Introductory Lectures on Optimization

Foundations of Smooth Optimization (1)

Hui Qian
qianhui@zju.edu.cn

College of Computer Science, Zhejiang University

September 19, 2024

Outline

Part I
Relaxation and Approximation

Concepts of Relaxation and Approximation

The majority of general optimization methods are based on the idea of relaxation:

> We call the sequence $\{a_k\}_{k=0}^{\infty}$ a relaxation sequence if
>
> $$a_{k+1} \leq a_k, \quad \forall k \geq 0.$$

Consider several methods for solving the following unconstrained minimization problem

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x}), \tag{1}$$

where $f(\boldsymbol{x})$is a smooth function. In order to do so, we generate a relaxation sequence

$$\{f(\boldsymbol{x}_k)\}_{k=0}^{\infty}, \quad f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k), \; k = 0, 1, \ldots$$

Concepts of Relaxation and Approximation

This strategy has the following important advantages:

1. If $f(\boldsymbol{x})$ is bounded below on $\mathbb{R}^n$, then the sequence $\{f(\boldsymbol{x}_k)\}_{k=0}^{\infty}$ converges.
2. In any case we improve the initial value of the objective function.

However, it would be impossible to implement the idea of relaxation without employing another fundamental principle of numerical analysis, the approximation. In general,

to approximate means to replace an initial complex object by a simplified one, which is close by its properties to the original.

In nonlinear optimization we usually apply local approximations based on derivatives of nonlinear functions. These are commonly the first-order and the second-order approximations (or, the linear and quadratic approximations).

First Order Approximation

Let $f(\boldsymbol{x})$ be differentiable at $\bar{\boldsymbol{x}}$. Then for $\boldsymbol{y} \in \mathbb{R}^n$, we have

$$f(\boldsymbol{y}) = f(\bar{\boldsymbol{x}}) + \langle \nabla f(\bar{\boldsymbol{x}}),\ \boldsymbol{y} - \bar{\boldsymbol{x}} \rangle + o(\|\boldsymbol{y} - \bar{\boldsymbol{x}}\|),$$

where $o(r)$is some function of $r \geq 0$, such that

$$\lim_{r \downarrow 0} \frac{1}{r} o(r) = 0,\ \ o(0) = 0.$$

In the sequel we fix the notation $\|\cdot\|$ for the standard Euclidean norm on $\mathbb{R}^n$:

$$\|\boldsymbol{x}\| = \left[\sum_{i=1}^{n} \left(\boldsymbol{x}^{(i)} \right)^2 \right]^{1/2}.$$

The linear function $f(\bar{\boldsymbol{x}}) + \langle \nabla f(\bar{\boldsymbol{x}}),\ \boldsymbol{y} - \bar{\boldsymbol{x}} \rangle$ is called the linear approximation of f at $\bar{\boldsymbol{x}}$.

First Order Approximation

- The vector $\nabla f(\boldsymbol{x})$ is called the gradient of function f at x.
  Considering the points $\boldsymbol{y}_i = \bar{\boldsymbol{x}} + \epsilon e_i$, where e_i is the i-th coordinate vector in $\mathbb{R}^n$, and taking the limit in $\epsilon \to 0$, we obtain the following coordinate representation of the gradient:

$$\nabla f(\boldsymbol{x}) = \left(\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}^{(1)}}, \ldots, \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}^{(n)}} \right)^\top.$$

- Denote by $\mathcal{L}_f(\alpha)$ the level set of $f(\boldsymbol{x})$:

$$\mathcal{L}_f(\alpha) = \{\boldsymbol{x} \in \mathbb{R}^n | f(\boldsymbol{x}) \leq \alpha\}.$$

- Consider the set of directions that are tangent to $\mathcal{L}_f(f(\bar{\boldsymbol{x}}))$ at $\bar{\boldsymbol{x}}$:

$$S_f(\bar{\boldsymbol{x}}) = \left\{ \boldsymbol{s} \in \mathbb{R}^n | \boldsymbol{s} = \lim_{\boldsymbol{y}_k \to \bar{\boldsymbol{x}}, f(\boldsymbol{y}_k) = f(\bar{\boldsymbol{x}})} \frac{\boldsymbol{y}_k - \bar{\boldsymbol{x}}}{\|\boldsymbol{y}_k - \bar{\boldsymbol{x}}\|} \right\}.$$

First Order Approximation

$$S_f(\bar{\boldsymbol{x}}) = \left\{ \boldsymbol{s} \in \mathbb{R}^n \middle| \boldsymbol{s} = \lim_{\boldsymbol{y}_k \to \bar{\boldsymbol{x}}, f(\boldsymbol{y}_k) = f(\bar{\boldsymbol{x}})} \frac{\boldsymbol{y}_k - \bar{\boldsymbol{x}}}{\|\boldsymbol{y}_k - \bar{\boldsymbol{x}}\|} \right\}.$$

Lemma 1 (Lemma.1.2.1 of Nesterov [2003])

If $\boldsymbol{s} \in S_f(\bar{\boldsymbol{x}})$, then $\langle \nabla f(\bar{\boldsymbol{x}}), \ \boldsymbol{s} \rangle = 0$.

Proof. For $f(\boldsymbol{y}_k) = f(\bar{\boldsymbol{x}})$, we have

$$\begin{aligned} f(\boldsymbol{y}_k) &= f(\bar{\boldsymbol{x}}) + \langle \nabla f(\bar{\boldsymbol{x}}), \ \boldsymbol{y}_k - \bar{\boldsymbol{x}} \rangle + o(\|\boldsymbol{y}_k - \bar{\boldsymbol{x}}\|) \\ &= f(\bar{\boldsymbol{x}}). \end{aligned}$$

Therefore, $\langle \nabla f(\bar{\boldsymbol{x}}), \ \boldsymbol{y}_k - \bar{\boldsymbol{x}} \rangle + o(\|\boldsymbol{y}_k - \bar{\boldsymbol{x}}\|) = 0$. Dividing this equation by $\|\boldsymbol{y}_k - \bar{\boldsymbol{x}}\|$, and taking the limit $\boldsymbol{y}_k \to \bar{\boldsymbol{x}}$, we obtain the result. $\square$

First Order Approximation — Fastest Local Decrease

The direction $-\nabla f(\bar{\boldsymbol{x}})$ (the antigradient) is the direction of the fastest local decrease of $f(\boldsymbol{x})$ at point $\bar{\boldsymbol{x}}$.

Remark. Let $\boldsymbol{s}$ be a direction in $\mathbb{R}^n$, $\|\boldsymbol{s}\| = 1$. Consider the local decrease of $f(\boldsymbol{x})$ along $\boldsymbol{s}$:

$$\Delta(\boldsymbol{s}) = \lim_{\alpha \downarrow 0} \frac{1}{\alpha}[f(\bar{\boldsymbol{x}} + \alpha \boldsymbol{s}) - f(\bar{\boldsymbol{x}})].$$

Note that $f(\bar{\boldsymbol{x}} + \alpha \boldsymbol{s}) - f(\bar{\boldsymbol{x}}) = \alpha \langle \nabla f(\bar{\boldsymbol{x}}),\ \boldsymbol{s} \rangle + o(\alpha \|\boldsymbol{s}\|)$. Therefore, we have

$$\Delta(\boldsymbol{s}) = \langle \nabla f(\bar{\boldsymbol{x}}),\ \boldsymbol{s} \rangle.$$

First Order Approximation — Fastest Local Decrease

The direction $-\nabla f(\bar{\boldsymbol{x}})$ (the antigradient) is the direction of the fastest local decrease of $f(\boldsymbol{x})$ at point $\bar{\boldsymbol{x}}$.

Remark. (Continued.) By leveraging the Cauchy-Schwartz inequality, that is $-\|\boldsymbol{x}\| \cdot \|\boldsymbol{y}\| \leq \langle \boldsymbol{x}, \boldsymbol{y} \rangle \leq \|\boldsymbol{x}\| \cdot \|\boldsymbol{y}\|$, we obtain

$$\Delta(\boldsymbol{s}) = \langle \nabla f(\bar{\boldsymbol{x}}), \boldsymbol{s} \rangle \geq -\|\nabla f(\bar{\boldsymbol{x}})\|.$$

For the lower bound $\bar{\boldsymbol{s}} = -\nabla f(\bar{\boldsymbol{x}}) / \|\nabla f(\bar{\boldsymbol{x}})\|$ （取到下界） , we have

$$\Delta(\bar{\boldsymbol{s}}) = -\langle \nabla f(\bar{\boldsymbol{x}}), \nabla f(\bar{\boldsymbol{x}}) \rangle / \|\nabla f(\bar{\boldsymbol{x}})\| = -\|\nabla f(\bar{\boldsymbol{x}})\|.$$

$\square$

First Order Approximation — First-order Optimality Condition

Theorem 2 (First-order optimality condition.)
Let $\boldsymbol{x}^*$ be a local minimum of differentiable function $f(\boldsymbol{x})$. Then $\nabla f(\boldsymbol{x}^*) = 0$.

Proof. Since $\boldsymbol{x}^*$ is a local minimum of $f(\boldsymbol{x})$, then there exists $r > 0$ such that for all $\boldsymbol{y}$, $\|\boldsymbol{y} - \boldsymbol{x}^*\| \leq r$, we have $f(\boldsymbol{y}) \geq f(\boldsymbol{x}^*)$. Since f is differentiable, this implies that

$$f(\boldsymbol{y}) = f(\boldsymbol{x}^*) + \langle \nabla f(\boldsymbol{x}^*),\ \boldsymbol{y} - \boldsymbol{x}^* \rangle + o(\|\boldsymbol{y} - \boldsymbol{x}^*\|) \geq f(\boldsymbol{x}^*).$$

Thus, for all $\boldsymbol{s}$, $\|\boldsymbol{s}\| = 1$, we have $\langle \nabla f(\boldsymbol{x}^*),\ \boldsymbol{s} \rangle \geq 0$. Consider the directions $\boldsymbol{s}$ and $-\boldsymbol{s}$; we get

$$\langle \nabla f(\boldsymbol{x}^*),\ \boldsymbol{s} \rangle = 0,\ \forall \boldsymbol{s}, \|\boldsymbol{s}\| = 1.$$

Finally, choosing $\boldsymbol{s} = e_i, i = 1 \ldots n$, where e_i is the ith coordinate vector in $\mathbb{R}^n$, we obtain $\nabla f(\boldsymbol{x}^*) = 0$. $\square$

First Order Approximation — First-order Optimality Condition

Note that we have proved only a necessary condition of a local minimum. The points satisfying this condition are called the stationary points of function f.

In order to see that such points are not always the local minima, it is enough to look at the following simple example.

Example 3

$f(x) = x^3, x \in \mathbb{R}^1$, at $x = 0$. (Non-Isolated Critical Points)

First Order Approximation — Useful Corollary

Corollary 4 (Corollary 1.2.1 of Nesterov [2003])

Let $\boldsymbol{x}^*$ be a local minimum of a differentiable function $f(\boldsymbol{x})$ subject to linear equality constraints

$$\boldsymbol{x} \in \mathcal{L} \equiv \{\boldsymbol{x} \in \mathbb{R}^n | A\boldsymbol{x} = b\} \neq \emptyset$$

where A is an $m \times n$ matrix and $\boldsymbol{b} \in \mathbb{R}^m$, $m < n$. Then there exists a vector of multipliers λ^* such that

$$\nabla f(\boldsymbol{x}^*) = A^\top \lambda^*. \tag{2}$$

First Order Approximation — Useful Corollary

Proof. Consider some vectors $\boldsymbol{u}_i, i = 1 \dots k$, that form a basis of the NULL space of matrix A. Then any $\boldsymbol{x} \in \mathcal{L}$ can be represented as follows:

$$\boldsymbol{x} = \boldsymbol{x}(\boldsymbol{y}) \equiv \boldsymbol{x}^* + \sum_{i=1}^{k} \boldsymbol{y}^{(i)} \boldsymbol{u}_i, \ \boldsymbol{y} \in \mathbb{R}^k.$$

Moreover, the point $\boldsymbol{y} = 0$ is a local minimum of the function $\phi(\boldsymbol{y}) = f(\boldsymbol{x}(\boldsymbol{y}))$. In view of Theorem 2, $\nabla\phi(0) = 0$. This means that

$$\frac{\partial \phi(0)}{\partial \boldsymbol{y}^{(i)}} = \frac{\partial \phi(0)}{\partial \boldsymbol{x}(\boldsymbol{y})} \cdot \frac{\partial \boldsymbol{x}(\boldsymbol{y})}{\partial \boldsymbol{y}^{(i)}} = \langle \nabla f(\boldsymbol{x}^*), \ \boldsymbol{u}_i \rangle = 0, \ i = 1 \dots k,$$

and (2) follows. （因为零空间和行空间正交）。 $\square$

Second Order Approximation

Let function $f(\boldsymbol{x})$ be twice differentiable at $\bar{\boldsymbol{x}}$. Then

$$f(\boldsymbol{y}) = f(\bar{\boldsymbol{x}}) + \langle \nabla f(\bar{\boldsymbol{x}}),\ \boldsymbol{y} - \bar{\boldsymbol{x}} \rangle + \frac{1}{2} \langle \nabla^2 f(\bar{\boldsymbol{x}})(\boldsymbol{y} - \bar{\boldsymbol{x}}),\ \boldsymbol{y} - \bar{\boldsymbol{x}} \rangle + o(\|\boldsymbol{y} - \bar{\boldsymbol{x}}\|^2).$$

The quadratic function

$$\boxed{f(\bar{\boldsymbol{x}}) + \langle \nabla f(\bar{\boldsymbol{x}}),\ \boldsymbol{y} - \bar{\boldsymbol{x}} \rangle + \frac{1}{2} \langle \nabla^2 f(\bar{\boldsymbol{x}})(\boldsymbol{y} - \bar{\boldsymbol{x}}),\ \boldsymbol{y} - \bar{\boldsymbol{x}} \rangle}$$

is called the quadratic (or second-order) approximation of function f at $\bar{\boldsymbol{x}}$.

Second Order Approximation

Recall that the $(n \times n)$ matrix $\nabla^2 f(\boldsymbol{x})$ has the following entries:

$$(\nabla^2 f(\boldsymbol{x}))^{(i,j)} = \frac{\partial^2 f(\boldsymbol{x})}{\partial \boldsymbol{x}^{(i)} \partial \boldsymbol{x}^{(j)}}.$$

This matrix is called Hessian of function f at $\boldsymbol{x}$.

Note that the Hessian is a symmetric matrix:

$$\nabla^2 f(\boldsymbol{x}) = [\nabla^2 f(\boldsymbol{x})]^\top.$$

Second Order Approximation

Theorem 5 (Second-order optimality condition.)
Let $\boldsymbol{x}^*$ be a local minimum of a twice differentiable function $f(\boldsymbol{x})$. Then

$$\nabla f(\boldsymbol{x}^*) = 0, \;\; \nabla^2 f(\boldsymbol{x}^*) \succeq 0.$$

Remark. In what follows notation $A \succeq 0$ means that A is positive semidefinite:

$$\langle A\boldsymbol{x}, \; \boldsymbol{x} \rangle \geq 0 \; \forall \boldsymbol{x} \in \mathbb{R}^n.$$

Notation $A \succ 0$ means that A is positive definite (above inequality must be strict for $\boldsymbol{x} \neq 0$).

Second Order Approximation

Theorem 5

Let $\boldsymbol{x}^*$ be a local minimum of a twice differentiable function $f(\boldsymbol{x})$. Then

$$\nabla f(\boldsymbol{x}^*) = 0, \ \nabla^2 f(\boldsymbol{x}^*) \succeq 0.$$

Proof. Since $\boldsymbol{x}^*$ is a local minimum of function $f(\boldsymbol{x})$, there exists $r > 0$ such that for all $\boldsymbol{y}$, $\|\boldsymbol{y} - \boldsymbol{x}^*\| \leq r$, we have

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}^*).$$

In view of Theorem 2, $\nabla f(\boldsymbol{x}^*) = 0$.

Second Order Approximation

Proof. (Continued.) Therefore, for any such $\boldsymbol{y}$,

$$f(\boldsymbol{y}) = f(\boldsymbol{x}^*) + \frac{1}{2}\langle \nabla^2 f(\boldsymbol{x}^*)(\boldsymbol{y}-\boldsymbol{x}^*),\ \boldsymbol{y}-\boldsymbol{x}^*\rangle + o(\|\boldsymbol{y}-\boldsymbol{x}^*\|^2) \geq f(\boldsymbol{x}^*).$$

Thus, $\langle \nabla^2 f(\boldsymbol{x}^*)\boldsymbol{s},\ \boldsymbol{s}\rangle \geq 0$, for all s, $\|\boldsymbol{s}\| = 1$. $\square$

两项都除以 $\|y-x^*\|^2$，调整 y，使得 o 项为零。

Second Order Approximation

Theorem 6

Let function $f(\boldsymbol{x})$ be twice differentiable on $\mathbb{R}^n$ and let $\boldsymbol{x}^*$ satisfy the following conditions:

$$\nabla f(\boldsymbol{x}^*) = 0, \quad \nabla^2 f(\boldsymbol{x}^*) \succ 0.$$

Then $\boldsymbol{x}^*$ is a strict local minimum of $f(\boldsymbol{x})$.

Remark. A point $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ is an unconstrained strict local minimum of a function $f : \mathbb{R}^n \to \mathbb{R}$ if $\exists \epsilon > 0$ such that $f(\bar{\boldsymbol{x}}) < f(\boldsymbol{x})$ for all $\boldsymbol{x} \in B(\bar{\boldsymbol{x}}, \epsilon)$, $\boldsymbol{x} \neq \bar{\boldsymbol{x}}$, where $B(\bar{\boldsymbol{x}}, \epsilon) := \{\boldsymbol{x} | \, \|\boldsymbol{x} - \bar{\boldsymbol{x}}\| \leq \epsilon\}$.

Second Order Approximation

Proof.

Note that in a small neighborhood of point $\boldsymbol{x}^*$ function $f(\boldsymbol{y})$ can be represented as

$$f(\boldsymbol{y}) = f(\boldsymbol{x}^*) + \frac{1}{2}\langle \nabla^2 f(\boldsymbol{x}^*)(\boldsymbol{y} - \boldsymbol{x}^*),\ \boldsymbol{y} - \boldsymbol{x}^* \rangle + o(\|\boldsymbol{y} - \boldsymbol{x}^*\|^2).$$

Let $r = \|\boldsymbol{y} - \boldsymbol{x}^*\|$. Since $\frac{o(r^2)}{r^2} \to 0$ when $r^2 \downarrow 0$, there exists a value $\bar{r}$ such that for all $r \in [0, \bar{r}]$ we have

$$|o(r^2)| \leq \frac{r^2}{4}\lambda_1\left(\nabla^2 f(\boldsymbol{x}^*)\right),$$

where $\lambda_1\left(\nabla^2 f(\boldsymbol{x}^*)\right)$ is the smallest eigenvalue of matrix $\nabla^2 f(\boldsymbol{x}^*)$.

Recall, that in view of our assumption, this eigenvalue is positive.

Second Order Approximation

Proof. (Continued.)

Therefore, for any y, $\|\boldsymbol{y}-\boldsymbol{x}^*\| \leq \bar{r}$. We have

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}^*) + \underbrace{\frac{1}{2}\lambda_1(\nabla^2 f(\boldsymbol{x}^*))\,\|\boldsymbol{y}-\boldsymbol{x}^*\|^2}_{(1)} + \underbrace{o(\|\boldsymbol{y}-\boldsymbol{x}^*\|^2)}_{(2)}$$

$$\geq f(\boldsymbol{x}^*) + \frac{1}{4}\lambda_1(\nabla^2 f(\boldsymbol{x}^*))\,\|\boldsymbol{y}-\boldsymbol{x}^*\|^2 > f(\boldsymbol{x}^*).$$

$\square$

Second Order Approximation

(1)For a symmetric (real) matrix $A \in \mathbb{R}^{n\times n}$, we have

$$\lambda_1(A) \cdot \boldsymbol{x}^\top \boldsymbol{x} \;\leq\; \boldsymbol{x}^\top A \boldsymbol{x} \;\leq\; \lambda_{max}(A) \cdot \boldsymbol{x}^\top \boldsymbol{x}.$$

Therefore, we arrive at

$$\frac{1}{2}\langle \nabla^2 f(\boldsymbol{x}^*)(\boldsymbol{y}-\boldsymbol{x}^*),\, \boldsymbol{y}-\boldsymbol{x}^* \rangle \geq \frac{1}{2}\lambda_1\left(\nabla^2 f(\boldsymbol{x}^*)\right) \|\boldsymbol{y}-\boldsymbol{x}^*\|^2 .$$

(2) According to the settings described above,

$$|o(r^2)| \leq \frac{r^2}{4}\lambda_1\left(\nabla^2 f(\boldsymbol{x}^*)\right).$$

Thus,

$$o(r^2) \geq -\frac{r^2}{4}\lambda_1\left(\nabla^2 f(\boldsymbol{x}^*)\right).$$

The last two items can be combined.

Part II
Classes of differentiable functions

Class $C_L^{k,p}(\mathbb{R}^n)$

Consider a classes of differentiable functions which meet a Lipschitz conditon for a derivative of certain order.

Let Q be a subset of $\mathbb{R}^n$. We denote by $C_L^{k,p}(Q)$ the class of functions with the following properties:

- any $f \in C_L^{k,p}(Q)$ is k times continuously differentiable on Q.
- Its p-th derivative is Lipschitz continuous on Q with the constant L:

$$\left\| f^{(p)}(\boldsymbol{x}) - f^{(p)}(\boldsymbol{y}) \right\| \leq L \|\boldsymbol{x} - \boldsymbol{y}\|$$

for all $x, y \in Q$.

Class $C_L^{k,p}(\mathbb{R}^n)$

Clearly, we always have

1. $p \leq k$。显然成立。
2. if $q \geq k$, then $C_L^{q,p}(Q) \subseteq C_L^{k,p}(Q)$. 例如 $C_L^{2,1}(Q) \subseteq C_L^{1,1}(Q)$。
3. Note also that these classes possess the following property: if $f_1 \in C_{L_1}^{k,p}(Q)$, $f_2 \in C_{L_2}^{k,p}(Q)$ and $\alpha, \beta \in \mathbb{R}^1$, then for

$$L_3 = |\alpha| L_1 + |\beta| L_2,$$

we have $\alpha f_1 + \beta f_2 \in C_{L_3}^{k,p}(Q)$.

Remark. We use notation $f \in C^k(Q)$ for a function f which is k times continuously differentiable on Q.

Class $C_L^{1,1}(\mathbb{R}^n)$

Consider $C_L^{1,1}(\mathbb{R}^n)$, the class of functions with Lipschitz continuous gradient. By definition, the inclusion $f \in C_L^{1,1}(\mathbb{R}^n)$ implies that, for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$,

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq L \|\boldsymbol{x} - \boldsymbol{y}\| . \tag{3}$$

Let us give a sufficient condition for that inclusion.

Lemma 7 (Lemma 1.2.2 of Nesterov [2003])

A function $f(\boldsymbol{x})$ belongs to $C_L^{2,1}(\mathbb{R}^n) \subset C_L^{1,1}(\mathbb{R}^n)$, if and only if

$$\left\|\nabla^2 f(\boldsymbol{x})\right\| \leq L, \quad \forall \boldsymbol{x} \in \mathbb{R}^n. \tag{4}$$

函数的 Lipschitz 性质对应的是更高一阶的导数的界。

Class $C_L^{1,1}(\mathbb{R}^n)$

Lemma 7

A function $f(\boldsymbol{x})$ belongs to $C_L^{2,1}(\mathbb{R}^n) \subset C_L^{1,1}(\mathbb{R}^n)$, if and only if

$$\left\|\nabla^2 f(\boldsymbol{x})\right\| \leq L, \quad \forall \boldsymbol{x} \in \mathbb{R}^n. \tag{4}$$

Proof. Indeed, for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, we have

$$\begin{aligned}
\nabla f(\boldsymbol{y}) &= \nabla f(\boldsymbol{x}) + \int_0^1 \nabla^2 f(\boldsymbol{x} + \tau(\boldsymbol{y} - \boldsymbol{x}))(\boldsymbol{y} - \boldsymbol{x}) d\tau \\
&= \nabla f(\boldsymbol{x}) + \left(\int_0^1 \nabla^2 f(\boldsymbol{x} + \tau(\boldsymbol{y} - \boldsymbol{x})) d\tau \right) \cdot (\boldsymbol{y} - \boldsymbol{x}).
\end{aligned}$$

Class $C_L^{1,1}(\mathbb{R}^n)$

Lemma 7

A function $f(\boldsymbol{x})$ belongs to $C_L^{2,1}(\mathbb{R}^n) \subset C_L^{1,1}(\mathbb{R}^n)$, if and only if

$$\left\|\nabla^2 f(\boldsymbol{x})\right\| \leq L, \quad \forall \boldsymbol{x} \in \mathbb{R}^n. \tag{4}$$

Proof. (Continued.) Therefore, if condition (4) is satisfied then

$$\begin{aligned}
\|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\| &= \left\|\left(\int_0^1 \nabla^2 f(\boldsymbol{x} + \tau(\boldsymbol{y} - \boldsymbol{x}))d\tau\right) \cdot (\boldsymbol{y} - \boldsymbol{x})\right\| \\
&\leq \left\|\int_0^1 \nabla^2 f(\boldsymbol{x} + \tau(\boldsymbol{y} - \boldsymbol{x}))d\tau\right\| \cdot \|(\boldsymbol{y} - \boldsymbol{x})\| \\
&\leq \int_0^1 \underbrace{\left\|\nabla^2 f(\boldsymbol{x} + \tau(\boldsymbol{y} - \boldsymbol{x}))\right\|}_{\leq L} d\tau \cdot \|(\boldsymbol{y} - \boldsymbol{x})\| \leq L \left\|(\boldsymbol{y} - \boldsymbol{x})\right\|.
\end{aligned}$$

Class $C_L^{1,1}(\mathbb{R}^n)$

Lemma 7

A function $f(\boldsymbol{x})$ belongs to $C_L^{2,1}(\mathbb{R}^n) \subset C_L^{1,1}(\mathbb{R}^n)$, if and only if

$$\left\|\nabla^2 f(\boldsymbol{x})\right\| \leq L, \quad \forall \boldsymbol{x} \in \mathbb{R}^n. \tag{4}$$

Proof. (Continued.)

On the other hand, if $f \in C_L^{2,1}(\mathbb{R}^n)$, then for any $\boldsymbol{s} \in \mathbb{R}^n$ and $\alpha > 0$, we have

$$\left\|\left(\int_0^\alpha \nabla^2 f(\boldsymbol{x} + \tau \boldsymbol{s}) d\tau\right) \cdot \boldsymbol{s}\right\| = \|\nabla f(\boldsymbol{x} + \alpha \boldsymbol{s}) - \nabla f(\boldsymbol{x})\| \leq \alpha L \|\boldsymbol{s}\|. \tag{5}$$

Class $C_L^{1,1}(\mathbb{R}^n)$

Proof. (Continued.) Let $\Phi(\alpha) = \left(\int_0^\alpha \nabla^2 f(\boldsymbol{x} + \tau \boldsymbol{s}) d\tau\right)$. Dividing both sides of (5) by α and take $\alpha \downarrow 0$, we have

$$\left\|\left(\lim_{\alpha\downarrow 0} \frac{\Phi(\alpha)}{\alpha}\right) \cdot \boldsymbol{s}\right\| = \left\|\left(\lim_{\alpha\downarrow 0} \frac{\Phi(\alpha) - \Phi(0)}{\alpha - 0}\right) \cdot \boldsymbol{s}\right\| = \left\|\Phi'(0) \cdot \boldsymbol{s}\right\| \leq L \left\|\boldsymbol{s}\right\|.$$

Since $\Phi'(\alpha) = \nabla^2 f(\boldsymbol{x} + \alpha \boldsymbol{s})$, there is $\Phi'(0) = \nabla^2 f(\boldsymbol{x})$ and (4) holds. □

注 1： 矩阵 2 范数：

$$\|A\| = \sup_{\boldsymbol{x}\neq 0} \frac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|}$$

注 2： s 是任意的，因此，上界 $\leq L$ 也成立。

Class $C_L^{1,1}(\mathbb{R}^n)$

Example 8 (Example 1.2.1 of Nesterov [2003])

1. Linear function $f(\boldsymbol{x}) = \alpha + \langle \boldsymbol{a},\ \boldsymbol{x}\rangle \in C_0^{1,1}(\mathbb{R}^n)$, since

$$\nabla f(x) = \boldsymbol{a}, \quad \nabla^2 f(\boldsymbol{x}) = \mathbf{0}$$

2. For the quadratic function $f(\boldsymbol{x}) = \alpha + \langle \boldsymbol{a},\ \boldsymbol{x}\rangle + \frac{1}{2}\langle A\boldsymbol{x},\ \boldsymbol{x}\rangle$ with $A = A^\top$, we have

$$\nabla f(x) = \boldsymbol{a} + A\boldsymbol{x}, \quad \nabla^2 f(\boldsymbol{x}) = A.$$

Therefore $f(\boldsymbol{x}) \in C_L^{1,1}(\mathbb{R}^n)$ with $L = \|A\|$.

Class $C_L^{1,1}(\mathbb{R}^n)$

Example 8

3 Consider the function of one variable $f(x) = \sqrt{1+x^2}$, $x \in \mathbb{R}^1$. We have

$$\nabla f(x) = \frac{x}{\sqrt{1+x^2}}, \quad \nabla^2 f(x) = \frac{1}{(1+x^2)^{3/2}} \leq 1.$$

Therefore $f(x) \in C_1^{1,1}(\mathbb{R}^n)$ with $n = 1$.

Class $C_L^{1,1}(\mathbb{R}^n)$: Geometric Interpretation

The next statement is important for the geometric interpretation of function from $C_L^{1,1}(\mathbb{R}^n)$.

Lemma 9 (Lemma 1.2.3 of Nesterov [2003])
Let $f \in C_L^{1,1}(\mathbb{R}^n)$. Then for any x, y from $\mathbb{R}^n$, we have

$$|f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}),\ \boldsymbol{y} - \boldsymbol{x} \rangle| \leq \frac{L}{2} \|\boldsymbol{y} - \boldsymbol{x}\|^2 . \tag{6}$$

Remark. $f(\boldsymbol{y})$ 和其一阶逼近 $g(\boldsymbol{y}) = f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}),\ \boldsymbol{y} - \boldsymbol{x} \rangle$ 的距离的上界。

Class $C_L^{1,1}(\mathbb{R}^n)$: Geometric Interpretation

$$\left| f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}),\ \boldsymbol{y} - \boldsymbol{x} \rangle \right| \leq \frac{L}{2} \left\| \boldsymbol{y} - \boldsymbol{x} \right\|^2 . \tag{6}$$

Proof.

For all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, we have

$$\begin{aligned} f(\boldsymbol{y}) &= f(\boldsymbol{x}) + \int_0^1 \langle \nabla f(\boldsymbol{x} + \tau(\boldsymbol{y} - \boldsymbol{x})),\ \boldsymbol{y} - \boldsymbol{x} \rangle d\tau \\ &= f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}),\ \boldsymbol{y} - \boldsymbol{x} \rangle + \int_0^1 \langle \nabla f(\boldsymbol{x} + \tau(\boldsymbol{y} - \boldsymbol{x})) - \nabla f(\boldsymbol{x}),\ \boldsymbol{y} - \boldsymbol{x} \rangle d\tau. \end{aligned}$$

Class $C_L^{1,1}(\mathbb{R}^n)$: Geometric Interpretation

Proof. (Continued.) Therefore

$$
\begin{aligned}
|f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}),\, \boldsymbol{y} - \boldsymbol{x}\rangle| &= \left| \int_0^1 \langle \nabla f(\boldsymbol{x} + \tau(\boldsymbol{y} - \boldsymbol{x})) - \nabla f(\boldsymbol{x}),\, \boldsymbol{y} - \boldsymbol{x}\rangle d\tau \right| \\
&\leq \int_0^1 |\langle \nabla f(\boldsymbol{x} + \tau(\boldsymbol{y} - \boldsymbol{x})) - \nabla f(\boldsymbol{x}),\, \boldsymbol{y} - \boldsymbol{x}\rangle| \, d\tau \\
&\leq \int_0^1 \underbrace{\|\nabla f(\boldsymbol{x} + \tau(\boldsymbol{y} - \boldsymbol{x})) - \nabla f(\boldsymbol{x})\|}_{\text{Lipschitz continuous}} \cdot \|\boldsymbol{y} - \boldsymbol{x}\| \, d\tau \\
&\leq \int_0^1 \tau L \|\boldsymbol{y} - \boldsymbol{x}\|^2 \, d\tau = \frac{L}{2} \|\boldsymbol{y} - \boldsymbol{x}\|^2 .
\end{aligned}
$$

$\square$

Class $C_L^{1,1}(\mathbb{R}^n)$: Geometric Interpretation

Consider a function f from $C_L^{1,1}(\mathbb{R}^n)$. Let us fix some $x_0 \in \mathbb{R}^n$ and define two quadratic functions

$$\phi_1(\boldsymbol{x}) = \underbrace{f(\boldsymbol{x}_0) + \langle \nabla f(\boldsymbol{x}_0), \ \boldsymbol{x} - \boldsymbol{x}_0 \rangle}_{\hat{f}(\boldsymbol{x})} + \frac{L}{2} \|\boldsymbol{x} - \boldsymbol{x}_0\|^2 \text{ and}$$

$$\phi_2(\boldsymbol{x}) = f(\boldsymbol{x}_0) + \langle \nabla f(\boldsymbol{x}_0), \ \boldsymbol{x} - \boldsymbol{x}_0 \rangle - \frac{L}{2} \|\boldsymbol{x} - \boldsymbol{x}_0\|^2 .$$

The graph of the function f is located between the graph of ϕ_1 and ϕ_2

$$\phi_1(\boldsymbol{x}) \geq f(\boldsymbol{x}) \geq \phi_2(\boldsymbol{x}), \ \forall \boldsymbol{x} \in \mathbb{R}^n.$$

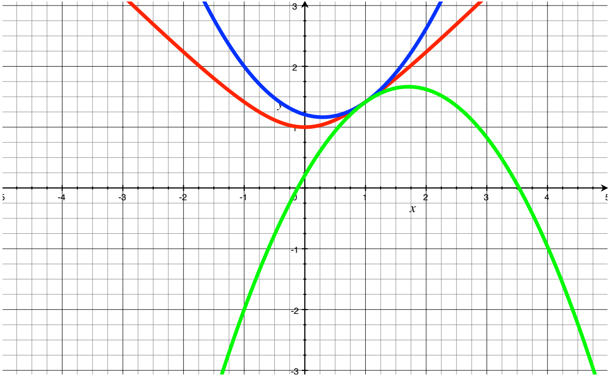
直接推论: $\left| f(\boldsymbol{x}) - \hat{f}(\boldsymbol{x}) \right| \leq \frac{L}{2} \|\boldsymbol{x} - \boldsymbol{x}_0\|^2$

Class $C_L^{1,1}(\mathbb{R}^n)$: Geometric Interpretation

$$f(x) = \sqrt{1+x^2}$$

$$\Phi_1(x) = \sqrt{2} + \frac{1}{\sqrt{2}}(x-1) + \frac{1}{2}(x-1)^2$$

$$\Phi_2(x) = \sqrt{2} + \frac{1}{\sqrt{2}}(x-1) - \frac{1}{2}(x-1)^2$$



Class $C_M^{2,2}(\mathbb{R}^n)$

Consider class $C_M^{2,2}(\mathbb{R}^n)$. That is, for all $x, y \in \mathbb{R}^n$, we have

$$\left\|\nabla^2 f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{y})\right\| \leq M \left\|\boldsymbol{x} - \boldsymbol{y}\right\|. \tag{7}$$

Lemma 10

Let $f \in C_M^{2,2}(\mathbb{R}^n)$. Then for any $\boldsymbol{x}, \boldsymbol{y}$ from $\mathbb{R}^n$ we have

$$\left\|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x})\right\| \leq \frac{M}{2} \left\|\boldsymbol{y} - \boldsymbol{x}\right\|^2. \tag{8}$$

另有一个引理可参见 Lemma 1.2.4 of *Introductory Lectures on Convex Optimization* by Yurii Nesterov。

Class $C_M^{2,2}(\mathbb{R}^n)$

Proof. Let us fix some $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. Then

$$\begin{aligned}
\nabla f(\boldsymbol{y}) &= \nabla f(\boldsymbol{x}) + \int_0^1 \nabla^2 f(\boldsymbol{x} + \tau(\boldsymbol{y} - \boldsymbol{x}))(\boldsymbol{y} - \boldsymbol{x}) d\tau \\
&= \nabla f(\boldsymbol{x}) + \nabla^2 f(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}) \\
&+ \int_0^1 \left(\nabla^2 f(\boldsymbol{x} + \tau(\boldsymbol{y} - \boldsymbol{x})) - \nabla^2 f(\boldsymbol{x})\right)(\boldsymbol{y} - \boldsymbol{x}) d\tau.
\end{aligned}$$

Class $C_M^{2,2}(\mathbb{R}^n)$

Proof. (Continued.) Therefore

$$
\begin{aligned}
\left\|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x})\right\| &= \left\| \int_0^1 \left(\nabla^2 f(\boldsymbol{x}+\tau(\boldsymbol{y}-\boldsymbol{x})) - \nabla^2 f(\boldsymbol{x})\right)(\boldsymbol{y}-\boldsymbol{x}) d\tau \right\| \\
&\leq \int_0^1 \left\| \left(\nabla^2 f(\boldsymbol{x}+\tau(\boldsymbol{y}-\boldsymbol{x})) - \nabla^2 f(\boldsymbol{x})\right)(\boldsymbol{y}-\boldsymbol{x}) \right\| d\tau \\
&\leq \int_0^1 \left\| \nabla^2 f(\boldsymbol{x}+\tau(\boldsymbol{y}-\boldsymbol{x})) - \nabla^2 f(\boldsymbol{x}) \right\| \cdot \|\boldsymbol{y}-\boldsymbol{x}\| \, d\tau \\
&\leq \int_0^1 \tau M \|\boldsymbol{y}-\boldsymbol{x}\|^2 \, d\tau = \frac{M}{2} \|\boldsymbol{y}-\boldsymbol{x}\|^2 .
\end{aligned}
$$

$\square$

Class $C_M^{2,2}(\mathbb{R}^n)$

Corollary 11 (Corollary 1.2.2 of Nesterov [2003])

Let $f \in C_M^{2,2}(\mathbb{R}^n)$ and $\|\boldsymbol{y} - \boldsymbol{x}\| = r$. Then

$$\nabla^2 f(\boldsymbol{x}) - MrI_n \preceq \nabla^2 f(\boldsymbol{y}) \preceq \nabla^2 f(\boldsymbol{x}) + MrI_n,$$

where I_n is the unit matrix in $\mathbb{R}^n$.

（回忆一下，对于矩阵 A 和 B，我们写 $A \succeq B$，如果 $A - B \succeq 0$。）

Proof.

Denote $G = \nabla^2 f(\boldsymbol{y}) - \nabla^2 f(\boldsymbol{x})$. Since G is also a symetric matrix, we have

$$\|G\| = \max |\lambda_i(G)|, \quad i = 1, 2, \cdots, n.$$

Class $C_M^{2,2}(\mathbb{R}^n)$

Proof. (Continued.)

Since $f \in C_M^{2,2}(\mathbb{R}^n)$, we have $\|G\| \leq Mr$. Therefore

$$Mr \geq \|G\| \geq |\lambda_i(G)|, \quad i = 1, 2, \cdots, n.$$

For $\lambda_1(G)$ and $\lambda_{max}(G)$, we have

$$-Mr \leq \lambda_1(G) \leq Mr, \text{ and}$$
$$-Mr \leq \lambda_{max}(G) \leq Mr.$$

Since

$$\lambda_1(G) \cdot \boldsymbol{z}^\top \boldsymbol{z} \;\leq\; \boldsymbol{z}^\top G \boldsymbol{z} \;\leq\; \lambda_{max}(G) \cdot \boldsymbol{z}^\top \boldsymbol{z},$$

Class $C_M^{2,2}(\mathbb{R}^n)$

Proof. (Continued.)

we arrive at

$$-Mr\boldsymbol{z}^\top\boldsymbol{z} \le \lambda_1(G)\cdot\boldsymbol{z}^\top\boldsymbol{z} \;\le\; \boldsymbol{z}^\top G\boldsymbol{z} \;\le\; \lambda_{max}(G)\cdot\boldsymbol{z}^\top\boldsymbol{z} \le Mr\boldsymbol{z}^\top\boldsymbol{z},$$

$$\Rightarrow$$

$$\boldsymbol{z}^\top(-MrI_n)\boldsymbol{z} \le \lambda_1(G)\cdot\boldsymbol{z}^\top\boldsymbol{x} \;\le\; \boldsymbol{z}^\top G\boldsymbol{z} \;\le\; \lambda_{max}(G)\cdot\boldsymbol{z}^\top\boldsymbol{z} \le \boldsymbol{z}^\top(MrI_n)\boldsymbol{z}.$$

Therefore, $-MrI_n \preceq \{G \equiv \nabla^2 f(\boldsymbol{y}) - \nabla^2 f(\boldsymbol{x})\} \preceq MrI_n.$ $\square$

References I

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 2003.

Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

Stephen J Wright and Benjamin Recht. *Optimization for data analysis*. Cambridge University Press, 2022.

Appendix: Cauchy-Schwartz inequality

The Cauchy-Schwarz inequality is a fundamental result in linear algebra and analysis that applies to vectors in an inner product space. It provides an upper bound on the absolute value of the inner product of two vectors. Formally, for any vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ in an inner product space, the inequality states that:

$$|\langle \boldsymbol{u}, \ \boldsymbol{v} \rangle| \leq \|\boldsymbol{u}\| \|\boldsymbol{v}\|.$$

Here:

1. $\langle \boldsymbol{u}, \ \boldsymbol{v} \rangle$denotes the inner product of $\boldsymbol{u}$ and $\boldsymbol{v}$,
2. And $\|\boldsymbol{u}\|$ and $\|\boldsymbol{v}\|$ represent the norms (magnitudes) of $\boldsymbol{u}$ and $\boldsymbol{v}$, respectively.

The inequality shows that the absolute value of the inner product of two vectors is less than or equal to the product of their magnitudes. Equality holds if and only if the vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ are linearly dependent, meaning one is a scalar multiple of the other.

Appendix: Hession Is Symetrical

The Hessian matrix is symmetric because of the Schwarz (Clairaut) theorem on the equality of mixed partial derivatives. This theorem states that if f is a scalar function with continuous second partial derivatives, then the order of differentiation does not matter:

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}.$$

Appendix: Spectral Decomposition of A Symmetric Matrix

If A is a symmetric matrix (i.e., $A = A^\top$), the spectral decomposition theorem states that it can be decomposed as:

$$A = Q\Lambda Q^\top,$$

where:

1. Q is an orthogonal matrix whose columns are the normalized eigenvectors of A, that is $Q = [\boldsymbol{v}_1\ \boldsymbol{v}_2 \cdots \boldsymbol{v}_n]$. Since Q is orthogonal, $Q^\top Q = QQ^\top = I$;
2. $Alpha$ is a diagonal matrix whose entries are the eigenvalues of A. If the eigenvalues are $\lambda_1, \lambda_2, \ldots, \lambda_n$, then $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$;
3. Also, $A = \lambda_1 \boldsymbol{v}_1 \boldsymbol{v}_1^\top + \lambda_2 \boldsymbol{v}_2 \boldsymbol{v}_2^\top \cdots \lambda_n \boldsymbol{v}_n \boldsymbol{v}_n^\top$.

Appendix: Positive Definition Matrix

A matrix A is positive definite if for any non-zero vector $\boldsymbol{x}$, the following condition holds:

$$\boldsymbol{x}^\top A\boldsymbol{x} > 0.$$

This property implies that the matrix A has certain characteristics regarding its eigenvalues:

1. A positive definite matrix is always symmetric.
2. All eigenvalues λ_i of a positive definite matrix A are positive. This can be shown using the Rayleigh quotient:

$$\lambda_{\text{min}} = \min_{\boldsymbol{x}\neq 0} \frac{\boldsymbol{x}^\top A\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}}.$$

   Since A is positive definition, $\boldsymbol{x}^\top A\boldsymbol{x} > 0$ for all non-zero $\boldsymbol{x}$, implying that the minumum eigenvalue λ_{min} must be positive.

Thank You!

Email:qianhui@zju.edu.cn