If $\mu < L$, let us define $\phi(\boldsymbol{x}) = f(\boldsymbol{x}) - \frac{\mu}{2}\|\boldsymbol{x}\|_2^2$. Then $\nabla\phi(\boldsymbol{x}) = \nabla f(\boldsymbol{x}) - \mu\boldsymbol{x}$ and $\langle\nabla\phi(\boldsymbol{x}) - \nabla\phi(\boldsymbol{y}), \boldsymbol{x}-\boldsymbol{y}\rangle = \langle\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x}-\boldsymbol{y}\rangle - \mu\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 \leq (L-\mu)\|\boldsymbol{x}-\boldsymbol{y}\|_2^2$ since $f \in \mathcal{C}_L^{1,1}(\mathbb{R}^n)$. Also $\langle\nabla\phi(\boldsymbol{x}) - \nabla\phi(\boldsymbol{y}), \boldsymbol{x}-\boldsymbol{y}\rangle \geq \mu\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 - \mu\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 = 0$ due to Theorem 6.18. Therefore, from Theorem 6.13, $\phi \in \mathcal{F}_{L-\mu}^{1,1}(\mathbb{R}^n)$.

We have now $\langle\nabla\phi(\boldsymbol{x}) - \nabla\phi(\boldsymbol{y}), \boldsymbol{x}-\boldsymbol{y}\rangle \geq \frac{1}{L-\mu}\|\nabla\phi(\boldsymbol{x}) - \nabla\phi(\boldsymbol{y})\|_2^2$ from Theorem 6.13. Therefore

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

and the result follows after some simplifications. ∎

6.5 Exercises

1. Given a convex set $S \subseteq \mathbb{R}^n$ and an arbitrarily norm $\|\cdot\|$ in $\mathbb{R}^n$, define the distance of a point $\boldsymbol{x} \in \mathbb{R}^n$ to the set S as
$$\operatorname{dist}(\boldsymbol{x}, S) := \inf_{\boldsymbol{y}\in S} \|\boldsymbol{x}-\boldsymbol{y}\|.$$
Show that the distance function $\operatorname{dist}(\boldsymbol{x}, S)$ is convex on $\boldsymbol{x}$.

2. Give an example of a function $f : \mathbb{R} \to \mathbb{R}$ and a nonempty set $C \subseteq \mathbb{R}$ illustrating each of the following facts:
   (a) f is non convex on $\mathbb{R}$, C is convex, and f is convex on C.
   (b) f is non convex on $\mathbb{R}$, C is non convex, and f is convex on C.

3. Prove Theorem 6.4.

4. Prove Theorem 6.7.

5. Prove Theorem 6.8.

6. Prove Lemma 6.9.

7. Prove Corollary 6.12.

8. Prove Corollary 6.17.

9. Prove Theorem 6.18.

10. Prove Theorem 6.20.

11. Prove Corollary 6.21.

7 Worse Case Analysis for Gradient Based Methods

7.1 Lower Complexity Bound for the class $\mathcal{F}_L^{\infty,1}(\mathbb{R}^n)$

Gradient Based Method: Iterative method $\mathcal{M}$ generated by a sequence such that
$$\boldsymbol{x}_k \in \boldsymbol{x}_0 + \operatorname{Lin}\{\nabla f(\boldsymbol{x}_0), \nabla f(\boldsymbol{x}_1), \ldots, \nabla f(\boldsymbol{x}_{k-1})\}, \quad k \geq 1.$$

Consider the problem class as follows

Model:	$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$ $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$
Oracle:	Only function and gradient values are available
Approximate solution:	Find $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ such that $f(\bar{\boldsymbol{x}}) - f(\boldsymbol{x}^*) < \varepsilon$

Theorem 7.1 For any $1 \le k \le \frac{n-1}{2}$, 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 gradient based method of type $\mathcal{M}$, we have

$$\begin{aligned} f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) &\ge \frac{3L\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2}{32(k+1)^2}, \\ \|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2 &\ge \frac{1}{8}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2, \end{aligned}$$

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

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 $\boldsymbol{x}_0 = \mathbf{0}$.

Consider the family of quadratic functions

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

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 $k = 3$, $f_3(\boldsymbol{x}) = \frac{L}{4}([\boldsymbol{x}]_1^2 + [\boldsymbol{x}]_2^2 + [\boldsymbol{x}]_3^2 - [\boldsymbol{x}]_1[\boldsymbol{x}]_2 - [\boldsymbol{x}]_2[\boldsymbol{x}]_3 - [\boldsymbol{x}]_1)$.
Therefore, $f_k(\boldsymbol{x}) = \frac{L}{4}\left[\frac{1}{2}\langle \boldsymbol{A}_k\boldsymbol{x}, \boldsymbol{x}\rangle - \langle \boldsymbol{e}_1, \boldsymbol{x}\rangle\right]$, where $\boldsymbol{e}_1 = (1, 0, \ldots, 0)^T$, and

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

Also, $\boldsymbol{\nabla f}_k(\boldsymbol{x}) = \frac{L}{4}(\boldsymbol{A}_k\boldsymbol{x} - \boldsymbol{e}_1)$ and $\boldsymbol{\nabla}^2\boldsymbol{f}_k(\boldsymbol{x}) = \frac{L}{4}\boldsymbol{A}_k$. After some calculations, we can show that $L\boldsymbol{I} \succeq \boldsymbol{\nabla}^2\boldsymbol{f}_k(\boldsymbol{x}) \succeq \boldsymbol{O}$ for $k = 1, 2, \ldots, n$, and therefore, $f_k(\boldsymbol{x}) \in \mathcal{F}_L^{\infty,1}(\mathbb{R}^n)$, for $k = 1, 2, \ldots, n$, due to Corollary 6.12.

Then

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

are the minimum value and the minimal solution for $f_k(\cdot)$, respectively.

Now, for $1 \le k \le \frac{n-1}{2}$, let us define $f(\boldsymbol{x}) := f_{2k+1}(\boldsymbol{x})$, and therefore $\boldsymbol{x}^* := \overline{\boldsymbol{x}_{2k+1}}$.

Note that $\boldsymbol{x}_k \in \boldsymbol{x}_0 + \mathrm{Lin}\{\boldsymbol{\nabla f}(\boldsymbol{x}_0), \boldsymbol{\nabla f}(\boldsymbol{x}_1), \ldots, \boldsymbol{\nabla f}(\boldsymbol{x}_{k-1})\}$ for $\boldsymbol{x}_0 = \mathbf{0}$. Moreover, since $\boldsymbol{\nabla f}_k(\boldsymbol{x}) = \frac{L}{4}(\boldsymbol{A}_k\boldsymbol{x} - \boldsymbol{e}_1)$, $[\boldsymbol{x}_k]_p = 0$ for $p > k$. Therefore, $f_p(\boldsymbol{x}_k) = f_k(\boldsymbol{x}_k)$ for $p \geq k$.

Then for $k = 1, 2, \ldots, \lfloor \frac{n-1}{2} \rfloor$,

$$\begin{aligned} f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) &= f_{2k+1}(\boldsymbol{x}_k) - f_{2k+1}(\overline{\boldsymbol{x}_{2k+1}}) = f_k(\boldsymbol{x}_k) - \frac{L}{8}\left(-1 + \frac{1}{2k+2}\right) \\ &\geq f_k(\overline{\boldsymbol{x}_k}) - \frac{L}{8}\left(-1 + \frac{1}{2k+2}\right) = \frac{L}{8}\left(-1 + \frac{1}{k+1}\right) - \frac{L}{8}\left(-1 + \frac{1}{2k+2}\right) \\ &= \frac{L}{16(k+1)}. \end{aligned}$$

After some calculations [Nesterov03], we obtain

$$\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2 = \|\boldsymbol{x}_0 - \overline{\boldsymbol{x}_{2k+1}}\|_2^2 = \sum_{i=1}^{2k+1}\left(1 - \frac{i}{2k+2}\right)^2 \leq \frac{2(k+1)}{3}.$$

Then

$$\frac{f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*)}{\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2} \geq \frac{L}{16(k+1)}\frac{3}{2(k+1)}.$$

Also $\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2 = \|\boldsymbol{x}_k - \overline{\boldsymbol{x}_{2k+1}}\|_2^2 \geq \sum_{i=k+1}^{2k+1}([\overline{\boldsymbol{x}_{2k+1}}]_i)^2 = \sum_{i=k+1}^{2k+1}\left(1 - \frac{i}{2k+2}\right)^2$ and with more calculations [Nesterov03], 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 function value can be expected to decrease fast.
- The convergence to the optimal solution $\boldsymbol{x}^*$ can be arbitrarily slow.

7.2 Lower Complexity Bound for the class $\mathcal{S}_{\mu,L}^{\infty,1}(\mathbb{R}^\infty)$

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

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

Let us define

$$\mathbb{R}^\infty := \ell_2 := \left\{ \{x_i\}_{i=1}^\infty \;\middle|\; \sum_{i=1}^\infty x_i^2 < \infty \right\}.$$

Consider the problem class as follows

Model:	$\min_{\boldsymbol{x}\in\mathbb{R}^\infty} f(\boldsymbol{x})$ $f \in \mathcal{S}_{\mu,L}^{\infty,1}(\mathbb{R}^\infty)$
Oracle:	Only function and gradient values are available
Approximate solution:	Find $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ such that $\begin{cases} f(\bar{\boldsymbol{x}}) - f(\boldsymbol{x}^*) < \varepsilon \\ \|\bar{\boldsymbol{x}} - \boldsymbol{x}^*\|_2^2 < \varepsilon \end{cases}$