- Note that the previous result for the steepest descent method, Theorem 5.12, was only a local result. Theorems 8.1 and 8.3 guarantee that the steepest descent method converges for any starting point $\boldsymbol{x}_0 \in \mathbb{R}^n$.
- 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 8.1, Corollary 8.2, and 8.3, respectively) with their lower complexity bounds (Theorems 7.1 and 7.2, respectively), we possible have a huge gap.

8.1 Exercises

1. Prove Corollary 8.2.

9 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], he gives a reinterpretation of the algorithm and provides another justification of it which attains the same complexity bound of the original article.

Definition 9.1 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

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

Lemma 9.2 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 9.3 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

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

is an estimate sequence.

[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.

Proof:
Let us prove by induction on 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 $f \in \mathcal{S}^1_\mu(\mathbb{R}^n)$,

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

The remaining part is left for exercise. ∎

Lemma 9.4 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 the previous lemma:

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

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

$$\phi_{k+1}(\boldsymbol{x}) = \phi_{k+1}^* + \frac{\gamma_{k+1}}{2}\|\boldsymbol{x} - \boldsymbol{v}_{k+1}\|_2^2 \tag{12}$$

for

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

Proof:
We will use again the induction hypothesis in k. Note that $\phi_0''(\boldsymbol{x}) = \gamma_0\boldsymbol{I}$. Now, for any $k \geq 0$,

$$\phi_{k+1}''(\boldsymbol{x}) = (1-\alpha_k)\phi_k''(\boldsymbol{x}) + \alpha_k\mu\boldsymbol{I} = ((1-\alpha_k)\gamma_k + \alpha_k\mu)\,\boldsymbol{I} = \gamma_{k+1}\boldsymbol{I}.$$

Therefore, $\phi_{k+1}(\boldsymbol{x})$ is a quadratic function of the form (12). 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

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