# 1. Nonlinear Equations

This lecture note excerpted parts from Michael Heath and Max Gunzburger.

Given function $\mathbf{f}$, we seek value $\mathbf{x}$ for which

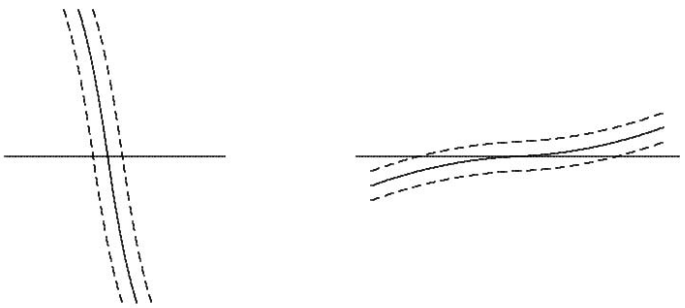
$$\mathbf{f}(\mathbf{x}) = 0$$

where $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^n$ is nonlinear.

http://www.cse.illinois.edu/heath/scicomp/notes/chap05.pdf p3-p9

Remark:

1. Conditioning of root finding problem is opposite to that for evaluating function.
   - Absolute condition number of root finding problem of $f : \mathbb{R} \to \mathbb{R}$ is $1/|f'(x^*)|$.
   - Root is ill-conditioned if tangent line is nearly horizontal. In particular, multiple root is ill-conditioned.
   - Absolute condition number of root finding problem of $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ is $\|(J_f)^{-1}(\mathbf{x}^*)\|$. Root is ill-conditioned if Jacobian matrix is nearly singular.



well-conditioned ill-conditioned

2. Numerical methods for finding the roots are iterative process
   Given an initial guess $\mathbf{x}_0$, compute $\mathbf{x}_1, \mathbf{x}_2, \cdots$ based on the algorithm
   - Accuracy of approximate solution
     small error $\|\hat{\mathbf{x}} - \mathbf{x}^*\| \approx 0$ or small residual $\|\mathbf{f}(\hat{\mathbf{x}})\| \approx 0$
     small residual implies accurate solution only if problem is well-conditioned
   - speed of convergence

**Definition**

*A sequence of iterates* $\{\mathbf{x}_k | k \geq 0\}$ *is said to converge with order* $r \geq 1$ *to* $\mathbf{x}^*$ *if*

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\| \leq C\|\mathbf{x}_k - \mathbf{x}^*\|^r, \quad k \geq 0$$

*for some* $C > 0$. *If* $r = 1$, *we require* $C < 1$.

- $r = 1$: linear convergence
- $r > 1$: superlinear convergence
- $r = 2$: quadratic convergence

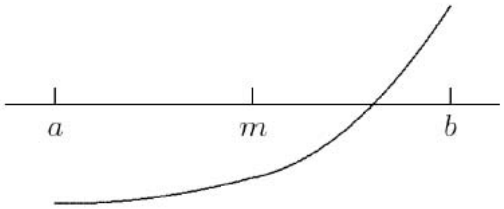
## 1.1 The Bisection Method

Assume that $f(x)$ is continuous on $[a, b]$ and

$$f(a)f(b) < 0.$$

Intermediate Value Theorem implies there exists at least one root in $[a, b]$.

**while** $((b-a) > tol)$ **do**
  $m = a + (b-a)/2$
  **if** $\text{sign}(f(a)) = \text{sign}(f(m))$ **then**
    $a = m$
  **else**
    $b = m$
  **end**
**end**




### [Speed of convergence]

At each iteration, length of interval containing solution reduced by half,

$$|x_k - x^*| \leq \left(\frac{1}{2}\right)^k (b-a) \quad : \text{ linear convergence with } C = \frac{1}{2}$$

- Bisection method makes no use of magnitudes of function values, only their signs
- Bisection is guaranteed to converge, but does very slowly
- achieving error tolerance of $\epsilon$ requires

$$\left\lceil \log_2\left(\frac{b-a}{\epsilon}\right) \right\rceil$$

  iterations, regardless of function $f$ involved.

Example: Use bisection method to find root of

$$f(x) = x^2 - 4sinx = 0$$

| $a$ | $f(a)$ | $b$ | $f(b)$ |
|---|---|---|---|
| 1.000000 | -2.365884 | 3.000000 | 8.435520 |
| 1.000000 | -2.365884 | 2.000000 | 0.362810 |
| 1.500000 | -1.739980 | 2.000000 | 0.362810 |
| 1.750000 | -0.873444 | 2.000000 | 0.362810 |
| 1.875000 | -0.300718 | 2.000000 | 0.362810 |
| 1.875000 | -0.300718 | 1.937500 | 0.019849 |
| 1.906250 | -0.143255 | 1.937500 | 0.019849 |
| 1.921875 | -0.062406 | 1.937500 | 0.019849 |
| 1.929688 | -0.021454 | 1.937500 | 0.019849 |
| 1.933594 | -0.000846 | 1.937500 | 0.019849 |
| 1.933594 | -0.000846 | 1.935547 | 0.009491 |
| 1.933594 | -0.000846 | 1.934570 | 0.004320 |
| 1.933594 | -0.000846 | 1.934082 | 0.001736 |
| ⋮ | ⋮ | ⋮ | ⋮ |
| 1.933750 | -0.000019 | 1.933754 | 0.000001 |
| 1.933752 | -0.000009 | 1.933754 | 0.000001 |
| 1.933753 | -0.000004 | 1.933754 | 0.000001 |

## 1.2 Newton's Method

Assume $f \in \mathcal{C}^2([a,b])$. Consider the Taylor's expansion of $f$ :

$$f(x) = f(x_k) + (x - x_k)f'(x_k) + \frac{(x - x_k)^2}{2!}f''(\xi),$$

where $\xi$ is in between $x$ and $x_k$. Set $x = x^*$ (the root of $f$), then

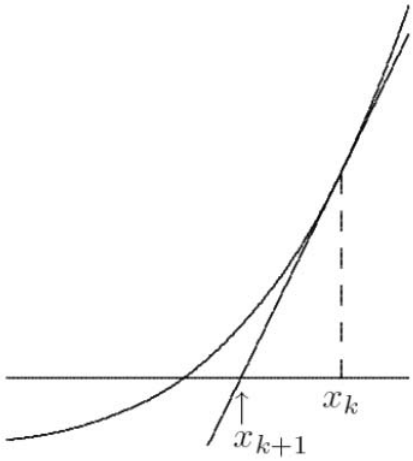
$$0 = f(x^*) = f(x_k) + (x^* - x_k)f'(x_k) + \frac{(x^* - x_k)^2}{2!}f''(\xi)$$
$$\Longrightarrow x^* = x_k - \frac{f(x_k)}{f'(x_k)} - \frac{(x^* - x_k)^2}{2!}\frac{f''(\xi)}{f'(x_k)}.$$

If $f'(x^*) \neq 0$, then

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \quad k \geq 1 \quad : \text{ Newton's method}$$

with

$$x_{k+1} - x^* = (x_k - x^*)^2 \frac{f''(\xi)}{2f'(x_k)}, \quad k \geq 1.$$



[Speed of convergence]

If $x_0$ is close enough to $x^*$ and $f'(x^*) \neq 0$ then

$$|x_{k+1} - x^*| \leq C\ |x_k - x^*|^2 \quad : \quad \text{quadratic convergence}$$

- Convergence of Newton's method for simple root is quadratic
- But iterations must start close enough to root to converge
- It requires knowledge of $f'(x)$ explicitly, which may not always possible

Example: Use Newton's method to find root of $f(x) = x^2 - 4sinx = 0$.

$$x_{k+1} = x_k - \frac{(x_k)^2 - 4sin(x_k)}{2x_k - 4cos(x_k)}$$

Taking $\mathbf{x}_0 = 3$ as starting value, we obtain

| $x_k$ | $f(x_k)$ | $f'(x_k)$ | $h$ |
|---|---|---|---|
| 3.000000 | 8.435520 | 9.959970 | -0.846942 |
| 2.153058 | 1.294772 | 6.505771 | -0.199019 |
| 1.954039 | 0.108438 | 5.403795 | -0.020067 |
| 1.933972 | 0.001152 | 5.288919 | -0.000218 |
| 1.933754 | 0.000000 | 5.287670 | 0.000000 |

- For multiple root, convergence rate of Newton's method is only linear, with constant $C = 1 - 1/m$, where $m$ is multiplicity of the root, e.g.

| $k$ | $f(x) = x^2 - 1$ | $f(x) = x^2 - 2x + 1$ |
|---|---|---|
| 0 | 2.0 | 2.0 |
| 1 | 1.25 | 1.5 |
| 2 | 1.025 | 1.25 |
| 3 | 1.0003 | 1.125 |
| 4 | 1.00000005 | 1.0625 |
| 5 | 1.0 | 1.03125 |

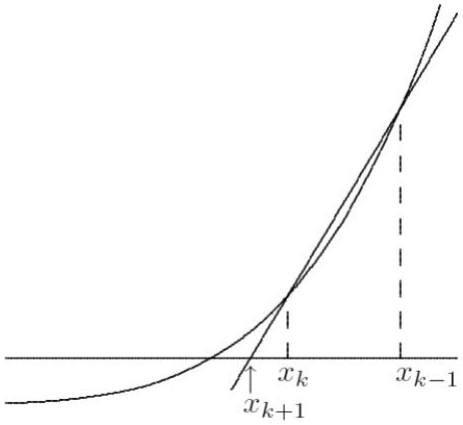
## 1.3 The Secant Method

- For each iteration, Newton's method requires evaluation of both function and its derivative, which may be inconvenient or expensive
- In secant method, derivative is approximated by finite difference using two successive iterates,

$$f'(x_k) \approx \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}.$$

so iteration becomes

$$x_{k+1} = x_k - f(x_k) \cdot \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})}, \quad k \geq 1 \quad : \text{ Secant method}$$




Remark:

1. The iterate $x_{k+1}$ in the Secant method is a solution to the linearized equation

$$l(x) = \left( \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}} \right) (x - x_k) + f(x_k) = 0$$

One usually views $l(x)$ as an approximation to the tangent line

$$l_t(x) = f'(x_k)(x - x_k) + f(x_k) = 0$$

2. $l(x)$ can be viewed as the linear interpolation of $f$ between the points $(x_{k-1}, f(x_{k-1}))$ and $(x_k, f(x_k))$. Then $x_{k+1}$ is the point where $l(x)$ intersects $x$-axis.

[Speed of convergence]

$$x_{k+1} - x^* = x_k - x^* + f(x_k) \cdot \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})},$$

yields

$$x_{k+1} - x^* = (x_k - x^*)(x_{k-1} - x^*)\frac{f[x_{k-1}, x_k, x^*]}{f[x_{k-1}, x_k]},$$

where $f[x_0, x_1] = \frac{f(x_1)-f(x_0)}{x_1-x_0}$ and $f[x_0, x_1, x_2] = \frac{f[x_1,x_2]-f[x_0,x_1]}{x_2-x_0}$.

Assume $f, f'$ and $f''$ are continuous in some interval containing $x^*$ and $f'(x^*) \neq 0$. Then if the initial guesses $x_0$ and $x_1$ are sufficiently close to $x^*$, then $x_k$ will converge to $x^*$ with order $r = (1 + \sqrt{5})/2 \approx 1.62$.

- Secant method converges faster than bisection method and slower than Newton's method
- It does not require a knowledge of $f'(x)$ and requires only one function evaluation per iterate
- Iterations must start close enough to root to converge

Example: Use secant method to find root of

$$f(x) = x^2 - 4sinx = 0$$

Taking $x^{(0)} = 1$ and $x^{(1)} = 3$ as starting guesses, we obtain

| $x_k$ | $f(x_k)$ | $h$ |
|---|---|---|
| 1.000000 | .2.365884 | |
| 3.000000 | 8.435520 | -1.561930 |
| 1.438070 | .1.896774 | 0.286735 |
| 1.724805 | .0.977706 | 0.305029 |
| 2.029833 | 0.534305 | -0.107789 |
| 1.922044 | .0.061523 | 0.011130 |
| 1.933174 | .0.003064 | 0.000583 |
| 1.933757 | 0.000019 | -0.000004 |
| 1.933754 | 0.000000 | 0.000000 |

## 1.4 Fixed-Point Iteration Methods

Fixed point of given function $g : \mathbb{R} \to \mathbb{R}$ is value $x$ such that

$$x = g(x)$$

Many iterative methods for solving nonlinear equations use fixed-point iteration scheme of form

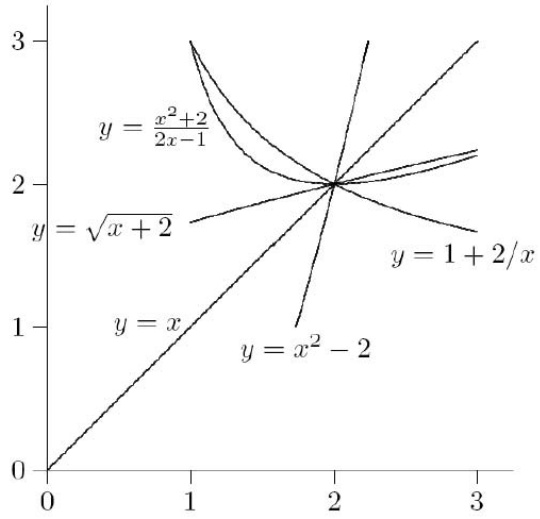
$$x_{k+1} = g(x_k), \quad k \geq 0.$$

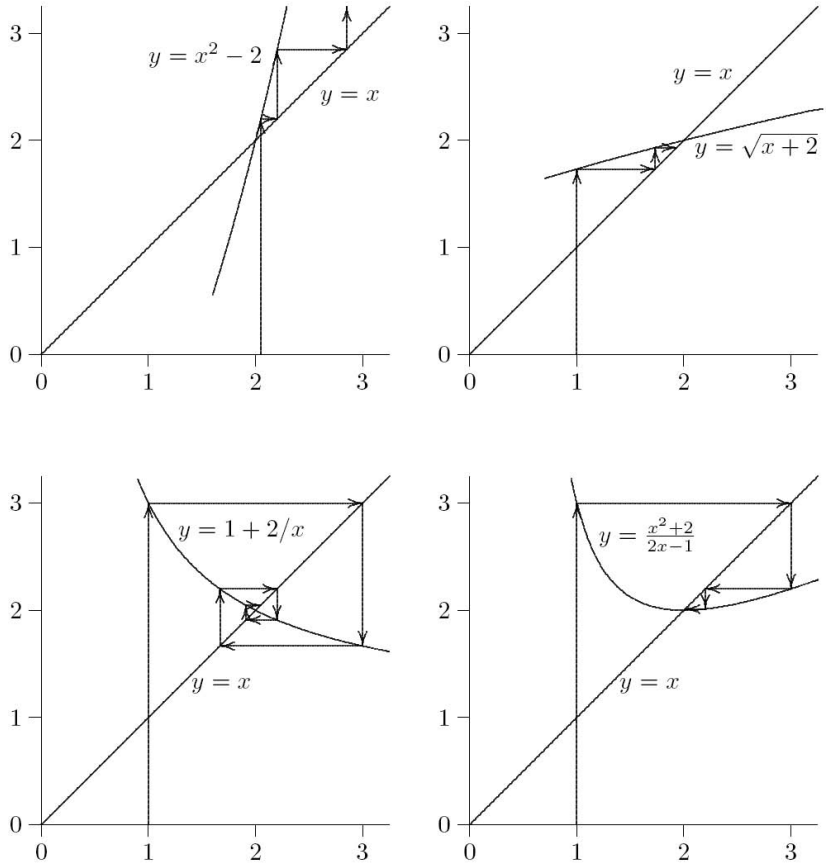
where fixed points for $g$ are solutions for $f(x) = 0$. Given equation $f(x) = 0$, there may be many equivalent fixed-point problems $x = g(x)$ with different choices for $g$.

Example: If $f(x) = x^2 - x - 2$, then fixed points of each of functions

$$\begin{aligned} g(x) &= x^2 - 2 \\ g(x) &= \sqrt{x+2} \\ g(x) &= 1 + \frac{2}{x} \\ g(x) &= \frac{x^2+2}{2x-1} \end{aligned}$$

are solutions to equation $f(x) = 0$.






**Theorem**

*Assume* $g(x) \in \mathcal{C}([a,b])$, $g([a,b]) \subset [a,b]$, *and*

$$\lambda = \max_{a \le x \le b} |g'(x)| < 1.$$

*Then* $x = g(x)$ *has a unique solution* $x^*$ *in* $[a,b]$. *Also, the iterates*

$$x_k = g(x_{k-1}), \quad k \ge 1$$

*will converge to* $x^*$ *for any choice of* $x_0$ *in* $[a,b]$ *and*

$$|x^* - x_k| \le \lambda^n |x^* - x_0| \le \frac{\lambda^n}{1-\lambda} |x_1 - x_0|.$$

*Moreover,*

$$\lim_{n\to\infty} \frac{x^* - x_{k+1}}{x^* - x_k} = \lim_{n\to\infty} \frac{g(x^*) - g(x_k)}{x^* - x_k} = \lim_{n\to\infty} \frac{(x^* - x_k) g'(\xi_n)}{x^* - x_k} = g'(x^*). \tag{1}$$

- Asymptotic convergence rate of fixed-point iteration is linear, with constant $C = |g'(x^*)|$
- But if $g'(x^*) = 0$, then convergence rate is at least quadratic
- Newton's method transforms nonlinear equation $f(x) = 0$ into fixed-point problem $x = g(x)$, where

$$g(x) = x - f(x)/f'(x)$$

- If $x^*$ is simple root, then $g'(x^*) = 0$ and hence Newton's method converges quadratically for simple root.

## 1.5 Systems of Nonlinear Equations

### [Intoduction]

For simplicity of presentation and ease of understanding, consider only two equations :

$$\mathbf{f}(\mathbf{x}) = \mathbf{0},$$

where $\mathbf{f}(\mathbf{x}) = \begin{bmatrix} f_1(x_1, x_2) \\ f_2(x_1, x_2) \end{bmatrix}$ and $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, that is

$$f_1(x_1, x_2) = 0, \qquad f_2(x_1, x_2) = 0. \tag{2}$$

Transform equations (2) into the following forms :

$$x_1 = g_1(x_1, x_2) \quad \text{and} \quad x_2 = g_2(x_1, x_2)$$
$$\text{e.g., } 0 = f_1(x_1, x_2) \Rightarrow x_1 = x_1 + f_1(x_1, x_2) \equiv g_1(x_1, x_2)$$

Denote its solution by

$$\mathbf{x}^* = \begin{bmatrix} x_1^* \\ x_2^* \end{bmatrix}$$

We study the fixed point iteration

$$\mathbf{x}_{k+1} = \mathbf{g}(\mathbf{x}_n),$$

where $x_{1,n+1} = g_1(x_{1,n}, x_{2,n})$ and $x_{2,n+1} = g_2(x_{1,n}, x_{2,n})$.

To analyze the convergence, consider the corresponding equations

$$x_1^* = g_1(x_1^*, x_2^*), \quad x_2^* = g_2(x_1^*, x_2^*).$$

By mean value theorem, we have

$$g_1(x_1^*, x_2^*) - g_1(x_{1,n}, x_{2,n}) = \frac{\partial g_1(\xi_{1,n}, \eta_{1,n})}{\partial x_1}(x_1^* - x_{1,n}) + \frac{\partial g_1(\xi_{1,n}, \eta_{1,n})}{\partial x_2}(x_2^* - x_{2,n})$$

which implies

$$x_1^* - x_{1,n+1} = \frac{\partial g_1(\xi_{1,n}, \eta_{1,n})}{\partial x_1}(x_1^* - x_{1,n}) + \frac{\partial g_1(\xi_{1,n}, \eta_{1,n})}{\partial x_2}(x_2^* - x_{2,n}),$$

where $(\xi_{1,n}, \eta_{1,n})$ is in between $\mathbf{x}^*$ and $\mathbf{x}_n$.

In matrix form, these error equations become

$$\begin{bmatrix} x_1^* - x_{1,n+1} \\ x_2^* - x_{2,n+1} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial g_1(\xi_{1,n}, \eta_{1,n})}{\partial x_1} & \dfrac{\partial g_1(\xi_{1,n}, \eta_{1,n})}{\partial x_2} \\ \dfrac{\partial g_2(\xi_{2,n}, \eta_{2,n})}{\partial x_1} & \dfrac{\partial g_2(\xi_{2,n}, \eta_{2,n})}{\partial x_2} \end{bmatrix} \begin{bmatrix} x_1^* - x_{1,n} \\ x_2^* - x_{2,n} \end{bmatrix}.$$

We can rewrite the above equation

$$\mathbf{x}^* - \mathbf{x}_{k+1} = G_n(\mathbf{x}^* - \mathbf{x}_n), \tag{3}$$

Denote the Jacobian matrix for $g_1, g_2$ by

$$\mathbf{g}'(\mathbf{x}) = \begin{bmatrix} \frac{\partial g_1(\mathbf{x})}{\partial x} & \frac{\partial g_1(\mathbf{x})}{\partial y} \\ \frac{\partial g_2(\mathbf{x})}{\partial x} & \frac{\partial g_2(\mathbf{x})}{\partial y} \end{bmatrix}$$

- If $\mathbf{x}_n$ is close to $\mathbf{x}^*$ then $G_n$ will be close to $\mathbf{g}'(\mathbf{x}^*)$. This will make the size or norm of $\mathbf{g}'(\mathbf{x}^*)$ crucial in analyzing the convergence in (3). The matrix $\mathbf{g}'(\mathbf{x}^*)$ plays the role of $g'(x^*)$ in the 1-d problem.

[Derivatives and Mean Value Theorem]

$f : \mathbb{R} \to \mathbb{R}$ is differentiable at $x$ if there exists $a$ such that

$$f'(x) = \lim_{t \to 0} \frac{1}{t}[f(x+t) - f(x)] = a$$

$$\lim_{t \to 0} \frac{1}{t}[f(x+t) - f(x) - ta] = 0$$

Note that if $f$ is differentiable then $f$ is continuous.

For $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^m$, we define

- Partial derivative (P-differentiable)
- Directional derivative (D-differentiable)
- Gateaux derivative (G-differentiable)
- Frechet derivtive (F-differentiable)

**Definition**

*A mapping* $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^m$ *is Gateaux differentiable at an interior point* $\mathbf{x}$ *of* $D$ *if there exists a linear mapping* $A \in L(\mathbb{R}^n, \mathbb{R}^m)$ *such that for any* $\mathbf{h} \in \mathbb{R}^n$

$$\lim_{t \to 0} \frac{1}{t} \|\mathbf{f}(\mathbf{x} + t\mathbf{h}) - \mathbf{f}(\mathbf{x}) - tA\mathbf{h}\| = 0$$

*If* $\mathbf{f}$ *is G-differentiable at* $\mathbf{x}$ *then the map* $A$ *is unique, which is called the Gateaux derivative and is denoted by* $A = \mathbf{f}'(\mathbf{x})$.

Remark:

1. $A$ is unique:

$$\begin{aligned} \|A_1\mathbf{h} - A_2\mathbf{h}\| &\leq \frac{1}{t}\|\mathbf{f}(\mathbf{x} + t\mathbf{h}) - \mathbf{f}(\mathbf{x}) - tA_2\mathbf{h}\| \\ &+ \frac{1}{t}\|\mathbf{f}(\mathbf{x} + t\mathbf{h}) - \mathbf{f}(\mathbf{x}) - tA_1\mathbf{h}\| \to 0 \end{aligned}$$

2. Let $\mathbf{f}'(\mathbf{x})_{ij} = A_{ij}$ and take $\mathbf{h} = \mathbf{e}^{(j)}$

$$\lim_{t \to 0} \frac{1}{t} \|\mathbf{f}(\mathbf{x} + t\mathbf{e}^{(j)}) - \mathbf{f}(\mathbf{x}) - tA\mathbf{e}^{(j)}\| = 0$$

$$\lim_{t \to 0} \frac{1}{t} |f_i(x + t\mathbf{e}^{(j)}) - f_i(x) - tA_{ij}\mathbf{e}^{(j)}| = 0 \text{ for } i = 1 \cdots m$$

   If the Gateaux derivative of $\mathbf{f}$ exists at $\mathbf{x}$ then $(\mathbf{f}'(\mathbf{x}))_{ij} = A_{ij} = \frac{\partial f_i}{\partial x_j}(\mathbf{x})$ for $i = 1 \cdots m$, $j = 1 \cdots n$. i.e. $\mathbf{f}'(\mathbf{x})$ is represented by Jacobian matrix

$$A = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n} \end{pmatrix}$$

3. Existence of Gateaux derivative implies the existence of directional derivatives and thus the partial derivatives. It does not imply the continuity.

**Definition**

*A mapping* $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^m$ *is Frechet differentiable at an interior point* $\mathbf{x}$ *of* $D$ *if there exists a linear mapping* $A \in L(\mathbb{R}^n, \mathbb{R}^m)$ *such that for any* $\mathbf{h} \in \mathbb{R}^n$

$$\lim_{\mathbf{h}\to 0} \frac{1}{\|\mathbf{h}\|} \|\mathbf{f}(\mathbf{x}+t\mathbf{h}) - \mathbf{f}(\mathbf{x}) - A\mathbf{h}\| = 0$$

*If* $\mathbf{f}$ *is F-differentiable at* $\mathbf{x}$ *then the map* $A$ *is unique, which is called the Frechet derivative and is denoted by* $A = \mathbf{f}'(\mathbf{x})$.

Remark:

1. If $\mathbf{f}$ is Frechet differentiable then it is Gateaux differentiable.
2. If $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^m$ is F-differentiable at $\mathbf{x}$ then $\mathbf{f}$ is continuous at $\mathbf{x}$: $\forall \epsilon > 0$, $\exists \delta > 0$ such that $\|\mathbf{h}\| \le \delta \ \to \ \|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - \mathbf{f}'(\mathbf{x})\mathbf{h}\| \le \epsilon\|\mathbf{h}\|$. Then $\|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x})\| \le (\|\mathbf{f}'(\mathbf{x})\| + \epsilon)\|\mathbf{h}\|$.
3. If $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^m$ is G-differentiable at $\mathbf{x}$ and $\mathbf{f}'(x)$ is continuous at $\mathbf{x}$ then $\mathbf{f}$ is F-differentiable at $\mathbf{x}$ and both derivatives coincide.

If $f : D \subset \mathbb{R}^n \to \mathbb{R}^1$ is differentiable (G or F) on a convex set $D_0 \subset D$ then, given $\mathbf{x}$, $\mathbf{y} \subset D_0$ there exists a $t$ such that

$$f(\mathbf{y}) - f(\mathbf{x}) = f'(t\mathbf{x} + (1-t)\mathbf{y})(\mathbf{y} - \mathbf{x})$$

which does not hold for $\mathbf{f} : D \subset R^n \to R^m$ for $m > 1$.

**Proposition**

*Assume that* $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^m$ *is G-differentiable on a convex set* $D_0 \subset D$. *Then* $\forall \mathbf{x}, \mathbf{y} \in D_0$

$$\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\| \le \sup_{0\le t\le 1} \|\mathbf{f}'(t\mathbf{x} + (1-t)\mathbf{y})\| \|\mathbf{x} - \mathbf{y}\|$$

**Proposition**

*Let* $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^m$ *have a continuous G-derivative at each point of a convex set* $D_0 \subset D$. *Then if* $\mathbf{x}, \mathbf{y} \in D_0$,

$$\mathbf{f}(\mathbf{y}) - \mathbf{f}(\mathbf{x}) = \int_0^1 \mathbf{f}'(t\mathbf{x} + (1-t)\mathbf{y})(\mathbf{y} - \mathbf{x})dt$$

## 1.6 Fixed Point Iteration

Fixed-point problem for $\mathbf{g} : D \subset \mathbb{R}^n \to \mathbb{R}^n$ is to find vector $x$ such that

$$\mathbf{x} = \mathbf{g}(\mathbf{x})$$

where $\mathbf{g} : D \subset \mathbb{R}^n \to \mathbb{R}^n$ is chosen such that solutions of $\mathbf{x} = \mathbf{g}(\mathbf{x})$ coincide with the zeros of $\mathbf{f}(\mathbf{x})$. Corresponding fixed-point iteration is

$$\mathbf{x}_{k+1} = \mathbf{g}(\mathbf{x}_k) \tag{4}$$

Example:

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

Consider two different choices of $\mathbf{g}(\mathbf{x})$:

$$\mathbf{g}_1(\mathbf{x}) = \begin{pmatrix} x_2^2 - 1 \\ x_1^3 \end{pmatrix} \quad \mathbf{g}_2(\mathbf{x}) = \begin{pmatrix} x_2^{1/3} \\ \sqrt{1 + x_1} \end{pmatrix}$$

Then fixed-point iteration using $\mathbf{g}_1(\mathbf{x})$ does not converge but, using $\mathbf{g}_2(\mathbf{x})$ converge.

**Definition**

*Let $\mathbf{g} : D \subset \mathbb{R}^n \to \mathbb{R}^n$. Then $\mathbf{x}^*$ is a point of attraction of the iteration $\mathbf{x}_{k+1} = \mathbf{g}(\mathbf{x}_k)$ if there is open neighborhood $S \subset D$ of $\mathbf{x}^*$ such that for any $\mathbf{x}_0 \in S$, the iterates $\mathbf{x}_k \in D$ for all $k$ and $\mathbf{x}_k \to \mathbf{x}^*$.*

**Theorem(Local Convergence)**

*Let $\mathbf{g} : D \subset \mathbb{R}^n \to \mathbb{R}^n$ and suppose that there is a ball $S = S(\mathbf{x}^*, \delta) \in D$ and a constant $\alpha < 1$ such that*

$$\|\mathbf{g}(\mathbf{x}) - \mathbf{x}^*\| \le \alpha \|\mathbf{x} - \mathbf{x}^*\| \quad \forall \mathbf{x} \in S$$

*Then for any $\mathbf{x}_0 \in S$ the iterates defined by (4) lie in $S$ and converges to $\mathbf{x}^*$. Hence, $\mathbf{x}^*$ is a point of attraction of the iteration (4).*

pf: Let $\mathbf{x}_0 \in S$.

$$\begin{aligned} \|\mathbf{x}_1 - \mathbf{x}^*\| &= \|\mathbf{g}(\mathbf{x}_0) - \mathbf{x}^*\| \le \alpha \|\mathbf{x}_0 - \mathbf{x}^*\| \to \mathbf{x}_1 \in S \\ \|\mathbf{x}_{k+1} - \mathbf{x}^*\| &= \|\mathbf{g}(\mathbf{x}_k) - \mathbf{x}^*\| \le \alpha \|\mathbf{x}_k - \mathbf{x}^*\| \le \alpha^{k+1} \|\mathbf{x}_0 - \mathbf{x}^*\| \to 0 \end{aligned}$$

Hence, $\mathbf{x}_k \in S$ for all $k$ and $\mathbf{x}_k \to x^*$.

**Theorem**

*Assume that* $\mathbf{g} : D \subset \mathbb{R}^n \to \mathbb{R}^n$ *has a fixed point* $\mathbf{x}^* \in D$ *and that* $\mathbf{g}$ *has a Frechet derivative* $\mathbf{g}'$ *at* $\mathbf{x}^*$. *If the spectral radius of* $\mathbf{g}'(\mathbf{x}^*)$ *satisfies* $\rho(\mathbf{g}'(\mathbf{x}^*)) = \sigma < 1$ *then* $\mathbf{x}^*$ *is a point of attraction of the iteration (4).*

pf: Since $inf\|\mathbf{g}'(\mathbf{x}^*)\| = \rho(\mathbf{g}'(\mathbf{x}^*)) = \sigma$, there is a norm such that

$$\|\mathbf{g}'(\mathbf{x}^*)\| \leq \sigma + \epsilon \quad \forall \epsilon$$

$\mathbf{g}$ is F-differentiable at $\mathbf{x}^*$:

$$\lim_{|\mathbf{h}|\to 0} \frac{1}{\|\mathbf{h}\|} \|\mathbf{g}(\mathbf{x}^* + \mathbf{h}) - \mathbf{g}(\mathbf{x}^*) - \mathbf{g}'(\mathbf{x}^*)\mathbf{h}\| = 0$$

i.e. there exists a $\delta = \delta(\mathbf{x}^*, \epsilon)$ such that $S = S(\mathbf{x}^*, \delta) \in D$ and

$$\|\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{x}^*) - \mathbf{g}'(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}^*)\| \leq \epsilon\|\mathbf{x} - \mathbf{x}^*\| \quad \forall \mathbf{x} \in S$$

Then

$$\begin{aligned} \|\mathbf{g}(\mathbf{x}) - \mathbf{x}^*\| &\leq \|\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{x}^*) - \mathbf{g}'(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}^*)\| + \|\mathbf{g}'(\mathbf{x}^*)\|\|\mathbf{x} - \mathbf{x}^*\| \\ &\leq (\sigma + 2\epsilon)\|\mathbf{x} - \mathbf{x}^*\| \end{aligned}$$

Choose $\epsilon$ such that $\sigma + 2\epsilon < 1$ completes the proof.

**Definition**

*A mapping* $\mathbf{g} : D \subset \mathbb{R}^n \to \mathbb{R}^n$ *is a contraction on a set* $D_0 \in D$ *if there is* $\alpha < 1$ *such that*

$$\|\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{y})\| \leq \alpha\|\mathbf{x} - \mathbf{y}\| \quad \forall \mathbf{x}, \mathbf{y} \in D_0$$

**Theorem (Contraction Mapping Theorem)**

*Suppose* $\mathbf{g} : D \subset \mathbb{R}^n \to \mathbb{R}^n$ *is a contraction on a closed set* $D_0 \subset D$ *and that* $\mathbf{g}(D_0) \subset D_0$ *(i.e. if* $\mathbf{x} \in D_0$ *then* $\mathbf{g}(\mathbf{x}) \in D_0$ *also). Then* $\mathbf{g}$ *has a unique fixed point* $\mathbf{x}^* \in D_0$. *Moreover, for any* $\mathbf{x}_0 \in D_0$, *the iterates defined by (4) converges to* $\mathbf{x}^*$. *And*

$$\begin{aligned} \|\mathbf{x}_k - \mathbf{x}^*\| &\leq \frac{\alpha}{1-\alpha}\|\mathbf{x}_k - \mathbf{x}_{k-1}\| \quad k = 1, 2, ... \\ \|\mathbf{x}_k - \mathbf{x}^*\| &\leq \frac{\alpha^k}{1-\alpha}\|\mathbf{g}(\mathbf{x}_0) - \mathbf{x}_0\| \quad k = 1, 2, ... \end{aligned}$$

pf: Let $\mathbf{x}_0 \in D_0$. Then $\mathbf{x}_k \in D_0$ since $\mathbf{g}(D_0) \subset D_0$.

$$\begin{aligned}
\|\mathbf{x}_{k+1} - \mathbf{x}_k\| &= \|\mathbf{g}(\mathbf{x}_k) - \mathbf{g}(\mathbf{x}_{k-1})\| \le \alpha\|\mathbf{x}_k - \mathbf{x}_{k-1}\| \\
\|\mathbf{x}_{k+p} - \mathbf{x}_k\| &\le \sum_{i=1}^{p} \|\mathbf{x}_{k+i} - \mathbf{x}_{k+i-1}\| \\
&\le (\alpha^p + \cdots + \alpha^2 + \alpha)\|\mathbf{x}_k - \mathbf{x}_{k-1}\| \\
&\le \frac{\alpha}{1-\alpha}\|\mathbf{x}_k - \mathbf{x}_{k-1}\| \le \frac{\alpha^k}{1-\alpha}\|\mathbf{x}_1 - \mathbf{x}_0\|
\end{aligned}$$

Therefore $\mathbf{x}_k$ is a Cauchy sequence in $D_0$ and converges to a limit $\mathbf{x}^* \in D_0$.

$$\|\mathbf{x}^* - \mathbf{g}(\mathbf{x}^*)\| \le \|\mathbf{x}^* - \mathbf{x}_{k+1}\| + \|\mathbf{g}(\mathbf{x}_k) - \mathbf{g}(\mathbf{x}^*)\| \le \|\mathbf{x}^* - \mathbf{x}_{k+1}\| + \alpha\|\mathbf{x}_k - \mathbf{x}^*\| \to 0$$

Hence, $\mathbf{x}^*$ is a fixed point and this is unique since

$$\|\mathbf{x}^* - \mathbf{y}^*\| = \|\mathbf{g}(\mathbf{x}^*) - \mathbf{g}(\mathbf{y}^*)\| \le \alpha\|\mathbf{x}^* - \mathbf{y}^*\| \text{ with } \alpha < 1 \to \mathbf{x}^* = \mathbf{y}^*$$

Remark:
We note that the contraction property is norm dependent so that $\mathbf{g}$ may be a contraction for one norm but not for another. Contraction mapping theorem requires that $\mathbf{g}(D_0) \subset D_0$ on which $\mathbf{g}$ is a contraction and global convergence theorem is given when $D_0 = \mathbb{R}^n$.

**Proposition**

*Assume that $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^n$ has a G-derivative which satisfies $\|\mathbf{f}'(x)\| \le \alpha < 1$ for all $\mathbf{x}$ in a convex set $D_0 \in D$. Then $\mathbf{f}$ is a contraction on $D_0$.*

**Proposition**

*Assume that $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^n$ has a continuous Frechet derivative in an open neighborhood $S$ of $\mathbf{x}$ and that $\rho(\mathbf{f}'(\mathbf{x})) < 1$. Then there is another neighborhood $S_1$ of $\mathbf{x}$ and a norm on $R^n$ such that $\mathbf{f}$ is a contraction on $S_1$.*

## 1.7 Newton's Method

We consider the fixed point iterations of the form

$$\mathbf{x}_{k+1} = \mathbf{x}_k - (A\mathbf{x}_k)^{-1}\mathbf{f}(\mathbf{x}_k) \tag{5}$$

i.e. $\mathbf{g}(\mathbf{x}) = \mathbf{x} - (A\mathbf{x})^{-1}\mathbf{f}(\mathbf{x})$ for the solutions of $\mathbf{f}(\mathbf{x}) = 0$. The special case

$$A\mathbf{x} = \mathbf{f}'(\mathbf{x})$$

is Newton's method:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{f}'(\mathbf{x}_k)^{-1}\mathbf{f}(\mathbf{x}_k)$$

Lemma: Let $A \in L(\mathbb{R}^n, \mathbb{R}^n)$ be a nonsingular matrix. Suppose $B \in L(\mathbb{R}^n, \mathbb{R}^n)$ is such that $\|A^{-1}\|\|B\| < 1$. Then $A + B$ is nonsingular and

$$\|(A+B)^{-1}\| \le \frac{\|A^{-1}\|}{1 - \|A^{-1}\|\|B\|}$$

**Proposition**

*Suppose that $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^n$ is F-differentiable at $\mathbf{x}^* \in D$ for which $\mathbf{f}(\mathbf{x}^*) = 0$. Let $A : S_0 \to L(\mathbb{R}^n, \mathbb{R}^n)$ be defined on an open neighborhood $S_0 \subset D$ of $\mathbf{x}^*$ and let $A$ be continuous at $\mathbf{x}^*$ and $A\mathbf{x}^*$ be nonsingular. Then there exists a closed ball $\overline{S}(\mathbf{x}^*, \delta) \subset S_0$, $\delta > 0$ on which $\mathbf{g} : \overline{S} \to \mathbb{R}^n$, $\mathbf{g}(\mathbf{x}) = \mathbf{x} - (A\mathbf{x})^{-1}\mathbf{f}(\mathbf{x})$, is well-defined for $\mathbf{x} \in \overline{S}$. Moreover $\mathbf{g}$ is F-differentiable at $\mathbf{x}^*$ and*

$$\mathbf{g}'(\mathbf{x}^*) = I - (A\mathbf{x}^*)^{-1}\mathbf{f}'(\mathbf{x}^*)$$

**Corollary**

*Assume the hypotheses of previous Proposition hold. Suppose that*

$$\rho(\mathbf{g}'(\mathbf{x}^*)) = \rho(I - (A\mathbf{x}^*)^{-1}\mathbf{f}'(\mathbf{x}^*)) = \sigma < 1$$

*Then $\mathbf{x}^*$ is a point of attraction of the iteration (5).*

**Theorem(Newton Attraction Theorem)**

*Assume that* $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^n$ *is F-differentiable on an open neighborhood* $S_0 \subset D$ *of* $\mathbf{x}^* \in D$ *for which* $\mathbf{f}(\mathbf{x}^*) = 0$. *Also assume that* $\mathbf{f}'$ *is continuous at* $\mathbf{x}^*$ *and that* $\mathbf{f}'(\mathbf{x}^*)$ *is nonsingular. Then* $\mathbf{x}^*$ *is a point of attraction of the Newton iteration*

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

pf: Use Proposition with $A(\mathbf{x}) = \mathbf{f}'(\mathbf{x})$ and Corollary applies since $\mathbf{g}'(\mathbf{x}^*) = 0$.

**Theorem**

*Assume that the hypotheses of Newton attraction theorem hold. Then for the point of attraction of the Newton iteration, we have*

$$\lim_{k\to\infty} \frac{\|\mathbf{x}_{k+1} - \mathbf{x}^*\|}{\|\mathbf{x}_k - \mathbf{x}^*\|} = 0$$

*Moreover, if for some constant* $\hat{C}$

$$\|\mathbf{f}'(x) - \mathbf{f}'(\mathbf{x}^*)\| \le \hat{C}\|\mathbf{x} - \mathbf{x}^*\| \tag{6}$$

*for all* $\mathbf{x}$ *in some neighborhood of* $\mathbf{x}^*$, *then there exists a positive constant* $C$ *such that*

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\| \le C\|\mathbf{x}_k - \mathbf{x}^*\|^2$$

pf:

$$\lim_{k\to\infty} \frac{\|\mathbf{x}_{k+1} - \mathbf{x}^*\|}{\|\mathbf{x}_k - \mathbf{x}^*\|} = \lim_{k\to\infty} \frac{\|\mathbf{g}\mathbf{x}_k - \mathbf{g}\mathbf{x}^* - \mathbf{g}'(\mathbf{x}^*)(\mathbf{x}_k - \mathbf{x}^*)\|}{\|\mathbf{x}_k - \mathbf{x}^*\|} = 0$$

Let $\mathbf{x}_k$ for $k \ge k_0$ belong to the neighborhood of $\mathbf{x}^*$ in which (6) hold. Consider the convex set consisting of segment between $\mathbf{x}_k$ and $\mathbf{x}^*$.

$$\|\mathbf{f}(\mathbf{x}_k) - \mathbf{f}(\mathbf{x}^*) - \mathbf{f}'(\mathbf{x}^*)(\mathbf{x}_k - \mathbf{x}^*)\| \le \frac{\hat{C}}{2}\|\mathbf{x}_k - \mathbf{x}^*\|^2$$

$$\begin{aligned} (\because) \quad lhs &= \|\int_0^1 \mathbf{f}'(\mathbf{x}^* + t(\mathbf{x}_k - \mathbf{x}^*))(\mathbf{x}_k - \mathbf{x}^*) - \mathbf{f}'(\mathbf{x}^*)(\mathbf{x}_k - \mathbf{x}^*)dt\| \\ &\le \|\mathbf{x}_k - \mathbf{x}^*\| \int_0^1 \hat{C}t\|\mathbf{x}_k - \mathbf{x}^*\|dt \end{aligned}$$

$$\begin{aligned}\|\mathbf{x}_{k+1}-\mathbf{x}^*\| &= \|\mathbf{g}(\mathbf{x}_k)-\mathbf{x}^*\| = \|\mathbf{x}_k-\mathbf{f}'(\mathbf{x}_k)^{-1}\mathbf{f}(\mathbf{x}_k)-\mathbf{x}^*\| \\ &\le \|-\mathbf{f}'(\mathbf{x}_k)^{-1}(\mathbf{f}(\mathbf{x}_k)-\mathbf{f}(\mathbf{x}^*)-\mathbf{f}'(\mathbf{x}^*)(\mathbf{x}_k-\mathbf{x}^*))\| \\ &+ \|\mathbf{f}'(\mathbf{x}_k)^{-1}(\mathbf{f}'(\mathbf{x}_k)-\mathbf{f}'(\mathbf{x}^*))(\mathbf{x}_k-\mathbf{x}^*)\| \\ &\le \left(\frac{\hat{C}}{2}+\hat{C}\right)\|\mathbf{f}'(\mathbf{x}_k)^{-1}\|\|\mathbf{x}_k-\mathbf{x}^*\|^2\end{aligned}$$

Let $\|\mathbf{f}'(\mathbf{x}^*)^{-1}\| = \beta$ and apply lemma with $A = \mathbf{f}'(\mathbf{x}^*), B = -\mathbf{f}'(\mathbf{x}^*) + \mathbf{f}'(\mathbf{x})$

$$\|\mathbf{f}'(\mathbf{x}^*)^{-1}\|\|\mathbf{f}'(\mathbf{x})-\mathbf{f}'(\mathbf{x}^*)\| \le \beta\epsilon < \frac{1}{2}$$

since $\mathbf{f}'$ is continuous. Then $\mathbf{f}'(\mathbf{x})$ is invertible and

$$\|\mathbf{f}'(\mathbf{x})^{-1}\| \le 2\|\mathbf{f}'(\mathbf{x}^*)^{-1}\|$$

which completes the proof.

Remark:

1. Convergence rate of Newton's method for nonlinear systems is quadratic, provided Jacobian matrix $\mathbf{f}'(\mathbf{x})$ is nonsingular, but it must be started close enough to solution to converge
$$\|\mathbf{x}^*-\mathbf{x}_{k+1}\|_\infty \le C\|\mathbf{x}^*-\mathbf{x}_k\|_\infty^2$$
2. For general iteration $\mathbf{x}_{k+1} = \mathbf{g}(\mathbf{x}_k)$ in local convergence theorem,
$$\|\mathbf{x}_{k+1}-\mathbf{x}^*\| = \|\mathbf{g}(\mathbf{x}_k)-\mathbf{x}^*\| \le \alpha\|\mathbf{x}_k-\mathbf{x}^*\|$$
for $k$ sufficiently large
$$\lim_{k\to\infty}\frac{\|\mathbf{x}_{k+1}-\mathbf{x}^*\|}{\|\mathbf{x}_k-\mathbf{x}^*\|} \le \alpha$$
It is contrasted with superlinear convergence of Newton's method.
3. In practice, we solve
$$\mathbf{f}'(\mathbf{x}_k)\delta\mathbf{x}_k = -\mathbf{f}(\mathbf{x}_k), \qquad \mathbf{x}_{k+1} = \mathbf{x}_k + \delta\mathbf{x}_k$$
Cost per iteration of Newton's method for dense problem in n dimensions is: Computing Jacobian matrix costs $n^2$ scalar function evaluations and solving linear system costs $O(n^3)$ operations. We may modify the Newton's method to to reduce cost with loss of quadratic convergence.

- Simplified Newton iteration

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{f}'(\mathbf{x}_0)^{-1}\mathbf{f}(\mathbf{x}_k)$$

  requires only forward and backward solving for each iteration.

- Newton-Gauss-Seidel method

$$\mathbf{f}'(\mathbf{x}) = D(\mathbf{x}) - L(\mathbf{x}) - U(\mathbf{x})$$
$$\mathbf{x}_{k+1} = \mathbf{x}_k - (D\mathbf{x}_k - L\mathbf{x}_k)^{-1}\mathbf{f}(\mathbf{x}_k)$$

  requires the solution of a triangular system.

**Theorem(Semi-Local Convergence)**

*Let $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^n$ be F-differentiable on a convex set $D_0 \subset D$ and assume that for some constant $\gamma > 0$*

$$\|\mathbf{f}'(\mathbf{x}) - \mathbf{f}'(\mathbf{y})\| \le \gamma\|\mathbf{x} - \mathbf{y}\| \quad \forall \mathbf{x}, \mathbf{y} \in D_0$$

*Suppose that there is a point $\mathbf{x}_0 \in D_0$ such that $\mathbf{f}'(\mathbf{x}_0)$ is invertible. Moreover suppose for some constants $\beta$, $\eta$*

$$\|\mathbf{f}'(\mathbf{x}_0)^{-1}\| \le \beta, \quad \|\mathbf{f}'(\mathbf{x}_0)^{-1}\mathbf{f}(\mathbf{x}_0)\| \le \eta, \quad \alpha = \beta\gamma\eta \le \frac{1}{2}$$

*Set*

$$t^* = \frac{1 - (1 - 2\alpha)^{\frac{1}{2}}}{\beta\gamma}, \quad \hat{t} = \frac{1 + (1 - 2\alpha)^{\frac{1}{2}}}{\beta\gamma}$$

*and assume that $\bar{S}(x^{(0)}, t^*) \subset D_0$. Then the Newton iteration*

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{f}'(\mathbf{x}_k)^{-1}\mathbf{f}(\mathbf{x}_k), \quad k = 0, 1, \cdots$$

*are well-defined in $\bar{S}(\mathbf{x}_0, t^*)$ and converges to a solution $\mathbf{x}^*$ of $\mathbf{f}(\mathbf{x}) = 0$ which is unique in $S(x^{(0)}, \hat{t}) \bigcap D_0$. Moreover we have*

$$\|\mathbf{x}^* - \mathbf{x}_1\| \le 2\beta\gamma\|\mathbf{x}_1 - \mathbf{x}_0\|^2$$

$$\|\mathbf{x}^* - \mathbf{x}_k\| \le \frac{(2\alpha)^{2^k}}{\beta\gamma 2^k}$$

The mapping $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^n$ is called convex on the convex set $D_0 \subset D$ if

$$\mathbf{f}(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \le \lambda\mathbf{f}(\mathbf{x}) + (1-\lambda)\mathbf{f}(\mathbf{y})$$

for all $\mathbf{x}, \mathbf{y} \in D_0$ and $\lambda \in [0, 1]$.

Lemma: Let $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^n$ be F-differentiable on the convex set $D_0 \subset D$. Then $\mathbf{f}$ is convex on $D_0$ iff

$$\mathbf{f}(\mathbf{y}) - \mathbf{f}(\mathbf{x}) \ge \mathbf{f}'(\mathbf{x})(\mathbf{y} - \mathbf{x}) \quad \forall \mathbf{x}, \mathbf{y} \in D_0$$

**Theorem(Global Convergence)**

*Let $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^n$ be continuously F-differentiable and convex on $\mathbb{R}^n$. Suppose that $\mathbf{f}'(\mathbf{x})^{-1}$ exists for all $\mathbf{x} \in \mathbb{R}^n$ and that $\mathbf{f}'(\mathbf{x})^{-1} \ge 0$ for all $\mathbf{x} \in \mathbb{R}^n$. Let $\mathbf{f}(\mathbf{x}) = 0$ have a solution $\mathbf{x}^*$. Then $\mathbf{x}^*$ is unique and the Newton iteration converges to $\mathbf{x}^*$ for any initial choice $\mathbf{x}_0 \in \mathbb{R}^n$. Moreover, for all $k > 0$,*

$$\mathbf{x}^* \le \mathbf{x}_{k+1} \le \mathbf{x}_k$$

pf: Let $\mathbf{x}_0 \in \mathbb{R}^n$. $\mathbf{x}_1 = \mathbf{x}_0 - \mathbf{f}'(\mathbf{x}_0)^{-1}\mathbf{f}(\mathbf{x}_0)$.

$$\mathbf{f}(\mathbf{x}_1) - \mathbf{f}(\mathbf{x}_0) \ge \mathbf{f}'(\mathbf{x}_0)(\mathbf{x}_1 - \mathbf{x}_0) = -\mathbf{f}(\mathbf{x}_0) \to \mathbf{f}(\mathbf{x}_1) \ge 0$$

But, $0 = \mathbf{f}(\mathbf{x}^*) \ge \mathbf{f}(\mathbf{x}_1) + \mathbf{f}'(\mathbf{x}_1)(\mathbf{x}^* - \mathbf{x}_1)$. Since $\mathbf{f}'(\mathbf{x}_1)^{-1} \ge 0$, $\mathbf{f}'(\mathbf{x}_1)^{-1}\mathbf{f}(\mathbf{x}_1) + \mathbf{f}'(\mathbf{x}_1)^{-1}\mathbf{f}'(\mathbf{x}_1)(\mathbf{x}^* - \mathbf{x}_1) \le 0$ and $\mathbf{x}^* \le \mathbf{x}_1$. Similarly

$$\mathbf{x}^* \le \mathbf{x}_k \text{ and } \mathbf{f}(\mathbf{x}_k) \ge 0$$

Then $\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{f}'(\mathbf{x}_k)^{-1}\mathbf{f}(\mathbf{x}_k) \le \mathbf{x}_k$. Since $\{x_{i,k}\}$ is monotone decreasing and bounded below by $x_i^*$, $x_{i,k} \to y_i$ as $k \to \infty$.
Note that $\mathbf{f}$ is F-differentiable, continuous and $\mathbf{f}'(\mathbf{x})$ is continuous.

$$\mathbf{f}(\mathbf{y}) = \lim_{k\to\infty} \mathbf{f}(\mathbf{x}_k) = \lim_{k\to\infty} \mathbf{f}'(\mathbf{x}_k)(\mathbf{x}_{k+1} - \mathbf{x}_k) = \mathbf{f}'(\mathbf{y}) \cdot 0 = 0$$

Hence, $\mathbf{x}_k$ converges to a solution $\mathbf{y}$ of $\mathbf{f}(\mathbf{x}) = 0$.
Let $\mathbf{x}^*$ and $\mathbf{y}^*$ be two solutions of $\mathbf{f}(\mathbf{x}) = 0$.

$$0 = \mathbf{f}(\mathbf{x}^*) - \mathbf{f}(\mathbf{y}^*) \ge \mathbf{f}'(\mathbf{y}^*)(\mathbf{x}^* - \mathbf{y}^*) \to \mathbf{x}^* \le \mathbf{y}^*$$

and reversing the roles yields $\mathbf{x}^* \ge \mathbf{y}^*$.

## 1.8 Nonlinear Gauss-Seidel and SOR

### [Newton-SOR method]

We recall the SOR applied to the linear system $A\mathbf{x} = b$ with $A = D - L - U$.

$$\begin{aligned} x_{i,k+1} &= (1-\omega)x_{i,k} + \frac{\omega}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_{j,k+1} - \sum_{j=i+1}^{n} a_{ij}x_{j,k}\right) \\ \mathbf{x}_{k+1} &= (D-\omega L)^{-1}[(1-\omega)D + \omega U]\mathbf{x}_k + \omega(D-\omega L)^{-1}b \\ \mathbf{x}_{k+1} &= \mathbf{x}_k - \omega(D-\omega L)^{-1}(A\mathbf{x}_k - b) \end{aligned}$$

Iterate $\mathbf{x}_{k+1}$ can be expressed in terms of $\mathbf{x}_0$:
Let $B = \omega^{-1}(D-\omega L)$ , $C = \omega^{-1}[(1-\omega)D + \omega U]$ and $H = B^{-1}C$.
Note that $A = B - C$

$$\begin{aligned} \mathbf{x}_{k+1} &= H\ \mathbf{x}_k + B^{-1}b \\ &= H(H\mathbf{x}_{k-1} + B^{-1}b) + B^{-1}b \\ &= H^2\mathbf{x}_{k-1} + (H+I)B^{-1}b \\ &= \cdots \\ &= H^{k+1}\mathbf{x}_0 + (H^k + \cdots + I)B^{-1}b \\ &= (H^{k+1} - I)\mathbf{x}_0 + \mathbf{x}_0 + (H^k + \cdots + I)B^{-1}(A\mathbf{x}_0 - A\mathbf{x}_0 + b) \end{aligned}$$

Using $B^{-1}A = B^{-1}(B-C) = I - H$ and $(I + \cdots + H^k)(I-H) = I - H^{k+1}$,

$$\mathbf{x}_{k+1} = \mathbf{x}_0 - \omega(H^k + \cdots + I)(D-\omega L)^{-1}(A\mathbf{x}_0 - b)$$

One way to use SOR for nonlinear system is as a means of solving the linear system at each iterative step of Newton's method. In this case, the primary iteration is Newton's method and the secondary iteration is SOR.

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{x}_k - \mathbf{f}'(\mathbf{x}_k)^{-1}\mathbf{f}(\mathbf{x}_k) \\ \mathbf{f}'(\mathbf{x}_k)\mathbf{x}_{k+1} &= \mathbf{f}'(\mathbf{x}_k)\mathbf{x}_k - \mathbf{f}(\mathbf{x}_k) \end{aligned}$$

We decompose $\mathbf{f}'(\mathbf{x}_k) = D_k - L_k - U_k$. Assume that the diagonal elements of $D_k$ are nonzero and $0 < \omega_k < 2$. We define

$$\mathbf{h}_k = (D_k - \omega_k L_k)^{-1}[(1-\omega_k)D_k + \omega_k U_k]$$

Applying SOR, we obtain Newton-SOR method

$$\begin{aligned}\mathbf{x}_{k,m} &= (D_k - \omega_k L_k)^{-1}[(1-\omega_k)D_k + \omega_k U_k]\mathbf{x}_{k,m-1} \\ &+ \omega_k (D_k - \omega_k L_k)^{-1}(\mathbf{f}'(\mathbf{x}_k)\mathbf{x}_k - \mathbf{f}(\mathbf{x}_k))\end{aligned}$$

with $\mathbf{x}_{k,0}$ is given.

$$\begin{aligned}\mathbf{x}_{k,m} &= \mathbf{x}_{k,0} \\ &- \omega_k (H_k^{m-1} + \cdots + I)(D_k - \omega_k L_k)^{-1}(\mathbf{f}'(\mathbf{x}_k)(\mathbf{x}_{k,0} - \mathbf{x}_{(k)}) + \mathbf{f}(\mathbf{x}_k))\end{aligned}$$

Natural way of picking $\mathbf{x}_{k,0}$ is to set it equal to $\mathbf{x}_k$.

$$\mathbf{x}_{k,m} = \mathbf{x}_{k,0} - \omega_k (H_k^{m-1} + \cdots + I)(D_k - \omega_k L_k)^{-1}\mathbf{f}(\mathbf{x}_k)$$

$m = 1$ : one step Newton-SOR iteration

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \omega_k (D_k - \omega_k L_k)^{-1}\mathbf{f}(\mathbf{x}_k)$$

requires the evaluation of $n(n+1)/2$ partial derivatives and the solution of one lower triangular linear system.

There is a second way of extending an iterative method for linear system into one for nonlinear system.

**[Jacobi-Newton method]**

Jacobi method for $A\mathbf{x} = b$

$$x_{i,k+1} = \frac{1}{a_{ii}}\left(b - \sum_{j \neq i} a_{ij} x_{j,k}\right)$$

with $x_{i,0}$ given.

We can interpret this as follows: We solve the $i^{th}$ equation of $A\mathbf{x} = b$ for $x_i$ with all other elements $x_j,\ j \neq i$ held their values at $k^{th}$ level i.e. $x_{j,k}$. Consider now the map $\mathbf{f} : D \in \mathbb{R}^n \to \mathbb{R}^n$ whose root $\mathbf{x}^*$ we are seeking. Let $\mathbf{f} = (f_1 \cdots f_n)^T$ where $f_i : D \subset \mathbb{R}^n \to \mathbb{R}^1$. The analogue of the linear Jacobi method would be to keep $x_j,\ j \neq i$ at the $k^{th}$ level and solve the $i^{th}$ equation of $\mathbf{f}(\mathbf{x}) = 0$, for $x_i$. In other words, we solve the nonlinear equation

$$f_i(x_{1,k}, \cdots, x_{i-1,k}, x_{i,k+1}, x_{i+1,k}, \cdots x_{n,k}) = 0$$

Applying Newton's method to solve this equation, we obtain Jacobi-Newton method

$$x_{i,j} = x_{i,j-1} - \frac{f_i(x_{1,k}, \cdots, x_{i-1,k}, x_{i,j-1}, x_{i+1,k}, \cdots, x_{n,k})}{\frac{\partial f_i}{\partial x_i}(x_{1,k}, \cdots, x_{i-1,k}, x_{i,j-1}, x_{i+1,k}, \cdots, x_{n,k})}$$

## [Gauss-Seidel Newton method]

Gauss-Seidel method for $A\mathbf{x} = b$

$$x_{i,k+1} = \frac{1}{a_{ii}} \left( b - \sum_{j=1}^{i-1} a_{ij} x_{j,k+1} - \sum_{j=i+1}^{n} a_{ij} x_{j,k} \right)$$

with $x_{i,0}$ given.
We solve the $i^{th}$ equation of $A\mathbf{x} = b$ for $x_i$ keeping $x_j,\ j > i$ at $k^{th}$ level and using updated values of $x_j,\ j < i$ i.e. $x_{j,k+1}$. We solve the nonlinear equation

$$f_i(x_{1,k+1}, \cdots, x_{i-1,k+1}, x_{i,k+1}, x_{i+1,k}, \cdots, x_{n,k}) = 0$$

Applying the Newton's method, we obtain Gauss-Seidel Newton method

$$x_{i,j} = x_{i,j-1} - \frac{f_i(x_{1,k+1}, \cdots, x_{i-1,k+1}, x_{i,j-1}, x_{i+1,k}, \cdots, x_{n,k})}{\frac{\partial f_i}{\partial x_i}(x_{1,k+1}, \cdots, x_{i-1,k+1}, x_{i,j-1}, x_{i+1,k}, \cdots, x_{n,k})} \tag{7}$$

## [SOR-Newton method]

If we set, after finding $x_i$ from (7),

$$x_{i,k+1} = x_{i,k} + \omega_k (x_i - x_{i,k})$$

for some parameter $\omega_k$ we obtain SOR-Newton method.

Remark: It can be shown that the composite methods don't have the superlinear convergence rate of Newton's method. In fact, they all converge linearly.

## [Local Convergence of Newton-SOR]

Consider a general process of the form

$$\mathbf{x}_{k+1} = \mathbf{x}_k - (H^{m-1} + \cdots + I)(B\mathbf{x}_k)^{-1}\mathbf{f}(\mathbf{x}_k) \tag{8}$$

where

$$\mathbf{f}'(\mathbf{x}) = B(\mathbf{x}) - C(\mathbf{x}),\ H(\mathbf{x}) = B(\mathbf{x})^{-1}C(\mathbf{x})$$

**Theorem**

*Let $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^n$ be F-differentiable in an open neighborhood $S_0 \subset D$ of $\mathbf{x}^* \in D$ where $\mathbf{f}'$ is continuous and $\mathbf{f}(\mathbf{x}^*) = 0$. Suppose that $B : S_0 \to L(\mathbb{R}^n, \mathbb{R}^n)$ is continuous at $\mathbf{x}^*$ , that $B(\mathbf{x}^*)^{-1}$ exists and $\rho(H(\mathbf{x}^*)) < 1$. Then, for any $m \geq 1$, $\mathbf{x}^*$ is a point of attraction of (8).*

pf: Write (8) as $\mathbf{x}_{k+1} = \mathbf{g}(\mathbf{x}_k)$, $\mathbf{g} = \mathbf{x} - (A\mathbf{x})^{-1}\mathbf{f}(\mathbf{x})$ in Proposition. Then

$$A = B(I + \cdots + H^{m-1})^{-1}$$

Since $B(\mathbf{x}^*)^{-1}$ exists and $B$ is continuous at $\mathbf{x}^*$, $B(\mathbf{x})^{-1}$ exist in some ball $S(\mathbf{x}^*, \delta) \subset S_0$ and is continuous at $\mathbf{x}^*$ by Lemma. Then $H$ is continuous at $\mathbf{x}^*$. Now,

$$I - H^m = (I - H)(I + \cdots + H^{m-1})$$

Note that $\rho(H(\mathbf{x}^*)) < 1$ implies $I - H$ is nonsingular and $\rho(H^m(\mathbf{x}^*)) = \rho(H(\mathbf{x}^*))^m < 1$ implies $I - H^m$ is nonsingular. Therefore, $I + \cdots + H^{m-1}$ is nonsingular at $\mathbf{x}^*$. Since $H$ is continuous at $\mathbf{x}^*$, $A$ is continuous at $\mathbf{x}^*$ and $(A(\mathbf{x}))^{-1}$ exists. By Proposition, $\mathbf{g}$ is F-differentiable and

$$\mathbf{g}'(\mathbf{x}^*) = I - (I + \cdots + H(\mathbf{x}^*)^{m-1})B(\mathbf{x}^*)^{-1}B(\mathbf{x}^*)(I - H(\mathbf{x}^*)) = H(\mathbf{x}^*)^m$$

Hence, $\rho(\mathbf{g}'(\mathbf{x}^*)) = \rho(H(\mathbf{x}^*)^m) < 1$ and $\mathbf{x}^*$ is a fixed point. And it is a point of attraction by the Corollary.

**Corollary**

*Let $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^n$ be F-differentiable in an open neighborhood $S_0 \subset D$ of $\mathbf{x}^* \in D$ where $\mathbf{f}'$ is continuous and $\mathbf{f}(\mathbf{x}^*) = 0$. Let $\mathbf{f}'(\mathbf{x}) = D\mathbf{x} - L\mathbf{x} - U\mathbf{x}$ and assume that $D(\mathbf{x}^*)^{-1}$ exists. Consider Newton-SOR method*

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \omega(H(\mathbf{x}_k)^{m-1} + \cdots + I)(D\mathbf{x}_k - \omega L\mathbf{x}_k)^{-1}\mathbf{f}(\mathbf{x}_k)$$

$$H = (D - \omega L)^{-1}[(1 - \omega)D + \omega U]$$

*If $\rho(H(\mathbf{x}^*) < 1$ , then $\mathbf{x}^*$ is point of attraction.*

## [Local Convergence of SOR-Newton]

An $n \times n$ matrix is called M-matrix if $A^{-1} \geq 0$ and $a_{ij} \leq 0$ for $i \neq j$.

**Proposition**

*Let $\mathbf{f} : D \subset \mathbb{R}^n \to \mathbb{R}^n$ be continuously F-differentiable in an open neighborhood $S_0 \subset D$ of $\mathbf{x}^* \in D$ for which $\mathbf{f}(x^*) = 0$. Assume that $\mathbf{f}'(\mathbf{x}^*)$ is an M-matrix. Then $\mathbf{x}^*$ is a point of attraction of SOR-Newton for any $0 \leq \omega \leq 1$.*

## 1.9 Variations of Newton's method

We consider methods for which the derivatives making up the Jacobian are not evaluated. An obvious alternative is to replace the partial derivatives $\frac{\partial f_i}{\partial x_j}$ by difference quotients. e.g.

$$\frac{\partial f_i}{\partial x_j}(\mathbf{x}) \approx \frac{1}{h_{ij}}(f_i(\mathbf{x} + h_{ij}e^j) - f_i(\mathbf{x}))$$

Let $(\mathbf{h})_{ij} = h_{ij}$ and $J(\mathbf{x}, \mathbf{h})_{ij}$ denote a difference approximation to $\frac{\partial f_i}{\partial x_j}$ i.e.

$$\lim_{\mathbf{h}\to 0} J(\mathbf{x}, \mathbf{h})_{ij} = \frac{\partial f_i}{\partial x_j}$$

Then the discretized Newton's method is obtained.

$$\mathbf{x}_{k+1} = \mathbf{x}_k - J(\mathbf{x}_k, \mathbf{h}_k)^{-1}\mathbf{f}(\mathbf{x}_k)$$

Remark: (choice of $\mathbf{h}_n$)

1. $\mathbf{h}_k = \mathbf{h}$ (independent of $k$) : linear convergence
2. method become more accurate if $\lim_{k\to\infty} \mathbf{h}_k = 0$
   e.g. let $\mathbf{h}_k = \gamma_k \mathbf{h}$ where $\mathbf{h}$ fixed and $\lim_{k\to\infty} \gamma_k = 0$ in $\mathbb{R}$.
   methods of great interest choose $\mathbf{h}_k$ to be functions of the iterates $\mathbf{x}_k$.

Illustration in one dimension:

$$x_{k+1} = x_k - \left(\frac{f(x_k + \mathbf{h}_k) - f(x_k)}{h_k}\right)^{-1} f(x_k) \quad k = 0, 1, \cdots \tag{9}$$

- $\mathbf{h}_k = \overline{x} - x_k$ where $\overline{x}$ is a fixed number leads to the regular false method

$$x_{k+1} = x_k - \left(\frac{f(\overline{x}) - f(x_k)}{\overline{x} - x_k}\right)^{-1} f(x_k) \quad k = 0, 1, \cdots$$

- $\mathbf{h}_k = x_{k-1} - x_k$ gives the secant method

$$x_{k+1} = x_k - \left(\frac{f(x_{k-1}) - f(x_k)}{x_{k-1} - x_k}\right)^{-1} f(x_k) \quad k = 0, 1, \cdots$$

  which requires two starting values $x_0$ and $x_1$.

[Broyden's Method]

- Secant updating methods reduce cost by
  Using function values at successive iterates to build approximate Jacobian to avoid explicit evaluation of derivatives and to reduce the function evaluations
  Updating factorization of approximate Jacobian rather than refactoring it each iteration or updating approximate inverse of Jacobian
- Most secant updating methods have superlinear but not quadratic convergence rate
- Secant updating methods often cost less overall than Newton's method because of lower cost per iteration
- Broyden's method is typical secant updating method

If $\bar{B}$ is to denote our approximation to $\mathbf{f}(\bar{\mathbf{x}})$, then

$$\bar{B}(\bar{\mathbf{x}} - \mathbf{x}) = \mathbf{f}(\bar{\mathbf{x}}) - \mathbf{f}(\mathbf{x}) \quad \text{quasi-Newton equation}$$

In 1-dimension, $\bar{B}$ is completely determined and the approximation reduces to secant method. For $n > 1$, $\bar{B}$ is not uniquely determined since the only information is in the direction of $\mathbf{s} = \bar{\mathbf{x}} - \mathbf{x}$.
Suppose we have an approximation $B$ to $\mathbf{f}(\bar{\mathbf{x}})$ in hand and we require that

$$\bar{B}\mathbf{z} = B\mathbf{z} \text{ if } \mathbf{z}^T\mathbf{s} = 0$$

Then

$$\bar{B} = B + \frac{(\mathbf{y} - B\mathbf{s})\mathbf{s}^T}{\mathbf{s}^T\mathbf{s}} \quad \text{Broyden update} \tag{10}$$

where, $\mathbf{y} = \mathbf{f}(\bar{\mathbf{x}}) - \mathbf{f}(\mathbf{x})$, satisfy both relations.

Basic Broyden's method:

$$\begin{aligned} B_k\mathbf{s}_k &= -\mathbf{f}(\mathbf{x}_k) \\ \mathbf{x}_{k+1} &= \mathbf{x}_k + \mathbf{s}_k \\ \mathbf{y}_k &= \mathbf{f}(\mathbf{x}_{k+1}) - \mathbf{f}(\mathbf{x}_k) \\ B_{k+1} &= B_k + \frac{(\mathbf{y}_k - B_k\mathbf{s}_k){\mathbf{s}_k}^T}{{\mathbf{s}_k}^T\mathbf{s}_k} \end{aligned}$$

Broyden's method can be carried out with $n$ scalar function evaluations per iteration, but we still need to solve the linear system $B_k\mathbf{s}_k = -\mathbf{f}(\mathbf{x}_k)$.

Lemma 1: Let $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$ be given. Then

$$det(I + \mathbf{v}\mathbf{w}^T) = 1 + \mathbf{v}^T\mathbf{w}$$

Lemma 2: Let $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ and assume that $A \in \mathcal{L}(\mathbb{R}^n)$ is nonsingular. Then $A + \mathbf{u}\mathbf{v}^T$ is nonsingular iff $\sigma = 1 + \mathbf{v}^T A^{-1}\mathbf{u} \neq 0$. Moreover, if $\sigma \neq 0$ then

$$(A + \mathbf{u}\mathbf{v}^T)^{-1} = A^{-1} - \frac{1}{\sigma}A^{-1}\mathbf{u}\mathbf{v}^T A^{-1}$$

From lemma 2, if $H_k = (B_k)^{-1}$ and $H_{k+1} = (B_{k+1})^{-1}$ then

$$H_{k+1} = H_k - \frac{1}{\sigma_k}H_k\frac{(\mathbf{y}_k - B_k\mathbf{s}_k){\mathbf{s}_k}^T}{{\mathbf{s}_k}^T\mathbf{s}_k}H_k$$

with $\sigma_n = 1 + \mathbf{s}_k^T H_k(\mathbf{y}_k - B_k\mathbf{s}_k)/({\mathbf{s}_k}^T\mathbf{s}_k)$, which yields

$$H_{k+1} = H_k - \frac{(H_k\mathbf{y}_k - \mathbf{s}_k){\mathbf{s}_k}^T H_k}{{\mathbf{s}_k}^T H_k\mathbf{y}_k}$$

provided ${\mathbf{s}_k}^T H_k\mathbf{y}_k \neq 0$.

Broyden's method:

Given $\mathbf{x}_0$ and $H_0 \approx F(\mathbf{x}_0)^{-1}$

For $n = 0, 1, \cdots$

$$\begin{aligned}
\mathbf{x}_{k+1} &= \mathbf{x}_k - H_k\mathbf{f}(\mathbf{x}_k) \\
\mathbf{s}_k &= \mathbf{x}_{k+1} - \mathbf{x}_k \\
\mathbf{y}_k &= \mathbf{f}(\mathbf{x}_{k+1}) - \mathbf{f}(\mathbf{x}_k) \\
H_{k+1} &= H_k - \frac{(H_k\mathbf{y}_k - \mathbf{s}_k){\mathbf{s}_k}^T H_k}{{\mathbf{s}_k}^T H_k\mathbf{y}_k}
\end{aligned}$$

**Theorem**

*Let $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ be continuously differentiable in an open convex set $D$. Let $\mathbf{x}^*$ be such that $\mathbf{f}(\mathbf{x}^*) = 0$ and $\mathbf{f}(\mathbf{x}^*)$ is nonsingular. Suppose that there exists a constant $\hat{C}$ such that $\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}^*)\| \leq \hat{C}\|\mathbf{x} - \mathbf{x}^*\|$ for $\mathbf{x} \in D$. Then Broyden's method is locally and superlinearly convergent at $\mathbf{x}^*$.*