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

4 NEWTON'S METHOD

4.1 INTRODUCTION

- The method of steepest descent uses **only the first derivatives** (gradients) in selecting a suitable search direction. This strategy is not always the most effective. If higher derivatives are used, the resulting iterative algorithm may perform better than the steepest descent method.

- Newton's method (sometimes called the Newton-Raphson method) uses the **first** and the **second derivatives** and indeed does perform better than the steepest descent method if the initial point is close to the minimizer.

- We can obtain a quadratic approximation to the twice continuously differentiable objection function $f: \mathbb{R}^n \to \mathbb{R}$ using the Taylor series expansion of f about the current point $\boldsymbol{x}^{(k)}$, neglecting terms of order three and higher:

$$f(x) \approx f\left(\boldsymbol{x}^{(k)}\right) + \left(\boldsymbol{x} - \boldsymbol{x}^{(k)}\right)^{\top} \boldsymbol{g}^{(k)} + \frac{1}{2}\left(\boldsymbol{x} - \boldsymbol{x}^{(k)}\right)^{\top} \boldsymbol{F}\left(\boldsymbol{x}^{(k)}\right)\left(\boldsymbol{x} - \boldsymbol{x}^{(k)}\right) \stackrel{\text{def}}{=} q(\boldsymbol{x}),$$

where, for simplicity, we use the notation $\boldsymbol{g}^{(k)} = \nabla f\left(\boldsymbol{x}^{(k)}\right)$. Applying the FONC to q yields

$$0 = \nabla q(\boldsymbol{x}) = \boldsymbol{g}^{(k)} + \boldsymbol{F}\left(\boldsymbol{x}^{(k)}\right)\left(\boldsymbol{x} - \boldsymbol{x}^{(k)}\right).$$

- If $\boldsymbol{F}\left(\boldsymbol{x}^{(k)}\right) > 0$, then q achieves a minimum at

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \boldsymbol{F}\left(\boldsymbol{x}^{(k)}\right)^{-1} \boldsymbol{g}^{(k)}.$$

This recursive formula represents **Newton's method**.

Example 4.1.1. Use Newton's method to minimize the Powell function:

$$f(x_1, x_2, x_3, x_4) = (x_1 + 10x_2)^2 + 5(x_3 - x_4)^2 + (x_2 - 2x_3)^4 + 10(x_1 - x_4)^4.$$

Use as the starting point $\boldsymbol{x}^{(0)} = [3, -1, 0, 1]^{\top}$. Perform three iterations.

Note that $f\left(\boldsymbol{x}^{(0)}\right) = 215$. We have

$$\nabla f(\boldsymbol{x}) = \begin{bmatrix} 2(x_1 + 10x_2) + 40(x_1 - x_4)^3 \\ 20(x_1 + 10x_2) + 4(x_2 - 2x_3)^3 \\ 10(x_3 - x_4) - 8(x_2 - 2x_3)^3 \\ -10(x_3 - x_4) - 40(x_1 - x_4)^3 \end{bmatrix},$$

and $\boldsymbol{F}(\boldsymbol{x})$ is given by

$$\begin{bmatrix} 2+120(x_1-x_4)^2 & 20 & 0 & -120(x_1-x_4)^2 \\ 20 & 200+12(x_2-2x_3)^2 & -24(x_2-2x_3)^2 & 0 \\ 0 & -24(x_2-2x_3)^2 & 10+48(x_2-2x_3)^2 & -10 \\ -120(x_1-x_4)^2 & 0 & -10 & 10+120(x_1-x_4)^2 \end{bmatrix}.$$

Iteration 1.

$$\boldsymbol{g}^{(0)} = [306, -144, -2, -310]^\top,$$

$$\boldsymbol{F}(\boldsymbol{x}^{(0)}) = \begin{bmatrix} 482 & 20 & 0 & -480 \\ 20 & 212 & -24 & 0 \\ 0 & -24 & 58 & -10 \\ -480 & 0 & -10 & 490 \end{bmatrix},$$

$$\boldsymbol{F}(\boldsymbol{x}^{(0)})^{-1} = \begin{bmatrix} 0.1126 & -0.0089 & 0.0154 & 0.1106 \\ -0.0089 & 0.0057 & 0.0008 & -0.0087 \\ 0.0154 & 0.0008 & 0.0203 & 0.0155 \\ 0.1106 & -0.0087 & 0.0155 & 0.1107 \end{bmatrix},$$

$$\boldsymbol{F}(\boldsymbol{x}^{(0)})^{-1}\boldsymbol{g}^{(0)} = [1.4127, -0.8413, -0.2540, 0.7460]^\top.$$

Hence,

$$\boldsymbol{x}^{(1)} = \boldsymbol{x}^{(0)} - \boldsymbol{F}(\boldsymbol{x}^{(0)})^{-1}\boldsymbol{g}^{(0)} = [-1.5873, -0.1587, 0.2540, 0.2540]^\top,$$

$$f(\boldsymbol{x}^{(1)}) = 31.8.$$

Iteration 2.

$$\boldsymbol{g}^{(1)} = [94.81, -1.179, 2.371, -94.81]^\top,$$

$$\boldsymbol{F}(\boldsymbol{x}^{(1)}) = \begin{bmatrix} 215.3 & 20 & 0 & -213.3 \\ 20 & 205.3 & -10.67 & 0 \\ 0 & -10.67 & 31.34 & -10 \\ -213.3 & 0 & -10 & 223.3 \end{bmatrix},$$

$$\boldsymbol{F}(\boldsymbol{x}^{(1)})^{-1}\boldsymbol{g}^{(1)} = [0.5291, -0.0529, 0.0846, 0.0846]^\top.$$

Hence,

$$\boldsymbol{x}^{(2)} = \boldsymbol{x}^{(1)} - \boldsymbol{F}(\boldsymbol{x}^{(1)})^{-1}\boldsymbol{g}^{(1)} = [1.0582, -0.1058, 0.1694, 0.1694]^\top,$$

$$f(\boldsymbol{x}^{(2)}) = 6.28.$$

Iteration 3.

$$\boldsymbol{g}^{(2)} = [28.09, -0.3475, 0.7031, -28.08]^\top,$$

$$\boldsymbol{F}(\boldsymbol{x}^{(2)}) = \begin{bmatrix} 96.80 & 20 & 0 & -94.80 \\ 20 & 202.4 & -4.744 & 0 \\ 0 & -4.744 & 19.49 & -10 \\ -94.80 & 0 & -10 & 104.80 \end{bmatrix},$$

$$\boldsymbol{x}^{(3)} = \boldsymbol{x}^{(2)} - \boldsymbol{F}\left(\boldsymbol{x}^{(2)}\right)^{-1}\boldsymbol{g}^{(2)} = [0.7037, -0.0704, 0.1121, 0.1111]^{\top},$$

$$f\left(\boldsymbol{x}^{(3)}\right) = 1.24$$

- Observe that the kth iteration of Newton's method can be written in two steps as

  1. Solve $F\left(\boldsymbol{x}^{(k)}\right)\boldsymbol{d}^{(k)} = -\boldsymbol{g}^{(k)}$ for $\boldsymbol{d}^{(k)}$.
  2. Set $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \boldsymbol{d}^{(k)}$.

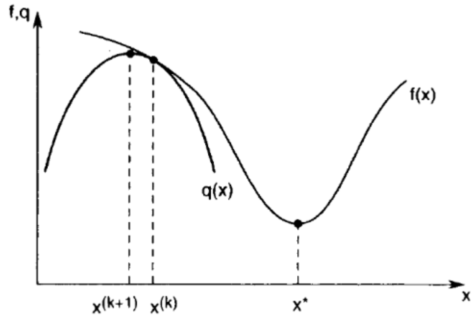
- As in the one-variable case, Newton's method can also be viewed as a technique for iteratively solving the equation

$$\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{0},$$

  where $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{g}: \mathbb{R}^n \to \mathbb{R}^n$.

4.2 ANALYSIS OF NEWTON'S METHOD

- As in the one-variable case there is no guarantee that Newton's algorithm heads in the direction of decreasing values of the objective function if $\boldsymbol{F}\left(\boldsymbol{x}^{(k)}\right)$ is not positive definite.



- Moreover, even if $\boldsymbol{F}\left(\boldsymbol{x}^{(k)}\right) > 0$, Newton's method may not be a descent method; that is, it is possible that $f\left(\boldsymbol{x}^{(k+1)}\right) \geq f\left(\boldsymbol{x}^{(k)}\right)$. For example, this may occur if our starting point $\boldsymbol{x}^{(0)}$ is far away from the solution.

- **Theorem 4.2.1** Suppose that $f \in C^3$ and $\boldsymbol{x}^* \in \mathbb{R}^n$ is a point such that $\nabla f(\boldsymbol{x}^*) = 0$ and $\boldsymbol{F}(\boldsymbol{x}^*)$ is invertible. Then, for all $\boldsymbol{x}^{(0)}$ sufficiently close to $\boldsymbol{x}^*$, Newton's method is well defined for all k and converges to $\boldsymbol{x}^*$ with an order of convergence at least 2.

- ***Warning:*** In the Theorem 4.2.1, we did not state that $\boldsymbol{x}^*$ is a local minimizer. For example, if $\boldsymbol{x}^*$ is a local maximizer, then provided that $f \in C^3$ and $\boldsymbol{F}(\boldsymbol{x}^*)$is invertible, Newton's method converges to $\boldsymbol{x}^*$ if we start close enough to it.

Example 4.2.1. Suppose f is a quadratic function such that

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

where $\boldsymbol{Q} = \boldsymbol{Q}^\top$ is invertible. Show that Newton's method reaches the point $\boldsymbol{x}^*$ such that $\nabla f(\boldsymbol{x}^*) = 0$ in just one step starting from any initial point $\boldsymbol{x}^{(0)}$. Determine the order of convergence.

Note that $\boldsymbol{g}(\boldsymbol{x}) = \nabla f(\boldsymbol{x}) = \boldsymbol{Q}\boldsymbol{x} - \boldsymbol{b}$ and $\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{Q}$.

Hence, given any initial point $\boldsymbol{x}^{(0)}$, by Newton's algorithm

$$\begin{aligned}\boldsymbol{x}^{(1)} &= \boldsymbol{x}^{(0)} - \boldsymbol{F}\left(\boldsymbol{x}^{(0)}\right)^{-1}\boldsymbol{g}^{(0)} \\ &= \boldsymbol{x}^{(0)} - \boldsymbol{Q}^{-1}\left(\boldsymbol{Q}\boldsymbol{x}^{(0)} - \boldsymbol{b}\right) \\ &= \boldsymbol{Q}^{-1}\boldsymbol{b} \\ &= \boldsymbol{x}^*\end{aligned}$$

In this case, if for all $p > 0$,

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

Therefore, the order of convergence is ∞.

- The Newton's method may not be a descent method; that is, it is possible that $f\left(\boldsymbol{x}^{(k+1)}\right) \geq f\left(\boldsymbol{x}^{(k)}\right)$. Fortunately, it is possible to modify the algorithm such that the descent property holds.

- **Theorem 4.2.2** Let $\left\{\boldsymbol{x}^{(k)}\right\}$ be the sequence generated by Newton's method for minimizing a given objective function $f(\boldsymbol{x})$. If the Hessian $\boldsymbol{F}\left(\boldsymbol{x}^{(k)}\right) \succ 0$ and $\boldsymbol{g}^{(k)} = \nabla f\left(\boldsymbol{x}^{(k)}\right) \neq \boldsymbol{0}$, then the search direction

$$\boldsymbol{d}^{(k)} = -\boldsymbol{F}\left(\boldsymbol{x}^{(k)}\right)^{-1}\boldsymbol{g}^{(k)} = \boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}$$

from $\boldsymbol{x}^{(k)}$ to $\boldsymbol{x}^{(k+1)}$ is a descent direction for f in the sense that there exists an $\bar{\alpha} > 0$ such that for all $\alpha \in (0, \bar{\alpha})$,

$$f\left(\boldsymbol{x}^{(k)} + \alpha \boldsymbol{d}^{(k)}\right) < f\left(\boldsymbol{x}^{(k)}\right).$$

Proof Let

$$\phi(\alpha) = f\left(\boldsymbol{x}^{(k)} + \alpha \boldsymbol{d}^{(k)}\right).$$

Then, using the chain rule, we obtain

$$\phi'(\alpha) = \nabla f\left(\boldsymbol{x}^{(k)} + \alpha \boldsymbol{d}^{(k)}\right)^\top \boldsymbol{d}^{(k)}.$$

Hence,

$$\phi'(0) = \nabla f\left(\boldsymbol{x}^{(k)}\right)^\top \boldsymbol{d}^{(k)} = -\boldsymbol{g}^{(k)\top} \boldsymbol{F}\left(\boldsymbol{x}^{(k)}\right)^{-1} \boldsymbol{g}^{(k)} < 0,$$

because $\boldsymbol{F}\left(\boldsymbol{x}^{(k)}\right)^{-1} > 0$ and $\boldsymbol{g}^{(k)} \neq \boldsymbol{0}$. Thus, there exists an $\bar{\alpha} > 0$ so that for all $\alpha \in (0, \bar{\alpha})$, $\phi(\alpha) < \phi(0)$. This implies that for all $\alpha \in (0, \bar{\alpha})$,

$$f\left(\boldsymbol{x}^{(k)} + \alpha \boldsymbol{d}^{(k)}\right) < f\left(\boldsymbol{x}^{(k)}\right).$$

- Theorem 4.2.2 motivates the following modification of Newton's method:

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

where

$$\alpha_k = \arg\min_{\alpha \geq 0} f\left(\boldsymbol{x}^{(k)} - \alpha \boldsymbol{F}\left(\boldsymbol{x}^{(k)}\right)^{-1} \boldsymbol{g}^{(k)}\right);$$

that is, at each iteration, we perform a line search in the direction $-\boldsymbol{F}\left(\boldsymbol{x}^{(k)}\right)^{-1} \boldsymbol{g}^{(k)}$. By Theorem 4.2.2 we conclude that the modified Newton's method has the descent property; that is,

$$f\left(\boldsymbol{x}^{(k+1)}\right) < f\left(\boldsymbol{x}^{(k)}\right)$$

whenever $\boldsymbol{g}^{(k)} \neq \boldsymbol{0}$.

- A drawback of Newton's method is that evaluation of $\boldsymbol{F}\left(\boldsymbol{x}^{(k)}\right)$ for large n can be computationally expensive. Furthermore, we have to solve the set of n linear equations $\boldsymbol{F}\left(\boldsymbol{x}^{(k)}\right)\boldsymbol{d}^{(k)} = -\boldsymbol{g}^{(k)}$.

- Another source of potential problems in Newton's method arises from the Hessian matrix not being positive definite.

4.3 LEVENBERG-MARQUARDT MODIFICATION

- If the Hessian matrix $\boldsymbol{F}\left(\boldsymbol{x}^{(k)}\right)$ is not positive definite, then the search direction $\boldsymbol{d}^{(k)} = -\boldsymbol{F}\left(\boldsymbol{x}^{(k)}\right)^{-1} \boldsymbol{g}^{(k)}$ may not point in a descent direction. A simple technique to ensure that the search direction is a descent direction is to introduce the **Levenberg-Marquardt modification** of Newton's algorithm:

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha_k \left(\boldsymbol{F}\left(\boldsymbol{x}^{(k)}\right) + \mu_k \boldsymbol{I}\right)^{-1} \boldsymbol{g}^{(k)},$$

where $\mu_k \geq 0$.

- The idea underlying the Levenberg-Marquardt modification is as follows. Consider a symmetric matrix $\boldsymbol{F}$, which may not be positive definite. Let $\lambda_1, \ldots, \lambda_n$

be the eigenvalues of $\boldsymbol{F}$ with corresponding eigenvectors $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$. The eigenvalues $\lambda_1, \dots, \lambda_n$ are real, but may not all be positive. Next, consider the matrix $\boldsymbol{G} = \boldsymbol{F} + \mu \boldsymbol{I}$, where $\mu \geq 0$. Note that the eigenvalues of $\boldsymbol{G}$ are $\lambda_1 + \mu, \dots, \lambda_n + \mu$. Indeed,

$$\boldsymbol{G}\boldsymbol{v}_i = (\boldsymbol{F} + \mu \boldsymbol{I})\boldsymbol{v}_i = \boldsymbol{F}\boldsymbol{v}_i + \mu \boldsymbol{I}\boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i + \mu \boldsymbol{v}_i = (\lambda_i + \mu)\boldsymbol{v}_i,$$

which shows that for all $i = 1, \dots, n$, v_i is also an eigenvector of $\boldsymbol{G}$ with eigenvalue $\lambda_i + \mu$. Therefore, if μ is sufficiently large, then all the eigenvalues of $\boldsymbol{G}$ are positive and $\boldsymbol{G}$ is positive definite.

- The Levenberg-Marquardt modification of Newton's algorithm can be made to approach the behavior of the pure Newton's method by letting $\mu_k \to 0$. On the other hand, by letting $\mu_k \to \infty$, the algorithm approaches a pure gradient method with small step size.

4.4 NEWTON'S METHOD FOR NONLINEAR LEAST SQUARES

- We now examine a particular class of optimization problems and the use of Newton's method for solving them. Consider the following problem:

$$\text{minimize} \sum_{i=1}^{m} \left(r_i(\boldsymbol{x})\right)^2,$$

where $r_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \dots, m$, are given functions. This particular problem is called a **nonlinear least-squares problem**.

Example 4.4.1 Suppose that we are given m measurements of a process at m points in time (here, $m = 21$). Let $t_1, \dots, t_m$ denote the measurement times and $y_1, \dots, y_m$ the measurement values. Note that $t_1 = 0$ while $t_{21} = 10$. We wish to fit a sinusoid to the measurement data. The equation of the sinusoid is

$$y = A \sin(\omega t + \phi)$$

with appropriate choices of the parameters A, ω, and ϕ. Formulate the data-fitting problem and derive the Newton's method.

To formulate the data-fitting problem, we construct the objective function

$$\sum_{i=1}^{m} (y_i - A \sin(\omega t_i + \phi))^2,$$

representing the sum of the squared errors between the measurement values and the function values at the corresponding points in time. Let $\boldsymbol{x} = [A, \omega, \phi]^\top$ represent the vector of decision variables. We therefore obtain a nonlinear least-squares problem with

$$r_i(\boldsymbol{x}) = y_i - A \sin(\omega t_i + \phi).$$

Defining $\boldsymbol{r} = [r_1, \ldots, r_m]^\top$, we write the objective function as $f(x) = r(\boldsymbol{x})^\top r(\boldsymbol{x})$. To apply Newton's method, we need to compute the gradient and the Hessian of f. The jth component of $\nabla f(\boldsymbol{x})$ is

$$\left(\nabla f(\boldsymbol{x})\right)_j = \frac{\partial f}{\partial x_j}(\boldsymbol{x}) = 2\sum_{i=1}^{m} r_i(\boldsymbol{x}) \frac{\partial r_i}{\partial x_j}(\boldsymbol{x}).$$

Then, the gradient of f can be represented as

$$\nabla f(\boldsymbol{x}) = 2\boldsymbol{J}(\boldsymbol{x})^\top \boldsymbol{r}(\boldsymbol{x})$$

where

$$\boldsymbol{J}(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial r_1}{\partial x_1}(\boldsymbol{x}) & \frac{\partial r_1}{\partial x_2}(\boldsymbol{x}) & \frac{\partial r_1}{\partial x_3}(\boldsymbol{x}) \\ \vdots & \vdots & \vdots \\ \frac{\partial r_m}{\partial x_1}(\boldsymbol{x}) & \frac{\partial r_m}{\partial x_2}(\boldsymbol{x}) & \frac{\partial r_m}{\partial x_3}(\boldsymbol{x}) \end{bmatrix}.$$

Next, we compute the Hessian matrix of f. The (k, j)th component of the Hessian is given by

$$\begin{aligned} \frac{\partial^2 f}{\partial x_k \partial x_j}(\boldsymbol{x}) &= \frac{\partial}{\partial x_k}\left(\frac{\partial f}{\partial x_j}(\boldsymbol{x})\right) \\ &= \frac{\partial}{\partial x_k}\left(2\sum_{i=1}^{m} r_i(\boldsymbol{x}) \frac{\partial r_i}{\partial x_j}(\boldsymbol{x})\right) \\ &= 2\sum_{i=1}^{m}\left(\frac{\partial r_i}{\partial x_k}(\boldsymbol{x}) \frac{\partial r_i}{\partial x_j}(\boldsymbol{x}) + r_i(\boldsymbol{x}) \frac{\partial^2 r_i}{\partial x_k \partial x_j}(\boldsymbol{x})\right) \end{aligned}$$

Letting $\boldsymbol{S}(\boldsymbol{x})$ be the matrix whose (k, j)th component is

$$\sum_{i=1}^{m} r_i(\boldsymbol{x}) \frac{\partial^2 r_i}{\partial x_k \partial x_j}(\boldsymbol{x}),$$

we write the Hessian matrix as

$$\boldsymbol{F}(\boldsymbol{x}) = 2\left(\boldsymbol{J}(\boldsymbol{x})^\top \boldsymbol{J}(\boldsymbol{x}) + \boldsymbol{S}(\boldsymbol{x})\right).$$

Therefore, Newton's method applied to the nonlinear least-squares problem is given by

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \left(\boldsymbol{J}(\boldsymbol{x})^\top \boldsymbol{J}(\boldsymbol{x}) + \boldsymbol{S}(\boldsymbol{x})\right)^{-1} \boldsymbol{J}(\boldsymbol{x})^\top \boldsymbol{r}(\boldsymbol{x}).$$