ERG 2012B
Advanced Engineering Mathematics II

Part III
Introduction to Numerical Methods

Lecture #17
Numerical Method Basics

Numerical Methods

Fixed Point System: all numbers are given with *a fixed number of decimal places* e.g. 62.358, 0.013, 1.000

Floating Point System: all numbers are given with a *fixed number of significant digits*. e.g.

0.6238×10^{3} 0.1714×10^{-13} -0.2000×10^{1}

0.6238E03 0.1714E-13 -0.2000E01

Significant digit of a number *c* is any given digit of *c,* except possibly for zeros to the left of the first nonzero digit that serve only to fix the position of the decimal point:

1360

1.360 4 significant digits (4 S)

0.001360

Numerical Methods

Chopping: discarding all decimals from some decimal place on

1.618 → 1.61 or 1.6 or 1

Rounding: to keep the number of digits of a number to k decimals or k significant digits according to the **round-off rule**

Round-off rule: discard the $(k+1)^{th}$ and all subsequent decimals

(a) If the number thus discarded is less than half a unit in the k^{th} place leave the k^{th} decimal unchanged (*rounding down*)

1.648 → 1.6

(b) If it is greater than half a unit in the k^{th} place, add one to the k^{th} decimal (*rounding up*)   1.648 → 1.65

(c) If it is exactly half a unit, round off to the nearest **even** decimal (average the chance) (e.g. 3.45 → 3.4 and 3.55→3.6)

In practice, most computers that use rounding-off *always* round up in case (c) of the rule, this is easier technically.

Errors of Numerical Results

Round-off errors: results from rounding

Experimental errors: are errors of given data (probably arising from the measurements)

Truncating errors: results from truncating

- If $\tilde{a}$ is an approximate value of a quantity whose exact value is a the difference $\varepsilon = a - \tilde{a}$ is called **the error of $\tilde{a}$**
- Hence $a = \tilde{a} + \varepsilon$ (True value = Approximation + Error)
- The **relative error** ε_r of $\tilde{a}$ is defined by

$$\varepsilon_r = \frac{\varepsilon}{a} = \frac{a - \tilde{a}}{a} = \frac{\text{Error}}{\text{True value}} \qquad (a \neq 0)$$

- if $|\varepsilon|$ is much less than $|\tilde{a}|$ then $\varepsilon_r \approx \varepsilon / \tilde{a}$
- **Error bound** for $\tilde{a}$ is a number β such that $|\varepsilon| \leq \beta$
- **Error bound for the relative error:** a number β_r such that $|\varepsilon_r| \leq \beta_r$

Error Propagation

Theorem 1: (a) In addition and subtraction, an error bound for the results is given by the sum of the error bounds of the terms

(b) In multiplication and division, an error bound for the **relative** error of the results is given (approximately) by the sum of error bound for the **relative** errors of the given numbers.

Proof: (a) if $x = \tilde{x} + \varepsilon_1, \quad y = \tilde{y} + \varepsilon_2, \; |\varepsilon_1| \le \beta_1, \; |\varepsilon_2| \le \beta_2$

Then for the error ε of the *difference* we get

$$\begin{aligned} |\varepsilon| &= |x - y - (\tilde{x} - \tilde{y})| \\ &= |x - \tilde{x} - (y - \tilde{y})| \\ &= |\varepsilon_1 - \varepsilon_2| \le |\varepsilon_1| + |\varepsilon_2| \le \beta_1 + \beta_2 \end{aligned}$$

The proof for the *sum* is similar.

Error Propagation

Theorem 1: (a) In addition and subtraction, an error bound for the results is given by the sum of the error bounds of the terms

(b) In multiplication and division, an error bound for the **relative** error of the results is given (approximately) by the sum of error bound for the **relative** errors of the given numbers.

Proof: (b) for the relative error ε_r of $\tilde{x}\tilde{y}$ from the relative errors ε_{r1} and ε_{r2} of $\tilde{x}, \tilde{y}$ and the bounds β_{r1}, β_{r2}

$$|\varepsilon_r| = \left|\frac{xy - \tilde{x}\tilde{y}}{xy}\right| = \left|\frac{xy - (x-\varepsilon_1)(y-\varepsilon_2)}{xy}\right| = \left|\frac{\varepsilon_1 y + \varepsilon_2 x - \varepsilon_1\varepsilon_2}{xy}\right|$$

$$\approx \left|\frac{\varepsilon_1 y + \varepsilon_2 x}{xy}\right| \le |\varepsilon_{r1}| + |\varepsilon_{r2}| \le \beta_{r1} + \beta_{r2}$$

Approximately means we ignore $\varepsilon_1\varepsilon_2$ as small. The quotient is similar

Iteration

Solution of Equations by Iteration

Given the equation $f(x)=0$ ....(*)

Fixed-point iteration method

Transform (*) *algebraically* into the form $x=g(x)$

Then choose an x_0 and compute $x_1=g(x_0)$, $x_2=g(x_1)$.....

and in general $x_{n+1} = g(x_n)$ $(n = 0,1,.....)$

A solution of $x=g(x)$ is called a **fixed point** of g - hence the name - and is a solution of (*)

From (*) we can get several different forms for $x=g(x)$ the behaviour of the corresponding iterative sequences, x_0, x_1,.. may differ in their speed of convergence

An iteration process is called **convergent** for x_0 if the corresponding sequence x_0, x_1, ... is convergent

Example 1

An iteration process

Set up an iteration process for the equation $f(x)=x^2-3x+1=0$

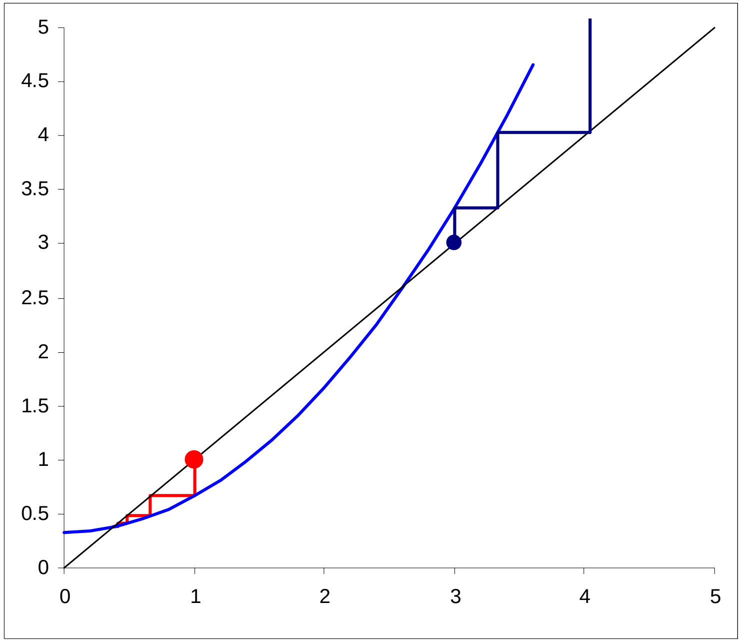
We know the solutions $x=1.5\pm\sqrt{1.25}$ or 2.618034 and 0.381966 so can watch the behaviour of the iteration process

Solution: The equation can be written:

$$x = g_1(x) = 1/3(x^2+1) \quad \text{thus } x_{n+1} = 1/3(x_n{}^2+1)$$

If we choose $x_0=1$ we get the sequence:
$x_0=1.000$, $x_1=0.667$, $x_2=0.481$, $x_3=0.411$
$x_4=0.390$,..... getting closer to the lower solution

If we choose $x_0=3$ we get the sequence:
$x_0=3.000$, $x_1=3.333$, $x_2=4.037$, $x_3=5.767$
$x_4=11.415$,..... diverging



Example 1

An iteration process

Set up an iteration process for the equation $f(x)=x^2-3x+1=0$

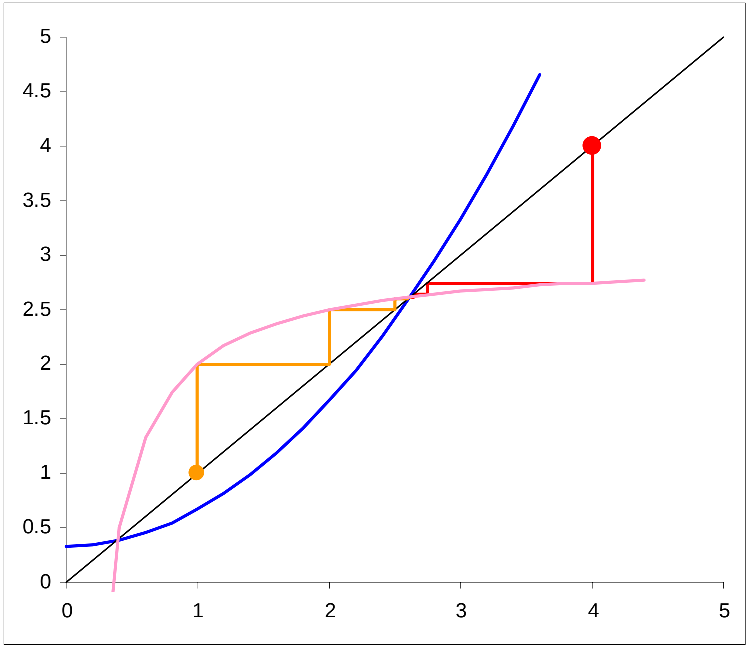
We know the solutions $x=1.5\pm\sqrt{1.25}$ or 2.618034 and 0.381966 so can watch the behaviour of the iteration process

Solution: The equation can also be written:

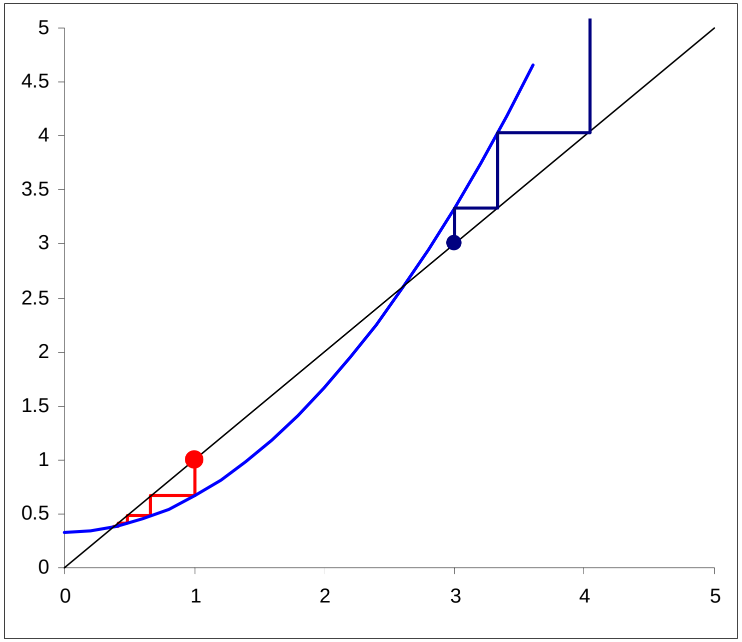
$$x = g_2(x) = 3 - 1/x \quad \text{thus } x_{n+1} = 3-1/x_n$$

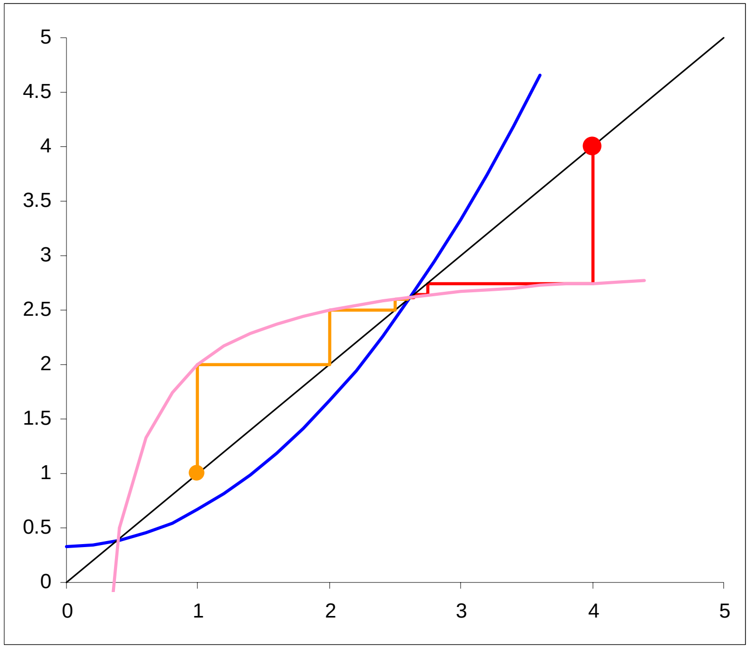
If we choose $x_0=1$ we get the sequence $x_0=1.000$, $x_1=2.000$, $x_2=2.500$, $x_3=2.600$ $x_4=2.615$, approaching the larger solution

If we choose $x_0=3$ we get the sequence: $x_0=3.000$, $x_1=2.667$, $x_2=2.626$, $x_3=2.619$ $x_4=2.618$, approaching the same solution.



Convergence





Notice that in the first figure the slope of the $g_1(x)$ is less than that of $y=x$ around the lower root and greater around the upper root. In the second figure this is the other way around.

It appears that convergence to a root is dependent upon the slope of the curve at that point compared with $y=x$

Convergence

Theorem 1: Convergence of fixed-point iteration.

Let $x = s$ be a solution of $x=g(x)$ and suppose that g has continuous derivative in some interval J containing s. Then if $|g'(x)| \le K < 1$ in J, the iteration process outlined above converges for any x_0 in J.

Proof: since $g(s) = s$ and $x_1 = g(x_0), x_2 = g(x_1), \ldots\ldots$ we can write

$$|x_n - s| = |g(x_{n-1}) - g(s)|$$

from the mean value theorem of calculus there is a t between x and s such that

$$g(x) - g(s) = g'(t)(x - s) \text{ and so}$$

$$\begin{aligned} |x_n - s| &= |g'(t)||x_{n-1} - s| \le K|x_{n-1} - s| \\ &= K|g(x_{n-2}) - g(s)| \\ &= K|g'(t)||x_{n-2} - s| \le K^2|x_{n-2} - s| \\ &\ldots\ldots. \\ &\le K^n|x_0 - s| \end{aligned}$$

if $K < 1$ $K^n \to 0$ as $n \to \infty$

Example 2

An Iteration process.

Find a solution of $f(x) = x^3 + x - 1 = 0$ by iteration

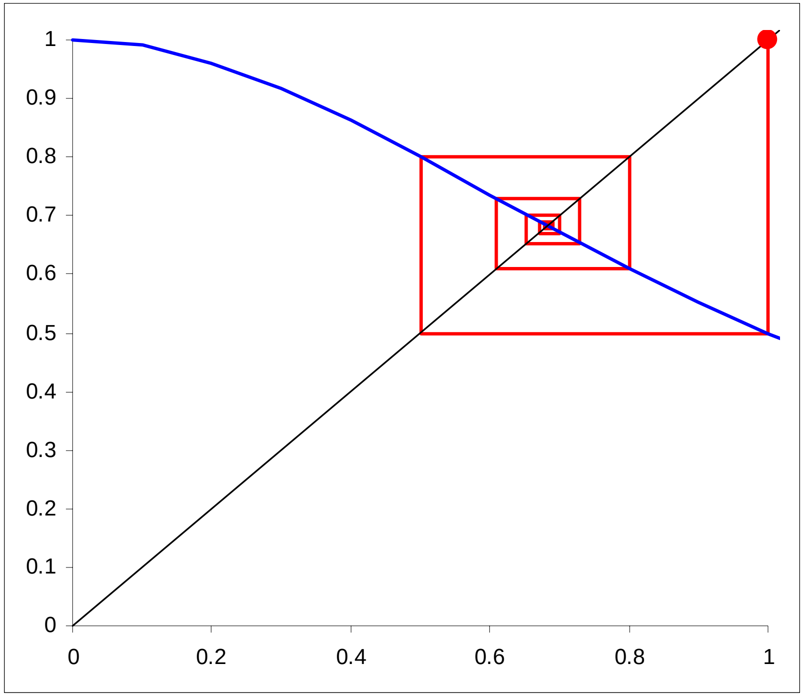
Solution: A rough sketch shows that a real solution lies between x=0 and 1 (f(1) = 1; f(0) = -1).

We can write the equation in the form $x=g_1(x)=1/(1+x^2)$ so that $x_{n+1}=1/(1+x_n{}^2)$

Then $|g'_1(x)| = 2|x|/(1+x^2)^2 < 1$ for any x as $4x^2/(1+x^2)^4 =4x^2/(1+4x^2+...) < 1$ for all x.

Choosing $x_0=1$ we get:

$x_1=0.500$, $x_2=0.800$, $x_3=0.610$, $x_4=0.729$, $x_5=0.653$, $x_6=0.701$,....



The solution to 6 decimal places is 0.682328

Example 2

An Iteration process.

Find a solution of $f(x) = x^3 + x - 1 = 0$ by iteration

Solution: A rough sketch shows that a real solution lies between x=0 and 1 (f(1) = 1; f(0) = -1).

We can also write the equation in the form $x=g_2(x)=1-x^3$ so that $x_{n+1}=1-x_n^{\ 3}$

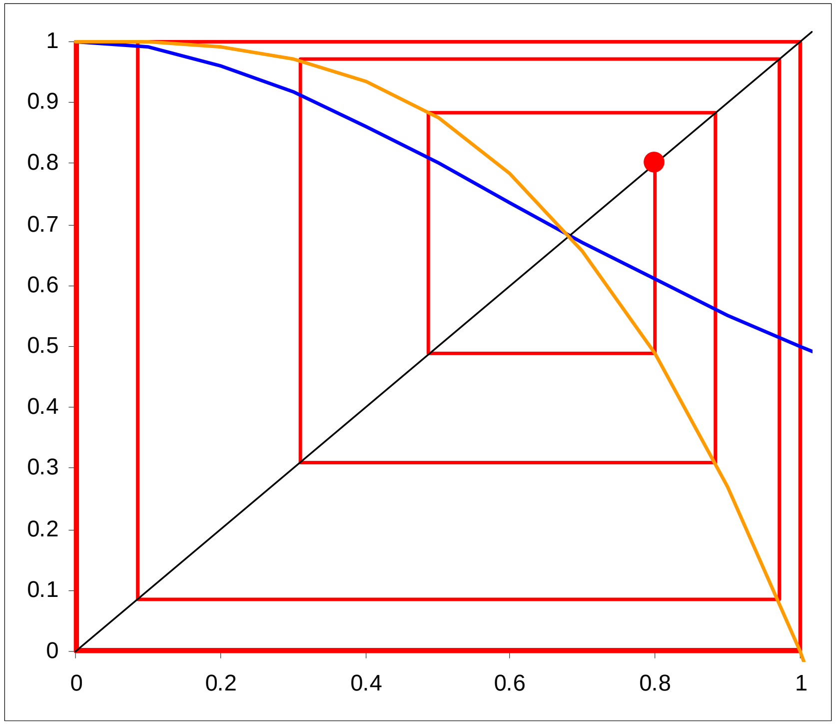
Then $|g'_2(x)| = 3x^2 > 1$ near the solution - can't expect convergence

Choosing $x_0=1$ we get:

$x_1=0$, $x_2=1$, *etc.*

Choosing $x_0=0.8$ we get

$x_1=0.488$, $x_2=0884$, $x_3=0.310$, $x_4=0.970$, $x_5=0.0864$,....



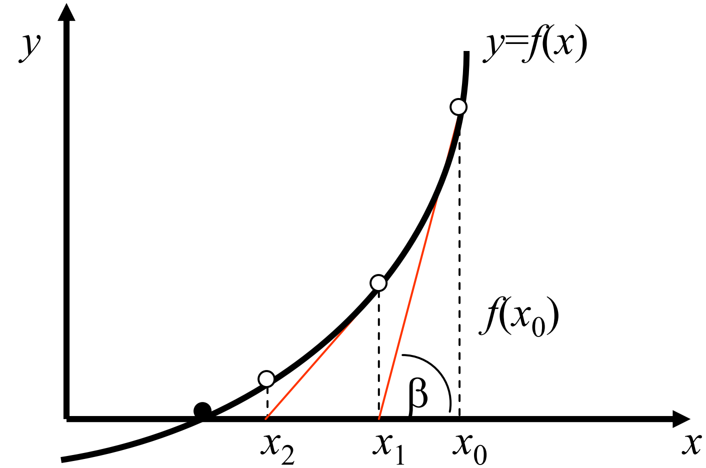
Newton's Method

Newton's Method for Solving Equations *f*(*x*)=0

The **Newton** or **Newton-Raphson method** is another iteration method for solving equations $f(x)=0$, where f is assumed to have a continuous derivative f'. The method is commonly used because of its simplicity and great speed.

$$\tan\beta = f'(x) = \frac{f(x_0)}{x_0 - x_1}$$

hence $$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$



In the second step we compute $x_2=x_1-f(x_1)/f'(x_1)$ in the third step x_3 from x_2 with the same formula and so on.

Leads to a simple computational algorithm.

Example 3

Square Root

Setup a Newton iteration for computing the square root of x of a given positive number c and apply it to c=2

Solution: We have $x=\sqrt{c}$ hence $f(x) = x^2 - c = 0$ and $f^{/}(x) = 2x$ and the iteration formula is:

$$x_{n+1} = x_n - \frac{x_n{}^2 - c}{2x_n} = \frac{1}{2}\left(x_n + \frac{c}{x_n}\right)$$

For $c = 2$, choosing x_0=1, we get:

x_1=1.500000, x_2=1.416667, x_3=1.414216, x_4=1.414214.......

and x_4 is already exact to 6 decimal places.

Example 4

Iteration of a transcendental equation

Find the positive solution of $2\sin x = x$

Solution: Setting $f(x) = x - 2\sin x$ we have $f'(x) = 1 - 2\cos x$ and the iteration formula is:

$$x_{n+1} = x_n - \frac{x_n - 2\sin x_n}{1 - 2\cos x_n} = \frac{2(\sin x_n - x_n \cos x_n)}{1 - 2\cos x_n} = \frac{N_n}{D_n}$$

Choosing $x_0=2$, we get:

n	x_n	N_n	D_n	x_{n+1}
0	2.00000	3.48318	1.83229	1.90100
1	1.90100	3.12470	1.64847	1.89552
2	1.89552	3.10500	1.63809	1.89550
3	1.89550	3.10493	1.63806	**1.89549**

Example 5

Newton's method applied to an algebraic equation

Apply Newton's method to the equation $f(x) = x^3 + x - 1 = 0$

Solution: the iteration formula is:

$$x_{n+1} = x_n - \frac{x_n^{\ 3} + x_n - 1}{3x_n^{\ 2} + 1} = \frac{2x_n^{\ 3} + 1}{3x_n^{\ 2} + 1}$$

Choosing $x_0=1$, we get:

$x_1 = 0.750000$, $x_2 = 0.686047$, $x_3 = 0.682340$, $x_4 = 0.682328$.....

and x_4 is exact to 6 decimal places.

Speed of Convergence

Let $x_{n+1} = g(x_n)$ define an iteration method and let x_n approximate a solution s of $x = g(x)$. Then $x_n = s - \varepsilon_n$; where ε_n is the error of x_n. Suppose that g is differentiable a number of times, so that the Taylor formula gives:

$$\begin{aligned} x_{n+1} = g(x_n) &= g(s) + g'(s)(x_n - s) + \tfrac{1}{2}g''(s)(x_n - s)^2 + \ldots \\ &= g(s) - g'(s)\varepsilon_n + \tfrac{1}{2}g''(s)\varepsilon_n^2 + \ldots \end{aligned}$$

The exponent of ε_n in the first non-vanishing term after $g(s)$ is called the **order** of the iteration process defined by g

The order measures the speed of convergence

subtract $g(s)=s$ on both sides then

on the left $x_{n+1} - s = -\varepsilon_{n+1}$ – the error in x_{n+1}

the expression on the right is ≈ its first nonzero term as $|\varepsilon_n|$ is small in convergence.

Speed of Convergence

Thus:

a) $\varepsilon_{n+1} \approx +g'(s)\,\varepsilon_n$ in the case of 1st order

b) $\varepsilon_{n+1} \approx -\tfrac{1}{2}g''(s)\,\varepsilon_n^2$ in the case of 2nd order

So that if $\varepsilon_n = 10^{-k}$ in some step, then for 2nd order, $\varepsilon_{n+1} = \text{cnst.}\,10^{-2k}$ and number of significant digits ~doubles in each step

For Newton's method, $g(x) = x - f(x)/f'(x)$ and by differentiation

$$g'(x) = 1 - \frac{f'(x)^2 - f(x)f''(x)}{f'(x)^2} = \frac{f(x)f''(x)}{f'(x)^2}$$

Since $f(s) = 0$ this shows that also $g'(s) = 0$. Hence Newton's Method is at least 2nd order. Differentiate again and

$$g''(s) = \frac{f''(s)}{f'(s)}$$

which in general will not be zero

Convergence of Newton's Method

Theorem 2: If $f(x)$ is three times differentiable and f' and f'' are not zero at a solution s of $f(x) = 0$ then for x_0 sufficiently close to s Newton's method is of second order.

Notice for Newton's method

$$\varepsilon_{n+1} \approx \frac{f''(s)}{2f'(s)} \varepsilon_n$$

Difficulties can arise if $|f'(x)|$ is very small near a solution s. So that values of $x = \hat{s}$ far away from the solution s can still have small values $R(\hat{s}) = f(\hat{s})$

In this case we call the equation $f(x) = 0$ **ill-conditioned**. $R(\hat{s})$ is called the **residual** of $f(x) = 0$ at s.

Thus a small residual only guarantees a small error of $\hat{s}$ if the equation is ***not*** ill-conditioned.