Scheme for the Optimal Gradient Method" is an optimal method in terms of complexity for the dominant term $\ln(\varepsilon^{-1})$.

∎

Remark 9.8 Many times, you will find in articles that a method has "optimal rate of convergence". In our case, if we apply the "General Scheme for the Optimal Gradient Method" to $\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$, the number of iterations of this method to obtain $f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) < \varepsilon$ is $k = k(L, \boldsymbol{x}_0, \boldsymbol{x}^*, \varepsilon) = \mathcal{O}\left(\sqrt{\frac{L\|\boldsymbol{x}_0-\boldsymbol{x}^*\|_2^2}{\varepsilon}}\right)$ and $k = k(L, \mu, \boldsymbol{x}_0, \boldsymbol{x}^*, \varepsilon) = \mathcal{O}\left(\sqrt{\frac{L}{\mu}}\ln\frac{L\|\boldsymbol{x}_0-\boldsymbol{x}^*\|_2^2}{\varepsilon}\right)$ for $f(\boldsymbol{x}) \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$ and $\mathcal{S}_{L,\mu}^{1,1}(\mathbb{R}^n)$, respectively.

It is extremely important to note that this value is the maximum number of iterations in the worse case scenario. To obtain the total complexity of the method, you need to multiply the above number by the number of floating-point operations per iteration. This value also vary according to the method.

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 $\boldsymbol{x}_{k+1} := \boldsymbol{y}_k - \frac{1}{L}\boldsymbol{\nabla f}(\boldsymbol{y}_k)$ (see proof of Theorem 9.5). From the calculations given at Exercise 1, 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 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 $\boldsymbol{\nabla f}(\boldsymbol{y}_k)$.
Step 2:	Set $\boldsymbol{x}_{k+1} := \boldsymbol{y}_k - \frac{1}{L}\boldsymbol{\nabla f}(\boldsymbol{y}_k)$.
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 $\{\boldsymbol{x}_k\}_{k=0}^{\infty}$ and $\{\boldsymbol{y}_k\}_{k=0}^{\infty}$ generated by the "General Scheme" and the "Constant Step Scheme for the Optimal Gradient Methods" are exactly the same[4] if we choose $\boldsymbol{x}_{k+1} := \boldsymbol{y}_k - \frac{1}{L}\boldsymbol{\nabla f}(\boldsymbol{y}_k)$ in the former method. Therefore, the result of Theorem 9.6 is still valid for $\gamma_0 := \alpha_0(\alpha_0 L - \mu)/(1-\alpha_0)$.

Also, 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.

9.1 Discussion on Particular Cases

9.1.1 Accelerated Gradient Method for Smooth (Differentiable) Strongly Convex Functions

In this case, we have $\mu > 0$ and choosing $\gamma_0 := \alpha_0(\alpha_0 L - \mu)/(1-\alpha_0) = \mu$, we can have further simplifications:

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

[4]strictly speaking, there is a one index difference between $\boldsymbol{y}_k$'s on these two methods due to the order $\boldsymbol{y}_k$ is defined in the loop.

Accelerated Gradient Method for Smooth Strongly Convex Function	
Step 0:	Choose $\boldsymbol{x}_0 \in \mathbb{R}^n$, set $\boldsymbol{y}_0 := \boldsymbol{x}_0$ and $k := 0$.
Step 1:	Compute $\boldsymbol{\nabla f}(\boldsymbol{y}_k)$.
Step 2:	Set $\boldsymbol{x}_{k+1} := \boldsymbol{y}_k - \frac{1}{L}\boldsymbol{\nabla f}(\boldsymbol{y}_k)$.
Step 3:	Set $\boldsymbol{y}_{k+1} := \boldsymbol{x}_{k+1} + \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k)$, $k := k+1$ and go to Step 1.

9.1.2 Accelerated Gradient Method for Smooth (Differentiable) Convex Functions

In the case $\mu = 0$, there are much simpler variation of the method[5].

Nesterov's Accelerated Gradient Method for Smooth Convex Function	
Step 0:	Choose $\boldsymbol{x}_0 \in \mathbb{R}^n$, set $\boldsymbol{y}_0 := \boldsymbol{x}_0$, $t_0 := 1$, and $k := 0$.
Step 1:	Compute $\boldsymbol{\nabla f}(\boldsymbol{y}_k)$.
Step 2:	Set $\boldsymbol{x}_{k+1} := \boldsymbol{y}_k - \frac{1}{L}\boldsymbol{\nabla f}(\boldsymbol{y}_k)$.
Step 3:	$t_{i+1} := \dfrac{1+\sqrt{1+4t_i^2}}{2}$.
Step 4:	Set $\boldsymbol{y}_{k+1} := \boldsymbol{x}_{k+1} + \dfrac{t_i - 1}{t_{i+1}}(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k)$, $k := k+1$ and go to Step 1.

Moreover, this is equivalent to the following update as well.

Nesterov's Accelerated Gradient Method for Smooth Convex Function	
Step 0:	Choose $\boldsymbol{x}_0 \in \mathbb{R}^n$, set $\boldsymbol{y}_0 := \boldsymbol{x}_0$ and $k := 1$.
Step 1:	Compute $\boldsymbol{\nabla f}(\boldsymbol{y}_{k-1})$.
Step 2:	Set $\boldsymbol{x}_k := \boldsymbol{y}_{k-1} - \frac{1}{L}\boldsymbol{\nabla f}(\boldsymbol{y}_{k-1})$.
Step 3:	Set $\boldsymbol{y}_k := \boldsymbol{x}_k + \dfrac{k-1}{k+2}(\boldsymbol{x}_k - \boldsymbol{x}_{k-1})$, $k := k+1$ and go to Step 1.

The Nesterov's Accelerated Gradient Method for $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$ generates a sequence $\{\boldsymbol{x}_k\}_{k=0}^{\infty}$ such that

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \leq \frac{2L\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2}{(k+1)^2}.$$

Recently, itt was shown that an extension of this method guarantee a $o(k^{-2})$ convergence for $f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*)$ by Attouch and Peypouquet[6].

9.2 Exercises

1. We want to justify the Constant Step Scheme of the Optimal Gradient Method. This is a particular case of the General Scheme for the Optimal Gradient Method for the following choice:

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

[5] 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. It also has a scheme to estimate L in the case this constant in unknown.

[6] Hedy Attouch and Juan Peypouquet, "The rate of convergence of Nesterovs accelerated forward-backward method is actually faster than $1/k^2$," *SIAM Journal on Optimization* **26** (2016), pp. 1824-1834.

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

10 Extension of the Optimal Gradient Method (First-Order Method, Accelerated Gradient Method, Fast Gradient Method) for the Min-Max Problems over Simple Closed Convex Sets

Suppose we are given Q a closed convex subset of $\mathbb{R}^n$, simple enough to have an easy projection onto it. *E.g.*, positive orthant, n-dimensional box, simplex, Euclidean ball, ellipsoids, *etc.*

Given $f_i \in \mathcal{S}^{1,1}_{\mu,L}(Q)$ $(i = 1, 2, \dots, m)$, we define the following function $f : Q \to \mathbb{R}$,

$$f(\boldsymbol{x}) := \max_{1 \le i \le m} f_i(\boldsymbol{x}) \qquad \text{for} \qquad \boldsymbol{x} \in Q. \tag{15}$$

This function is non-differentiable in general, but convex (see Theorem 6.6). We will see that the method discussed so far can be easily adapted for the following min-max-type convex optimization problem.

$$\begin{cases} \text{minimize} & f(\boldsymbol{x}) \\ \text{subject to} & \boldsymbol{x} \in Q, \end{cases} \tag{16}$$

where Q is a closed convex set with a simple structure, and $f(\boldsymbol{x})$ is defined as above.

For a given $\bar{\boldsymbol{x}} \in Q$, let us define the following linearization of $f(\boldsymbol{x})$ at $\bar{\boldsymbol{x}}$.

$$f(\bar{\boldsymbol{x}}; \boldsymbol{x}) := \max_{1 \le i \le m} \left[f_i(\bar{\boldsymbol{x}}) + \langle \boldsymbol{\nabla f}_i(\bar{\boldsymbol{x}}), \boldsymbol{x} - \bar{\boldsymbol{x}} \rangle \right], \qquad \text{for } \boldsymbol{x} \in \mathbb{R}^n.$$

Lemma 10.1 Let $f_i \in \mathcal{S}^{1,1}_{\mu,L}(Q)$ $(i = 1, 2, \dots, m)$. For $\boldsymbol{x} \in Q$, we have

$$f(\boldsymbol{x}) \ge f(\bar{\boldsymbol{x}}; \boldsymbol{x}) + \frac{\mu}{2}\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|_2^2,$$
$$f(\boldsymbol{x}) \le f(\bar{\boldsymbol{x}}; \boldsymbol{x}) + \frac{L}{2}\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|_2^2.$$

Proof:
It follows from the properties of $f_i \in \mathcal{S}^{1,1}_{\mu,L}(Q)$. ∎

Theorem 10.2 A point $\boldsymbol{x}^* \in Q$ is an optimal solution of (16) with $f_i \in \mathcal{S}^{1,1}_{\mu,L}(Q)$ $(i = 1, 2, \dots, m)$ if and only if

$$f(\boldsymbol{x}^*; \boldsymbol{x}) \ge f(\boldsymbol{x}^*; \boldsymbol{x}^*) = f(\boldsymbol{x}^*), \quad \forall \boldsymbol{x} \in Q.$$

Proof:
Indeed, if the inequality is true, it follows from Lemma 10.1 that

$$f(\boldsymbol{x}) \ge f(\boldsymbol{x}^*; \boldsymbol{x}) + \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2 \ge f(\boldsymbol{x}^*) + \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2 \ge f(\boldsymbol{x}^*), \quad \forall \boldsymbol{x} \in Q.$$

For the converse, let $\boldsymbol{x}^*$ be an optimal solution of the minimization problem (16). Assume by contradiction that there is a $\boldsymbol{x} \in Q$ such that $f(\boldsymbol{x}^*; \boldsymbol{x}) < f(\boldsymbol{x}^*)$.