathbf{xy}}(\mathbf{x},\mathbf{y}) \\ \nabla^2 f_{\mathbf{yx}}(\mathbf{x},\mathbf{y}) & \nabla^2 f_{\mathbf{yy}}(\mathbf{x},\mathbf{y}) \end{bmatrix}$$

is indefinite.

Quasi-Newton Methods for Minimax Optimization

Denote $\mathbf{z} = (\mathbf{x}, \mathbf{y})$, $\mathbf{g}_k = \nabla f(\mathbf{x}_k, \mathbf{y}_k)$, $\hat{\mathbf{H}}_k = \nabla^2 f(\mathbf{x}_k, \mathbf{y}_k)$ and $\mathbf{H}_k = \hat{\mathbf{H}}_k^2$.

We rewrite the Newton update

$$\begin{aligned}\mathbf{z}_{k+1} =& \mathbf{z}_k - \hat{\mathbf{H}}_k^{-1}\mathbf{g}_k \\ =& \mathbf{z}_k - \big(\hat{\mathbf{H}}_k^{-1}\hat{\mathbf{H}}_k^{-1}\big)\big(\hat{\mathbf{H}}_k\mathbf{g}_k\big) \\ =& \mathbf{z}_k - \big(\mathbf{H}_k^{-1}\big)\hat{\mathbf{H}}_k\mathbf{g}_k.\end{aligned}$$

Some good news

- The square of Hessian is $(2\kappa^2 L_2/L)$-Lipschitz continuous
- $\mathbf{H}_k = \hat{\mathbf{H}}_k^2 \succeq \mu^2\mathbf{I}$
- $\hat{\mathbf{H}}_k\mathbf{g}_k$ and $\mathbf{H}_k\mathbf{g}_k$ can be computed efficiently

Approximating $\mathbf{H}_k$ leads to quasi-Newton for minimax!

Quasi-Newton Methods for Minimax Optimization

Algorithm Quasi-Newton for Minimax

```
Input: $\mathbf{z}_0 \in \mathbb{R}^n$, $\mathbf{G}_0 \succeq L^2\mathbf{I}$, $\tau_k \in [0, 1]$ and $M \geq 0$.
for $k = 0, 1 \ldots$
    $\mathbf{z}_{k+1} = \mathbf{z}_k - \mathbf{G}_k^{-1}\hat{\mathbf{H}}_k\mathbf{g}_k$
    $r_k = \|\mathbf{z}_{k+1} - \mathbf{z}_k\|$
    $\tilde{\mathbf{G}}_k = (1 + Mr_k)\mathbf{G}_k$
    $\mathbf{u}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
    $\mathbf{G}_{k+1} = \mathrm{Broyd}_{\tau_k}(\tilde{\mathbf{G}}_k, \mathbf{H}_{k+1}, \mathbf{u}_k)$.
end for
```

Quasi-Newton Methods for Minimax Optimization

The different choice of τ

- SR1 method: $\tau_k = 0$
- BFGS method: $\tau_k = \dfrac{\mathbf{u}_k^\top \mathbf{H}_{k+1}\mathbf{u}_k}{\mathbf{u}_k^\top \tilde{\mathbf{G}}_k \mathbf{u}_k} \in [0,1]$
- DFP method: $\tau_k = 1$

For BFGS/SR1 methods, we have

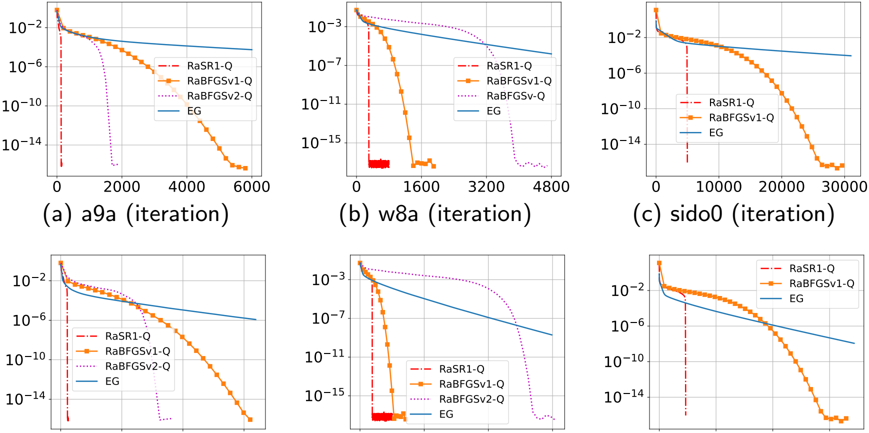
$$\lambda_{k+k_0} \leq \left(1 - \frac{1}{n}\right)^{\frac{k(k-1)}{2}} \left(\frac{1}{2}\right)^{k} \left(1 - \frac{1}{4\kappa^2}\right)^{k_0}$$

with probability $1-\delta$, where $\kappa = L/\mu$, $n = n_x + n_y$,

$$\lambda_k = \|\nabla f(\mathbf{x}_k, \mathbf{y}_k)\|_2 \quad \text{and} \quad k_0 = \mathcal{O}\left(\left(n+\kappa^2\right)\ln\left(\frac{n\kappa}{\delta}\right)\right).$$

We only require $\mathbf{z}_0$ is close to $\mathbf{z}^*$, rather than $\nabla^2 f(\mathbf{z}_0)$ is close to $\nabla^2 f(\mathbf{z}^*)$.

Experiments on AUC maximization



(a) a9a (iteration)

(b) w8a (iteration)

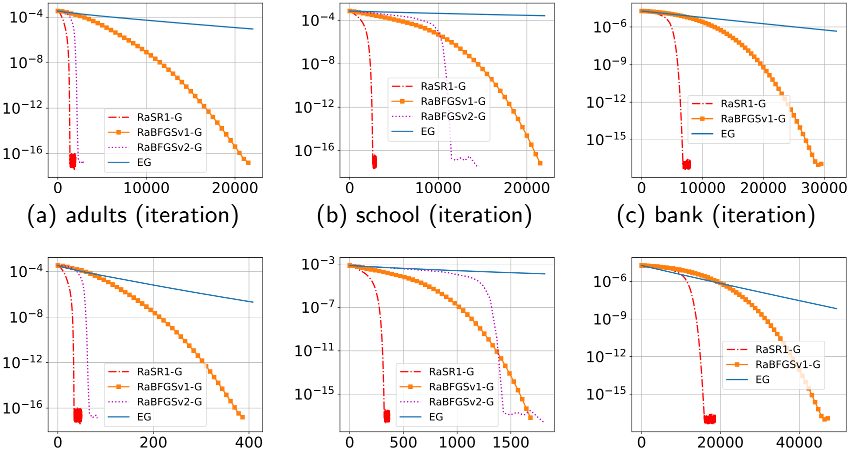
(c) sido0 (iteration)

(d) a9a (time)

(e) w8a (time)

(f) sido0 (time)

Experiments on Adversarial Debiasing



(a) adults (iteration)

(b) school (iteration)

(c) bank (iteration)

(d) adults (time)

(e) school (time)

(f) bank (time)

Some Open Problem

The theory of quasi-Newton is not perfect :

1. How to establish the global convergence?
2. What is the lower bound for superlinear convergence?
3. What is the local convergence of limited-memory quasi-Newton?
4. How to solve minimax problem only by using gradient?