Chapter 5

Gradient Method

This chapter introduces an important class of algorithms for solving unconstrained optimization problems. The central concept is that of a descent direction.

5.1 Order of Convergence of a Sequence

In this section, we introduce the notion of the order of convergence of a numerical sequence $\{x_k\}_{k\in\mathbb{N}}$, which will be useful in the remainder of the course. The higher the order of convergence, the faster the method converges and the less computational effort is needed to determine the solution.

Definition 5.1.1 (Order of Convergence)**.** Let $\{x_k\}_{k\in\mathbb{N}}$ be a convergent sequence with limit x^*. The order of convergence of $\{x_k\}$ is the positive integer p (if it exists) such that

$$0 \leq \lim_{k\to\infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|^p} = r \leq \infty.$$

The constant r is called the **rate of convergence**.

We distinguish the following cases:

- If $p = 1$ and $0 \leq r < 1$, then the convergence is **linear**.
- If $p = 1$ and $r = 0$, then the convergence is **superlinear**.
- If $p = 1$ and $r = 1$, then the convergence is **sublinear**.
- If $p = 2$, then the convergence is **quadratic**.
- If $p = 3$, then the convergence is **cubic**.

Example 5.1.2. Let $x_k = 2^{-k}$. Then

$$\lim_{k\to\infty} \frac{|x_{k+1} - x^*|}{|x_k - x^*|^p} = \begin{cases} 0, & p = 0, \\ 1, & p = 1, \\ +\infty, & p \geq 2. \end{cases}$$

Thus, the convergence is sublinear of order 1.

Example 5.1.3. Let $x_k = \dfrac{1}{3^k}$. Then

$$\lim_{k\to\infty} \frac{|x_{k+1} - x^*|}{|x_k - x^*|^p} = 3^{\,k(p-1)-1}.$$

Therefore,

$$\lim_{k\to\infty} \frac{|x_{k+1} - x^*|}{|x_k - x^*|^p} = \begin{cases} 0, & p = 0, \\ \frac{1}{3}, & p = 1, \\ +\infty, & p \geq 2. \end{cases}$$

Hence, the convergence is linear of order 1 with rate $r = \frac{1}{3}$.

5.2 Descent Method

To define descent methods, we need the notion of a descent direction.

Definition 5.2.1 (Descent Direction)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$. A vector $d \in \mathbb{R}^n$ is called a **descent direction** at $x \in \mathbb{R}^n$ if there exists $\alpha^* > 0$ such that

$$f(x + \alpha d) \leq f(x), \quad \forall \alpha \in (0, \alpha^*].$$

Example 5.2.2. Consider the function $f(x, y) = x^2 + y^2$. Let $\alpha^* = \frac{1}{2}$, $\hat{x} = 1$, $\hat{y} = 1$, and $d^{(1)} = (-1, -1)^T$. Then

$$f\left(\begin{bmatrix}\hat{x}\\ \hat{y}\end{bmatrix} + \alpha d^{(1)}\right) = (1-\alpha)^2 + (1-\alpha)^2 \leq f(\hat{x}, \hat{y}) = 2, \quad \forall \alpha \in (0, \alpha^*].$$

Thus, $d^{(1)}$ is a descent direction at $(\hat{x}, \hat{y})$.

On the other hand, the vector $d^{(2)} = (1, 1)^T$ is not a descent direction at $(\hat{x}, \hat{y})$, since for any $\alpha > 0$,

$$f\left(\begin{bmatrix}\hat{x}\\ \hat{y}\end{bmatrix} + \alpha d^{(2)}\right) = (1+\alpha)^2 + (1+\alpha)^2 \geq f(\hat{x}, \hat{y}) = 2, \quad \forall \alpha \in (0, \alpha^*].$$

The following result gives an important characterization of descent directions at a point x:

Proposition 5.2.3. Let f be continuously differentiable on $\mathbb{R}^n$, and let $x, d \in \mathbb{R}^n$. Then:

1. If d is a descent direction at x, then

$$\langle \nabla f(x), d \rangle \leq 0.$$

2. If $\langle \nabla f(x), d \rangle < 0$, then d is a descent direction at x.

Proof. **(1)** Suppose d is a descent direction at x. By definition, there exists $\alpha^* > 0$ such that

$$f(x + \alpha d) \leq f(x), \quad \forall \alpha \in (0, \alpha^*].$$

Define $\varphi(\alpha) = f(x + \alpha d)$. For $\alpha \in (0, \alpha^*]$, we have $\varphi(\alpha) \leq \varphi(0)$. Hence,

$$\frac{\varphi(\alpha) - \varphi(0)}{\alpha} \leq 0.$$

Taking the limit as $\alpha \to 0^+$, we obtain

$$\varphi'(0) = \langle \nabla f(x), d \rangle \leq 0.$$

(2) Exercise: Show that if $\langle \nabla f(x), d \rangle < 0$, then d is a descent direction at x. □

Example 5.2.4. We return to Example 5.3. We have

$$\nabla f(\hat{x}, \hat{y}) = (2, 2)^T.$$

Thus,

$$\langle \nabla f(\hat{x}, \hat{y}), d^{(1)} \rangle = -4 \leq 0,$$

which shows that $d^{(1)}$ is a descent direction at $(\hat{x}, \hat{y})$.

For the other direction $d^{(2)}$, we have

$$\langle \nabla f(\hat{x}, \hat{y}), d^{(2)} \rangle = 4 \geq 0,$$

which implies that $d^{(2)}$ is not a descent direction at $(\hat{x}, \hat{y})$.

In fact, the set of descent directions at $(\hat{x}, \hat{y})$ is the set of vectors that form an **obtuse angle** with the gradient vector $\nabla f(\hat{x}, \hat{y})$, as illustrated in Figure 5.1.

Remark 5.2.5. Descent methods differ according to the choice of d and α. The general algorithm for a descent method is the following:

[H] Descent Method

1. Step I: Set $k = 0$, choose a tolerance ε and an initial point $x^{(0)}$.
2. Step II:
   - Find the descent direction $d^{(k)}$.
   - Find the step size α_k.

Remark 5.1. Descent methods differ by the choice of d and α. The general algorithm for a descent method is the following:

Algorithm 5.1. Descent Method

Step I: $k = 0$, ϵ and $x^{(0)}$.

Step II:

- Find the descent direction $d^{(k)}$.
- Find the step size α_k.
- Compute $x^{(k+1)} = x^{(k)} + \alpha_k d^{(k)}$.

Step III: If $\|\nabla f(x^{(k+1)})\| \leq \epsilon$ then stop, $x^* = x^{(k+1)}$, otherwise set $k = k + 1$ and go to Step II.

5.3 Gradient Method

In the gradient method, we choose $d^{(k)} = -\nabla f(x^{(k)})$ as the descent direction, because if $\nabla f(x^{(k)}) \neq 0$, then

$$\langle -\nabla f(x^{(k)}), \nabla f(x^{(k)}) \rangle = -\|\nabla f(x^{(k)})\|^2 \leq 0.$$

There are many ways to use this descent direction. If we use a fixed step, i.e. $\alpha_k = \alpha$, we obtain the gradient method with fixed step size:

$$d^{(k)} = -\nabla f(x^{(k)}), \quad x^{(k+1)} = x^{(k)} + \alpha d^{(k)}.$$

In the case of a quadratic function

$$f(x) = \frac{1}{2}\langle x, Ax \rangle + \langle x, b \rangle + c,$$

with $A \in M_n$ symmetric positive definite, $b \in \mathbb{R}^n$ and $c \in \mathbb{R}$, we have the following convergence result:

Theorem 5.3.1 (5.1)**.** *Let λ be the smallest eigenvalue of A and L the largest eigenvalue of A. If $\alpha \in \left]0, \frac{2}{L}\right[$, then the convergence of the gradient method with fixed step size is linear with a rate bounded above by*

$$\max\big(|1 - \alpha L|, |1 - \alpha\lambda|\big).$$

Proof. We have

$$x^{(k+1)} = x^{(k)} - \alpha(Ax^{(k)} + b), \tag{5.1}$$

and

$$x^{(k)} = x^{(k-1)} - \alpha(Ax^{(k-1)} + b). \tag{5.2}$$

Subtracting relations (5.1)(5.2), we obtain

$$x^{(k+1)} - x^{(k)} = x^{(k)} - x^{(k-1)} - \alpha A(x^{(k)} - x^{(k-1)}),$$

which implies

$$x^{(k+1)} - x^{(k)} = (I_n - \alpha A)(x^{(k)} - x^{(k-1)}).$$

Hence,

$$\|x^{(k+1)} - x^{(k)}\| = \|(I_n - \alpha A)(x^{(k)} - x^{(k-1)})\| \leq \max_{i=1,\ldots,n} |\lambda_i(I_n - \alpha A)| \cdot \|x^{(k)} - x^{(k-1)}\|.$$

In this case,

$$1 - \alpha L \leq \lambda_i(I_n - \alpha A) \leq 1 - \alpha\lambda,$$

and therefore,

$$\max_{i=1,\ldots,n} |\lambda_i(I_n - \alpha A)| = \max(|1 - \alpha L|, |1 - \alpha\lambda|).$$

Consequently, the sequence $\{x^{(k)}\}_{k\in\mathbb{N}}$ converges if $\max(|1 - \alpha L|, |1 - \alpha\lambda|) < 1$, which corresponds to

$$\alpha \in]0, \tfrac{2}{L}[\cap]0, \tfrac{2}{\lambda}[.$$

□

Example 5.3.2 (5.5)**.** Let the quadratic function

$$f(x,y) = 2x^2 + 3y^2 - 2xy + 5x - 6.$$

The Hessian matrix of f is

$$D^2 f(x,y) = \begin{pmatrix} 4 & -2 \\ -2 & 6 \end{pmatrix}.$$

The eigenvalues are: $L = \sqrt{5}+5$ and $\lambda = 5-\sqrt{5}$. According to Theorem 5.1, for $\alpha \in \left]0, \frac{2}{\sqrt{5}+5}\right[$, the sequence generated by the gradient method with fixed step size converges linearly to

$$(x^*, y^*)^T = \left(-\tfrac{3}{2}, -\tfrac{1}{2}\right)^T,$$

with a rate at most

$$\max\Big(\big|1-\alpha(\sqrt{5}+5)\big|,\ \big|1-\alpha(5-\sqrt{5})\big|\Big).$$

The results obtained by the method for $\alpha = 0.1382$ and initial point $(x_0, y_0) = (1, \frac{5}{2})$ are shown in Figure 5.2.

Theorem 5.3.3 (5.2)**.** *Let f be a continuously differentiable function on $\mathbb{R}^n$, bounded below, and let $x^{(0)} \in \mathbb{R}^n$. If $\nabla f(x)$ is L-Lipschitz continuous and $0 \le \alpha \le \frac{2}{L}$, then the sequence*

$$\Big\{ f\big(x^{(k)} - \alpha \nabla f(x^{(k)})\big) \Big\}_{k\in\mathbb{N}}$$

converges to a finite limit. Moreover, the sequence

$$\Big\{ \nabla f(x^{(k)}) \Big\}_{k\in\mathbb{N}}$$

converges to zero in $\mathbb{R}^n$.

To prove this result, we need the following lemma:

Lemma 5.3.4 (5.1, [**?**])**.** *Let f be a continuously differentiable function on $\mathbb{R}^n$. If $\nabla f(x)$ is L-Lipschitz continuous, then for all $x, y \in \mathbb{R}^n$,*

$$f(y) \le f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y-x\|^2.$$

Proof. Let $x^{(k)} \in S_0 = \{x \in \mathbb{R}^n : f(x) \le f(x^{(0)})\}$. Applying Lemma 5.1 with $x = x^{(k)}$ and $y = x^{(k)} - \alpha\nabla f(x^{(k)})$, we obtain

$$f\big(x^{(k)} - \alpha\nabla f(x^{(k)})\big) - f(x^{(k)}) \le -\alpha\Big(1 - \frac{\alpha L}{2}\Big)\|\nabla f(x^{(k)})\|^2.$$

For $\alpha \in \big[0, \frac{2}{L}\big]$, this inequality becomes

$$f\big(x^{(k)} - \alpha\nabla f(x^{(k)})\big) - f(x^{(k)}) \le -\alpha\Big(1 - \frac{\alpha L}{2}\Big)\|\nabla f(x^{(k)})\|^2 \le 0. \tag{5.3}$$

Hence, the sequence $\{f(x^{(k)})\}_{k\in\mathbb{N}}$ is strictly decreasing. Since it is bounded below, it converges to a finite limit. Taking the limit in inequality (5.3), we conclude that

$$\lim_{k\to\infty} \nabla f(x^{(k)}) = 0_{\mathbb{R}^n}.$$

□

We now consider the case where the step size α_k is chosen optimally, in the sense that the function

$$\Phi_k(\alpha) = f\big(x^{(k)} + \alpha d^{(k)}\big)$$

decreases as much as possible with respect to α.

Definition 5.3.5 (5.3, Gradient method with optimal step size)**.** Let f be a continuously differentiable function on $\mathbb{R}^n$ and $x^{(0)} \in \mathbb{R}^n$. The gradient method with optimal step size is defined by

$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)}),$$

where

$$\alpha_k = \arg\min_{\alpha \geq 0} \; f\big(x^{(k)} - \alpha \nabla f(x^{(k)})\big).$$

The function Φ_0 has a single critical point at $\alpha_0 = \frac{1}{2}$, which is a global minimum since

$$\Phi_0''(\alpha) = 64 \geq 0.$$

Thus,

$$x^{(1)} = x^{(0)} - \alpha_0 \nabla f(x^{(0)}) = \begin{pmatrix} 2 \\ 3 \end{pmatrix} - \tfrac{1}{2} \begin{pmatrix} 4 \\ 4 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Computing the gradient of f at the point $x^{(1)}$, we find

$$\nabla f(x^{(1)}) = (-4, 4)^T \neq 0_{\mathbb{R}^2}.$$

At iteration $k = 1$, we have

$$\Phi_1(\alpha) = f\big(x^{(1)} - \alpha \nabla f(x^{(1)})\big) = f(4\alpha, 1 - 4\alpha).$$

The derivative of this function is

$$\Phi_1'(\alpha) = -\big\langle \nabla f(x^{(1)} - \alpha \nabla f(x^{(1)})), \nabla f(x^{(1)}) \big\rangle = -\left\langle \begin{pmatrix} 8(4\alpha) - 4(1 - 4\alpha) \\ 4(1 - 4\alpha) - 4(4\alpha) \end{pmatrix}, \begin{pmatrix} -4 \\ 4 \end{pmatrix} \right\rangle = 32(10\alpha - 1).$$

Thus, the function Φ_1 has a single critical point at $\alpha_1 = \frac{1}{10}$, which is a global minimum since

$$\Phi_1''(\alpha) = 320 \geq 0.$$

Therefore,

$$x^{(2)} = x^{(1)} - \alpha_1 \nabla f(x^{(1)}) = \begin{pmatrix} 0 \\ 1 \end{pmatrix} - \tfrac{1}{10} \begin{pmatrix} -4 \\ 4 \end{pmatrix} = \begin{pmatrix} \frac{4}{10} \\ \frac{6}{10} \end{pmatrix}.$$

In the same way, we obtain

$$x^{(3)} = \begin{pmatrix} 0 \\ \frac{2}{10} \end{pmatrix}.$$

Indeed, the function f has a global minimum at the point

$$x^\star = (0, 0)^T.$$

If we plot the sequence $\{x^{(k)}\}_{k\in\mathbb{N}}$, we see that this method follows a *zigzag trajectory at right angles* towards $x^\star$ (see Figure 5.3). In the gradient method with optimal step size, the successive directions are orthogonal, as shown by the following result:

Proposition 5.2. If α_k is optimal, then $\nabla f(x^{(k+1)})$ and $\nabla f(x^{(k)})$ are orthogonal.

Proof. If α_k is optimal, then

$$\phi_k'(\alpha_k) = 0,$$

that is,

$$\langle \nabla f(x^{(k)} - \alpha_k \nabla f(x^{(k)})), \nabla f(x^{(k)}) \rangle = 0.$$

1. $\|\nabla f(x^{(k+1)})\| \leq \varepsilon$
2. $\|x^{(k+1)} - x^{(k)}\| \leq \varepsilon$
3. $\dfrac{\|x^{(k+1)} - x^{(k)}\|}{\|x^{(k)}\|} \leq \varepsilon$
4. $\left|f(x^{(k+1)}) - f(x^{(k)})\right| \leq \varepsilon$
5. $\dfrac{\left|f(x^{(k+1)}) - f(x^{(k)})\right|}{|f(x^{(k)})|} \leq \varepsilon$

The general algorithm for the gradient method with optimal step size is the following:

Algorithm 5.2. Algorithm for the gradient method with optimal step size.

Step I: $k = 0$, choose ε and $x^{(0)}$.

Step II:

- $d^{(k)} = -\nabla f(x^{(k)})$
- $\alpha_k = \arg\min_{\alpha \geq 0} f(x^{(k)} + \alpha d^{(k)})$
- Compute $x^{(k+1)} = x^{(k)} + \alpha_k d^{(k)}$

Step III: If $\|\nabla f(x^{(k+1)})\| \leq \varepsilon$ then stop, $x^* = x^{(k+1)}$. Otherwise, set $k = k + 1$ and go to Step II.

3.1 The Quadratic Case

In general, it is not easy to determine the exact value of the optimal step size. However, for a positive definite quadratic functional we have the following result:

Lemma 5.2. Let f be a positive definite quadratic function,

$$f(x) = \frac{1}{2}\langle x, Ax\rangle + \langle x, b\rangle + c,$$

where A is a symmetric positive definite matrix, $b \in \mathbb{R}^n$, and $c \in \mathbb{R}$. with $A^T = A$ a symmetric square matrix of order n, positive definite ($A \succ 0$), $b \in \mathbb{R}^n$, and $c \in \mathbb{R}$.

Then, the optimal step size at each iteration of the steepest descent method is given by:

$$\alpha_k = \frac{\langle Ax^{(k)} + b,\ Ax^{(k)} + b\rangle}{\langle A(Ax^{(k)} + b),\ Ax^{(k)} + b\rangle} \geq 0.$$

Proof. We know that

$$\nabla f(x) = Ax + b.$$

If α_k is optimal, then it satisfies the first-order optimality condition:

$$\frac{d}{d\alpha} f\big(x^{(k)} - \alpha(Ax^{(k)} + b)\big)\Big|_{\alpha=\alpha_k} = 0.$$

Hence,

$$\frac{d}{d\alpha} f\big(x^{(k)} - \alpha(Ax^{(k)} + b)\big)\Big|_{\alpha=\alpha_k} = -\langle \nabla f(x^{(k)} - \alpha_k(Ax^{(k)} + b)),\ Ax^{(k)} + b\rangle.$$

$$\begin{aligned} &= -\langle A(x^{(k)} - \alpha_k(Ax^{(k)} + b)) + b,\ Ax^{(k)} + b\rangle \\ &= -\langle Ax^{(k)} + b - \alpha_k A(Ax^{(k)} + b),\ Ax^{(k)} + b\rangle \\ &= -\langle Ax^{(k)} + b,\ Ax^{(k)} + b\rangle + \alpha_k \langle A(Ax^{(k)} + b),\ Ax^{(k)} + b\rangle = 0. \end{aligned}$$

Since $Ax^{(k)} + b \neq 0$ and $A \succ 0$, we have $\langle A(Ax^{(k)} + b),\ Ax^{(k)} + b\rangle \neq 0$, which yields

$$\alpha_k = \frac{\langle Ax^{(k)} + b,\ Ax^{(k)} + b\rangle}{\langle A(Ax^{(k)} + b),\ Ax^{(k)} + b\rangle} \geq 0.$$

Example 5.7. Consider the optimization problem:

$$\min_{x\in\mathbb{R}^2} f(x) = x_1 + \tfrac{1}{2}x_2 + \tfrac{1}{2}x_1^2 + x_2^2 - 3.$$

We compute

$$\nabla^2 f(x) = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \succeq 0.$$

Thus, problem (5.7) admits a unique optimal solution x^* satisfying

$$\nabla f(x^*) = 0, \quad \text{where} \quad x^* = \begin{pmatrix} -1 \\ -\frac{1}{4} \end{pmatrix}.$$

We can find this solution by applying the steepest descent method with optimal step size. Starting from $x^{(0)} = (0, 0)^T$ with $\epsilon = 7 \times 10^{-3}$, we compute $\|\nabla f(x^{(0)})\| = 1.118 \geq \epsilon$.

First iteration:

$$\alpha_0 = \frac{\langle Ax^{(0)} + b,\ Ax^{(0)} + b\rangle}{\langle A(Ax^{(0)} + b),\ Ax^{(0)} + b\rangle} = \tfrac{5}{6}, \quad \text{with } A = \nabla^2 f(x),\ b = (1, 0.5)^T.$$

$$x^{(1)} = x^{(0)} - \alpha_0(Ax^{(0)} + b) = -(0.83,\ 0.41)^T.$$

We now compute successive iterations.

First iteration:

$$\|\nabla f(x^{(1)})\| = 0.37 \geq \epsilon.$$

Second iteration:

$$\alpha_1 = 0.56, \qquad x^{(2)} = x^{(1)} - \alpha_1(Ax^{(1)} + b) = (-0.93, -0.23)^T,$$

$$\|\nabla f(x^{(2)})\| = 0.08 \geq \epsilon.$$

Third iteration:

$$\alpha_2 = 0.833, \qquad x^{(3)} = x^{(2)} - \alpha_2(Ax^{(2)} + b) = (-0.983, -0.26)^T,$$

$$\|\nabla f(x^{(3)})\| = 0.02 \geq \epsilon.$$

Fourth iteration:

$$\alpha_3 = 0.5, \qquad x^{(4)} = x^{(3)} - \alpha_3(Ax^{(3)} + b) = (-0.9945, -0.2486)^T,$$

$$\|\nabla f(x^{(4)})\| = 0.006 \leq \epsilon.$$

Hence, the algorithm stops and the approximate solution is

$$x^* \approx x^{(4)}.$$

5 Exercises

Exercise 5.1. Let the function

$$f(x_1, x_2) = x_1^2 + 2x_1x_2 + 3x_2^2.$$

We want to minimize f on $\mathbb{R}^2$ using the descent method, where the main iteration is given by

$$x^{k+1} = x^k + \alpha_k d^k.$$

1. Starting from the initial point $x^{(0)} = (2, 2)^T$, are we at the optimum? 2. Consider $d^{(0)} = (1, 1)^T$ and $d^{(1)} = (-1, -1)^T$ as two possible directions. Which one is a descent direction at $x^{(0)}$? 3. After using the correct descent direction, we obtain the point $x^{(1)} = \left(\frac{3}{2}, \frac{3}{2}\right)^T$. Is the chosen step size α an optimal step?

Exercise 5.2. We consider the gradient method with a fixed step size α applied to minimize the functions $f : \mathbb{R}^2 \to \mathbb{R}$ given below. In each case, find the largest interval of α values for which the algorithm converges.

(a)

$$f(x) = 3x_1^2 + 3x_2^2 + 4x_1x_2 + 2x_1 + 1.$$

(b)

$$f(x) = x^T \begin{pmatrix} 3 & 3 \\ 1 & 3 \end{pmatrix} x - x^T \begin{pmatrix} 16 \\ 23 \end{pmatrix}.$$

Exercise 5.3. Use the Conjugate Gradient Method to solve the following optimization problem:

$$\min_{x \in \mathbb{R}^2} f(x) = \tfrac{1}{2} x^T A x + x^T b,$$

Solutions

Solution 5.1

Let

$$f(x_1, x_2) = x_1^2 - 2x_1x_2 - 3x_2^2.$$

Then

$$\nabla f(x) = \begin{pmatrix} 2x_1 - 2x_2 \\ 2x_1 - 6x_2 \end{pmatrix}.$$

1. For the initial point $x^{(0)} = (2, 2)^\top$ we have

$$\nabla f(x^{(0)}) = \begin{pmatrix} 2 \cdot 2 - 2 \cdot 2 \\ 2 \cdot 2 - 6 \cdot 2 \end{pmatrix} = \begin{pmatrix} 0 \\ -8 \end{pmatrix} \neq 0.$$

   Hence $x^{(0)}$ is not an optimum.

2. Compute the inner products with the two directions:

$$\langle d^{(0)}, \nabla f(x^{(0)}) \rangle = \langle (1, 1)^\top, (0, -8)^\top \rangle = 1 \cdot 0 + 1 \cdot (-8) = -8 < 0,$$

   so $d^{(0)} = (1, 1)^\top$ *is* a descent direction at $x^{(0)}$.

$$\langle d^{(1)}, \nabla f(x^{(0)}) \rangle = \langle (-1, -1)^\top, (0, -8)^\top \rangle = 0 + 8 = 8 > 0,$$

   so $d^{(1)} = (-1, -1)^\top$ is *not* a descent direction.

3. After using the correct descent direction one obtains

$$x^{(1)} = \begin{pmatrix} -\frac{3}{2} \\ \frac{3}{2} \end{pmatrix}.$$

   The step used is not optimal because

$$\langle d^{(0)}, \nabla f(x^{(1)}) \rangle = \left\langle (1, 1)^\top, \nabla f(-\tfrac{3}{2}, \tfrac{3}{2}) \right\rangle = -18 \neq 0,$$

   so the directional derivative along the chosen direction at $x^{(1)}$ is nonzero.

Solution 5.2 (a)

For

$$f(x) = 3x_1^2 + 3x_2^2 + 4x_1x_2 + 2x_1 + 1$$

the Hessian is

$$\nabla^2 f(x) = \begin{pmatrix} 6 & 4 \\ 4 & 6 \end{pmatrix}.$$

The eigenvalues are $\lambda_1 = 10$ and $\lambda_2 = 2$, so $\nabla^2 f \succ 0$ (positive definite). By Theorem 5.1 (fixed-step gradient for a quadratic with eigenvalues $\lambda_{\min} = \lambda_2 = 2$ and $\lambda_{\max} = \lambda_1 = 10$), the gradient method with constant step α converges provided

$$0 < \alpha < \frac{2}{\lambda_{\max}} = \frac{2}{10} = \frac{1}{5}.$$

Chapter 6

Newton's Method

6.1 Introduction

In this chapter, we introduce a well-known and widely used second-order method: Newton's method. This method belongs to one of the main classes of unconstrained optimization methods. We consider the minimization problem

$$\min_{x \in \mathbb{R}^n} f(x), \tag{6.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a function of class C^2.

At each iteration of Newton's method, we approximate f by a quadratic form using the first two terms of its Taylor expansion:

$$f(x) \approx q_k(x) = f(x^{(k)}) + (x - x^{(k)})^T \nabla f(x^{(k)}) + \frac{1}{2}(x - x^{(k)})^T H\big(x^{(k)}\big)(x - x^{(k)}), \tag{6.2}$$

with

$$H\big(x^{(k)}\big) = \nabla^2 f\big(x^{(k)}\big).$$

A necessary condition for a minimum of the quadratic function q_k is $\nabla q_k(x) = 0$, which implies

$$x^{(k+1)} = x^{(k)} - H\big(x^{(k)}\big)^{-1} \nabla f(x^{(k)}).$$

This solution exists if and only if the following conditions are satisfied:

(a) The Hessian is nonsingular.

(b) The approximation in equation (6.2) is valid in the neighborhood of the point $x^{(k)}$.

The general algorithm for Newton's method is the following:

[H] Newton's Method

1. Step I: Initialize $k = 0$, tolerance ϵ, and starting point $x^{(0)}$.

2. Step II: Compute

$$x^{(k+1)} = x^{(k)} - H\big(x^{(k)}\big)^{-1} \nabla f(x^{(k)}).$$

3. Step III: If $\|\nabla f(x^{(k+1)})\| \leq \epsilon$, stop and set $x^* = x^{(k+1)}$. Otherwise, set $k = k+1$ and return to Step II.

Example 6.1 Consider the function

$$f(x_1, x_2) = (x_1 - 2)^4 + (x_1 - 2)^2 x_2^2 + (x_1 + 1)^2,$$

$$\nabla f(x) = \begin{pmatrix} 4(x_1 - 2)^3 + 2(x_1 - 2)x_2^2 + 2(x_1 + 1) \\ 2(x_1 - 2)^2 x_2 \end{pmatrix},$$

$$\nabla^2 f(x) = \begin{pmatrix} 12(x_1 - 2)^2 + 2x_2^2 + 2 & 4(x_1 - 2)x_2 \\ 4(x_1 - 2)x_2 & 2(x_1 - 2)^2 \end{pmatrix}.$$

Let

$$x^{(0)} = (1, 1)^T, \quad x^* = (2, -1)^T.$$

The optimal value of the problem is $f(2, -1) = 0$.

By applying Newton's method, we obtain: **Iterations:**

$k = 0$:

$$x^{(1)} = x^{(0)} - \left(\nabla^2 f(x^{(0)})\right)^{-1} \nabla f(x^{(0)}) = \begin{pmatrix} 1 \\ 0.5 \end{pmatrix}, \quad f(x^{(0)}) = 6.$$

$k = 1$:

$$x^{(2)} = x^{(1)} - \left(\nabla^2 f(x^{(1)})\right)^{-1} \nabla f(x^{(1)}) = \begin{pmatrix} 1 \\ 0.5 \end{pmatrix}, \quad f(x^{(1)}) = 1.5.$$

$k = 2$:

$$x^{(3)} = x^{(2)} - \left(\nabla^2 f(x^{(2)})\right)^{-1} \nabla f(x^{(2)}) = \begin{pmatrix} 1.39 \\ -0.696 \end{pmatrix}, \quad f(x^{(2)}) = 4.09 \times 10^{-1}.$$

$k = 3$:

$$x^{(4)} = x^{(3)} - \left(\nabla^2 f(x^{(3)})\right)^{-1} \nabla f(x^{(3)}) = \begin{pmatrix} 1.746 \\ -0.949 \end{pmatrix}, \quad f(x^{(3)}) = 6.49 \times 10^{-2}.$$

For the case of positive definite quadratic functions, we have the following result:

Theorem 6.1.1. *If f is a positive definite quadratic form, then Newton's method converges to the optimal solution in a single iteration, for any starting point $x^{(0)} \in \mathbb{R}^n$.*

Proof. Let

$$f(x) = \tfrac{1}{2} x^T H x + x^T b + c,$$

with $H = H^T$, $H \succ 0$. Then

$$\nabla f(x) = Hx + b, \quad \nabla^2 f(x) = H.$$

In this case,

$$x^{(1)} = x^{(0)} - H^{-1} \nabla f(x^{(0)}) = x^{(0)} - H^{-1}(Hx^{(0)} + b) = -H^{-1} b.$$

□

Newton's Method: Convergence Analysis

As we have seen previously, in the case of a positive definite quadratic function, Newton's method reaches the minimum in a single iteration. However, for the general case where the function to minimize is not quadratic, there is no guarantee that Newton's method will converge. One of the advantages of this method is that if the starting point is close to the optimal solution $x^\star$, then the method will converge rapidly.

The following theorem shows the local convergence and gives the convergence rate of Newton's method.

Theorem 6.1.2 ([, 6])**.** *Let $f \in C^3$, $x^\star \in \mathbb{R}^n$ such that $\nabla f(x^\star) = 0$ and $H(x^\star)$ is invertible. Then, for any $x^{(0)}$ sufficiently close to $x^\star$, Newton's method converges to $x^\star$ with order $p > 2$.*

Example 6.1.3. Consider the optimization problem

$$\min_{x \in \mathbb{R}^2} (x_1 - 2)^4 + (x_1 - 2x_2)^2. \tag{6.3}$$

It is easy to see that the optimal solution of problem (6.3) is $x^\star = (2, 1)^T$. The results are presented in the following tables with a tolerance $\varepsilon = 10^{-4}$. In the first table, we start from an initial point close to the solution, while in the second table, we start from a point farther away from the solution.

Case 1: Initial point close to the solution

k	$x^{(k)}$	$\|\nabla f(x^{(k)})\|$
0	$(1, 0)^T$	4.4721
1	$(1.3333,\ 0.6667)^T$	1.1852
2	$(1.5556,\ 0.7778)^T$	0.3512
3	$(1.7037,\ 0.8519)^T$	0.1040
4	$(1.8025,\ 0.9012)^T$	0.0308
5	$(1.8683,\ 0.9342)^T$	0.0091
6	$(1.9122,\ 0.9561)^T$	0.0027
7	$(1.9415,\ 0.9707)^T$	0.0008

Case 2: Initial point far from the solution

k	$x^{(k)}$	$\|\nabla f(x^{(k)})\|$
0	$(-10,\ 60)^T$	7190.8264
1	$(-6,\ -3)^T$	2048606.8148
2	$(-3.3333,\ -1.6667)^T$	179.7970
3	$(-1.5556,\ -0.7778)^T$	53.2732
4	$(-0.3704,\ -0.1852)^T$	15.7846
5	$(0.4198,\ 0.2099)^T$	4.6769
6	$(0.9465,\ 0.4733)^T$	1.3858
7	$(1.2977,\ 0.6488)^T$	0.4106
8	$(1.5318,\ 0.7659)^T$	0.1217
9	$(1.6879,\ 0.8439)^T$	0.0360
10	$(1.7919,\ 0.8960)^T$	0.0107
11	$(1.8613,\ 0.9306)^T$	0.0032
12	$(1.9075,\ 0.9538)^T$	0.0009

If the Hessian matrix is positive definite at each point $x^{(k)}$, then Newton's method is a descent method.

Theorem 6.3

Let $(x^{(k)})$ be the sequence generated by Newton's method for the problem

$$\min_{x\in\mathbb{R}^n} f(x).$$

If $\forall k \in \mathbb{N},\ H(x^{(k)}) \succeq 0$ and $\nabla f(x^{(k)}) \neq 0$, then

$$d^{(k)} = -H(x^{(k)})^{-1}\nabla f(x^{(k)}) = x^{(k+1)} - x^{(k)}$$

is a descent direction at $x^{(k)}$.

Proof. Since

$$H(x^{(k)}) \succeq 0,$$

we also have

$$H(x^{(k)})^{-1} \succeq 0.$$

Therefore,

$$\langle d^{(k)}, \nabla f(x^{(k)})\rangle = -\langle H(x^{(k)})^{-1}\nabla f(x^{(k)}), \nabla f(x^{(k)})\rangle \leq 0.$$

Thus, $d^{(k)}$ is indeed a descent direction.

Example 6.1.4 (Example 6.3)**.** Let the function f be defined as

$$f(x) = 2x_1^3 + 3x_1^2 + 12x_1x_2 + 3x_2^2 - 6x_2 + 6.$$

We compute

$$\nabla f(x) = \begin{pmatrix} 6x_1^2 + 6x_1 + 12x_2 \\ 12x_1 + 6x_2 - 6 \end{pmatrix}, \quad \nabla^2 f(x) = \begin{pmatrix} 12x_1 + 6 & 12 \\ 12 & 6 \end{pmatrix}.$$

Applying Newton's method, we obtain the following results:

k	$x^{(k)}$	$\nabla f(x^{(k)})$	$\nabla^2 f(x^{(k)})$	$\left\|\nabla f(x^{(k)})\right\|$
0	$\begin{pmatrix} 1 \\ -4 \end{pmatrix}$	$\begin{pmatrix} -36 \\ -18 \end{pmatrix}$	$\begin{pmatrix} 18 & 12 \\ 12 & 12 \end{pmatrix}$	40.2492
1	$\begin{pmatrix} 4 \\ -5.5 \end{pmatrix}$	$\begin{pmatrix} 54 \\ 9 \end{pmatrix}$	$\begin{pmatrix} 54 & 12 \\ 12 & 12 \end{pmatrix}$	54.7449
2	$\begin{pmatrix} 2.9286 \\ -5.1786 \end{pmatrix}$	$\begin{pmatrix} 6.8878 \\ -1.9286 \end{pmatrix}$	$\begin{pmatrix} 41.1429 & 12 \\ 12 & 12 \end{pmatrix}$	7.1527
3	$\begin{pmatrix} 2.2661 \\ -4.7153 \end{pmatrix}$	$\begin{pmatrix} 0.5491 \\ -2.7794 \end{pmatrix}$	$\begin{pmatrix} 37.5126 & 12 \\ 12 & 12 \end{pmatrix}$	2.8331
4	$\begin{pmatrix} 2.4956 \\ -4.3533 \end{pmatrix}$	$\begin{pmatrix} 0.1021 \\ -2.1725 \end{pmatrix}$	$\begin{pmatrix} 35.947 & 12 \\ 12 & 12 \end{pmatrix}$	2.1749

Advantages of Newton's Method.

1. If the initial point $x^{(0)}$ is close to the solution $x^\star$, then Newton's method converges quadratically to $x^\star$, as stated in Theorem 6.2. In this case, the convergence is said to be *local*.

2. If the optimization problem is quadratic, then Newton's method converges in a single iteration.

Disadvantages of Newton's Method.

1. For many problems, convergence is not guaranteed. In particular, one must choose $x^{(0)}$ sufficiently close to the solution; hence global convergence is not assured.

2. The cost of each iteration is high: one must evaluate N first derivatives and N^2 second derivatives.

Theorem 6.3

Let $(x^{(k)})$ be the sequence generated by Newton's method for the problem

$$\min_{x \in \mathbb{R}^n} f(x).$$

If $\forall k \in \mathbb{N},\ H(x^{(k)}) \succeq 0$ and $\nabla f(x^{(k)}) \neq 0$, then

$$d^{(k)} = -H(x^{(k)})^{-1} \nabla f(x^{(k)}) = x^{(k+1)} - x^{(k)}$$

is a descent direction at $x^{(k)}$.

Proof. Since

$$H(x^{(k)}) \succeq 0,$$

we also have

$$H(x^{(k)})^{-1} \succeq 0.$$

Therefore,

$$\langle d^{(k)}, \nabla f(x^{(k)})\rangle = -\langle H(x^{(k)})^{-1}\nabla f(x^{(k)}), \nabla f(x^{(k)})\rangle \leq 0.$$

Thus, $d^{(k)}$ is indeed a descent direction.

Example 6.1.5 (Example 6.3)**.** Let the function f be defined as

$$f(x) = 2x_1^3 + 3x_1^2 + 12x_1x_2 + 3x_2^2 - 6x_2 + 6.$$

We compute

$$\nabla f(x) = \begin{pmatrix} 6x_1^2 + 6x_1 + 12x_2 \\ 12x_1 + 6x_2 - 6 \end{pmatrix}, \quad \nabla^2 f(x) = \begin{pmatrix} 12x_1 + 6 & 12 \\ 12 & 6 \end{pmatrix}.$$

Applying Newton's method, we obtain the following results:

k	$x^{(k)}$	$\nabla f(x^{(k)})$	$\nabla^2 f(x^{(k)})$	$\left\|\nabla f(x^{(k)})\right\|$
0	$\begin{pmatrix} 1 \\ -4 \end{pmatrix}$	$\begin{pmatrix} -36 \\ -18 \end{pmatrix}$	$\begin{pmatrix} 18 & 12 \\ 12 & 12 \end{pmatrix}$	40.2492
1	$\begin{pmatrix} 4 \\ -5.5 \end{pmatrix}$	$\begin{pmatrix} 54 \\ 9 \end{pmatrix}$	$\begin{pmatrix} 54 & 12 \\ 12 & 12 \end{pmatrix}$	54.7449
2	$\begin{pmatrix} 2.9286 \\ -5.1786 \end{pmatrix}$	$\begin{pmatrix} 6.8878 \\ -1.9286 \end{pmatrix}$	$\begin{pmatrix} 41.1429 & 12 \\ 12 & 12 \end{pmatrix}$	7.1527
3	$\begin{pmatrix} 2.2661 \\ -4.7153 \end{pmatrix}$	$\begin{pmatrix} 0.5491 \\ -2.7794 \end{pmatrix}$	$\begin{pmatrix} 37.5126 & 12 \\ 12 & 12 \end{pmatrix}$	2.8331
4	$\begin{pmatrix} 2.4956 \\ -4.3533 \end{pmatrix}$	$\begin{pmatrix} 0.1021 \\ -2.1725 \end{pmatrix}$	$\begin{pmatrix} 35.947 & 12 \\ 12 & 12 \end{pmatrix}$	2.1749

Advantages of Newton's Method.

1. If the initial point $x^{(0)}$ is close to the solution $x^\star$, then Newton's method converges quadratically to $x^\star$, as stated in Theorem 6.2. In this case, the convergence is said to be *local*.

2. If the optimization problem is quadratic, then Newton's method converges in a single iteration.

Disadvantages of Newton's Method.

1. For many problems, convergence is not guaranteed. In particular, one must choose $x^{(0)}$ sufficiently close to the solution; hence global convergence is not assured.

2. The cost of each iteration is high: one must evaluate N first derivatives and N^2 second derivatives.

Theorem 6.3

Let $(x^{(k)})$ be the sequence generated by Newton's method for the problem

$$\min_{x\in\mathbb{R}^n} f(x).$$

If $\forall k \in \mathbb{N},\ H(x^{(k)}) \succeq 0$ and $\nabla f(x^{(k)}) \neq 0$, then

$$d^{(k)} = -H(x^{(k)})^{-1}\nabla f(x^{(k)}) = x^{(k+1)} - x^{(k)}$$

is a descent direction at $x^{(k)}$.

Proof. Since

$$H(x^{(k)}) \succeq 0,$$

we also have

$$H(x^{(k)})^{-1} \succeq 0.$$

Therefore,

$$\langle d^{(k)}, \nabla f(x^{(k)})\rangle = -\langle H(x^{(k)})^{-1}\nabla f(x^{(k)}), \nabla f(x^{(k)})\rangle \leq 0.$$

Thus, $d^{(k)}$ is indeed a descent direction.

Example 6.1.6 (Example 6.3)**.** Let the function f be defined as

$$f(x) = 2x_1^3 + 3x_1^2 + 12x_1x_2 + 3x_2^2 - 6x_2 + 6.$$

We compute

$$\nabla f(x) = \begin{pmatrix} 6x_1^2 + 6x_1 + 12x_2 \\ 12x_1 + 6x_2 - 6 \end{pmatrix}, \quad \nabla^2 f(x) = \begin{pmatrix} 12x_1 + 6 & 12 \\ 12 & 6 \end{pmatrix}.$$

Applying Newton's method, we obtain the following results:

k	$x^{(k)}$	$\nabla f(x^{(k)})$	$\nabla^2 f(x^{(k)})$	$\left\|\nabla f(x^{(k)})\right\|$
0	$\begin{pmatrix} 1 \\ -4 \end{pmatrix}$	$\begin{pmatrix} -36 \\ -18 \end{pmatrix}$	$\begin{pmatrix} 18 & 12 \\ 12 & 12 \end{pmatrix}$	40.2492
1	$\begin{pmatrix} 4 \\ -5.5 \end{pmatrix}$	$\begin{pmatrix} 54 \\ 9 \end{pmatrix}$	$\begin{pmatrix} 54 & 12 \\ 12 & 12 \end{pmatrix}$	54.7449
2	$\begin{pmatrix} 2.9286 \\ -5.1786 \end{pmatrix}$	$\begin{pmatrix} 6.8878 \\ -1.9286 \end{pmatrix}$	$\begin{pmatrix} 41.1429 & 12 \\ 12 & 12 \end{pmatrix}$	7.1527
3	$\begin{pmatrix} 2.2661 \\ -4.7153 \end{pmatrix}$	$\begin{pmatrix} 0.5491 \\ -2.7794 \end{pmatrix}$	$\begin{pmatrix} 37.5126 & 12 \\ 12 & 12 \end{pmatrix}$	2.8331
4	$\begin{pmatrix} 2.4956 \\ -4.3533 \end{pmatrix}$	$\begin{pmatrix} 0.1021 \\ -2.1725 \end{pmatrix}$	$\begin{pmatrix} 35.947 & 12 \\ 12 & 12 \end{pmatrix}$	2.1749

Advantages of Newton's Method.

1. If the initial point $x^{(0)}$ is close to the solution $x^\star$, then Newton's method converges quadratically to $x^\star$, as stated in Theorem 6.2. In this case, the convergence is said to be *local*.

2. If the optimization problem is quadratic, then Newton's method converges in a single iteration.

Disadvantages of Newton's Method.

1. For many problems, convergence is not guaranteed. In particular, one must choose $x^{(0)}$ sufficiently close to the solution; hence global convergence is not assured.
2. The cost of each iteration is high: one must evaluate N first derivatives and N^2 second derivatives.

Theorem 6.3

Let $(x^{(k)})$ be the sequence generated by Newton's method for the problem

$$\min_{x\in\mathbb{R}^n} f(x).$$

If $\forall k \in \mathbb{N},\ H(x^{(k)}) \succeq 0$ and $\nabla f(x^{(k)}) \neq 0$, then

$$d^{(k)} = -H(x^{(k)})^{-1}\nabla f(x^{(k)}) = x^{(k+1)} - x^{(k)}$$

is a descent direction at $x^{(k)}$.

Proof. Since

$$H(x^{(k)}) \succeq 0,$$

we also have

$$H(x^{(k)})^{-1} \succeq 0.$$

Therefore,

$$\langle d^{(k)}, \nabla f(x^{(k)})\rangle = -\langle H(x^{(k)})^{-1}\nabla f(x^{(k)}), \nabla f(x^{(k)})\rangle \leq 0.$$

Thus, $d^{(k)}$ is indeed a descent direction.

Example 6.1.7 (Example 6.3)**.** Let the function f be defined as

$$f(x) = 2x_1^3 + 3x_1^2 + 12x_1x_2 + 3x_2^2 - 6x_2 + 6.$$

We compute

$$\nabla f(x) = \begin{pmatrix} 6x_1^2 + 6x_1 + 12x_2 \\ 12x_1 + 6x_2 - 6 \end{pmatrix}, \quad \nabla^2 f(x) = \begin{pmatrix} 12x_1 + 6 & 12 \\ 12 & 6 \end{pmatrix}.$$

Applying Newton's method, we obtain the following results:

k	$x^{(k)}$	$\nabla f(x^{(k)})$	$\nabla^2 f(x^{(k)})$	$\left\|\nabla f(x^{(k)})\right\|$
0	$\begin{pmatrix} 1 \\ -4 \end{pmatrix}$	$\begin{pmatrix} -36 \\ -18 \end{pmatrix}$	$\begin{pmatrix} 18 & 12 \\ 12 & 12 \end{pmatrix}$	40.2492
1	$\begin{pmatrix} 4 \\ -5.5 \end{pmatrix}$	$\begin{pmatrix} 54 \\ 9 \end{pmatrix}$	$\begin{pmatrix} 54 & 12 \\ 12 & 12 \end{pmatrix}$	54.7449
2	$\begin{pmatrix} 2.9286 \\ -5.1786 \end{pmatrix}$	$\begin{pmatrix} 6.8878 \\ -1.9286 \end{pmatrix}$	$\begin{pmatrix} 41.1429 & 12 \\ 12 & 12 \end{pmatrix}$	7.1527
3	$\begin{pmatrix} 2.2661 \\ -4.7153 \end{pmatrix}$	$\begin{pmatrix} 0.5491 \\ -2.7794 \end{pmatrix}$	$\begin{pmatrix} 37.5126 & 12 \\ 12 & 12 \end{pmatrix}$	2.8331
4	$\begin{pmatrix} 2.4956 \\ -4.3533 \end{pmatrix}$	$\begin{pmatrix} 0.1021 \\ -2.1725 \end{pmatrix}$	$\begin{pmatrix} 35.947 & 12 \\ 12 & 12 \end{pmatrix}$	2.1749

Advantages of Newton's Method.

1. If the initial point $x^{(0)}$ is close to the solution $x^\star$, then Newton's method converges quadratically to $x^\star$, as stated in Theorem 6.2. In this case, the convergence is said to be *local*.

2. If the optimization problem is quadratic, then Newton's method converges in a single iteration.

Disadvantages of Newton's Method.

1. For many problems, convergence is not guaranteed. In particular, one must choose $x^{(0)}$ sufficiently close to the solution; hence global convergence is not assured.

2. The cost of each iteration is high: one must evaluate N first derivatives and N^2 second derivatives.

Bibliography

[1] Andreas, A. and Lu, W.-S. L. (2007), *Practical Optimization: Algorithms and Engineering Applications*, Springer, New York.

[2] Avez, A. (1983), *Calcul diffrentiel*, Masson, Paris New York Barcelone Milan Mexico Sao Paulo.

[3] Berkovitz, L.D. (2002), *Convexity and Optimization in* $\mathbb{R}^n$, John Wiley & Sons, Inc.

[4] Biegler, L.T. (2010), *Nonlinear Programming: Concepts, Algorithms and Applications to Chemical Processes*, SIAM, Philadelphia.

[5] Bierlaire, M. (2006), *Introduction loptimisation diffrentiable*, premire dition, Presses Polytechniques et Universitaires Romandes.

[6] Chong, E.K.P. and Zak, S.H. (2013), *An Introduction to Optimization*, 4th Edition, John Wiley & Sons, Inc., Hoboken, New Jersey.

[7] Griva, I., Nash, S.G. and Sofer, A. (2009), *Linear and Nonlinear Optimization*, 2nd Edition, Society for Industrial and Applied Mathematics, Philadelphia.

[8] Luenberger, D.G. and Ye, Y. (2008), *Linear and Nonlinear Programming*, 4th Edition, Springer.

[9] Sundaram, R.K. (1996), *A First Course in Optimization Theory*, Cambridge University Press, New York.

[10] Sun, W. and Yuan, Y.-X. (2006), *Optimization Theory and Methods: Nonlinear Programming*, Springer-Verlag.

[11] Yang, X.S. (2018), *Optimization Techniques and Applications with Examples*, Wiley.