UNIT II NUMERICAL METHODS

2.Definition

An equation that consists of derivatives is called a differential equation. Differential equations have applications in all areas of science and engineering. Mathematical formulation of most of the physical and engineering problems lead to differential equations. So, it is important for engineers and scientists to know how to set up differential equations and solve them.

Differential equations are of two types

1) ordinary differential equation (ODE)
2) partial differential equations (PDE).

An ordinary differential equation is that in which all the derivatives are with respect to a single independent variable. Examples of ordinary differential equation include

1) $\frac{d^2y}{dx^2}+2\frac{dy}{dx}+y=0, \frac{dy}{dx}(0)=2, y(0)=4,$
2) $\frac{d^3y}{dx^3}+3\frac{d^2y}{dx^2}+5\frac{dy}{dx}+y=\sin x, \frac{d^2y}{dx^2}(0)=12, \frac{dy}{dx}(0)=2, y(0)=4$

Example 1

$$\frac{dy}{dx}+2y=1.3e^{-x}, y(0)=5$$

is rewritten as

$$\frac{dy}{dx}=1.3e^{-x}-2y, y(0)=5$$

In this case

$$f(x,y)=1.3e^{-x}-2y$$

Example 2

$$e^y\frac{dy}{dx}+x^2y^2=2Sin(3x), y(0)=5$$

is rewritten as

$$\frac{dy}{dx}=\frac{2Sin(3x)-x^2y^2}{e^y}, y(0)=5$$

In this case

$$f(x,y)=\frac{2Sin(3x)-x^2y^2}{e^y}$$

2.1 NEWTON'S METHOD

Let's say we want to evaluate the cube root of 467. That is, we want to find a value of x such that $x^3=467$. Put another way, we want to find a *root* of the following equation:

$$f(x)=x^3-467=0.$$

If $f(x)$ were a straight line, then $f(x_1) = f(x_o) + \dfrac{df(x = x_o)}{dx}(x_1 - x_o) = 0$.

In fact, $f(x_1) \neq 0$, but let's say that $f(x_1) \cong 0$ and solve for x_1.

$$x_1 = x_o + \frac{f(x_1) - f(x_o)}{\dfrac{df(x_o)}{dx}} \cong x_o - \frac{f(x_o)}{f'(x_o)}.$$

Note that we are using $f'(x_o) = \dfrac{df(x = x_o)}{dx}$.

Having now obtained a new estimate for the root, we repeat the process to obtain a sequence of estimated roots which we hope converges on the exact or correct root.

$$x_2 \cong x_1 - \frac{f(x_1)}{f'(x_1)}$$

$$x_3 \cong x_2 - \frac{f(x_2)}{f'(x_2)}$$

etc.

In our example, $f(x) = x^3 - 467$ and $f'(x) = 3x^2$. If we take our *initial guess* to be $x_o = 6$, then by *iterating* the formula above, we generate the following table:

i	x_i	$f(x_i)$	$f'(x_i)$
0	6	-251	108
1	8.324	109.7718	207.8706
2	7.796	6.8172	182.3316
3	7.759	0.108	0.0350

$$x_1 \cong x_o - \frac{f(x_o)}{f'(x_{1o})} = 6 - \frac{-251}{108} = 8.32407$$

$$x_2 \cong x_1 - \frac{f(x_1)}{f'(x_1)} = 8.32407 - \frac{109.7768}{207.8706} = 7.79597$$

$$x_3 \cong x_2 - \frac{f(x_2)}{f'(x_2)} = 7.79597 - \frac{6.817273}{182.33156} = 7.75858$$

Example In Chemical Engineering

You have a spherical storage tank containing oil. The tank has a diameter of $6\,\text{ft}$. You are asked to calculate the height h to which a dipstick 8 ft long would be wet with oil when immersed in the tank when it contains $6\,\text{ft}^3$ of oil.

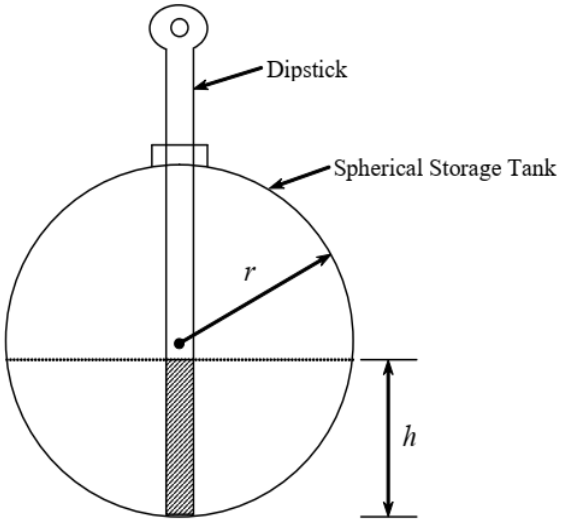


Figure 1 Spherical storage tank problem.

The equation that gives the height h of the liquid in the spherical tank for the given volume and radius is given by

$$f(h) = h^3 - 9h^2 + 3.8197 = 0$$

Use the Newton-Raphson method of finding roots of equations to find the height h to which the dipstick is wet with oil. Conduct three iterations to estimate the root of the above equation. Find the absolute relative approximate error at the end of each iteration and the number of significant digits at least correct at the end of each iteration.

Solution

$$f(h) = h^3 - 9h^2 + 3.8197$$

$$f'(h) = 3h^2 - 18h$$

Let us take the initial guess of the root of $f(h) = 0$ as $h_0 = 1$.

Iteration 1

The estimate of the root is

$$h_1 = h_0 - \frac{f(h_0)}{f'(h_0)}$$

$$= 1 - \frac{(1)^3 - 9(1)^2 + 3.8197}{3(1)^2 - 18(1)}$$

$$= 1 - \frac{-4.1803}{-15}$$

$$= 1 - (0.27869)$$

$$= 0.72131$$

The absolute relative approximate error $|\epsilon_a|$ at the end of Iteration 1 is

$$\left|\epsilon_a\right| = \left|\frac{h_1 - h_0}{h_1}\right| \times 100$$
$$= \left|\frac{0.72131 - 1}{0.72131}\right| \times 100$$
$$= 38.636\%$$

The number of significant digits at least correct is 0, as you need an absolute relative approximate error of 5% or less for one significant digit to be correct in your result.

Iteration 2

The estimate of the root is

$$h_2 = h_1 - \frac{f(h_1)}{f'(h_1)}$$
$$= 0.72131 - \frac{(0.72131)^3 - 9(0.72131)^2 + 3.8197}{3(0.72131)^2 - 18(0.72131)}$$
$$= 0.72131 - \frac{-0.48764}{-11.423}$$
$$= 0.72131 - (0.042690)$$
$$= 0.67862$$

The absolute relative approximate error $\left|\epsilon_a\right|$ at the end of Iteration 2 is

$$\left|\epsilon_a\right| = \left|\frac{h_2 - h_1}{h_2}\right| \times 100$$
$$= \left|\frac{0.67862 - 0.72131}{0.67862}\right| \times 100$$
$$= 6.2907\%$$

The number of significant digits at least correct is 0.

Iteration 3

The estimate of the root is

$$h_3 = h_2 - \frac{f(h_2)}{f'(h_2)}$$
$$= 0.67862 - \frac{(0.67862)^3 - 9(0.67862)^2 + 3.8197}{3(0.67862)^2 - 18(0.67862)}$$
$$= 0.67862 - \frac{-0.012536}{-10.834}$$
$$= 0.67862 - (0.0011572)$$
$$= 0.67747$$

The absolute relative approximate error $|\epsilon_a|$ at the end of Iteration 3 is

$$|\epsilon_a| = \left|\frac{h_3 - h_2}{h_3}\right| \times 100$$

$$= \left|\frac{0.67747 - 0.67862}{0.67747}\right| \times 100$$

$$= 0.17081\%$$

Hence the number of significant digits at least correct is given by the largest value of m for which

$$|\epsilon_a| \le 0.5 \times 10^{2-m}$$

$$0.17081 \le 0.5 \times 10^{2-m}$$

$$0.34162 \le 10^{2-m}$$

$$\log(0.34162) \le 2 - m$$

$$m \le 2 - \log(0.34162) = 2.4665$$

So

$$m = 2$$

The number of significant digits at least correct in the estimated root 0.67747 is 2.

2.2 RUNGE-KUTTA 4TH ORDER

Runge-Kutta 4th order method is based on the following

$$y_{i+1} = y_i + (a_1 k_1 + a_2 k_2 + a_3 k_3 + a_4 k_4)h$$

where knowing the value of $y = y_i$ at x_i, we can find the value of $y=y_{i+1}$ at x_{i+1}, and

$$h = x_{i+1} - x_i$$

The above equation is equated to the first five terms of Taylor series

$$y_{i+1} = y_i + \frac{dy}{dx}\Big|_{x_i, y_i}(x_{i+1} - x_i) + \frac{1}{2!}\frac{d^2y}{dx^2}\Big|_{x_i, y_i}(x_{i+1} - x_i)^2 + \frac{1}{3!}\frac{d^3y}{dx^3}\Big|_{x_i, y_i}(x_{i+1} - x_i)^3$$

$$+ \frac{1}{4!}\frac{d^4y}{dx^4}\Big|_{x_i, y_i}(x_{i+1} - x_i)^4$$

Knowing that $\frac{dy}{dx} = f(x, y)$ and $x_{i+1} - x_i = h$

$$y_{i+1} = y_i + f(x_i, y_i)h + \frac{1}{2!}f'(x_i, y_i)h^2 + \frac{1}{3!}f''(x_i, y_i)h^3 + \frac{1}{4!}f'''(x_i, y_i)h^4$$

Based on equating the above equations, one of the popular solutions used is

$$\boxed{y_{i+1} = y_i + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)h}$$

$$k_1 = f(x_i, y_i)$$
$$k_2 = f\left(x_i + \frac{1}{2}h, y_i + \frac{1}{2}k_1 h\right)$$
$$k_3 = f\left(x_i + \frac{1}{2}h, y_i + \frac{1}{2}k_2 h\right)$$
$$k_4 = f(x_i + h, y_i + k_3 h)$$

Example In Chemical Engineering

The concentration of salt x in a home made soap maker is given as a function of time by

$$\frac{dx}{dt} = 37.5 - 3.5x$$

At the initial time, $t = 0$, the salt concentration in the tank is 50 g/L Using Runge-Kutta 4th order method and a step size of, $h = 1.5$ min , what is the salt concentration after 3 minutes?

Solution

$$\frac{dx}{dt} = 37.5 - 3.5x$$
$$f(t, x) = 37.5 - 3.5x$$
$$x_{i+1} = x_i + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)h$$

For $i = 0$, $t_0 = 0$, $x_0 = 50$

$$k_1 = f(t_0, x_0)$$
$$= f(0,50)$$
$$= 37.5 - 3.5(50)$$
$$= -137.5$$

$$k_2 = f\left(t_0 + \frac{1}{2}h, x_0 + \frac{1}{2}k_1 h\right)$$
$$= f\left(0 + \frac{1}{2}1.5, 50 + \frac{1}{2}(-137.5)1.5\right)$$
$$= f(0.75, -53.125)$$
$$= 37.5 - 3.5(-53.125)$$
$$= 223.44$$

$$k_3 = f\left(t_0 + \frac{1}{2}h, x_0 + \frac{1}{2}k_2 h\right)$$

$$= f\left(0+\frac{1}{2}1.5, 50+\frac{1}{2}(223.44)1.5\right)$$
$$= f(0.75, 217.58)$$
$$= 37.5 - 3.5(217.58)$$
$$= -724.02$$
$$k_4 = f(t_0 + h, x_0 + k_3 h)$$
$$= f(0+1.5, 50+(-724.03)1.5)$$
$$= f(1.5, -1036.0)$$
$$= 37.5 - 3.5(-1036.0)$$
$$= 3663.6$$
$$x_1 = x_0 + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)h$$
$$= 50 + \frac{1}{6}(-137.5 + 2(223.44) + 2(-724.02) + (3663.6))1.5$$
$$= 50 + \frac{1}{6}(2525.0)1.5$$
$$= 681.24 \text{ g/L}$$

x_1 is the approximate concentration of salt at

$$t = t_1 = t_0 + h = 0 + 1.5 = 1.5$$
$$x(1.5) \approx x_1 = 681.24 \text{ g/L}$$

For $i = 1,\ t_1 = 1.5,\ x_1 = 681.24$

$$k_1 = f(t_1, x_1)$$
$$= f(1.5, 681.24)$$
$$= 37.5 - 3.5(681.24)$$
$$= -2346.8$$
$$k_2 = f\left(t_1 + \frac{1}{2}h, x_1 + \frac{1}{2}k_1 h\right)$$
$$= f\left(1.5 + \frac{1}{2}1.5, 681.24 + \frac{1}{2}(-2346.8)1.5\right)$$
$$= f(2.25, -1078.9)$$
$$= 37.5 - 3.5(-1078.9)$$
$$= 3813.6$$
$$k_3 = f\left(t_1 + \frac{1}{2}h, x_1 + \frac{1}{2}k_2 h\right)$$
$$= f\left(1.5 + \frac{1}{2}1.5, 681.24 + \frac{1}{2}(3813.6)1.5\right)$$

$$= f(2.25, 3541.4)$$
$$= 37.5 - 3.5(3541.4)$$
$$= -12358$$
$$k_4 = f(t_1 + h, x_1 + k_3 h)$$
$$= f(1.5 + 1.5, 681.24 + (-12358)1.5)$$
$$= f(3, -17855)$$
$$= 37.5 - 3.5(-17855)$$
$$= 62530$$
$$x_2 = x_1 + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)h$$
$$= 681.24 + \frac{1}{6}(-2346.8 + 2(3813.6) + 2(-12358) + 62530)1.5$$
$$= 681.24 + \frac{1}{6}(43096)1.5$$
$$= 11455 \text{ g/L}$$

x_2 is the approximate concentration of salt at

$$t_2 = t_1 + h = 1.5 + 1.5 = 3 \text{ min}$$
$$x(3) \approx x_2 = 11455 \text{ g/L}$$

The exact solution of the ordinary differential equation is given by

$$x(t) = 10.714 + 39.286e^{-3.5t}$$

The solution to this nonlinear equation at $t = 3 \text{ min}$ is

$$x(3) = 10.715 \text{ g/L}$$

2.3 LINEAR ALGEBRA

2.3.1 Matrices

A matrix is an *n* x *m* array of numbers. In these notes a matrix is symbolized by a letter with a line on top, $\overline{B}$; *n* is the number of rows and *m* is the number of columns. If *n* = *m*, the matrix is said to be a *square matrix*. If the matrix has only one column(row) it is said to be a *column(row) matrix*. The *j*th *element* in the *i*th row of a matrix is indicated by subscripts, b_{ij}. Mathematically, an entity like a matrix is defined by a list of properties and operations, for instance the rules for adding or multiplying two matrices. Also, matrices can be regarded as one way to represent members of a group in Group Theory.

$$\vec{B} = \begin{bmatrix} b_{11} & b_{12} & b_{13} & b_{14} \\ b_{21} & b_{22} & b_{23} & b_{24} \\ b_{31} & b_{32} & b_{33} & b_{34} \end{bmatrix} \qquad \vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

2.3.2 Addition & Subtraction

a. Definition

The addition is carried out by adding the respective matrix elements.

$\vec{C} = \vec{A} + \vec{B}$

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

b. Rules

The sum of two matrices is also a matrix. Only matrices having the same number of rows and the same number of columns may be added. Matrix addition is commutative and associative.

$\vec{A} + \vec{B} = \vec{B} + \vec{A} \qquad (\vec{A} + \vec{B}) + \vec{C} = \vec{A} + (\vec{B} + \vec{C})$

2.3.3 Multiplication

a. Definition

$\vec{C} = \vec{A}\vec{B}$

$c_{ij} = \sum_k a_{ik} b_{kj} = a_{i1} b_{1j} + a_{i2} b_{2j} + a_{i3} b_{3j} + \cdots$

b. Rules

The product of two matrices is also a matrix. The number of elements in a row of $\vec{A}$ must equal the number of elements in a column of $\vec{B}$. Matrix multiplication is not commutative.

$\vec{A}\vec{B} \neq \vec{B}\vec{A}$

A matrix may be multiplied by a constant, thusly: $c_{ij} = q \cdot a_{ij}$. The result is also a matrix.

2.3.4 Inverse Matrix

a. Unit matrix

The *unit matrix* is a square matrix with the diagonal elements equal to one and the off-diagonal elements all equal to zero. Here's a 3x3 unit matrix:

$$\vec{U} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

b. Inverse

The *inverse* of a matrix, $\vec{B}$, (denoted $\vec{B}^{-1}$) is a matrix such that $\vec{B}\vec{B}^{-1} = \vec{B}^{-1}\vec{B} = \vec{U}$. The inverse of a particular matrix may not exist, in which case the matrix is said to be *singular*.

The solution of a system of simultaneous equations in effect is a problem of evaluating the inverse of a square matrix.

2.3.5 Simultaneous Linear Equations

1. The Problem

a. Simultaneous equations

We wish to solve a system of n linear equations in n unknowns.

$$b_{11}x_1 + b_{12}x_2 + \cdots b_{1n}x_n = c_1$$
$$b_{21}x_1 + b_{22}x_2 + \cdots b_{2n}x_n = c_2$$
$$\vdots$$
$$b_{n1}x_1 + b_{n2}x_2 + \cdots b_{nn}x_n = c_n$$

where the $\{b_{ij}\}$ and the $\{c_i\}$ are constants.

b. Matrix notation

The system of equations can be written as a matrix multiplication.

$$\vec{B}\vec{x} = \vec{c}, \text{ where } \vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \vec{c} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} \text{ and } \vec{B} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{bmatrix}.$$

When n is small ($n \le 40$, say) a direct or *one-step* method is used. For larger systems, iterative methods are preferred.

2.4 GAUSSIAN ELIMINATION

In a one-step approach, we seek to evaluate the inverse of the $\vec{B}$ matrix.

$$\vec{B}\vec{x} = \vec{c}$$
$$\vec{B}^{-1}\vec{B}\vec{x} = \vec{x} = \vec{B}^{-1}\vec{c}$$

The solution is obtained by carrying out the matrix multiplication $\vec{B}^{-1}\vec{c}$.

a. Elimination

You may have seen this in high school algebra. For brevity's sake, let's let $n = 3$.

$b_{11}x_1 + b_{12}x_2 + b_{13}x_3 = c_1$ 1

$b_{21}x_1 + b_{22}x_2 + b_{23}x_3 = c_2$ 2

$b_{31}x_1 + b_{32}x_2 + b_{33}x_3 = c_3$ 3

In essence, we wish to eliminate unknowns from the equations by a sequence of algebraic steps.

normalization i) multiply eqn. 1 by $-\frac{b_{21}}{b_{11}}$ and add to eqn. 2; replace eqn. 2.

reduction ii) multiply eqn 1 by $-\frac{b_{31}}{b_{11}}$ and add to eqn. 3; replace eqn. 3.

$b_{11}x_1 + b_{12}x_2 + b_{13}x_3 = c_1$

$b'_{22}x_2 + b'_{23}x_3 = c'_2$

$b'_{32}x_2 + b'_{33}x_3 = c'_3$

iii) multiply eqn. 2 by $-\frac{b'_{32}}{b'_{22}}$ and add to eqn. 3; replace eqn. 3.

$b_{11}x_1 + b_{12}x_2 + b_{13}x_3 = c_1$

$b'_{22}x_2 + b'_{23}x_3 = c'_2$

$b''_{33}x_3 = c''_3$

We have eliminated x_1 and x_2 from eqn.3 and x_1 from eqn. 2.

back substitution iv) solve eqn. 3 for x_3, substitute in eqn. 2 & 1.
solve eqn. 2 for x_2, substitute in eqn. 1.
solve eqn. 1 for x_1.

b. Pivoting

Due to the finite number of digits carried along by the machine, we have to worry about the relative magnitudes of the matrix elements, especially the diagonal elements. In other words, the inverse matrix, $\vec{B}^{-1}$ may be effectively singular even if not actually so. To minimize this possibility, we commonly rearrange the set of equations to place the largest coefficients on the diagonal, to the extent possible. This process is called *pivoting*.

e.g.

$37x_2 - 3x_3 = 4$

$19x_1 - 2x_2 + 48x_3 = 99$

$7x_1 + 0.6x_2 + 15x_3 = -9$

rearrange

$19x_1 - 2x_2 + 48x_3 = 99$

$37x_2 - 3x_3 = 4$

$7x_1 + 0.6x_2 + 15x_3 = -9$
or
$7x_1 + 0.6x_2 + 15x_3 = -9$
$37x_2 - 3x_3 = 4$
$19x_1 - 2x_2 + 48x_3 = 99$

2.4.1 Matrix Operations

In preparation for writing a computer program, we'll cast the elimination and back substitutic the form of matrix multiplications.

a. Augmented matrix

$$\vec{A} = \left[\vec{B} : \vec{c}\right] = \begin{bmatrix} b_{11} & b_{12} & b_{13} & c_1 \\ b_{21} & b_{22} & b_{23} & c_2 \\ b_{31} & b_{32} & b_{33} & c_3 \end{bmatrix}$$

b. Elementary matrices

Each single step is represented by a single matrix multiplication.

The elimination steps:

$$\vec{S}_1 = \begin{bmatrix} 1 & 0 & 0 \\ -\dfrac{b_{21}}{b_{11}} & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \vec{S}_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -\dfrac{b_{31}}{b_{11}} & 0 & 1 \end{bmatrix} \qquad \vec{S}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -\dfrac{b'_{32}}{b'_{22}} & 1 \end{bmatrix}$$

$$\vec{S}_3\vec{S}_2\vec{S}_1\vec{A} = \begin{bmatrix} b_{11} & b_{12} & b_{13} & c_1 \\ 0 & b'_{22} & b'_{23} & c'_2 \\ 0 & 0 & b''_{33} & c''_3 \end{bmatrix}$$

The first back substitution step:

$$\vec{Q}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \dfrac{1}{b''_{33}} \end{bmatrix}$$

$$\vec{Q}_1\vec{S}_3\vec{S}_2\vec{S}_1\vec{A} = \begin{bmatrix} b_{11} & b_{12} & b_{13} & c_1 \\ 0 & b'_{22} & b'_{23} & c'_2 \\ 0 & 0 & 1 & x_3 \end{bmatrix}$$

This completes one cycle. Next we eliminate one unknown from the second row using

$$\vec{S}_4 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -b'_{23} \\ 0 & 0 & 1 \end{bmatrix}$$

$$\vec{S}_4\vec{Q}_1\vec{S}_3\vec{S}_2\vec{S}_1\vec{A} = \begin{bmatrix} b_{11} & b_{12} & b_{13} & c_1 \\ 0 & b''_{22} & 0 & c''_2 \\ 0 & 0 & 1 & x_3 \end{bmatrix}$$

$$\vec{Q}_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \dfrac{1}{b''_{22}} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\vec{Q}_2\vec{S}_4\vec{Q}_1\vec{S}_3\vec{S}_2\vec{S}_1\vec{A} = \begin{bmatrix} b_{11} & b_{12} & b_{13} & c_1 \\ 0 & 1 & 0 & x_2 \\ 0 & 0 & 1 & x_3 \end{bmatrix}$$

This completes the second cycle. The final cycle is

$$\vec{S}_5 = \begin{bmatrix} 1 & 0 & -b_{13} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \vec{S}_6 = \begin{bmatrix} 1 & -b_{12} & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \vec{Q}_3 = \begin{bmatrix} \dfrac{1}{b_{11}} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\vec{Q}_3\vec{S}_6\vec{S}_5\vec{Q}_2\vec{S}_4\vec{Q}_1\vec{S}_3\vec{S}_2\vec{S}_1 = \begin{bmatrix} 1 & 0 & 0 & x_1 \\ 0 & 1 & 0 & x_2 \\ 0 & 0 & 1 & x_3 \end{bmatrix}$$

We identify the inverse matrix $\vec{B}^{-1} = \vec{Q}_3\vec{S}_6\vec{S}_5\vec{Q}_2\vec{S}_4\vec{Q}_1\vec{S}_3\vec{S}_2\vec{S}_1$. Notice that the order of the matrix multiplications is significant. Naturally, we want to automate this process, and generalize to *n* equations.

Example in chemical engineering

A liquid-liquid extraction process conducted in the Electrochemical Materials Laboratory involved the extraction of nickel from the aqueous phase into an organic phase. A typical set of experimental data from the laboratory is given below.

Ni aqueous phase, a $(\mathrm{g/l})$	2	2.5	3
Ni organic phase, g $(\mathrm{g/l})$	8.57	10	12

Assuming g is the amount of Ni in the organic phase and a is the amount of Ni in the aqueous phase, the quadratic interpolant that estimates g is given by

$$g = x_1 a^2 + x_2 a + x_3,\ 2 \le a \le 3$$

The solution for the unknowns x_1, x_2, and x_3 is given by

$$\begin{bmatrix} 4 & 2 & 1 \\ 6.25 & 2.5 & 1 \\ 9 & 3 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 8.57 \\ 10 \\ 12 \end{bmatrix}$$

Find the values of x_1, x_2, and x_3 using naïve Gauss elimination. Estimate the amount of nickel in the organic phase when $2.3\ \text{g/l}$ is in the aqueous phase using quadratic interpolation.

Solution

Forward Elimination of Unknowns

Since there are three equations, there will be two steps of forward elimination of unknowns.

First step

Divide Row 1 by 4 and then multiply it by 6.25, that is, multiply Row 1 by $6.25/4 = 1.5625$.

$$\text{Row } 1 \times (1.5625) = [6.25 \quad 3.125 \quad 1.5625] \qquad [13.391]$$

Subtract the result from Row 2 to get

$$\begin{bmatrix} 4 & 2 & 1 \\ 0 & -0.625 & -0.5625 \\ 9 & 3 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 8.57 \\ -3.3906 \\ 12 \end{bmatrix}$$

Divide Row 1 by 4 and then multiply it by 9, that is, multiply Row 1 by $9/4 = 2.25$.

$$\text{Row } 1 \times (2.25) = [9 \quad 4.5 \quad 2.25] \qquad [19.283]$$

Subtract the result from Row 3 to get

$$\begin{bmatrix} 4 & 2 & 1 \\ 0 & -0.625 & -0.5625 \\ 0 & -1.5 & -1.25 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 8.57 \\ -3.3906 \\ -7.2825 \end{bmatrix}$$

Second step

We now divide Row 2 by −0.625 and then multiply it by −1.5, that is, multiply Row 2 by $-1.5/-0.625 = 2.4$.

$$\text{Row } 2 \times (2.4) = [0 \quad -1.5 \quad -1.35] \qquad [-8.1375]$$

Subtract the result from Row 3 to get

$$\begin{bmatrix} 4 & 2 & 1 \\ 0 & -0.625 & -0.5625 \\ 0 & 0 & 0.1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 8.57 \\ -3.3906 \\ 0.855 \end{bmatrix}$$

Back Substitution

From the third equation,

$$0.1x_3 = 0.855$$

$$x_3 = \frac{0.855}{0.1}$$

$$= 8.55$$

Substituting the value of x_3 in the second equation,

$$(-0.625)x_2 + (-0.5625)x_3 = -3.3906$$

$$x_2 = \frac{-3.3906 - (-0.5625)x_3}{-0.625}$$

$$= \frac{-3.3906 - (-0.5625) \times 8.55}{-0.625}$$

$$= -2.27$$

Substituting the values of x_2 and x_3 in the first equation,

$$4x_1 + 2x_2 + x_3 = 8.57$$

$$x_1 = \frac{8.57 - 2x_2 - x_3}{4}$$

$$= \frac{8.57 - 2 \times (-2.27) - 8.55}{4}$$

$$= 1.14$$

Hence the solution vector is

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1.14 \\ -2.27 \\ 8.55 \end{bmatrix}$$

The polynomial that passes through the three data points is then

$$g(a) = x_1 a^2 + x_2 a + x_3$$

$$= 1.14a^2 + (-2.27)a + 8.55$$

where g is the amount of nickel in the organic phase and a is the amount of nickel in the aqueous phase.

When $2.3\,\text{g/l}$ is in the aqueous phase, using quadratic interpolation, the estimated amount of nickel in the organic phase is

$$g(2.3) = 1.14 \times (2.3)^2 + (-2.27) \times (2.3) + 8.55$$

$$= 9.3596 \text{ g/l}$$

2.5 GAUSS-SEIDEL METHOD

Example in Chemical Engineering

A liquid-liquid extraction process conducted in the Electrochemical Materials Laboratory involved the extraction of nickel from the aqueous phase into an organic phase. A typical set of experimental data from the laboratory is given below.

Ni aqueous phase, $a(\text{g/l})$	2	2.5	3
Ni organic phase, $g(\text{g/l})$	8.57	10	12

Assuming g is the amount of Ni in the organic phase and a is the amount of Ni in the aqueous phase, the quadratic interpolant that estimates g is given by

$$g = x_1 a^2 + x_2 a + x_3, 2 \le a \le 3$$

The solution for the unknowns x_1, x_2, and x_3 is given by

$$\begin{bmatrix} 4 & 2 & 1 \\ 6.25 & 2.5 & 1 \\ 9 & 3 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 8.57 \\ 10 \\ 12 \end{bmatrix}$$

Find the values of x_1, x_2, and x_3 using the Gauss-Seidel method. Estimate the amount of nickel in the organic phase when 2.3 g/l is in the aqueous phase using quadratic interpolation. Use

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

as the initial guess and conduct two iterations.

Solution

Rewriting the equations gives

$$x_1 = \frac{8.57 - 2x_2 - x_3}{4}$$

$$x_2 = \frac{10 - 6.25x_1 - x_3}{2.5}$$

$$x_3 = \frac{12 - 9x_1 - 3x_2}{1}$$

Iteration #1

Given the initial guess of the solution vector as

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

we get

$$x_1 = \frac{8.57 - 2 \times 1 - 1}{4}$$
$$= 1.3925$$
$$x_2 = \frac{10 - 6.25 \times 1.3925 - 1}{2.5}$$
$$= 0.11875$$
$$x_3 = \frac{12 - 9 \times 1.3925 - 3 \times 0.11875}{1}$$
$$= -0.88875$$

The absolute relative approximate error for each x_i then is

$$|\epsilon_a|_1 = \left|\frac{1.3925 - 1}{1.3925}\right| \times 100$$
$$= 28.187\%$$
$$|\epsilon_a|_2 = \left|\frac{0.11875 - 1}{0.11875}\right| \times 100$$
$$= 742.11\%$$
$$|\epsilon_a|_3 = \left|\frac{-0.88875 - 1}{-0.88875}\right| \times 100$$
$$= 212.52\%$$

At the end of the first iteration, the estimate of the solution vector is

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1.3925 \\ 0.11875 \\ -0.88875 \end{bmatrix}$$

and the maximum absolute relative approximate error is 742.11%.

Iteration #2

The estimate of the solution vector at the end of Iteration #1 is

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1.3925 \\ 0.11875 \\ -0.88875 \end{bmatrix}$$

Now we get

$$x_1 = \frac{8.57 - 2 \times 0.11875 - (-0.88875)}{4}$$

$$= 2.3053$$

$$x_2 = \frac{10 - 6.25 \times 2.3053 - (-0.88875)}{2.5}$$

$$= -1.4078$$

$$x_3 = \frac{12 - 9 \times 2.3053 - 3 \times (-1.4078)}{1}$$

$$= -4.5245$$

The absolute relative approximate error for each x_i then is

$$|\epsilon_a|_1 = \left|\frac{2.3053 - 1.3925}{2.3053}\right| \times 100$$

$$= 39.596\%$$

$$|\epsilon_a|_2 = \left|\frac{-1.4078 - 0.11875}{-1.4078}\right| \times 100$$

$$= 108.44\%$$

$$|\epsilon_a|_3 = \left|\frac{-4.5245 - (-0.88875)}{-4.5245}\right| \times 100$$

$$= 80.357\%$$

At the end of the second iteration, the estimate of the solution vector is

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2.3053 \\ -1.4078 \\ -4.5245 \end{bmatrix}$$

and the maximum absolute relative approximate error is 108.44%.

Conducting more iterations gives the following values for the solution vector and the corresponding absolute relative approximate errors.

Iteration	x_1	$\lvert\epsilon_a\rvert_1\%$	x_2	$\lvert\epsilon_a\rvert_2\%$	x_3	$\lvert\epsilon_a\rvert_3\%$
1	1.3925	28.1867	0.11875	742.1053	–0.88875	212.52
2	2.3053	39.5960	–1.4078	108.4353	–4.5245	80.357
3	3.9775	42.041	–4.1340	65.946	–11.396	60.296
4	7.0584	43.649	–9.0877	54.510	–24.262	53.032
5	12.752	44.649	–18.175	49.999	–48.243	49.708
6	23.291	45.249	–34.930	47.967	–92.827	48.030

After six iterations, the absolute relative approximate errors are not decreasing much. In fact, conducting more iterations reveals that the absolute relative approximate error converges to a value of 46.070% for all three values with the solution vector diverging from the exact solution drastically.

Iteration	x_1	$\lvert\epsilon_a\rvert_1\%$	x_2	$\lvert\epsilon_a\rvert_2\%$	x_3	$\lvert\epsilon_a\rvert_3\%$
32	2.1428×10^{8}	46.0703	-3.3920×10^{8}	46.0703	-9.1095×10^{8}	46.0703

The exact solution vector is

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1.14 \\ -2.27 \\ 8.55 \end{bmatrix}$$

To correct this, the coefficient matrix needs to be more diagonally dominant. To achieve a more diagonally dominant coefficient matrix, rearrange the system of equations by exchanging equations one and three.

$$\begin{bmatrix} 9 & 3 & 1 \\ 6.25 & 2.5 & 1 \\ 4 & 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 12 \\ 10 \\ 8.57 \end{bmatrix}$$

Iteration #1

Given the initial guess of the solution vector as

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

we get

$$x_1 = \frac{12-3\times1-1}{9}$$
$$= 0.88889$$
$$x_2 = \frac{10-6.25\times0.88889-1}{2.5}$$
$$= 1.3778$$
$$x_3 = \frac{8.57-4\times0.88889-2\times1.3778}{1}$$
$$= 2.2589$$

The absolute relative approximate error for each x_i then is

$$\lvert\epsilon_a\rvert_1 = \left|\frac{0.88889-1}{0.88889}\right|\times100$$
$$= 12.5\%$$
$$\lvert\epsilon_a\rvert_2 = \left|\frac{1.3778-1}{1.3778}\right|\times100$$
$$= 27.419\%$$

$$|\epsilon_a|_3 = \left|\frac{2.2589-1}{2.2589}\right| \times 100$$
$$= 55.730\%$$

At the end of the first iteration, the estimate of the solution vector is

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0.88889 \\ 1.3778 \\ 2.2589 \end{bmatrix}$$

and the maximum absolute relative approximate error is 55.730%.

Iteration #2

The estimate of the solution vector at the end of Iteration #1 is

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0.88889 \\ 1.3778 \\ 2.2589 \end{bmatrix}$$

Now we get

$$x_1 = \frac{12 - 3 \times 1.3778 - 1 \times 2.2589}{9}$$
$$= 0.62309$$
$$x_2 = \frac{10 - 6.25 \times 0.62309 - 1 \times 2.2589}{2.5}$$
$$= 1.5387$$
$$x_3 = \frac{8.57 - 4 \times 0.62309 - 2 \times 1.5387}{1}$$
$$= 3.0002$$

The absolute relative approximate error for each x_i then is

$$|\epsilon_a|_1 = \left|\frac{0.62309 - 0.88889}{0.62309}\right| \times 100$$
$$= 42.659\%$$
$$|\epsilon_a|_2 = \left|\frac{1.5387 - 1.3778}{1.5387}\right| \times 100$$
$$= 10.460\%$$
$$|\epsilon_a|_3 = \left|\frac{3.0002 - 2.2589}{3.0002}\right| \times 100$$
$$= 24.709\%$$

At the end of the second iteration, the estimate of the solution is

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0.62309 \\ 1.5387 \\ 3.0002 \end{bmatrix}$$

and the maximum absolute relative approximate error is 42.659%.

Conducting more iterations gives the following values for the solution vector and the corresponding absolute relative approximate errors.

Iteration	x_1	$\lvert\epsilon_a\rvert_1\%$	x_2	$\lvert\epsilon_a\rvert_2\%$	x_3	$\lvert\epsilon_a\rvert_3\%$
1	0.88889	12.5	1.3778	27.419	2.2589	55.730
2	0.62309	42.659	1.5387	10.456	3.0002	24.709
3	0.48707	27.926	1.5822	2.7506	3.4572	13.220
4	0.42178	15.479	1.5627	1.2537	3.7576	7.9928
5	0.39494	6.7960	1.5096	3.5131	3.9710	5.3747
6	0.38890	1.5521	1.4393	4.8828	4.1357	3.9826

After six iterations, the absolute relative approximate errors seem to be decreasing. Conducting more iterations allows the absolute relative approximate error decrease to an acceptable level.

Iteration	x_1	$\lvert\epsilon_a\rvert_1\%$	x_2	$\lvert\epsilon_a\rvert_2\%$	x_3	$\lvert\epsilon_a\rvert_3\%$
199	1.1335	0.014412	–2.2389	0.034871	8.5139	0.010666
200	1.1337	0.014056	–2.2397	0.034005	8.5148	0.010403

This is close to the exact solution vector of

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1.14 \\ -2.27 \\ 8.55 \end{bmatrix}$$

The polynomial that passes through the three data points is then

$$\begin{aligned} g(a) &= x_1(a)^2 + x_2(a) + x_3 \\ &= 1.1337(a)^2 + (-2.2397)(a) + 8.5148 \end{aligned}$$

where g is the amount of nickel in the organic phase and a is the amount of nickel in the aqueous phase.

When $2.3\,g/l$ is in the aqueous phase, using quadratic interpolation, the estimated amount of nickel in the organic phase is

$$\begin{aligned} g(2.3) &= 1.1337(2.3)^2 + (-2.2397) \times (2.3) + 8.5148 \\ &= 9.3608\,\text{g/l} \end{aligned}$$

2.6 SUCCESSIVE OVER-RELAXATION

Relaxation moves towards solution faster:

$$x_i^{(k+1)} = x_i^{(k)} + \omega R_i$$

From Gauss-Seidel:

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

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

This can be written: $x_i^{k+1} = x_i^k + R_i$

Where R_i is the "Residual" (error or change)

Now, use relaxation (ω) to speed convergence:

$$x_i^{(k+1)} = x_i^{(k)} + \omega R_i$$

2.7 LINEAR TRANSFORMATIONS

A **function** F from the set V to the set W is a rule that assigns to each element x in V exactly one element F(x) in W.

If V and W are vector spaces , the function F is called a **mapping** or a **transformation** from the vector space V to the vector space W. It is denoted by F: $V \rightarrow W$.
In this case, the vector **w** = F(**v**) is called the image of the vector **v** under the transformation F.

Example 1: Let **u** = (x, y, z) be any vector in $\Re^3$. Then

$T(\mathbf{u}) = (3x - 2y + z, x + y - 2z)$

defines a transformation T from $\Re^3$ to $\Re^2$.

Observe that if

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

then the transformation T: $\Re^3 \rightarrow \Re^2$ given above, can be written as

$T(\mathbf{u}) = A\mathbf{u}$.

In this case, we call the transformation T: $\Re^3 \rightarrow \Re^2$ a **matrix transformation**, and the 2 x 3 matrix A is called the **matrix of the transformation**.

Definition: If V and W are vector spaces, then the mapping T: $V \rightarrow W$ is a **linear transformation** provided that

(a) $T(\mathbf{u} + \mathbf{v}) = T(\mathbf{u}) + T(\mathbf{v})$, and
(b) $T(c\mathbf{u}) = cT(\mathbf{u})$.

Theorem 1: If V and W are vector spaces, then the transformation T: $V \rightarrow W$ is a linear if and only if

$T(a\mathbf{u} + b\mathbf{v}) = aT(\mathbf{u}) + bT(\mathbf{v})$

for all pairs of vectors **u** and **v** in V and all pairs of scalars a and b, i.e. a transformation between two vector spaces is linear if, and only if, it preserves linear combinations of pairs of vectors.

Theorem 2: The mapping T: $\Re^n \rightarrow \Re^m$ is a linear transformation if and only if it is a matrix transformation. The matrix A of the transformation T is given by

$$A = [T(e_1) \;\; T(e_2) \;\; \ldots \;\; T(e_n)]$$

where $T(e_j)$ is the image under T of the *jth standard unit basis vector* $e_j = (0,...,1,...,0)$ with 1 in the j-th position.

Importance of Theorem 2: If the effect of the transformation on each standard unit basis vector is known, we can apply Theorem 2 to obtain the matrix of the transformation and thereby obtain a formula for the transformation.

Describing a linear transformation T: $\Re^2 \rightarrow \Re^2$ in geometrical terms:

I. Reflection in the axes

Consider the effect of the transformation T on each standard unit basis vector of $\Re^2$ given by $T(e_1) = (1, 0)$, and $T(e_2) = (0, -1)$.
The matrix of the transformation is

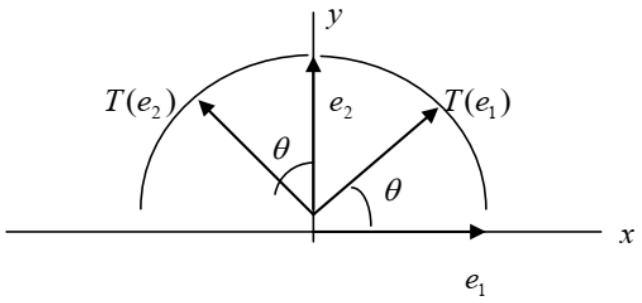
$$A = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix},$$

and
$T(x, y) = (x, -y)$.

II. Rotation

Consider the effect of the transformation T on each standard unit basis vector of $\Re^2$ given by $T(e_1) = (\cos\theta, \sin\theta)$, and $T(e_2) = (\cos(\theta + \frac{\pi}{2}), \sin(\theta + \frac{\pi}{2})) = (-\sin\theta, \cos\theta)$.

Figure:



The matrix of the transformation is

$$A = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix},$$

and
$T(x, y) = (x\cos\theta - y\sin\theta,\ x\sin\theta + y\cos\theta)$.

III. Expansion(Compression) in the *x*-direction

Consider the effect of the transformation T on each standard unit basis vector of $\Re^2$ given by $T(e_1) = (c, 0)$, and $T(e_2) = (0, 1)$, $c > 0$
The matrix of the transformation is

$$A = \begin{bmatrix} c & 0 \\ 0 & 1 \end{bmatrix},$$

and
$T(x, y) = (cx, y)$.

- If $c > 1$, the transformation is an expansion
- If $0 < c < 1$, the transformation is a compression
- If $c = 1$, the transformation is the **Identity transformation**

IV. Shear in the *x*-direction

Consider the effect of the transformation T on each standard unit basis vector of $\Re^2$ given by $T(e_1) = (1, 0)$, and $T(e_2) = (c, 1)$.
The matrix of the transformation is

$$A = \begin{bmatrix} 1 & c \\ 0 & 1 \end{bmatrix},$$

and

$T(x, y) = (x + cy, y)$.

Theorem: Suppose that the linear transformation $\boldsymbol{T}: \Re^2 \rightarrow \Re^2$ corresponds to a nonsingular matrix A. Then T is a finite composition of reflections, expansions, compressions, and shears.

Example: Suppose that $\boldsymbol{T}: \Re^2 \rightarrow \Re^2$ is defined by $\boldsymbol{T}(\mathbf{x}) = A\mathbf{x}$ where

$$A = \begin{bmatrix} 2 & 6 \\ 1 & 4 \end{bmatrix}$$

We reduce A to I as follows:

$$\begin{bmatrix} 2 & 6 \\ 1 & 4 \end{bmatrix} \rightarrow (swap\ r_1\ and\ r_2) \rightarrow \begin{bmatrix} 1 & 4 \\ 2 & 6 \end{bmatrix} \rightarrow (-2r_1 + r_2) \rightarrow \begin{bmatrix} 1 & 4 \\ 0 & -2 \end{bmatrix} \rightarrow (-.5r_2) \rightarrow$$
$$\begin{bmatrix} 1 & 4 \\ 0 & 1 \end{bmatrix} \rightarrow (-4r_2 + r_1) \rightarrow \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

The four elementary matrices corresponding to the four row operations used above are:

$$E_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad E_2 = \begin{bmatrix} 1 & 0 \\ -2 & 1 \end{bmatrix}, \quad E_3 = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{bmatrix}, \quad E_4 = \begin{bmatrix} 1 & -4 \\ 0 & 1 \end{bmatrix}.$$

We know that

$$E_4E_3E_2E_1A = I$$

Hence

$$A = E_1^{-1}E_2^{-1}E_3^{-1}E_4^{-1}$$

i.e.

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & 4 \\ 0 & 1 \end{bmatrix}$$

2.8 SINGULAR VALUE DECOMPOSITION

The *singular value decomposition*, or SVD, is a very powerful and useful matrix decomposition, particularly in the context of data analysis, dimension reducing transformations of images, satellite data etc, and is the method of choice for solving most *linear least–squares* problems.

SVD methods are based on the following theorem of linear algebra (whose proof may be sought elsewhere):

Theorem *Singular Value Decomposition (SVD). Let* **A** *be a real* $m \times n$ *matrix. There exist orthogonal matrices* **S** *and* **C** *such that*

$$\mathbf{A} = \mathbf{S}\boldsymbol{\Sigma}\mathbf{C}^T \qquad (1)$$

where **S** *is* $m \times m$, **C** *is* $n \times n$, *and* $\boldsymbol{\Sigma}$ *is* $m \times n$ and *has the special diagonal form*

when $m > n$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{pmatrix}$$

or *when* $m < n$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1 & & 0 & 0 & \cdots & 0 \\ & \ddots & & \vdots & & \vdots \\ 0 & & \sigma_m & 0 & \cdots & 0 \end{pmatrix}$$

The entries of $\boldsymbol{\Sigma}$ *are ordered in descending order according to*

$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_l \geq 0$, *where* $l = \min\{m, n\}$

The columns of **S** *are called the left–singular vectors, the columns of* **C** *the right–singular vectors, and the diagonal elements of* $\boldsymbol{\Sigma}$ *the singular values of the matrix* **A**.

To establish the decomposition given by Equation (1), we first (matrix) multiply from the right by **C** to obtain

$$\mathbf{AC} = \mathbf{S\Sigma} \qquad (2)$$

The *i*th column of this relationship is

$$\mathbf{Ac}_i = \sigma_i \mathbf{s}_i \qquad (3)$$

for $i = 1,\ldots,n$. Note that Equation (3) shows that $\mathbf{s}_i$ may be calculated directly from knowledge of $\mathbf{A}$, $\mathbf{c}_i$, and σ_i.

We get another relation by taking the transpose of Equation (1)

$$\mathbf{A}^T = \mathbf{C\Sigma}^T\mathbf{S}^T \qquad (4)$$

and then (matrix) multiply from the right by **S** to obtain

$$\mathbf{A}^T\mathbf{S} = \mathbf{C\Sigma}^T \qquad (5)$$

The *i*th column of this relationship is

$$\mathbf{A}^T\mathbf{s}_i = \sigma_i \mathbf{c}_i \qquad (6)$$

where again $i = 1,\ldots,n$. Note that Equation (6) shows that $\mathbf{c}_i$ may be calculated directly from knowledge of $\mathbf{A}$, $\mathbf{s}_i$, and σ_i.

The associated eigenvalue problems.

There are two eigenvalue problems that can be obtained from the SVD. For the first eigenvalue problem we start with Equation (2) and multiply from the left by $\mathbf{A}^T$

$$\begin{aligned}\mathbf{A}^T\mathbf{AC} &= \mathbf{A}^T\mathbf{S\Sigma} \\ &= (\mathbf{S\Sigma C}^T)^T\mathbf{S\Sigma} \\ &= \mathbf{C\Sigma}^T\mathbf{S}^T\mathbf{S\Sigma} \\ &= \mathbf{C\Sigma}^T\mathbf{\Sigma} \\ &= \mathbf{C\Sigma}^2\end{aligned} \qquad (7)$$

where (assuming $m > n$)

$$\boldsymbol{\Sigma}^2 = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{pmatrix} \qquad (8)$$

Let $\mathbf{R}_1 = \mathbf{A}^T\mathbf{A}$ and $\boldsymbol{\Lambda}_1 = \boldsymbol{\Sigma}^2$, then we can write Equation (7) as the eigenvalue problem

$$\mathbf{R}_1\mathbf{C} = \mathbf{C}\Lambda_1 \qquad (9)$$

For the second eigenvalue problem we start with Equation (5) and multiply from the left by $\mathbf{A}$

$$\begin{aligned} \mathbf{A}\mathbf{A}^T\mathbf{S} &= (\mathbf{S}\boldsymbol{\Sigma}\mathbf{C}^T)\mathbf{C}\boldsymbol{\Sigma}^T \\ &= \mathbf{S}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T \end{aligned} \qquad (10)$$

Let $\mathbf{R}_2 = \mathbf{A}\mathbf{A}^T$ and $\boldsymbol{\Lambda}_2 = \boldsymbol{\Sigma}\boldsymbol{\Sigma}^T$, then we can write Equation (10) as the eigenvalue problem

$$\mathbf{R}_2\mathbf{S} = \mathbf{S}\Lambda_2 \qquad (11)$$

The Thin SVD

Since

$$\boldsymbol{\Lambda}_2 = \boldsymbol{\Sigma}\boldsymbol{\Sigma}^T = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} \sigma_1 & & 0 & 0 & \cdots & 0 \\ & \ddots & & \vdots & & \vdots \\ 0 & & \sigma_n & 0 & \cdots & 0 \end{pmatrix}$$

which generates a square $m \times m$ matrix with diagonal elements

$$\mathbf{\Lambda}_2 = \begin{pmatrix} \sigma_1^2 & & 0 & 0 & \cdots & 0 \\ & \ddots & & \vdots & & \vdots \\ 0 & & \sigma_n^2 & \vdots & & \vdots \\ 0 & \cdots & \cdots & 0 & \cdots & 0 \\ \vdots & & & & \ddots & \vdots \\ 0 & \cdots & \cdots & & \cdots & 0 \end{pmatrix}$$

Because the diagonal elements $\Lambda_{kk} = 0$ for $k = n+1, \ \ldots \ , m$, the eigenvectors (singular vectors) $\mathbf{s}_{n+1}, \ \ldots \ , \mathbf{s}_m$ are of no importance. As a result we define a new $m \times n$ matrix $\hat{\mathbf{S}}$ (it is $\mathbf{S}$ with the last $m-n$ columns deleted) and a new $n \times n$ diagonal matrix $\hat{\mathbf{\Sigma}}$ (whose diagonal elements are $\sigma_1, \ \ldots \ , \sigma_n$) and write the *thin* SVD (or *reduced* SVD) of $\mathbf{A}$ as

$$\mathbf{A} = \hat{\mathbf{S}}\hat{\mathbf{\Sigma}}\mathbf{C}^T \qquad (12)$$