Solution of linear systems $\mathbf{A}x = \mathbf{B}$

Professor Henry Arguello

Universidad Industrial de Santander
Colombia

June 16, 2020

High Dimensional Signal Processing Group
www.hdspgroup.com
henarfu@uis.edu.co

Universidad Industrial de Santander

Outline

Coordinate form of a real N-dimensional vector,

$$X = (x_1, x_2, ..., x_N).$$

$x_1, x_2, ...,$ and x_N are the ***components of X***.

- The set consisting of all N-dimensional vectors is called ***N-dimensional space***.
- When a vector is used to denote a point or position in space, it is called a ***position vector***.
- When it is used to denote a movement between two points in space, it is called a ***displacement vector***.

- Two vectors X and Y are said to be equal if and only if each corresponding coordinate is the same

$$X = Y \text{ if and only if } x_j = y_j \text{ for } j = 1, 2, ..., N.$$

- Two vectors X and Y are said to be equal if and only if each corresponding coordinate is the same

$$X = Y \text{ if and only if } x_j = y_j \text{ for } j = 1, 2, ..., N.$$

- The sum of the vectors X and Y is computed component by component, using the definition

$$X + Y = (x_1 + y_1, x_2 + y_2, ..., x_N + y_N).$$

- Two vectors X and Y are said to be equal if and only if each corresponding coordinate is the same

$$X = Y \text{ if and only if } x_j = y_j \text{ for } j = 1, 2, ..., N.$$

- The sum of the vectors X and Y is computed component by component, using the definition

$$X + Y = (x_1 + y_1, x_2 + y_2, ..., x_N + y_N).$$

- The negative of the vector X is obtained by replacing each coordinate with its negative:

$$-X = (-x_1, -x_2, ..., -x_N).$$

- Two vectors X and Y are said to be equal if and only if each corresponding coordinate is the same

$$X = Y \text{ if and only if } x_j = y_j \text{ for } j = 1, 2, ..., N.$$

- The sum of the vectors X and Y is computed component by component, using the definition

$$X + Y = (x_1 + y_1, x_2 + y_2, ..., x_N + y_N).$$

- The negative of the vector X is obtained by replacing each coordinate with its negative:

$$-X = (-x_1, -x_2, ..., -x_N).$$

- The difference $Y - X$ is formed by taking the difference in each coordinate:

$$Y - X = (y_1 - x_1, y_2 - x_2, ..., y_N - x_N).$$

Introduction to Vectors and Matrices

- Vectors in N-dimensional space obey the algebraic property
$$Y - X = Y + (-X).$$

Introduction to Vectors and Matrices

- Vectors in N-dimensional space obey the algebraic property
$$Y - X = Y + (-X).$$
- If c is a real number (scalar), we define ***scalar multiplication*** cX as follows:
$$cX = (cx_1, cx_2, ..., cx_N).$$

- Vectors in N-dimensional space obey the algebraic property

$$Y - X = Y + (-X).$$

- If c is a real number (scalar), we define ***scalar multiplication*** cX as follows:

$$cX = (cx_1, cx_2, ..., cx_N).$$

- If c and d are scalars, then the weighted sum $cX + dY$ is called a ***linear combination*** of X and Y, and we write

$$cX + dY = (cx_1 + dy_1, cx_2 + dy_2, ..., cx_N + dy_N).$$

Introduction to Vectors and Matrices

- Vectors in N-dimensional space obey the algebraic property
$$Y - X = Y + (-X).$$
- If c is a real number (scalar), we define ***scalar multiplication*** cX as follows:
$$cX = (cx_1, cx_2, ..., cx_N).$$
- If c and d are scalars, then the weighted sum $cX + dY$ is called a ***linear combination*** of X and Y, and we write
$$cX + dY = (cx_1 + dy_1, cx_2 + dy_2, ..., cx_N + dy_N).$$
- The ***dot product*** of the two vectors X and Y is a scalar quantity (real number) defined by the equation
$$X \cdot Y = x_1y_1 + x_2y_2 + \cdots + x_Ny_N. \quad (1)$$

The ***norm*** (or ***length***) of the vector X is defined by

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

The ***norm*** (or ***length***) of the vector X is defined by

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

Equation (2) is referred to as the ***Euclidean norm*** (or ***length***) of the vector X. Scalar multiplication cX stretches the vector X when $| c | > 1$ and shrinks the vector when $| c | < 1$. This is shown by using equation (norm):

$$\begin{aligned} \| cX \| &= (c^2 x_1^2 + c^2 x_2^2 + \cdots + c^2 x_N^2)^{1/2}. \\ &= | c | \, (x_1^2 + x_2^2 + \cdots + x_N^2)^{1/2} = | c | \| X \| . \end{aligned}$$

The ***norm*** (or ***length***) of the vector X is defined by

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

Equation (2) is referred to as the ***Euclidean norm*** (or ***length***) of the vector X. Scalar multiplication cX stretches the vector X when $| c | > 1$ and shrinks the vector when $| c | < 1$. This is shown by using equation (norm):

$$\begin{aligned} \| cX \| &= (c^2 x_1^2 + c^2 x_2^2 + \cdots + c^2 x_N^2)^{1/2}. \\ &= | c | \, (x_1^2 + x_2^2 + \cdots + x_N^2)^{1/2} = | c | \| X \| . \end{aligned}$$

An important relationship exists between the dot product and the norm of a vector. If both sides of equation (2) are squared and equation (1) is used, with Y being replaced with X, we have

$$\| X \|^2 = x_1^2 + x_2^2 + \cdots + x_N^2 = X \cdot X.$$

If X and Y are position vectors that locate the two points $(x_1, x_2, ..., x_N)$ and $(y_1, y_2, ..., y_N)$ in N-dimensional space, then the ***displacement vector*** from X to Y is given by the difference

$Y - X$ (displacement from position X to position Y) (3).

If X and Y are position vectors that locate the two points $(x_1, x_2, ..., x_N)$ and $(y_1, y_2, ..., y_N)$ in N-dimensional space, then the ***displacement vector*** from X to Y is given by the difference

$Y - X$ (displacement from position X to position Y) (3).

Notice that if a particle starts at the position X and moves through the displacement $Y - X$, its new position is Y. This can be obtained by the following vector sum:

$$Y = X + (Y - X).$$

If X and Y are position vectors that locate the two points $(x_1, x_2, ..., x_N)$ and $(y_1, y_2, ..., y_N)$ in N-dimensional space, then the ***displacement vector*** from X to Y is given by the difference

$$Y - X \text{ (displacement from position } X \text{ to position } Y\text{) (3).}$$

Notice that if a particle starts at the position X and moves through the displacement $Y - X$, its new position is Y. This can be obtained by the following vector sum:

$$Y = X + (Y - X).$$

Using equation (2) and (3), we can write down the formula for the distance between two points in N-space.

$$\| Y - X \| = \left((y_1 - x_1)^2 + (y_2 - x_2)^2 + \cdots + (y_N - x_N)^2 \right)^{1/2}.$$

When the distance between points is computed using formula above, we say that the points lie in ***N-dimensional Euclidean space***.

Example:

Let $X = (2, -3, 5, -1)$ and $Y = (6, 1, 2, -4)$. The concepts mentioned above are now illustrated for vectors in 4-space.

Example:

Let $X = (2, -3, 5, -1)$ and $Y = (6, 1, 2, -4)$. The concepts mentioned above are now illustrated for vectors in 4-space.

Sum	$X + Y = (8, -2, 7, -5)$
Difference	$X - Y = (-4, -4, 3, 3)$
Scalar multiple	$3X = (6, -9, 15, -3)$
Length	$\parallel X \parallel = (4 + 9 + 25 + 1)^{1/2} = 39^{1/2}$
Dot product	$X \cdot Y = 12 - 3 + 10 + 4 = 23$
Displacement from X to Y	$Y - X = (4, 4, -3, -3)$
Distance from X to Y	$\parallel Y - X \parallel = (16 + 16 + 9 + 9)^{1/2} = 50^{1/2}$

Introduction to Vectors and Matrices

It is sometimes useful to write vectors as columns instead of rows. For example,

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \qquad \text{and} \qquad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

It is sometimes useful to write vectors as columns instead of rows. For example,

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \qquad \text{and} \qquad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

Then the linear combination $cX + dY$ is

$$cX + dY = \begin{bmatrix} cx_1 + dy_1 \\ cx_2 + dy_2 \\ \vdots \\ cx_N + dy_N \end{bmatrix}$$

It is sometimes useful to write vectors as columns instead of rows. For example,

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \quad \text{and} \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

Then the linear combination $cX + dY$ is

$$cX + dY = \begin{bmatrix} cx_1 + dy_1 \\ cx_2 + dy_2 \\ \vdots \\ cx_N + dy_N \end{bmatrix}$$

The set of vectors has a zero element **0**, which is defined by

$$\mathbf{0} = (0, 0, ..., 0).$$

Theorem 1. Vector Algebra.

Suppose that X, Y, and Z are N-dimensional vectors and a and b are scalars (real numbers). The following properties of vector addition and scalar multiplication hold:

$Y + X = X + Y$	commutative property
$0 + X = X + 0$	additive identity
$X - X = X + (-X) = 0$	additive inverse
$(X + Y) + Z = X + (Y + Z)$	associative property
$(a + b)X = aX + bX$	distributive property for scalars
$a(X + Y) = aX + aY$	distributive property for vectors
$a(bX) = (ab)X$	associative property for scalars

Matrices and Two-dimensional Arrays

A matrix is a rectangular array of numbers that is arranged systematically in rows and columns. A matrix having M rows and N columns is called an M x N (read "M by N") matrix. The capital letter A denotes a matrix and the lowercase subscripted letter a_{ij} denotes one of the numbers forming the matrix. We write

$$A = [a_{ij}]_{MxN} \qquad \text{for} \quad 1 \leq i \leq M, 1 \leq j \leq N.$$

where a_{ij} is the number in location (i, j) (i.e.. stored in the ith row and jth column of the matrix). We refer to a_{ij} as the element in location (i, j). In expanded form we write

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2N} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ij} & \cdots & a_{iN} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{M1} & a_{M2} & \cdots & a_{Mj} & \cdots & a_{MN} \end{bmatrix} = A.$$

where a_{ij} is the number in location (i, j) (i.e.. stored in the ith row and jth column of the matrix). We refer to a_{ij} as the element in location (i, j). In expanded form we write

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2N} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ij} & \cdots & a_{iN} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{M1} & a_{M2} & \cdots & a_{Mj} & \cdots & a_{MN} \end{bmatrix} = A.$$

The rows of the M x N matrix A are N-dimensional vectors:

$$V_i = (a_{i1}, a_{i2}, ..., a_{iN}) \qquad \text{for} \quad i = 1, 2, ..., M.$$

Upper-triangular Linear Systems

Back-substitution algorithm: it solves a linear system of equations that has an upper-triangular coefficient matrix.

Back-substitution algorithm: it solves a linear system of equations that has an upper-triangular coefficient matrix.

Definition 1.

An N x N matrix $A = [a_{ij}]$ is called ***upper triangular*** provided that the elements satisfy $a_{ij} = 0$ whenever $i < j$.

Upper-triangular Linear Systems

Back-substitution algorithm: it solves a linear system of equations that has an upper-triangular coefficient matrix.

Definition 1.

An N x N matrix $A = [a_{ij}]$ is called ***upper triangular*** provided that the elements satisfy $a_{ij} = 0$ whenever $i < j$.

If A is an upper-triangular matrix, then $AX = B$ is said to be an ***upper-triangular system*** of linear equations and has the form

$$\begin{aligned}
a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots + a_{1N-1}x_{N-1} + a_{1N}x_N &= b_1 \\
a_{22}x_2 + a_{23}x_3 + \cdots + a_{2N-1}x_{N-1} + a_{2N}x_N &= b_2 \\
a_{33}x_3 + \cdots + a_{3N-1}x_{N-1} + a_{3N}x_N &= b_3 \\
\vdots \qquad\qquad\quad & \;\;\vdots \\
a_{N-1N-1}x_{N-1} + a_{N-1N}x_N &= b_{N-1} \\
a_{NN}x_N &= b_N.
\end{aligned}$$

Upper-triangular Linear Systems

Theorem 2. Back Substitution.

Suppose that $AX = B$ is an upper-triangular system. If

$$a_{kk} \neq 0 \qquad ; \text{for} \quad k = 1, 2, ..., N.$$

then there exists a unique solution.

Constructive Proof. The last equation involves only x_N, so we solve it first:

$$x_N = \frac{b_N}{a_{NN}}.$$

Upper-triangular Linear Systems

Constructive Proof. The last equation involves only x_N, so we solve it first:

$$x_N = \frac{b_N}{a_{NN}}.$$

Now x_N is known and it can be used in the next-to-last equation:

$$x_{N-1} = \frac{b_{N-1} - a_{N-1N}x_N}{a_{N-1N-1}}.$$

Upper-triangular Linear Systems

Constructive Proof. The last equation involves only x_N, so we solve it first:

$$x_N = \frac{b_N}{a_{NN}}.$$

Now x_N is known and it can be used in the next-to-last equation:

$$x_{N-1} = \frac{b_{N-1} - a_{N-1N}x_N}{a_{N-1N-1}}.$$

Now x_N and x_{N-1} are used to find x_{N-2}:

$$x_{N-2} = \frac{b_{N-2} - a_{N-2N-1}x_{N-1} - a_{N-2N}x_N}{a_{N-2N-2}}.$$

Upper-triangular Linear Systems

Constructive Proof. The last equation involves only x_N, so we solve it first:

$$x_N = \frac{b_N}{a_{NN}}.$$

Now x_N is known and it can be used in the next-to-last equation:

$$x_{N-1} = \frac{b_{N-1} - a_{N-1N}x_N}{a_{N-1N-1}}.$$

Now x_N and x_{N-1} are used to find x_{N-2}:

$$x_{N-2} = \frac{b_{N-2} - a_{N-2N-1}x_{N-1} - a_{N-2N}x_N}{a_{N-2N-2}}.$$

Once the values $x_N, X_{N-1}, ..., x_{k+1}$ are known, the general step is

$$x_k = \frac{b_k - \sum_{j=k+1}^{N} a_{kj}x_j}{a_{kk}} \qquad \text{for} \qquad k = N-1, N-2, ..., 1.$$

The Nth equation implies that b_N/a_{NN} is the only possible value of x_N. Then finite induction is used to establish that $x_{N-1}, x_{N-2}, ..., x_1$ are **unique**.

Upper-triangular Linear Systems

Example:

Use back substitution to solve the linear system

$$\begin{aligned} 4x_1 - x_2 + 2x_3 + 3x_4 &= 20 \\ -2x_2 + 7x_3 - 4x_4 &= -7 \\ 6x_3 + 5x_4 &= 4 \\ 3x_4 &= 6. \end{aligned}$$

Example:

Use back substitution to solve the linear system

$$
\begin{aligned}
4x_1 - x_2 + 2x_3 + 3x_4 &= 20 \\
-2x_2 + 7x_3 - 4x_4 &= -7 \\
6x_3 + 5x_4 &= 4 \\
3x_4 &= 6.
\end{aligned}
$$

Solving for x_4 in the last equation yields

Example:

Use back substitution to solve the linear system

$$\begin{aligned} 4x_1 - x_2 + 2x_3 + 3x_4 &= 20 \\ -2x_2 + 7x_3 - 4x_4 &= -7 \\ 6x_3 + 5x_4 &= 4 \\ 3x_4 &= 6. \end{aligned}$$

Solving for x_4 in the last equation yields

$$x_4 = \frac{6}{3} = 2.$$

Upper-triangular Linear Systems

Example:

Use back substitution to solve the linear system

$$\begin{aligned} 4x_1 - x_2 + 2x_3 + 3x_4 &= 20 \\ -2x_2 + 7x_3 - 4x_4 &= -7 \\ 6x_3 + 5x_4 &= 4 \\ 3x_4 &= 6. \end{aligned}$$

Solving for x_4 in the last equation yields

$$x_4 = \frac{6}{3} = 2.$$

Using $x_4 = 2$ in the third equation.

Upper-triangular Linear Systems

Example:

Use back substitution to solve the linear system

$$\begin{aligned} 4x_1 - x_2 + 2x_3 + 3x_4 &= 20 \\ -2x_2 + 7x_3 - 4x_4 &= -7 \\ 6x_3 + 5x_4 &= 4 \\ 3x_4 &= 6. \end{aligned}$$

Solving for x_4 in the last equation yields

$$x_4 = \frac{6}{3} = 2.$$

Using $x_4 = 2$ in the third equation.

$$x_3 = \frac{4 - 5(2)}{6} = -1.$$

Upper-triangular Linear Systems

Example:

Use back substitution to solve the linear system

$$\begin{aligned} 4x_1 - x_2 + 2x_3 + 3x_4 &= 20 \\ -2x_2 + 7x_3 - 4x_4 &= -7 \\ 6x_3 + 5x_4 &= 4 \\ 3x_4 &= 6. \end{aligned}$$

Solving for x_4 in the last equation yields

$$x_4 = \frac{6}{3} = 2.$$

Using $x_4 = 2$ in the third equation.

$$x_3 = \frac{4 - 5(2)}{6} = -1.$$

Now $x_3 = -1$ and $x_4 = 2$ are used to find x_2 in the second equation:

Upper-triangular Linear Systems

Example:

Use back substitution to solve the linear system

$$\begin{aligned} 4x_1 - x_2 + 2x_3 + 3x_4 &= 20 \\ -2x_2 + 7x_3 - 4x_4 &= -7 \\ 6x_3 + 5x_4 &= 4 \\ 3x_4 &= 6. \end{aligned}$$

Solving for x_4 in the last equation yields

$$x_4 = \frac{6}{3} = 2.$$

Using $x_4 = 2$ in the third equation.

$$x_3 = \frac{4 - 5(2)}{6} = -1.$$

Now $x_3 = -1$ and $x_4 = 2$ are used to find x_2 in the second equation:

$$x_2 = \frac{-7 - 7(-1) + 4(2)}{-2} = -4.$$

Upper-triangular Linear Systems

Example:

Use back substitution to solve the linear system

$$\begin{aligned} 4x_1 - x_2 + 2x_3 + 3x_4 &= 20 \\ -2x_2 + 7x_3 - 4x_4 &= -7 \\ 6x_3 + 5x_4 &= 4 \\ 3x_4 &= 6. \end{aligned}$$

Solving for x_4 in the last equation yields

$$x_4 = \frac{6}{3} = 2.$$

Using $x_4 = 2$ in the third equation.

$$x_3 = \frac{4 - 5(2)}{6} = -1.$$

Now $x_3 = -1$ and $x_4 = 2$ are used to find x_2 in the second equation:

$$x_2 = \frac{-7 - 7(-1) + 4(2)}{-2} = -4.$$

Finally, x_1 is obtained using the first equation:

Upper-triangular Linear Systems

Example:

Use back substitution to solve the linear system

$$\begin{aligned} 4x_1 - x_2 + 2x_3 + 3x_4 &= 20 \\ -2x_2 + 7x_3 - 4x_4 &= -7 \\ 6x_3 + 5x_4 &= 4 \\ 3x_4 &= 6. \end{aligned}$$

Solving for x_4 in the last equation yields

$$x_4 = \frac{6}{3} = 2.$$

Using $x_4 = 2$ in the third equation.

$$x_3 = \frac{4 - 5(2)}{6} = -1.$$

Now $x_3 = -1$ and $x_4 = 2$ are used to find x_2 in the second equation:

$$x_2 = \frac{-7 - 7(-1) + 4(2)}{-2} = -4.$$

Finally, x_1 is obtained using the first equation:

$$x_1 = \frac{20 + 1(-4) - 2(-1) - 3(2)}{4} = 3.$$

Gaussian Elimination and Pivoting

Gaussian method: The goal is to construct an equivalent upper-triangular system $UX = Y$ that can be solved. Two linear systems of dimension N x N are said to be ***equivalent*** provided that their solution sets are the same.

Gaussian Elimination and Pivoting

Gaussian method: The goal is to construct an equivalent upper-triangular system $UX = Y$ that can be solved. Two linear systems of dimension N x N are said to be ***equivalent*** provided that their solution sets are the same.

Theorem 3. Elementary Transformations.

The following operations applied to a linear system yield an equivalent system:

(1) Interchanges:	The order of two equations can be changed.
(2) Scaling:	Multiplying an equation by a nonzero constant.
(3) Replacement:	An equation can be replaced by the sum of itself and a nonzero multiple of any other equation.

.

It is common to use (3) by replacing an equation with the difference of that equation and a multiple of another equation.

Example:

Find the parabola $y = A + Bx + Cx^2$ that passes through the three points $(1,1)$, $(2,-1)$, and $(3,1)$.

Example:

Find the parabola $y = A + Bx + Cx^2$ that passes through the three points $(1,1)$, $(2,-1)$, and $(3,1)$.
For each point we obtain an equation relating the value of x to the value of y. The result is the linear system

Example:

Find the parabola $y = A + Bx + Cx^2$ that passes through the three points $(1,1)$, $(2,-1)$, and $(3,1)$.
For each point we obtain an equation relating the value of x to the value of y. The result is the linear system

$$\begin{aligned} A + \ \ B + \ \ C &= \ \ 1 \qquad \text{at}(1,1) \\ A + 2B + 4C &= -1 \qquad \text{at}(2,-1) \\ A + 3B + 9C &= \ \ 1 \qquad \text{at}(3,1). \end{aligned}$$

Example:

Find the parabola $y = A + Bx + Cx^2$ that passes through the three points $(1, 1)$, $(2, -1)$, and $(3, 1)$.
For each point we obtain an equation relating the value of x to the value of y. The result is the linear system

$$\begin{aligned} A + \ \ B + \ \ C &= \ \ 1 \qquad \text{at}(1, 1) \\ A + 2B + 4C &= -1 \qquad \text{at}(2, -1) \\ A + 3B + 9C &= \ \ 1 \qquad \text{at}(3, 1). \end{aligned}$$

The variable A is eliminated from the second and third equations by subtracting the first equation from them. This is an application of the **replacement** transformation, and the resulting equivalent linear system

Example:

Find the parabola $y = A + Bx + Cx^2$ that passes through the three points $(1,1)$, $(2,-1)$, and $(3,1)$.
For each point we obtain an equation relating the value of x to the value of y. The result is the linear system

$$\begin{aligned} A + B + C &= 1 \qquad \text{at}(1,1) \\ A + 2B + 4C &= -1 \qquad \text{at}(2,-1) \\ A + 3B + 9C &= 1 \qquad \text{at}(3,1). \end{aligned}$$

The variable A is eliminated from the second and third equations by subtracting the first equation from them. This is an application of the **replacement** transformation, and the resulting equivalent linear system

$$\begin{aligned} A + B + C &= 1 \\ B + 3C &= -2 \\ 2B + 8C &= 0. \end{aligned}$$

$$\begin{aligned} A + B + C &= 1 \\ B + 3C &= -2 \\ 2B + 8C &= 0. \end{aligned}$$

The variable B is eliminated from the third equation by subtracting from it two times the second equation.

Gaussian Elimination and Pivoting

$$\begin{aligned} A + B + C &= 1 \\ B + 3C &= -2 \\ 2B + 8C &= 0. \end{aligned}$$

The variable B is eliminated from the third equation by subtracting from it two times the second equation.

$$\begin{aligned} A + B + C &= 1 \\ B + 3C &= -2 \\ 2C &= 4. \end{aligned}$$

$$\begin{aligned} A + B + C &= 1 \\ B + 3C &= -2 \\ 2B + 8C &= 0. \end{aligned}$$

The variable B is eliminated from the third equation by subtracting from it two times the second equation.

$$\begin{aligned} A + B + C &= 1 \\ B + 3C &= -2 \\ 2C &= 4. \end{aligned}$$

We arrive at the equivalent upper-triangular system. The back-substitution algorithm is now used to find the coefficients $C = 4/2 = 2, B = -2 - 3(2) = -8$, and $A = 1 - (-8) - 2 = 7$, and the equation of the parabola is $y = 7 - 8x + 2x^2$.

Gaussian Elimination and Pivoting

The ***augmented matrix*** is denoted $[A|B]$ and the linear system is represented as follows:

$$[A|B] = \left[\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1N} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2N} & b_2 \\ \vdots & \vdots & & \vdots & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NN} & b_N \end{array}\right].$$

Gaussian Elimination and Pivoting

The ***augmented matrix*** is denoted $[A|B]$ and the linear system is represented as follows:

$$[A|B] = \left[\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1N} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2N} & b_2 \\ \vdots & \vdots & & \vdots & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NN} & b_N \end{array}\right].$$

Theorem 4. Elementary Row Operations.

The following operations applied to the augmented matrix yield an equivalent linear system.:

(1) Interchanges: The order of two rows can be changed.
(2) Scaling: Multiplying a row by a nonzero constant.
(3) Replacement: The row can be replaced by the sum of that row and a nonzero multiple of any other row; that is:
. row$_r$ $=$ row$_r$$-m_{rp}$ x row$_p$.
.

Definition 2. Pivot.

The number a_{11} in the coefficient matrix A that is used to eliminate a_{kr}, where $k = r+1, r+2, ..., N$, is called the rth ***pivotal element***, and the rth row is called the ***pivot row***.

The following example illustrates how to use the operations in Theorem 4 to obtain an equivalent upper-triangular system $UX - Y$ from a linear system $AX = B$ where A is an N x N matrix.

Gaussian Elimination and Pivoting

The following example illustrates how to use the operations in Theorem 4 to obtain an equivalent upper-triangular system $UX - Y$ from a linear system $AX = B$ where A is an N x N matrix.

The following example illustrates how to use the operations in Theorem 4 to obtain an equivalent upper-triangular system $UX - Y$ from a linear system $AX = B$ where A is an N x N matrix.

Example:

Express the following system in augmented matrix form and find an equivalent upper-triangular system and the solution.

$$\begin{aligned} x_1 + 2x_2 + x_3 + 4x_4 &= 13 \\ 2x_1 + 0x_2 + 4x_3 + 3x_4 &= 28 \\ 4x_1 + 2x_x + 2x_3 + x_4 &= 20 \\ -3x_1 + x_2 + 3x_3 + 2x_4 &= 6. \end{aligned}$$

The augmented matrix is

The following example illustrates how to use the operations in Theorem 4 to obtain an equivalent upper-triangular system $UX - Y$ from a linear system $AX = B$ where A is an N x N matrix.

Example:

Express the following system in augmented matrix form and find an equivalent upper-triangular system and the solution.

$$\begin{aligned} x_1 + 2x_2 + x_3 + 4x_4 &= 13 \\ 2x_1 + 0x_2 + 4x_3 + 3x_4 &= 28 \\ 4x_1 + 2x_x + 2x_3 + x_4 &= 20 \\ -3x_1 + x_2 + 3x_3 + 2x_4 &= 6. \end{aligned}$$

The augmented matrix is

$$\begin{array}{r} pivot \rightarrow \\ m_{21} = 2 \\ m_{31} = 4 \\ m_{41} = -3 \end{array} \left[\begin{array}{cccc|c} 1 & 2 & 1 & 4 & 13 \\ 2 & 0 & 4 & 3 & 28 \\ 4 & 2 & 2 & 1 & 20 \\ -3 & 1 & 3 & 2 & 6 \end{array}\right].$$

Gaussian Elimination and Pivoting

$$
\begin{array}{r}
pivot \rightarrow \\
m_{21} = 2 \\
m_{31} = 4 \\
m_{41} = -3
\end{array}
\left[\begin{array}{cccc|c}
1 & 2 & 1 & 4 & 13 \\
2 & 0 & 4 & 3 & 28 \\
4 & 2 & 2 & 1 & 20 \\
-3 & 1 & 3 & 2 & 6
\end{array}\right].
$$

The first row is used to eliminate elements in the first column below the diagonal. We refer to the first row as the **pivotal row** and the element $a_{11} = 1$ is called the **pivotal element.** The values m_{k1} are the **multiples** of row 1 that are to be subtracted from row k for $k = 2, 3, 4$. the result after elimination is

Gaussian Elimination and Pivoting

$$\begin{array}{r} pivot \rightarrow \\ m_{21} = 2 \\ m_{31} = 4 \\ m_{41} = -3 \end{array} \left[\begin{array}{cccc|c} 1 & 2 & 1 & 4 & 13 \\ 2 & 0 & 4 & 3 & 28 \\ 4 & 2 & 2 & 1 & 20 \\ -3 & 1 & 3 & 2 & 6 \end{array}\right].$$

The first row is used to eliminate elements in the first column below the diagonal. We refer to the first row as the **pivotal row** and the element $a_{11} = 1$ is called the **pivotal element.** The values m_{k1} are the **multiples** of row 1 that are to be subtracted from row k for $k = 2, 3, 4$. the result after elimination is

$$\begin{array}{r} \\ pivot \rightarrow \\ m_{32} = 1.5 \\ m_{42} = -1.75 \end{array} \left[\begin{array}{cccc|c} 1 & 2 & 1 & 4 & 13 \\ 0 & -4 & 2 & -5 & 2 \\ 0 & -6 & -2 & -15 & -32 \\ 0 & 7 & 6 & 14 & 45 \end{array}\right].$$

Gaussian Elimination and Pivoting

$$\begin{array}{r} \\ pivot \rightarrow \\ m_{32} = 1.5 \\ m_{42} = -1.75 \end{array} \left[\begin{array}{cccc|c} 1 & 2 & 1 & 4 & 13 \\ 0 & -4 & 2 & -5 & 2 \\ 0 & -6 & -2 & -15 & -32 \\ 0 & 7 & 6 & 14 & 45 \end{array}\right].$$

The second row is used to eliminate elements in the second column that lie below the diagonal. The second row is the pivotal row and the values m_{k2} are the multiplies of row 2 that are to be subtracted from row k for $k = 3, 4$. The result after elimination is

$$\begin{array}{l} \\ pivot \rightarrow \\ m_{32} = 1.5 \\ m_{42} = -1.75 \end{array} \left[\begin{array}{cccc|c} 1 & 2 & 1 & 4 & 13 \\ 0 & -4 & 2 & -5 & 2 \\ 0 & -6 & -2 & -15 & -32 \\ 0 & 7 & 6 & 14 & 45 \end{array}\right].$$

The second row is used to eliminate elements in the second column that lie below the diagonal. The second row is the pivotal row and the values m_{k2} are the multiplies of row 2 that are to be subtracted from row k for $k = 3, 4$. The result after elimination is

$$\begin{array}{l} \\ \\ pivot \rightarrow \\ m_{43} = -1.9 \end{array} \left[\begin{array}{cccc|c} 1 & 2 & 1 & 4 & 13 \\ 0 & -4 & 2 & -5 & 2 \\ 0 & 0 & -5 & -7.5 & -35 \\ 0 & 0 & 9.5 & 5.25 & 48.5 \end{array}\right].$$

Gaussian Elimination and Pivoting

$$\begin{array}{r} \\ \\ pivot \rightarrow \\ m_{43} = -1.9 \end{array} \left[\begin{array}{cccc|c} 1 & 2 & 1 & 4 & 13 \\ 0 & -4 & 2 & -5 & 2 \\ 0 & 0 & -5 & -7.5 & -35 \\ 0 & 0 & 9.5 & 5.25 & 48.5 \end{array}\right].$$

Finally, the multiple $m_{43} = -1.9$ of the third row is subtracted from the fourth row and the result is the upper-triangular system

Gaussian Elimination and Pivoting

$$\begin{array}{r} \\ \\ pivot \rightarrow \\ m_{43} = -1.9 \end{array} \left[\begin{array}{cccc|c} 1 & 2 & 1 & 4 & 13 \\ 0 & -4 & 2 & -5 & 2 \\ 0 & 0 & -5 & -7.5 & -35 \\ 0 & 0 & 9.5 & 5.25 & 48.5 \end{array}\right].$$

Finally, the multiple $m_{43} = -1.9$ of the third row is subtracted from the fourth row and the result is the upper-triangular system

$$\left[\begin{array}{cccc|c} 1 & 2 & 1 & 4 & 13 \\ 0 & -4 & 2 & -5 & 2 \\ 0 & 0 & -5 & -7.5 & -35 \\ 0 & 0 & 0 & -9 & -18 \end{array}\right].$$

$$\begin{array}{l} \\ \\ pivot \rightarrow \\ m_{43} = -1.9 \end{array} \left[\begin{array}{cccc|c} 1 & 2 & 1 & 4 & 13 \\ 0 & -4 & 2 & -5 & 2 \\ 0 & 0 & -5 & -7.5 & -35 \\ 0 & 0 & 9.5 & 5.25 & 48.5 \end{array}\right].$$

Finally, the multiple $m_{43} = -1.9$ of the third row is subtracted from the fourth row and the result is the upper-triangular system

$$\left[\begin{array}{cccc|c} 1 & 2 & 1 & 4 & 13 \\ 0 & -4 & 2 & -5 & 2 \\ 0 & 0 & -5 & -7.5 & -35 \\ 0 & 0 & 0 & -9 & -18 \end{array}\right].$$

The back-substitution algorithm can be used to solve the system.

$$x_4 = 2, \quad x_3 = 4, \quad x_2 = -1 \quad x_1 = 3.$$

The process described above is called ***Gaussian elimination***.

Gaussian Elimination with Back Substitution

If A is an N x N nonsingular matrix, then there exists a system $UX = Y$, equivalent to $AX = B$, where U is an upper-triangular matrix with $u_{kk} \neq 0$. After U and Y are constructed, back substitution can be used to solve $UX = Y$ for X.

Gaussian Elimination and Pivoting

Gaussian Elimination with Back Substitution

If A is an N x N nonsingular matrix, then there exists a system $UX = Y$, equivalent to $AX = B$, where U is an upper-triangular matrix with $u_{kk} \neq 0$. After U and Y are constructed, back substitution can be used to solve $UX = Y$ for X.

Proof. We will use the augmented matrix with B stored in column $N+1$:

$$AX = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1N}^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} & a_{23}^{(1)} & \cdots & a_{2N}^{(1)} \\ a_{31}^{(1)} & a_{32}^{(1)} & a_{33}^{(1)} & \cdots & a_{3N}^{(13)} \\ \vdots & \vdots & \vdots & & \vdots \\ a_{N1}^{(1)} & a_{N2}^{(1)} & a_{N3}^{(1)} & \cdots & a_{NN}^{(1)} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} a_{1N+1}^{(1)} \\ a_{2N+1}^{(1)} \\ a_{3N+1}^{(1)} \\ \vdots \\ a_{NN+1}^{(1)} \end{bmatrix} = B.$$

Gaussian Elimination and Pivoting

Then we will construct an equivalent upper-triangular system $UX = Y$:

$$UX = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1N}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2N}^{(2)} \\ 0 & 0 & a_{33}^{(3)} & \cdots & a_{3N}^{(3)} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & a_{NN}^{(N)} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} a_{1N+1}^{(1)} \\ a_{2N+1}^{(2)} \\ a_{3N+1}^{(3)} \\ \vdots \\ a_{NN+1}^{(N)} \end{bmatrix} = Y.$$

Then we will construct an equivalent upper-triangular system $UX = Y$:

$$UX = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1N}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2N}^{(2)} \\ 0 & 0 & a_{33}^{(3)} & \cdots & a_{3N}^{(3)} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & a_{NN}^{(N)} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} a_{1N+1}^{(1)} \\ a_{2N+1}^{(2)} \\ a_{3N+1}^{(3)} \\ \vdots \\ a_{NN+1}^{(N)} \end{bmatrix} = Y.$$

$Step 1$. Store the coefficients in the augmented matrix. The superscript on $a_{rc}^{(1)}$ means that this is the first time that a number is stored in location (r, c):

$$\left[\begin{array}{ccccc|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1N}^{(1)} & a_{1N+1}^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} & a_{23}^{(1)} & \cdots & a_{2N}^{(1)} & a_{2N+1}^{(1)} \\ a_{31}^{(1)} & a_{32}^{(1)} & a_{33}^{(1)} & \cdots & a_{3N}^{(1)} & a_{3N+1}^{(1)} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ a_{N1}^{(1)} & a_{N2}^{(1)} & a_{N3}^{(1)} & \cdots & a_{NN}^{(1)} & a_{NN+1}^{(1)} \end{array}\right].$$

*Step*2. If necessary, switch rows so that $a_{11}^{(1)} \neq 0$; then eliminate x_1 in rows 2 through$N$. In this process, m_{r1} is the multiple of row 1 that is subtracted from row r.

for $r = 2 : N$
  $m_{r1} = a_{r1}^{(1)}/a_{11}^{(1)};$
  $a_{r1}^{(2)} = 0;$
  for $c = 2 : N + 1$
    $a_{rc}^{(2)} = a_{rc}^{(1)} - m_{r1} * a_{1c}^{(1)};$
  end
end

Gaussian Elimination and Pivoting

*Step*2. If necessary, switch rows so that $a_{11}^{(1)} \neq 0$; then eliminate x_1 in rows 2 through$N$. In this process, m_{r1} is the multiple of row 1 that is subtracted from row r.

for $r = 2 : N$
  $m_{r1} = a_{r1}^{(1)} / a_{11}^{(1)}$;
  $a_{r1}^{(2)} = 0$;
  for $c = 2 : N + 1$
    $a_{rc}^{(2)} = a_{rc}^{(1)} - m_{r1} * a_{1c}^{(1)}$;
  end
end

The new elements are written $a_{rc}^{(2)}$ to indicate that this is the second time that a number has been stored in the matrix at location (r, c). The result after step 2 is

$$\left[\begin{array}{ccccc|c}
a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1N}^{(1)} & a_{1N+1}^{(1)} \\
0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2N}^{(2)} & a_{2N+1}^{(2)} \\
0 & a_{32}^{(2)} & a_{33}^{(2)} & \cdots & a_{3N}^{(2)} & a_{3N+1}^{(2)} \\
\vdots & \vdots & \vdots & & \vdots & \vdots \\
0 & a_{N2}^{(2)} & a_{N3}^{(2)} & \cdots & a_{NN}^{(2)} & a_{NN+1}^{(2)}
\end{array}\right].$$

Gaussian Elimination and Pivoting

*Step*3. If necessary, switch the second row with some row below it so that $a_{22}^{(2)} \neq 0$; then eliminate x_2 in rows 3 through N. In this process, m_{r2} is the multiple of row 2 that is subtracted from row r.

for $r = 3 : N$
    $m_{r2} = a_{r2}^{(2)} / a_{22}^{(2)};$
    $a_{r2}^{(3)} = 0;$
    for $c = 3 : N + 1$
        $a_{rc}^{(3)} = a_{rc}^{(2)} - m_{r2} * a_{2c}^{(2)};$
    end
end

*Step*3. If necessary, switch the second row with some row below it so that $a_{22}^{(2)} \neq 0$; then eliminate x_2 in rows 3 through N. In this process, m_{r2} is the multiple of row 2 that is subtracted from row r.

for $r = 3 : N$
    $m_{r2} = a_{r2}^{(2)} / a_{22}^{(2)};$
    $a_{r2}^{(3)} = 0;$
    for $c = 3 : N + 1$
        $a_{rc}^{(3)} = a_{rc}^{(2)} - m_{r2} * a_{2c}^{(2)};$
    end
end

The new elements are written $a_{rc}^{(3)}$ to indicate that this is the third time that a number has been stored in the matrix at location (r, c). The result is

$$\left[\begin{array}{cccccc|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1N}^{(1)} & a_{1N+1}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2N}^{(2)} & a_{2N+1}^{(2)} \\ 0 & 0 & a_{33}^{(3)} & \cdots & a_{3N}^{(3)} & a_{3N+1}^{(3)} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & a_{N3}^{(3)} & \cdots & a_{NN}^{(3)} & a_{NN+1}^{(3)} \end{array}\right].$$

Gaussian Elimination and Pivoting

Step $p+1$. This is the general step. If necessary, switch row p with some row beneath it so that $a_{pp}^{(p)} \neq 0$; then eliminate x_p in rows $p+1$ through N. Here m_{rp} is the multiple of row p that is subtracted from row r.

for $r = p+1 : N$
    $m_{rp} = a_{rp}^{(p)} / a_{pp}^{(p)}$;
    $a_{rp}^{(p+1)} = 0$;
    for $c = p+1 : N+1$
        $a_{rc}^{(p+1)} = a_{rc}^{(p)} - m_{rp} * a_{pc}^{(p)}$;
    end
end

Gaussian Elimination and Pivoting

Step $p+1$. This is the general step. If necessary, switch row p with some row beneath it so that $a_{pp}^{(p)} \neq 0$; then eliminate x_p in rows $p+1$ through N. Here m_{rp} is the multiple of row p that is subtracted from row r.

for $r = p+1 : N$
  $m_{rp} = a_{rp}^{(p)} / a_{pp}^{(p)}$;
  $a_{rp}^{(p+1)} = 0$;
  for $c = p+1 : N+1$
    $a_{rc}^{(p+1)} = a_{rc}^{(p)} - m_{rp} * a_{pc}^{(p)}$;
  end
end

The final result after x_{N-1} has been eliminated from row N is

$$\left[\begin{array}{ccccc|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1N}^{(1)} & a_{1N+1}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2N}^{(2)} & a_{2N+1}^{(2)} \\ 0 & 0 & a_{33}^{(3)} & \cdots & a_{3N}^{(3)} & a_{3N+1}^{(3)} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & a_{NN}^{(N)} & a_{NN+1}^{(N)} \end{array}\right].$$

The upper-triangularization process is now complete.

Pivoting to Avoid $a_{pp}^{(p)} = 0$

If $a_{pp}^{(p)} = 0$, row p cannot be used to eliminate the elements in column p below the main diagonal. It is necessary to find row k, where $a_{kp}^{(p)} \neq 0$ and $k > p$, and then interchange row p and row k so that a nonzero pivot element is obtained. This process is called pivoting and the criterion for deciding which row to choose is called a pivoting strategy.

The ***trivial pivoting*** strategy is as follows. If $a_{pp}^{(p)} \neq 0$, do not switch rows. If $a_{pp}^{(p)} = 0$, locate the first row below p in which $a_{kp}^{(p)} \neq 0$ and switch rows k and p. This will result in a new element $a_{pp}^{(p)} \neq 0$, which is a nonzero pivot element.

Pivoting to Reduce Error

Because the computer uses fixed-precision arithmetic, it is possible that a small error will be introduced each time that an arithmetic operation is performed. The following example illustrates how the use of the trivial pivoting strategy in Gaussian elimination can lead to significant error in the solution of a linear system of equations.

Example:

The values $x_1 = x_2 = 1.000$ are the solutions to

$$\begin{aligned} 1.133x_1 + 5.281x_2 &= 6.414 \\ 24.14x_1 - 1.210x_2 &= 22.93. \end{aligned}$$

Example:

The values $x_1 = x_2 = 1.000$ are the solutions to

$$\begin{aligned} 1.133x_1 + 5.281x_2 &= 6.414 \\ 24.14x_1 - 1.210x_2 &= 22.93. \end{aligned}$$

Use four-digit arithmetic and Gaussian elimination with trivial pivoting to find a computed approximate solution to the system. The multiple $m_{21} = 24.14/1.133 = 21.31$ of row 1 is to be subtracted from row 2 to obtain the upper-triangular system.

Using four digits in the calculations, we obtain the new coefficients

$$\begin{aligned} a_{22}^{(2)} &= -1.210 - 21.31(5.281) = -1.210 - 112.5 = -113.7 \\ a_{23}^{(2)} &= 22.93 - 21.31(6.414) = 22.93 - 136.7 = -113.8. \end{aligned}$$

$$a_{22}^{(2)} = -1.210 - 21.31(5.281) = -1.210 - 112.5 = -113.7$$
$$a_{23}^{(2)} = 22.93 - 21.31(6.414) = 22.93 - 136.7 = -113.8.$$

The computed upper-triangular system is

$$1.133x_1 + 5.281x_2 = 6.414$$
$$-113.7x_2 = -113.8.$$

$$a_{22}^{(2)} = -1.210 - 21.31(5.281) = -1.210 - 112.5 = -113.7$$
$$a_{23}^{(2)} = 22.93 - 21.31(6.414) = 22.93 - 136.7 = -113.8.$$

The computed upper-triangular system is

$$\begin{aligned} 1.133x_1 + 5.281x_2 &= 6.414 \\ -113.7x_2 &= -113.8. \end{aligned}$$

Back substitution is used to compute $x_2 = -113.8/(-113.7) = 1.001$, and $x_1 = (6.414 - 5.281(1.001))/(1.133) = (6.414 - 5.286)/(1.233) = 0.9956$.

The error in the solution of the linear system is due to the magnitude of the multiplier $m_{21} = 21.31$. In the next example the magnitude of the multiplier m_{21} is reduced by first interchanging the first and second equations in the linear system and the using the trivial pivoting strategy in Gaussian elimination to solve the system.

Ill conditioning

A matrix A is called ***ill conditioned*** if there exists a matrix B for which small perturbations in the coefficients of A or B will produce large changes in $X = A^{-1}B$. The system $AX = B$ is said to be ill conditioned when A is ill conditioned. In this case, numerical methods for computing an approximate solution are prone to have more error.

One circumstance involving ill conditioning occurs when A is "nearly singular" and the determinant of A is close to zero. Ill conditioning can also occur in systems ot two equations when two lines are nearly parallel (or in three equations when three planes are nearly parallel). A consequence of ill conditioning is that substitution of erroneous values may appear to be genuine solutions.

For example, consider the two equations

$$\begin{aligned} x + 2y - 2.00 &= 0 \\ 2x + 3y - 3.40 &= 0. \end{aligned}$$

For example, consider the two equations

$$x + 2y - 2.00 = 0$$
$$2x + 3y - 3.40 = 0.$$

Substitution of $x_0 = 1.00$ and $y_0 = 0.48$ into these equations "almost produces zeros":

$$1 + 2(0.48) - 2.00 = 1.96 - 2.00 = -0.04 \approx 0$$
$$2 + 3(0.48) - 3.40 = 3.44 - 3.30 = 0.04 \approx 0.$$

For example, consider the two equations

$$x + 2y - 2.00 = 0$$
$$2x + 3y - 3.40 = 0.$$

Substitution of $x_0 = 1.00$ and $y_0 = 0.48$ into these equations "almost produces zeros":

$$1 + 2(0.48) - 2.00 = 1.96 - 2.00 = -0.04 \approx 0$$
$$2 + 3(0.48) - 3.40 = 3.44 - 3.30 = 0.04 \approx 0.$$

Here the discrepancy from 0 is only ± 0.04. However, the true solution to the this linear system is $x = 0.8$ and $y = 0.6$, so the errors in the approximate solution are $x - x_0 = 0.80 - 1.00 = -0.20$ and $y - y_0 = 0.60 - 0.48 = 0.12$. Thus, merely substituting values into a set of equations is not a reliable test for accuracy.

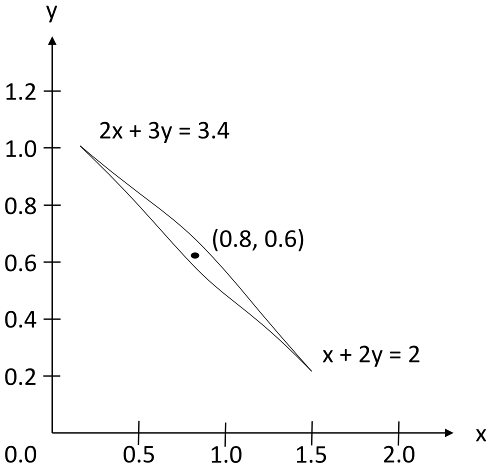

Figure: A region where two equations are "almost satisfied".

The rhombus-shaped region R in figure represents a set where both equations are "almost satisfied":

$$R = \{(x, y).|x + 2y - 2.00| < 0.1 \text{ and } |2x + 3y - 3.40| < 0.2\}.$$

The rhombus-shaped region R in figure represents a set where both equations are "almost satisfied":

$$R = \{(x, y).|x + 2y - 2.00| < 0.1 \text{ and } |2x + 3y - 3.40| < 0.2\}.$$

There are points in R that are far away from the solution point $(0.8, 0.6)$ and yet produce small values when substituted into the equations. If it is suspected that a linear system is ill conditioned, computations should be carried out in multiple-precision arithmetic.
Ill conditioning has more drastic consequences when several equations are in volved. Consider the problem of finding the cubic polynomial $y = c_1x^3 + c_2x^2 + c_3x + c_4$ that passes through the four points $(2, 8), (3, 27), (4, 64)$, and $(5, 125)$ (clearly $y = x^3$ is the desired cubic polynomial).

Triangular Factorization

A can be factorized into the product of a lower-triangular matrix L that has 1's along the main diagonal and an upper-triangular matrix U with nonzero diagonal elements. For ease of notation we illustrate the concepts with matrices of dimension 4 x 4, but they apply to an arbitrary system of dimension N x N.

A can be factorized into the product of a lower-triangular matrix L that has 1's along the main diagonal and an upper-triangular matrix U with nonzero diagonal elements. For ease of notation we illustrate the concepts with matrices of dimension 4 x 4, but they apply to an arbitrary system of dimension N x N.

Definition 3.

The nonsingular matrix A has a ***triangular factorization*** if it can be expressed as the product of a lower-triangular matrix L and an upper-triangular matrix U:

$$A = LU.$$

In matrix form, this is written as

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ m_{21} & 1 & 0 & 0 \\ m_{31} & m_{32} & 1 & 0 \\ m_{41} & m_{42} & m_{43} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ 0 & u_{22} & u_{23} & u_{24} \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{bmatrix}.$$

Solution of a Linear System

Suppose that the coefficient matrix A for the linear system $AX = B$ has a triangular factorization; then the solution to

$$LUX = B$$

Solution of a Linear System

Suppose that the coefficient matrix A for the linear system $AX = B$ has a triangular factorization; then the solution to

$$LUX = B$$

can be obtained by defining $Y = UX$ and then solving two systems:

first solve $LY = B$ for Y; then solve $UX = Y$ for X.

Triangular Factorization

Solution of a Linear System

Suppose that the coefficient matrix A for the linear system $AX = B$ has a triangular factorization; then the solution to

$$LUX = B$$

can be obtained by defining $Y = UX$ and then solving two systems:

first solve $LY = B$ for Y; then solve $UX = Y$ for X.

In equation form, we must first solve the lower-triangular system

$$\begin{aligned}
y_1 \qquad\qquad\qquad\qquad\qquad\qquad &= b_1 \\
m_{21}y_1 + \quad y_2 \qquad\qquad\qquad\qquad &= b_2 \\
m_{31}y_1 + m_{32}y_2 + \quad y_3 \qquad\quad &= b_3 \\
m_{41}y_1 + m_{42}y_2 + m_{43}y_3 + y_4 &= b_4
\end{aligned}$$

Triangular Factorization

Solution of a Linear System

Suppose that the coefficient matrix A for the linear system $AX = B$ has a triangular factorization; then the solution to

$$LUX = B$$

can be obtained by defining $Y = UX$ and then solving two systems:

first solve $LY = B$ for Y; then solve $UX = Y$ for X.

In equation form, we must first solve the lower-triangular system

$$\begin{aligned}
y_1 \qquad\qquad\qquad\qquad\qquad\qquad &= b_1 \\
m_{21}y_1 + \quad y_2 \qquad\qquad\qquad\qquad &= b_2 \\
m_{31}y_1 + m_{32}y_2 + \quad y_3 \qquad\qquad &= b_3 \\
m_{41}y_1 + m_{42}y_2 + m_{43}y_3 + y_4 &= b_4
\end{aligned}$$

to obtain y_1, y_2, y_3, and y_4 and use them in solving the upper-triangular system

$$\begin{aligned}
u_{11}x_1 + u_{12}x_2 + u_{13}x_3 + u_{14}x_4 &= y_1 \\
u_{22}x_2 + u_{23}x_3 + u_{24}x_4 &= y_2 \\
u_{33}x_3 + u_{34}x_4 &= y_3 \\
u_{44}x_4 &= y_4.
\end{aligned}$$

Triangular Factorization

Example:

Solve the system below using the triangular factorization technique.

$$\begin{aligned}
x_1 + 2x_2 + 4x_3 + x_4 &= 21\\
2x_1 + 8x_2 + 6x_3 + 4x_4 &= 52\\
3x_1 + 10x_2 + 8x_3 + 8x_4 &= 79\\
4x_1 + 12x_2 + 10x_3 + 6x_4 &= 82.
\end{aligned}$$

Triangular Factorization

Example:

Solve the system below using the triangular factorization technique.

$$\begin{aligned} x_1 + \ \ 2x_2 + \ \ 4x_3 + \ \ x_4 &= 21 \\ 2x_1 + \ \ 8x_2 + \ \ 6x_3 + 4x_4 &= 52 \\ 3x_1 + 10x_2 + \ \ 8x_3 + 8x_4 &= 79 \\ 4x_1 + 12x_2 + 10x_3 + 6x_4 &= 82. \end{aligned}$$

The triangular factorization of the matrix is,

Triangular Factorization

Example:

Solve the system below using the triangular factorization technique.

$$\begin{aligned} x_1 + \ \ 2x_2 + \ \ 4x_3 + \ \ x_4 &= 21 \\ 2x_1 + \ \ 8x_2 + \ \ 6x_3 + 4x_4 &= 52 \\ 3x_1 + 10x_2 + \ \ 8x_3 + 8x_4 &= 79 \\ 4x_1 + 12x_2 + 10x_3 + 6x_4 &= 82. \end{aligned}$$

The triangular factorization of the matrix is,

$$A = \begin{bmatrix} 1 & 2 & 4 & 1 \\ 2 & 8 & 6 & 4 \\ 3 & 10 & 8 & 8 \\ 4 & 12 & 10 & 6 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 1 & 1 & 0 \\ 4 & 1 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 4 & 1 \\ 0 & 4 & -2 & 2 \\ 0 & 0 & -2 & 3 \\ 0 & 0 & 0 & -6 \end{bmatrix} = LU.$$

Triangular Factorization

Example:

Solve the system below using the triangular factorization technique.

$$\begin{aligned} x_1 + \ \ 2x_2 + \ \ 4x_3 + \ \ x_4 &= 21 \\ 2x_1 + \ \ 8x_2 + \ \ 6x_3 + 4x_4 &= 52 \\ 3x_1 + 10x_2 + \ \ 8x_3 + 8x_4 &= 79 \\ 4x_1 + 12x_2 + 10x_3 + 6x_4 &= 82. \end{aligned}$$

The triangular factorization of the matrix is,

$$A = \begin{bmatrix} 1 & 2 & 4 & 1 \\ 2 & 8 & 6 & 4 \\ 3 & 10 & 8 & 8 \\ 4 & 12 & 10 & 6 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 1 & 1 & 0 \\ 4 & 1 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 4 & 1 \\ 0 & 4 & -2 & 2 \\ 0 & 0 & -2 & 3 \\ 0 & 0 & 0 & -6 \end{bmatrix} = LU.$$

Use the forward-substitution method to solve $LY = B$:

Triangular Factorization

Example:

Solve the system below using the triangular factorization technique.

$$\begin{aligned} x_1 + \ \ 2x_2 + \ \ 4x_3 + \ \ x_4 &= 21 \\ 2x_1 + \ \ 8x_2 + \ \ 6x_3 + 4x_4 &= 52 \\ 3x_1 + 10x_2 + \ \ 8x_3 + 8x_4 &= 79 \\ 4x_1 + 12x_2 + 10x_3 + 6x_4 &= 82. \end{aligned}$$

The triangular factorization of the matrix is,

$$A = \begin{bmatrix} 1 & 2 & 4 & 1 \\ 2 & 8 & 6 & 4 \\ 3 & 10 & 8 & 8 \\ 4 & 12 & 10 & 6 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 1 & 1 & 0 \\ 4 & 1 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 4 & 1 \\ 0 & 4 & -2 & 2 \\ 0 & 0 & -2 & 3 \\ 0 & 0 & 0 & -6 \end{bmatrix} = LU.$$

Use the forward-substitution method to solve $LY = B$:

$$\begin{aligned} y_1 \qquad\qquad\qquad &= 21 \\ 2y_1 + y_2 \qquad\qquad &= 52 \\ 3y_1 + y_2 + \ \ y_3 \qquad &= 79 \\ 4y_1 + y_2 + 2y_3 + y_4 &= 82. \end{aligned}$$

Triangular Factorization

$$\begin{aligned}
y_1 &= 21 \\
2y_1 + y_2 &= 52 \\
3y_1 + y_2 + \quad y_3 &= 79 \\
4y_1 + y_2 + 2y_3 + y_4 &= 82.
\end{aligned}$$

Compute the values $y_1 = 21, y_2 = 52-2(21) = 10, y_3 = 79-3(21)-10 = 6$, and $y_4 = 82 - 4(21) - 10 - 2(6) = -24$, or $Y = [21 \;\; 10 \;\; 6 \;\; -24]'$. Next write the system $UX = Y$:

Triangular Factorization

$$\begin{aligned}
y_1 \qquad\qquad\qquad &= 21 \\
2y_1 + y_2 \qquad\qquad &= 52 \\
3y_1 + y_2 + \quad y_3 \qquad &= 79 \\
4y_1 + y_2 + 2y_3 + y_4 &= 82.
\end{aligned}$$

Compute the values $y_1 = 21, y_2 = 52-2(21) = 10, y_3 = 79-3(21)-10 = 6$, and $y_4 = 82-4(21)-10-2(6) = -24$, or $Y = [21 \;\; 10 \;\; 6 \;\; -24]'$. Next write the system $UX = Y$:

$$\begin{aligned}
x_1 + 2x_2 + 4x_3 + \quad x_4 &= \quad 21 \\
4x_2 - 2x_3 + 2x_4 &= \quad 10 \\
-2x_3 + 3x_4 &= \quad 6 \\
-6x_4 &= -24.
\end{aligned}$$

Triangular Factorization

$$
\begin{aligned}
y_1 \qquad\qquad\qquad\quad &= 21 \\
2y_1 + y_2 \qquad\qquad\quad &= 52 \\
3y_1 + y_2 + \quad y_3 \qquad &= 79 \\
4y_1 + y_2 + 2y_3 + y_4 &= 82.
\end{aligned}
$$

Compute the values $y_1 = 21, y_2 = 52 - 2(21) = 10, y_3 = 79 - 3(21) - 10 = 6$, and $y_4 = 82 - 4(21) - 10 - 2(6) = -24$, or $Y = [21 \;\; 10 \;\; 6 \;\; -24]'$. Next write the system $UX = Y$:

$$
\begin{aligned}
x_1 + 2x_2 + 4x_3 + \quad x_4 &= \quad 21 \\
4x_2 - 2x_3 + 2x_4 &= \quad 10 \\
-2x_3 + 3x_4 &= \quad\; 6 \\
-6x_4 &= -24.
\end{aligned}
$$

Now use back substitution and compute the solution $x_4 = -24/8 - 6 = 4, x_3 = (6 - 3(4))/(-2) = 3, x_2 = (10 - 2(4) + 2(3))/4 = 2$, and $x_1 = 21 - 4 - 4(3) - 2(2) = 1$, or $X = [1 \;\; 2 \;\; 3 \;\; 4]$.

Triangular Factorization

If row interchanges are not necessary when using Gaussian elimination, the multipliers m_{ij} are the subdiagonal entries in L.

Example:

Use Gaussian elimination to construct the triangular factorization of the matrix

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

Triangular Factorization

If row interchanges are not necessary when using Gaussian elimination, the multipliers m_{ij} are the subdiagonal entries in L.

Example:

Use Gaussian elimination to construct the triangular factorization of the matrix

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

The matrix L will be constructed from an identity matrix placed at the left. For each row operation used to construct the upper-triangular matrix, the multipliers m_{ij} will be put in their proper places at the left. Start with

Triangular Factorization

If row interchanges are not necessary when using Gaussian elimination, the multipliers m_{ij} are the subdiagonal entries in L.

Example:

Use Gaussian elimination to construct the triangular factorization of the matrix

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

The matrix L will be constructed from an identity matrix placed at the left. For each row operation used to construct the upper-triangular matrix, the multipliers m_{ij} will be put in their proper places at the left. Start with

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 4 & 3 & -1 \\ -2 & -4 & 5 \\ 1 & 2 & 6 \end{bmatrix}.$$

Triangular Factorization

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 4 & 3 & -1 \\ \mathbf{-2} & -4 & 5 \\ \mathbf{1} & 2 & 6 \end{bmatrix}.$$

Row 1 is used to eliminate the elements of A in column 1 below a_{11}. The multiples $m_{21} = -0.5$ and $m_{31} = 0.25$ of row 1 are subtracted from rows 2 and 3, respectively. These multipliers are put in the matrix at the left and the result is

Triangular Factorization

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 4 & 3 & -1 \\ \mathbf{-2} & -4 & 5 \\ \mathbf{1} & 2 & 6 \end{bmatrix}.$$

Row 1 is used to eliminate the elements of A in column 1 below a_{11}. The multiples $m_{21} = -0.5$ and $m_{31} = 0.25$ of row 1 are subtracted from rows 2 and 3, respectively. These multipliers are put in the matrix at the left and the result is

$$A = \begin{bmatrix} 1 & 0 & 0 \\ -0.5 & 1 & 0 \\ 0.25 & 0 & 1 \end{bmatrix} \begin{bmatrix} 4 & 3 & -1 \\ 0 & \mathbf{-2.5} & 4.5 \\ 0 & \mathbf{1.25} & 6.25 \end{bmatrix}.$$

Triangular Factorization

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 4 & 3 & -1 \\ \mathbf{-2} & -4 & 5 \\ \mathbf{1} & 2 & 6 \end{bmatrix}.$$

Row 1 is used to eliminate the elements of A in column 1 below a_{11}. The multiples $m_{21} = -0.5$ and $m_{31} = 0.25$ of row 1 are subtracted from rows 2 and 3, respectively. These multipliers are put in the matrix at the left and the result is

$$A = \begin{bmatrix} 1 & 0 & 0 \\ -0.5 & 1 & 0 \\ 0.25 & 0 & 1 \end{bmatrix} \begin{bmatrix} 4 & 3 & -1 \\ 0 & \mathbf{-2.5} & 4.5 \\ 0 & \mathbf{1.25} & 6.25 \end{bmatrix}.$$

Row 2 is used to eliminate the elements of A in column 2 below a_{22}. The multiple $m_{32} = -0.5$ of the second row is subtracted from row 3, and the multiplier is entered in the matrix at the left and we have the desired triangular factorization of A.

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 4 & 3 & -1 \\ \mathbf{-2} & -4 & 5 \\ \mathbf{1} & 2 & 6 \end{bmatrix}.$$

Row 1 is used to eliminate the elements of A in column 1 below a_{11}. The multiples $m_{21} = -0.5$ and $m_{31} = 0.25$ of row 1 are subtracted from rows 2 and 3, respectively. These multipliers are put in the matrix at the left and the result is

$$A = \begin{bmatrix} 1 & 0 & 0 \\ -0.5 & 1 & 0 \\ 0.25 & 0 & 1 \end{bmatrix} \begin{bmatrix} 4 & 3 & -1 \\ 0 & \mathbf{-2.5} & 4.5 \\ 0 & \mathbf{1.25} & 6.25 \end{bmatrix}.$$

Row 2 is used to eliminate the elements of A in column 2 below a_{22}. The multiple $m_{32} = -0.5$ of the second row is subtracted from row 3, and the multiplier is entered in the matrix at the left and we have the desired triangular factorization of A.

$$A = \begin{bmatrix} 1 & 0 & 0 \\ -0.5 & 1 & 0 \\ 0.25 & -0.5 & 1 \end{bmatrix} \begin{bmatrix} 4 & 3 & -1 \\ 0 & -2.5 & 4.5 \\ 0 & 0 & 8.5 \end{bmatrix}.$$

Triangular Factorization

Theorem 5. Direct Factorization $A = LU$.. No Row Interchanges.

Suppose that Gaussian elimination, without row interchanges, can be successfully performed to solve the general linear system $AX = B$. Then the matrix A can be factored as the product of a lower-triangular matrix L and an upper-triangular matrix U:

$$A = LU.$$

Furthermore, L can be constructed to have 1's on its diagonal and U will have nonzero diagonal elements. After finding L and U. the solution X is computed in two steps:

1. Solve $LU = B$ for Y using forward substitution.
2. Solve $UX = Y$ for X using back substitution.

Proof. When the Gaussian elimination process is followed and B is stored in column $N+1$ of the augmented matrix, the result after the upper-triangularization step is the equivalent upper-triangular system $UX = Y$. The matrices L, U, B, and Y will have the form

$$L = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ m_{21} & 1 & 0 & \cdots & 0 \\ m_{31} & m_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ m_{N1} & m_{N2} & m_{N3} & \cdots & 1 \end{bmatrix}, B = \begin{bmatrix} a_{1N+1}^{(1)} \\ a_{2N+1}^{(2)} \\ a_{3N+1}^{(3)} \\ \vdots \\ a_{NN+1}^{(N)} \end{bmatrix}$$

$$U = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1N}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2N}^{(2)} \\ 0 & 0 & a_{33}^{(3)} & \cdots & a_{3N}^{(3)} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & a_{NN}^{(N)} \end{bmatrix}, Y = \begin{bmatrix} a_{1N+1}^{(1)} \\ a_{2N+1}^{(2)} \\ a_{3N+1}^{(3)} \\ \vdots \\ a_{NN+1}^{(N)} \end{bmatrix}$$

Triangular Factorization

Remark. To find just L and U, the $(N+1)$st column is not needed.

$Step 1$. Store the coefficients in the augmented matrix. The superscript on $a_{rc}^{(1)}$ means that this is the first time that a number is stored in location (r, c).

$$\left[\begin{array}{ccccc|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1N}^{(1)} & a_{1N+1}^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} & a_{23}^{(1)} & \cdots & a_{2N}^{(1)} & a_{2N+1}^{(1)} \\ a_{31}^{(1)} & a_{32}^{(1)} & a_{33}^{(1)} & \cdots & a_{3N}^{(1)} & a_{3N+1}^{(1)} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ a_{N1}^{(1)} & a_{N2}^{(1)} & a_{N3}^{(1)} & \cdots & a_{NN}^{(1)} & a_{NN+1}^{(1)} \end{array}\right].$$

Triangular Factorization

Remark. To find just L and U, the $(N+1)$st column is not needed.

*Step*1. Store the coefficients in the augmented matrix. The superscript on $a_{rc}^{(1)}$ means that this is the first time that a number is stored in location (r,c).

$$\left[\begin{array}{ccccc|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1N}^{(1)} & a_{1N+1}^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} & a_{23}^{(1)} & \cdots & a_{2N}^{(1)} & a_{2N+1}^{(1)} \\ a_{31}^{(1)} & a_{32}^{(1)} & a_{33}^{(1)} & \cdots & a_{3N}^{(1)} & a_{3N+1}^{(1)} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ a_{N1}^{(1)} & a_{N2}^{(1)} & a_{N3}^{(1)} & \cdots & a_{NN}^{(1)} & a_{NN+1}^{(1)} \end{array}\right].$$

*Step*2. Eliminate x_1 in rows 2 through N and store the multiplier m_{r1}, used to eliminate x_1 in row r, in the matrix at location $(r,1)$.

for $r = 2 : N$
    $m_{r1} = a_{r1}^{(1)} / a_{11}^{(1)};$
    $a_{r1} = m_{r1};$
    for $c = 2 : N+1$
        $a_{rc}^{(2)} = a_{rc}^{(1)} - m_{r1} * a_{1c}^{(1)};$
    end
end

Triangular Factorization

The new elements are written $a_{rc}^{(2)}$ to indicate that this is the second time that a number has been stored in the matrix at location (r, c). The result after step 2 is

$$\left[\begin{array}{ccccc|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1N}^{(1)} & a_{1N+1}^{(1)} \\ m_{21} & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2N}^{(2)} & a_{2N+1}^{(2)} \\ m_{31} & a_{32}^{(2)} & a_{33}^{(2)} & \cdots & a_{3N}^{(2)} & a_{3N+1}^{(2)} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ m_{N1} & a_{N2}^{(2)} & a_{N3}^{(2)} & \cdots & a_{NN}^{(2)} & a_{NN+1}^{(2)} \end{array}\right].$$

Triangular Factorization

The new elements are written $a_{rc}^{(2)}$ to indicate that this is the second time that a number has been stored in the matrix at location (r, c). The result after step 2 is

$$\left[\begin{array}{ccccc|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1N}^{(1)} & a_{1N+1}^{(1)} \\ m_{21} & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2N}^{(2)} & a_{2N+1}^{(2)} \\ m_{31} & a_{32}^{(2)} & a_{33}^{(2)} & \cdots & a_{3N}^{(2)} & a_{3N+1}^{(2)} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ m_{N1} & a_{N2}^{(2)} & a_{N3}^{(2)} & \cdots & a_{NN}^{(2)} & a_{NN+1}^{(2)} \end{array}\right].$$

$Step 3$. Eliminate x_2 in rows 3 through N and store the multiplier m_{r2}, used to eliminate x_2 in row r, in the matrix at location $(r, 2)$.

for $r = 3 : N$
  $m_{r2} = a_{r2}^{(2)} / a_{22}^{(2)}$;
  $a_{r2} = m_{r2}$;
  for $c = 3 : N + 1$
    $a_{rc}^{(3)} = a_{rc}^{(2)} - m_{r2} * a_{2c}^{(2)}$;
  end
end

Triangular Factorization

The new elements are written $a_{rc}^{(3)}$ to indicate that this is the third time that a number has been stored in the matrix at the location (r, c).

$$\left[\begin{array}{ccccc|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1N}^{(1)} & a_{1N+1}^{(1)} \\ m_{21} & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2N}^{(2)} & a_{2N+1}^{(2)} \\ m_{31} & m_{32} & a_{33}^{(3)} & \cdots & a_{3N}^{(3)} & a_{3N+1}^{(3)} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ m_{N1} & m_{N2} & a_{N3}^{(3)} & \cdots & a_{NN}^{(3)} & a_{NN+1}^{(3)} \end{array}\right].$$

Triangular Factorization

The new elements are written $a_{rc}^{(3)}$ to indicate that this is the third time that a number has been stored in the matrix at the location (r, c).

$$\left[\begin{array}{ccccc|c}
a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1N}^{(1)} & a_{1N+1}^{(1)} \\
m_{21} & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2N}^{(2)} & a_{2N+1}^{(2)} \\
m_{31} & m_{32} & a_{33}^{(3)} & \cdots & a_{3N}^{(3)} & a_{3N+1}^{(3)} \\
\vdots & \vdots & \vdots & & \vdots & \vdots \\
m_{N1} & m_{N2} & a_{N3}^{(3)} & \cdots & a_{NN}^{(3)} & a_{NN+1}^{(3)}
\end{array}\right].$$

Step $p+1$. This is the general step. Eliminate x_p in rows $p+1$ through N and store the multipliers at the location (r, p).

for $r = p+1 : N$
  $m_{rp} = a_{rp}^{(p)} / a_{pp}^{(p)}$;
  $a_{rp} = m_{rp}$;
  for $c = p+1 : N+1$
    $a_{rc}^{(p+1)} = a_{rc}^{(p)} - m_{rp} * a_{pc}^{(p)}$;
  end
end

Triangular Factorization

The final result after x_{N-1} has been eliminated form row N is

$$\left[\begin{array}{ccccc|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1N}^{(1)} & a_{1N+1}^{(1)} \\ m_{21} & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2N}^{(2)} & a_{2N+1}^{(2)} \\ m_{31} & m_{32} & a_{33}^{(3)} & \cdots & a_{3N}^{(3)} & a_{3N+1}^{(3)} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ m_{N1} & m_{N2} & m_{N3} & \cdots & a_{NN}^{(N)} & a_{NN+1}^{(N)} \end{array}\right].$$

Triangular Factorization

The final result after x_{N-1} has been eliminated form row N is

$$\left[\begin{array}{ccccc|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1N}^{(1)} & a_{1N+1}^{(1)} \\ m_{21} & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2N}^{(2)} & a_{2N+1}^{(2)} \\ m_{31} & m_{32} & a_{33}^{(3)} & \cdots & a_{3N}^{(3)} & a_{3N+1}^{(3)} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ m_{N1} & m_{N2} & m_{N3} & \cdots & a_{NN}^{(N)} & a_{NN+1}^{(N)} \end{array}\right].$$

The upper-triangular process is now complete. Notice that one array is used to store the elements of both L and U. The 1's of L are not stored, nor are the 0's of L and U that lie above and below the diagonal, respectively. Only the essential coefficients needed to reconstruct L and U are stored!

We must now verify that the product $LU = A$. Suppose that $D = LU$ and consider the case when $r \leq c$. Then d_{rc} is

$$(1)\ d_{rc} = m_{r1}a_{1c}^{(1)} + m_{r2}a_{2c}^{(2)} + \cdots + m_{rr-1}a_{r-1c}^{(r-1)} + a_{rc}^{(r)}.$$

Triangular Factorization

Using the replacement equations in steps 1 through $p+1=r$, we obtain the following substitutions: (2)

$$\begin{aligned} m_{r1}a_{1c}^{(1)} &= a_{rc}^{(1)} - a_{rc}^{(2)}, \\ m_{r2}a_{2c}^{(2)} &= a_{rc}^{(2)} - a_{rc}^{(3)}, \\ &\vdots \\ m_{rr-1}a_{r-1c}^{(r-1)} &= a_{rc}^{(r-1)} - a_{rc}^{(r)}. \end{aligned}$$

Triangular Factorization

Using the replacement equations in steps 1 through $p+1=r$, we obtain the following substitutions: (2)

$$
\begin{aligned}
m_{r1}a_{1c}^{(1)} &= a_{rc}^{(1)} - a_{rc}^{(2)}, \\
m_{r2}a_{2c}^{(2)} &= a_{rc}^{(2)} - a_{rc}^{(3)}, \\
&\vdots \\
m_{rr-1}a_{r-1c}^{(r-1)} &= a_{rc}^{(r-1)} - a_{rc}^{(r)}.
\end{aligned}
$$

When the substitutions in (2) are used in (1), the result is

$$d_{rc} = a_{rc}^{(1)} - a_{rc}^{(2)} + a_{rc}^{(2)} - a_{rc}^{(3)} + \cdots + a_{rc}^{(r-1)} - a_{rc}^{(r)} + a_{rc}^{(r)} = a_{rc}^{(1)}.$$

The other case, $r > c$, is similar to prove.

Computational Complexity

At the first N columns of the augmented matrix in Theorem 5 the outer loop of step $p+1$ requires $N-p = N-(p+1)+1$ divisions to compute the multipliers m_{rp}. Inside the loops, but for the first N columns only, a total of $(N-p)(N-p)$ multiplications and the same number of subtractions are required to compute the new row elements a_{rc}^{p+1}. This process is for $p = 1, 2, ..., N-1$. The triangulation factorization portion of $A = LU$ requires:

$$\sum_{p=1}^{N-1}(N-p)(N-p+1) = \frac{(N^3-N)}{3} \text{ multiplications and divisions} \tag{1}$$

and

$$\sum_{p=1}^{N-1}(N-p)(N-p) = \frac{(2N^3-3N^2+N)}{6} \text{ subtractions.} \tag{2}$$

Computational Complexity

Once the triangular factorization $A = LU$ has been obtained, the solution to the lower-triangular system $LY = B$ will require $0+1+...+N-1 = (N^2 - N)/2$ multiplications and subtractions; no divisions are required because the diagonal elements of L are $1's$. Then the solution of the upper-triangular system $UX = Y$ requires $1 + 2 + ... + N = (N^2 + N)/2$ multiplications and divisions and $(N^2 - N)/2$ subtractions. Therefore, finding the solution to $LUX = B$ requires

N^2 multiplications and divisions, and $N^2 - N$ subtractions.

The bulk of the calculation lies in the triangularization portion of the solution. If the linear system is to be solved many times, with the same coefficients matrix A but with different column matrices B, it is not necessary to triangularize the matrix each time if the factors are saved.

The $A = LU$ factorization in Theorem 5 assumes that there are no row interchanges. It is possible that a nonsingular matrix A cannot be factored directly as $A = LU$.

Example:

Show that the following matrix cannot be factored directly as $A = LU$:

$$A = \begin{bmatrix} 1 & 2 & 6 \\ 4 & 8 & -1 \\ -2 & 3 & 5 \end{bmatrix}.$$

Permutation Matrices

The $A = LU$ factorization in Theorem 5 assumes that there are no row interchanges. It is possible that a nonsingular matrix A cannot be factored directly as $A = LU$.

Example:

Show that the following matrix cannot be factored directly as $A = LU$:

$$A = \begin{bmatrix} 1 & 2 & 6 \\ 4 & 8 & -1 \\ -2 & 3 & 5 \end{bmatrix}.$$

Suppose that A has a direct factorization LU; then

$$\begin{bmatrix} 1 & 2 & 6 \\ 4 & 8 & -1 \\ -2 & 3 & 5 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ m_{21} & 1 & 0 \\ m_{31} & m_{32} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix} \tag{3}$$

Permutation Matrices

The matrices L and U on the right-hand side of (3) can be multiplied and each element of the product compared with the corresponding element of the matrix A.

$1 = 1u_{11}$
$4 = m_{21}u_{11} = m_{21}$
$-2 = m_{31}u_{11} = m_{31}$
$2 = 1u_{12}$
$8 = m_{21}u_{12} = (4)(2) + u_{22}$, then $u_{22} = 0$
$3 = m_{31}u_{12} + m_{32}u_{22} = (-2)(2) + m_{32}(0) = -4$, which is a contradiction.
Therefore, A does not have a LU factorization.

Permutation Matrices

A permutation of the first N positive integers $1, 2, ..., N$ is an arrangement $k_1, k_2, ..., k_N$ of these integers in a definite order. For example, $1, 4, 2, 3, 5$ is a permutation of $1, 2, 3, 4, 5$. The standard base vectors $E_i = [0\,0 \ldots 0\,1_i\,0 \ldots 0]$,

Definition 4.

An $N \times N$ ***permutation matrix*** $\mathbf{P}$ is a matrix with precisely one entry whose value is 1 in each column and row, and all of whose other entries are 0. The rows of P are a permutation of the rows of the identity matrix and can be written as

$$P = [E'_{k_1}\ E'_{k_2}\ \ldots\ E'_{k_N}]'. \tag{4}$$

The elements of $P = [p_{ij}]$ have the form

$$p_{ij} = \begin{cases} 1 & j = k_i, \\ 0 & \text{otherwise.} \end{cases}$$

Permutation Matrices

For example, the following 4×4 matrix is a permutation matrix,

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} = [E_2' \ E_1' \ E_4' \ E_3']'.$$

Theorem 6.

Suppose that $P = [E_{k_1}' \ E_{k_2}' \ \ldots \ E_{k_N}']'$ is a permutation matrix. The product PA is a new matrix whose rows consist of the rows of A rearranged in the order row$_{k1}A$, row$_{k2}A$, ..., row$_{kN}A$.

Iterative Methods for Linear Systems

We consider an extension of fixed-point iteration that applies to systems of linear equations.

Example: Jacobi Iteration

Consider the system of equations

$$\begin{aligned} 4x - y + z &= 7 \\ 4x - 8y + z &= -21 \\ -2x + y + 5z &= 15. \end{aligned}$$

These equations can be written in the form

$$\begin{aligned} x &= \frac{7 + y - z}{4} \\ y &= \frac{21 + 4x + z}{8} \\ z &= \frac{15 + 2x - y}{5}. \end{aligned}$$

This suggests the following Jacobi iterative process:

$$x = \frac{7 + y - z}{4}$$
$$y = \frac{21 + 4x + z}{8}$$
$$z = \frac{15 + 2x - y}{5}.$$

Let us show that if we start with $P_0 = (x_0, y_0, z_0) = (1, 2, 2)$, then the iteration in (3) appears to converge to the solution (2, 4, 3).
Substitute $x_0 = 1$, $y_0 = 2$, and $z_0 = 2$ into the right-hand side of each equation in (3) to obtain the new values

$$x_1 = \frac{7 + 2 - 2}{4} = 1.75$$
$$y_1 = \frac{21 + 4 + 2}{8} = 3.375$$
$$z_1 = \frac{15 + 2 - 2}{5} = 3.00$$

The new point $P_1 = (1.75, 3.375, 3.00)$ is closer to (2, 4, 3) then P_0. Iteration using (3) generates a sequence of points P_k that converges to the solution (2, 4, 3)

k	x_k	y_k	z_k
0	1.0	2.0	2.0
1	1.75	3.375	3.0
2	1.84375	3.875	3.025
3	1.9625	3.925	2.9625
4	1.99062500	3.97656250	3.00000000
5	1.99414063	3.99531250	3.00093750
⋮	⋮	⋮	⋮
15	1.99999993	3.99999985	2.99999993
⋮	⋮	⋮	⋮
19	2.00000000	4.00000000	3.00000000

This process is called ***Jacobi iteration*** and can be used to solve certain types of linear systems. After 19 steps, the iteration has converged to the nine-digit machine approximation (2.00000000,4.00000000,3.0000000

Linear systems with as many as 100,000 variables often arise in the solution of partial differential equations. The coefficient matrices for these systems are sparses that is, a large percentage of the entries of the coefficient matrix are zero. If there is a pattern to the nonzero entries (i.e., tridiagonal systems), the an iterative process provides an efficient method for solving these large systems.

Iterative Methods for Linear Systems

Linear systems with as many as 100,000 variables often arise in the solution of partial differential equations. The coefficient matrices for these systems are sparses that is, a large percentage of the entries of the coefficient matrix are zero. If there is a pattern to the nonzero entries (i.e., tridiagonal systems), the an iterative process provides an efficient method for solving these large systems.

Sometimes the jacobi method does not work. Let us experiment and see that a rearrangement of the original linear system can result in a system of iteration equations that will produce a divergent sequence of points.

Example 2: Let the linear system be rearranged as follows: (4)

$$\begin{aligned} -2x + y + 5z &= 15 \\ 4x - 8y + z &= -21 \\ 4x - y + z &= 7. \end{aligned}$$

These equations can be written in the form (5)

$$\begin{aligned} x &= \frac{-15 + y + 5z}{2} \\ y &= \frac{21 + 4x + z}{8} \\ z &= 7 - 4x + y. \end{aligned}$$

This suggests the following Jacobi iterative process: (6)

$$\begin{aligned} x_{k+1} &= \frac{-15 + y_k + 5z_k}{2} \\ y_{k+1} &= \frac{21 + 4x_k + z_k}{8} \\ z_{k+1} &= 7 - 4x_k + y_k. \end{aligned}$$

See that if we start with $P_0 = (x_0, y_0, z_0) = (1, 2, 2)$ then the iteration using (6) will diverge away from the solution (2, 4, 3).

Substitute $x_0 = 1, y_0 = 2$, and $z_0 = 2$ into the right-hand side of each equation in (6) to obtain the new values x_1, y_1, and z_1:

$$x_1 = \frac{-15 + 2 + 10}{2} = -1.5$$

$$y_1 = \frac{21 + 4 + 2}{8} = 3.375$$

$$z_1 = 7 - 4 + 2 = 5.00.$$

The new point $P_1 = (?1.75, 3.375, 5.00)$ is farther away from the solution (2, 4, 3) than P_0 . Iteration using the equations in (6) produces a divergent sequence.

k	x_k	y_k	z_k
0	1.0	2.0	2.0
1	−1.5	3.375	5.0
2	6.6875	2.5	16.375
3	34.6875	8.015625	−17.25
4	−46.617188	17.8125	−123.73438
5	−307.929688	−36.150391	211.28125
6	502.62793	−124.929688	1202.56836
⋮	⋮	⋮	⋮

Gauss-Seidel Iteration

Sometimes the convergence can be speeded up. Observe that the Jacobi iterative process yields three sequence x_k, y_k, and z_k that converge to 2, 4, and 3, respectively. It seems reasonable that x_{k+1} could be used in place of x_k in the computation of y_{k+1}. Similarly, x_{k+1} and y_{k+1} might be used in the computation of z_{k+1}. The next example shows what happens when this is applied to the equations in Example 1.

Example

Consider the system of equations given and Gauss-Seidel iterative process suggested by: (7)

$$x_{k+1} = \frac{7 + y_k - z_k}{4}$$
$$y_{k+1} = \frac{21 + 4x_{k+1} + z_k}{8}$$
$$z_{k+1} = \frac{15 + 2x_{k+1} - y_{k+1}}{5}.$$

Gauss-Seidel Iteration

See that if we start with $P_0 = (x_0, y_0, z_0)$ = (1, 2, 2). then iteration using (7) will converge to the solution (2, 4, 3).
Substitute $y_0 = 2$ and $z_0 = 2$ into the first equation of (7) and obtain

$$x_1 = \frac{7 + 2 - 2}{4} = 1.75$$

Then substitute $x_1 = 1.75$ and $z_0 = 2$ into the second equation and get

$$y_1 = \frac{21 + 4(1.75) + 2}{8} = 3.75$$

Finally, Substitute $x_1 = 1.75$ and $y_1 = 3.75$ into the third equation to get

$$z_1 = \frac{15 + 2(1.75) + 3.75}{5} = 2.95$$

Gauss-Seidel Iteration

The new point $P_1 = (1.75, 3.75, 2.95)$ is closer to (2, 4, 3) than P_0 and is better than the value given in Example 1. Iteration using (7) generates a sequence P_k that converges to (2, 4, 3).

k	x_k	y_k	z_k
0	1.0	2.0	2.0
1	1.75	3.75	2.95
2	1.95	3.96875	2.98625
3	1.995625	3.99609375	2.99903125
⋮	⋮	⋮	⋮
8	1.99999983	3.99999988	2.99999996
9	1.99999998	3.99999999	3.00000000
10	2.00000000	4.00000000	3.00000000

In view of Example 1 and 2, it is necessary to have some criterion to determine whether the jacobi iteration will converge. Hence we make the following definition.

Definition 5

A matrix A of dimension N x N is said to be strictly diagonally dominant provided that (8)

$$|a_{kk}| > \sum_{j=1}^{N} |a_{kj}| \quad , \quad j \neq k \qquad \text{for } k = 1, 2, ..., N$$

This means that in each row of the matrix the magnitude of the element on the main diagonal must exceed the sum of the magnitudes of all other elements in the row. The coefficient matrix of the linear system (1) in Example 1 is strictly diagonally dominant because

In row 1: $|4| > |-1| + |1|$
In row 2:$|-8| > |4| + |1|$
In row 3: $|5| > |-2| + |1|$.

All the rows satisfy relation (8) in Definition 5; therefore, the coefficient matrix A for the linear system (1) is strictly diagonally dominant.
The coefficient matrix A of the linear system (4) in Example 2 is not strictly diagonally dominant because

In row 1: $|-2| < |1| + |5|$
In row 2: $|-8| > |4| + |1|$
In row 3: $|1| < |4| + |-1|$.

Gauss-Seidel Iteration

Rows 1 and 3 do not satisfy relation (8) in Definition 5; therefore, the coefficient matrix A for the linear system (4) is not strictly diagonally dominant.

We now generalize the Jacobi and Gauss-Seidel iteration processes. Suppose that the given linear system is (9)

$$
\begin{array}{ccccccccccc}
a_{11}x_1 & + & a_{12}x_2 & + \cdots + & a_{1}x_j & + \cdots + & a_{1N}x_N & = & b_1 \\
a_{21}x_1 & + & a_{22}x_2 & + \cdots + & a_{2j}x_j & + \cdots + & a_{2N}x_N & = & b_2 \\
\vdots & & \vdots & & \vdots & & \vdots & & \vdots \\
a_{N1}x_1 & + & a_{N2}x_2 & + \cdots + & a_{jj}x_j & + \cdots + & a_{jN}x_N & = & b_j \\
\vdots & & \vdots & & \vdots & & \vdots & & \vdots \\
a_{N1}x_1 & + & a_{N2}x_2 & + \cdots + & a_{Nj}x_j & + \cdots + & a_{NN}xN & = & b_N.
\end{array}
$$

Let the kth point be $P_k = (x_1^{(k)}, x_2^{(k)}, ..., x_j^{(k)}, ..., x_N^{(k)})$; then the next point is $P_{k+1} = (x_1^{(k+1)}, x_2^{(k+1)}, ..., x_j^{(k+1)}, ..., x_N^{(k+1)})$. The superscript (k) on the coordinates of P_k enables us to identify the coordinates that belong to this point.

Gauss-Seidel Iteration

The iteration formulas use row j of (9) to solve for $x_j^{(k+1)}$ in terms of a linear combination of the previous values $x_1^{(k)}, x_2^{(k)}, ..., x_j^{(k)}, ..., x_N^{(k)}$:
Jacobi iteration: (10)

$$x_j^{(k+1)} = \frac{b_j - a_{j1}x_1^{(k)} - \cdots - a_{jj} - 1x_{j-1}^{(k)} - a_{jj} + 1x_{j+1}^{(k)} - \cdots - a_{jN}x_N^{(k)}}{a_{jj}}$$

for $j = 1, 2, ..., N$.

Jacobi iteration uses all old coordinates to generate all new coordinates, whereas Gauss-Seidel iteration uses the new coordinates as they become available:
Gauss-Seidel Iteration: (11)

$$x_j^{(k+1)} = \frac{b_j - a_{j1}x_1^{k+1} - \cdots - a_{jj-1}x_{j-1}^{(k+1)} - a_{jj+1}x_{j+1}^{(k)} - \cdots - a_{jN}x_N^{(k)}}{a_{jj}}$$

for $j = 1, 2, ..., N$.

Gauss-Seidel Iteration.

Theorem 7. Jacobi Iteration.

Suppose that A is a strictly diagonally dominant matrix. Then $AX = B$ has a unique solution $X = P$. Iteration using formula (10) will produce a sequence of vectors P_k that will converge to P for any choice of the starting vector P_0.

Proff The proof can be found in advanced texts on numerical analysis. It can be proved that the Gauss-Seidel method will also converge when the matrix A is strictly diagonally dominant. In many cases the Gauss-Seidel method will converge faster than the Jacobi method; hence it is usually preferred. It is important to understand the slight modification of formula (10) that has been made to obtain formula (11). In some cases the Jacobi method will converge even though the Gauss-Seidel method will not.

Convergence

A measure of the closeness between vectors is needed so that we can determine if P_k is converging to P. The Euclidean distance between $P = (x_1, x_2, ..., x_N)$ and $Q = (y_1, y_2, ..., y_N)$ is

$$||P - Q|| = \left(\sum_{j=1}^{N} (x_j - y_j)^2 \right)^{1/2} . \quad (5)$$

Its disadvantage is that it requires considerable computing effort. Hence we introduce a different norm, $||X||_1$:

$$||X||_1 = \sum_{j=1}^{N} |x_j|. \quad (6)$$

The following result ensures that $||X||_1$ has the mathematical structure of a metric and hence is suitable to use as a generalized "distance formula." From the study of linear algebra we know that on a finite-dimensional vector space all norms are equivalent; that is, if two vectors are close in the $|| * ||_1$ norm, then they are also close in the Euclidean norm $|| * ||$.

Convergence

Theorem 8.

Let X and Y be N-dimensional vectors and c be a scalar. Then the function $||X||_1$ has the following properties:
$||X||_1 \leq 0$,
$||X||_1 = 0$ if and only if $X = 0$,
$||cX||_1 = |c| \ ||X||_1$,
$||X + Y||_1 \leq ||X||_1 + ||Y||_1$.

Proof We prove and leave the others as exercises. For each j, the triangle inequality for real numbers states that $|x_j + y_j| \leq |x_j| + |y_j|$. Summing these yields inequality:

$$||X + Y||_1 = \sum_{j=1}^{N} |x_j + y_j| \leq \sum_{j=1}^{N} |x_j| + \sum_{j=1}^{N} |y_j| = ||X||_1 + ||Y||_1. \qquad (7)$$

The norm given can be used to define the distance between points.

Definition 6.

Suppose that X and Y are two points in *N*-dimensional space. We define the distance between X and Y in the $||*||_1$ norm as

$$||X-Y||_1 = \sum_{j=1}^{N} |x_j - y_j|.$$

Example Determine the Euclidean distance and $||*||_1$ distance between the points $P=(2,4,3)$ and $Q=(1.75,3.75,2.95)$.
The Euclidean distance is

$$||P-Q||_2 = ((2-1.75)^2 + (4-3.75)^2 + (3-2.95)^2)^{1/2} = 0.3570. \quad (8)$$

The $||*||_1$ distance is

$$||P-Q||_1 = |2-1.75| + |4-3.75| + |3-2.95| = 0.55 \quad (9)$$

The $||*||_1$ is easier to compute and use for determining convergence in *N*-dimensional space.

Iteration for Nonlinear: Seidel and Newton

Iterative techniques will now be discussed that extend the methods to the case of systems of nonlinear functions. Consider the functions

$$(1) \qquad \begin{aligned} f_1(x,y) &= x^2 - 2x - y + 0.5 \\ f_2(x,y) &= x^2 + 4y^2 - 4 \end{aligned}$$

We seek a method of solution for the system of nonlinear equations

$$(2) \qquad f_1(x,y) = 0 \text{ and } f_2(x,y) = 0.$$

The equations $f_1(x,y) = 0$ and $f_2(x,y) = 0$ implicitly define curves in the xy-plane. Hence a solution of the system (2) is a point (p,q) where the two curves cross (i.e., both $f_1(p,q) = 0$ and $f_2(p,q) = 0$). The curves for the system in (1) are well known:

$$(3) \qquad \begin{aligned} x^2 - 2x + 0.5 = 0 &\text{ is the graph of a parabola,} \\ x^2 + 4xy^2 - 4 = 0 &\text{ is the graph on a ellipse.} \end{aligned}$$

Iteration for Nonlinear: Seidel and Newton

The graphs show that there are two solution points and that they are in the vicinity of (-0.2, 1.0) and (1.9, 0.3).
The first techniques is fixed-point iteration. A method must be devised for generating a sequence (pk, qk) that converges to the solution (p, q). The first equation in (3) can be used to solve directly for x. However, a multiple of y can be added to each side of the second equation to get $x^2 + 4y^2 - 8y - 4 = -8y$. The choice of adding $-8y$ is crucial and will be explained later.

Iteration for Nonlinear: Seidel and Newton

We now have an equivalent system of equations:

$$(4) \qquad x = \frac{x^2 - y + 0.5}{2}$$
$$y = \frac{-x^2 - 4y^2 + 8y + 4}{8}$$

These two equations are used to write the recursive formulas. Star with an initial point (p_0, q_0), and then compute the sequence $\{(pk+1, qk+1)\}$ using

$$(5) \qquad p_{k+1} = g_1(p_k, q_k) = \frac{p_k^2 - q_k + 0.5}{2}$$
$$q_{k+1} = g_2(p_k, q_k) = \frac{-p_k^2 - 4q_k^2 + 8q_k + 4}{8}$$

Iteration for Nonlinear: Seidel and Newton

Case(i) : If we use the starting value (p_0, q_0) = (0, 1), then

$$p_1 = \frac{0^2 - 1 + 0.5}{2} = -0.25 \text{ and } q_1 = \frac{-0^2 - 4(1)^2 + 8(1) + 4}{8} = 1.0.$$

In this case the sequence converges to the solution that lies near the starting value(0, 1).
Case(ii) : If we use the starting value (p_0, q_0) = (2, 0), then

$$p_1 = \frac{2^2 - 0 + 0.5}{2} = 2.25 \text{ and } q_1 = \frac{-2^2 - 4(0)^2 + 8(0) + 4-}{8} = 0.0.$$

In this case the sequence diverges away from the solution

Iteration for Nonlinear: Seidel and Newton

Case (i): Start with (0, 1)			Case (ii): Start with (2, 0)		
k	p_k	q_k	k	p_k	q_k
0	0.00	1.00	0	2.00	0.00
1	-0.25	1.00	1	2.25	0.00
2	-0.21875	0.9921875	2	2.78125	-0.1328125
3	-0.2221680	0.9939880	3	4.184082	-0.6085510
4	-0.2223147	0.9938121	4	9.307547	-2.4820360
5	-0.2221941	0.9938029	5	44.80623	-15.891091
6	-0.2222163	0.9938095	6	1011.995	-392.60426
7	-0.2222147	0.9938083	7	512263.2	-205477.82
8	-0.2222145	0.9938084	This sequence is diverging.		
9	-0.2222146	0.9938084			

Iteration for Nonlinear: Seidel and Newton

Iteration using formulas (5) cannot be used to find the second solution (1.900677, 0.3112186). To find this point a different pair of iteration formulas are needed. Start with equation (3) and add $-2x$ to the first equation and $-11y$ to the second equation and get

$$x^2 - 4x - y + 0.5 = -2x \text{ and } x^2 + 4y^2 - 11y - 4 = -11y.$$

These equations can then be used to obtain the iteration formulas

$$\text{(6)} \quad \begin{aligned} p_{k+1} &= g_1(p_k, q_k) = \frac{-p_k^2 + 4p_k + q_k - 0.5}{2} \\ q_{k+1} &= g_2(p_k, q_k) = \frac{-p_k^2 - 4q_k^2 + 11q_k + 4}{11}. \end{aligned}$$

Iteration for Nonlinear: Seidel and Newton

k	p_k	q_k
0	2.00	0.00
1	1.75	0.00
2	1.71875	0.0852273
3	1.753063	0.1776676
4	1.808345	0.2504410
8	1.903595	0.3160782
12	1.900924	0.3112267
16	1.900652	0.3111994
20	1.900677	0.3112196
24	1.900677	0.3112186

Definition 7. Jacobian Matrix.

Assume that $f_1(x,y)$ and $f_2(x,y)$ are functions of the independent variables x and y; then their Jacobian matrix $J(x,y)$ is

$$\begin{bmatrix} \frac{\partial f_1}{\partial x} & \frac{\partial f_1}{\partial y} \\ \frac{\partial f_2}{\partial x} & \frac{\partial f_2}{\partial y} \end{bmatrix}$$

Similarly, if $f_1(x,y,z)$,$f_2(x,y,z)$, and $f_3(x,y,z)$ are functions of the independent variables x,y, and z, then their 3 x 3 Jacobian matrix $J(x,y,z)$ is defined as follows:

$$\begin{bmatrix} \frac{\partial f_1}{\partial x} & \frac{\partial f_1}{\partial y} & \frac{\partial f_1}{\partial z} \\ \frac{\partial f_2}{\partial x} & \frac{\partial f_2}{\partial y} & \frac{\partial f_2}{\partial z} \\ \frac{\partial f_3}{\partial x} & \frac{\partial f_3}{\partial y} & \frac{\partial f_3}{\partial z} \end{bmatrix}$$

Generalized Differential

For a function of several variables, the differential is used to show how changes of the independent variables affect the change in the dependent variables. Suppose that we have

$$u = f_1(x,y,z), v = f_2(x,y,z) \text{ and } w = f_3(x,y,z).$$

Suppose that the values of the functions in (9) are known at the point (x_0, y_0, z_0) and we wish to predict their value at a nearby point (x,y,z). Let du, dv, and dw denote differential changes in the dependent variables and dx, dy, and dz denote differential changes in the independent variables.

These changes obey the relationships

$$
\begin{aligned}
du &= \frac{\partial f_1}{\partial x}(x_0, y_0, z_0)dx + \frac{\partial f_1}{\partial y}(x_0, y_0, z_0)dy + \frac{\partial f_1}{\partial z}(x_0, y_0, z_0)dz,\\
dv &= \frac{\partial f_2}{\partial x}(x_0, y_0, z_0)dx + \frac{\partial f_2}{\partial y}(x_0, y_0, z_0)dy + \frac{\partial f_2}{\partial z}(x_0, y_0, z_0)dz,\\
dw &= \frac{\partial f_3}{\partial x}(x_0, y_0, z_0)dx + \frac{\partial f_3}{\partial y}(x_0, y_0, z_0)dy + \frac{\partial f_3}{\partial z}(x_0, y_0, z_0)dz,
\end{aligned} \tag{10}
$$

If vector notation is used, (10) can be compactly written by using the Jacobian matrix. The function changes are dF and the changes in the variables are denoted dX.

$$
dF = \begin{bmatrix} du \\ dv \\ dw \end{bmatrix} = J(x_0, y_0, z_0) \begin{bmatrix} dx \\ dy \\ dz \end{bmatrix} = J(x_0, y_0, z_0)dX.
$$

Convergence Near Fixed Points

The extensions of the definitions and theorems to the case of two and three dimensions are now given. The notation for *N*-dimensional functions has not been used. The reader can easily find these extensions is many books on numerical analysis.

Definition 8.

A fixed point for the system of two equations

$$x = g_1(x, y) \text{ and } y = g_2(x, y) \tag{12}$$

is a point (p, q) such that $p = g_1(p, q)$ and $q = g_2(p, q)$. Similarly, in three dimensions a fixed point for the system

$$x = g_1(x, y, z), y = g_2(x, y, z) \text{ and } z = g_3(x, y, z) \tag{13}$$

is a point (p, q, r) such that $p = g_1(p, q, r), q = g_2(p, q, r)$ and $r = g_3(p, q, r)$.

Iteration for Nonlinear: Seidel and Newton

Definition 9.

For the functions (12), ***fixed-point iteration*** is

$$p_{k+1} = g_1(pk, qk) and q_{k+1} = g_2(pk, qk) \tag{14}$$

for $k = 0, 1, ...$ Similarly, for the functions (13), ***fixed-point iteration is***

$$\begin{aligned} p_{k+1} &= g_1(pk, qk, rk) \\ q_{k+1} &= g_2(pk, qk, rk) \\ r_{k+1} &= g_3(pk, qk, rk) \end{aligned} \tag{15}$$

for $k = 0, 1, ...$

Iteration for Nonlinear: Seidel and Newton

Theorem 9. Fixed-Point Iteration.

Assume that the functions in (12) and (13) and their first partial derivatives are continuous on a region that contains the fixed point (p, q) or (p, q, r), respectively. If the starting point is chosen sufficiently close to the fixed point, then one of the following cases applies.

Case(i): Two dimensions. If (p_0, q_0) is sufficiently close to (p, q) and if

$$\begin{aligned} \left|\frac{\partial g_1}{\partial x}(p,q)\right| + \left|\frac{\partial g_1}{\partial y}(p,q)\right| &< 1. \\ \left|\frac{\partial g_2}{\partial x}(p,q)\right| + \left|\frac{\partial g_2}{\partial y}(p,q)\right| &< 1. \end{aligned} \tag{16}$$

then the iteration in (14) converges to the fixed point (p, q).

Case(ii): Three dimensions. If (p_0, q_0, r_0) is sufficiently close to (p, q, r) and if

$$\begin{aligned} \left|\frac{\partial g_1}{\partial x}(p,q,r)\right| + \left|\frac{\partial g_1}{\partial y}(p,q,r)\right| + \left|\frac{\partial g_1}{\partial z}(p,q,r)\right| &< 1. \\ \left|\frac{\partial g_2}{\partial x}(p,q,r)\right| + \left|\frac{\partial g_2}{\partial y}(p,q,r)\right| + \left|\frac{\partial g_2}{\partial z}(p,q,r)\right| &< 1. \\ \left|\frac{\partial g_3}{\partial x}(p,q,r)\right| + \left|\frac{\partial g_3}{\partial y}(p,q,r)\right| + \left|\frac{\partial g_3}{\partial z}(p,q,r)\right| &< 1. \end{aligned} \tag{17}$$

then the iteration in (15) converges to the fixed point (p, q, r).

Iteration for Nonlinear: Seidel and Newton

If conditions (16) or (17) are not met, the iteration might diverge. This will usually be the case if the sum of the magnitudes of the partial derivatives is much larger than 1. Theorem 9 can be used to show why the iteration (5) converged to the fixed point near (-0.2, 1.0). The partial derivatives are

$$\frac{\partial}{\partial x} g_1(x,y) = x, \qquad \frac{\partial}{\partial y} g_1(x,y) = -\frac{1}{2}.$$
$$\frac{\partial}{\partial x} g_2(x,y) = -\tfrac{x}{4}, \qquad \frac{\partial}{\partial y} g_2(x,y) = -y+1.$$

Indeed, for all (x,y) satisfying -0.5<x<0.5 and 0.5<y<1.5, the partial derivatives satisfy

$$\left|\frac{\partial}{\partial x} g_1(x,y)\right| + \left|\frac{\partial}{\partial y} g_1(x,y)\right| = |x| + |-0.5| < 1,$$
$$\left|\frac{\partial}{\partial x} g_2(x,y)\right| + \left|\frac{\partial}{\partial y} g_2(x,y)\right| = \frac{|-x|}{4} + |-y+1| < 0.625 < 1.$$

Therefore, the partial derivative conditions in (16) are met and Theorem 9 implies that fixed-point iteration will converge to $(p, q) \approx (0.2222146, 0.9)$ Notice that near the other fixed point (1.90068, 0.31122) the partial derivatives do not meet the conditions in (16); hence convergence is not guaranteed. That is,

$$\left|\frac{\partial}{\partial x} g_1(1.90068, 0.31122)\right| + \left|\frac{\partial}{\partial y} g_1(1.90068, 0.31122)\right| = 2.40068 > 1,$$
$$\left|\frac{\partial}{\partial x} g_2(1.90068, 0.31122)\right| + \left|\frac{\partial}{\partial y} g_2(1.90068, 0.31122)\right| = 1.16395 > 1.$$

Seidel Iteration

An improvement, analogous to the Gauss-Seidel method for linear systems, of fixed point iteration can be made. Suppose that $p_k + 1$ is used in the calculation of $q_k + 1$ (in three dimensions both $p_k + 1$ and $q_k + 1$ are used to compute $r_k + 1$). When these modifications are incorporated in formulas (14) and (15), the method is called Seidel iteration:

(18) $$p_{k+1} = g_1(p_k, q_k) \text{ and } q_{k+1} = g_2(p_{k+1}, q_k),$$

and

(19) $$\begin{aligned} p_{k+1} &= g_1(p_k, q_k, r_k) \\ q_{k+1} &= g_2(p_{k+1}, q_k, r_k) \\ r_{k+1} &= g_3(p_{k+1}, q_{k+1}, r_k) \end{aligned}$$

Iteration for Nonlinear: Seidel and Newton

We now outline the derivation of Newton's method in two dimensions. Newton's method can easily be extended to higher dimensions. Consider the system

$$u = f_1(x, y)$$
$$v = f_2(x, y) \tag{20}$$

which can be considered a transformation from the xy-plane to the uv-plane. We are interested in the behavior of this transformation near the point (x_0, y_0) whose image is the point (u_0, v_0). If the two functions have continuous partial derivatives, then the differential can be used to write a system of linear approximations that is valid near the point (x_0, y_0):

$$\begin{aligned} u - u_0 &= \frac{\partial}{\partial x} f_1(x_0, y_0)(x - x_0) + \frac{\partial}{\partial y} f_1(x_0, y_0)(y - y_0), \\ v - v_0 &= \frac{\partial}{\partial x} f_2(x_0, y_0)(x - x_0) + \frac{\partial}{\partial y} f_2(x_0, y_0)(y - y_0) \end{aligned} \tag{21}$$

Iteration for Nonlinear: Seidel and Newton

The system (21) is a local linear transformation that relates small changes in the independent variables to small changes in the dependent variable. When the Jacobian matrix $J(x_0, y_0)$ is used, this relationship is easier to visualize:

$$(22) \qquad \begin{bmatrix} u - u_0 \\ v - v_0 \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial x} f_1(x_0, y_0) & \frac{\partial}{\partial y} f_1(x_0, y_0) \\ \frac{\partial}{\partial x} f_2(x_0, y_0) & \frac{\partial}{\partial y} f_2(x_0, y_0) \end{bmatrix} \begin{bmatrix} x - x_0 \\ y - y_0 \end{bmatrix}.$$

If the system in (20) is written as a vector function $V = F(X)$, the Jacobian $J(x, y)$ is the two-dimensional analog of the derivative, because (22) can be written as

$$(23) \qquad \Delta F \approx J(x_0, y_0)\Delta X$$

We now use (23) to derive Newton?s method in two dimensions

Iteration for Nonlinear: Seidel and Newton

Consider the system (20) with u and v set equal to zero:

$$\text{(24)} \qquad \begin{aligned} 0 &= f_1(x, y) \\ 0 &= f_2(x, y). \end{aligned}$$

Suppose that (p, q) is a solution of (24); that is

$$\text{(25)} \qquad \begin{aligned} 0 &= f_1(p, q) \\ 0 &= f_2(p, q). \end{aligned}$$

To develop Newton's method for solving (24), we need to consider small changes in the functions near the point (p_0, q_0):

$$(26) \qquad \begin{aligned} \Delta u &= u - u_0, \quad & \Delta p &= x - p_0. \\ \Delta v &= v - v_0, \quad & \Delta q &= y - q_0. \end{aligned}$$

Set $(x, y) = (p, q)$ in (20) and use (25) to see that $(u, v) = (0, 0)$. Hence the changes in the dependent variables are

$$\text{(27)} \qquad \begin{aligned} u - u_0 &= f_1(p, q) - f_1(p_0, q_0) = 0 - f_1(p_0, q_0) \\ v - v_0 &= f_2(p, q) - f_2(p_0, q_0) = 0 - f_2(p_0, q_0). \end{aligned}$$

Iteration for Nonlinear: Seidel and Newton

Use the result of (27) in (22) to get the linear transformation

$$(28)\qquad \begin{bmatrix} \frac{\partial}{\partial x} f_1(p_0, q_0) & \frac{\partial}{\partial y} f_1(p_0, q_0) \\ \frac{\partial}{\partial x} f_2(p_0, q_0) & \frac{\partial}{\partial y} f_2(p_0, q_0) \end{bmatrix} \begin{bmatrix} \Delta p \\ \Delta q \end{bmatrix} \approx - \begin{bmatrix} f_1(p_0, q_0) \\ f_2(p_0, q_0) \end{bmatrix}$$

If the Jacobian $J(p_0, q_0)$ in (28) is nonsingular, we can solve for $\Delta P = [\Delta p \Delta q]' = [pq]' - [p_0 q_0]'$ as follows:

$$(29)\qquad \Delta P \approx -J(p_0, q_0)^{-1} F(p_0, q_0).$$

Then the next approximation P_1 to the solution P is

$$(30)\qquad P_1 = p_0 + \Delta P = P_= - J(p_0, q_0)^{-1} F(p_0, q_0).$$

Notice that (30) is the generalization of Newton's method for the one variable cases that is $p_1 = p_0 - f(p_0)/f'(p_0)$.

Outline of Newton's Method

Suppose that P_k has been obtained.

Step 1. Evaluate the function

$$F(p_k) = \begin{bmatrix} f_1(p_k, q_) \\ f_2(p_k, q_) \end{bmatrix}$$

Step 2. Evaluate the Jacobian

$$J(p_k) = \begin{bmatrix} \frac{\partial}{\partial x} f_1(p_k, q_k) & \frac{\partial}{\partial y} f_1(p_0, q_0) \\ \frac{\partial}{\partial x} f_2(p_0, q_0) & \frac{\partial}{\partial y} f_2(p_0, q_0) \end{bmatrix}$$

Step 3. Solve the linear system

$$J(P_k)\Delta P = -F(P_k) \text{ for } \Delta P$$

Now, repeat the process.

Example: Consider the nonlinear system

$$\begin{aligned} 0 &= x^2 - 2x - y + 0.5 \\ 0 &= x^2 + 4y^2 - 4. \end{aligned}$$

Use Newton's method with the starting value $(p_0, q_0) = (2.00, 0.25)$ and compute $(p_1, q_1), (p_2, q_2)$, and $(p3, q3)$.

The function vector and Jacobian matrix are

$$F(x,y) = \begin{bmatrix} x^2 - 2x - y + 0.5 \\ x^2 + 4y^2 - 4 \end{bmatrix}, \quad J(x,y) = \begin{bmatrix} 2x - 2 & 1 \\ 2x & 8y \end{bmatrix}.$$

At the point (2.00, 0.25) the take on the values

$$F(2.00, 0.25) = \begin{bmatrix} 0.25 \\ 0.25 \end{bmatrix} \quad J(2.00, 0, 25) = \begin{bmatrix} 2.0 & -1.0 \\ 4.0 & 2.0 \end{bmatrix}.$$

The differentials Δp and Δq are solutions of the linear system

$$\begin{bmatrix} 2.0 & -1.0 \\ 4.0 & 2.0 \end{bmatrix} \begin{bmatrix} \Delta p \\ \Delta q \end{bmatrix} = - \begin{bmatrix} 0.25 \\ 0.25 \end{bmatrix}.$$

Iteration for Nonlinear: Seidel and Newton

A straightforward calculation reveals that

$$\Delta P = \begin{bmatrix} \Delta p \\ \Delta q \end{bmatrix} = \begin{bmatrix} -0.09375 \\ 0.0625 \end{bmatrix}.$$

The next point in the iteration is

$$P_1 = P_0 + \Delta P = \begin{bmatrix} 2.00 \\ 0.25 \end{bmatrix} + \begin{bmatrix} -0.09375 \\ 0.0625 \end{bmatrix} = \begin{bmatrix} 1.90625 \\ 0.3125 \end{bmatrix}.$$

Similarly, the next two points are

$$P_2 = \begin{bmatrix} 1.900691 \\ 0.311213 \end{bmatrix} \text{ and } P_3 = \begin{bmatrix} 1.900677 \\ 0.311219 \end{bmatrix}.$$

Iteration for Nonlinear: Seidel and Newton

The coordinates of P3 are accurate to six decimal places. Calculations for finding P_2 and P_3 are summarized

P_k	Solution of the linear system $J(P_k)\Delta P = -F(P_k)$	$P_k + \Delta P$
$\begin{bmatrix} 2.00 \\ 0.25 \end{bmatrix}$	$\begin{bmatrix} 2.0 & -1.0 \\ 4.0 & 2.0 \end{bmatrix} \begin{bmatrix} 0.09375 \\ 0.0625 \end{bmatrix} = - \begin{bmatrix} 0.25 \\ 0.25 \end{bmatrix}$	$\begin{bmatrix} 1.90625 \\ 0.3125 \end{bmatrix}$
$\begin{bmatrix} 1.90625 \\ 0.3125 \end{bmatrix}$	$\begin{bmatrix} 1.8125 & -1.0 \\ 3.8125 & 2.5 \end{bmatrix} \begin{bmatrix} 0.005559 \\ 0.001287 \end{bmatrix} = - \begin{bmatrix} 0.008789 \\ 0.024414 \end{bmatrix}$	$\begin{bmatrix} 1.900691 \\ 0.311213 \end{bmatrix}$
$\begin{bmatrix} 1.900691 \\ 0.311213 \end{bmatrix}$	$\begin{bmatrix} 1.801381 & -1.000000 \\ 3.801381 & 2.489700 \end{bmatrix} \begin{bmatrix} 1.900691 \\ 0.311213 \end{bmatrix} \begin{bmatrix} -0.000014 \\ 0.000006 \end{bmatrix} = - \begin{bmatrix} 0.000031 \\ 0.000038 \end{bmatrix}$	$\begin{bmatrix} 1.900677 \\ 0.311219 \end{bmatrix}$