Theorem 9.6 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 general scheme of the optimal gradient method generates a sequence $\{\boldsymbol{x}_k\}_{k=0}^{\infty}$ such that

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \leq \lambda_k \left[f(\boldsymbol{x}_0) + \frac{\gamma_0}{2}\|\boldsymbol{x}^* - \boldsymbol{x}_0\|_2^2 - f(\boldsymbol{x}^*) \right],$$

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

$$\lambda_k \leq \min\left\{ \left(1 - \sqrt{\frac{\mu}{L}}\right)^k, \frac{4L}{(2\sqrt{L} + k\sqrt{\gamma_0})^2} \right\}.$$

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

Proof:
The first part is obvious from the definition and Lemma 9.2.
We already know that $\alpha_k \geq \sqrt{\frac{\mu}{L}}$ $(k = 0, 1, \ldots)$, 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 a meaning if $\mu > 0$. For the case $\mu = 0$, let us prove first that $\gamma_k = \gamma_0\lambda_k$. Obviously $\gamma_0 = \gamma_0\lambda_0$, 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

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

Thus

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

and we have the result. ∎

Theorem 9.7 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 general scheme of the "optimal" gradient method generates a sequence $\{\boldsymbol{x}_k\}_{k=0}^{\infty}$ such that

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

This means that it is "optimal" for the class of functions from $\mathcal{S}^{1,1}_{\mu,L}(\mathbb{R}^n)$ with $\mu > 0$, or $\mathcal{F}^{1,1}_L(\mathbb{R}^n)$.

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

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

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

Proof:

The first inequality follows from the previous theorem, $f(\boldsymbol{x}_0) - f(\boldsymbol{x}^*) \leq \langle f'(\boldsymbol{x}^*), \boldsymbol{x}_0 - \boldsymbol{x}^* \rangle + \frac{L}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2$, and the fact that $f'(\boldsymbol{x}^*) = \mathbf{0}$.

For the case $\mu = 0$, the conclusion is obvious from Theorem 7.1.

Let us analyze first the case when $\mu > 0$. From Theorem 7.2, we know that we can find functions 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 above result

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

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

Finally, for $\mu > 0$, since $\frac{\mu}{2}\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2 + f(\boldsymbol{x}^*) \leq f(\boldsymbol{x}_k)$ from the definition, we have the second inequality. ∎

Now, instead of doing line search at Step 4 of the general scheme for the optimal gradient method, let us consider the constant step size iteration $\boldsymbol{x}_{k+1} := \boldsymbol{y}_k - \frac{1}{L}f'(\boldsymbol{y}_k)$ (See proof of Theorem 9.5). From the calculations given at Exercise 2, 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 I for the 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 $f'(\boldsymbol{y}_k)$.
Step 2:	Set $\boldsymbol{x}_{k+1} := \boldsymbol{y}_k - \frac{1}{L}f'(\boldsymbol{y}_k)$.
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 generated by the General Scheme and the Constant Step Scheme I for the Optimal Gradient Methods are different. However, the rate of convergence of the above method is similar to Theorem 9.6 for $\gamma_0 := \alpha_0(\alpha_0 L - \mu)/(1-\alpha_0)$. 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 9.7:

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

Finally, if $\mu > 0$ and we choose $\gamma_0 := \alpha_0(\alpha_0 L - \mu)/(1-\alpha_0) = \mu$, we have a further simplification.

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

and we end up with