Optimization Theory

Lecture 12

Fudan University

luoluo@fudan.edu.cn

Outline

Outline

1. Greedy and Randomized Quasi-Newton Methods

2. Block Quasi-Newton Methods

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.

Greedy and Randomized Directions

The update $\mathbf{G}_{t+1} = \mathrm{Broyd}_\tau(\mathbf{G}, \mathbf{A}, \mathbf{u})$ with $\mathbf{A} = \nabla^2 f(\mathbf{x}_{t+1})$ satisfies

$$\mathbf{G}_{t+1}\mathbf{u} = \nabla^2 f(\mathbf{x}_{t+1})\mathbf{u}$$

for any $\mathbf{u} \in \mathbb{R}^d$.

We can construct $\mathbf{G}_{t+1}$ by the following choice of $\mathbf{u}$.

1. Greedy strategy: $\mathbf{u} = \arg\max_{\mathbf{v}\in\{\mathbf{e}_1,\dots,\mathbf{e}_d\}} \mathbf{v}^\top(\mathbf{G}_t - \nabla^2 f(\mathbf{x}_{t+1}))\mathbf{v}$;
2. Randomized strategy: $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Greedy and Randomized Quasi-Newton Methods

Algorithm 1 Greedy and Randomized Quasi-Newton Methods

1: **Input:** $\mathbf{G}_0 \in \mathbb{R}^{d \times d}$, $M > 0$

2: **for** $t = 0, 1 \ldots$

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

4: $\quad r_t = \|\mathbf{x}_{t+1} - \mathbf{x}_t\|_{\nabla^2 f(\mathbf{x}_t)}$

5: $\quad \tilde{\mathbf{G}}_t = (1 + M r_t) \mathbf{G}_t$

6: $\quad$ Construct $\mathbf{u}_t \in \mathbb{R}^d$ by

(a) randomized strategy: $[\mathbf{u}_t]_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, 1)$

(b) greedy strategy: $\mathbf{u}_t = \arg\max_{\mathbf{v} \in \{\mathbf{e}_1, \ldots, \mathbf{e}_d\}} \mathbf{v}^\top (\mathbf{G}_t - \nabla^2 f(\mathbf{x}_{t+1})) \mathbf{v}$

7: $\quad \mathbf{G}_{t+1} = \mathrm{Broyd}_\tau(\tilde{\mathbf{G}}_t, \nabla^2 f(\mathbf{x}_{t+1}), \mathbf{u}_t)$

8: **end for**

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 greedy/randomized Broyden family method, we have

$$\mathbb{E}[\lambda_t] \leq \mathcal{O}\left(\left(1 - \frac{1}{\kappa d}\right)^{t(t-1)}\right).$$

2. For greedy/randomized SR1 method, we have

$$\mathbb{E}[\lambda_t] \leq \mathcal{O}\left(\left(1 - \frac{1}{d}\right)^{t(t-1)}\right).$$

3. The rate $\mathbb{E}[\lambda_{t+1}/\lambda_t]$ converges to 0 linearly.

Outline

Multiple Directions

Recall that we have used the fact

$$\mathbf{G}_{t+1}\mathbf{u} = \nabla^2 f(\mathbf{x}_{t+1})\mathbf{u}$$

of Broyden family update to construct $\mathbf{G}_{t+1} \in \mathbb{R}^{d\times d}$.

Can we use multiple directions to construct $\mathbf{G}_{t+1}$? Such as

$$\mathbf{G}_{t+1}\mathbf{U} = \nabla^2 f(\mathbf{x}_{t+1})\mathbf{U}$$

for some $\mathbf{U} \in \mathbb{R}^{d\times k}$, where $k \ll d$.

Symmetric Rank-k Update

Recall that SR1 update can be written as

$$\mathrm{SR1}(\mathbf{G}, \mathbf{A}, \mathbf{u}) = \mathbf{G} - \frac{(\mathbf{G}-\mathbf{A})\mathbf{u}\mathbf{u}^\top(\mathbf{G}-\mathbf{A})}{\mathbf{u}^\top(\mathbf{G}-\mathbf{A})\mathbf{u}}.$$

for given $\mathbf{G} \in \mathbb{R}^{d\times d}$, $\mathbf{A} \in \mathbb{R}^{d\times d}$ and some $\mathbf{u} \in \mathbb{R}^{d}$.

We generalized SR1 to SR-k as

$$\mathrm{SR}\text{-}k(\mathbf{G}, \mathbf{A}, \mathbf{U}) = \mathbf{G} - (\mathbf{G}-\mathbf{A})\mathbf{U}(\mathbf{U}^\top(\mathbf{G}-\mathbf{A})\mathbf{U})^{-1}\mathbf{U}^\top(\mathbf{G}-\mathbf{A})$$

for given $\mathbf{G} \in \mathbb{R}^{d\times d}$, $\mathbf{A} \in \mathbb{R}^{d\times d}$ and some $\mathbf{U} \in \mathbb{R}^{d\times k}$.

Symmetric Rank-k Method

Algorithm 2 Symmetric Rank-k Method

Input: $\mathbf{G}_0 \in \mathbb{R}^{d \times d}$, $M \geq 0$ and $k \in [d]$.
for $t = 0, 1 \ldots$
    $\mathbf{x}_{t+1} = \mathbf{x}_t - \mathbf{G}_t^{-1} \nabla f(\mathbf{x}_t)$
    $r_t = \|\mathbf{x}_{t+1} - \mathbf{x}_t\|_{\nabla^2 f(\mathbf{x}_t)}$
    $\tilde{\mathbf{G}}_t = (1 + M r_t) \mathbf{G}_t$
    construct $\mathbf{U}_t \in \mathbb{R}^{d \times k}$ by $[\mathbf{U}_t]_{ij} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, 1)$
    $\mathbf{G}_{t+1} = \mathrm{SR}\text{-}k(\tilde{\mathbf{G}}_t, \nabla^2 f(\mathbf{x}_{t+1}), \mathbf{U}_t)$
end for

1. SR-k method has the local convergence rate $\mathbb{E}[\lambda_t] \leq \mathcal{O}\big((1 - k/d)^{t(t-1)} \big)$.
2. For quadratic problems, we set $M = 0$ and it has global linear convergence.

Symmetric Rank-k Update

Lemma

For any positive-definite matrices $\mathbf{A} \in \mathbb{R}^{d\times d}$ *and* $\mathbf{G} \in \mathbb{R}^{d\times d}$ *with*

$$\mathbf{A} \preceq \mathbf{G} \preceq \eta\mathbf{A}$$

for some $\eta \geq 1$, *we let* $\mathbf{G}_+ = \mathrm{SR}\text{-}k(\mathbf{G}, \mathbf{A}, \mathbf{U})$ *for some full rank matrix* $\mathbf{U} \in \mathbb{R}^{d\times k}$. *Then it holds that*

$$\mathbf{A} \preceq \mathbf{G}_+ \preceq \eta\mathbf{A}.$$

If we can construct $\{\eta_t\}$ such that

$$\nabla^2 f(\mathbf{x}_t) \preceq \mathbf{G}_t \preceq \eta_t \nabla^2 f(\mathbf{x}_t) \quad \text{and} \quad \lim_{t\to+\infty} \eta_t = 1.$$

Then the update $\mathbf{G}_{t+1} = \mathrm{SR}\text{-}k(\mathbf{G}_t, \nabla f(\mathbf{x}_{t+1}), \mathbf{U}_t)$ leads to

$$\lim_{t\to+\infty} (\mathbf{G}_t - \nabla^2 f(\mathbf{x}_t)) = \mathbf{0}.$$

Convergence Analysis

We introduce the quantity

$$\tau_{\mathbf{A}}(\mathbf{G}) \triangleq \mathrm{tr}(\mathbf{G} - \mathbf{A})$$

to characterize the difference between $\mathbf{A}$ and $\mathbf{G}$.

Theorem

Let $\mathbf{G}_+ = \mathrm{SR}\text{-}k(\mathbf{G}, \mathbf{A}, \mathbf{U})$ *with* $\mathbf{G} \succeq \mathbf{A} \in \mathbb{R}^{d\times d}$ *and select* $\mathbf{U} \in \mathbb{R}^{d\times k}$ *by drawing each entry of* $\mathbf{U}$ *according to* $\mathcal{N}(0,1)$ *independently. Then*

$$\mathbb{E}\left[\tau_{\mathbf{A}}(\mathbf{G}_+)\right] \leq \left(1 - \frac{k}{d}\right)\tau_{\mathbf{A}}(\mathbf{G}).$$

Lemma

Assume $\mathbf{P} \in \mathbb{R}^{d\times k}$ *is column orthonormal* $(k \le d)$ *and* $\mathbf{p} \sim \mathcal{N}(\mathbf{0}, \mathbf{P}\mathbf{P}^\top)$ *is a* d*-dimensional multivariate normal distributed vector. Then we have*

$$\mathbb{E}\left[\frac{\mathbf{p}\mathbf{p}^\top}{\mathbf{p}^\top\mathbf{p}}\right] = \frac{1}{k}\mathbf{P}\mathbf{P}^\top.$$

Lemma

Let $\mathbf{U} \in \mathbb{R}^{d\times k}$ *be a random matrix and each of its entry is independent and identically distributed according to* $\mathcal{N}(0,1)$*, then it holds that*

$$\mathbb{E}\left[\mathbf{U}(\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{U}^\top\right] = \frac{k}{d}\mathbf{I}_d.$$

Lemma

For positive semi-definite matrix $\mathbf{B} \in \mathbb{R}^{d\times d}$ *and full rank matrix* $\mathbf{U} \in \mathbb{R}^{d\times k}$ *with* $k \le d$*, it holds that*

$$\mathrm{tr}\left(\mathbf{B}\mathbf{U}\left(\mathbf{U}^\top\mathbf{B}\mathbf{U}\right)^{-1}\mathbf{U}^\top\mathbf{B}\right) \ge \mathrm{tr}\left(\mathbf{U}\left(\mathbf{U}^\top\mathbf{U}\right)^{-1}\mathbf{U}^\top\mathbf{B}\right).$$

Lemma

Suppose the twice differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is M-strongly self-concordant and the positive definite matrix $\mathbf{G}_t \in \mathbb{R}^{d\times d}$ satisfies

$$\nabla^2 f(\mathbf{x}_t) \preceq \mathbf{G}_t \preceq \eta_t \nabla^2 f(\mathbf{x}_t)$$

for some $\eta_t \geq 1$ and $M\lambda_t \leq 2$. Then the update formula

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

holds that

$$\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_{\nabla^2 f(\mathbf{x}_t)} \leq \lambda_t \quad \textit{and} \quad \lambda_{t+1} \leq \left(1 - \frac{1}{\eta_t}\right)\lambda_t + \frac{M}{2}\lambda_t^2 + \frac{M^2}{4\eta_t}\lambda_t^3.$$

Convergence Analysis

Lemma

If twice differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is M-strongly self-concordant and μ-strongly convex and positive definite matrix $\mathbf{G} \in \mathbb{R}^{d\times d}$ and $\mathbf{x} \in \mathbb{R}^d$ satisfy

$$\nabla^2 f(\mathbf{x}) \preceq \mathbf{G} \preceq \eta \nabla^2 f(\mathbf{x})$$

for some $\eta \geq 1$, then we have

$$\nabla^2 f(\mathbf{x}_+) \preceq \tilde{\mathbf{G}} \preceq \eta(1+Mr)^2 \nabla^2 f(\mathbf{x}_+)$$

for any $\mathbf{x}_+ \in \mathbb{R}^d$, where $\tilde{\mathbf{G}} = (1+Mr)\mathbf{G}$, $r = \|\mathbf{x} - \mathbf{x}_+\|_{\nabla^2 f(\mathbf{x})}$.

This implies

$$\nabla^2 f(\mathbf{x}_t) \preceq \mathbf{G}_t \preceq (1+\delta_t)\nabla^2 f(\mathbf{x}_t),$$

where

$$\delta_t = \frac{d\kappa \operatorname{tr}(\mathbf{G}_t - \nabla^2 f(\mathbf{x}_t))}{\operatorname{tr}(\nabla^2 f(\mathbf{x}_t))}.$$

Convergence Analysis

Theorem

Suppose $f : \mathbb{R}^d \to \mathbb{R}$ *is L-smooth,* μ*-strongly-convex and M-strongly self-concordant, we run* SR*-k with initial* $\mathbf{x}_0$ *and* $\mathbf{G}_0$ *such that*

$$\lambda_0 \leq \frac{\ln(3/2)}{4\eta_0 M} \qquad \text{and} \qquad \nabla^2 f(\mathbf{x}_0) \preceq \mathbf{G}_0 \preceq \eta_0 \nabla^2 f(\mathbf{x}_0)$$

for some $\eta_0 \geq 1$*. Then it holds that*

$$\nabla^2 f(\mathbf{x}_t) \preceq \mathbf{G}_t \preceq \frac{3\eta_0}{2} \nabla^2 f(\mathbf{x}_t) \qquad \text{and} \qquad \lambda_t \leq \left(1 - \frac{1}{2\eta_0}\right)^t \lambda_0.$$

Convergence Analysis

Theorem

Suppose function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth, μ-strongly-convex and M-strongly self-concordant, if we run SR-k with $k < d$ and set the initial $\mathbf{x}_0$ and $\mathbf{G}_0$ such that

$$\lambda_0 \leq \frac{\ln 2}{2} \cdot \frac{(d-k)}{M\eta_0 d^2 \kappa} \qquad \text{and} \qquad \nabla^2 f(\mathbf{x}_0) \preceq \mathbf{G}_0 \preceq \eta_0 \nabla^2 f(\mathbf{x}_0)$$

for some $\eta_0 \geq 1$. Then we have

$$\mathbb{E}\left[\frac{\lambda_{t+1}}{\lambda_t}\right] \leq 2d\kappa\eta_0 \left(1 - \frac{k}{d}\right)^t,$$

which naturally indicates the following two stage convergence

$$\lambda_{t_0+t} \leq \left(1 - \frac{k}{d+k}\right)^{t(t-1)/2} \cdot \left(\frac{1}{2}\right)^t \cdot \left(1 - \frac{1}{2\eta_0}\right)^{t_0} \lambda_0,$$

with probability at least $1-\delta$ for some $\delta \in (0,1)$, where $t_0 = \mathcal{O}(d \ln(\eta_0 \kappa d/\delta)/k)$.

Convergence Analysis

Corollary

Under the assumptions of above theorem, we run SR-k *with* $k = d$ *and set the initial* $\mathbf{x}_0$ *and* $\mathbf{G}_0$ *such that*

$$\lambda(\mathbf{x}_0) \leq \frac{\ln(3/2)}{4M\eta_0} \quad \textit{and} \quad \nabla^2 f(\mathbf{x}_0) \preceq \mathbf{G}_0 \preceq \eta_0 \nabla^2 f(\mathbf{x}_0).$$

Then we have $\mathbb{E}[\lambda_{t+1}] \leq M\lambda_t^2$ *for any* $t \geq 1$.

SR-k builds a bridge between the theories of standard quasi-Newton methods and Newton's method.