Course Learning Outcomes (CLOs)

Course Title: **Engineering Numerical Methods**

Course Code: **ME 3202**

Course Units: **3 Credits**

Course Category: **Department Requirement**

Course Instructor: **Asst. Prof. Dr. Ghalib R. Ibrahim**

Course Learning Outcomes:

By the end of successful completion of this course, the student will be able to:

1. To gain experience in error analysis.
2. Understanding the different numerical methods to solve systems of linear and nonlinear equations.
3. Understanding the different numerical methods for differentiation, integration, and solving a set of ordinary differential equations.
4. Understanding how numerical methods can be implemented in MATLAB software.

Numerical analysis Course Code: ME 3202

Topics

- Error Analysis
- Roots of equations
- Solving system of linear equations
- Integration and differentiation
- Ordinary differential equations

Measuring 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 $h = 0.3$,

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

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

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

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

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

c) True error is calculated as

$$E_t = \text{True value} - \text{Approximate value}$$
$$= 9.5140 - 10.265$$
$$= -0.75061$$

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,

$$E_t = \text{True value} - \text{Approximate value}$$
$$= 9.5140 - 10.265$$
$$= -0.75061$$

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 $h = 0.3$,

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

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

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

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

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

$$= \frac{f(2.15) - f(2)}{0.15}$$
$$= \frac{7e^{0.5(2.15)} - 7e^{0.5(2)}}{0.15}$$
$$= \frac{20.50 - 19.028}{0.15}$$
$$= 9.8799$$

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 $f(x) = 7\times10^{-6}e^{0.5x}$, the approximate error in calculating $f^{'}(2)$ with $h = 0.15$ would be $E_a = -0.38474\times10^{-6}$. 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

$$|\in_a| = |-0.038942|$$

$$= 0.038942 \text{ or } 3.8942\%$$

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, 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, $|\in_a| \le \in_s$, 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, $|\in_a| \le 0.5 \times 10^{2-m}\%$.

Example 5

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

Solution

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

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

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

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

to signify the correct number of significant digits.

Example 5

Give some examples of showing the number of significant digits.

Solution

a) 0.0459 has three significant digits
b) 4.590 has four significant digits
c) 4008 has four significant digits
d) 4008.0 has five significant digits
e) 1.079×10^3 has four significant digits
f) 1.0790×10^3 has five significant digits
g) 1.07900×10^3 has six significant digits

Reference

INTRODUCTION, APPROXIMATION AND ERRORS	
Topic	Measuring Errors
Summary	Textbook notes on measuring errors
Major	General Engineering
Authors	Autar Kaw
Date	February 26, 2022
Web Site	http://numericalmethods.eng.usf.edu

Sources of 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 $\frac{1}{3} - 0.333333 = 0.00000033\overline{3}$. 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 is truncation error?

A: Truncation error is defined as the error caused by truncating a mathematical procedure. For example, the Maclaurin series for e^x is given as

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

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

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

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

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

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

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	$\|\in_a\|\%$
1	1	-	-
2	2.2	1.2	54.546
3	2.92	0.72	24.658
4	3.208	0.288	8.9776
5	3.2944	0.0864	2.6226
6	3.3151	0.020736	0.62550

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

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

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

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

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,

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

$$= \lim_{\Delta x \to 0} \frac{(x + \Delta x)^2 - (x)^2}{\Delta x}$$

$$= \lim_{\Delta x \to 0} \frac{x^2 + 2x\Delta x + (\Delta x)^2 - x^2}{\Delta x}$$
$$= \lim_{\Delta x \to 0} (2x + \Delta x)$$
$$= 2x$$

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

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

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

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

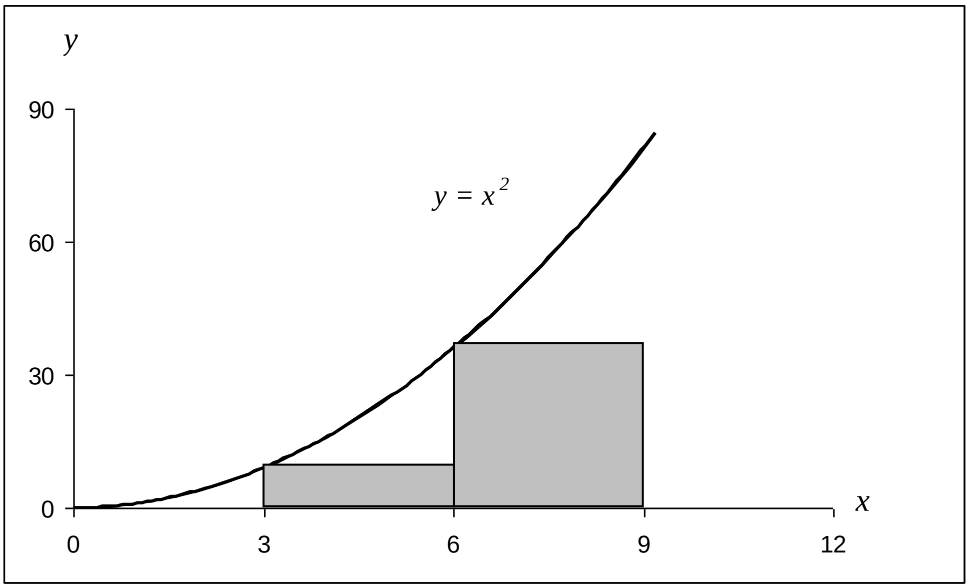

Figure 2 Plot of $y = x^2$ showing the approximate area under the curve from $x = 3$ to $x = 9$ using two rectangles.

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?

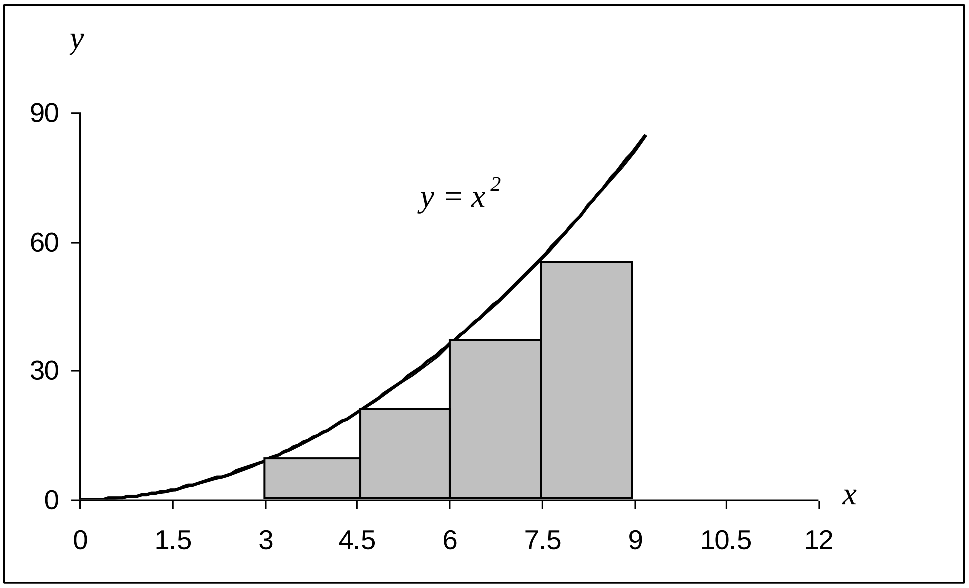

Figure 3 Plot of $y = x^2$ showing the approximate area under the curve from $x = 3$ to $x = 9$ using four rectangles.

Reference

INTRODUCTION, APPROXIMATION AND ERRORS	
Topic	Sources of error
Summary	Textbook notes on sources of error
Major	General Engineering
Authors	Autar Kaw
Date	February 28, 2022
Web Site	http://numericalmethods.eng.usf.edu

Bisection Method of Solving a Nonlinear Equation

What is the bisection method and what is it based on?

One of the first numerical methods developed to find the root of a nonlinear equation $f(x)=0$ was the bisection method (also called *binary-search* method). The method is based on the following theorem.

Theorem

An equation $f(x)=0$, where $f(x)$ is a real continuous function, has at least one root between x_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 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.

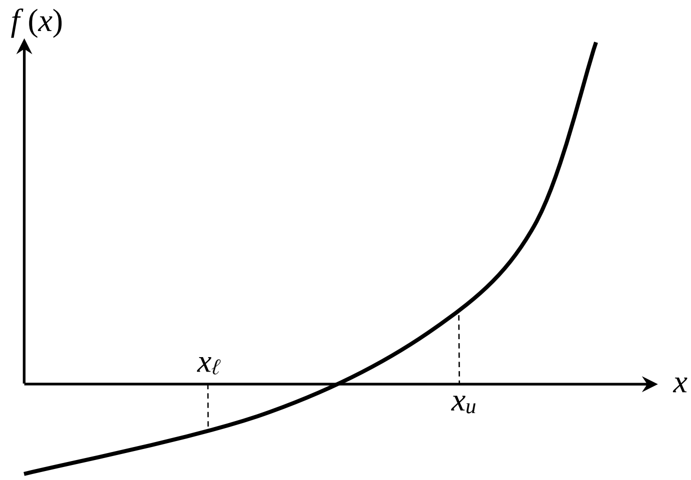

Figure 1 At least one root exists between the two points if the function is real, continuous, and changes sign.

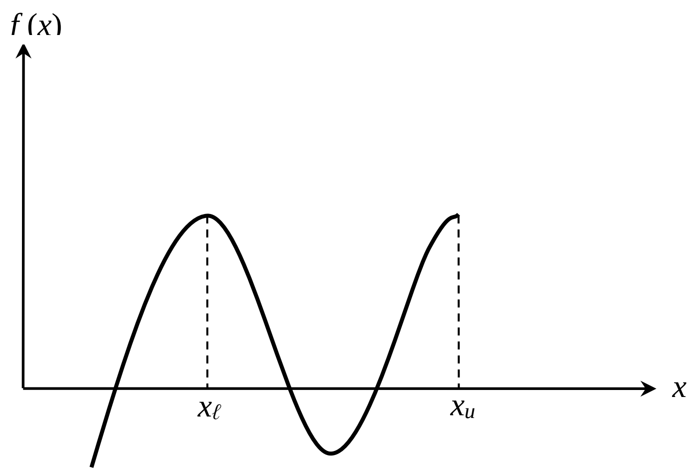

Figure 2 If the function $f(x)$ does not change sign between the two points, roots of the equation $f(x) = 0$ may still exist between the two points.

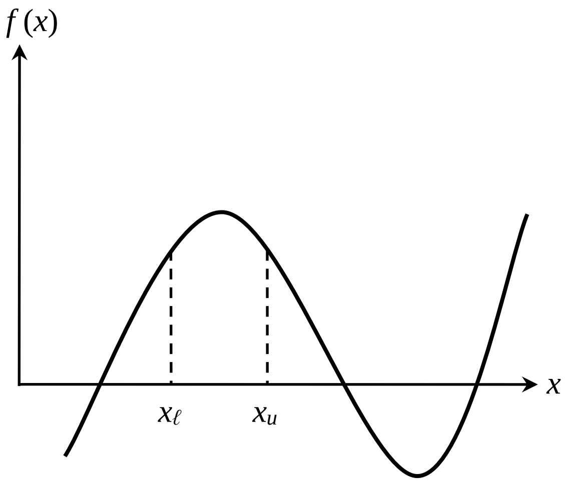

Figure 3 If the function $f(x)$ does not change sign between two points, there may not be any roots for the equation $f(x)=0$ between the two points.

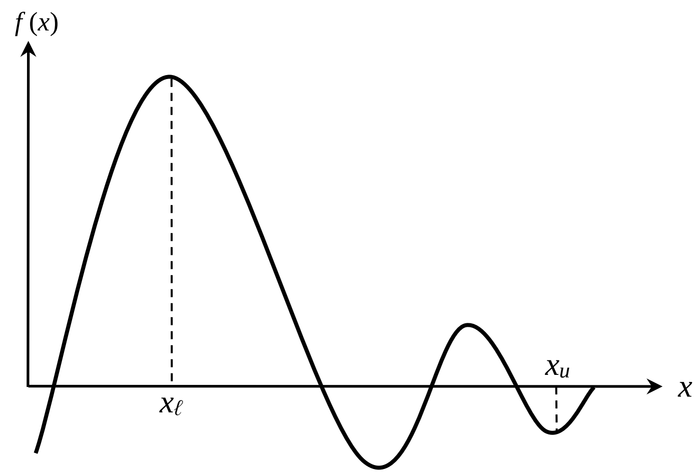

Figure 4 If the function $f(x)$ changes sign between the two points, more than one root for the equation $f(x)=0$ may exist between the two points.

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 from present iteration

   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. 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

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

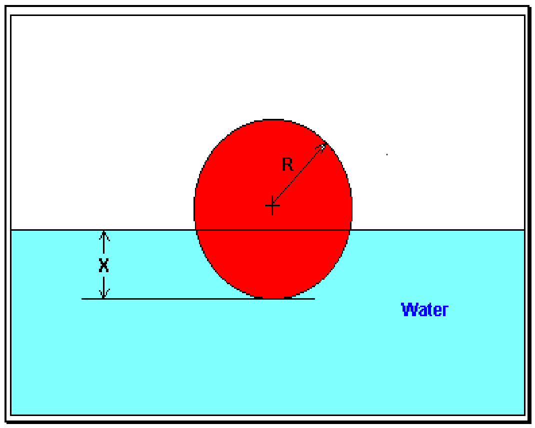

Figure 5 Floating ball problem.

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

$$x_m = \frac{x_\ell + x_u}{2}$$
$$= \frac{0 + 0.11}{2}$$
$$= 0.055$$

$$f(x_m) = f(0.055) = (0.055)^3 - 0.165(0.055)^2 + 3.993\times10^{-4} = 6.655\times10^{-5}$$
$$f(x_\ell)f(x_m) = f(0)f(0.055) = (3.993\times10^{-4})(6.655\times10^{-4}) > 0$$

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.

<u>Iteration 2</u>

The estimate of the root is

$$x_m = \frac{x_\ell + x_u}{2}$$
$$= \frac{0.055 + 0.11}{2}$$
$$= 0.0825$$

$$f(x_m) = f(0.0825) = (0.0825)^3 - 0.165(0.0825)^2 + 3.993\times10^{-4} = -1.622\times10^{-4}$$

$$f(x_\ell)f(x_m) = f(0.055)f(0.0825) = \left(6.655\times10^{-5}\right)\times\left(-1.622\times10^{-4}\right) < 0$$

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

$$|\in_a| = \left|\frac{x_m^{\text{new}} - x_m^{\text{old}}}{x_m^{\text{new}}}\right|\times100$$
$$= \left|\frac{0.0825 - 0.055}{0.0825}\right|\times100$$
$$= 33.33\%$$

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%.

<u>Iteration 3</u>

$$x_m = \frac{x_\ell + x_u}{2}$$
$$= \frac{0.055 + 0.0825}{2}$$
$$= 0.06875$$

$$f(x_m) = f(0.06875) = (0.06875)^3 - 0.165(0.06875)^2 + 3.993\times10^{-4} = -5.563\times10^{-5}$$

$$f(x_\ell)f(x_m) = f(0.055)f(0.06875) = (6.655\times10^{5})\times(-5.563\times10^{-5}) < 0$$

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,

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

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

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

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

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

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

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

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.

Advantages of bisection method

a) The bisection method is always convergent. Since the method brackets the root, the method is guaranteed to converge.
b) As iterations are conducted, the interval gets halved. So one can guarantee the error in the solution of the equation.

Drawbacks of bisection method

a) The convergence of the bisection method is slow as it is simply based on halving the interval.

b) If one of the initial guesses is closer to the root, it will take larger number of iterations to reach the root.

c) If a function $f(x)$ is such that it just touches the x-axis (Figure 6) such as

$$f(x) = x^2 = 0$$

it will be unable to find the lower guess, x_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.

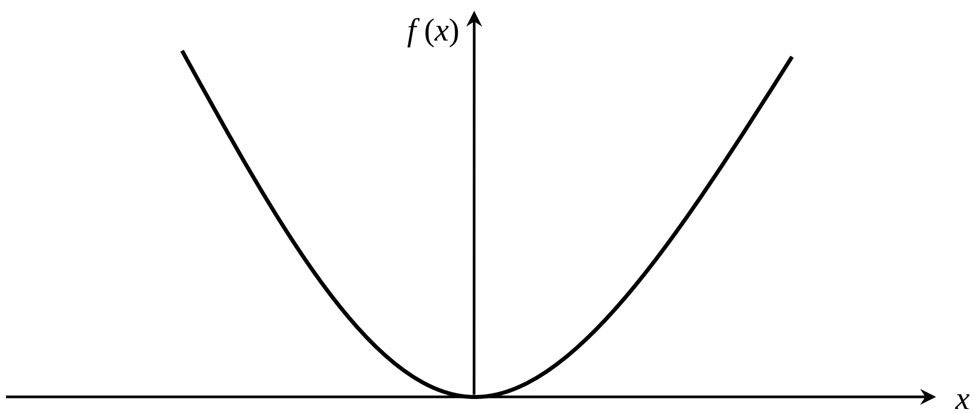

Figure 6 The equation $f(x) = x^2 = 0$ has a single root at $x = 0$ that cannot be bracketed.

[1] 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.

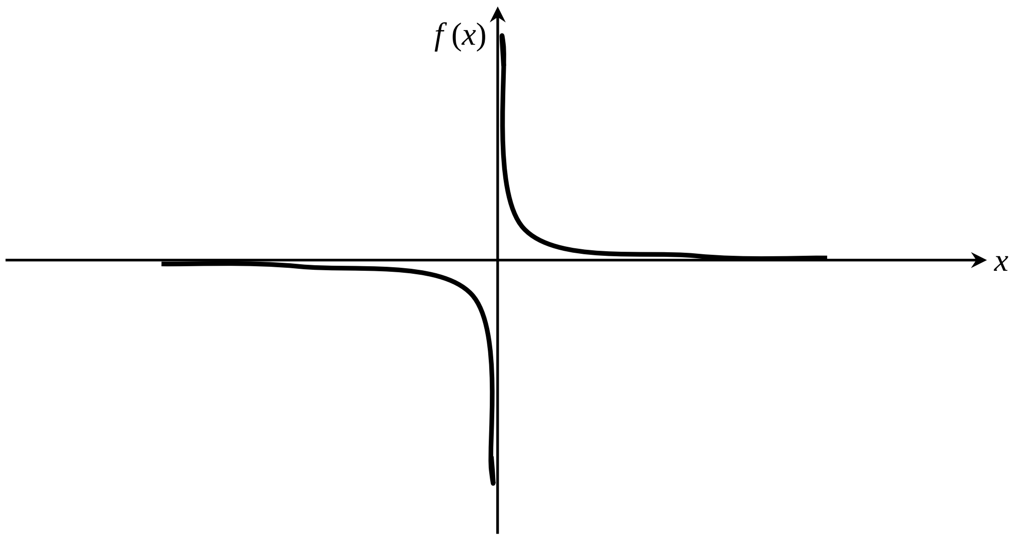

Figure 7 The equation $f(x)=\frac{1}{x}=0$ has no root but changes sign.

Reference

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	March 5, 2022
Web Site	http://numericalmethods.eng.usf.edu

False-Position Method of Solving a Nonlinear Equation

Introduction

In the previous lecture, the bisection method was described as one of the simple bracketing methods of solving a nonlinear equation of the general form

$$f(x) = 0 \quad (1)$$

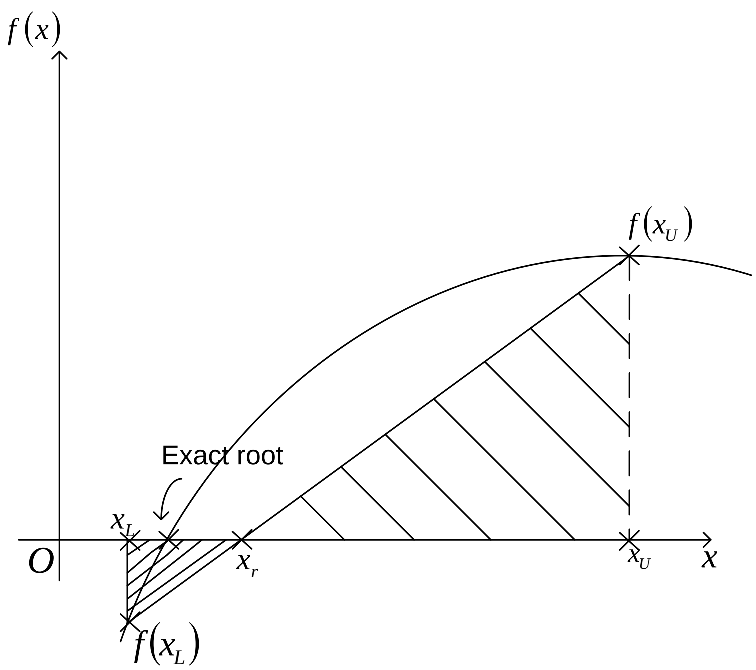

Figure 1 False-Position Method

The above nonlinear equation can be stated as finding the value of *x* such that Equation (1) is satisfied.

In the bisection method, we identify proper values of x_L (lower bound value) and x_U (upper bound value) for the current bracket, such that

$$f(x_L)\, f(x_U) < 0 . \quad (2)$$

The next predicted/improved root x_r can be computed as the midpoint between x_L and x_U as

$$x_r = \frac{x_L + x_U}{2} \quad (3)$$

The new upper and lower bounds are then established, and the procedure is repeated until the convergence is achieved (such that the new lower and upper bounds are sufficiently close to each other).

However, in the example shown in Figure 1, the bisection method may not be efficient because it does not take into consideration that $f(x_L)$ is much closer to the zero of the function $f(x)$ as compared to $f(x_U)$. In other words, the next predicted root x_r would be closer to x_L (in the example as shown in Figure 1), than the mid-point between x_L and x_U. The false-position method takes advantage of this observation mathematically by drawing a secant from the function value at x_L to the function value at x_U, and estimates the root as where it crosses the *x*-axis.

False-Position Method

Based on two similar triangles, shown in Figure 1, one gets

$$\frac{0-f(x_L)}{x_r-x_L}=\frac{0-f(x_U)}{x_r-x_U} \tag{4}$$

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

$$x_r=x_U-\frac{f(x_U)}{\left\{\frac{f(x_L)-f(x_U)}{x_L-x_U}\right\}} \tag{6}$$

or

$$x_r=x_L-\frac{f(x_L)}{\left\{\frac{f(x_U)-f(x_L)}{x_U-x_L}\right\}} \tag{7}$$

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

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

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.

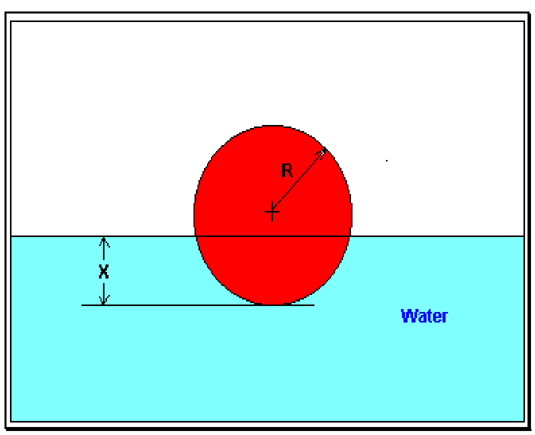

Figure 2 Floating ball problem.

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

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

Hence

$$f(x_L)f(x_U) = f(0)f(0.11) = (3.993\times10^{-4})(-2.662\times10^{-4}) < 0$$

Therefore, there is at least one root between x_L and x_U, that is between 0 and 0.11.

Iteration 1

The estimate of the root is

$$x_r = \frac{x_U f(x_L) - x_L f(x_U)}{f(x_L) - f(x_U)}$$
$$= \frac{0.11\times3.993\times10^{-4} - 0\times(-2.662\times10^{-4})}{3.993\times10^{-4} - (-2.662\times10^{-4})}$$
$$= 0.0660$$

$$f(x_r) = f(0.0660)$$
$$= (0.0660)^3 - 0.165(0.0660)^2 + (3.993\times10^{-4})$$
$$= -3.1944\times10^{-5}$$
$$f(x_L)f(x_r) = f(0)f(0.0660) = (+)(-) < 0$$

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

$$x_r = \frac{x_U f(x_L) - x_L f(x_U)}{f(x_L) - f(x_U)}$$
$$= \frac{0.0660\times3.993\times10^{-4} - 0\times(-3.1944\times10^{-5})}{3.993\times10^{-4} - (-3.1944\times10^{-5})}$$
$$= 0.0611$$

The absolute relative approximate error for this iteration is

$$\in_a = \left|\frac{0.0611 - 0.0660}{0.0611}\right| \times 100 \cong 8\%$$

$$f(x_r) = f(0.0611)$$
$$= (0.0611)^3 - 0.165(0.0611)^2 + (3.993\times10^{-4})$$
$$= 1.1320\times10^{-5}$$
$$f(x_L)f(x_r) = f(0)f(0.0611) = (+)(+) > 0$$

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

$$x_r = \frac{x_U f(x_L) - x_L f(x_U)}{f(x_L) - f(x_U)}$$
$$= \frac{0.0660\times1.132\times10^{-5} - 0.0611\times(-3.1944\times10^{-5})}{1.132\times10^{-5} - (-3.1944\times10^{-5})}$$
$$= 0.0624$$

The absolute relative approximate error for this iteration is

$$\in_a = \left|\frac{0.0624 - 0.0611}{0.0624}\right| \times 100 \cong 2.05\%$$
$$f(x_r) = -1.1313\times10^{-7}$$
$$f(x_L)f(x_r) = f(0.0611)f(0.0624) = (+)(-) < 0$$

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

$$|\in_a| \le 0.5\times10^{2-m}$$
$$2.05 \le 0.5\times10^{2-m}$$
$$m \le 1.387$$

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 $f(x) = x^3 - 0.165x^2 + 3.993\times10^{-4} = 0$ for false-position method.

Iteration	x_L	x_U	x_r	$\|\in_a\|\%$	$f(x_m)$
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,

$$\lvert\in_a\rvert \le 0.5\times 10^{2-m}$$

$$0.02706 \le 0.5\times 10^{2-m}$$

$$m \le 3.2666$$

Hence, at least 3 significant digits can be trusted to be accurate at the end of the fifth iteration.

Reference

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	March 6, 2022

Secant Method of Solving Nonlinear Equations

What is the secant method and why would I want to use it instead of the Newton-Raphson method?

The Newton-Raphson method of solving a nonlinear equation $f(x)=0$ is given by the iterative formula

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

One of the drawbacks of the Newton-Raphson method is that you have to evaluate the derivative of the function. With availability of symbolic manipulators such as Maple, MathCAD, MATHEMATICA and MATLAB, this process has become more convenient. However, it still can be a laborious process, and even intractable if the function is derived as part of a numerical scheme. To overcome these drawbacks, the derivative of the function, $f(x)$ is approximated as

$$f'(x_i)=\frac{f(x_i)-f(x_{i-1})}{x_i-x_{i-1}} \quad (2)$$

Substituting Equation (2) in Equation (1) gives

$$x_{i+1}=x_i-\frac{f(x_i)(x_i-x_{i-1})}{f(x_i)-f(x_{i-1})} \quad (3)$$

The above equation is called the secant method. This method now requires two initial guesses, but unlike the bisection method, the two initial guesses do not need to bracket the root of the equation. The secant method is an open method and may or may not converge. However, when secant method converges, it will typically converge faster than the bisection method. However, since the derivative is approximated as given by Equation (2), it typically converges slower than the Newton-Raphson method.

The secant method can also be derived from geometry, as shown in Figure 1. Taking two initial guesses, x_{i-1} and x_i, one draws a straight line between $f(x_i)$ and $f(x_{i-1})$ passing through the x-axis at x_{i+1}. ABE and DCE are similar triangles.

Hence

$$\frac{AB}{AE}=\frac{DC}{DE}$$

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

On rearranging, the secant method is given as

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

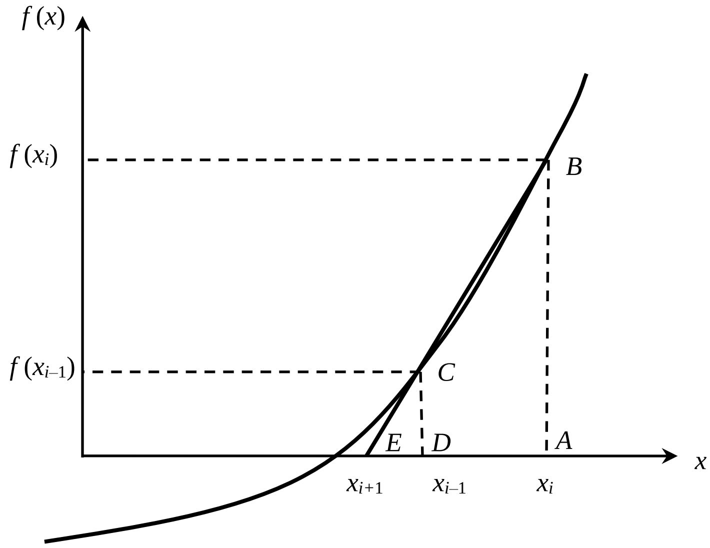

Figure 1 Geometrical representation of the secant method.

Example 1

You are working for 'DOWN THE TOILET COMPANY' that makes floats (Figure 2) for ABC commodes. The floating ball has a specific gravity of 0.6 and a radius of 5.5 cm. You are asked to find the depth to which the ball is submerged when floating in water.

The equation that gives the depth x to which the ball is submerged under water is given by

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

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 $x_0 = 0.05$.

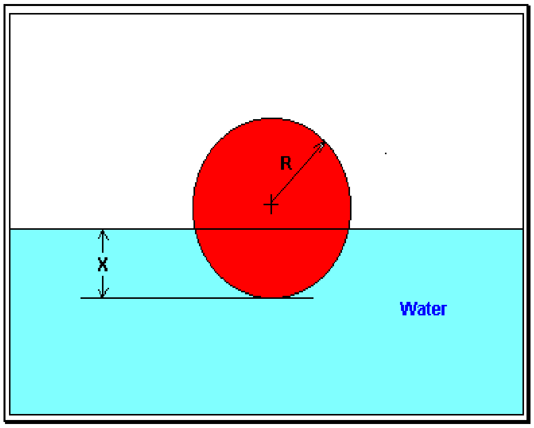

Figure 2 Floating ball problem.

Iteration 1

The estimate of the root is

$$x_1 = x_0 - \frac{f(x_0)(x_0 - x_{-1})}{f(x_0) - f(x_{-1})}$$

$$= x_0 - \frac{\left(x_0^3 - 0.165x_0^2 + 3.993\times10^{-4}\right)\times(x_0 - x_{-1})}{\left(x_0^3 - 0.165x_0^2 + 3.993\times10^{-4}\right) - \left(x_{-1}^3 - 0.165x_{-1}^2 + 3.993\times10^{-4}\right)}$$

$$= 0.05 - \frac{\left[0.05^3 - 0.165(0.05)^2 + 3.993\times10^{-4}\right]\times[0.05 - 0.02]}{\left[0.05^3 - 0.165(0.05)^2 + 3.993\times10^{-4}\right] - \left[0.02^3 - 0.165(0.02)^2 + 3.993\times10^{-4}\right]}$$

$$= 0.06461$$

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

$$x_2 = x_1 - \frac{f(x_1)(x_1 - x_0)}{f(x_1) - f(x_0)}$$

$$= x_1 - \frac{\left(x_1^3 - 0.165x_1^2 + 3.993\times10^{-4}\right)\times(x_1 - x_0)}{\left(x_1^3 - 0.165x_1^2 + 3.993\times10^{-4}\right) - \left(x_0^3 - 0.165x_0^2 + 3.993\times10^{-4}\right)}$$

$$= 0.06461 - \frac{\left[0.06461^3 - 0.165(0.06461)^2 + 3.993\times10^{-4}\right]\times(0.06461 - 0.05)}{\left[0.06461^3 - 0.165(0.06461)^2 + 3.993\times10^{-4}\right] - \left[0.05^3 - 0.165(0.05)^2 + 3.993\times10^{-4}\right]}$$

$$= 0.06241$$

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

$$|\in_a| = \left|\frac{x_2 - x_1}{x_2}\right|\times100$$

$$= \left|\frac{0.06241 - 0.06461}{0.06241}\right|\times100$$

$$= 3.525\%$$

The number of significant digits at least correct is 1, as you need an absolute relative approximate error of 5% or less.

<u>Iteration 3</u>

$$x_3 = x_2 - \frac{f(x_2)(x_2 - x_1)}{f(x_2) - f(x_1)}$$

$$= x_2 - \frac{\left(x_2^3 - 0.165x_2^2 + 3.993\times10^{-4}\right)\times(x_2 - x_1)}{\left(x_2^3 - 0.165x_2^2 + 3.993\times10^{-4}\right) - \left(x_1^3 - 0.165x_1^2 + 3.993\times10^{-4}\right)}$$

$$= 0.06241 - \frac{\left[0.06241^3 - 0.165(0.06241)^2 + 3.993\times10^{-4}\right]\times(0.06241 - 0.06461)}{\left[0.06241^3 - 0.165(0.06241)^2 + 3.993\times10^{-4}\right] - \left[0.06461^3 - 0.165(0.06461)^2 + 3.993\times10^{-4}\right]}$$

$$= 0.06238$$

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

$$|\in_a| = \left|\frac{x_3 - x_2}{x_3}\right|\times100$$

$$= \left|\frac{0.06238 - 0.06241}{0.06238}\right|\times100$$

$$= 0.0595\%$$

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}	$\lvert\in_a\rvert\%$	$f(x_{i+1})$
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}

Reference

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	March 11, 2022

Newton-Raphson Method of Solving a Nonlinear Equation

Introduction

Methods such as the bisection method and the false position method of finding roots of a nonlinear equation $f(x)=0$ require bracketing of the root by two guesses. Such methods are called *bracketing methods*. These methods are always convergent since they are based on reducing the interval between the two guesses so as to zero in on the root of the equation.

In the Newton-Raphson method, the root is not bracketed. In fact, only one initial guess of the root is needed to get the iterative process started to find the root of an equation. The method hence falls in the category of *open methods*. Convergence in open methods is not guaranteed but if the method does converge, it does so much faster than the bracketing methods.

Derivation

The Newton-Raphson method is based on the principle that if the initial guess of the root of $f(x)=0$ is at x_i, then if one draws the tangent to the curve at $f(x_i)$, the point x_{i+1} where the tangent crosses the x-axis is an improved estimate of the root (Figure 1).

Using the definition of the slope of a function, at $x=x_i$

$$f'(x_i) = \tan\theta$$

$$= \frac{f(x_i)-0}{x_i - x_{i+1}},$$

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.

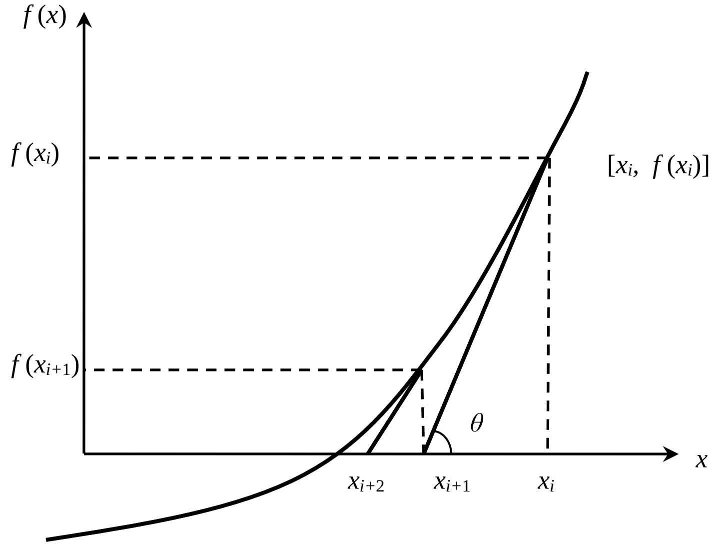

Figure 1 Geometrical illustration of the Newton-Raphson method.

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.

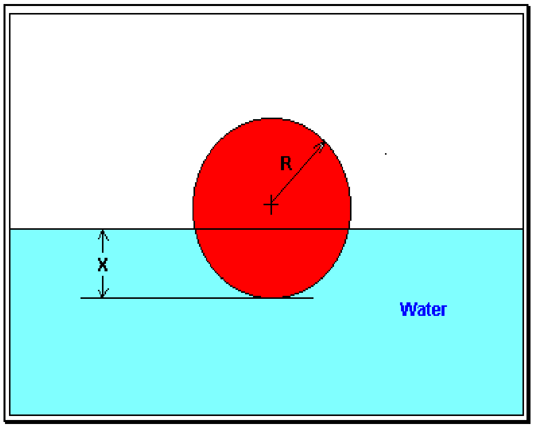

Figure 2 Floating ball problem.

The equation that gives the depth x in meters to which the ball is submerged under water is given by

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

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

$$f(x) = x^3 - 0.165x^2 + 3.993\times10^{-4}$$
$$f'(x) = 3x^2 - 0.33x$$

Let us assume the initial guess of the root of $f(x)=0$ is $x_0 = 0.05$ m. This is a reasonable guess (discuss why $x=0$ and $x=0.11$m are not good choices) as the extreme values of the depth x would be 0 and the diameter (0.11 m) of the ball.

<u>Iteration 1</u>

The estimate of the root is

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$
$$= 0.05 - \frac{(0.05)^3 - 0.165(0.05)^2 + 3.993\times10^{-4}}{3(0.05)^2 - 0.33(0.05)}$$
$$= 0.05 - \frac{1.118\times10^{-4}}{-9\times10^{-3}}$$
$$= 0.05 - (-0.01242)$$
$$= 0.06242$$

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

$$|\in_a| = \left|\frac{x_1 - x_0}{x_1}\right|\times100$$

$$= \left| \frac{0.06242 - 0.05}{0.06242} \right| \times 100$$
$$= 19.90\%$$

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.

<u>Iteration 2</u>

The estimate of the root is

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}$$
$$= 0.06242 - \frac{(0.06242)^3 - 0.165(0.06242)^2 + 3.993 \times 10^{-4}}{3(0.06242)^2 - 0.33(0.06242)}$$
$$= 0.06242 - \frac{-3.97781 \times 10^{-7}}{-8.90973 \times 10^{-3}}$$
$$= 0.06242 - \left(4.4646 \times 10^{-5}\right)$$
$$= 0.06238$$

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

$$|\in_a| = \left| \frac{x_2 - x_1}{x_2} \right| \times 100$$
$$= \left| \frac{0.06238 - 0.06242}{0.06238} \right| \times 100$$
$$= 0.0716\%$$

The maximum value of m for which $|\in_a| \leq 0.5 \times 10^{2-m}$ is 2.844. Hence, the number of significant digits at least correct in the answer is 2.

<u>Iteration 3</u>

The estimate of the root is

$$x_3 = x_2 - \frac{f(x_2)}{f'(x_2)}$$
$$= 0.06238 - \frac{(0.06238)^3 - 0.165(0.06238)^2 + 3.993 \times 10^{-4}}{3(0.06238)^2 - 0.33(0.06238)}$$
$$= 0.06238 - \frac{4.44 \times 10^{-11}}{-8.91171 \times 10^{-3}}$$
$$= 0.06238 - \left(-4.9822 \times 10^{-9}\right)$$
$$= 0.06238$$

The absolute relative approximate error $|\in_a|$ 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

If the selection of the initial guess or an iterated value of the root turns out to be close to the inflection point (see the definition in the appendix of this chapter) of the function $f(x)$ in the equation $f(x)=0$, Newton-Raphson method may start diverging away from the root. It may then start converging back to the root. For example, to find the root of the equation

$$f(x)=(x-1)^3+0.512=0$$

the Newton-Raphson method reduces to

$$x_{i+1}=x_i-\frac{(x_i^3-1)^3+0.512}{3(x_i-1)^2}$$

Starting with an initial guess of $x_0=5.0$, Table 1 shows the iterated values of the root of the equation. As you can observe, the root starts to diverge at Iteration 6 because the previous estimate of 0.92589 is close to the inflection point of $x=1$ (the value of $f'(x)$ is zero at the inflection point). Eventually, after 12 more iterations the root converges to the exact value of $x=0.2$.

Table 1 Divergence near inflection point.

Iteration Number	x_i
0	5.0000
1	3.6560
2	2.7465
3	2.1084
4	1.6000
5	0.92589
6	–30.119
7	–19.746
8	–12.831
9	–8.2217
10	–5.1498
11	–3.1044
12	–1.7464
13	–0.85356
14	–0.28538
15	0.039784
16	0.17475
17	0.19924
18	0.2

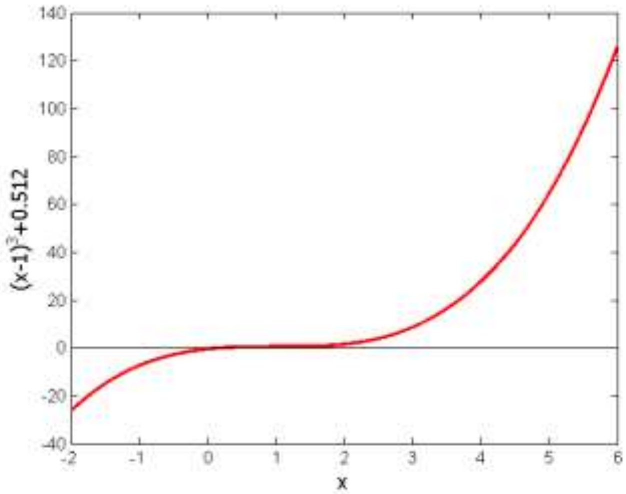

Figure 3 Divergence at inflection point for $f(x)=(x-1)^3=0$.

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 in the formula is a small number. For this case, as given in Table 2, even after 9 iterations, the Newton-Raphson method does not converge.

Table 2 Division by near zero in Newton-Raphson method.

Iteration Number	x_i	$f(x_i)$	$\lvert\in_a\rvert\%$
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

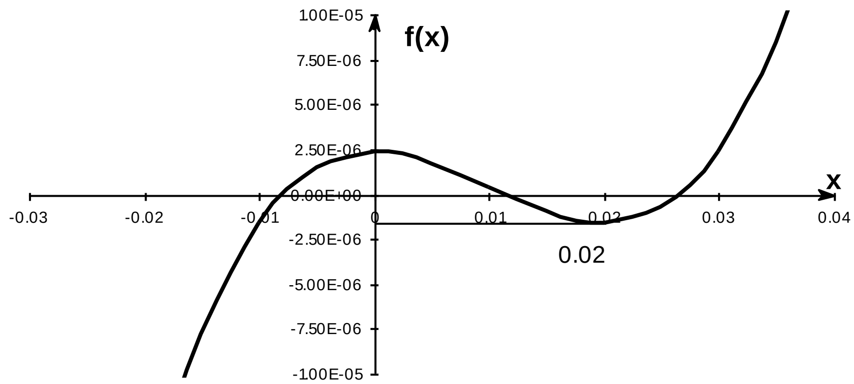

Figure 4 Pitfall of division by zero or a near zero number.

3. Oscillations near local maximum and minimum

Results obtained from the Newton-Raphson method may oscillate about the local maximum or minimum without converging on a root but converging on the local maximum or minimum. Eventually, it may lead to division by a number close to zero and may diverge. For example, for

$$f(x) = x^2 + 2 = 0$$

the equation has no real roots (Figure 5 and Table 3).

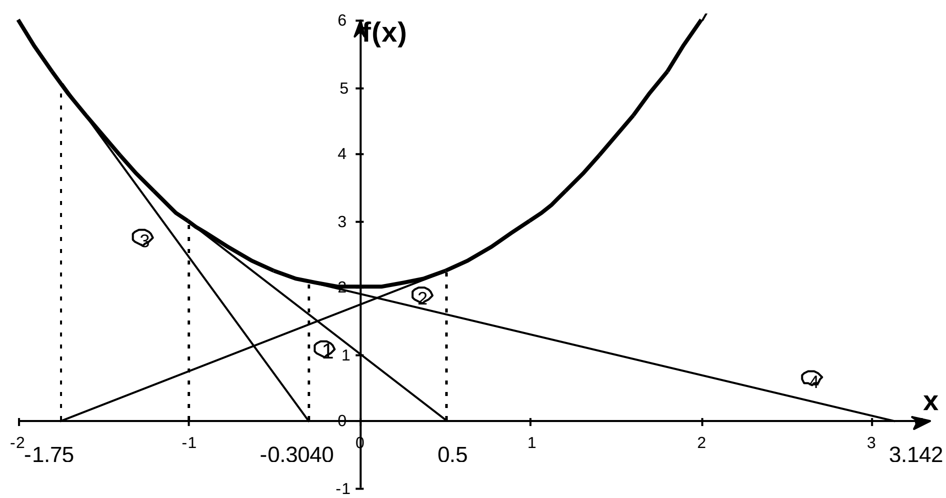

Figure 5 Oscillations around local minima for $f(x) = x^2 + 2$.

Table 3 Oscillations near local maxima and minima in Newton-Raphson method.

Iteration Number	x_i	$f(x_i)$	$\lvert\in_a\rvert\%$
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

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}

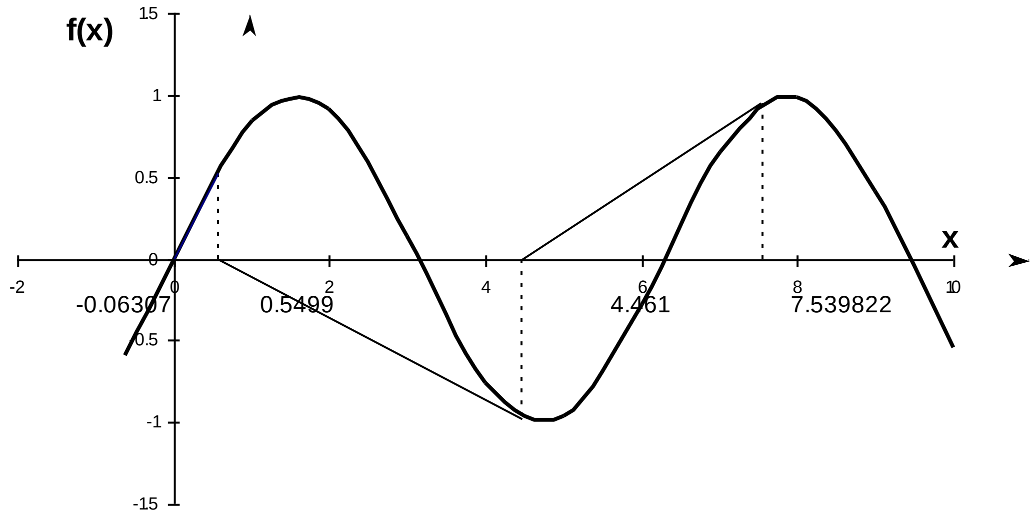

Figure 6 Root jumping from intended location of root for $f(x) = \sin x = 0$.

Appendix A. What is an inflection point?

For a function $f(x)$, the point where the concavity changes from up-to-down or down-to-up is called its inflection point. For example, for the function $f(x) = (x-1)^3$, the concavity changes at $x = 1$ (see Figure 3), and hence (1,0) is an inflection point.

An inflection points MAY exist at a point where $f''(x) = 0$ and where $f''(x)$ does not exist. The reason we say that it MAY exist is because if $f''(x) = 0$, it only makes it a possible inflection point. For example, for $f(x) = x^4 - 16$, $f''(0) = 0$, but the concavity does not change at $x = 0$. Hence the point (0, –16) is not an inflection point of $f(x) = x^4 - 16$.

For $f(x) = (x-1)^3$, $f''(x)$ changes sign at $x = 1$ ($f''(x) < 0$ for $x < 1$, and $f''(x) > 0$ for $x > 1$), and thus brings up the *Inflection Point Theorem* for a function $f(x)$ that states the following.

"If $f'(c)$ exists and $f''(c)$ changes sign at $x = c$, then the point $(c, f(c))$ is an inflection point of the graph of f."

Appendix B. Derivation of Newton-Raphson method from Taylor series

Newton-Raphson method can also be derived from Taylor series. For a general function $f(x)$, the Taylor series is

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

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.

Reference

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	March 11, 2022

Trapezoidal Rule of Integration

What is integration?

Integration is the process of measuring the area under a function plotted on a graph. Why would we want to integrate a function? Among the most common examples are finding the velocity of a body from an acceleration function, and displacement of a body from a velocity function. Throughout many engineering fields, there are (what sometimes seems like) countless applications for integral calculus. You can read about some of these applications in Chapters 07.00A-07.00G.

Sometimes, the evaluation of expressions involving these integrals can become daunting, if not indeterminate. For this reason, a wide variety of numerical methods has been developed to simplify the integral.

Here, we will discuss the trapezoidal rule of approximating integrals of the form

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

where

$f(x)$ is called the integrand,

$a =$ lower limit of integration

$b =$ upper limit of integration

What is the trapezoidal rule?

The trapezoidal rule is based on the Newton-Cotes formula that if one approximates the integrand by an n^{th} order polynomial, then the integral of the function is approximated by the integral of that n^{th} order polynomial. Integrating polynomials is simple and is based on the calculus formula.

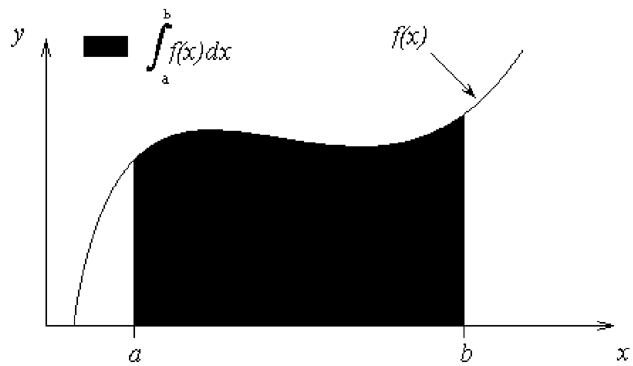

Figure 1 Integration of a function

$$\int_a^b x^n dx = \left(\frac{b^{n+1} - a^{n+1}}{n+1}\right), \; n \neq -1 \tag{1}$$

So if we want to approximate the integral

$$I = \int_a^b f(x)dx \tag{2}$$

to find the value of the above integral, one assumes

$$f(x) \approx f_n(x) \tag{3}$$

where

$$f_n(x) = a_0 + a_1 x + \ldots + a_{n-1}x^{n-1} + a_n x^n. \tag{4}$$

where $f_n(x)$ is a n^{th} order polynomial. The trapezoidal rule assumes $n = 1$, that is, approximating the integral by a linear polynomial (straight line),

$$\int_a^b f(x)dx \approx \int_a^b f_1(x)dx$$

Derivation of the Trapezoidal Rule

Method 1: Derived from Calculus

$$\int_a^b f(x)dx \approx \int_a^b f_1(x)dx$$

$$= \int_a^b (a_0 + a_1 x)dx$$

$$= a_0(b-a) + a_1\left(\frac{b^2 - a^2}{2}\right) \tag{5}$$

But what is a_0 and a_1? Now if one chooses, $(a, f(a))$ and $(b, f(b))$ as the two points to approximate $f(x)$ by a straight line from a to b,

$$f(a) = f_1(a) = a_0 + a_1 a \tag{6}$$

$$f(b) = f_1(b) = a_0 + a_1 b \tag{7}$$

Solving the above two equations for a_1 and a_0,

$$a_1 = \frac{f(b) - f(a)}{b - a}$$

$$a_0 = \frac{f(a)b - f(b)a}{b - a} \tag{8a}$$

Hence from Equation (5),

$$\int_a^b f(x)dx \approx \frac{f(a)b - f(b)a}{b - a}(b - a) + \frac{f(b) - f(a)}{b - a}\frac{b^2 - a^2}{2} \tag{8b}$$

$$= (b - a)\left[\frac{f(a) + f(b)}{2}\right] \tag{9}$$

<u>Method 2: Also Derived from Calculus</u>

$f_1(x)$ can also be approximated by using Newton's divided difference polynomial as

$$f_1(x) = f(a) + \frac{f(b) - f(a)}{b - a}(x - a) \tag{10}$$

Hence

$$\int_a^b f(x)dx \approx \int_a^b f_1(x)dx$$

$$= \int_a^b \left[f(a) + \frac{f(b) - f(a)}{b - a}(x - a)\right]dx$$

$$= \left[f(a)x + \frac{f(b) - f(a)}{b - a}\left(\frac{x^2}{2} - ax\right)\right]_a^b$$

$$= f(a)b - f(a)a + \left(\frac{f(b) - f(a)}{b - a}\right)\left(\frac{b^2}{2} - ab - \frac{a^2}{2} + a^2\right)$$

$$= f(a)b - f(a)a + \left(\frac{f(b) - f(a)}{b - a}\right)\left(\frac{b^2}{2} - ab + \frac{a^2}{2}\right)$$

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

$$= f(a)b - f(a)a + \frac{1}{2}\left(f(b) - f(a)\right)(b - a)$$

$$= f(a)b - f(a)a + \frac{1}{2}f(b)b - \frac{1}{2}f(b)a - \frac{1}{2}f(a)b + \frac{1}{2}f(a)a$$

$$= \frac{1}{2}f(a)b - \frac{1}{2}f(a)a + \frac{1}{2}f(b)b - \frac{1}{2}f(b)a$$

$$= (b-a)\left[\frac{f(a)+f(b)}{2}\right] \tag{11}$$

This gives the same result as Equation (10) because they are just different forms of writing the same polynomial.

Method 3: Derived from Geometry

The trapezoidal rule can also be derived from geometry. Look at Figure 2. The area under the curve $f_1(x)$ is the area of a trapezoid. The integral

$$\int_a^b f(x)dx \approx \text{Area of trapezoid}$$

$$= \frac{1}{2}(\text{Sum of length of parallel sides})(\text{Perpendicular distance between parallel sides})$$

$$= \frac{1}{2}\left(f(b)+f(a)\right)(b-a)$$

$$= (b-a)\left[\frac{f(a)+f(b)}{2}\right] \tag{12}$$

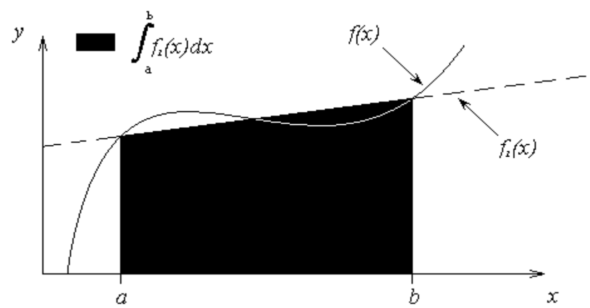

Figure 2 Geometric representation of trapezoidal rule.

Method 4: Derived from Method of Coefficients

The trapezoidal rule can also be derived by the method of coefficients. The formula

$$\int_a^b f(x)dx \approx \frac{b-a}{2}f(a)+\frac{b-a}{2}f(b) \tag{13}$$

$$= \sum_{i=1}^{2} c_i f(x_i)$$

where

$$c_1 = \frac{b-a}{2}$$

$$c_2 = \frac{b-a}{2}$$

$$x_1 = a$$

$$x_2 = b$$

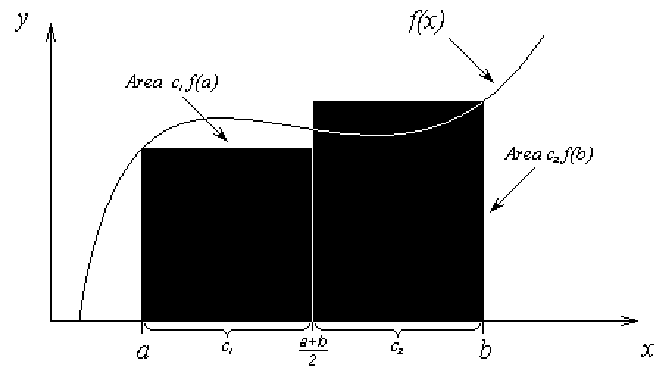

Figure 3 Area by method of coefficients.

The interpretation is that $f(x)$ is evaluated at points a and b, and each function evaluation is given a weight of $\frac{b-a}{2}$. Geometrically, Equation (12) is looked at as the area of a trapezoid, while Equation (13) is viewed as the sum of the area of two rectangles, as shown in Figure 3. How can one derive the trapezoidal rule by the method of coefficients?

Assume

$$\int_a^b f(x)dx = c_1 f(a) + c_2 f(b) \tag{14}$$

Let the right hand side be an exact expression for integrals of $\int_a^b 1dx$ and $\int_a^b xdx$, that is, the formula will then also be exact for linear combinations of $f(x)=1$ and $f(x)=x$, that is, for $f(x) = a_0(1) + a_1(x)$.

$$\int_a^b 1dx = b - a = c_1 + c_2 \tag{15}$$

$$\int_a^b xdx = \frac{b^2 - a^2}{2} = c_1 a + c_2 b \tag{16}$$

Solving the above two equations gives

$$c_1 = \frac{b-a}{2}$$

$$c_2 = \frac{b-a}{2} \tag{17}$$

Hence

$$\int_a^b f(x)dx \approx \frac{b-a}{2} f(a) + \frac{b-a}{2} f(b) \tag{18}$$

Method 5: Another approach on the Method of Coefficients

The trapezoidal rule can also be derived by the method of coefficients by another approach

$$\int_a^b f(x)dx \approx \frac{b-a}{2} f(a) + \frac{b-a}{2} f(b)$$

Assume

$$\int_a^b f(x)dx = c_1 f(a) + c_2 f(b) \quad (19)$$

Let the right hand side be exact for integrals of the form

$$\int_a^b (a_0 + a_1 x)dx$$

So

$$\int_a^b (a_0 + a_1 x)dx = \left(a_0 x + a_1 \frac{x^2}{2} \right)_a^b$$

$$= a_0 (b-a) + a_1 \left(\frac{b^2 - a^2}{2} \right) \quad (20)$$

But we want

$$\int_a^b (a_0 + a_1 x)dx = c_1 f(a) + c_2 f(b) \quad (21)$$

to give the same result as Equation (20) for $f(x) = a_0 + a_1 x$.

$$\int_a^b (a_0 + a_1 x)dx = c_1 (a_0 + a_1 a) + c_2 (a_0 + a_1 b)$$

$$= a_0 (c_1 + c_2) + a_1 (c_1 a + c_2 b) \quad (22)$$

Hence from Equations (20) and (22),

$$a_0 (b-a) + a_1 \left(\frac{b^2 - a^2}{2} \right) = a_0 (c_1 + c_2) + a_1 (c_1 a + c_2 b)$$

Since a_0 and a_1 are arbitrary for a general straight line

$$c_1 + c_2 = b - a$$

$$c_1 a + c_2 b = \frac{b^2 - a^2}{2} \quad (23)$$

Again, solving the above two equations (23) gives

$$c_1 = \frac{b-a}{2}$$

$$c_2 = \frac{b-a}{2} \quad (24)$$

Therefore

$$\int_a^b f(x)dx \approx c_1 f(a) + c_2 f(b)$$

$$= \frac{b-a}{2} f(a) + \frac{b-a}{2} f(b) \quad (25)$$

Example 1

The vertical distance covered by a rocket from $t = 8$ to $t = 30$ seconds is given by

$$x = \int_8^{30} \left(2000 \ln \left[\frac{140000}{140000 - 2100t} \right] - 9.8t \right) dt$$

a) Use the single segment trapezoidal rule to find the distance covered for $t = 8$ to $t = 30$ seconds.

b) Find the true error, E_t for part (a).

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

Solution

a) $I \approx (b-a)\left[\frac{f(a) + f(b)}{2} \right]$, where

$a = 8$

$b = 30$

$$f(t) = 2000 \ln \left[\frac{140000}{140000 - 2100t} \right] - 9.8t$$

$$f(8) = 2000 \ln \left[\frac{140000}{140000 - 2100(8)} \right] - 9.8(8)$$

$$= 177.27 \text{ m/s}$$

$$f(30) = 2000 \ln \left[\frac{140000}{140000 - 2100(30)} \right] - 9.8(30)$$

$$= 901.67 \text{ m/s}$$

$$I \approx (30 - 8)\left[\frac{177.27 + 901.67}{2} \right]$$

$$= 11868 \text{ m}$$

b) The exact value of the above integral is

$$x = \int_8^{30} \left(2000 \ln \left[\frac{140000}{140000 - 2100t} \right] - 9.8t \right) dt$$

$$= 11061 \text{ m}$$

so the true error is

$E_t =$ True Value – Approximate Value

$= 11061 - 11868$

$= -807$ m

c) The absolute relative true error, $|\in_t|$, would then be

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

$$= \left|\frac{11061 - 11868}{11061}\right| \times 100$$

$$= 7.2958\%$$

Multiple-Segment Trapezoidal Rule

In Example 1, the true error using a single segment trapezoidal rule was large. We can divide the interval $[8,30]$ into $[8,19]$ and $[19,30]$ intervals and apply the trapezoidal rule over each segment.

$$f(t) = 2000\ln\left(\frac{140000}{140000 - 2100t}\right) - 9.8t$$

$$\int_8^{30} f(t)dt = \int_8^{19} f(t)dt + \int_{19}^{30} f(t)dt$$

$$\approx (19-8)\left[\frac{f(8) + f(19)}{2}\right] + (30-19)\left[\frac{f(19) + f(30)}{2}\right]$$

$$f(8) = 177.27 \text{ m/s}$$

$$f(19) = 2000\ln\left(\frac{140000}{140000 - 2100(19)}\right) - 9.8(19) = 484.75 \text{ m/s}$$

$$f(30) = 901.67 \text{ m/s}$$

Hence

$$\int_8^{30} f(t)dt \approx (19-8)\left[\frac{177.27 + 484.75}{2}\right] + (30-19)\left[\frac{484.75 + 901.67}{2}\right]$$

$$= 11266 \text{ m}$$

The true error, E_t is

$$E_t = 11061 - 11266$$

$$= -205 \text{m}$$

The true error now is reduced from 807m to 205m. Extending this procedure to dividing $[a,b]$ into n equal segments and applying the trapezoidal rule over each segment, the sum of the results obtained for each segment is the approximate value of the integral.

Divide $(b-a)$ into n equal segments as shown in Figure 4. Then the width of each segment is

$$h = \frac{b-a}{n} \qquad (26)$$

The integral I can be broken into h integrals as

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

$$= \int_{a}^{a+h} f(x)dx + \int_{a+h}^{a+2h} f(x)dx + \ldots + \int_{a+(n-2)h}^{a+(n-1)h} f(x)dx + \int_{a+(n-1)h}^{b} f(x)dx \qquad (27)$$

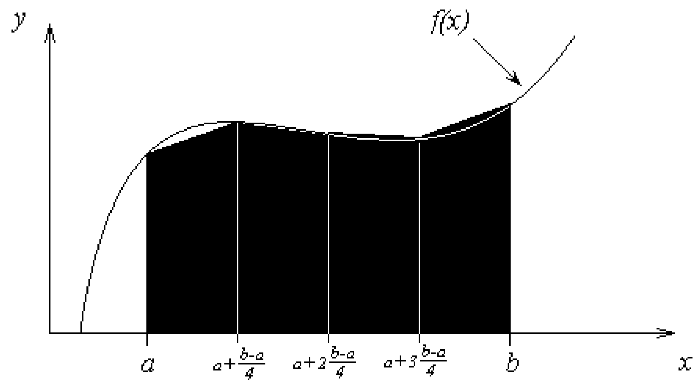

Figure 4 Multiple ($n = 4$) segment trapezoidal rule

Applying trapezoidal rule Equation (27) on each segment gives

$$\int_a^b f(x)dx = [(a+h)-a]\left[\frac{f(a)+f(a+h)}{2}\right]$$

$$+[(a+2h)-(a+h)]\left[\frac{f(a+h)+f(a+2h)}{2}\right]$$

$$+ \ldots\ldots\ldots\ldots\ldots + [(a+(n-1)h)-(a+(n-2)h)]\left[\frac{f(a+(n-2)h)+f(a+(n-1)h)}{2}\right]$$

$$+[b-(a+(n-1)h)]\left[\frac{f(a+(n-1)h)+f(b)}{2}\right]$$

$$= h\left[\frac{f(a)+f(a+h)}{2}\right] + h\left[\frac{f(a+h)+f(a+2h)}{2}\right] + \ldots\ldots\ldots\ldots\ldots\ldots$$

$$+ h\left[\frac{f(a+(n-2)h)+f(a+(n-1)h)}{2}\right] + h\left[\frac{f(a+(n-1)h)+f(b)}{2}\right]$$

$$= h\left[\frac{f(a)+2f(a+h)+2f(a+2h)+\ldots+2f(a+(n-1)h)+f(b)}{2}\right]$$

$$= \frac{h}{2}\left[f(a)+2\left\{\sum_{i=1}^{n-1} f(a+ih)\right\}+f(b)\right]$$

$$= \frac{b-a}{2n}\left[f(a)+2\left\{\sum_{i=1}^{n-1} f(a+ih)\right\}+f(b)\right] \qquad (28)$$

Example 2

The vertical distance covered by a rocket from $t = 8$ to $t = 30$ seconds is given by

$$x = \int_{8}^{30} \left(2000 \ln\left[\frac{140000}{140000 - 2100t} \right] - 9.8t \right) dt$$

a) Use the two-segment trapezoidal rule to find the distance covered from $t = 8$ to $t = 30$ seconds.

b) Find the true error, E_t for part (a).

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

Solution

a) The solution using 2-segment Trapezoidal rule is

$$I \approx \frac{b-a}{2n}\left[f(a) + 2\left\{ \sum_{i=1}^{n-1} f(a+ih) \right\} + f(b) \right]$$

$$n = 2$$
$$a = 8$$
$$b = 30$$
$$h = \frac{b-a}{n}$$
$$= \frac{30-8}{2}$$
$$= 11$$
$$I \approx \frac{30-8}{2(2)}\left[f(8) + 2\left\{ \sum_{i=1}^{2-1} f(8+11i) \right\} + f(30) \right]$$
$$= \frac{22}{4}\left[f(8) + 2f(19) + f(30) \right]$$
$$= \frac{22}{4}\left[177.27 + 2(484.75) + 901.67 \right]$$
$$= 11266 \text{ m}$$

b) The exact value of the above integral is

$$x = \int_{8}^{30} \left(2000 \ln\left[\frac{140000}{140000 - 2100t} \right] - 9.8t \right) dt$$
$$= 11061 \text{ m}$$

so the true error is

$$E_t = \text{True Value} - \text{Approximate Value}$$
$$= 11061 - 11266$$
$$= -205 \text{ m}$$

c) The absolute relative true error, $|\in_t|$, would then be

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

$$= \left|\frac{11061 - 11266}{11061}\right| \times 100$$

$$= 1.8537\%$$

Table 1 Values obtained using multiple-segment trapezoidal rule for

$$x = \int_8^{30} \left(2000 \ln\left[\frac{140000}{140000 - 2100t}\right] - 9.8t \right) dt$$

n	Approximate Value	E_t	$\lvert\in_t\rvert\%$	$\lvert\in_a\rvert\%$
1	11868	-807	7.296	---
2	11266	-205	1.853	5.343
3	11153	-91.4	0.8265	1.019
4	11113	-51.5	0.4655	0.3594
5	11094	-33.0	0.2981	0.1669
6	11084	-22.9	0.2070	0.09082
7	11078	-16.8	0.1521	0.05482
8	11074	-12.9	0.1165	0.03560

Example 3

Use the multiple-segment trapezoidal rule to find the area under the curve

$$f(x) = \frac{300x}{1+e^x}$$

from $x = 0$ to $x = 10$.

Solution

Using two segments, we get

$$h = \frac{10-0}{2} = 5$$

$$f(0) = \frac{300(0)}{1+e^0} = 0$$

$$f(5) = \frac{300(5)}{1+e^5} = 10.039$$

$$f(10) = \frac{300(10)}{1+e^{10}} = 0.136$$

$$I \approx \frac{b-a}{2n}\left[f(a) + 2\left\{\sum_{i=1}^{n-1} f(a+ih)\right\} + f(b)\right]$$

$$= \frac{10-0}{2(2)}\left[f(0) + 2\left\{\sum_{i=1}^{2-1} f(0+5)\right\} + f(10)\right]$$

$$= \frac{10}{4}[f(0) + 2f(5) + f(10)]$$

$$= \frac{10}{4}[0 + 2(10.039) + 0.136] \quad = 50.537$$

So what is the true value of this integral?

$$\int_0^{10} \frac{300x}{1+e^x} dx = 246.59$$

Making the absolute relative true error

$$|\in_t| = \left| \frac{246.59 - 50.535}{246.59} \right| \times 100$$

$$= 79.506\%$$

Why is the true value so far away from the approximate values? Just take a look at Figure 5. As you can see, the area under the "trapezoids" (yeah, they really look like triangles now) covers a small portion of the area under the curve. As we add more segments, the approximated value quickly approaches the true value.

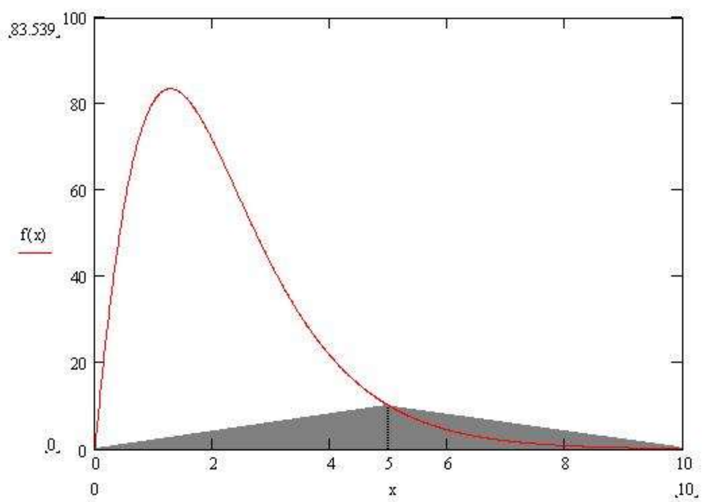

Figure 5 2-segment trapezoidal rule approximation.

Table 2 Values obtained using multiple-segment trapezoidal rule for $\int_0^{10} \frac{300x}{1+e^x} dx$.

n	Approximate Value	E_t	$\lvert \in_t \rvert$
1	0.681	245.91	99.724%
2	50.535	196.05	79.505%
4	170.61	75.978	30.812%
8	227.04	19.546	7.927%
16	241.70	4.887	1.982%
32	245.37	1.222	0.495%
64	246.28	0.305	0.124%

Example 4

Use multiple-segment trapezoidal rule to find

$$I = \int_0^2 \frac{1}{\sqrt{x}} dx$$

Solution

We cannot use the trapezoidal rule for this integral, as the value of the integrand at $x = 0$ is infinite. However, it is known that a discontinuity in a curve will not change the area under it. We can assume any value for the function at $x = 0$. The algorithm to define the function so that we can use the multiple-segment trapezoidal rule is given below.

> Function $f(x)$
> If $x = 0$ Then $f = 0$
> If $x \neq 0$ Then $f = x \wedge (-0.5)$
> End Function

Basically, we are just assigning the function a value of zero at $x = 0$. Everywhere else, the function is continuous. This means the true value of our integral will be just that—true. Let's see what happens using the multiple-segment trapezoidal rule.

Using two segments, we get

$$h = \frac{2-0}{2} = 1$$

$$f(0) = 0$$

$$f(1) = \frac{1}{\sqrt{1}} = 1$$

$$f(2) = \frac{1}{\sqrt{2}} = 0.70711$$

$$I \approx \frac{b-a}{2n}\left[f(a) + 2\left\{\sum_{i=1}^{n-1} f(a+ih)\right\} + f(b)\right]$$

$$= \frac{2-0}{2(2)}\left[f(0) + 2\left\{\sum_{i=1}^{2-1} f(0+1)\right\} + f(2)\right]$$

$$= \frac{2}{4}\left[f(0) + 2f(1) + f(2)\right]$$

$$= \frac{2}{4}\left[0 + 2(1) + 0.70711\right]$$

$$= 1.3536$$

So what is the true value of this integral?

$$\int_0^2 \frac{1}{\sqrt{x}}\,dx = 2.8284$$

Thus making the absolute relative true error

$$|\in_t| = \left|\frac{2.8284 - 1.3536}{2.8284}\right| \times 100$$

$$= 52.145\%$$

Table 3 Values obtained using multiple-segment trapezoidal rule for $\int_0^2 \frac{1}{\sqrt{x}}\,dx$.

n	Approximate Value	E_t	$\lvert\in_t\rvert$
2	1.354	1.474	52.14%
4	1.792	1.036	36.64%
8	2.097	0.731	25.85%
16	2.312	0.516	18.26%
32	2.463	0.365	12.91%
64	2.570	0.258	9.128%
128	2.646	0.182	6.454%
256	2.699	0.129	4.564%
512	2.737	0.091	3.227%
1024	2.764	0.064	2.282%
2048	2.783	0.045	1.613%
4096	2.796	0.032	1.141%

Error in Multiple-segment Trapezoidal Rule

The true error for a single segment Trapezoidal rule is given by

$$E_t = -\frac{(b-a)^3}{12} f''(\zeta), \quad a < \zeta < b$$

Where ζ is some point in $[a,b]$.

What is the error then in the multiple-segment trapezoidal rule? It will be simply the sum of the errors from each segment, where the error in each segment is that of the single segment trapezoidal rule. The error in each segment is

$$E_1 = -\frac{[(a+h)-a]^3}{12} f''(\zeta_1), \quad a < \zeta_1 < a+h$$

$$= -\frac{h^3}{12} f''(\zeta_1)$$

$$E_2 = -\frac{[(a+2h)-(a+h)]^3}{12} f''(\zeta_2), \quad a+h < \zeta_2 < a+2h$$

$$= -\frac{h^3}{12} f''(\zeta_2)$$

.
.
.

$$E_i = -\frac{[(a+ih)-(a+(i-1)h)]^3}{12} f''(\zeta_i), \quad a+(i-1)h < \zeta_i < a+ih$$

$$= -\frac{h^3}{12} f''(\zeta_i)$$

.
.
.

$$E_{n-1} = -\frac{[\{a+(n-1)h\}-\{a+(n-2)h\}]^3}{12} f''(\zeta_{n-1}), \quad a+(n-2)h < \zeta_{n-1} < a+(n-1)h$$

$$= -\frac{h^3}{12} f''(\zeta_{n-1})$$

$$E_n = -\frac{[b-\{a+(n-1)h\}]^3}{12} f''(\zeta_n), \quad a+(n-1)h < \zeta_n < b$$

$$= -\frac{h^3}{12} f''(\zeta_n)$$

Hence the total error in the multiple-segment trapezoidal rule is

$$E_t = \sum_{i=1}^{n} E_i$$

$$= -\frac{h^3}{12} \sum_{i=1}^{n} f''(\zeta_i)$$

$$= -\frac{(b-a)^3}{12n^3} \sum_{i=1}^{n} f''(\zeta_i)$$

$$= -\frac{(b-a)^3}{12n^2} \frac{\sum_{i=1}^{n} f''(\zeta_i)}{n}$$

The term $\frac{\sum_{i=1}^{n} f''(\zeta_i)}{n}$ is an approximate average value of the second derivative $f''(x), a < x < b$.

Hence

$$E_t = -\frac{(b-a)^3}{12n^2}\frac{\sum_{i=1}^{n} f''(\zeta_i)}{n}$$

In Table 4, the approximate value of the integral

$$\int_8^{30}\left(2000\ln\left[\frac{140000}{140000-2100t}\right]-9.8t\right)dt$$

is given as a function of the number of segments. You can visualize that as the number of segments are doubled, the true error gets approximately quartered.

Table 4 Values obtained using multiple-segment trapezoidal rule for

$$x = \int_8^{30}\left(2000\ln\left[\frac{140000}{140000-2100t}\right]-9.8t\right)dt.$$

n	Approximate Value	E_t	$\left\|\in_t\right\|\%$	$\left\|\in_a\right\|\%$
2	11266	-205	1.853	5.343
4	11113	-52	0.4701	0.3594
8	11074	-13	0.1175	0.03560
16	11065	-4	0.03616	0.00401

For example, for the 2-segment trapezoidal rule, the true error is -205, and a quarter of that error is -51.25. That is close to the true error of -48 for the 4-segment trapezoidal rule.

Can you answer the question w*hy is the true error not exactly -51.25?* How does this information help us in numerical integration? You will find out that this forms the basis of Romberg integration based on the trapezoidal rule, where we use the argument that true error gets approximately quartered when the number of segments is doubled. Romberg integration based on the trapezoidal rule is computationally more efficient than using the trapezoidal rule by itself in developing an automatic integration scheme.

Reference

INTEGRATION	
Topic	Trapezoidal Rule
Summary	These are textbook notes of trapezoidal rule of integration
Major	General Engineering
Authors	Autar Kaw, Michael Keteltas

Simpson's 1/3 Rule of Integration

What is integration?

Integration is the process of measuring the area under a function plotted on a graph. Why would we want to integrate a function? Among the most common examples are finding the velocity of a body from an acceleration function, and displacement of a body from a velocity function. Throughout many engineering fields, there are (what sometimes seems like) countless applications for integral calculus. Sometimes, the evaluation of expressions involving these integrals can become daunting, if not indeterminate. For this reason, a wide variety of numerical methods has been developed to simplify the integral. Here, we will discuss Simpson's 1/3 rule of integral approximation, which improves upon the accuracy of the trapezoidal rule.
Here, we will discuss the Simpson's 1/3 rule of approximating integrals of the form

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

where

$f(x)$ is called the integrand,

$a =$ lower limit of integration

$b =$ upper limit of integration

Simpson's 1/3 Rule

The trapezoidal rule was based on approximating the integrand by a first order polynomial, and then integrating the polynomial over interval of integration. Simpson's 1/3 rule is an extension of Trapezoidal rule where the integrand is approximated by a second order polynomial.

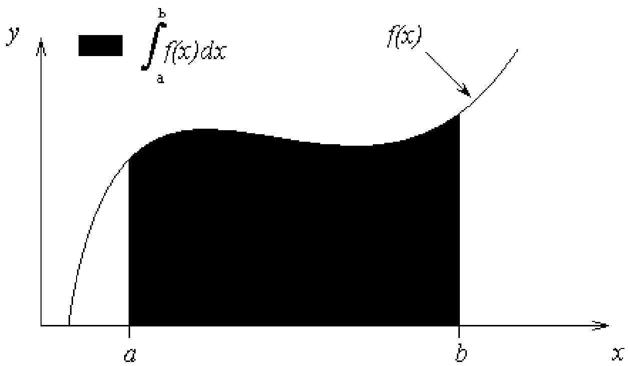

Figure 1 Integration of a function

Method 1:
Hence

$$I = \int_a^b f(x)dx \approx \int_a^b f_2(x)dx$$

where $f_2(x)$ is a second order polynomial given by

$$f_2(x) = a_0 + a_1 x + a_2 x^2.$$

Choose

$$(a, f(a)), \left(\frac{a+b}{2}, f\left(\frac{a+b}{2}\right)\right), \text{ and } (b, f(b))$$

as the three points of the function to evaluate a_0, a_1 and a_2.

$$f(a) = f_2(a) = a_0 + a_1 a + a_2 a^2$$

$$f\left(\frac{a+b}{2}\right) = f_2\left(\frac{a+b}{2}\right) = a_0 + a_1\left(\frac{a+b}{2}\right) + a_2\left(\frac{a+b}{2}\right)^2$$

$$f(b) = f_2(b) = a_0 + a_1 b + a_2 b^2$$

Solving the above three equations for unknowns, a_0, a_1 and a_2 give

$$a_0 = \frac{a^2 f(b) + abf(b) - 4abf\left(\frac{a+b}{2}\right) + abf(a) + b^2 f(a)}{a^2 - 2ab + b^2}$$

$$a_1 = -\frac{af(a) - 4af\left(\frac{a+b}{2}\right) + 3af(b) + 3bf(a) - 4bf\left(\frac{a+b}{2}\right) + bf(b)}{a^2 - 2ab + b^2}$$

$$a_2 = \frac{2\left(f(a) - 2f\left(\frac{a+b}{2}\right) + f(b)\right)}{a^2 - 2ab + b^2}$$

Then

$$I \approx \int_a^b f_2(x)dx$$

$$= \int_a^b \left(a_0 + a_1 x + a_2 x^2\right)dx$$

$$= \left[a_0 x + a_1 \frac{x^2}{2} + a_2 \frac{x^3}{3}\right]_a^b$$

$$= a_0(b-a) + a_1 \frac{b^2 - a^2}{2} + a_2 \frac{b^3 - a^3}{3}$$

Substituting values of a_0, a_1 and a_2 give

$$\int_a^b f_2(x)dx = \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right]$$

Since for Simpson 1/3 rule, the interval $[a,b]$ is broken into 2 segments, the segment width

$$h = \frac{b-a}{2}$$

Hence the Simpson's 1/3 rule is given by

$$\int_a^b f(x)dx \approx \frac{h}{3}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right]$$

Since the above form has 1/3 in its formula, it is called Simpson's 1/3 rule.

<u>Method 2:</u>

Simpson's 1/3 rule can also be derived by approximating $f(x)$ by a second order polynomial using Newton's divided difference polynomial as

$$f_2(x) = b_0 + b_1(x-a) + b_2(x-a)\left(x - \frac{a+b}{2}\right)$$

where

$$b_0 = f(a)$$

$$b_1 = \frac{f\left(\frac{a+b}{2}\right) - f(a)}{\frac{a+b}{2} - a}$$

$$b_2 = \frac{\dfrac{f(b) - f\left(\frac{a+b}{2}\right)}{b - \frac{a+b}{2}} - \dfrac{f\left(\frac{a+b}{2}\right) - f(a)}{\frac{a+b}{2} - a}}{b-a}$$

Integrating Newton's divided difference polynomial gives us

$$\int_a^b f(x)dx \approx \int_a^b f_2(x)dx$$

$$= \int_a^b \left[b_0 + b_1(x-a) + b_2(x-a)\left(x - \frac{a+b}{2}\right)\right]dx$$

$$= \left[b_0 x + b_1\left(\frac{x^2}{2} - ax\right) + b_2\left(\frac{x^3}{3} - \frac{(3a+b)x^2}{4} + \frac{a(a+b)x}{2}\right)\right]_a^b$$

$$= b_0(b-a) + b_1\left(\frac{b^2-a^2}{2} - a(b-a)\right)$$

$$+ b_2\left(\frac{b^3-a^3}{3} - \frac{(3a+b)(b^2-a^2)}{4} + \frac{a(a+b)(b-a)}{2}\right)$$

Substituting values of b_0, b_1, and b_2 into this equation yields the same result as before

$$\int_a^b f(x)dx \approx \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right]$$

$$= \frac{h}{3}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right]$$

Method 3:

One could even use the Lagrange polynomial to derive Simpson's formula. Notice any method of three-point quadratic interpolation can be used to accomplish this task. In this case, the interpolating function becomes

$$f_2(x) = \frac{\left(x - \frac{a+b}{2}\right)(x-b)}{\left(a - \frac{a+b}{2}\right)(a-b)} f(a) + \frac{(x-a)(x-b)}{\left(\frac{a+b}{2} - a\right)\left(\frac{a+b}{2} - b\right)} f\left(\frac{a+b}{2}\right) + \frac{(x-a)\left(x - \frac{a+b}{2}\right)}{(b-a)\left(b - \frac{a+b}{2}\right)} f(b)$$

Integrating this function gets

$$\int_a^b f_2(x)dx = \left[\begin{array}{l} \dfrac{\frac{x^3}{3} - \frac{(a+3b)x^2}{4} + \frac{b(a+b)x}{2}}{\left(a - \frac{a+b}{2}\right)(a-b)} f(a) + \dfrac{\frac{x^3}{3} - \frac{(a+b)x^2}{2} + abx}{\left(\frac{a+b}{2} - a\right)\left(\frac{a+b}{2} - b\right)} f\left(\frac{a+b}{2}\right) \\ + \dfrac{\frac{x^3}{3} - \frac{(3a+b)x^2}{4} + \frac{a(a+b)x}{2}}{(b-a)\left(b - \frac{a+b}{2}\right)} f(b) \end{array} \right]_a^b$$

$$= \frac{\frac{b^3 - a^3}{3} - \frac{(a+3b)(b^2 - a^2)}{4} + \frac{b(a+b)(b-a)}{2}}{\left(a - \frac{a+b}{2}\right)(a-b)} f(a)$$

$$+ \frac{\frac{b^3 - a^3}{3} - \frac{(a+b)(b^2 - a^2)}{2} + ab(b-a)}{\left(\frac{a+b}{2} - a\right)\left(\frac{a+b}{2} - b\right)} f\left(\frac{a+b}{2}\right)$$

$$+ \frac{\frac{b^3 - a^3}{3} - \frac{(3a+b)(b^2 - a^2)}{4} + \frac{a(a+b)(b-a)}{2}}{(b-a)\left(b - \frac{a+b}{2}\right)} f(b)$$

Believe it or not, simplifying and factoring this large expression yields you the same result as before

$$\int_a^b f(x)dx \approx \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right]$$

$$= \frac{h}{3}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right].$$

<u>Method 4:</u>

Simpson's 1/3 rule can also be derived by the method of coefficients. Assume

$$\int_a^b f(x)dx \approx c_1 f(a) + c_2 f\left(\frac{a+b}{2}\right) + c_3 f(b)$$

Let the right-hand side be an exact expression for the integrals $\int_a^b 1dx,\ \int_a^b xdx,$ and $\int_a^b x^2dx$. This implies that the right hand side will be exact expressions for integrals of any linear combination of the three integrals for a general second order polynomial. Now

$$\int_a^b 1dx = b - a = c_1 + c_2 + c_3$$

$$\int_a^b xdx = \frac{b^2 - a^2}{2} = c_1 a + c_2 \frac{a+b}{2} + c_3 b$$

$$\int_a^b x^2dx = \frac{b^3 - a^3}{3} = c_1 a^2 + c_2\left(\frac{a+b}{2}\right)^2 + c_3 b^2$$

Solving the above three equations for c_0, c_1 and c_2 give

$$c_1 = \frac{b-a}{6}$$

$$c_2 = \frac{2(b-a)}{3}$$
$$c_3 = \frac{b-a}{6}$$

This gives

$$\int_a^b f(x)dx \approx \frac{b-a}{6} f(a) + \frac{2(b-a)}{3} f\left(\frac{a+b}{2}\right) + \frac{b-a}{6} f(b)$$
$$= \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right]$$
$$= \frac{h}{3}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right]$$

The integral from the first method

$$\int_a^b f(x)dx \approx \int_a^b (a_0 + a_1 x + a_2 x^2)dx$$

can be viewed as the area under the second order polynomial, while the equation from Method 4

$$\int_a^b f(x)dx \approx \frac{b-a}{6} f(a) + \frac{2(b-a)}{3} f\left(\frac{a+b}{2}\right) + \frac{b-a}{6} f(b)$$

can be viewed as the sum of the areas of three rectangles.

Example 1

The distance covered by a rocket in meters from $t = 8s$ to $t = 30s$ is given by

$$x = \int_8^{30} \left(2000\ln\left[\frac{140000}{140000 - 2100t}\right] - 9.8t \right) dt$$

a) Use Simpson’s 1/3 rule to find the approximate value of x.
b) Find the true error, E_t.
c) Find the absolute relative true error, $|\in_t|$.

Solution

a) $$x \approx \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right]$$

$a = 8$
$b = 30$
$\frac{a+b}{2} = 19$

$$f(t) = 2000\ln\left[\frac{140000}{140000 - 2100t}\right] - 9.8t$$

$$f(8) = 2000\ln\left[\frac{140000}{140000 - 2100(8)}\right] - 9.8(8) = 177.27 m/s$$

$$f(30) = 2000\ln\left[\frac{140000}{140000 - 2100(30)}\right] - 9.8(30) = 901.67 m/s$$

$$f(19) = 2000\ln\left(\frac{140000}{140000 - 2100(19)}\right) - 9.8(19) = 484.75 m/s$$

$$x \approx \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right]$$

$$= \left(\frac{30-8}{6}\right)\left[f(8) + 4f(19) + f(30)\right]$$

$$= \frac{22}{6}\left[177.27 + 4 \times 484.75 + 901.67\right]$$

=11065.72 m

b) The exact value of the above integral is

$$x = \int_{8}^{30}\left(2000\ln\left[\frac{140000}{140000 - 2100t}\right] - 9.8t\right)dt$$

=11061.34 m

So the true error is

$$E_t = True\ Value - Approximate\ Value$$

=11061.34-11065.72

$$= -4.38\ m$$

c) The absolute relative true error is

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

$$= \left|\frac{-4.38}{11061.34}\right| \times 100$$

$$= 0.0396\%$$

Multiple-segment Simpson's 1/3 Rule

Just like in multiple-segment trapezoidal rule, one can subdivide the interval $[a,b]$ into n segments and apply Simpson's 1/3 rule repeatedly over every two segments. Note that n needs to be even. Divide interval $[a,b]$ into n equal segments, so that the segment width is given by

$$h = \frac{b-a}{n}.$$

Now

$$\int_a^b f(x)dx = \int_{x_0}^{x_n} f(x)dx$$

where

$$x_0 = a$$

$$x_n = b$$

$$\int_a^b f(x)dx = \int_{x_0}^{x_2} f(x)dx + \int_{x_2}^{x_4} f(x)dx + ... + \int_{x_{n-4}}^{x_{n-2}} f(x)dx + \int_{x_{n-2}}^{x_n} f(x)dx$$

Apply Simpson's 1/3rd Rule over each interval,

$$\int_a^b f(x)dx \cong (x_2 - x_0)\left[\frac{f(x_0) + 4f(x_1) + f(x_2)}{6}\right] + (x_4 - x_2)\left[\frac{f(x_2) + 4f(x_3) + f(x_4)}{6}\right] + ...$$

$$+ (x_{n-2} - x_{n-4})\left[\frac{f(x_{n-4}) + 4f(x_{n-3}) + f(x_{n-2})}{6}\right] + (x_n - x_{n-2})\left[\frac{f(x_{n-2}) + 4f(x_{n-1}) + f(x_n)}{6}\right]$$

Since

$$x_i - x_{i-2} = 2h$$

$$i = 2, 4, ..., n$$

then

$$\int_a^b f(x)dx \cong 2h\left[\frac{f(x_0) + 4f(x_1) + f(x_2)}{6}\right] + 2h\left[\frac{f(x_2) + 4f(x_3) + f(x_4)}{6}\right] + ...$$

$$+ 2h\left[\frac{f(x_{n-4}) + 4f(x_{n-3}) + f(x_{n-2})}{6}\right] + 2h\left[\frac{f(x_{n-2}) + 4f(x_{n-1}) + f(x_n)}{6}\right]$$

$$= \frac{h}{3}\left[f(x_0) + 4\{f(x_1) + f(x_3) + ... + f(x_{n-1})\} + 2\{f(x_2) + f(x_4) + ... + f(x_{n-2})\} + f(x_n)\right]$$

$$= \frac{h}{3}\left[f(x_0) + 4\sum_{\substack{i=1 \\ i=odd}}^{n-1} f(x_i) + 2\sum_{\substack{i=2 \\ i=even}}^{n-2} f(x_i) + f(x_n)\right]$$

$$\boxed{\int_a^b f(x)dx \cong \frac{b-a}{3n}\left[f(x_0) + 4\sum_{\substack{i=1 \\ i=odd}}^{n-1} f(x_i) + 2\sum_{\substack{i=2 \\ i=even}}^{n-2} f(x_i) + f(x_n)\right]}$$

Example 2

Use 4-segment Simpson's 1/3 rule to approximate the distance covered by a rocket in meters from $t = 8$s to $t = 30$s as given by

$$x = \int_8^{30}\left(2000\ln\left[\frac{140000}{140000 - 2100t}\right] - 9.8t\right)dt$$

a) Use four segment Simpson's 1/3rd Rule to estimate *x*.
b) Find the true error, E_t for part (a).
c) Find the absolute relative true error, $|\in_t|$ for part (a).

Solution:

a) Using n segment Simpson's 1/3 rule,

$$x \approx \frac{b-a}{3n}\left[f(t_0) + 4\sum_{\substack{i=1\\i=odd}}^{n-1} f(t_i) + 2\sum_{\substack{i=2\\i=even}}^{n-2} f(t_i) + f(t_n)\right]$$

$$n = 4$$
$$a = 8$$
$$b = 30$$
$$h = \frac{b-a}{n}$$
$$= \frac{30-8}{4}$$
$$= 5.5$$
$$f(t) = 2000\ln\left[\frac{140000}{140000-2100t}\right] - 9.8t$$

So

$$f(t_0) = f(8)$$
$$f(8) = 2000\ln\left[\frac{140000}{140000-2100(8)}\right] - 9.8(8) = 177.27 m/s$$

$$f(t_1) = f(8+5.5) = f(13.5)$$

$$f(13.5) = 2000\ln\left[\frac{140000}{140000-2100(13.5)}\right] - 9.8(13.5) = 320.25 m/s$$

$$f(t_2) = f(13.5+5.5) = f(19)$$

$$f(19) = 2000\ln\left(\frac{140000}{140000-2100(19)}\right) - 9.8(19) = 484.75 m/s$$

$$f(t_3) = f(19+5.5) = f(24.5)$$

$$f(24.5) = 2000\ln\left[\frac{140000}{140000-2100(24.5)}\right] - 9.8(24.5) = 676.05 m/s$$

$$f(t_4) = f(t_n) = f(30)$$

$$f(30) = 2000\ln\left[\frac{140000}{140000 - 2100(30)}\right] - 9.8(30) = 901.67 m/s$$

$$x = \frac{b-a}{3n}\left[f(t_0) + 4\sum_{\substack{i=1 \\ i=odd}}^{n-1} f(t_i) + 2\sum_{\substack{i=2 \\ i=even}}^{n-2} f(t_i) + f(t_n)\right]$$

$$= \frac{30-8}{3(4)}\left[f(8) + 4\sum_{\substack{i=1 \\ i=odd}}^{3} f(t_i) + 2\sum_{\substack{i=2 \\ i=even}}^{2} f(t_i) + f(30)\right]$$

$$= \frac{22}{12}\left[f(8) + 4f(t_1) + 4f(t_3) + 2f(t_2) + f(30)\right]$$

$$= \frac{11}{6}\left[f(8) + 4f(13.5) + 4f(24.5) + 2f(19) + f(30)\right]$$

$$= \frac{11}{6}\left[177.27 + 4(320.25) + 4(676.05) + 2(484.75) + 901.67\right]$$

$$= 11061.64\ m$$

b) The exact value of the above integral is

$$x = \int_8^{30}\left(2000\ln\left[\frac{140000}{140000 - 2100t}\right] - 9.8t\right)dt$$

=11061.34 m

So the true error is

$$E_t = True\ Value - Approximate\ Value$$

$$E_t = 11061.34 - 11061.64$$

$$= -0.30\ m$$

c) The absolute relative true error is

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

$$= \left|\frac{-0.3}{11061.34}\right| \times 100$$

= 0.0027%

Table 1 Values of Simpson's 1/3 rule for Example 2 with multiple-segments

n	Approximate Value	E_t	$\|\in_t\|$
2	11065.72	-4.38	0.0396%
4	11061.64	-0.30	0.0027%
6	11061.40	-0.06	0.0005%

8	11061.35	-0.02	0.0002%
10	11061.34	-0.01	0.0001%

Error in Multiple-segment Simpson's 1/3 rule

The true error in a single application of Simpson's 1/3rd Rule is given[1] by

$$E_t = -\frac{(b-a)^5}{2880} f^{(4)}(\zeta), \quad a < \zeta < b$$

In multiple-segment Simpson's 1/3 rule, the error is the sum of the errors in each application of Simpson's 1/3 rule. The error in the *n* segments Simpson's 1/3rd Rule is given by

$$E_1 = -\frac{(x_2 - x_0)^5}{2880} f^{(4)}(\zeta_1), \quad x_0 < \zeta_1 < x_2$$

$$= -\frac{h^5}{90} f^{(4)}(\zeta_1)$$

$$E_2 = -\frac{(x_4 - x_2)^5}{2880} f^{(4)}(\zeta_2), \quad x_2 < \zeta_2 < x_4$$

$$= -\frac{h^5}{90} f^{(4)}(\zeta_2)$$

$$\vdots$$

$$E_i = -\frac{(x_{2i} - x_{2(i-1)})^5}{2880} f^{(4)}(\zeta_i), \quad x_{2(i-1)} < \zeta_i < x_{2i}$$

$$= -\frac{h^5}{90} f^{(4)}(\zeta_i)$$

$$\vdots$$

$$E_{\frac{n}{2}-1} = -\frac{(x_{n-2} - x_{n-4})^5}{2880} f^{(4)}\left(\zeta_{\frac{n}{2}-1}\right), \quad x_{n-4} < \zeta_{\frac{n}{2}-1} < x_{n-2}$$

$$= -\frac{h^5}{90} f^{(4)}\left(\zeta_{\frac{n}{2}-1}\right)$$

$$E_{\frac{n}{2}} = -\frac{(x_n - x_{n-2})^5}{2880} f^{(4)}\left(\zeta_{\frac{n}{2}}\right), x_{n-2} < \zeta_{\frac{n}{2}} < x_n$$

Hence, the total error in the multiple-segment Simpson's 1/3 rule is

$$= -\frac{h^5}{90} f^{(4)}\left(\zeta_{\frac{n}{2}}\right)$$

---

[1] The $f^{(4)}$ in the true error expression stands for the fourth derivative of the function $f(x)$.

$$
\begin{aligned}
E_t &= \sum_{i=1}^{\frac{n}{2}} E_i \\
&= -\frac{h^5}{90} \sum_{i=1}^{\frac{n}{2}} f^{(4)}(\zeta_i) \\
&= -\frac{(b-a)^5}{90n^5} \sum_{i=1}^{\frac{n}{2}} f^{(4)}(\zeta_i) \\
&= -\frac{(b-a)^5}{180n^4} \frac{\sum_{i=1}^{\frac{n}{2}} f^{(4)}(\zeta_i)}{\frac{n}{2}},
\end{aligned}
$$

The term $\frac{\sum_{i=1}^{\frac{n}{2}} f^{(4)}(\zeta_i)}{\frac{n}{2}}$ is an approximate average value of $f^{(4)}(x), a < x < b$. Hence

$$
E_t = -\frac{(b-a)^5}{180n^4} \overline{f}^{(4)}
$$

where

$$
\overline{f}^{(4)} = \frac{\sum_{i=1}^{\frac{n}{2}} f^{(4)}(\zeta_i)}{\frac{n}{2}}
$$

Reference

INTEGRATION	
Topic	Simpson’s 1/3 rule
Summary	Textbook notes of Simpson’s 1/3 rule
Major	General Engineering
Authors	Autar Kaw, Michael Keteltas
Date	March 22, 2022

Simpson 3/8 Rule for Integration

Introduction

The main objective of this chapter is to develop appropriate formulas for approximating the integral of the form

$$I = \int_a^b f(x)dx \tag{1}$$

Most (if not all) of the developed formulas for integration are based on a simple concept of approximating a given function $f(x)$ by a simpler function (usually a polynomial function) $f_i(x)$, where i represents the order of the polynomial function. In Chapter 07.03, Simpsons 1/3 rule for integration was derived by approximating the integrand $f(x)$ with a 2[nd] order (quadratic) polynomial function. $f_2(x)$

$$f_2(x) = a_0 + a_1 x + a_2 x^2 \tag{2}$$

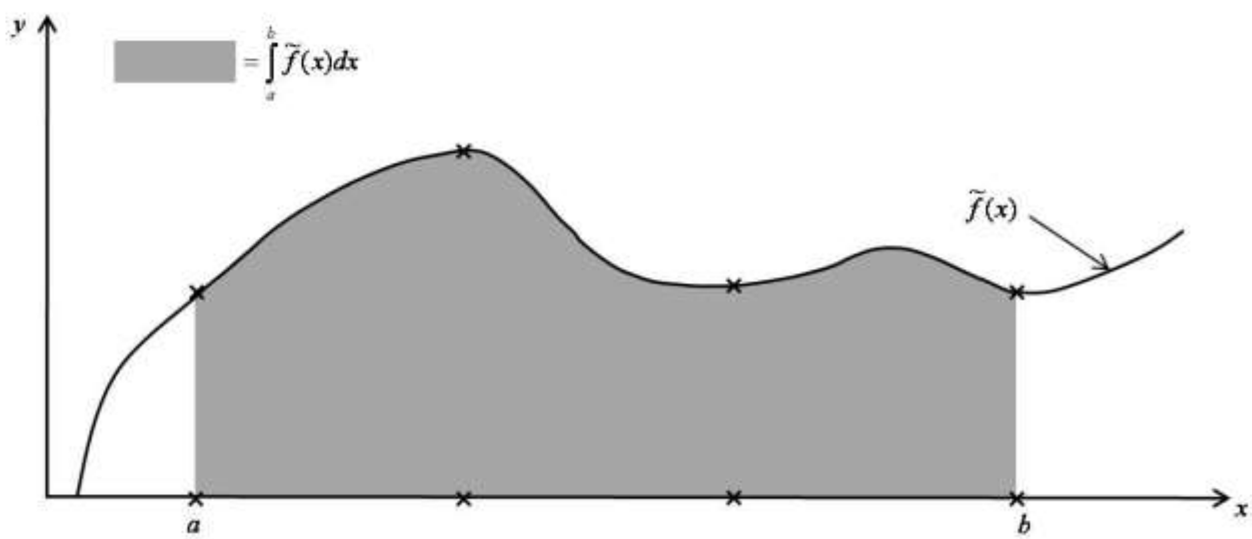

Figure 1 $\tilde{f}(x)$ Cubic function.

In a similar fashion, Simpson 3/8 rule for integration can be derived by approximating the given function $f(x)$ with the 3[rd] order (cubic) polynomial $f_3(x)$

$$f_3(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 \\ = \{1, x, x^2, x^3\} \times \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} \quad (3)$$

which can also be symbolically represented in Figure 1.

<u>Method 1</u>

The unknown coefficients a_0, a_1, a_2 and a_3 in Equation (3) can be obtained by substituting 4 known coordinate data points $\{x_0, f(x_0)\}, \{x_1, f(x_1)\}, \{x_2, f(x_2)\}$ and $\{x_3, f(x_3)\}$ into Equation (3) as follows.

$$\left.\begin{aligned} f(x_0) &= a_0 + a_1 x_0 + a_2 x_0^2 + a_3 x_0^2 \\ f(x_1) &= a_0 + a_1 x_1 + a_2 x_1^2 + a_3 x_1^2 \\ f(x_2) &= a_0 + a_1 x_2 + a_2 x_2^2 + a_3 x_2^2 \\ f(x_3) &= a_0 + a_1 x_3 + a_2 x_3^2 + a_3 x_3^2 \end{aligned}\right\} \quad (4)$$

Equation (4) can be expressed in matrix notation as

$$\begin{bmatrix} 1 & x_0 & x_0^2 & x_0^3 \\ 1 & x_1 & x_1^2 & x_1^3 \\ 1 & x_2 & x_2^2 & x_2^3 \\ 1 & x_3 & x_3^2 & x_3^3 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} f(x_0) \\ f(x_1) \\ f(x_2) \\ f(x_3) \end{bmatrix} \quad (5)$$

The above Equation (5) can symbolically be represented as

$$[A]_{4\times4}\, \vec{a}_{4\times1} = \vec{f}_{4\times1} \quad (6)$$

Thus,

$$\vec{a} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} = [A]^{-1} \times \vec{f} \quad (7)$$

Substituting Equation (7) into Equation (3), one gets

$$f_3(x) = \{1, x, x^2, x^3\} \times [A]^{-1} \times \vec{f} \quad (8)$$

As indicated in Figure 1, one has

$$\left.\begin{aligned}
x_0 &= a \\
x_1 &= a + h \\
&= a + \frac{b-a}{3} \\
&= \frac{2a+b}{3} \\
x_2 &= a + 2h \\
&= a + \frac{2b-2a}{3} \\
&= \frac{a+2b}{3} \\
x_3 &= a + 3h \\
&= a + \frac{3b-3a}{3} \\
&= b
\end{aligned}\right\} \tag{9}$$

With the help from MATLAB [Ref. 2], the unknown vector $\vec{a}$ (shown in Equation 7) can be solved for symbolically.

<u>Method 2</u>

Using Lagrange interpolation, the cubic polynomial function $f_3(x)$ that passes through 4 data points (see Figure 1) can be explicitly given as

$$\begin{aligned}
f_3(x) = {} & \frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} \times f(x_0) + \frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)} \times f(x_1) \\
& + \frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} \times f(x_3) + \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)} \times f(x_3)
\end{aligned} \tag{10}$$

Simpsons 3/8 Rule for Integration

Substituting the form of $f_3(x)$ from Method (1) or Method (2),

$$\begin{aligned}
I &= \int_a^b f(x)dx \\
&\approx \int_a^b f_3(x)dx \\
&= (b-a) \times \frac{\{f(x_0) + 3f(x_1) + 3f(x_2) + f(x_3)\}}{8}
\end{aligned} \tag{11}$$

Since

$$h = \frac{b-a}{3}$$

$$b-a=3h$$

and Equation (11) becomes

$$I \approx \frac{3h}{8}\times\{f(x_0)+3f(x_1)+3f(x_2)+f(x_3)\} \quad (12)$$

Note the 3/8 in the formula, and hence the name of method as the Simpson's 3/8 rule.
The true error in Simpson 3/8 rule can be derived as [Ref. 1]

$$E_t=-\frac{(b-a)^5}{6480}\times f''''(\zeta)\text{ , where } a\le\zeta\le b \quad (13)$$

Example 1

The vertical distance in meters covered by a rocket from $t=8$ to $t=30$ seconds is given by

$$s=\int_8^{30}\left(2000\ln\left[\frac{140000}{140000-2100t}\right]-9.8t\right)dt$$

Use Simpson 3/8 rule to find the approximate value of the integral.

Solution

$$h=\frac{b-a}{n}$$

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

$$=\frac{30-8}{3}$$

$$=7.3333$$

$$f(t)=2000\ln\left[\frac{140000}{140000-2100t}\right]-9.8t$$

$$I\approx\frac{3h}{8}\times\{f(t_0)+3f(t_1)+3f(t_2)+f(t_3)\}$$

$$t_0=8$$

$$f(t_0)=2000\ln\left(\frac{140000}{140000-2100\times8}\right)-9.8\times8$$

$$=177.2667$$

$$\left\{\begin{aligned}t_1&=t_0+h\\&=8+7.3333\\&=15.3333\\f(t_1)&=2000\ln\left(\frac{140000}{140000-2100\times15.3333}\right)-9.8\times15.3333\\&=372.4629\end{aligned}\right.$$

$$\begin{cases} t_2 = t_0 + 2h \\ \quad = 8 + 2(7.3333) \\ \quad = 22.6666 \\ f(t_2) = 2000\ln\left(\frac{140000}{140000 - 2100 \times 22.6666}\right) - 9.8 \times 22.6666 \\ \qquad = 608.8976 \end{cases}$$

$$\begin{cases} t_3 = t_0 + 3h \\ \quad = 8 + 3(7.3333) \\ \quad = 30 \\ f(t_3) = 2000\ln\left(\frac{140000}{140000 - 2100 \times 30}\right) - 9.8 \times 30 \\ \qquad = 901.6740 \end{cases}$$

Applying Equation (12), one has

$$I = \frac{3}{8} \times 7.3333 \times \{177.2667 + 3 \times 372.4629 + 3 \times 608.8976 + 901.6740\}$$
$$= 11063.3104m$$

The exact answer can be computed as

$$I_{exact} = 11061.34m$$

Multiple Segments for Simpson 3/8 Rule

Using n = number of equal segments, the width h can be defined as

$$h = \frac{b-a}{n} \qquad (14)$$

The number of segments need to be an integer multiple of 3 as a single application of Simpson 3/8 rule requires 3 segments.

The integral shown in Equation (1) can be expressed as

$$I = \int_a^b f(x)dx$$
$$\approx \int_a^b f_3(x)dx$$
$$\approx \int_{x_0=a}^{x_3} f_3(x)dx + \int_{x_3}^{x_6} f_3(x)dx + \ldots\ldots + \int_{x_{n-3}}^{x_n=b} f_3(x)dx \qquad (15)$$

Using Simpson 3/8 rule (See Equation 12) into Equation (15), one gets

$$I = \frac{3h}{8}\begin{Bmatrix} f(x_0)+3f(x_1)+3f(x_2)+f(x_3)+f(x_3)+3f(x_4)+3f(x_5)+f(x_6) \\ +.....+f(x_{n-3})+3f(x_{n-2})+3f(x_{n-1})+f(x_n) \end{Bmatrix} \quad (16)$$

$$= \frac{3h}{8}\left\{ f(x_0)+3\sum_{i=1,4,7,..}^{n-2} f(x_i)+3\sum_{i=2,5,8,..}^{n-1} f(x_i)+2\sum_{i=3,6,9,..}^{n-3} f(x_i)+f(x_n)\right\} \quad (17)$$

Example 2

The vertical distance in meters covered by a rocket from $t = 8$ to $t = 30$ seconds is given by

$$s = \int_8^{30}\left(2000\ln\left[\frac{140000}{140000-2100t}\right]-9.8t\right)dt$$

Use Simpson 3/8 multiple segments rule with six segments to estimate the vertical distance.

Solution

In this example, one has (see Equation 14):

$$f(t) = 2000\ln\left[\frac{140000}{140000-2100t}\right]-9.8t$$

$$h = \frac{30-8}{6} = 3.6666$$

$$\{t_0, f(t_0)\} = \{8, 177.2667\}$$

$$\{t_1, f(t_1)\} = \{11.6666, 270.4104\} \text{ where } t_1 = t_0 + h = 8 + 3.6666 = 11.6666$$

$$\{t_2, f(t_2)\} = \{15.3333, 372.4629\} \text{ where } t_2 = t_0 + 2h = 15.3333$$

$$\{t_3, f(t_3)\} = \{19, 484.7455\} \text{ where } t_3 = t_0 + 3h = 19$$

$$\{t_4, f(t_4)\} = \{22.6666, 608.8976\} \text{ where } t_4 = t_0 + 4h = 22.6666$$

$$\{t_5, f(t_5)\} = \{26.3333, 746.9870\} \text{ where } t_5 = t_0 + 5h = 26.3333$$

$$\{t_6, f(t_6)\} = \{30, 901.6740\} \text{ where } t_6 = t_0 + 6h = 30$$

Applying Equation (17), one obtains:

$$I = \frac{3}{8}(3.6666)\left\{177.2667+3\sum_{i=1,4,..}^{n-2=4} f(t_i)+3\sum_{i=2,5,..}^{n-1=5} f(t_i)+2\sum_{i=3,6,..}^{n-3=3} f(t_i)+901.6740\right\}$$

$$= (1.3750)\begin{Bmatrix} 177.2667+3(270.4104+608.8976) \\ +3(372.4629+746.9870)+2(484.7455)+901.6740 \end{Bmatrix}$$

$$= 11{,}601.4696m$$

Example 3

Compute

$$I = \int_8^{30}\left\{2000\ln\left(\frac{140000}{140000-2100t}\right)-9.8t\right\}dt,$$

using Simpson 1/3 rule (with $n_1 = 4$), and Simpson 3/8 rule (with $n_2 = 3$).

Solution

The segment width is

$$h = \frac{b-a}{n}$$
$$= \frac{b-a}{n_1 + n_2}$$
$$= \frac{30-8}{(4+3)}$$
$$= 3.1429$$

$$f(t) = 2000\ln\left[\frac{140000}{140000 - 2100t}\right] - 9.8t$$

$$\left.\begin{aligned} t_0 &= a = 8 \\ t_1 &= x_0 + 1h = 8 + 3.1429 = 11.1429 \\ t_2 &= t_0 + 2h = 8 + 2(3.1429) = 14.2857 \\ t_3 &= t_0 + 3h = 8 + 3(3.1429) = 17.4286 \\ t_4 &= t_0 + 4h = 8 + 4(3.1429) = 20.5714 \end{aligned}\right\} \text{Simpson's 1/3 rule}$$
$$t_5 = t_0 + 5h = 8 + 5(3.1429) = 23.7143$$
$$t_6 = t_0 + 6h = 8 + 6(3.1429) = 26.8571$$
$$t_7 = t_0 + 7h = 8 + 7(3.1429) = 30$$

Now

$$f(t_0 = 8) = 2000\ln\left(\frac{140{,}000}{140{,}000 - 2100 \times 8}\right) - 9.8 \times 8$$
$$= 177.2667$$

Similarly:

$$f(t_1) = 256.5863$$
$$f(t_2) = 342.3241$$
$$f(t_3) = 435.2749$$
$$f(t_4) = 536.3909$$
$$f(t_5) = 646.8260$$
$$f(t_6) = 767.9978$$
$$f(t_7) = 901.6740$$

For multiple segments $(n_1 = \text{first 4 segments})$, using Simpson 1/3 rule, one obtains (See Equation 19):

$$I_1 = \left(\frac{h}{3}\right)\left\{ f(t_0) + 4\sum_{i=1,3,\ldots}^{n_1-1=3} f(t_i) + 2\sum_{i=2,\ldots}^{n_1-2=2} f(t_i) + f\left(t_{n_1}\right)\right\}$$
$$= \left(\frac{h}{3}\right)\{ f(t_0) + 4(f(t_1) + f(t_3)) + 2 f(t_2) + f(t_4)\}$$
$$= \left(\frac{3.1429}{3}\right)\{177.2667 + 4(256.5863 + 435.2749) + 2(342.3241) + 536.3909\}$$
$$= 4364.1197$$

For multiple segments $(n_2 = \text{last 3 segments})$, using Simpson 3/8 rule, one obtains (See Equation 17):

$$I_2 = \left(\frac{3h}{8}\right)\left\{ f(t_0) + 3\sum_{i=1,3,\ldots}^{n_2-2=1} f(t_i) + 3\sum_{i=2,\ldots}^{n_2-1=2} f(t_i) + 2\sum_{i=3,6,..}^{n_2-3=0} f(t_i) + f\left(t_{n_1}\right)\right\}$$
$$= \left(\frac{3h}{8}\right)\{ f(t_0) + 3 f(t_1) + 3 f(t_2) + 2(\text{no contribution}) + f(t_3)\}$$
$$= \left(\frac{3h}{8}\right)\{ f(t_4) + 3 f(t_5) + 3 f(t_6) + f(t_7)\}$$
$$= \left(\frac{3}{8}\times 3.1429\right)\{536.3909 + 3(646.8260) + 3(767.9978) + 901.6740\}$$
$$= 6697.3663$$

The mixed (combined) Simpson 1/3 and 3/8 rules give

$$I = I_1 + I_2$$
$$= 4364.1197 + 6697.3663$$
$$= 11061m$$

Comparing the truncated error of Simpson 1/3 rule

$$E_t = -\frac{(b-a)^5}{2880} \times f''''(\zeta) \tag{18}$$

With Simpson 3/8 rule (See Equation 12), it seems to offer slightly more accurate answer than the former. However, the cost associated with Simpson 3/8 rule (using 3rd order polynomial function) is significantly higher than the one associated with Simpson 1/3 rule (using 2nd order polynomial function).

The number of multiple segments that can be used in the conjunction with Simpson 1/3 rule is 2, 4, 6, 8, … (any even numbers) for

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

$$\approx\left(\frac{h}{3}\right)\{f(x_0)+4f(x_1)+f(x_2)+f(x_2)+4f(x_3)+f(x_4)+.....+f(x_{n-2})+4f(x_{n-1})+f(x_n)\}$$

$$=\left(\frac{h}{3}\right)\left\{f(x_0)+4\sum_{i=1,3,\ldots}^{n-1}f(x_i)+2\sum_{i=2,4,6\ldots}^{n-2}f(x_i)+f(x_n)\right\} \quad (19)$$

However, Simpson 3/8 rule can be used with the number of segments equal to 3,6,9,12,.. (can be certain integers that are multiples of 3).

If the user wishes to use, say 7 segments, then the mixed Simpson 1/3 rule (for the first 4 segments), and Simpson 3/8 rule (for the last 3 segments) would be appropriate.

Computer Algorithm for Mixed Simpson 1/3 and 3/8 Rule for Integration

Based on the earlier discussion on (single and multiple segments) Simpson 1/3 and 3/8 rules, the following "pseudo" step-by-step mixed Simpson rules for estimating

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

can be given as

Step 1

User inputs information, such as

$f(x)$ = integrand

n_1 = number of segments in conjunction with Simpson 1/3 rule (a multiple of 2 (any even numbers)

n_2 = number of segments in conjunction with Simpson 3/8 rule (a multiple of 3)

Step 2

Compute

$$n=n_1+n_2$$

$$h=\frac{b-a}{n}$$

$$x_0=a$$

$$x_1=a+1h$$

$$x_2=a+2h$$

.

.

$$x_i=a+ih$$

.

.

$$x_n=a+nh=b$$

Step 3

Compute result from multiple-segment Simpson 1/3 rule (See Equation 19)

$$I_1 = \left(\frac{h}{3}\right)\left\{ f(x_0) + 4\sum_{i=1,3,\ldots}^{n_1-1} f(x_i) + 2\sum_{i=2,4,6\ldots}^{n_1-2} f(x_i) + f(x_{n_1}) \right\} \qquad \text{(19, repeated)}$$

<u>Step 4</u>

Compute result from multiple segment Simpson 3/8 rule (See Equation 17)

$$I_2 = \left(\frac{3h}{8}\right)\left\{ f(x_0) + 3\sum_{i=1,4,7\ldots}^{n_2-2} f(x_i) + 3\sum_{i=2,5,8\ldots}^{n_2-1} f(x_i) + 2\sum_{i=3,6,9,\ldots}^{n_2-3} f(x_i) + f(x_{n_2}) \right\} \qquad \text{(17, repeated)}$$

<u>Step 5</u>

$$I \approx I_1 + I_2 \qquad (20)$$

and print out the final approximated answer for I .

Reference

SIMPSON'S 3/8 RULE FOR INTEGRATION	
Topic	Simpson 3/8 Rule for Integration
Summary	Textbook Chapter of Simpson's 3/8 Rule for Integration
Major	General Engineering
Authors	Duc Nguyen
Date	March 28, 2022

Euler's Method for Ordinary Differential Equations

What is Euler's method?

Euler's method is a numerical technique to solve ordinary differential equations of the form

$$\frac{dy}{dx} = f(x, y), y(0) = y_0 \qquad (1)$$

So only first order ordinary differential equations can be solved by using Euler's method. In another chapter we will discuss how Euler's method is used to solve higher order ordinary differential equations or coupled (simultaneous) differential equations. How does one write a first order differential equation in the above form?

Example 1

Rewrite

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

in

$$\frac{dy}{dx} = f(x, y), \ y(0) = y_0 \text{ form.}$$

Solution

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

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

In this case

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

Example 2

Rewrite

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

in

$$\frac{dy}{dx} = f(x, y), \ y(0) = y_0 \text{ form.}$$

Solution

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

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

In this case

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

Derivation of Euler's method

At $x = 0$, we are given the value of $y = y_0$. Let us call $x = 0$ as x_0. Now since we know the slope of y with respect to x, that is, $f(x, y)$, then at $x = x_0$, the slope is $f(x_0, y_0)$. Both x_0 and y_0 are known from the initial condition $y(x_0) = y_0$.

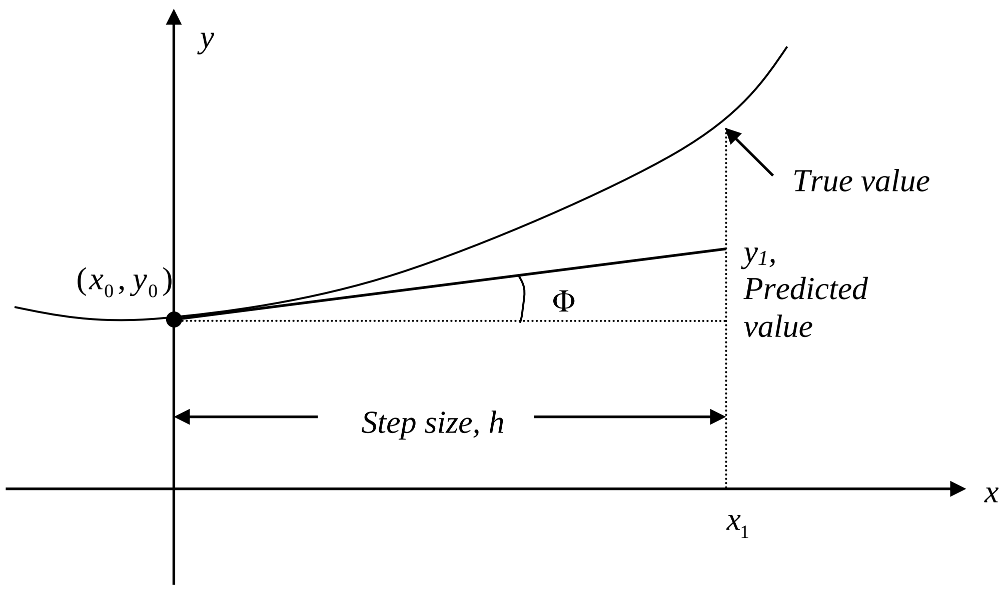

Figure 1 Graphical interpretation of the first step of Euler's method.

So the slope at $x = x_0$ as shown in Figure 1 is

$$\text{Slope} = \frac{Rise}{Run} = \frac{y_1 - y_0}{x_1 - x_0} = f(x_0, y_0)$$

From here

$$y_1 = y_0 + f(x_0, y_0)(x_1 - x_0)$$

Calling $x_1 - x_0$ the step size h, we get

$$y_1 = y_0 + f(x_0, y_0)h \tag{2}$$

One can now use the value of y_1 (an approximate value of y at $x = x_1$) to calculate y_2, and that would be the predicted value at x_2, given by

$$y_2 = y_1 + f(x_1, y_1)h$$

$$x_2 = x_1 + h$$

Based on the above equations, if we now know the value of $y = y_i$ at x_i, then

$$y_{i+1} = y_i + f(x_i, y_i)h \tag{3}$$

This formula is known as Euler's method and is illustrated graphically in Figure 2. In some books, it is also called the Euler-Cauchy method.

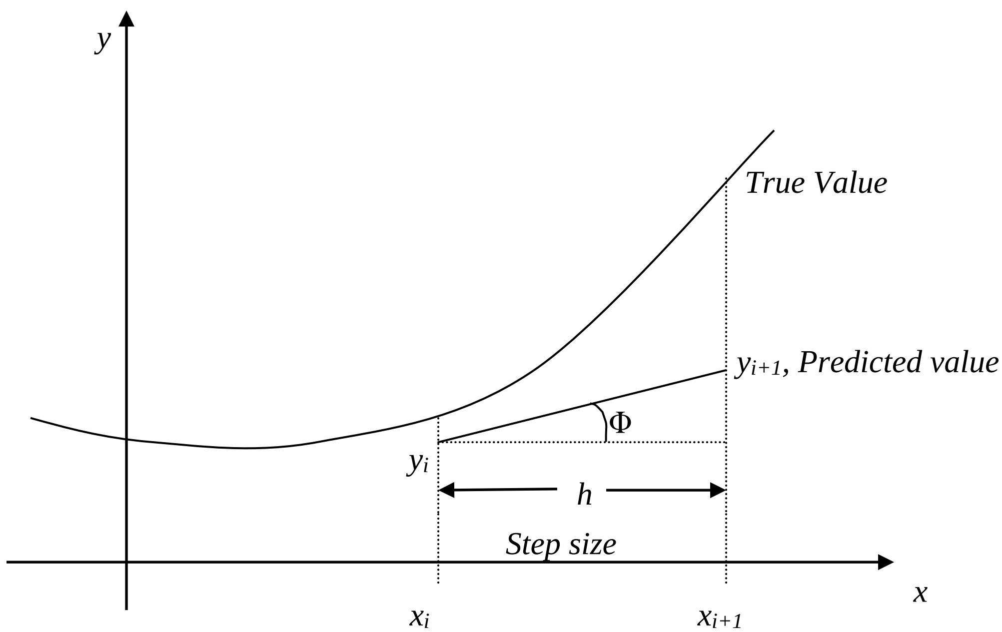

Figure 2 General graphical interpretation of Euler's method.

Example 3

A ball at 1200K is allowed to cool down in air at an ambient temperature of 300K. Assuming heat is lost only due to radiation, the differential equation for the temperature of the ball is given by

$$\frac{d\theta}{dt} = -2.2067 \times 10^{-12}\left(\theta^4 - 81 \times 10^8\right), \ \theta(0) = 1200\text{K}$$

where θ is in K and t in seconds. Find the temperature at $t = 480$ seconds using Euler's method. Assume a step size of $h = 240$ seconds.

Solution

$$\frac{d\theta}{dt} = -2.2067\times10^{-12}\left(\theta^4 - 81\times10^8\right)$$

$$f(t,\theta) = -2.2067\times10^{-12}\left(\theta^4 - 81\times10^8\right)$$

Per Equation (3), Euler's method reduces to

$$\theta_{i+1} = \theta_i + f(t_i, \theta_i)h$$

For $i = 0$, $t_0 = 0$, $\theta_0 = 1200$

$$\begin{aligned}\theta_1 &= \theta_0 + f(t_0, \theta_0)h \\ &= 1200 + f(0,1200)\times240 \\ &= 1200 + \left(-2.2067\times10^{-12}\left(1200^4 - 81\times10^8\right)\right)\times240 \\ &= 1200 + (-4.5579)\times240 \\ &= 106.09\text{K}\end{aligned}$$

θ_1 is the approximate temperature at

$$t = t_1 = t_0 + h = 0 + 240 = 240$$

$$\theta_1 = \theta(240) \approx 106.09\text{K}$$

For $i = 1$, $t_1 = 240$, $\theta_1 = 106.09$

$$\begin{aligned}\theta_2 &= \theta_1 + f(t_1, \theta_1)h \\ &= 106.09 + f(240,106.09)\times240 \\ &= 106.09 + \left(-2.2067\times10^{-12}\left(106.09^4 - 81\times10^8\right)\right)\times240 \\ &= 106.09 + (0.017595)\times240 \\ &= 110.32\text{K}\end{aligned}$$

θ_2 is the approximate temperature at

$$t = t_2 = t_1 + h = 240 + 240 = 480$$

$$\theta_2 = \theta(480) \approx 110.32\text{K}$$

Figure 3 compares the exact solution with the numerical solution from Euler's method for the step size of $h = 240$.

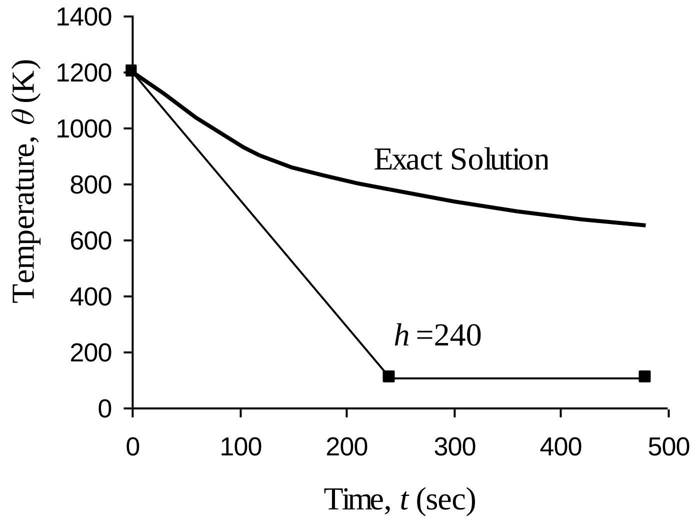

Figure 3 Comparing the exact solution and Euler's method.

The problem was solved again using a smaller step size. The results are given below in Table 1.

Table 1 Temperature at 480 seconds as a function of step size, h.

Step size, h	$\theta(480)$	E_t	$\lvert\in_t\rvert\,\%$
480	-987.81	1635.4	252.54
240	110.32	537.26	82.964
120	546.77	100.80	15.566
60	614.97	32.607	5.0352
30	632.77	14.806	2.2864

Figure 4 shows how the temperature varies as a function of time for different step sizes.

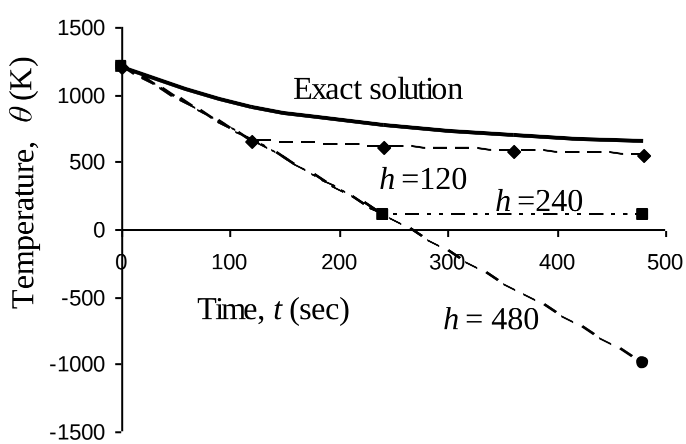

Figure 4 Comparison of Euler's method with the exact solution for different step sizes.

The values of the calculated temperature at $t = 480\text{s}$ as a function of step size are plotted in Figure 5.

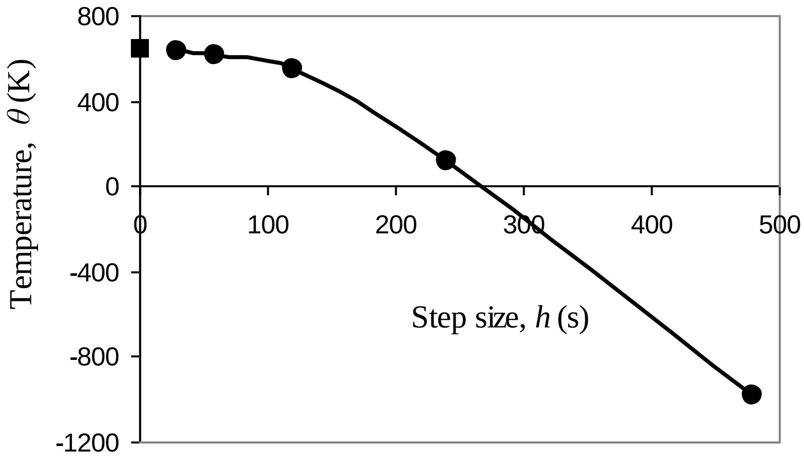

Figure 5 Effect of step size in Euler's method.

The exact solution of the ordinary differential equation is given by the solution of a non-linear equation as

$$0.92593 \ln \frac{\theta - 300}{\theta + 300} - 1.8519 \tan^{-1}\left(0.333 \times 10^{-2}\theta\right) = -0.22067 \times 10^{-3} t - 2.9282 \qquad (4)$$

The solution to this nonlinear equation is

$$\theta = 647.57\text{K}$$

It can be seen that Euler's method has large errors. This can be illustrated using the Taylor series.

$$y_{i+1} = y_i + \left.\frac{dy}{dx}\right|_{x_i, y_i} (x_{i+1} - x_i) + \frac{1}{2!}\left.\frac{d^2 y}{dx^2}\right|_{x_i, y_i} (x_{i+1} - x_i)^2 + \frac{1}{3!}\left.\frac{d^3 y}{dx^3}\right|_{x_i, y_i} (x_{i+1} - x_i)^3 + \ldots \qquad (5)$$

$$= y_i + f(x_i, y_i)(x_{i+1} - x_i) + \frac{1}{2!} f'(x_i, y_i)(x_{i+1} - x_i)^2 + \frac{1}{3!} f''(x_i, y_i)(x_{i+1} - x_i)^3 + \ldots \qquad (6)$$

As you can see the first two terms of the Taylor series

$$y_{i+1} = y_i + f(x_i, y_i)h$$

are Euler's method.

The true error in the approximation is given by

$$E_t = \frac{f'(x_i, y_i)}{2!} h^2 + \frac{f''(x_i, y_i)}{3!} h^3 + \ldots \qquad (7)$$

The true error hence is approximately proportional to the square of the step size, that is, as the step size is halved, the true error gets approximately quartered. However from Table 1, we see that as the step size gets halved, the true error only gets approximately halved. This is because the true error, being proportioned to the square of the step size, is the local truncation error, that is, error from one point to the next. The global truncation error is however proportional only to the step size as the error keeps propagating from one point to another.

Can one solve a definite integral using numerical methods such as Euler's method of solving ordinary differential equations?

Let us suppose you want to find the integral of a function $f(x)$

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

Both fundamental theorems of calculus would be used to set up the problem so as to solve it as an ordinary differential equation.

The first fundamental theorem of calculus states that if f is a continuous function in the interval [a,b], and F is the antiderivative of f, then

$$\int_a^b f(x)dx = F(b) - F(a)$$

The second fundamental theorem of calculus states that if f is a continuous function in the open interval D, and a is a point in the interval D, and if

$$F(x) = \int_a^x f(t)dt$$

then

$$F'(x) = f(x)$$

at each point in D.

Asked to find $\int_a^b f(x)dx$, we can rewrite the integral as the solution of an ordinary differential equation (here is where we are using the second fundamental theorem of calculus)

$$\frac{dy}{dx} = f(x), \ y(a) = 0,$$

where then $y(b)$ (here is where we are using the first fundamental theorem of calculus) will give the value of the integral $\int_a^b f(x)dx$.

Example 4

Find an approximate value of

$$\int_5^8 6x^3 dx$$

using Euler's method of solving an ordinary differential equation. Use a step size of $h = 1.5$.

Solution

Given $\int_5^8 6x^3 dx$, we can rewrite the integral as the solution of an ordinary differential equation

$$\frac{dy}{dx} = 6x^3, \; y(5) = 0$$

where $y(8)$ will give the value of the integral $\int_5^8 6x^3 dx$.

$$\frac{dy}{dx} = 6x^3 = f(x, y), \; y(5) = 0$$

The Euler's method equation is

$$y_{i+1} = y_i + f(x_i, y_i)h$$

<u>Step 1</u>

$$i = 0, \; x_0 = 5, \; y_0 = 0$$

$$h = 1.5$$

$$\begin{aligned} x_1 &= x_0 + h \\ &= 5 + 1.5 \\ &= 6.5 \end{aligned}$$

$$\begin{aligned} y_1 &= y_0 + f(x_0, y_0)h \\ &= 0 + f(5,0) \times 1.5 \\ &= 0 + (6 \times 5^3) \times 1.5 \\ &= 1125 \\ &\approx y(6.5) \end{aligned}$$

<u>Step 2</u>

$$i = 1, \; x_1 = 6.5, \; y_1 = 1125$$

$$\begin{aligned} x_2 &= x_1 + h \\ &= 6.5 + 1.5 \\ &= 8 \end{aligned}$$

$$\begin{aligned} y_2 &= y_1 + f(x_1, y_1)h \\ &= 1125 + f(6.5, 1125) \times 1.5 \\ &= 1125 + (6 \times 6.5^3) \times 1.5 \\ &= 3596.625 \\ &\approx y(8) \end{aligned}$$

Hence

$$\begin{aligned} \int_5^8 6x^3 dx &= y(8) - y(5) \\ &\approx 3596.625 - 0 \\ &= 3596.625 \end{aligned}$$

Runge-Kutta 2nd Order Method for Ordinary Differential Equations

What is the Runge-Kutta 2nd order method?

The Runge-Kutta 2nd order method is a numerical technique used to solve an ordinary differential equation of the form

$$\frac{dy}{dx} = f(x, y), y(0) = y_0$$

Only first order ordinary differential equations can be solved by using the Runge-Kutta 2nd order method. In other sections, we will discuss how the Euler and Runge-Kutta methods are used to solve higher order ordinary differential equations or coupled (simultaneous) differential equations.

How does one write a first order differential equation in the above form?

Example 1

Rewrite

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

in

$$\frac{dy}{dx} = f(x, y), \ y(0) = y_0 \text{ form.}$$

Solution

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

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

In this case

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

Example 2

Rewrite

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

in

$$\frac{dy}{dx} = f(x, y), \ y(0) = y_0 \text{ form.}$$

Solution

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

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

In this case

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

Runge-Kutta 2nd order method

Euler's method is given by

$$y_{i+1}=y_i+f(x_i,y_i)h \tag{1}$$

where

$$x_0=0$$

$$y_0=y(x_0)$$

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

To understand the Runge-Kutta 2nd order method, we need to derive Euler's method from the Taylor series.

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

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

As you can see the first two terms of the Taylor series

$$y_{i+1}=y_i+f(x_i,y_i)h$$

are Euler's method and hence can be considered to be the Runge-Kutta 1st order method.

The true error in the approximation is given by

$$E_t=\frac{f'(x_i,y_i)}{2!}h^2+\frac{f''(x_i,y_i)}{3!}h^3+\ldots \tag{3}$$

So what would a 2nd order method formula look like. It would include one more term of the Taylor series as follows.

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

Let us take a generic example of a first order ordinary differential equation

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

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

Now since y is a function of x,

$$f'(x,y)=\frac{\partial f(x,y)}{\partial x}+\frac{\partial f(x,y)}{\partial y}\frac{dy}{dx} \tag{5}$$

$$= \frac{\partial}{\partial x}\left(e^{-2x} - 3y\right) + \frac{\partial}{\partial y}\left[\left(e^{-2x} - 3y\right)\right]\left(e^{-2x} - 3y\right)$$
$$= -2e^{-2x} + (-3)\left(e^{-2x} - 3y\right)$$
$$= -5e^{-2x} + 9y$$

The 2nd order formula for the above example would be

$$y_{i+1} = y_i + f(x_i, y_i)h + \frac{1}{2!} f'(x_i, y_i)h^2$$
$$= y_i + \left(e^{-2x_i} - 3y_i\right)h + \frac{1}{2!}\left(-5e^{-2x_i} + 9y_i\right)h^2$$

However, we already see the difficulty of having to find $f'(x, y)$ in the above method. What Runge and Kutta did was write the 2nd order method as

$$y_{i+1} = y_i + (a_1 k_1 + a_2 k_2)h \tag{6}$$

where

$$k_1 = f(x_i, y_i)$$
$$k_2 = f(x_i + p_1 h, y_i + q_{11} k_1 h) \tag{7}$$

This form allows one to take advantage of the 2nd order method without having to calculate $f'(x, y)$.

So how do we find the unknowns a_1, a_2, p_1 and q_{11}. Without proof (see Appendix for proof), equating Equation (4) and (6) , gives three equations.

$$a_1 + a_2 = 1$$
$$a_2 p_1 = \frac{1}{2}$$
$$a_2 q_{11} = \frac{1}{2}$$

Since we have 3 equations and 4 unknowns, we can assume the value of one of the unknowns. The other three will then be determined from the three equations. Generally the value of a_2 is chosen to evaluate the other three constants. The three values generally used for a_2 are $\frac{1}{2}$, 1 and $\frac{2}{3}$, and are known as Heun's Method, the midpoint method and Ralston's method, respectively.

<u>Heun's Method</u>

Here $a_2 = \frac{1}{2}$ is chosen, giving

$$a_1 = \frac{1}{2}$$
$$p_1 = 1$$
$$q_{11} = 1$$

resulting in

$$y_{i+1} = y_i + \left(\frac{1}{2}k_1 + \frac{1}{2}k_2\right)h \tag{8}$$

where

$$k_1 = f\left(x_i, y_i\right) \tag{9a}$$

$$k_2 = f\left(x_i + h, y_i + k_1 h\right) \tag{9b}$$

This method is graphically explained in Figure 1.

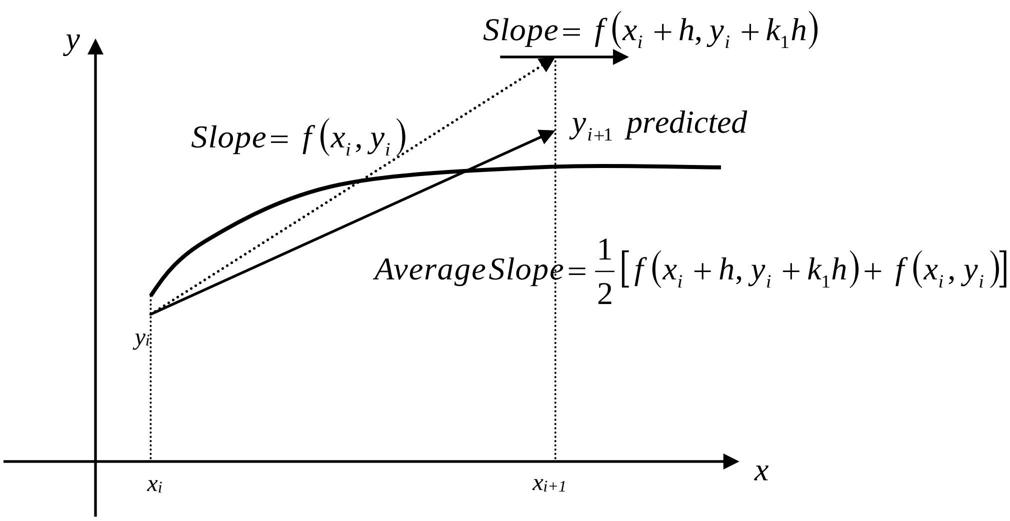

Figure 1 Runge-Kutta 2nd order method (Heun's method).

<u>Midpoint Method</u>

Here $a_2 = 1$ is chosen, giving

$$a_1 = 0$$

$$p_1 = \frac{1}{2}$$

$$q_{11} = \frac{1}{2}$$

resulting in

$$y_{i+1} = y_i + k_2 h \tag{10}$$

where

$$k_1 = f\left(x_i, y_i\right) \tag{11a}$$

$$k_2 = f\left(x_i + \frac{1}{2}h, y_i + \frac{1}{2}k_1 h\right) \tag{11b}$$

<u>Ralston's Method</u>

Here $a_2 = \frac{2}{3}$ is chosen, giving

$$a_1 = \frac{1}{3}$$

$$p_1 = \frac{3}{4}$$

$$q_{11} = \frac{3}{4}$$

resulting in

$$y_{i+1} = y_i + \left(\frac{1}{3}k_1 + \frac{2}{3}k_2\right)h \qquad (12)$$

where

$$k_1 = f\left(x_i, y_i\right) \qquad (13a)$$

$$k_2 = f\left(x_i + \frac{3}{4}h, y_i + \frac{3}{4}k_1 h\right) \qquad (13b)$$

Example 3

A ball at 1200K is allowed to cool down in air at an ambient temperature of 300K. Assuming heat is lost only due to radiation, the differential equation for the temperature of the ball is given by

$$\frac{d\theta}{dt} = -2.2067\times10^{-12}\ (\theta^4 - 81\times10^8)$$

where θ is in K and t in seconds. Find the temperature at $t = 480$ seconds using Runge-Kutta 2nd order method. Assume a step size of $h = 240$ seconds.

Solution

$$\frac{d\theta}{dt} = -2.2067\times10^{-12}\left(\theta^4 - 81\times10^8\right)$$

$$f\left(t,\theta\right) = -2.2067\times10^{-12}\left(\theta^4 - 81\times10^8\right)$$

Per Heun's method given by Equations (8) and (9)

$$\theta_{i+1} = \theta_i + \left(\frac{1}{2}k_1 + \frac{1}{2}k_2\right)h$$

$$k_1 = f\left(t_i, \theta_i\right)$$

$$k_2 = f\left(t_i + h, \theta_i + k_1 h\right)$$

$$i = 0, t_0 = 0, \theta_0 = \theta(0) = 1200$$

$$\begin{aligned}
k_1 &= f\left(t_0, \theta_o\right) \\
&= f\left(0,1200\right) \\
&= -2.2067\times10^{-12}\left(1200^4 - 81\times10^8\right) \\
&= -4.5579
\end{aligned}$$

$$\begin{aligned}
k_2 &= f\left(t_0 + h, \theta_0 + k_1 h\right) \\
&= f\left(0+240, 1200 + \left(-4.5579\right)240\right) \\
&= f\left(240, 106.09\right) \\
&= -2.2067\times10^{-12}\left(106.09^4 - 81\times10^8\right) \\
&= 0.017595
\end{aligned}$$

$$\theta_1 = \theta_0 + \left(\frac{1}{2}k_1 + \frac{1}{2}k_2\right)h$$

$$= 1200 + \left(\frac{1}{2}(-4.5579) + \frac{1}{2}(0.017595)\right)240$$

$$= 1200 + (-2.2702)240$$

$$= 655.16\text{K}$$

$$i = 1, t_1 = t_0 + h = 0 + 240 = 240, \theta_1 = 655.16\text{K}$$

$$k_1 = f(t_1, \theta_1)$$

$$= f(240, 655.16)$$

$$= -2.2067\times10^{-12}(655.16^4 - 81\times10^8)$$

$$= -0.38869$$

$$k_2 = f(t_1 + h, \theta_1 + k_1 h)$$

$$= f(240 + 240, 655.16 + (-0.38869)240)$$

$$= f(480, 561.87)$$

$$= -2.2067\times10^{-12}(561.87^4 - 81\times10^8)$$

$$= -0.20206$$

$$\theta_2 = \theta_1 + \left(\frac{1}{2}k_1 + \frac{1}{2}k_2\right)h$$

$$= 655.16 + \left(\frac{1}{2}(-0.38869) + \frac{1}{2}(-0.20206)\right)240$$

$$= 655.16 + (-0.29538)240$$

$$= 584.27\text{K}$$

$$\theta_2 = \theta(480) = 584.27\text{K}$$

The results from Heun's method are compared with exact results in Figure 2.

The exact solution of the ordinary differential equation is given by the solution of a non-linear equation as

$$0.92593\ln\frac{\theta - 300}{\theta + 300} - 1.8519\tan^{-1}(0.0033333\theta) = -0.22067\times10^{-3}t - 2.9282$$

The solution to this nonlinear equation at $t = 480\text{s}$ is

$$\theta(480) = 647.57\text{ K}$$

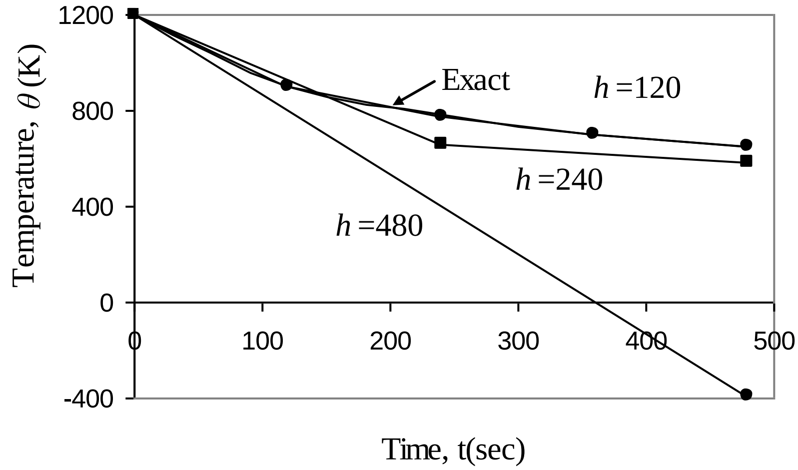

Figure 2 Heun's method results for different step sizes.

Using a smaller step size would increase the accuracy of the result as given in Table 1 and Figure 3 below.

Table 1 Effect of step size for Heun's method

Step size, h	$\theta(480)$	E_t	$\lvert \in_t \rvert \%$
480	-393.87	1041.4	160.82
240	584.27	63.304	9.7756
120	651.35	-3.7762	0.58313
60	649.91	-2.3406	0.36145
30	648.21	-0.63219	0.097625

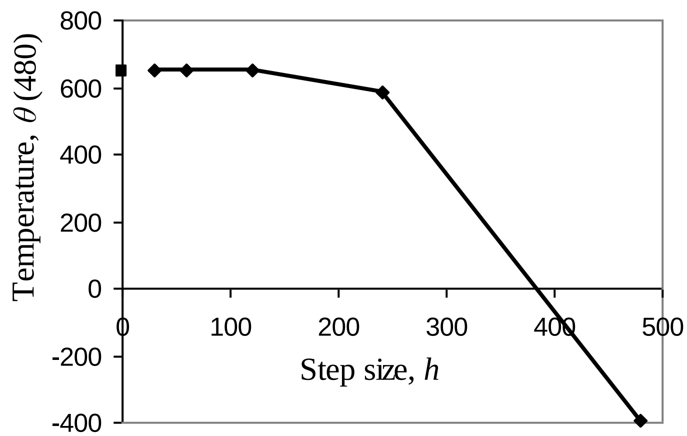

Figure 3 Effect of step size in Heun's method.

In Table 2, Euler's method and the Runge-Kutta 2nd order method results are shown as a function of step size,

Table 2 Comparison of Euler and the Runge-Kutta methods

Step size, h	$\theta(480)$			
	Euler	Heun	Midpoint	Ralston
480	-987.84	-393.87	1208.4	449.78
240	110.32	584.27	976.87	690.01
120	546.77	651.35	690.20	667.71
60	614.97	649.91	654.85	652.25
30	632.77	648.21	649.02	648.61

while in Figure 4, the comparison is shown over the range of time.

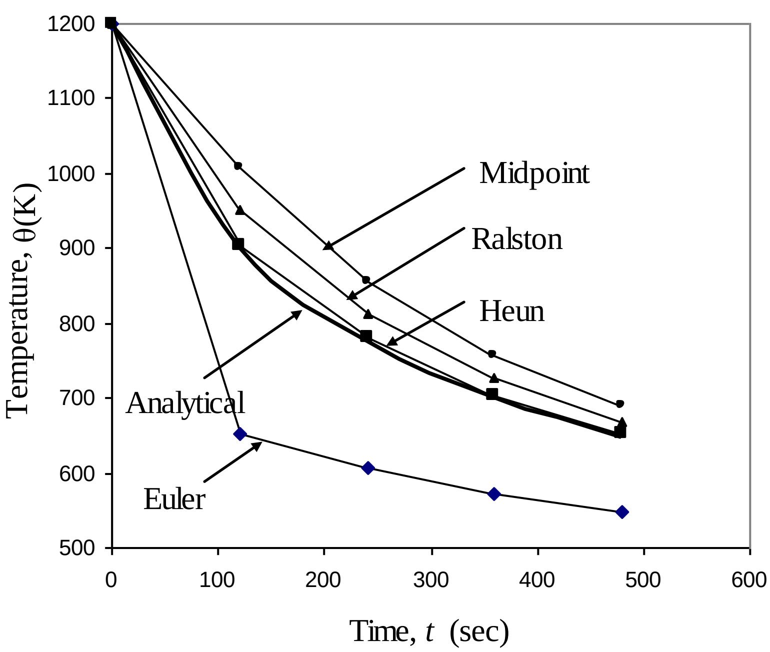

Figure 4 Comparison of Euler and Runge Kutta methods with exact results over time.

How do these three methods compare with results obtained if we found $f'(x,y)$ directly?

Of course, we know that since we are including the first three terms in the series, if the solution is a polynomial of order two or less (that is, quadratic, linear or constant), any of the three methods are exact. But for any other case the results will be different.

Let us take the example of

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

If we directly find $f'(x,y)$, the first three terms of the Taylor series gives

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

where

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

$$f'(x, y) = -5e^{-2x} + 9y$$

For a step size of $h = 0.2$, using Heun's method, we find

$$y(0.6) = 1.0930$$

The exact solution

$$y(x) = e^{-2x} + 4e^{-3x}$$

gives

$$y(0.6) = e^{-2(0.6)} + 4e^{-3(0.6)}$$

$$= 0.96239$$

Then the absolute relative true error is

$$|\in_t| = \left|\frac{0.96239 - 1.0930}{0.96239}\right| \times 100$$

$$= 13.571\%$$

For the same problem, the results from Euler's method and the three Runge-Kutta methods are given in Table 3.

Table 3 Comparison of Euler's and Runge-Kutta 2nd order methods

	y(0.6)					
	Exact	Euler	Direct 2nd	Heun	Midpoint	Ralston
Value	0.96239	0.4955	1.0930	1.1012	1.0974	1.0994
$\|\in_t\|$ %		48.514	13.571	14.423	14.029	14.236

Reference

ORDINARY DIFFERENTIAL EQUATIONS	
Topic	Runge 2nd Order Method for Ordinary Differential Equations
Summary	Textbook notes on Runge 2nd order method for ODE
Major	General Engineering
Authors	Autar Kaw

Runge-Kutta 4th Order Method for Ordinary Differential Equations

What is the Runge-Kutta 4th order method?

Runge-Kutta 4^{th} order method is a numerical technique used to solve ordinary differential equation of the form

$$\frac{dy}{dx} = f(x, y), y(0) = y_0$$

So only first order ordinary differential equations can be solved by using the Runge-Kutta 4^{th} order method. In other sections, we have discussed how Euler and Runge-Kutta methods are used to solve higher order ordinary differential equations or coupled (simultaneous) differential equations.

How does one write a first order differential equation in the above form?

Example 1

Rewrite

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

in

$$\frac{dy}{dx} = f(x, y), \ y(0) = y_0 \text{ form.}$$

Solution

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

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

In this case

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

Example 2

Rewrite

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

in

$$\frac{dy}{dx}= f(x,y),\ \ y(0)=y_0 \text{ form.}$$

Solution

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

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

In this case

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

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

$$y_{i+1}=y_i+(a_1k_1+a_2k_2+a_3k_3+a_4k_4)h \tag{1}$$

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

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

Equation (1) is equated to the first five terms of Taylor series

$$\begin{aligned} y_{i+1}=y_i+\frac{dy}{dx}\Big|_{x_i,y_i}(x_{i+1}-x_i)+\frac{1}{2!}\frac{d^2y}{dx^2}\Big|_{x_i,y_i}(x_{i+1}-x_i)^2+\frac{1}{3!}\frac{d^3y}{dx^3}\Big|_{x_i,y_i}(x_{i+1}-x_i)^3 \\ +\frac{1}{4!}\frac{d^4y}{dx^4}\Big|_{x_i,y_i}(x_{i+1}-x_i)^4 \end{aligned} \tag{2}$$

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

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

Based on equating Equation (2) and Equation (3), one of the popular solutions used is

$$y_{i+1}=y_i+\frac{1}{6}(k_1+2k_2+2k_3+k_4)h \tag{4}$$

$$k_1=f(x_i,y_i) \tag{5a}$$

$$k_2=f\left(x_i+\frac{1}{2}h,y_i+\frac{1}{2}k_1h\right) \tag{5b}$$

$$k_3=f\left(x_i+\frac{1}{2}h,y_i+\frac{1}{2}k_2h\right) \tag{5c}$$

$$k_4=f(x_i+h,y_i+k_3h) \tag{5d}$$

Example 3

A ball at 1200 K is allowed to cool down in air at an ambient temperature of 300 K. Assuming heat is lost only due to radiation, the differential equation for the temperature of the ball is given by

$$\frac{d\theta}{dt} = -2.2067\times10^{-12}\left(\theta^4 - 81\times10^8\right), \theta(0) = 1200\text{K}$$

where θ is in K and t in seconds. Find the temperature at $t = 480$ seconds using Runge-Kutta 4th order method. Assume a step size of $h = 240$ seconds.

Solution

$$\frac{d\theta}{dt} = -2.2067\times10^{-12}\left(\theta^4 - 81\times10^8\right)$$

$$f(t,\theta) = -2.2067\times10^{-12}\left(\theta^4 - 81\times10^8\right)$$

$$\theta_{i+1} = \theta_i + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)h$$

For $i = 0$, $t_0 = 0$, $\theta_0 = 1200\text{K}$

$$\begin{aligned}
k_1 &= f(t_0, \theta_0) \\
&= f(0,1200) \\
&= -2.2067\times10^{-12}\left(1200^4 - 81\times10^8\right) \\
&= -4.5579
\end{aligned}$$

$$\begin{aligned}
k_2 &= f\left(t_0 + \frac{1}{2}h, \theta_0 + \frac{1}{2}k_1 h\right) \\
&= f\left(0 + \frac{1}{2}(240), 1200 + \frac{1}{2}(-4.5579)\times 240\right) \\
&= f(120, 653.05) \\
&= -2.2067\times10^{-12}\left(653.05^4 - 81\times10^8\right) \\
&= -0.38347
\end{aligned}$$

$$\begin{aligned}
k_3 &= f\left(t_0 + \frac{1}{2}h, \theta_0 + \frac{1}{2}k_2 h\right) \\
&= f\left(0 + \frac{1}{2}(240), 1200 + \frac{1}{2}(-0.38347)\times 240\right) \\
&= f(120, 1154.0) \\
&= -2.2067\times10^{-12}\left(1154.0^4 - 81\times10^8\right) \\
&= -3.8954
\end{aligned}$$

$$\begin{aligned}
k_4 &= f(t_0 + h, \theta_0 + k_3 h) \\
&= f(0 + 240, 1200 + (-3.894)\times 240) \\
&= f(240, 265.10) \\
&= -2.2067\times10^{-12}\left(265.10^4 - 81\times10^8\right)
\end{aligned}$$

$$= 0.0069750$$

$$\theta_1 = \theta_0 + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)h$$

$$= 1200 + \frac{1}{6}(-4.5579 + 2(-0.38347) + 2(-3.8954) + (0.069750))240$$

$$= 1200 + (-2.1848) \times 240$$

$$= 675.65\text{K}$$

θ_1 is the approximate temperature at

$$t = t_1$$

$$= t_0 + h$$

$$= 0 + 240$$

$$= 240$$

$$\theta_1 = \theta(240)$$

$$\approx 675.65\text{K}$$

For $i = 1, t_1 = 240, \theta_1 = 675.65\text{K}$

$$k_1 = f(t_1, \theta_1)$$

$$= f(240, 675.65)$$

$$= -2.2067 \times 10^{-12}(675.65^4 - 81 \times 10^8)$$

$$= -0.44199$$

$$k_2 = f\left(t_1 + \frac{1}{2}h, \theta_1 + \frac{1}{2}k_1 h\right)$$

$$= f\left(240 + \frac{1}{2}(240), 675.65 + \frac{1}{2}(-0.44199)240\right)$$

$$= f(360, 622.61)$$

$$= -2.2067 \times 10^{-12}(622.61^4 - 81 \times 10^8)$$

$$= -0.31372$$

$$k_3 = f\left(t_1 + \frac{1}{2}h, \theta_1 + \frac{1}{2}k_2 h\right)$$

$$= f\left(240 + \frac{1}{2}(240), 675.65 + \frac{1}{2}(-0.31372) \times 240\right)$$

$$= f(360, 638.00)$$

$$= -2.2067 \times 10^{-12}(638.00^4 - 81 \times 10^8)$$

$$= -0.34775$$

$$k_4 = f(t_1 + h, \theta_1 + k_3 h)$$

$$= f(240 + 240, 675.65 + (-0.34775) \times 240)$$

$$= f(480, 592.19)$$

$$= 2.2067 \times 10^{-12}(592.19^4 - 81 \times 10^8)$$

$$= -0.25351$$

$$\begin{aligned}\theta_2 &= \theta_1 + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)h \\ &= 675.65 + \frac{1}{6}\left(-0.44199 + 2(-0.31372) + 2(-0.34775) + (-0.25351)\right)\times 240 \\ &= 675.65 + \frac{1}{6}(-2.0184)\times 240 \\ &= 594.91\text{K}\end{aligned}$$

θ_2 is the approximate temperature at

$$\begin{aligned}t &= t_2 \\ &= t_1 + h \\ &= 240 + 240 \\ &= 480\end{aligned}$$

$$\begin{aligned}\theta_2 &= \theta(480) \\ &\approx 594.91\text{K}\end{aligned}$$

Figure 1 compares the exact solution with the numerical solution using the Runge-Kutta 4th order method with different step sizes.

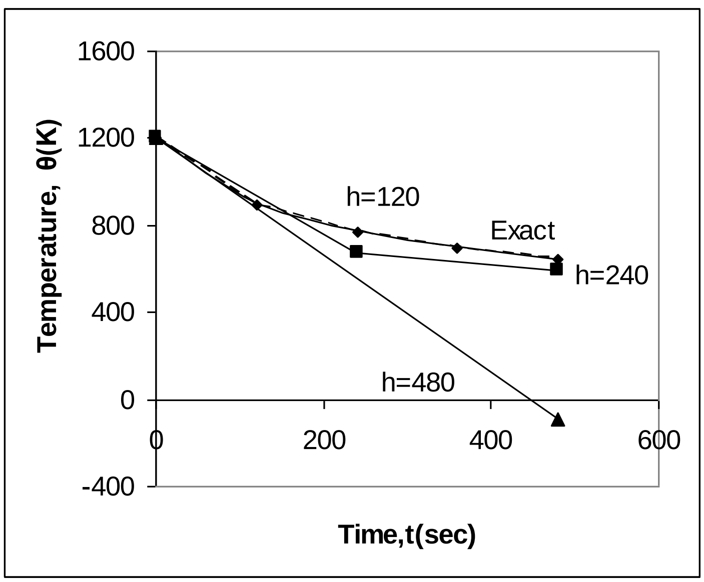

Figure 1 Comparison of Runge-Kutta 4th order method with exact solution for different step sizes.

Table 1 and Figure 2 show the effect of step size on the value of the calculated temperature at $t = 480$ seconds.

Table 1 Value of temperature at time, $t = 480\text{s}$ for different step sizes

Step size, h	$\theta(480)$	E_t	$\lvert \varepsilon_t \rvert \%$
480	-90.278	737.85	113.94
240	594.91	52.660	8.1319
120	646.16	1.4122	0.21807
60	647.54	0.033626	0.0051926
30	647.57	0.00086900	0.00013419

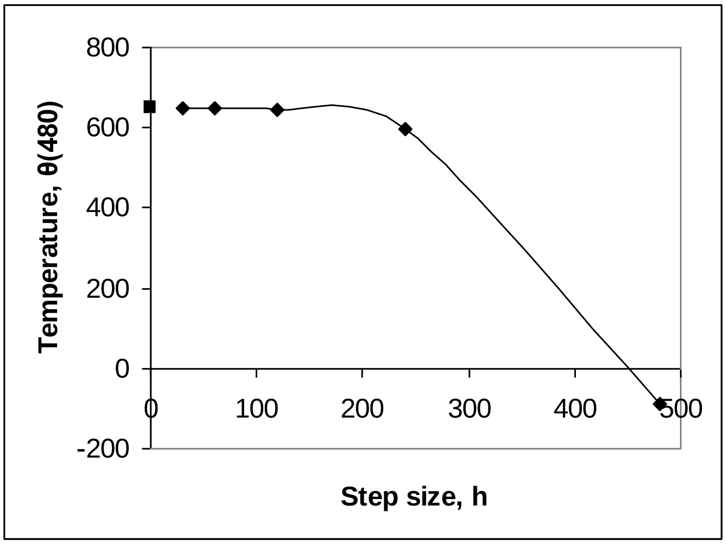

Figure 2 Effect of step size in Runge-Kutta 4th order method.

In Figure 3, we are comparing the exact results with Euler's method (Runge-Kutta 1st order method), Heun's method (Runge-Kutta 2nd order method), and Runge-Kutta 4th order method.

The formula described in this chapter was developed by Runge. This formula is same as Simpson's 1/3 rule, if $f(x, y)$ were only a function of x. There are other versions of the 4th order method just like there are several versions of the second order methods. The formula developed by Kutta is

$$y_{i+1} = y_i + \frac{1}{8}(k_1 + 3k_2 + 3k_3 + k_4)h \tag{6}$$

where

$$k_1 = f(x_i, y_i) \tag{7a}$$

$$k_2 = f\left(x_i + \frac{1}{3}h, y_i + \frac{1}{3}hk_1\right) \tag{7b}$$

$$k_3 = f\left(x_i + \frac{2}{3}h, y_i - \frac{1}{3}hk_1 + hk_2\right) \tag{7c}$$

$$k_4 = f(x_i + h, y_i + hk_1 - hk_2 + hk_3) \tag{7d}$$

This formula is the same as the Simpson's 3/8 rule, if $f(x, y)$ is only a function of x.

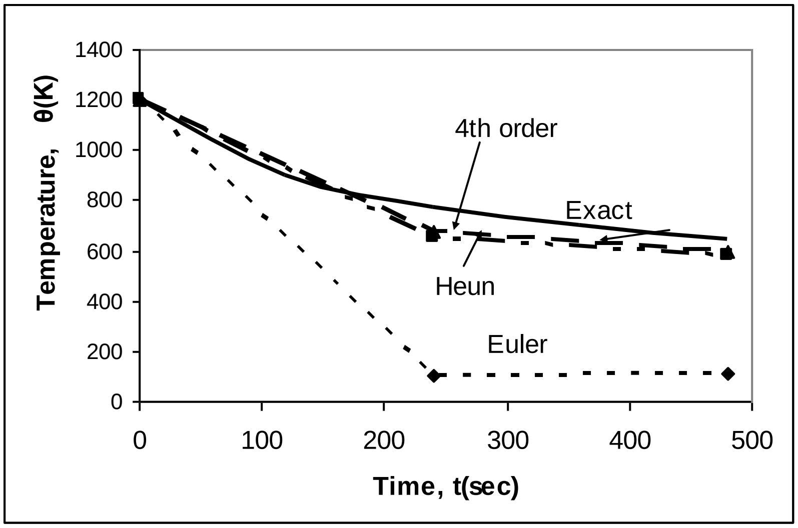

Figure 3 Comparison of Runge-Kutta methods of 1st (Euler), 2nd, and 4th order.

Reference

ORDINARY DIFFERENTIAL EQUATIONS	
Topic	Runge-Kutta 4th order method
Summary	Textbook notes on the Runge-Kutta 4th order method for solving ordinary differential equations.
Major	General Engineering
Authors	Autar Kaw
Last Revised	April 11, 2022

On Solving Higher Order Equations for Ordinary Differential Equations

We have learned Euler's and Runge-Kutta methods to solve first order ordinary differential equations of the form

$$\frac{dy}{dx} = f(x, y),\ y(0) = y_0 \tag{1}$$

What do we do to solve simultaneous (coupled) differential equations, or differential equations that are higher than first order? For example an n^{th} order differential equation of the form

$$a_n \frac{d^n y}{dx^n} + a_{n-1}\frac{d^{n-1}y}{dx^{n-1}} + \ldots + a_1 \frac{dy}{dx} + a_o y = f(x) \tag{2}$$

with $n-1$ initial conditions can be solved by assuming

$$y = z_1 \tag{3.1}$$

$$\frac{dy}{dx} = \frac{dz_1}{dx} = z_2 \tag{3.2}$$

$$\frac{d^2 y}{dx^2} = \frac{dz_2}{dx} = z_3 \tag{3.3}$$

$$\vdots$$

$$\frac{d^{n-1}y}{dx^{n-1}} = \frac{dz_{n-1}}{dx} = z_n \tag{3.n}$$

$$\frac{d^n y}{dx^n} = \frac{dz_n}{dx}$$

$$= \frac{1}{a_n}\left(-a_{n-1}\frac{d^{n-1}y}{dx^{n-1}} \ldots - a_1 \frac{dy}{dx} - a_0 y + f(x)\right)$$

$$= \frac{1}{a_n}\left(-a_{n-1}z_n \ldots - a_1 z_2 - a_0 z_1 + f(x)\right) \tag{3.n+1}$$

The above Equations from (3.1) to (3.n+1) represent n first order differential equations as follows

$$\frac{dz_1}{dx} = z_2 = f_1(z_1, z_2, \ldots, x) \tag{4.1}$$

$$\frac{dz_2}{dx} = z_3 = f_2(z_1, z_2, \ldots, x) \tag{4.2}$$

$$\vdots$$

$$\frac{dz_n}{dx} = \frac{1}{a_n}\left(-a_{n-1}z_n \ldots - a_1 z_2 - a_0 z_1 + f(x)\right) \tag{4.n}$$

Each of the n first order ordinary differential equations are accompanied by one initial condition. These first order ordinary differential equations are simultaneous in nature but can be solved by the methods used for solving first order ordinary differential equations that we have already learned.

Example 1

Rewrite the following differential equation as a set of first order differential equations.

$$3\frac{d^2y}{dx^2} + 2\frac{dy}{dx} + 5y = e^{-x},\ y(0) = 5,\ y'(0) = 7$$

Solution

The ordinary differential equation would be rewritten as follows. Assume

$$\frac{dy}{dx} = z,$$

Then

$$\frac{d^2y}{dx^2} = \frac{dz}{dx}$$

Substituting this in the given second order ordinary differential equation gives

$$3\frac{dz}{dx} + 2z + 5y = e^{-x}$$

$$\frac{dz}{dx} = \frac{1}{3}\left(e^{-x} - 2z - 5y\right)$$

The set of two simultaneous first order ordinary differential equations complete with the initial conditions then is

$$\frac{dy}{dx} = z,\ y(0) = 5$$

$$\frac{dz}{dx} = \frac{1}{3}\left(e^{-x} - 2z - 5y\right),\ z(0) = 7.$$

Now one can apply any of the numerical methods used for solving first order ordinary differential equations.

Example 2

Given

$$\frac{d^2y}{dt^2} + 2\frac{dy}{dt} + y = e^{-t},\ y(0) = 1,\ \frac{dy}{dt}(0) = 2,$$ find by Euler's method

a) $y(0.75)$

b) the absolute relative true error for part(a), if $y(0.75)\big|_{exact} = 1.668$

c) $\frac{dy}{dt}(0.75)$

Use a step size of $h = 0.25$.

Solution

First, the second order differential equation is written as two simultaneous first-order differential equations as follows. Assume

$$\frac{dy}{dt} = z$$

then

$$\frac{dz}{dt} + 2z + y = e^{-t}$$

$$\frac{dz}{dt} = e^{-t} - 2z - y$$

So the two simultaneous first order differential equations are

$$\frac{dy}{dt} = z = f_1(t,y,z),\ y(0) = 1 \qquad \text{(E2.1)}$$

$$\frac{dz}{dt} = e^{-t} - 2z - y = f_2(t,y,z),\ \ z(0) = 2 \qquad \text{(E2.2)}$$

Using Euler's method on Equations (E2.1) and (E2.2), we get

$$y_{i+1} = y_i + f_1(t_i, y_i, z_i)h \qquad \text{(E2.3)}$$

$$z_{i+1} = z_i + f_2(t_i, y_i, z_i)h \qquad \text{(E2.4)}$$

a) To find the value of $y(0.75)$ and since we are using a step size of 0.25 and starting at $t = 0$, we need to take three steps to find the value of $y(0.75)$.

For $i = 0, t_0 = 0,\ y_0 = 1,\ z_0 = 2$,

From Equation (E2.3)

$$\begin{aligned} y_1 &= y_0 + f_1(t_0, y_0, z_0)h \\ &= 1 + f_1(0,1,2)(0.25) \\ &= 1 + 2(0.25) \\ &= 1.5 \end{aligned}$$

y_1 is the approximate value of y at

$$t = t_1 = t_0 + h = 0 + 0.25 = 0.25$$

$$y_1 = y(0.25) \approx 1.5$$

From Equation (E2.4)

$$\begin{aligned} z_1 &= z_0 + f_2(t_0, y_0, z_0)h \\ &= 2 + f_2(0,1,2)(0.25) \\ &= 2 + \left(e^{-0} - 2(2) - 1\right)(0.25) \\ &= 1 \end{aligned}$$

z_1 is the approximate value of z (same as $\frac{dy}{dt}$) at $t = 0.25$

$$z_1 = z(0.25) \approx 1$$

For $i=1,\ t_1=0.25,\ y_1=1.5,\ z_1=1$,

From Equation (E2.3)

$$y_2 = y_1 + f_1(t_1, y_1, z_1)h$$
$$=1.5+ f_1(0.25,1.5,1)(0.25)$$
$$=1.5+(1)(0.25)$$
$$=1.75$$

y_2 is the approximate value of y at

$$t=t_2=t_1+h=0.25+0.25=0.50$$
$$y_2 = y(0.5)\approx 1.75$$

From Equation (E2.4)

$$z_2 = z_1 + f_2(t_1, y_1, z_1)h$$
$$=1+ f_2(0.25,1.5,1)(0.25)$$
$$=1+\left(e^{-0.25}-2(1)-1.5\right)(0.25)$$
$$=1+(-2.7211)(0.25)$$
$$=0.31970$$

z_2 is the approximate value of z at

$$t=t_2=0.5$$
$$z_2=z(0.5)\approx 0.31970$$

For $i=2,\ t_2=0.5,\ y_2=1.75,\ z_2=0.31970$,

From Equation (E2.3)

$$y_3 = y_2 + f_1(t_2, y_2, z_2)h$$
$$=1.75+ f_1(0.50,1.75,0.31970)(0.25)$$
$$=1.75+(0.31970)(0.25)$$
$$=1.8299$$

y_3 is the approximate value of y at

$$t=t_3=t_2+h=0.5+0.25=0.75$$
$$y_3=y(0.75)\approx 1.8299$$

From Equation (E2.4)

$$z_3 = z_2 + f_2(t_2, y_2, z_2)h$$
$$=0.31972+ f_2(0.50,1.75,0.31970)(0.25)$$
$$=0.31972+\left(e^{-0.50}-2(0.31970)-1.75\right)(0.25)$$
$$=0.31972+(-1.7829)(0.25)$$
$$=-0.1260$$

z_3 is the approximate value of z at

$$t=t_3=0.75$$
$$z_3=z(0.75)\approx -0.12601$$
$$y(0.75)\approx y_3=1.8299$$

b) The exact value of $y(0.75)$ is

$$y(0.75)\big|_{exact} = 1.668$$

The absolute relative true error in the result from part (a) is

$$|\epsilon_t| = \left|\frac{1.668 - 1.8299}{1.668}\right| \times 100$$

$$= 9.7062\%$$

c) $\frac{dy}{dx}(0.75) = z_3 \approx -0.12601$

Example 3

Given

$$\frac{d^2y}{dt^2} + 2\frac{dy}{dt} + y = e^{-t}, y(0) = 1, \frac{dy}{dt}(0) = 2,$$

find by Heun's method

a) $y(0.75)$

b) $\frac{dy}{dx}(0.75)$.

Use a step size of $h = 0.25$.

Solution

First, the second order differential equation is rewritten as two simultaneous first-order differential equations as follows. Assume

$$\frac{dy}{dt} = z$$

then

$$\frac{dz}{dt} + 2z + y = e^{-t}$$

$$\frac{dz}{dt} = e^{-t} - 2z - y$$

So the two simultaneous first order differential equations are

$$\frac{dy}{dt} = z = f_1(t, y, z), y(0) = 1 \quad \text{(E3.1)}$$

$$\frac{dz}{dt} = e^{-t} - 2z - y = f_2(t, y, z), z(0) = 2 \quad \text{(E3.2)}$$

Using Heun's method on Equations (1) and (2), we get

$$y_{i+1} = y_i + \frac{1}{2}\left(k_1^y + k_2^y\right)h \quad \text{(E3.3)}$$

$$k_1^y = f_1\left(t_i, y_i, z_i\right) \quad \text{(E3.4a)}$$

$$k_2^y = f_1\left(t_i + h, y_i + hk_1^y, z_i + hk_1^z\right) \quad \text{(E 3.4b)}$$

$$z_{i+1} = z_i + \frac{1}{2}\left(k_1^z + k_2^z\right)h \quad \text{(E3.5)}$$

$$k_1^z = f_2\left(t_i, y_i, z_i\right) \quad \text{(E3.6a)}$$

$$k_2^z = f_2\left(t_i + h, y_i + hk_1^y, z_i + hk_1^z\right) \tag{E3.6b}$$

For $i = 0,\ t_o = 0,\ y_o = 1,\ z_o = 2$

From Equation (E3.4a)

$$\begin{aligned} k_1^y &= f_1(t_o, y_o, z_o) \\ &= f_1(0,1,2) \\ &= 2 \end{aligned}$$

From Equation (E3.6a)

$$\begin{aligned} k_1^z &= f_2(t_0, y_0, z_0) \\ &= f_2(0,1,2) \\ &= e^{-0} - 2(2) - 1 \\ &= -4 \end{aligned}$$

From Equation (E3.4b)

$$\begin{aligned} k_2^y &= f_1\left(t_0 + h, y_0 + hk_1^y, z_0 + hk_1^z\right) \\ &= f_1\left(0 + 0.25, 1 + (0.25)(2), 2 + (0.25)(-4)\right) \\ &= f_1(0.25, 1.5, 1) \\ &= 1 \end{aligned}$$

From Equation (E3.6b)

$$\begin{aligned} k_2^z &= f_2\left(t_0 + h, y_0 + hk_1^y, z_0 + hk_1^z\right) \\ &= f_2\left(0 + 0.25, 1 + (0.25)(2), 2 + (0.25)(-4)\right) \\ &= f_2(0.25, 1.5, 1) \\ &= e^{-0.25} - 2(1) - 1.5 \\ &= -2.7212 \end{aligned}$$

From Equation (E3.3)

$$\begin{aligned} y_1 &= y_0 + \frac{1}{2}\left(k_1^y + k_2^y\right)h \\ &= 1 + \frac{1}{2}(2 + 1)(0.25) \\ &= 1.375 \end{aligned}$$

y_1 is the approximate value of y at

$$t = t_1 = t_0 + h = 0 + 0.25 = 0.25$$

$$y_1 = y(0.25) \cong 1.375$$

From Equation (E3.5)

$$\begin{aligned} z_1 &= z_0 + \frac{1}{2}\left(k_1^z + k_2^z\right)h \\ &= 2 + \frac{1}{2}(-4 + (-2.7212))(0.25) \\ &= 1.1598 \end{aligned}$$

z_1 is the approximate value of z at

$$t = t_1 = 0.25$$

$$z_1 = z(0.25) \approx 1.1598$$

For $i=1,\ t_1=0.25,\ y_1=1.375,\ z_1=1.1598$

From Equation (E3.4a)

$$k_1^y = f_1(t_1, y_1, z_1)$$
$$= f_1(0.25, 1.375, 1.1598)$$
$$= 1.1598$$

From Equation (E3.6a)

$$k_1^z = f_2(t_1, y_1, z_1)$$
$$= f_2(0.25, 1.375, 1.1598)$$
$$= e^{-0.25} - 2(1.1598) - 1.375$$
$$= -2.9158$$

From Equation (E3.4b)

$$k_2^y = f_1\left(t_1 + h, y_1 + hk_1^y, z_1 + hk_1^z\right)$$
$$= f_1(0.25 + 0.25, 1.375 + (0.25)(1.1598), 1.1598 + (0.25)(-2.9158))$$
$$= f_1(0.50, 1.6649, 0.43087)$$
$$= 0.43087$$

From Equation (E3.6b)

$$k_2^z = f_2\left(t_1 + h, y_1 + hk_1^y, z_1 + hk_1^z\right)$$
$$= f_2(0.25 + 0.25, 1.375 + (0.25)(1.1598), 1.1598 + (0.25)(-2.9158))$$
$$= f_2(0.50, 1.6649, 0.43087)$$
$$= e^{-0.50} - 2(0.43087) - 1.6649$$
$$= -1.9201$$

From Equation (E3.3)

$$y_2 = y_1 + \frac{1}{2}\left(k_1^y + k_2^y\right)h$$
$$= 1.375 + \frac{1}{2}(1.1598 + 0.43087)(0.25)$$
$$= 1.5738$$

y_2 is the approximate value of y at

$$t = t_2 = t_1 + h = 0.25 + 0.25 = 0.50$$
$$y_2 = y(0.50) \approx 1.5738$$

From Equation (E3.5)

$$z_2 = z_1 + \frac{1}{2}\left(k_1^z + k_2^z\right)h$$
$$= 1.1598 + \frac{1}{2}(-2.9158 + (-1.9201))(0.25)$$
$$= 0.55533$$

z_2 is the approximate value of z at

$$t = t_2 = 0.50$$
$$z_2 = z(0.50) \approx 0.55533$$

For $i=2,\ t_2=0.50,\ y_2=1.57384,\ z_2=0.55533$

From Equation (E3.4a)

$$k_1^y = f_1(t_2, y_2, z_2)$$
$$= f_1(0.50, 1.5738, 0.55533)$$
$$= 0.55533$$

From Equation (E3.6a)

$$k_1^z = f_2(t_2, y_2, z_2)$$
$$= f_2(0.50, 1.5738, 0.55533)$$
$$= e^{-0.50} - 2(0.55533) - 1.5738$$
$$= -2.0779$$

From Equation (E3.4b)

$$k_2^y = f_2(t_2 + h, y_2 + hk_1^y, z_2 + hk_1^z)$$
$$= f_1(0.50 + 0.25, 1.5738 + (0.25)(0.55533), 0.55533 + (0.25)(-2.0779))$$
$$= f_1(0.75, 1.7126, 0.035836)$$
$$= 0.035836$$

From Equation (E3.6b)

$$k_2^z = f_2(t_2 + h, y_2 + hk_1^y, z_2 + hk_1^z)$$
$$= f_2(0.50 + 0.25, 1.5738 + (0.25)(0.55533), 0.55533 + (0.25)(-2.0779))$$
$$= f_2(0.75, 1.7126, 0.035836)$$
$$= e^{-0.75} - 2(0.035836) - 1.7126$$
$$= -1.3119$$

From Equation (E3.3)

$$y_3 = y_2 + \frac{1}{2}(k_1^y + k_2^y)h$$
$$= 1.5738 + \frac{1}{2}(0.55533 + 0.035836)(0.25)$$
$$= 1.6477$$

y_3 is the approximate value of y at

$$t = t_3 = t_2 + h = 0.50 + 0.25 = 0.75$$
$$y_3 = y(0.75) \approx 1.6477$$

b) From Equation (E3.5)

$$z_3 = z_2 + \frac{1}{2}(k_1^z + k_2^z)h$$
$$= 0.55533 + \frac{1}{2}(-2.0779 + (-1.3119))(0.25)$$
$$= 0.13158$$

z_3 is the approximate value of z at

$$t = t_3 = 0.75$$
$$z_3 = z(0.75) \cong 0.13158$$

The intermediate and the final results are shown in Table 1.

Table 1 Intermediate results of Heun's method.

i	0	1	2
t_i	0	0.25	0.50
y_i	1	1.3750	1.5738
z_i	2	1.1598	0.55533
k_1^y	2	1.1598	0.55533
k_1^z	-4	-2.9158	-2.0779
k_2^y	1	0.43087	0.035836
k_2^z	-2.7211	-1.9201	-1.3119
y_{i+1}	1.3750	1.5738	1.6477
z_{i+1}	1.1598	0.55533	0.13158

Reference

ORDINARY DIFFERENTIAL EQUATIONS	
Topic	Higher Order Equations
Summary	Textbook notes on higher order differential equations
Major	General Engineering
Authors	Autar Kaw
Last Revised	April 12, 2022

Finite Difference Method for Ordinary Differential Equations

What is the finite difference method?

The finite difference method is used to solve ordinary differential equations that have conditions imposed on the boundary rather than at the initial point. These problems are called boundary-value problems. In this chapter, we solve second-order ordinary differential equations of the form

$$\frac{d^2y}{dx^2} = f(x, y, y'), a \le x \le b, \tag{1}$$

with boundary conditions

$$y(a) = y_a \text{ and } y(b) = y_b \tag{2}$$

Many academics refer to boundary value problems as position-dependent and initial value problems as time-dependent. That is not necessarily the case as illustrated by the following examples.

The differential equation that governs the deflection y of a simply supported beam under uniformly distributed load (Figure 1) is given by

$$\frac{d^2y}{dx^2} = \frac{qx(L - x)}{2EI} \tag{3}$$

where

$x =$ location along the beam (in)
$E =$ Young's modulus of elasticity of the beam (psi)
$I =$ second moment of area (in^4)
$q =$ uniform loading intensity (lb/in)
$L =$ length of beam (in)

The conditions imposed to solve the differential equation are

$$y(x = 0) = 0 \tag{4}$$
$$y(x = L) = 0$$

Clearly, these are boundary values and hence the problem is considered a boundary-value problem.

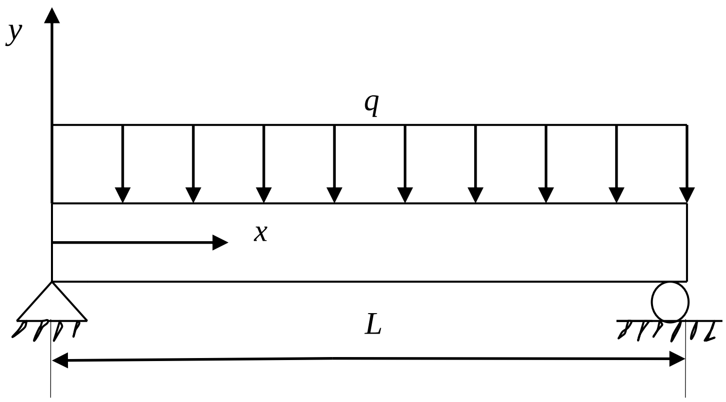

Figure 1 Simply supported beam with uniform distributed load.

Now consider the case of a cantilevered beam with a uniformly distributed load (Figure 2). The differential equation that governs the deflection y of the beam is given by

$$\frac{d^2y}{dx^2} = \frac{q(L-x)^2}{2EI} \tag{5}$$

where

$x =$ location along the beam (in)
$E =$ Young's modulus of elasticity of the beam (psi)
$I =$ second moment of area (in^4)
$q =$ uniform loading intensity (lb/in)
$L =$ length of beam (in)

The conditions imposed to solve the differential equation are

$$y(x=0) = 0 \tag{6}$$

$$\frac{dy}{dx}(x=0) = 0$$

Clearly, these are initial values and hence the problem needs to be considered as an initial value problem.

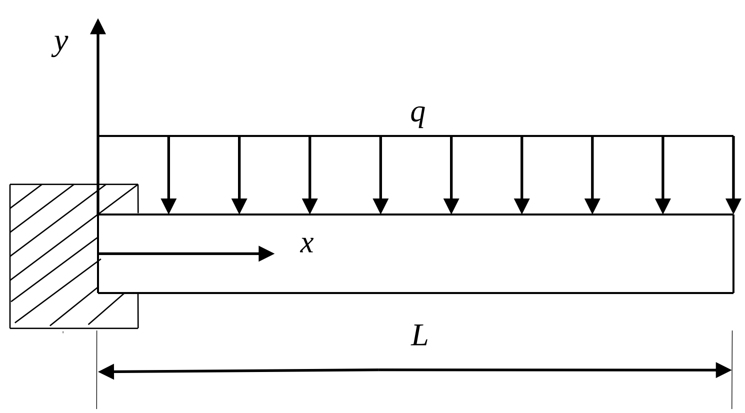

Figure 2 Cantilevered beam with a uniformly distributed load.

Example 1

The deflection y in a simply supported beam with a uniform load q and a tensile axial load T is given by

$$\frac{d^2y}{dx^2} - \frac{Ty}{EI} = \frac{qx(L-x)}{2EI} \quad \text{(E1.1)}$$

where

$x =$ location along the beam (in)
$T =$ tension applied (lbs)
$E =$ Young's modulus of elasticity of the beam (psi)
$I =$ second moment of area (in^4)
$q =$ uniform loading intensity (lb/in)
$L =$ length of beam (in)

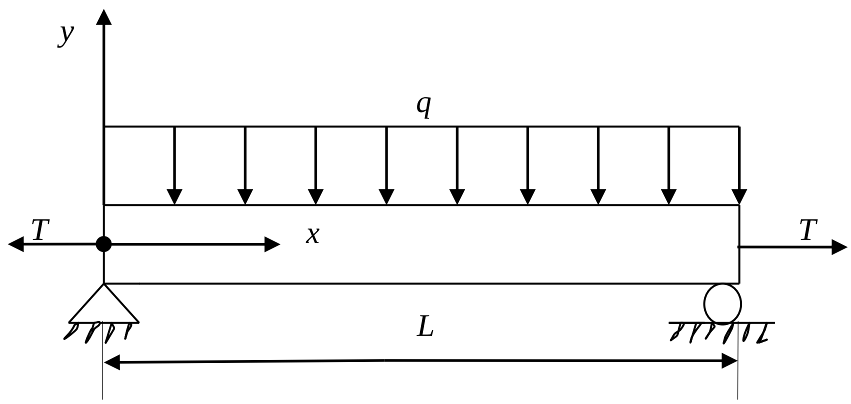

Figure 3 Simply supported beam for Example 1.

Given,

$T = 7200\text{lbs},\ q = 5400\text{lbs/in},\ L = 75\,\text{in},\ E = 30\,\text{Msi}$, and $I = 120\,\text{in}^4$,

a) Find the deflection of the beam at $x = 50''$. Use a step size of $\Delta x = 25''$ and approximate the derivatives by central divided difference approximation.
b) Find the relative true error in the calculation of $y(50)$.

Solution

a) Substituting the given values,

$$\frac{d^2y}{dx^2} - \frac{7200y}{(30\times10^6)(120)} = \frac{(5400)x(75-x)}{2(30\times10^6)(120)}$$

$$\frac{d^2y}{dx^2} - 2\times10^{-6}y = 7.5\times10^{-7}x(75-x) \quad \text{(E1.2)}$$

Approximating the derivative $\frac{d^2y}{dx^2}$ at node i by the central divided difference approximation,

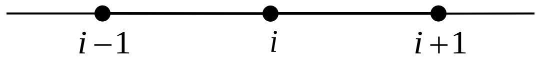

Figure 4 Illustration of finite difference nodes using central divided difference method.

$$\frac{d^2y}{dx^2} \approx \frac{y_{i+1} - 2y_i + y_{i-1}}{(\Delta x)^2} \quad \text{(E1.3)}$$

We can rewrite the equation as

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{(\Delta x)^2} - 2\times10^{-6} y_i = 7.5\times10^{-7} x_i (75 - x_i) \quad \text{(E1.4)}$$

Since $\Delta x = 25$, we have 4 nodes as given in Figure 3

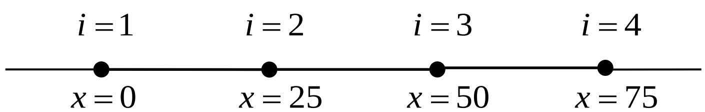

Figure 5 Finite difference method from $x = 0$ to $x = 75$ with $\Delta x = 25$.

The location of the 4 nodes then is

$$x_0 = 0$$

$$x_1 = x_0 + \Delta x = 0 + 25 = 25$$

$$x_2 = x_1 + \Delta x = 25 + 25 = 50$$

$$x_3 = x_2 + \Delta x = 50 + 25 = 75$$

Writing the equation at each node, we get

Node 1: From the simply supported boundary condition at $x = 0$, we obtain

$$y_1 = 0 \quad \text{(E1.5)}$$

Node 2: Rewriting equation (E1.4) for node 2 gives

$$\frac{y_3 - 2y_2 + y_1}{(25)^2} - 2\times10^{-6} y_2 = 7.5\times10^{-7} x_2 (75 - x_2)$$

$$0.0016y_1 - 0.003202y_2 + 0.0016y_3 = 7.5\times10^{-7}(25)(75-25)$$

$$0.0016y_1 - 0.003202y_2 + 0.0016y_3 = 9.375\times10^{-4} \quad \text{(E1.6)}$$

Node 3: Rewriting equation (E1.4) for node 3 gives

$$\frac{y_4 - 2y_3 + y_2}{(25)^2} - 2\times10^{-6} y_3 = 7.5\times10^{-7} x_3 (75 - x_3)$$

$$0.0016y_2 - 0.003202y_3 + 0.0016y_4 = 7.5\times10^{-7}(50)(75-50)$$

$$0.0016y_2 - 0.003202y_3 + 0.0016y_4 = 9.375\times10^{-4} \quad \text{(E1.7)}$$

Node 4: From the simply supported boundary condition at $x = 75$, we obtain

$$y_4 = 0 \quad \text{(E1.8)}$$

Equations (E1.5-E1.8) are 4 simultaneous equations with 4 unknowns and can be written in matrix form as

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.0016 & -0.003202 & 0.0016 & 0 \\ 0 & 0.0016 & -0.003202 & 0.0016 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 9.375\times10^{-4} \\ 9.375\times10^{-4} \\ 0 \end{bmatrix}$$

The above equations have a coefficient matrix that is tridiagonal (we can use Thomas' algorithm to solve the equations) and is also strictly diagonally dominant (convergence is guaranteed if we use iterative methods such as the Gauss-Siedel method). Solving the equations we get,

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 0 \\ -0.5852 \\ -0.5852 \\ 0 \end{bmatrix}$$

$$y(50) = y(x_2) \approx y_2 = -0.5852'$$

The exact solution of the ordinary differential equation is derived as follows. The homogeneous part of the solution is given by solving the characteristic equation

$$m^2 - 2\times10^{-6} = 0$$

$$m = \pm 0.0014142$$

Therefore,

$$y_h = K_1 e^{0.0014142x} + K_2 e^{-0.0014142x}$$

The particular part of the solution is given by

$$y_p = Ax^2 + Bx + C$$

Substituting the differential equation (E1.2) gives

$$\frac{d^2 y_p}{dx^2} - 2\times10^{-6} y_p = 7.5\times10^{-7} x(75-x)$$

$$\frac{d^2}{dx^2}(Ax^2 + Bx + C) - 2\times10^{-6}(Ax^2 + Bx + C) = 7.5\times10^{-7} x(75-x)$$

$$2A - 2\times10^{-6}(Ax^2 + Bx + C) = 7.5\times10^{-7} x(75-x)$$

$$-2\times10^{-6} Ax^2 - 2\times10^{-6} Bx + (2A - 2\times10^{-6} C) = 5.625\times10^{-5} x - 7.5\times10^{-7} x^2$$

Equating terms gives

$$-2\times10^{-6} A = -7.5\times10^{-7}$$

$$-2\times10^{-6} B = -5.625\times10^{-5}$$

$$2A - 2\times10^{-6} C = 0$$

Solving the above equation gives

$$A = 0.375$$

$$B = -28.125$$

$$C = 3.75\times10^{5}$$

The particular solution then is

$$y_p = 0.375x^2 - 28.125x + 3.75\times10^5$$

The complete solution is then given by

$$y = 0.375x^2 - 28.125x + 3.75\times10^5 + K_1e^{0.0014142x} + K_2e^{-0.0014142x}$$

Applying the following boundary conditions

$$y(x=0) = 0$$

$$y(x=75) = 0$$

we obtain the following system of equations

$$K_1 + K_2 = -3.75\times10^5$$

$$1.1119K_1 + 0.89937K_2 = -3.75\times10^5$$

These equations are represented in matrix form by

$$\begin{bmatrix} 1 & 1 \\ 1.1119 & 0.89937 \end{bmatrix}\begin{bmatrix} K_1 \\ K_2 \end{bmatrix} = \begin{bmatrix} -3.75\times10^5 \\ -3.75\times10^5 \end{bmatrix}$$

A number of different numerical methods may be utilized to solve this system of equations such as the Gaussian elimination. Using any of these methods yields

$$\begin{bmatrix} K_1 \\ K_2 \end{bmatrix} = \begin{bmatrix} -1.775656226\times10^5 \\ -1.974343774\times10^5 \end{bmatrix}$$

Substituting these values back into the equation gives

$$y = 0.375x^2 - 28.125x + 3.75\times10^5 - 1.775656266\times10^5e^{0.0014142x} - 1.974343774\times10^5e^{-0.0014142x}$$

Unlike other examples in this chapter and in the book, the above expression for the deflection of the beam is displayed with a larger number of significant digits. This is done to minimize the round-off error because the above expression involves subtraction of large numbers that are close to each other.

b) To calculate the relative true error, we must first calculate the value of the exact solution at $y = 50$.

$$y(50) = 0.375(50)^2 - 28.125(50) + 3.75\times10^5 - 1.775656266\times10^5e^{0.0014142(50)}$$
$$-1.974343774\times10^5e^{-0.0014142(50)}$$

$$y(50) = -0.5320$$

The true error is given by

$$E_t = \text{Exact Value} - \text{Approximate Value}$$

$$E_t = -0.5320 - (-0.5852)$$

$$E_t = 0.05320$$

The relative true error is given by

$$\in_t = \frac{\text{True Error}}{\text{True Value}}\times100\%$$

$$\in_t = \frac{0.05320}{-0.5320}\times100\%$$

$$\in_t = -10\%$$

Example 2

Take the case of a pressure vessel that is being tested in the laboratory to check its ability to withstand pressure. For a thick pressure vessel of inner radius a and outer radius b, the differential equation for the radial displacement u of a point along the thickness is given by

$$\frac{d^2u}{dr^2} + \frac{1}{r}\frac{du}{dr} - \frac{u}{r^2} = 0 \qquad \text{(E2.3)}$$

The inner radius $a = 5''$ and the outer radius $b = 8''$, and the material of the pressure vessel is ASTM A36 steel. The yield strength of this type of steel is 36 ksi. Two strain gages that are bonded tangentially at the inner and the outer radius measure normal tangential strain as

$$\in_{t/r=a} = 0.00077462$$

$$\in_{t/r=b} = 0.00038462 \qquad \text{(E2.4a,b)}$$

at the maximum needed pressure. Since the radial displacement and tangential strain are related simply by

$$\in_t = \frac{u}{r}, \qquad \text{(E2.5)}$$

then

$$u\big|_{r=a} = 0.00077462 \times 5 = 0.0038731'$$

$$u\big|_{r=b} = 0.00038462 \times 8 = 0.0030769'$$

The maximum normal stress in the pressure vessel is at the inner radius $r = a$ and is given by

$$\sigma_{\max} = \frac{E}{1-\nu^2}\left(\left.\frac{u}{r}\right|_{r=a} + \nu\left.\frac{du}{dr}\right|_{r=a}\right) \qquad \text{(E2.7)}$$

where

$E =$ Young's modulus of steel (E= 30 Msi)
$\nu =$ Poisson's ratio ($\nu = 0.3$)

The factor of safety, FS is given by

$$FS = \frac{\text{Yield strength of steel}}{\sigma_{\max}} \qquad \text{(E2.8)}$$

a) Divide the radial thickness of the pressure vessel into 6 equidistant nodes, and find the radial displacement profile
b) Find the maximum normal stress and factor of safety as given by equation (E2.8)
c) Find the exact value of the maximum normal stress as given by equation (E2.8) if it is given that the exact expression for radial displacement is of the form

$$u = C_1 r + \frac{C_2}{r}.$$

Calculate the relative true error.

Solution

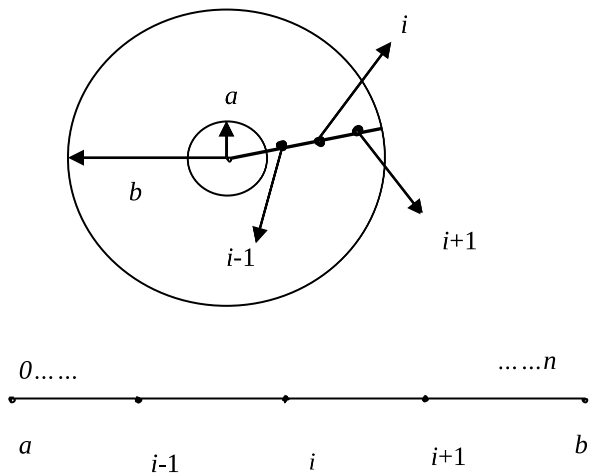

Figure 4 Nodes along the radial direction.

a) The radial locations from $r = a$ to $r = b$ are divided into n equally spaced segments, and hence resulting in $n+1$ nodes. This will allow us to find the dependent variable u numerically at these nodes.
At node i along the radial thickness of the pressure vessel,

$$\frac{d^2u}{dr^2} \approx \frac{u_{i+1} - 2u_i + u_{i-1}}{(\Delta r)^2} \quad \text{(E2.9)}$$

$$\frac{du}{dr} \approx \frac{u_{i+1} - u_i}{\Delta r} \quad \text{(E2.10)}$$

Such substitutions will convert the ordinary differential equation into a linear equation (but with more than one unknown). By writing the resulting linear equation at different points at which the ordinary differential equation is valid, we get simultaneous linear equations that can be solved by using techniques such as Gaussian elimination, the Gauss-Siedel method, etc.
Substituting these approximations from Equations (E2.9) and (E2.10) in Equation (E2.3)

$$\frac{u_{i+1} - 2u_i + u_{i-1}}{(\Delta r)^2} + \frac{1}{r_i}\frac{u_{i+1} - u_i}{\Delta r} - \frac{u_i}{r_i^2} = 0 \quad \text{(E2.11)}$$

$$\left(\frac{1}{(\Delta r)^2} + \frac{1}{r_i \Delta r}\right)u_{i+1} + \left(-\frac{2}{(\Delta r)^2} - \frac{1}{r_i \Delta r} - \frac{1}{r_i^2}\right)u_i + \frac{1}{(\Delta r)^2}u_{i-1} = 0 \quad \text{(E2.12)}$$

Let us break the thickness, $b - a$, of the pressure vessel into $n+1$ nodes, that is $r = a$ is node $i = 0$ and $r = b$ is node $i = n$. That means we have $n+1$ unknowns.
We can write the above equation for nodes $1, \ldots, n-1$. This will give us $n-1$ equations. At the edge nodes, $i = 0$ and $i = n$, we use the boundary conditions of

$$u_0 = u|_{r=a}$$

$$u_n = u|_{r=b}$$

This gives a total of $n+1$ equations. So we have $n+1$ unknowns and $n+1$ linear equations. These can be solved by any of the numerical methods used for solving simultaneous linear equations.

We have been asked to do the calculations for $n = 5$, that is a total of 6 nodes. This gives

$$\Delta r = \frac{b-a}{n}$$

$$= \frac{8-5}{5}$$

$$= 0.6''$$

At node $i = 0, r_0 = a = 5''$, $u_0 = 0.0038731''$ (E2.13)

At node $i = 1, r_1 = r_0 + \Delta r = 5 + 0.6 = 5.6''$ (E2.14)

$$\frac{1}{0.6^2}u_0 + \left(-\frac{2}{0.6^2} - \frac{1}{(5.6)(0.6)} - \frac{1}{(5.6)^2}\right)u_1 + \left(\frac{1}{0.6^2} + \frac{1}{(5.6)(0.6)}\right)u_2 = 0$$

$$2.7778u_0 - 5.8851u_1 + 3.0754u_2 = 0 \quad \text{(E2.15)}$$

At node $i = 2$, $r_2 = r_1 + \Delta r = 5.6 + 0.6 = 6.2''$

$$\frac{1}{0.6^2}u_1 + \left(-\frac{2}{0.6^2} - \frac{1}{(6.2)(0.6)} - \frac{1}{6.2^2}\right)u_2 + \left(\frac{1}{0.6^2} + \frac{1}{(6.2)(0.6)}\right)u_3 = 0$$

$$2.7778u_1 - 5.8504u_2 + 3.0466u_3 = 0 \quad \text{(E2.16)}$$

At node $i = 3$, $r_3 = r_2 + \Delta r = 6.2 + 0.6 = 6.8''$

$$\frac{1}{0.6^2}u_2 + \left(-\frac{2}{0.6^2} - \frac{1}{(6.8)(0.6)} - \frac{1}{6.8^2}\right)u_3 + \left(\frac{1}{0.6^2} + \frac{1}{(6.8)(0.6)}\right)u_4 = 0$$

$$2.7778u_2 - 5.8223u_3 + 3.0229u_4 = 0 \quad \text{(E2.17)}$$

At node $i = 4$, $r_4 = r_3 + \Delta r = 6.8 + 0.6 = 7.4''$

$$\frac{1}{0.6^2}u_3 + \left(-\frac{2}{0.6^2} - \frac{1}{(7.4)(0.6)} - \frac{1}{(7.4)^2}\right)u_4 + \left(\frac{1}{0.6^2} + \frac{1}{(7.4)(0.6)}\right)u_5 = 0$$

$$2.7778u_3 - 5.7990u_4 + 3.0030u_5 = 0 \quad \text{(E2.18)}$$

At node $i = 5$, $r_5 = r_4 + \Delta r = 7.4 + 0.6 = 8''$

$$u_5 = u|_{r=b} = 0.0030769'' \quad \text{(E2.19)}$$

Writing Equation (E2.13) to (E2.19) in matrix form gives

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 2.7778 & -5.8851 & 3.0754 & 0 & 0 & 0 \\ 0 & 2.7778 & -5.8504 & 3.0466 & 0 & 0 \\ 0 & 0 & 2.7778 & -5.8223 & 3.0229 & 0 \\ 0 & 0 & 0 & 2.7778 & -5.7990 & 3.0030 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{bmatrix} = \begin{bmatrix} 0.0038731 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0.0030769 \end{bmatrix}$$

The above equations are a tri-diagonal system of equations and special algorithms such as Thomas' algorithm can be used to solve such a system of equations.

$$u_0 = 0.0038731''$$
$$u_1 = 0.0036165''$$
$$u_2 = 0.0034222''$$
$$u_3 = 0.0032743''$$
$$u_4 = 0.0031618''$$
$$u_5 = 0.0030769''$$

b) To find the maximum stress, it is given by Equation (E2.7) as

$$\sigma_{\max} = \frac{E}{1-\nu^2}\left(\frac{u}{r}\bigg|_{r=a} + \nu \frac{du}{dr}\bigg|_{r=a}\right)$$

$$E = 30 \times 10^6 \text{ psi}$$
$$\nu = 0.3$$
$$u\big|_{r=a} = u_0 = 0.0038731''$$

$$\frac{du}{dr}\bigg|_{r=a} \approx \frac{u_1 - u_0}{\Delta r}$$
$$= \frac{0.0036165 - 0.0038731}{0.6}$$
$$= -0.00042767$$

The maximum stress in the pressure vessel then is

$$\sigma_{\max} = \frac{30 \times 10^6}{1-0.3^2}\left(\frac{0.0038731}{5} + 0.3(-0.00042767)\right)$$
$$= 2.1307 \times 10^4 \text{ psi}$$

So the factor of safety FS from Equation (E2.8) is

$$FS = \frac{36 \times 10^3}{2.1307 \times 10^4} = 1.6896$$

c) The differential equation has an exact solution and is given by the form

$$u = C_1 r + \frac{C_2}{r} \qquad \text{(E2.20)}$$

where C_1 and C_2 are found by using the boundary conditions at $r = a$ and $r = b$.

$$u(r=a)=u(r=5)=0.0038731=C_1(5)+\frac{C_2}{5}$$

$$u(r=b)=u(r=8)=0.0030769=C_1(8)+\frac{C_2}{8}$$

giving

$$C_1=0.00013462$$

$$C_2=0.016000$$

Thus

$$u=0.00013462r+\frac{0.016000}{r} \qquad \text{(E2.21)}$$

$$\frac{du}{dr}=0.00013462-\frac{0.016000}{r^2} \qquad \text{(E2.22)}$$

$$\sigma_{\max}=\frac{E}{1-\nu^2}\left(\left.\frac{u}{r}\right|_{r=a}+\nu\left.\frac{du}{dr}\right|_{r=a}\right)$$

$$=\frac{30\times10^6}{1-0.3^2}\left(\frac{0.00013462(5)+\frac{0.01600}{5}}{5}+0.3\left(0.0013462-\frac{0.016000}{5^2}\right)\right)$$

$$=2.0538\times10^4 \text{ psi}$$

The true error is

$$E_t=2.0538\times10^4-2.1307\times10^4$$

$$=-7.6859\times10^2$$

The absolute relative true error is

$$|\in_t|=\left|\frac{2.0538\times10^4-2.1307\times10^4}{2.0538\times10^4}\right|\times100$$

$$=3.744\%$$

Example 3

The approximation in Example 2

$$\frac{du}{dr}\approx\frac{u_{i+1}-u_i}{\Delta r}$$

is first order accurate, that is , the true error is of $O(\Delta r)$.

The approximation

$$\frac{d^2u}{dr^2}\approx\frac{u_{i+1}-2u_i+u_{i-1}}{(\Delta r)^2} \qquad \text{(E3.1)}$$

is second order accurate, that is , the true error is $O\left((\Delta r)^2\right)$

Mixing these two approximations will result in the order of accuracy of $O(\Delta r)$ and $O\left((\Delta r)^2\right)$, that is $O(\Delta r)$.

So it is better to approximate

$$\frac{du}{dr} \approx \frac{u_{i+1} - u_{i-1}}{2(\Delta r)} \tag{E3.2}$$

because this equation is second order accurate. Repeat Example 2 with the more accurate approximations.

Solution

a) Repeating the problem with this approximation, at node i in the pressure vessel,

$$\frac{d^2u}{dr^2} \approx \frac{u_{i+1} - 2u_i + u_{i-1}}{(\Delta r)^2} \tag{E3.3}$$

$$\frac{du}{dr} \approx \frac{u_{i+1} - u_{i-1}}{2\Delta r} \tag{E3.4}$$

Substituting Equations (E3.3) and (E3.4) in Equation (E2.3) gives

$$\frac{u_{i+1} - 2u_i + u_{i-1}}{(\Delta r)^2} + \frac{1}{r_i}\frac{u_{i+1} - u_{i-1}}{2(\Delta r)} - \frac{u_i}{r_i^2} = 0$$

$$\left(-\frac{1}{2r_i(\Delta r)} + \frac{1}{(\Delta r)^2}\right)u_{i-1} + \left(-\frac{2}{(\Delta r)^2} - \frac{1}{r_i^2}\right)u_i + \left(\frac{1}{(\Delta r)^2} + \frac{1}{2r_i\Delta r}\right)u_{i+1} = 0 \tag{E3.5}$$

At node $i = 0,\ r_0 = a = 5''$

$$u_0 = 0.0038731'' \tag{E3.6}$$

At node $i = 1,\ r_1 = r_0 + \Delta r = 5 + 0.6 = 5.6''$

$$\left(-\frac{1}{2(5.6)(0.6)} + \frac{1}{(0.6)^2}\right)u_0 + \left(-\frac{2}{(0.6)^2} - \frac{1}{(5.6)^2}\right)u_1 + \left(\frac{1}{0.6^2} + \frac{1}{2(5.6)(0.6)}\right)u_2 = 0$$

$$2.6297u_0 - 5.5874u_1 + 2.9266u_2 = 0 \tag{E3.7}$$

At node $i = 2,\ r_2 = r_1 + \Delta r = 5.6 + 0.6 = 6.2''$

$$\left(-\frac{1}{2(6.2)(0.6)} + \frac{1}{0.6^2}\right)u_1 + \left(-\frac{2}{0.6^2} - \frac{1}{6.2^2}\right)u_2 + \left(\frac{1}{0.6^2} + \frac{1}{2(6.2)(0.6)}\right)u_3 = 0 \tag{E3.8}$$

$$2.6434u_1 - 5.5816u_2 + 2.9122u_3 = 0$$

At node $i = 3,\ r_3 = r_2 + \Delta r = 6.2 + 0.6 = 6.8''$

$$\left(-\frac{1}{2(6.8)(0.6)} + \frac{1}{0.6^2}\right)u_2 + \left(-\frac{2}{0.6^2} - \frac{1}{6.8^2}\right)u_3 + \left(\frac{1}{0.6^2} + \frac{1}{2(6.8)(0.6)}\right)u_4 = 0 \tag{E3.9}$$

$$2.6552u_2 - 5.5772u_3 + 2.9003u_4 = 0$$

At node $i = 4,\ r_4 = r_3 + \Delta r = 6.8 + 0.6 = 7.4''$

$$\left(-\frac{1}{2(7.4)(0.6)} + \frac{1}{0.6^2}\right)u_3 + \left(-\frac{2}{0.6^2} - \frac{1}{(7.4)^2}\right)u_4 + \left(\frac{1}{0.6^2} + \frac{1}{2(7.4)(0.6)}\right)u_5 = 0 \tag{E3.10}$$

$$2.6651u_3 - 5.5738u_4 + 2.8903u_5 = 0$$

At node $i = 5,\ r_5 = r_4 + \Delta r = 7.4 + 0.6 = 8''$

$$u_5 = u/_{r=b} = 0.0030769'' \tag{E3.11}$$

Writing Equations (E3.6) thru (E3.11) in matrix form gives

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 2.6297 & -5.5874 & 2.9266 & 0 & 0 & 0 \\ 0 & 2.6434 & -5.5816 & 2.9122 & 0 & 0 \\ 0 & 0 & 2.6552 & -5.5772 & 2.9003 & 0 \\ 0 & 0 & 0 & 2.6651 & -5.5738 & 2.8903 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{bmatrix} = \begin{bmatrix} 0.0038731 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0.0030769 \end{bmatrix}$$

The above equations are a tri-diagonal system of equations and special algorithms such as Thomas' algorithm can be used to solve such equations.

$$u_0 = 0.0038731''$$
$$u_1 = 0.0036115''$$
$$u_2 = 0.0034159''$$
$$u_3 = 0.0032689''$$
$$u_4 = 0.0031586''$$
$$u_5 = 0.0030769''$$

b) $$\left.\frac{du}{dr}\right|_{r=a} \approx \frac{-3u_0 + 4u_1 - u_2}{2(\Delta r)}$$

$$= \frac{-3 \times 0.0038731 + 4 \times 0.0036115 - 0.0034159}{2(0.6)}$$

$$= -4.925 \times 10^{-4}$$

$$\sigma_{\max} = \frac{30 \times 10^6}{1 - 0.3^2}\left(\frac{0.0038731}{5} + 0.3\left(-4.925 \times 10^{-4}\right)\right)$$

$$= 2.0666 \times 10^4 \text{ psi}$$

Therefore, the factor of safety FS is

$$FS = \frac{36 \times 10^3}{2.0666 \times 10^4}$$

$$= 1.7420$$

c) The true error in calculating the maximum stress is

$$E_t = 2.0538 \times 10^4 - 2.0666 \times 10^4$$

$$= -128 \text{ psi}$$

The relative true error in calculating the maximum stress is

$$|\in_t| = \left|\frac{-128}{2.0538 \times 10^4}\right| \times 100$$

$$= 0.62323\%$$

Table 1 Comparisons of radial displacements from two methods.

r	u_{exact}	$u_{\text{1st order}}$	$\left\|\in_t\right\|$	$u_{\text{2nd order}}$	$\left\|\in_t\right\|$
5	0.0038731	0.0038731	0.0000	0.0038731	0.0000
5.6	0.0036110	0.0036165	1.5160×10^{-1}	0.0036115	1.4540×10^{-2}
6.2	0.0034152	0.0034222	2.0260×10^{-1}	0.0034159	1.8765×10^{-2}
6.8	0.0032683	0.0032743	1.8157×10^{-1}	0.0032689	1.6334×10^{-2}
7.4	0.0031583	0.0031618	1.0903×10^{-1}	0.0031586	9.5665×10^{-3}
8	0.0030769	0.0030769	0.0000	0.0030769	0.0000

Reference

ORDINARY DIFFERENTIAL EQUATIONS	
Topic	Finite Difference Methods of Solving Ordinary Differential Equations
Summary	Textbook notes of Finite Difference Methods of solving ordinary differential equations
Major	General Engineering
Authors	Autar Kaw, Cuong Nguyen, Luke Snyder
Date	April 25, 2022