Now, since $\boldsymbol{\nabla f}_{\mu,L}(\boldsymbol{x}_0) = -\frac{\mu(L/\mu-1)}{4}\boldsymbol{e}_1$, and $\boldsymbol{A}$ is a tridiagonal matrix, $[\boldsymbol{x}_k]_i = 0$ for $i = k+1, k+2, \ldots$, and

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2 \geq \sum_{i=k+1}^{\infty} [\boldsymbol{x}^*]_i^2 = \sum_{i=k+1}^{\infty} q^{2i} = \frac{q^{2(k+1)}}{1-q^2} = q^{2k}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2.$$

Finally, the first inequality follows from Corollary 5.17. ∎

7 The Steepest Descent Method for Differentiable Convex and Differentiable Strongly Convex Functions with Lipschitz Continuous Gradients

Let us consider the steepest descent method with constant step h.

Theorem 7.1 Let $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$, and $0 < h < \frac{2}{L}$. The steepest descent method with constant step generates a sequence which converges as follows:

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \leq \frac{2(f(\boldsymbol{x}_0) - f(\boldsymbol{x}^*))\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2}{2\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2 + kh(2-Lh)(f(\boldsymbol{x}_0) - f(\boldsymbol{x}^*))}.$$

Proof:
Denote $r_k = \|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2$. Then

$$\begin{aligned} r_{k+1}^2 &= \|\boldsymbol{x}_k - \boldsymbol{x}^* - h\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2 \\ &= r_k^2 - 2h\langle \boldsymbol{\nabla f}(\boldsymbol{x}_k), \boldsymbol{x}_k - \boldsymbol{x}^*\rangle + h^2\|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2 \\ &= r_k^2 - 2h\langle \boldsymbol{\nabla f}(\boldsymbol{x}_k) - \boldsymbol{\nabla f}(\boldsymbol{x}^*), \boldsymbol{x}_k - \boldsymbol{x}^*\rangle + h^2\|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2 \\ &\leq r_k^2 - h\left(\frac{2}{L} - h\right)\|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2, \end{aligned}$$

where the last inequality follows from Theorem 5.13.
Therefore, since $0 < h < \frac{2}{L}$, $r_{k+1} < r_k < \cdots < r_0$.
Now

$$\begin{aligned} f(\boldsymbol{x}_{k+1}) &\leq f(\boldsymbol{x}_k) + \langle \boldsymbol{\nabla f}(\boldsymbol{x}_k), \boldsymbol{x}_{k+1} - \boldsymbol{x}_k\rangle + \frac{L}{2}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|_2^2 \\ &= f(\boldsymbol{x}_k) - h\|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2 + \frac{L}{2}\| - h\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2 \qquad (12) \\ &= f(\boldsymbol{x}_k) - \omega\|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2 < f(\boldsymbol{x}_k), \qquad (13) \end{aligned}$$

where $\omega = h(1 - \frac{L}{2}h)$. Denoting by $\Delta_k = f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*)$, from the convexity of $f(\boldsymbol{x})$, Theorem 5.7, and the Cauchy-Schwarz inequality,

$$\Delta_k = f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \leq \langle \boldsymbol{\nabla f}(\boldsymbol{x}_k), \boldsymbol{x}_k - \boldsymbol{x}^*\rangle \leq \|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2 r_k \leq \|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2 r_0. \qquad (14)$$

Combining (13) and (14),

$$\Delta_{k+1} \leq \Delta_k - \frac{\omega}{r_0^2}\Delta_k^2.$$

Thus dividing by $\Delta_k\Delta_{k+1}$,

$$\frac{1}{\Delta_{k+1}} \geq \frac{1}{\Delta_k} + \frac{\omega}{r_0^2}\frac{\Delta_k}{\Delta_{k+1}} \geq \frac{1}{\Delta_k} + \frac{\omega}{r_0^2}.$$

since $\frac{\Delta_k}{\Delta_{k+1}} \geq 1$. Summing up these inequalities we get

$$\frac{1}{\Delta_{k+1}} \geq \frac{1}{\Delta_0} + \frac{\omega}{r_0^2}(k+1).$$

■

To obtain the optimal step size, it is sufficient to find the maximum of the function $\omega := \omega(h) = h(1 - \frac{L}{2}h)$ which is $h^* := 1/L$.

Corollary 7.2 If $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$, the steepest descent method with constant step $h = 1/L$ yields

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \leq \frac{2L\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2}{k+4}.$$

That is, $\{f(\boldsymbol{x}_k)\}_{k=0}^{\infty}$ converges R-sublinearly to $f(\boldsymbol{x}^*)$.

Proof:
Left for exercise. ■

Theorem 7.3 Let $f \in \mathcal{S}_{\mu,L}^{1,1}(\mathbb{R}^n)$, and $0 < h \leq \frac{2}{\mu+L}$. The steepest descent method with constant step generates a sequence which converges as follows:

$$\begin{aligned} f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) &\leq \frac{L}{2}\left(1 - \frac{2h\mu L}{\mu + L}\right)^k \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2, \\ \|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2 &\leq \left(1 - \frac{2h\mu L}{\mu + L}\right)^k \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2. \end{aligned}$$

If $h = \frac{2}{\mu+L}$, then

$$\begin{aligned} f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) &\leq \frac{L}{2}\left(\frac{L/\mu - 1}{L/\mu + 1}\right)^{2k} \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2, \\ \|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2 &\leq \left(\frac{L/\mu - 1}{L/\mu + 1}\right)^k \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2. \end{aligned}$$

That is, $\{\boldsymbol{x}_k\}_{k=0}^{\infty}$ and $\{f(\boldsymbol{x}_k)\}_{k=0}^{\infty}$ converges R-linearly to $\boldsymbol{x}^*$ and $f(\boldsymbol{x}^*)$, respectively.

Proof:
Denote $r_k = \|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2$. Then

$$\begin{aligned} r_{k+1}^2 &= \|\boldsymbol{x}_k - \boldsymbol{x}^* - h\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2 \\ &= r_k^2 - 2h\langle \boldsymbol{\nabla f}(\boldsymbol{x}_k), \boldsymbol{x}_k - \boldsymbol{x}^*\rangle + h^2\|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2 \\ &= r_k^2 - 2h\langle \boldsymbol{\nabla f}(\boldsymbol{x}_k) - \boldsymbol{\nabla f}(\boldsymbol{x}^*), \boldsymbol{x}_k - \boldsymbol{x}^*\rangle + h^2\|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2 \\ &\leq r_k^2 - 2h\left(\frac{\mu L}{\mu + L}r_k^2 + \frac{1}{\mu + L}\|\boldsymbol{\nabla f}(\boldsymbol{x}_k) - \boldsymbol{\nabla f}(\boldsymbol{x}^*)\|_2^2\right) + h^2\|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2 \\ &= \left(1 - \frac{2h\mu L}{\mu + L}\right) r_k^2 + h\left(h - \frac{2}{\mu + L}\right)\|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2 \end{aligned}$$

from Theorems 5.13 and 5.23, and it proves the first two inequalities.

Now, for $h = 2/(L + \mu)$ and again from Theorem 5.13,

$$\begin{aligned} f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) - \langle \boldsymbol{\nabla f}(\boldsymbol{x}^*), \boldsymbol{x}_k - \boldsymbol{x}^*\rangle &\leq \frac{L}{2}\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2 \\ &\leq \frac{L}{2}\left(\frac{L/\mu - 1}{L/\mu + 1}\right)^{2k} r_0^2. \end{aligned}$$

■

Theorem 7.4 (Yuan 2010) [2] In the special case of a strongly convex quadratic function $f(\boldsymbol{x}) = \frac{1}{2}\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x}\rangle + \langle \boldsymbol{a}, \boldsymbol{x}\rangle + \alpha$ with $\lambda_1(\boldsymbol{A}) = L \geq \lambda_n(\boldsymbol{A}) = \mu > 0$, we can obtain

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2 \leq \left(\frac{L/\mu - 1}{L/\mu + \sqrt{\frac{\mu}{2L}}} \right)^k \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2$$

for the steepest descent method with "exact line search".

- Note that the previous result for the steepest descent method, Theorem 4.18, was only a local result. Theorems 7.1 and 7.3 guarantee that the steepest descent method converges for any starting point $\boldsymbol{x}_0 \in \mathbb{R}^n$ (due to convexity).
- Comparing the rate of convergence of the steepest descent method for the classes $\mathcal{F}_L^{1,1}(\mathbb{R}^n)$ and $\mathcal{S}_{\mu,L}^{1,1}(\mathbb{R}^n)$ (Theorems 7.1, Corollary 7.2, and 7.3, respectively) with their lower complexity bounds (Theorems 6.1 and 6.2, respectively), we possible have a huge gap.

7.1 Exercises

1. Prove Corollary 7.2.
2. Consider a sequence $\{\beta_k\}_{k=0}^{\infty}$ which converges to zero.

   The sequence is said to converge *Q-sublinearly* if

$$\lim_{k\to\infty} \sup \left| \frac{\beta_{k+1}}{\beta_k} \right| = 1.$$

   A zero converging sequence $\{\beta_k\}_{k=0}^{\infty}$ is said to converge *R-sublinearly* if it is dominated by a Q-sublinearly converging sequence. That is, if there is a Q-sublinearly converging sequence $\{\hat{\beta}_k\}_{k=0}^{\infty}$ such that $0 \leq |\beta_k| \leq \hat{\beta}_k$.

   (a) Give an example of a Q-sublinear converging sequence which is not Q-linear converging sequence.

   (b) Give an example of a R-sublinear converging sequence which is not R-linear converging sequence.

8 The Optimal Gradient Method (First-Order Method, Accelerated Gradient Method, Fast Gradient Method)

This algorithm was proposed for the first time by Nesterov[3] in 1983. In [Nesterov03, Nesterov18], he gives a reinterpretation of the algorithm and provides another justification of it which attains the same complexity bound of the original article.

Definition 8.1 A pair of sequences $\{\phi_k(\boldsymbol{x})\}_{k=0}^{\infty}$ and $\{\lambda_k\}_{k=0}^{\infty}$ with $\lambda_k \geq 0$ is called an *estimate sequence* of the function $f(\boldsymbol{x})$ if

$$\lambda_k \to 0,$$

and for any $\boldsymbol{x} \in \mathbb{R}^n$ and any $k \geq 0$, we have

$$\phi_k(\boldsymbol{x}) \leq (1 - \lambda_k) f(\boldsymbol{x}) + \lambda_k \phi_0(\boldsymbol{x}).$$

[2] Y.-X. Yuan, "A short note on the Q-linear convergence of the steepest descent method", *Mathematical Programming* **123** (2010), pp. 339–343.

[3] Y. Nesterov, "A method for solving the convex programming problem with convergence rate $\mathcal{O}(1/k^2)$," *Dokl. Akad. Nauk SSSR* **269** (1983), pp. 543–547.

Lemma 8.2 Given an estimate sequence $\{\phi_k(\boldsymbol{x})\}_{k=0}^{\infty}$, $\{\lambda_k\}_{k=0}^{\infty}$, and if for some sequence $\{\boldsymbol{x}_k\}_{k=0}^{\infty}$ we have

$$f(\boldsymbol{x}_k) \leq \phi_k^* := \min_{\boldsymbol{x}\in\mathbb{R}^n} \phi_k(\boldsymbol{x})$$

then $f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \leq \lambda_k(\phi_0(\boldsymbol{x}^*) - f(\boldsymbol{x}^*)) \to 0$.

Proof:
It follows from the definition. ∎

Lemma 8.3 Assume that

1. $f \in \mathcal{S}_\mu^1(\mathbb{R}^n)$, possible with $\mu = 0$ (which means that $f \in \mathcal{F}^1(\mathbb{R}^n)$).
2. $\phi_0(\boldsymbol{x})$ is an arbitrary function on $\mathbb{R}^n$.
3. $\{\boldsymbol{y}_k\}_{k=0}^{\infty}$ is an arbitrary sequence in $\mathbb{R}^n$.
4. $\{\alpha_k\}_{k=-1}^{\infty}$ is an arbitrary sequence such that $\alpha_{-1} = 0$, $\alpha_k \in (0,1] \quad (k = 0,1,\ldots)$, and $\displaystyle\sum_{k=0}^{\infty} \alpha_k = \infty$.

Then the pair of sequences $\left\{\displaystyle\prod_{i=-1}^{k-1}(1-\alpha_i)\right\}_{k=0}^{\infty}$ and $\{\phi_k(\boldsymbol{x})\}_{k=0}^{\infty}$ recursively defined as

$$\phi_{k+1}(\boldsymbol{x}) \quad = \quad (1-\alpha_k)\phi_k(\boldsymbol{x}) + \alpha_k\left[f(\boldsymbol{y}_k) + \langle \boldsymbol{\nabla f}(\boldsymbol{y}_k), \boldsymbol{x} - \boldsymbol{y}_k\rangle + \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{y}_k\|_2^2\right]$$

is an estimate sequence.

Proof:
Let us prove by induction in k. For $k = 0$, $\phi_0(\boldsymbol{x}) = (1-(1-\alpha_{-1}))\,f(\boldsymbol{x}) + (1-\alpha_{-1})\phi_0(\boldsymbol{x})$ since $\alpha_{-1} = 0$. Suppose that the induction hypothesis is valid for any index equal or smaller than k. Since $f \in \mathcal{S}_\mu^1(\mathbb{R}^n)$,

$$\begin{aligned}
\phi_{k+1}(\boldsymbol{x}) &= (1-\alpha_k)\phi_k(\boldsymbol{x}) + \alpha_k\left[f(\boldsymbol{y}_k) + \langle \boldsymbol{\nabla f}(\boldsymbol{y}_k), \boldsymbol{x} - \boldsymbol{y}_k\rangle + \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{y}_k\|_2^2\right] \\
&\leq (1-\alpha_k)\phi_k(\boldsymbol{x}) + \alpha_k f(\boldsymbol{x}) \\
&= \left(1-(1-\alpha_k)\prod_{i=-1}^{k-1}(1-\alpha_i)\right)f(\boldsymbol{x}) + (1-\alpha_k)\left(\phi_k(\boldsymbol{x}) - \left(1-\prod_{i=-1}^{k-1}(1-\alpha_i)\right)f(\boldsymbol{x})\right) \\
&\leq \left(1-(1-\alpha_k)\prod_{i=-1}^{k-1}(1-\alpha_i)\right)f(\boldsymbol{x}) + (1-\alpha_k)\prod_{i=-1}^{k-1}(1-\alpha_i)\phi_0(\boldsymbol{x}) \\
&= \left(1-\prod_{i=-1}^{k}(1-\alpha_i)\right)f(\boldsymbol{x}) + \prod_{i=-1}^{k}(1-\alpha_i)\phi_0(\boldsymbol{x}).
\end{aligned}$$

Now, it remains to show that $\prod_{i=-1}^{k-1}(1-\alpha_i) \to 0$. This is equivalent to show that $\log\prod_{i=-1}^{k-1}(1-\alpha_i) \to -\infty$. Using the inequality $\log(1-a) \leq -a$ for $a \in (-\infty, 1)$, we have

$$\log\prod_{i=-1}^{k-1}(1-\alpha_i) = \sum_{i=-1}^{k-1}\log(1-\alpha_i) \leq -\sum_{i=-1}^{k-1}\alpha_i \to -\infty$$

due to our assumption. ∎

Lemma 8.4 Let $f : \mathbb{R}^n \to \mathbb{R}$ be an arbitrary continuously differentiable function. Also let $\phi_0^* \in \mathbb{R}$, $\mu \geq 0$, $\gamma_0 \geq 0$, $\boldsymbol{v}_0 \in \mathbb{R}^n$, $\{\boldsymbol{y}_k\}_{k=0}^{\infty}$, and $\{\alpha_k\}_{k=0}^{\infty}$ given arbitrarily sequences such that $\alpha_{-1} = 0$, $\alpha_k \in (0,1] \quad (k = 0,1,\ldots)$. In the special case of $\mu = 0$, we further assume that $\gamma_0 > 0$ and $\alpha_k < 1 \quad (k = 0,1,\ldots)$. Let $\phi_0(\boldsymbol{x}) = \phi_0^* + \frac{\gamma_0}{2}\|\boldsymbol{x} - \boldsymbol{v}_0\|_2^2$. If we define recursively $\phi_{k+1}(\boldsymbol{x})$ such as in the previous lemma:

$$\phi_{k+1}(\boldsymbol{x}) = (1-\alpha_k)\phi_k(\boldsymbol{x}) + \alpha_k \left[f(\boldsymbol{y}_k) + \langle \boldsymbol{\nabla f}(\boldsymbol{y}_k), \boldsymbol{x} - \boldsymbol{y}_k \rangle + \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{y}_k\|_2^2 \right],$$

$\phi_{k+1}(\boldsymbol{x})$ preserve the canonical form

$$\phi_{k+1}(\boldsymbol{x}) = \phi_{k+1}^* + \frac{\gamma_{k+1}}{2}\|\boldsymbol{x} - \boldsymbol{v}_{k+1}\|_2^2 \tag{15}$$

for

$$\begin{aligned}
\gamma_{k+1} &= (1-\alpha_k)\gamma_k + \alpha_k \mu, \\
\boldsymbol{v}_{k+1} &= \frac{1}{\gamma_{k+1}}[(1-\alpha_k)\gamma_k \boldsymbol{v}_k + \alpha_k \mu \boldsymbol{y}_k - \alpha_k \boldsymbol{\nabla f}(\boldsymbol{y}_k)], \\
\phi_{k+1}^* &= (1-\alpha_k)\phi_k^* + \alpha_k f(\boldsymbol{y}_k) - \frac{\alpha_k^2}{2\gamma_{k+1}}\|\boldsymbol{\nabla f}(\boldsymbol{y}_k)\|_2^2 \\
&\quad + \frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}} \left(\frac{\mu}{2}\|\boldsymbol{y}_k - \boldsymbol{v}_k\|_2^2 + \langle \boldsymbol{\nabla f}(\boldsymbol{y}_k), \boldsymbol{v}_k - \boldsymbol{y}_k \rangle \right).
\end{aligned}$$

Proof:

We will use again the induction hypothesis in k. Note that $\boldsymbol{\nabla}^2 \boldsymbol{\phi}_0(\boldsymbol{x}) = \gamma_0 \boldsymbol{I}$. Now, for any $k \geq 0$,

$$\boldsymbol{\nabla}^2 \boldsymbol{\phi}_{k+1}(\boldsymbol{x}) = (1-\alpha_k)\boldsymbol{\nabla}^2 \boldsymbol{\phi}_k(\boldsymbol{x}) + \alpha_k \mu \boldsymbol{I} = ((1-\alpha_k)\gamma_k + \alpha_k \mu)\,\boldsymbol{I} = \gamma_{k+1}\boldsymbol{I}.$$

Therefore, $\phi_{k+1}(\boldsymbol{x})$ is a quadratic function of the form (15). Also, $\gamma_{k+1} > 0$ since $\mu > 0$ and $\alpha_k > 0 \quad (k = 0,1,\ldots)$; or if $\mu = 0$, we assumed that $\gamma_0 > 0$ and $\alpha_k \in (0,1) \quad (k = 0,1,\ldots)$.

From the first-order optimality condition

$$\begin{aligned}
\boldsymbol{\nabla \phi}_{k+1}(\boldsymbol{x}) &= (1-\alpha_k)\boldsymbol{\nabla \phi}_k(\boldsymbol{x}) + \alpha_k \boldsymbol{\nabla f}(\boldsymbol{y}_k) + \alpha_k \mu(\boldsymbol{x} - \boldsymbol{y}_k) \\
&= (1-\alpha_k)\gamma_k(\boldsymbol{x} - \boldsymbol{v}_k) + \alpha_k \boldsymbol{\nabla f}(\boldsymbol{y}_k) + \alpha_k \mu(\boldsymbol{x} - \boldsymbol{y}_k) = 0.
\end{aligned}$$

Thus,

$$\boldsymbol{x} = \boldsymbol{v}_{k+1} = \frac{1}{\gamma_{k+1}}\left[(1-\alpha_k)\gamma_k \boldsymbol{v}_k + \alpha_k \mu \boldsymbol{y}_k - \alpha_k \boldsymbol{\nabla f}(\boldsymbol{y}_k)\right]$$

is the minimal optimal solution of $\phi_{k+1}(\boldsymbol{x})$.

Finally, from what we proved so far and from the definition

$$\begin{aligned}
\phi_{k+1}(\boldsymbol{y}_k) &= \phi_{k+1}^* + \tfrac{\gamma_{k+1}}{2}\|\boldsymbol{y}_k - \boldsymbol{v}_{k+1}\|_2^2 \\
&= (1-\alpha_k)\phi_k(\boldsymbol{y}_k) + \alpha_k f(\boldsymbol{y}_k) \\
&= (1-\alpha_k)\left(\phi_k^* + \tfrac{\gamma_k}{2}\|\boldsymbol{y}_k - \boldsymbol{v}_k\|_2^2\right) + \alpha_k f(\boldsymbol{y}_k).
\end{aligned} \tag{16}$$

Now,

$$\boldsymbol{v}_{k+1} - \boldsymbol{y}_k = \frac{1}{\gamma_{k+1}}\left[(1-\alpha_k)\gamma_k(\boldsymbol{v}_k - \boldsymbol{y}_k) - \alpha_k \boldsymbol{\nabla f}(\boldsymbol{y}_k)\right].$$

Therefore,

$$\begin{aligned}
\tfrac{\gamma_{k+1}}{2}\|\boldsymbol{v}_{k+1} - \boldsymbol{y}_k\|_2^2 &= \tfrac{1}{2\gamma_{k+1}}\left[(1-\alpha_k)^2\gamma_k^2\|\boldsymbol{v}_k - \boldsymbol{y}_k\|_2^2 + \alpha_k^2\|\boldsymbol{\nabla f}(\boldsymbol{y}_k)\|_2^2 \right. \\
&\quad \left. - 2\alpha_k(1-\alpha_k)\gamma_k\langle \boldsymbol{\nabla f}(\boldsymbol{y}_k), \boldsymbol{v}_k - \boldsymbol{y}_k \rangle\right].
\end{aligned} \tag{17}$$

Substituting (17) into (16), we obtain the expression for ϕ_{k+1}^*. ∎

Theorem 8.5 Let $L \geq \mu \geq 0$. Consider $f \in \mathcal{S}^{1,1}_{\mu,L}(\mathbb{R}^n)$, possible with $\mu = 0$ (which means that $f \in \mathcal{F}^{1,1}_L(\mathbb{R}^n)$). For given $\boldsymbol{x}_0 \in \mathbb{R}^n$, let us choose $\phi_0^* = f(\boldsymbol{x}_0)$ and $\boldsymbol{v}_0 := \boldsymbol{x}_0$. Consider also $\gamma_0 > 0$ such that $L \geq \gamma_0 \geq \mu \geq 0$. Define the sequences $\{\alpha_k\}_{k=-1}^{\infty}$, $\{\gamma_k\}_{k=0}^{\infty}$, $\{\boldsymbol{y}_k\}_{k=0}^{\infty}$, $\{\boldsymbol{x}_k\}_{k=0}^{\infty}$, $\{\boldsymbol{v}_k\}_{k=0}^{\infty}$, $\{\phi_k^*\}_{k=0}^{\infty}$, and $\{\phi_k(\boldsymbol{x})\}_{k=0}^{\infty}$ for the iteration k starting at $k := 0$:

$$\begin{array}{rll}
 & \alpha_{-1} = 0, & \\
\alpha_k \in (0,1] & \text{root of} & L\alpha_k^2 = (1-\alpha_k)\gamma_k + \alpha_k\mu := \gamma_{k+1}, \\
 & \boldsymbol{y}_k = & \dfrac{\alpha_k\gamma_k\boldsymbol{v}_k + \gamma_{k+1}\boldsymbol{x}_k}{\gamma_k + \alpha_k\mu}, \\
\boldsymbol{x}_{k+1} & \text{is such that} & f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{y}_k) - \dfrac{1}{2L}\|\boldsymbol{\nabla f}(\boldsymbol{y}_k)\|_2^2, \\
 & \boldsymbol{v}_{k+1} = & \dfrac{1}{\gamma_{k+1}}[(1-\alpha_k)\gamma_k\boldsymbol{v}_k + \alpha_k\mu\boldsymbol{y}_k - \alpha_k\boldsymbol{\nabla f}(\boldsymbol{y}_k)], \\
 & \phi_{k+1}^* = & (1-\alpha_k)\phi_k^* + \alpha_k f(\boldsymbol{y}_k) - \dfrac{\alpha_k^2}{2\gamma_{k+1}}\|\boldsymbol{\nabla f}(\boldsymbol{y}_k)\|_2^2 \\
 & & +\dfrac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\left(\dfrac{\mu}{2}\|\boldsymbol{y}_k - \boldsymbol{v}_k\|_2^2 + \langle\boldsymbol{\nabla f}(\boldsymbol{y}_k), \boldsymbol{v}_k - \boldsymbol{y}_k\rangle\right), \\
 & \phi_{k+1}(\boldsymbol{x}) = & \phi_{k+1}^* + \dfrac{\gamma_{k+1}}{2}\|\boldsymbol{x} - \boldsymbol{v}_{k+1}\|_2^2.
\end{array}$$

Then, we satisfy all the conditions of Lemma 8.2 for $\lambda_k = \displaystyle\prod_{i=-1}^{k-1}(1-\alpha_i)$.

Proof:

In fact, due to Lemmas 8.3 and 8.4, it just remains to show that $\alpha_k \in (0,1]$ for $(k = 0, 1, \ldots)$ such that $\displaystyle\sum_{k=0}^{\infty}\alpha_k = \infty$. In the special case of $\mu = 0$, we must show that $\alpha_k < 1 \quad (k = 0, 1, \ldots)$. And finally that $f(\boldsymbol{x}_k) \leq \phi_k^*$.

Let us show both using induction hypothesis.

Consider the quadratic equation in α, $q_0(\alpha) := L\alpha^2 + (\gamma_0 - \mu)\alpha - \gamma_0 = 0$. Notice that its discriminant $\Delta := (\gamma_0 - \mu)^2 + 4\gamma_0 L$ is always positive by the hypothesis. Also, $q_0(0) = -\gamma_0 < 0$, due to the hypothesis again. Therefore, this equation always has a root $\alpha_0 > 0$. Since $q_0(1) = L - \mu \geq 0$, $\alpha_0 \leq 1$, and we have $\alpha_0 \in (0,1]$. If $\mu = 0$, and $\alpha_0 = 1$, we will have $L = 0$ which implies $\gamma_0 = 0$ which contradicts our hypothesis. Then $\alpha_0 < 1$ in this case. In addition, $\gamma_1 := (1-\alpha_0)\gamma_0 + \alpha_0\mu > 0$ and $\gamma_0 + \alpha_0\mu > 0$. The same arguments are valid for any k. Therefore, $\alpha_k \in (0,1]$, and $\alpha_k < 1 \quad (k = 0, 1, \ldots,)$ if $\mu = 0$.

Finally, $L\alpha_k^2 = (1-\alpha_k)\gamma_k + \alpha_k\mu \geq (1-\alpha_k)\mu + \alpha_k\mu = \mu$. And we have $\alpha_k \geq \sqrt{\frac{\mu}{L}}$, and therefore, $\displaystyle\sum_{k=0}^{\infty}\alpha_k = \infty$, if $\mu > 0$. For the case $\mu = 0$, let us prove first that $\gamma_k = \gamma_0\lambda_k$. Obviously $\gamma_0 = \gamma_0\lambda_0 (= \gamma_0(1-\alpha_{-1}) = \gamma_0)$, and assuming the induction hypothesis,

$$\gamma_{k+1} = (1-\alpha_k)\gamma_k + \alpha_k\mu = (1-\alpha_k)\gamma_k = (1-\alpha_k)\gamma_0\lambda_k = \gamma_0\lambda_{k+1}.$$

Therefore, $L\alpha_k^2 = \gamma_{k+1} = \gamma_0\lambda_{k+1}$. Since λ_k is a decreasing sequence and $\lambda_k > 0$,

$$\begin{aligned}
\frac{1}{\sqrt{\lambda_{k+1}}} - \frac{1}{\sqrt{\lambda_k}} &= \frac{\sqrt{\lambda_k} - \sqrt{\lambda_{k+1}}}{\sqrt{\lambda_k\lambda_{k+1}}} = \frac{\lambda_k - \lambda_{k+1}}{\sqrt{\lambda_k\lambda_{k+1}}(\sqrt{\lambda_k} + \sqrt{\lambda_{k+1}})} \\
&\geq \frac{\lambda_k - \lambda_{k+1}}{\sqrt{\lambda_k\lambda_{k+1}}(\sqrt{\lambda_k} + \sqrt{\lambda_k})} = \frac{\lambda_k - \lambda_{k+1}}{2\lambda_k\sqrt{\lambda_{k+1}}} = \frac{\lambda_k - (1-\alpha_k)\lambda_k}{2\lambda_k\sqrt{\lambda_{k+1}}} \\
&= \frac{\alpha_k}{2\sqrt{\lambda_{k+1}}} = \frac{1}{2}\sqrt{\frac{\gamma_0}{L}}.
\end{aligned}$$

Thus

$$\frac{1}{\sqrt{\lambda_k}} \geq \frac{1}{\sqrt{\lambda_0}} + \frac{k}{2}\sqrt{\frac{\gamma_0}{L}} = 1 + \frac{k}{2}\sqrt{\frac{\gamma_0}{L}}.$$

Finally,

$$\lambda_k \leq \frac{4L}{(2\sqrt{L} + k\sqrt{\gamma_0})^2} \longrightarrow 0,$$

which is equivalent to $\sum_{k=0}^{\infty} \alpha_k = \infty$ as we saw before.

Now for $k = 0$, $f(\boldsymbol{x}_0) \leq \phi_0^*$. Suppose that the induction hypothesis is valid for any index equal or smaller than k. Due to the previous lemma,

$$\begin{aligned}
\phi_{k+1}^* &= (1-\alpha_k)\phi_k^* + \alpha_k f(\boldsymbol{y}_k) - \frac{\alpha_k^2}{2\gamma_{k+1}}\|\boldsymbol{\nabla f}(\boldsymbol{y}_k)\|_2^2 \\
&\quad + \frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\left(\frac{\mu}{2}\|\boldsymbol{y}_k - \boldsymbol{v}_k\|_2^2 + \langle \boldsymbol{\nabla f}(\boldsymbol{y}_k), \boldsymbol{v}_k - \boldsymbol{y}_k\rangle\right) \\
&\geq (1-\alpha_k) f(\boldsymbol{x}_k) + \alpha_k f(\boldsymbol{y}_k) - \frac{\alpha_k^2}{2\gamma_{k+1}}\|\boldsymbol{\nabla f}(\boldsymbol{y}_k)\|_2^2 \\
&\quad + \frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\left(\frac{\mu}{2}\|\boldsymbol{y}_k - \boldsymbol{v}_k\|_2^2 + \langle \boldsymbol{\nabla f}(\boldsymbol{y}_k), \boldsymbol{v}_k - \boldsymbol{y}_k\rangle\right).
\end{aligned}$$

Now, since $f(\boldsymbol{x})$ is convex, $f(\boldsymbol{x}_k) \geq f(\boldsymbol{y}_k) + \langle \boldsymbol{\nabla f}(\boldsymbol{y}_k), \boldsymbol{x}_k - \boldsymbol{y}_k\rangle$, and multiplying this inequality by $(1-\alpha_k)$ we have:

$$\phi_{k+1}^* \geq f(\boldsymbol{y}_k) - \frac{\alpha_k^2}{2\gamma_{k+1}}\|\boldsymbol{\nabla f}(\boldsymbol{y}_k)\|_2^2 + (1-\alpha_k)\langle \boldsymbol{\nabla f}(\boldsymbol{y}_k), \frac{\alpha_k\gamma_k}{\gamma_{k+1}}(\boldsymbol{v}_k - \boldsymbol{y}_k) + \boldsymbol{x}_k - \boldsymbol{y}_k\rangle + \frac{\alpha_k(1-\alpha_k)\gamma_k\mu}{2\gamma_{k+1}}\|\boldsymbol{y}_k - \boldsymbol{v}_k\|_2^2.$$

Recall that since $\boldsymbol{\nabla f}$ is L-Lipschitz continuous, if we apply Lemma 3.6 to $\boldsymbol{y}_k$ and $\boldsymbol{x}_{k+1} = \boldsymbol{y}_k - \frac{1}{L}\boldsymbol{\nabla f}(\boldsymbol{y}_k)$, we obtain

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

Therefore, if we impose

$$\frac{\alpha_k\gamma_k}{\gamma_{k+1}}(\boldsymbol{v}_k - \boldsymbol{y}_k) + \boldsymbol{x}_k - \boldsymbol{y}_k = \boldsymbol{0}$$

it justifies our choice for $\boldsymbol{y}_k$. And putting

$$\frac{\alpha_k^2}{2\gamma_{k+1}} = \frac{1}{2L}$$

it justifies our choice for α_k. Since $\frac{\alpha_k(1-\alpha_k)\gamma_k\mu}{\gamma_{k+1}} \geq 0$, we finally obtain $\phi_{k+1}^* \geq f(\boldsymbol{x}_{k+1})$ as wished. ∎

The above theorem suggests an algorithm to minimize $f \in \mathcal{S}_{\mu,L}^{1,1}(\mathbb{R}^n)$.

Notice that in the following method, we don't need the estimated sequence anymore.

Generic Scheme for the Nesterov's Optimal Gradient Method	
Step 0:	Choose $\boldsymbol{x}_0 \in \mathbb{R}^n$, let $\gamma_0 > 0$ such that $L \geq \gamma_0 \geq \mu \geq 0$. Set $\boldsymbol{v}_0 := \boldsymbol{x}_0$ and $k := 0$.
Step 1:	Compute $\alpha_k \in (0,1]$ from the equation $L\alpha_k^2 = (1-\alpha_k)\gamma_k + \alpha_k\mu$.
Step 2:	Set $\gamma_{k+1} := (1-\alpha_k)\gamma_k + \alpha_k\mu$, $\boldsymbol{y}_k := \frac{\alpha_k\gamma_k\boldsymbol{v}_k + \gamma_{k+1}\boldsymbol{x}_k}{\gamma_k + \alpha_k\mu}$.
Step 3:	Compute $f(\boldsymbol{y}_k)$ and $\boldsymbol{\nabla f}(\boldsymbol{y}_k)$.
Step 4:	Find $\boldsymbol{x}_{k+1}$ such that $f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{y}_k) - \frac{1}{2L}\|\boldsymbol{\nabla f}(\boldsymbol{y}_k)\|_2^2$ using "line search".
Step 5:	Set $\boldsymbol{v}_{k+1} := \frac{(1-\alpha_k)\gamma_k\boldsymbol{v}_k + \alpha_k\mu\boldsymbol{y}_k - \alpha_k\boldsymbol{\nabla f}(\boldsymbol{y}_k)}{\gamma_{k+1}}$, $k := k+1$ and go to Step 1.

Theorem 8.6 Consider $f \in \mathcal{S}^{1,1}_{\mu,L}(\mathbb{R}^n)$, possible with $\mu = 0$ (which means that $f \in \mathcal{F}^{1,1}_{L}(\mathbb{R}^n)$). The generic scheme of the Nesterov's optimal gradient method generates a sequence $\{\boldsymbol{x}_k\}_{k=0}^{\infty}$ such that

$$\begin{aligned} f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) &\leq \lambda_k \left[f(\boldsymbol{x}_0) + \frac{\gamma_0}{2} \|\boldsymbol{x}^* - \boldsymbol{x}_0\|_2^2 - f(\boldsymbol{x}^*) \right] \\ &\leq \min \left\{ \left(1 - \sqrt{\frac{\mu}{L}}\right)^k, \frac{4L}{(2\sqrt{L} + k\sqrt{\gamma_0})^2} \right\} \left[f(\boldsymbol{x}_0) + \frac{\gamma_0}{2} \|\boldsymbol{x}^* - \boldsymbol{x}_0\|_2^2 - f(\boldsymbol{x}^*) \right], \end{aligned}$$

where $\alpha_{-1} = 0$ and $\lambda_k = \prod_{i=-1}^{k-1} (1 - \alpha_i)$.

In other words, the sequence $\{f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*)\}_{k=0}^{\infty}$ converges R-sublinearly to zero if $\mu = 0$ and R-linearly to zero if $\mu > 0$.

In addition, if $\mu > 0$,

$$\begin{aligned} \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 &\leq \frac{2}{\mu} \lambda_k \left[f(\boldsymbol{x}_0) + \frac{\gamma_0}{2} \|\boldsymbol{x}^* - \boldsymbol{x}_0\|_2^2 - f(\boldsymbol{x}^*) \right] \\ &\leq \frac{2}{\mu} \min \left\{ \left(1 - \sqrt{\frac{\mu}{L}}\right)^k, \frac{4L}{(2\sqrt{L} + k\sqrt{\gamma_0})^2} \right\} \left[f(\boldsymbol{x}_0) + \frac{\gamma_0}{2} \|\boldsymbol{x}^* - \boldsymbol{x}_0\|_2^2 - f(\boldsymbol{x}^*) \right]. \end{aligned}$$

That is, $\{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2\}_{k=0}^{\infty}$ converges R-linearly to zero.

Proof:

The first inequality is obvious from the definitions and Lemma 8.2.

We already know that $\alpha_k \geq \sqrt{\frac{\mu}{L}}$ $(k = 0, 1, \dots)$ (see proof of Theorem 8.5), therefore,

$$\lambda_k = \prod_{i=-1}^{k-1} (1 - \alpha_i) = \prod_{i=0}^{k-1} (1 - \alpha_i) \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)^k,$$

which only has an effect if $\mu > 0$. For the case $\mu = 0$, we already proved in Theorem 8.5.

For $\mu > 0$, using the definition of strong convexity of $f(\boldsymbol{x})$, we obtain the upper bound for $\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2$. ∎

Corollary 8.7 Consider $f \in \mathcal{S}^{1,1}_{\mu,L}(\mathbb{R}^n)$, possible with $\mu = 0$ (which means that $f \in \mathcal{F}^{1,1}_{L}(\mathbb{R}^n)$). If we take $\gamma_0 = L$, the generic scheme of the Nesterov's optimal gradient method generates a sequence $\{\boldsymbol{x}_k\}_{k=0}^{\infty}$ such that

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \leq L \min \left\{ \left(1 - \sqrt{\frac{\mu}{L}}\right)^k, \frac{4}{(k+2)^2} \right\} \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2.$$

In other words, the sequence $\{f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*)\}_{k=0}^{\infty}$ converges R-sublinearly to zero if $\mu = 0$ and R-linearly to zero if $\mu > 0$.

In the particular case of $\mu > 0$, we have the following inequality:

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2 \leq \frac{2L}{\mu} \min \left\{ \left(1 - \sqrt{\frac{\mu}{L}}\right)^k, \frac{4}{(k+2)^2} \right\} \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2.$$

That means that the sequence $\{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2\}_{k=0}^{\infty}$ converges R-linearly to zero.

Proof:

The two inequalities follow from the previous theorem, $f(\boldsymbol{x}_0) - f(\boldsymbol{x}^*) \leq \langle \boldsymbol{\nabla f}(\boldsymbol{x}^*), \boldsymbol{x}_0 - \boldsymbol{x}^* \rangle + \frac{L}{2} \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2$, and the fact that $\boldsymbol{\nabla f}(\boldsymbol{x}^*) = \boldsymbol{0}$. ∎

Now, instead of doing a line search at Step 4 of the generic scheme for the Nesterov's optimal gradient method, let us consider the constant step size iteration $\boldsymbol{x}_{k+1} := \boldsymbol{y}_k - \frac{1}{L}\boldsymbol{\nabla f}(\boldsymbol{y}_k)$ (see proof of Theorem 8.5). From the calculations given at Exercise 1, we arrive to the following simplified scheme. Hereafter, we assume that $L > \mu$ to exclude the trivial case $L = \mu$ with finished in one iteration.

Constant Step Scheme for the Nesterov's Optimal Gradient Method	
Step 0:	Choose $\boldsymbol{x}_0 \in \mathbb{R}^n$, $\alpha_0 \in (0,1)$ such that $\frac{\alpha_0(\alpha_0 L-\mu)}{1-\alpha_0} > 0$, $\mu \leq \frac{\alpha_0(\alpha_0 L-\mu)}{1-\alpha_0} \leq L$, set $\boldsymbol{y}_0 := \boldsymbol{x}_0$ and $k := 0$.
Step 1:	Compute $\boldsymbol{\nabla f}(\boldsymbol{y}_k)$.
Step 2:	Set $\boldsymbol{x}_{k+1} := \boldsymbol{y}_k - \frac{1}{L}\boldsymbol{\nabla f}(\boldsymbol{y}_k)$.
Step 3:	Compute $\alpha_{k+1} \in (0,1)$ from the equation $\alpha_{k+1}^2 = (1-\alpha_{k+1})\alpha_k^2 + \frac{\mu}{L}\alpha_{k+1}$.
Step 4:	Set $\beta_k := \frac{\alpha_k(1-\alpha_k)}{\alpha_k^2+\alpha_{k+1}}$.
Step 5:	Set $\boldsymbol{y}_{k+1} := \boldsymbol{x}_{k+1} + \beta_k(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k)$, $k := k+1$ and go to Step 1.

Observe that the sequences $\{\boldsymbol{x}_k\}_{k=0}^{\infty}$ and $\{\boldsymbol{y}_k\}_{k=0}^{\infty}$ generated by the "Generic Scheme" and the "Constant Step Scheme" are exactly the same[4] if we choose $\boldsymbol{x}_{k+1} := \boldsymbol{y}_k - \frac{1}{L}\boldsymbol{\nabla f}(\boldsymbol{y}_k)$ in the former method. Therefore, the result of Theorem 8.6 is still valid for $\gamma_0 := \alpha_0(\alpha_0 L - \mu)/(1-\alpha_0)$.

Also, if we further impose $\gamma_0 = \alpha_0(\alpha_0 L - \mu)/(1-\alpha_0) = L$, we will have the rate of convergence of Theorem 8.7.

Theorem 8.8 Consider $f \in \mathcal{S}_{\mu,L}^{1,1}(\mathbb{R}^n)$, possible with $\mu = 0$ (which means that $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$). The constant step scheme of the Nesterov's optimal gradient method generates a sequence $\{\boldsymbol{x}_k\}_{k=0}^{\infty}$ such that

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \leq L \min\left\{\left(1-\sqrt{\frac{\mu}{L}}\right)^k, \frac{4}{(k+2)^2}\right\}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2,$$

and

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2 \leq \frac{2L}{\mu}\min\left\{\left(1-\sqrt{\frac{\mu}{L}}\right)^k, \frac{4}{(k+2)^2}\right\}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2.$$

This means that the method is "optimal" for the class of functions $\mathcal{F}_L^{1,1}(\mathbb{R}^n)$, and $\mathcal{S}_{\mu,L}^{1,1}(\mathbb{R}^n)$.

Proof: Since the inequalities above are already shown in the previous Corollary 8.7, it remains to show the "optimality" of the methods for each class of functions.

For the case $\mu = 0$, the "optimality" of the method is obvious from Theorem 6.1.

Let us analyze the case when $\mu > 0$. From Theorem 6.2, we know that we can find a function $f \in \mathcal{S}_{\mu,L}^{\infty,1}(\ell^2)$ such that

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \geq \frac{\mu}{2}\left(\frac{\sqrt{L/\mu}-1}{\sqrt{L/\mu}+1}\right)^{2k}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2 \geq \frac{\mu}{2}\exp\left(-\frac{4k}{\sqrt{L/\mu}-1}\right)\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2,$$

where the second inequality follows from $\ln(\frac{a-1}{a+1}) = -\ln(\frac{a+1}{a-1}) \geq 1 - \frac{a+1}{a-1} = -\frac{2}{a-1}$, for $a \in (1,+\infty)$. Therefore, the worst case bound to find $\boldsymbol{x}_k$ such that $f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) < \varepsilon$ can not be better than

$$k > \frac{\sqrt{L/\mu}-1}{4}\left(\ln\frac{1}{\varepsilon} + \ln\frac{\mu}{2} + 2\ln\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2\right).$$

On the other hand, from the inequality above

[4] strictly speaking, there is a one index difference between $\boldsymbol{y}_k$'s on these two methods due to the order $\boldsymbol{y}_k$ is defined in the loop.

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \leq L\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2 \left(1 - \sqrt{\frac{\mu}{L}}\right)^k \leq L\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2 \exp\left(-\frac{k}{\sqrt{L/\mu}}\right),$$

where the second inequality follows from $\ln(1-a) \leq -a$ for $a < 1$. Therefore, we can guarantee $f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) < \varepsilon$ for $k > \sqrt{L/\mu}\left(\ln \frac{1}{\varepsilon} + \ln L + 2 \ln \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2\right)$.

Now, let us analize the sequences $\{\boldsymbol{x}_k\}_{k=0}^{\infty}$ generated by the method. Again from Theorem 6.2, we can find a function $f \in \mathcal{S}_{\mu,L}^{\infty,1}(\ell^2)$ such that

$$\|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2 \geq \left(\frac{\sqrt{L/\mu} - 1}{\sqrt{L/\mu} + 1}\right)^{2k} \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2 \geq \exp\left(-\frac{4k}{\sqrt{L/\mu} - 1}\right) \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2.$$

Therefore, the worst case bound to find $\boldsymbol{x}_k$ such that $\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2 < \varepsilon$ can not be better than

$$k > \frac{\sqrt{L/\mu} - 1}{4}\left(\ln \frac{1}{\varepsilon} + 2 \ln \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2\right).$$

On the other hand, from the inequality above

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2 \leq \frac{2L}{\mu}\left(1 - \sqrt{\frac{\mu}{L}}\right)^k \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2 \leq \frac{2L}{\mu} \exp\left(-\frac{k}{\sqrt{L/\mu}}\right) \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2.$$

Therefore, we can guarantee $\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2 < \varepsilon$ for $k > \sqrt{L/\mu}\left(\ln \frac{1}{\varepsilon} + \ln 2L - \ln \mu + 2 \ln \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2\right)$.

This shows that the constant step scheme for the Nesterov's gradient method is an optimal method in terms of complexity for the dominant term $\ln(\varepsilon^{-1})$. ∎

Remark 8.9 Many times, you will find in articles that a method has "optimal rate of convergence". In our case, if we apply the constant step scheme for the Nesterov's optimal gradient method to $\min_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x})$, the number of iterations of this method to obtain $f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) < \varepsilon$ is $k = k(L, \boldsymbol{x}_0, \boldsymbol{x}^*, \varepsilon) = \mathcal{O}\left(\sqrt{\frac{L\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2}{\varepsilon}}\right)$ and $k = k(L, \mu, \boldsymbol{x}_0, \boldsymbol{x}^*, \varepsilon) = \mathcal{O}\left(\sqrt{\frac{L}{\mu}} \ln \frac{L\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2}{\varepsilon}\right)$ for $f(\boldsymbol{x}) \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$ and $\mathcal{S}_{L,\mu}^{1,1}(\mathbb{R}^n)$, respectively.

It is extremely important to note that this value is the maximum number of iterations in the worse case scenario.

To obtain the total complexity of the method, you need to multiply the above number by the number of floating-point operations per iteration. This value also vary according to the method.

8.1 Discussion on Particular Cases

8.1.1 Nesterov's Optimal Gradient Method for Smooth (Differentiable) Strongly Convex Functions

In this case, we have $\mu > 0$ and choosing $\gamma_0 := \alpha_0(\alpha_0 L - \mu)/(1 - \alpha_0) = \mu$, we can have further simplifications:

$$\alpha_k = \sqrt{\frac{\mu}{L}}, \quad \beta_k = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}.$$

Nesterov's Optimal Gradient Method for Smooth Strongly Convex Function	
Step 0:	Choose $\boldsymbol{x}_0 \in \mathbb{R}^n$, set $\boldsymbol{y}_0 := \boldsymbol{x}_0$ and $k := 0$.
Step 1:	Compute $\boldsymbol{\nabla f}(\boldsymbol{y}_k)$.
Step 2:	Set $\boldsymbol{x}_{k+1} := \boldsymbol{y}_k - \frac{1}{L}\boldsymbol{\nabla f}(\boldsymbol{y}_k)$.
Step 3:	Set $\boldsymbol{y}_{k+1} := \boldsymbol{x}_{k+1} + \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k)$, $k := k + 1$ and go to Step 1.