Theorem 7.2 For any $\boldsymbol{x}_0 \in \mathbb{R}^\infty$, there exists a function $f \in \mathcal{S}_{\mu,L}^{\infty,1}(\mathbb{R}^\infty)$ such that for any gradient based method of type $\mathcal{M}$, we have

$$\begin{aligned} f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) &\geq \frac{\mu}{2}\left(\frac{\sqrt{L/\mu}-1}{\sqrt{L/\mu}+1}\right)^{2k}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2, \\ \|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2 &\geq \left(\frac{\sqrt{L/\mu}-1}{\sqrt{L/\mu}+1}\right)^{2k}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2, \end{aligned}$$

where $\boldsymbol{x}^*$ is the minimum of $f(\boldsymbol{x})$.

Proof:

This type of methods are invariant with respect to a simultaneous shift of all objects in the space of variables. Therefore, we can assume that $\boldsymbol{x}_0 = \{0\}_{i=1}^\infty$.

Consider the following quadratic function

$$f_{\mu,L}(\boldsymbol{x}) = \frac{\mu(L/\mu - 1)}{8}\left\{[\boldsymbol{x}]_1^2 + \sum_{i=1}^{\infty}([\boldsymbol{x}]_i - [\boldsymbol{x}]_{i+1})^2 - 2[\boldsymbol{x}]_1\right\} + \frac{\mu}{2}\|\boldsymbol{x}\|_2^2.$$

Then

$$\boldsymbol{\nabla f}_{\mu,L}(\boldsymbol{x}) = \left(\frac{\mu(L/\mu-1)}{4}\boldsymbol{A} + \mu\boldsymbol{I}\right)\boldsymbol{x} - \frac{\mu(L/\mu-1)}{4}\boldsymbol{e}_1,$$

where $\boldsymbol{A}$ is the same tridiagonal matrix defined in Theorem 7.1, but with infinite dimension and $\boldsymbol{e}_1 \in \mathbb{R}^\infty$ is a vector where only the first element is one.

After some calculations, we can show that $\mu\boldsymbol{I} \preceq \boldsymbol{\nabla}^2\boldsymbol{f}(\boldsymbol{x}) \preceq L\boldsymbol{I}$ and therefore, $f(\boldsymbol{x}) \in \mathcal{S}_{\mu,L}^{\infty,1}(\mathbb{R}^\infty)$, due to Corollary 6.21.

The minimal optimal solution of this function is:

$$[\boldsymbol{x}^*]_i := q^i = \left(\frac{\sqrt{L/\mu}-1}{\sqrt{L/\mu}+1}\right)^i, \quad i = 1, 2, \ldots$$

Then

$$\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2 = \sum_{i=1}^{\infty}[\boldsymbol{x}^*]_i^2 = \sum_{i=1}^{\infty} q^{2i} = \frac{q^2}{1-q^2}.$$

Now, since $\boldsymbol{\nabla f}_{\mu,L}(\boldsymbol{x}_0) = -\frac{\mu(L/\mu-1)}{4}\boldsymbol{e}_1$, and $\boldsymbol{A}$ is a tridiagonal matrix, $[\boldsymbol{x}_k]_i = 0$ for $i = k+1, k+2, \ldots$, and

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2 \geq \sum_{i=k+1}^{\infty}[\boldsymbol{x}^*]_i^2 = \sum_{i=k+1}^{\infty} q^{2i} = \frac{q^{2(k+1)}}{1-q^2} = q^{2k}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2.$$

Finally, the first inequality follows from Corollary 6.17. ∎

8 The Steepest Descent Method for Differentiable Convex and Differentiable Strongly Convex Functions with Lipschitz Continuous Gradients

Let us consider the steepest descent method with constant step h.

Theorem 8.1 Let $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$, and $0 < h < \frac{2}{L}$. The steepest descent method with constant step generates a sequence which converges as follows:

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

Proof:

Denote $r_k = \|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2$. Then

$$\begin{aligned}
r_{k+1}^2 &= \|\boldsymbol{x}_k - \boldsymbol{x}^* - h\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2 \\
&= r_k^2 - 2h\langle \boldsymbol{\nabla f}(\boldsymbol{x}_k), \boldsymbol{x}_k - \boldsymbol{x}^*\rangle + h^2\|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2 \\
&= r_k^2 - 2h\langle \boldsymbol{\nabla f}(\boldsymbol{x}_k) - \boldsymbol{\nabla f}(\boldsymbol{x}^*), \boldsymbol{x}_k - \boldsymbol{x}^*\rangle + h^2\|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2 \\
&\leq r_k^2 - h\left(\frac{2}{L} - h\right)\|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2,
\end{aligned}$$

where the last inequality follows from Theorem 6.13.

Therefore, since $0 < h < \frac{2}{L}$, $r_{k+1} < r_k < \cdots < r_0$.

Now

$$\begin{aligned}
f(\boldsymbol{x}_{k+1}) &\leq f(\boldsymbol{x}_k) + \langle \boldsymbol{\nabla f}(\boldsymbol{x}_k), \boldsymbol{x}_{k+1} - \boldsymbol{x}_k\rangle + \frac{L}{2}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|_2^2 \\
&= f(\boldsymbol{x}_k) - \omega\|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2 < f(\boldsymbol{x}_k), \qquad (10)
\end{aligned}$$

where $\omega = h(1 - \frac{L}{2}h)$. Denoting by $\Delta_k = f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*)$, from the convexity of $f(\boldsymbol{x})$, Theorem 6.7, and the Cauchy-Schwarz inequality,

$$\Delta_k = f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \leq \langle \boldsymbol{\nabla f}(\boldsymbol{x}_k), \boldsymbol{x}_k - \boldsymbol{x}^*\rangle \leq \|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2 r_k \leq \|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2 r_0. \qquad (11)$$

Combining (10) and (11),

$$\Delta_{k+1} \leq \Delta_k - \frac{\omega}{r_0^2}\Delta_k^2.$$

Thus dividing by $\Delta_k\Delta_{k+1}$,

$$\frac{1}{\Delta_{k+1}} \geq \frac{1}{\Delta_k} + \frac{\omega}{r_0^2}\frac{\Delta_k}{\Delta_{k+1}} \geq \frac{1}{\Delta_k} + \frac{\omega}{r_0^2}.$$

since $\frac{\Delta_k}{\Delta_{k+1}} \geq 1$. Summing up these inequalities we get

$$\frac{1}{\Delta_{k+1}} \geq \frac{1}{\Delta_0} + \frac{\omega}{r_0^2}(k+1).$$

∎

To obtain the optimal step size, it is sufficient to find the maximum of the function $\omega := \omega(h) = h(1 - \frac{L}{2}h)$ which is $h^* := 1/L$.

Corollary 8.2 If $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$, the steepest descent method with constant step $h = 1/L$ yields

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

That is, $\{f(\boldsymbol{x}_k)\}_{k=0}^{\infty}$ converges R-sublinearly to $f(\boldsymbol{x}^*)$.

Proof:

Left for exercise. ∎

Theorem 8.3 Let $f \in \mathcal{S}_{\mu,L}^{1,1}(\mathbb{R}^n)$, and $0 < h \leq \frac{2}{\mu+L}$. The steepest descent method with constant step generates a sequence which converges as follows:

$$\begin{aligned}
\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2 &\leq \left(1 - \frac{2h\mu L}{\mu+L}\right)^k \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2, \\
f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) &\leq \frac{L}{2}\left(1 - \frac{2h\mu L}{\mu+L}\right)^k \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2.
\end{aligned}$$

If $h = \frac{2}{\mu+L}$, then

$$\begin{aligned} f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) &\leq \frac{L}{2}\left(\frac{L/\mu - 1}{L/\mu + 1}\right)^{2k} \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2, \\ \|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2 &\leq \left(\frac{L/\mu - 1}{L/\mu + 1}\right)^{k} \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2. \end{aligned}$$

That is, $\{\boldsymbol{x}_k\}_{k=0}^{\infty}$ and $\{f(\boldsymbol{x}_k)\}_{k=0}^{\infty}$ converges R-linearly to $\boldsymbol{x}^*$ and $f(\boldsymbol{x}^*)$, respectively.

Proof:
Denote $r_k = \|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2$. Then

$$\begin{aligned} r_{k+1}^2 &= \|\boldsymbol{x}_k - \boldsymbol{x}^* - h\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2 \\ &= r_k^2 - 2h\langle \boldsymbol{\nabla f}(\boldsymbol{x}_k), \boldsymbol{x}_k - \boldsymbol{x}^*\rangle + h^2\|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2 \\ &= r_k^2 - 2h\langle \boldsymbol{\nabla f}(\boldsymbol{x}_k) - \boldsymbol{\nabla f}(\boldsymbol{x}^*), \boldsymbol{x}_k - \boldsymbol{x}^*\rangle + h^2\|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2 \\ &\leq r_k^2 - 2h\left(\frac{\mu L}{\mu + L} r_k^2 + \frac{1}{\mu + L}\|\boldsymbol{\nabla f}(\boldsymbol{x}_k) - \boldsymbol{\nabla f}(\boldsymbol{x}^*)\|_2^2\right) + h^2\|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2 \\ &= \left(1 - \frac{2h\mu L}{\mu + L}\right) r_k^2 + h\left(h - \frac{2}{\mu + L}\right)\|\boldsymbol{\nabla f}(\boldsymbol{x}_k)\|_2^2 \end{aligned}$$

from Theorems 6.13 and 6.22, and it proves the first two inequalities.

Now, for $h = 2/(L+\mu)$ and again from Theorem 6.13,

$$\begin{aligned} f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) - \langle \boldsymbol{\nabla f}(\boldsymbol{x}^*), \boldsymbol{x}_k - \boldsymbol{x}^*\rangle &\leq \frac{L}{2}\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2 \\ &\leq \frac{L}{2}\left(\frac{L/\mu - 1}{L/\mu + 1}\right)^{2k} r_0^2. \end{aligned}$$

∎

Theorem 8.4 (Yuan 2010) [2] In the special case of a strongly convex quadratic function $f(\boldsymbol{x}) = \frac{1}{2}\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x}\rangle + \langle \boldsymbol{a}, \boldsymbol{x}\rangle + \alpha$ with $\lambda_1(\boldsymbol{A}) = L \geq \lambda_n(\boldsymbol{A}) = \mu > 0$, we can obtain

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

for the steepest descent method with "exact line search".

- 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 $\boldsymbol{x}_0 \in \mathbb{R}^n$ (due to convexity).
- 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.

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

8.1 Exercises

1. Prove Corollary 8.2.

2. Consider a sequence $\{\beta_k\}_{k=0}^{\infty}$ which converges to zero.

   The sequence is said to converge *Q-sublinearly* if

$$\lim_{k\to\infty} \sup \left| \frac{\beta_{k+1}}{\beta_k} \right| = 1.$$

   A zero converging sequence $\{\beta_k\}_{k=0}^{\infty}$ is said to converge *R-sublinearly* if it is dominated by a Q-sublinearly converging sequence. That is, if there is a Q-sublinearly converging sequence $\{\hat{\beta}_k\}_{k=0}^{\infty}$ such that $0 \leq |\beta_k| \leq \hat{\beta}_k$.

   (a) ~~Show that a Q-linear converging sequence is a Q-sublinear converging sequence.~~

   (b) Give an example of a Q-sublinear converging sequence which is not Q-linear converging sequence.

   (c) Give an example of a R-sublinear converging sequence which is not R-linear converging sequence.

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 $\mathbb{R}^n$.

[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.