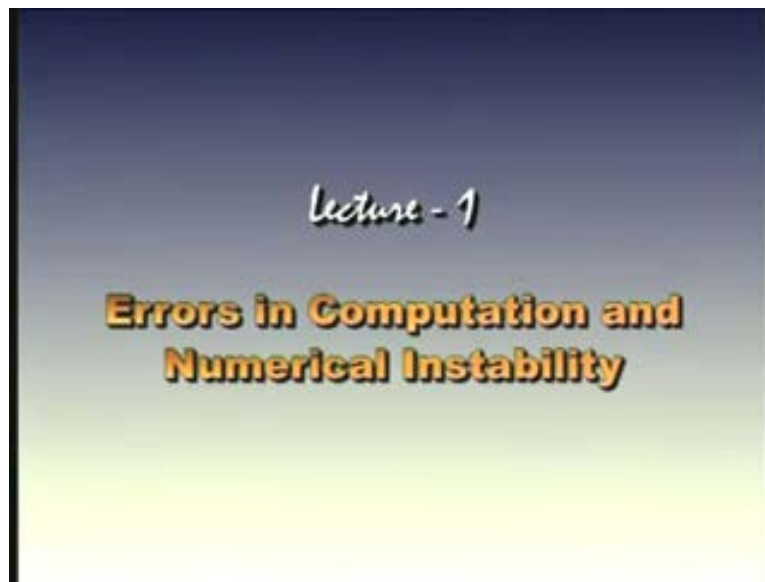


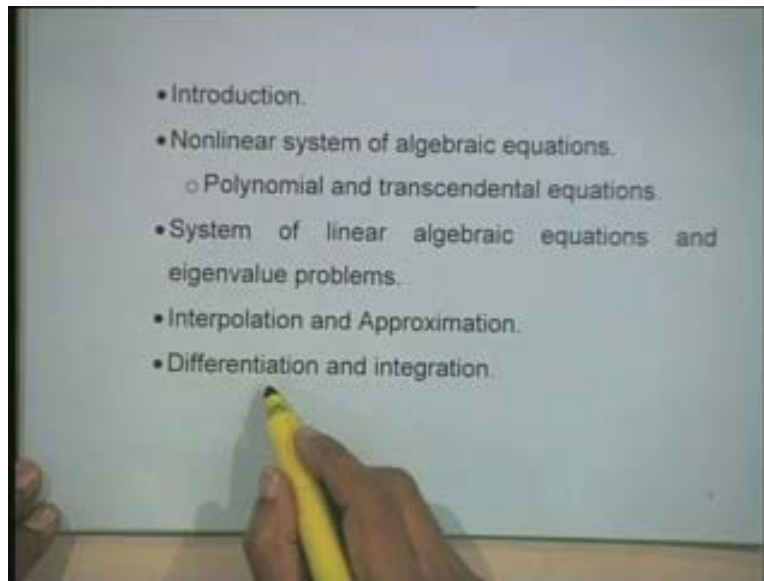
Numerical Methods and Computation
Prof. S.R.K. Iyengar
Department of Mathematics
Indian Institute of Technology Delhi
Lecture No # 1
Errors in Computation and Numerical Instability

(Refer Slide Time: 00:00:56 min)



In this course we shall derive and analyze numerical methods for the solution of various problems. We shall discuss the following topics.

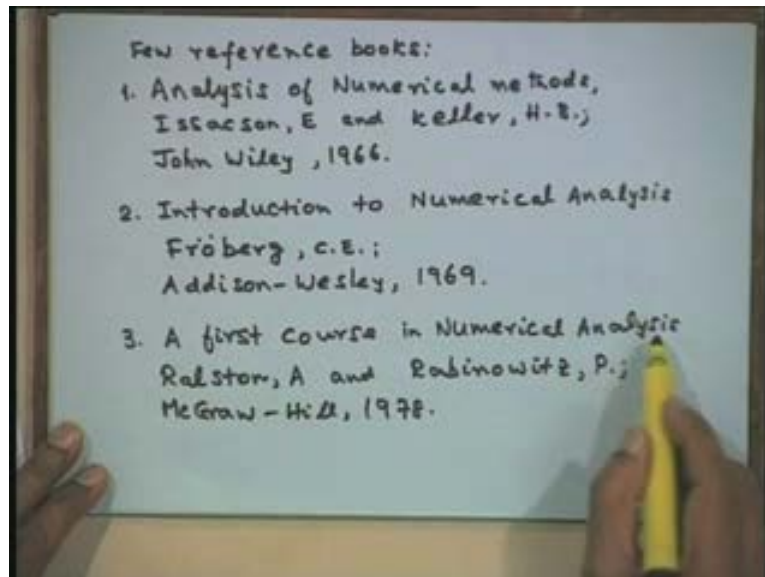
(Refer Slide Time: 00:01:13 min)



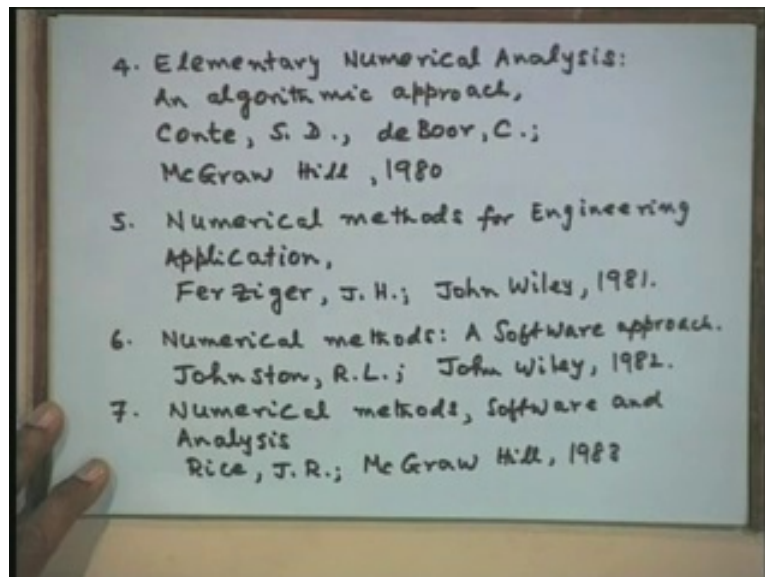
In the first topic we shall discuss the numerical solution of nonlinear system of algebraic equations. We shall construct methods for finding the roots or zeros of a transcendental or a polynomial equation in one variable. Then we shall extend the methods for the solution of a system of nonlinear equations. In the next topic we shall construct methods for the solution of linear algebraic equations and matrix eigenvalue problems. In the next topic we consider a data or a table of values and construct the polynomial that fits this data exactly. This polynomial can be used for interpolating or predicting the value of the function, represent the data at any intermediate point. This polynomial may also be used for various other operations like differentiation and integration.

In approximation we shall deal with approximation to a continuous function or to a function which represents the given data. In the next topic we shall use the interpolating polynomial of a given data to find the derivative if it exists of a function at a natal point or at any intermediate point. We shall also construct methods to numerically integrate a given function or to integrate a function that represents a given data. With the advent of high speed computers it is now possible to numerically solve many complex mathical models that represent the physical processes. A numerical method produces numbers as a required solution of the given problem. Many types of errors arise in solving a problem numerically and these errors must be kept under control otherwise errors dominate the solution finally. There are many books that are available on numerical methods few of the books are listed here.

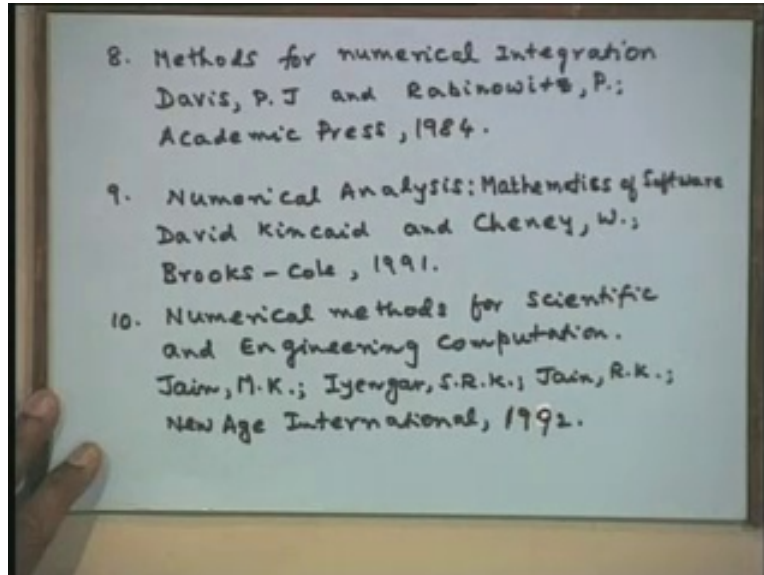
(Refer Slide Time: 00:03:02 min)



(Refer Slide Time: 00:03:11 min)



(Refer Slide Time: 00:03:26 min)

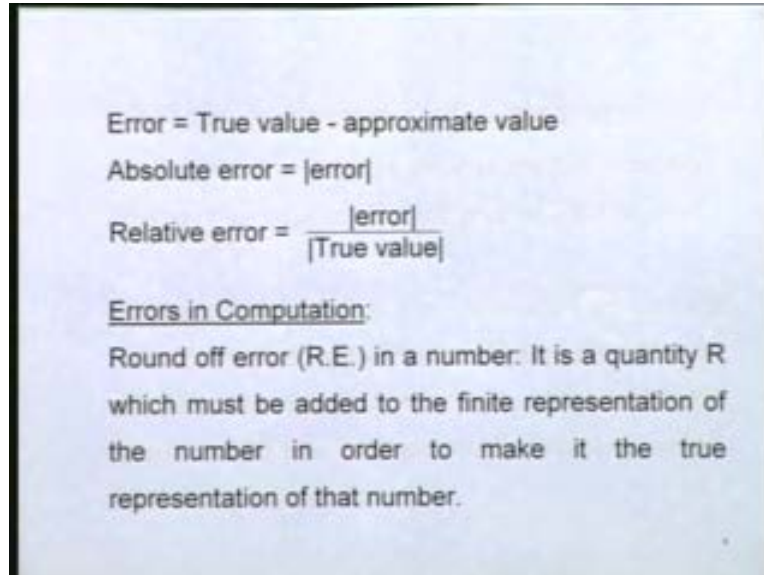


We shall follow the course material given in this last book.

Let us now define errors, discuss about the sources of errors and their effect on computation. Now in earlier years simple mathematical models were constructed to represent various physical processes using many approximations, so that the problem can be solved analytically. There were no avenues other than analytical methods in those days, but these are very approximate solutions. With the advent of the modern computers the models were now refined further and further, removed many approximations and ultimately they have come across a model which could be without any approximations. For example, you can have weather forecasting model or the very difficult problems or the nuclear reactor problems. The problems of very complex natures have been now constructed. But then these problems cannot be solved analytically. Therefore a new branch of mathematics, the numerical methods has come into existence to study or to solve these particular mathematical models.

What does numerical method give? A numerical method is constructed to give a typical problem. For example, if you said solution of linear algebraic equations we shall say that these are the methods that you can use for solving the system of linear algebraic equations. Similarly, a typical numerical method is constructed for a given typical problem. It gives you finally numbers as the answers. It says that these are the numbers and take them as the solution of your problem. Therefore it is necessary for us to see what kind of errors have gone into this or the numbers that we have got make any meaning or not are they correct solutions or are there embedded with some errors which is unknown to us. Therefore it is necessary to know before we actually construct numerical methods, what are errors, what are the sources of errors, do they get magnified; and if they are magnified, do they spoil the numerical solution and what is final outcome, how accurate or how reliable are the numbers that we are producing outputting from the computer.

(Refer Slide Time: 00:07:28 min)



Error = True value - approximate value

Absolute error = |error|

Relative error = $\frac{|\text{error}|}{|\text{True value}|}$

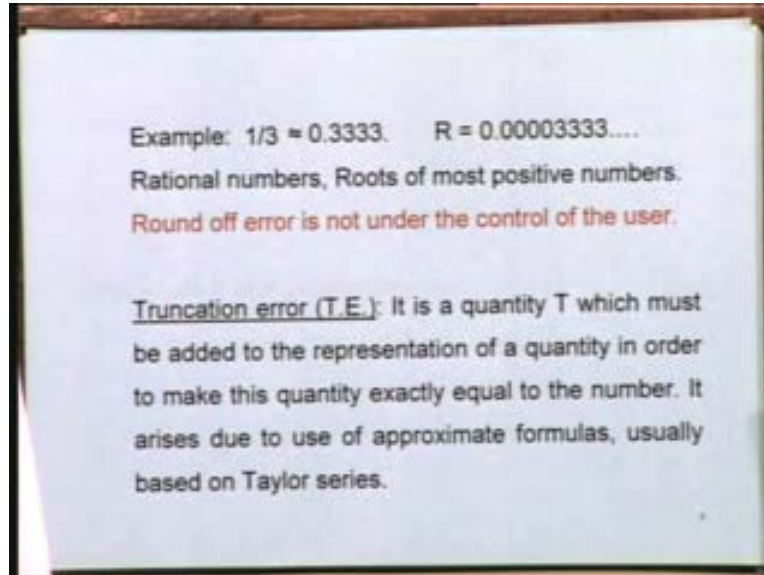
Errors in Computation:

Round off error (R.E.) in a number: It is a quantity R which must be added to the finite representation of the number in order to make it the true representation of that number.

Let us first of all define what error is and then what are the sources of errors? Now as you all know that error is true value minus approximate value. That is a simple definition of an error. However, the error could be positive or negative because the true value could be on either side of the approximate value, therefore, it could be positive or negative. We would like to study what is the absolute error, so that we can talk of the magnitude of the error and say this is the error that has been made. Very often these absolute errors could be - if the problem that we are solving has a solution which are very small or problems in which the solution is very large, and then absolute error is going to be very small or very large. Suppose the order of magnitude of a solution is 10^{-5} and you get absolute error is ten to the power of minus six, you may think that is a very good solution but it is not so because we have to talk of the magnitude of the solution also into account. Therefore a good indicator would be relative error, so that we have error divided by the true value so that it will take care of both the cases, when the solution is very small or a solution is a very large quantity.

Usually relative error is a good indicator. Then we would like to look at what are the errors in computation that would arise in general. The first thing is round off error. What is a round off error? It is a quantity R, which must be added to the finite representation of the number in order to make it the true representation of that number. That means we are given a number which we want to carry out for computations either by calculator or by the computer. So this number has to be given a finite representation and whatever is left out will be the round off error for that particular problem. Now let us take a simple example.

(Refer Slide Time: 00:08:58 min)



A simple example is you take one by three. So you may be keeping four significant digits. So we may be writing one by three as 0.333 and whatever is left out is an infinite representation. And what is left out of this is your round off error. For example, all rational numbers, the roots of most of positive numbers, all of them would give round off error. There is nothing we can do about round of error to stop it. Round off error is always there in any number. Round off error is not under the control of the user. What we mean is when we go to the computer and we are putting these numbers on the computer, the round off error is not under our control.

The other thing is the truncation error. Truncation error is a quantity T which must be added to the representation of a quantity in order to make this quantity exactly equal to that number. It arises due to use of approximate formulas, usually based on Taylor series. For example, if you want to get the value of exponential of x , for small x , i may approximate it as one plus x plus x square by factorial two. Now whatever is left out in this approximation will be the truncation error. For example, if you now look at the modern day computers most of the functions are embedded in there. All the major functions are embedded in the computer it automatically gives you. And those were evaluated by certain formulas and those formulas also have got their truncation error. So that part is called the truncation error.

(Refer Slide Time: 00:09:59 min)

The Taylor series of $f(x)$ about a point $x = x_0$ is given by

$$f(x) = f(x_0) + (x - x_0)f'(x_0) + \dots + \frac{(x - x_0)^n}{n!} f^{(n)}(x_0) + R$$

where R is the remainder

$$R = \frac{(x - x_0)^{n+1}}{(n+1)!} f^{(n+1)}(\xi), \quad x_0 < \xi < x$$

Now let us just see how this error comes about. Now if I write the Taylor series expansion of a function f_x about a point, x is equal to x_0 I can write it as $f(x) = f$ of x_0 plus $x - x_0$ f prime of x_0 and the n^{th} term plus R as a remainder, where R is the remainder given by $x - x_0$ to power of n plus one, factorial n plus one of x_i ; x_i lying between x_0 and x . There are various other forms of the remainder but however we shall use this particular remainder in our analysis later on.

Now what we were stating earlier is, if I take this Taylor series and truncate it upto certain terms then whatever I am leaving out will be the truncation error in that particular problem.

(Refer Slide Time: 00:10:59 min)

Examples:

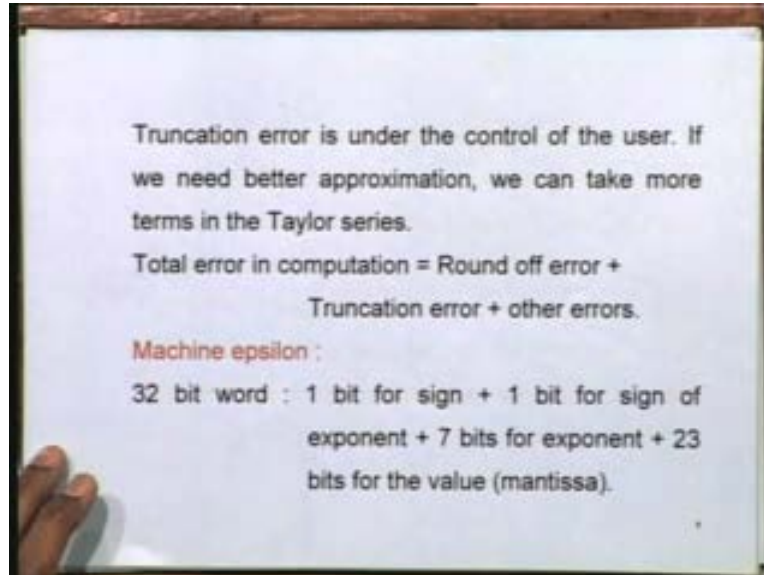
1. $\sin x \approx x - \frac{x^3}{3!}$, for small x
 If $f(x) = \sin x$, we have $f^{(4)}(x) = \sin x$
 $T.E = \frac{x^4}{4!} \sin \xi$, $0 < \xi < x$ and $|T.E| \leq \max \frac{|x|^4}{4!}$
2. $e^x \approx 1 + x + \frac{x^2}{2}$, for small x
 If $f(x) = e^x$, we have $f^{(3)}(x) = e^x$
 $T.E = \frac{x^3}{3!} e^\xi$, $0 < \xi < x$ and $|T.E| \leq \max \frac{|x|^3}{6} e^x$

Now let me take this simple example. If I take $\sin x$ and I write approximate it as x minus x to the power of three by factorial three for small x , then I would like to look at what would be truncation error in this particular problem. The truncation error, (go back to this slide), from this remainder I can find out what is the truncation error in my given problem. Therefore if I write this as this the truncation error would have the fourth derivative. Therefore, the fourth derivative is required. $f(x)$ equal to $\sin x$ gives the fourth derivative. Four times derivation gives me $\sin x$ back. Therefore truncation error is x to the power of four by factorial four into f four of x_i (i.e.) \sin of x_i . Now here you can see that we are expanding about the point 0. So this is $x - 0$, $x - 0$ whole cube by factorial three and hence here truncation error also will be x to the power of four by factorial four into \sin of x_i . Now I can always bound this by taking its magnitude. So if I write it as magnitude, I can write this maximum of magnitude of x to the power of four divided by factorial four, magnitude of $\sin x_i$ is always less than one. So I can simply write maximum of x to the power of four by factorial four. Now this maximum can be determined, when once in the problem it is given that we are using this approximation, for example x is from 0 to 0.001 or x is from 0 to .005. So when once you are given the interval in which this formula is being used you can find out its maximum, because maximum magnitude x to the power of four can immediately be determined and I can say that this is the largest truncation error that would be done while evaluating this in the given interval.

Similarly another example is if I take exponential of x (e^x) is one plus x plus x to the power of two by two again for small x , then I have retained only terms of this second order. The truncation error would contain the third order term. Therefore I need the third derivative of exponential of x (e^x) and the third derivative of exponential of x is e to power of x (e^x) and therefore the truncation error will become x to the power of three by factorial three into exponential of x_i . Again we are expanding about the point 0 therefore I will have here $x - 0$ to power of three. This and I can again bound the error, as magnitude of this is maximum of magnitude of x^3 by six into exponential of x_i . Again if once we are given the interval in which we are to evaluate using this for small x , say for example, from x is equal to 0 to one, then I can find that the maximum would be (in the interval 0 to 1), maximum of magnitude of x is 1, exponential of x_i will be 1, exponential of one, therefore this maximum will be one upon six into e to the power of one. So it is simply e by six will be the maximum bound of the error. So I can find out the bound for the errors given any particular formula and then write down its Taylor expansion and retain up to certain terms. We can always find out what will be the truncation error in our problem, that is why we put here magnitude of x . Truncation error will be $x - x$ 0 to power of four but here we have written magnitude. So it would be positive or negative.

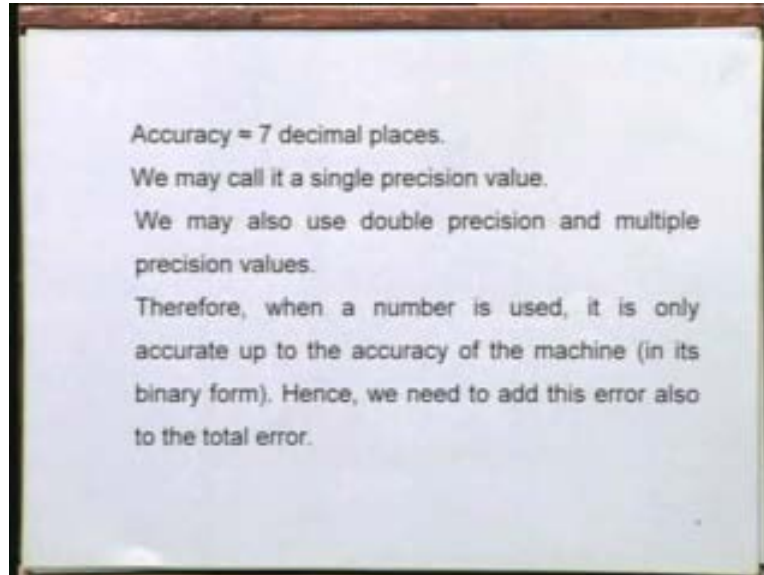
Now the next thing is - Truncation error is in the control of a user. That means truncation error is in our control. If we need better approximation you can take more terms in the Taylor series. Suppose you want an accuracy of ten to the power of minus six in your solutions we can find out the number of terms that we require in a Taylor series and then use that many terms viz., 10 terms or 11 terms so that we get the accuracy of six places. Therefore the truncation error is in the control of the user. So there is no difficulty about that one.

(Refer Slide Time: 00:15:28 min)



Now, what will be the total error in a computation? Total error in computation is the round off error plus truncation error and various other errors. We will now describe the other errors that can arise are. Besides the mistakes that we commit there will be some other errors. The most important error that would arise is the machine epsilon. Now what is machine epsilon? We know all our computers have got finite word lengths. Either we may have a thirty two bit word length or we may be having sixty four bit word length. If you take a thirty two bit word in the machine then 1 bit for sign is used, 1 bit for sign of exponent is used, 7 bits are used for exponent and 23 bits for the value (i.e.) mantissa. That is how the computer takes in. Similarly it will almost be doubled, this part will be more than doubled if you go for a 64 bit word and the reason why we are saying this is, from this we would know, when we put a number to a computer what is the accuracy that the computer is maintaining for this particular number. Now if I look what is this 23, so i can go about by taking x to the power of minus twenty three and find out what will be the approximate value for this one.

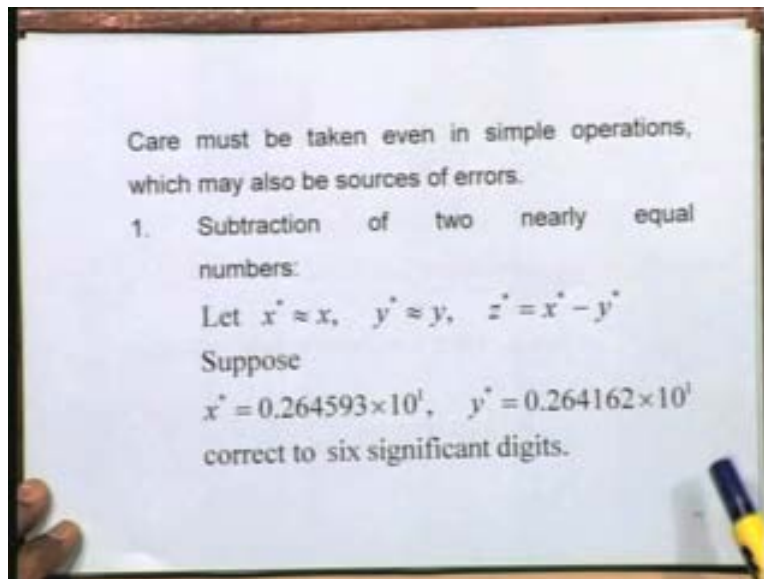
(Refer Slide Time: 00:17:47 min)



So the accuracy is two the power of minus twenty three, approximately 7 decimal place accuracy. So if I go to the computer (the 34 bit) then we can expect 7 decimal place accuracy in each number. Now we usually call this as a single precision value. If we want to have more accuracy then we can go for double precision value. We can define the double precision variable and then the double precision can be used then it will be more than double. We are using only one bit for sign, one bit for the sign of the exponent and seven bits for the exponent. Therefore if you go by 64 bit length this mantissa length is going to increase. Therefore instead of 14 decimal places, 15 to 16 decimal places accuracy will be obtained.

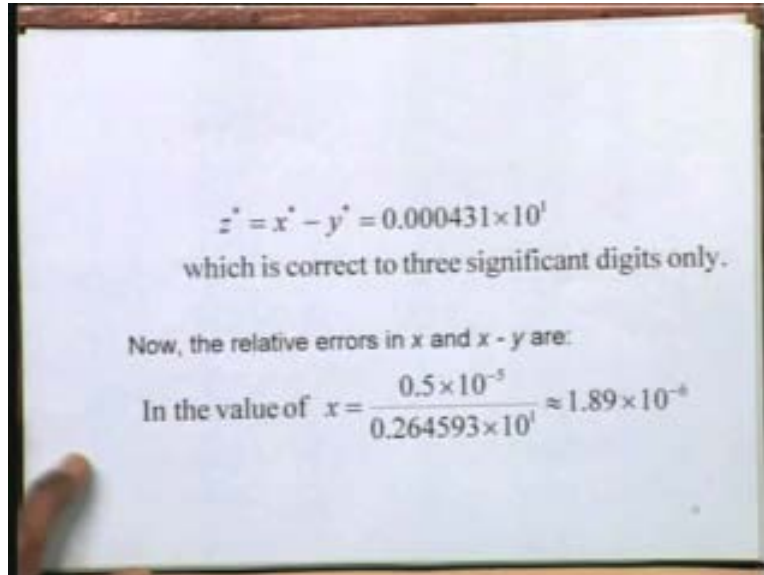
If for some problems the solution values are too small and you have to get them accurately we can go for multiple precision. We can also define the precision as four time's precision also, except that the time that is taken for computations are enormous. So the computer will allot four words for each particular number. So you can have a multiple precision also defined as this. Now therefore when a number is used it is only accurate up to the accuracy of its machine. We are taking its binary form. Therefore we need to add this error also to the total error. What we mean is, even if you are taking a number like 0.1 or 0.5 to the computer it is changed into binary form and then it is going to be cut at the 23 bits. Therefore the computer is automatically introducing an error which is the machine epsilon. So that is the accuracy of machine being included. Therefore this total error will now include the error which the computer is automatically making while converting a number into its binary form.

(Refer Slide Time: 00:18:38 min)



I also said that we should take some other errors. There are other sources of errors that could arise. Some sources of errors are some very simple calculations. Care must be taken even in simple operations which may also be source of errors. What are these simple operations? One is subtraction of two nearly equal numbers. If you are subtracting two nearly equal numbers in a computer, then we are going to have a big source of error because we are going to have loss of significant digits. Let me give it an example and then see what i mean by this. Let us say, this x and y are the exact numbers and x^* is an approximation to x , y^* is an approximation to this(y) and I want to subtract these two numbers and get the value. So what I would be getting in the computer is a z^* which is the subtracting $x^* - y^*$. Let us now take two numbers x^* , I will take to six decimal places 264593 into ten to the power one (i.e.) 2.64593 and y^* , I will take 2.64162. That is in this particular form both are connected to significant digits, because we are now talking in the normal form, that is what the computer uses. So we are using the normal form of a decimal number therefore we are putting this into exponent and ten to the power one into ten to the power one and this is corrected to six significant digits.

(Refer Slide Time: 00:20:12 min)


$$z^* = x^* - y^* = 0.000431 \times 10^1$$

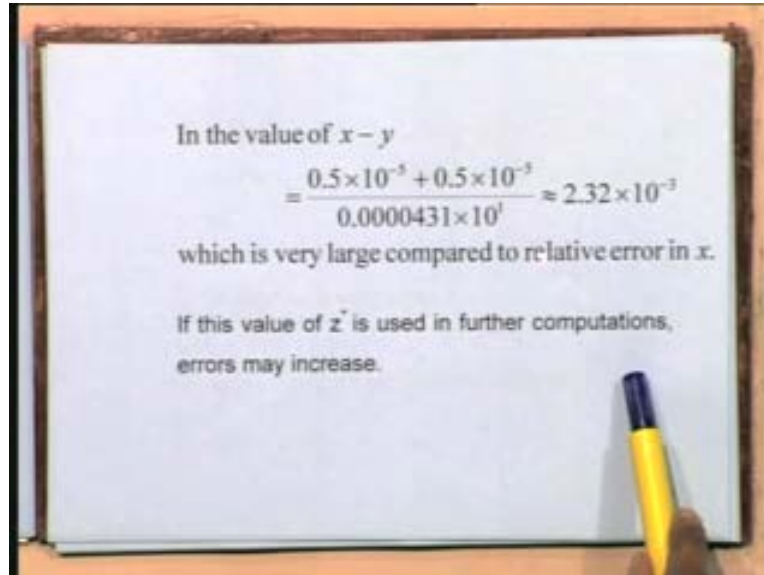
which is correct to three significant digits only.

Now, the relative errors in x and $x - y$ are:

$$\text{In the value of } x = \frac{0.5 \times 10^{-5}}{0.264593 \times 10^1} \approx 1.89 \times 10^{-6}$$

Now I want to subtract these two numbers. Now if I subtract these two numbers z^* will be $x^* - y^*$ which is 3.000431 into ten to the power one. Now when it converts to the normal form this goes as 0.031 into ten to the power minus two. This is in the usual form that we are writing. But when once it is written in the normal form it is going to be read as 0.431 into ten to the power minus two. Therefore this is correct only to 3 significant digits. Therefore we have lost 3 digits in this subtraction of these small numbers. Now how you are able to say that this is not accurate. Let us look at its relative errors. Let us look at what is the relative error in x and relative error in $x - y$ (i.e.) this number that we have got. That will give you an idea how accurate this computation has been made. Now if I take in the value of x (go back to the slide), this x^* is given these significant digits which this is 2.64593 , therefore I can immediately write down what would be the error in this. In this case of error, we are talking of the magnitude absolute value which could be on either side, so 0.5 into ten to the power of minus five comes from here and the true value is 264593 . If I divide these two then what I would get is 1.89 into ten to the power of minus six. This is the relative error in x when we have started the problem.

(Refer Slide Time: 00:22:30 min)



In the value of $x - y$

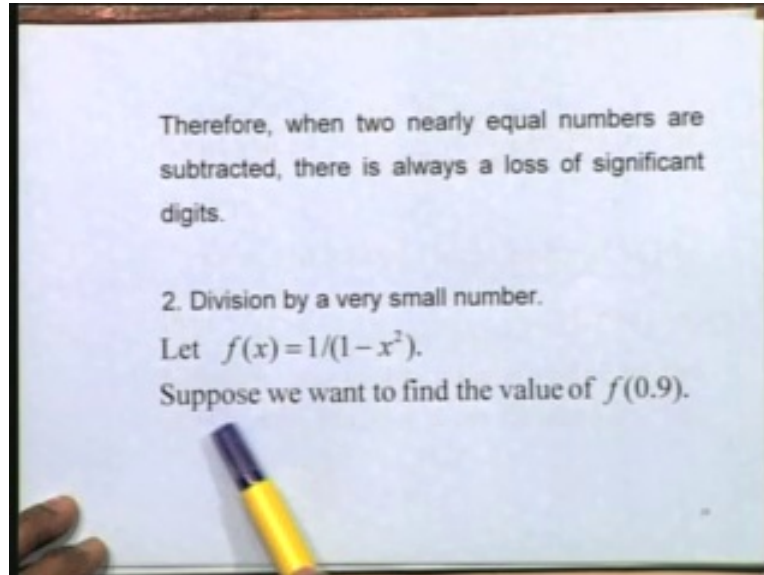
$$= \frac{0.5 \times 10^{-5} + 0.5 \times 10^{-5}}{0.0000431 \times 10^1} \approx 2.32 \times 10^{-3}$$

which is very large compared to relative error in x .

If this value of z^* is used in further computations, errors may increase.

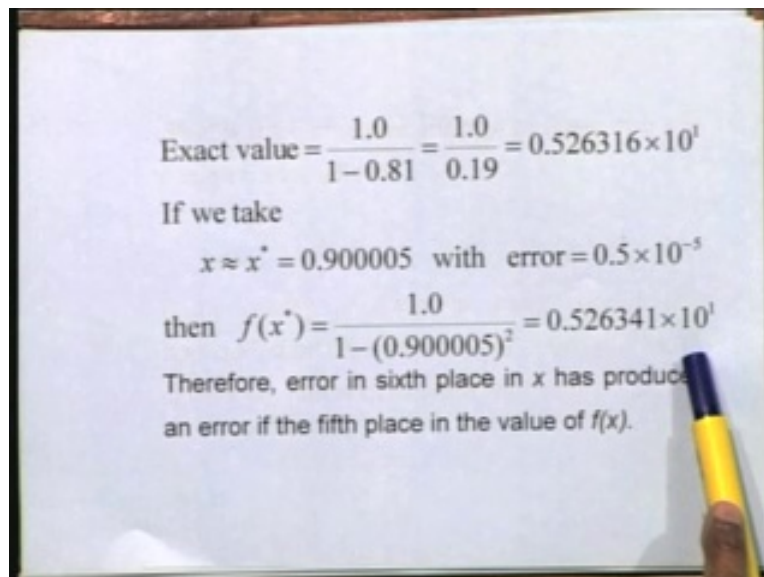
Now let us find out what is the relative error in $x - y$. Relative error in $x - y$ will be the error in x , we are talking of the absolute value, therefore a magnitude of the error in this, plus magnitude of the error in this and divided by the actual value. Now if I substitute this, we get it as 2.32 into ten to the power of minus three, which is very large compared to the relative error in x . The relative error in x was ten to the power of minus six, whereas here relative error is ten to the power of minus three. So it is now reflecting how inaccurate is this value, with this we obtain $x - y$. Normally this is very trivial and very difficult subtraction which we are doing and with this we are going to do tremendous amount of computations. Now if this value of z^* is used in further computations, error would definitely increase. Therefore this is a very simple case when loss of significant digits can occur.

(Refer Slide Time: 00:22:58 min)



The second example that comes is the division by a small number. When we divide by a very small number, and then also we are going to have tremendous amount of error in the problem. Let us take a simple way of illustrating it by taking an example. Let us suppose we want to evaluate this function $f(x)$ is equal to one upon one minus x to the power of two. Suppose we want to find the value of f at 0.9, so I would first find what its exact value is.

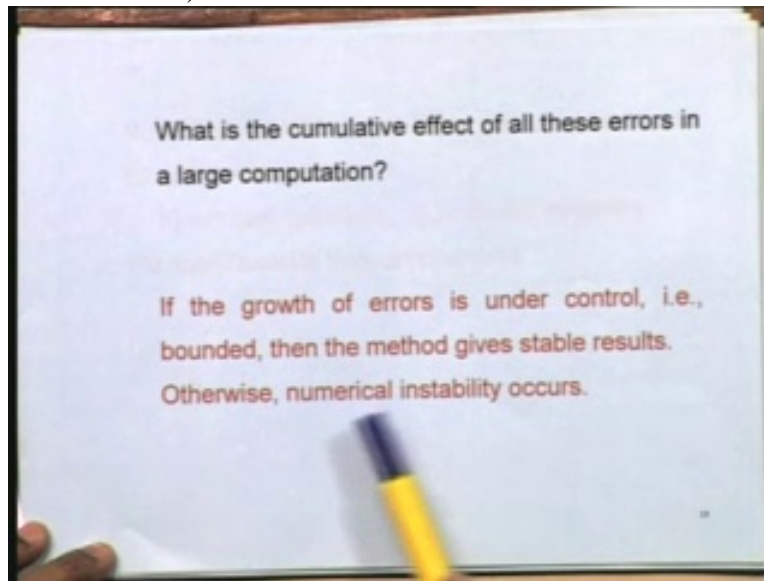
(Refer Slide Time: 00:24:07 min)



The exact value of this will be simply as one of upon one minus x to the power of two. So I compute this and write this as my exact value of the one upon one minus x to the power of two. Now let us commit an error in x , that means let us take x is equal to x^* as 0.9005 and the error is 0.5 into ten

to the power of minus five. If I now substitute this and evaluate it then I can evaluate it; f of x^* is one upon one – x to the power of two and I evaluate it 526341 into 10 to the power of one. Now the error in the sixth place has produced an error in the fifth place, which is this 5263, and 1 and here 4. We have committed error only in sixth place in x , whereas here this value function of a simple function one upon one – x to the power of two has produced an error in the fifth place. Now you can see how these errors are playing about, an accurate place up to six places has now reduced accuracy to five decimal places. Now these are being used elsewhere, so when you go on using them like this the accuracy that we have started with can go on decreasing further. Therefore when the huge computation is completed we really do not know whether, what our solution obtained is correct or not and how many decimals is correct. The computer may be outputting ten decimal places, are all 10 decimal place correct; or is it correct only upto two or three decimal places. Since we have lost so many significant digits, we do not know whether the solution obtained is correct or not.

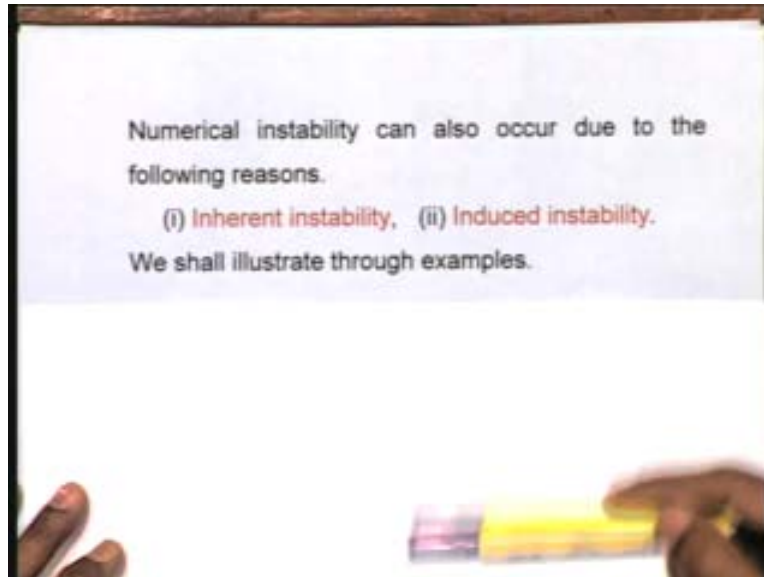
(Refer Slide Time: 00:25:08 min)



Now what is the cumulative effect of all these errors in a large computation? We are talking about the round off error, the truncation error, other sources of errors, a machine epsilon. There are so many errors that would be there in the computer. Now what is the cumulative effect of all these errors in a large computation? We will now define that, “if the growth of errors is under control (i.e.) bounded, then the method gives stable results, otherwise numerical instability occurs.” Now what do you mean by numerical instability in a computer? It gives you an overflow and it goes beyond the limit of the computer and it can be said that over flow has occurred and the computation will stop and then the output will be known to be exploded.

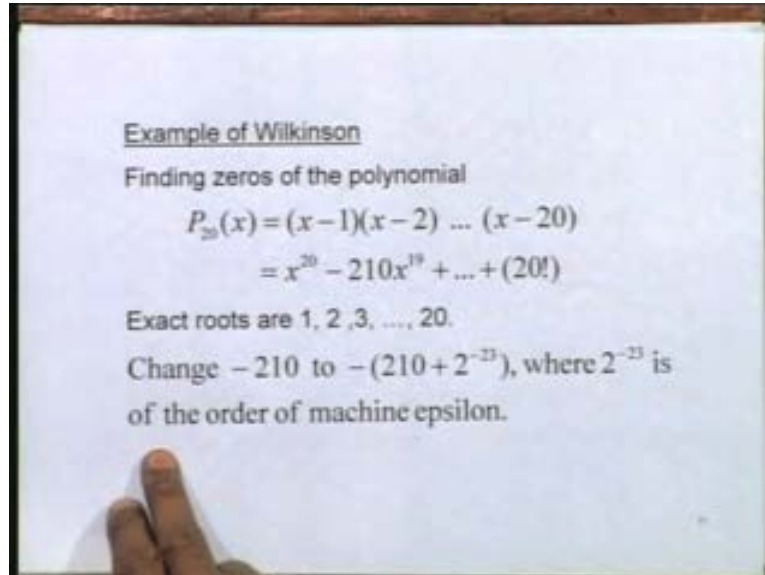
We talked about numbers. Now the next thing we would like to talk about in numerical instability is that we have a problem and we want to use a numerical method. In this case the numerical instability can occur not because of the numbers which you are using but because the problem itself was not constructed properly. Properly here means that, in our sense it may be proper, but when it has gone to the computer it has become improper or we might be using a wrong numerical method, because a problem may have many methods that we can use, of which only few methods would fit a particular problem.

(Refer Slide Time: 00:26:24 min)



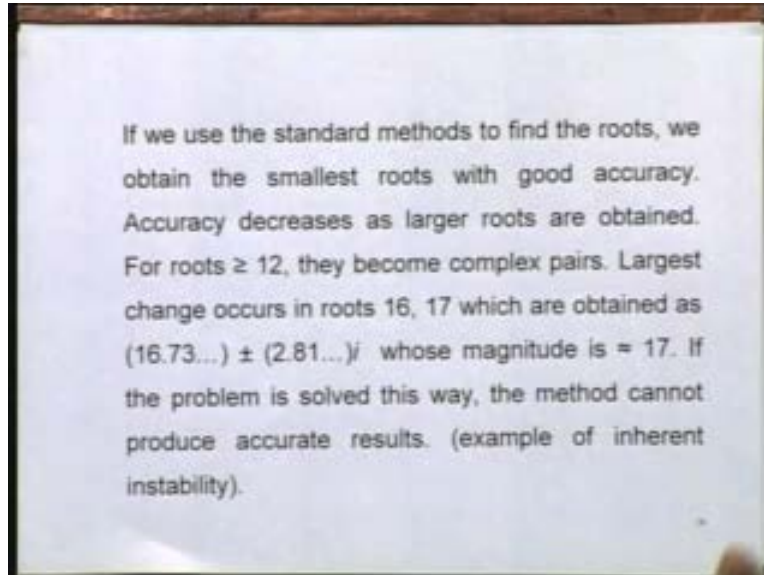
Let us take a simple example. You will be wondering how such things can happen when we go to the computer. The two instabilities that we will talk will be defined as inherent instability and induced instability. Inherent instability means, instability is inherent in the problem itself; whereas induced instability means instability induced by the user which means he is using the wrong method for solving that particular problem.

(Refer Slide Time: 00:27:24 min)



Now let me take this very famous example called Example of Wilkinson. It is finding the 0s of this polynomial. He has just taken the product of $x - 1$, $x - 2$, $x - 20$. If I expand it out, this will give you x to the power of twenty minus 210 i.e. sum of $1 + 2 + 3 + 20$ i.e. $219x$ to the power of nineteen and the constant is 1 into 2 into 20 i.e. factorial twenty. So this is the polynomial that we will have. We know its exact roots are $1, 2, 3$ and 20 when I find the roots of this particular polynomial. What we do now here is, we just change this only one coefficient of x to the power of nineteen from minus 210 to minus 210 plus 2 to the power of minus twenty three where 2 to the power of minus twenty is the order of machine epsilon. Therefore we are not committing an extra special error; we are just adding a number to it which is the accuracy of the machine i.e. 2 to the power of minus twenty as the error in the particular problem.

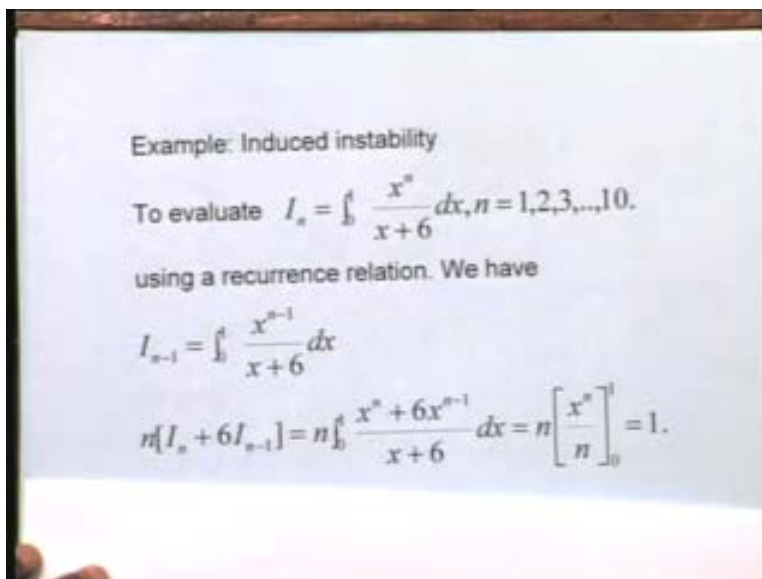
(Refer Slide Time: 00:28:06 min)



Now let us solve this problem. We shall discuss later on the number of methods used to solve this particular problem. The very interesting thing that happens is, if you solve by any method as discussed, all the smallest roots i.e. 1, 2, 3, 4, all these smallest roots are obtained with quite good accuracy. Now accuracy starts decreasing as larger roots are obtained. As we go up like 7, 8, 9, 10, for roots greater than or equal to twelve the roots become complex pairs which means they are no longer real, they are complex pairs. The largest change occurs in the roots 16 and 17 which are obtained as the complex pair. The real part is 16.733 and the imaginary part is 2.81 and so on. So this is a complex pair and this represents the roots of both of them 16 and 17; and if you take the magnitude of this it is approximately seventeen. So both these roots 16 and 17 come out as complex pairs, whose magnitude is approximately 17.

Now if you change the method from the given method it is not going to give us the solution. Therefore this is an example of inherent instability which means the problem taken was not proper. The way in which the solution adopted for finding the roots of the polynomial should be changed, so that the problem can be solved. Therefore this is an example in which the problem itself is having a difficulty.

(Refer Slide Time: 00:30:35 min)



Example: Induced instability

To evaluate $I_n = \int_0^1 \frac{x^n}{x+6} dx, n = 1, 2, 3, \dots, 10.$

using a recurrence relation. We have

$$I_{n-1} = \int_0^1 \frac{x^{n-1}}{x+6} dx$$
$$n[I_n + 6I_{n-1}] = n \int_0^1 \frac{x^n + 6x^{n-1}}{x+6} dx = n \left[\frac{x^n}{n} \right]_0^1 = 1.$$

The second example we shall give is of induced instability. Now when I say induced instability, given a problem it is not that we just pick up a method and try to solve that particular problem, we may have to see whether this method is suitable for solving that particular problem.

Let us see a very simple thing. We want to evaluate this integral, 0 to 1, x to the power of n , $x + 6$ for $n = 1, 2, 3 \dots, 10$. So that I_{10} is 0 to 1, integral x to the power of ten divided by $x + 6$, dx of x . I want to use a recurrence relation for this; so I build a recurrence relation for this so that I can evaluate this particular integral. Now I_{n-1} is 0 to 1, x to the power of $n - 1$, $x + 6$ dx . What I would do is I will add $I_n +$ six times I_{n-1} into n . I have been doing this so that I can integrate this easily. So the numerator becomes x to the power of n , six times x to the power of $n - 1$ by $x + 6$, dx . Now I can take out x to the power of $n - 1$ common here, so that $x + 6$ cancels here. So what is left out is x to the power of $n - 1$ of dx , therefore I can integrate x to the power of n by n zero to one, the value of this is equal to 1. You can see why we have multiplied by 6. I want to remove the denominator from this so that I can exactly integrate this as this. So I take this recurrence and I get this and get the value. Therefore I have the recurrence relation n times $I_n + 6$ times $I_{n-1} = 1$.

(Refer Slide Time: 00:32:16 min)

Now, to find the value of the integral, let us use the relation

$$I_n = \frac{1}{n} - 6I_{n-1} \quad \text{with}$$

$$I_0 = \int_0^1 \frac{dx}{x+6} = [\ln(x+6)]_0^1 = \ln\left[\frac{7}{6}\right] = 0.15415.$$

We obtain

$$I_1 = 1 - 6I_0 = 0.07510, I_2 = 0.04940, I_3 = 0.03693,$$

Now I want to use this recurrence relation starting with I_0 and so on get the value of I_{10} . So what I would first do, I can take this n to this side, 1 upon n and take this 6 to I_n - into the right hand side. So I simply have $I_n = 1$ upon $n-6$ and $n-1$ as a recurrence relation with I_0 . I_0 can be integrated exactly. So I can write down I_0 is 0 to $1dx$ upon $x+6$, I integrate it as logarithm of 7 by 6 and retain five decimal places 1.45 . So I want to get the values of the integral for n $1, 2, 3, \dots, 10$ using this recurrence relation with this starting value. Now I set $n = 1$, so I get I_1 is $1 - 6I_0$, so I substitute it here and get 07510 and I_2 will be equal to 1 by $2 - 6$ times I_1 . Now I_1 has been obtained so I can get I_2 is 04940 . I get the values of I_3 as this.

(Refer Slide Time: 00:35:12 min)

$\dots, I_8 = -1.09972, I_9 = 6.70943, I_{10} = -40.15658$

But $I_n \rightarrow 0$ as $n \rightarrow \infty$.

The solution oscillates and explodes. Instability occurs.

Now, use the relation $I_{n-1} = \frac{1}{6} \left[\frac{1}{n} - I_n \right]$

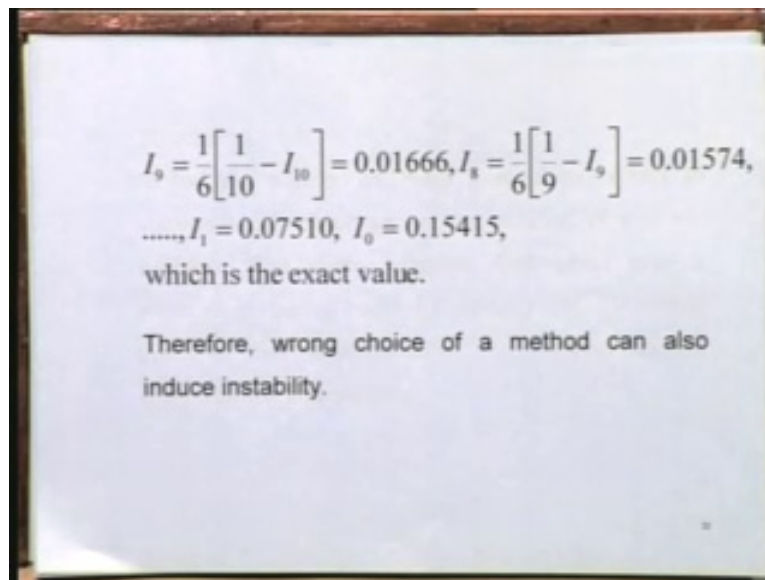
Since $I_n \rightarrow 0$ as $n \rightarrow \infty$, approximate $I_{10} = 0$.

We get

I proceed further and I land up finally at the value $I_8 = -100972$ and I_9 comes out to be 6.70943 and I_{10} comes out as -40.15658 . Let us just go back to the integral once more. If you look at the integral that is given to us, n tends to infinity. As n tends to infinity, we are integrating from 0 to 1, therefore x to the power of n is going to be 0; in other words as n tends to infinity; the value the integral I_n has to tend to 0. So we know from the given integral that I_n should tend to 0 as n tends to infinity. Now here the solutions, as we go up to I_{10} it has started oscillating between negative positive and negative positive. So they are oscillating and exploding. So it has already reached 40 and in the next step it is going to be beyond 150 and then few hundreds. So when we say that the errors are coming and it is going to explode this is what we mean. Explosion is if you now compute really up to I_{20} , I_{25} you will find that the number goes out of bound and it will give you an overflow on the computer. Therefore instability has occurred in such trivial and very simple example.

Now we shall say that this is a wrong choice of solving the recurrence relation that we have taken. So what we would here is, this recurrence relation, we will write it in the reverse form and write it as $I_n - 1$ in terms of I_n . So I will write down $I_n - 1$ is 1 upon 6, 1 upon n -, which means I would like to work backwards, the reason being I_n tends to 0; as n tends to infinity it is a decreasing value. So I will boldly commit an error here and say since I_n tends to 0, n tends to infinity let me approximate I_{10} as 0. Other numbers can also be taken but let us take 0 and see how it is going affect our solution.

(Refer Slide Time: 00:36:10 min)



$$I_9 = \frac{1}{6} \left[\frac{1}{10} - I_{10} \right] = 0.01666, I_8 = \frac{1}{6} \left[\frac{1}{9} - I_9 \right] = 0.01574,$$

$$\dots, I_1 = 0.07510, I_0 = 0.15415,$$

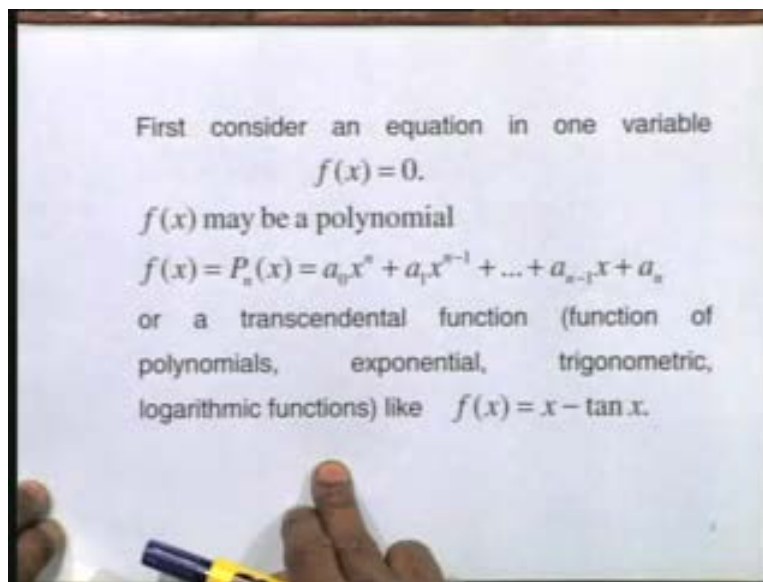
which is the exact value.

Therefore, wrong choice of a method can also induce instability.

So I can now work backwards. I_9 will be 1 by 6, 1 by $10 - I_{10}$. So this was taken as 0, this is simply 1 upon 60, that is 0.01666. Then I work backwards I_8 is equal to 1 by 6, 1 by $9 - I_9$; this is substituted here and I get 0.015174. And now I work backwards I get I_1 is equal to 0.07510. I_0 is 0.1545, which is the exact value. Here when we work backwards I have got exact value of I_0 . Therefore this would tell me that, if you are solving even a simple problem one should not take it

as an obvious problem and when we are putting it in the computer and going to solve the problem it is the correct choice or correct way of writing the problem that gives you the solution. Therefore wrong choice of a method can also induce instability. In fact many of you would be doing your projects later on, wherein you have to use a lot of numerical methods. These are some of the serious difficulties that you would encounter there, as in, you produce a solution and say that it does not match with what I was expecting and expecting a different graph of the solution. Therefore you would have to look at the various aspects of the errors that can arise in a problem including a proper method that is being used or not.

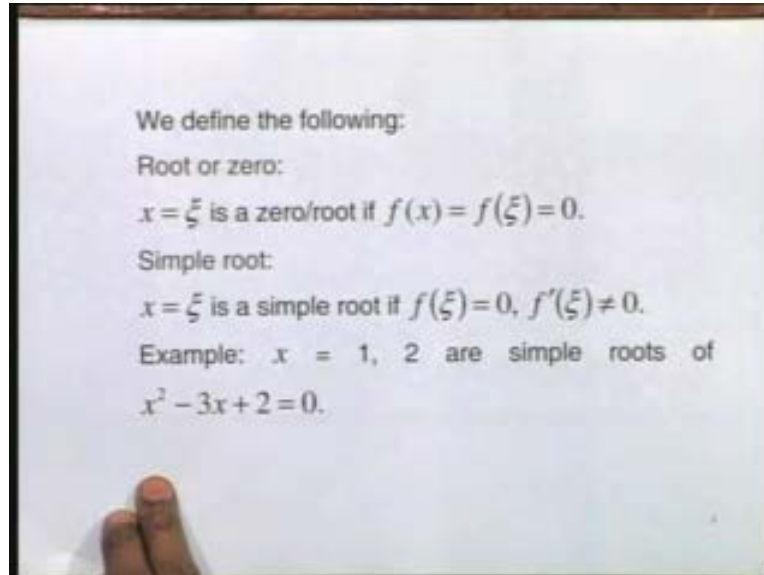
(Refer Slide Time: 00:37:57 min)



The first topic we would like to study is the Solution of Nonlinear Algebraic Equations. Now when we talk of the solution of the nonlinear algebraic equations we would like to first consider an equation in one variable. Later on we shall take system of nonlinear algebraic equations. Let us take the equation as $fx = 0$. Now here this fx may be a polynomial it may be taking polynomial that is a_0x to the power of n , a_1x to the power of n minus one, $a_{n-1}x + a_n$. So it could be a polynomial for which we are finding the result. For example, we earlier talked of finding the roots of a polynomial. So we are now talking of a polynomial or we may be talking of a transcendental function. fx could be a transcendental function, that means it will be a function of combination of polynomials, exponential functions, trigonometric functions, logarithms functions. A combination of all these functions will be called as a transcendental function.

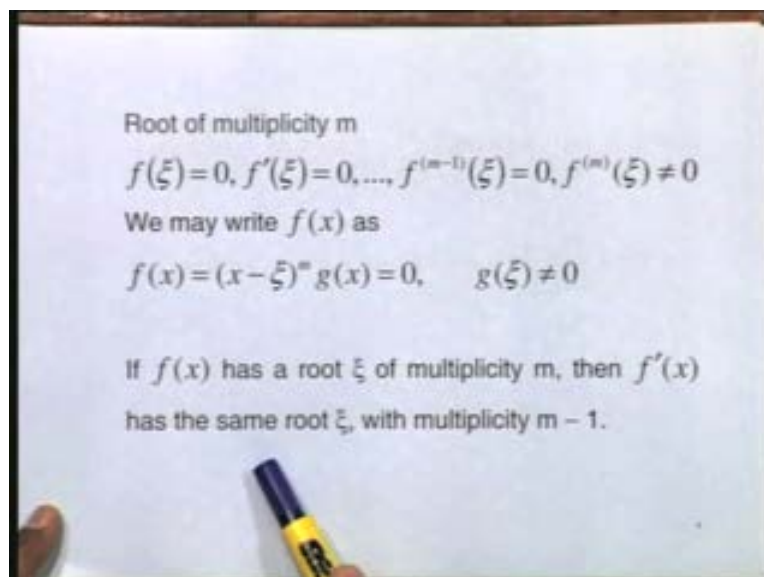
For example $fx = x - \tan x$. It is a transcendental function. It is a combination of polynomial and trigonometric function. So any combination of this shall be called as a transcendental function and otherwise we will be considering a polynomial.

(Refer Slide Time: 00:38:57 min)



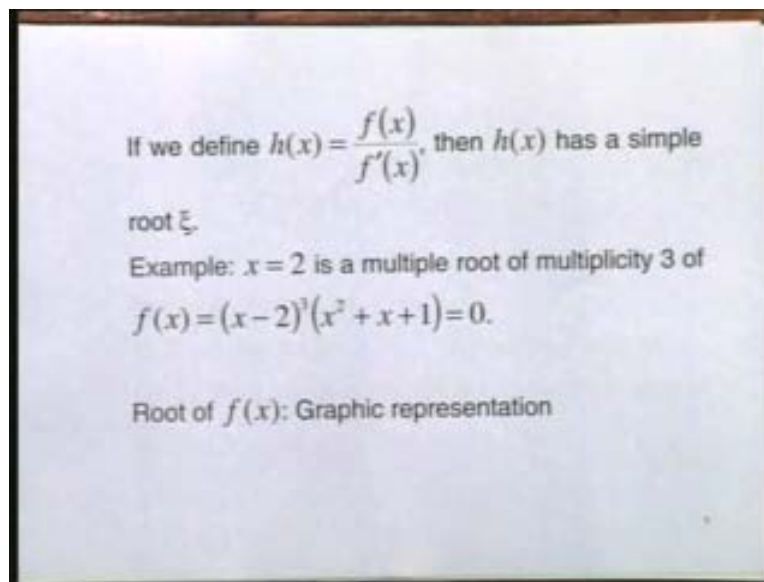
Now we define some of the concepts for the finding the roots of the polynomial. So we would like to first define a root or 0 which you know already but let us just revise what we mean by that. We shall take $x = x_i$, or an exact root or a 0. If $f(x)$ is equal to $f(x_i)$ which is equal to 0, so we will use both the words 0 or root for an equation. Now this root will be a simple root. Mathematically $f(x_i)$ is 0 but $f'(x_i)$ is not equal to 0. Its first derivative at that value is not equal to 0, therefore $x = x_i$ is a simple root. That means if you are factorizing the given polynomial or a given function, you can take out the factor $x - x_i$ as a simple factor. For example $x = 1$ and 2 are the simple roots of the equation, $x^2 - 3x + 2 = 0$.

(Refer Slide Time: 00:40:19 min)



Now we would like to talk of the root of the multiplicity m . So it is a multiple root of multiplicity m . So we define that, f of x_i is 0, f' of x_i is 0, $m-1^{\text{th}}$ derivative of f at x_i is also 0, but n^{th} derivative at x_i is not equal to 0. Therefore it will be a root of multiplicity m , therefore the derivatives upto $m - 1$ will be 0, n^{th} derivative will be equal to 0. In some cases we may be able to write $f(x)$ is equal to $(x - x_i)^m g(x)$, $g(x_i) \neq 0$. However we can also look at this same problem in a slightly different way. We know that, if a function $f(x)$ has a root x_i of multiplicity m , then it is the derivative $f'(x)$ has the same root x_i with multiplicity $m - 1$. So the derivative has got multiplicity $m - 1$. Similarly if I go further, the second derivative $f''(x)$ will have multiplicity of $m - 2$.

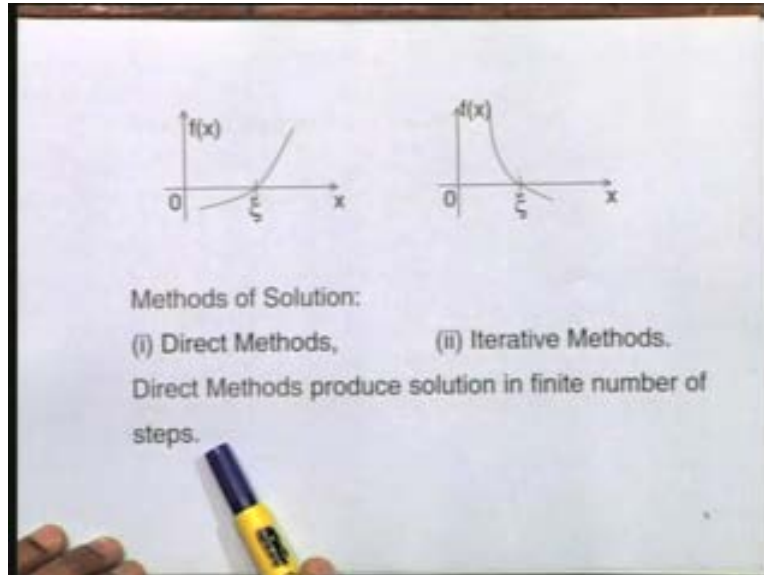
(Refer Slide Time: 00:41:37 min)



Now what I do is I will take ratio of the $f(x)$ and $f'(x)$ and then construct a new function, so that I can now define $h(x)$ as the ratio of $f(x)$ by $f'(x)$ and then $h(x)$ as a simple root x_i . Therefore we are now suggesting a method that, if I have a method for finding a simple root I can use the same method for obtaining the multiple root also. But if I use a new function $h(x) = f(x) \text{ upon } f'(x)$. This means if I know that a particular function has a multiple root and I do not know its multiplicity, then I would take the ratio of $f(x)$ by $f'(x)$. $f(x)$ by $f'(x)$ has to have only a simple root; therefore I can use the method that we are trying now to use for finding a simple root or also for the problem where it has a multiple root in which multiplicity is not known for us.

Now for example, $x = 2$ is a multiple root of multiplicity 3 for this particular equation. Before we talk of what is the method that we would like to construct, let us see what we mean by the root of a function $f(x)$. Let us just look at the graphic representation through which, we will be able to pictorially represent any method that we construct, give the meaning of a numerical method, what a numerical method is doing and how it is getting a root of this particular problem.

(Refer Slide Time: 00:42:48 min)



So if I just plot the graph of y is equal to fx , take x and then along the y axis we take fx , the graph cuts the x axis at a point and that is my exact root i.e. $fx = 0$. So $y = fx = 0$; therefore $y = 0$. So the graph is cutting the x axis at a point and that is the root. So what we are trying to find out is, to determine what my x_i is. Now in the methods for solution for finding the roots of simple root or multiple roots there are two types of methods. One is called the direct methods the other is the iterative methods. The direct methods produce solution in finite number of steps. So we are now trying to distinguish the direct method and iterative method.

(Refer Slide Time: 00:45:05 min)

Example: Roots of $a_0x^2 + a_1x + a_2 = 0$

$$x = \frac{1}{2a_0} \left[-a_1 \pm \sqrt{a_1^2 - 4a_0a_2} \right]$$

Operational count is possible.

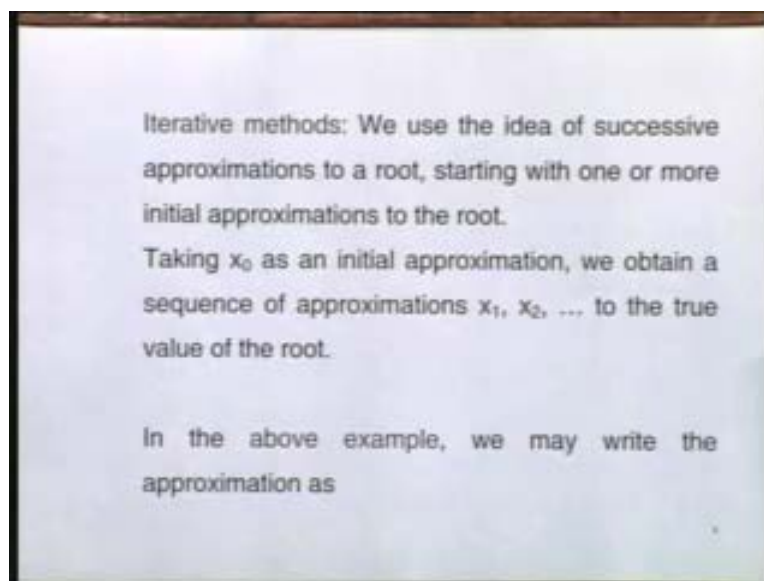
Multiplications: 3,	Divisions: 1,
Add/Sub: 2,	Square root: 1,

There are no powerful, direct methods available.

Let us take a very simple example. I want the roots of this simple quadratic. $a_0x^2 + a_1x + a_2 = 0$. So I can write down that x is equal to our standard formula. I can write down this formula and say these are the two roots. I would say that this is a direct method; I am getting both the roots x_1 and x_2 by just finding this particular value, finding its root and then getting this value. Now in a direct method the operational count is possible. Operational count is the total number of multiplications, divisions, additions and subtractions that are there in a problem; thereby you would be able to really say what will be the computed time taken by this particular method. Since we know the speed of the machine we can immediately count the major operations which are the multiplication and division, and the minor operation which are addition and subtraction. Therefore we can compute the whole time that is taken and then say this particular problem will take this much of time, so therefore operational count is possible.

For example here I can count what are the operations involved. I have two multiplications here a_0a_2 . Two multiplications are there. I can take a_1^2 as a multiplication which is a_1 into a_1 . So I can take this as a multiplication, so I can say there are three multiplications. There is one division. Here is one division here. And there is one addition here and one subtraction over here; and also one addition/subtraction. So I have got 2 additions and subtraction and 1 square root. So I now know this particular formula has this operational count and I would be able to say what will be the exact amount of computed time taken; whether it will take few micro seconds or few seconds. Therefore in a direct method always it is always possible for us to give an operational count. But there are no powerful direct methods available at all for us to solve a difficult problem of even single nonlinear equations, leave alone a system of nonlinear equations.

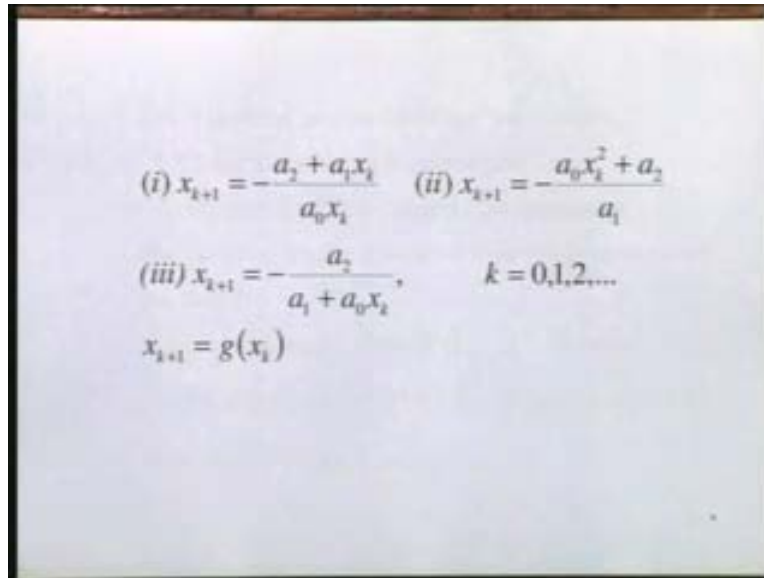
(Refer Slide Time: 00:46:11 min)



Now let us define what an iterative method is. In the iterative methods we use the idea of successive approximations to a root, starting with one or more initial approximation to a root. So what we mean by this is that we make an initial guess of the root and then try to refine it using a numerical method and refine it further, so that we will approach the exact root of this. We shall call this as an iterative method; which means taking x_0 as an initial approximation; we obtain a

sequence of approximations x_1, x_2, \dots to the true value of the root. Now let us go back and try to write down this example that we have now written a_0x to the power of two $+ a_1x + a_2 = 0$.

(Refer Slide Time: 00:47:56 min)



$$\begin{aligned}
 (i) \quad x_{k+1} &= -\frac{a_2 + a_1x_k}{a_0x_k} & (ii) \quad x_{k+1} &= -\frac{a_0x_k^2 + a_2}{a_1} \\
 (iii) \quad x_{k+1} &= -\frac{a_2}{a_1 + a_0x_k}, & k &= 0, 1, 2, \dots \\
 x_{k+1} &= g(x_k)
 \end{aligned}$$

Let us try to write down iterative method for obtaining the solution of this problem. Now there are three formulas I had given. If I take the first formula, what I have done here is I have taken these two terms to the right hand side, divided it by one of this x here and then retained this as x on the left hand side. So I will have here $-a_2 - a_1x_k$ divided by a_0x_k , so I would take this as an iteration formula; whether it is going to converge or diverge we will see later on. But this is an iteration formula starting with an initial approximation x_0 I can go and find x_1 , substitute it here x_1 , so iteratively I can get the sequence of approximations x_0, x_1, x_2, x_3 and so on.

Alternatively I can write another formula here in which I have taken the first and third terms to the right hand side and I have retained the middle term on the left hand side, so that I can divide by a_1 and write $x_k + 1 = -a_0x_k$ to the power of two $+ a_2$ divided by a_1 , so this is also another iterative formula that we can construct for this. Then thirdly if I take this particular term to the right hand side, retain these two terms on the left hand side and then I will take the common factor x here and then divide by this 1; so we will have $x_k + 1 - a_2$ upon $a_1 + a_0x_k$. Now all these formulas are of the form $x_k + 1$ is equal some function of $g(x_k)$. Therefore it is possible for us to have a large number of iterative methods. If given a problem it is possible for us to construct any number of iterative methods.

To summarize we have seen what a numerical method is, we have also defined what is error of approximation, what are the sources of errors, how serious or how important is the cumulative effect of all these errors. We have also defined what a nonlinear equation, single nonlinear equation is and also defined what a direct method iterative method is.