MA 5037: Optimization Methods and Applications
Newton's Method

Suh-Yuh Yang (楊肅煜)

Department of Mathematics, National Central University
Jhongli District, Taoyuan City 320317, Taiwan

First version: July 13, 2018/Last updated: June 15, 2025

Newton's method

- We consider the unconstrained minimization problem:

$$\min\{f(x) : x \in \mathbb{R}^n\},$$

where the objective function f *is twice continuously differentiable over* $\mathbb{R}^n$. We will consider a *second order method*, namely a method that uses, in addition to the information on function values and gradients, evaluations of the Hessian matrices.

- **Main idea of Newton's method:** *Given an iterate* x_k, *the next iterate* x_{k+1} *is chosen to minimize the quadratic approximation of the function* f *around* x_k,

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^n} \Big\{ f(x_k) + \nabla f(x_k)^\top (x - x_k) + \frac{1}{2}(x - x_k)^\top \nabla^2 f(x_k)(x - x_k) \Big\},$$

where we assume that $\nabla^2 f(x_k)$ *is positive definite, which implies the well-definedness of the above problem (§2.5, Lemma 2.41).*

Newton directions

- The unique minimizer of the above minimization problem is the unique stationary point, which implies that

$$\nabla f(\boldsymbol{x}_k) + \nabla^2 f(\boldsymbol{x}_k)(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k) = \mathbf{0},$$

  or equivalently,

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - (\nabla^2 f(\boldsymbol{x}_k))^{-1} \nabla f(\boldsymbol{x}_k).$$

- *The vector* $-(\nabla^2 f(\boldsymbol{x}_k))^{-1} \nabla f(\boldsymbol{x}_k)$ *is called the Newton direction,* and the algorithm induced by the update formula is called the *pure Newton's method.*

- *Note that when* $\nabla^2 f(\boldsymbol{x}_k)$ *is positive definite for any* k*, pure Newton's method is essentially a scaled gradient method with* $t_k = 1$ *for all* k*, and Newton's directions are descent directions since*

$$f'(\boldsymbol{x}_k; \underbrace{-(\nabla^2 f(\boldsymbol{x}_k))^{-1} \nabla f(\boldsymbol{x}_k)}_{\text{Newton direction}}) = -\nabla f(\boldsymbol{x}_k)^\top \underbrace{(\nabla^2 f(\boldsymbol{x}_k))^{-1}}_{\succ \mathbf{0}} \nabla f(\boldsymbol{x}_k) < 0.$$

Pure Newton's method

Input: $\varepsilon > 0$, tolerance parameter.

Initialization: Pick $\boldsymbol{x}_0 \in \mathbb{R}^n$ arbitrarily.

General step: For any $k = 0, 1, \cdots$, execute the following steps

(a) Compute the Newton direction $\boldsymbol{d}_k$, which is the solution to *the linear system* $\nabla^2 f(\boldsymbol{x}_k)\boldsymbol{d}_k = -\nabla f(\boldsymbol{x}_k)$.

(b) Set $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{d}_k$.

(c) If $\|\nabla f(\boldsymbol{x}_{k+1})\| \leq \varepsilon$ then stop, and $\boldsymbol{x}_{k+1}$ is the output.

Convergence of the sequence $\{x_k\}$

1. *Newton's method requires that $\nabla^2 f(\boldsymbol{x})$ is positive definite for every $\boldsymbol{x} \in \mathbb{R}^n$, which in particular implies that there exists a unique optimal solution $\boldsymbol{x}^*$* (In fact, we need the assumption $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$. Then, similar to the proof of Theorem 2.38, we can show that $\exists!$ minimizer $\boldsymbol{x}^*$). *However, this is not enough to guarantee convergence of the sequence $\{\boldsymbol{x}_k\}$.*

2. Consider the function $f(x) = \sqrt{1+x^2}$ defined over $\mathbb{R}$. The unique minimizer is of course $x^* = 0$. The first and second derivatives of f are

$$f'(x) = \frac{x}{\sqrt{1+x^2}} \quad \text{and} \quad f''(x) = \frac{1}{(1+x^2)^{3/2}} > 0.$$

Therefore, the pure Newton's method has the form

$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)} = x_k - x_k(1+x_k^2) = -x_k^3.$$

If $|x_0| \geq 1$ then the method diverges.

If $|x_0| < 1$ then the method converges very rapidly to $x^* = 0$.

Quadratic local convergence of Newton's method

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice continuously differentiable function. Assume

- $\exists\, m > 0$ *s.t.* $\nabla^2 f(\boldsymbol{x}) \succeq m\boldsymbol{I}$ *for any* $\boldsymbol{x} \in \mathbb{R}^n$,
- $\exists\, L > 0$ *s.t.* $\|\nabla^2 f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{y})\| \le L\|\boldsymbol{x} - \boldsymbol{y}\|$ *for any* $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$.

Let $\{\boldsymbol{x}_k\}$ be the sequence generated by Newton's method, and let $\boldsymbol{x}^*$ be the unique minimizer of f over $\mathbb{R}^n$. Then for any $k = 0, 1, \cdots$ the following inequality holds:

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\| \le \frac{L}{2m}\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2.$$

In addition, if $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| \le m/L$ then

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\| \le \frac{2m}{L}\left(\frac{1}{2}\right)^{2^k}, \quad k = 0, 1, \cdots. \qquad (\star)$$

Proof of the quadratic local convergence

Let k be a nonnegative integer. Then

$$\begin{aligned}
\mathbf{x}_{k+1}-\mathbf{x}^* &= \mathbf{x}_k-(\nabla^2 f(\mathbf{x}_k))^{-1}\nabla f(\mathbf{x}_k)-\mathbf{x}^* \\
&\overset{\nabla f(\mathbf{x}^*)=0}{=} \mathbf{x}_k-\mathbf{x}^*+(\nabla^2 f(\mathbf{x}_k))^{-1}(\nabla f(\mathbf{x}^*)-\nabla f(\mathbf{x}_k)) \\
&= \mathbf{x}_k-\mathbf{x}^*+(\nabla^2 f(\mathbf{x}_k))^{-1}\int_0^1 [\nabla^2 f(\mathbf{x}_k+t(\mathbf{x}^*-\mathbf{x}_k))](\mathbf{x}^*-\mathbf{x}_k)dt \\
&= (\nabla^2 f(\mathbf{x}_k))^{-1}\int_0^1 \left[\nabla^2 f(\mathbf{x}_k+t(\mathbf{x}^*-\mathbf{x}_k))-\nabla^2 f(\mathbf{x}_k)\right](\mathbf{x}^*-\mathbf{x}_k)dt.
\end{aligned}$$

Since $\nabla^2 f(\boldsymbol{x}_k) \succeq m\boldsymbol{I}$, it follows that $\|(\nabla^2 f(\boldsymbol{x}_k))^{-1}\| \le 1/m$. Hence,

$$\begin{aligned}
\|\mathbf{x}_{k+1}-\mathbf{x}^*\| &\le \|(\nabla^2 f(\mathbf{x}_k))^{-1}\| \left\| \int_0^1 \left[\nabla^2 f(\mathbf{x}_k+t(\mathbf{x}^*-\mathbf{x}_k))-\nabla^2 f(\mathbf{x}_k)\right](\mathbf{x}^*-\mathbf{x}_k)dt \right\| \\
&\le \|(\nabla^2 f(\mathbf{x}_k))^{-1}\| \int_0^1 \left\| \left[\nabla^2 f(\mathbf{x}_k+t(\mathbf{x}^*-\mathbf{x}_k))-\nabla^2 f(\mathbf{x}_k)\right](\mathbf{x}^*-\mathbf{x}_k) \right\| dt \\
&\le \|(\nabla^2 f(\mathbf{x}_k))^{-1}\| \int_0^1 \left\| \nabla^2 f(\mathbf{x}_k+t(\mathbf{x}^*-\mathbf{x}_k))-\nabla^2 f(\mathbf{x}_k) \right\| \cdot \|\mathbf{x}^*-\mathbf{x}_k\| dt \\
&\le \frac{L}{m}\int_0^1 t\|\mathbf{x}_k-\mathbf{x}^*\|^2 dt = \frac{L}{2m}\|\mathbf{x}_k-\mathbf{x}^*\|^2.
\end{aligned}$$

Proof of the quadratic local convergence (cont'd)

We use the mathematical induction to show $(\star)$. For $n = 0$, we have

$$\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| \leq \frac{m}{L} = \frac{2m}{L}\left(\frac{1}{2}\right)^{2^0}.$$

Assume for $n = k$, we have

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\| \leq \frac{2m}{L}\left(\frac{1}{2}\right)^{2^k}.$$

Then for $n = k + 1$,

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\| \leq \frac{L}{2m}\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 \leq \frac{L}{2m}\left(\frac{2m}{L}\left(\frac{1}{2}\right)^{2^k}\right)^2 = \frac{2m}{L}\left(\frac{1}{2}\right)^{2^{k+1}}.$$

This proves the desired result. $\square$

Example 1

Consider the minimization problem:

$$\min_{x,y} 100x^4 + 0.01y^4,$$

whose optimal solution is obviously $(x, y) = (0, 0)$.

1. The gradient method with initial vector $\boldsymbol{x}_0 = (1, 1)^\top$ and parameters $(s, \alpha, \beta, \varepsilon) = (1, 0.5, 0.5, 10^{-6})$ converges after the huge amount of 14612 iterations.
2. Invoking pure Newton's method, we obtain convergence after only 17 iterations.

   Note that the basic assumptions required for the convergence of Newton's method as described in the quadratic local convergence theorem are not satisfied. The Hessian is always positive semidefinite, but it is not always positive definite and does not satisfy a Lipschitz property.

Please see textbook, pages 86-87.

Example 2

Consider the minimization problem *(Please see textbook, pages 87-88):*

$$\min_{x,y}\left(\sqrt{x^2+1}+\sqrt{y^2+1}\right),$$

whose optimal solution is $\boldsymbol{x}=(0,0)$. The Hessian of the function is

$$\nabla^2 f(\boldsymbol{x})=\begin{bmatrix}\frac{1}{(x^2+1)^{3/2}} & 0\\ 0 & \frac{1}{(y^2+1)^{3/2}}\end{bmatrix}\succ \mathbf{0}.$$

Despite the fact that the Hessian is positive definite, there does not exist an $m>0$ for which $\nabla^2 f(\boldsymbol{x})\succeq m\boldsymbol{I}$.

1. If we employ Newton's method with initial vector $\boldsymbol{x}_0=(1,1)^\top$ and tolerance $\varepsilon=10^{-8}$ we obtain convergence after 37 iterations, but in the first 30 iterations the method is almost stuck.
2. The gradient method with backtracking and parameters $(s,\alpha,\beta)=(1,0.5,0.5)$ converges after only 7 iterations.
3. If $\boldsymbol{x}_0=(10,10)^\top$, the gradient method with backtracking converges after 13 iterations, but Newton's method diverges.

Damped Newton's method

As can be seen in the last example, pure Newton's method does not guarantee descent of the generated sequence of function values even when the Hessian is positive definite. *This drawback can be rectified by introducing a stepsize chosen by a certain line search procedure, leading to the so-called damped Newton's method.*

Input: $\alpha, \beta \in (0,1)$: parameters for the backtracking procedure, $\varepsilon > 0$: tolerance parameter.

Initialization: Pick $\boldsymbol{x}_0 \in \mathbb{R}^n$ arbitrarily.

General step: For any $k = 0, 1, \cdots$, execute the following steps

(a) compute the Newton direction $\boldsymbol{d}_k$, which is the solution to the linear system $\nabla^2 f(\boldsymbol{x}_k)\boldsymbol{d}_k = -\nabla f(\boldsymbol{x}_k)$.

(b) set $t_k = 1$. While $f(\boldsymbol{x}_k) - f(\boldsymbol{x}_k + t_k\boldsymbol{d}_k) < -\alpha t_k \nabla f(\boldsymbol{x}_k)^\top \boldsymbol{d}_k$, set $t_k := \beta t_k$.

(c) set $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + t_k\boldsymbol{d}_k$.

(d) if $\|\nabla f(\boldsymbol{x}_{k+1})\| \le \varepsilon$ then stop, and $\boldsymbol{x}_{k+1}$ is the output.

The Cholesky factorization

1. An important issue in Newton's method is whether the Hessian matrix is positive definite, and if it is, then how to solve the linear system $\nabla^2 f(\boldsymbol{x}_k)\boldsymbol{d}_k = -\nabla f(\boldsymbol{x}_k)$.
2. *Given an $n \times n$ positive definite matrix $\boldsymbol{A}$, a Cholesky factorization is a factorization of the form $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^\top$, where $\boldsymbol{L}$ is a lower triangular $n \times n$ matrix whose diagonal is positive.*
3. Given a Cholesky factorization, the task of solving $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ can be easily done by two steps:

   Step 1: Find the solution $\boldsymbol{u}$ of $\boldsymbol{L}\boldsymbol{u} = \boldsymbol{b}$.
   Step 2: Find the solution $\boldsymbol{x}$ of $\boldsymbol{L}^\top \boldsymbol{x} = \boldsymbol{u}$.

   Since $\boldsymbol{L}$ is a triangular matrix with a positive diagonal, steps 1 and 2 can be carried out by forward and backward substitutions, respectively, which both require $O(n^2)$ arithmetic operations.
4. However, the computation of the Cholesky factorization requires $O(n^3)$ operations.

How to compute the Cholesky factorization?

Consider the block matrix partition of the matrices $\boldsymbol{A}$ and $\boldsymbol{L}$:

$$\boldsymbol{A} = \begin{bmatrix} A_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{12}^\top & \boldsymbol{A}_{22} \end{bmatrix}, \quad \boldsymbol{L} = \begin{bmatrix} L_{11} & \boldsymbol{0} \\ \boldsymbol{L}_{21} & \boldsymbol{L}_{22} \end{bmatrix},$$

where $A_{11} \in \mathbb{R}$, $\boldsymbol{A}_{12} \in \mathbb{R}^{1\times(n-1)}$, $\boldsymbol{A}_{22} \in \mathbb{R}^{(n-1)\times(n-1)}$, $L_{11} \in \mathbb{R}$, $\boldsymbol{L}_{21} \in \mathbb{R}^{n-1}$, $\boldsymbol{L}_{22} \in \mathbb{R}^{(n-1)\times(n-1)}$. Since $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^\top$, we have

$$\begin{bmatrix} A_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{12}^\top & \boldsymbol{A}_{22} \end{bmatrix} = \begin{bmatrix} L_{11}^2 & L_{11}\boldsymbol{L}_{21}^\top \\ L_{11}\boldsymbol{L}_{21} & \boldsymbol{L}_{21}\boldsymbol{L}_{21}^\top + \boldsymbol{L}_{22}\boldsymbol{L}_{22}^\top \end{bmatrix}.$$

Therefore,

$$L_{11} = \sqrt{A_{11}}, \quad \boldsymbol{L}_{21} = \frac{1}{\sqrt{A_{11}}}\boldsymbol{A}_{12}^\top,$$

and we can thus also write

$$\boldsymbol{L}_{22}\boldsymbol{L}_{22}^\top = \boldsymbol{A}_{22} - \frac{1}{A_{11}}\boldsymbol{A}_{12}^\top\boldsymbol{A}_{12}.$$

We are left with the task of finding a Cholesky factorization of the $(n-1)\times(n-1)$ matrix $\boldsymbol{A}_{22} - \frac{1}{A_{11}}\boldsymbol{A}_{12}^\top\boldsymbol{A}_{12}$. Continuing in this way, we can compute the complete Cholesky factorization of matrix $\boldsymbol{A}$.

Example

Let the matrix $\boldsymbol{A}$ and the Cholesky factor $\boldsymbol{L}$ are respectively given by

$$\boldsymbol{A} = \begin{bmatrix} 9 & 3 & 3 \\ 3 & 17 & 21 \\ 3 & 21 & 107 \end{bmatrix} \quad \text{and} \quad \boldsymbol{L} = \begin{bmatrix} \ell_{11} & 0 & 0 \\ \ell_{21} & \ell_{22} & 0 \\ \ell_{31} & \ell_{32} & \ell_{33} \end{bmatrix}.$$

Then $\ell_{11} = \sqrt{9} = 3$ and

$$\begin{bmatrix} \ell_{21} \\ \ell_{31} \end{bmatrix} = \frac{1}{3} \begin{bmatrix} 3 \\ 3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

We now need to find the Cholesky factorization of

$$\boldsymbol{A}_{22} - \frac{1}{A_{11}} \boldsymbol{A}_{12}^{\top} \boldsymbol{A}_{12} = \begin{bmatrix} 17 & 21 \\ 21 & 107 \end{bmatrix} - \frac{1}{9} \begin{bmatrix} 3 \\ 3 \end{bmatrix} (3, 3) = \begin{bmatrix} 16 & 20 \\ 20 & 106 \end{bmatrix}.$$

Let $\boldsymbol{L}_{22} = \begin{bmatrix} \ell_{22} & 0 \\ \ell_{32} & \ell_{33} \end{bmatrix}$. Consequently, $\ell_{22} = \sqrt{16} = 4$ and $\ell_{32} = \frac{1}{\sqrt{16}} 20 = 5$. We are thus left with the task of finding the Cholesky factorization of $106 - \dfrac{1}{16}(20 \times 20) = 81$. Of course $\ell_{33} = \sqrt{81} = 9$ and then $\boldsymbol{L} = \begin{bmatrix} 3 & 0 & 0 \\ 1 & 4 & 0 \\ 1 & 5 & 9 \end{bmatrix}$.

A hybrid gradient-Newton method

How to employ Newton's method when the Hessian is not always positive definite? *The simplest one is to construct a hybrid method that employs either a Newton step at iterations in which the Hessian is positive definite or a gradient step when the Hessian is not positive definite.*

Input: $\alpha, \beta \in (0,1)$: parameters for the backtracking procedure, $\varepsilon > 0$: tolerance parameter.

Initialization: Pick $\boldsymbol{x}_0 \in \mathbb{R}^n$ arbitrarily.

General step: For any $k = 0, 1, \cdots$, execute the following steps

(a) if $\nabla^2 f(\boldsymbol{x}_k) \succ \mathbf{0}$ then take $\boldsymbol{d}_k$ as the Newton direction $\boldsymbol{d}_k$, which is the solution to the linear system $\nabla^2 f(\boldsymbol{x}_k)\boldsymbol{d}_k = -\nabla f(\boldsymbol{x}_k)$. Otherwise, set $\boldsymbol{d}_k = -\nabla f(\boldsymbol{x}_k)$.

(b) set $t_k = 1$. While $f(\boldsymbol{x}_k) - f(\boldsymbol{x}_k + t_k\boldsymbol{d}_k) < -\alpha t_k \nabla f(\boldsymbol{x}_k)^\top \boldsymbol{d}_k$, set $t_k := \beta t_k$.

(c) set $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + t_k\boldsymbol{d}_k$.

(d) if $\|\nabla f(\boldsymbol{x}_{k+1})\| \leq \varepsilon$ then stop, and $\boldsymbol{x}_{k+1}$ is the output.