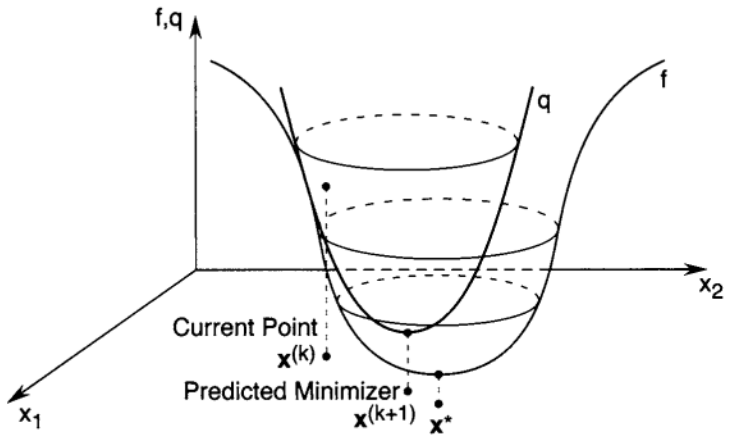
Chapter 9 Newton's Method

Newton's Method

motivation

- Steepest descent method only uses 1st-order derivative (gradient) as searching direction.
- If higher order derivative is applied to recursion, it may perform better than SD method.

1. Construct a quadratic function with the same 1st- and 2nd-order derivatives as the objective function at this point, which can be used as an approximation of the objective function.
2. Find the minimizer of quadratic function as the starting point of the next iteration.
3. Repeat the above procedure to derive the minimizer of the objective function.



Newton's Method

let $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuous differentiable

The Taylor expansion of f at $\boldsymbol{x}^{(k)}$ is

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x}^{(k)}) + (\boldsymbol{x} - \boldsymbol{x}^{(k)})^\top \boldsymbol{g}^{(k)} + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^{(k)})^\top \boldsymbol{F}(\boldsymbol{x}^{(k)})(\boldsymbol{x} - \boldsymbol{x}^{(k)}) =: q(\boldsymbol{x}),$$

where $\boldsymbol{g}^{(k)} = \nabla f(\boldsymbol{x}^{(k)})$ and $\boldsymbol{F}(\boldsymbol{x}^{(k)}) = \nabla^2 f(\boldsymbol{x}^{(k)})$.

By the optimality conditions, the local minima of $q(\boldsymbol{x})$ satisfies

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

If $\boldsymbol{F}(\boldsymbol{x}^{(k)}) \succ 0$, the minimum of q is

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \boldsymbol{F}(\boldsymbol{x}^{(k)})^{-1}\boldsymbol{g}^{(k)} \Longrightarrow \text{Newton's method}$$

Newton's Method

Example (minimize f by Newton's method)

$f(x_1, x_2, x_3, x_4) = (x_1 + 10x_2)^2 + 5(x_3 - x_4)^2 + (x_2 - 2x_3)^4 + 10(x_1 - x_4)^4$.
The initial point is $\boldsymbol{x}^{(0)} = [3, -1, 0, 1]^\top$. (Ans: see the textbook)

Example (minimize $f(\boldsymbol{x}) = x_1^2 + \gamma x_2^2$ by Newton's method)

Ans: (see the blackboard), $\boldsymbol{x}^* = (0, 0)$.

★ Newton's method can be divided into two steps

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \boldsymbol{F}(\boldsymbol{x}^{(k)})^{-1}\boldsymbol{g}^{(k)} \iff \begin{cases} \boldsymbol{d}^{(k)} = -\boldsymbol{F}(\boldsymbol{x}^{(k)})^{-1}\boldsymbol{g}^{(k)} \\ \boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \boldsymbol{d}^{(k)} \end{cases}$$

- Solve $\boldsymbol{F}(\boldsymbol{x}^{(k)})\boldsymbol{d}^{(k)} = -\boldsymbol{g}^{(k)}$ to obtain $\boldsymbol{d}^{(k)}$, which is an n-dimensional linear equations. $\boldsymbol{d}^{(k)}$ is called Newton direction.
- Determine the next iteration point $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \boldsymbol{d}^{(k)}$.

Newton's Method

Require: $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$ and $\boldsymbol{H}_0 \succ 0$. set $k := 0$;

1: **repeat**

2: solve linear equations $\boldsymbol{F}(\boldsymbol{x}^{(k)})\boldsymbol{d}^{(k)} = -\boldsymbol{g}^{(k)}$ for $\boldsymbol{d}^{(k)} \Longrightarrow$ Newton direction

3: $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \boldsymbol{d}^{(k)} \Longrightarrow$ updating iterative point.

4: **until** $\|\boldsymbol{g}^{(k)}\| \leq \epsilon$.

Example (apply Newton's method to minimize f)

$f(\boldsymbol{x}) = (x_1 - 2)^4 + (x_1 - 2x_2)^2$ with starting point $\boldsymbol{x}^{(0)} = [0, 3]^\top$.

Ans: S1. $\boldsymbol{g}^{(0)} = \begin{bmatrix} -44 \\ 24 \end{bmatrix}$ and $\boldsymbol{F}(\boldsymbol{x}^{(0)}) = \begin{bmatrix} 50 & -4 \\ -4 & 8 \end{bmatrix}$,

$$\boldsymbol{x}^{(1)} = \boldsymbol{x}^{(0)} - [\boldsymbol{F}(\boldsymbol{x}^{(0)})]^{-1}\boldsymbol{g}^{(0)} = \begin{bmatrix} 0 \\ 3 \end{bmatrix} + \begin{bmatrix} 50 & -4 \\ -4 & 8 \end{bmatrix}^{-1} \begin{bmatrix} -44 \\ 24 \end{bmatrix} = \begin{bmatrix} 0.67 \\ 0.33 \end{bmatrix}$$

S2. .....

k	1	2	3	4	5	6	7
$\boldsymbol{x}^{(k)}$	$[0, 3]$	$[0.67, 0.33]$	$[1.11, 0.56]$	$[1.41, 0.70]$	$[1.61, 0.80]$	$[1.74, 0.87]$	$[1.83, 0.91]$
$f(\boldsymbol{x}^{(k)})$	52	3.13	0.63	0.12	0.02	0.005	0.0009

Newton's Method

Example (apply Newton's method to minimize f)

$f(\boldsymbol{x}) = (x_1 - 2)^4 + (x_1 - 2)^2 x_2^2 + (x_2 + 1)^2$ with $\boldsymbol{x}^{(0)} = [1, 1]^\top$.

	$f(x_k)$
$x_0 = (1.0 \quad , \quad 1.0 \quad)^T$	6.0
$x_1 = (1.0 \quad , -0.5 \quad)^T$	1.5
$x_2 = (1.3913043 \quad , -0.69565217)^T$	4.09×10^{-1}
$x_3 = (1.7459441 \quad , -0.94879809)^T$	6.49×10^{-2}
$x_4 = (1.9862783 \quad , -1.0482081)^T$	2.53×10^{-3}
$x_5 = (1.9987342 \quad , -1.0001700)^T$	1.63×10^{-6}
$x_6 = (1.9999996 \quad , -1.0000016)^T$	2.75×10^{-12}

note: exact solution is $\boldsymbol{x}^* = [2, -1]^\top$.

analysis of Newton's method

- If the Hessian matrix $\boldsymbol{F}(\boldsymbol{x}^{(k)}) \not\succ 0$, Newton direction $\boldsymbol{d}^{(k)} = -\boldsymbol{F}(\boldsymbol{x}^{(k)})^{-1}\boldsymbol{g}^{(k)}$ is not necessarily descent direction.
- If the objective function f is quadratic, Newton's method converges to optima in one iteration for any initial point $\boldsymbol{x}^{(0)}$, i.e., the convergence rate of Newton's method on minimizing quadratic function is ∞.

Newton's Method

Newton's method for nonlinear equations

$$\boldsymbol{g}(\boldsymbol{x}) = \begin{bmatrix} g_1(\boldsymbol{x}) \\ g_2(\boldsymbol{x}) \\ \vdots \\ g_n(\boldsymbol{x}) \end{bmatrix} = 0,$$

where $\boldsymbol{x} \in \mathbb{R}^n$, $g_i : \mathbb{R}^n \to \mathbb{R}$ is continuous differentiable.
Let $\boldsymbol{F}(\boldsymbol{x}) \in \mathbb{R}^{n \times n}$ be the Jacobian of $\boldsymbol{g}$ at $\boldsymbol{x}$.
The (i,j)th element of $\boldsymbol{F}(\boldsymbol{x})$ is $\frac{\partial g_i}{\partial x_j}\boldsymbol{x}$.
iterative scheme: $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \boldsymbol{F}(\boldsymbol{x}^{(k)})^{-1}\boldsymbol{g}(\boldsymbol{x}^{(k)})$

★ if $n = 1$, iterative scheme reduces to: $x^{(k+1)} = x^{(k)} - \frac{g(x^{(k)})}{g'(x^{(k)})}$.

Example (apply Newton's method for nonlinear equation)

$g(x) = e^x - 1$ with starting point $x^{(0)} = -1$.

Ans:

k	0	1	2	3	4	5
$x^{(k)}$	-1.00000	0.71828	0.20587	0.01981	0.00019	0.00000
$g(x^{(k)})$	-0.63212	1.05091	0.22859	0.02000	0.00019	0.00000

★ Typically, 5-10 iterations can converge. Unfortunately, the classical Newton method may also diverge, and often does!

Convergence of Newton's Method

Theorem

If $f \in \mathcal{C}^3$, $\boldsymbol{x}^* \in \mathbb{R}^n$ satisfies $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$, 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 at least 2.

proof. Taylor expansion of gradient ∇f at $\boldsymbol{x}^{(0)}$:

$$\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}^{(0)}) - \boldsymbol{F}(\boldsymbol{x}^{(0)})(\boldsymbol{x} - \boldsymbol{x}^{(0)}) = O(\|\boldsymbol{x} - \boldsymbol{x}^{(0)}\|^2).$$

$\because$ $f \in \mathcal{C}^3$ and $\boldsymbol{F}(x^*)$ is invertible.

$\therefore$ $\exists\ \varepsilon > 0$, $c_1 > 0$, and $c_2 > 0$ such that: for all $\boldsymbol{x}^{(0)} \in \mathcal{B}(\boldsymbol{x}^*, \varepsilon)$, $\boldsymbol{x} \in \mathcal{B}(\boldsymbol{x}^*, \varepsilon)$,

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}^{(0)}) - \boldsymbol{F}(\boldsymbol{x}^{(0)})(\boldsymbol{x} - \boldsymbol{x}^{(0)})\| \le c_1 \|\boldsymbol{x} - \boldsymbol{x}^{(0)}\|^2. \tag{1}$$

$\because$ $\boldsymbol{F}(\boldsymbol{x})$ is invertible and $\|\boldsymbol{F}(\boldsymbol{x})^{-1}\| \le c_2$, by replacing $\boldsymbol{x}$ in (1) with $\boldsymbol{x}^*$ and using $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$, we have

$$\|\boldsymbol{F}(\boldsymbol{x}^{(0)})(\boldsymbol{x}^{(0)} - \boldsymbol{x}^*) - \nabla f(\boldsymbol{x}^{(0)})\| \le c_1 \|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\|^2.$$

Convergence of Newton's Method

$$\because \ \|\boldsymbol{x}^{(1)} - \boldsymbol{x}^*\| \overset{\text{why?}}{=\!=} \|\boldsymbol{x}^{(0)} - \boldsymbol{x}^* - \boldsymbol{F}(\boldsymbol{x}^{(0)})^{-1}\nabla f(\boldsymbol{x}^{(0)})\|$$
$$= \|\boldsymbol{F}(\boldsymbol{x}^{(0)})^{-1}[\boldsymbol{F}(\boldsymbol{x}^{(0)})(\boldsymbol{x}^{(0)} - \boldsymbol{x}^*) - \nabla f(\boldsymbol{x}^{(0)})]\|$$
$$\leq \|\boldsymbol{F}(\boldsymbol{x}^{(0)})^{-1}\| \cdot \|\boldsymbol{F}(\boldsymbol{x}^{(0)})(\boldsymbol{x}^{(0)} - \boldsymbol{x}^*) - \nabla f(\boldsymbol{x}^{(0)})\|$$
$$\leq \|\boldsymbol{F}(\boldsymbol{x}^{(0)})^{-1}\| \cdot c_1\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\|^2$$
$$\leq c_1c_2\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\|^2.$$

By choosing an $x^{(0)}$ satisfying $\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\| \leq \frac{\alpha}{c_1c_2}$ with $\alpha \in (0, 1)$, we have

$$\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^*\| \leq \alpha\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\|.$$

$\therefore \ \lim_{k\to\infty} \|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\| = 0 \Longrightarrow$ the sequence $\{\boldsymbol{x}^{(k)}\}$ converges to x^*.

$\because \ \|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*\| \leq c_1c_2\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|^2$, the order of convergence is at least 2.

★ If the initial point is close to the minima (maxima), Newton's method will have a good convergence. However, if it is far from the minima (maxima), Newton's method is not necessarily convergent.

Convergence of Newton's Method

Theorem (descent property of Newton's method)

Let $\boldsymbol{x}^{(k)}$ be the sequence generated by Newton's method for minimizing f. If $\boldsymbol{F}(\boldsymbol{x}^{(k)}) \succ \mathbf{0}$, and $\boldsymbol{g}^{(k)} = \nabla f(\boldsymbol{x}^{(k)}) \neq \mathbf{0}$, then the Newton direction $\boldsymbol{d}^{(k)} = -\boldsymbol{F}(\boldsymbol{x}^{(k)})^{-1}\boldsymbol{g}^{(k)} = \boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}$ is a descent direction in the sense that there exists a $\bar{\alpha} > 0$ such that $\alpha \in (0, \bar{\alpha})$, $f(\boldsymbol{x}^{(k)} + \alpha \boldsymbol{d}^{(k)}) < f(\boldsymbol{x}^{(k)})$.

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

$\because$ $\boldsymbol{F}(\boldsymbol{x}^{(k)})^{-1} \succ 0$ and $\boldsymbol{g}^{(k)} \neq \mathbf{0}$,

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

$\therefore$ $\exists \bar{\alpha} > 0$ such that $\phi(\alpha) < \phi(0)$ for all $\alpha \in (0, \bar{\alpha})$,

$\therefore$ $f(\boldsymbol{x}^{(k)} + \alpha d^{(k)}) < f(\boldsymbol{x}^{(k)})$ for all $\alpha \in (0, \bar{\alpha})$.

performance of Newton's method

- very fast when it converges (how fast? quadratic).
- may diverge (or worse, $\boldsymbol{F}(\boldsymbol{x}^{(k)})$ is nonsingular) when $\boldsymbol{x}^{(0)}$ is far from a nonsingular local optimum.

Question: how to modify Newton's method so that it converges globally?

Modification of Newton's Method

Newton's method with stepsize

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha_k \boldsymbol{F}(\boldsymbol{x}^{(k)})^{-1}\boldsymbol{g}^{(k)}, \text{ where } \alpha_k = \arg\min_{\alpha \geq 0} f(\boldsymbol{x}^{(k)} - \alpha \boldsymbol{F}(\boldsymbol{x}^{(k)})^{-1}\boldsymbol{g}^{(k)}).$$

★ Modified Newton method has the descent property $f(\boldsymbol{x}^{(k+1)}) < f(\boldsymbol{x}^{(k)})$.

Levenberg-Marquardt (L-M) modification

- If the Hessian $\boldsymbol{F}(\boldsymbol{x}^{(k)})$ is not necessarily positive definite, then $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha_k[\boldsymbol{F}(\boldsymbol{x}^{(k)}) + \mu_k \boldsymbol{I}]^{-1}\boldsymbol{g}^{(k)}$, where $\mu_k \geq 0$.
- If μ_k is large enough, e.g., $\mu_k > -\lambda_{\min}(\boldsymbol{F})$, then $\boldsymbol{G} = \boldsymbol{F} + \mu \boldsymbol{I} \succ 0$. Thus, the searching direction $\boldsymbol{d}^{(k)} = -(\boldsymbol{F}(\boldsymbol{x}^{(k)}) + \mu_k \boldsymbol{I})^{-1}\boldsymbol{g}^{(k)}$ is a descent direction.

Behavior of Parameter μ_k

1. if $\mu_k \to 0$, L-M modification can be made to approach the pure Newton's method.
2. if $\mu_k \to \infty$, L-M modification approaches the pure gradient descent method with a small step size.

Gauss-Newton Newton's Method

nonlinear least squares (NLS): e.g., applications to curve-fitting (see textbook)

$\min \sum_{i=1}^{m} [r_i(\boldsymbol{x})]^2$, where $r_i : \mathbb{R}^n \to \mathbb{R}$ $(i = 1, \cdots, m)$ are differentiable functions.

application of Newton's method to NLS

let $\boldsymbol{r} = [r_1, \cdots, r_m]^\top$ be vector-valued function from $\mathbb{R}^n \to \mathbb{R}^m$. The objective of NLS amounts to $f(\boldsymbol{x}) = \boldsymbol{r}(\boldsymbol{x})^\top \boldsymbol{r}(\boldsymbol{x})$.

- $\nabla f(\boldsymbol{x}) = 2\boldsymbol{J}(\boldsymbol{x})^\top \boldsymbol{r}(\boldsymbol{x})$ with $\boldsymbol{J}(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial r_1}{\partial x_1}(\boldsymbol{x}) & \cdots & \frac{\partial r_1}{\partial x_n}(\boldsymbol{x}) \\ \vdots & & \vdots \\ \frac{\partial r_m}{\partial x_1}(\boldsymbol{x}) & \cdots & \frac{\partial r_m}{\partial x_n}(\boldsymbol{x}) \end{bmatrix}$ as Jacobian of $\boldsymbol{r}$.
- $\nabla f(\boldsymbol{x})$ can also be rewritten by $(\nabla f(\boldsymbol{x}))_j = \frac{\partial f}{\partial x_j}(\boldsymbol{x}) = 2\sum_{i=1}^{m} r_i(\boldsymbol{x}) \frac{\partial r_i}{\partial x_j}(\boldsymbol{x})$.
- Hessian of f is $\boldsymbol{F}(\boldsymbol{x}) = 2[\boldsymbol{J}(\boldsymbol{x})^\top \boldsymbol{J}(\boldsymbol{x}) + \boldsymbol{S}(\boldsymbol{x})]$, where $\boldsymbol{S}(\boldsymbol{x})$ is a matrix whose (h, j)th element is $\sum_{i=1}^{m} r_i(\boldsymbol{x}) \frac{\partial^2 r_i}{\partial \boldsymbol{x}_k \partial \boldsymbol{x}_j}(\boldsymbol{x})$.

Gauss-Newton Newton's Method

recursion of Newton's method for solving NLS

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

★ As the matrix $\boldsymbol{S}(\boldsymbol{x})$ usually contains the 2nd derivative of $\boldsymbol{r}$, where the elements are all small.

Gauss-Newton method for NLS, i.e., neglecting $\boldsymbol{S}(\boldsymbol{x})$

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - [\boldsymbol{J}(\boldsymbol{x})^\top \boldsymbol{J}(\boldsymbol{x})]^{-1} \boldsymbol{J}(\boldsymbol{x})^\top \boldsymbol{r}(\boldsymbol{x}).$$

★ A pitfall of Gauss-Newton method is $\boldsymbol{J}(\boldsymbol{x})^\top \boldsymbol{J}(\boldsymbol{x})$ may not be positive definite.

★ By using Levenberg-Marquardt modification, Gauss-Newton method can be modified as

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - [\boldsymbol{J}(\boldsymbol{x})^\top \boldsymbol{J}(\boldsymbol{x}) + \mu_k \boldsymbol{I}]^{-1} \boldsymbol{J}(\boldsymbol{x})^\top \boldsymbol{r}(\boldsymbol{x}).$$

Homework

Exercise in the textbook: 9.4.