4. $0 \le \frac{1}{L}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2^2 \le \langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}\rangle$.
5. $0 \le \langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}\rangle \le L\|\boldsymbol{x} - \boldsymbol{y}\|_2^2$.
6. $f(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}) + \frac{\alpha(1-\alpha)}{2L}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2^2 \le \alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{y})$.
7. $0 \le \alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{y}) - f(\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y}) \le \alpha(1-\alpha)\frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2$.

Proof:

1⇒2 It follows from Lemmas 6.7 and 3.5.

2⇒3 Fix $\boldsymbol{x} \in \mathbb{R}^n$, and consider the function $\phi(\boldsymbol{y}) = f(\boldsymbol{y}) - \langle \nabla f(\boldsymbol{x}), \boldsymbol{y}\rangle$. Clearly $\phi(\boldsymbol{y})$ satisfies 2. Also, $\boldsymbol{y}^* = \boldsymbol{x}$ is a minimal solution. Therefore from 2,

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

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

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

3⇒4 Adding two copies of 3 with $\boldsymbol{x}$ and $\boldsymbol{y}$ interchanged, we obtain 4.

4⇒1 Applying the Cauchy-Schwarz inequality to 4, we obtain $\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2 \le L\|\boldsymbol{x} - \boldsymbol{y}\|_2$. Also from Theorem 6.7, $f(\boldsymbol{x})$ is convex.

2⇒5 Adding two copies of 2 with $\boldsymbol{x}$ and $\boldsymbol{y}$ interchanged, we obtain 5.

5⇒2

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

The non-negativity follows from Theorem 6.7.

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

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

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

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

Finally, using the inequality

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

we have the result.

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

$\boxed{6\Rightarrow3}$ Dividing both sides by $1-\alpha$ and tending α to 1, we obtain 3.
$\boxed{2\Rightarrow7}$ 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 α, 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 6.7.

$\boxed{7\Rightarrow2}$ Dividing both sides by $1-\alpha$ and tending α to 1, we obtain 2. The non-negativity follows from Theorem 6.7. ∎

6.4 Differentiable Strongly Convex Functions

Definition 6.14 A continuously differentiable function $f(\boldsymbol{x})$ 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 μ is called the *convexity parameter* of the function f.

Example 6.15 The following functions are 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. A sum of a convex and a strongly convex functions.

Remark 6.16

1. 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.
2. The ℓ_1-regularized logistic regression function $f(\boldsymbol{x}) = \log(1 + \exp(-\langle \boldsymbol{a}, \boldsymbol{x}\rangle)) + \lambda\|\boldsymbol{x}\|_1$ which is a sum of a convex function and a strongly convex (non-differentiable) function is strongly convex.

Corollary 6.17 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 6.18 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 6.19 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. ∎

The converse of Theorem 6.19 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 6.20 Let f be a twice continuously differentiable function. Then $f \in \mathcal{S}_\mu^2(\mathbb{R}^n)$ if and only if

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

Proof:
Left for exercise. ∎

Corollary 6.21 Let f be a twice continuously differentiable function. Then $f \in \mathcal{S}_{\mu,L}^{2,1}(\mathbb{R}^n)$ if and only if

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

Proof:
Left for exercise. ∎

Theorem 6.22 If $f \in \mathcal{S}_{\mu,L}^{1,1}(\mathbb{R}^n)$, then

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

Proof:
If $\mu = L$, from Theorem 6.18 and the definition of $\mathcal{C}_\mu^1(\mathbb{R}^n)$,

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

and the result follows.