Let us analyze the case when $\mu > 0$. From Theorem 6.2, we know that we can find a function $f \in \mathcal{S}^{\infty,1}_{\mu,L}(\ell^2)$ 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 inequality above

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \leq L\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2\left(1 - \sqrt{\frac{\mu}{L}}\right)^k \leq L\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2\exp\left(-\frac{k}{\sqrt{L/\mu}}\right),$$

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 $k > \sqrt{L/\mu}\left(\ln\frac{1}{\varepsilon} + \ln L + 2\ln\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2\right)$.

Now, let us analize the sequences $\{\boldsymbol{x}_k\}_{k=0}^{\infty}$ generated by the method. Again from Theorem 6.2, we can find a function $f \in \mathcal{S}^{\infty,1}_{\mu,L}(\ell^2)$ such that

$$\|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2 \geq \left(\frac{\sqrt{L/\mu}-1}{\sqrt{L/\mu}+1}\right)^{2k}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2 \geq \exp\left(-\frac{4k}{\sqrt{L/\mu}-1}\right)\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2.$$

Therefore, the worst case bound to find $\boldsymbol{x}_k$ such that $\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 < \varepsilon$ can not be better than

$$k > \frac{\sqrt{L/\mu}-1}{4}\left(\ln\frac{1}{\varepsilon} + 2\ln\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2\right).$$

On the other hand, from the inequality above

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2 \leq \frac{2L}{\mu}\left(1 - \sqrt{\frac{\mu}{L}}\right)^k \leq \frac{2L}{\mu}\exp\left(-\frac{k}{\sqrt{L/\mu}}\right).$$

Therefore, we can guarantee $\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2 < \varepsilon$ for $k > \sqrt{L/\mu}\left(\ln\frac{1}{\varepsilon} + \ln 2L - \ln\mu + 2\ln\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2\right)$.

This shows that the constant step scheme for the Nesterov's gradient method is an optimal method in terms of complexity for the dominant term $\ln(\varepsilon^{-1})$. ∎

**Remark 8.9** Many times, you will find in articles that a method has "optimal rate of convergence". In our case, if we apply the constant step scheme for the Nesterov's 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 $f(\boldsymbol{x}) \in \mathcal{F}^{1,1}_L(\mathbb{R}^n)$ and $\mathcal{S}^{1,1}_{L,\mu}(\mathbb{R}^n)$, 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.

## 8.1 Discussion on Particular Cases

### 8.1.1 Nesterov's Optimal 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:

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

| **Nesterov's Optimal 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. |

### 8.1.2 Optimal Gradient Method for Smooth (Differentiable) Convex Functions

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

| **Nesterov's Original Optimal 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:** | $t_{k+1} := \dfrac{1 + \sqrt{1 + 4t_k^2}}{2}$. |
| **Step 4:** | Set $\boldsymbol{y}_{k+1} := \boldsymbol{x}_{k+1} + \dfrac{t_k - 1}{t_{k+1}}(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k)$, $k := k + 1$ and go to Step 1. |

Moreover, there is a simpler variant of this method.

| **Variant of Nesterov's Optimal 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. |

All of above methods generate sequence $\{\boldsymbol{x}_k\}_{k=0}^{\infty}$ such that

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \leq \frac{4L\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2}{(k+1)^2}.$$

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

Recently, it 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].

[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.

| **Kim-Fessler's Optimal 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:** $t_{k+1} := \begin{cases} \frac{1+\sqrt{1+4t_k^2}}{2}, & \text{if } k < N-2 \\ \frac{1+\sqrt{1+8t_k^2}}{2}, & \text{if } k = N-1 \end{cases}$. |
| **Step 4:** Set $\boldsymbol{y}_{k+1} := \boldsymbol{x}_{k+1} + \frac{t_k - 1}{t_{k+1}}(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k) + \frac{t_k}{t_{k+1}}(\boldsymbol{x}_{k+1} - \boldsymbol{y}_k)$, $k := k+1$ and go to Step 1. |

It can be shown that the Kim-Fessler's method generate sequence $\{\boldsymbol{x}_k\}_{k=0}^N$ such that

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

for $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$[7].

### 8.2 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:

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

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

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

$$f(\boldsymbol{x}) := \max_{1 \leq i \leq m} f_i(\boldsymbol{x}) \qquad \text{for} \qquad \boldsymbol{x} \in Q. \tag{16}$$

[7]Donghwan Kim and Jeffrey A. Fessler, "Optimized first-order methods for smooth convex minimization," *Mathematical Programming* **159** (2016), pp. 81–107.