

# Chapter 01.01

## Introduction to Numerical Methods

*After reading this chapter, you should be able to:*

1. *understand the need for numerical methods, and*
2. *go through the stages (mathematical modeling, solving and implementation) of solving a particular physical problem.*

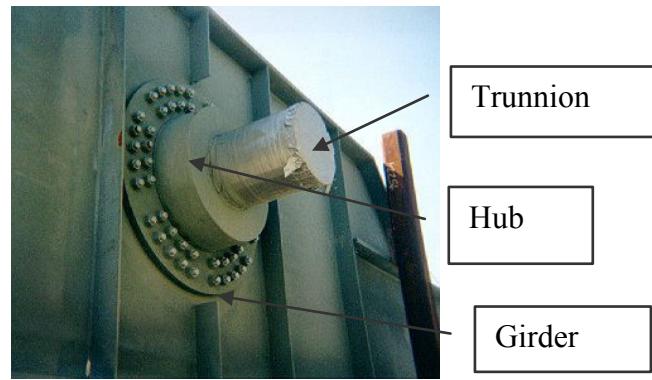
Mathematical models are an integral part in solving engineering problems. Many times, these mathematical models are derived from engineering and science principles, while at other times the models may be obtained from experimental data.

Mathematical models generally result in need of using mathematical procedures that include but are not limited to

- (A) differentiation,
- (B) nonlinear equations,
- (C) simultaneous linear equations,
- (D) curve fitting by interpolation or regression,
- (E) integration, and
- (F) differential equations.

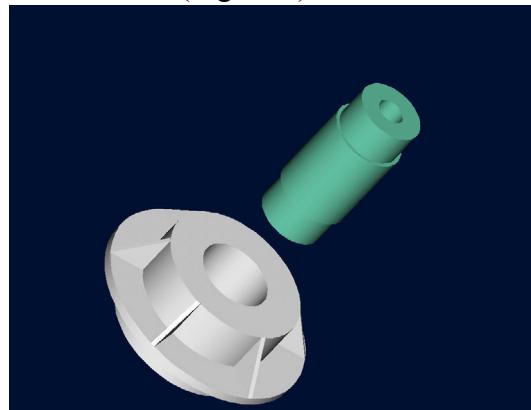
These mathematical procedures may be suitable to be solved exactly as you must have experienced in the series of calculus courses you have taken, but in most cases, the procedures need to be solved approximately using numerical methods. Let us see an example of such a need from a real-life physical problem.

To make the fulcrum (Figure 1) of a bascule bridge, a long hollow steel shaft called the trunnion is shrink fit into a steel hub. The resulting steel trunnion-hub assembly is then shrink fit into the girder of the bridge.



**Figure 1** Trunnion-Hub-Girder (THG) assembly.

This is done by first immersing the trunnion in a cold medium such as a dry-ice/alcohol mixture. After the trunnion reaches the steady state temperature of the cold medium, the trunnion outer diameter contracts. The trunnion is taken out of the medium and slid through the hole of the hub (Figure 2).



**Figure 2** Trunnion滑ed through the hub after contracting

When the trunnion heats up, it expands and creates an interference fit with the hub. In 1995, on one of the bridges in Florida, this assembly procedure did not work as designed. Before the trunnion could be inserted fully into the hub, the trunnion got stuck. Luckily, the trunnion was taken out before it got stuck permanently. Otherwise, a new trunnion and hub would needed to be ordered at a cost of \$50,000. Coupled with construction delays, the total loss could have been more than a hundred thousand dollars.

Why did the trunnion get stuck? This was because the trunnion had not contracted enough to slide through the hole. Can you find out why?

A hollow trunnion of outside diameter 12.363" is to be fitted in a hub of inner diameter 12.358". The trunnion was put in dry ice/alcohol mixture (temperature of the fluid - dry ice/alcohol mixture is  $-108^{\circ}\text{F}$ ) to contract the trunnion so that it can be slid through the hole of the hub. To slide the trunnion without sticking, a diametrical clearance of at least 0.01" is required between the trunnion and the hub. Assuming the room temperature is  $80^{\circ}\text{F}$ , is immersing the trunnion in dry-ice/alcohol mixture a correct decision?

To calculate the contraction in the diameter of the trunnion, the thermal expansion coefficient at room temperature is used. In that case the reduction  $\Delta D$  in the outer diameter of the trunnion is

$$\Delta D = D\alpha\Delta T \quad (1)$$

where

$D$  = outer diameter of the trunnion,

$\alpha$  = coefficient of thermal expansion coefficient at room temperature, and

$\Delta T$  = change in temperature,

Given

$$D = 12.363"$$

$$\alpha = 6.47 \times 10^{-6} \text{ in/in}/^{\circ}\text{F} \text{ at } 80^{\circ}\text{F}$$

$$\Delta T = T_{\text{fluid}} - T_{\text{room}}$$

$$= -108 - 80$$

$$= -188^{\circ}\text{F}$$

where

$T_{fluid}$  = temperature of dry-ice/alcohol mixture

$T_{room}$  = room temperature

the reduction in the outer diameter of the trunnion is given by

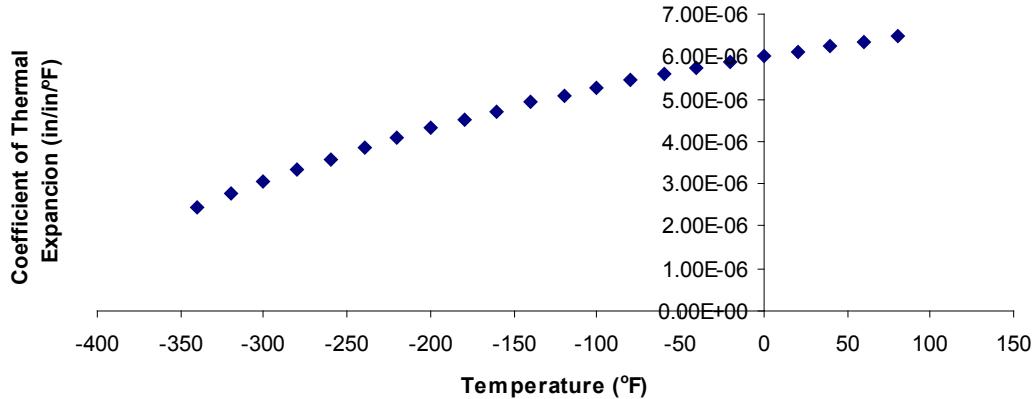
$$\begin{aligned}\Delta D &= (12.363)(6.47 \times 10^{-6})(-188) \\ &= -0.01504''\end{aligned}$$

So the trunnion is predicted to reduce in diameter by 0.01504". But, is this enough reduction in diameter? As per specifications, the trunnion needs to contract by

$$\begin{aligned}&= \text{trunnion outside diameter} - \text{hub inner diameter} + \text{diametric clearance} \\ &= 12.363 - 12.358 + 0.01 \\ &= 0.015''\end{aligned}$$

So according to his calculations, immersing the steel trunnion in dry-ice/alcohol mixture gives the desired contraction of greater than 0.015" as the predicted contraction is 0.01504". But, when the steel trunnion was put in the hub, it got stuck. Why did this happen? Was our mathematical model adequate for this problem or did we create a mathematical error?

As shown in Figure 3 and Table 1, the thermal expansion coefficient of steel decreases with temperature and is not constant over the range of temperature the trunnion goes through. Hence, Equation (1) would overestimate the thermal contraction.



**Figure 3** Varying thermal expansion coefficient as a function of temperature for cast steel.

The contraction in the diameter of the trunnion for which the thermal expansion coefficient varies as a function of temperature is given by

$$\Delta D = D \int_{T_{room}}^{T_{fluid}} \alpha dT \quad (2)$$

So one needs to curve fit the data to find the coefficient of thermal expansion as a function of temperature. This is done by regression where we best fit a curve through the data given in Table 1. In this case, we may fit a second order polynomial

$$\alpha = a_0 + a_1 \times T + a_2 \times T^2 \quad (3)$$

**Table 1** Instantaneous thermal expansion coefficient as a function of temperature.

Temperature °F	Instantaneous Thermal Expansion μin/in/°F
80	6.47
60	6.36
40	6.24
20	6.12
0	6.00
-20	5.86
-40	5.72
-60	5.58
-80	5.43
-100	5.28
-120	5.09
-140	4.91
-160	4.72
-180	4.52
-200	4.30
-220	4.08
-240	3.83
-260	3.58
-280	3.33
-300	3.07
-320	2.76
-340	2.45

The values of the coefficients in the above Equation (3) will be found by polynomial regression (we will learn how to do this later in Chapter 06.04). At this point we are just going to give you these values and they are

$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 6.0150 \times 10^{-6} \\ 6.1946 \times 10^{-9} \\ -1.2278 \times 10^{-11} \end{bmatrix}$$

to give the polynomial regression model (Figure 4) as

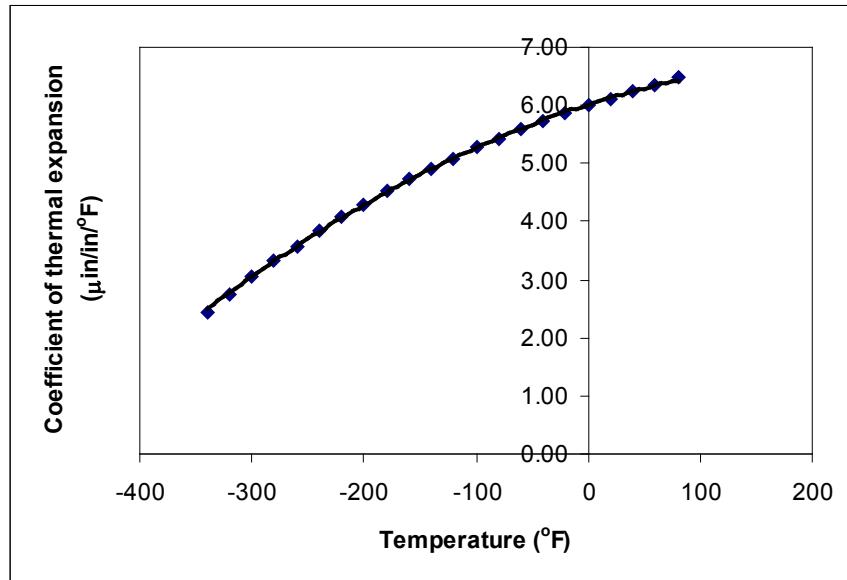
$$\begin{aligned} \alpha &= a_0 + a_1 T + a_2 T^2 \\ &= 6.0150 \times 10^{-6} + 6.1946 \times 10^{-9} T - 1.2278 \times 10^{-11} T^2 \end{aligned}$$

Knowing the values of  $a_0$ ,  $a_1$  and  $a_2$ , we can then find the contraction in the trunnion diameter as

$$\begin{aligned} \Delta D &= D \int_{T_{room}}^{T_{fluid}} (a_0 + a_1 T + a_2 T^2) dT \\ &= D[a_0(T_{fluid} - T_{room}) + a_1 \frac{(T_{fluid}^2 - T_{room}^2)}{2} + a_2 \frac{(T_{fluid}^3 - T_{room}^3)}{3}] \end{aligned} \quad (4)$$

which gives

$$\Delta D = 12.363 \left[ 6.0150 \times 10^{-6} \times (-108 - 80) + 6.1946 \times 10^{-9} \frac{((-108)^2 - (80)^2)}{2} - 1.2278 \times 10^{-12} \frac{((-108)^3 - (80)^3)}{3} \right] \\ = -0.013689"$$



**Figure 4** Second order polynomial regression model for coefficient of thermal expansion as a function of temperature.

What do we find here? The contraction in the trunnion is not enough to meet the required specification of 0.015".

So here are some questions that you may want to ask yourself?

1. What if the trunnion were immersed in liquid nitrogen (boiling temperature =  $-321^{\circ}\text{F}$ )? Will that cause enough contraction in the trunnion?
2. Rather than regressing the thermal expansion coefficient data to a second order polynomial so that one can find the contraction in the trunnion OD, how would you use Trapezoidal rule of integration for unequal segments? What is the relative difference between the two results?
3. We chose a second order polynomial for regression. Would a different order polynomial be a better choice for regression? Is there an optimum order of polynomial you can find?

As mentioned at the beginning of this chapter, we generally see mathematical procedures that require the solution of nonlinear equations, differentiation, solution of simultaneous linear equations, interpolation, regression, integration, and differential equations. A physical example to illustrate the need for each of these mathematical procedures is given in the beginning of each chapter. You may want to look at them now to understand better why we need numerical methods in everyday life.

---

**INTRODUCTION, APPROXIMATION AND ERRORS**

---

Topic      Introduction to Numerical Methods  
Summary    Textbook notes of Introduction to Numerical Methods  
Major      General Engineering  
Authors     Autar Kaw  
Date       January 27, 2011  

---

Web Site    <http://numericalmethods.eng.usf.edu>

---

## Chapter 01.02

# Measuring Errors

After reading this chapter, you should be able to:

1. find the true and relative true error,
2. find the approximate and relative approximate error,
3. relate the absolute relative approximate error to the number of significant digits at least correct in your answers, and
4. know the concept of significant digits.

In any numerical analysis, errors will arise during the calculations. To be able to deal with the issue of errors, we need to

- (A) identify where the error is coming from, followed by
- (B) quantifying the error, and lastly
- (C) minimize the error as per our needs.

In this chapter, we will concentrate on item (B), that is, how to quantify errors.

**Q:** What is true error?

**A:** True error denoted by  $E_t$  is the difference between the true value (also called the exact value) and the approximate value.

$$\text{True Error} = \text{True value} - \text{Approximate value}$$

### Example 1

The derivative of a function  $f(x)$  at a particular value of  $x$  can be approximately calculated by

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

of  $f'(2)$  For  $f(x) = 7e^{0.5x}$  and  $h = 0.3$ , find

- a) the approximate value of  $f'(2)$
- b) the true value of  $f'(2)$
- c) the true error for part (a)

### Solution

a) 
$$f'(x) \approx \frac{f(x+h) - f(x)}{h}$$

For  $x = 2$  and  $h = 0.3$ ,

$$\begin{aligned} f'(2) &\approx \frac{f(2+0.3) - f(2)}{0.3} \\ &= \frac{f(2.3) - f(2)}{0.3} \\ &= \frac{7e^{0.5(2.3)} - 7e^{0.5(2)}}{0.3} \\ &= \frac{22.107 - 19.028}{0.3} \\ &= 10.265 \end{aligned}$$

b) The exact value of  $f'(2)$  can be calculated by using our knowledge of differential calculus.

$$\begin{aligned} f(x) &= 7e^{0.5x} \\ f'(x) &= 7 \times 0.5 \times e^{0.5x} \\ &= 3.5e^{0.5x} \end{aligned}$$

So the true value of  $f'(2)$  is

$$\begin{aligned} f'(2) &= 3.5e^{0.5(2)} \\ &= 9.5140 \end{aligned}$$

c) True error is calculated as

$$\begin{aligned} E_t &= \text{True value} - \text{Approximate value} \\ &= 9.5140 - 10.265 \\ &= -0.75061 \end{aligned}$$

The magnitude of true error does not show how bad the error is. A true error of  $E_t = -0.722$  may seem to be small, but if the function given in the Example 1 were  $f(x) = 7 \times 10^{-6} e^{0.5x}$ , the true error in calculating  $f'(2)$  with  $h = 0.3$ , would be  $E_t = -0.75061 \times 10^{-6}$ . This value of true error is smaller, even when the two problems are similar in that they use the same value of the function argument,  $x = 2$  and the step size,  $h = 0.3$ . This brings us to the definition of relative true error.

**Q:** What is relative true error?

**A:** Relative true error is denoted by  $\epsilon_t$  and is defined as the ratio between the true error and the true value.

$$\text{Relative True Error} = \frac{\text{True Error}}{\text{True Value}}$$

## Example 2

The derivative of a function  $f(x)$  at a particular value of  $x$  can be approximately calculated by

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

For  $f(x) = 7e^{0.5x}$  and  $h = 0.3$ , find the relative true error at  $x = 2$ .

### Solution

From Example 1,

$$\begin{aligned} E_t &= \text{True value} - \text{Approximate value} \\ &= 9.5140 - 10.265 \\ &= -0.75061 \end{aligned}$$

Relative true error is calculated as

$$\begin{aligned} \epsilon_t &= \frac{\text{True Error}}{\text{True Value}} \\ &= \frac{-0.75061}{9.5140} \\ &= -0.078895 \end{aligned}$$

Relative true errors are also presented as percentages. For this example,

$$\begin{aligned} \epsilon_t &= -0.078895 \times 100\% \\ &= -7.58895\% \end{aligned}$$

Absolute relative true errors may also need to be calculated. In such cases,

$$\begin{aligned} |\epsilon_t| &= |-0.078895| \\ &= 0.078895 \\ &= 7.58895\% \end{aligned}$$

**Q:** What is approximate error?

**A:** In the previous section, we discussed how to calculate true errors. Such errors are calculated only if true values are known. An example where this would be useful is when one is checking if a program is in working order and you know some examples where the true error is known. But mostly we will not have the luxury of knowing true values as why would you want to find the approximate values if you know the true values. So when we are solving a problem numerically, we will only have access to approximate values. We need to know how to quantify error for such cases.

Approximate error is denoted by  $E_a$  and is defined as the difference between the present approximation and previous approximation.

$$\text{Approximate Error} = \text{Present Approximation} - \text{Previous Approximation}$$

### Example 3

The derivative of a function  $f(x)$  at a particular value of  $x$  can be approximately calculated by

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

For  $f(x) = 7e^{0.5x}$  and at  $x = 2$ , find the following

- a)  $f'(2)$  using  $h = 0.3$
- b)  $f'(2)$  using  $h = 0.15$
- c) approximate error for the value of  $f'(2)$  for part (b)

### Solution

a) The approximate expression for the derivative of a function is

$$f'(x) \approx \frac{f(x+h) - f(x)}{h}.$$

For  $x = 2$  and  $h = 0.3$ ,

$$\begin{aligned} f'(2) &\approx \frac{f(2+0.3) - f(2)}{0.3} \\ &= \frac{f(2.3) - f(2)}{0.3} \\ &= \frac{7e^{0.5(2.3)} - 7e^{0.5(2)}}{0.3} \\ &= \frac{22.107 - 19.028}{0.3} \\ &= 10.265 \end{aligned}$$

b) Repeat the procedure of part (a) with  $h = 0.15$ ,

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

For  $x = 2$  and  $h = 0.15$ ,

$$\begin{aligned} f'(2) &\approx \frac{f(2+0.15) - f(2)}{0.15} \\ &= \frac{f(2.15) - f(2)}{0.15} \\ &= \frac{7e^{0.5(2.15)} - 7e^{0.5(2)}}{0.15} \\ &= \frac{20.50 - 19.028}{0.15} \\ &= 9.8799 \end{aligned}$$

c) So the approximate error,  $E_a$  is

$$\begin{aligned} E_a &= \text{Present Approximation} - \text{Previous Approximation} \\ &= 9.8799 - 10.265 \\ &= -0.38474 \end{aligned}$$

The magnitude of approximate error does not show how bad the error is. An approximate error of  $E_a = -0.38300$  may seem to be small; but for  $f(x) = 7 \times 10^{-6} e^{0.5x}$ , the approximate error in calculating  $f'(2)$  with  $h = 0.15$  would be  $E_a = -0.38474 \times 10^{-6}$ . This value of approximate error is smaller, even when the two problems are similar in that they use the same value of the function argument,  $x = 2$ , and  $h = 0.15$  and  $h = 0.3$ . This brings us to the definition of relative approximate error.

**Q:** What is relative approximate error?

**A:** Relative approximate error is denoted by  $\epsilon_a$  and is defined as the ratio between the approximate error and the present approximation.

$$\text{Relative Approximate Error} = \frac{\text{Approximate Error}}{\text{Present Approximation}}$$

**Example 4**

The derivative of a function  $f(x)$  at a particular value of  $x$  can be approximately calculated by

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

For  $f(x) = 7e^{0.5x}$ , find the relative approximate error in calculating  $f'(2)$  using values from  $h = 0.3$  and  $h = 0.15$ .

**Solution**

From Example 3, the approximate value of  $f'(2) = 10.263$  using  $h = 0.3$  and  $f'(2) = 9.8800$  using  $h = 0.15$ .

$$\begin{aligned} E_a &= \text{Present Approximation} - \text{Previous Approximation} \\ &= 9.8799 - 10.265 \\ &= -0.38474 \end{aligned}$$

The relative approximate error is calculated as

$$\begin{aligned} \epsilon_a &= \frac{\text{Approximate Error}}{\text{Present Approximation}} \\ &= \frac{-0.38474}{9.8799} \\ &= -0.038942 \end{aligned}$$

Relative approximate errors are also presented as percentages. For this example,

$$\begin{aligned} \epsilon_a &= -0.038942 \times 100\% \\ &= -3.8942\% \end{aligned}$$

Absolute relative approximate errors may also need to be calculated. In this example

$$\begin{aligned} |\epsilon_a| &= |-0.038942| \\ &= 0.038942 \text{ or } 3.8942\% \end{aligned}$$

**Q:** While solving a mathematical model using numerical methods, how can we use relative approximate errors to minimize the error?

**A:** In a numerical method that uses iterative methods, a user can calculate relative approximate error  $\epsilon_a$  at the end of each iteration. The user may pre-specify a minimum acceptable tolerance called the pre-specified tolerance,  $\epsilon_s$ . If the absolute relative approximate error  $\epsilon_a$  is less than or equal to the pre-specified tolerance  $\epsilon_s$ , that is,  $|\epsilon_a| \leq \epsilon_s$ , then the acceptable error has been reached and no more iterations would be required.

Alternatively, one may pre-specify how many significant digits they would like to be correct in their answer. In that case, if one wants at least  $m$  significant digits to be correct in the answer, then you would need to have the absolute relative approximate error,  $|\epsilon_a| \leq 0.5 \times 10^{2-m}\%$ .

**Example 5**

If one chooses 6 terms of the Maclaurin series for  $e^x$  to calculate  $e^{0.7}$ , how many significant digits can you trust in the solution? Find your answer without knowing or using the exact answer.

**Solution**

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

Using 6 terms, we get the current approximation as

$$\begin{aligned} e^{0.7} &\approx 1 + 0.7 + \frac{0.7^2}{2!} + \frac{0.7^3}{3!} + \frac{0.7^4}{4!} + \frac{0.7^5}{5!} \\ &= 2.0136 \end{aligned}$$

Using 5 terms, we get the previous approximation as

$$\begin{aligned} e^{0.7} &\approx 1 + 0.7 + \frac{0.7^2}{2!} + \frac{0.7^3}{3!} + \frac{0.7^4}{4!} \\ &= 2.0122 \end{aligned}$$

The percentage absolute relative approximate error is

$$\begin{aligned} |\epsilon_a| &= \left| \frac{2.0136 - 2.0122}{2.0136} \right| \times 100 \\ &= 0.069527\% \end{aligned}$$

Since  $|\epsilon_a| \leq 0.5 \times 10^{-2}\%$ , at least 2 significant digits are correct in the answer of

$$e^{0.7} \approx 2.0136$$

**Q:** But what do you mean by significant digits?

**A:** Significant digits are important in showing the truth one has in a reported number. For example, if someone asked me what the population of my county is, I would respond, "The population of the Hillsborough county area is 1 million". But if someone was going to give me a \$100 for every citizen of the county, I would have to get an exact count. That count would have been 1,079,587 in year 2003. So you can see that in my statement that the population is 1 million, that there is only one significant digit, that is, 1, and in the statement that the population is 1,079,587, there are seven significant digits. So, how do we differentiate the number of digits correct in 1,000,000 and 1,079,587? Well for that, one may use scientific notation. For our data we show

$$1,000,000 = 1 \times 10^6$$

$$1,079,587 = 1.079587 \times 10^6$$

to signify the correct number of significant digits.

**Example 5**

Give some examples of showing the number of significant digits.

**Solution**

- a) 0.0459 has three significant digits
- b) 4.590 has four significant digits
- c) 4008 has four significant digits
- d) 4008.0 has five significant digits

- e)  $1.079 \times 10^3$  has four significant digits
- f)  $1.0790 \times 10^3$  has five significant digits
- g)  $1.07900 \times 10^3$  has six significant digits

---

#### INTRODUCTION, APPROXIMATION AND ERRORS

---

Topic Measuring Errors  
Summary Textbook notes on measuring errors  
Major General Engineering  
Authors Autar Kaw  
Date December 23, 2009  
Web Site <http://numericalmethods.eng.usf.edu>

---

## Chapter 01.03

### Sources of Error

*After reading this chapter, you should be able to:*

1. know that there are two inherent sources of error in numerical methods – round-off and truncation error,
2. recognize the sources of round-off and truncation error, and
3. know the difference between round-off and truncation error.

Error in solving an engineering or science problem can arise due to several factors. First, the error may be in the modeling technique. A mathematical model may be based on using assumptions that are not acceptable. For example, one may assume that the drag force on a car is proportional to the velocity of the car, but actually it is proportional to the square of the velocity of the car. This itself can create huge errors in determining the performance of the car, no matter how accurate the numerical methods you may use are. Second, errors may arise from mistakes in programs themselves or in the measurement of physical quantities. But, in applications of numerical methods itself, the two errors we need to focus on are

1. Round off error
2. Truncation error.

**Q:** What is round off error?

**A:** A computer can only represent a number approximately. For example, a number like  $\frac{1}{3}$  may be represented as 0.333333 on a PC. Then the round off error in this case is  $\frac{1}{3} - 0.333333 = 0.0000003\bar{3}$ . Then there are other numbers that cannot be represented exactly. For example,  $\pi$  and  $\sqrt{2}$  are numbers that need to be approximated in computer calculations.

**Q:** What problems can be created by round off errors?

**A:** Twenty-eight Americans were killed on February 25, 1991. An Iraqi Scud hit the Army barracks in Dhahran, Saudi Arabia. The patriot defense system had failed to track and intercept the Scud. What was the cause for this failure?

The Patriot defense system consists of an electronic detection device called the range gate. It calculates the area in the air space where it should look for a Scud. To find out where it

should aim next, it calculates the velocity of the Scud and the last time the radar detected the Scud. Time is saved in a register that has 24 bits length. Since the internal clock of the system is measured for every one-tenth of a second, 1/10 is expressed in a 24 bit-register as 0.00011001100110011001100. However, this is not an exact representation. In fact, it would need infinite numbers of bits to represent 1/10 exactly. So, the error in the representation in decimal format is



**Figure 1** Patriot missile (Courtesy of the US Armed Forces,  
<http://www.redstone.army.mil/history/archives/patriot/patriot.html>)

$$\begin{aligned}\frac{1}{10} - (0 \times 2^{-1} + 0 \times 2^{-2} + 0 \times 2^{-3} + 1 \times 2^{-4} + \dots + 1 \times 2^{-22} + 0 \times 2^{-23} + 0 \times 2^{-24}) \\ = 9.537 \times 10^{-8}\end{aligned}$$

The battery was on for 100 consecutive hours, hence causing an inaccuracy of

$$\begin{aligned}= 9.537 \times 10^{-8} \frac{\text{s}}{0.1\text{s}} \times 100 \text{ hr} \times \frac{3600\text{s}}{1\text{hr}} \\ = 0.3433\text{s}\end{aligned}$$

The shift calculated in the range gate due to 0.3433s was calculated as 687m. For the Patriot missile defense system, the target is considered out of range if the shift was going to more than 137m.

**Q:** What is truncation error?

**A:** Truncation error is defined as the error caused by truncating a mathematical procedure.

For example, the Maclaurin series for  $e^x$  is given as

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

This series has an infinite number of terms but when using this series to calculate  $e^x$ , only a finite number of terms can be used. For example, if one uses three terms to calculate  $e^x$ , then

$$e^x \approx 1 + x + \frac{x^2}{2!}.$$

the truncation error for such an approximation is

$$\begin{aligned}\text{Truncation error} &= e^x - \left(1 + x + \frac{x^2}{2!}\right), \\ &= \frac{x^3}{3!} + \frac{x^4}{4!} + \dots\end{aligned}$$

But, how can truncation error be controlled in this example? We can use the concept of relative approximate error to see how many terms need to be considered. Assume that one is calculating  $e^{1.2}$  using the Maclaurin series, then

$$e^{1.2} = 1 + 1.2 + \frac{1.2^2}{2!} + \frac{1.2^3}{3!} + \dots$$

Let us assume one wants the absolute relative approximate error to be less than 1%. In Table 1, we show the value of  $e^{1.2}$ , approximate error and absolute relative approximate error as a function of the number of terms,  $n$ .

$n$	$e^{1.2}$	$E_a$	$ E_a  \%$
1	1	-	-
2	2.2	1.2	54.546
3	2.92	0.72	24.658
4	3.208	0.288	8.9776
5	3.2944	0.0864	2.6226
6	3.3151	0.020736	0.62550

Using 6 terms of the series yields a  $|E_a| < 1\%$ .

**Q:** Can you give me other examples of truncation error?

**A:** In many textbooks, the Maclaurin series is used as an example to illustrate truncation error. This may lead you to believe that truncation errors are just chopping a part of the series. However, truncation error can take place in other mathematical procedures as well. For example to find the derivative of a function, we define

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

But since we cannot use  $\Delta x \rightarrow 0$ , we have to use a finite value of  $\Delta x$ , to give

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

So the truncation error is caused by choosing a finite value of  $\Delta x$  as opposed to a  $\Delta x \rightarrow 0$ .

For example, in finding  $f'(3)$  for  $f(x) = x^2$ , we have the exact value calculated as follows.

$$f(x) = x^2$$

From the definition of the derivative of a function,

$$\begin{aligned}f'(x) &= \lim_{\Delta x \rightarrow 0} \frac{f(x + \Delta x) - f(x)}{\Delta x} \\ &= \lim_{\Delta x \rightarrow 0} \frac{(x + \Delta x)^2 - (x)^2}{\Delta x}\end{aligned}$$

$$\begin{aligned}
 &= \lim_{\Delta x \rightarrow 0} \frac{x^2 + 2x\Delta x + (\Delta x)^2 - x^2}{\Delta x} \\
 &= \lim_{\Delta x \rightarrow 0} (2x + \Delta x) \\
 &= 2x
 \end{aligned}$$

This is the same expression you would have obtained by directly using the formula from your differential calculus class

$$\frac{d}{dx}(x^n) = nx^{n-1}$$

By this formula for

$$\begin{aligned}
 f(x) &= x^2 \\
 f'(x) &= 2x
 \end{aligned}$$

The exact value of  $f'(3)$  is

$$\begin{aligned}
 f'(3) &= 2 \times 3 \\
 &= 6
 \end{aligned}$$

If we now choose  $\Delta x = 0.2$ , we get

$$\begin{aligned}
 f'(3) &= \frac{f(3 + 0.2) - f(3)}{0.2} \\
 &= \frac{f(3.2) - f(3)}{0.2} \\
 &= \frac{3.2^2 - 3^2}{0.2} \\
 &= \frac{10.24 - 9}{0.2} \\
 &= \frac{1.24}{0.2} \\
 &= 6.2
 \end{aligned}$$

We purposefully chose a simple function  $f(x) = x^2$  with value of  $x = 2$  and  $\Delta x = 0.2$  because we wanted to have no round-off error in our calculations so that the truncation error can be isolated. The truncation error in this example is

$$6 - 6.2 = -0.2.$$

Can you reduce the truncate error by choosing a smaller  $\Delta x$ ?

Another example of truncation error is the numerical integration of a function,

$$I = \int_a^b f(x) dx$$

Exact calculations require us to calculate the area under the curve by adding the area of the rectangles as shown in Figure 2. However, exact calculations requires an infinite number of such rectangles. Since we cannot choose an infinite number of rectangles, we will have truncation error.

For example, to find

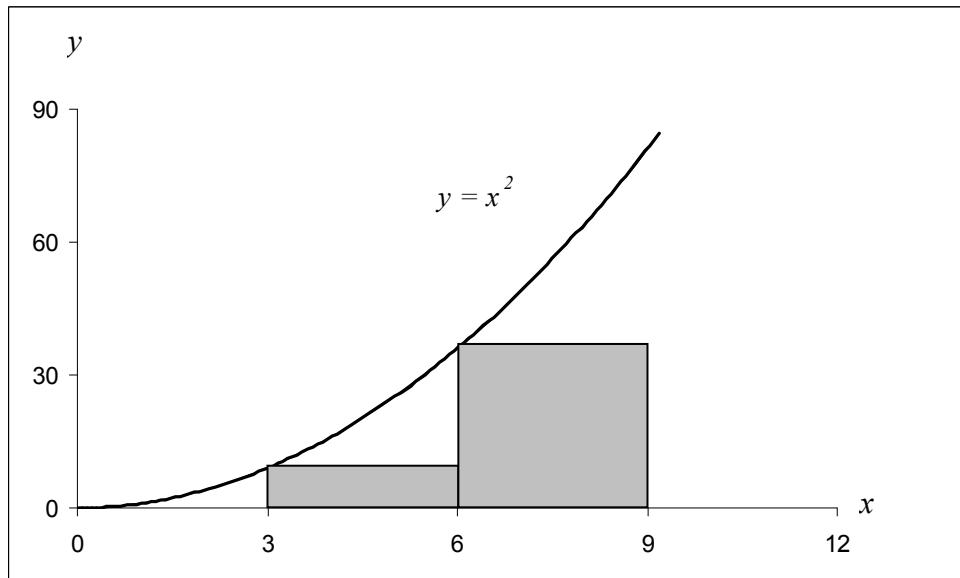
$$\int_3^9 x^2 dx,$$

we have the exact value as

$$\begin{aligned}\int_3^9 x^2 dx &= \left[ \frac{x^3}{3} \right]_3^9 \\ &= \left[ \frac{9^3 - 3^3}{3} \right] \\ &= 234\end{aligned}$$

If we now choose to use two rectangles of equal width to approximate the area (see Figure 2) under the curve, the approximate value of the integral

$$\begin{aligned}\int_3^9 x^2 dx &= (x^2)|_{x=3} (6-3) + (x^2)|_{x=6} (9-6) \\ &= (3^2)3 + (6^2)3 \\ &= 27 + 108 \\ &= 135\end{aligned}$$

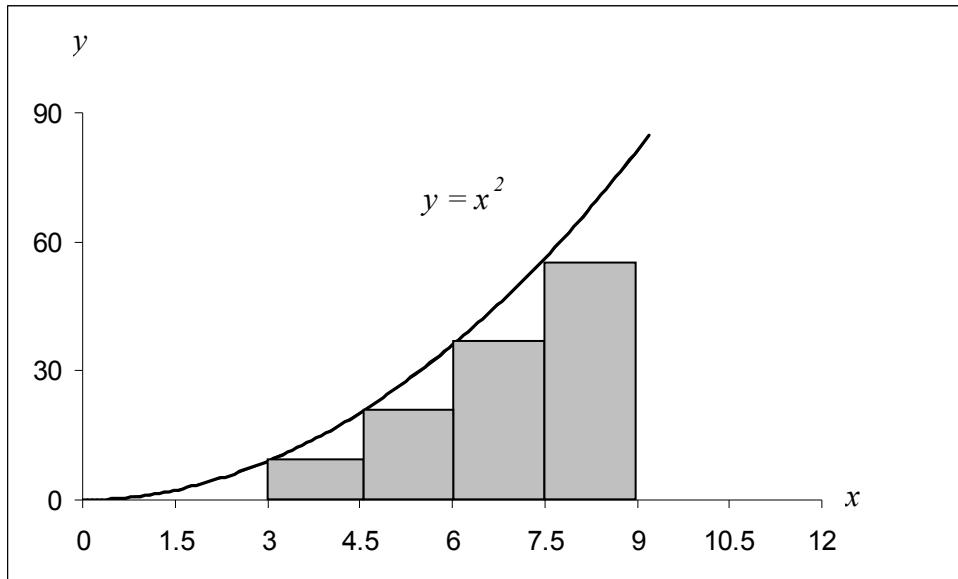


**Figure 2** Plot of  $y = x^2$  showing the approximate area under the curve from  $x = 3$  to  $x = 9$  using two rectangles.

Again, we purposefully chose a simple example because we wanted to have no round off error in our calculations. This makes the obtained error purely truncation. The truncation error is

$$234 - 135 = 99$$

Can you reduce the truncation error by choosing more rectangles as given in Figure 3? What is the truncation error?



**Figure 3** Plot of  $y = x^2$  showing the approximate area under the curve from  $x = 3$  to  $x = 9$  using four rectangles.

## References

“Patriot Missile Defense – Software Problem Led to System Failure at Dhahran, Saudi Arabia”, GAO Report, General Accounting Office, Washington DC, February 4, 1992.

---

### INTRODUCTION, APPROXIMATION AND ERRORS

---

Topic	Sources of error
Summary	Textbook notes on sources of error
Major	General Engineering
Authors	Autar Kaw
Date	December 23, 2009
Web Site	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

---

## Chapter 01.04

# Binary Representation of Numbers

*After reading this chapter, you should be able to:*

1. convert a base-10 real number to its binary representation,
2. convert a binary number to an equivalent base-10 number.

In everyday life, we use a number system with a base of 10. For example, look at the number 257.56. Each digit in 257.56 has a value of 0 through 9 and has a place value. It can be written as

$$257.76 = 2 \times 10^2 + 5 \times 10^1 + 7 \times 10^0 + 7 \times 10^{-1} + 6 \times 10^{-2}$$

In a binary system, we have a similar system where the base is made of only two digits 0 and 1. So it is a base 2 system. A number like (1011.0011) in base-2 represents the decimal number as

$$\begin{aligned}(1011.0011)_2 &= ((1 \times 2^3 + 0 \times 2^2 + 1 \times 2^1 + 1 \times 2^0) + (0 \times 2^{-1} + 0 \times 2^{-2} + 1 \times 2^{-3} + 1 \times 2^{-4}))_{10} \\ &= 11.1875\end{aligned}$$

in the decimal system.

To understand the binary system, we need to be able to convert binary numbers to decimal numbers and vice-versa.

We have already seen an example of how binary numbers are converted to decimal numbers. Let us see how we can convert a decimal number to a binary number. For example take the decimal number 11.1875. First, look at the integer part: 11.

1. Divide 11 by 2. This gives a quotient of 5 and a remainder of 1. Since the remainder is 1,  $a_0 = 1$ .
2. Divide the quotient 5 by 2. This gives a quotient of 2 and a remainder of 1. Since the remainder is 1,  $a_1 = 1$ .
3. Divide the quotient 2 by 2. This gives a quotient of 1 and a remainder of 0. Since the remainder is 0,  $a_2 = 0$ .
4. Divide the quotient 1 by 2. This gives a quotient of 0 and a remainder of 1. Since the remainder is ,  $a_3 = 1$ .

Since the quotient now is 0, the process is stopped. The above steps are summarized in Table 1.

**Table 1** Converting a base-10 integer to binary representation.

	Quotient	Remainder
11/2	5	$1 = a_0$
5/2	2	$1 = a_1$
2/2	1	$0 = a_2$
1/2	0	$1 = a_3$

Hence

$$\begin{aligned}(11)_{10} &= (a_3 a_2 a_1 a_0)_2 \\ &= (1011)_2\end{aligned}$$

For any integer, the algorithm for finding the binary equivalent is given in the flow chart on the next page.

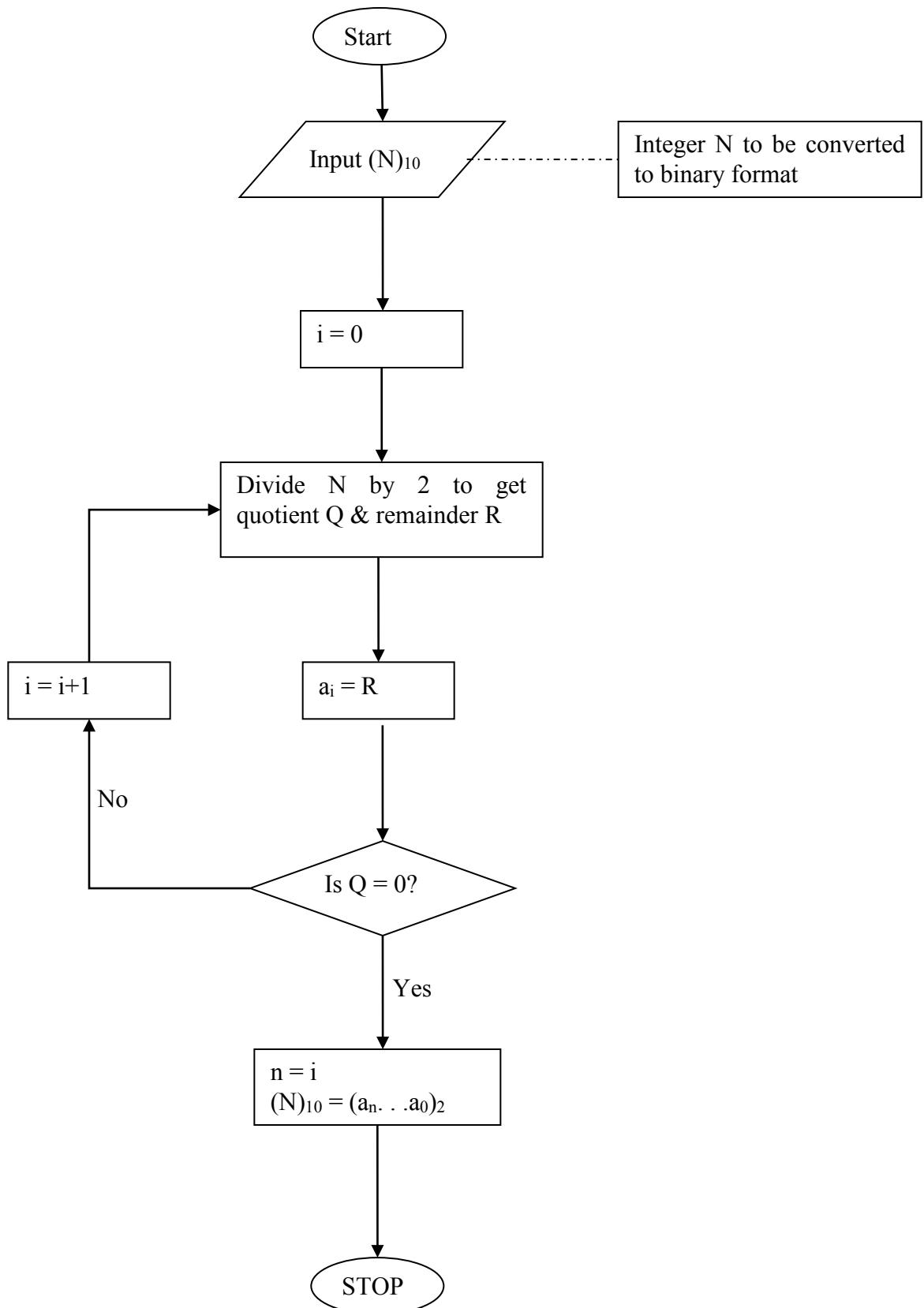
Now let us look at the decimal part, that is, 0.1875.

1. Multiply 0.1875 by 2. This gives 0.375. The number before the decimal is 0 and the number after the decimal is 0.375. Since the number before the decimal is 0,  $a_{-1} = 0$ .
2. Multiply the number after the decimal, that is, 0.375 by 2. This gives 0.75. The number before the decimal is 0 and the number after the decimal is 0.75. Since the number before the decimal is 0,  $a_{-2} = 0$ .
3. Multiply the number after the decimal, that is, 0.75 by 2. This gives 1.5. The number before the decimal is 1 and the number after the decimal is 0.5. Since the number before the decimal is 1,  $a_{-3} = 1$ .
4. Multiply the number after the decimal, that is, 0.5 by 2. This gives 1.0. The number before the decimal is 1 and the number after the decimal is 0. Since the number before the decimal is 1,  $a_{-4} = 1$ .

Since the number after the decimal is 0, the conversion is complete. The above steps are summarized in Table 2.

**Table 2.** Converting a base-10 fraction to binary representation.

	Number	Number after decimal	Number before decimal
$0.1875 \times 2$	0.375	0.375	$0 = a_{-1}$
$0.375 \times 2$	0.75	0.75	$0 = a_{-2}$
$0.75 \times 2$	1.5	0.5	$1 = a_{-3}$
$0.5 \times 2$	1.0	0.0	$1 = a_{-4}$



Hence

$$\begin{aligned}(0.1875)_{10} &= (a_{-1}a_{-2}a_{-3}a_{-4})_2 \\ &= (0.0011)_2\end{aligned}$$

The algorithm for any fraction is given in a flowchart on the next page.

Having calculated

$$(11)_{10} = (1011)_2$$

and

$$(0.1875)_{10} = (0.0011)_2,$$

we have

$$(11.1875)_{10} = (1011.0011)_2.$$

In the above example, when we were converting the fractional part of the number, we were left with 0 after the decimal number and used that as a place to stop. In many cases, we are never left with a 0 after the decimal number. For example, finding the binary equivalent of 0.3 is summarized in Table 3.

**Table 3.** Converting a base-10 fraction to approximate binary representation.

	Number	Number after decimal	Number before decimal
$0.3 \times 2$	0.6	0.6	$0 = a_{-1}$
$0.6 \times 2$	1.2	0.2	$1 = a_{-2}$
$0.2 \times 2$	0.4	0.4	$0 = a_{-3}$
$0.4 \times 2$	0.8	0.8	$0 = a_{-4}$
$0.8 \times 2$	1.6	0.6	$1 = a_{-5}$

As you can see the process will never end. In this case, the number can only be approximated in binary format, that is,

$$(0.3)_{10} \approx (a_{-1}a_{-2}a_{-3}a_{-4}a_{-5})_2 = (0.01001)_2$$

**Q:** But what is the mathematics behind this process of converting a decimal number to binary format?

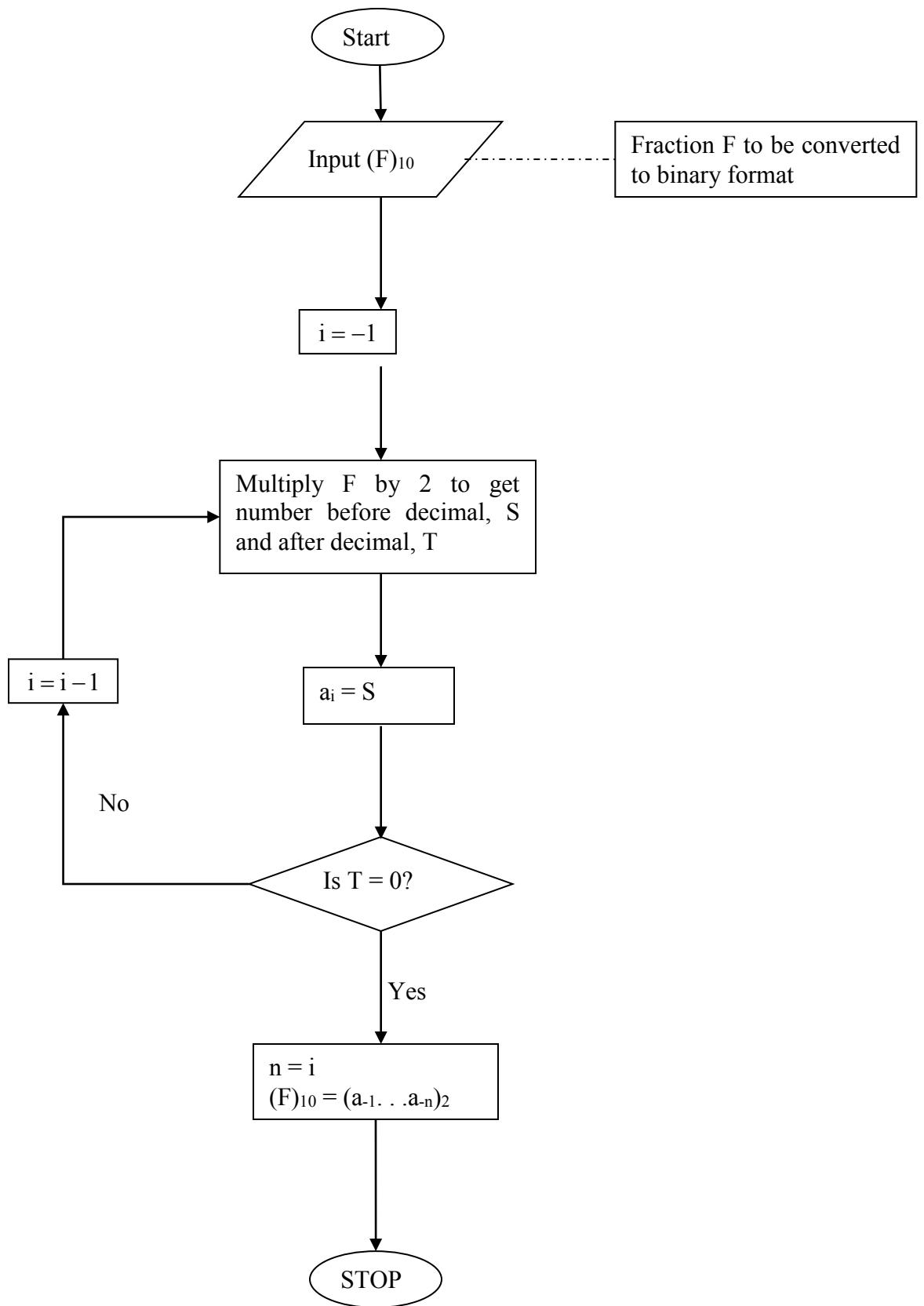
**A:** Let  $z$  be the decimal number written as

$$z = x.y$$

where

$x$  is the integer part and  $y$  is the fractional part.

We want to find the binary equivalent of  $x$ . So we can write



$$x = a_n 2^n + a_{n-1} 2^{n-1} + \dots + a_0 2^0$$

If we can now find  $a_0, \dots, a_n$  in the above equation then

$$(x)_{10} = (a_n a_{n-1} \dots a_0)_2$$

We now want to find the binary equivalent of  $y$ . So we can write

$$y = b_{-1} 2^{-1} + b_{-2} 2^{-2} + \dots + b_{-m} 2^{-m}$$

If we can now find  $b_{-1}, \dots, b_{-m}$  in the above equation then

$$(y)_{10} = (b_{-1} b_{-2} \dots b_{-m})_2$$

Let us look at this using the same example as before.

### Example 1

Convert  $(11.1875)_{10}$  to base 2.

#### Solution

To convert  $(11)_{10}$  to base 2, what is the highest power of 2 that is part of 11. That power is 3, as  $2^3 = 8$  to give

$$11 = 2^3 + 3$$

What is the highest power of 2 that is part of 3. That power is 1, as  $2^1 = 2$  to give

$$3 = 2^1 + 1$$

So

$$11 = 2^3 + 3 = 2^3 + 2^1 + 1$$

What is the highest power of 2 that is part of 1. That power is 0, as  $2^0 = 1$  to give

$$1 = 2^0$$

Hence

$$(11)_{10} = 2^3 + 2^1 + 1 = 2^3 + 2^1 + 2^0 = 1 \times 2^3 + 0 \times 2^2 + 1 \times 2^1 + 1 \times 2^0 = (1011)_2$$

To convert  $(0.1875)_{10}$  to the base 2, we proceed as follows. What is the smallest negative power of 2 that is less than or equal to 0.1875. That power is  $-3$  as  $2^{-3} = 0.125$ .

So

$$0.1875 = 2^{-3} + 0.0625$$

What is the next smallest negative power of 2 that is less than or equal to 0.0625. That power is  $-4$  as  $2^{-4} = 0.0625$ .

So

$$0.1875 = 2^{-3} + 2^{-4}$$

Hence

$$(0.1875)_{10} = 2^{-3} + 0.0625 = 2^{-3} + 2^{-4} = 0 \times 2^{-1} + 0 \times 2^{-2} + 1 \times 2^{-3} + 1 \times 2^{-4} = (0.0011)_2$$

Since

$$(11)_{10} = (1011)_2$$

and

$$(0.1875)_{10} = (0.0011)_2$$

we get

$$(11.1875)_{10} = (1011.0011)_2$$

Can you show this algebraically for any general number?

### Example 2

Convert  $(13.875)_{10}$  to base 2.

#### Solution

For  $(13)_{10}$ , conversion to binary format is shown in Table 4.

**Table 4.** Conversion of base-10 integer to binary format.

	Quotient	Remainder
$13/2$	6	$1 = a_0$
$6/2$	3	$0 = a_1$
$3/2$	1	$1 = a_2$
$1/2$	0	$1 = a_3$

So

$$(13)_{10} = (1101)_2.$$

Conversion of  $(0.875)_{10}$  to binary format is shown in Table 5.

**Table 5.** Converting a base-10 fraction to binary representation.

	Number	Number after decimal	Number before decimal
$0.875 \times 2$	1.75	0.75	$1 = a_{-1}$
$0.75 \times 2$	1.5	0.5	$1 = a_{-2}$
$0.5 \times 2$	1.0	0.0	$1 = a_{-3}$

So

$$(0.875)_{10} = (0.111)_2$$

Hence

$$(13.875)_{10} = (1101.111)_2$$

## INTRODUCTION TO NUMERICAL METHODS

Topic      Binary representation of number

Summary    Textbook notes on binary representation of numbers

Major      General Engineering

Authors     Autar Kaw

Date        September 4, 2014

Web Site    <http://numericalmethods.eng.usf.edu>

# Chapter 01.05

## Floating Point Representation

*After reading this chapter, you should be able to:*

1. convert a base-10 number to a binary floating point representation,
2. convert a binary floating point number to its equivalent base-10 number,
3. understand the IEEE-754 specifications of a floating point representation in a typical computer,
4. calculate the machine epsilon of a representation.

Consider an old time cash register that would ring any purchase between 0 and 999.99 units of money. Note that there are five (not six) working spaces in the cash register (the decimal number is shown just for clarification).

**Q:** How will the smallest number 0 be represented?

**A:** The number 0 will be represented as

0	0	0	.	0	0
---	---	---	---	---	---

**Q:** How will the largest number 999.99 be represented?

**A:** The number 999.99 will be represented as

9	9	9	.	9	9
---	---	---	---	---	---

**Q:** Now look at any typical number between 0 and 999.99, such as 256.78. How would it be represented?

**A:** The number 256.78 will be represented as

2	5	6	.	7	8
---	---	---	---	---	---

**Q:** What is the smallest change between consecutive numbers?

**A:** It is 0.01, like between the numbers 256.78 and 256.79.

**Q:** What amount would one pay for an item, if it costs 256.789?

**A:** The amount one would pay would be rounded off to 256.79 or chopped to 256.78. In either case, the maximum error in the payment would be less than 0.01.

**Q:** What magnitude of relative errors would occur in a transaction?

**A:** Relative error for representing small numbers is going to be high, while for large numbers the relative error is going to be small.

For example, for 256.786, rounding it off to 256.79 accounts for a round-off error of  $256.786 - 256.79 = -0.004$ . The relative error in this case is

$$\begin{aligned}\varepsilon_t &= \frac{-0.004}{256.786} \times 100 \\ &= -0.001558\%.\end{aligned}$$

For another number, 3.546, rounding it off to 3.55 accounts for the same round-off error of  $3.546 - 3.55 = -0.004$ . The relative error in this case is

$$\begin{aligned}\varepsilon_t &= \frac{-0.004}{3.546} \times 100 \\ &= -0.11280\%.\end{aligned}$$

**Q:** If I am interested in keeping relative errors of similar magnitude for the range of numbers, what alternatives do I have?

**A:** To keep the relative error of similar order for all numbers, one may use a floating-point representation of the number. For example, in floating-point representation, a number

- 256.78 is written as  $+2.5678 \times 10^2$ ,
- 0.003678 is written as  $+3.678 \times 10^{-3}$ , and
- 256.789 is written as  $-2.56789 \times 10^2$ .

The general representation of a number in base-10 format is given as

$$\text{sign} \times \text{mantissa} \times 10^{\text{exponent}}$$

or for a number  $y$ ,

$$y = \sigma \times m \times 10^e$$

Where

$\sigma$  = sign of the number, +1 or -1

$m$  = mantissa,  $1 \leq m < 10$

$e$  = integer exponent (also called ficand)

Let us go back to the example where we have five spaces available for a number. Let us also limit ourselves to positive numbers with positive exponents for this example. If we use the same five spaces, then let us use four for the mantissa and the last one for the exponent. So the smallest number that can be represented is 1 but the largest number would be  $9.999 \times 10^9$ . By using the floating-point representation, what we lose in accuracy, we gain in the range of numbers that can be represented. For our example, the maximum number represented changed from 999.99 to  $9.999 \times 10^9$ .

What is the error in representing numbers in the scientific format? Take the previous example of 256.78. It would be represented as  $2.568 \times 10^2$  and in the five spaces as

2	5	6	8	2
---	---	---	---	---

Another example, the number 576329.78 would be represented as  $5.763 \times 10^5$  and in five spaces as

5	7	6	3	5
---	---	---	---	---

So, how much error is caused by such representation. In representing 256.78, the round off error created is  $256.78 - 256.8 = -0.02$ , and the relative error is

$$\varepsilon_t = \frac{-0.02}{256.78} \times 100 = -0.0077888\%,$$

In representing  $576329.78$ , the round off error created is  $576329.78 - 5.763 \times 10^5 = 29.78$ , and the relative error is

$$\varepsilon_t = \frac{29.78}{576329.78} \times 100 = 0.0051672\%.$$

What you are seeing now is that although the errors are large for large numbers, but the relative errors are of the same order for both large and small numbers.

**Q:** How does this floating-point format relate to binary format?

**A:** A number  $y$  would be written as

$$y = \sigma \times m \times 2^e$$

Where

$\sigma$  = sign of number (negative or positive – use 0 for positive and 1 for negative),

$m$  = mantissa,  $(1)_2 \leq m < (10)_2$ , that is,  $(1)_{10} \leq m < (2)_{10}$ , and

$e$  = integer exponent.

### Example 1

Represent  $(54.75)_{10}$  in floating point binary format. Assuming that the number is written to a hypothetical word that is 9 bits long where the first bit is used for the sign of the number, the second bit for the sign of the exponent, the next four bits for the mantissa, and the next three bits for the exponent,

### Solution

$$(54.75)_{10} = (110110.11)_2 = (1.1011011)_2 \times 2^{(5)_{10}}$$

The exponent 5 is equivalent in binary format as

$$(5)_{10} = (101)_2$$

Hence

$$(54.75)_{10} = (1.1011011)_2 \times 2^{(101)_2}$$

The sign of the number is positive, so the bit for the sign of the number will have zero in it.

$$\sigma = 0$$

The sign of the exponent is positive. So the bit for the sign of the exponent will have zero in it.

The mantissa

$$m = 1011$$

(There are only 4 places for the mantissa, and the leading 1 is not stored as it is always expected to be there), and

the exponent

$$e = 101.$$

we have the representation as

0	0	1	0	1	1	1	0	1
---	---	---	---	---	---	---	---	---

**Example 2**

What number does the below given floating point format

0	1	1	0	1	1	1	1	0
---	---	---	---	---	---	---	---	---

represent in base-10 format. Assume a hypothetical 9-bit word, where the first bit is used for the sign of the number, second bit for the sign of the exponent, next four bits for the mantissa and next three for the exponent.

**Solution**

Given

Bit Representation	Part of Floating point number
0	Sign of number
1	Sign of exponent
1011	Magnitude of mantissa
110	Magnitude of exponent

The first bit is 0, so the number is positive.

The second bit is 1, so the exponent is negative.

The next four bits, 1011, are the magnitude of the mantissa, so

$$m = (1.1011)_2 = (1 \times 2^0 + 1 \times 2^{-1} + 0 \times 2^{-2} + 1 \times 2^{-3} + 1 \times 2^{-4})_{10} = (1.6875)_{10}$$

The last three bits, 110, are the magnitude of the exponent, so

$$e = (110)_2 = (1 \times 2^2 + 1 \times 2^1 + 0 \times 2^0)_{10} = (6)_{10}$$

The number in binary format then is

$$(1.1011)_2 \times 2^{-(110)_2}$$

The number in base-10 format is

$$= 1.6875 \times 2^{-6}$$

$$= 0.026367$$

**Example 3**

A machine stores floating-point numbers in a hypothetical 10-bit binary word. It employs the first bit for the sign of the number, the second one for the sign of the exponent, the next four for the exponent, and the last four for the magnitude of the mantissa.

- a) Find how 0.02832 will be represented in the floating-point 10-bit word.
- b) What is the decimal equivalent of the 10-bit word representation of part (a)?

**Solution**

- a) For the number, we have the integer part as 0 and the fractional part as 0.02832

Let us first find the binary equivalent of the integer part

$$\text{Integer part } (0)_{10} = (0)_2$$

Now we find the binary equivalent of the fractional part

$$\text{Fractional part: } \underline{.02832 \times 2}$$

$$\underline{0.05664 \times 2}$$

$$\underline{0.11328 \times 2}$$

$$\underline{0.22656 \times 2}$$

$$\begin{array}{r}
 \underline{0.45312 \times 2} \\
 \underline{0.90624 \times 2} \\
 \underline{1.81248 \times 2} \\
 \underline{1.62496 \times 2} \\
 \underline{1.24992 \times 2} \\
 \underline{0.49984 \times 2} \\
 \underline{0.99968 \times 2} \\
 \underline{1.99936}
 \end{array}$$

Hence

$$\begin{aligned}
 (0.02832)_{10} &\equiv (0.00000111001)_2 \\
 &= (1.11001)_2 \times 2^{-6} \\
 &\equiv (1.1100)_2 \times 2^{-6}
 \end{aligned}$$

The binary equivalent of exponent is found as follows

	Quotient	Remainder
6/2	3	0 = $a_0$
3/2	1	1 = $a_1$
1/2	0	1 = $a_2$

So

$$(6)_{10} = (110)_2$$

So

$$\begin{aligned}
 (0.02832)_{10} &= (1.1100)_2 \times 2^{-(110)_2} \\
 &= (1.1100)_2 \times 2^{-(0110)_2}
 \end{aligned}$$

Part of Floating point number	Bit Representation
Sign of number is positive	0
Sign of exponent is negative	1
Magnitude of the exponent	0110
Magnitude of mantissa	1100

The ten-bit representation bit by bit is

0	1	0	1	1	0	1	1	0	0
---	---	---	---	---	---	---	---	---	---

b) Converting the above floating point representation from part (a) to base 10 by following Example 2 gives

$$\begin{aligned}
 (1.1100)_2 \times 2^{-(0110)_2} \\
 &= (1 \times 2^0 + 1 \times 2^{-1} + 1 \times 2^{-2} + 0 \times 2^{-3} + 0 \times 2^{-4}) \times 2^{-(0 \times 2^3 + 1 \times 2^2 + 1 \times 2^1 + 0 \times 2^0)} \\
 &= (1.75)_{10} \times 2^{-(6)_{10}} \\
 &= 0.02734375
 \end{aligned}$$

**Q:** How do you determine the accuracy of a floating-point representation of a number?

**A:** The machine epsilon,  $\epsilon_{mach}$  is a measure of the accuracy of a floating point representation and is found by calculating the difference between 1 and the next number that can be represented. For example, assume a 10-bit hypothetical computer where the first bit is used for the sign of the number, the second bit for the sign of the exponent, the next four bits for the exponent and the next four for the mantissa.

We represent 1 as

0	0	0	0	0	0	0	0	0	0
---	---	---	---	---	---	---	---	---	---

and the next higher number that can be represented is

0	0	0	0	0	0	0	0	0	1
---	---	---	---	---	---	---	---	---	---

The difference between the two numbers is

$$\begin{aligned} & (1.0001)_2 \times 2^{(0000)_2} - (1.0000)_2 \times 2^{(0000)_2} \\ &= (0.0001)_2 \\ &= (1 \times 2^{-4})_{10} \\ &= (0.0625)_{10}. \end{aligned}$$

The machine epsilon is

$$\epsilon_{mach} = 0.0625.$$

The machine epsilon,  $\epsilon_{mach}$  is also simply calculated as two to the negative power of the number of bits used for mantissa. As far as determining accuracy, machine epsilon,  $\epsilon_{mach}$  is an upper bound of the magnitude of relative error that is created by the approximate representation of a number (See Example 4).

#### Example 4

A machine stores floating-point numbers in a hypothetical 10-bit binary word. It employs the first bit for the sign of the number, the second one for the sign of the exponent, the next four for the exponent, and the last four for the magnitude of the mantissa. Confirm that the magnitude of the relative true error that results from approximate representation of 0.02832 in the 10-bit format (as found in previous example) is less than the machine epsilon.

#### Solution

From Example 2, the ten-bit representation of 0.02832 bit-by-bit is

0	1	0	1	1	0	1	1	0	0
---	---	---	---	---	---	---	---	---	---

Again from Example 2, converting the above floating point representation to base-10 gives

$$\begin{aligned} & (1.1100)_2 \times 2^{-(0110)_2} \\ &= (1.75)_{10} \times 2^{-(6)_{10}} \\ &= (0.02734375)_{10} \end{aligned}$$

The absolute relative true error between the number 0.02832 and its approximate representation 0.02734375 is

$$\begin{aligned} |\varepsilon_t| &= \left| \frac{0.02832 - 0.02734375}{0.02832} \right| \\ &= 0.034472 \end{aligned}$$

which is less than the machine epsilon for a computer that uses 4 bits for mantissa, that is,

$$\begin{aligned}\varepsilon_{mach} &= 2^{-4} \\ &= 0.0625\end{aligned}$$

**Q:** How are numbers actually represented in floating point in a real computer?

**A:** In an actual typical computer, a real number is stored as per the IEEE-754 (Institute of Electrical and Electronics Engineers) floating-point arithmetic format. To keep the discussion short and simple, let us point out the salient features of the single precision format.

- A single precision number uses 32 bits.
- A number  $y$  is represented as

$$y = \sigma \times (1.a_1a_2 \cdots a_{23}) \cdot 2^e$$

where

$\sigma$  = sign of the number (positive or negative)

$a_i$  = entries of the mantissa, can be only 0 or 1,  $i = 1, \dots, 23$

$e$  = the exponent

Note the 1 before the radix point.

- The first bit represents the sign of the number (0 for positive number and 1 for a negative number).
- The next eight bits represent the exponent. Note that there is no separate bit for the sign of the exponent. The sign of the exponent is taken care of by normalizing by adding 127 to the actual exponent. For example in the previous example, the exponent was 6. It would be stored as the binary equivalent of  $127 + 6 = 133$ . Why is 127 and not some other number added to the actual exponent? Because in eight bits the largest integer that can be represented is  $(1111111)_2 = 255$ , and halfway of 255 is 127. This allows negative and positive exponents to be represented equally. The normalized (also called biased) exponent has the range from 0 to 255, and hence the exponent  $e$  has the range of  $-127 \leq e \leq 128$ .
- If instead of using the biased exponent, let us suppose we still used eight bits for the exponent but used one bit for the sign of the exponent and seven bits for the exponent magnitude. In seven bits, the largest integer that can be represented is  $(1111111)_2 = 127$  in which case the exponent  $e$  range would have been smaller, that is,  $-127 \leq e \leq 127$ . By biasing the exponent, the unnecessary representation of a negative zero and positive zero exponent (which are the same) is also avoided.
- Actually, the biased exponent range used in the IEEE-754 format is not 0 to 255, but 1 to 254. Hence, exponent  $e$  has the range of  $-126 \leq e \leq 127$ . So what are  $e = -127$  and  $e = 128$  used for? If  $e = 128$  and all the mantissa entries are zeros, the number is  $\pm\infty$  (the sign of infinity is governed by the sign bit), if  $e = 128$  and the mantissa entries are not zero, the number being represented is Not a Number (NaN). Because of the leading 1 in the floating point representation, the number zero cannot be represented exactly. That is why the number zero (0) is represented by  $e = -127$  and all the mantissa entries being zero.
- The next twenty-three bits are used for the mantissa.
- The largest number by magnitude that is represented by this format is

$$(1 \times 2^0 + 1 \times 2^{-1} + 1 \times 2^{-2} + \cdots + 1 \times 2^{-22} + 1 \times 2^{-23}) \times 2^{127} = 3.40 \times 10^{38}$$

The smallest number by magnitude that is represented, other than zero, is

$$(1 \times 2^0 + 0 \times 2^{-1} + 0 \times 2^{-2} + \dots + 0 \times 2^{-22} + 0 \times 2^{-23}) \times 2^{-126} = 1.18 \times 10^{-38}$$

- Since 23 bits are used for the mantissa, the machine epsilon,

$$\begin{aligned}\epsilon_{mach} &= 2^{-23} \\ &= 1.19 \times 10^{-7}\end{aligned}$$

**Q:** How are numbers represented in floating point in double precision in a computer?

**A:** In double precision IEEE-754 format, a real number is stored in 64 bits.

- The first bit is used for the sign,
- the next 11 bits are used for the exponent, and
- the rest of the bits, that is 52, are used for mantissa.

Can you find in double precision the

- range of the biased exponent,
- smallest number that can be represented,
- largest number that can be represented, and
- machine epsilon?

#### INTRODUCTION TO NUMERICAL METHODS

Topic	Floating Point Representation
Summary	Textbook notes on floating point representation
Major	General Engineering
Authors	Autar Kaw
Date	December 23, 2009
Web Site	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

# **Chapter 01.06**

## **Propagation of Errors**

If a calculation is made with numbers that are not exact, then the calculation itself will have an error. How do the errors in each individual number propagate through the calculations. Let's look at the concept via some examples.

### **Example 1**

Find the bounds for the propagation error in adding two numbers. For example if one is calculating  $X + Y$  where

$$X = 1.5 \pm 0.05,$$

$$Y = 3.4 \pm 0.04$$

### **Solution**

By looking at the numbers, the maximum possible value of X and Y are

$$X = 1.55 \text{ and } Y = 3.44$$

Hence

$$X + Y = 1.55 + 3.44 = 4.99$$

is the maximum value of  $X + Y$ .

The minimum possible value of X and Y are

$$X = 1.45 \text{ and } Y = 3.36.$$

Hence

$$\begin{aligned} X + Y &= 1.45 + 3.36 \\ &= 4.81 \end{aligned}$$

is the minimum value of  $X + Y$ .

Hence

$$4.81 \leq X + Y \leq 4.99.$$

One can find similar intervals of the bound for the other arithmetic operations of  $X - Y$ ,  $X * Y$ , and  $X / Y$ . What if the evaluations we are making are function evaluations instead? How do we find the value of the propagation error in such cases.

If  $f$  is a function of several variables  $X_1, X_2, X_3, \dots, X_{n-1}, X_n$ , then the maximum possible value of the error in  $f$  is

$$\Delta f \approx \left| \frac{\partial f}{\partial X_1} \Delta X_1 \right| + \left| \frac{\partial f}{\partial X_2} \Delta X_2 \right| + \dots + \left| \frac{\partial f}{\partial X_{n-1}} \Delta X_{n-1} \right| + \left| \frac{\partial f}{\partial X_n} \Delta X_n \right|$$

**Example 2**

The strain in an axial member of a square cross-section is given by

$$\epsilon = \frac{F}{h^2 E}$$

where

$F$  = axial force in the member, N

$h$  = length or width of the cross-section, m

$E$  = Young's modulus, Pa

Given

$$F = 72 \pm 0.9 \text{ N}$$

$$h = 4 \pm 0.1 \text{ mm}$$

$$E = 70 \pm 1.5 \text{ GPa}$$

Find the maximum possible error in the measured strain.

Solution

$$\begin{aligned} \epsilon &= \frac{72}{(4 \times 10^{-3})^2 (70 \times 10^9)} \\ &= 64.286 \times 10^{-6} \\ &= 64.286 \mu \end{aligned}$$

$$\Delta \epsilon = \left| \frac{\partial \epsilon}{\partial F} \Delta F \right| + \left| \frac{\partial \epsilon}{\partial h} \Delta h \right| + \left| \frac{\partial \epsilon}{\partial E} \Delta E \right|$$

$$\frac{\partial \epsilon}{\partial F} = \frac{1}{h^2 E}$$

$$\frac{\partial \epsilon}{\partial h} = -\frac{2F}{h^3 E}$$

$$\frac{\partial \epsilon}{\partial E} = -\frac{F}{h^2 E^2}$$

$$\begin{aligned} \Delta \epsilon &= \left| \frac{1}{h^2 E} \Delta F \right| + \left| \frac{2F}{h^3 E} \Delta h \right| + \left| \frac{F}{h^2 E^2} \Delta E \right| \\ &= \left| \frac{1}{(4 \times 10^{-3})^2 (70 \times 10^9)} \times 0.9 \right| + \left| \frac{2 \times 72}{(4 \times 10^{-3})^3 (70 \times 10^9)} \times 0.0001 \right| \\ &\quad + \left| \frac{72}{(4 \times 10^{-3})^2 (70 \times 10^9)^2} \times 1.5 \times 10^9 \right| \end{aligned}$$

$$= 8.0357 \times 10^{-7} + 3.2143 \times 10^{-6} + 1.3776 \times 10^{-6}$$

$$= 5.3955 \times 10^{-6}$$

$$= 5.3955 \mu$$

Hence

$$\epsilon = (64.286 \mu \pm 5.3955 \mu)$$

implying that the axial strain,  $\epsilon$  is between  $58.8905 \mu$  and  $69.6815 \mu$

**Example 3**

Subtraction of numbers that are nearly equal can create unwanted inaccuracies. Using the formula for error propagation, show that this is true.

**Solution**

Let

$$z = x - y$$

Then

$$\begin{aligned} |\Delta z| &= \left| \frac{\partial z}{\partial x} \Delta x + \frac{\partial z}{\partial y} \Delta y \right| \\ &= |(1)\Delta x| + |(-1)\Delta y| \\ &= |\Delta x| + |\Delta y| \end{aligned}$$

So the absolute relative change is

$$\left| \frac{\Delta z}{z} \right| = \frac{|\Delta x| + |\Delta y|}{|x - y|}$$

As  $x$  and  $y$  become close to each other, the denominator becomes small and hence create large relative errors.

For example if

$$x = 2 \pm 0.001$$

$$y = 2.003 \pm 0.001$$

$$\begin{aligned} \left| \frac{\Delta z}{z} \right| &= \frac{|0.001| + |0.001|}{|2 - 2.003|} \\ &= 0.6667 \\ &= 66.67\% \end{aligned}$$

## INTRODUCTION TO NUMERICAL METHODS

<b>Topic</b>	Propagation of Errors
<b>Summary</b>	Textbook notes on how errors propagate in arithmetic and function evaluations
<b>Major</b>	All Majors of Engineering
<b>Authors</b>	Autar Kaw
<b>Last Revised</b>	June 3, 2014
<b>Web Site</b>	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

# Chapter 01.07

## Taylor Theorem Revisited

After reading this chapter, you should be able to

1. understand the basics of Taylor's theorem,
2. write transcendental and trigonometric functions as Taylor's polynomial,
3. use Taylor's theorem to find the values of a function at any point, given the values of the function and all its derivatives at a particular point,
4. calculate errors and error bounds of approximating a function by Taylor series, and
5. revisit the chapter whenever Taylor's theorem is used to derive or explain numerical methods for various mathematical procedures.

The use of Taylor series exists in so many aspects of numerical methods that it is imperative to devote a separate chapter to its review and applications. For example, you must have come across expressions such as

$$\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \dots \quad (1)$$

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots \quad (2)$$

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots \quad (3)$$

All the above expressions are actually a special case of Taylor series called the Maclaurin series. Why are these applications of Taylor's theorem important for numerical methods? Expressions such as given in Equations (1), (2) and (3) give you a way to find the approximate values of these functions by using the basic arithmetic operations of addition, subtraction, division, and multiplication.

### Example 1

Find the value of  $e^{0.25}$  using the first five terms of the Maclaurin series.

#### Solution

The first five terms of the Maclaurin series for  $e^x$  is

$$e^x \approx 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!}$$

$$\begin{aligned} e^{0.25} &\approx 1 + 0.25 + \frac{0.25^2}{2!} + \frac{0.25^3}{3!} + \frac{0.25^4}{4!} \\ &= 1.2840 \end{aligned}$$

The exact value of  $e^{0.25}$  up to 5 significant digits is also 1.2840.

But the above discussion and example do not answer our question of what a Taylor series is.

Here it is, for a function  $f(x)$

$$f(x+h) = f(x) + f'(x)h + \frac{f''(x)}{2!}h^2 + \frac{f'''(x)}{3!}h^3 + \dots \quad (4)$$

provided all derivatives of  $f(x)$  exist and are continuous between  $x$  and  $x+h$ .

### What does this mean in plain English?

As Archimedes would have said (*without the fine print*), “*Give me the value of the function at a single point, and the value of all (first, second, and so on) its derivatives, and I can give you the value of the function at any other point*”.

It is very important to note that the Taylor series is not asking for the expression of the function and its derivatives, just the value of the function and its derivatives at a single point.

*Now the fine print:* Yes, all the derivatives have to exist and be continuous between  $x$  (the point where you are) to the point,  $x+h$  where you are wanting to calculate the function at. However, if you want to calculate the function approximately by using the  $n^{\text{th}}$  order Taylor polynomial, then  $1^{\text{st}}, 2^{\text{nd}}, \dots, n^{\text{th}}$  derivatives need to exist and be continuous in the closed interval  $[x, x+h]$ , while the  $(n+1)^{\text{th}}$  derivative needs to exist and be continuous in the open interval  $(x, x+h)$ .

### Example 2

Take  $f(x) = \sin(x)$ , we all know the value of  $\sin\left(\frac{\pi}{2}\right) = 1$ . We also know the  $f'(x) = \cos(x)$  and  $\cos\left(\frac{\pi}{2}\right) = 0$ . Similarly  $f''(x) = -\sin(x)$  and  $\sin\left(\frac{\pi}{2}\right) = 1$ . In a way, we know the value of  $\sin(x)$  and all its derivatives at  $x = \frac{\pi}{2}$ . We do not need to use any calculators, just plain differential calculus and trigonometry would do. Can you use Taylor series and this information to find the value of  $\sin(2)$ ?

### Solution

$$\begin{aligned} x &= \frac{\pi}{2} \\ x+h &= 2 \\ h &= 2-x \\ &= 2 - \frac{\pi}{2} \\ &= 0.42920 \end{aligned}$$

So

$$f(x+h) = f(x) + f'(x)h + f''(x)\frac{h^2}{2!} + f'''(x)\frac{h^3}{3!} + f''''(x)\frac{h^4}{4!} + \dots$$

$$x = \frac{\pi}{2}$$

$$h = 0.42920$$

$$f(x) = \sin(x), \quad f\left(\frac{\pi}{2}\right) = \sin\left(\frac{\pi}{2}\right) = 1$$

$$f'(x) = \cos(x), \quad f'\left(\frac{\pi}{2}\right) = 0$$

$$f''(x) = -\sin(x), \quad f''\left(\frac{\pi}{2}\right) = -1$$

$$f'''(x) = -\cos(x), \quad f'''\left(\frac{\pi}{2}\right) = 0$$

$$f''''(x) = \sin(x), \quad f''''\left(\frac{\pi}{2}\right) = 1$$

Hence

$$\begin{aligned} f\left(\frac{\pi}{2} + h\right) &= f\left(\frac{\pi}{2}\right) + f'\left(\frac{\pi}{2}\right)h + f''\left(\frac{\pi}{2}\right)\frac{h^2}{2!} + f'''\left(\frac{\pi}{2}\right)\frac{h^3}{3!} + f''''\left(\frac{\pi}{2}\right)\frac{h^4}{4!} + \dots \\ f\left(\frac{\pi}{2} + 0.42920\right) &= 1 + 0(0.42920) - 1\frac{(0.42920)^2}{2!} + 0\frac{(0.42920)^3}{3!} + 1\frac{(0.42920)^4}{4!} + \dots \\ &= 1 + 0 - 0.092106 + 0 + 0.00141393 + \dots \\ &\approx 0.90931 \end{aligned}$$

The value of  $\sin(2)$  I get from my calculator is 0.90930 which is very close to the value I just obtained. Now you can get a better value by using more terms of the series. In addition, you can now use the value calculated for  $\sin(2)$  coupled with the value of  $\cos(2)$  (which can be calculated by Taylor series just like this example or by using the  $\sin^2 x + \cos^2 x \equiv 1$  identity) to find value of  $\sin(x)$  at some other point. In this way, we can find the value of  $\sin(x)$  for any value from  $x = 0$  to  $2\pi$  and then can use the periodicity of  $\sin(x)$ , that is  $\sin(x) = \sin(x + 2n\pi), n = 1, 2, \dots$  to calculate the value of  $\sin(x)$  at any other point.

### Example 3

Derive the Maclaurin series of  $\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots$

### Solution

In the previous example, we wrote the Taylor series for  $\sin(x)$  around the point  $x = \frac{\pi}{2}$ .

Maclaurin series is simply a Taylor series for the point  $x = 0$ .

$$f(x) = \sin(x), \quad f(0) = 0$$

$$\begin{aligned}
 f'(x) &= \cos(x), f'(0) = 1 \\
 f''(x) &= -\sin(x), f''(0) = 0 \\
 f'''(x) &= -\cos(x), f'''(0) = -1 \\
 f''''(x) &= \sin(x), f''''(0) = 0 \\
 f'''''(x) &= \cos(x), f'''''(0) = 1
 \end{aligned}$$

Using the Taylor series now,

$$\begin{aligned}
 f(x+h) &= f(x) + f'(x)h + f''(x)\frac{h^2}{2!} + f'''(x)\frac{h^3}{3!} + f''''(x)\frac{h^4}{4!} + f'''''(x)\frac{h^5}{5!} + \dots \\
 f(0+h) &= f(0) + f'(0)h + f''(0)\frac{h^2}{2!} + f'''(0)\frac{h^3}{3!} + f''''(0)\frac{h^4}{4!} + f'''''(0)\frac{h^5}{5!} + \dots \\
 f(h) &= f(0) + f'(0)h + f''(0)\frac{h^2}{2!} + f'''(0)\frac{h^3}{3!} + f''''(0)\frac{h^4}{4!} + f'''''(0)\frac{h^5}{5!} + \dots \\
 &= 0 + 1(h) - 0\frac{h^2}{2!} - 1\frac{h^3}{3!} + 0\frac{h^4}{4!} + 1\frac{h^5}{5!} + \dots \\
 &= h - \frac{h^3}{3!} + \frac{h^5}{5!} + \dots
 \end{aligned}$$

So

$$\begin{aligned}
 f(x) &= x - \frac{x^3}{3!} + \frac{x^5}{5!} - \dots \\
 \sin(x) &= x - \frac{x^3}{3!} + \frac{x^5}{5!} - \dots
 \end{aligned}$$

#### Example 4

Find the value of  $f(6)$  given that  $f(4) = 125$ ,  $f'(4) = 74$ ,  $f''(4) = 30$ ,  $f'''(4) = 6$  and all other higher derivatives of  $f(x)$  at  $x = 4$  are zero.

#### Solution

$$\begin{aligned}
 f(x+h) &= f(x) + f'(x)h + f''(x)\frac{h^2}{2!} + f'''(x)\frac{h^3}{3!} + \dots \\
 x &= 4 \\
 h &= 6 - 4 \\
 &= 2
 \end{aligned}$$

Since fourth and higher derivatives of  $f(x)$  are zero at  $x = 4$ .

$$\begin{aligned}
 f(4+2) &= f(4) + f'(4)2 + f''(4)\frac{2^2}{2!} + f'''(4)\frac{2^3}{3!} \\
 f(6) &= 125 + 74(2) + 30\left(\frac{2^2}{2!}\right) + 6\left(\frac{2^3}{3!}\right) \\
 &= 125 + 148 + 60 + 8 \\
 &= 341
 \end{aligned}$$

Note that to find  $f(6)$  exactly, we only needed the value of the function and all its derivatives at some other point, in this case,  $x = 4$ . We did not need the expression for the function and all its derivatives. Taylor series application would be redundant if we needed to know the expression for the function, as we could just substitute  $x = 6$  in it to get the value of  $f(6)$ .

Actually the problem posed above was obtained from a known function  $f(x) = x^3 + 3x^2 + 2x + 5$  where  $f(4) = 125$ ,  $f'(4) = 74$ ,  $f''(4) = 30$ ,  $f'''(4) = 6$ , and all other higher derivatives are zero.

### Error in Taylor Series

As you have noticed, the Taylor series has infinite terms. Only in special cases such as a finite polynomial does it have a finite number of terms. So whenever you are using a Taylor series to calculate the value of a function, it is being calculated approximately.

The Taylor polynomial of order  $n$  of a function  $f(x)$  with  $(n+1)$  continuous derivatives in the domain  $[x, x+h]$  is given by

$$f(x+h) = f(x) + f'(x)h + f''(x)\frac{h^2}{2!} + \cdots + f^{(n)}(x)\frac{h^n}{n!} + R_n(x+h)$$

where the remainder is given by

$$R_n(x+h) = \frac{(h)^{n+1}}{(n+1)!} f^{(n+1)}(c).$$

where

$$x < c < x+h$$

that is,  $c$  is some point in the domain  $(x, x+h)$ .

### Example 5

The Taylor series for  $e^x$  at point  $x = 0$  is given by

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \cdots$$

- a) What is the truncation (true) error in the representation of  $e^1$  if only four terms of the series are used?
- b) Use the remainder theorem to find the bounds of the truncation error.

#### Solution

- a) If only four terms of the series are used, then

$$\begin{aligned} e^x &\approx 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} \\ e^1 &\approx 1 + 1 + \frac{1^2}{2!} + \frac{1^3}{3!} \\ &= 2.66667 \end{aligned}$$

The truncation (true) error would be the unused terms of the Taylor series, which then are

$$\begin{aligned}
 E_t &= \frac{x^4}{4!} + \frac{x^5}{5!} + \dots \\
 &= \frac{1^4}{4!} + \frac{1^5}{5!} + \dots \\
 &\approx 0.0516152
 \end{aligned}$$

- b) But is there any way to know the bounds of this error other than calculating it directly? Yes,

$$f(x+h) = f(x) + f'(x)h + \dots + f^{(n)}(x) \frac{h^n}{n!} + R_n(x+h)$$

where

$$R_n(x+h) = \frac{(h)^{n+1}}{(n+1)!} f^{(n+1)}(c), \quad x < c < x+h, \text{ and}$$

$c$  is some point in the domain  $(x, x+h)$ . So in this case, if we are using four terms of the Taylor series, the remainder is given by ( $x = 0, n = 3$ )

$$\begin{aligned}
 R_3(0+1) &= \frac{(1)^{3+1}}{(3+1)!} f^{(3+1)}(c) \\
 &= \frac{1}{4!} f^{(4)}(c) \\
 &= \frac{e^c}{24}
 \end{aligned}$$

Since

$$\begin{aligned}
 x < c < x+h \\
 0 < c < 0+1 \\
 0 < c < 1
 \end{aligned}$$

The error is bound between

$$\begin{aligned}
 \frac{e^0}{24} < R_3(1) &< \frac{e^1}{24} \\
 \frac{1}{24} < R_3(1) &< \frac{e}{24} \\
 0.041667 < R_3(1) &< 0.113261
 \end{aligned}$$

So the bound of the error is less than 0.113261 which does concur with the calculated error of 0.0516152.

### Example 6

The Taylor series for  $e^x$  at point  $x = 0$  is given by

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \dots$$

As you can see in the previous example that by taking more terms, the error bounds decrease and hence you have a better estimate of  $e^1$ . How many terms it would require to get an approximation of  $e^1$  within a magnitude of true error of less than  $10^{-6}$ ?

**Solution**

Using  $(n+1)$  terms of the Taylor series gives an error bound of

$$R_n(x+h) = \frac{(h)^{n+1}}{(n+1)!} f^{(n+1)}(c)$$

$$x = 0, h = 1, f(x) = e^x$$

$$R_n(1) = \frac{(1)^{n+1}}{(n+1)!} f^{(n+1)}(c)$$

$$= \frac{(1)^{n+1}}{(n+1)!} e^c$$

Since

$$x < c < x + h$$

$$0 < c < 0 + 1$$

$$0 < c < 1$$

$$\frac{1}{(n+1)!} < |R_n(1)| < \frac{e}{(n+1)!}$$

So if we want to find out how many terms it would require to get an approximation of  $e^1$  within a magnitude of true error of less than  $10^{-6}$ ,

$$\frac{e}{(n+1)!} < 10^{-6}$$

$$(n+1)! > 10^6 e$$

$$(n+1)! > 10^6 \times 3 \quad (\text{as we do not know the value of } e \text{ but it is less than 3}).$$

$$n \geq 9$$

So 9 terms or more will get  $e^1$  within an error of  $10^{-6}$  in its value.

We can do calculations such as the ones given above only for simple functions. To do a similar analysis of how many terms of the series are needed for a specified accuracy for any general function, we can do that based on the concept of absolute relative approximate errors discussed in Chapter 01.02 as follows.

We use the concept of absolute relative approximate error (see Chapter 01.02 for details), which is calculated after each term in the series is added. The maximum value of  $m$ , for which the absolute relative approximate error is less than  $0.5 \times 10^{2-m}\%$  is the least number of significant digits correct in the answer. It establishes the accuracy of the approximate value of a function without the knowledge of remainder of Taylor series or the true error.

---

---

INTRODUCTION TO NUMERICAL METHODS

---

Topic Taylor Theorem Revisited  
Summary These are textbook notes on Taylor Series  
Major All engineering majors  
Authors Autar Kaw  
Date June 3, 2014  
Web Site <http://numericalmethods.eng.usf.edu>

---

# Chapter 02.01

## Primer on Differentiation

*After reading this chapter, you should be able to:*

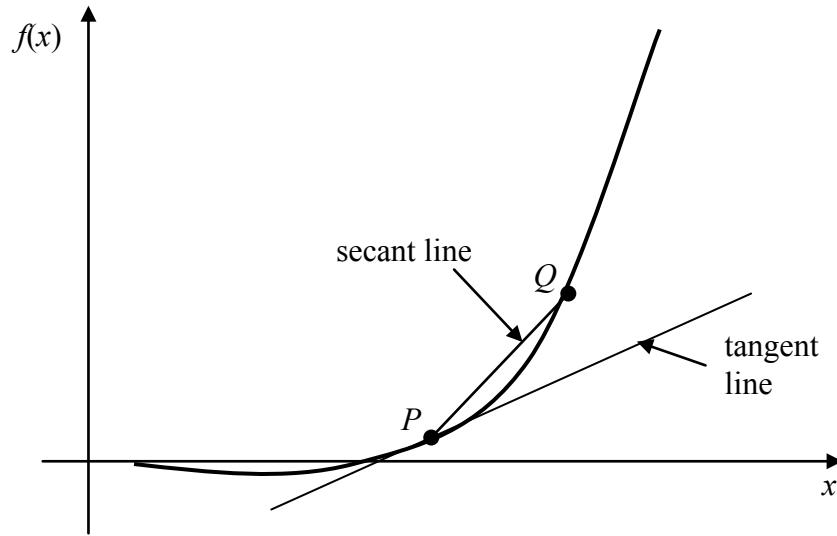
1. *understand the basics of differentiation,*
2. *relate the slopes of the secant line and tangent line to the derivative of a function,*
3. *find derivatives of polynomial, trigonometric and transcendental functions,*
4. *use rules of differentiation to differentiate functions,*
5. *find maxima and minima of a function, and*
6. *apply concepts of differentiation to real world problems.*

In this primer, we will review the concepts of differentiation you learned in calculus. Mostly those concepts are reviewed that are applicable in learning about numerical methods. These include the concepts of the secant line to learn about numerical differentiation of functions, the slope of a tangent line as a background to solving nonlinear equations using the Newton-Raphson method, finding maxima and minima of functions as a means of optimization, the use of the Taylor series to approximate functions, etc.

### Introduction

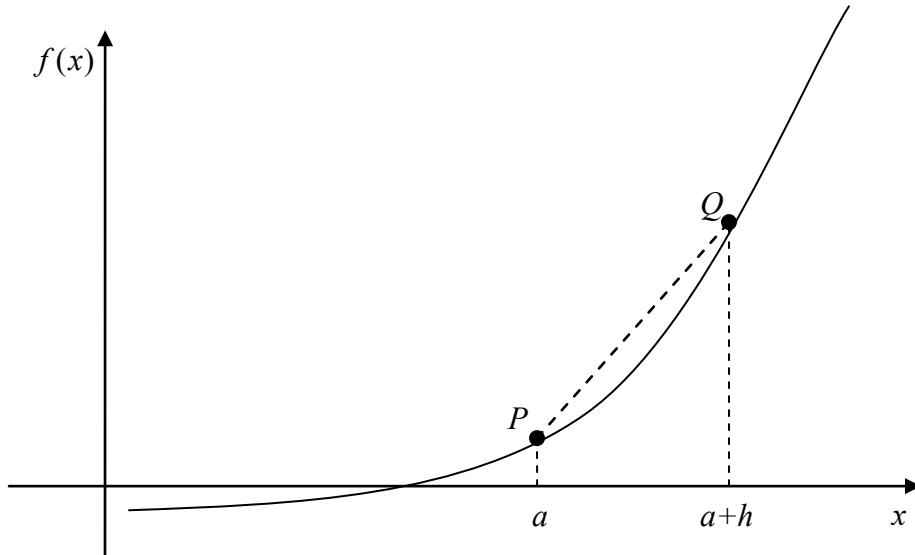
The derivative of a function represents the rate of change of a variable with respect to another variable. For example, the velocity of a body is defined as the rate of change of the location of the body with respect to time. The location is the *dependent* variable while time is the *independent* variable. Now if we measure the rate of change of velocity with respect to time, we get the acceleration of the body. In this case, the velocity is the *dependent* variable while time is the *independent* variable.

Whenever differentiation is introduced to a student, two concepts of the secant line and tangent line (Figure 1) are revisited.



**Figure 1** Function curve with tangent and secant lines.

Let  $P$  and  $Q$  be two points on the curve as shown in Figure 1. The secant line is the straight line drawn through  $P$  and  $Q$ .



**Figure 2** Calculation of the secant line.

The slope of the secant line (Figure 2) is then given as

$$m_{PQ,\text{secant}} = \frac{f(a+h) - f(a)}{(a+h) - a}$$

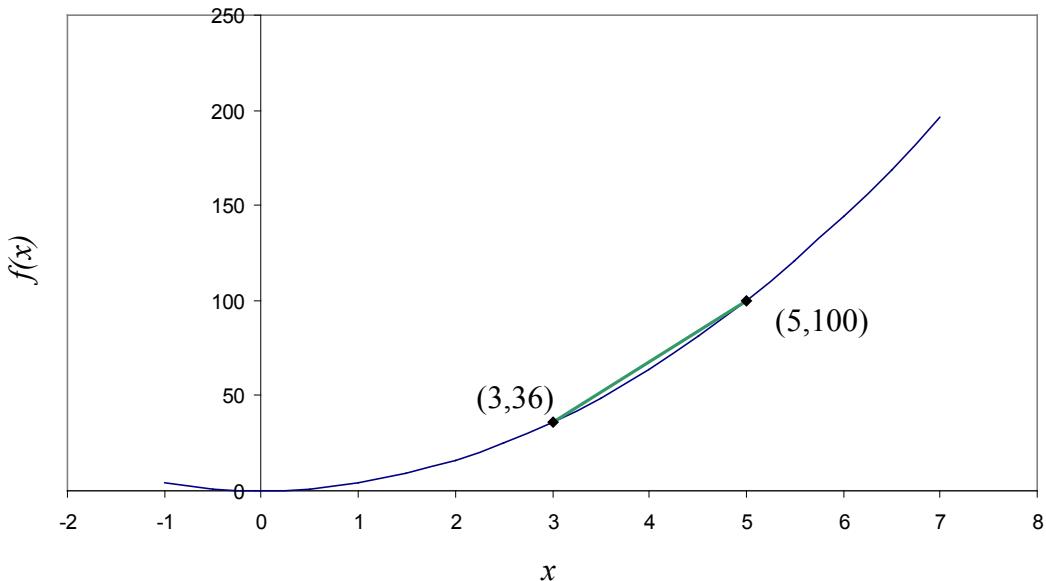
$$= \frac{f(a+h) - f(a)}{h}$$

As  $Q$  moves closer and closer to  $P$ , the limiting portion is called the tangent line. The slope of the tangent line  $m_{PQ,\text{tangent}}$  then is the limiting value of  $m_{PQ,\text{secant}}$  as  $h \rightarrow 0$ .

$$m_{PQ,\text{tangent}} = \lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h}$$

### Example 1

Find the slope of the secant line of the curve  $y = 4x^2$  between points  $(3, 36)$  and  $(5, 100)$ .



**Figure 3** Calculation of the secant line for the function  $y = 4x^2$ .

### Solution

The slope of the secant line between  $(3, 36)$  and  $(5, 100)$  is

$$\begin{aligned} m &= \frac{f(5) - f(3)}{5 - 3} \\ &= \frac{100 - 36}{5 - 3} \\ &= 32 \end{aligned}$$

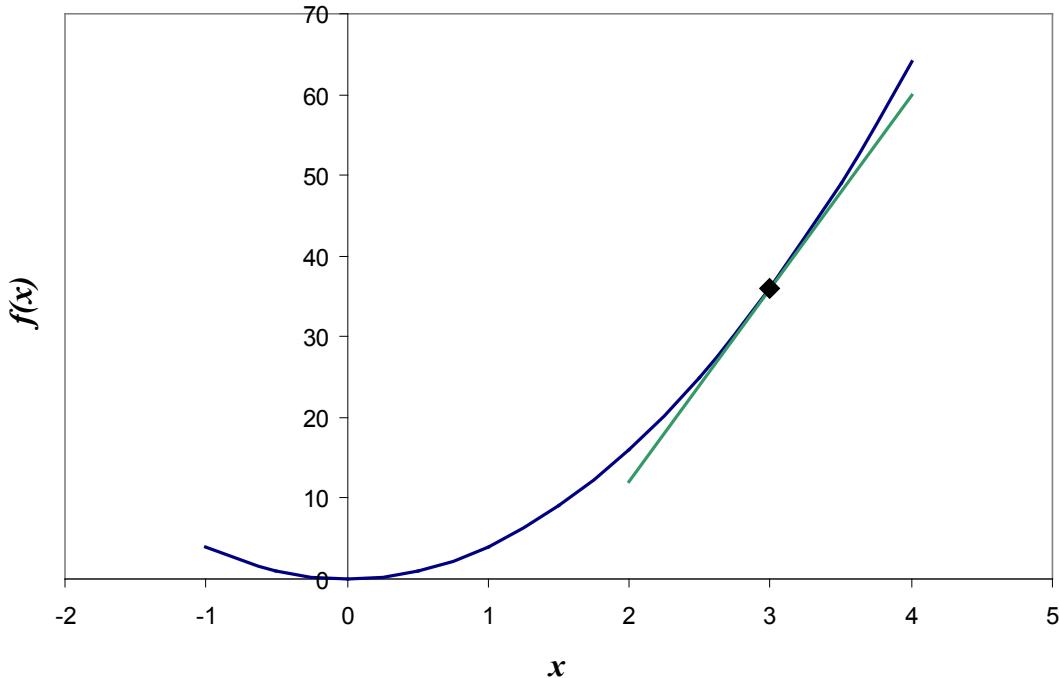
### Example 2

Find the slope of the tangent line of the curve  $y = 4x^2$  at point  $(3, 36)$ .

### Solution

The slope of the tangent line at  $(3, 36)$  is

$$\begin{aligned}
 m &= \lim_{h \rightarrow 0} \frac{f(3+h) - f(3)}{h} \\
 &= \lim_{h \rightarrow 0} \frac{4(3+h)^2 - 4(3)^2}{h} \\
 &= \lim_{h \rightarrow 0} \frac{4(9+h^2+6h)-36}{h} \\
 &= \lim_{h \rightarrow 0} \frac{36+4h^2+24h-36}{h} \\
 &= \lim_{h \rightarrow 0} \frac{h(4h+24)}{h} \\
 &= \lim_{h \rightarrow 0} (4h+24) \\
 &= 24
 \end{aligned}$$



**Figure 4** Calculation of the tangent line in the function  $y = 4x^2$ .

The slope of the tangent line is

$$\begin{aligned}
 m &= \lim_{h \rightarrow 0} \frac{f(3+h) - f(3)}{h} \\
 &= \lim_{h \rightarrow 0} \frac{4(3+h)^2 - 4(3)^2}{h} \\
 &= \lim_{h \rightarrow 0} \frac{36+24h+4h^2-36}{h}
 \end{aligned}$$

$$\begin{aligned}
 &= \lim_{h \rightarrow 0} \frac{h(24 + 4h)}{h} \\
 &= \lim_{h \rightarrow 0} (24 + 4h) \\
 &= 24
 \end{aligned}$$

### Derivative of a Function

Recall from calculus, the derivative of a function  $f(x)$  at  $x = a$  is defined as

$$f'(a) = \lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h}$$

#### Example 3

Find  $f'(3)$  if  $f(x) = 4x^2$ .

##### Solution

$$\begin{aligned}
 f'(3) &= \lim_{h \rightarrow 0} \frac{f(3+h) - f(3)}{h} \\
 &= \lim_{h \rightarrow 0} \frac{4(3+h)^2 - 4(3)^2}{h} \\
 &= \lim_{h \rightarrow 0} \frac{4(9 + h^2 + 6h) - 36}{h} \\
 &= \lim_{h \rightarrow 0} \frac{36 + 4h^2 + 24h - 36}{h} \\
 &= \lim_{h \rightarrow 0} \frac{h(4h + 24)}{h} \\
 &= \lim_{h \rightarrow 0} (4h + 24) \\
 &= 24
 \end{aligned}$$

#### Example 4

Find  $f'\left(\frac{\pi}{4}\right)$  if  $f(x) = \sin(2x)$

##### Solution

$$\begin{aligned}
 f'\left(\frac{\pi}{4}\right) &= \lim_{h \rightarrow 0} \frac{f\left(\frac{\pi}{4} + h\right) - f\left(\frac{\pi}{4}\right)}{h} \\
 &= \lim_{h \rightarrow 0} \frac{\sin\left(2\left(\frac{\pi}{4} + h\right)\right) - \sin\left(2\left(\frac{\pi}{4}\right)\right)}{h}
 \end{aligned}$$

$$\begin{aligned}
 &= \lim_{h \rightarrow 0} \frac{\sin\left(\frac{\pi}{2} + 2h\right) - \sin\left(\frac{\pi}{2}\right)}{h} \\
 &= \lim_{h \rightarrow 0} \frac{\sin\left(\frac{\pi}{2}\right)\cos(2h) + \cos\left(\frac{\pi}{2}\right)\sin(2h) - \sin\left(\frac{\pi}{2}\right)}{h} \\
 &= \lim_{h \rightarrow 0} \frac{\cos(2h) + 0 - 1}{h} \\
 &= \lim_{h \rightarrow 0} \frac{\cos(2h) - 1}{h} \\
 &= 0
 \end{aligned}$$

from knowing that

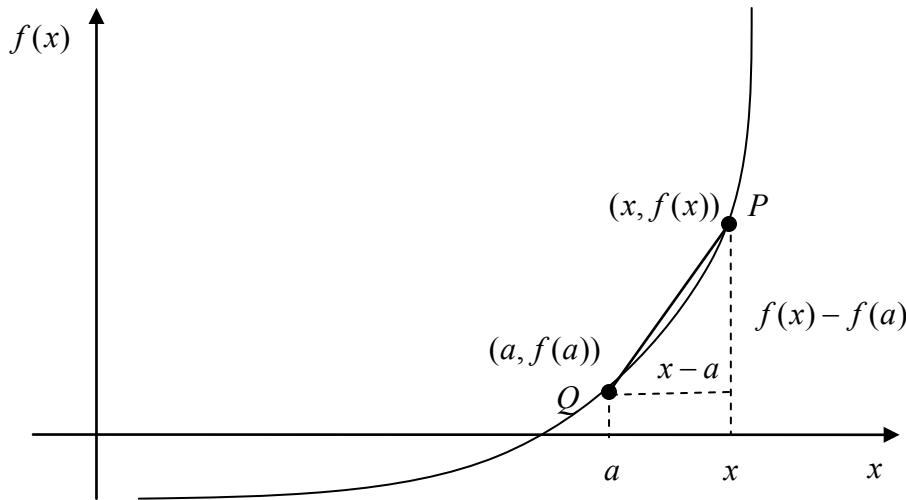
$$\lim_{h \rightarrow 0} \frac{1 - \cos(h)}{h} = 0$$

### Second Definition of Derivatives

There is another form of the definition of the derivative of a function. The derivative of the function  $f(x)$  at  $x = a$  is defined as

$$f'(a) = \lim_{x \rightarrow a} \frac{f(x) - f(a)}{x - a}$$

As  $x \rightarrow a$ , the definition is nothing but the slope of the tangent line at  $P$ .



**Figure 5** Graph showing the second definition of the derivative.

### Example 5

Find  $f'(3)$  if  $f(x) = 4x^2$  by using the form

$$f'(a) = \lim_{x \rightarrow a} \frac{f(x) - f(a)}{x - a}$$

of the definition of a derivative.

### Solution

$$\begin{aligned} f'(3) &= \lim_{x \rightarrow 3} \frac{f(x) - f(3)}{x - 3} \\ &= \lim_{x \rightarrow 3} \frac{4x^2 - 4(3)^2}{x - 3} \\ &= \lim_{x \rightarrow 3} \frac{4x^2 - 36}{x - 3} \\ &= \lim_{x \rightarrow 3} \frac{4(x^2 - 9)}{x - 3} \\ &= \lim_{x \rightarrow 3} \frac{4(x - 3)(x + 3)}{x - 3} \\ &= \lim_{x \rightarrow 3} 4(x + 3) \\ &= 4(3 + 3) \\ &= 24 \end{aligned}$$

### Finding equations of a tangent line

One of the numerical methods used to solve a nonlinear equation is called the *Newton-Raphson method*. This method is based on the knowledge of finding the tangent line to a curve at a point. Let us look at an example to illustrate finding the equation of the tangent line to a curve.

### Example 6

Find the equation of the line tangent to the function

$$f(x) = x^3 - 0.165x + 3.993 \times 10^{-4} \text{ at } x = 0.05.$$

### Solution

The line tangent is a straight line of the form

$$y = mx + c$$

To find the equation of the tangent line, let us first find the slope  $m$  of the straight line.

$$f'(x) = 3x^2 - 0.165$$

$$\begin{aligned} f'(0.05) &= 3(0.05)^2 - 0.165 \\ &= -0.1575 \end{aligned}$$

$$m = -0.1575$$

To find the value of the  $y$ -intercept  $c$  of the straight line, we first find the value of the function at  $x = 0.05$ .

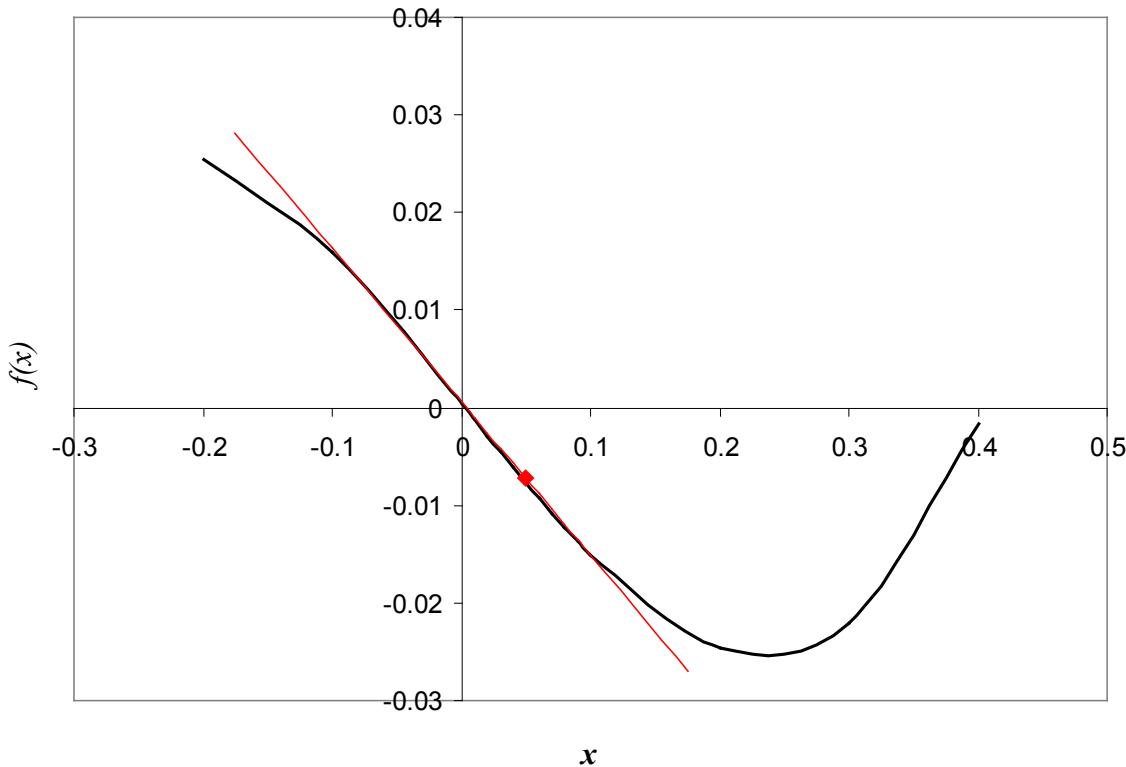
$$\begin{aligned} f(0.05) &= (0.05)^3 - 0.165(0.05) + 3.993 \times 10^{-4} \\ &= -0.0077257 \end{aligned}$$

The tangent line passes through the point  $(0.05, -0.0077257)$ , so

$$-0.0077257 = m(0.05) + c$$

$$-0.0077257 = -0.1575(0.05) + c$$

$$c = 0.0001493$$



**Figure 6** Graph of function  $f(x)$  and the tangent line at  $x = 0.05$ .

Hence,

$$y = mx + c$$

$$= -0.1575x + 0.0001493$$

is the equation of the tangent line.

### Other Notations of Derivatives

Derivatives can be denoted in several ways. For the first derivative, the notations are

$$f'(x), \frac{d}{dx}f(x), y', \text{ and } \frac{dy}{dx}$$

For the second derivative, the notations are

$$f''(x), \frac{d^2}{dx^2}f(x), y'', \text{ and } \frac{d^2y}{dx^2}$$

For the  $n^{th}$  derivative, the notations are

$$f^{(n)}(x), \frac{d^n}{dx^n} f(x), y^{(n)}, \frac{d^n y}{dx^n}$$

### Theorems of Differentiation

Several theorems of differentiation are given to show how one can find the derivative of different functions.

#### Theorem 1

The derivative of a constant is zero. If  $f(x) = k$ , where  $k$  is a constant,  $f'(x) = 0$ .

#### Example 7

Find the derivative of  $f(x) = 6$ .

#### Solution

$$\begin{aligned} f(x) &= 6 \\ f'(x) &= 0 \end{aligned}$$

#### Theorem 2

The derivative of  $f(x) = x^n$ , where  $n \neq 0$  is  $f'(x) = nx^{n-1}$ .

#### Example 8

Find the derivative of  $f(x) = x^6$ .

#### Solution

$$\begin{aligned} f(x) &= x^6 \\ f'(x) &= 6x^{6-1} \\ &= 6x^5 \end{aligned}$$

#### Example 9

Find the derivative of  $f(x) = x^{-6}$ .

#### Solution

$$\begin{aligned} f(x) &= x^{-6} \\ f'(x) &= -6x^{-6-1} \\ &= -6x^{-7} \\ &= -\frac{6}{x^7} \end{aligned}$$

#### Theorem 3

The derivative of  $f(x) = kg(x)$ , where  $k$  is a constant is  $f'(x) = kg'(x)$ .

**Example 10**

Find the derivative of  $f(x) = 10x^6$ .

**Solution**

$$\begin{aligned}f(x) &= 10x^6 \\f'(x) &= \frac{d}{dx}(10x^6) \\&= 10 \frac{d}{dx}x^6 \\&= 10(6x^5) \\&= 60x^5\end{aligned}$$

**Theorem 4**

The derivative of  $f(x) = u(x) \pm v(x)$  is  $f'(x) = u'(x) \pm v'(x)$ .

**Example 11**

Find the derivative of  $f(x) = 3x^3 + 8$ .

**Solution**

$$\begin{aligned}f(x) &= 3x^3 + 8 \\f'(x) &= \frac{d}{dx}(3x^3 + 8) \\&= \frac{d}{dx}(3x^3) + \frac{d}{dx}(8) \\&= 3 \frac{d}{dx}(x^3) + 0 \\&= 3(3x^2) \\&= 9x^2\end{aligned}$$

**Theorem 5**

The derivative of

$$f(x) = u(x)v(x)$$

is

$$f'(x) = u(x) \frac{d}{dx}v(x) + v(x) \frac{d}{dx}u(x). \quad (\text{Product Rule})$$

**Example 12**

Find the derivative of  $f(x) = (2x^2 - 6)(3x^3 + 8)$

### Solution

Using the product rule as given by Theorem 5 where,

$$f(x) = u(x)v(x)$$

$$f'(x) = u(x)\frac{d}{dx}v(x) + v(x)\frac{d}{dx}u(x)$$

$$f(x) = (2x^2 - 6)(3x^3 + 8)$$

$$u(x) = 2x^2 - 6$$

$$v(x) = 3x^3 + 8$$

Taking the derivative of  $u(x)$ ,

$$\begin{aligned}\frac{du}{dx} &= \frac{d}{dx}(2x^2 - 6) \\ &= \frac{d}{dx}(2x^2) - \frac{d}{dx}(6) \\ &= 2\frac{d}{dx}(x^2) - 0 \\ &= 2(2x) \\ &= 4x\end{aligned}$$

Taking the derivative of  $v(x)$ ,

$$\begin{aligned}\frac{dv}{dx} &= \frac{d}{dx}(3x^3 + 8) \\ &= \frac{d}{dx}(3x^3) + \frac{d}{dx}(8) \\ &= 3\frac{d}{dx}(x^3) + 0 \\ &= 3(3x^2) \\ &= 9x^2\end{aligned}$$

Using the formula for the product rule

$$\begin{aligned}f'(x) &= u(x)\frac{d}{dx}v(x) + v(x)\frac{d}{dx}u(x) \\ &= (2x^2 - 6)(9x^2) + (3x^3 + 8)(4x) \\ &= 18x^4 - 54x^2 + 12x^4 + 32x \\ &= 30x^4 - 54x^2 + 32x\end{aligned}$$

### Theorem 6

The derivative of

$$f(x) = \frac{u(x)}{v(x)}$$

is

$$f'(x) = \frac{v(x)\frac{d}{dx}u(x) - u(x)\frac{d}{dx}v(x)}{(v(x))^2} \quad (\text{Quotient Rule})$$

**Example 13**

Find the derivative of  $f(x) = \frac{(2x^2 - 6)}{(3x^3 + 8)}$ .

**Solution**

Use the quotient rule of Theorem 6, if

$$f(x) = \frac{u(x)}{v(x)}$$

then

$$f'(x) = \frac{v(x)\frac{d}{dx}u(x) - u(x)\frac{d}{dx}v(x)}{(v(x))^2}$$

From

$$f(x) = \frac{(2x^2 - 6)}{(3x^3 + 8)}$$

we have

$$u(x) = 2x^2 - 6$$

$$v(x) = 3x^3 + 8$$

Taking the derivative of  $u(x)$ ,

$$\begin{aligned} \frac{du}{dx} &= \frac{d}{dx}(2x^2 - 6) \\ &= \frac{d}{dx}(2x^2) - \frac{d}{dx}(6) \\ &= 2\frac{d}{dx}(x^2) - 0 \\ &= 2(2x) \\ &= 4x \end{aligned}$$

Taking the derivative of  $v(x)$ ,

$$\begin{aligned} \frac{dv}{dx} &= \frac{d}{dx}(3x^3 + 8) \\ &= \frac{d}{dx}(3x^3) + \frac{d}{dx}(8) \\ &= 3\frac{d}{dx}(x^3) + 0 \\ &= 3(3x^2) \end{aligned}$$

$$= 9x^2$$

Using the formula for the quotient rule,

$$\begin{aligned} f'(x) &= \frac{(3x^3 + 8)(4x) - (2x^2 - 6)(9x^2)}{(3x^3 + 8)^2} \\ &= \frac{12x^4 + 32x - 18x^4 + 54x^2}{9x^6 + 48x^3 + 64} \\ &= \frac{-6x^4 + 54x^2 + 32x}{9x^6 + 48x^3 + 64} \end{aligned}$$

### Table of Derivatives

$f(x)$	$f'(x)$
$x^n, n \neq 0$	$nx^{n-1}$
$kx^n, n \neq 0$	$knx^{n-1}$
$\sin(x)$	$\cos(x)$
$\cos(x)$	$-\sin(x)$
$\tan(x)$	$\sec^2(x)$
$\sinh(x)$	$\cosh(x)$
$\cosh(x)$	$\sinh(x)$
$\tanh(x)$	$1 - \tanh^2(x)$
$\sin^{-1}(x)$	$\frac{1}{\sqrt{1-x^2}}$
$\cos^{-1}(x)$	$\frac{-1}{\sqrt{1-x^2}}$
$\tan^{-1}(x)$	$\frac{1}{1+x^2}$
$\csc(x)$	$-\csc(x)\cot(x)$
$\sec(x)$	$\sec(x)\tan(x)$
$\cot(x)$	$-\csc^2(x)$
$\operatorname{csch}(x)$	$-\coth(x)\operatorname{csch}(x)$
$\operatorname{sech}(x)$	$-\tanh(x)\operatorname{sech}(x)$

$\coth(x)$	$1 - \coth^2(x)$
$\csc^{-1}(x)$	$-\frac{ x }{x^2\sqrt{x^2-1}}$
$\sec^{-1}(x)$	$\frac{ x }{x^2\sqrt{x^2-1}}$
$\cot^{-1}(x)$	$\frac{-1}{1+x^2}$
$a^x$	$\ln(a)a^x$
$\ln(x)$	$\frac{1}{x}$
$\log_a(x)$	$\frac{1}{x\ln(a)}$
$e^x$	$e^x$

### Chain Rule of Differentiation

Sometimes functions that need to be differentiated do not fall in the form of simple functions or the forms described previously. Such functions can be differentiated using the chain rule if they are of the form  $f(g(x))$ . The chain rule states

$$\frac{d}{dx}(f(g(x))) = f'(g(x))g'(x)$$

For example, to find  $f'(x)$  of  $f(x) = (3x^2 - 2x)^4$ , one could use the chain rule.

$$g(x) = (3x^2 - 2x)$$

$$g'(x) = 6x - 2$$

$$f'(g(x)) = 4(g(x))^3$$

$$\frac{d}{dx}((3x^2 - 2x)^4) = 4(3x^2 - 2x)^3(6x - 2)$$

### Implicit Differentiation

Sometimes, the function to be differentiated is not given explicitly as an expression of the independent variable. In such cases, how do we find the derivatives? We will discuss this via examples.

### Example 14

Find  $\frac{dy}{dx}$  if  $x^2 + y^2 = 2xy$

**Solution**

$$\begin{aligned}
 x^2 + y^2 &= 2xy \\
 \frac{d}{dx}(x^2 + y^2) &= \frac{d}{dx}(2xy) \\
 \frac{d}{dx}(x^2) + \frac{d}{dx}(y^2) &= \frac{d}{dx}(2xy) \\
 2x + 2y \frac{dy}{dx} &= 2x \frac{dy}{dx} + 2y \\
 2y \frac{dy}{dx} - 2x \frac{dy}{dx} &= 2y - 2x \\
 (2y - 2x) \frac{dy}{dx} &= 2y - 2x \\
 \frac{dy}{dx} &= \frac{2y - 2x}{2y - 2x} \\
 \frac{dy}{dx} &= 1
 \end{aligned}$$

**Example 15**

If  $x^2 - xy + y^2 = 5$ , find the value of  $y'$ .

**Solution**

$$\begin{aligned}
 x^2 - xy + y^2 &= 5 \\
 \frac{d}{dx}(x^2 - xy + y^2) &= \frac{d}{dx}(5) \\
 \frac{d}{dx}(x^2) - \frac{d}{dx}(xy) + \frac{d}{dx}(y^2) &= 0 \\
 2x - x \frac{dy}{dx} - y + 2y \frac{dy}{dx} &= 0 \\
 (-x + 2y) \frac{dy}{dx} &= -2x + y \\
 \frac{dy}{dx} &= \frac{y - 2x}{2y - x} \\
 y' &= \frac{y - 2x}{2y - x}
 \end{aligned}$$

**Higher order derivatives**

So far, we have limited our discussion to calculating first derivative,  $f'(x)$  of a function  $f(x)$ . What if we are asked to calculate higher order derivatives of  $f(x)$ .

A simple example of this is finding acceleration of a body from a function that gives the location of the body as a function of time. The derivative of the location with respect to time

is the velocity of the body, followed by the derivative of velocity with respect to time being the acceleration. Hence, the second derivative of the location function gives the acceleration function of the body.

### Example 16

Given  $f(x) = 3x^3 - 2x - 7$ , find the second derivative,  $f''(x)$  and the third derivative,  $f'''(x)$ .

#### Solution

Given

$$f(x) = 3x^3 - 2x - 7$$

we have

$$\begin{aligned} f'(x) &= 3(3x^2) - 2 \\ &= 9x^2 - 2 \end{aligned}$$

$$\begin{aligned} f''(x) &= \frac{d}{dx}(f'(x)) \\ &= \frac{d}{dx}(9x^2 - 2) \\ &= 9(2x) \\ &= 18x \end{aligned}$$

$$\begin{aligned} f'''(x) &= \frac{d}{dx}(f''(x)) \\ &= \frac{d}{dx}(18x) \\ &= 18 \end{aligned}$$

### Example 17

If  $x^2 - xy + y^2 = 5$ , find the value of  $y''$ .

#### Solution

From Example 15 we obtain

$$y' = \frac{y - 2x}{2y - x},$$

$$(2y - x)y' = y - 2x$$

$$\frac{d}{dx}((2y - x)y') = \frac{d}{dx}(y - 2x)$$

$$(2y - x)\frac{d}{dx}(y') + y'\frac{d}{dx}(2y - x) = \frac{d}{dx}(y) - \frac{d}{dx}(2x)$$

$$y''(2y - x) + y'(2y' - 1) = y' - 2$$

$$y'' = \frac{2y' - 2 - 2y'^2}{2y - x}$$

After substitution of  $y'$ ,

$$\begin{aligned}y'' &= \frac{2\frac{y-2x}{2y-x}-2-2\left(\frac{y-2x}{2y-x}\right)^2}{2y-x} \\&= -\frac{6(y^2-xy+x^2)}{(2y-x)^3}\end{aligned}$$

### Finding maximum and minimum of a function

The knowledge of first derivative and second derivative of a function is used to find the minimum and maximum of a function. First, let us define what the maximum and minimum of a function are. Let  $f(x)$  be a function in domain  $D$ , then

$f(a)$  is the maximum of the function if  $f(a) \geq f(x)$  for all values of  $x$  in the domain  $D$ .

$f(a)$  is the minimum of the function if  $f(a) \leq f(x)$  for all values of  $x$  in the domain  $D$ .

The minimum and maximum of a function are also the critical values of a function.

An extreme value can occur in the interval  $[c,d]$  at

end points  $x = c, x = d$ .

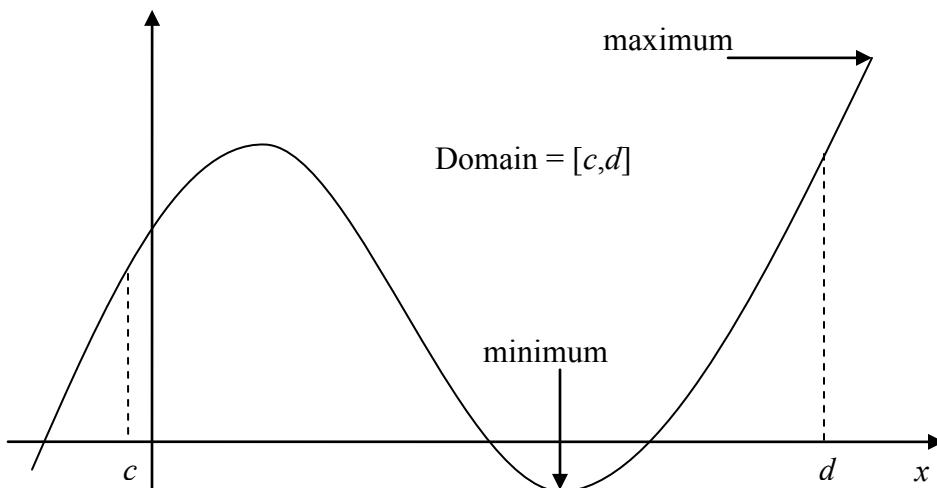
a point in  $[c,d]$  where  $f'(x) = 0$ .

a point in  $[c,d]$  where  $f'(x)$  does not exist.

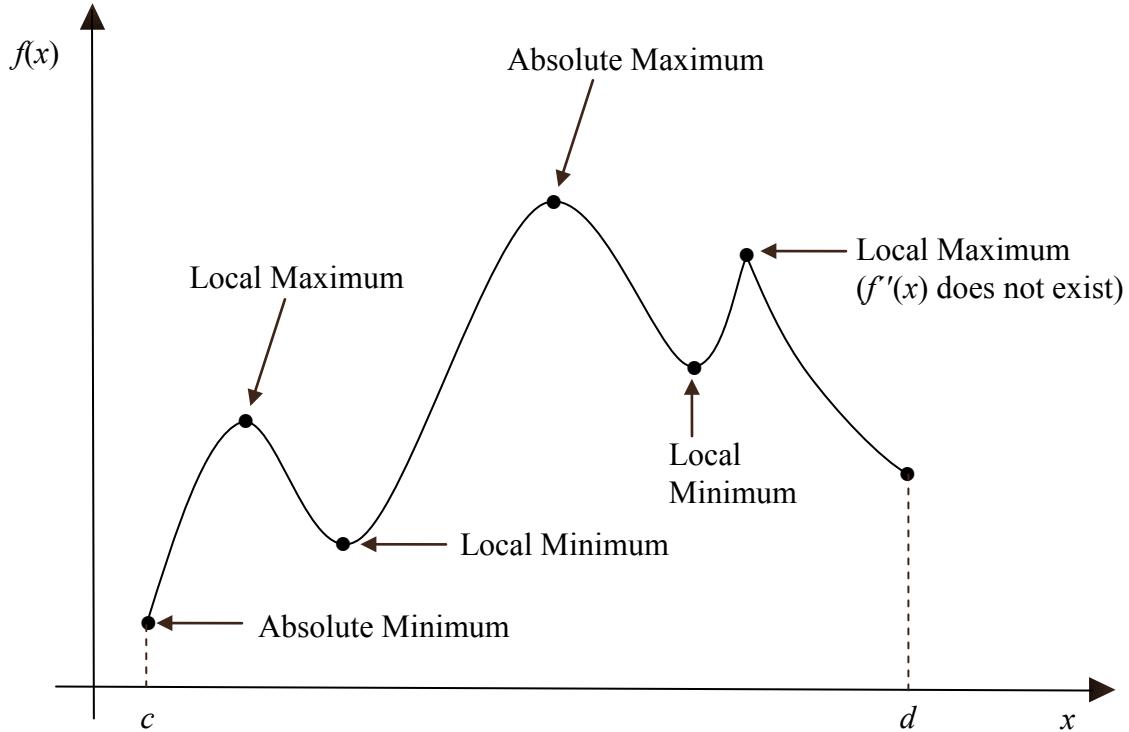
These critical points can be the local maximas and minimas of the function (See Figure 8).

### Example 18

Find the minimum and maximum value of  $f(x) = x^2 - 2x - 5$  in the interval  $[0,5]$ .



**Figure 7** Graph illustrating the concepts of maximum and minimum.



**Figure 8** The plot shows critical points of  $f(x)$  in  $[c, d]$ .

### Solution

$$f(x) = x^2 - 2x - 5$$

$$f'(x) = 2x - 2$$

$$f'(x) = 0 \text{ at } x = 1.$$

$f'(x)$  exists everywhere in  $[0, 5]$ .

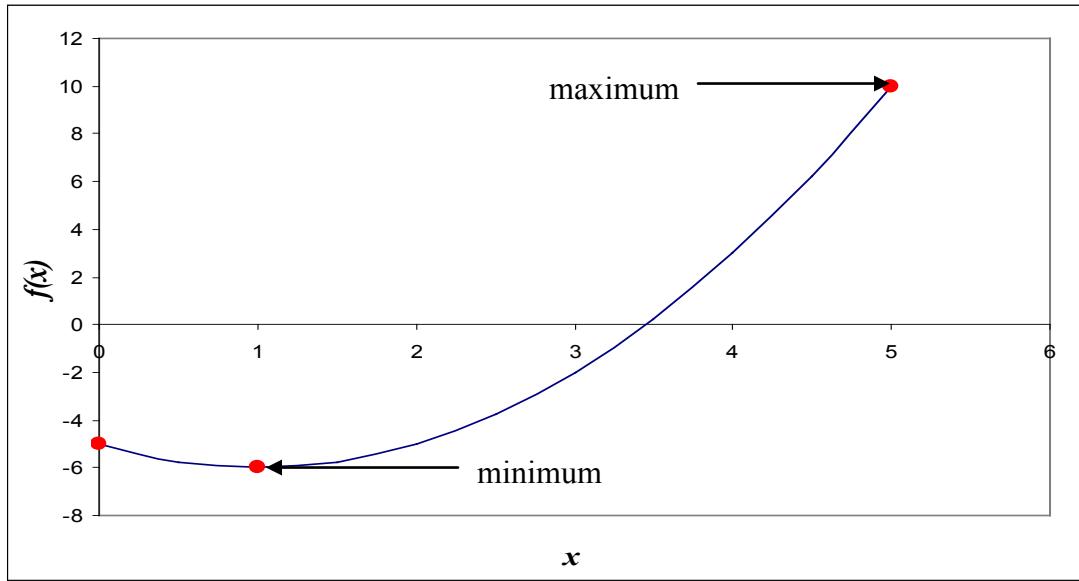
So the critical points are  $x = 0, x = 1, x = 5$ .

$$f(0) = (0)^2 - 2(0) - 5 = -5$$

$$f(1) = (1)^2 - 2(1) - 5 = -6$$

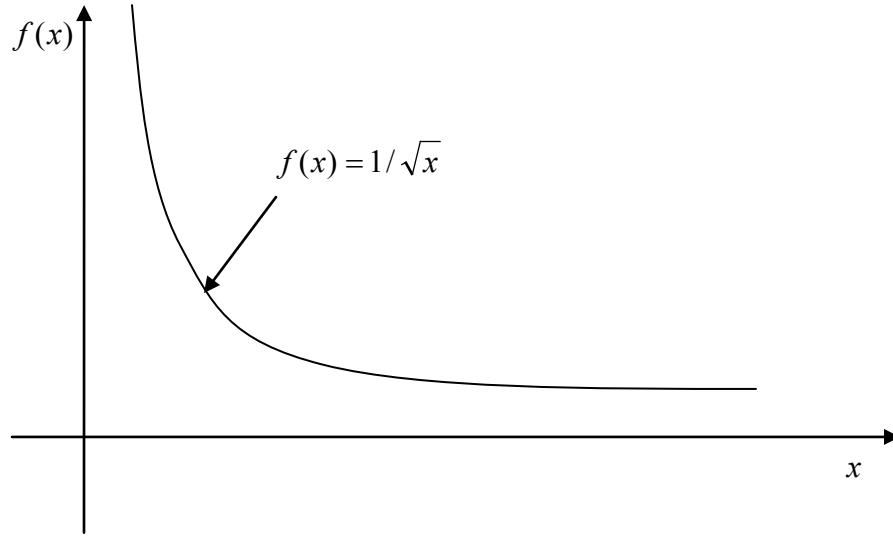
$$f(5) = (5)^2 - 2(5) - 5 = 10$$

Hence, the minimum value of  $f(x)$  occurs at  $x = 1$ , and the maximum value occurs at  $x = 5$ .



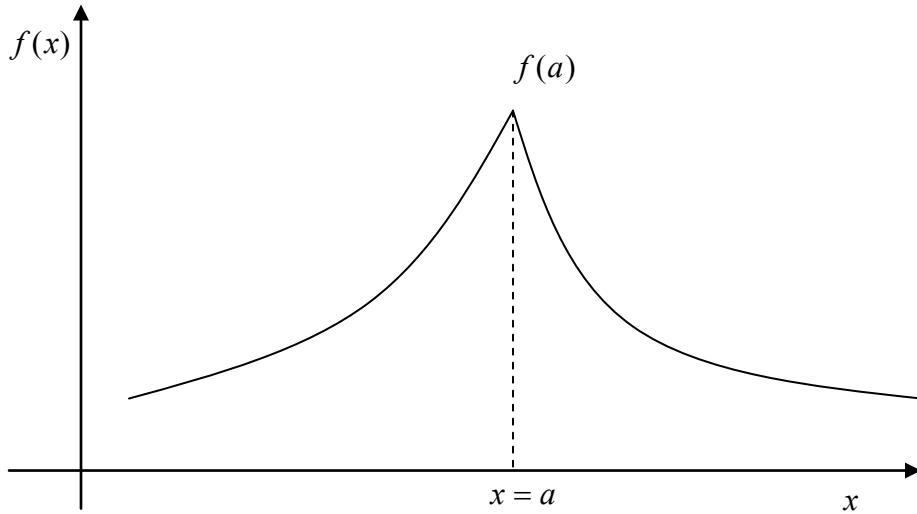
**Figure 9** Maximum and minimum values of  $f(x) = x^2 - 2x - 5$  over interval  $[0,5]$ .

Figure 10 shows an example of a function that has no minimum or maximum value in the domain  $(0, \infty)$ .



**Figure 10** Function that has no maximum or minimum.

Figure 11 shows the maximum of the function occurring at a singular point. The function  $f(x)$  has a sharp corner at  $x = a$ .



**Figure 11** Graph demonstrates the concept of a singular point with discontinuous slope at  $x = a$

### Example 19

Find the maximum and minimum of  $f(x) = 2x$  in the interval  $[0,5]$ .

#### Solution

$$f(x) = 2x$$

$$f'(x) = 2$$

$f'(x) \neq 0$  on  $[0,5]$ .

So the critical points are  $x = 0$  and  $x = 5$ .

$$f(x) = 2x$$

$$f(0) = 2(0) = 0$$

$$f(5) = 2(5) = 10$$

So the minimum value of  $f(x) = 2x$  is at  $x = 0$ , and the maximum value is at  $x = 5$ .

The point(s) where the second derivative of a function becomes zero is a way to know whether the critical point found in the first derivative test is a local minimum or maximum.

Let  $f(x)$  be a function in the interval  $(c,d)$  and  $f(a) = 0$ .

$f(a)$  is a local maximum of the function if  $f''(a) < 0$ .

$f(a)$  is a local minimum of the function if  $f''(a) > 0$ .

If  $f''(a) = 0$ , then the second derivative does not offer any insight into the local maxima or minima.

### Example 20

Remember Example 18 where we found  $f'(x) = 0$  at  $x = 1$  for  $f(x) = x^2 - 2x - 5$  in the interval  $[0,5]$ . Is  $x = 1$  a local maxima or minima of the function?

**Solution**

$$f(x) = x^2 - 2x - 5$$

$$f'(x) = 2x - 2$$

$$f'(x) = 0 \text{ at } x = 1$$

$$f''(x) = 2$$

$$f''(1) = 2 > 0$$

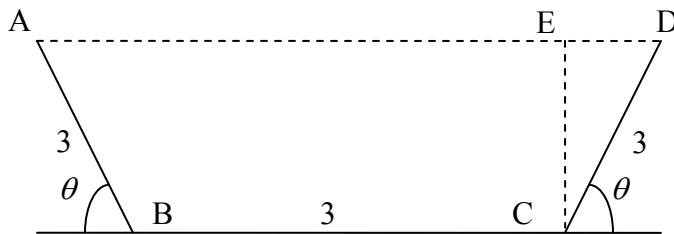
So the  $f(1)$  is the local minimum of the function.

**Applications of Derivatives**

Below are some examples to show real-life applications of differentiation.

**Example 21**

A rain gutter cross-section is shown below.



**Figure 12** Gutter dimensions for Example 21.

What angle of  $\theta$  would make the cross-sectional area of ABCD maximum? Note that common sense or intuition may lead us to believe that  $\theta = \pi/4$  would maximize the cross-sectional area of ABCD. Question your intuition.

**Solution**

$$\text{Area} = \frac{1}{2}(\overline{BC} + \overline{AD}) \times \overline{CE}$$

$$\begin{aligned}\overline{CE} &= \overline{CD} \sin(\theta) \\ &= 3 \sin(\theta)\end{aligned}$$

$$\overline{BC} = 3$$

$$\overline{AD} = \overline{BC} + \overline{CD} \cos(\theta) + \overline{AB} \cos(\theta)$$

$$\overline{AD} = 3 + 3 \cos(\theta) + 3 \cos(\theta)$$

$$\overline{AD} = 3 + 6 \cos(\theta)$$

$$\text{Area} = \frac{1}{2}(3 + 3 + 6 \cos(\theta))(3 \sin(\theta))$$

$$\begin{aligned}
 &= 9\sin(\theta) + 9\sin(\theta)\cos(\theta) \\
 &= 9\sin(\theta) + \frac{9}{2}\sin(2\theta) \\
 \frac{dA}{d\theta} &= 9\cos(\theta) + \frac{9}{2} \times 2\cos(2\theta) \\
 &= 9\cos(\theta) + 9\cos(2\theta)
 \end{aligned}$$

When is

$$\frac{dA}{d\theta} = 0?$$

$$9\cos(\theta) + 9\cos(2\theta) = 0$$

$$\theta = \frac{\pi}{3}$$

The angle at which the area is maximum is  $\theta = 60^\circ$ .

$$\begin{aligned}
 \text{Area}\left(\frac{\pi}{3}\right) &= 9\sin\left(\frac{\pi}{3}\right) + \frac{9}{2}\sin\left(2\left(\frac{\pi}{3}\right)\right) \\
 &= 9\left(\frac{\sqrt{3}}{2}\right) + \frac{9}{2}\left(\frac{\sqrt{3}}{2}\right) \\
 &= \frac{27}{4}\sqrt{3}
 \end{aligned}$$

For the interval of  $\theta = [0, \pi]$ , the area at the end points is

$$\text{Area}(0) = 0$$

$$\text{Area}(\pi) = 0$$

### Example 22

A classic example of the application of differentiation is to find the dimensions of a circular cylinder for a specific volume but which uses the least amount of material. Do this classic problem for a volume of  $9m^3$ .

#### Solution

The total surface area,  $A$  of the cylinder is

$$\begin{aligned}
 A &= \text{top surface} + \text{side surface} + \text{bottom surface} \\
 &= \pi r^2 + 2\pi rh + \pi r^2 \\
 &= 2\pi r^2 + 2\pi rh
 \end{aligned}$$

The volume,  $V$  of the cylinder is

$$V = \pi r^2 h$$

since

$$V = 9m^3.$$

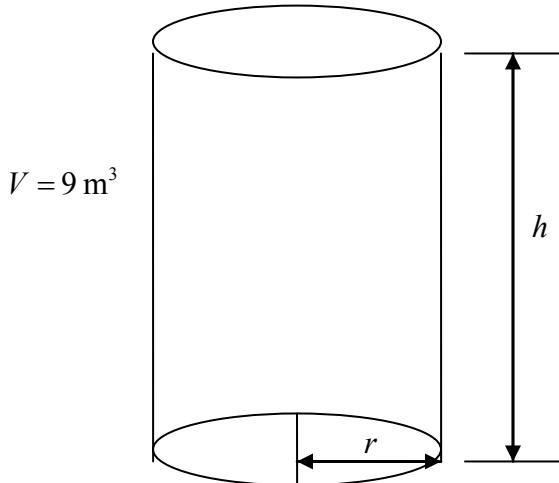
We can write

$$9 = \pi r^2 h$$

$$h = \frac{9}{\pi r^2}$$

This gives the surface area just in terms of  $r$  as

$$\begin{aligned} A &= 2\pi r^2 + 2\pi r \left( \frac{9}{\pi r^2} \right) \\ &= 2\pi r^2 + \frac{18}{r} \\ &= 2\pi r^2 + 18r^{-1} \end{aligned}$$



**Figure 13** Cylinder drawing for Example 20.

To find the minimum, take the first derivative of  $A$  with respect to  $r$  as

$$\begin{aligned} \frac{dA}{dr} &= 4\pi r + 18(-1)r^{-2} \\ &= 4\pi r - \frac{18}{r^2} \end{aligned}$$

Solving for

$$\frac{dA}{dr} = 0,$$

$$4\pi r - \frac{18}{r^2} = 0$$

$$4\pi r^3 - 18 = 0$$

$$r^3 = \frac{18}{4\pi}$$

$$r = \left( \frac{18}{4\pi} \right)^{\frac{1}{3}}$$

$$= 1.12725 \text{ m}$$

Since

$$h = \frac{9}{\pi r^2},$$

$$h = \frac{9}{\pi(1.12725)^2}$$

$$= 2.25450 \text{ m}$$

But does this value of  $r$  correspond to a minimum?

$$\frac{d^2 A}{dr^2} = 4\pi - 18(-2)r^{-3}$$

$$= 4\pi + \frac{36}{r^3}$$

$$= 4\pi + \frac{36}{1.12725}$$

$$= 44.5025$$

This value  $\frac{d^2 A}{dr^2} > 0$  for  $r = 1.12725 \text{ m}$ . As per the second derivative test,  $r = 1.12725 \text{ m}$  corresponds to a minimum.

## DIFFERENTIATION

Topic	Primer on Differentiation
Summary	These are textbook notes of a primer on differentiation
Major	General Engineering
Authors	Autar Kaw, Luke Snyder
Date	July 17, 2008
Web Site	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

## Chapter 02.02

# Differentiation of Continuous Functions

*After reading this chapter, you should be able to:*

1. derive formulas for approximating the first derivative of a function,
2. derive formulas for approximating derivatives from Taylor series,
3. derive finite difference approximations for higher order derivatives, and
4. use the developed formulas in examples to find derivatives of a function.

The derivative of a function at  $x$  is defined as

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

To be able to find a derivative numerically, one could make  $\Delta x$  finite to give,

$$f'(x) \approx \frac{f(x + \Delta x) - f(x)}{\Delta x}.$$

Knowing the value of  $x$  at which you want to find the derivative of  $f(x)$ , we choose a value of  $\Delta x$  to find the value of  $f'(x)$ . To estimate the value of  $f'(x)$ , three such approximations are suggested as follows.

### Forward Difference Approximation of the First Derivative

From differential calculus, we know

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

For a finite  $\Delta x$ ,

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

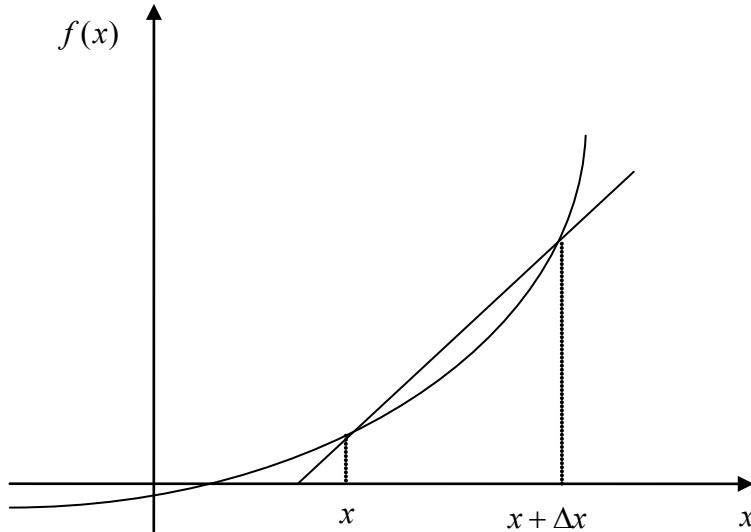
The above is the forward divided difference approximation of the first derivative. It is called forward because you are taking a point ahead of  $x$ . To find the value of  $f'(x)$  at  $x = x_i$ , we may choose another point  $\Delta x$  ahead as  $x = x_{i+1}$ . This gives

$$f'(x_i) \approx \frac{f(x_{i+1}) - f(x_i)}{\Delta x}$$

$$= \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i}$$

where

$$\Delta x = x_{i+1} - x_i$$



**Figure 1** Graphical representation of forward difference approximation of first derivative.

### Example 1

The velocity of a rocket is given by

$$v(t) = 2000 \ln \left[ \frac{14 \times 10^4}{14 \times 10^4 - 2100t} \right] - 9.8t, \quad 0 \leq t \leq 30$$

where  $v$  is given in m/s and  $t$  is given in seconds. At  $t = 16$ s,

- a) use the forward difference approximation of the first derivative of  $v(t)$  to calculate the acceleration. Use a step size of  $\Delta t = 2$ s.
- b) find the exact value of the acceleration of the rocket.
- c) calculate the absolute relative true error for part (b).

### Solution

$$(a) \quad a(t_i) \approx \frac{v(t_{i+1}) - v(t_i)}{\Delta t}$$

$$t_i = 16$$

$$\Delta t = 2$$

$$t_{i+1} = t_i + \Delta t$$

$$= 16 + 2$$

$$= 18$$

$$\begin{aligned}
 a(16) &\approx \frac{v(18) - v(16)}{2} \\
 v(18) &= 2000 \ln \left[ \frac{14 \times 10^4}{14 \times 10^4 - 2100(18)} \right] - 9.8(18) \\
 &= 453.02 \text{ m/s} \\
 v(16) &= 2000 \ln \left[ \frac{14 \times 10^4}{14 \times 10^4 - 2100(16)} \right] - 9.8(16) \\
 &= 392.07 \text{ m/s}
 \end{aligned}$$

Hence

$$\begin{aligned}
 a(16) &\approx \frac{v(18) - v(16)}{2} \\
 &= \frac{453.02 - 392.07}{2} \\
 &= 30.474 \text{ m/s}^2
 \end{aligned}$$

(b) The exact value of  $a(16)$  can be calculated by differentiating

$$v(t) = 2000 \ln \left[ \frac{14 \times 10^4}{14 \times 10^4 - 2100t} \right] - 9.8t$$

as

$$a(t) = \frac{d}{dt}[v(t)]$$

Knowing that

$$\begin{aligned}
 \frac{d}{dt}[\ln(t)] &= \frac{1}{t} \quad \text{and} \quad \frac{d}{dt}\left[\frac{1}{t}\right] = -\frac{1}{t^2} \\
 a(t) &= 2000 \left( \frac{14 \times 10^4 - 2100t}{14 \times 10^4} \right) \frac{d}{dt} \left( \frac{14 \times 10^4}{14 \times 10^4 - 2100t} \right) - 9.8 \\
 &= 2000 \left( \frac{14 \times 10^4 - 2100t}{14 \times 10^4} \right) \left( -1 \right) \left( \frac{14 \times 10^4}{(14 \times 10^4 - 2100t)^2} \right) (-2100) - 9.8 \\
 &= \frac{-4040 - 29.4t}{-200 + 3t} \\
 a(16) &= \frac{-4040 - 29.4(16)}{-200 + 3(16)} \\
 &= 29.674 \text{ m/s}^2
 \end{aligned}$$

(c) The absolute relative true error is

$$|\epsilon_t| = \left| \frac{\text{True Value} - \text{Approximate Value}}{\text{True Value}} \right| \times 100$$

$$= \left| \frac{29.674 - 30.474}{29.674} \right| \times 100 \\ = 2.6967\%$$

### Backward Difference Approximation of the First Derivative

We know

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

For a finite  $\Delta x$ ,

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

If  $\Delta x$  is chosen as a negative number,

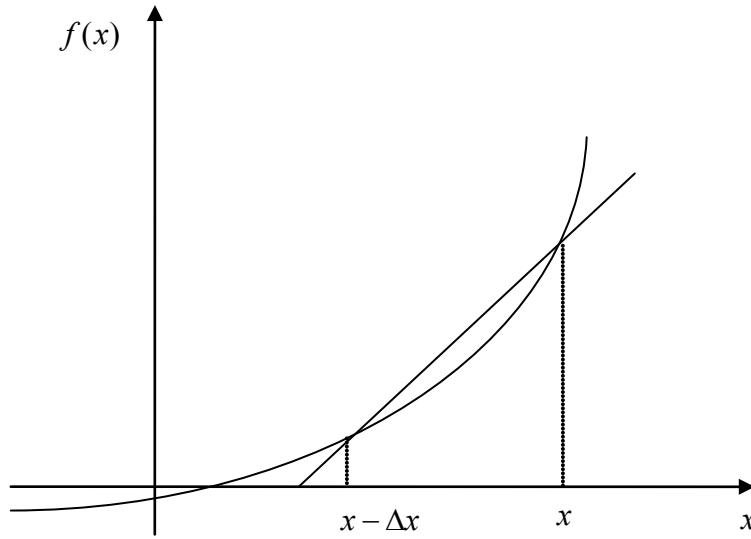
$$f'(x) \approx \frac{f(x + \Delta x) - f(x)}{\Delta x} \\ = \frac{f(x) - f(x - \Delta x)}{\Delta x}$$

This is a backward difference approximation as you are taking a point backward from  $x$ . To find the value of  $f'(x)$  at  $x = x_i$ , we may choose another point  $\Delta x$  behind as  $x = x_{i-1}$ . This gives

$$f'(x_i) \approx \frac{f(x_i) - f(x_{i-1})}{\Delta x} \\ = \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}}$$

where

$$\Delta x = x_i - x_{i-1}$$



**Figure 2** Graphical representation of backward difference approximation of first derivative.

### Example 2

The velocity of a rocket is given by

$$v(t) = 2000 \ln \left[ \frac{14 \times 10^4}{14 \times 10^4 - 2100t} \right] - 9.8t, \quad 0 \leq t \leq 30$$

(a) Use the backward difference approximation of the first derivative of  $v(t)$  to calculate the acceleration at  $t = 16\text{s}$ . Use a step size of  $\Delta t = 2\text{s}$ .

(b) Find the absolute relative true error for part (a).

### Solution

$$a(t) \approx \frac{v(t_i) - v(t_{i-1})}{\Delta t}$$

$$t_i = 16$$

$$\Delta t = 2$$

$$\begin{aligned} t_{i-1} &= t_i - \Delta t \\ &= 16 - 2 \\ &= 14 \end{aligned}$$

$$a(16) \approx \frac{v(16) - v(14)}{2}$$

$$\begin{aligned} v(16) &= 2000 \ln \left[ \frac{14 \times 10^4}{14 \times 10^4 - 2100(16)} \right] - 9.8(16) \\ &= 392.07 \text{ m/s} \end{aligned}$$

$$v(14) = 2000 \ln \left[ \frac{14 \times 10^4}{14 \times 10^4 - 2100(14)} \right] - 9.8(14)$$

$$= 334.24 \text{ m/s}$$

$$\begin{aligned} a(16) &\approx \frac{v(16) - v(14)}{2} \\ &= \frac{392.07 - 334.24}{2} \\ &= 28.915 \text{ m/s}^2 \end{aligned}$$

(b) The exact value of the acceleration at  $t = 16\text{s}$  from Example 1 is

$$a(16) = 29.674 \text{ m/s}^2$$

The absolute relative true error for the answer in part (a) is

$$\begin{aligned} |\epsilon_t| &= \left| \frac{29.674 - 28.915}{29.674} \right| \times 100 \\ &= 2.5584\% \end{aligned}$$

### Forward Difference Approximation from Taylor Series

Taylor's theorem says that if you know the value of a function  $f(x)$  at a point  $x_i$  and all its derivatives at that point, provided the derivatives are continuous between  $x_i$  and  $x_{i+1}$ , then

$$f(x_{i+1}) = f(x_i) + f'(x_i)(x_{i+1} - x_i) + \frac{f''(x_i)}{2!}(x_{i+1} - x_i)^2 + \dots$$

Substituting for convenience  $\Delta x = x_{i+1} - x_i$

$$f(x_{i+1}) = f(x_i) + f'(x_i)\Delta x + \frac{f''(x_i)}{2!}(\Delta x)^2 + \dots$$

$$f'(x_i) = \frac{f(x_{i+1}) - f(x_i)}{\Delta x} - \frac{f''(x_i)}{2!}(\Delta x) + \dots$$

$$f'(x_i) = \frac{f(x_{i+1}) - f(x_i)}{\Delta x} + O(\Delta x)$$

The  $O(\Delta x)$  term shows that the error in the approximation is of the order of  $\Delta x$ .

Can you now derive from the Taylor series the formula for the backward divided difference approximation of the first derivative?

As you can see, both forward and backward divided difference approximations of the first derivative are accurate on the order of  $O(\Delta x)$ . Can we get better approximations? Yes, another method to approximate the first derivative is called the **central difference approximation of the first derivative**.

From the Taylor series

$$f(x_{i+1}) = f(x_i) + f'(x_i)\Delta x + \frac{f''(x_i)}{2!}(\Delta x)^2 + \frac{f'''(x_i)}{3!}(\Delta x)^3 + \dots \quad (1)$$

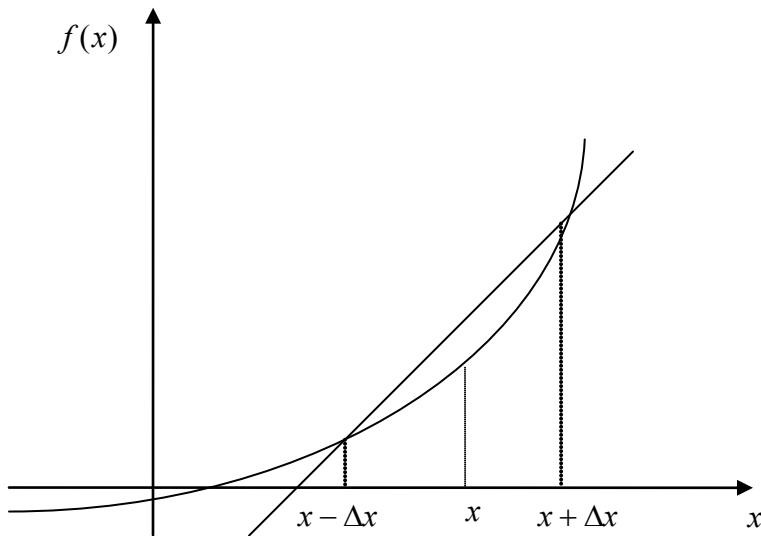
and

$$f(x_{i-1}) = f(x_i) - f'(x_i)\Delta x + \frac{f''(x_i)}{2!}(\Delta x)^2 - \frac{f'''(x_i)}{3!}(\Delta x)^3 + \dots \quad (2)$$

Subtracting Equation (2) from Equation (1)

$$\begin{aligned}
 f(x_{i+1}) - f(x_{i-1}) &= f'(x_i)(2\Delta x) + \frac{2f'''(x_i)}{3!}(\Delta x)^3 + \dots \\
 f'(x_i) &= \frac{f(x_{i+1}) - f(x_{i-1})}{2\Delta x} - \frac{f'''(x_i)}{3!}(\Delta x)^2 + \dots \\
 &= \frac{f(x_{i+1}) - f(x_{i-1})}{2\Delta x} + O(\Delta x)^2
 \end{aligned}$$

hence showing that we have obtained a more accurate formula as the error is of the order of  $O(\Delta x)^2$ .



**Figure 3** Graphical representation of central difference approximation of first derivative.

### Example 3

The velocity of a rocket is given by

$$v(t) = 2000 \ln \left[ \frac{14 \times 10^4}{14 \times 10^4 - 2100t} \right] - 9.8t, \quad 0 \leq t \leq 30.$$

- (a) Use the central difference approximation of the first derivative of  $v(t)$  to calculate the acceleration at  $t = 16$  s. Use a step size of  $\Delta t = 2$  s.
- (b) Find the absolute relative true error for part (a).

### Solution

$$a(t_i) \approx \frac{v(t_{i+1}) - v(t_{i-1})}{2\Delta t}$$

$$t_i = 16$$

$$\Delta t = 2$$

$$\begin{aligned}
t_{i+1} &= t_i + \Delta t \\
&= 16 + 2 \\
&= 18 \\
t_{i-1} &= t_i - \Delta t \\
&= 16 - 2 \\
&= 14 \\
a(16) &\approx \frac{v(18) - v(14)}{2(2)} \\
&= \frac{v(18) - v(14)}{4} \\
v(18) &= 2000 \ln \left[ \frac{14 \times 10^4}{14 \times 10^4 - 2100(18)} \right] - 9.8(18) \\
&= 453.02 \text{ m/s} \\
v(14) &= 2000 \ln \left[ \frac{14 \times 10^4}{14 \times 10^4 - 2100(14)} \right] - 9.8(14) \\
&= 334.24 \text{ m/s} \\
a(16) &\approx \frac{v(18) - v(14)}{4} \\
&= \frac{453.02 - 334.24}{4} \\
&= 29.694 \text{ m/s}^2
\end{aligned}$$

(b) The exact value of the acceleration at  $t = 16 \text{ s}$  from Example 1 is

$$a(16) = 29.674 \text{ m/s}^2$$

The absolute relative true error for the answer in part (a) is

$$\begin{aligned}
|\epsilon_t| &= \left| \frac{29.674 - 29.694}{29.674} \right| \times 100 \\
&= 0.069157\%
\end{aligned}$$

The results from the three difference approximations are given in Table 1.

**Table 1** Summary of  $a(16)$  using different difference approximations

Type of difference approximation	$a(16)$ ( $\text{m/s}^2$ )	$ \epsilon_t  \%$
Forward	30.475	2.6967
Backward	28.915	2.5584
Central	29.695	0.069157

Clearly, the central difference scheme is giving more accurate results because the order of accuracy is proportional to the square of the step size. In real life, one would not

know the exact value of the derivative – so how would one know how accurately they have found the value of the derivative? A simple way would be to start with a step size and keep on halving the step size until the absolute relative approximate error is within a pre-specified tolerance.

Take the example of finding  $v'(t)$  for

$$v(t) = 2000 \ln \left[ \frac{14 \times 10^4}{14 \times 10^4 - 2100t} \right] - 9.8t$$

at  $t = 16$  using the backward difference scheme. Given in Table 2 are the values obtained using the backward difference approximation method and the corresponding absolute relative approximate errors.

**Table 2** First derivative approximations and relative errors for different  $\Delta t$  values of backward difference scheme.

$\Delta t$	$v'(t)$	$ e_a  \%$
2	28.915	
1	29.289	1.2792
0.5	29.480	0.64787
0.25	29.577	0.32604
0.125	29.625	0.16355

From the above table, one can see that the absolute relative approximate error decreases as the step size is reduced. At  $\Delta t = 0.125$ , the absolute relative approximate error is 0.16355%, meaning that at least 2 significant digits are correct in the answer.

### Finite Difference Approximation of Higher Derivatives

One can also use the Taylor series to approximate a higher order derivative. For example, to approximate  $f''(x)$ , the Taylor series is

$$f(x_{i+2}) = f(x_i) + f'(x_i)(2\Delta x) + \frac{f''(x_i)}{2!}(2\Delta x)^2 + \frac{f'''(x_i)}{3!}(2\Delta x)^3 + \dots \quad (3)$$

where

$$\begin{aligned} x_{i+2} &= x_i + 2\Delta x \\ f(x_{i+1}) &= f(x_i) + f'(x_i)(\Delta x) + \frac{f''(x_i)}{2!}(\Delta x)^2 + \frac{f'''(x_i)}{3!}(\Delta x)^3 \dots \end{aligned} \quad (4)$$

where

$$x_{i-1} = x_i - \Delta x$$

Subtracting 2 times Equation (4) from Equation (3) gives

$$f(x_{i+2}) - 2f(x_{i+1}) = -f(x_i) + f''(x_i)(\Delta x)^2 + f'''(x_i)(\Delta x)^3 \dots$$

$$\begin{aligned} f''(x_i) &= \frac{f(x_{i+2}) - 2f(x_{i+1}) + f(x_i)}{(\Delta x)^2} - f'''(x_i)(\Delta x) + \dots \\ f''(x_i) &\approx \frac{f(x_{i+2}) - 2f(x_{i+1}) + f(x_i)}{(\Delta x)^2} + O(\Delta x) \end{aligned} \quad (5)$$

**Example 4**

The velocity of a rocket is given by

$$v(t) = 2000 \ln \left[ \frac{14 \times 10^4}{14 \times 10^4 - 2100t} \right] - 9.8t, \quad 0 \leq t \leq 30$$

Use the forward difference approximation of the second derivative of  $v(t)$  to calculate the jerk at  $t = 16$  s. Use a step size of  $\Delta t = 2$  s.

**Solution**

$$j(t_i) \approx \frac{v(t_{i+2}) - 2v(t_{i+1}) + v(t_i)}{(\Delta t)^2}$$

$$t_i = 16$$

$$\Delta t = 2$$

$$t_{i+1} = t_i + \Delta t$$

$$= 16 + 2$$

$$= 18$$

$$t_{i+2} = t_i + 2(\Delta t)$$

$$= 16 + 2(2)$$

$$= 20$$

$$j(16) \approx \frac{v(20) - 2v(18) + v(16)}{(2)^2}$$

$$\begin{aligned} v(20) &= 2000 \ln \left[ \frac{14 \times 10^4}{14 \times 10^4 - 2100(20)} \right] - 9.8(20) \\ &= 517.35 \text{ m/s} \end{aligned}$$

$$\begin{aligned} v(18) &= 2000 \ln \left[ \frac{14 \times 10^4}{14 \times 10^4 - 2100(18)} \right] - 9.8(18) \\ &= 453.02 \text{ m/s} \end{aligned}$$

$$\begin{aligned} v(16) &= 2000 \ln \left[ \frac{14 \times 10^4}{14 \times 10^4 - 2100(16)} \right] - 9.8(16) \\ &= 392.07 \text{ m/s} \end{aligned}$$

$$\begin{aligned} j(16) &\approx \frac{517.35 - 2(453.02) + 392.07}{4} \\ &= 0.84515 \text{ m/s}^3 \end{aligned}$$

The exact value of  $j(16)$  can be calculated by differentiating

$$v(t) = 2000 \ln \left[ \frac{14 \times 10^4}{14 \times 10^4 - 2100t} \right] - 9.8t$$

twice as

$$a(t) = \frac{d}{dt}[v(t)] \text{ and}$$

$$j(t) = \frac{d}{dt}[a(t)]$$

Knowing that

$$\frac{d}{dt}[\ln(t)] = \frac{1}{t} \text{ and}$$

$$\frac{d}{dt}\left[\frac{1}{t}\right] = -\frac{1}{t^2}$$

$$\begin{aligned} a(t) &= 2000 \left( \frac{14 \times 10^4 - 2100t}{14 \times 10^4} \right) \frac{d}{dt} \left( \frac{14 \times 10^4}{14 \times 10^4 - 2100t} \right) - 9.8 \\ &= 2000 \left( \frac{14 \times 10^4 - 2100t}{14 \times 10^4} \right) \left( -1 \right) \left( \frac{14 \times 10^4}{(14 \times 10^4 - 2100t)^2} \right) (-2100) - 9.8 \\ &= \frac{-4040 - 29.4t}{-200 + 3t} \end{aligned}$$

Similarly it can be shown that

$$\begin{aligned} j(t) &= \frac{d}{dt}[a(t)] \\ &= \frac{18000}{(-200 + 3t)^2} \end{aligned}$$

$$\begin{aligned} j(16) &= \frac{18000}{[-200 + 3(16)]^2} \\ &= 0.77909 \text{ m/s}^3 \end{aligned}$$

The absolute relative true error is

$$\begin{aligned} |\epsilon_r| &= \left| \frac{0.77909 - 0.84515}{0.77909} \right| \times 100 \\ &= 8.4797\% \end{aligned}$$

The formula given by Equation (5) is a forward difference approximation of the second derivative and has an error of the order of  $O(\Delta x)$ . Can we get a formula that has a better accuracy? Yes, we can derive the central difference approximation of the second derivative.

The Taylor series is

$$f(x_{i+1}) = f(x_i) + f'(x_i)\Delta x + \frac{f''(x_i)}{2!}(\Delta x)^2 + \frac{f'''(x_i)}{3!}(\Delta x)^3 + \frac{f''''(x_i)}{4!}(\Delta x)^4 + \dots \quad (6)$$

where

$$\begin{aligned} x_{i+1} &= x_i + \Delta x \\ f(x_{i-1}) &= f(x_i) - f'(x_i)\Delta x + \frac{f''(x_i)}{2!}(\Delta x)^2 - \frac{f'''(x_i)}{3!}(\Delta x)^3 + \frac{f''''(x_i)}{4!}(\Delta x)^4 - \dots \end{aligned} \quad (7)$$

where

$$x_{i-1} = x_i - \Delta x$$

Adding Equations (6) and (7), gives

$$\begin{aligned} f(x_{i+1}) + f(x_{i-1}) &= 2f(x_i) + f''(x_i)(\Delta x)^2 + f'''(x_i)\frac{(\Delta x)^4}{12} + \dots \\ f''(x_i) &= \frac{f(x_{i+1}) - 2f(x_i) + f(x_{i-1})}{(\Delta x)^2} - \frac{f''''(x_i)(\Delta x)^2}{12} + \dots \\ &= \frac{f(x_{i+1}) - 2f(x_i) + f(x_{i-1})}{(\Delta x)^2} + O(\Delta x)^2 \end{aligned}$$

### Example 5

The velocity of a rocket is given by

$$v(t) = 2000 \ln \left[ \frac{14 \times 10^4}{14 \times 10^4 - 2100t} \right] - 9.8t, \quad 0 \leq t \leq 30,$$

- (a) Use the central difference approximation of the second derivative of  $v(t)$  to calculate the jerk at  $t = 16$  s. Use a step size of  $\Delta t = 2$  s.

### Solution

The second derivative of velocity with respect to time is called jerk. The second order approximation of jerk then is

$$j(t_i) \approx \frac{v(t_{i+1}) - 2v(t_i) + v(t_{i-1})}{(\Delta t)^2}$$

$$t_i = 16$$

$$\Delta t = 2$$

$$t_{i+1} = t_i + \Delta t$$

$$= 16 + 2$$

$$= 18$$

$$t_{i+2} = t_i - \Delta t$$

$$= 16 - 2$$

$$= 14$$

$$j(16) \approx \frac{v(18) - 2v(16) + v(14)}{(2)^2}$$

$$v(18) = 2000 \ln \left[ \frac{14 \times 10^4}{14 \times 10^4 - 2100(18)} \right] - 9.8(18)$$

$$= 453.02 \text{ m/s}$$

$$v(16) = 2000 \ln \left[ \frac{14 \times 10^4}{14 \times 10^4 - 2100(16)} \right] - 9.8(16)$$

$$= 392.07 \text{ m/s}$$

$$v(14) = 2000 \ln \left[ \frac{14 \times 10^4}{14 \times 10^4 - 2100(14)} \right] - 9.8(14)$$

$$= 334.24 \text{ m/s}$$

$$j(16) \approx \frac{v(18) - 2v(16) + v(14)}{(2)^2}$$

$$= \frac{453.02 - 2(392.07) + 334.24}{4}$$

$$= 0.77969 \text{ m/s}^3$$

The absolute relative true error is

$$|\epsilon_t| = \left| \frac{0.77908 - 0.77969}{0.77908} \right| \times 100$$

$$= 0.077992\%$$

## DIFFERENTIATION

Topic	Differentiation of Continuous functions
Summary	These are textbook notes of differentiation of continuous functions
Major	General Engineering
Authors	Autar Kaw, Luke Snyder
Date	February 2, 2012
Web Site	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

## Chapter 02.03

# Differentiation of Discrete Functions

*After reading this chapter, you should be able to:*

1. *find approximate values of the first derivative of functions that are given at discrete data points, and*
2. *use Lagrange polynomial interpolation to find derivatives of discrete functions.*

To find the derivatives of functions that are given at discrete points, several methods are available. Although these methods are mainly used when the data is spaced unequally, they can be used for data that is spaced equally as well.

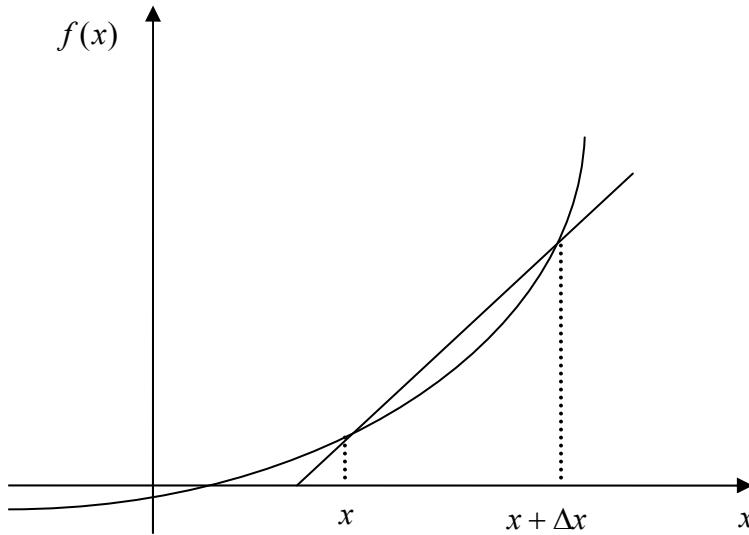
### Forward Difference Approximation of the First Derivative

We know

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

For a finite  $\Delta x$ ,

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



**Figure 1** Graphical representation of forward difference approximation of first derivative.

So given  $n+1$  data points  $(x_0, y_0), (x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ , the value of  $f'(x)$  for  $x_i \leq x \leq x_{i+1}$ ,  $i = 0, \dots, n-1$ , is given by

$$f'(x_i) \approx \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i}$$

### Example 1

The upward velocity of a rocket is given as a function of time in Table 1.

**Table 1** Velocity as a function of time.

$t$ (s)	$v(t)$ (m/s)
0	0
10	227.04
15	362.78
20	517.35
22.5	602.97
30	901.67

Using forward divided difference, find the acceleration of the rocket at  $t = 16$  s.

### Solution

To find the acceleration at  $t = 16$  s, we need to choose the two values of velocity closest to  $t = 16$  s, that also bracket  $t = 16$  s to evaluate it. The two points are  $t = 15$  s and  $t = 20$  s

$$\begin{aligned}
 a(t_i) &\approx \frac{v(t_{i+1}) - v(t_i)}{\Delta t} \\
 t_i &= 15 \\
 t_{i+1} &= 20 \\
 \Delta t &= t_{i+1} - t_i \\
 &= 20 - 15 \\
 &= 5 \\
 a(16) &\approx \frac{v(20) - v(15)}{5} \\
 &= \frac{517.35 - 362.78}{5} \\
 &= 30.914 \text{ m/s}^2
 \end{aligned}$$

### Direct Fit Polynomials

In this method, given  $n+1$  data points  $(x_0, y_0), (x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ , one can fit a  $n^{\text{th}}$  order polynomial given by

$$P_n(x) = a_0 + a_1 x + \dots + a_{n-1} x^{n-1} + a_n x^n$$

To find the first derivative,

$$P'_n(x) = \frac{dP_n(x)}{dx} = a_1 + 2a_2 x + \dots + (n-1)a_{n-1} x^{n-2} + n a_n x^{n-1}$$

Similarly, other derivatives can also be found.

### Example 2

The upward velocity of a rocket is given as a function of time in Table 2.

**Table 2** Velocity as a function of time.

$t$ (s)	$v(t)$ (m/s)
0	0
10	227.04
15	362.78
20	517.35
22.5	602.97
30	901.67

Using a third order polynomial interpolant for velocity, find the acceleration of the rocket at  $t = 16 \text{ s}$ .

### Solution

For the third order polynomial (also called cubic interpolation), we choose the velocity given by

$$v(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3$$

Since we want to find the velocity at  $t = 16\text{ s}$ , and we are using a third order polynomial, we need to choose the four points closest to  $t = 16$  and that also bracket  $t = 16$  to evaluate it. The four points are  $t_0 = 10$ ,  $t_1 = 15$ ,  $t_2 = 20$  and  $t_3 = 22.5$ .

$$t_0 = 10, \quad v(t_0) = 227.04$$

$$t_1 = 15, \quad v(t_1) = 362.78$$

$$t_2 = 20, \quad v(t_2) = 517.35$$

$$t_3 = 22.5, \quad v(t_3) = 602.97$$

such that

$$v(10) = 227.04 = a_0 + a_1(10) + a_2(10)^2 + a_3(10)^3$$

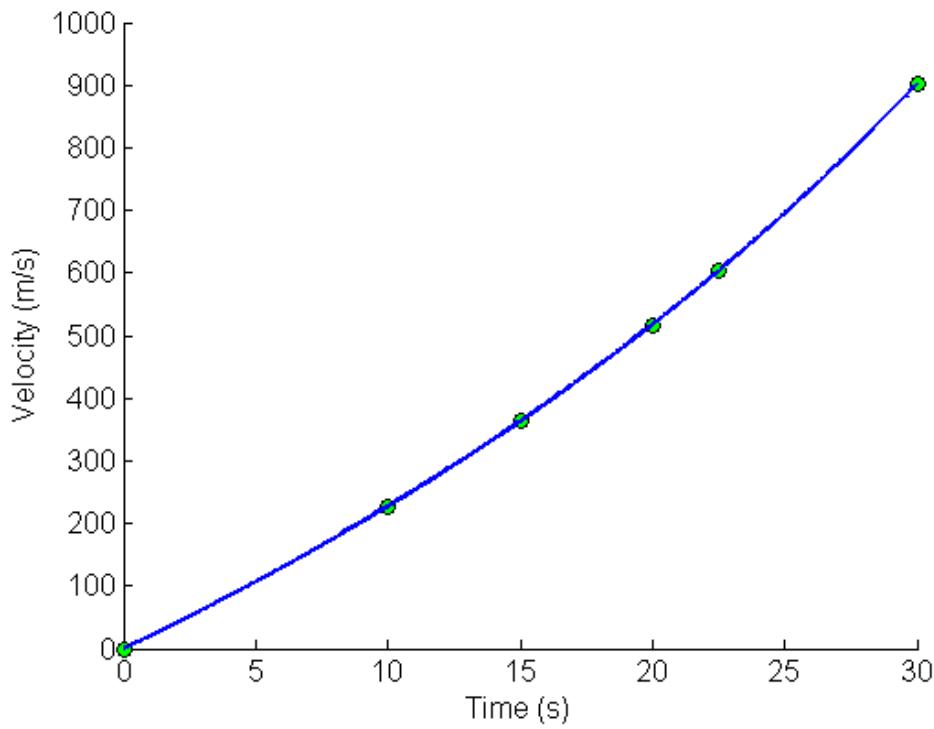
$$v(15) = 362.78 = a_0 + a_1(15) + a_2(15)^2 + a_3(15)^3$$

$$v(20) = 517.35 = a_0 + a_1(20) + a_2(20)^2 + a_3(20)^3$$

$$v(22.5) = 602.97 = a_0 + a_1(22.5) + a_2(22.5)^2 + a_3(22.5)^3$$

Writing the four equations in matrix form, we have

$$\begin{bmatrix} 1 & 10 & 100 & 1000 \\ 1 & 15 & 225 & 3375 \\ 1 & 20 & 400 & 8000 \\ 1 & 22.5 & 506.25 & 11391 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 227.04 \\ 362.78 \\ 517.35 \\ 602.97 \end{bmatrix}$$



**Figure 2** Graph of upward velocity of the rocket vs. time.

Solving the above four equations gives

$$a_0 = -4.3810$$

$$a_1 = 21.289$$

$$a_2 = 0.13065$$

$$a_3 = 0.0054606$$

Hence

$$\begin{aligned} v(t) &= a_0 + a_1 t + a_2 t^2 + a_3 t^3 \\ &= -4.3810 + 21.289t + 0.13065t^2 + 0.0054606t^3, \quad 10 \leq t \leq 22.5 \end{aligned}$$

The acceleration at  $t = 16$  is given by

$$a(16) = \frac{d}{dt} v(t) \Big|_{t=16}$$

Given that  $v(t) = -4.3810 + 21.289t + 0.13065t^2 + 0.0054606t^3, \quad 10 \leq t \leq 22.5$ ,

$$\begin{aligned} a(t) &= \frac{d}{dt} v(t) \\ &= \frac{d}{dt} (-4.3810 + 21.289t + 0.13065t^2 + 0.0054606t^3) \\ &= 21.289 + 0.26130t + 0.016382t^2, \quad 10 \leq t \leq 22.5 \end{aligned}$$

$$\begin{aligned} a(16) &= 21.289 + 0.26130(16) + 0.016382(16)^2 \\ &= 29.664 \text{ m/s}^2 \end{aligned}$$

### Lagrange Polynomial

In this method, given  $(x_0, y_0), \dots, (x_n, y_n)$ , one can fit a  $n^{\text{th}}$  order Lagrangian polynomial given by

$$f_n(x) = \sum_{i=0}^n L_i(x) f(x_i)$$

where  $n$  in  $f_n(x)$  stands for the  $n^{\text{th}}$  order polynomial that approximates the function  $y = f(x)$  and

$$L_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^n \frac{x - x_j}{x_i - x_j}$$

$L_i(x)$  is a weighting function that includes a product of  $n-1$  terms with terms of  $j = i$  omitted.

Then to find the first derivative, one can differentiate  $f_n(x)$  once, and so on for other derivatives.

For example, the second order Lagrange polynomial passing through  $(x_0, y_0), (x_1, y_1)$ , and  $(x_2, y_2)$  is

$$f_2(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} f(x_0) + \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} f(x_1) + \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} f(x_2)$$

Differentiating the above equation gives

$$f_2'(x) = \frac{2x-(x_1+x_2)}{(x_0-x_1)(x_0-x_2)} f(x_0) + \frac{2x-(x_0+x_2)}{(x_1-x_0)(x_1-x_2)} f(x_1) + \frac{2x-(x_0+x_1)}{(x_2-x_0)(x_2-x_1)} f(x_2)$$

Differentiating again would give the second derivative as

$$f_2''(x) = \frac{2}{(x_0-x_1)(x_0-x_2)} f(x_0) + \frac{2}{(x_1-x_0)(x_1-x_2)} f(x_1) + \frac{2}{(x_2-x_0)(x_2-x_1)} f(x_2)$$

### Example 3

The upward velocity of a rocket is given as a function of time in Table 3.

**Table 3** Velocity as a function of time.

$t$ (s)	$v(t)$ (m/s)
0	0
10	227.04
15	362.78
20	517.35
22.5	602.97
30	901.67

Determine the value of the acceleration at  $t = 16\text{ s}$  using second order Lagrangian polynomial interpolation for velocity.

### Solution

$$v(t) = \left( \frac{t-t_1}{t_0-t_1} \right) \left( \frac{t-t_2}{t_0-t_2} \right) v(t_0) + \left( \frac{t-t_0}{t_1-t_0} \right) \left( \frac{t-t_2}{t_1-t_2} \right) v(t_1) + \left( \frac{t-t_0}{t_2-t_0} \right) \left( \frac{t-t_1}{t_2-t_1} \right) v(t_2)$$

$$a(t) = \frac{2t-(t_1+t_2)}{(t_0-t_1)(t_0-t_2)} v(t_0) + \frac{2t-(t_0+t_2)}{(t_1-t_0)(t_1-t_2)} v(t_1) + \frac{2t-(t_0+t_1)}{(t_2-t_0)(t_2-t_1)} v(t_2)$$

$$a(16) = \frac{2(16)-(15+20)}{(10-15)(10-20)} (227.04) + \frac{2(16)-(10+20)}{(15-10)(15-20)} (362.78)$$

$$+ \frac{2(16)-(10+15)}{(20-10)(20-15)} (517.35)$$

$$= -0.06(227.04) - 0.08(362.78) + 0.14(517.35)$$

$$= 29.784 \text{ m/s}^2$$

---

## DIFFERENTIATION

---

Topic      Differentiation of Discrete Functions  
Summary    These are textbook notes differentiation of discrete functions  
Major      General Engineering  
Authors     Autar Kaw, Luke Snyder  
Date       December 23, 2009  

---

Web Site    <http://numericalmethods.eng.usf.edu>

---

# Chapter 03.01

## Solution of Quadratic Equations

After reading this chapter, you should be able to:

1. find the solutions of quadratic equations,
2. derive the formula for the solution of quadratic equations,
3. solve simple physical problems involving quadratic equations.

### What are quadratic equations and how do we solve them?

A quadratic equation has the form

$$ax^2 + bx + c = 0, \text{ where } a \neq 0$$

The solution to the above quadratic equation is given by

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

So the equation has two roots, and depending on the value of the discriminant,  $b^2 - 4ac$ , the equation may have real, complex or repeated roots.

If  $b^2 - 4ac < 0$ , the roots are complex.

If  $b^2 - 4ac > 0$ , the roots are real.

If  $b^2 - 4ac = 0$ , the roots are real and repeated.

### Example 1

Derive the solution to  $ax^2 + bx + c = 0$ .

#### Solution

$$ax^2 + bx + c = 0$$

Dividing both sides by  $a$ , ( $a \neq 0$ ), we get

$$x^2 + \frac{b}{a}x + \frac{c}{a} = 0$$

Note if  $a = 0$ , the solution to

$$ax^2 + bx + c = 0$$

is

$$x = -\frac{c}{b}$$

Rewrite

$$x^2 + \frac{b}{a}x + \frac{c}{a} = 0$$

as

$$\left(x + \frac{b}{2a}\right)^2 - \frac{b^2}{4a^2} + \frac{c}{a} = 0$$

$$\begin{aligned} \left(x + \frac{b}{2a}\right)^2 &= \frac{b^2}{4a^2} - \frac{c}{a} \\ &= \frac{b^2 - 4ac}{4a^2} \end{aligned}$$

$$\begin{aligned} x + \frac{b}{2a} &= \pm \sqrt{\frac{b^2 - 4ac}{4a^2}} \\ &= \pm \frac{\sqrt{b^2 - 4ac}}{2a} \end{aligned}$$

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

### Example 2

A ball is thrown down at 50 mph from the top of a building. The building is 420 feet tall. Derive the equation that would let you find the time the ball takes to reach the ground.

#### Solution

The distance  $s$  covered by the ball is given by

$$s = ut + \frac{1}{2}gt^2$$

where

$u$  = initial velocity (ft/s)

$g$  = acceleration due to gravity (ft/s<sup>2</sup>)

$t$  = time (s)

Given

$$u = 50 \frac{\text{miles}}{\text{hour}} \times \frac{1 \text{ hour}}{3600 \text{ s}} \times \frac{5280 \text{ ft}}{1 \text{ mile}}$$

$$= 73.33 \frac{\text{ft}}{\text{s}}$$

$$g = 32.2 \frac{\text{ft}}{\text{s}^2}$$

$$s = 420 \text{ ft}$$

we have

$$420 = 73.33t + \frac{1}{2}(32.2)t^2$$

$$16.1t^2 + 73.33t - 420 = 0$$

The above equation is a quadratic equation, the solution of which would give the time it would take the ball to reach the ground. The solution of the quadratic equation is

$$t = \frac{-73.33 \pm \sqrt{73.33^2 - 4 \times 16.1 \times (-420)}}{2(16.1)}$$
$$= 3.315, -7.870$$

Since  $t > 0$ , the valid value of time  $t$  is 3.315 s.

---

### NONLINEAR EQUATIONS

---

Topic	Solution of quadratic equations
Summary	Textbook notes on solving quadratic equations
Major	General Engineering
Authors	Autar Kaw
Date	December 23, 2009
Web Site	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

---

## Chapter 03.02

# Solution of Cubic Equations

After reading this chapter, you should be able to:

1. find the exact solution of a general cubic equation.

### How to Find the Exact Solution of a General Cubic Equation

In this chapter, we are going to find the exact solution of a general cubic equation

$$ax^3 + bx^2 + cx + d = 0 \quad (1)$$

To find the roots of Equation (1), we first get rid of the quadratic term ( $x^2$ ) by making the substitution

$$x = y - \frac{b}{3a} \quad (2)$$

to obtain

$$a\left(y - \frac{b}{3a}\right)^3 + b\left(y - \frac{b}{3a}\right)^2 + c\left(y - \frac{b}{3a}\right) + d = 0 \quad (3)$$

Expanding Equation (3) and simplifying, we obtain the following equation

$$ay^3 + \left(c - \frac{b^2}{3a}\right)y + \left(d + \frac{2b^3}{27a^2} - \frac{bc}{3a}\right) = 0 \quad (4)$$

Equation (4) is called the depressed cubic since the quadratic term is absent. Having the equation in this form makes it easier to solve for the roots of the cubic equation (Click [here](#) to know the history behind solving cubic equations exactly).

First, convert the depressed cubic Equation (4) into the form

$$\begin{aligned} y^3 + \frac{1}{a}\left(c - \frac{b^2}{3a}\right)y + \frac{1}{a}\left(d + \frac{2b^3}{27a^2} - \frac{bc}{3a}\right) &= 0 \\ y^3 + ey + f &= 0 \end{aligned} \quad (5)$$

where

$$e = \frac{1}{a}\left(c - \frac{b^2}{3a}\right)$$

$$f = \frac{1}{a} \left( d + \frac{2b^3}{27a^2} - \frac{bc}{3a} \right)$$

Now, reduce the above equation using Vieta's substitution

$$y = z + \frac{s}{z} \quad (6)$$

For the time being, the constant  $s$  is undefined. Substituting into the depressed cubic Equation (5), we get

$$\left( z + \frac{s}{z} \right)^3 + e \left( z + \frac{s}{z} \right) + f = 0 \quad (7)$$

Expanding out and multiplying both sides by  $z^3$ , we get

$$z^6 + (3s + e)z^4 + fz^3 + s(3s + e)z^2 + s^3 = 0 \quad (8)$$

Now, let  $s = -\frac{e}{3}$  ( $s$  is no longer undefined) to simplify the equation into a tri-quadratic equation.

$$z^6 + fz^3 - \frac{e^3}{27} = 0 \quad (9)$$

By making one more substitution,  $w = z^3$ , we now have a general quadratic equation which can be solved using the quadratic formula.

$$w^2 + fw - \frac{e^3}{27} = 0 \quad (10)$$

Once you obtain the solution to this quadratic equation, back substitute using the previous substitutions to obtain the roots to the general cubic equation.

$$w \rightarrow z \rightarrow y \rightarrow x$$

where we assumed

$$w = z^3 \quad (11)$$

$$y = z + \frac{s}{z} \quad (12)$$

$$s = -\frac{e}{3}$$

$$x = y - \frac{b}{3a}$$

Note: You will get two roots for  $w$  as Equation (10) is a quadratic equation. Using Equation (11) would then give you three roots for each of the two roots of  $w$ , hence giving you six root values for  $z$ . But the six root values of  $z$  would give you six values of  $y$  (Equation (6)); but three values of  $y$  will be identical to the other three. So one gets only three values of  $y$ , and hence three values of  $x$ . (Equation (2))

### Example 1

Find the roots of the following cubic equation.

$$x^3 - 9x^2 + 36x - 80 = 0$$

**Solution**

For the general form given by Equation (1)

$$ax^3 + bx^2 + cx + d = 0$$

we have

$$a = 1, b = -9, c = 36, d = -80$$

in

$$x^3 - 9x^2 + 36x - 80 = 0 \quad (\text{E1-1})$$

Equation (E1-1) is reduced to

$$y^3 + ey + f = 0$$

where

$$\begin{aligned} e &= \frac{1}{a} \left( c - \frac{b^2}{3a} \right) \\ &= \frac{1}{1} \left( 36 - \frac{(-9)^2}{3(1)} \right) \\ &= 9 \end{aligned}$$

and

$$\begin{aligned} f &= \frac{1}{a} \left( d + \frac{2b^3}{27a^2} - \frac{bc}{3a} \right) \\ &= \frac{1}{1} \left( -80 + \frac{2(-9)^3}{27(1)^2} - \frac{(-9)(36)}{3(1)} \right) \\ &= -26 \end{aligned}$$

giving

$$y^3 + 9y - 26 = 0 \quad (\text{E1-2})$$

For the general form given by Equation (5)

$$y^3 + ey + f = 0$$

we have

$$e = 9, f = -26$$

in Equation (E1-2).

From Equation (12)

$$\begin{aligned} s &= -\frac{e}{3} \\ &= -\frac{9}{3} \\ &= -3 \end{aligned}$$

From Equation (10)

$$w^2 + fw - \frac{e^3}{27} = 0$$

$$w^2 - 26w - \frac{9^3}{27} = 0$$

$$w^2 - 26w - 27 = 0$$

where

$$w = z^3$$

and

$$\begin{aligned} y &= z + \frac{s}{z} \\ &= z - \frac{3}{z} \\ w &= \frac{-(-26) \pm \sqrt{(-26)^2 - 4(1)(-27)}}{2(1)} \\ &= 27, -1 \end{aligned}$$

The solution is

$$w_1 = 27$$

$$w_2 = -1$$

Since

$$w = z^3$$

$$z^3 = w$$

For  $w = w_1$

$$\begin{aligned} z^3 &= w_1 \\ &= 27 \\ &= 27e^{i0} \end{aligned}$$

Since

$$w = z^3$$

$$re^{i\theta} = (ue^{i\alpha})^3 = u^3 e^{3i\alpha}$$

$$r(\cos\theta + i\sin\theta) = u^3(\cos 3\alpha + i\sin 3\alpha)$$

resulting in

$$r = u^3$$

$$\cos\theta = \cos 3\alpha$$

$$\sin\theta = \sin 3\alpha$$

Since  $\sin\theta$  and  $\cos\theta$  are periodic of  $2\pi$ ,

$$3\alpha = \theta + 2\pi k$$

$$\alpha = \frac{\theta + 2\pi k}{3}$$

$k$  will take the value of 0, 1 and 2 before repeating the same values of  $\alpha$ .

So,

$$\alpha = \frac{\theta + 2\pi k}{3}, k = 0, 1, 2$$

$$\alpha_1 = \frac{\theta}{3}$$

$$\alpha_2 = \frac{(\theta + 2\pi)}{3}$$

$$\alpha_3 = \frac{(\theta + 4\pi)}{3}$$

So roots of  $w = z^3$  are

$$z_1 = r^{\frac{1}{3}} \left( \cos \frac{\theta}{3} + i \sin \frac{\theta}{3} \right)$$

$$z_2 = r^{\frac{1}{3}} \left( \cos \frac{\theta + 2\pi}{3} + i \sin \frac{\theta + 2\pi}{3} \right)$$

$$z_3 = r^{\frac{1}{3}} \left( \cos \frac{\theta + 4\pi}{3} + i \sin \frac{\theta + 4\pi}{3} \right)$$

gives

$$z_1 = (27)^{1/3} \left( \cos \frac{0}{3} + i \sin \frac{0}{3} \right)$$

$$= 3$$

$$z_2 = (27)^{1/3} \left( \cos \frac{0 + 2\pi}{3} + i \sin \frac{0 + 2\pi}{3} \right)$$

$$= 3 \left( \cos \frac{2\pi}{3} + i \sin \frac{2\pi}{3} \right)$$

$$= 3 \left( -\frac{1}{2} + i \frac{\sqrt{3}}{2} \right)$$

$$= -\frac{3}{2} + i \frac{3\sqrt{3}}{2}$$

$$z_3 = (27)^{1/3} \left( \cos \frac{0 + 4\pi}{3} + i \sin \frac{0 + 4\pi}{3} \right)$$

$$= 3 \left( \cos \frac{4\pi}{3} + i \sin \frac{4\pi}{3} \right)$$

$$= 3 \left( -\frac{1}{2} - i \frac{\sqrt{3}}{2} \right)$$

$$= -\frac{3}{2} - i \frac{3\sqrt{3}}{2}$$

Since

$$y = z - \frac{3}{z}$$

$$y_1 = z_1 - \frac{3}{z_1}$$

$$= 3 - \frac{3}{3}$$

$$= 2$$

$$\begin{aligned}
y_2 &= z_2 - \frac{3}{z_2} \\
&= \left( -\frac{3}{2} + i\frac{3\sqrt{3}}{2} \right) - \frac{3}{\left( -\frac{3}{2} + i\frac{3\sqrt{3}}{2} \right)} \\
&= -\frac{5+3i\sqrt{3}}{-1+i\sqrt{3}} \\
&= -\frac{5+3i\sqrt{3}}{-1+i\sqrt{3}} \times \frac{-1-i\sqrt{3}}{-1-i\sqrt{3}} \\
&= -1+i2\sqrt{3} \\
y_3 &= z_3 - \frac{3}{z_3} \\
&= \left( -\frac{3}{2} - i\frac{3\sqrt{3}}{2} \right) - \frac{3}{\left( -\frac{3}{2} - i\frac{3\sqrt{3}}{2} \right)} \\
&= \frac{5-i3\sqrt{3}}{1+i\sqrt{3}} \\
&= \frac{5-i3\sqrt{3}}{1+i\sqrt{3}} \times \frac{1-i\sqrt{3}}{1-i\sqrt{3}} \\
&= -1-i2\sqrt{3}
\end{aligned}$$

Since

$$\begin{aligned}
x &= y + 3 \\
x_1 &= y_1 + 3 \\
&= 2 + 3 \\
&= 5 \\
x_2 &= y_2 + 3 \\
&= (-1+i2\sqrt{3}) + 3 \\
&= 2+i2\sqrt{3} \\
x_3 &= y_3 + 3 \\
&= (-1-i2\sqrt{3}) + 3 \\
&= 2-i2\sqrt{3}
\end{aligned}$$

The roots of the original cubic equation

$$x^3 - 9x^2 + 36x - 80 = 0$$

are  $x_1, x_2$ , and  $x_3$ , that is,

$$5, 2+i2\sqrt{3}, 2-i2\sqrt{3}$$

Verifying

$$(x - 5)(x - (2 + i2\sqrt{3}))(x - (2 - i2\sqrt{3})) = 0$$

gives

$$x^3 - 9x^2 + 36x - 80 = 0$$

Using

$$w_2 = -1$$

would yield the same values of the three roots of the equation. Try it.

### Example 2

Find the roots of the following cubic equation

$$x^3 - 0.03x^2 + 2.4 \times 10^{-6} = 0$$

#### Solution

For the general form

$$ax^3 + bx^2 + cx + d = 0$$

$$a = 1, b = -0.03, c = 0, d = 2.4 \times 10^{-6}$$

Depress the cubic equation by letting (Equation (2))

$$\begin{aligned} x &= y - \frac{b}{3a} \\ &= y - \frac{(-0.03)}{3(1)} \\ &= y + 0.01 \end{aligned}$$

Substituting the above equation into the cubic equation and simplifying, we get

$$y^3 - (3 \times 10^{-4})y + (4 \times 10^{-7}) = 0$$

That gives  $e = -3 \times 10^{-4}$  and  $f = 4 \times 10^{-7}$  for Equation (5), that is,  $y^3 + ey + f = 0$ .

Now, solve the depressed cubic equation by using Vieta's substitution as

$$y = z + \frac{s}{z}$$

to obtain

$$z^6 + (3s - 3 \times 10^{-4})z^4 + (4 \times 10^{-7})z^3 + s(3s - 3 \times 10^{-4})z^2 + s^3 = 0$$

Letting

$$s = -\frac{e}{3} = -\frac{-3 \times 10^{-4}}{3} = 10^{-4}$$

we get the following tri-quadratic equation

$$z^6 + (4 \times 10^{-7})z^3 + 1 \times 10^{-12} = 0$$

Using the following conversion,  $w = z^3$ , we get a general quadratic equation

$$w^2 + (4 \times 10^{-7})w + (1 \times 10^{-12}) = 0$$

Using the quadratic equation, the solutions for  $w$  are

$$w = \frac{-4 \times 10^{-7} \pm \sqrt{(4 \times 10^{-7})^2 - 4(1)(1 \times 10^{-12})}}{2(1)}$$

giving

$$\begin{aligned}w_1 &= -2 \times 10^{-7} + i(9.79795897113 \times 10^{-7}) \\w_2 &= -2 \times 10^{-7} - i(9.79795897113 \times 10^{-7})\end{aligned}$$

Each solution of  $w = z^3$  yields three values of  $z$ . The three values of  $z$  from  $w_1$  are in rectangular form.

Since

$$w = z^3$$

Then

$$z = w^{\frac{1}{3}}$$

Let

$$w = r(\cos \theta + i \sin \theta) = re^{i\theta}$$

then

$$z = u(\cos \alpha + i \sin \alpha) = ue^{i\alpha}$$

This gives

$$\begin{aligned}w &= z^3 \\re^{i\theta} &= (ue^{i\alpha})^3 = u^3 e^{3i\alpha} \\r(\cos \theta + i \sin \theta) &= u^3 (\cos 3\alpha + i \sin 3\alpha)\end{aligned}$$

resulting in

$$r = u^3$$

$$\cos \theta = \cos 3\alpha$$

$$\sin \theta = \sin 3\alpha$$

Since  $\sin \theta$  and  $\cos \theta$  are periodic of  $2\pi$ ,

$$3\alpha = \theta + 2\pi k$$

$$\alpha = \frac{\theta + 2\pi k}{3}$$

$k$  will take the value of 0, 1 and 2 before repeating the same values of  $\alpha$ .

So,

$$\alpha = \frac{\theta + 2\pi k}{3}, k = 0, 1, 2$$

$$\alpha_1 = \frac{\theta}{3}$$

$$\alpha_2 = \frac{(\theta + 2\pi)}{3}$$

$$\alpha_3 = \frac{(\theta + 4\pi)}{3}$$

So the roots of  $w = z^3$  are

$$z_1 = r^{\frac{1}{3}} \left( \cos \frac{\theta}{3} + i \sin \frac{\theta}{3} \right)$$

$$z_2 = r^{\frac{1}{3}} \left( \cos \frac{\theta + 2\pi}{3} + i \sin \frac{\theta + 2\pi}{3} \right)$$

$$z_3 = r^{\frac{1}{3}} \left( \cos \frac{\theta + 4\pi}{3} + i \sin \frac{\theta + 4\pi}{3} \right)$$

So for

$$\begin{aligned} w_1 &= -2 \times 10^{-7} + i(9.79795897113 \times 10^{-7}) \\ r &= \sqrt{(-2 \times 10^{-7})^2 + (9.79795897113 \times 10^{-7})^2} \\ &= 1 \times 10^{-6} \end{aligned}$$

$$\begin{aligned} \theta &= \tan^{-1} \frac{9.79795897113 \times 10^{-7}}{-2 \times 10^{-7}} \\ &= 1.772154248 \text{ (2nd quadrant because } y \text{ (the numerator) is positive and } x \text{ (the denominator) is negative)} \end{aligned}$$

$$z_1 = (1 \times 10^{-6})^{\frac{1}{3}} \left( \cos \frac{1.772154248}{3} + i \sin \frac{1.772154248}{3} \right)$$

$$= 0.008305409517 + i0.005569575635$$

$$\begin{aligned} z_2 &= (1 \times 10^{-6})^{\frac{1}{3}} \left( \cos \frac{1.772154248 + 2\pi}{3} + i \sin \frac{1.772154248 + 2\pi}{3} \right) \\ &= -0.008976098746 + i0.004407907815 \end{aligned}$$

$$\begin{aligned} z_3 &= (1 \times 10^{-6})^{\frac{1}{3}} \left( \cos \frac{1.772154248 + 4\pi}{3} + i \sin \frac{1.772154248 + 4\pi}{3} \right) \\ &= 0.0006706892313 - i0.009977483448 \end{aligned}$$

Compiling

$$z_1 = 0.008305409518 + i0.005569575634$$

$$z_2 = -0.008976098746 + i0.004407907814$$

$$z_3 = 6.70689228525 \times 10^{-4} - i0.009977483448$$

Similarly, the three values of  $z$  from  $w_2$  in rectangular form are

$$z_4 = 0.008305409518 - i0.005569575634$$

$$z_5 = -0.008976098746 - i0.004407907814$$

$$z_6 = 6.70689228525 \times 10^{-4} + i0.009977483448$$

Using Vieta's substitution (Equation (6)),

$$\begin{aligned} y &= z + \frac{s}{z} \\ y &= z + \frac{(1 \times 10^{-4})}{z} \end{aligned}$$

we back substitute to find three values for  $y$ .

For example, choosing

$$z_1 = 0.008305409518 + i0.005569575634$$

gives

$$y_1 = 0.008305409518 + i0.005569575634 + \frac{1 \times 10^{-4}}{0.008305409518 + i0.005569575634}$$

$$\begin{aligned}
 &= 0.008305409518 + i0.005569575634 \\
 &\quad + \frac{1 \times 10^{-4}}{0.008305409078 + i0.00556957634} \times \frac{0.008305409518 - i0.00556957634}{0.008305409518 - i0.00556957634} \\
 &= 0.008305409518 + i0.005569575634 \\
 &\quad + \frac{1 \times 10^{-4}}{1 \times 10^{-4}} (0.008305409518 - i0.00556957634) \\
 &= 0.016610819036
 \end{aligned}$$

The values of  $z_1$ ,  $z_2$  and  $z_3$  give

$$y_1 = 0.016610819036$$

$$y_2 = -0.01795219749$$

$$y_3 = 0.001341378457$$

respectively. The three other  $z$  values of  $z_4$ ,  $z_5$  and  $z_6$  give the same values as  $y_1$ ,  $y_2$  and  $y_3$ , respectively.

Now, using the substitution of

$$x = y + 0.01$$

the three roots of the given cubic equation are

$$\begin{aligned}
 x_1 &= 0.016610819036 + 0.01 \\
 &= 0.026610819036
 \end{aligned}$$

$$\begin{aligned}
 x_2 &= -0.01795219749 + 0.01 \\
 &= -0.00795219749
 \end{aligned}$$

$$\begin{aligned}
 x_3 &= 0.001341378457 + 0.01 \\
 &= 0.011341378457
 \end{aligned}$$

## NONLINEAR EQUATIONS

Topic	Exact Solution to Cubic Equations
Summary	Textbook notes on finding the exact solution to a cubic equation.
Major	General Engineering
Authors	Autar Kaw
Last Revised	July 3, 2009
Web Site	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

# **Chapter 03.03**

## **Bisection Method of Solving a Nonlinear Equation**

*After reading this chapter, you should be able to:*

1. follow the algorithm of the bisection method of solving a nonlinear equation,
2. use the bisection method to solve examples of finding roots of a nonlinear equation, and
3. enumerate the advantages and disadvantages of the bisection method.

### **What is the bisection method and what is it based on?**

One of the first numerical methods developed to find the root of a nonlinear equation  $f(x) = 0$  was the bisection method (also called *binary-search* method). The method is based on the following theorem.

#### **Theorem**

An equation  $f(x) = 0$ , where  $f(x)$  is a real continuous function, has at least one root between  $x_\ell$  and  $x_u$  if  $f(x_\ell)f(x_u) < 0$  (See Figure 1).

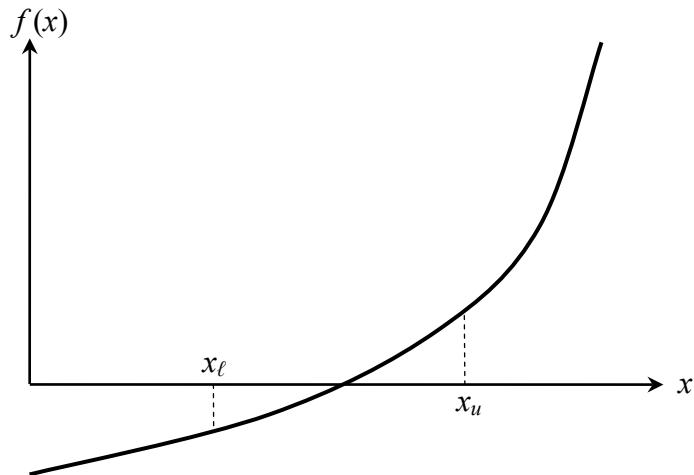
Note that if  $f(x_\ell)f(x_u) > 0$ , there may or may not be any root between  $x_\ell$  and  $x_u$  (Figures 2 and 3). If  $f(x_\ell)f(x_u) < 0$ , then there may be more than one root between  $x_\ell$  and  $x_u$  (Figure 4). So the theorem only guarantees one root between  $x_\ell$  and  $x_u$ .

### **Bisection method**

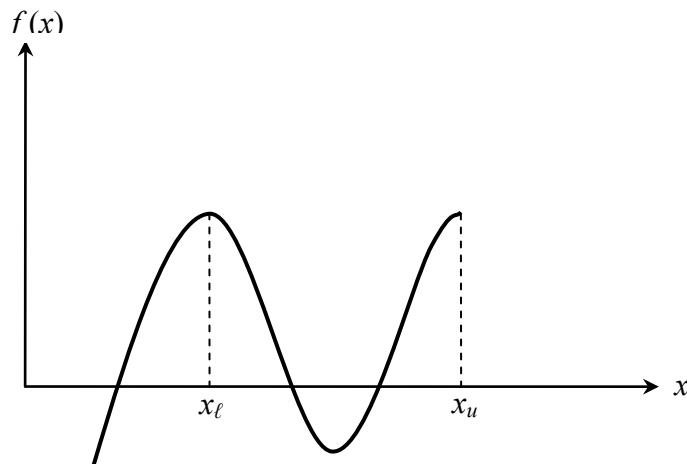
Since the method is based on finding the root between two points, the method falls under the category of bracketing methods.

Since the root is bracketed between two points,  $x_\ell$  and  $x_u$ , one can find the mid-point,  $x_m$  between  $x_\ell$  and  $x_u$ . This gives us two new intervals

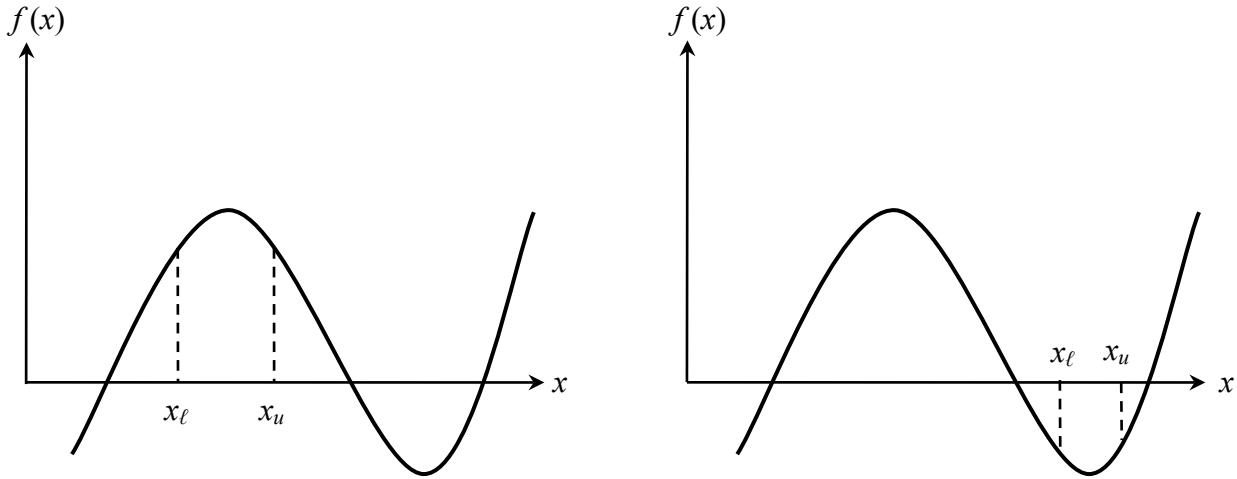
1.  $x_\ell$  and  $x_m$ , and
2.  $x_m$  and  $x_u$ .



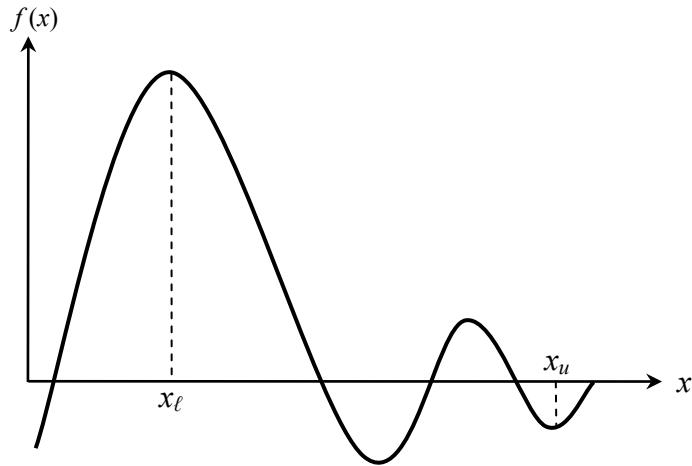
**Figure 1** At least one root exists between the two points if the function is real, continuous, and changes sign.



**Figure 2** If the function  $f(x)$  does not change sign between the two points, roots of the equation  $f(x) = 0$  may still exist between the two points.



**Figure 3** If the function  $f(x)$  does not change sign between two points, there may not be any roots for the equation  $f(x) = 0$  between the two points.



**Figure 4** If the function  $f(x)$  changes sign between the two points, more than one root for the equation  $f(x) = 0$  may exist between the two points.

Is the root now between  $x_l$  and  $x_m$  or between  $x_m$  and  $x_u$ ? Well, one can find the sign of  $f(x_l)f(x_m)$ , and if  $f(x_l)f(x_m) < 0$  then the new bracket is between  $x_l$  and  $x_m$ , otherwise, it is between  $x_m$  and  $x_u$ . So, you can see that you are literally halving the interval. As one repeats this process, the width of the interval  $[x_l, x_u]$  becomes smaller and smaller, and you can zero in to the root of the equation  $f(x) = 0$ . The algorithm for the bisection method is given as follows.

### Algorithm for the bisection method

The steps to apply the bisection method to find the root of the equation  $f(x) = 0$  are

1. Choose  $x_\ell$  and  $x_u$  as two guesses for the root such that  $f(x_\ell)f(x_u) < 0$ , or in other words,  $f(x)$  changes sign between  $x_\ell$  and  $x_u$ .
2. Estimate the root,  $x_m$ , of the equation  $f(x) = 0$  as the mid-point between  $x_\ell$  and  $x_u$  as

$$x_m = \frac{x_\ell + x_u}{2}$$

3. Now check the following
  - a) If  $f(x_\ell)f(x_m) < 0$ , then the root lies between  $x_\ell$  and  $x_m$ ; then  $x_\ell = x_m$  and  $x_u = x_m$ .
  - b) If  $f(x_\ell)f(x_m) > 0$ , then the root lies between  $x_m$  and  $x_u$ ; then  $x_u = x_m$  and  $x_\ell = x_m$ .
  - c) If  $f(x_\ell)f(x_m) = 0$ ; then the root is  $x_m$ . Stop the algorithm if this is true.
4. Find the new estimate of the root

$$x_m = \frac{x_\ell + x_u}{2}$$

Find the absolute relative approximate error as

$$|\epsilon_a| = \left| \frac{x_m^{\text{new}} - x_m^{\text{old}}}{x_m^{\text{new}}} \right| \times 100$$

where

$x_m^{\text{new}}$  = estimated root from present iteration

$x_m^{\text{old}}$  = estimated root from previous iteration

5. Compare the absolute relative approximate error  $|\epsilon_a|$  with the pre-specified relative error tolerance  $\epsilon_s$ . If  $|\epsilon_a| > \epsilon_s$ , then go to Step 3, else stop the algorithm. Note one should also check whether the number of iterations is more than the maximum number of iterations allowed. If so, one needs to terminate the algorithm and notify the user about it.

### Example 1

You are working for ‘DOWN THE TOILET COMPANY’ that makes floats for ABC commodes. The floating ball has a specific gravity of 0.6 and has a radius of 5.5 cm. You are asked to find the depth to which the ball is submerged when floating in water.

The equation that gives the depth  $x$  to which the ball is submerged under water is given by

$$x^3 - 0.165x^2 + 3.993 \times 10^{-4} = 0$$

Use the bisection method of finding roots of equations to find the depth  $x$  to which the ball is submerged under water. 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

From the physics of the problem, the ball would be submerged between  $x = 0$  and  $x = 2R$ , where

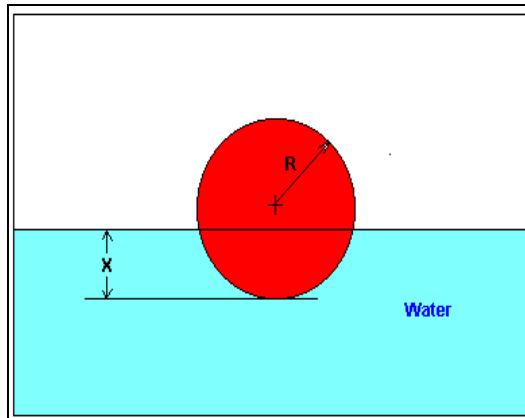
$R$  = radius of the ball,

that is

$$0 \leq x \leq 2R$$

$$0 \leq x \leq 2(0.055)$$

$$0 \leq x \leq 0.11$$



**Figure 5** Floating ball problem.

Lets us assume

$$x_\ell = 0, x_u = 0.11$$

Check if the function changes sign between  $x_\ell$  and  $x_u$ .

$$f(x_\ell) = f(0) = (0)^3 - 0.165(0)^2 + 3.993 \times 10^{-4} = 3.993 \times 10^{-4}$$

$$f(x_u) = f(0.11) = (0.11)^3 - 0.165(0.11)^2 + 3.993 \times 10^{-4} = -2.662 \times 10^{-4}$$

Hence

$$f(x_\ell)f(x_u) = f(0)f(0.11) = (3.993 \times 10^{-4})(-2.662 \times 10^{-4}) < 0$$

So there is at least one root between  $x_\ell$  and  $x_u$ , that is between 0 and 0.11.

#### Iteration 1

The estimate of the root is

$$\begin{aligned} x_m &= \frac{x_\ell + x_u}{2} \\ &= \frac{0 + 0.11}{2} \\ &= 0.055 \end{aligned}$$

$$f(x_m) = f(0.055) = (0.055)^3 - 0.165(0.055)^2 + 3.993 \times 10^{-4} = 6.655 \times 10^{-5}$$

$$f(x_\ell)f(x_m) = f(0)f(0.055) = (3.993 \times 10^{-4})(6.655 \times 10^{-5}) > 0$$

Hence the root is bracketed between  $x_m$  and  $x_u$ , that is, between 0.055 and 0.11. So, the lower and upper limit of the new bracket is

$$x_\ell = 0.055, x_u = 0.11$$

At this point, the absolute relative approximate error  $|e_a|$  cannot be calculated as we do not have a previous approximation.

### Iteration 2

The estimate of the root is

$$\begin{aligned} x_m &= \frac{x_\ell + x_u}{2} \\ &= \frac{0.055 + 0.11}{2} \\ &= 0.0825 \end{aligned}$$

$$f(x_m) = f(0.0825) = (0.0825)^3 - 0.165(0.0825)^2 + 3.993 \times 10^{-4} = -1.622 \times 10^{-4}$$

$$f(x_\ell)f(x_m) = f(0.055)f(0.0825) = (6.655 \times 10^{-5}) \times (-1.622 \times 10^{-4}) < 0$$

Hence, the root is bracketed between  $x_\ell$  and  $x_m$ , that is, between 0.055 and 0.0825. So the lower and upper limit of the new bracket is

$$x_\ell = 0.055, x_u = 0.0825$$

The absolute relative approximate error  $|e_a|$  at the end of Iteration 2 is

$$\begin{aligned} |e_a| &= \left| \frac{x_m^{\text{new}} - x_m^{\text{old}}}{x_m^{\text{new}}} \right| \times 100 \\ &= \left| \frac{0.0825 - 0.055}{0.0825} \right| \times 100 \\ &= 33.33\% \end{aligned}$$

None of the significant digits are at least correct in the estimated root of  $x_m = 0.0825$  because the absolute relative approximate error is greater than 5%.

### Iteration 3

$$\begin{aligned} x_m &= \frac{x_\ell + x_u}{2} \\ &= \frac{0.055 + 0.0825}{2} \\ &= 0.06875 \end{aligned}$$

$$f(x_m) = f(0.06875) = (0.06875)^3 - 0.165(0.06875)^2 + 3.993 \times 10^{-4} = -5.563 \times 10^{-5}$$

$$f(x_\ell)f(x_m) = f(0.055)f(0.06875) = (6.655 \times 10^{-5}) \times (-5.563 \times 10^{-5}) < 0$$

Hence, the root is bracketed between  $x_\ell$  and  $x_m$ , that is, between 0.055 and 0.06875. So the lower and upper limit of the new bracket is

$$x_\ell = 0.055, x_u = 0.06875$$

The absolute relative approximate error  $|e_a|$  at the ends of Iteration 3 is

$$\begin{aligned} |e_a| &= \left| \frac{x_m^{\text{new}} - x_m^{\text{old}}}{x_m^{\text{new}}} \right| \times 100 \\ &= \left| \frac{0.06875 - 0.0825}{0.06875} \right| \times 100 \\ &= 20\% \end{aligned}$$

Still none of the significant digits are at least correct in the estimated root of the equation as the absolute relative approximate error is greater than 5%.

Seven more iterations were conducted and these iterations are shown in Table 1.

**Table 1** Root of  $f(x) = 0$  as function of number of iterations for bisection method.

Iteration	$x_l$	$x_u$	$x_m$	$ e_a  \%$	$f(x_m)$
1	0.00000	0.11	0.055	-----	$6.655 \times 10^{-5}$
2	0.055	0.11	0.0825	33.33	$-1.622 \times 10^{-4}$
3	0.055	0.0825	0.06875	20.00	$-5.563 \times 10^{-5}$
4	0.055	0.06875	0.06188	11.11	$4.484 \times 10^{-6}$
5	0.06188	0.06875	0.06531	5.263	$-2.593 \times 10^{-5}$
6	0.06188	0.06531	0.06359	2.702	$-1.0804 \times 10^{-5}$
7	0.06188	0.06359	0.06273	1.370	$-3.176 \times 10^{-6}$
8	0.06188	0.06273	0.0623	0.6897	$6.497 \times 10^{-7}$
9	0.0623	0.06273	0.06252	0.3436	$-1.265 \times 10^{-6}$
10	0.0623	0.06252	0.06241	0.1721	$-3.0768 \times 10^{-7}$

At the end of 10<sup>th</sup> iteration,

$$|e_a| = 0.1721\%$$

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

$$|e_a| \leq 0.5 \times 10^{2-m}$$

$$0.1721 \leq 0.5 \times 10^{2-m}$$

$$0.3442 \leq 10^{2-m}$$

$$\log(0.3442) \leq 2 - m$$

$$m \leq 2 - \log(0.3442) = 2.463$$

So

$$m = 2$$

The number of significant digits at least correct in the estimated root of 0.06241 at the end of the 10<sup>th</sup> iteration is 2.

### Advantages of bisection method

- a) The bisection method is always convergent. Since the method brackets the root, the method is guaranteed to converge.
- b) As iterations are conducted, the interval gets halved. So one can guarantee the error in the solution of the equation.

### Drawbacks of bisection method

- a) The convergence of the bisection method is slow as it is simply based on halving the interval.
- b) If one of the initial guesses is closer to the root, it will take larger number of iterations to reach the root.
- c) If a function  $f(x)$  is such that it just touches the  $x$ -axis (Figure 6) such as

$$f(x) = x^2 = 0$$

it will be unable to find the lower guess,  $x_\ell$ , and upper guess,  $x_u$ , such that

$$f(x_\ell)f(x_u) < 0$$

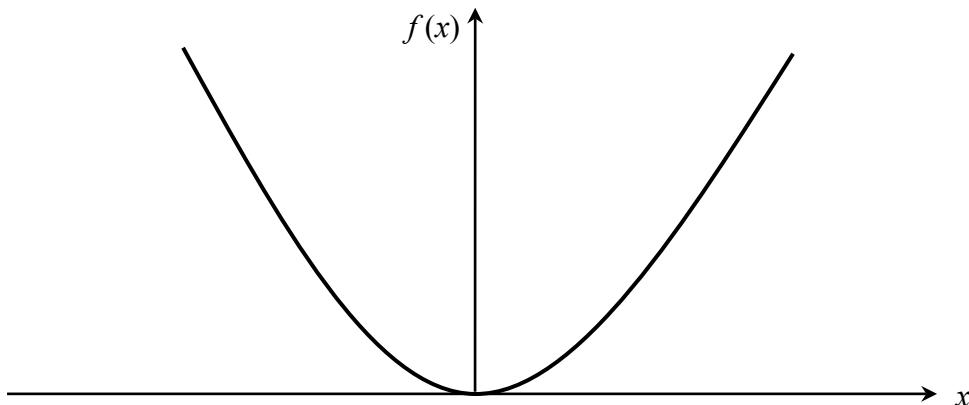
- d) For functions  $f(x)$  where there is a singularity<sup>1</sup> and it reverses sign at the singularity, the bisection method may converge on the singularity (Figure 7). An example includes

$$f(x) = \frac{1}{x}$$

where  $x_\ell = -2$ ,  $x_u = 3$  are valid initial guesses which satisfy

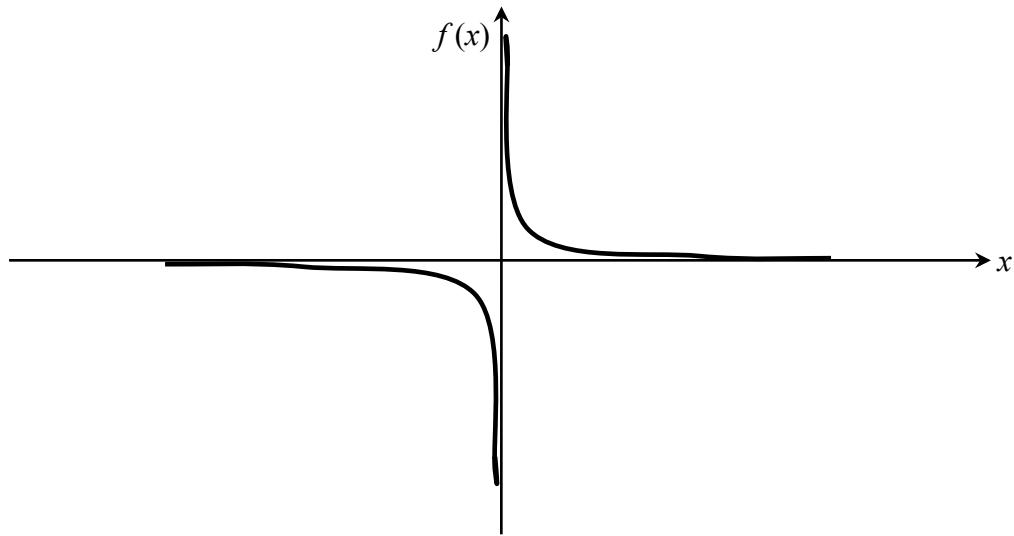
$$f(x_\ell)f(x_u) < 0$$

However, the function is not continuous and the theorem that a root exists is also not applicable.



**Figure 6** The equation  $f(x) = x^2 = 0$  has a single root at  $x = 0$  that cannot be bracketed.

<sup>1</sup> A singularity in a function is defined as a point where the function becomes infinite. For example, for a function such as  $1/x$ , the point of singularity is  $x = 0$  as it becomes infinite.



**Figure 7** The equation  $f(x) = \frac{1}{x} = 0$  has no root but changes sign.

---

#### NONLINEAR EQUATIONS

---

Topic	Bisection method of solving a nonlinear equation
Summary	These are textbook notes of bisection method of finding roots of nonlinear equation, including convergence and pitfalls.
Major	General Engineering
Authors	Autar Kaw
Date	January 15, 2012
Web Site	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

---

# **Chapter 03.04**

## **Newton-Raphson Method of Solving a Nonlinear Equation**

*After reading this chapter, you should be able to:*

1. derive the Newton-Raphson method formula,
2. develop the algorithm of the Newton-Raphson method,
3. use the Newton-Raphson method to solve a nonlinear equation, and
4. discuss the drawbacks of the Newton-Raphson method.

### **Introduction**

Methods such as the bisection method and the false position method of finding roots of a nonlinear equation  $f(x) = 0$  require bracketing of the root by two guesses. Such methods are called *bracketing methods*. These methods are always convergent since they are based on reducing the interval between the two guesses so as to zero in on the root of the equation.

In the Newton-Raphson method, the root is not bracketed. In fact, only one initial guess of the root is needed to get the iterative process started to find the root of an equation. The method hence falls in the category of *open methods*. Convergence in open methods is not guaranteed but if the method does converge, it does so much faster than the bracketing methods.

### **Derivation**

The Newton-Raphson method is based on the principle that if the initial guess of the root of  $f(x) = 0$  is at  $x_i$ , then if one draws the tangent to the curve at  $f(x_i)$ , the point  $x_{i+1}$  where the tangent crosses the  $x$ -axis is an improved estimate of the root (Figure 1).

Using the definition of the slope of a function, at  $x = x_i$

$$\begin{aligned} f'(x_i) &= \tan \theta \\ &= \frac{f(x_i) - 0}{x_i - x_{i+1}}, \end{aligned}$$

which gives

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} \quad (1)$$

Equation (1) is called the Newton-Raphson formula for solving nonlinear equations of the form  $f(x)=0$ . So starting with an initial guess,  $x_i$ , one can find the next guess,  $x_{i+1}$ , by using Equation (1). One can repeat this process until one finds the root within a desirable tolerance.

### Algorithm

The steps of the Newton-Raphson method to find the root of an equation  $f(x)=0$  are

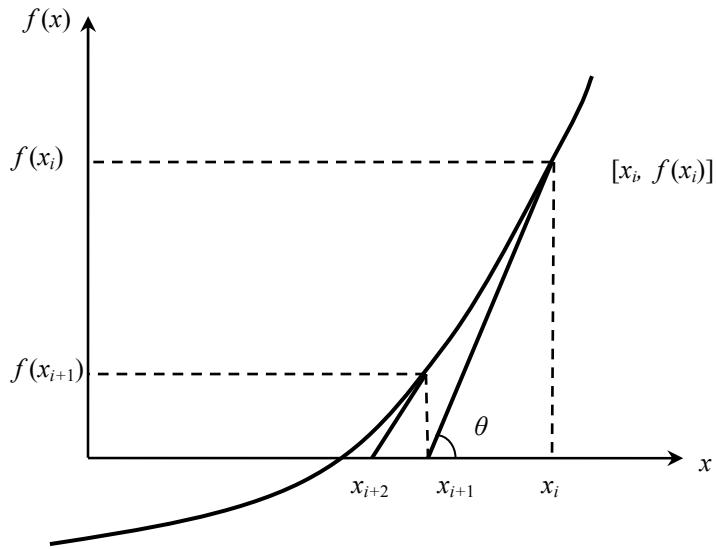
1. Evaluate  $f'(x)$  symbolically
2. Use an initial guess of the root,  $x_i$ , to estimate the new value of the root,  $x_{i+1}$ , as

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}$$

3. Find the absolute relative approximate error  $|e_a|$  as

$$|e_a| = \left| \frac{x_{i+1} - x_i}{x_{i+1}} \right| \times 100$$

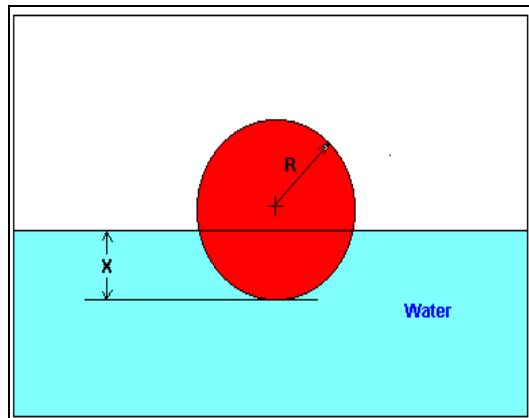
4. Compare the absolute relative approximate error with the pre-specified relative error tolerance,  $e_s$ . If  $|e_a| > e_s$ , then go to Step 2, else stop the algorithm. Also, check if the number of iterations has exceeded the maximum number of iterations allowed. If so, one needs to terminate the algorithm and notify the user.



**Figure 1** Geometrical illustration of the Newton-Raphson method.

**Example 1**

You are working for ‘DOWN THE TOILET COMPANY’ that makes floats for ABC commodes. The floating ball has a specific gravity of 0.6 and has a radius of 5.5 cm. You are asked to find the depth to which the ball is submerged when floating in water.



**Figure 2** Floating ball problem.

The equation that gives the depth  $x$  in meters to which the ball is submerged under water is given by

$$x^3 - 0.165x^2 + 3.993 \times 10^{-4} = 0$$

Use the Newton-Raphson method of finding roots of equations to find

- the depth  $x$  to which the ball is submerged under water. Conduct three iterations to estimate the root of the above equation.
- 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**

$$f(x) = x^3 - 0.165x^2 + 3.993 \times 10^{-4}$$

$$f'(x) = 3x^2 - 0.33x$$

Let us assume the initial guess of the root of  $f(x)=0$  is  $x_0 = 0.05$  m. This is a reasonable guess (discuss why  $x=0$  and  $x=0.11$  m are not good choices) as the extreme values of the depth  $x$  would be 0 and the diameter (0.11 m) of the ball.

Iteration 1

The estimate of the root is

$$\begin{aligned} x_1 &= x_0 - \frac{f(x_0)}{f'(x_0)} \\ &= 0.05 - \frac{(0.05)^3 - 0.165(0.05)^2 + 3.993 \times 10^{-4}}{3(0.05)^2 - 0.33(0.05)} \\ &= 0.05 - \frac{1.118 \times 10^{-4}}{-9 \times 10^{-3}} \\ &= 0.05 - (-0.01242) \\ &= 0.06242 \end{aligned}$$

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

$$\begin{aligned} |\epsilon_a| &= \left| \frac{x_1 - x_0}{x_1} \right| \times 100 \\ &= \left| \frac{0.06242 - 0.05}{0.06242} \right| \times 100 \\ &= 19.90\% \end{aligned}$$

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

### Iteration 2

The estimate of the root is

$$\begin{aligned} x_2 &= x_1 - \frac{f(x_1)}{f'(x_1)} \\ &= 0.06242 - \frac{(0.06242)^3 - 0.165(0.06242)^2 + 3.993 \times 10^{-4}}{3(0.06242)^2 - 0.33(0.06242)} \\ &= 0.06242 - \frac{-3.97781 \times 10^{-7}}{-8.90973 \times 10^{-3}} \\ &= 0.06242 - (4.4646 \times 10^{-5}) \\ &= 0.06238 \end{aligned}$$

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

$$\begin{aligned} |\epsilon_a| &= \left| \frac{x_2 - x_1}{x_2} \right| \times 100 \\ &= \left| \frac{0.06238 - 0.06242}{0.06238} \right| \times 100 \\ &= 0.0716\% \end{aligned}$$

The maximum value of  $m$  for which  $|\epsilon_a| \leq 0.5 \times 10^{2-m}$  is 2.844. Hence, the number of significant digits at least correct in the answer is 2.

### Iteration 3

The estimate of the root is

$$\begin{aligned} x_3 &= x_2 - \frac{f(x_2)}{f'(x_2)} \\ &= 0.06238 - \frac{(0.06238)^3 - 0.165(0.06238)^2 + 3.993 \times 10^{-4}}{3(0.06238)^2 - 0.33(0.06238)} \\ &= 0.06238 - \frac{4.44 \times 10^{-11}}{-8.91171 \times 10^{-3}} \\ &= 0.06238 - (-4.9822 \times 10^{-9}) \\ &= 0.06238 \end{aligned}$$

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

$$\begin{aligned} |\epsilon_a| &= \left| \frac{0.06238 - 0.06238}{0.06238} \right| \times 100 \\ &= 0 \end{aligned}$$

The number of significant digits at least correct is 4, as only 4 significant digits are carried through in all the calculations.

### Drawbacks of the Newton-Raphson Method

#### 1. Divergence at inflection points

If the selection of the initial guess or an iterated value of the root turns out to be close to the inflection point (see the definition in the appendix of this chapter) of the function  $f(x)$  in the equation  $f(x) = 0$ , Newton-Raphson method may start diverging away from the root. It may then start converging back to the root. For example, to find the root of the equation

$$f(x) = (x - 1)^3 + 0.512 = 0$$

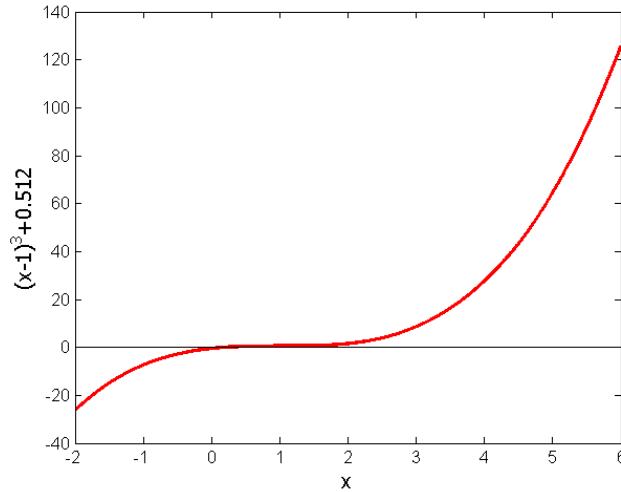
the Newton-Raphson method reduces to

$$x_{i+1} = x_i - \frac{(x_i^3 - 1)^3 + 0.512}{3(x_i - 1)^2}$$

Starting with an initial guess of  $x_0 = 5.0$ , Table 1 shows the iterated values of the root of the equation. As you can observe, the root starts to diverge at Iteration 6 because the previous estimate of 0.92589 is close to the inflection point of  $x = 1$  (the value of  $f'(x)$  is zero at the inflection point). Eventually, after 12 more iterations the root converges to the exact value of  $x = 0.2$ .

**Table 1** Divergence near inflection point.

Iteration Number	$x_i$
0	5.0000
1	3.6560
2	2.7465
3	2.1084
4	1.6000
5	0.92589
6	-30.119
7	-19.746
8	-12.831
9	-8.2217
10	-5.1498
11	-3.1044
12	-1.7464
13	-0.85356
14	-0.28538
15	0.039784
16	0.17475
17	0.19924
18	0.2



**Figure 3** Divergence at inflection point for  $f(x) = (x-1)^3 + 0.512$ .

## 2. Division by zero

For the equation

$$f(x) = x^3 - 0.03x^2 + 2.4 \times 10^{-6} = 0$$

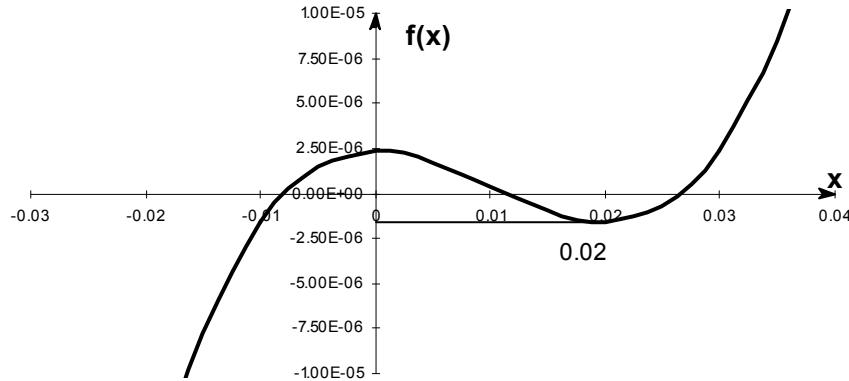
the Newton-Raphson method reduces to

$$x_{i+1} = x_i - \frac{x_i^3 - 0.03x_i^2 + 2.4 \times 10^{-6}}{3x_i^2 - 0.06x_i}$$

For  $x_0 = 0$  or  $x_0 = 0.02$ , division by zero occurs (Figure 4). For an initial guess close to 0.02 such as  $x_0 = 0.01999$ , one may avoid division by zero, but then the denominator in the formula is a small number. For this case, as given in Table 2, even after 9 iterations, the Newton-Raphson method does not converge.

**Table 2** Division by near zero in Newton-Raphson method.

Iteration Number	$x_i$	$f(x_i)$	$ e_a  \%$
0	0.019990	$-1.60000 \times 10^{-6}$	—
1	-2.6480	18.778	100.75
2	-1.7620	-5.5638	50.282
3	-1.1714	-1.6485	50.422
4	-0.77765	-0.48842	50.632
5	-0.51518	-0.14470	50.946
6	-0.34025	-0.042862	51.413
7	-0.22369	-0.012692	52.107
8	-0.14608	-0.0037553	53.127
9	-0.094490	-0.0011091	54.602



**Figure 4** Pitfall of division by zero or a near zero number.

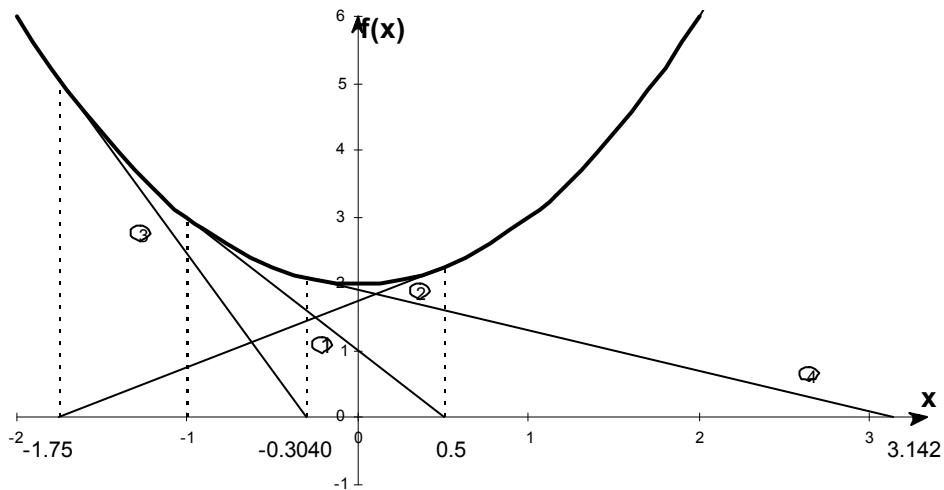
### 3. Oscillations near local maximum and minimum

Results obtained from the Newton-Raphson method may oscillate about the local maximum or minimum without converging on a root but converging on the local maximum or minimum. Eventually, it may lead to division by a number close to zero and may diverge.

For example, for

$$f(x) = x^2 + 2 = 0$$

the equation has no real roots (Figure 5 and Table 3).



**Figure 5** Oscillations around local minima for  $f(x) = x^2 + 2$ .

**Table 3** Oscillations near local maxima and minima in Newton-Raphson method.

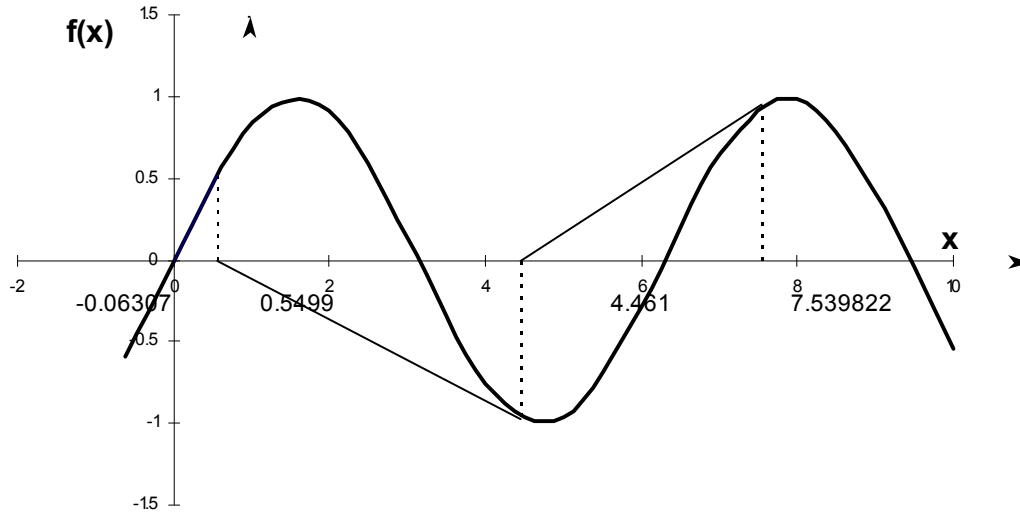
Iteration Number	$x_i$	$f(x_i)$	$ e_a  \%$
0	-1.0000	3.00	—
1	0.5	2.25	300.00
2	-1.75	5.063	128.571
3	-0.30357	2.092	476.47
4	3.1423	11.874	109.66
5	1.2529	3.570	150.80
6	-0.17166	2.029	829.88
7	5.7395	34.942	102.99
8	2.6955	9.266	112.93
9	0.97678	2.954	175.96

**4. Root jumping**

In some case where the function  $f(x)$  is oscillating and has a number of roots, one may choose an initial guess close to a root. However, the guesses may jump and converge to some other root. For example for solving the equation  $\sin x = 0$  if you choose  $x_0 = 2.4\pi = (7.539822)$  as an initial guess, it converges to the root of  $x = 0$  as shown in Table 4 and Figure 6. However, one may have chosen this as an initial guess to converge to  $x = 2\pi = 6.2831853$ .

**Table 4** Root jumping in Newton-Raphson method.

Iteration Number	$x_i$	$f(x_i)$	$ e_a  \%$
0	7.539822	0.951	—
1	4.462	-0.969	68.973
2	0.5499	0.5226	711.44
3	-0.06307	-0.06303	971.91
4	$8.376 \times 10^{-4}$	$8.375 \times 10^{-5}$	$7.54 \times 10^4$
5	$-1.95861 \times 10^{-13}$	$-1.95861 \times 10^{-13}$	$4.28 \times 10^{10}$



**Figure 6** Root jumping from intended location of root for  $f(x) = \sin x = 0$ .

#### Appendix A. What is an inflection point?

For a function  $f(x)$ , the point where the concavity changes from up-to-down or down-to-up is called its inflection point. For example, for the function  $f(x) = (x-1)^3$ , the concavity changes at  $x=1$  (see Figure 3), and hence  $(1,0)$  is an inflection point.

An inflection points MAY exist at a point where  $f''(x)=0$  and where  $f''(x)$  does not exist. The reason we say that it MAY exist is because if  $f''(x)=0$ , it only makes it a possible inflection point. For example, for  $f(x) = x^4 - 16$ ,  $f''(0)=0$ , but the concavity does not change at  $x=0$ . Hence the point  $(0, -16)$  is not an inflection point of  $f(x) = x^4 - 16$ .

For  $f(x) = (x-1)^3$ ,  $f''(x)$  changes sign at  $x=1$  ( $f''(x) < 0$  for  $x < 1$ , and  $f''(x) > 0$  for  $x > 1$ ), and thus brings up the *Inflection Point Theorem* for a function  $f(x)$  that states the following.

"If  $f'(c)$  exists and  $f''(c)$  changes sign at  $x=c$ , then the point  $(c, f(c))$  is an inflection point of the graph of  $f$ ."

#### Appendix B. Derivation of Newton-Raphson method from Taylor series

Newton-Raphson method can also be derived from Taylor series. For a general function  $f(x)$ , the Taylor series is

$$f(x_{i+1}) = f(x_i) + f'(x_i)(x_{i+1} - x_i) + \frac{f''(x_i)}{2!}(x_{i+1} - x_i)^2 + \dots$$

As an approximation, taking only the first two terms of the right hand side,

$$f(x_{i+1}) \approx f(x_i) + f'(x_i)(x_{i+1} - x_i)$$

and we are seeking a point where  $f(x)=0$ , that is, if we assume

$$f(x_{i+1}) = 0,$$

$$0 \approx f(x_i) + f'(x_i)(x_{i+1} - x_i)$$

which gives

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}$$

This is the same Newton-Raphson method formula series as derived previously using the geometric method.

---

## NONLINEAR EQUATIONS

---

Topic	Newton-Raphson Method of Solving Nonlinear Equations
Summary	Text book notes of Newton-Raphson method of finding roots of nonlinear equation, including convergence and pitfalls.
Major	General Engineering
Authors	Autar Kaw
Date	December 23, 2009
Web Site	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

---

# Chapter 03.05

## Secant Method of Solving Nonlinear Equations

After reading this chapter, you should be able to:

1. derive the secant method to solve for the roots of a nonlinear equation,
2. use the secant method to numerically solve a nonlinear equation.

**What is the secant method and why would I want to use it instead of the Newton-Raphson method?**

The Newton-Raphson method of solving a nonlinear equation  $f(x)=0$  is given by the iterative formula

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} \quad (1)$$

One of the drawbacks of the Newton-Raphson method is that you have to evaluate the derivative of the function. With availability of symbolic manipulators such as Maple, MathCAD, MATHEMATICA and MATLAB, this process has become more convenient. However, it still can be a laborious process, and even intractable if the function is derived as part of a numerical scheme. To overcome these drawbacks, the derivative of the function,  $f(x)$  is approximated as

$$f'(x_i) = \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}} \quad (2)$$

Substituting Equation (2) in Equation (1) gives

$$x_{i+1} = x_i - \frac{f(x_i)(x_i - x_{i-1})}{f(x_i) - f(x_{i-1})} \quad (3)$$

The above equation is called the secant method. This method now requires two initial guesses, but unlike the bisection method, the two initial guesses do not need to bracket the root of the equation. The secant method is an open method and may or may not converge. However, when secant method converges, it will typically converge faster than the bisection method. However, since the derivative is approximated as given by Equation (2), it typically converges slower than the Newton-Raphson method.

The secant method can also be derived from geometry, as shown in Figure 1. Taking two initial guesses,  $x_{i-1}$  and  $x_i$ , one draws a straight line between  $f(x_i)$  and  $f(x_{i-1})$  passing through the  $x$ -axis at  $x_{i+1}$ .  $ABE$  and  $DCE$  are similar triangles.

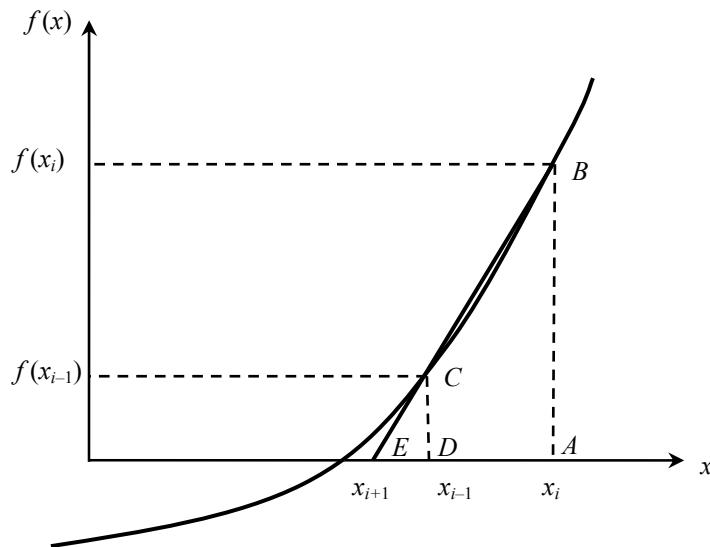
Hence

$$\frac{AB}{AE} = \frac{DC}{DE}$$

$$\frac{f(x_i)}{x_i - x_{i+1}} = \frac{f(x_{i-1})}{x_{i-1} - x_{i+1}}$$

On rearranging, the secant method is given as

$$x_{i+1} = x_i - \frac{f(x_i)(x_i - x_{i-1})}{f(x_i) - f(x_{i-1})}$$



**Figure 1** Geometrical representation of the secant method.

### Example 1

You are working for ‘DOWN THE TOILET COMPANY’ that makes floats (Figure 2) for ABC commodes. The floating ball has a specific gravity of 0.6 and a radius of 5.5 cm. You are asked to find the depth to which the ball is submerged when floating in water.

The equation that gives the depth  $x$  to which the ball is submerged under water is given by

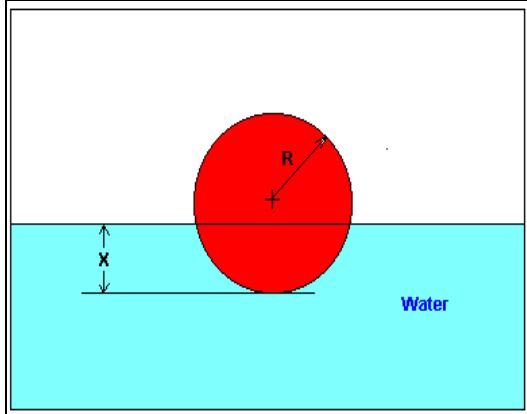
$$x^3 - 0.165x^2 + 3.993 \times 10^{-4} = 0$$

Use the secant method of finding roots of equations to find the depth  $x$  to which the ball is submerged under water. Conduct three iterations to estimate the root of the above equation. Find the absolute relative approximate error and the number of significant digits at least correct at the end of each iteration.

**Solution**

$$f(x) = x^3 - 0.165x^2 + 3.993 \times 10^{-4}$$

Let us assume the initial guesses of the root of  $f(x) = 0$  as  $x_{-1} = 0.02$  and  $x_0 = 0.05$ .



**Figure 2** Floating ball problem.

Iteration 1

The estimate of the root is

$$\begin{aligned} x_1 &= x_0 - \frac{f(x_0)(x_0 - x_{-1})}{f(x_0) - f(x_{-1})} \\ &= x_0 - \frac{(x_0^3 - 0.165x_0^2 + 3.993 \times 10^{-4}) \times (x_0 - x_{-1})}{(x_0^3 - 0.165x_0^2 + 3.993 \times 10^{-4}) - (x_{-1}^3 - 0.165x_{-1}^2 + 3.993 \times 10^{-4})} \\ &= 0.05 - \frac{[0.05^3 - 0.165(0.05)^2 + 3.993 \times 10^{-4}] \times [0.05 - 0.02]}{[0.05^3 - 0.165(0.05)^2 + 3.993 \times 10^{-4}] - [0.02^3 - 0.165(0.02)^2 + 3.993 \times 10^{-4}]} \\ &= 0.06461 \end{aligned}$$

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

$$\begin{aligned} |e_a| &= \left| \frac{x_1 - x_0}{x_1} \right| \times 100 \\ &= \left| \frac{0.06461 - 0.05}{0.06461} \right| \times 100 \\ &= 22.62\% \end{aligned}$$

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

$$\begin{aligned} x_2 &= x_1 - \frac{f(x_1)(x_1 - x_0)}{f(x_1) - f(x_0)} \\ &= x_1 - \frac{(x_1^3 - 0.165x_1^2 + 3.993 \times 10^{-4}) \times (x_1 - x_0)}{(x_1^3 - 0.165x_1^2 + 3.993 \times 10^{-4}) - (x_0^3 - 0.165x_0^2 + 3.993 \times 10^{-4})} \end{aligned}$$

$$\begin{aligned}
&= 0.06461 - \frac{\left[0.06461^3 - 0.165(0.06461)^2 + 3.993 \times 10^{-4}\right] \times (0.06461 - 0.05)}{\left[0.06461^3 - 0.165(0.06461)^2 + 3.993 \times 10^{-4}\right] - \left[0.05^3 - 0.165(0.05)^2 + 3.993 \times 10^{-4}\right]} \\
&= 0.06241
\end{aligned}$$

The absolute relative approximate error  $|e_a|$  at the end of Iteration 2 is

$$\begin{aligned}
|e_a| &= \left| \frac{x_2 - x_1}{x_2} \right| \times 100 \\
&= \left| \frac{0.06241 - 0.06461}{0.06241} \right| \times 100 \\
&= 3.525\%
\end{aligned}$$

The number of significant digits at least correct is 1, as you need an absolute relative approximate error of 5% or less.

### Iteration 3

$$\begin{aligned}
x_3 &= x_2 - \frac{f(x_2)(x_2 - x_1)}{f(x_2) - f(x_1)} \\
&= x_2 - \frac{(x_2^3 - 0.165x_2^2 + 3.993 \times 10^{-4}) \times (x_2 - x_1)}{(x_2^3 - 0.165x_2^2 + 3.993 \times 10^{-4}) - (x_1^3 - 0.165x_1^2 + 3.993 \times 10^{-4})} \\
&= 0.06241 - \frac{\left[0.06241^3 - 0.165(0.06241)^2 + 3.993 \times 10^{-4}\right] \times (0.06241 - 0.06461)}{\left[0.06241^3 - 0.165(0.06241)^2 + 3.993 \times 10^{-4}\right] - \left[0.06461^3 - 0.165(0.06461)^2 + 3.993 \times 10^{-4}\right]} \\
&= 0.06238
\end{aligned}$$

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

$$\begin{aligned}
|e_a| &= \left| \frac{x_3 - x_2}{x_3} \right| \times 100 \\
&= \left| \frac{0.06238 - 0.06241}{0.06238} \right| \times 100 \\
&= 0.0595\%
\end{aligned}$$

The number of significant digits at least correct is 2, as you need an absolute relative approximate error of 0.5% or less. Table 1 shows the secant method calculations for the results from the above problem.

**Table 1** Secant method results as a function of iterations.

Iteration Number, $i$	$x_{i-1}$	$x_i$	$x_{i+1}$	$ e_a  \%$	$f(x_{i+1})$
1	0.02	0.05	0.06461	22.62	$-1.9812 \times 10^{-5}$
2	0.05	0.06461	0.06241	3.525	$-3.2852 \times 10^{-7}$
3	0.06461	0.06241	0.06238	0.0595	$2.0252 \times 10^{-9}$
4	0.06241	0.06238	0.06238	$-3.64 \times 10^{-4}$	$-1.8576 \times 10^{-13}$

---

## NONLINEAR EQUATIONS

---

Topic	Secant Method for Solving Nonlinear Equations.
Summary	These are textbook notes of secant method of finding roots of nonlinear equations. Derivations and examples are included.
Major	General Engineering
Authors	Autar Kaw
Date	December 23, 2009
Web Site	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

---

# Chapter 03.06

## False-Position Method of Solving a Nonlinear Equation

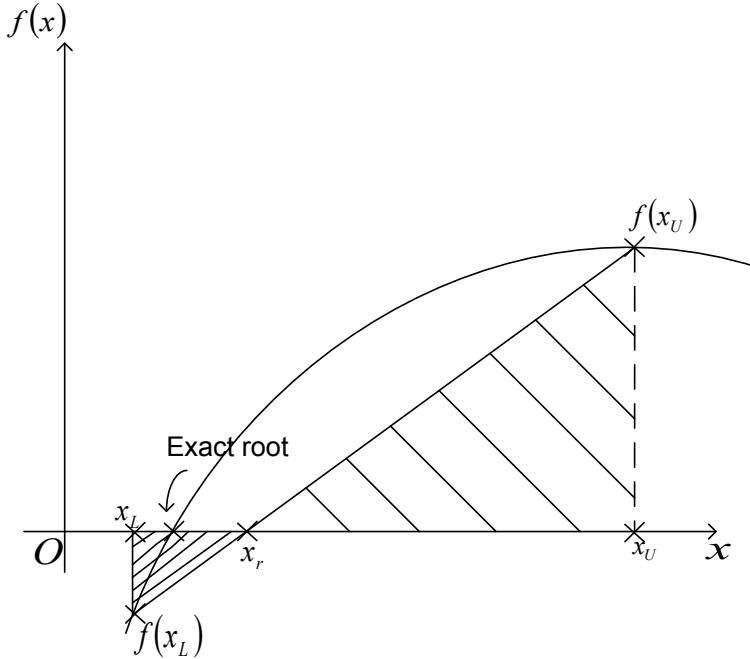
After reading this chapter, you should be able to

1. follow the algorithm of the false-position method of solving a nonlinear equation,
2. apply the false-position method to find roots of a nonlinear equation.

### Introduction

In Chapter 03.03, the bisection method was described as one of the simple bracketing methods of solving a nonlinear equation of the general form

$$f(x) = 0 \quad (1)$$



**Figure 1** False-Position Method

The above nonlinear equation can be stated as finding the value of  $x$  such that Equation (1) is satisfied.

In the bisection method, we identify proper values of  $x_L$  (lower bound value) and  $x_U$  (upper bound value) for the current bracket, such that

$$f(x_L)f(x_U) < 0. \quad (2)$$

The next predicted/improved root  $x_r$  can be computed as the midpoint between  $x_L$  and  $x_U$  as

$$x_r = \frac{x_L + x_U}{2} \quad (3)$$

The new upper and lower bounds are then established, and the procedure is repeated until the convergence is achieved (such that the new lower and upper bounds are sufficiently close to each other).

However, in the example shown in Figure 1, the bisection method may not be efficient because it does not take into consideration that  $f(x_L)$  is much closer to the zero of the function  $f(x)$  as compared to  $f(x_U)$ . In other words, the next predicted root  $x_r$  would be closer to  $x_L$  (in the example as shown in Figure 1), than the mid-point between  $x_L$  and  $x_U$ . The false-position method takes advantage of this observation mathematically by drawing a secant from the function value at  $x_L$  to the function value at  $x_U$ , and estimates the root as where it crosses the  $x$ -axis.

### False-Position Method

Based on two similar triangles, shown in Figure 1, one gets

$$\frac{0 - f(x_L)}{x_r - x_L} = \frac{0 - f(x_U)}{x_r - x_U} \quad (4)$$

From Equation (4), one obtains

$$\begin{aligned} (x_r - x_L)f(x_U) &= (x_r - x_U)f(x_L) \\ x_U f(x_L) - x_L f(x_U) &= x_r \{f(x_L) - f(x_U)\} \end{aligned}$$

The above equation can be solved to obtain the next predicted root  $x_m$  as

$$x_r = \frac{x_U f(x_L) - x_L f(x_U)}{f(x_L) - f(x_U)} \quad (5)$$

The above equation, through simple algebraic manipulations, can also be expressed as

$$x_r = x_U - \frac{f(x_U)}{\left\{ \frac{f(x_L) - f(x_U)}{x_L - x_U} \right\}} \quad (6)$$

or

$$x_r = x_L - \frac{f(x_L)}{\left\{ \frac{f(x_U) - f(x_L)}{x_U - x_L} \right\}} \quad (7)$$

Observe the resemblance of Equations (6) and (7) to the secant method.

### False-Position Algorithm

The steps to apply the false-position method to find the root of the equation  $f(x) = 0$  are as follows.

1. Choose  $x_L$  and  $x_U$  as two guesses for the root such that  $f(x_L)f(x_U) < 0$ , or in other words,  $f(x)$  changes sign between  $x_L$  and  $x_U$ .

2. Estimate the root,  $x_r$ , of the equation  $f(x) = 0$  as

$$x_r = \frac{x_U f(x_L) - x_L f(x_U)}{f(x_L) - f(x_U)}$$

3. Now check the following

If  $f(x_L)f(x_r) < 0$ , then the root lies between  $x_L$  and  $x_r$ ; then  $x_L = x_L$  and  $x_U = x_r$ .

If  $f(x_L)f(x_r) > 0$ , then the root lies between  $x_r$  and  $x_U$ ; then  $x_L = x_r$  and  $x_U = x_U$ .

If  $f(x_L)f(x_r) = 0$ , then the root is  $x_r$ . Stop the algorithm.

4. Find the new estimate of the root

$$x_r = \frac{x_U f(x_L) - x_L f(x_U)}{f(x_L) - f(x_U)}$$

Find the absolute relative approximate error as

$$|\epsilon_a| = \left| \frac{x_r^{new} - x_r^{old}}{x_r^{new}} \right| \times 100$$

where

$x_r^{new}$  = estimated root from present iteration

$x_r^{old}$  = estimated root from previous iteration

5. Compare the absolute relative approximate error  $|\epsilon_a|$  with the pre-specified relative error tolerance  $\epsilon_s$ . If  $|\epsilon_a| > \epsilon_s$ , then go to step 3, else stop the algorithm. Note one should also check whether the number of iterations is more than the maximum number of iterations allowed. If so, one needs to terminate the algorithm and notify the user about it.

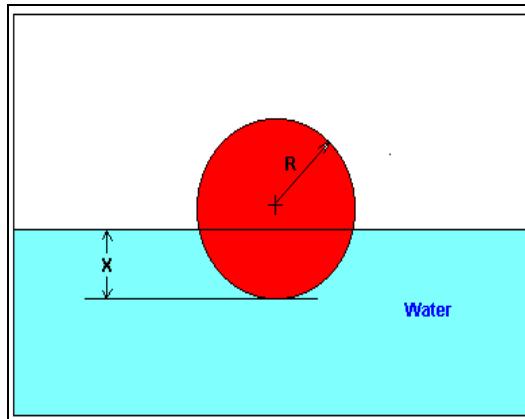
Note that the false-position and bisection algorithms are quite similar. The only difference is the formula used to calculate the new estimate of the root  $x_r$  as shown in steps #2 and #4!

### Example 1

You are working for “DOWN THE TOILET COMPANY” that makes floats for ABC commodes. The floating ball has a specific gravity of 0.6 and has a radius of 5.5cm. You are asked to find the depth to which the ball is submerged when floating in water. The equation that gives the depth  $x$  to which the ball is submerged under water is given by

$$x^3 - 0.165x^2 + 3.993 \times 10^{-4} = 0$$

Use the false-position method of finding roots of equations to find the depth  $x$  to which the ball is submerged under water. 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 third iteration.



**Figure 2** Floating ball problem.

### Solution

From the physics of the problem, the ball would be submerged between  $x = 0$  and  $x = 2R$ , where

$R$  = radius of the ball,

that is

$$0 \leq x \leq 2R$$

$$0 \leq x \leq 2(0.055)$$

$$0 \leq x \leq 0.11$$

Let us assume

$$x_L = 0, x_U = 0.11$$

Check if the function changes sign between  $x_L$  and  $x_U$

$$f(x_L) = f(0) = (0)^3 - 0.165(0)^2 + 3.993 \times 10^{-4} = 3.993 \times 10^{-4}$$

$$f(x_U) = f(0.11) = (0.11)^3 - 0.165(0.11)^2 + 3.993 \times 10^{-4} = -2.662 \times 10^{-4}$$

Hence

$$f(x_L)f(x_U) = f(0)f(0.11) = (3.993 \times 10^{-4})(-2.662 \times 10^{-4}) < 0$$

Therefore, there is at least one root between  $x_L$  and  $x_U$ , that is between 0 and 0.11.

Iteration 1

The estimate of the root is

$$\begin{aligned} x_r &= \frac{x_U f(x_L) - x_L f(x_U)}{f(x_L) - f(x_U)} \\ &= \frac{0.11 \times 3.993 \times 10^{-4} - 0 \times (-2.662 \times 10^{-4})}{3.993 \times 10^{-4} - (-2.662 \times 10^{-4})} \\ &= 0.0660 \end{aligned}$$

$$\begin{aligned} f(x_r) &= f(0.0660) \\ &= (0.0660)^3 - 0.165(0.0660)^2 + (3.993 \times 10^{-4}) \\ &= -3.1944 \times 10^{-5} \end{aligned}$$

$$f(x_L)f(x_r) = f(0)f(0.0660) = (+)(-) < 0$$

Hence, the root is bracketed between  $x_L$  and  $x_r$ , that is, between 0 and 0.0660. So, the lower and upper limits of the new bracket are  $x_L = 0$ ,  $x_U = 0.0660$ , respectively.

### Iteration 2

The estimate of the root is

$$\begin{aligned} x_r &= \frac{x_U f(x_L) - x_L f(x_U)}{f(x_L) - f(x_U)} \\ &= \frac{0.0660 \times 3.993 \times 10^{-4} - 0 \times (-3.1944 \times 10^{-5})}{3.993 \times 10^{-4} - (-3.1944 \times 10^{-5})} \\ &= 0.0611 \end{aligned}$$

The absolute relative approximate error for this iteration is

$$\epsilon_a = \left| \frac{0.0611 - 0.0660}{0.0611} \right| \times 100 \approx 8\%$$

$$\begin{aligned} f(x_r) &= f(0.0611) \\ &= (0.0611)^3 - 0.165(0.0611)^2 + (3.993 \times 10^{-4}) \\ &= 1.1320 \times 10^{-5} \end{aligned}$$

$$f(x_L)f(x_r) = f(0)f(0.0611) = (+)(+) > 0$$

Hence, the lower and upper limits of the new bracket are  $x_L = 0.0611$ ,  $x_U = 0.0660$ , respectively.

### Iteration 3

The estimate of the root is

$$\begin{aligned} x_r &= \frac{x_U f(x_L) - x_L f(x_U)}{f(x_L) - f(x_U)} \\ &= \frac{0.0660 \times 1.132 \times 10^{-5} - 0.0611 \times (-3.1944 \times 10^{-5})}{1.132 \times 10^{-5} - (-3.1944 \times 10^{-5})} \\ &= 0.0624 \end{aligned}$$

The absolute relative approximate error for this iteration is

$$\epsilon_a = \left| \frac{0.0624 - 0.0611}{0.0624} \right| \times 100 \approx 2.05\%$$

$$f(x_r) = -1.1313 \times 10^{-7}$$

$$f(x_L)f(x_r) = f(0.0611)f(0.0624) = (+)(-) < 0$$

Hence, the lower and upper limits of the new bracket are  $x_L = 0.0611$ ,  $x_U = 0.0624$

All iterations results are summarized in Table 1. To find how many significant digits are at least correct in the last iterative value

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

$$2.05 \leq 0.5 \times 10^{2-m}$$

$$m \leq 1.387$$

The number of significant digits at least correct in the estimated root of 0.0624 at the end of 3<sup>rd</sup> iteration is 1.

**Table 1** Root of  $f(x) = x^3 - 0.165x^2 + 3.993 \times 10^{-4} = 0$  for false-position method.

Iteration	$x_L$	$x_U$	$x_r$	$ \epsilon_a  \%$	$f(x_m)$
1	0.0000	0.1100	0.0660	----	$-3.1944 \times 10^{-5}$
2	0.0000	0.0660	0.0611	8.00	$-1.1320 \times 10^{-5}$
3	0.0611	0.0660	0.0624	2.05	$-1.1313 \times 10^{-7}$

### Example 2

Find the root of  $f(x) = (x - 4)^2(x + 2) = 0$ , using the initial guesses of  $x_L = -2.5$  and  $x_U = -1.0$ , and a pre-specified tolerance of  $\epsilon_s = 0.1\%$ .

### Solution

The individual iterations are not shown for this example, but the results are summarized in Table 2. It takes five iterations to meet the pre-specified tolerance.

**Table 2** Root of  $f(x) = (x - 4)^2(x + 2) = 0$  for false-position method.

Iteration	$x_L$	$x_U$	$f(x_L)$	$f(x_U)$	$x_r$	$ \epsilon_a  \%$	$f(x_m)$
1	-2.5	-1	-21.13	25.00	-1.813	N/A	6.319
2	-2.5	-1.813	-21.13	6.319	-1.971	8.024	1.028
3	-2.5	-1.971	-21.13	1.028	-1.996	1.229	0.1542
4	-2.5	-1.996	-21.13	0.1542	-1.999	0.1828	0.02286
5	-2.5	-1.999	-21.13	0.02286	-2.000	0.02706	0.003383

To find how many significant digits are at least correct in the last iterative answer,

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

$$0.02706 \leq 0.5 \times 10^{2-m}$$

$$m \leq 3.2666$$

Hence, at least 3 significant digits can be trusted to be accurate at the end of the fifth iteration.

---

### FALSE-POSITION METHOD OF SOLVING A NONLINEAR EQUATION

---

Topic	False-Position Method of Solving a Nonlinear Equation
Summary	Textbook Chapter of False-Position Method
Major	General Engineering
Authors	Duc Nguyen
Date	September 4, 2012

---

# Chapter 04.01

## Introduction

After reading this chapter, you should be able to

1. *define what a matrix is.*
2. *identify special types of matrices, and*
3. *identify when two matrices are equal.*

### What does a matrix look like?

Matrices are everywhere. If you have used a spreadsheet such as Excel or written numbers in a table, you have used a matrix. Matrices make presentation of numbers clearer and make calculations easier to program. Look at the matrix below about the sale of tires in a Blowout'r's store – given by quarter and make of tires.

	Q1	Q2	Q3	Q4
Tirestone	25	20	3	2
Michigan	5	10	15	25
Copper	6	16	7	27

If one wants to know how many *Copper* tires were sold in *Quarter 4*, we go along the row *Copper* and column *Q4* and find that it is 27.

### So what is a matrix?

A *matrix* is a rectangular array of elements. The elements can be symbolic expressions or/and numbers. Matrix  $[A]$  is denoted by

$$[A] = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}$$

Row  $i$  of  $[A]$  has  $n$  elements and is

$$\begin{bmatrix} a_{i1} & a_{i2} \dots a_{in} \end{bmatrix}$$

and column  $j$  of  $[A]$  has  $m$  elements and is

$$\begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{bmatrix}$$

Each matrix has rows and columns and this defines the size of the matrix. If a matrix  $[A]$  has  $m$  rows and  $n$  columns, the size of the matrix is denoted by  $m \times n$ . The matrix  $[A]$  may also be denoted by  $[A]_{m \times n}$  to show that  $[A]$  is a matrix with  $m$  rows and  $n$  columns.

Each entry in the matrix is called the entry or element of the matrix and is denoted by  $a_{ij}$  where  $i$  is the row number and  $j$  is the column number of the element.

The matrix for the tire sales example could be denoted by the matrix  $[A]$  as

$$[A] = \begin{bmatrix} 25 & 20 & 3 & 2 \\ 5 & 10 & 15 & 25 \\ 6 & 16 & 7 & 27 \end{bmatrix}.$$

There are 3 rows and 4 columns, so the size of the matrix is  $3 \times 4$ . In the above  $[A]$  matrix,  $a_{34} = 27$ .

### What are the special types of matrices?

**Vector:** A vector is a matrix that has only one row or one column. There are two types of vectors – row vectors and column vectors.

#### Row Vector:

If a matrix  $[B]$  has one row, it is called a row vector  $[B] = [b_1 \ b_2 \ \dots \ b_n]$  and  $n$  is the dimension of the row vector.

#### Example 1

Give an example of a row vector.

#### Solution

$$[B] = [25 \ 20 \ 3 \ 2 \ 0]$$

is an example of a row vector of dimension 5.

#### Column vector:

If a matrix  $[C]$  has one column, it is called a column vector

$$[C] = \begin{bmatrix} c_1 \\ \vdots \\ c_m \end{bmatrix}$$

and  $m$  is the dimension of the vector.

### Example 2

Give an example of a column vector.

#### Solution

$$[C] = \begin{bmatrix} 25 \\ 5 \\ 6 \end{bmatrix}$$

is an example of a column vector of dimension 3.

### Submatrix:

If some row(s) or/and column(s) of a matrix  $[A]$  are deleted (no rows or columns may be deleted), the remaining matrix is called a submatrix of  $[A]$ .

### Example 3

Find some of the submatrices of the matrix

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

#### Solution

$$\begin{bmatrix} 4 & 6 & 2 \\ 3 & -1 & 2 \end{bmatrix}, \begin{bmatrix} 4 & 6 \\ 3 & -1 \end{bmatrix}, [4 \ 6 \ 2], [4], \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

are some of the submatrices of  $[A]$ . Can you find other submatrices of  $[A]$ ?

### Square matrix:

If the number of rows  $m$  of a matrix is equal to the number of columns  $n$  of a matrix  $[A]$ , that is,  $m = n$ , then  $[A]$  is called a square matrix. The entries  $a_{11}, a_{22}, \dots, a_{nn}$  are called the *diagonal elements* of a square matrix. Sometimes the diagonal of the matrix is also called the *principal or main of the matrix*.

### Example 4

Give an example of a square matrix.

**Solution**

$$[A] = \begin{bmatrix} 25 & 20 & 3 \\ 5 & 10 & 15 \\ 6 & 15 & 7 \end{bmatrix}$$

is a square matrix as it has the same number of rows and columns, that is, 3. The diagonal elements of  $[A]$  are  $a_{11} = 25$ ,  $a_{22} = 10$ ,  $a_{33} = 7$ .

**Upper triangular matrix:**

A  $n \times n$  matrix for which  $a_{ij} = 0$ ,  $i > j$  for all  $i, j$  is called an upper triangular matrix. That is, all the elements below the diagonal entries are zero.

**Example 5**

Give an example of an upper triangular matrix.

**Solution**

$$[A] = \begin{bmatrix} 10 & -7 & 0 \\ 0 & -0.001 & 6 \\ 0 & 0 & 15005 \end{bmatrix}$$

is an upper triangular matrix.

**Lower triangular matrix:**

A  $n \times n$  matrix for which  $a_{ij} = 0$ ,  $j > i$  for all  $i, j$  is called a lower triangular matrix. That is, all the elements above the diagonal entries are zero.

**Example 6**

Give an example of a lower triangular matrix.

**Solution**

$$[A] = \begin{bmatrix} 1 & 0 & 0 \\ 0.3 & 1 & 0 \\ 0.6 & 2.5 & 1 \end{bmatrix}$$

is a lower triangular matrix.

**Diagonal matrix:**

A square matrix with all non-diagonal elements equal to zero is called a diagonal matrix, that is, only the diagonal entries of the square matrix can be non-zero, ( $a_{ij} = 0$ ,  $i \neq j$ ).

**Example 7**

Give examples of a diagonal matrix.

**Solution**

$$[A] = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 2.1 & 0 \\ 0 & 0 & 5 \end{bmatrix}$$

is a diagonal matrix.

Any or all the diagonal entries of a diagonal matrix can be zero. For example

$$[A] = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 2.1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

is also a diagonal matrix.

**Identity matrix:**

A diagonal matrix with all diagonal elements equal to 1 is called an identity matrix, ( $a_{ij} = 0$ ,  $i \neq j$  for all  $i, j$  and  $a_{ii} = 1$  for all  $i$ ).

**Example 8**

Give an example of an identity matrix.

**Solution**

$$[A] = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

is an identity matrix.

**Zero matrix:**

A matrix whose all entries are zero is called a zero matrix, ( $a_{ij} = 0$  for all  $i$  and  $j$ ).

**Example 9**

Give examples of a zero matrix.

**Solution**

$$[A] = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$[B] = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$[C] = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$[D] = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$$

are all examples of a zero matrix.

### Tridiagonal matrices:

A tridiagonal matrix is a square matrix in which all elements not on the following are zero - the major diagonal, the diagonal above the major diagonal, and the diagonal below the major diagonal.

### Example 10

Give an example of a tridiagonal matrix.

#### Solution

$$[A] = \begin{bmatrix} 2 & 4 & 0 & 0 \\ 2 & 3 & 9 & 0 \\ 0 & 0 & 5 & 2 \\ 0 & 0 & 3 & 6 \end{bmatrix}$$

is a tridiagonal matrix.

### Do non-square matrices have diagonal entries?

Yes, for a  $m \times n$  matrix  $[A]$ , the diagonal entries are  $a_{11}, a_{22}, \dots, a_{k-1,k-1}, a_{kk}$  where  $k = \min\{m, n\}$ .

### Example 11

What are the diagonal entries of

$$[A] = \begin{bmatrix} 3.2 & 5 \\ 6 & 7 \\ 2.9 & 3.2 \\ 5.6 & 7.8 \end{bmatrix}$$

#### Solution

The diagonal elements of  $[A]$  are  $a_{11} = 3.2$  and  $a_{22} = 7$ .

### Diagonally Dominant Matrix:

A  $n \times n$  square matrix  $[A]$  is a diagonally dominant matrix if

$$|a_{ii}| \geq \sum_{\substack{j=1 \\ i \neq j}}^n |a_{ij}| \text{ for } i = 1, 2, \dots, n \text{ and}$$

$$|a_{ii}| > \sum_{\substack{j=1 \\ i \neq j}}^n |a_{ij}| \text{ for at least one } i,$$

that is, for each row, the absolute value of the diagonal element is greater than or equal to the sum of the absolute values of the rest of the elements of that row, and that the inequality is strictly greater than for at least one row. Diagonally dominant matrices are important in ensuring convergence in iterative schemes of solving simultaneous linear equations.

### Example 12

Give examples of diagonally dominant matrices and not diagonally dominant matrices.

#### Solution

$$[A] = \begin{bmatrix} 15 & 6 & 7 \\ 2 & -4 & -2 \\ 3 & 2 & 6 \end{bmatrix}$$

is a diagonally dominant matrix as

$$\begin{aligned} |a_{11}| &= |15| = 15 \geq |a_{12}| + |a_{13}| = |6| + |7| = 13 \\ |a_{22}| &= |-4| = 4 \geq |a_{21}| + |a_{23}| = |2| + |-2| = 4 \\ |a_{33}| &= |6| = 6 \geq |a_{31}| + |a_{32}| = |3| + |2| = 5 \end{aligned}$$

and for at least one row, that is Rows 1 and 3 in this case, the inequality is a strictly greater than inequality.

$$[B] = \begin{bmatrix} -15 & 6 & 9 \\ 2 & -4 & 2 \\ 3 & -2 & 5.001 \end{bmatrix}$$

is a diagonally dominant matrix as

$$\begin{aligned} |b_{11}| &= |-15| = 15 \geq |b_{12}| + |b_{13}| = |6| + |9| = 15 \\ |b_{22}| &= |-4| = 4 \geq |b_{21}| + |b_{23}| = |2| + |2| = 4 \\ |b_{33}| &= |5.001| = 5.001 \geq |b_{31}| + |b_{32}| = |3| + |-2| = 5 \end{aligned}$$

The inequalities are satisfied for all rows and it is satisfied strictly greater than for at least one row (in this case it is Row 3).

$$[C] = \begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix}$$

is not diagonally dominant as

$$|c_{22}| = |8| = 8 \leq |c_{21}| + |c_{23}| = |64| + |1| = 65$$

### When are two matrices considered to be equal?

Two matrices  $[A]$  and  $[B]$  are equal if the size of  $[A]$  and  $[B]$  is the same (number of rows and columns of  $[A]$  are same as that of  $[B]$ ) and  $a_{ij} = b_{ij}$  for all  $i$  and  $j$ .

### Example 13

What would make

$$[A] = \begin{bmatrix} 2 & 3 \\ 6 & 7 \end{bmatrix}$$

to be equal to

$$[B] = \begin{bmatrix} b_{11} & 3 \\ 6 & b_{22} \end{bmatrix}$$

### Solution

The two matrices  $[A]$  and  $[B]$  could be equal if  $b_{11} = 2$  and  $b_{22} = 7$ .

### Key Terms:

- Matrix*
- Vector*
- Submatrix*
- Square matrix*
- Equal matrices*
- Zero matrix*
- Identity matrix*
- Diagonal matrix*
- Upper triangular matrix*
- Lower triangular matrix*
- Tri-diagonal matrix*
- Diagonally dominant matrix*

## Chapter 04.02

# Vectors

After reading this chapter, you should be able to:

1. define a vector,
2. add and subtract vectors,
3. find linear combinations of vectors and their relationship to a set of equations,
4. explain what it means to have a linearly independent set of vectors, and
5. find the rank of a set of vectors.

### What is a vector?

A vector is a collection of numbers in a definite order. If it is a collection of  $n$  numbers, it is called a  $n$ -dimensional vector. So the vector  $\vec{A}$  given by

$$\vec{A} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$$

is a  $n$ -dimensional column vector with  $n$  components,  $a_1, a_2, \dots, a_n$ . The above is a column vector. A row vector  $[B]$  is of the form  $\vec{B} = [b_1, b_2, \dots, b_n]$  where  $\vec{B}$  is a  $n$ -dimensional row vector with  $n$  components  $b_1, b_2, \dots, b_n$ .

### Example 1

Give an example of a 3-dimensional column vector.

#### Solution

Assume a point in space is given by its  $(x, y, z)$  coordinates. Then if the value of  $x = 3, y = 2, z = 5$ , the column vector corresponding to the location of the points is

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \\ 5 \end{bmatrix}.$$

### When are two vectors equal?

Two vectors  $\vec{A}$  and  $\vec{B}$  are equal if they are of the same dimension and if their corresponding components are equal.

Given

$$\vec{A} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$$

and

$$\vec{B} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

then  $\vec{A} = \vec{B}$  if  $a_i = b_i$ ,  $i = 1, 2, \dots, n$ .

### Example 2

What are the values of the unknown components in  $\vec{B}$  if

$$\vec{A} = \begin{bmatrix} 2 \\ 3 \\ 4 \\ 1 \end{bmatrix}$$

and

$$\vec{B} = \begin{bmatrix} b_1 \\ 3 \\ 4 \\ b_4 \end{bmatrix}$$

and  $\vec{A} = \vec{B}$ .

### Solution

$$b_1 = 2, b_4 = 1$$

### How do you add two vectors?

Two vectors can be added only if they are of the same dimension and the addition is given by

$$[A] + [B] = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

$$= \begin{bmatrix} a_1 + b_1 \\ a_2 + b_2 \\ \vdots \\ a_n + b_n \end{bmatrix}$$

**Example 3**

Add the two vectors

$$\vec{A} = \begin{bmatrix} 2 \\ 3 \\ 4 \\ 1 \end{bmatrix}$$

and

$$\vec{B} = \begin{bmatrix} 5 \\ -2 \\ 3 \\ 7 \end{bmatrix}$$

**Solution**

$$\vec{A} + \vec{B} = \begin{bmatrix} 2 \\ 3 \\ 4 \\ 1 \end{bmatrix} + \begin{bmatrix} 5 \\ -2 \\ 3 \\ 7 \end{bmatrix}$$

$$= \begin{bmatrix} 2+5 \\ 3-2 \\ 4+3 \\ 1+7 \end{bmatrix}$$

$$= \begin{bmatrix} 7 \\ 1 \\ 7 \\ 8 \end{bmatrix}$$

**Example 4**

A store sells three brands of tires: Firestone, Michigan and Copper. In quarter 1, the sales are given by the column vector

$$\vec{A}_1 = \begin{bmatrix} 25 \\ 5 \\ 6 \end{bmatrix}$$

where the rows represent the three brands of tires sold – Firestone, Michigan and Copper respectively. In quarter 2, the sales are given by

$$\vec{A}_2 = \begin{bmatrix} 20 \\ 10 \\ 6 \end{bmatrix}$$

What is the total sale of each brand of tire in the first half of the year?

**Solution**

The total sales would be given by

$$\begin{aligned} \vec{C} &= \vec{A}_1 + \vec{A}_2 \\ &= \begin{bmatrix} 25 \\ 5 \\ 6 \end{bmatrix} + \begin{bmatrix} 20 \\ 10 \\ 6 \end{bmatrix} \\ &= \begin{bmatrix} 25 + 20 \\ 5 + 10 \\ 6 + 6 \end{bmatrix} \\ &= \begin{bmatrix} 45 \\ 15 \\ 12 \end{bmatrix} \end{aligned}$$

So the number of Firestone tires sold is 45, Michigan is 15 and Copper is 12 in the first half of the year.

**What is a null vector?**

A null vector (also called zero vector) is where all the components of the vector are zero.

**Example 5**

Give an example of a null vector or zero vector.

**Solution**

The vector

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

is an example of a zero or null vector.

### What is a unit vector?

A unit vector  $\vec{U}$  is defined as

$$\vec{U} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

where

$$\sqrt{u_1^2 + u_2^2 + u_3^2 + \dots + u_n^2} = 1$$

### Example 6

Give examples of 3-dimensional unit column vectors.

#### Solution

Examples include

$$\begin{bmatrix} \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \text{etc.}$$

### How do you multiply a vector by a scalar?

If  $k$  is a scalar and  $\vec{A}$  is a  $n$ -dimensional vector, then

$$\begin{aligned} k\vec{A} &= k \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \\ &= \begin{bmatrix} ka_1 \\ ka_2 \\ \vdots \\ ka_n \end{bmatrix} \end{aligned}$$

**Example 7**

What is  $2\vec{A}$  if

$$\vec{A} = \begin{bmatrix} 25 \\ 20 \\ 5 \end{bmatrix}$$

**Solution**

$$\begin{aligned} 2\vec{A} &= 2 \begin{bmatrix} 25 \\ 20 \\ 5 \end{bmatrix} \\ &= \begin{bmatrix} 2 \times 25 \\ 2 \times 20 \\ 2 \times 5 \end{bmatrix} \\ &= \begin{bmatrix} 50 \\ 40 \\ 10 \end{bmatrix} \end{aligned}$$

**Example 8**

A store sells three brands of tires: Firestone, Michelin and Copper. In quarter 1, the sales are given by the column vector

$$\vec{A} = \begin{bmatrix} 25 \\ 25 \\ 6 \end{bmatrix}$$

If the goal is to increase the sales of all tires by at least 25% in the next quarter, how many of each brand should be sold?

**Solution**

Since the goal is to increase the sales by 25%, one would multiply the  $\vec{A}$  vector by 1.25,

$$\begin{aligned} \vec{B} &= 1.25 \begin{bmatrix} 25 \\ 25 \\ 6 \end{bmatrix} \\ &= \begin{bmatrix} 31.25 \\ 31.25 \\ 7.5 \end{bmatrix} \end{aligned}$$

Since the number of tires must be an integer, we can say that the goal of sales is

$$\vec{B} = \begin{bmatrix} 32 \\ 32 \\ 8 \end{bmatrix}$$

### What do you mean by a linear combination of vectors?

Given

$$\vec{A}_1, \vec{A}_2, \dots, \vec{A}_m$$

as  $m$  vectors of same dimension  $n$ , and if  $k_1, k_2, \dots, k_m$  are scalars, then

$$k_1 \vec{A}_1 + k_2 \vec{A}_2 + \dots + k_m \vec{A}_m$$

is a linear combination of the  $m$  vectors.

### Example 9

Find the linear combinations

a)  $\vec{A} - \vec{B}$  and

b)  $\vec{A} + \vec{B} - 3\vec{C}$

where

$$\vec{A} = \begin{bmatrix} 2 \\ 3 \\ 6 \end{bmatrix}, \vec{B} = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}, \vec{C} = \begin{bmatrix} 10 \\ 1 \\ 2 \end{bmatrix}$$

### Solution

$$\text{a) } \vec{A} - \vec{B} = \begin{bmatrix} 2 \\ 3 \\ 6 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}$$

$$= \begin{bmatrix} 2-1 \\ 3-1 \\ 6-2 \end{bmatrix}$$

$$= \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix}$$

$$\text{b) } \vec{A} + \vec{B} - 3\vec{C} = \begin{bmatrix} 2 \\ 3 \\ 6 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} - 3 \begin{bmatrix} 10 \\ 1 \\ 2 \end{bmatrix}$$

$$= \begin{bmatrix} 2+1-30 \\ 3+1-3 \\ 6+2-6 \end{bmatrix}$$

$$= \begin{bmatrix} -27 \\ 1 \\ 2 \end{bmatrix}$$

### What do you mean by vectors being linearly independent?

A set of vectors  $\vec{A}_1, \vec{A}_2, \dots, \vec{A}_m$  are considered to be linearly independent if

$$k_1 \vec{A}_1 + k_2 \vec{A}_2 + \dots + k_m \vec{A}_m = \vec{0}$$

has only one solution of

$$k_1 = k_2 = \dots = k_m = 0$$

### Example 10

Are the three vectors

$$\vec{A}_1 = \begin{bmatrix} 25 \\ 64 \\ 144 \end{bmatrix}, \vec{A}_2 = \begin{bmatrix} 5 \\ 8 \\ 12 \end{bmatrix}, \vec{A}_3 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

linearly independent?

#### Solution

Writing the linear combination of the three vectors

$$k_1 \begin{bmatrix} 25 \\ 64 \\ 144 \end{bmatrix} + k_2 \begin{bmatrix} 5 \\ 8 \\ 12 \end{bmatrix} + k_3 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

gives

$$\begin{bmatrix} 25k_1 + 5k_2 + k_3 \\ 64k_1 + 8k_2 + k_3 \\ 144k_1 + 12k_2 + k_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

The above equations have only one solution,  $k_1 = k_2 = k_3 = 0$ . However, how do we show that this is the only solution? This is shown below.

The above equations are

$$25k_1 + 5k_2 + k_3 = 0 \quad (1)$$

$$64k_1 + 8k_2 + k_3 = 0 \quad (2)$$

$$144k_1 + 12k_2 + k_3 = 0 \quad (3)$$

Subtracting Eqn (1) from Eqn (2) gives

$$39k_1 + 3k_2 = 0 \quad (4)$$

$$k_2 = -13k_1 \quad (4)$$

Multiplying Eqn (1) by 8 and subtracting it from Eqn (2) that is first multiplied by 5 gives

$$\begin{aligned} 120k_1 - 3k_3 &= 0 \\ k_3 &= 40k_1 \end{aligned} \tag{5}$$

Remember we found Eqn (4) and Eqn (5) just from Eqns (1) and (2).

Substitution of Eqns (4) and (5) in Eqn (3) for  $k_1$  and  $k_2$  gives

$$\begin{aligned} 144k_1 + 12(-13k_1) + 40k_1 &= 0 \\ 28k_1 &= 0 \\ k_1 &= 0 \end{aligned}$$

This means that  $k_1$  has to be zero, and coupled with (4) and (5),  $k_2$  and  $k_3$  are also zero. So the only solution is  $k_1 = k_2 = k_3 = 0$ . The three vectors hence are linearly independent.

### Example 11

Are the three vectors

$$\vec{A}_1 = \begin{bmatrix} 1 \\ 2 \\ 5 \end{bmatrix}, \vec{A}_2 = \begin{bmatrix} 2 \\ 5 \\ 7 \end{bmatrix}, \vec{A}_3 = \begin{bmatrix} 6 \\ 14 \\ 24 \end{bmatrix}$$

linearly independent?

#### Solution

By inspection,

$$\vec{A}_3 = 2\vec{A}_1 + 2\vec{A}_2$$

or

$$-2\vec{A}_1 - 2\vec{A}_2 + \vec{A}_3 = \vec{0}$$

So the linear combination

$$k_1\vec{A}_1 + k_2\vec{A}_2 + k_3\vec{A}_3 = \vec{0}$$

has a non-zero solution

$$k_1 = -2, k_2 = -2, k_3 = 1$$

Hence, the set of vectors is linearly dependent.

What if I cannot prove by inspection, what do I do? Put the linear combination of three vectors equal to the zero vector,

$$k_1 \begin{bmatrix} 1 \\ 2 \\ 5 \end{bmatrix} + k_2 \begin{bmatrix} 2 \\ 5 \\ 7 \end{bmatrix} + k_3 \begin{bmatrix} 6 \\ 14 \\ 24 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

to give

$$k_1 + 2k_2 + 6k_3 = 0 \tag{1}$$

$$2k_1 + 5k_2 + 14k_3 = 0 \tag{2}$$

$$5k_1 + 7k_2 + 24k_3 = 0 \tag{3}$$

Multiplying Eqn (1) by 2 and subtracting from Eqn (2) gives

$$k_2 + 2k_3 = 0$$

$$k_2 = -2k_3 \quad (4)$$

Multiplying Eqn (1) by 2.5 and subtracting from Eqn (2) gives

$$\begin{aligned} -0.5k_1 - k_3 &= 0 \\ k_1 &= -2k_3 \end{aligned} \quad (5)$$

Remember we found Eqn (4) and Eqn (5) just from Eqns (1) and (2).

Substitute Eqn (4) and (5) in Eqn (3) for  $k_1$  and  $k_2$  gives

$$\begin{aligned} 5(-2k_3) + 7(-2k_3) + 24k_3 &= 0 \\ -10k_3 - 14k_3 + 24k_3 &= 0 \\ 0 &= 0 \end{aligned}$$

This means any values satisfying Eqns (4) and (5) will satisfy Eqns (1), (2) and (3) simultaneously.

For example, chose

$$k_3 = 6, \text{ then}$$

$$k_2 = -12 \text{ from Eqn (4), and}$$

$$k_1 = -12 \text{ from Eqn (5).}$$

Hence we have a nontrivial solution of  $[k_1 \ k_2 \ k_3] = [-12 \ -12 \ 6]$ . This implies the three given vectors are linearly dependent. Can you find another nontrivial solution?

What about the following three vectors?

$$\begin{bmatrix} 1 \\ 2 \\ 5 \end{bmatrix}, \begin{bmatrix} 2 \\ 5 \\ 7 \end{bmatrix}, \begin{bmatrix} 6 \\ 14 \\ 25 \end{bmatrix}$$

Are they linearly dependent or linearly independent?

Note that the only difference between this set of vectors and the previous one is the third entry in the third vector. Hence, equations (4) and (5) are still valid. What conclusion do you draw when you plug in equations (4) and (5) in the third equation:  $5k_1 + 7k_2 + 25k_3 = 0$ ? What has changed?

### Example 12

Are the three vectors

$$\vec{A}_1 = \begin{bmatrix} 25 \\ 64 \\ 89 \end{bmatrix}, \vec{A}_2 = \begin{bmatrix} 5 \\ 8 \\ 13 \end{bmatrix}, \vec{A}_3 = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}$$

linearly independent?

### Solution

Writing the linear combination of the three vectors and equating to zero vector

$$k_1 \begin{bmatrix} 25 \\ 64 \\ 89 \end{bmatrix} + k_2 \begin{bmatrix} 5 \\ 8 \\ 13 \end{bmatrix} + k_3 \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

gives

$$\begin{bmatrix} 25k_1 + 5k_2 + k_3 \\ 64k_1 + 8k_2 + k_3 \\ 89k_1 + 13k_2 + 2k_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

In addition to  $k_1 = k_2 = k_3 = 0$ , one can find other solutions for which  $k_1, k_2, k_3$  are not equal to zero. For example,  $k_1 = 1, k_2 = -13, k_3 = 40$  is also a solution as

$$1 \begin{bmatrix} 25 \\ 64 \\ 89 \end{bmatrix} - 13 \begin{bmatrix} 5 \\ 8 \\ 13 \end{bmatrix} + 40 \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Hence  $\vec{A}_1, \vec{A}_2, \vec{A}_3$  are linearly dependent.

### What do you mean by the rank of a set of vectors?

From a set of  $n$ -dimensional vectors, the maximum number of linearly independent vectors in the set is called the rank of the set of vectors. *Note that the rank of the vectors can never be greater than the vectors dimension.*

### Example 13

What is the rank of

$$\vec{A}_1 = \begin{bmatrix} 25 \\ 64 \\ 144 \end{bmatrix}, \vec{A}_2 = \begin{bmatrix} 5 \\ 8 \\ 12 \end{bmatrix}, \vec{A}_3 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}?$$

### Solution

Since we found in [Example 2.10](#) that  $\vec{A}_1, \vec{A}_2, \vec{A}_3$  are linearly independent, the rank of the set of vectors  $\vec{A}_1, \vec{A}_2, \vec{A}_3$  is 3. If we were given another vector  $\vec{A}_4$ , the rank of the set of the vectors  $\vec{A}_1, \vec{A}_2, \vec{A}_3, \vec{A}_4$  would still be 3 as the rank of a set of vectors is always less than or equal to the dimension of the vectors and that at least  $\vec{A}_1, \vec{A}_2, \vec{A}_3$  are linearly independent.

### Example 14

What is the rank of

$$\vec{A}_1 = \begin{bmatrix} 25 \\ 64 \\ 89 \end{bmatrix}, \vec{A}_2 = \begin{bmatrix} 5 \\ 8 \\ 13 \end{bmatrix}, \vec{A}_3 = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}?$$

**Solution**

In Example 2.12, we found that  $\vec{A}_1, \vec{A}_2, \vec{A}_3$  are linearly dependent, the rank of  $\vec{A}_1, \vec{A}_2, \vec{A}_3$  is hence not 3, and is less than 3. Is it 2? Let us choose two of the three vectors

$$\vec{A}_1 = \begin{bmatrix} 25 \\ 64 \\ 89 \end{bmatrix}, \vec{A}_2 = \begin{bmatrix} 5 \\ 8 \\ 13 \end{bmatrix}$$

Linear combination of  $\vec{A}_1$  and  $\vec{A}_2$  equal to zero has only one solution – the trivial solution. Therefore, the rank is 2.

**Example 15**

What is the rank of

$$\vec{A}_1 = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}, \vec{A}_2 = \begin{bmatrix} 2 \\ 2 \\ 4 \end{bmatrix}, \vec{A}_3 = \begin{bmatrix} 3 \\ 3 \\ 5 \end{bmatrix}?$$

**Solution**

From inspection,

$$\vec{A}_2 = 2\vec{A}_1,$$

that implies

$$2\vec{A}_1 - \vec{A}_2 + 0\vec{A}_3 = \vec{0}.$$

Hence

$$k_1\vec{A}_1 + k_2\vec{A}_2 + k_3\vec{A}_3 = \vec{0}.$$

has a nontrivial solution.

So  $\vec{A}_1, \vec{A}_2, \vec{A}_3$  are linearly dependent, and hence the rank of the three vectors is not 3. Since

$$\vec{A}_2 = 2\vec{A}_1,$$

$\vec{A}_1$  and  $\vec{A}_2$  are linearly dependent, but

$$k_1\vec{A}_1 + k_3\vec{A}_3 = \vec{0}.$$

has trivial solution as the only solution. So  $\vec{A}_1$  and  $\vec{A}_3$  are linearly independent. The rank of the above three vectors is 2.

**Prove that if a set of vectors contains the null vector, the set of vectors is linearly dependent.**

Let  $\vec{A}_1, \vec{A}_2, \dots, \vec{A}_m$  be a set of  $n$ -dimensional vectors, then

$$k_1\vec{A}_1 + k_2\vec{A}_2 + \dots + k_m\vec{A}_m = \vec{0}$$

is a linear combination of the  $m$  vectors. Then assuming if  $\vec{A}_1$  is the zero or null vector, any value of  $k_1$  coupled with  $k_2 = k_3 = \dots = k_m = 0$  will satisfy the above equation. Hence, the

set of vectors is linearly dependent as more than one solution exists.

**Prove that if a set of  $m$  vectors is linearly independent, then a subset of the  $m$  vectors also has to be linearly independent.**

Let this subset of vectors be

$$\vec{A}_{a1}, \vec{A}_{a2}, \dots, \vec{A}_{ap}$$

where  $p < m$ .

Then if this subset of vectors is linearly dependent, the linear combination

$$k_1 \vec{A}_{a1} + k_2 \vec{A}_{a2} + \dots + k_p \vec{A}_{ap} = \vec{0}$$

has a non-trivial solution.

So

$$k_1 \vec{A}_{a1} + k_2 \vec{A}_{a2} + \dots + k_p \vec{A}_{ap} + 0 \vec{A}_{a(p+1)} + \dots + 0 \vec{A}_{am} = \vec{0}$$

also has a non-trivial solution too, where  $\vec{A}_{a(p+1)}, \dots, \vec{A}_{am}$  are the rest of the  $(m-p)$  vectors.

However, this is a contradiction. Therefore, a subset of linearly independent vectors cannot be linearly dependent.

**Prove that if a set of vectors is linearly dependent, then at least one vector can be written as a linear combination of others.**

Let  $\vec{A}_1, \vec{A}_2, \dots, \vec{A}_m$  be linearly dependent set of vectors, then there exists a set of scalars  $k_1, \dots, k_m$  not all of which are zero for the linear combination equation

$$k_1 \vec{A}_1 + k_2 \vec{A}_2 + \dots + k_m \vec{A}_m = \vec{0}.$$

Let  $k_p$  be one of the non-zero values of  $k_i, i = 1, \dots, m$ , that is,  $k_p \neq 0$ , then

$$\vec{A}_p = -\frac{k_2}{k_p} \vec{A}_2 - \dots - \frac{k_{p-1}}{k_p} \vec{A}_{p-1} - \frac{k_{p+1}}{k_p} \vec{A}_{p+1} - \dots - \frac{k_m}{k_p} \vec{A}_m.$$

and that proves the theorem.

**Prove that if the dimension of a set of vectors is less than the number of vectors in the set, then the set of vectors is linearly dependent.**

Can you prove it?

**How can vectors be used to write simultaneous linear equations?**

If a set of  $m$  simultaneous linear equations with  $n$  unknowns is written as

$$a_{11}x_1 + \dots + a_{1n}x_n = c_1$$

$$a_{21}x_1 + \dots + a_{2n}x_n = c_2$$

$$\vdots \quad \vdots$$

$$\vdots \quad \vdots$$

$$a_{m1}x_1 + \dots + a_{mn}x_n = c_n$$

where

$x_1, x_2, \dots, x_n$  are the unknowns, then in the vector notation they can be written as

$$x_1 \vec{A}_1 + x_2 \vec{A}_2 + \dots + x_n \vec{A}_n = \vec{C}$$

where

$$\vec{A}_1 = \begin{bmatrix} a_{11} \\ \vdots \\ a_{m1} \end{bmatrix}$$

where

$$\vec{A}_1 = \begin{bmatrix} a_{11} \\ \vdots \\ a_{m1} \end{bmatrix}$$

$$\vec{A}_2 = \begin{bmatrix} a_{12} \\ \vdots \\ a_{m2} \end{bmatrix}$$

$$\vec{A}_n = \begin{bmatrix} a_{1n} \\ \vdots \\ a_{mn} \end{bmatrix}$$

$$\vec{C}_1 = \begin{bmatrix} c_1 \\ \vdots \\ c_m \end{bmatrix}$$

The problem now becomes whether you can find the scalars  $x_1, x_2, \dots, x_n$  such that the linear combination

$$x_1 \vec{A}_1 + \dots + x_n \vec{A}_n$$

is equal to the  $\vec{C}$ , that is

$$x_1 \vec{A}_1 + \dots + x_n \vec{A}_n = \vec{C}$$

### Example 16

Write

$$25x_1 + 5x_2 + x_3 = 106.8$$

$$64x_1 + 8x_2 + x_3 = 177.2$$

$$144x_1 + 12x_2 + x_3 = 279.2$$

as a linear combination of set of vectors equal to another vector.

**Solution**

$$\begin{bmatrix} 25x_1 & + 5x_2 & + x_3 \\ 64x_1 & + 8x_2 & + x_3 \\ 144x_1 & + 12x_2 & + x_3 \end{bmatrix} = \begin{bmatrix} 106.8 \\ 177.2 \\ 279.2 \end{bmatrix}$$

$$x_1 \begin{bmatrix} 25 \\ 64 \\ 144 \end{bmatrix} + x_2 \begin{bmatrix} 5 \\ 8 \\ 12 \end{bmatrix} + x_3 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 106.8 \\ 177.2 \\ 279.2 \end{bmatrix}$$

**What is the definition of the dot product of two vectors?**

Let  $\vec{A} = [a_1, a_2, \dots, a_n]$  and  $\vec{B} = [b_1, b_2, \dots, b_n]$  be two  $n$ -dimensional vectors. Then the dot product of the two vectors  $\vec{A}$  and  $\vec{B}$  is defined as

$$\vec{A} \cdot \vec{B} = a_1 b_1 + a_2 b_2 + \dots + a_n b_n = \sum_{i=1}^n a_i b_i$$

A dot product is also called an inner product.

**Example 17**

Find the dot product of the two vectors  $\vec{A} = [4, 1, 2, 3]$  and  $\vec{B} = [3, 1, 7, 2]$ .

**Solution**

$$\begin{aligned} \vec{A} \cdot \vec{B} &= [4, 1, 2, 3] \cdot [3, 1, 7, 2] \\ &= (4)(3) + (1)(1) + (2)(7) + (3)(2) \\ &= 33 \end{aligned}$$

**Example 18**

A product line needs three types of rubber as given in the table below.

Rubber Type	Weight (lbs)	Cost per pound (\$)
A	200	20.23
B	250	30.56
C	310	29.12

Use the definition of a dot product to find the total price of the rubber needed.

**Solution**

The weight vector is given by

$$\vec{W} = [200, 250, 310]$$

and the cost vector is given by

$$\vec{C} = [20.23, 30.56, 29.12].$$

The total cost of the rubber would be the dot product of  $\vec{W}$  and  $\vec{C}$ .

$$\vec{W} \cdot \vec{C} = [200, 250, 310] \cdot [20.23, 30.56, 29.12]$$

$$\begin{aligned} &= (200)(20.23) + (250)(30.56) + (310)(29.12) \\ &= 4046 + 7640 + 9027.2 \\ &= \$20713.20 \end{aligned}$$

**Key Terms:**

*Vector*

*Addition of vectors*

*Rank*

*Dot Product*

*Subtraction of vectors*

*Unit vector*

*Scalar multiplication of vectors*

*Null vector*

*Linear combination of vectors*

*Linearly independent vectors*

## Chapter 04.03

# Binary Matrix Operations

After reading this chapter, you should be able to

1. add, subtract, and multiply matrices, and
2. apply rules of binary operations on matrices.

### How do you add two matrices?

Two matrices  $[A]$  and  $[B]$  can be added only if they are the same size. The addition is then shown as

$$[C] = [A] + [B]$$

where

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

### Example 1

Add the following two matrices.

$$[A] = \begin{bmatrix} 5 & 2 & 3 \\ 1 & 2 & 7 \end{bmatrix} \quad [B] = \begin{bmatrix} 6 & 7 & -2 \\ 3 & 5 & 19 \end{bmatrix}$$

### Solution

$$\begin{aligned} [C] &= [A] + [B] \\ &= \begin{bmatrix} 5 & 2 & 3 \\ 1 & 2 & 7 \end{bmatrix} + \begin{bmatrix} 6 & 7 & -2 \\ 3 & 5 & 19 \end{bmatrix} \\ &= \begin{bmatrix} 5+6 & 2+7 & 3-2 \\ 1+3 & 2+5 & 7+19 \end{bmatrix} \\ &= \begin{bmatrix} 11 & 9 & 1 \\ 4 & 7 & 26 \end{bmatrix} \end{aligned}$$

### Example 2

Blowout r'us store has two store locations  $A$  and  $B$ , and their sales of tires are given by make (in rows) and quarters (in columns) as shown below.

$$[A] = \begin{bmatrix} 25 & 20 & 3 & 2 \\ 5 & 10 & 15 & 25 \\ 6 & 16 & 7 & 27 \end{bmatrix}$$

$$[B] = \begin{bmatrix} 20 & 5 & 4 & 0 \\ 3 & 6 & 15 & 21 \\ 4 & 1 & 7 & 20 \end{bmatrix}$$

where the rows represent the sale of Tirestone, Michigan and Copper tires respectively and the columns represent the quarter number: 1, 2, 3 and 4. What are the total tire sales for the two locations by make and quarter?

### Solution

$$\begin{aligned} [C] &= [A] + [B] \\ &= \begin{bmatrix} 25 & 20 & 3 & 2 \\ 5 & 10 & 15 & 25 \\ 6 & 16 & 7 & 27 \end{bmatrix} + \begin{bmatrix} 20 & 5 & 4 & 0 \\ 3 & 6 & 15 & 21 \\ 4 & 1 & 7 & 20 \end{bmatrix} \\ &= \begin{bmatrix} (25+20) & (20+5) & (3+4) & (2+0) \\ (5+3) & (10+6) & (15+15) & (25+21) \\ (6+4) & (16+1) & (7+7) & (27+20) \end{bmatrix} \\ &= \begin{bmatrix} 45 & 25 & 7 & 2 \\ 8 & 16 & 30 & 46 \\ 10 & 17 & 14 & 47 \end{bmatrix} \end{aligned}$$

So if one wants to know the total number of Copper tires sold in quarter 4 at the two locations, we would look at Row 3 – Column 4 to give  $c_{34} = 47$ .

### How do you subtract two matrices?

Two matrices  $[A]$  and  $[B]$  can be subtracted only if they are the same size. The subtraction is then given by

$$[D] = [A] - [B]$$

Where

$$d_{ij} = a_{ij} - b_{ij}$$

### Example 3

Subtract matrix  $[B]$  from matrix  $[A]$ .

$$[A] = \begin{bmatrix} 5 & 2 & 3 \\ 1 & 2 & 7 \end{bmatrix}$$

$$[B] = \begin{bmatrix} 6 & 7 & -2 \\ 3 & 5 & 19 \end{bmatrix}$$

**Solution**

$$\begin{aligned}
 [D] &= [A] - [B] \\
 &= \begin{bmatrix} 5 & 2 & 3 \\ 1 & 2 & 7 \end{bmatrix} - \begin{bmatrix} 6 & 7 & -2 \\ 3 & 5 & 19 \end{bmatrix} \\
 &= \begin{bmatrix} (5-6) & (2-7) & (3-(-2)) \\ (1-3) & (2-5) & (7-19) \end{bmatrix} \\
 &= \begin{bmatrix} -1 & -5 & 5 \\ -2 & -3 & -12 \end{bmatrix}
 \end{aligned}$$

**Example 4**

Blowout r'us has two store locations  $A$  and  $B$  and their sales of tires are given by make (in rows) and quarters (in columns) as shown below.

$$\begin{aligned}
 [A] &= \begin{bmatrix} 25 & 20 & 3 & 2 \\ 5 & 10 & 15 & 25 \\ 6 & 16 & 7 & 27 \end{bmatrix} \\
 [B] &= \begin{bmatrix} 20 & 5 & 4 & 0 \\ 3 & 6 & 15 & 21 \\ 4 & 1 & 7 & 20 \end{bmatrix}
 \end{aligned}$$

where the rows represent the sale of Tirestone, Michigan and Copper tires respectively and the columns represent the quarter number: 1, 2, 3, and 4. How many more tires did store  $A$  sell than store  $B$  of each brand in each quarter?

**Solution**

$$\begin{aligned}
 [D] &= [A] - [B] \\
 &= \begin{bmatrix} 25 & 20 & 3 & 2 \\ 5 & 10 & 15 & 25 \\ 6 & 16 & 7 & 27 \end{bmatrix} - \begin{bmatrix} 20 & 5 & 4 & 0 \\ 3 & 6 & 15 & 21 \\ 4 & 1 & 7 & 20 \end{bmatrix} \\
 &= \begin{bmatrix} 25-20 & 20-5 & 3-4 & 2-0 \\ 5-3 & 10-6 & 15-15 & 25-21 \\ 6-4 & 16-1 & 7-7 & 27-20 \end{bmatrix} \\
 &= \begin{bmatrix} 5 & 15 & -1 & 2 \\ 2 & 4 & 0 & 4 \\ 2 & 15 & 0 & 7 \end{bmatrix}
 \end{aligned}$$

So if you want to know how many more Copper tires were sold in quarter 4 in store  $A$  than store  $B$ ,  $d_{34} = 7$ . Note that  $d_{13} = -1$  implies that store  $A$  sold 1 less Michigan tire than store  $B$  in quarter 3.

### How do I multiply two matrices?

Two matrices  $[A]$  and  $[B]$  can be multiplied only if the number of columns of  $[A]$  is equal to the number of rows of  $[B]$  to give

$$[C]_{m \times n} = [A]_{m \times p} [B]_{p \times n}$$

If  $[A]$  is a  $m \times p$  matrix and  $[B]$  is a  $p \times n$  matrix, the resulting matrix  $[C]$  is a  $m \times n$  matrix.

So how does one calculate the elements of  $[C]$  matrix?

$$\begin{aligned} c_{ij} &= \sum_{k=1}^p a_{ik} b_{kj} \\ &= a_{i1} b_{1j} + a_{i2} b_{2j} + \dots + a_{ip} b_{pj} \end{aligned}$$

for each  $i = 1, 2, \dots, m$  and  $j = 1, 2, \dots, n$ .

To put it in simpler terms, the  $i^{th}$  row and  $j^{th}$  column of the  $[C]$  matrix in  $[C] = [A][B]$  is calculated by multiplying the  $i^{th}$  row of  $[A]$  by the  $j^{th}$  column of  $[B]$ , that is,

$$\begin{aligned} c_{ij} &= \left[ a_{i1} \ a_{i2} \ \dots \ a_{ip} \right] \begin{bmatrix} b_{1j} \\ b_{2j} \\ \vdots \\ b_{pj} \end{bmatrix} \\ &= a_{i1} b_{1j} + a_{i2} b_{2j} + \dots + a_{ip} b_{pj}. \\ &= \sum_{k=1}^p a_{ik} b_{kj} \end{aligned}$$

### Example 5

Given

$$[A] = \begin{bmatrix} 5 & 2 & 3 \\ 1 & 2 & 7 \end{bmatrix}$$

$$[B] = \begin{bmatrix} 3 & -2 \\ 5 & -8 \\ 9 & -10 \end{bmatrix}$$

Find

$$[C] = [A][B]$$

### Solution

$c_{12}$  can be found by multiplying the first row of  $[A]$  by the second column of  $[B]$ ,

$$c_{12} = [5 \ 2 \ 3] \begin{bmatrix} -2 \\ -8 \\ -10 \end{bmatrix}$$

$$\begin{aligned}
 &= (5)(-2) + (2)(-8) + (3)(-10) \\
 &= -56
 \end{aligned}$$

Similarly, one can find the other elements of  $[C]$  to give

$$[C] = \begin{bmatrix} 52 & -56 \\ 76 & -88 \end{bmatrix}$$

### Example 6

Blowout r'us store location  $A$  and the sales of tires are given by make (in rows) and quarters (in columns) as shown below

$$[A] = \begin{bmatrix} 25 & 20 & 3 & 2 \\ 5 & 10 & 15 & 25 \\ 6 & 16 & 7 & 27 \end{bmatrix}$$

where the rows represent the sale of Tirestone, Michigan and Copper tires respectively and the columns represent the quarter number: 1, 2, 3, and 4. Find the per quarter sales of store  $A$  if the following are the prices of each tire.

Tirestone = \$33.25

Michigan = \$40.19

Copper = \$25.03

### Solution

The answer is given by multiplying the price matrix by the quantity of sales of store  $A$ . The price matrix is  $[33.25 \ 40.19 \ 25.03]$ , so the per quarter sales of store  $A$  would be given by

$$[C] = [33.25 \ 40.19 \ 25.03] \begin{bmatrix} 25 & 20 & 3 & 2 \\ 5 & 10 & 15 & 25 \\ 6 & 16 & 7 & 27 \end{bmatrix}$$

$$\begin{aligned}
 c_{ij} &= \sum_{k=1}^3 a_{ik} b_{kj} \\
 c_{11} &= \sum_{k=1}^3 a_{1k} b_{k1} \\
 &= a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} \\
 &= (33.25)(25) + (40.19)(5) + (25.03)(6) \\
 &= \$1182.38
 \end{aligned}$$

Similarly

$$c_{12} = \$1467.38$$

$$c_{13} = \$877.81$$

$$c_{14} = \$1747.06$$

Therefore, each quarter sales of store  $A$  in dollars is given by the four columns of the row vector

$$[C] = [1182.38 \ 1467.38 \ 877.81 \ 1747.06]$$

Remember since we are multiplying a  $1 \times 3$  matrix by a  $3 \times 4$  matrix, the resulting matrix is a  $1 \times 4$  matrix.

### What is the scalar multiplication of a matrix?

If  $[A]$  is a  $m \times n$  matrix and  $k$  is a real number, then the multiplication  $[A]$  by a scalar  $k$  is another  $m \times n$  matrix  $[B]$ , where

$$b_{ij} = k a_{ij} \text{ for all } i, j.$$

### Example 7

Let

$$[A] = \begin{bmatrix} 2.1 & 3 & 2 \\ 5 & 1 & 6 \end{bmatrix}$$

Find  $2[A]$

#### Solution

$$\begin{aligned} 2[A] &= 2 \begin{bmatrix} 2.1 & 3 & 2 \\ 5 & 1 & 6 \end{bmatrix} \\ &= \begin{bmatrix} 2 \times 2.1 & 2 \times 3 & 2 \times 2 \\ 2 \times 5 & 2 \times 1 & 2 \times 6 \end{bmatrix} \\ &= \begin{bmatrix} 4.2 & 6 & 4 \\ 10 & 2 & 12 \end{bmatrix} \end{aligned}$$

### What is a linear combination of matrices?

If  $[A_1], [A_2], \dots, [A_p]$  are matrices of the same size and  $k_1, k_2, \dots, k_p$  are scalars, then

$$k_1[A_1] + k_2[A_2] + \dots + k_p[A_p]$$

is called a linear combination of  $[A_1], [A_2], \dots, [A_p]$ .

### Example 8

$$\text{If } [A_1] = \begin{bmatrix} 5 & 6 & 2 \\ 3 & 2 & 1 \end{bmatrix}, [A_2] = \begin{bmatrix} 2.1 & 3 & 2 \\ 5 & 1 & 6 \end{bmatrix}, [A_3] = \begin{bmatrix} 0 & 2.2 & 2 \\ 3 & 3.5 & 6 \end{bmatrix}$$

find

$$[A_1] + 2[A_2] - 0.5[A_3]$$

#### Solution

$$\begin{aligned} &[A_1] + 2[A_2] - 0.5[A_3] \\ &= \begin{bmatrix} 5 & 6 & 2 \\ 3 & 2 & 1 \end{bmatrix} + 2 \begin{bmatrix} 2.1 & 3 & 2 \\ 5 & 1 & 6 \end{bmatrix} - 0.5 \begin{bmatrix} 0 & 2.2 & 2 \\ 3 & 3.5 & 6 \end{bmatrix} \\ &= \begin{bmatrix} 5 & 6 & 2 \\ 3 & 2 & 1 \end{bmatrix} + \begin{bmatrix} 4.2 & 6 & 4 \\ 10 & 2 & 12 \end{bmatrix} - \begin{bmatrix} 0 & 1.1 & 1 \\ 1.5 & 1.75 & 3 \end{bmatrix} \\ &= \begin{bmatrix} 9.2 & 10.9 & 5 \\ 11.5 & 2.25 & 10 \end{bmatrix} \end{aligned}$$

### What are some of the rules of binary matrix operations?

#### Commutative law of addition

If  $[A]$  and  $[B]$  are  $m \times n$  matrices, then

$$[A] + [B] = [B] + [A]$$

#### Associative law of addition

If  $[A]$ ,  $[B]$  and  $[C]$  are all  $m \times n$  matrices, then

$$[A] + ([B] + [C]) = ([A] + [B]) + [C]$$

#### Associative law of multiplication

If  $[A]$ ,  $[B]$  and  $[C]$  are  $m \times n$ ,  $n \times p$  and  $p \times r$  size matrices, respectively, then

$$[A]([B][C]) = ([A][B])[C]$$

and the resulting matrix size on both sides of the equation is  $m \times r$ .

#### Distributive law

If  $[A]$  and  $[B]$  are  $m \times n$  size matrices, and  $[C]$  and  $[D]$  are  $n \times p$  size matrices

$$[A]([C] + [D]) = [A][C] + [A][D]$$

$$([A] + [B])[C] = [A][C] + [B][C]$$

and the resulting matrix size on both sides of the equation is  $m \times p$ .

#### Example 9

Illustrate the associative law of multiplication of matrices using

$$[A] = \begin{bmatrix} 1 & 2 \\ 3 & 5 \\ 0 & 2 \end{bmatrix}, \quad [B] = \begin{bmatrix} 2 & 5 \\ 9 & 6 \end{bmatrix}, \quad [C] = \begin{bmatrix} 2 & 1 \\ 3 & 5 \end{bmatrix}$$

#### Solution

$$\begin{aligned} [B][C] &= \begin{bmatrix} 2 & 5 \\ 9 & 6 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 3 & 5 \end{bmatrix} \\ &= \begin{bmatrix} 19 & 27 \\ 36 & 39 \end{bmatrix} \end{aligned}$$

$$\begin{aligned} [A]([B][C]) &= \begin{bmatrix} 1 & 2 \\ 3 & 5 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 19 & 27 \\ 36 & 39 \end{bmatrix} \\ &= \begin{bmatrix} 91 & 105 \\ 237 & 276 \\ 72 & 78 \end{bmatrix} \end{aligned}$$

$$\begin{aligned}
 [A][B] &= \begin{bmatrix} 1 & 2 \\ 3 & 5 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 2 & 5 \\ 9 & 6 \end{bmatrix} \\
 &= \begin{bmatrix} 20 & 17 \\ 51 & 45 \\ 18 & 12 \end{bmatrix} \\
 ([A][B])[C] &= \begin{bmatrix} 20 & 17 \\ 51 & 45 \\ 18 & 12 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 3 & 5 \end{bmatrix} \\
 &= \begin{bmatrix} 91 & 105 \\ 237 & 276 \\ 72 & 78 \end{bmatrix}
 \end{aligned}$$

The above illustrates the associative law of multiplication of matrices.

### Is $[A][B] = [B][A]$ ?

If  $[A][B]$  exists, number of columns of  $[A]$  has to be same as the number of rows of  $[B]$  and if  $[B][A]$  exists, number of columns of  $[B]$  has to be same as the number of rows of  $[A]$ . Now for  $[A][B]=[B][A]$ , the resulting matrix from  $[A][B]$  and  $[B][A]$  has to be of the same size. This is only possible if  $[A]$  and  $[B]$  are square and are of the same size. Even then in general  $[A][B] \neq [B][A]$

### Example 10

Determine if

$$[A][B] = [B][A]$$

for the following matrices

$$[A] = \begin{bmatrix} 6 & 3 \\ 2 & 5 \end{bmatrix}, \quad [B] = \begin{bmatrix} -3 & 2 \\ 1 & 5 \end{bmatrix}$$

### Solution

$$\begin{aligned}
 [A][B] &= \begin{bmatrix} 6 & 3 \\ 2 & 5 \end{bmatrix} \begin{bmatrix} -3 & 2 \\ 1 & 5 \end{bmatrix} \\
 &= \begin{bmatrix} -15 & 27 \\ -1 & 29 \end{bmatrix} \\
 [B][A] &= \begin{bmatrix} -3 & 2 \\ 1 & 5 \end{bmatrix} \begin{bmatrix} 6 & 3 \\ 2 & 5 \end{bmatrix}
 \end{aligned}$$

$$= \begin{bmatrix} -14 & 1 \\ 16 & 28 \end{bmatrix}$$
$$[A][B] \neq [B][A]$$

**Key Terms:***Addition of matrices**Subtraction of matrices**Multiplication of matrices**Scalar Product of matrices**Linear Combination of Matrices**Rules of Binary Matrix Operation*

# Chapter 04.04

## Unary Matrix Operations

After reading this chapter, you should be able to:

1. know what unary operations are,
2. find the transpose of a square matrix and its relationship to symmetric matrices,
3. find the trace of a matrix, and
4. find the determinant of a matrix by the cofactor method.

### What is the transpose of a matrix?

Let  $[A]$  be a  $m \times n$  matrix. Then  $[B]$  is the transpose of the  $[A]$  if  $b_{ji} = a_{ij}$  for all  $i$  and  $j$ . That is, the  $i^{\text{th}}$  row and the  $j^{\text{th}}$  column element of  $[A]$  is the  $j^{\text{th}}$  row and  $i^{\text{th}}$  column element of  $[B]$ . Note,  $[B]$  would be a  $n \times m$  matrix. The transpose of  $[A]$  is denoted by  $[A]^T$ .

### Example 1

Find the transpose of

$$[A] = \begin{bmatrix} 25 & 20 & 3 & 2 \\ 5 & 10 & 15 & 25 \\ 6 & 16 & 7 & 27 \end{bmatrix}$$

### Solution

The transpose of  $[A]$  is

$$[A]^T = \begin{bmatrix} 25 & 5 & 6 \\ 20 & 10 & 16 \\ 3 & 15 & 7 \\ 2 & 25 & 27 \end{bmatrix}$$

Note, the transpose of a row vector is a column vector and the transpose of a column vector is a row vector.

Also, note that the transpose of a transpose of a matrix is the matrix itself, that is,  $([A]^T)^T = [A]$ . Also,  $(A + B)^T = A^T + B^T$ ;  $(cA)^T = cA^T$ .

### What is a symmetric matrix?

A square matrix  $[A]$  with real elements where  $a_{ij} = a_{ji}$  for  $i = 1, 2, \dots, n$  and  $j = 1, 2, \dots, n$  is called a symmetric matrix. This is same as saying that if  $[A] = [A]^T$ , then  $[A]^T$  is a symmetric matrix.

### Example 2

Give an example of a symmetric matrix.

#### Solution

$$[A] = \begin{bmatrix} 21.2 & 3.2 & 6 \\ 3.2 & 21.5 & 8 \\ 6 & 8 & 9.3 \end{bmatrix}$$

is a symmetric matrix as  $a_{12} = a_{21} = 3.2$ ,  $a_{13} = a_{31} = 6$  and  $a_{23} = a_{32} = 8$ .

### What is a skew-symmetric matrix?

A  $n \times n$  matrix is skew symmetric if  $a_{ij} = -a_{ji}$  for  $i = 1, \dots, n$  and  $j = 1, \dots, n$ . This is same as

$$[A] = -[A]^T.$$

### Example 3

Give an example of a skew-symmetric matrix.

#### Solution

$$\begin{bmatrix} 0 & 1 & 2 \\ -1 & 0 & -5 \\ -2 & 5 & 0 \end{bmatrix}$$

is skew-symmetric as

$a_{12} = -a_{21} = 1$ ;  $a_{13} = -a_{31} = 2$ ;  $a_{23} = -a_{32} = -5$ . Since  $a_{ii} = -a_{ii}$  only if  $a_{ii} = 0$ , all the diagonal elements of a skew-symmetric matrix have to be zero.

### What is the trace of a matrix?

The trace of a  $n \times n$  matrix  $[A]$  is the sum of the diagonal entries of  $[A]$ , that is,

$$\text{tr}[A] = \sum_{i=1}^n a_{ii}$$

### Example 4

Find the trace of

$$[A] = \begin{bmatrix} 15 & 6 & 7 \\ 2 & -4 & 2 \\ 3 & 2 & 6 \end{bmatrix}$$

**Solution**

$$\begin{aligned}\text{tr}[A] &= \sum_{i=1}^3 a_{ii} \\ &= (15) + (-4) + (6) \\ &= 17\end{aligned}$$

**Example 5**

The sales of tires are given by make (rows) and quarters (columns) for Blowout r'us store location  $A$ , as shown below.

$$[A] = \begin{bmatrix} 25 & 20 & 3 & 2 \\ 5 & 10 & 15 & 25 \\ 6 & 16 & 7 & 27 \end{bmatrix}$$

where the rows represent the sale of Tirestone, Michigan and Copper tires, and the columns represent the quarter number 1, 2, 3, 4.

Find the total yearly revenue of store  $A$  if the prices of tires vary by quarters as follows.

$$[B] = \begin{bmatrix} 33.25 & 30.01 & 35.02 & 30.05 \\ 40.19 & 38.02 & 41.03 & 38.23 \\ 25.03 & 22.02 & 27.03 & 22.95 \end{bmatrix}$$

where the rows represent the cost of each tire made by Tirestone, Michigan and Copper, and the columns represent the quarter numbers.

**Solution**

To find the total tire sales of store  $A$  for the whole year, we need to find the sales of each brand of tire for the whole year and then add to find the total sales. To do so, we need to rewrite the price matrix so that the quarters are in rows and the brand names are in the columns, that is, find the transpose of  $[B]$ .

$$\begin{aligned}[C] &= [B]^T \\ &= \begin{bmatrix} 33.25 & 30.01 & 35.02 & 30.05 \\ 40.19 & 38.02 & 41.03 & 38.23 \\ 25.03 & 22.02 & 27.03 & 22.95 \end{bmatrix}^T \\ &= \begin{bmatrix} 33.25 & 40.19 & 25.03 \\ 30.01 & 38.02 & 22.02 \\ 35.02 & 41.03 & 27.03 \\ 30.05 & 38.23 & 22.95 \end{bmatrix}\end{aligned}$$

Recognize now that if we find  $[A][C]$ , we get

$$[D] = [A][C]$$

$$\begin{aligned}
 &= \begin{bmatrix} 25 & 20 & 3 & 2 \\ 5 & 10 & 15 & 25 \\ 6 & 16 & 7 & 27 \end{bmatrix} \begin{bmatrix} 33.25 & 40.19 & 25.03 \\ 30.01 & 38.02 & 22.02 \\ 35.02 & 41.03 & 27.03 \\ 30.05 & 38.23 & 22.95 \end{bmatrix} \\
 &= \begin{bmatrix} 1597 & 1965 & 1193 \\ 1743 & 2152 & 1325 \\ 1736 & 2169 & 1311 \end{bmatrix}
 \end{aligned}$$

The diagonal elements give the sales of each brand of tire for the whole year, that is

$$d_{11} = \$1597 \text{ (Tirestone sales)}$$

$$d_{22} = \$2152 \text{ (Michigan sales)}$$

$$d_{33} = \$1311 \text{ (Cooper sales)}$$

The total yearly sales of all three brands of tires are

$$\begin{aligned}
 \sum_{i=1}^3 d_{ii} &= 1597 + 2152 + 1311 \\
 &= \$5060
 \end{aligned}$$

and this is the trace of the matrix  $[D]$ .

### Define the determinant of a matrix.

The determinant of a square matrix is a single unique real number corresponding to a matrix. For a matrix  $[A]$ , determinant is denoted by  $|A|$  or  $\det(A)$ . So do not use  $[A]$  and  $|A|$  interchangeably.

For a  $2 \times 2$  matrix,

$$[A] = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

$$\det(A) = a_{11}a_{22} - a_{12}a_{21}$$

### How does one calculate the determinant of any square matrix?

Let  $[A]$  be  $n \times n$  matrix. The minor of entry  $a_{ij}$  is denoted by  $M_{ij}$  and is defined as the determinant of the  $(n-1 \times n-1)$  submatrix of  $[A]$ , where the submatrix is obtained by deleting the  $i^{th}$  row and  $j^{th}$  column of the matrix  $[A]$ . The determinant is then given by

$$\det(A) = \sum_{j=1}^n (-1)^{i+j} a_{ij} M_{ij} \text{ for any } i = 1, 2, \dots, n$$

or

$$\det(A) = \sum_{i=1}^n (-1)^{i+j} a_{ij} M_{ij} \text{ for any } j = 1, 2, \dots, n$$

Coupled that with  $\det(A) = a_{11}$  for a  $1 \times 1$  matrix  $[A]$ , we can always reduce the determinant of a matrix to determinants of  $1 \times 1$  matrices. The number  $(-1)^{i+j} M_{ij}$  is called the cofactor of  $a_{ij}$  and is denoted by  $c_{ij}$ . The formula for the determinant can then be written as

$$\det(A) = \sum_{j=1}^n a_{ij} C_{ij} \text{ for any } i = 1, 2, \dots, n$$

or

$$\det(A) = \sum_{i=1}^n a_{ij} C_{ij} \text{ for any } j = 1, 2, \dots, n$$

Determinants are not generally calculated using this method as it becomes computationally intensive for large matrices. For a  $n \times n$  matrix, it requires arithmetic operations proportional to  $n!$ .

### Example 6

Find the determinant of

$$[A] = \begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix}$$

### Solution

#### Method 1:

$$\det(A) = \sum_{j=1}^3 (-1)^{i+j} a_{ij} M_{ij} \text{ for any } i = 1, 2, 3$$

Let us choose  $i = 1$  in the formula

$$\begin{aligned} \det(A) &= \sum_{j=1}^3 (-1)^{1+j} a_{1j} M_{1j} \\ &= (-1)^{1+1} a_{11} M_{11} + (-1)^{1+2} a_{12} M_{12} + (-1)^{1+3} a_{13} M_{13} \\ &= a_{11} M_{11} - a_{12} M_{12} + a_{13} M_{13} \end{aligned}$$

$$M_{11} = \begin{vmatrix} 8 & 1 \\ 12 & 1 \end{vmatrix}$$

$$= -4$$

$$M_{12} = \begin{vmatrix} 64 & 1 \\ 144 & 1 \end{vmatrix}$$

$$= -80$$

$$M_{13} = \begin{vmatrix} 64 & 8 \\ 144 & 12 \end{vmatrix}$$

$$= -384$$

$$\det(A) = a_{11} M_{11} - a_{12} M_{12} + a_{13} M_{13}$$

$$\begin{aligned}
 &= 25(-4) - 5(-80) + 1(-384) \\
 &= -100 + 400 - 384 \\
 &= -84
 \end{aligned}$$

Also for  $i = 1$ ,

$$\begin{aligned}
 \det(A) &= \sum_{j=1}^3 a_{1j} C_{1j} \\
 C_{11} &= (-1)^{1+1} M_{11} \\
 &= M_{11} \\
 &= -4 \\
 C_{12} &= (-1)^{1+2} M_{12} \\
 &= -M_{12} \\
 &= 80 \\
 C_{13} &= (-1)^{1+3} M_{13} \\
 &= M_{13} \\
 &= -384 \\
 \det(A) &= a_{11}C_{11} + a_{21}C_{21} + a_{31}C_{31} \\
 &= (25)(-4) + (5)(80) + (1)(-384) \\
 &= -100 + 400 - 384 \\
 &= -84
 \end{aligned}$$

### Method 2:

$$\det(A) = \sum_{i=1}^3 (-1)^{i+j} a_{ij} M_{ij} \text{ for any } j = 1, 2, 3$$

Let us choose  $j = 2$  in the formula

$$\begin{aligned}
 \det(A) &= \sum_{i=1}^3 (-1)^{i+2} a_{i2} M_{i2} \\
 &= (-1)^{1+2} a_{12} M_{12} + (-1)^{2+2} a_{22} M_{22} + (-1)^{3+2} a_{32} M_{32} \\
 &= -a_{12} M_{12} + a_{22} M_{22} - a_{32} M_{32} \\
 M_{12} &= \begin{vmatrix} 64 & 1 \\ 144 & 1 \end{vmatrix} \\
 &= -80 \\
 M_{22} &= \begin{vmatrix} 25 & 1 \\ 144 & 1 \end{vmatrix} \\
 &= -119 \\
 M_{32} &= \begin{vmatrix} 25 & 1 \\ 64 & 1 \end{vmatrix} \\
 &= -39
 \end{aligned}$$

$$\det(A) = -a_{12} M_{12} + a_{22} M_{22} - a_{32} M_{32}$$

$$\begin{aligned}
 &= -5(-80) + 8(-119) - 12(-39) \\
 &= 400 - 952 + 468 \\
 &= -84
 \end{aligned}$$

In terms of cofactors for  $j = 2$ ,

$$\begin{aligned}
 \det(A) &= \sum_{i=1}^3 a_{i2} C_{i2} \\
 C_{12} &= (-1)^{1+2} M_{12} \\
 &= -M_{12} \\
 &= 80 \\
 C_{22} &= (-1)^{2+2} M_{22} \\
 &= M_{22} \\
 &= -119 \\
 C_{32} &= (-1)^{3+2} M_{32} \\
 &= -M_{32} \\
 &= 39 \\
 \det(A) &= a_{12}C_{12} + a_{22}C_{22} + a_{32}C_{32} \\
 &= (5)(80) + (8)(-119) + (12)(39) \\
 &= 400 - 952 + 468 \\
 &= -84
 \end{aligned}$$

**Is there a relationship between  $\det(AB)$ , and  $\det(A)$  and  $\det(B)$ ?**

Yes, if  $[A]$  and  $[B]$  are square matrices of same size, then

$$\det(AB) = \det(A)\det(B)$$

**Are there some other theorems that are important in finding the determinant of a square matrix?**

**Theorem 1:** If a row or a column in a  $n \times n$  matrix  $[A]$  is zero, then  $\det(A) = 0$ .

**Theorem 2:** Let  $[A]$  be a  $n \times n$  matrix. If a row is proportional to another row, then  $\det(A) = 0$ .

**Theorem 3:** Let  $[A]$  be a  $n \times n$  matrix. If a column is proportional to another column, then  $\det(A) = 0$ .

**Theorem 4:** Let  $[A]$  be a  $n \times n$  matrix. If a column or row is multiplied by  $k$  to result in matrix  $B$ , then  $\det(B) = k \det(A)$ .

**Theorem 5:** Let  $[A]$  be a  $n \times n$  upper or lower triangular matrix, then  $\det(A) = \prod_{i=1}^n a_{ii}$ .

### Example 7

What is the determinant of

$$[A] = \begin{bmatrix} 0 & 2 & 6 & 3 \\ 0 & 3 & 7 & 4 \\ 0 & 4 & 9 & 5 \\ 0 & 5 & 2 & 1 \end{bmatrix}$$

**Solution**

Since one of the columns (first column in the above example) of  $[A]$  is a zero,  $\det(A) = 0$ .

**Example 8**

What is the determinant of

$$[A] = \begin{bmatrix} 2 & 1 & 6 & 4 \\ 3 & 2 & 7 & 6 \\ 5 & 4 & 2 & 10 \\ 9 & 5 & 3 & 18 \end{bmatrix}$$

**Solution**

$\det(A)$  is zero because the fourth column

$$\begin{bmatrix} 4 \\ 6 \\ 10 \\ 18 \end{bmatrix}$$

is 2 times the first column

$$\begin{bmatrix} 2 \\ 3 \\ 5 \\ 9 \end{bmatrix}$$

**Example 9**

If the determinant of

$$[A] = \begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix}$$

is  $-84$ , then what is the determinant of

$$[B] = \begin{bmatrix} 25 & 10.5 & 1 \\ 64 & 16.8 & 1 \\ 144 & 25.2 & 1 \end{bmatrix}$$

**Solution**

Since the second column of  $[B]$  is 2.1 times the second column of  $[A]$

$$\begin{aligned}\det(B) &= 2.1 \det(A) \\ &= (2.1)(-84) \\ &= -176.4\end{aligned}$$

**Example 10**

Given the determinant of

$$[A] = \begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix}$$

is  $-84$ , what is the determinant of

$$[B] = \begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 144 & 12 & 1 \end{bmatrix}$$

**Solution**

Since  $[B]$  is simply obtained by subtracting the second row of  $[A]$  by 2.56 times the first row of  $[A]$ ,

$$\begin{aligned}\det(B) &= \det(A) \\ &= -84\end{aligned}$$

**Example 11**

What is the determinant of

$$[A] = \begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & 0 & 0.7 \end{bmatrix}$$

**Solution**

Since  $[A]$  is an upper triangular matrix

$$\begin{aligned}\det(A) &= \prod_{i=1}^3 a_{ii} \\ &= a_{11} \times a_{22} \times a_{33} \\ &= 25 \times (-4.8) \times 0.7 \\ &= -84\end{aligned}$$

**Key Terms:**

*Transpose*

*Symmetric Matrix*

*Skew-Symmetric Matrix*

*Trace of Matrix  
Determinant*

# Chapter 04.05

## System of Equations

After reading this chapter, you should be able to:

1. setup simultaneous linear equations in matrix form and vice-versa,
2. understand the concept of the inverse of a matrix,
3. know the difference between a consistent and inconsistent system of linear equations, and
4. learn that a system of linear equations can have a unique solution, no solution or infinite solutions.

**Matrix algebra is used for solving systems of equations. Can you illustrate this concept?**

Matrix algebra is used to solve a system of simultaneous linear equations. In fact, for many mathematical procedures such as the solution to a set of nonlinear equations, interpolation, integration, and differential equations, the solutions reduce to a set of simultaneous linear equations. Let us illustrate with an example for interpolation.

### Example 1

The upward velocity of a rocket is given at three different times on the following table.

**Table 5.1.** Velocity vs. time data for a rocket

Time, $t$ (s)	Velocity, $v$ (m/s)
5	106.8
8	177.2
12	279.2

The velocity data is approximated by a polynomial as

$$v(t) = at^2 + bt + c, \quad 5 \leq t \leq 12.$$

Set up the equations in matrix form to find the coefficients  $a, b, c$  of the velocity profile.

### Solution

The polynomial is going through three data points  $(t_1, v_1), (t_2, v_2)$ , and  $(t_3, v_3)$  where from table 5.1.

$$t_1 = 5, v_1 = 106.8$$

$$t_2 = 8, v_2 = 177.2$$

$$t_3 = 12, v_3 = 279.2$$

Requiring that  $v(t) = at^2 + bt + c$  passes through the three data points gives

$$v(t_1) = v_1 = at_1^2 + bt_1 + c$$

$$v(t_2) = v_2 = at_2^2 + bt_2 + c$$

$$v(t_3) = v_3 = at_3^2 + bt_3 + c$$

Substituting the data  $(t_1, v_1), (t_2, v_2)$ , and  $(t_3, v_3)$  gives

$$a(5^2) + b(5) + c = 106.8$$

$$a(8^2) + b(8) + c = 177.2$$

$$a(12^2) + b(12) + c = 279.2$$

or

$$25a + 5b + c = 106.8$$

$$64a + 8b + c = 177.2$$

$$144a + 12b + c = 279.2$$

This set of equations can be rewritten in the matrix form as

$$\begin{bmatrix} 25a & 5b & c \\ 64a & 8b & c \\ 144a & 12b & c \end{bmatrix} = \begin{bmatrix} 106.8 \\ 177.2 \\ 279.2 \end{bmatrix}$$

The above equation can be written as a linear combination as follows

$$a \begin{bmatrix} 25 \\ 64 \\ 144 \end{bmatrix} + b \begin{bmatrix} 5 \\ 8 \\ 12 \end{bmatrix} + c \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 106.8 \\ 177.2 \\ 279.2 \end{bmatrix}$$

and further using matrix multiplication gives

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 106.8 \\ 177.2 \\ 279.2 \end{bmatrix}$$

The above is an illustration of why matrix algebra is needed. The complete solution to the set of equations is given later in this chapter.

A general set of  $m$  linear equations and  $n$  unknowns,

$$a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n = c_1$$

$$a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n = c_2$$

.....

.....

$$a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mn}x_n = c_m$$

can be rewritten in the matrix form as

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_m \end{bmatrix}$$

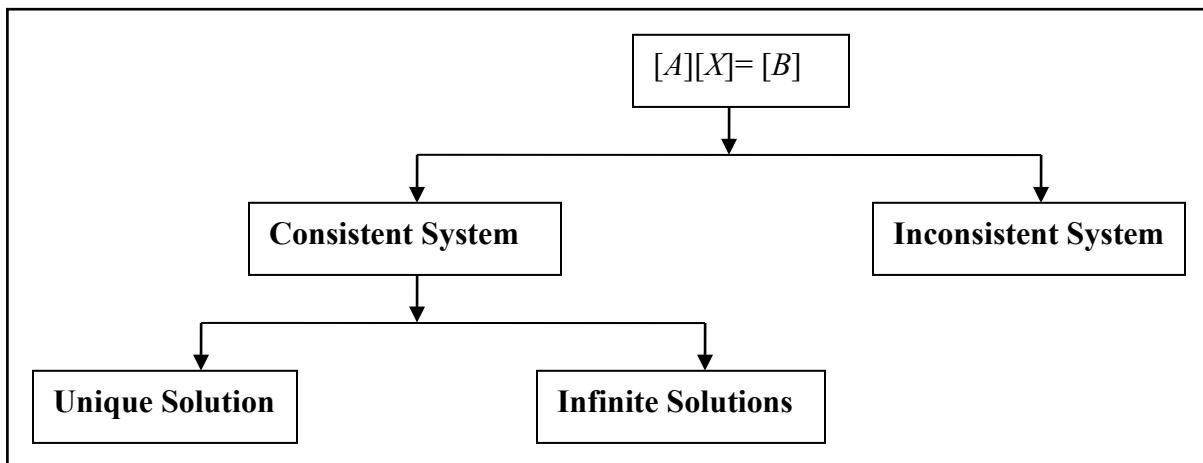
Denoting the matrices by  $[A]$ ,  $[X]$ , and  $[C]$ , the system of equation is  $[A][X]=[C]$ , where  $[A]$  is called the coefficient matrix,  $[C]$  is called the right hand side vector and  $[X]$  is called the solution vector.

Sometimes  $[A][X]=[C]$  systems of equations are written in the augmented form. That is

$$[A:C] = \left[ \begin{array}{cccc|c} a_{11} & a_{12} & \dots & a_{1n} & c_1 \\ a_{21} & a_{22} & \dots & a_{2n} & c_2 \\ \vdots & & & & \vdots \\ \vdots & & & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} & c_n \end{array} \right]$$

**A system of equations can be consistent or inconsistent. What does that mean?**

A system of equations  $[A][X]=[C]$  is consistent if there is a solution, and it is inconsistent if there is no solution. However, a consistent system of equations does not mean a unique solution, that is, a consistent system of equations may have a unique solution or infinite solutions (Figure 1).



**Figure 5.1.** Consistent and inconsistent system of equations flow chart.

### Example 2

Give examples of consistent and inconsistent system of equations.

#### Solution

- a) The system of equations

$$\begin{bmatrix} 2 & 4 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 6 \\ 4 \end{bmatrix}$$

is a consistent system of equations as it has a unique solution, that is,

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

b) The system of equations

$$\begin{bmatrix} 2 & 4 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 6 \\ 3 \end{bmatrix}$$

is also a consistent system of equations but it has infinite solutions as given as follows.

Expanding the above set of equations,

$$2x + 4y = 6$$

$$x + 2y = 3$$

you can see that they are the same equation. Hence, any combination of  $(x, y)$  that satisfies

$$2x + 4y = 6$$

is a solution. For example  $(x, y) = (1, 1)$  is a solution. Other solutions include  $(x, y) = (0.5, 1.25)$ ,  $(x, y) = (0, 1.5)$ , and so on.

c) The system of equations

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

is inconsistent as no solution exists.

### How can one distinguish between a consistent and inconsistent system of equations?

A system of equations  $[A][X] = [C]$  is *consistent* if the rank of  $A$  is equal to the rank of the augmented matrix  $[A:C]$

A system of equations  $[A][X] = [C]$  is *inconsistent* if the rank of  $A$  is less than the rank of the augmented matrix  $[A:C]$ .

### But, what do you mean by rank of a matrix?

The rank of a matrix is defined as the order of the largest square submatrix whose determinant is not zero.

### Example 3

What is the rank of

$$[A] = \begin{bmatrix} 3 & 1 & 2 \\ 2 & 0 & 5 \\ 1 & 2 & 3 \end{bmatrix} ?$$

### Solution

The largest square submatrix possible is of order 3 and that is  $[A]$  itself. Since  $\det(A) = -23 \neq 0$ , the rank of  $[A] = 3$ .

**Example 4**

What is the rank of

$$[A] = \begin{bmatrix} 3 & 1 & 2 \\ 2 & 0 & 5 \\ 5 & 1 & 7 \end{bmatrix}?$$

**Solution**

The largest square submatrix of  $[A]$  is of order 3 and that is  $[A]$  itself. Since  $\det(A) = 0$ , the rank of  $[A]$  is less than 3. The next largest square submatrix would be a  $2 \times 2$  matrix. One of the square submatrices of  $[A]$  is

$$[B] = \begin{bmatrix} 3 & 1 \\ 2 & 0 \end{bmatrix}$$

and  $\det(B) = -2 \neq 0$ . Hence the rank of  $[A]$  is 2. There is no need to look at other  $2 \times 2$  submatrices to establish that the rank of  $[A]$  is 2.

**Example 5**

How do I now use the concept of rank to find if

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 106.8 \\ 177.2 \\ 279.2 \end{bmatrix}$$

is a consistent or inconsistent system of equations?

**Solution**

The coefficient matrix is

$$[A] = \begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix}$$

and the right hand side vector is

$$[C] = \begin{bmatrix} 106.8 \\ 177.2 \\ 279.2 \end{bmatrix}$$

The augmented matrix is

$$[B] = \begin{bmatrix} 25 & 5 & 1 & : & 106.8 \\ 64 & 8 & 1 & : & 177.2 \\ 144 & 12 & 1 & : & 279.2 \end{bmatrix}$$

Since there are no square submatrices of order 4 as  $[B]$  is a  $3 \times 4$  matrix, the rank of  $[B]$  is at most 3. So let us look at the square submatrices of  $[B]$  of order 3; if any of these square

submatrices have determinant not equal to zero, then the rank is 3. For example, a submatrix of the augmented matrix  $[B]$  is

$$[D] = \begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix}$$

has  $\det(D) = -84 \neq 0$ .

Hence the rank of the augmented matrix  $[B]$  is 3. Since  $[A]=[D]$ , the rank of  $[A]$  is 3. Since the rank of the augmented matrix  $[B]$  equals the rank of the coefficient matrix  $[A]$ , the system of equations is consistent.

### Example 6

Use the concept of rank of matrix to find if

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 89 & 13 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 106.8 \\ 177.2 \\ 284.0 \end{bmatrix}$$

is consistent or inconsistent?

#### Solution

The coefficient matrix is given by

$$[A] = \begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 89 & 13 & 2 \end{bmatrix}$$

and the right hand side

$$[C] = \begin{bmatrix} 106.8 \\ 177.2 \\ 284.0 \end{bmatrix}$$

The augmented matrix is

$$[B] = \begin{bmatrix} 25 & 5 & 1 & :106.8 \\ 64 & 8 & 1 & :177.2 \\ 89 & 13 & 2 & :284.0 \end{bmatrix}$$

Since there are no square submatrices of order 4 as  $[B]$  is a  $4 \times 3$  matrix, the rank of the augmented  $[B]$  is at most 3. So let us look at square submatrices of the augmented matrix  $[B]$  of order 3 and see if any of these have determinants not equal to zero. For example, a square submatrix of the augmented matrix  $[B]$  is

$$[D] = \begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 89 & 13 & 2 \end{bmatrix}$$

has  $\det(D) = 0$ . This means, we need to explore other square submatrices of order 3 of the augmented matrix  $[B]$  and find their determinants.

That is,

$$[E] = \begin{bmatrix} 5 & 1 & 106.8 \\ 8 & 1 & 177.2 \\ 13 & 2 & 284.0 \end{bmatrix}$$

$$\det(E) = 0$$

$$[F] = \begin{bmatrix} 25 & 5 & 106.8 \\ 64 & 8 & 177.2 \\ 89 & 13 & 284.0 \end{bmatrix}$$

$$\det(F) = 0$$

$$[G] = \begin{bmatrix} 25 & 1 & 106.8 \\ 64 & 1 & 177.2 \\ 89 & 2 & 284.0 \end{bmatrix}$$

$$\det(G) = 0$$

All the square submatrices of order  $3 \times 3$  of the augmented matrix  $[B]$  have a zero determinant. So the rank of the augmented matrix  $[B]$  is less than 3. Is the rank of augmented matrix  $[B]$  equal to 2? One of the  $2 \times 2$  submatrices of the augmented matrix  $[B]$  is

$$[H] = \begin{bmatrix} 25 & 5 \\ 64 & 8 \end{bmatrix}$$

and

$$\det(H) = -120 \neq 0$$

So the rank of the augmented matrix  $[B]$  is 2.

Now we need to find the rank of the coefficient matrix  $[B]$ .

$$[A] = \begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 89 & 13 & 2 \end{bmatrix}$$

and

$$\det(A) = 0$$

So the rank of the coefficient matrix  $[A]$  is less than 3. A square submatrix of the coefficient matrix  $[A]$  is

$$[J] = \begin{bmatrix} 5 & 1 \\ 8 & 1 \end{bmatrix}$$

$$\det(J) = -3 \neq 0$$

So the rank of the coefficient matrix  $[A]$  is 2.

Hence, rank of the coefficient matrix  $[A]$  equals the rank of the augmented matrix  $[B]$ . So the system of equations  $[A][X]=[C]$  is consistent.

### Example 7

Use the concept of rank to find if

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 89 & 13 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 106.8 \\ 177.2 \\ 280.0 \end{bmatrix}$$

is consistent or inconsistent.

### Solution

The augmented matrix is

$$[B] = \begin{bmatrix} 25 & 5 & 1 & : 106.8 \\ 64 & 8 & 1 & : 177.2 \\ 89 & 13 & 2 & : 280.0 \end{bmatrix}$$

Since there are no square submatrices of order  $4 \times 4$  as the augmented matrix  $[B]$  is a  $4 \times 3$  matrix, the rank of the augmented matrix  $[B]$  is at most 3. So let us look at square submatrices of the augmented matrix  $(B)$  of order 3 and see if any of the  $3 \times 3$  submatrices have a determinant not equal to zero. For example, a square submatrix of order  $3 \times 3$  of  $[B]$

$$[D] = \begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 89 & 13 & 2 \end{bmatrix}$$

$$\det(D) = 0$$

So it means, we need to explore other square submatrices of the augmented matrix  $[B]$

$$[E] = \begin{bmatrix} 5 & 1 & 106.8 \\ 8 & 1 & 177.2 \\ 13 & 2 & 280.0 \end{bmatrix}$$

$$\det(E) = 12.0 \neq 0.$$

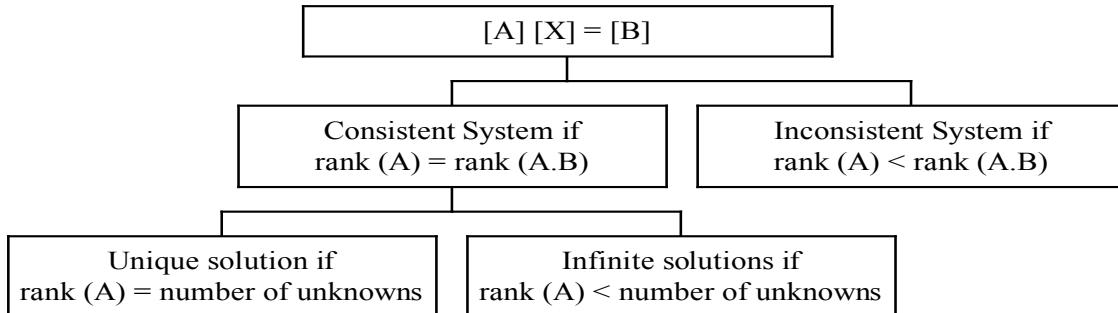
So the rank of the augmented matrix  $[B]$  is 3.

The rank of the coefficient matrix  $[A]$  is 2 from the previous example.

Since the rank of the coefficient matrix  $[A]$  is less than the rank of the augmented matrix  $[B]$ , the system of equations is inconsistent. Hence, no solution exists for  $[A][X]=[C]$ .

### If a solution exists, how do we know whether it is unique?

In a system of equations  $[A][X]=[C]$  that is consistent, the rank of the coefficient matrix  $[A]$  is the same as the augmented matrix  $[A|C]$ . If in addition, the rank of the coefficient matrix  $[A]$  is same as the number of unknowns, then the solution is unique; if the rank of the coefficient matrix  $[A]$  is less than the number of unknowns, then infinite solutions exist.



**Figure 5.2.** Flow chart of conditions for consistent and inconsistent system of equations.

### Example 8

We found that the following system of equations

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 106.8 \\ 177.2 \\ 279.2 \end{bmatrix}$$

is a consistent system of equations. Does the system of equations have a unique solution or does it have infinite solutions?

#### Solution

The coefficient matrix is

$$[A] = \begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix}$$

and the right hand side is

$$[C] = \begin{bmatrix} 106.8 \\ 177.2 \\ 279.2 \end{bmatrix}$$

While finding out whether the above equations were consistent in an earlier example, we found that the rank of the coefficient matrix ( $A$ ) equals rank of augmented matrix  $[A:C]$  equals 3.

The solution is unique as the number of unknowns = 3 = rank of ( $A$ ).

### Example 9

We found that the following system of equations

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 89 & 13 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 106.8 \\ 177.2 \\ 284.0 \end{bmatrix}$$

is a consistent system of equations. Is the solution unique or does it have infinite solutions.

### Solution

While finding out whether the above equations were consistent, we found that the rank of the coefficient matrix  $[A]$  equals the rank of augmented matrix  $(A:C)$  equals 2

Since the rank of  $[A] = 2 <$  number of unknowns = 3, infinite solutions exist.

**If we have more equations than unknowns in  $[A] [X] = [C]$ , does it mean the system is inconsistent?**

No, it depends on the rank of the augmented matrix  $[A:C]$  and the rank of  $[A]$ .

a) For example

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \\ 89 & 13 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 106.8 \\ 177.2 \\ 279.2 \\ 284.0 \end{bmatrix}$$

is consistent, since

$$\begin{aligned} \text{rank of augmented matrix} &= 3 \\ \text{rank of coefficient matrix} &= 3 \end{aligned}$$

Now since the rank of  $(A) = 3 =$  number of unknowns, the solution is not only consistent but also unique.

b) For example

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \\ 89 & 13 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 106.8 \\ 177.2 \\ 279.2 \\ 280.0 \end{bmatrix}$$

is inconsistent, since

$$\begin{aligned} \text{rank of augmented matrix} &= 4 \\ \text{rank of coefficient matrix} &= 3 \end{aligned}$$

c) For example

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 50 & 10 & 2 \\ 89 & 13 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 106.8 \\ 177.2 \\ 213.6 \\ 280.0 \end{bmatrix}$$

is consistent, since

$$\begin{aligned} \text{rank of augmented matrix} &= 2 \\ \text{rank of coefficient matrix} &= 2 \end{aligned}$$

But since the rank of  $[A] = 2 <$  the number of unknowns = 3, infinite solutions exist.

**Consistent systems of equations can only have a unique solution or infinite solutions. Can a system of equations have more than one but not infinite number of solutions?**

No, you can only have either a unique solution or infinite solutions. Let us suppose  $[A][X]=[C]$  has two solutions  $[Y]$  and  $[Z]$  so that

$$[A][Y]=[C]$$

$$[A][Z]=[C]$$

If  $r$  is a constant, then from the two equations

$$r[A][Y]=r[C]$$

$$(1-r)[A][Z]=(1-r)[C]$$

Adding the above two equations gives

$$r[A][Y]+(1-r)[A][Z]=r[C]+(1-r)[C]$$

$$[A](r[Y]+(1-r)[Z])=[C]$$

Hence

$$r[Y]+(1-r)[Z]$$

is a solution to

$$[A][X]=[C]$$

Since  $r$  is any scalar, there are infinite solutions for  $[A][X]=[C]$  of the form

$$r[Y]+(1-r)[Z]$$

### Can you divide two matrices?

If  $[A][B]=[C]$  is defined, it might seem intuitive that  $[A]=\frac{[C]}{[B]}$ , but matrix division is not defined like that. However an inverse of a matrix can be defined for certain types of square matrices. The inverse of a square matrix  $[A]$ , if existing, is denoted by  $[A]^{-1}$  such that

$$[A][A]^{-1}=[I]=[A]^{-1}[A]$$

Where  $[I]$  is the identity matrix.

In other words, let  $[A]$  be a square matrix. If  $[B]$  is another square matrix of the same size such that  $[B][A]=[I]$ , then  $[B]$  is the inverse of  $[A]$ .  $[A]$  is then called to be invertible or nonsingular. If  $[A]^{-1}$  does not exist,  $[A]$  is called noninvertible or singular.

If  $[A]$  and  $[B]$  are two  $n \times n$  matrices such that  $[B][A]=[I]$ , then these statements are also true

- $[B]$  is the inverse of  $[A]$
- $[A]$  is the inverse of  $[B]$
- $[A]$  and  $[B]$  are both invertible
- $[A][B]=[I]$ .
- $[A]$  and  $[B]$  are both nonsingular
- all columns of  $[A]$  and  $[B]$  are linearly independent
- all rows of  $[A]$  and  $[B]$  are linearly independent.

### Example 10

Determine if

$$[B]=\begin{bmatrix} 3 & 2 \\ 5 & 3 \end{bmatrix}$$

is the inverse of

$$[A] = \begin{bmatrix} -3 & 2 \\ 5 & -3 \end{bmatrix}$$

**Solution**

$$\begin{aligned}[B][A] &= \begin{bmatrix} 3 & 2 \\ 5 & 3 \end{bmatrix} \begin{bmatrix} -3 & 2 \\ 5 & -3 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\ &= [I]\end{aligned}$$

Since

$$[B][A] = [I],$$

[B] is the inverse of [A] and [A] is the inverse of [B].

But, we can also show that

$$\begin{aligned}[A][B] &= \begin{bmatrix} -3 & 2 \\ 5 & -3 \end{bmatrix} \begin{bmatrix} 3 & 2 \\ 5 & 3 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\ &= [I]\end{aligned}$$

to show that [A] is the inverse of [B].

**Can I use the concept of the inverse of a matrix to find the solution of a set of equations  
[A] [X] = [C]?**

Yes, if the number of equations is the same as the number of unknowns, the coefficient matrix [A] is a square matrix.

Given

$$[A][X] = [C]$$

Then, if  $[A]^{-1}$  exists, multiplying both sides by  $[A]^{-1}$ .

$$[A]^{-1}[A][X] = [A]^{-1}[C]$$

$$[I][X] = [A]^{-1}[C]$$

$$[X] = [A]^{-1}[C]$$

This implies that if we are able to find  $[A]^{-1}$ , the solution vector of  $[A][X] = [C]$  is simply a multiplication of  $[A]^{-1}$  and the right hand side vector, [C].

**How do I find the inverse of a matrix?**

If  $[A]$  is a  $n \times n$  matrix, then  $[A]^{-1}$  is a  $n \times n$  matrix and according to the definition of inverse of a matrix

$$[A][A]^{-1} = [I]$$

Denoting

$$[A] = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

$$[A]^{-1} = \begin{bmatrix} \dot{a}_{11} & \dot{a}_{12} & \cdots & \dot{a}_{1n} \\ \dot{a}_{21} & \dot{a}_{22} & \cdots & \dot{a}_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ \dot{a}_{n1} & \dot{a}_{n2} & \cdots & \dot{a}_{nn} \end{bmatrix}$$

$$[I] = \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & & & 0 \\ 0 & & \ddots & & \vdots \\ \vdots & & & 1 & \vdots \\ 0 & \cdots & \cdots & \cdots & 1 \end{bmatrix}$$

Using the definition of matrix multiplication, the first column of the  $[A]^{-1}$  matrix can then be found by solving

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} \dot{a}_{11} \\ \dot{a}_{21} \\ \vdots \\ \dot{a}_{n1} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Similarly, one can find the other columns of the  $[A]^{-1}$  matrix by changing the right hand side accordingly.

### Example 11

The upward velocity of the rocket is given by

**Table 5.2.** Velocity vs time data for a rocket

Time, $t$ (s)	Velocity, $v$ (m/s)
5	106.8
8	177.2
12	279.2

In an earlier example, we wanted to approximate the velocity profile by

$$v(t) = at^2 + bt + c, \quad 5 \leq t \leq 12$$

We found that the coefficients  $a, b$ , and  $c$  in  $v(t)$  are given by

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 106.8 \\ 177.2 \\ 279.2 \end{bmatrix}$$

First, find the inverse of

$$[A] = \begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix}$$

and then use the definition of inverse to find the coefficients  $a, b$ , and  $c$ .

### Solution

If

$$[A]^{-1} = \begin{bmatrix} \dot{a_{11}} & \dot{a_{12}} & \dot{a_{13}} \\ \dot{a_{21}} & \dot{a_{22}} & \dot{a_{23}} \\ \dot{a_{31}} & \dot{a_{32}} & \dot{a_{33}} \end{bmatrix}$$

is the inverse of  $[A]$ , then

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix} \begin{bmatrix} \dot{a_{11}} & \dot{a_{12}} & \dot{a_{13}} \\ \dot{a_{21}} & \dot{a_{22}} & \dot{a_{23}} \\ \dot{a_{31}} & \dot{a_{32}} & \dot{a_{33}} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

gives three sets of equations

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix} \begin{bmatrix} \dot{a_{11}} \\ \dot{a_{21}} \\ \dot{a_{31}} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix} \begin{bmatrix} \dot{a_{12}} \\ \dot{a_{22}} \\ \dot{a_{32}} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix} \begin{bmatrix} \dot{a_{13}} \\ \dot{a_{23}} \\ \dot{a_{33}} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

Solving the above three sets of equations separately gives

$$\begin{bmatrix} \dot{a_{11}} \\ \dot{a_{21}} \\ \dot{a_{31}} \end{bmatrix} = \begin{bmatrix} 0.04762 \\ -0.9524 \\ 4.571 \end{bmatrix}$$

$$\begin{bmatrix} \dot{a_{12}} \\ \dot{a_{22}} \\ \dot{a_{32}} \end{bmatrix} = \begin{bmatrix} -0.08333 \\ 1.417 \\ -5.000 \end{bmatrix}$$

$$\begin{bmatrix} a_{13} \\ a_{23} \\ a_{33} \end{bmatrix} = \begin{bmatrix} 0.03571 \\ -0.4643 \\ 1.429 \end{bmatrix}$$

Hence

$$[A]^{-1} = \begin{bmatrix} 0.04762 & -0.08333 & 0.03571 \\ -0.9524 & 1.417 & -0.4643 \\ 4.571 & -5.000 & 1.429 \end{bmatrix}$$

Now

$$[A][X] = [C]$$

where

$$[X] = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

$$[C] = \begin{bmatrix} 106.8 \\ 177.2 \\ 279.2 \end{bmatrix}$$

Using the definition of  $[A]^{-1}$ ,

$$[A]^{-1}[A][X] = [A]^{-1}[C]$$

$$[X] = [A]^{-1}[C]$$

$$\begin{bmatrix} 0.04762 & -0.08333 & 0.03571 \\ -0.9524 & 1.417 & -0.4643 \\ 4.571 & -5.000 & 1.429 \end{bmatrix} \begin{bmatrix} 106.8 \\ 177.2 \\ 279.2 \end{bmatrix}$$

Hence

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 0.2905 \\ 19.69 \\ 1.086 \end{bmatrix}$$

So

$$v(t) = 0.2905t^2 + 19.69t + 1.086, \quad 5 \leq t \leq 12$$

### Is there another way to find the inverse of a matrix?

For finding the inverse of small matrices, the inverse of an invertible matrix can be found by

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

where

$$\text{adj}(A) = \begin{bmatrix} C_{11} & C_{12} & \cdots & C_{1n} \\ C_{21} & C_{22} & & C_{2n} \\ \vdots & & & \vdots \\ C_{n1} & C_{n2} & \cdots & C_{nn} \end{bmatrix}^T$$

where  $C_{ij}$  are the cofactors of  $a_{ij}$ . The matrix

$$\begin{bmatrix} C_{11} & C_{12} & \cdots & C_{1n} \\ C_{21} & C_{22} & \cdots & C_{2n} \\ \vdots & & & \vdots \\ C_{n1} & \cdots & \cdots & C_{nn} \end{bmatrix}$$

itself is called the matrix of cofactors from  $[A]$ . Cofactors are defined in [Chapter 4](#).

### Example 12

Find the inverse of

$$[A] = \begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix}$$

### Solution

From Example 4.6 in Chapter 04.06, we found

$$\det(A) = -84$$

Next we need to find the adjoint of  $[A]$ . The cofactors of  $A$  are found as follows.

The minor of entry  $a_{11}$  is

$$\begin{aligned} M_{11} &= \begin{vmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{vmatrix} \\ &= \begin{vmatrix} 8 & 1 \\ 12 & 1 \end{vmatrix} \\ &= -4 \end{aligned}$$

The cofactors of entry  $a_{11}$  is

$$\begin{aligned} C_{11} &= (-1)^{1+1} M_{11} \\ &= M_{11} \\ &= -4 \end{aligned}$$

The minor of entry  $a_{12}$  is

$$M_{12} = \begin{vmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{vmatrix}$$

$$= \begin{vmatrix} 64 & 1 \\ 144 & 1 \end{vmatrix} \\ = -80$$

The cofactor of entry  $a_{12}$  is

$$\begin{aligned} C_{12} &= (-1)^{1+2} M_{12} \\ &= -M_{12} \\ &= -(-80) \\ &= 80 \end{aligned}$$

Similarly

$$\begin{aligned} C_{13} &= -384 \\ C_{21} &= 7 \\ C_{22} &= -119 \\ C_{23} &= 420 \\ C_{31} &= -3 \\ C_{32} &= 39 \\ C_{33} &= -120 \end{aligned}$$

Hence the matrix of cofactors of  $[A]$  is

$$[C] = \begin{bmatrix} -4 & 80 & -384 \\ 7 & -119 & 420 \\ -3 & 39 & -120 \end{bmatrix}$$

The adjoint of matrix  $[A]$  is  $[C]^T$ ,

$$\begin{aligned} adj(A) &= [C]^T \\ &= \begin{bmatrix} -4 & 7 & -3 \\ 80 & -119 & 39 \\ -384 & 420 & -120 \end{bmatrix} \end{aligned}$$

Hence

$$\begin{aligned} [A]^{-1} &= \frac{1}{\det(A)} adj(A) \\ &= \frac{1}{-84} \begin{bmatrix} -4 & 7 & -3 \\ 80 & -119 & 39 \\ -384 & 420 & -120 \end{bmatrix} \\ &= \begin{bmatrix} 0.04762 & -0.08333 & 0.03571 \\ -0.9524 & 1.417 & -0.4643 \\ 4.571 & -5.000 & 1.429 \end{bmatrix} \end{aligned}$$

**If the inverse of a square matrix [A] exists, is it unique?**

Yes, the inverse of a square matrix is unique, if it exists. The proof is as follows. Assume that the inverse of  $[A]$  is  $[B]$  and if this inverse is not unique, then let another inverse of  $[A]$  exist called  $[C]$ .

If  $[B]$  is the inverse of  $[A]$ , then

$$[B][A]=[I]$$

Multiply both sides by  $[C]$ ,

$$[B][A][C]=[I][C]$$

$$[B][A][C]=[C]$$

Since  $[C]$  is inverse of  $[A]$ ,

$$[A][C]=[I]$$

Multiply both sides by  $[B]$ ,

$$[B][I]=[C]$$

$$[B]=[C]$$

This shows that  $[B]$  and  $[C]$  are the same. So the inverse of  $[A]$  is unique.

**Key Terms:**

*Consistent system*

*Inconsistent system*

*Infinite solutions*

*Unique solution*

*Rank*

*Inverse*

# Chapter 04.06

## Gaussian Elimination

After reading this chapter, you should be able to:

1. solve a set of simultaneous linear equations using Naïve Gauss elimination,
2. learn the pitfalls of the Naïve Gauss elimination method,
3. understand the effect of round-off error when solving a set of linear equations with the Naïve Gauss elimination method,
4. learn how to modify the Naïve Gauss elimination method to the Gaussian elimination with partial pivoting method to avoid pitfalls of the former method,
5. find the determinant of a square matrix using Gaussian elimination, and
6. understand the relationship between the determinant of a coefficient matrix and the solution of simultaneous linear equations.

### How is a set of equations solved numerically?

One of the most popular techniques for solving simultaneous linear equations is the Gaussian elimination method. The approach is designed to solve a general set of  $n$  equations and  $n$  unknowns

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \dots + a_{1n}x_n = b_1$$

$$a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \dots + a_{2n}x_n = b_2$$

⋮

⋮

$$a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \dots + a_{nn}x_n = b_n$$

Gaussian elimination consists of two steps

1. Forward Elimination of Unknowns: In this step, the unknown is eliminated in each equation starting with the first equation. This way, the equations are *reduced* to one equation and one unknown in each equation.
2. Back Substitution: In this step, starting from the last equation, each of the unknowns is found.

### Forward Elimination of Unknowns:

In the first step of forward elimination, the first unknown,  $x_1$  is eliminated from all rows below the first row. The first equation is selected as the pivot equation to eliminate  $x_1$ . So,

to eliminate  $x_1$  in the second equation, one divides the first equation by  $a_{11}$  (hence called the pivot element) and then multiplies it by  $a_{21}$ . This is the same as multiplying the first equation by  $a_{21}/a_{11}$  to give

$$a_{21}x_1 + \frac{a_{21}}{a_{11}}a_{12}x_2 + \dots + \frac{a_{21}}{a_{11}}a_{1n}x_n = \frac{a_{21}}{a_{11}}b_1$$

Now, this equation can be subtracted from the second equation to give

$$\left( a_{22} - \frac{a_{21}}{a_{11}}a_{12} \right)x_2 + \dots + \left( a_{2n} - \frac{a_{21}}{a_{11}}a_{1n} \right)x_n = b_2 - \frac{a_{21}}{a_{11}}b_1$$

or

$$a'_{22}x_2 + \dots + a'_{2n}x_n = b'_2$$

where

$$a'_{22} = a_{22} - \frac{a_{21}}{a_{11}}a_{12}$$

$$\vdots$$

$$a'_{2n} = a_{2n} - \frac{a_{21}}{a_{11}}a_{1n}$$

This procedure of eliminating  $x_1$ , is now repeated for the third equation to the  $n^{\text{th}}$  equation to reduce the set of equations as

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \dots + a_{1n}x_n = b_1$$

$$a'_{22}x_2 + a'_{23}x_3 + \dots + a'_{2n}x_n = b'_2$$

$$a'_{32}x_2 + a'_{33}x_3 + \dots + a'_{3n}x_n = b'_3$$

$$\cdot \quad \cdot \quad \cdot$$

$$\cdot \quad \cdot \quad \cdot$$

$$\cdot \quad \cdot \quad \cdot$$

$$a'_{n2}x_2 + a'_{n3}x_3 + \dots + a'_{nn}x_n = b'_n$$

This is the end of the first step of forward elimination. Now for the second step of forward elimination, we start with the second equation as the pivot equation and  $a'_{22}$  as the pivot element. So, to eliminate  $x_2$  in the third equation, one divides the second equation by  $a'_{22}$  (the pivot element) and then multiply it by  $a'_{32}$ . This is the same as multiplying the second equation by  $a'_{32}/a'_{22}$  and subtracting it from the third equation. This makes the coefficient of  $x_2$  zero in the third equation. The same procedure is now repeated for the fourth equation till the  $n^{\text{th}}$  equation to give

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \dots + a_{1n}x_n = b_1$$

$$a'_{22}x_2 + a'_{23}x_3 + \dots + a'_{2n}x_n = b'_2$$

$$a''_{32}x_2 + \dots + a''_{3n}x_n = b''_3$$

$$\cdot \quad \cdot \quad \cdot$$

$$\cdot \quad \cdot \quad \cdot$$

$$a''_{n2}x_2 + \dots + a''_{nn}x_n = b''_n$$

The next steps of forward elimination are conducted by using the third equation as a pivot equation and so on. That is, there will be a total of  $n - 1$  steps of forward elimination. At the end of  $n - 1$  steps of forward elimination, we get a set of equations that look like

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \dots + a_{1n}x_n &= b_1 \\ a'_{22}x_2 + a'_{23}x_3 + \dots + a'_{2n}x_n &= b'_2 \\ a''_{33}x_3 + \dots + a''_{3n}x_n &= b''_3 \\ &\vdots &&\vdots \\ &\vdots &&\vdots \\ a_{nn}^{(n-1)}x_n &= b_n^{(n-1)} \end{aligned}$$

### Back Substitution:

Now the equations are solved starting from the last equation as it has only one unknown.

$$x_n = \frac{b_n^{(n-1)}}{a_{nn}^{(n-1)}}$$

Then the second last equation, that is the  $(n-1)^{\text{th}}$  equation, has two unknowns:  $x_n$  and  $x_{n-1}$ , but  $x_n$  is already known. This reduces the  $(n-1)^{\text{th}}$  equation also to one unknown. Back substitution hence can be represented for all equations by the formula

$$x_i = \frac{b_i^{(i-1)} - \sum_{j=i+1}^n a_{ij}^{(i-1)}x_j}{a_{ii}^{(i-1)}} \quad \text{for } i = n-1, n-2, \dots, 1$$

and

$$x_n = \frac{b_n^{(n-1)}}{a_{nn}^{(n-1)}}$$

### Example 1

The upward velocity of a rocket is given at three different times in Table 1.

**Table 1** Velocity vs. time data.

Time, $t$ (s)	Velocity, $v$ (m/s)
5	106.8
8	177.2
12	279.2

The velocity data is approximated by a polynomial as

$$v(t) = a_1t^2 + a_2t + a_3, \quad 5 \leq t \leq 12$$

The coefficients  $a_1$ ,  $a_2$ , and  $a_3$  for the above expression are given by

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 106.8 \\ 177.2 \\ 279.2 \end{bmatrix}$$

Find the values of  $a_1$ ,  $a_2$ , and  $a_3$  using the Naïve Gauss elimination method. Find the velocity at  $t = 6, 7.5, 9, 11$  seconds.

### 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 25 and then multiply it by 64, that is, multiply Row 1 by  $64/25 = 2.56$ .

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 12.8 & 2.56 \end{bmatrix} \begin{bmatrix} 106.8 \\ 177.2 \\ 279.2 \end{bmatrix} \times 2.56 \text{ gives Row 1 as}$$

$$\begin{bmatrix} 64 & 12.8 & 2.56 \\ 64 & 12.8 & 2.56 \end{bmatrix} \begin{bmatrix} 273.408 \\ 273.408 \end{bmatrix}$$

Subtract the result from Row 2

$$\begin{array}{r} [64 \quad 8 \quad 1] \quad [177.2] \\ - [64 \quad 12.8 \quad 2.56] \quad [273.408] \\ \hline 0 \quad -4.8 \quad -1.56 \quad -96.208 \end{array}$$

to get the resulting equations as

$$\begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 144 & 12 & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 106.8 \\ -96.208 \\ 279.2 \end{bmatrix}$$

Divide Row 1 by 25 and then multiply it by 144, that is, multiply Row 1 by  $144/25 = 5.76$ .

$$\begin{bmatrix} 25 & 5 & 1 \\ 144 & 28.8 & 5.76 \end{bmatrix} \begin{bmatrix} 106.8 \\ 279.2 \\ 279.2 \end{bmatrix} \times 5.76 \text{ gives Row 1 as}$$

$$\begin{bmatrix} 144 & 28.8 & 5.76 \\ 144 & 28.8 & 5.76 \end{bmatrix} \begin{bmatrix} 615.168 \\ 615.168 \end{bmatrix}$$

Subtract the result from Row 3

$$\begin{array}{r} [144 \quad 12 \quad 1] \quad [279.2] \\ - [144 \quad 28.8 \quad 5.76] \quad [615.168] \\ \hline 0 \quad -16.8 \quad -4.76 \quad -335.968 \end{array}$$

to get the resulting equations as

$$\begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & -16.8 & -4.76 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 106.8 \\ -96.208 \\ -335.968 \end{bmatrix}$$

##### Second step

We now divide Row 2 by  $-4.8$  and then multiply by  $-16.8$ , that is, multiply Row 2 by  $-16.8/-4.8 = 3.5$ .

$$\begin{bmatrix} 0 & -4.8 & -1.56 \\ 0 & -16.8 & -5.46 \end{bmatrix} \begin{bmatrix} -96.208 \\ -335.968 \end{bmatrix} \times 3.5 \text{ gives Row 2 as}$$

$$\begin{bmatrix} 0 & -4.8 & -1.56 \\ 0 & -16.8 & -5.46 \end{bmatrix} \begin{bmatrix} -336.728 \\ -336.728 \end{bmatrix}$$

Subtract the result from Row 3

$$\begin{array}{r} \left[ \begin{array}{ccc|c} 0 & -16.8 & -4.76 & -335.968 \\ - \left[ \begin{array}{ccc|c} 0 & -16.8 & -5.46 & -336.728 \\ \hline 0 & 0 & 0.7 & 0.76 \end{array} \right] \end{array} \right. \\ \left. \begin{array}{c} 106.8 \\ -96.208 \\ 0.76 \end{array} \right] \end{array}$$

to get the resulting equations as

$$\left[ \begin{array}{ccc|c} 25 & 5 & 1 & 106.8 \\ 0 & -4.8 & -1.56 & -96.208 \\ 0 & 0 & 0.7 & 0.76 \end{array} \right] \left[ \begin{array}{c} a_1 \\ a_2 \\ a_3 \end{array} \right] = \left[ \begin{array}{c} 106.8 \\ -96.208 \\ 0.76 \end{array} \right]$$

### Back substitution

From the third equation

$$\begin{aligned} 0.7a_3 &= 0.76 \\ a_3 &= \frac{0.76}{0.7} \\ &= 1.08571 \end{aligned}$$

Substituting the value of  $a_3$  in the second equation,

$$\begin{aligned} -4.8a_2 - 1.56a_3 &= -96.208 \\ a_2 &= \frac{-96.208 + 1.56a_3}{-4.8} \\ &= \frac{-96.208 + 1.56 \times 1.08571}{-4.8} \\ &= 19.6905 \end{aligned}$$

Substituting the value of  $a_2$  and  $a_3$  in the first equation,

$$\begin{aligned} 25a_1 + 5a_2 + a_3 &= 106.8 \\ a_1 &= \frac{106.8 - 5a_2 - a_3}{25} \\ &= \frac{106.8 - 5 \times 19.6905 - 1.08571}{25} \\ &= 0.290472 \end{aligned}$$

Hence the solution vector is

$$\left[ \begin{array}{c} a_1 \\ a_2 \\ a_3 \end{array} \right] = \left[ \begin{array}{c} 0.290472 \\ 19.6905 \\ 1.08571 \end{array} \right]$$

The polynomial that passes through the three data points is then

$$\begin{aligned} v(t) &= a_1 t^2 + a_2 t + a_3 \\ &= 0.290472t^2 + 19.6905t + 1.08571, \quad 5 \leq t \leq 12 \end{aligned}$$

Since we want to find the velocity at  $t = 6, 7.5, 9$  and  $11$  seconds, we could simply substitute each value of  $t$  in  $v(t) = 0.290472t^2 + 19.6905t + 1.08571$  and find the corresponding velocity. For example, at  $t = 6$

$$\begin{aligned}v(6) &= 0.290472(6)^2 + 19.6905(6) + 1.08571 \\&= 129.686 \text{ m/s}\end{aligned}$$

However we could also find all the needed values of velocity at  $t = 6, 7.5, 9, 11$  seconds using matrix multiplication.

$$v(t) = [0.290472 \quad 19.6905 \quad 1.08571] \begin{bmatrix} t^2 \\ t \\ 1 \end{bmatrix}$$

So if we want to find  $v(6), v(7.5), v(9), v(11)$ , it is given by

$$\begin{aligned}[v(6) \ v(7.5) \ v(9) \ v(11)] &= [0.290472 \quad 19.6905 \quad 1.08571] \begin{bmatrix} 6^2 & 7.5^2 & 9^2 & 11^2 \\ 6 & 7.5 & 9 & 11 \\ 1 & 1 & 1 & 1 \end{bmatrix} \\&= [0.290472 \quad 19.6905 \quad 1.08571] \begin{bmatrix} 36 & 56.25 & 81 & 121 \\ 6 & 7.5 & 9 & 11 \\ 1 & 1 & 1 & 1 \end{bmatrix} \\&= [129.686 \quad 165.104 \quad 201.828 \quad 252.828]\end{aligned}$$

$$v(6) = 129.686 \text{ m/s}$$

$$v(7.5) = 165.104 \text{ m/s}$$

$$v(9) = 201.828 \text{ m/s}$$

$$v(11) = 252.828 \text{ m/s}$$

## Example 2

Use Naïve Gauss elimination to solve

$$20x_1 + 15x_2 + 10x_3 = 45$$

$$-3x_1 - 2.249x_2 + 7x_3 = 1.751$$

$$5x_1 + x_2 + 3x_3 = 9$$

Use six significant digits with chopping in your calculations.

### Solution

Working in the matrix form

$$\begin{bmatrix} 20 & 15 & 10 \\ -3 & -2.249 & 7 \\ 5 & 1 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 45 \\ 1.751 \\ 9 \end{bmatrix}$$

### Forward Elimination of Unknowns

#### First step

Divide Row 1 by 20 and then multiply it by  $-3$ , that is, multiply Row 1 by  $-3/20 = -0.15$ .

$([20 \ 15 \ 10] \ [45]) \times -0.15$  gives Row 1 as

$$[-3 \ -2.25 \ -1.5] \quad [-6.75]$$

Subtract the result from Row 2

$$\begin{array}{r} \left[ \begin{array}{ccc|c} -3 & -2.249 & 7 & 1.751 \end{array} \right] \\ - \left[ \begin{array}{ccc|c} -3 & -2.25 & -1.5 & -6.75 \end{array} \right] \\ \hline \left[ \begin{array}{ccc|c} 0 & 0.001 & 8.5 & 8.501 \end{array} \right] \end{array}$$

to get the resulting equations as

$$\left[ \begin{array}{ccc|c} 20 & 15 & 10 & 45 \\ 0 & 0.001 & 8.5 & 8.501 \\ 5 & 1 & 3 & 9 \end{array} \right]$$

Divide Row 1 by 20 and then multiply it by 5, that is, multiply Row 1 by  $5/20 = 0.25$

$$\left[ \begin{array}{ccc|c} [20 & 15 & 10] & [45] \\ [5 & 3.75 & 2.5] & [11.25] \end{array} \right] \times 0.25$$

Subtract the result from Row 3

$$\begin{array}{r} \left[ \begin{array}{ccc|c} 5 & 1 & 3 & 9 \end{array} \right] \\ - \left[ \begin{array}{ccc|c} 5 & 3.75 & 2.5 & 11.25 \end{array} \right] \\ \hline \left[ \begin{array}{ccc|c} 0 & -2.75 & 0.5 & -2.25 \end{array} \right] \end{array}$$

to get the resulting equations as

$$\left[ \begin{array}{ccc|c} 20 & 15 & 10 & 45 \\ 0 & 0.001 & 8.5 & 8.501 \\ 0 & -2.75 & 0.5 & -2.25 \end{array} \right]$$

### Second step

Now for the second step of forward elimination, we will use Row 2 as the pivot equation and eliminate Row 3: Column 2.

Divide Row 2 by 0.001 and then multiply it by  $-2.75$ , that is, multiply Row 2 by  $-2.75/0.001 = -2750$ .

$$\left[ \begin{array}{ccc|c} [0 & 0.001 & 8.5] & [8.501] \\ [0 & -2.75 & -23375] & [-23377.75] \end{array} \right] \times -2750$$

Rewriting within 6 significant digits with chopping

$$\left[ \begin{array}{ccc|c} [0 & -2.75 & -23375] & [-23377.7] \end{array} \right]$$

Subtract the result from Row 3

$$\begin{array}{r} \left[ \begin{array}{ccc|c} 0 & -2.75 & 0.5 & -2.25 \end{array} \right] \\ - \left[ \begin{array}{ccc|c} 0 & -2.75 & -23375 & -23377.7 \end{array} \right] \\ \hline \left[ \begin{array}{ccc|c} 0 & 0 & 23375.5 & 23375.45 \end{array} \right] \end{array}$$

Rewriting within 6 significant digits with chopping

$$\left[ \begin{array}{ccc|c} [0 & 0 & 23375.5] & [-23375.4] \end{array} \right]$$

to get the resulting equations as

$$\left[ \begin{array}{ccc|c} 20 & 15 & 10 & 45 \\ 0 & 0.001 & 8.5 & 8.501 \\ 0 & 0 & 23375.5 & 23375.4 \end{array} \right]$$

This is the end of the forward elimination steps.

### Back substitution

We can now solve the above equations by back substitution. From the third equation,

$$23375.5x_3 = 23375.4$$

$$\begin{aligned} x_3 &= \frac{23375.4}{23375.5} \\ &= 0.999995 \end{aligned}$$

Substituting the value of  $x_3$  in the second equation

$$0.001x_2 + 8.5x_3 = 8.501$$

$$\begin{aligned} x_2 &= \frac{8.501 - 8.5x_3}{0.001} \\ &= \frac{8.501 - 8.5 \times 0.999995}{0.001} \\ &= \frac{8.501 - 8.49995}{0.001} \\ &= \frac{0.00105}{0.001} \\ &= 1.05 \end{aligned}$$

Substituting the value of  $x_3$  and  $x_2$  in the first equation,

$$20x_1 + 15x_2 + 10x_3 = 45$$

$$\begin{aligned} x_1 &= \frac{45 - 15x_2 - 10x_3}{20} \\ &= \frac{45 - 15 \times 1.05 - 10 \times 0.999995}{20} \\ &= \frac{45 - 15.75 - 9.99995}{20} \\ &= \frac{29.25 - 9.99995}{20} \\ &= \frac{19.2500}{20} \\ &= 0.9625 \end{aligned}$$

Hence the solution is

$$\begin{aligned} [X] &= \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \\ &= \begin{bmatrix} 0.9625 \\ 1.05 \\ 0.999995 \end{bmatrix} \end{aligned}$$

Compare this with the exact solution of

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

$$= \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

### Are there any pitfalls of the Naïve Gauss elimination method?

Yes, there are two pitfalls of the Naïve Gauss elimination method.

**Division by zero:** It is possible for division by zero to occur during the beginning of the  $n-1$  steps of forward elimination.

For example

$$5x_2 + 6x_3 = 11$$

$$4x_1 + 5x_2 + 7x_3 = 16$$

$$9x_1 + 2x_2 + 3x_3 = 15$$

will result in division by zero in the first step of forward elimination as the coefficient of  $x_1$  in the first equation is zero as is evident when we write the equations in matrix form.

$$\left[ \begin{array}{ccc|c} 0 & 5 & 6 \\ 4 & 5 & 7 \\ 9 & 2 & 3 \end{array} \right] \left[ \begin{array}{c} x_1 \\ x_2 \\ x_3 \end{array} \right] = \left[ \begin{array}{c} 11 \\ 16 \\ 15 \end{array} \right]$$

But what about the equations below: Is division by zero a problem?

$$5x_1 + 6x_2 + 7x_3 = 18$$

$$10x_1 + 12x_2 + 3x_3 = 25$$

$$20x_1 + 17x_2 + 19x_3 = 56$$

Written in matrix form,

$$\left[ \begin{array}{ccc|c} 5 & 6 & 7 \\ 10 & 12 & 3 \\ 20 & 17 & 19 \end{array} \right] \left[ \begin{array}{c} x_1 \\ x_2 \\ x_3 \end{array} \right] = \left[ \begin{array}{c} 18 \\ 25 \\ 56 \end{array} \right]$$

there is no issue of division by zero in the first step of forward elimination. The pivot element is the coefficient of  $x_1$  in the first equation, 5, and that is a non-zero number. However, at the end of the first step of forward elimination, we get the following equations in matrix form

$$\left[ \begin{array}{ccc|c} 5 & 6 & 7 \\ 0 & 0 & -11 \\ 0 & -7 & -9 \end{array} \right] \left[ \begin{array}{c} x_1 \\ x_2 \\ x_3 \end{array} \right] = \left[ \begin{array}{c} 18 \\ -11 \\ -16 \end{array} \right]$$

Now at the beginning of the 2<sup>nd</sup> step of forward elimination, the coefficient of  $x_2$  in Equation 2 would be used as the pivot element. That element is zero and hence would create the division by zero problem.

So it is important to consider that the possibility of division by zero can occur at the beginning of any step of forward elimination.

**Round-off error:** The Naïve Gauss elimination method is prone to round-off errors. This is true when there are large numbers of equations as errors propagate. Also, if there is subtraction of numbers from each other, it may create large errors. See the example below.

### Example 3

Remember Example 2 where we used Naïve Gauss elimination to solve

$$\begin{aligned} 20x_1 + 15x_2 + 10x_3 &= 45 \\ -3x_1 - 2.249x_2 + 7x_3 &= 1.751 \\ 5x_1 + x_2 + 3x_3 &= 9 \end{aligned}$$

using six significant digits with chopping in your calculations? Repeat the problem, but now use five significant digits with chopping in your calculations.

#### Solution

Writing in the matrix form

$$\left[ \begin{array}{ccc|c} 20 & 15 & 10 & 45 \\ -3 & -2.249 & 7 & 1.751 \\ 5 & 1 & 3 & 9 \end{array} \right]$$

#### Forward Elimination of Unknowns

##### First step

Divide Row 1 by 20 and then multiply it by  $-3$ , that is, multiply Row 1 by  $-3/20 = -0.15$ .

$$\begin{array}{cc} ([20 \ 15 \ 10] \ [45]) \times -0.15 & \text{gives Row 1 as} \\ [-3 \ -2.25 \ -1.5] & [-6.75] \end{array}$$

Subtract the result from Row 2

$$\begin{array}{r} \begin{array}{ccc|c} & [-3 \ -2.249 \ 7] & [1.751] \\ - & [-3 \ -2.25 \ -1.5] & [-6.75] \\ \hline & 0 \ 0.001 \ 8.5 & 8.501 \end{array} \end{array}$$

to get the resulting equations as

$$\left[ \begin{array}{ccc|c} 20 & 15 & 10 & 45 \\ 0 & 0.001 & 8.5 & 8.501 \\ 5 & 1 & 3 & 9 \end{array} \right]$$

Divide Row 1 by 20 and then multiply it by 5, that is, multiply Row 1 by  $5/20 = 0.25$ .

$$\begin{array}{cc} ([20 \ 15 \ 10] \ [45]) \times 0.25 & \text{gives Row 1 as} \\ [5 \ 3.75 \ 2.5] & [11.25] \end{array}$$

Subtract the result from Row 3

$$\begin{array}{r} \begin{array}{ccc|c} & [5 \ 1 \ 3] & [9] \\ - & [5 \ 3.75 \ 2.5] & [11.25] \\ \hline & 0 \ -2.75 \ 0.5 & -2.25 \end{array} \end{array}$$

to get the resulting equations as

$$\begin{bmatrix} 20 & 15 & 10 \\ 0 & 0.001 & 8.5 \\ 0 & -2.75 & 0.5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 45 \\ 8.501 \\ -2.25 \end{bmatrix}$$

### Second step

Now for the second step of forward elimination, we will use Row 2 as the pivot equation and eliminate Row 3: Column 2.

Divide Row 2 by 0.001 and then multiply it by  $-2.75$ , that is, multiply Row 2 by  $-2.75/0.001 = -2750$ .

$$([0 \ 0.001 \ 8.5] [8.501]) \times -2750 \text{ gives Row 2 as} \\ [0 \ -2.75 \ -23375] \quad [-23377.75]$$

Rewriting within 5 significant digits with chopping

$$[0 \ -2.75 \ -23375] \quad [-23377]$$

Subtract the result from Row 3

$$\begin{array}{r} [0 \ -2.75 \ 0.5] \quad [-2.25] \\ - [0 \ -2.75 \ -23375] \quad [-23377] \\ \hline 0 \ 0 \ 23375 \ 23374 \end{array}$$

Rewriting within 6 significant digits with chopping

$$[0 \ 0 \ 23375] \quad [-23374]$$

to get the resulting equations as

$$\begin{bmatrix} 20 & 15 & 10 \\ 0 & 0.001 & 8.5 \\ 0 & 0 & 23375 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 45 \\ 8.501 \\ 23374 \end{bmatrix}$$

This is the end of the forward elimination steps.

### **Back substitution**

We can now solve the above equations by back substitution. From the third equation,

$$23375x_3 = 23374$$

$$x_3 = \frac{23374}{23375} \\ = 0.99995$$

Substituting the value of  $x_3$  in the second equation

$$\begin{aligned} 0.001x_2 + 8.5x_3 &= 8.501 \\ x_2 &= \frac{8.501 - 8.5x_3}{0.001} \\ &= \frac{8.501 - 8.5 \times 0.99995}{0.001} \\ &= \frac{8.501 - 8.499575}{0.001} \\ &= \frac{8.501 - 8.4995}{0.001} \end{aligned}$$

$$\begin{aligned} &= \frac{0.0015}{0.001} \\ &= 1.5 \end{aligned}$$

Substituting the value of  $x_3$  and  $x_2$  in the first equation,

$$\begin{aligned} 20x_1 + 15x_2 + 10x_3 &= 45 \\ x_1 &= \frac{45 - 15x_2 - 10x_3}{20} \\ &= \frac{45 - 15 \times 1.5 - 10 \times 0.99995}{20} \\ &= \frac{45 - 22.5 - 9.9995}{20} \\ &= \frac{22.5 - 9.9995}{20} \\ &= \frac{12.5005}{20} \\ &= \frac{12.500}{20} \\ &= 0.625 \end{aligned}$$

Hence the solution is

$$\begin{aligned} [X] &= \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \\ &= \begin{bmatrix} 0.625 \\ 1.5 \\ 0.99995 \end{bmatrix} \end{aligned}$$

Compare this with the exact solution of

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

### What are some techniques for improving the Naïve Gauss elimination method?

As seen in Example 3, round off errors were large when five significant digits were used as opposed to six significant digits. One method of decreasing the round-off error would be to use more significant digits, that is, use double or quad precision for representing the numbers. However, this would not avoid possible division by zero errors in the Naïve Gauss elimination method. To avoid division by zero as well as reduce (not eliminate) round-off error, Gaussian elimination with partial pivoting is the method of choice.

### How does Gaussian elimination with partial pivoting differ from Naïve Gauss elimination?

The two methods are the same, except in the beginning of each step of forward elimination, a row switching is done based on the following criterion. If there are  $n$  equations, then there are  $n-1$  forward elimination steps. At the beginning of the  $k^{\text{th}}$  step of forward elimination, one finds the maximum of

$$|a_{kk}|, |a_{k+1,k}|, \dots, |a_{nn}|$$

Then if the maximum of these values is  $|a_{pk}|$  in the  $p^{\text{th}}$  row,  $k \leq p \leq n$ , then switch rows  $p$  and  $k$ .

The other steps of forward elimination are the same as the Naïve Gauss elimination method. The back substitution steps stay exactly the same as the Naïve Gauss elimination method.

#### Example 4

In the previous two examples, we used Naïve Gauss elimination to solve

$$\begin{aligned} 20x_1 + 15x_2 + 10x_3 &= 45 \\ -3x_1 - 2.249x_2 + 7x_3 &= 1.751 \\ 5x_1 + x_2 + 3x_3 &= 9 \end{aligned}$$

using five and six significant digits with chopping in the calculations. Using five significant digits with chopping, the solution found was

$$\begin{aligned} [X] &= \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \\ &= \begin{bmatrix} 0.625 \\ 1.5 \\ 0.99995 \end{bmatrix} \end{aligned}$$

This is different from the exact solution of

$$\begin{aligned} [X] &= \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \\ &= \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \end{aligned}$$

Find the solution using Gaussian elimination with partial pivoting using five significant digits with chopping in your calculations.

#### Solution

$$\begin{bmatrix} 20 & 15 & 10 \\ -3 & -2.249 & 7 \\ 5 & 1 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 45 \\ 1.751 \\ 9 \end{bmatrix}$$

### Forward Elimination of Unknowns

Now for the first step of forward elimination, the absolute value of the first column elements below Row 1 is

$$|20|, |-3|, |5|$$

or

$$20, 3, 5$$

So the largest absolute value is in the Row 1. So as per Gaussian elimination with partial pivoting, the switch is between Row 1 and Row 1 to give

$$\begin{bmatrix} 20 & 15 & 10 \\ -3 & -2.249 & 7 \\ 5 & 1 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 45 \\ 1.751 \\ 9 \end{bmatrix}$$

Divide Row 1 by 20 and then multiply it by  $-3$ , that is, multiply Row 1 by  $-3/20 = -0.15$ .

$$([20 \ 15 \ 10] [45]) \times -0.15 \text{ gives Row 1 as}$$

$$\begin{bmatrix} -3 & -2.25 & -1.5 \\ -3 & -2.25 & -1.5 \end{bmatrix} \begin{bmatrix} 1.751 \\ -6.75 \end{bmatrix}$$

Subtract the result from Row 2

$$\begin{array}{r} \begin{bmatrix} -3 & -2.249 & 7 \end{bmatrix} \begin{bmatrix} 1.751 \end{bmatrix} \\ - \begin{bmatrix} -3 & -2.25 & -1.5 \end{bmatrix} \begin{bmatrix} -6.75 \end{bmatrix} \\ \hline 0 & 0.001 & 8.5 & 8.501 \end{array}$$

to get the resulting equations as

$$\begin{bmatrix} 20 & 15 & 10 \\ 0 & 0.001 & 8.5 \\ 5 & 1 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 45 \\ 8.501 \\ 9 \end{bmatrix}$$

Divide Row 1 by 20 and then multiply it by 5, that is, multiply Row 1 by  $5/20 = 0.25$ .

$$([20 \ 15 \ 10] [45]) \times 0.25 \text{ gives Row 1 as}$$

$$\begin{bmatrix} 5 & 3.75 & 2.5 \\ 5 & 3.75 & 2.5 \end{bmatrix} \begin{bmatrix} 11.25 \end{bmatrix}$$

Subtract the result from Row 3

$$\begin{array}{r} \begin{bmatrix} 5 & 1 & 3 \end{bmatrix} \begin{bmatrix} 9 \end{bmatrix} \\ - \begin{bmatrix} 5 & 3.75 & 2.5 \end{bmatrix} \begin{bmatrix} 11.25 \end{bmatrix} \\ \hline 0 & -2.75 & 0.5 & -2.25 \end{array}$$

to get the resulting equations as

$$\begin{bmatrix} 20 & 15 & 10 \\ 0 & 0.001 & 8.5 \\ 0 & -2.75 & 0.5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 45 \\ 8.501 \\ -2.25 \end{bmatrix}$$

This is the end of the first step of forward elimination.

Now for the second step of forward elimination, the absolute value of the second column elements below Row 1 is

$$|0.001|, |-2.75|$$

or

$$0.001, 2.75$$

So the largest absolute value is in Row 3. So Row 2 is switched with Row 3 to give

$$\begin{bmatrix} 20 & 15 & 10 \\ 0 & -2.75 & 0.5 \\ 0 & 0.001 & 8.5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 45 \\ -2.25 \\ 8.501 \end{bmatrix}$$

Divide Row 2 by  $-2.75$  and then multiply it by 0.001, that is, multiply Row 2 by  $0.001/-2.75 = -0.00036363$ .

$$\begin{bmatrix} 0 & -2.75 & 0.5 \\ 0 & 0.00099998 & -0.00018182 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -2.25 \\ 0.00081816 \end{bmatrix}$$

Subtract the result from Row 3

$$\begin{array}{r} [0 \quad 0.001 \quad 8.5] \quad [8.501] \\ - [0 \quad 0.00099998 \quad -0.00018182] \quad [0.00081816] \\ \hline 0 \quad 0 \quad 8.50018182 \quad 8.50018184 \end{array}$$

Rewriting within 5 significant digits with chopping

$$\begin{bmatrix} 0 & 0 & 8.5001 \end{bmatrix} \quad [8.5001]$$

to get the resulting equations as

$$\begin{bmatrix} 20 & 15 & 10 \\ 0 & -2.75 & 0.5 \\ 0 & 0 & 8.5001 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 45 \\ -2.25 \\ 8.5001 \end{bmatrix}$$

### Back substitution

$$8.5001x_3 = 8.5001$$

$$\begin{aligned} x_3 &= \frac{8.5001}{8.5001} \\ &= 1 \end{aligned}$$

Substituting the value of  $x_3$  in Row 2

$$-2.75x_2 + 0.5x_3 = -2.25$$

$$\begin{aligned} x_2 &= \frac{-2.25 - 0.5x_3}{-2.75} \\ &= \frac{-2.25 - 0.5 \times 1}{-2.75} \\ &= \frac{-2.25 - 0.5}{-2.75} \\ &= \frac{-2.75}{-2.75} \\ &= 1 \end{aligned}$$

Substituting the value of  $x_3$  and  $x_2$  in Row 1

$$20x_1 + 15x_2 + 10x_3 = 45$$

$$x_1 = \frac{45 - 15x_2 - 10x_3}{20}$$

$$\begin{aligned}
 &= \frac{45 - 15 \times 1 - 10 \times 1}{20} \\
 &= \frac{45 - 15 - 10}{20} \\
 &= \frac{30 - 10}{20} \\
 &= \frac{20}{20} \\
 &= 1
 \end{aligned}$$

So the solution is

$$\begin{aligned}
 [X] &= \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \\
 &= \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}
 \end{aligned}$$

This, in fact, is the exact solution. By coincidence only, in this case, the round-off error is fully removed.

### Can we use Naïve Gauss elimination methods to find the determinant of a square matrix?

One of the more efficient ways to find the determinant of a square matrix is by taking advantage of the following two theorems on a determinant of matrices coupled with Naïve Gauss elimination.

#### Theorem 1:

Let  $[A]$  be a  $n \times n$  matrix. Then, if  $[B]$  is a  $n \times n$  matrix that results from adding or subtracting a multiple of one row to another row, then  $\det(A) = \det(B)$  (The same is true for column operations also).

#### Theorem 2:

Let  $[A]$  be a  $n \times n$  matrix that is upper triangular, lower triangular or diagonal, then

$$\begin{aligned}
 \det(A) &= a_{11} \times a_{22} \times \dots \times a_{ii} \times \dots \times a_{nn} \\
 &= \prod_{i=1}^n a_{ii}
 \end{aligned}$$

This implies that if we apply the forward elimination steps of the Naïve Gauss elimination method, the determinant of the matrix stays the same according to Theorem 1. Then since at the end of the forward elimination steps, the resulting matrix is upper triangular, the determinant will be given by Theorem 2.

**Example 5**

Find the determinant of

$$[A] = \begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix}$$

**Solution**

Remember in Example 1, we conducted the steps of forward elimination of unknowns using the Naïve Gauss elimination method on  $[A]$  to give

$$[B] = \begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & 0 & 0.7 \end{bmatrix}$$

According to Theorem 2

$$\begin{aligned} \det(A) &= \det(B) \\ &= 25 \times (-4.8) \times 0.7 \\ &= -84.00 \end{aligned}$$

**What if I cannot find the determinant of the matrix using the Naïve Gauss elimination method, for example, if I get division by zero problems during the Naïve Gauss elimination method?**

Well, you can apply Gaussian elimination with partial pivoting. However, the determinant of the resulting upper triangular matrix may differ by a sign. The following theorem applies in addition to the previous two to find the determinant of a square matrix.

**Theorem 3:**

Let  $[A]$  be a  $n \times n$  matrix. Then, if  $[B]$  is a matrix that results from switching one row with another row, then  $\det(B) = -\det(A)$ .

**Example 6**

Find the determinant of

$$[A] = \begin{bmatrix} 10 & -7 & 0 \\ -3 & 2.099 & 6 \\ 5 & -1 & 5 \end{bmatrix}$$

**Solution**

The end of the forward elimination steps of Gaussian elimination with partial pivoting, we would obtain

$$[B] = \begin{bmatrix} 10 & -7 & 0 \\ 0 & 2.5 & 5 \\ 0 & 0 & 6.002 \end{bmatrix}$$

$$\det(B) = 10 \times 2.5 \times 6.002$$

$$= 150.05$$

Since rows were switched once during the forward elimination steps of Gaussian elimination with partial pivoting,

$$\begin{aligned}\det(A) &= -\det(B) \\ &= -150.05\end{aligned}$$

### Example 7

Prove

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

### Solution

$$\begin{aligned}[A][A]^{-1} &= [I] \\ \det(A) \det(A^{-1}) &= \det(I) \\ \det(A) \det(A^{-1}) &= 1 \\ \det(A) &= \frac{1}{\det(A^{-1})}\end{aligned}$$

If  $[A]$  is a  $n \times n$  matrix and  $\det(A) \neq 0$ , what other statements are equivalent to it?

1.  $[A]$  is invertible.
2.  $[A]^{-1}$  exists.
3.  $[A][X] = [C]$  has a unique solution.
4.  $[A][X] = [0]$  solution is  $[X] = [\bar{0}]$ .
5.  $[A][A]^{-1} = [I] = [A]^{-1}[A]$ .

### Key Terms:

*Naïve Gauss Elimination*

*Partial Pivoting*

*Determinant*

## Chapter 04.07

# LU Decomposition

After reading this chapter, you should be able to:

1. identify when LU decomposition is numerically more efficient than Gaussian elimination,
2. decompose a nonsingular matrix into LU, and
3. show how LU decomposition is used to find the inverse of a matrix.

**I hear about LU decomposition used as a method to solve a set of simultaneous linear equations. What is it?**

We already studied two numerical methods of finding the solution to simultaneous linear equations – Naïve Gauss elimination and Gaussian elimination with partial pivoting. Then, why do we need to learn another method? To appreciate why LU decomposition could be a better choice than the Gauss elimination techniques in some cases, let us discuss first what LU decomposition is about.

For a nonsingular matrix  $[A]$  on which one can successfully conduct the Naïve Gauss elimination forward elimination steps, one can always write it as

$$[A] = [L][U]$$

where

$[L]$  = Lower triangular matrix

$[U]$  = Upper triangular matrix

Then if one is solving a set of equations

$$[A][X] = [C],$$

then

$$[L][U][X] = [C] \text{ as } ([A] = [L][U])$$

Multiplying both sides by  $[L]^{-1}$ ,

$$[L]^{-1}[L][U][X] = [L]^{-1}[C]$$

$$[I][U][X] = [L]^{-1}[C] \text{ as } ([L]^{-1}[L] = [I])$$

$$[U][X] = [L]^{-1}[C] \text{ as } ([I][U] = [U])$$

Let

$$[L]^{-1}[C] = [Z]$$

then

$$[L][Z]=[C] \quad (1)$$

and

$$[U][X]=[Z] \quad (2)$$

So we can solve Equation (1) first for  $[Z]$  by using forward substitution and then use Equation (2) to calculate the solution vector  $[X]$  by back substitution.

**This is all exciting but LU decomposition looks more complicated than Gaussian elimination. Do we use LU decomposition because it is computationally more efficient than Gaussian elimination to solve a set of n equations given by  $[A][X]=[C]$ ?**

For a square matrix  $[A]$  of  $n \times n$  size, the computational time<sup>1</sup>  $CT|_{DE}$  to decompose the  $[A]$  matrix to  $[L][U]$  form is given by

$$CT|_{DE} = T\left(\frac{8n^3}{3} + 4n^2 - \frac{20n}{3}\right),$$

where

$$T = \text{clock cycle time}^2.$$

The computational time  $CT|_{FS}$  to solve by forward substitution  $[L][Z]=[C]$  is given by

$$CT|_{FS} = T(4n^2 - 4n)$$

The computational time  $CT|_{BS}$  to solve by back substitution  $[U][X]=[Z]$  is given by

$$CT|_{BS} = T(4n^2 + 12n)$$

So, the total computational time to solve a set of equations by LU decomposition is

$$\begin{aligned} CT|_{LU} &= CT|_{DE} + CT|_{FS} + CT|_{BS} \\ &= T\left(\frac{8n^3}{3} + 4n^2 - \frac{20n}{3}\right) + T(4n^2 - 4n) + T(4n^2 + 12n) \\ &= T\left(\frac{8n^3}{3} + 12n^2 + \frac{4n}{3}\right) \end{aligned}$$

Now let us look at the computational time taken by Gaussian elimination. The computational time  $CT|_{FE}$  for the forward elimination part,

$$CT|_{FE} = T\left(\frac{8n^3}{3} + 8n^2 - \frac{32n}{3}\right),$$

<sup>1</sup> The time is calculated by first separately calculating the number of additions, subtractions, multiplications, and divisions in a procedure such as back substitution, etc. We then assume 4 clock cycles each for an add, subtract, or multiply operation, and 16 clock cycles for a divide operation as is the case for a typical AMD®-K7 chip.

[http://www.isi.edu/~draper/papers/mwscas07\\_kw.pdf](http://www.isi.edu/~draper/papers/mwscas07_kw.pdf)

<sup>2</sup> As an example, a 1.2 GHz CPU has a clock cycle of  $1/(1.2 \times 10^9) = 0.833333\text{ns}$

and the computational time  $CT|_{BS}$  for the back substitution part is

$$CT|_{BS} = T(4n^2 + 12n)$$

So, the total computational time  $CT|_{GE}$  to solve a set of equations by Gaussian Elimination is

$$\begin{aligned} CT|_{GE} &= CT|_{FE} + CT|_{BS} \\ &= T\left(\frac{8n^3}{3} + 8n^2 - \frac{32n}{3}\right) + T(4n^2 + 12n) \\ &= T\left(\frac{8n^3}{3} + 12n^2 + \frac{4n}{3}\right) \end{aligned}$$

The computational time for Gaussian elimination and LU decomposition is identical.

**This has confused me further! Why learn LU decomposition method when it takes the same computational time as Gaussian elimination, and that too when the two methods are closely related. Please convince me that LU decomposition has its place in solving linear equations!**

We have the knowledge now to convince you that LU decomposition method has its place in the solution of simultaneous linear equations. Let us look at an example where the LU decomposition method is computationally more efficient than Gaussian elimination. Remember in trying to find the inverse of the matrix  $[A]$  in Chapter 04.05, the problem reduces to solving  $n$  sets of equations with the  $n$  columns of the identity matrix as the RHS vector. For calculations of each column of the inverse of the  $[A]$  matrix, the coefficient matrix  $[A]$  matrix in the set of equation  $[A][X]=[C]$  does not change. So if we use the LU decomposition method, the  $[A]=[L][U]$  decomposition needs to be done only once, the forward substitution (Equation 1)  $n$  times, and the back substitution (Equation 2)  $n$  times.

Therefore, the total computational time  $CT|_{inverseLU}$  required to find the inverse of a matrix using LU decomposition is

$$\begin{aligned} CT|_{inverseLU} &= 1 \times CT|_{DE} + n \times CT|_{FS} + n \times CT|_{BS} \\ &= 1 \times T\left(\frac{8n^3}{3} + 4n^2 - \frac{20n}{3}\right) + n \times T(4n^2 - 4n) + n \times T(4n^2 + 12n) \\ &= T\left(\frac{32n^3}{3} + 12n^2 - \frac{20n}{3}\right) \end{aligned}$$

In comparison, if Gaussian elimination method were used to find the inverse of a matrix, the forward elimination as well as the back substitution will have to be done  $n$  times. The total computational time  $CT|_{inverseGE}$  required to find the inverse of a matrix by using Gaussian elimination then is

$$\begin{aligned} CT|_{inverseGE} &= n \times CT|_{FE} + n \times CT|_{BS} \\ &= n \times T\left(\frac{8n^3}{3} + 8n^2 - \frac{32n}{3}\right) + n \times T(4n^2 + 12n) \end{aligned}$$

$$= T \left( \frac{8n^4}{3} + 12n^3 + \frac{4n^2}{3} \right)$$

Clearly for large  $n$ ,  $CT|_{\text{inverseGE}} \gg CT|_{\text{inverseLU}}$  as  $CT|_{\text{inverseGE}}$  has the dominating terms of  $n^4$  and  $CT|_{\text{inverseLU}}$  has the dominating terms of  $n^3$ . For large values of  $n$ , Gaussian elimination method would take more computational time (approximately  $n/4$  times – prove it) than the LU decomposition method. Typical values of the ratio of the computational time for different values of  $n$  are given in Table 1.

**Table 1** Comparing computational times of finding inverse of a matrix using LU decomposition and Gaussian elimination.

$n$	10	100	1000	10000
$CT _{\text{inverseGE}}/CT _{\text{inverseLU}}$	3.28	25.83	250.8	2501

Are you convinced now that LU decomposition has its place in solving systems of equations? We are now ready to answer other curious questions such as

- 1) How do I find LU matrices for a nonsingular matrix  $[A]$ ?
- 2) How do I conduct forward and back substitution steps of Equations (1) and (2), respectively?

### How do I decompose a non-singular matrix $[A]$ , that is, how do I find $[A] = [L][U]$ ?

If forward elimination steps of the Naïve Gauss elimination methods can be applied on a nonsingular matrix, then  $[A]$  can be decomposed into LU as

$$\begin{aligned} [A] &= \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \dots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & \dots & 0 \\ \ell_{21} & 1 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ \ell_{n1} & \ell_{n2} & \dots & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & \dots & u_{1n} \\ 0 & u_{22} & \dots & u_{2n} \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & u_{nn} \end{bmatrix} \end{aligned}$$

The elements of the  $[U]$  matrix are exactly the same as the coefficient matrix one obtains at the end of the forward elimination steps in Naïve Gauss elimination.

The lower triangular matrix  $[L]$  has 1 in its diagonal entries. The non-zero elements on the non-diagonal elements in  $[L]$  are multipliers that made the corresponding entries zero in the upper triangular matrix  $[U]$  during forward elimination.

Let us look at this using the same example as used in Naïve Gaussian elimination.

**Example 1**

Find the LU decomposition of the matrix

$$[A] = \begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix}$$

**Solution**

$$\begin{aligned} [A] &= [L][U] \\ &= \begin{bmatrix} 1 & 0 & 0 \\ \ell_{21} & 1 & 0 \\ \ell_{31} & \ell_{32} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix} \end{aligned}$$

The  $[U]$  matrix is the same as found at the end of the forward elimination of Naïve Gauss elimination method, that is

$$[U] = \begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & 0 & 0.7 \end{bmatrix}$$

To find  $\ell_{21}$  and  $\ell_{31}$ , find the multiplier that was used to make the  $a_{21}$  and  $a_{31}$  elements zero in the first step of forward elimination of the Naïve Gauss elimination method. It was

$$\begin{aligned} \ell_{21} &= \frac{64}{25} \\ &= 2.56 \\ \ell_{31} &= \frac{144}{25} \\ &= 5.76 \end{aligned}$$

To find  $\ell_{32}$ , what multiplier was used to make  $a_{32}$  element zero? Remember  $a_{32}$  element was made zero in the second step of forward elimination. The  $[A]$  matrix at the beginning of the second step of forward elimination was

$$\begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & -16.8 & -4.76 \end{bmatrix}$$

So

$$\begin{aligned} \ell_{32} &= \frac{-16.8}{-4.8} \\ &= 3.5 \end{aligned}$$

Hence

$$[L] = \begin{bmatrix} 1 & 0 & 0 \\ 2.56 & 1 & 0 \\ 5.76 & 3.5 & 1 \end{bmatrix}$$

Confirm  $[L][U]=[A]$ .

$$\begin{aligned}[L][U] &= \begin{bmatrix} 1 & 0 & 0 \\ 2.56 & 1 & 0 \\ 5.76 & 3.5 & 1 \end{bmatrix} \begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & 0 & 0.7 \end{bmatrix} \\ &= \begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix}\end{aligned}$$

**Example 2**

Use the LU decomposition method to solve the following simultaneous linear equations.

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 106.8 \\ 177.2 \\ 279.2 \end{bmatrix}$$

**Solution**

Recall that

$$[A][X] = [C]$$

and if

$$[A] = [L][U]$$

then first solving

$$[L][Z] = [C]$$

and then

$$[U][X] = [Z]$$

gives the solution vector  $[X]$ .

Now in the previous example, we showed

$$[A] = [L][U]$$

$$\begin{bmatrix} 1 & 0 & 0 \\ 2.56 & 1 & 0 \\ 5.76 & 3.5 & 1 \end{bmatrix} \begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & 0 & 0.7 \end{bmatrix}$$

First solve

$$[L][Z] = [C]$$

$$\begin{bmatrix} 1 & 0 & 0 \\ 2.56 & 1 & 0 \\ 5.76 & 3.5 & 1 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} 106.8 \\ 177.2 \\ 279.2 \end{bmatrix}$$

to give

$$z_1 = 106.8$$

$$2.56z_1 + z_2 = 177.2$$

$$5.76z_1 + 3.5z_2 + z_3 = 279.2$$

Forward substitution starting from the first equation gives

$$z_1 = 106.8$$

$$\begin{aligned}
 z_2 &= 177.2 - 2.56z_1 \\
 &= 177.2 - 2.56 \times 106.8 \\
 &= -96.208 \\
 z_3 &= 279.2 - 5.76z_1 - 3.5z_2 \\
 &= 279.2 - 5.76 \times 106.8 - 3.5 \times (-96.208) \\
 &= 0.76
 \end{aligned}$$

Hence

$$\begin{aligned}
 [Z] &= \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} \\
 &= \begin{bmatrix} 106.8 \\ -96.208 \\ 0.76 \end{bmatrix}
 \end{aligned}$$

This matrix is same as the right hand side obtained at the end of the forward elimination steps of Naïve Gauss elimination method. Is this a coincidence?

Now solve

$$\begin{aligned}
 [U][X] &= [Z] \\
 \begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & 0 & 0.7 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} &= \begin{bmatrix} 106.8 \\ -96.208 \\ 0.76 \end{bmatrix} \\
 25a_1 + 5a_2 + a_3 &= 106.8 \\
 -4.8a_2 - 1.56a_3 &= -96.208 \\
 0.7a_3 &= 0.76
 \end{aligned}$$

From the third equation

$$\begin{aligned}
 0.7a_3 &= 0.76 \\
 a_3 &= \frac{0.76}{0.7} \\
 &= 1.0857
 \end{aligned}$$

Substituting the value of  $a_3$  in the second equation,

$$\begin{aligned}
 -4.8a_2 - 1.56a_3 &= -96.208 \\
 a_2 &= \frac{-96.208 + 1.56a_3}{-4.8} \\
 &= \frac{-96.208 + 1.56 \times 1.0857}{-4.8} \\
 &= 19.691
 \end{aligned}$$

Substituting the value of  $a_2$  and  $a_3$  in the first equation,

$$\begin{aligned}
 25a_1 + 5a_2 + a_3 &= 106.8 \\
 a_1 &= \frac{106.8 - 5a_2 - a_3}{25}
 \end{aligned}$$

$$= \frac{106.8 - 5 \times 19.691 - 1.0857}{25} \\ = 0.29048$$

Hence the solution vector is

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 0.29048 \\ 19.691 \\ 1.0857 \end{bmatrix}$$

### How do I find the inverse of a square matrix using LU decomposition?

A matrix  $[B]$  is the inverse of  $[A]$  if

$$[A][B] = [I] = [B][A].$$

How can we use LU decomposition to find the inverse of the matrix? Assume the first column of  $[B]$  (the inverse of  $[A]$ ) is

$$[b_{11} \ b_{12} \ \dots \ \dots \ b_{n1}]^T$$

Then from the above definition of an inverse and the definition of matrix multiplication

$$[A] \begin{bmatrix} b_{11} \\ b_{21} \\ \vdots \\ b_{n1} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Similarly the second column of  $[B]$  is given by

$$[A] \begin{bmatrix} b_{12} \\ b_{22} \\ \vdots \\ b_{n2} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}$$

Similarly, all columns of  $[B]$  can be found by solving  $n$  different sets of equations with the column of the right hand side being the  $n$  columns of the identity matrix.

### Example 3

Use LU decomposition to find the inverse of

$$[A] = \begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix}$$

### Solution

Knowing that

$$[A] = [L][U] \\ = \begin{bmatrix} 1 & 0 & 0 \\ 2.56 & 1 & 0 \\ 5.76 & 3.5 & 1 \end{bmatrix} \begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & 0 & 0.7 \end{bmatrix}$$

We can solve for the first column of  $[B] = [A]^{-1}$  by solving for

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix} \begin{bmatrix} b_{11} \\ b_{21} \\ b_{31} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

First solve

$$[L][Z] = [C],$$

that is

$$\begin{bmatrix} 1 & 0 & 0 \\ 2.56 & 1 & 0 \\ 5.76 & 3.5 & 1 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

to give

$$z_1 = 1$$

$$2.56z_1 + z_2 = 0$$

$$5.76z_1 + 3.5z_2 + z_3 = 0$$

Forward substitution starting from the first equation gives

$$z_1 = 1$$

$$z_2 = 0 - 2.56z_1$$

$$= 0 - 2.56(1)$$

$$= -2.56$$

$$z_3 = 0 - 5.76z_1 - 3.5z_2$$

$$= 0 - 5.76(1) - 3.5(-2.56)$$

$$= 3.2$$

Hence

$$\begin{bmatrix} Z \end{bmatrix} = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} 1 \\ -2.56 \\ 3.2 \end{bmatrix}$$

Now solve

$$[U][X] = [Z]$$

that is

$$\begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & 0 & 0.7 \end{bmatrix} \begin{bmatrix} b_{11} \\ b_{21} \\ b_{31} \end{bmatrix} = \begin{bmatrix} 1 \\ -2.56 \\ 3.2 \end{bmatrix}$$

$$25b_{11} + 5b_{21} + b_{31} = 1$$

$$-4.8b_{21} - 1.56b_{31} = -2.56$$

$$0.7b_{31} = 3.2$$

Backward substitution starting from the third equation gives

$$\begin{aligned} b_{31} &= \frac{3.2}{0.7} \\ &= 4.571 \\ b_{21} &= \frac{-2.56 + 1.56b_{31}}{-4.8} \\ &= \frac{-2.56 + 1.56(4.571)}{-4.8} \\ &= -0.9524 \\ b_{11} &= \frac{1 - 5b_{21} - b_{31}}{25} \\ &= \frac{1 - 5(-0.9524) - 4.571}{25} \\ &= 0.04762 \end{aligned}$$

Hence the first column of the inverse of  $[A]$  is

$$\begin{bmatrix} b_{11} \\ b_{21} \\ b_{31} \end{bmatrix} = \begin{bmatrix} 0.04762 \\ -0.9524 \\ 4.571 \end{bmatrix}$$

Similarly by solving

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix} \begin{bmatrix} b_{12} \\ b_{22} \\ b_{32} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \text{ gives } \begin{bmatrix} b_{12} \\ b_{22} \\ b_{32} \end{bmatrix} = \begin{bmatrix} -0.08333 \\ 1.417 \\ -5.000 \end{bmatrix}$$

and solving

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix} \begin{bmatrix} b_{13} \\ b_{23} \\ b_{33} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \text{ gives } \begin{bmatrix} b_{13} \\ b_{23} \\ b_{33} \end{bmatrix} = \begin{bmatrix} 0.03571 \\ -0.4643 \\ 1.429 \end{bmatrix}$$

Hence

$$[A]^{-1} = \begin{bmatrix} 0.04762 & -0.08333 & 0.03571 \\ -0.9524 & 1.417 & -0.4643 \\ 4.571 & -5.000 & 1.429 \end{bmatrix}$$

Can you confirm the following for the above example?

$$[A][A]^{-1} = [I] = [A]^{-1}[A]$$

### Key Terms:

*LU decomposition*  
*Inverse*

# **Chapter 04.08**

## **Gauss-Seidel Method**

*After reading this chapter, you should be able to:*

1. *solve a set of equations using the Gauss-Seidel method,*
2. *recognize the advantages and pitfalls of the Gauss-Seidel method, and*
3. *determine under what conditions the Gauss-Seidel method always converges.*

### **Why do we need another method to solve a set of simultaneous linear equations?**

In certain cases, such as when a system of equations is large, iterative methods of solving equations are more advantageous. Elimination methods, such as Gaussian elimination, are prone to large round-off errors for a large set of equations. Iterative methods, such as the Gauss-Seidel method, give the user control of the round-off error. Also, if the physics of the problem are well known, initial guesses needed in iterative methods can be made more judiciously leading to faster convergence.

What is the algorithm for the Gauss-Seidel method? Given a general set of  $n$  equations and  $n$  unknowns, we have

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \dots + a_{1n}x_n = c_1$$

$$a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \dots + a_{2n}x_n = c_2$$

.

.

$$a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \dots + a_{nn}x_n = c_n$$

If the diagonal elements are non-zero, each equation is rewritten for the corresponding unknown, that is, the first equation is rewritten with  $x_1$  on the left hand side, the second equation is rewritten with  $x_2$  on the left hand side and so on as follows

$$\begin{aligned}
 x_1 &= \frac{c_1 - a_{12}x_2 - a_{13}x_3 - \dots - a_{1n}x_n}{a_{11}} \\
 x_2 &= \frac{c_2 - a_{21}x_1 - a_{23}x_3 - \dots - a_{2n}x_n}{a_{22}} \\
 &\vdots \\
 &\vdots \\
 x_{n-1} &= \frac{c_{n-1} - a_{n-1,1}x_1 - a_{n-1,2}x_2 - \dots - a_{n-1,n-2}x_{n-2} - a_{n-1,n}x_n}{a_{n-1,n-1}} \\
 x_n &= \frac{c_n - a_{n1}x_1 - a_{n2}x_2 - \dots - a_{n,n-1}x_{n-1}}{a_{nn}}
 \end{aligned}$$

These equations can be rewritten in a summation form as

$$\begin{aligned}
 x_1 &= \frac{c_1 - \sum_{\substack{j=1 \\ j \neq 1}}^n a_{1j}x_j}{a_{11}} \\
 x_2 &= \frac{c_2 - \sum_{\substack{j=1 \\ j \neq 2}}^n a_{2j}x_j}{a_{22}} \\
 &\vdots \\
 &\vdots \\
 x_{n-1} &= \frac{c_{n-1} - \sum_{\substack{j=1 \\ j \neq n-1}}^n a_{n-1,j}x_j}{a_{n-1,n-1}} \\
 x_n &= \frac{c_n - \sum_{\substack{j=1 \\ j \neq n}}^n a_{nj}x_j}{a_{nn}}
 \end{aligned}$$

Hence for any row  $i$ ,

$$x_i = \frac{c_i - \sum_{\substack{j=1 \\ j \neq i}}^n a_{ij}x_j}{a_{ii}}, \quad i = 1, 2, \dots, n.$$

Now to find  $x_i$ 's, one assumes an initial guess for the  $x_i$ 's and then uses the rewritten equations to calculate the new estimates. Remember, one always uses the most recent estimates to calculate the next estimates,  $x_i$ . At the end of each iteration, one calculates the absolute relative approximate error for each  $x_i$  as

$$|e_a|_i = \left| \frac{x_i^{\text{new}} - x_i^{\text{old}}}{x_i^{\text{new}}} \right| \times 100$$

where  $x_i^{\text{new}}$  is the recently obtained value of  $x_i$ , and  $x_i^{\text{old}}$  is the previous value of  $x_i$ .

When the absolute relative approximate error for each  $x_i$  is less than the pre-specified tolerance, the iterations are stopped.

### Example 1

The upward velocity of a rocket is given at three different times in the following table

**Table 1** Velocity vs. time data.

Time, $t$ (s)	Velocity, $v$ (m/s)
5	106.8
8	177.2
12	279.2

The velocity data is approximated by a polynomial as

$$v(t) = a_1 t^2 + a_2 t + a_3, \quad 5 \leq t \leq 12$$

Find the values of  $a_1$ ,  $a_2$ , and  $a_3$  using the Gauss-Seidel method. Assume an initial guess of the solution as

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 5 \end{bmatrix}$$

and conduct two iterations.

### Solution

The polynomial is going through three data points  $(t_1, v_1)$ ,  $(t_2, v_2)$ , and  $(t_3, v_3)$  where from the above table

$$t_1 = 5, \quad v_1 = 106.8$$

$$t_2 = 8, \quad v_2 = 177.2$$

$$t_3 = 12, \quad v_3 = 279.2$$

Requiring that  $v(t) = a_1 t^2 + a_2 t + a_3$  passes through the three data points gives

$$v(t_1) = v_1 = a_1 t_1^2 + a_2 t_1 + a_3$$

$$v(t_2) = v_2 = a_1 t_2^2 + a_2 t_2 + a_3$$

$$v(t_3) = v_3 = a_1 t_3^2 + a_2 t_3 + a_3$$

Substituting the data  $(t_1, v_1)$ ,  $(t_2, v_2)$ , and  $(t_3, v_3)$  gives

$$a_1(5^2) + a_2(5) + a_3 = 106.8$$

$$a_1(8^2) + a_2(8) + a_3 = 177.2$$

$$a_1(12^2) + a_2(12) + a_3 = 279.2$$

or

$$25a_1 + 5a_2 + a_3 = 106.8$$

$$64a_1 + 8a_2 + a_3 = 177.2$$

$$144a_1 + 12a_2 + a_3 = 279.2$$

The coefficients  $a_1$ ,  $a_2$ , and  $a_3$  for the above expression are given by

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 106.8 \\ 177.2 \\ 279.2 \end{bmatrix}$$

Rewriting the equations gives

$$a_1 = \frac{106.8 - 5a_2 - a_3}{25}$$

$$a_2 = \frac{177.2 - 64a_1 - a_3}{8}$$

$$a_3 = \frac{279.2 - 144a_1 - 12a_2}{1}$$

### Iteration #1

Given the initial guess of the solution vector as

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 5 \end{bmatrix}$$

we get

$$a_1 = \frac{106.8 - 5(2) - (5)}{25}$$

$$= 3.6720$$

$$a_2 = \frac{177.2 - 64(3.6720) - (5)}{8}$$

$$= -7.8150$$

$$a_3 = \frac{279.2 - 144(3.6720) - 12(-7.8150)}{1}$$

$$= -155.36$$

The absolute relative approximate error for each  $x_i$  then is

$$|\epsilon_a|_1 = \left| \frac{3.6720 - 1}{3.6720} \right| \times 100$$

$$= 72.76\%$$

$$|\epsilon_a|_2 = \left| \frac{-7.8150 - 2}{-7.8150} \right| \times 100$$

$$= 125.47\%$$

$$|\epsilon_a|_3 = \left| \frac{-155.36 - 5}{-155.36} \right| \times 100$$

$$= 103.22\%$$

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

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 3.6720 \\ -7.8510 \\ -155.36 \end{bmatrix}$$

and the maximum absolute relative approximate error is 125.47%.

### Iteration #2

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

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 3.6720 \\ -7.8510 \\ -155.36 \end{bmatrix}$$

Now we get

$$\begin{aligned} a_1 &= \frac{106.8 - 5(-7.8510) - (-155.36)}{25} \\ &= 12.056 \\ a_2 &= \frac{177.2 - 64(12.056) - (-155.36)}{8} \\ &= -54.882 \\ a_3 &= \frac{279.2 - 144(12.056) - 12(-54.882)}{1} \\ &= -798.34 \end{aligned}$$

The absolute relative approximate error for each  $x_i$  then is

$$\begin{aligned} |e_a|_1 &= \left| \frac{12.056 - 3.6720}{12.056} \right| \times 100 \\ &= 69.543\% \\ |e_a|_2 &= \left| \frac{-54.882 - (-7.8510)}{-54.882} \right| \times 100 \\ &= 85.695\% \\ |e_a|_3 &= \left| \frac{-798.34 - (-155.36)}{-798.34} \right| \times 100 \\ &= 80.540\% \end{aligned}$$

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

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 12.056 \\ -54.882 \\ -798.34 \end{bmatrix}$$

and the maximum absolute relative approximate error is 85.695%.

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

Iteration	$a_1$	$ e_a _1 \%$	$a_2$	$ e_a _2 \%$	$a_3$	$ e_a _3 \%$
1	3.6720	72.767	-7.8510	125.47	-155.36	103.22
2	12.056	69.543	-54.882	85.695	-798.34	80.540
3	47.182	74.447	-255.51	78.521	-3448.9	76.852
4	193.33	75.595	-1093.4	76.632	-14440	76.116
5	800.53	75.850	-4577.2	76.112	-60072	75.963
6	3322.6	75.906	-19049	75.972	-249580	75.931

As seen in the above table, the solution estimates are not converging to the true solution of

$$a_1 = 0.29048$$

$$a_2 = 19.690$$

$$a_3 = 1.0857$$

### The above system of equations does not seem to converge. Why?

Well, a pitfall of most iterative methods is that they may or may not converge. However, the solution to a certain classes of systems of simultaneous equations does always converge using the Gauss-Seidel method. This class of system of equations is where the coefficient matrix  $[A]$  in  $[A][X]=[C]$  is diagonally dominant, that is

$$|a_{ii}| \geq \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}| \text{ for all } i$$

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}| \text{ for at least one } i$$

If a system of equations has a coefficient matrix that is not diagonally dominant, it may or may not converge. Fortunately, many physical systems that result in simultaneous linear equations have a diagonally dominant coefficient matrix, which then assures convergence for iterative methods such as the Gauss-Seidel method of solving simultaneous linear equations.

### Example 2

Find the solution to the following system of equations using the Gauss-Seidel method.

$$12x_1 + 3x_2 - 5x_3 = 1$$

$$x_1 + 5x_2 + 3x_3 = 28$$

$$3x_1 + 7x_2 + 13x_3 = 76$$

Use

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

as the initial guess and conduct two iterations.

### Solution

The coefficient matrix

$$[A] = \begin{bmatrix} 12 & 3 & -5 \\ 1 & 5 & 3 \\ 3 & 7 & 13 \end{bmatrix}$$

is diagonally dominant as

$$|a_{11}| = |12| = 12 \geq |a_{12}| + |a_{13}| = |3| + |-5| = 8$$

$$|a_{22}| = |5| = 5 \geq |a_{21}| + |a_{23}| = |1| + |3| = 4$$

$$|a_{33}| = |13| = 13 \geq |a_{31}| + |a_{32}| = |3| + |7| = 10$$

and the inequality is strictly greater than for at least one row. Hence, the solution should converge using the Gauss-Seidel method.

Rewriting the equations, we get

$$x_1 = \frac{1 - 3x_2 + 5x_3}{12}$$

$$x_2 = \frac{28 - x_1 - 3x_3}{5}$$

$$x_3 = \frac{76 - 3x_1 - 7x_2}{13}$$

Assuming an initial guess of

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

Iteration #1

$$x_1 = \frac{1 - 3(0) + 5(1)}{12}$$

$$= 0.50000$$

$$x_2 = \frac{28 - (0.50000) - 3(1)}{5}$$

$$= 4.9000$$

$$x_3 = \frac{76 - 3(0.50000) - 7(4.9000)}{13}$$

$$= 3.0923$$

The absolute relative approximate error at the end of the first iteration is

$$|\epsilon_a|_1 = \left| \frac{0.50000 - 1}{0.50000} \right| \times 100$$

$$= 100.00\%$$

$$|\epsilon_a|_2 = \left| \frac{4.9000 - 0}{4.9000} \right| \times 100$$

$$= 100.00\%$$

$$|\epsilon_a|_3 = \left| \frac{3.0923 - 1}{3.0923} \right| \times 100$$

$$= 67.662\%$$

The maximum absolute relative approximate error is 100.00%

Iteration #2

$$\begin{aligned}x_1 &= \frac{1 - 3(4.9000) + 5(3.0923)}{12} \\&= 0.14679 \\x_2 &= \frac{28 - (0.14679) - 3(3.0923)}{5} \\&= 3.7153 \\x_3 &= \frac{76 - 3(0.14679) - 7(3.7153)}{13} \\&= 3.8118\end{aligned}$$

At the end of second iteration, the absolute relative approximate error is

$$\begin{aligned}|\epsilon_a|_1 &= \left| \frac{0.14679 - 0.50000}{0.14679} \right| \times 100 \\&= 240.61\% \\|\epsilon_a|_2 &= \left| \frac{3.7153 - 4.9000}{3.7153} \right| \times 100 \\&= 31.889\% \\|\epsilon_a|_3 &= \left| \frac{3.8118 - 3.0923}{3.8118} \right| \times 100 \\&= 18.874\%\end{aligned}$$

The maximum absolute relative approximate error is 240.61%. This is greater than the value of 100.00% we obtained in the first iteration. Is the solution diverging? No, as you conduct more iterations, the solution converges as follows.

Iteration	$x_1$	$ \epsilon_a _1$ %	$x_2$	$ \epsilon_a _2$ %	$x_3$	$ \epsilon_a _3$ %
1	0.50000	100.00	4.9000	100.00	3.0923	67.662
2	0.14679	240.61	3.7153	31.889	3.8118	18.874
3	0.74275	80.236	3.1644	17.408	3.9708	4.0064
4	0.94675	21.546	3.0281	4.4996	3.9971	0.65772
5	0.99177	4.5391	3.0034	0.82499	4.0001	0.074383
6	0.99919	0.74307	3.0001	0.10856	4.0001	0.00101

This is close to the exact solution vector of

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \\ 4 \end{bmatrix}$$

**Example 3**

Given the system of equations

$$3x_1 + 7x_2 + 13x_3 = 76$$

$$x_1 + 5x_2 + 3x_3 = 28$$

$$12x_1 + 3x_2 - 5x_3 = 1$$

find the solution using the Gauss-Seidel method. Use

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

as the initial guess.

### Solution

Rewriting the equations, we get

$$x_1 = \frac{76 - 7x_2 - 13x_3}{3}$$

$$x_2 = \frac{28 - x_1 - 3x_3}{5}$$

$$x_3 = \frac{1 - 12x_1 - 3x_2}{-5}$$

Assuming an initial guess of

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

the next six iterative values are given in the table below.

Iteration	$x_1$	$ e_a _1\%$	$x_2$	$ e_a _2\%$	$x_3$	$ e_a _3\%$
1	21.000	95.238	0.80000	100.00	50.680	98.027
2	-196.15	110.71	14.421	94.453	-462.30	110.96
3	1995.0	109.83	-116.02	112.43	4718.1	109.80
4	-20149	109.90	1204.6	109.63	-47636	109.90
5	$2.0364 \times 10^5$	109.89	-12140	109.92	$4.8144 \times 10^5$	109.89
6	$-2.0579 \times 10^6$	109.89	$1.2272 \times 10^5$	109.89	$-4.8653 \times 10^6$	109.89

You can see that this solution is not converging and the coefficient matrix is not diagonally dominant. The coefficient matrix

$$[A] = \begin{bmatrix} 3 & 7 & 13 \\ 1 & 5 & 3 \\ 12 & 3 & -5 \end{bmatrix}$$

is not diagonally dominant as

$$|a_{11}| = |3| = 3 \leq |a_{12}| + |a_{13}| = |7| + |13| = 20$$

Hence, the Gauss-Seidel method may or may not converge.

However, it is the same set of equations as the previous example and that converged. The only difference is that we exchanged first and the third equation with each other and that made the coefficient matrix not diagonally dominant.

Therefore, it is possible that a system of equations can be made diagonally dominant if one exchanges the equations with each other. However, it is not possible for all cases. For example, the following set of equations

$$\begin{aligned}x_1 + x_2 + x_3 &= 3 \\2x_1 + 3x_2 + 4x_3 &= 9 \\x_1 + 7x_2 + x_3 &= 9\end{aligned}$$

cannot be rewritten to make the coefficient matrix diagonally dominant.

**Key Terms:**

*Gauss-Seidel method*

*Convergence of Gauss-Seidel method*

*Diagonally dominant matrix*

## Chapter 04.09

# Adequacy of Solutions

*After reading this chapter, you should be able to:*

1. know the difference between ill-conditioned and well-conditioned systems of equations,
2. define and find the norm of a matrix
3. define and evaluate the condition number of an invertible square matrix
4. relate the condition number of a coefficient matrix to the ill or well conditioning of the system of simultaneous linear equations, that is, how much can you trust the solution of the simultaneous linear equations.

### What do you mean by ill-conditioned and well-conditioned system of equations?

A system of equations is considered to be **well-conditioned** if a small change in the coefficient matrix or a small change in the right hand side results in a small change in the solution vector.

A system of equations is considered to be **ill-conditioned** if a small change in the coefficient matrix or a small change in the right hand side results in a large change in the solution vector.

### Example 1

Is this system of equations well-conditioned?

$$\begin{bmatrix} 1 & 2 \\ 2 & 3.999 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 7.999 \end{bmatrix}$$

### Solution

The solution to the above set of equations is

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

Make a small change in the right hand side vector of the equations

$$\begin{bmatrix} 1 & 2 \\ 2 & 3.999 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4.001 \\ 7.998 \end{bmatrix}$$

gives

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -3.999 \\ 4.000 \end{bmatrix}$$

Make a small change in the coefficient matrix of the equations

$$\begin{bmatrix} 1.001 & 2.001 \\ 2.001 & 3.998 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 7.999 \end{bmatrix}$$

gives

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 3.994 \\ 0.001388 \end{bmatrix}$$

This last systems of equation “looks” ill-conditioned because a small change in the coefficient matrix or the right hand side resulted in a large change in the solution vector.

### Example 2

Is this system of equations well-conditioned?

$$\begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 7 \end{bmatrix}$$

#### Solution

The solution to the above equations is

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

Make a small change in the right hand side vector of the equations.

$$\begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4.001 \\ 7.001 \end{bmatrix}$$

gives

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1.999 \\ 1.001 \end{bmatrix}$$

Make a small change in the coefficient matrix of the equations.

$$\begin{bmatrix} 1.001 & 2.001 \\ 2.001 & 3.001 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 7 \end{bmatrix}$$

gives

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2.003 \\ 0.997 \end{bmatrix}$$

This system of equation “looks” well conditioned because small changes in the coefficient matrix or the right hand side resulted in small changes in the solution vector.

### So what if the system of equations is ill conditioned or well conditioned?

Well, if a system of equations is ill-conditioned, we cannot trust the solution as much. Revisit the velocity problem, Example 5.1 in Chapter 5. The values in the coefficient matrix  $[A]$  are squares of time, etc. For example, if instead of  $a_{11} = 25$ , you used  $a_{11} = 24.99$ , would you want this small change to make a huge difference in the solution vector. If it did, would you trust the solution?

Later we will see how much (quantifiable terms) we can trust the solution in a system of equations. Every invertible square matrix has a **condition number** and coupled with the **machine epsilon**, we can quantify how many significant digits one can trust in the solution.

### To calculate the condition number of an invertible square matrix, I need to know what the norm of a matrix means. How is the norm of a matrix defined?

Just like the determinant, the norm of a matrix is a simple unique scalar number. However, the norm is always positive and is defined for all matrices – square or rectangular, and invertible or noninvertible square matrices.

One of the popular definitions of a norm is the row sum norm (also called the uniform-matrix norm). For a  $m \times n$  matrix  $[A]$ , the row sum norm of  $[A]$  is defined as

$$\|A\|_\infty = \max_{1 \leq i \leq m} \sum_{j=1}^n |a_{ij}|$$

that is, find the sum of the absolute value of the elements of each row of the matrix  $[A]$ . The maximum out of the  $m$  such values is the row sum norm of the matrix  $[A]$ .

### Example 3

Find the row sum norm of the following matrix  $[A]$ .

$$A = \begin{bmatrix} 10 & -7 & 0 \\ -3 & 2.099 & 6 \\ 5 & -1 & 5 \end{bmatrix}$$

**Solution**

$$\begin{aligned}\|A\|_{\infty} &= \max_{1 \leq i \leq 3} \sum_{j=1}^3 |a_{ij}| \\ &= \max[(|10| + |-7| + |0|), (|-3| + |2.099| + |6|), (|5| + |-1| + |5|)] \\ &= \max[(10 + 7 + 0), (3 + 2.099 + 6), (5 + 1 + 5)] \\ &= \max[17, 11.099, 11] \\ &= 17.\end{aligned}$$

**How is the norm related to the conditioning of the matrix?**

Let us start answering this question using an example. Go back to the *ill-conditioned* system of equations,

$$\begin{bmatrix} 1 & 2 \\ 2 & 3.999 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 7.999 \end{bmatrix}$$

that gives the solution as

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

Denoting the above set of equations as

$$[A][X] = [C]$$

$$\|X\|_{\infty} = 2$$

$$\|C\|_{\infty} = 7.999$$

Making a small change in the right hand side,

$$\begin{bmatrix} 1 & 2 \\ 2 & 3.999 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4.001 \\ 7.998 \end{bmatrix}$$

gives

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -3.999 \\ 4.000 \end{bmatrix}$$

Denoting the above set of equations by

$$[A][X'] = [C']$$

right hand side vector is found by

$$[\Delta C] = [C'] - [C]$$

and the change in the solution vector is found by

$$[\Delta X] = [X'] - [X]$$

then

$$\begin{aligned} [\Delta C] &= \begin{bmatrix} 4.001 \\ 7.998 \end{bmatrix} - \begin{bmatrix} 4 \\ 7.999 \end{bmatrix} \\ &= \begin{bmatrix} 0.001 \\ -0.001 \end{bmatrix} \end{aligned}$$

and

$$\begin{aligned} [\Delta X] &= \begin{bmatrix} -3.999 \\ 4.000 \end{bmatrix} - \begin{bmatrix} 2 \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} -5.999 \\ 3.000 \end{bmatrix} \end{aligned}$$

then

$$\|\Delta C\|_\infty = 0.001$$

$$\|\Delta X\|_\infty = 5.999$$

The relative change in the norm of the solution vector is

$$\begin{aligned} \frac{\|\Delta X\|_\infty}{\|X\|_\infty} &= \frac{5.999}{2} \\ &= 2.9995 \end{aligned}$$

The relative change in the norm of the right hand side vector is

$$\begin{aligned} \frac{\|\Delta C\|_\infty}{\|C\|_\infty} &= \frac{0.001}{7.999} \\ &= 1.250 \times 10^{-4} \end{aligned}$$

See the small relative change of  $1.250 \times 10^{-4}$  in the right hand side vector norm results in a large relative change in the solution vector norm of 2.9995.

In fact, the ratio between the relative change in the norm of the solution vector and the relative change in the norm of the right hand side vector is

$$\begin{aligned} \frac{\|\Delta X\|_\infty / \|X\|_\infty}{\|\Delta C\|_\infty / \|C\|_\infty} &= \frac{2.9995}{1.250 \times 10^{-4}} \\ &= 23993 \end{aligned}$$

Let us now go back to the *well-conditioned* system of equations.

$$\begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 7 \end{bmatrix}$$

gives

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

Denoting the system of equations by

$$[A][X] = [C]$$

$$\|X\|_{\infty} = 2$$

$$\|C\|_{\infty} = 7$$

Making a small change in the right hand side vector

$$\begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4.001 \\ 7.001 \end{bmatrix}$$

gives

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1.999 \\ 1.001 \end{bmatrix}$$

Denoting the above set of equations by

$$[A][X'] = [C']$$

the change in the right hand side vector is then found by

$$[\Delta C] = [C'] - [C]$$

and the change in the solution vector is

$$[\Delta X] = [X'] - [X]$$

then

$$\begin{aligned} [\Delta C] &= \begin{bmatrix} 4.001 \\ 7.001 \end{bmatrix} - \begin{bmatrix} 4 \\ 7 \end{bmatrix} \\ &= \begin{bmatrix} 0.001 \\ 0.001 \end{bmatrix} \end{aligned}$$

and

$$\begin{aligned} [\Delta X] &= \begin{bmatrix} 1.999 \\ 1.001 \end{bmatrix} - \begin{bmatrix} 2 \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} -0.001 \\ 0.001 \end{bmatrix} \end{aligned}$$

then

$$\|\Delta C\|_{\infty} = 0.001$$

$$\|\Delta X\|_{\infty} = 0.001$$

The relative change in the norm of solution vector is

$$\frac{\|\Delta X\|_{\infty}}{\|X\|_{\infty}} = \frac{0.001}{2}$$

$$= 5 \times 10^{-4}$$

The relative change in the norm of the right hand side vector is

$$\frac{\|\Delta C\|_{\infty}}{\|C\|_{\infty}} = \frac{0.001}{7}$$

$$= 1.429 \times 10^{-4}$$

See the small relative change in the right hand side vector norm of  $1.429 \times 10^{-4}$  results in the small relative change in the solution vector norm of  $5 \times 10^{-4}$ .

In fact, the ratio between the relative change in the norm of the solution vector and the relative change in the norm of the right hand side vector is

$$\begin{aligned} \frac{\|\Delta X\|_{\infty}/\|X\|_{\infty}}{\|\Delta C\|_{\infty}/\|C\|_{\infty}} &= \frac{5 \times 10^{-4}}{1.429 \times 10^{-4}} \\ &= 3.5 \end{aligned}$$

### What are some of the properties of norms?

1. For a matrix  $[A]$ ,  $\|A\| \geq 0$
2. For a matrix  $[A]$  and a scalar  $k$ ,  $\|kA\| = |k|\|A\|$
3. For two matrices  $[A]$  and  $[B]$  of same order,  $\|A + B\| \leq \|A\| + \|B\|$
4. For two matrices  $[A]$  and  $[B]$  that can be multiplied as  $[A][B]$ ,  $\|AB\| \leq \|A\|\|B\|$

**Is there a general relationship that exists between  $\|\Delta X\|/\|X\|$  and  $\|\Delta C\|/\|C\|$  or between  $\|\Delta X\|/\|X\|$  and  $\|\Delta A\|/\|A\|$ ? If so, it could help us identify well-conditioned and ill conditioned system of equations.**

If there is such a relationship, will it help us quantify the conditioning of the matrix? That is, will it tell us how many significant digits we could trust in the solution of a system of simultaneous linear equations?

There is a relationship that exists between

$$\frac{\|\Delta X\|}{\|X\|} \text{ and } \frac{\|\Delta C\|}{\|C\|}$$

and between

$$\frac{\|\Delta X\|}{\|X\|} \text{ and } \frac{\|\Delta A\|}{\|A\|}$$

These relationships are

$$\frac{\|\Delta X\|}{\|X\|} \leq \|A\| \|A^{-1}\| \frac{\|\Delta C\|}{\|C\|}$$

and

$$\frac{\|\Delta X\|}{\|X + \Delta X\|} \leq \|A\| \|A^{-1}\| \frac{\|\Delta A\|}{\|A\|}$$

The above two inequalities show that the relative change in the norm of the right hand side vector or the coefficient matrix can be amplified by as much as  $\|A\| \|A^{-1}\|$ .

This number  $\|A\| \|A^{-1}\|$  is called the **condition number** of the matrix and coupled with the machine epsilon, we can quantify the accuracy of the solution of  $[A][X] = [C]$ .

### Prove for

$$[A][X] = [C]$$

that

$$\frac{\|\Delta X\|}{\|X + \Delta X\|} \leq \|A\| \|A^{-1}\| \frac{\|\Delta A\|}{\|A\|}.$$

### Proof

Let

$$[A][X] = [C] \quad (1)$$

Then if  $[A]$  is changed to  $[A']$ , the  $[X]$  will change to  $[X']$ , such that

$$[A'][X'] = [C] \quad (2)$$

From Equations (1) and (2),

$$[A][X] = [A'][X']$$

Denoting change in  $[A]$  and  $[X]$  matrices as  $[\Delta A]$  and  $[\Delta X]$ , respectively

$$[\Delta A] = [A'] - [A]$$

$$[\Delta X] = [X'] - [X]$$

then

$$[A][X] = ([A] + [\Delta A])([X] + [\Delta X])$$

Expanding the above expression

$$[A][X] = [A][X] + [A][\Delta X] + [\Delta A][X] + [\Delta A][\Delta X]$$

$$[0] = [A][\Delta X] + [\Delta A](X + \Delta X)$$

$$-[A][\Delta X] = [\Delta A](X + \Delta X)$$

$$[\Delta X] = -[A]^{-1}[\Delta A](X + \Delta X)$$

Applying the theorem of norms, that the norm of multiplied matrices is less than the multiplication of the individual norms of the matrices,

$$\|\Delta X\| \leq \|A^{-1}\| \|\Delta A\| \|X + \Delta X\|$$

Multiplying both sides by  $\|A\|$

$$\|A\| \|\Delta X\| \leq \|A\| \|A^{-1}\| \|\Delta A\| \|X + \Delta X\|$$

$$\frac{\|\Delta X\|}{\|X + \Delta X\|} \leq \|A\| \|A^{-1}\| \frac{\|\Delta A\|}{\|A\|}$$

### How do I use the above theorems to find how many significant digits are correct in my solution vector?

The relative error in a solution vector norm is  $\leq \text{Cond}(A) \times$  relative error in right hand side vector norm.

The possible relative error in the solution vector norm is  $\leq \text{Cond}(A) \times \epsilon_{mach}$

Hence  $\text{Cond}(A) \times \epsilon_{mach}$  should give us the number of significant digits,  $m$  that are at least correct in our solution by finding out the largest value of  $m$  for which  $\text{Cond}(A) \times \epsilon_{mach}$  is less than  $0.5 \times 10^{-m}$ .

### Example 4

How many significant digits can I trust in the solution of the following system of equations?

$$\begin{bmatrix} 1 & 2 \\ 2 & 3.999 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$$

**Solution**

For

$$[A] = \begin{bmatrix} 1 & 2 \\ 2 & 3.999 \end{bmatrix}$$

it can be shown

$$[A]^{-1} = \begin{bmatrix} -3999 & 2000 \\ 2000 & -1000 \end{bmatrix}$$

$$\|A\|_{\infty} = 5.999$$

$$\|A^{-1}\|_{\infty} = 5999$$

$$\begin{aligned} \text{Cond}(A) &= \|A\| \|A^{-1}\| \\ &= 5.999 \times 5999.4 \\ &= 35990 \end{aligned}$$

Assuming single precision with 23 bits used in the mantissa for real numbers, the machine epsilon is

$$\begin{aligned} \epsilon_{mach} &= 2^{-23} \\ &= 0.119209 \times 10^{-6} \end{aligned}$$

$$\begin{aligned} \text{Cond}(A) \times \epsilon_{mach} &= 35990 \times 0.119209 \times 10^{-6} \\ &= 0.4290 \times 10^{-2} \end{aligned}$$

For what maximum positive value of  $m$ , would  $\text{Cond}(A) \times \epsilon_{mach}$  be less than or equal to  $0.5 \times 10^{-m}$

$$0.4290 \times 10^{-2} \leq 0.5 \times 10^{-m}$$

$$0.8580 \times 10^{-2} \leq 10^{-m}$$

$$\log(0.8580 \times 10^{-2}) \leq \log(10^{-m})$$

$$-2.067 \leq -m$$

$$m \leq 2.067$$

$$m \leq 2$$

So two significant digits are at least correct in the solution vector.

**Example 5**

How many significant digits can I trust in the solution of the following system of equations?

$$\begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 7 \end{bmatrix}$$

**Solution**

For

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

it can be shown

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

Then

$$\|A\|_{\infty} = 5,$$

$$\|A^{-1}\|_{\infty} = 5.$$

$$\begin{aligned} \text{Cond}(A) &= \|A\| \|A^{-1}\| \\ &= 5 \times 5 \\ &= 25 \end{aligned}$$

Assuming single precision with 23 bits used in the mantissa for real numbers, the machine epsilon

$$\begin{aligned} \epsilon_{mach} &= 2^{-23} \\ &= 0.119209 \times 10^{-6} \end{aligned}$$

$$\begin{aligned} \text{Cond}(A) \times \epsilon_{mach} &= 25 \times 0.119209 \times 10^{-6} \\ &= 0.2980 \times 10^{-5} \end{aligned}$$

For what maximum positive value of  $m$ , would  $\text{Cond}(A) \times \epsilon_{mach}$  be less than or equal to  $0.5 \times 10^{-m}$

$$0.2980 \times 10^{-5} \leq 0.5 \times 10^{-m}$$

$$m \leq 5$$

So five significant digits are at least correct in the solution vector.

**Key Terms:**

*Ill-Conditioned matrix*

*Well-Conditioned matrix*

*Norm*

*Condition Number*

*Machine Epsilon*

*Significant Digits*

# Chapter 04.10

## Eigenvalues and Eigenvectors

After reading this chapter, you should be able to:

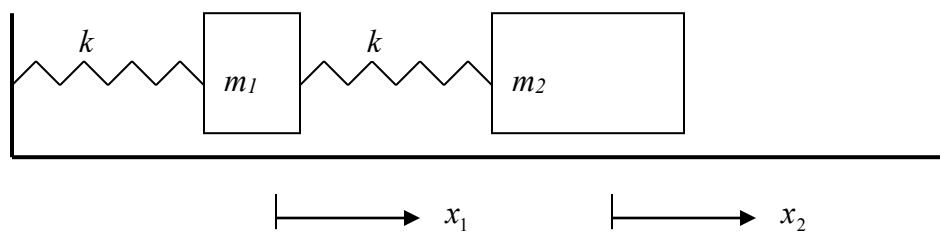
1. define eigenvalues and eigenvectors of a square matrix,
2. find eigenvalues and eigenvectors of a square matrix,
3. relate eigenvalues to the singularity of a square matrix, and
4. use the power method to numerically find the largest eigenvalue in magnitude of a square matrix and the corresponding eigenvector.

### What does eigenvalue mean?

The word eigenvalue comes from the German word *Eigenwert* where Eigen means *characteristic* and Wert means *value*. However, what the word means is not on your mind! You want to know why I need to learn about eigenvalues and eigenvectors. Once I give you an example of an application of eigenvalues and eigenvectors, you will want to know how to find these eigenvalues and eigenvectors.

### Can you give me a physical example application of eigenvalues and eigenvectors?

Look at the spring-mass system as shown in the picture below.



Assume each of the two mass-displacements to be denoted by  $x_1$  and  $x_2$ , and let us assume each spring has the same spring constant  $k$ . Then by applying Newton's 2<sup>nd</sup> and 3<sup>rd</sup> law of motion to develop a force-balance for each mass we have

$$m_1 \frac{d^2x_1}{dt^2} = -kx_1 + k(x_2 - x_1)$$

$$m_2 \frac{d^2x_2}{dt^2} = -k(x_2 - x_1)$$

Rewriting the equations, we have

$$m_1 \frac{d^2x_1}{dt^2} - k(-2x_1 + x_2) = 0$$

$$m_2 \frac{d^2x_2}{dt^2} - k(x_1 - x_2) = 0$$

Let  $m_1 = 10, m_2 = 20, k = 15$

$$10 \frac{d^2x_1}{dt^2} - 15(-2x_1 + x_2) = 0$$

$$20 \frac{d^2x_2}{dt^2} - 15(x_1 - x_2) = 0$$

From vibration theory, the solutions can be of the form

$$x_i = A_i \sin(\omega t - \theta)$$

where

$A_i$  = amplitude of the vibration of mass  $i$ ,

$\omega$  = frequency of vibration,

$\theta$  = phase shift.

then

$$\frac{d^2x_i}{dt^2} = -A_i \omega^2 \sin(\omega t - \theta)$$

Substituting  $x_i$  and  $\frac{d^2x_i}{dt^2}$  in equations,

$$-10A_1 \omega^2 - 15(-2A_1 + A_2) = 0$$

$$-20A_2 \omega^2 - 15(A_1 - A_2) = 0$$

gives

$$(-10\omega^2 + 30)A_1 - 15A_2 = 0$$

$$-15A_1 + (-20\omega^2 + 15)A_2 = 0$$

or

$$(-\omega^2 + 3)A_1 - 1.5A_2 = 0$$

$$-0.75A_1 + (-\omega^2 + 0.75)A_2 = 0$$

In matrix form, these equations can be rewritten as

$$\begin{bmatrix} -\omega^2 + 3 & -1.5 \\ -0.75 & -\omega^2 + 0.75 \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 3 & -1.5 \\ -0.75 & 0.75 \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} - \omega^2 \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Let  $\omega^2 = \lambda$

$$[A] = \begin{bmatrix} 3 & -1.5 \\ -0.75 & 0.75 \end{bmatrix}$$

$$[X] = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}$$

$$[A][X] - \lambda[X] = 0$$

$$[A][X] = \lambda[X]$$

In the above equation,  $\lambda$  is the eigenvalue and  $[X]$  is the eigenvector corresponding to  $\lambda$ . As you can see, if we know  $\lambda$  for the above example we can calculate the natural frequency of the vibration

$$\omega = \sqrt{\lambda}$$

Why are the natural frequencies of vibration important? Because you do not want to have a forcing force on the spring-mass system close to this frequency as it would make the amplitude  $A_i$  very large and make the system unstable.

### What is the general definition of eigenvalues and eigenvectors of a square matrix?

If  $[A]$  is a  $n \times n$  matrix, then  $[X] \neq \vec{0}$  is an eigenvector of  $[A]$  if

$$[A][X] = \lambda[X]$$

where  $\lambda$  is a scalar and  $[X] \neq 0$ . The scalar  $\lambda$  is called the eigenvalue of  $[A]$  and  $[X]$  is called the eigenvector corresponding to the eigenvalue  $\lambda$ .

### How do I find eigenvalues of a square matrix?

To find the eigenvalues of a  $n \times n$  matrix  $[A]$ , we have

$$[A][X] = \lambda[X]$$

$$[A][X] - \lambda[X] = 0$$

$$[A][X] - \lambda[I][X] = 0$$

$$([A] - [\lambda][I])[X] = 0$$

Now for the above set of equations to have a nonzero solution,

$$\det([A] - \lambda[I]) = 0$$

This left hand side can be expanded to give a polynomial in  $\lambda$  and solving the above equation would give us values of the eigenvalues. The above equation is called the characteristic equation of  $[A]$ .

For a  $[A]$   $n \times n$  matrix, the characteristic polynomial of  $A$  is of degree  $n$  as follows

$$\det([A] - \lambda[I]) = 0$$

giving

$$\lambda^n + c_1\lambda^{n-1} + c_2\lambda^{n-2} + \dots + c_n = 0$$

Hence, this polynomial has  $n$  roots.

### Example 1

Find the eigenvalues of the physical problem discussed in the beginning of this chapter, that is, find the eigenvalues of the matrix

$$[A] = \begin{bmatrix} 3 & -1.5 \\ -0.75 & 0.75 \end{bmatrix}$$

**Solution**

$$\begin{aligned}
 [A] - \lambda[I] &= \begin{bmatrix} 3-\lambda & -1.5 \\ -0.75 & 0.75-\lambda \end{bmatrix} \\
 \det([A] - \lambda[I]) &= (3-\lambda)(0.75-\lambda) - (-0.75)(-1.5) = 0 \\
 2.25 - 0.75\lambda - 3\lambda + \lambda^2 - 1.125 &= 0 \\
 \lambda^2 - 3.75\lambda + 1.125 &= 0 \\
 \lambda &= \frac{-(-3.75) \pm \sqrt{(-3.75)^2 - 4(1)(1.125)}}{2(1)} \\
 &= \frac{3.75 \pm 3.092}{2} \\
 &= 3.421, 0.3288
 \end{aligned}$$

So the eigenvalues are 3.421 and 0.3288.

**Example 2**

Find the eigenvectors of

$$A = \begin{bmatrix} 3 & -1.5 \\ -0.75 & 0.75 \end{bmatrix}$$

Solution

The eigenvalues have already been found in Example 1 as

$$\lambda_1 = 3.421, \lambda_2 = 0.3288$$

Let

$$[X] = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

be the eigenvector corresponding to

$$\lambda_1 = 3.421$$

Hence

$$\begin{aligned}
 ([A] - \lambda_1[I])[X] &= 0 \\
 \left\{ \begin{bmatrix} 3 & -1.5 \\ -0.75 & 0.75 \end{bmatrix} - 3.421 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right\} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} &= 0 \\
 \begin{bmatrix} -0.421 & -1.5 \\ -0.75 & -2.671 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} &= \begin{bmatrix} 0 \\ 0 \end{bmatrix}
 \end{aligned}$$

If

$$x_1 = s$$

then

$$-0.421s - 1.5x_2 = 0$$

$$x_2 = -0.2808s$$

The eigenvector corresponding to  $\lambda_1 = 3.421$  then is

$$[X] = \begin{bmatrix} s \\ -0.2808s \end{bmatrix} \\ = s \begin{bmatrix} 1 \\ -0.2808 \end{bmatrix}$$

The eigenvector corresponding to

$$\lambda_1 = 3.421$$

is

$$\begin{bmatrix} 1 \\ -0.2808 \end{bmatrix}$$

Similarly, the eigenvector corresponding to

$$\lambda_2 = 0.3288$$

is

$$\begin{bmatrix} 1 \\ 1.781 \end{bmatrix}$$

### Example 3

Find the eigenvalues and eigenvectors of

$$[A] = \begin{bmatrix} 1.5 & 0 & 1 \\ -0.5 & 0.5 & -0.5 \\ -0.5 & 0 & 0 \end{bmatrix}$$

### Solution

The characteristic equation is given by

$$\det([A] - \lambda[I]) = 0$$

$$\det \begin{bmatrix} 1.5 - \lambda & 0 & 1 \\ -0.5 & 0.5 - \lambda & -0.5 \\ -0.5 & 0 & -\lambda \end{bmatrix} = 0$$

$$(1.5 - \lambda)[(0.5 - \lambda)(-\lambda) - (-0.5)(0)] + (1)[(-0.5)(0) - (-0.5)(0.5 - \lambda)] = 0$$

$$-\lambda^3 + 2\lambda^2 - 1.25\lambda + 0.25 = 0$$

To find the roots of the characteristic polynomial equation

$$-\lambda^3 + 2\lambda^2 - 1.25\lambda + 0.25 = 0$$

we find that the first root by observation is

$$\lambda = 1$$

as substitution of  $\lambda = 1$  gives

$$(-1)^3 + 2(1)^2 - 1.25(1) + 0.25 = 0$$

$$0 = 0$$

So

$$(\lambda - 1)$$

is a factor of

$$-\lambda^3 + 2\lambda^2 - 1.25\lambda + 0.25.$$

To find the other factors of the characteristic polynomial, we first conduct long division

$$\begin{array}{r} -\lambda^2 + \lambda + 0.25 \\ \lambda - 1 \overline{) -\lambda^3 + 2\lambda^2 - 1.25\lambda + 0.25} \\ -\lambda^3 + \lambda^2 \\ \hline \lambda^2 - 1.25\lambda + 0.25 \\ \lambda^2 - \lambda \\ \hline -0.25\lambda + 0.25 \\ -0.25\lambda + 0.25 \\ \hline \end{array}$$

Hence

$$-\lambda^3 + 2\lambda^2 - 1.25\lambda + 0.25 = (\lambda - 1)(-\lambda^2 + \lambda + 0.25)$$

To find zeroes of  $-\lambda^2 + \lambda + 0.25$ , we solve the quadratic equation,

$$-\lambda^2 + \lambda + 0.25 = 0$$

to give

$$\begin{aligned} \lambda &= \frac{-(1) \pm \sqrt{(1)^2 - (4)(-1)(0.25)}}{2(-1)} \\ &= \frac{-1 \pm \sqrt{0}}{-2} \\ &= 0.5, 0.5 \end{aligned}$$

So

$\lambda = 0.5$  and  $\lambda = 0.5$  are the zeroes of

$$-\lambda^2 + \lambda + 0.25$$

giving

$$-\lambda^2 + \lambda + 0.25 = -(\lambda - 0.5)(\lambda - 0.5)$$

Hence

$$-\lambda^3 + 2\lambda^2 - 1.25\lambda + 0.25 = 0$$

can be rewritten as

$$-(\lambda - 1)(\lambda - 0.5)(\lambda - 0.5) = 0$$

to give the roots as

$$\lambda = 1, 0.5, 0.5$$

These are the three roots of the characteristic polynomial equation and hence the eigenvalues of matrix [A].

Note that there are eigenvalues that are repeated. Since there are only two distinct eigenvalues, there are only two eigenspaces. But, corresponding to  $\lambda = 0.5$  there should be two eigenvectors that form a basis for the eigenspace corresponding to  $\lambda = 0.5$ .

Given

$$[(A - \lambda I)][X] = 0$$

then

$$\begin{bmatrix} 1.5 - \lambda & 0 & 1 \\ -0.5 & 0.5 - \lambda & -0.5 \\ -0.5 & 0 & -\lambda \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

For  $\lambda = 0.5$ ,

$$\begin{bmatrix} 1 & 0 & 1 \\ -0.5 & 0 & -0.5 \\ -0.5 & 0 & -0.5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Solving this system gives

$$x_1 = -a, x_2 = b, x_3 = a$$

So

$$\begin{aligned} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} &= \begin{bmatrix} -a \\ b \\ a \end{bmatrix} \\ &= \begin{bmatrix} -a \\ 0 \\ a \end{bmatrix} + \begin{bmatrix} 0 \\ b \\ 0 \end{bmatrix} \\ &= a \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} + b \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \end{aligned}$$

So the vectors  $\begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$  and  $\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$  form a basis for the eigenspace for the eigenvalue  $\lambda = 0.5$  and

are the two eigenvectors corresponding to  $\lambda = 0.5$ .

For  $\lambda = 1$ ,

$$\begin{bmatrix} 0.5 & 0 & 1 \\ -0.5 & -0.5 & -0.5 \\ -0.5 & 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Solving this system gives

$$x_1 = a, x_2 = -0.5a, x_3 = -0.5a$$

The eigenvector corresponding to  $\lambda = 1$  is

$$\begin{bmatrix} a \\ -0.5a \\ -0.5a \end{bmatrix} = a \begin{bmatrix} 1 \\ -0.5 \\ -0.5 \end{bmatrix}$$

Hence the vector

$$\begin{bmatrix} 1 \\ -0.5 \\ -0.5 \end{bmatrix}$$

is a basis for the eigenspace for the eigenvalue of  $\lambda = 1$ , and is the eigenvector corresponding to  $\lambda = 1$ .

### What are some of the theorems of eigenvalues and eigenvectors?

Theorem 1: If  $[A]$  is a  $n \times n$  triangular matrix – upper triangular, lower triangular or diagonal, the eigenvalues of  $[A]$  are the diagonal entries of  $[A]$ .

Theorem 2:  $\lambda = 0$  is an eigenvalue of  $[A]$  if  $[A]$  is a singular (noninvertible) matrix.

Theorem 3:  $[A]$  and  $[A]^T$  have the same eigenvalues.

Theorem 4: Eigenvalues of a symmetric matrix are real.

Theorem 5: Eigenvectors of a symmetric matrix are orthogonal, but only for distinct eigenvalues.

Theorem 6:  $|\det(A)|$  is the product of the absolute values of the eigenvalues of  $[A]$ .

### Example 4

What are the eigenvalues of

$$[A] = \begin{bmatrix} 6 & 0 & 0 & 0 \\ 7 & 3 & 0 & 0 \\ 9 & 5 & 7.5 & 0 \\ 2 & 6 & 0 & -7.2 \end{bmatrix}$$

### Solution

Since the matrix  $[A]$  is a lower triangular matrix, the eigenvalues of  $[A]$  are the diagonal elements of  $[A]$ . The eigenvalues are

$$\lambda_1 = 6, \lambda_2 = 3, \lambda_3 = 7.5, \lambda_4 = -7.2$$

### Example 5

One of the eigenvalues of

$$[A] = \begin{bmatrix} 5 & 6 & 2 \\ 3 & 5 & 9 \\ 2 & 1 & -7 \end{bmatrix}$$

is zero. Is  $[A]$  invertible?

**Solution**

$\lambda = 0$  is an eigenvalue of  $[A]$ , that implies  $[A]$  is singular and is not invertible.

**Example 6**

Given the eigenvalues of

$$[A] = \begin{bmatrix} 2 & -3.5 & 6 \\ 3.5 & 5 & 2 \\ 8 & 1 & 8.5 \end{bmatrix}$$

are

$$\lambda_1 = -1.547, \lambda_2 = 12.33, \lambda_3 = 4.711$$

What are the eigenvalues of  $[B]$  if

$$[B] = \begin{bmatrix} 2 & 3.5 & 8 \\ -3.5 & 5 & 1 \\ 6 & 2 & 8.5 \end{bmatrix}$$

**Solution**

Since  $[B] = [A]^T$ , the eigenvalues of  $[A]$  and  $[B]$  are the same. Hence eigenvalues of  $[B]$  also are

$$\lambda_1 = -1.547, \lambda_2 = 12.33, \lambda_3 = 4.711$$

**Example 7**

Given the eigenvalues of

$$[A] = \begin{bmatrix} 2 & -3.5 & 6 \\ 3.5 & 5 & 2 \\ 8 & 1 & 8.5 \end{bmatrix}$$

are

$$\lambda_1 = -1.547, \lambda_2 = 12.33, \lambda_3 = 4.711$$

Calculate the magnitude of the determinant of the matrix.

**Solution**

Since

$$\begin{aligned} |\det[A]| &= |\lambda_1||\lambda_2||\lambda_3| \\ &= |-1.547||12.33||4.711| \\ &= 89.88 \end{aligned}$$

**How does one find eigenvalues and eigenvectors numerically?**

One of the most common methods used for finding eigenvalues and eigenvectors is the power method. It is used to find the largest eigenvalue in an absolute sense. Note that if this

largest eigenvalues is repeated, this method will not work. Also this eigenvalue needs to be distinct. The method is as follows:

1. Assume a guess  $[X^{(0)}]$  for the eigenvector in

$$[A][X] = \lambda[X]$$

equation. One of the entries of  $[X^{(0)}]$  needs to be unity.

2. Find

$$[Y^{(1)}] = [A][X^{(0)}]$$

3. Scale  $[Y^{(1)}]$  so that the chosen unity component remains unity.

$$[Y^{(1)}] = \lambda^{(1)}[X^{(1)}]$$

4. Repeat steps (2) and (3) with

$$[X] = [X^{(1)}] \text{ to get } [X^{(2)}].$$

5. Repeat the steps 2 and 3 until the value of the eigenvalue converges.

If  $E_s$  is the pre-specified percentage relative error tolerance to which you would like the answer to converge to, keep iterating until

$$\left| \frac{\lambda^{(i+1)} - \lambda^{(i)}}{\lambda^{(i+1)}} \right| \times 100 \leq E_s$$

where the left hand side of the above inequality is the definition of absolute percentage relative approximate error, denoted generally by  $E_s$ . A pre-specified percentage relative tolerance of  $0.5 \times 10^{-m}$  implies at least  $m$  significant digits are current in your answer. When the system converges, the value of  $\lambda$  is the largest (in absolute value) eigenvalue of  $[A]$ .

### Example 8

Using the power method, find the largest eigenvalue and the corresponding eigenvector of

$$[A] = \begin{bmatrix} 1.5 & 0 & 1 \\ -0.5 & 0.5 & -0.5 \\ -0.5 & 0 & 0 \end{bmatrix}$$

### Solution

Assume

$$[X^{(0)}] = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

$$[A][X^{(0)}] = \begin{bmatrix} 1.5 & 0 & 1 \\ -0.5 & 0.5 & -0.5 \\ -0.5 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} 2.5 \\ -0.5 \\ -0.5 \end{bmatrix}$$

$$Y^{(1)} = 2.5 \begin{bmatrix} 1 \\ -0.2 \\ -0.2 \end{bmatrix}$$

$$\lambda^{(1)} = 2.5$$

We will choose the first element of  $[X^{(0)}]$  to be unity.

$$[X^{(1)}] = \begin{bmatrix} 1 \\ -0.2 \\ -0.2 \end{bmatrix}$$

$$[A][X^{(1)}] = \begin{bmatrix} 1.5 & 0 & 1 \\ -0.5 & 0.5 & -0.5 \\ -0.5 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ -0.2 \\ -0.2 \end{bmatrix}$$

$$= \begin{bmatrix} 1.3 \\ -0.5 \\ -0.5 \end{bmatrix}$$

$$[X^{(2)}] = 1.3 \begin{bmatrix} 1 \\ -0.3846 \\ -0.3846 \end{bmatrix}$$

$$\lambda^{(2)} = 1.3$$

$$[X^{(2)}] = \begin{bmatrix} 1 \\ -0.3846 \\ -0.3846 \end{bmatrix}$$

The absolute relative approximate error in the eigenvalues is

$$|\varepsilon_a| = \left| \frac{\lambda^{(2)} - \lambda^{(1)}}{\lambda^{(2)}} \right| \times 100$$

$$= \left| \frac{1.3 - 1.5}{1.5} \right| \times 100$$

$$= 92.307\%$$

Conducting further iterations, the values of  $\lambda^{(i)}$  and the corresponding eigenvectors is given in the table below

$i$	$\lambda^{(i)}$	$[X^{(i)}]$	$ \varepsilon_a  (\%)$
-----	-----------------	-------------	------------------------

1	2.5	$\begin{bmatrix} 1 \\ -0.2 \\ -0.2 \end{bmatrix}$	_____
2	1.3	$\begin{bmatrix} 1 \\ -0.38462 \\ -0.38462 \end{bmatrix}$	92.307
3	1.1154	$\begin{bmatrix} 1 \\ -0.44827 \\ -0.44827 \end{bmatrix}$	16.552
4	1.0517	$\begin{bmatrix} 1 \\ -0.47541 \\ -0.47541 \end{bmatrix}$	6.0529
5	1.02459	$\begin{bmatrix} 1 \\ -0.48800 \\ -0.48800 \end{bmatrix}$	1.2441

The exact value of the eigenvalue is  $\lambda = 1$

and the corresponding eigenvector is

$$[X] = \begin{bmatrix} 1 \\ -0.5 \\ -0.5 \end{bmatrix}$$

### Key Terms:

*Eigenvalue*

*Eigenvectors*

*Power method*

# Chapter 04.11

## Cholesky and $LDL^T$ Decomposition

After reading this chapter, you should be able to:

1. understand why the  $LDL^T$  algorithm is more general than the Cholesky algorithm,
2. understand the differences between the factorization phase and forward solution phase in the Cholesky and  $LDL^T$  algorithms,
3. find the factorized  $[L]$  and  $[D]$  matrices,
4. obtain the forward solution phase,
5. obtain the diagonal scaling phase,
6. obtain the backward solution phase,
7. solve a set of simultaneous linear equations using  $LDL^T$  algorithm.

### Introduction

Solving large (and sparse) system of simultaneous linear equations (SLE) has been (and continues to be) a major challenging problem for many real-world engineering/science applications [1-2]. In matrix notation, a set of SLE can be represented as:

$$[A][x] = [b] \quad (1)$$

where

$[A]$ = known coefficient matrix, with dimension  $n \times n$

$[b]$ = known right-hand-side (RHS)  $n \times 1$  vector

$[x]$ = unknown  $n \times 1$  vector.

### Symmetrical Positive Definite (SPD) SLE

For many practical SLE, the coefficient matrix  $[A]$  (see Equation (1)) is Symmetric Positive Definite (SPD). In this case, the efficient a 3-step Cholesky algorithm [1-2] can be used. A symmetric matrix  $[A]_{n \times n}$  is SPD if either of the following conditions is satisfied:

- (a) If each and every determinant of sub-matrix  $A_{ii}$  ( $i = 1, 2, \dots, n$ ) is positive, or..
- (b) If  $y^T A y > 0$  for any given vector  $[y]_{n \times 1} \neq \vec{0}$

**Example 1**

Find if

$$[A] = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}$$

is SPD?

**Solution**

*Criterion a:* If each and every determinant of sub-matrix  $A_{ii}$  ( $i = 1, 2, \dots, n$ ) is positive.

The given  $3 \times 3$  matrix  $[A]$  is symmetrical, because  $a_{ij} = a_{ji}$ . Furthermore, one has

$$\det[A]_{1 \times 1} = |2| = 2 > 0$$

$$\begin{aligned} \det[A]_{2 \times 2} &= \begin{vmatrix} 2 & -1 \\ -1 & 2 \end{vmatrix} \\ &= 3 > 0 \end{aligned}$$

$$\begin{aligned} \det[A]_{3 \times 3} &= \begin{vmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{vmatrix} \\ &= 1 > 0 \end{aligned}$$

Hence  $[A]$  is SPD.

*Criterion (b):* If  $y^T A y > 0$  for any given vector  $[y]_{n \times 1} \neq \vec{0}$

For any given vector

$$\vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \neq \vec{0},$$

one computes

$$\begin{aligned} y^T A y &= [y_1 \ y_2 \ y_3] \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \\ &= (2y_1^2 - 2y_1y_2 + 2y_2^2) + (y_3^2 - 2y_2y_3) \\ &= (y_1 - y_2)^2 + y_1^2 + y_2^2 + (y_3^2 - 2y_2y_3) \\ &= (y_1 - y_2)^2 + y_1^2 + (y_2 - y_3)^2 > 0 \end{aligned}$$

Since the above scalar is always positive, hence matrix  $[A]$  is SPD.

### Step 1: Matrix Factorization phase

In this step, the coefficient matrix  $[A]$  that is SPD can be decomposed (or factorized) into

$$[A] = [U]^T [U] \quad (2)$$

where,

$[U]$  is a  $n \times n$  upper triangular matrix.

The following simple  $3 \times 3$  matrix example will illustrate how to find the matrix  $[U]$ .

Various terms of the factorized matrix  $[U]$  can be computed/derived as follows (see Equation (2)):

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} u_{11} & 0 & 0 \\ u_{12} & u_{22} & 0 \\ u_{13} & u_{23} & u_{33} \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix} \quad (3)$$

Multiplying two matrices on the right-hand-side (RHS) of Equation (3), and then equating each upper-triangular RHS terms to the corresponding ones on the upper-triangular left-hand-side (LHS), one gets the following 6 equations for the 6 unknowns in the factorized matrix  $[U]$ .

$$u_{11} = \sqrt{a_{11}} ; u_{12} = \frac{a_{12}}{u_{11}} ; u_{13} = \frac{a_{13}}{u_{11}} \quad (4)$$

$$u_{22} = (a_{22} - u_{12}^2)^{\frac{1}{2}} ; u_{23} = \frac{a_{23} - u_{12}u_{13}}{u_{22}} ; u_{33} = (a_{33} - u_{13}^2 - u_{23}^2)^{\frac{1}{2}} \quad (5)$$

In general, for a  $n \times n$  matrix, the diagonal and off-diagonal terms of the factorized matrix  $[U]$  can be computed from the following formulas:

$$u_{ii} = \left( a_{ii} - \sum_{k=1}^{i-1} (u_{ki})^2 \right)^{\frac{1}{2}} \quad (6)$$

$$u_{ij} = \frac{a_{ij} - \sum_{k=1}^{i-1} u_{ki}u_{kj}}{u_{ii}} \quad (7)$$

It is noted that if  $i = j$ , then the numerator of Equation (7) becomes identical to the terms under the square root in Equation (6). In other words, to factorize a general term  $u_{ij}$ , one simply needs to do the following steps:

Step 1.1: Compute the numerator of Equation (7), such as

$$\text{Sum} = a_{ij} - \sum_{k=1}^{i-1} u_{ki}u_{kj}$$

Step 1.2 If  $u_{ij}$  is an off-diagonal term (say,  $i < j$ ) then from Equation (7)

$$u_{ij} = \frac{\text{Sum}}{u_{ii}}$$

else, if  $u_{ij}$  is a diagonal term (that is,  $i = j$ ), then from Equation (6)

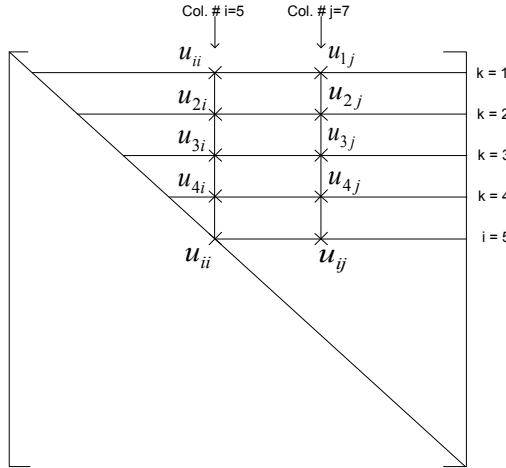
$$u_{ii} = \sqrt{Sum}$$

As a quick example, one computes:

$$u_{57} = \frac{a_{57} - u_{15}u_{17} - u_{25}u_{27} - u_{35}u_{37} - u_{45}u_{47}}{u_{55}} \quad (8)$$

Thus, for computing  $u(i=5, j=7)$ , one only needs to use the (already factorized) data in columns # $i=5$ , and # $j=7$  of  $[U]$ , respectively.

In general, to find the (off-diagonal) factorized term  $u_{ij}$ , one only needs to utilize the “already factorized” columns # $i$ , and # $j$  information (see Figure 1). For example, if  $i=5$ , and  $j=7$ , then Figure 1 will lead to the same formula as shown earlier in Equation (7), or in Equation (8). Similarly, to find the (diagonal) factorized term  $u_{ii}$ , one simply needs to utilize columns # $i$ , and # $i$  (again!) information (see Figure 1). In this case, Figure 1 will lead to the same formula as shown earlier in Equation (6).



**Figure 1** Cholesky Factorization for the term  $u_{ij}$

Since the square root operation involved during the Cholesky factorization phase (see Equation (6)), one must make sure the term under the square root is non-negative. This requirement satisfied by  $[A]$  being SPD.

### Step 2: Forward Solution phase

Substituting Equation (2) into Equation (1), one gets

$$[U]^T [U] [x] = [b] \quad (9)$$

Let us define

$$[U] [x] \equiv [y] \quad (10)$$

Then, Equation (9) becomes

$$[U]^T [y] = [b] \quad (11)$$

Since  $[U]^T$  is a lower triangular matrix, Equation (11) can be efficiently solved for the intermediate unknown vector  $[y]$ , according to the order

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

hence the name “forward solution”.

As a quick example, one has from Equation (11)

$$\begin{bmatrix} u_{11} & 0 & 0 \\ u_{12} & u_{22} & 0 \\ u_{13} & u_{23} & u_{33} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \quad (12)$$

From the 1st row of Equation (12), one gets

$$\begin{aligned} u_{11}y_1 &= b_1 \\ y_1 &= \frac{b_1}{u_{11}} \end{aligned} \quad (13)$$

From the 2<sup>nd</sup> row of Equation (12), one gets

$$\begin{aligned} u_{12}y_1 + u_{22}y_2 &= b_2 \\ y_2 &= b_2 - \frac{u_{12}y_1}{u_{22}} \end{aligned} \quad (14)$$

Similarly

$$y_3 = \frac{b_3 - u_{13}y_1 - u_{23}y_2}{u_{33}} \quad (15)$$

In general, from the  $j^{th}$  row of Equation (12), one has

$$y_j = \frac{b_j - \sum_{i=1}^{j-1} u_{ij}y_i}{u_{jj}} \quad (16)$$

### Step 3: Backward Solution phase

Since  $[U]$  is an upper triangular matrix, Equation (10) can be efficiently solved for the original unknown vector  $[x]$ , according to the order

$$\begin{bmatrix} x_n \\ x_{n-1} \\ x_{n-2} \\ \vdots \\ x_1 \end{bmatrix}$$

and hence the name “backward solution”.

As a quick example, one has from Equation (10)

$$\begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \quad (17)$$

From the last (or  $n^{th} = 3^{rd}$ ) row of Equation (17), one has

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

hence

$$x_3 = \frac{y_3}{u_{33}} \quad (18)$$

Similarly

$$x_2 = \frac{y_2 - u_{23}x_3}{u_{22}} \quad (19)$$

and

$$x_1 = \frac{y_1 - u_{12}x_2 - u_{13}x_3}{u_{11}} \quad (20)$$

In general, one has

$$x_j = \frac{y_j - \sum_{i=j+1}^n u_{ji}x_i}{u_{jj}} \quad (21)$$

Amongst the above 3-step Cholesky algorithms, factorization phase in step 1 consumes about 95% of the total SLE solution time.

If the coefficient matrix  $[A]$  is symmetrical but not necessarily positive definite, then the above Cholesky algorithms will not be valid. In this case, the following  $LDL^T$  factorized algorithms can be employed

$$[A] = [L][D][L]^T \quad (22)$$

For example

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix} \begin{bmatrix} d_{11} & 0 & 0 \\ 0 & d_{22} & 0 \\ 0 & 0 & d_{33} \end{bmatrix} \begin{bmatrix} 1 & l_{21} & l_{31} \\ 0 & 1 & l_{32} \\ 0 & 0 & 1 \end{bmatrix} \quad (23)$$

Multiplying the three matrices on the RHS of Equation (23), then equating the resulting upper-triangular RHS terms of Equation (23) to the corresponding ones on the LHS, one obtains the following formulas for the diagonal  $[D]$ , and lower-triangular  $[L]$  matrices

$$d_{jj} = a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2 d_{kk} \quad (24)$$

$$l_{ij} = \left( a_{ij} - \sum_{k=1}^{j-1} l_{ik} d_{kk} l_{jk} \right) \times \left( \frac{1}{d_{jj}} \right) \quad (25)$$

Thus, the  $LDL^T$  algorithms can be summarized by the following step-by-step procedures.

**Step1: Factorization phase**

$$[A] = [L][D][L]^T \quad (22, \text{repeated})$$

**Step 2: Forward solution and diagonal scaling phase**

Substituting Equation (22) into Equation (1), one gets

$$[L][D][L]^T[x] = [b] \quad (26)$$

Let us define

$$\begin{aligned} [L]^T[x] &= [y] \\ \begin{bmatrix} 1 & l_{21} & l_{31} \\ 0 & 1 & l_{32} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} &= \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \end{aligned} \quad (27)$$

$$x_i = y_i - \sum_{k=i+1}^n l_{ki} x_k; \text{ for } i = n, n-1, \dots, 2, 1 \quad (28)$$

Also, define

$$\begin{aligned} [D][y] &= [z] \\ \begin{bmatrix} d_{11} & 0 & 0 \\ 0 & d_{22} & 0 \\ 0 & 0 & d_{33} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} &= \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} \end{aligned} \quad (29)$$

$$y_i = \frac{z_i}{d_{ii}}, \text{ for } i = 1, 2, 3, \dots, n \quad (30)$$

Then Equation (26) becomes

$$\begin{aligned} [L][z] &= [b] \\ \begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} &= \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \end{aligned} \quad (31)$$

$$z_i = b_i - \sum_{k=1}^{i-1} L_{ik} z_k \quad \text{for } i = 1, 2, 3, \dots, n \quad (32)$$

Equation (31) can be efficiently solved for the vector  $[z]$ , and then Equation (29) can be conveniently (and trivially) solved for the vector  $[y]$ .

**Step 3: Backward solution phase**

In this step, Equation (27) can be efficiently solved for the original unknown vector  $[x]$ .

**Example 2**

Using the Cholesky algorithm, solve the following SLE system for the unknown vector  $[x]$ .

$$[A][x] = [b]$$

where

$$[A] = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}$$

$$[b] = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

**Solution**

The factorized, upper triangular matrix  $[U]$  can be computed by either referring to Equations (6-7), or looking at Figure 1, as following:

Row 1 of  $[U]$  is given below.

$$\begin{aligned} u_{11} &= \sqrt{a_{11}} \\ &= \sqrt{2} \\ &= 1.414 \end{aligned}$$

$$\begin{aligned} u_{12} &= \frac{a_{12}}{u_{11}} \\ &= \frac{-1}{1.414} \\ &= -0.7071 \end{aligned}$$

$$\begin{aligned} u_{13} &= \frac{a_{13}}{u_{11}} \\ &= \frac{0}{1.414} \\ &= 0 \end{aligned}$$

Row 2 of  $[U]$  is given below

$$\begin{aligned} u_{22} &= \left\{ a_{22} - \sum_{k=1}^{i-1=1} (u_{ki})^2 \right\}^{\frac{1}{2}} \\ &= \left\{ 2 - (u_{12})^2 \right\}^{\frac{1}{2}} \\ &= \sqrt{2 - (-0.7071)^2} \\ &= 1.225 \end{aligned}$$

$$\begin{aligned}
 u_{23} &= \frac{a_{23} - \sum_{k=1}^{i-1=1} u_{ki} u_{kj}}{U_{22}} \\
 &= \frac{-1 - u_{12} \times u_{13}}{1.225} \\
 &= \frac{-1 - (-0.7071)(0)}{1.225} \\
 &= -0.8165
 \end{aligned}$$

Row 3 of [U] is given below

$$\begin{aligned}
 u_{33} &= \left\{ a_{33} - \sum_{k=1}^{i-1=2} (u_{ki})^2 \right\}^{\frac{1}{2}} \\
 &= \left\{ a_{33} - u_{13}^2 - u_{23}^2 \right\}^{\frac{1}{2}} \\
 &= \sqrt{1 - (0)^2 - (-0.8165)^2} \\
 &= 0.5774
 \end{aligned}$$

Thus, the factorized matrix

$$[U] = \begin{bmatrix} 1.414 & -0.7071 & 0 \\ 0 & 1.225 & -0.8165 \\ 0 & 0 & 0.5774 \end{bmatrix}$$

The forward solution phase, shown in Equation (11), becomes

$$\begin{bmatrix} U^T \\ [y] \end{bmatrix} = \begin{bmatrix} b \\ [y] \end{bmatrix}$$

$$\begin{bmatrix} 1.414 & 0 & 0 \\ -0.7071 & 1.225 & 0 \\ 0 & -0.8165 & 0.5774 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

Thus, Equation (16) can be used to solve for [y] as

$$\begin{aligned}
 y_1 &= \frac{b_1}{u_{11}} \\
 &= \frac{1}{1.414} \\
 &= 0.7071 \\
 y_2 &= \frac{b_2 - \sum_{i=1}^{j-1=1} u_{ij} y_i}{u_{22}} \\
 &= \frac{0 - (u_{12} = -0.7071)(y_1 = 0.7071)}{(u_{22} = 1.225)} \\
 &= 0.4082
 \end{aligned}$$

$$\begin{aligned}
y_3 &= \frac{b_3 - \sum_{i=1}^{j-1=2} u_{ij} y_i}{u_{jj}} \\
&= \frac{0 - (u_{13} = 0)(y_1 = 0.7071) - (u_{23} = -0.8165)(y_2 = 0.4082)}{(u_{33} = 0.5774)} \\
&= 0.5774
\end{aligned}$$

The backward solution phase, shown in Equation (10), becomes:

$$[U][x] = [y]$$

$$\begin{bmatrix} 1.414 & -0.7071 & 0 \\ 0 & 1.225 & -0.8165 \\ 0 & 0 & 0.5774 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0.7071 \\ 0.4082 \\ 0.5774 \end{bmatrix}$$

Thus, Equation (21) can be used to solve

$$\begin{aligned}
x_3 &= \frac{y_j}{u_{jj}} \\
&= \frac{y_3}{u_{33}} \\
&= \frac{0.5774}{0.5774} \\
&= 1 \\
x_2 &= \frac{y_j - \sum_{i=j+1=3}^{N=3} u_{ji} x_i}{u_{jj}} \\
&= \frac{y_2 - u_{23} x_3}{u_{22}} \\
&= \frac{0.4082 - (-0.8165)(1)}{1.225} \\
&= 1 \\
x_1 &= \frac{y_j - \sum_{i=j+1=2}^{N=3} u_{ji} x_i}{u_{jj}} \\
&= \frac{y_1 - u_{12} x_2 - u_{13} x_3}{u_{11}} \\
&= \frac{0.7071 - (-0.7071)(1) - (0)(1)}{1.414} \\
&= 1
\end{aligned}$$

Hence

$$[x] = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

### Example 3

Using the  $LDL^T$  algorithm, solve the following SLE system for the unknown vector  $[x]$ .

$$[A][x] = [b]$$

where

$$[A] = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}$$

$$[b] = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

### Solution

The factorized matrices  $[D]$  and  $[L]$  can be computed from Equation (24) and Equation (25), respectively.

$$\left. \begin{array}{l} d_{11} = a_{11} - \sum_{k=1}^{j-1=0} l_{jk}^2 d_{kk} \\ \quad = a_{11} \\ \quad = 2 \\ l_{11} = 1 \text{ (always !)} \\ l_{21} = \frac{a_{21} - \sum_{k=1}^{j-1=0} l_{ik} d_{kk} l_{jk}}{d_{11}} \\ \quad = \frac{a_{21}}{d_{11}} \\ \quad = \frac{-1}{2} \\ \quad = -0.5 \\ l_{31} = \frac{a_{31}}{d_{11}} \\ \quad = \frac{0}{2} \\ \quad = 0 \end{array} \right\} \text{Column 1 of matrices of } [D] \text{ and } [L]$$

$$\left. \begin{array}{l}
 d_{22} = a_{22} - \sum_{k=1}^{j-1=1} l_{jk}^2 d_{kk} \\
 = 2 - l_{21}^2 d_{11} \\
 = 2 - (-0.5)^2 (2) \\
 = 1.5 \\
 l_{22} = 1 \text{ (always !)} \\
 l_{32} = \frac{a_{32} - \sum_{k=1}^{j-1=1} l_{31} d_{11} l_{21}}{d_{22}} \\
 = \frac{-1 - (0)(2)(-0.5)}{1.5} \\
 = -0.6667
 \end{array} \right\} \text{Column 2 of matrices } [D] \text{ and } [L]$$
  

$$\left. \begin{array}{l}
 d_{33} = a_{33} - \sum_{k=1}^{j-1=2} l_{jk}^2 d_{kk} \\
 = 1 - l_{31}^2 d_{11} - l_{32}^2 d_{22} \\
 = 1 - (0)^2 (2) - (-0.6667)^2 (1.5) \\
 = 0.3333
 \end{array} \right\} \text{Column 3 of matrices } [D] \text{ and } [L]$$

Hence

$$[D] = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1.5 & 0 \\ 0 & 0 & 0.3333 \end{bmatrix}$$

and

$$[L] = \begin{bmatrix} 1 & 0 & 0 \\ -0.5 & 1 & 0 \\ 0 & -0.6667 & 1 \end{bmatrix}$$

The forward solution shown in Equation (31) becomes:

$$\begin{aligned}
 [L][z] &= [b] \\
 \begin{bmatrix} 1 & 0 & 0 \\ -0.5 & 1 & 0 \\ 0 & -0.6667 & 1 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} &= \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \text{ or}
 \end{aligned}$$

$$z_i = b_i - \sum_{k=1}^{i-1} l_{ik} z_k \quad (32, \text{ repeated})$$

Hence

$$\begin{aligned}
 z_1 &= b_1 = 1 \\
 z_2 &= b_2 - L_{21}z_1 \\
 &= 0 - (-0.5)(1) \\
 &= 0.5 \\
 z_3 &= b_3 - L_{31}z_1 - L_{32}z_2 \\
 &= 0 - (0)(1) - (-0.6667)(0.5) \\
 &= 0.3333
 \end{aligned}$$

The diagonal scaling phase, shown in Equation (29) becomes

$$\begin{bmatrix} D \end{bmatrix} \begin{bmatrix} y \end{bmatrix} = \begin{bmatrix} z \end{bmatrix}$$

$$\begin{bmatrix} 2 & 0 & 0 \\ 0 & 1.5 & 0 \\ 0 & 0 & 0.3333 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0.5 \\ 0.3333 \end{bmatrix}, \text{ or}$$

$$y_i = \frac{z_i}{d_{ii}}$$

Hence

$$\begin{aligned}
 y_1 &= \frac{z_1}{d_{11}} \\
 &= \frac{1}{2} \\
 &= 0.5
 \end{aligned}$$

$$\begin{aligned}
 y_2 &= \frac{z_2}{d_{22}} \\
 &= \frac{0.5}{1.5} \\
 &= 0.3333
 \end{aligned}$$

$$\begin{aligned}
 y_3 &= \frac{z_3}{d_{33}} \\
 &= \frac{0.3333}{0.3333} \\
 &= 1
 \end{aligned}$$

The backward solution phase can be found by referring to Equation (27)

$$\begin{bmatrix} L \end{bmatrix}^T \begin{bmatrix} x \end{bmatrix} = \begin{bmatrix} y \end{bmatrix}$$

$$\begin{bmatrix} 1 & -0.5 & 0 \\ 0 & 1 & -0.667 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.333 \\ 1 \end{bmatrix}$$

$$x_i = y_i - \sum_{k=i+1}^N l_{ki}x_k \tag{28, repeated}$$

Hence

$$\begin{aligned}x_3 &= y_3 \\&= 1 \\x_2 &= y_2 - l_{32}x_3 \\&= 0.3333 - (-0.6667) \times 1 \\x_2 &= 1 \\x_1 &= y_1 - l_{21}x_2 - l_{31}x_3 \\x_1 &= 0.5 - (-0.5)(1) - (0)(1) \\&= 1\end{aligned}$$

Hence

$$\begin{aligned}[x] &= \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \\&= \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}\end{aligned}$$

Through this numerical example, one clearly sees that the “square root operations” have NOT been involved during the entire  $LDL^T$  algorithms. Thus, the coefficient matrix  $[A]$ , shown in Equation (1) is NOT required to be SPD.

### Re-ordering Algorithms For Minimizing Fill-in Terms [1,2].

During the factorization phase (of Cholesky, or  $LDL^T$  algorithms), many “zero” terms in the original/given matrix  $[A]$  will become “non-zero” terms in the factored matrix  $[U]$ . These new non-zero terms are often called as “fill-in” terms (indicated by the symbol  $F$ ). It is, therefore, highly desirable to minimize these fill-in terms, so that both computational time/effort and computer memory requirements can be substantially reduced. For example, the following matrix  $[A]$  and vector  $[b]$  are given:

$$[A] = \begin{bmatrix} 112 & 7 & 0 & 0 & 0 & 2 \\ 7 & 110 & 5 & 4 & 3 & 0 \\ 0 & 5 & 88 & 0 & 0 & 1 \\ 0 & 4 & 0 & 66 & 0 & 0 \\ 0 & 3 & 0 & 0 & 44 & 0 \\ 2 & 0 & 1 & 0 & 0 & 11 \end{bmatrix} \quad (33)$$

$$[b] = \begin{bmatrix} 121 \\ 129 \\ 94 \\ 70 \\ 47 \\ 14 \end{bmatrix} \quad (34)$$

The Cholesky factorization matrix  $[U]$ , based on the original matrix  $[A]$  (see Equation 33) and Equations (6-7), or Figure 1, can be symbolically computed as

$$[U] = \begin{bmatrix} \times & \times & 0 & 0 & 0 & \times \\ 0 & \times & \times & \times & \times & F \\ 0 & 0 & \times & F & F & \times \\ 0 & 0 & 0 & \times & F & F \\ 0 & 0 & 0 & 0 & \times & F \\ 0 & 0 & 0 & 0 & 0 & \times \end{bmatrix} \quad (35)$$

In Equation (35), the symbols  $\times$ , and  $F$  represents the “non-zero” and “fill-in” terms, respectively.

In practical applications, however, it is always a necessary step to rearrange the original matrix  $[A]$  through re-ordering algorithms (or subroutines) [Refs 1-2] and produce the following integer mapping array

$$IPERM(\text{new equation \#}) = \{\text{old equation \#}\} \quad (36)$$

such as, for this particular example:

$$IPERM \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{bmatrix} = \begin{bmatrix} 6 \\ 5 \\ 4 \\ 3 \\ 2 \\ 1 \end{bmatrix} \quad (37)$$

Using the above results (see Equation 37), one will be able to construct the following rearranged matrices:

$$[A^*] = \begin{bmatrix} 11 & 0 & 0 & 1 & 0 & 2 \\ 0 & 44 & 0 & 0 & 3 & 0 \\ 0 & 0 & 66 & 0 & 4 & 0 \\ 1 & 0 & 0 & 88 & 5 & 0 \\ 0 & 3 & 4 & 5 & 110 & 7 \\ 2 & 0 & 0 & 0 & 7 & 112 \end{bmatrix} \quad (38)$$

and

$$[b^*] = \begin{bmatrix} 14 \\ 47 \\ 70 \\ 94 \\ 129 \\ 121 \end{bmatrix} \quad (39)$$

In the original matrix  $A$  (shown in Equation 33), the nonzero term  $A$  (old row 1, old column 2) = 7 will move to new location of the new matrix  $A^*$  (new row 6, new column 5) = 7, etc.

The non zero term  $A$  (old row 3, old column 3) = 88 will move to  $A^*$  (new row 4, new column 4) = 88, etc.

The value of  $b$  (old row 4) = 70 will be moved to (or located at)  $b^*$  (new row 3) = 70, etc.

Now, one would like to solve the following modified system of linear equations (SLE) for  $[x^*]$ ,

$$[A^*][x^*] = [b^*] \quad (40)$$

rather than to solve the original SLE (see Equation (1)). The original unknown vector  $\{x\}$  can be easily recovered from  $[x^*]$  and  $[IPERM]$ , shown in Equation (37).

The factorized matrix  $[U^*]$  can be “symbolically” computed from  $[A^*]$  as (by referring to either Figure 1 or Equations 6-7):

$$[U^*] = \begin{bmatrix} \times & 0 & 0 & \times & 0 & \times \\ 0 & \times & 0 & 0 & \times & 0 \\ 0 & 0 & \times & 0 & \times & 0 \\ 0 & 0 & 0 & \times & \times & F \\ 0 & 0 & 0 & 0 & \times & \times \\ 0 & 0 & 0 & 0 & 0 & \times \end{bmatrix} \quad (41)$$

You can clearly see the big benefits of solving the SLE shown in Equation (40), instead of solving the original Equation (1), since the factorized matrix  $[U^*]$  has only 1 fill-in term (see the symbol “F” in Equation 41), as compared to six fill-in-terms occurred in the factorized matrix  $[U]$  as shown in Equation 35.

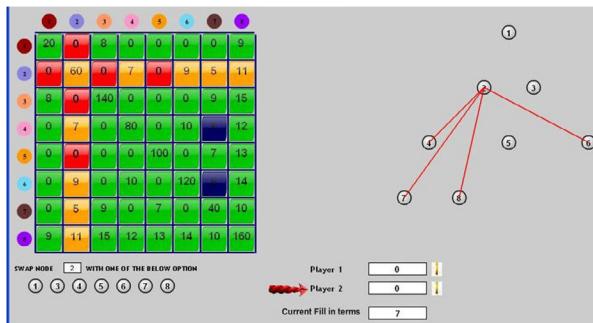
#### On-Line Chess-Like Game For Reordering/Factorized Phase [4].

Based on the discussions presented in the previous section 2 (about factorization phase), and section 3 (about reordering phase), one can easily see the similar operations between the symbolic, numerical factorization and reordering (to minimize the number of fill-in terms) phases of sparse SLE.

In practical computer implementation for the solution of SLE, the reordering phase is usually conducted first (to produce the mapping between “old-new” equation numbers, as indicated in the integer array  $IPERM(-)$ , see Equations 36-37).

Then, the sparse *symbolic* factorization phase is followed by using either Cholesky Equations 6-7, or the  $LDL^T$  Equations 24-25 (without requiring the actual/numerical values to be computed). The reason is because during the *symbolic factorization* phase, one only wishes to find the number (and the location) of non-zero *fill-in terms*. This *symbolic* factorization process is necessary for allocating the “computer memory” requirement for the “numerical factorization” phase which will actually compute the exact numerical values of  $[U^*]$ , based on the same Cholesky Equations (6-7) (or the  $LDL^T$  Equations (24-25)).

In this work, a chess-like game (shown in Figure 2, Ref. [4]) has been designed with the following objectives:



**Figure 2** A Chess-Like Game For Learning to Solve SLE.

- (A) Teaching undergraduates the process how to use the reordering output IPERM(-), see Equations (36-37) for converting the original/given matrix  $[A]$ , see Equation (33), into the new/modified matrix  $[A^*]$ , see Equation (38). This step is reflected in Figure 2, when the “Game Player” decides to swap node (or equation)  $i$  (say  $i = 2$ ) with another node (or equation)  $j$ , and click the CONFIRM icon! Since node  $i=2$  is currently connected to nodes  $j=4, 6, 7, 8$ , swapping node  $i = 2$  with the above nodes  $j$  will *NOT* change the number/pattern of the *fill-in* terms. However, if node  $i = 2$  is swapped with node  $j=1, 3$  or  $5$ , then the fill-in terms pattern may change (for better or worse)!
- (B) Helping undergraduates to understand the “symbolic” factorization phase by symbolically utilizing the Cholesky factorized Equations (6-7). This step is illustrated in Figure 2, for which the “game player” will see (and also hear the computer animated sound, and human voice) the non-zero terms (including fill-in terms) of the original matrix  $[A]$  to move to the new locations in the new/modified matrix  $[A^*]$ .
- (C) Helping undergraduates to understand the *numerical factorization* phase, by numerically utilizing the same Cholesky factorized Equations (6-7).
- (D) Teaching undergraduates to *understand existing reordering concepts*, or to *discover new reordering algorithms*.

### Further Explanation on the Developed Game

1. In the above chess-like game, which is available on-line [4], powerful features of FLASH computer environment [3], such as animated sound, human voice, motions, graphical colors etc... have been incorporated and programmed into the developed game-software for more appeal to game players/learners.
2. In the developed chess-like game, fictitious monetary (or any kind of ‘scoring system’) is rewarded (and broadcasted by computer animated human voice) to game players, based on how he/she swaps the node (or equation) numbers, and consequently based on how many fill-in  $F$  terms occurred. In general, less fill-in terms introduced will result in more rewards.
3. Based on the original/given matrix  $[A]$ , and existing re-ordering algorithms (such as the Reverse Cuthill-McKee, or RCM algorithms [1-2]) the number of fill-in terms  $F$  can be computed using RCM algorithms. This internally generated information will be used to judge how good the players/learners are, and/or broadcast “congratulations

message” to a particular player who discovers a new “chess-like move” (or, swapping node) strategies which are even better than RCM algorithms.

4. Initially, the player(s) will select the matrix size ( $8 \times 8$ , or larger is recommended), and the percentage (50%, or larger is suggested) of zero-terms (or sparsity of the matrix). Then, the *START Game* icon will be clicked by the player.
5. The player will then CLICK one of the selected node  $i$  (or equation) numbers appearing on the computer screen. The player will see those nodes  $j$  which are connected to node  $i$  (based on the given/generated matrix  $[A]$ ). The player then has to decide to swap node  $i$  with one of the possible node  $j$ . After *confirming* the player’s decision, the outcomes/results will be announced by the computer animated human voice, and the monetary-award will (or will not) be given to the players/learners, accordingly. In this software, a maximum of \$1,000,000 can be earned by the player, and the exact dollar amount will be *inversely* proportional to the number of fill-in terms occurred (based on the player’s decision on how to swap node  $i$  with another node  $j$ ).
6. The next player will continue to play, with his/her move (meaning to swap the  $i^{th}$  node with the  $j^{th}$  node) based on the *current best* non-zero terms pattern of the matrix.

## References

---

### **CHOLESKY AND $LDL^T$ DECOMPOSITION**

---

Topic	Cholesky and $LDL^T$ Decomposition
Summary	Textbook chapter of Cholesky and $LDL^T$ Decomposition
Major	General Engineering
Authors	Duc Nguyen
Date	July 29, 2010
Web Site	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

---

## 05.01

# History of Interpolation

*After reading this chapter, you should be able to:*

1. *Know the history of Interpolation and its current uses by the HNMI.*

### History

Sir Edmund Whittaker, a professor of Numerical Mathematics at the University of Edinburgh from 1913 to 1923, observed “the most common form of interpolation occurs when we seek data from a table which does not have the exact values we want.” Throughout history, interpolation has been used in one form or another for just about every purpose under the sun.

Speaking of the sun, some of the first surviving evidence of the use of interpolation came from ancient Babylon and Greece. Around 300 BC, they were using not only linear, but also more complex forms of interpolation to predict the positions of the sun, moon, and the planets they knew of. Farmers, timing the planting of their crops, were the primary users of these predictions. Also in Greece sometime around 150 BC, Hipparchus of Rhodes used linear interpolation to construct a “chord function”, which is similar to a sinusoidal function, to compute positions of celestial bodies.

Farther east, Chinese evidence of interpolation dates back to around 600 AD. Liu Zhuo used the equivalent of second order Gregory-Newton interpolation to construct an “Imperial Standard Calendar”. In 625 AD, Indian astronomer and mathematician Brahmagupta introduced a method for second order interpolation of the sine function and, later on, a method for interpolation of unequal-interval data.

Many similar land-based purposes were found for interpolation over the ages, but ocean navigation was found to be one of the most important applications for centuries. Tables of special function values were constructed using numerical methods, and seafarers used certain ones to determine latitude and longitude values. The French government started production on an extensive set of such tables when the metric system was introduced. Ideally, one would want mathematicians to construct a large set of tables due to their proficiency at the subject. However, the primary source of work on the project ended up being hairdressers who had lost their gaudy-wigged customers to the guillotine.

The unfortunate truth about special function tables is that most of them were plagiarized. Since the “computers”, the workers who carried out and recorded the calculations, were prone to making many errors during the creation of these daunting tables,

plagiarism only propagated more errors. Charles Babbage tried to solve this problem with the invention of his “difference engine”, a mechanical computer programmed by the use of punch cards. On the side, Babbage also tried inventing a system that would choose winning horse race numbers, hoping to raise extra money. Although he was not short of funds, his life ran short and never saw the completion of the invention. Over a century and a quarter later, as we plunge into the nano-technology era, Babbage is now considered the grandfather of modern computing.

During the Great Depression, one final burst of manual table-making found its way into the United States. The Works Progress Administration began the Mathematical Tables Project shortly before World War II. As with the French project, the desired “mathematician” workers ended up being unskilled—this time to the point that negative numbers were puzzling. The solution: black pencils for positive numbers and red ones for negative numbers. Having each calculation in this project iterated twice (each by a different person), and extensive proof reading carried out, these tables were “possibly the most accurate ever produced”. Many of them were collected in a book by Milton Abramowitz and Irene Stegun, which is still in worldwide use today. With computers (not the people type, either), tables are no longer manually constructed, but the Australian Government produces life tables which describe mortality rates. Relevant to the life insurance industry and the study of demography, “these tables are extended using modern interpolation methods.” No matter how advanced or extensive, interpolation will always be needed to find values in modern tables due to their nature. Since they aren’t continuous functions, there will be infinitely many missing values.

Two of the methods of interpolation taught at the HNMI are credited to Newton and Lagrange. Newton began his work on the subject in 1675, which “laid the foundation of classical interpolation theory”. In 1795, Lagrange published the interpolation formula now known under his name, despite the fact that Waring had already produced the same formula sixteen years earlier.

## Bibliography

Kahaner, David, Cleve Moler, and Stephen Nash. Numerical Methods and Software. Englewood Cliffs, NJ: Prentice Hall, 1989.

Meijering, Erik. “A Chronology of Interpolation: From Ancient Astronomy to Modern Signal and Image Processing.” Proceedings of the IEEE. vol. 90, no. 3, pp. 319-42. March 2002.

Mills, Terry. “Historical Notes.” Join the Dots and See the World. La Trobe University, Bendigo, Australia. <http://www.bendigo.latrobe.edu.au/rahdo/research/worner96.htm>.

---

**INTERPOLATION**

---

Topic	History of Interpolation
Summary	Textbook notes on the history of Interpolation and its current uses in the HNMI.
Major	All Majors of Engineering
Authors	Autar Kaw, Michael Keteltas
Last Revised	December 23, 2009
Web Site	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

---

## Chapter 05.02

# Direct Method of Interpolation

After reading this chapter, you should be able to:

1. apply the direct method of interpolation,
2. solve problems using the direct method of interpolation, and
3. use the direct method interpolants to find derivatives and integrals of discrete functions.

### What is interpolation?

Many times, data is given only at discrete points such as  $(x_0, y_0)$ ,  $(x_1, y_1)$ , ...,  $(x_{n-1}, y_{n-1})$ ,  $(x_n, y_n)$ . So, how then does one find the value of  $y$  at any other value of  $x$ ? Well, a continuous function  $f(x)$  may be used to represent the  $n+1$  data values with  $f(x)$  passing through the  $n+1$  points (Figure 1). Then one can find the value of  $y$  at any other value of  $x$ . This is called *interpolation*.

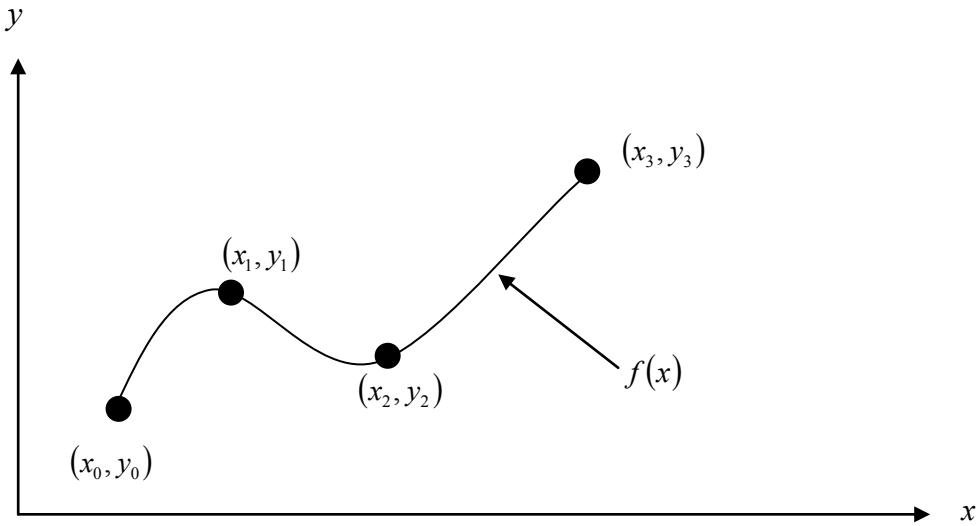
Of course, if  $x$  falls outside the range of  $x$  for which the data is given, it is no longer interpolation but instead is called *extrapolation*.

So what kind of function  $f(x)$  should one choose? A polynomial is a common choice for an interpolating function because polynomials are easy to

- (A) evaluate,
- (B) differentiate, and
- (C) integrate

relative to other choices such as a trigonometric and exponential series.

Polynomial interpolation involves finding a polynomial of order  $n$  that passes through the  $n+1$  points. One of the methods of interpolation is called the direct method. Other methods include Newton's divided difference polynomial method and the Lagrangian interpolation method. We will discuss the direct method in this chapter.



**Figure 1** Interpolation of discrete data.

### Direct Method

The direct method of interpolation is based on the following premise. Given  $n+1$  data points, fit a polynomial of order  $n$  as given below

$$y = a_0 + a_1 x + \dots + a_n x^n \quad (1)$$

through the data, where  $a_0, a_1, \dots, a_n$  are  $n+1$  real constants. Since  $n+1$  values of  $y$  are given at  $n+1$  values of  $x$ , one can write  $n+1$  equations. Then the  $n+1$  constants,  $a_0, a_1, \dots, a_n$  can be found by solving the  $n+1$  simultaneous linear equations. To find the value of  $y$  at a given value of  $x$ , simply substitute the value of  $x$  in Equation 1.

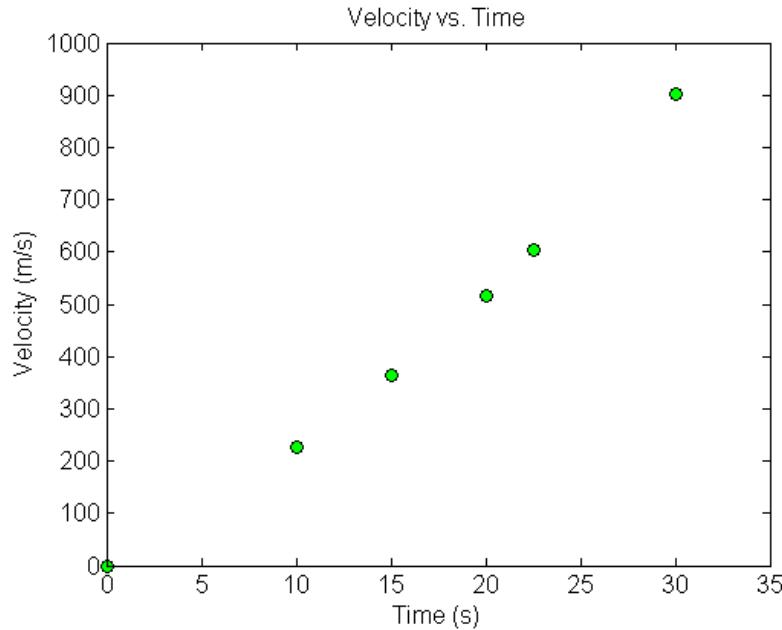
But, it is not necessary to use all the data points. How does one then choose the order of the polynomial and what data points to use? This concept and the direct method of interpolation are best illustrated using examples.

### Example 1

The upward velocity of a rocket is given as a function of time in Table 1.

**Table 1** Velocity as a function of time.

$t$ (s)	$v(t)$ (m/s)
0	0
10	227.04
15	362.78
20	517.35
22.5	602.97
30	901.67



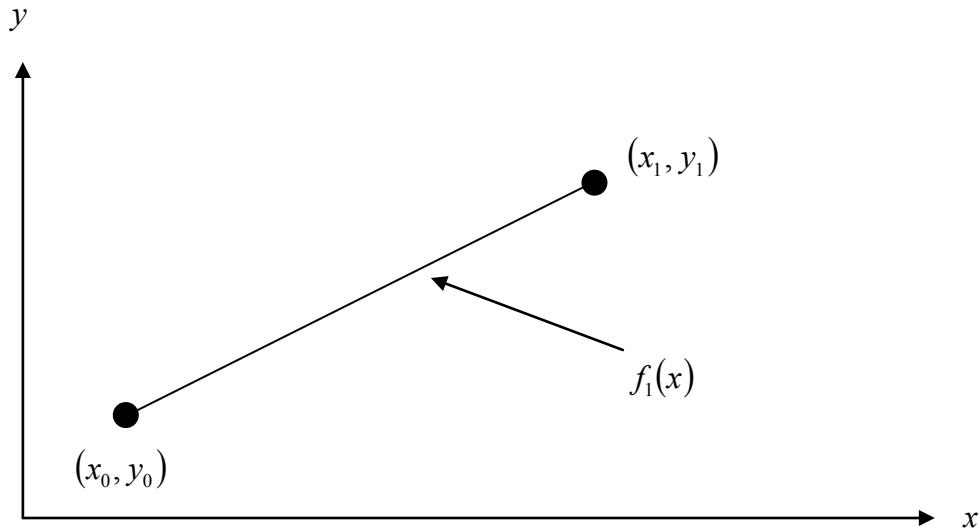
**Figure 2** Graph of velocity vs. time data for the rocket example.

Determine the value of the velocity at  $t = 16$  seconds using the direct method of interpolation and a first order polynomial.

### Solution

For first order polynomial interpolation (also called linear interpolation), the velocity given by

$$v(t) = a_0 + a_1 t$$



**Figure 3** Linear interpolation.

Since we want to find the velocity at  $t = 16$ , and we are using a first order polynomial, we need to choose the two data points that are closest to  $t = 16$  that also bracket  $t = 16$  to evaluate it. The two points are  $t_0 = 15$  and  $t_1 = 20$ .

Then

$$t_0 = 15, \quad v(t_0) = 362.78$$

$$t_1 = 20, \quad v(t_1) = 517.35$$

gives

$$v(15) = a_0 + a_1(15) = 362.78$$

$$v(20) = a_0 + a_1(20) = 517.35$$

Writing the equations in matrix form, we have

$$\begin{bmatrix} 1 & 15 \\ 1 & 20 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} 362.78 \\ 517.35 \end{bmatrix}$$

Solving the above two equations gives

$$a_0 = -100.93$$

$$a_1 = 30.914$$

Hence

$$\begin{aligned} v(t) &= a_0 + a_1 t \\ &= -100.93 + 30.914t, \quad 15 \leq t \leq 20 \end{aligned}$$

At  $t = 16$ ,

$$\begin{aligned} v(16) &= -100.93 + 30.914 \times 16 \\ &= 393.7 \text{ m/s} \end{aligned}$$

## Example 2

The upward velocity of a rocket is given as a function of time in Table 2.

**Table 2** Velocity as a function of time.

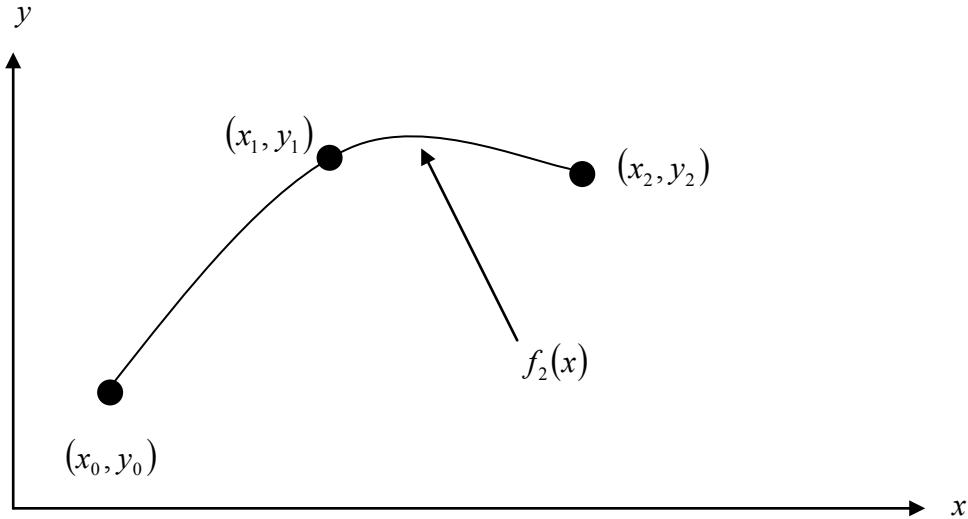
$t$ (s)	$v(t)$ (m/s)
0	0
10	227.04
15	362.78
20	517.35
22.5	602.97
30	901.67

Determine the value of the velocity at  $t = 16$  seconds using the direct method of interpolation and a second order polynomial.

### Solution

For second order polynomial interpolation (also called quadratic interpolation), the velocity is given by

$$v(t) = a_0 + a_1 t + a_2 t^2$$



**Figure 4** Quadratic interpolation.

Since we want to find the velocity at  $t = 16$ , and we are using a second order polynomial, we need to choose the three data points that are closest to  $t = 16$  that also bracket  $t = 16$  to evaluate it. The three points are  $t_0 = 10$ ,  $t_1 = 15$ , and  $t_2 = 20$ .

Then

$$t_0 = 10, \quad v(t_0) = 227.04$$

$$t_1 = 15, \quad v(t_1) = 362.78$$

$$t_2 = 20, \quad v(t_2) = 517.35$$

gives

$$v(10) = a_0 + a_1(10) + a_2(10)^2 = 227.04$$

$$v(15) = a_0 + a_1(15) + a_2(15)^2 = 362.78$$

$$v(20) = a_0 + a_1(20) + a_2(20)^2 = 517.35$$

Writing the three equations in matrix form, we have

$$\begin{bmatrix} 1 & 10 & 100 \\ 1 & 15 & 225 \\ 1 & 20 & 400 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 227.04 \\ 362.78 \\ 517.35 \end{bmatrix}$$

Solving the above three equations gives

$$a_0 = 12.05$$

$$a_1 = 17.733$$

$$a_2 = 0.3766$$

Hence

$$v(t) = 12.05 + 17.733t + 0.3766t^2, \quad 10 \leq t \leq 20$$

At  $t = 16$ ,

$$\begin{aligned} v(16) &= 12.05 + 17.733(16) + 0.3766(16)^2 \\ &= 392.19 \text{ m/s} \end{aligned}$$

The absolute relative approximate error  $|e_a|$  obtained between the results from the first and second order polynomial is

$$\begin{aligned}|e_a| &= \left| \frac{392.19 - 393.70}{392.19} \right| \times 100 \\ &= 0.38410\%\end{aligned}$$

### Example 3

The upward velocity of a rocket is given as a function of time in Table 3.

**Table 3** Velocity as a function of time.

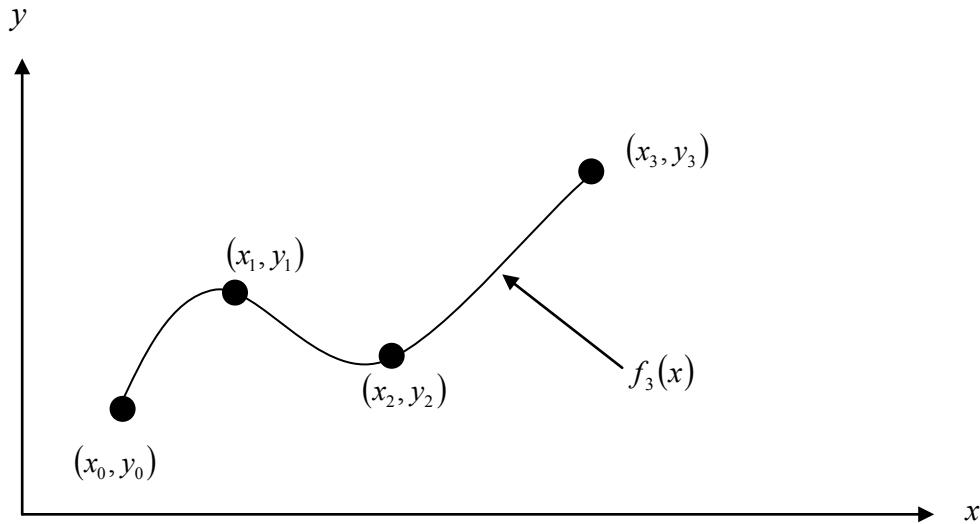
$t$ (s)	$v(t)$ (m/s)
0	0
10	227.04
15	362.78
20	517.35
22.5	602.97
30	901.67

- a) Determine the value of the velocity at  $t = 16$  seconds using the direct method of interpolation and a third order polynomial.
- b) Find the absolute relative approximate error for the third order polynomial approximation.
- c) Using the third order polynomial interpolant for velocity from part (a), find the distance covered by the rocket from  $t = 11$ s to  $t = 16$ s .
- d) Using the third order polynomial interpolant for velocity from part (a), find the acceleration of the rocket at  $t = 16$ s .

### Solution

- a) For third order polynomial interpolation (also called cubic interpolation), we choose the velocity given by

$$v(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3$$



**Figure 5** Cubic interpolation.

Since we want to find the velocity at  $t = 16$ , and we are using a third order polynomial, we need to choose the four data points closest to  $t = 16$  that also bracket  $t = 16$  to evaluate it.

The four points are  $t_0 = 10$ ,  $t_1 = 15$ ,  $t_2 = 20$  and  $t_3 = 22.5$ .

Then

$$\begin{aligned} t_0 &= 10, \quad v(t_0) = 227.04 \\ t_1 &= 15, \quad v(t_1) = 362.78 \\ t_2 &= 20, \quad v(t_2) = 517.35 \\ t_3 &= 22.5, \quad v(t_3) = 602.97 \end{aligned}$$

gives

$$\begin{aligned} v(10) &= a_0 + a_1(10) + a_2(10)^2 + a_3(10)^3 = 227.04 \\ v(15) &= a_0 + a_1(15) + a_2(15)^2 + a_3(15)^3 = 362.78 \\ v(20) &= a_0 + a_1(20) + a_2(20)^2 + a_3(20)^3 = 517.35 \\ v(22.5) &= a_0 + a_1(22.5) + a_2(22.5)^2 + a_3(22.5)^3 = 602.97 \end{aligned}$$

Writing the four equations in matrix form, we have

$$\begin{bmatrix} 1 & 10 & 100 & 1000 \\ 1 & 15 & 225 & 3375 \\ 1 & 20 & 400 & 8000 \\ 1 & 22.5 & 506.25 & 11391 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 227.04 \\ 362.78 \\ 517.35 \\ 602.97 \end{bmatrix}$$

Solving the above four equations gives

$$a_0 = -4.2540$$

$$a_1 = 21.266$$

$$a_2 = 0.13204$$

$$a_3 = 0.0054347$$

Hence

$$\begin{aligned} v(t) &= a_0 + a_1 t + a_2 t^2 + a_3 t^3 \\ &= -4.2540 + 21.266t + 0.13204t^2 + 0.0054347t^3, \quad 10 \leq t \leq 22.5 \\ v(16) &= -4.2540 + 21.266(16) + 0.13204(16)^2 + 0.0054347(16)^3 \\ &= 392.06 \text{ m/s} \end{aligned}$$

b) The absolute percentage relative approximate error  $|\epsilon_a|$  for the value obtained for  $v(16)$  between second and third order polynomial is

$$\begin{aligned} |\epsilon_a| &= \left| \frac{392.06 - 392.19}{392.06} \right| \times 100 \\ &= 0.033269\% \end{aligned}$$

c) The distance covered by the rocket between  $t = 11$  s and  $t = 16$  s can be calculated from the interpolating polynomial

$$v(t) = -4.2540 + 21.266t + 0.13204t^2 + 0.0054347t^3, \quad 10 \leq t \leq 22.5$$

Note that the polynomial is valid between  $t = 10$  and  $t = 22.5$  and hence includes the limits of integration of  $t = 11$  and  $t = 16$ .

So

$$\begin{aligned} s(16) - s(11) &= \int_{11}^{16} v(t) dt \\ &= \int_{11}^{16} (-4.2540 + 21.266t + 0.13204t^2 + 0.0054347t^3) dt \\ &= \left[ -4.2540t + 21.266 \frac{t^2}{2} + 0.13204 \frac{t^3}{3} + 0.0054347 \frac{t^4}{4} \right]_{11}^{16} \\ &= 1605 \text{ m} \end{aligned}$$

d) The acceleration at  $t = 16$  is given by

$$a(16) = \frac{d}{dt} v(t) \Big|_{t=16}$$

Given that

$$\begin{aligned} v(t) &= -4.2540 + 21.266t + 0.13204t^2 + 0.0054347t^3, \quad 10 \leq t \leq 22.5 \\ a(t) &= \frac{d}{dt} v(t) \\ &= \frac{d}{dt} (-4.2540 + 21.266t + 0.13204t^2 + 0.0054347t^3) \\ &= 21.266 + 0.26408t + 0.016304t^2, \quad 10 \leq t \leq 22.5 \\ a(16) &= 21.266 + 0.26408(16) + 0.016304(16)^2 \\ &= 29.665 \text{ m/s}^2 \end{aligned}$$

---

**INTERPOLATION**

---

Topic      Direct Method of Interpolation  
Summary    Textbook notes on the direct method of interpolation.  
Major      General Engineering  
Authors     Autar Kaw, Peter Warr, Michael Keteltas  
Date       June 17, 2012

---

Web Site    <http://numericalmethods.eng.usf.edu>

---

# Chapter 05.03

## Newton's Divided Difference Interpolation

After reading this chapter, you should be able to:

1. derive Newton's divided difference method of interpolation,
2. apply Newton's divided difference method of interpolation, and
3. apply Newton's divided difference method interpolants to find derivatives and integrals.

### What is interpolation?

Many times, data is given only at discrete points such as  $(x_0, y_0)$ ,  $(x_1, y_1)$ , ...,  $(x_{n-1}, y_{n-1})$ ,  $(x_n, y_n)$ . So, how then does one find the value of  $y$  at any other value of  $x$ ? Well, a continuous function  $f(x)$  may be used to represent the  $n+1$  data values with  $f(x)$  passing through the  $n+1$  points (Figure 1). Then one can find the value of  $y$  at any other value of  $x$ . This is called *interpolation*.

Of course, if  $x$  falls outside the range of  $x$  for which the data is given, it is no longer interpolation but instead is called *extrapolation*.

So what kind of function  $f(x)$  should one choose? A polynomial is a common choice for an interpolating function because polynomials are easy to

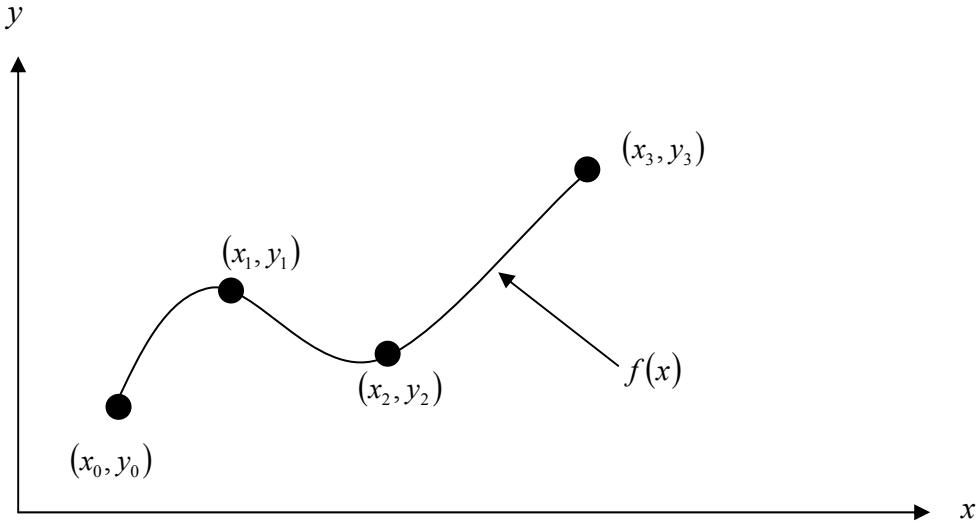
- (A) evaluate,
- (B) differentiate, and
- (C) integrate,

relative to other choices such as a trigonometric and exponential series.

Polynomial interpolation involves finding a polynomial of order  $n$  that passes through the  $n+1$  points. One of the methods of interpolation is called Newton's divided difference polynomial method. Other methods include the direct method and the Lagrangian interpolation method. We will discuss Newton's divided difference polynomial method in this chapter.

### Newton's Divided Difference Polynomial Method

To illustrate this method, linear and quadratic interpolation is presented first. Then, the general form of Newton's divided difference polynomial method is presented. To illustrate the general form, cubic interpolation is shown in Figure 1.



**Figure 1** Interpolation of discrete data.

### Linear Interpolation

Given  $(x_0, y_0)$  and  $(x_1, y_1)$ , fit a linear interpolant through the data. Noting  $y = f(x)$  and  $y_1 = f(x_1)$ , assume the linear interpolant  $f_1(x)$  is given by (Figure 2)

$$f_1(x) = b_0 + b_1(x - x_0)$$

Since at  $x = x_0$ ,

$$f_1(x_0) = f(x_0) = b_0 + b_1(x_0 - x_0) = b_0$$

and at  $x = x_1$ ,

$$\begin{aligned} f_1(x_1) &= f(x_1) = b_0 + b_1(x_1 - x_0) \\ &= f(x_0) + b_1(x_1 - x_0) \end{aligned}$$

giving

$$b_1 = \frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

So

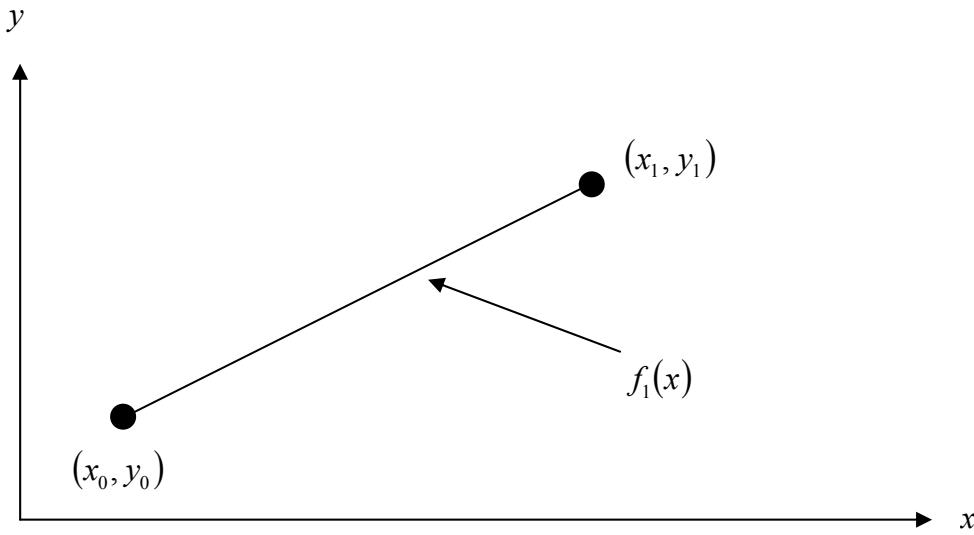
$$b_0 = f(x_0)$$

$$b_1 = \frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

giving the linear interpolant as

$$f_1(x) = b_0 + b_1(x - x_0)$$

$$f_1(x) = f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_0)$$



**Figure 2** Linear interpolation.

### Example 1

The upward velocity of a rocket is given as a function of time in Table 1 (Figure 3).

**Table 1** Velocity as a function of time.

$t$ (s)	$v(t)$ (m/s)
0	0
10	227.04
15	362.78
20	517.35
22.5	602.97
30	901.67

Determine the value of the velocity at  $t = 16$  seconds using first order polynomial interpolation by Newton's divided difference polynomial method.

### Solution

For linear interpolation, the velocity is given by

$$v(t) = b_0 + b_1(t - t_0)$$

Since we want to find the velocity at  $t = 16$ , and we are using a first order polynomial, we need to choose the two data points that are closest to  $t = 16$  that also bracket  $t = 16$  to evaluate it. The two points are  $t = 15$  and  $t = 20$ .

Then

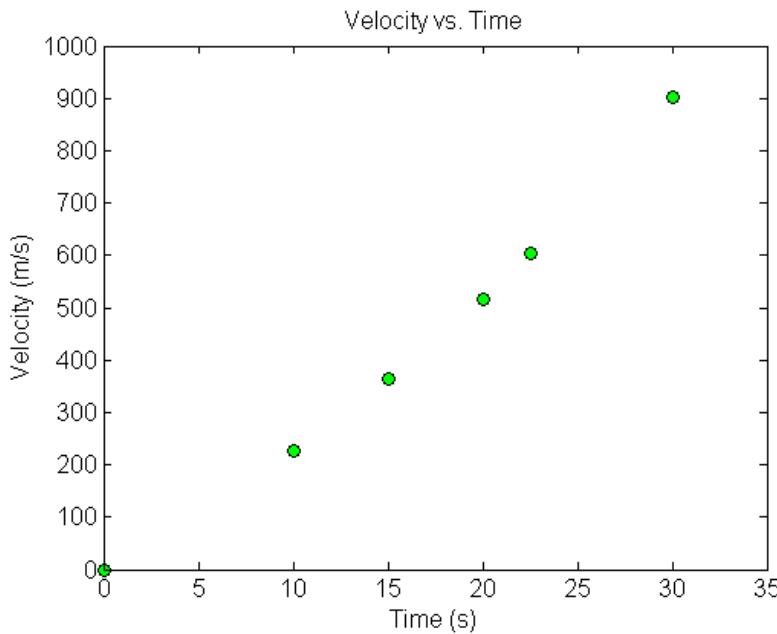
$$t_0 = 15, v(t_0) = 362.78$$

$$t_1 = 20, v(t_1) = 517.35$$

gives

$$b_0 = v(t_0)$$

$$\begin{aligned}
 &= 362.78 \\
 b_1 &= \frac{v(t_1) - v(t_0)}{t_1 - t_0} \\
 &= \frac{517.35 - 362.78}{20 - 15} \\
 &= 30.914
 \end{aligned}$$



**Figure 3** Graph of velocity vs. time data for the rocket example.

Hence

$$\begin{aligned}
 v(t) &= b_0 + b_1(t - t_0) \\
 &= 362.78 + 30.914(t - 15), \quad 15 \leq t \leq 20
 \end{aligned}$$

At  $t = 16$ ,

$$\begin{aligned}
 v(16) &= 362.78 + 30.914(16 - 15) \\
 &= 393.69 \text{ m/s}
 \end{aligned}$$

If we expand

$$v(t) = 362.78 + 30.914(t - 15), \quad 15 \leq t \leq 20$$

we get

$$v(t) = -100.93 + 30.914t, \quad 15 \leq t \leq 20$$

and this is the same expression as obtained in the direct method.

### Quadratic Interpolation

Given  $(x_0, y_0)$ ,  $(x_1, y_1)$ , and  $(x_2, y_2)$ , fit a quadratic interpolant through the data. Noting  $y = f(x)$ ,  $y_0 = f(x_0)$ ,  $y_1 = f(x_1)$ , and  $y_2 = f(x_2)$ , assume the quadratic interpolant  $f_2(x)$  is given by

$$f_2(x) = b_0 + b_1(x - x_0) + b_2(x - x_0)(x - x_1)$$

At  $x = x_0$ ,

$$\begin{aligned}f_2(x_0) &= f(x_0) = b_0 + b_1(x_0 - x_0) + b_2(x_0 - x_0)(x_0 - x_1) \\&= b_0 \\b_0 &= f(x_0)\end{aligned}$$

At  $x = x_1$

$$\begin{aligned}f_2(x_1) &= f(x_1) = b_0 + b_1(x_1 - x_0) + b_2(x_1 - x_0)(x_1 - x_1) \\f(x_1) &= f(x_0) + b_1(x_1 - x_0)\end{aligned}$$

giving

$$b_1 = \frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

At  $x = x_2$

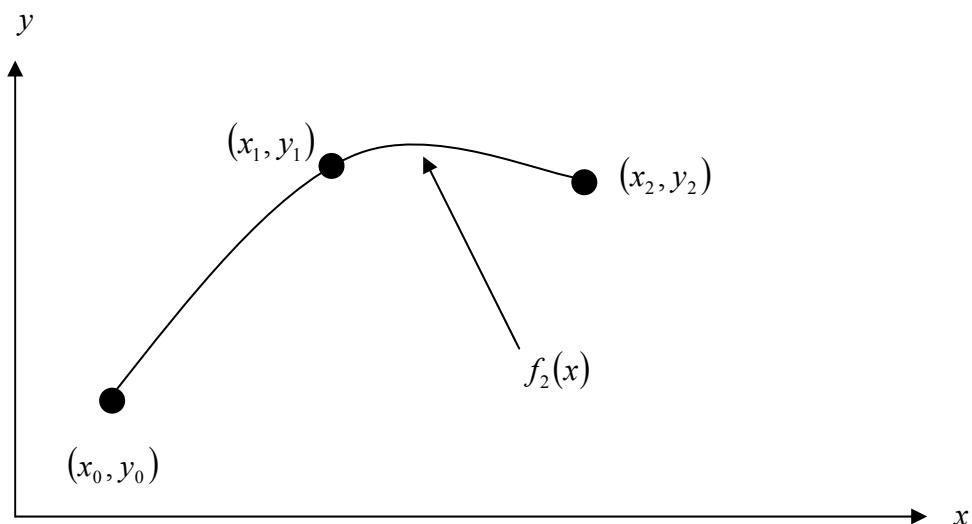
$$\begin{aligned}f_2(x_2) &= f(x_2) = b_0 + b_1(x_2 - x_0) + b_2(x_2 - x_0)(x_2 - x_1) \\f(x_2) &= f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x_2 - x_0) + b_2(x_2 - x_0)(x_2 - x_1)\end{aligned}$$

Giving

$$b_2 = \frac{\frac{f(x_2) - f(x_1)}{x_2 - x_1} - \frac{f(x_1) - f(x_0)}{x_1 - x_0}}{x_2 - x_0}$$

Hence the quadratic interpolant is given by

$$\begin{aligned}f_2(x) &= b_0 + b_1(x - x_0) + b_2(x - x_0)(x - x_1) \\&= f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_0) + \frac{\frac{f(x_2) - f(x_1)}{x_2 - x_1} - \frac{f(x_1) - f(x_0)}{x_1 - x_0}}{x_2 - x_0}(x - x_0)(x - x_1)\end{aligned}$$



**Figure 4** Quadratic interpolation.

**Example 2**

The upward velocity of a rocket is given as a function of time in Table 2.

**Table 2** Velocity as a function of time.

$t$ (s)	$v(t)$ (m/s)
0	0
10	227.04
15	362.78
20	517.35
22.5	602.97
30	901.67

Determine the value of the velocity at  $t = 16$  seconds using second order polynomial interpolation using Newton's divided difference polynomial method.

**Solution**

For quadratic interpolation, the velocity is given by

$$v(t) = b_0 + b_1(t - t_0) + b_2(t - t_0)(t - t_1)$$

Since we want to find the velocity at  $t = 16$ , and we are using a second order polynomial, we need to choose the three data points that are closest to  $t = 16$  that also bracket  $t = 16$  to evaluate it. The three points are  $t_0 = 10$ ,  $t_1 = 15$ , and  $t_2 = 20$ .

Then

$$t_0 = 10, v(t_0) = 227.04$$

$$t_1 = 15, v(t_1) = 362.78$$

$$t_2 = 20, v(t_2) = 517.35$$

gives

$$b_0 = v(t_0)$$

$$= 227.04$$

$$b_1 = \frac{v(t_1) - v(t_0)}{t_1 - t_0}$$

$$= \frac{362.78 - 227.04}{15 - 10}$$

$$= 27.148$$

$$b_2 = \frac{v(t_2) - v(t_1)}{t_2 - t_1} - \frac{v(t_1) - v(t_0)}{t_1 - t_0}$$

$$= \frac{517.35 - 362.78}{20 - 15} - \frac{362.78 - 227.04}{15 - 10}$$

$$= \frac{30.914 - 27.148}{10}$$

$$= 0.37660$$

Hence

$$\begin{aligned} v(t) &= b_0 + b_1(t - t_0) + b_2(t - t_0)(t - t_1) \\ &= 227.04 + 27.148(t - 10) + 0.37660(t - 10)(t - 15), \quad 10 \leq t \leq 20 \end{aligned}$$

At  $t = 16$ ,

$$\begin{aligned} v(16) &= 227.04 + 27.148(16 - 10) + 0.37660(16 - 10)(16 - 15) \\ &= 392.19 \text{ m/s} \end{aligned}$$

If we expand

$$v(t) = 227.04 + 27.148(t - 10) + 0.37660(t - 10)(t - 15), \quad 10 \leq t \leq 20$$

we get

$$v(t) = 12.05 + 17.733t + 0.37660t^2, \quad 10 \leq t \leq 20$$

This is the same expression obtained by the direct method.

### General Form of Newton's Divided Difference Polynomial

In the two previous cases, we found linear and quadratic interpolants for Newton's divided difference method. Let us revisit the quadratic polynomial interpolant formula

$$f_2(x) = b_0 + b_1(x - x_0) + b_2(x - x_0)(x - x_1)$$

where

$$\begin{aligned} b_0 &= f(x_0) \\ b_1 &= \frac{f(x_1) - f(x_0)}{x_1 - x_0} \\ b_2 &= \frac{\frac{f(x_2) - f(x_1)}{x_2 - x_1} - \frac{f(x_1) - f(x_0)}{x_1 - x_0}}{x_2 - x_0} \end{aligned}$$

Note that  $b_0$ ,  $b_1$ , and  $b_2$  are finite divided differences.  $b_0$ ,  $b_1$ , and  $b_2$  are the first, second, and third finite divided differences, respectively. We denote the first divided difference by

$$f[x_0] = f(x_0)$$

the second divided difference by

$$f[x_1, x_0] = \frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

and the third divided difference by

$$\begin{aligned} f[x_2, x_1, x_0] &= \frac{f[x_2, x_1] - f[x_1, x_0]}{x_2 - x_0} \\ &= \frac{\frac{f(x_2) - f(x_1)}{x_2 - x_1} - \frac{f(x_1) - f(x_0)}{x_1 - x_0}}{x_2 - x_0} \end{aligned}$$

where  $f[x_0]$ ,  $f[x_1, x_0]$ , and  $f[x_2, x_1, x_0]$  are called bracketed functions of their variables enclosed in square brackets.

Rewriting,

$$f_2(x) = f[x_0] + f[x_1, x_0](x - x_0) + f[x_2, x_1, x_0](x - x_0)(x - x_1)$$

This leads us to writing the general form of the Newton's divided difference polynomial for  $n+1$  data points,  $(x_0, y_0), (x_1, y_1), \dots, (x_{n-1}, y_{n-1}), (x_n, y_n)$ , as

$$f_n(x) = b_0 + b_1(x - x_0) + \dots + b_n(x - x_0)(x - x_1)\dots(x - x_{n-1})$$

where

$$b_0 = f[x_0]$$

$$b_1 = f[x_1, x_0]$$

$$b_2 = f[x_2, x_1, x_0]$$

⋮

$$b_{n-1} = f[x_{n-1}, x_{n-2}, \dots, x_0]$$

$$b_n = f[x_n, x_{n-1}, \dots, x_0]$$

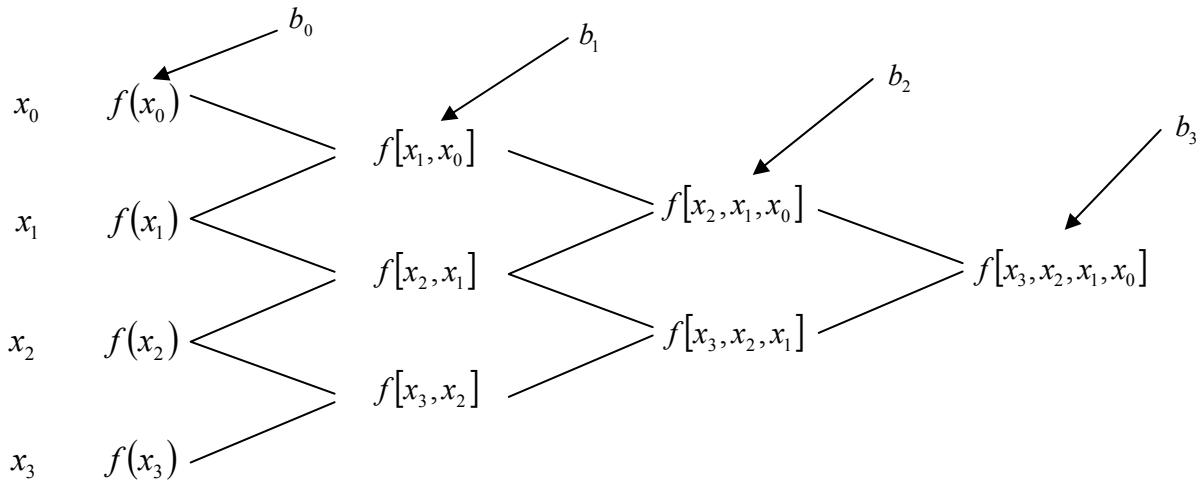
where the definition of the  $m^{\text{th}}$  divided difference is

$$\begin{aligned} b_m &= f[x_m, \dots, x_0] \\ &= \frac{f[x_m, \dots, x_1] - f[x_{m-1}, \dots, x_0]}{x_m - x_0} \end{aligned}$$

From the above definition, it can be seen that the divided differences are calculated recursively.

For an example of a third order polynomial, given  $(x_0, y_0)$ ,  $(x_1, y_1)$ ,  $(x_2, y_2)$ , and  $(x_3, y_3)$ ,

$$\begin{aligned} f_3(x) &= f[x_0] + f[x_1, x_0](x - x_0) + f[x_2, x_1, x_0](x - x_0)(x - x_1) \\ &\quad + f[x_3, x_2, x_1, x_0](x - x_0)(x - x_1)(x - x_2) \end{aligned}$$



**Figure 5** Table of divided differences for a cubic polynomial.

### Example 3

The upward velocity of a rocket is given as a function of time in Table 3.

**Table 3** Velocity as a function of time.

$t$ (s)	$v(t)$ (m/s)
0	0
10	227.04
15	362.78
20	517.35
22.5	602.97
30	901.67

- a) Determine the value of the velocity at  $t = 16$  seconds with third order polynomial interpolation using Newton's divided difference polynomial method.  
 b) Using the third order polynomial interpolant for velocity, find the distance covered by the rocket from  $t = 11\text{ s}$  to  $t = 16\text{ s}$ .  
 c) Using the third order polynomial interpolant for velocity, find the acceleration of the rocket at  $t = 16\text{ s}$ .

**Solution**

a) For a third order polynomial, the velocity is given by

$$v(t) = b_0 + b_1(t - t_0) + b_2(t - t_0)(t - t_1) + b_3(t - t_0)(t - t_1)(t - t_2)$$

Since we want to find the velocity at  $t = 16$ , and we are using a third order polynomial, we need to choose the four data points that are closest to  $t = 16$  that also bracket  $t = 16$  to evaluate it. The four data points are  $t_0 = 10$ ,  $t_1 = 15$ ,  $t_2 = 20$ , and  $t_3 = 22.5$ .

Then

$$t_0 = 10, \quad v(t_0) = 227.04$$

$$t_1 = 15, \quad v(t_1) = 362.78$$

$$t_2 = 20, \quad v(t_2) = 517.35$$

$$t_3 = 22.5, \quad v(t_3) = 602.97$$

gives

$$b_0 = v[t_0]$$

$$= v(t_0)$$

$$= 227.04$$

$$b_1 = v[t_1, t_0]$$

$$= \frac{v(t_1) - v(t_0)}{t_1 - t_0}$$

$$= \frac{362.78 - 227.04}{15 - 10}$$

$$= 27.148$$

$$b_2 = v[t_2, t_1, t_0]$$

$$= \frac{v[t_2, t_1] - v[t_1, t_0]}{t_2 - t_0}$$

$$v[t_2, t_1] = \frac{v(t_2) - v(t_1)}{t_2 - t_1}$$

$$= \frac{517.35 - 362.78}{20 - 15}$$

$$= 30.914$$

$$v[t_1, t_0] = 27.148$$

$$b_2 = \frac{v[t_2, t_1] - v[t_1, t_0]}{t_2 - t_0}$$

$$= \frac{30.914 - 27.148}{20 - 10}$$

$$= 0.37660$$

$$b_3 = v[t_3, t_2, t_1, t_0]$$

$$= \frac{v[t_3, t_2, t_1] - v[t_2, t_1, t_0]}{t_3 - t_0}$$

$$v[t_3, t_2, t_1] = \frac{v[t_3, t_2] - v[t_2, t_1]}{t_3 - t_1}$$

$$v[t_3, t_2] = \frac{v(t_3) - v(t_2)}{t_3 - t_2}$$

$$= \frac{602.97 - 517.35}{22.5 - 20}$$

$$= 34.248$$

$$v[t_2, t_1] = \frac{v(t_2) - v(t_1)}{t_2 - t_1}$$

$$= \frac{517.35 - 362.78}{20 - 15}$$

$$= 30.914$$

$$v[t_3, t_2, t_1] = \frac{v[t_3, t_2] - v[t_2, t_1]}{t_3 - t_1}$$

$$= \frac{34.248 - 30.914}{22.5 - 15}$$

$$= 0.44453$$

$$v[t_2, t_1, t_0] = 0.37660$$

$$b_3 = \frac{v[t_3, t_2, t_1] - v[t_2, t_1, t_0]}{t_3 - t_0}$$

$$= \frac{0.44453 - 0.37660}{22.5 - 10}$$

$$= 5.4347 \times 10^{-3}$$

Hence

$$v(t) = b_0 + b_1(t - t_0) + b_2(t - t_0)(t - t_1) + b_3(t - t_0)(t - t_1)(t - t_2)$$

$$= 227.04 + 27.148(t-10) + 0.37660(t-10)(t-15) \\ + 5.5347 \times 10^{-3}(t-10)(t-15)(t-20)$$

At  $t = 16$ ,

$$v(16) = 227.04 + 27.148(16-10) + 0.37660(16-10)(16-15) \\ + 5.5347 \times 10^{-3}(16-10)(16-15)(16-20) \\ = 392.06 \text{ m/s}$$

b) The distance covered by the rocket between  $t = 11 \text{ s}$  and  $t = 16 \text{ s}$  can be calculated from the interpolating polynomial

$$v(t) = 227.04 + 27.148(t-10) + 0.37660(t-10)(t-15) \\ + 5.5347 \times 10^{-3}(t-10)(t-15)(t-20) \\ = -4.2541 + 21.265t + 0.13204t^2 + 0.0054347t^3, \quad 10 \leq t \leq 22.5$$

Note that the polynomial is valid between  $t = 10$  and  $t = 22.5$  and hence includes the limits of  $t = 11$  and  $t = 16$ .

So

$$s(16) - s(11) = \int_{11}^{16} v(t) dt \\ = \int_{11}^{16} (-4.2541 + 21.265t + 0.13204t^2 + 0.0054347t^3) dt \\ = \left[ -4.2541t + 21.265 \frac{t^2}{2} + 0.13204 \frac{t^3}{3} + 0.0054347 \frac{t^4}{4} \right]_{11}^{16} \\ = 1605 \text{ m}$$

c) The acceleration at  $t = 16$  is given by

$$a(16) = \frac{d}{dt} v(t) \Big|_{t=16} \\ a(t) = \frac{d}{dt} v(t) \\ = \frac{d}{dt} \left( -4.2541 + 21.265t + 0.13204t^2 + 0.0054347t^3 \right) \\ = 21.265 + 0.26408t + 0.016304t^2 \\ a(16) = 21.265 + 0.26408(16) + 0.016304(16)^2 \\ = 29.664 \text{ m/s}^2$$

## INTERPOLATION

Topic	Newton's Divided Difference Interpolation
Summary	Textbook notes on Newton's divided difference interpolation.
Major	General Engineering
Authors	Autar Kaw, Michael Keteltas
Last Revised	December 23, 2009
Web Site	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

## Chapter 05.04

# Lagrangian Interpolation

After reading this chapter, you should be able to:

1. derive Lagrangian method of interpolation,
2. solve problems using Lagrangian method of interpolation, and
3. use Lagrangian interpolants to find derivatives and integrals of discrete functions.

### What is interpolation?

Many times, data is given only at discrete points such as  $(x_0, y_0)$ ,  $(x_1, y_1)$ , ...,  $(x_{n-1}, y_{n-1})$ ,  $(x_n, y_n)$ . So, how then does one find the value of  $y$  at any other value of  $x$ ? Well, a continuous function  $f(x)$  may be used to represent the  $n+1$  data values with  $f(x)$  passing through the  $n+1$  points (Figure 1). Then one can find the value of  $y$  at any other value of  $x$ . This is called *interpolation*.

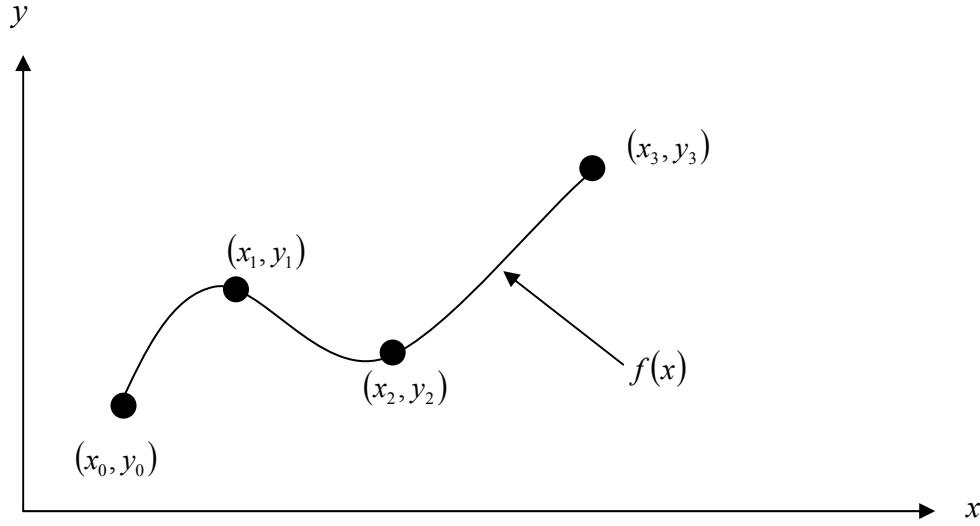
Of course, if  $x$  falls outside the range of  $x$  for which the data is given, it is no longer interpolation but instead is called *extrapolation*.

So what kind of function  $f(x)$  should one choose? A polynomial is a common choice for an interpolating function because polynomials are easy to

- (A) evaluate,
- (B) differentiate, and
- (C) integrate,

relative to other choices such as a trigonometric and exponential series.

Polynomial interpolation involves finding a polynomial of order  $n$  that passes through the  $n+1$  data points. One of the methods used to find this polynomial is called the Lagrangian method of interpolation. Other methods include Newton's divided difference polynomial method and the direct method. We discuss the Lagrangian method in this chapter.



**Figure 1** Interpolation of discrete data.

The Lagrangian interpolating polynomial is given by

$$f_n(x) = \sum_{i=0}^n L_i(x)f(x_i)$$

where  $n$  in  $f_n(x)$  stands for the  $n^{\text{th}}$  order polynomial that approximates the function  $y = f(x)$  given at  $n+1$  data points as  $(x_0, y_0), (x_1, y_1), \dots, (x_{n-1}, y_{n-1}), (x_n, y_n)$ , and

$$L_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^n \frac{x - x_j}{x_i - x_j}$$

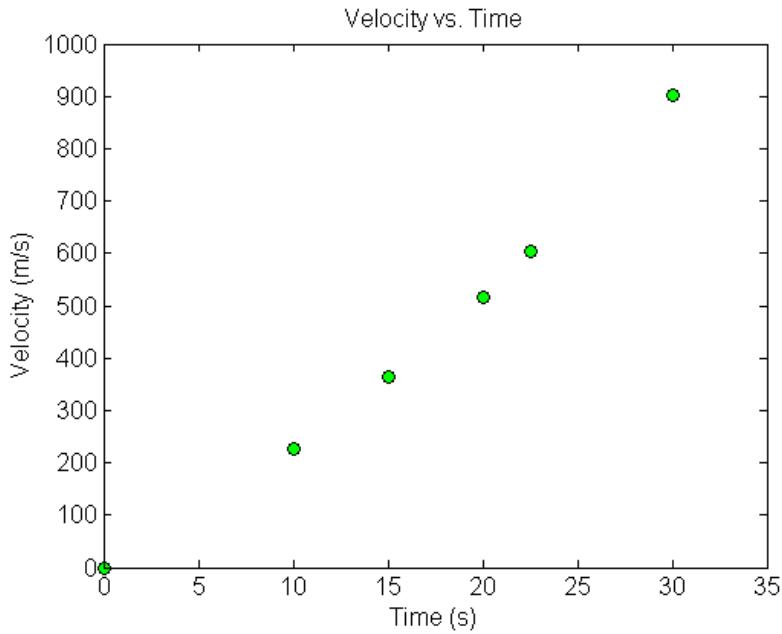
$L_i(x)$  is a weighting function that includes a product of  $n-1$  terms with terms of  $j=i$  omitted. The application of Lagrangian interpolation will be clarified using an example.

### Example 1

The upward velocity of a rocket is given as a function of time in Table 1.

**Table 1** Velocity as a function of time.

$t$ (s)	$v(t)$ (m/s)
0	0
10	227.04
15	362.78
20	517.35
22.5	602.97
30	901.67



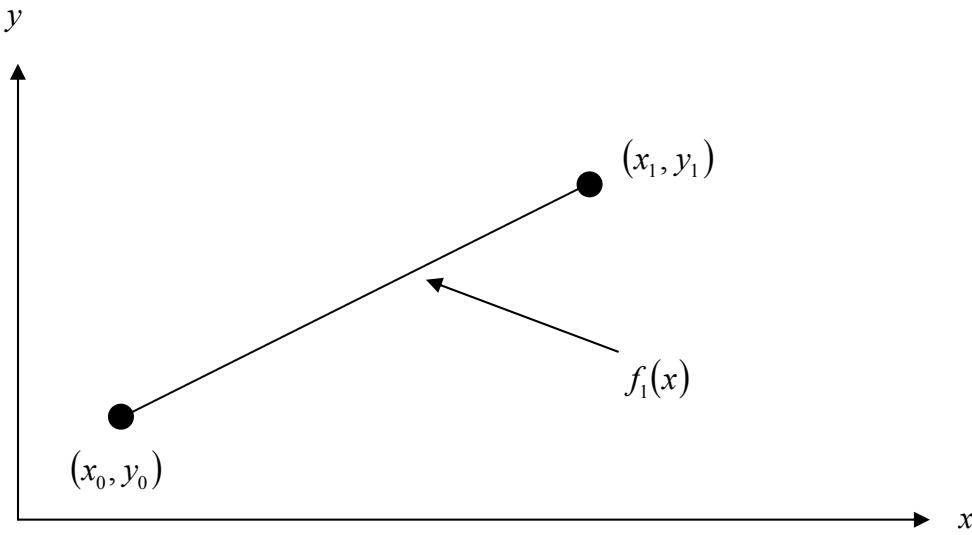
**Figure 2** Graph of velocity vs. time data for the rocket example.

Determine the value of the velocity at  $t = 16$  seconds using a first order Lagrange polynomial.

### Solution

For first order polynomial interpolation (also called linear interpolation), the velocity is given by

$$\begin{aligned} v(t) &= \sum_{i=0}^1 L_i(t)v(t_i) \\ &= L_0(t)v(t_0) + L_1(t)v(t_1) \end{aligned}$$



**Figure 3** Linear interpolation.

Since we want to find the velocity at  $t = 16$ , and we are using a first order polynomial, we need to choose the two data points that are closest to  $t = 16$  that also bracket  $t = 16$  to evaluate it. The two points are  $t_0 = 15$  and  $t_1 = 20$ .

Then

$$\begin{aligned} t_0 &= 15, \quad v(t_0) = 362.78 \\ t_1 &= 20, \quad v(t_1) = 517.35 \end{aligned}$$

gives

$$\begin{aligned} L_0(t) &= \prod_{\substack{j=0 \\ j \neq 0}}^1 \frac{t - t_j}{t_0 - t_j} \\ &= \frac{t - t_1}{t_0 - t_1} \\ L_1(t) &= \prod_{\substack{j=0 \\ j \neq 1}}^1 \frac{t - t_j}{t_1 - t_j} \\ &= \frac{t - t_0}{t_1 - t_0} \end{aligned}$$

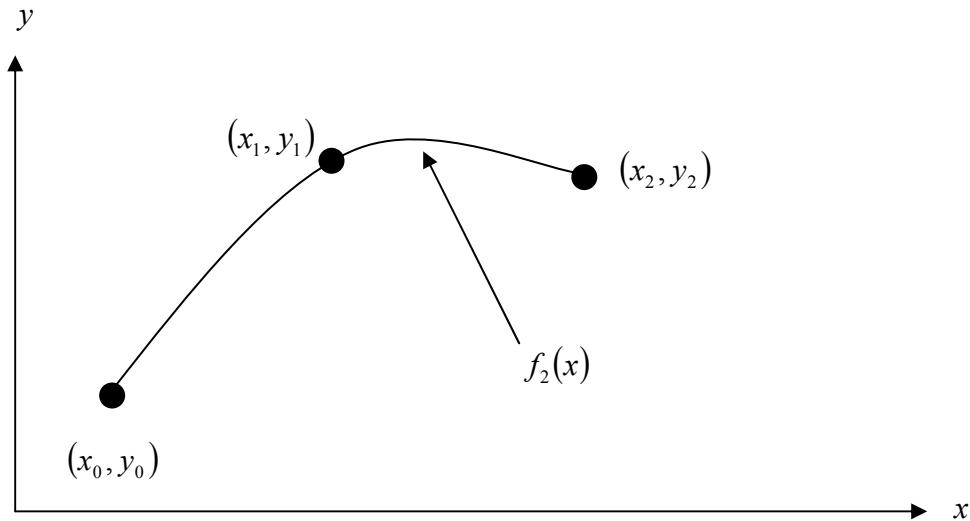
Hence

$$\begin{aligned} v(t) &= \frac{t - t_1}{t_0 - t_1} v(t_0) + \frac{t - t_0}{t_1 - t_0} v(t_1) \\ &= \frac{t - 20}{15 - 20} (362.78) + \frac{t - 15}{20 - 15} (517.35), \quad 15 \leq t \leq 20 \\ v(16) &= \frac{16 - 20}{15 - 20} (362.78) + \frac{16 - 15}{20 - 15} (517.35) \end{aligned}$$

$$\begin{aligned}
 &= 0.8(362.78) + 0.2(517.35) \\
 &= 393.69 \text{ m/s}
 \end{aligned}$$

You can see that  $L_0(t) = 0.8$  and  $L_1(t) = 0.2$  are like weightages given to the velocities at  $t = 15$  and  $t = 20$  to calculate the velocity at  $t = 16$ .

### Quadratic Interpolation



**Figure 4** Quadratic interpolation.

### **Example 2**

The upward velocity of a rocket is given as a function of time in Table 2.

**Table 2** Velocity as a function of time.

$t$ (s)	$v(t)$ (m/s)
0	0
10	227.04
15	362.78
20	517.35
22.5	602.97
30	901.67

- Determine the value of the velocity at  $t = 16$  seconds with second order polynomial interpolation using Lagrangian polynomial interpolation.
- Find the absolute relative approximate error for the second order polynomial approximation.

**Solution**

a) For second order polynomial interpolation (also called quadratic interpolation), the velocity is given by

$$\begin{aligned} v(t) &= \sum_{i=0}^2 L_i(t)v(t_i) \\ &= L_0(t)v(t_0) + L_1(t)v(t_1) + L_2(t)v(t_2) \end{aligned}$$

Since we want to find the velocity at  $t = 16$ , and we are using a second order polynomial, we need to choose the three data points that are closest to  $t = 16$  that also bracket  $t = 16$  to evaluate it. The three points are  $t_0 = 10$ ,  $t_1 = 15$ , and  $t_2 = 20$ .

Then

$$\begin{aligned} t_0 &= 10, \quad v(t_0) = 227.04 \\ t_1 &= 15, \quad v(t_1) = 362.78 \\ t_2 &= 20, \quad v(t_2) = 517.35 \end{aligned}$$

gives

$$\begin{aligned} L_0(t) &= \prod_{\substack{j=0 \\ j \neq 0}}^2 \frac{t - t_j}{t_0 - t_j} \\ &= \left( \frac{t - t_1}{t_0 - t_1} \right) \left( \frac{t - t_2}{t_0 - t_2} \right) \\ L_1(t) &= \prod_{\substack{j=0 \\ j \neq 1}}^2 \frac{t - t_j}{t_1 - t_j} \\ &= \left( \frac{t - t_0}{t_1 - t_0} \right) \left( \frac{t - t_2}{t_1 - t_2} \right) \\ L_2(t) &= \prod_{\substack{j=0 \\ j \neq 2}}^2 \frac{t - t_j}{t_2 - t_j} \\ &= \left( \frac{t - t_0}{t_2 - t_0} \right) \left( \frac{t - t_1}{t_2 - t_1} \right) \end{aligned}$$

Hence

$$\begin{aligned} v(t) &= \left( \frac{t - t_1}{t_0 - t_1} \right) \left( \frac{t - t_2}{t_0 - t_2} \right) v(t_0) + \left( \frac{t - t_0}{t_1 - t_0} \right) \left( \frac{t - t_2}{t_1 - t_2} \right) v(t_1) + \left( \frac{t - t_0}{t_2 - t_0} \right) \left( \frac{t - t_1}{t_2 - t_1} \right) v(t_2), \quad t_0 \leq t \leq t_2 \\ v(16) &= \frac{(16-15)(16-20)}{(10-15)(10-20)} (227.04) + \frac{(16-10)(16-20)}{(15-10)(15-20)} (362.78) \\ &\quad + \frac{(16-10)(16-15)}{(20-10)(20-15)} (517.35) \\ &= (-0.08)(227.04) + (0.96)(362.78) + (0.12)(517.35) \\ &= 392.19 \text{ m/s} \end{aligned}$$

b) The absolute relative approximate error  $|\epsilon_a|$  for the second order polynomial is calculated by considering the result of the first order polynomial (Example 1) as the previous approximation.

$$\begin{aligned} |\epsilon_a| &= \left| \frac{392.19 - 393.69}{392.19} \right| \times 100 \\ &= 0.38410\% \end{aligned}$$

### Example 3

The upward velocity of a rocket is given as a function of time in Table 3.

**Table 3** Velocity as a function of time

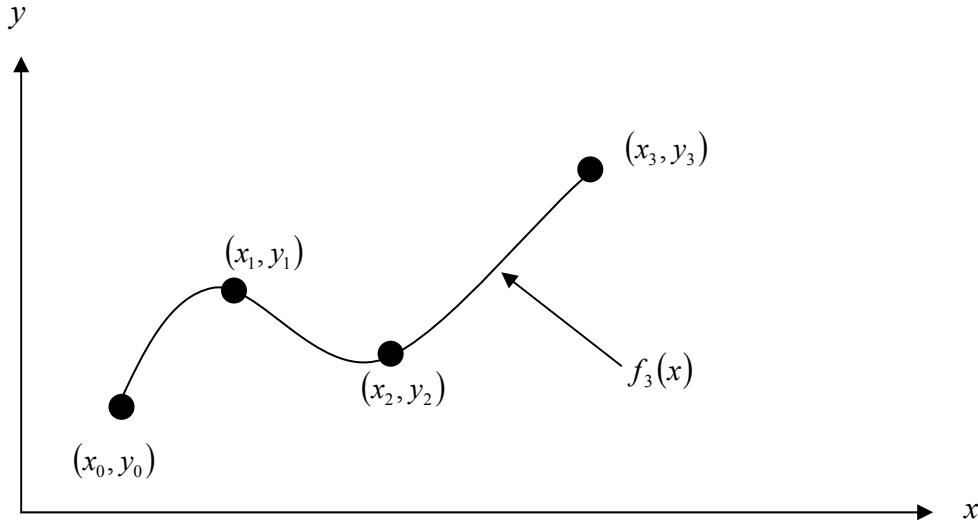
$t$ (s)	$v(t)$ (m/s)
0	0
10	227.04
15	362.78
20	517.35
22.5	602.97
30	901.67

- a) Determine the value of the velocity at  $t = 16$  seconds using third order Lagrangian polynomial interpolation.
- b) Find the absolute relative approximate error for the third order polynomial approximation.
- c) Using the third order polynomial interpolant for velocity, find the distance covered by the rocket from  $t = 11$  s to  $t = 16$  s.
- d) Using the third order polynomial interpolant for velocity, find the acceleration of the rocket at  $t = 16$  s.

### Solution

- a) For third order polynomial interpolation (also called cubic interpolation), the velocity is given by

$$\begin{aligned} v(t) &= \sum_{i=0}^3 L_i(t)v(t_i) \\ &= L_0(t)v(t_0) + L_1(t)v(t_1) + L_2(t)v(t_2) + L_3(t)v(t_3) \end{aligned}$$



**Figure 5** Cubic interpolation.

Since we want to find the velocity at  $t = 16$ , and we are using a third order polynomial, we need to choose the four data points closest to  $t = 16$  that also bracket  $t = 16$  to evaluate it. The four points are  $t_0 = 10$ ,  $t_1 = 15$ ,  $t_2 = 20$  and  $t_3 = 22.5$ .

Then

$$\begin{aligned} t_0 &= 10, \quad v(t_0) = 227.04 \\ t_1 &= 15, \quad v(t_1) = 362.78 \\ t_2 &= 20, \quad v(t_2) = 517.35 \\ t_3 &= 22.5, \quad v(t_3) = 602.97 \end{aligned}$$

gives

$$\begin{aligned} L_0(t) &= \prod_{\substack{j=0 \\ j \neq 0}}^3 \frac{t - t_j}{t_0 - t_j} \\ &= \left( \frac{t - t_1}{t_0 - t_1} \right) \left( \frac{t - t_2}{t_0 - t_2} \right) \left( \frac{t - t_3}{t_0 - t_3} \right) \\ L_1(t) &= \prod_{\substack{j=0 \\ j \neq 1}}^3 \frac{t - t_j}{t_1 - t_j} \\ &= \left( \frac{t - t_0}{t_1 - t_0} \right) \left( \frac{t - t_2}{t_1 - t_2} \right) \left( \frac{t - t_3}{t_1 - t_3} \right) \\ L_2(t) &= \prod_{\substack{j=0 \\ j \neq 2}}^3 \frac{t - t_j}{t_2 - t_j} \\ &= \left( \frac{t - t_0}{t_2 - t_0} \right) \left( \frac{t - t_1}{t_2 - t_1} \right) \left( \frac{t - t_3}{t_2 - t_3} \right) \end{aligned}$$

$$\begin{aligned}
L_3(t) &= \prod_{\substack{j=0 \\ j \neq 3}}^3 \frac{t - t_j}{t_3 - t_j} \\
&= \left( \frac{t - t_0}{t_3 - t_0} \right) \left( \frac{t - t_1}{t_3 - t_1} \right) \left( \frac{t - t_2}{t_3 - t_2} \right)
\end{aligned}$$

Hence

$$\begin{aligned}
v(t) &= \left( \frac{t - t_1}{t_0 - t_1} \right) \left( \frac{t - t_2}{t_0 - t_2} \right) \left( \frac{t - t_3}{t_0 - t_3} \right) v(t_0) + \left( \frac{t - t_0}{t_1 - t_0} \right) \left( \frac{t - t_2}{t_1 - t_2} \right) \left( \frac{t - t_3}{t_1 - t_3} \right) v(t_1) \\
&\quad + \left( \frac{t - t_0}{t_2 - t_0} \right) \left( \frac{t - t_1}{t_2 - t_1} \right) \left( \frac{t - t_3}{t_2 - t_3} \right) v(t_2) + \left( \frac{t - t_0}{t_3 - t_0} \right) \left( \frac{t - t_1}{t_3 - t_1} \right) \left( \frac{t - t_2}{t_3 - t_2} \right) v(t_3), \quad t_0 \leq t \leq t_3 \\
v(16) &= \frac{(16-15)(16-20)(16-22.5)}{(10-15)(10-20)(10-22.5)} (227.04) + \frac{(16-10)(16-20)(16-22.5)}{(15-10)(15-20)(15-22.5)} (362.78) \\
&\quad + \frac{(16-10)(16-15)(16-22.5)}{(20-10)(20-15)(20-22.5)} (517.35) \\
&\quad + \frac{(16-10)(16-15)(16-20)}{(22.5-10)(22.5-15)(22.5-20)} (602.97) \\
&= (-0.0416)(227.04) + (0.832)(362.78) + (0.312)(517.35) + (-0.1024)(602.97) \\
&= 392.06 \text{ m/s}
\end{aligned}$$

b) The absolute percentage relative approximate error,  $|e_a|$  for the value obtained for  $v(16)$  can be obtained by comparing the result with that obtained using the second order polynomial (Example 2)

$$\begin{aligned}
|e_a| &= \left| \frac{392.06 - 392.19}{392.06} \right| \times 100 \\
&= 0.033269\%
\end{aligned}$$

c) The distance covered by the rocket between  $t = 11 \text{ s}$  to  $t = 16 \text{ s}$  can be calculated from the interpolating polynomial as

$$\begin{aligned}
v(t) &= \frac{(t-15)(t-20)(t-22.5)}{(10-15)(10-20)(10-22.5)} (227.04) + \frac{(t-10)(t-20)(t-22.5)}{(15-10)(15-20)(15-22.5)} (362.78) \\
&\quad + \frac{(t-10)(t-15)(t-22.5)}{(20-10)(20-15)(20-22.5)} (517.35) \\
&\quad + \frac{(t-10)(t-15)(t-20)}{(22.5-10)(22.5-15)(22.5-20)} (602.97), \quad 10 \leq t \leq 22.5 \\
&= \frac{(t^2 - 35t + 300)(t - 22.5)}{(-5)(-10)(-12.5)} (227.04) + \frac{(t^2 - 30t + 200)(t - 22.5)}{(5)(-5)(-7.5)} (362.78) \\
&\quad + \frac{(t^2 - 25t + 150)(t - 22.5)}{(10)(5)(-2.5)} (517.35) + \frac{(t^2 - 25t + 150)(t - 20)}{(12.5)(7.5)(2.5)} (602.97)
\end{aligned}$$

$$\begin{aligned}
&= (t^3 - 57.5t^2 + 1087.5t - 6750)(-0.36326) + (t^3 - 52.5t^2 + 875t - 4500)(1.9348) \\
&\quad + (t^3 - 47.5t^2 + 712.5t - 3375)(-4.1388) + (t^3 - 45t^2 + 650t - 3000)(2.5727) \\
&= -4.245 + 21.265t + 0.13195t^2 + 0.00544t^3, \quad 10 \leq t \leq 22.5
\end{aligned}$$

Note that the polynomial is valid between  $t = 10$  and  $t = 22.5$  and hence includes the limits of  $t = 11$  and  $t = 16$ .

So

$$\begin{aligned}
s(16) - s(11) &= \int_{11}^{16} v(t) dt \\
&= \int_{11}^{16} (-4.245 + 21.265t + 0.13195t^2 + 0.00544t^3) dt \\
&= \left[ -4.245t + 21.265 \frac{t^2}{2} + 0.13195 \frac{t^3}{3} + 0.00544 \frac{t^4}{4} \right]_{11}^{16} \\
&= 1605 \text{ m}
\end{aligned}$$

d) The acceleration at  $t = 16$  is given by

$$a(16) = \frac{d}{dt} v(t) \Big|_{t=16}$$

Given that

$$v(t) = -4.245 + 21.265t + 0.13195t^2 + 0.00544t^3, \quad 10 \leq t \leq 22.5$$

$$\begin{aligned}
a(t) &= \frac{d}{dt} v(t) \\
&= \frac{d}{dt} (-4.245 + 21.265t + 0.13195t^2 + 0.00544t^3) \\
&= 21.265 + 0.26390t + 0.01632t^2, \quad 10 \leq t \leq 22.5 \\
a(16) &= 21.265 + 0.26390(16) + 0.01632(16)^2 \\
&= 29.665 \text{ m/s}^2
\end{aligned}$$

Note: There is no need to get the simplified third order polynomial expression to conduct the differentiation. An expression of the form

$$L_0(t) = \left( \frac{t - t_1}{t_0 - t_1} \right) \left( \frac{t - t_2}{t_0 - t_2} \right) \left( \frac{t - t_3}{t_0 - t_3} \right)$$

gives the derivative without expansion as

$$\frac{d}{dt} (L_0(t)) = \left( \frac{t - t_1}{t_0 - t_1} \right) \left( \frac{t - t_2}{t_0 - t_2} \right) + \left( \frac{t - t_2}{t_0 - t_2} \right) \left( \frac{t - t_3}{t_0 - t_3} \right) + \left( \frac{t - t_3}{t_0 - t_3} \right) \left( \frac{t - t_1}{t_0 - t_1} \right)$$

---

**INTERPOLATION**

---

Topic	Lagrange Interpolation
Summary	Textbook notes on the Lagrangian method of interpolation
Major	General Engineering
Authors	Autar Kaw, Michael Keteltas
Last Revised	December 23, 2009
Web Site	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

---

# **Chapter 05.05**

## **Spline Method of Interpolation**

*After reading this chapter, you should be able to:*

1. *interpolate data using spline interpolation, and*
2. *understand why spline interpolation is important.*

### **What is interpolation?**

Many times, data is given only at discrete points such as  $(x_0, y_0)$ ,  $(x_1, y_1)$ , ...,  $(x_{n-1}, y_{n-1})$ ,  $(x_n, y_n)$ . So, how then does one find the value of  $y$  at any other value of  $x$ ? Well, a continuous function  $f(x)$  may be used to represent the  $n+1$  data values with  $f(x)$  passing through the  $n+1$  points (Figure 1). Then one can find the value of  $y$  at any other value of  $x$ . This is called *interpolation*.

Of course, if  $x$  falls outside the range of  $x$  for which the data is given, it is no longer interpolation but instead is called *extrapolation*.

So what kind of function  $f(x)$  should one choose? A polynomial is a common choice for an interpolating function because polynomials are easy to

- (A) evaluate,
- (B) differentiate, and
- (C) integrate

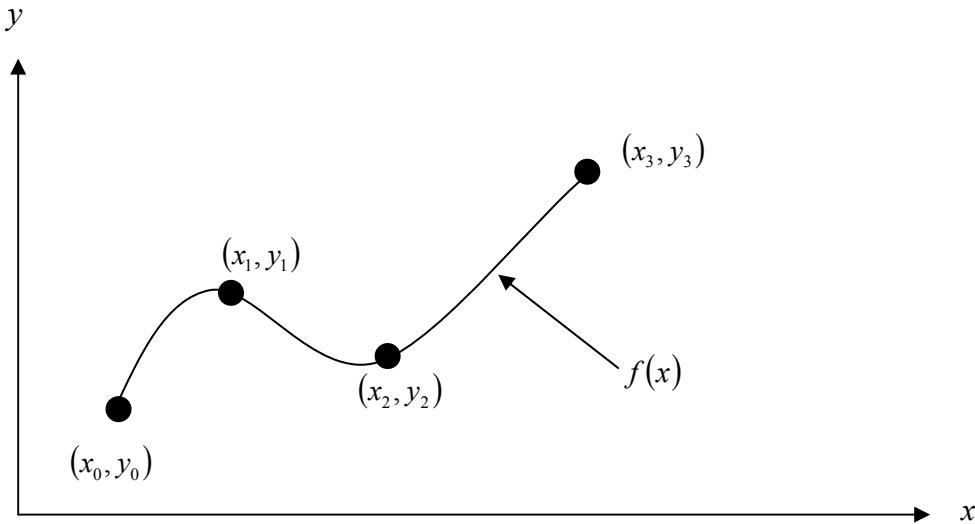
relative to other choices such as a trigonometric and exponential series.

Polynomial interpolation involves finding a polynomial of order  $n$  that passes through the  $n+1$  points. Several methods to obtain such a polynomial include the direct method, Newton's divided difference polynomial method and the Lagrangian interpolation method.

So is the spline method yet another method of obtaining this  $n^{\text{th}}$  order polynomial. .... NO! Actually, when  $n$  becomes large, in many cases, one may get oscillatory behavior in the resulting polynomial. This was shown by Runge when he interpolated data based on a simple function of

$$y = \frac{1}{1 + 25x^2}$$

on an interval of  $[-1, 1]$ . For example, take six equidistantly spaced points in  $[-1, 1]$  and find  $y$  at these points as given in Table 1.



**Figure 1** Interpolation of discrete data.

**Table 1** Six equidistantly spaced points in  $[-1, 1]$ .

$x$	$y = \frac{1}{1 + 25x^2}$
-1.0	0.038461
-0.6	0.1
-0.2	0.5
0.2	0.5
0.6	0.1
1.0	0.038461

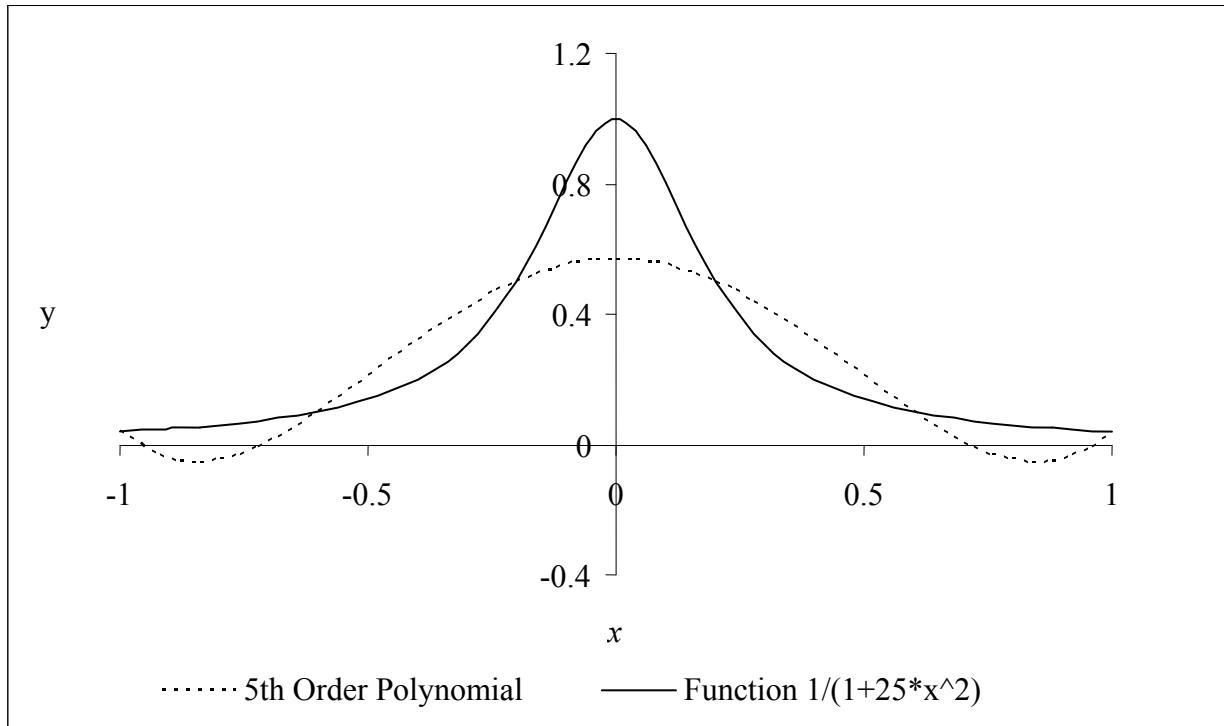
Now through these six points, one can pass a fifth order polynomial

$$f_5(x) = 3.1378 \times 10^{-11} x^5 + 1.2019 x^4 - 3.3651 \times 10^{-11} x^3 - 1.7308 x^2 + 1.0004 \times 10^{-11} x + 5.6731 \times 10^{-1},$$

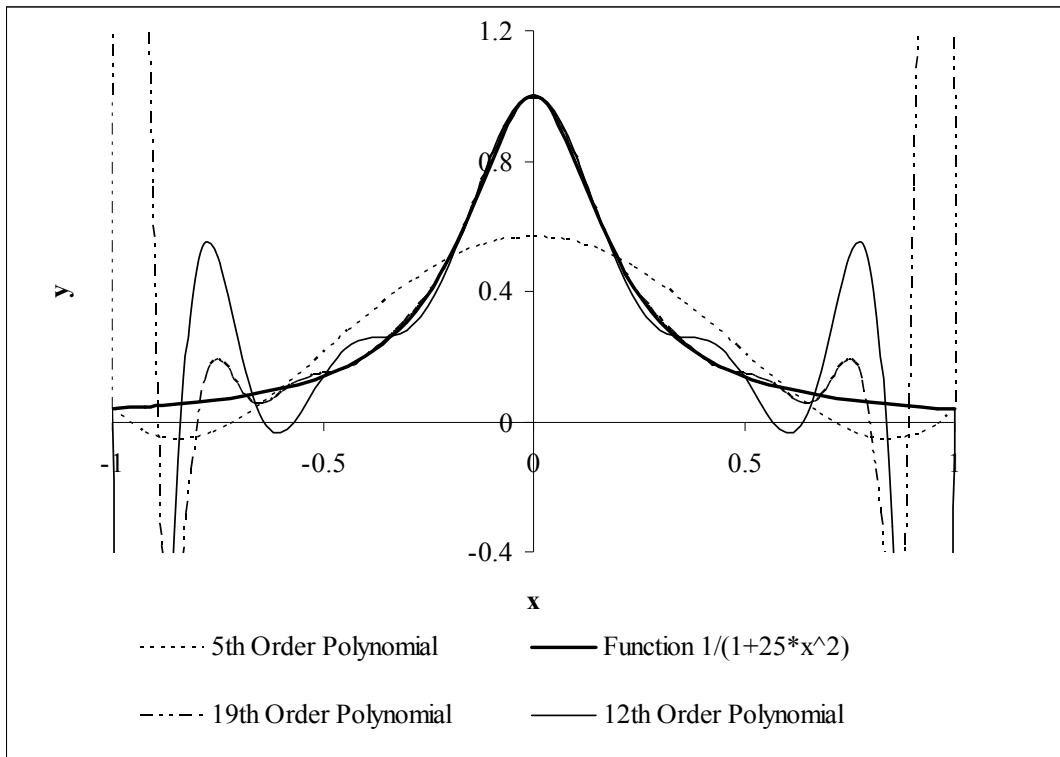
$$-1 \leq x \leq 1$$

through the six data points. On plotting the fifth order polynomial (Figure 2) and the original function, one can see that the two do not match well. One may consider choosing more points in the interval  $[-1, 1]$  to get a better match, but it diverges even more (see Figure 3), where 20 equidistant points were chosen in the interval  $[-1, 1]$  to draw a 19th order polynomial. In fact, Runge found that as the order of the polynomial becomes infinite, the polynomial diverges in the interval of  $-1 < x < -0.726$  and  $0.726 < x < 1$ .

So what is the answer to using information from more data points, but at the same time keeping the function true to the data behavior? The answer is in spline interpolation. The most common spline interpolations used are linear, quadratic, and cubic splines.



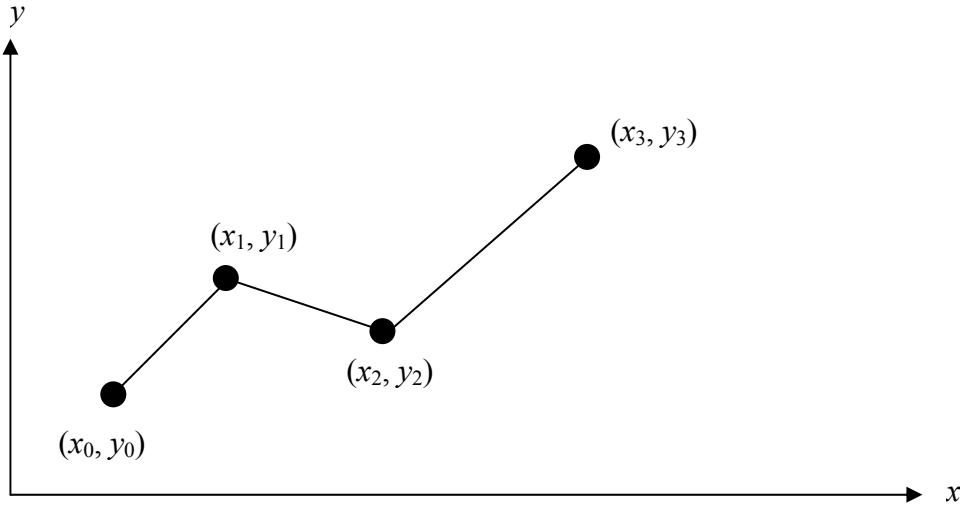
**Figure 2** 5th order polynomial interpolation with six equidistant points.



**Figure 3** Higher order polynomial interpolation is a bad idea.

### Linear Spline Interpolation

Given  $(x_0, y_0), (x_1, y_1), \dots, (x_{n-1}, y_{n-1}), (x_n, y_n)$ , fit linear splines (Figure 4) to the data. This simply involves forming the consecutive data through straight lines. So if the above data is given in an ascending order, the linear splines are given by  $y_i = f(x_i)$ .



**Figure 4** Linear splines.

$$\begin{aligned}
 f(x) &= f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_0), & x_0 \leq x \leq x_1 \\
 &= f(x_1) + \frac{f(x_2) - f(x_1)}{x_2 - x_1}(x - x_1), & x_1 \leq x \leq x_2 \\
 &\quad \vdots \\
 &= f(x_{n-1}) + \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}(x - x_{n-1}), & x_{n-1} \leq x \leq x_n
 \end{aligned}$$

Note the terms of

$$\frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}}$$

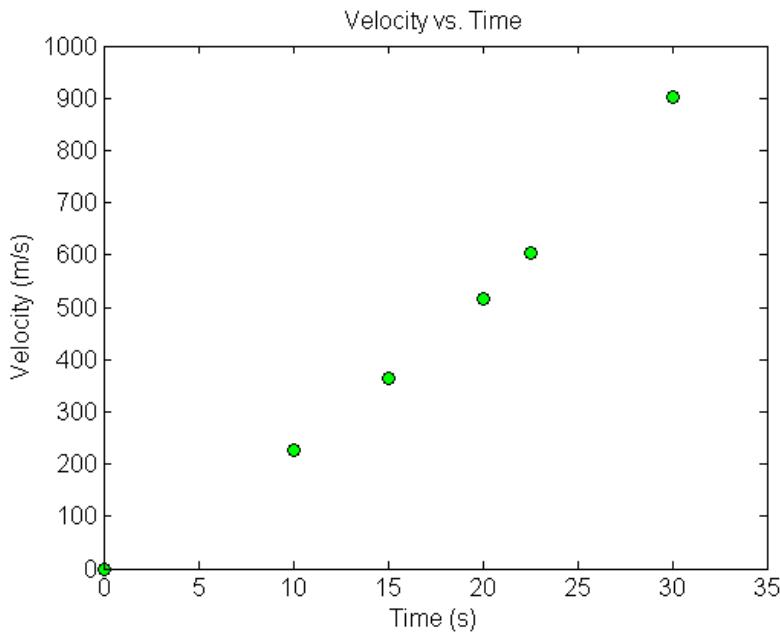
in the above function are simply slopes between  $x_{i-1}$  and  $x_i$ .

**Example 1**

The upward velocity of a rocket is given as a function of time in Table 2 (Figure 5).

**Table 2** Velocity as a function of time.

$t$ (s)	$v(t)$ (m/s)
0	0
10	227.04
15	362.78
20	517.35
22.5	602.97
30	901.67

**Figure 5** Graph of velocity vs. time data for the rocket example.

Determine the value of the velocity at  $t = 16$  seconds using linear splines.

**Solution**

Since we want to evaluate the velocity at  $t = 16$ , and we are using linear splines, we need to choose the two data points closest to  $t = 16$  that also bracket  $t = 16$  to evaluate it. The two points are  $t_0 = 15$  and  $t_1 = 20$ .

Then

$$t_0 = 15, \quad v(t_0) = 362.78$$

$$t_1 = 20, \quad v(t_1) = 517.35$$

gives

$$\begin{aligned}
 v(t) &= v(t_0) + \frac{v(t_1) - v(t_0)}{t_1 - t_0}(t - t_0) \\
 &= 362.78 + \frac{517.35 - 362.78}{20 - 15}(t - 15) \\
 &= 362.78 + 30.913(t - 15), \quad 15 \leq t \leq 20
 \end{aligned}$$

At  $t = 16$ ,

$$\begin{aligned}
 v(16) &= 362.78 + 30.913(16 - 15) \\
 &= 393.7 \text{ m/s}
 \end{aligned}$$

Linear spline interpolation is no different from linear polynomial interpolation. Linear splines still use data only from the two consecutive data points. Also at the interior points of the data, the slope changes abruptly. This means that the first derivative is not continuous at these points. So how do we improve on this? We can do so by using quadratic splines.

### Quadratic Splines

In these splines, a quadratic polynomial approximates the data between two consecutive data points. Given  $(x_0, y_0), (x_1, y_1), \dots, (x_{n-1}, y_{n-1}), (x_n, y_n)$ , fit quadratic splines through the data. The splines are given by

$$\begin{aligned}
 f(x) &= a_1 x^2 + b_1 x + c_1, & x_0 \leq x \leq x_1 \\
 &= a_2 x^2 + b_2 x + c_2, & x_1 \leq x \leq x_2 \\
 &\vdots \\
 &\vdots \\
 &= a_n x^2 + b_n x + c_n, & x_{n-1} \leq x \leq x_n
 \end{aligned}$$

So how does one find the coefficients of these quadratic splines? There are  $3n$  such coefficients

$$\begin{aligned}
 a_i, \quad i &= 1, 2, \dots, n \\
 b_i, \quad i &= 1, 2, \dots, n \\
 c_i, \quad i &= 1, 2, \dots, n
 \end{aligned}$$

To find  $3n$  unknowns, one needs to set up  $3n$  equations and then simultaneously solve them. These  $3n$  equations are found as follows.

1. Each quadratic spline goes through two consecutive data points

$$\begin{aligned}
 a_1 x_0^2 + b_1 x_0 + c_1 &= f(x_0) \\
 a_1 x_1^2 + b_1 x_1 + c_1 &= f(x_1)
 \end{aligned}$$

$\vdots$

$$\begin{aligned}
 a_i x_{i-1}^2 + b_i x_{i-1} + c_i &= f(x_{i-1}) \\
 a_i x_i^2 + b_i x_i + c_i &= f(x_i)
 \end{aligned}$$

$\vdots$

$$\begin{aligned} a_n x_{n-1}^2 + b_n x_{n-1} + c_n &= f(x_{n-1}) \\ a_n x_n^2 + b_n x_n + c_n &= f(x_n) \end{aligned}$$

This condition gives  $2n$  equations as there are  $n$  quadratic splines going through two consecutive data points.

2. The first derivatives of two quadratic splines are continuous at the interior points. For example, the derivative of the first spline

$$a_1 x^2 + b_1 x + c_1$$

is

$$2a_1 x + b_1$$

The derivative of the second spline

$$a_2 x^2 + b_2 x + c_2$$

is

$$2a_2 x + b_2$$

and the two are equal at  $x = x_1$  giving

$$2a_1 x_1 + b_1 = 2a_2 x_1 + b_2$$

$$2a_1 x_1 + b_1 - 2a_2 x_1 - b_2 = 0$$

Similarly at the other interior points,

$$2a_2 x_2 + b_2 - 2a_3 x_2 - b_3 = 0$$

$$2a_i x_i + b_i - 2a_{i+1} x_i - b_{i+1} = 0$$

$$2a_{n-1} x_{n-1} + b_{n-1} - 2a_n x_{n-1} - b_n = 0$$

Since there are  $(n-1)$  interior points, we have  $(n-1)$  such equations. So far, the total number of equations is  $(2n) + (n-1) = (3n-1)$  equations. We still then need one more equation.

We can assume that the first spline is linear, that is

$$a_1 = 0$$

This gives us  $3n$  equations and  $3n$  unknowns. These can be solved by a number of techniques used to solve simultaneous linear equations.

### Example 2

The upward velocity of a rocket is given as a function of time as

**Table 3** Velocity as a function of time.

$t$ (s)	$v(t)$ (m/s)
0	0
10	227.04
15	362.78
20	517.35
22.5	602.97
30	901.67

- a) Determine the value of the velocity at  $t = 16$  seconds using quadratic splines.
- b) Using the quadratic splines as velocity functions, find the distance covered by the rocket from  $t = 11\text{s}$  to  $t = 16\text{s}$ .
- c) Using the quadratic splines as velocity functions, find the acceleration of the rocket at  $t = 16\text{s}$ .

### Solution

a) Since there are six data points, five quadratic splines pass through them.

$$\begin{aligned} v(t) &= a_1 t^2 + b_1 t + c_1, \quad 0 \leq t \leq 10 \\ &= a_2 t^2 + b_2 t + c_2, \quad 10 \leq t \leq 15 \\ &= a_3 t^2 + b_3 t + c_3, \quad 15 \leq t \leq 20 \\ &= a_4 t^2 + b_4 t + c_4, \quad 20 \leq t \leq 22.5 \\ &= a_5 t^2 + b_5 t + c_5, \quad 22.5 \leq t \leq 30 \end{aligned}$$

The equations are found as follows.

1. Each quadratic spline passes through two consecutive data points.

$a_1 t^2 + b_1 t + c_1$  passes through  $t = 0$  and  $t = 10$ .

$$a_1(0)^2 + b_1(0) + c_1 = 0 \tag{1}$$

$$a_1(10)^2 + b_1(10) + c_1 = 227.04 \tag{2}$$

$a_2 t^2 + b_2 t + c_2$  passes through  $t = 10$  and  $t = 15$ .

$$a_2(10)^2 + b_2(10) + c_2 = 227.04 \tag{3}$$

$$a_2(15)^2 + b_2(15) + c_2 = 362.78 \tag{4}$$

$a_3 t^2 + b_3 t + c_3$  passes through  $t = 15$  and  $t = 20$ .

$$a_3(15)^2 + b_3(15) + c_3 = 362.78 \tag{5}$$

$$a_3(20)^2 + b_3(20) + c_3 = 517.35 \tag{6}$$

$a_4 t^2 + b_4 t + c_4$  passes through  $t = 20$  and  $t = 22.5$ .

$$a_4(20)^2 + b_4(20) + c_4 = 517.35 \tag{7}$$

$$a_4(22.5)^2 + b_4(22.5) + c_4 = 602.97 \tag{8}$$

$a_5 t^2 + b_5 t + c_5$  passes through  $t = 22.5$  and  $t = 30$ .

$$a_5(22.5)^2 + b_5(22.5) + c_5 = 602.97 \quad (9)$$

$$a_5(30)^2 + b_5(30) + c_5 = 901.67 \quad (10)$$

2. Quadratic splines have continuous derivatives at the interior data points.

At  $t = 10$

$$2a_1(10) + b_1 - 2a_2(10) - b_2 = 0 \quad (11)$$

At  $t = 15$

$$2a_2(15) + b_2 - 2a_3(15) - b_3 = 0 \quad (12)$$

At  $t = 20$

$$2a_3(20) + b_3 - 2a_4(20) - b_4 = 0 \quad (13)$$

At  $t = 22.5$

$$2a_4(22.5) + b_4 - 2a_5(22.5) - b_5 = 0 \quad (14)$$

3. Assuming the first spline  $a_1 t^2 + b_1 t + c_1$  is linear,

$$a_1 = 0 \quad (15)$$

Combining Equation (1) –(15) in matrix form gives

$$\left[ \begin{array}{cccccccccccccc|c} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & a_1 \\ 100 & 10 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & b_1 \\ 0 & 0 & 0 & 100 & 10 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & c_1 \\ 0 & 0 & 0 & 225 & 15 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & a_2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 225 & 15 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & b_2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 400 & 20 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & c_2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 400 & 20 & 1 & 0 & 0 & 0 & a_3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 506.25 & 22.5 & 1 & 0 & 0 & b_3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 506.25 & 22.5 & 1 & c_3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & a_4 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & b_4 \\ 20 & 1 & 0 & -20 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & c_4 \\ 0 & 0 & 0 & 30 & 1 & 0 & -30 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & a_5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 40 & 1 & 0 & -40 & -1 & 0 & 0 & 0 & 0 & b_5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 45 & 1 & 0 & -45 & -1 & 0 & c_5 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \right] = \left[ \begin{array}{c} 0 \\ 227.04 \\ 227.04 \\ 362.78 \\ 362.78 \\ 517.35 \\ 517.35 \\ 602.97 \\ 602.97 \\ 901.67 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{array} \right]$$

Solving the above 15 equations give the 15 unknowns as

$i$	$a_i$	$b_i$	$c_i$
1	0	22.704	0
2	0.8888	4.928	88.88
3	-0.1356	35.66	-141.61
4	1.6048	-33.956	554.55
5	0.20889	28.86	-152.13

Therefore, the splines are given by

$$\begin{aligned}
 v(t) &= 22.704t, & 0 \leq t \leq 10 \\
 &= 0.8888t^2 + 4.928t + 88.88, & 10 \leq t \leq 15 \\
 &= -0.1356t^2 + 35.66t - 141.61, & 15 \leq t \leq 20 \\
 &= 1.6048t^2 - 33.956t + 554.55, & 20 \leq t \leq 22.5 \\
 &= 0.20889t^2 + 28.86t - 152.13, & 22.5 \leq t \leq 30
 \end{aligned}$$

At  $t = 16$  s

$$\begin{aligned}
 v(16) &= -0.1356(16)^2 + 35.66(16) - 141.61 \\
 &= 394.24 \text{ m/s}
 \end{aligned}$$

b) The distance covered by the rocket between 11 and 16 seconds can be calculated as

$$s(16) - s(11) = \int_{11}^{16} v(t) dt$$

But since the splines are valid over different ranges, we need to break the integral accordingly as

$$\begin{aligned}
 v(t) &= 0.8888t^2 + 4.928t + 88.88, \quad 10 \leq t \leq 15 \\
 &= -0.1356t^2 + 35.66t - 141.61, \quad 15 \leq t \leq 20 \\
 \int_{11}^{16} v(t) dt &= \int_{11}^{15} v(t) dt + \int_{15}^{16} v(t) dt \\
 s(16) - s(11) &= \int_{11}^{15} (0.8888t^2 + 4.928t + 88.88) dt + \int_{15}^{16} (-0.1356t^2 + 35.66t - 141.61) dt \\
 &= \left[ 0.8888 \frac{t^3}{3} + 4.928 \frac{t^2}{2} + 88.88t \right]_{11}^{15} \\
 &\quad + \left[ -0.1356 \frac{t^3}{3} + 35.66 \frac{t^2}{2} - 141.61t \right]_{15}^{16} \\
 &= 1217.35 + 378.53 \\
 &= 1595.9 \text{ m}
 \end{aligned}$$

c) What is the acceleration at  $t = 16$ ?

$$\begin{aligned}
 a(16) &= \frac{d}{dt} v(t) \Big|_{t=16} \\
 a(t) &= \frac{d}{dt} v(t) = \frac{d}{dt} (-0.1356t^2 + 35.66t - 141.61) \\
 &= -0.2712t + 35.66, \quad 15 \leq t \leq 20 \\
 a(16) &= -0.2712(16) + 35.66 \\
 &= 31.321 \text{ m/s}^2
 \end{aligned}$$

---

**INTERPOLATION**

---

Topic Spline Method of Interpolation  
Summary Textbook notes on the spline method of interpolation  
Major General Engineering  
Authors Autar Kaw, Michael Keteltas  
Date December 23, 2009  

---

Web Site <http://numericalmethods.eng.usf.edu>

---

# **Chapter 05.06**

## **Extrapolation is a Bad Idea**

*After reading this chapter, you should be able to:*

1. understand why using extrapolation can be a bad idea.

### **Example**

(Due to certain reasons, this student wishes to remain anonymous.)

This takes place in Summer Session B – July 2001

**Student:** “Hey, Dr. Kaw! Look at this cool new cell phone I just got!”

**Kaw:** “That’s nice. It better not ring in my class or it’s mine.”

**Student:** “What would you think about getting stock in this company?”

**Kaw:** “What company is that?”

**Student:** “WorldCom! They’re the world’s leading global data and internet company.”

**Kaw:** “So?”

**Student:** “They’ve just closed the deal today to merge with Intermedia Communications, based right here in Tampa!”

**Kaw:** “Yeah, and ...?”

**Student:** “The stock’s booming! It’s at \$14.11 per share and promised to go only one way—up! We’ll be millionaires if we invest now!”

**Kaw:** “You might not want to assume their stock will keep rising ... besides, I’m skeptical of their success. I don’t want you putting yourself in financial ‘jeopardy!’ over some silly extrapolation. Take a look at these NASDAQ composite numbers (Table 1)”

**Student:** “That’s only up to two years ago ...”

**Kaw:** “That’s right. Looking at this data, don’t you think you should’ve invested back then?”

**Student:** “Well, didn’t the composite drop after that?”

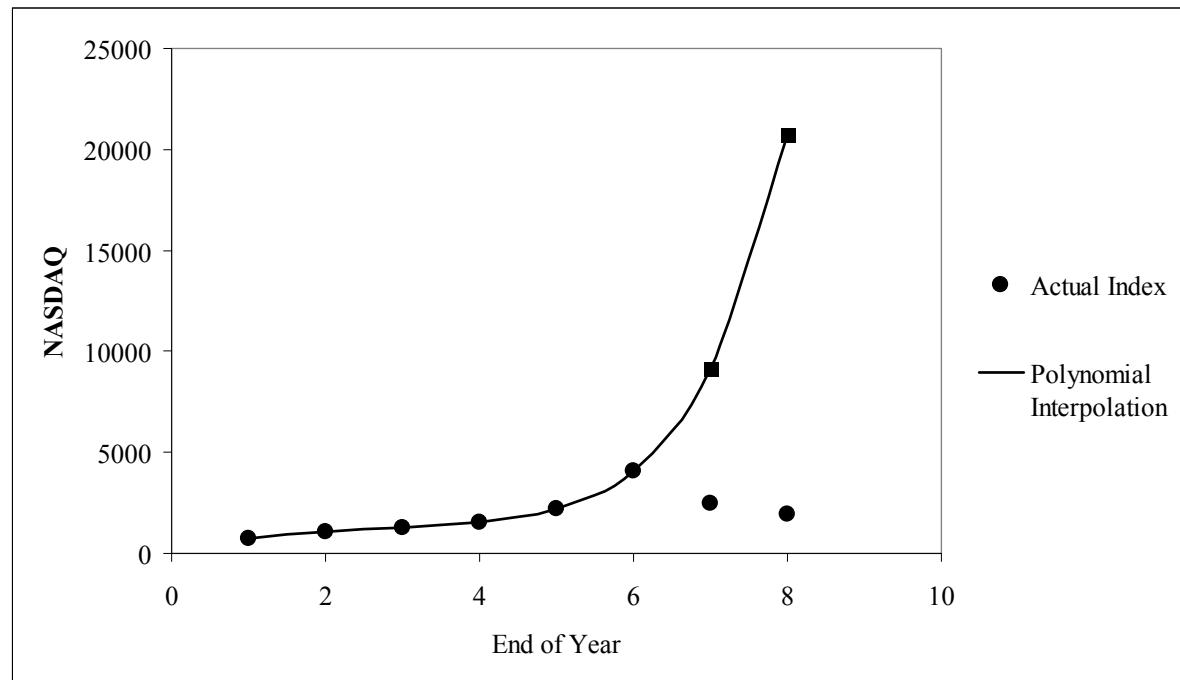
**Kaw:** “Right again, but look what you would’ve hoped for if you had depended on that trend continuing (Figure 1).”

**Student:** “So you’re saying that ...?”

**Kaw:** “You should seldom depend on extrapolation as a source of approximation! Just take a look at how wrong you would have been (Table 2).”

**Table 1.** End of year NASDAQ composite data

End of year <sup>1</sup>	NASDAQ
1	751.96
2	1052.13
3	1291.03
4	1570.35
5	2192.69
6	4069.31

**Figure 1** Data from 1994 to 1999 extrapolated to yield results for 2000 and 2001 using polynomial extrapolation.

---

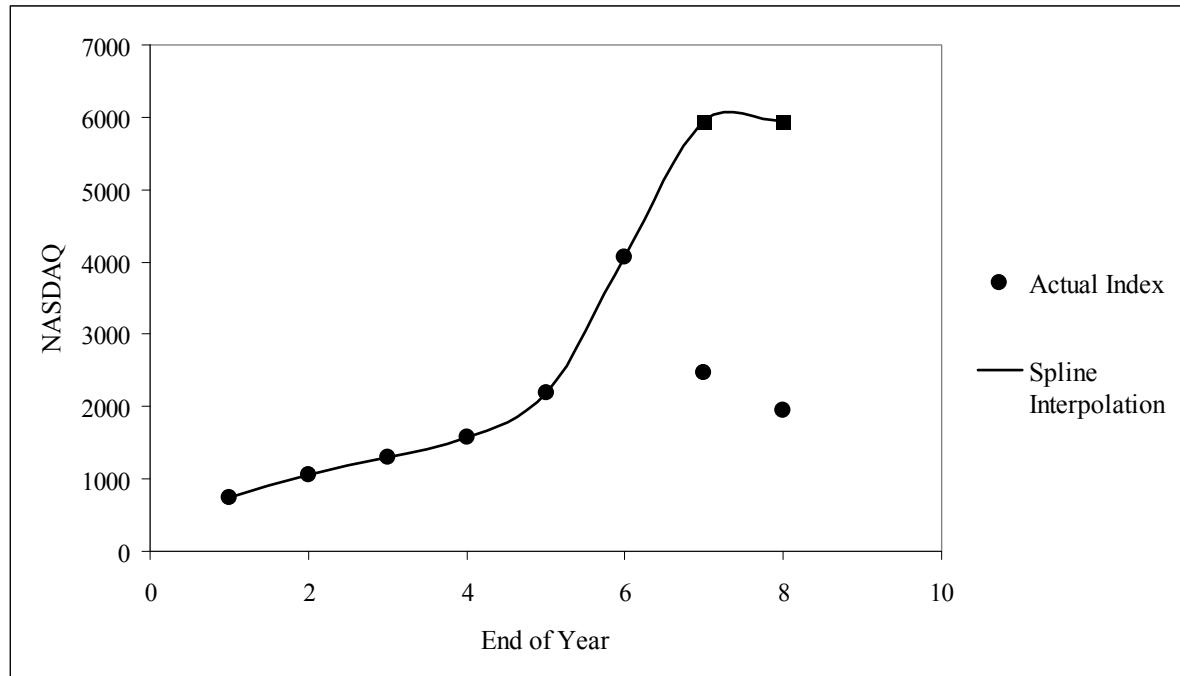
<sup>1</sup> Range of years actually between 1994 (Year 1) and 1999 (Year 9). Numbers start from 1 to avoid round-off errors and near singularity in matrix calculations.

**Table 2** Absolute relative true error of polynomial interpolation.

End of Year	Actual	Fifth order polynomial interpolation	Absolute relative true error
2000	2471	9128	269.47 %
2001	1950	20720	962.36 %

**Student:** “Now wait a sec! I wouldn’t have been quite that wrong. What if I had used cubic splines instead of a fifth order interpolant?”

**Kaw:** “Let’s find out.”

**Figure 2** Data from 1994 to 1999 extrapolated to yield results for 2000 and 2001 using cubic spline interpolation.**Table 3** Absolute relative true error of cubic spline interpolation

End of Year	Actual	Cubic spline interpolation	Absolute relative true error
2000	2471	5945.9	140.63 %
2001	1950	5947.4	204.99 %

**Student:** “There you go. That didn’t take so long (Figure 2 and Table 3).”

**Kaw:** “Well, let’s think about what this data means. If you had gone ahead and invested, thinking your projected yield would follow the spline, you would have only been 205% (Table 3) wrong, as opposed to being 962% (Table 2) wrong by following the polynomial. That’s not so bad, is it?”

**Student:** “Okay, you’ve got a point. Maybe I’ll hold off on being an investor and just use the cell phone.”

**Kaw:** “You’ve got a point, too—you’re brighter than you look … that is if you turn off the phone before coming to class.”

\* \* \* \*

<One year later … July 2002>

**Student:** “Hey, Dr. Kaw! Whatcha got for me today?”

**Kaw:** “The Computational Methods students just took their interpolation test today, so here you go. <hands stack of tests to student> Time to grade them!”

**Student:** <Grunt!> “That’s a lot of paper! Boy, interpolation … learned that a while ago.”

**Kaw:** “You haven’t forgotten my lesson to you about not extrapolating, have you?”

**Student:** “Of course not! Haven’t you seen the news? WorldCom just closed down 93% from 83¢ on June 25 to 6¢ per share! They’ve had to recalculate their earnings, so your skepticism really must’ve spread. Did you have an “in” on what was going on?”

**Kaw:** “Oh, of course not. I’m just an ignorant numerical methods professor.”

---

## INTERPOLATION

---

Topic	Extrapolation is a bad idea
Summary	Textbook notes on errors that can occur when extrapolating data
Major	All majors of engineering
Authors	Autar Kaw
Last Revised	December 23, 2009
Web Site	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

---

# Chapter 05.07

## Higher Order Interpolation is a Bad Idea

*After reading this chapter, you should be able to:*

1. *Understand why higher order interpolation is a bad idea*

### Example

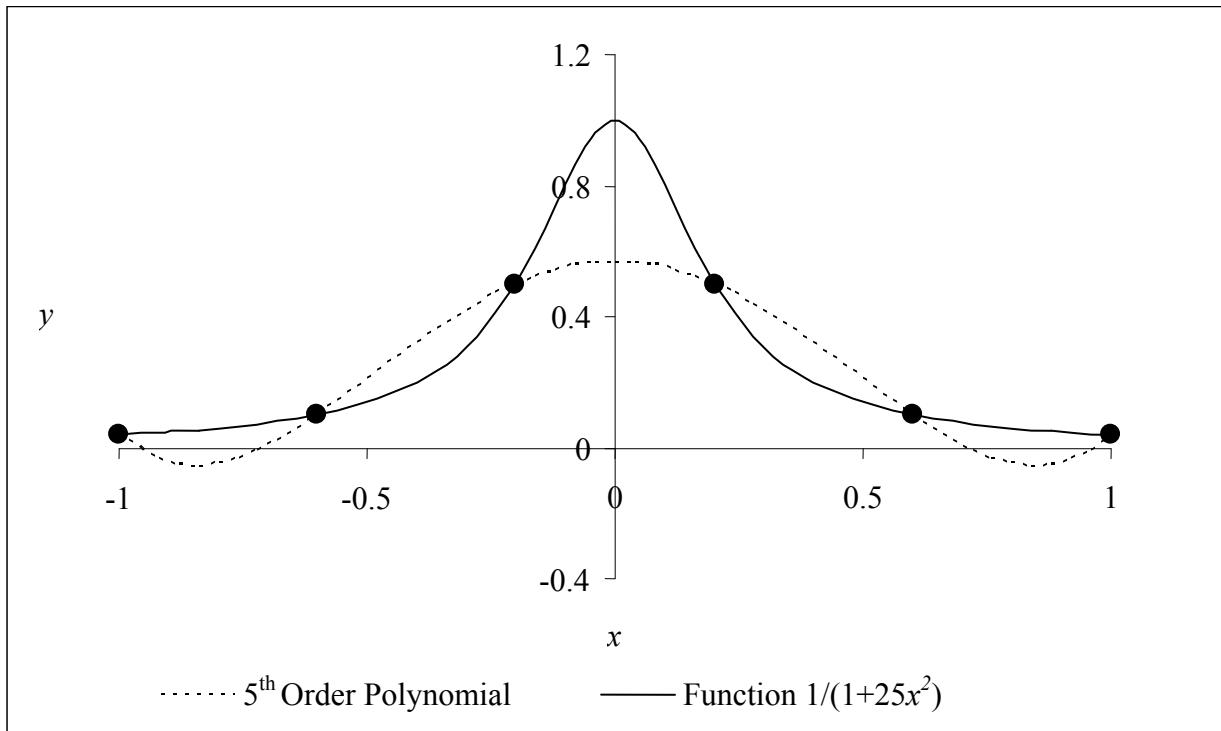
**Peter:** “Dr. Kaw, what is this you were talking about in class that higher order interpolation is a bad idea? More points, more accuracy; isn’t that the way it works?”

**Kaw:** “Come on in. In 1901, Runge wanted to show that higher order interpolation is a bad idea. He took this function,  $f(x) = 1/(1 + 25x^2)$  in the domain  $[-1, 1]$ .”

**Table 1.** Six equidistant points of  $f(x) = 1/(1 + 25x^2)$

$x$	$y$
-1	0.038462
-0.6	0.1
-0.2	0.5
0.2	0.5
0.6	0.1
1	0.038462

Let us choose 6 points equidistantly between  $-1$  and  $1$  as given in Table 1. You can interpolate these 6 data points by a 5th order polynomial. In Figure 1, I am then plotting the fifth order polynomial and the original function. See the oscillations in the interpolating polynomial. The polynomial does go through the six points, but at many other points it does not even come close to the original function. Just look at  $x = 0.85$ , the value of the function is 0.052459, but the fifth order polynomial gives you a value of -0.055762. That is a whopping 206.30 % relative error and also note the opposite sign.



**Figure 1** 5<sup>th</sup> order polynomial interpolation with six equidistant points.

**Peter:** “Maybe you are not taking enough points. Six points may be too small a number to approximate the function.”

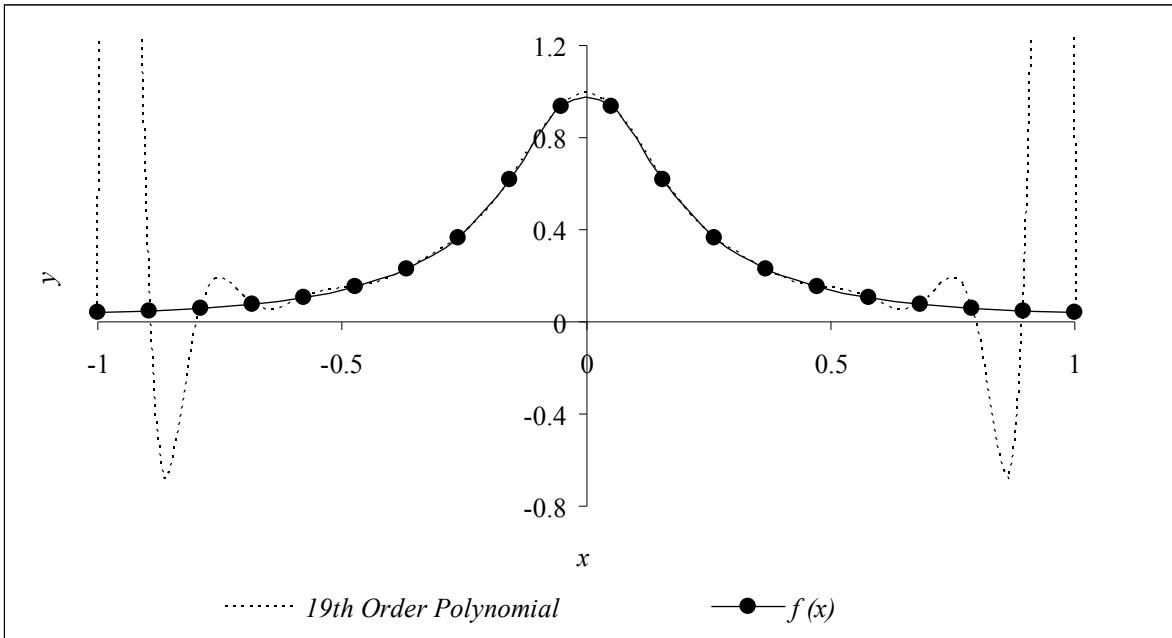
**Kaw:** “How many points do you want to choose?”

**Peter:** “Ok! Let’s get crazy. How about 20? That will give us a 19<sup>th</sup> order polynomial”

**Kaw:** “I chose 20 points equidistantly in  $[-1, 1]$ . It is not any better; the oscillations continue and get worse near the end points (Figure 2).”

**Peter:** “Yes, it is wild. It, however, did do a better job of approximating the data except near the ends, but at the ends it is worse than before. At our chosen point,  $x = 0.85$ , the value of the function is 0.052459, while we get  $-0.62944$  from the 19<sup>th</sup> order polynomial, and that is a big whopper error of 1299.9 %. Higher order interpolation is a bad idea. What is the solution to the problem then? What if we choose more points close to the end points?”

**Kaw:** “You are on to something. But, I need to go to teach my other class. You can get your question answered by seeing other anecdotes on the numerical methods web site. Just choose any interpolation module. You will get the answers to the questions you just asked.”



**Figure 2** 19th order polynomial interpolation with twenty equidistant points

---

## INTERPOLATION

---

Topic	Higher order interpolation is a bad idea
Summary	Textbook notes on errors which occur when using higher order interpolation.
Major	All majors of engineering
Authors	Autar Kaw
Last Revised	December 23, 2009
Web Site	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

---

## Chapter 05.08

# Why Do we Need Splines?

*After reading this chapter, you should be able to:*

1. *understand why we use splines for interpolation.*

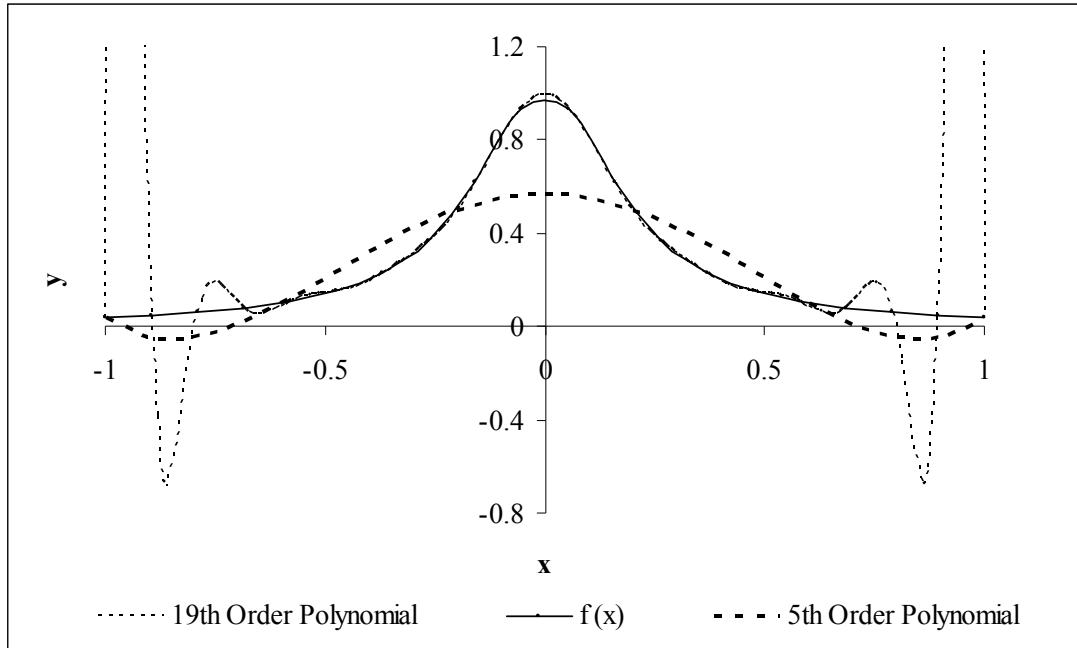
### Example

**Peter:** “Dr. Kaw, in class, you were talking about higher order interpolation being a bad idea and then telling us that taking more points is not going to get you a better approximation.”

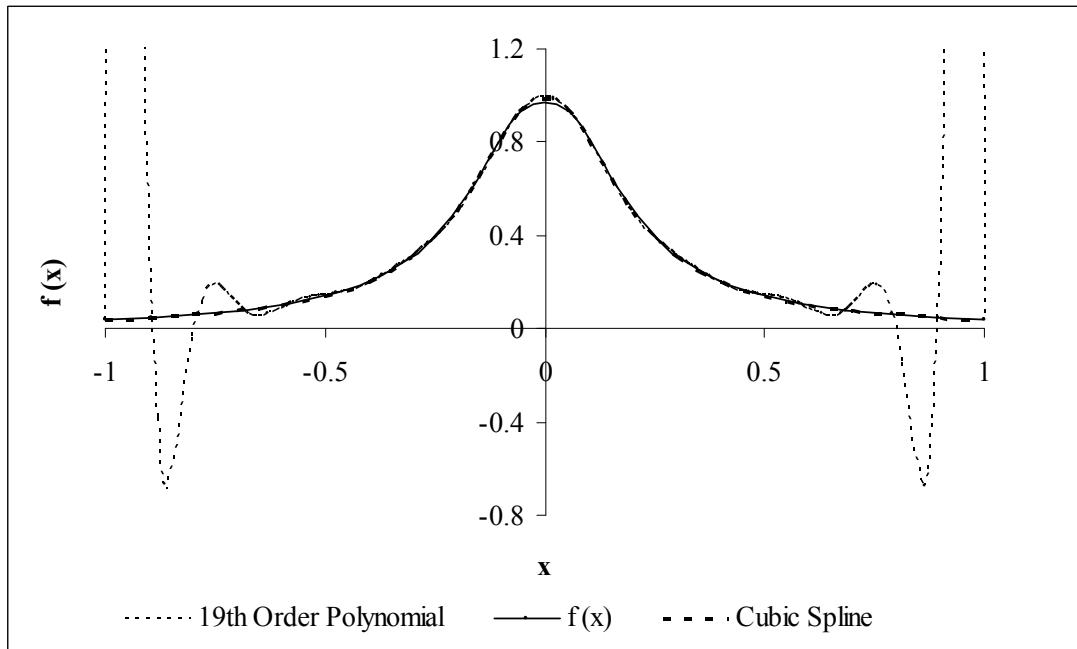
**Kaw:** “Yes, we were talking in class about the classic example taken by Runge. He took  $f(x) = 1/(1 + 25x^2)$  in the domain  $[-1, 1]$ . Choosing 20 equidistant points (Figure 1) on  $[-1, 1]$  to approximate the function by a 19th order polynomial gave worse results than when we chose 6 equidistant points to approximate the function by a 5th order polynomial.”

**Peter:** “Yes, it was wild. So what do we do? Accept this fact and roll over?”

**Kaw:** “Now, we do not have to do that. We can use interpolation such as cubic splines. Cubic splines approximate data between consecutive data points by cubic polynomials but at the same time use all the data to approximate the function. You can see from Figure 2 how cubic splines do a better job of approximating the data. The thin dash line is a 19<sup>th</sup> order polynomial approximation of the function by choosing 20 equidistant data points in  $[-1, 1]$ , while the thick dash line is the cubic spline approximation of the data. See how close the cubic splines are to the original function (continuous line).”



**Figure 1**  $5^{\text{th}}$  and  $19^{\text{th}}$  order polynomial approximations of Runge's function.



**Figure 2** Approximating Runge's function by a  $19^{\text{th}}$  order polynomial and a cubic spline.

## INTERPOLATION

Topic	Why do we need splines?
Summary	Textbook notes on understanding why we use splines for interpolation.

---

Major	All majors of engineering
Authors	Autar Kaw
Last Revised	December 23, 2009
Web Site	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

---

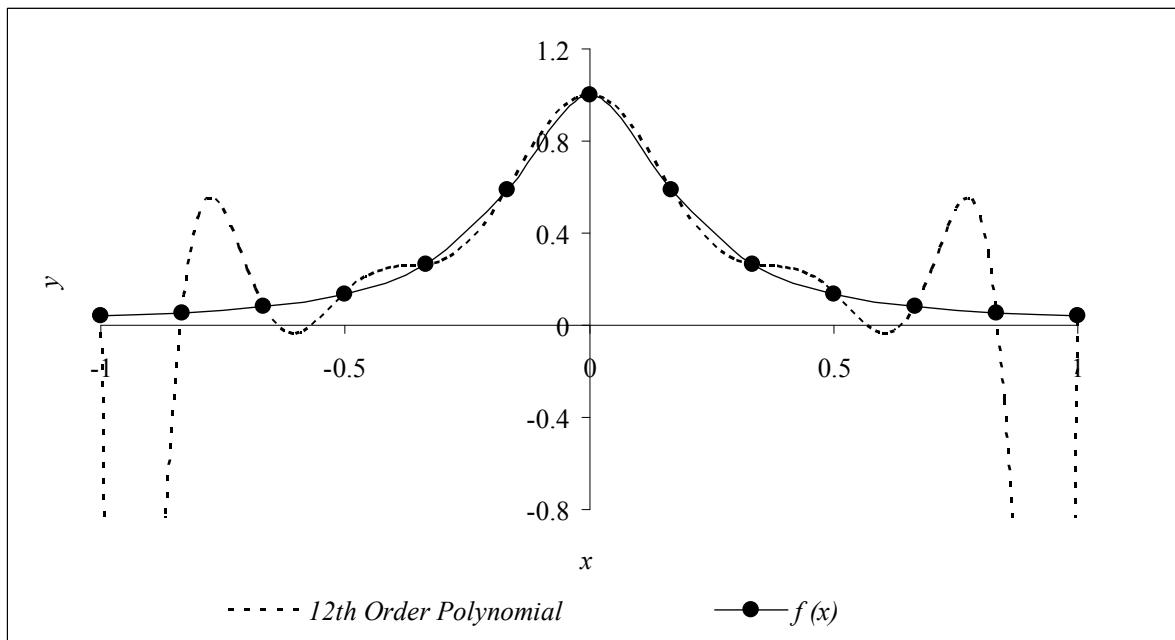
## Chapter 05.09 Choice of Points

After reading this chapter, you should be able to:

1. see how your choice of points affects interpolation.

### Example

**Peter:** “Dr. Kaw, in the last class you showed us that higher order interpolation is a bad idea. But when you took equidistant points between  $[-1,1]$  to approximate the function  $f(x) = 1/(1+25x^2)$ , it seemed that as you increased the number of points for approximation, the approximation was getting closer for the range of  $[-0.5,0.5]$ .

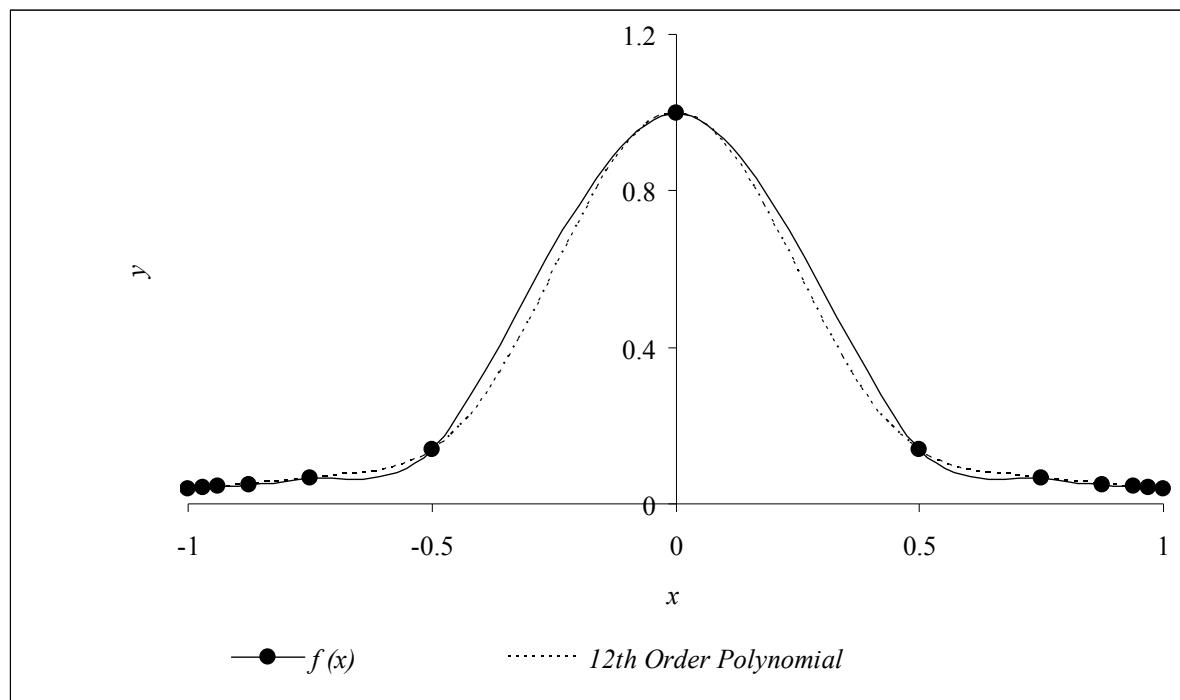


**Figure 1** 12<sup>th</sup> order polynomial interpolation with equidistantly spaced points

**Kaw:** “But it did give highly oscillatory results outside that range. In Figure 1, you are approximating the function by a 12<sup>th</sup> order polynomial (dash curve), and it matches the original function (continuous curve) very well between [-0.5, 0.5].

**Peter:** “What if we choose more points close to the end points?”

**Kaw:** “You are on to something. Yes, it would make a difference and Runge found that if you choose more points close to the ends, you do get a better approximation. Let us choose points not equidistantly but closer to the ends. You can see in Figure 2, how much closer the 12th order polynomial (dash curve) is to the original function (continuous curve). This is not to say that this choice of points will work for every case. The choice of points is dependent on the value of the possible derivatives of a function, but this concept is beyond the scope of the course.”



**Figure 2** 12<sup>th</sup> order polynomial interpolation with more point at the end

---

## INTERPOLATION

---

Topic	Choice of points affects interpolation
Summary	Textbook notes on how your choice of points can affect interpolation.
Major	All majors of engineering
Authors	Autar Kaw
Last Revised	December 23, 2009
Web Site	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

---

## Chapter 05.10

### Shortest Path of a Robot

*After reading this chapter, you should be able to:*

1. *find the shortest smooth path through consecutive points, and*
2. *compare the lengths of different paths.*

#### Example

**Peter:** “Dr. Kaw, I am taking a course in manufacturing. We are solving the following problem. A robot arm with a rapid laser is used to do a quick quality check, such as the radius of hole, on six holes on a rectangular plate 15” $\times$ 10” at several points as shown in Table 1 and Figure 1.

**Table 1** The coordinate values of six holes on a rectangle plate.

$x$	$y$
2.00	7.2
4.5	7.1
5.25	6.0
7.81	5.0
9.20	3.5
10.60	5.0

I am using Excel to fit a fifth order polynomial through the 6 points. But, when I plot the polynomial, it is taking a long path! (Figure 2)”

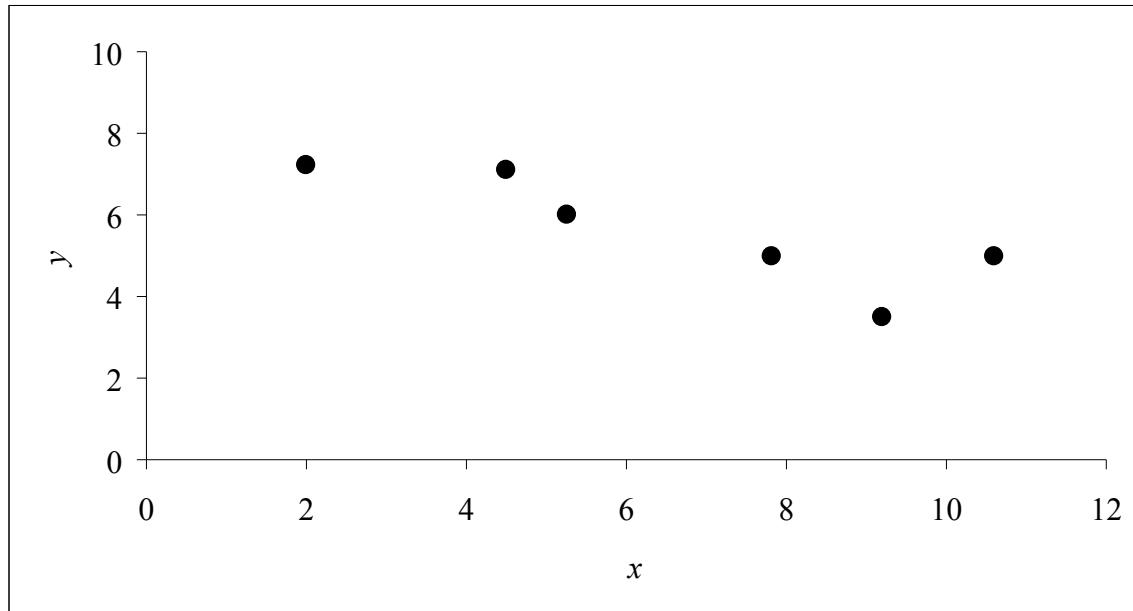
**Kaw:** “Why do you not just join the consecutive points by a straight line; just like the kids do at Pizza Hut™ with those ‘Connect the dots’ activities?”

**Peter:** “You are making me hungry and I wish it were that easy. The path of the robot going from one point to another point needs to be smooth so as to avoid sharp jerks in the arm that can otherwise create premature wear and tear of the robot arm.”

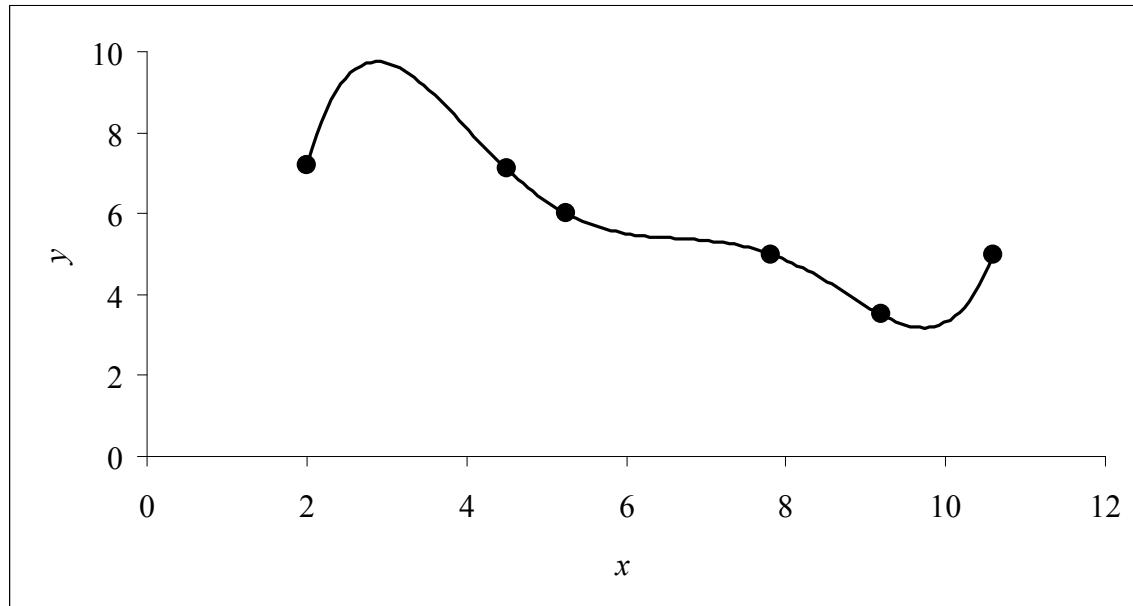
**Kaw:** “As I recall, you took my course in Numerical Methods. What was that ..... one year ago?”

**Peter:** “Yes, your memory is sharp, but my retention from that course – can we not talk about that?!?”

**Kaw:** “Come into my office. I wrote this program using Maple. See this function,  $f(x) = 1/(1 + 25x^2)$ . I am choosing 7 points equidistantly (Table 2) between  $-1$  and  $1$ .



**Figure 1** Locations of holes on the rectangular plate.



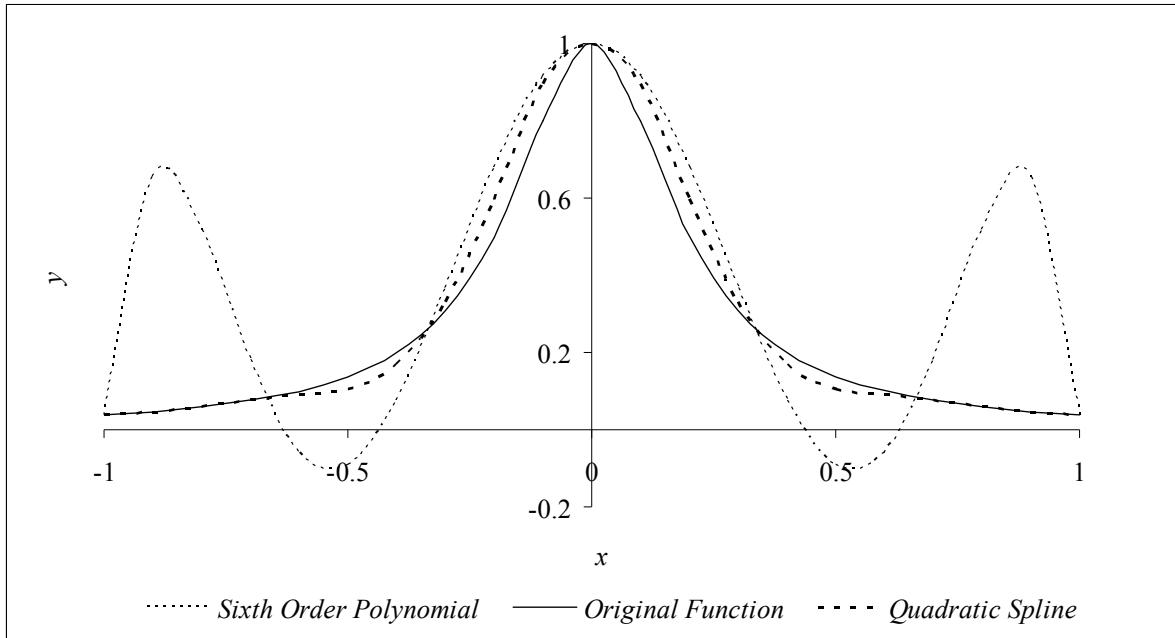
**Figure 2** Approximating the path of the robot using 5th order polynomial.

Now look at the sixth order interpolating polynomial and the original function (Figure 3). See the oscillations in the interpolating polynomial. In 1901, Runge used this example function to show that higher order interpolation is a bad idea. One of the solutions to your

robot path problem is to use quadratic or cubic spline interpolation. That will give you a smooth curve with fewer oscillations, and a smoother and shorter path.”

**Table 2** The coordinate values of 7 equidistantly spaced points.

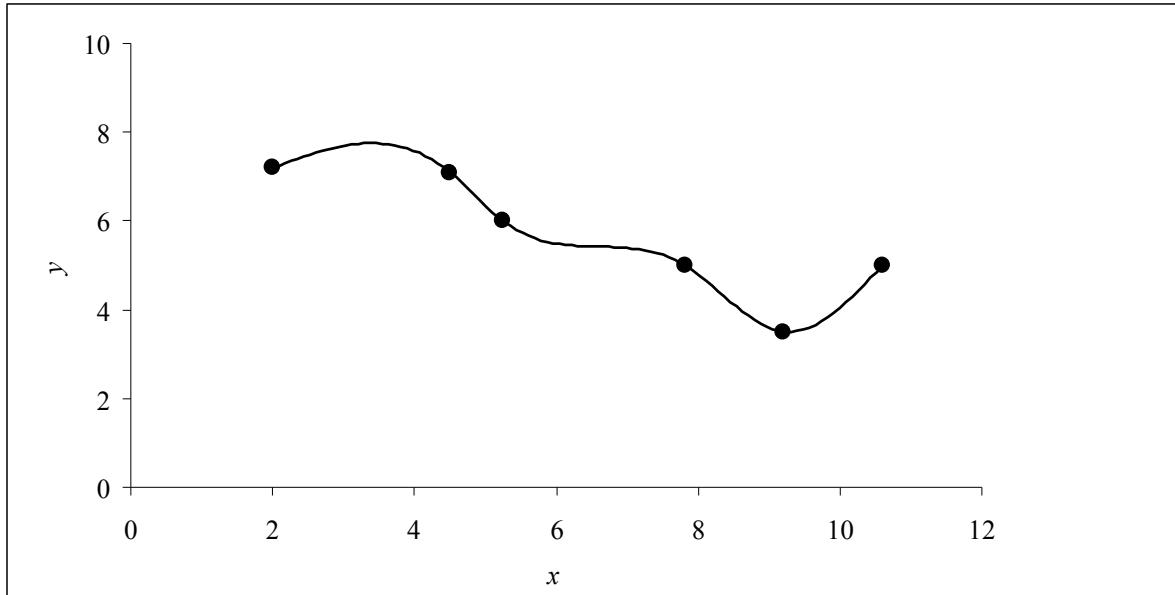
$x$	$y = \frac{1}{1+25x^2}$
-1	0.038462
-0.66667	0.0826
-0.33333	0.264706
0	1
0.333333	0.264706
0.666667	0.082569
1	0.0385



**Figure 3** Runge’s function interpolated.

**Peter:** “Okay. Let’s give that a try.”

**Kaw:** “Now, let’s try generating a set of cubic splines to go through the data.”



**Figure 4** Path of the robot arm using cubic spline interpolation.

**Peter:** “Wow! That (Figure 4) looks much better!”

**Kaw:** “It may look better, but let’s find out for sure. See if you can combine the two plots (Figure 5) and compare the lengths of each path.”

**Peter:** “The length of a path  $S$  if  $y = f(x)$  from  $a$  to  $b$  is given by

$$S = \int_a^b \sqrt{1 + \left(\frac{df}{dx}\right)^2} dx$$

Right?”

**Table 3** Comparison of the length of curves.

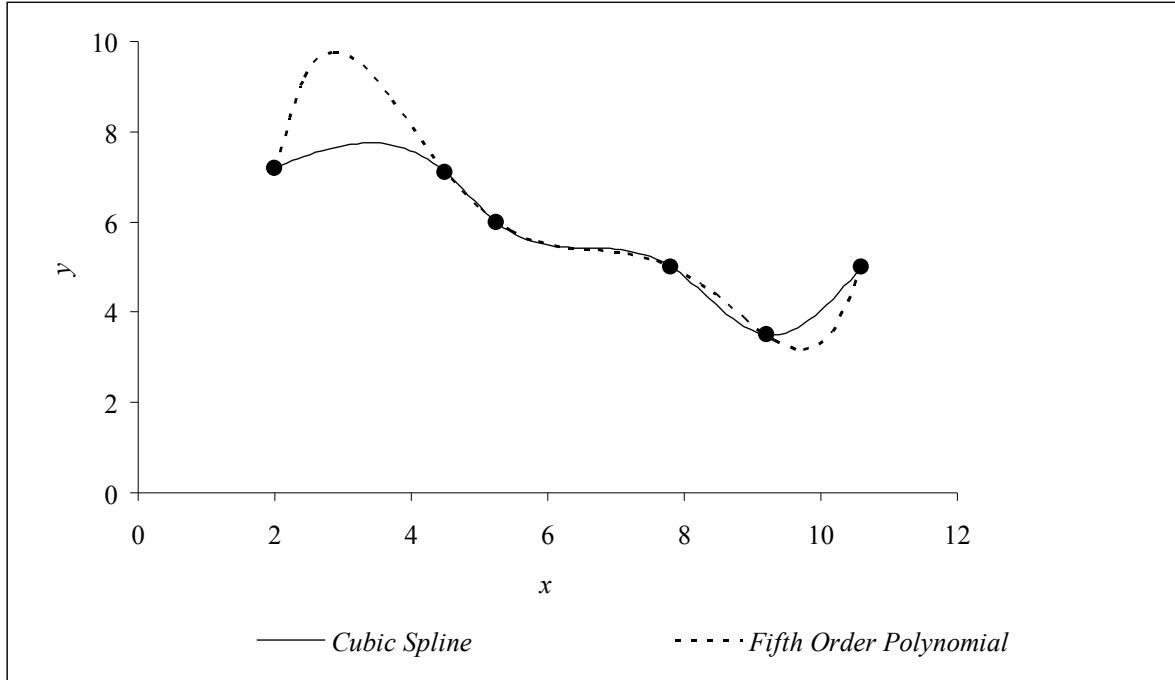
Type of interpolation	5th order polynomial	Cubic Spline
Length of Curve	14.919”	11.248”

**Kaw:** “Yes! You solved the problem. See Table 3 for answers.”

**Peter:** “I guess your class was good for something after all, Dr. Kaw.”

**Kaw:** “Are you sure? You could have always fallen back on the connecting-the-dots method. Besides, you don’t want to grow up ... you’re a Pizza Hut™ kid, right?”

**Peter:** “That’s a Toys R’ Us™ kid. You’ll do anything to be reminded of songs, won’t you?”



**Figure 5** Path of robot arm compared using polynomial interpolation and cubic spline interpolation.

---

#### INTERPOLATION

---

Topic	Shortest path of a robot
Summary	An example of interpolation: A robot arm path needs to be developed over several points on a flat plate. The path needs to be smooth to avoid sudden jerky motion and at the same time needs to be short.
Major	General Engineering
Authors	Autar Kaw, Michael Keteltas
Date	December 23, 2009
Web Site	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

---

# **Chapter 06.01**

## **Statistics Background of Regression Analysis**

*After reading this chapter, you should be able to:*

1. *review the statistics background needed for learning regression, and*
2. *know a brief history of regression.*

### **Review of Statistical Terminologies**

Although the language of statistics may be used at an elementary and descriptive level in this chapter, it makes an integral part of our every day discussions. When two friends talk about the weather (whether it will rain or not - probability), or the time it takes to drive from point A to point B (speed - mean or average), or baseball facts (all time career RBI or home runs of a sportsman - sorting, range), or about class grades (lowest and highest score - range and sorting), they are invariably using statistical tools. From the foregoing, it is imperative then that we review some of the statistical terminologies that we may encounter in studying the topic of regression. Some key terms we need to review are sample, arithmetic mean (average), error or deviation, standard deviation, variance, coefficient of variation, probability, Gaussian or normal distribution, degrees of freedom, and hypothesis.

### **Elementary Statistics**

A statistical sample is a fraction or a portion of the whole (population) that is studied. This is a concept that may be confusing to many and is best illustrated with examples. Consider that a chemical engineer is interested in understanding the relationship between the rate of a reaction and temperature. It is impractical for the engineer to test all possible and measurable temperatures. Apart from the fact that the instrument for temperature measurement have limited temperature ranges for which they can function, the sheer number of hours required to measure every possible temperature makes it impractical. What the engineer does is choose a temperature range (based on his/her knowledge of the chemistry of the system) in which to study. Within the chosen temperature range, the engineer further chooses specific temperatures that span the range within which to conduct the experiments. These chosen temperatures for study constitute the sample while all possible temperatures are the population. In statistics, the sample is the fraction of the population chosen for study.

The location of the center of a distribution - the mean or average - is an item of interest in our every day lives. We use the concept when we talk about the average income, the class average for a test, the average height of some persons or about one being overweight (based on the average weight expected of an individual with similar

characteristics) or not. The arithmetic mean of a sample is a measure of its central tendency and is evaluated by dividing the sum of individual data points by the number of points.

Consider Table 1 which 14 measurements of the concentration of sodium chlorate produced in a chemical reactor operated at a pH of 7.0.

**Table 1** Chlorate ion concentration in mmol/cm<sup>3</sup>

12.0	15.0	14.1	15.9	11.5	14.8	11.2	13.7	15.9	12.6	14.3	12.6	12.1	14.8
------	------	------	------	------	------	------	------	------	------	------	------	------	------

The arithmetic mean  $\bar{y}$  is mathematically defined as

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n} \quad (1)$$

which is the sum of the individual data points  $y_i$  divided by the number of data points  $n$ .

One of the measures of the spread of the data is the range of the data. The range  $R$  is defined as the difference between the maximum and minimum value of the data as

$$R = y_{\max} - y_{\min} \quad (2)$$

where

$y_{\max}$  is the maximum of the values of  $y_i$ ,  $i = 1, 2, \dots, n$ ,

$y_{\min}$  is the minimum of the values of  $y_i$ ,  $i = 1, 2, \dots, n$ .

However, range may not give a good idea of the spread of the data as some data points may be far away from most other data points (such data points are called outliers). That is why the deviation from the average or arithmetic mean is looked as a better way to measure the spread. The residual between the data point and the mean is defined as

$$e_i = y_i - \bar{y} \quad (3)$$

The difference of each data point from the mean can be negative or positive depending on which side of the mean the data point lies (recall the mean is centrally located) and hence if one calculates the sum of such differences to find the overall spread, the differences may simply cancel each other. That is why the sum of the square of the differences is considered a better measure. The sum of the squares of the differences, also called summed squared error (SSE),  $S_t$ , is given by

$$S_t = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (4)$$

Since the magnitude of the summed squared error is dependent on the number of data points, an average value of the summed squared error is defined as the variance,  $\sigma^2$

$$\sigma^2 = \frac{S_t}{n-1} = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1} \quad (5)$$

The variance,  $\sigma^2$  is sometimes written in two different convenient formulas as

$$\sigma^2 = \frac{\sum_{i=1}^n y_i^2 - \left(\sum_{i=1}^n y_i\right)^2 / n}{n-1} \quad (6)$$

or

$$\sigma^2 = \frac{\sum_{i=1}^n y_i^2 - n\bar{y}^2}{n-1} \quad (7)$$

However, why is the variance divided by  $(n-1)$  and not  $n$  as we have  $n$  data points? This is because with the use of the mean in calculating the variance, we lose the independence of one of the data points. That is, if you know the mean of  $n$  data points, then the value of one of the  $n$  data points can be calculated by knowing the other  $(n-1)$  data points.

To bring the variation back to the same level of units as the original data, a new term called standard deviation,  $\sigma$ , is defined as

$$\sigma = \sqrt{\frac{S_t}{n-1}} = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}} \quad (8)$$

Furthermore, the ratio of the standard deviation to the mean, known as the coefficient variation  $c.v$  is also used to normalize the spread of a sample.

$$c.v = \frac{\sigma}{\bar{y}} \times 100 \quad (9)$$

### Example 1

Use the data in Table 1 to calculate the

- mean chlorate concentration,
- range of data,
- residual of each data point,
- sum of the square of the residuals.
- sample standard deviation,
- variance, and
- coefficient of variation.

### Solution

Set up a table (see Table 2) containing the data, the residual for each data point and the square of the residuals.

**Table 2** Data and data summations for statistical calculations.

$i$	$y_i$	$y_i^2$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
1	12	144	-1.6071	2.5829
2	15	225	1.3929	1.9401
3	14.1	198.81	0.4929	0.24291

4	15.9	252.81	2.2929	5.2572
5	11.5	132.25	-2.1071	4.4401
6	14.8	219.04	1.1929	1.4229
7	11.2	125.44	-2.4071	5.7943
8	13.7	187.69	0.0929	0.0086224
9	15.9	252.81	2.2929	5.2572
10	12.6	158.76	-1.0071	1.0143
11	14.3	204.49	0.6929	0.48005
12	12.6	158.76	-1.0071	1.0143
13	12.1	146.41	-1.5071	2.2715
14	14.8	219.04	1.1929	1.4229
$\sum_{i=1}^{14}$	190.50	2625.3	0.0000	33.149

- a) Mean chlorate concentration as from Equation (1)

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n} = \frac{190.5}{14} = 13.607$$

- b) The range of data as per Equation (2) is

$$\begin{aligned} R &= y_{\max} - y_{\min} \\ &= 15.9 - 11.2 \\ &= 4.7 \end{aligned}$$

- c) Residual at each point is shown in Table 2. For example, at the first data point as per Equation (3)

$$\begin{aligned} e_1 &= y_1 - \bar{y} \\ &= 12.0 - 13.607 \\ &= -1.6071 \end{aligned}$$

- d) The sum of the square of the residuals as from Equation (4) is

$$\begin{aligned} S_t &= \sum_{i=1}^n (y_i - \bar{y})^2 \\ &= 33.149 \text{ (See Table 2)} \end{aligned}$$

- e) The standard deviation as per Equation (8) is

$$\begin{aligned} \sigma &= \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}} \\ &= \sqrt{\frac{33.149}{14-1}} \\ &= 1.5969 \end{aligned}$$

- f) The variance is calculated as from Equation (5)

$$\begin{aligned} \sigma^2 &= (1.597)^2 \\ &= 2.5499 \end{aligned}$$

The variance can be calculated using Equation (6)

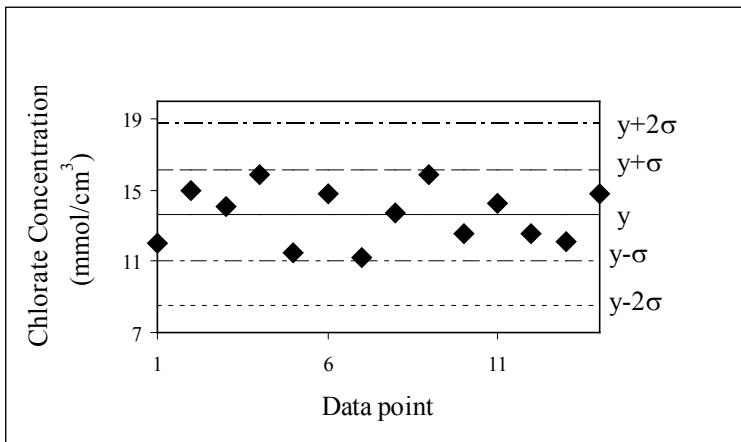
$$\begin{aligned}\sigma^2 &= \frac{\sum_{i=1}^n y_i^2 - \frac{\left(\sum_{i=1}^n y_i\right)^2}{n}}{n-1} \\ &= \frac{2625.31 - \frac{(190.5)^2}{14}}{14-1} \\ &= 2.5499\end{aligned}$$

or by using Equation (7)

$$\begin{aligned}\sigma^2 &= \frac{\sum_{i=1}^n y_i^2 - n\bar{y}^2}{n-1} \\ &= \frac{2625.3 - 14 \times 13.607^2}{14-1} \\ &= 2.5499\end{aligned}$$

g) The coefficient of variation,  $c.v$  as from Equation (9) is

$$\begin{aligned}c.v &= \frac{\sigma}{\bar{y}} \times 100 \\ &= \frac{1.5969}{13.607} \times 100 \\ &= 11.735\%\end{aligned}$$



**Figure 1** Chlorate concentration data points.

### A Brief History of Regression

Anyone who is familiar with the Pearson Product Moment Correlation (PPMC) will no doubt associate regression principles with the name of Pearson. Although this association may be right, the concept of linear regression was largely due to the work of Galton, a cousin of Charles Darwin of the evolution theory fame. Sir Galton's work on inherited

characteristics of sweet peas led to the initial conception of linear regression. His treatment of regression was not mathematically rigorous. The mathematical rigor and subsequent development of multiple regression were due largely to the contributions of his assistant and co-worker - Karl Pearson.

It is however instructive to note for historical accuracy that the development of regression could be attributed to the attempt at answering the question of hereditary - how and what characteristics offspring acquire from their progenitor. Sweet peas were used by Galton in his observations of characteristics of next generations of a given species. Despite his poor choice of descriptive statistics and limited mathematical rigor, Galton was able to generalize his work over a variety of hereditary problems. He further arrived at the idea that the differences in regression slopes were due to differences in variability between different sets of measurements. In today's appreciation of this, one can say that Galton recognized the ratio of variability of two measures was a key factor in determining the slope of the regression line.

The first rigorous treatment of correlation and regression was the work of Pearson in 1896. In the paper in the Philosophical Transactions of the Royal Society of London, Pearson showed that the optimum values of both the regression slope and the correlation coefficient for a straight line could be evaluated from the product-moment,

$$\sum_{i=1}^n \frac{(x_i - \bar{x})(y_i - \bar{y})}{n},$$

where  $\bar{x}$  and  $\bar{y}$  are the means of observed  $x$  and  $y$  values, respectively. In the 1896 paper, Pearson had attributed the initial mathematical formula for correlation to Auguste Bravais' work fifty years earlier. Pearson stated that although Bravais did demonstrate the use of product-moment for calculating the correlation coefficient, he did not show that it provided the best fit for the data.

---

## REGRESSION

---

<b>Topic</b>	Statistics Background for Regression
<b>Summary</b>	Textbook notes for the background of regression
<b>Major</b>	All engineering majors
<b>Authors</b>	Egwu Kalu, Autar Kaw
<b>Date</b>	October 11, 2008
<b>Web Site</b>	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

---

## **Chapter 06.02**

### **Introduction of Regression Analysis**

*After reading this chapter, you should be able to:*

1. know what regression analysis is,
2. know the effective use of regression, and
3. enumerate uses and abuses of regression.

#### **What is regression analysis?**

Regression analysis gives information on the relationship between a response (dependent) variable and one or more (predictor) independent variables to the extent that information is contained in the data. The goal of regression analysis is to express the response variable as a function of the predictor variables. The duality of fit and the accuracy of conclusion depend on the data used. Hence non-representative or improperly compiled data result in poor fits and conclusions. Thus, for effective use of regression analysis one must

1. investigate the data collection process,
2. discover any limitations in data collected, and
3. restrict conclusions accordingly.

Once a regression analysis relationship is obtained, it can be used to predict values of the response variable, identify variables that most affect the response, or verify hypothesized causal models of the response. The value of each predictor variable can be assessed through statistical tests on the estimated coefficients (multipliers) of the predictor variables.

An example of a regression model is the linear regression model which is a linear relationship between response variable,  $y$  and the predictor variable,  $x_i, i = 1, 2, \dots, n$  of the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \varepsilon \quad (1)$$

where

$\beta_0, \beta_1, \dots, \beta_n$  are regression coefficients (unknown model parameters), and  
 $\varepsilon$  is the error due to variability in the observed responses.

### Example 1

In the transformation of raw or uncooked potato to cooked potato, heat is applied for some specific time. One might postulate that the amount of untransformed portion of the starch ( $y$ ) inside the potato is a linear function of time ( $t$ ) and temperature ( $\theta$ ) of cooking. This is represented as

$$y = \beta_0 + \beta_1 t + \beta_2 \theta + \varepsilon \quad (2)$$

Linear as used in linear regression refers to the form of occurrence of the unknown parameters,  $\beta_1$  and  $\beta_2$  as simple linear multipliers of the predictor variable. Thus, the two equations below are also both linear.

$$y = \beta_0 + \beta_1 t + \beta_2 t \theta + \beta_3 \theta + \varepsilon \quad (3)$$

$$y = \beta_0 + \beta_1 t \theta + \beta_2 \theta + \varepsilon \quad (4)$$

### Comparison of Regression and Correlation

Unlike regression, correlation analysis assesses the simultaneous variability of a collection of variables. The relationship is not directional and interest is not on how some variables respond to others but on how they are mutually associated. Thus, simultaneous variability of a collection of variables is referred to as correlation analysis.

### Uses of Regression Analysis

Three uses for regression analysis are for

1. prediction
2. model specification and
3. parameter estimation.

Regression analysis equations are designed only to make predictions. Good predictions will not be possible if the model is not correctly specified and accuracy of the parameter not ensured. However, accurate prediction and model specification require that all relevant variables be accounted for in the data and the prediction equation be defined in the correct functional form for all predictor variables.

Parameter estimation is the most difficult to perform because not only is the model required to be correctly specified, the prediction must also be accurate and the data should allow for good estimation. For example, multicollinearity creates a problem and requires that some estimators may not be used. Thus, limitations of data and inability to measure all predictor variables relevant in a study restrict the use of prediction equations.

### Abuses of Regression Analysis

Let us examine three common abuses of regression analysis.

1. Extrapolation
2. Generalization
3. Causation

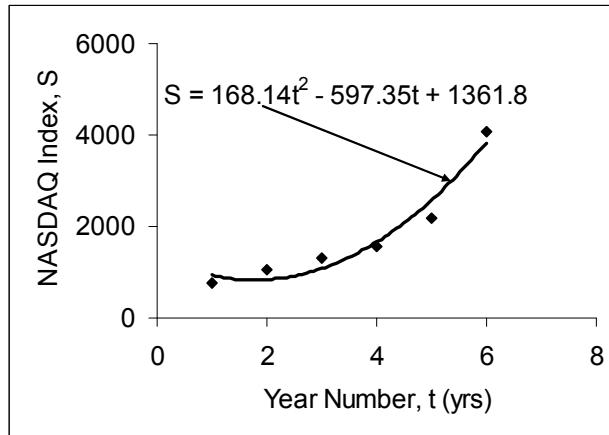
## Extrapolation

If you were dealing in the stock market or even interested in it, then you might remember the stock market crash of March 2000. During 1997-1999, investors thought they would double their money every year. They started buying fancy cars and houses on credit, and living the high life. Little did they know that the whole market was hyped on speculation and little economic sense. The Enron and MCI financial fiascos soon followed.

Let us look if we could have safely extrapolated the NASDAQ index<sup>1</sup> from past years. Below is the table of NASDAQ index,  $S$ , as a function of end of year number,  $t$  (Year 1 is the end of year 1994, and Year 6 is the end of year 1999).

**Table 1** NASDAQ index as a function of year number.

Year Number ( $t$ )	NASDAQ Index ( $S$ )
1 (1994)	752
2 (1995)	1052
3 (1996)	1291
4 (1997)	1570
5 (1998)	2193
6 (1999)	4069



**Figure 1** The regression line of NASDAQ Index as a function of year number.

---

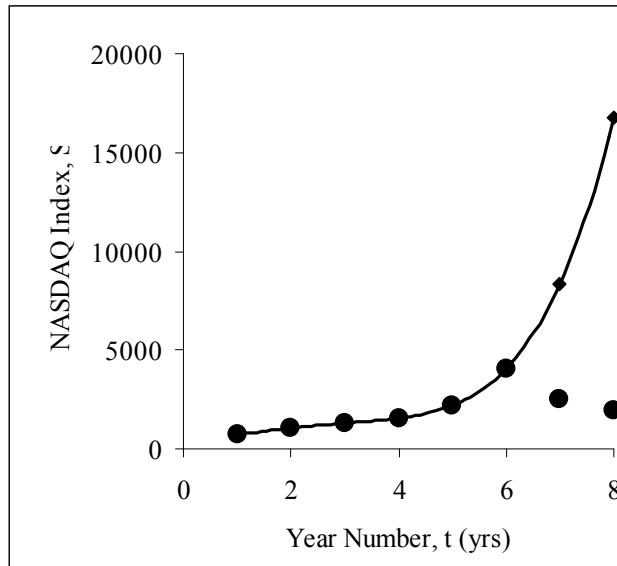
<sup>1</sup> NASDAQ (National Association of Securities Dealers Automated Quotations) index is a composite index based on the stock market value of 3,000 companies. The NASDAQ index began on February 5, 1971 with a base value of 100. Twenty years later in 1995, NASDAQ index crossed the 1000 mark. It rose as high as 5132 on March 10, 2000 and currently is at a value of 2282 (February 19, 2006).

A relationship  $S = a_0 + a_1t + a_2t^2$  between the NASDAQ index,  $S$ , and the year number,  $t$ , is developed using least square regression and is found to be  $S = 168.14t^2 - 597.37t + 1361.8$ . The data and the regression line are shown in Figure 1. The data is given only for Years 1 through 6 and it is desired to calculate the value for  $t > 6$ . This is extrapolation outside the model data. The error inherent in this model is shown in Table 2 and Figure 2. Look at the Year 7 and 8 that was not included in the data – the error between the predicted and actual values is 119% and 277%, respectively.

**Table 2** NASDAQ index as a function of year number.

Year Number ( $t$ )	NASDAQ Index ( $S$ )	Predicted Index	Absolute Relative True Error (%)
1 (1994)	752	933	24
2 (1995)	1052	840	20
3 (1996)	1291	1083	16
4 (1997)	1570	1663	6
5 (1998)	2193	2579	18
6 (1999)	4069	3831	6
7 (2000)	2471	5419	119
8 (2001)	1951	7344	276

This illustration is not exaggerated and it is important that a careful use of any given model equations is always employed. At all times, it is imperative to infer the domain of independent variables for which a given equation is valid.



**Figure 2** Extrapolated curve and actual data for Years 7 and 8.

### Generalization

Generalization could arise when unsupported or over exaggerated claims are made. It is not often possible to measure all predictor variables relevant in a study. For example, a

study carried out about the behavior of men might have inadvertently restricted the survey to Caucasian men only. Shall we then generalize the result as the attributes of all men irrespective of race? Such use of regression equation is an abuse since the limitations imposed by the data restrict the use of the prediction equations to Caucasian men.

### Misidentification

Finally, misidentification of causation is a classic abuse of regression analysis equations. Regression analysis can only aid in the confirmation or refutation of a causal model - the model must however have a theoretical basis. In a chemical reacting system in which two species react to form a product, the amount of product formed or amount of reacting species vary with time. Although a regression equation of species concentration and time can be obtained, one cannot attribute time as the causal agent for the varying species concentration. Regression analysis cannot prove causality, rather it can only substantiate or contradict causal assumptions. Anything outside this is an abuse of regression analysis method.

### Least Squares Methods

This is the most popular method of parameter estimation for coefficients of regression models. It has well known probability distributions and gives unbiased estimators of regression parameters with the smallest variance.

We wish to predict the response to  $n$  data points  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  by a regression model given by

$$y = f(x) \quad (6)$$

where, the function  $f(x)$  has regression constants that need to be estimated.

For example

$f(x) = a_0 + a_1x$  is a straight-line regression model with constants  $a_0$  and  $a_1$

$f(x) = a_0 e^{a_1 x}$  is an exponential model with constants  $a_0$  and  $a_1$

$f(x) = a_0 + a_1x + a_2x^2$  is a quadratic model with constants  $a_0$ ,  $a_1$  and  $a_2$

A measure of goodness of fit, that is how the regression model  $f(x)$  predicts the response variable  $y$  is the magnitude of the residual,  $E_i$  at each of the  $n$  data points.

$$E_i = y_i - f(x_i), i = 1, 2, \dots, n \quad (7)$$

Ideally, if all the residuals  $E_i$  are zero, one may have found an equation in which all the points lie on a model. Thus, minimization of the residual is an objective of obtaining regression coefficients. In the least squares method, estimates of the constants of the models are chosen such that minimization of the sum of the squared residuals is achieved, that is minimize  $\sum_{i=1}^n E_i^2$ .

### Why minimize the sum of the square of the residuals?

Why not for instance minimize the sum of the residual errors,  $\sum_{i=1}^n E_i$ , or the sum of the absolute values of the residuals,  $\sum_{i=1}^n |E_i|$ ? Alternatively, constants of the model can be chosen such that the average residual is zero without making individual residuals small. Will any of these criteria yield unbiased parameters with the smallest variance? All of these questions will be answered when we discuss linear regression in the next chapter (Chapter 06.03).

---

### Regression

---

<b>Topic</b>	Introduction to Regression
<b>Summary</b>	Textbook notes for the introduction to regression
<b>Major</b>	All engineering majors
<b>Authors</b>	Egwu Kalu, Autar Kaw
<b>Date</b>	October 11, 2008
<b>Web Site</b>	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

---

## Chapter 06.03

### Linear Regression

After reading this chapter, you should be able to

1. define regression,
2. use several minimizing of residual criteria to choose the right criterion,
3. derive the constants of a linear regression model based on least squares method criterion,
4. use in examples, the derived formulas for the constants of a linear regression model, and
5. prove that the constants of the linear regression model are unique and correspond to a minimum.

Linear regression is the most popular regression model. In this model, we wish to predict response to  $n$  data points  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  by a regression model given by

$$y = a_0 + a_1 x \quad (1)$$

where  $a_0$  and  $a_1$  are the constants of the regression model.

A measure of goodness of fit, that is, how well  $a_0 + a_1 x$  predicts the response variable  $y$  is the magnitude of the residual  $\varepsilon_i$  at each of the  $n$  data points.

$$E_i = y_i - (a_0 + a_1 x_i) \quad (2)$$

Ideally, if all the residuals  $\varepsilon_i$  are zero, one may have found an equation in which all the points lie on the model. Thus, minimization of the residual is an objective of obtaining regression coefficients.

The most popular method to minimize the residual is the least squares methods, where the estimates of the constants of the models are chosen such that the sum of the squared residuals is minimized, that is minimize  $\sum_{i=1}^n E_i^2$ .

**Why minimize the sum of the square of the residuals?** Why not, for instance, minimize the sum of the residual errors or the sum of the absolute values of the residuals? Alternatively, constants of the model can be chosen such that the average residual is zero without making individual residuals small. Will any of these criteria yield unbiased

parameters with the smallest variance? All of these questions will be answered below. Look at the data in Table 1.

**Table 1** Data points.

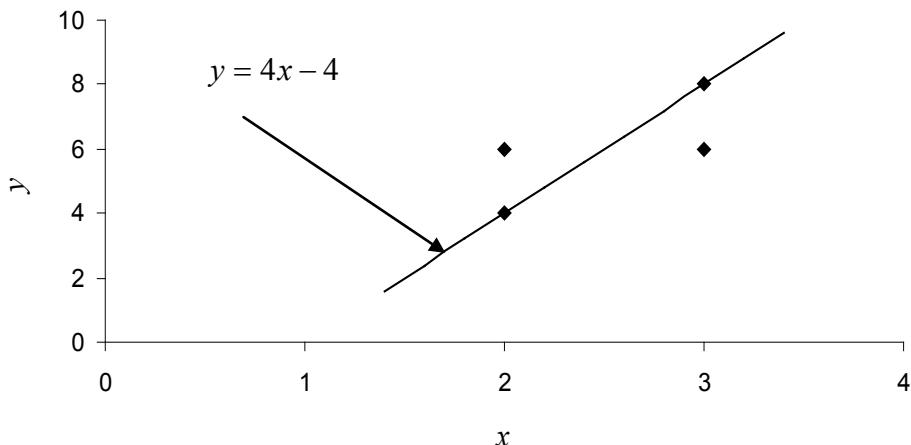
$x$	$y$
2.0	4.0
3.0	6.0
2.0	6.0
3.0	8.0

To explain this data by a straight line regression model,

$$y = a_0 + a_1 x \quad (3)$$

and using minimizing  $\sum_{i=1}^n E_i$  as a criteria to find  $a_0$  and  $a_1$ , we find that for (Figure 1)

$$y = 4x - 4 \quad (4)$$



**Figure 1** Regression curve  $y = 4x - 4$  for  $y$  vs.  $x$  data.

the sum of the residuals,  $\sum_{i=1}^4 E_i = 0$  as shown in the Table 2.

**Table 2** The residuals at each data point for regression model  $y = 4x - 4$ .

$x$	$y$	$y_{predicted}$	$\varepsilon = y - y_{predicted}$
2.0	4.0	4.0	0.0
3.0	6.0	8.0	-2.0
2.0	6.0	4.0	2.0
3.0	8.0	8.0	0.0
			$\sum_{i=1}^4 \varepsilon_i = 0$

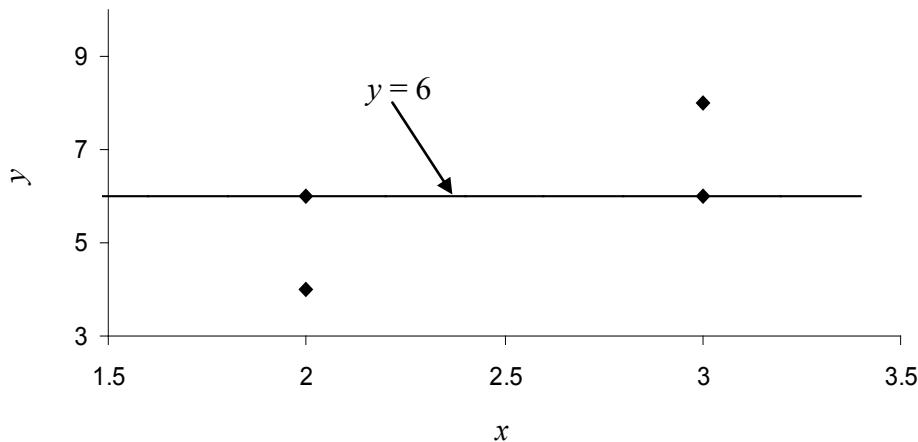
So does this give us the smallest error? It does as  $\sum_{i=1}^4 E_i = 0$ . But it does not give unique values for the parameters of the model. A straight-line of the model

$$y = 6 \quad (5)$$

also makes  $\sum_{i=1}^4 E_i = 0$  as shown in the Table 3.

**Table 3** The residuals at each data point for regression model  $y = 6$

$x$	$y$	$y_{predicted}$	$\epsilon = y - y_{predicted}$
2.0	4.0	6.0	-2.0
3.0	6.0	6.0	0.0
2.0	6.0	6.0	0.0
3.0	8.0	6.0	2.0
			$\sum_{i=1}^4 E_i = 0$



**Figure 2** Regression curve  $y = 6$  for  $y$  vs.  $x$  data.

Since this criterion does not give a unique regression model, it cannot be used for finding the regression coefficients. Let us see why we cannot use this criterion for any general data. We want to minimize

$$\sum_{i=1}^n E_i = \sum_{i=1}^n (y_i - a_0 - a_1 x_i) \quad (6)$$

Differentiating Equation (6) with respect to  $a_0$  and  $a_1$ , we get

$$\frac{\partial \sum_{i=1}^n E_i}{\partial a_0} = -\sum_{i=1}^n 1 = -n \quad (7)$$

$$\frac{\partial \sum_{i=1}^n E_i}{\partial a_1} = -\sum_{i=1}^n x_i = -n \bar{x} \quad (8)$$

Putting these equations to zero, give  $n = 0$  but that is not possible. Therefore, unique values of  $a_0$  and  $a_1$  do not exist.

You may think that the reason the minimization criterion  $\sum_{i=1}^n E_i$  does not work is that negative residuals cancel with positive residuals. So is minimizing  $\sum_{i=1}^n |E_i|$  better? Let us look at the data given in the Table 2 for equation  $y = 4x - 4$ . It makes  $\sum_{i=1}^4 |E_i| = 4$  as shown in the following table.

**Table 4** The absolute residuals at each data point when employing  $y = 4x - 4$ .

$x$	$y$	$y_{predicted}$	$\varepsilon = y - y_{predicted}$
2.0	4.0	4.0	0.0
3.0	6.0	8.0	2.0
2.0	6.0	4.0	2.0
3.0	8.0	8.0	0.0
$\sum_{i=1}^4  \varepsilon_i  = 4$			

The value of  $\sum_{i=1}^4 |E_i| = 4$  also exists for the straight line model  $y = 6$ . No other straight line model for this data has  $\sum_{i=1}^4 |E_i| < 4$ . Again, we find the regression coefficients are not unique, and hence this criterion also cannot be used for finding the regression model.

Let us use the least squares criterion where we minimize

$$S_r = \sum_{i=1}^n E_i^2 = \sum_{i=1}^n (y_i - a_0 - a_1 x_i)^2 \quad (9)$$

$S_r$  is called the sum of the square of the residuals.

To find  $a_0$  and  $a_1$ , we minimize  $S_r$  with respect to  $a_0$  and  $a_1$ .

$$\frac{\partial S_r}{\partial a_0} = 2 \sum_{i=1}^n (y_i - a_0 - a_1 x_i)(-1) = 0 \quad (10)$$

$$\frac{\partial S_r}{\partial a_1} = 2 \sum_{i=1}^n (y_i - a_0 - a_1 x_i)(-x_i) = 0 \quad (11)$$

giving

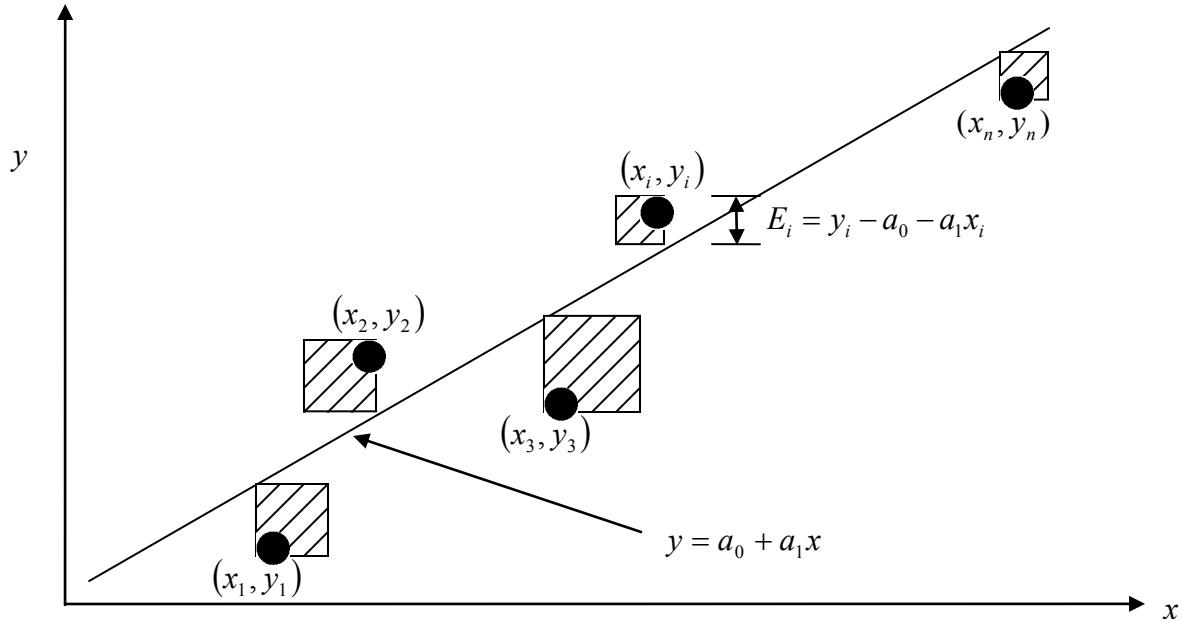
$$-\sum_{i=1}^n y_i + \sum_{i=1}^n a_0 + \sum_{i=1}^n a_1 x_i = 0 \quad (12)$$

$$-\sum_{i=1}^n y_i x_i + \sum_{i=1}^n a_0 x_i + \sum_{i=1}^n a_1 x_i^2 = 0 \quad (13)$$

Noting that  $\sum_{i=1}^n a_0 = a_0 + a_0 + \dots + a_0 = n a_0$

$$n a_0 + a_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i \quad (14)$$

$$a_0 \sum_{i=1}^n x_i + a_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i \quad (15)$$



**Figure 3** Linear regression of  $y$  vs.  $x$  data showing residuals and square of residual at a typical point,  $x_i$ .

Solving the above Equations (14) and (15) gives

$$a_1 = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2} \quad (16)$$

$$a_0 = \frac{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2} \quad (17)$$

Redefining

$$S_{xy} = \sum_{i=1}^n x_i y_i - n \bar{x} \bar{y} \quad (18)$$

$$S_{xx} = \sum_{i=1}^n x_i^2 - n \bar{x}^2 \quad (19)$$

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad (20)$$

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n} \quad (21)$$

we can rewrite

$$a_1 = \frac{S_{xy}}{S_{xx}} \quad (22)$$

$$a_0 = \bar{y} - a_1 \bar{x} \quad (23)$$

### Example 1

The torque  $T$  needed to turn the torsional spring of a mousetrap through an angle,  $\theta$  is given below

**Table 5** Torque versus angle for a torsion spring.

Angle, $\theta$ Radians	Torque, $T$ N · m
0.698132	0.188224
0.959931	0.209138
1.134464	0.230052
1.570796	0.250965
1.919862	0.313707

Find the constants  $k_1$  and  $k_2$  of the regression model

$$T = k_1 + k_2 \theta$$

### Solution

Table 6 shows the summations needed for the calculation of the constants of the regression model.

**Table 6** Tabulation of data for calculation of needed summations.

$i$	$\theta$	$T$	$\theta^2$	$T\theta$
1	radians	N · m	radians <sup>2</sup>	N · m
2	0.698132	0.188224	$4.87388 \times 10^{-1}$	$1.31405 \times 10^{-1}$
3	0.959931	0.209138	$9.21468 \times 10^{-1}$	$2.00758 \times 10^{-1}$
4	1.134464	0.230052	1.2870	$2.60986 \times 10^{-1}$

5	1.570796	0.250965	2.4674	$3.94215 \times 10^{-1}$
6	1.919862	0.313707	3.6859	$6.02274 \times 10^{-1}$
$\sum_{i=1}^5$	<b>6.2831</b>	<b>1.1921</b>	<b>8.8491</b>	<b>1.5896</b>

$$n = 5$$

$$k_2 = \frac{n \sum_{i=1}^5 \theta_i T_i - \sum_{i=1}^5 \theta_i \sum_{i=1}^5 T_i}{n \sum_{i=1}^5 \theta_i^2 - \left( \sum_{i=1}^5 \theta_i \right)^2}$$

$$= \frac{5(1.5896) - (6.2831)(1.1921)}{5(8.8491) - (6.2831)^2}$$

$$= 9.6091 \times 10^{-2} \text{ N - m/rad}$$

$$\bar{T} = \frac{\sum_{i=1}^5 T_i}{n}$$

$$= \frac{1.1921}{5}$$

$$= 2.3842 \times 10^{-1} \text{ N-m}$$

$$\bar{\theta} = \frac{\sum_{i=1}^5 \theta_i}{n}$$

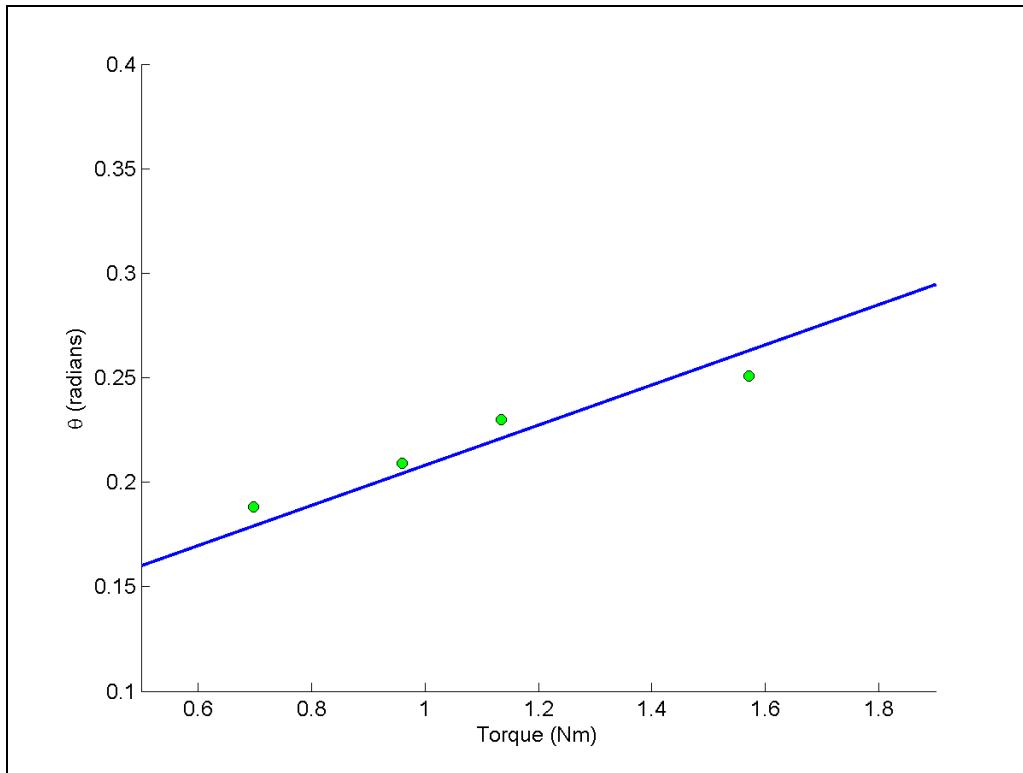
$$= \frac{6.2831}{5}$$

$$= 1.2566 \text{ radians}$$

$$k_1 = \bar{T} - k_2 \bar{\theta}$$

$$= 2.3842 \times 10^{-1} - (9.6091 \times 10^{-2})(1.2566)$$

$$= 1.1767 \times 10^{-1} \text{ N - m}$$



**Figure 4** Linear regression of torque vs. angle data

### Example 2

To find the longitudinal modulus of a composite material, the following data, as given in Table 7, is collected.

**Table 7** Stress vs. strain data for a composite material.

Strain (%)	Stress (MPa)
0	0
0.183	306
0.36	612
0.5324	917
0.702	1223
0.867	1529
1.0244	1835
1.1774	2140
1.329	2446
1.479	2752
1.5	2767
1.56	2896

Find the longitudinal modulus  $E$  using the regression model.

$$\sigma = E\varepsilon \quad (24)$$

### Solution

Rewriting data from Table 7, stresses versus strain data in Table 8

**Table 8** Stress vs strain data for a composite in SI system of units

Strain (m/m)	Stress (Pa)
0.0000	0.0000
$1.8300 \times 10^{-3}$	$3.0600 \times 10^8$
$3.6000 \times 10^{-3}$	$6.1200 \times 10^8$
$5.3240 \times 10^{-3}$	$9.1700 \times 10^8$
$7.0200 \times 10^{-3}$	$1.2230 \times 10^9$
$8.6700 \times 10^{-3}$	$1.5290 \times 10^9$
$1.0244 \times 10^{-2}$	$1.8350 \times 10^9$
$1.1774 \times 10^{-2}$	$2.1400 \times 10^9$
$1.3290 \times 10^{-2}$	$2.4460 \times 10^9$
$1.4790 \times 10^{-2}$	$2.7520 \times 10^9$
$1.5000 \times 10^{-2}$	$2.7670 \times 10^9$
$1.5600 \times 10^{-2}$	$2.8960 \times 10^9$

Applying the least square method, the residuals  $\gamma_i$  at each data point is

$$\gamma_i = \sigma_i - E\varepsilon_i$$

The sum of square of the residuals is

$$\begin{aligned} S_r &= \sum_{i=1}^n \gamma_i^2 \\ &= \sum_{i=1}^n (\sigma_i - E\varepsilon_i)^2 \end{aligned}$$

Again, to find the constant  $E$ , we need to minimize  $S_r$  by differentiating with respect to  $E$  and then equating to zero

$$\frac{dS_r}{dE} = \sum_{i=1}^n 2(\sigma_i - E\varepsilon_i)(-\varepsilon_i) = 0$$

From there, we obtain

$$E = \frac{\sum_{i=1}^n \sigma_i \varepsilon_i}{\sum_{i=1}^n \varepsilon_i^2} \quad (25)$$

Note, Equation (25) only so far has shown that it corresponds to a local minimum or maximum. Can you show that it corresponds to an absolute minimum.

The summations used in Equation (25) are given in the Table 9.

**Table 9** Tabulation for Example 2 for needed summations

$i$	$\varepsilon$	$\sigma$	$\varepsilon^2$	$\varepsilon\sigma$
1	0.0000	0.0000	0.0000	0.0000
2	$1.8300 \times 10^{-3}$	$3.0600 \times 10^8$	$3.3489 \times 10^{-6}$	$5.5998 \times 10^5$
3	$3.6000 \times 10^{-3}$	$6.1200 \times 10^8$	$1.2960 \times 10^{-5}$	$2.2032 \times 10^6$
4	$5.3240 \times 10^{-3}$	$9.1700 \times 10^8$	$2.8345 \times 10^{-5}$	$4.8821 \times 10^6$
5	$7.0200 \times 10^{-3}$	$1.2230 \times 10^9$	$4.9280 \times 10^{-5}$	$8.5855 \times 10^6$
6	$8.6700 \times 10^{-3}$	$1.5290 \times 10^9$	$7.5169 \times 10^{-5}$	$1.3256 \times 10^7$
7	$1.0244 \times 10^{-2}$	$1.8350 \times 10^9$	$1.0494 \times 10^{-4}$	$1.8798 \times 10^7$
8	$1.1774 \times 10^{-2}$	$2.1400 \times 10^9$	$1.3863 \times 10^{-4}$	$2.5196 \times 10^7$
9	$1.3290 \times 10^{-2}$	$2.4460 \times 10^9$	$1.7662 \times 10^{-4}$	$3.2507 \times 10^7$
10	$1.4790 \times 10^{-2}$	$2.7520 \times 10^9$	$2.1874 \times 10^{-4}$	$4.0702 \times 10^7$
11	$1.5000 \times 10^{-2}$	$2.7670 \times 10^9$	$2.2500 \times 10^{-4}$	$4.1505 \times 10^7$
12	$1.5600 \times 10^{-2}$	$2.8960 \times 10^9$	$2.4336 \times 10^{-4}$	$4.5178 \times 10^7$
$\sum_{i=1}^{12}$			$1.2764 \times 10^{-3}$	$2.3337 \times 10^8$

$$n = 12$$

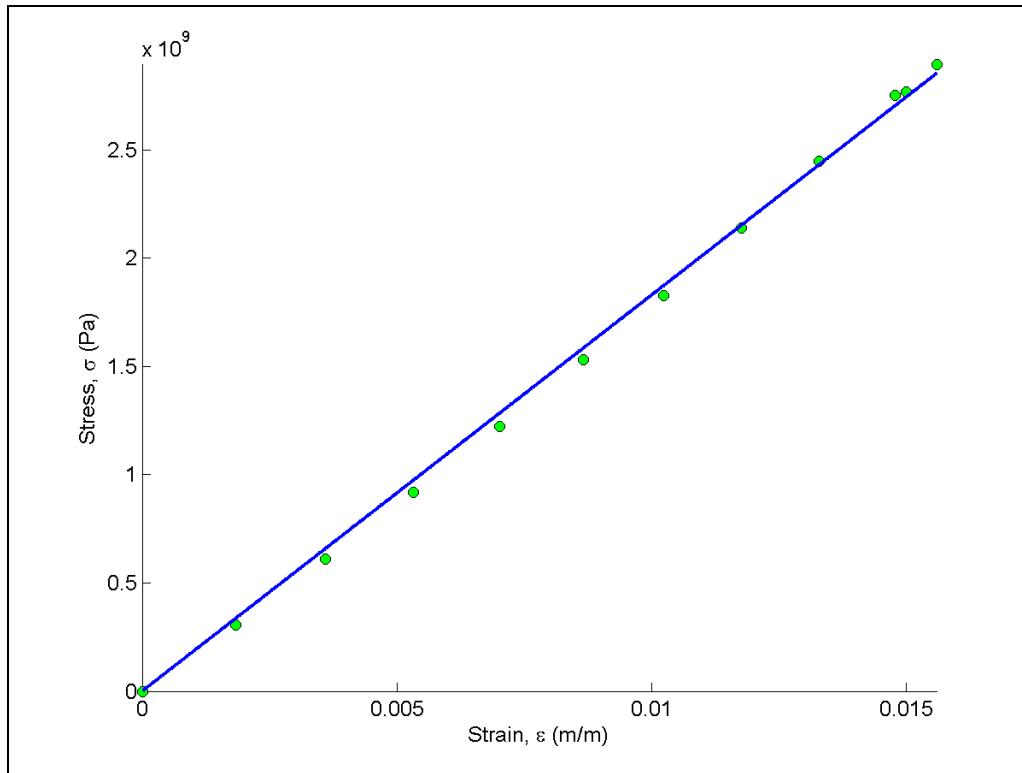
$$\sum_{i=1}^{12} \varepsilon_i^2 = 1.2764 \times 10^{-3}$$

$$\sum_{i=1}^{12} \sigma_i \varepsilon_i = 2.3337 \times 10^8$$

$$E = \frac{\sum_{i=1}^{12} \sigma_i \varepsilon_i}{\sum_{i=1}^{12} \varepsilon_i^2}$$

$$= \frac{2.3337 \times 10^8}{1.2764 \times 10^{-3}}$$

$$= 182.84 \text{ GPa}$$



**Figure 5** Linear regression model of stress vs. strain for a composite material.

**QUESTION:**

Given  $n$  data pairs,  $(x_1, y_1), \dots, (x_n, y_n)$ , do the values of the two constants  $a_0$  and  $a_1$  in the least squares straight-line regression model  $y = a_0 + a_1 x$  correspond to the absolute minimum of the sum of the squares of the residuals? Are these constants of regression unique?

**ANSWER:**

Given  $n$  data pairs  $(x_1, y_1), \dots, (x_n, y_n)$ , the best fit for the straight-line regression model

$$y = a_0 + a_1 x \quad (\text{A.1})$$

is found by the method of least squares. Starting with the sum of the squares of the residuals  $S_r$

$$S_r = \sum_{i=1}^n (y_i - a_0 - a_1 x_i)^2 \quad (\text{A.2})$$

and using

$$\frac{\partial S_r}{\partial a_0} = 0 \quad (\text{A.3})$$

$$\frac{\partial S_r}{\partial a_1} = 0 \quad (\text{A.4})$$

gives two simultaneous linear equations whose solution is

$$a_1 = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2} \quad (\text{A.5a})$$

$$a_0 = \frac{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2} \quad (\text{A.5b})$$

But do these values of  $a_0$  and  $a_1$  give the absolute minimum of value of  $S_r$  (Equation (A.2))? The first derivative analysis only tells us that these values give a local minima or maxima of  $S_r$ , and not whether they give an absolute minimum or maximum. So, we still need to figure out if they correspond to an absolute minimum.

We need to first conduct a second derivative test to find out whether the point  $(a_0, a_1)$  from Equation (A.5) gives a local minimum or local maximum of  $S_r$ . Only then can we proceed to show if this local minimum (or maximum) also corresponds to the absolute minimum (or maximum).

*What is the second derivative test for a local minimum of a function of two variables?*

If you have a function  $f(x, y)$  and we found a critical point  $(a, b)$  from the first derivative test, then  $(a, b)$  is a minimum point if

$$\frac{\partial^2 f}{\partial x^2} \frac{\partial^2 f}{\partial y^2} - \left( \frac{\partial^2 f}{\partial x \partial y} \right)^2 > 0, \text{ and} \quad (\text{A.6})$$

$$\frac{\partial^2 f}{\partial x^2} > 0 \text{ OR } \frac{\partial^2 f}{\partial y^2} > 0 \quad (\text{A.7})$$

From Equation (A.2)

$$\begin{aligned} \frac{\partial S_r}{\partial a_0} &= \sum_{i=1}^n 2(y_i - a_0 - a_1 x_i)(-1) \\ &= -2 \sum_{i=1}^n (y_i - a_0 - a_1 x_i) \end{aligned} \quad (\text{A.8})$$

$$\begin{aligned} \frac{\partial S_r}{\partial a_1} &= \sum_{i=1}^n 2(y_i - a_0 - a_1 x_i)(-x_i) \\ &= -2 \sum_{i=1}^n (x_i y_i - a_0 x_i - a_1 x_i^2) \end{aligned} \quad (\text{A.9})$$

then

$$\frac{\partial^2 S_r}{\partial a_0^2} = -2 \sum_{i=1}^n -1 = 2n \quad (\text{A.10})$$

$$\frac{\partial^2 S_r}{\partial a_1^2} = 2 \sum_{i=1}^n x_i^2 \quad (\text{A.11})$$

$$\frac{\partial^2 S_r}{\partial a_0 \partial a_1} = 2 \sum_{i=1}^n x_i \quad (\text{A.12})$$

So, we satisfy condition (A.7) because from Equation (A.10) we see that  $2n$  is a positive number. Although not required, from Equation (A.11) we see that  $2 \sum_{i=1}^n x_i^2$  is also a positive number as assuming that all  $x$  data points are NOT zero is reasonable.

Is the other condition (Equation (A.6)) for  $S_r$  being a minimum met? Yes, we can show (*proof not given that the term is positive*)

$$\begin{aligned} \frac{\partial^2 S_r}{\partial a_0^2} \frac{\partial^2 S_r}{\partial a_1^2} - \left( \frac{\partial^2 S_r}{\partial a_0 \partial a_1} \right)^2 &= (2n) \left( 2 \sum_{i=1}^n x_i^2 \right) - \left( 2 \sum_{i=1}^n x_i \right)^2 \\ &= 4 \left[ n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2 \right] \\ &= 4 \sum_{\substack{i=1 \\ i < j}}^n (x_i - x_j)^2 > 0 \end{aligned} \quad (\text{A.13})$$

So the values of  $a_0$  and  $a_1$  that we have in Equation (A.5) do correspond to a local minimum of  $S_r$ . But, is this local minimum also an absolute minimum. Yes, as given by Equation (A.5), the first derivatives of  $S_r$  are zero at *only one* point. This observation also makes the straight-line regression model based on least squares to be unique.

As a side note, the denominator in Equations (A.5) is nonzero as shown by Equation (A.13). This shows that the values of  $a_0$  and  $a_1$  are finite.

## LINEAR REGRESSION

Topic	Linear Regression
Summary	Textbook notes of Linear Regression
Major	General Engineering
Authors	Egwu Kalu, Autar Kaw, Cuong Nguyen
Date	August 13, 2012
Web Site	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

## Chapter 06.04

# Nonlinear Models for Regression

After reading this chapter, you should be able to

1. derive constants of nonlinear regression models,
2. use in examples, the derived formula for the constants of the nonlinear regression model, and
3. linearize (transform) data to find constants of some nonlinear regression models.

From fundamental theories, we may know the relationship between two variables. An example in chemical engineering is the Clausius-Clapeyron equation that relates vapor pressure  $P$  of a vapor to its absolute temperature,  $T$ .

$$\log(P) = A + \frac{B}{T} \quad (1)$$

where  $A$  and  $B$  are the unknown parameters to be determined. The above equation is not linear in the unknown parameters. Any model that is not linear in the unknown parameters is described as a nonlinear regression model.

### Nonlinear models using least squares

The development of the least squares estimation for nonlinear models does not generally yield equations that are linear and hence easy to solve. An example of a nonlinear regression model is the exponential model.

#### Exponential model

Given  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ , best fit  $y = ae^{bx}$  to the data. The variables  $a$  and  $b$  are the constants of the exponential model. The residual at each data point  $x_i$  is

$$E_i = y_i - ae^{bx_i} \quad (2)$$

The sum of the square of the residuals is

$$\begin{aligned} S_r &= \sum_{i=1}^n E_i^2 \\ &= \sum_{i=1}^n (y_i - ae^{bx_i})^2 \end{aligned} \quad (3)$$

To find the constants  $a$  and  $b$  of the exponential model, we minimize  $S_r$  by differentiating with respect to  $a$  and  $b$  and equating the resulting equations to zero.

$$\begin{aligned}\frac{\partial S_r}{\partial a} &= \sum_{i=1}^n 2(y_i - ae^{bx_i})(-e^{bx_i}) = 0 \\ \frac{\partial S_r}{\partial b} &= \sum_{i=1}^n 2(y_i - ae^{bx_i})(-ax_i e^{bx_i}) = 0\end{aligned}\quad (4a,b)$$

or

$$\begin{aligned}-\sum_{i=1}^n y_i e^{bx_i} + a \sum_{i=1}^n e^{2bx_i} &= 0 \\ \sum_{i=1}^n y_i x_i e^{bx_i} - a \sum_{i=1}^n x_i e^{2bx_i} &= 0\end{aligned}\quad (5a,b)$$

Equations (5a) and (5b) are nonlinear in  $a$  and  $b$  and thus not in a closed form to be solved as was the case for linear regression. In general, iterative methods (such as Gauss-Newton iteration method, method of steepest descent, Marquardt's method, direct search, etc) must be used to find values of  $a$  and  $b$ .

However, in this case, from Equation (5a),  $a$  can be written explicitly in terms of  $b$  as

$$a = \frac{\sum_{i=1}^n y_i e^{bx_i}}{\sum_{i=1}^n e^{2bx_i}} \quad (6)$$

Substituting Equation (6) in (5b) gives

$$\sum_{i=1}^n y_i x_i e^{bx_i} - \frac{\sum_{i=1}^n y_i e^{bx_i}}{\sum_{i=1}^n e^{2bx_i}} \sum_{i=1}^n x_i e^{2bx_i} = 0 \quad (7)$$

This equation is still a nonlinear equation in  $b$  and can be solved best by numerical methods such as the bisection method or the secant method.

### Example 1

Many patients get concerned when a test involves injection of a radioactive material. For example for scanning a gallbladder, a few drops of Technetium-99m isotope is used. Half of the technetium-99m would be gone in about 6 hours. It, however, takes about 24 hours for the radiation levels to reach what we are exposed to in day-to-day activities. Below is given the relative intensity of radiation as a function of time.

**Table 1** Relative intensity of radiation as a function of time

$t$ (hrs)	0	1	3	5	7	9
$\gamma$	1.000	0.891	0.708	0.562	0.447	0.355

If the level of the relative intensity of radiation is related to time via an exponential formula  $\gamma = Ae^{\lambda t}$ , find

- the value of the regression constants  $A$  and  $\lambda$ ,
- the half-life of Technium-99m, and
- the radiation intensity after 24 hours.

### Solution

a) The value of  $\lambda$  is given by solving the nonlinear Equation (7),

$$f(\lambda) = \sum_{i=1}^n \gamma_i t_i e^{\lambda t_i} - \frac{\sum_{i=1}^n \gamma_i e^{\lambda t_i}}{\sum_{i=1}^n e^{2\lambda t_i}} \sum_{i=1}^n t_i e^{2\lambda t_i} = 0 \quad (8)$$

and then the value of  $A$  from Equation (6),

$$A = \frac{\sum_{i=1}^n \gamma_i e^{\lambda t_i}}{\sum_{i=1}^n e^{2\lambda t_i}} \quad (9)$$

Equation (8) can be solved for  $\lambda$  using bisection method. To estimate the initial guesses, we assume  $\lambda = -0.120$  and  $\lambda = -0.110$ . We need to check whether these values first bracket the root of  $f(\lambda) = 0$ . At  $\lambda = -0.120$ , the table below shows the evaluation of  $f(-0.120)$ .

**Table 2** Summation value for calculation of constants of model

$i$	$t_i$	$\gamma_i$	$\gamma_i t_i e^{\lambda t_i}$	$\gamma_i e^{\lambda t_i}$	$e^{2\lambda t_i}$	$t_i e^{2\lambda t_i}$
1	0	1	0.00000	1.00000	1.00000	0.00000
2	1	0.891	0.79205	0.79205	0.78663	0.78663
3	3	0.708	1.4819	0.49395	0.48675	1.4603
4	5	0.562	1.5422	0.30843	0.30119	1.5060
5	7	0.447	1.3508	0.19297	0.18637	1.3046
6	9	0.355	1.0850	0.12056	0.11533	1.0379
$\sum_{i=1}^6$			6.2501	2.9062	2.8763	6.0954

From Table 2

$$n = 6$$

$$\sum_{i=1}^6 \gamma_i t_i e^{-0.120 t_i} = 6.2501$$

$$\sum_{i=1}^6 \gamma_i e^{-0.120 t_i} = 2.9062$$

$$\sum_{i=1}^6 e^{2(-0.120)t_i} = 2.8763$$

$$\sum_{i=1}^6 t_i e^{2(-0.120)t_i} = 6.0954$$

$$\begin{aligned} f(-0.120) &= (6.2501) - \frac{2.9062}{2.8763}(6.0954) \\ &= 0.091357 \end{aligned}$$

Similarly

$$f(-0.110) = -0.10099$$

Since

$$f(-0.120) \times f(-0.110) < 0,$$

the value of  $\lambda$  falls in the bracket of  $[-0.120, -0.110]$ . The next guess of the root then is

$$\begin{aligned} \lambda &= \frac{-0.120 + (-0.110)}{2} \\ &= -0.115 \end{aligned}$$

Continuing with the bisection method, the root of  $f(\lambda) = 0$  is found as  $\lambda = -0.11508$ . This value of the root was obtained after 20 iterations with an absolute relative approximate error of less than 0.000008%.

From Equation (9),  $A$  can be calculated as

$$\begin{aligned} A &= \frac{\sum_{i=1}^6 \gamma_i e^{\lambda t_i}}{\sum_{i=1}^6 e^{2\lambda t_i}} \\ &= \frac{1 \times e^{-0.11508(0)} + 0.891 \times e^{-0.11508(1)} + 0.708 \times e^{-0.11508(3)} +}{e^{2(-0.11508)(0)} + e^{2(-0.11508)(1)} + e^{2(-0.11508)(3)} +} \\ &\quad \frac{0.562 \times e^{-0.11508(5)} + 0.447 \times e^{-0.11508(7)} + 0.355 \times e^{-0.11508(9)}}{e^{2(-0.11508)(5)} + e^{2(-0.11508)(7)} + e^{2(-0.11508)(9)}} \\ &= \frac{2.9373}{2.9378} \\ &= 0.99983 \end{aligned}$$

The regression formula is hence given by

$$\gamma = 0.99983 e^{-0.11508t}$$

b) Half life of Technetium-99m is when  $\gamma = \frac{1}{2} \gamma \Big|_{t=0}$

$$0.99983 \times e^{-0.11508t} = \frac{1}{2} (0.99983) e^{-0.11508(0)}$$

$$e^{-0.11508t} = 0.5$$

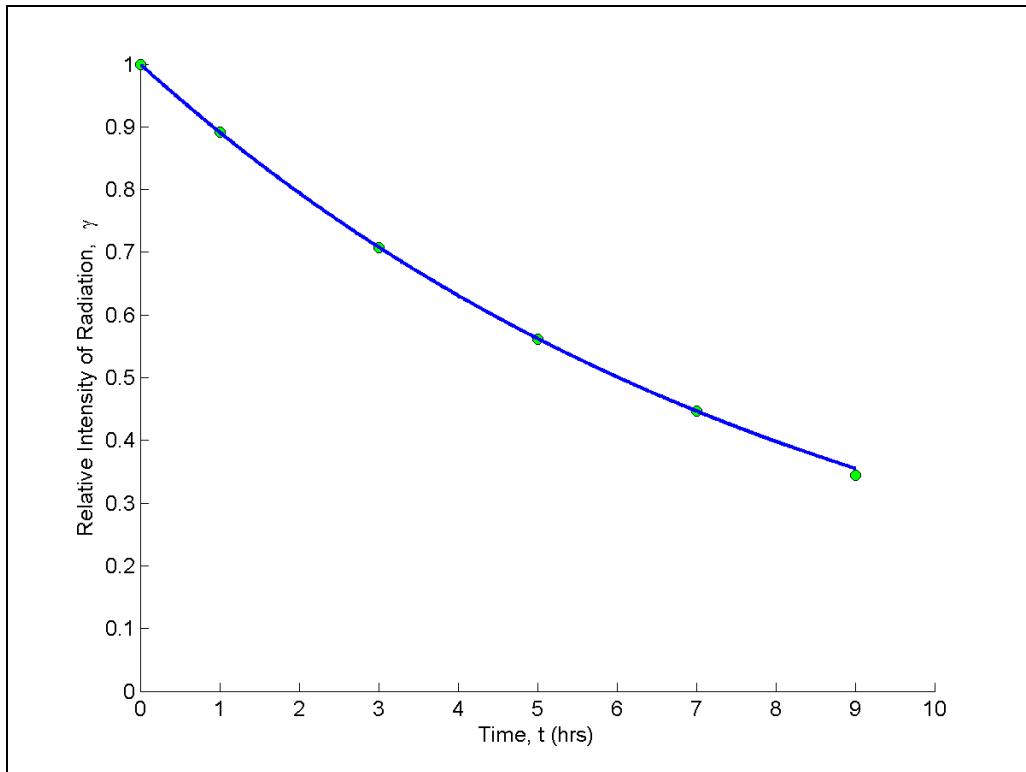
$$-0.11508t = \ln(0.5)$$

$$t = 6.0232 \text{ hours}$$

c) The relative intensity of the radiation after 24 hrs is

$$\begin{aligned}\gamma &= 0.99983 \times e^{-0.11508(24)} \\ &= 6.3160 \times 10^{-2}\end{aligned}$$

This implies that only  $\frac{6.3160 \times 10^{-2}}{0.99983} \times 100 = 6.3171\%$  of the initial radioactive intensity is left after 24 hrs.



**Figure 1** Relative intensity of radiation as a function of temperature using an exponential regression model.

### Growth model

Growth models common in scientific fields have been developed and used successfully for specific situations. The growth models are used to describe how something grows with changes in the regressor variable (often the time). Examples in this category include growth of thin films or population with time. Growth models include

$$y = \frac{a}{1 + be^{-cx}} \quad (10)$$

where  $a, b$  and  $c$  are the constants of the model. At  $x = 0$ ,  $y = \frac{a}{1 + b}$  and as  $x \rightarrow \infty$ ,  $y \rightarrow a$ .

The residuals at each data point  $x_i$ , are

$$E_i = y_i - \frac{a}{1 + be^{-cx_i}} \quad (11)$$

The sum of the square of the residuals is

$$\begin{aligned} S_r &= \sum_{i=1}^n E_i^2 \\ &= \sum_{i=1}^n \left( y_i - \frac{a}{1 + be^{-cx_i}} \right)^2 \end{aligned} \quad (12)$$

To find the constants  $a$ ,  $b$  and  $c$  we minimize  $S_r$  by differentiating with respect to  $a$ ,  $b$  and  $c$ , and equating the resulting equations to zero.

$$\begin{aligned} \frac{\partial S_r}{\partial a} &= \sum_{i=1}^n \left( \frac{2e^{cx_i} [ae^{cx_i} - y_i(e^{cx_i} + b)]}{(e^{cx_i} + b)^2} \right) = 0, \\ \frac{\partial S_r}{\partial b} &= \sum_{i=1}^n \left( \frac{2ae^{cx_i} [by_i + e^{cx_i}(y_i - a)]}{(e^{cx_i} + b)^3} \right) = 0, \\ \frac{\partial S_r}{\partial c} &= \sum_{i=1}^n \left( \frac{-2abx_i e^{cx_i} [by_i + e^{cx_i}(y_i - a)]}{(e^{cx_i} + b)^3} \right) = 0. \end{aligned} \quad (13a,b,c)$$

One can use the Newton-Raphson method to solve the above set of simultaneous nonlinear equations for  $a$ ,  $b$  and  $c$ .

### Example 2

The height of a child is measured at different ages as follows.

**Table 3** Height of the child at different ages.

$t$ (yrs)	0	5.0	8	12	16	18
$H$ (in)	20	36.2	52	60	69.2	70

Estimate the height of the child as an adult of 30 years of age using the growth model,

$$H = \frac{a}{1 + be^{-ct}}$$

### Solution

The saturation growth model of height,  $H$  vs. age,  $t$  is given as

$$H = \frac{a}{1 + be^{-ct}}$$

where the constants  $a$ ,  $b$  and  $c$  are the roots of the simultaneous nonlinear equation system

$$\begin{aligned}
 \sum_{i=1}^6 \left( \frac{2e^{ct_i} [ae^{ct_i} - H_i(e^{ct_i} + b)]}{(e^{ct_i} + b)^2} \right) &= 0 \\
 \sum_{i=1}^6 \left( \frac{2ae^{ct_i} [bH_i + e^{ct_i}(H_i - a)]}{(e^{ct_i} + b)^3} \right) &= 0 \\
 \sum_{i=1}^6 \left( \frac{-2abt_i e^{ct_i} [bH_i + e^{ct_i}(H_i - a)]}{(e^{ct_i} + b)^3} \right) &= 0
 \end{aligned} \tag{14a,b,c}$$

We need initial guesses of the roots to get the iterative process started to find the root of those equations. Suppose we use three of the given data points such as (0, 20), (12, 60) and (18, 70) to find the initial guesses of roots; we have

$$\begin{aligned}
 20 &= \frac{a}{1 + be^{-c(0)}} \\
 60 &= \frac{a}{1 + be^{-c(12)}} \\
 70 &= \frac{a}{1 + be^{-c(18)}}
 \end{aligned}$$

One can solve three unknowns  $a$ ,  $b$  and  $c$  for the initial guesses from the three equations as

$$\begin{aligned}
 a &= 7.5534 \times 10^1 \\
 b &= 2.7767 \\
 c &= 1.9772 \times 10^{-1}
 \end{aligned}$$

Applying the Newton-Raphson method for simultaneous nonlinear equations with the above initial guesses, one can get the roots

$$\begin{aligned}
 a &= 7.4321 \times 10^1 \\
 b &= 2.8233 \\
 c &= 2.1715 \times 10^{-1}
 \end{aligned}$$

The saturation growth model of the height of the child then is

$$H = \frac{7.4321 \times 10^1}{1 + 2.8233e^{-2.1715 \times 10^{-1} t}}$$

The height of the child as an adult of 30 years of age is

$$\begin{aligned}
 H &= \frac{7.4321 \times 10^1}{1 + 2.8233e^{-2.1715 \times 10^{-1} \times (30)}} \\
 &= 74 "
 \end{aligned}$$

### Polynomial Models

Given  $n$  data points  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  use least squares method to regress the data to an  $m^{\text{th}}$  order polynomial.

$$y = a_0 + a_1 x + a_2 x^2 + \dots + a_m x^m, m < n \tag{15}$$

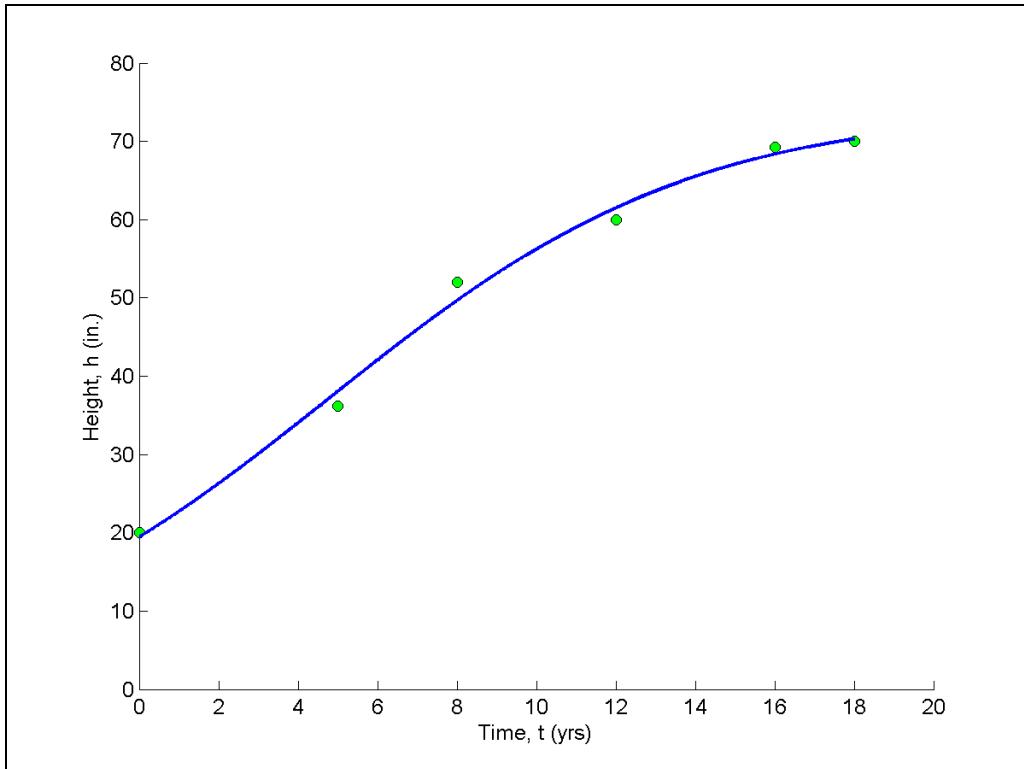
The residual at each data point is given by

$$E_i = y_i - a_0 - a_1 x_i - \dots - a_m x_i^m \tag{16}$$

The sum of the square of the residuals is given by

$$S_r = \sum_{i=1}^n E_i^2 \\ = \sum_{i=1}^n (y_i - a_0 - a_1 x_i - \dots - a_m x_i^m)^2 \quad (17)$$

To find the constants of the polynomial regression model, we put the derivatives with respect to  $a_i$  to zero, that is,



**Figure 2** Height of child as a function of age saturation growth model.

Setting those equations in matrix form gives

$$\begin{bmatrix} n & \left( \sum_{i=1}^n x_i \right) & \dots & \left( \sum_{i=1}^n x_i^m \right) \\ \left( \sum_{i=1}^n x_i \right) & \left( \sum_{i=1}^n x_i^2 \right) & \dots & \left( \sum_{i=1}^n x_i^{m+1} \right) \\ \vdots & \vdots & \ddots & \vdots \\ \left( \sum_{i=1}^n x_i^m \right) & \left( \sum_{i=1}^n x_i^{m+1} \right) & \dots & \left( \sum_{i=1}^n x_i^{2m} \right) \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \\ \vdots \\ \sum_{i=1}^n x_i^m y_i \end{bmatrix} \quad (19)$$

The above are solved for  $a_0, a_1, \dots, a_m$

### Example 3

To find contraction of a steel cylinder, one needs to regress the thermal expansion coefficient data to temperature

**Table 4** The thermal expansion coefficient at given different temperatures

Temperature, $T$ ( $^{\circ}\text{F}$ )	Coefficient of thermal expansion, $\alpha$ (in/in/ $^{\circ}\text{F}$ )
80	$6.47 \times 10^{-6}$
40	$6.24 \times 10^{-6}$
-40	$5.72 \times 10^{-6}$
-120	$5.09 \times 10^{-6}$
-200	$4.30 \times 10^{-6}$
-280	$3.33 \times 10^{-6}$
-340	$2.45 \times 10^{-6}$

Fit the above data to  $\alpha = a_0 + a_1 T + a_2 T^2$

### Solution

Since  $\alpha = a_0 + a_1 T + a_2 T^2$  is the quadratic relationship between the thermal expansion coefficient and the temperature, the coefficients  $a_0, a_1, a_2$  are found as follows

$$\begin{bmatrix} n & \left( \sum_{i=1}^n T_i \right) & \left( \sum_{i=1}^n T_i^2 \right) \\ \left( \sum_{i=1}^n T_i \right) & \left( \sum_{i=1}^n T_i^2 \right) & \left( \sum_{i=1}^n T_i^3 \right) \\ \left( \sum_{i=1}^n T_i^2 \right) & \left( \sum_{i=1}^n T_i^3 \right) & \left( \sum_{i=1}^n T_i^4 \right) \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n \alpha_i \\ \sum_{i=1}^n T_i \alpha_i \\ \sum_{i=1}^n T_i^2 \alpha_i \end{bmatrix}$$

**Table 5** Summations for calculating constants of model

$i$	$T(\text{ }^{\circ}\text{F})$	$\alpha (\text{in/in/ }^{\circ}\text{F})$	$T^2$	$T^3$
1	80	$6.4700 \times 10^{-6}$	$6.4000 \times 10^3$	$5.1200 \times 10^5$
2	40	$6.2400 \times 10^{-6}$	$1.6000 \times 10^3$	$6.4000 \times 10^4$
3	-40	$5.7200 \times 10^{-6}$	$1.6000 \times 10^3$	$-6.4000 \times 10^4$
4	-120	$5.0900 \times 10^{-6}$	$1.4400 \times 10^4$	$-1.7280 \times 10^6$
5	-200	$4.3000 \times 10^{-6}$	$4.0000 \times 10^4$	$-8.0000 \times 10^6$
6	-280	$3.3300 \times 10^{-6}$	$7.8400 \times 10^4$	$-2.1952 \times 10^7$
7	-340	$2.4500 \times 10^{-6}$	$1.1560 \times 10^5$	$-3.9304 \times 10^7$
$\sum_{i=1}^7$	$-8.6000 \times 10^2$	$3.3600 \times 10^{-5}$	$2.5800 \times 10^5$	$-7.0472 \times 10^7$

**Table 5 (cont)**

$i$	$T^4$	$T \times \alpha$	$T^2 \times \alpha$
1	$4.0960 \times 10^7$	$5.1760 \times 10^{-4}$	$4.1408 \times 10^{-2}$
2	$2.5600 \times 10^6$	$2.4960 \times 10^{-4}$	$9.9840 \times 10^{-3}$
3	$2.5600 \times 10^6$	$-2.2880 \times 10^{-4}$	$9.1520 \times 10^{-3}$
4	$2.0736 \times 10^8$	$-6.1080 \times 10^{-4}$	$7.3296 \times 10^{-2}$
5	$1.6000 \times 10^9$	$-8.6000 \times 10^{-4}$	$1.7200 \times 10^{-1}$
6	$6.1466 \times 10^9$	$-9.3240 \times 10^{-4}$	$2.6107 \times 10^{-1}$
7	$1.3363 \times 10^{10}$	$-8.3300 \times 10^{-4}$	$2.8322 \times 10^{-1}$
$\sum_{i=1}^7$	$2.1363 \times 10^{10}$	$-2.6978 \times 10^{-3}$	$8.5013 \times 10^{-1}$

$$n = 7$$

$$\sum_{i=1}^7 T_i = -8.6000 \times 10^{-2}$$

$$\sum_{i=1}^7 T_i^2 = 2.5580 \times 10^5$$

$$\sum_{i=1}^7 T_i^3 = -7.0472 \times 10^7$$

$$\sum_{i=1}^7 T_i^4 = 2.1363 \times 10^{10}$$

$$\sum_{i=1}^7 \alpha_i = 3.3600 \times 10^{-5}$$

$$\sum_{i=1}^7 T_i \alpha_i = -2.6978 \times 10^{-3}$$

$$\sum_{i=1}^7 T_i^2 \alpha_i = 8.5013 \times 10^{-1}$$

We have

$$\begin{bmatrix} 7.0000 & -8.6000 \times 10^2 & 2.5800 \times 10^5 \\ -8.600 \times 10^2 & 2.5800 \times 10^5 & -7.0472 \times 10^7 \\ 2.5800 \times 10^5 & -7.0472 \times 10^7 & 2.1363 \times 10^{10} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 3.3600 \times 10^{-5} \\ -2.6978 \times 10^{-3} \\ 8.5013 \times 10^{-1} \end{bmatrix}$$

Solving the above system of simultaneous linear equations, we get

$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 6.0217 \times 10^{-6} \\ 6.2782 \times 10^{-9} \\ -1.2218 \times 10^{-11} \end{bmatrix}$$

The polynomial regression model is

$$\begin{aligned} \alpha &= a_0 + a_1 T + a_2 T^2 \\ &= 6.0217 \times 10^{-6} + 6.2782 \times 10^{-9} T - 1.2218 \times 10^{-11} T^2 \end{aligned}$$

### Transforming the data to use linear regression formulas

Examination of the nonlinear models above shows that in general iterative methods are required to estimate the values of the model parameters. It is sometimes useful to use simple linear regression formulas to estimate the parameters of a nonlinear model. This involves first transforming the given data such as to regress it to a linear model. Following the transformation of the data, the evaluation of model parameters lends itself to a direct solution approach using the least squares method. Data for nonlinear models such as exponential, power, and growth can be transformed.

#### Exponential Model

As given in Example 1, many physical and chemical processes are governed by the exponential function.

$$\gamma = ae^{bx} \quad (20)$$

Taking natural log of both sides of Equation (20) gives

$$\ln \gamma = \ln a + bx \quad (21)$$

Let

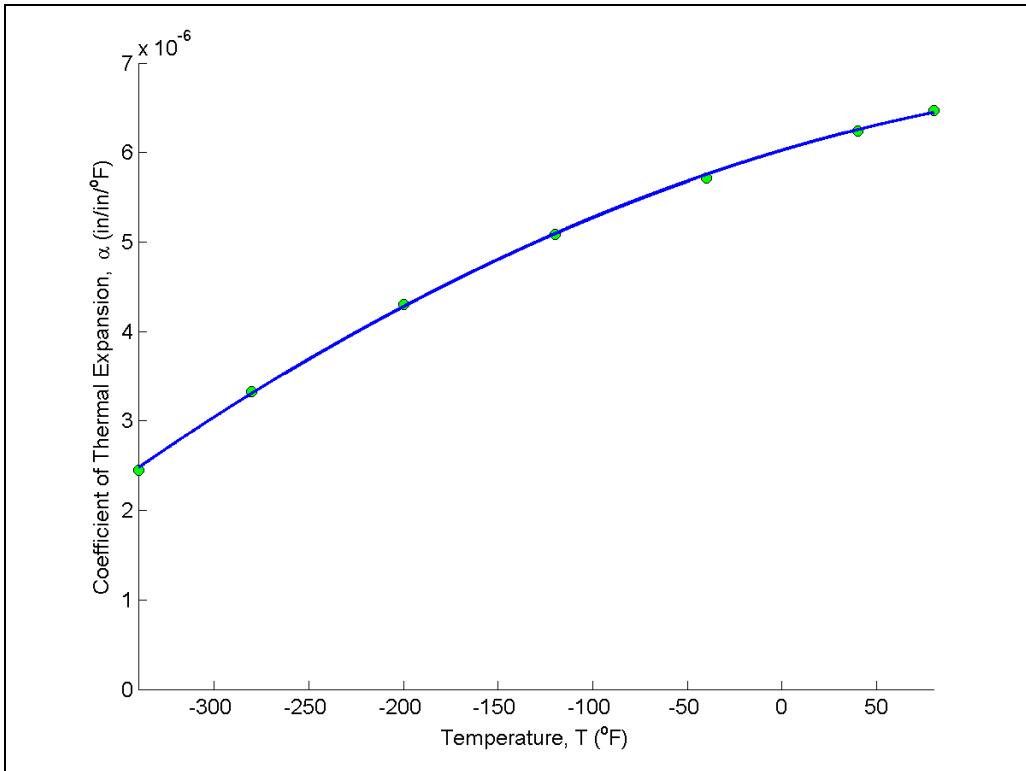
$$z = \ln \gamma$$

$$a_0 = \ln a \text{ implying } a = e^{a_0}$$

$$a_1 = b$$

then

$$z = a_0 + a_1 x \quad (22)$$



**Figure 3** Second-order polynomial regression model for coefficient of thermal expansion as a function of temperature.

The data  $z$  versus  $x$  is now a linear model. The constants  $a_0$  and  $a_1$  can be found using the equation for the linear model as

$$a_1 = \frac{n \sum_{i=1}^n x_i z_i - \sum_{i=1}^n x_i \sum_{i=1}^n z_i}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2} \quad (23a,b)$$

$$a_0 = \bar{z} - a_1 \bar{x}$$

Now since  $a_0$  and  $a_1$  are found, the original constants with the model are found as

$$\begin{aligned} b &= a_1 \\ a &= e^{a_0} \end{aligned} \tag{24a,b}$$

### Example 4

Repeat Example 1 using linearization of data.

### Solution

$$\begin{aligned} \gamma &= Ae^{\lambda t} \\ \ln(\gamma) &= \ln(A) + \lambda t \end{aligned}$$

Assuming

$$\begin{aligned} y &= \ln \gamma \\ a_0 &= \ln(A) \\ a_1 &= \lambda \end{aligned}$$

We get

$$y = a_0 + a_1 t$$

This is a linear relationship between  $y$  and  $t$ .

$$\begin{aligned} a_1 &= \frac{n \sum_{i=1}^n t_i y_i - \sum_{i=1}^n t_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n t_i^2 - \left( \sum_{i=1}^n t_i \right)^2} \\ a_0 &= \bar{y} - a_1 \bar{t} \end{aligned} \tag{25a,b}$$

**Table 6** Summations of data to calculate constants of model.

$i$	$t_i$	$\gamma_i$	$y_i = \ln \gamma_i$	$t_i y_i$	$t_i^2$
1	0	1	0.00000	0.0000	0.0000
2	1	0.891	-0.11541	-0.11541	1.0000
3	3	0.708	-0.34531	-1.0359	9.0000
4	5	0.562	-0.57625	-2.8813	25.000
5	7	0.447	-0.80520	-5.6364	49.000
6	9	0.355	-1.0356	-9.3207	81.000
$\sum_{i=1}^6$	25.000		-2.8778	-18.990	165.00

$$n = 6$$

$$\sum_{i=1}^6 t_i = 25.000$$

$$\sum_{i=1}^6 y_i = -2.8778$$

$$\sum_{i=1}^6 t_i y_i = -18.990$$

$$\sum_{i=1}^6 t_i^2 = 165.00$$

From Equation (25a,b) we have

$$a_1 = \frac{6(-18.990) - (25)(-2.8778)}{6(165.00) - (25)^2} \\ = -0.11505$$

$$a_0 = \frac{-2.8778}{6} - (-0.11505) \frac{25}{6} \\ = -2.6150 \times 10^{-4}$$

Since

$$a_0 = \ln(A)$$

$$A = e^{a_0} \\ = e^{-2.6150 \times 10^{-4}} \\ = 0.99974$$

$$\lambda = a_1 = -0.11505$$

The regression formula then is

$$\gamma = 0.99974 \times e^{-0.11505t}$$

Compare the formula to the one obtained without data linearization,

$$\gamma = 0.99983 \times e^{-0.11508t}$$

b) Half-life is when

$$\gamma = \frac{1}{2} \gamma \Big|_{t=0}$$

$$0.99974 \times e^{-0.11505t} = \frac{1}{2}(0.99974)e^{-0.11505(0)}$$

$$e^{-0.11508t} = 0.5$$

$$-0.11505t = \ln(0.5)$$

$$t = 6.0248 \text{ hours}$$

c) The relative intensity of radiation, after 24 hours is

$$\gamma = 0.99974 e^{-0.11505(24)} \\ = 0.063200$$

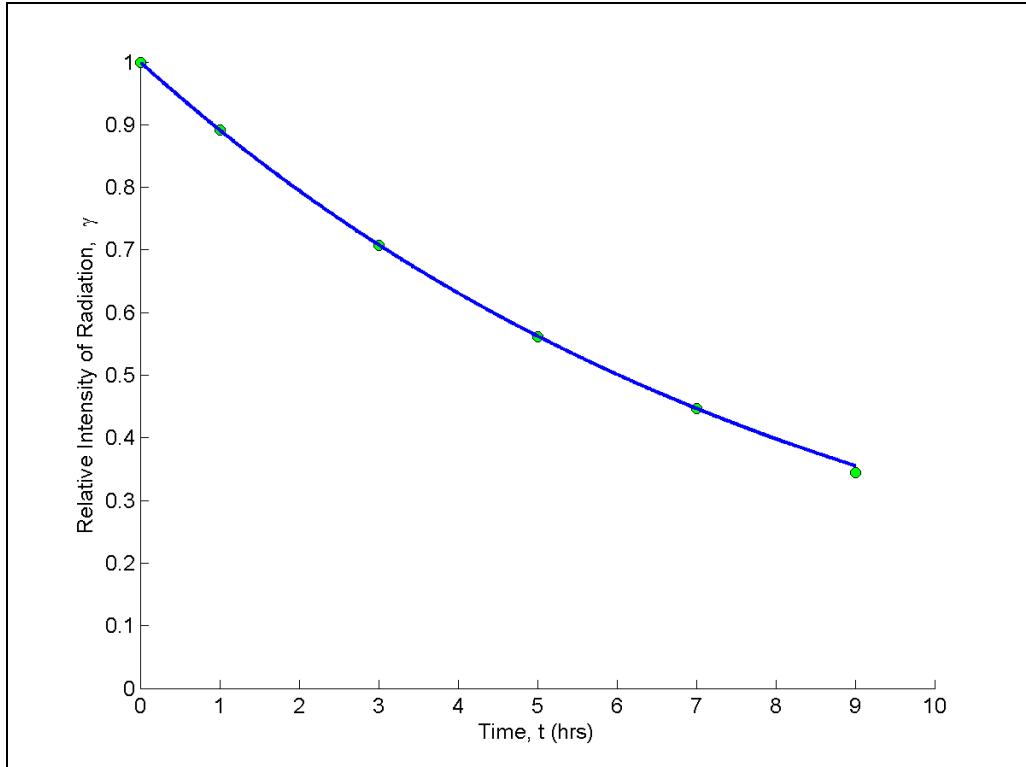
This implies that only  $\frac{6.3200 \times 10^{-2}}{0.99974} \times 100 = 6.3216\%$  of the initial radioactivity is left after 24 hours.

### Logarithmic Functions

The form for the log regression models is

$$y = \beta_0 + \beta_1 \ln(x) \tag{26}$$

This is a linear function between  $y$  and  $\ln(x)$  and the usual least squares method applies in which  $y$  is the response variable and  $\ln(x)$  is the regressor.



**Figure 4** Exponential regression model with transformed data for relative intensity of radiation as a function of temperature.

### Example 5

Sodium borohydride is a potential fuel for fuel cell. The following overpotential ( $\eta$ ) vs. current ( $i$ ) data was obtained in a study conducted to evaluate its electrochemical kinetics.

**Table 7** Electrochemical Kinetics of borohydride data.

$\eta$ (V)	-0.29563	-0.24346	-0.19012	-0.18772	-0.13407	-0.0861
$i$ (A)	0.00226	0.00212	0.00206	0.00202	0.00199	0.00195

At the conditions of the study, it is known that the relationship that exists between the overpotential ( $\eta$ ) and current ( $i$ ) can be expressed as

$$\eta = a + b \ln i \quad (27)$$

where  $a$  is an electrochemical kinetics parameter of borohydride on the electrode. Use the data in Table 7 to evaluate the values of  $a$  and  $b$ .

### Solution

Following the least squares method, Table 8 is tabulated where

$$\begin{aligned}x &= \ln i \\y &= \eta\end{aligned}$$

We obtain

$$y = a + bx \quad (28)$$

This is a linear relationship between  $y$  and  $x$ , and the coefficients  $b$  and  $a$  are found as follow

$$\begin{aligned}b &= \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2} \\a &= \bar{y} - b \bar{x}\end{aligned} \quad (29a,b)$$

**Table 8** Summation values for calculating constants of model

#	$i$	$y = \eta$	$x = \ln(i)$	$x^2$	$x \times y$
1	0.00226	-0.29563	-6.0924	37.117	1.8011
2	0.00212	-0.24346	-6.1563	37.901	1.4988
3	0.00206	-0.19012	-6.1850	38.255	1.1759
4	0.00202	-0.18772	-6.2047	38.498	1.1647
5	0.00199	-0.13407	-6.2196	38.684	0.83386
6	0.00195	-0.08610	-6.2399	38.937	0.53726
$\sum_{i=1}^6$	0.012400	-1.1371	-37.098	229.39	7.0117

$$n = 6$$

$$\sum_{i=1}^6 x_i = -37.098$$

$$\sum_{i=1}^6 y_i = -1.1371$$

$$\sum_{i=1}^6 x_i y_i = 7.0117$$

$$\sum_{i=1}^6 x_i^2 = 229.39$$

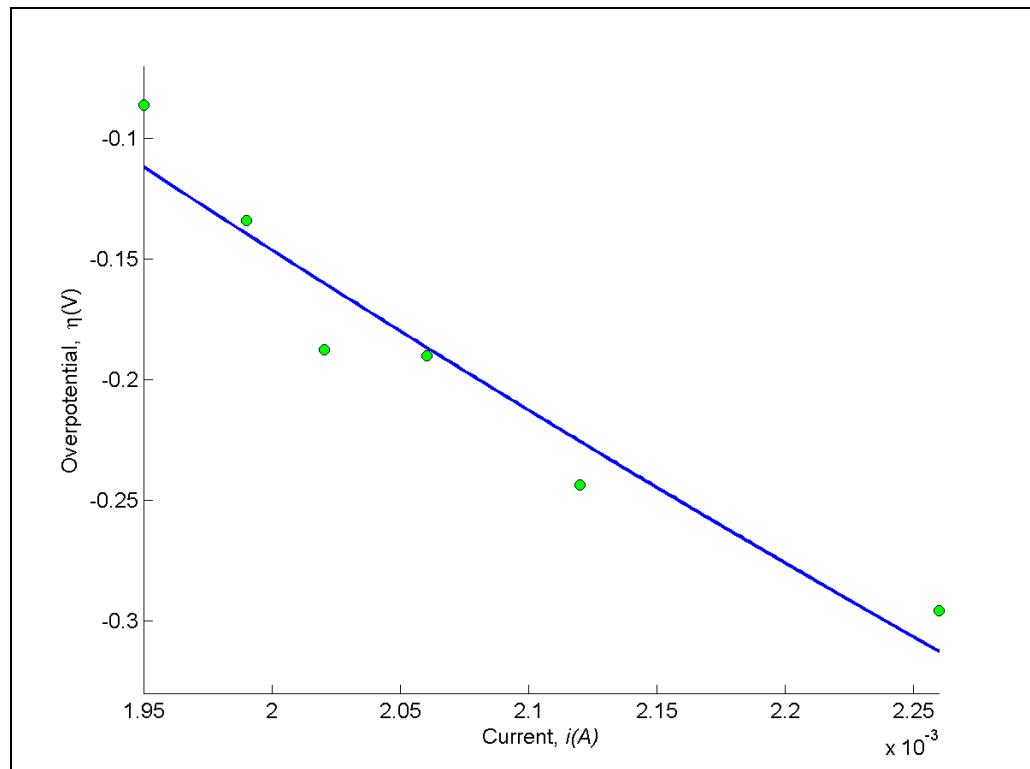
$$b = \frac{6(7.0117) - (-37.098)(-1.1371)}{6(229.39) - (-37.098)^2}$$

$$= -1.3601$$

$$a = \frac{-1.1371}{6} - (-1.3601) \frac{-37.098}{6} \\ = -8.5990$$

Hence

$$\eta = -8.5990 - 1.3601 \times \ln i$$



**Figure 5** Overpotential as a function of current.  $\eta(V)$

### Power Functions

The power function equation describes many scientific and engineering phenomena. In chemical engineering, the rate of chemical reaction is often written in power function form as

$$y = ax^b \quad (30)$$

The method of least squares is applied to the power function by first linearizing the data (the assumption is that  $b$  is not known). If the only unknown is  $a$ , then a linear relation exists between  $x^b$  and  $y$ . The linearization of the data is as follows.

$$\ln(y) = \ln(a) + b \ln(x) \quad (31)$$

The resulting equation shows a linear relation between  $\ln(y)$  and  $\ln(x)$ .

Let

$$z = \ln y$$

$$w = \ln(x)$$

$$a_0 = \ln a \text{ implying } a = e^{a_0}$$

$$a_1 = b$$

we get

$$z = a_0 + a_1 w \quad (32)$$

$$a_1 = \frac{n \sum_{i=1}^n w_i z_i - \sum_{i=1}^n w_i \sum_{i=1}^n z_i}{n \sum_{i=1}^n w_i^2 - \left( \sum_{i=1}^n w_i \right)^2} \quad (33a,b)$$

$$a_0 = \frac{\sum_{i=1}^n z_i}{n} - a_1 \frac{\sum_{i=1}^n w_i}{n}$$

Since  $a_0$  and  $a_1$  can be found, the original constants of the model are

$$\begin{aligned} b &= a_1 \\ a &= e^{a_0} \end{aligned} \quad (34a,b)$$

### Example 6

The progress of a homogeneous chemical reaction is followed and it is desired to evaluate the rate constant and the order of the reaction. The rate law expression for the reaction is known to follow the power function form

$$-r = kC^n \quad (35)$$

Use the data provided in the table to obtain  $n$  and  $k$ .

**Table 9** Chemical kinetics.

$C_A$ (gmol/l)	4	2.25	1.45	1.0	0.65	0.25	0.006
$-r_A$ (gmol/l·s)	0.398	0.298	0.238	0.198	0.158	0.098	0.048

### Solution

Taking the natural log of both sides of Equation (35), we obtain

$$\ln(-r) = \ln(k) + n \ln(C)$$

Let

$$z = \ln(-r)$$

$$w = \ln(C)$$

$$a_0 = \ln(k) \text{ implying that } k = e^{a_0} \quad (36)$$

$$a_1 = n \quad (37)$$

We get

$$z = a_0 + a_1 w$$

This is a linear relation between  $z$  and  $w$ , where

$$a_1 = \frac{n \sum_{i=1}^n w_i z_i - \sum_{i=1}^n w_i \sum_{i=1}^n z_i}{n \sum_{i=1}^n w_i^2 - \left( \sum_{i=1}^n w_i \right)^2}$$

$$a_0 = \left( \frac{\sum_{i=1}^n z_i}{n} \right) - a_1 \left( \frac{\sum_{i=1}^n w_i}{n} \right) \quad (38a,b)$$

**Table 10** Kinetics rate law using power function

$i$	$C$	$-r$	$w$	$z$	$w \times z$	$w^2$
1	4	0.398	1.3863	-0.92130	-1.2772	1.9218
2	2.25	0.298	0.8109	-1.2107	-0.9818	0.65761
3	1.45	0.238	0.3716	-1.4355	-0.5334	0.13806
4	1	0.198	0.0000	-1.6195	0.0000	0.00000
5	0.65	0.158	-0.4308	-1.8452	0.7949	0.18557
6	0.25	0.098	-1.3863	-2.3228	3.2201	1.9218
7	0.006	0.048	-5.1160	-3.0366	15.535	26.173
$\sum_{i=1}^7$			-4.3643	-12.391	16.758	30.998

$$n = 7$$

$$\sum_{i=1}^7 w_i = -4.3643$$

$$\sum_{i=1}^7 z_i = -12.391$$

$$\sum_{i=1}^7 w_i z_i = 16.758$$

$$\sum_{i=1}^7 w_i^2 = 30.998$$

From Equation (38a,b)

$$a_1 = \frac{7 \times (16.758) - (-4.3643) \times (-12.391)}{7 \times (30.998) - (-4.3643)^2}$$

$$= 0.31943$$

$$a_0 = \frac{-12.391}{7} - (0.31943) \frac{-4.3643}{7} \\ = -1.5711$$

From Equation (36) and (37), we obtain

$$k = e^{-1.5711}$$

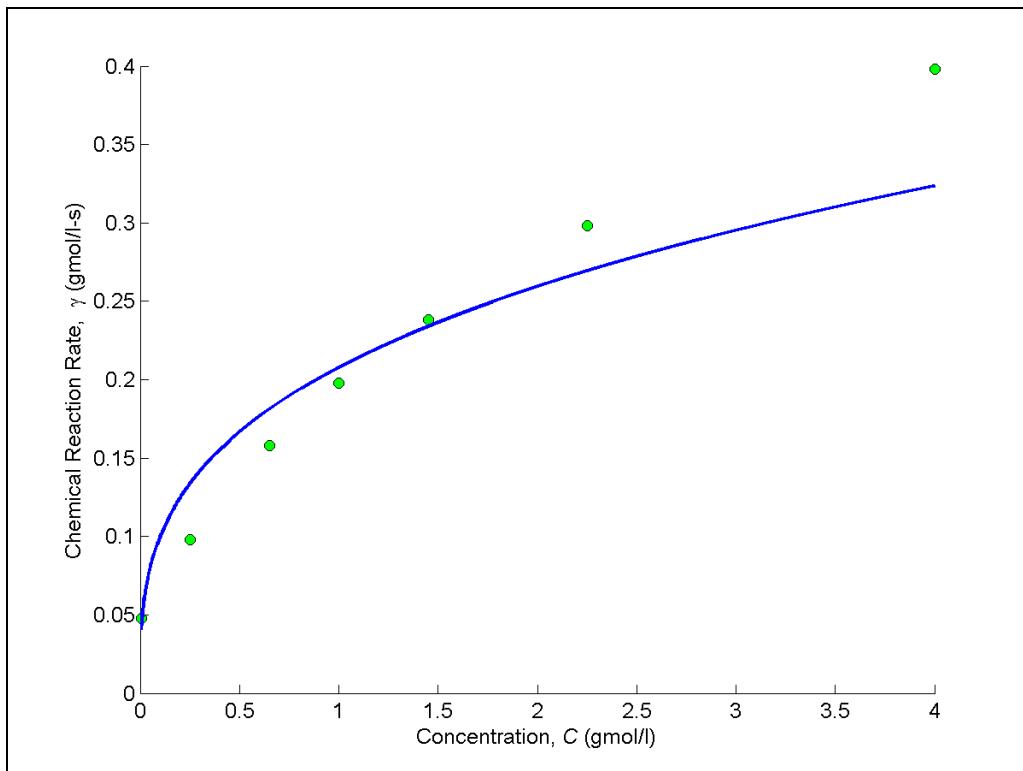
$$= 0.20782$$

$$n = a_1$$

$$= 0.31941$$

Finally, the model of progress of that chemical reaction is

$$-r = 0.20782 \times C^{0.31941}$$



**Figure 6** Kinetic chemical reaction rate as a function of concentration.

### Growth Model

Growth models common in scientific fields have been developed and used successfully for specific situations. The growth models are used to describe how something grows with changes in a regressor variable (often the time). Examples in this category include growth of thin films or population with time. In the logistic growth model, an example of a growth model in which a measurable quantity  $y$  varies with some quantity  $x$  is

$$y = \frac{ax}{b+x} \tag{39}$$

For  $x = 0$ ,  $y = 0$  while as  $x \rightarrow \infty$ ,  $y \rightarrow a$ . To linearize the data for this method,

$$\begin{aligned} \frac{1}{y} &= \frac{b+x}{ax} \\ &= \frac{b}{a} \frac{1}{x} + \frac{1}{a} \end{aligned} \quad (40)$$

Let

$$\begin{aligned} z &= \frac{1}{y} \\ w &= \frac{1}{x}, \\ a_0 &= \frac{1}{a} \text{ implying that } a = \frac{1}{a_0} \\ a_1 &= \frac{b}{a} \text{ implying } b = a_1 \times a = \frac{a_1}{a_0} \end{aligned}$$

Then

$$z = a_0 + a_1 w \quad (41)$$

The relationship between  $z$  and  $w$  is linear with the coefficients  $a_0$  and found as follows.

$$\begin{aligned} a_1 &= \frac{n \sum_{i=1}^n w_i z_i - \sum_{i=1}^n w_i \sum_{i=1}^n z_i}{n \sum_{i=1}^n w_i^2 - \left( \sum_{i=1}^n w_i \right)^2} \\ a_0 &= \left( \frac{\sum_{i=1}^n z_i}{n} \right) - a_1 \left( \frac{\sum_{i=1}^n w_i}{n} \right) \end{aligned} \quad (42a,b)$$

Finding  $a_0$  and  $a_1$ , then gives the constants of the original growth model as

$$\begin{aligned} a &= \frac{1}{a_0} \\ b &= \frac{a_1}{a_0} \end{aligned} \quad (43a,b)$$

---

## NONLINEAR REGRESSION

---

Topic Nonlinear Regression  
Summary Textbook notes of Nonlinear Regression  
Major General Engineering  
Authors Egwu Kalu, Autar Kaw, Cuong Nguyen  
Date June 17, 2015  
Web Site <http://numericalmethods.eng.usf.edu>

---

# Chapter 06.05

## Adequacy of Models for Regression

After reading this chapter, you should be able to

1. determine if a linear regression model is adequate
2. determine how well the linear regression model predicts the response variable.

### Quality of Fitted Model

In the application of regression models, one objective is to obtain an equation  $y = f(x)$  that best describes the  $n$  response data points  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ . Consequently, we are faced with answering two basic questions.

1. Does the model  $y = f(x)$  describe the data adequately, that is, is there an adequate fit?
2. How well does the model predict the response variable (predictability)?

To answer these questions, let us limit our discussion to straight line models as nonlinear models require a different approach. Some authors [1] claim that nonlinear model parameters are not unbiased.

To exemplify our discussion, we will take example data to go through the process of model evaluation. Given below is the data for the coefficient of thermal expansion vs. temperature for steel. We assume a linear relationship between the data as

$$\alpha(T) = a_0 + a_1 T$$

**Table 1** Values of coefficient of thermal expansion vs. temperature.

$T(^{\circ}\text{F})$	$\alpha (\mu\text{in/in}/^{\circ}\text{F})$
-340	2.45
-260	3.58
-180	4.52
-100	5.28
-20	5.86
60	6.36

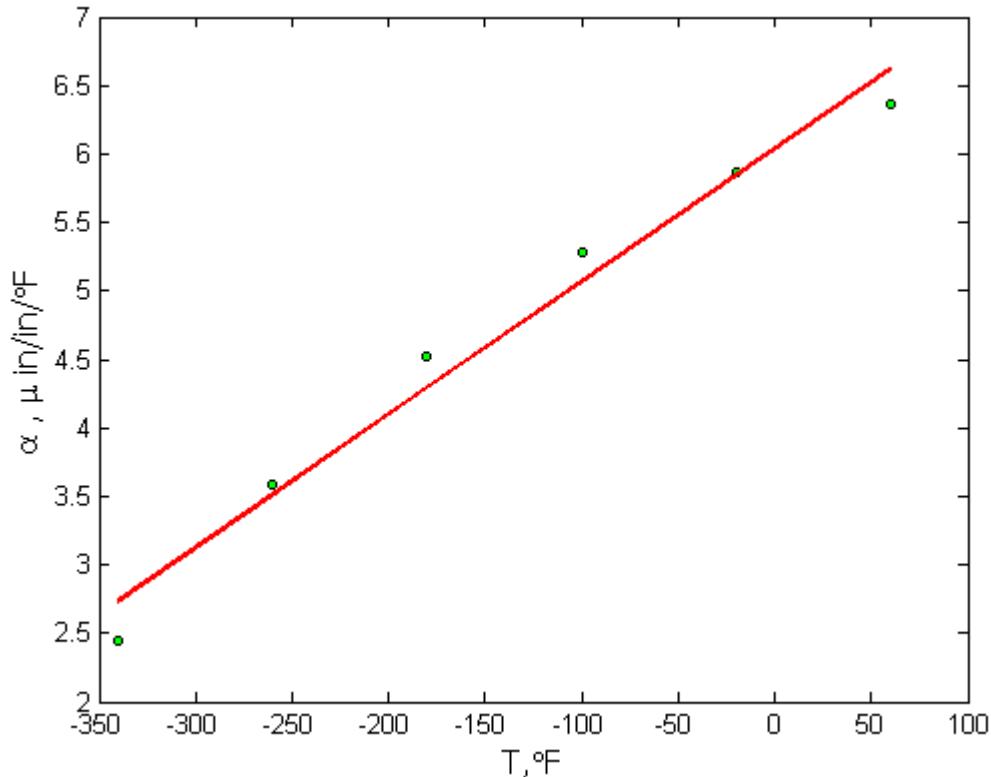
Following the procedure for conducting linear regression as given in Chapter 06.03, we get

$$\alpha(T) = 6.0325 + 0.0096964T$$

Let us now look at how we can evaluate the adequacy of a linear regression model.

### 1. Plot the data and the regression model.

Figure 1 shows the data and the regression model. From a visual check, it looks like the model explains the data adequately.



**Figure 1** Plot of coefficient of thermal expansion vs. temperature data points and regression line.

## 2. Calculate the standard error of estimate.

The standard error of estimate is defined as

$$s_{\alpha/T} = \sqrt{\frac{S_r}{n-2}}$$

where

$$S_r = \sum_{i=1}^n (\alpha_i - a_0 - a_1 T_i)^2$$

**Table 2** Residuals for data.

$T_i$	$\alpha_i$	$a_0 + a_1 T_i$	$\alpha_i - a_0 - a_1 T_i$
-340	2.45	2.7357	-0.28571
-260	3.58	3.5114	0.068571
-180	4.52	4.2871	0.23286
-100	5.28	5.0629	0.21714
20	5.86	5.8386	0.021429
60	6.36	6.6143	-0.25429

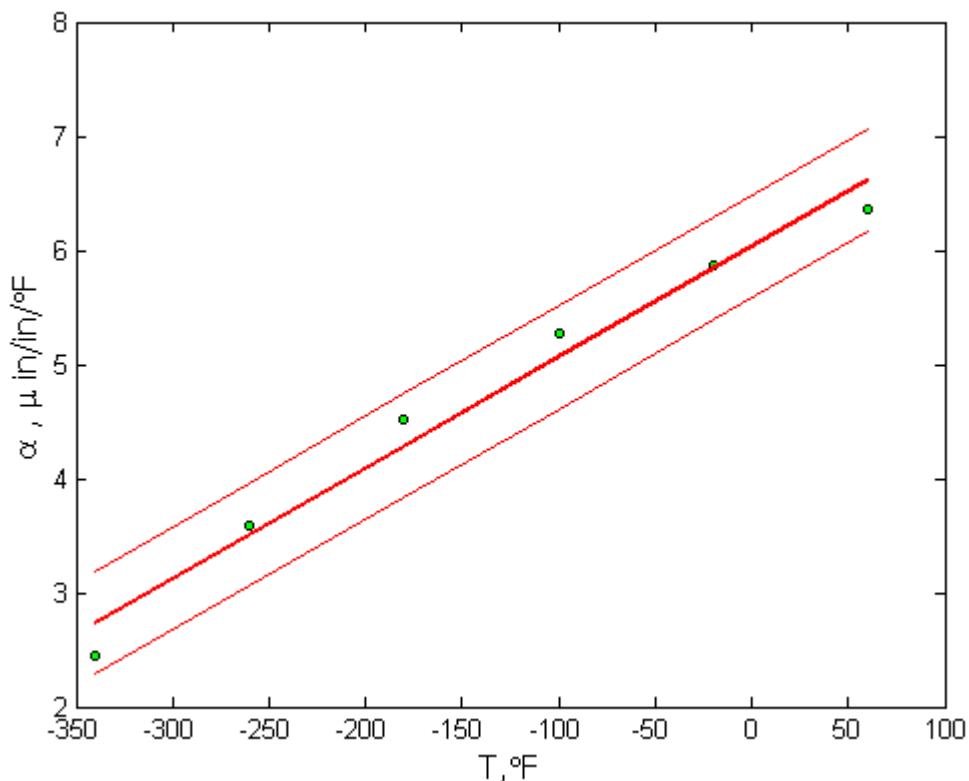
Table 2 shows the residuals of the data to calculate the sum of the square of residuals as

$$\begin{aligned} S_r &= (-0.28571)^2 + (0.068571)^2 + (0.23286)^2 + (0.21714)^2 \\ &\quad + (0.021429)^2 + (-0.25429)^2 \\ &= 0.25283 \end{aligned}$$

The standard error of estimate

$$\begin{aligned} s_{\alpha/T} &= \sqrt{\frac{S_r}{n-2}} \\ &= \sqrt{\frac{0.25283}{6-2}} \\ &= 0.25141 \end{aligned}$$

The units of  $s_{\alpha/T}$  are same as the units of  $\alpha$ . How is the value of the standard error of estimate interpreted? We may say that on average the difference between the observed and predicted values is  $0.25141 \mu\text{in/in/}^{\circ}\text{F}$ . Also, we can look at the value as follows. About 95% of the observed  $\alpha$  values are between  $\pm 2s_{\alpha/T}$  of the predicted value (see Figure 2). This would lead us to believe that the value of  $\alpha$  in the example is expected to be accurate within  $\pm 2s_{\alpha/T} = \pm 2 \times 0.25141 = \pm 0.50282 \mu\text{in/in/}^{\circ}\text{F}$ .



**Figure 2** Plotting the linear regression line and showing the regression standard error.

One can also look at this criterion as finding if 95% of the scaled residuals for the model are in the domain [-2,2], that is

$$\text{Scaled residual} = \frac{\alpha_i - a_0 - a_1 T_i}{s_{\alpha/T}}$$

For the example,

$$s_{\alpha/T} = 0.25141$$

**Table 4** Residuals and scaled residuals for data.

$T_i$	$\alpha_i$	$\alpha_i - a_0 - a_1 T_i$	Scaled Residuals
-340	2.45	-0.28571	-1.1364
-260	3.58	0.068571	0.27275
-180	4.52	0.23286	0.92622
-100	5.28	0.21714	0.86369
-20	5.86	0.021429	0.085235
60	6.36	-0.25429	-1.0115

and the scaled residuals are calculated in Table 4. All the scaled residuals are in the [-2,2] domain.

### 3. Calculate the coefficient of determination.

Denoted by  $r^2$ , the coefficient of determination is another criterion to use for checking the adequacy of the model.

To answer the above questions, let us start from the examination of some measures of discrepancies between the whole data and some key central tendency. Look at the two equations given below.

$$S_r = \sum_{i=1}^n (\alpha_i - \hat{\alpha}_i)^2 \quad (1)$$

$$= \sum_{i=1}^n (\alpha_i - a_0 - a_1 T_i)^2 \quad (2)$$

where

$$\bar{\alpha} = \frac{\sum_{i=1}^n \alpha_i}{n}$$

For the example data

$$\begin{aligned} \bar{\alpha} &= \frac{\sum_{i=1}^6 \alpha_i}{6} \\ &= \frac{2.45 + 3.58 + 4.52 + 5.28 + 5.86 + 6.36}{6} \\ &= 4.6750 \mu\text{in/in}^{\circ}\text{F} \end{aligned}$$

$$\begin{aligned}
 S_t &= \sum_{i=1}^n (\alpha_i - \bar{\alpha})^2 \\
 &= (-2.2250)^2 + (-1.0950)^2 + (-0.15500)^2 + (0.60500)^2 + (1.1850)^2 + (1.6850)^2 \\
 &= 10.783
 \end{aligned}$$

**Table 5** Difference between observed and average value.

$T_i$	$\alpha_i$	$\alpha_i - \bar{\alpha}$
-340	2.45	-2.2250
-260	3.58	-1.0950
-180	4.52	-0.15500
-100	5.28	0.60500
-20	5.86	1.1850
60	6.36	1.6850

where  $S_r$  is the sum of the square of the residuals (residual is the difference between the observed value and the predicted value), and  $S_t$  is the sum of the square of the difference between the observed value and the average value.

What inferences can we make about the two equations? Equation (2) measures the discrepancy between the data and the mean. Recall that the mean of the data is a measure of a single point that measures the central tendency of the whole data. Equation (2) contrasts with Equation (1) as Equation (1) measures the discrepancy between the vertical distance of the point from the regression line (another measure of central tendency). This line obtained by the least squares method gives the best estimate of a line with least sum of deviation.  $S_r$  as calculated quantifies the spread around the regression line.

The objective of least squares method is to obtain a compact equation that best describes all the data points. The mean can also be used to describe all the data points. The magnitude of the sum of squares of deviation from the mean or from the least squares line **is therefore a good indicator of how well the mean or least squares characterizes the whole data.** We can liken the sum of squares deviation around the mean,  $S_t$  as the error or variability in  $y$  without considering the regression variable  $x$ , while  $S_r$ , the sum of squares deviation around the least square regression line is error or variability in  $y$  remaining after the dependent variable  $x$  has been considered.

The difference between these two parameters measures the error due to describing or characterizing the data in one form instead of the other. A relative comparison of this difference ( $S_t - S_r$ ), with the sum of squares deviation associated with the mean  $S_t$  describes a quantity called **coefficient of determination**,  $r^2$

$$\begin{aligned}
 r^2 &= \frac{S_t - S_r}{S_t} \\
 &= \frac{10.783 - 0.25283}{10.783} \\
 &= 0.97655
 \end{aligned} \tag{5}$$

Based on the value obtained above, we can claim that 97.7% of the original uncertainty in the value of  $\alpha$  can be explained by the straight-line regression model of  $\alpha(T) = 6.0325 + 0.0096964T$ .

Going back to the definition of the coefficient of determination, one can see that  $S_t$  is the variation without any relationship of  $y$  vs.  $x$ , while  $S_r$  is the variation with the straight-line relationship.

The limits of the values of  $r^2$  are between 0 and 1. What do these limiting values of  $r^2$  mean? If  $r^2 = 0$ , then  $S_t = S_r$ , which means that regressing the data to a straight line does nothing to explain the data any further. If  $r^2 = 1$ , then  $S_r = 0$ , which means that the straight line is passing through all the data points and is a perfect fit.

### *Caution in the use of $r^2$*

- a) The coefficient of determination,  $r^2$  can be made larger (assumes no collinear points) by adding more terms to the model. For instance,  $n - 1$  terms in a regression equation for which  $n$  data points are used will give an  $r^2$  value of 1 if there are no collinear points.
- b) The magnitude of  $r^2$  also depends on the range of variability of the regressor ( $x$ ) variable. Increase in the spread of  $x$  increases  $r^2$  while a decrease in the spread of  $x$  decreases  $r^2$ .
- c) Large regression slope will also yield artificially high  $r^2$ .
- d) The coefficient of determination,  $r^2$  does not measure the appropriateness of the linear model.  $r^2$  may be large for nonlinearly related  $x$  and  $y$  values.
- e) Large coefficient of determination  $r^2$  value does not necessarily imply the regression will predict accurately.
- f) The coefficient of determination,  $r^2$  does not measure the magnitude of the regression slope.
- g) These statements above imply that one should not choose a regression model solely based on consideration of  $r^2$ .

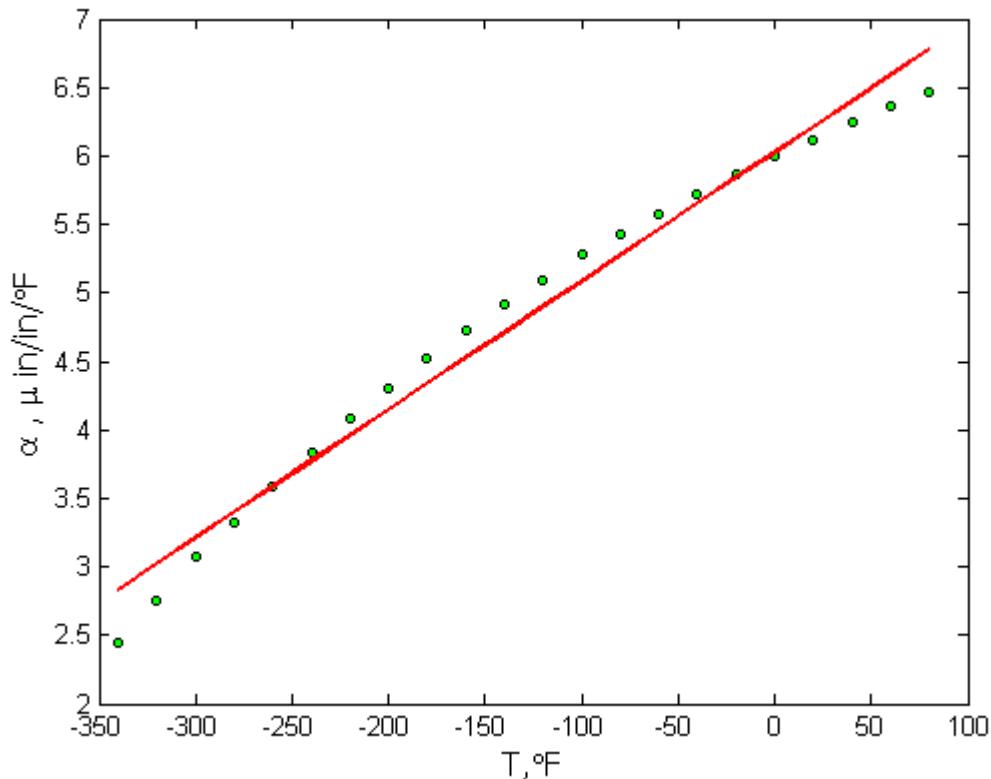
#### **4. Find if the model meets the assumptions of random errors.**

These assumptions include that the residuals are negative as well as positive to give a mean of zero, the variation of the residuals as a function of the independent variable is random, the residuals follow a normal distribution, and that there is no auto correlation between the data points.

To illustrate this better, we have an extended data set for the example that we took. Instead of 6 data points, this set has 22 data points (Table 6). Drawing conclusions from small or large data sets for checking assumption of random error is not recommended.

**Table 6** Instantaneous thermal expansion coefficient as a function of temperature.

Temperature °F	Instantaneous Thermal Expansion μin/in/°F
80	6.47
60	6.36
40	6.24
20	6.12
0	6.00
-20	5.86
-40	5.72
-60	5.58
-80	5.43
-100	5.28
-120	5.09
-140	4.91
-160	4.72
-180	4.52
-200	4.30
-220	4.08
-240	3.83
-260	3.58
-280	3.33
-300	3.07
-320	2.76
-340	2.45



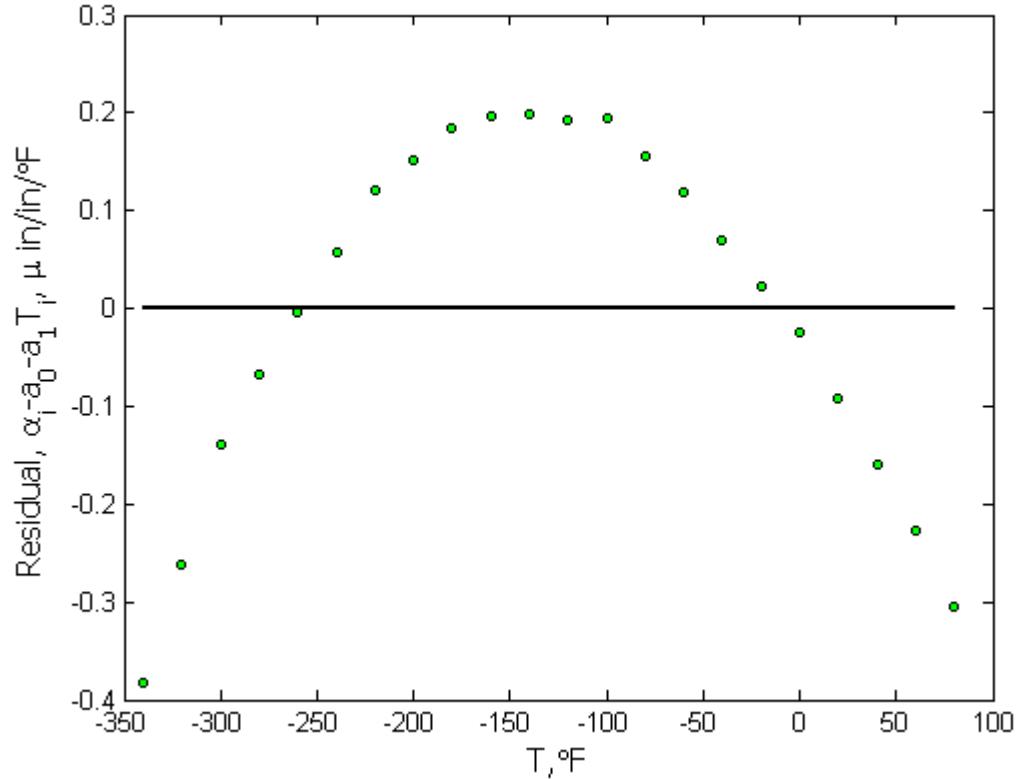
**Figure 3** Plot of thermal expansion coefficient vs. temperature data points and regression line for more data points.

Regressing the data from Table 2 to the straight line regression line

$$\alpha(T) = a_0 + a_1 T$$

and following the procedure for conducting linear regression as given in Chapter 06.03, we get (Figure 3)

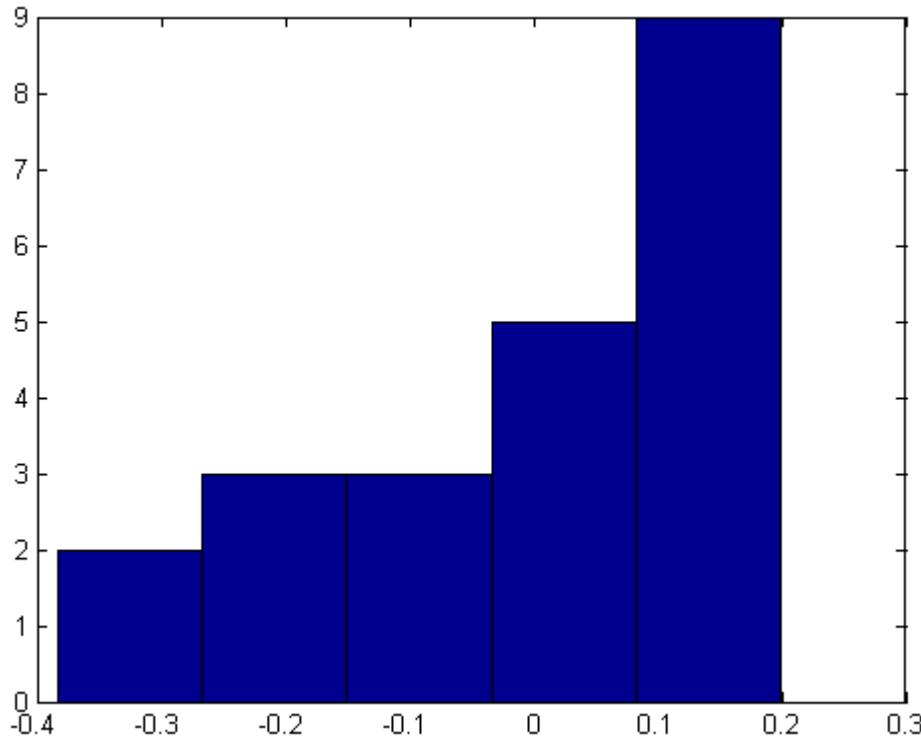
$$\alpha = 6.0248 + 0.0093868T$$



**Figure 4** Plot of residuals.

Figure 4 shows the residuals for the example as a function of temperature. Although the residuals seem to average to zero, but within a range, they do not exhibit this zero mean. For an initial value of  $T$ , the averages are below zero. For the middle values of  $T$ , the averages are below zero, and again for the final values of  $T$ , the averages are below zero. This may be considered a violation of the model assumption.

Figure 4 also shows the residuals for the example are following a nonlinear variance. This is a clear violation of the model assumption of constant variance.



**Figure 5** Histogram of residuals.

Figure 5 shows the histogram of the residuals. Clearly, the histogram is not showing a normal distribution, and hence violates the model assumption of normality.

To check that there is no autocorrelation between observed values, the following rule of thumb can be used. If  $n$  is the number of data points, and  $q$  is the number of times the sign of the residual changes, then if

$$\frac{(n-1)}{2} - \sqrt{n-1} \leq q \leq \frac{n-1}{2} + \sqrt{n-1},$$

you most likely do not have an autocorrelation. For the example,  $n = 22$ , then

$$\frac{(22-1)}{2} - \sqrt{22-1} \leq q \leq \frac{22-1}{2} + \sqrt{22-1}$$

$$5.9174 \leq q \leq 15.083$$

is not satisfied as  $q = 2$ . So this model assumption is violated.

## References

---

### ADEQUACY OF REGRESSION MODELS

---

Topic	Adequacy of Regression Models
Summary	Textbook notes of Adequacy of Regression Models
Major	General Engineering

---

---

Authors      Autar Kaw, Egwu Kalu  
Date          May 31, 2013  
Web Site     <http://numericalmethods.eng.usf.edu>

---

# Chapter 07.01

## Primer on Integration

After reading this chapter, you should be able to:

1. define an integral,
2. use Riemann's sum to approximately calculate integrals,
3. use Riemann's sum and its limit to find the exact expression of integrals, and
4. find exact integrals of different functions such as polynomials, trigonometric function and transcendental functions.

### What is integration?

The dictionary definition of *integration* is combining parts so that they work together or form a whole. Mathematically, integration stands for finding the area under a curve from one point to another. It is represented by

$$\int_a^b f(x)dx$$

where the symbol  $\int$  is an integral sign, and  $a$  and  $b$  are the lower and upper limits of integration, respectively, the function  $f$  is the integrand of the integral, and  $x$  is the variable of integration. Figure 1 represents a graphical demonstration of the concept.

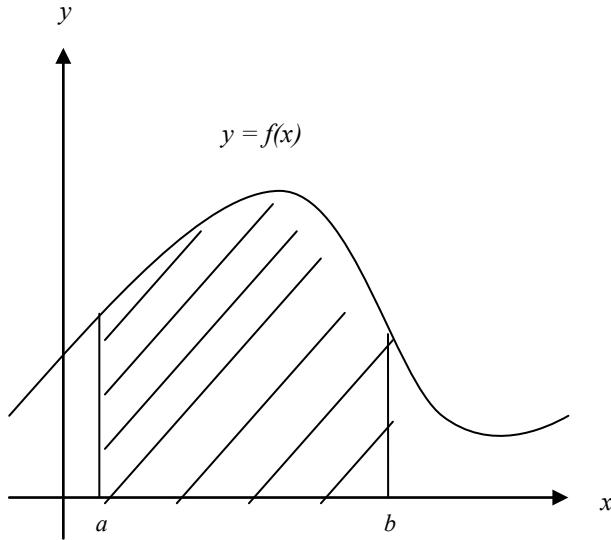
### Riemann Sum

Let  $f$  be defined on the closed interval  $[a,b]$ , and let  $\Delta$  be an arbitrary partition of  $[a,b]$  such as:  $a = x_0 < x_1 < x_2 < \dots < x_{n-1} < x_n = b$ , where  $\Delta x_i$  is the length of the  $i^{\text{th}}$  subinterval (Figure 2).

If  $c_i$  is any point in the  $i^{\text{th}}$  subinterval, then the sum

$$\sum_{i=1}^n f(c_i)\Delta x_i, x_{i-1} \leq c_i \leq x_i$$

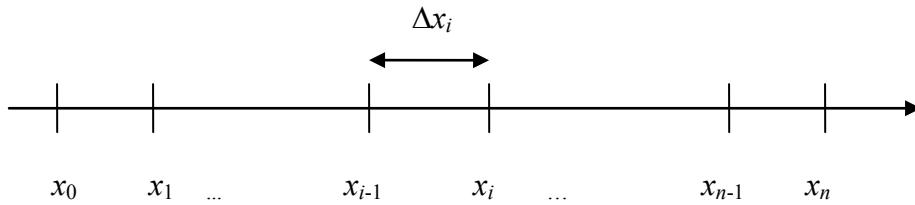
is called a Riemann sum of the function  $f$  for the partition  $\Delta$  on the interval  $[a,b]$ . For a given partition  $\Delta$ , the length of the longest subinterval is called the norm of the partition. It is denoted by  $\|\Delta\|$  (the norm of  $\Delta$ ). The following limit is used to define the definite integral.



**Figure 1** The definite integral as the area of a region under the curve,  $\text{Area} = \int_a^b f(x)dx$ .

If  $c_i$  is any point in the  $i^{\text{th}}$  subinterval, then the sum

$$\sum_{i=1}^n f(c_i)\Delta x_i, x_{i-1} \leq c_i \leq x_i$$



**Figure 2** Division of interval into  $n$  segments.

is called a Riemann sum of the function  $f$  for the partition  $\Delta$  on the interval  $[a,b]$ . For a given partition  $\Delta$ , the length of the longest subinterval is called the norm of the partition. It is denoted by  $\|\Delta\|$  (the norm of  $\Delta$ ). The following limit is used to define the definite integral.

$$\lim_{\|\Delta\| \rightarrow 0} \sum_{i=1}^n f(c_i)\Delta x_i = I$$

This limit exists if and only if for any positive number  $\varepsilon$ , there exists a positive number  $\delta$  such that for every partition  $\Delta$  of  $[a,b]$  with  $\|\Delta\| < \delta$ , it follows that

$$\left| I - \sum_{i=1}^n f(c_i) \Delta x_i \right| < \varepsilon$$

for any choice of  $c_i$  in the  $i^{\text{th}}$  subinterval of  $\Delta$ .

If the limit of a Riemann sum of  $f$  exists, then the function  $f$  is said to be integrable over  $[a,b]$  and the Riemann sum of  $f$  on  $[a,b]$  approaches the number  $I$ .

$$\lim_{\|\Delta\| \rightarrow 0} \sum_{i=1}^n f(c_i) \Delta x_i = I$$

where

$$I = \int_a^b f(x) dx$$

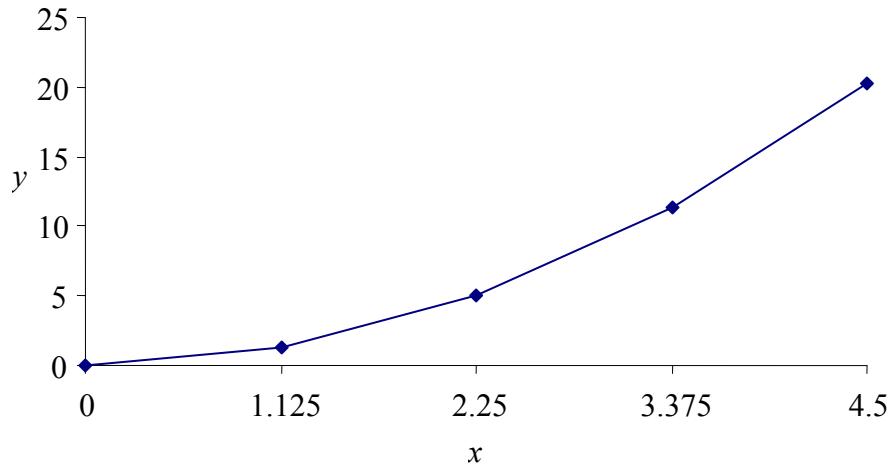
### Example 1

Find the area of the region between the parabola  $y = x^2$  and the  $x$ -axis on the interval  $[0,4.5]$ . Use Riemann's sum with four partitions.

### Solution

We evaluate the integral for the area as a limit of Riemann sums. We sketch the region (Figure 3), and partition  $[0,4.5]$  into four subintervals of length

$$\Delta x = \frac{4.5 - 0}{4} = 1.125.$$



**Figure 3** Graph of the function  $y = x^2$ .

The points of partition are

$$x_0 = 0, x_1 = 1.125, x_2 = 2.25, x_3 = 3.375, x_4 = 4.5$$

Let's choose  $c_i$ 's to be right hand endpoint of its subinterval. Thus,

$$c_1 = x_1 = 1.125, c_2 = x_2 = 2.25, c_3 = x_3 = 3.375, c_4 = x_4 = 4.5$$

The rectangles defined by these choices have the following areas:

$$f(c_1)\Delta x = f(1.125) \times (1.125) = (1.125)^2 (1.125) = 1.4238$$

$$f(c_2)\Delta x = f(2.25) \times (1.125) = (2.25)^2 (1.125) = 5.6953$$

$$f(c_3)\Delta x = f(3.375) \times (1.125) = (3.375)^2 (1.125) = 12.814$$

$$f(c_4)\Delta x = f(4.5) \times (1.125) = (4.5)^2 (1.125) = 22.781$$

The sum of the areas then is

$$\begin{aligned} \int_0^{4.5} x^2 dx &\approx \sum_{i=1}^4 f(c_i)\Delta x, \\ &= 1.4238 + 5.6953 + 12.814 + 22.781 \\ &= 42.715 \end{aligned}$$

How does this compare with the exact value of the integral  $\int_0^{4.5} x^2 dx$ ?

### Example 2

Find the exact area of the region between the parabola  $y = x^2$  and the  $x$ -axis on the interval  $[0, b]$ . Use Riemann's sum.

### Solution

Note that in Example 1 for  $y = x^2$  that

$$f(c_i)\Delta x = i^2(\Delta x)^3$$

Thus, the sum of these areas, if the interval is divided into  $n$  equal segments is

$$\begin{aligned} S_n &= \sum_{i=1}^n f(c_i)\Delta x \\ &= \sum_{i=1}^n i^2(\Delta x)^3 \\ &= (\Delta x)^3 \sum_{i=1}^n i^2 \end{aligned}$$

Since

$$\Delta x = \frac{b}{n}, \text{ and}$$

$$\sum_{i=1}^n i^2 = \frac{n(n+1)(2n+1)}{6}$$

then

$$\begin{aligned} S_n &= \frac{b^3}{n^3} \frac{n(n+1)(2n+1)}{6} \\ &= \frac{b^3}{6} \frac{2n^2 + n + 2n + 1}{n^2} \end{aligned}$$

$$= \frac{b^3}{6} \left( 2 + \frac{3}{n} + \frac{1}{n^2} \right)$$

The definition of a definite integral can now be used

$$\int_a^b f(x) dx = \lim_{\|\Delta x\| \rightarrow 0} \sum_{i=1}^n f(c_i) \Delta x$$

To find the area under the parabola from  $x = 0$  to  $x = b$ , we have

$$\begin{aligned} \int_0^b x^2 dx &= \lim_{|\Delta| \rightarrow 0} \sum_{i=1}^n f(c_i) \Delta x \\ &= \lim_{n \rightarrow \infty} S_n \\ &= \lim_{n \rightarrow \infty} \frac{b^3}{6} \left( 2 + \frac{3}{n} + \frac{1}{n^2} \right) \\ &= \frac{b^3}{6} (2 + 0 + 0) \\ &= \frac{b^3}{3} \end{aligned}$$

For the value of  $b = 4.5$  as given in Example 1,

$$\begin{aligned} \int_0^{4.5} x^2 dx &= \frac{4.5^3}{3} \\ &= 30.375 \end{aligned}$$

### The Mean Value Theorem for Integrals

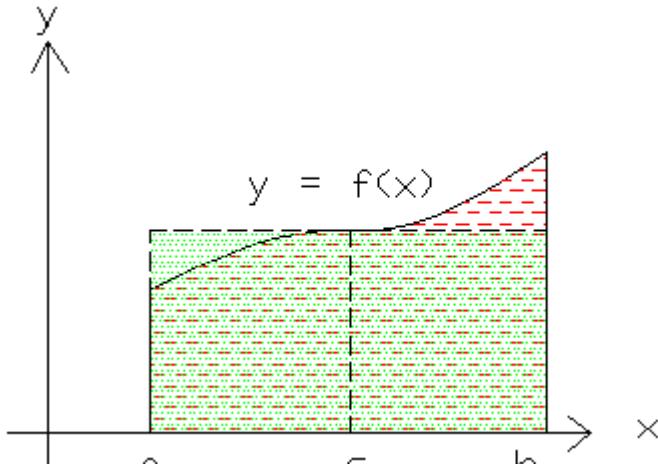
The area of a region under a curve is usually greater than the area of an inscribed rectangle and less than the area of a circumscribed rectangle. The mean value theorem for integrals states that somewhere between these two rectangles, there exists a rectangle whose area is exactly equal to the area of the region under the curve, as shown in Figure 4. Another variation states that if a function  $f$  is continuous between  $a$  and  $b$ , then there is at least one point in  $[a, b]$  where the function equals the average value of the function  $f$  over  $[a, b]$ .

Theorem: If the function  $f$  is continuous on the closed interval  $[a, b]$ , then there exists a number  $c$  in  $[a, b]$  such that:

$$f(c) = \frac{1}{b-a} \int_a^b f(x) dx$$

### Example 3

Graph the function  $f(x) = (x - 1)^2$ , and find its average value over the interval  $[0, 3]$ . At what point in the given interval does the function assume its average value?

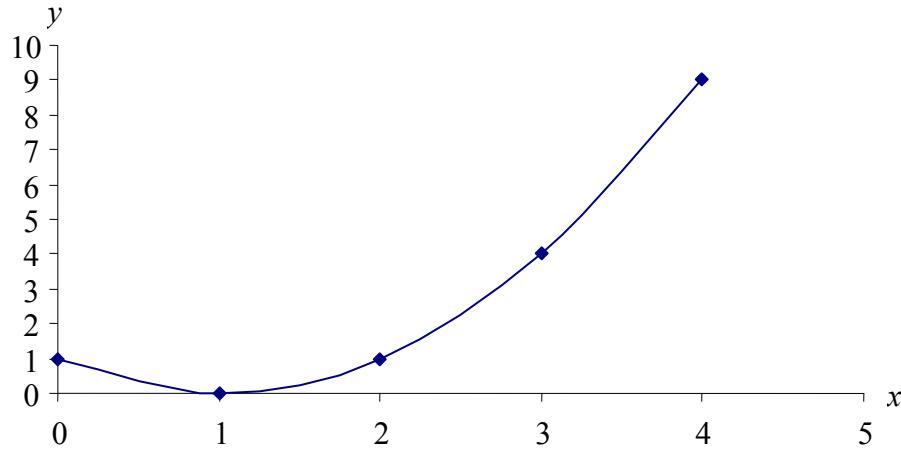
**Figure 4** Mean value rectangle.**Solution**

$$\begin{aligned}
 \text{Average}(f) &= \frac{1}{b-a} \int_a^b f(x) dx \\
 &= \frac{1}{3-0} \int_0^3 (x-1)^2 dx \\
 &= \frac{1}{3} \int_0^3 (x^2 - 2x + 1) dx \\
 &= \frac{1}{3} \left[ \left( \frac{1}{3} \times 27 - 9 + 3 \right) - 0 \right] \\
 &= 1
 \end{aligned}$$

The average value of the function  $f$  over the interval  $[0,3]$  is 1. Thus, the function assumes its average value at

$$\begin{aligned}
 f(c) &= 1 \\
 (c-1)^2 &= 1 \\
 c &= 0, 2
 \end{aligned}$$

The connection between integrals and area can be exploited in two ways. When a formula for the area of the region between the  $x$ -axis and the graph of a continuous function is known, it can be used to evaluate the integral of the function. However, if the area of region is not known, the integral of the function can be used to define and calculate the area. Table 1 lists a number of standard indefinite integral forms.

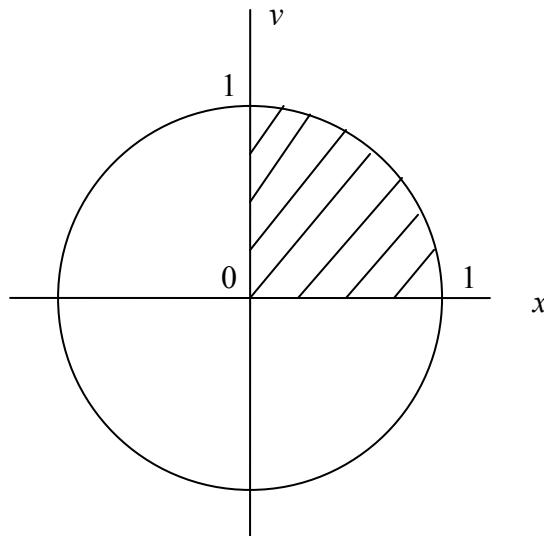


**Figure 5** The function  $f(x) = (x - 1)^2$ .

#### Example 4

Find the area of the region between the circle  $x^2 + y^2 = 1$  and the  $x$ -axis on the interval  $[0,1]$  (the shaded region) in two different ways.

#### Solution



**Figure 6** Graph of the function  $x^2 + y^2 = 1$ .

The first and easy way to solve this problem is by recognizing that it is a quarter circle. Hence the area of the shaded area is

$$A = \frac{1}{4}\pi r^2$$

$$\begin{aligned}
 &= \frac{1}{4}\pi(1)^2 \\
 &= \frac{\pi}{4}
 \end{aligned}$$

The second way is to use the integrals and the trigonometric functions. First, let's simplify the function  $x^2 + y^2 = 1$ .

$$\begin{aligned}
 x^2 + y^2 &= 1 \\
 y^2 &= 1 - x^2 \\
 y &= \sqrt{1 - x^2}
 \end{aligned}$$

The area of the shaded region is the equal to

$$A = \int_0^1 \sqrt{1 - x^2} dx$$

We set  $x = \sin \theta$ ,  $dx = \cos \theta d\theta$

$$\begin{aligned}
 A &= \int_0^1 \sqrt{1 - x^2} dx \\
 &= \int_0^{\pi/2} \sqrt{(1 - \sin^2 \theta)} \cos \theta d\theta
 \end{aligned}$$

$$\begin{aligned}
 &= \int_0^{\pi/2} \sqrt{(\cos^2 \theta)} \cos \theta d\theta \\
 &= \int_0^{\pi/2} \cos^2 \theta d\theta
 \end{aligned}$$

By using the following formula

$$\cos^2 \theta = \frac{1 + \cos 2\theta}{2},$$

we have

$$\begin{aligned}
 A &= \int_0^{\pi/2} \frac{1 + \cos 2\theta}{2} d\theta \\
 &= \int_0^{\pi/2} \left( \frac{1}{2} + \frac{\cos 2\theta}{2} \right) d\theta \\
 &= \left[ \frac{1}{2}\theta + \frac{\sin 2\theta}{4} \right]_0^{\pi/2} \\
 &= \left( \frac{\pi}{4} + 0 \right) - (0 + 0) \\
 &= \frac{\pi}{4}
 \end{aligned}$$

The following are some more examples of exact integration. You can use the brief table of integrals given in Table 1.

**Table 1** A brief table of integrals

$\int dx = x + C$	$\int \sin x dx = -\cos x + C$
$\int a f(x) dx = a \int f(x) dx + C$	$\int \cos x dx = \sin x + C$
$\int [u(x) \pm v(x)] dx = \int u(x) dx \pm \int v(x) dx + C$	$\int \tan x dx = -\ln \cos x  + C = \ln \sec x  + C$
$\int x^n dx = \frac{x^{n+1}}{n+1} + C$	$\int \sec(ax) dx = \frac{1}{a} \ln \sec(ax) + \tan(ax)  + C$
$\int u dv = u v - \int v du + C$	$\int \cot x dx = -\ln \csc x  + C = \ln \sin x  + C$
$\int \frac{dx}{ax+b} = \frac{1}{a} \ln ax+b  + C$	$\int \sec^2 ax dx = \frac{1}{a} \tan(ax) + C$
$\int a^x dx = \frac{a^x}{\ln a} + C$	$\int \sec(x) \tan(x) dx = \sec(x) + C$
$\int e^{ax} dx = \frac{e^{ax}}{a} + C$	$\int \csc(x) \cot(x) dx = -\csc(x) + C$

### Example 5

Evaluate the following integral

$$\int_0^1 2xe^{-x^2} dx$$

**Solution**

Let  $u = -x^2$ ,  $du = -2xdx$

At  $x = 0$ ,  $u = -(0)^2 = 0$

At  $x = 1$ ,  $u = -(1)^2 = -1$

$$\begin{aligned}\int_0^1 2xe^{-x^2} dx &= \int_0^1 (-e^{-x^2})(-2xdx) \\ &= \int_0^{-1} (-e^u)(du) \\ &= \left[ -e^u \right]_0^{-1} \\ &= -e^{-1} - (-e^0) \\ &= 0.6321\end{aligned}$$

**Example 6**

Evaluate

$$\int_0^{\pi/4} \frac{1 + \sin x}{\cos^2 x} dx$$

**Solution**

$$\begin{aligned}\int_0^{\pi/4} \frac{1 + \sin x}{\cos^2 x} dx &= \int_0^{\pi/4} \left( \frac{1}{\cos^2 x} + \frac{\sin x}{\cos^2 x} \right) dx \\ &= \int_0^{\pi/4} (\sec^2 x + \sec x \tan x) dx \\ &= \int_0^{\pi/4} (\sec^2 x) dx + \int_0^{\pi/4} (\sec x)(\tan x) dx \\ &= [\tan x]_0^{\pi/4} + [\sec x]_0^{\pi/4} \\ &= (1 - 0) + (\sqrt{2} - 1) \\ &= \sqrt{2}\end{aligned}$$

**Example 7**

Evaluate  $\int x \sec^2 x dx$

**Solution**

We use the formula

$$\int u dv = uv - \int v du$$

Let  $u = x$ ,  $du = dx$ , and  $dv = \sec^2 x dx$ ,  $v = \tan x$

So the new integral is

$$\begin{aligned}\int x \sec^2 x dx &= x \tan x - \int \tan x dx \\ &= x \tan x + \ln|\cos x| + C\end{aligned}$$

**Example 8**

Evaluate

$$\int_1^2 x \ln x dx$$

**Solution**

Let  $u = \ln x$ ,  $du = \frac{1}{x} dx$  and  $dv = x dx$ ,  $v = \frac{x^2}{2}$

Using the formula  $\int u dv = uv - \int v du$ , the new integral is

$$\begin{aligned}\int_1^2 (x)(\ln x) dx &= \left[ \ln x \times \frac{x^2}{2} \right]_1^2 - \int_1^2 \left( \frac{x^2}{2} \right) \left( \frac{1}{x} dx \right) \\ &= \left[ \ln x \times \frac{x^2}{2} \right]_1^2 - \int_1^2 \frac{x}{2} dx \\ &= \left[ \ln x \times \frac{x^2}{2} \right]_1^2 - \left[ \frac{x^2}{4} \right]_1^2 \\ &= \left[ \left( \ln 2 \times \frac{2^2}{2} \right) - \left( \ln 1 \times \frac{1^2}{2} \right) \right] - \left[ \left( \frac{2^2}{4} \right) - \left( \frac{1^2}{4} \right) \right] \\ &= \left[ (2 \ln 2) - \left( \frac{1}{2} \ln 1 \right) \right] - \left[ \left( \frac{4}{4} \right) - \left( \frac{1}{4} \right) \right] \\ &= \left[ (2 \ln 2) - \left( \frac{1}{2} \times 0 \right) \right] - \left[ 1 - \frac{1}{4} \right] \\ &= 0.6362\end{aligned}$$

**Example 9**

Evaluate

$$\int_0^1 \frac{5x}{(4+x^2)^2} dx$$

**Solution**

We use the formula  $\int_a^b f(g(x))g'(x) dx = \int_{g(a)}^{g(b)} f(u) du$ , by substituting  $u = g(x)$ ,  $du = g'(x)dx$

then integrating from  $g(a)$  to  $g(b)$ .

Let

$$u = g(x) = 4 + x^2,$$

so

$$g(0) = 4, g(1) = 5, \text{ and}$$

$$du = (2x)dx$$

The new integral is

$$\begin{aligned} \int_0^1 \frac{5x}{(4+x^2)^2} dx &= \int_0^1 \frac{1}{(4+x^2)^2} \times \frac{5}{2} \times (2x) dx \\ &= \frac{5}{2} \int_4^5 \frac{1}{u^2} du \\ &= \frac{5}{2} \left[ -\frac{1}{u} \right]_4^5 \\ &= \frac{5}{2} \left[ \left(-\frac{1}{5}\right) - \left(-\frac{1}{4}\right) \right] \\ &= \frac{5}{2} \times \frac{1}{20} \\ &= 0.125 \end{aligned}$$

### Example 10

Evaluate

$$\int_0^4 |2x-1| dx$$

### Solution

First, let's analyze the expression  $|2x-1|$ .

$$|2x-1| = -(2x-1), x < \frac{1}{2}$$

$$= (2x-1), x \geq \frac{1}{2}$$

$$\begin{aligned} \int_0^4 |2x-1| dx &= \int_0^{1/2} -(2x-1) dx + \int_{1/2}^4 (2x-1) dx \\ &= \left[ -x^2 + x \right]_0^{1/2} + \left[ x^2 - x \right]_{1/2}^4 \\ &= \left[ \left(-\frac{1}{4} + \frac{1}{2}\right) - 0 \right] + \left[ (16 - 4) - \left(\frac{1}{4} - \frac{1}{2}\right) \right] \\ &= 12.5 \end{aligned}$$

### Example 11

Evaluate

$$\int_{-\infty}^{-2} \frac{2}{x^2 - 1} dx$$

**Solution**

$$\begin{aligned}
\int_{-\infty}^{-2} \frac{2}{x^2 - 1} dx &= \int_{-\infty}^{-2} \frac{2}{(x-1) \times (x+1)} dx \\
&= \int_{-\infty}^{-2} \frac{(x+1) - (x-1)}{(x-1) \times (x+1)} dx \\
&= \int_{-\infty}^{-2} \frac{x+1}{(x-1) \times (x+1)} - \frac{x-1}{(x-1) \times (x+1)} dx \\
&= \int_{-\infty}^{-2} \frac{1}{x-1} dx - \int_{-\infty}^{-2} \frac{1}{x+1} dx \\
&= \lim_{b \rightarrow -\infty} [\ln|x-1|]_{-2}^{-2} - \lim_{b \rightarrow -\infty} [\ln|x+1|]_{-2}^{-2} \\
&= \lim_{b \rightarrow -\infty} \left[ \ln \left| \frac{x-1}{x+1} \right| \right]_{-2}^b \\
&= \lim_{b \rightarrow -\infty} \left[ \ln \left| \frac{-3}{-1} \right| - \ln \left| \frac{b-1}{b+1} \right| \right] \\
&= \ln(3) - \ln \left( \lim_{b \rightarrow -\infty} \left| \frac{b-1}{b+1} \right| \right) \\
&= \ln(3) - \ln(1) \\
&= \ln(3) \\
&= 1.0986
\end{aligned}$$

**Example 12**

Graph the function  $y = \frac{1}{3}(x^2 + 2)^{3/2}$ , and find the length of the curve from  $x = 0$  to  $x = 3$ .

**Solution**

We use the equation

$$L = \int_a^b \sqrt{1 + \left(\frac{dy}{dx}\right)^2} dx$$

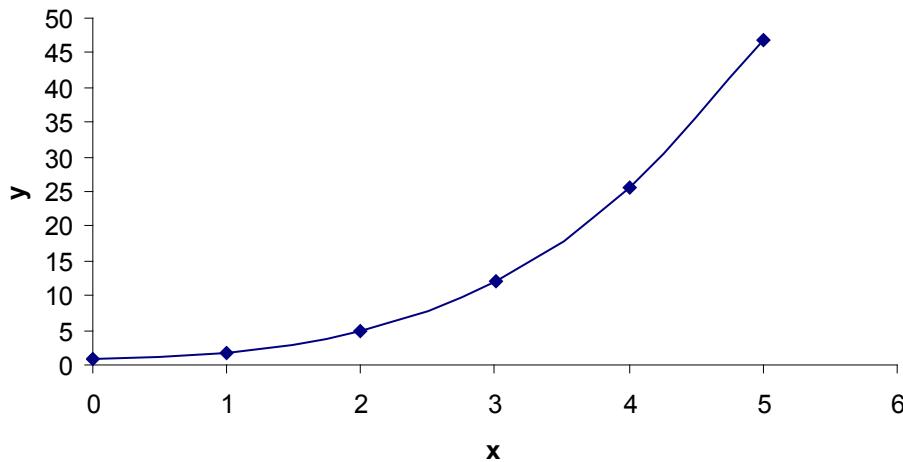
We have:

$$y = \frac{1}{3}(x^2 + 2)^{3/2}$$

So,

$$\begin{aligned}
\frac{dy}{dx} &= \left(\frac{1}{3}\right) \times \left(\frac{3}{2}\right) \times (x^2 + 2)^{3/2-1} \times (2x) \\
&= x\sqrt{x^2 + 2}
\end{aligned}$$

$$L = \int_0^3 \sqrt{1 + (x\sqrt{x^2 + 2})^2} dx$$

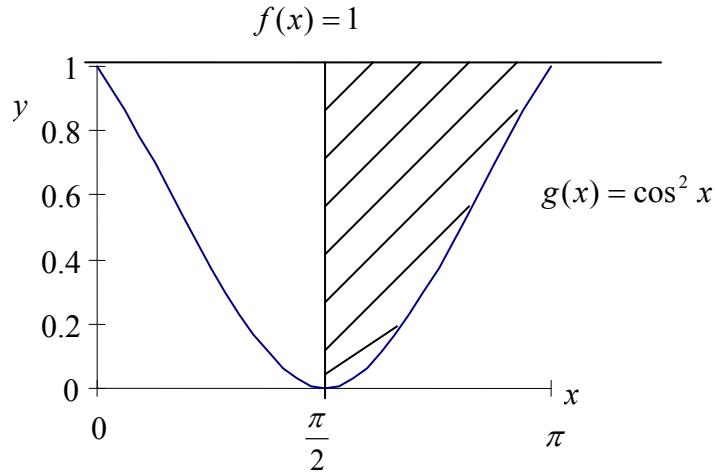


**Figure 7** Graph of the function  $y = \frac{1}{3}(x^2 + 2)^{3/2}$

$$\begin{aligned} &= \int_0^3 \sqrt{1 + x^2(x^2 + 2)} dx \\ &= \int_0^3 \sqrt{1 + x^4 + 2x^2} dx \\ &= \int_0^3 \sqrt{(x^2 + 1)^2} dx \\ &= \int_0^3 (x^2 + 1) dx \\ &= \left[ \frac{x^3}{3} + x \right]_0^3 \\ &= 12 \end{aligned}$$

### Example 13

Find the area of the shaded region given in Figure 8.



**Figure 8** Graph of the function  $\cos^2 x$ .

### Solution

For the sketch given,

$$a = \frac{\pi}{2}, b = \pi, \text{ and}$$

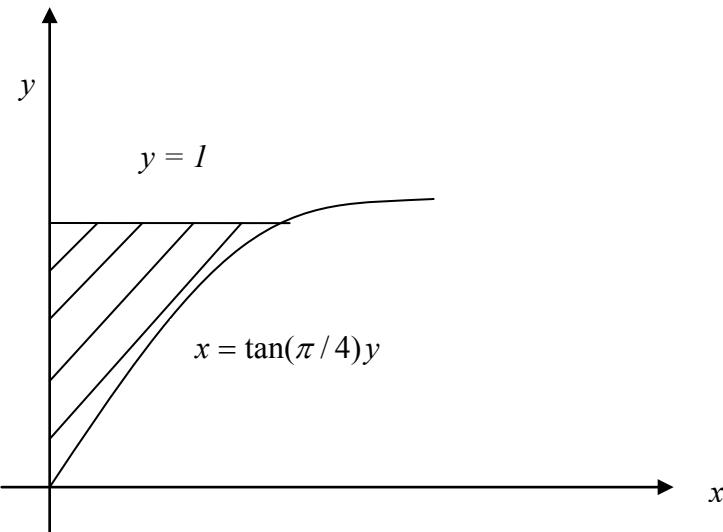
$$f(x) - g(x) = 1 - \cos^2 x = \sin^2 x$$

$$\begin{aligned} A &= \int_{\pi/2}^{\pi} \sin^2(x) dx \\ &= \int_{\pi/2}^{\pi} \frac{1 - \cos 2x}{2} dx \\ &= \int_{\pi/2}^{\pi} \left[ \frac{1}{2} - \frac{\cos 2x}{2} \right] dx \\ &= \left[ \frac{x}{2} - \frac{\sin 2x}{4} \right]_{\pi/2}^{\pi} \\ &= \left[ \left( \frac{\pi}{2} - \frac{\sin(2\pi)}{4} \right) - \left( \frac{\pi}{4} - \frac{\sin 2\left(\frac{\pi}{2}\right)}{4} \right) \right] \\ &= \left[ \left( \frac{\pi}{2} - 0 \right) - \left( \frac{\pi}{4} - 0 \right) \right] \end{aligned}$$

$$= \frac{\pi}{4}$$

**Example 14**

Find the volume of the solid generated by revolving the shaded region in Figure 9 about the  $y$ -axis.



**Figure 9** Volume generated by revolving shaded region.

**Solution**

We use the formula  $V = \int_a^b \pi (\text{radius})^2 dy$

Let

$$u = \frac{\pi}{4}y, \quad du = \frac{\pi}{4}dy.$$

Therefore, at  $y = 0, u = 0$

$$y = 1, \quad u = \frac{\pi}{4}$$

$$\begin{aligned} V &= \int_0^1 \pi [R(y)]^2 dy \\ &= \pi \int_0^1 \left[ \tan\left(\frac{\pi}{4}y\right) \right]^2 dy \\ &= \pi \times \frac{4}{\pi} \int_0^1 \left[ \tan\left(\frac{\pi}{4}y\right) \right]^2 \frac{\pi}{4} dy \end{aligned}$$

$$\begin{aligned} &= 4 \int_0^{\pi/4} (\tan u)^2 du \quad (\text{Choosing } u = \frac{\pi}{4}y) \\ &= 4 \int_0^{\pi/4} (-1 + \sec^2 u) du \\ &= 4 \left[ -u + \tan u \right]_0^{\pi/4} \\ &= 4 \left[ \left( -\frac{\pi}{4} + \tan \frac{\pi}{4} \right) - (0 + \tan 0) \right] \\ &= 4 \left[ \left( -\frac{\pi}{4} + 1 \right) - (0 + 0) \right] \\ &= 0.8584 \end{aligned}$$

---

## INTEGRATION

---

Topic	Primer on integration
Summary	These are textbook notes of a primer on integration.
Major	General Engineering
Authors	Autar Kaw, Loubna Guennoun
Date	July 2, 2010
Web Site	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

---

# **Chapter 07.02**

## **Trapezoidal Rule of Integration**

*After reading this chapter, you should be able to:*

1. derive the trapezoidal rule of integration,
2. use the trapezoidal rule of integration to solve problems,
3. derive the multiple-segment trapezoidal rule of integration,
4. use the multiple-segment trapezoidal rule of integration to solve problems, and
5. derive the formula for the true error in the multiple-segment trapezoidal rule of integration.

### **What is integration?**

Integration is the process of measuring the area under a function plotted on a graph. Why would we want to integrate a function? Among the most common examples are finding the velocity of a body from an acceleration function, and displacement of a body from a velocity function. Throughout many engineering fields, there are (what sometimes seems like) countless applications for integral calculus. You can read about some of these applications in Chapters 07.00A-07.00G.

Sometimes, the evaluation of expressions involving these integrals can become daunting, if not indeterminate. For this reason, a wide variety of numerical methods has been developed to simplify the integral.

Here, we will discuss the trapezoidal rule of approximating integrals of the form

$$I = \int_a^b f(x)dx$$

where

$f(x)$  is called the integrand,

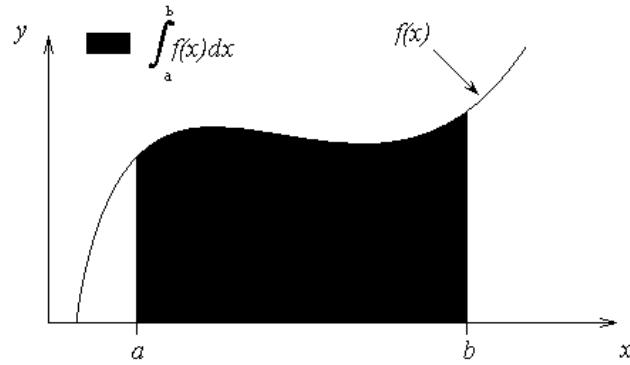
$a$  = lower limit of integration

$b$  = upper limit of integration

### **What is the trapezoidal rule?**

The trapezoidal rule is based on the Newton-Cotes formula that if one approximates the integrand by an  $n^{\text{th}}$  order polynomial, then the integral of the function is approximated by

the integral of that  $n^{\text{th}}$  order polynomial. Integrating polynomials is simple and is based on the calculus formula.



**Figure 1** Integration of a function

$$\int_a^b x^n dx = \left( \frac{b^{n+1} - a^{n+1}}{n+1} \right), n \neq -1 \quad (1)$$

So if we want to approximate the integral

$$I = \int_a^b f(x) dx \quad (2)$$

to find the value of the above integral, one assumes

$$f(x) \approx f_n(x) \quad (3)$$

where

$$f_n(x) = a_0 + a_1x + \dots + a_{n-1}x^{n-1} + a_nx^n. \quad (4)$$

where  $f_n(x)$  is a  $n^{\text{th}}$  order polynomial. The trapezoidal rule assumes  $n=1$ , that is, approximating the integral by a linear polynomial (straight line),

$$\int_a^b f(x) dx \approx \int_a^b f_1(x) dx$$

### Derivation of the Trapezoidal Rule

#### Method 1: Derived from Calculus

$$\begin{aligned} \int_a^b f(x) dx &\approx \int_a^b f_1(x) dx \\ &= \int_a^b (a_0 + a_1x) dx \\ &= a_0(b-a) + a_1 \left( \frac{b^2 - a^2}{2} \right) \end{aligned} \quad (5)$$

But what is  $a_0$  and  $a_1$ ? Now if one chooses,  $(a, f(a))$  and  $(b, f(b))$  as the two points to approximate  $f(x)$  by a straight line from  $a$  to  $b$ ,

$$f(a) = f_1(a) = a_0 + a_1 a \quad (6)$$

$$f(b) = f_1(b) = a_0 + a_1 b \quad (7)$$

Solving the above two equations for  $a_1$  and  $a_0$ ,

$$\begin{aligned} a_1 &= \frac{f(b) - f(a)}{b - a} \\ a_0 &= \frac{f(a)b - f(b)a}{b - a} \end{aligned} \quad (8a)$$

Hence from Equation (5),

$$\int_a^b f(x) dx \approx \frac{f(a)b - f(b)a}{b - a} (b - a) + \frac{f(b) - f(a)}{b - a} \frac{b^2 - a^2}{2} \quad (8b)$$

$$= (b - a) \left[ \frac{f(a) + f(b)}{2} \right] \quad (9)$$

### Method 2: Also Derived from Calculus

$f_1(x)$  can also be approximated by using Newton's divided difference polynomial as

$$f_1(x) = f(a) + \frac{f(b) - f(a)}{b - a} (x - a) \quad (10)$$

Hence

$$\begin{aligned} \int_a^b f(x) dx &\approx \int_a^b f_1(x) dx \\ &= \int_a^b \left[ f(a) + \frac{f(b) - f(a)}{b - a} (x - a) \right] dx \\ &= \left[ f(a)x + \frac{f(b) - f(a)}{b - a} \left( \frac{x^2}{2} - ax \right) \right]_a^b \\ &= f(a)b - f(a)a + \left( \frac{f(b) - f(a)}{b - a} \right) \left( \frac{b^2}{2} - ab - \frac{a^2}{2} + a^2 \right) \\ &= f(a)b - f(a)a + \left( \frac{f(b) - f(a)}{b - a} \right) \left( \frac{b^2}{2} - ab + \frac{a^2}{2} \right) \\ &= f(a)b - f(a)a + \left( \frac{f(b) - f(a)}{b - a} \right) \frac{1}{2} (b - a)^2 \\ &= f(a)b - f(a)a + \frac{1}{2} (f(b) - f(a))(b - a) \end{aligned}$$

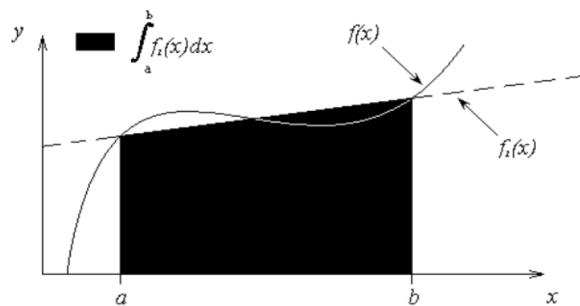
$$\begin{aligned}
&= f(a)b - f(a)a + \frac{1}{2}f(b)b - \frac{1}{2}f(b)a - \frac{1}{2}f(a)b + \frac{1}{2}f(a)a \\
&= \frac{1}{2}f(a)b - \frac{1}{2}f(a)a + \frac{1}{2}f(b)b - \frac{1}{2}f(b)a \\
&= (b-a) \left[ \frac{f(a) + f(b)}{2} \right]
\end{aligned} \tag{11}$$

This gives the same result as Equation (10) because they are just different forms of writing the same polynomial.

### Method 3: Derived from Geometry

The trapezoidal rule can also be derived from geometry. Look at Figure 2. The area under the curve  $f_1(x)$  is the area of a trapezoid. The integral

$$\begin{aligned}
\int_a^b f(x)dx &\approx \text{Area of trapezoid} \\
&= \frac{1}{2}(\text{Sum of length of parallel sides})(\text{Perpendicular distance between parallel sides}) \\
&= \frac{1}{2}(f(b) + f(a))(b-a) \\
&= (b-a) \left[ \frac{f(a) + f(b)}{2} \right]
\end{aligned} \tag{12}$$



**Figure 2** Geometric representation of trapezoidal rule.

### Method 4: Derived from Method of Coefficients

The trapezoidal rule can also be derived by the method of coefficients. The formula

$$\begin{aligned}
\int_a^b f(x)dx &\approx \frac{b-a}{2} f(a) + \frac{b-a}{2} f(b) \\
&= \sum_{i=1}^2 c_i f(x_i)
\end{aligned} \tag{13}$$

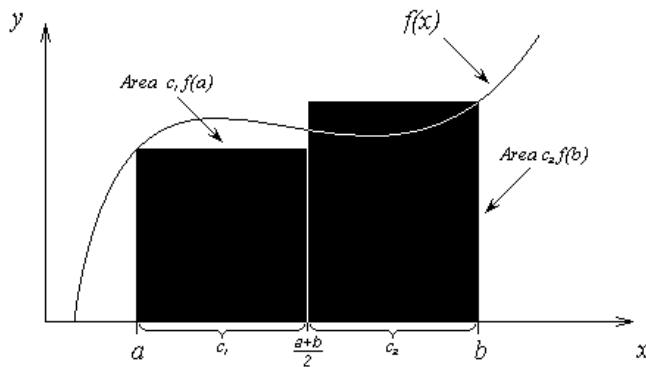
where

$$c_1 = \frac{b-a}{2}$$

$$c_2 = \frac{b-a}{2}$$

$$x_1 = a$$

$$x_2 = b$$



**Figure 3** Area by method of coefficients.

The interpretation is that  $f(x)$  is evaluated at points  $a$  and  $b$ , and each function evaluation is given a weight of  $\frac{b-a}{2}$ . Geometrically, Equation (12) is looked at as the area of a trapezoid, while Equation (13) is viewed as the sum of the area of two rectangles, as shown in Figure 3. How can one derive the trapezoidal rule by the method of coefficients?

Assume

$$\int_a^b f(x) dx = c_1 f(a) + c_2 f(b) \quad (14)$$

Let the right hand side be an exact expression for integrals of  $\int_a^b 1 dx$  and  $\int_a^b x dx$ , that is, the formula will then also be exact for linear combinations of  $f(x)=1$  and  $f(x)=x$ , that is, for  $f(x)=a_0(1)+a_1(x)$ .

$$\int_a^b 1 dx = b - a = c_1 + c_2 \quad (15)$$

$$\int_a^b x dx = \frac{b^2 - a^2}{2} = c_1 a + c_2 b \quad (16)$$

Solving the above two equations gives

$$\begin{aligned} c_1 &= \frac{b-a}{2} \\ c_2 &= \frac{b-a}{2} \end{aligned} \quad (17)$$

Hence

$$\int_a^b f(x)dx \approx \frac{b-a}{2}f(a) + \frac{b-a}{2}f(b) \quad (18)$$

### Method 5: Another approach on the Method of Coefficients

The trapezoidal rule can also be derived by the method of coefficients by another approach

$$\int_a^b f(x)dx \approx \frac{b-a}{2}f(a) + \frac{b-a}{2}f(b)$$

Assume

$$\int_a^b f(x)dx = c_1 f(a) + c_2 f(b) \quad (19)$$

Let the right hand side be exact for integrals of the form

$$\int_a^b (a_0 + a_1 x)dx$$

So

$$\begin{aligned} \int_a^b (a_0 + a_1 x)dx &= \left( a_0 x + a_1 \frac{x^2}{2} \right)_a^b \\ &= a_0(b-a) + a_1 \left( \frac{b^2 - a^2}{2} \right) \end{aligned} \quad (20)$$

But we want

$$\int_a^b (a_0 + a_1 x)dx = c_1 f(a) + c_2 f(b) \quad (21)$$

to give the same result as Equation (20) for  $f(x) = a_0 + a_1 x$ .

$$\begin{aligned} \int_a^b (a_0 + a_1 x)dx &= c_1(a_0 + a_1 a) + c_2(a_0 + a_1 b) \\ &= a_0(c_1 + c_2) + a_1(c_1 a + c_2 b) \end{aligned} \quad (22)$$

Hence from Equations (20) and (22),

$$a_0(b-a) + a_1 \left( \frac{b^2 - a^2}{2} \right) = a_0(c_1 + c_2) + a_1(c_1 a + c_2 b)$$

Since  $a_0$  and  $a_1$  are arbitrary for a general straight line

$$\begin{aligned} c_1 + c_2 &= b-a \\ c_1 a + c_2 b &= \frac{b^2 - a^2}{2} \end{aligned} \quad (23)$$

Again, solving the above two equations (23) gives

$$\begin{aligned} c_1 &= \frac{b-a}{2} \\ c_2 &= \frac{b-a}{2} \end{aligned} \quad (24)$$

Therefore

$$\begin{aligned} \int_a^b f(x)dx &\approx c_1 f(a) + c_2 f(b) \\ &= \frac{b-a}{2} f(a) + \frac{b-a}{2} f(b) \end{aligned} \quad (25)$$

### Example 1

The vertical distance covered by a rocket from  $t = 8$  to  $t = 30$  seconds is given by

$$x = \int_8^{30} \left( 2000 \ln \left[ \frac{140000}{140000 - 2100t} \right] - 9.8t \right) dt$$

- a) Use the single segment trapezoidal rule to find the distance covered for  $t = 8$  to  $t = 30$  seconds.
- b) Find the true error,  $E_t$  for part (a).
- c) Find the absolute relative true error for part (a).

### Solution

a)  $I \approx (b-a) \left[ \frac{f(a) + f(b)}{2} \right]$ , where

$$a = 8$$

$$b = 30$$

$$f(t) = 2000 \ln \left[ \frac{140000}{140000 - 2100t} \right] - 9.8t$$

$$f(8) = 2000 \ln \left[ \frac{140000}{140000 - 2100(8)} \right] - 9.8(8)$$

$$= 177.27 \text{ m/s}$$

$$f(30) = 2000 \ln \left[ \frac{140000}{140000 - 2100(30)} \right] - 9.8(30)$$

$$= 901.67 \text{ m/s}$$

$$I \approx (30-8) \left[ \frac{177.27 + 901.67}{2} \right]$$

$$= 11868 \text{ m}$$

- b) The exact value of the above integral is

$$\begin{aligned} x &= \int_8^{30} \left( 2000 \ln \left[ \frac{140000}{140000 - 2100t} \right] - 9.8t \right) dt \\ &= 11061 \text{ m} \end{aligned}$$

so the true error is

$$E_t = \text{True Value} - \text{Approximate Value}$$

$$= 11061 - 11868$$

$$= -807 \text{ m}$$

c) The absolute relative true error,  $|\epsilon_t|$ , would then be

$$\begin{aligned} |\epsilon_t| &= \left| \frac{\text{True Error}}{\text{True Value}} \right| \times 100 \\ &= \left| \frac{11061 - 11868}{11061} \right| \times 100 \\ &= 7.2958\% \end{aligned}$$

### Multiple-Segment Trapezoidal Rule

In Example 1, the true error using a single segment trapezoidal rule was large. We can divide the interval [8,30] into [8,19] and [19,30] intervals and apply the trapezoidal rule over each segment.

$$\begin{aligned} f(t) &= 2000 \ln\left(\frac{140000}{140000 - 2100t}\right) - 9.8t \\ \int_8^{30} f(t) dt &= \int_8^{19} f(t) dt + \int_{19}^{30} f(t) dt \\ &\approx (19 - 8) \left[ \frac{f(8) + f(19)}{2} \right] + (30 - 19) \left[ \frac{f(19) + f(30)}{2} \right] \\ f(8) &= 177.27 \text{ m/s} \\ f(19) &= 2000 \ln\left(\frac{140000}{140000 - 2100(19)}\right) - 9.8(19) = 484.75 \text{ m/s} \\ f(30) &= 901.67 \text{ m/s} \end{aligned}$$

Hence

$$\begin{aligned} \int_8^{30} f(t) dt &\approx (19 - 8) \left[ \frac{177.27 + 484.75}{2} \right] + (30 - 19) \left[ \frac{484.75 + 901.67}{2} \right] \\ &= 11266 \text{ m} \end{aligned}$$

The true error,  $E_t$ , is

$$\begin{aligned} E_t &= 11061 - 11266 \\ &= -205 \text{ m} \end{aligned}$$

The true error now is reduced from 807 m to 205 m. Extending this procedure to dividing  $[a,b]$  into  $n$  equal segments and applying the trapezoidal rule over each segment, the sum of the results obtained for each segment is the approximate value of the integral.

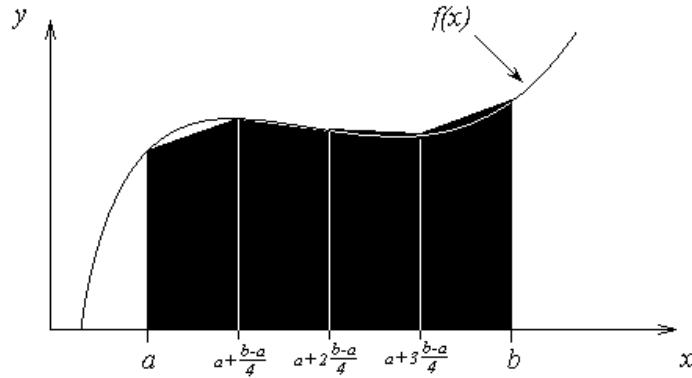
Divide  $(b-a)$  into  $n$  equal segments as shown in Figure 4. Then the width of each segment is

$$h = \frac{b-a}{n} \quad (26)$$

The integral  $I$  can be broken into  $h$  integrals as

$$I = \int_a^b f(x) dx$$

$$= \int_a^{a+h} f(x)dx + \int_{a+h}^{a+2h} f(x)dx + \dots + \int_{a+(n-2)h}^{a+(n-1)h} f(x)dx + \int_{a+(n-1)h}^b f(x)dx \quad (27)$$



**Figure 4** Multiple ( $n = 4$ ) segment trapezoidal rule

Applying trapezoidal rule Equation (27) on each segment gives

$$\begin{aligned}
 \int_a^b f(x)dx &= [(a+h)-a] \left[ \frac{f(a)+f(a+h)}{2} \right] \\
 &\quad + [(a+2h)-(a+h)] \left[ \frac{f(a+h)+f(a+2h)}{2} \right] \\
 &\quad + \dots \dots \dots + [(a+(n-1)h)-(a+(n-2)h)] \left[ \frac{f(a+(n-2)h)+f(a+(n-1)h)}{2} \right] \\
 &\quad + [b-(a+(n-1)h)] \left[ \frac{f(a+(n-1)h)+f(b)}{2} \right] \\
 &= h \left[ \frac{f(a)+f(a+h)}{2} \right] + h \left[ \frac{f(a+h)+f(a+2h)}{2} \right] + \dots \dots \dots \\
 &\quad + h \left[ \frac{f(a+(n-2)h)+f(a+(n-1)h)}{2} \right] + h \left[ \frac{f(a+(n-1)h)+f(b)}{2} \right] \\
 &= h \left[ \frac{f(a)+2f(a+h)+2f(a+2h)+\dots+2f(a+(n-1)h)+f(b)}{2} \right] \\
 &= \frac{h}{2} \left[ f(a) + 2 \left\{ \sum_{i=1}^{n-1} f(a+ih) \right\} + f(b) \right] \\
 &= \frac{b-a}{2n} \left[ f(a) + 2 \left\{ \sum_{i=1}^{n-1} f(a+ih) \right\} + f(b) \right]
 \end{aligned} \quad (28)$$

**Example 2**

The vertical distance covered by a rocket from  $t = 8$  to  $t = 30$  seconds is given by

$$x = \int_8^{30} \left( 2000 \ln \left[ \frac{140000}{140000 - 2100t} \right] - 9.8t \right) dt$$

- a) Use the two-segment trapezoidal rule to find the distance covered from  $t = 8$  to  $t = 30$  seconds.
- b) Find the true error,  $E_t$ , for part (a).
- c) Find the absolute relative true error for part (a).

**Solution**

a) The solution using 2-segment Trapezoidal rule is

$$I \approx \frac{b-a}{2n} \left[ f(a) + 2 \left\{ \sum_{i=1}^{n-1} f(a+ih) \right\} + f(b) \right]$$

$$n = 2$$

$$a = 8$$

$$b = 30$$

$$h = \frac{b-a}{n}$$

$$= \frac{30-8}{2}$$

$$= 11$$

$$I \approx \frac{30-8}{2(2)} \left[ f(8) + 2 \left\{ \sum_{i=1}^{2-1} f(8+11i) \right\} + f(30) \right]$$

$$= \frac{22}{4} [f(8) + 2f(19) + f(30)]$$

$$= \frac{22}{4} [177.27 + 2(484.75) + 901.67]$$

$$= 11266 \text{ m}$$

b) The exact value of the above integral is

$$x = \int_8^{30} \left( 2000 \ln \left[ \frac{140000}{140000 - 2100t} \right] - 9.8t \right) dt$$

$$= 11061 \text{ m}$$

so the true error is

$$\begin{aligned} E_t &= \text{True Value} - \text{Approximate Value} \\ &= 11061 - 11266 \\ &= -205 \text{ m} \end{aligned}$$

c) The absolute relative true error,  $|\epsilon_t|$ , would then be

$$\begin{aligned} |e_t| &= \left| \frac{\text{True Error}}{\text{True Value}} \right| \times 100 \\ &= \left| \frac{11061 - 11266}{11061} \right| \times 100 \\ &= 1.8537\% \end{aligned}$$

**Table 1** Values obtained using multiple-segment trapezoidal rule for

$$x = \int_8^{30} \left( 2000 \ln \left[ \frac{140000}{140000 - 2100t} \right] - 9.8t \right) dt$$

$n$	Approximate Value	$E_t$	$ e_t  \%$	$ e_a  \%$
1	11868	-807	7.296	---
2	11266	-205	1.853	5.343
3	11153	-91.4	0.8265	1.019
4	11113	-51.5	0.4655	0.3594
5	11094	-33.0	0.2981	0.1669
6	11084	-22.9	0.2070	0.09082
7	11078	-16.8	0.1521	0.05482
8	11074	-12.9	0.1165	0.03560

**Example 3**

Use the multiple-segment trapezoidal rule to find the area under the curve

$$f(x) = \frac{300x}{1 + e^x}$$

from  $x = 0$  to  $x = 10$ .

**Solution**

Using two segments, we get

$$h = \frac{10 - 0}{2} = 5$$

$$f(0) = \frac{300(0)}{1 + e^0} = 0$$

$$f(5) = \frac{300(5)}{1 + e^5} = 10.039$$

$$f(10) = \frac{300(10)}{1 + e^{10}} = 0.136$$

$$\begin{aligned} I &\approx \frac{b-a}{2n} \left[ f(a) + 2 \left\{ \sum_{i=1}^{n-1} f(a+ih) \right\} + f(b) \right] \\ &= \frac{10-0}{2(2)} \left[ f(0) + 2 \left\{ \sum_{i=1}^{2-1} f(0+5) \right\} + f(10) \right] \end{aligned}$$

$$\begin{aligned}
 &= \frac{10}{4} [f(0) + 2f(5) + f(10)] \\
 &= \frac{10}{4} [0 + 2(10.039) + 0.136] = 50.537
 \end{aligned}$$

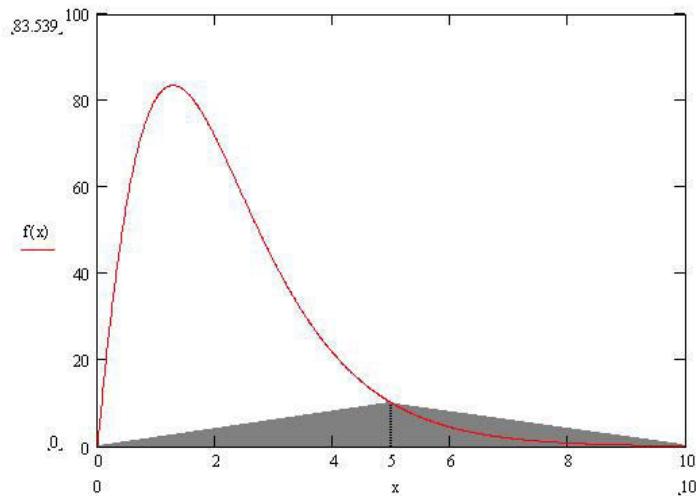
So what is the true value of this integral?

$$\int_0^{10} \frac{300x}{1+e^x} dx = 246.59$$

Making the absolute relative true error

$$\begin{aligned}
 |\epsilon_t| &= \left| \frac{246.59 - 50.535}{246.59} \right| \times 100 \\
 &= 79.506\%
 \end{aligned}$$

Why is the true value so far away from the approximate values? Just take a look at Figure 5. As you can see, the area under the “trapezoids” (yeah, they really look like triangles now) covers a small portion of the area under the curve. As we add more segments, the approximated value quickly approaches the true value.



**Figure 5** 2-segment trapezoidal rule approximation.

**Table 2** Values obtained using multiple-segment trapezoidal rule for  $\int_0^{10} \frac{300x}{1+e^x} dx$ .

$n$	Approximate Value	$E_t$	$ \epsilon_t $
1	0.681	245.91	99.724%
2	50.535	196.05	79.505%
4	170.61	75.978	30.812%
8	227.04	19.546	7.927%

16	241.70	4.887	1.982%
32	245.37	1.222	0.495%
64	246.28	0.305	0.124%

**Example 4**

Use multiple-segment trapezoidal rule to find

$$I = \int_0^2 \frac{1}{\sqrt{x}} dx$$

**Solution**

We cannot use the trapezoidal rule for this integral, as the value of the integrand at  $x = 0$  is infinite. However, it is known that a discontinuity in a curve will not change the area under it. We can assume any value for the function at  $x = 0$ . The algorithm to define the function so that we can use the multiple-segment trapezoidal rule is given below.

Function  $f(x)$

If  $x = 0$  Then  $f = 0$

If  $x \neq 0$  Then  $f = x^{-0.5}$

End Function

Basically, we are just assigning the function a value of zero at  $x = 0$ . Everywhere else, the function is continuous. This means the true value of our integral will be just that—true. Let's see what happens using the multiple-segment trapezoidal rule.

Using two segments, we get

$$h = \frac{2 - 0}{2} = 1$$

$$f(0) = 0$$

$$f(1) = \frac{1}{\sqrt{1}} = 1$$

$$f(2) = \frac{1}{\sqrt{2}} = 0.70711$$

$$I \approx \frac{b-a}{2n} \left[ f(a) + 2 \left\{ \sum_{i=1}^{n-1} f(a+ih) \right\} + f(b) \right]$$

$$= \frac{2-0}{2(2)} \left[ f(0) + 2 \left\{ \sum_{i=1}^{2-1} f(0+i) \right\} + f(2) \right]$$

$$= \frac{2}{4} [f(0) + 2f(1) + f(2)]$$

$$= \frac{2}{4} [0 + 2(1) + 0.70711]$$

$$= 1.3536$$

So what is the true value of this integral?

$$\int_0^2 \frac{1}{\sqrt{x}} dx = 2.8284$$

Thus making the absolute relative true error

$$|\epsilon_t| = \left| \frac{2.8284 - 1.3536}{2.8284} \right| \times 100 \\ = 52.145\%$$

**Table 3** Values obtained using multiple-segment trapezoidal rule for  $\int_0^2 \frac{1}{\sqrt{x}} dx$ .

$n$	Approximate Value	$E_t$	$ \epsilon_t $
2	1.354	1.474	52.14%
4	1.792	1.036	36.64%
8	2.097	0.731	25.85%
16	2.312	0.516	18.26%
32	2.463	0.365	12.91%
64	2.570	0.258	9.128%
128	2.646	0.182	6.454%
256	2.699	0.129	4.564%
512	2.737	0.091	3.227%
1024	2.764	0.064	2.282%
2048	2.783	0.045	1.613%
4096	2.796	0.032	1.141%

### Error in Multiple-segment Trapezoidal Rule

The true error for a single segment Trapezoidal rule is given by

$$E_t = -\frac{(b-a)^3}{12} f''(\zeta), \quad a < \zeta < b$$

Where  $\zeta$  is some point in  $[a, b]$ .

What is the error then in the multiple-segment trapezoidal rule? It will be simply the sum of the errors from each segment, where the error in each segment is that of the single segment trapezoidal rule. The error in each segment is

$$E_1 = -\frac{[(a+h)-a]^3}{12} f''(\zeta_1), \quad a < \zeta_1 < a+h \\ = -\frac{h^3}{12} f''(\zeta_1)$$

$$E_2 = -\frac{[(a+2h)-(a+h)]^3}{12} f''(\zeta_2), \quad a+h < \zeta_2 < a+2h \\ = -\frac{h^3}{12} f''(\zeta_2)$$

$$\begin{aligned}
E_i &= -\frac{[(a+ih)-(a+(i-1)h)]^3}{12} f''(\zeta_i), \quad a+(i-1)h < \zeta_i < a+ih \\
&= -\frac{h^3}{12} f''(\zeta_i) \\
&\dots \\
E_{n-1} &= -\frac{[a+(n-1)h] - [a+(n-2)h]^3}{12} f''(\zeta_{n-1}), \quad a+(n-2)h < \zeta_{n-1} < a+(n-1)h \\
&= -\frac{h^3}{12} f''(\zeta_{n-1}) \\
E_n &= -\frac{[b-a+(n-1)h]^3}{12} f''(\zeta_n), \quad a+(n-1)h < \zeta_n < b \\
&= -\frac{h^3}{12} f''(\zeta_n)
\end{aligned}$$

Hence the total error in the multiple-segment trapezoidal rule is

$$\begin{aligned}
E_t &= \sum_{i=1}^n E_i \\
&= -\frac{h^3}{12} \sum_{i=1}^n f''(\zeta_i) \\
&= -\frac{(b-a)^3}{12n^3} \sum_{i=1}^n f''(\zeta_i) \\
&= -\frac{(b-a)^3}{12n^2} \frac{\sum_{i=1}^n f''(\zeta_i)}{n} \\
&= -\frac{(b-a)^3}{12n^2} \overline{f''(\zeta_i)}
\end{aligned}$$

The term  $\frac{\sum_{i=1}^n f''(\zeta_i)}{n}$  is an approximate average value of the second derivative  $f''(x)$ ,  $a < x < b$ .

Hence

$$E_t = -\frac{(b-a)^3}{12n^2} \frac{\sum_{i=1}^n f''(\zeta_i)}{n}$$

In Table 4, the approximate value of the integral

$$\int_8^{30} \left( 2000 \ln \left[ \frac{140000}{140000 - 2100t} \right] - 9.8t \right) dt$$

is given as a function of the number of segments. You can visualize that as the number of segments are doubled, the true error gets approximately quartered.

**Table 4** Values obtained using multiple-segment trapezoidal rule for

$$x = \int_8^{30} \left( 2000 \ln \left[ \frac{140000}{140000 - 2100t} \right] - 9.8t \right) dt.$$

$n$	Approximate Value	$E_t$	$ E_t  \%$	$ E_a  \%$
2	11266	-205	1.853	5.343
4	11113	-52	0.4701	0.3594
8	11074	-13	0.1175	0.03560
16	11065	-4	0.03616	0.00401

For example, for the 2-segment trapezoidal rule, the true error is -205, and a quarter of that error is -51.25. That is close to the true error of -48 for the 4-segment trapezoidal rule.

Can you answer the question *why is the true error not exactly -51.25?* How does this information help us in numerical integration? You will find out that this forms the basis of Romberg integration based on the trapezoidal rule, where we use the argument that true error gets approximately quartered when the number of segments is doubled. Romberg integration based on the trapezoidal rule is computationally more efficient than using the trapezoidal rule by itself in developing an automatic integration scheme.

---

## INTEGRATION

---

Topic	Trapezoidal Rule
Summary	These are textbook notes of trapezoidal rule of integration
Major	General Engineering
Authors	Autar Kaw, Michael Keteltas
Date	January 15, 2012
Web Site	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

---

## Chapter 07.03

### Simpson's 1/3 Rule of Integration

*After reading this chapter, you should be able to*

1. derive the formula for Simpson's 1/3 rule of integration,
2. use Simpson's 1/3 rule it to solve integrals,
3. develop the formula for multiple-segment Simpson's 1/3 rule of integration,
4. use multiple-segment Simpson's 1/3 rule of integration to solve integrals, and
5. derive the true error formula for multiple-segment Simpson's 1/3 rule.

#### What is integration?

Integration is the process of measuring the area under a function plotted on a graph. Why would we want to integrate a function? Among the most common examples are finding the velocity of a body from an acceleration function, and displacement of a body from a velocity function. Throughout many engineering fields, there are (what sometimes seems like) countless applications for integral calculus. You can read about some of these applications in Chapters 07.00A-07.00G.

Sometimes, the evaluation of expressions involving these integrals can become daunting, if not indeterminate. For this reason, a wide variety of numerical methods has been developed to simplify the integral. Here, we will discuss Simpson's 1/3 rule of integral approximation, which improves upon the accuracy of the trapezoidal rule.

Here, we will discuss the Simpson's 1/3 rule of approximating integrals of the form

$$I = \int_a^b f(x)dx$$

where

$f(x)$  is called the integrand,

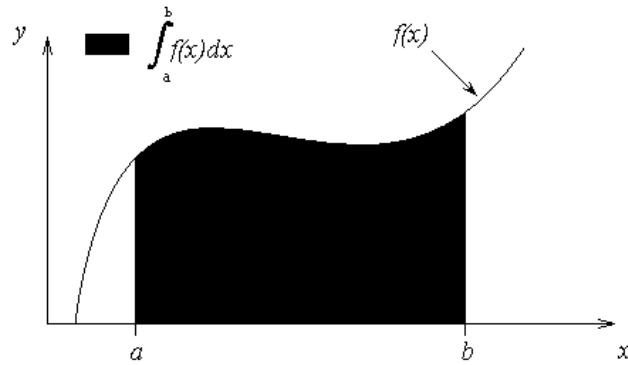
$a$  = lower limit of integration

$b$  = upper limit of integration

#### Simpson's 1/3 Rule

The trapezoidal rule was based on approximating the integrand by a first order polynomial, and then integrating the polynomial over interval of integration. Simpson's 1/3 rule is an

extension of Trapezoidal rule where the integrand is approximated by a second order polynomial.



**Figure 1** Integration of a function

### Method 1:

Hence

$$I = \int_a^b f(x) dx \approx \int_a^b f_2(x) dx$$

where  $f_2(x)$  is a second order polynomial given by

$$f_2(x) = a_0 + a_1x + a_2x^2$$

Choose

$$(a, f(a)), \left( \frac{a+b}{2}, f\left(\frac{a+b}{2}\right) \right), \text{ and } (b, f(b))$$

as the three points of the function to evaluate  $a_0$ ,  $a_1$  and  $a_2$ .

$$f(a) = f_2(a) = a_0 + a_1a + a_2a^2$$

$$f\left(\frac{a+b}{2}\right) = f_2\left(\frac{a+b}{2}\right) = a_0 + a_1\left(\frac{a+b}{2}\right) + a_2\left(\frac{a+b}{2}\right)^2$$

$$f(b) = f_2(b) = a_0 + a_1b + a_2b^2$$

Solving the above three equations for unknowns,  $a_0$ ,  $a_1$  and  $a_2$  give

$$a_0 = \frac{a^2 f(b) + abf(b) - 4abf\left(\frac{a+b}{2}\right) + abf(a) + b^2 f(a)}{a^2 - 2ab + b^2}$$

$$a_1 = -\frac{af(a) - 4af\left(\frac{a+b}{2}\right) + 3af(b) + 3bf(a) - 4bf\left(\frac{a+b}{2}\right) + bf(b)}{a^2 - 2ab + b^2}$$

$$a_2 = \frac{2\left(f(a) - 2f\left(\frac{a+b}{2}\right) + f(b)\right)}{a^2 - 2ab + b^2}$$

Then

$$\begin{aligned} I &\approx \int_a^b f_2(x)dx \\ &= \int_a^b (a_0 + a_1x + a_2x^2)dx \\ &= \left[ a_0x + a_1 \frac{x^2}{2} + a_2 \frac{x^3}{3} \right]_a^b \\ &= a_0(b-a) + a_1 \frac{b^2 - a^2}{2} + a_2 \frac{b^3 - a^3}{3} \end{aligned}$$

Substituting values of  $a_0$ ,  $a_1$  and  $a_2$  give

$$\int_a^b f_2(x)dx = \frac{b-a}{6} \left[ f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right]$$

Since for Simpson 1/3 rule, the interval  $[a, b]$  is broken into 2 segments, the segment width

$$h = \frac{b-a}{2}$$

Hence the Simpson's 1/3 rule is given by

$$\int_a^b f(x)dx \approx \frac{h}{3} \left[ f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right]$$

Since the above form has 1/3 in its formula, it is called Simpson's 1/3 rule.

### Method 2:

Simpson's 1/3 rule can also be derived by approximating  $f(x)$  by a second order polynomial using Newton's divided difference polynomial as

$$f_2(x) = b_0 + b_1(x-a) + b_2(x-a)\left(x - \frac{a+b}{2}\right)$$

where

$$\begin{aligned} b_0 &= f(a) \\ b_1 &= \frac{f\left(\frac{a+b}{2}\right) - f(a)}{\frac{a+b}{2} - a} \end{aligned}$$

$$b_2 = \frac{\frac{f(b) - f\left(\frac{a+b}{2}\right)}{b - \frac{a+b}{2}} - \frac{f\left(\frac{a+b}{2}\right) - f(a)}{\frac{a+b}{2} - a}}{b - a}$$

Integrating Newton's divided difference polynomial gives us

$$\begin{aligned} \int_a^b f(x)dx &\approx \int_a^b f_2(x)dx \\ &= \int_a^b \left[ b_0 + b_1(x-a) + b_2(x-a)\left(x-\frac{a+b}{2}\right) \right] dx \\ &= \left[ b_0x + b_1\left(\frac{x^2}{2} - ax\right) + b_2\left(\frac{x^3}{3} - \frac{(3a+b)x^2}{4} + \frac{a(a+b)x}{2}\right) \right]_a^b \\ &= b_0(b-a) + b_1\left(\frac{b^2 - a^2}{2} - a(b-a)\right) \\ &\quad + b_2\left(\frac{b^3 - a^3}{3} - \frac{(3a+b)(b^2 - a^2)}{4} + \frac{a(a+b)(b-a)}{2}\right) \end{aligned}$$

Substituting values of  $b_0$ ,  $b_1$ , and  $b_2$  into this equation yields the same result as before

$$\begin{aligned} \int_a^b f(x)dx &\approx \frac{b-a}{6} \left[ f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right] \\ &= \frac{h}{3} \left[ f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right] \end{aligned}$$

### Method 3:

One could even use the Lagrange polynomial to derive Simpson's formula. Notice any method of three-point quadratic interpolation can be used to accomplish this task. In this case, the interpolating function becomes

$$f_2(x) = \frac{\left(x - \frac{a+b}{2}\right)(x-b)}{\left(a - \frac{a+b}{2}\right)(a-b)} f(a) + \frac{(x-a)(x-b)}{\left(\frac{a+b}{2} - a\right)\left(\frac{a+b}{2} - b\right)} f\left(\frac{a+b}{2}\right) + \frac{(x-a)\left(x - \frac{a+b}{2}\right)}{(b-a)\left(b - \frac{a+b}{2}\right)} f(b)$$

Integrating this function gets

$$\int_a^b f_2(x)dx = \left[ \frac{\frac{x^3}{3} - \frac{(a+3b)x^2}{4} + \frac{b(a+b)x}{2}}{\left(a - \frac{a+b}{2}\right)(a-b)} f(a) + \frac{\frac{x^3}{3} - \frac{(a+b)x^2}{2} + abx}{\left(\frac{a+b}{2} - a\right)\left(\frac{a+b}{2} - b\right)} f\left(\frac{a+b}{2}\right) \right]_a^b$$

$$+ \frac{\frac{x^3}{3} - \frac{(3a+b)x^2}{4} + \frac{a(a+b)x}{2}}{(b-a)\left(b - \frac{a+b}{2}\right)} f(b)$$

$$= \frac{\frac{b^3 - a^3}{3} - \frac{(a+3b)(b^2 - a^2)}{4} + \frac{b(a+b)(b-a)}{2}}{\left(a - \frac{a+b}{2}\right)(a-b)} f(a)$$

$$+ \frac{\frac{b^3 - a^3}{3} - \frac{(a+b)(b^2 - a^2)}{2} + ab(b-a)}{\left(\frac{a+b}{2} - a\right)\left(\frac{a+b}{2} - b\right)} f\left(\frac{a+b}{2}\right)$$

$$+ \frac{\frac{b^3 - a^3}{3} - \frac{(3a+b)(b^2 - a^2)}{4} + \frac{a(a+b)(b-a)}{2}}{(b-a)\left(b - \frac{a+b}{2}\right)} f(b)$$

Believe it or not, simplifying and factoring this large expression yields you the same result as before

$$\int_a^b f(x)dx \approx \frac{b-a}{6} \left[ f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right]$$

$$= \frac{h}{3} \left[ f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right].$$

#### Method 4:

Simpson's 1/3 rule can also be derived by the method of coefficients. Assume

$$\int_a^b f(x)dx \approx c_1 f(a) + c_2 f\left(\frac{a+b}{2}\right) + c_3 f(b)$$

Let the right-hand side be an exact expression for the integrals  $\int_a^b 1dx$ ,  $\int_a^b xdx$ , and  $\int_a^b x^2dx$ . This

implies that the right hand side will be exact expressions for integrals of any linear combination of the three integrals for a general second order polynomial. Now

$$\int_a^b 1dx = b - a = c_1 + c_2 + c_3$$

$$\int_a^b x dx = \frac{b^2 - a^2}{2} = c_1 a + c_2 \frac{a+b}{2} + c_3 b$$

$$\int_a^b x^2 dx = \frac{b^3 - a^3}{3} = c_1 a^2 + c_2 \left( \frac{a+b}{2} \right)^2 + c_3 b^2$$

Solving the above three equations for  $c_0$ ,  $c_1$  and  $c_2$  give

$$c_1 = \frac{b-a}{6}$$

$$c_2 = \frac{2(b-a)}{3}$$

$$c_3 = \frac{b-a}{6}$$

This gives

$$\begin{aligned} \int_a^b f(x) dx &\approx \frac{b-a}{6} f(a) + \frac{2(b-a)}{3} f\left(\frac{a+b}{2}\right) + \frac{b-a}{6} f(b) \\ &= \frac{b-a}{6} \left[ f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right] \\ &= \frac{h}{3} \left[ f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right] \end{aligned}$$

The integral from the first method

$$\int_a^b f(x) dx \approx \int_a^b (a_0 + a_1 x + a_2 x^2) dx$$

can be viewed as the area under the second order polynomial, while the equation from Method 4

$$\int_a^b f(x) dx \approx \frac{b-a}{6} f(a) + \frac{2(b-a)}{3} f\left(\frac{a+b}{2}\right) + \frac{b-a}{6} f(b)$$

can be viewed as the sum of the areas of three rectangles.

### Example 1

The distance covered by a rocket in meters from  $t = 8\text{ s}$  to  $t = 30\text{ s}$  is given by

$$x = \int_8^{30} \left( 2000 \ln \left[ \frac{140000}{140000 - 2100t} \right] - 9.8t \right) dt$$

- Use Simpson's 1/3 rule to find the approximate value of  $x$ .
- Find the true error,  $E_t$ .
- Find the absolute relative true error,  $|e_r|$ .

**Solution**

$$\text{a) } x \approx \frac{b-a}{6} \left[ f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right]$$

$$a = 8$$

$$b = 30$$

$$\frac{a+b}{2} = 19$$

$$f(t) = 2000 \ln \left[ \frac{140000}{140000 - 2100t} \right] - 9.8t$$

$$f(8) = 2000 \ln \left[ \frac{140000}{140000 - 2100(8)} \right] - 9.8(8) = 177.27 \text{ m/s}$$

$$f(30) = 2000 \ln \left[ \frac{140000}{140000 - 2100(30)} \right] - 9.8(30) = 901.67 \text{ m/s}$$

$$f(19) = 2000 \ln \left[ \frac{140000}{140000 - 2100(19)} \right] - 9.8(19) = 484.75 \text{ m/s}$$

$$x \approx \frac{b-a}{6} \left[ f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right]$$

$$= \left( \frac{30-8}{6} \right) [f(8) + 4f(19) + f(30)]$$

$$= \frac{22}{6} [177.27 + 4 \times 484.75 + 901.67]$$

$$= 11065.72 \text{ m}$$

b) The exact value of the above integral is

$$\begin{aligned} x &= \int_8^{30} \left( 2000 \ln \left[ \frac{140000}{140000 - 2100t} \right] - 9.8t \right) dt \\ &= 11061.34 \text{ m} \end{aligned}$$

So the true error is

$$\begin{aligned} E_t &= \text{True Value} - \text{Approximate Value} \\ &= 11061.34 - 11065.72 \\ &= -4.38 \text{ m} \end{aligned}$$

c) The absolute relative true error is

$$\begin{aligned} |E_r| &= \left| \frac{\text{True Error}}{\text{True Value}} \right| \times 100 \\ &= \left| \frac{-4.38}{11061.34} \right| \times 100 \end{aligned}$$

$$= 0.0396\%$$

### Multiple-segment Simpson's 1/3 Rule

Just like in multiple-segment trapezoidal rule, one can subdivide the interval  $[a, b]$  into  $n$  segments and apply Simpson's 1/3 rule repeatedly over every two segments. Note that  $n$  needs to be even. Divide interval  $[a, b]$  into  $n$  equal segments, so that the segment width is given by

$$h = \frac{b-a}{n}.$$

Now

$$\int_a^b f(x)dx = \int_{x_0}^{x_n} f(x)dx$$

where

$$x_0 = a$$

$$x_n = b$$

$$\int_a^b f(x)dx = \int_{x_0}^{x_2} f(x)dx + \int_{x_2}^{x_4} f(x)dx + \dots + \int_{x_{n-4}}^{x_n} f(x)dx$$

Apply Simpson's 1/3rd Rule over each interval,

$$\begin{aligned} \int_a^b f(x)dx &\approx (x_2 - x_0) \left[ \frac{f(x_0) + 4f(x_1) + f(x_2)}{6} \right] + (x_4 - x_2) \left[ \frac{f(x_2) + 4f(x_3) + f(x_4)}{6} \right] + \dots \\ &+ (x_{n-2} - x_{n-4}) \left[ \frac{f(x_{n-4}) + 4f(x_{n-3}) + f(x_{n-2})}{6} \right] + (x_n - x_{n-2}) \left[ \frac{f(x_{n-2}) + 4f(x_{n-1}) + f(x_n)}{6} \right] \end{aligned}$$

Since

$$x_i - x_{i-2} = 2h$$

$$i = 2, 4, \dots, n$$

then

$$\begin{aligned} \int_a^b f(x)dx &\approx 2h \left[ \frac{f(x_0) + 4f(x_1) + f(x_2)}{6} \right] + 2h \left[ \frac{f(x_2) + 4f(x_3) + f(x_4)}{6} \right] + \dots \\ &+ 2h \left[ \frac{f(x_{n-4}) + 4f(x_{n-3}) + f(x_{n-2})}{6} \right] + 2h \left[ \frac{f(x_{n-2}) + 4f(x_{n-1}) + f(x_n)}{6} \right] \end{aligned}$$

$$= \frac{h}{3} [f(x_0) + 4\{f(x_1) + f(x_3) + \dots + f(x_{n-1})\} + 2\{f(x_2) + f(x_4) + \dots + f(x_{n-2})\} + f(x_n)]$$

$$\begin{aligned}
 &= \frac{h}{3} \left[ f(x_0) + 4 \sum_{\substack{i=1 \\ i=odd}}^{n-1} f(x_i) + 2 \sum_{\substack{i=2 \\ i=even}}^{n-2} f(x_i) + f(x_n) \right] \\
 \int_a^b f(x) dx &\equiv \frac{b-a}{3n} \left[ f(x_0) + 4 \sum_{\substack{i=1 \\ i=odd}}^{n-1} f(x_i) + 2 \sum_{\substack{i=2 \\ i=even}}^{n-2} f(x_i) + f(x_n) \right]
 \end{aligned}$$

**Example 2**

Use 4-segment Simpson's 1/3 rule to approximate the distance covered by a rocket in meters from  $t = 8$  s to  $t = 30$  s as given by

$$x = \int_8^{30} \left( 2000 \ln \left[ \frac{140000}{140000 - 2100t} \right] - 9.8t \right) dt$$

- a) Use four segment Simpson's 1/3rd Rule to estimate  $x$ .
- b) Find the true error,  $E_t$  for part (a).
- c) Find the absolute relative true error,  $|e_t|$  for part (a).

**Solution:**

- a) Using  $n$  segment Simpson's 1/3 rule,

$$x \approx \frac{b-a}{3n} \left[ f(t_0) + 4 \sum_{\substack{i=1 \\ i=odd}}^{n-1} f(t_i) + 2 \sum_{\substack{i=2 \\ i=even}}^{n-2} f(t_i) + f(t_n) \right]$$

$$n = 4$$

$$a = 8$$

$$b = 30$$

$$h = \frac{b-a}{n}$$

$$= \frac{30-8}{4}$$

$$= 5.5$$

$$f(t) = 2000 \ln \left[ \frac{140000}{140000 - 2100t} \right] - 9.8t$$

So

$$f(t_0) = f(8)$$

$$f(8) = 2000 \ln \left[ \frac{140000}{140000 - 2100(8)} \right] - 9.8(8) = 177.27 \text{ m/s}$$

$$f(t_1) = f(8 + 5.5) = f(13.5)$$

$$f(13.5) = 2000 \ln \left[ \frac{140000}{140000 - 2100(13.5)} \right] - 9.8(13.5) = 320.25 \text{ m/s}$$

$$f(t_2) = f(13.5 + 5.5) = f(19)$$

$$f(19) = 2000 \ln \left( \frac{140000}{140000 - 2100(19)} \right) - 9.8(19) = 484.75 \text{ m/s}$$

$$f(t_3) = f(19 + 5.5) = f(24.5)$$

$$f(24.5) = 2000 \ln \left[ \frac{140000}{140000 - 2100(24.5)} \right] - 9.8(24.5) = 676.05 \text{ m/s}$$

$$f(t_4) = f(t_n) = f(30)$$

$$f(30) = 2000 \ln \left[ \frac{140000}{140000 - 2100(30)} \right] - 9.8(30) = 901.67 \text{ m/s}$$

$$\begin{aligned} x &= \frac{b-a}{3n} \left[ f(t_0) + 4 \sum_{\substack{i=1 \\ i=odd}}^{n-1} f(t_i) + 2 \sum_{\substack{i=2 \\ i=even}}^{n-2} f(t_i) + f(t_n) \right] \\ &= \frac{30-8}{3(4)} \left[ f(8) + 4 \sum_{\substack{i=1 \\ i=odd}}^3 f(t_i) + 2 \sum_{\substack{i=2 \\ i=even}}^2 f(t_i) + f(30) \right] \\ &= \frac{22}{12} [f(8) + 4f(t_1) + 4f(t_3) + 2f(t_2) + f(30)] \\ &= \frac{11}{6} [f(8) + 4f(13.5) + 4f(24.5) + 2f(19) + f(30)] \\ &= \frac{11}{6} [177.27 + 4(320.25) + 4(676.05) + 2(484.75) + 901.67] \\ &= 11061.64 \text{ m} \end{aligned}$$

b) The exact value of the above integral is

$$\begin{aligned} x &= \int_8^{30} \left( 2000 \ln \left[ \frac{140000}{140000 - 2100t} \right] - 9.8t \right) dt \\ &= 11061.34 \text{ m} \end{aligned}$$

So the true error is

$$E_t = \text{True Value} - \text{Approximate Value}$$

$$E_t = 11061.34 - 11061.64$$

$$= -0.30 \text{ m}$$

c) The absolute relative true error is

$$\begin{aligned} |\epsilon_t| &= \left| \frac{\text{True Error}}{\text{True Value}} \right| \times 100 \\ &= \left| \frac{-0.3}{11061.34} \right| \times 100 \\ &= 0.0027\% \end{aligned}$$

**Table 1** Values of Simpson's 1/3 rule for Example 2 with multiple-segments

$n$	Approximate Value	$E_t$	$ \epsilon_t $
2	11065.72	-4.38	0.0396%
4	11061.64	-0.30	0.0027%
6	11061.40	-0.06	0.0005%
8	11061.35	-0.02	0.0002%
10	11061.34	-0.01	0.0001%

### Error in Multiple-segment Simpson's 1/3 rule

The true error in a single application of Simpson's 1/3rd Rule is given<sup>1</sup> by

$$E_t = -\frac{(b-a)^5}{2880} f^{(4)}(\zeta), \quad a < \zeta < b$$

In multiple-segment Simpson's 1/3 rule, the error is the sum of the errors in each application of Simpson's 1/3 rule. The error in the  $n$  segments Simpson's 1/3rd Rule is given by

$$\begin{aligned} E_1 &= -\frac{(x_2 - x_0)^5}{2880} f^{(4)}(\zeta_1), \quad x_0 < \zeta_1 < x_2 \\ &= -\frac{h^5}{90} f^{(4)}(\zeta_1) \\ E_2 &= -\frac{(x_4 - x_2)^5}{2880} f^{(4)}(\zeta_2), \quad x_2 < \zeta_2 < x_4 \\ &= -\frac{h^5}{90} f^{(4)}(\zeta_2) \\ &\vdots \\ E_i &= -\frac{(x_{2i} - x_{2(i-1)})^5}{2880} f^{(4)}(\zeta_i), \quad x_{2(i-1)} < \zeta_i < x_{2i} \\ &= -\frac{h^5}{90} f^{(4)}(\zeta_i) \\ &\vdots \end{aligned}$$


---

<sup>1</sup> The  $f^{(4)}$  in the true error expression stands for the fourth derivative of the function  $f(x)$ .

$$\begin{aligned} E_{\frac{n}{2}-1} &= -\frac{(x_{n-2} - x_{n-4})^5}{2880} f^{(4)}\left(\zeta_{\frac{n}{2}-1}\right), \quad x_{n-4} < \zeta_{\frac{n}{2}-1} < x_{n-2} \\ &= -\frac{h^5}{90} f^{(4)}\left(\zeta_{\frac{n}{2}-1}\right) \\ E_{\frac{n}{2}} &= -\frac{(x_n - x_{n-2})^5}{2880} f^{(4)}\left(\zeta_{\frac{n}{2}}\right), \quad x_{n-2} < \zeta_{\frac{n}{2}} < x_n \end{aligned}$$

Hence, the total error in the multiple-segment Simpson's 1/3 rule is

$$\begin{aligned} &= -\frac{h^5}{90} f^{(4)}\left(\zeta_{\frac{n}{2}}\right) \\ E_t &= \sum_{i=1}^{\frac{n}{2}} E_i \\ &= -\frac{h^5}{90} \sum_{i=1}^{\frac{n}{2}} f^{(4)}(\zeta_i) \\ &= -\frac{(b-a)^5}{90n^5} \sum_{i=1}^{\frac{n}{2}} f^{(4)}(\zeta_i) \\ &= -\frac{(b-a)^5}{180n^4} \frac{\sum_{i=1}^{\frac{n}{2}} f^{(4)}(\zeta_i)}{\frac{n}{2}}, \end{aligned}$$

$$\sum_{i=1}^{\frac{n}{2}} f^{(4)}(\zeta_i)$$

The term  $\frac{\sum_{i=1}^{\frac{n}{2}} f^{(4)}(\zeta_i)}{\frac{n}{2}}$  is an approximate average value of  $f^{(4)}(x)$ ,  $a < x < b$ . Hence

$$E_t = -\frac{(b-a)^5}{180n^4} \bar{f}^{(4)}$$

where

$$\bar{f}^{(4)} = \frac{\sum_{i=1}^{\frac{n}{2}} f^{(4)}(\zeta_i)}{\frac{n}{2}}$$

## INTEGRATION

Topic	Simpson's 1/3 rule
Summary	Textbook notes of Simpson's 1/3 rule
Major	General Engineering
Authors	Autar Kaw, Michael Keteltas

---

Date December 3, 2017  
Web Site <http://numericalmethods.eng.usf.edu>

---

# **Chapter 07.04**

## **Romberg Rule of Integration**

*After reading this chapter, you should be able to:*

1. *derive the Romberg rule of integration, and*
2. *use the Romberg rule of integration to solve problems.*

### **What is integration?**

Integration is the process of measuring the area under a function plotted on a graph. Why would we want to integrate a function? Among the most common examples are finding the velocity of a body from an acceleration function, and displacement of a body from a velocity function. Throughout many engineering fields, there are (what sometimes seems like) countless applications for integral calculus. You can read about some of these applications in Chapters 07.00A-07.00G.

Sometimes, the evaluation of expressions involving these integrals can become daunting, if not indeterminate. For this reason, a wide variety of numerical methods has been developed to simplify the integral.

Here, we will discuss the Romberg rule of approximating integrals of the form

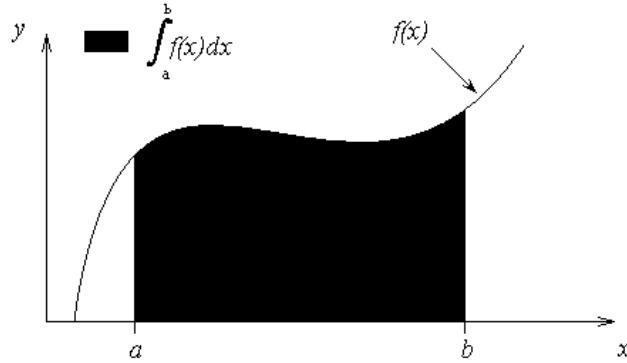
$$I = \int_a^b f(x)dx \quad (1)$$

where

$f(x)$  is called the integrand

$a$  = lower limit of integration

$b$  = upper limit of integration



**Figure 1** Integration of a function.

### Error in Multiple-Segment Trapezoidal Rule

The true error obtained when using the multiple segment trapezoidal rule with  $n$  segments to approximate an integral

$$\int_a^b f(x) dx$$

is given by

$$E_t = -\frac{(b-a)^3}{12n^2} \frac{\sum_{i=1}^n f''(\xi_i)}{n} \quad (2)$$

where for each  $i$ ,  $\xi_i$  is a point somewhere in the domain  $[a + (i-1)h, a + ih]$ , and

the term  $\frac{\sum_{i=1}^n f''(\xi_i)}{n}$  can be viewed as an approximate average value of  $f''(x)$  in  $[a, b]$ . This leads us to say that the true error  $E_t$  in Equation (2) is approximately proportional to

$$E_t \approx \alpha \frac{1}{n^2} \quad (3)$$

for the estimate of  $\int_a^b f(x) dx$  using the  $n$ -segment trapezoidal rule.

Table 1 shows the results obtained for

$$\int_8^{30} \left( 2000 \ln \left[ \frac{140000}{140000 - 2100t} \right] - 9.8t \right) dt$$

using the multiple-segment trapezoidal rule.

**Table 1** Values obtained using multiple segment trapezoidal rule for

$$x = \int_8^{30} \left( 2000 \ln \left[ \frac{140000}{140000 - 2100t} \right] - 9.8t \right) dt .$$

$n$	Approximate Value	$E_t$	$ e_t  \%$	$ e_a  \%$
1	11868	-807	7.296	---
2	11266	-205	1.854	5.343
3	11153	-91.4	0.8265	1.019
4	11113	-51.5	0.4655	0.3594
5	11094	-33.0	0.2981	0.1669
6	11084	-22.9	0.2070	0.09082
7	11078	-16.8	0.1521	0.05482
8	11074	-12.9	0.1165	0.03560

The true error for the 1-segment trapezoidal rule is  $-807$ , while for the 2-segment rule, the true error is  $-205$ . The true error of  $-205$  is approximately a quarter of  $-807$ . The true error gets approximately quartered as the number of segments is doubled from 1 to 2. The same trend is observed when the number of segments is doubled from 2 to 4 (the true error for 2-segments is  $-205$  and for four segments is  $-51.5$ ). This follows Equation (3). This information, although interesting, can also be used to get a better approximation of the integral. That is the basis of Richardson's extrapolation formula for integration by the trapezoidal rule.

### Richardson's Extrapolation Formula for Trapezoidal Rule

The true error,  $E_t$ , in the  $n$ -segment trapezoidal rule is estimated as

$$\begin{aligned} E_t &\approx \alpha \frac{1}{n^2} \\ E_t &\approx \frac{C}{n^2} \end{aligned} \tag{4}$$

where  $C$  is an approximate constant of proportionality.

Since

$$E_t = TV - I_n \tag{5}$$

where

$TV$  = true value

$I_n$  = approximate value using  $n$ -segments

Then from Equations (4) and (5),

$$\frac{C}{n^2} \approx TV - I_n \tag{6}$$

If the number of segments is doubled from  $n$  to  $2n$  in the trapezoidal rule,

$$\frac{C}{(2n)^2} \approx TV - I_{2n} \tag{7}$$

Equations (6) and (7) can be solved simultaneously to get

$$TV \approx I_{2n} + \frac{I_{2n} - I_n}{3} \quad (8)$$

### Example 1

The vertical distance in meters covered by a rocket from  $t = 8$  to  $t = 30$  seconds is given by

$$x = \int_8^{30} \left( 2000 \ln \left[ \frac{140000}{140000 - 2100t} \right] - 9.8t \right) dt$$

- a) Use Romberg's rule to find the distance covered. Use the 2-segment and 4-segment trapezoidal rule results given in Table 1.
- b) Find the true error for part (a).
- c) Find the absolute relative true error for part (a).

### Solution

a)  $I_2 = 11266 \text{ m}$

$I_4 = 11113 \text{ m}$

Using Richardson's extrapolation formula for the trapezoidal rule, the true value is given by

$$TV \approx I_{2n} + \frac{I_{2n} - I_n}{3}$$

and choosing  $n=2$ ,

$$\begin{aligned} TV &\approx I_4 + \frac{I_4 - I_2}{3} \\ &= 11113 + \frac{11113 - 11266}{3} \\ &= 11062 \text{ m} \end{aligned}$$

b) The exact value of the above integral is

$$\begin{aligned} x &= \int_8^{30} \left( 2000 \ln \left[ \frac{140000}{140000 - 2100t} \right] - 9.8t \right) dt \\ &= 11061 \text{ m} \end{aligned}$$

so the true error

$$\begin{aligned} E_t &= \text{True Value} - \text{Approximate Value} \\ &= 11061 - 11062 \\ &= -1 \text{ m} \end{aligned}$$

c) The absolute relative true error,  $|e_t|$ , would then be

$$\begin{aligned} |e_t| &= \left| \frac{\text{True Error}}{\text{True Value}} \right| \times 100 \\ &= \left| \frac{11061 - 11062}{11061} \right| \times 100 \\ &= 0.00904\% \end{aligned}$$

Table 2 shows the Richardson's extrapolation results using 1, 2, 4, and 8 segments. Results are compared with those of the trapezoidal rule.

**Table 2** Values obtained using Richardson's extrapolation formula for the trapezoidal rule for

$$x = \int_8^{30} \left( 2000 \ln \left[ \frac{140000}{140000 - 2100t} \right] - 9.8t \right) dt.$$

$n$	Trapezoidal Rule	$ E_t /\%$ for Trapezoidal Rule	Richardson's Extrapolation	$ E_t /\%$ for Richardson's Extrapolation
1	11868	7.296	--	--
2	11266	1.854	11065	0.03616
4	11113	0.4655	11062	0.009041
8	11074	0.1165	11061	0.0000

### Romberg Integration

Romberg integration is the same as Richardson's extrapolation formula as given by Equation (8). However, Romberg used a recursive algorithm for the extrapolation as follows.

The estimate of the true error in the trapezoidal rule is given by

$$E_t = -\frac{(b-a)^3}{12n^2} \sum_{i=1}^n f''(\xi_i)$$

Since the segment width,  $h$ , is given by

$$h = \frac{b-a}{n}$$

Equation (2) can be written as

$$E_t = -\frac{h^2(b-a)}{12} \sum_{i=1}^n f''(\xi_i) \quad (9)$$

The estimate of true error is given by

$$E_t \approx Ch^2 \quad (10)$$

It can be shown that the exact true error could be written as

$$E_t = A_1 h^2 + A_2 h^4 + A_3 h^6 + \dots \quad (11)$$

and for small  $h$ ,

$$E_t = A_1 h^2 + O(h^4) \quad (12)$$

Since we used  $E_t \approx Ch^2$  in the formula (Equation (12)), the result obtained from Equation (10) has an error of  $O(h^4)$  and can be written as

$$\begin{aligned} (I_{2n})_R &= I_{2n} + \frac{I_{2n} - I_n}{3} \\ &= I_{2n} + \frac{I_{2n} - I_n}{4^{2-1} - 1} \end{aligned} \quad (13)$$

where the variable  $TV$  is replaced by  $(I_{2n})_R$  as the value obtained using Richardson's extrapolation formula. Note also that the sign  $\approx$  is replaced by the sign  $=$ . Hence the estimate of the true value now is

$$TV \approx (I_{2n})_R + Ch^4$$

Determine another integral value with further halving the step size (doubling the number of segments),

$$(I_{4n})_R = I_{4n} + \frac{I_{4n} - I_{2n}}{3} \quad (14)$$

then

$$TV \approx (I_{4n})_R + C\left(\frac{h}{2}\right)^4$$

From Equation (13) and (14),

$$\begin{aligned} TV &\approx (I_{4n})_R + \frac{(I_{4n})_R - (I_{2n})_R}{15} \\ &= (I_{4n})_R + \frac{(I_{4n})_R - (I_{2n})_R}{4^{3-1} - 1} \end{aligned} \quad (15)$$

The above equation now has the error of  $O(h^6)$ . The above procedure can be further improved by using the new values of the estimate of the true value that has the error of  $O(h^8)$  to give an estimate of  $O(h^8)$ .

Based on this procedure, a general expression for Romberg integration can be written as

$$I_{k,j} = I_{k-1,j+1} + \frac{I_{k-1,j+1} - I_{k-1,j}}{4^{k-1} - 1}, \quad k \geq 2 \quad (16)$$

The index  $k$  represents the order of extrapolation. For example,  $k=1$  represents the values obtained from the regular trapezoidal rule,  $k=2$  represents the values obtained using the true error estimate as  $O(h^2)$ , etc. The index  $j$  represents the more and less accurate estimate of the integral. The value of an integral with a  $j+1$  index is more accurate than the value of the integral with a  $j$  index.

For  $k=2, j=1$ ,

$$\begin{aligned} I_{2,1} &= I_{1,2} + \frac{I_{1,2} - I_{1,1}}{4^{2-1} - 1} \\ &= I_{1,2} + \frac{I_{1,2} - I_{1,1}}{3} \end{aligned}$$

For  $k=3, j=1$ ,

$$I_{3,1} = I_{2,2} + \frac{I_{2,2} - I_{2,1}}{4^{3-1} - 1}$$

$$= I_{2,2} + \frac{I_{2,2} - I_{2,1}}{15} \quad (17)$$

**Example 2**

The vertical distance in meters covered by a rocket from  $t = 8$  to  $t = 30$  seconds is given by

$$x = \int_8^{30} \left( 2000 \ln \left[ \frac{140000}{140000 - 2100t} \right] - 9.8t \right) dt$$

Use Romberg's rule to find the distance covered. Use the 1, 2, 4, and 8-segment trapezoidal rule results as given in Table 1.

**Solution**

From Table 1, the needed values from the original the trapezoidal rule are

$$I_{1,1} = 11868$$

$$I_{1,2} = 11266$$

$$I_{1,3} = 11113$$

$$I_{1,4} = 11074$$

where the above four values correspond to using 1, 2, 4 and 8 segment trapezoidal rule, respectively. To get the first order extrapolation values,

$$\begin{aligned} I_{2,1} &= I_{1,2} + \frac{I_{1,2} - I_{1,1}}{3} \\ &= 11266 + \frac{11266 - 11868}{3} \\ &= 11065 \end{aligned}$$

Similarly

$$\begin{aligned} I_{2,2} &= I_{1,3} + \frac{I_{1,3} - I_{1,2}}{3} \\ &= 11113 + \frac{11113 - 11266}{3} \\ &= 11062 \\ I_{2,3} &= I_{1,4} + \frac{I_{1,4} - I_{1,3}}{3} \\ &= 11074 + \frac{11074 - 11113}{3} \\ &= 11061 \end{aligned}$$

For the second order extrapolation values,

$$\begin{aligned} I_{3,1} &= I_{2,2} + \frac{I_{2,2} - I_{2,1}}{15} \\ &= 11062 + \frac{11062 - 11065}{15} \\ &= 11062 \end{aligned}$$

Similarly

$$\begin{aligned}
 I_{3,2} &= I_{2,3} + \frac{I_{2,3} - I_{2,2}}{15} \\
 &= 11061 + \frac{11061 - 11062}{15} \\
 &= 11061
 \end{aligned}$$

For the third order extrapolation values,

$$\begin{aligned}
 I_{4,1} &= I_{3,2} + \frac{I_{3,2} - I_{3,1}}{63} \\
 &= 11061 + \frac{11061 - 11062}{63} \\
 &= 11061 \text{ m}
 \end{aligned}$$

Table 3 shows these increasingly correct values in a tree graph.

**Table 3** Improved estimates of the value of an integral using Romberg integration.

		First Order	Second Order	Third Order
1-segment	11868			
2-segment	11266	11065		
4-segment	11113	11062	11062	
8-segment	11074	11061	11061	11061

---

## INTEGRATION

---

Topic      Romberg Rule

Summary    Textbook notes of Romberg Rule of integration.

Major      General Engineering

Authors    Autar Kaw

Date       December 23, 2009

Web Site   <http://numericalmethods.eng.usf.edu>

---

## **Chapter 07.05**

### **Gauss Quadrature Rule of Integration**

*After reading this chapter, you should be able to:*

1. derive the Gauss quadrature method for integration and be able to use it to solve problems, and
2. use Gauss quadrature method to solve examples of approximate integrals.

#### **What is integration?**

Integration is the process of measuring the area under a function plotted on a graph. Why would we want to integrate a function? Among the most common examples are finding the velocity of a body from an acceleration function, and displacement of a body from a velocity function. Throughout many engineering fields, there are (what sometimes seems like) countless applications for integral calculus. You can read about some of these applications in Chapters 07.00A-07.00G.

Sometimes, the evaluation of expressions involving these integrals can become daunting, if not indeterminate. For this reason, a wide variety of numerical methods has been developed to simplify the integral.

Here, we will discuss the Gauss quadrature rule of approximating integrals of the form

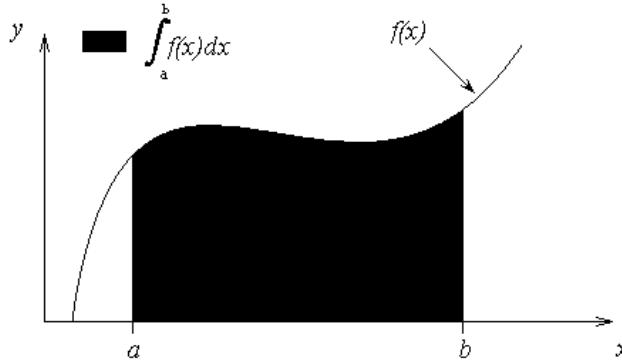
$$I = \int_a^b f(x)dx$$

where

$f(x)$  is called the integrand,

$a$  = lower limit of integration

$b$  = upper limit of integration



**Figure 1** Integration of a function.

### Gauss Quadrature Rule

#### Background:

To derive the trapezoidal rule from the method of undetermined coefficients, we approximated

$$\int_a^b f(x) dx \approx c_1 f(a) + c_2 f(b) \quad (1)$$

Let the right hand side be exact for integrals of a straight line, that is, for an integrated form of

$$\int_a^b (a_0 + a_1 x) dx$$

So

$$\begin{aligned} \int_a^b (a_0 + a_1 x) dx &= \left[ a_0 x + a_1 \frac{x^2}{2} \right]_a^b \\ &= a_0(b-a) + a_1 \left( \frac{b^2 - a^2}{2} \right) \end{aligned} \quad (2)$$

But from Equation (1), we want

$$\int_a^b (a_0 + a_1 x) dx = c_1 f(a) + c_2 f(b) \quad (3)$$

to give the same result as Equation (2) for  $f(x) = a_0 + a_1 x$ .

$$\begin{aligned} \int_a^b (a_0 + a_1 x) dx &= c_1(a_0 + a_1 a) + c_2(a_0 + a_1 b) \\ &= a_0(c_1 + c_2) + a_1(c_1 a + c_2 b) \end{aligned} \quad (4)$$

Hence from Equations (2) and (4),

$$a_0(b-a) + a_1 \left( \frac{b^2 - a^2}{2} \right) = a_0(c_1 + c_2) + a_1(c_1 a + c_2 b)$$

Since  $a_0$  and  $a_1$  are arbitrary constants for a general straight line

$$c_1 + c_2 = b - a \quad (5a)$$

$$c_1 a + c_2 b = \frac{b^2 - a^2}{2} \quad (5b)$$

Multiplying Equation (5a) by  $a$  and subtracting from Equation (5b) gives

$$c_2 = \frac{b - a}{2} \quad (6a)$$

Substituting the above found value of  $c_2$  in Equation (5a) gives

$$c_1 = \frac{b - a}{2} \quad (6b)$$

Therefore

$$\begin{aligned} \int_a^b f(x) dx &\approx c_1 f(a) + c_2 f(b) \\ &= \frac{b - a}{2} f(a) + \frac{b - a}{2} f(b) \end{aligned} \quad (7)$$

### Derivation of two-point Gauss quadrature rule

#### Method 1:

The two-point Gauss quadrature rule is an extension of the trapezoidal rule approximation where the arguments of the function are not predetermined as  $a$  and  $b$ , but as unknowns  $x_1$  and  $x_2$ . So in the two-point Gauss quadrature rule, the integral is approximated as

$$\begin{aligned} I &= \int_a^b f(x) dx \\ &\approx c_1 f(x_1) + c_2 f(x_2) \end{aligned}$$

There are four unknowns  $x_1$ ,  $x_2$ ,  $c_1$  and  $c_2$ . These are found by assuming that the formula gives exact results for integrating a general third order polynomial,  $f(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3$ . Hence

$$\begin{aligned} \int_a^b f(x) dx &= \int_a^b (a_0 + a_1 x + a_2 x^2 + a_3 x^3) dx \\ &= \left[ a_0 x + a_1 \frac{x^2}{2} + a_2 \frac{x^3}{3} + a_3 \frac{x^4}{4} \right]_a^b \\ &= a_0 (b - a) + a_1 \left( \frac{b^2 - a^2}{2} \right) + a_2 \left( \frac{b^3 - a^3}{3} \right) + a_3 \left( \frac{b^4 - a^4}{4} \right) \end{aligned} \quad (8)$$

The formula would then give

$$\begin{aligned} \int_a^b f(x) dx &\approx c_1 f(x_1) + c_2 f(x_2) = \\ &c_1 (a_0 + a_1 x_1 + a_2 x_1^2 + a_3 x_1^3) + c_2 (a_0 + a_1 x_2 + a_2 x_2^2 + a_3 x_2^3) \end{aligned} \quad (9)$$

Equating Equations (8) and (9) gives

$$\begin{aligned} & a_0(b-a) + a_1\left(\frac{b^2-a^2}{2}\right) + a_2\left(\frac{b^3-a^3}{3}\right) + a_3\left(\frac{b^4-a^4}{4}\right) \\ &= c_1(a_0 + a_1x_1 + a_2x_1^2 + a_3x_1^3) + c_2(a_0 + a_1x_2 + a_2x_2^2 + a_3x_2^3) \\ &= a_0(c_1 + c_2) + a_1(c_1x_1 + c_2x_2) + a_2(c_1x_1^2 + c_2x_2^2) + a_3(c_1x_1^3 + c_2x_2^3) \end{aligned} \quad (10)$$

Since in Equation (10), the constants  $a_0$ ,  $a_1$ ,  $a_2$ , and  $a_3$  are arbitrary, the coefficients of  $a_0$ ,  $a_1$ ,  $a_2$ , and  $a_3$  are equal. This gives us four equations as follows.

$$\begin{aligned} b-a &= c_1 + c_2 \\ \frac{b^2-a^2}{2} &= c_1x_1 + c_2x_2 \\ \frac{b^3-a^3}{3} &= c_1x_1^2 + c_2x_2^2 \\ \frac{b^4-a^4}{4} &= c_1x_1^3 + c_2x_2^3 \end{aligned} \quad (11)$$

Without proof (see Example 1 for proof of a related problem), we can find that the above four simultaneous nonlinear equations have only one acceptable solution

$$\begin{aligned} c_1 &= \frac{b-a}{2} \\ c_2 &= \frac{b-a}{2} \\ x_1 &= \left(\frac{b-a}{2}\right)\left(-\frac{1}{\sqrt{3}}\right) + \frac{b+a}{2} \\ x_2 &= \left(\frac{b-a}{2}\right)\left(\frac{1}{\sqrt{3}}\right) + \frac{b+a}{2} \end{aligned} \quad (12)$$

Hence

$$\begin{aligned} \int_a^b f(x)dx &\approx c_1f(x_1) + c_2f(x_2) \\ &= \frac{b-a}{2}f\left(\frac{b-a}{2}\left(-\frac{1}{\sqrt{3}}\right) + \frac{b+a}{2}\right) + \frac{b-a}{2}f\left(\frac{b-a}{2}\left(\frac{1}{\sqrt{3}}\right) + \frac{b+a}{2}\right) \end{aligned} \quad (13)$$

### Method 2:

We can derive the same formula by assuming that the expression gives exact values for the individual integrals of  $\int_a^b 1dx$ ,  $\int_a^b xdx$ ,  $\int_a^b x^2dx$ , and  $\int_a^b x^3dx$ . The reason the formula can also be

derived using this method is that the linear combination of the above integrands is a general third order polynomial given by  $f(x) = a_0 + a_1x + a_2x^2 + a_3x^3$ .

These will give four equations as follows

$$\begin{aligned} \int_a^b 1 dx &= b - a = c_1 + c_2 \\ \int_a^b x dx &= \frac{b^2 - a^2}{2} = c_1 x_1 + c_2 x_2 \\ \int_a^b x^2 dx &= \frac{b^3 - a^3}{3} = c_1 x_1^2 + c_2 x_2^2 \\ \int_a^b x^3 dx &= \frac{b^4 - a^4}{4} = c_1 x_1^3 + c_2 x_2^3 \end{aligned} \quad (14)$$

These four simultaneous nonlinear equations can be solved to give a single acceptable solution

$$\begin{aligned} c_1 &= \frac{b-a}{2} \\ c_2 &= \frac{b-a}{2} \\ x_1 &= \left( \frac{b-a}{2} \right) \left( -\frac{1}{\sqrt{3}} \right) + \frac{b+a}{2} \\ x_2 &= \left( \frac{b-a}{2} \right) \left( \frac{1}{\sqrt{3}} \right) + \frac{b+a}{2} \end{aligned} \quad (15)$$

Hence

$$\int_a^b f(x) dx \approx \frac{b-a}{2} f\left(\frac{b-a}{2}\left(-\frac{1}{\sqrt{3}}\right) + \frac{b+a}{2}\right) + \frac{b-a}{2} f\left(\frac{b-a}{2}\left(\frac{1}{\sqrt{3}}\right) + \frac{b+a}{2}\right) \quad (16)$$

Since two points are chosen, it is called the two-point Gauss quadrature rule. Higher point versions can also be developed.

### Higher point Gauss quadrature formulas

For example

$$\int_a^b f(x) dx \approx c_1 f(x_1) + c_2 f(x_2) + c_3 f(x_3) \quad (17)$$

is called the three-point Gauss quadrature rule. The coefficients  $c_1$ ,  $c_2$  and  $c_3$ , and the function arguments  $x_1$ ,  $x_2$  and  $x_3$  are calculated by assuming the formula gives exact expressions for integrating a fifth order polynomial

$$\int_a^b (a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4 + a_5x^5) dx .$$

General  $n$ -point rules would approximate the integral

$$\int_a^b f(x)dx \approx c_1f(x_1) + c_2f(x_2) + \dots + c_nf(x_n) \quad (18)$$

### Arguments and weighing factors for $n$ -point Gauss quadrature rules

In handbooks (see Table 1), coefficients and arguments given for  $n$ -point Gauss quadrature rule are given for integrals of the form

$$\int_{-1}^1 g(x)dx \approx \sum_{i=1}^n c_i g(x_i) \quad (19)$$

**Table 1** Weighting factors  $c$  and function arguments  $x$  used in Gauss quadrature formulas

Points	Weighting Factors	Function Arguments
2	$c_1 = 1.000000000$	$x_1 = -0.577350269$
	$c_2 = 1.000000000$	$x_2 = 0.577350269$
3	$c_1 = 0.555555556$	$x_1 = -0.774596669$
	$c_2 = 0.888888889$	$x_2 = 0.000000000$
	$c_3 = 0.555555556$	$x_3 = 0.774596669$
4	$c_1 = 0.347854845$	$x_1 = -0.861136312$
	$c_2 = 0.652145155$	$x_2 = -0.339981044$
	$c_3 = 0.652145155$	$x_3 = 0.339981044$
	$c_4 = 0.347854845$	$x_4 = 0.861136312$
5	$c_1 = 0.236926885$	$x_1 = -0.906179846$
	$c_2 = 0.478628670$	$x_2 = -0.538469310$
	$c_3 = 0.568888889$	$x_3 = 0.000000000$
	$c_4 = 0.478628670$	$x_4 = 0.538469310$
	$c_5 = 0.236926885$	$x_5 = 0.906179846$
6	$c_1 = 0.171324492$	$x_1 = -0.932469514$
	$c_2 = 0.360761573$	$x_2 = -0.661209386$
	$c_3 = 0.467913935$	$x_3 = -0.238619186$
	$c_4 = 0.467913935$	$x_4 = 0.238619186$

	$c_5 = 0.360761573$	$x_5 = 0.661209386$
	$c_6 = 0.171324492$	$x_6 = 0.932469514$

So if the table is given for  $\int_{-1}^1 g(x)dx$  integrals, how does one solve  $\int_a^b f(x)dx$  ?

The answer lies in that any integral with limits of  $[a, b]$  can be converted into an integral with limits  $[-1, 1]$ . Let

$$x = mt + c \quad (20)$$

If  $x = a$ , then  $t = -1$

If  $x = b$ , then  $t = +1$

such that

$$\begin{aligned} a &= m(-1) + c \\ b &= m(1) + c \end{aligned} \quad (21)$$

Solving the two Equations (21) simultaneously gives

$$\begin{aligned} m &= \frac{b-a}{2} \\ c &= \frac{b+a}{2} \end{aligned} \quad (22)$$

Hence

$$\begin{aligned} x &= \frac{b-a}{2}t + \frac{b+a}{2} \\ dx &= \frac{b-a}{2}dt \end{aligned}$$

Substituting our values of  $x$  and  $dx$  into the integral gives us

$$\int_a^b f(x)dx = \int_{-1}^1 f\left(\frac{b-a}{2}t + \frac{b+a}{2}\right) \frac{b-a}{2} dt \quad (23)$$

### Example 1

For an integral  $\int_{-1}^1 f(x)dx$ , show that the two-point Gauss quadrature rule approximates to

$$\int_{-1}^1 f(x)dx \approx c_1 f(x_1) + c_2 f(x_2)$$

where

$$c_1 = 1$$

$$c_2 = 1$$

$$x_1 = -\frac{1}{\sqrt{3}}$$

$$x_2 = \frac{1}{\sqrt{3}}$$

**Solution**

Assuming the formula

$$\int_{-1}^1 f(x)dx = c_1 f(x_1) + c_2 f(x_2) \quad (\text{E1.1})$$

gives exact values for integrals  $\int_{-1}^1 1dx$ ,  $\int_{-1}^1 xdx$ ,  $\int_{-1}^1 x^2dx$ , and  $\int_{-1}^1 x^3dx$ . Then

$$\int_{-1}^1 1dx = 2 = c_1 + c_2 \quad (\text{E1.2})$$

$$\int_{-1}^1 xdx = 0 = c_1 x_1 + c_2 x_2 \quad (\text{E1.3})$$

$$\int_{-1}^1 x^2dx = \frac{2}{3} = c_1 x_1^2 + c_2 x_2^2 \quad (\text{E1.4})$$

$$\int_{-1}^1 x^3dx = 0 = c_1 x_1^3 + c_2 x_2^3 \quad (\text{E1.5})$$

Multiplying Equation (E1.3) by  $x_1^2$  and subtracting from Equation (E1.5) gives

$$c_2 x_2 (x_1^2 - x_2^2) = 0 \quad (\text{E1.6})$$

The solution to the above equation is

$$c_2 = 0, \text{ or/and}$$

$$x_2 = 0, \text{ or/and}$$

$$x_1 = x_2, \text{ or/and}$$

$$x_1 = -x_2.$$

- I.  $c_2 = 0$  is not acceptable as Equations (E1.2-E1.5) reduce to  $c_1 = 2$ ,  $c_1 x_1 = 0$ ,  $c_1 x_1^2 = \frac{2}{3}$ , and  $c_1 x_1^3 = 0$ . But since  $c_1 = 2$ , then  $x_1 = 0$  from  $c_1 x_1 = 0$ , but  $x_1 = 0$  conflicts with  $c_1 x_1^2 = \frac{2}{3}$ .
- II.  $x_2 = 0$  is not acceptable as Equations (E1.2-E1.5) reduce to  $c_1 + c_2 = 2$ ,  $c_1 x_1 = 0$ ,  $c_1 x_1^2 = \frac{2}{3}$ , and  $c_1 x_1^3 = 0$ . Since  $c_1 x_1 = 0$ , then  $c_1$  or  $x_1$  has to be zero but this violates  $c_1 x_1^2 = \frac{2}{3} \neq 0$ .
- III.  $x_1 = x_2$  is not acceptable as Equations (E1.2-E1.5) reduce to  $c_1 + c_2 = 2$ ,  $c_1 x_1 + c_2 x_1 = 0$ ,  $c_1 x_1^2 + c_2 x_1^2 = \frac{2}{3}$ , and  $c_1 x_1^3 + c_2 x_1^3 = 0$ . If  $x_1 \neq 0$ , then  $c_1 x_1 + c_2 x_1 = 0$

gives  $c_1 + c_2 = 0$  and that violates  $c_1 + c_2 = 2$ . If  $x_1 = 0$ , then that violates  $c_1 x_1^2 + c_2 x_1^2 = \frac{2}{3} \neq 0$ .

That leaves the solution of  $x_1 = -x_2$  as the only possible acceptable solution and in fact, it does not have violations (see it for yourself)

$$x_1 = -x_2 \quad (\text{E1.7})$$

Substituting (E1.7) in Equation (E1.3) gives

$$c_1 = c_2 \quad (\text{E1.8})$$

From Equations (E1.2) and (E1.8),

$$c_1 = c_2 = 1 \quad (\text{E1.9})$$

Equations (E1.4) and (E1.9) gives

$$x_1^2 + x_2^2 = \frac{2}{3} \quad (\text{E1.10})$$

Since Equation (E1.7) requires that the two results be of opposite sign, we get

$$x_1 = -\frac{1}{\sqrt{3}}$$

$$x_2 = \frac{1}{\sqrt{3}}$$

Hence

$$\begin{aligned} \int_{-1}^1 f(x) dx &= c_1 f(x_1) + c_2 f(x_2) \\ &= f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right) \end{aligned} \quad (\text{E1.11})$$

## Example 2

For an integral  $\int_a^b f(x) dx$ , derive the one-point Gauss quadrature rule.

### Solution

The one-point Gauss quadrature rule is

$$\int_a^b f(x) dx \approx c_1 f(x_1) \quad (\text{E2.1})$$

Assuming the formula gives exact values for integrals  $\int_{-1}^1 1 dx$ , and  $\int_{-1}^1 x dx$

$$\begin{aligned} \int_a^b 1 dx &= b - a = c_1 \\ \int_a^b x dx &= \frac{b^2 - a^2}{2} = c_1 x_1 \end{aligned} \quad (\text{E2.2})$$

Since  $c_1 = b - a$ , the other equation becomes

$$(b-a)x_1 = \frac{b^2 - a^2}{2}$$

$$x_1 = \frac{b+a}{2} \quad (E2.3)$$

Therefore, one-point Gauss quadrature rule can be expressed as

$$\int_a^b f(x)dx \approx (b-a)f\left(\frac{b+a}{2}\right) \quad (E2.4)$$

### Example 3

What would be the formula for

$$\int_a^b f(x)dx = c_1 f(a) + c_2 f(b)$$

if you want the above formula to give you exact values of  $\int_a^b (a_0x + b_0x^2)dx$ , that is, a linear combination of  $x$  and  $x^2$ .

### Solution

If the formula is exact for a linear combination of  $x$  and  $x^2$ , then

$$\int_a^b xdx = \frac{b^2 - a^2}{2} = c_1 a + c_2 b$$

$$\int_a^b x^2 dx = \frac{b^3 - a^3}{3} = c_1 a^2 + c_2 b^2 \quad (E3.1)$$

Solving the two Equations (E3.1) simultaneously gives

$$\begin{bmatrix} a & b \\ a^2 & b^2 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} \frac{b^2 - a^2}{2} \\ \frac{b^3 - a^3}{3} \end{bmatrix}$$

$$c_1 = -\frac{1}{6} \frac{-ab - b^2 + 2a^2}{a}$$

$$c_2 = -\frac{1}{6} \frac{a^2 + ab - 2b^2}{b} \quad (E3.2)$$

So

$$\int_a^b f(x)dx = -\frac{1}{6} \frac{-ab - b^2 + 2a^2}{a} f(a) - \frac{1}{6} \frac{a^2 + ab - 2b^2}{b} f(b) \quad (E3.3)$$

Let us see if the formula works.

Evaluate  $\int_2^5 (2x^2 - 3x)dx$  using Equation(E3.3)

$$\begin{aligned} \int_2^5 (2x^2 - 3x) dx &\approx c_1 f(a) + c_2 f(b) \\ &= -\frac{1}{6} \frac{-(2)(5) - 5^2 + 2(2)^2}{2} [2(2)^2 - 3(2)] - \frac{1}{6} \frac{2^2 + 2(5) - 2(5)^2}{5} [2(5)^2 - 3(5)] \\ &= 46.5 \end{aligned}$$

The exact value of  $\int_2^5 (2x^2 - 3x) dx$  is given by

$$\begin{aligned} \int_2^5 (2x^2 - 3x) dx &= \left[ \frac{2x^3}{3} - \frac{3x^2}{2} \right]_2^5 \\ &= 46.5 \end{aligned}$$

Any surprises?

Now evaluate  $\int_2^5 3dx$  using Equation (E3.3)

$$\begin{aligned} \int_2^5 3dx &\approx c_1 f(a) + c_2 f(b) \\ &= -\frac{1}{6} \frac{-2(5) - 5^2 + 2(2)^2}{2} (3) - \frac{1}{6} \frac{2^2 + 2(5) - 2(5)^2}{5} (3) \\ &= 10.35 \end{aligned}$$

The exact value of  $\int_2^5 3dx$  is given by

$$\begin{aligned} \int_2^5 3dx &= [3x]_2^5 \\ &= 9 \end{aligned}$$

Because the formula will only give exact values for linear combinations of  $x$  and  $x^2$ , it does not work exactly even for a simple integral of  $\int_2^5 3dx$ .

Do you see now why we choose  $a_0 + a_1x$  as the integrand for which the formula

$$\int_a^b f(x) dx \approx c_1 f(a) + c_2 f(b)$$

gives us exact values?

#### Example 4

Use two-point Gauss quadrature rule to approximate the distance covered by a rocket from  $t = 8$  to  $t = 30$  as given by

$$x = \int_8^{30} \left( 2000 \ln \left[ \frac{140000}{140000 - 2100t} \right] - 9.8t \right) dt$$

Also, find the absolute relative true error.

**Solution**

First, change the limits of integration from  $[8, 30]$  to  $[-1, 1]$  using Equation(23) gives

$$\begin{aligned}\int_8^{30} f(t)dt &= \frac{30-8}{2} \int_{-1}^1 f\left(\frac{30-8}{2}x + \frac{30+8}{2}\right)dx \\ &= 11 \int_{-1}^1 f(11x + 19)dx\end{aligned}$$

Next, get weighting factors and function argument values from Table 1 for the two point rule,

$$c_1 = 1.000000000.$$

$$x_1 = -0.577350269$$

$$c_2 = 1.000000000$$

$$x_2 = 0.577350269$$

Now we can use the Gauss quadrature formula

$$\begin{aligned}11 \int_{-1}^1 f(11x + 19)dx &\approx 11[c_1 f(x_1) + c_2 f(x_2)] \\ &= 11[f(11(-0.5773503) + 19) + f(11(0.5773503) + 19)] \\ &= 11[f(12.64915) + f(25.35085)] \\ &= 11[(296.8317) + (708.4811)] \\ &= 11058.44 \text{ m}\end{aligned}$$

since

$$\begin{aligned}f(12.64915) &= 2000 \ln\left[\frac{140000}{140000 - 2100(12.64915)}\right] - 9.8(12.64915) \\ &= 296.8317\end{aligned}$$

$$\begin{aligned}f(25.35085) &= 2000 \ln\left[\frac{140000}{140000 - 2100(25.35085)}\right] - 9.8(25.35085) \\ &= 708.4811\end{aligned}$$

The absolute relative true error,  $|\epsilon_t|$ , is (True value = 11061.34 m)

$$\begin{aligned}|\epsilon_t| &= \left| \frac{11061.34 - 11058.44}{11061.34} \right| \times 100 \\ &= 0.0262\%\end{aligned}$$

**Example 5**

Use three-point Gauss quadrature rule to approximate the distance covered by a rocket from  $t = 8$  to  $t = 30$  as given by

$$x = \int_8^{30} \left( 2000 \ln\left[\frac{140000}{140000 - 2100t}\right] - 9.8t \right) dt$$

Also, find the absolute relative true error.

**Solution**

First, change the limits of integration from  $[8, 30]$  to  $[-1, 1]$  using Equation (23) gives

$$\begin{aligned}\int_8^{30} f(t)dt &= \frac{30-8}{2} \int_{-1}^1 f\left(\frac{30-8}{2}x + \frac{30+8}{2}\right)dx \\ &= 11 \int_{-1}^1 f(11x + 19)dx\end{aligned}$$

The weighting factors and function argument values are

$$c_1 = 0.555555556$$

$$x_1 = -0.774596669$$

$$c_2 = 0.888888889$$

$$x_2 = 0.000000000$$

$$c_3 = 0.555555556$$

$$x_3 = 0.774596669$$

and the formula is

$$\begin{aligned}11 \int_{-1}^1 f(11x + 19)dx &\approx 11[c_1 f(11x_1 + 19) + c_2 f(11x_2 + 19) + c_3 f(11x_3 + 19)] \\ &= 11 \left[ 0.5555556 f(11(-0.7745967) + 19) + 0.8888889 f(11(0.0000000) + 19) \right. \\ &\quad \left. + 0.5555556 f(11(0.7745967) + 19) \right] \\ &= 11[0.55556 f(10.47944) + 0.88889 f(19.00000) + 0.55556 f(27.52056)] \\ &= 11[0.55556 \times 239.3327 + 0.88889 \times 484.7455 + 0.55556 \times 795.1069] \\ &= 11061.31 \text{ m}\end{aligned}$$

since

$$\begin{aligned}f(10.47944) &= 2000 \ln \left[ \frac{140000}{140000 - 2100(10.47944)} \right] - 9.8(10.47944) \\ &= 239.3327\end{aligned}$$

$$\begin{aligned}f(19.00000) &= 2000 \ln \left[ \frac{140000}{140000 - 2100(19.00000)} \right] - 9.8(19.00000) \\ &= 484.7455\end{aligned}$$

$$\begin{aligned}f(27.52056) &= 2000 \ln \left[ \frac{140000}{140000 - 2100(27.52056)} \right] - 9.8(27.52056) \\ &= 795.1069\end{aligned}$$

The absolute relative true error,  $|\epsilon_t|$ , is (True value = 11061.34 m)

$$\begin{aligned}|\epsilon_t| &= \left| \frac{11061.34 - 11061.31}{11061.34} \right| \times 100 \\ &= 0.0003\%\end{aligned}$$

---

**INTEGRATION**

---

Topic      Gauss quadrature rule  
Summary    These are textbook notes of Gauss quadrature rule  
Major      General Engineering  
Authors     Autar Kaw, Michael Keteltas  
Date        August 11, 2010  
Web Site    <http://numericalmethods.eng.usf.edu>

---

## Chapter 07.06

# Integrating Discrete Functions

*After reading this chapter, you should be able to:*

1. integrate discrete functions by several methods,
2. derive the formula for trapezoidal rule with unequal segments, and
3. solve examples of finding integrals of discrete functions.

### What is integration?

Integration is the process of measuring the area under a function plotted on a graph. Why would we want to integrate a function? Among the most common examples are finding the velocity of a body from an acceleration function, and displacement of a body from a velocity function. Throughout many engineering fields, there are (what sometimes seems like) countless applications for integral calculus. You can read about a few of these applications in different engineering majors in Chapters 07.00A-07.00G.

Sometimes, the function to be integrated is given at discrete data points, and the area under the curve is needed to be approximated. Here, we will discuss the integration of such discrete functions,

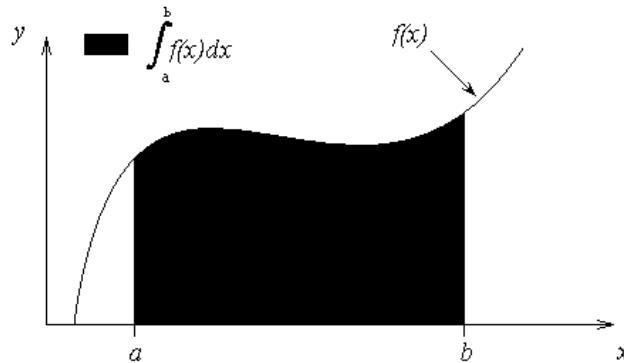
$$I = \int_a^b f(x)dx$$

where

$f(x)$  is called the integrand and is given at discrete value of  $x$ ,

$a$  = lower limit of integration

$b$  = upper limit of integration



**Figure 1** Integration of a function

### Integrating discrete functions

Multiple methods of integrating discrete functions are shown below using an example.

#### Example 1

The upward velocity of a rocket is given as a function of time in Table 1.

**Table 1** Velocity as a function of time.

$t$ (s)	$v(t)$ (m/s)
0	0
10	227.04
15	362.78
20	517.35
22.5	602.97
30	901.67

Determine the distance,  $s$ , covered by the rocket from  $t = 11$  to  $t = 16$  using the velocity data provided and use any applicable numerical technique.

#### Solution

##### Method 1: Average Velocity Method

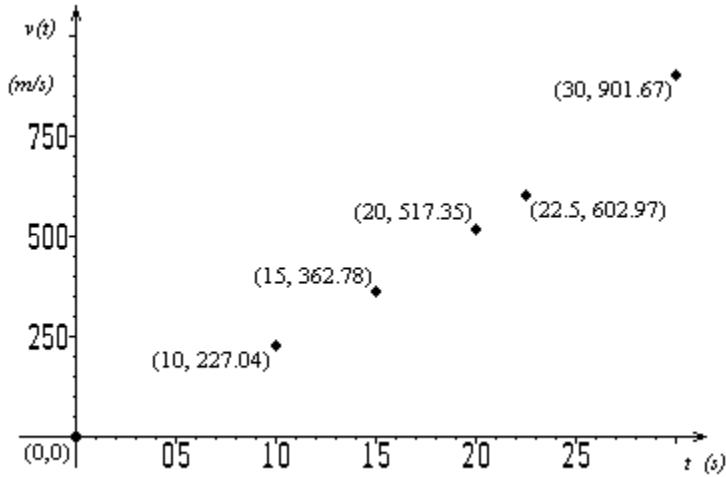
The velocity of the rocket is not provided at  $t = 11$  and  $t = 16$ , so we will have to use an interval that includes  $[11, 16]$  to find the average velocity of the rocket within that range. In this case, the interval  $[10, 20]$  will suffice.

$$v(10) = 227.04$$

$$v(15) = 362.78$$

$$v(20) = 517.35$$

$$\begin{aligned} \text{Average Velocity} &= \frac{v(10) + v(15) + v(20)}{3} \\ &= \frac{227.04 + 362.78 + 517.35}{3} \\ &= 369.06 \text{ m/s} \end{aligned}$$



**Figure 1** Velocity vs. time data for the rocket example

Using

$$s = \bar{v}\Delta t,$$

we get

$$s = (369.06)(16 - 11) = 1845.3 \text{ m}$$

### Method 2: Trapezoidal Rule

If we were finding the distance traveled between times in the data table, we would simply find the area of the trapezoids with the corner points given as the velocity and time data points. For example

$$\int_{10}^{20} v(t) dt = \int_{10}^{15} v(t) dt + \int_{15}^{20} v(t) dt$$

and applying the trapezoidal rule over each of the above integrals gives

$$\int_{10}^{20} v(t) dt \approx \frac{15 - 10}{2} [v(10) + v(15)] + \frac{20 - 15}{2} [v(15) + v(20)]$$

The values of  $v(10)$ ,  $v(15)$  and  $v(20)$  are given in Table 1.

However, we are interested in finding

$$\int_{11}^{16} v(t) dt = \int_{11}^{15} v(t) dt + \int_{15}^{16} v(t) dt$$

and applying the trapezoidal rule over each of the above integrals gives

$$\begin{aligned}\int_{11}^{16} v(t) dt &\approx \frac{15-11}{2}[v(11) + v(15)] + \frac{16-15}{2}[v(15) + v(16)] \\ &= \frac{15-11}{2}(v(11) + 362.78) + \frac{16-15}{2}(362.78 + v(16))\end{aligned}$$

How do we find  $v(11)$  and  $v(16)$ ? We use linear interpolation. To find  $v(11)$ ,

$$\begin{aligned}v(t) &= 227.04 + 27.148(t-10), \quad 10 \leq t \leq 15 \\ v(11) &= 227.04 + 27.148(11-10) \\ &= 254.19 \text{ m/s}\end{aligned}$$

and to find  $v(16)$

$$\begin{aligned}v(t) &= 362.78 + 30.913(t-15), \quad 15 \leq t \leq 20 \\ v(16) &= 362.78 + 30.913(16-15) \\ &= 393.69 \text{ m/s}\end{aligned}$$

Then

$$\begin{aligned}\int_{11}^{16} v(t) dt &\approx \frac{15-11}{2}(v(11) + 362.78) + \frac{16-15}{2}(362.78 + v(16)) \\ &= \frac{15-11}{2}(254.19 + 362.78) + \frac{16-15}{2}(362.78 + 393.69) \\ &= 1612.2 \text{ m}\end{aligned}$$

#### Method 3: Polynomial interpolation to find the velocity profile

Because we are finding the area under the curve from  $[10, 20]$ , we must use three points,  $t = 10$ ,  $t = 15$ , and  $t = 20$ , to fit a quadratic polynomial through the data. Using polynomial interpolation, our resulting velocity function is (refer to notes on direct method of interpolation)

$$v(t) = 12.05 + 17.733t + 0.3766t^2, \quad 10 \leq t \leq 20.$$

Now, we simply take the integral of the quadratic within our limits, giving us

$$\begin{aligned}s &\approx \int_{11}^{16} (12.05 + 17.733t + 0.3766t^2) dt \\ &= \left[ 12.05t + \frac{17.733t^2}{2} + \frac{0.3766t^3}{3} \right]_{11}^{16} \\ &= 12.05(16-11) + \frac{17.733}{2}(16^2 - 11^2) + \frac{0.3766}{3}(16^3 - 11^3) \\ &= 1604.3 \text{ m}\end{aligned}$$

#### Method 4: Spline interpolation to find the velocity profile

Fitting quadratic splines (refer to notes on spline method of interpolation) through the data results in the following set of quadratics

$$\begin{aligned}v(t) &= 22.704t, & 0 \leq t \leq 10 \\ &= 0.8888t^2 + 4.928t + 88.88, & 10 \leq t \leq 15\end{aligned}$$

$$\begin{aligned}
 &= -0.1356t^2 + 35.66t - 141.61, & 15 \leq t \leq 20 \\
 &= 1.6048t^2 - 33.956t + 554.55, & 20 \leq t \leq 22.5 \\
 &= 0.20889t^2 + 28.86t - 152.13, & 22.5 \leq t \leq 30
 \end{aligned}$$

The value of the integral would then simply be

$$\begin{aligned}
 s &= \int_{11}^{15} v(t) dt + \int_{15}^{16} v(t) dt \\
 &\approx \int_{11}^{15} (0.8888t^2 + 4.928t + 88.88) dt + \int_{15}^{16} (-0.1356t^2 + 35.66t - 141.61) dt \\
 &= \left[ \frac{0.8888t^3}{3} + \frac{4.928t^2}{2} + 88.88t \right]_{11}^{15} + \left[ \frac{-0.1356t^3}{3} + \frac{35.66t^2}{2} - 141.61t \right]_{15}^{16} \\
 &= \frac{0.8888}{3} (15^3 - 11^3) + \frac{4.928}{2} (15^2 - 11^2) + 88.88(15 - 11) \\
 &\quad + \frac{-0.1356}{3} (16^3 - 15^3) + \frac{35.66}{2} (16^2 - 15^2) - 141.61(16 - 15) \\
 &= 1595.9 \text{ m}
 \end{aligned}$$

### Example 2

What is the absolute relative true error for each of the four methods used in Example 1 if the data in Table 1 was actually obtained from the velocity profile of

$$v(t) = \left( 2000 \ln \left[ \frac{140000}{140000 - 2100t} \right] - 9.8t \right),$$

where  $v$  is given in m/s and  $t$  in s.

### Solution

The distance covered between  $t = 11$  and  $t = 16$  is

$$\begin{aligned}
 s &= \int_{11}^{16} \left( 2000 \ln \left[ \frac{140000}{140000 - 2100t} \right] - 9.8t \right) dt \\
 &= 1604.9 \text{ m}
 \end{aligned}$$

#### Method 1

The approximate value obtained using average velocity method was 1845.3 m. Hence, the absolute relative true error,  $|\epsilon_t|$ , is

$$\begin{aligned}
 |\epsilon_t| &= \left| \frac{1604.9 - 1845.3}{1604.9} \right| \times 100\% \\
 &= 14.976\%
 \end{aligned}$$

#### Method 2:

The approximate value obtained using the trapezoidal rule was 1612.2 m. Hence, the absolute relative true error,  $|\epsilon_t|$ , is

$$|\epsilon_t| = \left| \frac{1604.9 - 1612.2}{1604.9} \right| \times 100\% \\ = 0.451\%$$

**Method 3:**

The approximate value obtained using the direct polynomial was 1604.3 m. Hence, the absolute relative true error,  $|\epsilon_t|$ , is

$$|\epsilon_t| = \left| \frac{1604.9 - 1604.3}{1604.9} \right| \times 100\% \\ = 0.037\%$$

**Method 4:**

The approximate value obtained using the spline interpolation was 1595.9 m, hence, the absolute relative true error,  $|\epsilon_t|$ , is

$$|\epsilon_t| = \left| \frac{1604.9 - 1595.9}{1604.9} \right| \times 100\% \\ = 0.564\%$$

**Table 2** Comparison of discrete function methods of numerical integration

Method	Approximate Value	$ \epsilon_t $
Average Velocity	1845.3	14.976%
Trapezoidal Rule	1612.2	0.451%
Polynomial Interpolation	1604.3	0.037%
Spline Interpolation	1595.9	0.564%

**Trapezoidal Rule for Discrete Functions with Unequal Segments**

For a general case of a function given at  $n$  data points  $(x_1, f(x_1)), (x_2, f(x_2)), (x_3, f(x_3)), \dots, (x_n, f(x_n))$ , where,  $x_1, x_2, \dots, x_n$  are in an ascending order, the approximate value of the integral  $\int_{x_1}^{x_n} f(x) dx$  is given by

$$\int_{x_1}^{x_n} f(x) dx = \int_{x_1}^{x_2} f(x) dx + \int_{x_2}^{x_3} f(x) dx + \dots + \int_{x_{n-1}}^{x_n} f(x) dx \\ \approx (x_2 - x_1) \frac{f(x_1) + f(x_2)}{2} + (x_3 - x_2) \frac{f(x_2) + f(x_3)}{2} + \dots \\ \dots + (x_n - x_{n-1}) \frac{f(x_{n-1}) + f(x_n)}{2}$$

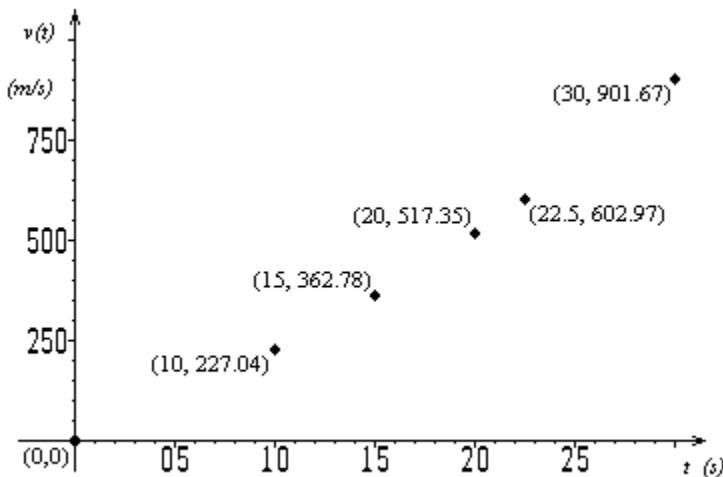
This approach uses the trapezoidal rule in the intervals  $[x_1, x_2], [x_2, x_3], \dots, [x_{n-1}, x_n]$  and then adds the obtained values.

**Example 3**

The upward velocity of a rocket is given as a function of time in Table 3.

**Table 3.** Velocity as a function of time.

t	v(t)
s	m/s
0	0
10	227.04
15	362.78
20	517.35
22.5	602.97
30	901.67

**Figure 2** Velocity vs. time data for the rocket example

Determine the distance,  $s$ , covered by the rocket from  $t = 0$  to  $t = 30$  using the velocity data provided and the trapezoidal rule for discrete data with unequal segments.

**Solution**

$$\begin{aligned}
 \int_0^{30} v(t) dt &= \int_0^{10} v(t) dt + \int_{10}^{15} v(t) dt + \int_{15}^{20} v(t) dt + \int_{20}^{22.5} v(t) dt + \int_{22.5}^{30} v(t) dt \\
 &= (10 - 0) \frac{v(0) + v(10)}{2} + (15 - 10) \frac{v(10) + v(15)}{2} \\
 &\quad + (20 - 15) \frac{v(15) + v(20)}{2} + (22.5 - 20) \frac{v(20) + v(22.5)}{2}
 \end{aligned}$$

$$\begin{aligned}
 & + (30 - 22.5) \frac{v(22.5) + v(30)}{2} \\
 & = (10) \frac{0 + 227.04}{2} + (5) \frac{227.04 + 362.78}{2} \\
 & \quad + (5) \frac{362.78 + 517.35}{2} + (2.5) \frac{517.35 + 602.97}{2} \\
 & \quad + (7.5) \frac{602.97 + 901.67}{2} \\
 & = 1135.2 + 1474.55 + 2200.325 + 1399.9 + 5642.4 \\
 & = 11852 \text{ m}
 \end{aligned}$$

Can you find the value of  $\int_{10}^{20} v(t)dt$ ?

### INTEGRATION

Topic	Integrating discrete functions
Summary	Textbook notes on integrating discrete functions
Major	All Majors of Engineering
Authors	Autar Kaw
Last Revised	December 23, 2009
Web Site	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

# Chapter 07.07

## Integrating Improper Functions

After reading this chapter, you should be able to:

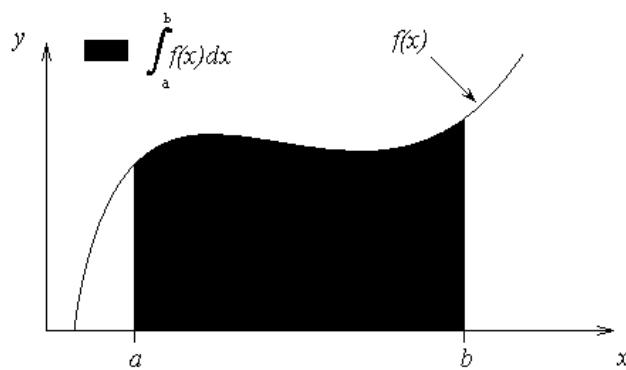
1. integrate improper functions using methods such as the trapezoidal rule and Gaussian Quadrature schemes.

### What is integration?

Integration is the process of measuring the area under a function plotted on a graph. Why would we want to integrate a function? Among the most common examples are finding the velocity of a body from an acceleration function, and displacement of a body from a velocity function. Throughout many engineering fields, there are (what sometimes seems like) countless applications for integral calculus. You can read about some of these applications in Chapters 07.00A-07.00G.

Sometimes, the evaluation of expressions involving these integrals can become daunting, if not indeterminate. For this reason, a wide variety of numerical methods has been developed to simplify the integral.

Here, we will discuss the incorporation of these numerical methods into improper integrals.



**Figure 1** Integration of a function

### What is an improper integral?

An integral is improper if

- a) the integrand becomes infinite in the interval of integration (including end points)  
or/and
- b) the interval of integration has an infinite bound.

### Example 1

Give some examples of improper integrals

#### Solution

The integral

$$I = \int_0^2 \frac{x}{\sqrt{4-x^2}} dx$$

is improper because the integrand becomes infinite at  $x = 2$ .

The integral

$$I = \int_0^2 \frac{x}{\sqrt{1-x}} dx$$

is improper because the integrand becomes infinite at  $x = 1$ .

The integral

$$I = \int_0^\infty e^{-t} t dt$$

is improper because the interval of integration has an infinite bound.

The integral

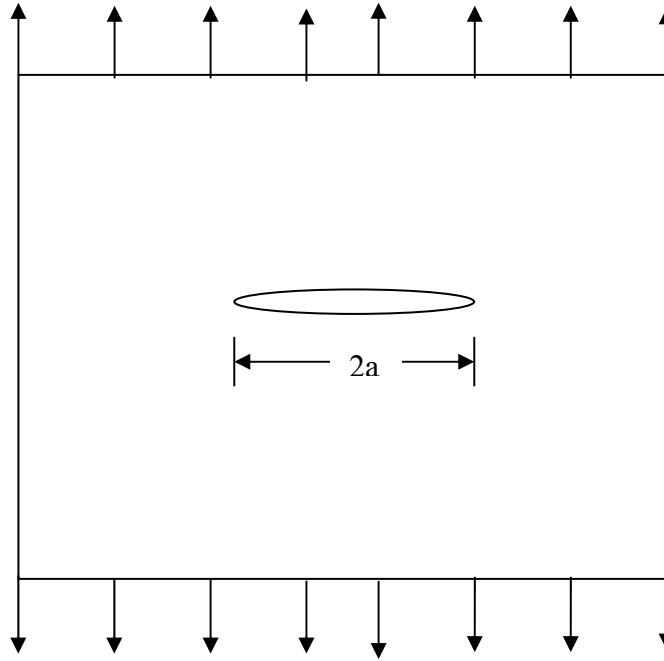
$$I = \int_0^\infty \frac{e^{-t}}{\sqrt{1-t}} dt$$

is improper because the interval of integration has an infinite bound and the integrand is infinite at  $t = 1$ .

If the integrand is undefined at a finite number of points, the value of the area under the curve does not change. Hence such integrals could theoretically be solved either by assuming any value of the integrand at such points. Also, methods such as Gauss quadrature rule do not use the value of the integrand at end points, and hence integrands that are undefined at end points can be integrated using such methods.

For the case where there is an infinite interval of integration, one may make a change of variables that transforms the infinite range of integration to a finite one.

Let us illustrate these two cases with examples.



**Figure 2** A plate with a crack under a uniform axial load

### Example 2

In analyzing fracture of metals, one wants to know the opening displacement of cracks. In a large plate, if there is a crack length of  $2a$  meters, then the maximum crack opening displacement (MCOD) is given by

$$\text{MCOD} = \frac{2\sigma}{E} \int_0^a \frac{x}{\sqrt{a^2 - x^2}} dx$$

where

$\sigma$  = remote normal applied stress

$E$  = Young's modulus

Assume

$$a = 0.02 \text{ m}$$

$$E = 210 \text{ GPa and}$$

$$\sigma = 70 \text{ MPa} .$$

Find the exact value of the maximum crack opening displacement.

### Solution

The maximum crack opening displacement (MCOD) is given by

$$\text{MCOD} = \frac{2\sigma}{E} \int_0^a \frac{x}{\sqrt{a^2 - x^2}} dx$$

Substituting  $a = 0.02 \text{ m}$ ,  $E = 210 \text{ GPa}$  and  $\sigma = 70 \text{ MPa}$  gives

$$\begin{aligned} \text{MCOD} &= \frac{2(70 \times 10^6)}{210 \times 10^9} \int_0^{0.02} \frac{x}{\sqrt{(0.02)^2 - x^2}} dx \\ &= \frac{1}{1500} \int_0^{0.02} \frac{x}{\sqrt{0.0004 - x^2}} dx \end{aligned}$$

The exact value of the integral then is

$$\begin{aligned} \text{MCOD} &= \frac{1}{1500} \left[ -\sqrt{0.0004 - x^2} \right]_0^{0.02} \\ &= \frac{1}{1500} (-0 + 0.02) \\ &= 1.3333 \times 10^{-5} \text{ m} \end{aligned}$$

### Example 3

Any of the Newton-Cotes formulas, such as Trapezoidal rule and Simpson's 1/3 rule, cannot be used directly for integrals where the integrands become infinite at the ends of the intervals. Since Gauss quadrature rule does not require calculation of the integrand at the end points, it could be used directly to calculate such integrals. Knowing this, find the value of the integral

$$\frac{1}{1500} \int_0^{0.02} \frac{x}{\sqrt{0.0004 - x^2}} dx$$

from Example 2 by using two-point Gauss quadrature rule.

### Solution

We will change the limits of integration from [0,0.02] to [-1,1], such that we may use the tabulated values of  $c_1$ ,  $c_2$ ,  $x_1$ , and  $x_2$ . Assigning

$$f(x) = \frac{x}{\sqrt{0.0004 - x^2}},$$

we get

$$\begin{aligned} \frac{1}{1500} \int_0^{0.02} f(x) dx &= \frac{1}{1500} \frac{0.02 - 0}{2} \int_{-1}^1 f\left(\frac{0.02 - 0}{2}x + \frac{0.02 + 0}{2}\right) dx \\ &= \frac{1}{150000} \int_{-1}^1 f(0.01x + 0.01) dx \end{aligned}$$

The function arguments and weighting factors for two-point Gauss quadrature rule are

$$c_1 = 1.000000000$$

$$x_1 = -0.577350269$$

$$c_2 = 1.000000000$$

$$x_2 = 0.577350269$$

Giving us a formula of

$$\begin{aligned}
\frac{1}{150000} \int_{-1}^1 f(0.01x + 0.01) dx &\approx \frac{1}{150000} c_1 f(0.01x_1 + 0.01) + \frac{1}{150000} c_2 f(0.01x_2 + 0.01) \\
&= \frac{1}{150000} f(0.01(-0.57735) + 0.01) + \frac{1}{150000} f(0.01(0.57735) + 0.01) \\
&= \frac{1}{150000} f(0.0042265) + \frac{1}{150000} f(0.0157735) \\
&= \frac{1}{150000} (0.21621) + \frac{1}{150000} (1.28279) \\
&= 9.9934 \times 10^{-6} \text{ m}
\end{aligned}$$

since

$$\begin{aligned}
f(0.0042265) &= \frac{0.0042265}{\sqrt{0.0004 - (0.0042265)^2}} = 0.21621 \\
f(0.0157735) &= \frac{0.0157735}{\sqrt{0.0004 - (0.0157735)^2}} = 1.28279
\end{aligned}$$

The absolute relative true error,  $|\epsilon_t|$ , is

$$\begin{aligned}
|\epsilon_t| &= \left| \frac{1.3333 \times 10^{-5} - 9.9934 \times 10^{-6}}{1.3333 \times 10^{-5}} \right| \times 100\% \\
&= 25.048\%
\end{aligned}$$

#### Example 4

The value of the integral

$$\frac{1}{1500} \int_0^{0.02} \frac{x}{\sqrt{0.0004 - x^2}} dx$$

in Example 3 by using two-point Gauss quadrature rule has a large absolute relative true error of more than 25%. Use the double-segment two-point Gauss quadrature rule to find the value of the integral. Take the interval  $[0, 0.02]$  and split it into two equal segments of  $[0, 0.01]$  and  $[0.01, 0.02]$ , and then apply the two-point Gauss quadrature rule over each segment.

#### Solution

Write the integral with interval of  $[0, 0.02]$  as sum of two integrals with intervals  $[0, 0.01]$  and  $[0.01, 0.02]$  gives

$$\begin{aligned}
\frac{1}{1500} \int_0^{0.02} f(x) dx &= \frac{1}{1500} \int_0^{0.01} f(x) dx + \frac{1}{1500} \int_{0.01}^{0.02} f(x) dx \\
&= \frac{1}{1500} \frac{0.01 - 0}{2} \int_{-1}^1 f\left(\frac{0.01 - 0}{2}x + \frac{0.01 + 0}{2}\right) dx \\
&\quad + \frac{1}{1500} \frac{0.02 - 0.01}{2} \int_{-1}^1 f\left(\frac{0.02 - 0.01}{2}x + \frac{0.02 + 0.01}{2}\right) dx
\end{aligned}$$

$$= \frac{1}{300000} \int_{-1}^1 f(0.005x + 0.005) dx + \frac{1}{300000} \int_{-1}^1 f(0.005x + 0.015) dx$$

Using the two-point Gauss quadrature rule, this becomes

$$\begin{aligned} \frac{1}{1500} \int_0^{0.02} f(x) dx &\approx \frac{1}{300000} c_1 f(0.005x_1 + 0.005) + \frac{1}{300000} c_2 f(0.005x_2 + 0.005) \\ &\quad + \frac{1}{300000} c_1 f(0.005x_1 + 0.015) + \frac{1}{300000} c_2 f(0.005x_2 + 0.015) \end{aligned}$$

Using the same arguments and weighting factors as before

$$\begin{aligned} \frac{1}{1500} \int_0^{0.02} f(x) dx &\approx \frac{1}{300000} f(0.005(-0.57735) + 0.005) + \frac{1}{300000} f(0.005(0.57735) + 0.005) \\ &\quad + \frac{1}{300000} f(0.005(-0.57735) + 0.015) + \frac{1}{300000} f(0.005(0.57735) + 0.015) \\ &= \frac{1}{300000} f(0.0021132) + \frac{1}{300000} f(0.0078868) \\ &\quad + \frac{1}{300000} f(0.0121132) + \frac{1}{300000} f(0.0178868) \\ &= \frac{1}{300000} (0.10626 + 0.42911 + 0.76115 + 1.99900) \\ &= 1.0985 \times 10^{-5} \text{ m} \end{aligned}$$

since

$$f(0.0021133) = \frac{0.0021133}{\sqrt{0.0004 - (0.0021133)^2}} = 0.10626$$

$$f(0.0078868) = \frac{0.0078868}{\sqrt{0.0004 - (0.0078868)^2}} = 0.42911$$

$$f(0.0121133) = \frac{0.0121133}{\sqrt{0.0004 - (0.0121133)^2}} = 0.76115$$

$$f(0.0178868) = \frac{0.0178868}{\sqrt{0.0004 - (0.0178868)^2}} = 1.99900$$

The absolute relative true error,  $|\epsilon_t|$ , is

$$\begin{aligned} |\epsilon_t| &= \left| \frac{1.3333 \times 10^{-5} - 1.0985 \times 10^{-5}}{1.3333 \times 10^{-5}} \right| \times 100\% \\ &= 17.610\% \end{aligned}$$

Repeating this process by splitting the interval into progressively more equal segments and applying the two-point Gaussian quadrature rule over each segment will obtain the data displayed in Table 1.

**Table 1** Gauss quadrature rule on an improper integral

$$\left( \frac{1}{1500} \int_0^{0.02} \frac{x}{\sqrt{0.0004 - x^2}} dx \right)$$

Number of Segments	Value	$ \epsilon_i  \%$
1	$9.9934 \times 10^{-6}$	25.05
2	$1.0985 \times 10^{-5}$	17.61
3	$1.1420 \times 10^{-5}$	14.35
4	$1.1679 \times 10^{-5}$	12.41
5	$1.1855 \times 10^{-5}$	11.09
6	$1.1984 \times 10^{-5}$	10.12
7	$1.2085 \times 10^{-5}$	9.365
8	$1.2166 \times 10^{-5}$	8.758

As evident from Table 1, the integral does not converge rapidly to the true value with an increase in number of quadrature points. Since the integrand becomes infinite at the end point  $x = 0.02$ , its value changes rapidly near  $x = 0.02$ . Since the multiple-segment two-point Gauss quadrature rule is non-adaptive, it will take a large number of segments to reach a converging value.

### Example 5

Euler's constant in mathematics is defined as

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt$$

Find  $\Gamma(2.4)$  using two and three-point Gauss quadrature rules. Also, find the absolute relative true error for each case.

#### Solution

$$\begin{aligned} \Gamma(2.4) &= \int_0^\infty e^{-t} t^{2.4-1} dt \\ &= \int_0^\infty e^{-t} t^{1.4} dt \end{aligned}$$

To solve the above improper integral, one may make a change of variables as

$$y = \frac{1}{1+t}$$

giving

$$\begin{aligned} t &= \frac{1}{y} - 1 \\ dt &= -\frac{1}{y^2} dy \end{aligned}$$

At  $t = 0, y = 1$ , at  $t = \infty, y = 0$ . So the integral can be re-written as

$$\Gamma(2.4) = \int_1^0 e^{-\left(\frac{1}{y}-1\right)} \left(\frac{1}{y}-1\right)^{1.4} \left(-\frac{1}{y^2}\right) dy$$

First, assigning

$$f(y) = e^{-\left(\frac{1}{y}-1\right)} \left(\frac{1}{y}-1\right)^{1.4} \left(-\frac{1}{y^2}\right)$$

and then changing the limits of integration, we get

$$\begin{aligned} \Gamma(2.4) &= \frac{0-1}{2} \int_{-1}^1 f\left(\frac{0-1}{2}y + \frac{0+1}{2}\right) dy \\ &= -0.5 \int_{-1}^1 f(-0.5y + 0.5) dy \end{aligned}$$

Now, one can use two-point Gauss Quadrature Rule to find the value of  $\Gamma(2.4)$  with weighting factors and function arguments of

$$c_1 = 1.000000000$$

$$y_1 = -0.577350269$$

$$c_2 = 1.000000000$$

$$y_2 = 0.577350269$$

$$\begin{aligned} \Gamma(2.4) &\approx -0.5c_1 f(-0.5y_1 + 0.5) - 0.5c_2 f(-0.5y_2 + 0.5) \\ &= -0.5f(-0.5(-0.57735) + 0.5) - 0.5f(-0.5(0.57735) + 0.5) \\ &= -0.5f(0.78868) - 0.5f(0.21133) \\ &= -0.5(-0.19458) - 0.5(-3.38857) \\ &= 1.7916 \end{aligned}$$

since

$$\begin{aligned} f(0.78868) &= e^{-\left(\frac{1}{0.78868}-1\right)} \left(\frac{1}{0.78868}-1\right)^{1.4} \left(-\frac{1}{(0.78868)^2}\right) \\ &= -0.19458 \\ f(0.21133) &= e^{-\left(\frac{1}{0.21133}-1\right)} \left(\frac{1}{0.21133}-1\right)^{1.4} \left(-\frac{1}{(0.21133)^2}\right) \\ &= -3.38857 \end{aligned}$$

The true value of the integral

$$\Gamma(2.4) = \int_0^\infty e^{-t} t^{1.4} dt = 1.2422$$

so the absolute relative true error,  $|\epsilon_t|$ , is

$$\begin{aligned} |\epsilon_t| &= \left| \frac{1.2422 - 1.7916}{1.2422} \right| \times 100\% \\ &= 44.230\% \end{aligned}$$

For three-point Gauss Quadrature Rule, the weighting factors and function arguments are

$$\begin{aligned}
 c_1 &= 0.555555556 \\
 y_1 &= -0.774596669 \\
 c_2 &= 0.888888889 \\
 y_2 &= 0.000000000 \\
 c_3 &= 0.555555556 \\
 y_3 &= 0.774596669
 \end{aligned}$$

The limits of integration and  $f(y)$  remain the same as for the two-point rule, so

$$\begin{aligned}
 \Gamma(2.4) &\approx -0.5c_1f(-0.5y_1 + 0.5) - 0.5c_2f(-0.5y_2 + 0.5) - 0.5c_3f(-0.5y_3 + 0.5) \\
 &= -0.5(0.55556)f(-0.5(-0.77460) + 0.5) \\
 &\quad - 0.5(0.88889)f(-0.5(0) + 0.5) - 0.5(0.55556)f(-0.5(0.77460) + 0.5) \\
 &= -0.27778f(0.88730) - 0.44444f(0.5) - 0.27778f(0.11270) \\
 &= -0.27778(-0.06224) - 0.44444(-1.47152) - 0.27778(-0.53890) \\
 &= 0.82100
 \end{aligned}$$

since

$$\begin{aligned}
 f(0.88730) &= e^{-\left(\frac{1}{0.88730}-1\right)} \left( \frac{1}{0.88730} - 1 \right)^{1.4} \left( -\frac{1}{(0.88730)^2} \right) \\
 &= -0.06224 \\
 f(0.5) &= e^{-\left(\frac{1}{0.5}-1\right)} \left( \frac{1}{0.5} - 1 \right)^{1.4} \left( -\frac{1}{(0.5)^2} \right) \\
 &= -1.47152 \\
 f(0.11270) &= e^{-\left(\frac{1}{0.11270}-1\right)} \left( \frac{1}{0.11270} - 1 \right)^{1.4} \left( -\frac{1}{(0.11270)^2} \right) \\
 &= -0.53894
 \end{aligned}$$

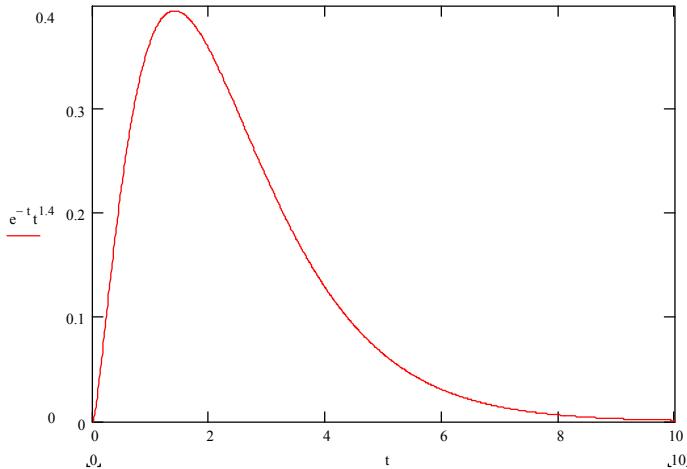
The absolute relative true error,  $|\epsilon_t|$ , is

$$\begin{aligned}
 |\epsilon_t| &= \left| \frac{1.2422 - 0.82099}{1.2422} \right| \times 100\% \\
 &= 33.906\%
 \end{aligned}$$

### Example 6

As you can see from the plot given in Figure 3 for the integrand in  $\int_0^\infty e^{-t} t^{1.4} dt$  of Example 5,

once the value of  $t$  exceeds 10, the area under the curve looks insignificant. What would happen if you used the two-segment two-point Gauss quadrature rule within the significant range of  $[0, 10]$ ?

**Figure 3** Plot of integrand**Solution**

In doing this, no change of variables is necessary—only a change in the limits of each segment is needed to apply Gauss quadrature rule. Observe

$$\begin{aligned}\Gamma(2.4) &= \int_0^\infty e^{-t} t^{1.4} dt \\ &\approx \int_0^{10} e^{-t} t^{1.4} dt \\ &= \int_0^{2.4} e^{-t} t^{1.4} dt + \int_{2.4}^{10} e^{-t} t^{1.4} dt\end{aligned}$$

Setting  $f(t) = e^{-t} t^{1.4}$  to make the change of variables, we get

$$\begin{aligned}\Gamma(2.4) &\approx \frac{2.4 - 0}{2} \int_{-1}^1 f\left(\frac{2.4 - 0}{2}t + \frac{2.4 + 0}{2}\right) dt + \frac{10 - 2.4}{2} \int_{-1}^1 f\left(\frac{10 - 2.4}{2}t + \frac{10 + 2.4}{2}\right) dt \\ &= 1.2 \int_{-1}^1 f(1.2t + 1.2) dt + 3.8 \int_{-1}^1 f(3.8t + 6.2) dt\end{aligned}$$

Applying two-point Gauss quadrature rule gets

$$c_1 = 1.000000000$$

$$t_1 = -0.577350269$$

$$c_2 = 1.000000000$$

$$t_2 = 0.577350269$$

$$\begin{aligned}\Gamma(2.4) &\approx 1.2c_1 f(1.2t_1 + 1.2) + 1.2c_2 f(1.2t_2 + 1.2) + 3.8c_1 f(3.8t_1 + 6.2) + 3.8c_2 f(3.8t_2 + 6.2) \\ &= 1.2f(1.2(-0.57735) + 1.2) + 1.2f(1.2(0.57735) + 1.2) \\ &\quad + 3.8f(3.8(-0.57735) + 6.2) + 3.8f(3.8(0.57735) + 6.2) \\ &= 1.2f(0.50718) + 1.2f(1.89282) + 3.8f(4.00607) + 3.8f(8.39393) \\ &= 1.2(0.23279) + 1.2(0.36805) + 3.8(0.12706) + 3.8(0.00445) \\ &= 1.2207\end{aligned}$$

since

$$f(0.50718) = e^{-0.50718} 0.50718^{1.4} = 0.23279$$

$$f(1.89282) = e^{-1.89282} 1.89282^{1.4} = 0.36805$$

$$f(4.00607) = e^{-4.00607} 4.00607^{1.4} = 0.12706$$

$$f(8.39393) = e^{-8.39393} 8.39393^{1.4} = 0.00445$$

The absolute relative true error,  $|\epsilon_t|$ , is

$$\begin{aligned} |\epsilon_t| &= \left| \frac{1.2422 - 1.2207}{1.2422} \right| \times 100\% \\ &= 1.731\% \end{aligned}$$

---

## INTEGRATION

---

Topic	Integrating improper functions
Summary	These are textbook notes of integrating improper functions
Major	General Engineering
Authors	Autar Kaw, Michael Keteltas
Last Revised	December 23, 2009
Web Site	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

---

# Chapter 07.08

## Simpson 3/8 Rule for Integration

After reading this chapter, you should be able to

1. derive the formula for Simpson's 3/8 rule of integration,
2. use Simpson's 3/8 rule it to solve integrals,
3. develop the formula for multiple-segment Simpson's 3/8 rule of integration,
4. use multiple-segment Simpson's 3/8 rule of integration to solve integrals,
5. compare true error formulas for multiple-segment Simpson's 1/3 rule and multiple-segment Simpson's 3/8 rule, and
6. use a combination of Simpson's 1/3 rule and Simpson's 3/8 rule to approximate integrals.

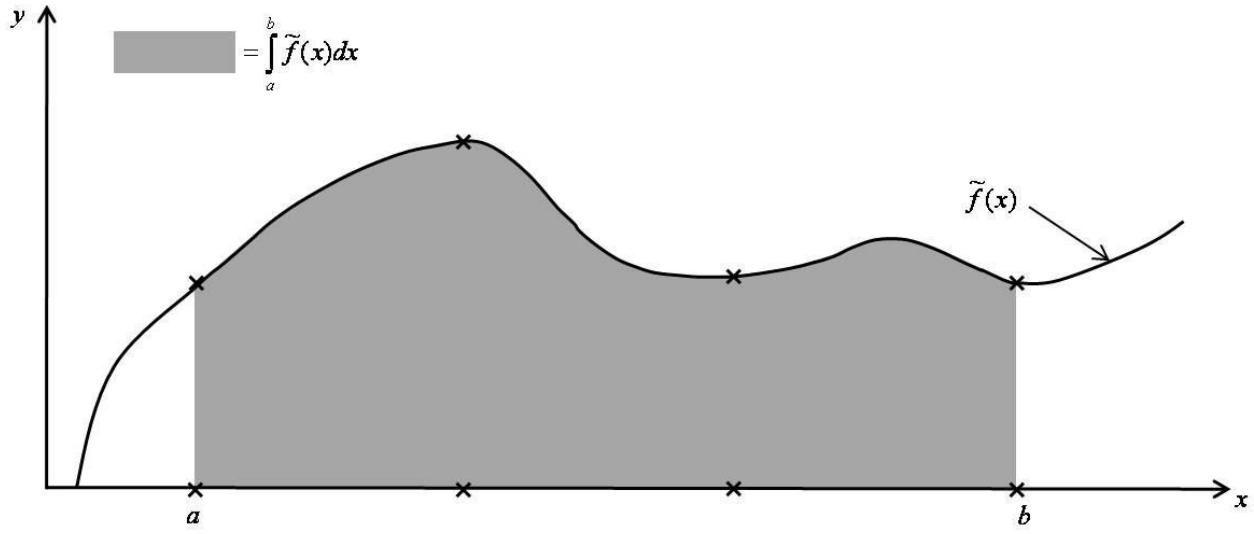
### Introduction

The main objective of this chapter is to develop appropriate formulas for approximating the integral of the form

$$I = \int_a^b f(x)dx \quad (1)$$

Most (if not all) of the developed formulas for integration are based on a simple concept of approximating a given function  $f(x)$  by a simpler function (usually a polynomial function)  $f_i(x)$ , where  $i$  represents the order of the polynomial function. In Chapter 07.03, Simpsons 1/3 rule for integration was derived by approximating the integrand  $f(x)$  with a 2<sup>nd</sup> order (quadratic) polynomial function.  $f_2(x)$

$$f_2(x) = a_0 + a_1x + a_2x^2 \quad (2)$$



**Figure 1**  $\tilde{f}(x)$  Cubic function.

In a similar fashion, Simpson 3/8 rule for integration can be derived by approximating the given function  $f(x)$  with the 3<sup>rd</sup> order (cubic) polynomial  $f_3(x)$

$$\left. \begin{aligned} f_3(x) &= a_0 + a_1x + a_2x^2 + a_3x^3 \\ &= \{1, x, x^2, x^3\} \times \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} \end{aligned} \right\} \quad (3)$$

which can also be symbolically represented in Figure 1.

#### Method 1

The unknown coefficients  $a_0, a_1, a_2$  and  $a_3$  in Equation (3) can be obtained by substituting 4 known coordinate data points  $\{x_0, f(x_0)\}, \{x_1, f(x_1)\}, \{x_2, f(x_2)\}$  and  $\{x_3, f(x_3)\}$  into Equation (3) as follows.

$$\left. \begin{aligned} f(x_0) &= a_0 + a_1x_0 + a_2x_0^2 + a_3x_0^3 \\ f(x_1) &= a_0 + a_1x_1 + a_2x_1^2 + a_3x_1^3 \\ f(x_2) &= a_0 + a_1x_2 + a_2x_2^2 + a_3x_2^3 \\ f(x_3) &= a_0 + a_1x_3 + a_2x_3^2 + a_3x_3^3 \end{aligned} \right\} \quad (4)$$

Equation (4) can be expressed in matrix notation as

$$\begin{bmatrix} 1 & x_0 & x_0^2 & x_0^3 \\ 1 & x_1 & x_1^2 & x_1^3 \\ 1 & x_2 & x_2^2 & x_2^3 \\ 1 & x_3 & x_3^2 & x_3^3 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} f(x_0) \\ f(x_1) \\ f(x_2) \\ f(x_3) \end{bmatrix} \quad (5)$$

The above Equation (5) can symbolically be represented as

$$[A]_{4 \times 4} \vec{a}_{4 \times 1} = \vec{f}_{4 \times 1} \quad (6)$$

Thus,

$$\vec{a} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} = [A]^{-1} \times \vec{f} \quad (7)$$

Substituting Equation (7) into Equation (3), one gets

$$f_3(x) = \{1, x, x^2, x^3\} \times [A]^{-1} \times \vec{f} \quad (8)$$

As indicated in Figure 1, one has

$$\left. \begin{aligned} x_0 &= a \\ x_1 &= a + h \\ &= a + \frac{b-a}{3} \\ &= \frac{2a+b}{3} \\ x_2 &= a + 2h \\ &= a + \frac{2b-2a}{3} \\ &= \frac{a+2b}{3} \\ x_3 &= a + 3h \\ &= a + \frac{3b-3a}{3} \\ &= b \end{aligned} \right\} \quad (9)$$

With the help from MATLAB [Ref. 2], the unknown vector  $\vec{a}$  (shown in Equation 7) can be solved for symbolically.

### Method 2

Using Lagrange interpolation, the cubic polynomial function  $f_3(x)$  that passes through 4 data points (see Figure 1) can be explicitly given as

$$f_3(x) = \frac{(x - x_1)(x - x_2)(x - x_3)}{(x_0 - x_1)(x_0 - x_2)(x_0 - x_3)} \times f(x_0) + \frac{(x - x_0)(x - x_2)(x - x_3)}{(x_1 - x_0)(x_1 - x_2)(x_1 - x_3)} \times f(x_1) \\ + \frac{(x - x_0)(x - x_1)(x - x_3)}{(x_2 - x_0)(x_2 - x_1)(x_2 - x_3)} \times f(x_3) + \frac{(x - x_0)(x - x_1)(x - x_2)}{(x_3 - x_0)(x_3 - x_1)(x_3 - x_2)} \times f(x_3) \quad (10)$$

### Simpsons 3/8 Rule for Integration

Substituting the form of  $f_3(x)$  from Method (1) or Method (2),

$$I = \int_a^b f(x) dx \\ \approx \int_a^b f_3(x) dx \\ = (b - a) \times \frac{\{f(x_0) + 3f(x_1) + 3f(x_2) + f(x_3)\}}{8} \quad (11)$$

Since

$$h = \frac{b - a}{3}$$

$$b - a = 3h$$

and Equation (11) becomes

$$I \approx \frac{3h}{8} \times \{f(x_0) + 3f(x_1) + 3f(x_2) + f(x_3)\} \quad (12)$$

Note the 3/8 in the formula, and hence the name of method as the Simpson's 3/8 rule.

The true error in Simpson 3/8 rule can be derived as [Ref. 1]

$$E_t = -\frac{(b - a)^5}{6480} \times f''''(\zeta), \text{ where } a \leq \zeta \leq b \quad (13)$$

### Example 1

The vertical distance in meters covered by a rocket from  $t = 8$  to  $t = 30$  seconds is given by

$$s = \int_8^{30} \left( 2000 \ln \left[ \frac{140000}{140000 - 2100t} \right] - 9.8t \right) dt$$

Use Simpson 3/8 rule to find the approximate value of the integral.

**Solution**

$$h = \frac{b-a}{n}$$

$$= \frac{b-a}{3}$$

$$= \frac{30-8}{3}$$

$$= 7.3333$$

$$f(t) = 2000 \ln \left[ \frac{140000}{140000 - 2100t} \right] - 9.8t$$

$$I \approx \frac{3h}{8} \times \{f(t_0) + 3f(t_1) + 3f(t_2) + f(t_3)\}$$

$$t_0 = 8$$

$$f(t_0) = 2000 \ln \left( \frac{140000}{140000 - 2100 \times 8} \right) - 9.8 \times 8$$

$$= 177.2667$$

$$\begin{cases} t_1 = t_0 + h \\ = 8 + 7.3333 \\ = 15.3333 \\ f(t_1) = 2000 \ln \left( \frac{140000}{140000 - 2100 \times 15.3333} \right) - 9.8 \times 15.3333 \\ = 372.4629 \end{cases}$$

$$\begin{cases} t_2 = t_0 + 2h \\ = 8 + 2(7.3333) \\ = 22.6666 \\ f(t_2) = 2000 \ln \left( \frac{140000}{140000 - 2100 \times 22.6666} \right) - 9.8 \times 22.6666 \\ = 608.8976 \end{cases}$$

$$\begin{cases} t_3 = t_0 + 3h \\ \quad = 8 + 3(7.3333) \\ \quad = 30 \\ f(t_3) = 2000 \ln\left(\frac{140000}{140000 - 2100 \times 30}\right) - 9.8 \times 30 \\ \quad = 901.6740 \end{cases}$$

Applying Equation (12), one has

$$\begin{aligned} I &= \frac{3}{8} \times 7.3333 \times \{177.2667 + 3 \times 372.4629 + 3 \times 608.8976 + 901.6740\} \\ &= 11063.3104 \text{ m} \end{aligned}$$

The exact answer can be computed as

$$I_{exact} = 11061.34 \text{ m}$$

### Multiple Segments for Simpson 3/8 Rule

Using  $n$  = number of equal segments, the width  $h$  can be defined as

$$h = \frac{b-a}{n} \quad (14)$$

The number of segments need to be an integer multiple of 3 as a single application of Simpson 3/8 rule requires 3 segments.

The integral shown in Equation (1) can be expressed as

$$\begin{aligned} I &= \int_a^b f(x) dx \\ &\approx \int_a^b f_3(x) dx \\ &\approx \int_{x_0=a}^{x_3} f_3(x) dx + \int_{x_3}^{x_6} f_3(x) dx + \dots + \int_{x_{n-3}}^{x_n=b} f_3(x) dx \end{aligned} \quad (15)$$

Using Simpson 3/8 rule (See Equation 12) into Equation (15), one gets

$$I = \frac{3h}{8} \left\{ f(x_0) + 3f(x_1) + 3f(x_2) + f(x_3) + f(x_4) + 3f(x_5) + f(x_6) + \dots + f(x_{n-3}) + 3f(x_{n-2}) + 3f(x_{n-1}) + f(x_n) \right\} \quad (16)$$

$$I = \frac{3h}{8} \left\{ f(x_0) + 3 \sum_{i=1,4,7,\dots}^{n-2} f(x_i) + 3 \sum_{i=2,5,8,\dots}^{n-1} f(x_i) + 2 \sum_{i=3,6,9,\dots}^{n-3} f(x_i) + f(x_n) \right\} \quad (17)$$

### Example 2

The vertical distance in meters covered by a rocket from  $t = 8$  to  $t = 30$  seconds is given by

$$s = \int_8^{30} \left( 2000 \ln \left[ \frac{140000}{140000 - 2100t} \right] - 9.8t \right) dt$$

Use Simpson 3/8 multiple segments rule with six segments to estimate the vertical distance.

### Solution

In this example, one has (see Equation 14):

$$f(t) = 2000 \ln \left[ \frac{140000}{140000 - 2100t} \right] - 9.8t$$

$$h = \frac{30 - 8}{6} = 3.6666$$

$$\{t_0, f(t_0)\} = \{8, 177.2667\}$$

$$\{t_1, f(t_1)\} = \{11.6666, 270.4104\} \text{ where } t_1 = t_0 + h = 8 + 3.6666 = 11.6666$$

$$\{t_2, f(t_2)\} = \{15.3333, 372.4629\} \text{ where } t_2 = t_0 + 2h = 15.3333$$

$$\{t_3, f(t_3)\} = \{19.4847455\} \text{ where } t_3 = t_0 + 3h = 19$$

$$\{t_4, f(t_4)\} = \{22.6666, 608.8976\} \text{ where } t_4 = t_0 + 4h = 22.6666$$

$$\{t_5, f(t_5)\} = \{26.3333, 746.9870\} \text{ where } t_5 = t_0 + 5h = 26.3333$$

$$\{t_6, f(t_6)\} = \{30.9016740\} \text{ where } t_6 = t_0 + 6h = 30$$

Applying Equation (17), one obtains:

$$\begin{aligned} I &= \frac{3}{8} (3.6666) \left\{ 177.2667 + 3 \sum_{i=1,4,\dots}^{n-2=4} f(t_i) + 3 \sum_{i=2,5,\dots}^{n-1=5} f(t_i) + 2 \sum_{i=3,6,\dots}^{n-3=3} f(t_i) + 901.6740 \right\} \\ &= (1.3750) \left\{ 177.2667 + 3(270.4104 + 608.8976) + 3(372.4629 + 746.9870) + 2(484.7455) + 901.6740 \right\} \\ &= 11,601.4696 m \end{aligned}$$

### Example 3

Compute

$$I = \int_8^{30} \left( 2000 \ln \left[ \frac{140000}{140000 - 2100t} \right] - 9.8t \right) dt,$$

using Simpson 1/3 rule (with  $n_1 = 4$ ), and Simpson 3/8 rule (with  $n_2 = 3$ ).

### Solution

The segment width is

$$\begin{aligned} h &= \frac{b-a}{n} \\ &= \frac{b-a}{n_1 + n_2} \end{aligned}$$

$$\begin{aligned} &= \frac{30 - 8}{(4 + 3)} \\ &= 3.1429 \end{aligned}$$

$$\left. \begin{aligned} f(t) &= 2000 \ln \left[ \frac{140000}{140000 - 2100t} \right] - 9.8t \\ t_0 &= a = 8 \\ t_1 &= x_0 + 1h = 8 + 3.1429 = 11.1429 \\ t_2 &= t_0 + 2h = 8 + 2(3.1429) = 14.2857 \\ t_3 &= t_0 + 3h = 8 + 3(3.1429) = 17.4286 \\ t_4 &= t_0 + 4h = 8 + 4(3.1429) = 20.5714 \\ t_5 &= t_0 + 5h = 8 + 5(3.1429) = 23.7143 \\ t_6 &= t_0 + 6h = 8 + 6(3.1429) = 26.8571 \\ t_7 &= t_0 + 7h = 8 + 7(3.1429) = 30 \end{aligned} \right\} \text{Simpson's 1/3 rule}$$

Now

$$\begin{aligned} f(t_0 = 8) &= 2000 \ln \left( \frac{140,000}{140,000 - 2100 \times 8} \right) - 9.8 \times 8 \\ &= 177.2667 \end{aligned}$$

Similarly:

$$\begin{aligned} f(t_1) &= 256.5863 \\ f(t_2) &= 342.3241 \\ f(t_3) &= 435.2749 \\ f(t_4) &= 536.3909 \\ f(t_5) &= 646.8260 \\ f(t_6) &= 767.9978 \\ f(t_7) &= 901.6740 \end{aligned}$$

For multiple segments ( $n_1$  = first 4 segments), using Simpson 1/3 rule, one obtains (See Equation 19):

$$\begin{aligned} I_1 &= \left( \frac{h}{3} \right) \left\{ f(t_0) + 4 \sum_{i=1,3,\dots}^{n_1-1=3} f(t_i) + 2 \sum_{i=2,\dots}^{n_1-2=2} f(t_i) + f(t_{n_1}) \right\} \\ &= \left( \frac{h}{3} \right) \{ f(t_0) + 4(f(t_1) + f(t_3)) + 2f(t_2) + f(t_4) \} \\ &= \left( \frac{3.1429}{3} \right) \{ 177.2667 + 4(256.5863 + 435.2749) + 2(342.3241) + 536.3909 \} \\ &= 4364.1197 \end{aligned}$$

For multiple segments ( $n_2 = \text{last 3 segments}$ ), using Simpson 3/8 rule, one obtains (See Equation 17):

$$\begin{aligned}
I_2 &= \left( \frac{3h}{8} \right) \left\{ f(t_0) + 3 \sum_{i=1,3,\dots}^{n_2-2=1} f(t_i) + 3 \sum_{i=2,\dots}^{n_2-1=2} f(t_i) + 2 \sum_{i=3,6,\dots}^{n_2-3=0} f(t_i) + f(t_{n_1}) \right\} \\
&= \left( \frac{3h}{8} \right) \{ f(t_0) + 3f(t_1) + 3f(t_2) + 2(\text{no contribution}) + f(t_3) \} \\
&= \left( \frac{3h}{8} \right) \{ f(t_4) + 3f(t_5) + 3f(t_6) + f(t_7) \} \\
&= \left( \frac{3}{8} \times 3.1429 \right) \{ 536.3909 + 3(646.8260) + 3(767.9978) + 901.6740 \} \\
&= 6697.3663
\end{aligned}$$

The mixed (combined) Simpson 1/3 and 3/8 rules give

$$\begin{aligned}
I &= I_1 + I_2 \\
&= 4364.1197 + 6697.3663 \\
&= 11061m
\end{aligned}$$

Comparing the truncated error of Simpson 1/3 rule

$$E_t = -\frac{(b-a)^5}{2880} \times f''''(\zeta) \quad (18)$$

With Simpson 3/8 rule (See Equation 12), it seems to offer slightly more accurate answer than the former. However, the cost associated with Simpson 3/8 rule (using 3rd order polynomial function) is significantly higher than the one associated with Simpson 1/3 rule (using 2nd order polynomial function).

The number of multiple segments that can be used in the conjunction with Simpson 1/3 rule is 2, 4, 6, 8, ... (any even numbers) for

$$\begin{aligned}
I &= \int_a^b f(x)dx \\
&\approx \left( \frac{h}{3} \right) \{ f(x_0) + 4f(x_1) + f(x_2) + f(x_3) + 4f(x_4) + \dots + f(x_{n-2}) + 4f(x_{n-1}) + f(x_n) \} \\
&= \left( \frac{h}{3} \right) \left\{ f(x_0) + 4 \sum_{i=1,3,\dots}^{n-1} f(x_i) + 2 \sum_{i=2,4,6,\dots}^{n-2} f(x_i) + f(x_n) \right\}
\end{aligned} \quad (19)$$

However, Simpson 3/8 rule can be used with the number of segments equal to 3, 6, 9, 12, .. (can be certain integers that are multiples of 3).

If the user wishes to use, say 7 segments, then the mixed Simpson 1/3 rule (for the first 4 segments), and Simpson 3/8 rule (for the last 3 segments) would be appropriate.

### Computer Algorithm for Mixed Simpson 1/3 and 3/8 Rule for Integration

Based on the earlier discussion on (single and multiple segments) Simpson 1/3 and 3/8 rules, the following “pseudo” step-by-step mixed Simpson rules for estimating

$$I = \int_a^b f(x) dx$$

can be given as

#### Step 1

User inputs information, such as

$$f(x) = \text{integrand}$$

$n_1$  = number of segments in conjunction with Simpson 1/3 rule (a multiple of 2 (any even numbers))

$n_2$  = number of segments in conjunction with Simpson 3/8 rule (a multiple of 3)

#### Step 2

Compute

$$n = n_1 + n_2$$

$$h = \frac{b - a}{n}$$

$$x_0 = a$$

$$x_1 = a + 1h$$

$$x_2 = a + 2h$$

$$\dots$$

$$x_i = a + ih$$

$$\dots$$

$$x_n = a + nh = b$$

#### Step 3

Compute result from multiple-segment Simpson 1/3 rule (See Equation 19)

$$I_1 = \left( \frac{h}{3} \right) \left\{ f(x_0) + 4 \sum_{i=1,3,\dots}^{n_1-1} f(x_i) + 2 \sum_{i=2,4,6,\dots}^{n_1-2} f(x_i) + f(x_{n_1}) \right\} \quad (19, \text{ repeated})$$

#### Step 4

Compute result from multiple segment Simpson 3/8 rule (See Equation 17)

$$I_2 = \left( \frac{3h}{8} \right) \left\{ f(x_0) + 3 \sum_{i=1,4,7,\dots}^{n_2-2} f(x_i) + 3 \sum_{i=2,5,8,\dots}^{n_2-1} f(x_i) + 2 \sum_{i=3,6,9,\dots}^{n_2-3} f(x_i) + f(x_{n_2}) \right\} \quad (17, \text{ repeated})$$

#### Step 5

$$I \approx I_1 + I_2 \quad (20)$$

and print out the final approximated answer for  $I$ .

---

### **SIMPSON'S 3/8 RULE FOR INTEGRATION**

---

Topic	Simpson 3/8 Rule for Integration
Summary	Textbook Chapter of Simpson's 3/8 Rule for Integration
Major	General Engineering
Authors	Duc Nguyen
Date	July 9, 2017

---

Web Site <http://numericalmethods.eng.usf.edu>

---

# Chapter 08.01

## Primer for Ordinary Differential Equations

After reading this chapter, you should be able to:

1. define an ordinary differential equation,
2. differentiate between an ordinary and partial differential equation, and
3. solve linear ordinary differential equations with fixed constants by using classical solution and Laplace transform techniques.

### Introduction

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 leads 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

- (A) ordinary differential equations (ODE)
- (B) 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 equations include

$$\frac{d^2y}{dx^2} + 2 \frac{dy}{dx} + y = 0, \quad \frac{dy}{dx}(0) = 2, \quad y(0) = 4,$$

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

Ordinary differential equations are classified in terms of order and degree. *Order* of an ordinary differential equation is the same as the highest derivative and the *degree* of an ordinary differential equation is the power of highest derivative.

Thus the differential equation,

$$x^3 \frac{d^3y}{dx^3} + x^2 \frac{d^2y}{dx^2} + x \frac{dy}{dx} + xy = e^x$$

is of order 3 and degree 1, whereas the differential equation

$$\left(\frac{dy}{dx} + 1\right)^2 + x^2 \frac{dy}{dx} = \sin x$$

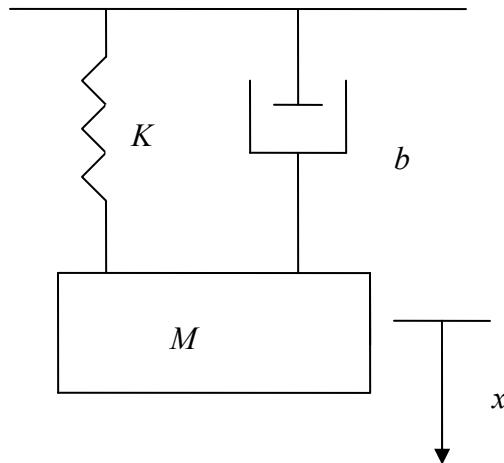
is of order 1 and degree 2.

An engineer's approach to differential equations is different from a mathematician. While, the latter is interested in the mathematical solution, an engineer should be able to interpret the result physically. So, an engineer's approach can be divided into three phases:

- a) formulation of a differential equation from a given physical situation,
- b) solving the differential equation and evaluating the constants, using given conditions, and
- c) interpreting the results physically for implementation.

### Formulation of differential equations

As discussed above, the formulation of a differential equation is based on a given physical situation. This can be illustrated by a spring-mass-damper system.



**Figure 1** Spring-mass damper system.

Above is the schematic diagram of a spring-mass-damper system. A block is suspended freely using a spring. As most physical systems involve some kind of damping - viscous damping, dry damping, magnetic damping, etc., a damper or dashpot is attached to account for viscous damping.

Let the mass of the block be \$M\$, the spring constant be \$K\$, and the damper coefficient be \$b\$. If we measure displacement from the static equilibrium position we need not consider gravitational force as it is balanced by tension in the spring at equilibrium.

Below is the free body diagram of the block at static and dynamic equilibrium. So, the equation of motion is given by

$$Ma = F_s + F_d \quad (1)$$

where

\$F\_s\$ is the restoring force due to spring.

$F_D$  is the damping force due to the damper.

$a$  is the acceleration.

The restoring force in the spring is given by

$$F_S = -Kx \quad (2)$$

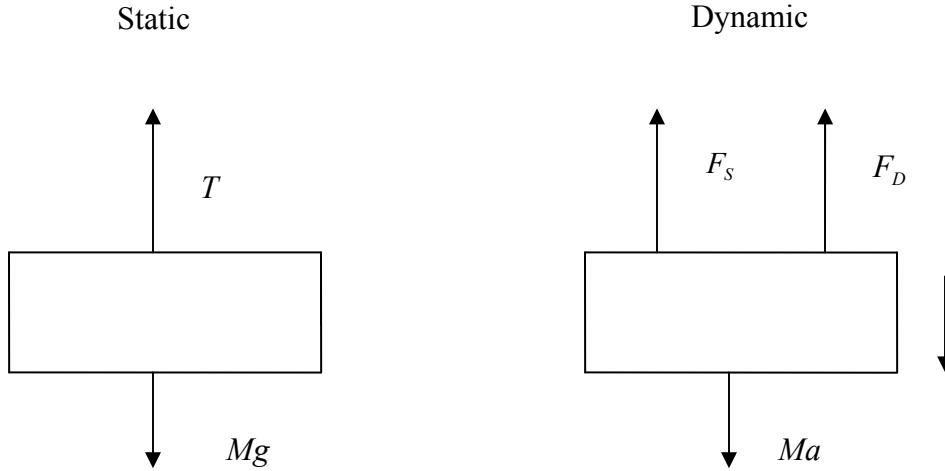
as the restoring force is proportional to displacement and it is negative as it opposes the motion. The damping force in the damper is given by

$$F_D = -bv \quad (3)$$

as the damping force is directly proportional to velocity and also opposes motion.

Therefore, the equation of motion can be written as

$$Ma = -Kx - bv \quad (4)$$



**Figure 2** Free body diagram of spring-mass-damper system.

Since

$$a = \frac{d^2 x}{dt^2} \text{ and } v = \frac{dx}{dt}$$

from Equation (4), we get

$$\begin{aligned} M \frac{d^2 x}{dt^2} &= -Kx - b \frac{dx}{dt} \\ M \frac{d^2 x}{dt^2} + b \frac{dx}{dt} + Kx &= 0 \end{aligned} \quad (5)$$

This is an ordinary differential equation of second order and of degree one.

### Solution to linear ordinary differential equations

In this section we discuss two techniques used to solve ordinary differential equations

- (A) Classical technique
- (B) Laplace transform technique

### Classical Technique

The general form of a linear ordinary differential equation with constant coefficients is given by

$$\frac{d^n y}{dx^n} + k_n \frac{d^{n-1} y}{dx^{n-1}} + \dots + k_3 \frac{d^2 y}{dx^2} + k_2 \frac{dy}{dx} + k_1 y = F(x) \quad (6)$$

The general solution contains two parts

$$y = y_H + y_P \quad (7)$$

where

$y_H$  is the homogeneous part of the solution and

$y_P$  is the particular part of the solution.

The homogeneous part of the solution  $y_H$  is that part of the solution that gives zero when substituted in the left hand side of the equation. So,  $y_H$  is solution of the equation

$$\frac{d^n y}{dx^n} + k_n \frac{d^{n-1} y}{dx^{n-1}} + \dots + k_3 \frac{d^2 y}{dx^2} + k_2 \frac{dy}{dx} + k_1 y = 0 \quad (8)$$

The above equation can be symbolically written as

$$D^n y + k_n D^{n-1} y + \dots + k_2 D y + k_1 y = 0 \quad (9)$$

$$(D^n + k_n D^{n-1} + \dots + k_2 D + k_1) y = 0 \quad (10)$$

where,

$$D^n = \frac{d^n}{dx^n} \quad (11)$$

$$D^{n-1} = \frac{d^{n-1}}{dx^{n-1}}$$

.

.

operating on  $y$  is the same as

$$(D - r_1), (D - r_2), (D - r_n)$$

operating one after the other in any order, where

$$(D - r_1), (D - r_2), \dots, (D - r_n)$$

are factors of

$$D^n + k_n D^{n-1} + \dots + k_2 D + k_1 = 0 \quad (12)$$

To illustrate

$$(D^2 - 3D + 2)y = 0$$

is same as

$$(D - 2)(D - 1)y = 0$$

$$(D - 1)(D - 2)y = 0$$

Therefore,

$$(D^n + k_n D^{n-1} + \dots + k_2 D + k_1)y = 0 \quad (13)$$

is same as

$$(D - r_n)(D - r_{n-1}) \dots (D - r_1)y = 0 \quad (14)$$

operating one after the other in any order.

### Case 1: Roots are real and distinct

The entire left hand side becomes zero if  $(D - r_1)y = 0$ . Therefore, the solution to  $(D - r_1)y = 0$  is a solution to a homogeneous equation.  $(D - r_1)y = 0$  is called Leibnitz's linear differential equation of first order and its solution is

$$(D - r_1)y = 0 \quad (15)$$

$$\frac{dy}{dx} = r_1 y \quad (16)$$

$$\frac{dy}{y} = r_1 dx \quad (17)$$

Integrating both sides we get

$$\ln y = r_1 x + c \quad (18)$$

$$y = ce^{r_1 x} \quad (19)$$

Since any of the  $n$  factors can be placed before  $y$ , there are  $n$  different solutions corresponding to  $n$  different factors given by

$$C_n e^{r_n x}, C_{n-1} e^{r_{n-1} x}, \dots, C_2 e^{r_2 x}, C_1 e^{r_1 x}$$

where

$r_n, r_{n-1}, \dots, r_2, r_1$  are the roots of Equation (12) and

$C_n, C_{n-1}, \dots, C_2, C_1$  are constants.

We get the general solution for a homogeneous equation by superimposing the individual Leibnitz's solutions. Therefore

$$y_H = C_1 e^{r_1 x} + C_2 e^{r_2 x} + \dots + C_{n-1} e^{r_{n-1} x} + C_n e^{r_n x} \quad (20)$$

### Case 2: Roots are real and identical

If two roots of a homogeneous equation are equal, say  $r_1 = r_2$ , then

$$(D - r_n)(D - r_{n-1}) \dots (D - r_1)(D - r_1)y = 0 \quad (21)$$

Let's work at

$$(D - r_1)(D - r_1)y = 0 \quad (22)$$

If

$$(D - r_1)y = z \quad (23)$$

then

$$(D - r_1)z = 0$$

$$z = C_2 e^{r_1 x} \quad (24)$$

Now substituting the solution from Equation (24) in Equation (23)

$$\begin{aligned} (D - r_1)y &= C_2 e^{r_1 x} \\ \frac{dy}{dx} - r_1 y &= C_2 e^{r_1 x} \\ e^{-r_1 x} \frac{dy}{dx} - r_1 e^{-r_1 x} y &= C_2 \\ \frac{d(e^{-r_1 x} y)}{dx} &= C_2 \\ d(e^{-r_1 x} y) &= C_2 dx \end{aligned} \quad (25)$$

Integrating both sides of Equation (25), we get

$$\begin{aligned} e^{-r_1 x} y &= C_2 x + C_1 \\ y &= (C_2 x + C_1) e^{r_1 x} \end{aligned} \quad (26)$$

Therefore the final homogeneous solution is given by

$$y_H = (C_1 + C_2 x) e^{r_1 x} + C_3 e^{r_3 x} + \dots + C_n e^{r_n x} \quad (27)$$

Similarly, if  $m$  roots are equal the solution is given by

$$y_H = (C_1 + C_2 x + C_3 x^2 + \dots + C_m x^{m-1}) e^{r_m x} + C_{m+1} e^{r_{m+1} x} + \dots + C_n e^{r_n x} \quad (28)$$

### Case 3: Roots are complex

If one pair of roots is complex, say  $r_1 = \alpha + i\beta$  and  $r_2 = \alpha - i\beta$ ,

where

$$i = \sqrt{-1}$$

then

$$y_H = C_1 e^{(\alpha+i\beta)x} + C_2 e^{(\alpha-i\beta)x} + C_3 e^{r_3 x} + \dots + C_n e^{r_n x} \quad (29)$$

Since

$$e^{i\beta x} = \cos \beta x + i \sin \beta x, \text{ and} \quad (30a)$$

$$e^{-i\beta x} = \cos \beta x - i \sin \beta x \quad (30b)$$

then

$$\begin{aligned} y_H &= C_1 e^{\alpha x} (\cos \beta x + i \sin \beta x) + C_2 e^{\alpha x} (\cos \beta x - i \sin \beta x) + C_3 e^{r_3 x} + \dots + C_n e^{r_n x} \\ &= (C_1 + C_2) e^{\alpha x} \cos \beta x + i(C_1 - C_2) e^{\alpha x} \sin \beta x + C_3 e^{r_3 x} + \dots + C_n e^{r_n x} \\ &= e^{\alpha x} (A \cos \beta x + B \sin \beta x) + C_3 e^{r_3 x} + \dots + C_n e^{r_n x} \end{aligned} \quad (31)$$

where

$$\begin{aligned} A &= C_1 + C_2 \text{ and} \\ B &= i(C_1 - C_2) \end{aligned} \quad (32)$$

Now, let us look at how the particular part of the solution is found. Consider the general form of the ordinary differential equation

$$(D^n + k_n D^{n-1} + k_{n-1} D^{n-2} + \dots + k_1) y = X \quad (33)$$

The particular part of the solution  $y_p$  is that part of solution that gives  $X$  when substituted for  $y$  in the above equation, that is,

$$(D^n + k_n D^{n-1} + k_{n-1} D^{n-2} + \dots + k_1) y_p = X \quad (34)$$

**Sample Case 1**

When  $X = e^{ax}$ , the particular part of the solution is of the form  $Ae^{ax}$ . We can find  $A$  by substituting  $y = Ae^{ax}$  in the left hand side of the differential equation and equating coefficients.

**Example 1**

Solve

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

**Solution**

The homogeneous solution for the above equation is given by

$$(3D + 2)y = 0$$

The characteristic equation for the above equation is given by

$$3r + 2 = 0$$

The solution to the equation is

$$r = -0.666667$$

$$y_H = Ce^{-0.666667x}$$

The particular part of the solution is of the form  $Ae^{-x}$

$$3\frac{d(Ae^{-x})}{dx} + 2Ae^{-x} = e^{-x}$$

$$-3Ae^{-x} + 2Ae^{-x} = e^{-x}$$

$$-Ae^{-x} = e^{-x}$$

$$A = -1$$

Hence the particular part of the solution is

$$y_p = -e^{-x}$$

The complete solution is given by

$$\begin{aligned} y &= y_H + y_p \\ &= Ce^{-0.666667x} - e^{-x} \end{aligned}$$

The constant  $C$  can be obtained by using the initial condition  $y(0) = 5$

$$y(0) = Ce^{-0.666667 \times 0} - e^{-0} = 5$$

$$C - 1 = 5$$

$$C = 6$$

The complete solution is

$$y = 6e^{-0.666667x} - e^{-x}$$

**Example 2**

Solve

$$2\frac{dy}{dx} + 3y = e^{-1.5x}, \quad y(0) = 5$$

### Solution

The homogeneous solution for the above equation is given by

$$(2D + 3)y = 0$$

The characteristic equation for the above equation is given by

$$2r + 3 = 0$$

The solution to the equation is

$$r = -1.5$$

$$y_H = Ce^{-1.5x}$$

Based on the forcing function of the ordinary differential equations, the particular part of the solution is of the form  $Ae^{-1.5x}$ , but since that is part of the form of the homogeneous part of the solution, we need to choose the next independent solution, that is,

$$y_P = Axe^{-1.5x}$$

To find  $A$ , we substitute this solution in the ordinary differential equation as

$$2 \frac{d(Axe^{-1.5x})}{dx} + 3Axe^{-1.5x} = e^{-1.5x}$$

$$2Ae^{-1.5x} - 3Axe^{-1.5x} + 3Axe^{-1.5x} = e^{-1.5x}$$

$$2Ae^{-1.5x} = e^{-1.5x}$$

$$A = 0.5$$

Hence the particular part of the solution is

$$y_P = 0.5xe^{-1.5x}$$

The complete solution is given by

$$\begin{aligned} y &= y_H + y_P \\ &= Ce^{-1.5x} + 0.5xe^{-1.5x} \end{aligned}$$

The constant  $C$  is obtained by using the initial condition  $y(0) = 5$ .

$$y(0) = Ce^{-1.5(0)} + 0.5(0)e^{-1.5(0)} = 5$$

$$C + 0 = 5$$

$$C = 5$$

The complete solution is

$$y = 5e^{-1.5x} + 0.5xe^{-1.5x}$$

### Sample Case 2

When

$$X = \sin(ax) \text{ or } \cos(ax),$$

the particular part of the solution is of the form

$$A\sin(ax) + B\cos(ax).$$

We can get  $A$  and  $B$  by substituting  $y = A\sin(ax) + B\cos(ax)$  in the left hand side of the differential equation and equating coefficients.

### Example 3

Solve

$$2\frac{d^2y}{dx^2} + 3\frac{dy}{dx} + 3.125y = \sin x, \quad y(0) = 5, \quad \frac{dy}{dx}(x=0) = 3$$

**Solution**

The homogeneous equation is given by

$$(2D^2 + 3D + 3.125)y = 0$$

The characteristic equation is

$$2r^2 + 3r + 3.125 = 0$$

The roots of the characteristic equation are

$$\begin{aligned} r &= \frac{-3 \pm \sqrt{3^2 - 4 \times 2 \times 3.125}}{2 \times 2} \\ &= \frac{-3 \pm \sqrt{9 - 25}}{4} \\ &= \frac{-3 \pm \sqrt{-16}}{4} \\ &= \frac{-3 \pm 4i}{4} \\ &= -0.75 \pm i \end{aligned}$$

Therefore the homogeneous part of the solution is given by

$$y_H = e^{-0.75x}(K_1 \cos x + K_2 \sin x)$$

The particular part of the solution is of the form

$$y_P = A \sin x + B \cos x$$

$$2\frac{d^2}{dx^2}(A \sin x + B \cos x) + 3\frac{d}{dx}(A \sin x + B \cos x) + 3.125(A \sin x + B \cos x) = \sin x$$

$$2\frac{d}{dx}(A \cos x - B \sin x) + 3(A \cos x - B \sin x) + 3.125(A \sin x + B \cos x) = \sin x$$

$$2(-A \sin x - B \cos x) + 3(A \cos x - B \sin x) + 3.125(A \sin x + B \cos x) = \sin x$$

$$(1.125A - 3B) \sin x + (1.125B + 3A) \cos x = \sin x$$

Equating coefficients of  $\sin x$  and  $\cos x$  on both sides, we get

$$1.125A - 3B = 1$$

$$1.125B + 3A = 0$$

Solving the above two simultaneous linear equations we get

$$A = 0.109589$$

$$B = -0.292237$$

Hence

$$y_P = 0.109589 \sin x - 0.292237 \cos x$$

The complete solution is given by

$$y = e^{-0.75x}(K_1 \cos x + K_2 \sin x) + (0.109589 \sin x - 0.292237 \cos x)$$

To find  $K_1$  and  $K_2$  we use the initial conditions

$$y(0) = 5, \quad \frac{dy}{dx}(x=0) = 3$$

From  $y(0) = 5$  we get

$$5 = e^{-0.75(0)}(K_1 \cos(0) + K_2 \sin(0)) + (0.109589 \sin(0) - 0.292237 \cos(0))$$

$$5 = K_1 - 0.292237$$

$$K_1 = 5.292237$$

$$\begin{aligned} \frac{dy}{dx} &= -0.75e^{-0.75x}(K_1 \cos x + K_2 \sin x) + e^{-0.75x}(-K_1 \sin x + K_2 \cos x) \\ &\quad + 0.109589 \cos x + 0.292237 \sin x \end{aligned}$$

From

$$\frac{dy}{dx}(x = 0) = 3,$$

we get

$$3 = -0.75e^{-0.75(0)}(K_1 \cos(0) + K_2 \sin(0)) + e^{-0.75(0)}(-K_1 \sin(0) + K_2 \cos(0))$$

$$+ 0.109589 \cos(0) + 0.292237 \sin(0)$$

$$3 = -0.75K_1 + K_2 + 0.109589$$

$$3 = -0.75(5.292237) + K_2 + 0.109589$$

$$K_2 = 6.859588$$

The complete solution is

$$y = e^{-0.75x}(5.292237 \cos x + 6.859588 \sin x) + 0.109589 \sin x - 0.292237 \cos x$$

#### Example 4

Solve

$$2\frac{d^2y}{dx^2} + 6\frac{dy}{dx} + 3.125y = \cos(x), \quad y(0) = 5, \quad \frac{dy}{dx}(x = 0) = 3$$

#### Solution

The homogeneous part of the equations is given by

$$(2D^2 + 6D + 3.125)y = 0$$

The characteristic equation is given by

$$\begin{aligned} 2r^2 + 6r + 3.125 &= 0 \\ r &= \frac{-6 \pm \sqrt{6^2 - 4(2)(3.125)}}{2(2)} \\ &= \frac{-6 \pm \sqrt{36 - 25}}{4} \\ &= \frac{-6 \pm \sqrt{11}}{4} \\ &= -1.5 \pm 0.829156 \\ &= -0.670844, -2.329156 \end{aligned}$$

Therefore, the homogeneous solution  $y_H$  is given by

$$y_H = K_1 e^{-0.670845x} + K_2 e^{-2.329156x}$$

The particular part of the solution is of the form

$$y_P = A \sin x + B \cos x$$

Substituting the particular part of the solution in the differential equation,

$$\begin{aligned} 2 \frac{d^2}{dx^2}(A \sin x + B \cos x) + 6 \frac{d}{dx}(A \sin x + B \cos x) \\ + 3.125(A \sin x + B \cos x) = \cos x \\ 2 \frac{d}{dx}(A \cos x - B \sin x) + 6(A \cos x - B \sin x) \\ + 3.125(A \sin x + B \cos x) = \cos x \\ 2(-A \sin x - B \cos x) + 6(A \cos x - B \sin x) \\ + 3.125(A \sin x + B \cos x) = \cos x \\ (1.125A - 6B) \sin x + (1.125B + 6A) \cos x = \cos x \end{aligned}$$

Equating coefficients of  $\cos x$  and  $\sin x$  we get

$$1.125B + 6A = 1$$

$$1.125A - 6B = 0$$

The solution to the above two simultaneous linear equations are

$$A = 0.161006$$

$$B = 0.0301887$$

Hence the particular part of the solution is

$$y_p = 0.161006 \sin x + 0.0301887 \cos x$$

Therefore the complete solution is

$$\begin{aligned} y &= y_h + y_p \\ y &= (K_1 e^{-0.670845x} + K_2 e^{-2.329156x}) + 0.161006 \sin x + 0.0301887 \cos x \end{aligned}$$

Constants  $K_1$  and  $K_2$  can be determined using initial conditions. From  $y(0) = 5$ ,

$$y(0) = K_1 + K_2 + 0.0301887 = 5$$

$$K_1 + K_2 = 5 - 0.0301887 = 4.969811$$

Now

$$\begin{aligned} \frac{dy}{dx} &= -0.670845K_1 e^{-(0.670845)x} - 2.329156K_2 e^{-(2.329156)x} \\ &\quad + 0.161006 \cos x - 0.0301887 \sin x \end{aligned}$$

$$\text{From } \frac{dy}{dx}(x=0) = 3$$

$$-0.670845K_1 - 2.329156K_2 + 0.161006 = 3$$

$$0.670845K_1 + 2.329156K_2 = -3 + 0.161006$$

$$0.670845K_1 + 2.329156K_2 = -2.838994$$

We have two linear equations with two unknowns

$$K_1 + K_2 = 4.969811$$

$$0.670845K_1 + 2.329156K_2 = -2.838994$$

Solving the above two simultaneous linear equations, we get

$$K_1 = 8.692253$$

$$K_2 = -3.722442$$

The complete solution is

$$y = (8.692253e^{-0.670845x} - 3.722442e^{-2.329156x}) + 0.161006\sin x + 0.0301887\cos x.$$

**Sample Case 3**

When

$$X = e^{ax} \sin bx \text{ or } e^{ax} \cos bx,$$

the particular part of the solution is of the form

$$e^{ax}(A\sin bx + B\cos bx),$$

we can get  $A$  and  $B$  by substituting

$$y = e^{ax}(A\sin bx + B\cos bx)$$

in the left hand side of differential equation and equating coefficients.

**Example 5**

Solve

$$2\frac{d^2y}{dx^2} + 5\frac{dy}{dx} + 3.125y = e^{-x} \sin x, y(0) = 5, \frac{dy}{dx}(x=0) = 3$$

**Solution**

The homogeneous equation is given by

$$(2D^2 + 5D + 3.125)y = 0$$

The characteristic equation is given by

$$\begin{aligned} 2r^2 + 5r + 3.125 &= 0 \\ r &= \frac{-5 \pm \sqrt{5^2 - 4(2)(3.125)}}{2(2)} \\ &= \frac{-5 \pm \sqrt{25 - 25}}{4} \\ &= \frac{-5 \pm 0}{4} \\ &= -1.25, -1.25 \end{aligned}$$

Since roots are repeated, the homogeneous solution  $y_H$  is given by

$$y_H = (K_1 + K_2x)e^{(-1.25)x}$$

The particular part of the solution is of the form

$$y_p = e^{-x}(A\sin x + B\cos x)$$

Substituting the particular part of the solution in the ordinary differential equation

$$\begin{aligned}
& 2 \frac{d^2}{dx^2} \{e^{-x}(A \sin x + B \cos x)\} + 5 \frac{d}{dx} \{e^{-x}(A \sin x + B \cos x)\} \\
& + 3.125 \{e^{-x}(A \sin x + B \cos x)\} = e^{-x} \sin x \\
& 2 \frac{d}{dx} \{-e^{-x}(A \sin x + B \cos x) + e^{-x}(A \cos x - B \sin x)\} \\
& + 5 \{-e^{-x}(A \sin x + B \cos x) + e^{-x}(A \cos x - B \sin x)\} + 3.125e^{-x}(A \sin x + B \cos x) = e^{-x} \sin x \\
& 2 \{e^{-x}(A \sin x + B \cos x) - e^{-x}(A \cos x - B \sin x) - e^{-x}(A \cos x - B \sin x) - e^{-x}(A \sin x + B \cos x)\} \\
& + 5 \{-e^{-x}(A \sin x + B \cos x) + e^{-x}(A \cos x - B \sin x)\} + 3.125e^{-x}(A \sin x + B \cos x) = e^{-x} \sin x \\
& -1.875e^{-x}(A \sin x + B \cos x) + e^{-x}(A \cos x - B \sin x) = e^{-x} \sin x \\
& -1.875(A \sin x + B \cos x) + (A \cos x - B \sin x) = \sin x \\
& -(1.875A + B) \sin x + (A - 1.875B) \cos x = \sin x
\end{aligned}$$

Equating coefficients of  $\cos x$  and  $\sin x$  on both sides we get

$$A - 1.875B = 0$$

$$1.875A + B = -1$$

Solving the above two simultaneous linear equations we get

$$A = -0.415224 \text{ and}$$

$$B = -0.221453$$

Hence,

$$y_P = -e^{-x}(0.415224 \sin x + 0.221453 \cos x)$$

Therefore complete solution is given by

$$y = y_H + y_P$$

$$y = (K_1 + xK_2)e^{-1.25x} - e^{-x}(0.415224 \sin x + 0.221453 \cos x)$$

Constants  $K_1$  and  $K_2$  can be determined using initial conditions,

From  $y(0) = 5$ , we get

$$K_1 - 0.221453 = 5$$

$$K_1 = 5.221453$$

Now

$$\begin{aligned}
\frac{dy}{dx} &= -1.25K_1e^{-1.25x} - 1.25K_2xe^{-1.25x} + K_2e^{-1.25x} - \\
& e^{-x}(0.415224 \cos x - 0.221453 \sin x) + e^{-x}(0.415224 \sin x + 0.221453 \cos x)
\end{aligned}$$

From  $\frac{dy}{dx}(0) = 3$ , we get

$$\begin{aligned}
& -1.25K_1e^{-1.25(0)} - 1.25K_2(0)e^{-1.25(0)} + K_2e^{-1.25(0)} \\
& - e^0(0.415224 \cos(0) - 0.221453 \sin(0)) + e^0(0.415224 \sin(0) + 0.221453 \cos(0)) = 3 \\
& -1.25K_1 + K_2 + 0.221453 - 0.415224 = 3 \\
& -1.25K_1 + K_2 = 3.193771 \\
& -1.25(5.221453) + K_2 = 3.193771 \\
& K_2 = 9.720582
\end{aligned}$$

Substituting

$$K_1 = 5.221453 \text{ and}$$

$$K_2 = 9.720582$$

in the solution, we get

$$y = (5.221453 + 9.720582x)e^{-1.25x} - e^{-x}(0.415224 \sin x + 0.221453 \cos x)$$

The forms of the particular part of the solution for different right hand sides of ordinary differential equations are given below

$X$	$y_p(x)$
$a_0 + a_1x + a_2x^2$	$b_0 + b_1x + b_2x^2$
$e^{ax}$	$Ae^{ax}$
$\sin(bx)$	$A \sin(bx) + B \cos(bx)$
$e^{ax} \sin(bx)$	$e^{ax}(A \sin(bx) + B \cos(bx))$
$\cos(bx)$	$A \sin(bx) + B \cos(bx)$
$e^{ax} \cos(bx)$	$e^{ax}(A \sin(bx) + B \cos(bx))$

### Laplace Transforms

If  $y = f(x)$  is defined at all positive values of  $x$ , the Laplace transform denoted by  $Y(s)$  is given by

$$Y(s) = L\{f(x)\} = \int_0^\infty e^{-sx} f(x) dx \quad (35)$$

where  $s$  is a parameter, which can be a real or complex number. We can get back  $f(x)$  by taking the inverse Laplace transform of  $Y(s)$ .

$$L^{-1}\{Y(s)\} = f(x) \quad (36)$$

Laplace transforms are very useful in solving differential equations. They give the solution directly without the necessity of evaluating arbitrary constants separately.

The following are Laplace transforms of some elementary functions

$$L(1) = \frac{1}{s}$$

$$L(x^n) = \frac{n!}{s^{n+1}}, \text{ where } n = 0, 1, 2, 3, \dots$$

$$L(e^{ax}) = \frac{1}{s-a}$$

$$L(\sin ax) = \frac{a}{s^2 + a^2}$$

$$L(\cos ax) = \frac{s}{s^2 + a^2}$$

$$L(\sinh ax) = \frac{a}{s^2 - a^2}$$

$$L(\cosh ax) = \frac{s}{s^2 - a^2} \quad (37)$$

The following are the inverse Laplace transforms of some common functions

$$L^{-1}\left(\frac{1}{s}\right) = 1$$

$$L^{-1}\left(\frac{1}{s-a}\right) = e^{ax}$$

$$L^{-1}\left(\frac{1}{s^n}\right) = \frac{x^{n-1}}{(n-1)!}, \text{ where } n=1,2,3,\dots$$

$$L^{-1}\left(\frac{1}{(s-a)^n}\right) = \frac{e^{ax}x^{n-1}}{(n-1)!}$$

$$L^{-1}\left(\frac{1}{s^2 + a^2}\right) = \frac{1}{a} \sin ax$$

$$L^{-1}\left(\frac{s}{s^2 + a^2}\right) = \cos ax$$

$$L^{-1}\left(\frac{1}{s^2 - a^2}\right) = \frac{1}{a} \sinh ax$$

$$L^{-1}\left(\frac{s}{s^2 - a^2}\right) = \cosh at$$

$$L^{-1}\left(\frac{1}{(s-a)^2 + b^2}\right) = \frac{1}{b} e^{ax} \sin bx$$

$$L^{-1}\left(\frac{s-a}{(s-a)^2 + b^2}\right) = e^{ax} \cos bx$$

$$L^{-1}\left(\frac{s}{(s^2 + a^2)^2}\right) = \frac{1}{2a} x \sin ax \quad (38)$$

## Properties of Laplace transforms

### Linear property

If  $a, b, c$  are constants and  $f(x), g(x)$ , and  $h(x)$  are functions of  $x$  then

$$L[af(x) + bg(x) + ch(x)] = aL(f(x)) + bL(g(x)) + cL(h(x)) \quad (39)$$

### Shifting property

If

$$L\{f(x)\} = Y(s) \quad (40)$$

then

$$L\{e^{at} f(x)\} = Y(s-a) \quad (41)$$

Using shifting property we get

$$\begin{aligned}
 L(e^{ax}x^n) &= \frac{n!}{(s-a)^{n+1}}, \quad n \geq 0 \\
 L(e^{ax}\sin bx) &= \frac{b}{(s-a)^2 + b^2} \\
 L(e^{ax}\cos bx) &= \frac{s-a}{(s-a)^2 + b^2} \\
 L(e^{ax}\sinh bx) &= \frac{b}{(s-a)^2 - b^2} \\
 L(e^{ax}\cosh bx) &= \frac{s-a}{(s-a)^2 - b^2}
 \end{aligned} \tag{42}$$

Scaling property

If

$$L\{f(x)\} = Y(s) \tag{43}$$

then

$$L\{f(ax)\} = \frac{1}{a} Y\left(\frac{s}{a}\right) \tag{44}$$

**Laplace transforms of derivatives**If the first  $n$  derivatives of  $f(x)$  are continuous then

$$L\{f^n(x)\} = \int_0^\infty e^{-sx} f^n(x) dx \tag{45}$$

Using integration by parts we get

$$\begin{aligned}
 \int_0^\infty e^{-sx} f^n(x) dx &= \left[ e^{-sx} f^{n-1}(x) - (-s)e^{-sx} f^{n-2}(x) \right]_0^\infty \\
 &\quad + (-s)^2 e^{-sx} f^{n-3}(x) + \dots + (-1)^{n-1} (-s)^{n-1} e^{-sx} f(x) \\
 &= -f^{n-1}(0) - sf^{n-2}(0) - s^2 f^{n-3}(0) - \dots - s^{n-1} f(0) + s^n \int_0^\infty e^{-sx} f(x) dx \\
 &= s^n Y(s) - f^{n-1}(0) - sf^{n-2}(0) - s^2 f^{n-3}(0) - \dots - s^{n-1} f(0)
 \end{aligned} \tag{46}$$

**Laplace transform technique to solve ordinary differential equations**

The following are steps to solve ordinary differential equations using the Laplace transform method

- (A) Take the Laplace transform of both sides of ordinary differential equations.
- (B) Express  $Y(s)$  as a function of  $s$ .
- (C) Take the inverse Laplace transform on both sides to get the solution.

Let us solve Examples 1 through 4 using the Laplace transform method.

**Example 6**

Solve

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

**Solution**

Taking the Laplace transform of both sides, we get

$$L\left(3\frac{dy}{dx} + 2y\right) = L(e^{-x})$$

$$3[sY(s) - y(0)] + 2Y(s) = \frac{1}{s+1}$$

Using the initial condition,  $y(0) = 5$  we get

$$3[sY(s) - 5] + 2Y(s) = \frac{1}{s+1}$$

$$(3s+2)Y(s) = \frac{1}{s+1} + 15$$

$$(3s+2)Y(s) = \frac{15s+16}{s+1}$$

$$Y(s) = \frac{15s+16}{(s+1)(3s+2)}$$

Writing the expression for  $Y(s)$  in terms of partial fractions

$$\frac{15s+16}{(s+1)(3s+2)} = \frac{A}{s+1} + \frac{B}{3s+2}$$

$$\frac{15s+16}{(s+1)(3s+2)} = \frac{3As+2A+Bs+B}{(s+1)(3s+2)}$$

$$15s+16 = 3As+2A+Bs+B$$

Equating coefficients of  $s^1$  and  $s^0$  gives

$$3A+B=15$$

$$2A+B=16$$

The solution to the above two simultaneous linear equations is

$$A=-1$$

$$B=18$$

$$\begin{aligned} Y(s) &= \frac{-1}{s+1} + \frac{18}{3s+2} \\ &= \frac{-1}{s+1} + \frac{6}{s+0.666667} \end{aligned}$$

Taking the inverse Laplace transform on both sides

$$L^{-1}\{Y(s)\} = L^{-1}\left(\frac{-1}{s+1}\right) + L^{-1}\left(\frac{6}{s+0.666667}\right)$$

Since

$$L^{-1}\left(\frac{1}{s+a}\right) = e^{-at}$$

The solution is given by

$$y(x) = -e^{-x} + 6e^{-0.666667x}$$

### Example 7

Solve

$$2\frac{dy}{dx} + 3y = e^{-1.5x}, \quad y(0) = 5$$

### Solution

Taking the Laplace transform of both sides, we get

$$L\left(2\frac{dy}{dx} + 3y\right) = L(e^{-1.5x})$$

$$2[sY(s) - y(0)] + 3Y(s) = \frac{1}{s+1.5}$$

Using the initial condition  $y(0) = 5$ , we get

$$2[sY(s) - 5] + 3Y(s) = \frac{1}{s+1.5}$$

$$(2s+3)Y(s) = \frac{1}{s+1.5} + 10$$

$$(2s+3)Y(s) = \frac{10s+16}{s+1.5}$$

$$Y(s) = \frac{10s+16}{(s+1.5)(2s+3)}$$

$$= \frac{10s+16}{2(s+1.5)(s+1.5)}$$

$$= \frac{10s+16}{2(s+1.5)^2}$$

$$= \frac{5s+8}{(s+1.5)^2}$$

Writing the expression for  $Y(s)$  in terms of partial fractions

$$\frac{5s+8}{(s+1.5)^2} = \frac{A}{s+1.5} + \frac{B}{(s+1.5)^2}$$

$$\frac{5s+8}{(s+1.5)^2} = \frac{As+1.5A+B}{(s+1.5)^2}$$

$$5s+8 = As+1.5A+B$$

Equating coefficients of  $s^1$  and  $s^0$  gives

$$A = 5$$

$$1.5A + B = 8$$

The solution to the above two simultaneous linear equations is

$$\begin{aligned}A &= 5 \\B &= 0.5 \\Y(s) &= \frac{5}{s+1.5} + \frac{0.5}{(s+1.5)^2}\end{aligned}$$

Taking the inverse Laplace transform on both sides

$$L^{-1}\{Y(s)\} = L^{-1}\left(\frac{5}{s+1.5}\right) + L^{-1}\left(\frac{0.5}{(s+1.5)^2}\right)$$

Since

$$L^{-1}\left(\frac{1}{s+a}\right) = e^{-ax} \text{ and } L^{-1}\left(\frac{1}{(s+a)^2}\right) = xe^{-ax}$$

The solution is given by

$$y(x) = 5e^{-1.5x} + 0.5xe^{-1.5x}$$

### Example 8

Solve

$$2\frac{d^2y}{dx^2} + 3\frac{dy}{dx} + 3.125y = \sin x, \quad y(0) = 5, \quad \frac{dy}{dx}(x=0) = 3$$

### Solution

Taking the Laplace transform of both sides

$$L\left(2\frac{d^2y}{dx^2} + 3\frac{dy}{dx} + 3.125y\right) = L(\sin x)$$

and knowing

$$L\left(\frac{d^2y}{dx^2}\right) = s^2Y(s) - sy(0) - \frac{dy}{dx}(x=0)$$

$$L\left(\frac{dy}{dx}\right) = sY(s) - y(0)$$

$$L(\sin x) = \frac{1}{s^2 + 1}$$

we get

$$2\left[s^2Y(s) - sy(0) - \frac{dy}{dx}(x=0)\right] + 3[sY(s) - y(0)] + 3.125Y(s) = \frac{1}{s^2 + 1}$$

$$2[s^2Y(s) - 5s - 3] + 3[sY(s) - 5] + 3.125Y(s) = \frac{1}{s^2 + 1}$$

$$[s(2s+3) + 3.125]Y(s) - 10s - 21 = \frac{1}{s^2 + 1}$$

$$[s(2s+3) + 3.125]Y(s) = \frac{1}{s^2 + 1} + 10s + 21$$

$$[2s^2 + 3s + 3.125]Y(s) = \frac{22 + 10s^3 + 10s + 21s^2}{(s^2 + 1)}$$

$$Y(s) = \frac{10s^3 + 21s^2 + 10s + 22}{(s^2 + 1)(2s^2 + 3s + 3.125)}$$

Writing the expression for  $Y(s)$  in terms of partial fractions

$$\begin{aligned} \frac{As + B}{(2s^2 + 3s + 3.125)} + \frac{Cs + D}{(s^2 + 1)} &= \frac{10s^3 + 21s^2 + 10s + 22}{(s^2 + 1)(2s^2 + 3s + 3.125)} \\ \frac{As^3 + As + Bs^2 + B + 2Cs^3 + 3Cs^2 + 3.125Cs + 2Ds^2 + 3Ds + 3.125D}{(2s^2 + 3s + 3.125)(s^2 + 1)} \\ &= \frac{10s^3 + 21s^2 + 10s + 22}{(s^2 + 1)(2s^2 + 3s + 3.125)} \\ \frac{(A + 2C)s^3 + (B + 3C + 2D)s^2 + (A + 3.125C + 3D)s + (B + 3.125D)}{(s^2 + 1)(2s^2 + 3s + 3.125)} \\ &= \frac{10s^3 + 21s^2 + 10s + 22}{(s^2 + 1)(2s^2 + 3s + 3.125)} \end{aligned}$$

Equating terms of  $s^3$ ,  $s^2$ ,  $s^1$  and  $s^0$  gives

$$A + 2C = 10$$

$$B + 3C + 2D = 21$$

$$A + 3.125C + 3D = 10$$

$$B + 3.125D = 22$$

The solution to the above four simultaneous linear equations is

$$A = 10.584474$$

$$B = 21.657534$$

$$C = -0.292237$$

$$D = 0.109589$$

Hence

$$\begin{aligned} Y(s) &= \frac{10.584474s + 21.657534}{2s^2 + 3s + 3.125} + \frac{-0.292237s + 0.109589}{s^2 + 1} \\ (2s^2 + 3s + 3.125) &= 2\{(s^2 + 1.5s + 0.5625) + 1\} = 2\{(s + 0.75)^2 + 1\} \\ Y(s) &= \frac{10.584474(s + 0.75) + 13.719179}{2\{(s + 0.75)^2 + 1\}} + \frac{-0.292237s + 0.109589}{s^2 + 1} \\ &= \frac{5.292237(s + 0.75)}{\{(s + 0.75)^2 + 1\}} + \frac{6.859589}{\{(s + 0.75)^2 + 1\}} - \frac{0.292237s}{(s^2 + 1)} + \frac{0.109589}{(s^2 + 1)} \end{aligned}$$

Taking the inverse Laplace transform of both sides

$$\begin{aligned} L^{-1}\{Y(s)\} &= L^{-1}\left(\frac{5.292237(s + 0.75)}{\{(s + 0.75)^2 + 1\}}\right) + L^{-1}\left(\frac{6.859589}{\{(s + 0.75)^2 + 1\}}\right) \\ &\quad - L^{-1}\left(\frac{0.292237s}{s^2 + 1}\right) + L^{-1}\left(\frac{0.109589}{s^2 + 1}\right) \end{aligned}$$

$$\begin{aligned} L^{-1}\{Y(s)\} &= 5.292237L^{-1}\left(\frac{s+0.75}{\{(s+0.75)^2+1\}}\right) + 6.859589L^{-1}\left(\frac{1}{\{(s+0.75)^2+1\}}\right) \\ &\quad - 0.292237L^{-1}\left(\frac{s}{s^2+1}\right) + 0.109589L^{-1}\left(\frac{1}{s^2+1}\right) \end{aligned}$$

Since

$$\begin{aligned} L^{-1}\left(\frac{s+a}{(s+a)^2+b^2}\right) &= e^{-ax} \cos bx \\ L^{-1}\left(\frac{b}{(s+a)^2+b^2}\right) &= e^{-ax} \sin bx \\ L^{-1}\left(\frac{1}{s^2+a^2}\right) &= \sin ax \\ L^{-1}\left(\frac{s}{s^2+a^2}\right) &= \cos ax \end{aligned}$$

The complete solution is

$$\begin{aligned} y(x) &= 5.292237e^{-0.75x} \cos x + 6.859589e^{-0.75x} \sin x \\ &\quad - 0.292237 \cos x + 0.109589 \sin x \\ &= e^{-0.75x}(5.292237 \cos x + 6.859589 \sin x) - 0.292237 \cos x + 0.109589 \sin x \end{aligned}$$

### Example 9

Solve

$$2\frac{d^2y}{dx^2} + 6\frac{dy}{dx} + 3.125y = \cos x, \quad y(0) = 5, \quad \frac{dy}{dx}(x=0) = 3$$

### Solution

Taking the Laplace transform of both sides

$$L\left(2\frac{d^2y}{dx^2} + 6\frac{dy}{dx} + 3.125y\right) = L(\cos x)$$

and knowing

$$\begin{aligned} L\left(\frac{d^2y}{dx^2}\right) &= s^2Y(s) - sy(0) - \frac{dy}{dx}(x=0) \\ L\left(\frac{dy}{dx}\right) &= sY(s) - y(0) \end{aligned}$$

$$L(\cos x) = \frac{s}{s^2+1}$$

we get

$$\begin{aligned} 2\left[s^2Y(s) - sy(0) - \frac{dy}{dx}(x=0)\right] + 6[sY(s) - y(0)] + 3.125Y(s) &= \frac{s}{s^2+1} \\ 2[s^2Y(s) - 5s - 3] + 6[sY(s) - 5] + 3.125Y(s) &= \frac{s}{s^2+1} \end{aligned}$$

$$\begin{aligned}[s(2s+6)+3.125]Y(s) &= \frac{s}{s^2+1} + 10s + 36 \\ [2s^2 + 6s + 3.125]Y(s) &= \frac{36 + 10s^3 + 11s + 36s^2}{s^2+1} \\ Y(s) &= \frac{10s^3 + 36s^2 + 11s + 36}{(s^2+1)(2s^2 + 6s + 3.125)}\end{aligned}$$

Writing the expression for  $Y(s)$  in terms of partial fractions

$$\begin{aligned}\frac{As+B}{(2s^2 + 6s + 3.125)} + \frac{Cs+D}{(s^2+1)} &= \frac{10s^3 + 36s^2 + 11s + 36}{(s^2+1)(2s^2 + 6s + 3.125)} \\ \frac{As^3 + As + Bs^2 + B + 2Cs^3 + 6Cs^2 + 3.125Cs + 2Ds^2 + 6Ds + 3.125D}{(2s^2 + 6s + 3.125)(s^2+1)} \\ &= \frac{10s^3 + 36s^2 + 11s + 36}{(s^2+1)(2s^2 + 6s + 3.125)} \\ \frac{(A+2C)s^3 + (B+6C+2D)s^2 + (A+3.125C+6D)s + (B+3.125D)}{(s^2+1)(2s^2 + 6s + 3.125)} \\ &= \frac{10s^3 + 36s^2 + 11s + 36}{(s^2+1)(2s^2 + 6s + 3.125)}\end{aligned}$$

Equating terms of  $s^3, s^2, s^1$  and  $s^0$  gives

$$\begin{aligned}A + 2C &= 10 \\ B + 6C + 2D &= 36 \\ A + 3.125C + 6D &= 11 \\ B + 3.125D &= 36\end{aligned}$$

The solution to the above four simultaneous linear equations is

$$\begin{aligned}A &= 9.939622 \\ B &= 35.496855 \\ C &= 0.0301886 \\ D &= 0.161006\end{aligned}$$

Then

$$\begin{aligned}Y(s) &= \frac{9.939622s + 35.496855}{2s^2 + 6s + 3.125} + \frac{0.0301886s + 0.161006}{s^2 + 1} \\ (2s^2 + 6s + 3.125) &= 2\{(s^2 + 3s + 2.25) - 0.6875\} = 2\{(s+1.5)^2 - 0.829156^2\} \\ Y(s) &= \frac{9.939622(s+1.5) + 20.587422}{2\{(s+1.5)^2 - 0.829156^2\}} + \frac{0.0301886s + 0.161006}{s^2 + 1} \\ &= \frac{4.969811(s+1.5)}{\{(s+1.5)^2 - 0.829156^2\}} + \frac{10.293711}{\{(s+1.5)^2 - 0.829156^2\}} \\ &\quad + \frac{0.0301886s}{s^2 + 1} + \frac{0.161006}{s^2 + 1}\end{aligned}$$

Taking the inverse Laplace transform on both sides

$$\begin{aligned}
L^{-1}\{Y(s)\} &= L^{-1}\left(\frac{4.969811(s+1.5)}{\{(s+1.5)^2 - 0.829156^2\}}\right) + L^{-1}\left(\frac{10.293711}{\{(s+1.5)^2 - 0.829156^2\}}\right) \\
&\quad + L^{-1}\left(\frac{0.0301886s}{s^2 + 1}\right) + L^{-1}\left(\frac{0.161006}{s^2 + 1}\right) \\
&= 4.969811L^{-1}\left(\frac{(s+1.5)}{(s+1.5)^2 - 0.829156^2}\right) + 10.293711L^{-1}\left(\frac{1}{(s+1.5)^2 - 0.829156^2}\right) \\
&\quad + 0.0301886L^{-1}\left(\frac{s}{(s^2 + 1)}\right) + 0.161006L^{-1}\left(\frac{1}{(s^2 + 1)}\right)
\end{aligned}$$

Since

$$\begin{aligned}
L^{-1}\left(\frac{s+a}{(s+a)^2 - b^2}\right) &= e^{-ax} \cosh bx \\
L^{-1}\left(\frac{1}{(s+a)^2 - b^2}\right) &= \frac{1}{b} e^{-ax} \sinh bx \\
L^{-1}\left(\frac{1}{s^2 + a^2}\right) &= \frac{1}{a} \sin ax \\
L^{-1}\left(\frac{s}{s^2 + a^2}\right) &= \cos ax
\end{aligned}$$

The complete solution is

$$\begin{aligned}
y(x) &= 4.969811e^{-1.5x} \cosh(0.829156x) + \frac{10.293711}{0.829156} e^{-1.5x} \sinh(0.829156x) \\
&\quad + 0.0301886 \cos x + 0.161006 \sin x \\
&= e^{-1.5x} \left( 4.969811 \left( \frac{e^{0.829156x} + e^{-0.829156x}}{2} \right) + 12.414685 \left( \frac{e^{0.829156x} - e^{-0.829156x}}{2} \right) \right) \\
&\quad + 0.0301886 \cos x + 0.161006 \sin x \\
&= e^{-1.5x} (8.692248e^{0.829156x} - 3.722437e^{-0.829156x}) + 0.0301886 \cos x \\
&\quad + 0.161006 \sin x
\end{aligned}$$

### Example 10

Solve

$$2\frac{d^2y}{dx^2} + 5\frac{dy}{dx} + 3.125y = e^{-x} \sin x, \quad y(0) = 5, \quad \frac{dy}{dx}(x=0) = 3$$

### Solution

Taking the Laplace transform of both sides

$$L\left(2\frac{d^2y}{dx^2} + 5\frac{dy}{dx} + 3.125y\right) = L(e^{-x} \sin x)$$

knowing

$$L\left(\frac{d^2y}{dx^2}\right) = s^2Y(s) - sy(0) - \frac{dy}{dx}(x=0)$$

$$L\left(\frac{dy}{dx}\right) = sY(s) - y(0)$$

$$L(e^{-x} \sin x) = \frac{1}{(s+1)^2 + 1}$$

we get

$$2\left[s^2Y(s) - sy(0) - \frac{dy}{dx}(x=0)\right] + 5[sY(s) - y(0)] + 3.125Y(s) = \frac{1}{(s+1)^2 + 1}$$

$$2[s^2Y(s) - 5s - 3] + 5[sY(s) - 5] + 3.125Y(s) = \frac{1}{(s+1)^2 + 1}$$

$$[s(2s+5) + 3.125]Y(s) - 10s - 31 = \frac{1}{(s+1)^2 + 1}$$

$$[s(2s+5) + 3.125]Y(s) = \frac{1}{(s+1)^2 + 1} + 10s + 31$$

$$[2s^2 + 5s + 3.125]Y(s) = \frac{63 + 10s^3 + 82s + 51s^2}{s^2 + 2s + 2}$$

$$Y(s) = \frac{10s^3 + 51s^2 + 82s + 63}{(s^2 + 2s + 2)(2s^2 + 5s + 3.125)}$$

Writing the expression for  $Y(s)$  in terms of partial fractions

$$\begin{aligned} & \frac{As+B}{2s^2 + 5s + 3.125} + \frac{Cs+D}{s^2 + 2s + 2} = \frac{10s^3 + 51s^2 + 82s + 63}{(s^2 + 2s + 2)(2s^2 + 5s + 3.125)} \\ & \frac{2Cs^3 + 5Cs^2 + 3.125Cs + 2Ds^2 + 5Ds + 3.125D + As^3 + 2As^2 + 2As + Bs^2 + 2Bs + 2B}{(2s^2 + 5s + 3.125)(s^2 + 2s + 2)} \\ & = \frac{10s^3 + 51s^2 + 82s + 63}{(s^2 + 2s + 2)(2s^2 + 5s + 3.125)} \end{aligned}$$

$$\begin{aligned} & \frac{(2C+A)s^3 + (5C+2D+2A+B)s^2 + (3.125C+5D+2A+2B)s + (3.125D+2B)}{(s^2 + 2s + 2)(2s^2 + 5s + 3.125)} \\ & = \frac{10s^3 + 51s^2 + 82s + 63}{(s^2 + 2s + 2)(2s^2 + 5s + 3.125)} \end{aligned}$$

Equating terms of  $s^3, s^2, s^1$  and  $s^0$  gives four simultaneous linear equations

$$2C + A = 10$$

$$5C + 2D + 2A + B = 51$$

$$3.125C + 5D + 2A + 2B = 82$$

$$3.125D + 2B = 63$$

The solution to the above four simultaneous linear equations is

$$A = 10.442906$$

$$B = 32.494809$$

$$C = -0.221453$$

$$D = -0.636678$$

Then

$$\begin{aligned} Y(s) &= \frac{10.442906s + 32.494809}{2s^2 + 5s + 3.125} + \frac{-0.221453s - 0.636678}{s^2 + 2s + 2} \\ (2s^2 + 5s + 3.125) &= 2\{(s^2 + 2.5s + 1.5625)\} = 2(s + 1.25)^2 \\ Y(s) &= \frac{10.442906(s + 1.25) + 19.441176}{2(s + 1.25)^2} + \frac{-0.221453(s + 1) - 0.415225}{(s + 1)^2 + 1} \\ &= \frac{5.221453(s + 1.25)}{(s + 1.25)^2} + \frac{9.720588}{(s + 1.25)^2} - \frac{0.221453(s + 1)}{(s + 1)^2 + 1} - \frac{0.415225}{(s + 1)^2 + 1} \end{aligned}$$

Taking the inverse Laplace transform on both sides

$$\begin{aligned} L^{-1}\{Y(s)\} &= L^{-1}\left(\frac{5.221453}{(s + 1.25)}\right) + L^{-1}\left(\frac{9.720588}{(s + 1.25)^2}\right) \\ &\quad - L^{-1}\left(\frac{0.221453(s + 1)}{(s + 1)^2 + 1}\right) - L^{-1}\left(\frac{0.415225}{(s + 1)^2 + 1}\right) \\ &= 5.221453L^{-1}\left(\frac{1}{(s + 1.25)}\right) + 9.720588L^{-1}\left(\frac{1}{(s + 1.25)^2}\right) \\ &\quad - 0.221453L^{-1}\left(\frac{(s + 1)}{(s + 1)^2 + 1}\right) - 0.415225L^{-1}\left(\frac{1}{(s + 1)^2 + 1}\right) \end{aligned}$$

Since

$$L^{-1}\left(\frac{s + a}{(s + a)^2 + b^2}\right) = e^{-ax} \cos bx$$

$$L^{-1}\left(\frac{b}{(s + a)^2 + b^2}\right) = e^{-ax} \sin bx$$

$$L^{-1}\left(\frac{1}{s + a}\right) = e^{-ax}$$

$$L^{-1}\left(\frac{1}{(s + a)^n}\right) = \frac{e^{-ax} x^{n-1}}{(n-1)!}$$

The complete solution is

$$\begin{aligned} y(x) &= 5.221453e^{-1.25x} + 9.720588e^{-1.25x}x - 0.221453e^{-x} \cos x \\ &\quad - 0.415225e^{-x} \sin x \\ &= e^{-1.25x}(5.221453 + 9.720588x) + e^x(-0.221453 \cos x - 0.415225 \sin x) \end{aligned}$$

---

## ORDINARY DIFFERENTIAL EQUATIONS

---

Topic	A Primer on ordinary differential equations
Summary	Textbook notes of a primer on solution of ordinary differential equations
Major	All majors of engineering
Authors	Autar Kaw, Praveen Chalasani
Date	April 24, 2009
Web Site	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

---

# Chapter 08.02

## Euler's Method for Ordinary Differential Equations

After reading this chapter, you should be able to:

1. develop Euler's Method for solving ordinary differential equations,
2. determine how the step size affects the accuracy of a solution,
3. derive Euler's formula from Taylor series, and
4. use Euler's method to find approximate values of integrals.

### What is Euler's method?

Euler's method is a numerical technique to solve ordinary differential equations of the form

$$\frac{dy}{dx} = f(x, y), y(0) = y_0 \quad (1)$$

So only first order ordinary differential equations can be solved by using Euler's method. In another chapter we will discuss how Euler's method is used to solve higher order ordinary differential equations or coupled (simultaneous) differential equations. How does one write a first order differential equation in the above form?

### Example 1

Rewrite

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

in

$$\frac{dy}{dx} = f(x, y), y(0) = y_0 \text{ form.}$$

### Solution

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

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

In this case

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

### Example 2

Rewrite

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

in

$$\frac{dy}{dx} = f(x, y), \quad y(0) = y_0 \text{ form.}$$

### Solution

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

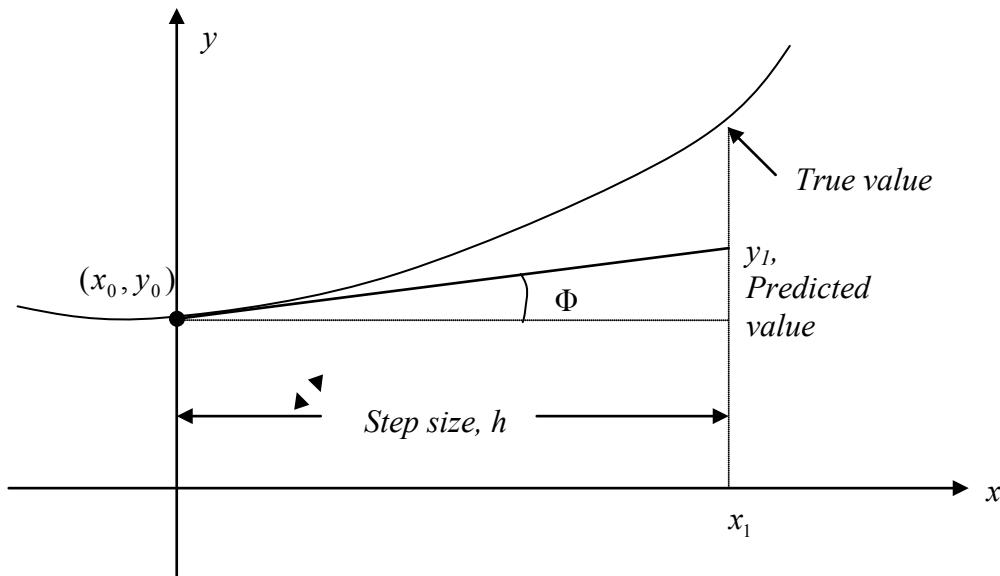
$$\frac{dy}{dx} = \frac{2 \sin(3x) - x^2 y^2}{e^y}, \quad y(0) = 5$$

In this case

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

### Derivation of Euler's method

At  $x = 0$ , we are given the value of  $y = y_0$ . Let us call  $x = 0$  as  $x_0$ . Now since we know the slope of  $y$  with respect to  $x$ , that is,  $f(x, y)$ , then at  $x = x_0$ , the slope is  $f(x_0, y_0)$ . Both  $x_0$  and  $y_0$  are known from the initial condition  $y(x_0) = y_0$ .



**Figure 1** Graphical interpretation of the first step of Euler's method.

So the slope at  $x = x_0$  as shown in Figure 1 is

$$\begin{aligned}\text{Slope} &= \frac{\text{Rise}}{\text{Run}} \\ &= \frac{y_1 - y_0}{x_1 - x_0} \\ &= f(x_0, y_0)\end{aligned}$$

From here

$$y_1 = y_0 + f(x_0, y_0)(x_1 - x_0)$$

Calling  $x_1 - x_0$  the step size  $h$ , we get

$$y_1 = y_0 + f(x_0, y_0)h \quad (2)$$

One can now use the value of  $y_1$  (an approximate value of  $y$  at  $x = x_1$ ) to calculate  $y_2$ , and that would be the predicted value at  $x_2$ , given by

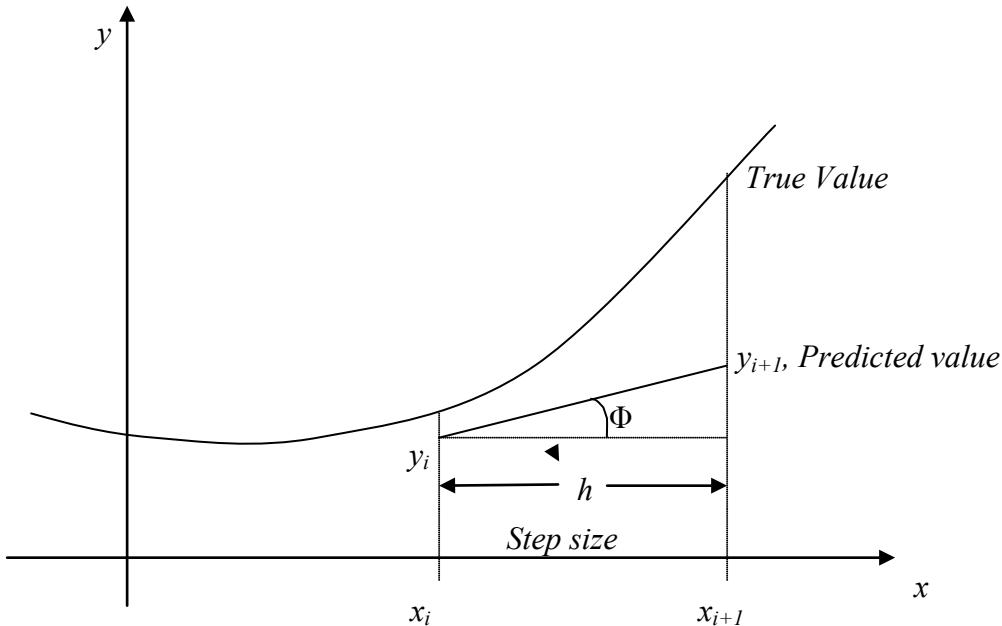
$$y_2 = y_1 + f(x_1, y_1)h$$

$$x_2 = x_1 + h$$

Based on the above equations, if we now know the value of  $y = y_i$  at  $x_i$ , then

$$y_{i+1} = y_i + f(x_i, y_i)h \quad (3)$$

This formula is known as Euler's method and is illustrated graphically in Figure 2. In some books, it is also called the Euler-Cauchy method.



**Figure 2** General graphical interpretation of Euler's method.

**Example 3**

A ball at 1200K is allowed to cool down in air at an ambient temperature of 300K. Assuming heat is lost only due to radiation, the differential equation for the temperature of the ball is given by

$$\frac{d\theta}{dt} = -2.2067 \times 10^{-12} (\theta^4 - 81 \times 10^8), \quad \theta(0) = 1200\text{K}$$

where  $\theta$  is in K and  $t$  in seconds. Find the temperature at  $t = 480$  seconds using Euler's method. Assume a step size of  $h = 240$  seconds.

**Solution**

$$\begin{aligned}\frac{d\theta}{dt} &= -2.2067 \times 10^{-12} (\theta^4 - 81 \times 10^8) \\ f(t, \theta) &= -2.2067 \times 10^{-12} (\theta^4 - 81 \times 10^8)\end{aligned}$$

Per Equation (3), Euler's method reduces to

$$\theta_{i+1} = \theta_i + f(t_i, \theta_i)h$$

For  $i = 0$ ,  $t_0 = 0$ ,  $\theta_0 = 1200$

$$\begin{aligned}\theta_1 &= \theta_0 + f(t_0, \theta_0)h \\ &= 1200 + f(0, 1200) \times 240 \\ &= 1200 + (-2.2067 \times 10^{-12} (1200^4 - 81 \times 10^8)) \times 240 \\ &= 1200 + (-4.5579) \times 240 \\ &= 106.09\text{K}\end{aligned}$$

$\theta_1$  is the approximate temperature at

$$t = t_1 = t_0 + h = 0 + 240 = 240$$

$$\theta_1 = \theta(240) \approx 106.09\text{K}$$

For  $i = 1$ ,  $t_1 = 240$ ,  $\theta_1 = 106.09$

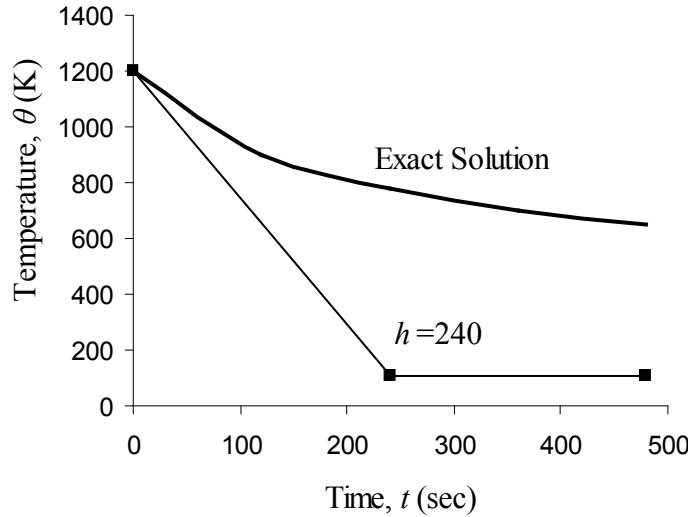
$$\begin{aligned}\theta_2 &= \theta_1 + f(t_1, \theta_1)h \\ &= 106.09 + f(240, 106.09) \times 240 \\ &= 106.09 + (-2.2067 \times 10^{-12} (106.09^4 - 81 \times 10^8)) \times 240 \\ &= 106.09 + (0.017595) \times 240 \\ &= 110.32\text{K}\end{aligned}$$

$\theta_2$  is the approximate temperature at

$$t = t_2 = t_1 + h = 240 + 240 = 480$$

$$\theta_2 = \theta(480) \approx 110.32\text{K}$$

Figure 3 compares the exact solution with the numerical solution from Euler's method for the step size of  $h = 240$ .



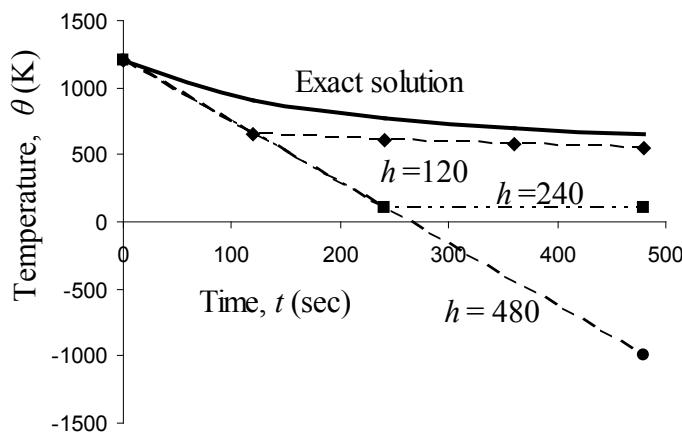
**Figure 3** Comparing the exact solution and Euler's method.

The problem was solved again using a smaller step size. The results are given below in Table 1.

**Table 1** Temperature at 480 seconds as a function of step size,  $h$ .

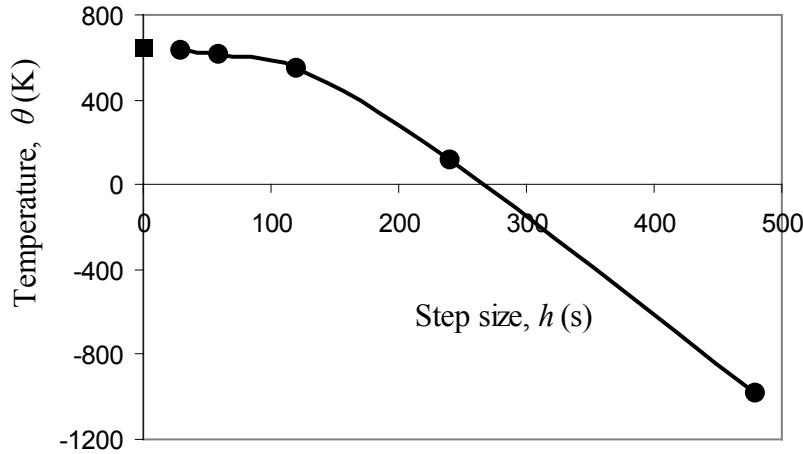
Step size, $h$	$\theta(480)$	$E_t$	$ e_t  \%$
480	-987.81	1635.4	252.54
240	110.32	537.26	82.964
120	546.77	100.80	15.566
60	614.97	32.607	5.0352
30	632.77	14.806	2.2864

Figure 4 shows how the temperature varies as a function of time for different step sizes.



**Figure 4** Comparison of Euler's method with the exact solution for different step sizes.

The values of the calculated temperature at  $t = 480$  s as a function of step size are plotted in Figure 5.



**Figure 5** Effect of step size in Euler's method.

The exact solution of the ordinary differential equation is given by the solution of a nonlinear equation as

$$0.92593 \ln \frac{\theta - 300}{\theta + 300} - 1.8519 \tan^{-1}(0.333 \times 10^{-2} \theta) = -0.22067 \times 10^{-3} t - 2.9282 \quad (4)$$

The solution to this nonlinear equation is

$$\theta = 647.57 \text{ K}$$

It can be seen that Euler's method has large errors. This can be illustrated using the Taylor series.

$$y_{i+1} = y_i + \left. \frac{dy}{dx} \right|_{x_i, y_i} (x_{i+1} - x_i) + \frac{1}{2!} \left. \frac{d^2 y}{dx^2} \right|_{x_i, y_i} (x_{i+1} - x_i)^2 + \frac{1}{3!} \left. \frac{d^3 y}{dx^3} \right|_{x_i, y_i} (x_{i+1} - x_i)^3 + \dots \quad (5)$$

$$= y_i + f(x_i, y_i)(x_{i+1} - x_i) + \frac{1}{2!} f'(x_i, y_i)(x_{i+1} - x_i)^2 + \frac{1}{3!} f''(x_i, y_i)(x_{i+1} - x_i)^3 + \dots \quad (6)$$

As you can see the first two terms of the Taylor series

$$y_{i+1} = y_i + f(x_i, y_i)h$$

are Euler's method.

The true error in the approximation is given by

$$E_t = \frac{f'(x_i, y_i)}{2!} h^2 + \frac{f''(x_i, y_i)}{3!} h^3 + \dots \quad (7)$$

The true error hence is approximately proportional to the square of the step size, that is, as the step size is halved, the true error gets approximately quartered. However from Table 1, we see that as the step size gets halved, the true error only gets approximately halved. This is because the true error, being proportioned to the square of the step size, is the local truncation

error, that is, error from one point to the next. The global truncation error is however proportional only to the step size as the error keeps propagating from one point to another.

**Can one solve a definite integral using numerical methods such as Euler's method of solving ordinary differential equations?**

Let us suppose you want to find the integral of a function  $f(x)$

$$I = \int_a^b f(x)dx .$$

Both fundamental theorems of calculus would be used to set up the problem so as to solve it as an ordinary differential equation.

The first fundamental theorem of calculus states that if  $f$  is a continuous function in the interval  $[a,b]$ , and  $F$  is the antiderivative of  $f$ , then

$$\int_a^b f(x)dx = F(b) - F(a)$$

The second fundamental theorem of calculus states that if  $f$  is a continuous function in the open interval  $D$ , and  $a$  is a point in the interval  $D$ , and if

$$F(x) = \int_a^x f(t)dt$$

then

$$F'(x) = f(x)$$

at each point in  $D$ .

Asked to find  $\int_a^b f(x)dx$ , we can rewrite the integral as the solution of an ordinary differential equation (here is where we are using the second fundamental theorem of calculus)

$$\frac{dy}{dx} = f(x), \quad y(a) = 0,$$

where then  $y(b)$  (here is where we are using the first fundamental theorem of calculus) will

give the value of the integral  $\int_a^b f(x)dx$ .

#### Example 4

Find an approximate value of

$$\int_5^8 6x^3 dx$$

using Euler's method of solving an ordinary differential equation. Use a step size of  $h = 1.5$ .

#### Solution

Given  $\int_5^8 6x^3 dx$ , we can rewrite the integral as the solution of an ordinary differential equation

$$\frac{dy}{dx} = 6x^3, \quad y(5) = 0$$

where  $y(8)$  will give the value of the integral  $\int_5^8 6x^3 dx$ .

$$\frac{dy}{dx} = 6x^3 = f(x, y), \quad y(5) = 0$$

The Euler's method equation is

$$y_{i+1} = y_i + f(x_i, y_i)h$$

### Step 1

$$i = 0, \quad x_0 = 5, \quad y_0 = 0$$

$$h = 1.5$$

$$x_1 = x_0 + h$$

$$= 5 + 1.5$$

$$= 6.5$$

$$y_1 = y_0 + f(x_0, y_0)h$$

$$= 0 + f(5, 0) \times 1.5$$

$$= 0 + (6 \times 5^3) \times 1.5$$

$$= 1125$$

$$\approx y(6.5)$$

### Step 2

$$i = 1, \quad x_1 = 6.5, \quad y_1 = 1125$$

$$x_2 = x_1 + h$$

$$= 6.5 + 1.5$$

$$= 8$$

$$y_2 = y_1 + f(x_1, y_1)h$$

$$= 1125 + f(6.5, 1125) \times 1.5$$

$$= 1125 + (6 \times 6.5^3) \times 1.5$$

$$= 3596.625$$

$$\approx y(8)$$

Hence

$$\int_5^8 6x^3 dx = y(8) - y(5)$$

$$\approx 3596.625 - 0$$

$$= 3596.625$$

---

**ORDINARY DIFFERENTIAL EQUATIONS**

---

Topic	Euler's Method for ordinary differential equations
Summary	Textbook notes on Euler's method for solving ordinary differential equations
Major	General Engineering
Authors	Autar Kaw
Last Revised	October 13, 2010
Web Site	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

---

# **Chapter 08.03**

## **Runge-Kutta 2nd Order Method for Ordinary Differential Equations**

*After reading this chapter, you should be able to:*

1. *understand the Runge-Kutta 2nd order method for ordinary differential equations and how to use it to solve problems.*

### **What is the Runge-Kutta 2nd order method?**

The Runge-Kutta 2nd order method is a numerical technique used to solve an ordinary differential equation of the form

$$\frac{dy}{dx} = f(x, y), y(0) = y_0$$

Only first order ordinary differential equations can be solved by using the Runge-Kutta 2nd order method. In other sections, we will discuss how the Euler and Runge-Kutta methods are used to solve higher order ordinary differential equations or coupled (simultaneous) differential equations.

How does one write a first order differential equation in the above form?

### **Example 1**

Rewrite

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

in

$$\frac{dy}{dx} = f(x, y), \quad y(0) = y_0 \text{ form.}$$

### **Solution**

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

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

In this case

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

**Example 2**

Rewrite

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

in

$$\frac{dy}{dx} = f(x, y), \quad y(0) = y_0 \text{ form.}$$

**Solution**

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

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

In this case

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

**Runge-Kutta 2<sup>nd</sup> order method**

Euler's method is given by

$$y_{i+1} = y_i + f(x_i, y_i)h \quad (1)$$

where

$$x_0 = 0$$

$$y_0 = y(x_0)$$

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

To understand the Runge-Kutta 2nd order method, we need to derive Euler's method from the Taylor series.

$$\begin{aligned} y_{i+1} &= y_i + \left. \frac{dy}{dx} \right|_{x_i, y_i} (x_{i+1} - x_i) + \frac{1}{2!} \left. \frac{d^2 y}{dx^2} \right|_{x_i, y_i} (x_{i+1} - x_i)^2 + \frac{1}{3!} \left. \frac{d^3 y}{dx^3} \right|_{x_i, y_i} (x_{i+1} - x_i)^3 + \dots \\ &= y_i + f(x_i, y_i)(x_{i+1} - x_i) + \frac{1}{2!} f'(x_i, y_i)(x_{i+1} - x_i)^2 + \frac{1}{3!} f''(x_i, y_i)(x_{i+1} - x_i)^3 + \dots \quad (2) \end{aligned}$$

As you can see the first two terms of the Taylor series

$$y_{i+1} = y_i + f(x_i, y_i)h$$

are Euler's method and hence can be considered to be the Runge-Kutta 1st order method.

The true error in the approximation is given by

$$E_t = \frac{f'(x_i, y_i)}{2!} h^2 + \frac{f''(x_i, y_i)}{3!} h^3 + \dots \quad (3)$$

So what would a 2nd order method formula look like. It would include one more term of the Taylor series as follows.

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

Let us take a generic example of a first order ordinary differential equation

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

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

Now since  $y$  is a function of  $x$ ,

$$\begin{aligned} f'(x, y) &= \frac{\partial f(x, y)}{\partial x} + \frac{\partial f(x, y)}{\partial y} \frac{dy}{dx} \\ &= \frac{\partial}{\partial x}(e^{-2x} - 3y) + \frac{\partial}{\partial y}[(e^{-2x} - 3y)](e^{-2x} - 3y) \\ &= -2e^{-2x} + (-3)(e^{-2x} - 3y) \\ &= -5e^{-2x} + 9y \end{aligned} \quad (5)$$

The 2nd order formula for the above example would be

$$\begin{aligned} y_{i+1} &= y_i + f(x_i, y_i)h + \frac{1}{2!}f'(x_i, y_i)h^2 \\ &= y_i + (e^{-2x_i} - 3y_i)h + \frac{1}{2!}(-5e^{-2x_i} + 9y_i)h^2 \end{aligned}$$

However, we already see the difficulty of having to find  $f'(x, y)$  in the above method. What Runge and Kutta did was write the 2nd order method as

$$y_{i+1} = y_i + (a_1 k_1 + a_2 k_2)h \quad (6)$$

where

$$\begin{aligned} k_1 &= f(x_i, y_i) \\ k_2 &= f(x_i + p_1 h, y_i + q_{11} k_1 h) \end{aligned} \quad (7)$$

This form allows one to take advantage of the 2nd order method without having to calculate  $f'(x, y)$ .

So how do we find the unknowns  $a_1$ ,  $a_2$ ,  $p_1$  and  $q_{11}$ . Without proof (see Appendix for proof), equating Equation (4) and (6), gives three equations.

$$a_1 + a_2 = 1$$

$$a_2 p_1 = \frac{1}{2}$$

$$a_2 q_{11} = \frac{1}{2}$$

Since we have 3 equations and 4 unknowns, we can assume the value of one of the unknowns. The other three will then be determined from the three equations. Generally the value of  $a_2$  is chosen to evaluate the other three constants. The three values generally used for  $a_2$  are  $\frac{1}{2}$ , 1 and  $\frac{2}{3}$ , and are known as Heun's Method, the midpoint method and Ralston's method, respectively.

Heun's Method

Here  $a_2 = \frac{1}{2}$  is chosen, giving

$$a_1 = \frac{1}{2}$$

$$p_1 = 1$$

$$q_{11} = 1$$

resulting in

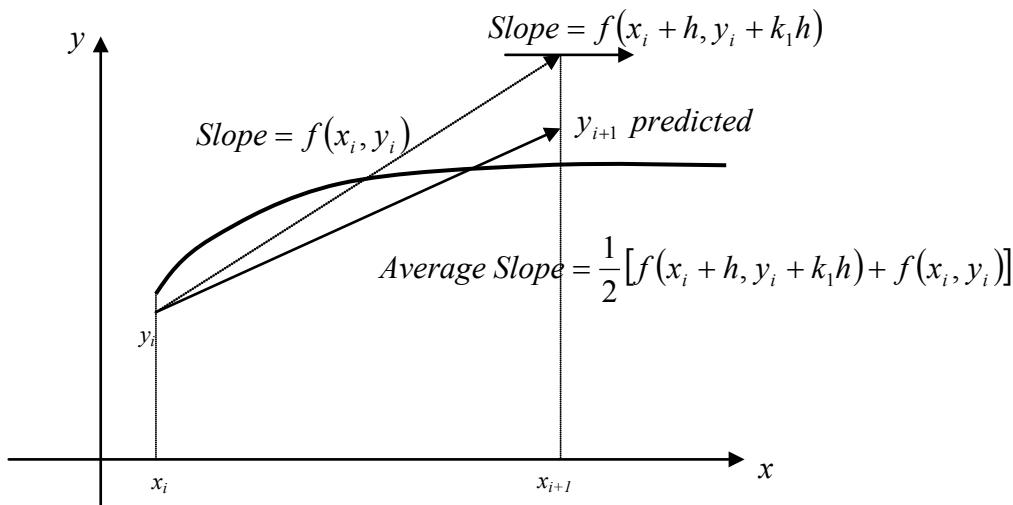
$$y_{i+1} = y_i + \left( \frac{1}{2}k_1 + \frac{1}{2}k_2 \right)h \quad (8)$$

where

$$k_1 = f(x_i, y_i) \quad (9a)$$

$$k_2 = f(x_i + h, y_i + k_1 h) \quad (9b)$$

This method is graphically explained in Figure 1.



**Figure 1** Runge-Kutta 2nd order method (Heun's method).

Midpoint Method

Here  $a_2 = 1$  is chosen, giving

$$a_1 = 0$$

$$p_1 = \frac{1}{2}$$

$$q_{11} = \frac{1}{2}$$

resulting in

$$y_{i+1} = y_i + k_2 h \quad (10)$$

where

$$k_1 = f(x_i, y_i) \quad (11a)$$

$$k_2 = f\left(x_i + \frac{1}{2}h, y_i + \frac{1}{2}k_1 h\right) \quad (11b)$$

Ralston's Method

Here  $a_2 = \frac{2}{3}$  is chosen, giving

$$a_1 = \frac{1}{3}$$

$$p_1 = \frac{3}{4}$$

$$q_{11} = \frac{3}{4}$$

resulting in

$$y_{i+1} = y_i + \left( \frac{1}{3}k_1 + \frac{2}{3}k_2 \right)h \quad (12)$$

where

$$k_1 = f(x_i, y_i) \quad (13a)$$

$$k_2 = f\left(x_i + \frac{3}{4}h, y_i + \frac{3}{4}k_1 h\right) \quad (13b)$$

**Example 3**

A ball at 1200K is allowed to cool down in air at an ambient temperature of 300K. Assuming heat is lost only due to radiation, the differential equation for the temperature of the ball is given by

$$\frac{d\theta}{dt} = -2.2067 \times 10^{-12} (\theta^4 - 81 \times 10^8)$$

where  $\theta$  is in K and  $t$  in seconds. Find the temperature at  $t = 480$  seconds using Runge-Kutta 2nd order method. Assume a step size of  $h = 240$  seconds.

**Solution**

$$\frac{d\theta}{dt} = -2.2067 \times 10^{-12} (\theta^4 - 81 \times 10^8)$$

$$f(t, \theta) = -2.2067 \times 10^{-12} (\theta^4 - 81 \times 10^8)$$

Per Heun's method given by Equations (8) and (9)

$$\theta_{i+1} = \theta_i + \left( \frac{1}{2}k_1 + \frac{1}{2}k_2 \right)h$$

$$k_1 = f(t_i, \theta_i)$$

$$k_2 = f(t_i + h, \theta_i + k_1 h)$$

$$i = 0, t_0 = 0, \theta_0 = \theta(0) = 1200$$

$$k_1 = f(t_0, \theta_0)$$

$$\begin{aligned}
&= f(0, 1200) \\
&= -2.2067 \times 10^{-12} (1200^4 - 81 \times 10^8) \\
&= -4.5579 \\
k_2 &= f(t_0 + h, \theta_0 + k_1 h) \\
&= f(0 + 240, 1200 + (-4.5579)240) \\
&= f(240, 106.09) \\
&= -2.2067 \times 10^{-12} (106.09^4 - 81 \times 10^8) \\
&= 0.017595 \\
\theta_1 &= \theta_0 + \left( \frac{1}{2} k_1 + \frac{1}{2} k_2 \right) h \\
&= 1200 + \left( \frac{1}{2}(-4.5579) + \frac{1}{2}(0.017595) \right) 240 \\
&= 1200 + (-2.2702)240 \\
&= 655.16 \text{ K} \\
i = 1, t_1 &= t_0 + h = 0 + 240 = 240, \theta_1 = 655.16 \text{ K} \\
k_1 &= f(t_1, \theta_1) \\
&= f(240, 655.16) \\
&= -2.2067 \times 10^{-12} (655.16^4 - 81 \times 10^8) \\
&= -0.38869 \\
k_2 &= f(t_1 + h, \theta_1 + k_1 h) \\
&= f(240 + 240, 655.16 + (-0.38869)240) \\
&= f(480, 561.87) \\
&= -2.2067 \times 10^{-12} (561.87^4 - 81 \times 10^8) \\
&= -0.20206 \\
\theta_2 &= \theta_1 + \left( \frac{1}{2} k_1 + \frac{1}{2} k_2 \right) h \\
&= 655.16 + \left( \frac{1}{2}(-0.38869) + \frac{1}{2}(-0.20206) \right) 240 \\
&= 655.16 + (-0.29538)240 \\
&= 584.27 \text{ K} \\
\theta_2 &= \theta(480) = 584.27 \text{ K}
\end{aligned}$$

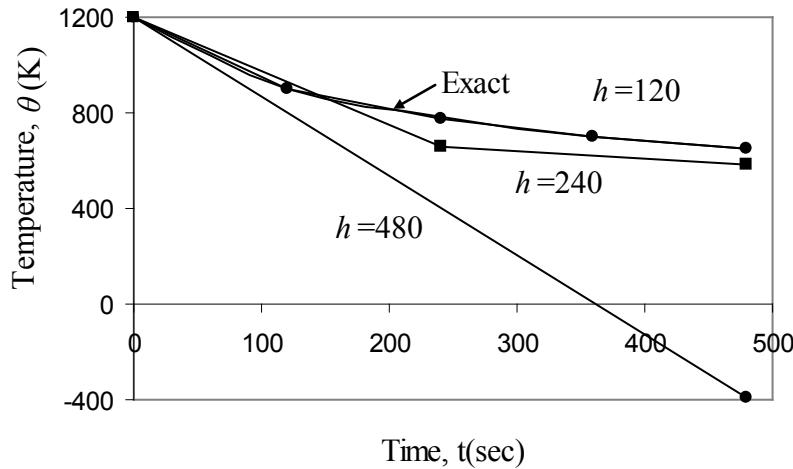
The results from Heun's method are compared with exact results in Figure 2.

The exact solution of the ordinary differential equation is given by the solution of a nonlinear equation as

$$0.92593 \ln \frac{\theta - 300}{\theta + 300} - 1.8519 \tan^{-1}(0.0033333\theta) = -0.22067 \times 10^{-3}t - 2.9282$$

The solution to this nonlinear equation at  $t = 480$  s is

$$\theta(480) = 647.57 \text{ K}$$

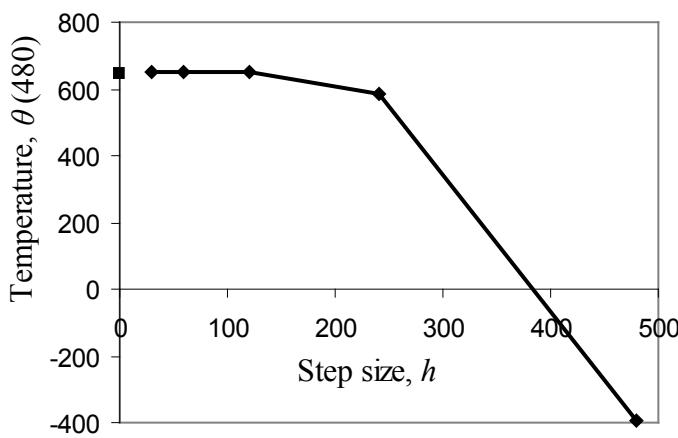


**Figure 2** Heun's method results for different step sizes.

Using a smaller step size would increase the accuracy of the result as given in Table 1 and Figure 3 below.

**Table 1** Effect of step size for Heun's method

Step size, $h$	$\theta(480)$	$E_t$	$ \epsilon_t  \%$
480	-393.87	1041.4	160.82
240	584.27	63.304	9.7756
120	651.35	-3.7762	0.58313
60	649.91	-2.3406	0.36145
30	648.21	-0.63219	0.097625



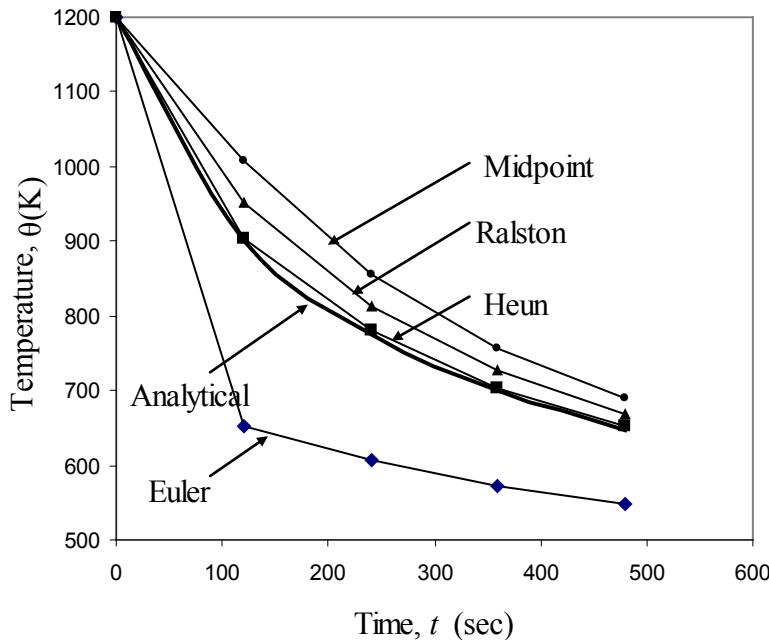
**Figure 3** Effect of step size in Heun's method.

In Table 2, Euler's method and the Runge-Kutta 2nd order method results are shown as a function of step size,

**Table 2** Comparison of Euler and the Runge-Kutta methods

Step size, $h$	$\theta(480)$			
	Euler	Heun	Midpoint	Ralston
480	-987.84	-393.87	1208.4	449.78
240	110.32	584.27	976.87	690.01
120	546.77	651.35	690.20	667.71
60	614.97	649.91	654.85	652.25
30	632.77	648.21	649.02	648.61

while in Figure 4, the comparison is shown over the range of time.



**Figure 4** Comparison of Euler and Runge Kutta methods with exact results over time.

**How do these three methods compare with results obtained if we found  $f'(x, y)$  directly?**

Of course, we know that since we are including the first three terms in the series, if the solution is a polynomial of order two or less (that is, quadratic, linear or constant), any of the three methods are exact. But for any other case the results will be different.

Let us take the example of

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

If we directly find  $f'(x, y)$ , the first three terms of the Taylor series gives

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

where

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

$$f'(x, y) = -5e^{-2x} + 9y$$

For a step size of  $h = 0.2$ , using Heun's method, we find

$$y(0.6) = 1.0930$$

The exact solution

$$y(x) = e^{-2x} + 4e^{-3x}$$

gives

$$\begin{aligned} y(0.6) &= e^{-2(0.6)} + 4e^{-3(0.6)} \\ &= 0.96239 \end{aligned}$$

Then the absolute relative true error is

$$\begin{aligned} |\epsilon_t| &= \left| \frac{0.96239 - 1.0930}{0.96239} \right| \times 100 \\ &= 13.571\% \end{aligned}$$

For the same problem, the results from Euler's method and the three Runge-Kutta methods are given in Table 3.

**Table 3** Comparison of Euler's and Runge-Kutta 2nd order methods

	y(0.6)					
	Exact	Euler	Direct 2nd	Heun	Midpoint	Ralston
Value	0.96239	0.4955	1.0930	1.1012	1.0974	1.0994
$ \epsilon_t  \%$		48.514	13.571	14.423	14.029	14.236

## Appendix A

### How do we get the 2nd order Runge-Kutta method equations?

We wrote the 2nd order Runge-Kutta equations without proof to solve

$$\frac{dy}{dx} = f(x, y), \quad y(0) = y_0 \quad (\text{A.1})$$

as

$$y_{i+1} = y_i + (a_1 k_1 + a_2 k_2)h \quad (\text{A.2})$$

where

$$k_1 = f(x_i, y_i) \quad (\text{A.3a})$$

$$k_2 = f(x_i + p_1 h, y_i + q_1 k_1 h) \quad (\text{A.3b})$$

and

$$a_1 + a_2 = 1$$

$$a_2 p_2 = \frac{1}{2}$$

$$a_2 q_{11} = \frac{1}{2} \quad (\text{A.4})$$

The advantage of using 2nd order Runge-Kutta method equations is based on not having to find the derivative of  $f(x, y)$  symbolically in the ordinary differential equation

So how do we get the above three Equations (A.4)? This is the question that is answered in this Appendix.

Writing out the first three terms of Taylor series are

$$y_{i+1} = y_i + \left. \frac{dy}{dx} \right|_{x_i, y_i} h + \frac{1}{2!} \left. \frac{d^2 y}{dx^2} \right|_{x_i, y_i} h^2 + O(h^3) \quad (\text{A.5})$$

where

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

Since

$$\frac{dy}{dx} = f(x, y)$$

we can rewrite the Taylor series as

$$y_{i+1} = y_i + f(x_i, y_i)h + \frac{1}{2!} f'(x_i, y_i)h^2 + O(h^3) \quad (\text{A.6})$$

Now

$$f'(x, y) = \frac{\partial f(x, y)}{\partial x} + \frac{\partial f(x, y)}{\partial y} \frac{dy}{dx}. \quad (\text{A.7})$$

Hence

$$\begin{aligned} y_{i+1} &= y_i + f(x_i, y_i)h + \frac{1}{2!} \left( \left. \frac{\partial f}{\partial x} \right|_{x_i, y_i} + \left. \frac{\partial f}{\partial y} \right|_{x_i, y_i} \times \left. \frac{dy}{dx} \right|_{x_i, y_i} \right) h^2 + O(h^3) \\ &= y_i + f(x_i, y_i)h + \frac{1}{2} \left. \frac{\partial f}{\partial x} \right|_{x_i, y_i} h^2 + \frac{1}{2} \left. \frac{\partial f}{\partial y} \right|_{x_i, y_i} f(x_i, y_i)h^2 + O(h^3) \end{aligned} \quad (\text{A.8})$$

Now the term used in the Runge-Kutta 2nd order method for  $k_2$  can be written as a Taylor series of two variables with the first three terms as

$$\begin{aligned} k_2 &= f(x_i + p_1 h, y_i + q_{11} k_1 h) \\ &= f(x_i, y_i) + p_1 h \left. \frac{\partial f}{\partial x} \right|_{x_i, y_i} + q_{11} k_1 h \left. \frac{\partial f}{\partial y} \right|_{x_i, y_i} + O(h^2) \end{aligned} \quad (\text{A.9})$$

Hence

$$\begin{aligned} y_{i+1} &= y_i + (a_1 k_1 + a_2 k_2)h \\ &= y_i + \left( a_1 f(x_i, y_i) + a_2 \left\{ f(x_i, y_i) + p_1 h \left. \frac{\partial f}{\partial x} \right|_{x_i, y_i} + q_{11} k_1 h \left. \frac{\partial f}{\partial y} \right|_{x_i, y_i} + O(h^2) \right\} \right) h \\ &= y_i + (a_1 + a_2)h f(x_i, y_i) + a_2 p_1 h^2 \left. \frac{\partial f}{\partial x} \right|_{x_i, y_i} + a_2 q_{11} f(x_i, y_i)h^2 \left. \frac{\partial f}{\partial y} \right|_{x_i, y_i} + O(h^3) \end{aligned}$$

(A.10)

Equating the terms in Equation (A.8) and Equation (A.10), we get

$$a_1 + a_2 = 1$$

$$a_2 p_1 = \frac{1}{2}$$

$$a_2 q_{11} = \frac{1}{2}$$

---

#### ORDINARY DIFFERENTIAL EQUATIONS

---

Topic	Runge 2nd Order Method for Ordinary Differential Equations
Summary	Textbook notes on Runge 2nd order method for ODE
Major	General Engineering
Authors	Autar Kaw
Last Revised	October 13, 2010
Web Site	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

---

# **Chapter 08.04**

## **Runge-Kutta 4th Order Method for Ordinary Differential Equations**

*After reading this chapter, you should be able to*

1. *develop Runge-Kutta 4<sup>th</sup> order method for solving ordinary differential equations,*
2. *find the effect size of step size has on the solution,*
3. *know the formulas for other versions of the Runge-Kutta 4<sup>th</sup> order method*

### **What is the Runge-Kutta 4th order method?**

Runge-Kutta 4<sup>th</sup> order method is a numerical technique used to solve ordinary differential equation of the form

$$\frac{dy}{dx} = f(x, y), y(0) = y_0$$

So only first order ordinary differential equations can be solved by using the Runge-Kutta 4<sup>th</sup> order method. In other sections, we have discussed how Euler and Runge-Kutta methods are used to solve higher order ordinary differential equations or coupled (simultaneous) differential equations.

### **How does one write a first order differential equation in the above form?**

#### **Example 1**

Rewrite

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

in

$$\frac{dy}{dx} = f(x, y), \quad y(0) = y_0 \text{ form.}$$

**Solution**

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

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

In this case

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

**Example 2**

Rewrite

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

in

$$\frac{dy}{dx} = f(x, y), \quad y(0) = y_0 \text{ form.}$$

**Solution**

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

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

In this case

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

The Runge-Kutta 4<sup>th</sup> order method is based on the following

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

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

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

Equation (1) is equated to the first five terms of Taylor series

$$\begin{aligned} y_{i+1} &= y_i + \frac{dy}{dx} \Big|_{x_i, y_i} (x_{i+1} - x_i) + \frac{1}{2!} \frac{d^2 y}{dx^2} \Big|_{x_i, y_i} (x_{i+1} - x_i)^2 + \frac{1}{3!} \frac{d^3 y}{dx^3} \Big|_{x_i, y_i} (x_{i+1} - x_i)^3 \\ &\quad + \frac{1}{4!} \frac{d^4 y}{dx^4} \Big|_{x_i, y_i} (x_{i+1} - x_i)^4 \end{aligned} \quad (2)$$

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

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

Based on equating Equation (2) and Equation (3), one of the popular solutions used is

$$y_{i+1} = y_i + \frac{1}{6} (k_1 + 2k_2 + 2k_3 + k_4) h \quad (4)$$

$$k_1 = f(x_i, y_i) \quad (5a)$$

$$k_2 = f\left(x_i + \frac{1}{2}h, y_i + \frac{1}{2}k_1 h\right) \quad (5b)$$

$$k_3 = f\left(x_i + \frac{1}{2}h, y_i + \frac{1}{2}k_2 h\right) \quad (5c)$$

$$k_4 = f(x_i + h, y_i + k_3 h) \quad (5d)$$

### Example 3

A ball at 1200 K is allowed to cool down in air at an ambient temperature of 300 K. Assuming heat is lost only due to radiation, the differential equation for the temperature of the ball is given by

$$\frac{d\theta}{dt} = -2.2067 \times 10^{-12} (\theta^4 - 81 \times 10^8), \theta(0) = 1200 \text{ K}$$

where  $\theta$  is in K and  $t$  in seconds. Find the temperature at  $t = 480$  seconds using Runge-Kutta 4th order method. Assume a step size of  $h = 240$  seconds.

#### Solution

$$\frac{d\theta}{dt} = -2.2067 \times 10^{-12} (\theta^4 - 81 \times 10^8)$$

$$f(t, \theta) = -2.2067 \times 10^{-12} (\theta^4 - 81 \times 10^8)$$

$$\theta_{i+1} = \theta_i + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)h$$

For  $i = 0$ ,  $t_0 = 0$ ,  $\theta_0 = 1200 \text{ K}$

$$\begin{aligned} k_1 &= f(t_0, \theta_0) \\ &= f(0, 1200) \\ &= -2.2067 \times 10^{-12} (1200^4 - 81 \times 10^8) \\ &= -4.5579 \end{aligned}$$

$$\begin{aligned} k_2 &= f\left(t_0 + \frac{1}{2}h, \theta_0 + \frac{1}{2}k_1 h\right) \\ &= f\left(0 + \frac{1}{2}(240), 1200 + \frac{1}{2}(-4.5579) \times 240\right) \\ &= f(120, 653.05) \\ &= -2.2067 \times 10^{-12} (653.05^4 - 81 \times 10^8) \\ &= -0.38347 \end{aligned}$$

$$\begin{aligned} k_3 &= f\left(t_0 + \frac{1}{2}h, \theta_0 + \frac{1}{2}k_2 h\right) \\ &= f\left(0 + \frac{1}{2}(240), 1200 + \frac{1}{2}(-0.38347) \times 240\right) \\ &= f(120, 1154.0) \end{aligned}$$

$$\begin{aligned}
&= -2.2067 \times 10^{-12} (1154.0^4 - 81 \times 10^8) \\
&= -3.8954 \\
k_4 &= f(t_0 + h, \theta_0 + k_3 h) \\
&= f(0 + 240, 1200 + (-3.894) \times 240) \\
&= f(240, 265.10) \\
&= -2.2067 \times 10^{-12} (265.10^4 - 81 \times 10^8) \\
&= 0.0069750
\end{aligned}$$

$$\begin{aligned}
\theta_1 &= \theta_0 + \frac{1}{6} (k_1 + 2k_2 + 2k_3 + k_4)h \\
&= 1200 + \frac{1}{6} (-4.5579 + 2(-0.38347) + 2(-3.8954) + (0.069750))240 \\
&= 1200 + (-2.1848) \times 240 \\
&= 675.65 \text{ K}
\end{aligned}$$

$\theta_1$  is the approximate temperature at

$$\begin{aligned}
t &= t_1 \\
&= t_0 + h \\
&= 0 + 240 \\
&= 240 \\
\theta_1 &= \theta(240) \\
&\approx 675.65 \text{ K}
\end{aligned}$$

For  $i = 1, t_1 = 240, \theta_1 = 675.65 \text{ K}$

$$\begin{aligned}
k_1 &= f(t_1, \theta_1) \\
&= f(240, 675.65) \\
&= -2.2067 \times 10^{-12} (675.65^4 - 81 \times 10^8) \\
&= -0.44199 \\
k_2 &= f\left(t_1 + \frac{1}{2}h, \theta_1 + \frac{1}{2}k_1 h\right) \\
&= f\left(240 + \frac{1}{2}(240), 675.65 + \frac{1}{2}(-0.44199)240\right) \\
&= f(360, 622.61) \\
&= -2.2067 \times 10^{-12} (622.61^4 - 81 \times 10^8) \\
&= -0.31372 \\
k_3 &= f\left(t_1 + \frac{1}{2}h, \theta_1 + \frac{1}{2}k_2 h\right) \\
&= f\left(240 + \frac{1}{2}(240), 675.65 + \frac{1}{2}(-0.31372) \times 240\right) \\
&= f(360, 638.00) \\
&= -2.2067 \times 10^{-12} (638.00^4 - 81 \times 10^8)
\end{aligned}$$

$$\begin{aligned}
 &= -0.34775 \\
 k_4 &= f(t_1 + h, \theta_1 + k_3 h) \\
 &= f(240 + 240, 675.65 + (-0.34775) \times 240) \\
 &= f(480, 592.19) \\
 &= 2.2067 \times 10^{-12} (592.19^4 - 81 \times 10^8) \\
 &= -0.25351
 \end{aligned}$$

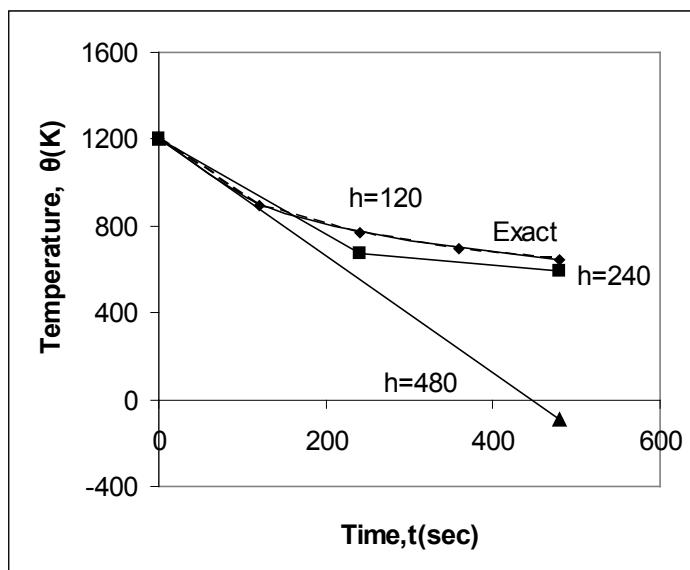
$$\begin{aligned}
 \theta_2 &= \theta_1 + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)h \\
 &= 675.65 + \frac{1}{6}(-0.44199 + 2(-0.31372) + 2(-0.34775) + (-0.25351)) \times 240 \\
 &= 675.65 + \frac{1}{6}(-2.0184) \times 240 \\
 &= 594.91 \text{ K}
 \end{aligned}$$

$\theta_2$  is the approximate temperature at

$$\begin{aligned}
 t &= t_2 \\
 &= t_1 + h \\
 &= 240 + 240 \\
 &= 480
 \end{aligned}$$

$$\begin{aligned}
 \theta_2 &= \theta(480) \\
 &\approx 594.91 \text{ K}
 \end{aligned}$$

Figure 1 compares the exact solution with the numerical solution using the Runge-Kutta 4th order method with different step sizes.

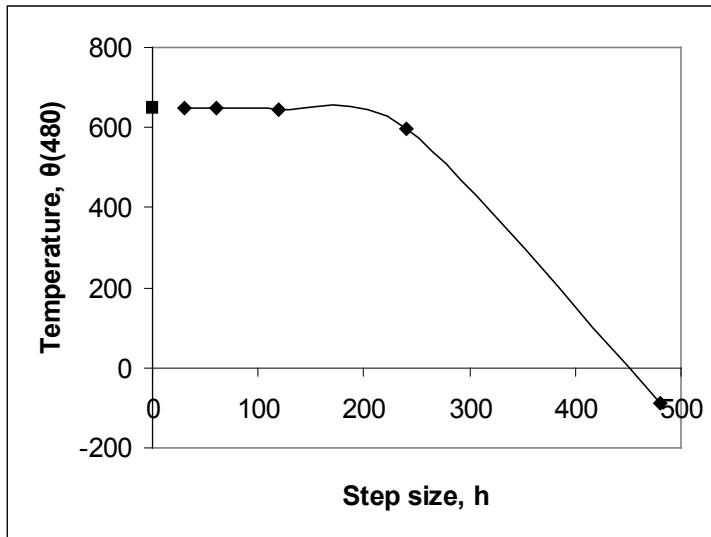


**Figure 1** Comparison of Runge-Kutta 4th order method with exact solution for different step sizes.

Table 1 and Figure 2 show the effect of step size on the value of the calculated temperature at  $t = 480$  seconds.

**Table 1** Value of temperature at time,  $t = 480$  s for different step sizes

Step size, $h$	$\theta(480)$	$E_t$	$ \varepsilon_t  \%$
480	-90.278	737.85	113.94
240	594.91	52.660	8.1319
120	646.16	1.4122	0.21807
60	647.54	0.033626	0.0051926
30	647.57	0.00086900	0.00013419



**Figure 2** Effect of step size in Runge-Kutta 4th order method.

In Figure 3, we are comparing the exact results with Euler's method (Runge-Kutta 1st order method), Heun's method (Runge-Kutta 2nd order method), and Runge-Kutta 4th order method.

The formula described in this chapter was developed by Runge. This formula is same as Simpson's 1/3 rule, if  $f(x, y)$  were only a function of  $x$ . There are other versions of the 4<sup>th</sup> order method just like there are several versions of the second order methods. The formula developed by Kutta is

$$y_{i+1} = y_i + \frac{1}{8}(k_1 + 3k_2 + 3k_3 + k_4)h \quad (6)$$

where

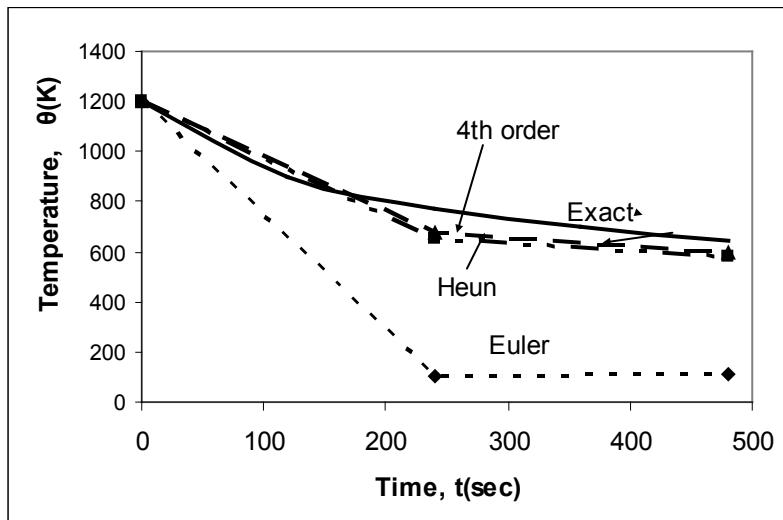
$$k_1 = f(x_i, y_i) \quad (7a)$$

$$k_2 = f\left(x_i + \frac{1}{3}h, y_i + \frac{1}{3}hk_1\right) \quad (7b)$$

$$k_3 = f\left(x_i + \frac{2}{3}h, y_i - \frac{1}{3}hk_1 + hk_2\right) \quad (7c)$$

$$k_4 = f(x_i + h, y_i + hk_1 - hk_2 + hk_3) \quad (7d)$$

This formula is the same as the Simpson's 3/8 rule, if  $f(x, y)$  is only a function of  $x$ .



**Figure 3** Comparison of Runge-Kutta methods of 1<sup>st</sup> (Euler), 2<sup>nd</sup>, and 4<sup>th</sup> order.

---

#### ORDINARY DIFFERENTIAL EQUATIONS

---

Topic	Runge-Kutta 4th order method
Summary	Textbook notes on the Runge-Kutta 4th order method for solving ordinary differential equations.
Major	General Engineering
Authors	Autar Kaw
Last Revised	October 13, 2010
Web Site	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

---

# Chapter 08.05

## On Solving Higher Order Equations for Ordinary Differential Equations

*After reading this chapter, you should be able to:*

1. solve higher order and coupled differential equations,

We have learned Euler's and Runge-Kutta methods to solve first order ordinary differential equations of the form

$$\frac{dy}{dx} = f(x, y), \quad y(0) = y_0 \quad (1)$$

What do we do to solve simultaneous (coupled) differential equations, or differential equations that are higher than first order? For example an  $n^{\text{th}}$  order differential equation of the form

$$a_n \frac{d^n y}{dx^n} + a_{n-1} \frac{d^{n-1} y}{dx^{n-1}} + \dots + a_1 \frac{dy}{dx} + a_0 y = f(x) \quad (2)$$

with  $n - 1$  initial conditions can be solved by assuming

$$y = z_1 \quad (3.1)$$

$$\frac{dy}{dx} = \frac{dz_1}{dx} = z_2 \quad (3.2)$$

$$\frac{d^2 y}{dx^2} = \frac{dz_2}{dx} = z_3 \quad (3.3)$$

⋮

$$\frac{d^{n-1} y}{dx^{n-1}} = \frac{dz_{n-1}}{dx} = z_n \quad (3.n)$$

$$\begin{aligned} \frac{d^n y}{dx^n} &= \frac{dz_n}{dx} \\ &= \frac{1}{a_n} \left( -a_{n-1} \frac{d^{n-1} y}{dx^{n-1}} - \dots - a_1 \frac{dy}{dx} - a_0 y + f(x) \right) \\ &= \frac{1}{a_n} (-a_{n-1} z_n - \dots - a_1 z_2 - a_0 z_1 + f(x)) \end{aligned} \quad (3.n+1)$$

The above Equations from (3.1) to (3.n+1) represent  $n$  first order differential equations as follows

$$\frac{dz_1}{dx} = z_2 = f_1(z_1, z_2, \dots, x) \quad (4.1)$$

$$\frac{dz_2}{dx} = z_3 = f_2(z_1, z_2, \dots, x) \quad (4.2)$$

⋮

$$\frac{dz_n}{dx} = \frac{1}{a_n}(-a_{n-1}z_n - \dots - a_1z_2 - a_0z_1 + f(x)) \quad (4.n)$$

Each of the  $n$  first order ordinary differential equations are accompanied by one initial condition. These first order ordinary differential equations are simultaneous in nature but can be solved by the methods used for solving first order ordinary differential equations that we have already learned.

### Example 1

Rewrite the following differential equation as a set of first order differential equations.

$$3\frac{d^2y}{dx^2} + 2\frac{dy}{dx} + 5y = e^{-x}, \quad y(0) = 5, \quad y'(0) = 7$$

#### Solution

The ordinary differential equation would be rewritten as follows. Assume

$$\frac{dy}{dx} = z,$$

Then

$$\frac{d^2y}{dx^2} = \frac{dz}{dx}$$

Substituting this in the given second order ordinary differential equation gives

$$3\frac{dz}{dx} + 2z + 5y = e^{-x}$$

$$\frac{dz}{dx} = \frac{1}{3}(e^{-x} - 2z - 5y)$$

The set of two simultaneous first order ordinary differential equations complete with the initial conditions then is

$$\frac{dy}{dx} = z, \quad y(0) = 5$$

$$\frac{dz}{dx} = \frac{1}{3}(e^{-x} - 2z - 5y), \quad z(0) = 7.$$

Now one can apply any of the numerical methods used for solving first order ordinary differential equations.

### Example 2

Given

$$\frac{d^2y}{dt^2} + 2\frac{dy}{dt} + y = e^{-t}, \quad y(0) = 1, \quad \frac{dy}{dt}(0) = 2, \quad \text{find by Euler's method}$$

a)  $y(0.75)$

- b) the absolute relative true error for part(a), if  $y(0.75)|_{exact} = 1.668$   
c)  $\frac{dy}{dt}(0.75)$

Use a step size of  $h = 0.25$ .

**Solution**

First, the second order differential equation is written as two simultaneous first-order differential equations as follows. Assume

$$\frac{dy}{dt} = z$$

then

$$\begin{aligned}\frac{dz}{dt} + 2z + y &= e^{-t} \\ \frac{dz}{dt} &= e^{-t} - 2z - y\end{aligned}$$

So the two simultaneous first order differential equations are

$$\frac{dy}{dt} = z = f_1(t, y, z), \quad y(0) = 1 \quad (\text{E2.1})$$

$$\frac{dz}{dt} = e^{-t} - 2z - y = f_2(t, y, z), \quad z(0) = 2 \quad (\text{E2.2})$$

Using Euler's method on Equations (E2.1) and (E2.2), we get

$$y_{i+1} = y_i + f_1(t_i, y_i, z_i)h \quad (\text{E2.3})$$

$$z_{i+1} = z_i + f_2(t_i, y_i, z_i)h \quad (\text{E2.4})$$

a) To find the value of  $y(0.75)$  and since we are using a step size of 0.25 and starting at  $t = 0$ , we need to take three steps to find the value of  $y(0.75)$ .

For  $i = 0, t_0 = 0, y_0 = 1, z_0 = 2$ ,

From Equation (E2.3)

$$\begin{aligned}y_1 &= y_0 + f_1(t_0, y_0, z_0)h \\ &= 1 + f_1(0, 1, 2)(0.25) \\ &= 1 + 2(0.25) \\ &= 1.5\end{aligned}$$

$y_1$  is the approximate value of  $y$  at

$$t = t_1 = t_0 + h = 0 + 0.25 = 0.25$$

$$y_1 = y(0.25) \approx 1.5$$

From Equation (E2.4)

$$\begin{aligned}z_1 &= z_0 + f_2(t_0, y_0, z_0)h \\ &= 2 + f_2(0, 1, 2)(0.25) \\ &= 2 + (e^{-0} - 2(2) - 1)(0.25) \\ &= 1\end{aligned}$$

$z_1$  is the approximate value of  $z$  (same as  $\frac{dy}{dt}$ ) at  $t = 0.25$

$$z_1 = z(0.25) \approx 1$$

For  $i = 1$ ,  $t_1 = 0.25$ ,  $y_1 = 1.5$ ,  $z_1 = 1$ ,

From Equation (E2.3)

$$\begin{aligned} y_2 &= y_1 + f_1(t_1, y_1, z_1)h \\ &= 1.5 + f_1(0.25, 1.5, 1)(0.25) \\ &= 1.5 + (1)(0.25) \\ &= 1.75 \end{aligned}$$

$y_2$  is the approximate value of  $y$  at

$$t = t_2 = t_1 + h = 0.25 + 0.25 = 0.50$$

$$y_2 = y(0.5) \approx 1.75$$

From Equation (E2.4)

$$\begin{aligned} z_2 &= z_1 + f_2(t_1, y_1, z_1)h \\ &= 1 + f_2(0.25, 1.5, 1)(0.25) \\ &= 1 + (e^{-0.25} - 2(1) - 1.5)(0.25) \\ &= 1 + (-2.7211)(0.25) \\ &= 0.31970 \end{aligned}$$

$z_2$  is the approximate value of  $z$  at

$$t = t_2 = 0.5$$

$$z_2 = z(0.5) \approx 0.31970$$

For  $i = 2$ ,  $t_2 = 0.5$ ,  $y_2 = 1.75$ ,  $z_2 = 0.31970$ ,

From Equation (E2.3)

$$\begin{aligned} y_3 &= y_2 + f_1(t_2, y_2, z_2)h \\ &= 1.75 + f_1(0.50, 1.75, 0.31970)(0.25) \\ &= 1.75 + (0.31970)(0.25) \\ &= 1.8299 \end{aligned}$$

$y_3$  is the approximate value of  $y$  at

$$t = t_3 = t_2 + h = 0.5 + 0.25 = 0.75$$

$$y_3 = y(0.75) \approx 1.8299$$

From Equation (E2.4)

$$\begin{aligned} z_3 &= z_2 + f_2(t_2, y_2, z_2)h \\ &= 0.31972 + f_2(0.50, 1.75, 0.31970)(0.25) \\ &= 0.31972 + (e^{-0.50} - 2(0.31970) - 1.75)(0.25) \\ &= 0.31972 + (-1.7829)(0.25) \\ &= -0.1260 \end{aligned}$$

$z_3$  is the approximate value of  $z$  at

$$t = t_3 = 0.75$$

$$z_3 = z(0.75) \approx -0.12601$$

$$y(0.75) \approx y_3 = 1.8299$$

b) The exact value of  $y(0.75)$  is

$$y(0.75)_{\text{exact}} = 1.668$$

The absolute relative true error in the result from part (a) is

$$\begin{aligned} |\epsilon_t| &= \left| \frac{1.668 - 1.8299}{1.668} \right| \times 100 \\ &= 9.7062\% \end{aligned}$$

c)  $\frac{dy}{dx}(0.75) = z_3 \approx -0.12601$

### Example 3

Given

$$\frac{d^2y}{dt^2} + 2 \frac{dy}{dt} + y = e^{-t}, y(0) = 1, \frac{dy}{dt}(0) = 2,$$

find by Heun's method

a)  $y(0.75)$

b)  $\frac{dy}{dx}(0.75)$ .

Use a step size of  $h = 0.25$ .

#### Solution

First, the second order differential equation is rewritten as two simultaneous first-order differential equations as follows. Assume

$$\frac{dy}{dt} = z$$

then

$$\frac{dz}{dt} + 2z + y = e^{-t}$$

$$\frac{dz}{dt} = e^{-t} - 2z - y$$

So the two simultaneous first order differential equations are

$$\frac{dy}{dt} = z = f_1(t, y, z), y(0) = 1 \quad (\text{E3.1})$$

$$\frac{dz}{dt} = e^{-t} - 2z - y = f_2(t, y, z), z(0) = 2 \quad (\text{E3.2})$$

Using Heun's method on Equations (1) and (2), we get

$$y_{i+1} = y_i + \frac{1}{2} (k_1^y + k_2^y) h \quad (\text{E3.3})$$

$$k_1^y = f_1(t_i, y_i, z_i) \quad (\text{E3.4a})$$

$$k_2^y = f_1(t_i + h, y_i + hk_1^y, z_i + hk_1^z) \quad (\text{E3.4b})$$

$$z_{i+1} = z_i + \frac{1}{2} (k_1^z + k_2^z) h \quad (\text{E3.5})$$

$$k_1^z = f_2(t_i, y_i, z_i) \quad (\text{E3.6a})$$

$$k_2^z = f_2(t_i + h, y_i + hk_1^y, z_i + hk_1^z) \quad (\text{E3.6b})$$

For  $i = 0, t_o = 0, y_o = 1, z_o = 2$

From Equation (E3.4a)

$$\begin{aligned} k_1^y &= f_1(t_o, y_o, z_o) \\ &= f_1(0, 1, 2) \\ &= 2 \end{aligned}$$

From Equation (E3.6a)

$$\begin{aligned} k_1^z &= f_2(t_o, y_o, z_o) \\ &= f_2(0, 1, 2) \\ &= e^{-0} - 2(2) - 1 \\ &= -4 \end{aligned}$$

From Equation (E3.4b)

$$\begin{aligned} k_2^y &= f_1(t_0 + h, y_0 + hk_1^y, z_0 + hk_1^z) \\ &= f_1(0 + 0.25, 1 + (0.25)(2), 2 + (0.25)(-4)) \\ &= f_1(0.25, 1.5, 1) \\ &= 1 \end{aligned}$$

From Equation (E3.6b)

$$\begin{aligned} k_2^z &= f_2(t_0 + h, y_0 + hk_1^y, z_0 + hk_1^z) \\ &= f_2(0 + 0.25, 1 + (0.25)(2), 2 + (0.25)(-4)) \\ &= f_2(0.25, 1.5, 1) \\ &= e^{-0.25} - 2(1) - 1.5 \\ &= -2.7212 \end{aligned}$$

From Equation (E3.3)

$$\begin{aligned} y_1 &= y_0 + \frac{1}{2}(k_1^y + k_2^y)h \\ &= 1 + \frac{1}{2}(2 + 1)(0.25) \\ &= 1.375 \end{aligned}$$

$y_1$  is the approximate value of  $y$  at

$$t = t_1 = t_0 + h = 0 + 0.25 = 0.25$$

$$y_1 = y(0.25) \approx 1.375$$

From Equation (E3.5)

$$\begin{aligned} z_1 &= z_0 + \frac{1}{2}(k_1^z + k_2^z)h \\ &= 2 + \frac{1}{2}(-4 + (-2.7212))(0.25) \\ &= 1.1598 \end{aligned}$$

$z_1$  is the approximate value of  $z$  at

$$t = t_1 = 0.25$$

$$z_1 = z(0.25) \approx 1.1598$$

For  $i = 1$ ,  $t_1 = 0.25$ ,  $y_1 = 1.375$ ,  $z_1 = 1.1598$

From Equation (E3.4a)

$$\begin{aligned} k_1^y &= f_1(t_1, y_1, z_1) \\ &= f_1(0.25, 1.375, 1.1598) \\ &= 1.1598 \end{aligned}$$

From Equation (E3.6a)

$$\begin{aligned} k_1^z &= f_2(t_1, y_1, z_1) \\ &= f_2(0.25, 1.375, 1.1598) \\ &= e^{-0.25} - 2(1.1598) - 1.375 \\ &= -2.9158 \end{aligned}$$

From Equation (E3.4b)

$$\begin{aligned} k_2^y &= f_1(t_1 + h, y_1 + hk_1^y, z_1 + hk_1^z) \\ &= f_1(0.25 + 0.25, 1.375 + (0.25)(1.1598), 1.1598 + (0.25)(-2.9158)) \\ &= f_1(0.50, 1.6649, 0.43087) \\ &= 0.43087 \end{aligned}$$

From Equation (E3.6b)

$$\begin{aligned} k_2^z &= f_2(t_1 + h, y_1 + hk_1^y, z_1 + hk_1^z) \\ &= f_2(0.25 + 0.25, 1.375 + (0.25)(1.1598), 1.1598 + (0.25)(-2.9158)) \\ &= f_2(0.50, 1.6649, 0.43087) \\ &= e^{-0.50} - 2(0.43087) - 1.6649 \\ &= -1.9201 \end{aligned}$$

From Equation (E3.3)

$$\begin{aligned} y_2 &= y_1 + \frac{1}{2}(k_1^y + k_2^y)h \\ &= 1.375 + \frac{1}{2}(1.1598 + 0.43087)(0.25) \\ &= 1.5738 \end{aligned}$$

$y_2$  is the approximate value of  $y$  at

$$t = t_2 = t_1 + h = 0.25 + 0.25 = 0.50$$

$$y_2 = y(0.50) \approx 1.5738$$

From Equation (E3.5)

$$\begin{aligned} z_2 &= z_1 + \frac{1}{2}(k_1^z + k_2^z)h \\ &= 1.1598 + \frac{1}{2}(-2.9158 + (-1.9201))(0.25) \\ &= 0.55533 \end{aligned}$$

$z_2$  is the approximate value of  $z$  at

$$t = t_2 = 0.50$$

$$z_2 = z(0.50) \approx 0.55533$$

For  $i = 2$ ,  $t_2 = 0.50$ ,  $y_2 = 1.57384$ ,  $z_2 = 0.55533$

From Equation (E3.4a)

$$\begin{aligned} k_1^y &= f_1(t_2, y_2, z_2) \\ &= f_1(0.50, 1.5738, 0.55533) \\ &= 0.55533 \end{aligned}$$

From Equation (E3.6a)

$$\begin{aligned} k_1^z &= f_2(t_2, y_2, z_2) \\ &= f_2(0.50, 1.5738, 0.55533) \\ &= e^{-0.50} - 2(0.55533) - 1.5738 \\ &= -2.0779 \end{aligned}$$

From Equation (E3.4b)

$$\begin{aligned} k_2^y &= f_1(t_2 + h, y_2 + hk_1^y, z_2 + hk_1^z) \\ &= f_1(0.50 + 0.25, 1.5738 + (0.25)(0.55533), 0.55533 + (0.25)(-2.0779)) \\ &= f_1(0.75, 1.7126, 0.035836) \\ &= 0.035836 \end{aligned}$$

From Equation (E3.6b)

$$\begin{aligned} k_2^z &= f_2(t_2 + h, y_2 + hk_1^y, z_2 + hk_1^z) \\ &= f_2(0.50 + 0.25, 1.5738 + (0.25)(0.55533), 0.55533 + (0.25)(-2.0779)) \\ &= f_2(0.75, 1.7126, 0.035836) \\ &= e^{-0.75} - 2(0.035836) - 1.7126 \\ &= -1.3119 \end{aligned}$$

From Equation (E3.3)

$$\begin{aligned} y_3 &= y_2 + \frac{1}{2}(k_1^y + k_2^y)h \\ &= 1.5738 + \frac{1}{2}(0.55533 + 0.035836)(0.25) \\ &= 1.6477 \end{aligned}$$

$y_3$  is the approximate value of  $y$  at

$$t = t_3 = t_2 + h = 0.50 + 0.25 = 0.75$$

$$y_3 = y(0.75) \approx 1.6477$$

b) From Equation (E3.5)

$$\begin{aligned} z_3 &= z_2 + \frac{1}{2}(k_1^z + k_2^z)h \\ &= 0.55533 + \frac{1}{2}(-2.0779 + (-1.3119))(0.25) \\ &= 0.13158 \end{aligned}$$

$z_3$  is the approximate value of  $z$  at

$$t = t_3 = 0.75$$

$$z_3 = z(0.75) \cong 0.13158$$

The intermediate and the final results are shown in Table 1.

**Table 1** Intermediate results of Heun's method.

$i$	0	1	2
$t_i$	0	0.25	0.50
$y_i$	1	1.3750	1.5738
$z_i$	2	1.1598	0.55533
$k_1^y$	2	1.1598	0.55533
$k_1^z$	-4	-2.9158	-2.0779
$k_2^y$	1	0.43087	0.035836
$k_2^z$	-2.7211	-1.9201	-1.3119
$y_{i+1}$	1.3750	1.5738	1.6477
$z_{i+1}$	1.1598	0.55533	0.13158

---

**ORDINARY DIFFERENTIAL EQUATIONS**

---

Topic	Higher Order Equations
Summary	Textbook notes on higher order differential equations
Major	General Engineering
Authors	Autar Kaw
Last Revised	October 13, 2010
Web Site	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

---

## 08.06

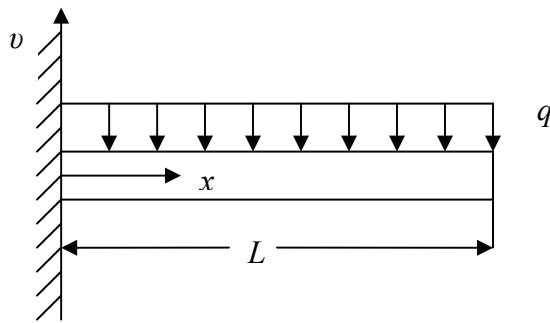
# Shooting Method for Ordinary Differential Equations

After reading this chapter, you should be able to

1. learn the shooting method algorithm to solve boundary value problems, and
2. apply shooting method to solve boundary value problems.

### What is the shooting method?

Ordinary differential equations are given either with initial conditions or with boundary conditions. Look at the problem below.



**Figure 1** A cantilevered uniformly loaded beam.

To find the deflection  $v$  as a function of location  $x$ , due to a uniform load  $q$ , the ordinary differential equation that needs to be solved is

$$\frac{d^2v}{dx^2} = \frac{q}{2EI}(L - x)^2 \quad (1)$$

where

$L$  is the length of the beam,

$E$  is the Young's modulus of the beam, and

$I$  is the second moment of area of the cross-section of the beam.

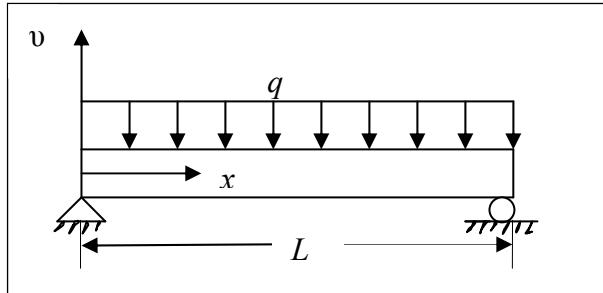
Two conditions are needed to solve the problem, and those are

$$v(0) = 0$$

$$\frac{dv}{dx}(0) = 0 \quad (2a,b)$$

as it is a cantilevered beam at  $x = 0$ . These conditions are *initial conditions* as they are given at an initial point,  $x = 0$ , so that we can find the deflection along the length of the beam.

Now consider a similar beam problem, where the beam is simply supported on the two ends



**Figure 2** A simply supported uniformly loaded beam.

To find the deflection  $v$  as a function of  $x$  due to the uniform load  $q$ , the ordinary differential equation that needs to be solved is

$$\frac{d^2v}{dx^2} = \frac{qx}{2EI} (x - L) \quad (3)$$

Two conditions are needed to solve the problem, and those are

$$\begin{aligned} v(0) &= 0 \\ v(L) &= 0 \end{aligned} \quad (4a,b)$$

as it is a simply supported beam at  $x = 0$  and  $x = L$ . These conditions are *boundary conditions* as they are given at the two boundaries,  $x = 0$  and  $x = L$ .

### The shooting method

The shooting method uses the same methods that were used in solving initial value problems. This is done by assuming initial values that would have been given if the ordinary differential equation were an initial value problem. The boundary value obtained is then compared with the actual boundary value. Using trial and error or some scientific approach, one tries to get as close to the boundary value as possible. This method is best explained by an example.

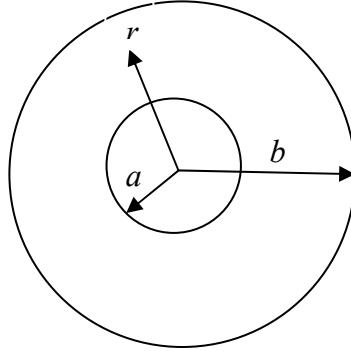
Take the case of a pressure vessel that is being tested in the laboratory to check its ability to withstand pressure. For a thick pressure vessel of inner radius  $a$  and outer radius  $b$ , the differential equation for the radial displacement  $u$  of a point along the thickness is given by

$$\frac{d^2u}{dr^2} + \frac{1}{r} \frac{du}{dr} - \frac{u}{r^2} = 0 \quad (5)$$

Assume that the inner radius  $a = 5"$  and the outer radius  $b = 8"$ , and the material of the pressure vessel is ASTM36 steel. The yield strength of this type of steel is 36 ksi. Two strain gages that are bonded tangentially at the inner and the outer radius measure the normal tangential strain in the pressure vessel as

$$\epsilon_{t/r=a} = 0.00077462$$

$$\epsilon_{t/r=b} = 0.00038462 \quad (6ab)$$



**Figure 3** Cross-sectional geometry of a pressure vessel.

at the maximum needed pressure. Since the radial displacement and tangential strain are related simply by

$$\epsilon_t = \frac{u}{r}, \quad (7)$$

then

$$\begin{aligned} u|_{r=a} &= 0.00077462 \times 5 = 0.0038731'' \\ u|_{r=b} &= 0.00038462 \times 8 = 0.0030770'' \end{aligned} \quad (8)$$

Starting with the ordinary differential equation

$$\frac{d^2u}{dr^2} + \frac{1}{r} \frac{du}{dr} - \frac{u}{r^2} = 0, \quad u(5) = 0.0038731, \quad u(8) = 0.0030770$$

Let

$$\frac{du}{dr} = w \quad (9)$$

Then

$$\frac{dw}{dr} + \frac{1}{r} w - \frac{u}{r^2} = 0 \quad (10)$$

giving us two first order differential equations as

$$\begin{aligned} \frac{du}{dr} &= w, \quad u(5) = 0.0038731'' \\ \frac{dw}{dr} &= -\frac{w}{r} + \frac{u}{r^2}, \quad w(5) = \text{not known} \end{aligned} \quad (11a,b)$$

Let us assume

$$w(5) = \frac{du}{dr}(5) \approx \frac{u(8) - u(5)}{8 - 5} = -0.00026538$$

Set up the initial value problem.

$$\frac{du}{dr} = w = f_1(r, u, w), \quad u(5) = 0.0038731''$$

$$\frac{dw}{dr} = -\frac{w}{r} + \frac{u}{r^2} = f_2(r, u, w), w(5) = -0.00026538 \quad (12a,b)$$

Using Euler's method,

$$\begin{aligned} u_{i+1} &= u_i + f_1(r_i, u_i, w_i)h \\ w_{i+1} &= w_i + f_2(r_i, u_i, w_i)h \end{aligned} \quad (13a,b)$$

Let us consider 4 segments between the two boundaries,  $r = 5"$  and  $r = 8"$ , then

$$h = \frac{8-5}{4} = 0.75"$$

$$i = 0, r_0 = 5, u_0 = 0.0038731", w_0 = -0.00026538$$

$$\begin{aligned} u_1 &= u_0 + f_1(r_0, u_0, w_0)h \\ &= 0.003871 + f_1(5, 0.003871, -0.00026538)(0.75) \\ &= 0.003871 + (-0.00026538)(0.75) \\ &= 0.0036741" \\ w_1 &= w_0 + f_2(r_0, u_0, w_0)h \\ &= -0.00026538 + f_2(5, 0.0038731, -0.00026538)(0.75) \\ &= -0.00026538 + \left( -\frac{-0.00026538}{5} + \frac{0.003871}{5^2} \right)(0.75) \\ &= -0.00010938 \end{aligned}$$

$$i = 1, r_1 = r_0 + h = 5 + 0.75 = 5.75",$$

$$u_1 = 0.0036741", w_1 = -0.00010940$$

$$\begin{aligned} u_2 &= u_1 + f_1(r_1, u_1, w_1)h \\ &= 0.0036741 + f_1(5.75, 0.0036741, -0.00010938)(0.75) \\ &= 0.0036741 + (-0.00010938)(0.75) \\ &= 0.0035920" \end{aligned}$$

$$\begin{aligned} w_2 &= w_1 + f_2(r_1, u_1, w_1)h \\ &= -0.00010938 + f_2(5.75, 0.0036741, -0.00010938)(0.75) \\ &= -0.00010938 + (0.00013015)(0.75) \\ &= -0.000011769 \end{aligned}$$

$$i = 2, r_2 = r_1 + h = 5.75 + 0.75 = 6.5"$$

$$u_2 = 0.0035920", w_2 = -0.000011785$$

$$\begin{aligned} u_3 &= u_2 + f_1(r_2, u_2, w_2)h \\ &= 0.0035920 + f_1(6.5, 0.0035920, -0.000011769)(0.75) \\ &= 0.0035920 + (-0.000011769)(0.75) \\ &= 0.0035832" \end{aligned}$$

$$\begin{aligned} w_3 &= w_2 + f_2(r_2, u_2, w_2)h \\ &= -0.000011769 + f_2(6.5, 0.0035920, -0.000011769)(0.75) \end{aligned}$$

$$\begin{aligned}
 &= -0.000011769 + (0.000086829)(0.75) \\
 &= 0.000053352
 \end{aligned}$$

$$\begin{aligned}
 i &= 3, r_3 = r_2 + h = 6.50 + 0.75 = 7.25" \\
 u_3 &= 0.0035832", w_3 = 0.000053352 \\
 u_4 &= u_3 + f_1(r_3, u_3, w_3)h \\
 &= 0.0035832 + f_1(7.25, 0.0035832, 0.000053352)(0.75) \\
 &= 0.0035832 + (0.000053352)(0.75) \\
 &= 0.0036232" \\
 w_4 &= w_3 + f_2(r_3, u_3, w_3)h \\
 &= -0.000011785 + f_2(7.25, 0.0035832, -0.000053352)(0.75) \\
 &= 0.000053352 + (0.000060811)(0.75) \\
 &= 0.000098961
 \end{aligned}$$

At

$$r = r_4 = r_3 + h = 7.25 + 0.75 = 8"$$

we have

$$u_4 = u(8) \approx 0.0036232"$$

While the given value of this boundary condition is

$$u_4 = u(8) = 0.003070"$$

Let us assume a new value for  $\frac{du}{dr}(5)$ . Based on the first assumed value, maybe using twice the value of initial guess.

$$w(5) = 2 \frac{du}{dr}(5) \approx 2 \frac{u(8) - u(5)}{8 - 5} = 2(-0.00026538) = -0.00053076$$

Using  $h = 0.75$ , and Euler's method, we get

$$u_4 = u(8) \approx 0.0029665"$$

While the given value of this boundary condition is

$$u_4 = u(8) = 0.0030770"$$

Can we use the results obtained from the two previous iterations to get a better estimate of the assumed initial condition of  $\frac{du}{dr}(5)$ ? One method is to use linear interpolation on the

obtained data for the two assumed values of  $\frac{du}{dr}(5)$ .

With

$$\frac{du}{dr}(5) \approx -0.00026538,$$

we obtained

$u(8) \approx 0.0036232"$ , and

with

$$\frac{du}{dr}(5) \approx -0.00053076,$$

we obtained

$$u(8) \approx 0.0029665"$$

so a better starting value of  $\frac{du}{dr}(5)$  knowing that the actual value at

$$u(8) = 0.00030770",$$

we get

$$\begin{aligned} \frac{du}{dr}(5) &\approx \frac{-0.00053076 - (-0.00026538)}{0.0029645 - 0.0036232} (0.0030770 - 0.0036232) + (-0.00026538) \\ &= -0.00048611 \end{aligned}$$

Using  $h = 0.75"$ , and repeating the Euler's method with

$$w(5) = -0.00048611,$$

we get

$$u_4 = u(8) \approx 0.0030769"$$

while the actual given value of this boundary condition is

$$u(8) = 0.0030770".$$

In this case, this value coincides with the actual value of  $u(8)$ . If that were not the case, one would continue to use linear interpolation to refine the value of  $u_4$  till one gets close to the actual value of  $u(8)$ . Note that the step size and the numerical method used would influence the accuracy for the obtained values. For the last case, the values are as follows

$$u_0 = u(5) = 0.0038731"$$

$$u_1 = u(5.75) \approx 0.0035085"$$

$$u_2 = u(6.50) \approx 0.0032858"$$

$$u_3 = u(7.25) \approx 0.0031518"$$

$$u_4 = u(8.00) \approx 0.0030770"$$

See Figure 4 for the comparison of the results with different initial guesses of the slope.

Using  $h = 0.75"$  and Runge-Kutta 4th order method,

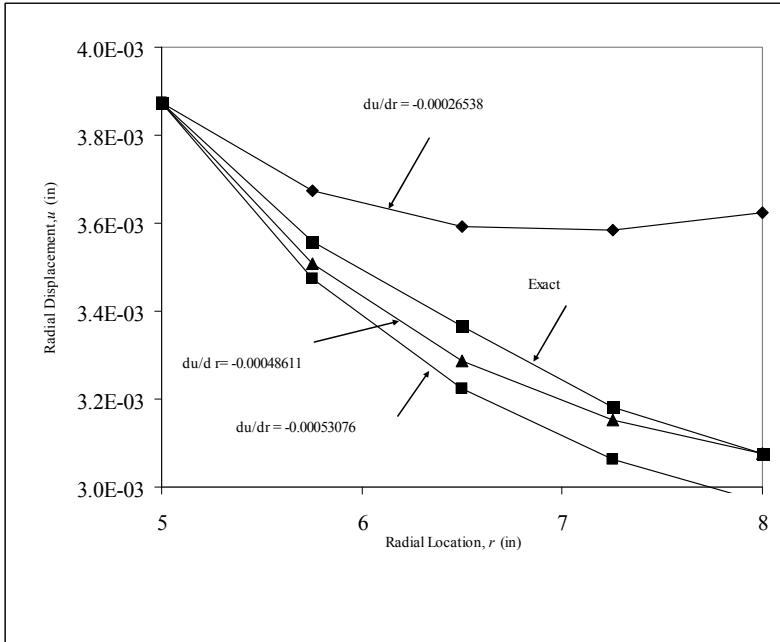
$$u_1 = u(5) = 0.0038731"$$

$$u_2 = u(5.75) \approx 0.0035554"$$

$$u_3 = u(6.50) \approx 0.0033341"$$

$$u_4 = u(7.25) \approx 0.00317923"$$

$$u_5 = u(8) \approx 0.0030723"$$



**Figure 4** Comparison of results with different initial guesses of slope

Table 1 shows the comparison of the results obtained using Euler's, Runge-Kutta and exact methods.

**Table 1** Comparison of Euler and Runge-Kutta results with exact results.

r (in)	Exact (in)	Euler (in)	$ \varepsilon_t $ (%)	Runge-Kutta (in)	$ \varepsilon_t $ (%)
5	$3.8731 \times 10^{-3}$	$3.8731 \times 10^{-3}$	0.0000	$3.8731 \times 10^{-3}$	0.0000
5.75	$3.5567 \times 10^{-3}$	$3.5085 \times 10^{-3}$	1.3731	$3.5554 \times 10^{-3}$	$3.5824 \times 10^{-2}$
6.5	$3.3366 \times 10^{-3}$	$3.2858 \times 10^{-3}$	1.5482	$3.3341 \times 10^{-3}$	$7.4037 \times 10^{-2}$
7.25	$3.1829 \times 10^{-3}$	$3.1518 \times 10^{-3}$	$9.8967 \times 10^{-1}$	$3.1792 \times 10^{-3}$	$1.1612 \times 10^{-1}$
8	$3.0770 \times 10^{-3}$	$3.0770 \times 10^{-3}$	1.9500	$3.0723 \times 10^{-3}$	$1.5168 \times 10^{-1}$

---

### ORDINARY DIFFERENTIAL EQUATIONS

---

Topic	Shooting method
Summary	Textbook notes on the shooting method for ODE.
Major	General Engineering
Authors	Autar Kaw
Last Revised	December 23, 2009
Web Site	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

---

# Chapter 08.07

## Finite Difference Method for Ordinary Differential Equations

After reading this chapter, you should be able to

1. Understand what the finite difference method is and how to use it to solve problems.

### What is the finite difference method?

The finite difference method is used to solve ordinary differential equations that have conditions imposed on the boundary rather than at the initial point. These problems are called boundary-value problems. In this chapter, we solve second-order ordinary differential equations of the form

$$\frac{d^2y}{dx^2} = f(x, y, y'), a \leq x \leq b, \quad (1)$$

with boundary conditions

$$y(a) = y_a \text{ and } y(b) = y_b \quad (2)$$

Many academics refer to boundary value problems as position-dependent and initial value problems as time-dependent. That is not necessarily the case as illustrated by the following examples.

The differential equation that governs the deflection  $y$  of a simply supported beam under uniformly distributed load (Figure 1) is given by

$$\frac{d^2y}{dx^2} = \frac{qx(L-x)}{2EI} \quad (3)$$

where

$x$  = location along the beam (in)

$E$  = Young's modulus of elasticity of the beam (psi)

$I$  = second moment of area (in<sup>4</sup>)

$q$  = uniform loading intensity (lb/in)

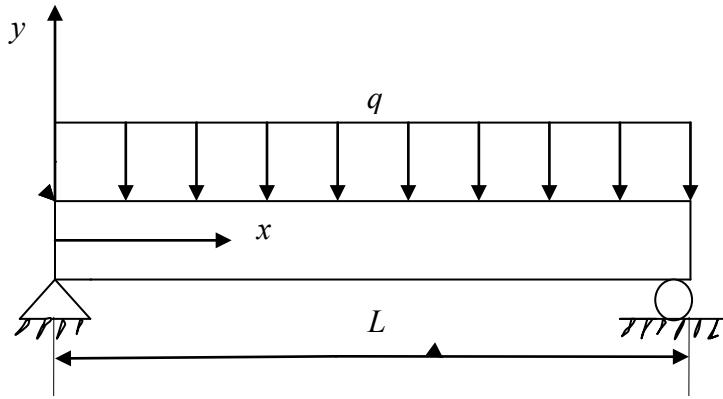
$L$  = length of beam (in)

The conditions imposed to solve the differential equation are

$$y(x = 0) = 0 \quad (4)$$

$$y(x = L) = 0$$

Clearly, these are boundary values and hence the problem is considered a boundary-value problem.



**Figure 1** Simply supported beam with uniform distributed load.

Now consider the case of a cantilevered beam with a uniformly distributed load (Figure 2). The differential equation that governs the deflection  $y$  of the beam is given by

$$\frac{d^2y}{dx^2} = \frac{q(L-x)^2}{2EI} \quad (5)$$

where

$x$  = location along the beam (in)

$E$  = Young's modulus of elasticity of the beam (psi)

$I$  = second moment of area ( $\text{in}^4$ )

$q$  = uniform loading intensity (lb/in)

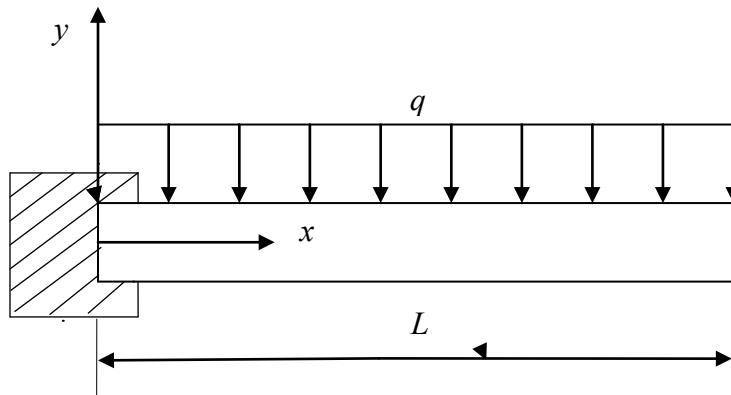
$L$  = length of beam (in)

The conditions imposed to solve the differential equation are

$$y(x=0) = 0 \quad (6)$$

$$\frac{dy}{dx}(x=0) = 0$$

Clearly, these are initial values and hence the problem needs to be considered as an initial value problem.



**Figure 2** Cantilevered beam with a uniformly distributed load.

**Example 1**

The deflection  $y$  in a simply supported beam with a uniform load  $q$  and a tensile axial load  $T$  is given by

$$\frac{d^2y}{dx^2} - \frac{Ty}{EI} = \frac{qx(L-x)}{2EI} \quad (\text{E1.1})$$

where

$x$  = location along the beam (in)

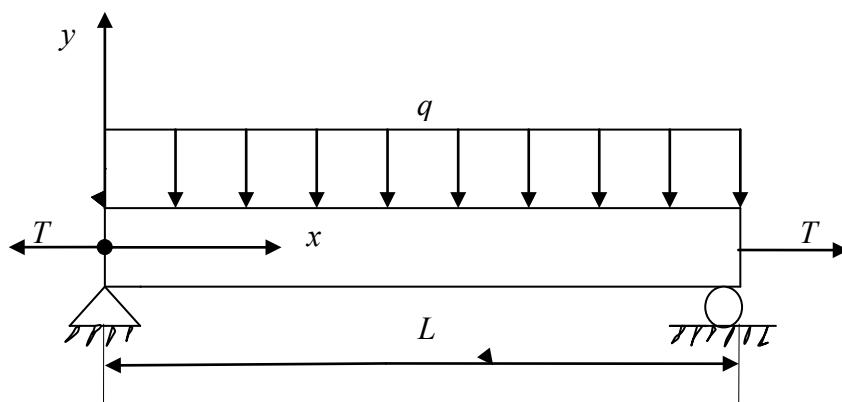
$T$  = tension applied (lbs)

$E$  = Young's modulus of elasticity of the beam (psi)

$I$  = second moment of area (in<sup>4</sup>)

$q$  = uniform loading intensity (lb/in)

$L$  = length of beam (in)



**Figure 3** Simply supported beam for Example 1.

Given,

$$T = 7200 \text{ lbs}, q = 5400 \text{ lbs/in}, L = 75 \text{ in}, E = 30 \text{ Ms}i, \text{ and } I = 120 \text{ in}^4,$$

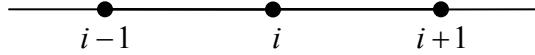
- Find the deflection of the beam at  $x = 50$ ". Use a step size of  $\Delta x = 25"$  and approximate the derivatives by central divided difference approximation.
- Find the relative true error in the calculation of  $y(50)$ .

**Solution**

- Substituting the given values,

$$\begin{aligned} \frac{d^2y}{dx^2} - \frac{7200y}{(30 \times 10^6)(120)} &= \frac{(5400)x(75-x)}{2(30 \times 10^6)(120)} \\ \frac{d^2y}{dx^2} - 2 \times 10^{-6}y &= 7.5 \times 10^{-7}x(75-x) \end{aligned} \quad (\text{E1.2})$$

Approximating the derivative  $\frac{d^2y}{dx^2}$  at node  $i$  by the central divided difference approximation,



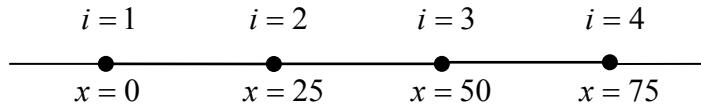
**Figure 4** Illustration of finite difference nodes using central divided difference method.

$$\frac{d^2y}{dx^2} \approx \frac{y_{i+1} - 2y_i + y_{i-1}}{(\Delta x)^2} \quad (\text{E1.3})$$

We can rewrite the equation as

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{(\Delta x)^2} - 2 \times 10^{-6} y_i = 7.5 \times 10^{-7} x_i (75 - x_i) \quad (\text{E1.4})$$

Since  $\Delta x = 25$ , we have 4 nodes as given in Figure 3



**Figure 5** Finite difference method from  $x = 0$  to  $x = 75$  with  $\Delta x = 25$ .

The location of the 4 nodes then is

$$x_0 = 0$$

$$x_1 = x_0 + \Delta x = 0 + 25 = 25$$

$$x_2 = x_1 + \Delta x = 25 + 25 = 50$$

$$x_3 = x_2 + \Delta x = 50 + 25 = 75$$

Writing the equation at each node, we get

Node 1: From the simply supported boundary condition at  $x = 0$ , we obtain

$$y_1 = 0 \quad (\text{E1.5})$$

Node 2: Rewriting equation (E1.4) for node 2 gives

$$\begin{aligned} \frac{y_3 - 2y_2 + y_1}{(25)^2} - 2 \times 10^{-6} y_2 &= 7.5 \times 10^{-7} x_2 (75 - x_2) \\ 0.0016y_1 - 0.003202y_2 + 0.0016y_3 &= 7.5 \times 10^{-7} (25)(75 - 25) \\ 0.0016y_1 - 0.003202y_2 + 0.0016y_3 &= 9.375 \times 10^{-4} \end{aligned} \quad (\text{E1.6})$$

Node 3: Rewriting equation (E1.4) for node 3 gives

$$\begin{aligned} \frac{y_4 - 2y_3 + y_2}{(25)^2} - 2 \times 10^{-6} y_3 &= 7.5 \times 10^{-7} x_3 (75 - x_3) \\ 0.0016y_2 - 0.003202y_3 + 0.0016y_4 &= 7.5 \times 10^{-7} (50)(75 - 50) \\ 0.0016y_2 - 0.003202y_3 + 0.0016y_4 &= 9.375 \times 10^{-4} \end{aligned} \quad (\text{E1.7})$$

Node 4: From the simply supported boundary condition at  $x = 75$ , we obtain

$$y_4 = 0 \quad (\text{E1.8})$$

Equations (E1.5-E1.8) are 4 simultaneous equations with 4 unknowns and can be written in matrix form as

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.0016 & -0.003202 & 0.0016 & 0 \\ 0 & 0.0016 & -0.003202 & 0.0016 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 9.375 \times 10^{-4} \\ 9.375 \times 10^{-4} \\ 0 \end{bmatrix}$$

The above equations have a coefficient matrix that is tridiagonal (we can use Thomas' algorithm to solve the equations) and is also strictly diagonally dominant (convergence is guaranteed if we use iterative methods such as the Gauss-Siedel method). Solving the equations we get,

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 0 \\ -0.5852 \\ -0.5852 \\ 0 \end{bmatrix}$$

$$y(50) = y(x_2) \approx y_2 = -0.5852"$$

The exact solution of the ordinary differential equation is derived as follows. The homogeneous part of the solution is given by solving the characteristic equation

$$\begin{aligned} m^2 - 2 \times 10^{-6} &= 0 \\ m &= \pm 0.0014142 \end{aligned}$$

Therefore,

$$y_h = K_1 e^{0.0014142x} + K_2 e^{-0.0014142x}$$

The particular part of the solution is given by

$$y_p = Ax^2 + Bx + C$$

Substituting the differential equation (E1.2) gives

$$\frac{d^2 y_p}{dx^2} - 2 \times 10^{-6} y_p = 7.5 \times 10^{-7} x(75 - x)$$

$$\frac{d^2}{dx^2}(Ax^2 + Bx + C) - 2 \times 10^{-6}(Ax^2 + Bx + C) = 7.5 \times 10^{-7} x(75 - x)$$

$$2A - 2 \times 10^{-6}(Ax^2 + Bx + C) = 7.5 \times 10^{-7} x(75 - x)$$

$$-2 \times 10^{-6} Ax^2 - 2 \times 10^{-6} Bx + (2A - 2 \times 10^{-6} C) = 5.625 \times 10^{-5} x - 7.5 \times 10^{-7} x^2$$

Equating terms gives

$$-2 \times 10^{-6} A = -7.5 \times 10^{-7}$$

$$-2 \times 10^{-6} B = -5.625 \times 10^{-5}$$

$$2A - 2 \times 10^{-6} C = 0$$

Solving the above equation gives

$$A = 0.375$$

$$B = -28.125$$

$$C = 3.75 \times 10^5$$

The particular solution then is

$$y_p = 0.375x^2 - 28.125x + 3.75 \times 10^5$$

The complete solution is then given by

$$y = 0.375x^2 - 28.125x + 3.75 \times 10^5 + K_1 e^{0.0014142x} + K_2 e^{-0.0014142x}$$

Applying the following boundary conditions

$$y(x=0) = 0$$

$$y(x=75) = 0$$

we obtain the following system of equations

$$K_1 + K_2 = -3.75 \times 10^5$$

$$1.1119K_1 + 0.89937K_2 = -3.75 \times 10^5$$

These equations are represented in matrix form by

$$\begin{bmatrix} 1 & 1 \\ 1.1119 & 0.89937 \end{bmatrix} \begin{bmatrix} K_1 \\ K_2 \end{bmatrix} = \begin{bmatrix} -3.75 \times 10^5 \\ -3.75 \times 10^5 \end{bmatrix}$$

A number of different numerical methods may be utilized to solve this system of equations such as the Gaussian elimination. Using any of these methods yields

$$\begin{bmatrix} K_1 \\ K_2 \end{bmatrix} = \begin{bmatrix} -1.775656226 \times 10^5 \\ -1.974343774 \times 10^5 \end{bmatrix}$$

Substituting these values back into the equation gives

$$y = 0.375x^2 - 28.125x + 3.75 \times 10^5 - 1.775656266 \times 10^5 e^{0.0014142x} - 1.974343774 \times 10^5 e^{-0.0014142x}$$

Unlike other examples in this chapter and in the book, the above expression for the deflection of the beam is displayed with a larger number of significant digits. This is done to minimize the round-off error because the above expression involves subtraction of large numbers that are close to each other.

b) To calculate the relative true error, we must first calculate the value of the exact solution at  $y = 50$ .

$$y(50) = 0.375(50)^2 - 28.125(50) + 3.75 \times 10^5 - 1.775656266 \times 10^5 e^{0.0014142(50)} - 1.974343774 \times 10^5 e^{-0.0014142(50)}$$

$$y(50) = -0.5320$$

The true error is given by

$$E_t = \text{Exact Value} - \text{Approximate Value}$$

$$E_t = -0.5320 - (-0.5852)$$

$$E_t = 0.05320$$

The relative true error is given by

$$\epsilon_t = \frac{\text{True Error}}{\text{True Value}} \times 100\%$$

$$\epsilon_t = \frac{0.05320}{-0.5320} \times 100\%$$

$$\epsilon_t = -10\%$$

**Example 2**

Take the case of a pressure vessel that is being tested in the laboratory to check its ability to withstand pressure. For a thick pressure vessel of inner radius  $a$  and outer radius  $b$ , the differential equation for the radial displacement  $u$  of a point along the thickness is given by

$$\frac{d^2u}{dr^2} + \frac{1}{r} \frac{du}{dr} - \frac{u}{r^2} = 0 \quad (\text{E2.3})$$

The inner radius  $a = 5"$  and the outer radius  $b = 8"$ , and the material of the pressure vessel is ASTM A36 steel. The yield strength of this type of steel is 36 ksi. Two strain gages that are bonded tangentially at the inner and the outer radius measure normal tangential strain as

$$\begin{aligned} \epsilon_{t/r=a} &= 0.00077462 \\ \epsilon_{t/r=b} &= 0.00038462 \end{aligned} \quad (\text{E2.4a,b})$$

at the maximum needed pressure. Since the radial displacement and tangential strain are related simply by

$$\epsilon_t = \frac{u}{r}, \quad (\text{E2.5})$$

then

$$\begin{aligned} u|_{r=a} &= 0.00077462 \times 5 = 0.0038731'' \\ u|_{r=b} &= 0.00038462 \times 8 = 0.0030769'' \end{aligned}$$

The maximum normal stress in the pressure vessel is at the inner radius  $r = a$  and is given by

$$\sigma_{\max} = \frac{E}{1-\nu^2} \left( \frac{u}{r} \Big|_{r=a} + \nu \frac{du}{dr} \Big|_{r=a} \right) \quad (\text{E2.7})$$

where

$E$  = Young's modulus of steel ( $E = 30$  Msi)

$\nu$  = Poisson's ratio ( $\nu = 0.3$ )

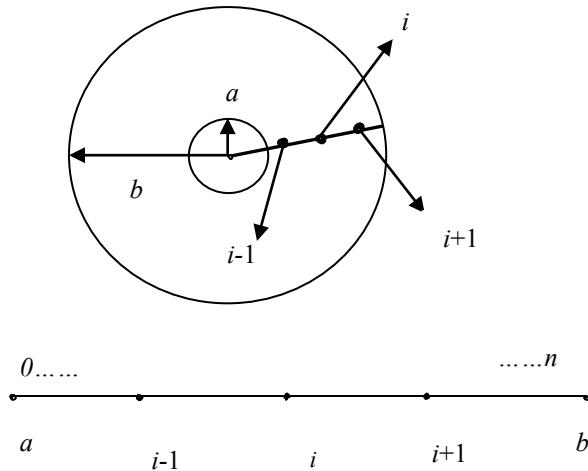
The factor of safety, FS is given by

$$FS = \frac{\text{Yield strength of steel}}{\sigma_{\max}} \quad (\text{E2.8})$$

- a) Divide the radial thickness of the pressure vessel into 6 equidistant nodes, and find the radial displacement profile
- b) Find the maximum normal stress and factor of safety as given by equation (E2.8)
- c) Find the exact value of the maximum normal stress as given by equation (E2.8) if it is given that the exact expression for radial displacement is of the form

$$u = C_1 r + \frac{C_2}{r}.$$

Calculate the relative true error.

**Solution****Figure 4** Nodes along the radial direction.

a) The radial locations from  $r = a$  to  $r = b$  are divided into  $n$  equally spaced segments, and hence resulting in  $n+1$  nodes. This will allow us to find the dependent variable  $u$  numerically at these nodes.

At node  $i$  along the radial thickness of the pressure vessel,

$$\frac{d^2u}{dr^2} \approx \frac{u_{i+1} - 2u_i + u_{i-1}}{(\Delta r)^2} \quad (\text{E2.9})$$

$$\frac{du}{dr} \approx \frac{u_{i+1} - u_i}{\Delta r} \quad (\text{E2.10})$$

Such substitutions will convert the ordinary differential equation into a linear equation (but with more than one unknown). By writing the resulting linear equation at different points at which the ordinary differential equation is valid, we get simultaneous linear equations that can be solved by using techniques such as Gaussian elimination, the Gauss-Siedel method, etc.

Substituting these approximations from Equations (E2.9) and (E2.10) in Equation (E2.3)

$$\frac{u_{i+1} - 2u_i + u_{i-1}}{(\Delta r)^2} + \frac{1}{r_i} \frac{u_{i+1} - u_i}{\Delta r} - \frac{u_i}{r_i^2} = 0 \quad (\text{E2.11})$$

$$\left( \frac{1}{(\Delta r)^2} + \frac{1}{r_i \Delta r} \right) u_{i+1} + \left( -\frac{2}{(\Delta r)^2} - \frac{1}{r_i \Delta r} - \frac{1}{r_i^2} \right) u_i + \frac{1}{(\Delta r)^2} u_{i-1} = 0 \quad (\text{E2.12})$$

Let us break the thickness,  $b - a$ , of the pressure vessel into  $n+1$  nodes, that is  $r = a$  is node  $i = 0$  and  $r = b$  is node  $i = n$ . That means we have  $n+1$  unknowns.

We can write the above equation for nodes  $1, \dots, n-1$ . This will give us  $n-1$  equations. At the edge nodes,  $i = 0$  and  $i = n$ , we use the boundary conditions of

$$u_0 = u|_{r=a}$$

$$u_n = u|_{r=b}$$

This gives a total of  $n+1$  equations. So we have  $n+1$  unknowns and  $n+1$  linear equations. These can be solved by any of the numerical methods used for solving simultaneous linear equations.

We have been asked to do the calculations for  $n=5$ , that is a total of 6 nodes. This gives

$$\begin{aligned}\Delta r &= \frac{b-a}{n} \\ &= \frac{8-5}{5} \\ &= 0.6\text{"}\end{aligned}$$

$$\text{At node } i=0, r_0 = a = 5\text{"}, \quad u_0 = 0.0038731\text{"} \quad (\text{E2.13})$$

$$\text{At node } i=1, r_1 = r_0 + \Delta r = 5 + 0.6 = 5.6\text{"} \quad (\text{E2.14})$$

$$\begin{aligned}\frac{1}{0.6^2}u_0 + \left(-\frac{2}{0.6^2} - \frac{1}{(5.6)(0.6)} - \frac{1}{(5.6)^2}\right)u_1 + \left(\frac{1}{0.6^2} + \frac{1}{(5.6)(0.6)}\right)u_2 &= 0 \\ 2.7778u_0 - 5.8851u_1 + 3.0754u_2 &= 0\end{aligned} \quad (\text{E2.15})$$

$$\text{At node } i=2, \quad r_2 = r_1 + \Delta r = 5.6 + 0.6 = 6.2\text{"}$$

$$\begin{aligned}\frac{1}{0.6^2}u_1 + \left(-\frac{2}{0.6^2} - \frac{1}{(6.2)(0.6)} - \frac{1}{6.2^2}\right)u_2 + \left(\frac{1}{0.6^2} + \frac{1}{(6.2)(0.6)}\right)u_3 &= 0 \\ 2.7778u_1 - 5.8504u_2 + 3.0466u_3 &= 0\end{aligned} \quad (\text{E2.16})$$

$$\text{At node } i=3, \quad r_3 = r_2 + \Delta r = 6.2 + 0.6 = 6.8\text{"}$$

$$\begin{aligned}\frac{1}{0.6^2}u_2 + \left(-\frac{2}{0.6^2} - \frac{1}{(6.8)(0.6)} - \frac{1}{6.8^2}\right)u_3 + \left(\frac{1}{0.6^2} + \frac{1}{(6.8)(0.6)}\right)u_4 &= 0 \\ 2.7778u_2 - 5.8223u_3 + 3.0229u_4 &= 0\end{aligned} \quad (\text{E2.17})$$

$$\text{At node } i=4, \quad r_4 = r_3 + \Delta r = 6.8 + 0.6 = 7.4\text{"}$$

$$\begin{aligned}\frac{1}{0.6^2}u_3 + \left(-\frac{2}{0.6^2} - \frac{1}{(7.4)(0.6)} - \frac{1}{7.4^2}\right)u_4 + \left(\frac{1}{0.6^2} + \frac{1}{(7.4)(0.6)}\right)u_5 &= 0 \\ 2.7778u_3 - 5.7990u_4 + 3.0030u_5 &= 0\end{aligned} \quad (\text{E2.18})$$

$$\text{At node } i=5, \quad r_5 = r_4 + \Delta r = 7.4 + 0.6 = 8\text{"}$$

$$u_5 = u|_{r=b} = 0.0030769\text{"} \quad (\text{E2.19})$$

Writing Equation (E2.13) to (E2.19) in matrix form gives

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 2.7778 & -5.8851 & 3.0754 & 0 & 0 & 0 \\ 0 & 2.7778 & -5.8504 & 3.0466 & 0 & 0 \\ 0 & 0 & 2.7778 & -5.8223 & 3.0229 & 0 \\ 0 & 0 & 0 & 2.7778 & -5.7990 & 3.0030 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{bmatrix} = \begin{bmatrix} 0.0038731 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0.0030769 \end{bmatrix}$$

The above equations are a tri-diagonal system of equations and special algorithms such as Thomas' algorithm can be used to solve such a system of equations.

$$u_0 = 0.0038731''$$

$$u_1 = 0.0036165''$$

$$u_2 = 0.0034222''$$

$$u_3 = 0.0032743''$$

$$u_4 = 0.0031618''$$

$$u_5 = 0.0030769''$$

b) To find the maximum stress, it is given by Equation (E2.7) as

$$\sigma_{\max} = \frac{E}{1-\nu^2} \left( \frac{u}{r} \Big|_{r=a} + \nu \frac{du}{dr} \Big|_{r=a} \right)$$

$$E = 30 \times 10^6 \text{ psi}$$

$$\nu = 0.3$$

$$u \Big|_{r=a} = u_0 = 0.0038731''$$

$$\begin{aligned} \frac{du}{dr} \Big|_{r=a} &\approx \frac{u_1 - u_0}{\Delta r} \\ &= \frac{0.0036165 - 0.0038731}{0.6} \\ &= -0.00042767 \end{aligned}$$

The maximum stress in the pressure vessel then is

$$\begin{aligned} \sigma_{\max} &= \frac{30 \times 10^6}{1 - 0.3^2} \left( \frac{0.0038731}{5} + 0.3(-0.00042767) \right) \\ &= 2.1307 \times 10^4 \text{ psi} \end{aligned}$$

So the factor of safety  $FS$  from Equation (E2.8) is

$$FS = \frac{36 \times 10^3}{2.1307 \times 10^4} = 1.6896$$

c) The differential equation has an exact solution and is given by the form

$$u = C_1 r + \frac{C_2}{r} \quad (\text{E2.20})$$

where  $C_1$  and  $C_2$  are found by using the boundary conditions at  $r = a$  and  $r = b$ .

## Finite Difference Method

08.07.11

$$u(r=a) = u(r=5) = 0.0038731 = C_1(5) + \frac{C_2}{5}$$

$$u(r=b) = u(r=8) = 0.0030769 = C_1(8) + \frac{C_2}{8}$$

giving

$$C_1 = 0.00013462$$

$$C_2 = 0.016000$$

Thus

$$u = 0.00013462r + \frac{0.016000}{r} \quad (\text{E2.21})$$

$$\frac{du}{dr} = 0.00013462 - \frac{0.016000}{r^2} \quad (\text{E2.22})$$

$$\sigma_{\max} = \frac{E}{1-\nu^2} \left( \frac{u}{r} \Big|_{r=a} + \nu \frac{du}{dr} \Big|_{r=a} \right)$$

$$= \frac{30 \times 10^6}{1-0.3^2} \left( \frac{0.00013462(5) + \frac{0.01600}{5}}{5} + 0.3 \left( 0.0013462 - \frac{0.016000}{5^2} \right) \right)$$

$$= 2.0538 \times 10^4 \text{ psi}$$

The true error is

$$E_t = 2.0538 \times 10^4 - 2.1307 \times 10^4$$

$$= -7.6859 \times 10^2$$

The absolute relative true error is

$$|\epsilon_t| = \left| \frac{2.0538 \times 10^4 - 2.1307 \times 10^4}{2.0538 \times 10^4} \right| \times 100$$

$$= 3.744\%$$

### Example 3

The approximation in Example 2

$$\frac{du}{dr} \approx \frac{u_{i+1} - u_i}{\Delta r}$$

is first order accurate, that is, the true error is of  $O(\Delta r)$ .

The approximation

$$\frac{d^2u}{dr^2} \approx \frac{u_{i+1} - 2u_i + u_{i-1}}{(\Delta r)^2} \quad (\text{E3.1})$$

is second order accurate, that is, the true error is  $O((\Delta r)^2)$

Mixing these two approximations will result in the order of accuracy of  $O(\Delta r)$  and  $O((\Delta r)^2)$ , that is  $O(\Delta r)$ .

So it is better to approximate

$$\frac{du}{dr} \approx \frac{u_{i+1} - u_{i-1}}{2(\Delta r)} \quad (\text{E3.2})$$

because this equation is second order accurate. Repeat Example 2 with the more accurate approximations.

### Solution

a) Repeating the problem with this approximation, at node  $i$  in the pressure vessel,

$$\frac{d^2u}{dr^2} \approx \frac{u_{i+1} - 2u_i + u_{i-1}}{(\Delta r)^2} \quad (\text{E3.3})$$

$$\frac{du}{dr} \approx \frac{u_{i+1} - u_{i-1}}{2\Delta r} \quad (\text{E3.4})$$

Substituting Equations (E3.3) and (E3.4) in Equation (E2.3) gives

$$\begin{aligned} \frac{u_{i+1} - 2u_i + u_{i-1}}{(\Delta r)^2} + \frac{1}{r_i} \frac{u_{i+1} - u_{i-1}}{2(\Delta r)} - \frac{u_i}{r_i^2} &= 0 \\ \left( -\frac{1}{2r_i(\Delta r)} + \frac{1}{(\Delta r)^2} \right) u_{i-1} + \left( -\frac{2}{(\Delta r)^2} - \frac{1}{r_i^2} \right) u_i + \left( \frac{1}{(\Delta r)^2} + \frac{1}{2r_i\Delta r} \right) u_{i+1} &= 0 \end{aligned} \quad (\text{E3.5})$$

At node  $i = 0$ ,  $r_0 = a = 5"$

$$u_0 = 0.0038731" \quad (\text{E3.6})$$

At node  $i = 1$ ,  $r_1 = r_0 + \Delta r = 5 + 0.6 = 5.6"$

$$\begin{aligned} \left( -\frac{1}{2(5.6)(0.6)} + \frac{1}{(0.6)^2} \right) u_0 + \left( -\frac{2}{(0.6)^2} - \frac{1}{(5.6)^2} \right) u_1 + \left( \frac{1}{(0.6)^2} + \frac{1}{2(5.6)(0.6)} \right) u_2 &= 0 \\ 2.6297u_0 - 5.5874u_1 + 2.9266u_2 &= 0 \end{aligned} \quad (\text{E3.7})$$

At node  $i = 2$ ,  $r_2 = r_1 + \Delta r = 5.6 + 0.6 = 6.2"$

$$\begin{aligned} \left( -\frac{1}{2(6.2)(0.6)} + \frac{1}{(0.6)^2} \right) u_1 + \left( -\frac{2}{(0.6)^2} - \frac{1}{(6.2)^2} \right) u_2 + \left( \frac{1}{(0.6)^2} + \frac{1}{2(6.2)(0.6)} \right) u_3 &= 0 \\ 2.6434u_1 - 5.5816u_2 + 2.9122u_3 &= 0 \end{aligned} \quad (\text{E3.8})$$

At node  $i = 3$ ,  $r_3 = r_2 + \Delta r = 6.2 + 0.6 = 6.8"$

$$\begin{aligned} \left( -\frac{1}{2(6.8)(0.6)} + \frac{1}{(0.6)^2} \right) u_2 + \left( -\frac{2}{(0.6)^2} - \frac{1}{(6.8)^2} \right) u_3 + \left( \frac{1}{(0.6)^2} + \frac{1}{2(6.8)(0.6)} \right) u_4 &= 0 \\ 2.6552u_2 - 5.5772u_3 + 2.9003u_4 &= 0 \end{aligned} \quad (\text{E3.9})$$

At node  $i = 4$ ,  $r_4 = r_3 + \Delta r = 6.8 + 0.6 = 7.4"$

$$\begin{aligned} \left( -\frac{1}{2(7.4)(0.6)} + \frac{1}{(0.6)^2} \right) u_3 + \left( -\frac{2}{(0.6)^2} - \frac{1}{(7.4)^2} \right) u_4 + \left( \frac{1}{(0.6)^2} + \frac{1}{2(7.4)(0.6)} \right) u_5 &= 0 \\ 2.6651u_3 - 5.5738u_4 + 2.8903u_5 &= 0 \end{aligned} \quad (\text{E3.10})$$

At node  $i = 5$ ,  $r_5 = r_4 + \Delta r = 7.4 + 0.6 = 8"$

$$u_5 = u|_{r=b} = 0.0030769" \quad (\text{E3.11})$$

Writing Equations (E3.6) thru (E3.11) in matrix form gives

## Finite Difference Method

08.07.13

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 2.6297 & -5.5874 & 2.9266 & 0 & 0 & 0 \\ 0 & 2.6434 & -5.5816 & 2.9122 & 0 & 0 \\ 0 & 0 & 2.6552 & -5.5772 & 2.9003 & 0 \\ 0 & 0 & 0 & 2.6651 & -5.5738 & 2.8903 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{bmatrix} = \begin{bmatrix} 0.0038731 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0.0030769 \end{bmatrix}$$

The above equations are a tri-diagonal system of equations and special algorithms such as Thomas' algorithm can be used to solve such equations.

$$u_0 = 0.0038731 "$$

$$u_1 = 0.0036115 "$$

$$u_2 = 0.0034159 "$$

$$u_3 = 0.0032689 "$$

$$u_4 = 0.0031586 "$$

$$u_5 = 0.0030769 "$$

b)

$$\begin{aligned} \frac{du}{dr} \Big|_{r=a} &\approx \frac{-3u_0 + 4u_1 - u_2}{2(\Delta r)} \\ &= \frac{-3 \times 0.0038731 + 4 \times 0.0036115 - 0.0034159}{2(0.6)} \\ &= -4.925 \times 10^{-4} \\ \sigma_{\max} &= \frac{30 \times 10^6}{1 - 0.3^2} \left( \frac{0.0038731}{5} + 0.3(-4.925 \times 10^{-4}) \right) \\ &= 2.0666 \times 10^4 \text{ psi} \end{aligned}$$

Therefore, the factor of safety  $FS$  is

$$\begin{aligned} FS &= \frac{36 \times 10^3}{2.0666 \times 10^4} \\ &= 1.7420 \end{aligned}$$

c) The true error in calculating the maximum stress is

$$\begin{aligned} E_t &= 2.0538 \times 10^4 - 2.0666 \times 10^4 \\ &= -128 \text{ psi} \end{aligned}$$

The relative true error in calculating the maximum stress is

$$\begin{aligned} |\epsilon_t| &= \left| \frac{-128}{2.0538 \times 10^4} \right| \times 100 \\ &= 0.62323 \% \end{aligned}$$

**Table 1** Comparisons of radial displacements from two methods.

$r$	$u_{\text{exact}}$	$u_{\text{1st order}}$	$ \epsilon_t $	$u_{\text{2nd order}}$	$ \epsilon_t $
-----	--------------------	------------------------	----------------	------------------------	----------------

5	0.0038731	0.0038731	0.0000	0.0038731	0.0000
5.6	0.0036110	0.0036165	$1.5160 \times 10^{-1}$	0.0036115	$1.4540 \times 10^{-2}$
6.2	0.0034152	0.0034222	$2.0260 \times 10^{-1}$	0.0034159	$1.8765 \times 10^{-2}$
6.8	0.0032683	0.0032743	$1.8157 \times 10^{-1}$	0.0032689	$1.6334 \times 10^{-2}$
7.4	0.0031583	0.0031618	$1.0903 \times 10^{-1}$	0.0031586	$9.5665 \times 10^{-3}$
8	0.0030769	0.0030769	0.0000	0.0030769	0.0000

---

**ORDINARY DIFFERENTIAL EQUATIONS**

---

Topic      Finite Difference Methods of Solving Ordinary Differential Equations  
 Summary     Textbook notes of Finite Difference Methods of solving ordinary differential equations  
 Major       General Engineering  
 Authors      Autar Kaw, Cuong Nguyen, Luke Snyder  
 Date        July 17, 2012  
 Web Site     <http://numericalmethods.eng.usf.edu>

---

# Chapter 09.01

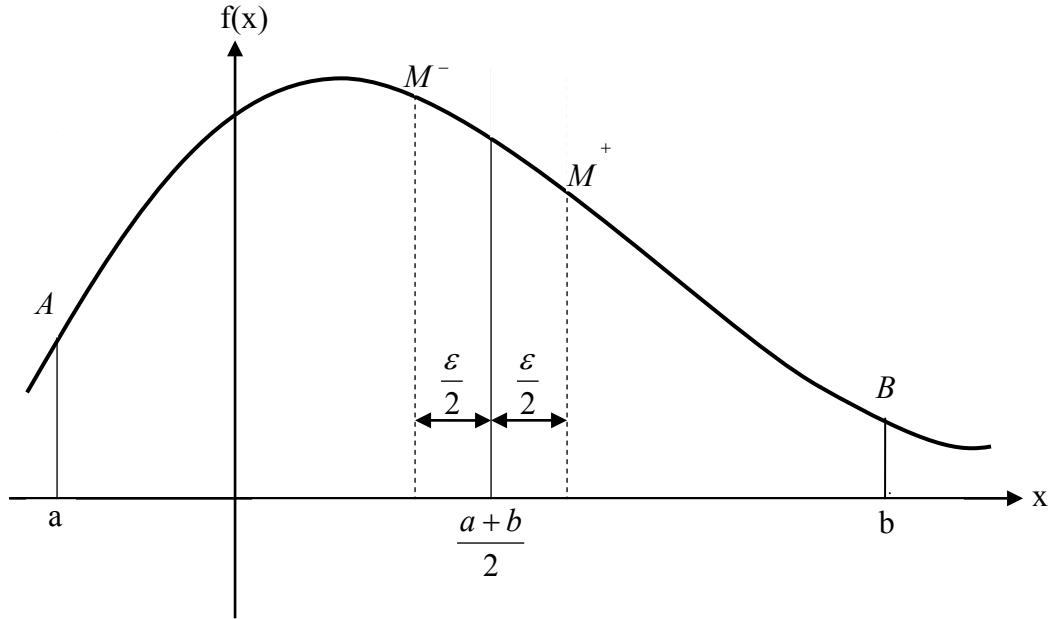
## Golden Section Search Method

After reading this chapter, you should be able to:

1. Understand the fundamentals of the Equal Interval Search method
2. Understand how the Golden Section Search method works
3. Learn about the Golden Ratio
4. Solve one-dimensional optimization problems using the Golden Section Search method

### Equal Interval Search Method

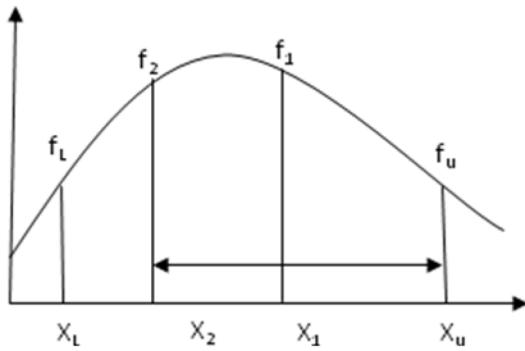
One of the simplest methods of finding the local maximum or local minimum is the Equal Interval Search method. Let's restrict our discussion to finding the local maximum of  $f(x)$  where the interval in which the local maximum occurs is  $[a,b]$ . As shown in Figure 1, let's choose an interval of  $\varepsilon$  over which we assume the maximum occurs. Then we can compute  $f\left(\frac{a+b}{2} + \frac{\varepsilon}{2}\right)$  and  $f\left(\frac{a+b}{2} - \frac{\varepsilon}{2}\right)$ . If  $f\left(\frac{a+b}{2} + \frac{\varepsilon}{2}\right) \geq f\left(\frac{a+b}{2} - \frac{\varepsilon}{2}\right)$ , then the interval in which the maximum occurs is  $\left[\frac{a+b}{2} - \frac{\varepsilon}{2}, b\right]$ , otherwise it occurs in  $\left[a, \frac{a+b}{2} + \frac{\varepsilon}{2}\right]$ . This reduces the interval in which the local maximum occurs. This procedure can be repeated until the interval is reduced to the level of our choice.



**Figure 1A** Equal interval search method (new **upper bound** can be identified).

Remarks:

As can be seen from the marked data points A,  $M^-$ ,  $M^+$ , and B on Figure 1A, the function values have increased from point A to point  $M^-$ , but then have decreased from point  $M^-$  to point  $M^+$ . Whenever there is a sudden change in the pattern, such as from increasing the function value to decreasing its value, as shown in Figure 1A (or vice versa, as shown in Figure 1B, where  $f_L < f_2 < f_1$  and then  $f_1 > f_u$ ), then the new lower and upper bound bracket values can be found. In this case, the new lower bound remains to be the same as its previous lower bound (at point A), and the new upper bound can be found (at point  $M^+$ ), as shown in Figure 1A.



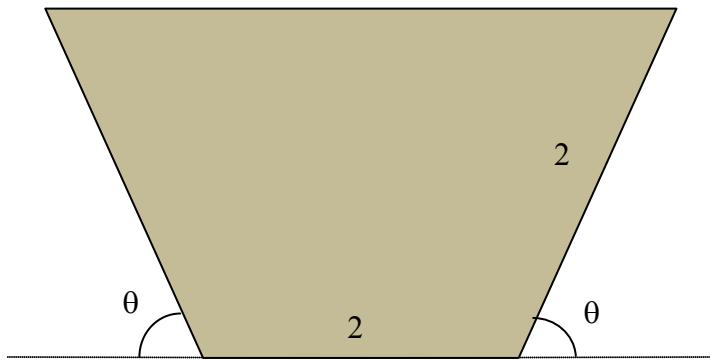
**Figure 1B** Equal interval search method (new **lower bound** can be identified).

**Example 1**

Consider Figure 2 below. The cross-sectional area  $A$  of a gutter with equal base and edge length of 2 is given by

$$A = 4 \sin \theta (1 + \cos \theta)$$

Using an initial interval of  $[0, \pi/2]$ , find the interval after 3 iterations. Use an initial interval  $\varepsilon = 0.2$ .



**Figure 2** Cross section of the gutter.

**Solution**

If we assume the initial interval to be  $[0, \pi/2] \cong [0, 1.5708]$  and choose  $\varepsilon = 0.2$ , then

$$\begin{aligned} f\left(\frac{a+b}{2} + \frac{\varepsilon}{2}\right) &= f\left(\frac{0+1.5708}{2} + \frac{0.2}{2}\right) \\ &= f(0.88540) \\ &= 5.0568 \end{aligned}$$

$$\begin{aligned} f\left(\frac{a+b}{2} - \frac{\varepsilon}{2}\right) &= f\left(\frac{0+1.5708}{2} - \frac{0.2}{2}\right) \\ &= f(0.6854) \\ &= 4.4921 \end{aligned}$$

Since  $f(0.88540) > f(0.68540)$ , the interval in which the local maximum occurs is  $[0.68540, 1.5708]$ .

Now

$$\begin{aligned} f\left(\frac{a+b}{2} + \frac{\varepsilon}{2}\right) &= f\left(\frac{0.68540+1.5708}{2} + \frac{0.2}{2}\right) \\ &= f(1.2281) \\ &= 5.0334 \end{aligned}$$

$$\begin{aligned}
 f\left(\frac{a+b}{2} - \frac{\varepsilon}{2}\right) &= f\left(\frac{0.68540 + 1.5708}{2} - \frac{0.2}{2}\right) \\
 &= f(1.0281) \\
 &= 5.1942
 \end{aligned}$$

Since  $f(1.2281) < f(1.0281)$ , the interval in which the local maximum occurs is  $[0.68540, 1.2281]$ .

Now

$$\begin{aligned}
 f\left(\frac{a+b}{2} + \frac{\varepsilon}{2}\right) &= f\left(\frac{0.68540 + 1.2281}{2} + \frac{0.2}{2}\right) \\
 &= f(1.0567) \\
 &= 5.1957
 \end{aligned}$$
  

$$\begin{aligned}
 f\left(\frac{a+b}{2} - \frac{\varepsilon}{2}\right) &= f\left(\frac{0.68540 + 1.2281}{2} - \frac{0.2}{2}\right) \\
 &= f(0.8567) \\
 &= 5.0025
 \end{aligned}$$

Since  $f(1.0567) > f(0.8567)$ , then the interval in which the local maximum occurs is  $(0.8567, 1.2281)$ . After sixteen iterations, the interval is reduced to 0.02 and the approximation of the maximum area is 5.1961 at an angle of 60.06 degrees.

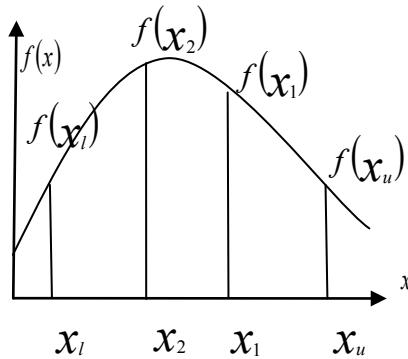
The exact answer is  $\theta = 1.0472$  for which  $f(\theta) = 5.1962$ .

### What is the Golden Section Search method used for and how does it work?

The Golden Section Search method is used to find the maximum or minimum of a unimodal function. (*A unimodal function contains only one minimum or maximum on the interval  $[a,b]$ .*) To make the discussion of the method simpler, let us assume that we are trying to find the maximum of a function. The previously introduced Equal Interval Search method is somewhat inefficient because if the interval is a small number it can take a long time to find the maximum of a function. To improve this efficiency, the Golden Section Search method is suggested.

As shown in Figure 3, choose three points  $x_l$ ,  $x_1$  and  $x_u$  ( $x_l < x_1 < x_u$ ) along the  $x$ -axis with corresponding values of the function  $f(x_l)$ ,  $f(x_1)$ , and  $f(x_u)$ , respectively. Since  $f(x_1) > f(x_l)$  and  $f(x_1) > f(x_u)$ , the maximum must lie between  $x_l$  and  $x_u$ . Now a fourth point denoted by  $x_2$  is chosen to be between the larger of the two intervals of  $[x_l, x_1]$  and  $[x_1, x_u]$ . Assuming that the interval  $[x_l, x_1]$  is larger than  $[x_1, x_u]$ , we would chose  $[x_l, x_1]$  as the interval in which  $x_2$  is chosen. If  $f(x_2) > f(x_1)$  then the new three points would be

$x_l < x_2 < x_1$ ; else if  $f(x_2) < f(x_1)$  then the new three points are  $x_2 < x_1 < x_u$ . This process is continued until the distance between the outer points is sufficiently small.

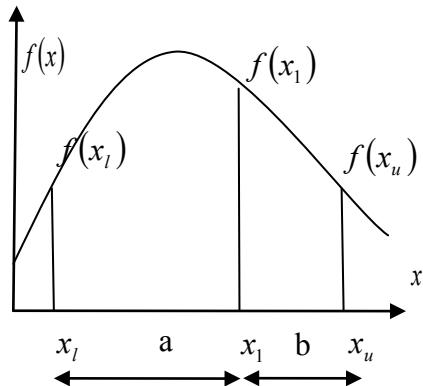


**Figure 3** Cross section of the gutter.

#### How are the intermediate points in the Golden Section Search determined?

We chose the first intermediate point  $x_l$  to equalize the ratio of the lengths as shown in Eq. (1) where  $a$  and  $b$  are distance as shown in Figure 4. Note that  $a+b$  is equal to the distance between the lower and upper boundary points  $x_l$  and  $x_u$ .

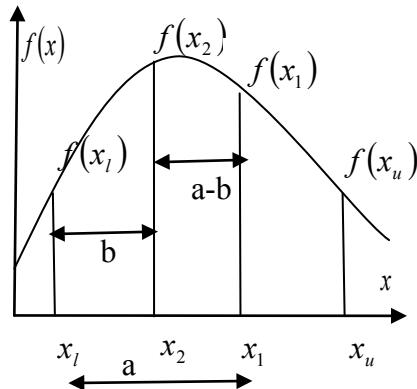
$$\frac{a}{a+b} = \frac{b}{a} \quad (1)$$



**Figure 4** Determining the first intermediate point

The second intermediate point  $x_2$  is chosen similarly in the interval  $a$  to satisfy the following ratio in Eq. (2) where the distances of  $a$  and  $b$  are shown in Figure 5.

$$\frac{b}{a} = \frac{a-b}{b} \quad (2)$$



**Figure 5** Determining the second intermediate point

### Does the Golden Section Search have anything to do with the Golden Ratio?

The ratios in Equations (1) and (2) are equal and have a special value known as the Golden Ratio. The Golden Ratio has been used since ancient times in various fields such as architecture, design, art and engineering. To determine the value of the Golden Ratio let  $R = a/b$ , then Eq. (1) can be written as

$$1 + R = \frac{1}{R}$$

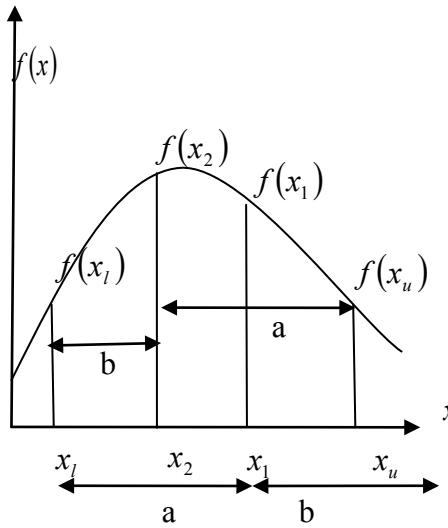
or

$$R^2 + R - 1 = 0 \quad (3)$$

Using the quadratic formula, the positive root of Eq. (3) is

$$\begin{aligned} R &= \frac{-1 + \sqrt{1 - 4(-1)}}{2} \\ &= \frac{\sqrt{5} - 1}{2} \\ &= 0.61803 \end{aligned} \quad (4)$$

In other words, the intermediate points  $x_1$  and  $x_2$  are chosen such that, the ratio of the distance from these points to the boundaries of the search region is equal to the golden ratio as shown in Figure 6.



**Figure 6** Intermediate points and their relation to boundary points

### What happens after choosing the first two intermediate points?

Next we determine a new and smaller interval where the maximum value of the function lies in. We know that the new interval is either  $[x_l, x_2, x_1]$  or  $[x_2, x_1, x_u]$ . To determine which of these intervals will be considered in the next iteration, the function is evaluated at the intermediate points  $x_2$  and  $x_1$ . If  $f(x_2) > f(x_1)$ , then the new region of interest will be  $[x_l, x_2, x_1]$ ; else if  $f(x_2) < f(x_1)$ , then the new region of interest will be  $[x_2, x_1, x_u]$ . In Figure 6, we see that  $f(x_2) > f(x_1)$ , therefore our new region of interest is  $[x_l, x_2, x_1]$ . We should point out that the boundaries of the new smaller region are now determined by  $x_l$  and  $x_1$ , and we already have one of the intermediate points, namely  $x_2$ , conveniently located at a point where the ratio of the distance to the boundaries is the Golden Ratio. All that is left to do is to determine the location of the second intermediate point. Can you determine if the second point will be closer to  $x_l$  or  $x_1$ ? This process of determining a new smaller region of interest and a new intermediate point will continue until the distance between the boundary points are sufficiently small.

### The Golden Section Search Algorithm

The following algorithm can be used to determine the maximum of a function  $f(x)$ .

#### Initialization:

Determine  $x_l$  and  $x_u$  which is known to contain the maximum of the function  $f(x)$ .

#### Step 1

Determine two intermediate points  $x_1$  and  $x_2$  such that

$$x_1 = x_l + d$$

$$x_2 = x_u - d$$

where

$$d = \frac{\sqrt{5}-1}{2}(x_u - x_l)$$

### Step 2

Evaluate  $f(x_1)$  and  $f(x_2)$ .

If  $f(x_1) > f(x_2)$ , then determine new  $x_l, x_1, x_2$  and  $x_u$  as shown in Equation set (5). Note that the only new calculation is done to determine the new  $x_1$ .

$$\begin{aligned} x_l &= x_2 \\ x_2 &= x_1 \\ x_u &= x_u \\ x_1 &= x_l + \frac{\sqrt{5}-1}{2}(x_u - x_l) \end{aligned} \tag{5}$$

If  $f(x_1) < f(x_2)$ , then determine new  $x_l, x_1, x_2$  and  $x_u$  as shown in Equation set (6). Note that the only new calculation is done to determine the new  $x_2$ .

$$\begin{aligned} x_l &= x_1 \\ x_u &= x_1 \\ x_1 &= x_2 \\ x_2 &= x_u - \frac{\sqrt{5}-1}{2}(x_u - x_l) \end{aligned} \tag{6}$$

### Step 3

If  $x_u - x_l < \varepsilon$  (a sufficiently small number), then the maximum occurs at  $\frac{x_u + x_l}{2}$  and stop iterating, else go to Step 2.

### **Further Remarks and Explanation About The Golden Section Search Algorithm**

The above discussion has assumed that the user can determine  $x_l$  and  $x_u$  which is known to contain the maximum of the function  $f(x)$ . In this section, the Golden Section algorithm is re-examined from a more rigorous viewpoint, and with the following 2 primary objectives:

- (a) Developing an automated procedure to determine the appropriated initial guesses for the lower and upper bounds, respectively.

- (b) Proving (in a more rigorous way) that we only needs to find/compute only 1 (not 2) intermediate point, based on the current bracket.

To start the Golden Section search process, a small (positive) parameter “ $\delta$ ” is defined by the user, say “ $\delta$ ” = 0.05. The function value  $g(\alpha = \delta) = g_1$  is initially computed. The second interval will be 1.618 times the previous (or first) interval (or  $1.618 * \delta$ ), therefore, the next computed function value  $g(\alpha = 2.618 * \delta) = g_2$  is computed.

Since  $g_2$  is smaller than the previous value  $g_1$  [see Figure 6A], one continues to consider the third interval which will be 1.618 times the second interval (or  $1.618 * 1.618 \delta = 1.618^2 \delta$ ), and the next computed function value  $g(\alpha = 5.232 * \delta) = g_3$  is computed. As indicated in Figure 6A,  $\alpha = (5.232 * \delta)$  is also labeled as the (j-2)-th point on the curve!

Since  $g_3$  is still smaller than the previous value  $g_2$  [see Figure 6A], one continues to consider the fourth interval which will be 1.618 times the third interval (or  $1.618 * 1.618^2 \delta = 1.618^3 \delta$ ) and the next computed function value  $g(\alpha = 9.468 * \delta) = g_4$  is computed. As indicated in Figure 6A,  $\alpha = (9.468 * \delta)$  is also labeled as the (j-1)-th point on the curve!

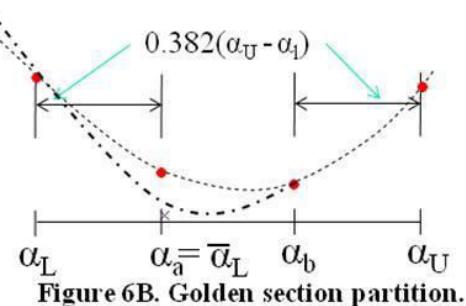
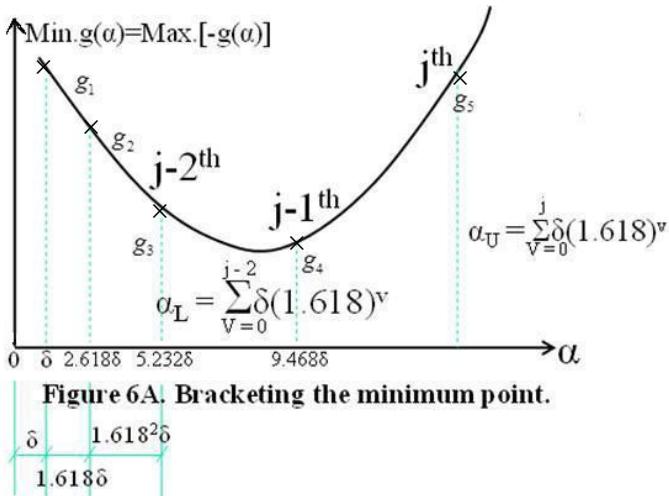
Since  $g_4$  is still smaller than the previous value  $g_3$  [see Figure 6A], one continues to consider the fifth interval which will be 1.618 times the fourth interval (or  $1.618 * 1.618^3 \delta = 1.618^4 \delta$ ), and the next computed function value  $g(\alpha = 16.3215 * \delta) = g_5$  is computed. As indicated in Figure 6A,  $\alpha = (16.3215 * \delta)$  is also labeled as the j-th point on the curve!

At this moment, since  $g_5 = g$ (at the j-th point) is larger than  $g_4 = g$ (at the j-1-th point), the “decreasing pattern” is no longer true, therefore, we can establish the initial lower bound and the initial upper bound to be equal to the values of  $\alpha$  at the (j-2)-th location and at the j-th location, respectively !

Based on the above observation and analysis, one can easily figure out the general formulas to compute and identify the initial lower and upper bound values for  $\alpha$  as indicated in Figure 6A.

Having found the initial lower and upper bounds for  $\alpha$ , the 2 intermediate points  $\alpha_a$  and  $\alpha_b$  need be inserted (with the same distance measured from the lower bound and upper bound, respectively) as shown in Figure 6B. Using Figure 6B,  $\alpha_a$  can be computed and displayed as shown in equation (7).

Finally, with trivial algebraic manipulations, the value for  $\alpha_a$  can be shown to be the same as the value for  $\alpha$  (at the j-1\_th point), as indicated in equation (8).



$$\alpha_a = \alpha_L + 0.382(\alpha_U - \alpha_L) = \sum_{v=0}^{j-2} \delta(1.618)^v + 0.382\delta(1.618)^{j-1}(1+1.618) \quad (7)$$

$$\alpha_a = \sum_{v=0}^{j-2} \delta(1.618)^v + 1\delta(1.618)^{j-1} = \sum_{v=0}^{j-1} \delta(1.618)^v = \text{already known!} \quad (8)$$

Based on Figures 6A and 6B, one observes that

- If  $g(\alpha_a) = g(\alpha_b)$  then the minimum will be between  $\alpha_a$  and  $\alpha_b$ .
- If  $g(\alpha_a) > g(\alpha_b)$  as shown in Figure 6B, then minimum will be between  $\alpha_a$  and  $\alpha_u$ .  
Hence,  $\overline{\alpha}_L = \text{new lower bound} = \alpha_a$
- Notice that:  $\overline{\alpha}_u - \overline{\alpha}_L = \alpha_u - \alpha_a = \delta(1.618)^j$

and

$$\begin{aligned} \alpha_b - \overline{\alpha}_L &= \alpha_b - \alpha_a = (1 - 2 \times 0.382)(\alpha_U - \alpha_L) = (0.236)(\delta[1.618]^{j-1} + \delta[1.618]^j) \\ &= (0.236)(\delta[1.618]^{j-1} \times [1+1.618]) = 0.618(\delta[1.618]^{j-1}) \times \frac{1.618}{1.618} \\ \alpha_b - \overline{\alpha}_L &= (0.382) \times (\delta[1.618]^j) = 0.382(\overline{\alpha}_U - \overline{\alpha}_L) \end{aligned}$$

Thus  $\alpha_b$  (with respect to  $\alpha_u$  and  $\alpha_L$ ) plays the same role as  $\alpha_a$  (with respect to  $\alpha_u$  and  $\alpha_L$ )!!  
The step-by-step Golden Section procedure can be summarized as:

### Step 1:

For a chosen small step size  $\delta$  in  $\alpha$  say,  $\delta = 0.05$ , let  $j$  be the smallest integer such that

$$g\left(\sum_{V=0}^j \delta(1.618)^V\right) g\left(\sum_{V=0}^{j-1} \delta(1.618)^V\right)$$

The upper and lower bound on  $\alpha^i$  are  $\alpha_U = \sum_{V=0}^j \delta(1.618)^V$  and  $\alpha_L = \sum_{V=0}^{j-2} \delta(1.618)^V$ .

### Step 2:

Compute  $g(\alpha_b)$ , where  $\alpha_a = \alpha_L + 0.382(\alpha_U - \alpha_L)$ , and  $\alpha_b = \alpha_L + 0.618(\alpha_U - \alpha_L)$

Note that  $\alpha_a = \sum_{V=0}^{j-1} \delta(1.618)^V$ , so  $g(\alpha_a)$  is already known.

### Step 3:

Compare  $g(\alpha_a)$  and  $g(\alpha_b)$  and go to Step 4, 5, or 6.

### Step 4:

If  $g(\alpha_a) < g(\alpha_b)$ , then  $\alpha_L \leq \alpha^i \leq \alpha_b$ . By choice of  $\alpha_a$  and  $\alpha_b$ , the new points  $\overline{\alpha}_L = \alpha_L$  and  $\overline{\alpha}_u = \alpha_b$  have  $\overline{\alpha}_b = \alpha_a$ .

Compute  $g(\overline{\alpha}_a)$ , where  $\overline{\alpha}_a = \overline{\alpha}_L + 0.382(\overline{\alpha}_u - \overline{\alpha}_L)$  and go to Step 7.

### Step 5:

If  $g(\alpha_a) > g(\alpha_b)$ , then  $\alpha_a \leq \alpha^i \leq \alpha_u$ . Similar to the procedure in Step 4, put  $\overline{\alpha}_L = \alpha_a$  and  $\overline{\alpha}_u = \alpha_b$ .

Compute  $g(\overline{\alpha}_b)$ , where  $\overline{\alpha}_b = \overline{\alpha}_L + 0.618(\overline{\alpha}_u - \overline{\alpha}_L)$  and go to Step 7.

### Step 6:

If  $g(\alpha_a) = g(\alpha_b)$  put  $\alpha_L = \alpha_a$  and  $\alpha_u = \alpha_b$  and return to Step 2.

### Step 7:

If  $\overline{\alpha}_u - \overline{\alpha}_L$  is suitably small, put  $\alpha^i = \frac{1}{2}(\overline{\alpha}_u + \overline{\alpha}_L)$  and stop.

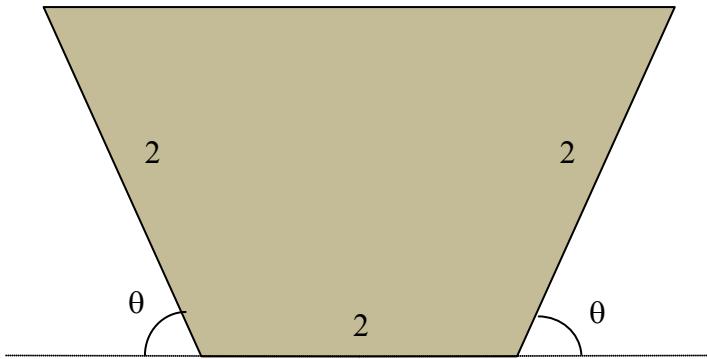
Otherwise, delete the bar symbols on  $\overline{\alpha}_L, \overline{\alpha}_a, \overline{\alpha}_b$  and  $\overline{\alpha}_u$  and return to Step 3.

## Example 2

Consider Figure 7 below. The cross-sectional area  $A$  of a gutter with equal base and edge length of 2 is given by

$$A = 4 \sin \theta (1 + \cos \theta)$$

Find the angle  $\theta$  which maximizes the cross-sectional area of the gutter. Using an initial interval of  $[0, \pi/2]$ , find the solution after 2 iterations. Use an initial  $\varepsilon = 0.05$ .

**Figure 7** Cross section of the gutter**Solution**

The function to be maximized is

$$f(\theta) = 4 \sin \theta (1 + \cos \theta)$$

Iteration 1:

Given the values for the boundaries of  $x_l = 0$  and  $x_u = \pi/2$ , we can calculate the initial intermediate points as follows:

$$x_1 = x_l + \frac{\sqrt{5}-1}{2}(x_u - x_l)$$

$$= 0 + \frac{\sqrt{5}-1}{2}(1.5708)$$

$$= 0.97080$$

$$x_2 = x_u - \frac{\sqrt{5}-1}{2}(x_u - x_l)$$

$$= 1.5708 - \frac{\sqrt{5}-1}{2}(1.5708)$$

$$= 0.60000$$

The function is evaluated at the intermediate points as  $f(0.9708) = 5.1654$  and  $f(0.60000) = 4.1227$ . Since  $f(x_1) > f(x_2)$ , we eliminate the region to the left of  $x_2$  and update the lower boundary point as  $x_l = x_2$ . The upper boundary point  $x_u$  remains unchanged. The second intermediate point  $x_2$  is updated to assume the value of  $x_1$  and finally the first intermediate point  $x_1$  is re-calculated as follows:

$$\begin{aligned}
 x_1 &= x_l + \frac{\sqrt{5}-1}{2}(x_u - x_l) \\
 &= 0.60000 + \frac{\sqrt{5}-1}{2}(1.5708 - 0.60000) \\
 &= 1.2000
 \end{aligned}$$

To check the stopping criteria the difference between  $x_u$  and  $x_l$  is calculated to be

$$x_u - x_l = 1.5708 - 0.60000 = 0.97080$$

which is greater than  $\varepsilon = 0.05$ . The process is repeated in the second iteration.

#### Iteration 2:

The values for the boundary and intermediate points used in this iteration were calculated in the previous iteration as shown below.

$$\begin{aligned}
 x_l &= 0.60000 \\
 x_u &= 1.5708 \\
 x_1 &= 1.2000 \\
 x_2 &= 0.97080
 \end{aligned}$$

Again the function is evaluated at the intermediate points as  $f(1.20000) = 5.0791$  and  $f(0.97080) = 5.1654$ . Since  $f(x_1) < f(x_2)$ , the opposite of the case seen in the first iteration, we eliminate the region to the right of  $x_1$  and update the upper boundary point as  $x_u = x_1$ . The lower boundary point  $x_l$  remains unchanged. The first intermediate point  $x_1$  is updated to assume the value of  $x_2$  and finally the second intermediate point  $x_2$  is re-calculated as follows:

$$\begin{aligned}
 x_2 &= x_u - \frac{\sqrt{5}-1}{2}(x_u - x_l) \\
 &= 1.2000 - \frac{\sqrt{5}-1}{2}(1.2000 - 0.60000) \\
 &= 0.82918
 \end{aligned}$$

To check the stopping criteria the difference between  $x_u$  and  $x_l$  is calculated to be

$$\begin{aligned}
 x_u - x_l &= 1.2000 - 0.60000 \\
 &= 0.60000
 \end{aligned}$$

which is greater than  $\varepsilon = 0.05$ . At the end of the second iteration the solution is

$$\frac{x_u + x_l}{2} = \frac{1.2000 + 0.60000}{2} \\ = 0.90000$$

Therefore, the maximum area occurs when  $\theta = 0.9$  radians or  $51.6^\circ$ .

The iterations will continue until the stopping criterion is met. Summary results of all the iterations are shown in Table 1. Note that at the end of the 9th iteration,  $\varepsilon < 0.05$  which causes the search to stop. The optimal value is calculated as the average of the upper and lower boundary points.

$$\frac{x_u + x_l}{2} = \frac{1.0249 + 1.0583}{2} \\ = 1.0416$$

which is about  $59.68^\circ$ . The area of the gutter at this angle is  $f(1.0416) = 5.1960$ . The theoretical optimal solution to the problem happens at exactly  $60^\circ$  which is 1.0472 radians and an area of 5.1962.

**Table 1** Summary of iterations for Example 1

Iteration	$x_l$	$x_u$	$x_1$	$x_2$	$f(x_1)$	$f(x_2)$	$\varepsilon$
1	0.00000	1.5708	0.97081	0.59999	5.1654	4.1226	1.5708
2	0.59999	1.5708	1.2000	0.97081	5.0791	5.1654	0.97081
3	0.59999	1.2000	0.97081	0.82917	5.1654	4.9418	0.59999
4	0.82917	1.2000	1.0583	0.97081	5.1955	5.1654	0.37081
5	0.97081	1.2000	1.1124	1.0583	5.1743	5.1955	0.22918
6	0.97081	1.1124	1.0583	1.0249	5.1955	5.1936	0.14164
7	1.0249	1.1124	1.0790	1.0583	5.1909	5.1955	0.08754
8	1.0249	1.0790	1.0583	1.0456	5.1955	5.1961	0.05410
9	1.0249	1.0583	1.0456	1.0377	5.1961	5.1957	0.03344

---

## OPTIMIZATION

---

Topic	Golden Search Method
Summary	Textbook notes for the golden search method
Major	All engineering majors
Authors	Ali Yalcin, Autar Kaw
Date	December 19, 2012
Web Site	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

---

## Chapter 09.02

# Newton's Method

After reading this chapter, you should be able to:

1. Understand how Newton's method is different from the Golden Section Search method
2. Understand how Newton's method works
3. Solve one-dimensional optimization problems using Newton's method

### How is the Newton's method different from the Golden Section Search method?

The Golden Section Search method requires explicitly indicating lower and upper boundaries for the search region in which the optimal solution lies. Such methods where the boundaries need to be specified are known as bracketing approaches in the sense that the optimal solution is bracketed by these boundaries.

Newton's method is an open (instead of bracketing) approach, where the optimum of the one-dimensional function  $f(x)$  is found using an initial guess of the optimal value without the need for specifying lower and upper boundary values for the search region.

Unlike the bracketing approaches, open approaches are not guaranteed to converge. However, if they do converge, they do so much faster than bracketed approaches. Therefore, open approaches are more useful if there is reasonable evidence that the initial guess is close to the optimal value. Otherwise, if there is doubt about the quality of the initial guess, it is advisable to use bracketing approaches to bring the guess closer to the optimal value and then use an open approach benefiting from the advantages presented by both techniques.

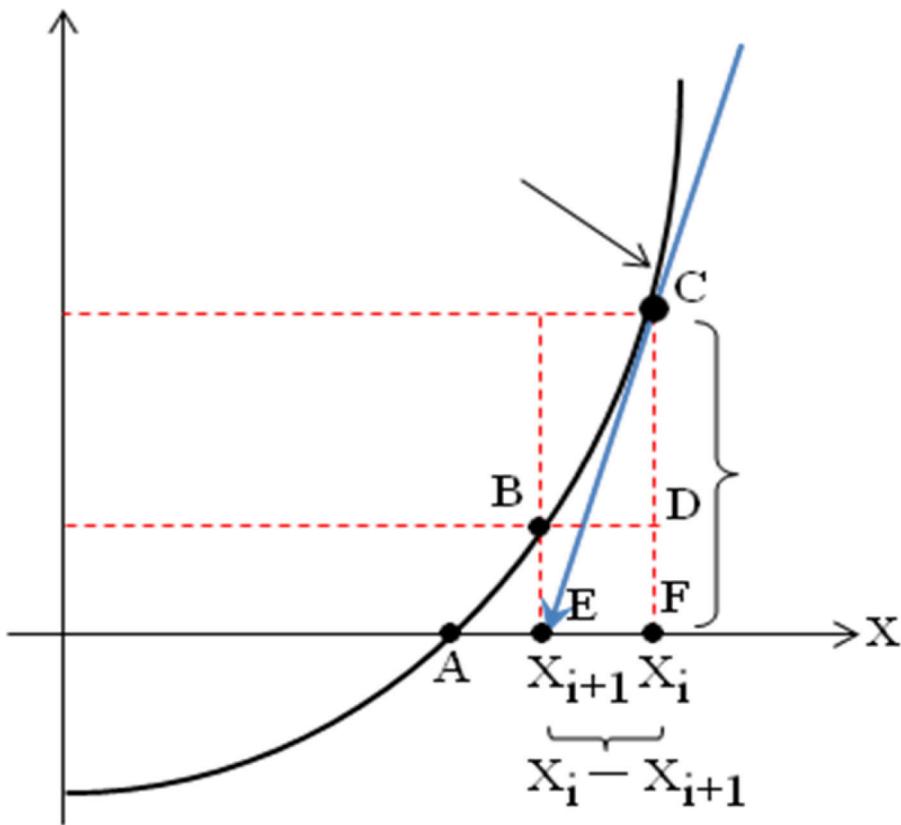
### What is the Newton's method and how does it work?

Newton's method is an open approach to find the minimum or the maximum of a function  $f(x)$ . It is very similar to the Newton-Raphson method [http://numericalmethods.eng.usf.edu/topics/newton\\_raphson.html](http://numericalmethods.eng.usf.edu/topics/newton_raphson.html) to find the roots of a function such that  $f(x)=0$ . Since the derivative of the function  $f(x)$ ,  $f'(x)=0$  at the functions maximum and minimum, the minima and the maxima can be found by applying the Newton-Raphson method to the derivative, essentially obtaining

$$x_{i+1} = x_i - \frac{f'(x_i)}{f''(x_i)} \quad (1)$$

We caution that before using Newton's method to determine the minimum or the maximum of a function, one should have a reasonably good estimate of the solution to ensure convergence, and that the function should be easily twice differentiable.

### Derivation of the Newton-Raphson Equation



Slope at point

$$C \approx \frac{F(X_i) - F(X_{i+1})}{X_i - X_{i+1}}$$

We "wish" that in the next iteration  $X_{i+1}$  will be the root, or  $F(X_{i+1}) = 0$ .

Thus:

Slope at point

$$C = \frac{F(X_i) - 0}{X_i - X_{i+1}}$$

or

$$F'(X_i) = \frac{F(X_i)}{X_i - X_{i+1}}$$

Hence :

$$X_{i+1} = X_i - \frac{F(X_i)}{F'(X_i)}$$

### Remarks:

1. If  $F(X) \equiv f'(X)$ , then  $X_{i+1} = X_i - \frac{f'(X_i)}{f''(X_i)}$
2. For Multi-variable case, then NR method becomes

$$\vec{X}_{i+1} = \vec{X}_i - [f''(\vec{X}_i)]^{-1} \times \nabla \vec{f}(\vec{X}_i)$$

### Step by step use of Newton's method

The following algorithm implements Newton's method to determine the maximum or minimum of a function  $f(x)$ .

#### Initialization

Determine a reasonably good estimate  $x_0$  for the maxima or the minima of the function  $f(x)$ .

#### Step 1

Determine  $f'(x)$  and  $f''(x)$ .

#### Step 2

Substitute  $x_{i+1}$ , the initial estimate  $x_0$  for the first iteration,  $f'(x)$  and  $f''(x)$  into Eqn. 1 to determine  $x_i$  and the function value in iteration  $i$ .

#### Step 3

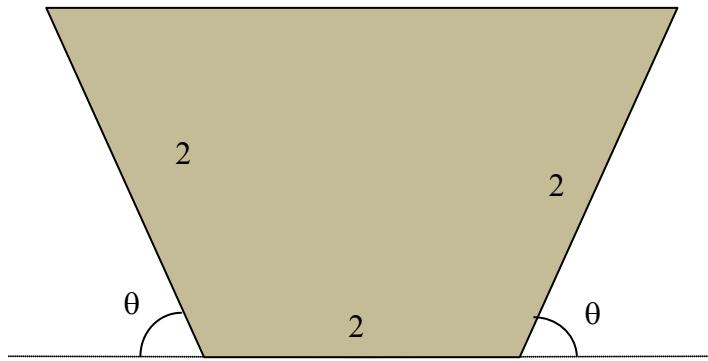
If the value of the first derivative of the function is zero, then you have reached the optimum (maxima or minima), otherwise repeat Step 2 with the new value of  $x_i$  until the absolute relative approximate error is less than the pre-specified tolerance.

### Example 1

Consider Figure 1 below. The cross-sectional area  $A$  of a gutter with equal base and edge length of 2 is given by

$$A = 4 \sin \theta (1 + \cos \theta)$$

Find the angle  $\theta$  which maximizes the cross-sectional area of the gutter.



**Figure 1:** Cross section of the gutter

### Solution

The function to be maximized is  $f(\theta) = 4 \sin \theta (1 + \cos \theta)$ . The first and second derivative of the function is shown below.

$$f'(\theta) = 4(\cos \theta + \cos^2 \theta - \sin^2 \theta)$$

$$f''(\theta) = -4 \sin \theta (1 + 4 \cos \theta)$$

Let us use  $\theta_0 = \pi/4$  as the initial estimate of  $\theta$ . Using Eqn. (1), we can calculate the first iteration follows:

$$\underline{i = 0}$$

$$\begin{aligned} \theta_1 &= \theta_0 - \frac{f'(\theta_0)}{f''(\theta_0)} \\ &= \frac{\pi}{4} - \frac{f'\left(\frac{\pi}{4}\right)}{f''\left(\frac{\pi}{4}\right)} \\ &= \frac{\pi}{4} - \frac{4\left(\cos \frac{\pi}{4} + \cos^2 \frac{\pi}{4} - \sin^2 \frac{\pi}{4}\right)}{-4 \sin \frac{\pi}{4} (1 + 4 \cos \frac{\pi}{4})} \\ &= 1.0466 \end{aligned}$$

The function is evaluated at the first estimate as  $f(1.0466) = 5.1962$ . The next iteration uses  $\theta_1 = 1.0466$  as the best estimate of  $\theta$ . Using Eqn(1) again, the second iteration is calculated as follows:

$$\underline{i = 1}$$

$$\begin{aligned}
 \theta_2 &= \theta_1 - \frac{f'(\theta_1)}{f''(\theta_1)} \\
 &= 1.0466 - \frac{f'(1.0466)}{f''(1.0466)} \\
 &= 1.0466 - \frac{4(\cos 1.0466 + \cos^2 1.0466 - \sin^2 1.0466)}{-4 \sin 1.0466(1 + 4 \cos 1.0466)} \\
 &= 1.0472
 \end{aligned}$$

The iterations will continue until the solution converges to a single optimal solution. Summary results of all the iterations are shown in Table 1.

Several important observations regarding the 5th iteration can be made. At each iteration, the magnitude of the first derivative gets smaller and approaches zero. A value of zero of the first derivative tells us that we have reached the optimal and we can stop. Also note that the sign of the second derivative is negative which tells us that we are at a maximum. This value would have been positive if we had reached a minimum. The solution tells us that the optimal angle is 1.0472. Remember that the actual solution to the problem is at 60 degrees or 1.0472 radians. See Example 2 in Golden Search Method [http://numericalmethods.eng.usf.edu/topics/opt\\_goldensearch.html](http://numericalmethods.eng.usf.edu/topics/opt_goldensearch.html).

**Table 1.** Summary of iterations for Example 1

Iteration	$\theta_i$	$f'(\theta_i)$	$f''(\theta_i)$	$\theta_{i+1}$	$f(\theta_{i+1})$
1	0.78540	2.8284	-10.828	1.0466	5.1962
2	1.0466	0.0061898	-10.396	1.0472	5.1962
3	1.0472	1.0613E-06	-10.392	1.0472	5.1962
4	1.0472	3.0642E-14	-10.392	1.0472	5.1962
5	1.0472	1.3323E-15	-10.392	1.0472	5.1962

---

## OPTIMIZATION

---

Topic	Newton's Method
Summary	Textbook notes for the Newton's method
Major	All engineering majors
Authors	Ali Yalcin
Date	August 17, 2011
Web Site	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

---

# **Chapter 09.03**

## **Multidimensional Direct Search Method**

*After reading this chapter, you should be able to:*

1. *Understand the fundamentals of the multidimensional direct search methods*
2. *Understand how the coordinate cycling search method works*
3. *Solve multi-dimensional optimization problems using the coordinate cycling search method*

### **Optimization Techniques**

Methods for finding optimal solutions in multidimensional spaces are not too different than their cousins used in finding optimal solutions in a single dimension. The trade-off between general applicability versus computational complexity also exists in multidimensional optimization. The multidimensional direct search methods we will cover in this chapter, like the one-dimensional Golden Section Search method ([http://numericalmethods.eng.usf.edu/topics/opt\\_goldensearch.html](http://numericalmethods.eng.usf.edu/topics/opt_goldensearch.html)), does not require a differentiable function. These methods are sometimes referred to as Zeroth Order Algorithms because it is not required to differentiate the optimization function.

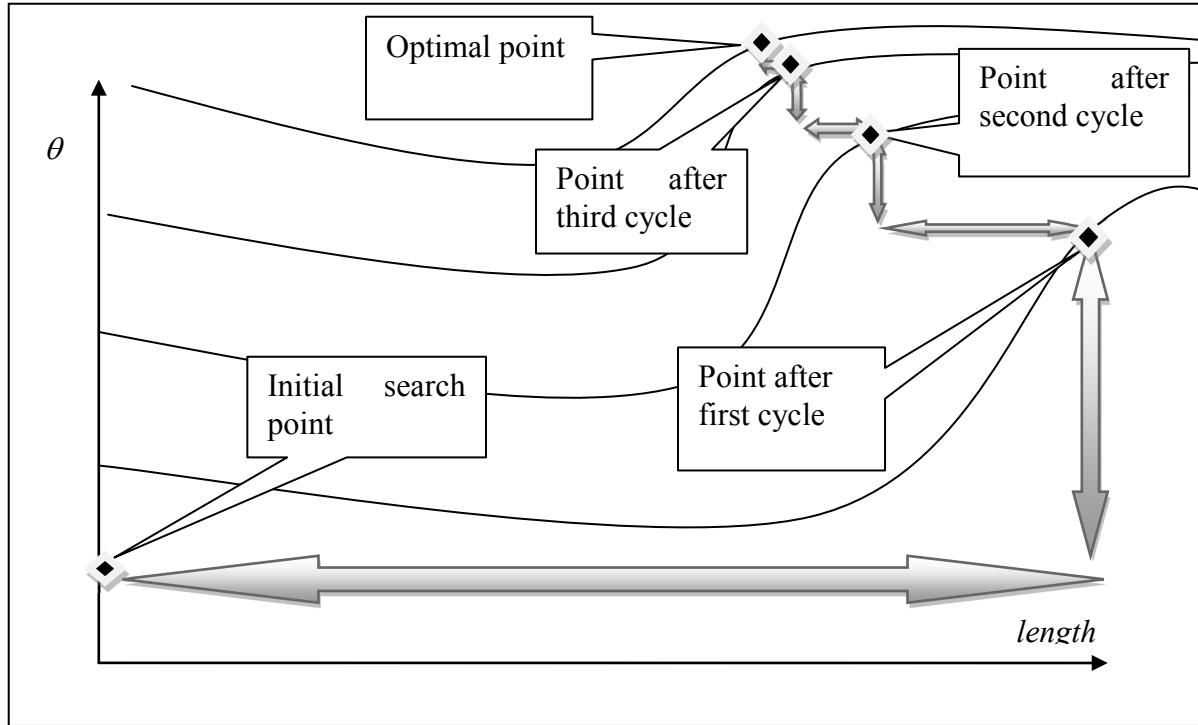
Probably the most obvious solution to an optimization problem in multidimensional space is to systematically evaluate every possible solution and select the maximum or the minimum depending on our objective. This is a very generally applicable approach and may even be useful if the solution space is relatively small. However, as the dimensions of the problem space, (number of independent variables), increase, the computational complexity of this solution approach quickly becomes unmanageable. Therefore, we are interested in methods that intelligently search through the solution space to find an optimal solution without enumerating all possible solutions.

It is important to note that some of the popular optimization techniques you may have heard of such as simulated annealing, tabu search, neural networks and genetic algorithms all fall under this family of optimization techniques.

### **What is the Coordinate Cycling Search Method and How Does it Work?**

The coordinate cycling search method, starts from an initial point and looks for an optimal solution along each coordinate direction iteratively. For example, using a function  $f(x,y)$  with two independent variables  $x$  and  $y$ , and starting at point  $(x_0, y_0)$ ; the first iteration will move along direction  $(1, 0)$ , until an optimal solution is found for the function  $f(x, y_0)$ . The next search involves searching along the direction  $(0, 1)$  to determine the optimal value for the function  $f(x_1, y)$  where  $x_1$  is the solution found in the previous search. Once searches in all directions are completed, the process is repeated in the next cycle. The search will

continue until convergence occurs or a predetermined error limit is met. The search along each coordinate direction can be conducted by using anyone of the one-dimensional search techniques previously covered. A visual representation of how the search converges is shown below in Figure 1.



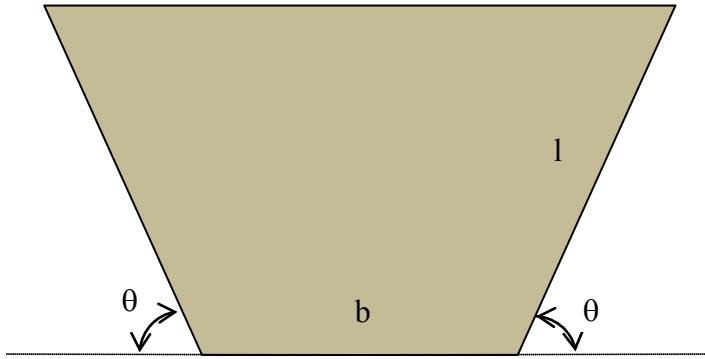
**Figure 1** Visual Representation of a Multidimensional Search

### Example 1

Consider Figure 2 below. The cross-sectional area  $A$  of a gutter with a base length  $b$  and an edge length of  $l$  is given by

$$A = \frac{1}{2}(b + b + 2l \cos \theta)l \sin \theta$$

Assuming that the width of the material to be bent into the gutter shape is 6 inches, find the angle  $\theta$  and edge length  $l$  which maximizes the cross-sectional area of the gutter.

**Figure 2** Cross section of the gutter**Solution**

Recognizing that the base length  $b$  can be expressed as  $b = 6 - 2l$ , we can re-write the area function to be optimized in terms of two independent variables giving

$$f(l, \theta) = (6 - 2l + l \cos \theta)l \sin \theta.$$

Let us consider an initial point  $(0, \frac{\pi}{6})$ . We will use the Golden Section Search method to determine the optimal solution along direction  $(1,0)$  namely the independent variable corresponding to the length of each side. To use the Golden Section Search method, we will use 0 and 3 as the lower and upper bounds, respectively for the search region (Can you determine why we are using 3 as the upper bound?) and look for the optimal solution of the function  $f(l, 0.52360)$  with a convergence limit of  $\varepsilon < 0.05$ . Table 1 below shows the iterations of the Golden Section Search method in the  $(1,0)$  direction. The maximum area of  $3.6964 \text{ in}^2$  is obtained at point  $(2.6459, 0.52360)$ .

**Table 1** Summary of the Golden Section Search iterations along direction  $(1,0)$  for Example 1. Here  $\theta = 0.52360$  and  $f(x_i) = (6 - 2l + l \cos(0.52360))l \sin(0.52360)$ )

Iteration	$x_l$	$x_u$	$x_1$	$x_2$	$f(x_1)$	$f(x_2)$	$\varepsilon$
1	0.0000	3.0000	1.8541	1.1459	3.6143	2.6941	3.0000
2	1.1459	3.0000	2.2918	1.8541	3.8985	3.6143	1.8541
3	1.8541	3.0000	2.5623	2.2918	3.9655	3.8985	1.1459
4	2.2918	3.0000	2.7295	2.5623	3.9654	3.9655	0.7082
5	2.2918	2.7295	2.5623	2.4590	3.9655	3.9497	0.4377
6	2.4590	2.7295	2.6262	2.5623	3.9692	3.9655	0.2705
7	2.5623	2.7295	2.6656	2.6262	3.9692	3.9692	0.1672
8	2.5623	2.6656	2.6262	2.6018	3.9692	3.9683	0.1033
9	2.6018	2.6656	2.6412	2.6262	3.9694	3.9692	0.0639
10	2.6262	2.6656	2.6506	2.6412	3.9694	3.9694	0.0395

To search along the  $(0,1)$  direction corresponding to the angle  $\theta$ , we again use the Golden Section Search method, but in this case using the function  $f(2.6459, \theta)$ . Table 2 below shows the iterations of the Golden Section Search method in the  $(0,1)$  direction. Note that at the new optimal point  $(2.6459, 0.8668)$ , the approximation of the maximum area is improved to  $4.8823 \text{ in}^2$ .

**Table 2** Summary of the Golden Section Search iterations along direction  $(0,1)$ . Here  $l = 2.6459$  and  $f(x_i) = (6 - 2 \times 2.6549 + 2.6549 \times \cos \theta) \times 2.6549 \times \sin \theta$

Iteration	$x_l$	$x_u$	$x_1$	$x_2$	$f(x_1)$	$f(x_2)$	$\varepsilon$
1	0.0000	1.5714	0.9712	0.6002	4.8084	4.3215	1.5714
2	0.6002	1.5714	1.2005	0.9712	4.1088	4.8084	0.9712
3	0.6002	1.2005	0.9712	0.8295	4.8084	4.8689	0.6002
4	0.6002	0.9712	0.8295	0.7419	4.8689	4.7533	0.3710
5	0.7419	0.9712	0.8836	0.8295	4.8816	4.8689	0.2293
6	0.8295	0.9712	0.9171	0.8836	4.8672	4.8816	0.1417
7	0.8295	0.9171	0.8836	0.8630	4.8816	4.8820	0.0876
8	0.8295	0.8836	0.8630	0.8502	4.8820	4.8790	0.0541
9	0.8502	0.8836	0.8708	0.8630	4.8826	4.8820	0.0334

After completing these two iterations, we use the optimal point to start a new cycle. Table 3 shows the first set of iterations for the second cycle.

**Table 3** Summary of the Golden Section Search iterations along direction  $(1,0)$

Iteration	$x_l$	$x_u$	$x_1$	$x_2$	$f(x_1)$	$f(x_2)$	$\varepsilon$
1	0.0000	3.0000	1.8541	1.1459	4.9354	3.8871	3.0000
2	1.1459	3.0000	2.2918	1.8541	5.0660	4.9354	1.8541
3	1.8541	3.0000	2.5623	2.2918	4.9491	5.0660	1.1459
4	1.8541	2.5623	2.2918	2.1246	5.0660	5.0627	0.7082
5	2.1246	2.5623	2.3951	2.2918	5.0391	5.0660	0.4377
6	2.1246	2.3951	2.2918	2.2279	5.0660	5.0715	0.2705
7	2.1246	2.2918	2.2279	2.1885	5.0715	5.0708	0.1672
8	2.1885	2.2918	2.2523	2.2279	5.0704	5.0715	0.1033
9	2.1885	2.2523	2.2279	2.2129	5.0715	5.0716	0.0639
10	2.1885	2.2279	2.2129	2.2035	5.0716	5.0714	0.0395

Here  $\theta = 0.8668$  and  $f(x_i) = (6 - 2l + l \cos(0.8668)/\sin(0.8668))$ . Note that we still use the initial intervals chosen for  $x_i$  and  $x_u$  values throughout the cycles.

Since this is a two-dimensional search problem, the two searches along the two dimensions completes the first cycle. In the next cycle, we return to the first dimension for which we conducted a search, namely  $l$ , and start the second cycle with a search along this dimension.

Namely, look for the optimal solution of the function  $f(l,0.8668)$ . Each cycle consists of enough iterations to satisfy the predetermined convergence limit.

After the fifth cycle, the optimal solution of  $(2.0016, 1.0420)$  with an area of  $5.1960 \text{ in}^2$  is obtained. The optimal solution to the problem happens at exactly  $60^\circ$  which is  $1.0472$  radians, having an edge and base length of  $2 \text{ in}$ . The area of the gutter at this point is  $5.1962 \text{ in}^2$ . Therefore folding the sheet metal in such a way that the base is  $2 \text{ in}$  and the sides are  $2 \text{ in}$  at an angle of  $60^\circ$  maximizes the area of the gutter.

---

## OPTIMIZATION

---

Topic	Multidimensional Direct Search Method
Summary	Textbook notes for the multidimensional direct search method
Major	All engineering majors
Authors	Ali Yalcin
Date	December 19, 2012
Web Site	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

---

# **Chapter 09.04**

## **Multidimensional Gradient Method**

*After reading this chapter, you should be able to:*

1. *Understand how multi-dimensional gradient methods are different from direct search methods*
2. *Understand the use of first and second derivatives in multi-dimensions*
3. *Understand how the steepest ascent/descent method works*
4. *Solve multi-dimensional optimization problems using the steepest ascent/descent method*

### **How do gradient methods differ from direct search methods in multi-dimensional optimization?**

The difference between gradient and direct search methods in multi-dimensional optimization is similar to the difference between these approaches in one-dimensional optimization. Direct search methods are useful when the derivative of the optimization function is not available to effectively guide the search for the optimum. While direct search methods explore the parameter space in a systematic manner, they are not computationally very efficient. On the other hand, gradient methods use information from the derivatives of the optimization function to more effectively guide the search and find optimum solutions much quicker.

### **Newton's Method**

When Newton's Method

([http://numericalmethods.eng.usf.edu/topics/opt\\_newtons\\_method.html](http://numericalmethods.eng.usf.edu/topics/opt_newtons_method.html)) was introduced as a one-dimensional optimization method, we discussed the use of the first and second derivative of the optimization function as sources of information to determine if we have reached an optimal point (where the value of the first derivative is zero). If that optimal point is a maximum, the second derivative is negative. If the point is a minimum, the second derivative is positive.

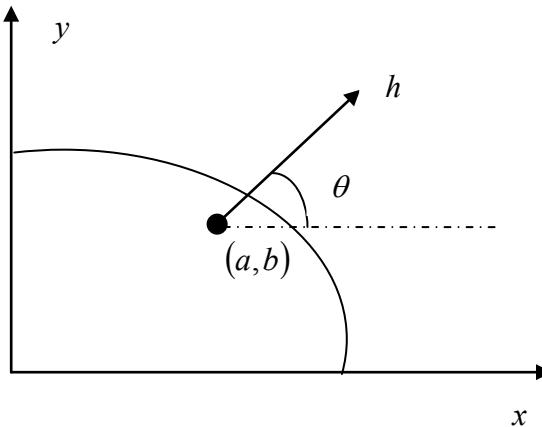
## What are Gradients and Hessians and how are Gradients and Hessians used in multi-dimensional optimization?

Gradients and Hessians describe the first and second derivatives of functions, respectively in multiple dimensions and are used frequently in various gradient methods for multi-dimensional optimization. We describe these two concepts of gradients and Hessians next.

### Gradient:

The gradient is a vector operator denoted by  $\nabla$  (referred to as “del”) which, when applied to a function  $f$ , represents its directional derivatives. For example, consider a two dimensional function  $f(x, y)$  which shows elevation above sea level at points  $x$  and  $y$ . If you wanted to move in the direction that would gain you the most elevation, this direction could be defined along a direction  $h$  which forms an angle  $\theta$  with the  $x$ -axis. For an illustration of this, see Figure 1. The elevation along this new axis can be described by a new function  $g(h)$  where your current location is the origin of the new coordinate axis or  $h=0$ . The slope in this direction can be calculated by taking the derivative of the new function  $g(h)$  at this point, namely  $g'(0)$ . The slope is then calculated by

$$g'(0) = \frac{\partial f}{\partial x} \cos \theta + \frac{\partial f}{\partial y} \sin \theta$$



**Figure 1.** Determining elevation along a new axis

The gradient is a special case where the direction of the vector gains the most elevation, or has the steepest ascent. If the goal was to decrease elevation, then this would be termed as the steepest descent.

The gradient of  $f(x, y)$  or  $\nabla f$  is the vector pointing in the direction of the steepest slope at that point. The gradient is calculated by

$$\nabla f = \frac{\partial f}{\partial x} \mathbf{i} + \frac{\partial f}{\partial y} \mathbf{j} \quad (1)$$

**Example 1**

Calculate the gradient to determine the direction of the steepest slope at point (2, 1) for the function  $f(x, y) = x^2 y^2$ .

**Solution**

To calculate the gradient; the partial derivatives must be evaluated as

$$\begin{aligned}\frac{\partial f}{\partial x} &= 2xy^2 \\ &= 2(2)(1)^2 \\ &= 4 \\ \frac{\partial f}{\partial y} &= 2x^2y \\ &= 2(2)^2(1) \\ &= 8\end{aligned}$$

which are used to determine the gradient at point (2,1) as

$$\nabla f = 4\mathbf{i} + 8\mathbf{j}$$

Traveling along this direction, we would gain elevation equal to the magnitude of the gradient which is  $\|\nabla f\| = \sqrt{4^2 + 8^2} = 8.94$ . Note that there is no other direction along which we can move to increase the slope.

**Hessians:**

The Hessian matrix or just the Hessian, is the Jacobian Matrix of the second-order partial derivatives of a function. For example, in a two dimensional function the Hessian matrix is simply

$$H = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix} \quad (2)$$

The determinant of the Hessian matrix is also referred to as the Hessian.

**Example 2**

Calculate the Hessian at point (2, 1) for the function  $f(x, y) = x^2 y^2$ .

**Solution**

To calculate the Hessian, we would need to calculate

$$\begin{aligned}\frac{\partial^2 f}{\partial^2 x^2} &= 2y^2 \\ &= 2(1)^2 \\ &= 2\end{aligned}$$

$$\frac{\partial^2 f}{\partial y^2} = 2x^2$$

$$= 2(2)^2$$

$$= 8$$

$$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x}$$

$$= 4xy$$

$$= 4(2)(1)$$

$$= 8$$

and the resulting Hessian matrix is

$$H = \begin{bmatrix} 2 & 8 \\ 8 & 8 \end{bmatrix}$$

Based on your knowledge of how second derivatives are used in one-dimensional optimization, you may guess that the value of the second derivatives in multi-dimensional optimization will tell us if we are at a maxima or minima. In multi-dimensional optimization, the Hessian of the optimization function contains the information of the second derivatives of the function, and is used to make such a determination. The determinant of the Hessian matrix denoted by  $|H|$  can have three cases:

1. If  $|H| > 0$  and  $\partial^2 f / \partial x^2 > 0$  then  $f(x, y)$  is a local minimum.
2. If  $|H| > 0$  and  $\partial^2 f / \partial x^2 < 0$  then  $f(x, y)$  is a local maximum.
3. If  $|H| < 0$  then  $f(x, y)$  is a saddle point.

Referring to Example 2, since  $|H| = -48 < 0$ , then point  $(2, 1)$  is a saddle point.

Hessians are also used in determining search trajectories in more advanced multi-dimensional gradient search techniques such as the Marquardt Method which is beyond the scope of this module.

### What is the Steepest Ascent (or Descent) method and how does it work?

In any multi-dimensional optimization algorithm, there are two key questions to be asked when searching for an optimal solution. The first one is about the direction of travel. The concept of gradients provides the answers to this question, at least in the short term. The second question is how long one should pursue a solution in this direction before reevaluating an alternative direction. If a re-evaluation strategy is executed too often, it increases computational costs. If it is executed too infrequently, one may end up pursuing a direction that may take the search away from the optimal solution. The steepest ascent algorithm proposes a simple solution to the second question by arbitrarily choosing a step size  $h$ . The special case where  $h = h^*$  is referred to as the optimal steepest descent where  $h^*$  brings us to the local maximum along the direction of the gradient. Consider the following

example which illustrates the application of the optimal steepest ascent method to a multi-dimensional optimization problem.

### Example 3

Determine the minimum of the function  $f(x, y) = x^2 + y^2 + 2x + 4$ . Use the point  $(2,1)$  as the initial estimate of the optimal solution.

#### Solution

##### Iteration 1:

To calculate the gradient; the partial derivatives must be evaluated as

$$\frac{\partial f}{\partial x} = 2x + 2 = 2(2) + 2 = 6$$

$$\frac{\partial f}{\partial y} = 2y = 2(1) = 2$$

which are used to determine the gradient at point  $(2,1)$  as

$$\nabla f = 6\mathbf{i} + 2\mathbf{j}$$

Now the function  $f(x, y)$  can be expressed along the direction of gradient as

$$f\left(x_0 + \frac{\partial f}{\partial x} h, y_0 + \frac{\partial f}{\partial y} h\right) = f(2 + 6h, 1 + 2h) = (2 + 6h)^2 + (1 + 2h)^2 + 2(2 + 6h) + 4$$

Multiplying out the terms we obtain the one dimensional function along the gradient as

$$g(h) = 40h^2 + 40h + 13$$

This is a simple function and it is easy to determine  $h^* = -0.5$  by taking the first derivative and solving for its roots. This means that traveling a step size of  $h = -0.5$  along the gradient reaches a minimum value for the function in this direction. These values are substituted back to calculate a new value for  $x$  and  $y$  as follows:

$$x = 2 + 6(-0.5) = -1$$

$$y = 1 + 2(-0.5) = 0$$

Calculating the new values of  $x$  and  $y$  concludes the first iteration. Note that  $f(-1, 0) = 3$  is less than  $f(2, 1) = 13$  which indicates a move in the right direction.

##### Iteration 2:

The new initial point is  $(-1, 0)$ . We calculate the gradient at this point as

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

$$\frac{\partial f}{\partial y} = 2y = 2(0) = 0$$

which are used to determine the gradient at point  $(-1, 0)$  as

$$\nabla f = 0\mathbf{i} + 0\mathbf{j}$$

This indicates that the current location is a local optimum and no improvement can be gained by moving in any direction. To ensure that we have reached a minimum, we can calculate the

Hessian of the function. To determine the Hessian, the second partial derivatives are determined and evaluated as follows

$$\frac{\partial^2 f}{\partial x^2} = 2$$

$$\frac{\partial^2 f}{\partial y^2} = 2$$

$$\frac{\partial^2 f}{\partial y \partial x} = 0$$

The resulting Hessian matrix and its determinant are

$$H = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \quad |H| = 4 - 0^2 = 4$$

Since  $|H| > 0$  and  $\partial^2 f / \partial^2 x^2 > 0$  then  $f(-1,0)$  is a local minimum.

## OPTIMIZATION

Topic	Multidimensional Gradient Method
Summary	Textbook notes for the multidimensional gradient method
Major	All engineering majors
Authors	Ali Yalcin
Date	December 22, 2012
Web Site	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

## Chapter 09.05

### Simplex Method

*After reading this chapter, you should be able to:*

1. Formulate constrained optimization problems as a linear program
2. Solve linear programs with graphical solution approaches
3. Solve constrained optimization problems using simplex method

#### What is linear programming?

Linear programming is an optimization approach that deals with problems that have specific constraints. The one-dimensional and multi-dimensional optimization problems previously discussed did not consider any constraints on the values of the independent variables. In linear programming, the independent variables which are frequently used to model concepts such as availability of resources or required ratio of resources are constrained to be more than, less than or equal to a specific value.

The simplest linear program requires an objective function and a set of constraints. The objective function is either a maximization or a minimization of a linear combination of the independent variables of the problem and is expressed as (for a maximization problem)

$$\max z = c_1x_1 + c_2x_2 + \dots + c_nx_n$$

where  $c_i$  expresses the contribution (e.g. cost, profit etc) of each unit of  $x_i$  to the objective of the problem, and  $x_i$  are the independent or more commonly referred to as the decision variables whose values are determined by the solution of the problem.

The constraints are also a linear combination of the decision variables commonly expressed as an inequality of the form

$$a_{i1}x_1 + a_{i2}x_2 + \dots + a_{in}x_n \leq b_i$$

where  $a_{ij}$  and  $b_i$  are constant coefficients determined from the problem description as they relate to the constraints on the availability, interaction, and use of the resources.

#### Example 1

A woodworker builds and sells band-saw boxes. He manufactures two types of boxes using a combination of three types of wood, maple, walnut and cherry. To construct the Type I box, the carpenter requires 2 board foot (bf) (The board foot is a specialized unit of measure for the volume of lumber. It is the volume of a one-foot length of a board one foot wide and one inch thick) maple and 1 bf walnut. To construct the Type II box, he requires 3 bf of cherry

and 1 bf of walnut. Given that he has 10 bf of maple, 5 bf of walnut and 11 bf of cherry and he can sell Type I of box for \$120 and Type II box for \$160, how many of each box type should he make to maximize his revenue? Assume that the woodworker can build the boxes in any size, therefore fractional solutions are acceptable.

### Solution

The decision variables in this problem are the number of Type I and II boxes to be built. They are denoted by  $x_1$  and  $x_2$  respectively. Since the goal is to maximize revenues and the revenues are a function of the number of boxes of each type sold, we can represent the objective function as

$$\max z = 120x_1 + 160x_2$$

One of the constraints in this problem is availability of different types of wood. Therefore, based on the number of boxes produced, the sum of the total wood requirement must be less than or equal to the available amount of wood for each type. We can represent this type of constraint with three inequalities referring to maple, cherry and walnut respectively as follows:

$$2x_1 \leq 10$$

$$3x_2 \leq 11$$

$$x_1 + x_2 \leq 5$$

In addition, there are the non-negativity constraints which ensure that our solution does not have negative number of boxes. These constraints are shown as

$$x_1, x_2 \geq 0$$

### Graphical Solutions to Linear Programs

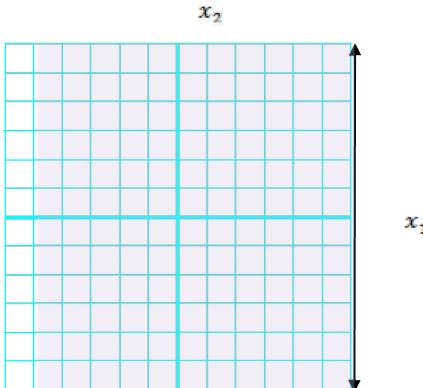
Linear programs of two or three dimensions can be solved using graphical solutions. While graphical solutions are not useful in addressing realistic size problems, they are particularly helpful in providing an intuitive explanation to the algebraic methodologies used to solve larger linear programs using computer algorithms. The graphical solution to linear programs is best explained by using an example.

### Example 2

Provide a graphical solution to the linear program in Example 1.

### Solution

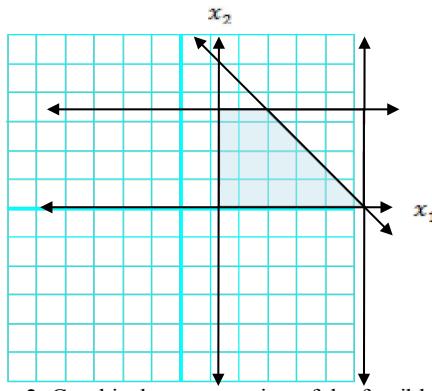
For a linear inequality of the form  $f(x_1, x_2) \leq b$  or  $f(x_1, x_2) \geq b$ , the points that satisfy the inequality includes the points on the line and the points on one side of the line. For example for the inequality  $2x_1 \leq 10$ , the shaded region in Figure 1 shows the points that satisfy this inequality. To determine which side of the line satisfies the inequality, simply test a single point in each region, such as the origin  $(0, 0)$  which satisfies the constraint and lies on the right side of the line in the shaded region.



**Figure 1.** Graphical representation of the points satisfying  $2x_1 \leq 10$ .

**Comment [AY1]:** Shading is not visible in these figures when printed. Maybe Russell can look into formatting it.

The set of points that satisfy all the constraints, including non-negativity constraints, from Example 1 are shown in Figure 2. The region which contains the points that satisfies all the constraint in a linear program is referred to as the feasible region.



**Figure 2.** Graphical representation of the feasible region.

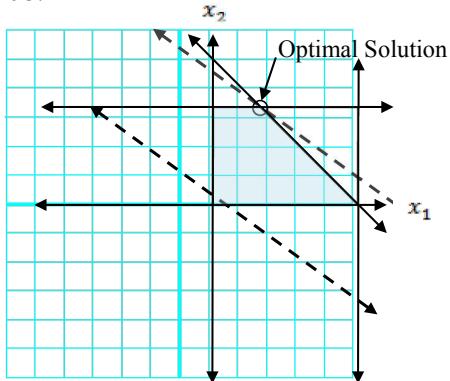
The objective function can also be represented by a line referred to as the isoprofit line (isocost line for minimization problems). To determine this line, simply assume a value for  $z$  such as  $z = 0$ . Then the objective function can be written as

$$0 = 120x_1 + 160x_2$$

$$x_2 = -\frac{120}{160}x_1 = -\frac{3}{4}x_1$$

where the isoprofit line has a slope of  $-\frac{3}{4}$ . The isoprofit line is shown as a dashed line through the origin in Figure 3. To determine the optimal solution, the isoprofit line is moved parallel to the original line drawn with slope  $-\frac{3}{4}$  in the direction that increases  $z$  until the

last point intersecting the feasible region is obtained. Such a point is reached at a single point  $\left(\frac{4}{3}, \frac{11}{3}\right)$  as shown in Figure 3.



**Figure 3.** Graphical representation of the optimal solution.

At the optimal solution, the value of the objective function is calculated as

$$120 \times \frac{4}{3} + 160 \times \frac{11}{3} = 746 \frac{2}{3}$$

The optimal solution when substituted back into the inequalities representing the structure of the problem reveals some additional important information about the problem. Below is the original set of constraints where the optimal solution to the problem is substituted in place of the decision variables. Note that the last two equations are now equalities indicating that the availability of the resources associated with these constraints (cherry and walnut) are preventing us from improving the value of the objective function. Such constraints are referred to as binding constraints. Note also that in the graphical solution, the optimal solution lies at the intersection of the binding constraints. On the other hand, the first inequality is a nonbinding constraint in the sense that the left-hand and the right-hand side of the constraint are unequal and this constraint does not pose a limitation to the optimal solution. In other words, if want to increase our revenues, we need to look into increasing the availability of cherry and walnut and not maple.

$$2 \times \frac{4}{3} < 10$$

$$3 \times \frac{11}{3} = 11$$

$$\frac{4}{3} + \frac{11}{3} = 5$$

### Solutions to Linear Programs

Solutions to linear programs can be one of two types as follows:

**1. Unique solution:**

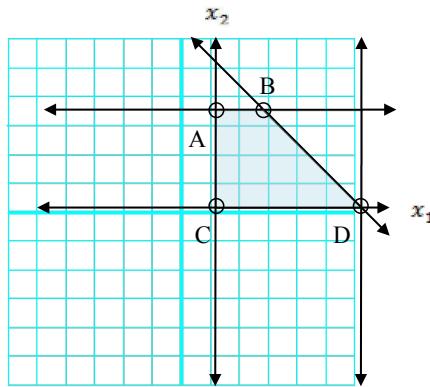
As seen in the solution to Example 2, there is a single point in the feasible region for which the maximum (or minimum in a minimization problem) value of the objective function is attainable. In graphical solutions, these points lie at the intersection of two or more lines which represent the constraints.

**2. Alternate Solutions:**

If the isoprofit (isocost) line is parallel to one of the lines representing the constraints, then the intersection would be an infinite number of points. In this case, any of such points would produce the maximum (minimum) value of the objective function.

A set of points  $S$  is said to be a convex set if the line segment joining any pair of points in  $S$  is also completely contained in  $S$ . For example, the feasible region shown in Figure 2 is a convex set. This is no coincidence. It can be shown that the feasible region of any linear program is a convex set.

Figure 4 shows the feasible region of Example 2 and highlights the corner points (also known as extreme points) of the convex set which occur where two or more constraints intersect within the feasible region. These extreme points are of special importance. Any linear program that has an optimal solution has an extreme point that is optimal. This is a very important result because it greatly reduces the number of points which may be optimal solutions to the linear program. For example, the entire feasible region shown in Figure 2 contains an infinite number of points, however the feasible region contains only four extreme points which may be the optimal solution to the linear program.



**Figure 4.** Graphical representation of the feasible region and its extreme points.

Once all the extreme points are determined, finding the optimal solution is trivial in the sense that the value of the objective function at each of these points can be calculated and, depending on the goal of the objective function, the extreme point resulting in the minimum or the maximum value is selected as the optimal solution. The simplex method which is the topic of next section is a much more efficient way of evaluating the extreme points in a convex set to determine the optimal solution.

### The Simplex Method

#### Converting a linear program to Standard Form

Before the simplex algorithm can be applied, the linear program must be converted into standard form where all the constraints are written as equations (no inequalities) and all variables are nonnegative (no unrestricted variables). This process of converting a linear program to its standard form requires the addition of slack variable  $s_i$  which represents the amount of the resource not used in the  $i$ th  $\leq$  constraint. Similarly,  $\geq$  constraints can be converted into standard form by subtracting excess variable  $e_i$ .

The standard form of any linear program can then be represented by the following linear system with  $n$  variables (including decision, slack and excess variables) and  $m$  constraints.

$$\begin{aligned} \max z = & c_1x_1 + c_2x_2 + \dots + c_nx_n \\ (\text{or min}) \quad & \\ \text{s.t.} \quad & a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n = b_1 \\ & a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n = b_2 \\ & \dots \quad \dots \quad \dots \quad \dots \\ & a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mn}x_n = b_m \end{aligned}$$

$$x_i \geq 0 \quad (i = 1, 2, \dots, n)$$

#### Example 3

Convert the linear program in Example 1 to its standard form.

#### Solution

For convenience, the linear program is reproduced below.

$$\begin{aligned} \text{Max } Z = & 120x_1 + 160x_2 \\ & 2x_1 \leq 10 \\ & 3x_2 \leq 11 \\ & x_1 + x_2 \leq 5 \\ & x_1, x_2 \geq 0 \end{aligned}$$

To convert the first constraint from an inequality to equality, we introduce the first slack variable  $s_1$  where

$$s_1 = 10 - 2x_1 \text{ or } 2x_1 + s_1 = 10.$$

Similarly after introducing  $s_2$  and  $s_3$ , we can convert the linear program into standard form as follows:

$$\begin{aligned}
 \max z = & 120x_1 + 160x_2 \\
 \text{s.t.} \quad & 2x_1 + s_1 = 10 \\
 & + 3x_2 + s_2 = 11 \\
 & x_1 + x_2 + s_3 = 5 \\
 & x_1, x_2, s_1, s_2, s_3 \geq 0
 \end{aligned}$$

### Basic and Nonbasic Variables, and Basic Feasible Solutions

If we define

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix} \text{ and } b = \begin{bmatrix} b_1 \\ b_2 \\ \dots \\ b_n \end{bmatrix},$$

the constraints of the standard form of a linear program can be simply represented by a system of simultaneous equations  $Ax = b$ .

A basic solution to system of  $m$  linear equations with  $n$  unknowns is found by setting  $n - m$  variables to zero and solving the  $m$  equations for the remaining  $m$  variables. The variables with zero values are referred to as the nonbasic variables and the remaining  $m$  variables are called the basic variables. Note that the choice of different nonbasic variables will lead to different solutions. If all basic variables are nonnegative, the solution is called a basic feasible solution. The optimum solution will be one of the basic feasible solutions. Let us illustrate this with an example.

#### **Example 3**

Determine a basic feasible solution for the linear program in Example 1.

#### **Solution**

The system of equation representing the constraints for this linear program is as follows:

$$\begin{aligned}
 2x_1 + s_1 &= 10 \\
 + 3x_2 + s_2 &= 11 \\
 x_1 + x_2 + s_3 &= 5
 \end{aligned}$$

where  $n = 5$  and  $m = 3$ . To obtain a basic feasible solution we need to set  $n - m = 2$  nonbasic variables to zero and solve the remaining system of  $3 \times 3$  linear equations. Let us start with setting values of  $x_1$  and  $x_2$  to zero. We can easily see that the solution to the system becomes

$$\begin{aligned}
 s_1 &= 10 \\
 s_2 &= 11 \\
 s_3 &= 5
 \end{aligned}$$

In this solution, all basic variables are nonnegative; therefore the solution is a basic feasible solution.

### Relationship between extreme points of a feasible region and basic feasible solutions

To establish the relationship between basic feasible solutions and extreme points of the feasible region, refer to Figure 4. The above basic feasible solution corresponds to the extreme point  $C$  at the origin since in this basic feasible solution  $x_1 = 0$  and  $x_2 = 0$ . Alternatively, if we set the values of  $s_3$  and  $x_2$  to zero, we see that we obtain the basic feasible solution where  $x_1 = 5$ ,  $s_1 = 5$ , and  $s_2 = 11$  which corresponds to the extreme point  $D$  in Figure 4.

There is a special relationship between extreme points  $C$  and  $D$  arising from their adjacency that is relevant to the simplex method. For a linear program with  $m$  constraints, two basic feasible solutions are adjacent if they have  $m - 1$  basic variables in common. In the basic feasible solutions corresponding to adjacent points  $C$  and  $D$ , the  $m - 1$  common basic variables are  $s_1 = 5$ , and  $s_2 = 11$ .

### The Simplex Algorithm

The simplex algorithm, instead of evaluating all basic feasible solutions (which can be prohibitive even for moderate-size problems), starts with a basic feasible solution and moves through other basic feasible solutions that successively improve the value of the objective function. The algorithm terminates once the optimal value is reached. Below we present a step-wise description of the simplex algorithm.

1. Convert the linear program into standard form.
2. Obtain a basic feasible solution from the standard form.
3. Determine if the basic feasible solution is optimal.
4. If the current basic feasible solution is not optimal, select a nonbasic variable that should become a basic variable and basic variable which should become a nonbasic variable to determine a new basic feasible solution with an improved objective function value.
5. Use elementary row operations to solve for the new basic feasible solution. Return to Step 3

Steps 1 and 2 of the algorithm have been previously discussed. Steps 3, 4 and 5 of the algorithm are best executed with the help of a tableau which is simply a table with a particular format that shows a summary of the key information regarding the linear program. For example the tableau shown in Table 1 below corresponds to the linear program described in Example 1 and the basic feasible solution in Example 3. There are several things to note about

Table 1.

1. The first row of the table (also called row 0) corresponds to the objective function where all the variables are on the left-hand side following the format
2.  $z - 120x_1 - 160x_2 = 0$
3. The basic feasible solution corresponds to the solution in Example 3. In addition note that variable  $z = 0$  is also considered as a basic variable.

4. In this particular example, the initial tableau where the decision variables  $x_1$  and  $x_2$  are considered as nonbasic variables leads to a basic feasible solution due to the fact that all the right hand side variables are nonnegative.
5. The tableau is in proper form which means the solution can be read directly by looking at the tableau and the RHS values. For a tableau to be in proper form it must meet all the following requirements:

one basic variable per row

the coefficient of all basic variables are +1 and the coefficients above and below the basic variables are zero

$z$  is the basic variable for row 0

**Table 1.**The initial tableau for example on in proper form

Basic	Z	$x_1$	$x_2$	$s_1$	$s_2$	$s_3$	RHS	Ratio
Z	1	-120	-160	0	0	0	0	
$s_1$	0	2	0	1	0	0	10	None
$s_2$	0	0	3	0	1	0	11	$\frac{11}{3}$
$s_3$	0	1	1	0	0	1	5	5

In step 3, to determine if a basic feasible solution is optimal, we need to determine if any of the nonbasic variables (who has value zero) can be increased to improve the value of the objective function. For example, in Table 1, since  $z = 120x_1 + 160x_2$ , increasing either one of the nonbasic variables  $x_1$  and  $x_2$  would increase the value of the objective function value. In the tableau, this equates to looking for negative coefficients in row 0 due to the format the objective function is written. The basic feasible solution shown in Table 1 is therefore not optimal since the coefficients of  $x_1$  and  $x_2$  are less than zero.

To improve the solution, we can increase the value of either  $x_1$  or  $x_2$ . We choose to increase  $x_2$  since the value of the objective function increases at a higher rate (160 vs. 120 per unit of increase). The nonbasic variable with the most negative coefficient (in a maximization problem) in row 0, in this case  $x_2$ , is called the entering variable and is always selected as the nonbasic variable that becomes a basic variable.

The basic variable that is replaced by the entering variable, also called the leaving variable, is determined by looking at the values in the “Ratio” column in the tableau. The values in this column are simply the ratio of the RHS values divided by the coefficient of the entering variable in that row. The leaving variable is selected to be the basic variable in the row with the smallest ratio. This is the highest value that the entering variable can have and still result in a basic feasible solution. For the tableau shown in Table 1, the leaving variable is  $s_2$  in row 3.

Once the entering and leaving variables are determined, we use [elementary row operations \(add link?\)](#) (EROs) to make the entering variable a basic variable in the row of the leaving variable by making its coefficient 1 in that row and 0 in all other rows. For example, for Table 1 where the entering and leaving variables are  $x_2$  and  $s_2$  respectively, after the

EROS, the tableau is shown in Table 2. The tableau shows a new basic feasible solution (note that all RHS are nonnegative) where

$$s_1 = 10$$

$$x_2 = \frac{1}{3}$$

$$s_3 = \frac{4}{3}$$

This basic feasible solution corresponds to the adjacent extreme point A in Figure 4 with coordinates  $x_1 = 0$  and  $x_2 = \frac{1}{3}$  and objective function value  $\frac{1760}{3}$ .

**Table 2.** The tableau for the basic feasible solution corresponding to extreme point A in proper form.

Basic	Z	$x_1$	$x_2$	$s_1$	$s_2$	$s_3$	RHS	Ratio
Z	1	-120	0	0	$\frac{160}{3}$	0	$\frac{1760}{3}$	
$s_1$	0	2	0	1	0	0	10	5
$x_2$	0	0	1	0	$\frac{1}{3}$	0	$\frac{1}{3}$	None
$s_3$	0	1	0	0	$-\frac{1}{3}$	1	$\frac{4}{3}$	$\frac{4}{3}$

After a new basic feasible solution is obtained, the algorithm returns to Step 3 to check if the new basic feasible solution is optimal. This cycle continues until the objective function value cannot be increased by increasing the value of any of the nonbasic variables. In other words, in a maximization problem, this is the same as having no negative valued coefficients in row 0.

#### A note about minimization problems:

It is important to note that the optimality condition of no negative valued coefficients in row 0 is only applicable in maximization problems. In a minimization problem, the optimality condition exists when none of the coefficients in row 0 are positive. Furthermore, in minimization problems, the entering variable is chosen to be the nonbasic variable with the highest positive coefficient in row 0.

Let us illustrate the simplex algorithm by solving the problem presented in Example 1.

#### **Example 4**

Solve the linear program in Example 1 using the simplex algorithm.

#### **Solution**

##### Step 1:

Convert the linear program into standard form.

The linear program in standard form is

$$\begin{array}{lll}
 \max z = & 120x_1 + 160x_2 \\
 \text{s.t.} & 2x_1 + s_1 = 10 \\
 & + 3x_2 + s_2 = 11 \\
 & x_1 + x_2 + s_3 = 5 \\
 & x_1, x_2, s_1, s_2, s_3 \geq 0
 \end{array}$$

**Step 2:**

Obtain a basic feasible solution from the standard form.

Previously we have shown that the solution where  $x_1 = 0$  and  $x_2 = 0$  is a basic feasible solution so we will start the algorithm here.

**Step 3:**

Determine if the basic feasible solution is optimal.

At this step we create the tableau for this basic feasible solution which was initially shown in Table 1. For convenience the table is reproduced as Table 3.

**Table 3.**The initial tableau in proper form

Basic	Z	$x_1$	$x_2$	$s_1$	$s_2$	$s_3$	RHS	Ratio
Z	1	-120	-160	0	0	0	0	
$s_1$	0	2	0	1	0	0	10	None
$s_2$	0	0	3	0	1	0	11	$\frac{11}{3}$
$s_3$	0	1	1	0	0	1	5	5

**Step 4:**

If the current basic feasible solution is not optimal, select a nonbasic variable that should become a basic variable and basic variable which should become a nonbasic variable to determine a new basic feasible solution with an improved objective function value.

The current solution is not optimal. There are negative coefficients in row 0. Since  $x_2$  has the most negative coefficient in row 0 and  $s_2$  has the lowest ratio, the entering and the leaving variables are  $x_2$  and  $s_2$ , respectively.

**Step 5:**

Use elementary row operations to solve for the new basic feasible solution. Return to Step 3. The new basic feasible solution is shown in Table 4, which is the same as Table 2.

**Table 4.**The tableau for the new basic feasible solution in the first iteration

Basic	Z	$x_1$	$x_2$	$s_1$	$s_2$	$s_3$	RHS	Ratio
Z	1	-120	0	0	$\frac{160}{3}$	0	$\frac{1760}{3}$	
$s_1$	0	2	0	1	0	0	10	5
$x_2$	0	0	1	0	$\frac{1}{3}$	0	$\frac{11}{3}$	None
$s_3$	0	1	0	0	$-\frac{1}{3}$	1	$\frac{4}{3}$	$\frac{4}{3}$

**Step 3:**

Determine if the basic feasible solution is optimal.

The basic solution in Table 4 is still not optimal as the objective function value can be increased by increasing the value of  $x_1$ .

**Step 4:**

If the current basic feasible solution is not optimal, select a nonbasic variable that should become a basic variable and basic variable which should become a nonbasic variable to determine a new basic feasible solution with an improved objective function value.

In the second iteration, since  $x_1$  has the most (and only) negative coefficient in row 0 and  $s_3$  has the lowest ratio, the entering and leaving variables are  $x_1$  and  $s_3$ , respectively.

**Step 5:**

Use elementary row operations to solve for the new basic feasible solution. Return to Step 3

The new basic feasible solution is shown in Table 5.

**Table 5.** The tableau for the basic feasible solution in the second iteration .

Basic	Z	$x_1$	$x_2$	$s_1$	$s_2$	$s_3$	RHS	Ratio
Z	1	0	0	0	$40/3$	120	$2240/3$	
$s_1$	0	0	0	1	$2/3$	-2	$22/3$	
$x_2$	0	0	1	0	$1/3$	0	$11/3$	
$x_1$	0	1	0	0	$-1/3$	1	$4/3$	

**Step 3:**

Determine if the basic feasible solution is optimal.

Since there are no negative coefficients in row 0, we have reached the optimal solution where the objective function value is  $2240/3$  and

$$s_1 = 22/3$$

$$x_2 = 11/3$$

$$x_1 = 4/3$$

$$s_2 = s_3 = 0$$

Note that all these values can be read from the tableau shown in Table 5. This solution also corresponds to the extreme point B in Figure 4 which was also determined to be optimal using the graphical solution approach.

Finally, the woodworker should build  $4/3$  Type I boxes and  $11/3$  Type II boxes to maximize his revenue to \$746.67.

---

**OPTIMIZATION**

---

Topic Simplex Method  
Summary Textbook notes for the Simplex method  
Major All engineering majors  
Authors Ali Yalcin, Autar Kaw,  
Date November 22, 2011  
Web Site <http://numericalmethods.eng.usf.edu>

---

# Chapter 10.01

## Introduction to Partial Differential Equations

After reading this chapter, you should be able to:

1. identify the difference between ordinary and partial differential equations.
2. identify different types of partial differential equations.

### What is a Partial Differential Equation (PDE)

A differential equation with one independent variable is called an ordinary differential equation. An example of such an equation would be

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

where  $y$  is the dependent variable, and  $x$  is the independent variable.

What if there is more than one independent variable? Then the differential equation is called a partial differential equation. An example of such an equation would be

$$3 \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = x^2 + y^2$$

subject to certain conditions: where  $u$  is the dependent variable, and  $x$  and  $y$  are the independent variables.

### From Ordinary to a Partial Differential Equation

Assume we put a spherical steel ball that is at room temperature in hot water. The temperature of the ball is going to increase with time. What if we wish to find what this temperature vs. time profile would look like for the ball? We would develop a mathematical model for this based on the law of conservation of heat energy. From an energy balance, Heat gained - Heat lost = Heat stored (1)

The energy stored in the mass is given by

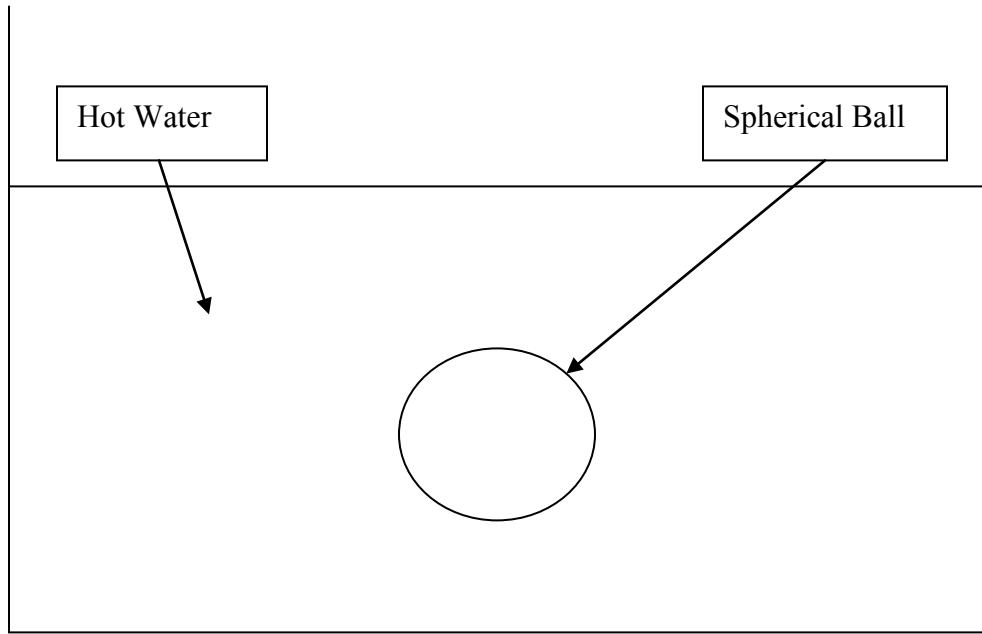
$$\text{Heat stored in the ball} = mC\theta \quad (2)$$

where

$m$  = mass of ball, kg

$C$  = specific heat of the ball,  $J/(kg \cdot K)$

$\theta$  = temperature of the ball at a given time, K



The rate of heat gained by the ball due to convection is

$$\text{Rate of heat gained due to convection} = hA(\theta - \theta_a), \quad (3)$$

where

$h$  = the convective cooling coefficient,  $\text{W}/(m^2 - K)$ .

$A$  = surface area of ball,  $m^2$

$\theta_a$  = ambient temperature of the hot water,  $K$

As you can see we have the expression for the rate at which heat is gained (not the heat gained), so we rewrite the heat energy balance as

$$\begin{aligned} &\text{Rate at which heat is gained} - \text{Rate at which heat is lost} \\ &= \text{Rate at which heat is stored} \end{aligned} \quad (4)$$

This gives us

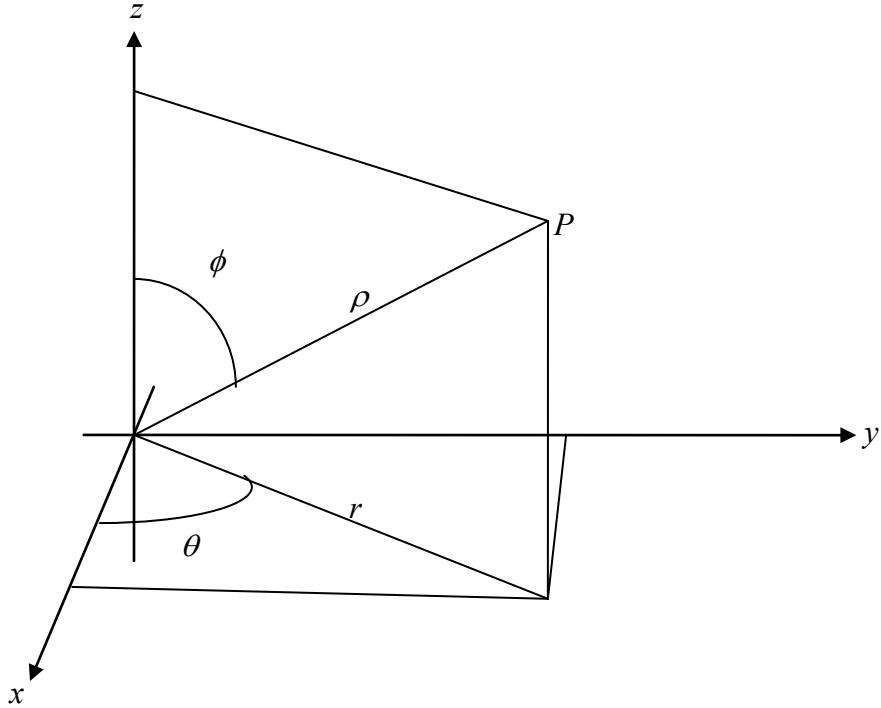
$$hA(\theta - \theta_a) = mC \frac{d\theta}{dt} \quad (5)$$

Equation (5) is a first order ordinary differential equation that when solved with the initial condition  $\theta(0) = \theta_0$ , would give us the temperature of the spherical ball as a function of time.

However, we made a large assumption in deriving Equation (5) - we assumed that the system is lumped. What does a lumped system mean? It implies that the internal conduction in the sphere is large enough that the temperature throughout the ball is uniform. This allows us to make the assumption that the temperature is only a function of time and not of the location in the spherical ball. The system being considered lumped for this case depends on: material of the ball, geometry, and heat exchange factor (convection coefficient) of the ball with its surroundings.

What happens if the system cannot be treated as a lumped system? In that case, the temperature of the ball will now be a function not only of time, but also the location.

In spherical co-ordinates, the location is given by  $r, \theta, \phi$  co-ordinates.



**Figure 1** Spherical Coordinate System.

The differential equation would now be a partial differential equation and is given as

$$\begin{aligned} \frac{k}{r^2} \frac{\partial}{\partial r} \left( r^2 \frac{\partial \theta}{\partial r} \right) + \frac{k}{r^2 \sin \theta} \frac{\partial}{\partial \theta} \left( \sin \theta \frac{\partial \theta}{\partial \theta} \right) + \frac{k}{r^2 \sin^2 \theta} \frac{\partial^2 \theta}{\partial \phi^2} &= \rho C \frac{\partial \theta}{\partial t}, \quad t \geq 0, \theta(0) = \theta_a \\ \frac{\partial \theta}{\partial r} + h(\theta - \theta_a) &= 0, \text{ at the surface} \end{aligned} \quad (6)$$

where

$k$  = thermal conductivity of material,  $W/(m \cdot K)$

$\rho$  = density of material,  $kg/m^3$

As an introduction to solve PDEs, most textbooks concentrate on linear second order PDEs with two independent variables and one dependent variable. The general form of such an equation is

$$A \frac{\partial^2 u}{\partial x^2} + B \frac{\partial^2 u}{\partial x \partial y} + C \frac{\partial^2 u}{\partial y^2} + D = 0 \quad (7)$$

Where  $A, B$ , and  $C$  are functions of  $x$  and  $y$  and  $D$  is a function of  $x, y, u$  and  $\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}$ .

Depending on the value of  $B^2 - 4AC$ , a 2<sup>nd</sup> order linear PDE can be classified into three categories.

1. if  $B^2 - 4AC < 0$ , it is called elliptic
2. if  $B^2 - 4AC = 0$ , it is called parabolic
3. if  $B^2 - 4AC > 0$ , it is called hyperbolic

### Elliptic Equation

The Laplace equation for steady state temperature in a plate is an example of an elliptic second order linear partial differential equation. The Laplace equation for steady state temperature in a plate is given by

$$\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} = 0 \quad (8)$$

Using the general form of second order linear PDEs with one dependent variable and two independent variables,

$$A \frac{\partial^2 u}{\partial x^2} + B \frac{\partial^2 u}{\partial x \partial y} + C \frac{\partial^2 u}{\partial y^2} + D = 0$$

$$A = 1, B = 0, C = 1, D = 0,$$

$$\begin{aligned} \text{gives } B^2 - 4AC &= 0 - 4(1)(1) \\ &= -4 \\ &= -4 < 0 \end{aligned}$$

This classifies Equation (8) as elliptic.

### Parabolic Equation

The heat conduction equation is an example of a parabolic second order linear partial differential equation. The heat conduction equation is given by

$$\frac{\partial T}{\partial t} = k \frac{\partial^2 T}{\partial x^2} \quad (9)$$

Using the general form of second order linear PDEs with one dependent variable and two independent variables,

$$A \frac{\partial^2 u}{\partial x^2} + B \frac{\partial^2 u}{\partial x \partial y} + C \frac{\partial^2 u}{\partial y^2} + D = 0$$

$$A = k, B = 0, C = 0, D = -1,$$

$$\begin{aligned} \text{gives } B^2 - 4AC &= 0 - 4(0)(k) \\ &= 0 \end{aligned}$$

This classifies Equation (9) as parabolic.

### Hyperbolic Equation

The wave equation is an example of a hyperbolic second order linear partial differential equation. The wave equation is given by

$$\frac{\partial^2 y}{\partial x^2} = \frac{1}{c^2} \frac{\partial^2 y}{\partial t^2} \quad (10)$$

Using the general form of second order linear PDEs with one dependent variable and two independent variables,

$$A \frac{\partial^2 u}{\partial x^2} + B \frac{\partial^2 u}{\partial x \partial y} + C \frac{\partial^2 u}{\partial y^2} + D = 0$$

$$A = 1, B = 0, C = -\frac{1}{c^2}, D = 0$$

$$\text{gives } B^2 - 4AC = 0 - 4(1)(-\frac{1}{c^2})$$

$$= \frac{4}{c^2}$$

$$= \frac{4}{c^2} > 0$$

This classifies Equation (10) as hyperbolic.

### PARTIAL DIFFERENTIAL EQUATIONS

Topic	Introduction to Partial Differential Equations
Summary	Textbook notes for the introduction of partial differential equations
Major	All engineering majors
Authors	Autar Kaw, Sri Harsha Garapati
Date	February 11, 2011
Web Site	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

## Chapter 10.02

# Parabolic Partial Differential Equations

After reading this chapter, you should be able to:

1. Use numerical methods to solve parabolic partial differential equations by explicit, implicit, and Crank-Nicolson methods.

The general second order linear PDE with two independent variables and one dependent variable is given by

$$A \frac{\partial^2 u}{\partial x^2} + B \frac{\partial^2 u}{\partial x \partial y} + C \frac{\partial^2 u}{\partial y^2} + D = 0 \quad (1)$$

where  $A, B, C$  are functions of the independent variables,  $x, y$ , and  $D$  can be a function of  $x, y, u, \frac{\partial u}{\partial x}$  and  $\frac{\partial u}{\partial y}$ . If  $B^2 - 4AC = 0$ , Equation (1) is called a parabolic partial differential equation. One of the simple examples of a parabolic PDE is the heat-conduction equation for a metal rod (Figure 1)

$$\alpha \frac{\partial^2 T}{\partial x^2} = \frac{\partial T}{\partial t} \quad (2)$$

where

$T$  = temperature as a function of location,  $x$  and time,  $t$   
in which the thermal diffusivity,  $\alpha$  is given by

$$\alpha = \frac{k}{\rho C}$$

where

$k$  = thermal conductivity of rod material,

$\rho$  = density of rod material,

$C$  = specific heat of the rod material.



**Figure 1:** A metal rod

### Explicit Method of Solving Parabolic PDEs

To numerically solve parabolic PDEs such as Equation (2), one can use finite difference approximations of the partial derivatives so that the dependent variable,  $T$  is now sought at particular nodes ( $x$ -location) and time ( $t$ ) (Figure 2). The left hand side second derivative is approximated by the central divided difference approximation as

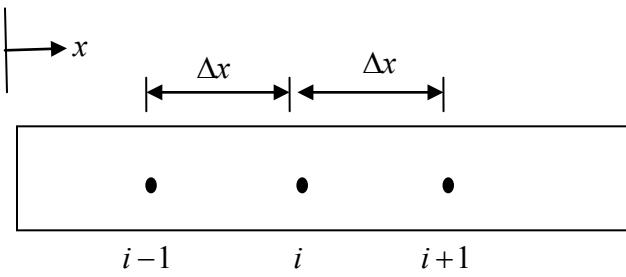
$$\frac{\partial^2 T}{\partial x^2} \Big|_{i,j} \cong \frac{T_{i+1}^j - 2T_i^j + T_{i-1}^j}{(\Delta x)^2} \quad (3)$$

where

$i$  = node number along the  $x$ -direction,  $i = 0, 1, \dots, n$ ,

$j$  = node number along the time,

$\Delta x$  = distance between nodes.



**Figure 2:** Schematic diagram showing the node representation in the model

For a rod of length  $L$  which is divided into  $n+1$  nodes,

$$\Delta x = \frac{L}{n} \quad (4)$$

The time is similarly broken into time steps of  $\Delta t$ . Hence  $T_i^j$  corresponds to the temperature at node  $i$ , that is,

$$x = (i)(\Delta x)$$

and time,

$$t = (j)(\Delta t),$$

where

$$\Delta t = \text{time step.}$$

The time derivative of the right hand side of Equation (2) is approximated by the forward divided difference approximation

$$\frac{\partial T}{\partial t} \Big|_{i,j} \cong \frac{T_i^{j+1} - T_i^j}{\Delta t} \quad (5)$$

Substituting the finite difference approximations given by Equations (3) and (5) in Equation (2) gives

$$\alpha \frac{T_{i+1}^j - 2T_i^j + T_{i-1}^j}{(\Delta x)^2} = \frac{T_i^{j+1} - T_i^j}{\Delta t}$$

Solving for the temperature at the time node  $j+1$ , gives

$$T_i^{j+1} = T_i^j + \alpha \frac{\Delta t}{(\Delta x)^2} (T_{i+1}^j - 2T_i^j + T_{i-1}^j)$$

Choosing

$$\lambda = \alpha \frac{\Delta t}{(\Delta x)^2} \quad (6)$$

$$T_i^{j+1} = T_i^j + \lambda (T_{i+1}^j - 2T_i^j + T_{i-1}^j) \quad (7)$$

Equation (7) can be solved explicitly because it can be written for each internal location node of the rod for time node  $j+1$  in terms of the temperature at time node  $j$ . In other words, if we know the temperature at node  $j=0$ , and knowing the boundary temperatures, which is the temperature at the external nodes, we can find the temperature at the next time step. We continue the process by first finding the temperature at all nodes  $j=1$ , and using these to find the temperature at the next time node,  $j=2$ . This process continues till we reach the time at which we are interested in finding the temperature.

### Example 1

A rod of steel is subjected to a temperature of  $100^\circ C$  on the left end and  $25^\circ C$  on the right end. If the rod is of length  $0.05m$ , use the explicit method to find the temperature distribution in the rod from  $t=0$  and  $t=9$  seconds. Use  $\Delta x = 0.01m$ ,  $\Delta t = 3s$ . Given:

$$k = 54 \frac{W}{m \cdot K}, \rho = 7800 \frac{kg}{m^3}, C = 490 \frac{J}{kg \cdot K}.$$

The initial temperature of the rod is  $20^\circ C$ .

### Solution

$$\begin{aligned} \alpha &= \frac{k}{\rho C} \\ &= \frac{54}{7800 \times 490} \\ &= 1.4129 \times 10^{-5} \text{ m}^2 / \text{s} \end{aligned}$$

Then

$$\lambda = \alpha \frac{\Delta t}{(\Delta x)^2}$$

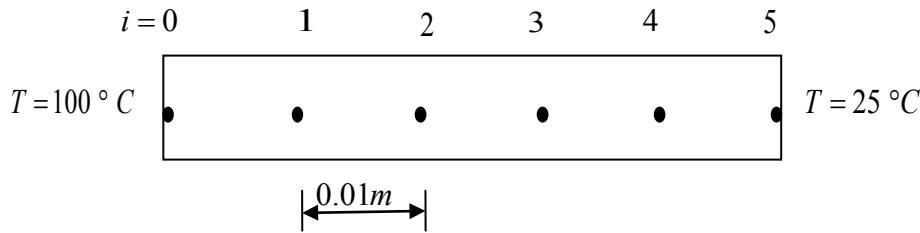
$$= 1.4129 \times 10^{-5} \frac{3}{(0.01)^2}$$

$$= 0.4239$$

$$\text{Number of time steps} = \frac{t_{final} - t_{initial}}{\Delta t}$$

$$= \frac{9 - 0}{3}$$

$$= 3$$



**Figure 3:** Schematic diagram showing the node distribution in the rod

The boundary conditions

$$\left. \begin{array}{l} T_0^j = 100^\circ C \\ T_5^j = 25^\circ C \end{array} \right\} \text{ for all } j = 0, 1, 2, 3 \quad (\text{E1.1})$$

The initial temperature of the rod is  $20^\circ C$ , that is, all the temperatures of the nodes inside the rod are at  $20^\circ C$  when time,  $t = 0 \text{ sec}$  except for the boundary nodes as given by Equation (E1.1). This could be represented as

$$T_i^0 = 20^\circ C, \text{ for all } i = 1, 2, 3, 4. \quad (\text{E1.2})$$

Initial temperature at the nodes inside the rod (when  $t=0 \text{ sec}$ )

$$T_0^0 = 100^\circ C \quad \text{from Equation (E1.1)}$$

$$\left. \begin{array}{l} T_1^0 = 20^\circ C \\ T_2^0 = 20^\circ C \\ T_3^0 = 20^\circ C \\ T_4^0 = 20^\circ C \end{array} \right\} \quad \text{from Equation (E1.2)}$$

$$T_5^0 = 25^\circ C \quad \text{from Equation (E1.1)}$$

Temperature at the nodes inside the rod when  $t=3 \text{ sec}$

Setting  $j = 0$  and  $i = 0, 1, 2, 3, 4, 5$  in Equation (7) gives the temperature of the nodes inside the rod when time,  $t = 3 \text{ sec}$ .

$$T_0^1 = 100^\circ C \quad \text{Boundary Condition (E1.1)}$$

$$\begin{aligned} T_1^1 &= T_1^0 + \lambda(T_2^0 - 2T_1^0 + T_0^0) \\ &= 20 + 0.4239(20 - 2(20) + 100) \\ &= 20 + 0.4239(80) \\ &= 20 + 33.912 \\ &= 53.912^\circ C \end{aligned}$$

$$\begin{aligned} T_2^1 &= T_2^0 + \lambda(T_3^0 - 2T_2^0 + T_1^0) \\ &= 20 + 0.4239(20 - 2(20) + 20) \\ &= 20 + 0.4239(0) \\ &= 20 + 0 \\ &= 20^\circ C \end{aligned}$$

$$\begin{aligned} T_3^1 &= T_3^0 + \lambda(T_4^0 - 2T_3^0 + T_2^0) \\ &= 20 + 0.4239(20 - 2(20) + 20) \\ &= 20 + 0.4239(0) \\ &= 20 + 0 \\ &= 20^\circ C \end{aligned}$$

$$\begin{aligned} T_4^1 &= T_4^0 + \lambda(T_5^0 - 2T_4^0 + T_3^0) \\ &= 20 + 0.4239(25 - 2(20) + 20) \\ &= 20 + 0.4239(5) \\ &= 20 + 2.1195 \\ &= 22.120^\circ C \end{aligned}$$

$$T_5^1 = 25^\circ C \quad \text{Boundary Condition (E1.1)}$$

Temperature at the nodes inside the rod when  $t=6$  sec

Setting  $j = 1$  and  $i = 0, 1, 2, 3, 4, 5$  in Equation (7) gives the temperature of the nodes inside the rod when time,  $t = 6$  sec

$$T_0^2 = 100^\circ C \quad \text{Boundary Condition (E1.1)}$$

$$\begin{aligned}
 T_1^2 &= T_1^1 + \lambda(T_2^1 - 2T_1^1 + T_0^1) \\
 &= 53.912 + 0.4239(20 - 2(53.912) + 100) \\
 &= 53.912 + 0.4239(12.176) \\
 &= 53.912 + 5.1614 \\
 &= 59.073^\circ C
 \end{aligned}$$

$$\begin{aligned}
 T_2^2 &= T_2^1 + \lambda(T_3^1 - 2T_2^1 + T_1^1) \\
 &= 20 + 0.4239(20 - 2(20) + 53.912) \\
 &= 20 + 0.4239(33.912) \\
 &= 20 + 14.375 \\
 &= 34.375^\circ C
 \end{aligned}$$

$$\begin{aligned}
 T_3^2 &= T_3^1 + \lambda(T_4^1 - 2T_3^1 + T_2^1) \\
 &= 20 + 0.4239(22.120 - 2(20) + 20) \\
 &= 20 + 0.4239(2.120) \\
 &= 20 + 0.89867 \\
 &= 20.899^\circ C
 \end{aligned}$$

$$\begin{aligned}
 T_4^2 &= T_4^1 + \lambda(T_5^1 - 2T_4^1 + T_3^1) \\
 &= 22.120 + 0.4239(25 - 2(22.120) + 20) \\
 &= 22.120 + 0.4239(0.76) \\
 &= 22.120 + 0.032220 \\
 &= 22.442^\circ C
 \end{aligned}$$

$$T_5^2 = 25^\circ C \quad \text{Boundary Condition (E1.1)}$$

Temperature at the nodes inside the rod when  $t=9$  sec

Setting  $j = 2$  and  $i = 0, 1, 2, 3, 4, 5$  in Equation (7) gives the temperature of the nodes inside the rod when time,  $t = 9$  sec

$$T_0^3 = 100^\circ C \quad \text{Boundary Condition (E1.1)}$$

$$\begin{aligned}
 T_1^3 &= T_1^2 + \lambda(T_2^2 - 2T_1^2 + T_0^2) \\
 &= 59.073 + 0.4239(34.375 - 2(59.073) + 100) \\
 &= 59.073 + 0.4239(16.229) \\
 &= 59.073 + 6.8795 \\
 &= 65.953^\circ C
 \end{aligned}$$

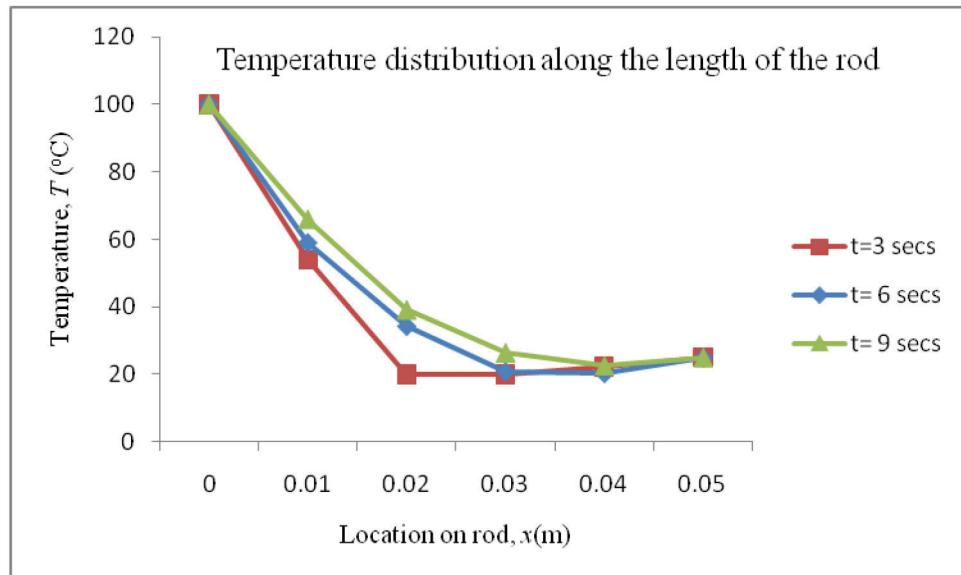
$$\begin{aligned}
 T_2^3 &= T_2^2 + \lambda(T_3^2 - 2T_2^2 + T_1^2) \\
 &= 34.375 + 0.4239(20.899 - 2(34.375) + 59.073) \\
 &= 34.375 + 0.4239(11.222) \\
 &= 34.375 + 4.7570 \\
 &= 39.132^\circ C
 \end{aligned}$$

$$\begin{aligned}
 T_3^3 &= T_3^2 + \lambda(T_4^2 - 2T_3^2 + T_2^2) \\
 &= 20.899 + 0.4239(22.442 - 2(20.899) + 34.375) \\
 &= 20.899 + 0.4239(15.019) \\
 &= 20.899 + 6.367 \\
 &= 27.266^\circ C
 \end{aligned}$$

$$\begin{aligned}
 T_4^3 &= T_4^2 + \lambda(T_5^2 - 2T_4^2 + T_3^2) \\
 &= 22.442 + 0.4239(25 - 2(22.442) + 20.899) \\
 &= 22.442 + 0.4239(1.0150) \\
 &= 22.442 + 0.4303 \\
 &= 22.872^\circ C
 \end{aligned}$$

$$T_5^3 = 25^\circ C \quad \text{Boundary Condition (E1.1)}$$

To better visualize the temperature variation at different locations at different times, temperature distribution along the length of the rod at different times is plotted in the Figure 4.



**Figure 4:** Temperature distribution from explicit method

### Implicit Method for Solving Parabolic PDEs

In the explicit method, one is able to find the solution at each node, one equation at a time. However, the solution at a particular node is dependent only on temperature from neighboring nodes from the previous time step. For example, in the solution of Example 1, the temperatures at node 2 and 3 artificially stay at the initial temperature at  $t = 3$  seconds. This is contrary to what we would expect physically from the problem.

Also the explicit method does not guarantee stability which depends on the value of the time step, location step and the parameters of the elliptic equation. For the PDE

$$\alpha \frac{\partial^2 T}{\partial x^2} = \frac{\partial T}{\partial t},$$

the explicit method is convergent and stable for

$$\frac{\alpha \Delta t}{(\Delta x)^2} \leq \frac{1}{2} \quad (8)$$

These issues are addressed by using the implicit method. Instead of the temperature being found one node at a time, the implicit method results in simultaneous linear equations for the temperature at all interior nodes for a particular time.

The implicit method to solve the parabolic PDE given by equation (2) is as follows. The second derivative on the left hand side of the equation is approximated by the central divided difference scheme at time level  $j+1$  at node  $i$  as

$$\left. \frac{\partial^2 T}{\partial x^2} \right|_{i,j+1} \approx \frac{T_{i+1}^{j+1} - 2T_i^{j+1} + T_{i-1}^{j+1}}{(\Delta x)^2} \quad (9)$$

The first derivative on the right hand side of the equation is approximated by backward divided difference approximation at time level  $j+1$  and node  $i$  as

$$\left. \frac{\partial T}{\partial t} \right|_{i,j+1} \approx \frac{T_i^{j+1} - T_i^j}{\Delta t} \quad (10)$$

Substituting Equations (9) and (10) in Equation (2) gives

$$\alpha \frac{T_{i+1}^{j+1} - 2T_i^{j+1} + T_{i-1}^{j+1}}{(\Delta x)^2} = \frac{T_i^{j+1} - T_i^j}{\Delta t}$$

giving

$$-\lambda T_{i-1}^{j+1} + (1 + 2\lambda)T_i^{j+1} - \lambda T_{i+1}^{j+1} = T_i^j \quad (11)$$

where

$$\lambda = \alpha \frac{\Delta t}{(\Delta x)^2}$$

Now Equation (11) can be written for all nodes (except the external nodes), at a particular time level. This results in simultaneous linear equations which can be solved to find the nodal temperature at a particular time.

### Example 2

A rod of steel is subjected to a temperature of  $100^\circ\text{C}$  on the left end and  $25^\circ\text{C}$  on the right end. If the rod is of length  $0.05\text{m}$ , use the implicit method to find the temperature distribution in the rod from  $t = 0$  to  $t = 9$  seconds. Use  $\Delta x = 0.01\text{m}$  and  $\Delta t = 3\text{s}$ .

Given

$$k = 54 \frac{W}{m \cdot K}, \rho = 7800 \frac{kg}{m^3}, C = 490 \frac{J}{kg \cdot K}.$$

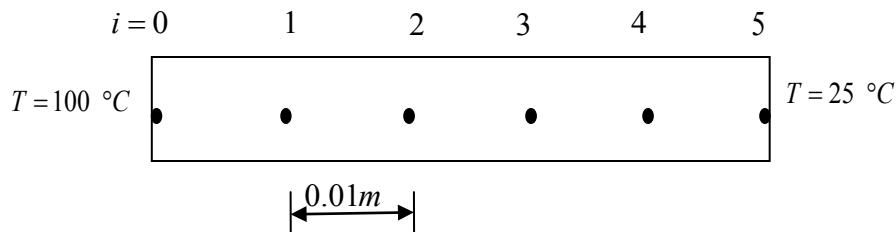
The initial temperature of the rod is  $20^\circ C$ .

**Solution**

$$\begin{aligned}\alpha &= \frac{k}{\rho C} \\ &= \frac{54}{7800 \times 490} \\ &= 1.4129 \times 10^{-5} \text{ } m^2 / \text{s}\end{aligned}$$

Then

$$\begin{aligned}\lambda &= \alpha \frac{\Delta t}{(\Delta x)^2} \\ &= 1.412 \times 10^{-5} \frac{3}{(0.01)^2} \\ &= 0.4239\end{aligned}$$



**Figure 5:** Schematic diagram showing the node representation in the model

The boundary conditions

$$\left. \begin{array}{l} T_0^j = 100^\circ C \\ T_5^j = 25^\circ C \end{array} \right\} \text{ for } j = 0, 1, 2, 3 \quad (\text{E2.1})$$

The initial temperature of the rod is  $20^\circ C$ , that is, the temperatures of all the nodes inside the rod are at  $20^\circ C$  when time,  $t = 0$  except for the boundary nodes where the temperatures are given by satisfying the Equation (E2.1). This could be represented as

$$T_i^0 = 20^\circ C, \text{ for } i = 1, 2, 3, 4. \quad (\text{E2.2})$$

Initial temperature at the nodes inside the rod (when  $t=0$  sec)

$$T_0^0 = 100^\circ C \quad \text{from Equation (E2.1)}$$

$$\left. \begin{array}{l} T_1^0 = 20^\circ C \\ T_2^0 = 20^\circ C \\ T_3^0 = 20^\circ C \\ T_4^0 = 20^\circ C \end{array} \right\} \text{from Equation (E2.2)}$$

$$T_5^0 = 25^\circ C \quad \text{from Equation (E2.1)}$$

Temperature at the nodes inside the rod when t=3 sec

$$\left. \begin{array}{l} T_0^1 = 100^\circ C \\ T_5^1 = 25^\circ C \end{array} \right\} \text{Boundary Condition (E2.1)}$$

For all the interior nodes, putting  $j = 0$  and  $i = 1, 2, 3, 4$  in Equation (11) gives the following equations

i=1

$$\begin{aligned} -\lambda T_0^1 + (1+2\lambda)T_1^1 - \lambda T_2^1 &= T_1^0 \\ (-0.4239 \times 100) + (1+2 \times 0.4239)T_1^1 - (0.4239 T_2^1) &= 20 \\ -42.39 + 1.8478 T_1^1 - 0.4239 T_2^1 &= 20 \\ 1.8478 T_1^1 - 0.4239 T_2^1 &= 62.390 \end{aligned} \quad (\text{E2.3})$$

i=2

$$\begin{aligned} -\lambda T_1^1 + (1+2\lambda)T_2^1 - \lambda T_3^1 &= T_2^0 \\ -0.4239 T_1^1 + 1.8478 T_2^1 - 0.4239 T_3^1 &= 20 \end{aligned} \quad (\text{E2.4})$$

i=3

$$\begin{aligned} -\lambda T_2^1 + (1+2\lambda)T_3^1 - \lambda T_4^1 &= T_3^0 \\ -0.4239 T_2^1 + 1.8478 T_3^1 - 0.4239 T_4^1 &= 20 \end{aligned} \quad (\text{E2.5})$$

i=4

$$\begin{aligned} -\lambda T_3^1 + (1+2\lambda)T_4^1 - \lambda T_5^1 &= T_4^0 \\ -0.4239 T_3^1 + 1.8478 T_4^1 - (0.4239 \times 25) &= 20 \\ -0.4239 T_3^1 + 1.8478 T_4^1 - 10.598 &= 20 \\ -0.4239 T_3^1 + 1.8478 T_4^1 &= 30.598 \end{aligned} \quad (\text{E2.6})$$

The simultaneous linear equations (E2.3) – (E2.6) can be written in matrix form as

$$\begin{bmatrix} 1.8478 & -0.4239 & 0 & 0 \\ -0.4239 & 1.8478 & -0.4239 & 0 \\ 0 & -0.4239 & 1.8478 & -0.4239 \\ 0 & 0 & -0.4239 & 1.8478 \end{bmatrix} \begin{bmatrix} T_1^1 \\ T_2^1 \\ T_3^1 \\ T_4^1 \end{bmatrix} = \begin{bmatrix} 62.390 \\ 20 \\ 20 \\ 30.598 \end{bmatrix}$$

The above coefficient matrix is tri-diagonal. Special algorithms such as Thomas' algorithm can be used to solve simultaneous linear equation with tri-diagonal coefficient matrices. The solution is given by

$$\begin{bmatrix} T_1^1 \\ T_2^1 \\ T_3^1 \\ T_4^1 \end{bmatrix} = \begin{bmatrix} 39.451 \\ 24.792 \\ 21.438 \\ 21.477 \end{bmatrix}$$

Hence, the temperature at all the nodes at time,  $t = 3$  sec is

$$\begin{bmatrix} T_0^1 \\ T_1^1 \\ T_2^1 \\ T_3^1 \\ T_4^1 \\ T_5^1 \end{bmatrix} = \begin{bmatrix} 100 \\ 39.451 \\ 24.792 \\ 21.438 \\ 21.477 \\ 25 \end{bmatrix}$$

Temperature at the nodes inside the rod when  $t=6$  sec

$$\left. \begin{array}{l} T_0^2 = 100^\circ C \\ T_5^2 = 25^\circ C \end{array} \right\} \text{Boundary Condition (E2.1)}$$

For all the interior nodes, putting  $j = 1$  and  $i = 1, 2, 3, 4$  in Equation (11) gives the following equations

$i=1$

$$\begin{aligned} -\lambda T_0^2 + (1+2\lambda)T_1^2 - \lambda T_2^2 &= T_1^1 \\ (-0.4239 \times 100) + (1+2 \times 0.4239)T_1^2 - 0.4239T_2^2 &= 39.451 \\ -42.39 + 1.8478T_1^2 - 0.4239T_2^2 &= 39.451 \\ 1.8478T_1^2 - 0.4239T_2^2 &= 81.841 \end{aligned} \quad (\text{E2.7})$$

$i=2$

$$\begin{aligned} -\lambda T_1^2 + (1+2\lambda)T_2^2 - \lambda T_3^2 &= T_2^1 \\ -0.4239T_1^2 + 1.8478T_2^2 - 0.4239T_3^2 &= 24.792 \end{aligned} \quad (\text{E2.8})$$

$i=3$

$$\begin{aligned} -\lambda T_2^2 + (1+2\lambda)T_3^2 - \lambda T_4^2 &= T_3^1 \\ -0.4239T_2^2 + 1.8478T_3^2 - 0.4239T_4^2 &= 21.438 \end{aligned} \quad (\text{E2.9})$$

$i=4$

$$\begin{aligned} -\lambda T_3^2 + (1+2\lambda)T_4^2 - \lambda T_5^2 &= T_4^1 \\ -0.4239T_3^2 + 1.8478T_4^2 - (0.4239 \times 25) &= 21.477 \\ -0.4239T_3^2 + 1.8478T_4^2 - 10.598 &= 21.477 \\ -0.4239T_3^2 + 1.8478T_4^2 &= 32.075 \end{aligned} \quad (\text{E2.10})$$

The simultaneous linear equations (E2.7) – (E2.10) can be written in matrix form as

$$\begin{bmatrix} 1.8478 & -0.4239 & 0 & 0 \\ -0.4239 & 1.8478 & -0.4239 & 0 \\ 0 & -0.4239 & 1.8478 & -0.4239 \\ 0 & 0 & -0.4239 & 1.8478 \end{bmatrix} \begin{bmatrix} T_1^2 \\ T_2^2 \\ T_3^2 \\ T_4^2 \end{bmatrix} = \begin{bmatrix} 81.841 \\ 24.792 \\ 21.438 \\ 32.075 \end{bmatrix}$$

The solution of the above set of simultaneous linear equation is

$$\begin{bmatrix} T_1^2 \\ T_2^2 \\ T_3^2 \\ T_4^2 \end{bmatrix} = \begin{bmatrix} 51.326 \\ 30.669 \\ 23.876 \\ 22.836 \end{bmatrix}$$

Hence, the temperature at all the nodes at time,  $t = 6$  sec is

$$\begin{bmatrix} T_0^2 \\ T_1^2 \\ T_2^2 \\ T_3^2 \\ T_4^2 \\ T_5^2 \end{bmatrix} = \begin{bmatrix} 100 \\ 51.326 \\ 30.669 \\ 23.876 \\ 22.836 \\ 25 \end{bmatrix}$$

Temperature at the nodes inside the rod when  $t=9$  sec

$$\left. \begin{array}{l} T_0^3 = 100^\circ C \\ T_5^3 = 25^\circ C \end{array} \right\} \text{Boundary Condition (E2.1)}$$

For all the interior nodes, setting  $j = 2$  and  $i = 1, 2, 3, 4$  in Equation (11) gives the following equations

$i=1$

$$\begin{aligned} -\lambda T_0^3 + (1+2\lambda)T_1^3 - \lambda T_2^3 &= T_1^2 \\ (-0.4239 \times 100) + (1+2 \times 0.4239)T_1^3 - (0.4239 T_2^3) &= 51.326 \\ -42.39 + 1.8478 T_1^3 - 0.4239 T_2^3 &= 51.326 \\ 1.8478 T_1^3 - 0.4239 T_2^3 &= 93.716 \end{aligned} \quad (\text{E2.11})$$

$i=2$

$$\begin{aligned} -\lambda T_1^3 + (1+2\lambda)T_2^3 - \lambda T_3^3 &= T_2^2 \\ -0.4239 T_1^3 + 1.8478 T_2^3 - 0.4239 T_3^3 &= 30.669 \end{aligned} \quad (\text{E2.12})$$

$i=3$

$$\begin{aligned} -\lambda T_2^3 + (1+2\lambda)T_3^3 - \lambda T_4^3 &= T_3^2 \\ -0.4239 T_2^3 + 1.8478 T_3^3 - 0.4239 T_4^3 &= 23.876 \end{aligned} \quad (\text{E2.13})$$

$i=4$

$$\begin{aligned}
 -\lambda T_3^3 + (1+2\lambda)T_4^3 - \lambda T_5^3 &= T_4^2 \\
 -0.4239T_3^3 + 1.8478T_4^3 - (0.4239 \times 25) &= 22.836 \\
 -0.4239T_3^3 + 1.8478T_4^3 - 10.598 &= 22.836 \\
 -0.4239T_3^3 + 1.8478T_4^3 &= 33.434
 \end{aligned} \tag{E2.14}$$

The simultaneous linear equations (E2.11) – (E2.14) can be written in matrix form as

$$\begin{bmatrix} 1.8478 & -0.4239 & 0 & 0 \\ -0.4239 & 1.8478 & -0.4239 & 0 \\ 0 & -0.4239 & 1.8478 & -0.4239 \\ 0 & 0 & -0.4239 & 1.8478 \end{bmatrix} \begin{bmatrix} T_1^3 \\ T_2^3 \\ T_3^3 \\ T_4^3 \end{bmatrix} = \begin{bmatrix} 93.716 \\ 30.669 \\ 23.876 \\ 33.434 \end{bmatrix}$$

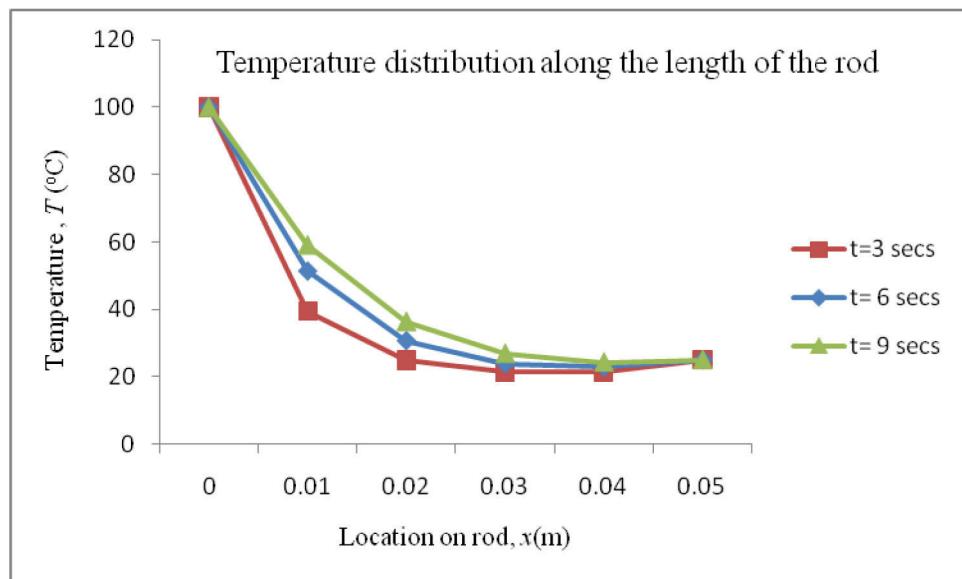
The solution of the above set of simultaneous linear equation is

$$\begin{bmatrix} T_1^3 \\ T_2^3 \\ T_3^3 \\ T_4^3 \end{bmatrix} = \begin{bmatrix} 59.043 \\ 36.292 \\ 26.809 \\ 24.243 \end{bmatrix}$$

Hence, the temperature at all the nodes at time,  $t = 9$  sec is

$$\begin{bmatrix} T_0^3 \\ T_1^3 \\ T_2^3 \\ T_3^3 \\ T_4^3 \\ T_5^3 \end{bmatrix} = \begin{bmatrix} 100 \\ 59.043 \\ 36.292 \\ 26.809 \\ 24.243 \\ 25 \end{bmatrix}$$

To better visualize the temperature variation at different locations at different times, the temperature distribution along the length of the rod at different times is plotted in Figure 6.



**Figure 6:** Temperature distribution in rod from implicit method

### Crank-Nicolson Method

The Crank-Nicolson method provides an alternative scheme to implicit method. The accuracy of Crank-Nicolson method is same in both space and time. In the implicit method,

the approximation of  $\frac{\partial^2 T}{\partial x^2}$  is of  $O(\Delta x)^2$  accuracy, while the approximation for  $\frac{\partial T}{\partial t}$  is of  $(\Delta t)$  accuracy.

The accuracy in the Crank-Nicolson method is achieved by approximating the derivative at the mid point of time step. To numerically solve PDEs such as Equation (2), one can use finite difference approximations of the partial derivatives. The left hand side of the second derivative is approximated at node  $i$  as the average value of the central divided difference approximation at time level  $j+1$  and time level  $j$ .

$$\left. \frac{\partial^2 T}{\partial x^2} \right|_{i,j} \approx \frac{1}{2} \left[ \frac{T_{i+1}^j - 2T_i^j + T_{i-1}^j}{(\Delta x)^2} + \frac{T_{i+1}^{j+1} - 2T_i^{j+1} + T_{i-1}^{j+1}}{(\Delta x)^2} \right] \quad (12)$$

The first derivative on the right side of Equation (2) is approximated using forward divided difference approximation at time level  $j+1$  and node  $i$  as

$$\left. \frac{\partial T}{\partial t} \right|_{i,j} \approx \frac{T_i^{j+1} - T_i^j}{\Delta t} \quad (13)$$

Substituting Equations (12) and (13) in Equation (2) gives

$$\alpha \cdot \frac{1}{2} \left[ \frac{T_{i+1}^j - 2T_i^j + T_{i-1}^j}{(\Delta x)^2} + \frac{T_{i+1}^{j+1} - 2T_i^{j+1} + T_{i-1}^{j+1}}{(\Delta x)^2} \right] = \frac{T_i^{j+1} - T_i^j}{\Delta t}$$

(14)

giving

$$-\lambda T_{i-1}^{j+1} + 2(1+\lambda)T_i^{j+1} - \lambda T_{i+1}^{j+1} = \lambda T_{i-1}^j + 2(1-\lambda)T_i^j + \lambda T_{i+1}^j \quad (15)$$

where

$$\lambda = \alpha \frac{\Delta t}{(\Delta x)^2}$$

Now Equation (15) is written for all nodes (except the external nodes). This will result in simultaneous linear equations that can be solved to find the temperature at a particular time.

### Example 3

A rod of steel is subjected to a temperature of  $100^\circ C$  on the left end and  $25^\circ C$  on the right end. If the rod is of length  $0.05m$ , use Crank-Nicolson method to find the temperature distribution in the rod from  $t = 0$  to  $t = 9$  seconds. Use  $\Delta x = 0.01m$ ,  $\Delta t = 3s$ .

Given

$$k = 54 \frac{W}{m \cdot K}, \rho = 7800 \frac{kg}{m^3}, C = 490 \frac{J}{kg \cdot K}.$$

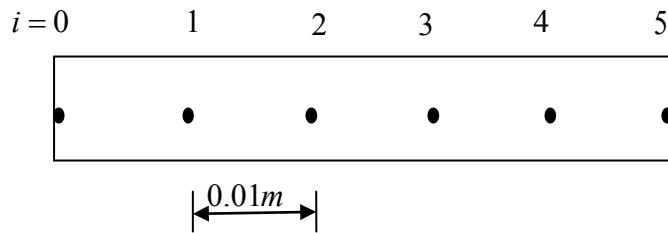
The initial temperature of the rod is  $20^\circ C$ .

**Solution**

$$\begin{aligned}\alpha &= \frac{k}{\rho C} \\ &= \frac{54}{7800 \times 490} \\ &= 1.4129 \times 10^{-5} \text{ m}^2 / \text{s}\end{aligned}$$

Then

$$\begin{aligned}\lambda &= \alpha \frac{\Delta t}{(\Delta x)^2} \\ &= 1.412 \times 10^{-5} \frac{3}{(0.01)^2} \\ &= 0.4239\end{aligned}$$



**Figure 7:** Schematic diagram showing the node representation in the model

The boundary conditions are

$$\left. \begin{array}{l} T_0^j = 100^\circ\text{C} \\ T_5^j = 25^\circ\text{C} \end{array} \right\} \text{ for } j = 0, 1, 2, 3 \quad (\text{E3.1})$$

The initial temperature of the rod is  $20^\circ\text{C}$ , that is, all the temperatures of the nodes inside the rod are at  $20^\circ\text{C}$  at,  $t = 0$  except for the boundary nodes given by Equation (E3.1). This could be represented as

$$T_i^0 = 20^\circ\text{C}, \text{ for } i = 1, 2, 3, 4. \quad (\text{E3.2})$$

Initial temperature at the nodes inside the rod (when  $t=0$  sec)

$$T_0^0 = 100^\circ\text{C} \quad \text{from Equation (E3.1)}$$

$$\left. \begin{array}{l} T_1^0 = 20^\circ\text{C} \\ T_2^0 = 20^\circ\text{C} \\ T_3^0 = 20^\circ\text{C} \\ T_4^0 = 20^\circ\text{C} \end{array} \right\} \text{ from Equation (E3.2)}$$

$$T_5^0 = 25^\circ C \quad \text{from Equation (E3.1)}$$

Temperature at the nodes inside the rod when  $t=3$  sec

$$\left. \begin{array}{l} T_0^1 = 100^\circ C \\ T_5^1 = 25^\circ C \end{array} \right\} \text{Boundary Condition (E3.1)}$$

For all the interior nodes, setting  $j = 0$  and  $i = 1, 2, 3, 4$  in Equation (15) gives the following equations

$i=1$

$$\begin{aligned} -\lambda T_0^1 + 2(1+\lambda)T_1^1 - \lambda T_2^1 &= \lambda T_0^0 + 2(1-\lambda)T_1^0 + \lambda T_2^0 \\ (-0.4239 \times 100) + 2(1+0.4239)T_1^1 - 0.4239 T_2^1 &= (0.4239)100 + 2(1-0.4239)20 + (0.4239)20 \\ -42.39 + 2.8478 T_1^1 - 0.4239 T_2^1 &= 42.39 + 23.044 + 8.478 \\ 2.8478 T_1^1 - 0.4239 T_2^1 &= 116.30 \end{aligned} \quad (\text{E3.3})$$

$i=2$

$$\begin{aligned} -\lambda T_1^1 + 2(1+\lambda)T_2^1 - \lambda T_3^1 &= \lambda T_1^0 + 2(1-\lambda)T_2^0 + \lambda T_3^0 \\ -0.4239 T_1^1 + 2(1+0.4239)T_2^1 - 0.4239 T_3^1 &= (0.4239)20 + 2(1-0.4239)20 + (0.4239)20 \\ -0.4239 T_1^1 + 2.8478 T_2^1 - 0.4239 T_3^1 &= 40.000 \end{aligned} \quad (\text{E3.4})$$

$i=3$

$$\begin{aligned} -\lambda T_2^1 + 2(1+\lambda)T_3^1 - \lambda T_4^1 &= \lambda T_2^0 + 2(1-\lambda)T_3^0 + \lambda T_4^0 \\ -0.4239 T_2^1 + 2(1+0.4239)T_3^1 - 0.4239 T_4^1 &= (0.4239)20 + 2(1-0.4239)20 + (0.4239)20 \\ -0.4239 T_2^1 + 2.8478 T_3^1 - 0.4239 T_4^1 &= 40.000 \end{aligned} \quad (\text{E3.5})$$

$i=4$

$$\begin{aligned} -\lambda T_3^1 + 2(1+\lambda)T_4^1 - \lambda T_5^1 &= \lambda T_3^0 + 2(1-\lambda)T_4^0 + \lambda T_5^0 \\ -0.4239 T_3^1 + 2(1+0.4239)T_4^1 - (0.4239)25 &= (0.4239)20 + 2(1-0.4239)20 + (0.4239)25 \\ -0.4239 T_3^1 + 2.8478 T_4^1 - 10.598 &= 8.478 + 23.044 + 10.598 \\ -0.4239 T_3^1 + 2.8478 T_4^1 &= 52.718 \end{aligned} \quad (\text{E3.6})$$

The coefficient matrix in the above set of equations is tridiagonal. Special algorithms such as Thomas' algorithm are used to solve equation with tridiagonal coefficient matrices

$$\begin{bmatrix} 2.8478 & -0.4239 & 0 & 0 \\ -0.4239 & 2.8478 & -0.4239 & 0 \\ 0 & -0.4239 & 2.8478 & -0.4239 \\ 0 & 0 & -0.4239 & 2.8478 \end{bmatrix} \begin{bmatrix} T_1^1 \\ T_2^1 \\ T_3^1 \\ T_4^1 \end{bmatrix} = \begin{bmatrix} 116.30 \\ 40.000 \\ 40.000 \\ 52.718 \end{bmatrix}$$

The above matrix is tridiagonal. Solving the above matrix we get

$$\begin{bmatrix} T_1^1 \\ T_2^1 \\ T_3^1 \\ T_4^1 \end{bmatrix} = \begin{bmatrix} 44.372 \\ 23.746 \\ 20.797 \\ 21.607 \end{bmatrix}$$

Hence, the temperature at all the nodes at time,  $t = 3$  sec is

$$\begin{bmatrix} T_0^1 \\ T_1^1 \\ T_2^1 \\ T_3^1 \\ T_4^1 \\ T_5^1 \end{bmatrix} = \begin{bmatrix} 100 \\ 44.372 \\ 23.746 \\ 20.797 \\ 21.607 \\ 25 \end{bmatrix}$$

#### Temperature at the nodes inside the rod when $t=6$ sec

$$\left. \begin{array}{l} T_0^2 = 100^\circ C \\ T_5^2 = 25^\circ C \end{array} \right\} \text{Boundary Condition (E3.1)}$$

For all the interior nodes, putting  $j = 1$  and  $i = 1, 2, 3, 4$  in Equation (15) gives the following equations

$i=1$

$$\begin{aligned} -\lambda T_0^2 + 2(1+\lambda)T_1^2 - \lambda T_2^2 &= \lambda T_0^1 + 2(1-\lambda)T_1^1 + \lambda T_2^1 \\ (-0.4239 \times 100) + 2(1+0.4239)T_1^2 - 0.4239T_2^2 &= \\ (0.4239)100 + 2(1-0.4239)44.372 + (0.4239)23.746 & \\ -42.39 + 2.8478T_1^2 - 0.4239T_2^2 &= 42.39 + 51.125 + 10.066 \\ 2.8478T_1^2 - 0.4239T_2^2 &= 145.971 \end{aligned} \quad (\text{E3.7})$$

$i=2$

$$\begin{aligned} -\lambda T_1^2 + 2(1+\lambda)T_2^2 - \lambda T_3^2 &= \lambda T_1^1 + 2(1-\lambda)T_2^1 + \lambda T_3^1 \\ -0.4239T_1^2 + 2(1+0.4239)T_2^2 - 0.4239T_3^2 &= \\ (0.4239)44.372 + 2(1-0.4239)23.746 + (0.4239)20.797 & \\ -0.4239T_1^2 + 2.8478T_2^2 - 0.4239T_3^2 &= 18.809 + 27.360 + 8.8158 \\ -0.4239T_1^2 + 2.8478T_2^2 - 0.4239T_3^2 &= 54.985 \end{aligned} \quad (\text{E3.8})$$

$i=3$

$$\begin{aligned} -\lambda T_2^2 + 2(1+\lambda)T_3^2 - \lambda T_4^2 &= \lambda T_2^1 + 2(1-\lambda)T_3^1 + \lambda T_4^1 \\ -0.4239T_2^2 + 2(1+0.4239)T_3^2 - 0.4239T_4^2 &= \\ (0.4239)23.746 + 2(1-0.4239)20.797 + (0.4239)21.607 & \\ -0.4239T_2^2 + 2.8478T_3^2 - 0.4239T_4^2 &= 10.066 + 23.962 + 9.1592 \\ -0.4239T_2^2 + 2.8478T_3^2 - 0.4239T_4^2 &= 43.187 \end{aligned} \quad (\text{E3.9})$$

$i=4$

$$\begin{aligned}
-\lambda T_3^2 + 2(1+\lambda)T_4^2 - \lambda T_5^2 &= \lambda T_3^1 + 2(1-\lambda)T_4^1 + \lambda T_5^1 \\
-0.4239T_3^2 + 2(1+0.4239)T_4^2 - (0.4239)25 &= \\
(0.4239)20.797 + 2(1-0.4239)21.607 + (0.4239)25 & \\
-0.4239T_3^2 + 2.8478T_4^2 - 10.598 &= 8.8158 + 24.896 + 10.598 \\
-0.4239T_3^2 + 2.8478T_4^2 &= 54.908 \tag{E3.10}
\end{aligned}$$

The simultaneous linear equations (E3.7) – (E3.10) can be written in matrix form as

$$\begin{bmatrix} 2.8478 & -0.4239 & 0 & 0 \\ -0.4239 & 2.8478 & -0.4239 & 0 \\ 0 & -0.4239 & 2.8478 & -0.4239 \\ 0 & 0 & -0.4239 & 2.8478 \end{bmatrix} \begin{bmatrix} T_1^2 \\ T_2^2 \\ T_3^2 \\ T_4^2 \end{bmatrix} = \begin{bmatrix} 145.971 \\ 54.985 \\ 43.187 \\ 54.908 \end{bmatrix}$$

Solving the above set of equations, we get

$$\begin{bmatrix} T_1^2 \\ T_2^2 \\ T_3^2 \\ T_4^2 \end{bmatrix} = \begin{bmatrix} 55.883 \\ 31.075 \\ 23.174 \\ 22.730 \end{bmatrix}$$

Hence, the temperature at all the nodes at time,  $t = 6$  sec is

$$\begin{bmatrix} T_0^2 \\ T_1^2 \\ T_2^2 \\ T_3^2 \\ T_4^2 \\ T_5^2 \end{bmatrix} = \begin{bmatrix} 100 \\ 55.883 \\ 31.075 \\ 23.174 \\ 22.730 \\ 25 \end{bmatrix}$$

Temperature at the nodes inside the rod when  $t=9$  sec

$$\left. \begin{array}{l} T_0^3 = 100^\circ C \\ T_5^3 = 25^\circ C \end{array} \right\} \text{Boundary Condition (E3.1)}$$

For all the interior nodes, setting  $j = 2$  and  $i = 1, 2, 3, 4$  in Equation (15) gives the following equations

$i=1$

$$\begin{aligned}
-\lambda T_0^3 + 2(1+\lambda)T_1^3 - \lambda T_2^3 &= \lambda T_0^2 + 2(1-\lambda)T_1^2 + \lambda T_2^2 \\
(-0.4239 \times 100) + 2(1+0.4239)T_1^3 - 0.4239T_2^3 &= \\
(0.4239)100 + 2(1-0.4239)55.883 + (0.4239)31.075 & \\
-42.39 + 2.8478T_1^3 - 0.4239T_2^3 &= 42.39 + 64.388 + 13.173 \\
2.8478T_1^3 - 0.4239T_2^3 &= 162.34 \tag{E3.11}
\end{aligned}$$

$i=2$

$$\begin{aligned}
-\lambda T_1^3 + 2(1+\lambda)T_2^3 - \lambda T_3^3 &= \lambda T_1^2 + 2(1-\lambda)T_2^2 + \lambda T_3^2 \\
-0.4239T_1^3 + 2(1+0.4239)T_2^3 - 0.4239T_3^3 &= \\
(0.4239)55.883 + 2(1-0.4239)31.075 + (0.4239)23.174 & \\
-0.4239T_1^3 + 2.8478T_2^3 - 0.4239T_3^3 &= 23.689 + 35.805 + 9.8235 \\
-0.4239T_1^3 + 2.8478T_2^3 - 0.4239T_3^3 &= 69.318
\end{aligned} \tag{E3.12}$$

i=3

$$\begin{aligned}
-\lambda T_2^3 + 2(1+\lambda)T_3^3 - \lambda T_4^3 &= \lambda T_2^2 + 2(1-\lambda)T_3^2 + \lambda T_4^2 \\
-0.4239T_2^3 + 2(1+0.4239)T_3^3 - 0.4239T_4^3 &= \\
(0.4239)31.075 + 2(1-0.4239)23.174 + (0.4239)22.730 & \\
-0.4239T_2^3 + 2.8478T_3^3 - 0.4239T_4^3 &= 13.173 + 26.701 + 9.635 \\
-0.4239T_2^3 + 2.8478T_3^3 - 0.4239T_4^3 &= 49.509
\end{aligned} \tag{E3.13}$$

i=4

$$\begin{aligned}
-\lambda T_3^3 + 2(1+\lambda)T_4^3 - \lambda T_5^3 &= \lambda T_3^2 + 2(1-\lambda)T_4^2 + \lambda T_5^2 \\
-0.4239T_3^3 + 2(1+0.4239)T_4^3 - (0.4239)25 &= \\
(0.4239)23.174 + 2(1-0.4239)22.730 + (0.4239)25 & \\
-0.4239T_3^3 + 2.8478T_4^3 - 10.598 &= 9.8235 + 26.190 + 10.598 \\
-0.4239T_3^3 + 2.8478T_4^3 &= 57.210
\end{aligned} \tag{E3.14}$$

The simultaneous linear equations (E3.11) – (E3.14) can be written in matrix form as

$$\begin{bmatrix} 2.8478 & -0.4239 & 0 & 0 \\ -0.4239 & 2.8478 & -0.4239 & 0 \\ 0 & -0.4239 & 2.8478 & -0.4239 \\ 0 & 0 & -0.4239 & 2.8478 \end{bmatrix} \begin{bmatrix} T_1^3 \\ T_2^3 \\ T_3^3 \\ T_4^3 \end{bmatrix} = \begin{bmatrix} 162.34 \\ 69.318 \\ 49.509 \\ 57.210 \end{bmatrix}$$

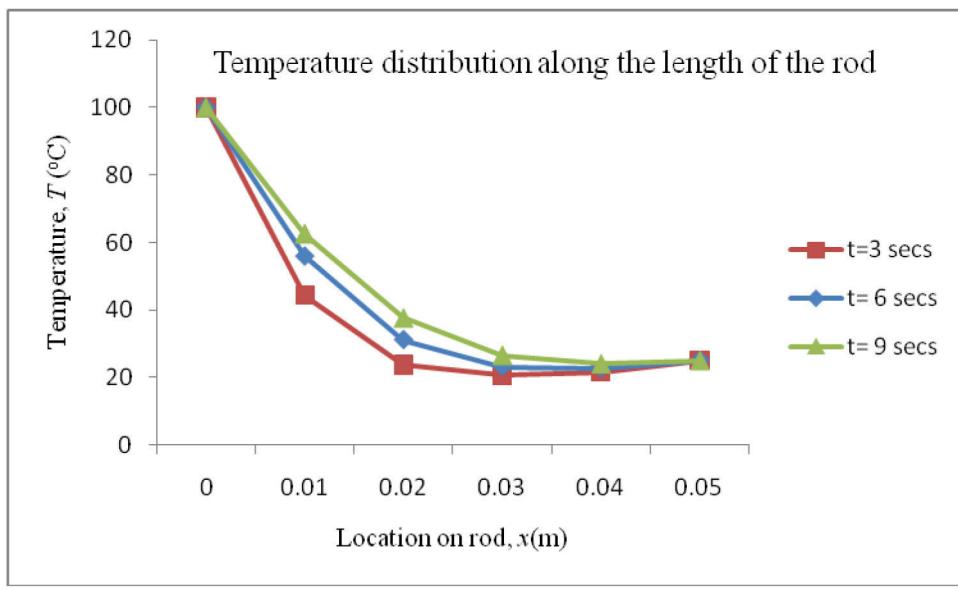
Solving the above set of equations, we get

$$\begin{bmatrix} T_1^3 \\ T_2^3 \\ T_3^3 \\ T_4^3 \end{bmatrix} = \begin{bmatrix} 62.604 \\ 37.613 \\ 26.562 \\ 24.042 \end{bmatrix}$$

Hence, the temperature at all the nodes at time,  $t = 9$  sec is

$$\begin{bmatrix} T_0^3 \\ T_1^3 \\ T_2^3 \\ T_3^3 \\ T_4^3 \\ T_5^3 \end{bmatrix} = \begin{bmatrix} 100 \\ 62.604 \\ 37.613 \\ 26.562 \\ 24.042 \\ 25 \end{bmatrix}$$

To better visualize the temperature variation at different locations at different times, the temperature distribution along the length of the rod at different times is plotted in Figure 8.



**Figure 8:** Temperature distribution in rod from Crank-Nicolson method

## Analytical Method

### Appendix A

The parabolic heat conduction equation given by Equation (2) is formulated as

$$\alpha \frac{\partial^2 T}{\partial x^2} = \frac{\partial T}{\partial t} \quad 0 < x < 0.05, \quad t > 0$$

with boundary conditions

$$T = 100^\circ C \text{ at } x = 0, \quad t > 0 \quad (16)$$

$$T = 25^\circ C \text{ at } x = 0.05, \quad t > 0 \quad (17)$$

and initial conditions

$$T = 20^\circ C \text{ at } t = 0, \quad 0 < x < 0.05 \quad (18)$$

We split the problem into a steady state problem and a transient (homogeneous) problem. The solutions of the steady state problem and transient problem are found separately and by applying the principle of superposition, the final solution would be obtained. This formulation can be represented as

$$T(x, t) = T_s(x) + T_h(x, t) \quad (19)$$

where

$T_s$  = solution for steady state problem,

$T_h$  = solution for transient problem.

#### Steady State Solution

Since the temperature at steady state is not changing,  $\frac{\partial T}{\partial t} = 0$ , the steady state problem is formulated as

$$\frac{d^2 T_s}{dx^2} = 0, \quad 0 < x < 0.05 \quad (20)$$

with boundary conditions

$$T_s = 100^\circ C \text{ at } x = 0 \quad (21)$$

$$T_s = 25^\circ C \text{ at } x = 0.05 \quad (22)$$

The solution to Equation (20) is given by integrating it on both sides to give

$$\frac{dT_s}{dx} = A$$

where  $A$  is a constant of integration and by integrating again to give

$$T_s = Ax + B \quad (23)$$

where  $B$  is another constant of integration. By substituting the boundary condition (21), we obtain

$$A(0) + B = 100$$

$$B = 100$$

By substituting the boundary condition (22), we obtain

$$A(0.05) + 100 = 25$$

$$\begin{aligned} A &= \frac{-75}{0.05} \\ &= -1500 \end{aligned}$$

Plugging back the values of  $A$  and  $B$  in Equation (23), we get the steady state solution as

$$T_s = -1500x + 100 \quad (24)$$

### Transient Solution

The transient problem is formulated as

$$\alpha \frac{\partial^2 T_h}{\partial x^2} = \frac{\partial T_h}{\partial t}, \quad 0 < x < 0.05 \quad (25)$$

with boundary conditions

$$T_h = 0^\circ C \text{ at } x = 0 \quad (26)$$

$$T_h = 0^\circ C \text{ at } x = 0.05 \quad (27)$$

Note: from Equation (19),

$$T(x, t) = T_s(x) + T_h(x, t)$$

and by substituting Equations (21) and (22), the boundary conditions of  $T_h$  are obtained.

Initial conditions for the transient problem are hence given by

$$\begin{aligned} T_h &= 20 - T_s, \quad t = 0, \quad 0 < x < 0.05 \\ &= 20 - (-1500x + 100) \\ &= 20 + 1500x - 100 \\ &= 1500x - 80, \quad t = 0, \quad 0 < x < 0.05 \end{aligned} \quad (28)$$

To obtain solution for the transient problem, let us assume  $T_h(x, t)$  is function of the product of a spatial function and a temperature function. That is

$$T_h(x, t) = X(x)\tau(t) \quad (29)$$

Substituting Equation (29) in Equation (25), we get

$$\begin{aligned} \alpha\tau \frac{d^2 X}{dx^2} &= X \frac{d\tau}{dt} \\ \frac{1}{X} \frac{d^2 X}{dx^2} &= \frac{1}{\alpha\tau} \frac{d\tau}{dt} \end{aligned} \quad (30)$$

The left hand side of Equation (30) represents the spatial term and the right hand side represents the temporal (time) term. We will attempt to find the solutions of the spatial and temporal term independently. To do so, let us assume that both the left hand side and the right hand side of the Equation (30) is equal to a constant  $-\beta^2$  (say)

$$\frac{1}{X} \frac{d^2 X}{dx^2} = \frac{1}{\alpha\tau} \frac{d\tau}{dt} = -\beta^2 \quad (31)$$

### *Spatial solution*

Taking just the spatial term from Equation (31), we have

$$\frac{1}{X} \frac{d^2 X}{dx^2} = -\beta^2$$

$$\frac{d^2 X}{dx^2} + \beta^2 X = 0 \quad (32)$$

The Equation (32) is a homogeneous second order ordinary differential equation. These type of equations have the solution of the form  $X(x) = e^{mx}$ . Substituting  $X(x) = e^{mx}$  in Equation (32) we get,

$$\begin{aligned} m^2 e^{mx} + \beta^2 e^{mx} &= 0 \\ e^{mx} (m^2 + \beta^2) &= 0 \\ m^2 + \beta^2 &= 0 \\ m_1, m_2 &= i\beta, -i\beta \end{aligned}$$

From the values of  $m_1$  and  $m_2$ , the solution of  $X(x)$  is written of the form

$$X(x) = C \cos(\beta x) + D \sin(\beta x) \quad (33)$$

### *Temporal solution*

Taking just the temporal term from Equation (31), we have

$$\begin{aligned} \frac{1}{\alpha\tau} \frac{d\tau}{dt} &= -\beta^2 \\ \frac{d\tau}{dt} + \alpha\tau\beta^2 &= 0 \end{aligned} \quad (34)$$

The above equation is a homogeneous first order ordinary differential equation. These type of equations have the solution of the form  $\tau(t) = e^{mt}$ . Substituting  $\tau(t) = e^{mt}$  in Equation (34) we get

$$\begin{aligned} me^{mt} + \alpha\beta^2 e^{mt} &= 0 \\ e^{mt} (m + \alpha\beta^2) &= 0 \\ m + \alpha\beta^2 &= 0 \\ m &= -\alpha\beta^2 \end{aligned}$$

From the value of  $m$ , the solution of  $\tau(t)$  is written as

$$\tau(t) = E e^{-\alpha\beta^2 t} \quad (35)$$

Substituting Equations (33) and (35) in Equation (29), we have

$$\begin{aligned} T_h(x, t) &= E e^{-\alpha\beta^2 t} [C \cos(\beta x) + D \sin(\beta x)] \\ T_h(x, t) &= e^{-\alpha\beta^2 t} [F \cos(\beta x) + G \sin(\beta x)] \end{aligned} \quad (36)$$

Substituting boundary condition represented by Equation (26) in Equation (36) gives

$$\begin{aligned} e^{-\alpha\beta^2 t} [F \cos(\beta \cdot 0) + G \sin(\beta \cdot 0)] &= 0 \\ e^{-\alpha\beta^2 t} [F \cdot 1 + G \cdot 0] &= 0 \\ e^{-\alpha\beta^2 t} [F] &= 0 \end{aligned}$$

Since,  $e^{-\alpha\beta^2 t}$  cannot be zero,  $F = 0$ . Now substituting  $F = 0$  in Equation (36) gives

$$T_h(x, t) = G e^{-\alpha\beta^2 t} \sin(\beta x) \quad (37)$$

Substituting boundary condition represented by Equation (27) in Equation (37) gives

$$Ge^{-\alpha\beta^2 t} \sin(0.4\beta) = 0$$

$$\sin(0.05\beta) = 0$$

$$0.05\beta = n\pi$$

$$\begin{aligned}\beta &= \frac{n\pi}{0.05} \\ &= 20n\pi\end{aligned}$$

Substituting the value of  $\beta$  in Equation (37) gives

$$T_h(x, t) = Ge^{-\alpha(20n\pi)^2 t} \sin(20n\pi x)$$

As the general solution can have any value of  $n$ ,

$$T_h(x, t) = \sum_{n=1}^{\infty} G_n e^{-\alpha(20n\pi)^2 t} \sin(20n\pi x) \quad (38)$$

Substituting the initial condition

$$T_h(x, 0) = (1500x - 80)^\circ C$$

from Equation (28) in Equation (38)

$$\sum_{n=1}^{\infty} G_n \sin(20n\pi x) = 1500x - 80$$

Multiplying both sides by  $\sin(20m\pi x)$  and integrating from 0 to 0.05 gives

$$\begin{aligned}\sum_{n=1}^{\infty} \int_0^{0.05} G_n \sin(20n\pi x) \sin(20m\pi x) dx &= \int_0^{0.05} (1500x - 80) \sin(20m\pi x) dx \\ \sum_{n=1}^{\infty} \frac{G_n}{2} \int_0^{0.05} 2 \sin(20n\pi x) \sin(20m\pi x) dx &= \int_0^{0.05} (1500x - 80) \sin(20m\pi x) dx \\ \sum_{n=1}^{\infty} \frac{G_n}{2} \left[ \int_0^{0.05} \cos(20(m-n)\pi x) dx - \int_0^{0.05} \cos(20(m+n)\pi x) dx \right] &= \\ 1500 \int_0^{0.05} x \sin(20m\pi x) dx - \int_0^{0.05} 80 \sin(20m\pi x) dx &\end{aligned}$$

Substituting the following in the above equation,

$$\int_0^{0.05} \cos(20(m-n)\pi x) dx = 0, \quad m \neq n$$

$$\int_0^{0.05} \cos(20(m-n)\pi x) dx = 0.05, \quad m = n$$

$$\int_0^{0.05} \cos(20(m+n)\pi x) dx = 0, \quad \text{for any } m$$

we get

$$\frac{G_m}{2} 0.05 = 1500 \int_0^{0.05} x \sin(20m\pi x) dx - \int_0^{0.05} 80 \sin(20m\pi x) dx$$

$$\begin{aligned}
&= 1500 \left[ \left[ \frac{-x \cos(20m\pi)}{20m\pi} \right]_0^{0.05} + \frac{1}{20m\pi} \int_0^{0.05} \cos(20m\pi) dx \right] - 80 \int_0^{0.05} \sin(20m\pi) dx \\
&= 1500 \left[ \left[ \frac{-x \cos(20m\pi)}{20m\pi} \right]_0^{0.05} + \frac{1}{20m\pi} \int_0^{0.05} \cos(20m\pi) dx \right] - 80 \int_0^{0.05} \sin(20m\pi) dx \\
&= 1500 \left[ \frac{-0.05 \cos(m\pi) + 0}{20m\pi} \right] + 80 \left[ \frac{(-1)^m - 1}{20m\pi} \right] \\
G_m &= \frac{-150(-1)^m}{m\pi} + \frac{160}{m\pi} [(-1)^m - 1] \\
&= \frac{10(-1)^m - 160}{m\pi}
\end{aligned} \tag{39}$$

Substituting Equation (39) in Equation (38), we get

$$T_h(x, t) = \sum_{m=1}^{\infty} \left\{ \frac{10(-1)^m - 160}{m\pi} \right\} e^{-\alpha[2.5m\pi]^2 t} \sin(2.5m\pi x) \tag{40}$$

Substituting Equations (40) and (24) in Equation (19) we have

$$T(x, t) = -1500x + 100 + \sum_{m=1}^{\infty} \left\{ \frac{10(-1)^m - 160}{m\pi} \right\} e^{-\alpha[20m\pi]^2 t} \sin(20m\pi x) \tag{41}$$

Now

$$\begin{aligned}
\alpha &= \frac{k}{\rho C} \\
&= \frac{54}{7800 \times 490} \\
&= 1.4129 \times 10^{-5} \text{ m}^2 / \text{s}
\end{aligned}$$

and substituting the value of  $\alpha$  in Equation (40) gives

$$T(x, t) = -1500x + 100 + \sum_{m=1}^{\infty} \left\{ \frac{10(-1)^m - 160}{m\pi} \right\} e^{-1.4129 \times 10^{-5} [20m\pi]^2 t} \sin(20m\pi x) \tag{42}$$

Equation (42) is the analytical solution of the problem. Substituting the values of  $x$  and  $t$  gives the temperature inside the rod at a particular location and time. For example using the analytical solution, we will find the temperature of the rod at the first node, that is,  $x = 0.01 \text{ m}$  when  $t = 9 \text{ secs}$ .

$$\begin{aligned}
T(0.01, 9) &= -1500(0.01) + 100 + \sum_{m=1}^{\infty} \left\{ \frac{10(-1)^m - 160}{m\pi} \right\} e^{-1.4129 \times 10^{-5} [20m\pi]^2 \times 9} \sin(0.2m\pi) \\
&= 62.510^\circ \text{C}
\end{aligned}$$

Similarly using Equation (42), the temperature of the rod at any location at any time can be found by substituting the corresponding values of  $x$  and  $t$ .

### Comparison of the three numerical methods

To compare all three numerical methods with the analytical solution, the temperature values obtained at all the interior nodes at time,  $t = 9$  sec are presented in the Table 1. From Table 1, it is clear that among the numerical methods used to solve partial differential equations, Crank-Nicolson method provides better accuracy compared to the other two numerical methods ( Explicit Method and Implicit Method) explained in this chapter.

**Table 1:** Comparison of temperature obtained at interior nodes using different methods discussed in this chapter (absolute true error is given in parenthesis)

Temperature at Nodes	Explicit Method (°C)	Implicit Method (°C)	Crank-Nicolson Method (°C)	Analytical Solution (°C)
$T_1^3$	65.953(3.443)	59.043(3.467)	62.604(0.094)	62.510
$T_2^3$	39.132(2.048)	36.292(0.792)	37.613(0.529)	37.084
$T_3^3$	27.266(1.422)	26.809(0.965)	26.562(0.282)	25.844
$T_4^3$	22.872(0.738)	24.243(0.633)	24.042(0.432)	23.610

---

### PARTIAL DIFFERENTIAL EQUATIONS

---

Topic	Parabolic Differential Equations
Summary	Textbook notes for the parabolic partial differential equations
Major	All engineering majors
Authors	Autar Kaw, Sri Harsha Garapati
Date	March 17, 2011
Web Site	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

---

## Chapter 10.03

# Elliptic Partial Differential Equations

After reading this chapter, you should be able to:

1. use numerical methods to solve elliptic partial differential equations by direct method, Gauss-Seidel method, and Gauss-Seidel method with over relaxation.

The general second order linear PDE with two independent variables and one dependent variable is given by

$$A \frac{\partial^2 u}{\partial x^2} + B \frac{\partial^2 u}{\partial x \partial y} + C \frac{\partial^2 u}{\partial y^2} + D = 0 \quad (1)$$

where  $A, B, C$  are functions of the independent variables  $x$  and  $y$ , and  $D$  can be a function of  $x, y, u, \frac{\partial u}{\partial x}$  and  $\frac{\partial u}{\partial y}$ . Equation (1) is considered to be elliptic if

$$B^2 - 4AC < 0 \quad (2)$$

One popular example of an elliptic second order linear partial differential equation is the Laplace equation which is of the form

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \quad (3)$$

As

$$A = 1, B = 0, C = 1, D = 0$$

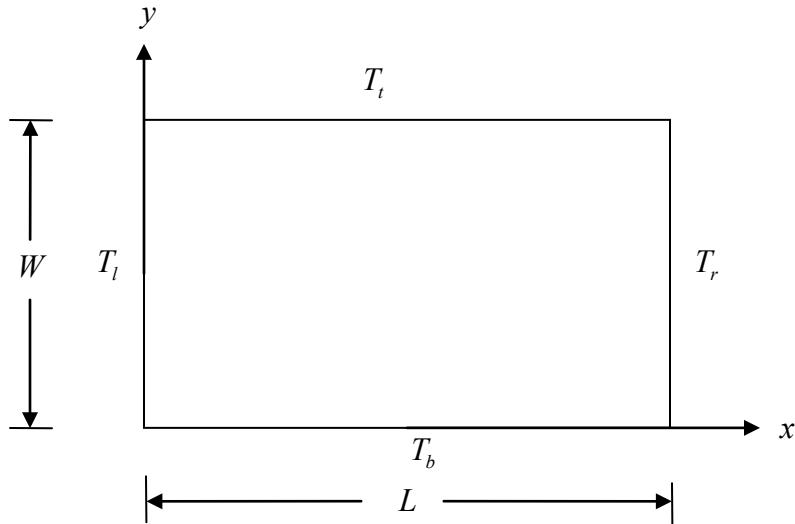
then

$$\begin{aligned} B^2 - 4AC &= 0 - 4(1)(1) \\ &= -4 < 0 \end{aligned}$$

Hence equation (3) is elliptic.

### The Direct Method of Solving Elliptic PDEs

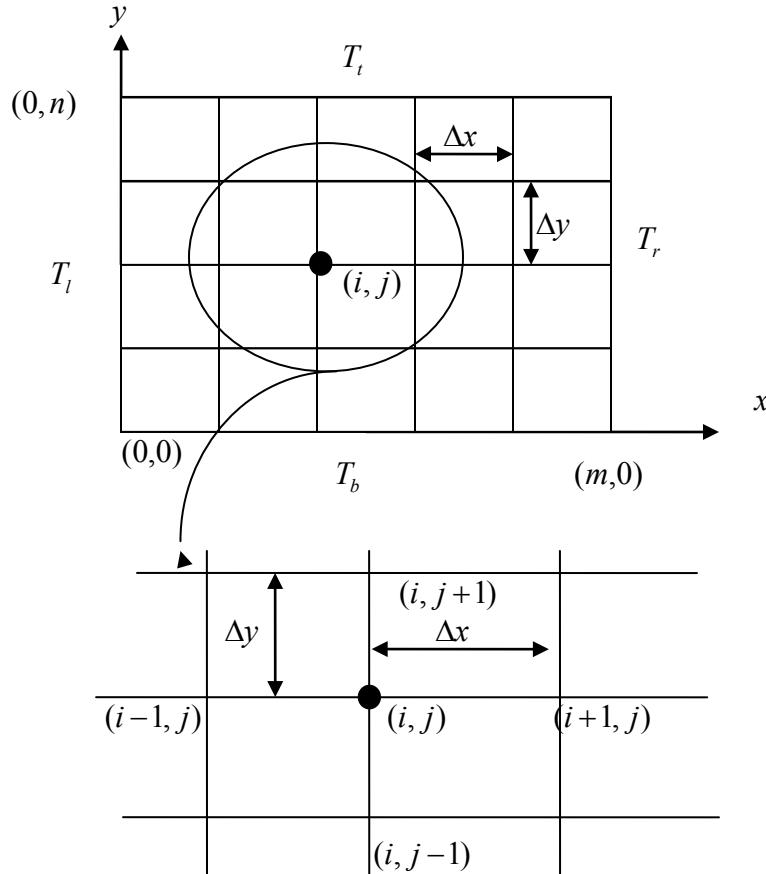
Let's find the solution via a specific physical example. Take a rectangular plate as shown in Fig. 1 where each side of the plate is maintained at a specific temperature. We are interested in finding the temperature within the plate at steady state. No heat sinks or sources exist in the problem.



**Figure 1:** Schematic diagram of the plate with the temperature boundary conditions  
The partial differential equation that governs the temperature  $T(x, y)$  is given by

$$\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} = 0 \quad (4)$$

To find the temperature within the plate, we divide the plate area by a grid as shown in Figure 2.



**Figure 2:** Plate area divided into a grid

The length  $L$  along the  $x$ -axis is divided into  $m$  equal segments, while the width  $W$  along the  $y$ -axis is divided into  $n$  equal segments, hence giving

$$\Delta x = \frac{L}{m} \quad (5)$$

$$\Delta y = \frac{W}{n} \quad (6)$$

Now we will apply the finite difference approximation of the partial derivatives at a general interior node  $(i, j)$ .

$$\left. \frac{\partial^2 T}{\partial x^2} \right|_{i,j} \cong \frac{T_{i+1,j} - 2T_{i,j} + T_{i-1,j}}{(\Delta x)^2} \quad (7)$$

$$\left. \frac{\partial^2 T}{\partial y^2} \right|_{i,j} \cong \frac{T_{i,j+1} - 2T_{i,j} + T_{i,j-1}}{(\Delta y)^2} \quad (8)$$

Equations (7) and (8) are central divided difference approximations of the second derivatives. Substituting Equations (7) and (8) in Equation (4), we get

$$\frac{T_{i+1,j} - 2T_{i,j} + T_{i-1,j}}{(\Delta x)^2} + \frac{T_{i,j+1} - 2T_{i,j} + T_{i,j-1}}{(\Delta y)^2} = 0 \quad (9)$$

For a grid with

$$\Delta x = \Delta y$$

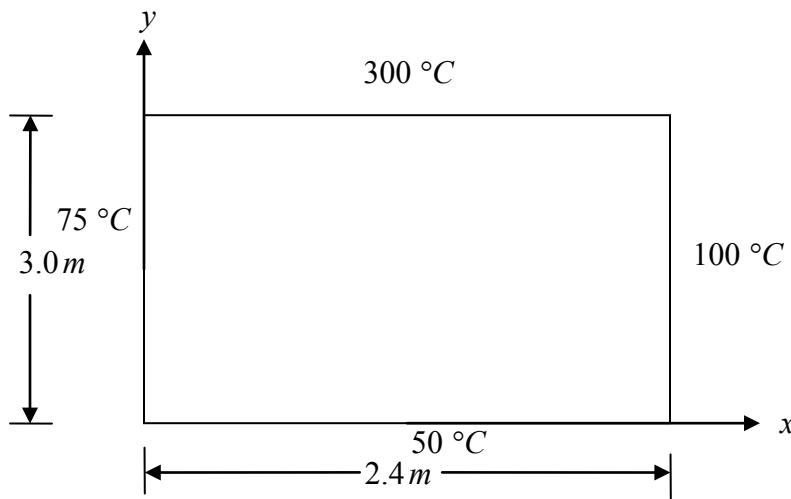
Equation (9) can be simplified as

$$T_{i+1,j} + T_{i-1,j} + T_{i,j+1} + T_{i,j-1} - 4T_{i,j} = 0 \quad (10)$$

Now we can write this equation at all the interior nodes of the plate, that is  $(m-1) \times (n-1)$  nodes. This will result in an equal number of equations and unknowns. The unknowns are the temperatures at the interior  $(m-1) \times (n-1)$  nodes. Solving these equations will give us the two-dimensional profile of the temperature inside the plate.

### Example 1

A plate  $2.4\text{ m} \times 3.0\text{ m}$  is subjected to temperatures as shown in Figure 3. Use a square grid length of  $0.6\text{ m}$ . Using the direct method, find the temperature at the interior nodes.



**Figure 3:** Plate with dimension and boundary temperatures

### Solution

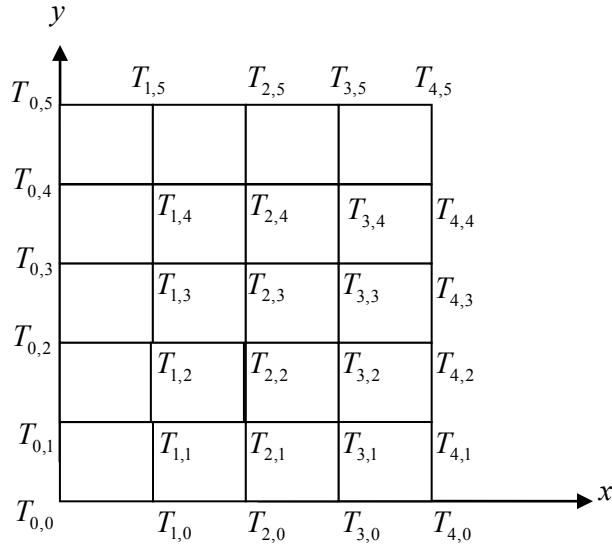
$$\Delta x = \Delta y = 0.6m$$

Re-writing Equations (5) and (6) we have

$$\begin{aligned} m &= \frac{L}{\Delta x} \\ &= \frac{2.4}{0.6} \\ &= 4 \end{aligned}$$

$$\begin{aligned} n &= \frac{W}{\Delta y} \\ &= \frac{3}{0.6} \\ &= 5 \end{aligned}$$

The nodes are shown in Figure 4.

**Figure 4:** Plate with nodes

All the nodes on the left and right boundary have an  $i$  value of zero and  $m$ , respectively. While all the nodes on the top and bottom boundary have a  $j$  value of zero and  $n$ , respectively.

From the boundary conditions

$$\left. \begin{array}{l} T_{0,j} = 75, j = 1, 2, 3, 4 \\ T_{4,j} = 100, j = 1, 2, 3, 4 \\ T_{i,0} = 50, i = 1, 2, 3 \\ T_{i,5} = 300, i = 1, 2, 3 \end{array} \right\} \quad (\text{E1.1})$$

The corner nodal temperature of  $T_{0,5}, T_{4,5}, T_{4,0}$  and  $T_{0,0}$  are not needed. Now to get the temperature at the interior nodes we have to write Equation (10) for all the combinations of  $i$  and  $j$ ,  $i = 1, \dots, m-1; j = 1, \dots, n-1$ .

$i=1$  and  $j=1$

$$\begin{aligned} T_{2,1} + T_{0,1} + T_{1,2} + T_{1,0} - 4T_{1,1} &= 0 \\ T_{2,1} + 75 + T_{1,2} + 50 - 4T_{1,1} &= 0 \\ -4T_{1,1} + T_{1,2} + T_{2,1} &= -125 \end{aligned} \quad (\text{E1.2})$$

$i=1$  and  $j=2$

$$\begin{aligned} T_{2,2} + T_{0,2} + T_{1,3} + T_{1,1} - 4T_{1,2} &= 0 \\ T_{2,2} + 75 + T_{1,3} + T_{1,1} - 4T_{1,2} &= 0 \\ T_{1,1} - 4T_{1,2} + T_{1,3} + T_{2,2} &= -75 \end{aligned} \quad (\text{E1.3})$$

$i=1$  and  $j=3$

$$\begin{aligned} T_{2,3} + T_{0,3} + T_{1,4} + T_{1,2} - 4T_{1,3} &= 0 \\ T_{2,3} + 75 + T_{1,4} + T_{1,2} - 4T_{1,3} &= 0 \end{aligned}$$

$$T_{1,2} - 4T_{1,3} + T_{1,4} + T_{2,3} = -75 \quad (\text{E1.4})$$

$i=1$  and  $j=4$

$$\begin{aligned} T_{2,4} + T_{0,4} + T_{1,5} + T_{1,3} - 4T_{1,4} &= 0 \\ T_{2,4} + 75 + 300 + T_{1,3} - 4T_{1,4} &= 0 \\ T_{1,3} - 4T_{1,4} + T_{2,4} &= -375 \end{aligned} \quad (\text{E1.5})$$

$i=2$  and  $j=1$

$$\begin{aligned} T_{3,1} + T_{1,1} + T_{2,2} + T_{2,0} - 4T_{2,1} &= 0 \\ T_{3,1} + T_{1,1} + T_{2,2} + 50 - 4T_{2,1} &= 0 \\ T_{1,1} - 4T_{2,1} + T_{2,2} + T_{3,1} &= -50 \end{aligned} \quad (\text{E1.6})$$

$i=2$  and  $j=2$

$$\begin{aligned} T_{3,2} + T_{1,2} + T_{2,3} + T_{2,1} - 4T_{2,2} &= 0 \\ T_{1,2} + T_{2,1} - 4T_{2,2} + T_{2,3} + T_{3,2} &= 0 \end{aligned} \quad (\text{E1.7})$$

$i=2$  and  $j=3$

$$\begin{aligned} T_{3,3} + T_{1,3} + T_{2,4} + T_{2,2} - 4T_{2,3} &= 0 \\ T_{1,3} + T_{2,2} - 4T_{2,3} + T_{2,4} + T_{3,3} &= 0 \end{aligned} \quad (\text{E1.8})$$

$i=2$  and  $j=4$

$$\begin{aligned} T_{3,4} + T_{1,4} + T_{2,5} + T_{2,3} - 4T_{2,4} &= 0 \\ T_{3,4} + T_{1,4} + 300 + T_{2,3} - 4T_{2,4} &= 0 \\ T_{1,4} + T_{2,3} - 4T_{2,4} + T_{3,4} &= -300 \end{aligned} \quad (\text{E1.9})$$

$i=3$  and  $j=1$

$$\begin{aligned} T_{4,1} + T_{2,1} + T_{3,2} + T_{3,0} - 4T_{3,1} &= 0 \\ 100 + T_{2,1} + T_{3,2} + 50 - 4T_{3,1} &= 0 \\ T_{2,1} - 4T_{3,1} + T_{3,2} &= -150 \end{aligned} \quad (\text{E1.10})$$

$i=3$  and  $j=2$

$$\begin{aligned} T_{4,2} + T_{2,2} + T_{3,3} + T_{3,1} - 4T_{3,2} &= 0 \\ 100 + T_{2,2} + T_{3,3} + T_{3,1} - 4T_{3,2} &= 0 \\ T_{2,2} + T_{3,1} - 4T_{3,2} + T_{3,3} &= -100 \end{aligned} \quad (\text{E1.11})$$

$i=3$  and  $j=3$

$$\begin{aligned} T_{4,3} + T_{2,3} + T_{3,4} + T_{3,2} - 4T_{3,3} &= 0 \\ 100 + T_{2,3} + T_{3,4} + T_{3,2} - 4T_{3,3} &= 0 \\ T_{2,3} + T_{3,2} - 4T_{3,3} + T_{3,4} &= -100 \end{aligned} \quad (\text{E1.12})$$

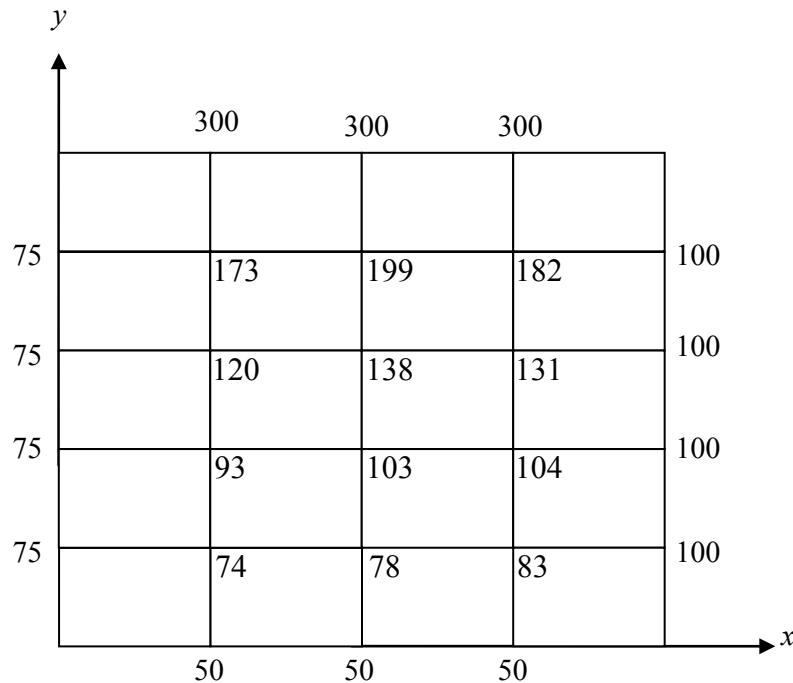
$i=3$  and  $j=4$

$$\begin{aligned} T_{4,4} + T_{2,4} + T_{3,5} + T_{3,3} - 4T_{3,4} &= 0 \\ 100 + T_{2,4} + 300 + T_{3,3} - 4T_{3,4} &= 0 \end{aligned}$$

$$T_{2,4} + T_{3,3} - 4T_{3,4} = -400 \quad (\text{E1.13})$$

Equations (E1.2) to (E1.13) represent a set of twelve simultaneous linear equations and solving them gives the temperature at the twelve interior nodes. The solution is

$$\begin{bmatrix} T_{1,1} \\ T_{1,2} \\ T_{1,3} \\ T_{1,4} \\ T_{2,1} \\ T_{2,2} \\ T_{2,3} \\ T_{2,4} \\ T_{3,1} \\ T_{3,2} \\ T_{3,3} \\ T_{3,4} \end{bmatrix} = \begin{bmatrix} 73.8924 \\ 93.0252 \\ 119.907 \\ 173.355 \\ 77.5443 \\ 103.302 \\ 138.248 \\ 198.512 \\ 82.9833 \\ 104.389 \\ 131.271 \\ 182.446 \end{bmatrix} {}^{\circ}\text{C}$$



**Figure 5:** Temperatures at the interior nodes of the plate

### Gauss-Seidel Method

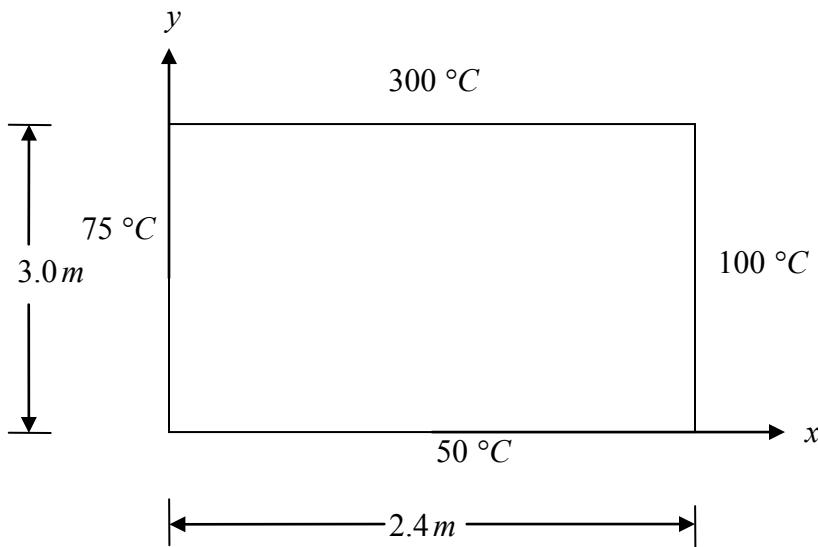
To take advantage of the sparseness of the coefficient matrix as seen in Example 1, the Gauss-Seidel method may provide a more efficient way of finding the solution. In this case, Equation (10) is written for all interior nodes as

$$T_{i,j} = \frac{T_{i+1,j} + T_{i-1,j} + T_{i,j+1} + T_{i,j-1}}{4}, i = 1, 2, 3, 4; j = 1, 2, 3, 4, 5 \quad (11)$$

Now Equation (11) is solved iteratively for all interior nodes until all the temperatures at the interior nodes are within a pre-specified tolerance.

### Example 2

A plate  $2.4\text{ m} \times 3.0\text{ m}$  is subjected to the temperatures as shown in Fig. 6. Use a square grid length of  $0.6\text{ m}$ . Using the Gauss-Seidel method, find the temperature at the interior nodes. Conduct two iterations at all interior nodes. Find the maximum absolute relative error at the end of the second iteration. Assume the initial temperature at all interior nodes to be  $0^\circ\text{C}$ .



**Figure 6:** A rectangular plate with the dimensions and boundary temperatures

### Solution

$$\Delta x = \Delta y = 0.6\text{ m}$$

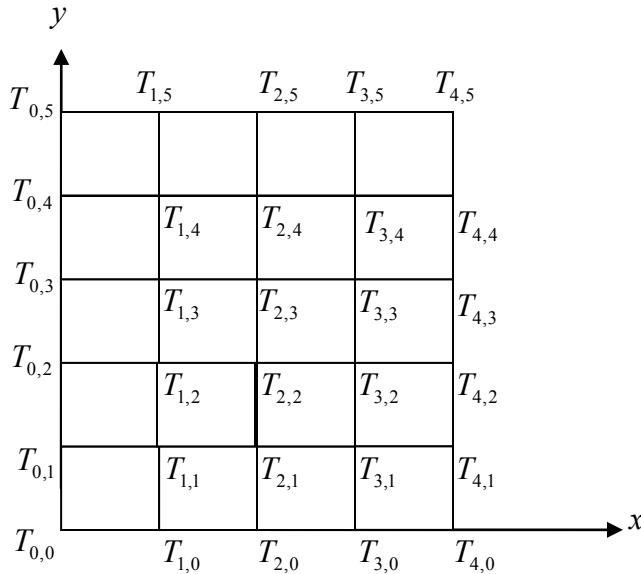
Re-writing Equations (5) and (6) we have

$$\begin{aligned} m &= \frac{L}{\Delta x} \\ &= \frac{2.4}{0.6} \\ &= 4 \end{aligned}$$

$$n = \frac{W}{\Delta y}$$

$$\begin{aligned}
 &= \frac{3}{0.6} \\
 &= 5
 \end{aligned}$$

The interior nodes are shown in Figure 7.



**Figure 7:** Plate with nodes

All the nodes on the left and right boundary have an  $i$  value of zero and  $m$ , respectively.  
All of the nodes on the top or bottom boundary have a  $j$  value of either zero or  $n$ , respectively.

From the boundary conditions

$$\left. \begin{array}{l} T_{0,j} = 75, j = 1, 2, 3, 4 \\ T_{4,j} = 100, j = 1, 2, 3, 4 \\ T_{i,0} = 50, i = 1, 2, 3 \\ T_{i,5} = 300, i = 1, 2, 3 \end{array} \right\} \quad (\text{E2.1})$$

The corner nodal temperature of  $T_{0,5}$ ,  $T_{4,5}$ ,  $T_{4,0}$  and  $T_{0,0}$  are not needed. Now to get the temperature at the interior nodes we have to write Equation (11) for all of the combinations of  $i$  and  $j$ ,  $i = 1, \dots, m-1; j = 1, \dots, n-1$ .

### Iteration 1

For iteration 1, we start with all of the interior nodes having a temperature of  $0^\circ C$ .

$i=1$  and  $j=1$

$$\begin{aligned}
 T_{1,1} &= \frac{T_{2,1} + T_{0,1} + T_{1,2} + T_{1,0}}{4} \\
 &= \frac{0 + 75 + 0 + 50}{4}
 \end{aligned}$$

$$= 31.2500^{\circ}C$$

i=1 and j=2

$$\begin{aligned} T_{1,2} &= \frac{T_{2,2} + T_{0,2} + T_{1,3} + T_{1,1}}{4} \\ &= \frac{0 + 75 + 0 + 31.2500}{4} \\ &= 26.5625^{\circ}C \end{aligned}$$

i=1 and j=3

$$\begin{aligned} T_{1,3} &= \frac{T_{2,3} + T_{0,3} + T_{1,4} + T_{1,2}}{4} \\ &= \frac{0 + 75 + 0 + 26.5625}{4} \\ &= 25.3906^{\circ}C \end{aligned}$$

i=1 and j=4

$$\begin{aligned} T_{1,4} &= \frac{T_{2,4} + T_{0,4} + T_{1,5} + T_{1,3}}{4} \\ &= \frac{0 + 75 + 300 + 25.3906}{4} \\ &= 100.098^{\circ}C \end{aligned}$$

i=2 and j=1

$$\begin{aligned} T_{2,1} &= \frac{T_{3,1} + T_{1,1} + T_{2,2} + T_{2,0}}{4} \\ &= \frac{0 + 31.2500 + 0 + 50}{4} \\ &= 20.3125^{\circ}C \end{aligned}$$

i=2 and j=2

$$\begin{aligned} T_{2,2} &= \frac{T_{3,2} + T_{1,2} + T_{2,3} + T_{2,1}}{4} \\ &= \frac{0 + 26.5625 + 0 + 20.3125}{4} \\ &= 11.7188^{\circ}C \end{aligned}$$

i=2 and j=3

$$\begin{aligned} T_{2,3} &= \frac{T_{3,3} + T_{1,3} + T_{2,4} + T_{2,2}}{4} \\ &= \frac{0 + 25.3906 + 0 + 11.7188}{4} \\ &= 9.27735^{\circ}C \end{aligned}$$

$i=2$  and  $j=4$ 

$$\begin{aligned} T_{2,4} &= \frac{T_{3,4} + T_{1,4} + T_{2,5} + T_{2,3}}{4} \\ &= \frac{0 + 100.098 + 300 + 9.27735}{4} \\ &= 102.344^\circ C \end{aligned}$$

 $i=3$  and  $j=1$ 

$$\begin{aligned} T_{3,1} &= \frac{T_{4,1} + T_{2,1} + T_{3,2} + T_{3,0}}{4} \\ &= \frac{100 + 20.3125 + 0 + 50}{4} \\ &= 42.5781^\circ C \end{aligned}$$

 $i=3$  and  $j=2$ 

$$\begin{aligned} T_{3,2} &= \frac{T_{4,2} + T_{2,2} + T_{3,3} + T_{3,1}}{4} \\ &= \frac{100 + 11.7188 + 0 + 42.5781}{4} \\ &= 38.5742^\circ C \end{aligned}$$

 $i=3$  and  $j=3$ 

$$\begin{aligned} T_{3,3} &= \frac{T_{4,3} + T_{2,3} + T_{3,4} + T_{3,2}}{4} \\ &= \frac{100 + 9.27735 + 0 + 38.5742}{4} \\ &= 36.9629^\circ C \end{aligned}$$

 $i=3$  and  $j=4$ 

$$\begin{aligned} T_{3,4} &= \frac{T_{4,4} + T_{2,4} + T_{3,5} + T_{3,3}}{4} \\ &= \frac{100 + 102.344 + 300 + 36.9629}{4} \\ &= 134.827^\circ C \end{aligned}$$

### Iteration 2

For iteration 2, we use the temperatures from iteration 1.

 $i=1$  and  $j=1$ 

$$\begin{aligned} T_{1,1} &= \frac{T_{2,1} + T_{0,1} + T_{1,2} + T_{1,0}}{4} \\ &= \frac{20.3125 + 75 + 26.5625 + 50}{4} \\ &= 42.9688^\circ C \end{aligned}$$

$$\begin{aligned} |\mathcal{E}_a|_{1,1} &= \left| \frac{T_{1,1}^{\text{present}} - T_{1,1}^{\text{previous}}}{T_{1,1}^{\text{present}}} \right| \times 100 \\ &= \left| \frac{42.9688 - 31.2500}{42.9688} \right| \times 100 \\ &= 27.27\% \end{aligned}$$

 $i=1$  and  $j=2$ 

$$\begin{aligned} T_{1,2} &= \frac{T_{2,2} + T_{0,2} + T_{1,3} + T_{1,1}}{4} \\ &= \frac{11.7188 + 75 + 25.3906 + 42.9688}{4} \\ &= 38.7696^\circ C \end{aligned}$$

$$\begin{aligned} |\mathcal{E}_a|_{1,2} &= \left| \frac{T_{1,2}^{\text{present}} - T_{1,2}^{\text{previous}}}{T_{1,2}^{\text{present}}} \right| \times 100 \\ &= \left| \frac{38.7696 - 26.5625}{38.7696} \right| \times 100 \\ &= 31.49\% \end{aligned}$$

 $i=1$  and  $j=3$ 

$$\begin{aligned} T_{1,3} &= \frac{T_{2,3} + T_{0,3} + T_{1,4} + T_{1,2}}{4} \\ &= \frac{9.27735 + 75 + 100.098 + 38.7696}{4} \\ &= 55.7862^\circ C \end{aligned}$$

$$\begin{aligned} |\mathcal{E}_a|_{1,3} &= \left| \frac{T_{1,3}^{\text{present}} - T_{1,3}^{\text{previous}}}{T_{1,3}^{\text{present}}} \right| \times 100 \\ &= \left| \frac{55.7862 - 25.3906}{55.7862} \right| \times 100 \\ &= 54.49\% \end{aligned}$$

 $i=1$  and  $j=4$ 

$$\begin{aligned} T_{1,4} &= \frac{T_{2,4} + T_{0,4} + T_{1,5} + T_{1,3}}{4} \\ &= \frac{102.344 + 75 + 300 + 55.7862}{4} \\ &= 133.283^\circ C \end{aligned}$$

$$\begin{aligned} |\mathcal{E}_a|_{1,4} &= \left| \frac{T_{1,4}^{\text{present}} - T_{1,4}^{\text{previous}}}{T_{1,4}^{\text{present}}} \right| \times 100 \\ &= \left| \frac{133.283 - 100.098}{133.283} \right| \times 100 \\ &= 24.90\% \end{aligned}$$

i=2 and j=1

$$\begin{aligned} T_{2,1} &= \frac{T_{3,1} + T_{1,1} + T_{2,2} + T_{2,0}}{4} \\ &= \frac{42.5781 + 42.9688 + 11.7188 + 50}{4} \\ &= 36.8164^\circ C \end{aligned}$$

$$\begin{aligned} |\mathcal{E}_a|_{2,1} &= \left| \frac{T_{2,1}^{\text{present}} - T_{2,1}^{\text{previous}}}{T_{2,1}^{\text{present}}} \right| \times 100 \\ &= \left| \frac{36.8164 - 20.3125}{36.8164} \right| \times 100 \\ &= 44.83\% \end{aligned}$$

i=2 and j=2

$$\begin{aligned} T_{2,2} &= \frac{T_{3,2} + T_{1,2} + T_{2,3} + T_{2,1}}{4} \\ &= \frac{38.5742 + 38.7696 + 9.27735 + 36.8164}{4} \\ &= 30.8594^\circ C \end{aligned}$$

$$\begin{aligned} |\mathcal{E}_a|_{2,2} &= \left| \frac{T_{2,2}^{\text{present}} - T_{2,2}^{\text{previous}}}{T_{2,2}^{\text{present}}} \right| \times 100 \\ &= \left| \frac{30.8594 - 11.7188}{30.8594} \right| \times 100 \\ &= 62.03\% \end{aligned}$$

i=2 and j=3

$$\begin{aligned} T_{2,3} &= \frac{T_{3,3} + T_{1,3} + T_{2,4} + T_{2,2}}{4} \\ &= \frac{36.9629 + 55.7862 + 102.344 + 30.8594}{4} \\ &= 56.4881^\circ C \end{aligned}$$

$$\begin{aligned} |\varepsilon_a|_{2,3} &= \left| \frac{T_{2,3}^{\text{present}} - T_{2,3}^{\text{previous}}}{T_{2,3}^{\text{present}}} \right| \times 100 \\ &= \left| \frac{56.4881 - 9.27735}{56.4881} \right| \times 100 \\ &= 83.58\% \end{aligned}$$

i=2 and j=4

$$\begin{aligned} T_{2,4} &= \frac{T_{3,4} + T_{1,4} + T_{2,5} + T_{2,3}}{4} \\ &= \frac{134.827 + 133.283 + 300 + 56.4881}{4} \\ &= 156.150^\circ C \\ |\varepsilon_a|_{2,4} &= \left| \frac{T_{2,4}^{\text{present}} - T_{2,4}^{\text{previous}}}{T_{2,4}^{\text{present}}} \right| \times 100 \\ &= \left| \frac{156.150 - 102.344}{156.150} \right| \times 100 \\ &= 34.46\% \end{aligned}$$

i=3 and j=1

$$\begin{aligned} T_{3,1} &= \frac{T_{4,1} + T_{2,1} + T_{3,2} + T_{3,0}}{4} \\ &= \frac{100 + 36.8164 + 38.5742 + 50}{4} \\ &= 56.3477^\circ C \\ |\varepsilon_a|_{3,1} &= \left| \frac{T_{3,1}^{\text{present}} - T_{3,1}^{\text{previous}}}{T_{3,1}^{\text{present}}} \right| \times 100 \\ &= \left| \frac{56.3477 - 42.5781}{56.3477} \right| \times 100 \\ &= 24.44\% \end{aligned}$$

i=3 and j=2

$$\begin{aligned} T_{3,2} &= \frac{T_{4,2} + T_{2,2} + T_{3,3} + T_{3,1}}{4} \\ &= \frac{100 + 30.8594 + 36.9629 + 56.3477}{4} \\ &= 56.0425^\circ C \end{aligned}$$

$$\begin{aligned} |\mathcal{E}_a|_{3,2} &= \left| \frac{T_{3,2}^{\text{present}} - T_{3,2}^{\text{previous}}}{T_{3,2}^{\text{present}}} \right| \times 100 \\ &= \left| \frac{56.0425 - 38.5742}{56.0425} \right| \times 100 \\ &= 31.70\% \end{aligned}$$

i=3 and j=3

$$\begin{aligned} T_{3,3} &= \frac{T_{4,3} + T_{2,3} + T_{3,4} + T_{3,2}}{4} \\ &= \frac{100 + 56.4881 + 134.827 + 56.0425}{4} \\ &= 86.8394^\circ C \end{aligned}$$

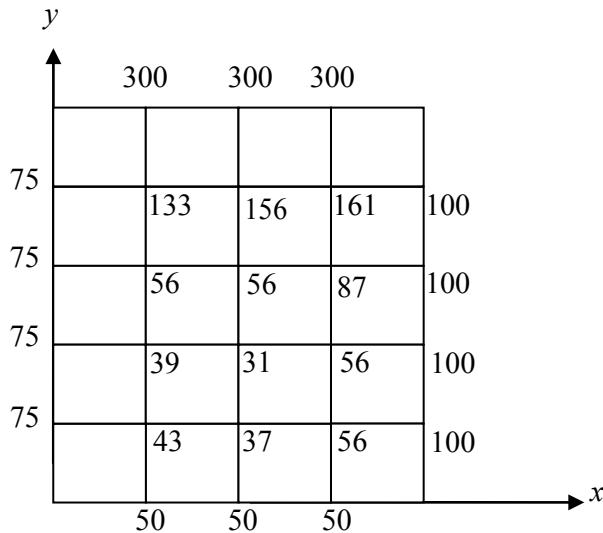
$$\begin{aligned} |\mathcal{E}_a|_{3,3} &= \left| \frac{T_{3,3}^{\text{present}} - T_{3,3}^{\text{previous}}}{T_{3,3}^{\text{present}}} \right| \times 100 \\ &= \left| \frac{86.8394 - 36.9629}{86.8394} \right| \times 100 \\ &= 57.44\% \end{aligned}$$

i=3 and j=4

$$\begin{aligned} T_{3,4} &= \frac{T_{4,4} + T_{2,4} + T_{3,5} + T_{3,3}}{4} \\ &= \frac{100 + 156.150 + 300 + 86.8394}{4} \\ &= 160.747^\circ C \end{aligned}$$

$$\begin{aligned} |\mathcal{E}_a|_{3,4} &= \left| \frac{T_{3,4}^{\text{present}} - T_{3,4}^{\text{previous}}}{T_{3,4}^{\text{present}}} \right| \times 100 \\ &= \left| \frac{160.747 - 134.827}{160.747} \right| \times 100 \\ &= 16.12\% \end{aligned}$$

The maximum absolute relative error at the end of iteration 2 is 83%.



**Figure 8:** Temperature distribution after two iterations

It took ten iterations to get all of the temperature values within 1% error. The table below lists the temperature values at the interior nodes at the end of each iteration:

Node	Number of Iterations				
	1	2	3	4	5
$T_{1,1}$	31.2500	42.9688	50.1465	56.1966	61.6376
$T_{1,2}$	26.5625	38.7695	52.9480	65.9264	76.5753
$T_{1,3}$	25.3906	55.7861	79.4296	96.8614	106.8163
$T_{1,4}$	100.0977	133.2825	152.6447	162.1695	167.1287
$T_{2,1}$	20.3125	36.8164	46.8384	55.6240	63.6980
$T_{2,2}$	11.7188	30.8594	53.0792	72.8024	85.3707
$T_{2,3}$	9.2773	56.4880	93.8744	113.5205	124.2410
$T_{2,4}$	102.3438	156.1493	176.8166	186.6986	191.8910
$T_{3,1}$	42.5781	56.3477	63.2202	70.3522	75.3468
$T_{3,2}$	38.5742	56.0425	75.7847	87.6890	94.6990
$T_{3,3}$	36.9629	86.8393	107.6015	118.0785	123.7836
$T_{3,4}$	134.8267	160.7471	171.1045	176.1943	178.9186

Node	Number of Iterations				
	6	7	8	9	10
$T_{1,1}$	66.3183	69.4088	71.2832	72.3848	73.0239
$T_{1,2}$	83.3763	87.4348	89.8017	91.1701	91.9585
$T_{1,3}$	112.4365	115.6295	117.4532	118.4980	119.0976
$T_{1,4}$	169.8319	171.3450	172.2037	172.6943	172.9755
$T_{2,1}$	69.2590	72.6980	74.7374	75.9256	76.6127
$T_{2,2}$	92.8938	97.2939	99.8423	102.3119	102.1577
$T_{2,3}$	130.2512	133.6661	135.6184	136.7377	137.3802
$T_{2,4}$	194.7504	196.3616	197.2791	197.8043	198.1055
$T_{3,1}$	78.4895	80.3724	81.4754	82.1148	82.4837
$T_{3,2}$	98.7917	101.1642	102.5335	103.3221	103.7757
$T_{3,3}$	126.9904	128.8164	129.8616	130.4612	130.8056
$T_{3,4}$	180.4352	181.2945	181.7852	182.0664	182.2278

### Successive Over Relaxation Method

The coefficient matrix for solving for temperatures given in Example 1 is diagonally dominant. Hence the Gauss-Siedel method is guaranteed to converge. To accelerate convergence to the solution, over relaxation is used. In this case

$$T_{i,j}^{relaxed} = \lambda T_{i,j}^{new} + (1-\lambda)T_{i,j}^{old} \quad (12)$$

where

$T_{i,j}^{new}$  = value of temperature from current iteration,

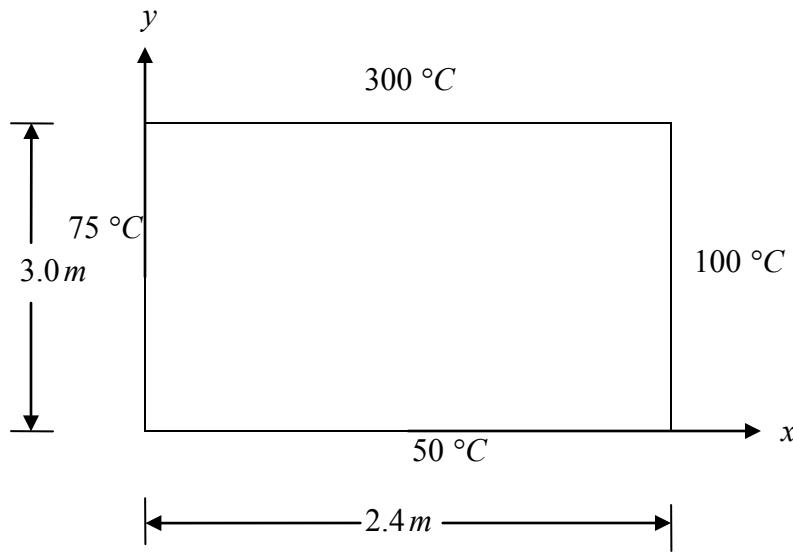
$T_{i,j}^{old}$  = value of temperature from previous iteration,

$\lambda$  = weighting factor,  $1 < \lambda < 2$ .

Again, these iterations are continued till the pre-specified tolerance is met for all nodal temperatures. This method is also called the Lieberman method.

### Example 3

A plate  $2.4\text{ m} \times 3.0\text{ m}$  is subjected to the temperatures as shown in Fig. 6. Use a square grid length of  $0.6\text{ m}$ . Use the Gauss-Seidel with successive over relaxation method with a weighting factor of 1.4 to find the temperature at the interior nodes. Conduct two iterations at all interior nodes. Find the maximum absolute relative error at the end of the second iteration. Assume the initial temperature at all interior nodes to be  $0^\circ\text{C}$ .



**Figure 9:** A rectangular plate with the dimensions and boundary temperatures

### Solution

$$\Delta x = \Delta y = 0.6\text{m}$$

Re-writing Equations (5) and (6) we have

$$m = \frac{L}{\Delta x}$$

$$= \frac{2.4}{0.6}$$

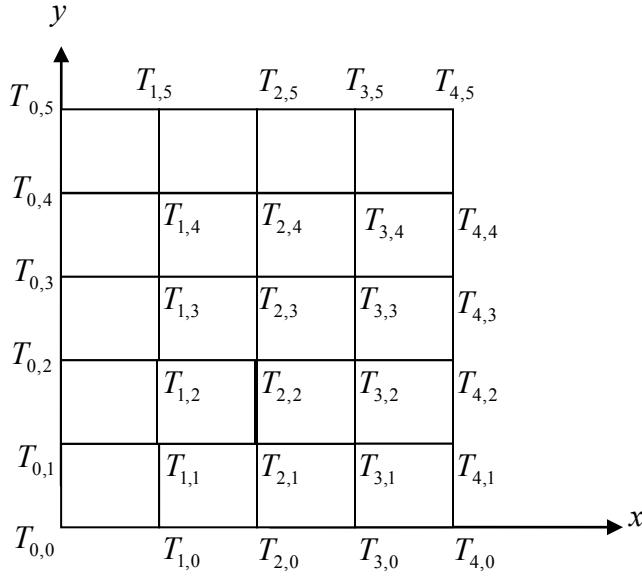
$$= 4$$

$$n = \frac{W}{\Delta y}$$

$$= \frac{3}{0.6}$$

$$= 5$$

The interior nodes are shown in the Figure 10.

**Figure 10:** Plate with nodes

All of the nodes on the left and right boundary have an  $i$  value of zero and  $m$ , respectively. All of the nodes on the top or bottom boundary have a  $j$  value of either zero or  $n$ , respectively.

From the boundary conditions

$$\left. \begin{array}{l} T_{0,j} = 75, j = 1, 2, 3, 4 \\ T_{4,j} = 100, j = 1, 2, 3, 4 \\ T_{i,0} = 50, i = 1, 2, 3 \\ T_{i,5} = 300, i = 1, 2, 3 \end{array} \right\} \quad (\text{E3.1})$$

The corner nodal temperature of  $T_{0,5}$ ,  $T_{4,5}$ ,  $T_{4,0}$  and  $T_{0,0}$  are not needed. Now to get the temperature at the interior nodes, we have to write Equation (11) for all of the combinations of  $i$  and  $j$ ,  $i = 1$  to  $m - 1$ ,  $j = 1$  to  $n - 1$ . After getting the temperature from Equation (11), we have to use Equation (12) to apply the over relaxation method.

### Iteration 1

For iteration 1, we start with all of the interior nodes having a temperature of  $0^\circ\text{C}$ .

#### $i=1$ and $j=1$

$$\begin{aligned} T_{1,1} &= \frac{T_{2,1} + T_{0,1} + T_{1,2} + T_{1,0}}{4} \\ &= \frac{0 + 75 + 0 + 50}{4} \\ &= 31.2500^\circ\text{C} \end{aligned}$$

$$T_{1,1}^{\text{relaxed}} = \lambda T_{1,1}^{\text{new}} + (1 - \lambda) T_{1,1}^{\text{old}}$$

$$\begin{aligned}
 &= 1.4(31.2500) + (1-1.4)0 \\
 &= 43.7500^\circ C
 \end{aligned}$$

i=1 and j=2

$$\begin{aligned}
 T_{1,2} &= \frac{T_{2,2} + T_{0,2} + T_{1,3} + T_{1,1}}{4} \\
 &= \frac{0 + 75 + 0 + 43.75}{4} \\
 &= 29.6875^\circ C
 \end{aligned}$$

$$\begin{aligned}
 T_{1,2}^{relaxed} &= \lambda T_{1,2}^{new} + (1-\lambda)T_{1,2}^{old} \\
 &= 1.4(29.6875) + (1-1.4)0 \\
 &= 41.5625^\circ C
 \end{aligned}$$

i=1 and j=3

$$\begin{aligned}
 T_{1,3} &= \frac{T_{2,3} + T_{0,3} + T_{1,4} + T_{1,2}}{4} \\
 &= \frac{0 + 75 + 0 + 41.5625}{4} \\
 &= 29.1406^\circ C
 \end{aligned}$$

$$\begin{aligned}
 T_{1,3}^{relaxed} &= \lambda T_{1,3}^{new} + (1-\lambda)T_{1,3}^{old} \\
 &= 1.4(29.1406) + (1-1.4)0 \\
 &= 40.7969^\circ C
 \end{aligned}$$

i=1 and j=4

$$\begin{aligned}
 T_{1,4} &= \frac{T_{2,4} + T_{0,4} + T_{1,5} + T_{1,3}}{4} \\
 &= \frac{0 + 75 + 300 + 40.7969}{4} \\
 &= 103.949^\circ C
 \end{aligned}$$

$$\begin{aligned}
 T_{1,4}^{relaxed} &= \lambda T_{1,4}^{new} + (1-\lambda)T_{1,4}^{old} \\
 &= 1.4(103.949) + (1-1.4)0 \\
 &= 145.529^\circ C
 \end{aligned}$$

i=2 and j=1

$$\begin{aligned}
 T_{2,1} &= \frac{T_{3,1} + T_{1,1} + T_{2,2} + T_{2,0}}{4} \\
 &= \frac{0 + 43.75 + 0 + 50}{4} \\
 &= 23.4375^\circ C
 \end{aligned}$$

$$T_{2,1}^{relaxed} = \lambda T_{2,1}^{new} + (1-\lambda)T_{2,1}^{old}$$

$$\begin{aligned}
 &= 1.4(23.4375) + (1-1.4)0 \\
 &= 32.8215^\circ C
 \end{aligned}$$

 $i=2$  and  $j=2$ 

$$\begin{aligned}
 T_{2,2} &= \frac{T_{3,2} + T_{1,2} + T_{2,3} + T_{2,1}}{4} \\
 &= \frac{0 + 41.5625 + 0 + 32.8125}{4} \\
 &= 18.5938^\circ C
 \end{aligned}$$

$$\begin{aligned}
 T_{2,2}^{relaxed} &= \lambda T_{2,2}^{new} + (1-\lambda) T_{2,2}^{old} \\
 &= 1.4(18.5938) + (1-1.4)0 \\
 &= 26.0313^\circ C
 \end{aligned}$$

 $i=2$  and  $j=3$ 

$$\begin{aligned}
 T_{2,3} &= \frac{T_{3,3} + T_{1,3} + T_{2,4} + T_{2,2}}{4} \\
 &= \frac{0 + 40.7969 + 0 + 26.0313}{4} \\
 &= 16.7071^\circ C
 \end{aligned}$$

$$\begin{aligned}
 T_{2,3}^{relaxed} &= \lambda T_{2,3}^{new} + (1-\lambda) T_{2,3}^{old} \\
 &= 1.4(16.7071) + (1-1.4)0 \\
 &= 23.3899^\circ C
 \end{aligned}$$

 $i=2$  and  $j=4$ 

$$\begin{aligned}
 T_{2,4} &= \frac{T_{3,4} + T_{1,4} + T_{2,5} + T_{2,3}}{4} \\
 &= \frac{0 + 145.529 + 300 + 23.3899}{4} \\
 &= 117.230^\circ C
 \end{aligned}$$

$$\begin{aligned}
 T_{2,4}^{relaxed} &= \lambda T_{2,4}^{new} + (1-\lambda) T_{2,4}^{old} \\
 &= 1.4(117.230) + (1-1.4)0 \\
 &= 164.122^\circ C
 \end{aligned}$$

 $i=3$  and  $j=1$ 

$$\begin{aligned}
 T_{3,1} &= \frac{T_{4,1} + T_{2,1} + T_{3,2} + T_{3,0}}{4} \\
 &= \frac{100 + 32.8125 + 0 + 50}{4} \\
 &= 45.7031^\circ C
 \end{aligned}$$

$$T_{3,1}^{relaxed} = \lambda T_{3,1}^{new} + (1-\lambda) T_{3,1}^{old}$$

$$\begin{aligned}
 &= 1.4(45.7031) + (1-1.4)0 \\
 &= 63.9844^\circ C
 \end{aligned}$$

i=3 and j=2

$$\begin{aligned}
 T_{3,2} &= \frac{T_{4,2} + T_{2,2} + T_{3,3} + T_{3,1}}{4} \\
 &= \frac{100 + 26.0313 + 0 + 63.9844}{4} \\
 &= 47.5039^\circ C \\
 T_{3,2}^{relaxed} &= \lambda T_{3,2}^{new} + (1-\lambda)T_{3,2}^{old} \\
 &= 1.4(47.5039) + (1-1.4)0 \\
 &= 66.5055^\circ C
 \end{aligned}$$

i=3 and j=3

$$\begin{aligned}
 T_{3,3} &= \frac{T_{4,3} + T_{2,3} + T_{3,4} + T_{3,2}}{4} \\
 &= \frac{100 + 23.3899 + 0 + 66.5055}{4} \\
 &= 47.4739^\circ C \\
 T_{3,3}^{relaxed} &= \lambda T_{3,3}^{new} + (1-\lambda)T_{3,3}^{old} \\
 &= 1.4(47.4739) + (1-1.4)0 \\
 &= 66.4634^\circ C
 \end{aligned}$$

i=3 and j=4

$$\begin{aligned}
 T_{3,4} &= \frac{T_{4,4} + T_{2,4} + T_{3,5} + T_{3,3}}{4} \\
 &= \frac{100 + 164.122 + 300 + 66.4634}{4} \\
 &= 157.646^\circ C \\
 T_{3,4}^{relaxed} &= \lambda T_{3,4}^{new} + (1-\lambda)T_{3,4}^{old} \\
 &= 1.4(157.646) + (1-1.4)0 \\
 &= 220.704^\circ C
 \end{aligned}$$

### Iteration 2

For iteration 2, we take the temperatures from iteration 1.

i=1 and j=1

$$\begin{aligned}
 T_{1,1} &= \frac{T_{2,1} + T_{0,1} + T_{1,2} + T_{1,0}}{4} \\
 &= \frac{32.8125 + 75 + 41.5625 + 50}{4} \\
 &= 49.8438^\circ C
 \end{aligned}$$

$$\begin{aligned}
T_{1,1}^{relaxed} &= \lambda T_{1,1}^{new} + (1-\lambda) T_{1,1}^{old} \\
&= 1.4(49.8438) + (1-1.4)43.75 \\
&= 52.2813^\circ C \\
|\mathcal{E}_a|_{1,1} &= \left| \frac{T_{1,1}^{present} - T_{1,1}^{previous}}{T_{1,1}^{present}} \right| \times 100 \\
&= \left| \frac{52.2813 - 43.7500}{52.2813} \right| \times 100 \\
&= 16.32\%
\end{aligned}$$

 $i=1$  and  $j=2$ 

$$\begin{aligned}
T_{1,2} &= \frac{T_{2,2} + T_{0,2} + T_{1,3} + T_{1,1}}{4} \\
&= \frac{26.0313 + 75 + 40.7969 + 52.2813}{4} \\
&= 48.5274^\circ C
\end{aligned}$$

$$\begin{aligned}
T_{1,2}^{relaxed} &= \lambda T_{1,2}^{new} + (1-\lambda) T_{1,2}^{old} \\
&= 1.4(48.5274) + (1-1.4)41.5625 \\
&= 51.3133^\circ C \\
|\mathcal{E}_a|_{1,2} &= \left| \frac{T_{1,2}^{present} - T_{1,2}^{previous}}{T_{1,2}^{present}} \right| \times 100 \\
&= \left| \frac{51.3133 - 41.5625}{51.3133} \right| \times 100 \\
&= 19.00\%
\end{aligned}$$

 $i=1$  and  $j=3$ 

$$\begin{aligned}
T_{1,3} &= \frac{T_{2,3} + T_{0,3} + T_{1,4} + T_{1,2}}{4} \\
&= \frac{23.3899 + 75 + 145.529 + 51.3133}{4} \\
&= 73.8103^\circ C
\end{aligned}$$

$$\begin{aligned}
T_{1,3}^{relaxed} &= \lambda T_{1,3}^{new} + (1-\lambda) T_{1,3}^{old} \\
&= 1.4(73.8103) + (1-1.4)40.7969 \\
&= 87.0157^\circ C
\end{aligned}$$

$$\begin{aligned} |\varepsilon_a|_{1,3} &= \left| \frac{T_{1,3}^{\text{present}} - T_{1,3}^{\text{previous}}}{T_{1,3}^{\text{present}}} \right| \times 100 \\ &= \left| \frac{87.0157 - 40.7969}{87.0157} \right| \times 100 \\ &= 53.12\% \end{aligned}$$

 $i=1$  and  $j=4$ 

$$\begin{aligned} T_{1,4} &= \frac{T_{2,4} + T_{0,4} + T_{1,5} + T_{1,3}}{4} \\ &= \frac{164.122 + 75 + 300 + 87.0157}{4} \\ &= 156.534^\circ C \end{aligned}$$

$$\begin{aligned} T_{1,4}^{\text{relaxed}} &= \lambda T_{1,4}^{\text{new}} + (1-\lambda) T_{1,4}^{\text{old}} \\ &= 1.4(156.534) + (1-1.4)145.529 \\ &= 160.936^\circ C \end{aligned}$$

$$\begin{aligned} |\varepsilon_a|_{1,4} &= \left| \frac{T_{1,4}^{\text{present}} - T_{1,4}^{\text{previous}}}{T_{1,4}^{\text{present}}} \right| \times 100 \\ &= \left| \frac{160.936 - 145.529}{160.936} \right| \times 100 \\ &= 9.57\% \end{aligned}$$

 $i=2$  and  $j=1$ 

$$\begin{aligned} T_{2,1} &= \frac{T_{3,1} + T_{1,1} + T_{2,2} + T_{2,0}}{4} \\ &= \frac{63.9844 + 52.2813 + 26.0313 + 50.000}{4} \\ &= 48.0743^\circ C \end{aligned}$$

$$\begin{aligned} T_{2,1}^{\text{relaxed}} &= \lambda T_{2,1}^{\text{new}} + (1-\lambda) T_{2,1}^{\text{old}} \\ &= 1.4(48.0743) + (1-1.4)32.8125 \\ &= 54.1790^\circ C \end{aligned}$$

$$\begin{aligned} |\varepsilon_a|_{2,1} &= \left| \frac{T_{2,1}^{\text{present}} - T_{2,1}^{\text{previous}}}{T_{2,1}^{\text{present}}} \right| \times 100 \\ &= \left| \frac{54.1790 - 32.8125}{54.1790} \right| \times 100 \\ &= 39.44\% \end{aligned}$$

$i=2$  and  $j=2$ 

$$\begin{aligned} T_{2,2} &= \frac{T_{3,2} + T_{1,2} + T_{2,3} + T_{2,1}}{4} \\ &= \frac{66.5055 + 51.3133 + 23.3899 + 54.1790}{4} \\ &= 48.8469^\circ C \end{aligned}$$

$$\begin{aligned} T_{2,2}^{relaxed} &= \lambda T_{2,2}^{new} + (1-\lambda) T_{2,2}^{old} \\ &= 1.4(48.8469) + (1-1.4)26.0313 \\ &= 57.9732^\circ C \\ |\mathcal{E}_a|_{2,2} &= \left| \frac{T_{2,2}^{present} - T_{2,2}^{previous}}{T_{2,2}^{present}} \right| \times 100 \\ &= \left| \frac{57.9732 - 26.0313}{57.9732} \right| \times 100 \\ &= 55.10\% \end{aligned}$$

 $i=2$  and  $j=3$ 

$$\begin{aligned} T_{2,3} &= \frac{T_{3,3} + T_{1,3} + T_{2,4} + T_{2,2}}{4} \\ &= \frac{66.4634 + 87.0157 + 164.122 + 57.9732}{4} \\ &= 93.8936^\circ C \end{aligned}$$

$$\begin{aligned} T_{2,3}^{relaxed} &= \lambda T_{2,3}^{new} + (1-\lambda) T_{2,3}^{old} \\ &= 1.4(93.8936) + (1-1.4)23.3899 \\ &= 122.095^\circ C \\ |\mathcal{E}_a|_{2,3} &= \left| \frac{T_{2,3}^{present} - T_{2,3}^{previous}}{T_{2,3}^{present}} \right| \times 100 \\ &= \left| \frac{122.095 - 23.3899}{122.095} \right| \times 100 \\ &= 80.84\% \end{aligned}$$

 $i=2$  and  $j=4$ 

$$\begin{aligned} T_{2,4} &= \frac{T_{3,4} + T_{1,4} + T_{2,5} + T_{2,3}}{4} \\ &= \frac{220.704 + 160.936 + 300 + 122.095}{4} \\ &= 200.934^\circ C \\ T_{2,4}^{relaxed} &= \lambda T_{2,4}^{new} + (1-\lambda) T_{2,4}^{old} \end{aligned}$$

$$= 1.4(200.934) + (1 - 1.4)164.122$$

$$= 215.659^{\circ}\text{C}$$

$$\begin{aligned} |\varepsilon_a|_{2,4} &= \left| \frac{T_{2,4}^{\text{present}} - T_{2,4}^{\text{previous}}}{T_{2,4}^{\text{present}}} \right| \times 100 \\ &= \left| \frac{215.659 - 164.122}{215.659} \right| \times 100 \\ &= 23.90\% \end{aligned}$$

i=3 and j=1

$$\begin{aligned} T_{3,1} &= \frac{T_{4,1} + T_{2,1} + T_{3,2} + T_{3,0}}{4} \\ &= \frac{100 + 54.1790 + 66.5055 + 50}{4} \\ &= 67.6711^{\circ}\text{C} \end{aligned}$$

$$\begin{aligned} T_{3,1}^{\text{relaxed}} &= \lambda T_{3,1}^{\text{new}} + (1 - \lambda) T_{3,1}^{\text{old}} \\ &= 1.4(67.6711) + (1 - 1.4)63.9844 \\ &= 69.1458^{\circ}\text{C} \end{aligned}$$

$$\begin{aligned} |\varepsilon_a|_{3,1} &= \left| \frac{T_{3,1}^{\text{present}} - T_{3,1}^{\text{previous}}}{T_{3,1}^{\text{present}}} \right| \times 100 \\ &= \left| \frac{69.1458 - 63.9844}{69.1458} \right| \times 100 \\ &= 7.46\% \end{aligned}$$

i=3 and j=2

$$\begin{aligned} T_{3,2} &= \frac{T_{4,2} + T_{2,2} + T_{3,3} + T_{3,1}}{4} \\ &= \frac{100 + 57.9732 + 66.4634 + 69.1458}{4} \\ &= 73.3956^{\circ}\text{C} \end{aligned}$$

$$\begin{aligned} T_{3,2}^{\text{relaxed}} &= \lambda T_{3,2}^{\text{new}} + (1 - \lambda) T_{3,2}^{\text{old}} \\ &= 1.4(73.3956) + (1 - 1.4)66.5055 \\ &= 76.1516^{\circ}\text{C} \end{aligned}$$

$$\begin{aligned} |\varepsilon_a|_{3,2} &= \left| \frac{T_{3,2}^{\text{present}} - T_{3,2}^{\text{previous}}}{T_{3,2}^{\text{present}}} \right| \times 100 \\ &= \left| \frac{76.1516 - 66.5055}{76.1516} \right| \times 100 \\ &= 12.67\% \end{aligned}$$

$i=3$  and  $j=3$ 

$$\begin{aligned} T_{3,3} &= \frac{T_{4,3} + T_{2,3} + T_{3,4} + T_{3,2}}{4} \\ &= \frac{100 + 122.095 + 220.704 + 76.1516}{4} \\ &= 129.738^\circ C \end{aligned}$$

$$\begin{aligned} T_{3,3}^{relaxed} &= \lambda T_{3,3}^{new} + (1 - \lambda) T_{3,3}^{old} \\ &= 1.4(129.738) + (1 - 1.4)66.4634 \\ &= 155.048^\circ C \end{aligned}$$

$$\begin{aligned} |\mathcal{E}_a|_{3,3} &= \left| \frac{T_{3,3}^{present} - T_{3,3}^{previous}}{T_{3,3}^{present}} \right| \times 100 \\ &= \left| \frac{155.048 - 66.4634}{155.048} \right| \times 100 \\ &= 57.13\% \end{aligned}$$

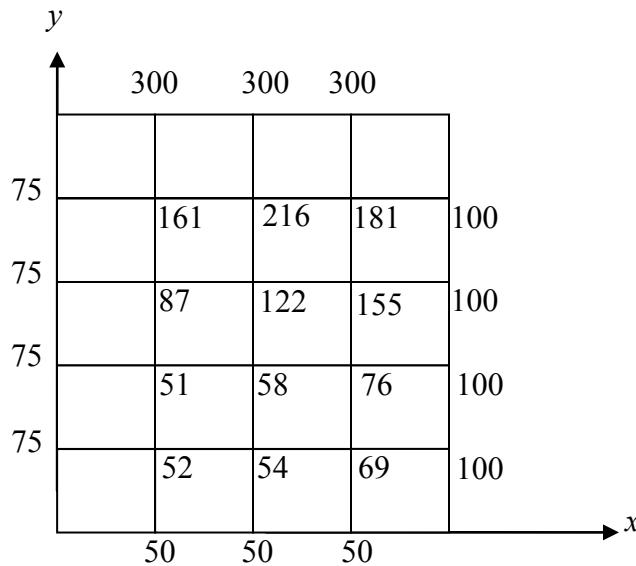
 $i=3$  and  $j=4$ 

$$\begin{aligned} T_{3,4} &= \frac{T_{4,4} + T_{2,4} + T_{3,5} + T_{3,3}}{4} \\ &= \frac{100 + 215.659 + 300 + 155.048}{4} \\ &= 192.677^\circ C \end{aligned}$$

$$\begin{aligned} T_{3,4}^{relaxed} &= \lambda T_{3,4}^{new} + (1 - \lambda) T_{3,4}^{old} \\ &= 1.4(192.677) + (1 - 1.4)220.704 \\ &= 181.466^\circ C \end{aligned}$$

$$\begin{aligned} |\mathcal{E}_a|_{3,4} &= \left| \frac{T_{3,4}^{present} - T_{3,4}^{previous}}{T_{3,4}^{present}} \right| \times 100 \\ &= \left| \frac{181.466 - 220.704}{181.466} \right| \times 100 \\ &= 21.62\% \end{aligned}$$

The maximum absolute relative error at the end of iteration 2 is 81%.



**Figure 11:** Temperature distribution after two iterations

It took nine iterations to get all of the temperature values within 1% error. The table below lists the temperature values at all nodes after each iteration.

Node	Number of Iterations				
	1	2	3	4	5
$T_{1,1}$	43.7500	52.2813	59.7598	68.3636	75.6025
$T_{1,2}$	41.5625	51.3133	77.3856	93.5293	101.8402
$T_{1,3}$	40.7969	87.0125	117.5901	130.5043	119.8434
$T_{1,4}$	145.5289	160.9353	183.5128	173.8030	173.3888
$T_{2,1}$	32.8125	54.1789	61.2360	75.6074	86.4009
$T_{2,2}$	26.0313	57.9731	94.7142	116.7560	105.9062
$T_{2,3}$	23.3898	122.0937	155.2159	140.9145	139.0181
$T_{2,4}$	164.1216	215.6582	200.8045	199.1851	198.6561
$T_{3,1}$	63.9844	69.1458	72.9273	90.9098	83.7806
$T_{3,2}$	66.5055	76.1516	117.4804	106.8690	105.2995
$T_{3,3}$	66.4634	155.0472	131.9376	133.3050	131.1769
$T_{3,4}$	220.7047	181.4650	183.8737	182.8220	182.3127

Node	Number of Iterations			
	6	7	8	9
$T_{1,1}$	79.3934	71.2937	74.2346	73.7832
$T_{1,2}$	92.3140	92.1224	93.0388	92.9758
$T_{1,3}$	119.9649	119.388	119.8366	119.9378
$T_{1,4}$	173.4118	173.0515	173.3665	173.3937
$T_{2,1}$	77.1177	76.4550	77.6097	77.5449
$T_{2,2}$	102.4498	102.4844	103.3554	103.3285
$T_{2,3}$	137.6794	137.7443	138.2932	138.3236
$T_{2,4}$	198.2290	198.2693	198.6060	198.5498
$T_{3,1}$	82.8338	82.4002	83.1150	82.9805
$T_{3,2}$	103.6414	104.0334	104.5308	104.3815
$T_{3,3}$	130.8010	131.0842	131.3876	131.2525
$T_{3,4}$	182.2354	182.3796	182.5459	182.4230

### Alternative Boundary Conditions

In Examples 1-3, the boundary conditions on the plate had a specified temperature on each edge. What if the conditions are different? For example; what if one of the edges of the plate is insulated? In this case, the boundary condition would be the derivative of the temperature (called the Neuman boundary condition). If the right edge of the plate is insulated, then the temperatures on the right edge nodes also become unknowns. The finite difference Equation (10) in this case for the right edge for the nodes  $(m, j), j = 1, 2, 3, \dots, n-1; i = 1, 2, \dots, m$

$$T_{m+1,j} + T_{m-1,j} + T_{m,j-1} + T_{m,j+1} - 4T_{m,j} = 0 \quad (13)$$

However, the node  $(m+1, j)$  is not inside the plate. The derivative boundary condition needs to be used to account for these additional unknown nodal temperatures on the right edge. This is done by approximating the derivative at the edge node  $(m, j)$  as

$$\frac{\partial T}{\partial x} \Big|_{m,j} \cong \frac{T_{m+1,j} - T_{m-1,j}}{2(\Delta x)} \quad (14)$$

giving

$$T_{m+1,j} = T_{m-1,j} + 2(\Delta x) \frac{\partial T}{\partial x} \Big|_{m,j} \quad (15)$$

substituting Equation (15) in Equation (13), gives

$$2T_{m-1,j} + 2(\Delta x) \frac{\partial T}{\partial x} \Big|_{m,j} + T_{m,j-1} + T_{m,j+1} - 4T_{m,j} = 0 \quad (16)$$

Now if the edge is insulated,

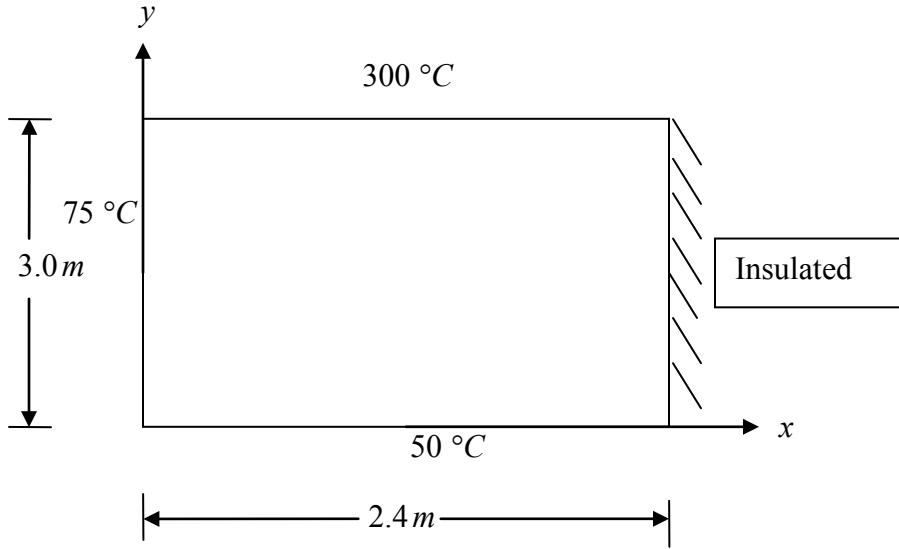
$$\frac{\partial T}{\partial x} \Big|_{m,j} = 0 \quad (17)$$

substituting Equation (17) in Equation (16), gives an equation to use at the Neuman Boundary condition

$$2T_{m-1,j} + T_{m,j-1} + T_{m,j+1} - 4T_{m,j} = 0 \quad (18)$$

### Example 4

A plate  $2.4\text{ m} \times 3.0\text{ m}$  is subjected to the temperatures and insulated boundary conditions as shown in Fig. 12. Use a square grid length of  $0.6\text{ m}$ . Assume the initial temperatures at all of the interior nodes to be  $0^\circ\text{C}$ . Find the temperatures at the interior nodes using the direct method.



**Figure 12:** Plate with the dimensions and boundary conditions

### Solution

$$\Delta x = \Delta y = 0.6 \text{ m}$$

Re-writing Equations (5) and (6) we have

$$m = \frac{L}{\Delta x}$$

$$= \frac{2.4}{0.6}$$

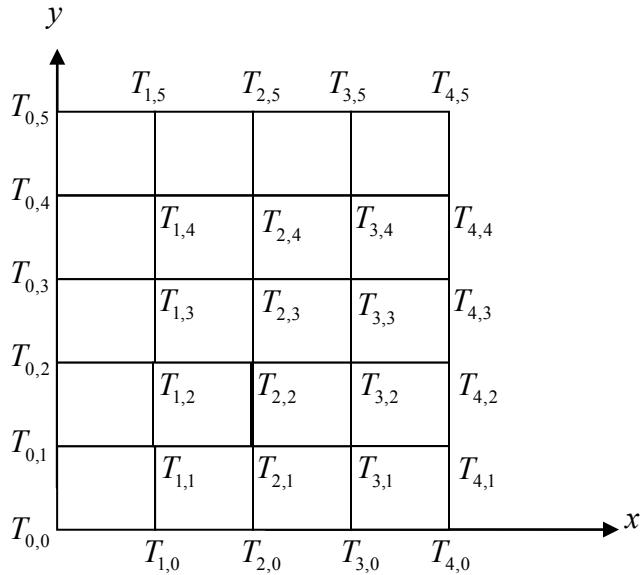
$$= 4$$

$$n = \frac{W}{\Delta y}$$

$$= \frac{3}{0.6}$$

$$= 5$$

The unknown temperature nodes are shown in Figure 13.

**Figure 13:** Plate with the nodes labeled

All of the nodes on the boundary have an  $i$  value of either zero or  $m$ . All of the nodes on the boundary have a  $j$  value of either zero or  $n$ .

From the boundary conditions

$$\left. \begin{array}{l} T_{0,j} = 75; j = 1,2,3,4 \\ T_{i,0} = 50; i = 1,2,3,4 \\ T_{i,5} = 300; i = 1,2,3,4 \\ \frac{\partial T}{\partial x} \Big|_{4,j} = 0; j = 1,2,3,4 \end{array} \right\} \quad (\text{E4.1})$$

Now in order to find the temperatures at the interior nodes, we have to write Equation (10) for all of the combinations of  $i$  and  $j$ . We express this using  $i$  from 1 to  $m-1$  and  $j$  from 1 to  $n-1$ . For the right side boundary nodes, where  $i = m = 4$ , we have to write Equation (18) for  $j = 1,2,3,4$ . This would give  $m \times n - 1$  simultaneous linear equations with  $m \times n - 1$  unknowns.

$i=1$  and  $j=1$

$$\begin{aligned} T_{2,1} + T_{0,1} + T_{1,2} + T_{1,0} - 4T_{1,1} &= 0 \\ T_{2,1} + 75 + T_{1,2} + 50 - 4T_{1,1} &= 0 \\ -4T_{1,1} + T_{1,2} + T_{2,1} &= -125 \end{aligned} \quad (\text{E4.2})$$

$i=1$  and  $j=2$

$$\begin{aligned} T_{2,2} + T_{0,2} + T_{1,3} + T_{1,1} - 4T_{1,2} &= 0 \\ T_{2,2} + 75 + T_{1,3} + T_{1,1} - 4T_{1,2} &= 0 \\ T_{1,1} - 4T_{1,2} + T_{1,3} + T_{2,2} &= -75 \end{aligned} \quad (\text{E4.3})$$

$i=1$  and  $j=3$ 

$$\begin{aligned} T_{2,3} + T_{0,3} + T_{1,4} + T_{1,2} - 4T_{1,3} &= 0 \\ T_{2,3} + 75 + T_{1,4} + T_{1,2} - 4T_{1,3} &= 0 \\ T_{1,2} - 4T_{1,3} + T_{1,4} + T_{2,3} &= -75 \end{aligned} \tag{E4.4}$$

 $i=1$  and  $j=4$ 

$$\begin{aligned} T_{2,4} + T_{0,4} + T_{1,5} + T_{1,3} - 4T_{1,4} &= 0 \\ T_{2,4} + 75 + 300 + T_{1,3} - 4T_{1,4} &= 0 \\ T_{1,3} - 4T_{1,4} + T_{2,4} &= -375 \end{aligned} \tag{E4.5}$$

 $i=2$  and  $j=1$ 

$$\begin{aligned} T_{3,1} + T_{1,1} + T_{2,2} + T_{2,0} - 4T_{2,1} &= 0 \\ T_{3,1} + T_{1,1} + T_{2,2} + 50 - 4T_{2,1} &= 0 \\ T_{1,1} - 4T_{2,1} + T_{2,2} + T_{3,1} &= -50 \end{aligned} \tag{E4.6}$$

 $i=2$  and  $j=2$ 

$$\begin{aligned} T_{3,2} + T_{1,2} + T_{2,3} + T_{2,1} - 4T_{2,2} &= 0 \\ T_{1,2} + T_{2,1} - 4T_{2,2} + T_{2,3} + T_{3,2} &= 0 \end{aligned} \tag{E4.7}$$

 $i=2$  and  $j=3$ 

$$\begin{aligned} T_{3,3} + T_{1,3} + T_{2,4} + T_{2,2} - 4T_{2,3} &= 0 \\ T_{1,3} + T_{2,2} - 4T_{2,3} + T_{2,4} + T_{3,3} &= 0 \end{aligned} \tag{E4.8}$$

 $i=2$  and  $j=4$ 

$$\begin{aligned} T_{3,4} + T_{1,4} + T_{2,5} + T_{2,3} - 4T_{2,4} &= 0 \\ T_{3,4} + T_{1,4} + 300 + T_{2,3} - 4T_{2,4} &= 0 \\ T_{1,4} + T_{2,3} - 4T_{2,4} + T_{3,4} &= -300 \end{aligned} \tag{E4.9}$$

 $i=3$  and  $j=1$ 

$$\begin{aligned} T_{4,1} + T_{2,1} + T_{3,2} + T_{3,0} - 4T_{3,1} &= 0 \\ T_{4,1} + T_{2,1} + T_{3,2} + 50 - 4T_{3,1} &= 0 \\ T_{2,1} - 4T_{3,1} + T_{3,2} + T_{4,1} &= -50 \end{aligned} \tag{E4.10}$$

 $i=3$  and  $j=2$ 

$$\begin{aligned} T_{4,2} + T_{2,2} + T_{3,3} + T_{3,1} - 4T_{3,2} &= 0 \\ T_{2,2} + T_{3,1} - 4T_{3,2} + T_{3,3} + T_{4,2} &= 0 \end{aligned} \tag{E4.11}$$

 $i=3$  and  $j=3$ 

$$\begin{aligned} T_{4,3} + T_{2,3} + T_{3,4} + T_{3,2} - 4T_{3,3} &= 0 \\ T_{2,3} + T_{3,2} - 4T_{3,3} + T_{3,4} + T_{4,3} &= 0 \end{aligned} \tag{E4.12}$$

 $i=3$  and  $j=4$ 

$$T_{4,4} + T_{2,4} + T_{3,5} + T_{3,3} - 4T_{3,4} = 0$$

$$\begin{aligned} T_{4,4} + T_{2,4} + 300 + T_{3,3} - 4T_{3,4} &= 0 \\ T_{2,4} + T_{3,3} - 4T_{3,4} + T_{4,4} &= -300 \end{aligned} \quad (\text{E4.13})$$

Now for  $i = 4$  (for this problem  $m = 4$ ), all of these nodes are on the right hand side boundary which is insulated, so we use Equation (18) for  $j = 1, 2, 3$  and  $4$ . Substituting  $i$  for  $m$  variables gives

$i=4$  and  $j=1$

$$\begin{aligned} 2T_{3,1} + T_{4,0} + T_{4,2} - 4T_{4,1} &= 0 \\ 2T_{3,1} + 50 + T_{4,2} - 4T_{4,1} &= 0 \\ 2T_{3,1} - 4T_{4,1} + T_{4,2} &= -50 \end{aligned} \quad (\text{E4.14})$$

$i=4$  and  $j=2$

$$\begin{aligned} 2T_{3,2} + T_{4,1} + T_{4,3} - 4T_{4,2} &= 0 \\ 2T_{3,2} + T_{4,1} - 4T_{4,2} + T_{4,3} &= 0 \end{aligned} \quad (\text{E4.15})$$

$i=4$  and  $j=3$

$$\begin{aligned} 2T_{3,3} + T_{4,2} + T_{4,4} - 4T_{4,3} &= 0 \\ 2T_{3,3} + T_{4,2} - 4T_{4,3} + T_{4,4} &= 0 \end{aligned} \quad (\text{E4.16})$$

$i=4$  and  $j=4$

$$\begin{aligned} 2T_{3,4} + T_{4,3} + T_{4,5} - 4T_{4,4} &= 0 \\ 2T_{3,4} + T_{4,3} + 300 - 4T_{4,4} &= 0 \\ 2T_{3,4} + T_{4,3} - 4T_{4,4} &= -300 \end{aligned} \quad (\text{E4.17})$$

Equations (E4.2) to (E4.17) represent a set of sixteen simultaneous linear equations, and solving them gives the temperature at sixteen interior nodes. The solution is

$$\begin{bmatrix} T_{1,1} \\ T_{1,2} \\ T_{1,3} \\ T_{1,4} \\ T_{2,1} \\ T_{2,2} \\ T_{2,3} \\ T_{2,4} \\ T_{3,1} \\ T_{3,2} \\ T_{3,3} \\ T_{3,4} \\ T_{4,1} \\ T_{4,2} \\ T_{4,3} \\ T_{4,4} \end{bmatrix} = \begin{bmatrix} 76.8254 \\ 99.4444 \\ 128.617 \\ 180.410 \\ 82.8571 \\ 117.335 \\ 159.614 \\ 218.021 \\ 87.2678 \\ 127.426 \\ 174.483 \\ 232.060 \\ 88.7882 \\ 130.617 \\ 178.830 \\ 232.738 \end{bmatrix} {}^{\circ}\text{C}$$

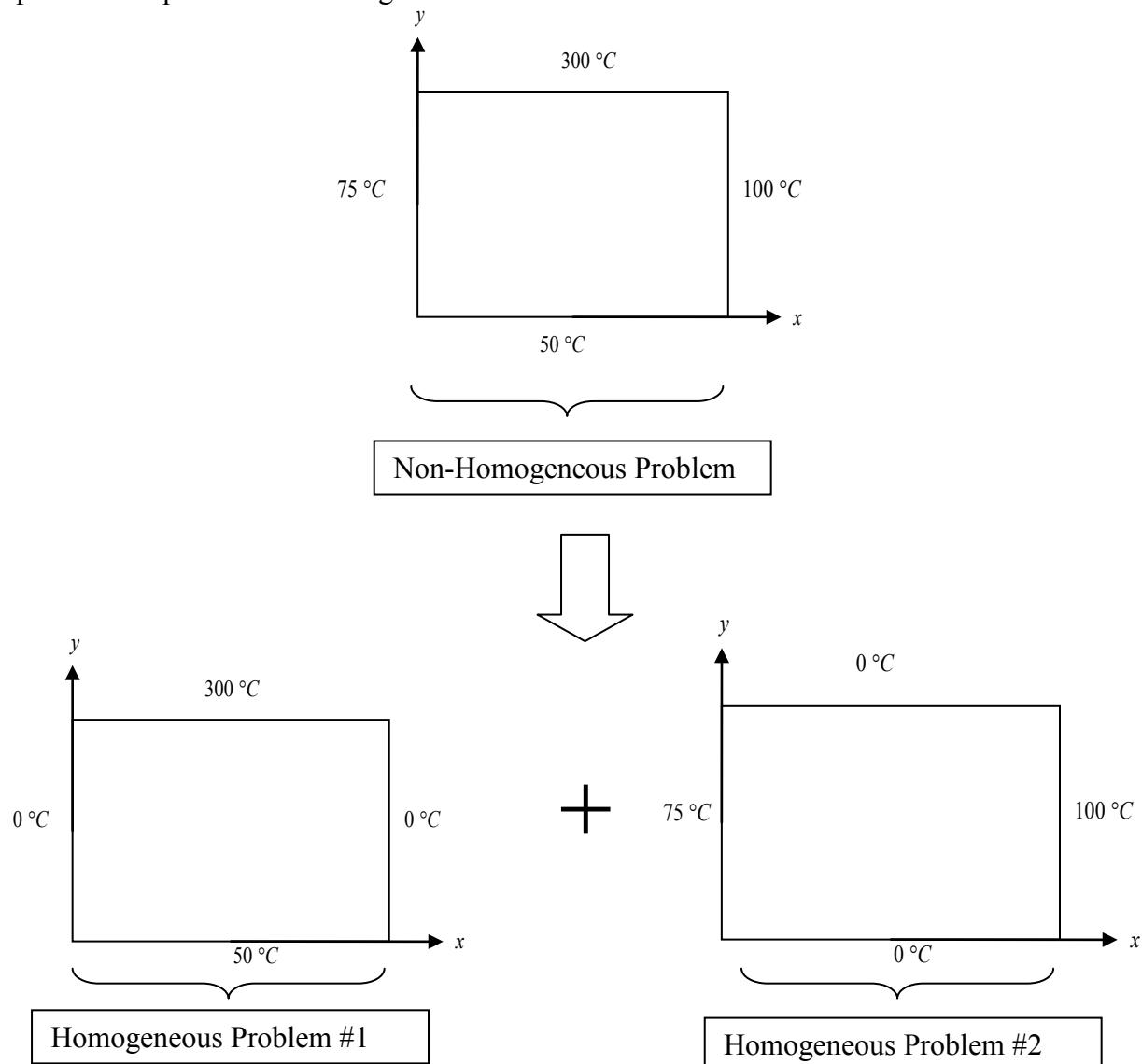
## APPENDIX A

### Analytical Solution of Example 1

The differential equation for Example 1 is

$$\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} = 0.$$

The temperature boundary conditions are given on the four sides of the plate (Dirichlet boundary conditions). This problem is too complex to solve analytically. To make this simple, we split the problem into two problems and using the principle of superposition. We then superimpose the solutions of the two simple problems to get the final solution. How the total problem is split is shown in Figure A.1.



**Figure A.1:** Splitting of non-homogeneous problem into two homogeneous problems

From Figure A.1, the total solution of the problem is obtained by the summation of the solutions of Problem 1 and Problem 2.

### Solution to Problem 1

Let the solution to problem 1 be  $T_1$ .

Then the differential equation is

$$\frac{\partial^2 T_1}{\partial x^2} + \frac{\partial^2 T_1}{\partial y^2} = 0 \quad 0 < x < L ; \quad 0 < y < W \quad (\text{A.1})$$

with boundary conditions

$$T_1(0, y) = 0 \quad (\text{A.2})$$

$$T_1(2.4, y) = 0 \quad (\text{A.3})$$

$$T_1(x, 0) = 50 \quad (\text{A.4})$$

$$T_1(x, 3.0) = 300 \quad (\text{A.5})$$

Let  $T_1$  be a function of  $X(x)$  and  $Y(y)$

$$T_1(x, y) = X(x)Y(y) \quad (\text{A.6})$$

Substituting Equation (A.6) in Equation (A.1), we have

$$\begin{aligned} X''Y + Y''X &= 0 \\ \frac{X''}{X} &= -\frac{Y''}{Y} \\ \frac{X''}{X} &= -\frac{Y''}{Y} = -\beta^2 \end{aligned} \quad (\text{A.7})$$

### Spatial Y solution

Now from Equation (A.7) we can write

$$\begin{aligned} \frac{Y''}{Y} &= \beta^2 \\ Y'' - \beta^2 Y &= 0 \end{aligned} \quad (\text{A.8})$$

Equation (A.8) is a homogeneous second order differential equation. These type of equations have the solution of the form  $Y(y) = e^{my}$ . Substituting  $Y(y) = e^{my}$  in Equation (A.8) we get,

$$m^2 e^{my} - \beta^2 e^{my} = 0$$

$$e^{my}(m^2 - \beta^2) = 0$$

$$m^2 - \beta^2 = 0$$

$$m_1, m_2 = \beta, -\beta$$

From the values of  $m_1$  and  $m_2$ , the solution of  $Y(y)$  is written as

$$Y(y) = A \cosh(\beta y) + B \sinh(\beta y) \quad (\text{A.9})$$

### Spatial X solution

Now from Equation (A.7) we can write

$$\begin{aligned} \frac{X''}{X} &= -\beta^2 \\ X'' + \beta^2 X &= 0 \end{aligned} \quad (\text{A.10})$$

Equation (A.10) is a homogeneous second order differential equation. These types of equations have the solution of the form  $X(x) = e^{mx}$ . Substituting  $X(x) = e^{mx}$  in Equation (A.10), we get

$$m^2 e^{mx} + \beta^2 e^{mx} = 0$$

$$e^{mx}(m^2 + \beta^2) = 0$$

$$m^2 + \beta^2 = 0$$

$$m_1, m_2 = i\beta, -i\beta$$

From the values of  $m_1$  and  $m_2$ , the solution of  $X(x)$  is written as

$$X(x) = C \cos(\beta x) + D \sin(\beta x) \quad (\text{A.11})$$

Substituting Equation (A.9) and Equation (A.11) in Equation (A.6) gives

$$T_1(x, y) = [C \cos(\beta x) + D \sin(\beta x)][A \cosh(\beta y) + B \sinh(\beta y)] \quad (\text{A.12})$$

To find the value of the constants we must use the boundary conditions. Applying boundary condition represented by Equation (A.2), we have

$$0 = C[A \cosh(\beta y) + B \sinh(\beta y)]$$

$$C = 0$$

Substituting  $C = 0$  in Equation (A.12), we have

$$\begin{aligned} T_1(x, y) &= D \sin(\beta x)[A \cosh(\beta y) + B \sinh(\beta y)] \\ &= \sin(\beta x)[A \cosh(\beta y) + B \sinh(\beta y)] \end{aligned} \quad (\text{A.13})$$

Applying the boundary condition represented by Equation (A.13), we have

$$0 = \sin(2.4\beta)[A \cosh(\beta y) + B \sinh(\beta y)]$$

$$0 = \sin(2.4\beta)$$

$$2.4\beta = n\pi$$

$$\beta = \frac{n\pi}{2.4} \quad (\text{A.14})$$

Substituting Equation (A.14) in Equation (A.13)

$$T_1(x, y) = \sin\left(\frac{n\pi}{2.4}x\right) \left[ A \cosh\left(\frac{n\pi}{2.4}y\right) + B \sinh\left(\frac{n\pi}{2.4}y\right) \right]$$

Since the general solution can have any value of  $n$ ,

$$T_1(x, y) = \sum_{n=1}^{\infty} \sin\left(\frac{n\pi}{2.4}x\right) \left[ A_n \cosh\left(\frac{n\pi}{2.4}y\right) + B_n \sinh\left(\frac{n\pi}{2.4}y\right) \right] \quad (\text{A.15})$$

Applying boundary condition represented by Equation (A.4), we have

$$50 = \sum_{n=1}^{\infty} A_n \sin\left(\frac{n\pi}{2.4}x\right) \quad (\text{A.16})$$

A half range sine series is given by

$$f(x) = \sum_{n=1}^{\infty} A_n \sin\left(\frac{n\pi}{L}x\right)$$

where

$$A_n = \frac{2}{L} \int_0^L f(x) \sin\left(\frac{n\pi}{L}x\right) dx$$

Comparing Equation (A.16) with half range sine series, Equation (A.16) is a half-range expression of 50 in sine series with  $L = 2.4$ . Therefore

$$\begin{aligned}
 A_n &= \frac{2}{2.4} \int_0^{2.4} 50 \sin\left(\frac{n\pi}{2.4}x\right) dx \\
 &= \frac{1}{1.2} 50 \int_0^{2.4} \sin\left(\frac{n\pi}{2.4}x\right) dx \\
 &= \frac{50}{\frac{n\pi}{2.4} 1.2} \left[ -\cos\left(\frac{n\pi}{2.4}x\right) \right]_0^{2.4} \\
 &= \frac{50 \times 2.4}{1.2 \times n\pi} [-\cos(n\pi) + 1] \\
 &= \frac{100}{n\pi} [-\cos(n\pi) + 1] \\
 &= \frac{100}{n\pi} [1 - (-1)^n]
 \end{aligned} \tag{A.17}$$

Applying boundary condition represented by Equation (A.5), we have

$$300 = \sum_{n=1}^{\infty} \sin\left(\frac{n\pi}{2.4}x\right) \left[ A_n \cosh\left(\frac{n\pi}{2.4}3.0\right) + B_n \sinh\left(\frac{n\pi}{2.4}3.0\right) \right] \tag{A.18}$$

Solving Equation (A.18) for  $B_n$  gives

$$\begin{aligned}
 B_n &= \frac{1}{\sin\left(\frac{3n\pi}{2.4}\right)} \left\{ \frac{2}{2.4} \int_0^{2.4} 300 \sin\left(\frac{n\pi}{2.4}x\right) dx - A_n \cos\left(\frac{3n\pi}{2.4}\right) \right\} \\
 &= \frac{1}{\sin\left(\frac{3n\pi}{2.4}\right)} \left\{ \frac{600}{2.4} \left[ \frac{-\cos\left(\frac{n\pi x}{2.4}\right)}{\frac{n\pi}{2.4}} \right]_0^{2.4} - A_n \cos\left(\frac{3n\pi}{2.4}\right) \right\} \\
 &= \frac{1}{\sin\left(\frac{3n\pi}{2.4}\right)} \left\{ \frac{600}{2.4} \frac{2.4}{n\pi} \left[ -\cos\left(\frac{n\pi x}{2.4}\right) \right]_0^{2.4} - A_n \cos\left(\frac{3n\pi}{2.4}\right) \right\} \\
 &= \frac{1}{\sin\left(\frac{3n\pi}{2.4}\right)} \left\{ \frac{600}{n\pi} [-\cos(n\pi) + 1] - A_n \cos\left(\frac{3n\pi}{2.4}\right) \right\} \\
 &= \frac{1}{\sin\left(\frac{3n\pi}{2.4}\right)} \left\{ \frac{600}{n\pi} [1 - (-1)^n] - A_n \cos\left(\frac{3n\pi}{2.4}\right) \right\}
 \end{aligned} \tag{A.19}$$

From Equations (A.15), (A.17) and (A.19), the solution  $T_1$  is given as

$$T_1(x, y) = \sum_{n=1}^{\infty} \sin\left(\frac{n\pi}{2.4}x\right) \left[ A_n \cosh\left(\frac{n\pi}{2.4}y\right) + B_n \sinh\left(\frac{n\pi}{2.4}y\right) \right] \quad (\text{A.20})$$

where

$$A_n = \frac{100}{n\pi} [1 - (-1)^n] \text{ and}$$

$$B_n = \frac{1}{\sin\left(\frac{3n\pi}{2.4}\right)} \left\{ \frac{600}{n\pi} [1 - (-1)^n] - A_n \cos\left(\frac{3n\pi}{2.4}\right) \right\}$$

### Solution Problem 2

Let the solution to Problem 2 be  $T_2$ . Problem 2 can be solved similarly as Problem 1. The solution to Problem 2 is

$$T_2(x, y) = \sum_{n=1}^{\infty} \sin\left(\frac{n\pi}{3}y\right) \left[ C_n \cosh\left(\frac{n\pi}{3}x\right) + D_n \sinh\left(\frac{n\pi}{3}x\right) \right] \quad (\text{A.21})$$

where

$$C_n = \frac{150}{n\pi} [1 - (-1)^n] \text{ and}$$

$$D_n = \frac{1}{\sin\left(\frac{2.4n\pi}{3}\right)} \left\{ \frac{200}{n\pi} [1 - (-1)^n] - C_n \cos\left(\frac{2.4n\pi}{3}\right) \right\}$$

### Overall Solution

The overall solution to the problem is

$$\begin{aligned} T(x, y) &= T_1(x, y) + T_2(x, y) \\ T(x, y) &= \sum_{n=1}^{\infty} \sin\left(\frac{n\pi}{2.4}x\right) \left[ A_n \cosh\left(\frac{n\pi}{2.4}y\right) + B_n \sinh\left(\frac{n\pi}{2.4}y\right) \right] + \\ &\quad \sum_{n=1}^{\infty} \sin\left(\frac{n\pi}{3}y\right) \left[ C_n \cosh\left(\frac{n\pi}{3}x\right) + D_n \sinh\left(\frac{n\pi}{3}x\right) \right] \end{aligned}$$

where

$$A_n = \frac{100}{n\pi} [1 - (-1)^n],$$

$$B_n = \frac{1}{\sin\left(\frac{3n\pi}{2.4}\right)} \left\{ \frac{600}{n\pi} [1 - (-1)^n] - A_n \cos\left(\frac{3n\pi}{2.4}\right) \right\},$$

$$C_n = \frac{150}{n\pi} [1 - (-1)^n],$$

$$D_n = \frac{1}{\sin\left(\frac{2.4n\pi}{3}\right)} \left\{ \frac{200}{n\pi} [1 - (-1)^n] - C_n \cos\left(\frac{2.4n\pi}{3}\right) \right\}.$$

---

**PARTIAL DIFFERENTIAL EQUATIONS**

---

Topic	Parabolic Differential Equations
Summary	Textbook notes for the parabolic partial differential equations
Major	All engineering majors
Authors	Autar Kaw, Sri Harsha Garapati, Frederik Schousboe
Date	February 21, 2011
Web Site	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

---

# Chapter 10.04

## Introduction to Finite Element Methods

*After reading this chapter, you should be able to:*

1. *Understand the basics of finite element methods using a one-dimensional problem.*

In the last fifty years, the use of approximation solution methods to solve complex problems in engineering and science has grown significantly. The widespread availability of powerful digital computers and commercial computational software based on these approximation methods with efficient solution algorithms has made them practical. In this chapter, we are introducing the student to finite methods of solving differential equations. We provide an elementary background on how finite element methods work, while using a single example to illustrate the approach, and discuss the accuracy and efficacy of the method.

The single example chosen is a classical problem of a uniformly pressurized thick-walled cylinder with an axis-symmetric response (Figure 1). This problem is chosen since it is simple enough to have an analytical solution, but complex enough such that its finite element method solution can be generalized for problems that are more complicated. We must first define the problem, and then develop the exact solution so that we may compare it with the finite element methods result.

### Thick-Wall Cylinder Problem

#### Problem Definition

Consider a thick-walled cylinder as depicted in Figure 1, with the following material properties:

Young's modulus  $E$ ,  
Poisson's ratio  $\nu$   
inner radius  $a$   
outer radius,  $b$   
uniform internal pressure  $p_i$   
external pressure,  $p_o$

Find the following variables in the cylinder. Plane stress state is assumed.

radial displacement,  $u$

radial stress,  $\sigma_r$   
tangential stress,  $\sigma_\theta$

### Numerical Example Problem

For demonstrating the use of approximate solution methods in solving the problem numerically, the following data is used:

$$a = 0.25 \text{ m}$$

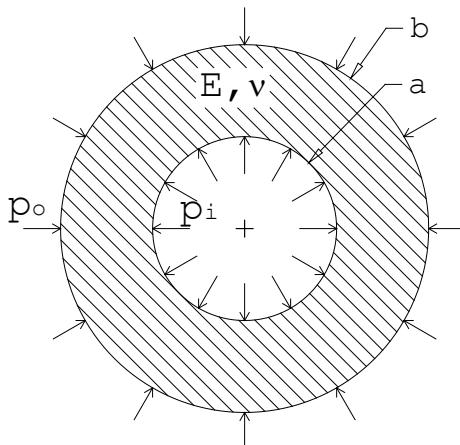
$$b = 0.5 \text{ m}$$

$$p_i = 200 \text{ MPa}$$

$$p_o = 0$$

$$E = 207 \text{ GPa}$$

$$\nu = 0.3$$



**Figure 1:** Pressured thick-wall cylinder problem

### Mathematical Formulation

The solution of the thick-wall cylinder problem can be found by solving the equation of compatibility in polar coordinates, which is a fourth order partial differential equation of Airy stress function (1), or by using axisymmetry conditions to formulate the problem as a second order differential equation of displacement (2), or equivalent forms (potential energy, integral equation, etc.). The last approach is adopted in this paper, as it is direct and does not require inverse or semi-inverse solution methods (1, 2). The details of this approach are given in (2) and the relevant formulas are summarized as follows. The radial strain,  $\varepsilon_r$ , tangential strain,  $\varepsilon_\theta$ , in terms of radial displacement,  $u$  are given as

$$\varepsilon_r = \frac{du}{dr} \quad (1)$$

$$\varepsilon_\theta = \frac{u}{r} \quad (2)$$

The radial stress,  $\sigma_r$ , and tangential stress,  $\sigma_\theta$ , in terms of radial displacement,  $u$ , are given as

$$\sigma_r = \frac{E}{1-\nu^2} \left( \frac{du}{dr} + \nu \frac{u}{r} \right) \quad (3)$$

$$\sigma_\theta = \frac{E}{1-\nu^2} \left( \nu \frac{du}{dr} + \frac{u}{r} \right) \quad (4)$$

The governing equation for radial displacement,  $u$ , is given by

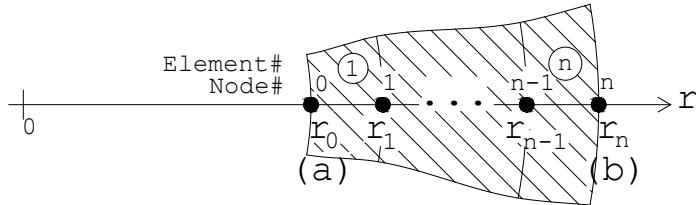
$$\frac{d^2u}{dr^2} + \frac{1}{r} \frac{du}{dr} - \frac{u}{r^2} = 0 \quad (5)$$

Using Equations 3-4, the boundary conditions  $\sigma_r(a) = -p_i$  and  $\sigma_r(b) = -p_o$  can be rewritten as

$$u'(a) + \nu \frac{u(a)}{a} = -\frac{1-\nu^2}{E} p_i \quad (6)$$

$$u'(b) + \nu \frac{u(b)}{b} = -\frac{1-\nu^2}{E} p_o \quad (7)$$

First, the exact solution is found, and then a finite element method is presented through solving the example problem. Nodal points chosen for the finite element method are uniformly spaced for convenience. Figure 2 shows how the nodal points and elements are numbered.



**Figure 2:** Numbering of nodal points and elements

### Exact Solution

The exact solution of displacement can be found directly by solving the governing differential equation, Equation (5), with associated boundary conditions, Equations (6-7), and then substituting it into Equations (3-4) to give an exact solution of stresses. The exact solutions (7) of radial displacement, radial stress, and tangential stress are obtained as

$$u = \frac{1-\nu}{E} \frac{(a^2 p_i - b^2 p_o)r}{b^2 - a^2} + \frac{1+\nu}{E} \frac{(p_i - p_o) a^2 b^2}{(b^2 - a^2)r} \quad (8)$$

$$\sigma_r = \frac{a^2 p_i - b^2 p_o}{b^2 - a^2} - \frac{(p_i - p_o) a^2 b^2}{(b^2 - a^2)r^2} \quad (9)$$

$$\sigma_\theta = \frac{a^2 p_i - b^2 p_o}{b^2 - a^2} + \frac{(p_i - p_o) a^2 b^2}{(b^2 - a^2)r^2} \quad (10)$$

### Solution for Example Problem

Substituting the numerical data into Equations (8-10), the exact solution for the example problem is

$$u = \left( 0.2254r + \frac{0.1047}{r} \right) \times 10^{-3} \quad (11)$$

$$\sigma_r = \left( 66.67 - \frac{16.67}{r^2} \right) \times 10^6 \quad (13)$$

$$\sigma_\theta = \left( 66.67 + \frac{16.67}{r^2} \right) \times 10^6 \quad (14)$$

Evaluating the solution at three nodal points (inner edge,  $r = 0.25\text{ m}$ ; mid-point,  $r = 0.375\text{ m}$ ; and outer edge,  $r = 0.5\text{ m}$ ) along the radial location for comparison, the resulted values are given in Table 1.

**Table 1:** Exact solution evaluated at nodal points

$r$ (m)	0.25	0.375	0.5
$u$ (mm)	0.4750	0.3637	0.3221
$\sigma_r$ (MPa)	-200	-51.85	0
$\sigma_\theta$ (MPa)	333.3	185.2	133.3

### What are Finite Element Methods?

The finite element method is a technique used to solve differential equations (ordinary or partial). They are mainly used to solve real world problems, as the differential equations that govern these problems cannot be solved exactly, or may be too intractable to be solved exactly.

The finite element methods use techniques to approximate the dependant variables of the differential equations by functions, and then reduce the unknowns in these functions to a set of simultaneous linear equations. These equations can then be solved by various numerical techniques. However, one needs to understand that finite element methods use a function, not the differential equation itself, to develop the approximate solution. This is unlike the finite difference methods, where the derivatives in the differential equations are approximated by finite divided difference methods. The functions used in the finite element methods are integral equations. In the case of the pressure vessel, these equations would model the total potential energy due to internal stresses and external loads

The Rayleigh-Ritz method can be viewed as a form of a finite element method where it reduces a continuous problem to a problem with a finite number of degrees of freedom. The Rayleigh-Ritz method is based on the principle of stationary potential energy, which states:

“Among all admissible configurations of a conservative system, those that satisfy the equations of equilibrium make the potential energy stationary with respect to small variations of displacement. If the stationary condition is a minimum, the equilibrium state is stable.”

Mathematically speaking, the Rayleigh-Ritz method is a variational method, based on the idea of finding a solution that minimizes a functional. For elasticity problems, the functional is the total potential energy. The solution must be admissible, that is, satisfying internal compatibility (e.g., continuity of displacement) and essential boundary conditions. For problems where displacements are primary unknowns, essential boundary conditions are prescriptions of displacement and non-essential boundary conditions are prescriptions of stress. Since the problem considered here, the thick-walled pressured cylinder problem where the primary unknown is radial displacement, has no prescription of displacement, there is no essential boundary condition.

### Potential Energy Formulation

The cylinder is assumed to be in a plane stress state which gives a strain energy density,  $U_0$  as

$$U_0 = \frac{1}{2}(\sigma_r \varepsilon_r + \sigma_\theta \varepsilon_\theta) \quad (15)$$

by using Equations (1-4), we get

$$U_0 = \frac{E}{2(1-\nu^2)} \left[ \left( \frac{du}{dr} \right)^2 + 2\nu \left( \frac{du}{dr} \right) \left( \frac{u}{r} \right) + \left( \frac{u}{r} \right)^2 \right] \quad (16)$$

Total strain energy,  $U$  of the cylinder is

$$U = \int_V U_0 dV = \int_0^L \int_0^{2\pi} \int_a^b U_0 r dr d\theta dz = 2\pi L \int_a^b U_0 r dr \quad (17)$$

where,

$L$  = cylinder length

Work done,  $W$  by external forces (internal and external pressures) is

$$W = \int_{S_i} p_i u(a) ds - \int_{S_o} p_o u(b) ds = 2\pi a L p_i u(a) - 2\pi b L p_o u(b) \quad (18)$$

where,

$S_i$  = inner cylinder surface

$S_o$  = outer cylinder surface

The total potential energy of the cylinder,  $\Pi$  is found as

$$\Pi = U - W = 2\pi L \left( \int_a^b U_0 r dr - ap_i u(a) + bp_o u(b) \right) \quad (19)$$

### Rayleigh-Ritz Method

The Rayleigh-Ritz method can be outlined as follows. The potential energy of the system is given as  $\Pi = \Pi(u', u, r)$ .

Assume a trial solution of the form:  $u = f(r, C_0, C_1, \dots, C_m)$

where  $C_i$ 's ( $i = 0..m$ ) are unknown parameters, and  $f$  is a known function. In this paper, we consider linear piecewise continuous functions.

Apply admissibility conditions to the trial solution. If there are  $m-n$  admissibility conditions, we have  $m-n$  equations of unknown parameters.

Solve the system of  $m-n$  equations for  $m-n$  unknowns  $C_{n+1} \dots C_m$ , and then plug them back into the trial solution, we obtain a new trial solution that is admissible and has fewer unknowns ( $n$  unknowns)  $u = f(r, C_0, C_1, \dots, C_n)$ .

Substitute the trial solution into the expression of potential energy. The stationary condition for potential energy  $\delta\Pi = 0$  gives

$$\left\{ \frac{\partial\Pi}{\partial C_i} = 0 \right\}, i = 0..n \quad (20)$$

Here we have a system of  $n$  algebraic equations with  $n$  unknowns. Solving this system of equations, we find the unknown parameters and thus the approximate solution for the radial displacement.

Substitute the found solution for radial displacement into Equations (3-4) to find the approximation solution for radial stress and tangential stress.

### Linear Piecewise Continuous Solution for Example Problem

Consider the case of  $n = 2$  with uniform spacing nodal points. The step size for locating nodal points is calculated as

$$\begin{aligned} h &= (b-a)/n \\ &= (0.5-0.25)/2 \\ &= 0.125. \end{aligned}$$

The radial coordinates of the nodal points are  $r_o = a = 0.25$ ,  $r_1 = 0.375$ ,  $r_2 = b = 0.5$ .

The displacement field is assumed to be a piecewise continuous function of two linear segments as

$$u = \begin{cases} C_0 + C_1 r, & 0.25 \leq r \leq 0.375 \\ C_3 + C_2 r, & 0.375 \leq r \leq 0.5 \end{cases} \quad (21)$$

To make the trial solution, Equation (21), admissible, it must be continuous at  $r = 0.375$ , which means

$$C_0 + 0.375C_1 = C_3 + 0.375C_2, \text{ or} \quad (22)$$

$$C_3 = C_0 + 0.375C_1 - 0.375C_2 \quad (23)$$

The trial solution, Equation (21), then becomes

$$u = \begin{cases} C_0 + C_1 r, & 0.25 \leq r \leq 0.375 \\ C_0 + 0.375C_1 - 0.375C_2 + C_2 r, & 0.375 \leq r \leq 0.5 \end{cases} \quad (24)$$

Substituting Equation (24) and the given numerical data into Equation (19), the total potential energy,  $\Pi$  in the cylinder is found as

$$\begin{aligned}\Pi = & 2\pi L(78.84C_0^2 + 61.50C_0C_1 + 12.42C_0C_2 + 16.15C_1^2 + 4.659C_1C_2 \\ & + 6.912C_2^2 - 0.05000C_0 - 0.01250C_1) \times 10^9\end{aligned}\quad (25)$$

The condition that the total potential energy  $\Pi$  is stationary,

$$\left\{ \frac{\partial \Pi}{\partial C_0} = 0, \frac{\partial \Pi}{\partial C_1} = 0, \frac{\partial \Pi}{\partial C_2} = 0 \right\}$$

Which gives a system of algebraic equations of the unknown coefficients as

$$\begin{cases} 157.7C_0 + 61.50C_1 + 12.42C_2 = 0.05000 \\ 61.50C_0 + 32.31C_1 + 4.659C_2 = 0.01250 \\ 12.42C_0 + 4.659C_1 + 13.82C_2 = 0 \end{cases} \quad (26)$$

The unknown coefficients are found as

$$\begin{cases} C_0 = 0.0006737 \\ C_1 = -0.0008496 \\ C_2 = -0.0003191 \end{cases} \quad (27)$$

Substituting Equation (27) into Equation (24), the approximate solution for radial displacement is

$$u = \begin{cases} 0.0006737 - 0.0008496r, & 0.25 \leq r \leq 0.375 \\ 0.0004748 - 0.0003191r, & 0.375 \leq r \leq 0.5 \end{cases} \quad (28)$$

Substituting the numerical data and displacement solution from Equation (28) into Equations (3-4), we find the radial and tangential stresses as

$$\sigma_r = \begin{cases} \frac{45.97}{r} - 251.2, & 0.25 < r < 0.375 \\ \frac{32.40}{r} - 94.38, & 0.375 < r < 0.5 \end{cases} \quad (29)$$

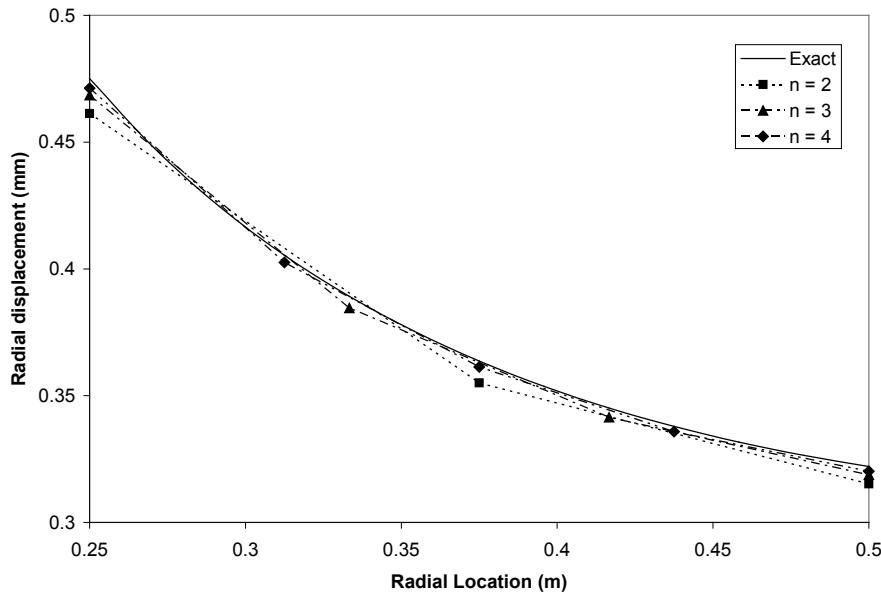
$$\sigma_\theta = \begin{cases} \frac{153.2}{r} - 251.2, & 0.25 < r < 0.375 \\ \frac{108.0}{r} - 94.38, & 0.375 < r < 0.5 \end{cases} \quad (30)$$

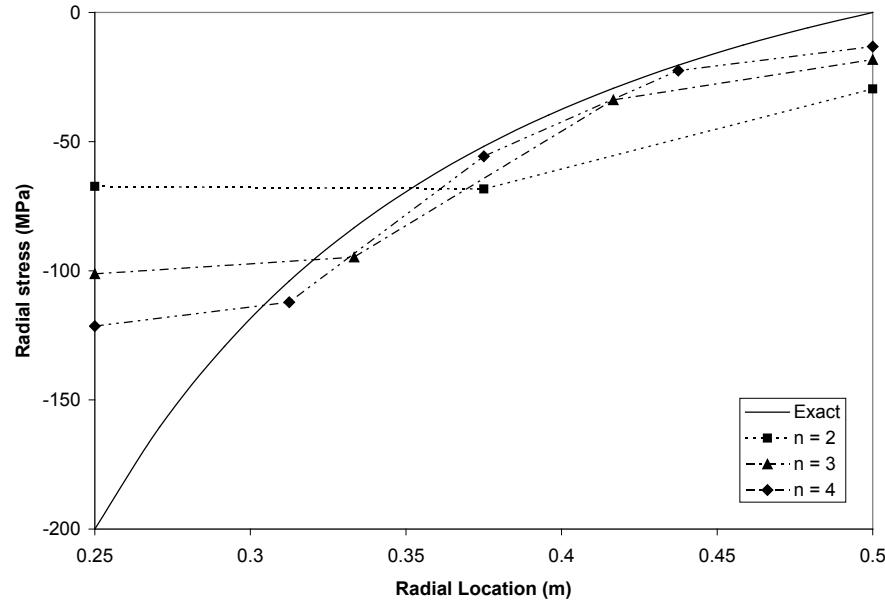
The solution of the radial displacement is continuous, since we have forced the trial solution to be admissible from the beginning, while the solutions for stresses are discontinuous at the interior knot ( $r = 0.375$ ) between the two segments (elements). To have reasonable results, in practice, the stress value at the interior knot is taken as the average of two stress values. The numerical solution with  $n = 2$  of the example problem is given in Table 4.

**Table 4:** Numerical solution, finite element method ( $n = 2$ )

$r$ (m)	0.25	0.375	0.5
$u$ (mm)	0.4613	0.3551	0.3152
$\sigma_r$ (MPa)	- 67.35	- 68.32	- 29.58
$\sigma_\theta$ (MPa)	361.7	175.5	121.6

The exact solution and numerical solutions with various values of number of nodal points,  $n = 2$ , 3, and 4, are given in Figure 5 for radial displacement, and Figure 6 for radial stress.

**Figure 5:** Radial displacement as a function of radial location (Finite element method)



**Figure 6:** Solution of radial stress as a function of radial location

The solution plots show that the approximate solutions approach the exact solution as the number of piecewise continuous functions increase. However, they do not satisfy the boundary conditions of radial stress. The assumption of the piecewise continuous solution as opposed to a continuous solution makes computation easier for a high number of segments in the piecewise functions, but it has the drawback of the discontinuity of stresses at the interior knots of the piecewise continuous function.

## REFERENCES

- [1] S.P. Timoshenko, J.N. Goodier, *Theory of Elasticity*, McGraw-Hill, 1970.
- [2] A.C. Ugural, S.K. Fenster, *Advanced Strength and Applied Elasticity*, 3rd Ed. Prentice-Hall PTR, 1995.
- [3] A.P. Boresi, K.P. Chong, *Approximate Solution Methods in Engineering Mechanics*, Elsevier Applied Science, 1991.
- [4] R.D. Cook, D.S. Malkus, M.E. Plesha, *Concepts and Applications of Finite Element Analysis*, 3rd Ed. Wiley, 1989.
- [5] W.S. Hall, *The Boundary Element Method*, Kluwer Academic Publishers, 1994.

---

## PARTIAL DIFFERENTIAL EQUATIONS

---

Topic	Introduction to Finite Element Methods
Summary	Textbook notes for the introduction of partial differential equations
Major	All engineering majors
Authors	Autar Kaw, Sun Ho
Date	July 11, 2011
Web Site	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

---

## Chapter 11.01

# Introduction to Fourier Series

In general, curve fitting interpolation through a set of data points can be done by a linear combination of polynomial functions, with based functions  $1, x, x^2, \dots, x^m$ . In this chapter, however, trigonometric functions such as  $1, \cos(x), \cos(2x), \dots, \cos(nx)$ , and  $\sin(x), \sin(2x), \dots, \sin(nx)$  will be used as based functions. In the former, the unknown coefficients of based functions can be found by solving the associated linear simultaneous equations (where the number of unknown coefficients will be matched with the same number of equations, provided by a set of given data points). In the latter, however, the unknown coefficients can be efficiently solved (by exploiting special properties of trigonometric functions) without requiring solving the expensive simultaneous linear equations (more details will be explained in Equation 6 of Chapter 11.05).

### Introduction

The following relationships can be readily established, and will be used in subsequent sections for derivation of useful formulas for the unknown Fourier coefficients, in both time and frequency domains.

$$\begin{aligned} \int_0^T \sin(kw_0 t) dt &= \int_0^T \cos(kw_0 t) dt \\ &= 0 \end{aligned} \tag{1}$$

$$\begin{aligned} \int_0^T \sin^2(kw_0 t) dt &= \int_0^T \cos^2(kw_0 t) dt \\ &= \frac{T}{2} \end{aligned} \tag{2}$$

$$\int_0^T \cos(kw_0 t) \sin(gw_0 t) dt = 0 \tag{3}$$

$$\int_0^T \sin(kw_0 t) \sin(gw_0 t) dt = 0 \tag{4}$$

$$\int_0^T \cos(kw_0 t) \cos(gw_0 t) dt = 0 \tag{5}$$

where

$$\omega_0 = 2\pi f \quad (6)$$

$$f = \frac{1}{T} \quad (7)$$

where  $f$  and  $T$  represents the frequency (in cycles/time) and period (in seconds) respectively. Also,  $k$  and  $g$  are integers.

A periodic function  $f(t)$  with a period  $T$  should satisfy the following equation

$$f(t+T) = f(t) \quad (8)$$

### Example 1

Prove that

$$\int_0^{\pi} \sin(k\omega_0 t) dt = 0$$

for

$$\omega_0 = 2\pi f$$

$$f = \frac{1}{T}$$

and  $k$  is an integer.

### Solution

Let

$$A = \int_0^T \sin(k\omega_0 t) dt \quad (9)$$

$$= -\left(\frac{1}{k\omega_0}\right) [\cos(k\omega_0 t)]_0^T$$

$$A = \left(\frac{-1}{k\omega_0}\right) [\cos(k\omega_0 T) - \cos(0)] \quad (10)$$

$$= \left(\frac{-1}{k\omega_0}\right) [\cos(k2\pi) - 1]$$

$$= 0$$

### Example 2

Prove that

$$\int_0^{\pi} \sin^2(k\omega_0 t) dt = \frac{T}{2}$$

for

$$\omega_0 = 2\pi f$$

$$f = \frac{1}{T}$$

and  $k$  is an integer.

### Solution

Let

$$B = \int_0^T \sin^2(kw_0 t) dt \quad (11)$$

Recall

$$\sin^2(\alpha) = \frac{1 - \cos(2\alpha)}{2} \quad (12)$$

Thus,

$$B = \int_0^T \left[ \frac{1}{2} - \frac{1}{2} \cos(2kw_0 t) \right] dt \quad (13)$$

$$= \left[ \left( \frac{1}{2} \right) t - \left( \frac{1}{2} \right) \left( \frac{1}{2kw_0} \right) \sin(2kw_0 t) \right]_0^T$$

$$B = \left[ \frac{T}{2} - \frac{1}{4kw_0} \sin(2kw_0 T) \right] - [0] \quad (14)$$

$$= \frac{T}{2} - \left( \frac{1}{4kw_0} \right) \sin(2k * 2\pi)$$

$$= \frac{T}{2}$$

### Example 3

Prove that

$$\int_0^\pi \sin(gw_0 t) \cos(kw_0 t) dt = 0$$

for

$$w_0 = 2\pi f$$

$$f = \frac{1}{T}$$

and  $k$  and  $g$  are integers.

### Solution

Let

$$C = \int_0^T \sin(gw_0 t) \cos(kw_0 t) dt \quad (15)$$

Recall that

$$\sin(\alpha + \beta) = \sin(\alpha)\cos(\beta) + \sin(\beta)\cos(\alpha) \quad (16)$$

Hence,

$$C = \int_0^T [\sin[(g+k)w_0 t] - \sin(kw_0 t) \cos(gw_0 t)] dt \quad (17)$$

$$= \int_0^T \sin[(g+k)w_0 t] dt - \int_0^T \sin(kw_0 t) \cos(gw_0 t) dt \quad (18)$$

From Equation (1),

$$\int_0^T \sin[(g+k)w_0 t] dt = 0$$

then

$$C = 0 - \int_0^T \sin(kw_0 t) \cos(gw_0 t) dt \quad (19)$$

Adding Equations (15), (19),

$$\begin{aligned} 2C &= \int_0^T \sin(gw_0 t) \cos(kw_0 t) dt - \int_0^T \sin(kw_0 t) \cos(gw_0 t) dt \\ &= \int_0^T \sin[(gw_0 t) - (kw_0 t)] dt = \int_0^T \sin[(g-k)w_0 t] dt \end{aligned} \quad (20)$$

$2C = 0$ , since the right side of the above equation is zero (see Equation 1). Thus,

$$\begin{aligned} C &= \int_0^T \sin(gw_0 t) \cos(kw_0 t) dt = 0 \\ &= 0 \end{aligned} \quad (21)$$

#### Example 4

Prove that

$$\int_0^T \sin(kw_0 t) \sin(gw_0 t) dt = 0$$

for

$$w_0 = 2\pi f$$

$$f = \frac{1}{T}$$

$$k, g = \text{integers}$$

#### Solution

$$\text{Let } D = \int_0^T \sin(kw_0 t) \sin(gw_0 t) dt \quad (22)$$

Since

$$\cos(\alpha + \beta) = \cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta)$$

or

$$\sin(\alpha)\sin(\beta) = \cos(\alpha)\cos(\beta) - \cos(\alpha + \beta)$$

Thus,

$$D = \int_0^T \cos(kw_0 t) \cos(gw_0 t) dt - \int_0^T \cos[(k+g)w_0 t] dt \quad (23)$$

From Equation (1)

$$\int_0^T \cos[(k+g)w_0 t] dt = 0$$

then

$$D = \int_0^T \cos(kw_0 t) \cos(gw_0 t) dt - 0 \quad (24)$$

Adding Equations (23), (26)

$$\begin{aligned} 2D &= \int_0^T \sin(kw_0 t) \sin(gw_0 t) + \int_0^T \cos(kw_0 t) \cos(gw_0 t) dt \\ &= \int_0^T \cos[kw_0 t - gw_0 t] dt \\ &= \int_0^T \cos[(k-g)w_0 t] dt \end{aligned} \quad (25)$$

$2D = 0$ , since the right side of the above equation is zero (see Equation 1). Thus,

$$D \equiv \int_0^T \sin(kw_0 t) \sin(gw_0 t) dt = 0 \quad (26)$$

### FAST FOURIER TRANSFORM

Topic	Introduction to Fourier Series
Summary	Textbook notes on an introduction to Fourier series
Major	General Engineering
Authors	Duc Nguyen
Date	July 25, 2010
Web Site	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

## Chapter 11.02

### Continuous Fourier Series

For a function with period  $T$ , a continuous Fourier series can be expressed as [1-5]

$$f(t) = a_0 + \sum_{k=1}^{\infty} a_k \cos(kw_0 t) + b_k \sin(kw_0 t) \quad (1)$$

The unknown Fourier coefficients  $a_0$ ,  $a_k$  and  $b_k$  can be computed as

$$a_0 = \left( \frac{1}{T} \right) \int_0^T f(t) dt \quad (2)$$

Thus,  $a_0$  can be interpreted as the “average” function value between the period interval  $[0, T]$ .

$$a_k = \left( \frac{2}{T} \right) \int_0^T f(t) \cos(kw_0 t) dt \quad (3)$$

$\equiv a_{-k}$  (hence  $a_k$  is an “even” function)

$$b_k = \left( \frac{2}{T} \right) \int_0^T f(t) \sin(kw_0 t) dt \quad (4)$$

$\equiv -b_{-k}$  (hence  $b_k$  is an “odd” function)

#### Derivation of formulas for $a_0$ , $a_k$ and $b_k$

Integrating both sides of Equation 1 with respect to time, one gets

$$\int_0^T f(t) dt = \int_0^T a_0 dt + \int_0^T \sum_{k=1}^{\infty} a_k \cos(kw_0 t) dt + \int_0^T \sum_{k=1}^{\infty} b_k \sin(kw_0 t) dt \quad (5)$$

The second and third terms on the right hand side of the above equations are both zeros, due to the result stated in Equation (1) of Chapter 11.01.

Thus,

$$\begin{aligned} \int_0^T f(t) dt &= \left[ a_0 t \right]_0^T \\ &= a_0 T \end{aligned} \quad (6)$$

Hence,

$$a_0 = \left( \frac{1}{T} \right) \int_0^T f(t) dt \quad (7)$$

Now, if both sides of Equation (1) are multiplied by  $\sin(mw_0t)$  and then integrated with respect to time, one obtains

$$\begin{aligned} \int_0^T f(t) \times \sin(mw_0t) dt &= \int_0^T a_0 \sin(mw_0t) dt + \int_0^T \sum_{k=1}^{\infty} a_k \cos(kw_0t) \sin(mw_0t) dt \\ &\quad + \int_0^T \sum_{k=1}^{\infty} b_k \sin(kw_0t) \sin(mw_0t) dt \end{aligned} \quad (8)$$

Due to Equations (1) and (3) of Chapter 11.01, the first and second terms on the right hand side (RHS) of Equation (8) are zero.

Due to Equation (4) of Chapter 11.01, the third RHS term of Equation (8) is also zero, with the exception when  $k = m$ , which will become (by referring to Equation (2) of Chapter 11.01)

$$\begin{aligned} \int_0^T f(t) \sin(kw_0t) dt &= 0 + 0 + \int_0^T b_k \sin^2(kw_0t) dt \\ &= b_k \times \frac{T}{2} \end{aligned} \quad (9)$$

Thus,

$$b_k = \left( \frac{2}{T} \right) \int_0^T f(t) \sin(kw_0t) dt$$

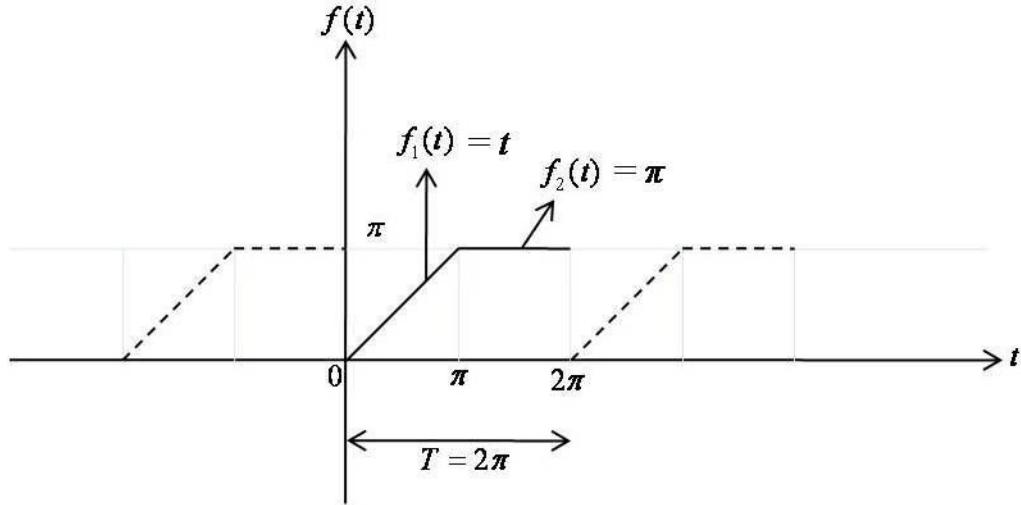
Similar derivation can be used to obtain  $a_k$ , as shown in Equation (3)

### A FORTRAN Program for finding Fourier Coefficients $a_0$ , $a_k$ , and $b_k$

Based upon the derived formulas for  $a_0$ ,  $a_k$  and  $b_k$  (shown in Equations 2-4), a FORTRAN/MATLAB computer program has been developed. (The program is available at [http://numericalmethods.eng.usf.edu/simulations/mlt/11fft/f\\_coeff\\_final.m](http://numericalmethods.eng.usf.edu/simulations/mlt/11fft/f_coeff_final.m))

### Example 1

Using the continuous Fourier series to approximate the following periodic function ( $T = 2\pi$  seconds) shown in Figure 1.



**Figure 1** A Periodic Function (Between 0 and  $2\pi$  ).

$$f(t) = \begin{cases} t & \text{for } 0 < t \leq \pi \\ \pi & \text{for } \pi \leq t < 2\pi \end{cases}$$

Specifically, find the Fourier coefficients  $a_0, a_1, \dots, a_8$  and  $b_1, \dots, b_8$ .

### Solution

The unknown Fourier coefficients  $a_0, a_k$  and  $b_k$  can be computed based on Equations (2–4); as following:

$$a_0 = \left( \frac{1}{T} \right) \int_0^{2\pi} f(t) dt$$

$$a_0 = \frac{1}{(2\pi)} \times \left\{ \int_0^{\pi} t dt + \int_{\pi}^{2\pi} \pi dt \right\}$$

$$a_0 = 2.35619$$

$$a_k = \left( \frac{2}{T} \right) \int_0^{2\pi} f(t) \cos(kw_0 t) dt$$

$$a_k = \left( \frac{2}{2\pi} \right) \times \left\{ \int_0^{\pi} t \cos\left(k \times \frac{2\pi}{T} \times t\right) dt + \int_{\pi}^{2\pi} \pi \cos\left(k \times \frac{2\pi}{T} \times t\right) dt \right\}$$

$$a_k = \left( \frac{1}{\pi} \right) \times \left\{ \int_0^{\pi} t \cos(kt) dt + \int_{\pi}^{2\pi} \pi \cos(kt) dt \right\}$$

The “integration by part” formula can be utilized to compute the first integral on the right-hand-side of the above equation.

For  $k = 1, 2, \dots, 8$ , the Fourier coefficients  $a_k$  can be computed as

$$a_1 = -0.6366257003116296$$

$$a_2 = -5.070352857678721 \times 10^{-6} \approx 0$$

$$a_3 = -0.07074100153210318$$

$$a_4 = -5.070320092569666 \times 10^{-6} \approx 0$$

$$a_5 = -0.025470225589332522$$

$$a_6 = -5.070265333302604 \times 10^{-6} \approx 0$$

$$a_7 = -0.0012997664818977102$$

$$a_8 = -5.070188612604695 \times 10^{-6} \approx 0$$

Similarly,

$$b_k = \left( \frac{2}{T} \right) \int_0^{2\pi} f(t) \sin(kw_0 t) dt$$

$$b_k = \left( \frac{1}{\pi} \right) \times \left\{ \int_0^{\pi} t \sin(kt) dt + \int_{\pi}^{2\pi} \pi \sin(kt) dt \right\}$$

For  $k = 1, 2, \dots, 8$ , the Fourier coefficients  $b_k$  can be computed as

$$b_1 = -0.9999986528958207$$

$$b_2 = -0.4999993232285269$$

$$b_3 = -0.3333314439509194$$

$$b_4 = -0.24999804122384547$$

$$b_5 = -0.19999713794872364$$

$$b_6 = -0.1666635603759553$$

$$b_7 = -0.14285324664625462$$

$$b_8 = -0.12499577981019251$$

Any periodic function  $f(t)$ , such as the one shown in Figure 1 can be represented by the Fourier series as

$$f(t) = a_0 + \sum_{k=1}^{\infty} \{a_k \cos(kw_0 t) + b_k \sin(kw_0 t)\}$$

where  $a_0$ ,  $a_k$  and  $b_k$  have already been computed (for  $k = 1, 2, \dots, 8$ );

and  $w_0 = 2\pi f$

$$= \frac{2\pi}{T}$$

$$= \frac{2\pi}{2\pi}$$

$$= 1$$

Thus, for  $k = 1$ , one obtains

$$\bar{f}_1(t) \approx a_0 + a_1 \cos(t) + b_1 \sin(t)$$

For  $k = 1 \rightarrow 2$ , one obtains

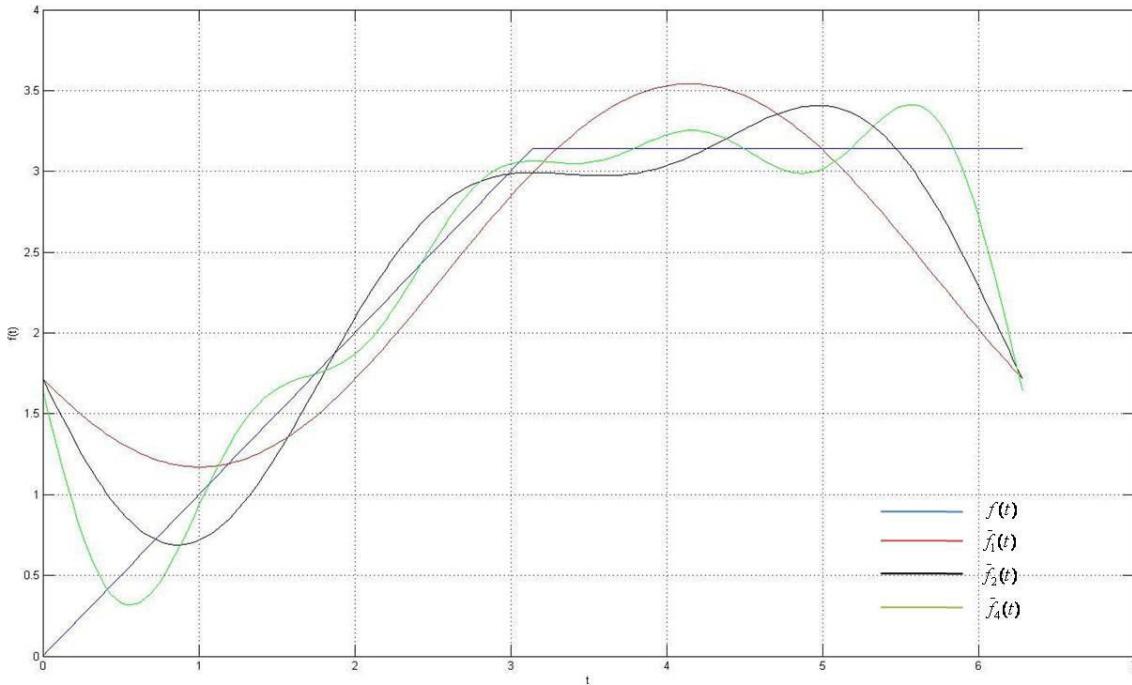
$$\bar{f}_2(t) \approx a_0 + a_1 \cos(t) + b_1 \sin(t) + a_2 \cos(2t) + b_2 \sin(2t)$$

For  $k = 1 \rightarrow 4$ , one obtains

$$\bar{f}_4(t) \approx a_0 + a_1 \cos(t) + b_1 \sin(t) + a_2 \cos(2t) + b_2 \sin(2t) + a_3 \cos(3t) + b_3 \sin(3t)$$

$$+ a_4 \cos(4t) + b_4 \sin(4t)$$

Plots for  $\bar{f}_1(t)$ ,  $\bar{f}_2(t)$  and  $\bar{f}_4(t)$  are shown in Figure 2.



**Figure 2** Fourier Approximated Functions (for Example 1).

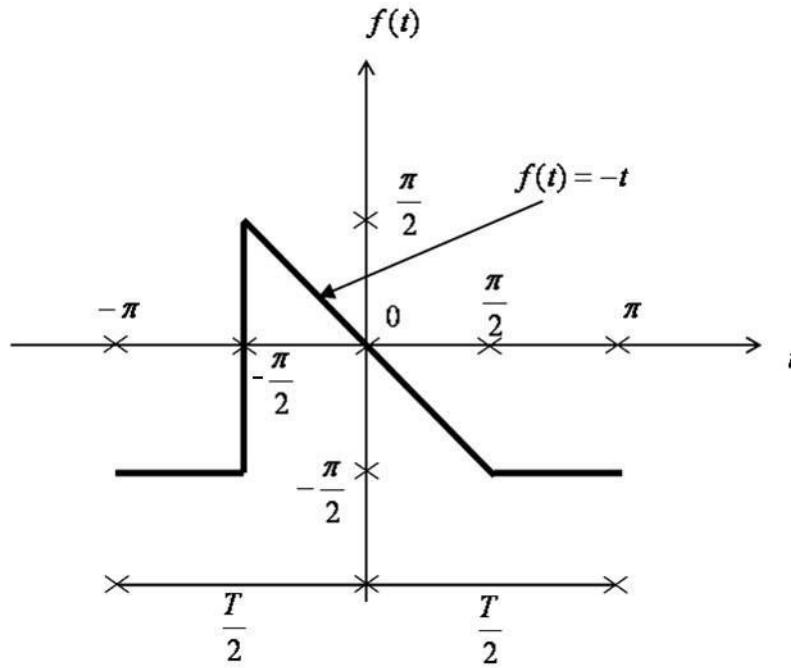
It can be observed from Figure 2 that as more terms are included in the Fourier series, the approximated Fourier functions are more closely resemble the original periodic function as shown in Figure 1.

### Example 2

The periodic triangular wave function  $f(t)$  is defined as

$$f(t) = \begin{cases} \frac{-\pi}{2} & \text{for } -\pi < t < \frac{-\pi}{2} \\ -t & \text{for } \frac{-\pi}{2} < t < \frac{\pi}{2} \\ \frac{\pi}{2} & \text{for } \frac{\pi}{2} < t < \pi \end{cases}$$

Find the Fourier coefficients  $a_0, a_1, \dots, a_8$  and  $b_1, \dots, b_8$  and approximate the periodic triangular wave function by the Fourier series.



**Figure 3** Periodic triangular wave function for Example 2.

### Solution

The unknown Fourier Coefficients  $a_0$ ,  $a_k$  and  $b_k$  can be computed based on Equations (2-4) as follows

$$a_0 = \left( \frac{1}{T} \right) \int_{-\pi}^{\pi} f(t) dt$$

$$a_0 = \frac{1}{(2\pi)} \times \left\{ \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \left( -\frac{\pi}{2} \right) dt + \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} (-t) dt + \int_{\frac{\pi}{2}}^{\pi} \left( -\frac{\pi}{2} \right) dt \right\}$$

$$a_0 = -0.78539753$$

$$a_k = \left( \frac{2}{T} \right) \int_{-\pi}^{\pi} f(t) \cos(kw_0 t) dt$$

where

$$w_0 = \frac{2\pi}{T}$$

$$= \frac{2\pi}{2\pi}$$

$$= 1$$

Hence,

$$a_k = \left( \frac{2}{T} \right) \int_{-\pi}^{\pi} f(t) \cos(kt) dt$$

or

$$a_k = \left( \frac{2}{2\pi} \right) \left\{ \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \left( -\frac{\pi}{2} \right) \cos(kt) dt + \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} (-t) \cos(kt) dt + \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \left( -\frac{\pi}{2} \right) \cos(kt) dt \right\}$$

Similarly,

$$b_k = \left( \frac{2}{T} \right) \int_{-\pi}^{\pi} f(t) \sin(kw_0 t) dt = \left( \frac{2}{T} \right) \int_{-\pi}^{\pi} f(t) \sin(kt) dt$$

or,

$$b_k = \left( \frac{2}{2\pi} \right) \left\{ \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \left( -\frac{\pi}{2} \right) \sin(kt) dt + \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} (-t) \sin(kt) dt + \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \left( -\frac{\pi}{2} \right) \sin(kt) dt \right\}$$

The “integration by part” formula can be utilized to compute the second integral on the right-hand-side of the above equations for  $a_k$  and  $b_k$ .

For  $k = 1, 2, \dots, 8$ , the Fourier coefficients  $a_k$  and  $b_k$  can be computed and summarized as following in Table 1

**Table 1** Fourier coefficients  $a_k$  and  $b_k$  for various  $k$  values.

$k$	$a_k$	$b_k$
1	0.999997	-0.63661936
2	0.00	-0.49999932
3	-0.33333355	0.07073466
4	0.00	0.2499980
5	0.1999968	-0.02546389
6	0.00	-0.16666356
7	-0.14285873	0.0126991327
8	0.00	0.12499578

The periodic function (shown in Example 1) can be approximated by Fourier series as

$$f(t) = a_0 + \sum_{k=1}^{\infty} \{a_k \cos(kt) + b_k \sin(kt)\}$$

Thus, for  $k = 1$ , one obtains:

$$\bar{f}_1(t) = a_0 + a_1 \cos(t) + b_1 \sin(t)$$

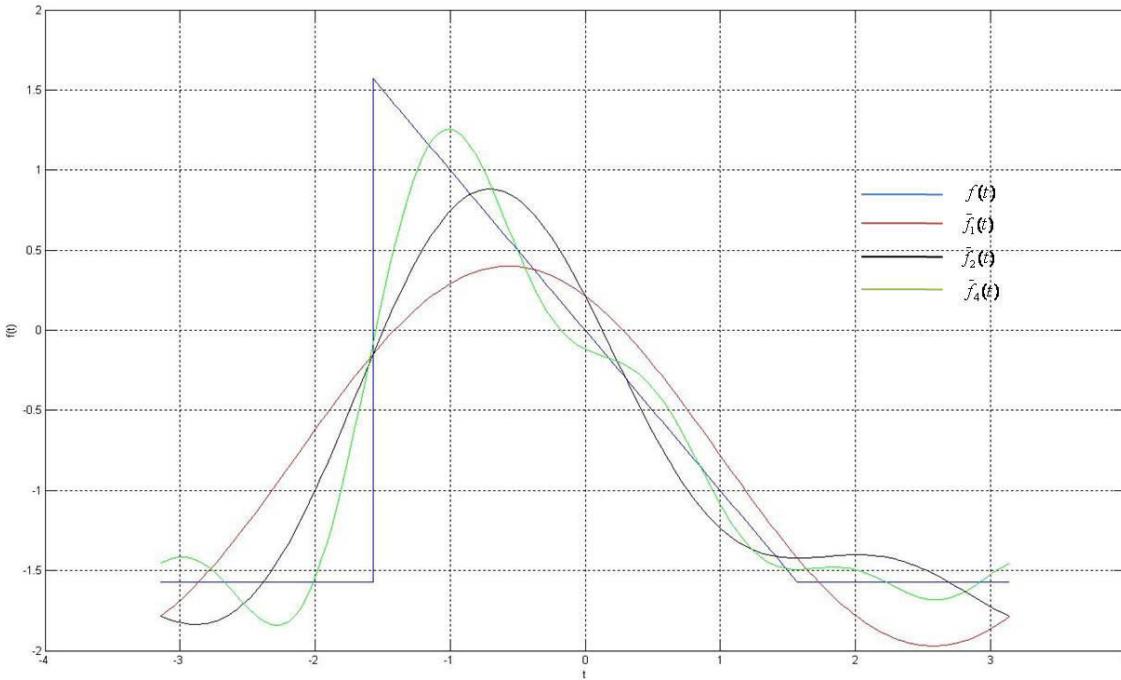
For  $k = 1 \rightarrow 2$ , one obtains:

$$\bar{f}_2(t) = a_0 + a_1 \cos(t) + b_1 \sin(t) + a_2 \cos(2t) + b_2 \sin(2t)$$

Similarly, for  $k = 1 \rightarrow 4$ , one has:

$$\begin{aligned} \bar{f}_4(t) &= a_0 + a_1 \cos(t) + b_1 \sin(t) + a_2 \cos(2t) + b_2 \sin(2t) + a_3 \cos(3t) + b_3 \sin(3t) \\ &\quad + a_4 \cos(4t) + b_4 \sin(4t) \end{aligned}$$

Plots for functions  $\bar{f}_1(t)$ ,  $\bar{f}_2(t)$  and  $\bar{f}_4(t)$  are shown in Figure 4.



**Figure 4** Fourier approximated functions for Example 2.

It can be observed from Figure 4 that as more terms are included in the Fourier series, the approximated Fourier functions closely resemble the original periodic function.

### Complex Form of the Fourier Series

Using Euler's identity,  $e^{ix} = \cos(x) + i \sin(x)$ , and  $e^{-ix} = \cos(x) - i \sin(x)$ , the sine and cosine can be expressed in the exponential form as

$$\sin(x) = \frac{e^{ix} - e^{-ix}}{2i} = \text{"odd" function, since } \sin(x) = -\sin(-x) \quad (10)$$

$$\cos(x) = \frac{e^{ix} + e^{-ix}}{2} = \text{"even" function, since } \cos(x) = \cos(-x) \quad (11)$$

Thus, the Fourier series (expressed in Equation 1) can be converted into the following form

$$f(t) = a_0 + \sum_{k=1}^{\infty} a_k \left( \frac{e^{ikw_0 t} + e^{-ikw_0 t}}{2} \right) + b_k \left( \frac{e^{ikw_0 t} - e^{-ikw_0 t}}{2i} \right) \quad (12)$$

or

$$f(t) = a_0 + \sum_{k=1}^{\infty} e^{ikw_0 t} \left( \frac{a_k}{2} + \frac{b_k}{2i} * \frac{i}{i} \right) + e^{-ikw_0 t} \left( \frac{a_k}{2} - \frac{b_k}{2i} * \frac{i}{i} \right)$$

or, since  $i^2 = -1$ , one obtains

$$f(t) = a_0 + \sum_{k=1}^{\infty} e^{ikw_0 t} \left( \frac{a_k - ib_k}{2} \right) + e^{-ikw_0 t} \left( \frac{a_k + ib_k}{2} \right) \quad (13)$$

Define the following constants

$$\tilde{C}_0 \equiv a_0 \quad (14)$$

$$\tilde{C}_k \equiv \frac{a_k - ib_k}{2} \quad (15)$$

Hence:

$$\tilde{C}_{-k} \equiv \frac{a_{-k} - ib_{-k}}{2} \quad (16)$$

Using the even and odd properties shown in Equations (3) and (4) respectively, Equation (16) becomes

$$\tilde{C}_{-k} \equiv \frac{a_k + ib_k}{2} \quad (17)$$

Substituting Equations (14), (15), (17) into Equation (13), one gets

$$\begin{aligned} f(t) &= \tilde{C}_0 + \sum_{k=1}^{\infty} \tilde{C}_k e^{ikw_0 t} + \sum_{k=1}^{\infty} \tilde{C}_{-k} e^{-ikw_0 t} \\ &= \sum_{k=0}^{\infty} \tilde{C}_k e^{ikw_0 t} + \sum_{k=-1}^{-\infty} \tilde{C}_k e^{ikw_0 t} \\ &= \sum_{k=0}^{\infty} \tilde{C}_k e^{ikw_0 t} + \sum_{k=-\infty}^{-1} \tilde{C}_k e^{ikw_0 t} \\ &= \sum_{k=-\infty}^{\infty} \tilde{C}_k e^{ikw_0 t} \end{aligned} \quad (18)$$

The coefficient  $\tilde{C}_k$  can be computed, by substituting Equations (3) and (4) into Equation (15) to obtain

$$\begin{aligned} \tilde{C}_k &= \left( \frac{1}{2} \right) \left( \frac{2}{T} \right) \left\{ \int_0^T f(t) \cos(kw_0 t) dt - i \int_0^T f(t) \sin(kw_0 t) dt \right\} \\ &= \left( \frac{1}{T} \right) \left\{ \int_0^T f(t) \times [\cos(kw_0 t) - i \sin(kw_0 t)] dt \right\} \end{aligned} \quad (19)$$

Substituting Equations (10, 11) into the above equation, one gets

$$\begin{aligned} \tilde{C}_k &= \left( \frac{1}{T} \right) \left\{ \int_0^T f(t) \times \left[ \frac{e^{ikw_0 t} + e^{-ikw_0 t}}{2} - i \times \frac{e^{ikw_0 t} - e^{-ikw_0 t}}{2i} \right] dt \right\} \\ &= \left( \frac{1}{T} \right) \left\{ \int_0^T f(t) \times e^{-ikw_0 t} dt \right\} \end{aligned} \quad (20)$$

Thus, Equations (18) and (20) are the equivalent complex version of Equations (1)-(4).

## References

- [1] E.Oran Brigham, The Fast Fourier Transform, Prentice-Hall, Inc. (1974).
- [2] S.C. Chapra, and R.P. Canale, Numerical Methods for Engineers, 4<sup>th</sup> Edition, Mc-Graw Hill (2002).
- [3] W.H . Press, B.P. Flannery, S.A. Tenkolsky, and W.T. Vetterling, Numerical Recipies, Cambridge University Press (1989), Chapter 12.

[4] M.T. Heath, Scientific Computing, Mc-Graw Hill (1997).

[5] H. Joseph Weaver, Applications of Discrete and Continuous Fourier Analysis, John Wiley & Sons, Inc. (1983).

---

#### FAST FOURIER TRANSFORM

---

Topic	Continuous Fourier Series
Summary	Textbook notes on continuous Fourier series
Major	General Engineering
Authors	Duc Nguyen
Date	July 25, 2010

---

Web Site <http://numericalmethods.eng.usf.edu>

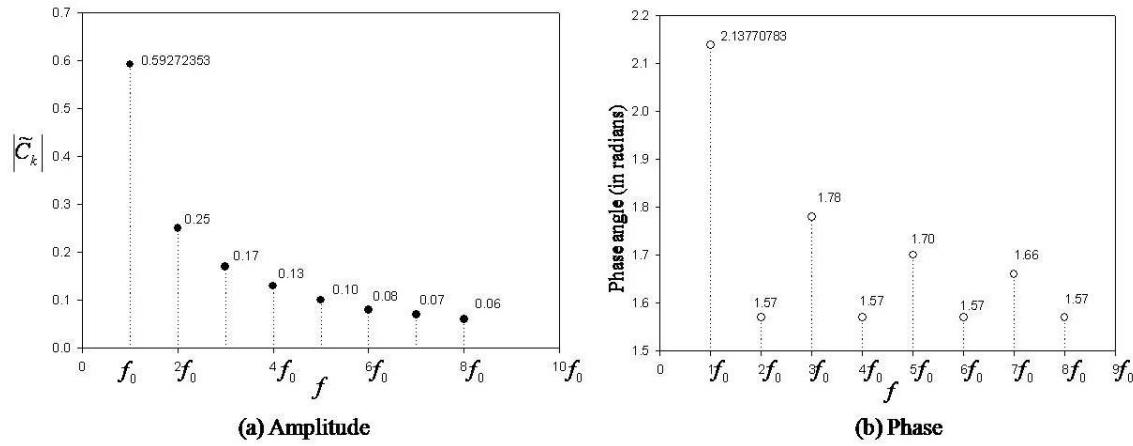
---

## Chapter 11.03

### Fourier Transform Pair: Frequency and Time Domain

#### Introduction

In Chapter 11.02, Fourier approximations were expressed in the time domain. The amplitude (vertical axis) of a given periodic function can be plotted versus time (horizontal axis), but it can also be plotted in the frequency domain [1-6] as shown in Figure 1.



**Figure 1** Periodic Function (see Example 1 in Chapter 11.02) In Frequency Domain.

The advantages of plotting the amplitude of a given periodic function in frequency domain (instead of time domain) are due to the following reasons:

For a specific value “ $k$ ” (say  $k = 2$ ) of the Fourier series in the time domain, one has to plot the entire curve to observe the amplitude of a given periodic function (recall  $\bar{f}_2(t) = a_0 + a_1 \cos(t) + b_1 \sin(t) + a_2 \cos(2t) + b_2 \sin(2t)$ , see Example 1 in Chapter 11.02). However, in the frequency domain, the amplitude can be plotted as a single point. (see Figure 1a).

In the frequency domain, one can easily identify which frequency (or corresponding to which value of “ $k$ ”) contributes the most to the amplitude [see Figure 1(a)], where such information is not readily available if time domain is used.

From the amplitude plot in frequency domain [see Figure 1(a)], one can easily identify that contributions to the amplitude beyond the 8th frequency ( $k > 8$ ) are not significant any more.

In real-life structural dynamics problems, such as the dynamical (time-dependent) response of a (building) structure subjected to oscillated loads (for example, the operational machines attached to the structures), the displacement superposition method is often used to predict the (time dependent) displacement response of the structure. This method basically transforms the original (large, coupled) equation of motion into a reduced (much smaller size, uncoupled) equation of motion by making use of the few free vibration mode shapes and its associated frequencies. Knowledge of which frequencies (and the corresponding mode shapes) that have the most contribution to the predicted dynamical response (such as nodal displacement response) plays crucial roles for the algorithms' efficiencies.

Detailed explanations on how to obtain Figures 1(a), and 1(b) are now presented in the following sections.

### Explanation of Figure 1(a) and 1(b)

One starts with Equation (18) and (20) of Chapter 11.02

$$f(t) = \sum_{k=-\infty}^{\infty} \tilde{C}_k e^{ikw_0 t}$$

where

$$\tilde{C}_k = \left( \frac{1}{T} \right) \left\{ \int_0^T f(t) \times e^{-ikw_0 t} dt \right\}$$

For the periodic function shown in Example 1 of Chapter 11.02 (or Figure 1 of Chapter 11.02), one has

$$\begin{aligned} w_0 &= 2\pi f \\ &= \frac{2\pi}{T} \\ &= \frac{2\pi}{2\pi} \\ &= 1 \end{aligned}$$

$$\tilde{C}_k = \left( \frac{1}{T} \right) \left\{ \int_0^\pi t \times e^{-ikt} dt + \int_\pi^{2\pi} \pi \times e^{-ikt} dt \right\}$$

Define, and using “integration by parts” formula

$$\begin{aligned} A &\equiv \int_0^\pi t \times e^{-ikt} dt = \left[ t \times \left( \frac{-1}{ik} \right) e^{-ikt} \right]_0^\pi + \int_0^\pi \left( \frac{1}{ik} \right) e^{-ikt} dt \\ A &= \left[ \left( \frac{-\pi}{ik} \right) e^{-ik\pi} \right]_0^\pi + \left( \frac{1}{ik} \right) \left[ \left( -\frac{1}{ik} \right) e^{-ikt} \right]_0^\pi \\ &= \left[ \left( \frac{-\pi}{ik} \right) e^{-ik\pi} \right] + \left( \frac{1}{k^2} \right) [e^{-ik\pi} - 1] \end{aligned}$$

$$\begin{aligned}
&= \left[ \left( \frac{\pi i}{k} \right) e^{-ik\pi} + \left( \frac{1}{k^2} \right) e^{-ik\pi} - \frac{1}{k^2} \right] \\
B &\equiv \pi \int_{-\pi}^{\pi} e^{-ikt} dt = \left[ \left( e^{-ikt} \right) \left( \frac{-\pi}{ik} \right) \right]_{-\pi}^{\pi} \\
&= \left( \frac{-\pi}{ik} \right) [e^{-ik2\pi} - e^{-ik\pi}] \\
&= \left( \frac{\pi i}{k} \right) [e^{-ik2\pi} - e^{-ik\pi}]
\end{aligned}$$

Thus,

$$\begin{aligned}
\tilde{C}_k &= \left( \frac{1}{2\pi} \right) \{A + B\} \\
&= \left( \frac{1}{2\pi} \right) \left\{ e^{-ik\pi} \left( \frac{\pi i}{k} + \frac{1}{k^2} - \frac{\pi i}{k} \right) - \frac{1}{k^2} + \left( \frac{\pi i}{k} \right) e^{-ik2\pi} \right\}
\end{aligned}$$

Using the following Euler identities

$$\begin{aligned}
e^{-ik\pi} &= \cos(-k\pi) + i \sin(-k\pi) \\
&= \cos(k\pi) - i \sin(k\pi) \\
&= \cos(k\pi) \\
e^{-ik(2\pi)} &= \cos(k(2\pi)) - i \sin(k(2\pi)) \\
&= \cos(k(2\pi))
\end{aligned}$$

Hence, one obtains (noting that  $\cos(k2\pi) = 1$ , for any integer  $k$ ):

$$\tilde{C}_k = \left( \frac{1}{2\pi} \right) \left\{ \cos(k\pi) \times \left( \frac{1}{k^2} \right) - \frac{1}{k^2} + \left( \frac{\pi i}{k} \right) \cos(k2\pi) \right\}$$

or,

$$\tilde{C}_k = \left( \frac{1}{2\pi} \right) \left\{ \left( \frac{1}{k^2} \right) \cos(k\pi) - \frac{1}{k^2} + \left( \frac{\pi i}{k} \right) \right\}$$

Also, since:

$$\cos(k\pi) = \begin{cases} -1 & \text{for } k = \text{odd integer } (= 1, 3, 5, 7, \dots) \\ +1 & \text{for } k = \text{even integer } (= 2, 4, 6, 8, \dots) \end{cases}$$

Hence:

$$\cos(k\pi) = (-1)^k$$

Thus,

$$\begin{aligned}
\tilde{C}_k &= \left( \frac{1}{2\pi} \right) \left\{ \frac{(-1)^k}{k^2} - \frac{1}{k^2} + \frac{\pi i}{k} \right\} \\
&= \left( \frac{1}{2\pi k^2} \right) [(-1)^k - 1] + \left( \frac{1}{2k} \right) i
\end{aligned}$$

From Equation (15) in Chapter 11.02, one has:

$$\tilde{C}_k = \frac{a_k - ib_k}{2}$$

Hence upon comparing the above 2 equations, one concludes

$$a_k \equiv \left( \frac{1}{\pi k^2} \right) [(-1)^k - 1]$$

$$b_k = \left( \frac{-1}{k} \right)$$

Remarks:

For  $k=1,2,3,4,\dots,8$ ; the values for  $a_k$  and  $b_k$  (based on the above 2 formulas) are exactly identical as the ones presented earlier in Example 1 in Chapter 11.02.

Thus

$$\tilde{C}_1 = \frac{a_1 - ib_1}{2}$$

$$= \frac{\frac{-2}{\pi} - i(-1)}{2}$$

$$= \frac{-1}{\pi} + \frac{1}{2}i$$

$$\tilde{C}_2 = \frac{a_2 - ib_2}{2}$$

$$= \frac{0 - i\left(-\frac{1}{2}\right)}{2}$$

$$= 0 + \frac{1}{4}i$$

$$\tilde{C}_3 = \frac{a_3 - ib_3}{2}$$

$$= \frac{\left(\frac{-2}{9\pi}\right) - i\left(\frac{-1}{3}\right)}{2}$$

$$= \left(\frac{-1}{9\pi}\right) + \frac{1}{6}i$$

$$\tilde{C}_4 = \frac{a_4 - ib_4}{2}$$

$$= \frac{0 - i\left(\frac{-1}{4}\right)}{2}$$

$$= 0 + \frac{1}{8}i$$

$$\tilde{C}_5 = \frac{a_5 - ib_5}{2}$$

$$\begin{aligned}
&= \frac{\left(\frac{-2}{25\pi}\right) - i\left(\frac{-1}{5}\right)}{2} \\
&= \left(\frac{-1}{25\pi}\right) + \frac{1}{10}i \\
\tilde{C}_6 &= \frac{a_6 - ib_6}{2} \\
&= \frac{0 - i\left(\frac{-1}{6}\right)}{2} \\
&= 0 + \frac{1}{12}i \\
\tilde{C}_7 &= \frac{a_7 - ib_7}{2} \\
&= \frac{\left(\frac{-2}{49\pi}\right) - i\left(\frac{-1}{7}\right)}{2} \\
&= \left(\frac{-1}{49\pi}\right) + \frac{1}{14}i \\
\tilde{C}_8 &= \frac{a_8 - ib_8}{2} \\
&= \frac{0 - i\left(\frac{-1}{8}\right)}{2} \\
&= 0 + \frac{1}{16}i
\end{aligned}$$

In general, one has

$$\tilde{C}_k = \begin{cases} \frac{-1}{k^2\pi} + \left(\frac{1}{2k}\right)i & \text{for } k = 1, 3, 5, 7, \dots = \text{odd integer} \\ \left(\frac{1}{2k}\right)i & \text{for } k = 2, 4, 6, 8, \dots = \text{even integer} \end{cases}$$

### Representation of a complex number in polar coordinates

In Cartesian (rectangular) coordinates, a complex number  $\tilde{C}_k$  can be expressed as:

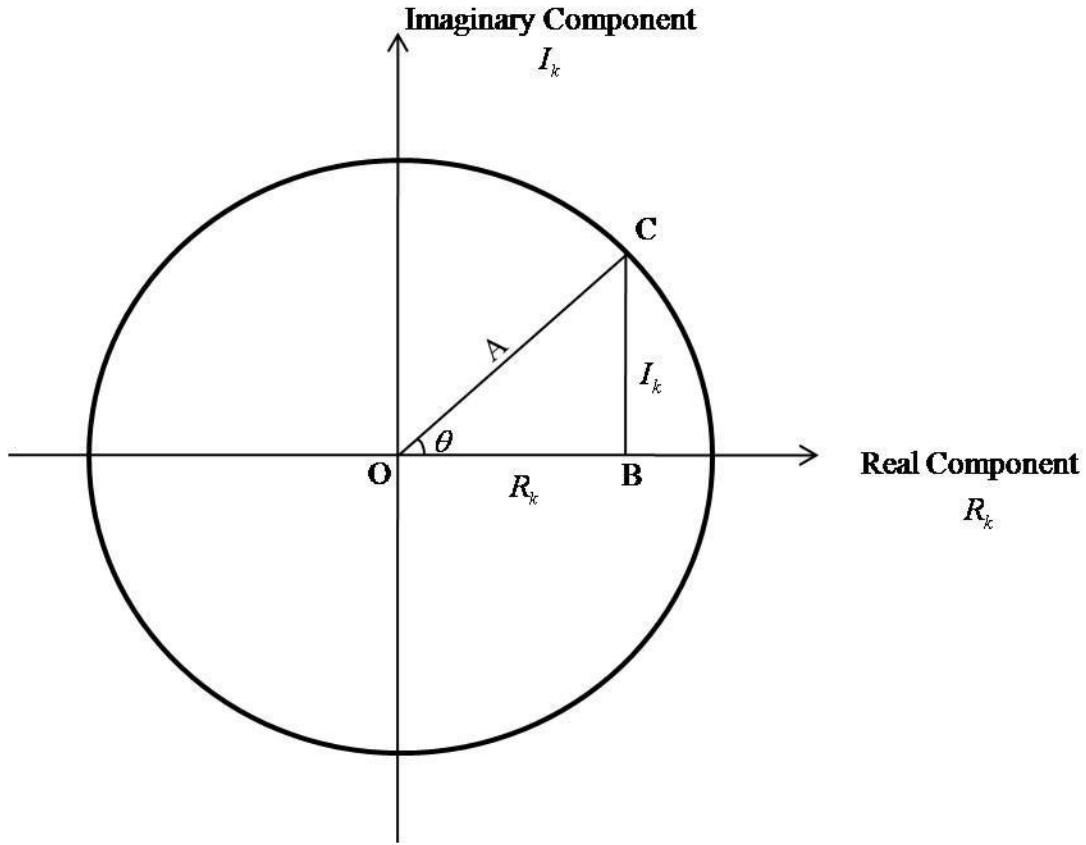
$$\tilde{C}_k = R_k + (I_k)i$$

where  $R_k$  and  $I_k$  represents the real and imaginary components of  $\tilde{C}_k$ , respectively.

In polar coordinates, a complex number  $\tilde{C}_k$  can be expressed as:

$$\tilde{C}_k = Ae^{i\theta} = A\{\cos(\theta) + i \sin(\theta)\} = \{A \cos(\theta)\} + \{A \sin(\theta)\}i$$

where  $A$  and  $\theta$  represents the amplitude and phase angle of  $\tilde{C}_k$ , respectively (see Figure 2).



**Figure 2** Representation of a complex number in polar coordinates

Thus, one obtains the following relations between the cartesian and polar coordinate systems:

$$R_k = A \cos(\theta)$$

$$I_k = A \sin(\theta)$$

Hence:

$$R_k^2 + I_k^2 = A^2 \cos^2(\theta) + A^2 \sin^2(\theta) = A^2 [\cos^2(\theta) + \sin^2(\theta)]$$

$$A^2 = R_k^2 + I_k^2$$

$$A = \sqrt{R_k^2 + I_k^2}$$

$$\cos(\theta) = \frac{R_k}{A} \text{ implies } \theta = \cos^{-1}\left(\frac{R_k}{A}\right)$$

$$\sin(\theta) = \frac{I_k}{A} \text{ implies } \theta = \sin^{-1}\left(\frac{I_k}{A}\right)$$

Based on the above 3 formulas, the complex numbers  $\tilde{C}_k$ , for  $k=1,2,3,\dots,8$  can be expressed as

$$\begin{aligned}\tilde{C}_1 &= \frac{-1}{\pi} + \left(\frac{1}{2}\right)i \\ &= (0.59272353)e^{i(2.13770783)}\end{aligned}$$

Hence, the amplitude  $A$  and Phase angle  $\theta$  for  $\tilde{C}_1$  are 0.59272353, and 2.13770783 radians, respectively. The readers should refer to Figures 1(a) and 1(b) to confirm the plotted values.

### Important Notes

If one uses the formula

$$\begin{aligned}\theta &= \cos^{-1}\left(\frac{R_k}{A}\right) \\ &= \cos^{-1}\left(\frac{\frac{-1}{\pi}}{0.59272353}\right) \\ &= 2.13770783 \text{ radians} \\ &= 122.48^\circ\end{aligned}$$

However, the other formula for  $\theta$  gives:

$$\begin{aligned}\theta &= \sin^{-1}\left(\frac{I_k}{A}\right) \\ &= \sin^{-1}\left(\frac{0.5}{0.59272353}\right) \\ &= 1.0038848 \text{ radians} \\ &= 57.52^\circ\end{aligned}$$

Since  $R_k$  is negative, and  $I_k$  is positive, the angle  $\theta$  must be in the 2nd (or upper left) quadrant of a circle (or  $90^\circ \leq \theta \leq 180^\circ$ ). Thus, the correct value for  $\theta$  should be 2.13770783 radians (or  $122.48^\circ$ ) and the other value for  $\theta = 1.0038848$  radians must be discarded.

Similarly, one obtains

$$\begin{aligned}\tilde{C}_2 &= 0 + \frac{1}{4}i \\ &= (0.25)e^{i\left(\frac{\pi}{2}\right)} \\ &= (0.25)e^{i(1.57079633)}\end{aligned}$$

$$\begin{aligned}
\tilde{C}_3 &= \left( \frac{-1}{9\pi} \right) + \frac{1}{6}i \\
&= (0.17037798)e^{i(1.77990097)} \\
\tilde{C}_4 &= 0 + \frac{1}{8}i \\
&= (0.125)e^{i\left(\frac{\pi}{2}\right)} \\
&= (0.125)e^{i(1.57079633)} \\
\tilde{C}_5 &= \left( \frac{-1}{25\pi} \right) + \frac{1}{10}i \\
&= (0.100807311)e^{i(1.69743886)} \\
\tilde{C}_6 &= 0 + \frac{1}{12}i \\
&= (0.08333333)e^{i\left(\frac{\pi}{2}\right)} \\
&= (0.08333333)e^{i(1.57079633)} \\
\tilde{C}_7 &= \left( \frac{-1}{49\pi} \right) + \frac{1}{14}i \\
&= (0.07172336)e^{i(1.66149251)} \\
\tilde{C}_8 &= 0 + \frac{1}{16}i \\
&= (0.0625)e^{i\left(\frac{\pi}{2}\right)}
\end{aligned}$$

In summary, the given periodic function (shown in Example 1 of Chapter 11.02) can also be expressed in complex number formats, in polar coordinate with the amplitudes and phase angles given in the following table (also refer to Figures 1(a), and 1(b)).

**Table 1** Amplitude and phase angle (in radians) for varying  $k$  values.

$k$	Amplitude	Phase Angle (radians)
1	0.59272353	2.13770783
2	0.25	$\frac{\pi}{2} = 1.57079633$
3	0.14037798	1.77990097
4	0.125	$\frac{\pi}{2}$
5	0.100807311	1.69743886
6	0.08333333	$\frac{\pi}{2}$
7	0.07172336	1.66149251
8	0.0625	$\frac{\pi}{2}$

### Non-Periodic Function

Recall that a periodic function can be expressed in terms of the exponential form, accordingly to Equations (18, 20) of Chapter 11.02 as

$$f(t) = \sum_{k=-\infty}^{\infty} \tilde{C}_k e^{ikw_0 t}$$

$$\tilde{C}_k = \left( \frac{1}{T} \right) \times \left\{ \int_0^T f(t) \times e^{-ikw_0 t} dt \right\}$$

Define the following function

$$\hat{F}(ikw_0) = \int_{-\frac{T}{2}}^{\frac{T}{2}} f(t) e^{-ikw_0 t} dt \quad (1)$$

where  $\hat{F}(ikw_0)$  is a function of  $i, k$ , and  $w_0$

Then, Equation (20) of Chapter 11.02 can be written as

$$\tilde{C}_k = \left( \frac{1}{T} \right) \times \hat{F}(ikw_0) \quad (2)$$

and Equation (18) of Chapter 11.02 becomes

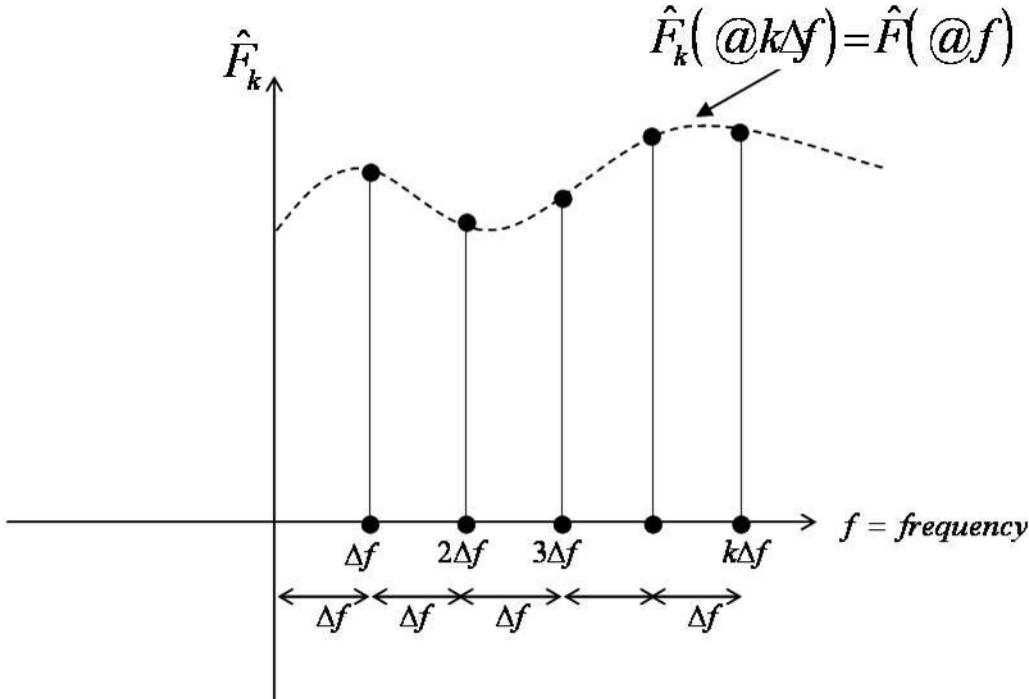
$$f(t) = \sum_{k=-\infty}^{\infty} \left( \frac{1}{T} \right) \times \hat{F}(ikw_0) e^{ikw_0 t} \quad (3)$$

A non-periodic function  $f_{np}$  can be considered as a periodic function, with the period

$$T \rightarrow \infty, \text{ or } \Delta f \equiv \frac{1}{T} \rightarrow 0 \text{ (see Figure 3)}$$

From Equations (6) and (7) from Chapter 11.01, one gets

$$\begin{aligned} w_0 &= 2\pi f \\ &= \frac{2\pi}{T} \\ &= 2\pi(\Delta f) \end{aligned} \quad (4)$$



**Figure 3** Discretization of frequency data.

From Equation (3), one obtains

$$\begin{aligned} f_{np}(t) &= \lim_{\substack{T \rightarrow \infty \\ \text{or } \Delta f \rightarrow 0}} f(t) \\ &= \lim_{\Delta f \rightarrow 0} \sum_{k=-\infty}^{\infty} (\Delta f) \times \hat{F}(ikw_0) e^{ikw_0 t} \end{aligned} \quad (5)$$

In the above equation, the subscript "np" denotes non-periodic function.

$$f_{np}(t) = \lim_{\Delta f \rightarrow 0} \sum_{k=-\infty}^{\infty} (\Delta f) \times \hat{F}(ik2\pi\Delta f) e^{ik2\pi\Delta f t} \quad (6)$$

Realizing that  $k\Delta f = f$  (See Figure 3), the above equation becomes

$$\begin{aligned} f_{np}(t) &= \int df \times \hat{F}(i2\pi f) e^{i2\pi f t} \\ f_{np}(t) &= \int \hat{F}(i2\pi f) e^{i2\pi f t} df \end{aligned} \quad (7)$$

Multiplying and dividing the right-hand-side of the equation by  $2\pi$ , one obtains

$$\begin{aligned} f_{np}(t) &= \left( \frac{1}{2\pi} \right) \int \hat{F}(i2\pi f) e^{i2\pi f t} d(2\pi f) \\ &= \left( \frac{1}{2\pi} \right) \int_{-\infty}^{\infty} \hat{F}(iw_0) e^{iw_0 t} d(w_0); \text{ inverse Fourier transform} \end{aligned} \quad (8)$$

Using the definition stated in Equation (1), one has

$$\hat{F}(iw_0) = \int_{-\infty}^{\infty} f_{np}(t) e^{-iw_0 t} d(t); \text{ Fourier transform} \quad (9)$$

Thus, Equations (9) and (8) will transform a non-periodic function from time domain to frequency domain, and from frequency domain to time domain, respectively.

## References

- [1] E.Oran Brigham, The Fast Fourier Transform, Prentice-Hall, Inc. (1974).
- [2] S.C. Chapra, and R.P. Canale, Numerical Methods for Engineers, 4<sup>th</sup> Edition, Mc-Graw Hill (2002).
- [3] W.H . Press, B.P. Flannery, S.A. Tenkolsky, and W.T. Vetterling, Numerical Recipies, Cambridge University Press (1989), Chapter 12.
- [4] M.T. Heath, Scientific Computing, Mc-Graw Hill (1997).
- [5] H. Joseph Weaver, Applications of Discrete and Continuous Fourier Analysis, John Wiley & Sons, Inc. (1983).
- [6] Larry N. Thibos, Fourier Analysis for Beginners, Email: thibos@indiana.edu (1993, 2000, 2003).

---

### FAST FOURIER TRANSFORM

---

Topic	Fourier Transform Pair: Frequency and Time Domain
Summary	Textbook notes on Fourier Transform Pair: Frequency and Time Domain
Major	General Engineering
Authors	Duc Nguyen
Date	July 25, 2010
Web Site	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

---

## Chapter 11.04

### Discrete Fourier Transform

#### Introduction

Recalled the exponential form of Fourier series (see Equations 18 and 20 from Chapter 11.02),

$$f(t) = \sum_{k=-\infty}^{\infty} \tilde{C}_k e^{ikw_0 t} \quad (18, \text{Ch. 11.02})$$

$$\tilde{C}_k = \left( \frac{1}{T} \right) \left\{ \int_0^T f(t) \times e^{-ikw_0 t} dt \right\} \quad (20, \text{Ch. 11.02})$$

While the above integral can be used to compute  $\tilde{C}_k$ , it is more preferable to have a discretized formula version to compute  $\tilde{C}_k$ . Furthermore, the Discrete Fourier Transform (or DFT) [1–5] will also facilitate the development of much more efficient algorithms for Fast Fourier Transform (or FFT), to be discussed in Chapters 11.05 and 11.06.

#### Derivations of DFT Formulas

If time “ $t$ ” is discretized at  $t_1 = \Delta t, t_2 = 2\Delta t, t_3 = 3\Delta t, \dots, t_n = n\Delta t$ ,

Then Equation (18, of Chapter 11.02) becomes

$$f(t_n) = \sum_{k=0}^{N-1} \tilde{C}_k e^{ikw_0 t_n} \quad (1)$$

To simplify the notation, define

$$t_n = n \quad (2)$$

Then, Equations (1) can be written as

$$f(n) = \sum_{k=0}^{N-1} \tilde{C}_k e^{ikw_0 n} \quad (3)$$

In the above formula, “ $n$ ” is an integer counter. However,  $f(n)$  and  $t_n$  do NOT have to be integer numbers.

Multiplying both sides of Equation (3) by  $e^{-ilw_0 n}$ , and performing the summation on “ $n$ ”, one obtains ( note:  $l$  = integer number)

$$\sum_{n=0}^{N-1} f(n) \times e^{-ilw_0 n} = \sum_{n=0}^{N-1} \sum_{k=0}^{N-1} \tilde{C}_k e^{ikw_0 n} \times e^{-ilw_0 n} \quad (4)$$

$$= \sum_{n=0}^{N-1} \sum_{k=0}^{N-1} \tilde{C}_k e^{i(k-l)w_0 n} \quad (5)$$

$$= \sum_{n=0}^{N-1} \sum_{k=0}^{N-1} \tilde{C}_k e^{i(k-l)\frac{2\pi}{N} n} \quad (6)$$

Switching the order of summations on the right-hand-side of Equation (6), one obtains

$$\sum_{n=0}^{N-1} f(n) \times e^{-il\left(\frac{2\pi}{N}\right)n} = \sum_{k=0}^{N-1} \tilde{C}_k \sum_{n=0}^{N-1} e^{i(k-l)\left(\frac{2\pi}{N}\right)n} \quad (7)$$

Define

$$A = \sum_{n=0}^{N-1} e^{i(k-l)\left(\frac{2\pi}{N}\right)n} \quad (8)$$

There are 2 possibilities for  $(k - l)$  to be considered in Equation (8)

**Case(1):**  $(k - l)$  is a multiple integer of  $N$ , such as

$$(k - l) = mN; \text{ or } k = l + mN \text{ where } m = 0, \pm 1, \pm 2, \dots$$

Thus, Equation (8) becomes:

$$\begin{aligned} A &= \sum_{n=0}^{N-1} e^{im2\pi n} \\ &= \sum_{n=0}^{N-1} \cos(mn2\pi) + i \sin(mn2\pi) \end{aligned} \quad (9)$$

Hence:

$$A = N \quad (10)$$

**Case(2):**  $(k - l)$  is NOT a multiple integer of  $N$

In this case, from Equation (8) one has

$$A = \sum_{n=0}^{N-1} \left\{ e^{i(k-l)\left(\frac{2\pi}{N}\right)n} \right\}^n \quad (11)$$

Define:

$$\begin{aligned} a &= e^{i(k-l)\frac{2\pi}{N}} \\ &= \cos\left((k-l)\frac{2\pi}{N}\right) + i \sin\left((k-l)\frac{2\pi}{N}\right) \end{aligned} \quad (12)$$

$$a \neq 1; \text{ because } (k - l) \text{ is "NOT" a multiple integer of } N \quad (13)$$

Then, Equation (11) can be expressed as

$$A = \sum_{n=0}^{N-1} a^n \quad (14)$$

From mathematical handbooks, the right side of Equation (14) represents the "geometric series", and can be expressed as

$$A = \sum_{n=0}^{N-1} a^n = N \text{ if } a = 1 \quad (15)$$

$$= \frac{1-a^N}{1-a} \text{ if } a \neq 1 \quad (16)$$

Because of Equation (13), hence Equation (16) should be used to compute  $A$ . Thus

$$\begin{aligned} A &= \frac{1-a^N}{1-a} \text{ (See Equation (12))} \\ &= \frac{1-e^{i(k-l)2\pi}}{1-a} \end{aligned} \quad (17)$$

Since  $(k-l)$  is still a multiple of  $2\pi$ , hence

$$\begin{aligned} e^{i(k-l)2\pi} &\equiv \cos\{(k-l)2\pi\} + i \sin\{(k-l)2\pi\} \\ &= 1 \end{aligned} \quad (18)$$

Substituting Equation (17) into Equation (18), one gets

$$A = 0 \quad (19)$$

Thus, combining the results of case (1) and case (2), one gets (see Equations (10) and Equation (19))

$$A = N + 0 \quad (20)$$

Substituting Equation (20) into Equation (8), and then referring to Equation (7), one gets

$$\sum_{n=0}^{N-1} f(n)e^{-ilw_0n} = \sum_{k=0}^{N-1} \tilde{C}_k \times N \quad (20a)$$

Recalled  $k = l + mN$  (where  $l, m$  are integer numbers), and since  $k$  must be in the range  $0 \rightarrow N-1$ , therefore  $m = 0$ . Thus:

$$k = l + mN \text{ becomes } k = l$$

Equation (20a) can, therefore, be simplified to

$$\sum_{n=0}^{N-1} f(n)e^{-ilw_0n} = \tilde{C}_l \times N \quad (20b)$$

Thus

$$\begin{aligned} \tilde{C}_l = \tilde{C}_k &= \left( \frac{1}{N} \right) \sum_{n=0}^{N-1} f(n)e^{-ikw_0n} \\ &= \left( \frac{1}{N} \right) \sum_{n=0}^{N-1} f(n) \{ \cos(kw_0n) - i \sin(kw_0n) \} \end{aligned} \quad (21)$$

where

$$n \equiv t_n$$

and

$$\begin{aligned} f(n) &= \sum_{k=0}^{N-1} \tilde{C}_k e^{ikw_0n} \\ &= \sum_{k=0}^{N-1} \tilde{C}_k \{ \cos(kw_0n) + i \sin(kw_0n) \} \end{aligned} \quad (3, \text{ repeated})$$

Remarks:

(a) Consider the exponential term in Equation (1). Let

$$E = e^{(ikw_0n)} = e^{(ik \times \frac{2\pi}{N} * n)}$$

If one replaces “ $n$ ” by “ $-(N - n)$ ” (or “ $(n - N)$ ”) into the above equation, then one obtains

$$\begin{aligned} e^{ik \times \frac{2\pi}{N} * (n - N)} &= e^{(ik \times \frac{2\pi}{N} * n)} \times [e^{(-ik \times 2\pi)} = 1] \\ &= E \end{aligned}$$

Thus, Equation (1) indicates that the force corresponding to frequencies of order “ $n$ ” and “ $-(N - n) = n - N$ ” have the same values. Hence

$$\begin{aligned} w_n &= n\bar{w} \text{ for } n \leq \frac{N}{2} \\ &= -(N - n)\bar{w} \text{ for } n > \frac{N}{2} \end{aligned}$$

and the frequency corresponding to  $n = \frac{N}{2}$  is the highest frequency that can be considered in the discrete Fourier series ( $w_{\frac{N}{2}}$  is called the Nyquist frequency). If there are harmonic (force) components above  $w_{\frac{N}{2}}$  in the original function, then these higher components will introduce distortions in the lower harmonic components (known as ALIASING phenomenon). Because of the ALIASING phenomenon, the number of ( $N$ ) data points should be “at least twice” the highest harmonic component presents in the (forcing) function, for sufficient computational accuracy. As an example, if the forcing function is given as

$$F(t) = \sum_{n=1}^{16} 100 \times \cos(2\pi nt)$$

then, the minimum value of  $N$  (= Number of sample data points) should be  $N_{\min} = 32$ .

(b) The factor  $\left(\frac{1}{N}\right)$ , shown in the DFT Equation (21), is merely a scale factor. It can also be placed in the inverse Fourier Transform Equation (1), but not both.

Thus, Equations (21) and (1) can be re written as

$$\tilde{C}_n = \sum_{k=0}^{N-1} f(k) e^{-ik \left( w_0 = \frac{2\pi}{N} \right) n} \quad (22)$$

$$f(k) = \left(\frac{1}{N}\right) \sum_{n=0}^{N-1} \tilde{C}_n e^{ik \left( w_0 = \frac{2\pi}{N} \right) n} \quad (23)$$

To avoid computation with “complex numbers”, Equation (22) can be expressed as

$$\tilde{C}_n^R + i\tilde{C}_n^I = \sum_{k=0}^{N-1} \left\{ f^R(k) + i f^I(k) \right\} \times \{\cos(\theta) - i \sin(\theta)\} \quad (22a)$$

where

$$\theta = k \left( w_0 = \frac{2\pi}{N} \right) n \quad (22b)$$

$$\tilde{C}_n^R + i\tilde{C}_n^I = \sum_{k=0}^{N-1} \left\{ f^R(k) \times \cos(\theta) + f^I(k) \sin(\theta) \right\} + i \left\{ f^I(k) \cos(\theta) - f^R(k) \sin(\theta) \right\}$$

The above “complex number” equation is equivalent to the following 2 “real number” equations

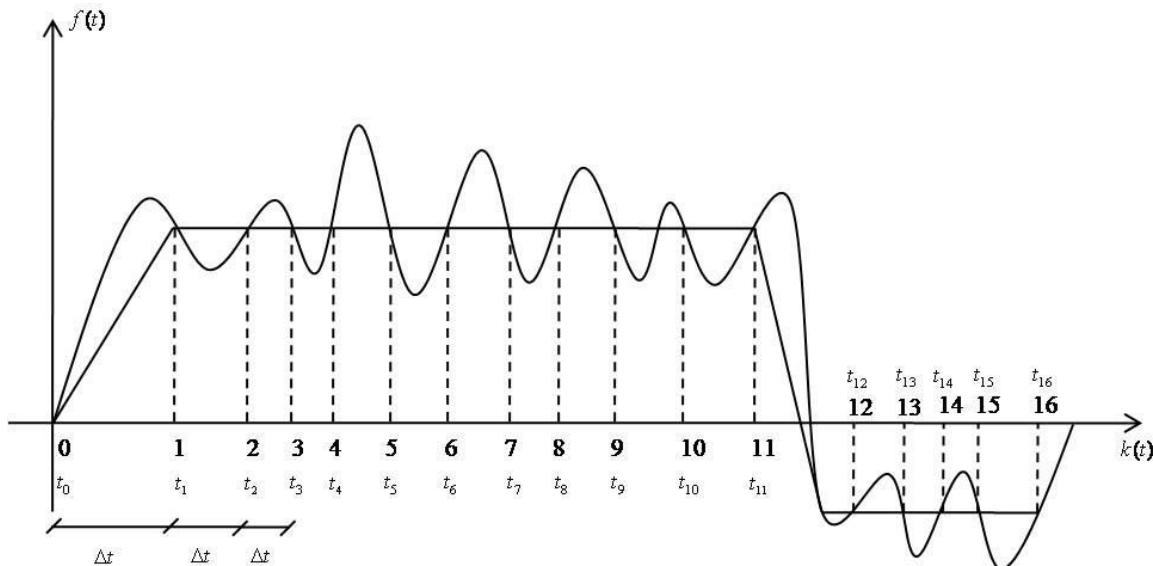
$$\tilde{C}_n^R = \sum_{k=0}^{N-1} \{f^R(k) \cos(\theta) + f^I(k) \sin(\theta)\} \quad (22c)$$

$$\tilde{C}_n^I = \sum_{k=0}^{N-1} \{f^I(k) \cos(\theta) - f^R(k) \sin(\theta)\} \quad (22d)$$

Computer program implementation for the DFT equations (22c, 22d) are given at <http://numericalmethods.eng.usf.edu/simulations/ml/11fft/dft.m>.

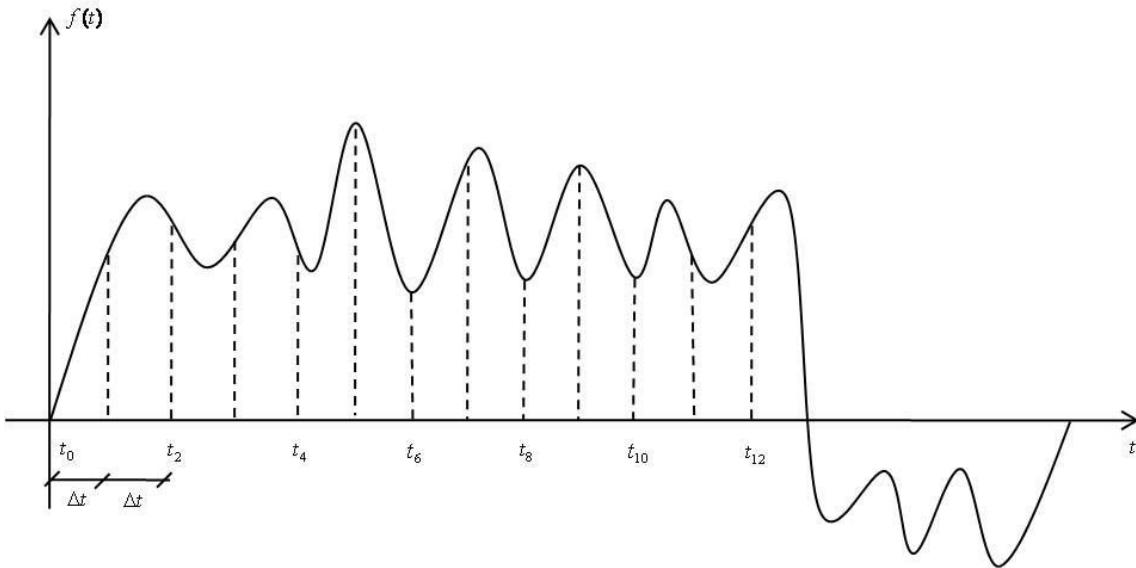
### Detailed Explanation About Aliasing Phenomenon, Nyquist Samples, Nyquist Rate.

When a function  $f(t)$ , which may represent the signals from some real-life phenomenon (shown in Figure 1), is sampled, it basically converts that function into a sequence  $\tilde{f}(k)$  at discrete locations of  $t$ . These discrete locations are assumed to have “equally spaced and the distance between any 2 samples is  $\Delta t$ . Thus,  $\tilde{f}(k)$  represents the value of  $f(t)$ , at  $t = t_0 + k\Delta t$ , where  $t_0$  is the location of the first sample (at  $k = 0$ ). If the sample locations were done properly, then the original function  $f(t)$ , can be recovered through interpolation process of these discrete sample values.



**Figure 1** Function to be Sampled and “Aliased” Sample Problem.

In Figure 1, the samples have been taken with a fairly large  $\Delta t$ . Thus, these sequence of discrete data will not be able to recover the original signal function  $f(t)$ . For example, if all discrete values of  $f(t)$ , were connected by piecewise linear fashion, then a nearly horizontal straight line will occur between  $t_1$  through  $t_{11}$ , and  $t_{12}$  through  $t_{16}$ , respectively (See Figure 1). These piecewise linear interpolation (or other interpolation schemes will NOT produce a curve which resemble well with the original function  $f(t)$ . This is the case where the data has been “ALIASED”.



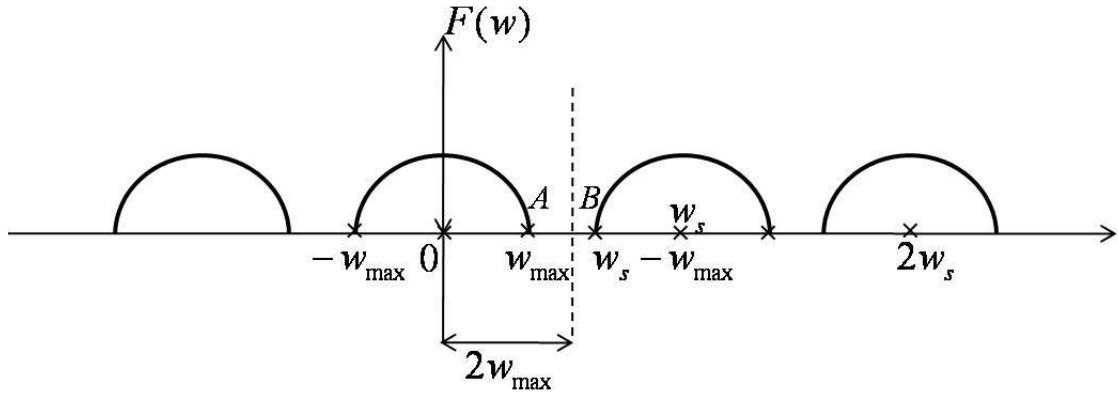
**Figure 2** Function to be sampled and “Windowing” Sample Problem.

Another potential difficulty in sampling the function is called “windowing” problem. As indicated in Figure 2, while  $\Delta t$  is small enough so that a piecewise linear interpolation for connecting these discrete values will adequately resemble the original function  $f(t)$ , however, only a portion of the function  $f(t)$  has been sampled (from  $t_1$  through  $t_{12}$ ) rather than the entire one. In other words, one has placed a “window” over the function.

To avoid aliased phenomenon, the sample space  $\Delta t$  should be small enough so that the discrete sequence will recover back the original function  $f(t)$ . The “sampling theorem” can be stated as:

“If the function  $f(t)$  is band-limited with bandwidth  $2w_{\max}$ ,  $F(w) \equiv$  Fourier transform of  $f(t) = 0$  for  $|w| \geq w_{\max} > 0$  then  $f(t)$  is uniquely determined by a knowledge of its values at uniformly spaced intervals  $\Delta t$  apart, with  $\Delta t = \frac{1}{2w_{\max}}$ .

The above “sampling theorem” can be loosely explained through the help of Figure 3.



**Figure 3** Frequency of sampling rate ( $w_s$ ) versus maximum frequency content ( $w_{\max}$ ).

To satisfy  $F(w) = 0$ , for  $|w| \geq w_{\max}$ , the frequency ( $w$ ) should be between points  $A$  and  $B$  of Figure 3.

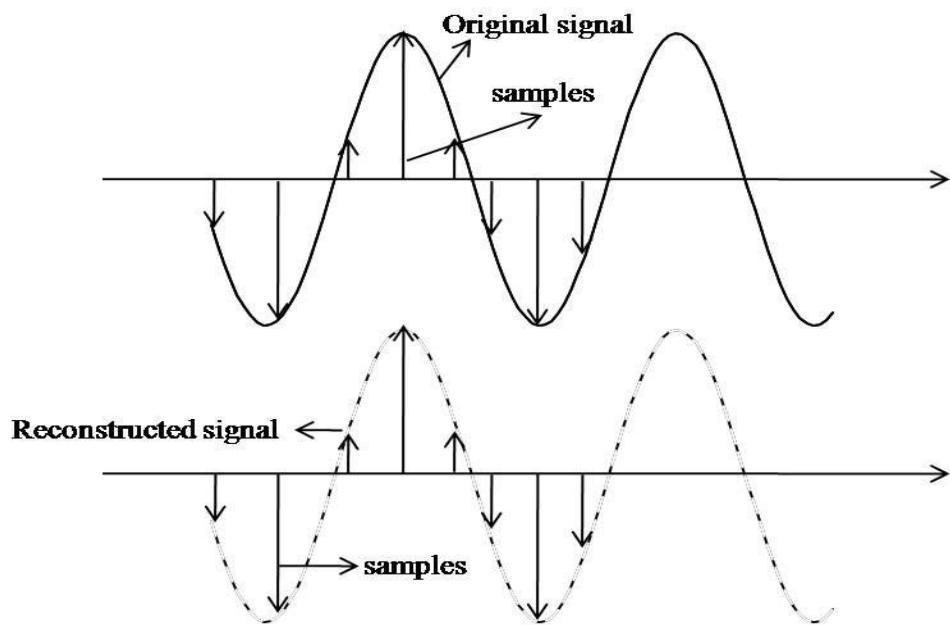
Hence

$$w_{\max} \leq w \leq w_s - w_{\max}$$

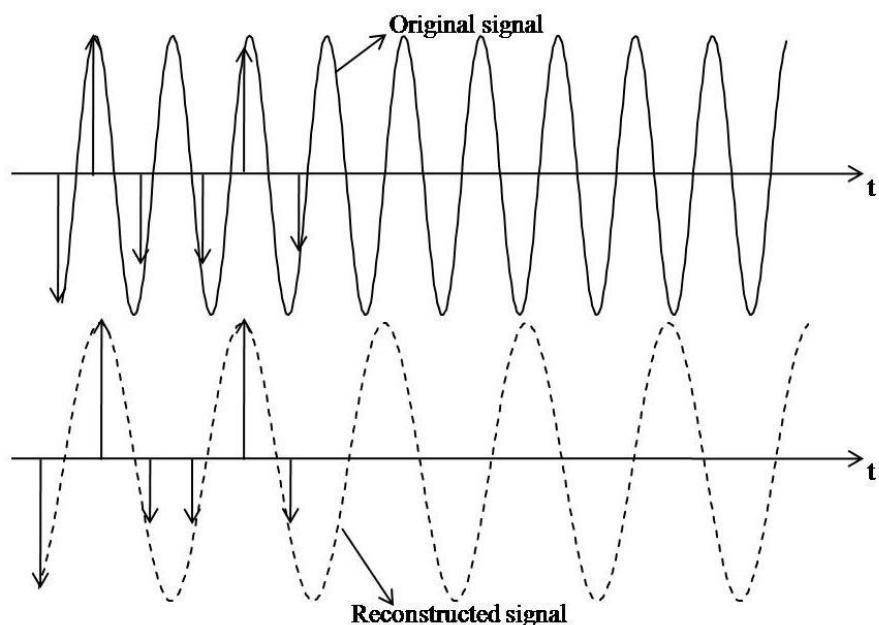
which implies

$$w_s \geq 2w_{\max}$$

Physically, the above equation states that one must have at least 2 samples per cycle of the highest frequency component present (Nyquist samples, Nyquist rate).



**Figure 4** Correctly reconstructed signal.



**Figure 5** Wrongly reconstructed signal.

In Figure 4, a sinusoidal signal is sampled at the rate of 6 samples per 1 cycle (or  $w_s = 6w_0$ ). Since this sampling rate does satisfy the sampling theorem requirement ( $w_s \geq 2w_{\max}$ ), the reconstructed signal does correctly represent the original signal. However, as indicated in Figure 5 a sinusoidal signal is sampled at the rate of 6 samples per 4 cycles (or  $w_s = \frac{6}{4}w_0$ ). Since this sampling rate does NOT satisfy the requirement ( $w_s \geq 2w_{\max}$ ), the reconstructed signal would wrongly represent the original signal.

## References

- [1] E.Oran Brigham, The Fast Fourier Transform, Prentice-Hall, Inc. (1974).
- [2] S.C. Chapra, and R.P. Canale, Numerical Methods for Engineers, 4<sup>th</sup> Edition, Mc-Graw Hill (2002).
- [3] W.H . Press, B.P. Flannery, S.A. Tenkolsky, and W.T. Vetterling, Numerical Recipies, Cambridge University Press (1989), Chapter 12.
- [4] M.T. Heath, Scientific Computing, Mc-Graw Hill (1997).
- [5] H. Joseph Weaver, Applications of Discrete and Continuous Fourier Analysis, John Wiley & Sons, Inc. (1983).

---

### FAST FOURIER TRANSFORM

---

Topic	Discrete Fourier Transform
Summary	Textbook notes on discrete Fourier transform
Major	General Engineering
Authors	Duc Nguyen
Date	July 25, 2010
Web Site	<a href="http://numericalmethods.eng.usf.edu">http://numericalmethods.eng.usf.edu</a>

---

# Chapter 11.05

## Informal Development of Fast Fourier Transform (FFT)

### Introduction

Recalled the DFT pairs of Equations (22) and (23) (of Chapter 11.04) and swapping the indices  $n, k$  one obtains:

$$\tilde{C}_n = \sum_{k=0}^{N-1} f(k) e^{-in\left(\frac{2\pi}{N}\right)k} \quad (1)$$

$$f(k) = \left(\frac{1}{N}\right) \sum_{n=0}^{N-1} \tilde{C}_n e^{in\left(\frac{2\pi}{N}\right)k} \quad (2)$$

$$\text{where } n, k = 0, 1, 2, 3, \dots, N-1 \quad (3)$$

While the above DFT pairs of equations are convenient for computer implementation, they still require substantial computation effort. The objective of this chapter, therefore, is to develop the improved version of DFT (namely Fast Fourier Transform, or FFT) so that much larger sampling data can be handled more efficiently.

Let

$$E = e^{-\frac{i2\pi}{N}} \text{ (hence } E^N = e^{-i2\pi} = \cos(2\pi) - i\sin(2\pi) = 1) \quad (4)$$

Then Equation (1) and Equation (2) become

$$\begin{aligned} \tilde{C}_n &= \tilde{C}(n) = \sum_{k=0}^{N-1} f(k) E^{nk} \\ f(k) &= \left(\frac{1}{N}\right) \sum_{n=0}^{N-1} \tilde{C}_n E^{-nk} \end{aligned} \quad (5)$$

It should be emphasized here that in performing interpolation, one usually has to solve a system of equations to determine the unknown coefficients of the linear combination of basis functions that fit the given data. For example, if  $N = 4$ , then one need to solve the following system (see the second part of Equation (5)), for obtaining  $\{\tilde{C}\}$ , with a given vector  $\{f\}$ .

$$\left(\frac{1}{N}\right) \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & E^{-1} & E^{-2} & E^{-3} \\ 1 & E^{-2} & E^{-4} & E^{-6} \\ 1 & E^{-3} & E^{-6} & E^{-9} \end{bmatrix} \begin{bmatrix} \tilde{C}(0) \\ \tilde{C}(1) \\ \tilde{C}(2) \\ \tilde{C}(3) \end{bmatrix} = \begin{bmatrix} f(0) \\ f(1) \\ f(2) \\ f(3) \end{bmatrix} \quad (5a)$$

However, the inverse of the above coefficient matrix can be easily obtained as

$$\left[ \left( \frac{1}{N} \right) \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & E^{-1} & E^{-2} & E^{-3} \\ 1 & E^{-2} & E^{-4} & E^{-6} \\ 1 & E^{-3} & E^{-6} & E^{-9} \end{bmatrix} \right]^{-1} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & E^1 & E^2 & E^3 \\ 1 & E^2 & E^4 & E^6 \\ 1 & E^3 & E^6 & E^9 \end{bmatrix}$$

Thus, the unknown vector  $\{\tilde{C}\}$  can be computed as matrix times vector operations, as following:

Assuming  $N = 4 = 2^{(r=2)}$ , then (see the first part of Equation (5))

$$\begin{Bmatrix} \tilde{C}(0) \\ \tilde{C}(1) \\ \tilde{C}(2) \\ \tilde{C}(3) \end{Bmatrix} = \begin{bmatrix} E^{(0)(0)} & E^{(0)(1)} & E^{(0)(2)} & E^{(0)(3)} \\ E^{(1)(0)} & E^{(1)(1)} & E^{(1)(2)} & E^{(1)(3)} \\ E^{(2)(0)} & E^{(2)(1)} & E^{(2)(2)} & E^{(2)(3)} \\ E^{(3)(0)} & E^{(3)(1)} & E^{(3)(2)} & E^{(3)(3)} \end{bmatrix} \begin{Bmatrix} f(0) \\ f(1) \\ f(2) \\ f(3) \end{Bmatrix} \quad (6)$$

$$\begin{Bmatrix} \tilde{C}(0) \\ \tilde{C}(1) \\ \tilde{C}(2) \\ \tilde{C}(3) \end{Bmatrix} = \begin{bmatrix} E^0 & E^0 & E^0 & E^0 \\ E^0 & E^1 & E^2 & E^3 \\ E^0 & E^2 & E^4 & E^6 \\ E^0 & E^3 & E^6 & E^9 \end{bmatrix} \begin{Bmatrix} f(0) \\ f(1) \\ f(2) \\ f(3) \end{Bmatrix} \quad (7)$$

For  $N = 4$ ,  $n = 2$  and  $k = 3$ , then

$$\begin{aligned} E^{nk} &= E^6 \\ &= [E^{(N=4)}]E^2 \\ &= \left( e^{\frac{-i2\pi}{N}} \right)^N E^2 \\ &= [e^{-i2\pi}]E^2 \\ &= E^2 \end{aligned}$$

The term inside the square bracket is equal to 1, since

$$\begin{aligned} [e^{-i2\pi}] &= \cos(-2\pi) + i \sin(-2\pi) \\ &= \cos(2\pi) - i \sin(2\pi) \\ &= 1 - i(0) = 1 \end{aligned}$$

For  $N = 4$ ,  $n = 3$  and  $k = 3$ , then

$$\begin{aligned} E^{nk} &= E^9 \\ &= [E^8]E^1 \\ &= [E^{2N}]E^1 \\ &= \left[ e^{\frac{-i2\pi \times 2N}{N}} \right]E^1 \\ &= [e^{-i4\pi}]E^1 \\ &= E^1 \end{aligned}$$

In the above equation, one should recall the following Euler identity

$$\begin{aligned} e^{-i4\pi} &= \cos(4\pi) - i\sin(4\pi) \\ &= 1 \end{aligned}$$

Thus, in general (for  $nk \geq N$ )

$$E^{nk} = E^U$$

where

$$\begin{aligned} U &= \text{mod}(nk, N) \\ &= \text{remainder of } \left(\frac{nk}{N}\right) \end{aligned} \tag{8}$$

Remarks:

Matrix times vector, shown in Equation (7), will require 16 (or  $N^2$ ) complex multiplications and 12 (or  $N * (N - 1)$ ) complex additions.

Usage of Equation (8) will help to reduce the number of operation counts, as explained in the next section.

### Factorized Matrix and Further Operation Count

Equation (7) can be factorized as:

$$\begin{Bmatrix} \tilde{C}(0) \\ \tilde{C}(2) \\ \tilde{C}(1) \\ \tilde{C}(3) \end{Bmatrix} = \begin{bmatrix} 1 & E^0 & 0 & 0 \\ 1 & E^2 & 0 & 0 \\ 0 & 0 & 1 & E^1 \\ 0 & 0 & 1 & E^3 \end{bmatrix} \begin{bmatrix} 1 & 0 & E^0 & 0 \\ 0 & 1 & 0 & E^0 \\ 1 & 0 & E^2 & 0 \\ 0 & 1 & 0 & E^2 \end{bmatrix} \begin{Bmatrix} f(0) \\ f(1) \\ f(2) \\ f(3) \end{Bmatrix} \tag{9}$$

Remarks:

The theory behind the 2 matrices on the right hand side (RHS) of Equation (9) will be clearly explained soon (see Equations 11 and 15, in chapter 11.06).

The order of the left-hand-side (LHS) vector has been changed, such as rows 2 and 3 have been swapped.

Let the row-interchanged LHS vector be defined as

$$\tilde{C}^*(n) = \begin{Bmatrix} \tilde{C}(0) \\ \tilde{C}(2) \\ \tilde{C}(1) \\ \tilde{C}(3) \end{Bmatrix} \tag{10}$$

Now performing the inner-product (matrix times vector) on the RHS of Equation (9), one obtains

$$\begin{Bmatrix} f_1(0) \\ f_1(1) \\ f_1(2) \\ f_1(3) \end{Bmatrix} = \begin{bmatrix} 1 & 0 & E^0 & 0 \\ 0 & 1 & 0 & E^0 \\ 1 & 0 & E^2 & 0 \\ 0 & 1 & 0 & E^2 \end{bmatrix} \begin{Bmatrix} f(0) \\ f(1) \\ f(2) \\ f(3) \end{Bmatrix} \tag{11}$$

or

$$f_1(0) = f(0) + E^0 f(2) \tag{11a}$$

$$f_1(1) = f(1) + E^0 f(3) \quad (11b)$$

$$f_1(2) = f(0) - E^0 f(2) \quad (11c)$$

since

$$\begin{aligned} E^2 &= e^{-i\frac{2\pi}{4}} \\ &= e^{-i\pi} \\ &= -1 \\ &= -E^0 \\ f_1(3) &= f(1) + E^2 f(3) \\ &= f(1) - E^0 f(3) \end{aligned} \quad (11d)$$

Equations (11a through 11d) for the “inner” matrix times vector requires 2 complex multiplications and 4 complex additions.

In Equations (11a–11d),  $E^0$  is intentionally not reduced to the numerical value of 1.0 to facilitate the discussions of more general cases.

Finally, performing the “outer” product (matrix times vector) on the RHS of Equation (9), one obtains:

$$\begin{Bmatrix} \tilde{C}(0) \\ \tilde{C}(2) \\ \tilde{C}(1) \\ \tilde{C}(3) \end{Bmatrix} = \begin{Bmatrix} f_2(0) \\ f_2(1) \\ f_2(2) \\ f_2(3) \end{Bmatrix} = \begin{bmatrix} 1 & E^0 & 0 & 0 \\ 1 & E^2 & 0 & 0 \\ 0 & 0 & 1 & E^1 \\ 0 & 0 & 1 & E^3 \end{bmatrix} \begin{Bmatrix} f_1(0) \\ f_1(1) \\ f_1(2) \\ f_1(3) \end{Bmatrix} \quad (13)$$

or

$$f_2(0) = f_1(0) + E^0 f_1(1) \quad (14a)$$

$$f_2(1) = f_1(0) + E^2 f_1(1) \quad (14b)$$

$$= f_1(0) - E^0 f_1(1)$$

$$f_2(2) = f_1(2) + E^1 f_1(3) \quad (14c)$$

$$f_2(3) = f_1(2) + E^3 f_1(3)$$

$$= f_1(2) + E^2 E^1 f_1(3)$$

$$= f_1(2) - E^1 f_1(3) \quad (14d)$$

Again, Equations (14a-14d) requires 2 complex multiplications and 4 complex additions.

Thus, the complete RHS of Equation (9) can be computed by only 4 complex multiplications

(or  $N \frac{r}{2} = 4 \frac{2}{2}$ ) and 8 complex additions (or  $Nr = 4 \times 2$ ). Since computational time is mainly

controlled by the number of multiplications, hence implementing Equation (9) will significantly reduce the number of multiplication, as compared to direct matrix times vector operations (as shown in Equation (7)).

For large value of data points ( $= N$ ), the ratio of complex multiplications by using Equation (7) and Equation (9) can be computed as

$$Ratio = \frac{N^2}{\left(\frac{Nr}{2}\right)} = \left(\frac{2N}{r}\right) \quad (15)$$

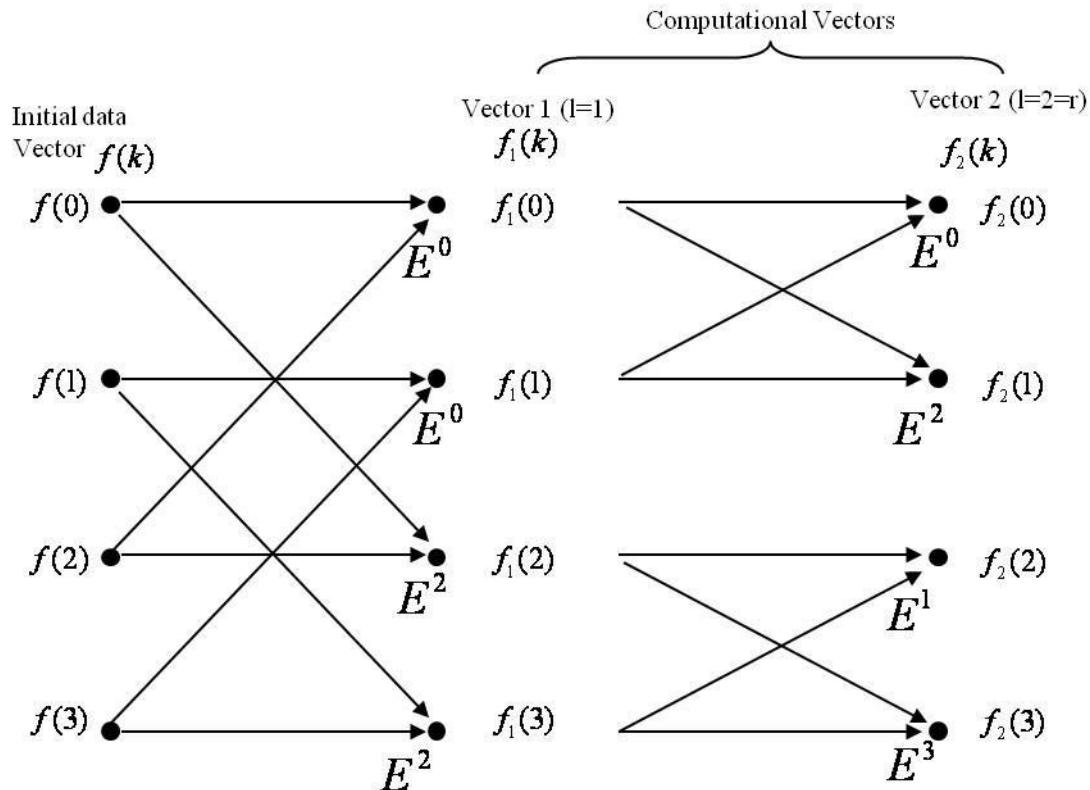
For  $N = 2048 = 2^{(r=11)}$ , Equation (15) gives

$$Ratio = \frac{2(2048)}{11} = 372.36,$$

which basically implies that the number of complex multiplications involved in Equation (9) is about 372 times less than the one involved in Equation (7).

### Graphical flow of Equation (9), for case $N = 2^r = 2^2 = 4$

Equation (9) can also be presented in the graphical form, as shown in Figure 1.



**Figure 1** Graphical form of FFT (Equation 9) for the case  $N = 2^r = 2^2 = 4$ .

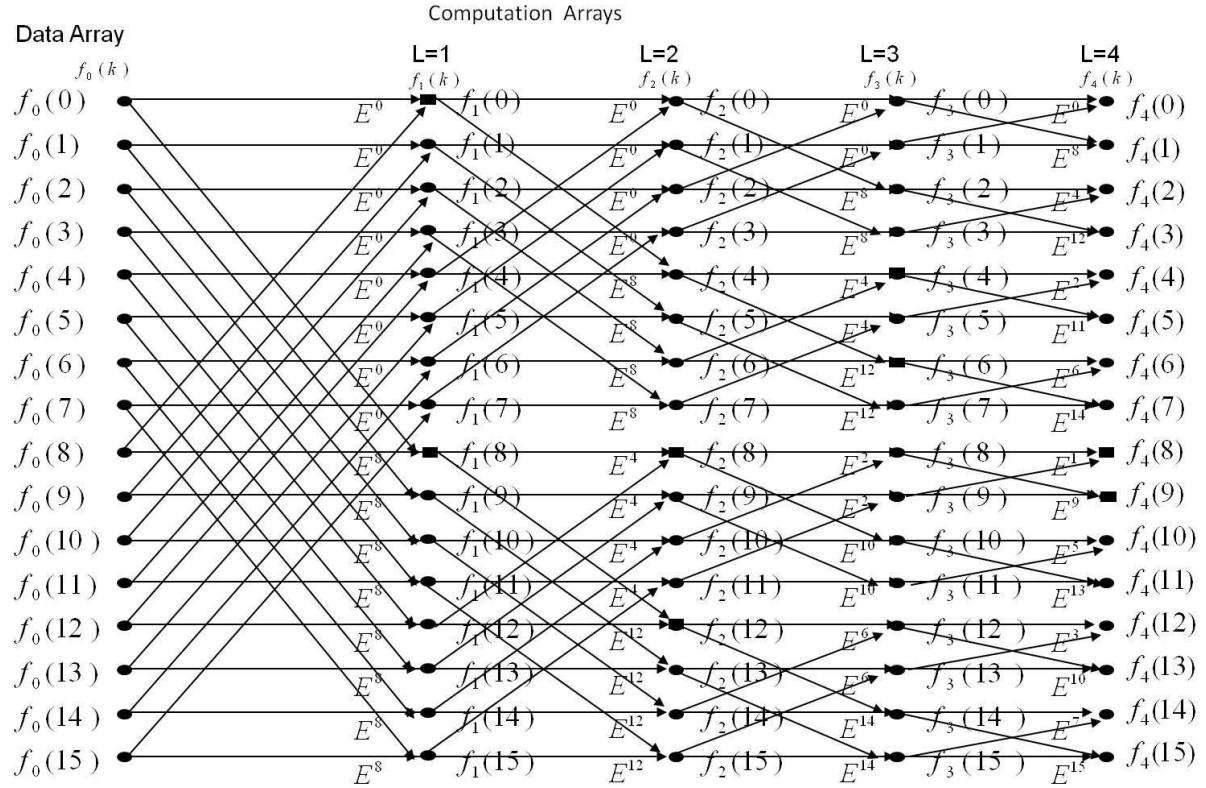
Remarks:

- Computed vector 1 does correspond to Equations (11a–11d).
- Computed vector 2 does correspond to Equations (14a–14d).
- Since  $r = 2$  in this example, one needs to compute 2 vectors  $\{= f_1(k) \text{ and } f_2(k)\}$
- Each node in the graph is computed from  $2 (= r)$  nodes in the “previous” vector.
- Factor  $E^U$  (such as  $E^0, E^1, E^2, E^3$ ) appears near the arrow head of the transmission path. Absence of  $E^U$  implies that  $E^U = E^0 = 1$ .

For example:  $f_2(2) = f_1(2) + f_1(3)E^1$ , which is the same as Equation (14c).

### Graphical Flow of Equation (9), for case $N = 2^r = 2^4 = 16$

To see a more detailed computational patterns of FFT, a slightly larger data size ( $N = 2^r = 2^4 = 16$ ) is shown in the graphical form, as indicated in Figure 2.



**Figure 2** Graphical form of FFT (Equation 9) for the case  $N = 2^r = 2^4 = 16$ .

### Companion Node Observation

Careful observation of Figure 2 reveals that for each computed  $l^{th}$ -vector (where  $l = 1, 2, \dots, r$ ; and  $N = 2^r = 2^4 = 16$ ), we can always find two (companion) nodes which came from the same pair of nodes in the previous vector. For example,  $f_1(0)$  and  $f_1(8)$  are computed in terms of  $f(0)$  and  $f(8)$ . Similarly, the companion nodes  $f_2(8)$  and  $f_2(12)$  are computed from the same pair of nodes  $f_1(8)$  and  $f_1(12)$ .

Furthermore, the computation of companion nodes is independent of other nodes (within the  $l^{th}$ -vector). Therefore, the computed  $f_1(0)$  and  $f_1(8)$  will override the original space of  $f(0)$  and  $f(8)$ . Similarly, the computed  $f_2(8)$  and  $f_2(12)$  will over ride the space occupied by  $f_1(8)$  and  $f_1(12)$ , which in turn, will occupy the original space of  $f(8)$  and  $f(12)$ . Hence, only one complex vector (or 2 real vectors) of length  $N$  are needed for the entire FFT process.

### Companion Node Spacing

Observing Figure 2, the following statements can be made:

- a) in the first vector ( $l = 1$ ), the companion nodes  $f_1(0)$  and  $f_1(8)$  is separated by  $k = 8$  (or  $\frac{N}{2^l} = \frac{16}{2^1} = 8$ ) spaces.
- b) In the second vector ( $l = 2$ ), the companion nodes  $f_2(8)$  and  $f_2(12)$  is separated by  $k = 4$  (or  $\frac{N}{2^l} = \frac{16}{2^2} = 4$ ).

### Companion Node Computation

The operation counts in any companion nodes (of the  $l^{th} = 2^{nd}$  vector), such as  $f_2(8)$  and  $f_2(12)$  can be explained as (see Figure 2):

$$f_2(8) = f_1(8) + f_1(12) \times E^4 \quad (16)$$

$$\begin{aligned} f_2(12) &= f_1(8) + f_1(12) \times E^{12} \\ &= f_1(8) + f_1(12) \times E^8 E^4 \\ &= f_1(8) + f_1(12) \left[ e^{-i\frac{2\pi}{(N=16)}} \right]^8 E^4 \\ &= f_1(8) + f_1(12) [e^{-i\pi}] E^4 \\ &= f_1(8) - f_1(12) \times E^4 \end{aligned} \quad (17)$$

Thus, the companion nodes  $f_2(8)$  and  $f_2(12)$  computation will require 1 complex multiplication and 2 complex additions (see Equations (16-17)). The weighting factors for the companion nodes [ $f_2(8)$  and  $f_2(12)$ ] are  $E^4$  (or  $E^U$ ) and  $E^{12}$  (or  $E^{\frac{U+N}{2}}$ ), respectively. Thus, in general

$$f_l(k) = f_{l-1}(k) + E^U f_{l-1}\left(k + \frac{N}{2^l}\right) \quad (18)$$

$$f_l\left(k + \frac{N}{2^l}\right) = f_{l-1}(k) - E^U f_{l-1}\left(k + \frac{N}{2^l}\right) \quad (19)$$

### Skipping computation of certain nodes

Because the pair of companion nodes “ $k$ ” and “ $k + \frac{N}{2^L}$ ” are separated by the “distance”. ( $= \frac{N}{2^L}$ ), hence, at the  $L^{th}$  level, after every  $\frac{N}{2^L}$  node computation, then the next  $\frac{N}{2^L}$  nodes will be skipped! (see Figure 2).

### Determination of $E^U$

The values of “ $U$ ” can be determined by the following steps:

Step 1: Express the index  $k$  ( $= 0,1,2,\dots,N-1$ ) in binary form, using  $r$  bits. For  $k = 8$ ,  $L = 2$  and  $r = 4$ ; or  $N = 2^r = 2^4 = 16$ , one obtains

$$\begin{aligned} k &= 8 \\ &= 1,0,0,0 \\ &= (1)2^{r-1=3} + (0)2^2 + (0)2^1 + (0)2^0 \end{aligned}$$

Step 2: Sliding this binary number “ $r - L = 4 - 2 = 2$ ” positions to the right, and fill in zeros, the results are

$$1,0,0,0 \rightarrow X, X, 1,0 \rightarrow 0,0,1,0$$

It is important to realize that the results of Step 2 (0,0,1,0) is equivalent to express an integer

$$\begin{aligned} M &= \frac{k}{2^{r-L}} \\ &= \frac{8}{2^{4-2}} \\ &= 2 \end{aligned}$$

in the binary formats. In other words,  $M = 2 = (0,0,1,0)$ .

Step 3: Reverse the order of the bits, then (0,0,1,0) becomes 0,1,0,0 =  $E$ . Thus,

$$\begin{aligned} U &= (0)2^3 + (1)2^2 + (0)2^1 + (0)2^0 \\ &= 4 \end{aligned}$$

It is “NOT” really necessary to perform Step 3, since the results of Step 2 can be used to compute “ $E$ ” as following

$$\begin{aligned} U &= (0)2^0 + (0)2^1 + (1)2^2 + (0)2^3 \\ &= 4 \end{aligned}$$

In conclusion, for  $N = 2^r = 2^4 = 16$ ;  $L = 2$ ;  $k = 8$  and  $U = 4$ ; the computation of companion nodes from general formulas (see Equations (18) and (19)) gives

$$f_2(8) = f_1(8) + E^4 f_1(12)$$

$$f_2(12) = f_1(8) - E^4 f_1(12)$$

The above 2 equations are identical to Equations (16) and (17).

### Computer Implementation to Find Value of “U” (in $E^U$ )

Based on the previous discussions (with the 3-step procedures), to find the value of “U”, one only needs a procedure to express an integer  $M = \frac{k}{2^{r-L}}$  in binary formats, with “r” bits.

Assuming  $M$  (a base 10 number) can be expressed as (assuming  $r = 4$  bits)

$$\begin{aligned} M &= a_4 a_3 a_2 a_1 \\ &= J_1 \end{aligned} \tag{20}$$

Divide  $M$  by 2 (say,  $J_2 = \frac{J_1}{2}$ ), multiply the truncated result by 2 (say,  $JJ_2 = J_2 \times 2$ ), and compute the difference between the original number  $= M = J_1$  and  $JJ_2$ .

$$IDIFF = J_1 - JJ_2 \left\{ = M - \left( \frac{M}{2} \right)_{Truncated} \times 2 \right\} \tag{21}$$

If  $IDIFF = 0$ , then the bit  $a_1 = 0$

If  $IDIFF \neq 0$ , then the bit  $a_1 = 1$

Once the bit  $a_1$  has been determined, the value of  $J_1$  is set to  $J_2$  (or value of  $J_1$  is reduced by a factor of 2), since the previous

$$\begin{aligned} J_1 &= M \\ &= a_4 a_3 a_2 a_1 \\ J_1 &= (a_1)2^0 + (a_2)2^1 + (a_3)2^2 + (a_4)2^3 \end{aligned}$$

and similar process can be used to determine the value of bit  $a_2$ , etc.

#### Example 1

For  $k = 8$ ;  $N = 16 = 2^r$ ;  $r = 4$  bits and  $L = 2$ , Find the value of  $U$ .

$$\begin{aligned} M &= \frac{k}{2^{r-L}} \\ &= \frac{8}{2^{4-2}} \\ &= 2 \\ &= J_1 \end{aligned}$$

Determine the bit  $a_1$ : (Index  $I = 1$ )

Initialize  $U = 0$

$$\begin{aligned} J_2 &= \frac{J_1}{2} \\ &= \frac{2}{2} \\ &= 1 \\ IDIFF &= J_1 - (JJ_2 = J_2 \times 2) \\ &= 2 - (1)(2) \\ &= 0 \end{aligned}$$

Thus

$$\begin{aligned}
 a_1 &= 0 \\
 U &= U \times 2 + IDIFF \\
 &= 0 \times 2 + 0 \\
 &= 0
 \end{aligned}$$

or

$$\begin{aligned}
 U &= U + (a_1)2^{r-1} \\
 &= 0 + (0)2^3 \\
 &= 0
 \end{aligned}$$

Determine the bit  $a_2$  [Index  $I = 2$ ]

$$J_1 = J_2$$

$$= 1$$

$$J_2 = \frac{J_1}{2}$$

$$= \frac{1}{2}$$

$$= 0$$

$$\begin{aligned}
 IDIFF &= J_1 - (JJ_2 = J_2 \times 2) \\
 &= 1 - (0 \times 2) \\
 &= 1
 \end{aligned}$$

Thus  $a_2 = 1$

$$\begin{aligned}
 U &= U \times 2 + IDIFF \\
 &= 0 \times 2 + 1 \\
 &= 1
 \end{aligned}$$

or

$$\begin{aligned}
 U &= U + (a_2)2^{r-1} \\
 &= 0 + (1)2^2 \\
 &= 4
 \end{aligned}$$

Determine the bit  $a_3$  [Index  $I = 3$ ]

$$\begin{aligned}
 J_1 &= J_2 \\
 &= 0
 \end{aligned}$$

$$J_2 = \frac{J_1}{2}$$

$$= \frac{0}{2}$$

$$= 0$$

$$\begin{aligned}
 IDIFF &= J_1 - (JJ_2 = J_2 \times 2) \\
 &= 0 - (0 \times 2) \\
 &= 0
 \end{aligned}$$

Thus  $a_3 = 0$

$$\begin{aligned}
 U &= U \times 2 + IDIFF \\
 &= 1 \times 2 + 0 \\
 &= 2
 \end{aligned}$$

or

$$\begin{aligned} U &= U + (a_3)2^{r-I} \\ &= 4 + (0)2^1 \\ &= 4 \end{aligned}$$

Determine the bit  $a_4$  [Index  $I = 4 = r$ ]

$$\begin{aligned} J_1 &= J_2 \\ &= 0 \end{aligned}$$

$$\begin{aligned} J_2 &= \frac{J_1}{2} \\ &= \frac{0}{2} \\ &= 0 \end{aligned}$$

$$\begin{aligned} IDIFF &= J_1 - (JJ_2 = J_2 \times 2) \\ &= 0 - (0) \times 2 \\ &= 0 \end{aligned}$$

Thus  $a_4 = 0$

$$\begin{aligned} U &= U \times 2 + IDIFF \\ &= 2 \times 2 + 0 \\ &= 4 \end{aligned}$$

or

$$\begin{aligned} U &= U + (a_4)2^{r-I} \\ &= 4 + (0)2^0 \\ &= 4 \end{aligned}$$

Remarks:

Although the “intermediate” results might be different, at the end of the do-loop process (computing  $a_4$ ), both formulas for “ $U$ ”, such as

$$U = U \times 2 + IDIFF; \text{ or} \quad (22)$$

$$U = U + (a_I)2^{r-I}; \text{ where } U = 1, 2, 3, \dots, r \quad (23)$$

will eventually give the same final answers for “ $U$ ”.

### Example 2

For  $k = 12$ ;  $N = 16 = 2^{r=4}$ ; and  $L = 3$ . Compute the corresponding value of  $U$ ?

One has

$$\begin{aligned} M &= \frac{k}{2^{r-L}} \\ &= \frac{12}{2^{4-3}} \end{aligned}$$

$$\begin{aligned} J_1 &= J_2 \\ &= 3 \end{aligned}$$

Determine the bit  $a_1$ : (Index  $I = 1$ )

Initialize  $U = 0$

$$\begin{aligned} J_2 &= \frac{J_1}{2} \\ &= \frac{6}{2} \\ &= 3 \end{aligned}$$

$$\begin{aligned} IDIFF &= J_1 - (JJ_2 = J_2 \times 2) \\ &= 6 - (3)(2) \\ &= 0 \end{aligned}$$

Thus

$$\begin{aligned} a_1 &= 0 \\ U &= U \times 2 + IDIFF \\ &= 0 \times 2 + 0 \\ &= 0 \end{aligned}$$

or

$$\begin{aligned} U &= U + (a_1)2^{r-1} \\ &= 0 + (0)2^3 \\ &= 0 \end{aligned}$$

Determine the bit  $a_2$  [Index  $I = 2$ ]

$$\begin{aligned} J_1 &= J_2 \\ &= 3 \\ J_2 &= \frac{J_1}{2} \\ &= \frac{3}{2} \\ &= 1 \end{aligned}$$

$$\begin{aligned} IDIFF &= J_1 - (JJ_2 = J_2 \times 2) \\ &= 3 - (1) \times 2 \\ &= 1 \end{aligned}$$

Thus  $a_2 = 1$

$$\begin{aligned} U &= U \times 2 + IDIFF \\ &= 0 \times 2 + 1 \\ &= 1 \end{aligned}$$

or

$$\begin{aligned} U &= U + (a_2)2^{r-2} \\ &= 0 + (1)2^2 \\ &= 4 \end{aligned}$$

Determine the bit  $a_3$  [Index  $I = 3$ ]

$$\begin{aligned} J_1 &= J_2 \\ &= 1 \\ J_2 &= \frac{J_1}{2} \end{aligned}$$

$$\begin{aligned}
 &= \frac{1}{2} \\
 &= 0 \\
 IDIFF &= J_1 - (J_2 = J_2 \times 2) \\
 &= 1 - (0) \times 2 \\
 &= 1
 \end{aligned}$$

Thus  $a_3 = 1$

$$\begin{aligned}
 U &= U \times 2 + IDIFF \\
 &= 1 \times 2 + 1 \\
 &= 3
 \end{aligned}$$

or

$$\begin{aligned}
 U &= U + (a_3)2^{r-3} \\
 &= 4 + (1)2^1 \\
 &= 6
 \end{aligned}$$

Determine the bit  $a_4$  [Index  $I = 4$ ]

$$\begin{aligned}
 J_1 &= J_2 \\
 &= 0 \\
 J_2 &= \frac{J_1}{2} \\
 &= \frac{0}{2} \\
 &= 0 \\
 IDIFF &= J_1 - (J_2 = J_2 \times 2) \\
 &= 0 - (0) \times 2 \\
 &= 0
 \end{aligned}$$

Thus  $a_4 = 0$

$$\begin{aligned}
 U &= U \times 2 + IDIFF \\
 &= 3 \times 2 + 0 \\
 &= 6
 \end{aligned}$$

or

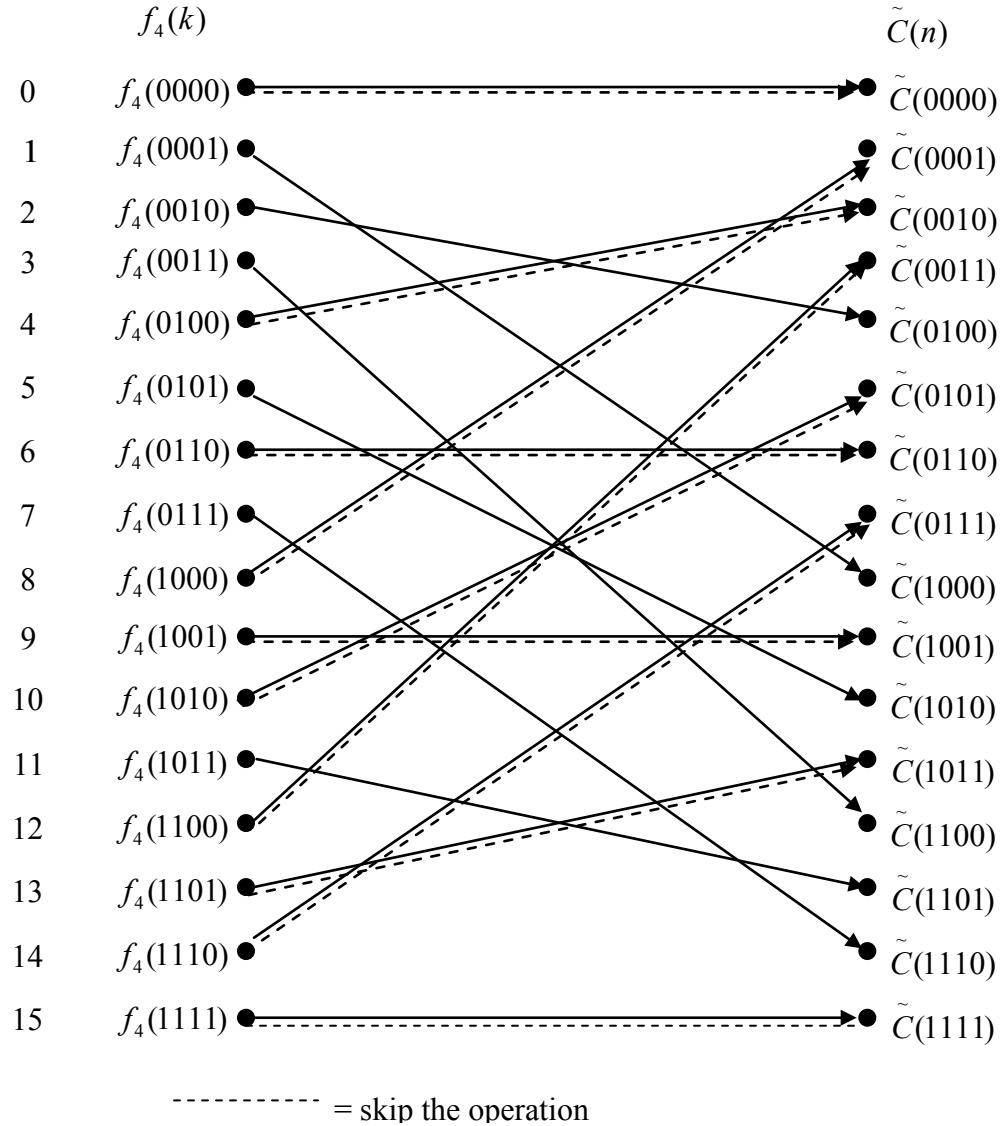
$$\begin{aligned}
 U &= U + (a_4)2^{r-4} \\
 &= 6 + (0)2^0 \\
 &= 6
 \end{aligned}$$

Remarks:

Although both formulas for “ $U$ ”, shown in Equations (22) and (23), will yield the same “final” value of “ $U$ ”. Implementation of Equation (22) will be more computationally efficient.

### Unscrambling the FFT

For the case  $N = 16 = 2^{r=4}$  (see Figure 2), the final ‘bit-reversing’ operation for FFT is shown in Figure 3.



**Figure 3** Final “bit-reversing” for FFT (with  $N = 2^r = 2^4 = 16$ ).

For do-loop index  $k = 0 = (0,0,0,0) \Rightarrow i = (0,0,0,0) = 0$

*If*( $i.GT.k$ )*Then*

$T = f_4(k)$

$f_4(k) = f_4(i)$

$f_4(i) = T$

*Endif*

Hence,  $f_4(0) = f_4(0)$ ; no swapping.

For  $k = 1 = (0,0,0,1) \Rightarrow i = (1,0,0,0) = 0$  = bit-reversion=8

If( $i.GT.k$ )Then

$$T = f_4(k)$$

$$f_4(k) = f_4(i)$$

$$f_4(i) = T$$

Endif

Hence,  $f_4(1)=f_4(8)$ ; are swapped.

For  $k = 2 = (0,0,1,0) \Rightarrow i = (0,1,0,0) = 4$

Hence,  $f_4(2)=f_4(4)$ ; are swapped.

For  $k = 3 = (0,0,1,1) \Rightarrow i = (1,1,0,0) = 12$

Hence,  $f_4(3)=f_4(12)$ ; are swapped.

For  $k = 4 = (0,1,0,0) \Rightarrow i = (0,0,1,0) = 2$

In this case, since “ $i$ ” is not greater than “ $k$ ”.

Hence, no swapping, since  $f_4(k = 2)$  and  $f_4(i = 4)$ ; had already been swapped earlier.

.

.

etc...

### Computer Implementation of FFT (for case $N = 2^r$ ).

The pair of companion nodes computation are given by Equations (18, 19). To avoid “complex number” operations, Equation (18) can be computed based on “real number” operations, as following:

$$\begin{aligned} \{f_L^R(k) + if_L^I(k)\} &= \{f_{L-1}^R(k) + if_{L-1}^I(k)\} \\ &\quad + \left\{E^{U,R} + iE^{U,I}\right\} \times \left\{f_{L-1}^R\left(k + \frac{N}{2^L}\right) + if_{L-1}^I\left(k + \frac{N}{2^L}\right)\right\} \end{aligned} \quad (24)$$

In Equation (24), the superscripts  $R$  and  $I$  denote real and imaginary components, respectively.

Multiplying the last 2 complex numbers, one obtains:

$$\begin{aligned} \{f_L^R(k) + if_L^I(k)\} &= \{f_{L-1}^R(k) + if_{L-1}^I(k)\} \\ &\quad + \left\{E^{U,R} \times f_{L-1}^R\left(k + \frac{N}{2^L}\right) - E^{U,I} \times f_{L-1}^I\left(k + \frac{N}{2^L}\right)\right\} \\ &\quad + i \left\{E^{U,R} \times f_{L-1}^I\left(k + \frac{N}{2^L}\right) + E^{U,I} \times f_{L-1}^R\left(k + \frac{N}{2^L}\right)\right\} \end{aligned} \quad (25)$$

Equating the real (and then, imaginary) components on the Left-Hand-Side (LHS), and the Right-Hand-Side (RHS) of Equation (25), one obtains

$$\{f_L^R(k)\} = \{f_{L-1}^R(k)\} + \left\{E^{U,R} \times f_{L-1}^R\left(k + \frac{N}{2^L}\right) - E^{U,I} \times f_{L-1}^I\left(k + \frac{N}{2^L}\right)\right\} \quad (26a)$$

$$\{f_L^I(k)\} = \{f_{L-1}^I(k)\} + \left\{ E^{U,R} \times f_{L-1}^I(k + \frac{N}{2^L}) + E^{U,I} \times f_{L-1}^R(k + \frac{N}{2^L}) \right\} \quad (26b)$$

Recall Equation (4)

$$E = e^{-i\frac{2\pi}{N}}$$

Hence

$$\begin{aligned} E^U &= \left( e^{-i\frac{2\pi}{N}} \right)^U \\ &= e^{-i\frac{2\pi U}{N}} \\ &= e^{-i\theta} \\ &= \cos(\theta) - i \sin(\theta) \end{aligned} \quad (27)$$

where

$$\begin{aligned} \theta &= \frac{2\pi U}{N} \\ &= \frac{6.28U}{N} \end{aligned} \quad (28)$$

Thus

$$E^{U,R} = \cos(\theta) \quad (29a)$$

$$E^{U,I} = -\sin(\theta) \quad (29b)$$

Substituting Equations (29a, 29b) into Equations (26a, 26b), one gets

$$\{f_L^R(k)\} = \{f_{L-1}^R(k)\} + \left\{ \cos(\theta) \times f_{L-1}^R(k + \frac{N}{2^L}) + \sin(\theta) \times f_{L-1}^I(k + \frac{N}{2^L}) \right\} \quad (30a)$$

$$\{f_L^I(k)\} = \{f_{L-1}^I(k)\} + \left\{ \cos(\theta) \times f_{L-1}^I(k + \frac{N}{2^L}) - \sin(\theta) \times f_{L-1}^R(k + \frac{N}{2^L}) \right\} \quad (30b)$$

Similarly, the single (complex number) Equation (19) can be expressed as 2 equivalent (real number) equations, such as equations (30a, 30b).

Listing of computer implementation of serial FFT algorithm is given at [http://numericalmethods.eng.usf.edu/simulations/ml/11fft/general\\_fft.m](http://numericalmethods.eng.usf.edu/simulations/ml/11fft/general_fft.m)

## References

- [1] E.Oran Brigham, The Fast Fourier Transform, Prentice-Hall, Inc. (1974).
- [2] S.C. Chapra, and R.P. Canale, Numerical Methods for Engineers, 4<sup>th</sup> Edition, Mc-Graw Hill (2002).
- [3] W.H . Press, B.P. Flannery, S.A. Tenkolsky, and W.T. Vetterling, Numerical Recipies, Cambridge University Press (1989), Chapter 12.
- [4] M.T. Heath, Scientific Computing, Mc-Graw Hill (1997).
- [5] H. Joseph Weaver, Applications of Discrete and Continuous Fourier Analysis, John Wiley & Sons, Inc. (1983).

---

## FAST FOURIER TRANSFORM

---

Topic      Informal Development of Fast Fourier Series  
Summary    Textbook notes on the informal development of fast Fourier series  
Major      General Engineering  
Authors     Duc Nguyen  
Date       February 9, 2012  
Web Site    <http://numericalmethods.eng.usf.edu>

---