Thus,

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

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

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

$$\begin{aligned} \phi_{k+1}(\boldsymbol{y}_k) &= \phi^*_{k+1} + \tfrac{\gamma_{k+1}}{2}\|\boldsymbol{y}_k - \boldsymbol{v}_{k+1}\|^2 \\ &= (1-\alpha_k)\phi_k(\boldsymbol{y}_k) + \alpha_k f(\boldsymbol{y}_k) \\ &= (1-\alpha_k)\left(\phi^*_k + \tfrac{\gamma_k}{2}\|\boldsymbol{y}_k - \boldsymbol{v}_k\|^2\right) + \alpha_k f(\boldsymbol{y}_k). \end{aligned} \tag{2.4}$$

Now,

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

Therefore,

$$\begin{aligned} \tfrac{\gamma_{k+1}}{2}\|\boldsymbol{v}_{k+1} - \boldsymbol{y}_k\|^2 &= \tfrac{1}{2\gamma_{k+1}}\left[(1-\alpha_k)^2\gamma_k^2\|\boldsymbol{v}_k - \boldsymbol{y}_k\|^2 + \alpha_k^2\|f'(\boldsymbol{y}_k)\|^2\right. \\ &\quad \left. -2\alpha_k(1-\alpha_k)\gamma_k\langle f'(\boldsymbol{y}_k), \boldsymbol{v}_k - \boldsymbol{y}_k\rangle\right]. \end{aligned} \tag{2.5}$$

Substituting (2.5) into (2.4), we obtain the expression for ϕ^*_{k+1}. ∎

Theorem 2.6.5 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 a given $\boldsymbol{x}_0, \boldsymbol{v}_0 \in \mathbb{R}^n$ and $L \geq \gamma_0 \geq \mu \geq 0$, let us choose $\phi^*_0 = f(\boldsymbol{x}_0)$. Define the sequences $\{\alpha_k\}_{k=0}^\infty$, $\{\gamma_k\}_{k=0}^\infty$, $\{\boldsymbol{y}_k\}_{k=0}^\infty$, $\{\boldsymbol{x}_k\}_{k=0}^\infty$, $\{\boldsymbol{v}_k\}_{k=0}^\infty$, $\{\phi^*_k\}_{k=0}^\infty$, and $\{\phi_k\}_{k=0}^\infty$ as follows:

$$\begin{aligned} \alpha_k \in (0,1) \quad \text{root of} \quad & L\alpha_k^2 = (1-\alpha_k)\gamma_k + \alpha_k\mu \equiv \gamma_{k+1}, \\ \boldsymbol{y}_k = \quad & \frac{\alpha_k\gamma_k\boldsymbol{v}_k + \gamma_{k+1}\boldsymbol{x}_k}{\gamma_k + \alpha_k\mu}, \\ \boldsymbol{x}_k \quad \text{is such that} \quad & f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{y}_k) - \frac{1}{2L}\|f'(\boldsymbol{y}_k)\|^2, \\ \boldsymbol{v}_{k+1} = \quad & \frac{1}{\gamma_{k+1}}[(1-\alpha_k)\gamma_k\boldsymbol{v}_k + \alpha_k\mu\boldsymbol{y}_k - \alpha_k f'(\boldsymbol{y}_k)], \\ \phi^*_{k+1} = \quad & (1-\alpha_k)\phi^*_k + \alpha_k f(\boldsymbol{y}_k) - \frac{\alpha_k^2}{2\gamma_{k+1}}\|f'(\boldsymbol{y}_k)\|^2 \\ & + \frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\left(\frac{\mu}{2}\|\boldsymbol{y}_k - \boldsymbol{v}_k\|^2 + \langle f'(\boldsymbol{y}_k), \boldsymbol{v}_k - \boldsymbol{y}_k\rangle\right), \\ \phi_{k+1}(\boldsymbol{x}) = \quad & \phi^*_{k+1} + \frac{\gamma_{k+1}}{2}\|\boldsymbol{x} - \boldsymbol{v}_{k+1}\|^2. \end{aligned}$$

Then, we satisfy all the conditions of Lemma 2.6.2.

Proof: In fact, it just remains to show that $f(\boldsymbol{x}_k) \leq \phi^*_k$ and $\sum_{k=1}^\infty \alpha_k = \infty$.

For $k = 0$, $f(\boldsymbol{x}_0) \leq \phi^*_0$. Suppose that induction hypothesis is valid for k, and due to the previous lemma,

$$\phi^*_{k+1} = (1-\alpha_k)\phi^*_k + \alpha_k f(\boldsymbol{y}_k) - \frac{\alpha_k^2}{2\gamma_{k+1}}\|f'(\boldsymbol{y}_k)\|^2$$

$$
\begin{aligned}
& +\frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\left(\langle f'(\boldsymbol{y}_k), \boldsymbol{v}_k-\boldsymbol{y}_k\rangle+\frac{\mu}{2}\|\boldsymbol{y}_k-\boldsymbol{v}_k\|^2\right) \\
\geq \; & (1-\alpha_k)f(\boldsymbol{x}_k)+\alpha_k f(\boldsymbol{y}_k)-\frac{\alpha_k^2}{2\gamma_{k+1}}\|f'(\boldsymbol{y}_k)\|^2 \\
& +\frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\left(\langle f'(\boldsymbol{y}_k), \boldsymbol{v}_k-\boldsymbol{y}_k\rangle+\frac{\mu}{2}\|\boldsymbol{y}_k-\boldsymbol{v}_k\|^2\right).
\end{aligned}
$$

Now, since $f(\boldsymbol{x})$ is convex, $f(\boldsymbol{x}_k) \geq f(\boldsymbol{y}_k)+\langle f'(\boldsymbol{y}_k), \boldsymbol{x}_k-\boldsymbol{y}_k\rangle$, and we have:

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

Recall that since f' is L-Lipschitz continuous, if we apply Theorem 2.1.8 to $\boldsymbol{y}_k$ and $\boldsymbol{x}_{k+1} = \boldsymbol{y}_k - \frac{1}{L}f'(\boldsymbol{y}_k)$, we obtain

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

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 $\mu \geq 0$, we finally obtain $\phi_{k+1}^* \geq f(\boldsymbol{x}_{k+1})$ as wished.

Now, $\gamma_{k+1} = L\alpha_k^2 = (1-\alpha_k)\gamma_k + \alpha_k\mu$, and since $L \geq \gamma_0 \geq \mu$, we have $\alpha_k \in [\sqrt{\frac{\mu}{L}}, 1)$ and $L \geq \gamma_k \geq \mu$. Therefore, $\sum_{k=1}^{\infty}\alpha_k = \infty$. ∎

We arrive finally at the following optimal gradient method

General Scheme for the Optimal Gradient Method	
Step 0:	Choose $\boldsymbol{x}_0 \in \mathbb{R}^n$, $L \geq \gamma_0 \geq \mu \geq 0$, set $\boldsymbol{v}_0 := \boldsymbol{x}_0$, $k := 0$
Step 1:	Compute $\alpha_k \in [\sqrt{\frac{\mu}{L}}, 1)$ 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 $f'(\boldsymbol{y}_k)$
Step 4:	Find $\boldsymbol{x}_{k+1}$ such that $f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{y}_k) - \frac{1}{2L}\|f'(\boldsymbol{y}_k)\|^2$ using "line search"
Step 5:	Set $\boldsymbol{v}_{k+1} := \frac{(1-\alpha_k)\gamma_k\boldsymbol{v}_k+\alpha_k\mu\boldsymbol{y}_k-\alpha_k f'(\boldsymbol{y}_k)}{\gamma_{k+1}}$, $k := k+1$ and go to Step 1

Theorem 2.6.6 Consider $f \in \mathcal{S}_{\mu,L}^{1,1}(\mathbb{R}^n)$, 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 $\{\boldsymbol{x}_k\}_{k=1}^{\infty}$ such that

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

where $\lambda_0 = 1$ and $\lambda_k = \Pi_{i=0}^{k-1}(1-\alpha_i)$. Moreover,

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

Proof: The first part is obvious from the definition and Lemma 2.6.2.
We already now that $\alpha_k \geq \sqrt{\frac{\mu}{L}}$, therefore,

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

Let us prove first that $\gamma_k \geq \gamma_0 \lambda_k$. Obviously $\gamma_0 = \gamma_0 \lambda_0$, and assuming the induction hypothesis,

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

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

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

Thus

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

and we have the result. ∎

Theorem 2.6.7 Consider $f \in \mathcal{S}_{\mu,L}^{1,1}(\mathbb{R}^n)$, 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 $\{\boldsymbol{x}_k\}_{k=1}^{\infty}$ such that

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

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 $\mathcal{F}_L^{1,1}(\mathbb{R}^n)$.

Proof: The inequality follows from the previous theorem and $f(\boldsymbol{x}_0) - f(\boldsymbol{x}^*) \leq \langle f'(\boldsymbol{x}^*), \boldsymbol{x}_0 - \boldsymbol{x}^*\rangle + \frac{L}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2$.
Let us analyze first the case when $\mu > 0$. From Theorem 2.4.1, we know that we can find functions such that

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

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

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

On the other hand, from the above result

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

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

For the case $\mu = 0$, the conclusion is obvious from Theorem 2.2.1. ∎

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 $\boldsymbol{x}_{k+1} = \boldsymbol{y}_k - \frac{1}{L}f'(\boldsymbol{y}_k)$. From the calculation given at Exercise 9, we arrive to the following simplified scheme:

Constant Step Scheme for the Optimal Gradient Method	
Step 0:	Choose $\boldsymbol{x}_0 \in \mathbb{R}^n$, $\alpha_0 \in (0,1)$ such that $\mu \leq \frac{\alpha_0(\alpha_0 L - \mu)}{1-\alpha_0} \leq L$, set $\boldsymbol{y}_0 := \boldsymbol{x}_0$, $k := 0$
Step 1:	Compute $f(\boldsymbol{y}_k)$ and $f'(\boldsymbol{y}_k)$
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 $\alpha_{k+1}^2 = (1-\alpha_{k+1})\alpha_k^2 + \mu\alpha_{k+1}/L$
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

The rate of convergence of the above method is the same as Theorem 2.6.6 for $\gamma_0 = \alpha_0(\alpha_0 L - \mu)/(1-\alpha_0)$, and of the Theorem 2.6.7.

2.7 Extension for "simple" convex sets

We are interested now to solve the following problem:

$$\begin{cases} \min & f(\boldsymbol{x}) \\ \boldsymbol{x} \in Q & \end{cases} \tag{2.6}$$

where Q is a closed convex set simple enough to have an easy projection onto it, *e.g.*, positive orthant, n dimensional box, simplex, Euclidean ball, *etc.*

Lemma 2.7.1 Let $f \in \mathcal{F}^1(\mathbb{R}^n)$ and Q be a closed convex set. The point $\boldsymbol{x}^*$ is a solution of (2.6) if and only if

$$\langle f'(\boldsymbol{x}^*), \boldsymbol{x} - \boldsymbol{x}^*\rangle \geq 0, \quad \forall \boldsymbol{x} \in Q.$$

Proof: Indeed, if the inequality is true,

$$f(\boldsymbol{x}) \geq f(\boldsymbol{x}^*) + \langle f'(\boldsymbol{x}^*), \boldsymbol{x} - \boldsymbol{x}^*\rangle \geq f(\boldsymbol{x}^*) \quad \forall \boldsymbol{x} \in Q.$$

Let $\boldsymbol{x}^*$ be an optimal solution of the minimization problem (2.6). Assume by contradiction that there is a $\boldsymbol{x} \in Q$ such that $\langle f'(\boldsymbol{x}^*), \boldsymbol{x} - \boldsymbol{x}^*\rangle < 0$. Consider the function $\phi(\alpha) = f(\boldsymbol{x}^* + \alpha(\boldsymbol{x} - \boldsymbol{x}^*))$ for $\alpha \in [0,1]$. Then, $\phi(0) = f(\boldsymbol{x}^*)$ and $\phi'(0) = \langle f'(\boldsymbol{x}^*), \boldsymbol{x} - \boldsymbol{x}^*\rangle < 0$. Therefore, for $\alpha > 0$ small enough, we have

$$f(\boldsymbol{x}^* + \alpha(\boldsymbol{x} - \boldsymbol{x}^*)) = \phi(\alpha) < \phi(0) = f(\boldsymbol{x}^*)$$

which is a contradiction. ∎

Definition 2.7.2 Let $f \in \mathcal{C}^1(\mathbb{R}^n)$, Q a closed convex set, $\bar{\boldsymbol{x}} \in \mathbb{R}^n$, and $\gamma > 0$. Denote by

$$\begin{aligned} \boldsymbol{x}_Q(\bar{\boldsymbol{x}};\gamma) &= \arg\min_{\boldsymbol{x}\in Q} \left[f(\bar{\boldsymbol{x}}) + \langle f'(\bar{\boldsymbol{x}}), \boldsymbol{x} - \bar{\boldsymbol{x}} \rangle + \frac{\gamma}{2}\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|^2 \right], \\ \boldsymbol{g}_Q(\bar{\boldsymbol{x}};\gamma) &= \gamma(\bar{\boldsymbol{x}} - \boldsymbol{x}_Q(\bar{\boldsymbol{x}};\gamma)). \end{aligned}$$

We call $\boldsymbol{g}_Q(\bar{\boldsymbol{x}};\gamma)$ the *gradient mapping of f on Q*.

Theorem 2.7.3 Let $f \in \mathcal{S}^{1,1}_{\mu,L}(\mathbb{R}^n)$, $\gamma \geq L$, and $\bar{\boldsymbol{x}} \in \mathbb{R}^n$. Then

$$f(\boldsymbol{x}) \geq f(\boldsymbol{x}_Q(\bar{\boldsymbol{x}};\gamma)) + \langle \boldsymbol{g}_Q(\bar{\boldsymbol{x}};\gamma), \boldsymbol{x} - \bar{\boldsymbol{x}} \rangle + \frac{1}{2\gamma}\|\boldsymbol{g}_Q(\bar{\boldsymbol{x}};\gamma)\|^2 + \frac{\mu}{2}\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|^2, \quad \forall \boldsymbol{x} \in Q.$$

Proof: Denote $\boldsymbol{x}_Q = \boldsymbol{x}_Q(\bar{\boldsymbol{x}};\gamma)$ and $\boldsymbol{g}_Q = \boldsymbol{g}_Q(\bar{\boldsymbol{x}};\gamma)$. Let $\phi(\boldsymbol{x}) = f(\bar{\boldsymbol{x}}) + \langle f'(\bar{\boldsymbol{x}}), \boldsymbol{x} - \bar{\boldsymbol{x}} \rangle + \frac{\gamma}{2}\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|^2$.

Then $\phi'(\boldsymbol{x}) = f'(\bar{\boldsymbol{x}}) + \gamma(\boldsymbol{x} - \bar{\boldsymbol{x}})$, and for $\forall \boldsymbol{x} \in Q$, we have

$$\langle f'(\bar{\boldsymbol{x}}) - \boldsymbol{g}_Q, \boldsymbol{x} - \boldsymbol{x}_Q \rangle = \langle \phi'(\boldsymbol{x}_Q), \boldsymbol{x} - \boldsymbol{x}_Q \rangle \geq 0,$$

due to Lemma 2.7.1.

Hence,

$$\begin{aligned} f(\boldsymbol{x}) - \frac{\mu}{2}\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|^2 &\geq f(\bar{\boldsymbol{x}}) + \langle f'(\bar{\boldsymbol{x}}), \boldsymbol{x} - \bar{\boldsymbol{x}} \rangle \\ &= f(\bar{\boldsymbol{x}}) + \langle f'(\bar{\boldsymbol{x}}), \boldsymbol{x} - \boldsymbol{x}_Q \rangle + \langle f'(\bar{\boldsymbol{x}}), \boldsymbol{x}_Q - \bar{\boldsymbol{x}} \rangle \\ &\geq f(\bar{\boldsymbol{x}}) + \langle \boldsymbol{g}_Q, \boldsymbol{x} - \boldsymbol{x}_Q \rangle + \langle f'(\bar{\boldsymbol{x}}), \boldsymbol{x}_Q - \bar{\boldsymbol{x}} \rangle \\ &= \phi(\boldsymbol{x}_Q) - \frac{\gamma}{2}\|\boldsymbol{x}_Q - \bar{\boldsymbol{x}}\|^2 + \langle \boldsymbol{g}_Q, \boldsymbol{x} - \boldsymbol{x}_Q \rangle \\ &= \phi(\boldsymbol{x}_Q) - \frac{1}{2\gamma}\|\boldsymbol{g}_Q\|^2 + \langle \boldsymbol{g}_Q, \boldsymbol{x} - \boldsymbol{x}_Q \rangle \\ &= \phi(\boldsymbol{x}_Q) - \frac{1}{2\gamma}\|\boldsymbol{g}_Q\|^2 + \langle \boldsymbol{g}_Q, \bar{\boldsymbol{x}} - \boldsymbol{x}_Q \rangle + \langle \boldsymbol{g}_Q, \boldsymbol{x} - \bar{\boldsymbol{x}} \rangle \\ &= \phi(\boldsymbol{x}_Q) + \frac{1}{2\gamma}\|\boldsymbol{g}_Q\|^2 + \langle \boldsymbol{g}_Q, \boldsymbol{x} - \bar{\boldsymbol{x}} \rangle. \end{aligned}$$

Since $\gamma \geq L$, $\phi(\boldsymbol{x}_Q) \geq f(\boldsymbol{x}_Q)$, and we have the result. ∎

We are ready to define our estimated sequence. Assume that $f \in \mathcal{S}^{1,1}_{\mu,L}(\mathbb{R}^n)$ possible with $\mu = 0$ (which means that $f \in \mathcal{F}^{1,1}_L(\mathbb{R}^n)$), $\boldsymbol{x}_0 \in Q$, and $\gamma_0 > 0$. Define

$$\begin{aligned} \phi_0(\boldsymbol{x}) &= f(\boldsymbol{x}_0) + \frac{\gamma_0}{2}\|\boldsymbol{x} - \boldsymbol{x}_0\|^2, \\ \phi_{k+1}(\boldsymbol{x}) &= (1-\alpha_k)\phi_k(\boldsymbol{x}) + \alpha_k \left[f(\boldsymbol{x}_Q(\boldsymbol{y}_k;L)) + \frac{1}{2L}\|\boldsymbol{g}_Q(\boldsymbol{y}_k;L)\|^2 + \langle \boldsymbol{g}_Q(\boldsymbol{y}_k;L), \boldsymbol{x} - \boldsymbol{y}_k \rangle \right. \\ &\quad \left. + \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{y}_k\|^2 \right], \end{aligned}$$

for the sequences $\{\alpha_k\}_{k=0}^{\infty}$ and $\{\boldsymbol{y}_k\}_{k=0}^{\infty}$ which will be defined later.

Similarly, we can prove that $\{\phi_k(\boldsymbol{x})\}_{k=0}^{\infty}$ can be written in the form

$$\phi_k(\boldsymbol{x}) = \phi_k^* + \frac{\gamma_k}{2}\|\boldsymbol{x} - \boldsymbol{v}_k\|^2$$

for $\phi_0^* = f(\boldsymbol{x}_0)$, $\boldsymbol{v}_0 = \boldsymbol{x}_0$:

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

Now, $\phi_0^* \geq f(\boldsymbol{x}_0)$. Assuming that $\phi_k^* \geq f(\boldsymbol{x}_k)$,

$$\begin{aligned}
\phi_{k+1}^* &\geq (1-\alpha_k)f(\boldsymbol{x}_k) + \alpha_k f(\boldsymbol{x}_Q(\boldsymbol{y}_k;L)) + \left(\frac{\alpha_k}{2L} - \frac{\alpha_k^2}{2\gamma_{k+1}}\right)\|\boldsymbol{g}_Q(\boldsymbol{y}_k;L)\|^2 \\
&\quad + \frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\langle\boldsymbol{g}_Q(\boldsymbol{y}_k;L), \boldsymbol{v}_k - \boldsymbol{y}_k\rangle \\
&\geq f(\boldsymbol{x}_Q(\boldsymbol{y}_k;L)) + \left(\frac{1}{2L} - \frac{\alpha_k^2}{2\gamma_{k+1}}\right)\|\boldsymbol{g}_Q(\boldsymbol{y}_k;L)\|^2 \\
&\quad + (1-\alpha_k)\langle\boldsymbol{g}_Q(\boldsymbol{y}_k;L), \frac{\alpha_k\gamma_k}{\gamma_{k+1}}(\boldsymbol{v}_k - \boldsymbol{y}_k) + \boldsymbol{x}_k - \boldsymbol{y}_k\rangle,
\end{aligned}$$

where the last inequality follows from Theorem 2.7.3.

Therefore, if we choose

$$\begin{aligned}
\boldsymbol{x}_{k+1} &= \boldsymbol{x}_Q(\boldsymbol{y}_k;L), \\
L\alpha_k^2 &= (1-\alpha_k)\gamma_k + \alpha_k\mu \equiv \gamma_{k+1}, \\
\boldsymbol{y}_k &= \frac{1}{\gamma_k + \alpha_k\mu}(\alpha_k\gamma_k\boldsymbol{v}_k + \gamma_{k+1}\boldsymbol{x}_k),
\end{aligned}$$

we obtain $\phi_{k+1}^* \geq f(\boldsymbol{x}_{k+1})$ as desired.

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

The rate of converge of this algorithm is exactly the same as the previous ones.

2.8 Further reading

1. Obviously, the first reading should be the continuation of [NESTEROV2004], where Nesterov extends the method for constrained minimization, min-max type problems, and non-differentiable problems.

2. A more general approach and variations can be found in [DASPREMONT2008, LLM2006, NESTEROV2005, NESTEROV2005-2, NESTEROV2007, NESTERVO2009, TSENG2010], *etc.*

2.9 Exercises

1. Prove Theorem 2.1.2.

2. Prove Lemma 2.1.3.

3. Prove Theorem 2.1.5.

4. Prove Corollary 2.3.3.

5. Prove Theorem 2.3.4.

6. Prove Theorem 2.3.6.

7. Prove Corollary 2.5.2.

8. Complete the prove of Lemma 2.6.3.

9. We want to justify the Constant Step Scheme of the Optimal Gradient Method. This is a particular case of the general optimal gradient method for the following choice:

$$\begin{aligned} \gamma_{k+1} &\equiv L\alpha_k^2 = (1-\alpha_k)\gamma_k + \alpha_k\mu \\ \boldsymbol{y}_k &= \frac{\alpha_k\gamma_k\boldsymbol{v}_k + \gamma_{k+1}\boldsymbol{x}_k}{\gamma_k + \alpha_k\mu} \\ \boldsymbol{x}_{k+1} &= \boldsymbol{y}_k - \frac{1}{L}f'(\boldsymbol{y}_k) \\ \boldsymbol{v}_{k+1} &= \frac{(1-\alpha_k)\gamma_k\boldsymbol{v}_k + \alpha_k\mu\boldsymbol{y}_k - \alpha_k f'(\boldsymbol{y}_k)}{\gamma_{k+1}}. \end{aligned}$$

   (a) Show that $\boldsymbol{v}_{k+1} = \boldsymbol{x}_k + \frac{1}{\alpha_k}(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k)$.

   (b) Show that $\boldsymbol{y}_{k+1} = \boldsymbol{x}_{k+1} + \beta_k(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k)$ for $\beta_k = \frac{\alpha_{k+1}\gamma_{k+1}(1-\alpha_k)}{\alpha_k(\gamma_{k+1}+\alpha_{k+1}\mu)}$.

   (c) Show that $\beta_k = \frac{\alpha_k(1-\alpha_k)}{\alpha_k^2+\alpha_{k+1}}$.

   (d) Explain why $\alpha_{k+1}^2 = (1-\alpha_{k+1})\alpha_k^2 + \frac{\mu}{L}\alpha_{k+1}$.

Bibliography

[DASPREMONT2008] A. d'Aspremont, "Smooth optimization with approximate gradient", *SIAM Journal on Optimization* **19** (2008), pp. 1171–1183.

[LLM2006] G. Lan, Z. Lu, and R. D. C. Monteiro, "Primal-dual first-order methods with $\mathcal{O}(1/\varepsilon)$ iteration-complexity for cone programming", *Mathematical Programming*, to appear.

[NESTEROV2004] Yu. Nesterov, *Introductory Lecture on Convex Optimization: A Basic Course*, (Kluwer Academic Publishers, Boston, 2004).

[NESTEROV2005] Yu. Nesterov, "Smooth minimization of non-smooth functions", *Mathematical Programming* **103** (2005), pp. 127–152.

[NESTEROV2005-2] Yu. Nesterov, "Excessive gap technique in nonsmooth convex minimization", *SIAM Journal on Optimization* **16** (2005), pp. 669–700.

[NESTEROV2007] Yu. Nesterov, "Smoothing technique and its applications in semidefinite optimization", *Mathematical Programming* **110** (2007), pp. 245–259.

[NESTERVO2009] Yu. Nesterov, "Primal-dual subgradient methods for convex problems", *Mathematical Programming* **120** (2009), pp. 221-259.

[NOCEDAL2006] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd edition, (Springer, New York, 2006).

[TSENG2010] P. Tseng, "Approximation accuracy, gradient methods, and error bound for structured convex optimization", *Mathematical Programming* **12** (2010), pp. 263–295.

[YUAN2010] Y.-X. Yuan, "A short note on the Q-linear convergence of the steepest descent method", *Mathematical Programming* **123** (2010), pp. 339–343.