Proof:

$\boxed{1\Rightarrow 2}$ It follows from the definition of convex function and Lemma 1.4.4.

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

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

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

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

$\boxed{3\Rightarrow 4}$ 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 $\|f'(\boldsymbol{x}) - f'(\boldsymbol{y})\| \le L\|\boldsymbol{x} - \boldsymbol{y}\|$. Also from Theorem 2.1.5, $f(\boldsymbol{x})$ is convex.

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

$\boxed{5\Rightarrow 2}$

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

The non-negativity follows from Theorem 2.1.5.

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

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

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

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

Finally, using the inequality

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

we have the result.

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

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

$\boxed{2\Rightarrow 7}$ From 2,

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

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

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

The non-negativity follows from Theorem 2.1.5.

$\boxed{7\Rightarrow 2}$ Dividing both sides by $1-\alpha$ and tending α to 1, we obtain 2. The non-negativity follows from Theorem 2.1.5. ∎

2.2 Lower complexity bound for $\mathcal{F}_L^{\infty,1}(\mathbb{R}^n)$ for gradient based methods

Gradient Method: Iterative method $\mathcal{M}$ generated by a sequence such that

$$\boldsymbol{x}_k \in \boldsymbol{x}_0 + \mathrm{Lin}\{f'(\boldsymbol{x}_0), f'(\boldsymbol{x}_1), \ldots, f'(\boldsymbol{x}_{k-1})\}, \quad k \geq 1.$$

Consider the problem class as follows

Model:	$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) \quad \text{where } f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$
Oracle:	First-order local black box
Approximate solution:	Find $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ such that $f(\bar{\boldsymbol{x}}) - f^* < \varepsilon$

Theorem 2.2.1 For any $1 \leq k \leq \frac{n-1}{2}$, and any $\boldsymbol{x}_0 \in \mathbb{R}^n$, there exists a function $f \in \mathcal{F}_L^{\infty,1}(\mathbb{R}^n)$ such that for any first-order method of type $\mathcal{M}$, we have

$$\begin{aligned} f(\boldsymbol{x}_k) - f^* &\geq \frac{3L\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2}{32(k+1)^2}, \\ \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 &\geq \frac{1}{8}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2, \end{aligned}$$

where $\boldsymbol{x}^*$ is the minimum of $f(\boldsymbol{x})$ and $f^* = f(\boldsymbol{x}^*)$.

Proof: This type of methods are invariant with respect to a simultaneous shift of all objects in the space of variables. Therefore, we can assume that $\boldsymbol{x}_0 = \boldsymbol{0}$.

Consider the family of quadratic functions

$$f_k(\boldsymbol{x}) = \frac{L}{4}\left\{\frac{1}{2}\left[[\boldsymbol{x}]_1^2 + \sum_{i=1}^{k-1}([\boldsymbol{x}]_i - [\boldsymbol{x}]_{i+1})^2 + [\boldsymbol{x}]_k^2\right] - [\boldsymbol{x}]_1\right\}, \quad k = 1, 2, \ldots, \lfloor\frac{n-1}{2}\rfloor.$$

We can see that

for $k=1$, $f_1(\boldsymbol{x}) = \frac{L}{4}([\boldsymbol{x}]_1^2 - [\boldsymbol{x}]_1)$,

for $k=2$, $f_2(\boldsymbol{x}) = \frac{L}{4}([\boldsymbol{x}]_1^2 + [\boldsymbol{x}]_2^2 - [\boldsymbol{x}]_1[\boldsymbol{x}]_2 - [\boldsymbol{x}]_1)$,

for $k=3$, $f_3(\boldsymbol{x}) = \frac{L}{4}([\boldsymbol{x}]_1^2 + [\boldsymbol{x}]_2^2 + [\boldsymbol{x}]_3^2 - [\boldsymbol{x}]_1[\boldsymbol{x}]_2 - [\boldsymbol{x}]_2[\boldsymbol{x}]_3 - [\boldsymbol{x}]_1)$.

Also, $f_k'(\boldsymbol{x}) = \frac{L}{4}(\boldsymbol{A}_k\boldsymbol{x} - \boldsymbol{e}_1)$, where $\boldsymbol{e}_1 = (1, 0, \ldots, 0)^T$, and

$$\boldsymbol{A}_k = \begin{pmatrix} 2 & -1 & 0 & 0 & \cdots & 0 & \\ -1 & 2 & -1 & 0 & \cdots & 0 & \\ 0 & -1 & 2 & -1 & \cdots & 0 & \boldsymbol{0}_{k,n-k} \\ \vdots & \ddots & \ddots & \ddots & \cdots & 0 & \\ 0 & \cdots & \cdots & 0 & -1 & 2 & \\ & & \boldsymbol{0}_{n-k,k} & & & & \boldsymbol{0}_{n-k,n-k} \end{pmatrix}.$$

After some calculations, we can show that $L\boldsymbol{I} \succeq f_k''(\boldsymbol{x}) \succeq \boldsymbol{O}$, $\quad k = 1, 2, \ldots, \lfloor \frac{n-1}{2} \rfloor$, and therefore, $f_k(\boldsymbol{x}) \in \mathcal{F}_L^{\infty,1}(\mathbb{R}^n)$, $\quad k = 1, 2, \ldots, \lfloor \frac{n-1}{2} \rfloor$.

Also

$$\begin{aligned} f_k^* \equiv f_k(\bar{\boldsymbol{x}}_k) &= \frac{L}{8}\left(-1 + \frac{1}{k+1}\right), \\ [\bar{\boldsymbol{x}}_k]_i &= \begin{cases} 1 - \frac{i}{k+1}, & i = 1, 2, \ldots, k \\ 0, & i = k+1, k+2, \ldots, n. \end{cases} \end{aligned}$$

Let us take $f(\boldsymbol{x}) = f_{2k+1}(\boldsymbol{x})$, and $\boldsymbol{x}^* = \boldsymbol{x}_{2\bar{k}+1}$.

Then

$$\begin{aligned} f(\boldsymbol{x}_k) - f^* &= f_{2k+1}(\boldsymbol{x}_k) - f_{2k+1}(\boldsymbol{x}_{2\bar{k}+1}) = f_k(\boldsymbol{x}_k) - \frac{L}{8}\left(-1 + \frac{1}{2k+2}\right) \\ &\geq f_k(\bar{\boldsymbol{x}}_k) - \frac{L}{8}\left(-1 + \frac{1}{2k+2}\right) = \frac{L}{8}\left(-1 + \frac{1}{k+1}\right) - \frac{L}{8}\left(-1 + \frac{1}{2k+2}\right). \end{aligned}$$

Note that $\boldsymbol{x}_k \in \{\boldsymbol{x}_0 + \mathrm{Lin}\{f'(\boldsymbol{x}_0), f'(\boldsymbol{x}_1), \ldots, f'(\boldsymbol{x}_{k-1})\}\}$ for $\boldsymbol{x}_0 = \boldsymbol{0}$, and $f_p(\boldsymbol{x}_k) = f_k(\boldsymbol{x}_k)$ for $p \geq k$.

Also $\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 = \|\boldsymbol{x}_k - \boldsymbol{x}_{2\bar{k}+1}\|^2 \geq \sum_{i=k+1}^{2k+1} ([\boldsymbol{x}_{2\bar{k}+1}]_i)^2$.

After some calculations [NESTEROV2004], we have the results. ∎

If we consider very large problems where we can not afford n number of iterations, the above theorem says that:

- The optimal value can be expected to decrease fast.
- The convergence to the optimal solution can be arbitrarily slow.

2.3 Strongly convex functions

Definition 2.3.1 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

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

The constant μ is called the *convexity parameter* of the function f.

Example 2.3.2 The following functions are strongly convex functions:

1. $f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|^2$.
2. $f(\boldsymbol{x}) = \alpha + \langle \boldsymbol{a}, \boldsymbol{x} \rangle + \frac{1}{2}\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x} \rangle$, for $\boldsymbol{A} \succeq \mu \boldsymbol{I}$.
3. A sum of a convex and a strongly convex functions.

Corollary 2.3.3 If $f \in \mathcal{S}_\mu^1(\mathbb{R}^n)$ and $f'(\boldsymbol{x}^*) = 0$, then

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

Proof: Left for exercise. ∎

Theorem 2.3.4 Let f be a continuously differentiable function. The following conditions are equivalent:

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

Proof: Left for exercise. ∎

Theorem 2.3.5 If $f \in \mathcal{S}_\mu^1(\mathbb{R}^n)$, we have

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

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

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

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

The converse of Theorem 2.3.5 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 $\mu > 0$.