Elementary Numerical Analysis
Prof. Rekha P. Kulkarni
Department of Mathematics
Indian Institute of Technology, Bombay

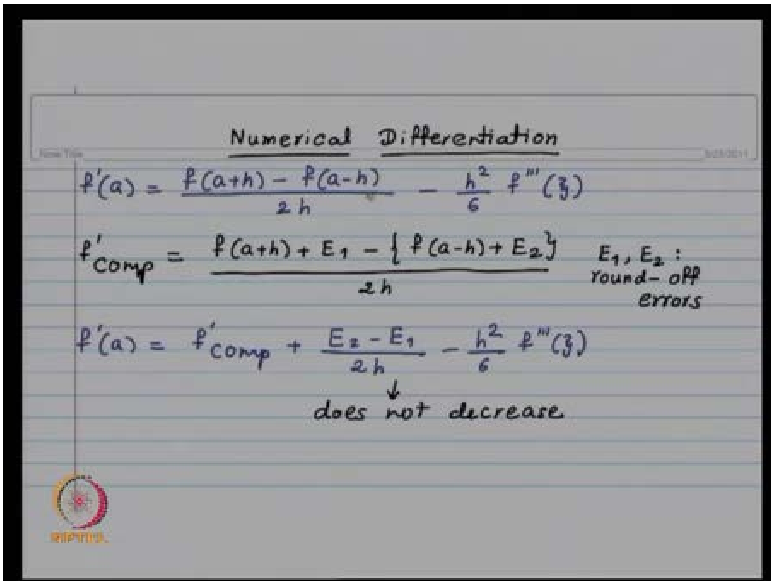
Lecture No. # 17

Gauss Elimination

Last time we have seen that, when we consider numerical differentiation, so when we try to approximate f dash a by derivative of an interpolating polynomial, then there is some problem which we face. So, when we go on reducing value of h, then there comes a stage when instead of the error being decreasing, it starts increasing and this phenomena is because the computations which we do these are done in finite precision. So, a number is represented on a computer and there is some round off error. Now, this problem does not come into picture when we consider numerical integration. So, I want to quickly explain the difference between numerical differentiation and numerical integration using a composite rule.

I am going to explain it for the derivative first derivative f dash at a and let us look at central difference formula, that means, f dash a we are going to approximate it by f of a plus h minus f of a minus h divided by 2 h.

(Refer Slide Time: 01:59)



So, we have f dash a to be equal to f of a plus h minus f of a minus h by 2 h minus h square by 6 f triple dash at some point psi. So, this formula tells us that, if you do exact computations, then as h tends to 0, the error will tend to 0 and the quotient f of a plus h minus f of a minus h by 2 h will tend to f dash a.

In practice, what happens is f of a plus h is not represented exactly. So, instead of that, we have f of a plus h plus some error even minus, there will be similar error in the quantity f of a minus h. So, the f dash computed, instead of computing the quotient, what we compute is f of a plus h plus E 1 minus f of a minus h plus E 2 by 2 h. So, if I substitute for f of a plus h minus f of a minus h 2 h to be equal to f dash computed plus E 2 minus E 1 by 2 h in this formula, what we get is f dash a, this is our f dash computed plus term E 2 minus E 1 by 2 h minus h square by 6 f triple dash of psi.

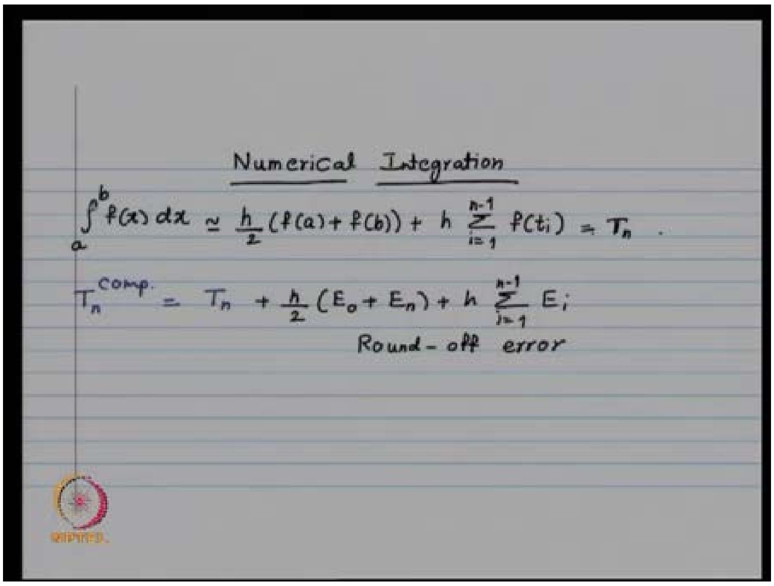
So, thus our error consists of two parts: one part is the discretization error, which is minus h square by 6 f triple dash psi and other error is E 2 minus E 1 by 2 h, which is because of the round off error. Now, there is no reason to expect that, the round of error E 1 and E 2, they will cancel. So, E 2 minus E 1 this need not be equal to 0.

Then E 2 minus E 1 divided by 2 h, when you increase h or whether we are actually letting h to tend to 0. So, when you decrease h, there will come a stage when this factor E 2 minus E 1 by 2 h, this will start dominating the other factor, which is minus h square by 6 f triple dash of psi and that is why when you are doing numerical differentiation, there is going to be a limit to the accuracy which we attend.

Theoretically yes, as h tends to 0, error should tend to 0, but in practice it does not happen. After a certain stage when you go on reducing value of h, after a certain stage the round off errors they start dominating the total error and they go on increasing, because you are diving by a small number. Now, in case of numerical integration, there are also we are letting h to tend to 0, because there h was length of our subinterval. When you look at a composite numerical quadrature rule, we have our interval A b, we sub divide it into smaller intervals and on each subinterval, we apply some basic rule. So, this is how we construct, the composite rules and then as the number of subintervals increase, n tends to infinity or equivalently when the length of the subinterval h, which is b minus a by n that tends to 0, our numerical quadrature formula is going to give us better and better approximation of integral a to b f x d x.

So, look at composite trapezoidal rule. So, you have integral a to b f x dx is approximately equal to h by 2 f a plus f b plus h times summation f t i; i goes from 1 to n minus 1 interior partition points. Now, each of the value f ti and f a and f b, there is going to be some round off error, but here you are multiplying by h; you are not diving by h.

(Refer Slide Time: 07:11)



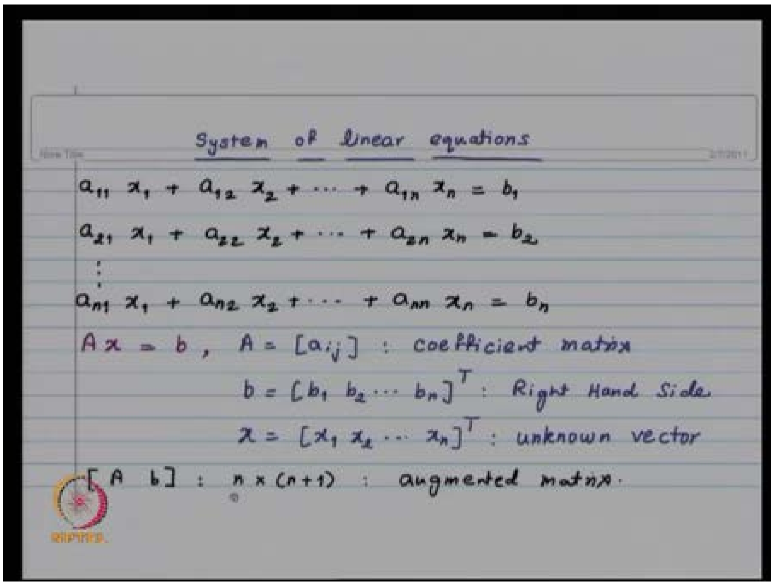
So, T n computed is going to be T n the trapezoidal rule plus this round off error. So, E 0 plus E n and then you have got E i and you are multiplying by a number h. So, this makes a difference and you are going to get a approximate value of integral a to b f x d x which is acceptable; there will be some part because of the round off errors, but that will not dominate the total error. So, we started with interpolating polynomials, we wrote down the divided difference form that allowed us to get an error formula and we integrated interpolating polynomial in order to get numerical quadrature rules. Differentiating this polynomial gives us a way to calculate approximate value of derivatives.

Now, we are going to start a new topic and that topic is solution of system of linear equations. So, we will have n equations in n unknowns; we are going to consider a square system. So, n equations in n unknowns, we will assume that the coefficient matrix is invertible, that means, our system is going to have a unique solution. So, that is our starting point.

Now, we want to solve this system; we are going to consider a real numbers. So, all the quantities involved, they are all going to be real number. There is a classical way of finding a solution; so that is known as Cramer’s rule. So, Cramer’s rule is theoretically important; it is an elegant formula for calculating the solution, but it is going to be very expensive. So, in practice, one does not use Cramer’s rule for finding solution of system of linear equations. Our starting point is going to be gauss elimination method, a simple method to describe.

So, this gauss elimination method, we will first describe, then we will show equivalence of gauss elimination method with l u decomposition of matrix A. What we will be interested in is the number of computations needed. So, we will calculate the numbers of operations needed in the gauss elimination method for solving system of linear equation. From l u decomposition for positive definite systems, we will show what is known as Choleskyed composition, then we are also going to consider the conditioning of the matrix. So, for that we need to define what a norm of a matrix is and so on.

(Refer Slide Time: 11:03)



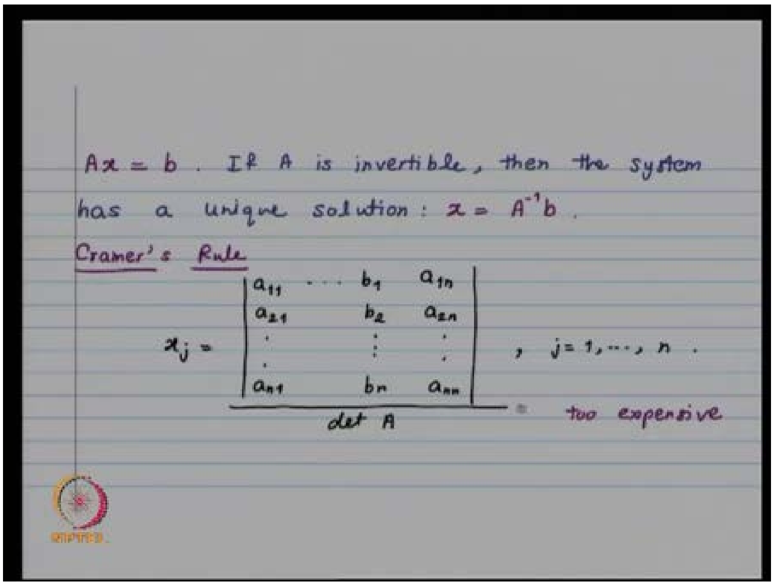
So, let us start with description of .Gauss elimination method; so we have our system. So, the system is a 11 x 1 plus a 1 2 x 2 plus a 1 n x n is equal to b 1, then the second equation is a 2 1 x 1 plus a 22 x 2 plus a2 n x n is equal to b 2 and a n 1 x 1 plus a n 2 x 2 plus a nn x n is equal to b n. The right side b 1 b 2 b n that is given to us ;a i j's they are also given. So, what is unknown is vector x. So, x 1, x 2, x n, we want to find x 1, x 2, x n which satisfy this system.

If you look at the powers of x i's, they are all equal to 1 and that is why it is known as a linear equation. This system in a compact form we write as A x is equal to b, where A is coefficient matrix a i j; i denotes the row index; j denotes the column index. So, A is going to be n by n coefficient matrix; our vector we denote by as column vectors. So, that is why b is equal to b 1 b 2 b n T, that means, transpose. So, this is a row vector and its transpose that is the right hand side; x is equal to x 1 x 2 x n transpose that is the unknown vector. So, this system is written as A x is equal to b a is the coefficient matrix and if to the matrix A, we Add one more column which consists of b 1 b 1 b n, then you get a matrix of size n rows and n plus 1 columns. So, that is known as the augmented matrix.

Now, we will assume that A is invertible matrix; if A is invertible, A x is equal to b whatever is the right hand side it is going to have unique solution; that unique solution will be given by x is equal to A inverse b. Now, let me first describe Cramer's rule. So, in Cramer's rule, we are assuming that A is invertible matrix that means determinant of A is not equal to 0.

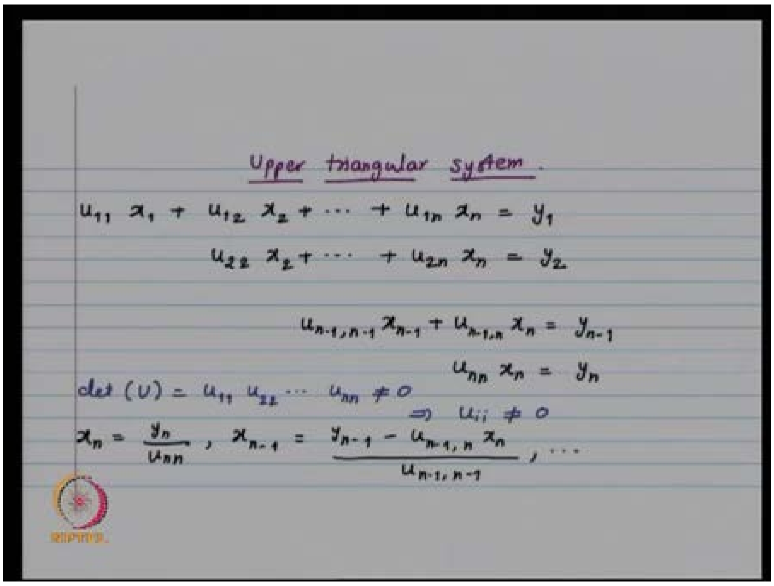
If you want to calculate x j, then what you have to do is, look at the determine look at the matrix A; in that matrix A, replace the jth column. So, the jth column of the matrix will be given by a 1j, a 2j, a nj. So, replace this by right hand side b 1, b 2, b n take the quotient that is going to be equal to x j. So, it is a very nice formula for calculating the unknown vector x 1, x 2 up to x n. So, x j is determinant of this matrix with jth column replaced by b 1, b 2, b n divided by determinant a; j is equal to 1 to up to n.

(Refer Slide Time: 14:45)



The problem with this formula is, it is too expensive. If you want to calculate determinant of A, where A is n by n matrix; you will need at least n factorial multiplications or plus n factorial additions slash subtraction and n factorial grows very fast with n. So, no matter how big your computer is, if you try to calculate determinant of A by traditional way, soon your computer memory will not be enough. So, this Cramer's rule is going to be too expensive to use in practice.

(Refer Slide Time: 15:45)



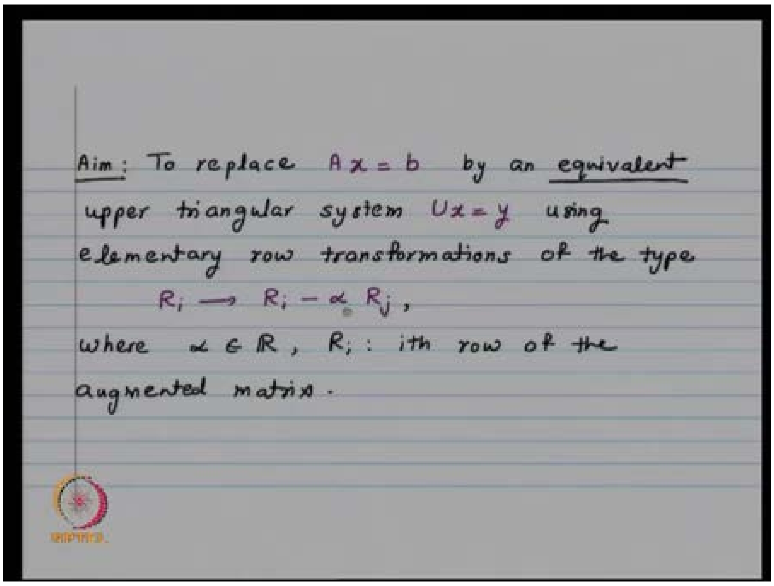
So, now, suppose your system is upper triangular, that means, in the first equation all unknowns appear x 1 x 2 x n; in the second equation, there is no x 1; in n minus first equation what is coming into picture is only x n minus 1 and x n and in the last equation you have got only x n. The coefficient matrix invertible will mean determinant of U which is product of diagonal entries u 11 u 22 u nn this is not equal to 0. So, all u ii's they are not equal to 2, what 1 can do is start with x n; x n is going to be equal to y n upon u nn, then having determined x n, go to this equation; now x n is known. So, take it on the other side and calculate x n minus 1 as y n minus 1 minus u n minus 1 n x n divided by you n minus 1 n minus 1 and. so on.

So, this is known as back substitution. So, if your coefficient matrix is upper triangular, then you can determine vector x in the order x n x n minus 1 x n minus 2 and so on. So, our aim will be to replace our system A x is equal to b by an equivalent system U x is equal to y what I mean by equivalent is both the systems they should have the same solution. We are interested in solution of A x is equal to b. So, if you replace by another system, then the new system should have the same solution.

Because we are interested in the system of original we are interested in the solution of the original system not some other system. So, coefficient matrix A, we want to reduce it to upper triangular form. Now, whatever you will do for A, you will have to do on the right hand side, because we do not want to change our system. So, we are going to use elementary row transformation, that means, what we can do is multiply a equation by a number alpha non zero number and subtract it from other equation; if you do this, you are going to get a system which is equivalent. Now, this equivalence we will show later on.

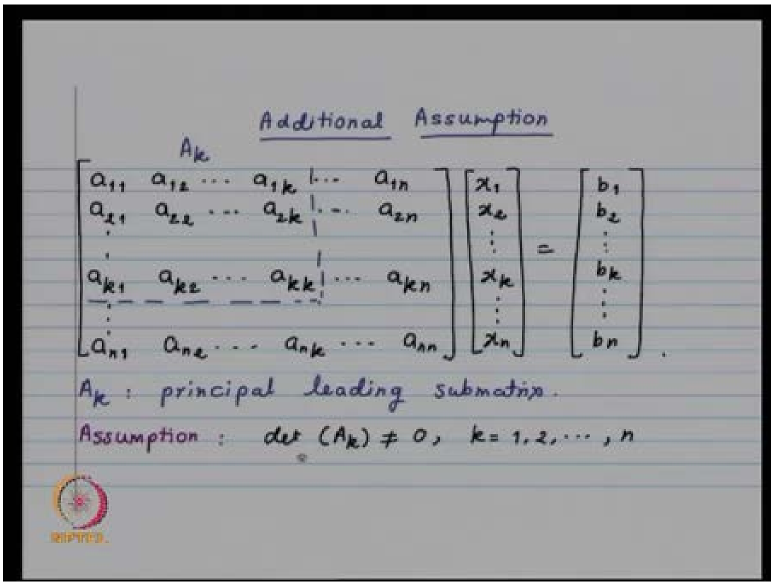
So, first we are going to describe the gauss elimination method, calculate the number of operations in it and then we will come back to equivalence. So, aim is to replace A x is equal to b by an equivalent upper triangular system U x is equal to y.

(Refer Slide Time: 19:30)



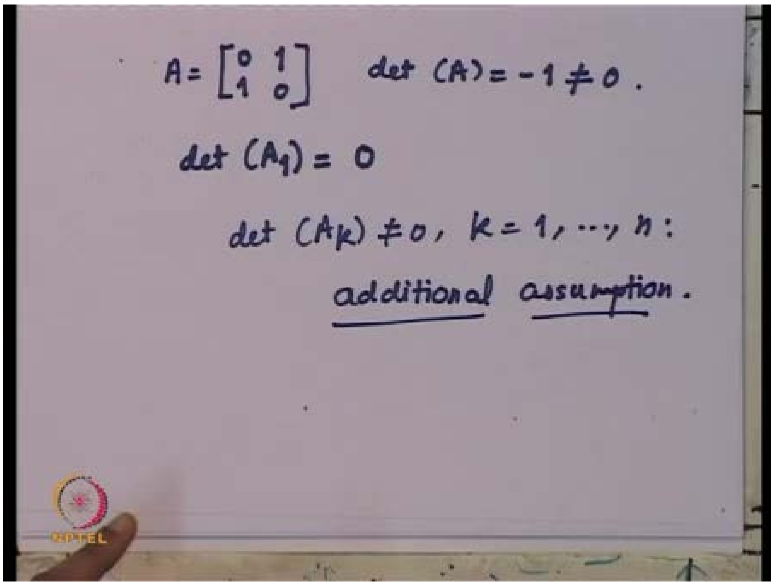
Using elementary row transformations of the type R i becomes R i minus alpha R j, that means, you are multiplying jth through by a constant alpha and subtracting it from R i. So, this was our original one original ith equation and this is the new equation.

(Refer Slide Time: 19:55)



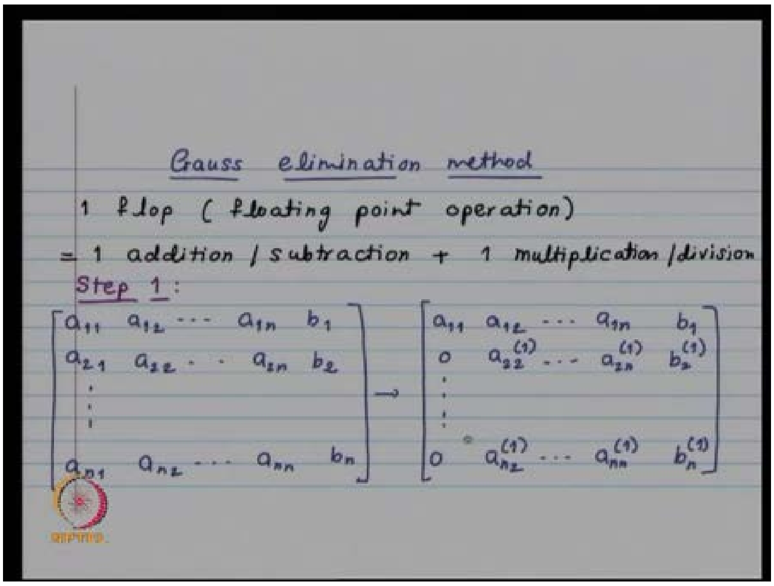
Now, we are going to make an additional assumption A is invertible, but we will look at principle leading sub matrix. So, that is formed by first k rows and first k columns. Our assumption is determinant A k not equal to 0 for k is equal to 1 to up to n. When k is equal to n, that is determinant of A not equal to 0 that we have already assumed. Now, we are saying something more, what should happen is a 11 should not be 0; determinant of a 1 a 1 a 1 a 12 a 21 a 22 should not be 0 and then. So, on this is additional assumption.

(Refer Slide Time: 20:55)



Like let me look at example, say 2 by 2 matrix 0 1 1 0 this matrix is such that determinant of A is equal to minus 1 which is not equal 0, but when I look at determinant of A 1 then that is equal to 0. So, this assumption that determinant of a k not equal to 0 for k is equal to 1 to up to n this is additional.

(Refer Slide Time: 21:57)

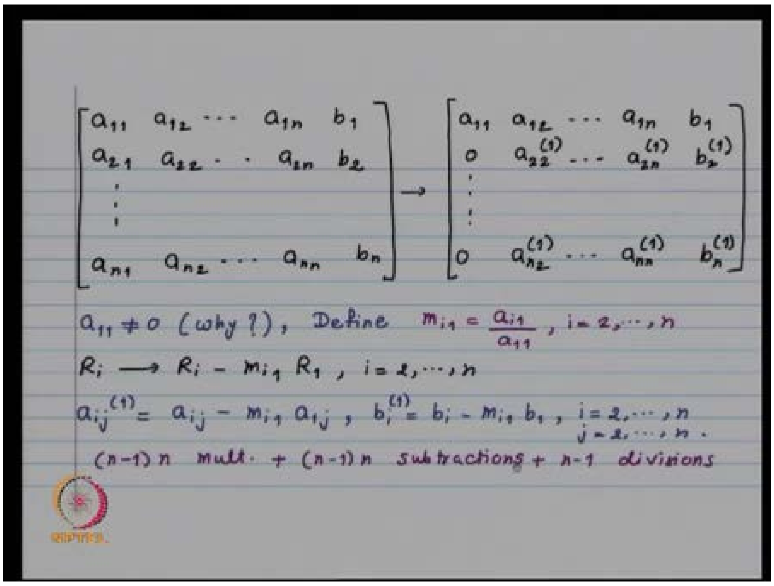


Now, we are going to show that there is a class of matrices for which the this condition is satisfied and those are known as positive definite matrices. By one flop, because as I said we are going to calculate the number of computations, so 1 flop, where flops stands for floating point operation, it consists of 1 addition or subtraction plus 1 multiplication slash division; addition and subtraction, they are considered on par and multiplication slash division they are also considered on par.

Generally, total number of multiplications slash divisions is same as total number of additions slashes subtractions. So, 1 flop is going to consists of this 1 addition slash subtraction plus 1 multiplication slash division. In the first type of gauss elimination method what we want to do is, we want to look at the entries in the first column below the diagonal entry and we want to make all of them 0; this entry a to 1 can be made 0 by multiplying the first equation by a 2 1 by a 1 1 and then subtract it and then so on.

So, when you do that, the first row will remain as it is; all these entries they will become 0 and the entries in this sub matrix, they are all going to get modified. So, that is going to be the first step; in the second step, we will work with this sub matrix; we will not touch the first row or the first column.

(Refer Slide Time: 24:04)

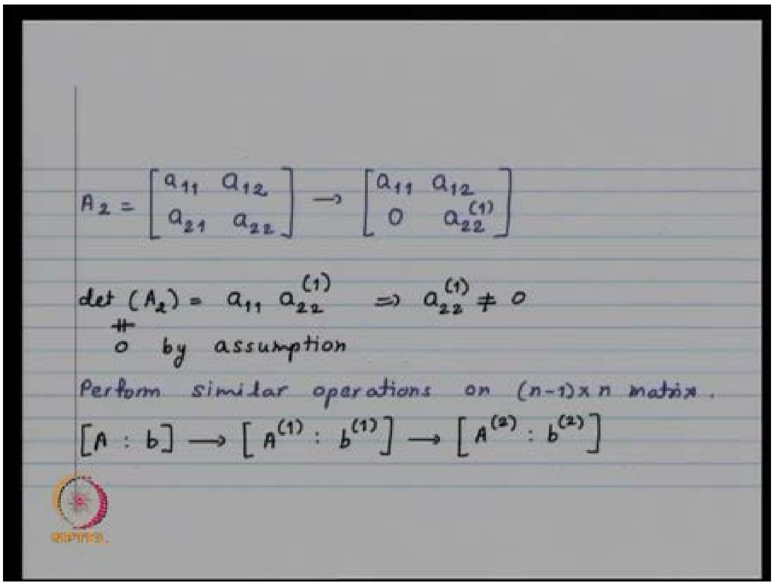


So, first we have to notice that a 11 is not equal to 0 that is because we are assuming that determinant a 1 is going to be not 0, next we define multipliers m i 1 to be a i1 divided by a 11 i goes from 2 up to n. a 11 is not equal to 0. So, m i 1's they are well defined and ith row, from the ith row we will subtract first row multiplied by m i 1 for i is equal to 2 up to n. We want to economize the computations. So, these 0's we will write down directly; we would not make the subtraction, because we have chosen the multipliers in such a manner that this operation will produce 0's in the first column below diagonal a i j 1, these will be given by a i j original entry minus m i 1 times the corresponding entry in the first column; the first suffix denotes the row. So, that is why you have got a and then, j. So, i jth entry minus m i 1 times corresponding entry in the first row.

You have to do it for the last column also. So, it is b i 1 is equal to b i minus m i 1 b 1; i going from 1 to n; j going from 2 to n. So, let us now calculate the number of operations. You are going to have two multiplications here and then one subtraction here; this you are going to do it for this whole square. So, that is going to be n minus 1 square, but you are also doing it here. So, it is becomes n minus 1 into n multiplications plus those many subtractions. So, n minus 1 into n subtractions plus in order to calculate m i 1, you need to do this division. So, there are n minus 1 division.

So, these many computations are needed for the first step. In the second step, we are going to work with a matrix which has got n minus 1 rows and n columns and exactly the same computations. So, now, the number of operations will be obtained by replacing in this relation n by n minus 1.

(Refer Slide Time: 27:19)

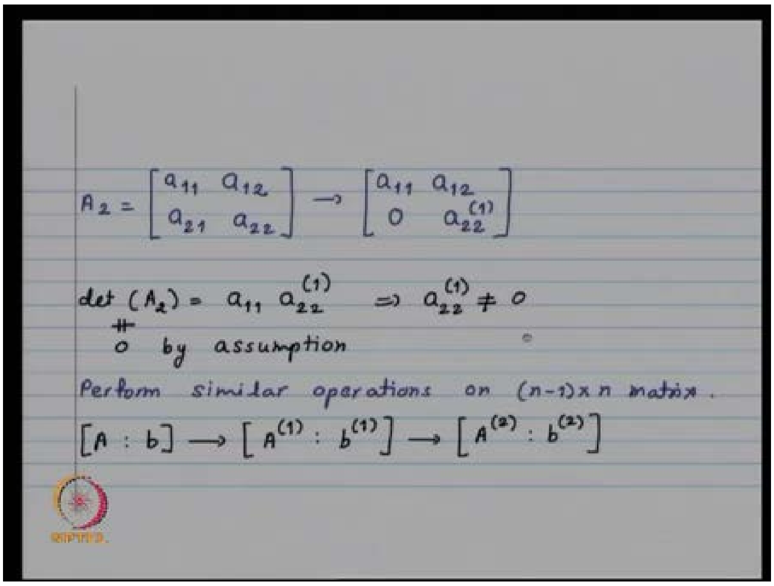


Now, we have to confirm that our procedure is going to work for the sub matrix also; to start with our matrix, in our matrix A, a 11 is not 0, because you are assuming determinant of A k not equal to 0 for k is equal to 1 to up to n. You do some operations, you get a new matrix; in the new matrix which consists of second, third up to nth row and second, third up to n plus first column, now the entry which we are going to divide by that is going to be a 221 the modified entry. So, we have to make sure that a 221 is not equal to 0. What we have done is, we have used elementary row transformation that means, we have multiplied a row by a fix real number and subtracted from other row. Now, this operation does not change the determinant of the matrix; the determinant of the matrix remains invariant. So, look at the 2 by 2 matrix which is formed by first 2 rows and first 2 columns. So, that is our capital A 2.

By assumption determinant A 2 is not equal to 0 and this 2 by 2 matrix after the first stage of our gauss elimination method, it is transformed to the first row as it is, a 11 a 12; second row 0; below the diagonal in the first column we have got 0 and then a 22 1. So, determinant A 2 will be a 11 times a 22 1; a 11 is not 0; a 22 1 also should not be 0, because determinant A 2 is not equal to 0.

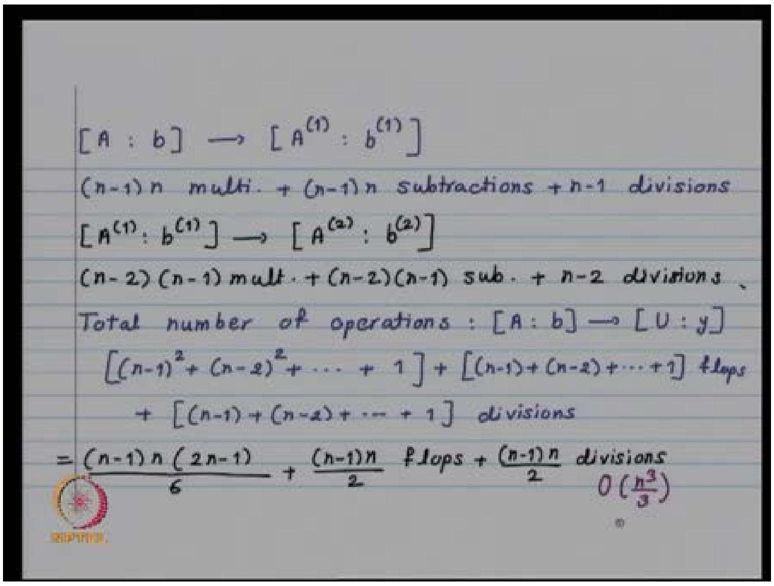
So that means our second step of gauss elimination method we can perform and then one uses the same argument that, you can do this process till your coefficient matrix A is reduced to an upper triangular form.

(Refer Slide Time: 30:10)



So, our starting matrix was A b augmented matrix; we transformed it to A 1 b 1, then using similar operations we have transformed it to A 2 b 2 and so on. In order to do do this this transformation, we need the fact that a 22 1 is not equal to 0.

(Refer Slide Time: 30:37)



So, now, the number of computations, we have seen that for the first step of gauss elimination method, we had n minus 1 into n multiplications plus n minus 1 into n subtractions plus n minus 1 divisions. In the second step, you are performing a similar operation on a sub matrix with n minus 1 rows and n columns.

So, everywhere you have to replace by n by n minus 1. So, it will be n minus 2 into n minus 1 multiplications plus n minus 2 into n minus 1 subtractions plus n minus 1 division. So, when you consider total number of operations which are needed to reduce the system A b to a system U y, where U is upper triangular matrix, it will be given by. So, here write this n as n minus 1 plus 1. So, you will have n minus 1 square. So, you will have n minus square plus n minus 2 square plus 1 plus this n we are writing as n minus 1 plus 1. So, that is why you will have plus n minus 1 from here, then n minus 2 plus 1 and now I am writing as flops number of multiplications and subtractions, they are the same and 1 multiplication plus 1 subtraction is 1 flop.

So, that takes cares of this and what is remaining is divisions. So, you have n minus 1 plus n minus 2 plus 1 division. Now, this is square of natural numbers; so we have got a formula. So, it is n minus 1 n 2 n minus 1 by 6 plus sum of these will be given by n minus 1 into n by 2; so, these many flops plus n minus 1 into n by 2 divisions. Now, this gauss elimination method which I have described it is for solving big system of linear equations using computer. If you are trying to solve a 3 by 3 system by hand, it will not matter which method you you use, you can use Cramer's rule, but in practice one comes across big systems of equation; so then one has to worry about the number of operations Now, we have calculated the number of operations; these number of operations, they are some constant times n cube plus another constant times n square plus constant times n plus constant.

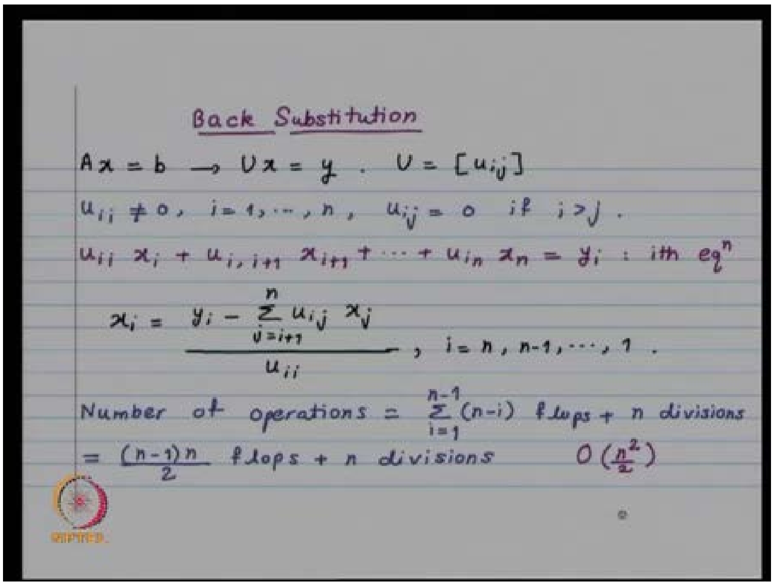
Now, when n is big what matters is, what will be coefficient of n cube like between n cube and n square. For n begin off you can ignore what is n square like, if your n is 10000, then n cube is going to be much bigger than n square. So, the dominating factor becomes n cube, but you should retain the coefficient of n cube, because there is a difference between n cube and 2 n cube the 2 n cube is double the number of operations. So, in this gauss elimination method coefficient of n cube is 1 by 3. So, 1 says that you can reduce your system to a upper triangular system by doing the number of operations of order n cube by 3.

So, now, we have reduced our system to upper triangular system, but still it remains to find a solution we do this, because we want to solve A x is equal to b. So, now, let us calculate the number of operations in the back substitution. A x is equal to b is our original system; this we have reduced it to upper triangular system U x is equal to y by performing number of operations of the order in cube by 3 and now let us see what effort is needed for getting vector x 1 x 2 x n from U x is equal to y. So, that we are going to do it by back substitution.

A x is equal to b A is invertible matrix. So, determinant of A is not equal to 0. All the operations which we have used, those are elementary row operation R i minus alpha times R j. So, this leaves the determinant in variant. So, determinant of U will be same as determinant of A U is a upper triangular matrix; for an upper triangular matrix, determinant is product of the diagonal entries. So, you are going to have u 12 u 22 into u nn as determinant of U; determinant of U not equal to 0 will mean that all the diagonal entries u ii, they are not equal to 0.

And we have already talked about the back substitution start with the last equation determine x n, then go to n minus first equation determine x n minus 1 and so on.

(Refer Slide Time: 37:17)



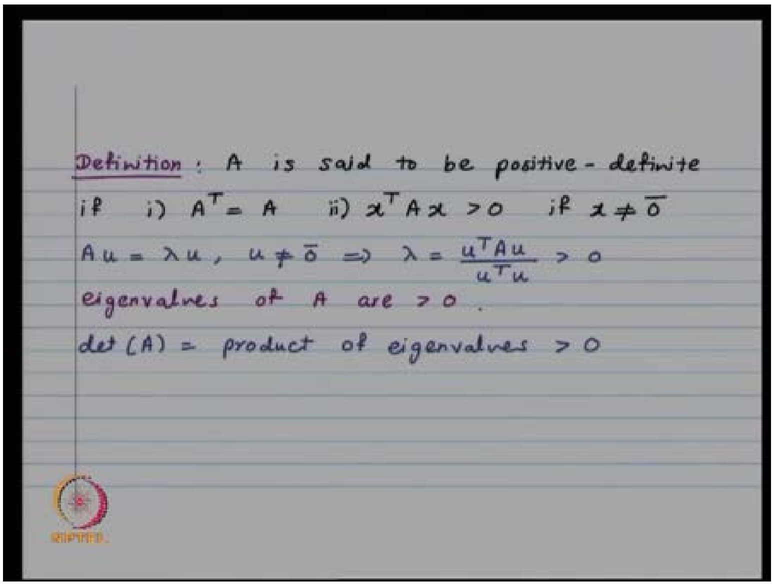
So, here U being an upper triangular matrix u ij will be 0 if i bigger than j and hence ith equation is going to be of the form u ii x i plus u ii plus 1 xi plus 1 plus u i n x n is equal to y i. So, now, x i will be equal to y i minus summation j is equal to i plus 1 to n u i j x j divided by u i i is equal to nn minus 1 up to 1. With the convention that if i is equal to n, then this summation is absent. Now, how many numbers of operations? So, here you have got j is equal to i plus 1 to n. So, you are doing n minus i multiplications u ij x j you have got n minus i terms when you add them up, you will need n minus i minus 1 additions, because when you want to add two numbers you do one addition, but then there is one subtraction here. So, that means, in order to calculate the numerator, you will need n minus i flops and there is going to be one division; this you will be doing it for i is equal to nn minus 1 up to 1.

So, the total number of operations they will be given by summation i goes from 1 to n minus 1 N minus i flops plus n divisions and this summation is nothing, but n minus 1 into n by 2 flops plus n division. So, that means, it is going to be of the order of n square by 2. So, thus the major chunk of our computations they are needed to replace our matrix A by an upper triangular matrix; the once you have done that, back substitution is of the order of n square or n square by 2, if we want to be really précised. So, suppose sometimes one wants to solve the same system of equations for different right hand side, so in that case what you should do is you should retain your say number of operation or whatever you are doing. Because see what you have now is A x is equal to b. So, the coefficient matrix A is the same, but on the right hand side you have got b 1 b 2 b n those will be different. So, you should keep the track of the operations reduce a to upper triangular for once for all, whatever are those operations you have to perform them on the right hand side also and then, solve system of linear equations. So, this is gauss elimination method; it is also known as gauss elimination method without partial pivoting.

We will see what pivoting means, but at present this is the simplest method. This method is applicable provided determinant A k is not equal to 0 for k is equal to 1 2 up to n. Because if at any stage, the number by which we are dividing in our multipliers we need to divide by certain number. So, if this becomes 0, then our system or our method will be it will fail then we cannot proceed. So, there is a big class of matrices for which this condition determinant A k not equal to 0 for k is equal to 1 2 up to n is satisfied and that class of matrices that is known as class of positive definite matrices.

So, our definition of positive definite matrix is going to be matrix is a real matrix; it should be a symmetric matrix A transpose is equal to A and in addition, if you look at a nonzero vector x, then x transpose A x should be bigger than 0; A is n by n matrix x is n by 1 vector. So, A x is going to be a n by 1 vector; our x transpose will be 1 by 1 n row vector. So, when you consider x transpose A x, what you get is a 1 by 1 matrix or a real number; if that number is bigger than 0, for all nonzero vectors, then our matrix is going to be positive definite.

(Refer Slide Time: 43:01)



Now, for the positive definite matrix when we look at its Eigen values, so we are going to look at the Eigen values of matrix little later, but at present lambda is an Eigen value of a provided, there exist a nonzero vector such that A u is equal to lambda u that is definition of Eigen value. Lambda is going to be equal to u transpose A u divided by u transpose u. Now, when A is positive definite, u transpose A u will be bigger than 0; u transpose u anyway it is going to be bigger than 0 and hence, the Eigen values of A, they are going to be strictly bigger than 0.

So, for positive definite matrices, Eigen values of A, they are bigger than 0 and determinant of A is product of Eigen values. So, determinant of A is bigger than 0. So, that means, for a positive definite matrix, determinant of A is going to be not only equal to 0, it will not be equal to 0, but it will be strictly bigger than 0. Now, what we want is additional condition; we want determinant A k should not be equal to 0. In order to show that positive definite matrices satisfy this property, what we are going to show is that A k which is a sub matrix formed by first k rows and first k columns of the matrix, they are also going to be positive definite. In order to show this we will have to go by definition.

So, for positive definite first it should be symmetric. So, checking symmetry of A k is no problem, but showing that y transpose A k y is bigger than 0 for y not equal to 0 that will need some proof. So, we will show next time that if A is positive definite, then all its principles sub matrices, they are also going to be positive definite, which will mean that determinant of A bigger than 0 for k is equal to 0 to up to n which means (( )) method which we described just now we can apply it to this class of matrices and then, we will also see some decomposition of positive definite matrix which is known as the colicky decomposition. So, we will continue next time.

Thank you.