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 $f \in \mathcal{S}_{\mu,L}^{1,1}(\mathbb{R}^n)$, possible with $\mu = 0$ (which means that $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$). The 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 $\alpha_k \geq \sqrt{\frac{\mu}{L}}$ $(k = 0, 1, \ldots)$ (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 $\mu = 0$, 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

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

and we have the result.

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 $f \in \mathcal{S}_{\mu,L}^{1,1}(\mathbb{R}^n)$, possible with $\mu = 0$ (which means that $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$). 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.

[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.