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

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. $0 \leq f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \nabla \boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \leq \frac{L}{2} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2$.
3. $f(\boldsymbol{x}) + \langle \nabla \boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{1}{2L} \|\nabla \boldsymbol{f}(\boldsymbol{x}) - \nabla \boldsymbol{f}(\boldsymbol{y})\|_2^2 \leq f(\boldsymbol{y})$.
4. $0 \leq \frac{1}{L} \|\nabla \boldsymbol{f}(\boldsymbol{x}) - \nabla \boldsymbol{f}(\boldsymbol{y})\|_2^2 \leq \langle \nabla \boldsymbol{f}(\boldsymbol{x}) - \nabla \boldsymbol{f}(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle$.
5. $0 \leq \langle \nabla \boldsymbol{f}(\boldsymbol{x}) - \nabla \boldsymbol{f}(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \leq L \|\boldsymbol{x} - \boldsymbol{y}\|_2^2$.
6. $f(\alpha \boldsymbol{x} + (1 - \alpha)\boldsymbol{y}) + \frac{\alpha(1-\alpha)}{2L} \|\nabla \boldsymbol{f}(\boldsymbol{x}) - \nabla \boldsymbol{f}(\boldsymbol{y})\|_2^2 \leq \alpha f(\boldsymbol{x}) + (1 - \alpha) f(\boldsymbol{y})$.
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 $\phi(\boldsymbol{y}) = f(\boldsymbol{y}) - \langle \nabla \boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y} \rangle$. Clearly $\phi(\boldsymbol{y})$ satisfies 2. Also, $\boldsymbol{y}^* = \boldsymbol{x}$ is a minimal solution. Therefore from 2,

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

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

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

$\boxed{3 \Rightarrow 4}$ 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 $\|\nabla \boldsymbol{f}(\boldsymbol{x}) - \nabla \boldsymbol{f}(\boldsymbol{y})\|_2 \leq L \|\boldsymbol{x} - \boldsymbol{y}\|_2$. 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.
$\boxed{5 \Rightarrow 2}$

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

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,

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

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

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

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 α to 1, we obtain 3.
$\boxed{2\Rightarrow 7}$ 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 5.7.
$\boxed{7\Rightarrow 2}$ Dividing both sides by $1-\alpha$ and tending α 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 $f(\boldsymbol{x})$ 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 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. ~~$|x| \in \mathcal{S}_1^1([0,1])$, but $|x| \notin \mathcal{S}_1^1(\mathbb{R})$.~~ $|x| \in \mathcal{S}_1^1(\{0\})$ (Function $|x|$ is strongly convex only at $0 \in \mathbb{R}$.
4. A sum of a convex and a strongly convex functions.
5. LASSO (Least Absolute Shrinkage and Selection Operator) with rank$(\boldsymbol{A}) = n$: $\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_2^2 + \lambda\|\boldsymbol{x}\|_1$ and $\lambda > 0$.
6. The ℓ_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.