9. In light of Theorem 4.21, show that under Assumption 4.20, if we want to obtain $\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2 < \varepsilon$, we need an order of $\ln(\ln \varepsilon^{-1})$ iterations for the Newton method.

10. In the Section 4.4.3, show that $\mathcal{L}_k = \{\boldsymbol{\delta}_0, \boldsymbol{\delta}_1, \dots, \boldsymbol{\delta}_{k-1}\}$.

11. In the same section, arrive at the expression (9) for a strictly convex quadratic function.

12. Show that the secant equation is valid for BFGS, DFP and symmetric-rank-one formulae.

13. Given $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$ and a non-singular matrix $\boldsymbol{M} \in \mathbb{R}^{n \times n}$, if $1 + \boldsymbol{v}^T \boldsymbol{M}^{-1} \boldsymbol{u} \neq 0$, then the following formula is valid:

$$(\boldsymbol{M} + \boldsymbol{u}\boldsymbol{v}^T)^{-1} = \boldsymbol{M}^{-1} - \frac{\boldsymbol{M}^{-1}\boldsymbol{u}\boldsymbol{v}^T\boldsymbol{M}^{-1}}{1 + \boldsymbol{v}^T\boldsymbol{M}^{-1}\boldsymbol{u}}. \quad \text{(Sherman-Morrison formula)}$$

Apply this formula to compute the inverses $\boldsymbol{B}_{k+1}$ of $\boldsymbol{H}_{k+1}$ for BFGS, DFP and symmetric-rank-one formulae.

14. Apply the quasi-Newton method with BFGS, DFP, and Symmetric-Rank-One updates for the strictly convex function $f(\boldsymbol{x}) = \alpha + \langle \boldsymbol{a}, \boldsymbol{x} \rangle + \frac{1}{2}\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x} \rangle$ with $\boldsymbol{A} \succ \boldsymbol{O}$.

# 5 Convex Functions and Extended Real-Valued Functions

## 5.1 Convex Functions

**Definition 5.1** Let $Q$ be a subset of $\mathbb{R}^n$. We denote by $\mathcal{F}^k(Q)$ the class of functions $f\colon \mathbb{R}^n \to \mathbb{R}$ with the following properties:

- Any $f \in \mathcal{F}^k(Q)$ is $k$ times continuously differentiable on $Q$;
- $f$ is convex on $Q$, *i.e.*, given $\forall \boldsymbol{x}, \boldsymbol{y} \in Q$ and $\forall \alpha \in [0, 1]$,

$$f(\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y}) \leq \alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{y}).$$

**Theorem 5.2** $f \in \mathcal{F}(\mathbb{R}^n)$ if and only if its epigraph $E := \{(\boldsymbol{x}, y) \in \mathbb{R}^{n+1} \mid f(\boldsymbol{x}) \leq y\}$ is a convex.

*Proof:*
$\boxed{\Rightarrow}$ Let $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2) \in E$. Then for any $0 \leq \alpha \leq 1$, we have

$$f(\alpha\boldsymbol{x}_1 + (1-\alpha)\boldsymbol{x}_2) \leq \alpha f(\boldsymbol{x}_1) + (1-\alpha) f(\boldsymbol{x}_2) \leq \alpha y_1 + (1-\alpha) y_2$$

and therefore $(\alpha\boldsymbol{x}_1 + (1-\alpha)\boldsymbol{x}_2, \alpha y_1 + (1-\alpha)y_2) \in E$.
$\boxed{\Leftarrow}$ Let $(\boldsymbol{x}_1, f(\boldsymbol{x}_1)), (\boldsymbol{x}_2, f(\boldsymbol{x}_2)) \in E$. By the convexity of $E$, for any $0 \leq \alpha \leq 1$,

$$f(\alpha\boldsymbol{x}_1 + (1-\alpha)\boldsymbol{x}_2) \leq \alpha f(\boldsymbol{x}_1) + (1-\alpha) f(\boldsymbol{x}_2)$$

and therefore, $f \in \mathcal{F}(\mathbb{R}^n)$. ∎

**Theorem 5.3** If $f \in \mathcal{F}(\mathbb{R}^n)$, then its $\lambda$-*level set* $L_\lambda := \{\boldsymbol{x} \in \mathbb{R}^n \mid f(\boldsymbol{x}) \leq \lambda\}$ is convex for each $\lambda \in \mathbb{R}$. But the converse is not true.

*Proof:*
For any $\lambda \in \mathbb{R}$, let $\boldsymbol{x}, \boldsymbol{y} \in L_\lambda$. Then for $\forall\alpha \in (0, 1)$, since $f \in \mathcal{F}(\mathbb{R}^n)$, $f(\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y}) \leq \alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{y}) \leq \alpha\lambda + (1-\alpha)\lambda = \lambda$. Therefore, $\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y} \in L_\lambda$.
For the converse, $L_\lambda = \{x \in \mathbb{R} \mid f(x) = x^3 \leq \lambda\}$ is convex for all $\lambda \in \mathbb{R}$, but $f \notin \mathcal{F}(\mathbb{R})$. ∎

**Example 5.4** The function $-\log x$ is convex on $(0, +\infty)$. Let $a, b \in (0, +\infty)$ and $0 \le \theta \le 1$. Then, from the definition of the convexity, we have

$$-\log(\theta a + (1-\theta)b) \le -\theta \log a - (1-\theta)\log b.$$

If we take the exponential of both sides, we obtain

$$a^\theta b^{1-\theta} \le \theta a + (1-\theta)b.$$

For $\theta = \frac{1}{2}$, we have the arithmetic-geometric mean inequality: $\sqrt{ab} \le \dfrac{a+b}{2}$.

Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n \backslash \{\boldsymbol{0}\}$, $p > 1$, and $q$ such that $\frac{1}{p} + \frac{1}{q} = 1$. Consider

$$a = \frac{|[\boldsymbol{x}]_i|^p}{\displaystyle\sum_{j=1}^n |[\boldsymbol{x}]_j|^p}, \; b = \frac{|[\boldsymbol{y}]_i|^q}{\displaystyle\sum_{j=1}^n |[\boldsymbol{y}]_j|^q}, \; \theta = \frac{1}{p}, \text{ and } (1-\theta) = \frac{1}{q}.$$

Then we have

$$\left( \frac{|[\boldsymbol{x}]_i|^p}{\displaystyle\sum_{j=1}^n |[\boldsymbol{x}]_j|^p} \right)^{\frac{1}{p}} \left( \frac{|[\boldsymbol{y}]_i|^q}{\displaystyle\sum_{j=1}^n |[\boldsymbol{y}]_j|^q} \right)^{\frac{1}{q}} \le \frac{|[\boldsymbol{x}]_i|^p}{p\displaystyle\sum_{j=1}^n |[\boldsymbol{x}]_j|^p} + \frac{|[\boldsymbol{y}]_i|^q}{q\displaystyle\sum_{j=1}^n |[\boldsymbol{y}]_j|^q}.$$

and summing over $i$, we obtain the Hölder inequality:

$$|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| \le \|\boldsymbol{x}\|_p \|\boldsymbol{y}\|_q$$

where $\|\boldsymbol{x}\|_p := \left( \displaystyle\sum_{i=1}^n |[\boldsymbol{x}]_i|^p \right)^{\frac{1}{p}}$.

**Theorem 5.5 (Jensen's inequality)** A function $f \colon \mathbb{R}^n \to \mathbb{R}$ is convex if and only if for any positive integer $m$, the following condition is valid

$$\left.\begin{array}{l} \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m \in \mathbb{R}^n \\ \alpha_1, \alpha_2, \ldots, \alpha_m \ge 0 \\ \displaystyle\sum_{i=1}^m \alpha_i = 1 \end{array} \right\} \Rightarrow f\left( \sum_{i=1}^m \alpha_i \boldsymbol{x}_i \right) \le \sum_{i=1}^m \alpha_i f(\boldsymbol{x}_i).$$

*Proof:*
Left for exercise. ∎

**Theorem 5.6** Let $\{f_i\}_{i \in I}$ be a family of (finite or infinite) functions which are bounded from above and $f_i \in \mathcal{F}(\mathbb{R}^n)$. Then, $f(\boldsymbol{x}) := \sup_{i \in I} f_i(\boldsymbol{x})$ is convex on $\mathbb{R}^n$.

*Proof:*
For each $i \in I$, since $f_i \in \mathcal{F}(\mathbb{R}^n)$, its epigraph $E_i = \{(\boldsymbol{x}, y) \in \mathbb{R}^{n+1} \mid f_i(\boldsymbol{x}) \le y\}$ is convex on $\mathbb{R}^{n+1}$ by Theorem 5.2. Also their intersection

$$\bigcap_{i \in I} E_i = \bigcap_{i \in I} \left\{ (\boldsymbol{x}, y) \in \mathbb{R}^{n+1} \mid f_i(\boldsymbol{x}) \le y \right\} = \left\{ (\boldsymbol{x}, y) \in \mathbb{R}^{n+1} \;\middle|\; \sup_{i \in I} f_i(\boldsymbol{x}) \le y \right\}$$

is convex by Exercise 2 of Section 1, which is exactly the epigraph of $f(\boldsymbol{x})$. ∎

## 5.2 Differentiable Convex Functions

**Theorem 5.7** Let $f$ be a continuously differentiable function. The following conditions are equivalent:

1. $f \in \mathcal{F}^1(\mathbb{R}^n)$.
2. $f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \boldsymbol{\nabla f}(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$.
3. $\langle \boldsymbol{\nabla f}(\boldsymbol{x}) - \boldsymbol{\nabla f}(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \geq 0, \ \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$.

*Proof:*
Left for exercise. ∎

**Theorem 5.8 (First-order sufficient optimality condition)** If $f \in \mathcal{F}^1(\mathbb{R}^n)$ and $\boldsymbol{\nabla f}(\boldsymbol{x}^*) = 0$, then $\boldsymbol{x}^*$ is the *global minimum* of $f(\boldsymbol{x})$ on $\mathbb{R}^n$.

*Proof:*
Left for exercise. ∎

**Lemma 5.9** If $f \in \mathcal{F}^1(\mathbb{R}^m)$, $\boldsymbol{b} \in \mathbb{R}^m$, and $\boldsymbol{A}\colon \mathbb{R}^n \to \mathbb{R}^m$, then

$$\phi(\boldsymbol{x}) = f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}) \in \mathcal{F}^1(\mathbb{R}^n).$$

*Proof:*
Left for exercise. ∎

**Example 5.10** The following functions are differentiable and convex:

1. $f(x) = e^x$
2. $f(x) = |x|^p, \quad p > 1$
3. $f(x) = \frac{x^2}{1+|x|}$
4. $f(x) = |x| - \ln(1 + |x|)$
5. $f(\boldsymbol{x}) = \sum_{i=1}^m e^{\alpha_i + \langle \boldsymbol{a}_i, \boldsymbol{x} \rangle}$
6. $f(\boldsymbol{x}) = \sum_{i=1}^m |\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle - b_i|^p, \quad p > 1$

**Theorem 5.11** Let $f$ be a twice continuously differentiable function. Then $f \in \mathcal{F}^2(\mathbb{R}^n)$ if and only if

$$\boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x}) \succeq \boldsymbol{O}, \quad \forall \boldsymbol{x} \in \mathbb{R}^n.$$

*Proof:*
Let $f \in \mathcal{F}^2(\mathbb{R}^n)$, and denote $\boldsymbol{x}_\tau = \boldsymbol{x} + \tau \boldsymbol{s}, \ \tau > 0$. Then, from the previous result

$$\begin{aligned} 0 &\leq \frac{1}{\tau^2} \langle \boldsymbol{\nabla f}(\boldsymbol{x}_\tau) - \boldsymbol{\nabla f}(\boldsymbol{x}), \boldsymbol{x}_\tau - \boldsymbol{x} \rangle = \frac{1}{\tau} \langle \boldsymbol{\nabla f}(\boldsymbol{x}_\tau) - \boldsymbol{\nabla f}(\boldsymbol{x}), \boldsymbol{s} \rangle \\ &= \frac{1}{\tau} \int_0^\tau \langle \boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x} + \lambda \boldsymbol{s}) \boldsymbol{s}, \boldsymbol{s} \rangle d\lambda \\ &= \frac{F(\tau) - F(0)}{\tau} \end{aligned}$$

where $F(\tau) = \int_0^\tau \langle \boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x} + \lambda \boldsymbol{s}) \boldsymbol{s}, \boldsymbol{s} \rangle d\lambda$. Therefore, tending $\tau$ to 0, we get $0 \leq F'(0) = \langle \boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x}) \boldsymbol{s}, \boldsymbol{s} \rangle$, and we have the result.

Conversely, $\forall \boldsymbol{x} \in \mathbb{R}^n$,

$$\begin{aligned} f(\boldsymbol{y}) &= f(\boldsymbol{x}) + \langle \boldsymbol{\nabla f}(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \int_0^1 \int_0^\tau \langle \boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x} + \lambda(\boldsymbol{y} - \boldsymbol{x}))(\boldsymbol{y} - \boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle d\lambda d\tau \\ &\geq f(\boldsymbol{x}) + \langle \boldsymbol{\nabla f}(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle. \end{aligned}$$

∎

## 5.3 Differentiable Convex Functions with Lipschitz Continuous Gradients

**Corollary 5.12** Let $f$ be a two times continuously differentiable function. $f \in \mathcal{F}_L^{2,1}(\mathbb{R}^n)$ if and only if $\boldsymbol{O} \preceq \boldsymbol{\nabla}^2 \boldsymbol{f}(\boldsymbol{x}) \preceq L\boldsymbol{I}, \quad \forall \boldsymbol{x} \in \mathbb{R}^n$.

*Proof:*
Left for exercise. ∎

**Theorem 5.13** Let $f$ be a continuously differentiable function on $\mathbb{R}^n$, $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, and $\alpha \in [0, 1]$. Then the following conditions are equivalent:

1. $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$.
2. $0 \leq f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \boldsymbol{\nabla f}(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \leq \frac{L}{2} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2$.
3. $f(\boldsymbol{x}) + \langle \boldsymbol{\nabla f}(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{1}{2L} \|\boldsymbol{\nabla f}(\boldsymbol{x}) - \boldsymbol{\nabla f}(\boldsymbol{y})\|_2^2 \leq f(\boldsymbol{y})$.
4. $0 \leq \frac{1}{L} \|\boldsymbol{\nabla f}(\boldsymbol{x}) - \boldsymbol{\nabla f}(\boldsymbol{y})\|_2^2 \leq \langle \boldsymbol{\nabla f}(\boldsymbol{x}) - \boldsymbol{\nabla f}(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle$.
5. $0 \leq \langle \boldsymbol{\nabla f}(\boldsymbol{x}) - \boldsymbol{\nabla f}(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \leq L \|\boldsymbol{x} - \boldsymbol{y}\|_2^2$.
6. $f(\alpha \boldsymbol{x} + (1 - \alpha)\boldsymbol{y}) + \frac{\alpha(1-\alpha)}{2L} \|\boldsymbol{\nabla f}(\boldsymbol{x}) - \boldsymbol{\nabla f}(\boldsymbol{y})\|_2^2 \leq \alpha f(\boldsymbol{x}) + (1 - \alpha) f(\boldsymbol{y})$.
7. $0 \leq \alpha f(\boldsymbol{x}) + (1 - \alpha) f(\boldsymbol{y}) - f(\alpha \boldsymbol{x} + (1 - \alpha)\boldsymbol{y}) \leq \alpha(1 - \alpha)\frac{L}{2} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2$.

*Proof:*
$\boxed{1 \Rightarrow 2}$ It follows from Lemmas 5.7 and 3.6.
$\boxed{2 \Rightarrow 3}$ Fix $\boldsymbol{x} \in \mathbb{R}^n$, and consider the function $\phi(\boldsymbol{y}) = f(\boldsymbol{y}) - \langle \boldsymbol{\nabla f}(\boldsymbol{x}), \boldsymbol{y} \rangle$. Clearly $\phi(\boldsymbol{y})$ satisfies 2. Also, $\boldsymbol{y}^* = \boldsymbol{x}$ is a minimal solution. Therefore from 2,

$$\begin{aligned} \phi(\boldsymbol{x}) &= \phi(\boldsymbol{y}^*) \leq \phi\left(\boldsymbol{y} - \frac{1}{L}\boldsymbol{\nabla \phi}(\boldsymbol{y})\right) \leq \phi(\boldsymbol{y}) + \frac{L}{2}\left\|\frac{1}{L}\boldsymbol{\nabla \phi}(\boldsymbol{y})\right\|_2^2 + \langle \boldsymbol{\nabla \phi}(\boldsymbol{y}), -\frac{1}{L}\boldsymbol{\nabla \phi}(\boldsymbol{y}) \rangle \\ &= \phi(\boldsymbol{y}) + \frac{1}{2L}\|\boldsymbol{\nabla \phi}(\boldsymbol{y})\|_2^2 - \frac{1}{L}\|\boldsymbol{\nabla \phi}(\boldsymbol{y})\|_2^2 = \phi(\boldsymbol{y}) - \frac{1}{2L}\|\boldsymbol{\nabla \phi}(\boldsymbol{y})\|_2^2. \end{aligned}$$

Since $\boldsymbol{\nabla \phi}(\boldsymbol{y}) = \boldsymbol{\nabla f}(\boldsymbol{y}) - \boldsymbol{\nabla f}(\boldsymbol{x})$, finally we have

$$f(\boldsymbol{x}) - \langle \boldsymbol{\nabla f}(\boldsymbol{x}), \boldsymbol{x} \rangle \leq f(\boldsymbol{y}) - \langle \boldsymbol{\nabla f}(\boldsymbol{x}), \boldsymbol{y} \rangle - \frac{1}{2L}\|\boldsymbol{\nabla f}(\boldsymbol{y}) - \boldsymbol{\nabla f}(\boldsymbol{x})\|_2^2.$$

$\boxed{3 \Rightarrow 4}$ Adding two copies of 3 with $\boldsymbol{x}$ and $\boldsymbol{y}$ interchanged, we obtain 4.
$\boxed{4 \Rightarrow 1}$ Applying the Cauchy-Schwarz inequality to 4, we obtain $\|\boldsymbol{\nabla f}(\boldsymbol{x}) - \boldsymbol{\nabla f}(\boldsymbol{y})\|_2 \leq L\|\boldsymbol{x} - \boldsymbol{y}\|_2$. Also from Theorem 5.7, $f(\boldsymbol{x})$ is convex.
$\boxed{2 \Rightarrow 5}$ Adding two copies of 2 with $\boldsymbol{x}$ and $\boldsymbol{y}$ interchanged, we obtain 5.
$\boxed{5 \Rightarrow 2}$

$$\begin{aligned} f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \boldsymbol{\nabla f}(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle &= \int_0^1 \langle \boldsymbol{\nabla f}(\boldsymbol{x} + \tau(\boldsymbol{y} - \boldsymbol{x})) - \boldsymbol{\nabla f}(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle d\tau \\ &\leq \int_0^1 \tau L \|\boldsymbol{y} - \boldsymbol{x}\|_2^2 d\tau = \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}\|_2^2. \end{aligned}$$

The non-negativity follows from Theorem 5.7.

$\boxed{3\Rightarrow 6}$ Denote $\boldsymbol{x}_\alpha = \alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y}$. From 3,

$$\begin{aligned} f(\boldsymbol{x}) &\geq f(\boldsymbol{x}_\alpha) + \langle \boldsymbol{\nabla f}(\boldsymbol{x}_\alpha), (1-\alpha)(\boldsymbol{x}-\boldsymbol{y})\rangle + \frac{1}{2L}\|\boldsymbol{\nabla f}(\boldsymbol{x}) - \boldsymbol{\nabla f}(\boldsymbol{x}_\alpha)\|_2^2 \\ f(\boldsymbol{y}) &\geq f(\boldsymbol{x}_\alpha) + \langle \boldsymbol{\nabla f}(\boldsymbol{x}_\alpha), \alpha(\boldsymbol{y}-\boldsymbol{x})\rangle + \frac{1}{2L}\|\boldsymbol{\nabla f}(\boldsymbol{y}) - \boldsymbol{\nabla f}(\boldsymbol{x}_\alpha)\|_2^2. \end{aligned}$$

Multiplying the first inequality by $\alpha$, the second by $1-\alpha$, and summing up, we have

$$\alpha f(\boldsymbol{x}) + (1-\alpha)f(\boldsymbol{y}) \geq f(\boldsymbol{x}_\alpha) + \frac{\alpha}{2L}\|\boldsymbol{\nabla f}(\boldsymbol{x}) - \boldsymbol{\nabla f}(\boldsymbol{x}_\alpha)\|_2^2 + \frac{1-\alpha}{2L}\|\boldsymbol{\nabla f}(\boldsymbol{y}) - \boldsymbol{\nabla f}(\boldsymbol{x}_\alpha)\|_2^2.$$

Finally, using the inequality

$$\alpha\|\boldsymbol{b}-\boldsymbol{d}\|_2^2 + (1-\alpha)\|\boldsymbol{c}-\boldsymbol{d}\|_2^2 \geq \alpha(1-\alpha)\|\boldsymbol{b}-\boldsymbol{c}\|_2^2$$

we have the result.

$$\left(\begin{array}{l} -\alpha(1-\alpha)\|\boldsymbol{b}-\boldsymbol{c}\|_2^2 \geq -\alpha(1-\alpha)(\|\boldsymbol{b}-\boldsymbol{d}\|_2 + \|\boldsymbol{c}-\boldsymbol{d}\|)_2^2 \\ \text{Therefore} \\ \alpha\|\boldsymbol{b}-\boldsymbol{d}\|_2^2 + (1-\alpha)\|\boldsymbol{c}-\boldsymbol{d}\|_2^2 - \alpha(1-\alpha)(\|\boldsymbol{b}-\boldsymbol{d}\|_2 + \|\boldsymbol{c}-\boldsymbol{d}\|_2)^2 \\ = (\alpha\|\boldsymbol{b}-\boldsymbol{d}\|_2 - (1-\alpha)\|\boldsymbol{c}-\boldsymbol{d}\|_2)^2 \geq 0 \end{array}\right)$$

$\boxed{6\Rightarrow 3}$ Dividing both sides by $1-\alpha$ and tending $\alpha$ to 1, we obtain 3.

$\boxed{2\Rightarrow 7}$ Denoting again $\boldsymbol{x}_\alpha = \alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y}$, from 2,

$$\begin{aligned} f(\boldsymbol{x}) &\leq f(\boldsymbol{x}_\alpha) + \langle \boldsymbol{\nabla f}(\boldsymbol{x}_\alpha), (1-\alpha)(\boldsymbol{x}-\boldsymbol{y})\rangle + \frac{L}{2}(1-\alpha)^2\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 \\ f(\boldsymbol{y}) &\leq f(\boldsymbol{x}_\alpha) + \langle \boldsymbol{\nabla f}(\boldsymbol{x}_\alpha), \alpha(\boldsymbol{y}-\boldsymbol{x})\rangle + \frac{L}{2}\alpha^2\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 \end{aligned}$$

Multiplying the first inequality by $\alpha$, the second by $1-\alpha$, and summing up, we have

$$\alpha f(\boldsymbol{x}) + (1-\alpha)f(\boldsymbol{y}) \leq f(\boldsymbol{x}_\alpha) + \frac{L}{2}\left(\alpha(1-\alpha)^2 + (1-\alpha)\alpha^2\right)\|\boldsymbol{x}-\boldsymbol{y}\|_2^2.$$

The non-negativity follows from Theorem 5.7.

$\boxed{7\Rightarrow 2}$ Dividing both sides by $1-\alpha$ and tending $\alpha$ to 1, we obtain 2. The non-negativity follows from Theorem 5.7. ∎

## 5.4 Differentiable Strongly Convex Functions

**Definition 5.14** A continuously differentiable function $f\colon \mathbb{R}^n \to \mathbb{R}$ is called *strongly convex* on $\mathbb{R}^n$ (notation $f \in \mathcal{S}_\mu^1(\mathbb{R}^n)$) if there exists a constant $\mu > 0$ such that

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \boldsymbol{\nabla f}(\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x}\rangle + \frac{1}{2}\mu\|\boldsymbol{y}-\boldsymbol{x}\|_2^2, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n.$$

The constant $\mu$ is called the *convexity parameter* of the function $f$.

**Example 5.15** The following functions are some examples of strongly convex functions:

1. $f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|_2^2$.
2. $f(\boldsymbol{x}) = \alpha + \langle \boldsymbol{a}, \boldsymbol{x}\rangle + \frac{1}{2}\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x}\rangle$, for $\boldsymbol{A} \succeq \mu\boldsymbol{I}$, $\mu > 0$.

3. $|x|$ (Although this function is not differentiable at $0 \in \mathbb{R}$, it is strongly convex only at the same point).

4. A sum of a convex and a strongly convex functions.

5. LASSO (Least Absolute Shrinkage and Selection Operator) with $\text{rank}(\boldsymbol{A}) = n$: $\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 + \lambda\|\boldsymbol{x}\|_1$ and $\lambda > 0$ (notice that this function is also not differentiable at $\boldsymbol{0} \in \mathbb{R}^n$).

6. The $\ell_2$-regularized logistic regression function $f(\boldsymbol{x}) = \log(1 + \exp(-\langle \boldsymbol{a}, \boldsymbol{x}\rangle)) + \lambda\|\boldsymbol{x}\|_2^2$, $\lambda > 0$, which is a sum of a convex function and a strongly convex function.

**Remark 5.16** Strongly convex functions are different from strictly convex functions. For instance, $f(x) = x^4$ is strictly convex at $x = 0$ but it is not strongly convex at the same point.

**Corollary 5.17** If $f \in \mathcal{S}_\mu^1(\mathbb{R}^n)$ and $\boldsymbol{\nabla f}(\boldsymbol{x}^*) = 0$, then

$$f(\boldsymbol{x}) \geq f(\boldsymbol{x}^*) + \frac{1}{2}\mu\|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2, \quad \forall \boldsymbol{x} \in \mathbb{R}^n.$$

*Proof:*
Left for exercise. ∎

**Theorem 5.18** Let $f$ be a continuously differentiable function. The following conditions are equivalent:

1. $f \in \mathcal{S}_\mu^1(\mathbb{R}^n)$.
2. $\mu\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \leq \langle \boldsymbol{\nabla f}(\boldsymbol{x}) - \boldsymbol{\nabla f}(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}\rangle, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$.
3. $f(\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y}) + \alpha(1-\alpha)\frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \leq \alpha f(\boldsymbol{x}) + (1-\alpha)f(\boldsymbol{y}),\ \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n,\ \forall \alpha \in [0, 1]$.

*Proof:*
Left for exercise. ∎

**Theorem 5.19** If $f \in \mathcal{S}_\mu^1(\mathbb{R}^n)$, we have

1. $f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \langle \boldsymbol{\nabla f}(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle + \frac{1}{2\mu}\|\boldsymbol{\nabla f}(\boldsymbol{x}) - \boldsymbol{\nabla f}(\boldsymbol{y})\|_2^2,\ \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$,
2. $\langle \boldsymbol{\nabla f}(\boldsymbol{x}) - \boldsymbol{\nabla f}(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}\rangle \leq \frac{1}{\mu}\|\boldsymbol{\nabla f}(\boldsymbol{x}) - \boldsymbol{\nabla f}(\boldsymbol{y})\|_2^2,\ \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$.

*Proof:*
Let us fix $\boldsymbol{x} \in \mathbb{R}^n$, and define the function $\phi(\boldsymbol{y}) = f(\boldsymbol{y}) - \langle \boldsymbol{\nabla f}(\boldsymbol{x}), \boldsymbol{y}\rangle$. Clearly, $\phi \in \mathcal{S}_\mu^1(\mathbb{R}^n)$. Also, one minimal solution is $\boldsymbol{x}$. Therefore,

$$\begin{aligned} \phi(\boldsymbol{x}) &= \min_{\boldsymbol{v} \in \mathbb{R}^n} \phi(\boldsymbol{v}) \geq \min_{\boldsymbol{v} \in \mathbb{R}^n} \left[\phi(\boldsymbol{y}) + \langle \boldsymbol{\nabla \phi}(\boldsymbol{y}), \boldsymbol{v} - \boldsymbol{y}\rangle + \frac{\mu}{2}\|\boldsymbol{v} - \boldsymbol{y}\|_2^2\right] \\ &= \phi(\boldsymbol{y}) - \frac{1}{2\mu}\|\boldsymbol{\nabla \phi}(\boldsymbol{y})\|_2^2 \end{aligned}$$

as wished. Adding two copies of the 1 with $\boldsymbol{x}$ and $\boldsymbol{y}$ interchanged, we get 2. ∎

**Remark 5.20** The converse of Theorem 5.19 is not valid. For instance, consider $f(x_1, x_2) = x_1^2 - x_2^2$, $\mu = 1$. Then the inequalities 1. and 2. are satisfied but $f \notin \mathcal{S}_\mu^1(\mathbb{R}^2)$ for any $\mu > 0$.

**Theorem 5.21** Let $f$ be a twice continuously differentiable function. Then $f \in \mathcal{S}_\mu^2(\mathbb{R}^n)$ if and only if

$$\boldsymbol{\nabla^2 f}(\boldsymbol{x}) \succeq \mu\boldsymbol{I}, \quad \forall \boldsymbol{x} \in \mathbb{R}^n.$$