Root Finding

Root Finding

Suppose you wish to find the wave number, k, of gravity water waves with a frequency, f, of 0.2 Hz. in water that is 5 meters deep. The circular frequency of 0.2 Hz. waves is $\omega = 2\pi f = 1.2566$ radians/second. The dispersion relation for gravity water waves is:

$$kg \tanh kh = \omega^2$$

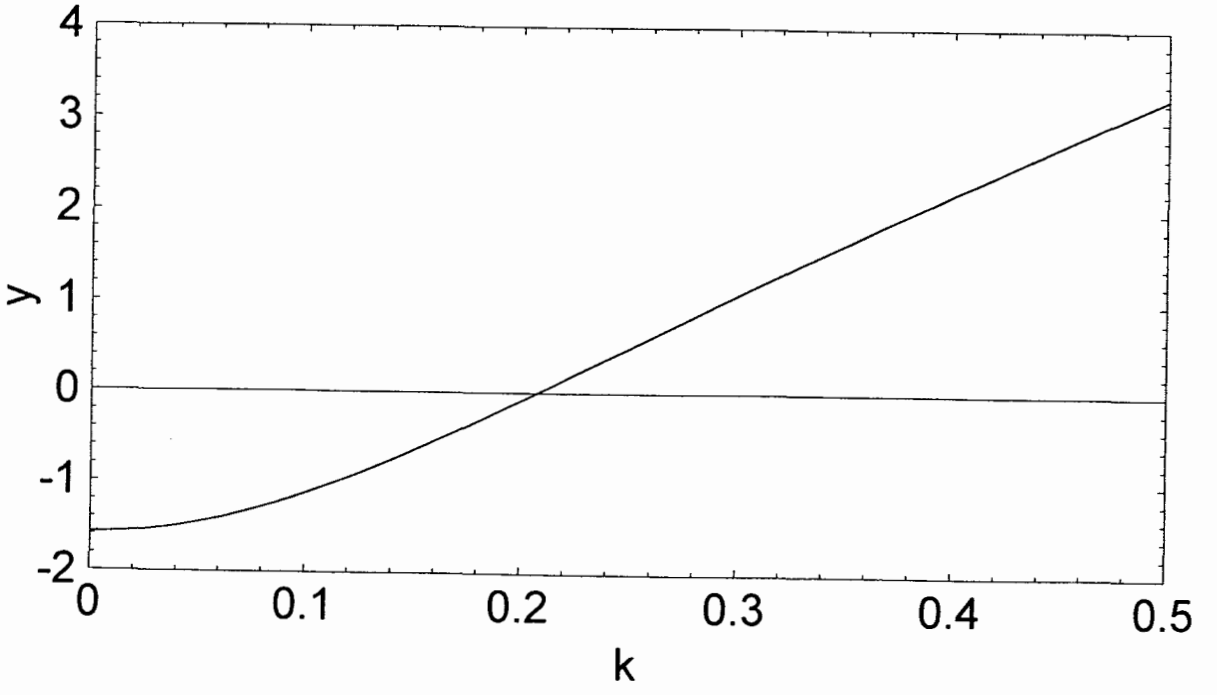
g is the acceleration of gravity, 9.81 m/s^2 and h is the water depth, 5 m.

This equation can be written as:

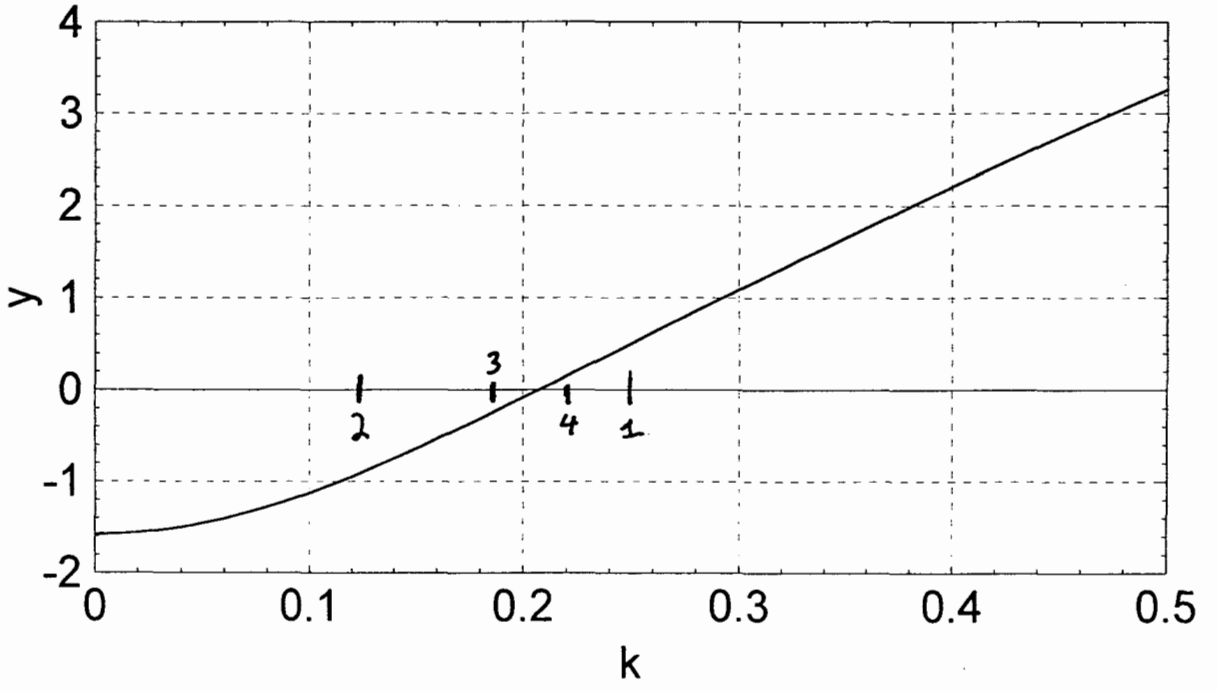
$$kg \tanh kh - \omega^2 = 0$$

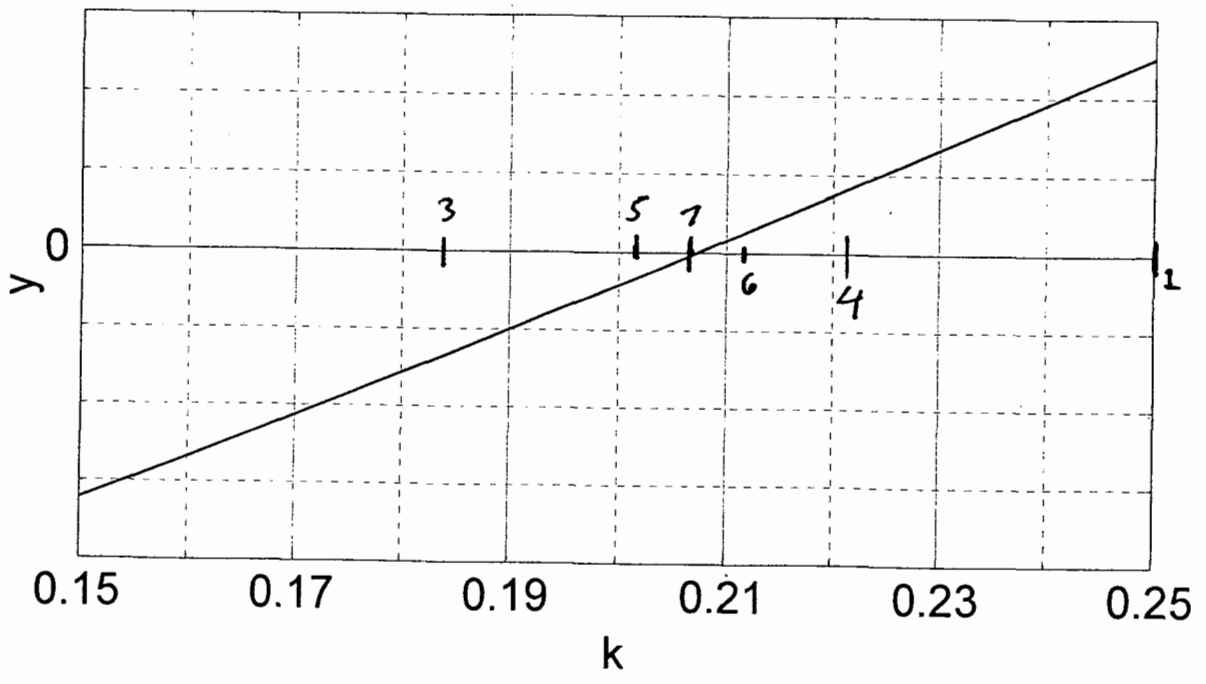
If we write an equation: $y(k) = kg \tanh kh - \omega^2$,

The problem at hand is the same as asking: "What is the value of k such that $y(k) = 0$? The value of a quantity that makes another equal to zero is called a root and the question above is called *Root Finding.*



Bisection Method

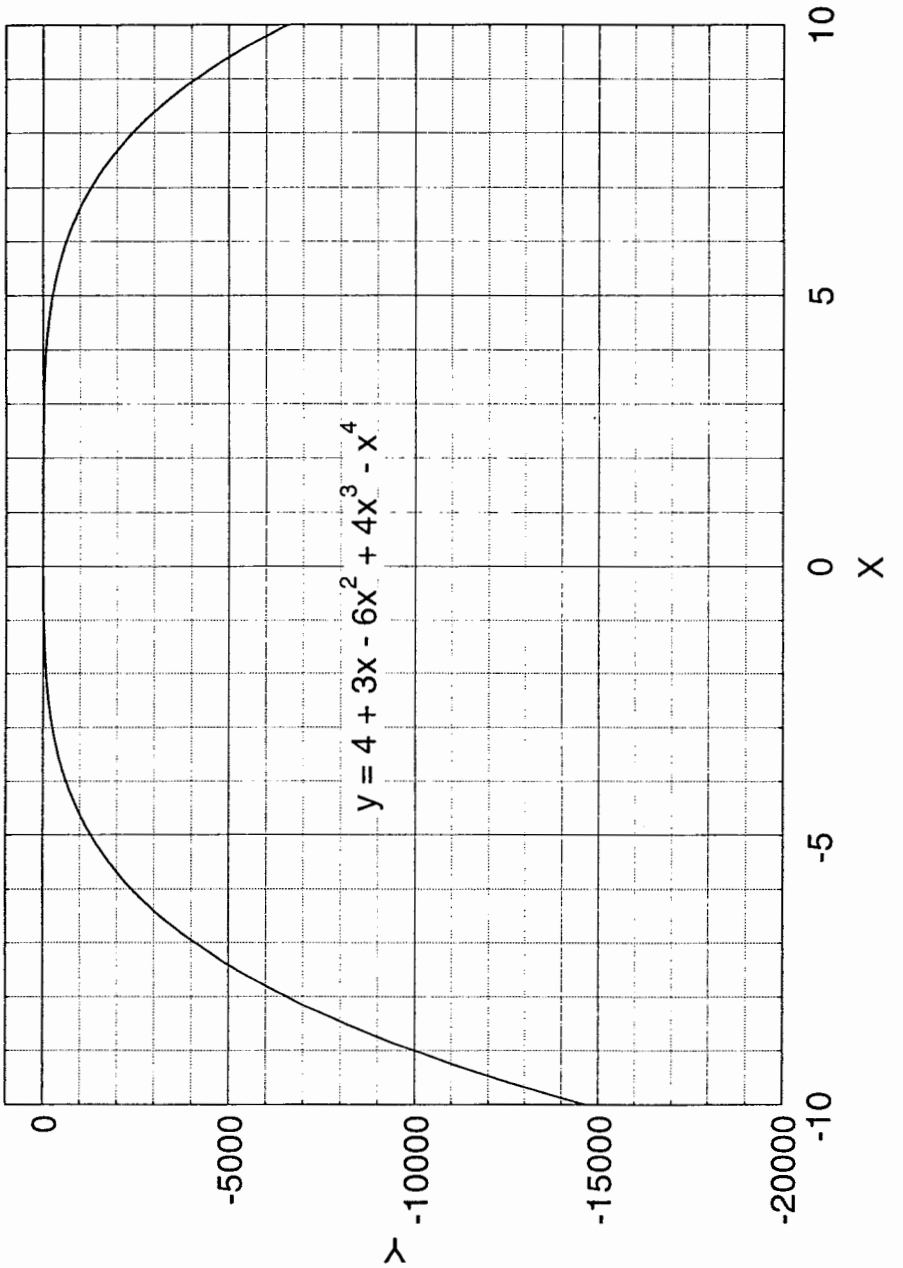




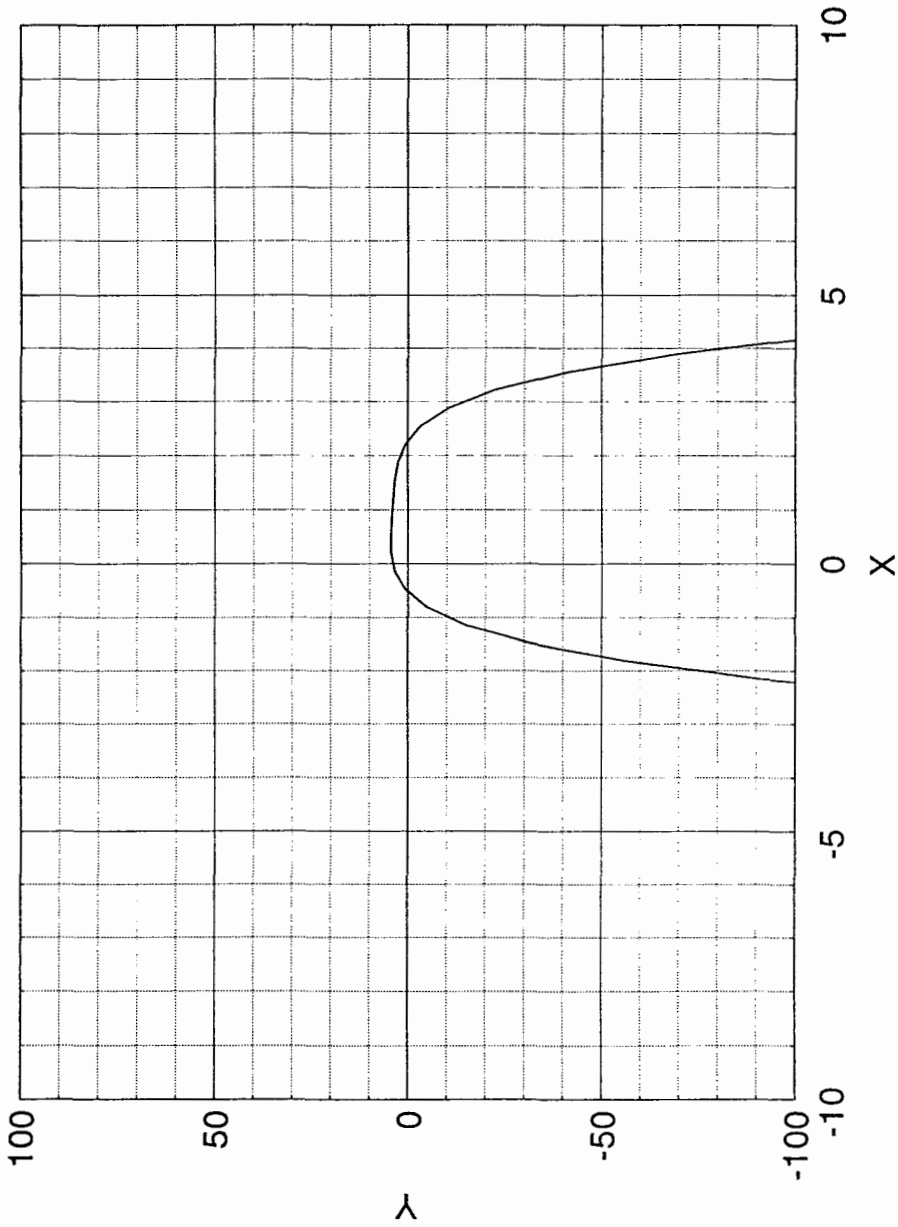
```
% biseck program to find k given omega
% using the bisection root finding method
om = 1.2566;
g = 9.81;
h = 5.0;
k1 = 0.0;
k2 = 0.5;
k3 = 0.25;
y = k3*g*tanh(k3*h) - om^2;
for m = 1:50
   y = k3*g*tanh(k3*h) - om^2;
  if (y*y < 1.0e-8);
      break
   end
   if(y >= 0.0);
      k2 = k3;
      k3 = 0.5*(k1+k2);
   elseif (y <= 0.0) ;
      k1 = k3;
      k3 = 0.5*(k1  + k2);
   else;
      fprintf(1,'there was no root')
   end;
end ;
fprintf(1,' k = %8.4f\n',k3);
fprintf(1,' Number of iterations =%3.0f\n',m);
```

```
>> biseck
 k =   0.2073
 Number of iterations = 15
>>
```

RF2a



RF2b



RF2C

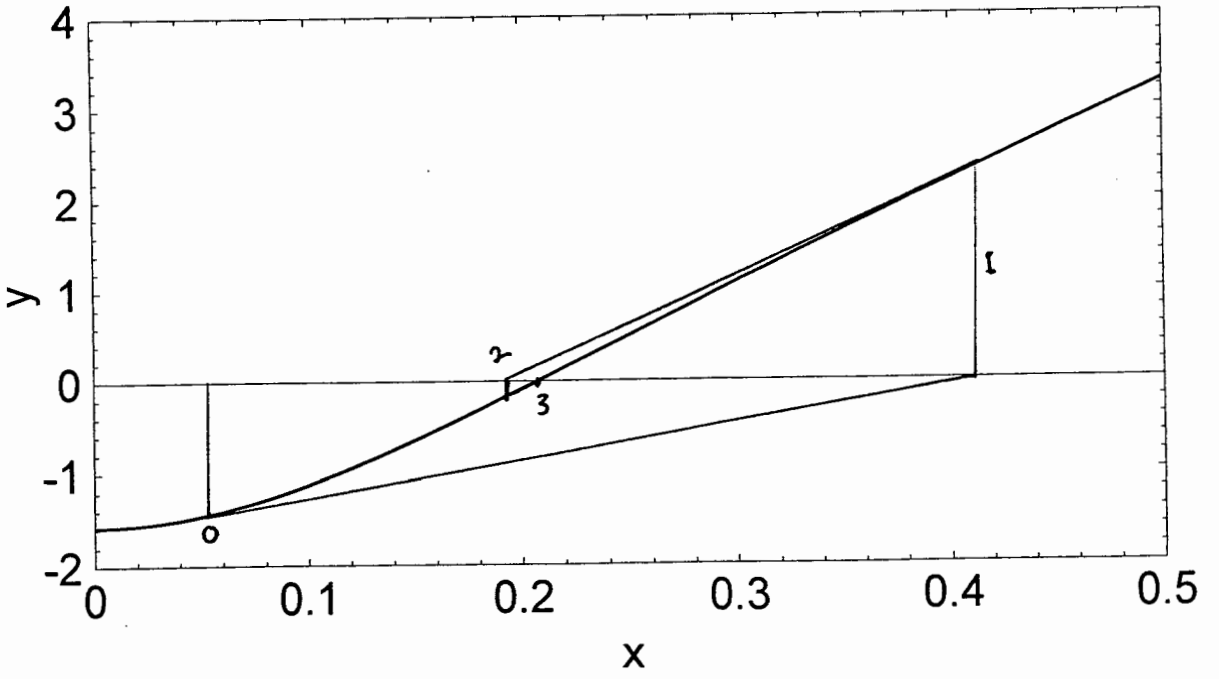
$y=4+3x-6x^2+4x^3-x^4$

1.00000	4.000000000	0.00000	4.000000000
2.00000	2.000000000	-1.00000	-10.000000000
3.00000	-14.000000000	-0.50000	0.437500000
2.50000	-2.562500000	-0.75000	-3.628906250
2.25000	0.308593750	-0.62500	-1.347900391
2.37500	-0.949462891	-0.56250	-0.397964478
2.31250	-0.280044556	-0.53125	0.033507347
2.28130	0.023423625	-0.54688	-0.178792423
2.29690	-0.125854491	-0.53906	-0.071706616
2.28910	-0.050608813	-0.53516	-0.018951368
2.28520	-0.013441856	-0.53320	0.007398979
2.28325	0.005028455	-0.53418	-0.005762632
2.28423	-0.004244654	-0.53369	0.000821562
2.28374	0.000394274	-0.53345	0.004044001
2.28398	-0.001877260	-0.53381	-0.000790267
2.28386	-0.000741351	-0.53363	0.001627324
2.28380	-0.000173503	-0.53354	0.002835777
2.28377	0.000110395	-0.53368	0.000955863
2.28379	-0.000078868	-0.53374	0.000150016
2.28378	0.000015764	-0.53378	-0.000387271
		-0.53376	-0.000118622
		-0.53375	0.000015699

Newton's Method for Finding Roots of $y(x)$

The approach taken in Newton's method is to take an estimate of the location of a root of $y(x)$ and then improve upon it. Thus, it is iterative, starting with a"guess" for x with each successive iteration being the result of the last iteration. The basic formula for each iteration is:

$$x_i = x_{i-1} - \frac{y_{i-1}}{y'_{i-1}} \qquad \text{where} \quad y'_{i-1} = \left|\frac{dy}{dx}\right|_{x=x_{i-1}}$$



$y = 4 + 3x - 6x^2 + 4x^3 - x^4 \quad y' = 3 - 12x + 12x^2 - 4x^3$

x	y	dy/dx	(y)/(dy/dx)
1.00000000	4.000000000	-1.000000000	-4
5.00000000	-256.000000000	-257.000000000	0.9961089
4.00389105	-80.424942796	-109.420778749	0.7350061
3.26888492	-24.769129200	-47.719415341	0.5190577
2.74982724	-7.125030491	-22.431151690	0.31764
2.43218726	-1.639446076	-12.750582959	0.1285781
2.30360914	-0.191558554	-9.861396677	0.0194251
2.28418404	-0.003809366	-9.471138816	0.0004022
2.28378183	-0.000001600	-9.463181800	1.691E-07
2.28378167	0.000000000	-9.463178456	3.003E-14

RF3c

Review of Matrix Algebra

An m x n matrix (m rows, n columns) ia said to be of order $m \times n$ and is written symbolically as:

$$\mathbf{A} = \underline{a} = \begin{bmatrix} a_{11} & a_{12} & \cdot & \cdot & \cdot & a_{1n} \\ a_{21} & a_{22} & \cdot & \cdot & \cdot & a_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{m1} & a_{m2} & \cdot & \cdot & \cdot & a_{mn} \end{bmatrix}$$

Each a_{ij} represents a numerical value. If the a_{ij}'s are real numbers, the matrix is called a *real matrix.* If the a_{ij}'s are complex numbers, the matrix is called a *complex matrix.*

The matrix is called *square* if $m = n$. Matrices $\mathbf{A}$ and $\mathbf{B}$ are called *equal* if $a_{ij} = b_{ij}$ for all i and j and they have the same number of rows and the same number of columns.

If $\mathbf{A}$ and $\mathbf{B}$ are both order $m \times n$ matrices, the matrix $\mathbf{C} = \mathbf{A} + \mathbf{B}$ is defined by the relations:

$$c_{ij} = a_{ij} + b_{ij}$$

The matrix $\mathbf{D} = \gamma\mathbf{A}$ is defined by the relations:

$$d_{ij} = \gamma a_{ij}$$

An important relation is the the *matrix product*, of two matrices $\mathbf{A}$ $(m \times n)$ and $\mathbf{B}$ $(n \times p)$ which is denoted by $\mathbf{C} = \mathbf{AB}$ whose elements are defined by:

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}, \qquad i = 1, 2, \cdots, m, \qquad j = 1, 2, \cdots, p$$

When $\mathbf{A}$ and $\mathbf{B}$ are square and of the same order, both $\mathbf{AB}$ and $\mathbf{BA}$ are defined, but except under special circumstances, $\mathbf{AB} \neq \mathbf{BA}$.

An $m \times 1$ matrix (m rows, 1 column) is called a *column vector* or a *vector* and is written symbolically as:

$$\mathbf{x} = \underline{x} = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_m \end{bmatrix}$$

Each x_i represents a numerical value.

The standard form for a set of linear equations is:

$$\mathbf{Ax} = \mathbf{b}$$

where $\mathbf{A}$ is an $m \times n$ matrix called the *coefficient matrix* and $\mathbf{x}$ is an *unknown vector* of length n. $\mathbf{b}$ is a known vector of length n.

Let $\mathbf{a_j}$ denote the j^{th} column of $\mathbf{A}$. Then the set of equations can be written as:

$$x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_n\mathbf{a}_n = \mathbf{b}$$

The vectors $\mathbf{a}_j$, $j = 1, 2, ..., n$ are *linearly dependent* if there is a set of numbers $x_1, x_2, ...x_n$, with at least one x_j being non zero such that:

$$x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_n\mathbf{a}_n = 0$$

In this instance, at least one $\mathbf{a}_j$ is a linear combination of the remaining $\mathbf{a}_j$'s.

The vectors $\mathbf{a}_j$, $j = 1, 2, ..., n$ are *linearly independent* if they are not linearly dependent.

If each of the m linear equations is *independent* then there is an exact solution if $m = n$. If $m > n$, there are more equations than unknowns and there is no exact solution. Rather, there is an approximate solution for $\mathbf{x}$ which is usually chosen to achieve minimum sum of the squared errors from each equation.

Example:

$$3x_1 + x_2 = 5$$
$$x_1 + 4x_2 = -3$$
$$-2x_1 + 3x_2 = -6$$

$$\begin{bmatrix} 3 & 1 \\ 1 & 4 \\ -2 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 5 \\ -3 \\ -6 \end{bmatrix}$$

Approximate Solution is: $x_1 = y_1$, $x_2 = y_2$: $\underline{x} = \underline{y}$

Need to find $\underline{y}$. Error Vector is $\underline{e} = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix}$

$$e_1 = 5 - 3y_1 - y_2$$
$$e_2 = -3 - y_1 - 4y_2$$
$$e_3 = -6 + 2y_1 - 3y_2$$

sum of squares of errors

$$E = e_1^2 + e_2^2 + e_3^2$$

$$E = (5 - 3y_1 - y_2)^2 + (-3 - y_1 - 4y_2)^2 + (-6 + 2y_1 - 3y_2)^2$$

We seek $\underline{y}$ such that $\frac{\partial E}{\partial y_1} = 0$ and $\frac{\partial E}{\partial y_2} = 0$

which is the same as; $\frac{1}{2}\frac{\partial E}{\partial y_1} = 0$ and $\frac{1}{2}\frac{\partial E}{\partial y_2} = 0$

$$\frac{1}{2}\frac{\partial E}{\partial y_1} = -3(5 - 3y_1 - y_2) - 1(-3 - y_1 - 4y_2) + 2(-6 + 2y_1 - 3y_2)$$

$$= \begin{array}{r} -15 + 9y_1 + 3y_2 \\ 3 + y_1 + 4y_2 \\ -12 + 4y_1 - 6y_2 \\ \hline -24 + 14y_1 + y_2 \end{array}$$

$$\frac{1}{2}\frac{\partial E}{\partial y_2} = 25 - y_1 + 26y_2$$

$0 = -24 + 14y_1 + y_2$ $\qquad 14y_1 + y_2 = 24$

$0 = 25 + y_1 + 26y_2$ $\qquad y_1 + 26y_2 = -25$

$$\begin{bmatrix} 14 & 1 \\ 1 & 26 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 24 \\ -25 \end{bmatrix}$$

$y_1 = 1.7879, \quad y_2 = -1.0303$

$e_1 = 5 - 5.3637 + 1.0303 = 0.6667$

$e_2 = -3 - 1.7879 + 4.1212 = -0.6667$

$e_3 = -6 + 3.5758 - 3.0909 = 0.6667$

Consider a square $n \times n$ matrix, $\mathbf{A}$. It is called the *identity matrix of order* n, $\mathbf{I}_n$ (or simply $\mathbf{I}$) if

$$a_{ij} = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases}$$

This is often written as: $a_{ij} = \delta_{ij}$

If $\mathbf{B}$ is an $n \times p$ matrix,

$$\mathbf{I_n B = B} \qquad \textbf{and} \qquad \mathbf{B I_n = B}$$

The j^{th} column of $\mathbf{I}$ is called the j^{th} *unit vector* and denoted by $\mathbf{e}_j$. Any n-vector $\mathbf{b}$ can be written as:

$$\mathbf{b} = \sum_{j=1}^{n} b_j \mathbf{e}_j$$

The j^{th} column of a matrix $\mathbf{B}$ is given by $\mathbf{Be}_j$. Therefore, if $\mathbf{C = AB}$, the j^{th} column of $\mathbf{C}$, called $\mathbf{c}_j$ is obtained as:

$$\mathbf{c}_j = \mathbf{Ce}_j = (\mathbf{AB})\mathbf{e}_j = \mathbf{A}(\mathbf{Be}_j) = \mathbf{Ab}_j$$

A collection V of linearly independent vectors in R^n (this means that the vectors are $1 \times n$) is called a *basis* for R^n if every n-vector can be written as a linear combination of the vectors in V. Obviously, the columns of $\mathbf{I}_n$, which are the $\mathbf{e}_j$'s form a basis for R^n. However, this is not the only set V of basis vectors. Any basis in R^n contains exactly n vectors.

THEOREM: The linear system $\mathbf{Ax = b}$ has a unique solution if and only if the only solution to $\mathbf{Ax = 0}$ is $\mathbf{x = 0}$.

THEOREM: If the homogeneous linear system $\mathbf{Ax = 0}$ has fewer equations than unknowns, it has nonzero (non trivial) solutions.

THEOREM: $\mathbf{A}$ is an $m \times n$ matrix. If the linear system $\mathbf{Ax = b}$ has a solution for every m-vector $\mathbf{b}$, then $m \leq n$.

Consider a square $n \times n$ matrix, $\mathbf{A}$. If there is a square $n \times n$ matrix, $\mathbf{B}$ such that $\mathbf{AB} = \mathbf{I}$, $\mathbf{B}$ is called the *inverse* of $\mathbf{A}$ and denoted by $\mathbf{A}^{-1}$. If $\mathbf{A}$ has an inverse $\mathbf{A}$ is called *nonsingular* and if it does not have an inverse, $\mathbf{A}$ is called *singular.*

Fact: If $\mathbf{A}$ and $\mathbf{B}$ are invertible, then, $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\,\mathbf{A}^{-1}$

Suppose there is a linear system of equations $\mathbf{Ax} = \mathbf{b}$ where $\mathbf{A}$ is invertible and $\mathbf{b}$ is a known vector and $\mathbf{x}$ is an unknown vector to be determined. Pre-multiplying the equation by $\mathbf{A}^{-1}$ gives:

$$\mathbf{A}^{-1}\,\mathbf{A}\,\mathbf{x} = \mathbf{A}^{-1}\,\mathbf{b}$$

$$\mathbf{x} = \mathbf{A}^{-1}\,\mathbf{b}$$

Thus, if the inverse of $\mathbf{A}$ is determined, the solution to the system of equations can be obtained by straightforward matrix multiplication. This is not a numerically efficient way to solve sets of linear equations, but it demonstrates theoretically that a solution exists of the *coefficient matrix* is nonsingular.

The principle of obtaining the inverse of a matrix can be demonstrated as follows: An $n \times n$ matrix $\mathbf{A}$ exists and the goal to to determine its inverse $\mathbf{B} \equiv \mathbf{A}^{-1}$ whose j^{th} column is called $\mathbf{b}_j$. $\mathbf{A}\,\mathbf{b}_j = \mathbf{e}_j$ is a system of n linear equations for the n elements of $\mathbf{b}_j$. Finding $\mathbf{A}^{-1}$ in this way requires solving a set of n equations for each of the n column vectors $\mathbf{b}_j$.

A more computationally efficient way to find the inverse of a nonsingular matrix will be shown subsequently.

Determinant of a Matrix

A matrix $\mathbf{A}$ has a *determinant* which is denoted by $\det(\mathbf{A})$ or $|\mathbf{A}|$. The determinant of a matrix plays a large theoretical role in linear algebra and a practical role in determining the inverse of a matrix. The determinant is defined as the sum of all signed elementary products from the matrix. An elementary product is the product of n elements all of which are from different rows and different columns. The sign is + if the number of inversions of the indices is even and − if the number of invertsions is odd.

$$|A| = \sum_{j,k,l,\ldots,q=1}^{n} (-1)^i a_{1j}\, a_{2k}\, a_{3l} \ldots a_{nq}$$

where $j, k, l, ..., q$ are all different and i is the number of inversions in the sequence $j, j, l, ..., q$.

A recursive definition of $|A|$ which includes a "prescription" of how to calculate it is as follows:

1. If $\mathbf{R}$ is a 1×1 matrix, $\mathbf{R} = [r]$, $\det(\mathbf{R}) \equiv r$.
2. $\mathbf{A}_{ij}$ is the *submatrix* of $\mathbf{A}$ obtained by deleting the i^{th} row and j^{th} column of $\mathbf{A}$. The *minor* m_{ij} (associated with the matrix $\mathbf{A}$) is the determinant of $\mathbf{A}_{ij}$
$$m_{ij} = \det(\mathbf{A}_{ij})$$
3. The *cofactor* c_{ij} is defined by $c_{ij} \equiv (-1)^{i+j}\, m_{ij}$
4. $$\det(\mathbf{A}) \equiv \sum_{j=1}^{n} a_{ij}\, c_{ij} \qquad \text{for any } i = 1, 2, ..., n, \text{ or}$$
$$\det(\mathbf{A}) \equiv \sum_{i=1}^{n} a_{ij}\, c_{ij} \qquad \text{for any } j = 1, 2, ..., n$$

- Adding a constant times one row of a matrix to another row does not change the determinant.
- Adding a constant times one column of a matrix to another column does not does not change the determinant.

Transpose of a Matrix

The transpose of a matrix $\mathbf{A}$ is called $\mathbf{A}^T$ and is obtained by making each row of $\mathbf{A}^T$ the corresponding column of $\mathbf{A}$. For example, if we define $\mathbf{C} \equiv \mathbf{A}^T$,

$$c_{ij} = a_{ji}$$

Calculating the Inverse of a Matrix

Consider a matrix $\mathbf{A}$. It's cofactors are c_{ij}. The matrix $\mathbf{C}$ is the matrix whose ij element is c_{ij}.

The matrix $\mathbf{C}^T$ is called the *adjugate* or the *adjoint* of $\mathbf{A}$ and is denoted by $\text{adj}(\mathbf{A})$.

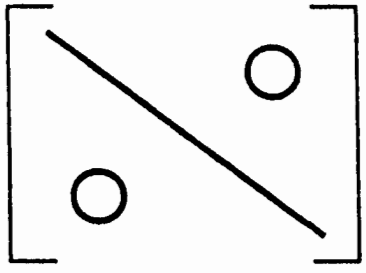
The inverse a $\mathbf{A}$ is given by:

$$A^{-1} = \frac{\text{adj}(\mathbf{A})}{\det(\mathbf{A})}$$

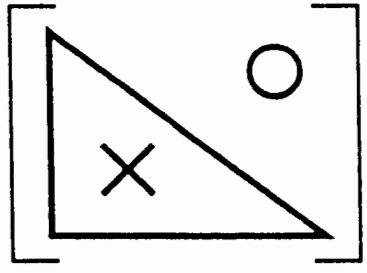
Cramer's Rule

Consider the system of linear equations, $\mathbf{A}\,\mathbf{x} = \mathbf{b}$. Define $\mathbf{A}^j$ as the matrix formed by replacing the j^{th} column of $\mathbf{A}$ by the column vector $\mathbf{b}$.

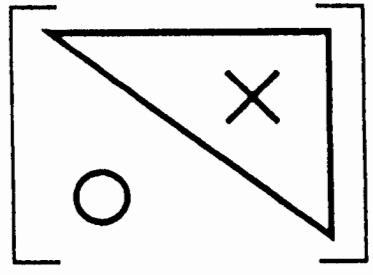
$$\text{Cramer's Rule is:} \quad x_j = \frac{|\mathbf{A}^j|}{|\mathbf{A}|}$$



Diagonal



Lower triangular



Upper triangular

The $\times$'s and the straight lines denote nonzero elements
and the $\bigcirc$'s denote zero elements.

Special matrices.

m 4 6 >

Matrix Norms

First we define vector norms $N(\mathbf{x})$ of the vector $\mathbf{x}$ in n dimensional space.. A vector norm has the following properties:

1. $N(\mathbf{x}) \geq 0$ for all n-vectors $\mathbf{x}$.
2. $N(\mathbf{x}) = 0$ if and only if $\mathbf{x} = 0$.
3. $N(\alpha\mathbf{x}) = |\alpha| N(\mathbf{x})$ for all real α and n-vectors $\mathbf{x}$.
4. $N(\mathbf{x}+\mathbf{y}) \leq N(\mathbf{x}) + N(\mathbf{y})$ for all n-vectors $\mathbf{x}$ and $\mathbf{y}$.

Norms are denoted by $||\cdot||$.

$$||\mathbf{x}||_1 = \sum_{i=1}^{n} |x_i|\,, \qquad ||\mathbf{x}||_2 = \sqrt{\sum_{i=1}^{n} x_i^2}\,, \qquad ||x||_\infty = \max_{1\leq i\leq n} |x_i|$$

An *operator norm* $N(\mathbf{A})$ of a real valued $n \times n$ matrix $\mathbf{A}$ is a real valued function having the following properties:

1. $N(\mathbf{A}) \geq 0$.
2. $N(\mathbf{A}) = 0$ if and only if all the elements of $\mathbf{A}$ are zero.
3. $N(\alpha\mathbf{A}) = |\alpha|\, N(\mathbf{A})$ for all real α.
4. $N(\mathbf{A}+\mathbf{B}) \leq N(\mathbf{A}) + N(\mathbf{B})$.
5. $N(\mathbf{AB}) \leq N(\mathbf{A})\, N(\mathbf{B})$.

There are a number of equivalent definitions for the v-norm of $\mathbf{A}$, $N(\mathbf{A}) = ||\mathbf{A}||_v$. One of them is:

For all n-vectors $\mathbf{z}$ such that $||\mathbf{z}||_v \leq 1$,

$$||\mathbf{A}||_v = \max_{||\mathbf{z}\leq 1||} ||\mathbf{Az}||_v$$

The Condition Number of A Matrix

Consider a set of linear equations, $\mathbf{Ax} = \mathbf{b}$ which is to be solved numerically. The exact solution is $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$. How sensitive is the accuracy of the solution to numerical errors? After obtaining an approximate numerical solution $\hat{\mathbf{x}}$, we can always compute the residual $\mathbf{r}$ as:

$$\mathbf{r} = \mathbf{b} - \mathbf{A}\hat{\mathbf{x}}$$

We would like to find some relationship between the computable residual $\mathbf{r}$ and the error $\mathbf{e} = \mathbf{x} - \hat{\mathbf{x}}$.

$$\mathbf{r} = \mathbf{b} - \mathbf{A}\hat{\mathbf{x}} = \mathbf{Ax} - \mathbf{A}\hat{\mathbf{x}} = \mathbf{A}(\mathbf{x} - \hat{\mathbf{x}}) = \mathbf{Ae}$$

$$\text{Since } ||\mathbf{r}|| \leq ||\mathbf{A}||\,||\mathbf{e}||, \qquad ||\mathbf{e}|| \geq \frac{||\mathbf{r}||}{||\mathbf{A}||}$$

$$\text{Then, since } ||\mathbf{x}|| \leq ||\mathbf{A}^{-1}||\,||\mathbf{b}|| \qquad \frac{||\mathbf{e}||}{||\mathbf{x}||} \geq \frac{1}{||\mathbf{A}||\,||\mathbf{A}^{-1}||}\frac{||\mathbf{r}||}{||\mathbf{b}||}$$

$$\text{Now we use: } ||\mathbf{e}|| \leq ||\mathbf{A}^{-1}||\,||\mathbf{r}|| \quad \text{and} \quad ||\mathbf{b}|| \leq ||\mathbf{A}||\,||\mathbf{x}||$$

$$\text{These give: } \frac{||\mathbf{e}||}{||\mathbf{x}||} \leq ||\mathbf{A}||\,||\mathbf{A}^{-1}||\,\frac{||\mathbf{r}||}{||\mathbf{b}||}$$

$$\text{Hence: } \frac{1}{||\mathbf{A}||\,||\mathbf{A}^{-1}||}\frac{||\mathbf{r}||}{||\mathbf{b}||} \leq \frac{||\mathbf{e}||}{||\mathbf{x}||} \leq ||\mathbf{A}||\,||\mathbf{A}^{-1}||\,\frac{||\mathbf{r}||}{||\mathbf{b}||}$$

$$\frac{1}{||\mathbf{A}||\,||\mathbf{A}^{-1}||}\frac{||\mathbf{r}||}{||\mathbf{b}||} \leq \frac{||\mathbf{e}||}{||\mathbf{x}||} \leq ||\mathbf{A}||\,||\mathbf{A}^{-1}||\,\frac{||\mathbf{r}||}{||\mathbf{b}||}$$

The *condition number* of an $n \times n$ matrix, $\mathbf{A}$, with respect to the operator norm $||\cdot||$ is called $\kappa(\mathbf{A})$ and defined by:

$$\kappa(\mathbf{A}) = ||\mathbf{A}||\;||\mathbf{A}^{-1}||$$

The condition number multiplied by the norm of the relative residual is an upper bound on the norm of the relative error. $1/\kappa(\mathbf{A})$ multiplied by the relative residual is a lower bound on the norm of the relative error.

For a large condition number, the relative residual is a poor indicator of the relative error. For $\kappa \approx 1$, the relative residual is a good measure of the relative error.

GAUSSIAN ELIMINATION

Triangular Systems

$$\begin{aligned} u_{11}x_1 + u_{12}x_2 + \cdots + u_{1,n-1}x_{n-1} + u_{1n}x_n &= f_1 \\ u_{22}x_2 + \cdots + u_{2,n-1}x_{n-1} + u_{2n}x_n &= f_2 \\ \ddots \qquad\qquad & \vdots \\ u_{n-1,n-1}x_{n-1} + u_{n-1,n}x_n &= f_{n-1} \\ u_{nn}x_n &= f_n \end{aligned}$$

back substitution

Solve the following upper triangular system by using back substitution.

$$\begin{aligned} 2x_1 - 5x_2 + x_3 &= 3 \\ 3x_2 - x_3 &= 7 \\ 4x_3 &= 8 \end{aligned}$$

Gaussian Elimination. This general solution technique is based on the following basic properties of linear systems:

1. Multiplying an equation by a constant does not alter the solution to the system.
2. Replacing an equation by a linear combination of itself with some other equations (one or more) in the system does not alter the solution to the system.
3. Interchanging the order of equations in the system does not affect the solution to the system.

These *Elementary Row Operations* are used to form an equivalent triangular system of equations from an original system of equations to be solved.

$$\begin{aligned} 3x_1 - x_2 + 2x_3 &= -3 \\ x_1 + x_2 + x_3 &= -4 \\ 2x_1 + x_2 - x_3 &= -3 \end{aligned}$$

Use 1^{st} equation to eliminate x_1 from 2^{nd} and 3^{rd} equation

$$\begin{aligned} 3x_1 - x_2 + 2x_3 &= -3 \\ \frac{4}{3}x_2 + \frac{1}{3}x_3 &= -3 \\ \frac{5}{3}x_2 - \frac{7}{3}x_3 &= -1 \end{aligned}$$

Use 2^{nd} equation to eliminate x_2 from 3^{rd} equation

$$\begin{aligned} 3x_1 - x_2 + 2x_3 &= -3 \\ \frac{4}{3}x_2 + \frac{1}{3}x_3 &= -3 \\ -\frac{11}{4}x_3 &= \frac{11}{4} \end{aligned}$$

Gaussian Elimination Operation Count for n Equations

The number of multiplications and divisions is called $\mathcal{M}$.

The number of additions and subtractions is called $\mathcal{A}$.

$$\mathcal{M} = \frac{n^3}{3} + n^2 - \frac{n}{3}$$

$$\mathcal{A} = \frac{n^3}{3} + \frac{n^2}{2} - \frac{5n}{6}$$

Errors in Numerical Solutions of Sets of Linear Equations

When terms that are subtracted from each other in a solution method have nearly the same magnitude, computational round off errors can result in large relative errors in the solution.

Computational errors are generally reduced if each equation is scaled as follows:

$$\begin{bmatrix} a_{11} & a_{12} & \cdot & \cdot & \cdot & a_{1n} \\ a_{21} & a_{22} & \cdot & \cdot & \cdot & a_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \cdot & \cdot & \cdot & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_m \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ b_m \end{bmatrix}$$

$$\text{Let} \qquad s_i = \max_j |a_{ij}|, \quad i = 1, 2, ..., n$$

Divide the i^{th} row of the **A** and **b** matrices by s_i.
This is called *scaling.*

In a Gaussian Elimination the element that is used to eliminate its column in the equations that do not contain it is called the *pivot element*

Errors in numerical solutions are generally reduced if pairs of equations are interchanged so the magnitude of the pivot element is the largest one possible.

Scaled Partial Pivoting Rule

Both of the above error-reduction steps can be incorporated in what is called the *Scaled Partial Pivoting Rule.*

1. Start by determining s_i for each row as explained above.
2. At the start of the k';th elimination step, scan the k^{th} column of **A** and determine the integer p such that:

$$\frac{|a_{pk}|}{s_p} \geq \frac{|a_{lk}|}{s_l}, \qquad l = k, k+1, ...n$$

3. If $p \neq k$, then interchange rows p and k.

This procedure removes the need to do the scaling explicitly which can be another source of round off error.

Scaling

$$\begin{bmatrix} 1 & 3 & 4 & 6 \\ 7 & 2 & 3 & 2 \\ 4 & 3 & 8 & 1 \\ 5 & 1 & -2 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 6 \\ 8 \\ 3 \\ 2 \end{bmatrix}$$

$$\Downarrow$$

$$\begin{bmatrix} 1/6 & 1/2 & 2/3 & 1 \\ 1 & 2/7 & 3/7 & 2/7 \\ 1/2 & 3/8 & 1 & 1/8 \\ 1 & 1/5 & -2/5 & 3/5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 8/7 \\ 3/8 \\ 2/5 \end{bmatrix}$$

Pivoting

$$\begin{bmatrix} 1 & 3 & 4 & 6 \\ 7 & 2 & 3 & 2 \\ 4 & 3 & 8 & 1 \\ 5 & 1 & -2 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 6 \\ 8 \\ 3 \\ 2 \end{bmatrix}$$

$$\Downarrow$$

$$\begin{bmatrix} 7 & 2 & 3 & 2 \\ 1 & 3 & 4 & 6 \\ 4 & 3 & 8 & 1 \\ 5 & 1 & -2 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 8 \\ 6 \\ 3 \\ 2 \end{bmatrix}$$

Scaled Partial Pivoting

a_{pi}/s_p	s_i						
1/6	6	1	3	4	6	x_1	b_1
1/3	6	2	6	3	2	x_2	b_2
2/3	6	4	3	6	1	x_3	b_3
1	4	4	2	-2	3	x_4	b_4

Interchange 1st and 4th rows (equations)

Solution of Linear Equations by LU Decomposition

Equation to Solve: $\mathbf{Ax} = \mathbf{b}$

$\mathbf{A}$ is presumed to be non-singular. Suppose we can decompose $\mathbf{A}$ into $\mathbf{A} = \mathbf{LU}$ where $\mathbf{L}$ is lower triangular with diagonal elements equal to 1 and $\mathbf{U}$ is upper triangular. Then the solution is straightforward.

$$\mathbf{LUx} = \mathbf{b}$$

Define: $\mathbf{y} \equiv \mathbf{Ux}$ $\qquad$ $\mathbf{Ly} = \mathbf{b}$

$$\begin{bmatrix} 1 & 0 & 0 \\ L_{21} & 1 & 0 \\ L_{31} & L_{32} & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

$$y_1 = b_1$$

$$L_{21}y_1 + y_2 = b_2 \qquad y_2 = b_2 - L_{21}b_1$$

$$L_{31}y_1 + L_{32}y_2 + y_3 = b_3 \qquad y_3 = b_3 - L_{31}y_1 - L_{32}y_2$$

$$\mathbf{Ux} = \mathbf{y}$$

$$\begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

$$x_3 = y_3/u_{33}$$

$$x_2 = (y_2 - u_{23}x_3)/u_{22}$$

$$x_1 = (y_1 - u_{12}x_2 - u_{13}x_3)/u_{11}$$

Procedure for Factorization of A

A is a nonsingular $n \times n$ matrix. $\mathbf{A} = \mathbf{LU}$. As an example, suppose:

$$\mathbf{A} = \begin{bmatrix} 3 & -1 & 3 \\ 6 & 0 & 9 \\ -12 & 0 & -10 \end{bmatrix}$$

From the second row we subtract $m_2^{(1)} = 2$ times the first row.
From the third row we subtract $m_3^{(1)} = -4$ times the first row.
The result is":

$$\mathbf{U}' = \begin{bmatrix} 3 & -1 & 3 \\ 0 & 2 & 3 \\ 0 & -4 & 2 \end{bmatrix}$$

From the third row we subtract $m_3^{(2)} = -2$ times the second row.
The result is the desired upper triangular matrix:

$$\mathbf{U} = \begin{bmatrix} 3 & -1 & 3 \\ 0 & 2 & 3 \\ 0 & 0 & 8 \end{bmatrix}$$

The lower triangular matrix with 1's on the diagonal is given by the formula $l_{ij} = m_i^{(j)}$ for $i > j$.

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -4 & -2 & 1 \end{bmatrix}$$

This results in $\mathbf{A} = \mathbf{LU}$