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 $\sum_{k=0}^{\infty} \alpha_k = \infty$.

Then the pair of sequences $\left\{ \prod_{i=-1}^{k-1} (1-\alpha_i) \right\}_{k=0}^{\infty}$ 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 $f \in \mathcal{S}^1_\mu(\mathbb{R}^n)$,

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

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 $\phi_0^* \in \mathbb{R}$, $\mu \ge 0$, $\gamma_0 \ge 0$, $\boldsymbol{v}_0 \in \mathbb{R}^n$, $\{\boldsymbol{y}_k\}_{k=0}^\infty$, and $\{\alpha_k\}_{k=0}^\infty$ given arbitrarily sequences such that $\alpha_{-1}=0$, $\alpha_k \in (0,1]$ $(k=0,1,\ldots)$. In the special case of $\mu=0$, we further assume that $\gamma_0>0$ and $\alpha_k<1$ $(k=0,1,\ldots)$. 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 $k \ge 0$,

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

Therefore, $\phi_{k+1}(\boldsymbol{x})$ is a quadratic function of the form (12). Also, $\gamma_{k+1}>0$ since $\mu>0$ and $\alpha_k>0$ $(k=0,1,\ldots)$; or if $\mu=0$, we assumed that $\gamma_0>0$ and $\alpha_k \in (0,1)$ $(k=0,1,\ldots)$.

From the first-order optimality condition

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

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

$$\begin{array}{rcl} \phi_{k+1}(\boldsymbol{y}_k) & = & \phi_{k+1}^* + \frac{\gamma_{k+1}}{2}\|\boldsymbol{y}_k - \boldsymbol{v}_{k+1}\|_2^2 \\ & = & (1-\alpha_k)\phi_k(\boldsymbol{y}_k) + \alpha_k f(\boldsymbol{y}_k) \\ & = & (1-\alpha_k)\left(\phi_k^* + \frac{\gamma_k}{2}\|\boldsymbol{y}_k - \boldsymbol{v}_k\|_2^2\right) + \alpha_k f(\boldsymbol{y}_k). \end{array} \tag{13}$$

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,

$$\begin{array}{rcl} \frac{\gamma_{k+1}}{2}\|\boldsymbol{v}_{k+1} - \boldsymbol{y}_k\|_2^2 & = & \frac{1}{2\gamma_{k+1}}\left[(1-\alpha_k)^2\gamma_k^2\|\boldsymbol{v}_k - \boldsymbol{y}_k\|_2^2 + \alpha_k^2\|f'(\boldsymbol{y}_k)\|_2^2\right. \\ & & \left. -2\alpha_k(1-\alpha_k)\gamma_k\langle f'(\boldsymbol{y}_k), \boldsymbol{v}_k - \boldsymbol{y}_k\rangle\right]. \end{array} \tag{14}$$

Substituting (14) into (13), we obtain the expression for ϕ_{k+1}^*. ∎

Theorem 9.5 Let $L \geq \mu \geq 0$. Consider $f \in \mathcal{S}_{\mu,L}^{1,1}(\mathbb{R}^n)$, possible with $\mu = 0$ (which means that $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$). For given $\boldsymbol{x}_0, \boldsymbol{v}_0 \in \mathbb{R}^n$, let us choose $\phi_0^* = f(\boldsymbol{x}_0)$. Consider also $\gamma_0 > 0$ such that $L \geq \gamma_0 \geq \mu \geq 0$. Define the sequences $\{\alpha_k\}_{k=-1}^\infty$, $\{\gamma_k\}_{k=0}^\infty$, $\{\boldsymbol{y}_k\}_{k=0}^\infty$, $\{\boldsymbol{x}_k\}_{k=0}^\infty$, $\{\boldsymbol{v}_k\}_{k=0}^\infty$, $\{\phi_k^*\}_{k=0}^\infty$, and $\{\phi_k(\boldsymbol{x})\}_{k=0}^\infty$ as follows:

$$\begin{array}{rcl} & \alpha_{-1} = 0, & \\ \alpha_k \in (0,1] & \text{root of} & L\alpha_k^2 = (1-\alpha_k)\gamma_k + \alpha_k\mu := \gamma_{k+1}, \\ & \boldsymbol{y}_k = & \dfrac{\alpha_k\gamma_k\boldsymbol{v}_k + \gamma_{k+1}\boldsymbol{x}_k}{\gamma_k + \alpha_k\mu}, \\ \boldsymbol{x}_k & \text{is such that} & f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{y}_k) - \dfrac{1}{2L}\|f'(\boldsymbol{y}_k)\|_2^2, \\ & \boldsymbol{v}_{k+1} = & \dfrac{1}{\gamma_{k+1}}[(1-\alpha_k)\gamma_k\boldsymbol{v}_k + \alpha_k\mu\boldsymbol{y}_k - \alpha_k f'(\boldsymbol{y}_k)], \\ & \phi_{k+1}^* = & (1-\alpha_k)\phi_k^* + \alpha_k f(\boldsymbol{y}_k) - \dfrac{\alpha_k^2}{2\gamma_{k+1}}\|f'(\boldsymbol{y}_k)\|_2^2 \\ & & +\dfrac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\left(\dfrac{\mu}{2}\|\boldsymbol{y}_k - \boldsymbol{v}_k\|_2^2 + \langle f'(\boldsymbol{y}_k), \boldsymbol{v}_k - \boldsymbol{y}_k\rangle\right), \\ & \phi_{k+1}(\boldsymbol{x}) = & \phi_{k+1}^* + \dfrac{\gamma_{k+1}}{2}\|\boldsymbol{x} - \boldsymbol{v}_{k+1}\|_2^2. \end{array}$$

Then, we satisfy all the conditions of Lemma 9.2 for the $\lambda_k = \displaystyle\prod_{i=-1}^{k-1}(1-\alpha_k)$.

Proof:

In fact, due to Lemmas 9.3 and 9.4, it just remains to show that $\alpha_k \in (0,1]$ for $(k = 0, 1, \ldots)$ such that $\displaystyle\sum_{k=0}^\infty \alpha_k = \infty$. In the special case of $\mu = 0$, we must show that $\alpha_k < 1 \quad (k = 0, 1, \ldots)$. And finally that $f(\boldsymbol{x}_k) \leq \phi_k^*$.

Let us show both using induction hypothesis.

Consider the quadratic equation in α, $q_0(\alpha) := L\alpha^2 + (\gamma_0 - \mu)\alpha - \gamma_0 = 0$. Notice that its discriminant $\Delta := (\gamma_0 - \mu)^2 + 4\gamma_0 L$ is always positive by the hypothesis. Also, $q_0(0) = -\gamma_0 < 0$, but due to the hypothesis again. Therefore, this equation always has a root $\alpha_0 > 0$. Since $q_0(1) = L - \mu \geq 0$, $\alpha_0 \leq 1$, and we have $\alpha_0 \in (0,1]$. If $\mu = 0$, and $\alpha_0 = 1$, we will have $L = 0$ which implies $\gamma_0 = 0$ which contradicts our hypothesis. Then $\alpha_0 < 1$. In addition, $\gamma_1 := (1-\alpha_0)\gamma_0 + \alpha_0\mu > 0$ and $\gamma_0 + \alpha_0\mu > 0$. The same arguments are valid for any k. Therefore, $\alpha_k \in (0,1]$, and $\alpha_k < 1 \quad (k = 0, 1, \ldots,)$ if $\mu = 0$.

Finally, $L\alpha_k^2 = (1-\alpha_k)\gamma_k + \alpha_k\mu \geq (1-\alpha_k)\mu + \alpha_k\mu = \mu$. And we have $\alpha_k \geq \sqrt{\frac{\mu}{L}}$, and therefore, $\sum_{k=0}^{\infty} \alpha_k = \infty$, if $\mu > 0$. For the case $\mu = 0$, the argument is the same as the proof of Theorem 9.6.

Now, suppose that 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,

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

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:

$$\phi_{k+1}^* \geq f(\boldsymbol{y}_k) - \frac{\alpha_k^2}{2\gamma_{k+1}}\|f'(\boldsymbol{y}_k)\|_2^2 + (1-\alpha_k)\langle f'(\boldsymbol{y}_k), \frac{\alpha_k\gamma_k}{\gamma_{k+1}}(\boldsymbol{v}_k - \boldsymbol{y}_k) + \boldsymbol{x}_k - \boldsymbol{y}_k\rangle + \frac{\alpha_k(1-\alpha_k)\gamma_k\mu}{2\gamma_{k+1}}\|\boldsymbol{y}_k - \boldsymbol{v}_k\|_2^2.$$

Recall that since f' is L-Lipschitz continuous, if we apply Lemma 3.4 to $\boldsymbol{y}_k$ and $\boldsymbol{x}_{k+1} = \boldsymbol{y}_k - \frac{1}{L}f'(\boldsymbol{y}_k)$, we obtain

$$f(\boldsymbol{y}_k) - \frac{1}{2L}\|f'(\boldsymbol{y}_k)\|_2^2 \geq f(\boldsymbol{x}_{k+1}).$$

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 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 $L \geq \gamma_0 \geq \mu \geq 0$. 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 $f'(\boldsymbol{y}_k)$.
Step 4:	Find $\boldsymbol{x}_{k+1}$ such that $f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{y}_k) - \frac{1}{2L}\|f'(\boldsymbol{y}_k)\|_2^2$ 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.