Chong, Edwin Kah Pin., and Stanislaw H. Żak. *An Introduction to Optimization*. Fourth edition., John Wiley & Sons, Inc., 2013.

3 GRADIENT METHODS

3.1 INTRODUCTION

- $\langle \nabla f(\boldsymbol{x}), \boldsymbol{d} \rangle$ with $\|\boldsymbol{d}\| = 1$, is the rate of increase of f in the direction $\boldsymbol{d}$ at the point $\boldsymbol{x}$. By the Cauchy-Schwarz inequality,

$$\langle \nabla f(\boldsymbol{x}), \boldsymbol{d} \rangle \leq \|\nabla f(\boldsymbol{x})\| \|\boldsymbol{d}\| = \|\nabla f(\boldsymbol{x})\|$$

because $\|\boldsymbol{d}\| = 1$.

- The direction of maximum rate of **increase** of f at $\boldsymbol{x}$ is $\boldsymbol{d} = \nabla f(\boldsymbol{x}) / \|\nabla f(\boldsymbol{x})\|$ because

$$\langle \nabla f(\boldsymbol{x}), \frac{\nabla f(\boldsymbol{x})}{\|\nabla f(\boldsymbol{x})\|} \rangle = \frac{\left(\nabla f(\boldsymbol{x})\right)^{\top} \nabla f(\boldsymbol{x})}{\|\nabla f(\boldsymbol{x})\|} = \frac{\|\nabla f(\boldsymbol{x})\|^2}{\|\nabla f(\boldsymbol{x})\|} = \|\nabla f(\boldsymbol{x})\|.$$

- The direction of **maximum rate of decrease** is $-\nabla f(\boldsymbol{x}) / \|\nabla f(\boldsymbol{x})\|$.

- Let $\boldsymbol{x}^{(0)}$ be a starting point and consider the point $\boldsymbol{x}^{(0)} - \alpha \nabla f\left(\boldsymbol{x}^{(0)}\right)$. The first-order Taylor expansion of a function f around a point $\boldsymbol{x}^{(0)}$ is given by

$$f(\boldsymbol{x}) = f\left(\boldsymbol{x}^{(0)}\right) + \nabla f\left(\boldsymbol{x}^{(0)}\right)^{\top} \left(\boldsymbol{x} - \boldsymbol{x}^{(0)}\right) + o\left(\boldsymbol{x} - \boldsymbol{x}^{(0)}\right)$$

- Then, we have

$$\begin{aligned}
f\left(\boldsymbol{x}^{(0)} - \alpha \nabla f\left(\boldsymbol{x}^{(0)}\right)\right) & \\
&= f\left(\boldsymbol{x}^{(0)}\right) + \nabla f\left(\boldsymbol{x}^{(0)}\right)^{\top} \left(\boldsymbol{x}^{(0)} - \alpha \nabla f\left(\boldsymbol{x}^{(0)}\right) - \boldsymbol{x}^{(0)}\right) \\
&\quad + o\left(\boldsymbol{x}^{(0)} - \alpha \nabla f\left(\boldsymbol{x}^{(0)}\right) - \boldsymbol{x}^{(0)}\right) \\
&= f\left(\boldsymbol{x}^{(0)}\right) - \alpha \nabla f\left(\boldsymbol{x}^{(0)}\right)^{\top} \nabla f\left(\boldsymbol{x}^{(0)}\right) + o\left(-\alpha \nabla f\left(\boldsymbol{x}^{(0)}\right)\right) \\
&= f\left(\boldsymbol{x}^{(0)}\right) - \alpha \left\|\nabla f\left(\boldsymbol{x}^{(0)}\right)\right\|^2 + o(\alpha).
\end{aligned}$$

- If $\nabla f\left(\boldsymbol{x}^{(0)}\right) \neq 0$, then for sufficiently small $\alpha > 0$, we have

$$f\left(\boldsymbol{x}^{(0)} - \alpha \nabla f\left(\boldsymbol{x}^{(0)}\right)\right) < f\left(\boldsymbol{x}^{(0)}\right).$$

- This means that the point $\boldsymbol{x}^{(0)} - \alpha \nabla f\left(\boldsymbol{x}^{(0)}\right)$ is an improvement over the point $\boldsymbol{x}^{(0)}$ if we are searching for a minimizer.

- Suppose that we are given a point $\boldsymbol{x}^{(k)}$. To find the next point $\boldsymbol{x}^{(k+1)}$ we start at $\boldsymbol{x}^{(k)}$ and move by an amount $-\alpha_k \nabla f\left(\boldsymbol{x}^{(k)}\right)$, where α_k is a positive scalar called the **step size**. This procedure leads to the gradient descent algorithm:

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha_k \nabla f\left(\boldsymbol{x}^{(k)}\right).$$

- $\left\{\boldsymbol{x}^{(k)}\right\}_{k=0}^{\infty}$ is a **steepest descent sequence** if $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha_k \nabla f\left(\boldsymbol{x}^{(k)}\right)$ where $\alpha_k = \arg\min_{\alpha \geq 0} f\left(\boldsymbol{x}^{(k)} - \alpha \nabla f\left(\boldsymbol{x}^{(k)}\right)\right)$.

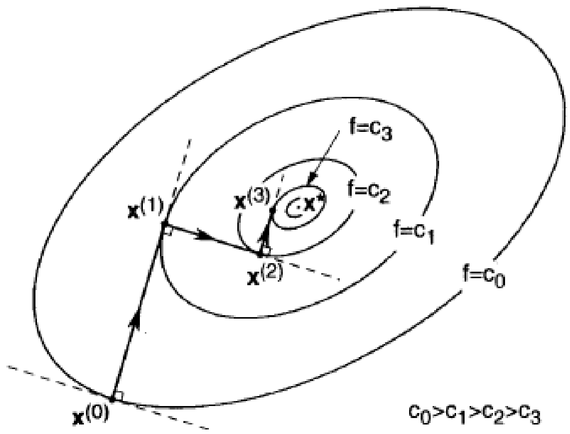
Chong, Edwin Kah Pin., and Stanislaw H. Żak. *An Introduction to Optimization*. Fourth edition., John Wiley & Sons, Inc., 2013.

3.2 THE METHOD OF STEEPEST DESCENT

- The method of steepest descent is a gradient algorithm where the step size α_k is chosen to achieve the **maximum amount of decrease** of the objective function at each individual step. Specifically, α_k is chosen to minimize $\phi_k(\alpha) \triangleq f\left(\boldsymbol{x}^{(k)} - \alpha\nabla f\left(\boldsymbol{x}^{(k)}\right)\right)$. In other words,

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

- **Proposition 3.2.1** If $\left\{\boldsymbol{x}^{(k)}\right\}_{k=0}^{\infty}$ is a steepest descent sequence for a given function $f\colon \mathbb{R}^n \to \mathbb{R}$, then for each k the vector $\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}$ is orthogonal to the vector $\boldsymbol{x}^{(k+2)} - \boldsymbol{x}^{(k+1)}$.



- **Proposition 3.2.2** If $\left\{\boldsymbol{x}^{(k)}\right\}_{k=0}^{\infty}$ is a steepest descent sequence for $f\colon \mathbb{R}^n \to \mathbb{R}$ and if $\nabla f\left(\boldsymbol{x}^{(k)}\right) \neq \boldsymbol{0}$, then $f\left(\boldsymbol{x}^{(k+1)}\right) < f\left(\boldsymbol{x}^{(k)}\right)$.

- If for some k, we have $\nabla f\left(\boldsymbol{x}^{(k)}\right) = \boldsymbol{0}$, then the point $\boldsymbol{x}^{(k)}$ satisfies the FONC. The condition $\nabla f\left(\boldsymbol{x}^{(k)}\right) = \boldsymbol{0}$, however, is not directly suitable as a practical stopping criterion, because the numerical computation of the gradient will rarely be identically equal to zero.

- A practical stopping criterion is to check if, given a prespecified threshold $\varepsilon > 0$,

    1. $\left\|\nabla f\left(\boldsymbol{x}^{(k)}\right)\right\| < \varepsilon$;
    2. $\left|f\left(\boldsymbol{x}^{(k+1)}\right) - f\left(\boldsymbol{x}^{(k)}\right)\right| < \varepsilon$;
    3. $\left\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\right\| < \varepsilon$;
    4. $\left|f\left(\boldsymbol{x}^{(k+1)}\right) - f\left(\boldsymbol{x}^{(k)}\right)\right| / \left|f\left(\boldsymbol{x}^{(k)}\right)\right| < \varepsilon$;
    5. $\left\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\right\| / \left\|\boldsymbol{x}^{(k)}\right\| < \varepsilon$.

- The criteria 4 and 5 above are preferable to the criteria 1, 2 and 3 because the relative criteria are "**scale-independent**." To avoid dividing by very small numbers in criteria 4 and 5, we can modify these stopping criteria as follows:

$$\frac{\left|f\left(\boldsymbol{x}^{(k+1)}\right) - f\left(\boldsymbol{x}^{(k)}\right)\right|}{\max\left\{1, \left|f\left(\boldsymbol{x}^{(k)}\right)\right|\right\}} < \varepsilon \quad \text{or} \quad \frac{\left\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\right\|}{\max\left\{1, \left\|\boldsymbol{x}^{(k)}\right\|\right\}} < \varepsilon.$$

Example 3.2.1. Perform three iterations of the method of steepest descent to find the minimizer of $f(x_1,x_2,x_3)=(x_1-4)^4+(x_2-3)^2+4(x_3+5)^4$. The initial point is $\boldsymbol{x}^{(0)}=[4,2,-1]^\top$.

Note that

$$\nabla f(\boldsymbol{x})=[4(x_1-4)^3,2(x_2-3),16(x_3+5)^3]^\top$$

$$\boldsymbol{x}-\alpha\nabla f(\boldsymbol{x})=[x_1-4\alpha(x_1-4)^3,x_2-2\alpha(x_2-3),x_3-16\alpha(x_3+5)^3]^\top$$

$$\begin{aligned}f\big(\boldsymbol{x}-\alpha\nabla f(\boldsymbol{x})\big)=(x_1-4\alpha(x_1-4)^3-4)^4+(x_2-2\alpha(x_2-3)-3)^2\\+4(x_3-16\alpha(x_3+5)^3+5)^4\end{aligned}$$

and $f\big(\boldsymbol{x}^{(0)}\big)=1025$.

Iteration 1:

To compute $\boldsymbol{x}^{(1)}$, we need

$$\alpha_0=\operatorname*{argmin}_{\alpha\ge 0} f\left(\boldsymbol{x}^{(0)}-\alpha\nabla f\big(\boldsymbol{x}^{(0)}\big)\right)$$

Let $\boldsymbol{h}(\alpha)=\boldsymbol{x}^{(0)}-\alpha\nabla f\big(\boldsymbol{x}^{(0)}\big)$ and $g(\alpha)=f\big(\boldsymbol{h}(\alpha)\big)$. Using the secant method, we obtain $\alpha_0=3.967\times10^{-3}$:

$$\alpha^{(k+1)}=\alpha^{(k)}-\frac{\alpha^{(k)}-\alpha^{(k-1)}}{g'(\alpha^{(k)})-g'(\alpha^{(k-1)})}g'\big(\alpha^{(k)}\big)$$

where $g'(\alpha)=\Big(\nabla f\big(\boldsymbol{h}(\alpha)\big)\Big)^\top\begin{bmatrix}h_1'(\alpha)\\ \vdots\\ h_1'(\alpha)\end{bmatrix}=-\left(\nabla f\left(\boldsymbol{x}^{(0)}-\alpha\nabla f\big(\boldsymbol{x}^{(0)}\big)\right)\right)^\top\nabla f\big(\boldsymbol{x}^{(0)}\big)$. Thus,

$$\begin{aligned}\boldsymbol{x}^{(1)}&=\boldsymbol{x}^{(0)}-\alpha_0\nabla f\big(\boldsymbol{x}^{(0)}\big)\\&=[4,2,-1]^\top-3.967\times10^{-3}[4(4-4)^3,2(2-3),16(-1+5)^3]^\top\\&=[4,2,-1]^\top-3.967\times10^{-3}[0,-2,1024]^\top\\&\approx[4.000,2.007934,-5.061568]^\top\\&\approx[4.000,2.008,-5.062]^\top,\end{aligned}$$

and $f\big(\boldsymbol{x}^{(1)}\big)\approx0.984$.

Iteration 2:

We find $\alpha_1\approx0.5000$, where $\alpha_1=\operatorname*{argmin}_{\alpha\ge0} f\left(\boldsymbol{x}^{(1)}-\alpha\nabla f\big(\boldsymbol{x}^{(1)}\big)\right)$. Thus,

$$\begin{aligned}\boldsymbol{x}^{(2)}&=\boldsymbol{x}^{(1)}-\alpha_1\nabla f\big(\boldsymbol{x}^{(1)}\big)\\&=[4.000,2.008,-5.062]^\top-0.500[4(4-4)^3,2(2.008-3),16(-5.062+5)^3]^\top\\&\approx[4.000,2.008,-5.062]^\top-0.500[0,-1.9841,-0.00877]^\top\\&\approx[4.000,3.000,-5.060]^\top,\end{aligned}$$

and $f\big(\boldsymbol{x}^{(2)}\big)=5.326\times10^{-5}$.

Iteration 3:

We find $\alpha_2 \approx 16.28$, where $\alpha_2 = \underset{\alpha \geq 0}{\operatorname{argmin}} f\left(\boldsymbol{x}^{(2)} - \alpha \nabla f\left(\boldsymbol{x}^{(2)}\right)\right)$. Thus,

$$\begin{aligned}\boldsymbol{x}^{(3)} &= \boldsymbol{x}^{(2)} - \alpha_2 \nabla f\left(\boldsymbol{x}^{(2)}\right) \\ &= [4.000, 3.000, -5.060]^\top - 16.28[4(4-4)^3, 2(3-3), 16(-5.060+5)^3]^\top \\ &\approx [4.000, 3.000, -5.060]^\top - 16.28[0.000, 0.000, 0.004]^\top \\ &\approx [4.000, 3.000, -5.003]^\top,\end{aligned}$$

and $f\left(\boldsymbol{x}^{(3)}\right) = 1.215 \times 10^{-8}$.

Example 3.2.2. Consider an objection function which is in quadratic form

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} - \boldsymbol{b}^\top \boldsymbol{x},$$

where $\boldsymbol{Q} \in \mathbb{R}^{n \times n}$ is a symmetric positive definite matrix, $\boldsymbol{b} \in \mathbb{R}^n$, and $\boldsymbol{x} \in \mathbb{R}^n$. Show that the method of steepest descent takes the form

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \frac{\boldsymbol{g}^{(k)\top}\boldsymbol{g}^{(k)}}{\boldsymbol{g}^{(k)\top}\boldsymbol{Q}\boldsymbol{g}^{(k)}}\boldsymbol{g}^{(k)},$$

where $\boldsymbol{g}^{(k)} = \nabla f\left(\boldsymbol{x}^{(k)}\right) = \boldsymbol{Q}\boldsymbol{x}^{(k)} - \boldsymbol{b}$.

There is no loss of generality in assuming $\boldsymbol{Q}$ to be a symmetric matrix. For if we are given a quadratic form $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$ and $\boldsymbol{A} \neq \boldsymbol{A}^\top$, then because the transposition of a scalar equals itself, we obtain

$$\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} = (\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x})^\top = \boldsymbol{x}^\top \boldsymbol{A}^\top \boldsymbol{x}.$$

Hence,

$$\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} + \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{A}^\top \boldsymbol{x} = \frac{1}{2}\boldsymbol{x}^\top (\boldsymbol{A} + \boldsymbol{A}^\top)\boldsymbol{x} \triangleq \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x}.$$

Note that $(\boldsymbol{A} + \boldsymbol{A}^\top)^\top = \boldsymbol{Q}^\top = (\boldsymbol{A} + \boldsymbol{A}^\top) = \boldsymbol{Q}$.

The unique minimizer of f can be found by setting the gradient of f to zero:

$$\nabla f(\boldsymbol{x}) = \frac{1}{2}(\boldsymbol{Q} + \boldsymbol{Q}^\top)^\top \boldsymbol{x} - \boldsymbol{b} = \boldsymbol{Q}\boldsymbol{x} - \boldsymbol{b} = \boldsymbol{0}.$$

The Hessian of f is $\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{Q} = \boldsymbol{Q}^\top > 0$. Let $\boldsymbol{g}^{(k)} = \nabla f\left(\boldsymbol{x}^{(k)}\right)$. The steepest descent algorithm for the quadratic function is

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha_k \boldsymbol{g}^{(k)},$$

where

$$\begin{aligned}\alpha_k &= \arg\min_{\alpha \geq 0} f\left(\boldsymbol{x}^{(k)} - \alpha \boldsymbol{g}^{(k)}\right) \\ &= \arg\min_{\alpha \geq 0} \left(\frac{1}{2}\left(\boldsymbol{x}^{(k)} - \alpha \boldsymbol{g}^{(k)}\right)^\top \boldsymbol{Q}\left(\boldsymbol{x}^{(k)} - \alpha \boldsymbol{g}^{(k)}\right) - \boldsymbol{b}^\top\left(\boldsymbol{x}^{(k)} - \alpha \boldsymbol{g}^{(k)}\right)\right).\end{aligned}$$

Because $\alpha_k \geq 0$ is a minimizer of $\phi_k(\alpha) = f\left(\boldsymbol{x}^{(k)} - \alpha \boldsymbol{g}^{(k)}\right)$, we apply the FONC to $\phi_k(\alpha)$ to obtain

$$\begin{aligned}
0 &= \phi_k'(\alpha) \\
0 &= \left(\boldsymbol{x}^{(k)} - \alpha \boldsymbol{g}^{(k)}\right)^\top \boldsymbol{Q}\left(-\boldsymbol{g}^{(k)}\right) - \boldsymbol{b}^\top\left(-\boldsymbol{g}^{(k)}\right) \\
0 &= -\boldsymbol{x}^{(k)\top}\boldsymbol{Q}\boldsymbol{g}^{(k)} + \alpha \boldsymbol{g}^{(k)\top}\boldsymbol{Q}\boldsymbol{g}^{(k)} + \boldsymbol{b}^\top \boldsymbol{g}^{(k)} \\
\boldsymbol{x}^{(k)\top}\boldsymbol{Q}\boldsymbol{g}^{(k)} - \boldsymbol{b}^\top \boldsymbol{g}^{(k)} &= \alpha \boldsymbol{g}^{(k)\top}\boldsymbol{Q}\boldsymbol{g}^{(k)} \\
\left(\boldsymbol{x}^{(k)\top}\boldsymbol{Q} - \boldsymbol{b}^\top\right)\boldsymbol{g}^{(k)} &= \alpha \boldsymbol{g}^{(k)\top}\boldsymbol{Q}\boldsymbol{g}^{(k)} \\
\boldsymbol{g}^{(k)\top}\boldsymbol{g}^{(k)} &= \alpha \boldsymbol{g}^{(k)\top}\boldsymbol{Q}\boldsymbol{g}^{(k)} \\
\alpha &= \frac{\boldsymbol{g}^{(k)\top}\boldsymbol{g}^{(k)}}{\boldsymbol{g}^{(k)\top}\boldsymbol{Q}\boldsymbol{g}^{(k)}}.
\end{aligned}$$

Therefore, the steepest descent algorithm for the quadratic function is

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \frac{\boldsymbol{g}^{(k)\top}\boldsymbol{g}^{(k)}}{\boldsymbol{g}^{(k)\top}\boldsymbol{Q}\boldsymbol{g}^{(k)}}\boldsymbol{g}^{(k)},$$

where $\boldsymbol{g}^{(k)} = \nabla f\left(\boldsymbol{x}^{(k)}\right) = \boldsymbol{Q}\boldsymbol{x}^{(k)} - \boldsymbol{b}$.

Example 3.2.3. State the steepest descent algorithms for each of the following objection functions:

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

(b) $f(x_1, x_2) = x_1^2/5 + x_2^2$.

(a) $f(x_1, x_2) = x_1^2 + x_2^2 = \frac{1}{2}\boldsymbol{x}^\top \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \boldsymbol{x}$

The steepest descent algorithm for the function is

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \frac{\boldsymbol{g}^{(k)\top}\boldsymbol{g}^{(k)}}{\boldsymbol{g}^{(k)\top}\boldsymbol{Q}\boldsymbol{g}^{(k)}}\boldsymbol{g}^{(k)},$$

where $\boldsymbol{g}^{(k)} = \nabla f\left(\boldsymbol{x}^{(k)}\right) = \boldsymbol{Q}\boldsymbol{x}^{(k)}$ and $\boldsymbol{Q} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$.

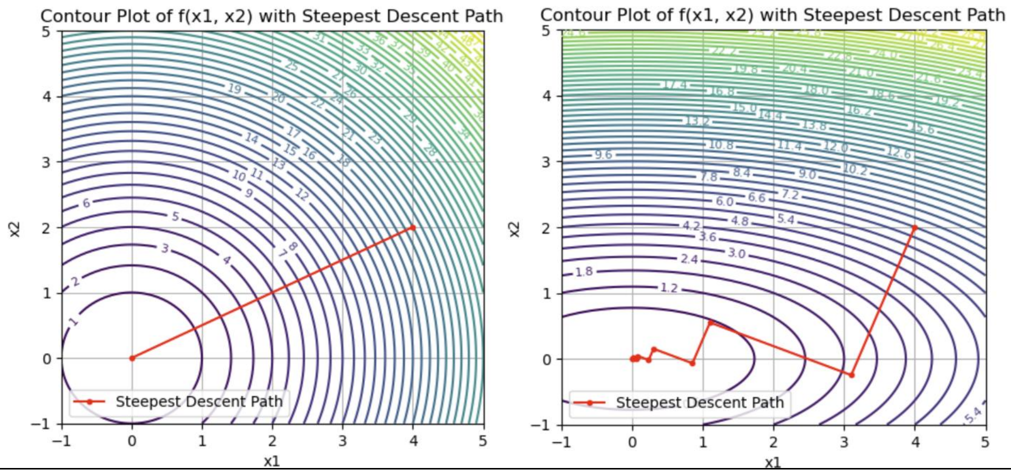
(b) $f(x_1, x_2) = \frac{x_1^2}{5} + x_2^2 = \frac{1}{2}\boldsymbol{x}^\top \begin{bmatrix} 2/5 & 0 \\ 0 & 2 \end{bmatrix} \boldsymbol{x}$

The steepest descent algorithm for the function is

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \frac{\boldsymbol{g}^{(k)\top}\boldsymbol{g}^{(k)}}{\boldsymbol{g}^{(k)\top}\boldsymbol{Q}\boldsymbol{g}^{(k)}}\boldsymbol{g}^{(k)},$$

where $\boldsymbol{g}^{(k)} = \nabla f\left(\boldsymbol{x}^{(k)}\right) = \boldsymbol{Q}\boldsymbol{x}^{(k)}$ and $\boldsymbol{Q} = \begin{bmatrix} 2/5 & 0 \\ 0 & 2 \end{bmatrix}$.

Chong, Edwin Kah Pin., and Stanislaw H. Żak. *An Introduction to Optimization*. Fourth edition., John Wiley & Sons, Inc., 2013.



3.3 THE CONDITION NUMBER

- Consider the quadratic minimization problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$$

where $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is a symmetric positive definite matrix. The optimal solution is obviously $\boldsymbol{x}^* = \boldsymbol{0}$. The steepest descent algorithm is

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha_k \boldsymbol{g}^{(k)},$$

where $\boldsymbol{g}^{(k)} = 2\boldsymbol{A}\boldsymbol{x}^{(k)}$ is the gradient of $f(\boldsymbol{x}) = \boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x}$ at $\boldsymbol{x}^{(k)}$ and the stepsize α_k is chosen by the minimization rule

$$\alpha_k = \frac{\boldsymbol{g}^{(k)\top}\boldsymbol{g}^{(k)}}{2\boldsymbol{g}^{(k)\top}\boldsymbol{A}\boldsymbol{g}^{(k)}}.$$

- Therefore,

$$\begin{aligned}
f\left(\boldsymbol{x}^{(k+1)}\right) &= \boldsymbol{x}^{(k+1)\top}\boldsymbol{A}\boldsymbol{x}^{(k+1)} = \left(\boldsymbol{x}^{(k)} - \alpha_k\boldsymbol{g}^{(k)}\right)^\top \boldsymbol{A}\left(\boldsymbol{x}^{(k)} - \alpha_k\boldsymbol{g}^{(k)}\right)\\
&= \boldsymbol{x}^{(k)\top}\boldsymbol{A}\boldsymbol{x}^{(k)} - 2\alpha_k\boldsymbol{g}^{(k)\top}\boldsymbol{A}\boldsymbol{x}^{(k)} + t_k^2\boldsymbol{g}^{(k)\top}\boldsymbol{A}\boldsymbol{g}^{(k)}\\
&= \boldsymbol{x}^{(k)\top}\boldsymbol{A}\boldsymbol{x}^{(k)} - \alpha_k\boldsymbol{g}^{(k)\top}\boldsymbol{g}^{(k)} + t_k^2\boldsymbol{g}^{(k)\top}\boldsymbol{A}\boldsymbol{g}^{(k)}\\
&= \boldsymbol{x}^{(k)\top}\boldsymbol{A}\boldsymbol{x}^{(k)} - \left(\frac{\boldsymbol{g}^{(k)\top}\boldsymbol{g}^{(k)}}{2\boldsymbol{g}^{(k)\top}\boldsymbol{A}\boldsymbol{g}^{(k)}}\right)\boldsymbol{g}^{(k)\top}\boldsymbol{g}^{(k)} + \left(\frac{\boldsymbol{g}^{(k)\top}\boldsymbol{g}^{(k)}}{2\boldsymbol{g}^{(k)\top}\boldsymbol{A}\boldsymbol{g}^{(k)}}\right)^2 \boldsymbol{g}^{(k)\top}\boldsymbol{A}\boldsymbol{g}^{(k)}\\
&= \boldsymbol{x}^{(k)\top}\boldsymbol{A}\boldsymbol{x}^{(k)} - \frac{1}{2}\frac{\left(\boldsymbol{g}^{(k)\top}\boldsymbol{g}^{(k)}\right)^2}{\boldsymbol{g}^{(k)\top}\boldsymbol{A}\boldsymbol{g}^{(k)}} + \frac{1}{4}\frac{\left(\boldsymbol{g}^{(k)\top}\boldsymbol{g}^{(k)}\right)^2}{\boldsymbol{g}^{(k)\top}\boldsymbol{A}\boldsymbol{g}^{(k)}}\\
&= \boldsymbol{x}^{(k)\top}\boldsymbol{A}\boldsymbol{x}^{(k)} - \frac{1}{4}\frac{\left(\boldsymbol{g}^{(k)\top}\boldsymbol{g}^{(k)}\right)^2}{\boldsymbol{g}^{(k)\top}\boldsymbol{A}\boldsymbol{g}^{(k)}}\frac{\boldsymbol{x}^{(k)\top}\boldsymbol{A}\boldsymbol{x}^{(k)}}{\boldsymbol{x}^{(k)\top}\boldsymbol{A}\boldsymbol{x}^{(k)}}\\
&= \boldsymbol{x}^{(k)\top}\boldsymbol{A}\boldsymbol{x}^{(k)}\left(1 - \frac{1}{4}\frac{\left(\boldsymbol{g}^{(k)\top}\boldsymbol{g}^{(k)}\right)^2}{\left(\boldsymbol{g}^{(k)\top}\boldsymbol{A}\boldsymbol{g}^{(k)}\right)\left(\boldsymbol{x}^{(k)\top}\boldsymbol{A}\boldsymbol{x}^{(k)}\right)}\right)\\
&= f\left(\boldsymbol{x}^{(k)}\right)\left(1 - \frac{\left(\boldsymbol{g}^{(k)\top}\boldsymbol{g}^{(k)}\right)^2}{\left(\boldsymbol{g}^{(k)\top}\boldsymbol{A}\boldsymbol{g}^{(k)}\right)\left(2\boldsymbol{x}^{(k)\top}\boldsymbol{A}\boldsymbol{A}^{-1}\left(2\boldsymbol{A}\boldsymbol{x}^{(k)}\right)\right)}\right)
\end{aligned}$$

$$= f\left(\boldsymbol{x}^{(k)}\right)\left(1 - \frac{\left(\boldsymbol{g}^{(k)\top}\boldsymbol{g}^{(k)}\right)^2}{\left(\boldsymbol{g}^{(k)\top}\boldsymbol{A}\boldsymbol{g}^{(k)}\right)\left(\boldsymbol{g}^{(k)\top}\boldsymbol{A}^{-1}\boldsymbol{g}^{(k)}\right)}\right)$$

- **Theorem 3.3.1 (Kantorovich inequality)** Let $\boldsymbol{A}$ be a positive definite $n \times n$ matrix. Then for any $\boldsymbol{x} \in \mathbb{R}^n \backslash \{\boldsymbol{0}\}$,

$$\frac{(\boldsymbol{x}^\top\boldsymbol{x})^2}{(\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x})(\boldsymbol{x}^\top\boldsymbol{A}^{-1}\boldsymbol{x})} \geq \frac{4\lambda_{\max}(\boldsymbol{A})\lambda_{\min}(\boldsymbol{A})}{\left(\lambda_{\max}(\boldsymbol{A}) + \lambda_{\min}(\boldsymbol{A})\right)^2}.$$

- Hence, by Kantorovich inequality,

$$\frac{\left(\boldsymbol{g}^{(k)\top}\boldsymbol{g}^{(k)}\right)^2}{\left(\boldsymbol{g}^{(k)\top}\boldsymbol{A}\boldsymbol{g}^{(k)}\right)\left(\boldsymbol{g}^{(k)\top}\boldsymbol{A}^\top\boldsymbol{g}^{(k)}\right)} \geq \frac{4\lambda_{\max}(\boldsymbol{A})\lambda_{\min}(\boldsymbol{A})}{\left(\lambda_{\max}(\boldsymbol{A}) + \lambda_{\min}(\boldsymbol{A})\right)^2}$$

  and

$$f\left(\boldsymbol{x}^{(k+1)}\right) \leq \left(1 - \frac{4Mm}{(M+m)^2}\right) f\left(\boldsymbol{x}^{(k)}\right) = \left(\frac{M-m}{M+m}\right)^2 f\left(\boldsymbol{x}^{(k)}\right) = \left(\frac{M/m-1}{M/m+1}\right)^2 f\left(\boldsymbol{x}^{(k)}\right),$$

  where $M = \lambda_{\max}(\boldsymbol{A})$, $m = \lambda_{\min}(\boldsymbol{A})$.

- The speed of convergence depends on the ratio $\lambda_{max}(\boldsymbol{A})/\lambda_{min}(\boldsymbol{A})$. As $\lambda_{max}(\boldsymbol{A})/\lambda_{min}(\boldsymbol{A})$ gets larger, the convergence speed becomes slower. The ratio $\lambda_{max}(\boldsymbol{A})/\lambda_{min}(\boldsymbol{A})$ is called the **condition number**.

- Matrices with large condition number are called **ill-conditioned**, and matrices with small condition number are call **well-conditioned**.

- The discussion assumes quadratic objective functions, where the Hessian matrix is constant. In practice, the matrix $\boldsymbol{A}$ is replaced by the Hessian matrix $\nabla^2 f(\boldsymbol{x}^*)$.

Example 3.3.1. The Rosenbrock function is defined as:

$$f(x_1, x_2) = 100(x_2 - x_1^2)^2 - (1 - x_1)^2.$$

Given that $\boldsymbol{x}^* = [1, 1]$ is the unique stationary point, find the condition number.

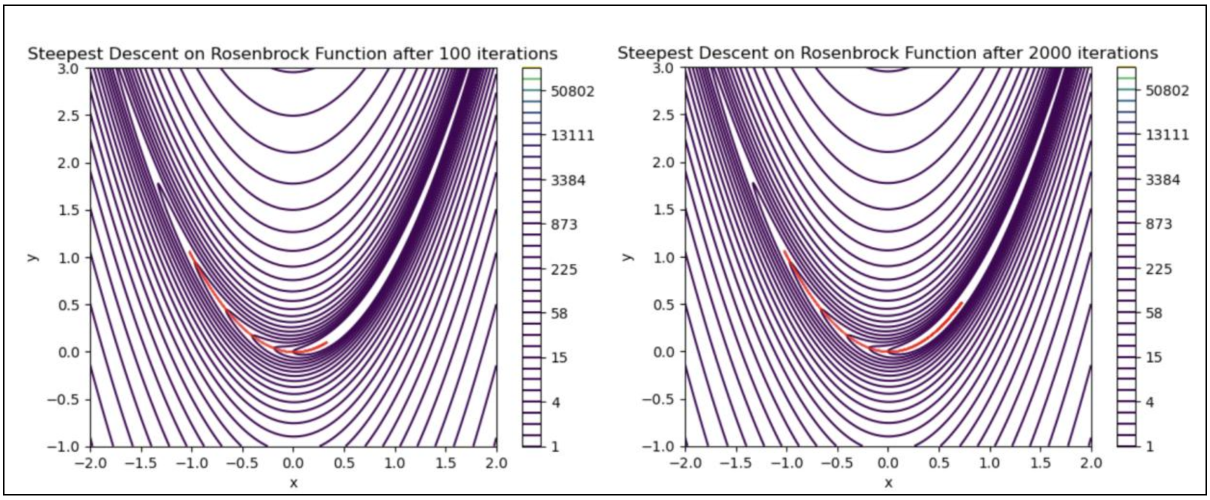
$$\nabla f(\boldsymbol{x}) = \begin{bmatrix} -400x_1(x_2 - x_1^2) - 2(1 - x_1) \\ 200(x_2 - x_1^2) \end{bmatrix} \quad \nabla^2 f(\boldsymbol{x}) = \begin{bmatrix} -400x_2 + 1200x_1^2 + 2 & -400x_1 \\ -400x_1 & 200 \end{bmatrix}$$

Hence,

$$\nabla^2 f(\boldsymbol{x}^*) = \begin{bmatrix} -400 + 1200 + 2 & -400 \\ -400 & 200 \end{bmatrix} = \begin{bmatrix} 802 & -400 \\ -400 & 200 \end{bmatrix}$$

Therefore, the condition number is

$$\frac{\lambda_{max}\left(\nabla^2 f(\boldsymbol{x}^*)\right)}{\lambda_{min}\left(\nabla^2 f(\boldsymbol{x}^*)\right)} \approx \frac{1001}{0.4} = 2502.5$$



3.3 THE ORDER OF CONVERGENCE

- The **order of convergence** of a sequence quantifies **how quickly** the sequence approaches its limit; a **higher order** indicates a **faster rate** of convergence.
- Let $\{\boldsymbol{x}^{(k)}\}$ be the sequence that converges to $\boldsymbol{x}^*$, that is, $\lim_{k\to\infty}\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^*\|=0$. We say that the order of convergence is p, where $p \in \mathbb{R}$, if

$$0 < \lim_{k\to\infty}\frac{\|\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^*\|}{\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^*\|^p} < \infty.$$

If, for all $p > 0$,

$$\lim_{k\to\infty}\frac{\|\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^*\|}{\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^*\|^p} = 0,$$

then we say that the order of convergence is ∞. Note that in the definition above, $0/0$ should be understood to be 0.

Example 3.3.1. Determine the order of convergence of each of the following sequences:

a) $x^{(k)} = 1/k$.
b) $x^{(k)} = \gamma^k$, where $0 < \gamma < 1$.
c) $x^{(k)} = \gamma^{(q^k)}$, where $q > 1$ and $0 < \gamma < 1$.
d) $x^{(k)} = 1$ for all k.

(a) Note that $x^{(k)} \to x^* = 0$. Then,

$$\frac{|x^{(k+1)}-x^*|}{|x^{(k)}-x^*|^p} = \frac{|x^{(k+1)}|}{|x^{(k)}|^p} = \frac{1/(k+1)}{1/k^p} = \frac{k^p}{k+1}.$$

If $p < 1$, the sequence converges to 0. If $p > 1$, it grows to ∞. If $p = 1$, the sequence converges to 1. Hence, the order of convergence is 1.

(b) Note that $x^{(k)} \to x^* = 0$. Then,

$$\frac{|x^{(k+1)} - x^*|}{|x^{(k)} - x^*|^p} = \frac{\gamma^{k+1}}{(\gamma^k)^p} = \gamma^{k+1-kp} = \gamma^{k(1-p)+1}.$$

If $p < 1$, the sequence converges to 0. If $p > 1$, it grows to ∞. If $p = 1$, the sequence converges to $\gamma < \infty$. Hence, the order of convergence is 1.

(c) Note that $x^{(k)} \to x^* = 0$. Then,

$$\frac{|x^{(k+1)} - x^*|}{|x^{(k)} - x^*|^p} = \frac{\gamma^{(q^{k+1})}}{\left(\gamma^{(q^k)}\right)^p} = \gamma^{(q^{k+1}-pq^k)} = \gamma^{(q-p)q^k}.$$

If $p < q$, the sequence converges to 0. If $p > q$, it grows to ∞. If $p = q$, the sequence converges to $1 < \infty$. Hence, the order of convergence is q.

(d) Note that $x^{(k)} \to x^* = 1$. Then,

$$\frac{|x^{(k+1)} - x^*|}{|x^{(k)} - x^*|^p} = \frac{|x^{(k+1)} - 1|}{|x^{(k)} - 1|^p} = \frac{0}{0^p} = 0$$

for all p. Hence, the order of convergence is ∞.

Example 3.3.2 Consider the problem of finding a minimizer of the function $f: \mathbb{R} \to \mathbb{R}$ given by

$$f(x) = x^2 - \frac{x^3}{3}.$$

Suppose that we use the algorithm $x^{(k+1)} = x^{(k)} - \alpha f'\left(x^{(k)}\right)$ with step size $\alpha = 1/2$ and initial condition $x^{(0)} = 1$. Show that the order of convergence is 2.

We first show that the algorithm converges to a local minimizer of f. Indeed, we have $f'(x) = 2x - x^2$. The fixed-step-size gradient algorithm with step size $\alpha = 1/2$ is therefore given by

$$x^{(k+1)} = x^{(k)} - \alpha f'\left(x^{(k)}\right) = \frac{1}{2}\left(x^{(k)}\right)^2.$$

With $x^{(0)} = 1$, we can derive the expression $x^{(k)} = (1/2)^{2^k-1}$. Hence, the algorithm converges to 0, a strict local minimizer of f.

When $p = 2$, we have

$$\frac{|x^{(k+1)}|}{|x^{(k)}|^p} = \frac{(1/2)^{2^{k+1}-1}}{\left((1/2)^{2^k-1}\right)^p} = (1/2)^{2^{k+1}-1-p2^k+2p} = (1/2)^{2^{k+1}-1-2\cdot 2^k+2} = \frac{1}{2}.$$

Therefore, the order of convergence is 2.

What happen if we try $p = 1$ or 3:

$$\frac{\left|x^{(k+1)}\right|}{\left|x^{(k)}\right|} = \frac{(1/2)^{2^{k+1}-1}}{(1/2)^{2^k-1}} = (1/2)^{2^{k+1}-1-2^k+1} = \frac{1}{2^2}.$$

$$\frac{\left|x^{(k+1)}\right|}{\left|x^{(k)}\right|^3} = \frac{(1/2)^{2^{k+1}-1}}{(1/2)^{3(2^k-1)}} = (1/2)^{2^{k+1}-1-3\cdot 2^k+3} = (1/2)^{-2^k+2} = 2^{2^k-2} \to \infty.$$

- In the analysis, we assume the objection function is a quadratic function of the form
$$f(x) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} - \boldsymbol{b}^\top \boldsymbol{x},$$
where $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ is a symmetric positive definite matrix, $\boldsymbol{b} \in \mathbb{R}^n$, and $\boldsymbol{x} \in \mathbb{R}^n$.

- In example 3.2.2, the steepest descent algorithm for the quadratic function is derived as
$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \frac{\boldsymbol{g}^{(k)\top}\boldsymbol{g}^{(k)}}{\boldsymbol{g}^{(k)\top}\boldsymbol{Q}\boldsymbol{g}^{(k)}}\boldsymbol{g}^{(k)},$$
where $\boldsymbol{g}^{(k)} = \nabla f\left(\boldsymbol{x}^{(k)}\right) = \boldsymbol{Q}\boldsymbol{x}^{(k)} - \boldsymbol{b}$.

- **Theorem 3.3.1** Let $\{\boldsymbol{x}^{(k)}\}$ be a convergent sequence of iterates of the steepest descent algorithm applied to a function f. Then, the order of convergence of is 1 in the worst case; that is, there exist a function f and an initial condition $\boldsymbol{x}^{(0)}$ such that the order of convergence of $\{\boldsymbol{x}^{(k)}\}$ is 1.