Chapter-3: Solution of Nonlinear Equations

-We want to find x such that f(x)=0.

Example: *f(x)=x-tan(x)=0*

Example: *x-a*sin(x)=b*

-There may be many approximate solutions even though the exact solution is unique. (Because of roundoff errors.)

Example: $P_4(x)=x^4-4x^3+6x^2-4x+1=(x-1)^4$

-If you use marc-32, you will find many zeros in the interval [0.975, 1.035]

3-1 Bisection (Interval Halving) Method

-If f(x) is a continuous function on the interval [a,b] and if f(a)f(b)<0, then f(x) must have a zero in (a,b).

Bisection Method:

1- Compute c=0.5*(a+b)

2- If f(a)f(c)<0 then f(x) has a zero in [a,c] $\Rightarrow$ $b \leftarrow c$ (assign c to b)

3- Else $\Rightarrow$ $a \leftarrow c$ (assign c to a $f(x)$ has a zero in $[c,b]$)

4- Stop if *f(c)=0*. (*c* is the zero of *f(x)*)

5- Go to step 1.

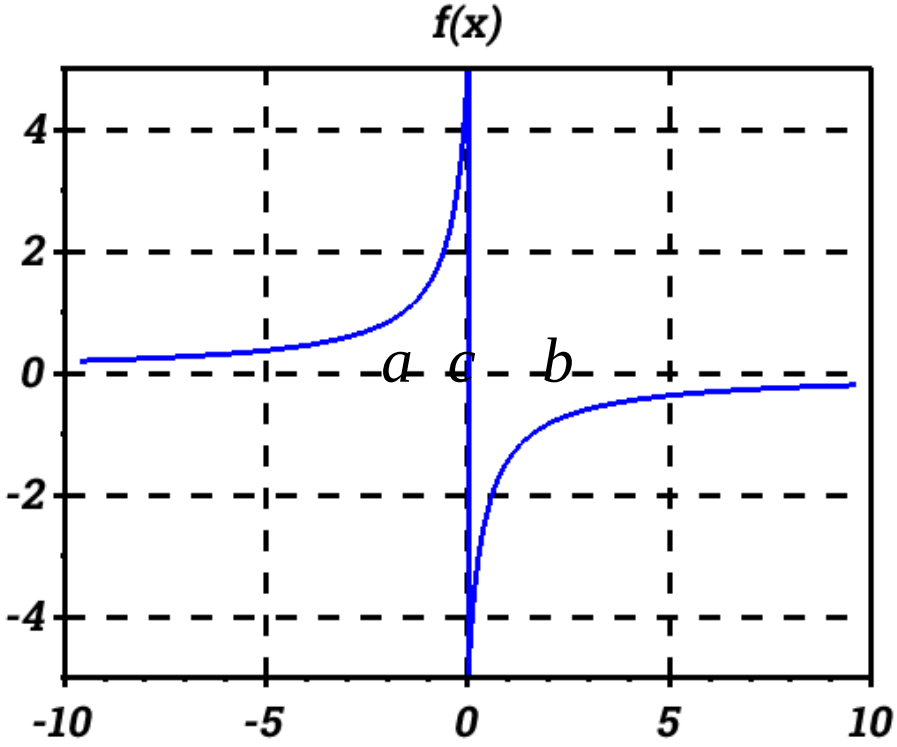
-It is quite unlikely that f(c) will be exactly 0 in the computer because of roundoff errors.

Stopping Criteria (stop when one of them is satisfied)

1-The maximum number of steps (M)

2- $|b-a|<\delta$ δ is a positive number.

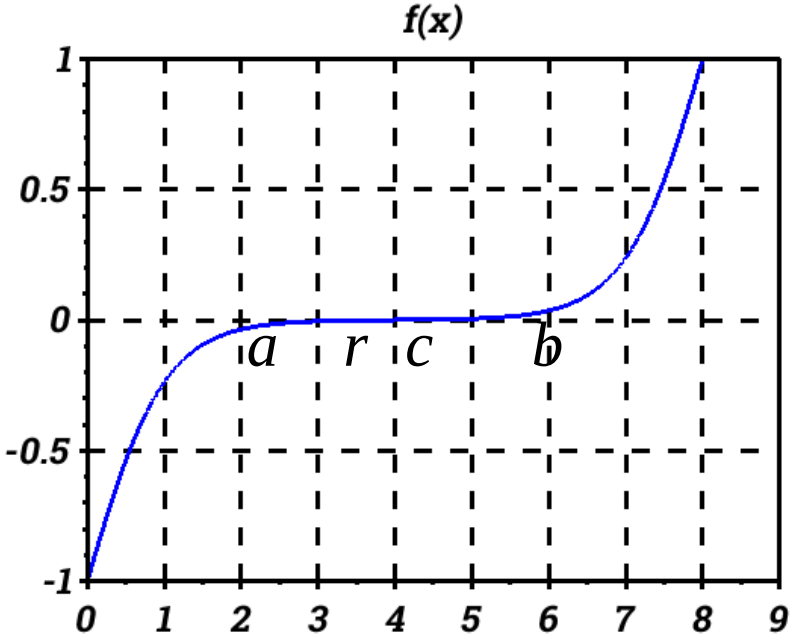
3- $|f(c)|<\epsilon$ ϵ is a positive number.



$|b-a|<\delta$ but $|f(c)|>\epsilon$

Criterion $|f(c)|<\epsilon$ failure

The function is not continuous but we can not verify in advance.



$|f(c)|<\epsilon$ but $|b-a|>\delta$

Criterion $|b-a|<\delta$ failure

-Error Analysis

-Let us denote the successive intervals that arise in the process by $[a_0, b_0]$, $[a_1, b_1]$ and so on.

$$\Rightarrow a_0 \le a_1 \le a_2 \le a_3 \le \ldots a_n \le b_0, \qquad b_0 \ge b_1 \ge b_2 \ge b_3 \ge \ldots \ge b_n \ge \ldots \ge a_0$$

$$\text{and } (b_{n+1} - a_{n+1}) = \frac{1}{2}(b_n - a_n) \qquad n \ge 0$$

-Since the sequence a_n is increasing and bounded above, it converges. Likewise, b_n converges.

$$\Rightarrow (b_n - a_n) = 2^{-n}(b_0 - a_0)$$

$$\Rightarrow \lim_{n \to \infty} (b_n - a_n) = \lim_{n \to \infty} 2^{-n}(b_0 - a_0) = 0$$

$$\Rightarrow r = \lim_{n \to \infty} a_n = \lim_{n \to \infty} b_n$$

$$f(a_n) f(b_n) \le 0 \quad \Rightarrow \lim_{n \to \infty} f(a_n) f(b_n) = (f(r))^2 \le 0 \quad \Rightarrow f(r) = 0$$

-The best estimate at the nth stage is not a_n or b_n but the midpoint of the interval $c_n = (a_n + b_n)/2$

Example: $2*x-\tan(x)$ in the interval $[0.5,1.5]$ Find the root.

Solution:

$a_0=0.5, \quad b_0=1.5 \quad \Rightarrow c_0=1, \quad f(a_0)=0.45, \quad f(b_0)=-11.1, \quad f(c_0)=0.44$

$a_1=1, \quad b_1=1.5 \quad \Rightarrow c_1=1.25, \quad f(a_1)=0.44, \quad f(b_1)=-11.1, \quad f(c_1)=-0.51$

$a_2=1, \quad b_2=1.25 \quad \Rightarrow c_2=1.125, \quad f(a_2)=0.44, \quad f(b_2)=-0.51, \quad f(c_2)=0.16$

$a_3=1.125, \quad b_3=1.25 \quad \Rightarrow c_3=1.1875, \quad f(a_3)=0.16, \quad f(b_3)=-0.51, \quad f(c_3)=-0.1$

$a_4=1.125, \quad b_4=1.1875 \quad \Rightarrow c_4=1.15625, \quad f(a_4)=0.16, \quad f(b_4)=-0.1, \quad f(c_4)=0.04$

$a_5=1.15625, \quad b_5=1.1875 \quad \Rightarrow c_5=1.171875, \quad f(a_5)=0.04, \quad f(b_5)=-0.1, \quad f(c_5)=-0.02$

$a_6=1.15625, \quad b_6=1.171875 \quad \Rightarrow c_6=1.1640625, \quad f(a_6)=0.04, \quad f(b_6)=-0.02, \quad f(c_6)=0.0066118$

-Example: $x^{-1}-2^{x}$ in the interval $[0,1]$

Solution:

$a_0=0,\quad b_0=1 \Rightarrow c_0=0.5,\quad f(a_0)=\infty,\quad f(b_0)=-1,\quad f(c_0)=0.58$

$a_1=0.5,\quad b_1=1 \Rightarrow c_1=0.75,\quad f(a_1)=0.58,\quad f(b_1)=-1,\quad f(c_1)=-0.35$

$a_2=0.5,\quad b_2=0.75 \Rightarrow c_2=0.625,\quad f(a_2)=0.58,\quad f(b_2)=-0.35,\quad f(c_2)=0.058$

$a_3=0.625,\quad b_3=0.75 \Rightarrow c_3=0.6875,\quad f(a_3)=0.058,\quad f(b_3)=-0.35,\quad f(c_3)=0.156$

$a_4=0.625,\quad b_4=0.6875 \Rightarrow c_4=0.65625,\quad f(a_4)=0.058,\quad f(b_4)=0.156,\quad f(c_4)=-0.0522$

$a_5=0.625,\quad b_5=0.65625 \Rightarrow c_5=0.640625,\quad f(a_5)=0.058,\quad f(b_5)=-0.0522,\quad f(c_5)=0.00197$

$a_6=0.640625,\quad b_6=0.65625 \Rightarrow c_6=0.6484375,\quad f(a_6)=0.00197,\quad f(b_6)=-0.0522,\quad f(c_6)=-0.025$

$a_7=0.640625,\quad b_7=0.6484375 \Rightarrow c_7=0.6445313,\quad f(a_7)=0.00197,\quad f(b_7)=-0.025,\quad f(c_7)=-0.0117$

3.2 Newton's Method

Newton's method is a general procedure that can be applied in many diverse situations. When applied to the problem of locating a zero of a real-valued function of a real variable, it is often called the Newton-Raphson iteration. It is faster than the bisection and the secant method (convergence is quadratic). Unfortunately, the method is not guaranteed always to converge. Therefore, it is combined with other slower methods.

-Let *r* be a zero of *f(x)* and let x be an approximation to *r.* If $f^{(2)}(x)$ exists and continuous, then by Taylor's Theorem,

$$0=f(r)=f(x)+hf'(x)+O(h^2)$$

where *h=r-x*. If *h* is small, we can ignore $O(h^2)$ $\Rightarrow$ $h\simeq-\frac{f(x)}{f'(x)}$

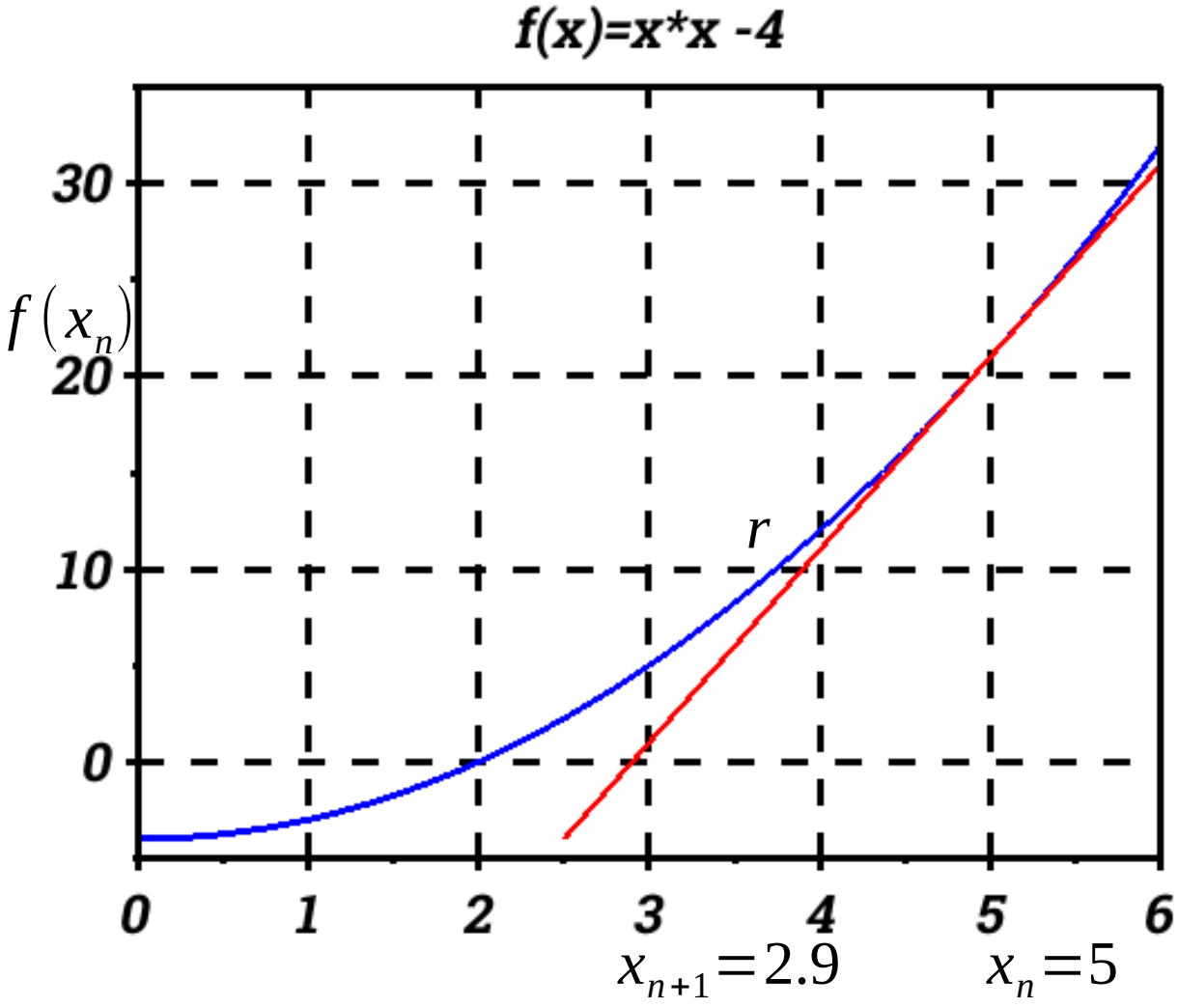
If *x* is approximation to *r*, then $x-\frac{f(x)}{f'(x)}$ should be a better approximation to *r*.

Newton's Method:

-Begin with an estimate x_0 of r and compute

$$x_{n+1}=x_n-\frac{f(x_n)}{f'(x_n)} \qquad n\geq 0$$

Graphical Interpretation



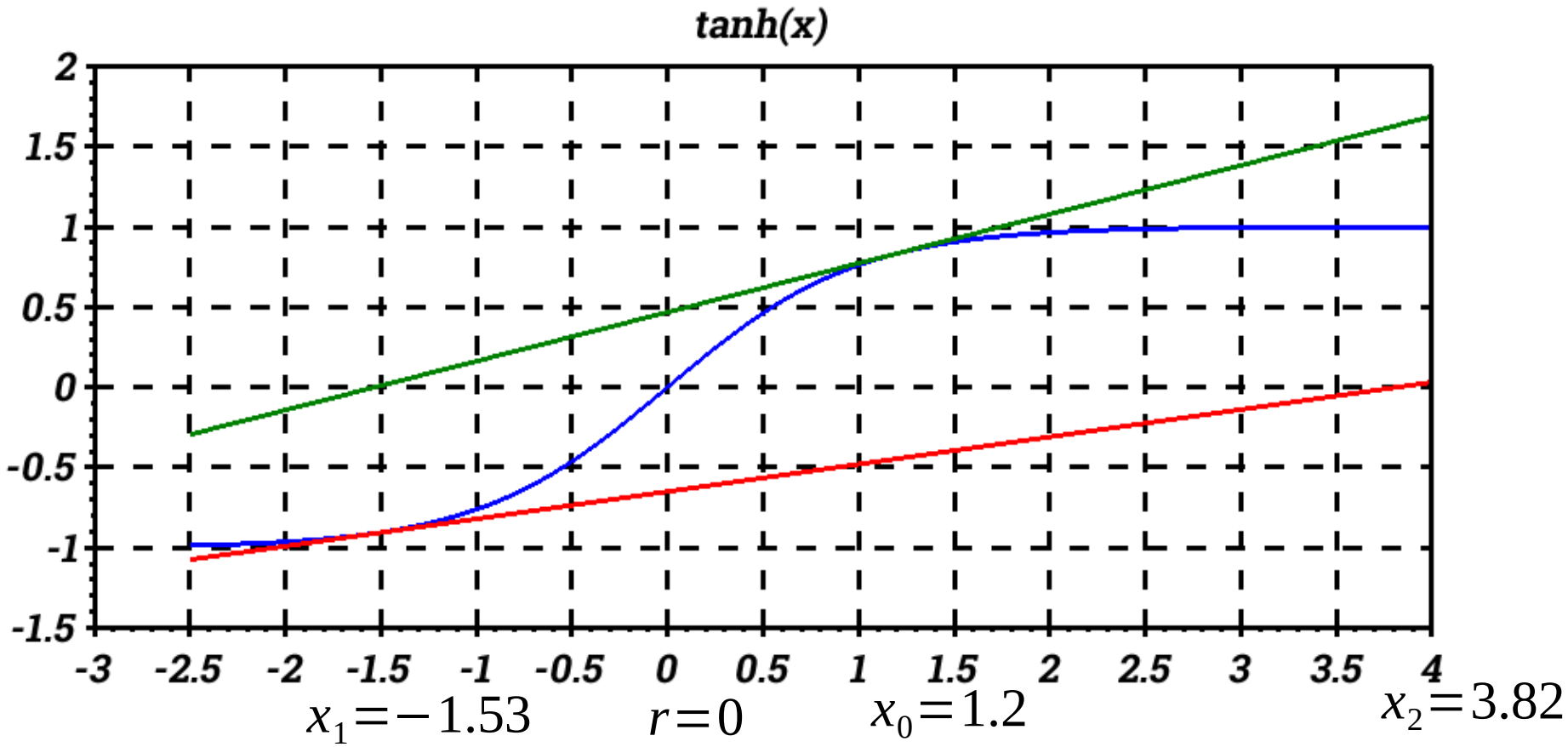
$l(x) = -29 + 10x$

$l(x)$ is the linearization of the function $f(x)$.
$l(x)$ is a good approximation of $f(x)$ if x close to c

$f(x) = f(c) + f'(c)(x-c) + 0.5f^{(2)}(c)(x-c)^2 + (1/6)f^{(3)}(c)(x-c)^3 + \ldots$ Taylor series ex.

$l(x) = f(c) + f'(c)(x-c)$ First Order Taylor series expansion.

$\Rightarrow l(x) = f(x_n) + f'(x_n)(x-x_n)$ $c = x_n$



Example of nonconvergence of Newton's method.

-For convergence of Newton's method, x_0 must be sufficiently close to zero, or $f(x)$ must have a prescribed shape (convex and increasing)

Error Analysis

Assume $f^{(2)}$ is continuous, $f(r)=0\neq f'(r)$

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

By Taylor's Theorem, we have

$$0=f(r)=f(x_n-e_n)=f(x_n)-e_n f'(x_n)+\frac{1}{2}e_n^2 f^{(2)}(\xi_n)$$

where ξ_n is a number between x_n and r.

$$\Rightarrow e_n f'(x_n)-f(x_n)=\frac{1}{2}f^{(2)}(\xi_n)e_n^2$$

$$\Rightarrow e_{n+1}=\frac{1}{2}\frac{f^{(2)}(\xi_n)}{f'(x_n)}e_n^2\simeq\frac{1}{2}\frac{f^{(2)}(r)}{f'(r)}e_n^2=Ce_n^2$$

$\Rightarrow$ The rate of convergence is quadratic. Does e_n converge to 0?

Choose δ small enough so that $\delta C(\delta)<1$. Having fixed δ, set $\rho=\delta C(\delta)$

Suppose we start iteration with a point x_0 satisfying $|x_0-r|\le\delta$.

Then $|e_0|\le\delta$ and $|\xi_0-r|\le\delta$

$$\Rightarrow|x_1-r|=|e_1|\le e_0^2C(\delta)=|e_0||e_0|C(\delta)\le|e_0|\delta C(\delta)=|e_0|\rho<|e_0|\le\delta$$

-If we repeat the argument

$$|e_1|\le\rho|e_0|$$
$$|e_2|\le\rho|e_1|\le\rho^2|e_0|$$
$$|e_3|\le\rho|e_2|\le\rho^3|e_0| \quad\Rightarrow\quad |e_n|\le\rho^n|e_0| \;\to 0.$$

-Theorem: Let $f^{(2)}(x)$ be continuous function and let *r* be a simple zero of *f(x)*. Then, there is a neighborhood of *r* and a constant *C* such that if Newton's method is started in that neighborhood, the successive points become steadily closer to *r* and satisfy

$$|x_{n+1}-r|\le C(x_n-r)^2 \qquad (n\ge 0)$$

-Theorem: If *f(x)* belongs to $C^2(R)$, is increasing, is convex, and has a zero, then the zero is unique, and the Newton's iteration will converge to it from any starting point.

Proof:

$f(x)$ is convex $\Rightarrow$ $f^{(2)}(x)>0$, increasing $\Rightarrow$ $f'(x)>0$.

$$\Rightarrow e_{n+1}=\frac{1}{2}\frac{f^{(2)}(\xi_n)}{f'(x_n)}e_n^2>0$$

$$\Rightarrow \quad x_n>r \text{ for } n\geq 1 \qquad (x_n=e_n+r)$$

Since $f(x)$ is increasing, $f(x_n)>f(r)=0$ $\qquad (n\geq 1)$

$$e_{n+1}=e_n-\frac{f(x_n)}{f'(x_n)} \quad \Rightarrow e_{n+1}<e_n \quad \Rightarrow \quad e_n \text{ is decreasing and bounded below by zero.}$$

$$\Rightarrow \lim_{n\to\infty} e_n=0 \quad \Rightarrow \quad \lim_{n\to\infty} x_n=r$$

Example: Find an efficient method for computing square roots based on the use of Newton's method.

Solution: Let $x=\sqrt{R}$ $\Rightarrow$ $x^2-R=0$

Use Newton's method on the function $f(x)=x^2-R$ $\Rightarrow$ $f'(x)=2x$.

$\Rightarrow$ the iteration formulea : $x_{n+1}=x_n-\frac{f(x_n)}{f'(x_n)}=x_n-\frac{x_n^2-R}{2x_n}=\frac{1}{2}\left(x_n+\frac{R}{x_n}\right)$

Lets compute $\sqrt{17}$ using the Newton's method. Let $x_0 = 4$.

$\Rightarrow \quad x_1 = 4.12, \quad x_2 = 4.123106, \quad x_3 = 4.1231056256177,$
$x_4 = 4.123105625617660549821409856,$

Implicit Functions

$G(x, y) = 0$ Find y values for the given x values.

$$\Rightarrow \quad y_{k+1} = y_k - \frac{G(x, y_k)}{\frac{\partial G}{\partial y}(x, y_k)}$$

-This method can be used to construct a table of the function *y(x)*

Example: Produce a table of *x* versus *y* for the function
$G(x, y) = 3x^7 + 2y^5 - x^3 + y^3 - 3 = 0$

Solution:

$$\frac{\partial G}{\partial y}(x, y) = 10y^4 + 3y^2$$

$$\Rightarrow y_{n+1} = y_n - \frac{3x^7 + 2y^5 - x^3 + y^3 - 3}{10y^4 + 3y^2}$$

i	x	y
0	0.0	1.000000
1	0.1	1.000077
2	0.2	1.000612
⋮	⋮	⋮
20	2.0	−2.810639
⋮	⋮	⋮
80	8.0	−19.92635
⋮	⋮	⋮
100	10.0	−27.23685

Systems of Nonlinear Equations

Newton's method for systems of nonlinear equations follows the same strategy that was used for a single equation.

Let $f_1(x_1, x_2) = 0$

$f_2(x_1, x_2) = 0$ find x_1 and x_2

Suppose that (x_1, x_2) is an approximate solution of previous equations. Let us compute h_1 and h_2 so that $(x_1 + h_1, x_2 + h_2)$ will be a better approximate solution.

-Using only linear terms in the Taylor expansion, we get

$$0 = f_1(x_1 + h_1, x_2 + h_2) \simeq f_1(x_{1,}, x_2) + h_1 \frac{\partial f_1}{\partial x_1} + h_2 \frac{\partial f_1}{\partial x_2}$$

$$0 = f_2(x_1 + h_1, x_2 + h_2) \simeq f_2(x_{1,}, x_2) + h_1 \frac{\partial f_2}{\partial x_1} + h_2 \frac{\partial f_2}{\partial x_2}$$

$$\Rightarrow \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} = - \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} \end{bmatrix}^{-1} \begin{bmatrix} f_1(x_1, x_2) \\ f_2(x_1, x_2) \end{bmatrix}, \qquad J = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} \end{bmatrix}$$

J is the Jacobian matrix of f_1 and f_2.

Newton's method fo two nonlinear equations in two variables is

$$\begin{bmatrix} x_1^{k+1} \\ x_2^{k+1} \end{bmatrix} = \begin{bmatrix} x_1^k \\ x_2^k \end{bmatrix} + \begin{bmatrix} h_1^k \\ h_2^k \end{bmatrix} \quad \text{where} \begin{bmatrix} h_1^k \\ h_2^k \end{bmatrix} = - J^{-1} \begin{bmatrix} f_1(x_1^k, x_2^k) \\ f_2(x_1^k, x_2^k) \end{bmatrix}, \quad J^{-1} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x_1^k, x_2^k) & \frac{\partial f_1}{\partial x_2}(x_1^k, x_2^k) \\ \frac{\partial f_2}{\partial x_1}(x_1^k, x_2^k) & \frac{\partial f_2}{\partial x_2}(x_1^k, x_2^k) \end{bmatrix}^{-1}$$

-The system with more than two variables:

$f_i(x_1, x_2, x_3, \ldots, x_n) = 0 \qquad (1 \le i \le n)$

Let $X = (x_1, x_2, x_3, \ldots, x_n)^T, \qquad F = (f_1, f_2, f_3, \ldots, f_n)^T$

Taylor expansion of F:

$0 = F(X+H) \simeq F(X) + F'(X)H, \qquad H = (h_1, h_2, h_3, \ldots, h_n)^T$

$F'(X)$ is the nxn Jacobian matrix $J(X)$ with elements $\frac{\partial f_i}{\partial x_j}$; namely,

$$F'(X) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix}$$

$\Rightarrow H = -F'(X)^{-1} F(X)$

$X^{(k+1)} = X^{(k)} + H^{(k)}$

Example:

-Starting with $(1,1,1)^T$ carry out six iteration of Newton's method for finding of the root of the nonlinear system:

$$xy = z^2 + 1$$
$$xyz + y^2 = x^2 + 2$$
$$e^x + z = e^y + 3$$

$$x_1 = x, \quad x_2 = y, \quad x_3 = z$$

$$F(X) = \begin{bmatrix} x_1 x_2 - x_3^2 - 1 \\ x_1 x_2 x_3 - x_1^2 + x_2^2 - 2 \\ e^{x_1} - e^{x_2} + x_3 - 3 \end{bmatrix}$$

Solution:

$$\Rightarrow F'(X) = \begin{bmatrix} x_2 & x_1 & -2x_3 \\ x_2 x_3 - 2x_1 & x_1 x_3 + 2x_2 & x_1 x_2 \\ e^{x_1} & -e^{x_2} & 1 \end{bmatrix}$$

n	x_1	x_2	x_3
1	1.7547911	1.5363988	1.1455950
⋮	⋮	⋮	⋮
6	1.7776717	1.4239606	1.2374710

3.3 Secant Method

Newton's method : $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$ (uses derivative. Need to compute derivative)

Steffensen's iteration: $x_{n+1} = x_n - \frac{f^2(x_n)}{f(x_n + f(x_n)) - f(x_n)}$

(Need to compute $f(x_n)$ and $f(x_n + f(x_n))$)

Secant method: Replaces $f'(x_n)$ with $f'(x_n) \simeq \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}$

$$\Rightarrow x_{n+1} = x_n - f(x_n)\left[\frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}\right]$$

To compute x_{n+1}, we need x_n and x_{n-1}. However evaluation of each new x_{n+1} requires only one evaluation of f.

Error Analysis

$$e_n = x_n - r$$

$$e_{n+1} = x_{n+1} - r = \frac{f(x_n)x_{n-1} - f(x_{n-1})x_n}{f(x_n) - f(x_{n-1})} - r = \frac{f(x_n)e_{n-1} - f(x_{n-1})e_n}{f(x_n) - f(x_{n-1})}$$

$$= \left[\frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}\right]\left[\frac{f(x_n)/e_n - f(x_{n-1})/e_{n-1}}{x_n - x_{n-1}}\right]e_n e_{n-1}$$

By Taylor's Theorem,

$$f(x_n) = f(r + e_n) = f(r) + e_n f'(r) + \frac{1}{2}e_n^2 f^{(2)}(r) + O(e_n^3), \qquad f(r) = 0.$$

$$\Rightarrow \frac{f(x_n)}{e_n} = f'(r) + \frac{1}{2}e_n f^{(2)}(r) + O(e_n^2)$$ Changing the index to $n-1$ yields

$$\frac{f(x_{n-1})}{e_{n-1}} = f'(r) + \frac{1}{2}e_{n-1} f^{(2)}(r) + O(e_{n-1}^2)$$

$$\Rightarrow f(x_n)/e_n - f(x_{n-1})/e_{n-1} = \frac{1}{2}(e_n - e_{n-1})f^{(2)}(r) + O(e_n^2) - O(e_{n-1}^2), \qquad x_n - x_{n-1} = e_n - e_{n-1}$$

$$\Rightarrow \left[\frac{f(x_n)/e_n - f(x_{n-1})/e_{n-1}}{x_n - x_{n-1}}\right] \simeq \frac{1}{2}f^{(2)}(r) \qquad \Rightarrow e_{n+1} \simeq \frac{1}{2}\frac{f^{(2)}(r)}{f'(r)}e_n e_{n-1} = C e_n e_{n-1}$$

What is the order of convergence?

Assume $|e_{n+1}| \sim A|e_n|^{\alpha}$ where A is a positive constant

This means taht the ratio $|e_{n+1}|/(A|e_n|^n)$ tends to 1 as $n \to \infty$ and implies $\alpha-$order convergence.

$\Rightarrow |e_n| \sim A|e_{n-1}|^{\alpha}$ and $|e_{n-1}| \sim (A^{-1}|e_n|)^{1/\alpha}$

$\Rightarrow |e_{n+1}| = A|e_n|^{\alpha} \sim |C||e_n| A^{-1/\alpha} |e_n|^{1/\alpha}$ $\Rightarrow A^{1+\frac{1}{\alpha}} |C|^{-1} \sim |e_n|^{1-\alpha+\frac{1}{\alpha}} \to 1$

$\Rightarrow A^{1+\frac{1}{\alpha}} |C|^{-1}$ is a nonzero constant while $e_n \to 0$ as $n \to \infty$.

$\Rightarrow 1-\alpha+\frac{1}{\alpha} = 0$ $\Rightarrow \alpha = (1+\sqrt{5})/2 \simeq 1.62$ $\Rightarrow$ the rate of convergence is superlinear.

$$A = |C|^{1/(1+\frac{1}{\alpha})} = |C|^{1/\alpha} = |C|^{\alpha-1} = |C|^{0.62} = \left| \frac{f^{(2)}(r)}{2f'(r)} \right|^{0.62}$$

$\Rightarrow |e_{n+1}| \simeq A|e_n|^{(1+\sqrt{5})/2}$.

-The Secant method converges slower than the Newton's method but Secant method requires evaluation of only *f(x)*. Newton's method requires evaluations of *f(x)* and *f'(x)* at every step.

3.5 Computing Zeros of Polynomials

$P(z)=a_n z^n+a_{n-1} z^{n-1}+\ldots+a_2 z^2+a_1 z+a_0$

a_k and the variable z may be comlex number or variables. If $a_n \neq 0$ then P has degree n. Find the zeros of P.

Theorem 1: (Fundamental Theorem of Algebra)

-Every non-constant polynomial has at least one zero in the complex field.

Let P is a nth degree polynomial.

$\Rightarrow P(z)=(z-c)q(z)+r$ where r is a complex number and q is a polynomial of degree $n-1$.

$P(c)=r$. If $r=0$ then c is a zero of P and we have $P(z)=(z-c)q(z)$.

Thus $(z-c)$ is a factor of $P(z)$.

Theorem : A polynomial of degree n has exactly n zeros in the complex plane (consequence of Fundamental Theorem of Algebra)

$P(z)=(z-r_1)(z-r_2)\ldots(z-r_n)q_n$. where q_n is a constant.

Theorem: All zeros of the *nth* degree polynomial lie in the open desk whose center is at the origin of the complex plane and whose radius is

$$\rho = 1 + |a_n|^{-1} \max_{0 \le k < n} |a_k|$$

Proof:

Let $c = \max_{0 \le k < n} |a_k|$ so that $c|a_n|^{-1} = \rho - 1$. If $c = 0$ our result is trivially true.

Hence, assume $c > 0$. Then $\rho > 1$. If $|z| \ge \rho$ If $d = e - f \Rightarrow |d| \le |e| + |-f| = |e| + |f|$

$$|P(z)| \ge |a_n z^n| - |a_{n-1} z^{n-1} + \ldots + a_1 z + a_0| \ge |a_n z^n| - c \sum_{k=0}^{n-1} |z|^k > |a_n z^n| - c|z^n|(|z| - 1)^{-1}$$

$$= |a_n z^n| \left\{ 1 - c|a_n|^{-1} (|z| - 1)^{-1} \right\} \ge |a_n z^n| \left\{ 1 - c|a_n|^{-1} (\rho - 1)^{-1} \right\} = 0$$

$$\Rightarrow \text{If } |z| \ge \rho \quad |P(z)| > 0 \quad (\text{ there is no zero if } |z| \ge \rho)$$

$$\sum_{k=0}^{n-1} a^k = \frac{1 - a^n}{1 - a}$$

Example: Find a disk centered at the origin that contains all the zeros of the polynomial $P(z) = z^4 - 4z^3 + 7z^2 - 5z - 2$

Solution:

$$\rho = 1 + |a_4|^{-1} \max_{0 \le k < 4} |a_k| = 8 \quad \Rightarrow |z_i| < 8 \quad \text{where } |z_i| \text{'s are the zeros of } P(z)$$

Let $S(z)=z^n P(1/z)=z^n\left[a_n z^{-n}+a_{n-1}z^{-(n-1)}+\ldots+a_2 z^{-2}+a_1 z^{-1}+a_0\right]$

$=a_n+a_{n-1}z+a_{n-2}z^2+\ldots+a_2 z^{n-2}+a_1 z^{n-1}+a_0 z^n$

The condition $P(z_0)=0$ is equivalent to the condition $S(1/z_0)=0$

Theorem: If all the zeros of $S(z)$ are in the disk $\{z:|z|\le\rho\}$, then all the nonzero zeros of $P(z)$ are outside the disk $\{z:|z|\ge\rho^{-1}\}$.

Example: $P(z)=z^4-4z^3+7z^2-5z-2$ Find a ring that all zeros of P(z) lie in it.

Solution:

$$\rho_p=1+|a_4|^{-1}\max_{0\le k<4}|a_k|=8$$

$$S(z)=1-4z+7z^2-5z^3-2z^4$$

$$\Rightarrow\rho_s=1+|a_4|^{-1}\max_{0\le k<4}|a_k|=1+\frac{7}{2}=\frac{9}{2}$$

$$\Rightarrow\frac{1}{\rho_s}<|z_i|<\rho_p. \quad\Rightarrow\quad \frac{2}{9}<|z_i|<8$$

where $|z_i|'$s are zeros of $P(z)$.

Horner's Algorithm

If a polynomial p and a complex number z_0 are given, Horner's algorithm will produce the polynomial $q(z)=\dfrac{p(z)-p(z_0)}{z-z_o}$.

The degree of polynomial *q* is one less than the degree of *p*.

$p(z)=(z-z_0)q(z)+p(z_0)$

Let q be represented by $q(z)=b_0+b_1 z+\ldots+b_{n-1}z^{n-1}$

$\Rightarrow b_{n-1}=a_n$

$b_{n-2}=a_{n-1}+z_0 b_{n-1}$

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

$b_0=a_1+z_0 b_1$

$p(z_0)=a_0+z_0 b_0$

	a_n	a_{n-1}	a_{n-2}	a_{n-3}	$\ldots$	a_1	a_0
z_0		$z_o b_{n-1}$	$z_0 b_{n-2}$	$z_0 b_{n-3}$	$\ldots$	$z_0 b_1$	$z_0 b_0$
	b_{n-1}	b_{n-2}	b_{n-3}	b_{n-4}	$\ldots$	b_0	b_{-1}

$b_{-1}=p(z_0)$

Example: $P(z)=z^4-4z^3+7z^2-5z-2.$ Evaluate $p(3)$ using Horner's algorithm.

	1	-4	7	-5	-2
3		3	-3	12	21
	1	-1	4	7	19

$\Rightarrow p(3)=19$

$\Rightarrow p(z)=(z-3)(z^3-z^2+4z+7)+19$

-Horner's algorithm is also used for deflation (factorization)

If z_0 is zero of p then $p(z)=(z-z_0)q(z)$

Example: Deflate the polynomial $P(z)=z^4-4z^3+7z^2-5z-2$ using the fact that 2 is one of its zero.

Solution:

	1	-4	7	-5	-2
2		2	-4	6	2
	1	-2	3	1	0

$\Rightarrow z^4-4z^3+7z^2-5z-2=(z-2)(z^3-2z^2+3z+1)$

-Horner’s algorithm can be used in finding the Taylor expansion of a polynomial about any point.

$$P(z)=a_n z^n+a_{n-1} z^{n-1}+\ldots+a_2 z^2+a_1 z+a_0=c_n(z-z_0)^n+c_{n-1}(z-z_0)^{n-1}+\ldots+c_0$$

$$\Rightarrow c_k=\frac{p^{(k)}(z_0)}{k!}$$

-If we apply the Horner's algorithm to the p with z_0

$$q(z)=\frac{p(z)-p(z_0)}{z-z_0}=c_n(z-z_0)^{n-1}+c_{n-1}(z-z_0)^{n-2}+\quad\ldots\quad+c_1$$

$$\Rightarrow c_1=q(z_0)$$

This process is repeated until all coefficients c_k are found.

Example:

Find the Taylor expansion of $p(z)=z^4-4z^3+7z^2-5z-2$ about the point $z_0=3$.

Solution:

	1	−4	7	−5	−2
3		3	−3	12	21
	1	−1	4	7	19
3		3	6	30	
	1	2	10	37	
3		3	15		
	1	5	25		
3		3			
	1	8			

$$\Rightarrow p(z)=(z-3)^4+8(z-3)^3+25(z-3)^2+37(z-3)+19$$

-Newton's iteration can be carried out on a polynomial.

$$z_{k+1}=z_k-\frac{f(z_k)}{f'(z_k)} \qquad \text{We know } c_0=p(z_0),\quad c_1=p'(z_0)$$

Example:

$p(z)=z^4-4z^3+7z^2-5z-2$ Start with $z_0=0$ and find a zero of the polynomial

	1	−4	7	−5	−2
0		0	0	0	0
	1	−4	7	−5	−2
0		0	0	0	
	1	−4	7	−5	

$$\Rightarrow c_0=-2=p(z_0) \qquad c_1=-5=p'(z_0)$$

$$\Rightarrow z_1=z_0-\frac{p(z_0)}{p'(z_0)}=0-\frac{-2}{-5}=-0.4$$

-If we Execute the algorithm on a computer, we get

k	$p(z_k)$	$p'(z_k)$	z_k
1	-2	-5	-0.4
$\vdots$	$\vdots$	$\vdots$	$\vdots$
5	0.00000	-9.85537	-0.27568

Theorem: Let z_k and z_{k+1} be two successive iterates when Newton's method is applied to a polynomial p of degree n. Then there is a zero of p within distance $n|z_k - z_{k+1}|$ of z_k in the complex plane.

Proof: Let $r_1, r_2, r_{3,}\ldots, r_n$ be the zeros of p then $p(z) = c\prod_{j=1}^{n}(z - r_j)$

$$p'(z) = c\sum_{k=1}^{n}\prod_{\substack{i=1\\ i\neq k}}^{n}(z - r_i) = \sum_{k=1}^{n}\frac{p(z)}{z - r_k} = p(z)\sum_{k=1}^{n}(z - r_k)^{-1}, \qquad z_{n+1} = z_n - \frac{p(z_n)}{p'(z_n)}$$

We want to show that for any $z(z_k)$ there is an index j for which $|z - r_j| \leq n|p(z)/p'(z)|$. If no index j satisfies the desired inequality, then for all j $|z - r_j| > n|p(z)/p'(z)|$.

$$|z - r_j|^{-1} < \frac{1}{n}|p'(z)/p(z)| = \frac{1}{n}|\sum_{k=1}^{n}(z - r_k)^{-1}| \leq \frac{1}{n}\sum_{k=1}^{n}|z - r_k|^{-1}$$

But this is not possible because average of n numbers cannot be greater than each of them.

-Laguerre Iteration

Converge faster than Newton' method (third-order convergence

$$A = -p'(z)/p(z)$$
$$B = A^2 - p^{(2)}(z)/p(z)$$
$$C = n^{-1}\left[A \pm \sqrt{(n-1)(nB - A^2)}\right]$$
$$z_{new} = z + 1/C$$

Theorem: İf *p* is a polynomial of degree *n*, if *z* is any complex number, and *C* is computed as in Laguerre's algorithm, then *p* has a zero in the complex plane within distance $\sqrt{n}/C$ of z.

CHAPTER-4 Solving Systems of Linear Equations

-We want to solve the systems of linear equations having the form

$$\begin{array}{l} a_{11}x_1+a_{12}x_2+a_{13}x_3+\ \dots\ +a_{1n}x_n=b_1 \\ \quad\vdots\quad\ \vdots\quad\ \vdots\quad\ \vdots\quad\ \vdots\quad\ \vdots \\ a_{n1}x_1+a_{n2}x_2+a_{n3}x_3+\ \dots\ +a_{nn}x_n=b_n \end{array}$$

in matrix form:

$$\underbrace{\begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}}_{X} = \underbrace{\begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}}_{b} \quad \Rightarrow AX=b$$

-A 1xn matrix is called a row vector.

-An mx1 matrix is called a column vector.

-If A is a matrix, the notation a_{ij}, $(A)_{ij}$, or A(i,j) is used to denote the element at the intersection of the i*th* row and j*th* column.

The transpose of a matrix is denoted by A^T and is the matrix defined by

$$\left(A^T\right)_{ij}=a_{ji}$$

-If $A^T = A$, we say that A is symetric.

-If A is a matrix and λ is a scalar, then λA is defined by $(\lambda A)_{ij} = \lambda a_{ij}$. If $A=(a_{ij})$ and $B=(b_{ij})$ are mxn matrices then $(A+B)$ is defined by $(A+B)_{ij}=a_{ij}+b_{ij}$.

-If A is a mxp matrix, B is a pxn matrix, then AB is an mxp matrix defined by

$$(AB)_{ij} = \sum_{k=1}^{p} a_{ik} b_{kj} \qquad (1 \le i \le m, \quad 1 \le j \le n)$$

-Let $Ax = b, \quad Bx = d$

-If two systems have precisely the same solution, we call them equivalent systems. Thus, to solve a system a system of equations, we can instead solve any equivalent system.

Elementary Operations:

i) Interchanging two equations in the system : $E_i \Leftrightarrow E_j$

ii) Multiplying an equation by a nonzero number : $\lambda E_i \rightarrow E_i$

iii) Adding to an equation a multiple of some other equation : $E_i + \lambda E_j \rightarrow E_i$

Theorem: If one system of equations is obtained from another by a finite sequence of elementary operations, then two systems are equivalent.

-Matrix Properties:

The nxn matrix $I=\begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}$ is called an identity matrix.

$IA=A=AI$ for any matrix A of size nxn.

-If *AB=I*, then we say that *B* is a right inverse of *A* and *A* is a left inverse of *B*.

$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \alpha & \beta \end{bmatrix}=\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}=AB$$

A is the left inverse of B and B is the right inverse of A

Theorem: A square matrix can poses at most one right inverse.

Theorem: If *A* and *B* are square matrices such that *AB=I*, then *BA=I*.

Proof: Let *C=BA-I+B* then *AC=ABA-AI+AB=A-A+I=I*, thus *C* is a right inverse of *A*. By the previous theorem *B=C*; hence, *BA=I* $\Rightarrow BA=AB=I$.

We then call *B* as the inverse of *A* and say that *A* is invertible or nonsingular.

-If A is invertible, then $Ax=b$ has the solution $x=A^{-1}b$

Elementary Matrix Operations

-An elementary matrix is defined to be *nxn* matrix that arises when an elementary operation is applied to the *nxn* identity matrix.

Elementary Operations (Matrix)

i) The interchange of two rows in A: $A_s \leftrightarrow A_t$

ii) Multiplying one row by a nonzero constant $\lambda A_s \rightarrow A_s$

iii) Adding to one row a multiple of another: $A_s + \lambda A_t \rightarrow A_s$

Examples:

$$\begin{bmatrix}1&0&0\\0&0&1\\0&1&0\end{bmatrix}\begin{bmatrix}a_{11}&a_{12}&a_{13}\\a_{21}&a_{22}&a_{23}\\a_{31}&a_{32}&a_{33}\end{bmatrix}=\begin{bmatrix}a_{11}&a_{12}&a_{13}\\a_{31}&a_{32}&a_{33}\\a_{21}&a_{22}&a_{23}\end{bmatrix} \qquad A_2 \leftrightarrow A_3$$

$$\begin{bmatrix}1&0&0\\0&\lambda&0\\0&0&1\end{bmatrix}\begin{bmatrix}a_{11}&a_{12}&a_{13}\\a_{21}&a_{22}&a_{23}\\a_{31}&a_{32}&a_{33}\end{bmatrix}=\begin{bmatrix}a_{11}&a_{12}&a_{13}\\\lambda a_{21}&\lambda a_{22}&\lambda a_{23}\\a_{31}&a_{32}&a_{33}\end{bmatrix} \qquad \lambda A_2 \rightarrow A_2$$

$$\begin{bmatrix}1&0&0\\0&1&0\\0&\lambda&1\end{bmatrix}\begin{bmatrix}a_{11}&a_{12}&a_{13}\\a_{21}&a_{22}&a_{23}\\a_{31}&a_{32}&a_{33}\end{bmatrix}=\begin{bmatrix}a_{11}&a_{12}&a_{13}\\a_{21}&a_{22}&a_{23}\\\lambda a_{21}+a_{31}&\lambda a_{22}+a_{32}&\lambda a_{23}+a_{33}\end{bmatrix} \qquad A_3+\lambda A_2 \rightarrow A_3$$

Elementary matrices

-If we apply elementary row operations m times to A, we get

$E_m E_{m-1} E_{m-2} \dots E_2 E_1 A$

If A is invertible $E_m E_{m-1} E_{m-2} \dots E_2 E_1 A = I$ (possible)

$\Rightarrow A^{-1} = E_m E_{m-1} E_{m-2} \dots E_2 E_1$

Example

$$A=\begin{bmatrix}1 & 2 & 3\\ 1 & 3 & 3\\ 2 & 4 & 7\end{bmatrix},\quad E_1=\begin{bmatrix}1 & 0 & 0\\ -1 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}\Rightarrow E_1A=\begin{bmatrix}1 & 2 & 3\\ 0 & 1 & 0\\ 2 & 4 & 7\end{bmatrix},\quad E_2=\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ -2 & 0 & 1\end{bmatrix}$$

$$\Rightarrow E_2E_1A=\begin{bmatrix}1 & 2 & 3\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix},\quad E_3=\begin{bmatrix}1 & -2 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix},\quad \Rightarrow E_3E_2E_1A=\begin{bmatrix}1 & 0 & 3\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}$$

$$E_4=\begin{bmatrix}1 & 0 & -3\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix},\quad \Rightarrow E_4E_3E_2E_1A=\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}$$

$$\Rightarrow A^{-1}=E_4E_3E_2E_1I=\begin{bmatrix}1 & 0 & -3\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix}1 & -2 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ -2 & 0 & 1\end{bmatrix}\begin{bmatrix}1 & 0 & 0\\ -1 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}$$

$$=\begin{bmatrix}9 & -2 & -3\\ -1 & 1 & 0\\ -2 & 0 & 1\end{bmatrix}$$

4.2 LU and Cholesky Factorizations

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{bmatrix} \qquad \Rightarrow AX = b \qquad \Rightarrow X = A^{-1}b$$

If A is a diagonal matrix

$$\begin{bmatrix} a_{11} & 0 & 0 & \dots & 0 \\ 0 & a_{22} & 0 & \dots & 0 \\ 0 & 0 & a_{33} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{bmatrix} \qquad \Rightarrow X = \begin{bmatrix} b_1/a_{11} \\ b_2/a_{22} \\ b_3/a_{33} \\ \vdots \\ b_n/a_{nn} \end{bmatrix}$$

If $a_{ii}=0$ for some index i, and if $b_i=0$, then x_i can be any real number. If $a_{ii}=0$, no solution of the system exists.

If A is a lower triangular matrix

$$\begin{bmatrix} a_{11} & 0 & 0 & \dots & 0 \\ a_{21} & a_{22} & 0 & \dots & 0 \\ a_{31} & a_{32} & a_{33} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{bmatrix}$$

$$x_1 = \frac{b_1}{a_{11}}$$

$$x_2 = \frac{-a_{21}x_1 + b_2}{a_{22}}$$

$$x_i = \frac{b_i - \sum_{j=1}^{i-1} a_{ij}x_j}{a_{ii}}$$

If A is a upper triangular matrix

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ 0 & a_{22} & a_{23} & \dots & a_{2n} \\ 0 & 0 & a_{33} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{bmatrix}$$

$$x_n = \frac{b_n}{a_{nn}}$$

$$x_{n-1} = \frac{b_{n-1} - a_{n-1,n}x_n}{a_{n-1,n-1}}$$

$$x_i = \frac{b_i - \sum_{j=i+1}^{n} a_{ij}x_j}{a_{ii}}$$

-There is still another simple type of systems.We want

Example:

$$\begin{bmatrix} a_{11} & a_{12} & 0 \\ a_{21} & a_{22} & a_{23} \\ a_{31} & 0 & 0 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}=\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \quad \Rightarrow \begin{bmatrix} a_{31} & 0 & 0 \\ a_{11} & a_{12} & 0 \\ a_{21} & a_{22} & a_{23} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}=\begin{bmatrix} b_3 \\ b_1 \\ b_2 \end{bmatrix}$$

We need to sove this system not in the usual order 1, 2, 3 but rather in the order 3, 1, 2.

-Assume p_1 is the row of A that has zeros in positions 2,3,...n
p_2 is the row of A that has zeros in positions 3,4,...n

:

p_n is the row of A that has zeros in positions n

If we reorder rows of A as $p_1, p_2, \ldots p_n,$ the resulting matrix would be lower triangular. $(p_1, p_2, \ldots p_n)$ is called permutation vector.

-LU Factorization

Suppose A=LU where L is the lower triangular matrix and U is the upper triangular matrix.

$Ax=b \quad \Rightarrow LUx=b.$ Let $Ux=z$

$\Rightarrow Ax=LUx=Lz=b$ (L is known and b is known.) We can find z.

Then $Ux=z$ (U and z are known.) We can find x.

Suppose $A=LU=\begin{bmatrix} l_{11} & 0 & 0 & \dots & 0 \\ l_{21} & l_{22} & 0 & \dots & 0 \\ l_{31} & l_{32} & l_{33} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ l_{n1} & l_{n2} & l_{n3} & \dots & l_{nn} \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} & \dots & u_{1n} \\ 0 & u_{22} & u_{23} & \dots & u_{2n} \\ 0 & 0 & u_{33} & \dots & u_{3n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & u_{nn} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} \end{bmatrix}$

When this is possible, we say that A has an LU decomposition.

-L and U are not uniquely determined.

If we choose $l_{ii}=1$ L is called unit lower triangular matrix. If we choose $u_{ii}=1$ U is called unit upper triangular matrix.

$$a_{ij}=\sum_{s=1}^{n} l_{is}u_{sj}=\sum_{s=1}^{min(i,j)} l_{is}u_{sj}$$

-Each step in this process determines one new raw of *U* and one new column in *L*.

At step *k*, we assume rows *1,2,3,…,k-1* have been computed in *U* and columns *1,2,3,…,k-1* have been computed in *L*.

$$\text{If } i=j=k \Rightarrow a_{kk}=\sum_{s=1}^{k-1} l_{ks}u_{sk}+l_{kk}u_{kk} \quad (l_{kk} \text{ or } u_{kk} \text{ can be cumputed using this formulea})$$

$$a_{kj}=\sum_{s=1}^{k-1} l_{ks}u_{sj}+l_{kk}u_{kj} \quad (k+1\le j\le n) \quad u_{kj} \text{ can be computed if } l_{kk}\neq 0$$

$$a_{ik}=\sum_{s=1}^{k-1} l_{is}u_{sk}+l_{ik}u_{kk} \quad (k+1\le i\le n) \quad l_{ik} \text{ can be computed if } u_{kk}\neq 0$$

If we choose $l_{ii}=1,$ the algorithm is called as Doolittle's factorization. If we choose $u_{ii}=1,$ the algorithm is called as Crout's algorithm. When $U=L^T,\ l_{ii}=u_{ii},$ the algorithm is called Choelsky's factorization. (A should be real, symmetric and positive definite)

Example: Find the Doolittle, Crout, and Choelsky factorizations of the matrix

$$A=\begin{bmatrix}60 & 30 & 20\\ 30 & 20 & 15\\ 20 & 15 & 12\end{bmatrix} \quad L=\begin{bmatrix}l_{11} & 0 & 0\\ l_{21} & l_{22} & 0\\ l_{31} & l_{32} & l_{33}\end{bmatrix} \quad U=\begin{bmatrix}u_{11} & u_{12} & u_{13}\\ 0 & u_{22} & u_{23}\\ 0 & 0 & u_{33}\end{bmatrix}$$

a) Dolittle's algorithm $(l_{11}=l_{22}=l_{33}=1)$

$$\Rightarrow l_{11}u_{11}=a_{11}=60 \quad \Rightarrow u_{11}=60$$

$$\Rightarrow l_{11}u_{12}=a_{12}=30 \quad \Rightarrow u_{12}=30, \quad \Rightarrow l_{11}u_{13}=a_{13}=20 \quad \Rightarrow u_{13}=20$$

$$\Rightarrow l_{21}u_{11}=a_{21}=30 \quad \Rightarrow l_{21}=0.5, \quad \Rightarrow l_{31}u_{11}=a_{31}=20 \quad \Rightarrow l_{31}=\frac{20}{60}=\frac{1}{3}$$

$$\Rightarrow l_{21}u_{12}+l_{22}u_{22}=a_{22}=20 \quad \Rightarrow u_{22}=20-l_{21}u_{12}=20-15=5$$

$$\Rightarrow l_{21}u_{13}+l_{22}u_{23}=a_{23}=15 \quad \Rightarrow u_{23}=15-0.5*20=5$$

$$\Rightarrow l_{31}u_{12}+l_{32}u_{22}=a_{32}=15 \quad \Rightarrow l_{32}=\left(15-\frac{1}{3}30\right)/5=1$$

$$\Rightarrow l_{31}u_{13}+l_{32}u_{23}+l_{33}u_{33}=a_{33}=12 \quad \Rightarrow u_{33}=\left(12-\frac{20}{3}-5\right)=\frac{1}{3}$$

$$\Rightarrow A = LU = \begin{bmatrix} 1 & 0 & 0 \\ 0.5 & 1 & 0 \\ 1/3 & 1 & 1 \end{bmatrix} \begin{bmatrix} 60 & 30 & 20 \\ 0 & 5 & 5 \\ 0 & 0 & 1/3 \end{bmatrix}$$

$$\Rightarrow U = \begin{bmatrix} 60 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 1/3 \end{bmatrix} \begin{bmatrix} 1 & 0.5 & 1/3 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \sqrt{60} & 0 & 0 \\ 0 & \sqrt{5} & 0 \\ 0 & 0 & \sqrt{3}/3 \end{bmatrix} \begin{bmatrix} \sqrt{60} & 0 & 0 \\ 0 & \sqrt{5} & 0 \\ 0 & 0 & \sqrt{3}/3 \end{bmatrix} \begin{bmatrix} 1 & 0.5 & 1/3 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\Rightarrow A = \begin{bmatrix} 60 & 0 & 0 \\ 30 & 5 & 0 \\ 20 & 5 & 1/3 \end{bmatrix} \begin{bmatrix} 1 & 0.5 & 1/3 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$ Crout's factorization

$$\Rightarrow A = \begin{bmatrix} 1 & 0 & 0 \\ 0.5 & 1 & 0 \\ 1/3 & 1 & 1 \end{bmatrix} \begin{bmatrix} \sqrt{60} & 0 & 0 \\ 0 & \sqrt{5} & 0 \\ 0 & 0 & \sqrt{3}/3 \end{bmatrix} \begin{bmatrix} \sqrt{60} & 0 & 0 \\ 0 & \sqrt{5} & 0 \\ 0 & 0 & \sqrt{3}/3 \end{bmatrix} \begin{bmatrix} 1 & 0.5 & 1/3 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} \sqrt{60} & 0 & 0 \\ \sqrt{15} & \sqrt{5} & 0 \\ \sqrt{60}/3 & \sqrt{5} & \sqrt{3}/3 \end{bmatrix} \begin{bmatrix} \sqrt{60} & \sqrt{15} & \sqrt{60}/3 \\ 0 & \sqrt{5} & \sqrt{5} \\ 0 & 0 & \sqrt{3}/3 \end{bmatrix} = \hat{L}\hat{L}^T$$ Choelsky's factorization

Theorem: If all n leading principle minors of the nxn matrix A are nonsingular, then A has a LU decomposition.

Choelsky Factorization

-If *A* is a real, symmetric and positive definite matrix, then it has a unique factorization, $A = L L^T$, in which *L* is lower triangular with a positive diagonal.

(A is symmetric if $A = A^T$. A is positive definite if $x^T A x > 0$ for every nonzero vector x.

$$a_{kk} = \sum_{s=1}^{k-1} l_{ks}^2 + l_{kk}^2$$

$$a_{ik} = \sum_{s=1}^{k-1} l_{is} k_{ks} + l_{ik} l_{kk} \quad (k+1 \le i \le n)$$

4.3 Pivoting and Construction an algorithm

-Basic Gaussian Elimination

Example $\begin{bmatrix} 6 & -2 & 2 & 4 \\ 12 & -8 & 6 & 10 \\ 3 & -13 & 9 & 3 \\ -6 & 4 & 1 & -18 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 12 \\ 34 \\ 27 \\ -38 \end{bmatrix}$

First step:

-Substract 2 times the first eqution from the second

-Substract $1/2$ times the first eqution from the third

-Substract -1 times the first eqution from the fourth

-The numbers $2, 1/2, -1$ are called multipliers for the first step.

6 (first row first column) is called the pivot element for this step.

In the first step, row-1 is called the pivot row.

$$\begin{bmatrix} 6 & -2 & 2 & 4 \\ 0 & -4 & 2 & 2 \\ 0 & -12 & 8 & 1 \\ 0 & 2 & 3 & -14 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 12 \\ 10 \\ 21 \\ -26 \end{bmatrix}$$

-Next step: (row 2 is used as the pivot row, -4 is the pivot element.)

-Substract 3 times the second row from the third row.

-Substract -0.5 times the second row from the fourth row.

-The multipliers: 3, -0.5

$$\begin{bmatrix} 6 & -2 & 2 & 4 \\ 0 & -4 & 2 & 2 \\ 0 & 0 & 2 & -5 \\ 0 & 0 & 4 & -13 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 12 \\ 10 \\ -9 \\ -21 \end{bmatrix}$$

-Final step: (row 3 is the pivot row, 2 is the pivot element.)

-Substract 2 times the third row from the fourth row.

-The multiplier: 2.

$$\Rightarrow \begin{bmatrix} 6 & -2 & 2 & 4 \\ 0 & -4 & 2 & 2 \\ 0 & 0 & 2 & -5 \\ 0 & 0 & 0 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 12 \\ 10 \\ -9 \\ -3 \end{bmatrix}$$

Equivalent to the original system.

$$\Rightarrow x=\begin{bmatrix}x_1\\x_2\\x_3\\x_4\end{bmatrix}=\begin{bmatrix}1\\-3\\-2\\1\end{bmatrix}\qquad L=\begin{bmatrix}1 & 0 & 0 & 0\\2 & 1 & 0 & 0\\0.5 & 3 & 1 & 0\\-1 & -0.5 & 2 & 1\end{bmatrix}\qquad U=\begin{bmatrix}6 & -2 & 2 & 4\\0 & -4 & 2 & 2\\0 & 0 & 2 & -5\\0 & 0 & 0 & -3\end{bmatrix}\qquad \Rightarrow A=LU$$

Multipliers at the first step

Multipliers at the second step

Multiplier at the third step

It is easy to show that $A=LU$ Let u_1,u_2,u_3,u_4 be the first, second, third and fourth rows of U and A_1,A_2,A_3,A_4 be the first, second, third and fourth rows of A, respectively.

$$\Rightarrow U_2=A_2-2A_1,\qquad A_1=U_1$$

$$\Rightarrow A_2=U_2+2A_1=U_2+2U_1=2U_1+U_2$$

$$\Rightarrow U_3=(A_3-0.5A_1)-3U_2\Rightarrow A_3=0.5A_1+3U_2+U_3=0.5U_1+3U_2+U_3$$

$$\Rightarrow U_4=(A_4+A_1)+0.5U_2-2U_3\Rightarrow A_4=-U_1-0.5U_2+2U_3+U_4$$

Theorem: If all the pivot elements are nonzero in the process of Basic Gauss Elimination, then *A=LU*.

Pivoting

-Basic Gauss Elimination method sometimes fail.

Example: $\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

-Fails. Because there is no way of adding a multiple of the first equation to the second in order to obtain 0 coefficient of x_1 in the second equation.

Example:

$\begin{bmatrix} \epsilon & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ ϵ is a small number. FAILS

$$\Rightarrow \begin{bmatrix} \epsilon & 1 \\ 0 & 1-\epsilon^{-1} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2-\epsilon^{-1} \end{bmatrix}$$

$$x_2 = \frac{2-\epsilon^{-1}}{1-\epsilon^{-1}} \simeq 1$$

$$x_1 = (1-x_2)\epsilon^{-1} \simeq 0$$

In computer, if ϵ is small enough, $2-\epsilon^{-1}$ will be computed to be the same as $-\epsilon^{-1}$.

The correct solution : $x_1 = \frac{1}{1-\epsilon} \simeq 1, \quad x_2 = \frac{1-2\epsilon}{1-\epsilon} \simeq 1.$

The problem is not the smallness of a_{11}. Rather, it is the smallness of a_{11} relative to the elements in its row.

Example:

$$\begin{bmatrix} 1 & \epsilon^{-1} \\ 1 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}=\begin{bmatrix} \epsilon^{-1} \\ 2 \end{bmatrix}$$

ϵ is a small number. Simple Gaussian algorithm produces:

$$\begin{bmatrix} 1 & \epsilon^{-1} \\ 0 & 1-\epsilon^{-1} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}=\begin{bmatrix} \epsilon^{-1} \\ 2-\epsilon^{-1} \end{bmatrix} \quad \Rightarrow x_2=\frac{2-\epsilon^{-1}}{1-\epsilon^{-1}}\simeq 1$$

$$x_1=\epsilon^{-1}-\epsilon^{-1}x_2\simeq 0 \qquad \text{wrong.}$$

If we change the order of equatioms:

$$\begin{bmatrix} 1 & 1 \\ \epsilon & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}=\begin{bmatrix} 2 \\ 1 \end{bmatrix} \quad \Rightarrow \begin{bmatrix} 1 & 1 \\ 0 & 1-\epsilon \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}=\begin{bmatrix} 2 \\ 1-2\epsilon \end{bmatrix} \quad \Rightarrow x_2=1, \quad x_1=1.$$

$$\begin{bmatrix} 1 & 1 \\ 1 & \epsilon^{-1} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}=\begin{bmatrix} 2 \\ \epsilon^{-1} \end{bmatrix} \quad \Rightarrow \begin{bmatrix} 1 & 1 \\ 0 & \epsilon^{-1}-1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}=\begin{bmatrix} 2 \\ \epsilon^{-1}-2 \end{bmatrix} \quad \Rightarrow x_2=1, \quad x_1=1.$$

-For the reason of economy in computing, we prefer not to move the rows of the matrix around in the computers's memory. Instead, we simply choose the pivot rows in a logical manner. Instead of using rows in the order 1, 2, 3, …,n-1 as pivot rows, we use p_1, p_2, p_3,…,p_{n-1}.

-In the first step multipliers of row p_1 will be subtracted from rows p_2,...,p_n

-In the next step multipliers of row p_2 will be subtracted from rows p_3,...,p_{n-1} ; and so on.

Gaussian Elimination with Scaled Row Pivoting

(Avoids choosing small pivot elements compared to other elements of the pivot row)

The algorithm consists of two parts: a factorization phase and a solution phase.

Factorization Phase:

$Ax=b \qquad PAx=Pb$ Permutation matrix P derived from the permutation array p.

$\Rightarrow PAx=LUx=Pb$ Let $Ux=z \quad \Rightarrow Lz=Pb$ (Find z) $\quad \Rightarrow Ux=z$ (Find x)

Factorization Phase:

$$s_i=\max_{1\le j\le n}|a_{ij}|=\max\{|a_{i1}|,|a_{i2}|,|a_{i3}|,\ldots,|a_{in}|\} \quad (1\le i\le n)$$

First step: Select the row as the pivot row for which $\frac{|a_{i1}|}{s_i}$ is largest.

Set the permutation vector $(p_1, p_2, \ldots, p_n)$ to $(1, 2, 3, \ldots, n)$. Then select an index j for which $\frac{|a_{p_j,1}|}{s_{p_j}}$ is maximal and interchange p_1 with p_j in the permutation array p. Then substract $\frac{a_{p_i,1}}{a_{p_1,1}}$ times p_1 from row p_i forr $2 \le i \le n$

kth step: Find an index j for which $|a_{p_j k}|/s_{p_j}$ is maximal. Interchange p_k with p_j in the array p, and then substract $a_{p_i k}/a_{p_k k}$ times row p_k from row p_i for $k+1 \le i \le n$.

Example: Find PA=LU using Gaussian Elimination with scaled row pivoting.

$$A=\begin{bmatrix}2 & 3 & -6\\ 1 & -6 & 8\\ 3 & -2 & 1\end{bmatrix} \qquad p=(1,2,3) \text{ and } s=(6,8,3)$$

$$\Rightarrow |a_{p_i 1}|/s_{p_i}=\{2/6,1/8,3/3\}$$

$$\Rightarrow \text{ row 3 is pivot row.} \qquad \Rightarrow p=(3,2,1)$$

$$A_1=\begin{bmatrix}2/3 & 13/3 & -20/3\\ 1/3 & -16/3 & 23/3\\ 3 & -2 & 1\end{bmatrix}$$

$$\Rightarrow |a_{p_2 2}|/s_{p_2}=(16/3)/8=16/24$$

$$\Rightarrow |a_{p_3 2}|/s_{p_3}=(13/3)/6=13/18 \Rightarrow (13/18)>(16/24)$$

$$\Rightarrow \text{ row 1 is pivot row.} \qquad \Rightarrow p=(3,1,2)$$

$$A_2=\begin{bmatrix}2/3 & 13/3 & -20/3\\ 1/3 & -16/13 & -7/13\\ 3 & -2 & 1\end{bmatrix}$$

$$\Rightarrow PA=\begin{bmatrix}1 & 0 & 0\\ 2/3 & 1 & 0\\ 1/3 & -16/13 & 1\end{bmatrix}\begin{bmatrix}3 & -2 & 1\\ 0 & 13/3 & -20/3\\ 0 & 0 & -7/13\end{bmatrix}=\begin{bmatrix}3 & -2 & 1\\ 2 & 3 & -6\\ 1 & -6 & 8\end{bmatrix}$$

-We can use the same algorithm to find the inverse of a matrix. Apply the elementary matrix operations to I. $(E_n E_{n-1} E_{n-2} \ldots E_2 E_1 I = A^{-1})$

-We want