NEW YORK CQF WORKSHOP

- Mathematical Methods: The Heat Equation; Laplace Transforms and Applications.

- Introduction to Numerical Methods: Numerical Integration; Root Finding; Numerical Linear Algebra; Finite Difference Method

Mathematical Methods

The Heat Equation

We have seen on numerous occasions the one dimensional heat/diffusion equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$$

for the unknown function of the form $u(x,t) = t^{\alpha}\phi\left(\frac{x}{t^{\beta}}\right)$. The corresponding solution derived using the similarity reduction technique is the *fundamental solution*

$$u(x,t) = \frac{1}{2\sqrt{\pi t}}\exp\left(-\frac{x^2}{4t}\right).$$

Some books refer to this as a *source solution*.

Let's consider the following integral

$$\lim_{t \longrightarrow 0} \int_{-\infty}^{\infty} u(y,t) f(y) \, dy$$

which can be simplified by the substitution

$$s = \frac{y}{2\sqrt{t}} \Longrightarrow 2\sqrt{t} ds = dy$$

to give

$$\lim_{t \longrightarrow 0} \frac{1}{2\sqrt{\pi t}} \int_{-\infty}^{\infty} \exp\left(-s^2\right) f\left(2\sqrt{t}s\right) 2\sqrt{t} ds.$$

In the limiting process we get

$$\begin{aligned} f(0) \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \exp\left(-s^2\right) ds &= f(0) \frac{1}{\sqrt{\pi}} \sqrt{\pi} \\ &= f(0). \end{aligned}$$

Hence

$$\lim_{t \longrightarrow 0} \int_{-\infty}^{\infty} u(y,t) f(y) \, dy = f(0).$$

A slight extension of the above shows that

$$\lim_{t \longrightarrow 0} \int_{-\infty}^{\infty} u\left(x-y,t\right) f\left(y\right) dy = f\left(x\right),$$

where

$$u\left(x-y,t\right) = \frac{1}{2\sqrt{\pi t}} \exp\left(-\frac{\left(x-y\right)^2}{4t}\right).$$

Let's derive the result above. As earlier we begin by writing $s = \frac{x-y}{2\sqrt{t}} \Longrightarrow y = x - 2\sqrt{t}s$ and hence $dy = -2\sqrt{t}ds$. Under this transformation the limits are

$$\begin{aligned} y &= \infty \longrightarrow s = -\infty \\ y &= -\infty \longrightarrow s = \infty \end{aligned}$$

$$\begin{aligned} & \frac{1}{2\sqrt{\pi t}} \int_{\infty}^{-\infty} \exp\left(-s^2\right) f\left(x - 2\sqrt{t}s\right) \left(-2\sqrt{t}ds\right) ds \\ = \; & \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \exp\left(-s^2\right) f\left(x - 2\sqrt{t}s\right) ds \end{aligned}$$

and $\lim_{t \longrightarrow 0} f\left(x - 2\sqrt{t}s\right) = f(x)$, so

$$\begin{aligned} &= f(x) \frac{1}{\sqrt{\pi}} \sqrt{\pi} \\ &= f(x) \end{aligned}$$

$$\lim_{t \longrightarrow 0} \int_{-\infty}^{\infty} u(x - y, t) f(y) \, dy = f(x).$$

Since the heat equation is a constant coefficient PDE, if $u(x, t)$ satisfies it, then $u(x - y, t)$ is also a solution for any y.

Recall what it means for an equation to be linear:

Since the heat equation is linear,

1. if $u(x - y, t)$ is a solution, so is a multiple $f(y) u(x - y, t)$

2. we can add up solutions. Since $f(y)\,u(x-y,t)$ is a solution for any y, so too is the integral

$$\int_{-\infty}^{\infty} u(x-y,t)\,f(y)\,dy.$$

Recall, adding can be done in terms of an integral. So we we can summarize by specifying the following initial value problem

$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x^2} \\ u(x,0) &= f(x) \end{aligned}$$

which has a solution

$$u(x,t) = \frac{1}{2\sqrt{\pi t}} \int_{\infty}^{-\infty} \exp\left(-\frac{(x-y)^2}{4t}\right) f(y)\,dy.$$

This satisfies the initial condition at $t = 0$ because we have shown that at that point the value of this integral is $f(x)$. Putting $t < 0$ gives a non-existent solution, i.e. the integrand will blow up.

Example Consider the IVP

$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x^2} \\ u(x,0) &= \begin{cases} 0 & \text{if } x > 0 \\ 1 & \text{if } x < 0 \end{cases} \end{aligned}$$

We can write down the solution as

$$\begin{aligned} u(x,t) &= \frac{1}{2\sqrt{\pi t}} \int_{\infty}^{-\infty} \exp\left(-\frac{(x-y)^2}{4t}\right) u(y,0)\, dy \\ &= \frac{1}{2\sqrt{\pi t}} \int_{-\infty}^{0} \exp\left(-\frac{1}{2}\frac{(x-y)^2}{2t}\right) .1 dy \end{aligned}$$

put

$$s = \frac{y-x}{\sqrt{2t}}$$

$\int_{-\infty}^{0}$ becomes $\int_{-\infty}^{\frac{-x}{2\sqrt{t}}}$

$$\begin{aligned} & \frac{1}{2\sqrt{\pi t}}\int_{-\infty}^{\frac{-x}{2\sqrt{t}}} \exp\left(-s^2/2\right)\sqrt{2t}ds \\ = \quad & \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\frac{-x}{2\sqrt{t}}} \exp\left(-s^2/2\right)ds \\ = \quad & N\left(\frac{-x}{2\sqrt{t}}\right) \end{aligned}$$

So we have expressed the solution in terms of the CDF.

This can also be solved by using the substitution

$$\widehat{s} = \frac{-(y-x)}{2\sqrt{t}} \longrightarrow -dy = 2\sqrt{t}d\widehat{s}$$

$\int_{-\infty}^{0}$ becomes $\int_{\infty}^{\frac{x}{2\sqrt{t}}}$

$$
\begin{aligned}
& -\frac{1}{2\sqrt{\pi t}}\int_{\infty}^{\frac{x}{2\sqrt{t}}} \exp\left(-\widehat{s}^2\right) 2\sqrt{t}d\widehat{s} \\
= \quad & \frac{1}{2}\frac{2}{\sqrt{\pi}}\int_{\frac{x}{2\sqrt{t}}}^{\infty} \exp\left(-\widehat{s}^2\right) d\widehat{s} = \\
= \quad & \frac{1}{2}\,\mathrm{erf\,c}\left(\frac{x}{2\sqrt{t}}\right)
\end{aligned}
$$

so now we have a solution in terms of the *complimentary error function*.

The Laplace Transform

Given a function defined $\forall\ t > 0$ then the *Laplace Transform* (LT) of $f(t)$, written $\mathcal{L}\{F(t)\}$ is defined by

$$\mathcal{L}\{f(t)\} = \int_0^\infty e^{-st} f(t)\, dt = F(s)$$

provided this integral exists.

$\mathcal{L}\{F(t)\}$ is a function of s which we denote by $F(s)$.

Lower case letters	Upper case letters
$f(t)$	$F(s)$
$g(t)$	$G(s)$
$h(t)$	$H(s)$

Example: If $f(t) = 1$ then $\mathcal{L}\{1\} = \int_0^\infty 1.e^{-st}dt.$ Look at rhs

$$\begin{aligned}
\lim_{T\to\infty}\int_0^T e^{-st}dt &= \lim_{T\to\infty}\left[-\frac{1}{s}e^{-st}\right]_0^T \quad (s=0)\\
&= \lim_{T\to\infty}\left(\frac{1}{s}\left[1-e^{-sT}\right]\right)\\
&\to \frac{1}{s} \text{ if } s>0
\end{aligned}$$

The limit does not exist for $s < 0$.

What about $s = 0$?

$$\int_0^\infty e^{0t}dt = \int_0^\infty 1dt$$

which does not exist. So for $f(t) = 1$ $F(s) = \frac{1}{s}$ $(s > 0)$ and L.T. does not exist for $s \leq 0$.

Laplace Transforms of Standard Functions

1. $\mathcal{L}\{1\} = \dfrac{1}{s} \quad (s > 0)$

2. $\mathcal{L}\left\{e^{at}\right\} = \displaystyle\int_0^\infty e^{at} e^{-st} dt =$

$$\lim_{T\to\infty} \int_0^T e^{-(s-a)t} dt = \lim_{T\to\infty} \left[\frac{1}{s-a}\left(1 - e^{-(s-a)T}\right)\right]$$
$$= \frac{1}{s-a} \quad (s > a)$$

3. $\mathcal{L}\{t\} = \displaystyle\int_0^\infty te^{-st} dt$ integration by parts

$$= -\frac{t}{s} e^{-st}\Big|_0^\infty + \frac{1}{s}\int_0^\infty e^{-st} dt$$

Now $te^{-st} \to 0$ as $t \to \infty \because s > 0$ therefore

$$\begin{aligned} \mathcal{L}\{t\} &= \frac{1}{s}\int_0^\infty e^{-st}dt = \frac{1}{s}\left[-\frac{e^{-st}}{s}\right]_0^\infty \\ &= \frac{1}{s^2} \end{aligned}$$

4. $\mathcal{L}\{t^n\} = \int_0^\infty t^n e^{-st}dt \qquad (s > 0)$, and put $u = st$ to give $dt = \frac{1}{s}du$, hence

$$\int_0^\infty t^n e^{-st}dt = \int_0^\infty \left(\frac{u}{s}\right)^n e^{-u}\frac{du}{s} = \frac{1}{s^{n+1}}\int_0^\infty u^n e^{-u}du$$

where

$$\int_0^\infty u^n e^{-u}du = \Gamma(n+1)$$

so we get

$$\frac{\Gamma(n+1)}{s^{n+1}} = \frac{n!}{s^{n+1}}$$

hence

$$\mathcal{L}\{t^n\} = \frac{n!}{s^{n+1}}.$$

5. Similarly $\mathcal{L}\{t^\alpha\} = \dfrac{\Gamma(\alpha+1)}{s^{\alpha+1}} \qquad (s>0,\ \alpha>-1)$

$$\begin{aligned}\mathcal{L}\left\{t^{-1/2}\right\} &= \frac{\Gamma(-1/2+1)}{s^{-1/2+1}} = \frac{\Gamma(1/2)}{s^{1/2}} \\ &= \frac{\sqrt{\pi}}{s^{1/2}} = \sqrt{\frac{\pi}{s}}\end{aligned}$$

$$\mathcal{L}\left\{t^{1/2}\right\} = \frac{\Gamma(1/2+1)}{s^{1/2+1}} = \frac{\Gamma(1/2+1)}{s^{3/2}}$$

recall $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$, so we can write $\Gamma(3/2) = (1/2)\Gamma(1/2) = (1/2)\sqrt{\pi}$

$$= \frac{\sqrt{\pi}}{2s^{3/2}} = \frac{1}{2s}\sqrt{\frac{\pi}{s}}.$$

6. $\mathcal{L}\{\sin t\} = \int_0^\infty e^{-st} \sin kt.dt$

Using Eulers Identity is fairly quick.

$$\begin{aligned}
\int_0^\infty e^{-st} \sin kt.dt &= \int_0^\infty e^{-st} \operatorname{Im} e^{ikt} dt \\
&= \operatorname{Im} \int_0^\infty e^{-(s-ik)t} dt \\
&= \operatorname{Im} \frac{1}{-(s-ik)} e^{-(s-ik)t} \Big|_0^\infty = \operatorname{Im} \frac{1}{(s-ik)} \\
&= \operatorname{Im} \frac{1}{(s-ik)} \frac{(s+ik)}{(s+ik)} = \operatorname{Im} \frac{(s+ik)}{s^2+k^2} \\
&= \frac{k}{k^2+s^2}
\end{aligned}$$

As an exercise repeat this for the Laplace Transform of $\cos kt$.

Simple Properties

1. $\mathcal{L}\{f(t)+g(t)\} = \mathcal{L}\{f(t)\} + \mathcal{L}\{g(t)\}$

2. $\mathcal{L}\{\lambda f(t)\} = \lambda\mathcal{L}\{f(t)\}$

3. $\mathcal{L}\{f(t)\cdot g(t)\} \neq \mathcal{L}\{f(t)\}\cdot\mathcal{L}\{g(t)\}$

A more elegant way of obtaining the result for trigonometric functions is to consider

$$\begin{aligned}\mathcal{L}\left\{e^{ikt}\right\} &= \frac{1}{s-ik} = \frac{s+ik}{s^2+k^2} \equiv \mathcal{L}\{\cos kt + i\sin kt\} \\ &= \frac{s}{s^2+k^2} + ik\frac{1}{s^2+k^2}\end{aligned}$$

Hence

$$\mathcal{L}\{\cos kt\} = \frac{s}{s^2+k^2} \quad , \qquad \mathcal{L}\{\sin kt\} = \frac{k}{s^2+k^2}$$

Shift Theorem

If $\mathcal{L}\{f(t)\} = F(s)$, then

$$\mathcal{L}\left\{e^{at} f(t)\right\} = F(s-a)$$

or $f(t) = t^n$ with $F(s) = \frac{n!}{s^{n+1}}$, then

$$\mathcal{L}\left\{e^{at} t^n\right\} = F(s-a) = \frac{n!}{(s-a)^{n+1}}$$

Table of Standard Forms

$f(t)$	$F(s)$
1	$\frac{1}{s}$
$e^{\pm at}$	$\frac{1}{s \mp a}$
t^n	$\frac{n!}{s^{n+1}}$
t^α	$\frac{\Gamma(\alpha+1)}{s^{\alpha+1}}$
$\sin kt$	$\frac{k}{k^2+s^2}$
$\cos kt$	$\frac{s}{s^2+k^2}$
$t \sin kt$	$\frac{2ks}{\left(k^2+s^2\right)^2}$
$t \cos kt$	$\frac{s^2-k^2}{\left(s^2+k^2\right)^2}$

The Inverse Laplace Transform

Given $F(s)$:

$\mathcal{L}^{-1}(F(s)) =$ the function $f(t)$ for which $\mathcal{L}\{f(t)\} = F(s)$.

Therefore we can read the table of Laplace Transforms in two ways

$$\begin{array}{cc} f(t) & F(s) \\ \mathcal{L} & \\ \longrightarrow & \\ & \mathcal{L}^{-1} \\ & \longleftarrow \end{array}$$

For example reading the earlier table from right to left we have

$$\begin{aligned} \mathcal{L}^{-1}\left(\frac{1}{s}\right) &= 1 \\ \mathcal{L}^{-1}\left(\frac{1}{s^{n+1}}\right) &= \frac{t^n}{n!} \end{aligned}$$

Example 1 : Evaluate $\mathcal{L}^{-1}\left(\dfrac{1}{s^2+2s+2}\right)$.

The structure should suggest some variation of

$$\mathcal{L}^{-1}\left(\frac{1}{s^2}\right) \quad \text{or} \quad \mathcal{L}^{-1}\left(\frac{1}{s^2+k^2}\right).$$

Now $s^2+2s+2=(s+1)^2+1^2$. Hence

$$\mathcal{L}^{-1}\left(\frac{1}{s^2+2s+2}\right) \equiv \mathcal{L}^{-1}\left(\frac{1}{(s+1)^2+1^2}\right) \quad \text{c.f.} \quad \mathcal{L}^{-1}\left(\frac{1}{s^2+k^2}\right)$$

with s replaced by $s+1$.

From the table:

$$\mathcal{L}\left\{e^{at}\sin kt\right\} = \frac{k}{(s+a)^2+k^2}$$

So $a=-1,\ k=1$

$$\therefore \mathcal{L}\left\{e^{-t}\sin t\right\} = \frac{1}{(s+1)^2+1^2} \text{ and } \mathcal{L}^{-1}\left(\frac{1}{s^2+2s+2}\right) = e^{-t}\sin t$$

Example 2 : $\mathcal{L}^{-1}\left(\dfrac{1}{(s+1)(s+2)}\right)$

Using partial fractions this becomes $\mathcal{L}^{-1}\left\{\dfrac{1}{(s+1)} - \dfrac{1}{(s+2)}\right\}$

$$= \mathcal{L}^{-1}\left\{\frac{1}{(s+1)}\right\} - \mathcal{L}^{-1}\left\{\frac{1}{(s+2)}\right\}$$

We know

$$\mathcal{L}\left\{e^{at}\right\} = \frac{1}{(s-a)} \therefore \mathcal{L}\left\{e^{-t}\right\} = \frac{1}{s+1} \;\&\; \mathcal{L}\left\{e^{-2t}\right\} = \frac{1}{s+2}$$

Hence we have

$$e^{-t} - e^{-2t}$$

Use of Laplace Transforms For Solving Differential Equations

Laplace Transforms provide us with a very useful technique for solving IVP's.

This is based on two results (which are also theorems):

1. $\mathcal{L}\{f'(t)\} = sF(s) - f(0)$

2. $\mathcal{L}\{f''(t)\} = s^2F(s) - sf(0) - f'(0)$

$\mathcal{L}\{f'(t)\} = \int_0^\infty e^{-st} f'(t)\,dt$ and we integrate by parts to give

$$f(t)\,e^{-st}\Big|_0^\infty + s\int_0^\infty e^{-st} f(t)\,dt = -f(0) + sF(s)$$

$\mathcal{L}\{f''(t)\} = \int_0^\infty e^{-st} f''(t)\,dt$ which requires integration by parts twice (and includes use of the earlier result)

$$\begin{aligned} f'(t)\,e^{-st}\Big|_0^\infty + s\int_0^\infty e^{-st} f'(t)\,dt &= -f'(0) + s\left(-f(0) + sF(s)\right) \\ &= s^2F(s) - sf(0) - f'(0) \end{aligned}$$

We can use these to transform an IVP such as

$$\left.\begin{array}{l} ay'' + by' + cy = f(t) \\ y(0) = \alpha;\quad y'(0) = \beta \end{array}\right\}$$

into an algebraists problem by taking Laplace Transforms of both sides of the ode. Let $Y(s) = \mathcal{L}\{y(t)\}$ then

$$a\left(s^2Y(s) - sy(0) - y'(0)\right) + b\left(sY(s) - y(0)\right) + cY(s) = F(s)$$

and simplifying

$$Y(s) = \frac{F(s) + ay'(0) + (as+b)\,y(0)}{as^2 + bs +}.$$

The final step in the working entails taking inverse Laplace Transforms to have $Y(s) \longrightarrow y(t)$.

Example: Solve the IVP

$$\left.\begin{array}{c} y'' + 2y' + y = 0 \\ y(0) = 2;\ y'(0) = -1. \end{array}\right\}$$

Start by taking LT of the whole equation, with $Y(s) = \mathcal{L}\{y(t)\}$

$$\begin{aligned} \mathcal{L}\left\{y'' + 2y' + y\right\} &= \mathcal{L}\{0\} \\ \mathcal{L}\left\{y''\right\} + 2\mathcal{L}\left\{y'\right\} + \mathcal{L}\{y\} &= 0 \end{aligned}$$

The initial condition simplifies the problem considerably to give

$$\begin{aligned} s^2Y(s) - sY(0) - Y'(0) + 2(sY(s) - Y(0)) + Y(s) &= 0 \\ Y(s)\left(s^2 + 2s + 1\right) - 2s - 3 &= 0 \\ Y(s)\left(s^2 + 2s + 1\right) &= 2s + 3 \\ Y(s) &= \frac{2s+3}{(s+1)^2} \end{aligned}$$

The right hand side can be written as

$$\frac{2s+3}{\left(s^2+2s+1\right)} \equiv \frac{A}{(s+1)} + \frac{B}{(s+1)^2}$$

for $A = 2,\ B = 1$ to give

$$Y(s) = \frac{2}{(s+1)} + \frac{1}{(s+1)^2}$$

Now taking inverse LT's

$$\mathcal{L}^{-1}\{Y(s)\} = \mathcal{L}^{-1}\left\{\frac{2}{(s+1)} + \frac{1}{(s+1)^2}\right\}$$

therefore

$$y(t) = 2\mathcal{L}^{-1}\left\{\frac{1}{(s+1)}\right\} + \mathcal{L}^{-1}\left\{\frac{1}{(s+1)^2}\right\}$$

Use the Shift Theorem for the second term

$$\mathcal{L}\left\{e^{at} f(t)\right\} = F(s-a)$$

we get a te^{-t}.

So the solution is

$$y(t) = 2e^{-t} + te^{-t}$$

Example: Solve the IVP

$$\left.\begin{array}{c} y'' + 4y' + 6y = 1 + e^{-t} \\ y(0) = y'(0) = 0. \end{array}\right\}$$

Start by taking LT of the whole equation, with $Y(s) = \mathcal{L}\{y(t)\}$

$$\begin{aligned} \mathcal{L}\left\{y'' + 4y' + 6y\right\} &= \mathcal{L}\left\{1 + e^{-t}\right\} \\ \mathcal{L}\left\{y''\right\} + 4\mathcal{L}\left\{y'\right\} + 6\mathcal{L}\{y\} &= \mathcal{L}\{1\} + \mathcal{L}\left\{e^{-t}\right\} \end{aligned}$$

The initial condition simplifies the problem considerably to give

$$\begin{aligned} s^2Y(s) + 4sY(s) + 6Y(s) &= \left(\frac{1}{s} + \frac{1}{s+1}\right) \\ Y(s) &= \frac{1}{s^2 + 4s + 6}\left(\frac{1}{s} + \frac{1}{s+1}\right) \end{aligned}$$

Now taking inverse LT's

$$\mathcal{L}^{-1}\{Y(s)\} = \mathcal{L}^{-1}\left\{\frac{1}{s\left(s^2+4s+6\right)} + \frac{1}{\left(s^2+4s+6\right)(s+1)}\right\}$$

after firstly using partial fractions, i.e. decompose as

$$\frac{1}{6s} + \frac{1}{3(s+1)} - \frac{s+2}{2\left(s^2+4s+6\right)} - \frac{2}{3\left(s^2+4s+6\right)}$$

therefore

$$\begin{aligned} y(t) &= \frac{1}{6}\mathcal{L}^{-1}\left\{\frac{1}{s}\right\} + \frac{1}{3}\mathcal{L}^{-1}\left\{\frac{1}{(s+1)}\right\} - \frac{1}{2}\mathcal{L}^{-1}\left\{\frac{s+2}{\left(s^2+4s+6\right)}\right\} \\ &\quad -\frac{2}{3}\mathcal{L}^{-1}\left\{\frac{1}{\left(s^2+4s+6\right)}\right\} \end{aligned}$$

We know $s^2+4s+6 \equiv (s+2)^2+2$. So use the Shift Theorem

$$\mathcal{L}\left\{e^{at}f(t)\right\} = F(s-a)$$

$$\mathcal{L}\{\cos kt\} = \frac{s}{s^2+k^2}; \quad \mathcal{L}\{\sin kt\} = \frac{k}{s^2+k^2}; \quad \mathcal{L}\{1\} = \frac{1}{s}$$

So the solution is

$$y(t) = \frac{1}{6} + \frac{1}{3}e^{-t} - \frac{1}{2}e^{-2t}\cos\sqrt{2}t - \frac{2}{3}e^{-2t}\cos\sqrt{2}t.$$

Theoretically, we can also extend this to an n^{th} order IVP (with constant coefficients) by finding $\mathcal{L}\left\{f^{(n)}(t)\right\}$.

Introduction to Numerical Methods

Numerical Integration

The need frequently arises for evaluating the definite integral of a function that has no explicit antiderivative or whose antiderivative is not easily obtainable.

Recall the fundamental theorem of calculus

$$\int_a^b f(x)\,dx = F(b) - F(a)$$

Consider the usual (Riemann) definition of a definite integral

$$\int_a^b f(x)\,dx$$

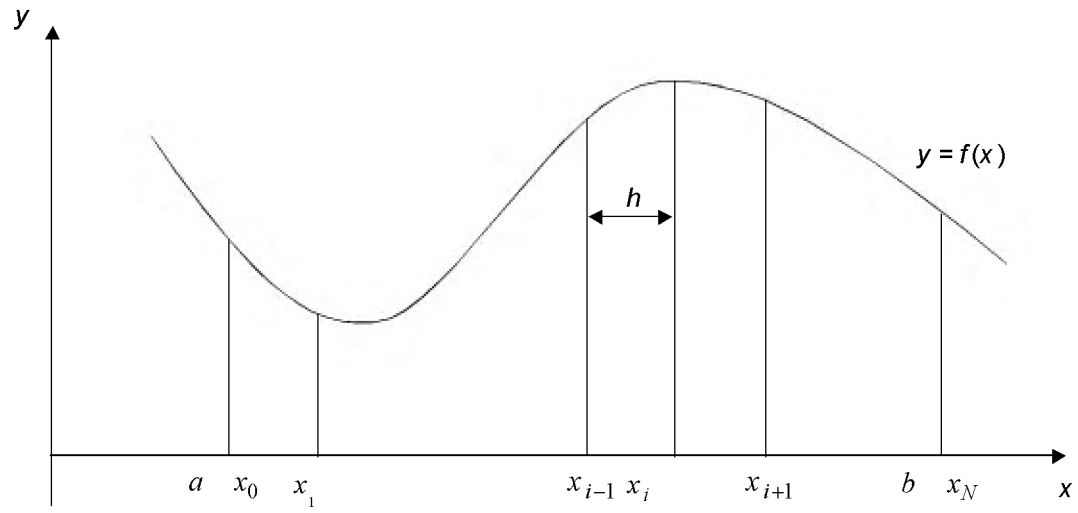
as the area under the curve where the curve is the graph of $f\ (x)$ plotted against x.

Assuming f is a "well behaved" function on $[a,\ b]$, we can define this in many different ways (which all lead to the same value for the definite integral).

Divide the interval $[a,\ b]$ into N intervals with end points $x_0 = a < x_1 < x_2 < \ ..... \ < x_{N-1} < x_N = b.$

Then $h = (b-a)/N$ is the constant length of an interval $x_{i+1} - x_i$, which tends to zero as $N \to \infty$.

$x_i = a + ih$ for each $j = 0,\ 1,\ ....., \ N.$



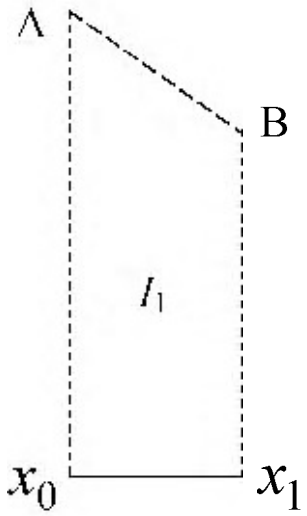
We could approximate the definite integral by *The trapezium rule*;

$$\int_a^b f(x)\,dx = \frac{h}{2}\left[f(a) + 2\sum_{i=1}^{N-1} f(x_i) + f(b)\right].$$

Each strip is labelled I_i for $1 \leq i \leq N$. So the area I under the curve between $x = a$ and $x = b$ is given by

$$I = I_1 + I_2 + ... + I_N \tag{1}$$

Put $f(x_0) = A$ and $f(x_1) = B$.Consider the first strip (a trapezium). That is, points A and B are joined with a straight line and the area of the trapezium ABx_1x_0 is assumed to be a good approximation to the area under the curve.



The area ABx_1x_0 is simply

$$I_1 \approx \frac{h}{2}(f_0 + f_1)$$

An approximation to

$$I_1 = \int_{x_0}^{x_1} f(x)\, dx$$

is now given by

$$I_1 \approxeq \frac{h}{2}(f_0 + f_1) \tag{2}$$

We can extend this to consider N strips. We can write (1) in integral form as

$$\begin{aligned}\int_{x_0}^{x_N} f(x)\,dx &= \int_{x_0}^{x_1} f(x)\,dx + \int_{x_1}^{x_2} f(x)\,dx + \ldots\ldots\ldots \\ &\quad + \int_{x_N-1}^{x_N} f(x)\,dx\end{aligned}$$

From (2)

$$I_1 = \frac{h}{2}(f_0 + f_1)$$

and by extension

$$\begin{aligned}I_2 &= \frac{h}{2}(f_1 + f_2) \\ I_3 &= \frac{h}{2}(f_2 + f_3), \ldots\ldots\ldots\end{aligned}$$

So summing these quantities gives the following $I =$

$$\frac{h}{2}\left[(f_0 + f_1) + (f_1 + f_2) + (f_2 + f_3) + ... (f_{N-1} + f_N)\right]$$

OR

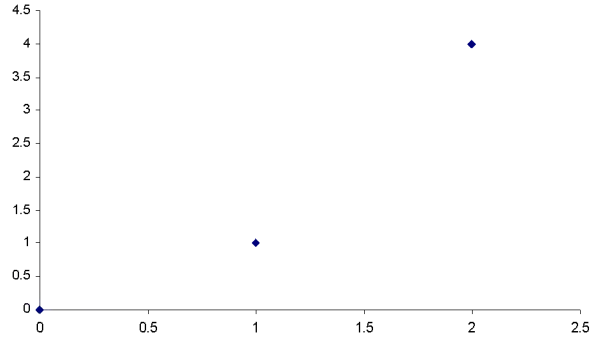
$$I = \frac{h}{2}\left[f_0 + 2f_1 + 2f_2 + ... + 2f_{N-1} + f_N\right] \qquad (3)$$

This is the *Composite Trapezoidal Rule*.

Lagrangian Interpolation

In general the term **interpolation** is the term used to describe the process in which a function, usually a polynomial function is constructed to pass through a given set of data points.

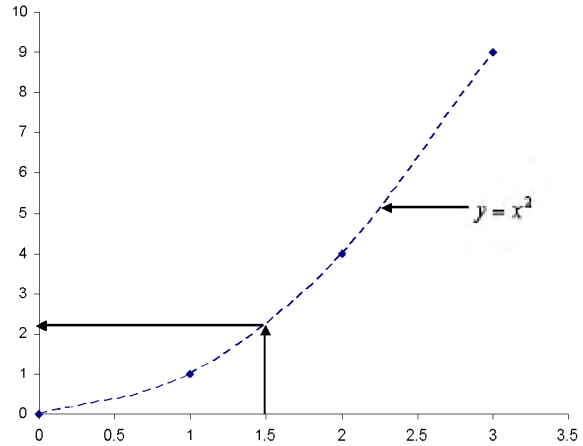
As an example consider the following data points (x_i, y_i) as follows: $(0, 0)$; $(1, 1)$; $(2, 4)$



In this example the function $y = x^2$ passes through these 3 data points.

Why is interpolation important and how is it used?

Suppose an estimate is required for the value of y at some x other than $x = 0, 1, 2$. In this case, the interpolating polynomial $y = x^2$ can be used to estimate such intermediate values, e.g. $x = 1.5$, which gives $y = 2.25$



Given the data points $(0,0)$; $(1,1)$; $(2,4)$ the values of y corresponding to any value of x between $x = 0$ and $x = 2$ can be estimated. This is known as *interpolation*.

Lagrange Polynomials

For a given set of $n+1$ nodes x_i,

i	0	1	2	.....	n
x_i	x_0	x_1	x_2	.....	x_i
y_i	y_0	y_1	y_2	.....	y_n

the Lagrange polynomials are the $n+1$ polynomials l_i defined by

$$l_i(x) = \prod_{j=0,\ j\neq i}^{n} \frac{x - x_j}{x_i - x_j} \tag{1}$$

where

$$l_i\left(x_j\right) = \delta_{ij} = \begin{cases} 0 \text{ if } i \neq j \\ 1 \text{ if } i = j \end{cases}$$

Then we define the interpolating polynomial as

$$p_n(x) = \sum_{i=0}^{n} y_i l_i(x)$$

If each Lagrange Polynomial is of degree at most n, then p_n also has this property. The Lagrange Polynomials can be characterized as follows:

By evaluating this product for each x_j, we see that this is indeed a characterization of the Lagrange Polynomials. Moreover, each polynomial is clearly the product of n "mononomials", and thus has degree no greater than n.

Example

Construct the polynomial interpolating the data

x	1	$\frac{1}{2}$	3
y	3	-10	2

by using Lagrange Polynomials.

Solution:

i	0	1	2
x_i	1	$\frac{1}{2}$	3
y_i	3	-10	2

We construct the Lagrange Polynomials:

So we start with $i = 0$, so need x_1, x_2

$$\begin{aligned} l_0(x) &= \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} \\ &= \frac{\left(x - \frac{1}{2}\right)(x - 3)}{\left(1 - \frac{1}{2}\right)(1 - 3)} \\ &= -\left(x - \frac{1}{2}\right)(x - 3) \end{aligned}$$

need x_0, x_2

$$\begin{aligned} l_1(x) &= \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} \\ &= \frac{(x - 1)(x - 3)}{\left(\frac{1}{2} - 1\right)\left(\frac{1}{2} - 3\right)} \\ &= \frac{4}{5}(x - 1)(x - 3) \end{aligned}$$

need x_0, x_1

$$\begin{aligned} l_2(x) &= \frac{(x-x_0)(x-x_1)}{(x_2-x_1)(x_2-x_1)} \\ &= \frac{(x-1)\left(x-\frac{1}{2}\right)}{(3-1)\left(3-\frac{1}{2}\right)} \\ &= \frac{1}{5}(x-1)\left(x-\frac{1}{2}\right) \end{aligned}$$

Then the interpolating polynomial, in "Lagrange Form" is

$$\begin{aligned} p_2(x) &= \sum_{i=0}^{n} y_i l_i(x) = y_0 l_0(x) + y_1 l_1(x) + y_2 l_2(x) \\ &= 3l_0(x) - 10l_1(x) + 2l_2(x) \\ &= -3\left(x - \frac{1}{2}\right)(x-3) - 8(x-1)(x-3) \\ &\quad + \frac{2}{5}(x-1)\left(x - \frac{1}{2}\right) \end{aligned}$$

A useful check on the correctness of the calculations performed is

$$\sum_{i=0}^{n} l_i(x) = 1$$

Simpson's Rule

Simpson's Rule is obtained by considering two strips together, fitting a quadratic polynomial through the points on the curve and then integrating the quadratic

(diagram 1)

Let

$$I = \int_{x_0}^{x_2} f(x)\, dx \tag{2}$$

To find I first fit a quadratic to the points ABC. By Lagrangian Interpolation of degree 2

$$f(x) \approxeq p_2(x) = f_0 l_0 + f_1 l_1 + f_2 l_2 \tag{3}$$

$$\begin{aligned} l_0 &= \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)}; \; l_1(x) = \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} \\ l_2(x) &= \frac{(x - x_0)(x - x_1)}{(x_2 - x_1)(x_2 - x_1)} \end{aligned}$$

Now

$$\begin{aligned}(x_0 - x_1) &= -h \; ; \; (x_0 - x_2) = -2h \\ (x_1 - x_0) &= h \; ; \; (x_1 - x_2) = -h \\ (x_2 - x_1) &= h \; ; \; (x_2 - x_0) = 2h\end{aligned}$$

So from (3)

$$\begin{aligned}f(x) \cong\; & p_2(x) = \frac{(x - x_1)(x - x_2)}{2h^2} f_0 - \\ & \frac{(x - x_0)(x - x_2)}{h^2} f_1 + \frac{(x - x_0)(x - x_1)}{2h^2} f_2\end{aligned}$$

$f(x)$ is now substituted into (2) to give

$$I \approx \frac{1}{2h^2}\int_{x_0}^{x_2}\{(x-x_1)(x-x_2)f_0 \quad (4)$$
$$-2(x-x_0)(x-x_2)f_1+(x-x_0)(x-x_1)f_2\}dx$$

(4) is integrated by parts to give

$$I \approx \frac{h}{3}(f_0+4f_1+f_2) \quad (5)$$

This is *Simpson's Rule*.

Composite Form of Simpson's Rule

This is similar in principle to the Composite Trapezoidal Rule but here the region of integration is split into pairs of strips as shown in the diagram

(diagram 2)

$$I = \int_{x_0}^{x_{2n}} f(x)\, dx$$

is now split into pairs of strips such that

$$I = \int_{x_0}^{x_2} f(x)\, dx + \int_{x_2}^{x_4} f(x)\, dx + ... + \int_{x_{2n-2}}^{x_{2n}} f(x)\, dx$$

Each integral on the right hand side of this expression above is now evaluated by Simpson's Rule.

This gives from (5) applied in the appropriate interval

$$I \approxeq \frac{h}{3}\left(\begin{array}{c} f_0 + 4f_1 + 2f_2 + 4f_3 + 2f_4 + ... \\ +2f_{2n-2} + 4f_{2n-1} + f_{2n} \end{array} \right)$$

so that

$$I \approxeq \frac{h}{3}\left(f_0 + f_{2n} + 4\sum_{i=1}^{n} f_{2i-1} + 2\sum_{i=1}^{n-1} f_{2i} \right) \tag{6}$$

This is the composite form of Simpson's Rule.

Numerical Linear Algebra

Definition An **upper triangular** matrix U has zero elements below the principal diagonal, i.e.

$$u_{ij} = 0 \quad \forall\, i > j.$$

For an upper triangular matrix U consider the system $U\underline{x} = \underline{b}$ in component form

$$\begin{aligned}
u_{11}x_1 + u_{12}x_2 + \ldots\ldots\ldots\ldots\ldots\ldots + u_{1n}x_n &= b_1 \\
u_{22}x_2 + \ldots\ldots\ldots\ldots\ldots\ldots + u_{2n}x_n &= b_2 \\
u_{33}x_3 + \ldots\ldots\ldots\ldots\ldots\ldots + u_{3n}x_n &= b_3 \\
\ddots \qquad\qquad\qquad\qquad\qquad\qquad\quad \vdots &= \vdots \\
\ddots \qquad\qquad\qquad\qquad\qquad\qquad\quad \vdots &= \vdots \\
\ddots \qquad\qquad\qquad\qquad\qquad\qquad\quad \vdots &= \vdots \\
u_{n-1,n-1}x_{n-1} + u_{n-1,n}x_n &= b_{n-1} \\
u_{nn}x_n &= b_n
\end{aligned}$$

The determinant of the matrix U , can now easily be obtained from

$$\det(U) = \prod_{i=1}^{n} u_{ii}.$$

Recall that U has an inverse if $\det(U) \neq 0$, hence for U to be non-singular requires the strict condition $u_{ii} \neq 0 \ \ \forall\ i$. This linear system can now easily be solved by **back-substitution**. The algorithm for this scheme becomes

$$\begin{aligned} x_n &= \frac{b_n}{u_{nn}} \\ x_i &= \frac{1}{u_{ii}} \left(b_i - \sum_{j=i+1}^{n} u_{ij} x_j \right) \text{ for } i = n-1,\ n-2,\ \ldots, 1 \end{aligned}$$

Definition A **lower triangular** matrix L has zero elements above the principal diagonal, i.e.

$$l_{ij} = 0 \ \ \forall\ i < j.$$

Now suppose L is a lower triangular matrix and consider the system $L\underline{x} = \underline{b}$.

A *lower triangular* matrix L has zero elements above the main diagonal, i.e.

$$l_{ij} = 0 \ \ \forall \ i < j.$$

Now suppose L is a lower triangular matrix and consider the system $L\underline{x} = \underline{b}$.

$$\begin{array}{rcl}
l_{11}x_1 & = & b_1 \\
l_{21}x_1 + l_{22}x_2 & = & b_2 \\
\vdots \quad \vdots \quad \ddots & = & \vdots \\
\vdots \quad \vdots \quad \quad \ddots & = & \vdots \\
\vdots \quad \vdots \quad \quad \quad \ddots & = & \vdots \\
l_{n1}x_1 + l_{n2}x_2 + \ldots\ldots\ldots + l_{nn}x_n & = & b_n
\end{array}$$

Again, because

$$\det(L) = \prod_{i=1}^{n} l_{ii},$$

L is non-singular iff $l_{ii} \neq 0 \ \forall\, i$. We solve this system by **forward-substitution**. The scheme is given by

$$\begin{aligned} x_1 &= \frac{b_1}{l_{11}} \\ x_i &= \frac{1}{l_{ii}} \left(b_i - \sum_{j=1}^{i-1} l_{ij} x_j \right) \quad \text{for } i = 2,\ 3....,\ n \end{aligned}$$

Definition An $n \times n$ matrix is called a *band matrix* if integers p and q exist such that $p,\ q \in (1, n)$ with the property that $a_{ij} = 0$ whenever $i + p \leq j$ or $j + q \leq i$. The band width w of a banded matrix is defined to be $w = p + q - 1$.

So for example, the matrix

$$A = \begin{pmatrix} 7 & 2 & 0 \\ 3 & 5 & -1 \\ 0 & -5 & -6 \end{pmatrix}$$

is a banded matrix with $p = q = 2$ and band width $w = 3$.

So the definition of a banded matrix forces such matrices to have all its non-zero elements to be clustered close to the main diagonal.

p is then called the *upper band-width* and q the *lower band-width*.

Solving equations with banded matrices can be much more efficient than with full matrices because the triangular factors in the LU decomposition are also banded. This results in greater efficiency in terms of storage and amount of computations required. Banded matrices arise

from the discretization of differential equations, because the finite difference method use to approximate derivatives only involve values at nearby mesh points.

From an applied mathematics perspective, the special case occurring when $p = q = 2$ and band width $w = 3$ is the most useful for us. These matrices are called *tridiagonal* since they have the form

$$\begin{pmatrix} a_{11} & a_{12} & 0 & \cdots & \cdots & 0 \\ a_{21} & a_{22} & a_{23} & \ddots & & \vdots \\ 0 & a_{32} & a_{33} & a_{34} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \ddots & a_{n-1,n} \\ 0 & \cdots & \cdots & 0 & a_{n,n-1} & a_{nn} \end{pmatrix}$$

The LU Decomposition

It is often advantageous to think of Gaussian elimination as constructing a lower tridiagonal matrix L and an upper triangular matrix U, so that $LU = A$. To illustrate this method consider the following *tridiagonal* linear system

$$\begin{pmatrix} b_0 & 0 & 0 & \cdots & \cdots & 0 \\ a_1 & b_1 & 1 & & & \vdots \\ 0 & a_2 & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & 0 \\ \vdots & & & a_{n-1} & b_{n-1} & n-1 \\ 0 & \cdots & \cdots & 0 & a_n & b_n \end{pmatrix} \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ \vdots \\ y_{n-1} \\ y_n \end{pmatrix} = \begin{pmatrix} p_0 \\ p_1 \\ p_2 \\ \vdots \\ p_{n-1} \\ p_n \end{pmatrix} \tag{1}$$

We can write this in the form

$$\mathbf{M}.\mathbf{y} = \mathbf{p}$$

and then consider the matrix factorisation $\mathbf{M} = \mathbf{LU}$

$$= \begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ l_1 & 1 & 0 & & & \vdots \\ 0 & l_2 & 1 & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & l_{n-1} & 1 & 0 \\ 0 & \cdots & \cdots & 0 & l_n & 1 \end{pmatrix} \begin{pmatrix} d_0 & u_0 & 0 & \cdots & \cdots & 0 \\ 0 & d_1 & u_1 & & & \vdots \\ 0 & 0 & d_2 & u_2 & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & 0 \\ \vdots & & & 0 & d_{n-1} & u_{n-1} \\ 0 & \cdots & \cdots & 0 & 0 & d_n \end{pmatrix}$$

$$= \begin{pmatrix} d_0 & u_0 & 0 & \cdots & \cdots & 0 \\ l_1 d_0 & l_1 u_0 + d_1 & u_1 & & & \vdots \\ 0 & l_2 d_1 & l_2 u_1 + d_2 & u_2 & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & 0 \\ \vdots & & & l_{n-1} d_{n-2} & l_{n-1} u_{n-2} + d_{n-1} & u_{n-1} \\ 0 & \cdots & \cdots & 0 & l_n d_{n-1} & l_n u_{n-1} + d_n \end{pmatrix}$$

where L is the lower triangular matrix above and U is the upper triangular matrix above.

Equating the elements of the original tridiagonal matrix M with the elements of the product matrix LU, we find that

$$\begin{aligned} d_0 &= b_0 \\ u_i &= {}_i, \quad i = 0, 1, 2, ..., n-1 \\ l_i &= \frac{a_i}{d_{i-1}}, \quad d_i = b_i - l_i u_{i-1}, \quad i = 1, 2, ..., n \end{aligned} \tag{2}$$

We then solve the original system given by (1) by solving two smaller problems. The tridiagonal system (1) is written in the form

$$\mathbf{M.y} = (\mathbf{LU}) \mathbf{.y} = \mathbf{L} (\mathbf{U.y}) = \mathbf{p}$$

We introduce an intermediate vector $\mathbf{z} = \mathbf{Uy}$ so that (1) becomes

$$\mathbf{Lz} = \mathbf{p}, \quad \mathbf{Uy} = \mathbf{z}.$$

First we solve the problem

$$\mathbf{Lz} = \mathbf{p}$$

for the intermediate vector $\mathbf{z}$, i.e. we solve the system

$$\begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ l_1 & 1 & 0 & & & \vdots \\ 0 & l_2 & 1 & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & l_{n-1} & 1 & 0 \\ 0 & \cdots & \cdots & 0 & l_n & 1 \end{pmatrix} \begin{pmatrix} z_0 \\ z_1 \\ \vdots \\ \vdots \\ z_{n-1} \\ z_n \end{pmatrix} = \begin{pmatrix} p_0 \\ p_1 \\ p_2 \\ \vdots \\ p_{n-1} \\ p_n \end{pmatrix} \tag{3}$$

Trivially, the z_i can be found by forward substitution so

$$z_0 = p_0, \quad z_i = p_i - l_i z_{i-1}, \quad i = 1, 2, ..., n$$

This determines the vector $\mathbf{z}$.

Having obtained $\mathbf{z}$ we find $\mathbf{y}$ by solving $\mathbf{U}\mathbf{y} = \mathbf{z}$;

$$\begin{pmatrix} d_0 & u_0 & 0 & \cdots & \cdots & 0 \\ 0 & d_1 & u_1 & & & \vdots \\ 0 & 0 & d_2 & u_2 & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & 0 \\ \vdots & & & 0 & d_{n-1} & u_{n-1} \\ 0 & \cdots & \cdots & 0 & 0 & d_n \end{pmatrix} \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ \vdots \\ y_{n-1} \\ y_n \end{pmatrix} = \begin{pmatrix} z_0 \\ z_1 \\ \vdots \\ \vdots \\ z_{n-1} \\ z_n \end{pmatrix} \tag{4}$$

The solution of (4) is trivially obtained by backward substitution

$$y_n = \frac{z_n}{d_n}, \quad y_i = \frac{z_i - u_i y_{i+1}}{d_i}, \quad i = n-1, n-2, ..., 2, 1$$

This gives us the solution, $\mathbf{y}$, of our original problem (1), $\mathbf{My} = \mathbf{p}$.

This method can be also extended to decompose non-sparse matrices, i.e. $A = LU$, thus opening up a wider class of associated methods.

$$= \begin{pmatrix} l_{11} & 0 & \cdots & \cdots & \cdots & 0 \\ l_{21} & l_{22} & 0 & & & \vdots \\ \vdots & l_{32} & & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & l_{n-1} & l_{n-1,n-1} & 0 \\ l_{n1} & l_{n2} & \cdots & \cdots & \cdots & l_{nn} \end{pmatrix} \times \begin{pmatrix} u_{11} & u_{12} & \cdots & \cdots & \cdots & u_{1n} \\ 0 & u_{22} & u_{23} & \cdots & \cdots & u_{2n} \\ 0 & 0 & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & 0 & u_{n-1,n-1} & u_{n-1,n} \\ 0 & \cdots & \cdots & 0 & 0 & u_{nn} \end{pmatrix}$$

If $l_{ii} = 1 \ \ \forall \ \ 1 \leq i \leq n$ as mentioned in the earlier discussion, this ensures a unique solution and is called *Doolittle's method*. *Crout's method* requires $u_{ii} = 1 \ \ \forall \ \ 1 \leq i \leq n$.

Example:

Consider the matrix $\begin{pmatrix} 2 & -1 & 1 \\ 3 & 3 & 9 \\ 3 & 3 & 5 \end{pmatrix}$ which can be factorised into

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 1.5 & 1 & 0 \\ 1.5 & 1 & 1 \end{pmatrix}, \quad U = \begin{pmatrix} 2 & -1 & 1 \\ 0 & 4.5 & 7.5 \\ 0 & 0 & -4 \end{pmatrix}$$

Iterative Techniques

Seldom used for low dimensional problems - Gaussian elimination preferred.

For large systems with a high percentage of zero entries these are efficient in terms of

1. Computer Storage

2. Computational Time

Consider $(n \times n)$ linear system: $A\underline{x} = \underline{b}$

Initial approximation $\underline{x}^{(0)}$ to solution $\underline{x}$ which generates a sequence of vectors $\left\{\underline{x}^{(k)}\right\}_{k=0}^{\infty}$ that converges to $\underline{x}$.

Most techniques involve a process which converts $A\underline{x} = \underline{b}$ to

$$\underline{x} = T\underline{x} + \_$$

where T is an $(n \times n)$ matrix and $\_$ is a vector.

The initial vector $\underline{x}^{(0)}$ is selected and a sequence of approximations generated by computing

$$\underline{x}^{(k)} = T\underline{x}^{(k-1)} + \_$$

for each $k = 1,\ 2,\ 3,\ ....$

$$\begin{aligned}
\mathsf{E}_1: &\quad 10x_1 - x_2 + 2x_3 &&= 6\\
\mathsf{E}_2: &\quad -x_1 + 11x_2 - x_3 + 3x_4 &&= 25\\
\mathsf{E}_3: &\quad 2x_1 - x_2 + 10x_3 - x_4 &&= -11\\
\mathsf{E}_4: &\quad 3x_2 - x_3 + 8x_4 &&= 15
\end{aligned}$$

Now consider $A\underline{x} = \underline{b}$ with

$$A = \begin{bmatrix} 10 & -1 & 2 & 0 \\ -1 & 11 & -1 & 3 \\ 2 & -1 & 10 & -1 \\ 0 & 3 & -1 & 8 \end{bmatrix}, \quad \underline{b} = \begin{bmatrix} 6 \\ 25 \\ -11 \\ 15 \end{bmatrix}$$

which has an exact solution

$$\underline{x} = \begin{bmatrix} 1 \\ 2 \\ -1 \\ 1 \end{bmatrix}$$

To convert $A\underline{x} = \underline{b}$ to the form $\underline{x} = T\underline{x} + \_$; solve E_i for each x_i $(i = 1,\ 2,\ 3,\ 4)$ to give

$$\left.\begin{aligned} x_1^{(k)} &= \tfrac{1}{10}\left\{x_2^{(k-1)} - 2x_3^{(k-1)} + 6\right\} \\ x_2^{(k)} &= \tfrac{1}{11}\left\{x_1^{(k-1)} + x_3^{(k-1)} - 3x_4^{(k-1)} + 25\right\} \\ x_3^{(k)} &= \tfrac{1}{10}\left\{-2x_1^{(k-1)} + x_2^{(k-1)} + x_4^{(k-1)} - 11\right\} \\ x_4^{(k)} &= \tfrac{1}{8}\left\{-3x_2^{(k-1)} + x_3^{(k-1)} + 15\right\} \end{aligned}\right\} \qquad (5)$$

So (5) can be written as

$$\left.\begin{aligned} x_1^{(k)} &= \tfrac{1}{10}\left\{x_2^{(k-1)} - 2x_3^{(k-1)}\right\} + \tfrac{3}{5} \\ x_2^{(k)} &= \tfrac{1}{11}\left\{x_1^{(k-1)} + x_3^{(k-1)} - 3x_4^{(k-1)}\right\} + \tfrac{25}{11} \\ x_3^{(k)} &= \tfrac{1}{10}\left\{-2x_1^{(k-1)} + x_2^{(k-1)} + x_4^{(k-1)}\right\} - \tfrac{11}{10} \\ x_4^{(k)} &= \tfrac{1}{8}\left\{-3x_2^{(k-1)} + x_3^{(k-1)}\right\} + \tfrac{15}{8} \end{aligned}\right\}$$

$\Rightarrow$

$$T = \begin{bmatrix} 0 & \frac{1}{10} & -\frac{1}{5} & 0 \\ \frac{1}{11} & 0 & \frac{1}{11} & -\frac{3}{11} \\ -\frac{1}{5} & \frac{1}{10} & 0 & \frac{1}{10} \\ 0 & -\frac{3}{8} & \frac{1}{8} & 0 \end{bmatrix} \qquad \_ = \begin{bmatrix} \frac{3}{5} \\ \frac{25}{11} \\ -\frac{11}{10} \\ \frac{15}{8} \end{bmatrix}$$

For initial approximation let $\underline{x}^{(0)} = \underline{0}^{\mathsf{T}}$ and generate $\underline{x}^{(1)}$ by substi-

tuting $(0,\ 0,\ 0,\ 0)^{\mathsf{T}}$ in (5) to get

$$(0.6,\ 2.2727, -1.1000,\ 1.8750) = \underline{x}^{(1)}$$

and $\underline{x}^{(2)}$ can now be obtained.

In general obtain $\underline{x}^{(k)} = \left(x_1^{(k)},\ x_2^{(k)},\ x_3^{(k)},\ x_4^{(k)}\right)^{\mathsf{T}}$.

By the very nature of iterative techniques, these continue indefinitely, hence one requires a convergence criterion to terminate the computation once the imposed condition is satisfied. So we need to know when to stop iterating, i.e. does the sequence converge to the solution of the given system of equations? To answer this, we need a means of measuring the distance between $n-$ dimensional vectors.

Vector Norms

In two and three dimensions, the size of vectors called the modulus is generally obtained by Pythagoras. So if

$$\underline{v} = (a_1,\ a_2,\ a_3)$$

then the modulus of this vector $|\underline{v}| = \sqrt{\left(a_1^2+,\ a_2^2,\ a_3^2\right)}$ which gives us the distance from the origin.

Now consider $\underline{x} = (x_1,\ x_2, \ldots\ldots\ldots,\ x_n)^{\mathsf{T}}$. Let $\mathbb{R}^n$ denote the set of all $n-$ dimensional vectors (so all vector components are real). To define distance in $\mathbb{R}^n$ we use the notion of a *norm*. A pair of double vertical lines $\|\cdot\|$ is used to denote a norm, i.e. size of an $n-$ dimensional vector.

More formally a *vector norm* on $\mathbb{R}^n$ is a function, $\|\cdot\|$, such that

$$\|\cdot\| : \mathbb{R}^n \longrightarrow \mathbb{R}$$

with the following properties:

(i) $\|\underline{x}\| \geq 0 \quad \forall \underline{x} \in \mathbb{R}^n$

(ii) $\|\underline{x}\| = 0 \quad$ iff $\underline{x} = \underline{0}$

(iii) $\|\alpha \underline{x}\| = |\alpha| \, \|\underline{x}\| \quad \forall \alpha \in \mathbb{R}$ and $\underline{x} \in \mathbb{R}^n$

(iv) $\left\|\underline{x} + \underline{y}\right\| \leq \|\underline{x}\| + \left\|\underline{y}\right\| \quad \forall \underline{x}, \ \underline{y} \in \mathbb{R}^n$ (Triangle inequality)

If $\underline{x} = (x_1, x_2, ..., x_n)^{\mathsf{T}}$, then the vector norm $\|\underline{x}\|_p$ for $p = 1, 2, ...$ is defined as

$$\|\underline{x}\|_p = \left\{ \sum_{i=1}^{n} |x_i|^p \right\}^{1/p}$$

Then there are a number of ways of defining a *norm*, depending on the value of p.

The special case $\|\underline{x}\|_\infty$ is given by

$$\|\underline{x}\|_\infty = \max_{1 \leq i \leq n} |x_i| \qquad\qquad l_\infty \text{ norm.}$$

The most commonly encountered vector norm (also known as the modulus of a vector) is $\|\underline{x}\|_2$ and defined by

$$\|\underline{x}\|_2 = \left\{ \sum_{i=1}^{n} x_i^2 \right\}^{1/2} \qquad l_2 \text{ norm}$$

For obvious reasons this is also called the *Euclidean norm*.

Examples

1. $\underline{x} = \left(\sin k,\ \cos k,\ 2^k\right)^{\mathsf{T}}$; k is a positive integer

$$\|\underline{x}\|_2 = \sqrt{\sin^2 k + \cos^2 k + 4^k} = \sqrt{1 + 4^k}$$

$$\|\underline{x}\|_\infty = \max\left|\sin k,\ \cos k,\ 2^k\right| = 2^k$$

2. $\underline{x} = (2,\ 1,\ 3,\ -4)^{\mathsf{T}}$

$$\|\underline{x}\|_2 = \sqrt{30} \qquad \text{and} \qquad \|\underline{x}\|_\infty = 4$$

These and other types of vector norms are further summarized in the following table, using a particular vector $\underline{v} = (1,\ 2,\ 3)$

Name	Symbol	value	numerical value
$l_1 - \text{norm}$	$\|\underline{x}\|_1$	6	6.000
$l_2 - \text{norm}$	$\|\underline{x}\|_2$	$\sqrt{14}$	3.742
$l_3 - \text{norm}$	$\|\underline{x}\|_3$	$6^{2/3}$	3.302
$l_4 - \text{norm}$	$\|\underline{x}\|_4$	$2^{1/4}\sqrt{7}$	3.146
$l_\infty - \text{norm}$	$\|\underline{x}\|_\infty$	3	3.00

If $\underline{x} = (x_1,\ x_2, \ldots\ldots\ldots,\ x_n)^{\mathsf{T}}$, $\underline{y} = (y_1,\ y_2, \ldots\ldots\ldots,\ y_n)^{\mathsf{T}} \in \mathbb{R}^n$, then the definitions for l_2 and l_∞ can be easily extended to consider distances between $\underline{x}$ and $\underline{y}$

$$\left\|\underline{x} - \underline{y}\right\|_2 = \left\{\sum_{i=1}^{n} (x_i - y_i)^2\right\}^{1/2}$$

$$\left\| \underline{x} - \underline{y} \right\|_\infty = \max_{1 \leq i \leq n} |x_i - y_i|$$

Having introduced the concept of a vector norm, we can now define the most appropriate form for our convergence criteria, by making use of the l_∞ norm. One such relative condition makes use of the infinity norm l_∞ as a test for convergence, where

$$\frac{\left\| \underline{x}^{(k)} - \underline{x}^{(k-1)} \right\|_\infty}{\left\| \underline{x}^{(k-1)} \right\|_\infty} < \varepsilon.$$

ε is the specified tolerance (> 0) and $\|\underline{x}\|_\infty = \max_{1 \leq i \leq n} |x_i|$.

In the numerical example, actual computations yield for $k = 9$ & 10

$k = 9$	$k = 10$
0.9997	1.0001
2.0004	1.9998
-1.0004	-0.9998
1.0006	0.9998

based on $\varepsilon = 10^{-3}$

$$\frac{\left\|\underline{x}^{(10)} - \underline{x}^{(9)}\right\|_\infty}{\left\|\underline{x}^{(9)}\right\|_\infty} = \frac{\left\|(4,\ -6,\ 6,\ -8)\,10^{-4}\right\|_\infty}{1.9998}$$

$$= 4 \times 10^{-4} < \varepsilon$$

Also $\left\| \underline{x}^{(10)} - \underline{x}^{(9)} \right\|_\infty = 2 \times 10^{-4} < \varepsilon$

The iterative technique used is called the *JACOBI* method.

It consists of solving the i^{th} equation in $A\underline{x} = \underline{b}$ for x_i (provided $a_{ii} \neq 0$) to obtain

$$x_i = \sum_{\substack{j=1 \\ j \neq i}}^{n} \left(\frac{-a_{i\ j} x_j}{a_{ii}} \right) + \frac{b_i}{a_{ii}} \qquad i = 1, ..., n \quad ; \quad j \neq i$$

and generating each $x_i^{(k)}$ from components of $\underline{x}^{(k-1)}$ $(k \geq 1)$ by

$$x_i^{(k)} = \frac{1}{a_{ii}} \left[\sum_{\substack{j=1 \\ j \neq i}}^{n} \left(-a_{i\ j}\ x_j^{(k-1)} \right) + b_i \right] \ , \ i = 1, ..., n; \ j \neq i$$

We can now refine this method very simply by using the most up-to-date x_i.

To compute $x_i^{(k)}$, components of $\underline{x}^{(k-1)}$ are used.

Since for $i > 1$, $x_1^{(k)}$, ..., $x_{i-1}^{(k)}$ have already been computed and are likely to be better approximations to the actual solutions x_1, ..., x_{i-1} than $x_1^{(k-1)}$, ..., $x_{i-1}^{(k-1)}$, it is far better to calculate $x_i^{(k)}$ using the most recently calculated values, i.e.

$$x_i^{(k)} = \frac{1}{a_{ii}} \left[-\sum_{j=1}^{i-1} \left(a_{i\ j}\ x_j^{(k)} \right) - \sum_{j=i+1}^{n} \left(a_{i\ j}\ x_j^{(k-1)} \right) + b_i \right] ,$$
$$(i = 1, ..., n)$$

This is called the *GAUSS-SEIDEL* method.

Re-doing the previous example using the G-S method iterative technique:

$$10x_1 - x_2 + 2x_3 = 6$$

$$-x_1 + 11x_2 - x_3 + 3x_4 = 25$$

$$2x_1 - x_2 + 10x_3 - x_4 = -11$$

$$3x_2 - x_3 + 8x_4 = 15$$

which is written as

$$\left.\begin{aligned} x_1^{(k)} &= \tfrac{1}{10}\left\{x_2^{(k-1)} - 2x_3^{(k-1)} + 6\right\} \\ x_2^{(k)} &= \tfrac{1}{11}\left\{x_1^{(k)} + x_3^{(k-1)} - 3x_4^{(k-1)} + 25\right\} \\ x_3^{(k)} &= \tfrac{1}{10}\left\{-2x_1^{(k)} + x_2^{(k)} + x_4^{(k-1)} - 11\right\} \\ x_4^{(k)} &= \tfrac{1}{8}\left\{-3x_2^{(k)} + x_3^{(k)} + 15\right\} \end{aligned}\right\}$$

Again letting $\underline{x}^{(0)} = \underline{0}^{\mathsf{T}}$

Here

$\underline{x}^{(4)} = (1.0009,\ 2.0003,\ -1.0003,\ 0.9999)$

$\underline{x}^{(5)} = (1.0001,\ 2.0000,\ -1.0000,\ 1.0000)$

$$\frac{\left\| \underline{x}^{(5)} - \underline{x}^{(4)} \right\|_\infty}{\left\| \underline{x}^{(4)} \right\|_\infty} = 4 \times 10^{-4} \qquad (< \quad \varepsilon)$$

We note that Jacobi's method required twice as many iterations for the same level of accuracy.

If A is strictly diagonally dominant than for any choice of $\underline{x}^{(0)}$ the sequence of solutions generated by both Gauss-Seidel and Jacobi converge to the unique solution.

Root Finding - Nonlinear Algebraic Equations

A fundamental problem in numerical analysis consists of obtaining the zero of a function. Given a function $y = f(x)$ obtain the root of $f(x) = 0$, i.e. find the value of $x =$ which satisfies $f(\ ) = 0$.

e.g.

$$f(x) = x - \sin x$$

or another example:- zero's of the polynomial

$$f(x) = a_0 + a_1 x + a_2 x_2^2 + ... + a_n x^n$$

Generally the various schemes can be grouped under four sections: -

(i) Methods which do not use derivatives of the function

(ii) Methods which do use $f'(x)$

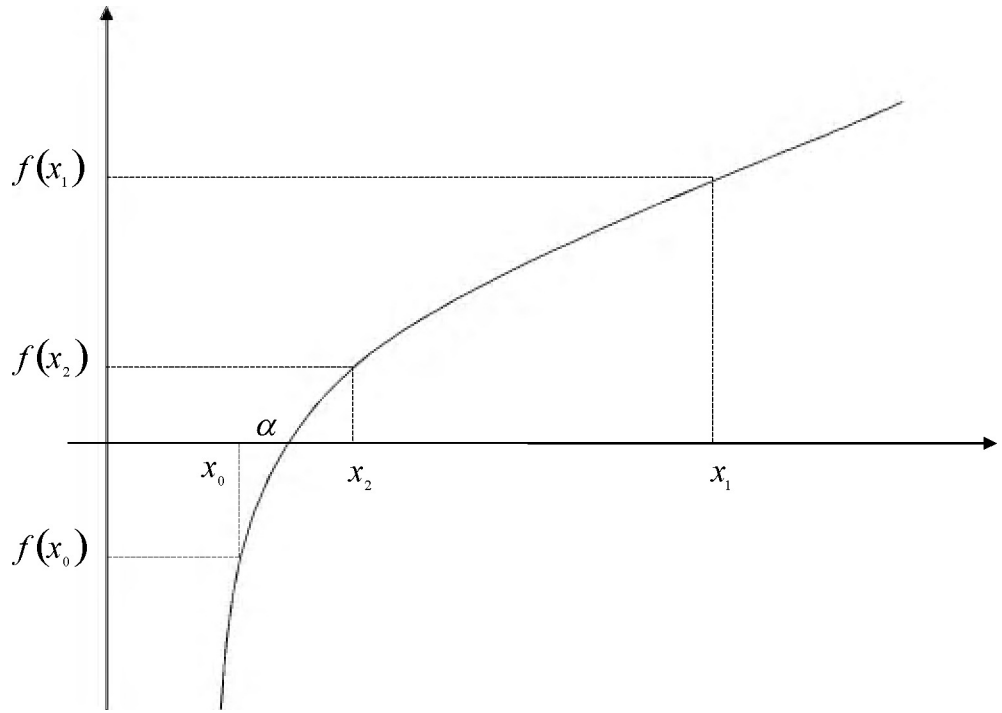
(iii) Methods for polynomials

(iv) Methods which deal with complex roots

Many computer programs (e.g. Mathematica, matlab and excel) have built-in equation solving algorithms, providing very accurate approximations to the solution. Two of the best known root finding algorithms are the bisection method and Newton's method. The latter uses the derivative and provides us with a highly efficient scheme. The one drawback is possible failure when the original estimate is too far from the root. The bisection method does not use any form of calculus and usually takes longer to converge - but is beautifully simple to use.

Methods Which Do Not Use Derivatives - Bisection Method

First we assume that an interval exists which contains the root which is required. We look at this graphically:



In this diagram

$$\begin{aligned} f(x_2) f(x_1) &> 0 \\ f(x_0) f(x_1) &< 0 \\ f(x_0) f(x_2) &< 0 \end{aligned}$$

Hence there is a root in $[x_0, x_1]$ as indicated by the fact that the product

$$f(x_0) f(x_1) \leq 0$$

The method of bisection continues by finding the midpoint of $[x_0, x_1]$

$$x_2 = \frac{(x_0 + x_1)}{2}$$

Then calculate the function value at x_2. If the product

$$f(x_0) f(x_2) \leq 0$$

then the root must lie in the interval $[x_0,\ x_2]$

(This is the case in the diagram given above)

If on the other hand

$$f(x_0) f(x_2) > 0$$

then the root lies in the interval $[x_2,\ x_1]$

Whichever interval is chosen,the new interval containing the root can be further subdivided. After n steps of bisection the interval containing the root will be reduced in size to

$$\frac{(x_1 - x_0)}{2^n}$$

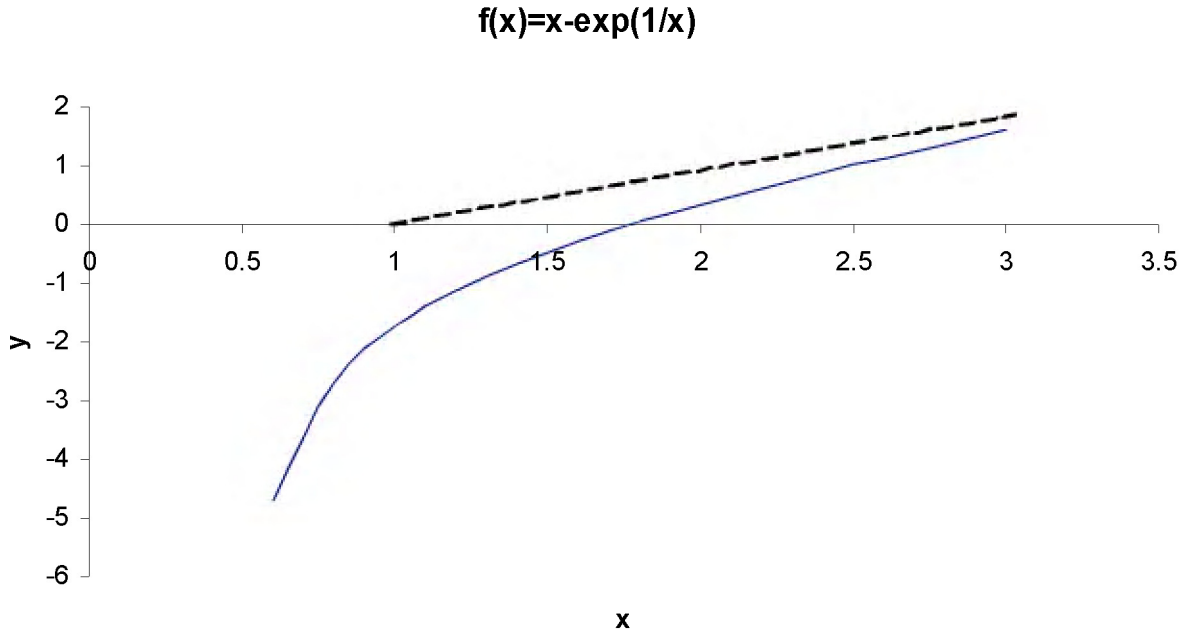
where in the example the value $x_1 = 2$ and $x_0 = 1$.

If the size of the interval becomes smaller than some specified tolerance, t, then the calculation stops.

Example

Consider

$$f(x) = x - e^{1/x}$$



There is a root in [1, 2]

Use the bisection method to show that the root of $f(x) = x - e^{1/x}$ in the interval [1, 2] is 1.763 (correct to 3 decimal places)

$$
\begin{aligned}
f(x_0) &= f(1) = 1 - e^1 < 0 \\
f(x_1) &= f(2) > 0
\end{aligned}
$$

$$
x_2 = \frac{x_0 + x_1}{2} = 1.5
$$

$$
f(1.5) = 1.5 - e^{2/3} = f(x_2) = -0.4477
$$

$$
f(x_0)\, f(x_2) > 0
$$

$\therefore$ root in $[x_2,\ x_1]$ i.e. in $[1.5,\ 2]$

$$
\begin{aligned}
x_3 &= \frac{x_2 + x_1}{2} = \frac{1.5 + 2}{2} = 1.75 \\
f(x_3) &= 1.75 - e^{1/1.75} = -0.0208 \\
f(x_3)\, f(x_2) &> 0
\end{aligned}
$$

$\therefore$ root in $[x_3,\ x_1]$ i.e. in $[1.75,\ 2]$

$$\begin{aligned} x_4 &= \frac{1.75+2}{2} = \frac{3.75}{2} = 1.875 \\ f(x_4) &= f(1.875) = 0.1704 \\ f(x_3)\, f(x_4) &< 0 \end{aligned}$$

$\therefore$ root in $[x_3,\ x_4]$

i.e. in $[1.75,\ 1.875]$

$$\begin{aligned} x_5 &= \frac{x_3+x_4}{2} = \frac{1.75+1.875}{2} = 1.8125 \\ f(x_5) &= f(1.8125) = 0.0763 \\ f(x_3)\, f(x_5) &< 0 \end{aligned}$$

$\therefore$ root in $[x_3,\ x_5]$

i.e. in $[1.75,\ 1.8125]$

$$\begin{aligned} x_6 &= \frac{x_3 + x_5}{2} = \frac{1.75 + 1.8125}{2} = 1.78125 \\ f(x_6) &= f(1.78125) > 0 \\ f(x_3)\, f(x_6) &< 0 \end{aligned}$$

$\therefore$ root in $[x_3,\ x_6]$

i.e. in $[1.75,\ 1.78125]$

$$\begin{aligned} x_7 &= \frac{x_3 + x_6}{2} = 1.765625 \\ f(x_7) &= f(1.765625) > 0 \\ f(x_3)\, f(x_7) &< 0 \end{aligned}$$

$\therefore$ root in $[x_3,\ x_7]$

i.e. in $[1.75,\ 1.765625]$

$$\begin{aligned} x_8 &= \frac{x_3 + x_7}{2} = 1.7578125 \\ f(x_8) &= f(1.7578125) < 0 \\ f(x_3)\, f(x_8) &> 0 \end{aligned}$$

$\therefore$ root in $[x_8,\ x_7]$

i.e. in $[1.7578125,\ 1.765625]$

$$\begin{aligned} x_9 &= \frac{x_8 + x_7}{2} = 1.76178 \\ f(x_9) &= f(1.76178) < 0 \\ f(x_8)\, f(x_9) &> 0 \end{aligned}$$

$\therefore$ root in $[x_8,\ x_7]$

i.e. in $[1.761718,\ 1.765625]$

$$\begin{aligned} x_{10} &= \frac{x_9 + x_7}{2} = 1.7636715 \\ f(x_{10}) &= f(1.7636715) > 0 \\ f(x_9)\, f(x_{10}) &< 0 \end{aligned}$$

$\therefore$ root in $[x_9,\ x_{10}]$

i.e. in $[1.761718,\ 1.7636715]$

$$\begin{aligned} x_{11} &= \frac{x_9 + x_{10}}{2} = 1.76269 \\ &= 1.763 \text{ to 3 decimal places} \end{aligned}$$

Methods Which Do Use The Derivative $f'(x)$

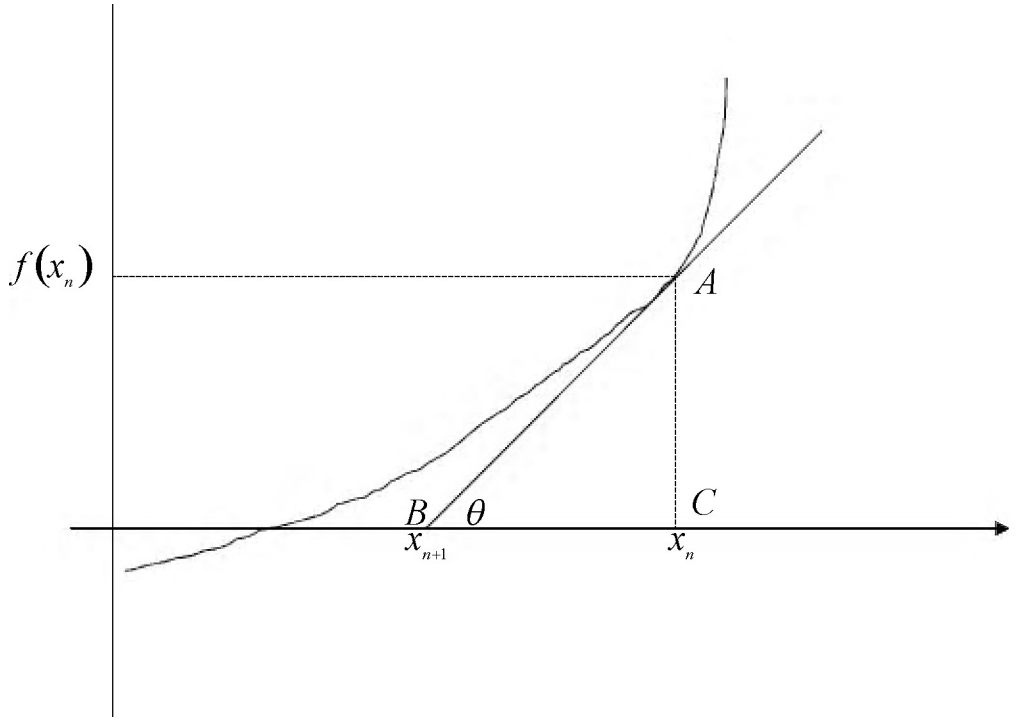
Newton-Raphson Method

For example

$$\begin{aligned} f(x) &= x - e^{1/x} \\ f'(x) &= 1 - e^{1/x}.\left(-\tfrac{1}{x^2}\right) \end{aligned}$$

$$f'(x) = 1 + \frac{e^{1/x}}{x^2}$$

From above we see that $\tan\theta = \dfrac{AC}{BC} = \dfrac{f(x_n)}{(x_n - x_{n+1})}$



But

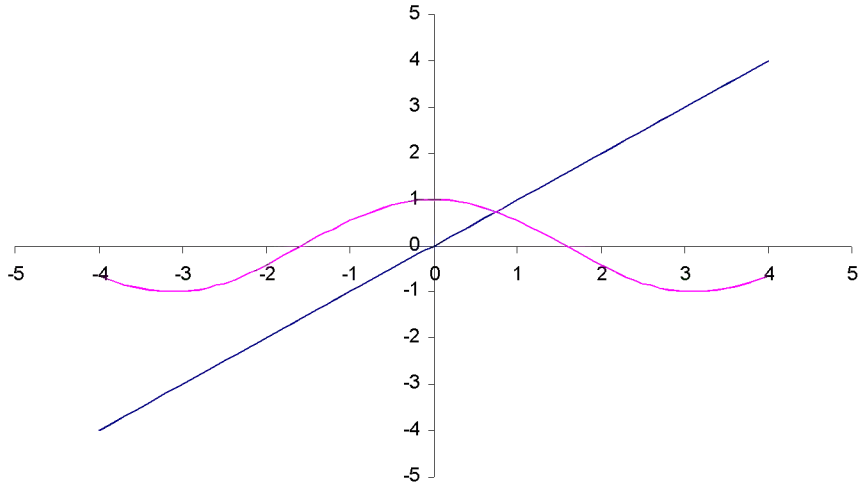
$$\begin{aligned} \tan\theta &= f'(x_n) \\ f'(x_n) &= \frac{f(x_n)}{(x_n - x_{n+1})} \end{aligned}$$

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

This is the *Newton-Raphson Technique*.

Example: Solve for roots, the function $f(x) = x - \cos x$.

Start by considering $x = \cos x$. That is draw $y = x$ and $y = \cos x$ to obtain an initial guess for the root(s).



Clearly the diagram above shows that there is only one root $\alpha \in (0,1)$.

We use the Newton formula

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \quad n = 0, 1, ...$$

where $n = 0$ is the initial guess. $f(x_n) = x_n - \cos x_n \longrightarrow f'(x_n) = 1 + \sin x_n$

x	0	1
$f(x)$	-1	0.75

so numerically we also see that $f(0) f(1) < 0 \Longrightarrow \alpha \in (0,1)$. NR formula for this function becomes

$$x_{n+1} = x_n - \frac{x_n - \cos x_n}{1 + \sin x_n}, \quad x_0 = 1$$

$$
\begin{aligned}
x_1 &= x_0 - \frac{x_0 - \cos x_0}{1 + \sin x_0} = 0.75036 \\
x_2 &= 0.75036 - \frac{0.75036 - \cos 0.75036}{1 + \sin 0.75036} = 0.73911 \\
x_3 &= 0.73911 - \frac{0.73911 - \cos 0.73911}{1 + \sin 0.73911} = 0.73909 \\
x_4 &= 0.73909 - \frac{00.73909 - \cos 0.73909}{1 + \sin 0.73909} = 0.73909
\end{aligned}
$$

which gives the root $\alpha \approx 0.73909$