Course Code: **CSC 251**
Course Title: **Numerical Analysis I**
Course Unit: **2**
Course Duration: **2**

COURSE DETAILS

Course Lecturer: A. J. Ikuomola
Department of Computer Science
College of Natural Science
University of Agriculture Abeokuta,
Ogun State, Nigeria

Emails: ikuomolaaj@unaab.edu.ng
Office Location: AMREC Building

COURSE CONTENT

Approximations, Significant Figures, Errors: Truncation & Round-Off Error, Recursive Computation (e.g. Horner's method and Synthetic division for polynomials), Polynomials and their zeroes (for at most degree 4), Bisection rule, Newton-Raphson rule, Computations of functions and series.

COURSE REQUIREMENTS

This is a compulsory course for students in the Department of Computer Science. In view of this, students are expected to participate in all the activities and have a minimum of 75% attendance to be able to write the final examination.

READING LIST

Bello, H.K. (2009). Introduction to Numerical Computations. Ibadan: Agunbay Publishers

Burden, R. L.; Faires, J. D. (1985), "2.1 The Bisection Algorithm", *Numerical Analysis* (3rd ed.), PWS Publishers, ISBN 0-87150-857-5.

Chartier, T. (2005). Devastating Round off Error." *Math. Horizons* **13**, No. 4, 11.

Corliss, George (1977), "Which root does the bisection algorithm find?", *SIAM Review* **19** (2): 325–327, doi:10.1137/1019044, ISSN 1095-7200.

Jain M.K., Iyengar, S.R.K., Jain, R.K. (2007). Numerical Methods for Scientific and Engineering Computation. New Delhi: New Age International Publishers

Kaw, Autar; Kalu, Egwu (2008), *Numerical Methods with Applications* (1st ed.). http://numericalmethods.eng.usf.edu/topics/textbook_index.html

Scheid F. (1988). Schaum's Outlines Numerical Analysis, Second Edition, McGraw Hill Co. Publisher.

Wilkinson, J. H. (1994). *Rounding Errors in Algebraic Processes.* New York: Dover

Wikipedia (2011). Round off Error

UNIVERSITY OF AGRICULTURE ABEOKUTA

READING LIST:

UNIT 1

NUMERICAL ANALYSIS

Numerical Analysis is the study of algorithms for the problems of continuous mathematics (as distinguished from discrete mathematics).

It is concerned with the mathematical derivation, description and analysis of methods of obtaining numerical solution of mathematical problems. It is an area of mathematics and computer science that creates, analyzes and implements algorithms for obtaining numerical solutions to problems involving continuous variables.

Numerical Method

Numerical method is a set of rules for solving a problem or problems of a particular type, involving only the operations of arithmetic.

COMPUTER ARITHMETIC

The decimal number system has the base 10. The decimal integer number 4987 actually means $(4987)_{10} = 4 \times 10^3 + 9 \times 10^2 + 8 \times 10^1 + 7 \times 10^0$ ---------- (1)

which represents a polynomial in the base 10. Similarly, a fractional decimal number 0.6251 means $(0.6251)_{10} = 6 \times 10^{-1} + 2 \times 10^{-2} + 5 \times 10^{-3} + 1 \times 10^{-4}$ -------- (2)

which is a polynomial in 10^{-1}.

Combining (1) and (2), we may write the number 4987.6251 in decimal system as:

$$4987.6251 = 4 \times 10^3 + 9 \times 10^2 + 8 \times 10^1 + 7 \times 10^0 + 6 \times 10^{-1} + 2 \times 10^{-2} + 5 \times 10^{-3} + 1 \times 10^{-4} \quad (3)$$

Binary Number System

Binary number system has base 2 with digits 0 and 1 called bits.

Example 1: Find the decimal number corresponding to the binary number $(111.011)_2$

$$111.011_2 = 1 \times 2^2 + 1 \times 2^1 + 1 \times 2^0 + 0 \times 2^{-1} + 1 \times 2^{-2} + 1 \times 2^{-3}$$

$$= 7.375_{10}$$

Example 2: Convert 58_{10}

2	58
2	29 r 0
2	14 r 1
2	7 r 0
2	3 r 1
2	1 r 1
2	0 r 1

$58_{10} = 111010_2$

Example 3: Convert 0.859375_{10} to the corresponding binary fraction.

0	0.859375
	x 2
1	0.718750
	x 2
1	0.437500
	x 2
0	0.875000
	x 2
1	0.750000
	x 2
1	0.500000
	x 2
1	0000000

The required binary fraction becomes; $0.859375_{10} = 0.110111_2$

Example 4: Convert 0.7_{10} to the corresponding binary fraction:

```
     0.7
   x   2
1    0.4
   x   2
0    0.8
   x   2
1    0.6
     x 2
1    0.2
     x 2
0    0.4
    x  2
0    0.8
    x  2
1    0.6
     x 2
1    0.2
     x 2
0    0.4
```

Thus we obtain $(0.7)_{10} = (.101100110...)_2$ which is a never ending sequence. If only 7 bits are retained in the binary fraction then the corresponding decimal number becomes

$$0.1011001_2 = 1 \times 2^{-1} + 0 \times 2^{-2} + 1 \times 2^{-3} + 1 \times 2^{-4} + 0 \times 2^{-5} + 0 \times 2^{-6} + 1 \times 2^{-7}$$

$$= 0.6953125$$

which is not exactly the same as the given number.

The difference

$$0.7 - 0.6953125 = 0.0046875$$

is the round-off error.

Octal System

The Octal system has base 8 and uses the digits 0, 1, 2, 3, 4, 5, 6, 7. A binary number can be converted to an Octal number by grouping the bits in groups of three to the right and left of the binary point by adding sufficient zeros to complete the groups and replacing each group of three bits by its Octal equivalent.

Example: Convert the binary number 1101001.1110011 to the octal system. We have

001| 101 | 001 . 111 | 001 | 100

1 5 1 . 7 1 4

$1101001.1110011 = 151.714_8$

Hexadecimal System

The hexadecimal system has base 16 and the digits 0 to 9 and A, B, C, D, E, F to represent 10, 11, 12, 13, 14, 15 respectively. To convert a binary number to a hexadecimal number, we form groups of four of the binary bits and replace it by the corresponding digit in the hexadecimal system.

Example: Convert the binary number 1101001.1110011 to the hexadecimal system.

0110	1001	1110	0110
6	9	E	6

$1101001.1110011 = 69E6_{16}$

UNIT 2

FLOATING POINT ARITHMETIC

In computing, **floating point** describes a system for representing real numbers which supports a wide range of values. Numbers are in general represented approximately to a fixed number of significant digits and scaled using an exponent. The base for the scaling is normally 2, 10 or 16. The typical number that can be represented exactly is of the form:

Significant digits × *base*^{exponent}

Definition 1: A floating point number is a number represented in the form

$$. d_1 d_2 ....d_t \text{ X } \beta^e \qquad \text{----------------} \qquad (1)$$

where $d_1, d_2, .., d_t$ are integers and satisfy $0 \leq d_i < \beta$ and the exponent e is such that $m \leq e \leq M$.

The fractional part $d_1 d_2 ....d_t$ is called the mantissa and it has between +1 and -1. The number 0 is written as

$$+ 0.0000 \text{ X } \beta^e$$

Definition 2: A non–zero floating point number as defined in (1) is in normal form if the value of the mantissa lies in the interval (-1, 1/ β] or in the interval [1/ β ,1).

Example:

Subtract the floating point number 0.36143447×10^7 and 0.36132346×10^7 .

Solution

$$\begin{array}{r} 0.36143447 \times 10^7 \\ \underline{-0.36132346 \times 10^7} \\ \underline{0.00011101 \times 10^7} \end{array}$$

The result is a floating point number, but not a normalized floating point number due to the presence of three leading zeros. Shifting the fractional part three places to the left, we get result $0.11101 \text{ X } 10^4$ which is normalized floating point number.

Definition 3:

A non-zero floating point number as defined in (1) is in t-digit– mantissa standard form if it is normalized and its mantissa consists of exactly t-digits. If a number x has the representation in the form

$$x = . d_1 d_2 ... d_t \, d_{t+1} ... \text{ x } \beta^e \qquad \text{---------------} \qquad (2).$$

then the floating point number fl(x) in t- digit mantissa standard form can be obtained in the following two ways:

(i) Chopping: We neglect d_{t+1}, d_{t+2} .... in (2) and obtain

$$fl(x) = .\,d_1 d_2 \ldots d_t \times \beta^e \qquad \text{--------------} \qquad (3).$$

(ii) Rounding: The fractional part in (2) is written as

$$.\,d_1 d_2 \ldots d_t d_{t+1} + \frac{1}{2}\beta \qquad \text{--------------} \qquad (4).$$

and the first t- digit are taken to write the floating point number.

Example :

Find the sum of $.123 \times 10^3$ and $.456 \times 10^2$ and write the result in three digit mantissa form.

Solution

The number of the smaller magnitude is adjusted so that its exponent is the same as that of the number of larger magnitude. We have

$.1230 \times 10^3$

$\underline{.0456 \times 10^3}$

$\underline{.1686 \times 10^3} = \left| \begin{array}{l} .\,168 \times 10^3 \text{ , for chopping} \\ .\,169 \times 10^3 \text{ , for rounding} \end{array} \right.$

Exercise

Evaluate : $f(x) = x3 - (6.1)x2 + (3.2)x-(1.5)$

Find the value of $f(x)$ at x=4.72 using 3 digit arithmetic

	x	x^2	x^3	6.1 x^2	3.2x
Exact	4.72	22.2784	105.154048	135.89824	15.104
Chopping	4.72	22.2	105	135	15.1
Rounding	4.72	22.3	105	136	15.1

Chopping → $f(x) = ?$

Rounding $\rightarrow f(x) = ?$

Range of floating-point numbers

By allowing the radix point to be adjustable, floating-point notation allows calculations over a wide range of magnitudes, using a fixed number of digits, while maintaining good precision. For example, in a decimal floating-point system with three digits, the multiplication that human would write as

$$0.12 \times 0.12 = 0.0144$$

would be expressed as

$$(1.2\times10^{-1}) \times (1.2\times10^{-1}) = (1.44\times10^{-2}).$$

In a fixed-point system with the decimal point at the left, it would be

$$0.120 \times 0.120 = 0.014.$$

A digit of the result was lost because of the inability of the digits and decimal point to 'float' relative to each other within the digit string.

Floating-point arithmetic operations

For ease of presentation and understanding, decimal radix with 7 digit precision will be used in the examples, as in the IEEE 754 *decimal32* format. The fundamental principles are the same in any radix or precision, except that normalization is optional (it does not affect the numerical value of the result). Here, *s* denotes the significant and *e* denotes the exponent.

Addition and subtraction

A simple method to add floating-point numbers is to first represent them with the same exponent. In the example below, the second number is shifted right by three digits and we then proceed with the usual addition method:

$123456.7 = 1.234567 \times 10^5$

$101.7654 = 1.017654 \times 10^2 = 0.001017654 \times 10^5$

Hence:

$$\begin{aligned}123456.7 + 101.7654 &= (1.234567 \times 10^5) + (1.017654 \times 10^2)\\ &= (1.234567 \times 10^5) + (0.001017654 \times 10^5)\\ &= (1.234567 + 0.001017654) \times 10^5\\ &= 1.235584654 \times 10^5\end{aligned}$$

In detail:

```
  e=5;  s=1.234567     (123456.7)
+ e=2;  s=1.017654     (101.7654)

  e=5;  s=1.234567
```

```
+ e=5;  s=0.001017654  (after shifting)
   ---------------------------
  e=5;  s=1.235584654  (true sum: 123558.4654)
```

This is the true result, the exact sum of the operands. It will be rounded to seven digits and then normalised if necessary. The final result is

```
  e=5;  s=1.235585    (final sum: 123558.5)
```

Note that the low 3 digits of the second operand (654) are essentially lost. This is round-off error. In extreme cases, the sum of two non-zero numbers may be equal to one of them:

```
   e =  5;   s=1.234567
+ e= −3;  s=9.876543

   e=5;  s=1.234567
+ e=5;  s=0.00000009876543 (after shifting)
----------------------
  e=5;  s=1.23456709876543 (true sum)
  e=5;  s=1.234567         (after rounding/normalization)
```

Another problem of loss of significance occurs when two close numbers are subtracted. In the following example $e = 5$; $s = 1.234571$ and $e = 5$; $s = 1.234567$ are representations of the rationals 123457.1467 and 123456.659.

```
   e=5;  s=1.234571
− e=5;  s=1.234567
-------------------------
   e=5;  s=0.000004

e=−1; s=4.000000 (after rounding/normalization)
```

The best representation of this difference is $e = -1$; $s = 4.877000$, which differs more than 20% from $e = -1$; $s = 4.000000$. In extreme cases, the final result may be zero even though an exact calculation may be several million.

Multiplication and division

To multiply, the significant are multiplied while the exponents are added and the result is rounded and normalized.

```
   e=3;  s = 4.734612
× e=5;  s =5.417242
-----------------------
  e=8;  s=25.648538980104 (true product)
  e=8;  s=25.64854       (after rounding)
  e=9;  s=2.564854       (after normalization)
```

Division is done similarly, but is more complicated

While floating-point addition and multiplication are both commutative ($a + b = b + a$ and $a \times b = b \times a$), they are not necessarily associative. That is, $(a + b) + c$ is not necessarily equal to $a + (b + c)$. Using 7-digit decimal arithmetic:

a = 1234.567, b = 45.67834, c = 0.0004

(a + b) + c:

```
   1234.567     (a)
 +   45.67834 (b)
 ______________
   1280.24534   rounds to   1280.245

  1280.245  (a + b)
 +    0.0004 (c)
  1280.2454   rounds to   1280.245  <--- (a + b) + c
```

a + (b + c):

```
  45.67834 (b)
+  0.0004  (c)
  45.67874

     45.67874 (b + c)
+ 1234.567     (a)
  1280.24574   rounds to   1280.246 <--- a + (b + c)
```

They are also not necessarily distributive. That is, $(a + b) \times c$ may not be the same as $a \times c + b \times c$:

1234.567 × 3.333333 = 4115.223
1.234567 × 3.333333 = 4.115223
4115.223 + 4.115223 = 4119.338
but
1234.567 + 1.234567 = 1235.802
1235.802 × 3.333333 = 4119.340

In addition to loss of significance, inability to represent numbers such as π and 0.1 exactly

UNIT 3

ERROR ANALYSIS

Errors come in a variety of ways; some are avoidable while some are not. How errors occur and how they affect the accuracy of calculation is very essential in understanding numerical methods. In applied mathematics, error is the difference between a true value and an estimate or approximation of that value.

E = True value – Approximate value

Source of Error

- Modeling Error: a wrong or inappropriate choice of model
- Measurement Error: incorrect or poor measurements
- Implementation Error: incorrect or poor choice of algorithms
- Simulation Error: error accumulated due to the execution of our model

Example

The value of π is 3.14159265….

The commonly used approximation to π is 22/7, what is the error in this approximation?

Solution

We must convert 22/7 to decimal form and find the difference.

True value of π is 3.14159265….

Approximated value of π is 3.14285714

Error = 3.14159265 - 3.14285714

= 0.00126449

Generally speaking, error can be introduced into numerical work by the following

- Mistake due to human error
- Error due to given data
- Round off (premature approximation)
- Error due to method employed

Types of Error

(1) **Inherent error**: is that quantity which is already present in the statement of the problem before it solution. The inherent error arises either due to the simplified assumption in the mathematical formulation of the problem or due to the errors in the physical measurements of the parameters of the problem.

(2) **Round off Error:** Rounding means to round by raising the last figure by 1 if the next figure would have been greater or less than 5.

Round-off Error is the quantity R which must be added to the finite representation of a computed number in order to make it the true representation of that number. It is due to representation of a number by a finite number of decimal digits e.g. approximation due to nearest whole number and also approximation to a certain decimal places.

A **round-off error**, also called **rounding error**, is the difference between the calculated approximation of a number and its exact mathematical value. Roundoff error is the difference between an approximation of a number used in computation and its exact (correct) value. Numerical analysis specifically tries to estimate this error when using approximation equations and/or algorithms, especially when using finitely many digits to represent real numbers (which in theory have infinitely many digits). This is a form of quantization error.
When a sequence of calculations subject to rounding error are made, errors may accumulate in certain cases known as ill-conditioned, sometimes to such an extent as to dominate the calculation and make the result meaningless.

Example
The approximation of 9.345 to the nearest whole number is 9
The approximation of 9.345 to 2 decimal points is 9.35
2/3 = 0.6666 rounded to three decimals places is 0.667

Round (17.5) = 18

Representation error

The error introduced by attempting to represent a number on the computer is called *representation error*. Some examples:

Notation	**Represent**	**Approximate**	**Error**
1/7	0.142 857	0.142 857	0.000 000 142 857
ln 2	0.693 147 180 559 945 309 41...	0.693 147	0.000 000 180 559 945 309 41...
$\log_{10} 2$	0.301 029 995 663 981 195 21...	0.3010	0.000 029 995 663 981 195 21...
$\sqrt[3]{2}$	1.259 921 049 894 873 164 76...	1.25992	0.000 001 049 894 873 164 76...
$\sqrt{2}$	1.414 213 562 373 095 048 80...	1.41421	0.000 003 562 373 095 048 80...
e	2.718 281 828 459 045 235 36...	2.718 281 828 459 045	0.000 000 000 000 000 235 36…
π	3.141 592 653 589 793 238 46...	3.141 592 653 589 793	0.000 000 000 000 000 238 46...

(3) **Truncation Error:** This is the quantity T which must be added to the true representation of the quantity in order for the result to be exactly equal to the quantity we are seeking to generate.

Truncate means to cut off and truncation error happen when a fraction is cut off a certain number of decimal or binary places.

In mathematics and computer science, **truncation** is the term for limiting the number of digits right of the decimal point, by discarding the least significant ones.
For example, consider the real numbers

5.6341432543653654
32.438191288
−6.3444444444444

To *truncate* these numbers to 4 decimal digits, we only consider the 4 digits to the right of the decimal point.

The result would be:

5.6341
32.4381
−6.3444

Note that in some cases, truncating would yield the same result as rounding, but truncation does not round up or round down the digits; it merely cuts off at the specified digit. The truncation error can be twice the maximum error in rounding.

Whenever a finite, y = f(x) is represented by an infinite series. Truncation error is defined on the error caused by truncating a mathematical procedure. The error in the value of f(x) due to deleting of the series after a finite number of terms is called truncation Error.

Example 2:

$Y = f(x) = x + x^2 + x^3 + x^4\ x^5 + \ldots + x^n + x^{n+1}$

The magnitude of truncation error equal the sum of all the discarded term represented by $x^4\ x^5 + \ldots + x^n + x^{n+1} + \ldots$ may be large and it may even exceed the sum of the term required.

Example 3:

The Maclaurin series for e^x is given as

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

The series has 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

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

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

Causes of Truncation

With computers, truncation can occur when a decimal number is typecast as an integer; it is truncated to zero decimal digits because integers cannot store real numbers (that are not themselves integers). Truncation may also occur when a number cannot be fully represented due to memory limitations.

Trunction errors in numerical integration are of two kinds:

- *local truncation errors* – the error caused by one iteration, and
- *global truncation errors* – the cumulative error cause by many iterations.

Local truncation error

The **local truncation error** is the error that the increment function, *A*, causes during a given iteration, assuming perfect knowledge of the true solution at the previous iteration.

More formally, the local truncation error, τ_n, at step n is defined by:

$$\tau_n = |y(t_n) - y(t_{n-1}) - h \cdot A(t_{n-1}, y(t_{n-1}), h, f)|$$

Global truncation error

The **global truncation error** is the accumulation of the *local truncation error* over all of the iterations, assuming perfect knowledge of the true solution at the initial time step.

More formally, the global truncation error, e_n, after N steps is defined by:

$$e_N = |y(t_N) - y_N|$$
$$= |y(t_N) - (y_0 + h \cdot A(t_0, y_0, h, f) + h \cdot A(t_1, y_1, h, f) + \cdots + h \cdot A(t_{N-1}, y_{N-1}, h, f))|$$

Techniques for Measuring Error

The quantity, true value – Approximation value is called the error.

In order to determine the accuracy in an approximate solution to a problem, either we find the bound of the

$$\text{Relative Error} = \frac{|Error|}{|True\ Value|} = \frac{|True\ Value - Approximation\ value|}{|True\ value|}$$

or of the

Absolute error = |Error|

Example:

Define the error of an approximation. The traditional definition is

True value = approximation + error

e.g. $\sqrt{2}$ = 1.414214 + Error

1.414213562373095 = 1.414214 + error

Error = 1.414213562373095 - 1.414214 = $-4.376269049512545e^{-7}$

π = 3.1415926536 + Error

Error = True value - approximation

e.g if 36.75 is the exact value of a number and if 37 is the approximated value then the error introduced is

|e | = 36.75 – 37 = |-0.25|

e = 0.25

Relative Error

This is error measure relative to the true value.

$$\text{Relative Error} = \frac{|Error|}{|True\ Value|} = \frac{|True\ Value - Approximation\ value|}{|True\ value|}$$

Example:

A resistor labeled as 240Ω is actually 243.32753Ω. What are the absolute and relative errors of the labeled value?

Solution

Absolute error = |True value – Approximation value|

= 243.32753 – 240

= 3.32753

$$\text{Relative Error} = \frac{|Error|}{|True\,Value|} = \frac{3.32753}{243.32753} = 0.0136751069638524 \approx 0.014$$

UNIT 4

SYNTHETIC DIVISION

Definition 1:

Synthetic division is a method of division in which you perform division on the coefficients, removing the variables and exponents. It allows you to add throughout the process rather than subtract (long division).

How to divide Polynomials using Synthetic Division

1. **For the purposes of this work**

   $(x^3 + 2x^2 - 4x + 8) \div (x + 2)$
   is the example for all steps.

2. **Reverse the sign of the constant in the divisor**.

   $(x + 2)$ is the divisor. The 2 (two) becomes a negative.

3. **Place this new number by itself and place a "backwards L" on its right side**.

   -2|

4. **To the right of this, write all of the coefficients (in standard form)**.

   -2| 1 2 -4 8

5. **Bring down the first coefficient**.

-2|1 2 -4 8

↓

1

6. **Multiply this by the new divisor and place it under the second coefficient**.

```
-2|1  2  -4  8
    -2
 1
```

7. **Combine the second coefficient and the product**.

```
-2|1  2  -4  8
      -2
 1   0
```

8. **Multiply this sum by the new divisor and place under the third coefficient**.

```
-2|1  2  -4  8
      -2   0
 1   0
```

9. **Combine these**.

```
-2|1  2  -4  8
     -2   0
 1  0  -4
```

10. **Continue in the same fashion until you have found the final sum**. This sum is the remainder.

```
-2|1  2  -4  8
    -2   0   8
 1 0  -4 |16
```

To write the answer, place each of the sums next to a variable of one less power than the original it is lined up with. In this case, the first sum is placed next to an *x* to the second power (one less than three), the second sum is zero, so it isn't part of the answer, and the negative four is not next to an *x*.

```
-2|1  2  -4  8
      -2  0 8
 1   0 -4 16
```

$x^2 + 0x - 4$R16

$x^2 - 4$ R16

If your remainder is 0, the original divisor was a factor of the polynomial.

- To check your answer, multiply the quotient by the divisor and add the remainder. It should be the same as the original polynomial.

  (divisor)(quotient)+(remainder)
  $(x + 2)(x^2 - 4) + 16$

  Using FOIL method, multiply.
  $(x^3 - 4x + 2x^2 - 8) + 16$
  $x^3 + 2x^2 - 4x - 8 + 16$
  $x^3 + 2x^2 - 4x + 8$

Definition 2:

Synthetic division is a shorthand method of dividing a polynomial by a binomial of the form x - a. For example, if $3x^4 + 2x^3 + 2x^2 - x - 6$ is to be divided by x - 1, the long form would be as follows:

```
              3x³ + 5x² + 7x + 6
x - 1 | 3x⁴ + 2x³ + 2x² - x - 6
        3x⁴ - 3x³
        ---------
            + 5x³ + 2x²
            + 5x³ - 5x²
            -----------
                  + 7x² - x
                  + 7x² - 7x
                  ----------
                        + 6x - 6
                        + 6x - 6
                        ========
```

Notice that every alternate line of work in this example contains a term which duplicates the one above it. Furthermore, when the subtraction is completed in each step, these duplicated terms cancel each other and thus have no effect on the final result. Another unnecessary duplication results when terms from the dividend are brought down and rewritten prior to subtraction. By omitting these duplications, the work may be condensed as follows:

	$3x^3$	$+5x^2$	$+7x$	$+6$	
$x - 1$	$3x^4$	$+2x^3$	$+2x^2$	$-x$	-6
		$-3x^3$	$-5x^2$	$-7x$	-6
		$+5x^3$	$+7x^2$	$+6x$	0

The coefficients of the dividend and the constant term of the divisor determine the results of each successive step of multiplication and subtraction. Therefore, we may condense still further by writing only the nonliteral factors, as follows:

	3	+5	+7	+6	
- 1	3	+2	+2	-1	-6
		-3	-5	-7	-6
	3	+5	+7	+6	0

Notice that if the coefficient of the first term in the dividend is brought down to the last line, then the numbers in the last line are the same as the coefficients of the terms in the quotient. Thus we do not really need to write a separate line of coefficients to represent the quotient. Instead, we bring down the first coefficient of the dividend and make the subtraction "subtotals" serve as coefficients for the rest of the quotient, as follows:

x - 1	3	2	2	-1	-6
		-3	-5	-7	-6
	3	5	7	6	0

The unnecessary writing of plus signs is also eliminated here.

The use of synthetic division is limited to divisors of the form x - a, in which the degree of x is 1. Thus the degree of each term in the quotient is 1 less than the degree of the corresponding term in the dividend. The quotient in this example is as follows:

$3x^3 + 5x^2 + 7x + 6$

The sequence of operations in synthetic division may be summarized as follows, using as an example the division of $3x - 4x^2 + x^4 - 3$ by $x + 2$:

2	1	0	-4	3	-3
		2	-4	0	6
	1	-2	0	3	-9

First, rearrange the terms of the dividend in descending powers of x. The dividend then becomes $x^4 - 4x^2 + 3x - 3$, with 1 understood as the coefficient of the first term. No x^3 term appears in the polynomial, but we supply a zero as a place holder for the x^3 position.

Second, bring down the 1 and multiply it by the +2 of the divisor. Place the result under the zero, and subtract. Multiply the result (-2) by the +2 of the divisor, place the product under the -4 of the dividend, and subtract. Continue this process, finally obtaining $x^3 - 2x^2 + 3$ as the quotient. The remainder is -9.

Practice problems. In the following problems, perform the indicated operations. In 4, 5, and 6, first use synthetic division and then check your work by long division:

1. $(a^3 - 3a^2 + a) \div a$

2. $\frac{x^6 - 7x^5 + 4x^4}{x^2}$
3. $(10x^3 - 7x^2y - 16xy^2 + 12y^3) \div (2x^2 + xy - 2y2)$
4. $(x^2 + 11x + 30) \div (x + 6)$
5. $(12 + x^2 - 7x) \div (x - 3)$
6. $(a^2 - 11a + 30) \div (a - 5)$

Answers:

1. $a^3 - 3a + 1$
2. $x^4 - 7x^3 + 4x^2$
3. $5x - 6y$
4. $x + 5$
5. $x - 4$
6. $a - 6$

UNIT 5

HORNER SCHEME

In numerical analysis, the Horner scheme (also known as Horner algorithm), named after William George Horner, is an algorithm for the efficient evaluation of polynomials in monomial form. Horner's method describes a manual process by which one may approximate the roots of a polynomial equation. The Horner scheme can also be viewed as a fast algorithm for dividing a polynomial by a linear polynomial with Ruffini's rule.

Using Horner to evaluate Polynomial functions

In a computer program it is sometimes necessary to evaluate polynomial functions. The basic method to evaluate the polynomial function is to "plug in" the value of x into the polynomial. Horner's method however results in fewer multiplications and additions and is faster and more precise when using float variables.

Univariate Polynomial function definition

A **univariate** polynomial function has the following form :

$$f(x) = \sum_{n=0}^{k} a_n x^n$$

For example if k = 4 the **order** of the polynomial is 4 and the function has the following form :

$$f(x) = a_4 x^4 + a_3 x^3 + a_2 x^2 + a_1 x + a_0$$

Description of the algorithm

Given the polynomial

$$p(x) = \sum_{i=0}^{n} a_i x^i = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots + a_n x^n,$$

where $a_0, \ldots, a_n$ are real numbers, we will evaluate the polynomial at a specific value of x, say x_0.

To accomplish this, we define a new sequence of constants as follows:

$$
\begin{aligned}
b_n &:= a_n \\
b_{n-1} &:= a_{n-1} + b_n x_0 \\
&\ \ \vdots \\
b_0 &:= a_0 + b_1 x_0.
\end{aligned}
$$

Then b_0 is the value of $p(x_0)$.

To see why this works, note that the polynomial can be written in the form

$$p(x) = a_0 + x(a_1 + x(a_2 + \cdots + x(a_{n-1} + a_n x)\cdots)).$$

Thus, by iteratively substituting the b_i into the expression,

$$
\begin{aligned}
p(x_0) &= a_0 + x_0(a_1 + x_0(a_2 + \cdots x_0(a_{n-1} + b_n x_0)\cdots)) \\
&= a_0 + x_0(a_1 + x_0(a_2 + \cdots x_0(b_{n-1})\ldots)) \\
&\ \ \vdots \\
&= a_0 + x_0(b_1) \\
&= b_0.
\end{aligned}
$$

Horner

Horner's method is commonly used to find the roots of a polynomial function. However it can also be used to evaluate the polynomial function for a given value of **x**.

The main goal of the Horner scheme is to reduce the number of multiplications needed by isolating the variable (in this case x).

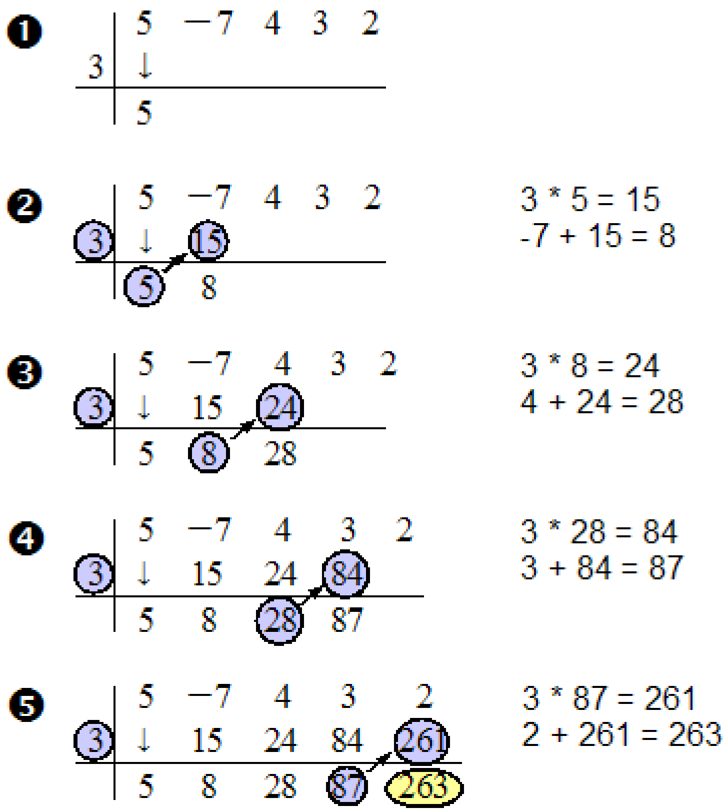
A concrete example :

$$
\begin{aligned}
f(x) &= (a_4 x^3 + a_3 x^2 + a_2 x \ + a_1) * x + a_o \\
f(x) &= (\ (a_4 x^2 + a_3 x^1 + a_2) * x + a_1 \ ) * x + a_0 \\
f(x) &= ((\ (a_4 x \ + a_3 \ ) * x + a_2 \ ) * x + a_1) * x + a_o
\end{aligned}
$$

The main advantage here is that when using Horner's method it is not necessary to use a power function to evaluate the variable **x**.

Examples

1. An example of the evaluation of a polynomial function with pen and paper is given below. The polynomial is $f(x) = 5x^4 - 7x^3 + 4x^2 + 3x + 2$ and the function is evaluated for **x = 3**.



Finally, the right bottom number is the result of evaluation of the polynomial function for x = 3, which is **263**.

2. Evaluate

$$f(x) = 2x^3 - 6x^2 + 2x - 1 \text{ for } x - 3.$$

We use synthetic division as follows:

x_0	x^3	x^2	x^1	x^0
3	2	-6	2	-1
		6	0	6
	2	0	2	5

The entries in the third row are the sum of those in the first two. Each entry in the second row is the product of the x-value (3 in this example) with the third-row entry immediately to the left. The entries in the first row are the coefficients of the polynomial to be evaluated. Then the remainder of $f(x)$ on division by $x - 3$ is 5.

But by the remainder theorem, we know that the remainder is $f(3)$. Thus $f(3) = 5$

In this example, if $a_3 = 2, a_2 = -6, a_1 = 2, a_0 = -1$ we can see that $b_3 = 2, b_2 = 0, b_1 = 2, b_0 = 5$, the entries in the third row. So, synthetic division is based on Horner Scheme.

As a consequence of the polynomial remainder theorem, the entries in the third row are the coefficients of the second-degree polynomial, the quotient of $f(x)$ on division by $x - 3$. The remainder is 5. This makes Horner's method useful for polynomial long division.

3. Divide $x^3 - 6x^2 + 11x - 6$ by $x - 2$:

2	1	-6	11	-6
		2	-8	6
	1	-4	3	0

The quotient is $x^2 - 4x + 3$.

4. Let $f_1(x) = 4x^4 - 6x^3 + 3x - 5$ and $f_2(x) = 2x - 1$. Divide $f_1(x)$ by $f_2(x)$ using Horner's scheme.

1	4	-6	0	3	-5
2		2	-2	-1	1
	2	-2	-1	1	-4

The third row is the sum of the first two rows, divided by 2. Each entry in the second row is the product of 1 with the third-row entry to the left. The answer is

$$\frac{f_1(x)}{f_2(x)} = 2x^3 - 2x^2 - x + 1 - \frac{4}{(2x-1)}.$$

Java implementation

The Java implementation is very straight forward :

```
package javaimpl;

public class Main {
    public static void main(String[] args) {
        // 2 + 3*x + 4*x^2 - 7*x^3 + 5*x^4
        Polynomial p = new Polynomial(2,3,4,-7,5);
        float result = p.evaluate(3);
        System.out.println(result);
    }
}

class Polynomial{
    float[] a;
    int order;

    Polynomial( float ... coefficients ){
        a = coefficients;
        order = coefficients.length - 1;
    }

    float evaluate(float x){
        float result = a[order];
        for (int i = order - 1 ; i >= 0 ; --i ){
            result = result * x + a[i];
        }
        return result;
    }
}
```

C++ Implementation

```
#include <iostream>
#include <vector>
using namespace std;

class Polynomial{
public:
    Polynomial(const vector<float> & coefficients)
    : m_a(coefficients.begin(),coefficients.end())
    {
        m_Order = m_a.size() - 1;
    }
```

```
    float evaluate( float x) {
        float result = m_a[m_Order];
        for ( int i = m_Order -1 ; i >= 0 ; --i ){
            result = result * x + m_a[i];
        }
        return result;
    }
private:
    vector<float> m_a;
    int m_Order;
};

int main()
{
    vector<float> cs;
    cs.push_back(2);
    cs.push_back(3);
    cs.push_back(4);
    cs.push_back(-7);
    cs.push_back(5);
    Polynomial p (cs);
    float result  = p.evaluate(3);
    cout << "Result is : " << result << endl;
}
```

C# implementation

```
using System;

namespace csharpimpl
{
    class Polynomial
    {
        private float[] a;
        private int order;

        public Polynomial ( params float[] coefficients){
            a = coefficients;
            order = coefficients.Length - 1;
        }

        public float Evaluate ( float x)
        {
            float result = a[order];
            for (int i = order - 1; i >= 0; --i)
```

```
            {
                result = result * x + a[i];
            }
            return result;
        }
    }
    class Program
    {
        static void Main(string[] args)
        {
            Polynomial p = new Polynomial( 2, 3, 4, -7, 5 );
            float result = p.Evaluate(3);
            Console.WriteLine("The result is : " + result);
        }
    }
}
```

UNIT 6

POLYNOMIALS AND THEIR ZEROES (FOR AT MOST DEGREE 4)

Imagine a general equation of the form

$$P(x) = \sum_{i=0}^{n} a_i x^i$$

$$= a_0 + a_1 x + a_2 x^2 + \ldots.. + a_n x^n$$

where all the a(s) are (constant) arbitrary real numbers.

The above equation is referred to as Polynomial equation of degree n. The highest power of a variable is called the degree of that particular variable.

Names of polynomials by degree
The following names are assigned to polynomials according to their degree:[1]

- Degree 0 – constant
- Degree 1 – linear
- Degree 2 – quadratic
- Degree 3 – cubic
- Degree 4 – quartic (or, less commonly, biquadratic) (or, a little more common, Fourth degree)
- Degree 5 – quintic
- Degree 6 – sextic (or, less commonly, hexic)
- Degree 7 – septic (or, less commonly, heptic)
- Degree 8 – octic
- Degree 9 – nonic
- Degree 10 – decic
- Degree 100 - hectic

The degree of the zero polynomial is either left explicitly undefined, or is defined to be negative (usually −1 or −∞).

To be able to get the solution to a particular problem, certain assumption must be taken:

(i) P(x) may be algebraic or transcendental

(ii) P(x) may be continuous with an interval of interest (domain)

(iii) P(x) is differentiable once or twice at least or more depending on the system of equation required.

There is a great problem in determining the roots of an equation of the form f(x) = 0.

The function f(x) may be given explicitly, for example

$$f(x) = P_n(x)$$
$$= a_0 x^n + a_1 x^{n-1} + ... + a_{n-1} x + a_{n,} \quad a_0 \neq 0$$

A polynomial of degree n in x or f(x) may be known only implicitly as a transcendental function.

Definition 1: A number ξ is a solution of f(x) = 0 if $f(\xi) \equiv 0.$ Such a solution ξ is a root or a zero of f(x) = 0

Geometrically, a root of the equation f(x) = 0 is the value of x at which the graph of y = f(x) intersects the x-axis.

Definition 2: If we can write f(x) = 0 as

$$f(x) = (x - \xi)^m g(x) = 0$$

Where g(x) is bounded and $g(\xi) \neq 0,$ then ξ is called a multiple root of multiplicity m.

In this case, $f(\xi) = f^{'}(\xi) = ... = f^{(m-1)}(\xi) = 0, \quad f^{(m)}(\xi) \neq 0.$

For m = 1, the number ξ is said to be a simple root.

There are generally two types of methods used to find the roots of the equation f(x) = 0. They are:

Direct Methods: These methods give the exact value of the roots in a finite number of steps. Further, the methods give all the roots at the same time. For example, a direct method gives the root of a linear or first degree equation

$$a_0 x + a_1 = 0, \quad a_0 \neq 0$$

as $x = -a_1 / a_0.$

Similarly, the roots of the quadratic equation

$$a_0 x^2 + a_1 x + a_2 = 0, \quad a_0 \neq 0$$

are given by

$$x = \frac{-a_1 \pm \sqrt{(a_1^2 - 4a_0 a_2)}}{2a_0}$$

http://www.unaab.edu.ng

Iterative Methods: These methods are based on the idea of successive approximations, i.e., starting with one or more initial approximations to the root, we obtain a sequence of approximations or iterates $\{x_k\}$, which in the limit converges to the root. The methods give only one root at a time.

Part 1: The Roots or Zeroes of a Polynomial

1. What is a polynomial equation?

   It is a polynomial set equal to 0. $P(x) = 0$.

   Example. $P(x) = 5x^3 - 4x^2 + 7x - 8 = 0$

2. What do we mean by a root, or zero, of a polynomial?

   It is a *solution* to the polynomial equation, $P(x) = 0$.
   It is that value of x that makes the polynomial equal to 0.

   In other words, the number r is a root of a polynomial $P(x)$ if and only if $P(r) = 0$.

   Example 1. Let $P(x) = 5x^3 - 4x^2 + 7x - 8$. Then a root of that polynomial is 1, because

$$P(1) = 5\cdot 1^3 - 4\cdot 1^2 + 7\cdot 1 - 8$$

$$= 5 - 4 + 7 - 8$$

$$= 0$$

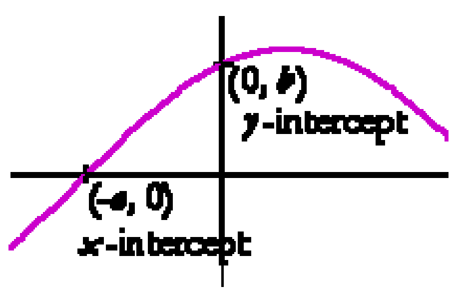
   It is traditional to speak of a root of a polynomial. Of a function in general, we speak of a zero.

   Example 2. The roots of this quadratic

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

   are -2 and 3. Those are the values of x that will make the polynomial equal to 0.

3. What are the x-intercept and y-intercept of a graph?

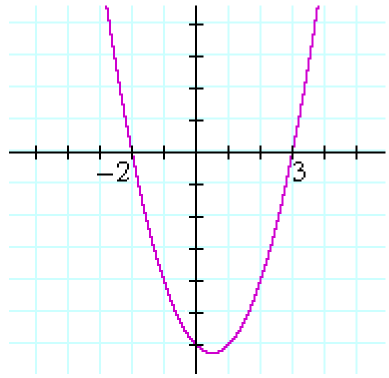


   The x-intercept is that value of x where the graph crosses or touches the x-axis. At the x-intercept -- on the x-axis -- $y = 0$.

The *y*-intercept is that value of *y* where the graph crosses the *y*-axis. At the *y*-intercept, $x = 0$.

4. What is the relationship between the root of a polynomial and the *x*-intercepts of its graph?

The roots *are* the *x*-intercepts!



The roots of $x^2 - x - 6$ are −2 and 3. Therefore, the graph of

$$y = x^2 - x - 6$$

will have the value 0 -- it will cross the *x*-axis -- at −2 and 3.

5. How do we find the *x*-intercepts of the graph of any function $y = f(x)$?

Solve the equation, $f(x) = 0$.

Examples

1. Write the polynomial with integer coefficients that has the following roots: −1, ¾.
Solution. Since −1 is a root, then $(x + 1)$ is a factor. As for the root ¾, we would have the solution

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

which implies

$$4x = 3$$

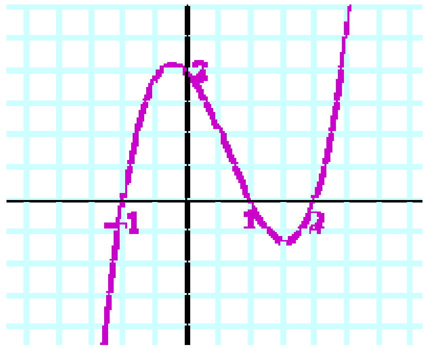
$$4x - 3 = 0$$

The factors are $(4x - 3)(x + 1)$.
The polynomial is $4x^2 + x - 3$.

2. Determine the polynomial whose roots are −1, 1, 2 and sketch its graph.

The factors are $(x + 1)(x - 1)(x - 2)$. On multiplying out, the polynomial is $(x^2 - 1)(x - 2) =$
$x^3 - 2x^2 - x + 2$.

Here is the graph:



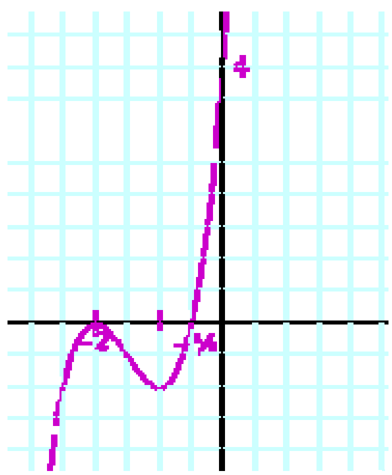
The *y*-intercept is the constant term 2. In every polynomial the *y*-intercept is the constant term, because the constant term is the value of *y* when $x = 0$.

3. Determine the polynomial with integer coefficients whose roots are −½, −2, −2, and sketch the graph.

The factors are $(2x + 1)(x + 2)^2$. On multiplying out, the polynomial is $(2x + 1)(x^2 + 4x + 4)$ =

$$2x^3 + 9x^2 + 12x + 4.$$

Here is the graph:



−2 is a double root. The graph does not cross the *x*-axis.

Question: If r is a root of a polynomial $p(x)$, then upon dividing $p(x)$ by $x - r$, what *remainder* do you expect?

0. Because r being a root will mean that $x - r$ is a *factor* of $p(x)$.

4. Is $x = 2$ a root of this polynomial:

$$x^6 - 3x^5 + 3x^4 - 3x^3 + 3x^2 - 3x + 2\ ?$$

Use synthetic division to divide the polynomial by $x - 2$, and look at the remainder.

1	−3	+3	−3	+3	−3	+2	2
	+2	−2	+2	−2	+2	−2	
1	−1	+1	−1	+1	−1	0	

The remainder is 0. 2 is a root of the polynomial.

5. Find the three roots of

$$P(x) = x^3 - 2x^2 - 9x + 18,$$

given that one root is 3.

Solution. Since 3 is a root of $P(x)$, then according to the factor theorem, $x - 3$ is a factor. Therefore, on dividing $P(x)$ by $x - 3$, we can find the other, quadratic factor.

1	−2	−9	+18	3
	+3	+3	−18	
1	+1	−6		

We have

$$x^3 - 2x^2 - 9x + 18 = (x^2 + x - 6)(x - 3)$$

$$= (x - 2)(x + 3)(x - 3)$$

The three roots are: 2, −3, 3.

Again, since $x - 3$ is a *factor* of $P(x)$, the *remainder* is 0.

6. Sketch the graph of this polynomial,

$$y = x^3 - 2x^2 - 5x + 6,$$

given that one root is −2.

Since −2 is a root, then $(x + 2)$ is a factor. To find the other, quadratic factor, divide the polynomial by $x + 2$. Note that the root −2 goes in the box:

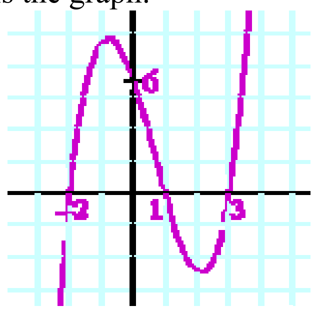
$$\begin{array}{rrrr|l} 1 & -2 & -5 & +6 & \underline{-2} \\ & -2 & +8 & -6 & \\ \hline 1 & -4 & +3 & & \end{array}$$

We have

$$x^3 - 2x^2 - 5x + 6 = (x^2 - 4x + 3)(x + 2)$$

$$= (x - 1)(x - 3)(x + 2)$$

The three roots are: 1, 3, −2. Here is the graph:



Zeroes of Polynomial Functions:
Rational Zero Theorem and Descartes's Rule of Signs

Introduction

In this unit, we will use the Rational Zero Theorem and Descartes's Rule of Signs to find the zeroes of polynomial functions. Basically when you are finding a zero of a function, you are looking for input values that cause your functional value to be equal to zero. Sometimes the polynomial has a degree of 3 or higher, which makes it hard or impossible to factor. We have to break down higher degree polynomial functions into workable factors. Synthetic division will be use to help us out with this process.

Rational Zero (or Root) Theorem

If $f(x) = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0$,where

$a_n, a_{n-1}, \ldots, a_1$ are integer coefficients and the reduced fraction $\frac{p}{q}$ is a rational zero, then *p* is a factor of the constant term a_0 and *q* is a factor of the leading coefficient a_n.

We can use this theorem to help us find all of the POSSIBLE rational zeroes or roots of a polynomial function.

Step 1: List all of the factors of the constant.

In the Rational Zero Theorem, *p* represents factors of the constant term.
Make sure that you include both the positive and negative factors.

Step 2: List all of the factors of the leading coefficient.

In the Rational Zero Theorem, *q* represents factors of the leading coefficient.

Make sure that you include both the positive and negative factors.

Step 3: List all the POSSIBLE rational zeros or roots.

This list comes from taking all the factors of the constant (*p*) and writing them ove all the factors of the leading coefficient (q), to get a list of $\frac{p}{q}$. Make sure that you get ever possible combination of these factors, written as $\frac{p}{q}$.

EXAMPLE 1:
Use the Rational Zero Theorem to list all the possible rational zeros for $f(x) = -x^3 + 4x^2 - 5x + 12$.

Solution:
Step 1: List all of the factors of the constant.
The factors of the constant term 12 are $\pm 1, \pm 2, \pm 3, \pm 4, \pm 6, \pm 12$.

Step 2: List all of the factors of the leading coefficient.
The factors of the leading coefficient -1 are ± 1.

Step 3: List all the POSSIBLE rational zeros or roots.
Writing the possible factors as $\frac{p}{q}$ we get:

$$\pm \frac{1}{1}, \pm \frac{2}{1}, \pm \frac{3}{1}, \pm \frac{4}{1}, \pm \frac{6}{1}, \pm \frac{12}{1}$$

OR

$$\pm 1, \pm 2, \pm 3, \pm 4, \pm 6, \pm 12$$

EXAMPLE 2:

Use the Rational Zero Theorem to list all the possible rational zeros for $f(x) = 6x^3 + 5x^2 + 7x - 20$.

Step 1: List all of the factors of the constant.
The factors of the constant term -20 are $\pm 1, \pm 2, \pm 4, \pm 5, \pm 10, \pm 20$.

Step 2: List all of the factors of the leading coefficient.
The factors of the leading coefficient 6 are $\pm 1, \pm 2, \pm 3, \pm 6$.

Step 3: List all the POSSIBLE rational zeros or roots.

Writing the possible factors as $\frac{p}{q}$ we get:

$$\pm\frac{1}{1}, \pm\frac{2}{1}, \pm\frac{4}{1}, \pm\frac{5}{1}, \pm\frac{10}{1}, \pm\frac{20}{1}, \pm\frac{1}{2}, \pm\frac{2}{2}, \pm\frac{4}{2}, \pm\frac{5}{2}, \pm\frac{10}{2}, \pm\frac{20}{2}$$

$$\pm\frac{1}{3}, \pm\frac{2}{3}, \pm\frac{4}{3}, \pm\frac{5}{3}, \pm\frac{10}{3}, \pm\frac{20}{3}, \pm\frac{1}{6}, \pm\frac{2}{6}, \pm\frac{4}{6}, \pm\frac{5}{6}, \pm\frac{10}{6}, \pm\frac{20}{6}$$

Note, that some of the fractions are repeated, so they need to be reduced.
Here is a final list of all the POSSIBLE rational zeros, each one written once and reduced:

$$\pm 1, \pm 2, \pm 4, \pm 5, \pm 10, \pm 20, \pm\frac{1}{2}, \pm\frac{5}{2},$$

$$\pm\frac{1}{3}, \pm\frac{2}{3}, \pm\frac{4}{3}, \pm\frac{5}{3}, \pm\frac{10}{3}, \pm\frac{20}{3}, \pm\frac{1}{6}, \pm\frac{5}{6}$$

EXAMPLE 3:

List all of the possible zeros, use synthetic division to test the possible zeros, find an actual zero and use the actual zero to find all the zeros of $f(x) = x^3 + 4x^2 - 25x - 100$
.

List all of the possible zeros:

The factors of the constant term 100 are $\pm 1, \pm 2, \pm 4, = 5, \pm 10, \pm 20, \pm 25, \pm 50, \pm 100$
.

The factors of the leading coefficient 1 are ± 1.

Writing the possible factors as $\frac{p}{q}$ we get:

$$\pm\frac{1}{1}, \pm\frac{2}{1}, \pm\frac{4}{1}, \pm\frac{5}{1}, \pm\frac{10}{1}, \pm\frac{20}{1}, \pm\frac{25}{1}, \pm\frac{50}{1}, \pm\frac{100}{1}$$

OR

$$+1, +2, +4, +5, +10, +20, +25, +50, +100$$

Use synthetic division to test the possible zeros and find an actual zero:
Recall that if you apply synthetic division and the remainder is 0, then *c* is a zero or root of the polynomial function.

At this point you can pick any POSSIBLE rational root from the list of $\frac{p}{q}$. I would suggest starting with smaller easier numbers and then going from there.

Let choose 2:

2	1	4	-25	100
		2	12	-28
	1	6	-13	-126
				↑ remainder

Since the reminder came out -126, this means *f*(2) = -126, which means *x* = 2 is NOT a zero for this polynomial function. We need to choose another number that comes from that same list of POSSIBLE rational roots.

Let's choose -2:

-2	1	4	-25	100
		-2	-4	58
	1	2	-29	-42
				↑ remainder

Again, the reminder is not 0, so *x* = -2 is not a zero of this polynomial function.

This time let's choose - 4:

-4	1	4	-25	-100
		-4	0	100
	1	0	-25	0
				↑ remainder

At last, we found a number that has a remainder of 0. This means that ***x* = - 4 is a zero or root of our polynomial function.**

Use the actual zero to find all the zeros:
Since, *x* = - 4 is a zero, that means *x* + 4 is a factor of our polynomial function.

Rewriting *f*(*x*) as (*x* + 4)(quotient) we get:

$$f(x) = x^3 + 4x^2 - 25x - 100 = (x+4)(x^2 - 25)$$

We need to finish this problem by setting this equal to zero and solving it:

$(x+4)(x^2-25)=0$

$(x+4)(x+5)(x-5)=0$

***Factor the difference of squares**

$x+4=0$

$x=-4$

***Set 1st factor = 0**

***Set 2nd factor = 0**

$x+5=0$

$x=-5$

***Set 3rd factor = 0**

$x-5=0$

$x=5$

The zeros of this function are *x* = - 4, -5, and 5.

EXAMPLE 4:

List all the possible zeroes, use synthetic division to test the possible zeroes, find an actual zero and use the actual zero to find all the zeroes of $x^4-5x^3+20x-16=0$.

List all the possible zeroes:

The factors of the constant term -16 are $\pm 1, \pm 2, \pm 4, \pm 8, \pm 16$.

The factors of the leading coefficient 1 are ± 1.

Writing the possible factors as $\frac{p}{q}$ we get:

$\pm\frac{1}{1}, \pm\frac{2}{1}, \pm\frac{4}{1}, \pm\frac{8}{1}, \pm\frac{16}{1}$

OR

$\pm 1, \pm 2, \pm 4, \pm 8, \pm 16$

Use synthetic division to test the possible zeros and find an actual zero:

Recall that if you apply synthetic division and the remainder is 0, then *c* is a zero or root of the polynomial function.

At this point you can pick any POSSIBLE rational root from the list of $\frac{p}{q}$. I would suggest you start with smaller easier numbers and then going from there.

Let's choose -1:

-1	1	-5	0	20	-16
		-1	6	-6	-14
	1	-6	6	14	-30

Since the reminder came out -30, this means *f*(-1) = -30, which means *x* = -1 is NOT a zero for this polynomial function. We need to choose another number that comes from that same list of POSSIBLE rational roots.

This time I'm going to choose 1:

1	1	-5	0	20	-16
		1	-4	-4	16
	1	-4	-4	16	0

At last, we found a number that has a remainder of 0. This means that ***x* = 1 is a zero or root of our polynomial function.**

Use the actual zero to find all the zeros:

Since, *x* = 1 is a zero, that means *x* - 1 is a factor of our polynomial function.

Rewriting *f*(*x*) as (*x* - 1)(quotient) we get:

$$x^4 - 5x^3 + 20x - 16 = (x - 1)(x^3 - 4x^2 - 4x + 16)$$

$$(x-1)(x^3-4x^2-4x+16)=0$$
$$(x-1)\left[x^2(x-4)-4(x-4)\right]=0$$
$$(x-1)(x-4)(x^2-4)=0$$

***Factor by grouping**

$$(x-1)(x-4)(x+2)(x-2)=0$$

***Factor the difference of squares**

$$x-1=0$$
$$x=1$$

***Set 1st factor = 0**

***Set 2nd factor = 0**

$$x-4=0$$
$$x=4$$

***Set 3rd factor = 0**

$$x+2=0$$
$$x=-2$$

***Set 4th factor = 0**

$$x-2=0$$
$$x=2$$

The zeros of this function are *x* = 1, 4, -2 and 2.

Descartes's Rule Of Signs

Let $f(x) = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_1 x + a_0$ be a polynomial where $a_n, a_{n-1}, \ldots, a_1$ are real coefficients.

The number of **POSITIVE REAL ZEROS of *f*** is either equal to the number of sign changes of successive terms of *f*(*x*) or is less than that number by an even number (until 1 or 0 is reached).
The number of **NEGATIVE REAL ZEROS of *f*** is either equal to the number of sign changes of successive terms of *f*(-*x*) or is less than that number by an even integer (until 1 or 0 is reached).

This can help narrow down your possibilities when you do go on to find the zeros.

Example 5:

Find the possible number of positive and negative real zeroes of $f(x) = 3x^4 - 5x^3 + 2x^2 - x + 10$ using Descartes's Rule of Signs.

In this problem it isn't asking for the zeroes themselves, but what the possible numbers of zeroes are. This can help narrow down your possibilities when you do go on to find the zeroes.

Possible number of positive real zeroes:

$$f(x) = 3x^4 \overset{\uparrow}{-} 5x^3 \overset{\uparrow}{+} 2x^2 \overset{\uparrow}{-} x \overset{\uparrow}{+} 10$$

The up arrows are showing where there are sign changes between successive terms, going left to right. The first arrow on the left shows a sign change from positive 3 to negative 5. The 2nd arrow shows a sign change from negative 5 to positive 2. The third arrow shows a sign change from positive 2 to negative 1. And the last arrow shows a sign change from negative 1 to positive 10.

There are 4 sign changes between successive terms, which mean that is the highest possible number of positive real zeros. To find the other possible number of positive real zeroes from these sign changes, you start with the number of changes, which in this case is 4 and then go down by even integers from that number until you get to 1 or 0.

Since we have 4 sign changes with *f*(*x*), then **there is a possibility of 4 or 4 - 2 = 2 or**

4 - 4 = 0 positive real zeroes.

Possible number of negative real zeros:

$$f(-x) = 3(-x)^4 - 5(-x)^3 + 2(-x)^2 - (-x) + 10$$
$$= 3x^4 + 5x^3 + 2x^2 + x + 10$$

Note how there are no sign changes between successive terms.

This means there are no negative real zeros.
Since we are counting the number of possible real zeros, 0 is the lowest number that we can have. This piece of information would be helpful when determining real zeroes for polynomial.

Example 6:
Find the possible number of positive and negative real zeros of
$f(x) = 2x^2 - 7x - 8$ using Descartes's Rule of Signs.

In this problem it isn't asking for the zeros themselves, but what the possible numbers of zeroes are. This can help narrow down your possibilities when you do go on to find the zeroes.

Possible number of positive real zeroes:

$$f(x) = 2x^2 - 7x - 8$$
↑

The up arrow is showing where there is a sign change between successive terms, going left to right. This arrow shows a sign change from positive 2 to negative 7.
There is only 1 sign change between successive terms, which means that is the highest possible number of positive real zeroes. To find the other possible number of positive real zeroes from these sign changes, you start with the number of changes, which in this case is 1 and then go down by even integers from that number until you get to 1 or 0.
If we went down by even integers from 1, we would be in the negative numbers, which is not a feasible answer, since we are looking for the possible number of positive real zeroes. In other words, we can't have a -1 of them.
Therefore, **there is exactly 1 positive real zero.**

Possible number of negative real zeroes:

$$f(-x) = 2(-x)^3 - 7(-x) - 8$$
$$= -2x^3 + 7x - 8$$
$$\quad\uparrow\quad\uparrow$$

The up arrows are showing where there are sign changes between successive terms, going left to right. The first arrow on the left shows a sign change from negative 2 to positive 7. The 2nd arrow shows a sign change from positive 7 to negative 8.

There are 2 sign changes between successive terms, which mean that is the highest possible number of negative real zeros. To find the other possible number of negative real zeros from these sign changes, you start with the number of changes, which in this case is 2, and then go down by even integers from that number until you get to 1 or 0.

Since we have 2 sign changes with $f(-x)$, then **there is a possibility of 2 or 2 - 2 = 0 negative real zeros.**

Example 7:
List all of the possible zeros, use Descartes's Rule of Signs to possibly narrow it down, use synthetic division to test the possible zeroes and find an actual zero, and use the actual zero to find all the zeros of $f(x) = 3x^3 - 8x^2 - 5x - 2$.

Solution:
List all the possible zeroes:
The factors of the constant term -2 are $\pm 1, \pm 2$.

The factors of the leading coefficient 3 are $\pm 1, \pm 3$.
Writing the possible factors as $\frac{p}{q}$ we get:

$$\pm\frac{1}{1}, \pm\frac{1}{3}, \pm\frac{2}{1}, \pm\frac{2}{3}$$
$$OR$$
$$\pm 1, \pm 2, \pm\frac{1}{3}, \pm\frac{2}{3}$$

Before we try any of these, let's apply Descartes's Rule of Signs to see if it can help narrow down our search for a rational zero.

Possible number of positive real zeros:

$$f(x) = 3x^3 - 8x^2 + 5x - 2$$

$$\qquad\quad \uparrow \quad \uparrow \quad \uparrow$$

The up arrows are showing where there are sign changes between successive terms, going left to right.

There are 3 sign changes between successive terms, which mean, that is the highest possible number of positive real zeroes. To find the other possible number of positive real zeroes from these sign changes, you start with the number of changes, which in this case is 3, and then go down by even integers from that number until you get to 1 or 0.

Since we have 3 sign changes with $f(x)$, then **there is a possibility of 3 or 3 - 2 = 1 positive real zeros.**

Possible number of negative real zeros:

$$f(-x) = 3(-x)^3 - 8(-x)^2 + 5(-x) - 2$$

$$= -3x^3 - 8x^2 - 5x - 2$$

Note that there are no sign changes between successive terms.

This means there are no negative real zeros.

Since we are counting the number of possible real zeros, 0 is the lowest number that we can have. This will help us narrow things down in the next step.

Use synthetic division to test the possible zeroes and find an actual zero:

Recall that if you apply synthetic division and the remainder is 0, then c is a zero or root of the polynomial function.

At this point you pick any POSSIBLE rational root from the list of $\frac{p}{q}$.

Above, we found that **there are NO negative rational zeroes**, so we do not have to bother with trying any negative numbers. See how Descartes's has helped us. I would suggest we start with smaller easier numbers and then go from there.

Let's choose 1:

1	3	-8	5	-2
		3	-5	0
	3	-5	0	-2
				↑ remainder

Since the reminder came out -2, this means $f(1) = -2$, which means $x = 1$ is NOT a zero for this polynomial function. We need to choose another number that comes from that same list of POSSIBLE rational roots.

This time I'm going to choose 2:

2	3	-8	5	-2
		6	-4	2
	3	-2	1	0
				↑ remainder

At last, we found a number that has a remainder of 0. This means that ***x* = 2 is a zero or root of our polynomial function.**

Use the actual zero to find all the zeroes:

Since, $x = 2$ is a zero, that means $x - 2$ is a factor of our polynomial function. Rewriting $f(x)$ as $(x - 2)$(quotient) we get:

$$f(x) = 3x^3 - 8x^2 + 5x - 2$$

$$= (x - 2)(3x^2 - 2x + 1)$$

We need to finish this problem by setting this equal to zero and solving it:

$$(x - 2)(3x^2 - 2x + 1) = 0$$

$$x - 2 = 0$$

$$x = 2$$

***Set 1st factor = 0**

***Set 2nd factor = 0**

***This is a quadratic that does not factor**

***Use the <u>quadratic formula</u>**

$$3x^2 - 2x + 1 = 0$$

$$x = \frac{-(-2) \pm \sqrt{(-2)^2 - 4(3)(1)}}{2(3)}$$

$$= \frac{2 \pm \sqrt{-8}}{6}$$

$$= \frac{2 \pm 2i\sqrt{2}}{6}$$

$$= \frac{1 \pm i\sqrt{2}}{3}$$

The zeros of this function are $x = 2$**,** $\frac{1+i\sqrt{2}}{3}$ **, and** $\frac{1-i\sqrt{2}}{3}$ **.**

Example 8: List all the possible zeroes, use Descartes's Rule of Signs to possibly narrow it down, use synthetic division to test the possible zeroes and find an actual zero, and use the actual zero to find all the zeros of $x^5 + 8x^4 + 20x^3 + 10x^2 - 21x - 18 = 0$.

List all the possible zeros:

The factors of the constant term -18 are $\pm 1, \pm 2, \pm 3, \pm 6, \pm 9, \pm 18$.

The factors of the leading coefficient 1 are ± 1.

Writing the possible factors as $\frac{p}{q}$ **we get:**

$$\pm\frac{1}{1}, \pm\frac{2}{1}, \pm\frac{3}{1}, \pm\frac{6}{1}, \pm\frac{9}{1}, \pm\frac{18}{1}$$

OR

$$\pm 1, \pm 2, \pm 3, \pm 6, \pm 9, \pm 18$$

Before we try any of these, let's apply Descartes's Rule of Signs to see if it can help narrow down our search for a rational zero.

Possible number of positive real zeroes:

$$x^5 + 8x^4 + 20x^3 + 10x^2 - 21x - 18$$

↑

The up arrow is showing where there is a sign change between successive terms, going left to right.
There is 1 sign change between successive terms, which means that is the highest possible number of positive real zeroes.
Since we have 1 sign change with $f(x)$, then **there is exactly 1 positive real zero.**

Possible number of negative real zeros:

$$f(-x) = (-x)^5 - 8(-x)^4 + 20(-x)^3 + 10(-x)^2 - 21(-x) - 18$$

$$= -x^5 + 8x^4 - 20x^3 + 10x^2 + 21x - 18$$

↑ ↑ ^ ↑

The up arrows are showing where there are sign changes between successive terms, going left to right. There are 4 sign changes between successive terms, which means that is the highest possible number of negative real zeros. To find the other possible number of negative real zeros from these sign changes, you start with the number of changes, which in this case is 4, and then go down by even integers from that number until you get to 1 or 0. Since we have 4 sign changes with $f(x)$, then there are possibility of 4, 4 - 2 = 2 or 4 - 4 = 0 negative real zeros.

Use synthetic division to test the possible zeros and find an actual zero:

At this point you can pick any POSSIBLE rational root from the list of $\frac{p}{q}$. Above, we found that there is exactly 1 positive rational zero. Since we know that there is 1 for sure, then we may want to go ahead and **start with trying positive rational roots.** I would suggest starting with smaller easier numbers and then go from there.

Let's choose 1 to try:

1	1	8	20	10	-21	-18
		1	9	29	39	18
	1	9	29	39	18	0 ↑ remainder

We found a number that has a remainder of 0. This means that ***x* = 1 is a zero or root of our polynomial function.**

Use the actual zero to find all the zeroes:

Since, $x = 1$ is a zero, that means $x - 1$ is a factor of our polynomial function. Rewriting $f(x)$ as $(x - 1)$(quotient) we get:

$$x^5 + 8x^4 + 20x^3 + 10x^2 - 21x - 18$$
$$= (x - 1)(x^4 + 9x^3 + 29x^2 + 39x + 18)$$

We need to finish this problem by setting this equal to zero and solving it:

$$(x - 1)(x^4 + 9x^3 + 29x^2 + 39x + 18) = 0$$

$$x - 1 = 0$$
$$x = 1$$

***Set 1st factor = 0**

$$x^4 + 9x^3 + 29x^2 + 39x + 18 = 0$$

Looks like we can't factor this one. We are going to have to repeat this process again, but this time we will use this factor that we found.

Recall, that in Descartes's Rule of Signs we already found that there is exactly one positive real zero. It looks like we already found that, so **when we are trying again we can focus on finding a negative real zero.** Note that we can still pick from the same list of $\frac{p}{q}$ numbers as we did above, since we are still looking at solving the same overall problem. However when we set up the synthetic division, we will just look at the remaining factor, to help us factor that down farther.

Let's try -1:

$$\begin{array}{r|rrrrr} -1 & 1 & 9 & 29 & 39 & 18 \\ & & -1 & -8 & -21 & -18 \\ \hline & 1 & 8 & 21 & 18 & 0 \end{array}$$

$\uparrow$ remainder

We found a number that has a remainder of 0. This means that x = -1 is a zero or root of our polynomial function.

Use the actual zero to find all the zeros:

Since, x = -1 is a zero, that means x +1 is a factor of our polynomial function.

Rewriting $f(x)$ as $(x - 1)(x + 1)$(quotient) we get:

$$x^5 + 8x^4 + 20x^3 + 10x^2 - 21x - 18$$
$$= (x-1)(x^4 + 9x^3 + 29x^2 + 39x + 18)$$
$$= (x-1)(x+1)(x^3 + 8x^2 + 21x + 18)$$

Looks like we can't factor this one. We are going to repeat this process again, but this time we will use this factor that we found.

Recall, that in Descartes's Rule of Signs we already found that there is exactly one positive real zero. It looks like we already found that, so **when we are trying again we can focus on finding a negative real zero.** Note that we can still pick from the same list of $\frac{p}{q}$ numbers as we did above, since we are still looking at solving the same overall problem.

However when we set up the synthetic division, we will just look at the remaining factor, to help us factor that down farther.

Let's choose -2 :

-2	1	8	21	18
		-2	-12	-18
	1	6	9	0
				↑ remainder

We found a number that has a remainder of 0. This means that x = -2 is a zero or root of our polynomial function.

Use the actual zero to find all the zeros:
Since, x = -2 is a zero, that means x + 2 is a factor of our polynomial function.

Rewriting $f(x)$ as $(x - 1)(x + 1)(x + 2)$(quotient) we get:

$$x^5 + 8x^4 + 20x^3 + 10x^2 - 21x - 18$$
$$= (x - 1)(x^4 + 9x^3 + 29x^2 + 39x + 18)$$
$$= (x - 1)(x + 1)(x^3 + 8x^2 + 21x + 18)$$
$$= (x - 1)(x + 1)(x + 2)(x^2 + 6x + 9)$$

We need to finish this problem by setting this equal to zero and solving it:

$(x-1)(x+1)(x+2)(x^2+6x+9)=0$

$(x-1)(x+1)(x+2)(x+3)(x+3)=0$

***Factor the trinomial**

$x-1=0$

$x=1$

***Set 1st factor = 0**

$x+1=0$

$x=-1$

***Set 2nd factor = 0**

$x+2=0$

$x=-2$

***Set 3rd factor = 0**

$x+3=0$

$x=-3$

***Set 4th factor = 0**

The zeros of this function are *x* = 1, -1, -2, and -3.

Exercises

1. Use the Rational Zero Theorem to list all the possible rational zeros for the given polynomial function.

$f(x) = -4x^2 + 7x + 10$

2. Find the possible number of positive and negative real zeros of the given polynomial function using Descartes's Rule of Signs.

$f(x) = -7x^4 + 8x^3 - 7x^2 + 9x + 10$

3. List all the possible zeroes, use Descartes's Rule of Signs to possibly narrow it down, use synthetic division to test the possible zeroes and find an actual zero, and use the actual zero to find all the zeroes of the given polynomial function.

a. $f(x) = x^3 - 10x^2 + 11x + 70$

b. $5x^5 - 28x^4 + 50x^3 - 46x^2 + 45x - 18 = 0$

Zeroes of Polynomial Functions:
Upper and Lower Bounds, Intermediate Value Theorem

The Upper and Lower Bound Theorem

Upper Bound

If you divide a polynomial function $f(x)$ by $(x - c)$, where $c > 0$, using synthetic division and this yields all positive numbers, then c is an upper bound to the real roots of the equation $f(x) = 0$.

Note that two things must occur for c to be an upper bound. One is $c > 0$ or positive. The other is that all the coefficients of the quotient as well as the remainder are positive.

Lower Bound

If you divide a polynomial function $f(x)$ by $(x - c)$, where $c < 0$, using synthetic division and this yields alternating signs, then c is a lower bound to the real roots of the equation $f(x) = 0$. Special note that zeroes can be either positive or negative.

Note that two things must occur for c to be a lower bound. One is $c < 0$ or negative. The other is that successive coefficients of the quotient and the remainder have alternating signs.

Example 1:

Show that all real roots of the equation $x^5 - 5x^3 - 10x^2 + 12x - 20 = 0$ lie between - 4 and 4.

Solution

In other words, we need to show that - 4 is a lower bound and 4 is an upper bound for real roots of the given equation.

Checking the Lower Bound:

Lets apply **synthetic division** with - 4 and see if we get alternating signs:

-4	1	0	-5	-10	12	-20
		-4	16	-44	216	-912
	1	-4	11	-54	228	-932

Note that $c = -4 < 0$ AND the successive signs in the bottom row of our synthetic division alternate.

You know what that means?

- 4 is a lower bound for the real roots of this equation.

Checking the Upper Bound:
Lets apply synthetic division with 4 and see if we get all positive:

4	1	0	-5	-10	12	-20
		4	16	44	136	592
	1	4	11	34	148	572

Note that $c = 4 > 0$ AND all the signs in the bottom row of our synthetic division are positive.

You know what that means?
4 is an upper bound for the real roots of this equation.

Since - 4 is a lower bound and 4 is an upper bound for the real roots of the equation, then that means **all real roots of the equation lie between - 4 and 4.**

The Intermediate Value Theorem

Theorem 1: If $f(x)$ is a polynomial function and $f(a)$ and $f(b)$ have different signs, then there is at least one value, c, between a and b such that $f(c) = 0$.

In other words, when you have a polynomial function and one input value causes the function to be positive and the other negative, then there has to be at least one value in between them that causes the polynomial function to be 0.

This works because 0 separates the positives from the negatives. So to go from positive to negative or vice - versa you would have to hit a point in between that goes through 0.

Example 2:

Show that $f(x) = x^3 - 3x^2 + 3x - 4$ has a real zero between 2 and 3. Use the Intermediate Value theorem to find an approximation for this zero to the nearest tenth.

Solution:

When finding functional values, you can either use synthetic division or directly plug the number into the function.

Finding *f*(2):

$f(2) = (2)^3 - 3(2)^2 + 3(2) - 4 = -2$

Finding *f*(3):

$f(3) = (3)^3 - 3(3)^2 + 3(3) - 4 = 5$

Since there is a sign change between *f*(2) = -2 and *f*(3) = 5, then according to the **Intermediate Value Theorem**, there is at least one value between 2 and 3 that is a zero of this polynomial function.

Checking functional values at intervals of one-tenth for a sign change:

$f(2) = (2)^3 - 3(2)^2 + 3(2) - 4 = -2$

Finding *f*(2.1):

$f(2.1) = (2.1)^3 - 3(2.1)^2 + 3(2.1) - 4 = -1.669$

Finding *f*(2.2):

$f(2.2) = (2.2)^3 - 3(2.2)^2 + 3(2.2) - 4 = -1.272$

Finding *f*(2.3):

$f(2.3) = (2.3)^3 - 3(2.3)^2 + 3(2.3) - 4 = -.803$

Finding *f*(2.4):

$f(2.4) = (2.4)^3 - 3(2.4)^2 + 3(2.4) - 4 = -.256$

Finding *f*(2.5):

$f(2.5) = (2.5)^3 - 3(2.5)^2 + 3(2.5) - 4 = .375$

Note that we have the sign change
Now we want to find the zero to the nearest tenth. So it is going to be $x = 2.4$ or $x = 2.5$. We can now check for the functional value which is closer to zero. We will need to dig a little bit deeper and go by intervals of one-hundredths:

$$f(2.4) = (2.4)^3 - 3(2.4)^2 + 3(2.4) - 4 = -.256$$

Finding *f*(2.41):

$$f(2.41) = (2.41)^3 - 3(2.41)^2 + 3(2.41) - 4 = -.196779$$

Finding *f*(2.42):

$$f(2.42) = (2.42)^3 - 3(2.42)^2 + 3(2.42) - 4 = -.136712$$

Finding *f*(2.43):

$$f(2.43) = (2.43)^3 - 3(2.43)^2 + 3(2.43) - 4 = -.075793$$

Finding *f*(2.44):

$$f(2.44) = (2.44)^3 - 3(2.44)^2 + 3(2.44) - 4 = -.014016$$

Finding *f*(2.45):

$$f(2.45) = (2.45)^3 - 3(2.45)^2 + 3(2.45) - 4 = .048625$$

Now at last we have gotten a sign change between successive hundredths. That means we have narrowed it down which makes it a little bit bitter. There is a zero between 2.44 and 2.45.

Since it would land slightly below 2.45, the nearest tenth would be 2.4.
The work is not hard but a little bit tedious.

Theorem 2:

If f(x) is a continuous function on some interval [a,b] and f(a) f(b) < 0, then the equation f(x) = 0 has at least one real root or an odd number of real roots in the interval (a,b).

We can set up a table of values of f(x) for various values of x and obtain a suitable initial approximation to the root.

Examples:

1. The equation $8x^3 - 12x^2 - 2x + 3 = 0$

has three real roots. Find the intervals each of unit length containing each one of these roots.

Solution:

We prepare a table of the values of the function f(x) for various values of x

x	-2	-1	0	1	2	3
f(x)	-105	-15	3	-3	15	105

From the table, we find that the equation f(x) = 0 has roots in the intervals (-1,0), (0,1) and (1,2). The exact roots are -0.5, 0.5 and 1.5

2. Obtain the interval which contains a root of the equation $f(x) = \cos x - xe^{x} = 0$

Solution:

We prepare a table of the values of the function f(x) for various values of x

x	0	0.5	1	1.5	2
f(x)	1	0.0532	-2.1780	-6.6518	-15.1942

From the table, we find that the equation f(x) = 0 has at least one roots in the intervals (0.5, 1). The exact root correct to ten decimal places is 0.5177573637.

UNIT 7

BISECTION METHOD

Bisection method is the simplest method of bracketing the roots of a function and requires an initial interval which is guaranteed to contain a root -- if a and b are the endpoints of the interval then f(a) must differ in sign from f(b). This ensures that the function crosses zero at least once in the interval. If a valid initial interval is used then these algorithm cannot fail, provided the function is well behaved.

On each iteration, the interval is bisected and the value of the function at the midpoint is calculated. The sign of this value is used to determine which half of the interval does not contain a root. That half is discarded to give a new, smaller interval containing the root. This method can be continued indefinitely until the interval is sufficiently small. At any time, the current estimate of the root is taken as the midpoint of the interval.

Bisection method has linear convergence. Linear convergence means that successive significant figures are won linearly with computational effort.

When an interval contains more than one root, the bisection method can find one of them. When an interval contains a singularity, the bisection method converges to that singularity.

Theorem (Bisection Theorem).
Assume that $f \in C[a, b]$ and that there exists a number $r \in [a, b]$ such that f(r) = 0.
If f(a) and f(b) have opposite signs, and $\{c_n\}$ represents the sequence of midpoints generated by the bisection process, then

$$\left| r - c_n \right| \leq \frac{b - a}{2^{n+1}} \quad \text{for} \quad n = 0, 1, \ldots,$$

and the sequence $\{c_n\}$ converges to the zero $x = r$.

That is, $\lim_{k \to \infty} c_n = r$.

Method

The method is applicable when we wish to solve the equation $f(x) = 0$ for the real variable x, where f is a continuous function defined on an interval $[a, b]$ and $f(a)$ and $f(b)$ have opposite signs. In this case a and b are said to bracket a root since, by the intermediate value theorem, the f must have at least one root in the interval (a, b).

At each step the method divides the interval in two by computing the midpoint $c = (a+b) / 2$ of the interval and the value of the function $f(c)$ at that point. Unless c is itself a root (which is very unlikely, but possible) there are now two possibilities: either $f(a)$ and $f(c)$ have opposite signs and bracket a root, or $f(c)$ and $f(b)$ have opposite signs and bracket a root. The method selects the subinterval that is a bracket as a new interval to be used in the next step. In this way the interval that contains a zero of f is reduced in width by 50% at each step. The process is continued until the interval is sufficiently small.

Explicitly, if $f(a)$ and $f(c)$ are opposite signs, then the method sets c as the new value for b, and if $f(b)$ and $f(c)$ are opposite signs then the method sets c as the new a. (If $f(c)$=0 then c may be taken as the solution and the process stops.) In both cases, the new $f(a)$ and $f(b)$ have opposite signs, so the method is applicable to this smaller interval.

Analysis

The method is guaranteed to converge to a root of f if f is a continuous function on the interval $[a, b]$ and $f(a)$ and $f(b)$ have opposite signs. The absolute error is halved at each step so the method converges linearly, which is comparatively slow.

Specifically, if $p_1 = (a+b)/2$ is the midpoint of the initial interval, and p_n is the midpoint of the interval in the nth step, then the difference between p_n and a solution p is bounded by

$$|p_n - p| \leq \frac{|b-a|}{2^n}.$$

This formula can be used to determine in advance the number of iterations that the bisection method would need to converge to a root to within a certain tolerance.

Example:

To find a root of $x^3 - 4x - 9 = 0$ correct to 3 decimal places using bisection method, take $f(x) = x^3 - 4x - 9$.

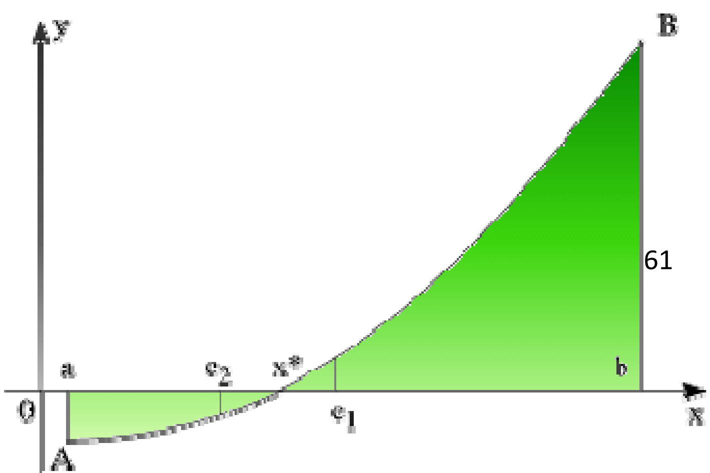
First, $f(2) = -9 < 0$ and $f(3) = 6 > 0$ so a root lies between 2 and 3.
The first approximation to the root is then $x_1 = 1/2(a + b) = 2.5$.
Then $f(x_1) = -3.375 < 0$, so the root lies between x_1 and 3.
Thus second approximation to the root is $x_2 = 1/2(2.5 + 3) = 2.75$.
$f(x_2) = 0.7969 > 0$, so the root lies between x_1 and x_2.
Continue until the size of the intervals is less than the required tolerance .001, so 10 steps are required.
This gives $x_{10} = 2.706$ as the required root

Bisection Algorithm

INPUT: Function *f*, endpoint values *a*, *b*, tolerance *TOL*, maximum iterations *NMAX*
CONDITIONS: *a* < *b*, either *f*(*a*) < 0 and *f*(*b*) > 0 or *f*(*a*) > 0 and *f*(*b*) < 0
OUTPUT: value which differs from a root of *f*(*x*)=0 by less than *TOL*

```
N ← 1
While N ≤ NMAX { limit iterations to prevent infinite loop
  c ← (a + b)/2 new midpoint
  If (f(c) = 0 or (b – a)/2 < TOL then { solution found
    Output(c)
    Stop
  }
  N ← N + 1 increment step counter
  If sign(f(c)) = sign(f(a)) then a ← c else b ← c new interval
}
Output("Method failed.") max number of steps exceeded
```

```
PROCEDURE Bisection(a,b,eps:Real; VAR xsol:Real);
{ Required condition: f(a)*f(b)<0 }
{ eps = accuracy of the root, e.g.: 0.000001 }
VAR
  c:Real;
BEGIN
  REPEAT
    c:=(a+b)/2;
    IF f(a)*f(c)<0 THEN b:=c
    ELSE a:=c
  UNTIL b-a<eps;
  xsol:=c
END; {Bisection method - Pascal code}
```



UNIT 8

NEWTON'S METHOD

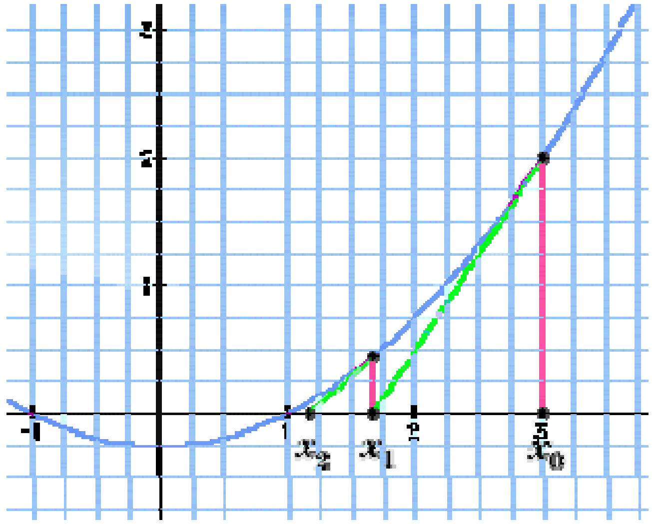
Newton method is the standard root-polishing algorithm. Newton method, also called the Newton-Raphson method, is a root-finding algorithm that uses the first few terms of the Taylor series of a function f(x) in the vicinity of a suspected root.

$$f(x+\delta) \approx f(x) + f'(x)\delta + \frac{f''(x)}{2}\delta^2 + \dots$$

Newton algorithm begins with an initial guess for the location of the root. On each iteration, a line tangent to the function f is drawn at that position. The point where this line crosses the x-axis becomes the new guess.

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}.$$

Newton method converges quadratically for single roots, and linearly for multiple roots.



Example: $f(x) = x^3 + 2x - 2$.

First, recall Newton's Method is for finding roots (or zeros) of functions. In order to use Newton's Method, you need to (1) make a first "guess" as to what you think the root is and (2) find the derivative of the function. You then use the following, easily-derived formula (where x_0 is your first guess) to arrive at your second guess, called x_1:

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

Again, we are trying to find when $f(x) = 0$.

A good first guess is $x = 1$, since $f(1) = 1$, and $f'(x) = 3x^2 + 2$.

Then

$$\begin{aligned} x_1 &= 1 - \frac{f(1)}{f'(1)} \\ &= 1 - \frac{1}{5} \\ &= 0.8 \end{aligned}$$

We can then continue this process iteratively, using x_1 as our next "guess":

$$x_2 = 0.8 - \frac{f(0.8)}{f'(0.8)}$$
$$= 0.8 - \frac{0.112}{3.92}$$
$$= 0.7714285714$$

Newton Alogorithms

The function below, written in Pascal and C, takes a simpler approach, ignoring the situation in which Newton method does not converge.

```
PROCEDURE Newton(c, eps:Real; VAR xsol:Real);
VAR
   d:Real;
BEGIN
{ df(x) = value of the first derivative }
{ eps = accuracy of the root, e.g.: 0.000001 }
   REPEAT
     d:=c;
     c:=c-f(c)/df(c)
   UNTIL Abs(d-c)<eps;
   xsol:=c
END; {Newton method - Pascal code}
```

Newton.c

```
// Implementation of the Newton algorithm in C

#include <stdio.h>
#include <math.h>

double newton(double x_0, double tol, int max_iters,
        int* iters_p, int* converged_p);
double f(double x);
double f_prime(double x);

int main() {
  double x_0;      /* Initial guess              */
  double x;        /* Approximate solution       */
  double tol;      /* Maximum error              */
```

```
  int    max_iters; /* Maximum number of iterations */
  int    iters;     /* Actual number of iterations  */
  int    converged; /* Whether iteration converged  */

  printf("Enter x_0, tol, and max_iters\n");
  scanf("%lf %lf %d", &x_0, &tol, &max_iters);

  x = newton(x_0, tol, max_iters, &iters, &converged);

  if (converged) {
   printf("Newton algorithm converged after %d steps.\n",
        iters);
   printf("The approximate solution is %19.16e\n", x);
   printf("f(%19.16e) = %19.16e\n", x, f(x));
  } else {
   printf("Newton algorithm didn't converge after %d steps.\n",
         iters);
   printf("The final estimate was %19.16e\n", x);
   printf("f(%19.16e) = %19.16e\n", x, f(x));
  }

  return 0;
}  /* main */


double newton(double x_0, double tol, int max_iters,
        int* iters_p, int* converged_p) {
  double x = x_0;
  double x_prev;
  int    iter = 0;

  do {
    iter++;
    x_prev = x;
    x = x_prev - f(x_prev)/f_prime(x_prev);
  } while (fabs(x - x_prev) > tol && iter < max_iters);

  if (fabs(x - x_prev) <= tol)
    *converged_p = 1;
  else
```

```
    *converged_p = 0;
  *iters_p = iter;

  return x;
}  /* newton algorithm */


double f(double x) {
  return x*x-2;
}  /* f */

double f_prime(double x) {
  return 2*x; //the derivative
}  /* f_prime */
```