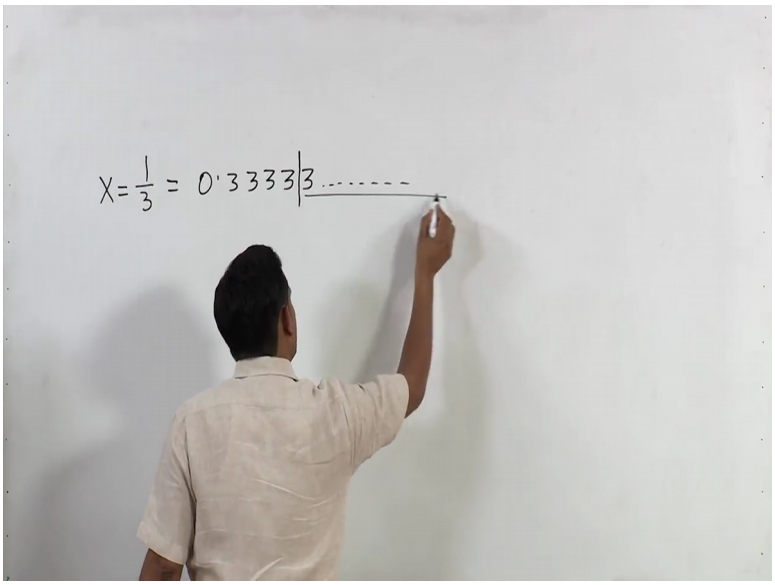
Numerical Methods
Dr. Sanjeev Kumar
Department of Mathematics
Indian Institute of Technology Roorkee
Lecture No 1
Introduction to Error Analysis and Linear Systems

Hello friends so this is the first lecture of this course, numerical methods and in this lecture I will talk about errors in numerical computations and in the later part of the lecture I will introduce system of linear equations. So first of all let us talk about error, so in the concept of error in numerical analysis we are having a term significant digits and that play a very important role or the accuracy of any numerical computation.

(Refer Slide Time: 1:02)

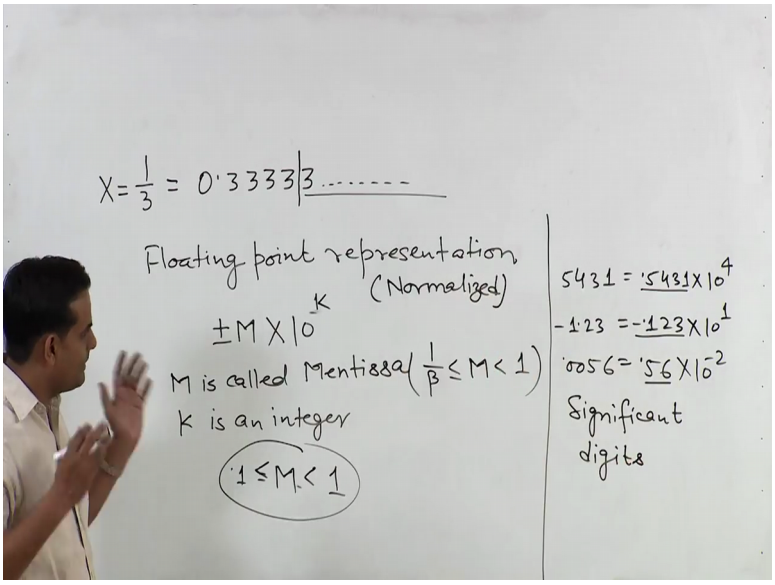


So for example suppose I want to perform a numerical computation with a real number let us say x equal to 1 by 3, however the computers generally stores the finite number of digits, so every computer forms the computation by taking finite number of digits. However we see this particular number, this particular number in decimal format can be written as like this and having the infinite number of digits after decimal, so the computer cannot perform the computation with such infinite long string of digits.

So what happens we need to cut down this number somewhere after some fixed decimal places, so for example if I cut down this number after 4 decimal places, so this portion after this particular thing will be ignored and hence we are introducing an error in our computation. If you see here, here this is a very small error up to fifth place of decimal however in for the computations this small error propagates and becomes a large error. So

now how to do it? Before that let us learn how the computer stores the number, so in computer each number can be stored by a fixed length, so fixed length means the fixed number of digits and it depends on computer that how many digits we can store in the computer, different computers can have different ability to store the number and to perform the computation with different number of digits.

(Refer Slide Time: 3:18)

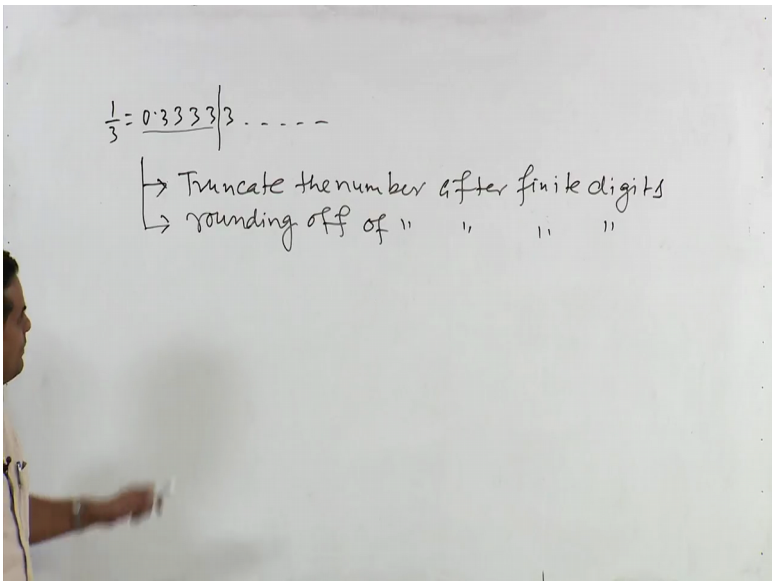


But in general we are having the floating point representation which generally each computer follow and in this representation each number can be written as plus minus M into 10 raised to power K if am talking about decimal number means numbers having best 10. Here this M is called mentissa and K is an integer. Now the range of this M will depend on the base of the number so beta is the base of representation in this will be… For example in decimal representation the range of M will be 0.1 M up to 1. In fact this particular representation is called normalised floating point representation.

So for example if I am having a number let us say 5431, so in floating point form this particular number can be written as 0.5431 into 10 raise to power 4. Similarly if I am having a number minus 1.23 then this particular number can be written as minus 0.123 into 10 raise to power 1, so here you can see the Mentissa 0.5431, here Mentissa 0.123 and each one is coming in this range and then K is 4 here and K is 1 here. Suppose I am having a number 0.0056 so here you can see that 0.0056 is less than 0.1 so what will happen I will write this number as 0.56 into 10 raise to power minus 2, so this is the floating point representation of a number in the computers and digit in mantissa are called significant digits.

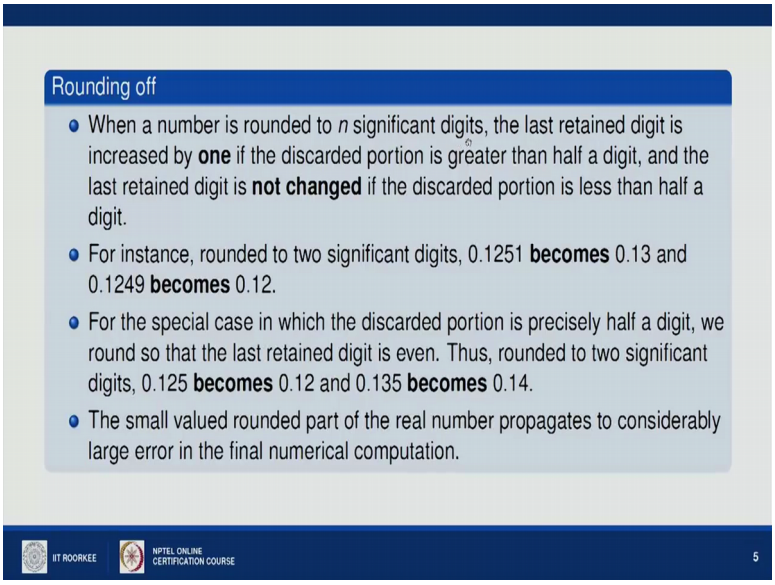
So if someone will give you a number and ask you tell me how many significant digit we are having in this number, so what you need to do, you need to write the number in the floating point representation and then the number of digits in the mantissa are called significant digits in the given number. Now as I told you that computer can perform the computation with finite number of digits, so after the finite digits, if the number is having more digits what we need to do? We have to cut down those digits.

(Refer Slide Time: 7:20)



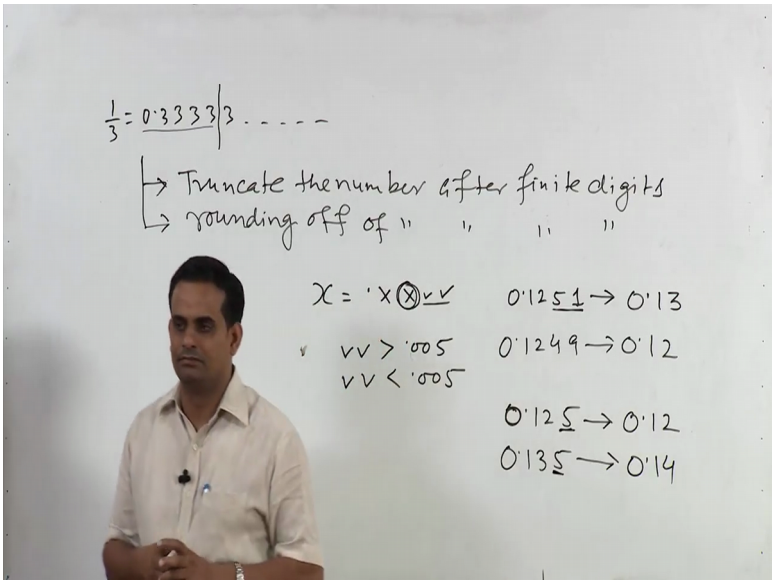
As I have given an example of 1 by 3 that is 0.3333 and so on and suppose I have taken only this portion now what will happen, so there will be some we are introducing some error in our competition by ignoring this part of the number, so this can be done in 2 way one is either we truncate the number after finite digit or we perform rounding of the number, so here truncation means that we will take this much term and after this whatever we are having we leave out as such, however rounding off each having a different procedure for cutting down a number after some finite digits.

(Refer Slide Time: 8:47)



So let me tell you how will perform rounding off, so when a number is rounded to n significant digits, the last retain digit is increased by 1 if the discarded portion is greater than half a digit in the last retain digit is not change if the discarded portion is less than half a digit. Or instance are rounded to 2 significant digit the number 0.1251 becomes 0.13 because please note that here the 51 is greater than 50 means this is 0.0051 is greater than 0.005, so what will happen if I want to rounded off this digit up to 2 significant digits after the decimal if I want to round off this particular number and it will become 0.13. Similarly if you take 0.1249 it becomes 0.12.

(Refer Slide Time: 9:45)

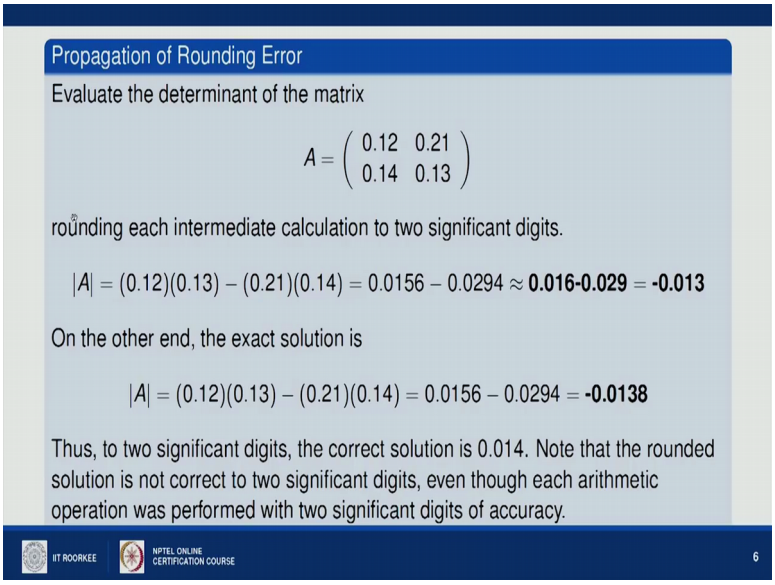


So what is happening, if I am having a number x which is having some digits let us say like this, so let us say 4 digits and I want to round it off this number up to 2 significant digits after decimal, so what will happen? I will see this particular portion, so if this particular portion is greater than 0.005 then what will happen in this particular number, this particular digit will be increased by 1, if it is less than 0.005 then this particular digit written as such no change in this digit, so for example I have taken 0.1251 so here 0.0051, so this particular digit is greater than half so after rounding off to 2 significant digits it will become 0.13 if I take 0.1249 it will become 0.12.

What will happen if it is exactly 0.125 and I want to round off it up to 2 significant digits? So what will happen this will be written 0.12 but suppose I am having a number 0.135 and I want to round off this up to 2 significant digits, it will become 0.14 why I am doing this? This number is same but here I am not increasing this digit by one but here I am increasing, this because if this is the case that in the last you are having exactly half of the number so what will happen we will round off the number in such a way that the last digit should be an even digit for a sample should be an even number example here the 0.125.

So 2 is an even, year 0.135 here 4 is an even, so I want to make this even so for making this even I want to increase it by 1. So this procedure is called rounding off a number of to given significant digits and this particular thing play a very important role in numerical computation and basically here we are introducing a very small error in the number by rounding off or truncated up to some finite number of digit at later on in the computation it propagates and become a large error.

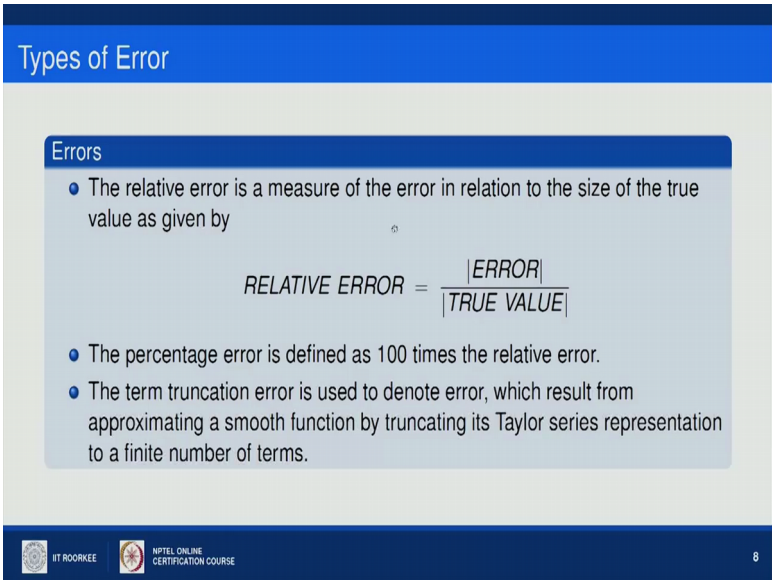
(Refer Slide Time: 13:05)



Let us take an example of that, so consider this particular matrix and for this matrix and I want to find out determinate of this by rounding each intermediate calculation up to 2 significant digits, so I have taken this matrix let us calculate the determinant so 0.12 into 0.13 minus 0.21 into 0.14, so this particular number becomes 0.0156 and this particular number becomes 0.0294. Now as I told you I need to round off each intermediate calculation to 2 significant digit, so when I rounding off it, it will become 0.016 because here 6 is half of the digit and so 0.016 this will remain as 0.029 and the final determinant is minus 0.013.

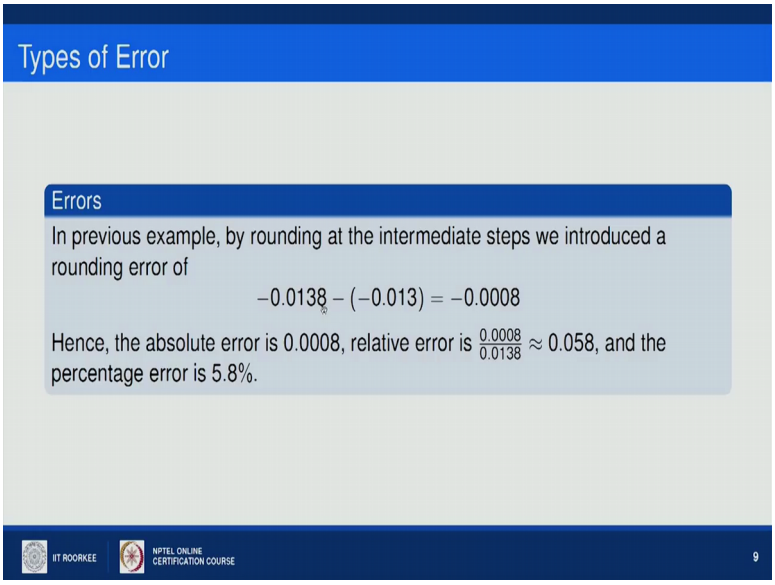
On the other hand the exact solution is this one minus this one and it is 0.0138. Now if I round off the final result up to 2 significant digits it will become minus 0.014. So you can see that up to 2 significant digits the correct solution is minus 0.014 wherever if I am performing in the intermediate calculation up to 2 significant digits, it is coming minus 0.013 so we are having a significant error in the calculation where I am rounding off each intermediate calculation up to 2 significant digits. Now what type of error we are having in numerical computation, so the 1st error can be defined as true value minus approximate value, approximate value means which we are calculating using the numerical methods, so if I am getting like the true value exact solution so error will be 0.

(Refer Slide Time: 15:28)



Now absolute error is called absolute value of the error, the relative error is a measure of the error in the relation of the size of the true value and it is given as the absolute value of the error that this is the absolute error upon absolute true value, the percentage error is defined by multiplying by 100 to the relative error or 100 times of the relative error. The term truncation error is used to denote error which results from approximating a smooth function by truncating its Taylor series representation to a finite number of terms.

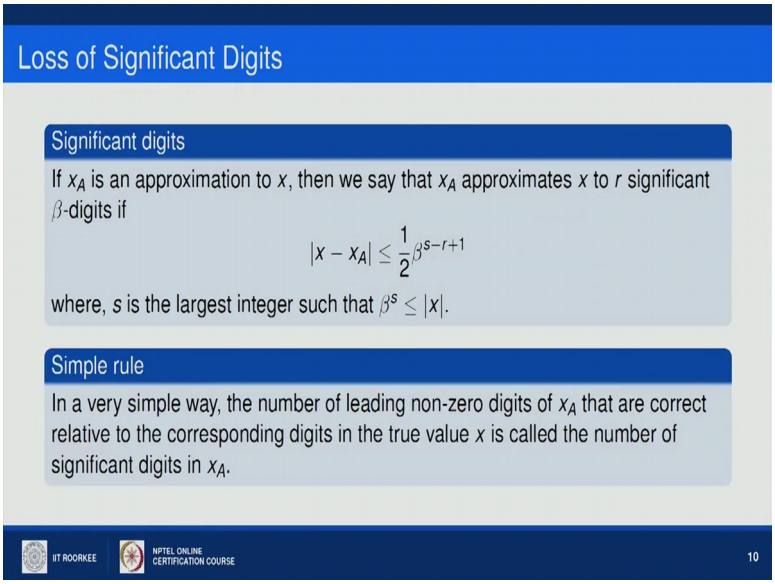
(Refer Slide Time: 16:07)



So if we take this particular example here the true value is minus 0.138 while the approximate solution which we are getting by rounding off each intermediate calculation up to 2 significant digits is minus 0.13, so the error here is minus 0.0008 absolute error is 0.0008

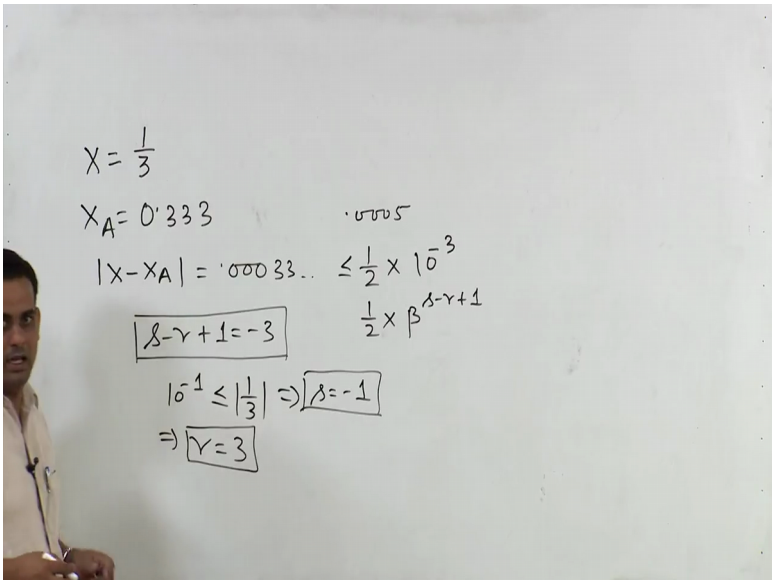
the relative error is 0.0008 upon 0.0138 and it becomes 0.058. It is looking small but when we see the percentage error it is 5.8 because it is 100 times of this particular number and which is now by looking at the error 5 percent error more than 5 percent error is quite significant. Now let us talk about significant digits in the approximate solution of approximation of a number.

(Refer Slide Time: 17:19)



So let x A be an approximation of a number x or when we say that x A approximate x to are significant beta digits where beta is the base of the number of, if x minus x A and the absolute value of this difference is less than equal to half-time beta rays to power s minus r plus 1, where x is the largest integer such that beta raise to power is less than equal to mode of x. So let us take some example of it.

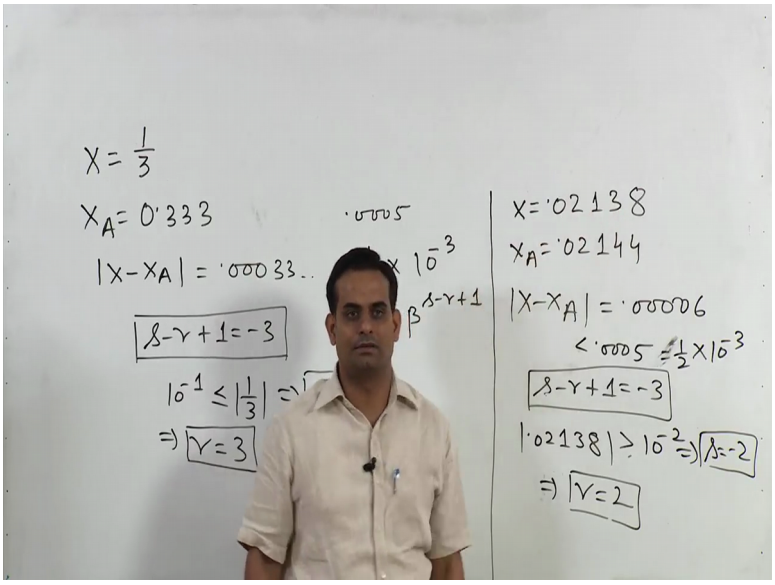
(Refer Slide Time: 17:49)



So let us say the true value is 1 by 3 and the approximate solution is 0.333, now I want to check up to how many significant digits this approximate solution is true for the means match with the exact solution, so first of all I will calculate x minus x A and it will be something 0.0033 and so on. Now if I check this this is less than equals to basically 0.0005 so 0.5 into 10 raise to power minus 3 and as I told you it is by the given definition so beta is the base so base is 10 here. Now s minus r plus 1 equals to minus 3, I got a relation.

Now I want to find out the value of s as a told you s is the largest integer such that beta raise to power s less than absolute value of true solutions. So here moreover we can have that 10 raise to power minus 1 there is one upon 10.1 which is less than 1 by 3, the true solution so this gives me as is equal to minus 1. So from these 2 equations what I can write r equals to 3, so hence I can say that the approximation 0.333 is having 3 significant digits to the exact value 1 upon 3 or match with the exact value with 3 significant digits.

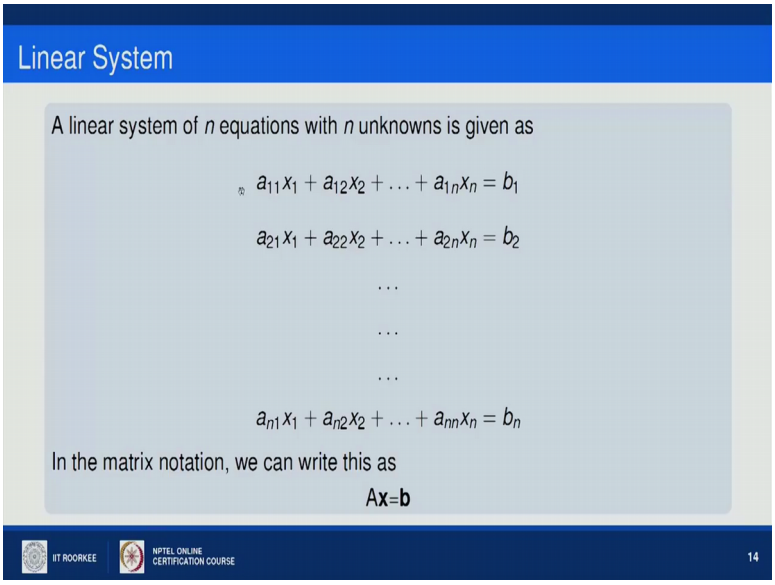
(Refer Slide Time: 20:33)



If you take one more example of the same thing so for example I am having true value is 0.02138 and approximation is let us say 0.02144, so here again following the same process this becomes 0.00006 which is less than 0.0005 or what I can say it is equals to 1 by 2 into 10 raise to power minus 3. So again like earlier one here s minus r plus one equals to minus 3. Moreover here the true value 0.02138 will be greater than 10 raise to power minus 2 that is 0.01. So here this gives me s equals to minus 2, so by substituting this value of s here minus 2 minus r plus 1 equals to minus 3, I got r equals to 2.

So here this approximate x A is having to significant digit to the true value 0.02138 and let me introduce system of linear equations. So in general we use to see the system of linear equations in number of applications in science and engineering, most of the problems in science and engineering and be formulated into nonlinear equations and that can be approximated into or converted into system of linear equations finally to solve the overall system.

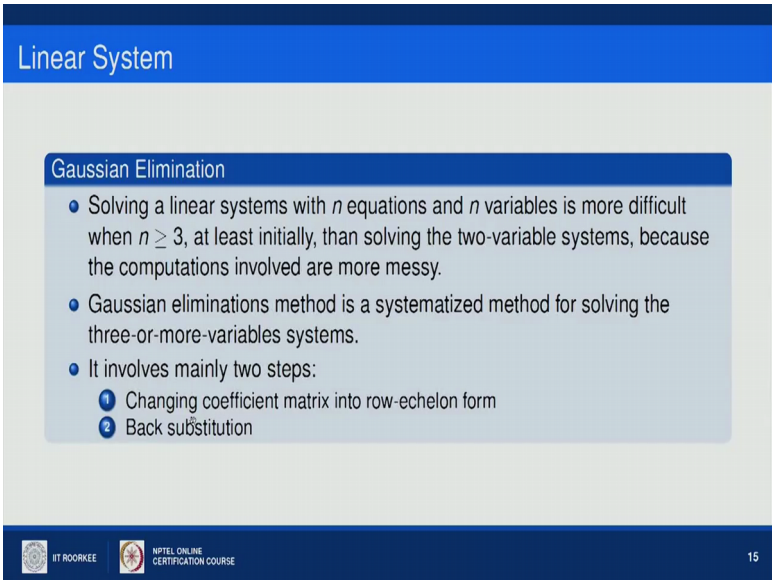
(Refer Slide Time: 22:44)



Now a linear system of n equation with n unknowns is given as by this type of so here this is the 1st the question here x 1, x 2, x n are the unknown variables a 11, a 12, a 11 or a i j are the coefficients in different equations and b 1, b 2, b n are the right-hand side vector. In the matrix notation we can write this particular system as A x equals to b. Now how to solve such a system?

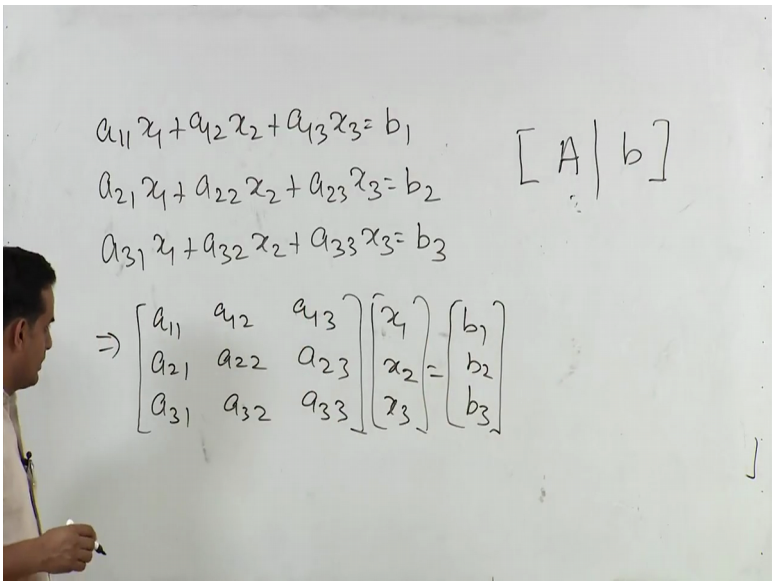
So solving linear system with n equation and n variables is more difficult when n is greater than equal to 3 because if I am having to equations with 2 unknown I can solve it directly but if I am having 3 equations or more than 3 equations in the same number of unknown variables, it becomes quite difficult and we need some systematic way of solving such system. So here I want to introduce a direct method that is called Gaussian elimination, so it is a systemised systematic method for solving the system of linear equations having 3 or more variables with the same number of questions.

(Refer Slide Time: 24:02)



It involves mainly 2 steps 1st one is changing coefficient matrix into row equivalent form, so basically coefficient matrix are having the coefficients of unknown variables from different equations and then finally back substitution. So let us take an example of 3 by 3 systems to explain this particular method and then we will solve an example of this.

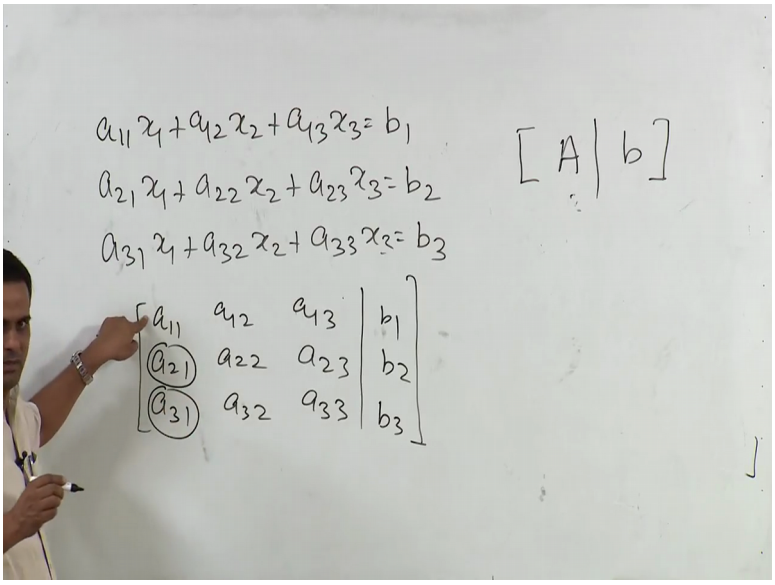
(Refer Slide Time: 24:30)

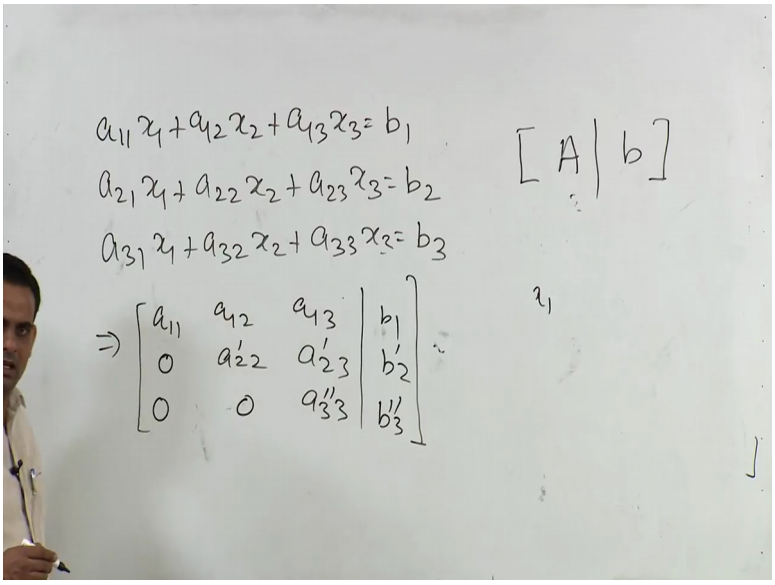


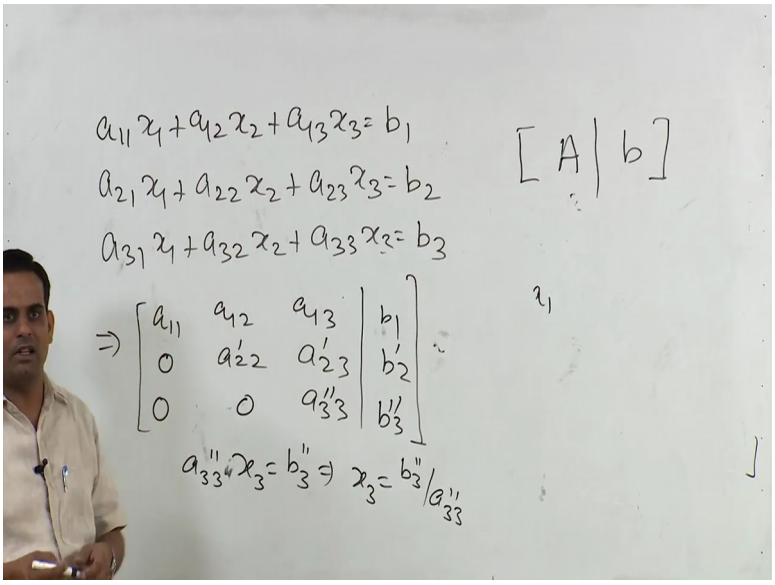
So suppose I am having 3 equations the 1st equation is like this. So 3 equations and 3 unknown in matrix form I can write it a 11, a 12, a 13, a 21, a 22, a 23, a 31, a 32, a 33 into the unknown vector x 1, x 2, x 3 equals to right-hand side vector b 1, b 2, b 3. So in Gaussian elimination method what we use to do, 1st of all we will write the augmented matrix which is given as the coefficient matrix say and we append the right-hand side vector as the last

column. So for example this is the efficient matrix here, what I will do? I will apprehend this b 1, b 2, b 3 here. Now this is my augmented matrix. Now what I will do? I will perform elementary row operations on this matrix to convert it into row equivalent form. So 1st of all to converted in row equivalent form which is basically I want to make it an upper triangular matrix something like this.

(Refer Slide Time: 26:30)







So 1st of all I will make this and this particular element 0 with the help of this element, so if this element is 0, what I will do? I will interchange it with some other row where 1st element the element in the 1st column is known 0, so then after that if I will do this some row operation on this so 1st row will not change so I can do it here itself, so 1st row will not change, so I will make this 2 entries 0 and then these numbers will change after that what I will do, I will make this particular anti-0 with the help of 2nd row, so I will make this anti-0 so these 2 entries will change.

Now if you see this particular thing in the system of linear equations from the last row I will be having a 33 double prime or double dash equals to into x 3 equals to b 3 double dash, so this will give me x 3 equals to b 3 double dash upon. So from here I will calculate the value of x 3, if I substitute this value of x 3 in the 2nd equation I will get the value of x 2 and if I substitute the value of x 3 and x 2 back into the 1st equation I will get the value of x 1 and this particular procedure is called Gaussian elimination.

(Refer Slide Time: 28:21)

Linear System

Example

Let us solve the following system using Gaussian elimination method on a computer using a floating-point representation with three digits in the mantissa and all operations will be rounded.

$$0.143x_1 + 0.357x_2 + 2.01x_3 = -5.173$$

$$-1.31x_1 + 0.911x_2 + 1.99x_3 = -5.458$$

$$11.2x_1 - 4.30x_2 - 0.605x_3 = 4.415$$

Linear System

Example

The augmented matrix for the above system can be written as

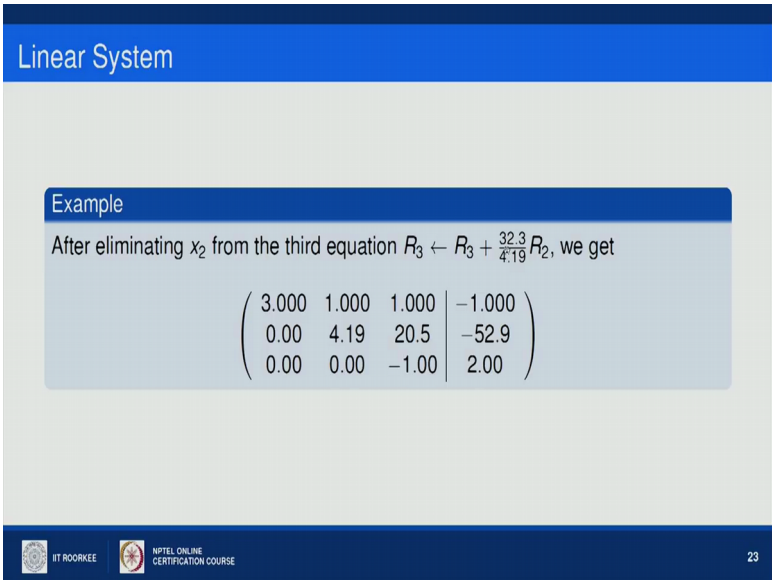
$$\left(\begin{array}{ccc|c} 0.143 & 0.357 & 2.01 & -5.17 \\ -1.31 & 0.911 & 1.99 & -5.46 \\ 11.2 & -4.30 & -0.605 & 4.42 \end{array}\right)$$

Linear System

Example

After eliminating x_1 from second and third equations using row operations $R_2 \leftarrow R_2 + \frac{1.31}{0.143}R_1$ and $R_3 \leftarrow R_3 - \frac{11.2}{0.143}R_1$, we get

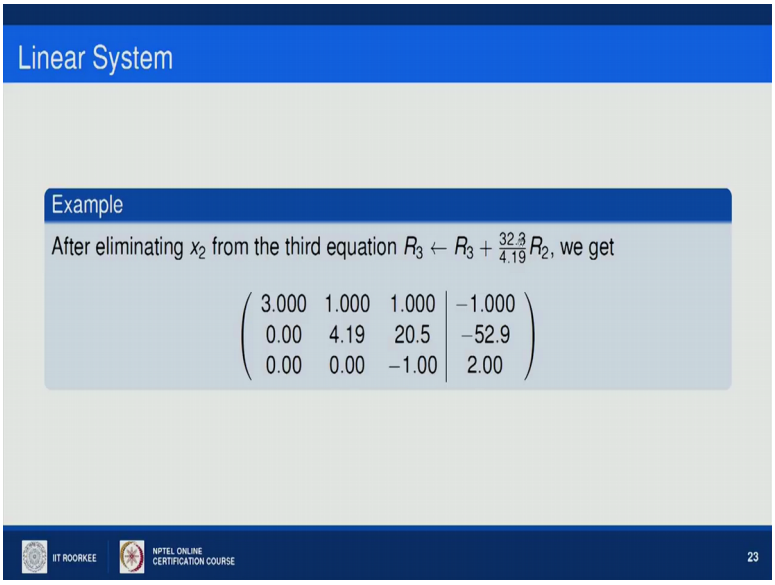
$$\left(\begin{array}{ccc|c} 0.143 & 0.357 & 2.01 & -5.17 \\ 0.00 & 4.19 & 20.5 & -52.9 \\ 0.00 & -32.3 & -159 & 409 \end{array}\right)$$

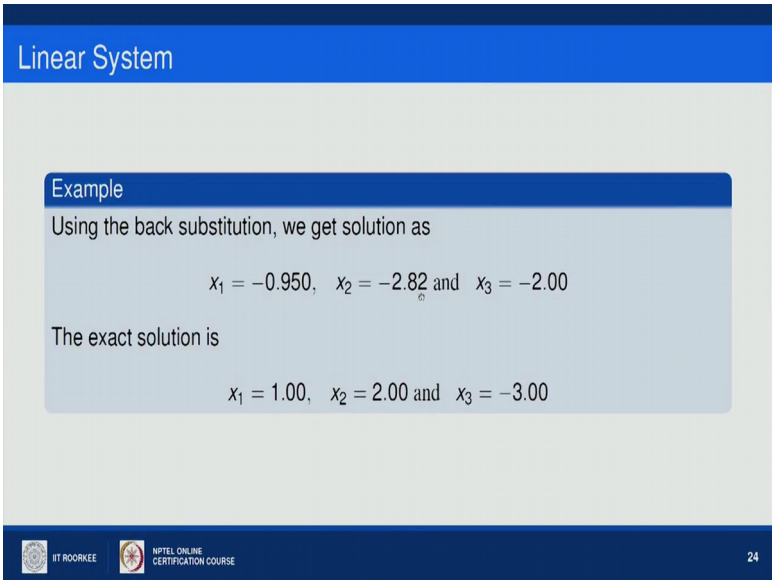


So let us take an example this particular system of equation, we need to solve the following system using Gaussian elimination on a computer using floating point representation with 3 digits in the Mentissa and all operations will be rounded. So I am having 3 equations and 3 unknown, so augmented matrix for this system is given by this particular 3 by 4 matrix, so this is the right-hand side vector.

Now what I will do 1st of all I need to make this and these 2 anti-0 with the helper of 1st one, so for that I will use the row operations like R 2 is replaced by R 2 plus 1.3 upon 0.143 R 1 and R 3 is placed by R 3 minus 11.2 up on 0.143 R 1, we get this particular matrix, so these 2 entries are 0 these entries are also changed. When I will make this particular term 0 with the help of 2nd row, so I need to perform the operation R 3 is replaced by R 3 plus 32.3 upon 4.19 R 2 from that I will get this one.

(Refer Slide Time: 29:31)





So now you can see this particular matrix in a coefficient matrix is an upper triangle matrix. Now from the 3rd of this I can say that x 3 is 2 upon minus 1 that is minus 2. If I substitute this value of x 3 in the 2nd equation I will get x 2 as minus 2.82 and finally if I substitute the value of x 3 and x 2 in the 1st equation I will get extra one as minus 0.950 however the exact solution of this system is like this and you can see you having huge error in our solution which we obtain using the Gaussian elimination method.

Here as you know we have put the procedure in a correct manner but still a big difference in the final solution, so how to overcome this? This can be done using the Gaussian elimination method using partial pivoting and that I will introduce in the coming lecture in the next lecture so in this lecture I told you about the errors in numerical computation, what type of

error we are having? What is the concept of significant digits? I have taken some examples in which will lose the significant digits in further computations and finally Gaussian elimination. So with this I will stop myself, thank you very much.