Optimization Theory

Lecture 13

Fudan University

luoluo@fudan.edu.cn

Outline

Outline

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

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

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

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

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

Classical BFGS Method

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

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

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

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

Asymptotic 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.

Broyden Family Update

The Broyden family update is

$$\mathrm{Broyd}_\tau(\mathbf{G}, \mathbf{A}, \mathbf{u}) \triangleq \tau \left[\mathbf{G} - \frac{\mathbf{A}\mathbf{u}\mathbf{u}^\top\mathbf{G} + \mathbf{G}\mathbf{u}\mathbf{u}^\top\mathbf{A}}{\mathbf{u}^\top\mathbf{A}\mathbf{u}} + \left(\frac{\mathbf{u}^\top\mathbf{G}\mathbf{u}}{\mathbf{u}^\top\mathbf{A}\mathbf{u}} + 1\right)\frac{\mathbf{A}\mathbf{u}\mathbf{u}^\top\mathbf{A}}{\mathbf{u}^\top\mathbf{A}\mathbf{u}}\right] + (1-\tau)\left[\mathbf{G} - \frac{(\mathbf{G}-\mathbf{A})\mathbf{u}\mathbf{u}^\top(\mathbf{G}-\mathbf{A})}{\mathbf{u}^\top(\mathbf{G}-\mathbf{A})\mathbf{u}}\right],$$

where $\mathbf{G} \in \mathbb{R}^{d\times d}$,$\mathbf{A} \in \mathbb{R}^{d\times d}$, $\mathbf{u} \in \mathbb{R}^d$ and $\tau \in [0, 1]$.

Let $\mathbf{G} = \mathbf{G}_t$, $\mathbf{A} = \int_0^1 \nabla^2 f(\mathbf{x}_t + t(\mathbf{x}_{t+1} - \mathbf{x}_t))\,\mathrm{d}t$ and $\mathbf{u} = \mathbf{x}_{t+1} - \mathbf{x}_t$.

- For $\tau = 0$, it is classical SR1 method.
- For $\tau = \dfrac{\mathbf{u}^\top\mathbf{A}\mathbf{u}}{\mathbf{u}^\top\mathbf{G}\mathbf{u}}$, it is classical BFGS method.
- For $\tau = 1$, it is classical DFP method.

Explicit Local Convergence Rate

Suppose the objective is μ-strongly-convex and L-smooth and let

$$\kappa = L/\mu \quad \text{and} \quad \lambda_t = \sqrt{\nabla f(\mathbf{x}_t)^\top (\nabla^2 f(\mathbf{x}_t))^{-1} \nabla f(\mathbf{x}_t)}.$$

1. For classical DFP method, we have
$$\lambda_t \leq \mathcal{O}\left(\left(\frac{\kappa^2 d}{t}\right)^{t/2}\right).$$
2. For classical BFGS method, we have
$$\lambda_t \leq \mathcal{O}\left(\left(\frac{\kappa d}{t}\right)^{t/2}\right).$$
3. For classical SR1 method, we have
$$\lambda_t \leq \mathcal{O}\left(\left(\frac{d \ln \kappa}{t}\right)^{t/2}\right).$$

Outline

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 $\mathcal{O}(d^2)$.

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 $\mathbf{H}_{k,0} = \delta_{k,0}\mathbf{I}$.

Limited-Memory BFGS (L-BFGS)

The update of L-BFGS can be written as

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

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

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