CS 314 Numerical Methods

Web page:
http://www.cs.purdue.edu/homes/cs314

Goal

- How to use the computer to solve or approximate the solution of non-linear and linear equations, integration differentiation, interpolation, and differential equations.

- Also you will learn the limitations of the computer, such as precision, computational errors etc.

Textbook
"Numerical Methods using Matlab"
Mathews and Fink (Third or fourth ed)

Grade Distribution

25% Midterm
25% Final
50% Projects & homeworks.

Syllabus

- Floating and Fixed Point representation of numbers and Error.
- Solution of non-linear equations

$$f(x) = 0$$

- Solution of linear Equations

$$\mathbb{A}\mathbf{x} = \mathbb{B}$$

- Interpolation
- Curve Fitting
- Numerical Diferentiation
- Numerical Integration
- Solution of Diferential Equations.

- Binary Numbers.

- Computers Use Binary numbers. It uses digits ϕ and 1

537.5_{10} ← base 10 (decimal)

$$\underset{(512)}{1\times 2^9} + 0\times 2^8 + 0\times 2^7 + 0\times 2^6 + 0\times 2^5 + \underset{(16)}{1\times 2^4} + \underset{(8)}{1\times 2^3} + 0\times 2^2$$

$$+ 0\times 2^1 + \underset{(1)}{1\times 2^0} + \underset{(.5)}{1\times 2^{-1}}$$

$512+16+8+1+.5 = \underline{537.5}$

$537.5_{10} = 1000011001.1_2$ ← base 2 (binary)

How to go from base 10 to base 2

Separate the integer part ≥ 1 from the fraction < 1

$$N.F$$

The integer part N has the form

$$N = b_0 + b_1 \times 2^1 + b_2 \times 2^2 + b_3 \times 2^3 \dots b_k 2^k$$

we wish to find digit b_0. We divide by 2

$$\frac{N}{2} = \underbrace{\frac{b_0}{2}}_{\text{reminder}} + \underbrace{b_1 + b_2 2^1 + b_3 2^2 \dots b_k 2^{k-1}}_{n}$$

The reminder of $\frac{N}{2}$ determines if b_0 is ϕ or 1

reminder $== 0 \rightarrow b_0 = 0$ Repeat until

reminder $!= 0 \rightarrow b_0 = 1$ n is ϕ

Example

$N = 73$

$$2\,\overline{)\,73}^{\;36}$$
$$13$$
$$1$$

$$2\,\overline{)\,36}^{\;18}$$
$$16$$
$$0$$

$$2\,\overline{)\,18}^{\;9}$$
$$40$$

$$2\,\overline{)\,9}^{\;4}$$
$$1$$

$$2\,\overline{)\,4}^{\;2} \quad 2\,\overline{)\,2}^{\;1} \quad 2\,\overline{)\,1}^{\;0} \quad \text{stop}$$
$$0 \qquad\quad 0 \qquad\quad 1$$

$$1001001_b$$

$$1 \times 2^6 \qquad 1 \times 2^3 \qquad 1 \times 2^0$$

$$64 + 8 + 1 = 73$$

To compute F in N.F

$$F = b_{-1}2^{-1} + b_{-2}2^{-2} + b_{-3}2^{-3} \cdots$$

$$2F = b_{-1} + b_{-2}2^{-1} + b_{-3}2^{-2} \cdots$$

IF $2F \geq 1$ then $b_{-1} = 1$

otherwise $2F < 1$ then $b_{-1} = 0$

Example

$F = .3$

$2F = 2(.3) = .6 \rightarrow b_{-1} = 0$

$2(.6) = 1.2 \rightarrow b_{-2} = 1$

$2(.2) = .4 \rightarrow b_{-2} = 0$

$2(.4) = .8 \rightarrow b_{-3} = 0$

$2(.8) = 1.6 \rightarrow b_{-4} = 1$

$2(.6) = 1.2$

$2(.2) = .4$

$F_2 = .01001100$
11001100

$14.2_{10} = X_2$

$N . F$

$N_{10} = 14$

$2\overline{)14}$ = 7, remainder 0

$2\overline{)7}$ = 3, remainder 1

$2\overline{)3}$ = 1, remainder 1

$2\overline{)1}$ = 0, remainder 1

$N_2 = 1110$

$F_{10} = .2$

$(.2)2 = .4 < 1$

$(.4)2 = .8 < 1$

$(.8)2 = 1.6 > 1$

$F_2 = .00110011\ldots$

$(.6)2 = 1.2 > 1$

$(.2)2 = .4 < 1$

$(.4)2 = .8 < 1$

$(.8)2 = 1.6 > 1$

$14.2_{10} = 1110.00110011\ldots_2$

Fixed Point vs. Floating Point

Fixed Point

The number of bits for integer part (N) and Fraction (F) is fixed.

NNNN.FF — 4 bits for integer part, 2 bits for fraction

Largest Number

1111.11 → 15.75 is max number

Smallest Number

0000.00 → 0 is smallest number

Advantages:

- Fast Arithmetic. You can use integer arithmetic for operations.
- Can be used is simple electronics like MP3 players.

$$
\begin{array}{r}
0101.01 \\
+\ 1000.10 \\
\hline
1101.11
\end{array}
$$

Disadvantage: Little Range/bad precision

Floating Point

+ The decimal point moves around to keep the desired precision.
+ It uses scientific notation ($A \times 10^x$) but in binary.
+ The mantissa and exponent are stored in different parts.

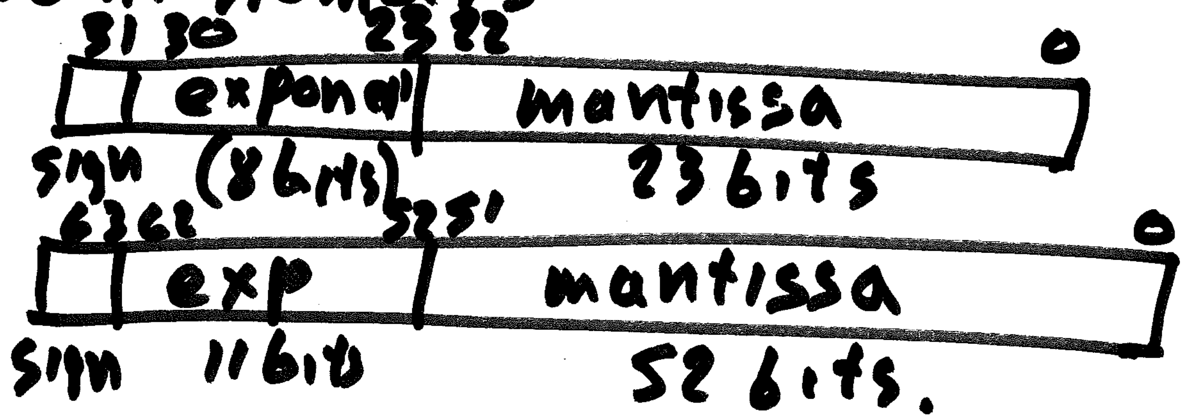
Standard Format of Floating Point Numbers

Float - Single precision
uses 4 bytes (32 bits)
24 bits mantissa and 8 bits exponent

Double - Double Precision
uses 8 bytes (64 bits)
53 bits mantissa, 11 exponent

Standard IEEE 754
for floating point numbers

float → 32 bits
double → 64 bits



exponent stored in memory does not contain sign. It is "biassed" using an added constant.

$1.M \times 2^{x}$

$exp_{(mem)} = x + bias$

bias = 127 float
= 1023 double

$7.8_{10} \rightarrow$ binary representation
using IEEE 754 double (64 bits)

$\Downarrow$ binary

111.1100110011001100 . . . . .

$$2\overline{)7}^{\,3} \;\; 1 \qquad 2\overline{)3}^{\,1} \;\; 1 \qquad 2\overline{)1}^{\,0} \;\; 1$$

$.8 \times 2 = 1.6$
$.6 \times 2 = 1.2$
$.2 \times 2 = .4$
$.4 \times 2 = .8$
$.8 \times 2 = 1.6$
$.6 \times 2 = 1.2$

Make it look like $1.M \times 2^{x}$

$$1.\underbrace{1111001100110011001100\ldots}_{\text{mantissa}} \times 2^{2}$$

$2^{10} + 2^{0}$

$exp = x + bias$
$= 2 + 1023 = 1025$

63	(11 bits)	52 bits
0	10000000001	1111001100110011001100 . . . . .
sign	exp	mantissa

∅ remove 1.

Computational Errors.

Errors arise because the result of an operation does not fit in the memory assigned to the resut.

Two types of error

Absolute Error $= |p - p^*|$

Relative Error $= \dfrac{|p - p^*|}{|p|}$ for $p \neq 0$.

p^* is the approximation of p

Propagation of Errors

Assume p and q are approximated by $*p$ and $*q$.

$$p^* = p + e_p$$
$$q^* = q + e_q$$

Sum

$$p^* + q^* = p + q + \underbrace{(e_p + e_q)}_{\text{error}}$$

Product

$$p^* q^* = (p + e_p)(q + e_q)$$

$$= pq + \underbrace{p e_q + q e_p + e_p e_q}_{\text{error}}$$

- In the sum the new error is the sum of the original errors.
- In the product the [redacted] errors are amplified by the magnitude of p and q.

An algorithm is

stable — If small initial errors stay small

unstable — If small initial errors get large.

Solution of Nonlinear Equations

$$f(x) = 0$$

- We will use "iteration" to solve the equations

starting with a value p_0 and a function $p_k = g(p_{k-1})$ we find

$$
\begin{aligned}
&p_0 \\
&p_1 = g(p_0) \\
&p_2 = g(p_1) \\
&\quad\vdots \\
&p_k = g(p_{k-1})
\end{aligned}
$$

we stop when $|p_k - p_{k-1}| < \epsilon$ for some ϵ

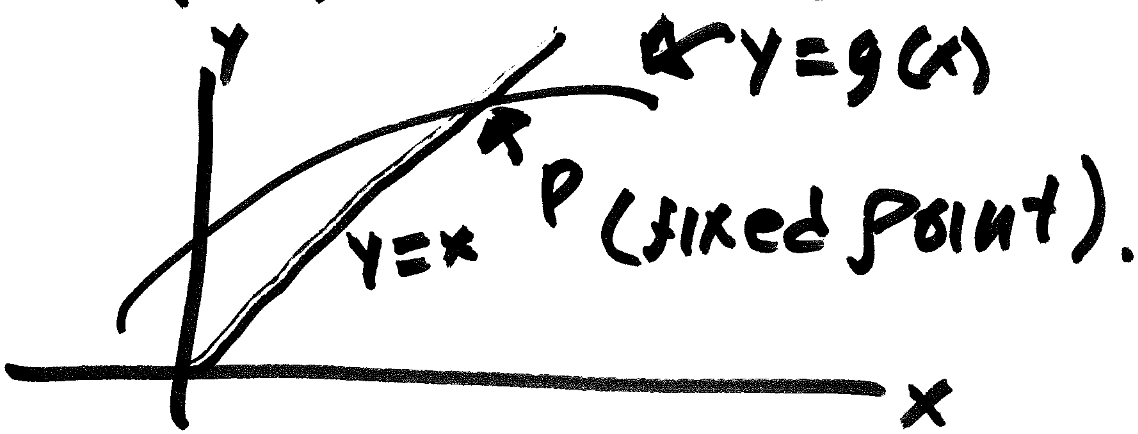
Definitions

- Fixed Point

  A fixed point of $g(x)$ is a number p such that

$$p = g(p)$$

- Geometrically the fixed point of a function $y = g(x)$ are the point of intersection between the line $y = g(x)$ and $y = x$



The iteration

$$p_{n+1} = g(p_n) \quad \text{for } n = 0, 1, 2 \ldots$$

is called a <u>fixed point iteration</u>

<u>Example</u>

$$\cancel{} \quad x^2 - 2x + 1 = 0$$

Lets obtain a fixed point iteration funtion for $x^2-2x+1=0$

$$x^2 - 2x + 1 = 0$$

$$x^2 + 1 = 2x$$

$$x = \frac{x^2+1}{2} \Rightarrow x_{k+1} = g(x_k)$$

$$x_{k+1} = \frac{x_k^2+1}{2}$$

Solution:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

$$= \frac{2 \pm \cancel{\sqrt{4-4}}}{2}$$

$$x = 1$$

Let $x_0 = 0$

$$x_1 = \frac{0^2+1}{2} = .5$$

$$x_2 = \frac{.5^2+1}{2} = .625$$

$$x_3 = \frac{.625^2+1}{2} = .695$$

$$x_4 = \frac{.695^2+1}{2} = .74$$

$x_5 = .77$ $x_6 = .79$ $x_7 = .82 \ldots x = 1$

Fixed Point Theorem

If $|g'(x)| \leq k < 1$ for all $x \in [a,b]$

and $g(x) \in [a,b]$

then the iteration $p_n = g(p_{n-1})$ will converge to a unique fixed point $p \in [a,b]$

In this case p is said to be an attractive fixed point.

E.g. $x_{k+1} = \dfrac{x_k^2+1}{2} = g(x_k)$

$$g(x) = \frac{x^2+1}{2} \Rightarrow g'(x) = \frac{2x}{2} = x$$

So if we start at $x_0 = 0$ $g'(x_0) = 0$

then it will converge

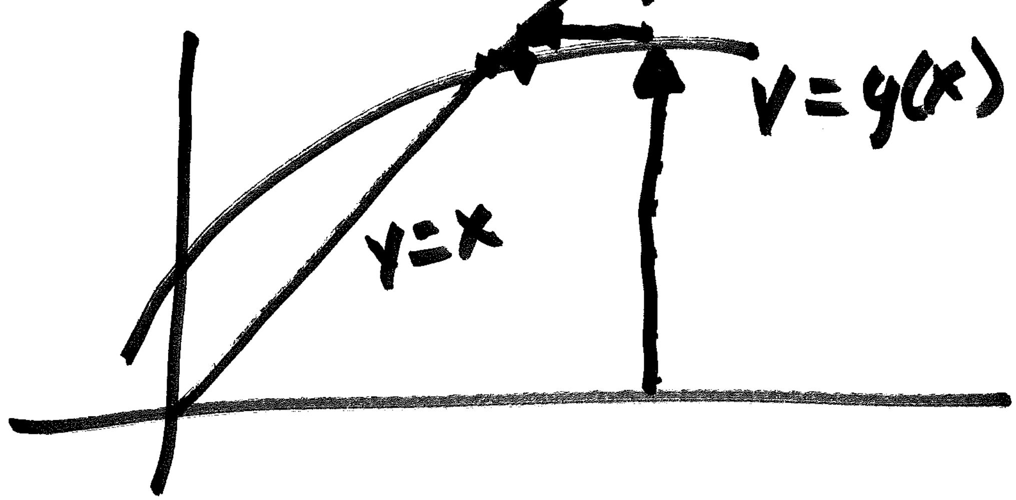
If $|g'(x)| > 1$ then the
Iteration $p_n = g(p_{n-1})$ will not converge.

Now lets start at $x_0 = 2$

$$x_0 = 2$$
$$x_1 = \frac{2^2+1}{2} = 2.5$$
$$x_2 = \frac{2.5^2+1}{2} = 3.625$$
$$x_3 = \frac{3.625^2+1}{2} = 7.07$$
$$\vdots$$
$$x_4 \doteq 25.495$$

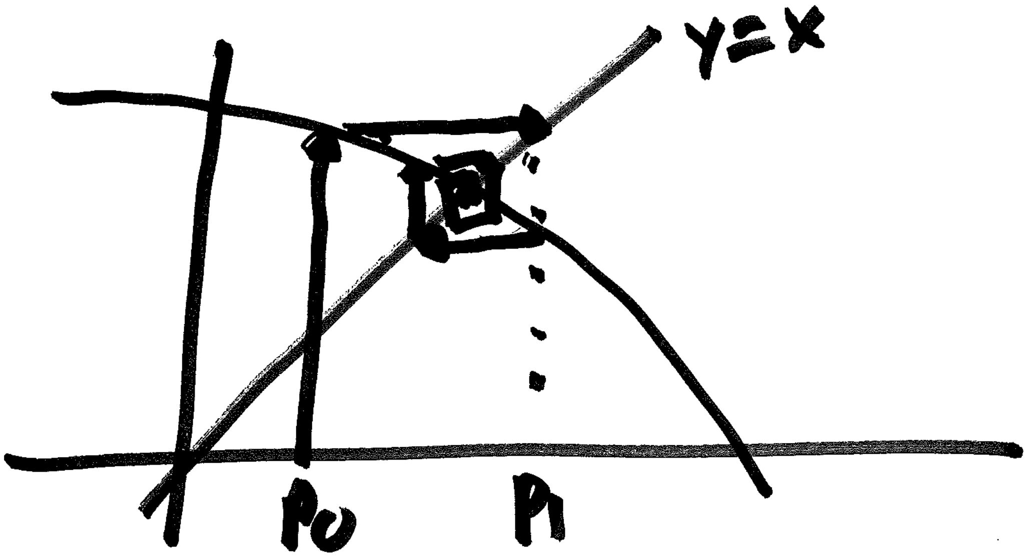
It does not converge.

Monotone Convergence



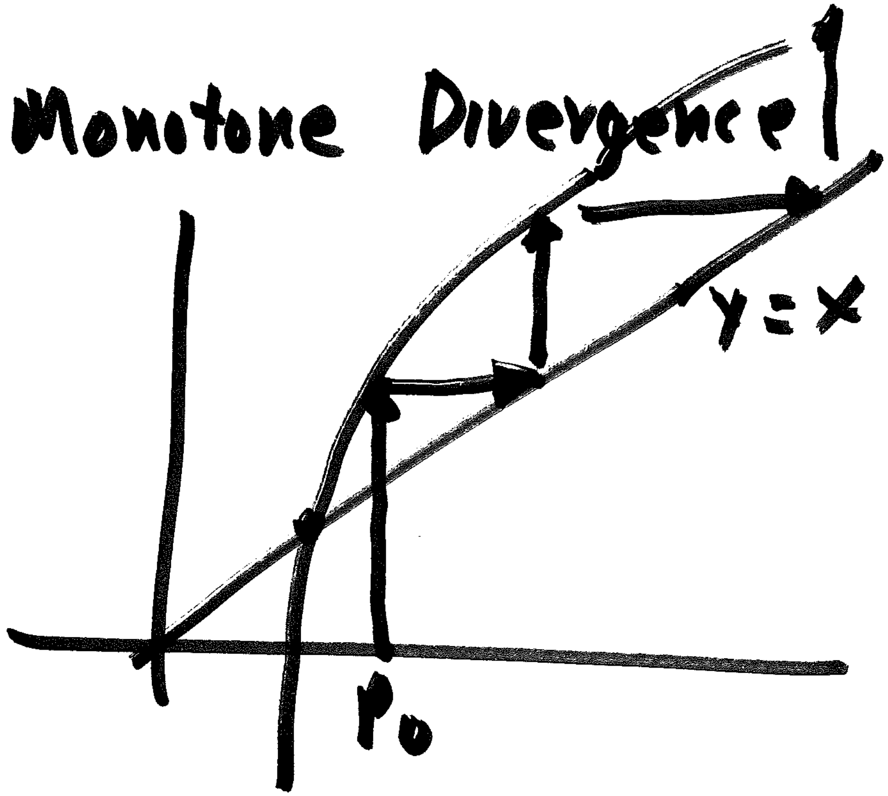
$0 < g'(p) < 1$

Oscilating Convergence



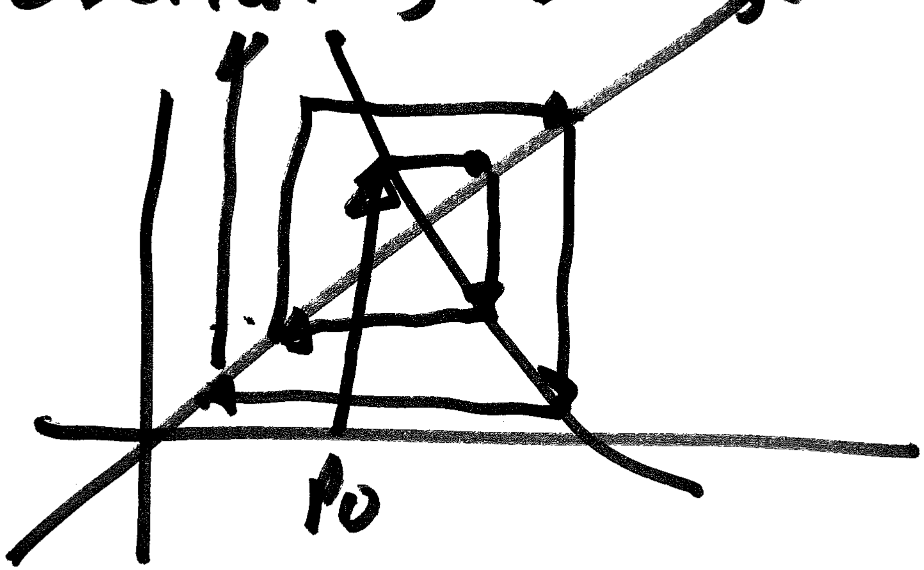
$-1 < g'(p) < 0$

Monotone Divergence



$1 < g'(p)$

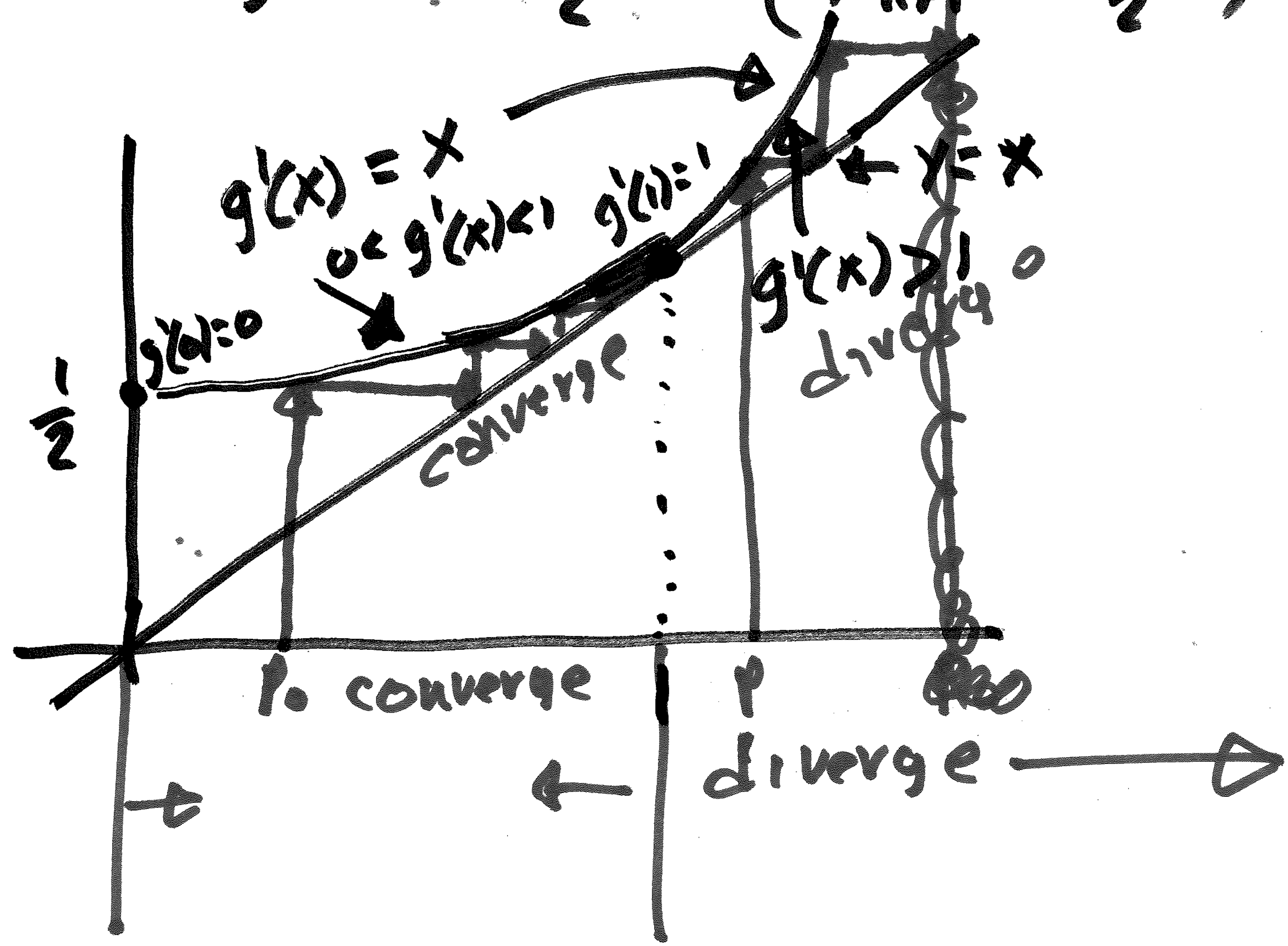
Oscilating Divergence



$g'(p) < -1$

Going back to the example

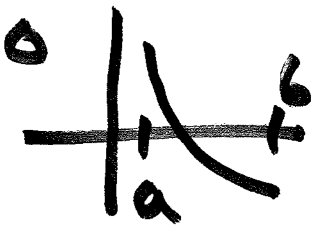
$$g(x) = \frac{x^2+1}{2} \quad \left(x_{k+1} = \frac{x_k^2+1}{2} \right)$$



Bisection Method

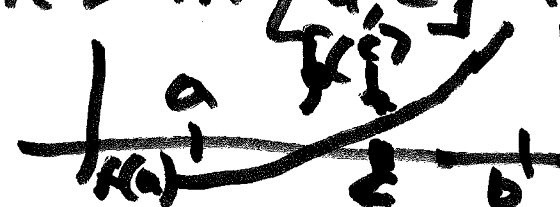
- Find the solution of $f(x) = 0$
- Similar to binary search.
- Starting with two values a, b such that $f(a) < 0$ and $f(b) > 0$

or $f(a) > 0$ and $f(b) < 0$



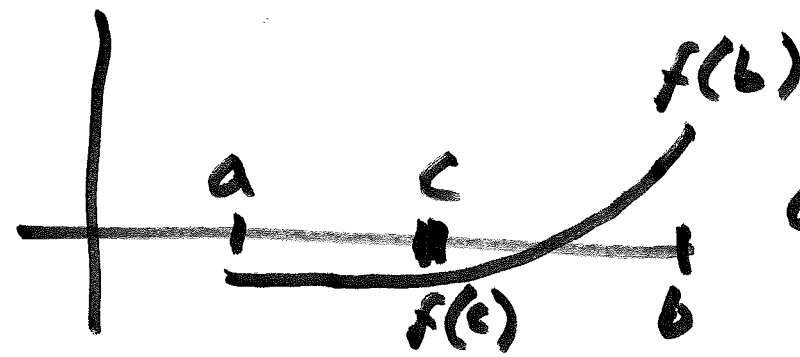
Since $f(x)$ is contiguous, theres is an x between a, b such that $f(x) = 0$

1. Get $c = \frac{a+b}{2}$ (middle point
   a) If $f(a)$ and $f(c)$ have opposite signs, then a zero lies in $[a, c]$ then assign $b \leftarrow c$



$b \leftarrow c$

b) If $f(c)$ and $f(b)$ have opposite signs, a zero lies in $[c, b]$ assign $a \leftarrow c$



$a \leftarrow c$

c) If $f(c) = 0$ then the zero is c. In practice this does not happen. We stope when

$$|a-b| < \varepsilon \text{ or}$$

$$|f(c)| < \varepsilon$$

Example:

$\sin(x) = \frac{1}{2}$ x is in degrees

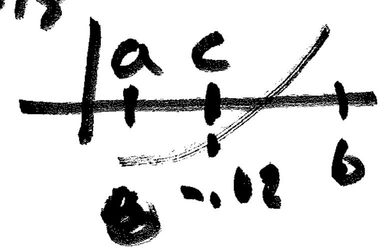
$f(x) = \sin(x) - \frac{1}{2} = 0$

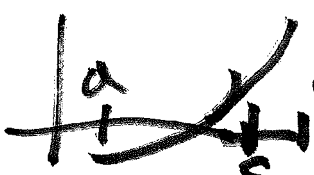
Exact Solution $x = 30^\circ$

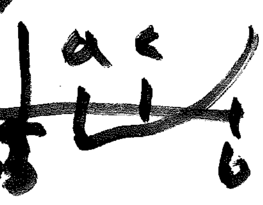
start with $a = 0$, $b = 50$

$f(0) = \sin(0) - .5 = -.5$ $f(50) = \sin(50) - .5 = .266$

a	b	c	f(c)
0	50	$\frac{0+50}{2} = 25$	$\sin(25) - .5 = -.077$
25	50	$\frac{25+50}{2} = 37.5$	$\sin(37.5) - .5 = .1087$
25	37.5	$\frac{25+37.5}{2} = 31.25$	$\sin(31.25) - .5 = .018$
25	31.25	$\frac{25+31.25}{2} = 28.125$	$\sin(28.125) - .5 = -.02$
28.125	31.25	$\frac{28.125+31.25}{2} = 29.6875$	$\sin(29.6875) - .5 = -.00473$
29.6875	31.25	$\frac{29.6875+31.25}{2} = 30.46875$	$\sin(30.46875) - .5 = 0.007069$
29.6875	30.46875		







How many steps do we need to arrive to a solution?

At each step, the range $|a-b|$ is reduced by half

Therefore, at step n

$$|a_n - b_n| = \frac{|a-b|}{2^n}$$

If we want an approximation error $< \varepsilon$ how many steps do we need?

$$|a_n - b_n| = \frac{|a-b|}{2^n} < \varepsilon$$

$$\frac{|a-b|}{\varepsilon} < 2^n$$

$$\cancel{\log\left(\frac{|a-b|}{\varepsilon}\right)}$$

$$\log\left[\frac{|a-b|}{\varepsilon}\right] < \log 2^n$$
$$= n \log 2$$

$$\frac{\log\left[\frac{|a-b|}{\varepsilon}\right]}{\log 2} < n$$

Example: if we need an approximation

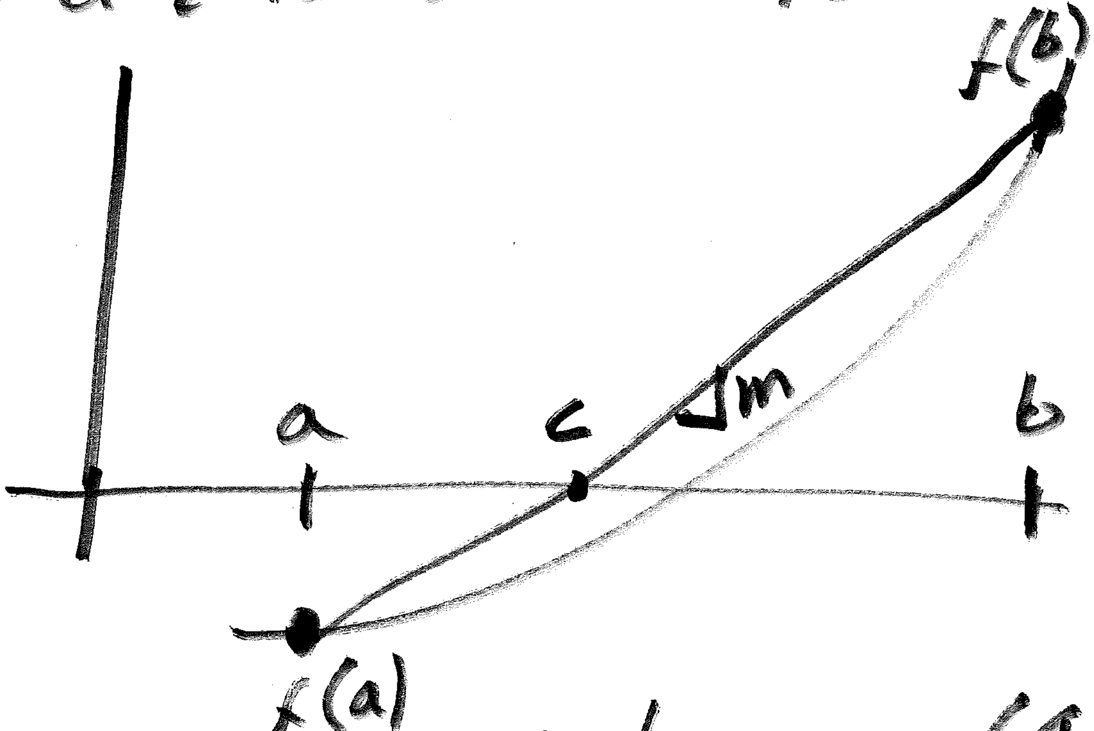
$\varepsilon = .01$ $\qquad |a_n - b_n| < \underset{\varepsilon}{.01}$

$$\frac{\log \frac{|\overset{a}{0} - \overset{b}{5}0|}{\underset{\varepsilon}{.01}}}{\log 2} < n$$

$$12.28 < n$$

False Position Method (Regula Falsi)

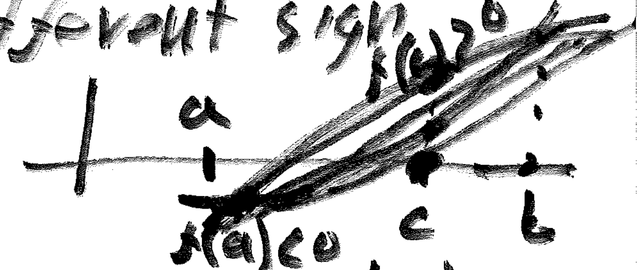
- It was created because ~~bo~~bisection was too slow
- It also needs [a,b] where there is a zero between a,b



- A line is subtended between (a, f(a)) and (b, f(b))
- c is chosen to be the intersection of the line with the x axis.

As is bisection

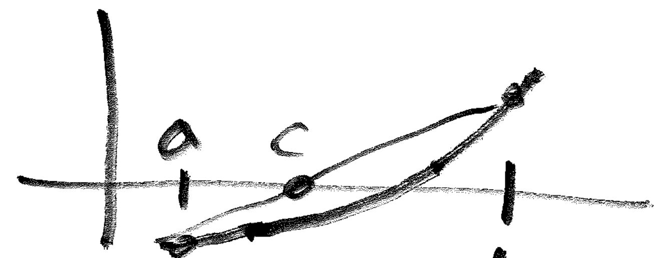
If $f(a)$ and $f(c)$ have different sign
then $b \leftarrow c$



$b \leftarrow c$

&

If $f(b)$ and $f(c)$ have different signs
then $a \leftarrow c$



$a \leftarrow c$

Compute value of c

① $$m = \frac{f(a) - f(b)}{a - b}$$

also

② $$m = \frac{f(b) - 0}{b - c}$$

Putting together ① and ②

$$\frac{f(a)-f(b)}{a-b} = \frac{f(b)-0}{b-c}$$

$$b-c = f(b)\,\frac{a-b}{f(a)-f(b)}$$

$$\boxed{c = b - f(b)\,\frac{a-b}{f(a)-f(b)}}$$

False Position Method (regula falsi)

Example:

$\sin(x) = \frac{1}{2}$

$f(x) = \sin(x) - \frac{1}{2} = 0$ x in degrees

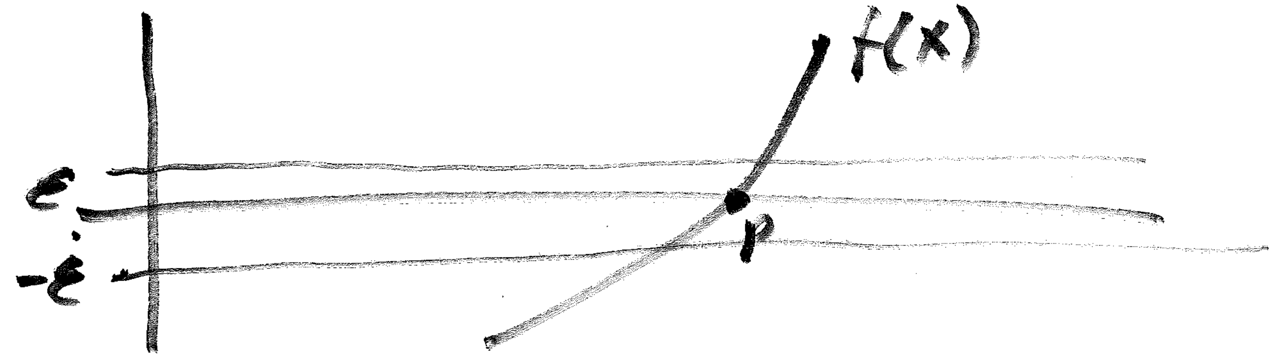
Start with $a=0$, $b=50$

i	a	b	f(a)	f(b)	c
0	0	50	$\sin(0) - .5 = -.5$	$\sin(50) - .5 = .266$	$50 - \frac{.266(0-50)}{-.5-.266} = 32.546$; $f(c) = \sin(32.546) - .5 = .03$
1	0	32.546	$-.5$	$.03$	$c = 32.546 - \frac{.03(0-32.546)}{-.5-.03}$; $c = 30.70$; $f(c) = \sin(30.70) - .5 = .01$
2	0	30.70	$-.5$	$.01$	

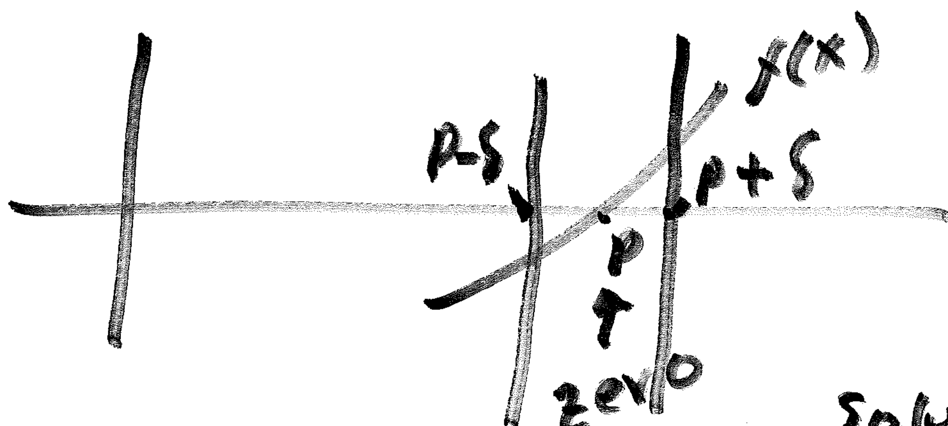
Checking for Convergence

1. Horizontal Convergence

Stop when $|f(x)| < \epsilon$



2. Vertical Convergence



Stop when

$$|x - p| < \delta$$

zero

$f(p) = 0.$

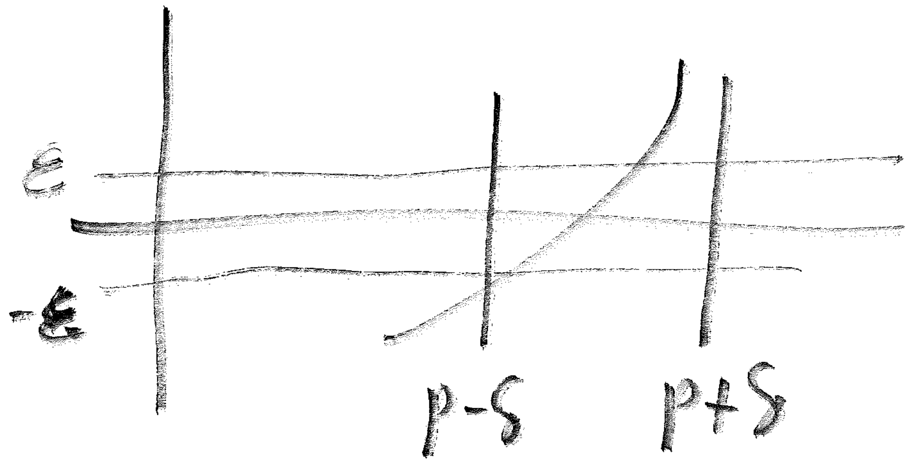
Problem: we don't know p

Solution: we approximate this criteria by stopping when

$$|p_n - p_{n-1}| < \delta$$

3. Both Vertical and Horizontal Convergency

30



Stop when both

$|p_n - p_{n-1}| < \delta$

and

$|f(x)| < \varepsilon$

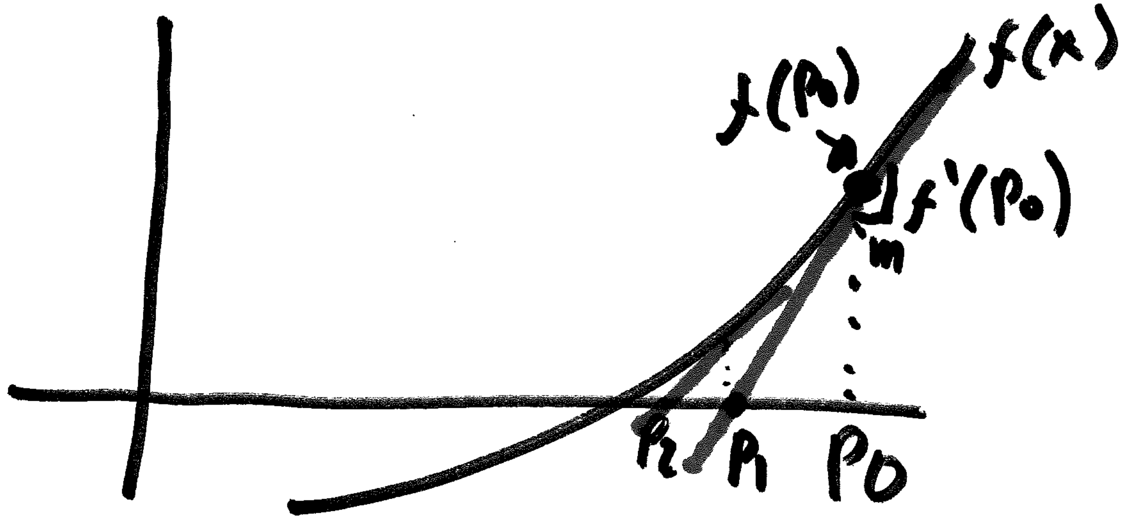
Troublesome Functions

If the graph $y=f(x)$ is close to vertical (90°) near the root $(p,0)$ then the root finding is well conditioned. i.e. The solution can be obtained with good precision.

If the graph $y=f(x)$ is shallow (close to horizontal) near $(p,0)$ then the root finding is "ill conditioned" i.e. the root finding may only have a few significant digits.

Newthon-Rapson

If $f(x)$, $f'(x)$ and $f''(x)$ are continuous then we can use this information to find the solution of $f(x)$.



① $m = f'(P_0)$

② $m = \dfrac{f(P_0) - 0}{P_0 - P_1}$

Putting ① and ② together we obtain

$$f'(P_0) = \frac{f(P_0) - 0}{P_0 - P_1}$$

$$P_0 - P_1 = \frac{f(P_0)}{f'(P_0)} \rightarrow \boxed{P_1 = P_0 - \frac{f(P_0)}{f'(P_0)}}$$

Newthon Rapson

Example

$\sin(x) = \frac{1}{2}$

X is in radians.
Otherwise, if X is in degrees $f'(x)$ is more complicated

$f(x) = \sin(x) - \frac{1}{2} = 0$

$f'(x) = \cos(x)$

start with $p_0 = 0$

i	p_k	$f(p)$	$f'(p)$	p_{k+1}
0	$p_0 = 0$	$\sin(0) - \frac{1}{2} = -.5$	$\cos(0) = 1$	$p_1 = 0 - \frac{(-.5)}{1} = .5$
1	$p_1 = .5$	$\sin(.5) - .5 = -.02$	$\cos(.5) = .8775$	$p_2 = .5 - \frac{-.02}{.8775} = .522$
2	$p_2 = .522$	$\sin(.522) - .5 = -.00139$	$\cos(.522) = .8668$	$p_3 = .522 - \frac{-.00139}{.8668}$

$p_3 = .5235$

Exact Solution

$\sin(x) - \frac{1}{2} = 0$ $\quad \sin(x) = \frac{1}{2}$

$x = 30^\circ = 30 \frac{\pi}{180} = .5235$

Relation between Newthon Rapson and Taylor Expansions

~~Ex~~

We can approximate a continuous function with a taylor polynomial expansion around p_0

$$f(x) = \underbrace{f(p_0) + f'(p_0)(x-p_0)}_{\text{two terms}} + \frac{f''(p_0)(x-p_0)^2}{2!} \dots$$

Newthon-Rapson method approximates $f(x)$ by using the taylor expansion up to the first two terms. It makes $f(x)=0$ and $x = p_1$

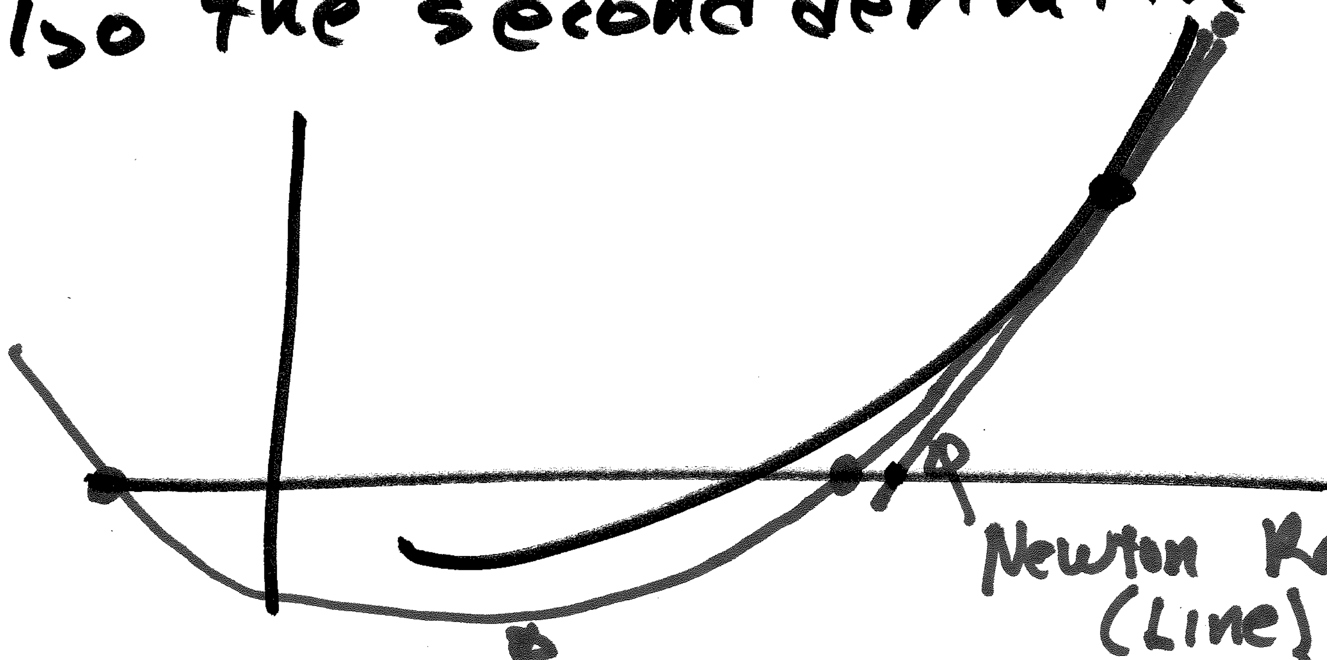
$$f(x) = 0$$

$$0 = f(p_0) - f'(p_0)(p_1 - p_0)$$

$$\frac{-f(p_0)}{f'(p_0)} = p_1 - p_0 \Rightarrow \boxed{p_1 = p_0 - \frac{f(p_0)}{f'(p_0)}}$$

Newton Raphson

In fact, we could use three terms of taylor expansion to obtain an improved Newthon-Rapson that uses also the second derivative



- Newton-Raphson is a very fast method but it needs the derivative of the equation.
- It is possible to obtain the derivative of an equation numerically

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x}$$

If we assume that ϵ is a small number, we can approximate a derivative in the following way

$$f'(x) = \frac{f(x+\epsilon) - f(x)}{\epsilon} \quad \text{for a small } \epsilon$$

Example:

$$f(x) = \sin(x) \implies f'(x) = \cos(x)$$

$$f'(x) \approx \frac{\sin(x+\epsilon) - \sin(x)}{\epsilon}$$

Exact Solution $f'(.5) = .8775$

If $\epsilon = .001$ $\quad x = 0$

$$f'(0) \approx \frac{\sin(0+.001) - \sin(0)}{.001} = 1$$

$$f'(.5) \approx \frac{\sin(.5+.001) - \sin(.5)}{.001} = .8773$$

Order of Convergence

Assume p is a zero of the function and $E = p - p_n$ is the approximation error

$$\lim_{n\to\infty} \frac{|p - p_{n+1}|}{|p - p_n|^R} = \lim_{n\to\infty} \frac{E_{n+1}}{|E_n|^R} = A$$

constant

Then the sequence is said to converge to p with order of convergence R

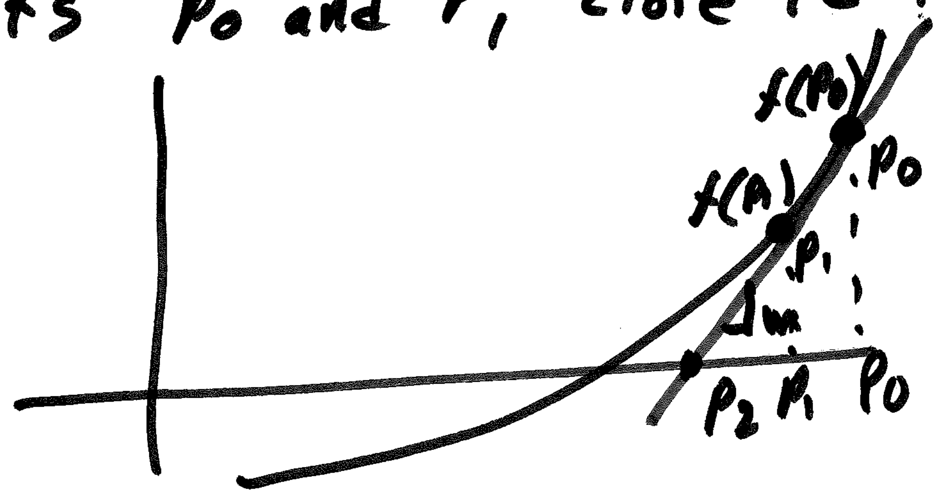
$$E_{n+1} = A|E_n|^R$$

If $R=1$ the convergence is called linear
If $R=2$ the convergence is called quadratic.
The larger R, the faster the convergence ($E_n < 1$).

If p is a simple root (one solution) then Newton-Raphson has $R=2$ (quadratic convergence)

Secant Method

- The Newton-Raphson method requires the evaluation of two functions at each iteration $f(p_k)$ and $f'(p_k)$
- The secant method will require only one evaluation of $f(x)$.
- If $f(x)$ has a simple root (one solution) then it has an order of convergence $R = 1.618$.
- The secant method uses two initial points p_0 and p_1 close to the root



P_2 will be the intersection with the x axis of the line between P_0, P_1

① $m = \dfrac{f(P_1) - f(P_0)}{P_1 - P_0}$ ② $m = \dfrac{0 - f(P_1)}{P_2 - P_1}$

From ① and ②

$$\frac{f(P_1) - f(P_0)}{P_1 - P_0} = \frac{0 - f(P_1)}{P_2 - P_1}$$

$$P_2 - P_1 = \frac{-f(P_1)}{f(P_1) - f(P_0)} (P_1 - P_0)$$

$$\boxed{P_2 = P_1 - \frac{f(P_1)}{f(P_1) - f(P_0)} (P_1 - P_0}$$

Secant Method

Example

$p_0 = 0 \quad p_1 = 1$

$f(x) = \sin(x) - .5 = 0$

$$p_{k+1} = p_k - \frac{f(p_k)(p_k - p_{k-1})}{f(p_k) - f(p_{k-1})}$$

i	p_k	$f(p_k)$	
0	$p_0 = 0$	$\sin(0) - .5 = -.5$	
1	$p_1 = 1$	$\sin(1) - .5 = .3415$	$p_2 = 1 - \frac{.3415(1-0)}{.3415 - (-.5)} = .594$
2	$.594$	$\sin(.594) - .5 = .0597$	$p_3 = .594 - \frac{.0597(.594 - 1)}{.0597 - .3415} = .50798$
3	$.50798$	$\sin(.50798) - .5 = -.0135$	$p_4 = .50798 - \frac{(-.0135)(.50798 - .594)}{-.0135 - .0597}$

$p_4 = \underline{.5238}$

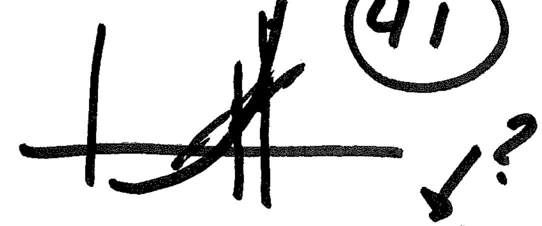
Exact Solution

$\sin(x) = .5$

$x = 30^\circ = 30 \frac{\pi}{180} = \underline{.5235}$

The solution of Linear Systems

$$\mathbb{A}\mathbb{X} = \mathbb{B} \to \begin{aligned} 6x+3y+2z&=29 \\ 3x+y+2z&=17 \\ 2x+3y+2z&=21 \end{aligned} \Rightarrow \underset{\mathbb{A}}{\begin{bmatrix} 6 & 3 & 2 \\ 3 & 1 & 2 \\ 2 & 3 & 2 \end{bmatrix}} \underset{\mathbb{X}}{\begin{bmatrix} x \\ y \\ z \end{bmatrix}} = \underset{\mathbb{B}}{\begin{bmatrix} 29 \\ 17 \\ 21 \end{bmatrix}}$$



Definitions

N Dimensional Vector

$$\mathbb{X} = (x_1, x_2, x_3 \ldots\ldots)$$

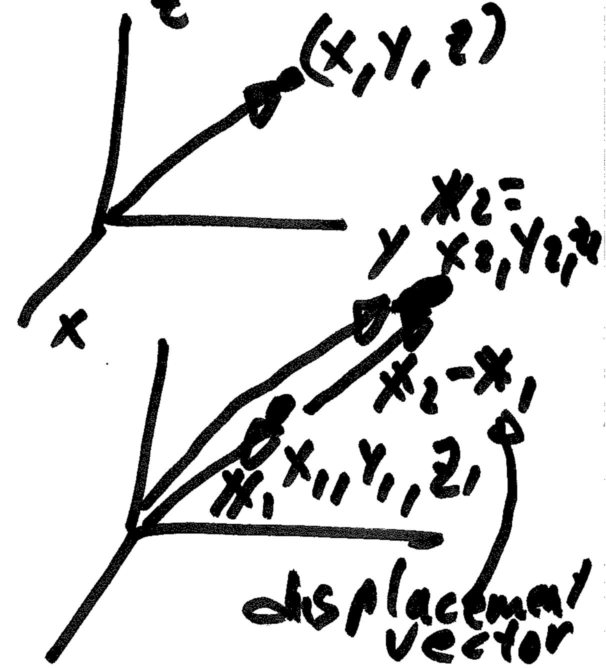
$x_1, x_2 \ldots x_N$ are called componets of $\mathbb{X}$

N-dimensional space

The set of all N dimensional vectors.

When a vector is used to determine a position it is called position vector

When a vector is used to denote movement between two points, it is called a displacement vector



Vector Properties

Equality

$\mathbf{X} = \mathbf{Y}$ if and only if $X_J = Y_J$ for $J = 1, 2 \ldots N$

Sum

$\mathbf{X} + \mathbf{Y} = (X_1 + Y_1, X_2 + Y_2, X_3 + Y_3 \ldots)$

Negative

$-\mathbf{X} = (-X_1, -X_2, -X_3 \ldots)$

Difference

$\mathbf{Y} - \mathbf{X} = \mathbf{Y} + (-\mathbf{X})$

Scalar Multiplication

$C\mathbf{X} = (CX_1, CX_2, CX_3 \ldots)$

Linear Combination

$C\mathbf{X} + d\mathbf{Y} = (CX_1 + dY_1, CX_2 + dY_2, \ldots)$

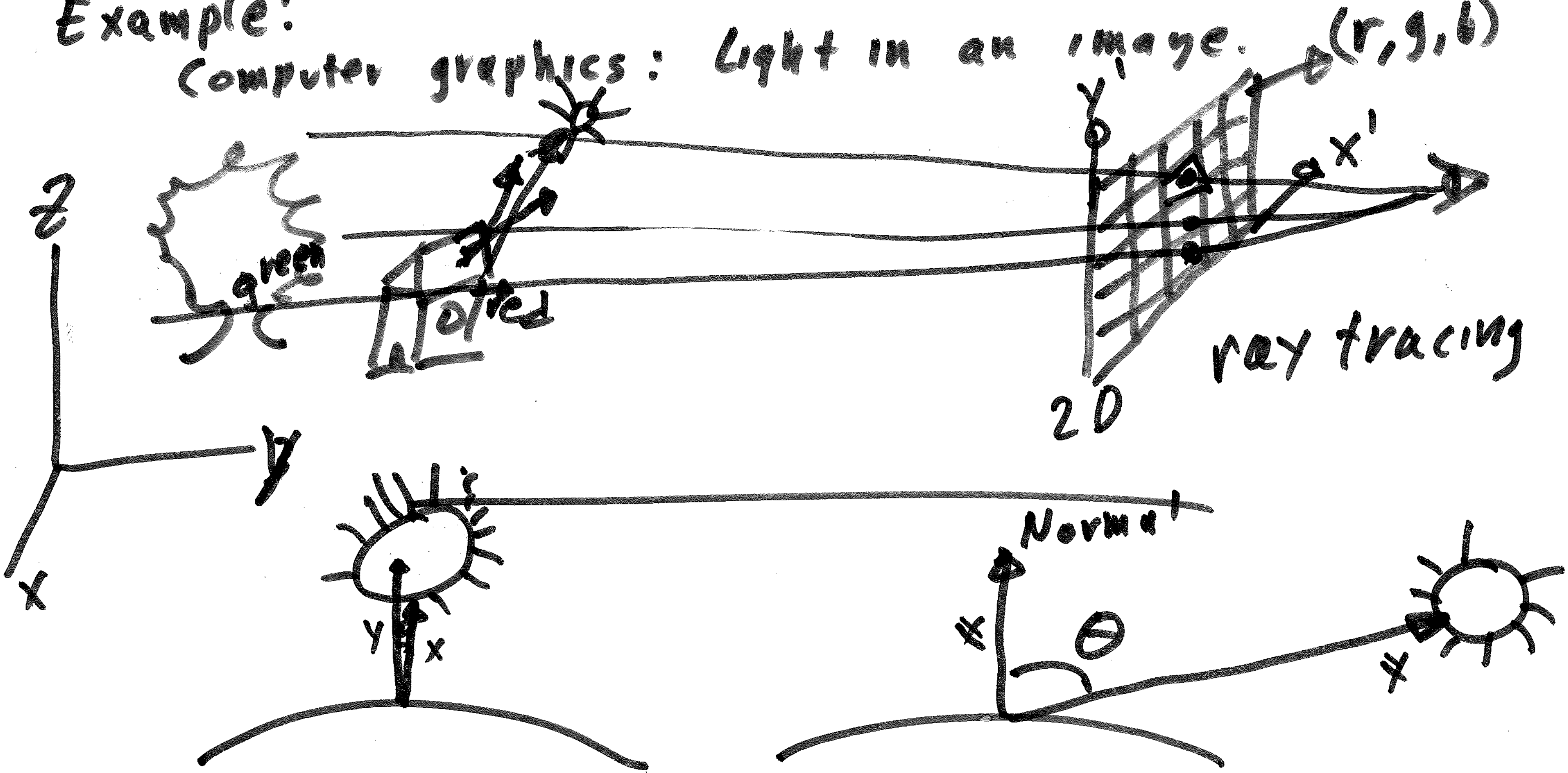
Dot Product

$$\mathbf{x}\cdot\mathbf{y} = x_1 y_1 + x_2 y_2 + x_3 y_3 \ldots$$

$$= |\mathbf{x}||\mathbf{y}|\cos(\theta_{\mathbf{x}\mathbf{y}}) = |\mathbf{x}||\mathbf{y}|\cos(\theta_{\mathbf{x}\mathbf{y}})$$

Example:

Computer graphics: Light in an image.

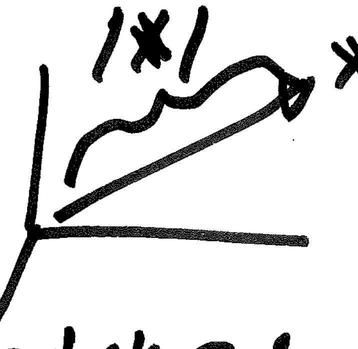


~~Ligh~~

$$\text{Ligh} = C\cos\theta$$
$$= C\,\mathbf{x}\cdot\mathbf{y}$$

The norm or Euclidean Norm of a Vector

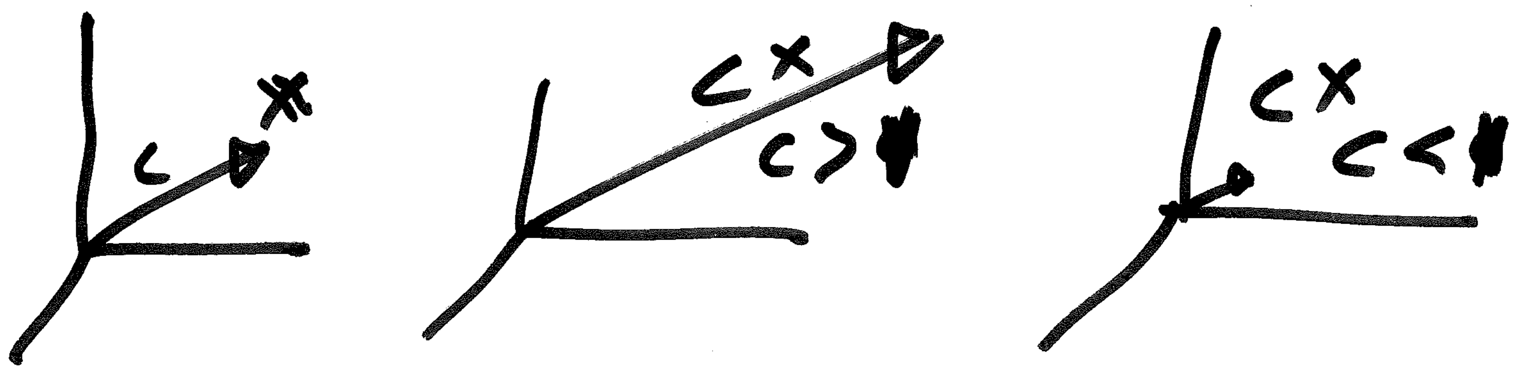
$$|X| = (x_1^2 + x_2^2 + \dots x_N^2)^{1/2}$$



Scalar multiplication cX stretches a vector when $|c| > 1$ and shrinks the vector when $|c| < 0$

$$\|cX\| = ((c^2 x_1)^2 + (c x_2)^2 + (c x_3)^2 \dots)^{1/2}$$

$$= |c| (x_1^2 + x_2^2 + x_3^2 + \dots)^{1/2}$$



Let $\mathbf{x} = (-1, 3, 2)$ $\quad \mathbf{y} = (3, 5, 2)$

Sum $\quad \mathbf{x} + \mathbf{y} = (-1+3,\ 3+5,\ 2+2) = (2, 8, 4)$

Difference $\quad \mathbf{x} - \mathbf{y} = (-1-3,\ 3-5,\ 2-2) = (-4, -2, 0)$

Scalar $\quad 2\mathbf{x} = (2\cdot(-1),\ 2\cdot 3,\ 2\cdot 2) = (-2, 6, 4)$

Length $\quad \|\mathbf{x}\| = \left((-1)^2 + (3)^2 + (2)^2\right)^{1/2} = (1+9+4)^{1/2} = \sqrt{14}$

Dot product $\quad \mathbf{x}\cdot\mathbf{y} = \left((-1)(3) + (3)(5) + (2)(2)\right) = -3+15+4 = 16$

Displacement from $\mathbf{x}$ to $\mathbf{y}$ $\quad \mathbf{y} - \mathbf{x} = (3-(-1),\ 5-3,\ 2-2) = (4, 2, 0)$

Distance from $\mathbf{x}$ to $\mathbf{y}$ $= |\mathbf{y} - \mathbf{x}| = (4^2 + 2^2 + 0^2)^{1/2} = (16+4)^{1/2} = \sqrt{20}$

Sometimes it is useful to write vectors as columns instead of rows

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

We use the superscript "$'$" for <u>transponse</u> to convert a vector from column to row and viceversa

$$(x_1, x_2 \cdots x_N)' = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}$$

Vector Algebra

$\mathbf{y} + \mathbf{x} = \mathbf{x} + \mathbf{y}$ commutative

$\mathbf{x} = 0 + \mathbf{x} = \mathbf{x} + 0$ additive identity

$\mathbf{x} - \mathbf{x} = \mathbf{x} + (-\mathbf{x}) = 0$ additive inverse

$(\mathbf{x} + \mathbf{y}) + \mathbf{z} = (\mathbf{x}) + (\mathbf{y} + \mathbf{z}) =$ associative

$(a+b)\mathbf{x} = a\mathbf{x} + b\mathbf{x}$ distributed for scalars

$a(\mathbf{x} + \mathbf{y}) = a\mathbf{x} + a\mathbf{y}$ distubuted for vectors

$a(b\mathbf{x}) = ab\mathbf{x}$ associative for scalars

Matrices

If $\mathbb{A}$ is a matrix, then the letter a_{ij} denotes the number in location a_{ij} (ith row and jth column)

If A is a $M \times N$ matrix then it has M rows and N columns

Example:

$$\mathbb{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{bmatrix}$$ $\mathbb{A}$ is a 3×4 Matrix

You can see a $M \times N$ matrix as M rows of N dimensional vectors

$$\mathbb{A} = \begin{bmatrix} \mathbb{V}_1 \\ \mathbb{V}_2 \\ \vdots \\ \mathbb{V}_M \end{bmatrix} = \left[\mathbb{V}_1, \mathbb{V}_2, \mathbb{V}_3 \ldots \mathbb{V}_M\right]^T$$

or it can be seen as N columns with N M-dimensional vectors

$$\mathbb{A} = \begin{bmatrix} \mathbb{C}_1 & \mathbb{C}_2 & \mathbb{C}_3 & \cdots & \mathbb{C}_N \end{bmatrix}$$

Operations

Equality

$\mathbb{A} = \mathbb{B}$ if and only if $a_{ij} = b_{ij}$
$1 \leq i \leq M, \ 1 \leq j \leq N$

Sum $\mathbb{A} + \mathbb{B} = [a_{ij} + b_{ij}]_{M \times N}$

Negation $-\mathbb{A} = [-a_{ij}]_{M \times N}$

Difference $\mathbb{A} - \mathbb{B} = [a_{ij} - b_{ij}]_{M \times N}$

Scalar Multiplication $c\mathbb{A} = [c\,a_{ij}]_{M \times N}$

Weighted Sum

$$p\mathbb{A} + q\mathbb{B} = [p\,a_{ij} + q\,b_{ij}]$$

Example

$$\mathbb{A} = \begin{bmatrix} 2 & 1 \\ 4 & 6 \\ 3 & 2 \end{bmatrix} \qquad \mathbb{B} = \begin{bmatrix} 1 & 7 \\ 6 & 2 \\ 2 & 1 \end{bmatrix}$$

$$3\mathbb{A} + 2\mathbb{B} = 3\begin{bmatrix} 2 & 1 \\ 4 & 6 \\ 3 & 2 \end{bmatrix} + 2\begin{bmatrix} 1 & 7 \\ 6 & 2 \\ 2 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 6 & 3 \\ 12 & 18 \\ 9 & 6 \end{bmatrix} + \begin{bmatrix} 2 & 14 \\ 12 & 4 \\ 4 & 2 \end{bmatrix}$$

$$3\mathbb{A} + 2\mathbb{B} = \begin{bmatrix} 8 & 17 \\ 24 & 22 \\ 13 & 8 \end{bmatrix}$$

Matrix Properties

$\mathbb{B}+\mathbb{A}=\mathbb{A}+\mathbb{B}$ commutative

$\mathbb{0}+\mathbb{A}=\mathbb{A}+\mathbb{0}$ additive identity

$\mathbb{A}-\mathbb{A}=\mathbb{A}+(-\mathbb{A})=0$ additive inverse

$(\mathbb{A}+\mathbb{B})+\mathbb{C}=\mathbb{A}+(\mathbb{B}+\mathbb{C})$ = associtive

$(p+q)\mathbb{A}=p\mathbb{A}+q\mathbb{A}$ distributed for scalars

$p(\mathbb{A}+\mathbb{B})=p\mathbb{A}+p\mathbb{B}$ distributed for matrices

Matrix Multiplication

$\mathbb{A}\mathbb{B}=\mathbb{C}$ $\quad \mathbb{A}=[a_{ik}]_{M\times N}$

$\mathbb{B}=[b_{kj}]_{N\times P}$

$$C_{ij}=\sum_{k=1}^{N} a_{ik}b_{kj}=a_{i1}b_{1j}+a_{i2}b_{2j}+\dots a_{iN}b_{Nj}$$

for $i=1,2\dots M$ $\quad j=1,2\dots P$

Example

$$\mathbb{A} = \begin{bmatrix} 3 & 1 & 2 \\ 2 & 6 & 5 \end{bmatrix} \qquad \mathbb{B} = \begin{bmatrix} 6 & 4 & 6 & 4 \\ 2 & 1 & 3 & 1 \\ 1 & 2 & 1 & 0 \end{bmatrix}$$

$$2 \times 3 \qquad\qquad 3 \times 4$$

$\mathbb{C} = \mathbb{A} \times \mathbb{B} =$

$$\begin{bmatrix} 3\cdot6+1\cdot2+2\cdot1 & 3\cdot4+1\cdot1+2\cdot2 & 3\cdot6+1\cdot3+2\cdot1 & 3\cdot4+1\cdot1+2\cdot0 \\ 2\cdot6+6\cdot2\cdot+5\cdot1 & 2\cdot4+6\cdot1+5\cdot2 & 2\cdot6+6\cdot3+5\cdot1 & 2\cdot4+6\cdot1+5\cdot0 \end{bmatrix}$$

$$\begin{bmatrix} 22 & 17 & 23 & 13 \\ 29 & 24 & 35 & 14 \end{bmatrix}$$

Special Matrices 52

$\mathbb{O} = [0]_{M \times N}$ Example $\mathbb{O}_{3\times 2} \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$

Identity Matrix

$I_N = [\delta_{ij}]_{N \times N}$ where $\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$

Example

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \times \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}$$

Matrix Multiplication $\longrightarrow$ No commutative $\mathbb{A}\mathbb{B} \neq \mathbb{B}\mathbb{A}$

$(\mathbb{A}\mathbb{B})\mathbb{C} = \mathbb{A}(\mathbb{B}\mathbb{C})$ Associative

$\mathbb{I}\mathbb{A} = \mathbb{A}\mathbb{I} = \mathbb{A}$ identity

$\mathbb{A}(\mathbb{B}+\mathbb{C}) = \mathbb{A}\mathbb{B} + \mathbb{A}\mathbb{C}$ left distributed

$(\mathbb{A}+\mathbb{B})\mathbb{C} = \mathbb{A}\mathbb{C} + \mathbb{B}\mathbb{C}$ right distributed

$c\mathbb{A}\mathbb{B} = (c\mathbb{A})\mathbb{B} = \mathbb{A}c\mathbb{B}$ scalar distributed

Inverse of a nonsingular matrix

A $N \times N$ matrix $\mathbb{A}$ is called nonsingular or invertible if there exists a $N \times N$ matrix $\mathbb{B}$ such that

$$\mathbb{A}\mathbb{B} = \mathbb{B}\mathbb{A} = \mathbb{I}$$

If there is such $\mathbb{B}$, $\mathbb{A}$ is said to be nonsingular and $\mathbb{B}$ is the inverse of $\mathbb{A}$

If $\mathbb{B}$ can be found, it can be written as

$$\mathbb{B} = \mathbb{A}^{-1} \qquad \mathbb{A}\mathbb{A}^{-1} = \mathbb{A}^{-1}\mathbb{A} = \mathbb{I}$$

Upper Triangular System of Equations

A matrix $A = [a_{ij}]$ is called <u>upper triangular</u> if $a_{ij} = \phi$ when $i > j$

$$
\begin{array}{rcrcccrl}
a_{11}x_1 + & a_{12}x_2 + & \cdots & \cdots & & a_{1N}x_N & = b_1 \\
 & a_{22}x_2 + & a_{23}x_3 & \cdots & & a_{2N}x_N & = b_2 \\
 & & a_{33}x_3 & \cdots & & a_{3N}x_N & = b_3 \\
 & \phi & & \ddots & & \vdots & \\
 & & & & a_{N-1,N-1}x_{N-1} + & a_{N-1,N}x_N & = b_{N-1} \\
 & & & & & a_{NN}x_N & = b_N
\end{array}
$$

To solve an Upper triangular system of equations we use "Back Substitution". 55

$$X_N = \frac{b_N}{a_{NN}}$$

$$X_{N-1} = \frac{b_{N-1} - a_{N-1,N} X_N}{a_{N-1,N-1}}$$

$$\vdots$$

$$X_1 = \frac{b_1 - a_{1N} X_N - a_{1,N-1} X_{N-1} \cdots a_{12} X_2}{a_{11}}$$

In general

$$X_K = \frac{b_K - \sum_{J=K+1}^{N} a_{KJ} X_J}{a_{KK}} \quad \text{for } K = N-1, N-2, \ldots 1$$

Gauss Elimination

Gauss elimination is an optimal method to solve any singular system of linear equations that converts the system into an upper triangular system first, using some valid transformations and then it uses back substitution.

Basic Transformations

1. Interchange:
   The order of two equations can be changed and it does not affect the solution.
2. Scaling:
   Multiplying an equation by a constant will not affect the solution
3. Replacement:
   An equation can be replaced by the sum of itself and a non-zero multiple of another equation without affecting the solution.

We will use these transformations to convert a system of linear equations to upper triangular and then solve it using back substitution. (57)

$$\begin{aligned} 1x + 3y + 4z &= 19 \\ 8x + 9y + 3z &= 35 \\ x + y + z &= 6 \end{aligned} \Rightarrow \begin{bmatrix} 1 & 3 & 4 \\ 8 & 9 & 3 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 19 \\ 35 \\ 6 \end{bmatrix}$$

$\mathbb{A}\mathbb{X} + \mathbb{B}$

$\mathbb{A} \quad \mathbb{X} : \mathbb{B}$

We put together $\mathbb{A}$ and $\mathbb{B}$ into a single matrix called "augmented matrix" to facilitate Gauss elimination

$$\begin{bmatrix} 1 & 3 & 4 & 19 \\ 8 & 9 & 3 & 35 \\ 1 & 1 & 1 & 6 \end{bmatrix}$$

we want to transfor this system into upper diagonal

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & b_1 \\ 0 & a_{22} & a_{23} & b_2 \\ 0 & 0 & a_{33} & b_3 \end{bmatrix}$$

pivot

$$\begin{bmatrix} 1 & 3 & 4 & 19 \\ 8 & 9 & 3 & 35 \\ 1 & 1 & 1 & 6 \end{bmatrix} \quad \frac{1}{\text{pivot}} = \frac{1}{1}$$

$$\begin{array}{c} A \\ B \\ C \end{array} \begin{bmatrix} 1 & 3 & 4 & 19 \\ 8 & 9 & 3 & 35 \\ 1 & 1 & 1 & 6 \end{bmatrix}$$

m

$$\begin{bmatrix} 1 & 3 & 4 & 19 \\ 8-8(1) & 9-8\cdot 3 & 3-8\cdot 4 & 35-8\cdot 19 \\ 1-(1)(1) & 1-1\cdot 3 & 1-1\cdot 4 & 6-1\cdot 19 \end{bmatrix} \begin{array}{l} \\ \leftarrow B-8A \\ \leftarrow C-(1)A \end{array}$$

pivot

$$\begin{bmatrix} 1 & 3 & 4 & 19 \\ 0 & -15 & -29 & -117 \\ 0 & -2 & -3 & -13 \end{bmatrix} \quad \frac{1}{-15} \leftarrow \text{pivot}$$

$$\begin{array}{c} \begin{bmatrix} 1 & 3 & 4 & 19 \\ 0 & 1 & +\frac{29}{15} & \frac{117}{15} \\ 0 & -2 & -3 & -13 \end{bmatrix} \end{array} \begin{array}{c} A \\ B \\ C \end{array}$$

$$\begin{bmatrix} 1 & 3 & 4 & 19 \\ 0 & 1 & 29/15 & \frac{117}{15} \\ 0 & -2+2\cdot 1 & -3+2\cdot\frac{29}{15} & -13+2\cdot\frac{117}{15} \end{bmatrix} \leftarrow C+2B$$

$$\begin{bmatrix} 1 & 3 & 4 & 19 \\ 0 & 1 & 29/15 & \frac{117}{15} \\ 0 & 0 & \underset{\text{pivot}}{.86666} & 2.6 \end{bmatrix} \leftarrow \frac{1}{.8666}$$

$$\begin{bmatrix} 1 & 3 & 4 & 19 \\ 0 & 1 & 1.9333 & 7.8 \\ 0 & 0 & 1 & 3 \end{bmatrix}$$

rebuilding system of linear equations

$$1x + 3y + 4z = 19$$
$$y + 1.9333z = 7.8$$
$$z = 3$$

Using backward sustitution.

$\underline{\underline{z = 3}}$

$$y = 7.8 - 1.9333(3)$$

$\underline{\underline{y = 2}}$

$$x = 19 - 3y - 4z$$
$$= 19 - 3(2) - 4(3)$$

$\underline{\underline{x = 1}}$

Notes on Gauss Elimination

- If pivot = 0 at some step then switch rows under the pivotal row to prevent division by zero
- If pivot = 0 in the last row then there is no solution.
- To reduce computing error you may use as the next pivot the largest number in the rows that are still left to transform. switch pivotal row withe that row.

$$\begin{bmatrix} 1 & 3 & 4 & 19 \\ 0 & 2 & 1 & 6 \\ 0 & 9 & 3 & 39 \end{bmatrix} \quad 9>2$$

switch these lines to have largest pivot

$$\Downarrow$$

$$\begin{bmatrix} 1 & 3 & 4 & 19 \\ 0 & 9 & 3 & 43 \\ 0 & 2 & 1 & 6 \end{bmatrix}$$

this reduces error.

Implementing Gauss Elimination in Matlab

gauss.m

```
function X = gauss(A,B)
% Input  - A is a NxN nonsingular matrix
%        - B is a Nx1 matrix
% Output - X is a Nx1 matrix with
%          the solution Ax=B
%

% Get dimensions of A
[N N] = size(A)

% Initialize X with zeroes
X = zeros(N,1)

% Obtain augmented matrix
Aug = [A B]
```

```
% Gauss elimination

% For all rows. Make upper triangular matrix
for p=1:N
   % Choose pivot. We ignore for now case when p=0
   piv= Aug(p,p)
   % Divide pivotal row by pivot
   for k=p:N+1
      Aug(p,k)=Aug(p,k)/piv
   end
   % Make 0's the other elements in pivotal column
   for k= p+1:N   % for all remaining rows
      m= Aug(k,p)
      for i=p:N+1   % for all remaining columns
         Aug(k,i) = Aug(k,i)-m*Aug(p,i)
      end
   end
end
```

It can be done with matlab short notation

Aug(p,p:N+1)= Aug(p,p:N+1)/piv

```
% Backward Sustitution
for K=N:1
    sum = 0   % Accumulate elements in upper triangle
    for i=K+1:N
        sum = sum + Aug(K,i)*X(i,1)
    end
    X(K,1)= (Aug(K,N+1)-Sum)/Aug(K,K)
end
```

Triangular Factorization

In some cases in scientific computing you have to solve multiple systems of linear equations that share the same $\mathbb{A}$

$$\mathbb{A}x_1 = b_1$$
$$\mathbb{A}x_2 = b_2$$
$$\mathbb{A}x_3 = b_3$$
$$\vdots$$
$$\mathbb{A}x_N = b_N$$

Instead of using gauss elimination for each system of equations (N times),

we can save some work by first factorizing $\mathbb{A}$ into two matrices $\mathbb{L}$ and $\mathbb{U}$ such that $\mathbb{A} = \mathbb{L}\mathbb{U}$ and then doing "forward" and then "backward" substitution.

A matrix $\mathbb{A}$ has a triangular factorization if it can be expressed as the product of a lower triangular matrix $\mathbb{L}$ and a upper triangular matrix $\mathbb{U}$.

$$\mathbb{A} = \mathbb{L}\,\mathbb{U}$$

$$\underbrace{\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & & & \\ a_{N1} & a_{N2} & \cdots & a_{NN} \end{bmatrix}}_{\mathbb{A}} = \underbrace{\begin{bmatrix} 1 & 0 & \cdots & 0 \\ m_{21} & 1 & \cdots & 0 \\ m_{31} & m_{32} & 1 \cdots & 0 \\ \vdots & & & \\ m_{N1} & m_{N2} & \cdots & 1 \end{bmatrix}}_{\mathbb{L}} \underbrace{\begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1N} \\ 0 & u_{22} & \cdots & u_{2N} \\ 0 & 0 & \cdots & u_{3N} \\ & & & \vdots \\ 0 & 0 & 0 \cdots & u_{NN} \end{bmatrix}}_{\mathbb{U}}$$

Assume the linear system $\mathbb{A}\,\mathbb{X} = \mathbb{B}$

$$\mathbb{A} = \mathbb{L}\mathbb{U}$$

$$\mathbb{A}\,\mathbb{X} = \mathbb{B}$$

$$\mathbb{L}\,\mathbb{U}\,\mathbb{X} = \mathbb{B}$$

Then we can define $\mathbb{Y} = \mathbb{U}\,\mathbb{X}$ ① so that

$$\mathbb{L}\,\mathbb{Y} = \mathbb{B} \quad ②$$

$$\begin{array}{c} \mathbb{L} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \cdots & 0 \\ m_{21} & 1 & \cdots & \cdots & 0 \\ m_{31} & m_{32} & 1 & \cdots & 0 \\ m_{N1} & m_{N2} & & \cdots & 1 \end{bmatrix} \end{array} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_N \end{bmatrix}$$

So we can solve ②
by using "forward substitution"
to find $y_1, y_2 \cdots y_N$

to obtain x_1, x_2, x_N we solve ①

$\mathbb{U}\mathbb{X} = \mathbb{Y}$

using backward substitution.

$$\begin{bmatrix} U_{11} & U_{12} & \cdots & U_{1N} \\ 0 & U_{22} & \cdots & U_{2N} \\ 0 & 0 & \cdots & U_{3N} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & U_{NN} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ N_N \end{bmatrix} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix}$$

So we can solve multiple M systems of equations $\mathbb{A}\mathbb{X} = \mathbb{B}_1$, $\mathbb{A}\mathbb{X} = \mathbb{B}_2 \cdots \mathbb{A}\mathbb{X} = \mathbb{B}_N$ by first

1) finding $\underline{\underline{\mathbb{L}\mathbb{U}}} = \mathbb{A}$ $O(N^3)$ and then for each system of equations $i = 1 \ldots \underline{M}$ ~~and~~ solve $\underline{\mathbb{L}\mathbb{Y}_i = \mathbb{B}_i}$ $O(N^2)$ and then $\mathbb{U}\mathbb{X} = \mathbb{Y}_i$.

$O(N^3) + O(N^2)M$

LU Factorization example

we want to solve the equations

$$6x_1 + 1x_2 - 4x_3 = 3$$
$$5x_1 + 3x_2 + 2x_3 = 21$$
$$1x_1 - 4x_2 + 3x_3 = 10$$

and

$$6x_1 + 1x_2 - 4x_3 = 3$$
$$5x_1 + 3x_2 + 2x_2 = 10$$
$$1x_1 - 4x_2 + 3x_3 = 0$$

these two systems share the same $\mathbb{A}$ but they have different $\mathbb{B}$

To do the LU factorization we start with the following

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \begin{array}{c} \begin{bmatrix} 6 & 1 & -4 \\ 5 & 3 & 2 \\ 1 & -4 & 3 \end{bmatrix} \end{array} \begin{array}{l} R_1 \\ R_2 \\ R_3 \end{array}$$

$$\mathbb{I} \Rightarrow \mathbb{L} \qquad \mathbb{A} \Rightarrow \mathbb{U}$$

we will do transformations to these matrices such that $\mathbb{I}$ will become $\mathbb{L}$ and $\mathbb{A}$ will become $\mathbb{U}$

We want to make $a_{21}=5$ to become ϕ

to do this we make $R_2 \Leftarrow R_2 - \left(\frac{5}{6}\right)R_1$ $\rightarrow m = \frac{5}{6}$

$$\begin{array}{r}\\ m \rightarrow \\ \\ \end{array}\begin{bmatrix}1&0&0\\5/6&1&0\\0&0&1\end{bmatrix}\begin{bmatrix}6&1&-4\\5-\frac{5}{6}\cdot 5&3-\frac{5}{6}\cdot 1&2-\frac{5}{6}\cdot(-4)\\1&-4&3\end{bmatrix}$$

$$3-\frac{5}{6}=\frac{13}{6}$$

$$2+\frac{20}{6}=\frac{32}{6}$$

$$\begin{bmatrix}1&0&0\\5/6&1&0\\0&0&1\end{bmatrix}\begin{bmatrix}6&1&-4\\0&13/6&32/6\\1&-4&3\end{bmatrix}$$

to make $a_{31}=1$ to become ϕ we make $R_3 \Leftarrow R_3 - \left(\frac{1}{6}\right)R_1$ $\rightarrow \underline{\underline{m}}$

$$\begin{array}{r}\\ \\ \underline{\underline{m}} \rightarrow \end{array}\begin{bmatrix}1&0&0\\5/6&1&0\\1/6&0&\phi\end{bmatrix}\begin{bmatrix}6&1&-4\\0&13/6&32/6\\1-\frac{1}{6}6&-4-\left(\frac{1}{6}\right)1&3-\frac{1}{6}(-4)\end{bmatrix}$$

$$\begin{bmatrix}1&0&0\\5/6&1&0\\1/6&0&1\end{bmatrix}\begin{array}{c}\begin{bmatrix}6&1&-4\\0&13/6&32/6\\0&-25/6&\frac{22}{6}\end{bmatrix}\end{array}\begin{array}{l}R_1\\R_2\\R_3\end{array}$$

Now we want $a_{32} = -\frac{25}{6}$ to become ϕ
so we do

$$R_3 \Leftarrow R_3 - \underbrace{\left(-\frac{25}{\cancel{6}}\right)\left(\frac{\cancel{6}}{13}\right)}_{m = -\frac{25}{13}} R_2$$

$$\begin{bmatrix} 1 & 0 & 0 \\ 5/6 & 1 & 0 \\ 1/6 & -\frac{25}{13} & 1 \end{bmatrix} \begin{bmatrix} 6 & 1 & -4 \\ 0 & 13/6 & 32/6 \\ 0 & -\frac{25}{6} - \left(-\frac{25}{\cancel{6}}\right)\left(\frac{\cancel{6}}{\cancel{13}}\right)\frac{\cancel{13}}{6} & \frac{22}{6} - \left(-\frac{25}{6}\right)\left(\frac{\cancel{6}}{13}\right)\frac{32}{\cancel{6}} \end{bmatrix}$$

m

$$A = \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 5/6 & 1 & 0 \\ 1/6 & -\frac{25}{13} & 1 \end{bmatrix}}_{L} \underbrace{\begin{bmatrix} 6 & 1 & -4 \\ 0 & 13/6 & 32/6 \\ 0 & 0 & \frac{1086}{78} \end{bmatrix}}_{U}$$

$$\frac{22}{6} + \frac{25}{6} \cdot \frac{32}{13} = \frac{1086}{78}$$

Now we use $\mathbb{L}\mathbb{U}$ to solve the first system of equations

$$6x_1 + 1x_2 - 4x_3 = 3$$
$$5x_1 + 3x_2 + 2x_3 = 21$$
$$1x_1 - 4x_2 + 3x_3 = 10$$

1st system of linear equations.

$\mathbb{A}\mathbb{X} = \mathbb{B}$

$\mathbb{L}\underbrace{\mathbb{U}\mathbb{X}}_{\mathbb{Y}} = \mathbb{B} \Rightarrow \mathbb{L}\mathbb{Y} = B$ ①

and $\mathbb{U}\mathbb{X} = \mathbb{Y}$ ②

First solve ①

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 5/6 & 1 & 0 \\ 1/6 & -\frac{25}{13} & \text{[redacted]} \end{bmatrix}}_{\mathbb{L}} \underbrace{\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix}}_{\mathbb{Y}} = \underbrace{\begin{bmatrix} 3 \\ 21 \\ 10 \end{bmatrix}}_{\mathbb{B}}$$

$$Y_1 = 3$$
$$5/6\, Y_1 + Y_2 = 21$$
$$1/6\, Y_1 - \frac{25}{13} Y_2 + Y_3 = 10$$

72

We use forward sustitution to solve this system of linear equations

$y_1 = 3$

21

$y_2 = 21 - \frac{5}{6}(3) = 2 - \frac{5}{2} = 21 - 2.5 = 18.5$

$y_3 = 10 - \frac{1}{6}(3) + \frac{25}{13}(18.5) = 45.0769$

Now we solve 2 $UX = Y$

$$\underbrace{\begin{bmatrix} 6 & 1 & -4 \\ 0 & 13/6 & 32/6 \\ 0 & 0 & \frac{1086}{78} \end{bmatrix}}_{U} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 21 \\ 18.5 \\ 45.0769 \end{bmatrix}$$

$$\begin{bmatrix} 6 & 1 & -4 \\ 0 & 2.1667 & 5.333 \\ 0 & 0 & 13.9231 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 21 \\ 18.5 \\ 45.0769 \end{bmatrix}$$

$6x_1 + x_2 - 4x_3 = 21$

$2.1667x_2 + 5.333x_3 = 18.5$

$13.9231x_3 = 45.0769$

We solve this system of equations using backward sustitution:

$$X_3 = \frac{45.0764}{13.9231} = 3.2376$$

$$X_2 = \frac{18.5 - 5.333(3.2376)}{2.1667} = .5695$$

$$X_1 = \frac{21 - .5695 + 4(3.2376)}{6} = 2.5635$$

Do the same to solve the second system of linear equations

$$6X_1 + 1X_2 - 4X_3 = 3$$
$$5X_1 + 3X_2 + 2X_3 = 10$$
$$1X_1 - 4X_2 + 3X_3 = 0$$

(You can solve it at home).

Complexity:

- Gauss elimination takes $O(N^3)$ time
  So if we want to solve M systems of linear equations $M\,O(N^3) = \underline{O(MN^3)}$

- Using LU factorization to solve M systems of linear equation takes

  LU factorization (only once) $\quad O(N^3)$

  forward sustitution (M times) $\quad M(O(N^2)) = O(MN^2)$

  backward sustitution (M times) $\quad M\,O(N^2) = O(MN^2)$

  $$\text{total} = \underline{O(N^3) + O(MN^2)}$$

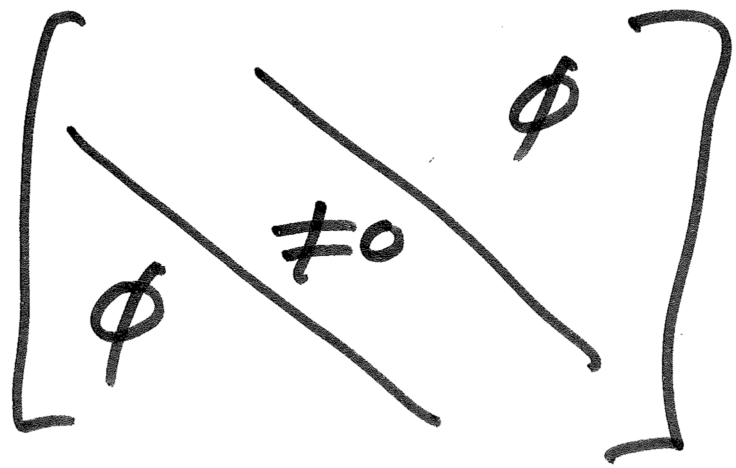
So it takes less time to use LU factorization to solve M systems of linear equations with the same A.

Iterative Methods for Linear Equations

Motivation

In the numerical solution of partial differential equations and othe problems, often we have linear systems with as many as 1,000,000 variables. The coefficients of these equations are mostly zeroes leading to a sparse matrix. A sparse matrix is one where a large percentage of the coefficients is zero

$$\begin{bmatrix} a_{11} & a_{12} & 0 & 0 & 0 & 0 & \phi \\ a_{21} & a_{22} & a_{23} & 0 & \cdots & & \\ 0 & a_{32} & a_{33} & a_{34} & \cdots & & \\ \phi & & & & & & \\ 0 & 0 & 0 & 0 & & a_{N,N-1} & a_{NN} \end{bmatrix}$$



Iterative Methods for Systems of Linear Equations

Jacobi Iteration

~~Let~~ We have the system

$$\text{(1)} \quad 3x + y = 5$$

$$\text{(2)} \quad x + 3y = 7$$

We can rewrite this system as the following iteration

From (1) $\quad x_{k+1} = \dfrac{5 - y_k}{3}$

From (2) $\quad y_{k+1} = \dfrac{7 - x_k}{3}$

Lets start at $x_0 = 0,\ y_0 = 0$

$x_0 = 0 \qquad y_0 = 0$

$x_1 = \frac{5-0}{3} = 1.667 \qquad y_1 = \frac{7-0}{3} = 2.33$

$x_2 = \frac{5-2.33}{3} = .889 \qquad y_2 = \frac{7-1.667}{3} = 1.777$

$x_3 = \frac{5-1.777}{3} = 1.074 \qquad y_3 = \frac{7-.889}{3} = 2.0377$

$x_4 = \frac{5-2.0377}{3} = .987 \qquad y_4 = \frac{7-1.074}{3} = 1.975$

$\vdots \qquad \vdots$

$x_N = 1 \qquad y_N = 2$

Gauss Seidel Iteration

We can speed up the convergence of the solution by using the values of the iteration as they are obtained. This is called "Gauss Seidel Iteration".

Jacobi

$$x_{k+1} = \frac{5 - y_k}{3}$$

$3x + y = 5$

$$y_{k+1} = \frac{7 - x_k}{3}$$

Gauss Seidel

$$x_{k+1} = \frac{5 - y_k}{3}$$

$$y_{k+1} = \frac{7 - x_{k+1}}{3}$$

See difference

Gauss Seidel Example

$x_0 = 0 \qquad y_0 = 0$

$x_1 = \dfrac{5-0}{3} = 1.667 \qquad y_1 = \dfrac{7-1.667}{3} = 1.778$

$x_2 = \dfrac{5-1.778}{3} = 1.074 \qquad y_2 = \dfrac{7-1.074}{3} = 1.975$

$x_3 = \dfrac{5-1.975}{3} = 1.008 \qquad y_3 = \dfrac{7-1.008}{3} = 1.997$

$\vdots \qquad\qquad \vdots$

$x_N = 1 \qquad y_N = 2$

Gauss Seidel converges faster than Jacobi

Using the estimated values as soon as they are produced, speeds up the convergence.

Jacobi and Gauss Seidel may not converge in some cases.

For example if we rearange the previous system of equations

$$3x + y = 5 \qquad \text{switch order} \qquad x + 3y = 7$$
$$x + 3y = 7 \qquad\qquad\qquad\quad 3x + y = 5$$

This new system will lead to the following iterative method

$$x_{k+1} = \frac{7 - 3y_k}{1} \quad \text{and} \quad y_{k+1} = \frac{5 - 3x_k}{1}$$

Assume $x_0 = 0$ $\qquad$ $y_0 = 0$

$$x_1 = \frac{7-0}{1} = 7 \qquad y_1 = \frac{5-0}{1} = 5$$

$$x_2 = \frac{7-3(5)}{1} = -8 \qquad y_2 = \frac{5-3(7)}{1} = -16$$

It does not converge!

$$x_3 = \frac{7-3(-16)}{1} = 55 \qquad y_3 = \frac{5-3(-8)}{1} = 29$$

$$\vdots \qquad\qquad\qquad \vdots$$

How is convergence ensured?

A matrix A of dimension $N \times N$ is said to be "strictly diagonal dominant"

if

$$|a_{KK}| > \sum_{\substack{J=1 \\ J \neq K}}^{N} |a_{KJ}| \quad K = 1, 2 \ldots N$$

Jacobi Iteration (and also Gauss Seidel) will converge if A is strictly diagonal

Example

Case 1

$$\begin{aligned} 3X + Y &= 5 \\ X + 3Y &= 7 \end{aligned} \Rightarrow \begin{aligned} X_{K+1} &= \frac{5 - Y_K}{3} \\ Y_{K+1} &= \frac{7 - X_K}{3} \end{aligned} \Rightarrow A = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$$

Diagonal Dominant a_{12}

$|a_{11}| = |3| > |1|$

$|a_{22}| = |3| > |1|$

This A is diagonal dominant so it will converge.

Case 2 $\quad x + 3y = 5$ $\quad \Rightarrow \quad x_{k+1} = 5 - 3y_k$ $\quad \Rightarrow A = \begin{bmatrix} 1 & 3 \\ 3 & 1 \end{bmatrix}$

$3x + y = 7 \quad\quad y_{k+1} = 7 - 3x_k$

$|a_{11}| = |1| \not> |3|$ Not diagonal dominant

A is not diagonal dominant $|a_{22}| = |1| \not> |3|$

There is no guarantee that this iteration will converge.

Example

$$5x + 2y + 1z = 8 \qquad x_{k+1} = \frac{8 - 2y_k - 1z_k}{5}$$

$$3x + 6y + 2z = 11 \quad \Rightarrow \quad y_{k+1} = \frac{11 - 3x_k - 2z_k}{6}$$

$$2x + 3y + 8z = 13 \qquad z_{k+1} = \frac{13 - 2x_k - 3y_k}{8}$$

$$A = \begin{bmatrix} 5 & 2 & 1 \\ 3 & 6 & 2 \\ 2 & 3 & 8 \end{bmatrix}$$

$|a_{11}| = |5| > |2| + |1|$

$|a_{22}| = |6| > |3| + |2|$

$|a_{33}| = |8| > |2| + |3|$

this iteration converges because A is diagonal dominant.

However if we change the order of the equations

$$5x + 2y + 1z = 8$$
$$2x + 3y + 8z = 13$$
$$3x + 6y + 2z = 11$$

$$\Rightarrow \quad x_{k+1} = \frac{8 - 2y_k - 1z_k}{5}, \quad y_{k+1} = \frac{13 - 2x_k - 8z_k}{3}, \quad z_{k+1} = \frac{11 - 3x_k - 6y_k}{2}$$

$$A = \begin{bmatrix} 5 & 2 & 1 \\ 2 & 3 & 8 \\ 3 & 6 & 2 \end{bmatrix}$$

$|a_{11}| = |5| > |2| + |1|$ ✓

$|a_{22}| = |3| \not> |2| + |8|$ X

$|a_{33}| = |2| \not> |3| + |6|$ X

A is not diagonal dominant and therefore this iteration may not converge

We have seen how to solve a system of linear equations iteratively.
Now how do we solve a system of non-linear equations iteratively.

Newton Method for Systems of Nonlinear Equations

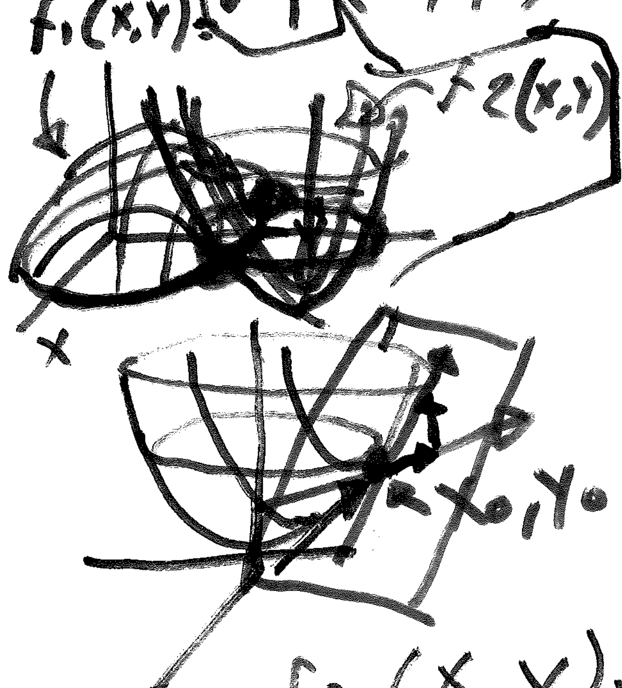
Assume that we have a non-linear system of equations such as:

$$f_1(x,y) = x^2 - y - .2 = 0$$
$$f_2(x,y) = y^2 - x - .3 = 0$$

- to solve this system we extend Newton's method for equations of two variables.
- Assume the taylor expansion around point x_0, y_0 of functions with two variables:

Taylor expansion of $f_1(X,Y)$ and $f_2(X,Y)$

84



$$f_1(X,Y) \cong f_1(X_0,Y_0) + \underbrace{\frac{\partial}{\partial X} f_1(X_0,Y_0)(X - X_0) + \frac{\partial}{\partial Y} f_1(X_0,Y_0)(Y-Y_0)}_{\text{approximate function with a plane}} \; \ldots\ldots$$

and

$$f_2(X,Y) \cong f_2(X_0,Y_0) + \frac{\partial}{\partial X} f_2(X_0,Y_0)(X-X_0) + \frac{\partial}{\partial Y} f_2(X_0,Y_0)(Y-Y_0)$$

we write this in matrix form

$$\begin{bmatrix} f_1(X,Y) \\ f_2(X,X) \end{bmatrix} \approx \begin{bmatrix} \frac{\partial f_1(X_0,Y_0)}{\partial X} & \frac{\partial f_1(X_0,Y_0)}{\partial Y} \\ \frac{\partial f_2(X_0,Y_0)}{\partial X} & \frac{\partial f_2(X_0,Y_0)}{\partial Y} \end{bmatrix} \begin{bmatrix} X - X_0 \\ Y - Y_0 \end{bmatrix} + \begin{bmatrix} f_1(X_0,Y_0) \\ f_2(X_0,Y_0) \end{bmatrix}$$

Jacobian = $\mathbb{J}$

Since we want to obtain X, Y such that $f_1(X,Y)=0$ and $f_2(X,Y)=0$ we make the leftside to be ϕ.

$$\begin{bmatrix} f_1(X,Y)=0 \\ f_2(X,Y)=0 \end{bmatrix} \approx \begin{bmatrix} \dfrac{\partial f_1(X_0,Y_0)}{\partial x} & \dfrac{\partial f_1(X_0,Y_0)}{\partial y} \\ \dfrac{\partial f_2(X_0,Y_0)}{\partial x} & \dfrac{\partial f_2(X_0,Y_0)}{\partial y} \end{bmatrix} \begin{bmatrix} X-X_0 \\ Y-Y_0 \end{bmatrix} + \begin{bmatrix} f_1(X_0,Y_0) \\ f_2(X_0,Y_0) \end{bmatrix}$$

$$-\begin{bmatrix} f_1(X_0,Y_0) \\ f_2(X_0,Y_0) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix} = J \begin{bmatrix} X-X_0 \\ Y-Y_0 \end{bmatrix} + \cancel{\begin{bmatrix} f_1(X_0,Y_0) \\ f_2(X_0,Y_0) \end{bmatrix}} - \cancel{\begin{bmatrix} f_1(X_0,Y_0) \\ f_2(X_0,Y_0) \end{bmatrix}}$$

$$(J)^{-1}\left(- \begin{bmatrix} f_1(X_0,Y_0) \\ f_2(X_0,Y_0) \end{bmatrix} \right) = (J)^{-1} J \begin{bmatrix} X-X_0 \\ Y-Y_0 \end{bmatrix}$$

$$-J^{-1} \begin{bmatrix} f_1(X_0,Y_0) \\ f_2(X_0,Y_0) \end{bmatrix} = \begin{bmatrix} X-X_0 \\ Y-Y_0 \end{bmatrix}$$

$$-J^{-1}\begin{bmatrix} f_1(x_0,y_0) \\ f_2(x_0,y_0) \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix} - \begin{bmatrix} x_0 \\ y_0 \end{bmatrix}$$

sustitute $x = x_1$ $y = y_1$

$$\boxed{\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} - J^{-1}\begin{bmatrix} f_1(x_0,y_0) \\ f_2(x_0,y_0) \end{bmatrix}}$$

Newton method for systems of non linear equations

Notice that this method (equation) is very similar to the one dimensional Newton Raphson method.

$$p_1 = p_0 - \frac{f(p_0)}{f'(p_0)}$$

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} - J^{-1}\begin{bmatrix} f_1(x_0,y_0) \\ f_2(x_0,y_0) \end{bmatrix}$$

Since obtaining J^{-1} the inverse of J takes time, we will solve instead the equations:

$$① \quad \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} f_1(x_0, y_0) \\ f_2(x_0, y_0) \end{bmatrix} + J \begin{bmatrix} x_1 - x_0 \\ y_1 - y_0 \end{bmatrix}$$

using Gaussian elimination.

Newton method for systems of non linear equations without computing J^{-1}. Find x_1, y_1 using Gauss elimination.

Example:

$$f_1(x,y) = x^2 - y - .2 = 0$$
$$f_2(x,y) = y^2 - x - .3 = 0$$

$$J = \begin{bmatrix} \frac{\partial f_1}{\partial x} & \frac{\partial f_1}{\partial y} \\ \frac{\partial f_2}{\partial x} & \frac{\partial f_2}{\partial y} \end{bmatrix} = \begin{bmatrix} 2x & -1 \\ -1 & 2y \end{bmatrix}$$

Assume $x_0 = 0$ $y_0 = 0$ we use 1

$i = 1$

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0^2 - 0 - .2 \\ 0^2 - 0 - .3 \end{bmatrix} + \begin{bmatrix} 2(0) & -1 \\ -1 & 2(0) \end{bmatrix} \begin{bmatrix} x_1 - 0 \\ y_1 - 0 \end{bmatrix}$$

Simplifying we obtain the equations

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} -.2 \\ -.3 \end{bmatrix} + \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$$

$$0 = -.2 - y_1 \quad \Rightarrow \quad y_1 = -.2$$
$$0 = -.3 - x_1 \quad \quad x_1 = -.3$$

$i=2$

$f_1 \to x^2-y-.2$ $\qquad J \to \begin{bmatrix} 2x & -1 \\ -1 & 2y \end{bmatrix}$

$$\begin{bmatrix}0\\0\end{bmatrix} = \begin{bmatrix} f_1(x_1,y_1) \\ f_2(x_1,y_1)\end{bmatrix} + J(x_1,y_1)\begin{bmatrix}x_2-x_1\\y_2-y_1\end{bmatrix}$$

$f_2 \to y^2-x-.3$

$$\begin{bmatrix}0\\0\end{bmatrix} = \begin{bmatrix} (-.3)^2 - \cancel{(-.2)-.2} \\ (-.2)^2 - \cancel{(-.3)-.3}\end{bmatrix} + \begin{bmatrix} 2(-.3) & -1 \\ -1 & 2(-.2)\end{bmatrix}\begin{bmatrix}x_2-(-.3)\\y_2-(-.2)\end{bmatrix}$$

$$\begin{bmatrix}0\\0\end{bmatrix} = \begin{bmatrix} .09 \\ .04\end{bmatrix} + \begin{bmatrix} -.6 & -1 \\ -1 & -.4\end{bmatrix}\begin{bmatrix}x_2+.3\\y_2+.2\end{bmatrix}$$

$$0 = .09 - .6(x_2+.3) - 1(y_2+.2)$$
$$0 = .04 - 1(x_2+.3) - .4(y_2+.2)$$

$$0 = \underline{.09} - .6x_2 - \underline{.18} - y_2 - \underline{.2}$$
$$0 = .04 - x_2 - .3 - .4y_2 - .08$$

$$0 = -.29 - .6X_2 - Y_2$$

$$0 = -.34 - X_2 - .4Y_2$$

$$-.6X_2 - Y_2 = .29$$

$$-X_2 - .4Y_2 = .34$$

$$\begin{array}{c c} R_1 & \\ R_2 & \end{array}\begin{bmatrix} -.6 & -1 \\ -1 & -.4 \end{bmatrix} \begin{bmatrix} X_2 \\ Y_2 \end{bmatrix} = \begin{bmatrix} .29 \\ .34 \end{bmatrix}$$

$$R2 - \frac{R_1}{-.6}(-1) \begin{bmatrix} -.6 & -1 \\ -1 - \frac{-.6}{-.6}(-1) & -.4 - \frac{-1}{-.6}(-1) \end{bmatrix} \begin{bmatrix} X_2 \\ Y_2 \end{bmatrix} = \begin{bmatrix} .29 \\ .34 - \frac{.29}{-.6}(-1) \end{bmatrix}$$

$\rightarrow .34 - .4833$

$.167$

$-.4 + \frac{1}{6}$

$$\begin{bmatrix} -.6 & -1 \\ 0 & -.233 \end{bmatrix} \begin{bmatrix} X_2 \\ Y_2 \end{bmatrix} = \begin{bmatrix} .29 \\ -.1433 \end{bmatrix}$$

with back sust.

$$Y_2 = \frac{-.1433}{-.233} = .6151$$

$$X_2 = [.29 + (.6151)]\frac{1}{-.6} = -1.5085$$

Interpolation and polinomial Approximation

- We want to approximate functions using polinomials.
- We used this for example to compute $\sin(x)$, $\cos(x)$, e^x etc in the computer

Taylor approximation

A polynomial $P_N(x)$ can be used to approximate $f(x)$

$$f(x) \approx P_N(x) = \sum_{k=0}^{N} \frac{f^{(k)}(x_0)}{k!}(x-x_0)^k$$

Examples:

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} \cdots$$

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} \cdots$$

Methods for evaluating a polinomial

$f(x) = 2x^4 + 4x^3 + 3x^2 + 2x + 1$

If we evaluate each term independently we will need:

$$f(x) = \underbrace{2\cdot x\cdot x\cdot x\cdot x}_{4\ \text{mul}} + \underbrace{4\cdot x\cdot x\cdot x}_{3\ \text{mul}} + \underbrace{3\cdot x\cdot x}_{2\ \text{mul}} + \underbrace{2\cdot x}_{1\ \text{mul}} + 1$$

10 multiplications + 4 sums

Horner's method, also called nested multiplication reduces the number of multiplications needed.

we factor x such that

$$f(x) = (((2\cdot x + 4)\cdot x + 3)\cdot x + 2)\cdot x + 1$$

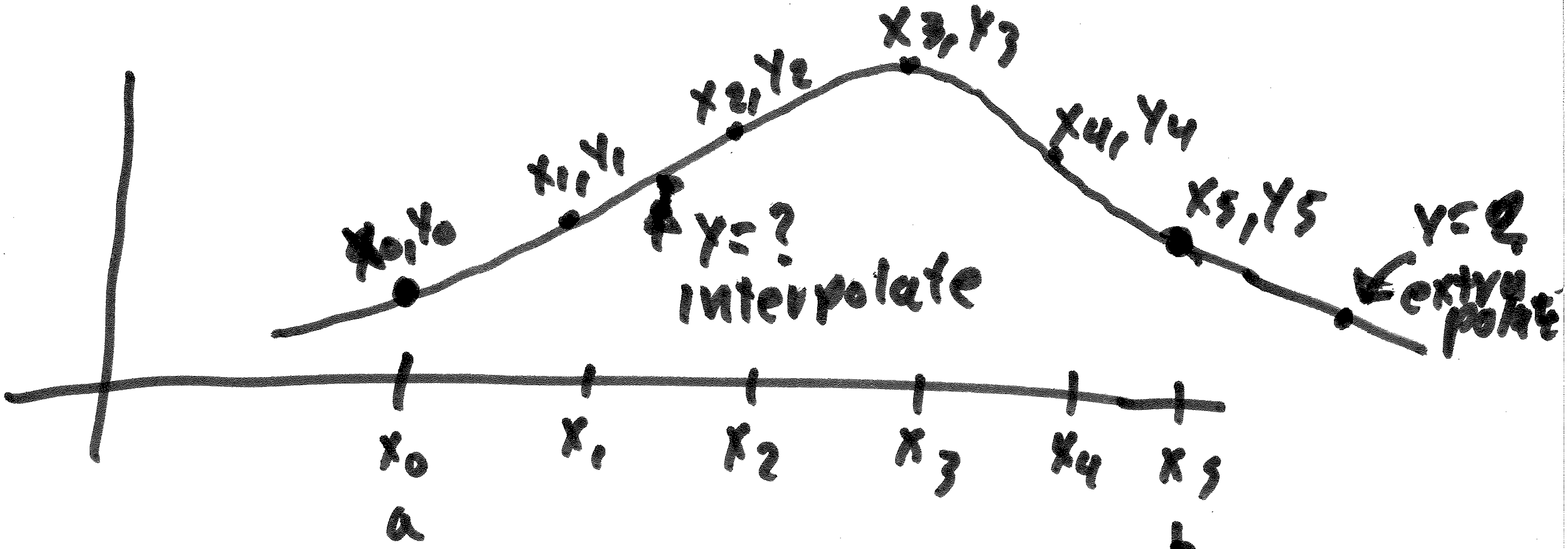
4 mul + 4 additions

Polynomial approximation

Suppose that a function $y=f(x)$ is known to have $N+1$ points $(x_0, y_0)\ldots(x_N, y_N)$ such that

$$a \leq x_0 < x_1 \ldots x_N < b \quad \text{and} \quad y_k = f(x_k)$$

then a polinomial $P(x)$ of degree N will be constructed that passes through these $N+1$ points



we want a polinomial that passes through these points

$$P(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4 + a_5 x^5$$

We want to build a polinomial either to:

interpolate:

Find a value ~~x~~ y from x assuming $x_0 < x < x_N$

extrapolate

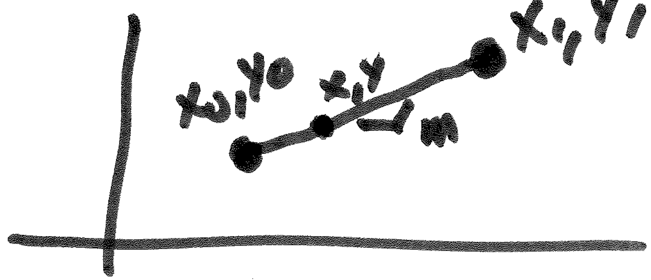
Find a value y from x assumming $x < x_0$ or $x > x_N$

We will see two methods that can be used to bold a polinomial $P_N(x)$ using $N+1$ points.

- Lagrange ~~method~~ approximation
- Newton polinomials

Lagrange approximation

Assume two points (X_0, Y_0), (X_1, Y_1)
the polynomial that passes through these two points is a straight line



$$m = \frac{Y - Y_0}{X - X_0} \quad (1)$$

$$m = \frac{Y_1 - Y_0}{X_1 - X_0} \quad (2)$$

From (1) and (2)

$$\frac{Y - Y_0}{X - X_0} = \frac{Y_1 - Y_0}{X_1 - X_0}$$

$$Y - Y_0 = \frac{Y_1 - X_0}{X_1 - X_0}(X - X_0)$$

$$Y = Y_0 + \frac{Y_1 - Y_0}{X_1 - X_0}(X - X_0)$$

equation of line that passes throug (X_0, Y_0). and (X_1, Y_1)

Lagrange uses the following approach.

$$P_1(x) = Y = Y_0\left[\frac{(x-x_1)}{(x_0-x_1)}\right] + Y_1\left[\frac{(x-x_0)}{(x_1-x_0)}\right]$$

$$\underbrace{\left[\frac{(x-x_1)}{(x_0-x_1)}\right]}_{A_0} \qquad \underbrace{\left[\frac{(x-x_0)}{(x_1-x_0)}\right]}_{A_1}$$

we want A_0 such that

$$A_0 = \begin{Bmatrix} 1 & x = x_0 \\ 0 & x = x_1 \end{Bmatrix} \qquad A_1 = \begin{Bmatrix} 1 & x = x_1 \\ 0 & x = x_0 \end{Bmatrix}$$

~~The~~

So we have that

$$P_1(x_0) = y_0\left[\cancelto{1}{\frac{x_0-x_1}{x_0-x_1}}\right] + y_1\left[\cancelto{0}{\frac{x_0-x_0}{x_1-x_0}}\right]$$

$$= Y_0$$

$$P_1(x_1) = y_0\left[\cancelto{0}{\frac{x_1-x_1}{x_0-x_1}}\right] + Y_1\left[\cancelto{1}{\frac{x_1-x_0}{x_1-x_0}}\right]$$

$$= Y_1$$

The terms

$$L_{1,0} = \frac{x - x_1}{x_0 - x_1} \quad \text{and} \quad L_{1,1} = \frac{x - x_0}{x_1 - x_0}$$

are called "Lagrange Polinomials".

For $P_2(x)$: we have points (x_0, y_0), (x_1, y_1) and (x_2, y_2) and we want to build a polinomial of degree 2 (quadratic).

ϕ when $x = x_1, x_2$
1 when $x = x_0$

$$P_2(x) = y_0 \left[\frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)}\right] + y_1 \left[\frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)}\right]$$

$$+ \; y_2 \left[\frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)}\right]$$

0 when $x = x_0, x_2$
1 when $x = x_1$

0 when $x = x_0, x_1$
1 when $x = x_2$

For $P_3(x)$ we have points (x_0,y_0), (x_1,y_1), (x_2,y_2), (x_3,y_3) and we want to build a polinomial of degree 3 (cubic) that passes through these points

$$P_3(x) = y_0\left[\frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)}\right] \begin{matrix} = 0, & x=x_1,x_2,x_3 \\ = 1, & x=x_0 \end{matrix}$$

$$+\ y_1\left[\frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)}\right] \begin{matrix} = 0, & x=x_0,x_2,x_3 \\ = 1, & x=x_1 \end{matrix}$$

$$+\ y_2\left[\frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)}\right] \begin{matrix} = 0, & x=x_0,x_1,x_3 \\ = 1, & x=x_2 \end{matrix}$$

$$+\ y_3\left[\frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)}\right] \begin{matrix} = 0, & x=x_0,x_1,x_2 \\ = 1, & x=x_3 \end{matrix}$$

In general

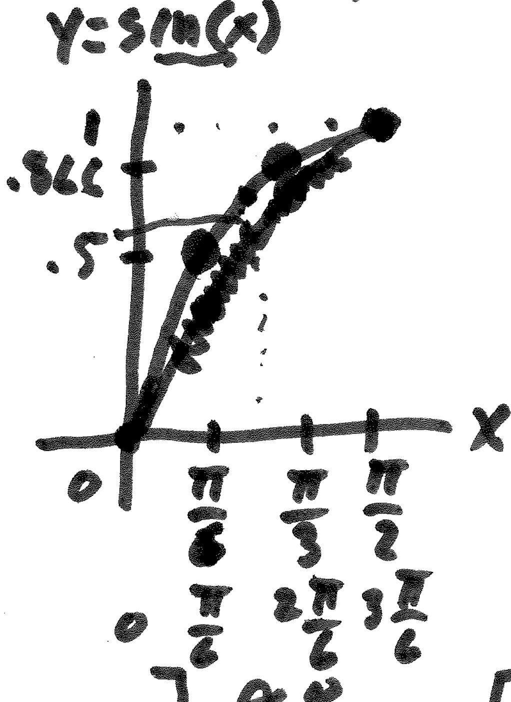
$$P_N(x) = \sum_{K=0}^{N} Y_K L_{N,K}(x)$$

where

$$L_{N,K}(x) = \frac{\prod_{\substack{J=0 \\ J\neq K}}^{N} (x - x_J)}{\prod_{\substack{J=0 \\ J=K}}^{N} (x_K - x_J)}$$

Example

We want to approximate sin(x) in the interval $0, \pi/2$ with a polynomial of degree 3 and 4 values.



i	X	$\sin(x)$
0	$x_0 = 0$	$y_0 = 0$
1	$x_1 = \frac{\pi}{6}$	$y_1 = .5$
2	$x_2 = \frac{\pi}{3}$	$y_2 = .866$
3	$x_3 = \frac{\pi}{2}$	$y_3 = 1$

$$P_3(x) = \cancel{0\left[\frac{(x-\frac{\pi}{6})(x-\frac{\pi}{3})(x-\frac{\pi}{2})}{(0-\frac{\pi}{6})(0-\frac{\pi}{3})(0-\frac{\pi}{2})}\right]} + .5\left[\frac{(x-0)(x-\frac{\pi}{3})(x-\frac{\pi}{2})}{(\frac{\pi}{6}-0)(\frac{\pi}{6}-\frac{\pi}{3})(\frac{\pi}{6}-\frac{\pi}{2})}\right]$$

$$+ .866\left[\frac{(x-0)(x-\frac{\pi}{6})(x-\frac{\pi}{2})}{(\frac{\pi}{3}-0)(\frac{\pi}{3}-\frac{\pi}{6})(\frac{\pi}{3}-\frac{\pi}{2})}\right] + 1\left[\frac{(x-0)(x-\frac{\pi}{6})(x-\frac{\pi}{3})}{(\frac{\pi}{2}-0)(\frac{\pi}{2}-\frac{\pi}{6})(\frac{\pi}{2}-\frac{\pi}{3})}\right]$$