Optimization Theory

Lecture 10

Fudan University

luoluo@fudan.edu.cn

Outline

1. Self-Concordant Functions

2. Classical Quasi-Newton Methods

3. Limited-Memory Quasi-Newton Methods

Outline

1. Self-Concordant Functions
2. Classical Quasi-Newton Methods
3. Limited-Memory Quasi-Newton Methods

Damped Newton Method

The damped Newton method is based on

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{1}{1 + M_f \lambda_f(\mathbf{x}_t)} \left(\nabla^2 f(\mathbf{x}_t)\right)^{-1} \nabla f(\mathbf{x}_t),$$

where $M_f > 0$ and

$$\lambda_f(\mathbf{x}_t) = \sqrt{\left\langle \nabla f(\mathbf{x}_t), \left(\nabla^2 f(\mathbf{x}_t)\right)^{-1} \nabla f(\mathbf{x}_t) \right\rangle}.$$

This method has global convergence guarantee under mild assumptions.

Self-Concordant Functions

Definition

We say $f : \mathbb{R}^d \to \mathbb{R}$ is M-strongly self-concordant, if it is twice differentiable and holds

$$\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y}) \preceq M \left\| \mathbf{x} - \mathbf{y} \right\|_{\nabla^2 f(\mathbf{z})} \nabla^2 f(\mathbf{w}),$$

for any $\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{w} \in \mathbb{R}^d$ and some $M > 0$.

1. The strong self-concordant property is affine invariant.
2. If $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex and has L_2-Lipschitz continuous Hessian, then it is M-strongly self-concordant with

$$M = \frac{L_2}{\mu^{3/2}}.$$

3. The M-strong self-concordance leads to $(M/2)$-self-concordance.

Self-Concordant Functions

Definition

A function $f : \mathbb{R}^d \to \mathbb{R}$ is called self-concordant if there exists a constant $M_f \geq 0$ such that the inequality

$$|D^3 f(\mathbf{x})[\mathbf{h}, \mathbf{h}, \mathbf{h}]| \leq 2M_f \, \|\mathbf{h}\|^3_{\nabla^2 f(\mathbf{x})}$$

holds for any $\mathbf{x}, \mathbf{h} \in \mathbb{R}^d$.

Lemma

A function $f : \mathbb{R}^d \to \mathbb{R}$ is self-concordant if and only if for any $\mathbf{x} \in \mathbb{R}^d$ and any triple of directions $\mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3 \in \mathbb{R}^d$, we have

$$|D^3 f(\mathbf{x})[\mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3]| \leq 2M_f \prod_{i=1}^{3} \|\mathbf{h}_i\|^3_{\nabla^2 f(\mathbf{x})}$$

Global Convergence

To the ease of presentation, we take $M = 2$ ($M_f = 1$). Then iteration

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{1}{1 + \lambda_f(\mathbf{x}_t)} \left(\nabla^2 f(\mathbf{x}_t)\right)^{-1} \nabla f(\mathbf{x}_t)$$

leads to global convergence of $\lambda_f(\mathbf{x}_t)$.

1. For $\lambda_f(\mathbf{x}_t) \geq 1/4$, we have

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leq -\frac{1}{38}.$$

2. For $\lambda_f(\mathbf{x}_t) \leq 1/4$, we have

$$\lambda_f(\mathbf{x}_{t+1}) \leq 2(\lambda_f(\mathbf{x}_t))^2.$$

Convergence Analysis

Let $\rho(z) = -\ln(1-z) - z$ and

$$\delta = \sqrt{(\mathbf{y}-\mathbf{x})^\top \nabla^2 f(\mathbf{x})(\mathbf{y}-\mathbf{x})} < 1,$$

then we have

$$\rho(-\delta) \leq f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x} \rangle \leq \rho(\delta),$$
$$(1-\delta)^2 \nabla^2 f(\mathbf{x}) \preceq \nabla^2 f(\mathbf{y}) \preceq \frac{1}{(1-\delta)^2} \nabla^2 f(\mathbf{x})$$

and

$$\left\| \nabla f(\mathbf{x})^{-1/2} \big(\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}) - \nabla^2 f(\mathbf{x})(\mathbf{y}-\mathbf{x}) \big) \right\|_2 \leq \frac{\delta^2}{1-\delta}.$$

Outline

1. Self-Concordant Functions

2. Classical Quasi-Newton Methods

3. Limited-Memory Quasi-Newton Methods

Secant Condition

For quadratic function

$$Q(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x},$$

we have $\nabla Q(\mathbf{x}_{t+1}) - \nabla Q(\mathbf{x}_t) = \nabla^2 Q(\mathbf{x}_{t+1})(\mathbf{x}_{t+1} - \mathbf{x}_t)$.

For general $f(\mathbf{x})$ with Lipschitz continuous Hessian, we have

$$\nabla f(\mathbf{x}_{t+1}) - \nabla f(\mathbf{x}_t) = \nabla^2 f(\mathbf{x}_{t+1})(\mathbf{x}_{t+1} - \mathbf{x}_t) + o(\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2),$$

which leads to

$$\nabla f(\mathbf{x}_{t+1}) - \nabla f(\mathbf{x}_t) \approx \nabla^2 f(\mathbf{x}_{t+1})(\mathbf{x}_{t+1} - \mathbf{x}_t).$$

Classical Quasi-Newton Methods

Motivated by

$$\nabla f(\mathbf{x}_{t+1}) - \nabla f(\mathbf{x}_t) \approx \nabla^2 f(\mathbf{x}_{t+1})(\mathbf{x}_{t+1} - \mathbf{x}_t),$$

classical Quasi-Newton methods target to find $\mathbf{G}_{t+1}$ such that

$$\nabla f(\mathbf{x}_{t+1}) - \nabla f(\mathbf{x}_t) = \mathbf{G}_{t+1}(\mathbf{x}_{t+1} - \mathbf{x}_t)$$

and update the variable as

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \mathbf{G}_t^{-1} \nabla f(\mathbf{x}_t).$$

We typically take $\mathbf{G}_0 = \delta_0 \mathbf{I}$ with some $\delta_0 > 0$.

For given $\mathbf{G}_t$ or $\mathbf{G}_t^{-1}$, we hope

1. $\{\mathbf{x}_t\}$ converges to $\mathbf{x}^*$ efficiently;
2. $\mathbf{G}_{t+1}$ is close to $\mathbf{G}_t$;
3. $\mathbf{G}_{t+1}$ or $\mathbf{G}_{t+1}^{-1}$ can be constructed/stored efficiently.

Woodbury Matrix Identity

The Woodbury matrix identity is

$$(\mathbf{A} + \mathbf{UCV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{C}^{-1} + \mathbf{VA}^{-1}\mathbf{U})^{-1}\mathbf{VA}^{-1},$$

where $\mathbf{A} \in \mathbb{R}^{d\times d}$, $\mathbf{C} \in \mathbb{R}^{k\times k}$, $\mathbf{U} \in \mathbb{R}^{d\times k}$ and $\mathbf{V} \in \mathbb{R}^{k\times d}$.

For $\mathbf{A} = \mathbf{G}_t$, $\mathbf{U} = \mathbf{Z}_t$, $\mathbf{V} = \mathbf{Z}_t^\top$ and $\mathbf{C} = \mathbf{I}$, we let

$$\mathbf{G}_{t+1} = \mathbf{G}_t + \mathbf{Z}_t\mathbf{Z}_t^\top,$$

then

$$\mathbf{G}_{t+1}^{-1} = \mathbf{G}_t^{-1} - \mathbf{G}_t^{-1}\mathbf{Z}_t(\mathbf{I} + \mathbf{Z}_t^\top\mathbf{G}_t^{-1}\mathbf{Z}_t)^{-1}\mathbf{Z}_t^\top\mathbf{G}_t^{-1}$$

can be computed within $\mathcal{O}(kd^2)$ flops for given $\mathbf{G}_t^{-1}$.

Classical SR1 Method

We consider secant condition and the symmetric rank one (SR1) update

$$\begin{cases} \mathbf{y}_t = \mathbf{G}_{t+1}\mathbf{s}_t, \\ \mathbf{G}_{t+1} = \mathbf{G}_t + \mathbf{z}_t\mathbf{z}_t^\top. \end{cases}$$

where $\mathbf{s}_t = \mathbf{x}_{t+1} - \mathbf{x}_t$ and $\mathbf{y}_t = \nabla f(\mathbf{x}_{t+1}) - \nabla f(\mathbf{x}_t)$.

It implies

$$\mathbf{G}_{t+1} = \mathbf{G}_t + \frac{(\mathbf{y}_t - \mathbf{G}_t\mathbf{s}_t)(\mathbf{y}_t - \mathbf{G}_t\mathbf{s}_t)^\top}{(\mathbf{y}_t - \mathbf{G}_t\mathbf{s}_t)^\top\mathbf{s}_t}.$$

and the corresponding update to inverse of Hessian estimator is

$$\mathbf{G}_{t+1}^{-1} = \mathbf{G}_t^{-1} + \frac{(\mathbf{s}_t - \mathbf{G}_t^{-1}\mathbf{y}_t)(\mathbf{s}_t - \mathbf{G}_t^{-1}\mathbf{y}_t)^\top}{(\mathbf{s}_t - \mathbf{G}_t^{-1}\mathbf{y}_t)^\top\mathbf{y}_t}.$$

Classical DFP Method

Let $\mathbf{G}_{t+1}$ be the solution of following matrix optimization problem

$$\min_{\mathbf{G}\in\mathbb{R}^{d\times d}} \|\mathbf{G}-\mathbf{G}_t\|_{\bar{\mathbf{G}}_t^{-1}}$$
$$\mathrm{s.t} \quad \mathbf{G}=\mathbf{G}^\top, \ \ \mathbf{G}\mathbf{s}_t=\mathbf{y}_t,$$

where the weighted norm $\|\cdot\|_{\bar{\mathbf{G}}_t}$ is defined as

$$\|\mathbf{A}\|_{\bar{\mathbf{G}}_t} = \left\|\bar{\mathbf{G}}_t^{-1/2}\mathbf{A}\bar{\mathbf{G}}_t^{-1/2}\right\|_F \quad \text{with} \quad \bar{\mathbf{G}}_t = \int_0^1 \nabla^2 f(\mathbf{x}_t+\tau(\mathbf{x}_{t+1}-\mathbf{x}_t))\,\mathrm{d}\tau.$$

It implies DFP update

$$\mathbf{G}_{t+1} = \left(\mathbf{I}-\frac{\mathbf{y}_t\mathbf{s}_t^\top}{\mathbf{y}_t^\top\mathbf{s}_t}\right)\mathbf{G}_t\left(\mathbf{I}-\frac{\mathbf{s}_t\mathbf{y}_t^\top}{\mathbf{y}_t^\top\mathbf{s}_t}\right)+\frac{\mathbf{y}_t\mathbf{y}_t^\top}{\mathbf{y}_t^\top\mathbf{s}_t}.$$

The corresponding update to inverse of Hessian estimator is

$$\mathbf{G}_{t+1}^{-1} = \mathbf{G}_t^{-1} - \frac{\mathbf{G}_t^{-1}\mathbf{y}_t\mathbf{y}_t^\top\mathbf{G}_t^{-1}}{\mathbf{y}_t^\top\mathbf{G}_t^{-1}\mathbf{y}_t}+\frac{\mathbf{s}_t\mathbf{s}_t^\top}{\mathbf{y}_t^\top\mathbf{s}_t}.$$

Classical BFGS Method

This algorithm is named after Charles G. Broyden, Roger Fletcher, Donald Goldfarb and David F. Shanno.

Broyden, Fletcher, Goldfarb, Shanno

Classical BFGS Method

Let $\mathbf{G}_{t+1}^{-1}$ be the solution of the following matrix optimization problem

$$\begin{aligned} &\min_{\mathbf{H}\in\mathbb{R}^{d\times d}} \|\mathbf{H}-\mathbf{H}_t\|_{\bar{\mathbf{G}}_t} \\ &\text{s.t} \quad \mathbf{H}=\mathbf{H}^\top, \ \ \mathbf{H}\mathbf{y}_t=\mathbf{s}_t, \end{aligned}$$

where $\mathbf{H}_t=\mathbf{G}_t^{-1}$ and the weighted norm $\|\cdot\|_{\bar{\mathbf{G}}_t}$ is defined as

$$\|\mathbf{A}\|_{\bar{\mathbf{G}}_t} = \left\|\bar{\mathbf{G}}_t^{1/2}\mathbf{A}\bar{\mathbf{G}}_t^{1/2}\right\|_F \quad \text{with} \quad \bar{\mathbf{G}}_t = \int_0^1 \nabla^2 f(\mathbf{x_t}+\tau(\mathbf{x}_{t+1}-\mathbf{x}_t))\,\mathrm{d}\tau.$$

It implies BFGS update

$$\mathbf{G}_{t+1}^{-1} = \left(\mathbf{I}-\frac{\mathbf{s}_t\mathbf{y}_t^\top}{\mathbf{y}_t^\top\mathbf{s}_t}\right)\mathbf{G}_t^{-1}\left(\mathbf{I}-\frac{\mathbf{y}_t\mathbf{s}_t^\top}{\mathbf{y}_t^\top\mathbf{s}_t}\right)+\frac{\mathbf{s}_t\mathbf{s}_t^\top}{\mathbf{y}_t^\top\mathbf{s}_t}.$$

The corresponding update to Hessian estimator is

$$\mathbf{G}_{t+1} = \mathbf{G}_t - \frac{\mathbf{G}_t\mathbf{s}_t\mathbf{s}_t^\top\mathbf{G}_t}{\mathbf{s}_t^\top\mathbf{G}_t\mathbf{s}_t} + \frac{\mathbf{y}_t\mathbf{y}_t^\top}{\mathbf{y}_t^\top\mathbf{s}_t}.$$

Superlinear Convergence

The following theorem implies SR1/DFP/BFGS converge superlinearly.

Theorem (Dennis–Moré Condition)

If sequence $\{\mathbf{x}_t\}$ converges to $\mathbf{x}^*$ such that $\nabla f(\mathbf{x}^*) = \mathbf{0}$ and $\nabla^2 f(\mathbf{x}^*) \succ \mathbf{0}$ and the search direction satisfies

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

Then $\{\mathbf{x}_t\}$ converges to $\mathbf{x}^*$ superlinearly.

For quasi-Newton iteration $\mathbf{x}_{t+1} = \mathbf{x}_t - \mathbf{G}_t^{-1}\nabla f(\mathbf{x}_t)$, the condition in above theorem can be written as

$$\lim_{t\to\infty} \frac{\left\|(\mathbf{G}_t - \nabla^2 f(\mathbf{x}_t))(\mathbf{x}_{t+1} - \mathbf{x}_t)\right\|_2}{\left\|\mathbf{x}_{t+1} - \mathbf{x}_t\right\|_2} = 0,$$

which only requires that $\mathbf{G}_t$ converges to Hessian along with the search direction.

Outline

1. Self-Concordant Functions

2. Classical Quasi-Newton Methods

3. Limited-Memory Quasi-Newton Methods

Quasi-Newton Methods

Classical quasi-Newton methods are too expensive for large d.

1. Each iteration requires $\mathcal{O}(d^2)$ complexity.
2. The space complexity is $\mathcal{O}(d^2)$.

Limited-Memory BFGS (L-BFGS)

The BFGS update can be written as

$$\mathbf{H}_{t+1} = \mathbf{V}_t^\top \mathbf{H}_t \mathbf{V}_t + \rho_t \mathbf{s}_t \mathbf{s}_t^\top,$$

where $\rho_t = (\mathbf{y}_t^\top \mathbf{s}_t)^{-1}$ and $\mathbf{V}_t = \mathbf{I} - \rho_t \mathbf{y}_t \mathbf{s}_t^\top$.

Limited-memory BFGS method keeps the m most recent vector pairs

$$\{\mathbf{s}_i, \mathbf{y}_i\}_{i=k-m}^{k-1}$$

and applying BFGS update m times on some initial estimator $\mathbf{H}_{k,0}$.

Limited-Memory BFGS (L-BFGS)

The update of L-BFGS can be written as

$$\begin{aligned}\mathbf{H}_k =&(\mathbf{V}_{k-1}^\top \dots \mathbf{V}_{k-m}^\top)\mathbf{H}_{k,0}(\mathbf{V}_{k-m} \dots \mathbf{V}_{k-1}) \\ &+ \rho_{k-m}(\mathbf{V}_{k-1}^\top \dots \mathbf{V}_{k-m+1}^\top)\mathbf{s}_{k-m}\mathbf{s}_{k-m}^\top(\mathbf{V}_{k-m+1} \dots \mathbf{V}_{k-1}) \\ &+ \rho_{k-m+1}(\mathbf{V}_{k-1}^\top \dots \mathbf{V}_{k-m+2}^\top)\mathbf{s}_{k-m+1}\mathbf{s}_{k-m+1}^\top(\mathbf{V}_{k-m+2} \dots \mathbf{V}_{k-1}) \\ &+ \dots \\ &+ \rho_{k-1}\mathbf{s}_{k-1}\mathbf{s}_{k-1}^\top.\end{aligned}$$

The iteration of L-BFGS is efficient for small m.

1. Computing $\mathbf{H}_k \nabla f(\mathbf{x}_k)$ requires $\mathcal{O}(md)$ flops for given $\nabla f(\mathbf{x}_k)$.
2. The storage of $\{\mathbf{s}_i, \mathbf{y}_i\}_{i=k-m}^{k-1}$ requires $\mathcal{O}(md)$ space complexity.
3. The idea also works for SR1 and DFP.

What is the convergence rate of L-BFGS?