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 $\alpha_k$. 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 optimal gradient method, we don't need the estimated sequence anymore.

| **General Scheme for the Optimal Gradient Method** | |
|---|---|
| **Step 0:** | Choose $\boldsymbol{x}_0 \in \mathbb{R}^n$, let $\gamma_0 > 0$ such that $L \geq \gamma_0 \geq \mu \geq 0$. <br> 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 9.6** 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 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 $R$-sublinearly to zero if $\mu = 0$ and $R$-linearly to zero if $\mu > 0$. In addition, if $\mu > 0$,

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

*Proof:*

The first part is obvious from the definition and Lemma 9.2.

We already know that $\alpha_k \geq \sqrt{\frac{\mu}{L}}$ $(k = 0, 1, \ldots)$ (see proof of Theorem 9.5), 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 $\gamma_0 = \gamma_0\lambda_0$, 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 $\lambda_k$ 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.

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

**Theorem 9.7** 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 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 $\mathcal{S}_{\mu,L}^{1,1}(\mathbb{R}^n)$ with $\mu > 0$, or $\mathcal{F}_L^{1,1}(\mathbb{R}^n)$.

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

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

That means that the sequence $\{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2\}_{k=0}^{\infty}$ 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 $\boldsymbol{\nabla f}(\boldsymbol{x}^*) = \boldsymbol{0}$.

For the case $\mu = 0$, the "optimality" of the method is obvious from Theorem 7.1.

Let us analyze the case when $\mu > 0$. From Theorem 7.2, we know that we can find functions $f \in \mathcal{S}_{\mu,L}^{\infty,1}(\mathbb{R}^\infty)$ 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 $\ln(1-a) \leq -a$ for $a < 1$. Therefore, we can guarantee $f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) < \varepsilon$ for $k > \sqrt{L/\mu}\left(\ln\frac{1}{\varepsilon} + \ln L + 2\ln\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2\right)$. This shows that the "General