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 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. $|x| \in \mathcal{S}_1^1(\{0\})$ (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 $\text{rank}(\boldsymbol{A}) = n$: $\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 + \lambda\|\boldsymbol{x}\|_1$ 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.

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

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

Proof:
Left for exercise. ∎

Corollary 5.22 Let f be a twice continuously differentiable function. Then $f \in \mathcal{S}^{2,1}_{\mu,L}(\mathbb{R}^n)$ 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 5.23 If $f \in \mathcal{S}^{1,1}_{\mu,L}(\mathbb{R}^n)$, 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 5.18 and the definition of $\mathcal{C}^1_\mu(\mathbb{R}^n)$,

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

and the result follows.

If $\mu < L$, let us define $\phi(\boldsymbol{x}) = f(\boldsymbol{x}) - \frac{\mu}{2}\|\boldsymbol{x}\|_2^2$. Then $\boldsymbol{\nabla \phi}(\boldsymbol{x}) = \boldsymbol{\nabla f}(\boldsymbol{x}) - \mu \boldsymbol{x}$ and $\langle \boldsymbol{\nabla \phi}(\boldsymbol{x}) - \boldsymbol{\nabla \phi}(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}\rangle = \langle \boldsymbol{\nabla f}(\boldsymbol{x}) - \boldsymbol{\nabla f}(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}\rangle - \mu\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \leq (L - \mu)\|\boldsymbol{x} - \boldsymbol{y}\|_2^2$ since $f \in \mathcal{C}^{1,1}_L(\mathbb{R}^n)$. Also $\langle \boldsymbol{\nabla \phi}(\boldsymbol{x}) - \boldsymbol{\nabla \phi}(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}\rangle \geq \mu\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 - \mu\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 = 0$ due to Theorem 5.18. Therefore, from Theorem 5.13, $\phi \in \mathcal{F}^{1,1}_{L-\mu}(\mathbb{R}^n)$.

We have now $\langle \boldsymbol{\nabla \phi}(\boldsymbol{x}) - \boldsymbol{\nabla \phi}(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}\rangle \geq \frac{1}{L-\mu}\|\boldsymbol{\nabla \phi}(\boldsymbol{x}) - \boldsymbol{\nabla \phi}(\boldsymbol{y})\|_2^2$ from Theorem 5.13. Therefore

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

and the result follows after some simplifications. ∎

5.5 Extended Real-Valued Functions

Definition 5.24 A function that can take values $-\infty$ or $+\infty$ is called an *extended real-valued function*. That is $f : \mathbb{R}^n \to [-\infty, +\infty]$. The *domain* of this function is defined by the set $\text{dom}(f) = \{\boldsymbol{x} \in \mathbb{R} \mid f(\boldsymbol{x}) < +\infty\}$.

Example 5.25 For an arbitrary set $S \subset \mathbb{R}^n$, the *indicator function* of S is defined by the following extended real-valued function:

$$\delta_S(\boldsymbol{x}) = \begin{cases} 0, & \boldsymbol{x} \in S, \\ +\infty, & \boldsymbol{x} \notin S. \end{cases}$$

Definition 5.26 A function $f : \mathbb{R}^n \to [-\infty, +\infty]$ is called *proper* if it does not attain the value $-\infty$ and $\text{dom}(f) \neq \emptyset$. This function is called *closed* if its epigraph is a closed set.