Introduction to Numerical Methods

1. Chapter 01.01 Introduction to Numerical Methods

PRE-REQUISITES (ön koşullar)

1. Be able to find integrals of a function (Primer for Integral Calculus).
2. Understand the concept of curve fitting.

OBJECTIVES (hedefler)

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

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 (ayrılmaz/bütünü) part in **solving engineering problems**. Many times, these mathematical models are derived (türetilmiş) from engineering and science principles, while at other times the models may be obtained (elde edilmiş/toplanmış) from experimental data.

Mathematical models generally result in need of using mathematical procedures that include but are not limited to (matematiksel modellerde matematiksel işlemlere gereksinim vardır)

(A) differentiation, (değişiklik/farklılaşma)
(B) nonlinear equations, (çizgisel olmayan eşitlikler)
(C) simultaneous linear equations, (aynı anda çözülen çizgisel eşitlikler)
(D) curve fitting by interpolation or regression, (interpolasyon/regresyon ile eğri uydurma)
(E) integration, (toplama) and
(F) differential equations (diferensiyel eşitlikler).

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** (derslerde matematik problemlerini analitik çözerken kesin sonuçlar elde etmişsinizdir, sayısal yöntemlerde ise yaklaşık çözümler elde edilir). Let us see an example of such a need from a real-life physical problem.

To make the fulcrum (dayanak/mesnet noktası) (Figure 1) of a bascule bridge (basküllü köprü), a long hollow steel shaft (içi boş çelik şaft) called the trunnion (mafsal/dayanak) is shrink fit into a steel hub. The resulting steel trunnion-hub assembly is then shrink fit into the girder (kiriş) of the bridge.

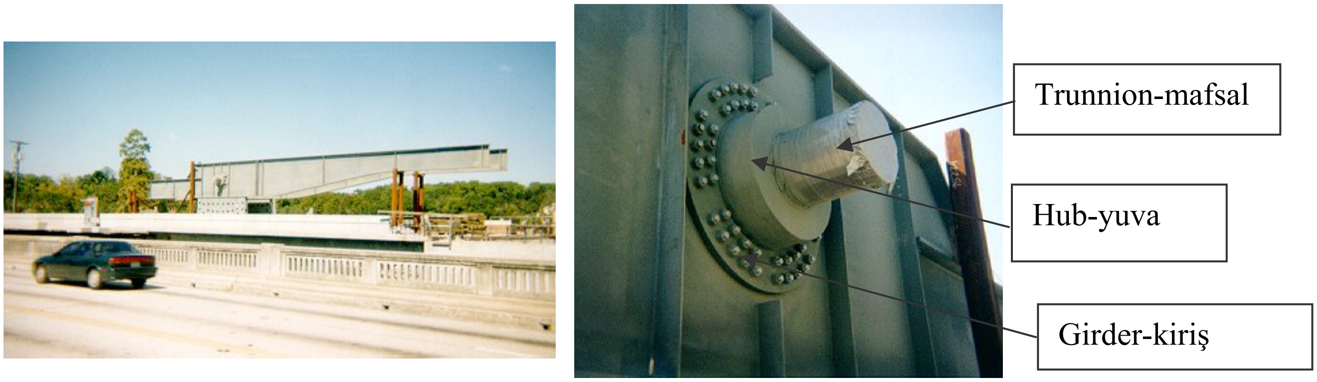


Figure 1 Trunnion-Hub-Girder (THG) assembly (mafsal-yuva-kiriş işlemi).

This is done by first immersing (daldırılmış) the trunnion (mafsal) 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) (mafsal kuru buz/alkol karışımı ortamında kararlı bir sıcaklığa ulaşana kadar soğutulduktan sonra yuvasına/deliğe geçirilir. Deliğin içine tam oturması sağlanır).

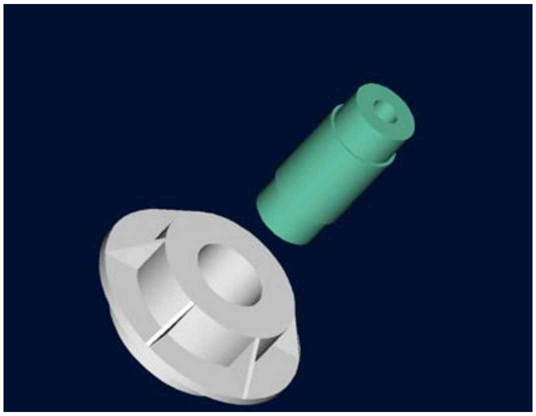


Figure 2 Trunnion (mafsal) slided through the hub after contracting (mafsal uygun haldeyken yuvasına yerleştirilir).

When the trunnion heats up, it expands and creates an interference fit with the hub (mafsal ısındıktan sonra genişler ve yuvasına oturur). 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 (sıkışmıştır). 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 (mafsal yuvada neden sıkışmıştır)? This was because the trunnion had not contracted enough to slide through the hole. Can you find out why?

A hollow trunnion (mafsalın) of outside diameter (dış çapı) 12.363" is to be fitted in a hub of inner diameter (iç çapı) 12.358" (olan hub-yuva içine sokulmak istenmiştir). The trunnion was put in dry ice/alcohol mixture (temperature of the fluid - dry ice/alcohol mixture is $-108°\text{F} = -42.2$ °C) to contract the trunnion so that it can be slid through the hole of the hub. To slide the trunnion without sticking (yapışma olmadan), a diametrical clearance of at least 0.01" is required between the trunnion and the hub (delik ve mafsal arasındaki açıklık). Assuming the room temperature is 80°F (=26.7°C), is immersing the trunnion in dry-ice/alcohol mixture a correct decision? (mafsalı kurubuz/alkol karışımına daldırmak çapını küçültmek adına doğru bir karar mıdır?)

To calculate the contraction (daralma) in the diameter of the trunnion (mafsal), the thermal expansion coefficient at room temperature is used. In that case the reduction ΔD in the outer diameter of the trunnion (mafsal) is

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

where

D = outer diameter of the trunnion,

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

$\Delta T =$ change in temperature,

Given

$D = 12.363''$

$\alpha = 6.47\times10^{-6}\,\text{in/in/°F}$ at $80°\text{F}$

$\Delta T = T_{fluid} - T_{room} = -108 - 80 = -188°\text{F}\ \ (=-122.2°\text{C})$

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

$$\Delta D = (12.363)\left(6.47\times10^{-6}\right)\left(-188\right) = -\,0.01504''$$

So the trunnion (mafsal) 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

= trunnion outside diameter – hub inner diameter + diametric clearance

= mafsalın dış çapı – deliğin iç çapı + arada kalan boşluk

= 12.363 – 12.358 + 0.01= $0.015''$

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 (bu nedenle, Eşitlik (1) ısısal daralma ile ilgili olarak yeterli olacaktır.).

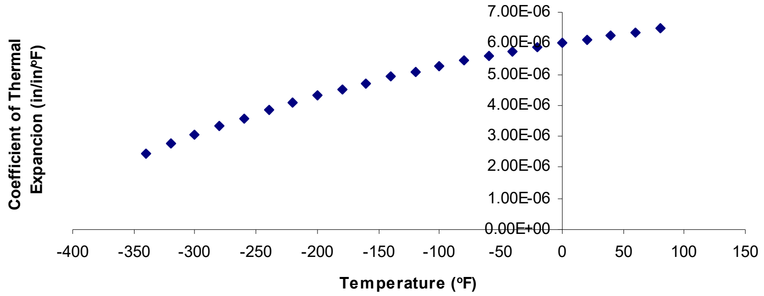


Figure 3 Varying thermal expansion coefficient as a function of temperature for cast (döküm) steel.

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

$$\Delta D = D\int_{T_{room}}^{T_{fluid}} \alpha dT \tag{2}$$

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

$$\alpha = a_0 + a_1 \times T + a_2 \times T^2 \tag{3}$$

Table 1 Instantaneous (anlık) thermal expansion coefficient (ısısal genleşme sabiti) as a function of temperature.

Temperature	**Instantaneous Thermal Expansion**
°F	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

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

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

$$\Delta D = D\int_{T_{room}}^{T_{fluid}} (a_0 + a_1 T + a_2 T^2)dT$$

$$= D[a_0(T_{fluid} - T_{room}) + a_1\frac{(T_{fluid}{}^2 - T_{room}{}^2)}{2} + a_2\frac{(T_{fluid}{}^3 - T_{room}{}^3)}{3}] \qquad (4)$$

which gives

$$\Delta D = 12.363\begin{bmatrix} 6.0150\times10^{-6}\times(-108-80) + 6.1946\times10^{-9}\dfrac{\left((-108)^2-(80)^2\right)}{2} \\ -1.2278\times10^{-12}\dfrac{((-108)^3-(80)^3)}{3}\end{bmatrix}$$

$$= -0.013689"$$

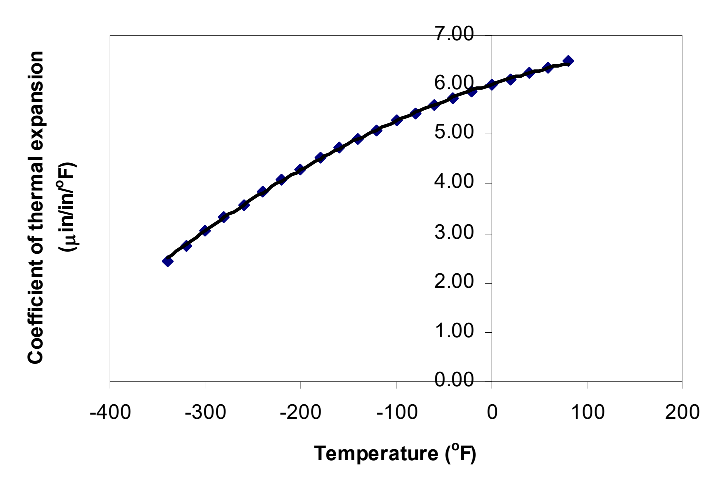


Figure 4 Second order polynomial regression model for coefficient of thermal expansion as a function of temperature.

What do we find here? The contraction (daralma/kasılma) in the trunnion (mafsal) 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$=-321$°F $=-$196.11°C)? Will that cause enough contraction in the trunnion? (mafsal sıvı azot içine daldırılırsa ne olur? Mafsalın yeteri kadar daralmasını sağlar mı?)
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? (mafsalın ısısal genleşme sabitini ikinci derecen polinoma fit etmek yerine eşit olmayan segmanlara sahip yamuk yöntemi kullanılabilir mi? İki sonuç arasındaki bağıl farkı nedir?)
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?

(Burada ikinci dereceden polinom seçilerek regresyon yapılmıştır. Başka polinomlar seçilmesi daha iyi sonuç alabilir misiniz? En uygun polinom derecesi nedir?)

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	Aralık 8, 2016
Web Site	http://numericalmethods.eng.usf.edu

1.1.1 Multiple-Choice Test Chapter 01.01 Introduction to Numerical Methods

1. Solving an engineering problem requires four steps. In order of sequence, the four steps are

(A) Formulate (formüle etmek), solve, interpret, implement
(B) Solve (çözmek), formulate, interpret, implement
(C) formulate, solve, implement (uygulamasını yapmak), interpret
(D) formulate, implement, solve, interpret (yorum yapmak)

2. One of the roots of the equation $x^3 - 3x^2 + x - 3 = 0$ is

(A) -1 (B) 1 (C) $\sqrt{3}$ (D) 3

3. The solution to the set of equations

$$25a + b + c = 25$$
$$64a + 8b + c = 71$$
$$144a + 12b + c = 155$$

most nearly is $(a, b, c) =$

(A) (1,1,1) (B) (1,-1,1) (C) (1,1,-1) (D) does not have a unique solution.

4. The exact integral of $\int_0^{\frac{\pi}{4}} 2\cos 2x dx$ is most nearly

(A) -1.000 (B) 1.000 (C) 0.000 (D) 2.000

5. The value of $\frac{dy}{dx}(1.0)$, given $y = 2\sin(3x)$ most nearly is

(A) -5.9399 (B) -1.980 (C) 0.31402 (D) 5.9918

6. The form of the exact solution of the ordinary differential equation

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

(A) $Ae^{-1.5x}+Be^{x}$ (B) $Ae^{-1.5x}+Be^{-x}$ (C) $Ae^{1.5x}+Be^{-x}$ (D) $Ae^{-1.5x}+Bxe^{-x}$

For a complete solution, refer to the links at the end of the book.

1.2 Chapter 01.02 Measuring Errors (ölçme hataları)

PRE-REQUISITES

1. Know the definition of a secant and first derivative of a function (Primer for Differential Calculus).
2. Understand the representation of trigonometric and transcendental functions as a Maclaurin series (Taylor Series Revisited).

OBJECTIVES

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 (anlamlı haneler).

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.

True Error $=$ True value – 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 $h = 0.3$,

$$f'(2) \approx \frac{f(2+0.3)-f(2)}{0.3}$$

$$= \frac{f(2.3)-f(2)}{0.3}$$

$$= \frac{7e^{0.5(2.3)} - 7e^{0.5(2)}}{0.3}$$

$$= \frac{22.107-19.028}{0.3} \quad = 10.265$$

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

$$f(x) = 7e^{0.5x}$$

$$f'(x) = 7 \times 0.5 \times e^{0.5x} \quad = 3.5e^{0.5x}$$

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

$$f'(2) = 3.5e^{0.5(2)} \quad = 9.5140$$

c) True error is calculated as

E_t= True value – Approximate value

$$= 9.5140 - 10.265 \quad = -0.75061$$

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 $\in_t$ 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,

E_t= True value – Approximate value

$$= 9.5140 - 10.265 = -0.75061$$

Relative true error is calculated as

$$\in_t = \frac{\text{True Error}}{\text{True Value}}$$
$$= \frac{-0.75061}{9.5140} = -0.078895$$

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

$$\in_t = -0.0758895 \times 100\% = -7.58895\%$$

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

$$|\in_t| = |-0.075888|$$
$$= 0.0758895 = 7.58895\%$$

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 $h = 0.3$,

$$f'(2) \approx \frac{f(2+0.3) - f(2)}{0.3}$$
$$= \frac{f(2.3) - f(2)}{0.3}$$
$$= \frac{7e^{0.5(2.3)} - 7e^{0.5(2)}}{0.3}$$

$$= \frac{22.107 - 19.028}{0.3} = 10.265$$

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 $h = 0.15$,

$$f^{'}(2) \approx \frac{f(2+0.15) - f(2)}{0.15}$$

$$= \frac{f(2.15) - f(2)}{0.15}$$

$$= \frac{7e^{0.5(2.15)} - 7e^{0.5(2)}}{0.15}$$

$$= \frac{20.50 - 19.028}{0.15} = 9.8799$$

c) So the approximate error, E_a is

$$E_a = \text{Present Approximation} - \text{Previous Approximation}$$

$$= 9.8799 - 10.265 = -0.38474$$

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 $\in_a$ 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 $h = 0.15$.

$$E_a = \text{Present Approximation} - \text{Previous Approximation}$$

$$= 9.8799 - 10.265$$
$$= -0.38474$$

The relative approximate error is calculated as

$$\in_a = \frac{\text{Approximate Error}}{\text{Present Approximation}}$$
$$= \frac{-0.38474}{9.8799} = -0.038942$$

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

$$\in_a = -0.038942 \times 100\%$$
$$= -3.8942\%$$

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

$$\left|\in_a\right| = |-0.038942| = 0.038942 \text{ or } 3.8942\%$$

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 (yinelemeli yöntem), a user can calculate relative approximate error $\in_a$ at the end of each iteration. The user may pre-specify a minimum acceptable tolerance called the pre-specified tolerance, $\in_s$. If the absolute relative approximate error $\in_a$ is less than or equal to the pre-specified tolerance $\in_s$, that is, $|\in_a| \leq \in_s$, 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, $|\in_a| \leq 0.5 \times 10^{2-m}$ %. (alternatif olarak cevabınızı anlamlı basamak/hane sayısı ile belirleyebilirsiniz. Bu durumda mutlak göreli yaklaşık hata % cinsinden $|\in_a| \leq 0.5 \times 10^{2-m}$ denklemini kullanarak m istenilen anlamlı basamak sayısı belirlenebilir.)

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

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

Using 6 terms, we get the current approximation as

$$e^{0.7} \cong 1 + 0.7 + \frac{0.7^2}{2!} + \frac{0.7^3}{3!} + \frac{0.7^4}{4!} + \frac{0.7^5}{5!} = 2.0136$$

Using 5 terms, we get the previous approximation as

$$e^{0.7} \cong 1 + 0.7 + \frac{0.7^2}{2!} + \frac{0.7^3}{3!} + \frac{0.7^4}{4!} = 2.0122$$

The percentage absolute relative approximate error is

$$|\in_a| = \left|\frac{2.0136-2.0122}{2.0136}\right|\times 100 = 0.069527\%$$

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

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

Q: But what do you mean by significant digits (anlamlı basamak ile ne anlatılmak isteniyor)?
A: Significant digits are important in showing the truth one has in a reported number (anlamlı basamaklar ifade edilmek istenilen sayıyı gerçek anlamında ifade edebilir). 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" (Örneğin birisi size yaşadığınız şehrin nüfusunu sorsa ona çok yaklaşık bir değer -1 milyon- söylersiniz). But if someone was going to give me a $100 for every citizen of the county, I would have to get an exact count (her yurttaş için 100$ verileceği söylense bu durumda kesin rakam -2003 yılı için 1,079,587 kişi şeklinde- belirtmek durumunda kalırsınız). 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 (1 milyon şeklinde söylediğinizde 1 anlamlı basamağı olan nüfus sayısı 1,079,587 rakamında 7 anlamlı nüfus sayısı ile ifade edilir). 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) 1.079×10^3 has four significant digits
(f) 1.0790×10^3 has five significant digits
(g) 1.07900×10^3 has six significant digits

INTRODUCTION, APPROXIMATION AND ERRORS	
Topic	Measuring Errors
Summary	Textbook notes on measuring errors
Major	General Engineering
Authors	Autar Kaw
Date	Aralık 8, 2016
Web Site	http://numericalmethods.eng.usf.edu

1.2.1 Multiple-Choice Test Chapter 01.02 Measuring Errors

1. True error is defined as
   (A) Present Approximation – Previous Approximation
   (B) True Value – Approximate Value
   (C) abs (True Value – Approximate Value)
   (D) abs (Present Approximation – Previous Approximation)

2. The expression for true error in calculating the derivative of $\sin(2x)$ at $x = \pi/4$ by using the approximate expression

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

is

(A) $\frac{h - \cos(2h) - 1}{h}$ (B) $\frac{h - \cos(h) - 1}{h}$ (C) $\frac{1 - \cos(2h)}{h}$ (D) $\frac{\sin(2h)}{h}$

3. The relative approximate error at the end of an iteration to find the root of an equation is 0.004%. The least number of significant digits we can trust in the solution is
   (A) 2 (B) 3 (C) 4 (D) 5

4. The number 0.01850×10^3 has ________ significant digits
   (A) 3 (B) 4 (C) 5 (D) 6

5. The following gas stations were cited for irregular dispensation by the Department of Agriculture. Which one cheated you the most?

Station	Actual gasoline dispensed	Gasoline reading at pump
Ser	9.90	10.00
Cit	19.90	20.00
Hus	29.80	30.00
She	29.95	30.00

(A) Ser (B) Cit (C) Hus (D) She

6. The number of significant digits in the number 219900 is
   (A) 4 (B) 5 (C) 6 (D) 4 or 5 or 6

For a complete solution, refer to the links at the end of the book.

1.3 Chapter 01.03 Sources of Error

PRE-REQUISITES

1. Binary representation of numbers (Binary representation of numbers)
2. Know the definition of a secant and first derivative of a function (Primer for Differential Calculus).
3. Know the Riemann sum concept of integration (Primer for Integral Calculus).
4. Understand the representation of trigonometric and transcendental functions as a Maclaurin series (Taylor Series Revisited).

OBJECTIVES

1. know that there are two inherent (tabiatından) sources of error in numerical methods – round-off (yuvarlama hatası) and truncation error (kesme hatası),
2. recognize the sources of round-off and truncation error, and
3. know the difference between round-off and truncation 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 $\frac{1}{3} - 0.333333 = 0.0000003\overline{3}$. Then there are other numbers that cannot be represented exactly. For example, π 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?

25 şubat 1991'de Irak'tan fırlatılan bir scud füzesi Suudi Arabistan'ın Dahran kentindeki abd'nin askeri kışlasına düşmüş, 28 amerikalı asker ölmüş ve 100'nü yaralamıştır. Abd'nin Patriot savunma sistemi scud'ları izleyememiş ve scud'ları havada tahrip edememiştir. Bu hatanın aslı ne idi?

The Patriot defense system consists of an electronic detection device called the range gate (erim kapısı). 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 0.0001 1001 1001 1001 1001 100. 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 Patriot savunma sistemi menzil aralığı/kapısı denilen elektronik algılama sistemidir. Bu elektronik sistem havada bir scud'un olup olmadığını belli bir alanı tarayarak hesaplamalar yapmaktadır. Sistemin amacı scud'un hızını belirlemek ve radarda tanımlanan füzenin en son anını kayıt etmektir. O an/zaman yani saniyenin 1/10'i 24 bitlik 0.00011001100110011001100 uzunlukta bir veri olarak sisteme kayıt edilmekteydi. Sistemin kendi saatine göre bu kayıtları saniyenin 1/10 zaman aralıklarında arka arkaya tekrarlanmakta ve kayıt altına alınmaktaydı (1/10'ları toplamaktaydı). Onluk sisteme göre her kayıtta

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

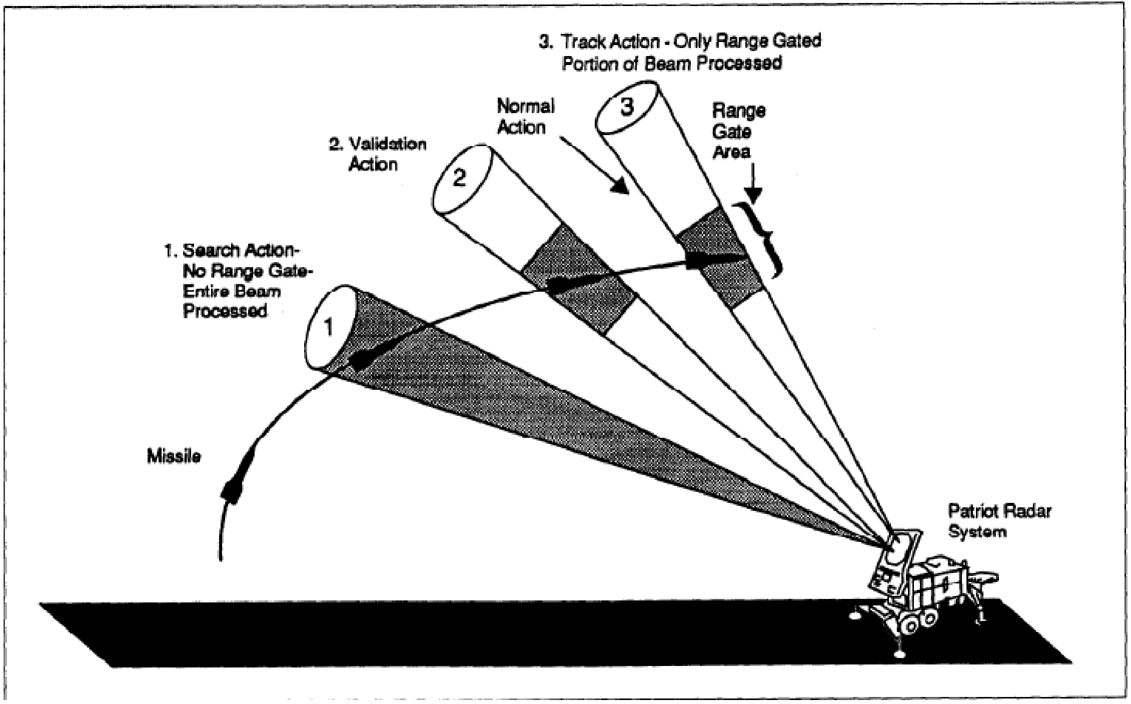
$$\frac{1}{10} - (0\times2^{-1} + 0\times2^{-2} + 0\times2^{-3} + 1\times2^{-4} + \ldots + 1\times2^{-22} + 0\times2^{-23} + 0\times2^{-24})$$
$$= 9.537\times10^{-8}$$

büyüklüğünde bir zaman farkı ortaya çıkmaktaydı. Sistemin akülü güç kaynağı 100 saat (yaklaşık 4 gün) boyunca sürekli kayıt yapıyordu. 100 saat sonunda (başlama anına göre) ortaya çıkan zaman farkı ise aşağıdaki gibidir:

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

$$= 9.537\times10^{-8}\,\frac{\text{s}}{0.1\text{s}}\times100\,\text{hr}\times\frac{3600\text{s}}{1\text{hr}}$$
$$= 0.3433\text{s}$$

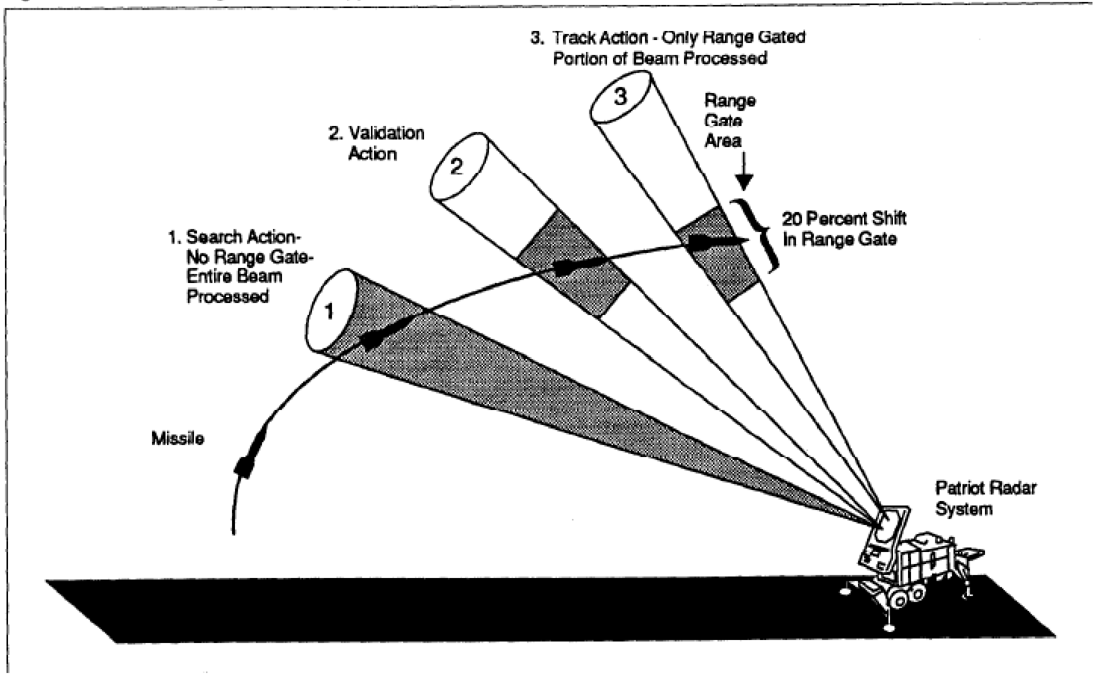
Figure 3: Correctly Calculated Range Gate


3. Track Action - Only Range Gated Portion of Beam Processed
Normal Action
3
Range Gate Area
2. Validation Action
2
1. Search Action- No Range Gate- Entire Beam Processed
1
Missile
Patriot Radar System

Patriot sistemi havada belirli bir açı içindeki kısmın taramasını (araştırma safhasısearch action) yapar (1) ve belirli bir süre (1/10 saniye) sonra aynı genişlikteki açı ile tekrar tarama (onama safhası-validation action range gate-menzil geçit genişliği belirlenir) yapılır (2). Bu taramalardan (yani scud'tan) sinyal gelirse bu sinyaller kayıt edilir. Bu iki taramadan yararlanarak Scud'un olası yerini (range gate area) tespit eder/tanımlar ve patriotları oraya yönlendirir (3) (Figure 3).

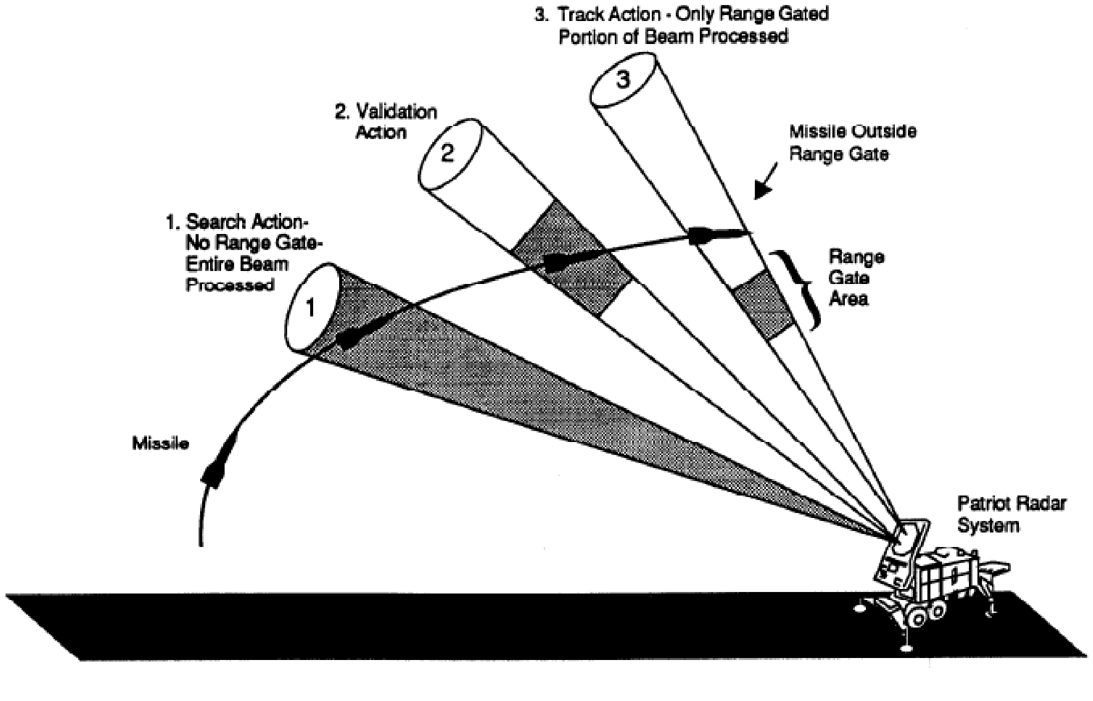
İsrailliler Patriot Projesi Ofisi'nden aldıkları verileri incelediklerinde sistemin 8 saat sürekli çalışınca menzil kapısı aralığında %20'lik bir hata yaptığını anladılar. Bu hata scud'un menzil kapısı aralığının merkezinde olmadığını gösteriyordu. Scud menzil aralığının merkezinde ise sistem başarılı olabiliyordu. Scud füzesi patriot sisteminin menzilinde ise patriotlar hemen ateşleniyordu (Figure 4).

Figure 4: Calculated Range Gate After Approximately 8 Hours


3. Track Action - Only Range Gated Portion of Beam Processed
Range Gate Area
2. Validation Action
20 Percent Shift In Range Gate
1. Search Action- No Range Gate- Entire Beam Processed
1
2
3
Missile
Patriot Radar System

Patriot Projesi Ofis çalışanları patriot sisteminin menzil kapısı aralığında %50'lik sapma durumunda scudları takip etmediğini belirtmişlerdir. %50'lik menzil kapısı merkezleme hatasına 20 saat sonra ulaşılmaktaydı. 20 saat sonra radar scud füzelerinin olası yerini yanlış yerde gösteriyor ve sistem harekete geçmiyord. Scudlar Patriot sisteminin radarlarının menzil kapısı aralığının dışında kalıyordu (Figure 5).

Figure 5: Incorrectly Calculated Range Gate


3. Track Action - Only Range Gated Portion of Beam Processed
2. Validation Action
Missile Outside Range Gate
Range Gate Area
1. Search Action- No Range Gate- Entire Beam Processed
1
2
3
Missile
Patriot Radar System

The shift calculated in the range gate due to 0.3433s was calculated as 583.61m. For the Patriot missile defense system, the target is considered out of range if the shift was going to more than 137m.
Scud'ların hızı yaklaşık 1700 metre/saniye civarındadır (https://pediaview.com/openpedia/Scud). Hesaplamadaki kapı aralıklarından kaynaklanan toplam kayma 0.3433 saniye ise bu 583.61metre'lik bir range-gate (kapı genişliği/menzil geçit genişliği) ile tarama yapılması demektir. Oysa Patriot füzeleri hedef 137m'den büyük kapı genişliği dışında ise etkisiz kalıyordu.
(https://www.ima.umn.edu/~arnold/disasters/patriot.html,
http://fas.org/spp/starwars/gao/im92026.htm)
Patriot sisteminin sürekli çalışması sonucu elde edilen değerler :

Hours	**Seconds**	**Calculated Time (Seconds)**	**Inaccuracy (Seconds)**	**Approximate Shift In Range Gate (Meters)**
0	0	0	0	0
1	3600	3599.9966	.0034	7
3	28800	28799.9725	.0275	55
20a	72000	71999.9313	.0687	137
48	172800	172799.8352	.1648	330
72	259200	259199.7528	.2472	494
100b	360000	359999.6667	.3433	687

a Sistem sürekli çalışırsa 20 saat sonra hedef menzil aralığının dışında kalıyor.
b Alfa aküleri 100 saat boyunca sürekli çalışırsa
United States General Accounting Office, GAO/IMTEC-92-26, February 1992.

Q: What is truncation error (Kesme hatası nedir)?
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

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

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

$$\text{Truncation error} = e^x - \left(1 + x + \frac{x^2}{2!}\right),$$

$$= \frac{x^3}{3!} + \frac{x^4}{4!} + ...$$

But, how can truncation error (kesme hatası) be controlled in this example? We can use the concept (kavram) 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

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

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	$\lvert\in_a\rvert\%$
1	$e^{1.2} = 1 = 1$	-	-
2	$e^{1.2} = 1 + 1.2 = 2.2$	1.2	54.546
3	$e^{1.2} = 1 + 1.2 + \frac{1.2^2}{2!} = 2.92$	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 $\lvert\in_a\rvert <$ 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

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

But since we cannot use $\Delta x \to 0$, we have to use a finite value of Δx, 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 Δx as opposed to a $\Delta x \to 0$.

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,

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

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

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

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

$$f'(3) = 2 \times 3$$
$$= 6$$

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 Δx?

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

$$\int_3^9 x^2 dx = \left[\frac{x^3}{3}\right]_3^9$$

$$= \left[\frac{9^3 - 3^3}{3} \right]$$

$$= 234$$

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

$$\int_3^9 x^2 dx = (x^2)\Big|_{x=3}(6-3) + (x^2)\Big|_{x=6}(9-6)$$

$$= (3^2)3 + (6^2)3$$

$$= 27 + 108$$

$$= 135$$

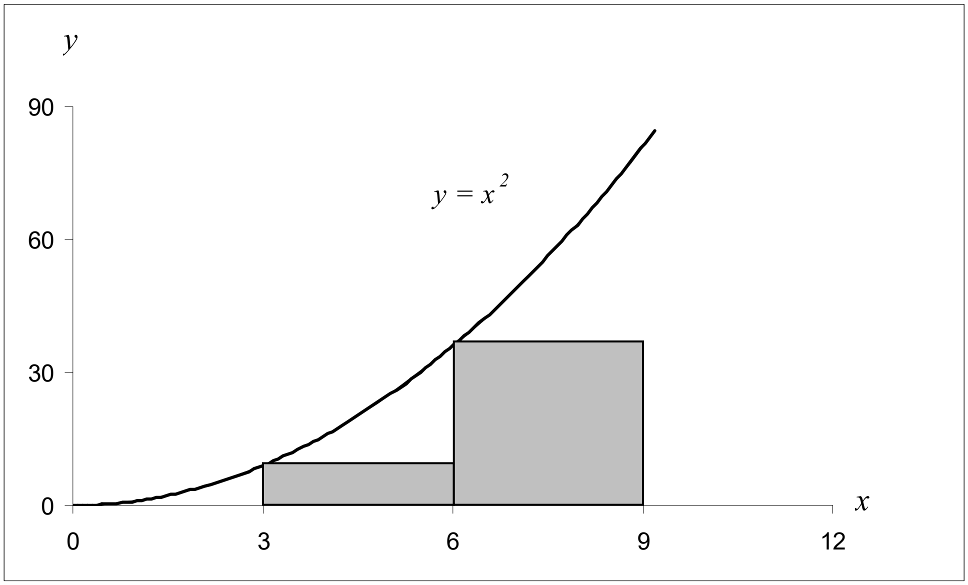


Figure 2 Plot of $y = x^2$ showing the approximate area under the curve from $x = 3$ to $x = 9$ using two rectangles.

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?

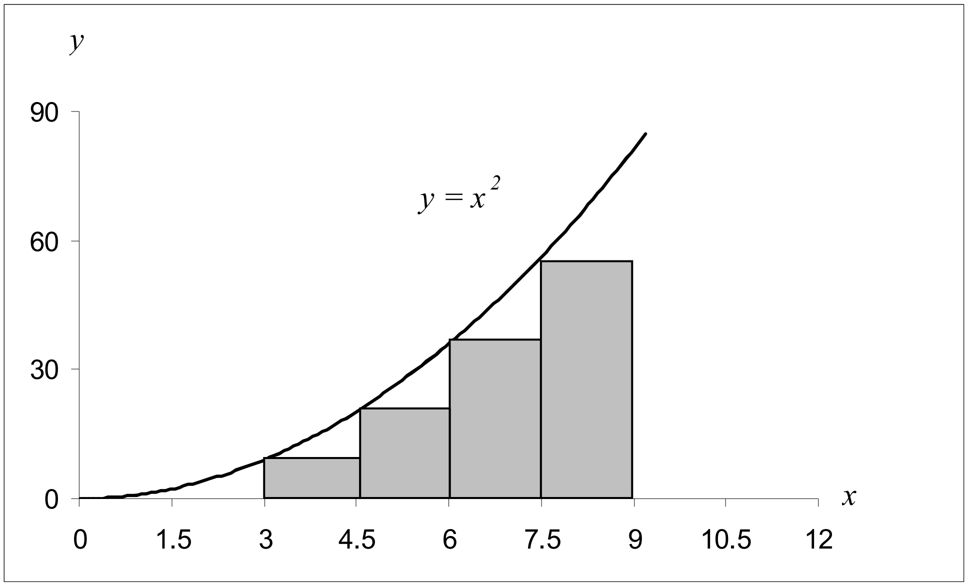


Figure 3 Plot of $y = x^2$ showing the approximate area under the curve from $x = 3$ to $x = 9$ using four rectangles.

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	Aralık 8, 2016
Web Site	http://numericalmethods.eng.usf.edu

1.3.1 Multiple-Choice Test Chapter 01.03 Sources of Error

1. Truncation error is caused by approximating
(A) irrational numbers (B) fractions (C) rational numbers (D) exact mathematical procedures

2. A computer that represents only 4 significant digits with chopping would calculate 66.666*33.333 as
(A) 2220 (B) 2221 (C) 2221.17778 (D) 2222

3. A computer that represents only 4 significant digits with rounding would calculate 66.666*33.333 as
(A) 2220 (B) 2221 (C) 2221.17778 (D) 2222

4. The truncation error in calculating $f'(2)$ for $f(x)=x^2$ by

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

with $h = 0.2$ is

(A) –0.2 (B) 0.2 (C) 4.0 (D) 4.2

5. The truncation error in finding $\int_{-3}^{9} x^3 dx$ using LRAM (left end point Riemann approximation) with equally portioned points $-3 < 0 < 3 < 6 < 9$ is

(A) 648 (B) 756 (C) 972 (D) 1620

6. The number 1/10 is registered in a fixed 6 bit-register with all bits used for the fractional part. The difference gets accumulated every 1/10th of a second for one day. The magnitude of the accumulated difference is

(A) 0.082 (B) 135 (C) 270 (D) 5400

For a complete solution, refer to the links at the end of the book.

1.4 Chapter 01.04 Binary Representation

PRE-REQUISITES

1. Long Division

OBJECTIVES

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

After reading this chapter, you should be able to:

3. *convert a base-10 real number to its binary representation,*
4. *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

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

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	$0 = a_2$
½	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.

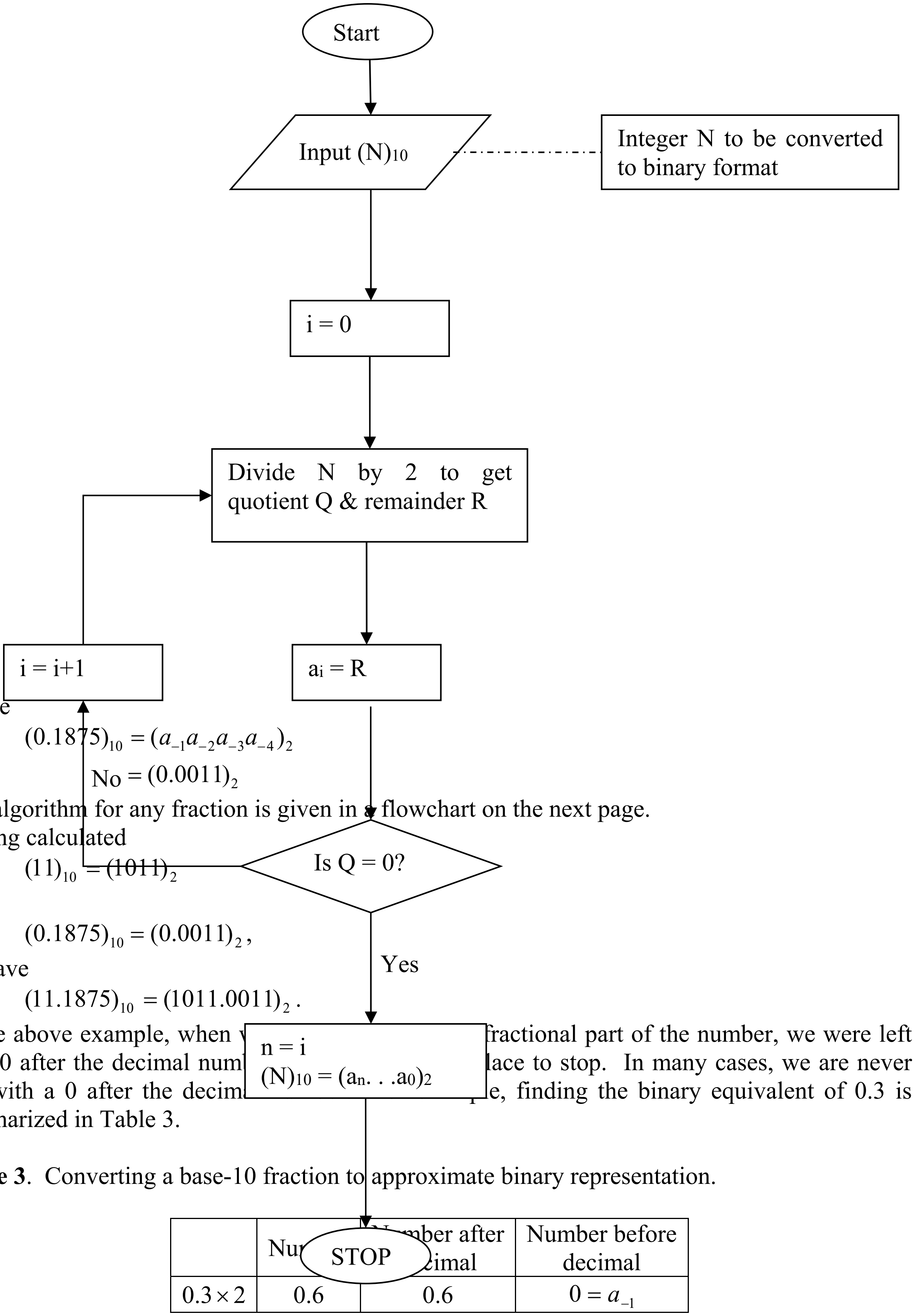
Now let us look at the decimal part, that is, 0.1875.

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
0.1875×2	0.375	0.375	$0 = a_{-1}$
0.375×2	0.75	0.75	$0 = a_{-2}$
0.75×2	1.5	0.5	$1 = a_{-3}$
0.5×2	1.0	0.0	$1 = a_{-4}$


Start
Input $(N)_{10}$
Integer N to be converted to binary format
i = 0
Divide N by 2 to get quotient Q & remainder R
i = i+1
a_i = R
Is Q = 0?
Yes
n = i
$(N)_{10} = (a_n. . .a_0)_2$
STOP

Hence

$$(0.1875)_{10} = (a_{-1}a_{-2}a_{-3}a_{-4})_2$$

$$\text{No} = (0.0011)_2$$

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 w fractional part of the number, we were left with 0 after the decimal numl lace to stop. In many cases, we are never left with a 0 after the decim e, finding the binary equivalent of 0.3 is summarized in Table 3.

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

	Nu	ber after imal	Number before decimal
0.3×2	0.6	0.6	$0 = a_{-1}$

0.6×2	1.2	0.2	$1=a_{-2}$
0.2×2	0.4	0.4	$0=a_{-3}$
0.4×2	0.8	0.8	$0=a_{-4}$
0.8×2	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 behinds this process of converting a decimal number to binary format?

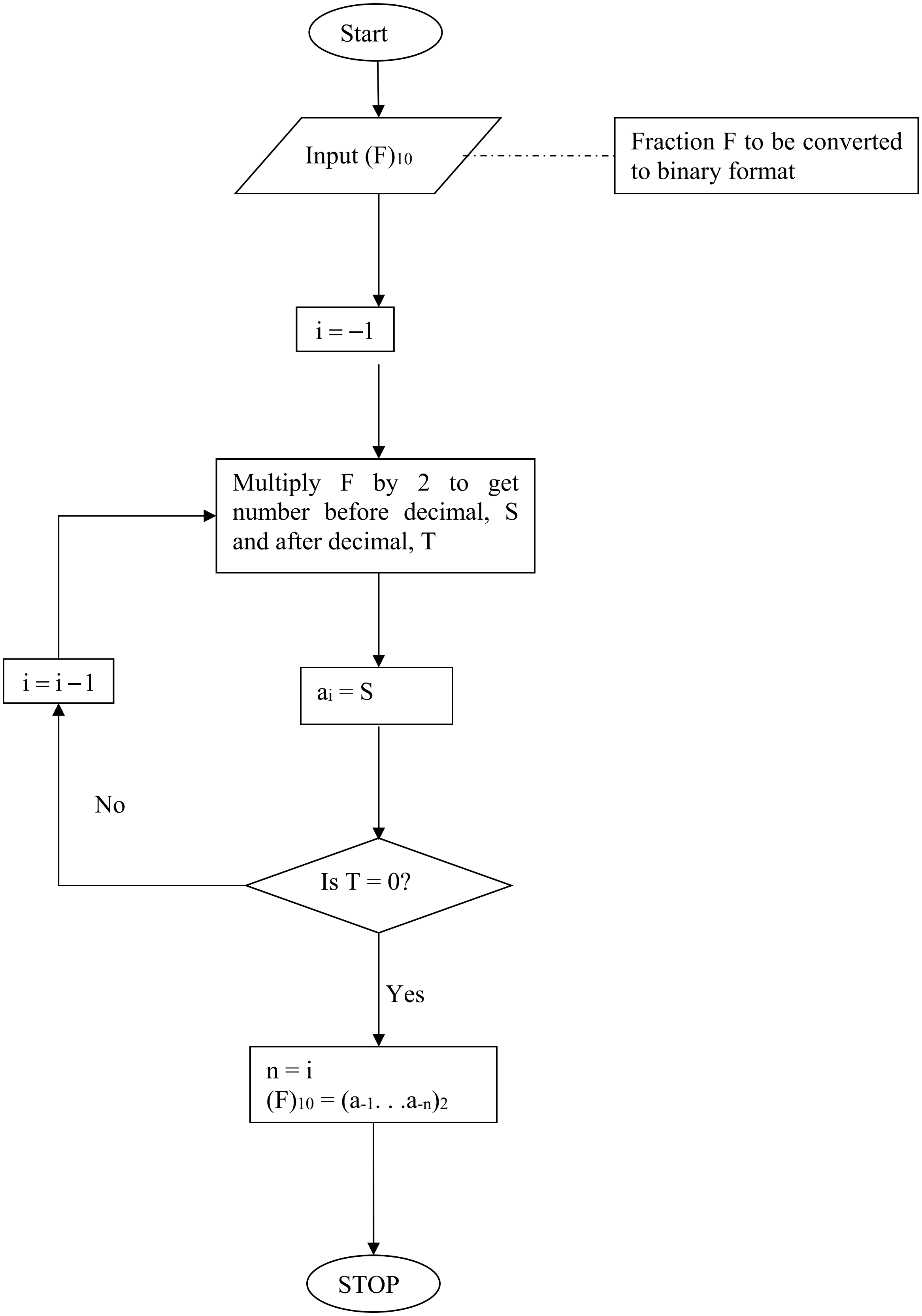
A: Let z be the decimal number written as

$$z = x.y$$

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


Start
Input (F)10
Fraction F to be converted to binary format
i = −1
Multiply F by 2 to get number before decimal, S and after decimal, T
i = i − 1
ai = S
No
Is T = 0?
Yes
n = i
(F)10 = (a-1. . .a-n)2
STOP

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

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

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

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

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

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

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

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
0.875×2	1.75	0.75	$1 = a_{-1}$
0.75×2	1.5	0.5	$1 = a_{-2}$
0.5×2	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	Aralık 8, 2016
Web Site	http://numericalmethods.eng.usf.edu

1.4.1 Multiple-Choice Test Chapter 01.04 Binary Representation

1. $(25)_{10} = (?)_2$

   (A) 100110 (B) 10011 (C) 11001 (D) 110010

2. $(1101)_2 = (?)_{10}$

   (A) 3 (B) 13 (C) 15 (D) 26

3. $(25.375)_{10} = (?.?)_2$

   (A) 100110.011 (B) 11001.011 (C) 10011.0011 (D) 10011.110

4. Representing $\sqrt{2}$ in a fixed point register with 2 bits for the integer part and 3 bits for the fractional part gives a round-off error of most nearly

   (A) -0.085709 (B) 0.0392 (C) 0.1642 (D) 0.2892

5. An engineer working for the Department of Defense is writing a program that transfers non-negative real numbers to integer format. To avoid overflow problems, the maximum non-negative integer that can be represented in a 5-bit integer word is

   (A) 16 (B) 31 (C) 63 (D) 64

6. For a numerically controlled machine, integers need to be stored in a memory location. The minimum number of bits needed for an integer word to represent all integers between 0 and 1024 is

   (A) 8 (B) 9 (C) 10 (D) 11

For a complete solution, refer to the links at the end of the book.

1.5 Chapter 01.05 Floating Point Representation

PRE-REQUISITES

1. Know how to represent numbers in binary format (Binary representation of numbers).
2. Know the definition of true error (Measuring Errors)

OBJECTIVES

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.

After reading this chapter, you should be able to:

2. *convert a base-10 number to a binary floating point representation,*
3. *convert a binary floating point number to its equivalent base-10 number,*
4. *understand the IEEE-754 specifications of a floating point representation in a typical computer,*
5. *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

$$\varepsilon_t = \frac{-0.004}{256.786} \times 100 = -0.001558\% .$$

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

$$\varepsilon_t = \frac{-0.004}{3.546} \times 100 = -0.11280\% .$$

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 -2.56789×10^2.

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

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

or for a number y,

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

Where

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

$m = \text{mantissa}, 1 \le m < 10$

$e = \text{integer exponent (also called ficand)}$

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 9.999×10^9. 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 9.999×10^9.

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

2	5	6	8	2

Another example, the number 576329.78 would be represented as 5.763×10^5 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 $576329.78-5.763\times10^5 = 29.78$, 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

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

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

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

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

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

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

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

Now we find the binary equivalent of the fractional part

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

$\underline{0.05664\times 2}$

$\underline{0.11328\times 2}$

$\underline{0.22656\times 2}$

$$\underline{0.45312\times 2}$$
$$\underline{0.90624\times 2}$$
$$\underline{1.81248\times 2}$$
$$\underline{1.62496\times 2}$$
$$\underline{1.24992\times 2}$$
$$\underline{0.49984\times 2}$$
$$\underline{0.99968\times 2}$$
$$\underline{1.99936}$$

Hence

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

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

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

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

A: The machine epsilon, $\in_{mach}$ 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

$$\in_{mach} = 0.0625\,.$$

The machine epsilon, $\in_{mach}$ 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, $\in_{mach}$ 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

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

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

$$|\varepsilon_t| = \left| \frac{0.02832 - 0.02734375}{0.02832} \right| = 0.034472$$

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

  $y = \sigma \times (1.a_1 a_2 \cdots a_{23}) \cdot 2^e$

where

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

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

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 $(11111111)_2 = 255$, 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 $-127 \le e \le 128$.
- 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, $-127 \le e \le 127$. 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 $-126 \le e \le 127$. 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

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

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

  $$\left(1 \times 2^0 + 0 \times 2^{-1} + 0 \times 2^{-2} + \cdots + 0 \times 2^{-22} + 0 \times 2^{-23}\right) \times 2^{-126} = 1.18 \times 10^{-38}$$
- Since 23 bits are used for the mantissa, the machine epsilon,

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

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	Aralık 8, 2016
Web Site	http://numericalmethods.eng.usf.edu

1.5.1 Multiple-Choice Test Chapter 01.05 Floating Point Representation

1. A hypothetical computer stores real numbers in floating point format in 8-bit words. The first bit is used for the sign of the number, the second bit for the sign of the exponent, the next two bits for the magnitude of the exponent, and the next four bits for the magnitude of the mantissa. The number $e \cong 2.718$ in the 8-bit format is ()
(A) 00010101 (B) 00011010 (C) 00010011 (D) 00101010

2. A hypothetical computer stores real numbers in floating point format in 8-bit words. The first bit is used for the sign of the number, the second bit for the sign of the exponent, the next two bits for the magnitude of the exponent, and the next four bits for the magnitude of the mantissa. The number that $(10100111)_2$ represented in the above given 8-bit format is
(A) -5.75 (B) -2.875 (C) -1.75 (D) -0.359375

3. A hypothetical computer stores floating point numbers in 8-bit words. The first bit is used for the sign of the number, the second bit for the sign of the exponent, the next two bits for the magnitude of the exponent, and the next four bits for the magnitude of the mantissa. The machine epsilon is most nearly
(A) 2^{-8} (B) 2^{-4} (C) 2^{-3} (D) 2^{-2}

4. A machine stores floating point numbers in 7-bit word. The first bit is used for the sign of the number, the next three for the biased exponent and the next three for the magnitude of the mantissa. The number $(0010110)_2$ represented in base-10 is
(A) 0.375 (B) 0.875 (C) 1.5 (D) 3.5

5. A machine stores floating point numbers in 7-bit words. The first bit is stored for the sign of the number, the next three for the biased exponent and the next three for the magnitude of the mantissa. You are asked to represent 33.35 in the above word. The error you will get in this case would be
(A) underflow (B) overflow (C) NaN (D) No error will be registered.

6. A hypothetical computer stores floating point numbers in 9-bit words. The first bit is used for the sign of the number, the second bit for the sign of the exponent, the next three bits for the magnitude of the exponent, and the next four bits for the magnitude of the mantissa. Every second, the error between 0.1 and its binary representation in the 9-bit word is accumulated. The accumulated error after one day most nearly is
(A) 0.002344 (B) 20.25 (C) 202.5 (D) 8640

For a complete solution, refer to the links at the end of the book.

1.6 Chapter 01.06 Propagation of Errors

PRE-REQUISITES

1. Know the definition of first derivative of a function (Primer for Differential Calculus).
2. Know how to find partial derivatives

OBJECTIVES

1. Find how errors propagate in arithmetic operations.
2. Quantify the errors based on individual components of an arithmetic operation or a mathematical formula.

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 $X_1, X_2, X_3, ..., X_{n-1}, X_n$, then the maximum possible value of the error in f is

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

Example 2

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

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

where

F =axial force in the member, N

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

E =Young's modulus, Pa

Given

$F = 72 \pm 0.9$ N

$h = 4 \pm 0.1$ mm

$E = 70 \pm 1.5$ GPa

Find the maximum possible error in the measured strain.

Solution

$$\in = \frac{72}{(4\times10^{-3})^2(70\times10^9)}$$

$$= 64.286\times10^{-6}$$

$$= 64.286\mu$$

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

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

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

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

$$\Delta\varepsilon = \left|\frac{1}{h^2 E}\Delta F\right| + \left|\frac{2F}{h^3 E}\Delta h\right| + \left|\frac{F}{h^2 E^2}\Delta E\right|$$

$$= \left|\frac{1}{(4\times10^{-3})^2(70\times10^9)}\times0.9\right| + \left|\frac{2\times72}{(4\times10^{-3})^3(70\times10^9)}\times0.0001\right|$$

$$+ \left|\frac{72}{(4\times10^{-3})^2(70\times10^9)^2}\times1.5\times10^9\right|$$

$$= 8.0357\times10^{-7} + 3.2143\times10^{-6} + 1.3776\times10^{-6}$$

$$= 5.3955\times10^{-6}$$

$$= 5.3955\mu$$

Hence

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

implying that the axial strain, $\in$ is between 58.8905μ and 69.6815μ

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

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

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	
Topic	Propagation of Errors
Summary	Textbook notes on how errors propagate in arithmetic and function evaluations
Major	All Majors of Engineering
Authors	Autar Kaw
Last Revised	Aralık 8, 2016
Web Site	http://numericalmethods.eng.usf.edu

1.6.1 Multiple-Choice Test Chapter 01.06 Propagation of Errors

1. If $A = 3.56 \pm 0.05$ and $B = 3.25 \pm 0.04$, the values of $A + B$ are
   (A) $6.81 \leq A + B \leq 6.90$ (B) $6.72 \leq A + B \leq 6.90$
   (C) $6.81 \leq A + B \leq 6.81$ (D) $6.71 \leq A + B \leq 6.91$

2. A number A is correctly rounded to 3.18 from a given number B. Then $|A - B| \leq C$, where C is
   (A) 0.005 (B) 0.01 (C) 0.18 (D) 0.09999

3. Two numbers A and B are approximated as C and D, respectively. The relative error in $C \times D$ is given by
   (A) $\left|\left(\frac{A-C}{A}\right)\right| \times \left|\left(\frac{B-D}{B}\right)\right|$
   (B) $\left|\left(\frac{A-C}{A}\right)\right| + \left|\left(\frac{B-D}{B}\right)\right| + \left|\left(\frac{A-C}{A}\right)\right| \times \left|\left(\frac{B-D}{B}\right)\right|$
   (C) $\left|\left(\frac{A-C}{A}\right)\right| + \left|\left(\frac{B-D}{B}\right)\right| - \left|\left(\frac{A-C}{A}\right)\right| \times \left|\left(\frac{B-D}{B}\right)\right|$
   (D) $\left(\frac{A-C}{A}\right) - \left(\frac{B-D}{B}\right)$

4. The formula for normal strain in a longitudinal bar is given by $\in = \frac{F}{AE}$ where
   F = normal force applied
   A = cross-sectional area of the bar
   E = Young's modulus
   If $F = 50 \pm 0.5\,\text{N}$, $A = 0.2 \pm 0.002\ \text{m}^2$, and $E = 210 \times 10^9 \pm 1 \times 10^9\ \text{Pa}$, the maximum error in the measurement of strain is
   (A) 10^{-12} (B) 2.95×10^{-11} (C) 1.22×10^{-9} (D) 1.19×10^{-9}

5. A wooden block is measured to be 60 mm by a ruler and the measurements are considered to be good to 1/4th of a millimeter. Then in the measurement of 60 mm, we have ________ significant digits.
   (A) 0 (B) 1 (C) 2 (D) 3

6. In the calculation of the volume of a cube of nominal size 5", the uncertainty in the measurement of each side is 10%. The uncertainty in the measurement of the volume would be
   (A) 5.477% (B) 10.00% (C) 17.32% (D) 30.00%

For a complete solution, refer to the links at the end of the book.

1.7 Chapter 01.07 Taylor's Theorem Revisited

PRE-REQUISITES

1. Know the definition of derivatives of a function (Primer for Differential Calculus).
2. Know the derivatives of trigonometric and transcendental functions (Primer for Differential Calculus).

OBJECTIVES

1. understand the basics of Taylor's theorem,
2. write transcendental (karışık) 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.

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 (yönleri) of numerical methods that it is imperative to devote (uzun süreçte) a separate chapter to its review and applications (Taylor serilerinin bütün yönleriyle sayısal yöntemlerde kullanıldığını görmek için ayrıntılı özetinin ve uygulamalarının ayrı bir bölüm halinde hazırlanması gerekmektedir). For example, you must have come across expressions such as (aşağıdaki ifadelerle karşılaşabilirsiniz)

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

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

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

All the above expressions are actually a special case (özel durumu) 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 (toplama, çıkarma, çarpma ve bölme gibi basit aritmetik işlemleriyle fonksiyonların – iki sayı sisteminde işlem yapılabilecek şekilde - yazılması sağlanır).

Example 1

Find the value of $e^{0.25}$ using the first five terms of the Maclaurin series (Maclaurin serisinin ilk beş terimi ile $e^{0.25}$ değerini hesaplayınız).

Solution

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

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

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

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 $f(x)$

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

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^{th} order Taylor polynomial, then $1^{st}, 2^{nd}, \ldots, n^{th}$ derivatives need to exist and be continuous in the closed interval $[x, x+h]$, while the $(n+1)^{th}$ 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

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

$$
\begin{aligned}
& x = \frac{\pi}{2} \\
& h = 0.42920 \\
& f(x) = \sin(x),\ f\left(\frac{\pi}{2}\right) = \sin\left(\frac{\pi}{2}\right) = 1 \\
& f'(x) = \cos(x),\ f'\left(\frac{\pi}{2}\right) = 0 \\
& f''(x) = -\sin(x),\ f''\left(\frac{\pi}{2}\right) = -1 \\
& f'''(x) = -\cos(x),\ f'''\left(\frac{\pi}{2}\right) = 0 \\
& f''''(x) = \sin(x),\ f''''\left(\frac{\pi}{2}\right) = 1
\end{aligned}
$$

Hence

$$f\left(\frac{\pi}{2} + h\right) = f\left(\frac{\pi}{2}\right) + f'\left(\frac{\pi}{2}\right)h + f''\left(\frac{\pi}{2}\right)\frac{h^2}{2!} + f'''\left(\frac{\pi}{2}\right)\frac{h^3}{3!} + f''''\left(\frac{\pi}{2}\right)\frac{h^4}{4!} + \cdots$$

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

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 2π and then can use the periodicity of $\sin(x)$, that is $\sin(x) = \sin(x + 2n\pi), n = 1,2,\ldots$ to calculate the value of $\sin(x)$ at any other point.

Example 3

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

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 $x = 0$.

$$f(x) = \sin(x),\ f(0) = 0$$
$$f'(x) = \cos(x), f'(0) = 1$$
$$f''(x) = -\sin(x), f''(0) = 0$$
$$f'''(x) = -\cos(x), f'''(0) = -1$$
$$f''''(x) = \sin(x), f''''(0) = 0$$
$$f'''''(x) = \cos(x), f'''''(0) = 1$$

Using the Taylor series now,

$$f(x+h) = f(x) + f'(x)h + f''(x)\frac{h^2}{2!} + f'''(x)\frac{h^3}{3!} + f''''(x)\frac{h^4}{4} + f'''''(x)\frac{h^5}{5} + \cdots$$
$$f(0+h) = f(0) + f'(0)h + f''(0)\frac{h^2}{2!} + f'''(0)\frac{h^3}{3!} + f''''(0)\frac{h^4}{4} + f'''''(0)\frac{h^5}{5} + \cdots$$
$$f(h) = f(0) + f'(0)h + f''(0)\frac{h^2}{2!} + f'''(0)\frac{h^3}{3!} + f''''(0)\frac{h^4}{4} + f'''''(0)\frac{h^5}{5} + \cdots$$
$$= 0 + 1(h) - 0\frac{h^2}{2!} - 1\frac{h^3}{3!} + 0\frac{h^4}{4} + 1\frac{h^5}{5} + \cdots$$
$$= h - \frac{h^3}{3!} + \frac{h^5}{5!} + \cdots$$

So

$$f(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots$$
$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots$$

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

$$f(x+h) = f(x) + f'(x)h + f''(x)\frac{h^2}{2!} + f'''(x)\frac{h^3}{3!} + \cdots$$
$$x = 4$$
$$h = 6 - 4$$
$$= 2$$

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

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

$$f(6) = 125 + 74(2) + 30\left(\frac{2^2}{2!}\right) + 6\left(\frac{2^3}{3!}\right)$$

$$= 125 + 148 + 60 + 8 = 341$$

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

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

$$e^1 \approx 1 + 1 + \frac{1^2}{2!} + \frac{1^3}{3!} = 2.66667$$

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

$$E_t = \frac{x^4}{4!} + \frac{x^5}{5!} + \cdots$$

$$= \frac{1^4}{4!} + \frac{1^5}{5!} + \cdots \cong 0.0516152$$

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

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

where

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

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 $(x = 0, n = 3)$

$$R_3(0+1) = \frac{(1)^{3+1}}{(3+1)!} f^{(3+1)}(c)$$

$$= \frac{1}{4!} f^{(4)}(c)$$

$$= \frac{e^c}{24}$$

Since

$$x < c < x+h$$
$$0 < c < 0+1$$
$$0 < c < 1$$

The error is bound between

$$\frac{e^0}{24} < R_3(1) < \frac{e^1}{24}$$

$$\frac{1}{24} < R_3(1) < \frac{e}{24}$$

$$0.041667 < R_3(1) < 0.113261$$

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

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

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},

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

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

$$(n+1)! > 10^6 \times 3$$ (as we do not know the value of e 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	Aralık 8, 2016
Web Site	http://numericalmethods.eng.usf.edu

1.7.1 Multiple-Choice Test Chapter 01.07 Taylors Series Revisited

1. The coefficient of the x^5 term in the Maclaurin polynomial for $\sin(2x)$ is

(A) 0 (B) 0.0083333 (C) 0.016667 (D) 0.26667

2. Given $f(3)=6$, $f'(3)=8$, $f''(3)=11$, and that all other higher order derivatives of $f(x)$ are zero at $x=3$, and assuming the function and all its derivatives exist and are continuous between $x=3$ and $x=7$, the value of $f(7)$ is

(A) 38.000 (B) 79.500 (C) 126.00 (D) 331.50

3. Given that $y(x)$ is the solution to $\frac{dy}{dx}=y^3+2$, $y(0)=3$ the value of $y(0.2)$ from a second order Taylor polynomial written around $x=0$ is

(A) 4.400 (B) 8.800 (C) 24.46 (D) 29.00

4. The series $\sum_{n=0}^{\infty}(-1)^n\frac{x^{2n}}{(2n)!}4^n$ is a Maclaurin series for the following function

(A) $\cos(x)$ (B) $\cos(2x)$ (C) $\sin(x)$ (D) $\sin(2x)$

5. The function

$$erf(x)=\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$$

is called the error function. It is used in the field of probability and cannot be calculated exactly for finite values of x. However, one can expand the integrand as a Taylor polynomial and conduct integration. The approximate value of $erf(2.0)$ using the first three terms of the Taylor series around $t=0$ is

(A) -0.75225 (B) 0.99532 (C) 1.5330 (D) 2.8586

6. Using the remainder of Maclaurin polynomial of n^{th} order for $f(x)$ defined as

$$R_n(x)=\frac{x^{n+1}}{(n+1)!}f^{(n+1)}(c), \quad n\geq 0,\ 0\leq c\leq x$$

the least order of the Maclaurin polynomial required to get an absolute true error of at most 10^{-6} in the calculation of $\sin(0.1)$ is (do not use the exact value of $\sin(0.1)$ or $\cos(0.1)$ to find the answer, but the knowledge that $|\sin(x)| \leq 1$ and $|\cos(x)| \leq 1$).

(A) 3 (B) 5 (C) 7 (D) 9

For a complete solution, refer to the links at the end of the book.

2 Chapter 02.01 Primer on Differentiation (Diferensiyelin başlangıç bilgileri)

PRE-REQUISITES

1. High School Algebra, Trigonometry and Geometry.

OBJECTIVES

1. understand the basics of differentiation,
2. relate the slopes (eğim) of the secant line (kiriş çizgisi) and tangent line (teğet çizgisi) 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.

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 (bu başlangıçta, girişte), we will review the concepts of differentiation you learned in calculus (matematik derslerinden gördüğünüz diferensiyel kavramları konusunda temel düzeyde kavramlar gözden geçirilecektir). 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.

2.1 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.

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.

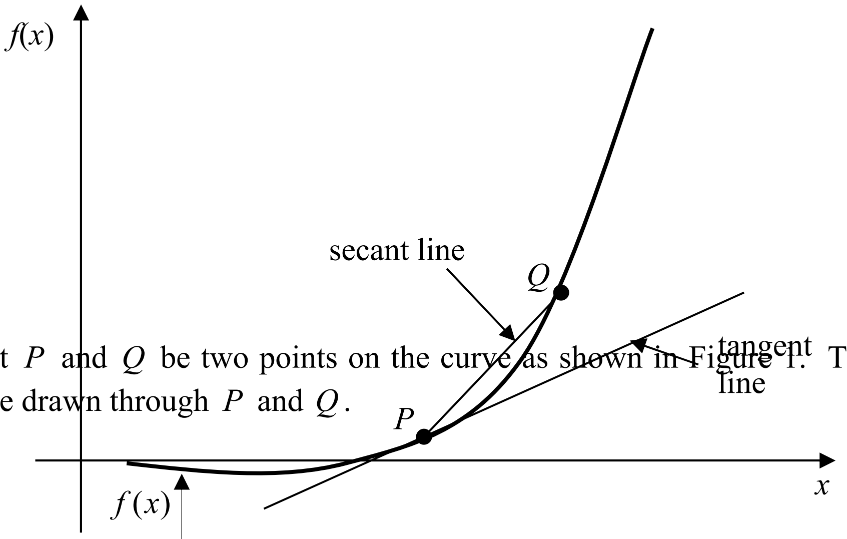


Figure 1 Function curve with tangent and secant lines.

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

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

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

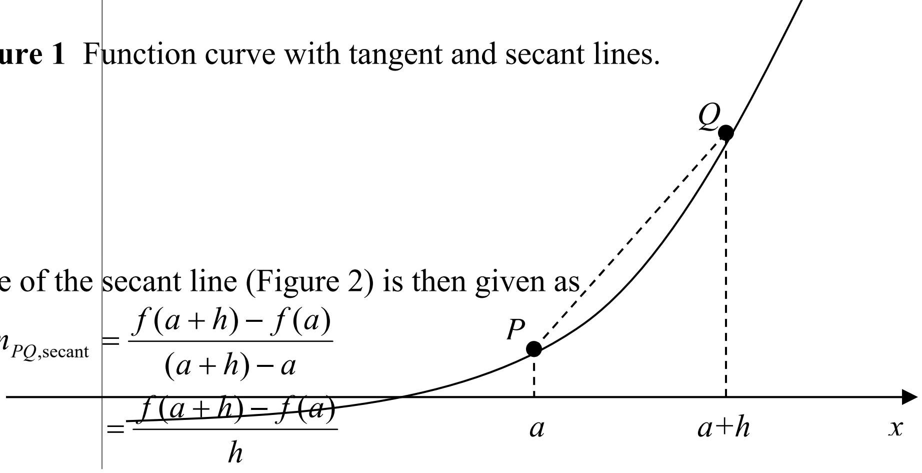


Figure 2 Calculation of the secant line.

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 $h \to 0$.

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

Example 1
Find the slope of the secant line of the curve $y = 4x^2$ between points (3,36) and (5,100).

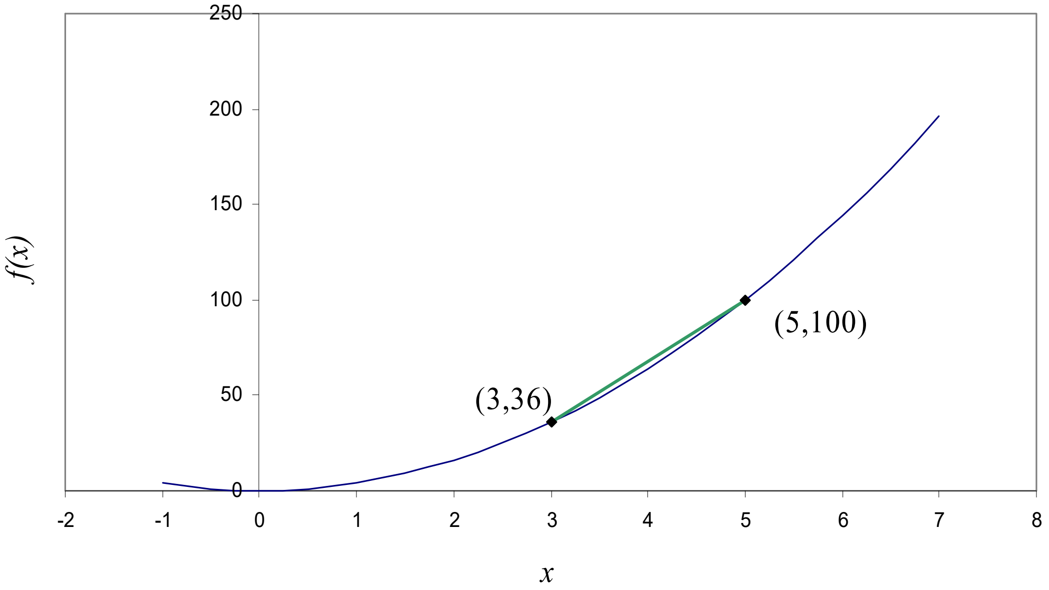


Figure 3 Calculation of the secant line for the function $y = 4x^2$.

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

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

$$= 24$$

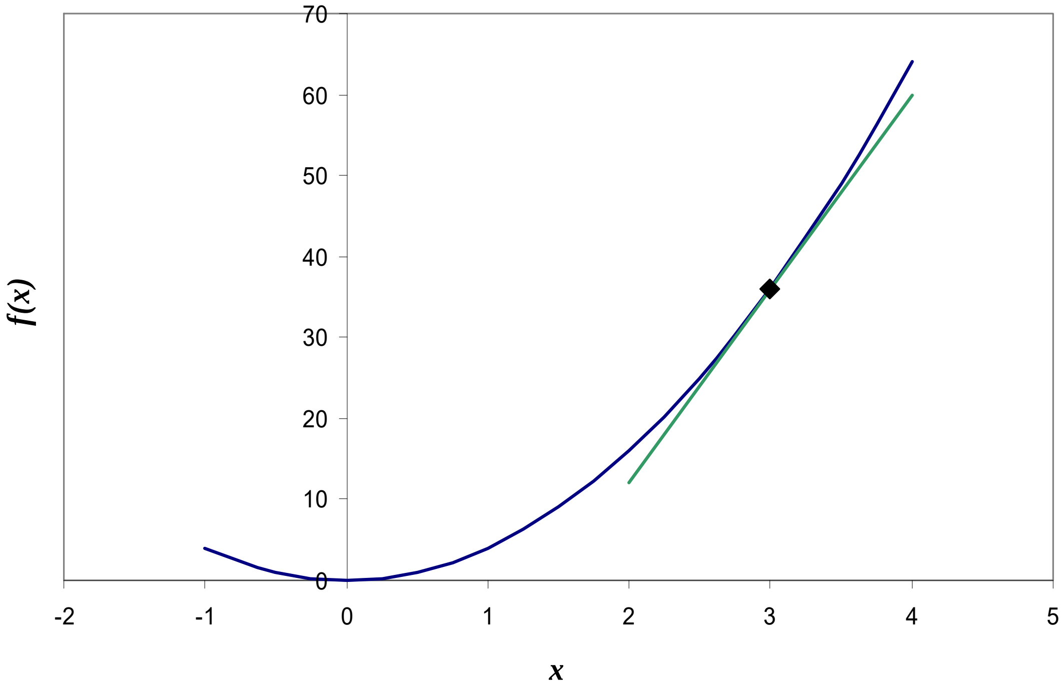


Figure 4 Calculation of the tangent line in the function $y = 4x^2$.

The slope of the tangent line is

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

Derivative of a Function

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

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

Example 3

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

Solution

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

Example 4

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

Solution

$$\begin{aligned}
f'\left(\frac{\pi}{4}\right) &= \lim_{h\to 0}\frac{f\left(\frac{\pi}{4}+h\right)-f\left(\frac{\pi}{4}\right)}{h} \\
&= \lim_{h\to 0}\frac{sin\left(2\left(\frac{\pi}{4}+h\right)\right)-sin\left(2\left(\frac{\pi}{4}\right)\right)}{h} \\
&= \lim_{h\to 0}\frac{sin\left(\frac{\pi}{2}+2h\right)-sin\left(\frac{\pi}{2}\right)}{h} \\
&= \lim_{h\to 0}\frac{sin\left(\frac{\pi}{2}\right)cos(2h)+cos\left(\frac{\pi}{2}\right)sin(2h)-sin\left(\frac{\pi}{2}\right)}{h} \\
&= \lim_{h\to 0}\frac{cos(2h)+0-1}{h} \\
&= \lim_{h\to 0}\frac{cos(2h)-1}{h} \\
&= 0
\end{aligned}$$

from knowing that

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

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

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

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

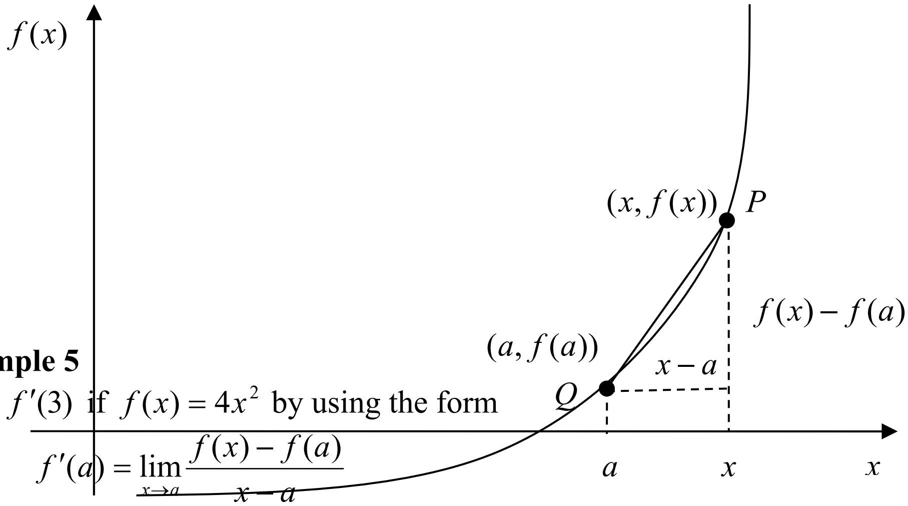


Figure 5 Graph showing the second definition of the derivative.

Example 5

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

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

of the definition of a derivative.

Solution

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

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.

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

$$f'(0.05) = 3(0.05)^2 - 0.165$$

$$= -0.1575$$

$$m = -0.1575$$

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

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

$$= -0.0077257$$

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

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

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

$$c = 0.0001493$$

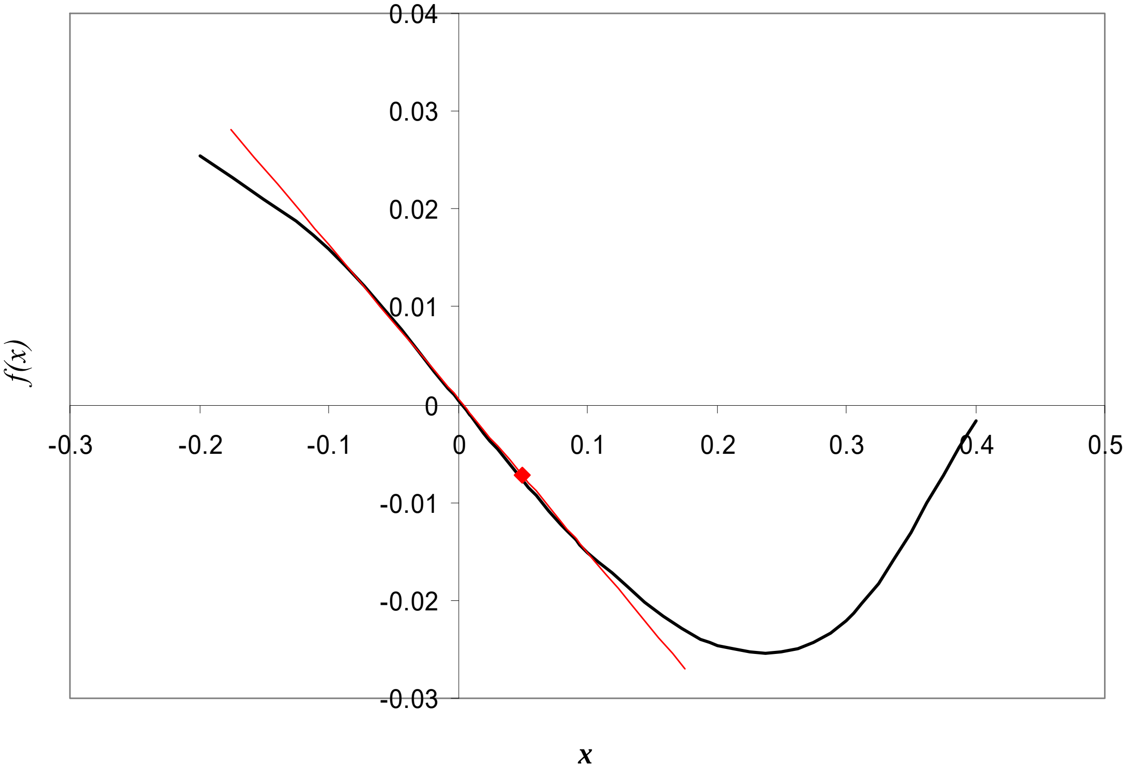


Figure 6 Graph of function $f(x)$ and the tangent line at $x = 0.05$.

Hence,

$$\begin{aligned} y &= mx + c \\ &= -0.1575x + 0.0001493 \end{aligned}$$

is the equation of the tangent line.

Other Notations of Derivatives

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

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

For the second derivative, the notations are

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

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

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

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

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

$$= 9x^2$$

Theorem 5

The derivative of

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

is

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

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 $u(x)$,

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

Taking the derivative of $v(x)$,

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

Using the formula for the product rule

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

$$\begin{aligned} &= (2x^2 - 6)(9x^2) + (3x^3 + 8)(4x) \\ &= 18x^4 - 54x^2 + 12x^4 + 32x \\ &= 30x^4 - 54x^2 + 32x \end{aligned}$$

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

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 $v(x)$,

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

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}
$sin(x)$	$cos(x)$
$cos(x)$	$-sin(x)$
$tan(x)$	$sec^2(x)$
$sinh(x)$	$cosh(x)$
$cosh(x)$	$sinh(x)$
$tanh(x)$	$1-tanh^2(x)$
$sin^{-1}(x)$	$\frac{1}{\sqrt{1-x^2}}$
$cos^{-1}(x)$	$\frac{-1}{\sqrt{1-x^2}}$
$tan^{-1}(x)$	$\frac{1}{1+x^2}$

$csc(x)$	$-csc(x)cot(x)$
$sec(x)$	$sec(x)tan(x)$
$cot(x)$	$-csc^2(x)$
$csch(x)$	$-coth(x)csch(x)$
$sech(x)$	$-tanh(x)sech(x)$
$coth(x)$	$1-coth^2(x)$
$csc^{-1}(x)$	$-\frac{\|x\|}{x^2\sqrt{x^2-1}}$
$sec^{-1}(x)$	$\frac{\|x\|}{x^2\sqrt{x^2-1}}$
$cot^{-1}(x)$	$\frac{-1}{1+x^2}$
a^x	$ln(a)a^x$
$ln(x)$	$\frac{1}{x}$
$log_a(x)$	$\frac{1}{xln(a)}$
e^x	e^x

Chain Rule of Differentiation (diferensiyelin zincir kuralı)

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 (zincir kuralı aşağıdaki gibidir)

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

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 (kapalı diferensiyel alma)

Sometimes, the function to be differentiated is not given explicitly (bağımsız değişkene bağlılık açık değildir) 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

$$x^2 + y^2 = 2xy$$

$$\frac{d}{dx}(x^2 + y^2) = \frac{d}{dx}(2xy)$$

$$\frac{d}{dx}(x^2) + \frac{d}{dx}(y^2) = \frac{d}{dx}(2xy)$$

$$2x + 2y\frac{dy}{dx} = 2x\frac{dy}{dx} + 2y$$

$$2y\frac{dy}{dx} - 2x\frac{dy}{dx} = 2y - 2x$$

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

$$\frac{dy}{dx} = \frac{2y - 2x}{2y - 2x}$$

$$\frac{dy}{dx} = 1$$

Example 15

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

Solution

$$x^2 - xy + y^2 = 5$$

$$\frac{d}{dx}(x^2 - xy + y^2) = \frac{d}{dx}(5)$$

$$\frac{d}{dx}(x^2) - \frac{d}{dx}(xy) + \frac{d}{dx}(y^2) = 0$$

$$2x - x\frac{dy}{dx} - y + 2y\frac{dy}{dx} = 0$$

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

$$\frac{dy}{dx} = \frac{y - 2x}{2y - x}$$

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

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

$$\begin{aligned} f'(x) &= 3(3x^2) - 2 \\ &= 9x^2 - 2 \\ f''(x) &= \frac{d}{dx}(f'(x)) \\ &= \frac{d}{dx}(9x^2 - 2) \\ &= 9(2x) \\ &= 18x \\ f'''(x) &= \frac{d}{dx}(f''(x)) \\ &= \frac{d}{dx}(18x) \\ &= 18 \end{aligned}$$

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',

$$y''=\frac{2\frac{y-2x}{2y-x}-2-2\left(\frac{y-2x}{2y-x}\right)^2}{2y-x}$$

$$=-\frac{6(y^2-xy+x^2)}{(2y-x)^3}$$

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

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

(E) $f(a)$ 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 $[0,5]$.

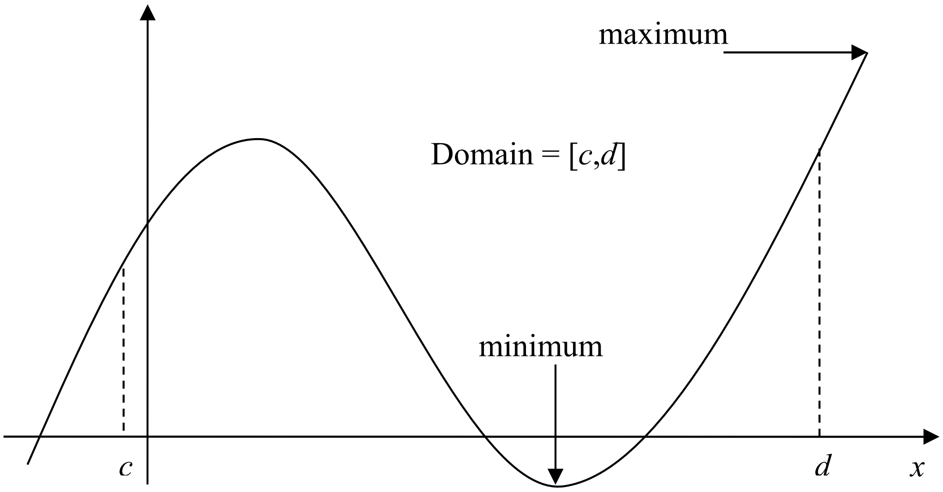


Figure 7 Graph illustrating the concepts of maximum and minimum.

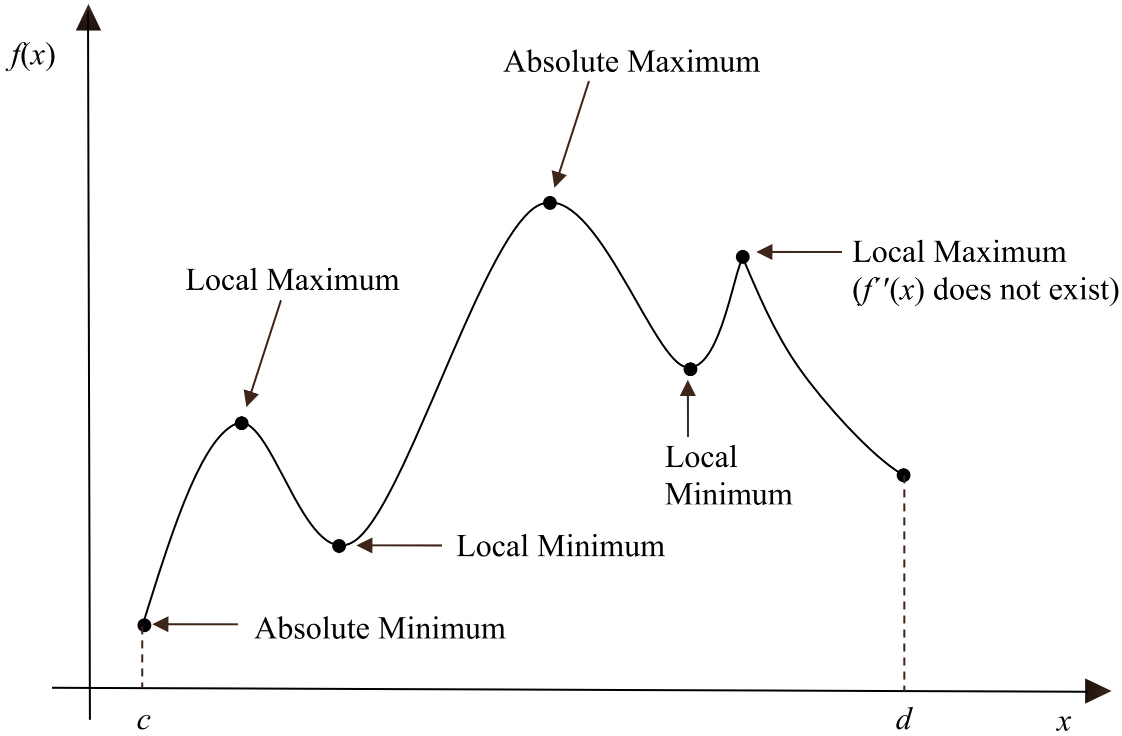


Figure 8 The plot shows critical points of $f(x)$ in $[c,d]$.

Solution

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

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

$f'(x)=0$ 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 $x = 5$.

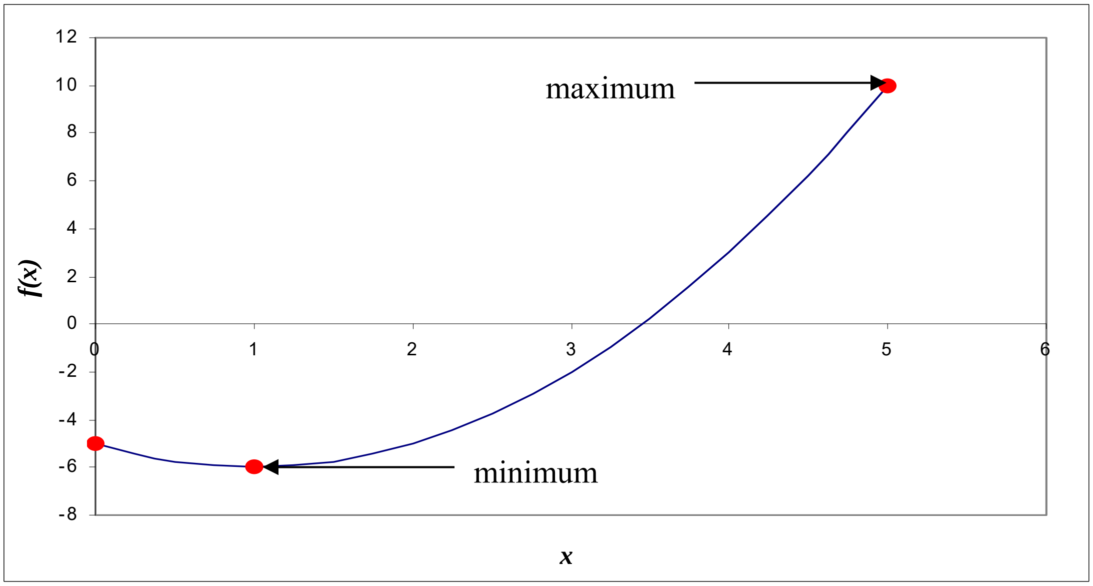


Figure 9 Maximum and minimum values of $f(x) = x^2 - 2x - 5$ over interval [0,5].

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

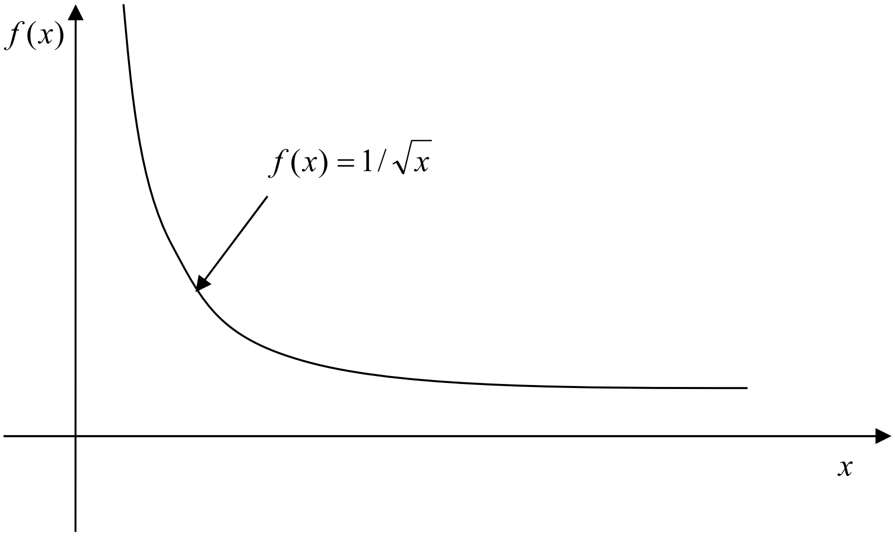


Figure 10 Function that has no maximum or minimum.

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

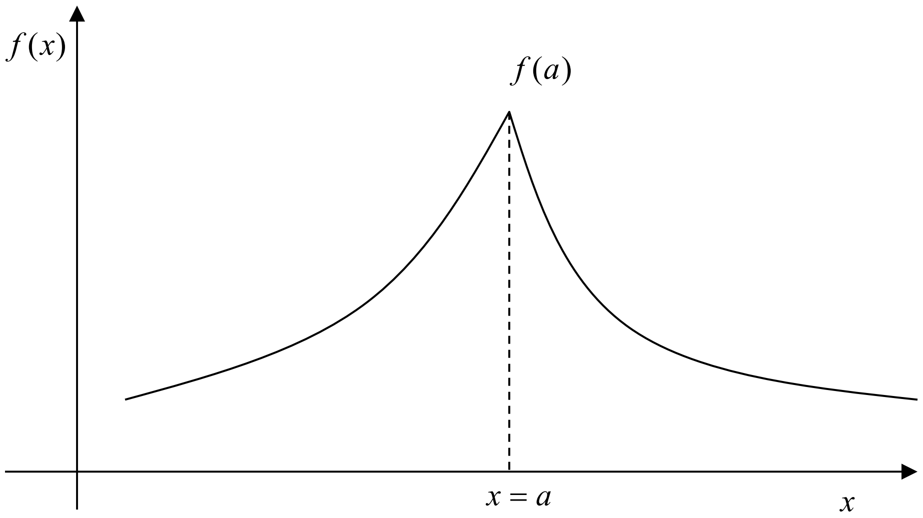


Figure 11 Graph demonstrates the concept of a singular point with discontinuous slope at $x = a$

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.

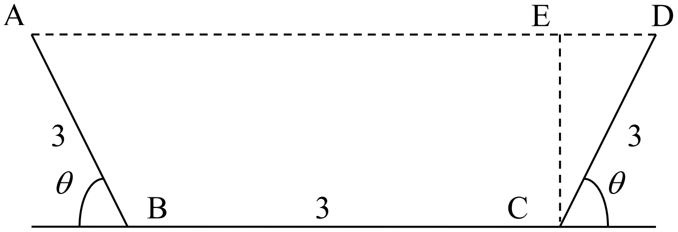


Figure 12 Gutter dimensions for Example 21.

What angle of θ 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

$$Area = \frac{1}{2}(\overline{BC} + \overline{AD}) \times \overline{CE}$$
$$\overline{CE} = \overline{CD} sin(\theta)$$
$$= 3sin(\theta)$$
$$\overline{BC} = 3$$
$$\overline{AD} = \overline{BC} + \overline{CD}\cos(\theta) + \overline{AB}\,cos(\theta)$$
$$\overline{AD} = 3 + 3cos(\theta) + 3cos(\theta)$$
$$\overline{AD} = 3 + 6cos(\theta)$$
$$Area = \frac{1}{2}(3 + 3 + 6cos(\theta))(3sin(\theta))$$
$$= 9sin(\theta) + 9sin(\theta)cos(\theta)$$

$$= 9sin(\theta) + \frac{9}{2}sin(2\theta)$$

$$\frac{dA}{d\theta} = 9cos(\theta) + \frac{9}{2} \times 2cos(2\theta)$$

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

When is

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

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

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

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

$$Area\left(\frac{\pi}{3}\right) = 9sin\left(\frac{\pi}{3}\right) + \frac{9}{2}sin\left(2\left(\frac{\pi}{3}\right)\right)$$

$$= 9\left(\frac{\sqrt{3}}{2}\right) + \frac{9}{2}\left(\frac{\sqrt{3}}{2}\right)$$

$$= \frac{27}{4}\sqrt{3}$$

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

$$Area(0) = 0$$

$$Area(\pi) = 0$$

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

$$A = \text{top surface + side surface + bottom surface}$$

$$= \pi r^2 + 2\pi rh + \pi r^2$$

$$= 2\pi r^2 + 2\pi rh$$

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

$$A = 2\pi r^2 + 2\pi r\left(\frac{9}{\pi r^2}\right)$$

$$= 2\pi r^2 + \frac{18}{r}$$

$$= 2\pi r^2 + 18r^{-1}$$

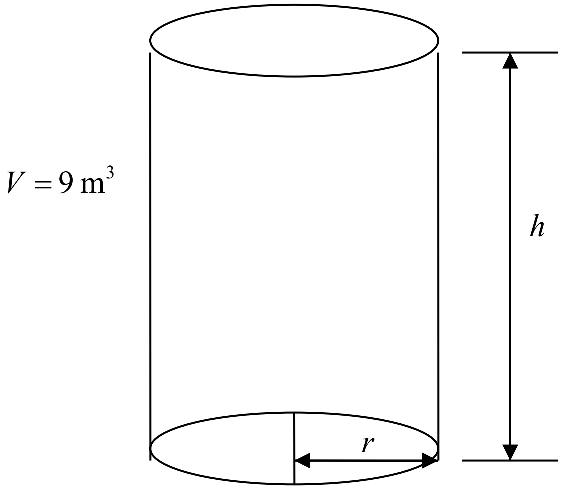


Figure 13 Cylinder drawing for Example 20.

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

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

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

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?

$$\begin{aligned}\frac{d^2A}{dr^2} &= 4\pi - 18(-2)r^{-3} \\ &= 4\pi + \frac{36}{r^3} \\ &= 4\pi + \frac{36}{1.12725} \\ &= 44.5025\end{aligned}$$

This value $\frac{d^2A}{dr^2} > 0$ 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	http://numericalmethods.eng.usf.edu

Multiple-Choice Test Chapter 02.01 A Primer on Differentiation

1. The definition of the first derivative of a function $f(x)$ is

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

(C) $f'(x) = \lim\limits_{\Delta x \to 0} \frac{f(x+\Delta x) + f(x)}{\Delta x}$ (D) $f'(x) = \lim\limits_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x}$

2. Given $y = 5e^{3x} + \sin x$, $\frac{dy}{dx}$ is

(A) $5e^{3x} + \cos x$ (B) $15e^{3x} + \cos x$ (C) $15e^{3x} - \cos x$ (D) $2.666e^{3x} - \cos x$

3. Given $y = \sin 2x$, $\frac{dy}{dx}$ at $x = 3$ is most nearly

(A) 0.9600 (B) 0.9945 (C) 1.920 (D) 1.989

4. Given $y = x^3 \ln x$, $\frac{dy}{dx}$ is

(A) $3x^2 \ln x$ (B) $3x^2 \ln x + x^2$ (C) x^2 (D) $3x$

5. The velocity of a body as a function of time is given as $v(t) = 5e^{-2t} + 4$, where t is in seconds, and v is in m/s. The acceleration in m/s^2 at $t = 0.6$ s is

(A) -3.012 (B) 5.506 (C) 4.147 (D) -10.00

6. If $x^2 + 2xy = y^2$, then $\frac{dy}{dx}$ is

(A) $\frac{x+y}{y-x}$ (B) $2x + 2y$ (C) $\frac{x+1}{y}$ (D) $-x$

For a complete solution, refer to the links at the end of the book.

2.2 Chapter 02.02 Differentiation of Continuous Functions

PRE-REQUISITES

1. Know the definition of a secant, tangent to a function, and derivative of a function (Primer for Differential Calculus).
2. Understand the representation of trigonometric and transcendental functions as a Maclaurin series (Taylor Series Revisited).

OBJECTIVES

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.

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

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

To be able to find a derivative numerically, one could make Δx 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 Δx 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

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

For a finite Δx,

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

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 Δx 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

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

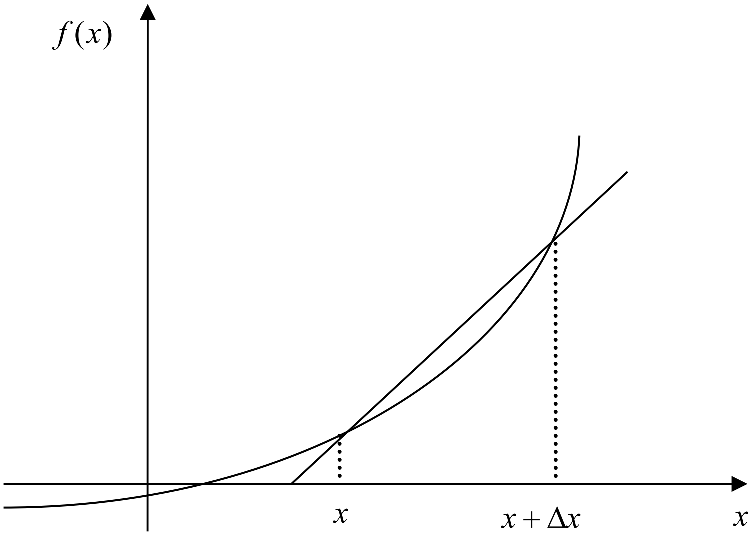


Figure 1 Graphical representation of forward difference approximation of first derivative.

Example 1

The velocity of a rocket is given by

$$v(t) = 2000\ln\left[\frac{14\times10^4}{14\times10^4 - 2100t}\right] - 9.8t,\ 0 \le t \le 30$$

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

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

b) find the exact value of the acceleration of the rocket.

c) calculate the absolute relative true error for part (b).

Solution

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

$$t_i = 16$$

$$\Delta t = 2$$

$$\begin{aligned} t_{i+1} &= t_i + \Delta t \\ &= 16 + 2 \\ &= 18 \end{aligned}$$

$$a(16) \approx \frac{v(18) - v(16)}{2}$$

$$v(18) = 2000\ln\left[\frac{14\times10^4}{14\times10^4 - 2100(18)}\right] - 9.8(18)$$

$$= 453.02 \text{ m/s}$$

$$v(16) = 2000\ln\left[\frac{14\times10^4}{14\times10^4 - 2100(16)}\right] - 9.8(16)$$

$$= 392.07 \text{ m/s}$$

Hence

$$a(16) \approx \frac{v(18) - v(16)}{2}$$

$$= \frac{453.02 - 392.07}{2}$$

$$= 30.474 \text{ m/s}^2$$

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

$$v(t) = 2000\ln\left[\frac{14\times10^4}{14\times10^4 - 2100t}\right] - 9.8t$$

as

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

Knowing that

$$\frac{d}{dt}[\ln(t)] = \frac{1}{t} \text{ and } \frac{d}{dt}\left[\frac{1}{t}\right] = -\frac{1}{t^2}$$

$$a(t) = 2000\left(\frac{14\times10^4 - 2100t}{14\times10^4}\right)\frac{d}{dt}\left(\frac{14\times10^4}{14\times10^4 - 2100t}\right) - 9.8$$

$$= 2000\left(\frac{14\times10^4 - 2100t}{14\times10^4}\right)(-1)\left(\frac{14\times10^4}{(14\times10^4 - 2100t)^2}\right)(-2100) - 9.8$$

$$= \frac{-4040 - 29.4t}{-200 + 3t}$$

$$a(16) = \frac{-4040 - 29.4(16)}{-200 + 3(16)}$$

$$= 29.674 \text{ m/s}^2$$

(c) The absolute relative true error is

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

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

$$= 2.6967\%$$

Backward Difference Approximation of the First Derivative

We know

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

For a finite Δx,

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

If Δx 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 Δx 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

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

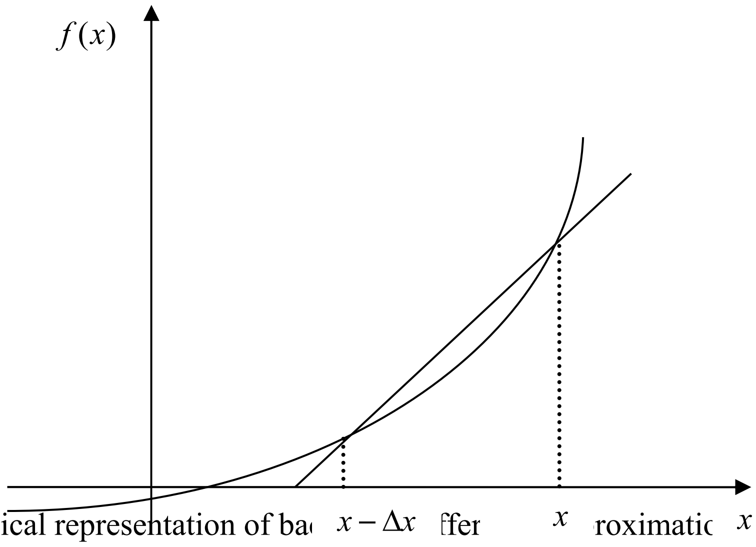


Figure 2 Graphical representation of ba ffer roximatic first derivative.

Example 2

The velocity of a rocket is given by

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

(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

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

$$t_i = 16$$

$$\Delta t = 2$$

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

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

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

$$\begin{aligned} v(14) &= 2000 \ln\left[\frac{14 \times 10^4}{14 \times 10^4 - 2100(14)}\right] - 9.8(14) \\ &= 334.24\,\text{m/s} \end{aligned}$$

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

(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

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

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

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

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

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

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

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

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

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

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

and

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

Subtracting Equation (2) from Equation (1)

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

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

$$= \frac{f(x_{i+1}) - f(x_{i-1})}{2\Delta x} + O(\Delta x)^2$$

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

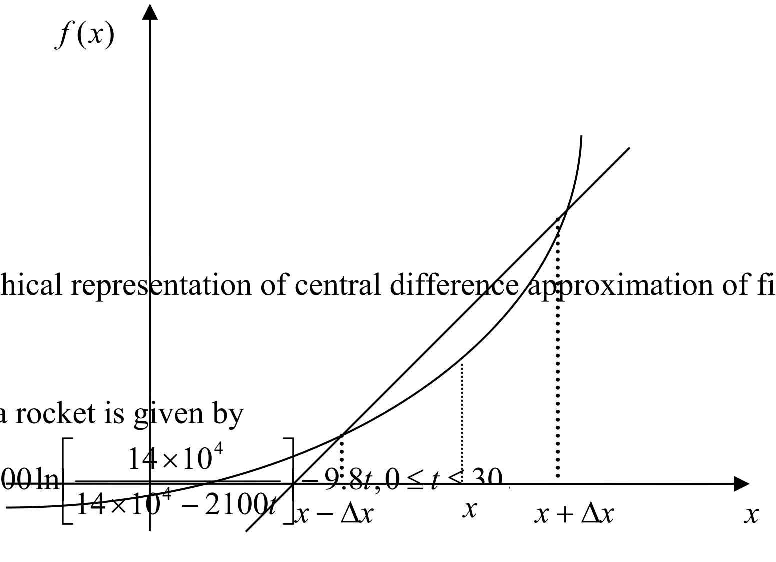


Figure 3 Graphical representation of central difference approximation of first derivative.

Example 3

The velocity of a rocket is given by

$$v(t) = 2000\ln\left[\frac{14\times10^4}{14\times10^4 - 2100t}\right] - 9.8t, 0 \le t \le 30$$

(a) Use the central 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

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

$$t_i = 16$$

$$\Delta t = 2$$

$$\begin{aligned} t_{i+1} &= t_i + \Delta t \\ &= 16 + 2 \\ &= 18 \end{aligned}$$

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

$$\begin{aligned} a(16) &\approx \frac{v(18) - v(14)}{2(2)} \\ &= \frac{v(18) - v(14)}{4} \end{aligned}$$

$$\begin{aligned} v(18) &= 2000\ln\left[\frac{14\times10^4}{14\times10^4 - 2100(18)}\right] - 9.8(18) \\ &= 453.02\,\text{m/s} \end{aligned}$$

$$\begin{aligned} v(14) &= 2000\ln\left[\frac{14\times10^4}{14\times10^4 - 2100(14)}\right] - 9.8(14) \\ &= 334.24\,\text{m/s} \end{aligned}$$

$$\begin{aligned} a(16) &\approx \frac{v(18) - v(14)}{4} \\ &= \frac{453.02 - 334.24}{4} \\ &= 29.694\ \text{m/s}^2 \end{aligned}$$

(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	$a(16)$ $\left(\text{m/s}^2\right)$	$\lvert\in_t\rvert\%$
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 Δt values of backward difference scheme.

Δt	$v'(t)$	$\lvert\in_a\rvert\%$
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

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

where

$$x_{i+2} = x_i + 2\Delta x$$

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

where

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

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

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

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

$$f''(x_i) \approx \frac{f(x_{i+2}) - 2f(x_{i+1}) + f(x_i)}{(\Delta x)^2} + O(\Delta x) \tag{5}$$

Example 4

The velocity of a rocket is given by

$$v(t) = 2000 \ln\left[\frac{14\times10^4}{14\times10^4 - 2100t}\right] - 9.8t, 0 \le t \le 30$$

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

Solution

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

$$t_i = 16$$

$$\Delta t = 2$$

$$\begin{aligned} t_{i+1} &= t_i + \Delta t \\ &= 16 + 2 \\ &= 18 \end{aligned}$$

$$\begin{aligned} t_{i+2} &= t_i + 2(\Delta t) \\ &= 16 + 2(2) \\ &= 20 \end{aligned}$$

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

$$\begin{aligned} v(20) &= 2000 \ln\left[\frac{14\times10^4}{14\times10^4 - 2100(20)}\right] - 9.8(20) \\ &= 517.35\,\text{m/s} \end{aligned}$$

$$\begin{aligned} v(18) &= 2000 \ln\left[\frac{14\times10^4}{14\times10^4 - 2100(18)}\right] - 9.8(18) \\ &= 453.02\,\text{m/s} \end{aligned}$$

$$\begin{aligned} v(16) &= 2000 \ln\left[\frac{14\times10^4}{14\times10^4 - 2100(16)}\right] - 9.8(16) \\ &= 392.07\,\text{m/s} \end{aligned}$$

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

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

$$\frac{d}{dt}[\ln(t)] = \frac{1}{t} \text{ and}$$
$$\frac{d}{dt}\left[\frac{1}{t}\right] = -\frac{1}{t^2}$$

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

Similarly it can be shown that

$$j(t) = \frac{d}{dt}[a(t)]$$
$$= \frac{18000}{(-200 + 3t)^2}$$
$$j(16) = \frac{18000}{[-200 + 3(16)]^2}$$
$$= 0.77909 \text{m/s}^3$$

The absolute relative true error is

$$|\in_t| = \left|\frac{0.77909 - 0.84515}{0.77909}\right| \times 100$$
$$= 8.4797\%$$

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

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

where

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

$$f(x_{i-1}) = f(x_i) - f'(x_i)\Delta x + \frac{f''(x_i)}{2!}(\Delta x)^2 - \frac{f'''(x_i)}{3!}(\Delta x)^3 + \frac{f''''(x_i)}{4!}(\Delta x)^4 - \ldots \qquad (7)$$

where

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

Adding Equations (6) and (7), gives

$$f(x_{i+1}) + f(x_{i-1}) = 2f(x_i) + f''(x_i)(\Delta x)^2 + f'''(x_i)\frac{(\Delta x)^4}{12} + \ldots$$

$$f''(x_i) = \frac{f(x_{i+1}) - 2f(x_i) + f(x_{i-1})}{(\Delta x)^2} - \frac{f''''(x_i)(\Delta x)^2}{12} + \ldots$$

$$= \frac{f(x_{i+1}) - 2f(x_i) + f(x_{i-1})}{(\Delta x)^2} + O(\Delta x)^2$$

Example 5

The velocity of a rocket is given by

$$v(t) = 2000\ln\left[\frac{14\times10^4}{14\times10^4 - 2100t}\right] - 9.8t,\ 0 \le t \le 30,$$

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

Solution

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

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

$$t_i = 16$$

$$\Delta t = 2$$

$$\begin{aligned} t_{i+1} &= t_i + \Delta t \\ &= 16 + 2 \\ &= 18 \end{aligned}$$

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

$$j(16) \approx \frac{v(18) - 2v(16) + v(14)}{(2)^2}$$

$$v(18) = 2000\ln\left[\frac{14\times10^4}{14\times10^4 - 2100(18)}\right] - 9.8(18)$$

$$= 453.02\,\text{m/s}$$

$$v(16) = 2000\ln\left[\frac{14\times10^4}{14\times10^4 - 2100(16)}\right] - 9.8(16)$$

$$= 392.07\,\text{m/s}$$

$$v(14) = 2000\ln\left[\frac{14\times10^4}{14\times10^4 - 2100(14)}\right] - 9.8(14)$$

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

$$j(16) \approx \frac{v(18) - 2v(16) + v(14)}{(2)^2}$$

$$= \frac{453.02 - 2(392.07) + 334.24}{4}$$

$$= 0.77969\,\text{m/s}^3$$

The absolute relative true error is

$$|\in_t| = \left|\frac{0.77908 - 0.77969}{0.77908}\right| \times 100$$

$$= 0.077992\%$$

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	Aralık 8, 2016
Web Site	http://numericalmethods.eng.usf.edu

2.3 Multiple-Choice Test Chapter 02.02 Differentiation of Continuous Functions

1. The definition of the first derivative of a function $f(x)$ is

(A) $f'(x) = \dfrac{f(x+\Delta x) + f(x)}{\Delta x}$ (B) $f'(x) = \dfrac{f(x+\Delta x) - f(x)}{\Delta x}$

(C) $f'(x) = \lim\limits_{\Delta x\to 0}\dfrac{f(x+\Delta x) + f(x)}{\Delta x}$ (D) $f'(x) = \lim\limits_{\Delta x\to 0}\dfrac{f(x+\Delta x) - f(x)}{\Delta x}$

2. The exact derivative of $f(x) = x^3$ at $x = 5$ is most nearly

(A) 25.00 (B) 75.00 (C) 106.25 (D) 125.00

3. Using the forwarded divided difference approximation with a step size of 0.2, the derivative of $f(x) = 5e^{2.3x}$ at $x = 1.25$ is

(A) 163.4 (B) 203.8 (C) 211.1 (D) 258.8

4. A student finds the numerical value of $\frac{d}{dx}(e^x) = 20.220$ at $x = 3$ using a step size of 0.2. Which of the following methods did the student use to conduct the differentiation?

(A) Backward divided difference (B) Calculus, that is, exact
(C) Central divided difference (D) Forward divided difference

5. Using the backward divided difference approximation, $\frac{d}{dx}(e^x) = 4.3715$ at $x = 1.5$ for a step size of 0.05. If you keep halving the step size to find $\frac{d}{dx}(e^x)$ at $x = 1.5$ before two significant digits can be considered to be at least correct in your answer, the step size would be (you cannot use the exact value to determine the answer)

(A) 0.05/2 (B) 0.05/4 (C) 0.05/8 (D) 0.05/16

6. The heat transfer rate q over a surface is given by

$$q = -kA\frac{dT}{dy}$$

where

k = thermal conductivity $\left(\frac{\text{J}}{\text{s}\cdot\text{m}\cdot\text{K}}\right)$
A = surface area $\left(\text{m}^2\right)$
T = temperature (K)
y = distance normal to the surface (m)

Given

$$k = 0.025\frac{\text{J}}{\text{s}\cdot\text{m}\cdot\text{K}}$$

$$A = 3\,\text{m}^2$$

the temperature T over the surface varies as

$$T = -1493y^3 + 2200y^2 - 1076y + 500$$

The heat transfer rate q at the surface most nearly is

(A) -1076 W (B) 37.5 W (C) 80.7 W (D) 500 W

For a complete solution, refer to the links at the end of the book.

2.4 Chapter 02.03 Differentiation of Discrete Functions

PRE-REQUISITES

1. Know the definition of a secant, tangent to a function, and derivative of a function (Primer for Differential Calculus).
2. Understand the representation of trigonometric and transcendental functions as a Maclaurin series (Taylor Series Revisited).

OBJECTIVES

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.

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

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

For a finite Δx,

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

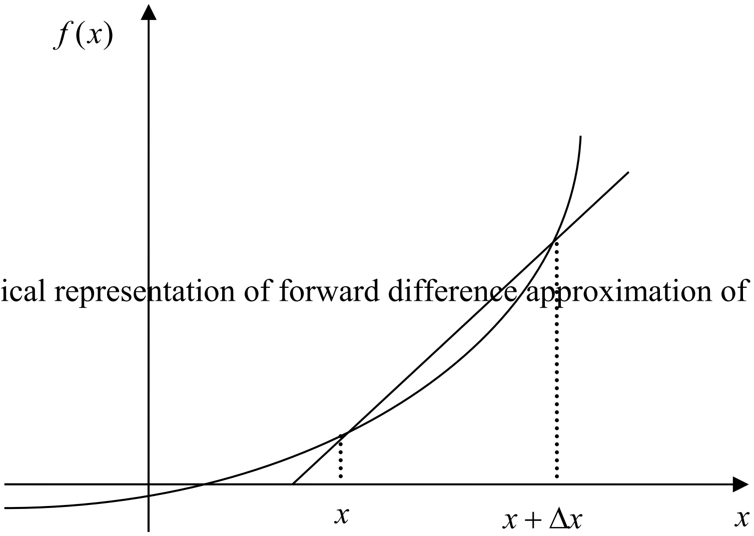


Figure 1 Graphical representation of forward difference approximation of first derivative.

So given $n+1$ data points $(x_0, y_0), (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, the value of $f'(x)$ for $x_i \le x \le x_{i+1}$, $i = 0, \ldots, n-1$, 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 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

Using forward divided difference, find the acceleration of the rocket at $t = 16\ \text{s}$.

Solution

To find the acceleration at $t = 16\ \text{s}$, we need to choose the two values of velocity closest to $t = 16\ \text{s}$, that also bracket $t = 16\ \text{s}$ to evaluate it. The two points are $t = 15\ \text{s}$ and $t = 20\ \text{s}$

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

$$t_i = 15$$

$$t_{i+1} = 20$$

$$\begin{aligned}\Delta t &= t_{i+1} - t_i \\ &= 20 - 15 \\ &= 5\end{aligned}$$

$$\begin{aligned}a(16) &\approx \frac{v(20) - v(15)}{5} \\ &= \frac{517.35 - 362.78}{5} \\ &= 30.914\ \text{m/s}^2\end{aligned}$$

Direct Fit Polynomials

In this method, given $n+1$ data points $(x_0, y_0), (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, one can fit a n^{th} order polynomial given by

$$P_n(x) = a_0 + a_1 x + \ldots\ldots + a_{n-1}x^{n-1} + a_n x^n$$

To find the first derivative,

$$P_n'(x) = \frac{dP_n(x)}{dx} = a_1 + 2a_2 x + \ldots\ldots + (n-1)a_{n-1}x^{n-2} + na_n x^{n-1}$$

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 4 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 $t_3 = 22.5$.

$$t_0 = 10,\ v(t_0) = 227.04$$
$$t_1 = 15,\ v(t_1) = 362.78$$
$$t_2 = 20,\ v(t_2) = 517.35$$
$$t_3 = 22.5,\ v(t_3) = 602.97$$

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

$$\begin{bmatrix} 1 & 10 & 100 & 1000 \\ 1 & 15 & 225 & 3375 \\ 1 & 20 & 400 & 8000 \\ 1 & 22.5 & 506.25 & 11391 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 227.04 \\ 362.78 \\ 517.35 \\ 602.97 \end{bmatrix}$$

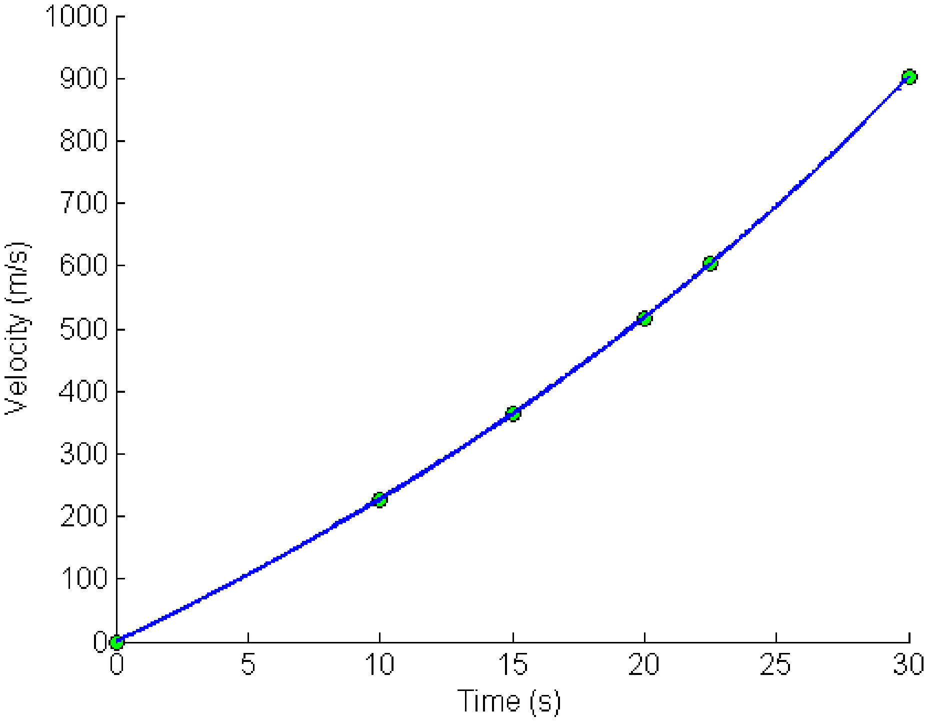


Figure 2 Graph of upward velocity of the rocket vs. time.

Solving the above four equations gives

$$a_0 = -4.3810$$
$$a_1 = 21.289$$
$$a_2 = 0.13065$$
$$a_3 = 0.0054606$$

Hence

$$\begin{aligned} v(t) &= a_0 + a_1 t + a_2 t^2 + a_3 t^3 \\ &= -4.3810 + 21.289t + 0.13065t^2 + 0.0054606t^3, \quad 10 \le t \le 22.5 \end{aligned}$$

The acceleration at $t = 16$ is given by

$$a(16) = \frac{d}{dt} v(t)\Big|_{t=16}$$

Given that $v(t) = -4.3810 + 21.289t + 0.13065t^2 + 0.0054606t^3, \quad 10 \le t \le 22.5$,

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

$$
\begin{aligned}
&= \frac{d}{dt}\left(-4.3810 + 21.289t + 0.13065t^2 + 0.0054606t^3\right) \\
&= 21.289 + 0.26130t + 0.016382t^2, \quad 10 \le t \le 22.5 \\
a(16) &= 21.289 + 0.26130(16) + 0.016382(16)^2 \\
&= 29.664 \text{ m/s}^2
\end{aligned}
$$

Lagrange Polynomial

In this method, given $(x_0, y_0), \ldots, (x_n, y_n)$, one can fit a n^{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^{th} order polynomial that approximates the function $y = f(x)$ and

$$L_i(x) = \prod_{\substack{j=0 \\ j \ne i}}^{n} \frac{x - x_j}{x_i - x_j}$$

$L_i(x)$ 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

$$v(t) = \left(\frac{t-t_1}{t_0-t_1}\right)\left(\frac{t-t_2}{t_0-t_2}\right)v(t_0) + \left(\frac{t-t_0}{t_1-t_0}\right)\left(\frac{t-t_2}{t_1-t_2}\right)v(t_1) + \left(\frac{t-t_0}{t_2-t_0}\right)\left(\frac{t-t_1}{t_2-t_1}\right)v(t_2)$$

$$a(t) = \frac{2t-(t_1+t_2)}{(t_0-t_1)(t_0-t_2)}v(t_0) + \frac{2t-(t_0+t_2)}{(t_1-t_0)(t_1-t_2)}v(t_1) + \frac{2t-(t_0+t_1)}{(t_2-t_0)(t_2-t_1)}v(t_2)$$

$$\begin{aligned}a(16) &= \frac{2(16)-(15+20)}{(10-15)(10-20)}(227.04) + \frac{2(16)-(10+20)}{(15-10)(15-20)}(362.78)\\ &\quad + \frac{2(16)-(10+15)}{(20-10)(20-15)}(517.35)\\ &= -0.06(227.04) - 0.08(362.78) + 0.14(517.35)\\ &= 29.784\,\text{m/s}^2\end{aligned}$$

DIFFERENTIATION

Topic	Differentiation of Discrete Functions
Summary	These are textbook notes differentiation of discrete functions
Major	General Engineering
Authors	Autar Kaw, Luke Snyder
Date	Aralık 8, 2016
Web Site	http://numericalmethods.eng.usf.edu

2.5 Multiple-Choice Test Chapter 02.03 Differentiation of Discrete Functions

1. The definition of the first derivative of a function $f(x)$ is

(A) $f'(x) = \dfrac{f(x+\Delta x)+f(x)}{\Delta x}$ (B) $f'(x) = \dfrac{f(x+\Delta x)-f(x)}{\Delta x}$

(C) $f'(x) = \lim\limits_{\Delta x\to 0}\dfrac{f(x+\Delta x)+f(x)}{\Delta x}$ (D) $f'(x) = \lim\limits_{\Delta x\to 0}\dfrac{f(x+\Delta x)-f(x)}{\Delta x}$

2. Using the forward divided difference approximation with a step size of 0.2, the derivative of the function at $x = 2$ is given as

x	1.8	2.0	2.2	2.4	2.6
$f(x)$	6.0496	7.3890	9.0250	11.023	13.464

(A) 6.697 (B) 7.389 (C) 7.438 (D) 8.180

3. A student finds the numerical value of $f'(x) = 20.220$ at $x = 3$ using a step size of 0.2. Which of the following methods did the student use to conduct the differentiation if $f(x)$ is given in the table below?

x	2.6	2.8	3.0	3.2	3.4	3.6
$f(x)$	$e^{2.6}$	$e^{2.8}$	e^{3}	$e^{3.2}$	$e^{3.4}$	$e^{3.6}$

(A) Backward divided difference (B) Calculus, that is, exact
(C) Central divided difference (D) Forward divided difference

4. The upward velocity of a body is given as a function of time as

t, s	10	15	20	22
v, m/s	22	36	57	10

To find the acceleration at $t = 17$ s, a scientist finds a second order polynomial approximation for the velocity, and then differentiates it to find the acceleration. The estimate of the acceleration in m/s^2 at $t = 17$ s is most nearly

(A) 4.060 (B) 4.200 (C) 8.157 (D) 8.498

5. The velocity of a rocket is given as a function of time as

t, s	0	0.5	1.2	1.5	1.8
v, m/s	0	213	223	275	300

Allowed to use the forward divided difference, backward divided difference or central divided difference approximation of the first derivative, your best estimate for the acceleration $\left(a = \dfrac{dv}{dt}\right)$ of the rocket in m/s^2 at $t = 1.5$ seconds is

(A) 83.33 (B) 128.33 (C) 173.33 (D) 183.33

6. In a circuit with an inductor of inductance L, a resistor with resistance R, and a variable voltage source $E(t)$,

$$E(t) = L\frac{di}{dt} + Ri$$

The current, i, is measured at several values of time as

Time, t (secs)	1.00	1.01	1.03	1.1
Current, i (amperes)	3.10	3.12	3.18	3.24

If $L = 0.98$ henries and $R = 0.142$ ohms, the most accurate expression for $E(1.00)$ is

(A) $0.98\left(\dfrac{3.24 - 3.10}{0.1}\right) + (0.142)(3.10)$

(B) 0.142×3.10

(C) $0.98\left(\dfrac{3.12 - 3.10}{0.01}\right) + (0.142)(3.10)$

(D) $0.98\left(\dfrac{3.12 - 3.10}{0.01}\right)$

For a complete solution, refer to the links at the end of the book.

3 Chapter 03.01 Cubic Equations - Solution of Quadratic Equations

3.1 Solution of Quadratic Equations

PRE-REQUISITES (ön koşullar)

1. High School Algebra – Equations
2. Complex Algebra – Concept of complex numbers, Euler's formula

OBJECTIVES (hedefler)

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.

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

$ax^2 + bx + c = 0$, 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

$$\left(x + \frac{b}{2a}\right)^2 - \frac{b^2}{4a^2} + \frac{c}{a} = 0$$

$$\left(x + \frac{b}{2a}\right)^2 = \frac{b^2}{4a^2} - \frac{c}{a}$$

$$= \frac{b^2 - 4ac}{4a^2}$$

$$x + \frac{b}{2a} = \pm\sqrt{\frac{b^2 - 4ac}{4a^2}}$$

$$= \pm\frac{\sqrt{b^2 - 4ac}}{2a}$$

$$x = -\frac{b}{2a} \pm \frac{\sqrt{b^2 - 4ac}}{2a}$$

$$= \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

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^2)

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 $3.315\ \text{s}$.

NONLINEAR EQUATIONS	
Topic	Solution of quadratic equations
Summary	Textbook notes on solving quadratic equations
Major	General Engineering
Authors	Autar Kaw
Date	Aralık 8, 2016
Web Site	http://numericalmethods.eng.usf.edu

3.1.1 Multiple-Choice Test Chapter 03.01 Background Nonlinear Equations

1. The value of x that satisfies $f(x) = 0$ is called the
(A) root of an equation $f(x) = 0$ (B) root of a function $f(x)$
(C) zero of an equation $f(x) = 0$ (D) none of the above

2. A quadratic equation has ________ root(s).
(A) one (B) two (C) three (D) four

3. For a certain cubic equation, at least one of the roots is known to be a complex root. How many total complex roots does the cubic equation have?
(A) one (B) two (C) three (D) cannot be determined

4. An equation such as $\tan x = x$ has _________ root(s).
(A) zero (B) one (C) two (D) infinite

5. A polynomial of order n has __________ zeros.
(A) $n-1$ (B) n (C) $n+1$ (D) $n+2$

6. The velocity of a body is given by $v(t) = 5e^{-t} + 4$, where t is in seconds and v is in m/s. The velocity of the body is $6\ \text{m/s}$ at $t =$ _____ seconds.
(A) 0.1823 (B) 0.3979 (C) 0.9163 (D) 1.609

For a complete solution, refer to the links at the end of the book.

3.2 Chapter 03.02 Solution of Cubic Equations

PRE-REQUISITES

1. Know how to manipulate equations.
2. Know how to find solution of quadratic equations (Solution of Quadratic Equations)
3. Know basics of complex numbers (Complex numbers – Trusted External link)
4. Know Euler's formula (Euler's formula – Trusted External link)

OBJECTIVES

1. find the exact solution of a general cubic equation.

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

$$ax^3+bx^2+cx+d=0 \tag{1}$$

To find the roots of Equation (1), we first get rid of the quadratic term $\left(x^2\right)$ by making the substitution

$$x=y-\frac{b}{3a} \tag{2}$$

to obtain

$$a\left(y-\frac{b}{3a}\right)^3+b\left(y-\frac{b}{3a}\right)^2+c\left(y-\frac{b}{3a}\right)+d=0 \tag{3}$$

Expanding Equation (3) and simplifying, we obtain the following equation

$$ay^3+\left(c-\frac{b^2}{3a}\right)y+\left(d+\frac{2b^3}{27a^2}-\frac{bc}{3a}\right)=0 \tag{4}$$

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

$$y^3+\frac{1}{a}\left(c-\frac{b^2}{3a}\right)y+\frac{1}{a}\left(d+\frac{2b^3}{27a^2}-\frac{bc}{3a}\right)=0$$

$$y^3+ey+f=0 \tag{5}$$

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

$$y = z + \frac{s}{z} \tag{6}$$

For the time being, the constant s is undefined. Substituting into the depressed cubic Equation (5), we get

$$\left(z + \frac{s}{z}\right)^3 + e\left(z + \frac{s}{z}\right) + f = 0 \tag{7}$$

Expanding out and multiplying both sides by z^3, we get

$$z^6 + (3s + e)z^4 + fz^3 + s(3s + e)z^2 + s^3 = 0 \tag{8}$$

Now, let $s = -\frac{e}{3}$ (s is no longer undefined) to simplify the equation into a tri-quadratic equation.

$$z^6 + fz^3 - \frac{e^3}{27} = 0 \tag{9}$$

By making one more substitution, $w = z^3$, we now have a general quadratic equation which can be solved using the quadratic formula.

$$w^2 + fw - \frac{e^3}{27} = 0 \tag{10}$$

Once you obtain the solution to this quadratic equation, back substitute using the previous substitutions to obtain the roots to the general cubic equation.

$$w \to z \to y \to x$$

where we assumed

$$w = z^3 \tag{11}$$

$$y = z + \frac{s}{z}$$

$$s = -\frac{e}{3} \tag{12}$$

$$x = y - \frac{b}{3a}$$

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

$$x^3 - 9x^2 + 36x - 80 = 0 \qquad \text{(E1-1)}$$

Equation (E1-1) is reduced to

$$y^3 + ey + f = 0$$

where

$$e = \frac{1}{a}\left(c - \frac{b^2}{3a}\right)$$

$$= \frac{1}{1}\left(36 - \frac{(-9)^2}{3(1)}\right)$$

$$= 9$$

and

$$f = \frac{1}{a}\left(d + \frac{2b^3}{27a^2} - \frac{bc}{3a}\right)$$

$$= \frac{1}{1}\left(-80 + \frac{2(-9)^3}{27(1)^2} - \frac{(-9)(36)}{3(1)}\right)$$

$$= -26$$

giving

$$y^3 + 9y - 26 = 0 \qquad \text{(E1-2)}$$

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)

$$s = -\frac{e}{3}$$

$$= -\frac{9}{3}$$

$$= -3$$

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

$$y = z + \frac{s}{z}$$

$$= z - \frac{3}{z}$$

$$w = \frac{-(-26) \pm \sqrt{(-26)^2 - 4(1)(-27)}}{2(1)}$$

$$= 27, -1$$

The solution is

$$w_1 = 27$$

$$w_2 = -1$$

Since

$$w = z^3$$

$$z^3 = w$$

For $w = w_1$

$$z^3 = w_1$$

$$= 27$$

$$= 27e^{i0}$$

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 2π ,

$$3\alpha = \theta + 2\pi k$$

$$\alpha = \frac{\theta + 2\pi k}{3}$$

k will take the value of 0, 1 and 2 before repeating the same values of α .

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

$$y = z - \frac{3}{z}$$
$$y_1 = z_1 - \frac{3}{z_1}$$
$$= 3 - \frac{3}{3}$$

$$= 2$$

$$y_2 = z_2 - \frac{3}{z_2}$$

$$= \left(-\frac{3}{2} + i\frac{3\sqrt{3}}{2}\right) - \frac{3}{\left(-\frac{3}{2} + i\frac{3\sqrt{3}}{2}\right)}$$

$$= -\frac{5 + 3i\sqrt{3}}{-1 + i\sqrt{3}}$$

$$= -\frac{5 + 3i\sqrt{3}}{-1 + i\sqrt{3}} \times \frac{-1 - i\sqrt{3}}{-1 - i\sqrt{3}}$$

$$= -1 + i2\sqrt{3}$$

$$y_3 = z_3 - \frac{3}{z_3}$$

$$= \left(-\frac{3}{2} - i\frac{3\sqrt{3}}{2}\right) - \frac{3}{\left(-\frac{3}{2} - i\frac{3\sqrt{3}}{2}\right)}$$

$$= \frac{5 - i3\sqrt{3}}{1 + i\sqrt{3}}$$

$$= \frac{5 - i3\sqrt{3}}{1 + i\sqrt{3}} \times \frac{1 - i\sqrt{3}}{1 - i\sqrt{3}}$$

$$= -1 - i2\sqrt{3}$$

Since

$$x = y + 3$$

$$x_1 = y_1 + 3$$
$$= 2 + 3$$
$$= 5$$

$$x_2 = y_2 + 3$$
$$= \left(-1 + i2\sqrt{3}\right) + 3$$
$$= 2 + i2\sqrt{3}$$

$$x_3 = y_3 + 3$$
$$= \left(-1 - i2\sqrt{3}\right) + 3$$
$$= 2 - i2\sqrt{3}$$

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

$$(x-5)\left(x-\left(2+i2\sqrt{3}\right)\right)\left(x-\left(2-i2\sqrt{3}\right)\right)=0$$

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

$$x^3-0.03x^2+2.4\times10^{-6}=0$$

Solution

For the general form

$$ax^3+bx^2+cx+d=0$$

$$a=1, b=-0.03, c=0, d=2.4\times10^{-6}$$

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

$$y^3-\left(3\times10^{-4}\right)y+\left(4\times10^{-7}\right)=0$$

That gives $e=-3\times10^{-4}$ and $f=4\times10^{-7}$ 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

$$z^6+\left(3s-3\times10^{-4}\right)z^4+\left(4\times10^{-7}\right)z^3+s\left(3s-3\times10^{-4}\right)z^2+s^3=0$$

Letting

$$s=-\frac{e}{3}=-\frac{-3\times10^{-4}}{3}=10^{-4}$$

we get the following tri-quadratic equation

$$z^6+\left(4\times10^{-7}\right)z^3+1\times10^{-12}=0$$

Using the following conversion, $w=z^3$, we get a general quadratic equation

$$w^2+\left(4\times10^{-7}\right)w+\left(1\times10^{-12}\right)=0$$

Using the quadratic equation, the solutions for w are

$$w=\frac{-4\times10^{-7}\pm\sqrt{\left(4\times10^{-7}\right)^2-4(1)\left(1\times10^{-12}\right)}}{2(1)}$$

giving

$$w_1 = -2\times10^{-7} + i\left(9.79795897113\times10^{-7}\right)$$
$$w_2 = -2\times10^{-7} - i\left(9.79795897113\times10^{-7}\right)$$

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

$$w = z^3$$
$$re^{i\theta} = \left(ue^{i\alpha}\right)^3 = u^3e^{3i\alpha}$$
$$r(\cos\theta + i\sin\theta) = u^3(\cos 3\alpha + i\sin 3\alpha)$$

resulting in

$$r = u^3$$
$$\cos\theta = \cos 3\alpha$$
$$\sin\theta = \sin 3\alpha$$

Since $\sin\theta$ and $\cos\theta$ are periodic of 2π,

$$3\alpha = \theta + 2\pi k$$
$$\alpha = \frac{\theta + 2\pi k}{3}$$

k will take the value of 0, 1 and 2 before repeating the same values of α.

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

$$w_1 = -2\times 10^{-7} + i\left(9.79795897113\times 10^{-7}\right)$$

$$r = \sqrt{\left(-2\times 10^{-7}\right)^2 + \left(9.79795897113\times 10^{-7}\right)^2}$$

$$= 1\times 10^{-6}$$

$$\theta = \tan^{-1}\frac{9.79795897113\times 10^{-7}}{-2\times 10^{-7}}$$

$= 1.772154248$ (2nd quadrant because y (the numerator) is positive and x (the denominator) is negative)

$$z_1 = \left(1\times 10^{-6}\right)^{\frac{1}{3}}\left(\cos\frac{1.772154248}{3} + i\sin\frac{1.772154248}{3}\right)$$

$$= 0.008305409517 + i0.005569575635$$

$$z_2 = \left(1\times 10^{-6}\right)^{\frac{1}{3}}\left(\cos\frac{1.772154248 + 2\pi}{3} + i\sin\frac{1.772154248 + 2\pi}{3}\right)$$

$$= -0.008976098746 + i0.004407907815$$

$$z_3 = \left(1\times 10^{-6}\right)^{\frac{1}{3}}\left(\cos\frac{1.772154248 + 4\pi}{3} + i\sin\frac{1.772154248 + 4\pi}{3}\right)$$

$$= 0.0006706892313 - i0.009977483448$$

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)),

$$y = z + \frac{s}{z}$$

$$y = z + \frac{\left(1\times 10^{-4}\right)}{z}$$

we back substitute to find three values for y.

For example, choosing

$$z_1 = 0.008305409518 + i0.005569575634$$

gives

$$y_1 = 0.008305409518 + i0.005569575634 + \frac{1\times 10^{-4}}{0.008305409518 + i0.005569575634}$$

$$= 0.008305409518 + i0.005569575634$$

$$+ \frac{1\times10^{-4}}{0.008305409078 + i0.00556957634} \times \frac{0.008305409518 - i0.00556957634}{0.008305409518 - i0.00556957634}$$

$$= 0.008305409518 + i0.005569575634$$

$$+ \frac{1\times10^{-4}}{1\times10^{-4}}(0.008305409518 - i0.00556957634)$$

$$= 0.016610819036$$

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

$$x_1 = 0.016610819036 + 0.01$$

$$= 0.026610819036$$

$$x_2 = -0.01795219749 + 0.01$$

$$= -0.00795219749$$

$$x_3 = 0.001341378457 + 0.01$$

$$= 0.011341378457$$

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	Aralık 8, 2016
Web Site	http://numericalmethods.eng.usf.edu

3.3 Chapter 03.03 Bisection (ikiye bölme) Method of Solving a Nonlinear Equation

PRE-REQUISITES

1. Know what a function of one variable is.
2. High School Algebra.

OBJECTIVES

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.

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? (İkiye bölme yöntemi nedir ve neye dayanır?)

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.

İkiye bölme yöntemi ilk sayısal yöntemlerden birisi olarak f(x)=0 şeklindeki çizgisel olmayan eşitliklerin köklerini bulmak için geliştirilmiştir (aynı zamanda ikili-arama yöntemi olarakta isimlendirilir). Yöntem aşağıdaki teoreme dayandırılır.

Theorem

An equation $f(x)=0$, where $f(x)$ is a real continuous function, has at least one root between x_ℓ 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_ℓ 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_ℓ and x_u (Figure 4). So the theorem only guarantees one root between x_ℓ 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 (aralık) methods.

Since the root is bracketed between two points, x_ℓ and x_u, one can find the mid-point, x_m between x_ℓ and x_u. This gives us two new intervals

1. x_ℓ and x_m, and

2. x_m and x_u.

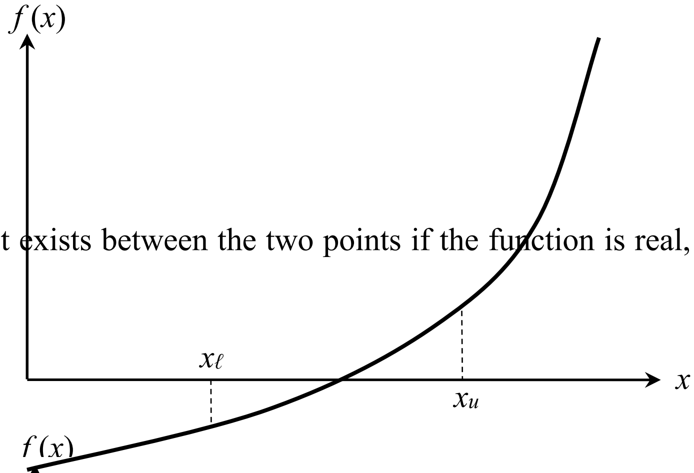


Figure 1 At least one root exists between the two points if the function is real, continuous, and changes sign.

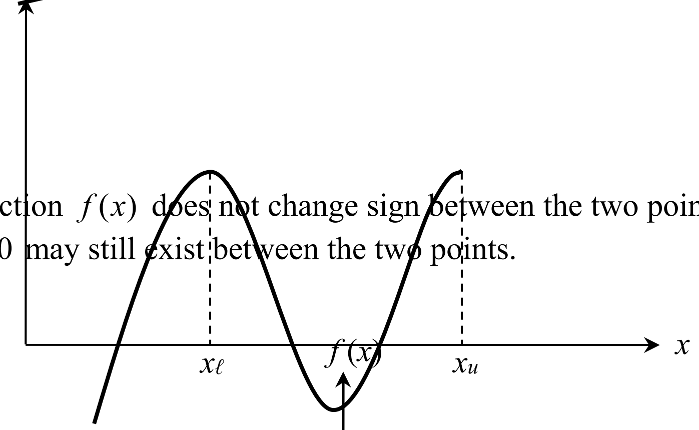


Figure 2 If the function $f(x)$ does not change sign between the two points, roots of the equation $f(x)=0$ may still exist between the two points.

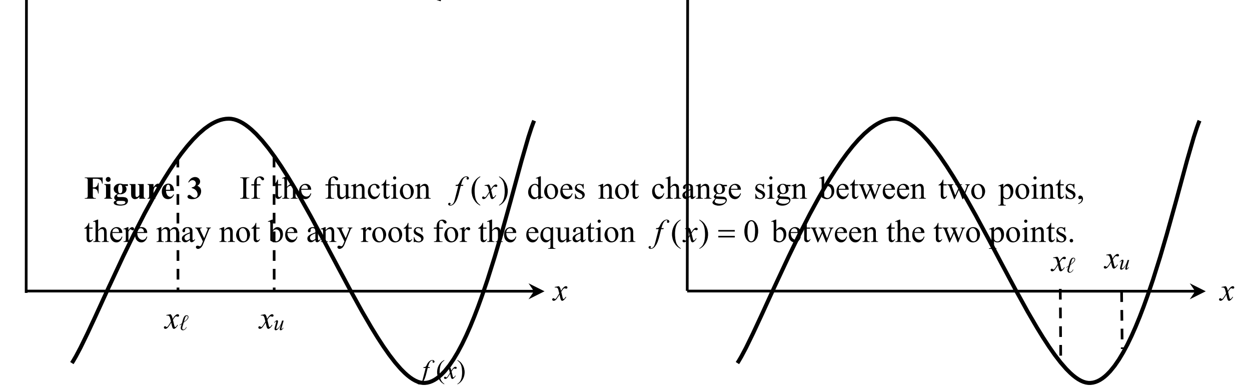


Figure 3 If the function $f(x)$ does not change sign between two points, there may not be any roots for the equation $f(x)=0$ between the two points.

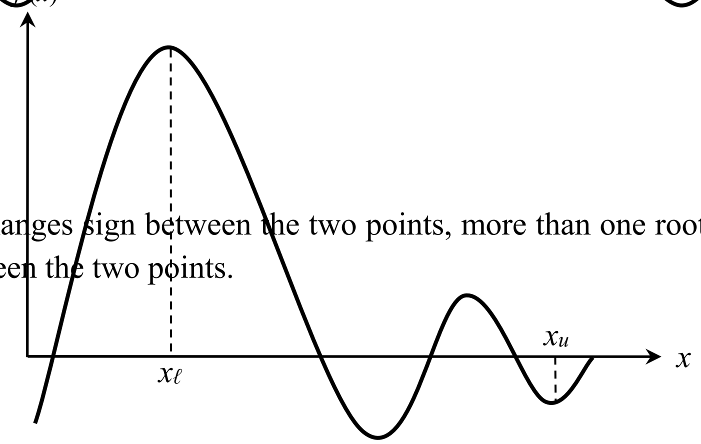


Figure 4 If the function $f(x)$ changes sign between the two points, more than one root for the equation $f(x)=0$ may exist between the two points.

Is the root now between x_ℓ and x_m or between x_m and x_u? Well, one can find the sign of $f(x_\ell)f(x_m)$, and if $f(x_\ell)f(x_m) < 0$ 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 $[x_\ell, x_u]$ 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_ℓ 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_ℓ and x_u.
2. Estimate the root, x_m, of the equation $f(x) = 0$ as the mid-point between x_ℓ 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_ℓ and x_m; then $x_\ell = x_\ell$ 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 $x_\ell = x_m$ and $x_u = x_u$.
   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
$$\left|\in_a\right| = \left|\frac{x_m^{\text{new}} - x_m^{\text{old}}}{x_m^{\text{new}}}\right| \times 100$$
where
x_m^{new} = estimated root (tahmini kök) from present iteration (yinelemek)
x_m^{old} = estimated root from previous iteration
5. Compare the absolute relative approximate error $\left|\in_a\right|$ with the pre-specified relative error tolerance $\in_s$. If $\left|\in_a\right| > \in_s$, 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 (klozet). 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

$$x^3 - 0.165x^2 + 3.993\times10^{-4} = 0$$

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

$$R = \text{radius of the ball,}$$

that is

$$0 \le x \le 2R$$
$$0 \le x \le 2(0.055)$$
$$0 \le x \le 0.11$$

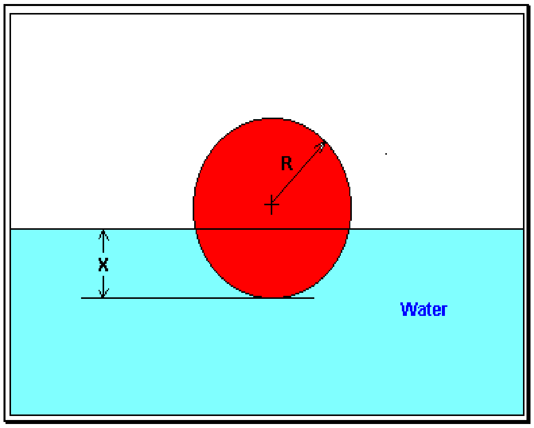


Figure 5 Floating ball problem.

Lets us assume

$$x_\ell = 0, x_u = 0.11$$

Check if the function changes sign between x_ℓ and x_u.

$$f(x_\ell) = f(0) = (0)^3 - 0.165(0)^2 + 3.993\times10^{-4} = 3.993\times10^{-4}$$
$$f(x_u) = f(0.11) = (0.11)^3 - 0.165(0.11)^2 + 3.993\times10^{-4} = -2.662\times10^{-4}$$

Hence

$$f(x_\ell)f(x_u) = f(0)f(0.11) = (3.993\times10^{-4})(-2.662\times10^{-4}) < 0$$

So 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

$$x_m = \frac{x_\ell + x_u}{2}$$
$$= \frac{0 + 0.11}{2}$$

$$= 0.055$$

$$f(x_m) = f(0.055) = (0.055)^3 - 0.165(0.055)^2 + 3.993 \times 10^{-4} = 6.655 \times 10^{-5}$$

$$f(x_\ell)f(x_m) = f(0)f(0.055) = (3.993 \times 10^{-4})(6.655 \times 10^{-4}) > 0$$

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 $|\in_a|$ cannot be calculated as we do not have a previous approximation.

Iteration 2

The estimate of the root is

$$x_m = \frac{x_\ell + x_u}{2}$$

$$= \frac{0.055 + 0.11}{2}$$

$$= 0.0825$$

$$f(x_m) = f(0.0825) = (0.0825)^3 - 0.165(0.0825)^2 + 3.993 \times 10^{-4} = -1.622 \times 10^{-4}$$

$$f(x_\ell)f(x_m) = f(0.055)f(0.0825) = (6.655 \times 10^{-5}) \times (-1.622 \times 10^{-4}) < 0$$

Hence, the root is bracketed between x_ℓ 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 $|\in_a|$ at the end of Iteration 2 is

$$|\in_a| = \left| \frac{x_m^{\text{new}} - x_m^{\text{old}}}{x_m^{\text{new}}} \right| \times 100$$

$$= \left| \frac{0.0825 - 0.055}{0.0825} \right| \times 100$$

$$= 33.33\%$$

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

$$x_m = \frac{x_\ell + x_u}{2}$$

$$= \frac{0.055 + 0.0825}{2}$$

$$= 0.06875$$

$$f(x_m) = f(0.06875) = (0.06875)^3 - 0.165(0.06875)^2 + 3.993 \times 10^{-4} = -5.563 \times 10^{-5}$$

$$f(x_\ell)f(x_m) = f(0.055)f(0.06875) = (6.655 \times 10^{5}) \times (-5.563 \times 10^{-5}) < 0$$

Hence, the root is bracketed between x_ℓ 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 $|\in_a|$ at the ends of Iteration 3 is

$$|\in_a| = \left|\frac{x_m^{\text{new}} - x_m^{\text{old}}}{x_m^{\text{new}}}\right| \times 100$$

$$= \left|\frac{0.06875 - 0.0825}{0.06875}\right| \times 100$$

$$= 20\%$$

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	$\|\in_a\|\%$	$f(x_m)$
1	0.00000	0.11	0.055	----------	6.655×10^{-5}
2	0.055	0.11	0.0825	33.33	-1.622×10^{-4}
3	0.055	0.0825	0.06875	20.00	-5.563×10^{-5}
4	0.055	0.06875	0.06188	11.11	4.484×10^{-6}
5	0.06188	0.06875	0.06531	5.263	-2.593×10^{-5}
6	0.06188	0.06531	0.06359	2.702	-1.0804×10^{-5}
7	0.06188	0.06359	0.06273	1.370	-3.176×10^{-6}
8	0.06188	0.06273	0.0623	0.6897	6.497×10^{-7}
9	0.0623	0.06273	0.06252	0.3436	-1.265×10^{-6}
10	0.0623	0.06252	0.06241	0.1721	-3.0768×10^{-7}

At the end of 10th iteration (10.yinelemenin sonunda),

$$|\in_a| = 0.1721\%$$

Hence the number of significant digits at least correct is given by the largest value of m for which (Bu durumda doğru olarak verilebilecek anlamlı hane sayısı m'nin alacağı en büyük değer kadardır)

$$|\in_a| \le 0.5\times10^{2-m}$$

$$0.1721 \le 0.5\times10^{2-m}$$

$$0.3442 \le 10^{2-m}$$

$$\log(0.3442) \le 2-m$$

$$m \le 2-\log(0.3442) = 2.463$$

So

$$m = 2$$

The number of significant digits at least correct in the estimated root of 0.06241 at the end of the 10^{th} iteration is 2.

Example 2 (Mechanical Engineering)

A trunnion has to be cooled before it is shrink fitted into a steel hub.

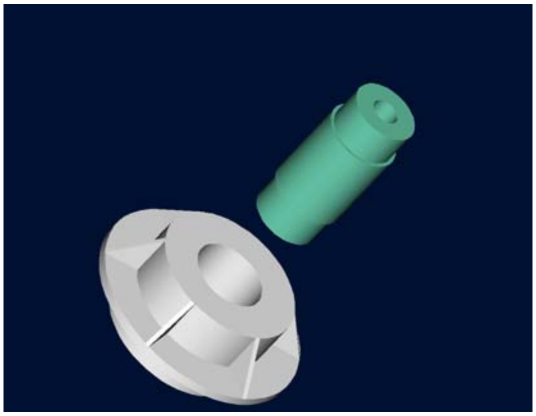


Figure 1 Trunnion to be slid through the hub after contracting.

The equation that gives the temperature T_f to which the trunnion has to be cooled to obtain the desired contraction is given by

$$f(T_f) = -0.50598\times10^{-10}T_f^3 + 0.38292\times10^{-7}T_f^2 + 0.74363\times10^{-4}T_f + 0.88318\times10^{-2} = 0$$

Use the bisection method of finding roots of equations to find the temperature T_f to which the trunnion has to be cooled. 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 designer's records for the previous bridge, the temperature to which the trunnion was cooled was $-108\,°\text{F}$. Hence assuming the temperature to be between $-100\,°\text{F}$ and $-150\,°\text{F}$, we have

$$T_{f,\ell} = -150\,°\text{F},\ T_{f,u} = -100\,°\text{F}$$

Check if the function changes sign between $T_{f,\ell}$ and $T_{f,u}$.

$$\begin{aligned} f\left(T_{f,\ell}\right) &= f(-150) \\ &= -0.50598\times10^{-10}(-150)^3 + 0.38292\times10^{-7}(-150)^2 \\ &\quad + 0.74363\times10^{-4}(-150) + 0.88318\times10^{-2} \\ &= -1.2903\times10^{-3} \end{aligned}$$

$$\begin{aligned} f\left(T_{f,u}\right) &= f(-100) \\ &= -0.50598\times10^{-10}(-100)^3 + 0.38292\times10^{-7}(-100)^2 \\ &\quad + 0.74363\times10^{-4}(-100) + 0.88318\times10^{-2} \\ &= 1.8290\times10^{-3} \end{aligned}$$

Hence

$$f\left(T_{f,\ell}\right)f\left(T_{f,u}\right) = f(-150)f(-100) = \left(-1.2903\times10^{-3}\right)\left(1.8290\times10^{-3}\right) < 0$$

So there is at least one root between $T_{f,\ell}$ and $T_{f,u}$ that is between -150 and -100 .

Iteration 1

The estimate of the root is

$$\begin{aligned}T_{f,m} &= \frac{T_{f,\ell} + T_{f,u}}{2}\\ &= \frac{-150 + (-100)}{2}\\ &= -125\end{aligned}$$

$$\begin{aligned}f\left(T_{f,m}\right) &= f\left(-125\right)\\ &= -0.50598\times 10^{-10}(-125)^3 + 0.38292\times 10^{-7}(-125)^2\\ &\quad + 0.74363\times 10^{-4}(-125) + 0.88318\times 10^{-2}\\ &= 2.3356\times 10^{-4}\end{aligned}$$

$$f\left(T_{f,\ell}\right)f\left(T_{f,m}\right) = f\left(-150\right)f\left(-125\right) = \left(-1.2903\times 10^{-3}\right)\left(2.3356\times 10^{-4}\right) < 0$$

Hence the root is bracketed between $T_{f,\ell}$ and $T_{f,m}$, that is, between -150 and -125.

So, the lower and upper limits of the new bracket are

$$T_{f,\ell} = -150,\ T_{f,u} = -125$$

At this point, the absolute relative approximate error $\left|\in_a\right|$ cannot be calculated, as we do not have a previous approximation.

Iteration 2

The estimate of the root is

$$\begin{aligned}T_{f,m} &= \frac{T_{f,\ell} + T_{f,u}}{2}\\ &= \frac{-150 + \left(-125\right)}{2}\\ &= -137.5\end{aligned}$$

$$\begin{aligned}f\left(T_{f,m}\right) &= f\left(-137.5\right)\\ &= -0.50598\times 10^{-10}(-137.5)^3 + 0.38292\times 10^{-7}(-137.5)^2\\ &\quad + 0.74363\times 10^{-4}(-137.5) + 0.88318\times 10^{-2}\\ &= -5.3762\times 10^{-4}\end{aligned}$$

$$f\left(T_{f,m}\right)f\left(T_{f,u}\right) = f\left(-137.5\right)f\left(-125\right) = \left(-5.3762\times 10^{-4}\right)\left(2.3356\times 10^{-4}\right) < 0$$

Hence, the root is bracketed between $T_{f,m}$ and $T_{f,u}$, that is, between -125 and -137.5.

So the lower and upper limits of the new bracket are

$$T_{f,\ell} = -137.5, T_{f,u} = -125$$

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

$$\left|\in_a\right| = \left|\frac{T_{f,m}^{\text{new}} - T_{f,m}^{\text{old}}}{T_{f,m}^{\text{new}}}\right|\times 100$$

$$= \left| \frac{-137.5 - (-125)}{-137.5} \right| \times 100$$

$$= 9.0909\%$$

None of the significant digits are at least correct in the estimated root of

$$T_{f,m} = -137.5$$

as the absolute relative approximate error is greater that 5%.

Iteration 3

The estimate of the root is

$$T_{f,m} = \frac{T_{f,\ell} + T_{f,u}}{2}$$

$$= \frac{-137.5 + (-125)}{2}$$

$$= -131.25$$

$$f\left(T_{f,m}\right) = f\left(-131.25\right)$$

$$= -0.50598 \times 10^{-10} (-131.25)^3 + 0.38292 \times 10^{-7} (-131.25)^2$$

$$+ 0.74363 \times 10^{-4} (-131.25) + 0.88318 \times 10^{-2}$$

$$= -1.54303 \times 10^{-4}$$

$$f\left(T_{f,\ell}\right) f\left(T_{f,m}\right) = f\left(-125\right) f\left(-131.25\right) = \left(2.3356 \times 10^{-4}\right)\left(-1.5430 \times 10^{-4}\right) < 0$$

Hence, the root is bracketed between $T_{f,\ell}$ and $T_{f,m}$, that is, between -125 and -131.25. So the lower and upper limits of the new bracket are

$$T_{f,\ell} = -131.25, T_{f,u} = -125$$

The absolute relative approximate error $\left|\in_a\right|$ at the ends of Iteration 3 is

$$\left|\in_a\right| = \left| \frac{T_{f,m}^{\text{new}} - T_{f,m}^{\text{old}}}{T_{f,m}^{\text{new}}} \right| \times 100$$

$$= \left| \frac{-131.25 - (-137.5)}{-131.25} \right| \times 100$$

$$= 4.7619\%$$

The number of significant digits at least correct is 1.

Seven more iterations were conducted and these iterations are shown in the Table 1 below.

Table 1 Root of $f\left(x\right) = 0$ as function of number of iterations for bisection method.

Iteration	$T_{f,\ell}$	$T_{f,u}$	$T_{f,m}$	$\lvert\in_a\rvert\%$	$f(T_{f,m})$
1	−150	−100	−125	---------	2.3356×10^{-4}
2	−150	−125	−137.5	9.0909	-5.3762×10^{-4}
3	−137.5	−125	−131.25	4.7619	-1.5430×10^{-4}
4	−131.25	−125	−128.13	2.4390	3.9065×10^{-5}
5	−131.25	−128.13	−129.69	1.2048	-5.7760×10^{-5}
6	−129.69	−123.13	−128.91	0.60606	-9.3826×10^{-6}
7	−128.91	−123.13	−128.52	0.30395	1.4838×10^{-5}
8	−128.91	−128.52	−128.71	0.15175	2.7228×10^{-6}
9	−128.91	−128.71	−128.81	0.075815	-3.3305×10^{-6}
10	−128.81	−128.71	−128.76	0.037922	-3.0396×10^{-7}

At the end of the 10^{th} iteration,

$$\lvert\in_a\rvert = 0.037922\%$$

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

$$\lvert\in_a\rvert \le 0.5\times10^{2-m}$$
$$0.037922 \le 0.5\times10^{2-m}$$
$$0.075844 \le 10^{2-m}$$
$$\log(0.075844) \le 2-m$$
$$m \le 2-\log(0.075844) = 3.1201$$

So

$$m = 3$$

The number of significant digits at least correct in the estimated root of -128.76 is 3.

Example 3 (Industrial Engineering)

You are working for a start-up computer assembly company and have been asked to determine the minimum number of computers that the shop will have to sell to make a profit.

The equation that gives the minimum number of computers n to be sold after considering the total costs and the total sales is

$$f(n) = 40n^{1.5} - 875n + 35000 = 0$$

Use the bisection method of finding roots of equations to find the minimum number of computers that need to be sold to make a profit. 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

Let us assume

$$n_\ell = 50,\ n_u = 100$$

Check if the function changes sign between n_ℓ and n_u.

$$f(n_\ell) = f(50) = 40(50)^{1.5} - 875(50) + 35000 = 5392.1$$
$$f(n_u) = f(100) = 40(100)^{1.5} - 875(100) + 35000 = -12500$$

Hence

$$f(n_\ell)f(n_u) = f(50)f(100) = (5392.1)(-12500) < 0$$

So there is at least one root between n_ℓ and n_u, that is, between 50 and 100.

Iteration 1

The estimate of the root is

$$\begin{aligned} n_m &= \frac{n_\ell + n_u}{2} \\ &= \frac{50+100}{2} \\ &= 75 \end{aligned}$$

$$f(n_m) = f(75) = 40(75)^{1.5} - 875(75) + 35000 = -4.6442 \times 10^3$$
$$f(n_\ell)f(n_m) = f(50)f(75) = (5392.1)(-4.6442 \times 10^3) < 0$$

Hence the root is bracketed between n_ℓ and n_m, that is, between 50 and 75. So, the lower and upper limits of the new bracket are

$$n_\ell = 50, n_u = 75$$

At this point, the absolute relative approximate error $|\in_a|$ cannot be calculated, as we do not have a previous approximation.

Iteration 2

The estimate of the root is

$$\begin{aligned} n_m &= \frac{n_\ell + n_u}{2} \\ &= \frac{50+75}{2} \\ &= 62.5 \end{aligned}$$

$$f(n_m) = f(62.5) = 40(62.5)^{1.5} - 875(62.5) + 35000 = 76.735$$
$$f(n_\ell)f(n_m) = f(50)f(62.5) = (5392.1)(76.735) > 0$$

Hence, the root is bracketed between n_m and n_u, that is, between 62.5 and 75. So the lower and upper limits of the new bracket are

$$n_\ell = 62.5, n_u = 75$$

The absolute relative approximate error, $|\in_a|$ at the end of Iteration 2 is

$$\begin{aligned} |\in_a| &= \left| \frac{n_m^{\text{new}} - n_m^{\text{old}}}{n_m^{\text{new}}} \right| \times 100 \\ &= \left| \frac{62.5 - 75}{62.5} \right| \times 100 \\ &= 20\% \end{aligned}$$

None of the significant digits are at least correct in the estimated root

$$n_m = 62.5$$

as the absolute relative approximate error is greater that 5%.

Iteration 3

The estimate of the root is

$$\begin{aligned} n_m &= \frac{n_\ell + n_u}{2} \\ &= \frac{62.5 + 75}{2} \\ &= 68.75 \end{aligned}$$

$$f(n_m) = f(68.75) = 40(68.75)^{1.5} - 875(68.75) + 35000 = -2.3545 \times 10^3$$

$$f(n_\ell) f(n_m) = f(62.5) f(68.75) = (76.735)(-2.3545 \times 10^3) < 0$$

Hence, the root is bracketed between n_ℓ and n_m, that is, between 62.5 and 68.75. So the lower and upper limits of the new bracket are

$$n_\ell = 62.5,\ n_u = 68.75$$

The absolute relative approximate error $\left|\in_a\right|$ at the end of Iteration 3 is

$$\begin{aligned} \left|\in_a\right| &= \left|\frac{n_m^{\text{new}} - n_m^{\text{old}}}{n_m^{\text{new}}}\right| \times 100 \\ &= \left|\frac{68.75 - 62.5}{68.75}\right| \times 100 \\ &= 9.0909\% \end{aligned}$$

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%. The estimated minimum number of computers that need to be sold to break even at the end of the third iteration is 69. Seven more iterations were conducted and these iterations are shown in the Table 1.

Table 1 Root of $f(x) = 0$ as a function of the number of iterations for bisection method.

Iteration	n_ℓ	n_u	n_m	$\left\|\in_a\right\|\%$	$f(n_m)$
1	50	100	75	----------	-4.6442×10^3
2	50	75	62.5	20	76.735
3	62.5	75	68.75	9.0909	-2.3545×10^3
4	62.5	68.75	65.625	4.7619	-1.1569×10^3
5	62.5	65.625	64.063	2.4390	−544.68
6	62.5	64.063	63.281	1.2346	−235.12
7	62.5	63.281	62.891	0.62112	−79.483
8	62.5	62.891	62.695	0.31153	−1.4459
9	62.5	62.695	62.598	0.15601	37.627
10	62.598	62.695	62.646	0.077942	18.086

At the end of the 10th iteration,

$$|\in_a| = 0.077942\%$$

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

$$|\in_a| \le 0.5\times10^{2-m}$$

$$0.077942 \le 0.5\times10^{2-m}$$

$$0.15588 \le 10^{2-m}$$

$$\log(0.15588) \le 2-m$$

$$m \le 2-\log(0.15588) = 2.8072$$

So

$$m = 2$$

The number of significant digits at least correct in the estimated root 62.646 is 2.

Example 1 (Electrical Engineering)

Thermistors are temperature-measuring devices based on the principle that the thermistor material exhibits a change in electrical resistance with a change in temperature. By measuring the resistance of the thermistor material, one can then determine the temperature.

For a 10K3A Betatherm thermistor,

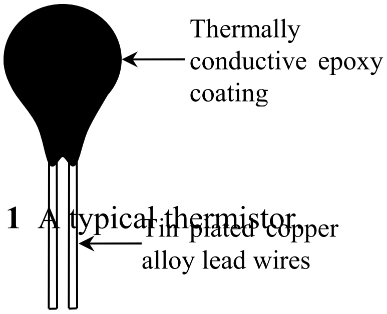

Thermally conductive epoxy coating
Tin plated copper alloy lead wires

Figure 1 A typical thermistor.

the relationship between the resistance R of the thermistor and the temperature is given by

$$\frac{1}{T} = 1.129241\times10^{-3} + 2.341077\times10^{-4}\ln(R) + 8.775468\times10^{-8}\{\ln(R)\}^3$$

where T is in Kelvin and R is in ohms.

A thermistor error of no more than $\pm 0.01°\text{C}$ is acceptable. To find the range of the resistance that is within this acceptable limit at $19\ °\text{C}$, we need to solve

$$\frac{1}{19.01+273.15} = 1.129241\times10^{-3} + 2.341077\times10^{-4}\ln(R) + 8.775468\times10^{-8}\{\ln(R)\}^3$$

and

$$\frac{1}{18.99+273.15} = 1.129241\times10^{-3} + 2.341077\times10^{-4}\ln(R) + 8.775468\times10^{-8}\{\ln(R)\}^3$$

Use the bisection method of finding roots of equations to find the resistance R at $18.99°\text{C}$. 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

Solving

$$\frac{1}{18.99+273.15}=1.129241\times10^{-3}+2.341077\times10^{-4}\ln(R)+8.775468\times10^{-8}\{\ln(R)\}^3$$

we get

$$f(R)=2.341077\times10^{-4}\ln(R)+8.775468\times10^{-8}\{\ln(R)\}^3-2.293775\times10^{-3}$$

Lets us assume

$$R_\ell=11000,\ R_u=14000$$

Check if the function changes sign between R_ℓ and R_u.

$$\begin{aligned}f(R_\ell)&=f(11000)\\&=2.341077\times10^{-4}\ln(11000)+8.775468\times10^{-8}\{\ln(11000)\}^3-2.293775\times10^{-3}\\&=-4.4536\times10^{-5}\end{aligned}$$

$$\begin{aligned}f(R_u)&=f(14000)\\&=2.341077\times10^{-4}\ln(14000)+8.775468\times10^{-8}\{\ln(14000)\}^3-2.293775\times10^{-3}\\&=1.7563\times10^{-5}\end{aligned}$$

Hence

$$f(R_\ell)f(R_u)=f(11000)f(14000)=(-4.4536\times10^{-5})(1.7563\times10^{-5})<0$$

So there is at least one root between R_ℓ and R_u, that is, between 11000 and 14000.

Iteration 1

The estimate of the root is

$$\begin{aligned}R_m&=\frac{R_\ell+R_u}{2}\\&=\frac{11000+14000}{2}\\&=12500\end{aligned}$$

$$\begin{aligned}f(R_m)&=f(12500)\\&=2.341077\times10^{-4}\ln(12500)+8.775468\times10^{-8}\{\ln(12500)\}^3-2.293775\times10^{-3}\\&=-1.1655\times10^{-5}\end{aligned}$$

$$f(R_\ell)f(R_m)=f(11000)f(12500)=(-4.4536\times10^{-5})(-1.1655\times10^{-5})>0$$

Hence the root is bracketed between R_m and R_u, that is, between 12500 and 14000. So, the lower and upper limits of the new bracket are

$$R_\ell=12500,\ R_u=14000$$

At this point, the absolute relative approximate error $\left|\in_a\right|$ cannot be calculated as we do not have a previous approximation.

Iteration 2

The estimate of the root is

$$
\begin{aligned}
R_m &= \frac{R_\ell + R_u}{2} \\
&= \frac{12500 + 14000}{2} \\
&= 13250
\end{aligned}
$$

$$
\begin{aligned}
f(R_m) &= f(13250) \\
&= 2.341077\times10^{-4}\ln(13250) + 8.775468\times10^{-8}\{\ln(13250)\}^3 - 2.293775\times10^{-3} \\
&= 3.3599\times10^{-6}
\end{aligned}
$$

$$f(R_\ell)f(R_m) = f(12500)f(13250) = \left(-1.1655\times10^{-5}\right)\left(3.3599\times10^{-6}\right) < 0$$

Hence, the root is bracketed between R_ℓ and R_m, that is, between 12500 and 13250.
So the lower and upper limits of the new bracket are

$$R_\ell = 12500,\ R_u = 13250$$

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

$$
\begin{aligned}
\left|\in_a\right| &= \left|\frac{R_m^{\text{new}} - R_m^{\text{old}}}{R_m^{\text{new}}}\right| \times 100 \\
&= \left|\frac{13250 - 12500}{13250}\right| \times 100 \\
&= 5.6604\%
\end{aligned}
$$

None of the significant digits are at least correct in the estimated root

$$R_m = 13250$$

as the absolute relative approximate error is greater than 5%.

Iteration 3

$$
\begin{aligned}
R_m &= \frac{R_\ell + R_u}{2} \\
&= \frac{12500 + 13250}{2} \\
&= 12875
\end{aligned}
$$

$$
\begin{aligned}
f(R_m) &= f(12875) \\
&= 2.341077\times10^{-4}\ln(12875) + 8.775468\times10^{-8}\{\ln(12875)\}^3 - 2.293775\times10^{-3} \\
&= -4.0403\times10^{-6}
\end{aligned}
$$

$$f(R_\ell)f(R_m) = f(12500)f(12875) = \left(\left(-1.1654\times10^{-5}\right)\right)\left(-4.0398\times10^{-6}\right) > 0$$

Hence, the root is bracketed between R_m and R_u, that is, between 12875 and 13250.

So, the lower and upper limits of the new bracket are

$$R_\ell = 12875,\ R_u = 13250$$

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

$$|\in_a| = \left|\frac{R_m^{\text{new}} - R_m^{\text{old}}}{R_m^{\text{new}}}\right| \times 100$$

$$= \left|\frac{12875 - 13250}{12875}\right| \times 100$$

$$= 2.9126\%$$

One of the significant digits is at least correct in the estimated root of the equation as the absolute relative approximate error is less than 5%.

Seven more iterations were conducted and these iterations are shown in the Table 1.

Table 1 Root of $f(x) = 0$ as a function of the number of iterations for bisection method.

Iteration	R_ℓ	R_u	R_m	$\lvert\in_a\rvert\%$	$f(R_m)$
1	11000	14000	12500	----------	1.1655×10^{-5}
2	12500	14000	13250	5.6604	3.3599×10^{-6}
3	12500	13250	12875	2.9126	-4.0403×10^{-6}
4	12875	13250	13063	1.4354	-3.1417×10^{-7}
5	13063	13250	13156	0.71259	1.5293×10^{-6}
6	13063	13156	13109	0.35757	6.0917×10^{-7}
7	13063	13109	13086	0.17910	1.4791×10^{-7}
8	13063	13086	13074	0.089633	-8.3022×10^{-8}
9	13074	13086	13080	0.044796	3.2470×10^{-8}
10	13074	13080	13077	0.022403	-2.5270×10^{-8}

At the end of the 10^{th} iteration,

$$|\in_a| = 0.022403\%$$

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

$$|\in_a| \le 0.5\times10^{2-m}$$

$$0.022403 \le 0.5\times10^{2-m}$$

$$0.044806 \le 10^{2-m}$$

$$\log(0.044806) \le 2 - m$$

$$m \le 2 - \log(0.044806) = 3.3487$$

So

$$m = 3$$

The number of significant digits at least correct in the estimated root 13077 is 3.

Advantages of bisection method (ikiye bölme yönteminin avantajları)

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 (ikiye bölme yönteminin dez avantajları)

a) The convergence of the bisection method is slow as it is simply based on halving the interval (ikiye bölme yöntemiyle sonuca yaklaşmak/yakınsamak uzun zaman alabilir).

b) If one of the initial guesses is closer to the root, it will take larger number of iterations to reach the root (Köke çok yakın başlangıç tahmininde bulunulursa, köke ulaşmak için daha fazla yineleme yapmak gerekebilir).

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_ℓ, and upper guess, x_u, such that

$$f(x_\ell)f(x_u) < 0$$

d) For functions $f(x)$ where there is a singularity[1] 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.

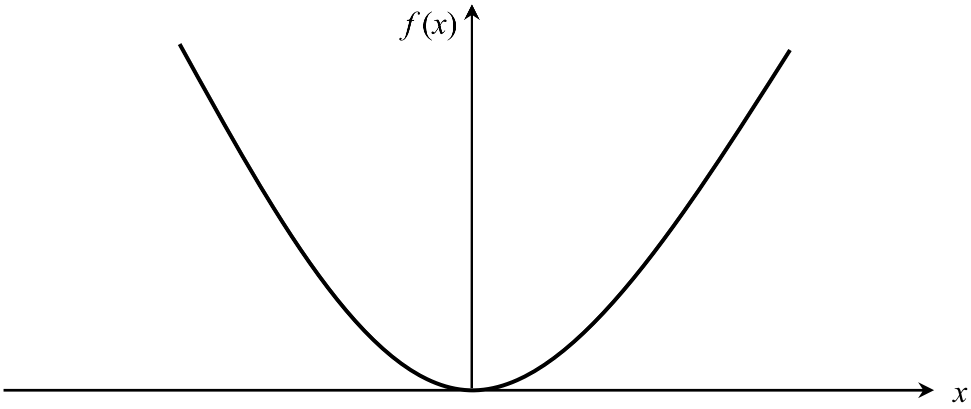


Figure 6 The equation $f(x) = x^2 = 0$ has a single root at $x = 0$ that cannot be bracketed.

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.

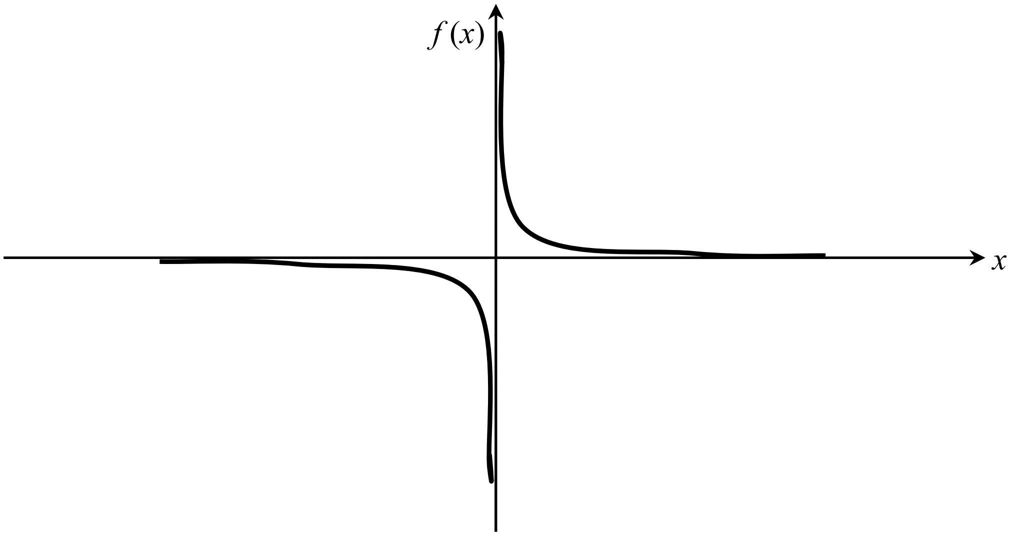


Figure 7 The equation $f(x) = \frac{1}{x} = 0$ has no root but changes sign.

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	Aralık 8, 2016
Web Site	http://numericalmethods.eng.usf.edu

3.3.1 Multiple-Choice Test Chapter 03.03 Bisection Method

1. The bisection method of finding roots of nonlinear equations falls under the category of a (an) __________ method.

(A) open (B) bracketing (C) random (D) graphical

2. If $f(x)$ is a real continuous function in $[a,b]$, and $f(a)f(b) < 0$, then for $f(x) = 0$, there is (are) _____________ in the domain $[a,b]$.

(A) one root (B) an undeterminable number of roots (C) no root (D) at least one root

3. Assuming an initial bracket of $[1,5]$, the second (at the end of 2 iterations) iterative value of the root of $te^{-t} - 0.3 = 0$ using the bisection method is

(A) 0 (B) 1.5 (C) 2 (D) 3

```
[1.00,5.00] birinci kök değeri  : 3.000
 fonksiyonun xl değeri          : 0.068
 fonksiyonun değeri             : -0.151
 kök hesabındaki hata           : ----
 [1.00,3.00]  ikinci kök değeri : 2.000
 fonksiyonun xl değeri          : 0.068
 fonksiyonun xm değeri          : -0.029
 kök hesabındaki hata           : 50.000
```

4. To find the root of $f(x) = 0$, a scientist is using the bisection method. At the beginning of an iteration, the lower and upper guesses of the root are x_l and x_u. At the end of the iteration, the absolute relative approximate error in the estimated value of the root would be

(A) $\left|\dfrac{x_u}{x_u + x_\ell}\right|$ (B) $\left|\dfrac{x_\ell}{x_u + x_\ell}\right|$ (C) $\left|\dfrac{x_u - x_\ell}{x_u + x_\ell}\right|$ (D) $\left|\dfrac{x_u + x_\ell}{x_u - x_\ell}\right|$

5. For an equation like $x^2 = 0$, a root exists at $x = 0$. The bisection method cannot be adopted to solve this equation in spite of the root existing at $x = 0$ because the function $f(x) = x^2$

(A) is a polynomial (B) has repeated roots at $x = 0$

(C) is always non-negative (D) has a slope equal to zero at $x = 0$

6. The ideal gas law is given by

$$pv = RT$$

where p is the pressure, v is the specific volume, R is the universal gas constant, and T is the absolute temperature. This equation is only accurate for a limited range of pressure and temperature. Vander Waals came up with an equation that was accurate for larger ranges of pressure and temperature given by

$$\left(p + \frac{a}{v^2}\right)(v - b) = RT$$

Where a and b are empirical constants dependent on a particular gas. Given the value of $R = 0.08$, $a = 3.592$, $b = 0.04267$, $p = 10$ and $T = 300$ (assume all units are consistent), one is going to find the specific volume, v, for the above values. Without finding the solution from the Vander Waals equation, what would be a good initial guess for v?

(A) 0 (B) 1.2 (C) 2.4 (D) 3.6

For a complete solution, refer to the links at the end of the book.

3.4 Chapter 03.04 Newton-Raphson Method of Solving a Nonlinear Equation

PRE-REQUISITES

1. Know the definition of a derivative of a function (Primer for Differential Calculus).
2. Be able to find derivatives of function (Primer for Differential Calculus).
3. Know what a tangent to a curve is and how to find the tangent line (Primer for Differential Calculus).

OBJECTIVES

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.

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 (f(x)=0 şeklindeki çizgisel olmayan bir eşitliğin bisection ve false position gibi yöntemlerde çözüm aralığının tahmin edilmesi gerekmektedir). Such methods are called *bracketing methods (aralık yöntemi)*. 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 (eşitliği sıfıra yaklaştıracak aralığı daralttıkça bu yöntemlerle bir sonuca ulaşılabilir).

In the Newton-Raphson method, the root is not bracketed (Newton-Raphson yönteminde kök bölgesi yoktur). In fact, only one initial guess of the root is needed to get the iterative process started to find the root of an equation (aslında bir başlangıç tahmini kök değeri yinelemeyi başlatmak için yeterli olacaktır). The method hence falls in the category of *open methods (yöntem açık yöntem kategorisine alınabilir)*. 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) (Newton-Raphson yöntemi f(x)=0'ın başlangıç kökü için x_i'yi alır ve bu noktadaki tanjantının/teğetinin yatay ekseni kestiği yeri x_{i+1} olarak daha iyi bir kök değerini bulmaya çalışır).

Using the definition of the slope of a function, at $x = x_i$

$$f'(x_i) = \tan\theta \qquad = \frac{f(x_i) - 0}{x_i - x_{i+1}},$$

which gives

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

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 $|\in_a|$ as
$$|\in_a| = \left|\frac{x_{i+1} - x_i}{x_{i+1}}\right| \times 100$$
4. Compare the absolute relative approximate error with the pre-specified relative error tolerance, $\in_s$. If $|\in_a| > \in_s$, 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.

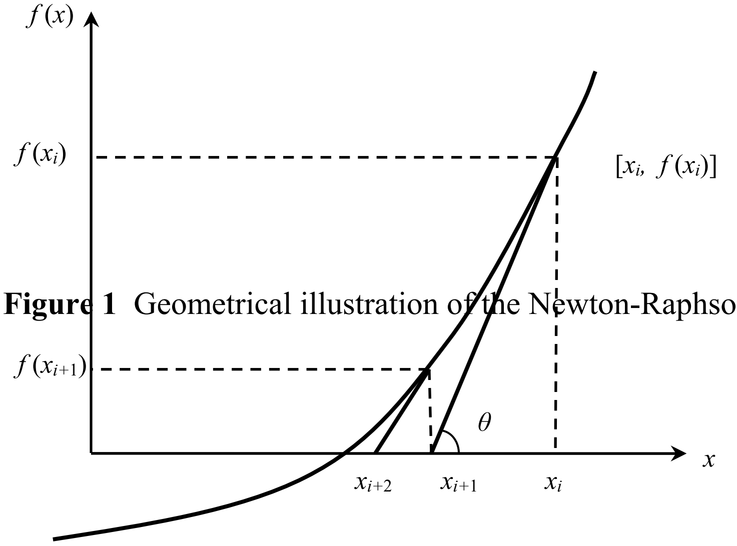


Figure 1 Geometrical illustration of the Newton-Raphson method.

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.

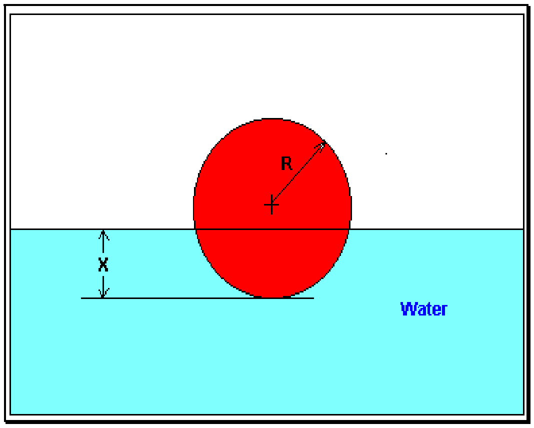


Figure 2 Floating ball problem.

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

a) the depth x to which the ball is submerged under water. Conduct three iterations to estimate the root of the above equation.
b) the absolute relative approximate error at the end of each iteration, and
c) 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 $x_0 = 0.05\text{ m}$. This is a reasonable guess (discuss why $x = 0$ and $x = 0.11\text{ m}$ 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 $|\in_a|$ at the end of Iteration 1 is

$$
\begin{aligned}
\left|\in_a\right| &= \left|\frac{x_1 - x_0}{x_1}\right| \times 100 \\
&= \left|\frac{0.06242 - 0.05}{0.06242}\right| \times 100 \\
&= 19.90\%
\end{aligned}
$$

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

$$
\begin{aligned}
x_2 &= x_1 - \frac{f(x_1)}{f'(x_1)} \\
&= 0.06242 - \frac{(0.06242)^3 - 0.165(0.06242)^2 + 3.993\times 10^{-4}}{3(0.06242)^2 - 0.33(0.06242)} \\
&= 0.06242 - \frac{-3.97781\times 10^{-7}}{-8.90973\times 10^{-3}} \\
&= 0.06242 - \left(4.4646\times 10^{-5}\right) \\
&= 0.06238
\end{aligned}
$$

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

$$
\begin{aligned}
\left|\in_a\right| &= \left|\frac{x_2 - x_1}{x_2}\right| \times 100 \\
&= \left|\frac{0.06238 - 0.06242}{0.06238}\right| \times 100 \\
&= 0.0716\%
\end{aligned}
$$

The maximum value of m for which $\left|\in_a\right| \le 0.5\times 10^{2-m}$ 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

$$
\begin{aligned}
x_3 &= x_2 - \frac{f(x_2)}{f'(x_2)} \\
&= 0.06238 - \frac{(0.06238)^3 - 0.165(0.06238)^2 + 3.993\times 10^{-4}}{3(0.06238)^2 - 0.33(0.06238)} \\
&= 0.06238 - \frac{4.44\times 10^{-11}}{-8.91171\times 10^{-3}} \\
&= 0.06238 - \left(-4.9822\times 10^{-9}\right) \\
&= 0.06238
\end{aligned}
$$

The absolute relative approximate error $\left|\in_a\right|$ at the end of Iteration 3 is

$$|\in_a| = \left|\frac{0.06238 - 0.06238}{0.06238}\right| \times 100$$
$$= 0$$

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 (dönüm noktalarındaki ıraksamalar)

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 (Başlangıç tahmin değeri veya iterasyon sonucunda elde edilen bir değer f(x)=0 eşitliğindeki fonksiyonun dönüm noktasına gelebilir, Newton-Rapson yöntemi ıraksamaya başlarsa kökten uzaklaşılabilir). It may then start converging back to the root (Sonra tekrar köke doğru yakınsayabilir). 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

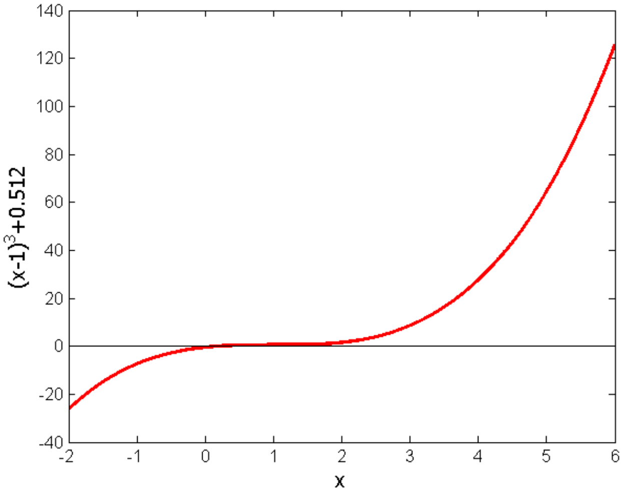


Figure 3 Divergence at inflection point for $f(x)=(x-1)^3=0$.

2. Division by zero

For the equation

$$f(x)=x^3-0.03x^2+2.4\times10^{-6}=0$$

the Newton-Raphson method reduces to

$$x_{i+1}=x_i-\frac{x_i^{\,3}-0.03x_i^{\,2}+2.4\times10^{-6}}{3x_i^{\,2}-0.06x_i}$$

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 (payda) in the formula is a small number. For this case, as given in Table 2, even after 9 iteration, the Newton-Raphson method does not converge (Çizelge 2’de verildiği gibi buradaki durumda, 9 yinelemeye rağmen Newton-Raphson yöntemi köke yakınsamamaktadır).

Table 2 Division by near zero in Newton-Raphson method.

Iteration Number	x_i	$f(x_i)$	$\lvert\in_a\rvert\%$
0	0.019990	-1.60000×10^{-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

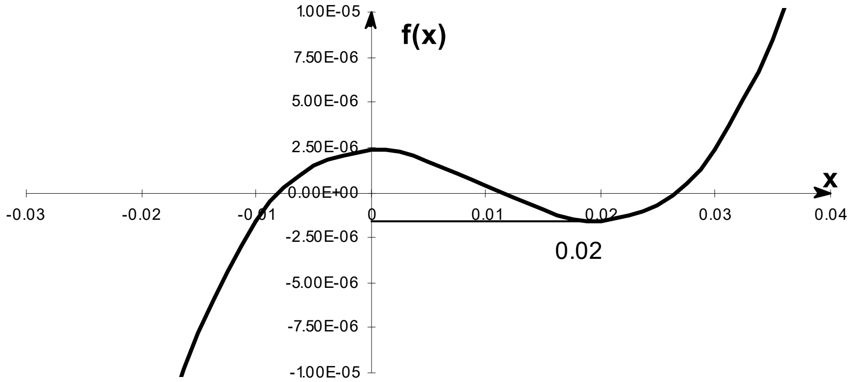


Figure 4 Pitfall of division by zero or a near zero number (Sıfıra bölme veya sıfıra yakın sayı tuzağı).

3. Oscillations near local maximum and minimum (maksimum ve minimum çevresinde salınım)

Results obtained from the Newton-Raphson method may oscillate about the local maximum or minimum without converging on a root but converging (yaklaşma) on the local maximum or minimum. Eventually (sonunda), it may lead to division by a number close to zero and may diverge (sapma).

For example, for

$$f(x) = x^2 + 2 = 0$$

the equation has no real roots (Figure 5 and Table 3).

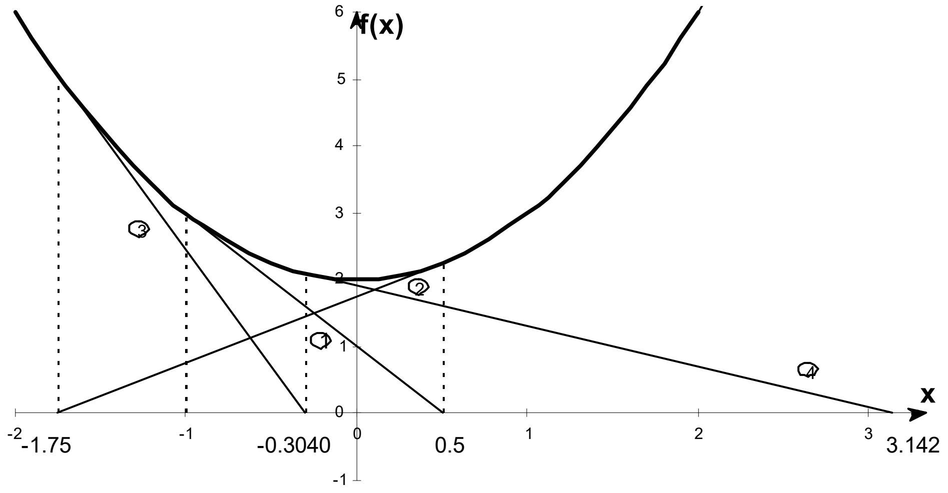


Figure 5 Oscillations around local minima for $f(x) = x^2 + 2$.

Table 3 Oscillations near local maxima and minima in Newton-Raphson method.

Iteration Number	x_i	$f(x_i)$	$\lvert\in_a\rvert\%$
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 (kökü atlamak)

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	$f(x_i)$	$\lvert\in_a\rvert\%$
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	8.376×10^{-4}	8.375×10^{-5}	7.54×10^{4}
5	-1.95861×10^{-13}	-1.95861×10^{-13}	4.28×10^{10}

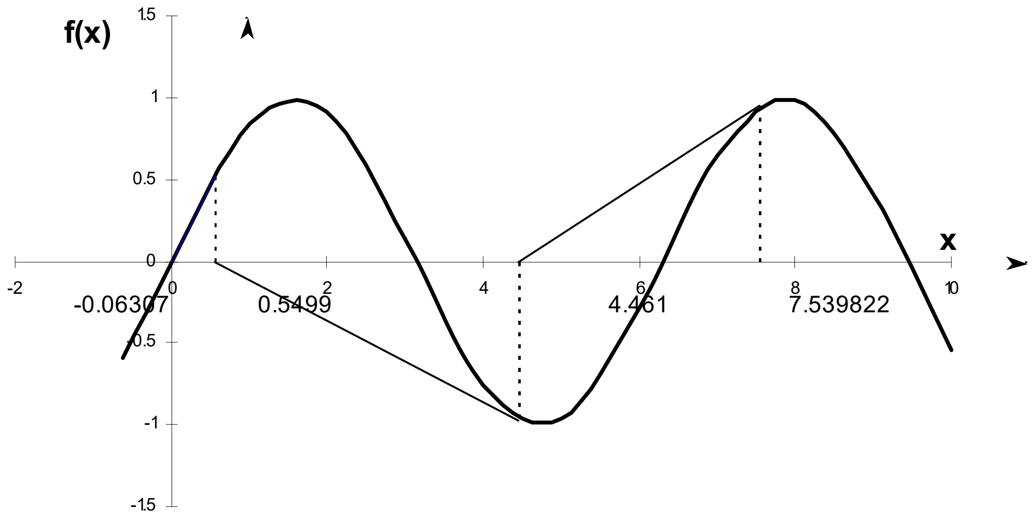


Figure 6 Root jumping from intended location of root for $f(x) = \sin x = 0$.

Appendix A. What is an inflection point? (dönüm noktası nedir)

For a function $f(x)$, the point where the concavity (çukurluk) 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 yönteminin Taylor serilerinden Elde edilmesi)

Newton-Raphson method can also be derived from Taylor series. For a general function $f(x)$, the Taylor series is

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

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	Aralık 8, 2016
Web Site	http://numericalmethods.eng.usf.edu

3.4.1 Multiple-Choice Test Chapter 03.04 Newton-Raphson Method

1. The Newton-Raphson method of finding roots of nonlinear equations falls under the category of _____________ methods.

(A) bracketing (B) open (C) random (D) graphical

2. The Newton-Raphson method formula for finding the square root of a real number R from the equation $x^2 - R = 0$ is,

(A) $x_{i+1} = \frac{x_i}{2}$ (B) $x_{i+1} = \frac{3x_i}{2}$ (C) $x_{i+1} = \frac{1}{2}\left(x_i + \frac{R}{x_i}\right)$ (D) $x_{i+1} = \frac{1}{2}\left(3x_i - \frac{R}{x_i}\right)$

3. The next iterative value of the root of $x^2 - 4 = 0$ using the Newton-Raphson method, if the initial guess is 3, is

(A) 1.5 (B) 2.067 (C) 2.167 (D) 3.000

4. The root of the equation $f(x) = 0$ is found by using the Newton-Raphson method. The initial estimate of the root is $x_0 = 3$, $f(3) = 5$. The angle the line tangent to the function $f(x)$ makes at $x = 3$ is $57°$ with respect to the x-axis. The next estimate of the root, x_1 most nearly is

(A) –3.2470 (B) −0.2470 (C) 3.2470 (D) 6.2470

5. The root of $x^3 = 4$ is found by using the Newton-Raphson method. The successive iterative values of the root are given in the table below.

Iteration Number	Value of Root
0	2.0000

1	1.6667
2	1.5911
3	1.5874
4	1.5874

The iteration number at which I would first trust at least two significant digits in the answer is

(A) 1 (B) 2 (C) 3 (D) 4

6. The ideal gas law is given by

$$pv = RT$$

where p is the pressure, v is the specific volume, R is the universal gas constant, and T is the absolute temperature. This equation is only accurate for a limited range of pressure and temperature. Vander Waals came up with an equation that was accurate for larger ranges of pressure and temperature given by

$$\left(p+\frac{a}{v^2}\right)(v-b)=RT$$

where a and b are empirical constants dependent on a particular gas. Given the value of $R = 0.08$, $a = 3.592$, $b = 0.04267$, $p = 10$ and $T = 300$ (assume all units are consistent), one is going to find the specific volume, v, for the above values. Without finding the solution from the Vander Waals equation, what would be a good initial guess for v?

(A) 0 (B) 1.2 (C) 2.4 (D) 3.6

For a complete solution, refer to the links at the end of the book.

3.5 Chapter 03.05 Secant Method of Solving a Nonlinear Equation

PRE-REQUISITES

1. Know the definition of a derivative of a function (Primer for Differential Calculus).
2. Be able to find derivatives of function (Primer for Differential Calculus).
3. Know what a secant line is and how to find the equation of the secant line (Primer for Differential Calculus).

OBJECTIVES

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.

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

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

One of the drawbacks (dez avantajları) 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 (Newton-Raphson yönteminin dez avantajlarından biri fonksiyonun türevinin olması gerektiğidir. Sembolik programlama dili kullanan Maple, MathCAD, MATHEMATICA ve MATLAB gibi programlar ile kök kolayca bulunabilir. Ancak, Newton-Raphson yönteminin uygulaması zahmetli ve fonksiyonun türevinin alınması gerektiğinden dolayı zordur. Bu dezavantajı)

$$f'(x_i) = \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}} \tag{2}$$

Substituting Equation (2) in Equation (1) gives

$$x_{i+1} = x_i - \frac{f(x_i)(x_i - x_{i-1})}{f(x_i) - f(x_{i-1})} \tag{3}$$

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

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

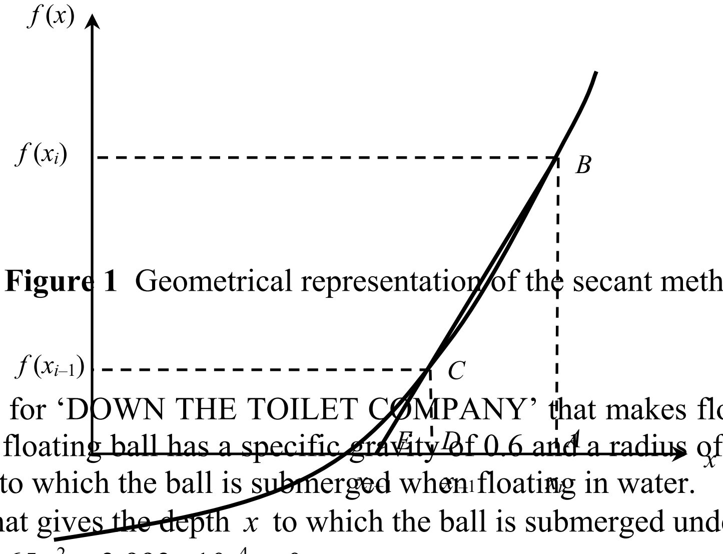


Figure 1 Geometrical representation of the secant method.

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

$$x^3 - 0.165x^2 + 3.993\times10^{-4} = 0$$

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

$$f(x) = x^3 - 0.165x^2 + 3.993\times10^{-4}$$

Let us assume the initial guesses of the root of $f(x) = 0$ as $x_{-1} = 0.02$ and $x_0 = 0.05$.

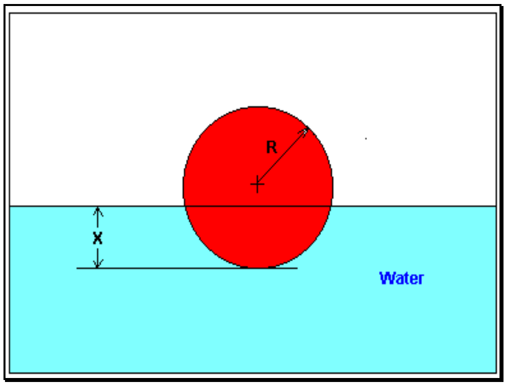


Figure 2 Floating ball problem.

Iteration 1

The estimate of the root is

$$x_1 = x_0 - \frac{f(x_0)(x_0 - x_{-1})}{f(x_0) - f(x_{-1})}$$

$$= x_0 - \frac{\left(x_0^3 - 0.165x_0^2 + 3.993\times10^{-4}\right)\times\left(x_0 - x_{-1}\right)}{\left(x_0^3 - 0.165x_0^2 + 3.993\times10^{-4}\right) - \left(x_{-1}^3 - 0.165x_{-1}^2 + 3.993\times10^{-4}\right)}$$

$$= 0.05 - \frac{\left[0.05^3 - 0.165(0.05)^2 + 3.993\times10^{-4}\right]\times\left[0.05 - 0.02\right]}{\left[0.05^3 - 0.165(0.05)^2 + 3.993\times10^{-4}\right] - \left[0.02^3 - 0.165(0.02)^2 + 3.993\times10^{-4}\right]}$$

$$= 0.06461$$

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

$$|\in_a| = \left|\frac{x_1 - x_0}{x_1}\right|\times100$$

$$= \left|\frac{0.06461 - 0.05}{0.06461}\right|\times100$$

$$= 22.62\%$$

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

$$x_2 = x_1 - \frac{f(x_1)(x_1 - x_0)}{f(x_1) - f(x_0)}$$

$$= x_1 - \frac{\left(x_1^3 - 0.165x_1^2 + 3.993\times10^{-4}\right)\times\left(x_1 - x_0\right)}{\left(x_1^3 - 0.165x_1^2 + 3.993\times10^{-4}\right) - \left(x_0^3 - 0.165x_0^2 + 3.993\times10^{-4}\right)}$$

$$= 0.06461 - \frac{\left[0.06461^3 - 0.165(0.06461)^2 + 3.993\times10^{-4}\right]\times\left(0.06461 - 0.05\right)}{\left[0.06461^3 - 0.165(0.06461)^2 + 3.993\times10^{-4}\right] - \left[0.05^3 - 0.165(0.05)^2 + 3.993\times10^{-4}\right]}$$

$$= 0.06241$$

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

$$\begin{aligned}\left|\in_a\right| &= \left|\frac{x_2 - x_1}{x_2}\right| \times 100 \\ &= \left|\frac{0.06241 - 0.06461}{0.06241}\right| \times 100 \\ &= 3.525\%\end{aligned}$$

The number of significant digits at least correct is 1, as you need an absolute relative approximate error of 5% or less.

Iteration 3

$$\begin{aligned}x_3 &= x_2 - \frac{f(x_2)(x_2 - x_1)}{f(x_2) - f(x_1)} \\ &= x_2 - \frac{\left(x_2^3 - 0.165x_2^2 + 3.993\times10^{-4}\right)\times\left(x_2 - x_1\right)}{\left(x_2^3 - 0.165x_2^2 + 3.993\times10^{-4}\right) - \left(x_1^3 - 0.165x_1^2 + 3.993\times10^{-4}\right)}\end{aligned}$$

$$= 0.06241 - \frac{\left[0.06241^3 - 0.165(0.06241)^2 + 3.993\times10^{-4}\right]\times\left(0.06241 - 0.06461\right)}{\left[0.06241^3 - 0.165(0.06241)^2 + 3.993\times10^{-4}\right] - \left[0.06461^3 - 0.165(0.06461)^2 + 3.993\times10^{-4}\right]}$$

$$= 0.06238$$

The absolute relative approximate error $\left|\in_a\right|$ at the end of Iteration 3 is

$$\begin{aligned}\left|\in_a\right| &= \left|\frac{x_3 - x_2}{x_3}\right| \times 100 \\ &= \left|\frac{0.06238 - 0.06241}{0.06238}\right| \times 100 \\ &= 0.0595\%\end{aligned}$$

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}	$\left\|\in_a\right\|\%$	$f\left(x_{i+1}\right)$
1	0.02	0.05	0.06461	22.62	-1.9812×10^{-5}
2	0.05	0.06461	0.06241	3.525	-3.2852×10^{-7}
3	0.06461	0.06241	0.06238	0.0595	2.0252×10^{-9}
4	0.06241	0.06238	0.06238	-3.64×10^{-4}	-1.8576×10^{-13}

Example 2 (Computer Science)

To find the inverse of a number a, one can use the equation

$$f(c) = a - \frac{1}{c} = 0$$

where c is the inverse of a.
Use the secant method of finding roots of equations to find the inverse of $a = 2.5$. Conduct three iterations to estimate the root of the above equation. Find the absolute relative approximate error at the end of each iteration and the number of significant digits at least correct at the end of each iteration.

Solution

$$f(c) = a - \frac{1}{c} = 0$$

$$\begin{aligned}
c_{i+1} &= c_i - \frac{\left(a - \frac{1}{c_i}\right)(c_i - c_{i-1})}{\left(a - \frac{1}{c_i}\right) - \left(a - \frac{1}{c_{i-1}}\right)} \\
&= c_i - \frac{\left(a - \frac{1}{c_i}\right)(c_i - c_{i-1})}{\frac{1}{c_{i-1}} - \frac{1}{c_i}} \\
&= c_i - \frac{\left(a - \frac{1}{c_i}\right)(c_i - c_{i-1})}{\frac{(c_i - c_{i-1})}{c_i c_{i-1}}} \\
&= c_i - c_i c_{i-1}\left(a - \frac{1}{c_i}\right) \\
&= c_i - c_{i-1}(ac_i - 1)
\end{aligned}$$

Let us take the initial guesses of the root of $f(c) = 0$ as $c_{-1} = 0.1$ and $c_0 = 0.6$.

Iteration 1
The estimate of the root is

$$\begin{aligned}
c_1 &= c_0 - c_{-1}(ac_0 - 1) \\
&= 0.6 - (0.1)(2.5(0.6) - 1) \\
&= 0.55
\end{aligned}$$

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

$$\begin{aligned}
|\in_a| &= \left|\frac{c_1 - c_0}{c_1}\right| \times 100 \\
&= \left|\frac{0.55 - 0.6}{0.55}\right| \times 100 \\
&= 9.0909\%
\end{aligned}$$

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

Iteration 2

The estimate of the root is

$$\begin{aligned} c_2 &= c_1 - c_0(ac_1 - 1) \\ &= 0.55 - (0.6)\big(2.5(0.55) - 1\big) \\ &= 0.325 \end{aligned}$$

The absolute relative approximate error $|\in_a|$ at the end of Iteration 2 is

$$\begin{aligned} |\in_a| &= \left|\frac{c_2 - c_1}{c_2}\right| \times 100 \\ &= \left|\frac{0.325 - 0.55}{0.325}\right| \times 100 \\ &= 69.231\% \end{aligned}$$

The number of significant digits at least correct is 0.

Iteration 3

The estimate of the root is

$$\begin{aligned} c_3 &= c_2 - c_1(ac_2 - 1) \\ &= 0.325 - (0.55)\big(2.5(0.325) - 1\big) \\ &= 0.42813 \end{aligned}$$

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

$$\begin{aligned} |\in_a| &= \left|\frac{c_3 - c_2}{c_3}\right| \times 100 \\ &= \left|\frac{0.42813 - 0.325}{0.42813}\right| \times 100 \\ &= 24.088\% \end{aligned}$$

The number of significant digits at least correct is 0.

Example 3 (Civil Engineering)

You are making a bookshelf to carry books that range from 8½" (21.59cm) to 11" (27.94cm) in height and would take up 29" (73.66cm) of space along the length. The material is wood having a Young's Modulus of 3.667Msi, thickness of 3/8" (0.9525cm) and width of 12" (30.48cm). You want to find the maximum vertical deflection of the bookshelf. The vertical deflection of the shelf is given by

$$v(x) = 0.42493\times10^{-4}x^3 - 0.13533\times10^{-8}x^5 - 0.66722\times10^{-6}x^4 - 0.018507x$$

where x is the position along the length of the beam. Hence to find the maximum deflection we need to find where $f(x) = \frac{dv}{dx} = 0$ and conduct the second derivative test.

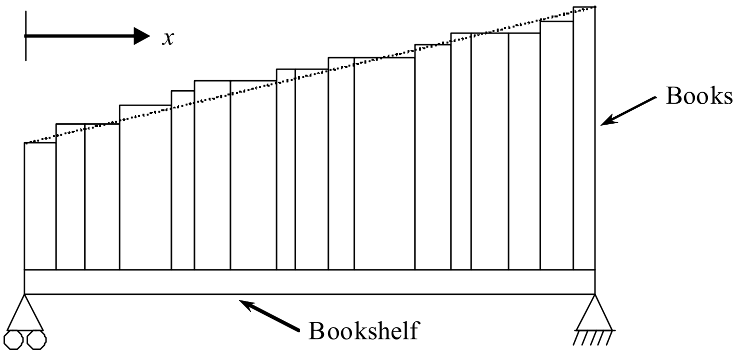


Figure 1 A loaded bookshelf.

The equation that gives the position x where the deflection is maximum is given by

$$-0.67665\times10^{-8}x^4 - 0.26689\times10^{-5}x^3 + 0.12748\times10^{-3}x^2 - 0.018507 = 0$$

Use the secant method of finding roots of equations to find the position x where the deflection is maximum. 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
Let us take the initial guesses of the root of $f(x) = 0$ as $x_{-1} = 10$ and $x_0 = 15.$

Iteration 1
The estimate of the root is

$$x_1 = x_0 - \frac{f(x_0)(x_0 - x_{-1})}{f(x_0) - f(x_{-1})}$$

$$\begin{aligned} f(x_0) &= -0.67665\times10^{-8}x_0^4 - 0.26689\times10^{-5}x_0^3 + 0.12748\times10^{-3}x_0^2 - 0.018507 \\ &= -0.67665\times10^{-8}(15)^4 - 0.26689\times10^{-5}(15)^3 + 0.12748\times10^{-3}(15)^2 - 0.018507 \\ &= 8.2591\times10^{-4} \end{aligned}$$

$$\begin{aligned} f(x_{-1}) &= -0.67665\times10^{-8}x_{-1}^4 - 0.26689\times10^{-5}x_{-1}^3 + 0.12748\times10^{-3}x_{-1}^2 - 0.018507 \\ &= -0.67665\times10^{-8}(10)^4 - 0.26689\times10^{-5}(10)^3 + 0.12748\times10^{-3}(10)^2 - 0.018507 \\ &= -8.4956\times10^{-3} \end{aligned}$$

$$\begin{aligned} x_1 &= 15 - \frac{\left(8.2591\times10^{-4}\right)\times(15-10)}{\left(8.2591\times10^{-4}\right) - \left(-8.4956\times10^{-3}\right)} \\ &= 14.557 \end{aligned}$$

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

$$
\begin{aligned}
\left|\in_a\right| &= \left|\frac{x_1 - x_0}{x_1}\right| \times 100 \\
&= \left|\frac{14.557 - 15}{14.557}\right| \times 100 \\
&= 3.0433\%
\end{aligned}
$$

The number of significant digits at least correct is 1, because the absolute relative approximate error is less than 5%.

Iteration 2

The estimate of the root is

$$
x_2 = x_1 - \frac{f(x_1)(x_1 - x_0)}{f(x_1) - f(x_0)}
$$

$$
\begin{aligned}
f(x_1) &= -0.67665 \times 10^{-8} x_1^4 - 0.26689 \times 10^{-5} x_1^3 + 0.12748 \times 10^{-3} x_1^2 - 0.018507 \\
&= -0.67665 \times 10^{-8} (14.557)^4 - 0.26689 \times 10^{-5} (14.557)^3 \\
&\quad + 0.12748 \times 10^{-3} (14.557)^2 - 0.018507 \\
&= -2.9870 \times 10^{-5}
\end{aligned}
$$

$$
\begin{aligned}
x_2 &= 15 - \frac{\left(-2.9870 \times 10^{-5}\right) \times (14.557 - 15)}{\left(-2.9870 \times 10^{-5}\right) - \left(8.2591 \times 10^{-4}\right)} \\
&= 14.572
\end{aligned}
$$

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

$$
\begin{aligned}
\left|\in_a\right| &= \left|\frac{x_2 - x_1}{x_2}\right| \times 100 \\
&= \left|\frac{14.572 - 14.557}{14.572}\right| \times 100 \\
&= 0.10611\%
\end{aligned}
$$

The number of significant digits at least correct is 2, because the absolute relative approximate error is less than 0.5%.

Iteration 3

The estimate of the root is

$$
x_3 = x_2 - \frac{f(x_2)(x_2 - x_1)}{f(x_2) - f(x_1)}
$$

$$
\begin{aligned}
f(x_2) &= -0.67665 \times 10^{-8} x_2^4 - 0.26689 \times 10^{-5} x_2^3 + 0.12748 \times 10^{-3} x_2^2 - 0.018507 \\
&= -0.67665 \times 10^{-8} (14.572)^4 - 0.26689 \times 10^{-5} (14.572)^3 \\
&\quad + 0.12748 \times 10^{-3} (14.572)^2 - 0.018507 \\
&= -6.0676 \times 10^{-9}
\end{aligned}
$$

$$
\begin{aligned}
x_3 &= 14.572 - \frac{\left(-6.0676 \times 10^{-9}\right) \times (14.572 - 14.557)}{\left(-6.0676 \times 10^{-9}\right) - \left(-2.9870 \times 10^{-5}\right)} \\
&= 14.572
\end{aligned}
$$

The absolute relative approximate error $\left|\in_a\right|$ at the end of Iteration 3 is

$$\begin{aligned}\left|\in_a\right| &= \left|\frac{x_3 - x_2}{x_3}\right| \times 100 \\ &= \left|\frac{14.572 - 14.572}{14.572}\right| \times 100 \\ &= 2.1559 \times 10^{-5}\%\end{aligned}$$

The number of significant digits at least correct is 6, because the absolute relative approximate error is less than 0.00005%.

Exercise . The average energy of vibration E of a molecule with frequency f depends on the temperature T according to the equation
$E =\ hf/(e^{hf/kT} -1) + hf$
Here, h = 6.626×10^{-27} erg/sec. is Planck's constant and k=1.38×10^{-16} erg/°K is Boltzmann's constant. Find the frequency f of a molecule for which E = 3.97×10^{-14} erg and T = 31O°K.

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	Aralık 8, 2016
Web Site	http://numericalmethods.eng.usf.edu

3.5.1 Multiple-Choice Test Secant Method Chapter 03.05

1. The secant method of finding roots of nonlinear equations falls under the category of ______________ methods.
   (A) bracketing (B) graphical (C) open (D) random

2. The secant method formula for finding the square root of a real number R from the equation $x^2 - R = 0$ is
   (A) $\frac{x_i x_{i-1} + R}{x_i + x_{i-1}}$ (B) $\frac{x_i x_{i-1}}{x_i + x_{i-1}}$ (C) $\frac{1}{2}\left(x_i + \frac{R}{x_i}\right)$ (D) $\frac{2x_i^2 + x_i x_{i-1} - R}{x_i + x_{i-1}}$

3. The next iterative value of the root of $x^2 - 4 = 0$ using secant method, if the initial guesses are 3 and 4, is
   (A) 2.2857 (B) 2.5000 (C) 5.5000 (D) 5.7143

4. The root of the equation $f(x)=0$ is found by using the secant method. Given one of the initial estimates is $x_0=3$, $f(3)=5$, and the angle the secant line makes with the x-axis is $57°$, the next estimate of the root, x_1, is

(A) –3.2470 (B) –0.24704 (C) 3.247 (D) 6.2470

5. For finding the root of $\sin x=0$ by the secant method, the following choice of initial guesses would not be appropriate.

(A) $\frac{\pi}{4}$ and $\frac{\pi}{2}$ (B) $\frac{\pi}{4}$ and $\frac{3\pi}{4}$ (C) $-\frac{\pi}{2}$ and $\frac{\pi}{2}$ (D) $\frac{\pi}{3}$ and $\frac{\pi}{2}$

6. When drugs are given orally to a patient, the drug concentration c in the blood stream at time t is given by a formula

$$c = Kte^{-at}$$

where K is dependent on parameters such as the dose administered while a is dependent on the absorption and elimination rates of the drug. If $K=2$ and $a=0.25$, and t is in seconds and c is in mg/ml, the time at which the maximum concentration is reached is given by the solution of the equation

(A) $2te^{-0.25t}=0$ (B) $2e^{-0.25t}-2te^{-0.25t}=0$ (C) $2e^{-0.25t}-0.5te^{-0.25t}=0$ (D) $2te^{-0.25t}=2$

For a complete solution, refer to the links at the end of the book.

3.6 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 (bölüm 03.03'te basit aralık yöntemlerinden ikiye bölme yöntemi ile genel formu aşağıda verilen çizgisel olmayan fonksiyonların köklerinin hesaplanması gösterilmişti.)

$$f(x) = 0 \tag{1}$$

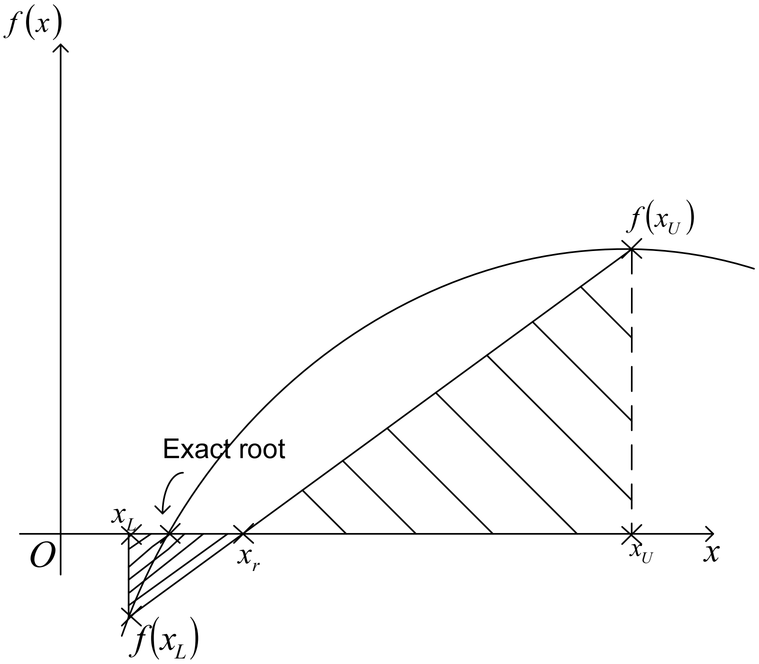


Figure 1 False-Position Method

The above nonlinear equation can be stated as finding the value of x such that Equation (1) is satisfied (yukarıdaki çizgisel olmayan eşitlikte x değerinin bulunması yeterli olacaktır).

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 (ikiye bölme yönteminde xL (alt değer) ve xU (üst değer) değerleri ile aralık tanımlanarak aşağıdaki durumun sağlanması yeterli olacaktır)

$$f(x_L)\,f(x_U) < 0. \tag{2}$$

The next predicted/improved root x_r can be computed as the midpoint between x_L and x_U as (aralık belirlendikten sonra xr kök değeri xL ile xU'nun tam ortasındaki noktadır)

$$x_r = \frac{x_L + x_U}{2} \tag{3}$$

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). (bu aşamadan sonra alt ve üst sınır değerleri yeniden tanımlanarak kök değerine yaklaşılır).

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. (Şekil 1'den görüleceği gibi ikiye bölme yönteminde f(xL) fonksiyonunun değeri sıfıra çok yakınsa f(x) fonksiyonunun değeri f(xU)'ya doğru yaklaşmasından dolayı bu yöntem doğru kökü bulmada yeterli olmayabilir. Başka bir deyişle xL ile xU arasında olması gereken xr kök değeri xL'ye doğru yaklaşacaktır. False-position yöntemi bu aşamada xL'deki fonksiyon değerinden xU'daki fonksiyon değerine yatay ekseni kesen bir kiriş çizerek daha avantajlı olmaktadır.)

False-Position Method

Based on two similar triangles, shown in Figure 1, one gets

$$\frac{0-f(x_L)}{x_r-x_L}=\frac{0-f(x_U)}{x_r-x_U} \tag{4}$$

From Equation (4), one obtains

$$(x_r-x_L)f(x_U)=(x_r-x_U)f(x_L)$$

$$x_U f(x_L)-x_L f(x_U)=x_r\{f(x_L)-f(x_U)\}$$

The above equation can be solved to obtain the next predicted root x_m as

$$x_r=\frac{x_U f(x_L)-x_L f(x_U)}{f(x_L)-f(x_U)} \tag{5}$$

The above equation, through simple algebraic manipulations, can also be expressed as

$$x_r=x_U-\frac{f(x_U)}{\left\{\dfrac{f(x_L)-f(x_U)}{x_L-x_U}\right\}} \tag{6}$$

or

$$x_r=x_L-\frac{f(x_L)}{\left\{\dfrac{f(x_U)-f(x_L)}{x_U-x_L}\right\}} \tag{7}$$

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

$$|\in_a| = \left|\frac{x_r^{new} - x_r^{old}}{x_r^{new}}\right| \times 100$$

where

x_r^{new} = estimated root from present iteration

x_r^{old} = estimated root from previous iteration

5. Compare the absolute relative approximate error $|\in_a|$ with the pre-specified relative error tolerance $\in_s$. If $|\in_a| > \in_s$, 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.

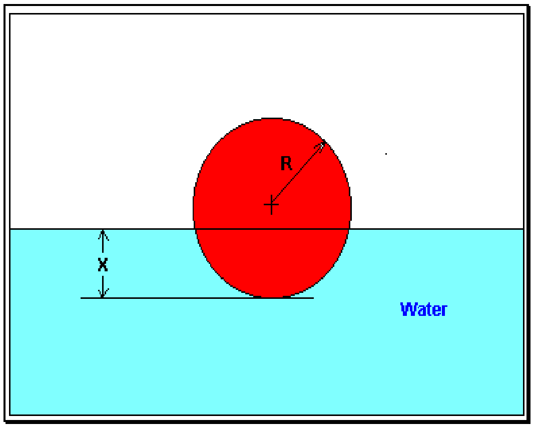


Figure 2 Floating ball problem.

Solution

From the physics of the problem, the ball would be submerged between $x = 0$ and $x = 2R$, where

$$R = \text{radius of the ball,}$$

that is

$$0 \le x \le 2R$$
$$0 \le x \le 2(0.055)$$
$$0 \le x \le 0.11$$

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

$$f(x_L)f(x_U) = f(0)f(0.11) = \left(3.993 \times 10^{-4}\right)\left(-2.662 \times 10^{-4}\right) < 0$$

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

$$x_r = \frac{x_U f(x_L) - x_L f(x_U)}{f(x_L) - f(x_U)}$$
$$= \frac{0.11 \times 3.993 \times 10^{-4} - 0 \times \left(-2.662 \times 10^{-4}\right)}{3.993 \times 10^{-4} - \left(-2.662 \times 10^{-4}\right)}$$
$$= 0.0660$$

$$f(x_r) = f(0.0660)$$
$$= (0.0660)^3 - 0.165(0.0660)^2 + \left(3.993 \times 10^{-4}\right)$$
$$= -3.1944 \times 10^{-5}$$

$$f(x_L)f(x_r) = f(0)f(0.0660) = (+)(-) < 0$$

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

$$
\begin{aligned}
x_r &= \frac{x_U f(x_L) - x_L f(x_U)}{f(x_L) - f(x_U)} \\
&= \frac{0.0660 \times 3.993 \times 10^{-4} - 0 \times \left(-3.1944 \times 10^{-5}\right)}{3.993 \times 10^{-4} - \left(-3.1944 \times 10^{-5}\right)} \\
&= 0.0611
\end{aligned}
$$

The absolute relative approximate error for this iteration is

$$
\in_a = \left| \frac{0.0611 - 0.0660}{0.0611} \right| \times 100 \cong 8\%
$$

$$
\begin{aligned}
f(x_r) &= f(0.0611) \\
&= (0.0611)^3 - 0.165(0.0611)^2 + \left(3.993 \times 10^{-4}\right) \\
&= 1.1320 \times 10^{-5}
\end{aligned}
$$

$$
f(x_L) f(x_r) = f(0) f(0.0611) = (+)(+) > 0
$$

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

$$
\begin{aligned}
x_r &= \frac{x_U f(x_L) - x_L f(x_U)}{f(x_L) - f(x_U)} \\
&= \frac{0.0660 \times 1.132 \times 10^{-5} - 0.0611 \times \left(-3.1944 \times 10^{-5}\right)}{1.132 \times 10^{-5} - \left(-3.1944 \times 10^{-5}\right)} \\
&= 0.0624
\end{aligned}
$$

The absolute relative approximate error for this iteration is

$$
\in_a = \left| \frac{0.0624 - 0.0611}{0.0624} \right| \times 100 \cong 2.05\%
$$

$$
f(x_r) = -1.1313 \times 10^{-7}
$$

$$
f(x_L) f(x_r) = f(0.0611) f(0.0624) = (+)(-) < 0
$$

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

$$
\left| \in_a \right| \le 0.5 \times 10^{2-m}
$$

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

$$
m \le 1.387
$$

The number of significant digits at least correct in the estimated root of 0.0624 at the end of 3rd 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	$\lvert \in_a \rvert \%$	$f(x_m)$
1	0.0000	0.1100	0.0660	----	-3.1944×10^{-5}
2	0.0000	0.0660	0.0611	8.00	-1.1320×10^{-5}
3	0.0611	0.0660	0.0624	2.05	-1.1313×10^{-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 $\in_s = 0.1\%$.

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	$\lvert \in_a \rvert \%$	$f(x_m)$
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,

$$\lvert \in_a \rvert \leq 0.5 \times 10^{2-m}$$

$$0.02706 \leq 0.5 \times 10^{2-m}$$

$$m \leq 3.2666$$

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	Aralık 8, 2016

3.6.1 Multiple-Choice Test Chapter 03.06 False-Position Method of Solving a Nonlinear Equation

1. The false-position method for finding roots of nonlinear equations belongs to a class of a (an) ____________ method.

(A) open (B) bracketing (C) random (D) graphical

2. The newly predicted root for false-position and secant method can be respectively given as (tahmini kök değerleri false-position ve kiriş yöntemleri sırasıyla aşağıdaki gibi verilmektedir)

$$x_r = x_U - \frac{f(x_U)\{x_U - x_L\}}{f(x_U) - f(x_L)}$$

and

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

While the appearance of the above 2 equations look essentially identical, and both methods require two initial guesses, the major difference between the above two formulas is

(A) false-position method is not guaranteed to converge.

(B) secant method is guaranteed to converge

(C) secant method requires the 2 initial guesses x_{i-1} *and* x_i to satisfy $f(x_{i-1}) \times f(x_i) < 0$

(D) false-position method requires the 2 initial guesses x_L *and* x_U to satisfy $f(x_L) \times f(x_U) < 0$

3. Given are the following nonlinear equation

$$e^{-2x} + 4x^2 - 36 = 0$$

two initial guesses, $x_L = 1$ and $x_U = 4$, and a pre-specified relative error tolerance of 0.1%. Using the false-position method, which of the following tables is correct (x_r = predicted root)?

(A)

Iteration	x_L	x_U	x_r
1	1	4	?
2	?	?	2.939

(B)

Iteration	x_L	x_U	x_r
1	1	4	?
2	?	?	2.500

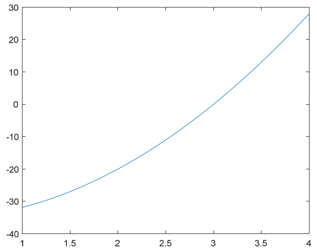
(C)

Iteration	x_L	x_U	x_r
1	1	4	?
2	?	?	1.500

(D)

Iteration	x_L	x_U	x_r
1	1	4	?
2	?	?	2.784

```
clear all
xL=1.0; % baslangic tahmin degerleri
xU=4.0;
i=1;
eA=100.0
xr_old=xU;
fxL=exp(-2.0*xL)+4.0*xL*xL-36.0;
fxU=exp(-2.0*xU)+4.0*xU*xU-36.0;
xr=(xU*fxL-xL*fxU)/(fxL-fxU);
xr_new=xr;
fxr=exp(-2.0*xr)+4.0*xr*xr-36.0;
fprintf(' Yineleme fxL    fxU    Tahmini      Hata   \n');
fprintf(' Sayisi                 Kök degeri   Degeri \n');
fprintf('--------------------------------------------\n');
if(fxL*fxr<0)
    xU=xr;
else
    xL=xr;
end;
eA=abs((xr_new-xr_old)*100/xr_new);
fprintf(' %3d     %5.3f  %5.3f   %5.3f      %5.3f      \n', i, fxL, fxU, xr, eA);
while (i<7)
  i=i+1;
  xr_old=xr;
  fxL=exp(-2.0*xL)+4.0*xL*xL-36.0;
  fxU=exp(-2.0*xU)+4.0*xU*xU-36.0;
  xr=(xU*fxL-xL*fxU)/(fxL-fxU);
  fxr=exp(-2.0*xr)+4.0*xr*xr-36.0;
  if(fxL*fxr<0)
      xU=xr;
  else
      xL=xr;
  end;
  xr_new=xr;
  eA=abs((xr_new-xr_old)*100/xr_new);
fprintf(' %3d     %5.3f  %5.3f   %5.3f      %5.3f      \n', i, fxL, fxU, xr, eA);
end;
x = 1.0 : 0.2 : 4.0;
y = exp(-2.0*x)+4.0*x.*x-36.0;
plot(x, y)
```



yukarıdaki MatLab kodunun çalıştırılması sonucunda aşağıdaki veriler elde edilmiştir:

```
Yineleme fxL    fxU    Tahmini      Hata
 Sayisi                 Kök degeri   Degeri
--------------------------------------------
   1     -31.865  28.000   2.597      54.034
   2     -9.020  28.000   2.939      11.634
   3     -1.453  28.000   2.991      1.750
   4     -0.211  28.000   2.999      0.252
   5     -0.030  28.000   3.000      0.036
   6     -0.004  28.000   3.000      0.005
   7     -0.001  28.000   3.000      0.001
```

4. Given are the following nonlinear equation

$$e^{-2x} + 4x^2 - 36 = 0$$

two initial guesses, $x_L = 1$ and $x_U = 4$, and a pre-specified relative error tolerance of 0.1%. Using the false-position method, which of the following tables is correct ($x_r =$ predicted root, $|\in_a| =$ percentage absolute relative approximate error).

(A)

Iteration	x_L	x_U	x_r	$\lvert\in_a\rvert$ %
1	1	4	?	?
2	?	?	?	11.63

(B)

Iteration	x_L	x_U	x_r	$\lvert\in_a\rvert$ %
1	1	4	?	?
2	?	?	?	6.11

(C)

Iteration	x_L	x_U	x_r	$\lvert\in_a\rvert$ %
1	1	4	?	?
2	?	?	?	5.14

(D)

Iteration	x_L	x_U	x_r	$\lvert\in_a\rvert$ %
1	1	4	?	?
2	?	?	?	4.15

5. The root of $(x-4)^2(x+2)=0$ was found using false-position method with initial guesses of $x_L = -2.5$ and $x_U = -1.0$, and a pre-specified relative error tolerance of 10^{-6}%. The final converged root was found as $x_r == -1.9999997$, and the corresponding percentage absolute relative approximate error was found as $|\in_a| = 8.7610979 \times 10^{-5}\%$. Based on the given information, the number of significant digits of the converged root x_r that can be trusted at least are

(A) 3 (B) 4 (C) 5 (D) 6

6. The false-position method may have difficulty in finding the root of $f(x) = x^2 - 7.4x + 13.69 = 0$ because

(A) $f(x)$ is a quadratic polynomial

(B) $f'(x)$ a straight line

(C) one cannot find initial guesses x_L and x_U that satisfy $f(x_L)\,f(x_U) < 0$

(D) the equation has two identical roots.

4 Chapter 04.01 Introduction to Matrix Algebra

PRE-REQUISITES

1. This is a primer. So all you need to know is high school algebra. If you had exposure to matrices, you may already know many of the concepts presented here.

OBJECTIVES

1. define what a matrix is.
2. identify special types of matrices, and
3. identify when two matrices are equal.
4. add, subtract, and multiply matrices, and
5. apply rules of binary operations on matrices.
6. know what unary operations means,
7. find the transpose of a square matrix and it’s relationship to symmetric matrices,
8. setup simultaneous linear equations in matrix form and vice-versa,
9. understand the concept of the inverse of a matrix.

4.1 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 (Matris nasıl bir şeydir)?

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 Blowoutr’us store – given by quarter and make of tires (Matrisler her yerde karşımıza çıkar. Excel çalışma sayfası açtıysanız veya sayılardan çizelge oluşturmuşsanız matris kullanıyorsunuzdur. Matrisler sayıların daha yalın görünmesini ve programın hesaplamaları kolayca yapmasını sağlar. Aşağıdaki çizelgeye/matise bakarsanız Blowout r’us mağazalarında yılın çeyreklerinde lastik ürünlerin satış miktarları verilmektedir) .

$$\begin{array}{l c}
 & \begin{array}{cccc} \text{Q1} & \text{Q2} & \text{Q3} & \text{Q4} \end{array} \\
\begin{array}{l} \textbf{Tirestone} \\ \textbf{Michigan} \\ \textbf{Copper} \end{array} & \begin{bmatrix} 25 & 20 & 3 & 2 \\ 5 & 10 & 15 & 25 \\ 6 & 16 & 7 & 27 \end{bmatrix}
\end{array}$$

If one wants to know how many *Copper* tires were sold in *Quarter* 4, we go along the row *Copper* and column *Q*4 and find that it is 27 (Copper marka lastiklerin yılın 4.cü çeyreğindeki satış adetini öğrenmek istersek matriste önce Copper markasının olduğu satıra gelip daha sonra Q4 yani dördüncü çeyreğin hizasındaki rakama bakılır).

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

$$[A] = \begin{bmatrix} a_{11} & a_{12} & .... & a_{1n} \\ a_{21} & a_{22} & .... & a_{2n} \\ \vdots & & & \vdots \\ a_{m1} & a_{m2} & .... & a_{mn} \end{bmatrix}$$

Row i of $[A]$ has n elements and is

$$\begin{bmatrix} a_{i1} & a_{i2} ....a_{in} \end{bmatrix}$$

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 3×4. 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 $[B] = [b_1 \; b_2 \ldots\ldots b_n]$ 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

$$[C]=\begin{bmatrix} c_1 \\ \vdots \\ \vdots \\ c_m \end{bmatrix}$$

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

$$\begin{bmatrix} 4 & 6 & 2 \\ 3 & -1 & 2 \end{bmatrix}, \begin{bmatrix} 4 & 6 \\ 3 & -1 \end{bmatrix}, \begin{bmatrix} 4 & 6 & 2 \end{bmatrix}, \begin{bmatrix} 4 \end{bmatrix}, \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

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 $a_{11}, a_{22}, \ldots, a_{nn}$ 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

$$[A]=\begin{bmatrix}0 & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}$$

$$[\mathrm{B}]=\begin{bmatrix}0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}$$

$$[C]=\begin{bmatrix}0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}$$

$$[D]=\begin{bmatrix}0 & 0 & 0\end{bmatrix}$$

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 $a_{11}, a_{22} ..., a_{k-1,k-1}, a_{kk}$ 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

$$\left|a_{ii}\right| \geq \sum_{\substack{j=1 \\ i \neq j}}^{n} | a_{ij} | \text{ for } i = 1,2,.....,n \text{ and}$$

$$\left|a_{ii}\right| > \sum_{\substack{j=1 \\ i \neq j}}^{n} | a_{ij} | \text{ for at least one } i,$$

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 $[B]$ ould 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

4.2 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, $a_1, a_2, ..., a_n$. The above is a column vector. A row vector $[B]$ is of the form $\vec{B} = [b_1, b_2, ..., b_n]$ where $\vec{B}$ is a n-dimensional row vector with n components $b_1, b_2, ..., b_n$.

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 $a_i = b_i$, $i = 1,2, ..., n$.

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: Tirestone, 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 – Tirestone, 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

$$\vec{C}=\vec{A}_1+\vec{A}_2$$

$$=\begin{bmatrix}25\\5\\6\end{bmatrix}+\begin{bmatrix}20\\10\\6\end{bmatrix}$$

$$=\begin{bmatrix}25+20\\5+10\\6+6\end{bmatrix}$$

$$=\begin{bmatrix}45\\15\\12\end{bmatrix}$$

So the number of Tirestone 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

$$\sqrt{u_1^2+u_2^2+u_3^2+\ldots+u_n^2}=1$$

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

$$k\vec{A} = k\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$$

$$= \begin{bmatrix} ka_1 \\ ka_2 \\ \vdots \\ ka_n \end{bmatrix}$$

Example 7

What is $2\vec{A}$ if

$$\vec{A} = \begin{bmatrix} 25 \\ 20 \\ 5 \end{bmatrix}$$

Solution

$$2\vec{A} = 2\begin{bmatrix} 25 \\ 20 \\ 5 \end{bmatrix}$$

$$= \begin{bmatrix} 2\times 25 \\ 2\times 20 \\ 2\times 5 \end{bmatrix}$$

$$= \begin{bmatrix} 50 \\ 40 \\ 10 \end{bmatrix}$$

Example 8

A store sells three brands of tires: Tirestone, Michigan 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,

$$\vec{B} = 1.25\begin{bmatrix} 25 \\ 25 \\ 6 \end{bmatrix}$$

$$= \begin{bmatrix} 31.25 \\ 31.25 \\ 7.5 \end{bmatrix}$$

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

$$\vec{A}_1, \vec{A}_2, ..., \vec{A}_m$$

as *m* vectors of same dimension *n*, and if $k_1, k_2, ..., k_m$ are scalars, then

$$k_1\vec{A}_1 + k_2\vec{A}_2 + ... + k_m\vec{A}_m$$

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

a) $\vec{A} - \vec{B} = \begin{bmatrix} 2 \\ 3 \\ 6 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}$

$$= \begin{bmatrix} 2-1 \\ 3-1 \\ 6-2 \end{bmatrix}$$

$$=\begin{bmatrix}1\\2\\4\end{bmatrix}$$

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 $\vec{A}_1,\vec{A}_2,\ldots,\vec{A}_m$ are considered to be linearly independent if

$$k_1\vec{A}_1+k_2\vec{A}_2+.......+k_m\vec{A}_m=\vec{0}$$

has only one solution of

$$k_1=k_2=......=k_m=0$$

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

$$25k_1 + 5k_2 + k_3 = 0 \tag{1}$$

$$64k_1 + 8k_2 + k_3 = 0 \tag{2}$$

$$144k_1 + 12k_2 + k_3 = 0 \tag{3}$$

Subtracting Eqn (1) from Eqn (2) gives

$$39k_1 + 3k_2 = 0$$

$$k_2 = -13k_1 \tag{4}$$

Multiplying Eqn (1) by 8 and subtracting it from Eqn (2) that is first multiplied by 5 gives

$$120k_1 - 3k_3 = 0$$

$$k_3 = 40k_1 \tag{5}$$

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

$$144k_1 + 12(-13k_1) + 40k_1 = 0$$

$$28k_1 = 0$$

$$k_1 = 0$$

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

$$\vec{A}_1 = \begin{bmatrix} 1 \\ 2 \\ 5 \end{bmatrix}, \vec{A}_2 = \begin{bmatrix} 2 \\ 5 \\ 7 \end{bmatrix}, A_3 = \begin{bmatrix} 6 \\ 14 \\ 24 \end{bmatrix}$$

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

$$k_1+2k_2+6k_3=0 \qquad (1)$$

$$2k_1+5k_2+14k_3=0 \qquad (2)$$

$$5k_1+7k_2+24k_3=0 \qquad (3)$$

Multiplying Eqn (1) by 2 and subtracting from Eqn (2) gives

$$k_2+2k_3=0$$

$$k_2=-2k_3 \qquad (4)$$

Multiplying Eqn (1) by 2.5 and subtracting from Eqn (2) gives

$$-0.5k_1-k_3=0$$

$$k_1=-2k_3 \qquad (5)$$

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

$$5(-2k_3)+7(-2k_3)+24k_3=0$$

$$-10k_3-14k_3+24k_3=0$$

$$0=0$$

This means any values satisfying Eqns (4) and (5) will satisfy Eqns (1), (2) and (3) simultaneously.

For example, chose

$k_3=6$, then

$k_2=-12$ from Eqn (4), and

$k_1=-12$ from Eqn (5).

Hence we have a nontrivial solution of $[k_1 \quad k_2 \quad k_3]=[-12 \quad -12 \quad 6]$. 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 $\vec{A}_1, \vec{A}_2, \ldots\ldots\ldots, \vec{A}_m$ be a set of n-dimensional vectors, then

$$k_1\vec{A}_1 + k_2\vec{A}_2 + \ldots + k_m\vec{A}_m = \vec{0}$$

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 $k_2 = k_3 = \ldots = k_m = 0$ 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

$$\vec{A}_{a1}, \vec{A}_{a2}, \ldots, \vec{A}_{ap}$$

where $p < m$.

Then if this subset of vectors is linearly dependent, the linear combination

$$k_1\vec{A}_{a1} + k_2\vec{A}_{a2} + \ldots + k_p\vec{A}_{ap} = \vec{0}$$

has a non-trivial solution.

So

$$k_1\vec{A}_{a1} + k_2\vec{A}_{a2} + \ldots + k_p\vec{A}_{ap} + 0\vec{A}_{a(p+1)} + \ldots\ldots. + 0\vec{A}_{am} = \vec{0}$$

also has a non-trivial solution too, where $\vec{A}_{a(p+1)}, \ldots, \vec{A}_{am}$ 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 $\vec{A}_1, \vec{A}_2, \ldots, \vec{A}_m$ be linearly dependent set of vectors, then there exists a set of scalars $k_1, \ldots, k_m$ not all of which are zero for the linear combination equation

$$k_1\vec{A}_1 + k_2\vec{A}_2 + \ldots + k_m\vec{A}_m = \vec{0}.$$

Let k_p be one of the non-zero values of $k_i, i = 1, \ldots, m$, that is, $k_p \neq 0$, then

$$A_p = -\frac{k_2}{k_p}\vec{A}_2 - \ldots - \frac{k_{p-1}}{k_p}\vec{A}_{p-1} - \frac{k_{p+1}}{k_p}\vec{A}_{p+1} - \ldots - \frac{k_m}{k_p}\vec{A}_m.$$

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

$$a_{11}x_1 + \ldots + a_{1n}x_n = c_1$$

$$a_{21}x_1 + \ldots + a_{2n}x_n = c_2$$

$$\vdots \qquad \vdots$$

$$\vdots \qquad \vdots$$

$$a_{m1}x_1 + \ldots + a_{mn}x_n = c_n$$

where

$x_1, x_2, \ldots, x_n$ are the unknowns, then in the vector notation they can be written as

$$x_1\vec{A}_1 + x_2\vec{A}_2 + \ldots + x_n\vec{A}_n = \vec{C}$$

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 $x_1, x_2, \ldots.., x_n$ such that the linear combination

$$x_1\vec{A}_1 + \ldots\ldots\ldots + x_n\vec{A}_n$$

is equal to the $\vec{C}$, that is

$$x_1\vec{A}_1 + \ldots\ldots\ldots + x_n\vec{A}_n = \vec{C}$$

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 $\vec{A} = [a_1, a_2, \ldots, a_n]$ and $\vec{B} = [b_1, b_2, \ldots, b_n]$ be two *n*-dimensional vectors. Then the dot product of the two vectors $\vec{A}$ and $\vec{B}$ is defined as

$$\vec{A} \cdot \vec{B} = a_1b_1 + a_2b_2 + \ldots + a_nb_n = \sum_{i=1}^{n} a_ib_i$$

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

$$\begin{aligned} \vec{A} \cdot \vec{B} &= [4,1,2,3].[3,1,7,2] \\ &= (4)(3)+(1)(1)+(2)(7)+(3)(2) \\ &= 33 \end{aligned}$$

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 $\vec{C}$.

$$\begin{aligned} \vec{W} \cdot \vec{C} &= [200,250,310] \cdot [20.23,30.56,29.12] \\ &= (200)(20.23) + (250)(30.56) + (310)(29.12) \\ &= 4046 + 7640 + 9027.2 \end{aligned}$$

$$= \$20713.20$$

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

4.3 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.

$$[A]=\begin{bmatrix}5 & 2 & 3\\ 1 & 2 & 7\end{bmatrix} \qquad [B]=\begin{bmatrix}6 & 7 & -2\\ 3 & 5 & 19\end{bmatrix}$$

Solution

$$[C]=[A]+[B]$$

$$=\begin{bmatrix}5 & 2 & 3\\ 1 & 2 & 7\end{bmatrix}+\begin{bmatrix}6 & 7 & -2\\ 3 & 5 & 19\end{bmatrix}$$

$$=\begin{bmatrix}5+6 & 2+7 & 3-2\\ 1+3 & 2+5 & 7+19\end{bmatrix}$$

$$= \begin{bmatrix} 11 & 9 & 1 \\ 4 & 7 & 26 \end{bmatrix}$$

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

$$[C] = [A] + [B]$$

$$= \begin{bmatrix} 25 & 20 & 3 & 2 \\ 5 & 10 & 15 & 25 \\ 6 & 16 & 7 & 27 \end{bmatrix} + \begin{bmatrix} 20 & 5 & 4 & 0 \\ 3 & 6 & 15 & 21 \\ 4 & 1 & 7 & 20 \end{bmatrix}$$

$$= \begin{bmatrix} (25+20) & (20+5) & (3+4) & (2+0) \\ (5+3) & (10+6) & (15+15) & (25+21) \\ (6+4) & (16+1) & (7+7) & (27+20) \end{bmatrix}$$

$$= \begin{bmatrix} 45 & 25 & 7 & 2 \\ 8 & 16 & 30 & 46 \\ 10 & 17 & 14 & 47 \end{bmatrix}$$

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.

$$[A]=\begin{bmatrix} 25 & 20 & 3 & 2 \\ 5 & 10 & 15 & 25 \\ 6 & 16 & 7 & 27 \end{bmatrix}$$

$$[B]=\begin{bmatrix} 20 & 5 & 4 & 0 \\ 3 & 6 & 15 & 21 \\ 4 & 1 & 7 & 20 \end{bmatrix}$$

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. How many more tires did store A sell than store B of each brand in each quarter?

Solution

$$\begin{aligned}[D]&=[A]-[B]\\ &=\begin{bmatrix} 25 & 20 & 3 & 2 \\ 5 & 10 & 15 & 25 \\ 6 & 16 & 7 & 27 \end{bmatrix}-\begin{bmatrix} 20 & 5 & 4 & 0 \\ 3 & 6 & 15 & 21 \\ 4 & 1 & 7 & 20 \end{bmatrix}\\ &=\begin{bmatrix} 25-20 & 20-5 & 3-4 & 2-0 \\ 5-3 & 10-6 & 15-15 & 25-21 \\ 6-4 & 16-1 & 7-7 & 27-20 \end{bmatrix}\end{aligned}$$

$$= \begin{bmatrix} 5 & 15 & -1 & 2 \\ 2 & 4 & 0 & 4 \\ 2 & 15 & 0 & 7 \end{bmatrix}$$

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?

$$\begin{aligned} c_{ij} &= \sum_{k=1}^{p} a_{ik} b_{kj} \\ &= a_{i1}b_{1j} + a_{i2}b_{2j} + \ldots\ldots + a_{ip}b_{pj} \end{aligned}$$

for each $i = 1,\ 2, \ldots\ldots, m$ and $j = 1,\ 2, \ldots\ldots, n$.

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,

$$\begin{aligned} c_{ij} &= \begin{bmatrix} a_{i1} & a_{i2} \ldots\ldots a_{ip} \end{bmatrix} \begin{bmatrix} b_{1j} \\ b_{2j} \\ \vdots \\ \vdots \\ b_{pj} \end{bmatrix} \\ &= a_{i1}\,b_{1j} + \ a_{i2}\,b_{2j} + \ldots\ldots.. + a_{ip}\,b_{pj}. \\ &= \sum_{k=1}^{p} a_{ik} b_{kj} \end{aligned}$$

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 $[B]$,

$$\begin{aligned} c_{12} &= \begin{bmatrix} 5 & 2 & 3 \end{bmatrix} \begin{bmatrix} -2 \\ -8 \\ -10 \end{bmatrix} \\ &= (5)(-2) + (2)(-8) + (3)(-10) \\ &= -56 \end{aligned}$$

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 $\begin{bmatrix} 33.25 & 40.19 & 25.03 \end{bmatrix}$, so the per quarter sales of store A would be given by

$$[C] = \begin{bmatrix} 33.25 & 40.19 & 25.03 \end{bmatrix} \begin{bmatrix} 25 & 20 & 3 & 2 \\ 5 & 10 & 15 & 25 \\ 6 & 16 & 7 & 27 \end{bmatrix}$$

$$c_{ij} = \sum_{k=1}^{3} a_{ik} b_{kj}$$

$$\begin{aligned} c_{11} &= \sum_{k=1}^{3} a_{1k} b_{k1} \\ &= a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} \\ &= (33.25)(25) + (40.19)(5) + (25.03)(6) \\ &= \$1182.38 \end{aligned}$$

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

$$[C] = [1182.38 \quad 1467.38 \quad 877.81 \quad 1747.06]$$

Remember since we are multiplying a 1×3 matrix by a 3×4 matrix, the resulting matrix is a 1×4 matrix.

What is the scalar multiplication of a matrix (matrisilerin skaler çarpımı nedir)?

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

$b_{ij} = k\,a_{ij}$ for all *i, j*.

Example 7

Let

$$[A] = \begin{bmatrix} 2.1 & 3 & 2 \\ 5 & 1 & 6 \end{bmatrix}$$

Find $2[A]$

Solution

$$2[A] = 2\begin{bmatrix} 2.1 & 3 & 2 \\ 5 & 1 & 6 \end{bmatrix}$$
$$= \begin{bmatrix} 2\times 2.1 & 2\times 3 & 2\times 2 \\ 2\times 5 & 2\times 1 & 2\times 6 \end{bmatrix}$$
$$= \begin{bmatrix} 4.2 & 6 & 4 \\ 10 & 2 & 12 \end{bmatrix}$$

What is a linear combination of matrices (matrisin çizgisel kombinasyonu nedir)?

If $[A_1],[A_2],.....[A_p]$ are matrices of the same size and $k_1, k_2,.....,k_p$ are scalars, then

$$k_1[A_1] + k_2[A_2] + ... + k_p[A_p]$$

is called a linear combination of $[A_1],[A_2],...,[A_p]$.

Example 8

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

$$[A_1] + 2[A_2] - 0.5[A_3]$$

$$=\begin{bmatrix}5 & 6 & 2\\ 3 & 2 & 1\end{bmatrix}+2\begin{bmatrix}2.1 & 3 & 2\\ 5 & 1 & 6\end{bmatrix}-0.5\begin{bmatrix}0 & 2.2 & 2\\ 3 & 3.5 & 6\end{bmatrix}$$

$$=\begin{bmatrix}5 & 6 & 2\\ 3 & 2 & 1\end{bmatrix}+\begin{bmatrix}4.2 & 6 & 4\\ 10 & 2 & 12\end{bmatrix}-\begin{bmatrix}0 & 1.1 & 1\\ 1.5 & 1.75 & 3\end{bmatrix}$$

$$=\begin{bmatrix}9.2 & 10.9 & 5\\ 11.5 & 2.25 & 10\end{bmatrix}$$

What are some of the rules of binary matrix operations (İkili matris işlemlerinin kuralları nelerdir)?

Commutative law of addition (yerdeğiştirme özelliği)

If $[A]$ and $[B]$ are $m\times n$ matrices, then

$$[A]+[B]=[B]+[A]$$

Associative law of addition (birleşme özelliği)

If [A], [B] and [C] are all $m\times n$ matrices, then

$$[A]+([B]+[C])=([A]+[B])+[C]$$

Associative law of multiplication (çarpma kuralı)

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 (dağılma kuralı)

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

$$[B][C]=\begin{bmatrix}2 & 5\\ 9 & 6\end{bmatrix}\begin{bmatrix}2 & 1\\ 3 & 5\end{bmatrix}$$

$$=\begin{bmatrix}19 & 27\\36 & 39\end{bmatrix}$$

$$[A]([B][C])=\begin{bmatrix}1 & 2\\3 & 5\\0 & 2\end{bmatrix}\begin{bmatrix}19 & 27\\36 & 39\end{bmatrix}$$

$$=\begin{bmatrix}91 & 105\\237 & 276\\72 & 78\end{bmatrix}$$

$$[A][B]=\begin{bmatrix}1 & 2\\3 & 5\\0 & 2\end{bmatrix}\begin{bmatrix}2 & 5\\9 & 6\end{bmatrix}$$

$$=\begin{bmatrix}20 & 17\\51 & 45\\18 & 12\end{bmatrix}$$

$$([A][B])[C]=\begin{bmatrix}20 & 17\\51 & 45\\18 & 12\end{bmatrix}\begin{bmatrix}2 & 1\\3 & 5\end{bmatrix}$$

$$=\begin{bmatrix}91 & 105\\237 & 276\\72 & 78\end{bmatrix}$$

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

$$[A][B] = \begin{bmatrix} 6 & 3 \\ 2 & 5 \end{bmatrix}\begin{bmatrix} -3 & 2 \\ 1 & 5 \end{bmatrix}$$

$$= \begin{bmatrix} -15 & 27 \\ -1 & 29 \end{bmatrix}$$

$$[B][A] = \begin{bmatrix} -3 & 2 \\ 1 & 5 \end{bmatrix}\begin{bmatrix} 6 & 3 \\ 2 & 5 \end{bmatrix}$$

$$= \begin{bmatrix} -14 & 1 \\ 16 & 28 \end{bmatrix}$$

$$[A][B] \neq [B][A]$$

Key Terms:

Addition of matrices
Subtraction of matrices
Multiplication of matrices
Scalar Product of matrices
Linear Combination of Matrices
Rules of Binary Matrix Operation

4.4 Chapter 04.04 Unary Matrix Operations (Tekil matris işlemleri)

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^{th} row and the j^{th} column element of $[A]$ is the j^{th} row and i^{th} column element of $[B]$. Note, $[B]$ would be a $n \times m$ matrix. The transpose of $[A]$ is denoted by $[A]^{\mathrm{T}}$.

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

$$[A]^{\mathrm{T}} = \begin{bmatrix} 25 & 5 & 6 \\ 20 & 10 & 16 \\ 3 & 15 & 7 \\ 2 & 25 & 27 \end{bmatrix}$$

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, $\left([A]^{\mathrm{T}}\right)^{\mathrm{T}} = [A]$. Also, $(A+B)^{\mathrm{T}} = A^{\mathrm{T}} + B^{\mathrm{T}}; (cA)^{\mathrm{T}} = cA^{\mathrm{T}}$.

What is a symmetric matrix (Simetrik matris nedir)?

A square matrix $[A]$ with real elements where $a_{ij} = a_{ji}$ for $i = 1,2,...,n$ and $j = 1,2,...,n$ is called a symmetric matrix. This is same as saying that if $[A] = [A]^{\mathrm{T}}$, then $[A]^{\mathrm{T}}$ 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 (çapraz-simetrik matris nedir)?

A $n \times n$ matrix is skew symmetric if $a_{ij} = -a_{ji}$ for $i = 1,...,n$ and $j = 1,...,n$. This is same as

$$[A] = -[A]^{\mathrm{T}}.$$

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 (Matrisin İzi Nedir)?
The trace of a $n \times n$ matrix $[A]$ is the sum of the diagonal entries of $[A]$, that is,

$$\mathrm{tr}[A] = \sum_{i=1}^{n} a_{ii}$$

Example 4
Find the trace of

$$[A] = \begin{bmatrix} 15 & 6 & 7 \\ 2 & -4 & 2 \\ 3 & 2 & 6 \end{bmatrix}$$

Solution

$$\begin{aligned} \mathrm{tr}[A] &= \sum_{i=1}^{3} a_{ii} \\ &= (15) + (-4) + (6) \\ &= 17 \end{aligned}$$

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 $[B]$.

$$[C]=[B]^{\mathrm{T}}$$

$$=\begin{bmatrix} 33.25 & 30.01 & 35.02 & 30.05 \\ 40.19 & 38.02 & 41.03 & 38.23 \\ 25.03 & 22.02 & 27.03 & 22.95 \end{bmatrix}^{\mathrm{T}}$$

$$=\begin{bmatrix} 33.25 & 40.19 & 25.03 \\ 30.01 & 38.02 & 22.02 \\ 35.02 & 41.03 & 27.03 \\ 30.05 & 38.23 & 22.95 \end{bmatrix}$$

Recognize now that if we find $[A][C]$, we get

$$[D]=[A][C]$$

$$=\begin{bmatrix} 25 & 20 & 3 & 2 \\ 5 & 10 & 15 & 25 \\ 6 & 16 & 7 & 27 \end{bmatrix}\begin{bmatrix} 33.25 & 40.19 & 25.03 \\ 30.01 & 38.02 & 22.02 \\ 35.02 & 41.03 & 27.03 \\ 30.05 & 38.23 & 22.95 \end{bmatrix}$$

$$=\begin{bmatrix} 1597 & 1965 & 1193 \\ 1743 & 2152 & 1325 \\ 1736 & 2169 & 1311 \end{bmatrix}$$

The diagonal elements give the sales of each brand of tire for the whole year, that is

$$d_{11}=\$1597 \quad \text{(Tirestone sales)}$$
$$d_{22}=\$2152 \quad \text{(Michigan sales)}$$
$$d_{33}=\$1311 \quad \text{(Cooper sales)}$$

The total yearly sales of all three brands of tires are

$$\sum_{i=1}^{3} d_{ii} = 1597+2152+1311$$
$$=\$5060$$

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 2×2 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

$$\det(A) = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} M_{ij} \text{ for any } i = 1, 2, \cdots, n$$

or

$$\det(A) = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} M_{ij} \text{ for any } j = 1, 2, \cdots, n$$

Coupled that with $\det(A) = a_{11}$ for a 1×1 matrix $[A]$, we can always reduce the determinant of a matrix to determinants of 1×1 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

$$\det(A) = \sum_{j=1}^{n} a_{ij} C_{ij} \text{ for any } i = 1, 2, \cdots, n$$

or

$$\det(A) = \sum_{i=1}^{n} a_{ij} C_{ij} \text{ for any } j = 1, 2, \cdots, n$$

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

$$\begin{aligned} \det(A) &= \sum_{j=1}^{3} (-1)^{1+j} a_{1j} M_{1j} \\ &= (-1)^{1+1} a_{11} M_{11} + (-1)^{1+2} a_{12} M_{12} + (-1)^{1+3} a_{13} M_{13} \\ &= a_{11} M_{11} - a_{12} M_{12} + a_{13} M_{13} \end{aligned}$$

$$M_{11} = \begin{vmatrix} 8 & 1 \\ 12 & 1 \end{vmatrix}$$
$$= -4$$
$$M_{12} = \begin{vmatrix} 64 & 1 \\ 144 & 1 \end{vmatrix}$$
$$= -80$$
$$M_{13} = \begin{vmatrix} 64 & 8 \\ 144 & 12 \end{vmatrix}$$
$$= -384$$
$$\det(A) = a_{11}M_{11} - a_{12}M_{12} + a_{13}M_{13}$$
$$= 25(-4) - 5(-80) + 1(-384)$$
$$= -100 + 400 - 384$$
$$= -84$$

Also for $i = 1$,

$$\det(A) = \sum_{j=1}^{3} a_{1j}C_{1j}$$
$$C_{11} = (-1)^{1+1} M_{11}$$
$$= M_{11}$$
$$= -4$$
$$C_{12} = (-1)^{1+2} M_{12}$$
$$= -M_{12}$$
$$= 80$$
$$C_{13} = (-1)^{1+3} M_{13}$$
$$= M_{13}$$
$$= -384$$
$$\det(A) = a_{11}C_{11} + a_{21}C_{21} + a_{31}C_{31}$$
$$= (25)(-4) + (5)(80) + (1)(-384)$$
$$= -100 + 400 - 384$$
$$= -84$$

Method 2:

$\det(A) = \sum_{i=1}^{3} (-1)^{i+j} a_{ij}M_{ij}$ for any $j = 1,2,3$

Let us choose $j = 2$ in the formula

$$\det(A) = \sum_{i=1}^{3} (-1)^{i+2} a_{i2}M_{i2}$$
$$= (-1)^{1+2} a_{12}M_{12} + (-1)^{2+2} a_{22}M_{22} + (-1)^{3+2} a_{32}M_{32}$$
$$= -a_{12}M_{12} + a_{22}M_{22} - a_{32}M_{32}$$

$$M_{12} = \begin{vmatrix} 64 & 1 \\ 144 & 1 \end{vmatrix}$$
$$= -80$$
$$M_{22} = \begin{vmatrix} 25 & 1 \\ 144 & 1 \end{vmatrix}$$
$$= -119$$
$$M_{32} = \begin{vmatrix} 25 & 1 \\ 64 & 1 \end{vmatrix}$$
$$= -39$$
$$\det(A) = -a_{12}M_{12} + a_{22}M_{22} - a_{32}M_{32}$$
$$= -5(-80) + 8(-119) - 12(-39)$$
$$= 400 - 952 + 468$$
$$= -84$$

In terms of cofactors for $j = 2$,

$$\det(A) = \sum_{i=1}^{3} a_{i2}C_{i2}$$
$$C_{12} = (-1)^{1+2} M_{12}$$
$$= -M_{12}$$
$$= 80$$
$$C_{22} = (-1)^{2+2} M_{22}$$
$$= M_{22}$$
$$= -119$$
$$C_{32} = (-1)^{3+2} M_{32}$$
$$= -M_{32}$$
$$= 39$$
$$\det(A) = a_{12}C_{12} + a_{22}C_{22} + a_{32}C_{32}$$
$$= (5)(80) + (8)(-119) + (12)(39)$$
$$= 400 - 952 + 468$$
$$= -84$$

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 k, 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 $[A]$,

$$\begin{aligned} \det(\mathrm{B}) &= \det(\mathrm{A}) \\ &= -84 \end{aligned}$$

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

$$\det(A) = \prod_{i=1}^{3} a_{ii}$$
$$= a_{11} \times a_{22} \times a_{33}$$
$$= 25 \times (-4.8) \times 0.7$$
$$= -84$$

Key Terms:
Transpose
Symmetric Matrix
Skew-Symmetric Matrix
Trace of Matrix
Determinant

4.5 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 (Matris cebiri eşitliklerin çözümünde kullanılabilir. Bunu tanımlayabilir misiniz)?

Matrix algebra is used to solve a system of simultaneous linear equations. In fact (aslında), 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. (Matris cebiri çizgisel eşitliklerin aynı anda çözümünde kullanılabilir. Aslında birçok matematiksel işlemde örneğin çizgisel olmayan eşitliklerin çözümünde, interpolasyonda, integral alma işleminde, diferensiyel eşitliklerde ve çizgisel eşitliklerin aynı anda çözümünde denklem sayılarının azaltılmasında kullanılabilir. Matrislerin interpolasyonda kullanımına bir bakalım.)

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*	Velocity, *v*
(s)	(m/s)
5	106.8
8	177.2
12	279.2

The velocity data is approximated by a polynomial as

$$v(t) = at^2 + bt + c, \quad 5 \le \mathrm{t} \le 12.$$

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. (Yukarıdaki gösterim matris cebirine neden ihtiyaç olduğunu göstermektedir. Eşitliklerin çözümü ile ilgili bilgiler bölümün sonuna doğru verilecektir.)

A general set of m linear equations and n unknowns,

$$a_{11}x_1 + a_{12}x_2 + \cdots\cdots + a_{1n}x_n = c_1$$

$$a_{21}x_1 + a_{22}x_2 + \cdots\cdots + a_{2n}x_n = c_2$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$a_{m1}x_1 + a_{m2}x_2 + \ldots\ldots + a_{mn}x_n = c_m$$

can be rewritten in the matrix form as

$$\begin{bmatrix} a_{11} & a_{12} & . & . & a_{1n} \\ a_{21} & a_{22} & . & . & a_{2n} \\ \vdots & & & & \vdots \\ \vdots & & & & \vdots \\ a_{m1} & a_{m2} & . & . & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_n \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ \cdot \\ \cdot \\ c_m \end{bmatrix}$$

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

$$[A \vdots C] = \begin{bmatrix} a_{11} & a_{12} & \ldots\ldots & a_{1n} & \vdots c_1 \\ a_{21} & a_{22} & \ldots\ldots & a_{2n} & \vdots c_2 \\ \vdots & & & & \vdots \\ \vdots & & & & \vdots \\ a_{m1} & a_{m2} & \ldots\ldots & a_{mn} & \vdots c_n \end{bmatrix}$$

A system of equations can be consistent or inconsistent. What does that mean? (eşitlikler sistemi tutarlı veya tutarsız olabilir. Bu ne anlama gelir?)

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).

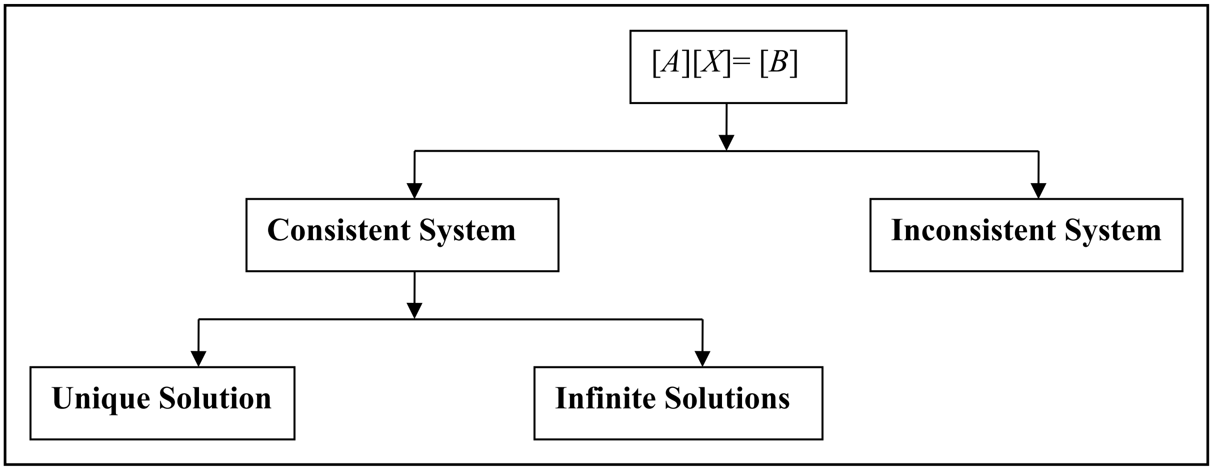


Figure 5.1. Consistent and inconsistent system of equations flow chart.

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 $[A\vdots C]$
A system of equations $[A][X]=[C]$ is *inconsistent* if the rank of A is less than the rank of the augmented matrix $[A\vdots C]$.

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 2×2 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 2×2 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

$$[B]=\begin{bmatrix} 25 & 5 & 1 & \vdots & 106.8 \\ 64 & 8 & 1 & \vdots & 177.2 \\ 144 & 12 & 1 & \vdots & 279.2 \end{bmatrix}$$

Since there are no square submatrices of order 4 as $[B]$ is a 3×4 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 4×3 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 3×3 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 2×2 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 (artan) 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 4×4 as the augmented matrix $[B]$ is a 4×3 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 3×3 submatrices have a determinant not equal to zero. For example, a square submatrix of order 3×3 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.

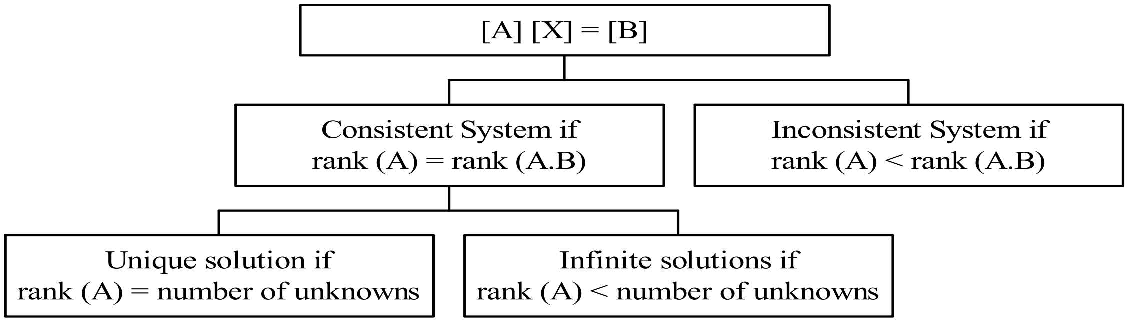

[A] [X] = [B]
Consistent System if rank (A) = rank (A.B)
Inconsistent System if rank (A) < rank (A.B)
Unique solution if rank (A) = number of unknowns
Infinite solutions if rank (A) < number of unknowns

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 $[A \vdots C]$ 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 $(A \vdots C)$ 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 $[A \vdots C]$ 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

rank of augmented matrix = 3

rank of coefficient matrix = 3

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

rank of augmented matrix = 4

rank of coefficient matrix = 3

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

rank of augmented matrix = 2
rank of coefficient matrix = 2

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 $[A]=\dfrac{[C]}{[B]}$, 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

$$[\mathrm{A}]=\begin{bmatrix}-3 & 2\\5 & -3\end{bmatrix}$$

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

$$\begin{aligned}[A][B] &= \begin{bmatrix}-3 & 2\\5 & -3\end{bmatrix}\begin{bmatrix}3 & 2\\5 & 3\end{bmatrix}\\ &= \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\\ &= [\mathrm{I}]\end{aligned}$$

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

$$[A]=\begin{bmatrix} a_{11} & a_{12} & \cdot & \cdot & a_{1n} \\ a_{21} & a_{22} & \cdot & \cdot & a_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \cdot & \cdot & a_{nn} \end{bmatrix}$$

$$[A]^{-1}=\begin{bmatrix} a'_{11} & a'_{12} & \cdot & \cdot & a'_{1n} \\ a'_{21} & a'_{22} & \cdot & \cdot & a'_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ a'_{n1} & a'_{n2} & \cdot & \cdot & a'_{nn} \end{bmatrix}$$

$$[I]=\begin{bmatrix} 1 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 1 & & & & 0 \\ 0 & & \cdot & & & \cdot \\ \cdot & & & 1 & & \cdot \\ \cdot & & & & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & 1 \end{bmatrix}$$

Using the definition of matrix multiplication, the first column of the $[A]^{-1}$ matrix can then be found by solving

$$\begin{bmatrix} a_{11} & a_{12} & \cdot & \cdot & a_{1n} \\ a_{21} & a_{22} & \cdot & \cdot & a_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \cdot & \cdot & a_{nn} \end{bmatrix}\begin{bmatrix} a'_{11} \\ a'_{21} \\ \cdot \\ \cdot \\ a'_{n1} \end{bmatrix}=\begin{bmatrix} 1 \\ 0 \\ \cdot \\ \cdot \\ 0 \end{bmatrix}$$

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

$$v(t) = at^2 + bt + c, \quad 5 \le t \le 12$$

We found that the coefficients $a, b, \text{and}\, c$ in $v(t)$ are given by

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix} \begin{bmatrix} \text{a} \\ \text{b} \\ \text{c} \end{bmatrix} = \begin{bmatrix} 106.8 \\ 177.2 \\ 279.2 \end{bmatrix}$$

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 $a, b, \text{and}\, c.$

Solution

If

$$[A]^{-1} = \begin{bmatrix} a'_{11} & a'_{12} & a'_{13} \\ a'_{21} & a'_{22} & a'_{23} \\ a'_{31} & a'_{32} & a'_{33} \end{bmatrix}$$

is the inverse of $[A]$, then

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix} \begin{bmatrix} a'_{11} & a'_{12} & a'_{13} \\ a'_{21} & a'_{22} & a'_{23} \\ a'_{31} & a'_{32} & a'_{33} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

gives three sets of equations

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix} \begin{bmatrix} a'_{11} \\ a'_{21} \\ a'_{31} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix} \begin{bmatrix} a'_{12} \\ a'_{22} \\ a'_{32} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix}\begin{bmatrix} a'_{13} \\ a'_{23} \\ a'_{33} \end{bmatrix}=\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

Solving the above three sets of equations separately gives

$$\begin{bmatrix} a'_{11} \\ a'_{21} \\ a'_{31} \end{bmatrix}=\begin{bmatrix} 0.04762 \\ -0.9524 \\ 4.571 \end{bmatrix}$$

$$\begin{bmatrix} a'_{12} \\ a'_{22} \\ a'_{32} \end{bmatrix}=\begin{bmatrix} -0.08333 \\ 1.417 \\ -5.000 \end{bmatrix}$$

$$\begin{bmatrix} a'_{13} \\ a'_{23} \\ a'_{33} \end{bmatrix}=\begin{bmatrix} 0.03571 \\ -0.4643 \\ 1.429 \end{bmatrix}$$

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 $[A]^{-1}$,

$$[A]^{-1}[A][X]=[A]^{-1}[C]$$

$$[\mathrm{X}]=[A]^{-1}[C]$$

$$\begin{bmatrix} 0.04762 & -0.08333 & 0.03571 \\ -0.9524 & 1.417 & -0.4643 \\ 4.571 & -5.000 & 1.429 \end{bmatrix}\begin{bmatrix} 106.8 \\ 177.2 \\ 279.2 \end{bmatrix}$$

Hence

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 0.2905 \\ 19.69 \\ 1.086 \end{bmatrix}$$

So

$$v(t) = 0.2905t^2 + 19.69t + 1.086, 5 \le t \le 12$$

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

$$[A]^{-1} = \frac{1}{\det(A)} adj(A)$$

where

$$adj(A) = \begin{bmatrix} C_{11} & C_{12} & \cdots & C_{1n} \\ C_{21} & C_{22} & & C_{2n} \\ \vdots & & & \\ C_{n1} & C_{n2} & \cdots & C_{nn} \end{bmatrix}^{\mathrm{T}}$$

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

$$M_{11} = \begin{vmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{vmatrix}$$

$$= \begin{vmatrix} 8 & 1 \\ 12 & 1 \end{vmatrix}$$

$$= -4$$

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 $[C]^{\mathrm{T}}$,

$$\begin{aligned} adj(A) &= [C]^{\mathrm{T}} \\ &= \begin{bmatrix} -4 & 7 & -3 \\ 80 & -119 & 39 \\ -384 & 420 & -120 \end{bmatrix} \end{aligned}$$

Hence

$$[A]^{-1} = \frac{1}{\det(A)} adj(A)$$

$$= \frac{1}{-84}\begin{bmatrix} -4 & 7 & -3 \\ 80 & -119 & 39 \\ -384 & 420 & -120 \end{bmatrix}$$

$$= \begin{bmatrix} 0.04762 & -0.08333 & 0.03571 \\ -0.9524 & 1.417 & -0.4643 \\ 4.571 & -5.000 & 1.429 \end{bmatrix}$$

If the inverse of a square matrix [A] exists, is it unique? (bir kare [A] matrisinin tersi varsa denklem sistemininçözümü de var mıdır?)

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

4.5.1 Multiple-Choice Test Chapter 04.01 Background Simultaneous Linear Equations

1. Given $[A]=\begin{bmatrix} 6 & 2 & 3 & 9 \\ 0 & 1 & 2 & 3 \\ 0 & 0 & 4 & 5 \\ 0 & 0 & 0 & 6 \end{bmatrix}$ then $[A]$ is a (an) ______________ matrix.

(A) diagonal (B) identity (C) lower triangular (D) upper triangular

2. A square matrix $[A]$ is lower triangular if

(A) $a_{ij} = 0, j > i$ (B) $a_{ij} = 0, i > j$ (C) $a_{ij} \neq 0, i > j$ (D) $a_{ij} \neq 0, j > i$

3. Given

$$[A] = \begin{bmatrix} 12.3 & -12.3 & 20.3 \\ 11.3 & -10.3 & -11.3 \\ 10.3 & -11.3 & -12.3 \end{bmatrix}, \quad [B] = \begin{bmatrix} 2 & 4 \\ -5 & 6 \\ 11 & -20 \end{bmatrix}$$

then if

$[C] = [A][B]$, then

$c_{31} =$ ______________________

(A) -58.2 (B) -37.6 (C) 219.4 (D) 259.4

4. The following system of equations has ____________ solution(s).

$x + y = 2$ (x=0, y=2; x=1, y=1; x=2, y=0; x=4, y=-2, x=6, y=-4...)

$6x + 6y = 12$

(A) infinite (B) no (C) two (D) unique

5. Consider there are only two computer companies in a country. The companies are named Dude and Imac. Each year, Dude keeps 1/5th of its customers, while the rest switch to Imac. Each year, Imac keeps 1/3rd of its customers, while the rest switch to Dude. If in 2003, Dude had 1/6th of the market and Imac had 5/6th of the market, what will be the share of Dude computers when the market becomes stable? (bir şehirde 2 bilgisayar şirketinin olduğunu kabul edelim. Bu şirketlerin adları Dude ve Imac olsun. Dude şirketi her yıl müşterilerinin 1/5'ni korumakta, geri kalanlar Imac'i tercih etmektedir. Imac firması ise her yıl müşterilerinin 1/3'nü koruyabilmekte geri kalan müşteriler Dude şirketini tercih etmektedir. 2003 yılında Dude pazarın 1/6'sına sahipken Imac 5/6'sına hakimdir. Pazar kararlı hale geldiğinde Dude şirketinin payı aşağıdakilerden hangisi olabilir?)

(A) 37/90 (B) 5/11 (C) 6/11 (D) 53/90

6. Three kids - Jim, Corey and David receive an inheritance of \$2,253,453. The money is put in three trusts but is not divided equally to begin with. Corey's trust is three times that of David's because Corey made an A in Dr. Kaw's class. Each trust is put in an interest generating investment. The three trusts of Jim, Corey and David pays an interest of 6%, 8%, 11%, respectively. The total interest of all the three trusts combined at the end of the first year is \$190,740.57. The equations to find the trust money of Jim (J), Corey (C) and David (D) in a matrix form is (Jim, Corey ve David 2,253,453\$'lık mirasa konmuşlardır. Para, çocuklar arasında eşit paylaştırılmamaktadır. Corey Kaw'ın dersinden A notunu aldığı için David'in payının 3 katı kadar para almaktadır. Her pay

faiz veren bir şirkete yatırılmıştır. Bu şirket Jim'e %6, Corey'e %8, ve David'e 11 kar vermektedir.)

(A) $\begin{bmatrix} 1 & 1 & 1 \\ 0 & 3 & -1 \\ 0.06 & 0.08 & 0.11 \end{bmatrix} \begin{bmatrix} J \\ C \\ D \end{bmatrix} = \begin{bmatrix} 2{,}253{,}453 \\ 0 \\ 190{,}740.57 \end{bmatrix}$ (B) $\begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & -3 \\ 0.06 & 0.08 & 0.11 \end{bmatrix} \begin{bmatrix} J \\ C \\ D \end{bmatrix} = \begin{bmatrix} 2{,}253{,}453 \\ 0 \\ 190{,}740.57 \end{bmatrix}$

(C) $\begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & -3 \\ 6 & 8 & 11 \end{bmatrix} \begin{bmatrix} J \\ C \\ D \end{bmatrix} = \begin{bmatrix} 2{,}253{,}453 \\ 0 \\ 190{,}740.57 \end{bmatrix}$ (D) $\begin{bmatrix} 1 & 1 & 1 \\ 0 & 3 & -1 \\ 6 & 8 & 11 \end{bmatrix} \begin{bmatrix} J \\ C \\ D \end{bmatrix} = \begin{bmatrix} 2{,}253{,}453 \\ 0 \\ 19{,}074{,}057 \end{bmatrix}$

For a complete solution, refer to the links at the end of the book.

4.6 Chapter 04.06 Gaussian Elimination (Gauss eleme yöntemi)

PRE-REQUISITES

1. Matrix Algebra Basics: Binary operations on matrices and inverse of a matrix (Primer for Matrix Algebra).

OBJECTIVES

1. solve a set of simultaneous linear equations using Naïve Gauss elimination (basit Gauss eleme),
2. learn the pitfalls of the Naïve Gauss elimination method (Gauss eleme yönteminde düşülen hatalar),
3. understand the effect of round-off error when solving a set of linear equations with the Naïve Gauss elimination method (basit Gauss eleme yöntemi ile çizgisel denklemler çözülürken yuvarlama hatalarının etkisinin belirlenmesi),
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 (basit Gauss eleme yönteminin kısmen pivotlamalı Gauss eleme yöntemine dönüştürülmesi),
5. find the determinant of a square matrix using Gaussian elimination, and (Gauss eleme yönteminde kullanılacak olan kare matrisin determinantının hesaplanması)
6. understand the relationship between the determinant of a coefficient matrix and the solution of simultaneous linear equations (katsayılar matrisinin determinantının ve çizgisel eşitliklerin aynı anda çözümü arasındaki ilişkinin anlaşılması).

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 (bir grup eşitliğin sayısal çözümü nasıl yapılır)?

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

$$
\begin{aligned}
&a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n = b_1 \\
&a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \ldots + a_{2n}x_n = b_2 \\
&\quad . \qquad\qquad . \\
&\quad . \qquad\qquad . \\
&\quad . \qquad\qquad . \\
&a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \ldots + a_{nn}x_n = b_n
\end{aligned}
$$

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 (bilinmeyenlerin ileri yönde elenmesi):

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

$$a_{21}x_1 + \frac{a_{21}}{a_{11}}a_{12}x_2 + \ldots + \frac{a_{21}}{a_{11}}a_{1n}x_n = \frac{a_{21}}{a_{11}}b_1$$

Now, this equation can be subtracted from the second equation to give

$$\left(a_{22} - \frac{a_{21}}{a_{11}}a_{12}\right)x_2 + \ldots + \left(a_{2n} - \frac{a_{21}}{a_{11}}a_{1n}\right)x_n = b_2 - \frac{a_{21}}{a_{11}}b_1$$

or

$$a'_{22}x_2 + \ldots + a'_{2n}x_n = b'_2$$

where

$$
\begin{aligned}
a'_{22} &= \ a_{22} - \frac{a_{21}}{a_{11}}a_{12} \\
&\vdots \\
a'_{2n} &= \ a_{2n} - \frac{a_{21}}{a_{11}}a_{1n}
\end{aligned}
$$

This procedure of eliminating x_1, is now repeated for the third equation to the n^{th} equation to reduce the set of equations as

$$
\begin{aligned}
&a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n = b_1 \\
&a'_{22}x_2 + a'_{23}x_3 + \ldots + a'_{2n}x_n = b'_2 \\
&a'_{32}x_2 + a'_{33}x_3 + \ldots + a'_{3n}x_n = b'_3 \\
&. \qquad\qquad . \qquad\qquad . \\
&. \qquad\qquad . \qquad\qquad . \\
&. \qquad\qquad . \qquad\qquad . \\
&a'_{n2}x_2 + a'_{n3}x_3 + \ldots + a'_{nn}x_n = b'_n
\end{aligned}
$$

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^{th} equation to give

$$
\begin{aligned}
&a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n = b_1 \\
&a'_{22}x_2 + a'_{23}x_3 + \ldots + a'_{2n}x_n = b'_2 \\
&a''_{33}x_3 + \ldots + a''_{3n}x_n = b''_3 \\
&. \qquad\qquad . \\
&. \qquad\qquad . \\
&. \qquad\qquad . \\
&a''_{n3}x_3 + \ldots + a''_{nn}x_n = b''_n
\end{aligned}
$$

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

$$
\begin{aligned}
a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n &= b_1 \\
a'_{22}x_2 + a'_{23}x_3 + \ldots + a'_{2n}x_n &= b'_2 \\
a''_{33}x_3 + \ldots + a''_{3n}x_n &= b''_3 \\
. \qquad\qquad . \\
. \qquad\qquad . \\
. \qquad\qquad . \\
a_{nn}^{(n-1)}x_n &= b_n^{(n-1)}
\end{aligned}
$$

Back Substitution (geriye yerine koyma):

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

$$x_i = \frac{b_i^{(i-1)} - \sum_{j=i+1}^{n} a_{ij}^{(i-1)} x_j}{a_{ii}^{(i-1)}} \qquad \text{for } i = n-1, n-2, \ldots, 1$$

and

$$x_n = \frac{b_n^{(n-1)}}{a_{nn}^{(n-1)}}$$

```
Başla
denklem sisteminin derecesini giriniz, n
a(n,n) tanımlayınız
katsayılar matrisinin değerlerini giriniz,
for i=1,n
 for j=1,n+1
  a(i,j)’ye değerleri oku
 next j
next i
pivot denklem (k=1)
for k=1,n-1
 for i=k+1,n
  for j=k, n+1
   a(i,j)=a(i,j)-a(k,j)*a(i,k)/a(k,k)
  next j
 next i
next k
x(n)=a(n,n+1)/a(n, n)
for i=n-1, 1, -1
t=0.0
for j=i+1, n
t=t+a(i,j)*x(j)
end
x(i)=(a(i,n+1)-t)/a(i,i)
end
```

```
clear all;
% Defining the augmented matrix [a].
a = [25 5 1 106.8; 64 8 1 177.2; 144 12 1 279.2];
c = [25 5 1 106.8; 64 8 1 177.2; 144 12 1 279.2];
n=3
%Conducting k, or (n-1) steps of forward elimination.
for k=1:(n-1)
 %Defining the proper row elements [c] .
 for i=k+1:n
  %Generating the value that is multiplied to each equation.
  r=c(i,k)/c(k,k)
  for j=k:n+1
   %Subtracting the product of the multiplier and
   %pivot equation from the ith row to generate new rows of [c] matrix.
   c(i,j)=c(i,j)-r*c(k,j)
  end
 end
end
x(n)=c(n,n+1)/c(n, n);
for i=n-1:-1:1
 t=0.0;
 for j=i+1:n
  t=t+c(i,j)*x(j);
 end
 x(i)=(c(i,n+1)-t)/c(i,i);
end
fprintf(' ', k, c);
fprintf('\n');
for i=1:n
fprintf('%3d = %0.3f \n',i, x(i));
end
```

```
c =

   25.0000    5.0000    1.0000  106.8000
         0   -4.8000   -1.5600  -96.2080
         0         0    0.7000    0.7600


  1 = 0.290
  2 = 19.690
  3 = 1.086
```

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

$$v(t)=a_1t^2+a_2t+a_3\,, \qquad 5\le t\le 12$$

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 $64/25 = 2.56$.

$$([25 \quad 5 \quad 1] \qquad [106.8]) \times 2.56 \text{ gives Row 1 as}$$

$$[64 \quad 12.8 \quad 2.56] \qquad [273.408]$$

Subtract the result from Row 2

$$\begin{array}{rrrrr} & [64 & 8 & 1] & [177.2] \\ - & [64 & 12.8 & 2.56] & [273.408] \\ \hline & 0 & -4.8 & -1.56 & -96.208 \end{array}$$

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 $144/25 = 5.76$.

$$([25 \quad 5 \quad 1] \qquad [106.8]) \times 5.76 \text{ gives Row 1 as}$$

$$[144 \quad 28.8 \quad 5.76] \qquad [615.168]$$

Subtract the result from Row 3

$$\begin{array}{rrrrr} & [144 & 12 & 1] & [279.2] \\ - & [144 & 28.8 & 5.76] & [615.168] \\ \hline & 0 & -16.8 & -4.76 & -335.968 \end{array}$$

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 $-16.8/-4.8 = 3.5$.

$$([0 \quad -4.8 \quad -1.56] \qquad [-96.208]) \times 3.5 \text{ gives Row 2 as}$$

$$[0 \quad -16.8 \quad -5.46] \qquad [-336.728]$$

Subtract the result from Row 3

$$\begin{array}{rrrrr} & [0 & -16.8 & -4.76] & [-335.968] \\ - & [0 & -16.8 & -5.46] & [-336.728] \\ \hline & 0 & 0 & 0.7 & 0.76 \end{array}$$

to get the resulting equations as

$$\begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & 0 & 0.7 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 106.8 \\ -96.208 \\ 0.76 \end{bmatrix}$$

Back substitution
From the third equation

$$0.7a_3 = 0.76$$
$$a_3 = \frac{0.76}{0.7}$$
$$= 1.08571$$

Substituting the value of a_3 in the second equation,

$$-4.8a_2 - 1.56a_3 = -96.208$$
$$a_2 = \frac{-96.208 + 1.56a_3}{-4.8}$$
$$= \frac{-96.208 + 1.56 \times 1.08571}{-4.8}$$
$$= 19.6905$$

Substituting the value of a_2 and a_3 in the first equation,

$$25a_1 + 5a_2 + a_3 = 106.8$$
$$a_1 = \frac{106.8 - 5a_2 - a_3}{25}$$
$$= \frac{106.8 - 5 \times 19.6905 - 1.08571}{25}$$
$$= 0.290472$$

Hence the solution vector is

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 0.290472 \\ 19.6905 \\ 1.08571 \end{bmatrix}$$

The polynomial that passes through the three data points is then

$$v(t) = a_1t^2 + a_2t + a_3$$
$$= 0.290472t^2 + 19.6905t + 1.08571, \ 5 \le t \le 12$$

Since we want to find the velocity at $t = 6, 7.5, 9 \text{ and } 11$ 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 $t = 6$

$$v(6) = 0.290472(6)^2 + 19.6905(6) + 1.08571$$
$$= 129.686 \text{ m/s}$$

However we could also find all the needed values of velocity at t = 6, 7.5, 9, 11 seconds using matrix multiplication.

$$v(t) = \begin{bmatrix} 0.290472 & 19.6905 & 1.08571 \end{bmatrix} \begin{bmatrix} t^2 \\ t \\ 1 \end{bmatrix}$$

So if we want to find $v(6), v(7.5), v(9), v(11),$ it is given by

$$\begin{bmatrix} v(6) \, v(7.5) \; v(9) \, v(11) \end{bmatrix} = \begin{bmatrix} 0.290472 & 19.6905 & 1.08571 \end{bmatrix} \begin{bmatrix} 6^2 & 7.5^2 & 9^2 & 11^2 \\ 6 & 7.5 & 9 & 11 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 0.290472 & 19.6905 & 1.08571 \end{bmatrix} \begin{bmatrix} 36 & 56.25 & 81 & 121 \\ 6 & 7.5 & 9 & 11 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 129.686 & 165.104 & 201.828 & 252.828 \end{bmatrix}$$

$v(6) = 129.686 \text{ m/s}$

$v(7.5) = 165.104 \text{ m/s}$

$v(9) = 201.828 \text{ m/s}$

$v(11) = 252.828 \text{ m/s}$

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 $-3/20 = -0.15$.

$(\begin{bmatrix} 20 & 15 & 10 \end{bmatrix} \; \begin{bmatrix} 45 \end{bmatrix}) \times -0.15$ gives Row 1 as

$\begin{bmatrix} -3 & -2.25 & -1.5 \end{bmatrix} \quad \begin{bmatrix} -6.75 \end{bmatrix}$

Subtract the result from Row 2

$$\begin{array}{rrrrr} & [-3 & -2.249 & 7] & [1.751] \\ - & [-3 & -2.25 & -1.5] & [-6.75] \\ \hline & 0 & 0.001 & 8.5 & 8.501 \end{array}$$

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 $5/20=0.25$

$([20 \quad 15 \quad 10] \quad [45])\times 0.25$ gives Row 1 as

$[5 \quad 3.75 \quad 2.5] \qquad [11.25]$

Subtract the result from Row 3

$$\begin{array}{rrrrr} & [5 & 1 & 3] & [9] \\ - & [5 & 3.75 & 2.5] & [11.25] \\ \hline & 0 & -2.75 & 0.5 & -2.25 \end{array}$$

to get the resulting equations as

$$\begin{bmatrix} 20 & 15 & 10 \\ 0 & 0.001 & 8.5 \\ 0 & -2.75 & 0.5 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}=\begin{bmatrix} 45 \\ 8.501 \\ -2.25 \end{bmatrix}$$

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 $-2.75/0.001=-2750$.

$([0 \quad 0.001 \quad 8.5] \quad [8.501])\times -2750$ gives Row 2 as

$[0 \quad -2.75 \quad -23375] \qquad [-23377.75]$

Rewriting within 6 significant digits with chopping

$[0 \quad -2.75 \quad -23375] \qquad [-23377.7]$

Subtract the result from Row 3

$$\begin{array}{rrrrr} & [0 & -2.75 & 0.5] & [-2.25] \\ - & [0 & -2.75 & -23375] & [-23377.7] \\ \hline & 0 & 0 & 23375.5 & 23375.45 \end{array}$$

Rewriting within 6 significant digits with chopping

$[0 \quad 0 \quad 23375.5] \qquad [-23375.4]$

to get the resulting equations as

$$\begin{bmatrix} 20 & 15 & 10 \\ 0 & 0.001 & 8.5 \\ 0 & 0 & 23375.5 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}=\begin{bmatrix} 45 \\ 8.501 \\ 23375.4 \end{bmatrix}$$

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,

$$23375.5x_3 = 23375.4$$

$$\begin{aligned} x_3 &= \frac{23375.4}{23375.5} \\ &= 0.999995 \end{aligned}$$

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,

$$20x_1 + 15x_2 + 10x_3 = 45$$

$$\begin{aligned} x_1 &= \frac{45 - 15\,x_2 - 10x_3}{20} \\ &= \frac{45 - 15 \times 1.05 - 10 \times 0.999995}{20} \\ &= \frac{45 - 15.75 - 9.99995}{20} \\ &= \frac{29.25 - 9.99995}{20} \\ &= \frac{19.2500}{20} \\ &= 0.9625 \end{aligned}$$

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

$$[X] = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

$$= \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

Are there any pitfalls of the Naïve Gauss elimination method? (Yalın Gauss eleme yönteminde karşılaşılabilecek sıkıntılar/tuzaklar)

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.

$$\begin{bmatrix} 0 & 5 & 6 \\ 4 & 5 & 7 \\ 9 & 2 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 11 \\ 16 \\ 15 \end{bmatrix}$$

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,

$$\begin{bmatrix} 5 & 6 & 7 \\ 10 & 12 & 3 \\ 20 & 17 & 19 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 18 \\ 25 \\ 56 \end{bmatrix}$$

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 (ileri yönde elemede sıfıra bölme sorunu yoktur. İlk eşitliğin pivot elemanı yani x_1'in katsayısı 5'tir. Ancak ilk aşamanın sonunda aşağıdaki gibi bir matris elde edilir)

$$\begin{bmatrix} 5 & 6 & 7 \\ 0 & 0 & -11 \\ 0 & -7 & -9 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 18 \\ -11 \\ -16 \end{bmatrix}$$

Now at the beginning of the 2nd 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 (basit Gauss eleme yöntemi yuvarlama hatasının oluşmasına eğilimlidir). This is true when there are large numbers of equations as errors propagate (bu durum büyük sayıda eşitliklerin olduğunda hata daha başat olmaktadır). 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

$$20x_1 + 15x_2 + 10x_3 = 45$$

$$-3x_1 - 2.249x_2 + 7x_3 = 1.751$$

$$5x_1 + x_2 + 3x_3 = 9$$

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

$$\begin{bmatrix} 20 & 15 & 10 \\ -3 & -2.249 & 7 \\ 5 & 1 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 45 \\ 1.751 \\ 9 \end{bmatrix}$$

Forward Elimination of Unknowns (bilinmeyenlerin ileri yönde elenmesi)

First step

Divide Row 1 by 20 and then multiply it by –3, that is, multiply Row 1 by $-3/20 = -0.15$.

$([20 \quad 15 \quad 10] \quad [45]) \times -0.15$ gives Row 1 as

$[-3 \quad -2.25 \quad -1.5] \qquad [-6.75]$

Subtract the result from Row 2

$$\begin{array}{rrrrr} & [-3 & -2.249 & 7] & [1.751] \\ - & [-3 & -2.25 & -1.5] & [-6.75] \\ \hline & 0 & 0.001 & 8.5 & 8.501 \end{array}$$

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 $5/20 = 0.25$.

$([20 \quad 15 \quad 10] \quad [45]) \times 0.25$ gives Row 1 as

$[5 \quad 3.75 \quad 2.5] \qquad [11.25]$

Subtract the result from Row 3

$$\begin{array}{rrrrr} & [5 & 1 & 3] & [9] \\ - & [5 & 3.75 & 2.5] & [11.25] \\ \hline & 0 & -2.75 & 0.5 & -2.25 \end{array}$$

to get the resulting equations as

$$\begin{bmatrix} 20 & 15 & 10 \\ 0 & 0.001 & 8.5 \\ 0 & -2.75 & 0.5 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}=\begin{bmatrix} 45 \\ 8.501 \\ -2.25 \end{bmatrix}$$

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 $-2.75/0.001=-2750$.

$$([0 \quad 0.001 \quad 8.5] \quad [8.501])\times -2750 \text{ gives Row 2 as}$$

$$[0 \quad -2.75 \quad -23375] \qquad [-23377.75]$$

Rewriting within 5 significant digits with chopping

$$[0 \quad -2.75 \quad -23375] \qquad [-23377]$$

Subtract the result from Row 3

$$\begin{array}{rrrrr} & [0 & -2.75 & 0.5] & [-2.25] \\ - & [0 & -2.75 & -23375] & [-23377] \\ \hline & 0 & 0 & 23375 & 23374 \end{array}$$

Rewriting within 6 significant digits with chopping

$$[0 \quad 0 \quad 23375] \qquad [-23374]$$

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

$$0.001x_2 + 8.5x_3 = 8.501$$

$$x_2 = \frac{8.501 - 8.5x_3}{0.001}$$

$$= \frac{8.501 - 8.5 \times 0.99995}{0.001}$$

$$= \frac{8.501 - 8.499575}{0.001}$$

$$= \frac{8.501 - 8.4995}{0.001}$$

$$= \frac{0.0015}{0.001}$$

$$= 1.5$$

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

$$20x_1 + 15x_2 + 10x_3 = 45$$

$$x_1 = \frac{45 - 15\,x_2 - 10x_3}{20}$$

$$= \frac{45 - 15 \times 1.5 - 10 \times 0.99995}{20}$$

$$= \frac{45 - 22.5 - 9.9995}{20}$$

$$= \frac{22.5 - 9.9995}{20}$$

$$= \frac{12.5005}{20}$$

$$= \frac{12.500}{20}$$

$$= 0.625$$

Hence the solution is

$$[X] = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

$$= \begin{bmatrix} 0.625 \\ 1.5 \\ 0.99995 \end{bmatrix}$$

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^{th} step of forward elimination, one finds the maximum of

$$|a_{kk}|, |a_{k+1,k}|, \ldots\ldots\ldots\ldots, |a_{nk}|$$

Then if the maximum of these values is $|a_{pk}|$ in the p^{th} row, $k \le p \le n$, 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

$$20x_1 + 15x_2 + 10x_3 = 45$$
$$-3x_1 - 2.249x_2 + 7x_3 = 1.751$$
$$5x_1 + x_2 + 3x_3 = 9$$

using five and six significant digits with chopping in the calculations. Using five significant digits with chopping, the solution found was

$$[X] = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$
$$= \begin{bmatrix} 0.625 \\ 1.5 \\ 0.99995 \end{bmatrix}$$

This is different from the exact solution of

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

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 $-3/20 = -0.15$.

$([20 \quad 15 \quad 10] \quad [45]) \times -0.15$ gives Row 1 as

$[-3 \quad -2.25 \quad -1.5] \qquad [-6.75]$

Subtract the result from Row 2

$$\begin{array}{rrrrr} & [-3 & -2.249 & 7] & [1.751] \\ - & [-3 & -2.25 & -1.5] & [-6.75] \\ \hline & 0 & 0.001 & 8.5 & 8.501 \end{array}$$

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 $5/20 = 0.25$.

$([20 \quad 15 \quad 10] \quad [45]) \times 0.25$ gives Row 1 as

$[5 \quad 3.75 \quad 2.5] \qquad [11.25]$

Subtract the result from Row 3

$$\begin{array}{rrrrr} & [5 & 1 & 3] & [9] \\ - & [5 & 3.75 & 2.5] & [11.25] \\ \hline & 0 & -2.75 & 0.5 & -2.25 \end{array}$$

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 $0.001/-2.75 = -0.00036363$.

$$([0 \quad -2.75 \quad 0.5] \quad [-2.25]) \times -0.00036363 \text{ gives Row 2 as}$$

$$[0 \quad 0.00099998 \quad -0.00018182] \qquad [0.00081816]$$

Subtract the result from Row 3

$$\begin{array}{rllll} & [0 & 0.001 & 8.5] & [8.501] \\ - & [0 & 0.00099998 & -0.00018182] & [0.00081816] \\ \hline & 0 & 0 & 8.50018182 & 8.50018184 \end{array}$$

Rewriting within 5 significant digits with chopping

$$[0 \quad 0 \quad 8.5001] \qquad [8.5001]$$

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

$$8.5001x_3 = 8.5001$$

$$x_3 = \frac{8.5001}{8.5001}$$

$$=1$$

Substituting the value of x_3 in Row 2

$$-2.75x_2 + 0.5x_3 = -2.25$$

$$x_2 = \frac{-2.25 - 0.5x_2}{-2.75}$$

$$= \frac{-2.25 - 0.5 \times 1}{-2.75}$$

$$= \frac{-2.25 - 0.5}{-2.75}$$

$$= \frac{-2.75}{-2.75}$$

$$= 1$$

Substituting the value of x_3 and x_2 in Row 1

$$20x_1 + 15x_2 + 10x_3 = 45$$

$$x_1 = \frac{45 - 15x_2 - 10x_3}{20}$$

$$= \frac{45 - 15 \times 1 - 10 \times 1}{20}$$

$$=\frac{45-15-10}{20}$$

$$=\frac{30-10}{20}$$

$$=\frac{20}{20}$$

$$=1$$

So the solution is

$$[X]=\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}$$

$$=\begin{bmatrix}1\\1\\1\end{bmatrix}$$

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). [A]'nın nxn şeklinde bir matris oldupunu kabul edelim. [B] matrisinin'de nxn şeklinde bir matris olduğunu kabul edelim ve bir satırının diğer satırı ile toplandığında veya çıkarıldığında det(A)=det(B) şeklinde bir sonuç veribilir.

Theorem 2:

Let $[A]$ be a $n\times n$ matrix that is upper triangular, lower triangular or diagonal, then

$$\det(A)=a_{11}\times a_{22}\times\ldots\times a_{ii}\times\ldots\times a_{nn}$$

$$=\prod_{i=1}^{n}a_{ii}$$

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. [A]'nın nxn şeklinde üst üçgen, alt üçgen veya diyagonal bir matris olduğunu kabul edelim. Bu durumda

$\det(A)=a_{11}xa_{22}xa_{33}x\ldots xa_{nn}=\Pi a_{ii}$

ifadesi yazılabilir. Bunun anlamı denklem sisteminin çözümü için basit gauss eleme yöntemi kullanılıyorsa matrisin determinantının Teorem 1'e göre aynı kalacağını göstermektedir.

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

$$[B]=\begin{bmatrix}10 & -7 & 0\\ 0 & 2.5 & 5\\ 0 & 0 & 6.002\end{bmatrix}$$

$$\begin{aligned}\det(B)&=10\times 2.5\times 6.002\\ &=150.05\end{aligned}$$

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

$$\det(A)=\frac{1}{\det\left(A^{-1}\right)}$$

Solution

$$[A][A]^{-1}=[I]$$

$$\det\left(A\,A^{-1}\right)=\det\left(I\right)$$

$$\det\left(A\right)\det\left(A^{-1}\right)=1$$

$$\det\left(A\right)=\frac{1}{\det\left(A^{-1}\right)}$$

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

4.6.1 Multiple-Choice Test Chapter 04.06 Gaussian Elimination

1. The goal of forward elimination steps in the Naïve Gauss elimination method is to reduce the coefficient matrix to a (an) _____________ matrix.
   (E) diagonal
   (F) identity

(G) lower triangular
(H) upper triangular

2. Division by zero during forward elimination steps in Naïve Gaussian elimination of the set of equations $[A][X]=[C]$ implies the coefficient matrix $[A]$

(A) is invertible
(B) is nonsingular
(C) may be singular or nonsingular
(D) is singular

3. Using a computer with four significant digits with chopping, the Naïve Gauss elimination solution to

$$0.0030x_1 + 55.23x_2 = 58.12$$
$$6.239x_1 - 7.123x_2 = 47.23$$

is

(A) $x_1 = 26.66;\ x_2 = 1.051$
(B) $x_1 = 8.769;\ x_2 = 1.051$
(C) $x_1 = 8.800;\ x_2 = 1.000$
(D) $x_1 = 8.771;\ x_2 = 1.052$

4. Using a computer with four significant digits with chopping, the Gaussian elimination with partial pivoting solution to

$$0.0030x_1 + 55.23x_2 = 58.12$$
$$6.239x_1 - 7.123x_2 = 47.23$$

is

(A) $x_1 = 26.66;\ x_2 = 1.051$
(B) $x_1 = 8.769;\ x_2 = 1.051$
(C) $x_1 = 8.800;\ x_2 = 1.000$
(D) $x_1 = 8.771;\ x_2 = 1.052$

5. At the end of the forward elimination steps of the Naïve Gauss elimination method on the following equations

$$\begin{bmatrix} 4.2857\times10^{7} & -9.2307\times10^{5} & 0 & 0 \\ 4.2857\times10^{7} & -5.4619\times10^{5} & -4.2857\times10^{7} & 5.4619\times10^{5} \\ -6.5 & -0.15384 & 6.5 & 0.15384 \\ 0 & 0 & 4.2857\times10^{7} & -3.6057\times10^{5} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix} = \begin{bmatrix} -7.887\times10^{3} \\ 0 \\ 0.007 \\ 0 \end{bmatrix}$$

the resulting equations in matrix form are given by

$$\begin{bmatrix} 4.2857\times10^{7} & -9.2307\times10^{5} & 0 & 0 \\ 0 & 3.7688\times10^{5} & -4.2857\times10^{7} & 5.4619\times10^{5} \\ 0 & 0 & -26.9140 & 0.579684 \\ 0 & 0 & 0 & 5.62500\times10^{5} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix} = \begin{bmatrix} -7.887\times10^{3} \\ 7.887\times10^{3} \\ 1.19530\times10^{-2} \\ 1.90336\times10^{4} \end{bmatrix}$$

The determinant of the original coefficient matrix is

(A) 0.00

(B) 4.2857×10^{7}

(C) 5.486×10^{19}

(D) -2.445×10^{20}

6. The following data is given for the velocity of the rocket as a function of time. To find the velocity at $t = 21\,\text{s}$, you are asked to use a quadratic polynomial, $v(t) = at^2 + bt + c$ to approximate the velocity profile.

t	(s)	0	14	15	20	30	35
$v(t)$	(m/s)	0	227.04	362.78	517.35	602.97	901.67

The correct set of equations that will find a, b and c are

(A) $\begin{bmatrix} 176 & 14 & 1 \\ 225 & 15 & 1 \\ 400 & 20 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 227.04 \\ 362.78 \\ 517.35 \end{bmatrix}$

(B) $\begin{bmatrix} 225 & 15 & 1 \\ 400 & 20 & 1 \\ 900 & 30 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 362.78 \\ 517.35 \\ 602.97 \end{bmatrix}$

(C) $\begin{bmatrix} 0 & 0 & 1 \\ 225 & 15 & 1 \\ 400 & 20 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 0 \\ 362.78 \\ 517.35 \end{bmatrix}$

(D) $\begin{bmatrix} 400 & 20 & 1 \\ 900 & 30 & 1 \\ 1225 & 35 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 517.35 \\ 602.97 \\ 901.67 \end{bmatrix}$

For a complete solution, refer to the links at the end of the book.

4.7 Chapter 04.07 LU Decomposition (LU ayrıştırması)

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?(LU ayrıştırmasınınçizgisel denklem sistemlerin aynı anda çözümünde kullanıldığı söylenmektedir. Bu ne demektir?)

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. (şimdiye kadar çizgisel denklem sistemlerinin çözümü için basit Gauss eleme ve kısmi pivotlu Gauss eleme yöntemlerini kullandık. Öyleyse başka yöntemlere neden gerek var dır?)

For a nonsingular matrix $[A]$ on which one can successfully conduct the Naïve Gauss elimination forward elimination steps, one can always write it as (singülerliği olmayan [A] matrisi için yalın Gauss eleme yöntemi denklemlerin çözümü için başarılı bir şekilde kullanılabilir.)

$$[A]=[L][U]$$

where

$[L]$= Lower triangular matrix (alt üçgen matris)

$[U]$ = Upper triangular matrix (üst üçgen matris)

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 $[L]^{-1}$,

$$[L]^{-1}[L][U][X]=[L]^{-1}[C]$$

$$[I][U][X]=[L]^{-1}[C] \text{ as } \left([L]^{-1}[L]=[I]\right)$$

$$[U][X]=[L]^{-1}[C] \text{ as } ([I][U]=[U])$$

Let

$$[L]^{-1}[C]=[Z]$$

then

$$[L][Z]=[C] \qquad (1)$$

and

$$[U][X]=[Z] \qquad (2)$$

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*]? (Yukarıdakiler Gauss eleme yöntemine göre ilginç gelebilir ama daha karmaşık bir algortitmaya sahiptir. Gauss eleme yöntemine göre daha etkili olan LU ayrıştırma yöntemi ile [A][X]=[C] şeklindeki denklemler çözülebilir mi?)

For a square matrix $[A]$ of $n \times n$ size, the computational time[1] $CT|_{DE}$ 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 = clock cycle time[2].

The computational time $CT|_{FS}$ to solve by forward substitution $[L][Z]=[C]$ is given by

$$CT|_{FS} = T\left(4n^2 - 4n\right)$$

The computational time $CT|_{BS}$ to solve by back substitution $[U][X]=[Z]$ is given by

$$CT|_{BS} = T\left(4n^2 + 12n\right)$$

So, the total computational time to solve a set of equations by LU decomposition is

$$\begin{aligned} CT|_{LU} &= CT|_{DE} + CT|_{FS} + CT|_{BS} \\ &= T\left(\frac{8n^3}{3} + 4n^2 - \frac{20n}{3}\right) + T\left(4n^2 - 4n\right) + T\left(4n^2 + 12n\right) \\ &= T\left(\frac{8n^3}{3} + 12n^2 + \frac{4n}{3}\right) \end{aligned}$$

Now let us look at the computational time taken by Gaussian elimination. The computational time $CT|_{FE}$ for the forward elimination part,

$$CT|_{FE} = T\left(\frac{8n^3}{3} + 8n^2 - \frac{32n}{3}\right),$$

and the computational time $CT|_{BS}$ for the back substitution part is

$$CT|_{BS} = T\left(4n^2 + 12n\right)$$

So, the total computational time $CT|_{GE}$ to solve a set of equations by Gaussian Elimination is

$$CT|_{GE} = CT|_{FE} + CT|_{BS}$$

[1] 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_kwon.pdf

[2] As an example, a 1.2 GHz CPU has a clock cycle of $1/(1.2\times10^9) = 0.833333\,\text{ns}$

$$= T\left(\frac{8n^3}{3} + 8n^2 - \frac{32n}{3}\right) + T\left(4n^2 + 12n\right)$$

$$= T\left(\frac{8n^3}{3} + 12n^2 + \frac{4n}{3}\right)$$

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 $CT|_{inverse\,LU}$ required to find the inverse of a matrix using LU decomposition is

$$CT|_{inverse\,LU} = 1\times CT|_{DE} + n\times CT|_{FS} + n\times CT|_{BS}$$

$$= 1\times T\left(\frac{8n^3}{3} + 4n^2 - \frac{20n}{3}\right) + n\times T\left(4n^2 - 4n\right) + n\times T\left(4n^2 + 12n\right)$$

$$= T\left(\frac{32n^3}{3} + 12n^2 - \frac{20n}{3}\right)$$

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 $CT|_{inverse\,GE}$ required to find the inverse of a matrix by using Gaussian elimination then is

$$CT|_{inverse\,GE} = n\times CT|_{FE} + n\times CT|_{BS}$$

$$= n\times T\left(\frac{8n^3}{3} + 8n^2 - \frac{32n}{3}\right) + n\times T\left(4n^2 + 12n\right)$$

$$= T\left(\frac{8n^4}{3} + 12n^3 + \frac{4n^2}{3}\right)$$

Clearly for large n, $CT|_{inverse\,GE} >> CT|_{inverse\,LU}$ as $CT|_{inverse\,GE}$ has the dominating terms of n^4 and $CT|_{inverse\,LU}$ 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
$CT\|_{inverse\,GE} / CT\|_{inverse\,LU}$	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 $[A]=[L][\mathrm{U}]$? (singüler olmayan bir [A] matrisini [L][U] matrislerine nasıl ayrıştırabilirim?)

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 (İleri yönde basit Gauss eleme yöntemi singüler olmayan bir matrise uygulanabilir ve matris LU şekline ayrıştırılabilir)

$$[A]=\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

$$=\begin{bmatrix} 1 & 0 & \dots & 0 \\ \ell_{21} & 1 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ \ell_{n1} & \ell_{n2} & \cdots & 1 \end{bmatrix}\begin{bmatrix} u_{11} & u_{12} & \dots & u_{1n} \\ 0 & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & u_{nn} \end{bmatrix}$$

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 ([U] matrisi ileri yönde yalın Gauss eleme yönteminde elde edilen matrise benzemektedir.).

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 (alt tarafı üçgen [L] matrisinin köşegen elemanları 1'dir ve köşegen üzerindeki elemanları ise sıfırdır).

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

$$[A]=[L][U]$$

$$=\begin{bmatrix} 1 & 0 & 0 \\ \ell_{21} & 1 & 0 \\ \ell_{31} & \ell_{32} & 1 \end{bmatrix}\begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix}$$

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 ℓ_{21} and ℓ_{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

$$\ell_{21}=\frac{64}{25}$$
$$=2.56$$
$$\ell_{31}=\frac{144}{25}$$
$$=5.76$$

To find ℓ_{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

$$\ell_{32}=\frac{-16.8}{-4.8}$$
$$=3.5$$

Hence

$$[L]=\begin{bmatrix} 1 & 0 & 0 \\ 2.56 & 1 & 0 \\ 5.76 & 3.5 & 1 \end{bmatrix}$$

Confirm $[L][U]=[A]$.

$$[L][U]=\begin{bmatrix} 1 & 0 & 0 \\ 2.56 & 1 & 0 \\ 5.76 & 3.5 & 1 \end{bmatrix}\begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & 0 & 0.7 \end{bmatrix}$$

$$= \begin{bmatrix} 25 & 5 & 1 \\ 64 & 8 & 1 \\ 144 & 12 & 1 \end{bmatrix}$$

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

$$z_1 = 106.8$$

$$\begin{aligned} z_2 &= 177.2 - 2.56z_1 \\ &= 177.2 - 2.56 \times 106.8 \\ &= -96.208 \end{aligned}$$

$$
\begin{aligned}
z_3 &= 279.2 - 5.76z_1 - 3.5z_2 \\
&= 279.2 - 5.76 \times 106.8 - 3.5 \times (-96.208) \\
&= 0.76
\end{aligned}
$$

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

$$[U][X] = [Z]$$

$$
\begin{bmatrix} 25 & 5 & 1 \\ 0 & -4.8 & -1.56 \\ 0 & 0 & 0.7 \end{bmatrix}
\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}
=
\begin{bmatrix} 106.8 \\ -96.208 \\ 0.76 \end{bmatrix}
$$

$$
\begin{aligned}
25a_1 + 5a_2 + a_3 &= 106.8 \\
-4.8a_2 - 1.56a_3 &= -96.208 \\
0.7a_3 &= 0.76
\end{aligned}
$$

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,

$$
\begin{aligned}
25a_1 + 5a_2 + a_3 &= 106.8 \\
a_1 &= \frac{106.8 - 5a_2 - a_3}{25} \\
&= \frac{106.8 - 5 \times 19.691 - 1.0857}{25} \\
&= 0.29048
\end{aligned}
$$

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

$$[b_{11}\, b_{12} \ldots \ldots\, b_{n1}]^{\mathrm{T}}$$

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

$$\begin{aligned} z_1 &= 1 \\ z_2 &= 0 - 2.56z_1 \\ &= 0 - 2.56(1) \\ &= -2.56 \\ z_3 &= 0 - 5.76z_1 - 3.5z_2 \\ &= 0 - 5.76(1) - 3.5(-2.56) \\ &= 3.2 \end{aligned}$$

Hence

$$\begin{aligned} [Z] &= \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} \\ &= \begin{bmatrix} 1 \\ -2.56 \\ 3.2 \end{bmatrix} \end{aligned}$$

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

4.7.1 Multiple-Choice Test Chapter 04.07 LU Decomposition Method

1.The $[L][U]$ decomposition method is computationally more efficient than Naïve Gauss elimination for solving

(A) a single set of simultaneous linear equations.

(B) multiple sets of simultaneous linear equations with different coefficient matrices and the same right hand side vectors.

(C) multiple sets of simultaneous linear equations with the same coefficient matrix and different right hand side vectors.

(D) less than ten simultaneous linear equations.

2.The lower triangular matrix $[L]$ in the $[L][U]$ decomposition of the matrix given below

$$\begin{bmatrix} 25 & 5 & 4 \\ 10 & 8 & 16 \\ 8 & 12 & 22 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \ell_{21} & 1 & 0 \\ \ell_{31} & \ell_{32} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix}$$

is

$$\text{(A)} \begin{bmatrix} 1 & 0 & 0 \\ 0.40000 & 1 & 0 \\ 0.32000 & 1.7333 & 1 \end{bmatrix} \text{(B)} \begin{bmatrix} 25 & 5 & 4 \\ 0 & 6 & 14.400 \\ 0 & 0 & -4.2400 \end{bmatrix} \text{(C)} \begin{bmatrix} 1 & 0 & 0 \\ 10 & 1 & 0 \\ 8 & 12 & 0 \end{bmatrix} \text{(D)} \begin{bmatrix} 1 & 0 & 0 \\ 0.40000 & 1 & 0 \\ 0.32000 & 1.5000 & 1 \end{bmatrix}$$

3.The upper triangular matrix $[U]$ in the $[L][U]$ decomposition of the matrix given below

$$\begin{bmatrix} 25 & 5 & 4 \\ 0 & 8 & 16 \\ 0 & 12 & 22 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \ell_{21} & 1 & 0 \\ \ell_{31} & \ell_{32} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix}$$

is

$$\text{(A)} \begin{bmatrix} 1 & 0 & 0 \\ 0.40000 & 1 & 0 \\ 0.32000 & 1.7333 & 1 \end{bmatrix} \text{(B)} \begin{bmatrix} 25 & 5 & 4 \\ 0 & 6 & 14.400 \\ 0 & 0 & -4.2400 \end{bmatrix} \text{(C)} \begin{bmatrix} 25 & 5 & 4 \\ 0 & 8 & 16 \\ 0 & 0 & -2 \end{bmatrix} \text{(D)} \begin{bmatrix} 1 & 0.2000 & 0.16000 \\ 0 & 1 & 2.4000 \\ 0 & 0 & -4.240 \end{bmatrix}$$

4.For a given 2000×2000 matrix $[A]$, assume that it takes about 15 seconds to find the inverse of $[A]$ by the use of the $[L][U]$ decomposition method, that is, finding the $[L][U]$ once, and then doing forward substitution and back substitution 2000 times using the 2000 columns of the identity matrix as the right hand side vector. The approximate time, in seconds, that it will take to find the inverse if found by repeated use of the Naive Gauss elimination method, that is, doing forward elimination and back substitution 2000 times by using the 2000 columns of the identity matrix as the right hand side vector is most nearly

(A) 300 (B) 1500 (C) 7500 (D) 30000

5.The algorithm for solving a set of n equations $[A][X]=[C]$, where $[A]=[L][U]$ involves solving $[L][Z]=[C]$ by forward substitution. The algorithm to solve $[L][Z]=[C]$ is given by

(A) $z_1 = c_1 / l_{11}$

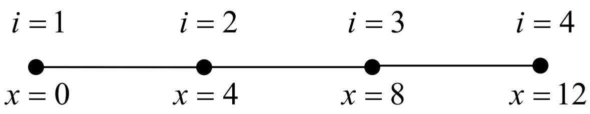
```
          for $i$ from 2 to $n$ do
          sum = 0
          for $j$ from 1 to $i$ do
          sum = sum + $l_{ij} * z_j$
          end do
          $z_i = (c_i - sum)/l_{ii}$
          end do
(B)  $z_1 = c_1/l_{11}$
          for $i$ from 2 to $n$ do
          sum = 0
          for $j$ from 1 to $(i-1)$ do
          sum = sum + $l_{ij} * z_j$
          end do
          $z_i = (c_i - sum)/l_{ii}$
          end do
(C)       $z_1 = c_1/l_{11}$
          for $i$ from 2 to $n$ do
          for $j$ from 1 to $(i-1)$do
          sum = sum + $l_{ij} * z_j$
          end do
          $z_i = (c_i - sum)/l_{ii}$
          end do
(D)  for $i$ from 2 to $n$ do
          sum = 0
          for $j$ from 1 to $(i-1)$ do
          sum = sum +$l_{ij} * z_j$
          end do
          $z_i = (c_i - sum)/l_{ii}$
          end do
```

6.To solve boundary value problems, a numerical method based on finite difference method is used. This results in simultaneous linear equations with tridiagonal coefficient matrices. These are solved using a specialized $[L][U]$ decomposition method.

Choose the set of equations that approximately solves the boundary value problem

$$\frac{d^2y}{dx^2} = 6x - 0.5x^2,\ y(0) = 0,\ y(12) = 0,\ 0 \le x \le 12$$

The second derivative in the above equation is approximated by the second order accurate central divided difference approximation as learned in the differentiation module (Chapter 02.02). A step size of $h = 4$ is used, and hence the value of y can be found approximately at equidistantly placed 4 nodes between x=0 and x=12.



(A) $$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.0625 & 0.125 & 0.0625 & 0 \\ 0 & 0.0625 & 0.125 & 0.0625 \\ 0 & 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}=\begin{bmatrix} 0 \\ 16.0 \\ 16.0 \\ 0 \end{bmatrix}$$

(B) $$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.0625 & -0.125 & 0.0625 & 0 \\ 0 & 0.0625 & -0.125 & 0.0625 \\ 0 & 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}=\begin{bmatrix} 0 \\ 16.0 \\ 16.0 \\ 0 \end{bmatrix}$$

(C) $$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0.0625 & -0.125 & 0.0625 & 0 \\ 0 & 0.0625 & -0.125 & 0.0625 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}=\begin{bmatrix} 0 \\ 0 \\ 16.0 \\ 16.0 \end{bmatrix}$$

(D) $$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0.0625 & 0.125 & 0.0625 & 0 \\ 0 & 0.0625 & 0.125 & 0.0625 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}=\begin{bmatrix} 0 \\ 0 \\ 16.0 \\ 16.0 \end{bmatrix}$$

For a complete solution, refer to the links at the end of the book.

4.8 Chapter 04.07 Gauss-Seidel Method

PRE-REQUISITES

Matrix Algebra Basics: Matrix multiplication, Diagonally dominant matrices (Primer for Matrix Algebra).

OBJECTIVES

1. solve a set of equations using the Gauss-Seidel method,
2. recognize the advantages and pitfalls of the Gauss-Seidel method, and determine under what conditions the Gauss-Seidel method always converges.

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? (çizgisel eşitlikleri çözmek için başka yönteme ihtiyacımız var mıdır?)

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 (bazı durumlarda örneğin denklem sistemi çok büyükse yineleme yöntemleri diğer yöntemlere göre daha avantajlıdır. Eleme yöntemlerinden Gauss eleme yöntemi yuvarlama hatalarını eklemelerle artırır. Yineleme yöntemlerinden Gauss-Seidel yöntemi yuvarlama hatalarının kontrolünde çok etkindir. Fizik problemi anlaşılırsa ve yineleme yöntemi kullanılıyorsa başlangıç için epey sayıda tahmini değer hazırlanarak sonuca daha kolayca ulaşılabilir).

What is the algorithm for the Gauss-Seidel method? Given a general set of n equations and n unknowns, we have (Gauss-Seidel yönteminin algoritması nasıldır? N bilinmeyenli n tane denklemden oluşan bir sistem aşağıdaki gibidir:)

$$\begin{array}{l}
a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n = c_1 \\
a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \ldots + a_{2n}x_n = c_2 \\
. \qquad\qquad . \\
. \qquad\qquad . \\
. \qquad\qquad . \\
a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \ldots + a_{nn}x_n = c_n
\end{array}$$

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 (diyagonal elementleri sıfırdan farklı ise eşitlikler bilinmeyenlere göre yeniden düzenlenir. yani x1 eşitliğin sol tarafında yalnız kalacak şekilde ve diğer terimlerde eşitliğin sağ tarafında kalacak şekilde yazılır. x2 eşitliğin solunda ve diğer terimler eşitliğin sağında olacak şekilde yeniden yazılır. Diğer bilinmeyenlerde benzer şekilde yazılarak denklem sisteminin bilinmeyenleri aşağıdaki çözülür.)

$$x_1 = \frac{c_1 - a_{12}x_2 - a_{13}x_3 \ldots\ldots - a_{1n}x_n}{a_{11}}$$

$$x_2 = \frac{c_2 - a_{21}x_1 - a_{23}x_3 \ldots\ldots - a_{2n}x_n}{a_{22}}$$

$$\vdots$$

$$\vdots$$

$$x_{n-1} = \frac{c_{n-1} - a_{n-1,1}x_1 - a_{n-1,2}x_2 \ldots\ldots - a_{n-1,n-2}x_{n-2} - a_{n-1,n}x_n}{a_{n-1,n-1}}$$

$$x_n = \frac{c_n - a_{n1}x_1 - a_{n2}x_2 - \ldots\ldots - a_{n,n-1}x_{n-1}}{a_{nn}}$$

These equations can be rewritten in a summation form as

$$x_1 = \frac{c_1 - \sum_{\substack{j=1 \\ j\neq 1}}^{n} a_{1j}x_j}{a_{11}}$$

$$x_2 = \frac{c_2 - \sum_{\substack{j=1 \\ j\neq 2}}^{n} a_{2j}x_j}{a_{22}}$$

.

.

.

$$x_{n-1} = \frac{c_{n-1} - \sum_{\substack{j=1 \\ j\neq n-1}}^{n} a_{n-1,j}x_j}{a_{n-1,n-1}}$$

$$x_n = \frac{c_n - \sum_{\substack{j=1 \\ j\neq n}}^{n} a_{nj}x_j}{a_{nn}}$$

Hence for any row i,

$$x_i = \frac{c_i - \sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}x_j}{a_{ii}}, i = 1,2,\ldots,n.$$

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 (artık xi'leri belirleyebilmek için xi'ler için başlangıç tahmin değerlerine ihtiyaç vardır ve bu değerlere göre yeni hesaplamalar yapılır. Her yineleme sonunda xi'ler için mutlak bağıl yaklaşık hata hesabı yapılır:)

$$\left|\in_a\right|_i = \left|\frac{x_i^{\text{new}} - x_i^{\text{old}}}{x_i^{\text{new}}}\right| \times 100$$

where x_i^{new} is the recently obtained value of x_i, and x_i^{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

$$v(t)=a_1t^2+a_2t+a_3\ , \qquad 5\le t\le 12$$

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 $v(t)=a_1t^2+a_2t+a_3$ passes through the three data points gives

$$v(t_1)=v_1=a_1t_1^2+a_2t_1+a_3$$
$$v(t_2)=v_2=a_1t_2^2+a_2t_2+a_3$$
$$v(t_3)=v_3=a_1t_3^2+a_2t_3+a_3$$

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

$$\begin{aligned} a_1 &= \frac{106.8 - 5(2) - (5)}{25} \\ &= 3.6720 \\ a_2 &= \frac{177.2 - 64(3.6720) - (5)}{8} \\ &= -7.8150 \\ a_3 &= \frac{279.2 - 144(3.6720) - 12(-7.8510)}{1} \\ &= -155.36 \end{aligned}$$

The absolute relative approximate error for each x_i then is

$$\begin{aligned} \left|\in_a\right|_1 &= \left|\frac{3.6720 - 1}{3.6720}\right| \times 100 \\ &= 72.76\% \\ \left|\in_a\right|_2 &= \left|\frac{-7.8510 - 2}{-7.8510}\right| \times 100 \\ &= 125.47\% \\ \left|\in_a\right|_3 &= \left|\frac{-155.36 - 5}{-155.36}\right| \times 100 \\ &= 103.22\% \end{aligned}$$

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

$$a_1 = \frac{106.8 - 5(-7.8510) - (-155.36)}{25}$$

$$= 12.056$$

$$a_2 = \frac{177.2 - 64(12.056) - (-155.36)}{8}$$

$$= -54.882$$

$$a_3 = \frac{279.2 - 144(12.056) - 12(-54.882)}{1}$$

$$= -798.34$$

The absolute relative approximate error for each x_i then is

$$|\in_a|_1 = \left|\frac{12.056 - 3.6720}{12.056}\right| \times 100$$

$$= 69.543\%$$

$$|\in_a|_2 = \left|\frac{-54.882 - (-7.8510)}{-54.882}\right| \times 100$$

$$= 85.695\%$$

$$|\in_a|_3 = \left|\frac{-798.34 - (-155.36)}{-798.34}\right| \times 100$$

$$= 80.540\%$$

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

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 12.056 \\ -54.882 \\ -798.54 \end{bmatrix}$$

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	$\|\in_a\|_1\%$	a_2	$\|\in_a\|_2\%$	a_3	$\|\in_a\|_3\%$
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? (yukarıdaki denklem sistemi bir sonuca yakınsamadı. Neden?)

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 (yineleme yöntemlerinin birçoğunda olduğu gibi bir tuzaktan dolayı bir sonuca ulaşmayabilir. Bununla birlikte belli sınıflardaki eşitlikler sistemlerinin Gauss-Seidel öteleme yöntemi ile bir sonuca yakınsanır/yaklaşılır. [A][X]=[C] şeklinde bu eşitlikler sistemlerinin [A] katsayılar matrisinin diyagonalleri diğer terimlere göre baskındır/dominanttır:)

$$|a_{ii}| \geq \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \text{ for all } i$$

$$|a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \text{ for at least one } i$$

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 (diyagonal elemanları diğer elemanlarına göre baskın olmayan katsayılar matrisinin kullanılmasıyla yapılan denklem sistemlerinin çözümleri bizi bir sonuca ulaştırabilir veya ulaştırmayabilir. Neyse ki birçok fiziksel sistemin katsayılar matrisi dominanttır ve Gauss-Seidel yöntemiyle çizgisel denklem sistemlerinin aynı anda çözümünde bir sonuca ulaşılır.).

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

$$|\in_a|_1 = \left|\frac{0.50000 - 1}{0.50000}\right| \times 100$$
$$= 100.00\%$$
$$|\in_a|_2 = \left|\frac{4.9000 - 0}{4.9000}\right| \times 100$$
$$= 100.00\%$$
$$|\in_a|_3 = \left|\frac{3.0923 - 1}{3.0923}\right| \times 100$$

$$= 67.662\%$$

The maximum absolute relative approximate error is 100.00%

Iteration #2

$$x_1 = \frac{1 - 3(4.9000) + 5(3.0923)}{12}$$

$$= 0.14679$$

$$x_2 = \frac{28 - (0.14679) - 3(3.0923)}{5}$$

$$= 3.7153$$

$$x_3 = \frac{76 - 3(0.14679) - 7(3.7153)}{13}$$

$$= 3.8118$$

At the end of second iteration, the absolute relative approximate error is

$$|\in_a|_1 = \left|\frac{0.14679 - 0.50000}{0.14679}\right| \times 100$$

$$= 240.61\%$$

$$|\in_a|_2 = \left|\frac{3.7153 - 4.9000}{3.7153}\right| \times 100$$

$$= 31.889\%$$

$$|\in_a|_3 = \left|\frac{3.8118 - 3.0923}{3.8118}\right| \times 100$$

$$= 18.874\%$$

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	$\|\in_a\|_1 \%$	x_2	$\|\in_a\|_2 \%$	x_3	$\|\in_a\|_3 \%$
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	$\left\|\in_a\right\|_1 \%$	x_2	$\left\|\in_a\right\|_2 \%$	x_3	$\left\|\in_a\right\|_3 \%$
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	2.0364×10^5	109.89	–12140	109.92	4.8144×10^5	109.89
6	-2.0579×10^6	109.89	1.2272×10^5	109.89	-4.8653×10^6	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

$$|a_{11}| = |3| = 3 \le |a_{12}| + |a_{13}| = |7| + |13| = 20$$

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

$$x_1 + x_2 + x_3 = 3$$

$$2x_1 + 3x_2 + 4x_3 = 9$$

$$x_1 + 7x_2 + x_3 = 9$$

cannot be rewritten to make the coefficient matrix diagonally dominant.

Example 8 (Chemical Engineering)

A liquid-liquid extraction process conducted in the Electrochemical Materials Laboratory involved the extraction of nickel from the aqueous phase into an organic phase. A typical set of experimental data from the laboratory is given below (Elektrokimyasal Malzemeler Laboratuvarı'nda sıvı-sıvı karışımındaki nikel organik fazı ile nikel su fazı ayrıştırma işlemi yapılmaktadır. Deneyle ilgili olarak Laboratuvardan aşağıdaki veri seti verilmektedir).

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

Assuming g is the amount of Ni in the organic phase and a is the amount of Ni in the aqueous phase, the quadratic interpolant that estimates g is given by (g'nin sıvı organik fazdaki nikelin ve a'nın da sıvı su fazındaki nikel değerlerini göstersin. Bu iki değer arasındaki ilişkinin g(a) şeklinde ikinci dereceden bir polinom olarak aşağıdaki gibi verildiğini kabul edelim)

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

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

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

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

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

as the initial guess and conduct two iterations.

Solution

Rewriting the equations gives

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

Iteration #1

Given the initial guess of the solution vector as

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

we get

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

The absolute relative approximate error for each x_i then is

$$\begin{aligned} \left|\in_a\right|_1 &= \left|\frac{1.3925 - 1}{1.3925}\right| \times 100 \\ &= 28.187\% \end{aligned}$$
$$\begin{aligned} \left|\in_a\right|_2 &= \left|\frac{0.11875 - 1}{0.11875}\right| \times 100 \\ &= 742.11\% \end{aligned}$$
$$\begin{aligned} \left|\in_a\right|_3 &= \left|\frac{-0.88875 - 1}{-0.88875}\right| \times 100 \\ &= 212.52\% \end{aligned}$$

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

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

and the maximum absolute relative approximate error is 742.11%.

Iteration #2

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

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

Now we get

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

$$= 2.3053$$

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

$$= -1.4078$$

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

$$= -4.5245$$

The absolute relative approximate error for each x_i then is

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

$$= 39.596\%$$

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

$$= 108.44\%$$

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

$$= 80.357\%$$

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

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

and the maximum absolute relative approximate error is 108.44%.

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

Iteration	x_1	$\|\in_a\|_1 \%$	x_2	$\|\in_a\|_2 \%$	x_3	$\|\in_a\|_3 \%$
1	1.3925	28.1867	0.11875	742.1053	–0.88875	212.52
2	2.3053	39.5960	–1.4078	108.4353	–4.5245	80.357
3	3.9775	42.041	–4.1340	65.946	–11.396	60.296
4	7.0584	43.649	–9.0877	54.510	–24.262	53.032
5	12.752	44.649	–18.175	49.999	–48.243	49.708
6	23.291	45.249	–34.930	47.967	–92.827	48.030

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

Iteration	x_1	$\left|\in_a\right|_1\%$	x_2	$\left|\in_a\right|_2\%$	x_3	$\left|\in_a\right|_3\%$
32	2.1428×10^8	46.0703	-3.3920×10^8	46.0703	-9.1095×10^8	46.0703

The exact solution vector is

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

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

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

Iteration #1

Given the initial guess of the solution vector as

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

we get

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

The absolute relative approximate error for each x_i then is

$$\left|\in_a\right|_1 = \left|\frac{0.88889-1}{0.88889}\right|\times100$$
$$= 12.5\%$$
$$\left|\in_a\right|_2 = \left|\frac{1.3778-1}{1.3778}\right|\times100$$
$$= 27.419\%$$

$$|\in_a|_3 = \left|\frac{2.2589 - 1}{2.2589}\right| \times 100$$

$$= 55.730\%$$

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

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

and the maximum absolute relative approximate error is 55.730%.

Iteration #2

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

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

Now we get

$$x_1 = \frac{12 - 3 \times 1.3778 - 1 \times 2.2589}{9}$$

$$= 0.62309$$

$$x_2 = \frac{10 - 6.25 \times 0.62309 - 1 \times 2.2589}{2.5}$$

$$= 1.5387$$

$$x_3 = \frac{8.57 - 4 \times 0.62309 - 2 \times 1.5387}{1}$$

$$= 3.0002$$

The absolute relative approximate error for each x_i then is

$$|\in_a|_1 = \left|\frac{0.62309 - 0.88889}{0.62309}\right| \times 100$$

$$= 42.659\%$$

$$|\in_a|_2 = \left|\frac{1.5387 - 1.3778}{1.5387}\right| \times 100$$

$$= 10.460\%$$

$$|\in_a|_3 = \left|\frac{3.0002 - 2.2589}{3.0002}\right| \times 100$$

$$= 24.709\%$$

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

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

and the maximum absolute relative approximate error is 42.659%.

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

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

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

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

This is close to the exact solution vector of

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

The polynomial that passes through the three data points is then

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

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

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

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

Example 9 (Civil Engineering)

To find the maximum stresses in a compound cylinder, the following four simultaneous linear equations need to be solved.

$$\begin{bmatrix} 4.2857\times10^7 & -9.2307\times10^5 & 0 & 0 \\ 4.2857\times10^7 & -5.4619\times10^5 & -4.2857\times10^7 & 5.4619\times10^5 \\ -6.5 & -0.15384 & 6.5 & 0.15384 \\ 0 & 0 & 4.2857\times10^7 & -3.6057\times10^5 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix} = \begin{bmatrix} -7.887\times10^3 \\ 0 \\ 0.007 \\ 0 \end{bmatrix}$$

In the compound cylinder, the inner cylinder has an internal radius of $a = 5"$, and an outer radius $c = 6.5"$, while the outer cylinder has an internal radius of $c = 6.5"$ and an outer radius of $b = 8"$. Given $E = 30\times10^6$ psi, $\nu = 0.3$, and that the hoop stress in the outer cylinder is given by

$$\sigma_\theta = \frac{E}{1-\nu^2}\left[c_3\left(1+\nu\right)+c_4\left(\frac{1-\nu}{r^2}\right)\right],$$

find the stress on the inside radius of the outer cylinder.

Find the values of c_1, c_2, c_3 and c_4 using the Gauss-Seidel Method. Use

$$\begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix} = \begin{bmatrix} -0.005 \\ 0.001 \\ 0.0002 \\ 0.03 \end{bmatrix}$$

as the initial guess and conduct two iterations.

Solution

Rewriting the equations gives

$$c_1 = \frac{-7.887\times10^3 - \left(-9.2307\times10^5\right)c_2 - 0c_3 - 0c_4}{4.2857\times10^7}$$

$$c_2 = \frac{0 - 4.2857\times10^7 c_1 - \left(-4.2857\times10^7\right)c_3 - 5.4619\times10^5 c_4}{-5.4619\times10^5}$$

$$c_3 = \frac{0.007 - (-6.5)c_1 - (-0.15384)c_2 - 0.15384c_4}{6.5}$$

$$c_4 = \frac{0 - 0c_1 - 0c_2 - 4.2857\times10^7 c_3}{-3.6057\times10^5}$$

Iteration #1

Given the initial guess of the solution vector as

$$\begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix} = \begin{bmatrix} -0.005 \\ 0.001 \\ 0.0002 \\ 0.03 \end{bmatrix}$$

we get

$$c_1 = \frac{-7.887\times10^3 - \left(-9.2307\times10^5\right)\times0.001}{4.2857\times10^7}$$
$$= -1.6249\times10^{-4}$$

$$c_2 = \frac{0 - 4.2857\times10^7 \times\left(-1.6249\times10^{-4}\right) - \left(-4.2857\times10^7\right)\times0.0002 - 5.4619\times10^5\times0.03}{-5.4619\times10^5}$$
$$= 1.5569\times10^{-3}$$

$$c_3 = \frac{0.007 - (-6.5)\times\left(-1.6249\times10^{-4}\right) - (-0.15384)\times1.5569\times10^{-3} - 0.15384\times0.03}{6.5}$$
$$= 2.4125\times10^{-4}$$

$$c_4 = \frac{0 - 4.2857 \times 10^7 \times 2.4125 \times 10^{-4}}{-3.6057 \times 10^5}$$
$$= 2.8675 \times 10^{-2}$$

The absolute relative approximate error for each c_i then is

$$\left|\in_a\right|_1 = \left|\frac{-1.6249 \times 10^{-4} - (-0.005)}{-1.6249 \times 10^{-4}}\right| \times 100$$
$$= 2977.1\%$$
$$\left|\in_a\right|_2 = \left|\frac{1.5569 \times 10^{-3} - 0.001}{1.5569 \times 10^{-3}}\right| \times 100$$
$$= 35.770\%$$
$$\left|\in_a\right|_3 = \left|\frac{2.4125 \times 10^{-4} - 0.002}{2.4125 \times 10^{-4}}\right| \times 100$$
$$= 17.098\%$$
$$\left|\in_a\right|_4 = \left|\frac{2.8675 \times 10^{-2} - 0.03}{2.8675 \times 10^{-2}}\right| \times 100$$
$$= 4.6223\%$$

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

$$\begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix} = \begin{bmatrix} -1.6249 \times 10^{-4} \\ 1.5569 \times 10^{-3} \\ 2.4125 \times 10^{-4} \\ 2.8675 \times 10^{-2} \end{bmatrix}$$

and the maximum absolute relative approximate error is 2977.1%.

Iteration #2

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

$$\begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix} = \begin{bmatrix} -1.6249 \times 10^{-4} \\ 1.5569 \times 10^{-3} \\ 2.4125 \times 10^{-4} \\ 2.8675 \times 10^{-2} \end{bmatrix}$$

Now we get

$$c_1 = \frac{-7.887 \times 10^3 - (-9.2307 \times 10^5) \times 1.5569 \times 10^{-3}}{4.2857 \times 10^7}$$
$$= -1.5050 \times 10^{-4}$$
$$c_2 = \frac{\begin{pmatrix} 0 - 4.2857 \times 10^7 \times (-1.5050 \times 10^{-4}) - (-4.2857 \times 10^7) \times 2.4125 \times 10^{-4} \\ -5.4619 \times 10^5 \times 2.8675 \times 10^{-2} \end{pmatrix}}{-5.4619 \times 10^5}$$
$$= -2.0639 \times 10^{-3}$$

$$c_3 = \frac{\begin{pmatrix} 0.007-(-6.5)\times\left(-1.5050\times10^{-4}\right)-(-0.15384)\times -2.0639\times10^{-3} \\ -0.15384\times2.8675\times10^{-2} \end{pmatrix}}{6.5}$$

$$= 1.9892\times10^{-4}$$

$$c_4 = \frac{0-4.2857\times10^{7}\times1.9892\times10^{-4}}{-3.6057\times10^{5}}$$

$$= 2.3643\times10^{-2}$$

The absolute relative approximate error for each c_i then is

$$\left|\in_a\right|_1 = \left|\frac{-1.5050\times10^{-4}-\left(-1.6249\times10^{-4}\right)}{-1.5050\times10^{-4}}\right|\times100$$

$$= 7.9702\%$$

$$\left|\in_a\right|_2 = \left|\frac{-2.0639\times10^{-3}-1.5569\times10^{-3}}{-2.0639\times10^{-3}}\right|\times100$$

$$= 175.44\%$$

$$\left|\in_a\right|_3 = \left|\frac{1.9892\times10^{-4}-2.4125\times10^{-4}}{1.9892\times10^{-4}}\right|\times100$$

$$= 21.281\%$$

$$\left|\in_a\right|_4 = \left|\frac{2.3643\times10^{-2}-2.8675\times10^{-2}}{2.3643\times10^{-2}}\right|\times100$$

$$= 21.281\%$$

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

$$\begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix} = \begin{bmatrix} -1.5050\times10^{-4} \\ -2.0639\times10^{-3} \\ 1.9892\times10^{-4} \\ 2.3643\times10^{-2} \end{bmatrix}$$

and the maximum absolute relative approximate error is 175.44%.

At the end of the second iteration the stress on the inside radius of the outer cylinder is calculated

$$\sigma_\theta = \frac{E}{1-\nu^2}\left[c_3(1+\nu)+c_4\left(\frac{1-\nu}{r^2}\right)\right]$$

$$= \frac{30\times10^6}{1-(0.3)^2}\left[1.9892\times10^{-4}(1+0.3)+2.3643\times10^{-2}\left(\frac{1-0.3}{(6.5)^2}\right)\right]$$

$$= 21439\,\text{psi}$$

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

Iteration	c_1	$\left|\in_a\right|_1\%$	c_2	$\left|\in_a\right|_2\%$
1	-1.6249×10^{-4}	2977.1	1.5569×10^{-3}	35.770
2	-1.5050×10^{-4}	7.9702	-2.0639×10^{-3}	175.44

3	-2.2848×10^{-4}	34.132	-9.8931×10^{-3}	79.138
4	-3.9711×10^{-4}	42.464	-2.8949×10^{-2}	65.826
5	-8.0755×10^{-4}	50.825	-6.9799×10^{-2}	58.524
6	-1.6874×10^{-3}	52.142	-1.7015×10^{-1}	58.978

Iteration	c_3	$\left\|\in_a\right\|_3 \%$	c_4	$\left\|\in_a\right\|_4 \%$
1	2.4125×10^{-4}	17.098	2.8675×10^{-2}	4.6223
2	1.9892×10^{-4}	21.281	2.3643×10^{-2}	21.281
3	5.4716×10^{-5}	263.55	6.5035×10^{-3}	263.55
4	-1.5927×10^{-4}	134.35	-1.8931×10^{-2}	134.35
5	-9.3454×10^{-4}	82.957	-1.1108×10^{-1}	82.957
6	-2.0085×10^{-3}	53.472	-2.3873×10^{-1}	53.472

After six iterations, the absolute relative approximate errors are not decreasing. In fact, conducting more iterations reveals that the absolute relative approximate error does not approach zero or converge to any other number.

Gauss-Seidel Method – More Examples Computer Engineering

Example 1

To infer the surface shape of an object from images taken of a surface from three different directions, one needs to solve the following set of equations.

$$\begin{bmatrix} 0.2425 & 0 & -0.9701 \\ 0 & 0.2425 & -0.9701 \\ -0.2357 & -0.2357 & -0.9428 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 247 \\ 248 \\ 239 \end{bmatrix}$$

The right hand side values are the light intensities from the middle of the images, while the coefficient matrix is dependent on the light source directions with respect to the camera. The unknowns are the incident intensities that will determine the shape of the object.

Find the values of x_1, x_2, and x_3 using the Gauss-Seidel method. Use

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 10 \\ 10 \\ 10 \end{bmatrix}$$

as the initial guess and conduct two iterations.

Solution

Rewriting the equations gives

$$x_1 = \frac{247 - 0x_2 - (-0.9701)x_3}{0.2425}$$

$$x_2 = \frac{248 - 0x_1 - (-0.9701)x_3}{0.2425}$$

$$x_3 = \frac{239-(-0.2357)x_1-(-0.2357)x_2}{-0.9428}$$

Iteration #1

Given the initial guess of the solution vector as

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 10 \\ 10 \\ 10 \end{bmatrix}$$

we get

$$x_1 = \frac{247-0\times 10-(-0.9701)\times 10}{0.2425}$$
$$= 1058.6$$
$$x_2 = \frac{248-0\times 1058.6-(-0.9701)\times 10}{0.2425}$$
$$= 1062.7$$
$$x_3 = \frac{239-(-0.2357)\times 1058.6-(-0.2357)\times 1062.7}{-0.9428}$$
$$= -783.81$$

The absolute relative approximate error for each x_i then is

$$|\in_a|_1 = \left|\frac{1058.6-10}{1058.6}\right|\times 100$$
$$= 99.055\%$$
$$|\in_a|_2 = \left|\frac{1062.7-10}{1062.7}\right|\times 100$$
$$= 99.059\%$$
$$|\in_a|_3 = \left|\frac{-783.81-10}{-783.81}\right|\times 100$$
$$= 101.28\%$$

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

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1058.6 \\ 1062.7 \\ -783.81 \end{bmatrix}$$

and the maximum absolute relative approximate error is 101.28%.

Iteration #2

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

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1058.6 \\ 1062.7 \\ -783.81 \end{bmatrix}$$

Now we get

$$x_1 = \frac{247 - 0 \times 1062 .685 - (-0.9701) \times (-783.8116)}{0.2425}$$

$$= -2117.0$$

$$x_2 = \frac{248 - 0 \times (-2117.0) - (-0.9701) \times (-783.81)}{0.2425}$$

$$= -2112.9$$

$$x_3 = \frac{239 - (-0.2357) \times (-2117.0) - (-0.2357) \times (-2112.9)}{-0.9428}$$

$$= 803.98$$

The absolute relative approximate error for each x_i then is

$$|\in_a|_1 = \left| \frac{(-2117.0) - 1058.6}{-2117.0} \right| \times 100$$

$$= 150.00\%$$

$$|\in_a|_2 = \left| \frac{(-2112.9) - 1062.7}{-2112.9} \right| \times 100$$

$$= 150.30\%$$

$$|\in_a|_3 = \left| \frac{803.98 - (-783.81)}{803.98} \right| \times 100$$

$$= 197.49\%$$

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

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -2117.0 \\ -2112.9 \\ 803.98 \end{bmatrix}$$

and the maximum absolute relative approximate error is 197.49%.

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

Iteration	x_1	$\|\in_a\|_1 \%$	x_2	$\|\in_a\|_2 \%$	x_3	$\|\in_a\|_3 \%$
1	1058.6	99.055	1062.7	99.059	–783.81	101.28
2	–2117.0	150.00	–2112.9	150.295	803.98	197.49
3	4234.8	149.99	4238.9	149.85	–2371.9	133.90
4	–8470.1	150.00	–8466.0	150.07	3980.5	159.59
5	16942	149.99	16946	149.96	–8725.7	145.62
6	–33888	150.00	–33884	150.01	16689	152.28

After six iterations, the absolute relative approximate errors are not decreasing. In fact, conducting more iterations reveals that the absolute relative approximate error does not approach zero but approaches 149.99%.

Gauss-Seidel Method – More Examples Electrical Engineering

Example 1

Three-phase loads are common in AC systems. When the system is balanced the analysis can be simplified to a single equivalent circuit model. However, when it is unbalanced the only practical solution involves the solution of simultaneous linear equations. In one model the following equations need to be solved (üç-fazlı elektrik ile bir AC sistemi yüklenmektedir. Sistem dengeye geldiğinde bir eşdeğer devre modeli üzerinden devre analizi yapılabilir. Devre dengesiz olduğu durumda olsa bile denklem sisteminin pratik çözümü vardır. Bu modelde aşağıdaki denklem sisteminin çözümlenmesi gerekmetedir:).

$$\begin{bmatrix} 0.7460 & -0.4516 & 0.0100 & -0.0080 & 0.0100 & -0.0080 \\ 0.4516 & 0.7460 & 0.0080 & 0.0100 & 0.0080 & 0.0100 \\ 0.0100 & -0.0080 & 0.7787 & -0.5205 & 0.0100 & -0.0080 \\ 0.0080 & 0.0100 & 0.5205 & 0.7787 & 0.0080 & 0.0100 \\ 0.0100 & -0.0080 & 0.0100 & -0.0080 & 0.8080 & -0.6040 \\ 0.0080 & 0.0100 & 0.0080 & 0.0100 & 0.6040 & 0.8080 \end{bmatrix} \begin{bmatrix} I_{ar} \\ I_{ai} \\ I_{br} \\ I_{bi} \\ I_{cr} \\ I_{ci} \end{bmatrix} = \begin{bmatrix} 120 \\ 0.000 \\ -60.00 \\ -103.9 \\ -60.00 \\ 103.9 \end{bmatrix}$$

Find the values of I_{ar}, I_{ai}, I_{br}, I_{bi}, I_{cr}, and I_{ci} using the Gauss-Seidel method. Use

$$\begin{bmatrix} I_{ar} \\ I_{ai} \\ I_{br} \\ I_{bi} \\ I_{cr} \\ I_{ci} \end{bmatrix} = \begin{bmatrix} 20 \\ 20 \\ 20 \\ 20 \\ 20 \\ 20 \end{bmatrix}$$

as the initial guess and conduct two iterations.

Solution

Rewriting the equations gives

$$I_{ar} = \frac{120 - (-0.4516)I_{ai} - 0.0100 I_{br} - (-0.0080)I_{bi} - 0.0100 I_{cr} - (-0.0080)I_{ci}}{0.7460}$$

$$I_{ai} = \frac{0.000 - 0.4516 I_{ar} - 0.0080 I_{br} - 0.0100 I_{bi} - 0.0080 I_{cr} - 0.0100 I_{ci}}{0.7460}$$

$$I_{br} = \frac{-60.00 - 0.0100 I_{ar} - (-0.0080)I_{ai} - (-0.5205)I_{bi} - 0.0100 I_{cr} - (-0.0080)I_{ci}}{0.7787}$$

$$I_{bi} = \frac{-103.9 - 0.0080 I_{ar} - 0.0100 I_{ai} - 0.5205 I_{br} - 0.0080 I_{cr} - 0.0100 I_{ci}}{0.7787}$$

$$I_{cr} = \frac{-60.00 - 0.0100 I_{ar} - (-0.0080)I_{ai} - 0.0100 I_{br} - (-0.0080)I_{bi} - (-0.6040)I_{ci}}{0.8080}$$

$$I_{ci} = \frac{103.9 - 0.0080 I_{ar} - 0.0100 I_{ai} - 0.0080 I_{br} - 0.0100 I_{bi} - 0.6040 I_{cr}}{0.8080}$$

Iteration #1

Given the initial guess of the solution vector as

$$\begin{bmatrix} I_{ar} \\ I_{ai} \\ I_{br} \\ I_{bi} \\ I_{cr} \\ I_{ci} \end{bmatrix} = \begin{bmatrix} 20 \\ 20 \\ 20 \\ 20 \\ 20 \\ 20 \end{bmatrix}$$

Substituting the guess values into the first equation

$$I_{ar} = \frac{120 - (-0.4516) I_{ai} - 0.0100 I_{br} - (-0.0080) I_{bi} - 0.0100 I_{cr} - (-0.0080) I_{ci}}{0.7460}$$

$$= 172.86$$

Substituting the new value of I_{ar} and the remaining guess values into the second equation

$$I_{ai} = \frac{0.000 - 0.4516 I_{ar} - 0.0080 I_{br} - 0.0100 I_{bi} - 0.0080 I_{cr} - 0.0100 I_{ci}}{0.7460}$$

$$= -105.61$$

Substituting the new values of I_{ar}, I_{ai}, and the remaining guess values into the third equation

$$I_{br} = \frac{-60.00 - 0.0100 I_{ar} - (-0.0080) I_{ai} - (-0.5205) I_{bi} - 0.0100 I_{cr} - (-0.0080) I_{ci}}{0.7787}$$

$$= -67.039$$

Substituting the new values of I_{ar}, I_{ai}, I_{br}, and the remaining guess values into the fourth equation

$$I_{bi} = \frac{-103.9 - 0.0080 I_{ar} - 0.0100 I_{ai} - 0.5205 I_{br} - 0.0080 I_{cr} - 0.0100 I_{ci}}{0.7787}$$

$$= -89.499$$

Substituting the new values of I_{ar}, I_{ai}, I_{br}, I_{bi}, and the remaining guess values into the fifth equation

$$I_{cr} = \frac{-60.00 - 0.0100 I_{ar} - (-0.0080) I_{ai} - 0.0100 I_{br} - (-0.0080) I_{bi} - (-0.6040) I_{ci}}{0.8080}$$

$$= -62.548$$

Substituting the new values of I_{ar}, I_{ai}, I_{br}, I_{bi}, I_{cr}, and the remaining guess value into the sixth equation

$$I_{ci} = \frac{103.9 - 0.0080 I_{ar} - 0.0100 I_{ai} - 0.0080 I_{br} - 0.0100 I_{bi} - 0.6040 I_{cr}}{0.8080}$$

$$= 176.71$$

The absolute relative approximate error for each I then is

$$\left|\in_a\right|_1 = \left|\frac{172.86-20}{172.86}\right|\times 100$$
$$= 88.430\%$$
$$\left|\in_a\right|_2 = \left|\frac{-105.61-20}{-105.61}\right|\times 100$$
$$= 118.94\%$$
$$\left|\in_a\right|_3 = \left|\frac{-67.039-20}{-67.039}\right|\times 100$$
$$= 129.83\%$$
$$\left|\in_a\right|_4 = \left|\frac{-89.499-20}{-89.499}\right|\times 100$$
$$= 122.35\%$$
$$\left|\in_a\right|_5 = \left|\frac{-62.548-20}{-62.548}\right|\times 100$$
$$= 131.98\%$$
$$\left|\in_a\right|_6 = \left|\frac{176.71-20}{176.71}\right|\times 100$$
$$= 88.682\%$$

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

$$\begin{bmatrix} I_{ar} \\ I_{ai} \\ I_{br} \\ I_{bi} \\ I_{cr} \\ I_{ci} \end{bmatrix} = \begin{bmatrix} 172.86 \\ -105.61 \\ -67.039 \\ -89.499 \\ -62.548 \\ 176.71 \end{bmatrix}$$

and the maximum absolute relative approximate error is 131.98%.

Iteration #2

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

$$\begin{bmatrix} I_{ar} \\ I_{ai} \\ I_{br} \\ I_{bi} \\ I_{cr} \\ I_{ci} \end{bmatrix} = \begin{bmatrix} 172.86 \\ -105.61 \\ -67.039 \\ -89.499 \\ -62.548 \\ 176.71 \end{bmatrix}$$

Substituting the values from Iteration #1 into the first equation

$$I_{ar} = \frac{120-(-0.4516)I_{ai}-0.0100I_{br}-(-0.0080)I_{bi}-0.0100I_{cr}-(-0.0080)I_{ci}}{0.7460}$$

$$= 99.600$$

Substituting the new value of I_{ar} and the remaining values from Iteration #1 into the second equation

$$I_{ai} = \frac{0.000 - 0.4516 I_{ar} - 0.0080 I_{br} - 0.0100 I_{bi} - 0.0080 I_{cr} - 0.0100 I_{ci}}{0.7460}$$

$$= -60.073$$

Substituting the new values of I_{ar}, I_{ai}, and the remaining values from Iteration #1 into the third equation

$$I_{br} = \frac{-60.00 - 0.0100 I_{ar} - (-0.0080) I_{ai} - (-0.5205) I_{bi} - 0.0100 I_{cr} - (-0.0080) I_{ci}}{0.7787}$$

$$= -136.15$$

Substituting the new values of I_{ar}, I_{ai}, I_{br}, and the remaining values from Iteration #1 into the fourth equation

$$I_{bi} = \frac{-103.9 - 0.0080 I_{ar} - 0.0100 I_{ai} - 0.5205 I_{br} - 0.0080 I_{cr} - 0.0100 I_{ci}}{0.7787}$$

$$= -44.299$$

Substituting the new values of I_{ar}, I_{ai}, I_{br}, I_{bi}, and the remaining values from Iteration #1 into the fifth equation

$$I_{cr} = \frac{-60.00 - 0.0100 I_{ar} - (-0.0080) I_{ai} - 0.0100 I_{br} - (-0.0080) I_{bi} - (-0.6040) I_{ci}}{0.8080}$$

$$= 57.259$$

Substituting the new values of I_{ar}, I_{ai}, I_{br}, I_{bi}, I_{cr}, and the remaining value from Iteration #1 into the sixth equation

$$I_{ci} = \frac{103.9 - 0.0080 I_{ar} - 0.0100 I_{ai} - 0.0080 I_{br} - 0.0100 I_{bi} - 0.6040 I_{cr}}{0.8080}$$

$$= 87.441$$

The absolute relative approximate error for each I then is

$$\left|\in_a\right|_1 = \left|\frac{99.600 - 172.86}{99.600}\right| \times 100$$

$$= 73.552\%$$

$$\left|\in_a\right|_2 = \left|\frac{-60.073 - (-105.61)}{-60.073}\right| \times 100$$

$$= 75.796\%$$

$$\left|\in_a\right|_3 = \left|\frac{-136.35 - (-67.039)}{-136.35}\right| \times 100$$

$$= 50.762\%$$

$$\left|\in_a\right|_4 = \left|\frac{-44.299 - (-89.499)}{-44.299}\right| \times 100$$

$$= 102.03\%$$

$$\left|\in_a\right|_5 = \left|\frac{57.259-\left(-62.548\right)}{57.259}\right|\times 100$$
$$= 209.24\%$$
$$\left|\in_a\right|_6 = \left|\frac{87.441-176.71}{87.441}\right|\times 100$$
$$= 102.09\%$$

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

$$\begin{bmatrix} I_{ar} \\ I_{ai} \\ I_{br} \\ I_{bi} \\ I_{cr} \\ I_{ci} \end{bmatrix} = \begin{bmatrix} 99.600 \\ -60.073 \\ -136.15 \\ -44.299 \\ 57.259 \\ 87.441 \end{bmatrix}$$

and the maximum absolute relative approximate error is 141.4087%.
Conducting more iterations gives the following values for the solution vector and the corresponding absolute relative approximate errors.

Iteration	I_{ar}	I_{ai}	I_{br}	I_{bi}	I_{cr}	I_{ci}
1	172.86	–105.61	–67.039	–89.499	–62.548	176.71
2	99.600	–60.073	–136.15	–44.299	57.259	87.441
3	126.01	–76.015	–108.90	–62.667	–10.478	137.97
4	117.25	–70.707	–119.62	–55.432	27.658	109.45
5	119.87	–72.301	–115.62	–58.141	6.2513	125.49
6	119.28	–71.936	–116.98	–57.216	18.241	116.53

Iteration	$\left\|\in_a\right\|_1\%$	$\left\|\in_a\right\|_2\%$	$\left\|\in_a\right\|_3\%$	$\left\|\in_a\right\|_4\%$	$\left\|\in_a\right\|_5\%$	$\left\|\in_a\right\|_6\%$
1	88.430	118.94	129.83	122.35	131.98	88.682
2	73.552	75.796	50.762	102.03	209.24	102.09
3	20.960	20.972	25.027	29.311	646.45	36.623
4	7.4738	7.5067	8.9631	13.053	137.89	26.001
5	2.1840	2.2048	3.4633	4.6595	342.43	12.742
6	0.49408	0.50789	1.1629	1.6170	65.729	7.6884

After six iterations, the absolute relative approximate errors are decreasing, but are still high. Allowing for more iteration, the relative approximate errors decrease significantly.

Iteration	I_{ar}	I_{ai}	I_{br}	I_{bi}	I_{cr}	I_{ci}
32	119.33	–71.973	–116.66	–57.432	13.940	119.74
33	119.33	–71.973	–116.66	–57.432	13.940	119.74

Iteration	$\left\|\in_a\right\|_1 \%$	$\left\|\in_a\right\|_2 \%$	$\left\|\in_a\right\|_3 \%$	$\left\|\in_a\right\|_4 \%$	$\left\|\in_a\right\|_5 \%$	$\left\|\in_a\right\|_6 \%$
32	3.0666×10^{-7}	3.0047×10^{-7}	4.2389×10^{-7}	5.7116×10^{-7}	2.0941×10^{-5}	1.8238×10^{-6}
33	1.7062×10^{-7}	1.6718×10^{-7}	2.3601×10^{-7}	3.1801×10^{-7}	1.1647×10^{-5}	1.0144×10^{-6}

After 33 iterations, the solution vector is

$$\begin{bmatrix} I_{ar} \\ I_{ai} \\ I_{br} \\ I_{bi} \\ I_{cr} \\ I_{ci} \end{bmatrix} = \begin{bmatrix} 119.33 \\ -71.973 \\ -116.66 \\ -57.432 \\ 13.940 \\ 119.74 \end{bmatrix}$$

The maximum absolute relative approximate error is $1.1647\times10^{-5}\%$.

Gauss-Seidel Method – More Examples Industrial Engineering

Example 1

To find the number of toys a company should manufacture per day to optimally use their injection-molding machine and the assembly line, one needs to solve the following set of equations. The unknowns are the number of toys for boys, x_1, the number of toys for girls, x_2, and the number of unisexual toys, x_3.

$$\begin{bmatrix} 0.3333 & 0.1667 & 0.6667 \\ 0.1667 & 0.6667 & 0.3333 \\ 1.05 & -1.00 & 0.00 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 756 \\ 1260 \\ 0 \end{bmatrix}$$

Find the values of x_1, x_2, and x_3 using the Gauss-Seidel method. Use

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1000 \\ 1000 \\ 1000 \end{bmatrix}$$

as the initial guess and conduct two iterations.

Solution

Rewriting the equations gives

$$x_1 = \frac{756 - 0.1667x_2 - 0.6667x_3}{0.3333}$$

$$x_2 = \frac{1260 - 0.1667x_1 - 0.3333x_3}{0.6667}$$

$$x_3 = \frac{0 - 1.05x_1 - (-1.00)x_2}{0}$$

The equation for x_3 is divided by 0 which is undefined. Therefore the order of the equations will need to be changed. Equation 3 and Equation 1 will be switched. By switching Equations 3 and 1, the matrix will also become diagonally dominant.

The system of equations becomes

$$\begin{bmatrix} 1.05 & -1.00 & 0.00 \\ 0.1667 & 0.6667 & 0.3333 \\ 0.3333 & 0.1667 & 0.6667 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1260 \\ 756 \end{bmatrix}$$

Rewriting the equations gives

$$x_1 = \frac{0-(-1.00)x_2 - 0x_3}{1.05}$$

$$x_2 = \frac{1260 - 0.1667x_1 - 0.3333x_3}{0.6667}$$

$$x_3 = \frac{756 - 0.3333x_1 - 0.1667x_2}{0.6667}$$

Iteration #1

Given the initial guess of the solution vector as

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1000 \\ 1000 \\ 100 \end{bmatrix}$$

we get

$$x_1 = \frac{0-(-1.00)\times 1000 - 0 \times 100}{1.05}$$

$$= 952.38$$

$$x_2 = \frac{1260 - 0.1667 \times 952.38 - 0.3333 \times 100}{0.6667}$$

$$= 1601.8$$

$$x_3 = \frac{756 - 0.3333 \times 952.38 - 0.1667 \times 1601.8}{0.6667}$$

$$= 257.32$$

The absolute relative approximate error for each x_i then is

$$|\in_a|_1 = \left| \frac{952.38 - 1000}{952.38} \right| \times 100$$

$$= 5\%$$

$$|\in_a|_2 = \left| \frac{1601.8 - 1000}{1601.8} \right| \times 100$$

$$= 37.570\%$$

$$|\in_a|_3 = \left| \frac{257.32 - 100}{257.32} \right| \times 100$$

$$= 61.138$$

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

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 952.38 \\ 1601.8 \\ 257.32 \end{bmatrix}$$

and the maximum absolute relative approximate error is 61.138%.

Iteration #2

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

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 952.38 \\ 1601.8 \\ 257.32 \end{bmatrix}$$

Now we get

$$x_1 = \frac{0-(-1.00)\times 1601.8 - 0\times 257.32}{1.05}$$

$$= 1525.5$$

$$x_2 = \frac{1260 - 0.1667\times 1525.5 - 0.3333\times 257.32}{0.6667}$$

$$= 1379.8$$

$$x_3 = \frac{756 - 0.3333\times 1525.5 - 0.1667\times 1379.8}{0.6667}$$

$$= 26.295$$

The absolute relative approximate error for each x_i then is

$$|\in_a|_1 = \left|\frac{1525.5 - 952.38}{1525.5}\right| \times 100$$

$$= 37.570\%$$

$$|\in_a|_2 = \left|\frac{1379.8 - 1601.8}{1379.8}\right| \times 100$$

$$= 16.085\%$$

$$|\in_a|_3 = \left|\frac{26.295 - 257.32}{26.295}\right| \times 100$$

$$= 878.59\%$$

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

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1525.5 \\ 1379.8 \\ 26.295 \end{bmatrix}$$

and the maximum absolute relative approximate error is 878.59%.

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

Iteration	x_1	$\lvert\in_a\rvert_1 \%$	x_2	$\lvert\in_a\rvert_2 \%$	x_3	$\lvert\in_a\rvert_3 \%$

1	952.38	5	1601.8	37.570	257.32	61.138
2	1525.5	37.570	1379.8	16.085	26.295	878.59
3	1314.1	16.085	1548.2	10.874	89.876	70.743
4	1474.5	10.874	1476.3	4.8686	27.694	224.53
5	1406.0	4.8686	1524.5	3.1618	49.863	44.459
6	1451.9	3.1618	1501.9	1.5021	32.554	53.170

After six iterations, the absolute relative approximate errors are decreasing, but they are still high. Allowing for more iterations, the absolute relative approximate errors decrease significantly.

Iteration	x1	$\left\|\in_a\right\|_1 \%$	x2	$\left\|\in_a\right\|_2 \%$	x3	$\left\|\in_a\right\|_3 \%$
20	1439.8	0.00064276	1511.8	0.00034987	36.115	0.0091495
21	1439.8	0.00034987	1511.8	0.00019257	36.114	0.0049578

This is close to the exact solution vector of

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1439.8 \\ 1511.8 \\ 36.113 \end{bmatrix}$$

SIMULTANEOUS LINEAR EQUATIONS	
Topic	Gauss-Seidel Method – More Examples
Summary	Examples of the Gauss-Seidel method
Major	Civil Engineering
Authors	Autar Kaw
Date	Aralık 8, 2016
Web Site	http://numericalmethods.eng.usf.edu

Key Terms:

Gauss-Seidel method

Convergence of Gauss-Seidel method

Diagonally dominant matrix

4.8.1 Multiple-Choice Test Chapter 04.08 Gauss-Seidel Method

1. A square matrix $[A]_{n\times n}$ is diagonally dominant if

A) $|a_{ii}| \geq \sum_{\substack{j=1 \\ i\neq j}}^{n} |a_{ij}|, \ i = 1,2,...,n$

(B) $|a_{ii}| \geq \sum_{\substack{j=1 \\ i\neq j}}^{n} |a_{ij}|, \ i = 1,2,...,n$ and $|a_{ii}| > \sum_{\substack{j=1 \\ i\neq j}}^{n} |a_{ij}|$, for any $i = 1,2,...,n$

(C) $\left|a_{ii}\right| \geq \sum_{j=1}^{n}\left|a_{ij}\right|,\ i=1,2,\ldots,n$ and $\left|a_{ii}\right| > \sum_{j=1}^{n}\left|a_{ij}\right|,$ for any $i=1,2,\ldots,n$

(D) $\left|a_{ii}\right| \geq \sum_{j=1}^{n}\left|a_{ij}\right|,\ i=1,2,\ldots,n$

2. Using $[x_1, x_2, x_3] = [1,3,5]$ as the initial guess, the values of $[x_1, x_2, x_3]$ after three iterations in the Gauss-Seidel method for

$$\begin{bmatrix} 12 & 7 & 3 \\ 1 & 5 & 1 \\ 2 & 7 & -11 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ -5 \\ 6 \end{bmatrix}$$

are

(A) [-2.8333 -1.4333 -1.9727] (B) [1.4959 -0.90464 -0.84914]
(C)[0.90666 -1.0115 -1.0243] (D) [1.2148 -0.72060 -0.82451]

3. To ensure that the following system of equations,

$$\begin{aligned} 2x_1 + 7x_2 - 11x_3 &= 6 \\ x_1 + 2x_2 + x_3 &= -5 \\ 7x_1 + 5x_2 + 2x_3 &= 17 \end{aligned}$$

converges using the Gauss-Seidel method, one can rewrite the above equations as follows:

(A) $\begin{bmatrix} 2 & 7 & -11 \\ 1 & 2 & 1 \\ 7 & 5 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 6 \\ -5 \\ 17 \end{bmatrix}$ (B) $\begin{bmatrix} 7 & 5 & 2 \\ 1 & 2 & 1 \\ 2 & 7 & -11 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 17 \\ -5 \\ 6 \end{bmatrix}$ (C) $\begin{bmatrix} 7 & 5 & 2 \\ 1 & 2 & 1 \\ 2 & 7 & -11 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 6 \\ -5 \\ 17 \end{bmatrix}$

(D) The equations cannot be rewritten in a form to ensure convergence.

4. For $\begin{bmatrix} 12 & 7 & 3 \\ 1 & 5 & 1 \\ 2 & 7 & -11 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 22 \\ 7 \\ -2 \end{bmatrix}$ and using $\begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 1 \end{bmatrix}$ as the initial guess, the values of $\begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}$ are found at the end of each iteration as

Iteration #	x_1	x_2	x_3
1	0.41667	1.1167	0.96818
2	0.93990	1.0184	1.0008
3	0.98908	1.0020	0.99931
4	0.99899	1.0003	1.0000

At what first iteration number would you trust at least 1 significant digit in your solution?

(A) 1 (B) 2 (C) 3 (D)4

5. The algorithm for the Gauss-Seidel method to solve [A][X]=[C] is given as follows when using nmax iterations. The initial value of [X] is stored in [X].

```
(A)   Sub Seidel(n,a,x,rhs,nmax)
      For k=1 To nmax
      For i=1 To n
      For j=1 To n
      If (i<>j) Then
      Sum = Sum + a(i,j)*x(j)
      endif
      Next j
      x(i)=(rhs(i)-Sum)/a(i,i)
      Next i
      Next j
      End Sub
```

```
(B)   Sub Seidel(n,a,x,rhs,nmax)
      For k = 1 To nmax
      For i = 1 To n
      Sum = 0
      For j=1 To n
      If (i<>j) Then
      Sum = Sum + a(i,j)*x(j)
      endif
      Next j
      x(i)=(rhs(i)-Sum)/a(i,i)
      Next i
      Next k
      End Sub
```

```
(D)   Sub Seidel(n,a,x,rhs,nmax)
      For k=1To nmax
      For i=1 To n
      Sum = 0
      For j=1 To n
      If (i<>j) Then
      Sum = Sum + a(i,j)*x(j)
      endif
      Next j
      x(i)=(rhs(i)-Sum)/a(i,i)
      Next i
      Next k
      End Sub
```

6. Thermistors measure temperature, have a nonlinear output and are valued for a limited range. So when a thermistor is manufactured, the manufacturer supplies a resistance vs. temperature curve. An accurate representation of the curve is generally given by (termistörler sıcaklık ölçerler, çizgisel olmayan bir çıkış verirler ve sınırlı aralıktaki değerleri okuyabilirler. Bir termistör üretildiğinde üretici dirence bağlı olarak sıcaklık grafiğini de vermektedir. Eğrinin doğru bir şekilde fonksiyonu aşağıdaki gibi verilmektedir:)

$$\frac{1}{T}=a_0+a_1\ln(R)+a_2\{\ln(R)\}^2+a_3\{\ln(R)\}^3$$

where T is temperature in Kelvin, R is resistance in ohms, and a_0, a_1, a_2, a_3 are constants of the calibration curve. Given the following for a thermistor (denklemdeki T Kelvin cinsinden sıcaklık değerlerini, R Ohm biriminde direnci temsil etmekte ve a_0, a_1, a_2 ve a_3 kalibrasyon grafiği sabitlerini göstermektedir. Aşağıdaki çizelgeyi kullanarak direnci 900 Ohm olarak ölçülen termistörün ölçtüğü sıcaklık aşağıdakilerden hangisi olabilir:)

R	T
ohm	°C
1101.0	25.113
911.3	30.131
636.0	40.120

451.1	50.128

the value of temperature in °C for a measured resistance of 900 ohms most nearly is

(A) 30.002 (B) 30.473 (C) 31.272 (D) 31.445

For a complete solution, refer to the links at the end of the book.

5 Chapter 05.01 Background of Interpolation

5.1 Chapter 05.01 Background of Interpolation

PRE-REQUISITES
1. Know simple co-ordinate geometry and graphing.

OBJECTIVES
1. Understand what Interpolation is.

Definition of Interpolation

After reading this chapter, you should be able to:
1. Understand what Interpolation is.

What is Interpolation?
Many a times, a function y=f(x) is given only at discrete points such as (x_0,y_0), (x_1,y_1), (x_2,y_2),... (x_{n-1},y_{n-1}), (x_n,y_n). How 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. 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.

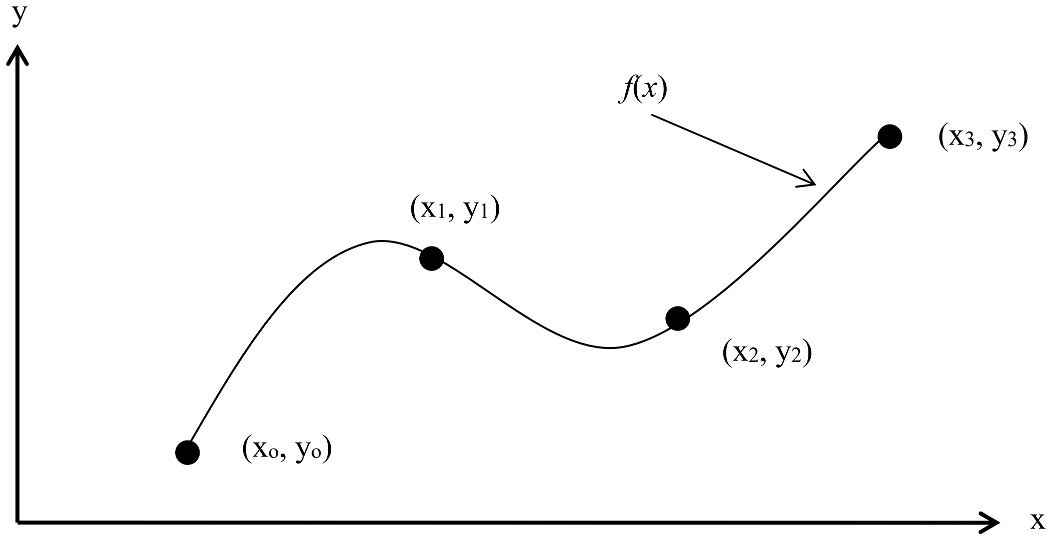


Figure 1 Interpolation of discrete data (kesikli verilerin interpolasyonu)

INTERPOLATION	
Topic	Definition of Interpolation
Summary	Textbook notes on the definition of interpolation, with graph.
Major	All Majors of Engineering
Authors	Autar Kaw
Last Revised	Aralık 8, 2016

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 (1913-1923 yılları arasında Edinburg Üniversitesi Sayısal Matematik bölümünde çalışan profesörü Sir Edmund Whittaker interpolasyonu "bir çizelgede olmayan veriye ihtiyaç duyulması" şeklinde açıklamıştır. Tarihte interpolasyonun çok değişik şekilde kullanıldığı görülmektedir. Antik Babil'de ve Yunan'da interpolasyonun kullanıldığı bilinmektedir. MÖ 300'lü yıllarda sadece çizgisel değil çizgisel olmayan formlara sahip interpolasyon kullanılarak güneşin, ayın ve gezegenlerin konumlarının belirlendiğini biliyoruz. Çiftçiler tohum ekme zamanlarını interpolasyon ile belirliyorlardı. MÖ 150'li yıllarda Rodoslu Hipparchus gök cisimlerinin konularını belirlemek için kullanılan sinüsel fonksiyonlar gibi çizgisel interpolasyonla "akor fonksiyonu" 'nu hesaplıyordu.).

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 (MS 600'lü yıllarda daha doğuda Çinlilerin interpolasyonu kullandığı biliniyordu. Liu Zhuo ikinci dereceden Gregory-Newton interpolasyonu ile "İmparatorluk Standart Takvimi"'ni oluşturmuştu. MS 650 yıllarda Hindli astronom ve matematikçi Brahmagupta ikinci dereceden sinüs interpolasyon formülünü kullanmış ve eşit aralıklı olmayan verilerin interpolasyonunu geliştirmiştir.).

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 (topraklarla ilgili birçok yerde interpolasyon kullanılmış, fakat denizlerdeki kullanımı ise daha önemlidir. Özel fonksiyonların değerleri sayısal yöntemler kullanılarak çizelgeler halinde hazırlanmış, denizciler tarafından enlem ve boylam değerlerini belirlemek için kullanılmıştır. Metrik sisteme geçildikten sonra fransız hükümeti bu çizelgelerin hazırlanmasını sağlamıştır. Matematikçilerden bu tür çizelgeler hazırlanması istenmiştir. Ancak projelerde çalışanların matematikçilerin çoğu giyotinle başlarını kaybetmişlerdi.).

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. (bu çizelgelerle ilgili en talihsiz durum bunların çoğunn başkalarının yaptıklarından aşırma olmalarıydı. Bilgisayarlar yani bu zor çizelgeleri elle hazırlayanların bilgileri başka kaynaklardan aşırmaları daha fazla hataya neden olmuştur. Charles Babbage bu problemi delikli kartlarla programlanmış/kullanan mekanik bilgisayar olan "fark makinesi" ile çözmeye çalışmıştır. Ayrıca Babbage yaptığı makineyi para kazanmak için at yarışı tahminlerinde kullanmıştır. Sadece makinesine fon bulmakta zorlanmamış hayatı boyunca kıt kanaat geçinmiş ve buluşunun tamamlanmasını göremeden ölmüştür. Ölümünden 125 yıl sonra nano teknoloji çağında Babbage modern bilgisayarların atası kabul edilmektedir.)

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 Abromowitz 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. (HNMI'da Newton ve Lagrange yöntemleri olmak üzere kredisi olan 2 interpolasyon yöntemi öğretilmektedir. Newton 1675'te "klasik interpolasyon teorisinin oluşturulması" çalışmasını yayınlamış, 1795'te Lagrange kendi ismiyle anılan interpolasyon yöntemini geliştirmiştir.)

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	Aralık 8, 2016
Web Site	http://numericalmethods.eng.usf.edu

5.2 Chapter 05.02 Direct Method of Interpolation (Doğrudan yöntemli interpolasyon)

PRE-REQUISITES

1. Introduction to matrix algebra – setting up equations in matrix form (Primer for Matrix Algebra).
2. Solving a set of simultaneous linear equations by methods such as Gauss elimination. (Gaussian elimination)

OBJECTIVES

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.

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? (interpolasyon nedir?)
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* (çoğu zaman veriler noktalar şeklinde (x_0, y_0), (x_1, y_1), ...,(x_{n-1}, y_{n-1}), (x_n, y_n) verilir. Herhangi bir x noktasına ait y değeri nasıl hesaplanabilir? Sürekli bir f(x) fonksiyonu n+1 veriden geçen bir fonksiyonu temsil etsin (Figure 1). Fonksiyon belirlendikten sonra herhangi bir x noktasındaki y değeri hesaplanabilir. Buna interpolasyon denir.).

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* (tabi x değeri sınır değerlerinin dışında olabilir. Bu durumda yapılacak işleme ekstrapolasyon denir).

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 (nasıl bir *f(x)* fonksiyonu seçilmelidir? Bir polinom bu seçim için aşağıdaki nedenlerden dolayı iyi olabilir:)

(A) evaluate, (geliştirilebilir)

(B) differentiate, and (türevlenebilir)

(C) integrate (integre edilebilir)

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.

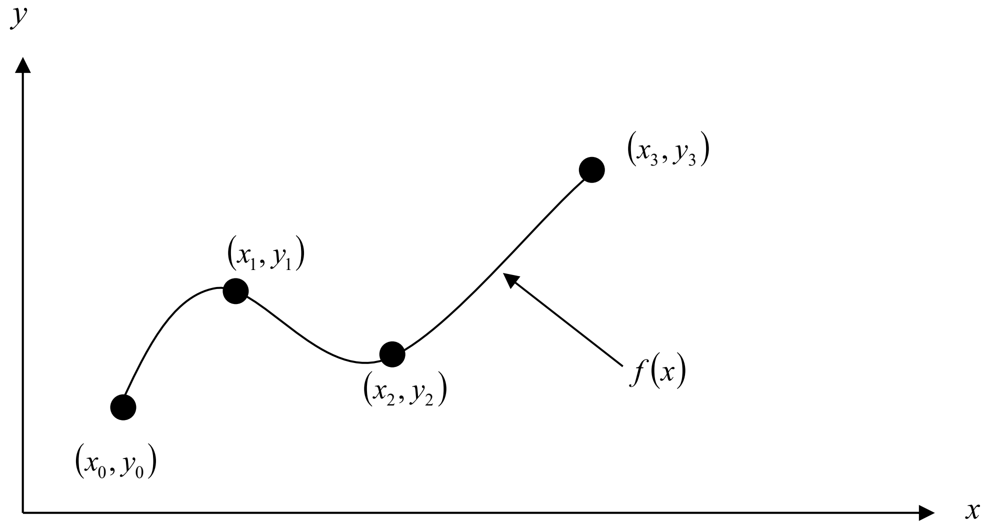


Figure 1 Interpolation of discrete data.

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 (doğrudan interpolasyon yöntemi aşağıdaki denklemdeki gibi verilebilir. *n+1* noktadan geçen polinomun derecesi n olmalıdır:)

$$y=a_0+a_1x+...+a_nx^n \qquad (1)$$

through the data, where $a_0, a_1,..., a_n$ 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, $a_0, a_1,..., a_n$ 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. (Polinomun veri noktalarından geçmesini sağlayan $a_0, a_1,..., a_n$ nicelikleri $n+1$ tane gerçel sabiti tanımlar ve seçilen noktalar arasında değişmezler. $n+1$ tane x değerinden ve $n+1$ tane y değerinden $n+1$ tane eşitlik yazılabilir. Böylece $a_0, a_1,..., a_n$ nicelikleri n+1 tane çizgisel denklem kullanılarak çözülebilir. Çözümleme işleminden sonra herhangi bir x değeri için y hesaplanabilir.)

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. (bütün veri noktalarının kullanılmasına gerek yoktur. Bu durumda polinomun derecesi nasıl seçilmeli ve bu polinomun için hangi veri noktaları kullanılmalıdır? Bu kavramlar ve doğrudan interpolasyon polinom yöntemi aşağıda örneklerle anlatılmaktadır.)

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

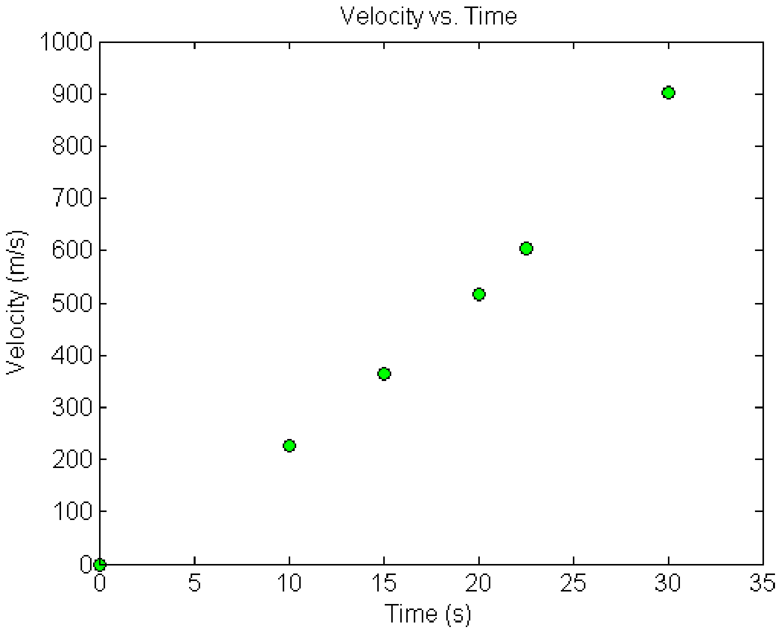


Figure 2 Graph of velocity vs. time data for the rocket example (roket örneğinin zamana bağlı hız grafiği verileri.).

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

$v(t)=a_0+a_1t$

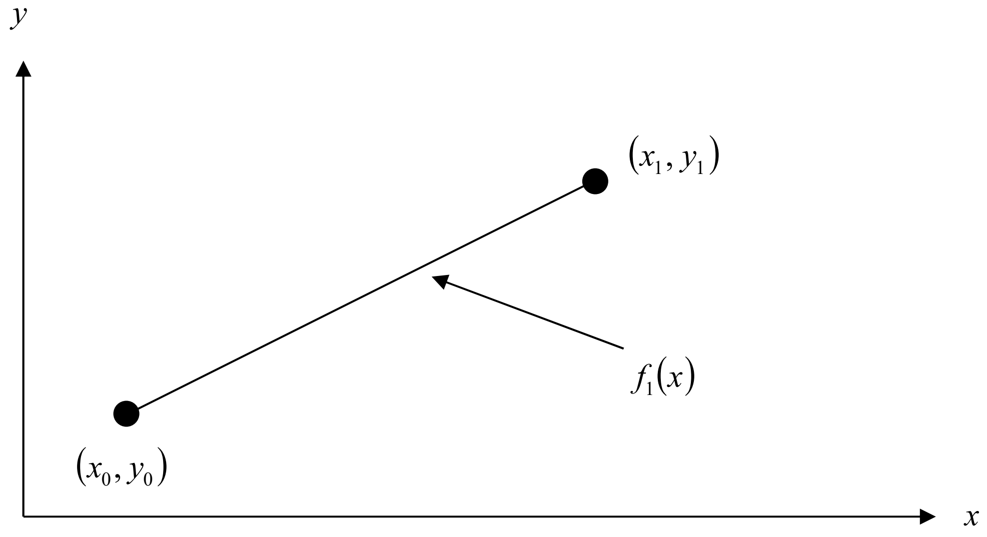


Figure 3 Linear interpolation.

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

$t_0=15,\ v(t_0)=362.78$

$t_1=20,\ v(t_1)=517.35$

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

$v(t)=a_0+a_1(t)$

$=-100.93+30.914t,\ 15 \le t \le 20$

At $t = 16$,

$$v(16) = -100.92 + 30.914 \times 16$$

$=393.7$ m/s

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

$$v(t) = a_0 + a_1 t + a_2 t^2$$

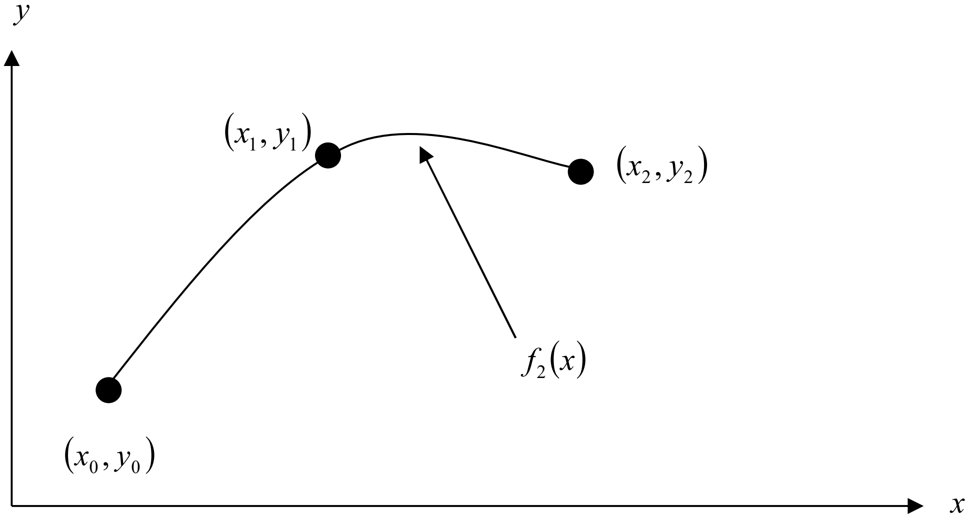


Figure 4 Quadratic interpolation.

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

$$t_0 = 10,\ v(t_0) = 227.04$$
$$t_1 = 15,\ v(t_1) = 362.78$$
$$t_2 = 20,\ v(t_2) = 517.35$$

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

$$v(t) = 12.05 + 17.733t + 0.3766t^2, \ 10 \le t \le 20$$

At $t = 16$,

$$v(16) = 12.05 + 17.733(16) + 0.3766(16)^2$$
$$= 392.19 \text{ m/s}$$

The absolute relative approximate error $|\in_a|$ obtained between the results from the first and second order polynomial is

$$|\in_a| = \left|\frac{392.19 - 393.70}{392.19}\right| \times 100$$
$$= 0.38410\%$$

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	0
10.0	227.04
15.0	362.78
20.0	517.35
22.5	602.97
30.0	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 $t = 11\text{s}$ to $t = 16\text{s}$.

d) Using the third order polynomial interpolant for velocity from part (a), find the acceleration of the rocket at $t = 16\text{s}$.

Solution

a) For third order polynomial interpolation (also called cubic interpolation), we choose the velocity given by

$$v(t)=a_0+a_1t+a_2t^2+a_3t^3$$

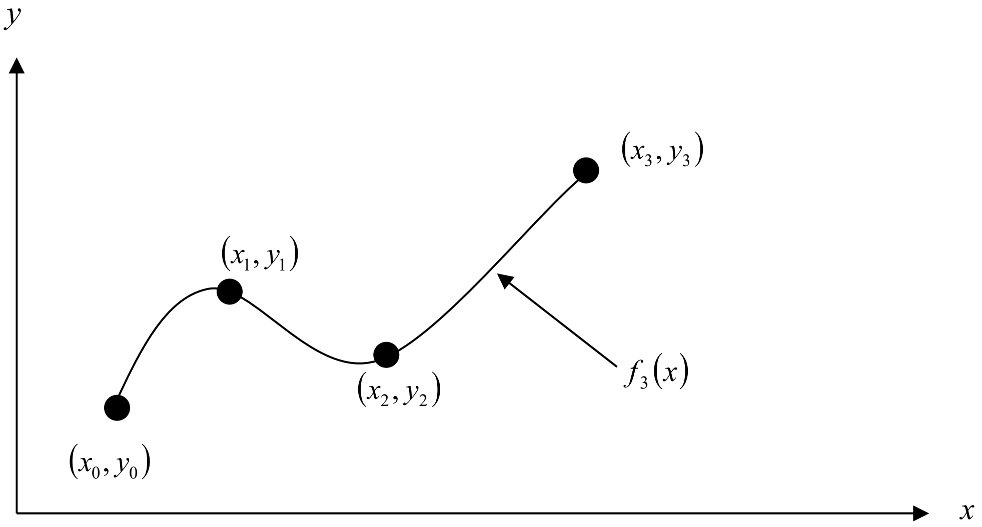


Figure 5 Cubic interpolation.

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

$$t_0 = 10, \quad v(t_0) = 227.04$$
$$t_1 = 15, \quad v(t_1) = 362.78$$
$$t_2 = 20, \quad v(t_2) = 517.35$$
$$t_3 = 22.5, \quad v(t_3) = 602.97$$

gives

$$v(10) = a_0 + a_1(10) + a_2(10)^2 + a_3(10)^3 = 227.04$$
$$v(15) = a_0 + a_1(15) + a_2(15)^2 + a_3(15)^3 = 362.78$$
$$v(20) = a_0 + a_1(20) + a_2(20)^2 + a_3(20)^3 = 517.35$$
$$v(22.5) = a_0 + a_1(22.5) + a_2(22.5)^2 + a_3(22.5)^3 = 602.97$$

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

$$\begin{aligned} v(t) &= a_0 + a_1 t + a_2 t^2 + a_3 t^3 \\ &= -4.2540 + 21.266t + 0.13204t^2 + 0.0054347t^3, \quad 10 \le t \le 22.5 \end{aligned}$$

$$\begin{aligned} v(16) &= -4.2540 + 21.266(16) + 0.13204(16)^2 + 0.0054347(16)^3 \\ &= 392.06 \text{ m/s} \end{aligned}$$

b) The absolute percentage relative approximate error $|\in_a|$ for the value obtained for $v(16)$ between second and third order polynomial is

$$\begin{aligned} |\in_a| &= \left| \frac{392.06 - 392.19}{392.06} \right| \times 100 \\ &= 0.033269\% \end{aligned}$$

c) The distance covered by the rocket between $t = 11\text{s}$ and $t = 16\text{s}$ can be calculated from the interpolating polynomial

$$v(t) = -4.2540 + 21.266t + 0.13204t^2 + 0.0054347t^3, \quad 10 \le t \le 22.5$$

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

$$v(t) = -4.2540 + 21.266t + 0.13204t^2 + 0.0054347t^3, \quad 10 \le t \le 22.5$$

$$
\begin{aligned}
a(t) &= \frac{d}{dt}v(t) \\
&= \frac{d}{dt}\left(-4.2540 + 21.266t + 0.13204t^2 + 0.0054347t^3\right) \\
&= 21.266 + 0.26408t + 0.016304t^2, \quad 10 \le t \le 22.5
\end{aligned}
$$

$$
\begin{aligned}
a(16) &= 21.266 + 0.26408(16) + 0.016304(16)^2 \\
&= 29.665\,\text{m/s}^2
\end{aligned}
$$

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	Aralık 8, 2016
Web Site	http://numericalmethods.eng.usf.edu

5.2.1 Multiple-Choice Test Chapter 05.01 Background on Interpolation

1. The number of polynomials that can go through two fixed data points (x_1, y_1) and (x_2, y_2) is
   (A) 0 (B) 1 (C) 2 (D) infinite

2. A unique polynomial of degree ________________ passes through $n+1$ data points.
   (A) *n*+*1* (B) *n*+*1* or less (C) *n* (D) *n* or less

3. The following function(s) can be used for interpolation:
   (A) polynomial (B) exponential (C) trigonometric (D) all of the above

4. Polynomials are the most commonly used functions for interpolation because they are easy to
   (A) evaluate (B) differentiate (C) integrate (D) evaluate, differentiate and integrate

5. Given *n*+*1* data points $(x_0, y_0), (x_1, y_1), \ldots\ldots, (x_{n-1}, y_{n-1}), (x_n, y_n)$, assume you pass a function $f(x)$ through all the data points. If now the value of the function $f(x)$ is required to be found outside the range of the given x-data, the procedure is called
   (A) extrapolation (B) interpolation (C) guessing (D) regression

6. Given three data points (1,6), (3,28), and (10, 231), it is found that the function $y = 2x^2 + 3x + 1$ passes through the three data points. Your estimate of y at $x = 2$ is most nearly
   (A) 6 (B) 15 (C) 17 (D) 28

7. A unique polynomial of degree ______________ passes through *n*+1 data points.
   (A) *n*+1 (B) *n*+1 or less (C) *n* (D) *n* or less

8. The following data of the velocity of a body is given as a function of time.

Time (s)	0	15	18	22	24
Velocity (m/s)	22	24	37	25	123

The velocity in m/s at 16 s using linear polynomial interpolation is most nearly

(A) 27.867 (B) 28.333 (C) 30.429 (D) 43.000

9. The following data of the velocity of a body is given as a function of time.

Time (s)	0	15	18	22	24
Velocity (m/s)	22	24	37	25	123

Using quadratic interpolation, the interpolant $v(t)$=8.6667t^2–349.67t+3523, 18≤t≤24 approximates the velocity of the body. From this information, the time in seconds at which the velocity of the body is 35 m/s during the above time interval of t=18 s to t=24 s is

(A) 18.667 (B) 20.850 (C) 22.200 (D) 22.294

10. The following data of the velocity of a body is given as a function of time.

Time (s)	0	15	18	22	24
Velocity (m/s)	22	24	37	25	123

One of the interpolant approximations for the velocity from the above data is given as $v(t)$=8.6667t^2–349.67t+3523, 18≤t≤24. Using the above interpolant, the distance in meters covered by the body between $t = 19\text{ s}$ and $t = 22\text{ s}$ is most nearly

(A) 10.337 (B) 88.500 (C) 93.000 (D) 168.00

11. The following data of the velocity of a body is given as a function of time.

Time (s)	0	15	18	22	24
Velocity (m/s)	22	24	37	25	123

If you were going to use quadratic interpolation to find the value of the velocity at $t = 14.9$ seconds, what three data points of time would you choose for interpolation?

(A) 0, 15, 18 (B) 15, 18, 22 (C) 0, 15, 22 (D) 0, 18, 24

Example 1 Direct Method of Interpolation – More Examples Chemical Engineering

To find how much heat is required to bring a kettle of water to its boiling point, you are asked to calculate the specific heat of water at $61°\text{C}$. The specific heat of water is given as a function of time in Table 1.

Table 1 Specific heat of water as a function of temperature.

Temperature, T $(°\text{C})$	Specific heat, C_p $\left(\frac{\text{J}}{\text{kg}-°\text{C}}\right)$
22	4181
42	4179
52	4186
82	4199
100	4217

Determine the value of the specific heat at $T = 61°\text{C}$ using the direct method of interpolation and a first order polynomial.

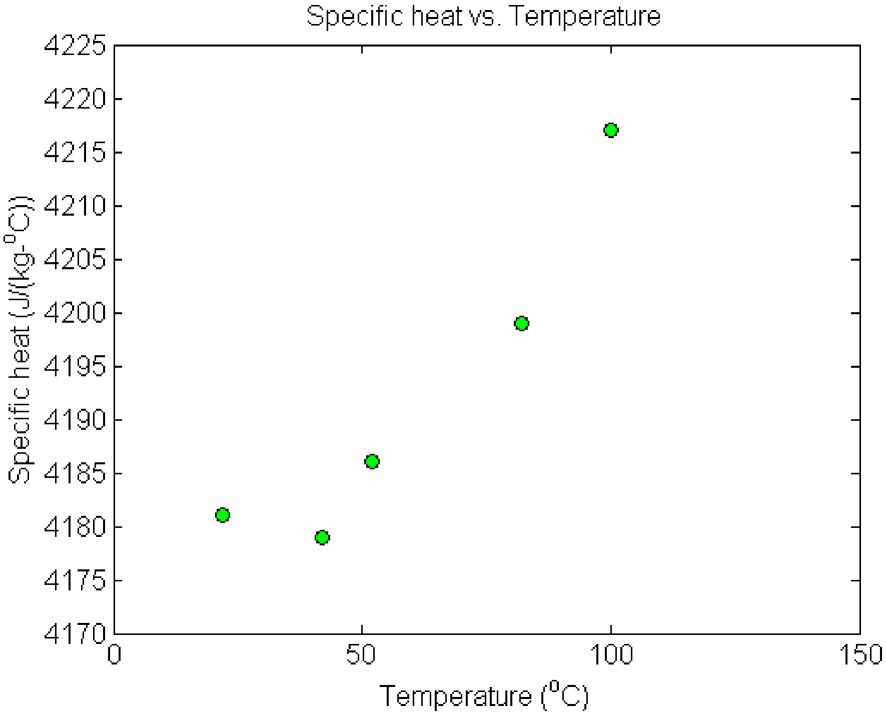


Figure 1 Specific heat of water vs. temperature.

Solution

For first order polynomial interpolation (also called linear interpolation), we choose the specific heat given by

$$C_p(T) = a_0 + a_1T$$

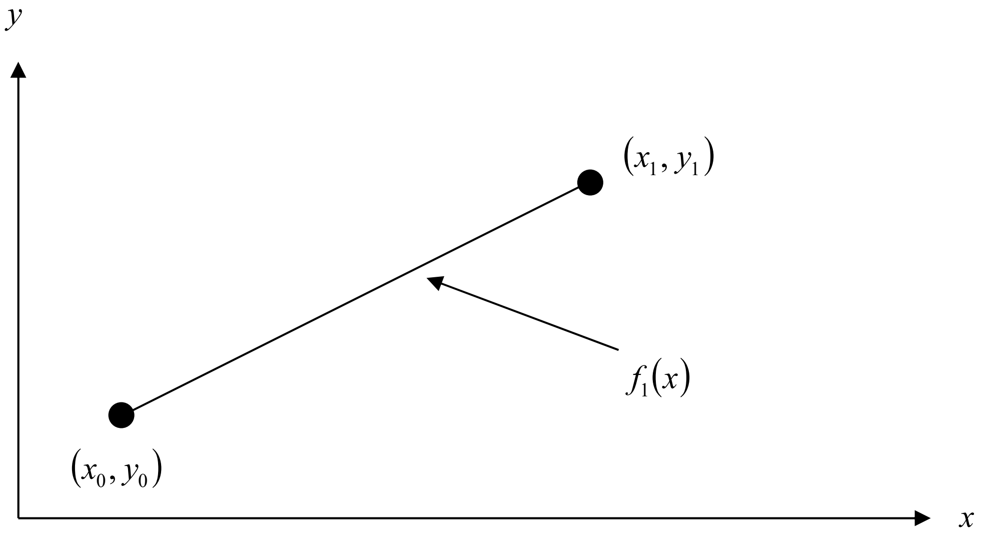


Figure 2 Linear interpolation.

Since we want to find the specific heat at $T = 61°\text{C}$, and we are using a first order polynomial, we need to choose the two data points that are closest to $T = 61°\text{C}$ that also bracket $T = 61°\text{C}$ to evaluate it. The two points are $T_0 = 52$ and $T_1 = 82$.

Then

$$T_0 = 52,\ C_p(T_0) = 4186$$
$$T_1 = 82,\ C_p(T_1) = 4199$$

gives

$$C_p(52) = a_0 + a_1(52) = 4186$$
$$C_p(82) = a_0 + a_1(82) = 4199$$

Writing the equations in matrix form, we have

$$\begin{bmatrix} 1 & 52 \\ 1 & 82 \end{bmatrix}\begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} 4186 \\ 4199 \end{bmatrix}$$

Solving the above two equations gives

$$a_0 = 4163.5$$
$$a_1 = 0.43333$$

Hence

$$\begin{aligned} C_p(T) &= a_0 + a_1 T \\ &= 4163.5 + 0.43333T,\ 52 \le T \le 82 \end{aligned}$$

At $T = 61$,

$$\begin{aligned} C_p(61) &= 4163.5 + 0.43333(61) \\ &= 4189.9 \frac{\text{J}}{\text{kg} - °\text{C}} \end{aligned}$$

Example 2 Direct Method of Interpolation – More Examples Chemical Engineering

To find how much heat is required to bring a kettle of water to its boiling point, you are asked to calculate the specific heat of water at $61°\text{C}$. The specific heat of water is given as a function of time in Table 2.

Table 2 Specific heat of water as a function of temperature.

Temperature, T $(°\text{C})$	Specific heat, C_p $\left(\frac{\text{J}}{\text{kg} - °\text{C}}\right)$
22	4181
42	4179
52	4186
82	4199
100	4217

Determine the value of the specific heat at $T = 61°\text{C}$ using the direct method of interpolation and a second order polynomial. Find the absolute relative approximate error for the second order polynomial approximation.

Solution
For second order polynomial interpolation (also called quadratic interpolation), we choose the specific heat given by

$$C_p(T) = a_0 + a_1 T + a_2 T^2$$

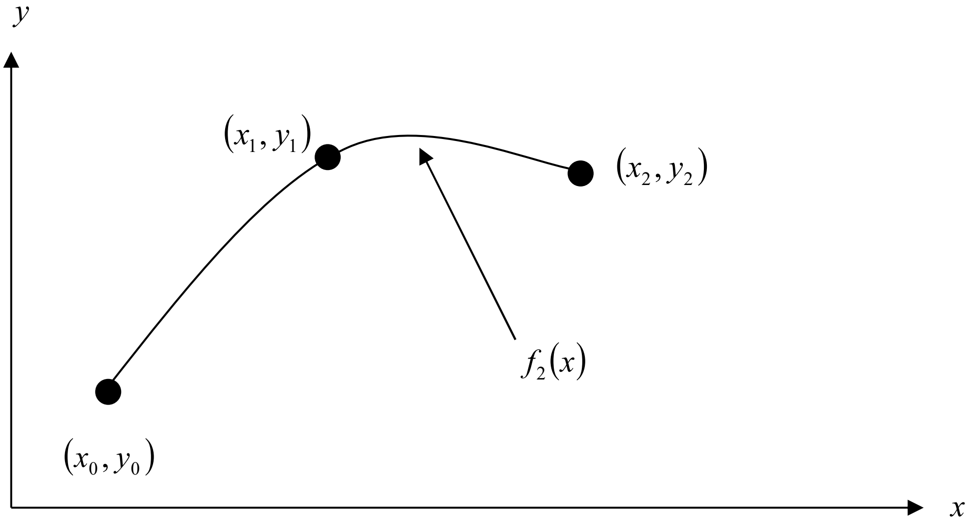


Figure 3 Quadratic interpolation.

Since we want to find the specific heat at $T = 61°\text{C}$, and we are using a second order polynomial, we need to choose the three data points that are closest to $T = 61°\text{C}$ that also bracket $T = 61°\text{C}$ to evaluate it. The three points are $T_0 = 42$, $T_1 = 52$, and $T_2 = 82$.
Then

$$T_0 = 42,\ C_p(T_0) = 4179$$
$$T_1 = 52,\ C_p(T_1) = 4186$$
$$T_2 = 82,\ C_p(T_2) = 4199$$

gives

$$C_p(42) = a_0 + a_1(42) + a_2(42)^2 = 4179$$
$$C_p(52) = a_0 + a_1(52) + a_2(52)^2 = 4186$$
$$C_p(82) = a_0 + a_1(82) + a_2(82)^2 = 4199$$

Writing the three equations in matrix form, we have

$$\begin{bmatrix} 1 & 42 & 1764 \\ 1 & 52 & 2704 \\ 1 & 82 & 6724 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 4179 \\ 4186 \\ 4199 \end{bmatrix}$$

Solving the above three equations gives

$$a_0 = 4135.0$$
$$a_1 = 1.3267$$
$$a_2 = -6.6667 \times 10^{-3}$$

Hence

$$C_p(T) = 4135.0 + 1.3267T - 6.6667 \times 10^{-3} T^2, \ 42 \le T \le 82$$

At $T = 61$,

$$C_p(61) = 4135.0 + 1.3267(61) - 6.6667 \times 10^{-3}(61)^2$$
$$= 4191.2 \frac{\text{J}}{\text{kg} - {}^\circ\text{C}}$$

The absolute relative approximate error $|\in_a|$ obtained between the results from the first and second order polynomial is

$$|\in_a| = \left|\frac{4191.2 - 4189.9}{4191.2}\right| \times 100$$
$$= 0.030063\%$$

Example 3 Direct Method of Interpolation – More Examples Chemical Engineering

To find how much heat is required to bring a kettle of water to its boiling point, you are asked to calculate the specific heat of water at 61°C. The specific heat of water is given as a function of time in Table 3. (bir su ısıtıcısındaki suyu kaynama noktasına çıkarabilmek için suyun 61°C'deki öz ısısının hesaplanması gerekmektedir. Çizelge 3'de suyun öz ısısı ile ilgili bilgiler verilmektedir.)

Table 3 Specific heat of water as a function of temperature.

Temperature, T (°C)	Specific heat, C_p $\left(\frac{\text{J}}{\text{kg} - {}^\circ\text{C}}\right)$
22	4181
42	4179
52	4186
82	4199
100	4217

Determine the value of the specific heat at $T = 61^\circ\text{C}$ using the direct method of interpolation and a third order polynomial. Find the absolute relative approximate error for the third order polynomial approximation.

Solution

For third order polynomial interpolation (also called cubic interpolation), we choose the specific heat given by $C_p(T)=a_0+a_1T+a_2T^2+a_3T^3$

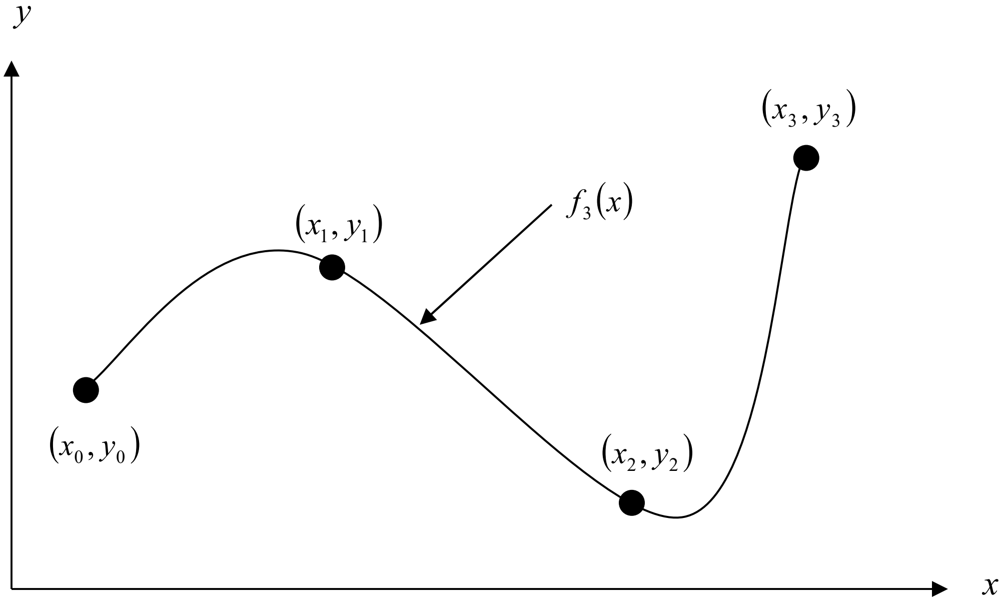


Figure 4 Cubic interpolation.

Since we want to find the specific heat at *T*=61°C, and we are using a third order polynomial, we need to choose the four data points closest to *T*=61°C that also bracket *T*=61°C to evaluate it. The four points are $T_0 = 42, T_1 = 52, T_2 = 82$ and $T_3 = 100$. (Choosing the four points as $T_0 = 22$, $T_1 = 42$, $T_2 = 52$ and $T_3 = 82$ is equally valid.)

Then

$$T_0 = 42, \quad C_p(T_0) = 4179$$
$$T_1 = 52, \quad C_p(T_1) = 4186$$
$$T_2 = 82, \quad C_p(T_2) = 4199$$
$$T_3 = 100, \quad C_p(T_3) = 4217$$

gives

$$C_p(42) = a_0 + a_1(42) + a_2(42)^2 + a_3(42)^3 = 4179$$
$$C_p(52) = a_0 + a_1(52) + a_2(52)^2 + a_3(52)^3 = 4186$$
$$C_p(82) = a_0 + a_1(82) + a_2(82)^2 + a_3(82)^3 = 4199$$
$$C_p(100) = a_0 + a_1(100) + a_2(100)^2 + a_3(100)^3 = 4217$$

Writing the four equations in matrix form, we have

$$\begin{bmatrix} 1 & 42 & 1764 & 7.4088\times10^4 \\ 1 & 52 & 2704 & 1.4061\times10^5 \\ 1 & 82 & 6724 & 5.5137\times10^5 \\ 1 & 100 & 10000 & 10^6 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 4179 \\ 4186 \\ 4199 \\ 4217 \end{bmatrix}$$

Solving the above four equations gives

$$a_0 = 4078.0$$
$$a_1 = 4.4771$$
$$a_2 = -0.062720$$
$$a_3 = 3.1849 \times 10^{-4}$$

Hence

$$C_p(T) = a_0 + a_1 T + a_2 T^2 + a_3 T^3$$
$$= 4078.0 + 4.4771T - 0.062720T^2 + 3.1849 \times 10^{-4} T^3, \quad 42 \le T \le 100$$
$$T(61) = 4078.0 + 4.4771(61) - 0.062720(61)^2 + 3.1849 \times 10^{-4}(61)^3$$
$$= 4190.0 \frac{\text{J}}{\text{kg} - {}^\circ\text{C}}$$

The absolute relative approximate error $|\in_a|$ obtained between the results from the second and third order polynomial is

$$|\in_a| = \left|\frac{4190.0 - 4191.2}{4190.0}\right| \times 100$$
$$= 0.027295\%$$

Example 4 Direct Method of Interpolation – More Examples Civil Engineering

To maximize a catch of bass in a lake, it is suggested to throw the line to the depth of the thermocline. The characteristic feature of this area is the sudden change in temperature. We are given the temperature vs. depth data for a lake in Table 1. (bir gölde levrek balığı avlama olasılığını artırmak için oltanın ucundaki yemin thermocline derinliğe kadar inmesi gerekmektedir. Bu bölgenin-derinliğin-karakteristiği göl sıcaklığının ani değiştiği alan olmasıdır. Çizelge 1'de gölün derinliğine bağlı olarak sıcaklığın değişimi verilmektedir.)

Table 1 Temperature vs. depth for a lake.

Temperature, T $(^\circ\text{C})$	Depth, z (m)
19.1	0
19.1	–1
19	–2
18.8	–3
18.7	–4
18.3	–5
18.2	–6
17.6	–7
11.7	–8
9.9	–9
9.1	–10

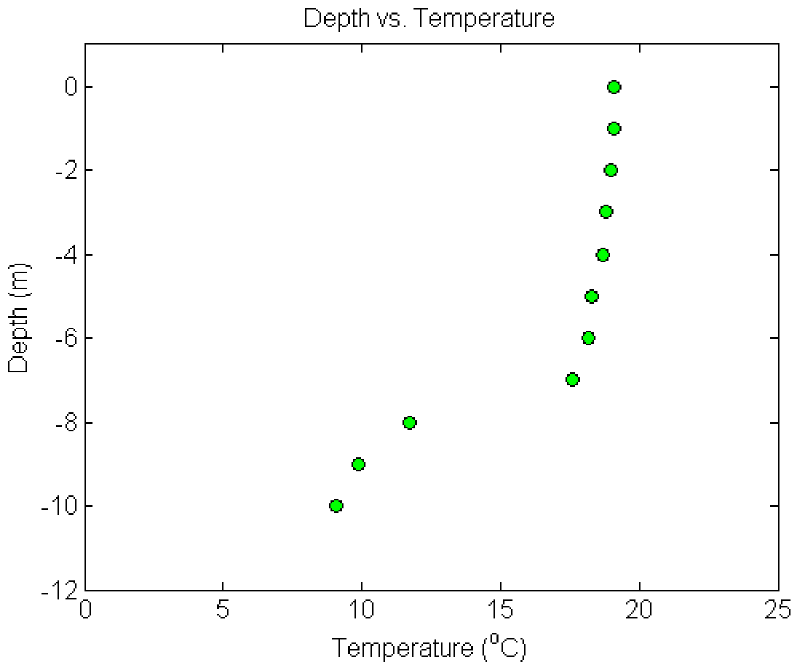


Figure 1 Temperature vs. depth of a lake.

Using the given data, we see the largest change in temperature is between $z = -8$ m and $z = -7$ m. Determine the value of the temperature at $z = -7.5$ m using the direct method of interpolation and a first order polynomial.

Solution

For first order polynomial interpolation (also called linear interpolation), we choose the temperature given by

$$T(z) = a_0 + a_1 z$$

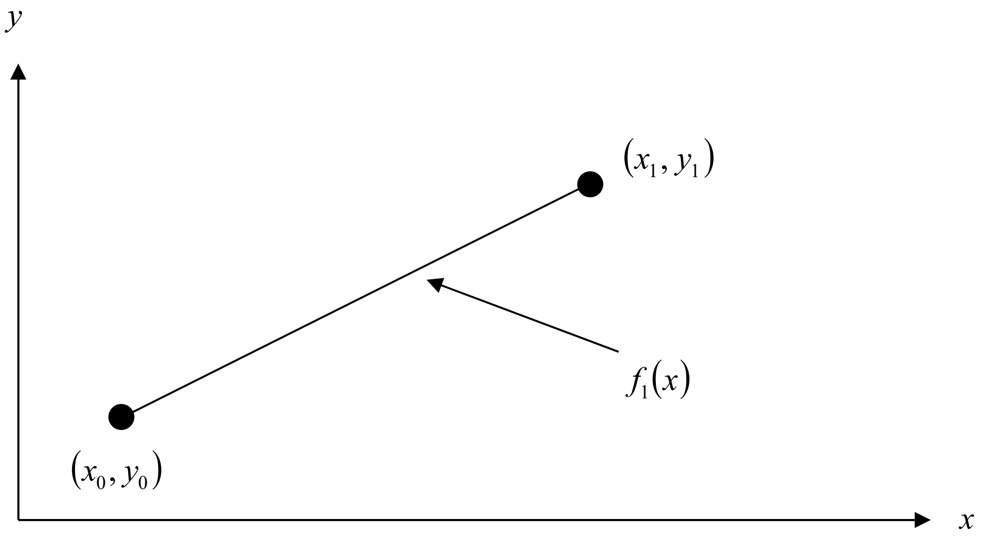


Figure 2 Linear interpolation.

Since we want to find the temperature at $z = -7.5$ m, and we are using a first order polynomial, we need to choose the two data points that are closest to $z = -7.5$ m that also bracket $z = -7.5$ m to evaluate it. The two points are $z_0 = -8$ and $z_1 = -7$.

Then

$$z_0 = -8,\ T(z_0) = 11.7$$
$$z_1 = -7,\ T(z_1) = 17.6$$

gives

$$T(-8) = a_0 + a_1(-8) = 11.7$$
$$T(-7) = a_0 + a_1(-7) = 17.6$$

Writing the equations in matrix form, we have

$$\begin{bmatrix} 1 & -8 \\ 1 & -7 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} 11.7 \\ 17.6 \end{bmatrix}$$

Solving the above two equations gives

$$a_0 = 58.9 \text{ and } a_1 = 5.9$$

Hence

$$T(z) = a_0 + a_1 z$$
$$T(z) = 58.9 + 5.9z,\ -8 \le z \le -7$$
$$T(-7.5) = 58.9 + 5.9(-7.5)$$
$$= 14.65\,°\mathrm{C}$$

Example 5 Direct Method of Interpolation – More Examples Civil Engineering

To maximize a catch of bass in a lake, it is suggested to throw the line to the depth of the thermocline. The characteristic feature of this area is the sudden change in temperature. We are given the temperature vs. depth data for a lake in Table 2. (bir gölde levrek balığı avlama olasılığını artırmak için oltanın ucundaki yemin thermocline derinliğe kadar inmesi gerekmektedir. Bu bölgenin-derinliğin-karakteristiği göl sıcaklığının ani değiştiği alan olmasıdır. Çizelge 2'de gölün derinliğine bağlı olarak sıcaklığın değişimi verilmektedir.)

Table 2 Temperature vs. depth for a lake.

Temperature, T (°C)	Depth, z (m)
19.1	0
19.1	–1
19	–2
18.8	–3
18.7	–4
18.3	–5
18.2	–6
17.6	–7
11.7	–8
9.9	–9

9.1	−10

Using the given data, we see the largest change in temperature is between $z=-8$ m and $z=-7$ m. Determine the value of the temperature at $z=-7.5$ m using the direct method of interpolation and a second order polynomial. Find the absolute relative approximate error for the second order polynomial approximation.

Solution

For second order polynomial interpolation (also called quadratic interpolation), we choose the velocity given by

$$v(t)=a_0+a_1t+a_2t^2$$

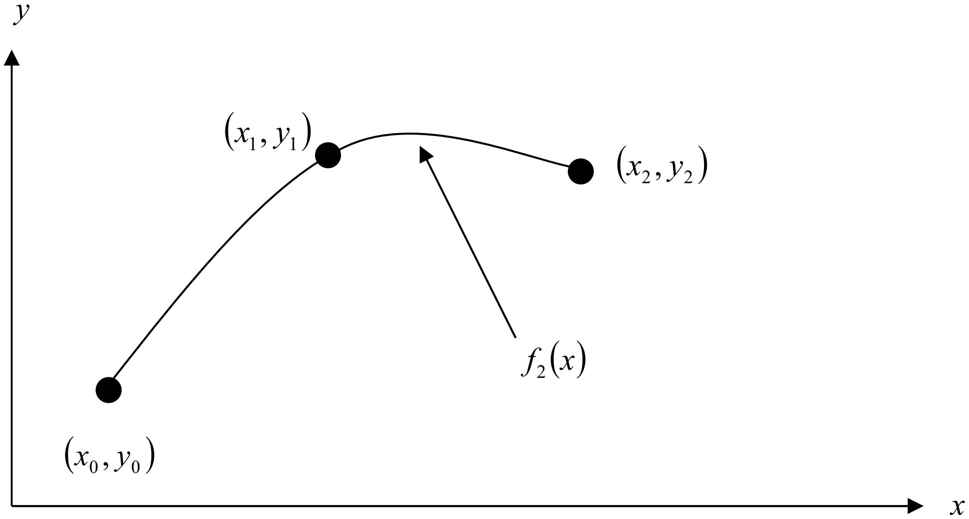


Figure 3 Quadratic interpolation.

Since we want to find the temperature at $z=-7.5$, and we are using a second order polynomial, we need to choose the three data points that are closest to $z=-7.5$ that also bracket $z=-7.5$ to evaluate it. The three points are $z_0=-9$, $z_1=-8$ and $z_2=-7$. (Choosing the three points as $z_0=-8$, $z_1=-7$ and $z_2=-6$ is equally valid.)

Then

$$z_0=-9,\ T(z_0)=9.9$$
$$z_1=-8,\ T(z_1)=11.7$$
$$z_2=-7,\ T(z_2)=17.6$$

gives

$$T(-9)=a_0+a_1(-9)+a_2(-9)^2=9.9$$
$$T(-8)=a_0+a_1(-8)+a_2(-8)^2=11.7$$
$$T(-7)=a_0+a_1(-7)+a_2(-7)^2=17.6$$

Writing the three equations in matrix form

$$\begin{bmatrix} 1 & -9 & 81 \\ 1 & -8 & 64 \\ 1 & -7 & 49 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 9.9 \\ 11.7 \\ 17.6 \end{bmatrix}$$

and the solution of the above three equations gives

$$a_0 = 173.7$$

$$a_1 = 36.65$$

$$a_2 = 2.05$$

Hence

$$T(z) = 173.7 + 36.65z + 2.05z^2, \ -9 \le z \le -7$$

At z=–7.5,

$$T(-7.5)=173.7+36.65(-7.5)+2.05(-7.5)^2$$
$$=14.138°\text{C}$$

The absolute relative approximate error $|\epsilon_a|$ obtained between the results from the first and second order polynomial is

$$|\epsilon_a| = \left| \frac{14.138 - 14.65}{14.138} \right| \times 100$$
$$= 3.6251\%$$

Example 6 Direct Method of Interpolation – More Examples Civil Engineering

To maximize a catch of bass in a lake, it is suggested to throw the line to the depth of the thermocline. The characteristic feature of this area is the sudden change in temperature. We are given the temperature vs. depth data for a lake in Table 3. (bir gölde levrek balığı avlama olasılığını artırmak için oltanın ucundaki yemin thermocline derinliğe kadar inmesi gerekmektedir. Bu bölgenin-derinliğin-karakteristiği göl sıcaklığının ani değiştiği alan olmasıdır. Çizelge 3'de gölün derinliğine bağlı olarak sıcaklığın değişimi verilmektedir.)

Table 3 Temperature vs. depth for a lake.

Temperature, T (°C)	Depth, z (m)
19.1	0
19.1	–1
19.0	–2
18.8	–3
18.7	–4
18.3	–5
18.2	–6
17.6	–7
11.7	–8
9.9	–9
9.1	–10

Using the given data, we see the largest change in temperature is between z=–8 m and z=–7 m.

a) Determine the value of the temperature at z=–7.5 m using the direct method of interpolation and a third order polynomial. Find the absolute relative approximate error for the third order polynomial approximation.

b) The position where the thermocline exists is given where $\frac{d^2T}{dz^2}=0$. Using the expression from part (a), what is the value of the depth at which the thermocline exists?

Solution

a) For third order polynomial interpolation (also called cubic interpolation), we choose the temperature given by

$$T(z)=a_0+a_1z+a_2z^2+a_3z^3$$

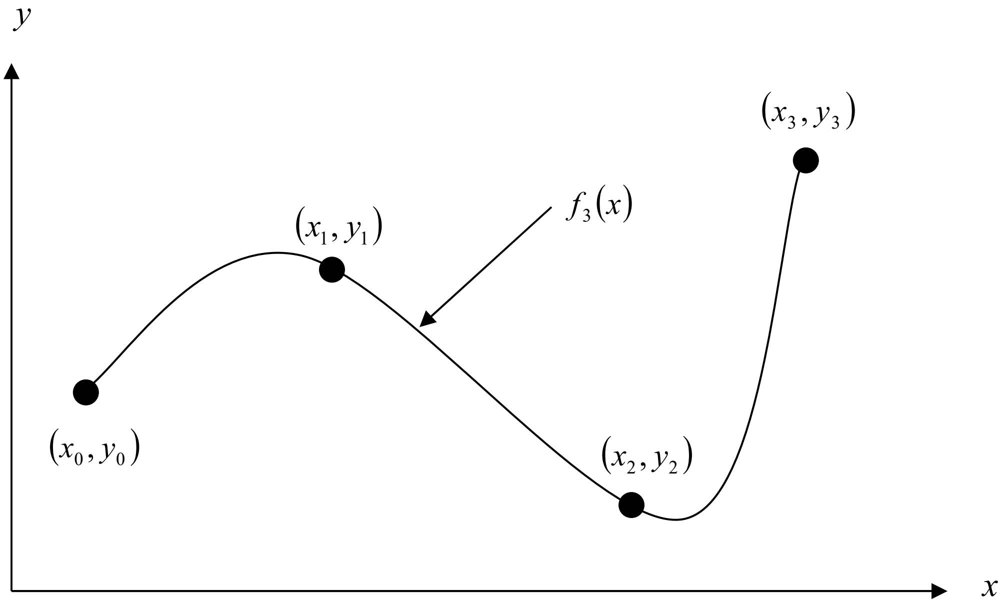


Figure 4 Cubic interpolation.

Since we want to find the temperature at $z=-7.5$, and we are using a third order polynomial, we need to choose the four data points closest to $z=-7.5$ that also bracket $z=-7.5$ to evaluate it. The four points are $z_0=-9$, $z_1=-8$, $z_2=-7$ and $z_3=-6$.

Then

$$z_0=-9,\ T(z_0)=9.9$$
$$z_1=-8,\ T(z_1)=11.7$$
$$z_2=-7,\ T(z_2)=17.6$$
$$z_3=-6,\ T(z_3)=18.2$$

gives

$$T(-9)=a_0+a_1(-9)+a_2(-9)^2+a_3(-9)^3=9.9$$

$$T(-8) = a_0 + a_1(-8) + a_2(-8)^2 + a_3(-8)^3 = 11.7$$

$$T(-7) = a_0 + a_1(-7) + a_2(-7)^2 + a_3(-7)^3 = 17.6$$

$$T(-6) = a_0 + a_1(-6) + a_2(-6)^2 + a_3(-6)^3 = 18.2$$

Writing the four equations in matrix form, we have

$$\begin{bmatrix} 1 & -9 & 81 & -729 \\ 1 & -8 & 64 & -512 \\ 1 & -7 & 49 & -343 \\ 1 & -6 & 36 & -216 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 9.9 \\ 11.7 \\ 17.6 \\ 18.2 \end{bmatrix}$$

Solving the above four equations gives

$$a_0 = -615.9$$

$$a_1 = -262.58$$

$$a_2 = -35.55$$

$$a_3 = -1.5667$$

Hence

$$T(z) = a_0 + a_1 z + a_2 z^2 + a_3 z^3$$

$$= -615.9 - 262.58z - 35.55z^2 - 1.5667z^3, \quad -9 \le z \le -6$$

$$T(-7.5) = -615.9 - 262.58(-7.5) - 35.55(-7.5)^2 - 1.5667(-7.5)^3$$

$$= 14.725 ^\circ\text{C}$$

The absolute relative approximate error $|\in_a|$ obtained between the results from the second and third order polynomial is

$$|\in_a| = \left| \frac{14.725 - 14.138}{14.725} \right| \times 100$$

$$= 3.9898\%$$

b) To find the position of the thermocline, we must find the points of inflection of the third order polynomial, given by $\frac{d^2T}{dz^2} = 0$

$$T(z) = -615.9 - 262.58z - 35.55z^2 - 1.5667z^3, \quad -9 \le z \le -6$$

$$\frac{dT}{dz} = -262.58 - 71.10z - 4.7z^2, \quad -9 \le z \le -6$$

$$\frac{d^2T}{dz^2} = -71.1 - 9.4z, \quad -9 \le z \le -6$$

Simply setting this expression equal to zero, we get

$$0 = -71.10 - 9.4z$$

$$z = -7.5638 \text{ m}$$

This answer can be verified due to the fact that it falls within the specified range of the third order polynomial and it also falls within the region of the greatest temperature change in the collected data from the lake.

Example 7 Direct Method of Interpolation – More Examples Mechanical Engineering

For the purpose of shrinking a trunnion into a hub, the reduction of diameter ΔD of a trunnion shaft by cooling it through a temperature change of ΔT is given by

$$\Delta D = D\,\alpha\,\Delta T$$

where

D= original diameter (in.)

α= coefficient of thermal expansion at average temperature (in/in/°F)

The trunnion is cooled from $80°\text{F}$ to $-108°\text{F}$, giving the average temperature as $-14°\text{F}$. The table of the coefficient of thermal expansion vs. temperature data is given in Table 1.

Table 1 Thermal expansion coefficient as a function of temperature.

Temperature, T (°F)	Thermal Expansion Coefficient, α (in/in/°F)
80	6.47×10^{-6}
0	6.00×10^{-6}
–60	5.58×10^{-6}
–160	4.72×10^{-6}
–260	3.58×10^{-6}
–340	2.45×10^{-6}

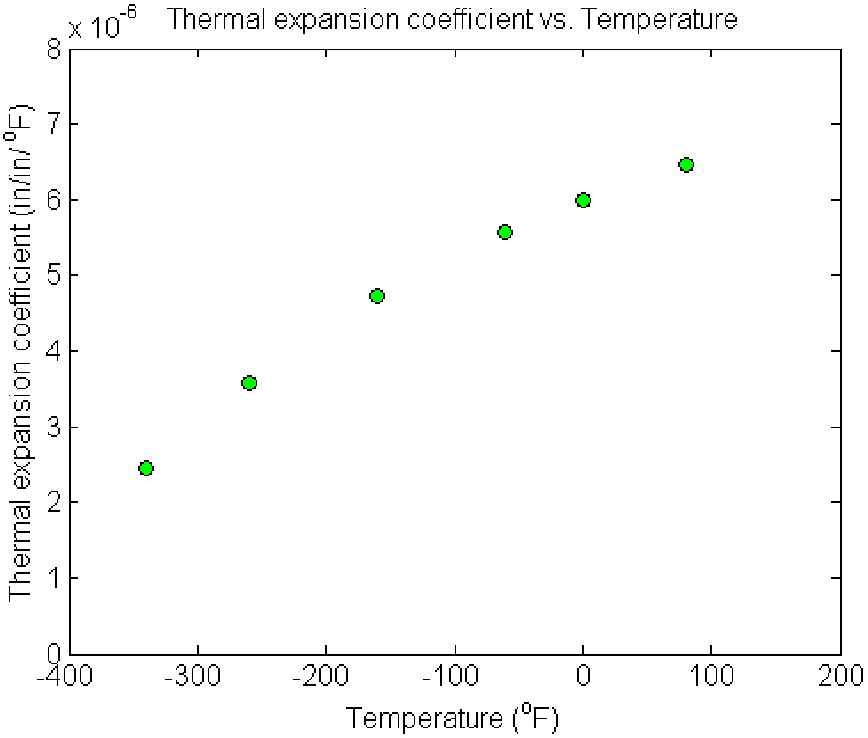


Figure 1 Thermal expansion coefficient vs. temperature.

If the coefficient of thermal expansion needs to be calculated at the average temperature of $-14\,°\text{F}$, determine the value of the coefficient of thermal expansion at $T = -14\,°\text{F}$ using the direct method of interpolation and a first order polynomial.

Solution

For first order polynomial interpolation (also called linear interpolation), we choose the coefficient of thermal expansion given by

$$\alpha(T) = a_0 + a_1 T$$

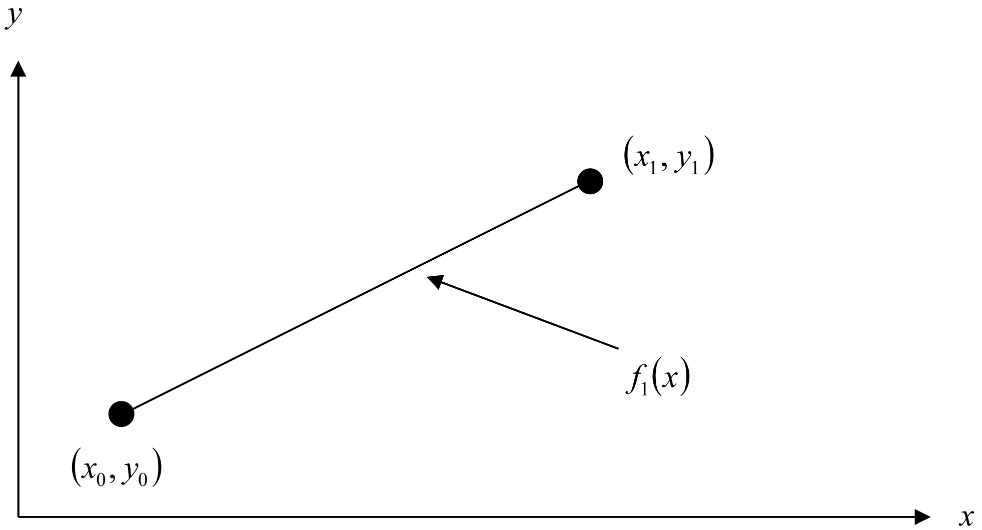


Figure 2 Linear interpolation.

Since we want to find the coefficient of thermal expansion at $T = -14\,°\text{F}$, and we are using a first order polynomial, we need to choose the two data points that are closest to $T = -14\,°\text{F}$ that also bracket $T = -14\,°\text{F}$ to evaluate it. The two points are $T_0 = 0\,°\text{F}$ and $T_1 = -60\,°\text{F}$.

Then

$$T_0 = 0, \quad \alpha(T_0) = 6.00 \times 10^{-6}$$
$$T_1 = -60, \alpha(T_1) = 5.58 \times 10^{-6}$$

gives

$$\alpha(0) = a_0 + a_1(0) = 6.00 \times 10^{-6}$$
$$\alpha(-60) = a_0 + a_1(-60) = 5.58 \times 10^{-6}$$

Writing the equations in matrix form, we have

$$\begin{bmatrix} 1 & 0 \\ 1 & -60 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} 6.00 \times 10^{-6} \\ 5.58 \times 10^{-6} \end{bmatrix}$$

Solving the above two equations gives

$$a_0 = 6.00 \times 10^{-6}$$
$$a_1 = 0.007 \times 10^{-6}$$

Hence

$$\alpha(T) = a_0 + a_1 T$$
$$= 6.00\times10^{-6} + 0.007\times10^{-6}T,\ -60 \le T \le 0$$

At $T = -14\,°\text{F}$,

$$\alpha(-14) = 6.00\times10^{-6} + 0.007\times10^{-6}(-14)$$
$$= 5.902\times10^{-6}\ \text{in/in/°F}$$

Example 8 Direct Method of Interpolation – More Examples Mechanical Engineering

For the purpose of shrinking a trunnion into a hub, the reduction of diameter ΔD of a trunnion shaft by cooling it through a temperature change of ΔT is given by (ΔD çapındaki mafsal şaftının ΔT kadar sıcaklığı değiştirilerek büzüşmesi amaçlanmaktadır.)

$\Delta D = D\alpha\ \Delta T$

where

D= original diameter (in.)

α= coefficient of thermal expansion at average temperature (in/in/°F)

The trunnion is cooled from $80\,°\text{F}$ to $-108\,°\text{F}$, giving the average temperature as $-14\,°\text{F}$. The table of the coefficient of thermal expansion vs. temperature data is given in Table 2.

Table 2 Thermal expansion coefficient as a function of temperature.

Temperature, T (°F)	Thermal Expansion Coefficient, α (in/in/°F)
80	6.47×10^{-6}
0	6.00×10^{-6}
–60	5.58×10^{-6}
–160	4.72×10^{-6}
–260	3.58×10^{-6}
–340	2.45×10^{-6}

If the coefficient of thermal expansion needs to be calculated at the average temperature of $-14\,°\text{F}$, determine the value of the coefficient of thermal expansion at $T = -14\,°\text{F}$ using the direct method of interpolation and a first order polynomial.

Solution

For second order polynomial interpolation (also called quadratic interpolation), we choose the coefficient of thermal expansion given by

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

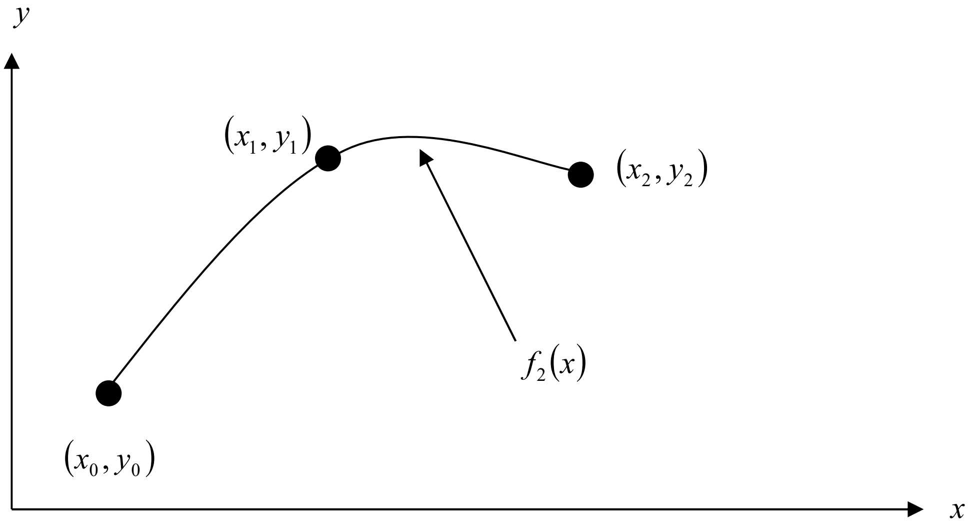


Figure 3 Quadratic interpolation.

Since we want to find the coefficient of thermal expansion at $T = -14°\text{F}$, and we are using a second order polynomial, we need to choose the three data points that are closest to $T = -14°\text{F}$ that also bracket $T = -14°\text{F}$ to evaluate it. These three points are $T_0 = 80\ °\text{F}$, $T_1 = 0°\text{F}$ and $T_2 = -60°\text{F}$.

Then

$$T_0 = 80, \quad \alpha(T_0) = 6.47 \times 10^{-6}$$
$$T_1 = 0, \quad \alpha(T_1) = 6.00 \times 10^{-6}$$
$$T_2 = -60, \quad \alpha(T_2) = 5.58 \times 10^{-6}$$

gives

$$\alpha(80) = a_0 + a_1(80) + a_2(80)^2 = 6.47 \times 10^{-6}$$
$$\alpha(0) = a_0 + a_1(0) + a_2(0)^2 = 6.00 \times 10^{-6}$$
$$\alpha(-60) = a_0 + a_1(-60) + a_2(-60)^2 = 5.58 \times 10^{-6}$$

Writing the three equations in matrix form, we have

$$\begin{bmatrix} 1 & 80 & 6400 \\ 1 & 0 & 0 \\ 1 & -60 & 3600 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 6.47 \times 10^{-6} \\ 6.00 \times 10^{-6} \\ 5.58 \times 10^{-6} \end{bmatrix}$$

Solving the above three equations gives

$$a_0 = 6.00 \times 10^{-6}$$
$$a_1 = 6.5179 \times 10^{-9}$$
$$a_2 = -8.0357 \times 10^{-12}$$

Hence

$$\alpha(T) = 6.00 \times 10^{-6} + 6.5179 \times 10^{-9} T - 8.0357 \times 10^{-12} T^2, \ -60 \le T \le 80$$

At $T = -14°\text{F}$,

$$\alpha(-14) = 6.00\times10^{-6} + 6.5179\times10^{-9}(-14) - 8.0357\times10^{-12}(-14)^2$$
$$= 5.9072\times10^{-6} \text{ in/in/°F}$$

The absolute relative approximate error $|\in_a|$ obtained between the results from the first and second order polynomial is

$$|\in_a| = \left|\frac{5.9072\times10^{-6} - 5.902\times10^{-6}}{5.9072\times10^{-6}}\right|\times100$$
$$= 0.087605\%$$

Example 9 Direct Method of Interpolation – More Examples Mechanical Engineering

For the purpose of shrinking a trunnion into a hub, the reduction of diameter ΔD of a trunnion shaft by cooling it through a temperature change of ΔT is given by

$$\Delta D = D\alpha\Delta T$$

where

D= original diameter (in.)

α= coefficient of thermal expansion at average temperature (in/in/°F)

The trunnion is cooled from $80°\text{F}$ to $-108°\text{F}$, giving the average temperature as $-14°\text{F}$. The table of the coefficient of thermal expansion vs. temperature data is given in Table 3.

Table 3 Thermal expansion coefficient as a function of temperature.

Temperature, T (°F)	Thermal Expansion Coefficient, α (in/in/°F)
80	6.47×10^{-6}
0	6.00×10^{-6}
–60	5.58×10^{-6}
–160	4.72×10^{-6}
–260	3.58×10^{-6}
–340	2.45×10^{-6}

a) If the coefficient of thermal expansion needs to be calculated at the average temperature of –14°F, determine the value of the coefficient of thermal expansion at T= –14°F using the direct method of interpolation and a first order polynomial. Find the absolute relative approximate error for the third order polynomial approximation.

b) The actual reduction in diameter is given by

$$\Delta D = D\int_{T_r}^{T_f} \alpha dT$$

where T_r= room temperature (°F)

T_f= temperature of cooling medium (°F)

Since

T_r=80°F

T_f= –108°F

$$\Delta D = D \int_{80}^{-108} \alpha dT$$

Find out the percentage difference in the reduction in the diameter by the above integral formula and the result using the thermal expansion coefficient from part (a).

Solution

a) For third order polynomial interpolation (also called cubic interpolation), we choose the coefficient of thermal expansion given by

$$\alpha(T) = a_0 + a_1 T + a_2 T^2 + a_3 T^3$$

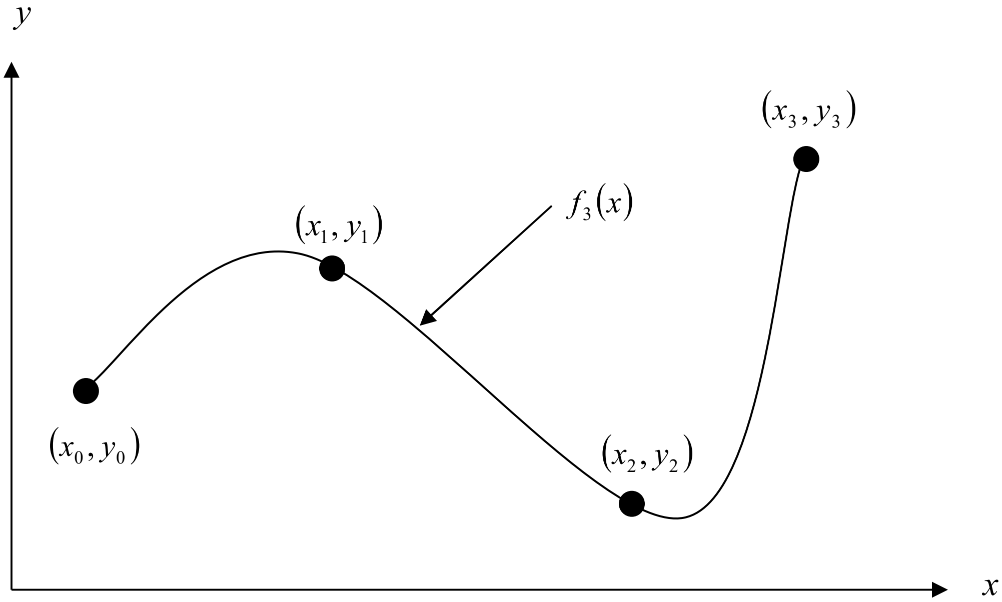


Figure 4 Cubic interpolation.

Since we want to find the coefficient of thermal expansion at $T = -14\,°\text{F}$, and we are using a third order polynomial, we need to choose the four data points closest to $T = -14\,°\text{F}$ that also bracket $T = -14\,°\text{F}$ to evaluate it. Then the four points are $\text{T}_0 = 80\,°\text{F}$, $\text{T}_1 = 0\,°\text{F}$, $T_2 = -60\,°\text{F}$ and $T_3 = -160\,°\text{F}$.

$$T_0 = 80, \quad \alpha(T_0) = 6.47 \times 10^{-6}$$
$$T_1 = 0, \quad \alpha(T_1) = 6.00 \times 10^{-6}$$
$$T_2 = -60, \quad \alpha(T_2) = 5.58 \times 10^{-6}$$
$$T_3 = -160, \quad \alpha(T_3) = 4.72 \times 10^{-6}$$

gives

$$\alpha(80) = a_0 + a_1(80) + a_2(80)^2 + a_3(80)^3 = 6.47 \times 10^{-6}$$
$$\alpha(0) = a_0 + a_1(0) + a_2(0)^2 + a_3(0)^3 = 6.00 \times 10^{-6}$$

$$\alpha(-60) = a_0 + a_1(-60) + a_2(-60)^2 + a_3(-60)^3 = 5.58\times10^{-6}$$

$$\alpha(-160) = a_0 + a_1(-160) + a_2(-160)^2 + a_3(-160)^3 = 4.72\times10^{-6}$$

Writing the four equations in matrix form, we have

$$\begin{bmatrix} 1 & 80 & 6400 & 5.12\times10^5 \\ 1 & 0 & 0 & 0 \\ 1 & -60 & 3600 & -2.16\times10^5 \\ 1 & -160 & 25600 & -4.096\times10^6 \end{bmatrix}\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 6.47\times10^{-6} \\ 6.00\times10^{-6} \\ 5.58\times10^{-6} \\ 4.72\times10^{-6} \end{bmatrix}$$

Solving the above four equations gives

$a_0 = 6.00\times10^{-6}$

$a_1=6.4786\times10^{-9}$

$a_2=-8.1994\times10^{-12}$

$a_3=8.1845\times10^{-15}$

Hence

$$\alpha(T)=a_0+a_1T+a_2T^2+a_3T^3$$

$$= 6.00\times10^{-6} + 6.4786\times10^{-9}T - 8.1994\times10^{-12}T^2 + 8.1845\times10^{-15}T^3, \quad -160 \le T \le 80$$

$$\alpha(-14) = 6.00\times10^{-6} + 6.4786\times10^{-9}(-14) - 8.1994\times10^{-12}(-14)^2 + 8.1845\times10^{-15}(-14)^3$$

$$= 5.9077\times10^{-6} \text{ in/in/°F}$$

The absolute relative approximate error $|\in_a|$ obtained between the results from the second and third order polynomial is

$$|\in_a| = \left|\frac{5.9077\times10^{-6} - 5.9072\times10^{-6}}{5.9077\times10^{-6}}\right|\times100$$

$$= 0.0083867\%$$

b) In finding the percentage difference in the reduction in diameter, we can rearrange the integral formula to

$$\frac{\Delta D}{D} = \int_{T_r}^{T_f} \alpha dT$$

and since we know from part (a) that

$\alpha(T) = 6.00\times10^{-6} + 6.4786\times10^{-9}T - 8.1994\times10^{-12}T^2 + 8.1845\times10^{-15}T^3, \quad -160 \le T \le 80$ we see that we can use the integral formula in the range from $T_f = -108$°F to $T_r = 80$°F

Therefore,

$$\frac{\Delta D}{D} = \int_{T_r}^{T_f} \alpha dT$$

$$= \int_{80}^{-108} \left(6.00\times10^{-6} + 6.4786\times10^{-9}T - 8.1994\times10^{-12}T^2 + 8.1845\times10^{-15}T^3\right)dT$$

$$= \left[6.00\times10^{-6}T + 6.4786\times10^{-9}\frac{T^2}{2} - 8.1994\times10^{-12}\frac{T^3}{3} + 8.1845\times10^{-15}\frac{T^4}{4}\right]_{80}^{-108}$$

$$= -1105.9 \times 10^{-6}$$

So $\frac{\Delta D}{D} = -1105.9 \times 10^{-6}$ in/in using the actual reduction in diameter integral formula. If we use the average value for the coefficient of thermal expansion from part (a), we get

$$\begin{aligned}\frac{\Delta D}{D} &= \alpha \Delta T \\ &= \alpha \left(T_f - T_r\right) \\ &= 5.9077 \times 10^{-6} \left(-108 - 80\right) \\ &= -1110.6 \times 10^{-6}\end{aligned}$$

and $\frac{\Delta D}{D} = -1110.6 \times 10^{-6}$ in/in using the average value of the coefficient of thermal expansion using a third order polynomial. Considering the integral to be the more accurate calculation, the percentage difference would be

$$\begin{aligned}\left|\in_a\right| &= \left|\frac{-1105.9 \times 10^{-6} - \left(-1110.6 \times 10^{-6}\right)}{-1105.9 \times 10^{-6}}\right| \times 100 \\ &= 0.42775\%\end{aligned}$$

INTERPOLATION	
Topic	Direct Method of Interpolation
Summary	Examples of direct method of interpolation.
Major	Chemical Engineering
Authors	Autar Kaw
Date	Aralık 8, 2016
Web Site	http://numericalmethods.eng.usf.edu