Notes on Optimization

Peize Sun

1 Preliminary

Theorem 1-1 Assume $\varphi(\lambda) = f(\boldsymbol{x} + \lambda \boldsymbol{p})$, then

$$\varphi'(\lambda) = \nabla^T f(\boldsymbol{x} + \lambda \boldsymbol{p})\ \boldsymbol{p}$$
$$\varphi''(\lambda) = \boldsymbol{p}^T\ \nabla^2 f(\boldsymbol{x} + \lambda \boldsymbol{p})\ \boldsymbol{p}$$

Theorem 1-2 Assume $f(\boldsymbol{x})$ is continuous second-order differentiable. Given one point $\boldsymbol{x}$ and one direction $\boldsymbol{p}$, then

$$f(\boldsymbol{x} + \lambda \boldsymbol{p}) = f(\boldsymbol{x}) + \nabla^T f(\boldsymbol{x})\ \boldsymbol{p}\ \lambda + \frac{1}{2}\boldsymbol{p}^T\ \nabla^2 f(\boldsymbol{x})\ \boldsymbol{p}\ \lambda^2 + \circ(\lambda^2)$$
$$f(\boldsymbol{x} + \lambda \boldsymbol{p}) = f(\boldsymbol{x}) + \nabla^T f(\boldsymbol{x})\ \boldsymbol{p}\ \lambda + \frac{1}{2}\boldsymbol{p}^T\ \nabla^2 f(\boldsymbol{\xi})\ \boldsymbol{p}\ \lambda^2$$

where $\boldsymbol{\xi}$ is one point between $\boldsymbol{x}$ and $\boldsymbol{x} + \lambda \boldsymbol{p}$.

Theorem 1-3 Assume $f(\boldsymbol{x})$ is continuous second-order differentiable. Given one point $\boldsymbol{x_0}$, then $\forall\ \boldsymbol{x}$:

$$f(\boldsymbol{x}) = f(\boldsymbol{x_0}) + \nabla^T f(\boldsymbol{x_0})\ (\boldsymbol{x} - \boldsymbol{x_0}) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x_0})^T\ \nabla^2 f(\boldsymbol{x_0})\ (\boldsymbol{x} - \boldsymbol{x_0}) + \circ(||\boldsymbol{x} - \boldsymbol{x_0}||^2)$$
$$f(\boldsymbol{x}) = f(\boldsymbol{x_0}) + \nabla^T f(\boldsymbol{x_0})\ (\boldsymbol{x} - \boldsymbol{x_0}) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x_0})^T\ \nabla^2 f(\boldsymbol{\xi})\ (\boldsymbol{x} - \boldsymbol{x_0})$$

where $\boldsymbol{\xi} = \boldsymbol{x_0} + \theta(\boldsymbol{x} - \boldsymbol{x_0})\ \ (0 < \theta < 1)$ is one point between $\boldsymbol{x_0}$ and $\boldsymbol{x}$.

2 Unconstrained optimization

Formulation Denote $f(\boldsymbol{x})$ is function of n-dimension variable $\boldsymbol{x}$, then unconstrained optimization is formulated as

$$\min_{\boldsymbol{x}} f(\boldsymbol{x})$$

Theorem 2-1 Assume $f(\boldsymbol{x})$ is differentiable, $\boldsymbol{x}^* \in R^n$ is one local minimum point, then

$$\nabla f(\boldsymbol{x}^*) = 0$$

Theorem 2-2 Assume $f(\boldsymbol{x})$ is differentiable, if one point $\boldsymbol{x}^*$, $\nabla f(\boldsymbol{x}^*) = 0$, and $\nabla^2 f(\boldsymbol{x}^*)$ is positive-definite, then $\boldsymbol{x}^*$ is one local mininum point.(Sufficient Condition)

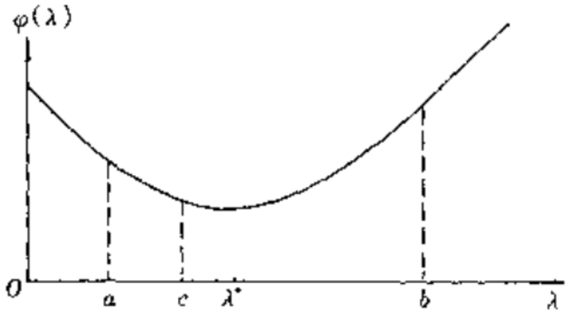
Theorem 2-3 Assume $f(\boldsymbol{x})$ is differentiable, if one point $\boldsymbol{x}^*$ is local mininum point, then $\nabla f(\boldsymbol{x}^*) = 0$, and $\nabla^2 f(\boldsymbol{x}^*)$ is positive semi-definite. (Necessary Condition)

3 Line Search

Formulation Assume $\varphi(\lambda) = f(\boldsymbol{x} + \lambda \boldsymbol{p})$, given one point $\boldsymbol{x}$ and one direction $\boldsymbol{p}$, then line search is formulated as

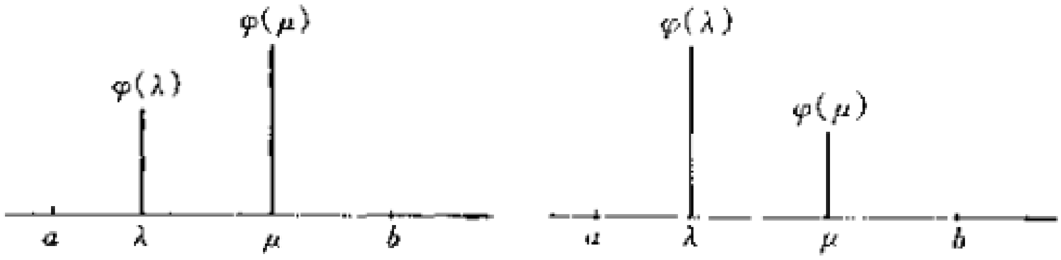
$$\lambda^* = \arg\min_{\lambda>0} \varphi(\lambda) = \arg\min_{\lambda>0} f(\boldsymbol{x} + \lambda \boldsymbol{p})$$

Golden Section Search
1. Assume $\varphi(\lambda)$ has "single valley" in range $[a, b]$: decreases in $[a, \lambda^*]$, increases in $[\lambda^*, b]$. Accordingly, find search range $[a, b]$ as follows[1]:



(a) $a = 0, c = a$+step length. If $\varphi(a) \leq \varphi(c)$, narrow step length, until $\varphi(a) > \varphi(c)$.
(b) $b = c$+step length. If $\varphi(c) \geq \varphi(b)$, push right $a, c \to c, b$(when $\varphi(a) > \varphi(c) > \varphi(b)$, it shows $\varphi(\lambda)$ is keeping decreasing, so search starting point a could be push right $a, c \to c, b$) and enlarge step length, until $\varphi(c) < \varphi(b)$.

2. Denote left point $\lambda = b - \tau(b - a)$, right point $\mu = a + \tau(b - a)$, $\tau > \frac{1}{2}$. If $\varphi(\lambda) < \varphi(\mu)$, then $\lambda^* \in [a, \mu]$, $a, b \to a, \mu$; else $a, b \to \lambda, b$. Narrow search range iteratively.



For τ, considering example of $[a, b] \to [a, \mu]$, new right point $\mu_{new} = a + \tau(\mu - a)$, making $\mu_{new} = \lambda$ could ease computation cost, so

$$\begin{aligned}
\mu_{new} &= \lambda \\
a + \tau(\mu - a) &= b - \tau(b - a) \\
a + \tau(a + \tau(b - a) - a) &= b - \tau(b - a) \\
(\tau^2 + \tau - 1)(b - a) &= 0 \\
(\tau^2 + \tau - 1) &= 0 \\
\tau &= \frac{1}{2}(-1 + \sqrt{5}) \approx 0.618
\end{aligned}$$

Two Point Cubic Interpolation Search
I don't like it. Skip it.

4 Conjugate Gradient

Steepest Desent Method

$$\boldsymbol{p} = -\nabla f(\boldsymbol{x})$$

.

Lemma 4-0 From $\boldsymbol{x}$, along any direction $\boldsymbol{p}$, execute line search one step

$$\varphi(\lambda^*) = \min_{\lambda} \varphi(\lambda) = \min_{\lambda} f(\boldsymbol{x} + \lambda \boldsymbol{p})$$

and obtain $\hat{\boldsymbol{x}} = \boldsymbol{x} + \lambda^* \boldsymbol{p}$, then $\nabla f(\hat{\boldsymbol{x}})$ is orthogonal to $\boldsymbol{p}$, i.e.

$$\nabla f(\hat{\boldsymbol{x}}) \cdot \boldsymbol{p} = 0$$

(Steepest descent method is searching by a way of "zigzag")

Definition 4-1 Suppose $\boldsymbol{p_0}, \boldsymbol{p_1}, \ldots, \boldsymbol{p_{k-1}} (k \leq n)$ are non-zero directions in R^n, and $\boldsymbol{A}$ is $n \times n$ positive-definite matrix. If $\forall i, j, i \neq j$

$$\boldsymbol{p_i}^T \boldsymbol{A}\ \boldsymbol{p_j} = 0$$

then $\boldsymbol{p_0}, \boldsymbol{p_1}, \ldots, \boldsymbol{p_{k-1}}$ is pairwise $\boldsymbol{A}$-conjugate.When $\boldsymbol{A}$ is identity matrix, $\boldsymbol{p_0}, \boldsymbol{p_1}, \ldots, \boldsymbol{p_{k-1}}$ is pairwise orthogonal. So conjuate is generalization of orthogonal.

Theorem 4-2 Suppose $\boldsymbol{A}$ is $n \times n$ positive-definite matrix, if non-zero $\boldsymbol{p_0}, \boldsymbol{p_1}, \ldots, \boldsymbol{p_{k-1}}$ is pairwise $\boldsymbol{A}$-conjugate, then $\boldsymbol{p_0}, \boldsymbol{p_1}, \ldots, \boldsymbol{p_{k-1}}$ is linear independent.

Theorem 4-3 Suppose $f(\boldsymbol{x})$ is positive-definite quadratic function

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^T \boldsymbol{x} + c$$

and non-zero $\boldsymbol{p_0}, \boldsymbol{p_1}, \ldots, \boldsymbol{p_{k-1}}$ is pairwise $\boldsymbol{A}$-conjugate. Starting from any point $\boldsymbol{x_0}$, along $\boldsymbol{p_0}, \boldsymbol{p_1}, \ldots, \boldsymbol{p_{k-1}}$ accordingly, execute line search and arrive at $\boldsymbol{x_k}$

$$\boldsymbol{x_0} \xrightarrow{\boldsymbol{p_0}} \boldsymbol{x_1} \xrightarrow{\boldsymbol{p_1}} \boldsymbol{x_2} \cdots \boldsymbol{x_{k-1}} \xrightarrow{\boldsymbol{p_{k-1}}} \boldsymbol{x_k}$$

then $\nabla f(\boldsymbol{x_k})$ is orthogonal to all previous search directions. i.e.

$$\boldsymbol{p_j^T} \cdot \nabla f(\boldsymbol{x_k}) = 0 \quad (j = 0, 1, \cdots, k-1)$$

Theorem 4-4 Suppose $f(\boldsymbol{x})$ is positive-definite quadratic function

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^T \boldsymbol{x} + c$$

and non-zero $\boldsymbol{p_0}, \boldsymbol{p_1}, \ldots, \boldsymbol{p_{n-1}}$ is pairwise $\boldsymbol{A}$-conjugate. Starting from any point $\boldsymbol{x_0}$, along $\boldsymbol{p_0}, \boldsymbol{p_1}, \ldots, \boldsymbol{p_{n-1}}$ accordingly, execute line search and arrive at $\boldsymbol{x_n}$, then $\boldsymbol{x_n}$ is minumum point.

Conjugate Gradient Method for positive-definite quadratic function $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{x} + c$,

$$
\begin{cases}
\boldsymbol{p_0} = -\nabla f(\boldsymbol{x}) \\
\boldsymbol{p_k} = -\nabla f(\boldsymbol{x_k}) + a_k \boldsymbol{p_{k-1}} \\
a_k = \dfrac{\boldsymbol{p_{k-1}^T} \boldsymbol{A} \nabla f(\boldsymbol{x_k})}{\boldsymbol{p_{k-1}^T} \boldsymbol{A} \boldsymbol{p_{k-1}}} \quad (k = 1, 2, \cdots, n-1)
\end{cases}
$$

Conjugate Gradient Method for any differentiable function $f(\boldsymbol{x})$,

$$
\begin{cases}
\boldsymbol{p_0} = -\nabla f(\boldsymbol{x}) \\
\boldsymbol{p_k} = -\nabla f(\boldsymbol{x_k}) + a_k \boldsymbol{p_{k-1}} \\
a_k = \dfrac{||\nabla f(\boldsymbol{x_k})||^2}{||\nabla f(\boldsymbol{x_{k-1}})||^2} \quad (k = 1, 2, \cdots, n-1)
\end{cases}
$$

More is coming....

Reference

1. Tashan Su. Optimization calculation principle and algorithm program design[M]. 2004.