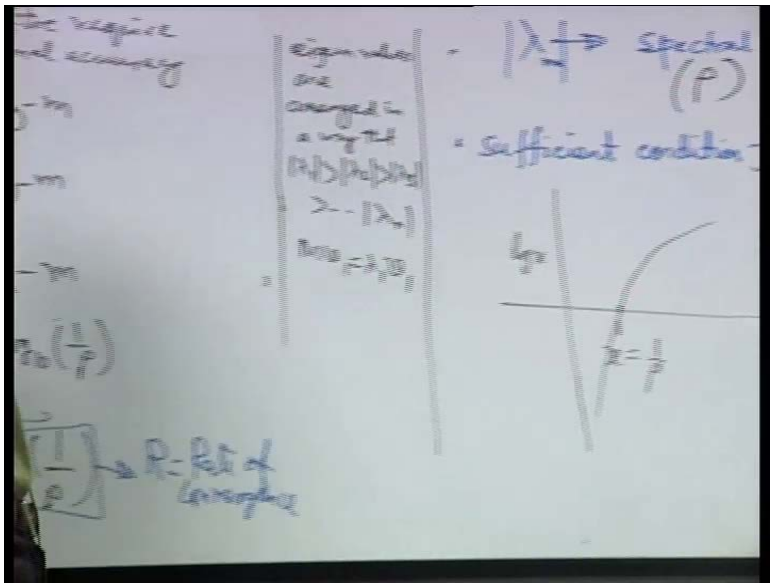
Computational Fluid Dynamics
Dr. Suman Chakraborty
Department of Mechanical Engineering
Indian Institute of Technology, Kharagpur

Lecture No. # 26
Iterative Methods for Numerical Solution of Systems of Linear Algebraic Equations (Contd.)

Towards the end of the previous lecture we were discussing about two important terminologies - the spectral radius of convergence and the rate of convergence and we will be continuing with that.

(Refer Slide Time: 00:36)



Let us try to focus a bit on first the spectral radius of convergence. So, as we have seen that if you want to have a very few number of iterations to satisfy convergence, then what should be the spectral radius of convergence? So, just keep in mind that essentially, the rate of convergence is related to log of 1 by the spectral radius of convergence.

If you make a plot of log x verses x, it will be something like this right now. That means, if x is 1 by rho as you have smaller and smaller values of rho, remember that we have the maximum value of rho as 1 to achieve the sufficient condition for convergence. That means, it is a positive fraction. So, 1 by the positive fraction is a number which is greater than 1. So, smaller the value of rho, larger is this value of x which is say 1 by rho and

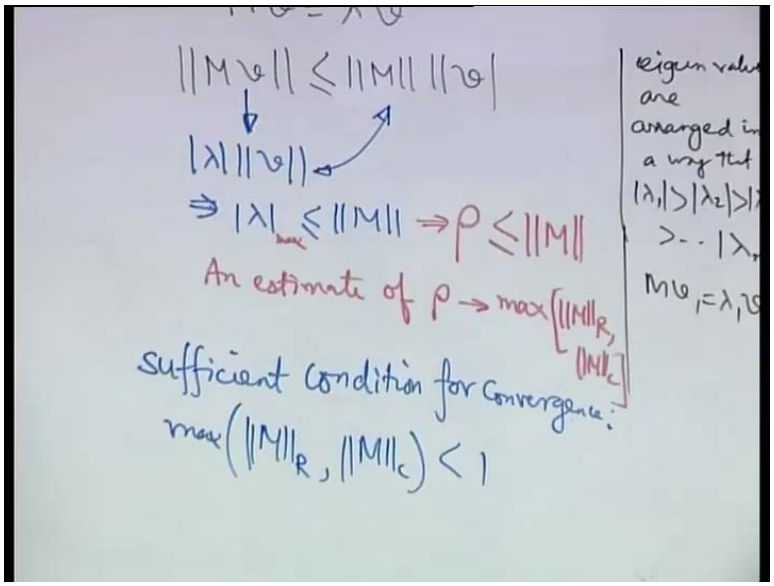
larger is log x, right. So, essentially you are dealing with variations towards the right side of x equal to 1 because rho is less than 1.

So, as you increase x, your log x increases and in the limiting case when rho is very small then log x is very large. That means, the rate of convergence is very large theoretically when rho tends to 0 x tends to infinity. So, log x also tends to a very large number. And, when log x tends to a very large number, that means, the rate of convergence is very large. So, you require a very few number of steps to achieve convergence. So, what we can infer from this is that, smaller you keep the spectral radius of convergence, better it is in terms of the rate of convergence.

To achieve a high rate of convergence your spectral radius of convergence should be as small as possible; that is the first and foremost thing that we can infer now. But, how to assess what is the spectral radius of convergence. There are several ways in which you can assess the spectral radius of convergence, but we have to keep in mind is that again to do that in a straight forward way we need to calculate the Eigen values of the metric sem.

Eigen value computation is also expensive just like we had seen that the power - the power computation of the metric semi. It is expensive to get rid of that problem we came down to the equivalent Eigen value representation, but equivalent Eigen value representation. If you want to use that you need to calculate the Eigen values of m that is also expensive. So, we need to go to something which is again equivalent to the Eigen value representation, but something which is computationally much more straight forward.

(Refer Slide Time: 04:52)



To do that, we will consider the norm of the matrix m. So, remember that m v is equal to lambda v lambda is an Eigen value of the matrix m. Now, we know that norm of m v is less than equal to norm of m into norm of v and because m v is equal to lambda v. Therefore, norm of m v is mod of lambda into norm of v right and from these two it follows that mod of lambda less than equal to norm of m.

Since, mod of lambda is less than equal to norm of m maximum of mod of lambda is also less than norm of m that is the spectral radius of convergence is less than equal to norm of m. So, if you can have the estimate of norm of m then that is good enough to give you a bound of the spectral radius of convergence and that can in effect be substituted in place of the Eigen value representation to get an estimate of that characteristics of convergence.

Now, when you say the norm of m again, the question comes that which norm I mean there we have we have seen that there can be several types of norms possible. But, some of the limiting norms are and some of the norms which are easy to calculate are the first norm and the infinity norm that is the column sum norm and the row sum norm.

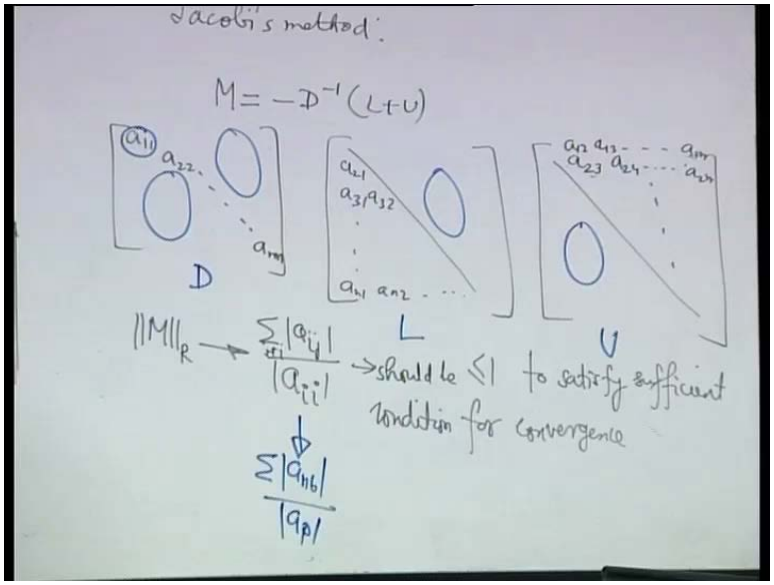
Conservative estimate and estimate of rho is given by the maximum of the row sum norm and the column sum norm. That is out of these two types of norms whatever norm is the maximum one. Remember, that you have to keep the maximum of the maximum of the lambda mod lambda maximum out of all lambdas that you need to keep less than 1

and norm of m is always greater than that 1. Out of all possible norms of m whatever is the maximum that if you keep less than 1 then; that means, you are maintaining the or you are satisfying the sufficient condition for convergence.

So, you are becoming very conservative you know that whatever is the norm of m that is greater than rho for you it was sufficient to satisfy rho less than equal to one, but if you satisfy norm of m less than equal to 1 then, implicitly you are satisfying rho less than 1 definitely. And, out of the two extreme norms, one of those whatever is the greater one that you satisfy as less than 1; that means, the other one will be automatically be less than 1.

So, your sufficient condition for convergence boils down to less than or equal to 1. Now, let us take an example and see that what can be the norm for the method that we have learnt. So far we will take an example of the Jacobi's method. Why we take the example of the Jacobi's method again for estimating the rate of convergence or any convergence criteria Jacobi's method will give you a conservative estimate because, Gauss Seidel is faster than the Jacobi's method. So, if it works for Jacobi in an efficient way you expect that for Gauss Seidel it will work in a more efficient way.

(Refer Slide Time: 09:50)



Now, for the Jacobi's method, what was m minus D inverse 1 plus u? This was m for the Jacobi's method, where let us write again what are the entries of l D and u. So, D is a 1 1 a 2 2. See, eventually the matrix m is a derived matrix the original matrix which was

given to us was the coefficient matrix. Therefore, we should be able to write everything in terms of the entries of the coefficient matrix. So, this is the d and l is a 2 1 a 3 1 a 3 2 and a n 1 a n 2 in this way this is l and u u is a 2 3 a 2 4 in this way, sorry, we will start with a 1 2 a 1 3 in this way a 1 n then a 2 3 a 2 4 in this way a 2 n and so on.

Now, d inverse into 1 plus u; so, D inverse roughly represents what it is; remember the diagonal for each row contains only one number. So, D inverse roughly represents the division by the diagonal component of the matrix for a single scalar number inverse is 1 by that 1 for a matrix where you have more number of entries. It is systematically calculated in a more general way, but the diagonal matrix is a special type of matrix where only diagonal matrix is diagonal term is there.

So, you can write this as 1 plus u means sum of log diagonal terms divided by the diagonal term. So, now, if you want to calculate the norm of this one you can calculate either the column sum norm or the row sum norm. let us consider the row sum norm as an example. So, if you consider the row sum norm then what will be that?

Sometimes out of the row sum and the column sum norm, the row sum norm is used as a criteria I mean, there is nothing hard and fast about it. One could use column sum or the max of those two, but usually any one of those is considered; it is just a matter of convention, there is nothing very fundamental about it.

It is just a convention that in many cases the row sum norm is used as a representation of the norm of the matrix. So, what will be the row sum norm? For any row we divide by the corresponding diagonal, so, a i i for the it row is the diagonal. So, you divide it by this 1. So, mod of that 1 in the numerator you have summation of mod of a i j where j is not equal to I and this you expect to be less than equal to 1 to satisfy sufficient condition for convergence.

Now, try to relate this with the c f d discretization in c f d discretization using the nomenclature that we have considered for the grid layout. This is summation of a n b that is naive ring as divided by summation of a p mod of both. So, the diagonal represents the grid point. The grid point p and the other 1s represent the off diagonal; 1s represent the naive rings. So, this is like the point p wherein the other sides of it you the effect of its naive rings. For a 1 dimensional system it is a tri diagonal matrix whereas, for a multi dimensional system it is for a 2 dimensional system considering the finite grid

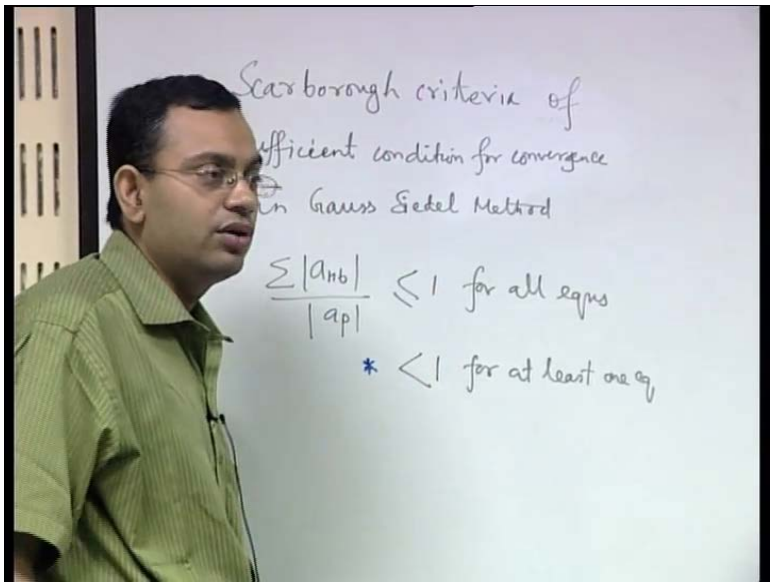
considering the finite volume grid layout that we had considered it is a pent diagonal matrix. So, you have another two diagonals on the two sides of it.

So, in the 3 dimensional you have another two. So, 7 total diagonals involved. So, it is possible to have several of diagonal or near diagonal terms, but the matrix is generally spars these types of matrices are called as spars matrices where the coefficients in many of the terms are 0s. So, you have essentially the diagonal coefficients which are non-zero and you have some of the off diagonal coefficients which are non-zero depending on I mean, if it is a tri diagonal system then 3 diagonal see if it is a pent diagonal systems - 5 diagonals. But, these are not the only terms. So, you have majority of the terms which are away from the diagonal these are 0s. So, the matrix is densely populated with 0s and these types of matrices are called spars matrices and it particularly important in computational fluid dynamics to understand how to handle these spars matrices.

Spars matrices are actually very irritating because if you are trying to represent the matrix as it is then you are wasting a lot of storage space for storing zero's you we have seen that how cleverly that can be avoided for a tri diagonal matrix by using the tri diagonal matrix algorithm by converting the storage from a 2 dimensional to a 1 dimensional storage. So, it is important to take care of that in the formulation of the solution procedure.

Now, coming back to this effect of the naives; so, you have effect of the naives as summation of the mod of the naive ring terms divided by the grid point contribution and that needs to be less than equal to one to satisfy the sufficient condition for convergence keeping this in view proposed a criteria which is called a Scarborough criterion for sufficient condition of convergence for the Gauss Seidel method by taking clue from here and let us state that.

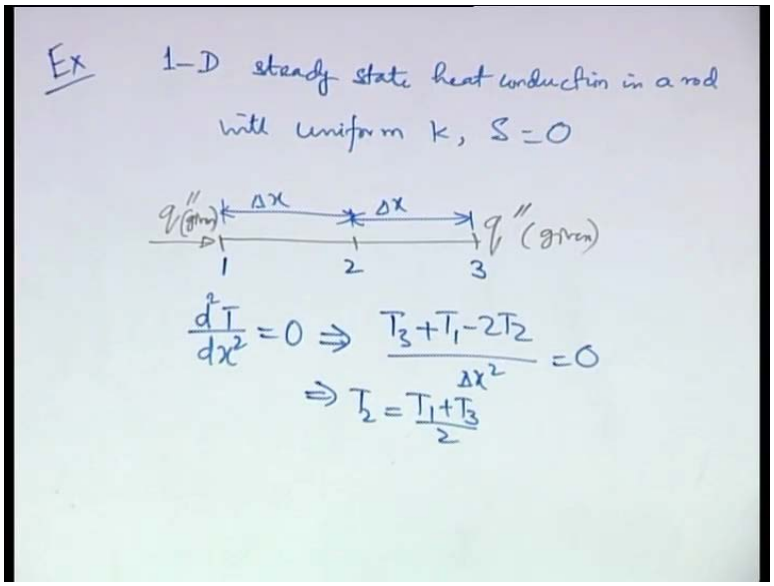
(Refer Slide Time: 18:58)



So, Scarborough criteria for sufficient condition for convergence Scarborough criteria of sufficient condition for convergence in Gauss Seidel method it states that you must have summation of mod of a n b by mod of a p that is less than equal to one for all equations and in addition one physically based criterion that is less than 1 for at least one equation this is a very important consideration which we will try to discuss in more details.

So, from the mathematical perspective that we have discussed. So, far less than equal to one for all equations would have been good enough because; that means, that if it is less than equal to one for all equations; that means, for all rows considered together this is less than equal to one. So, maximum row sum norm that is less than equal to one now why less than 1 for at least one equation. So, the criterion says that it is not good enough to satisfy this less than equal to one that inequality should be a strict inequality in a sense that it should be less than 1 for at least one of the equations that you are solving for.

(Refer Slide Time: 21:39)



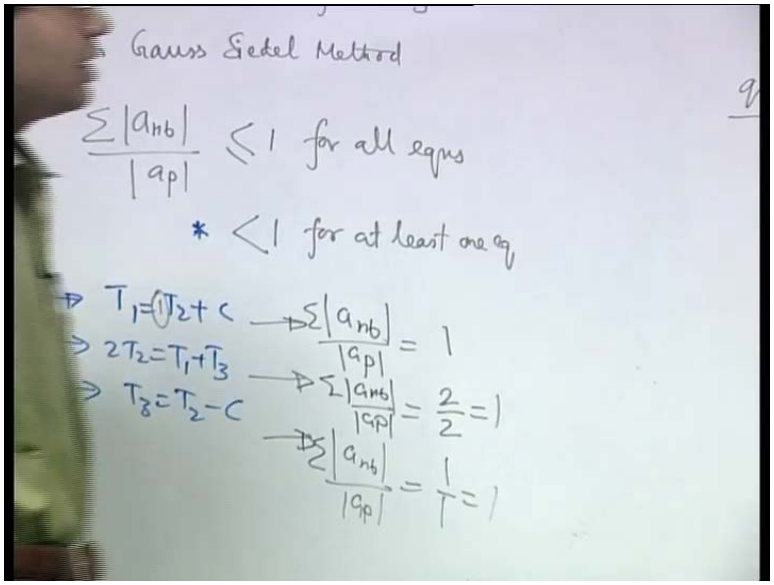
To understand why it is necessary let us go back to one of our old examples of 1 dimensional steady state heat conduction in a rod with uniform thermal conductivity and the source term equal to 0. The Scarborough criteria is generally I mean because people use Gauss Seidel method more of10 than the Jacobi's method it is in10ded for the Gauss Seidel method, but with an understanding that its essential conception is based on the understanding based on the theoretical derivations we have for the Jacobi's method. Why it is based on the understanding from the Jacobi's method is because, the same evaluation is more tedious for the Gauss Seidel method by virtue of the form of the matrix; but at the same time because the Gauss Seidel method is supposed to act in a more efficient way than the Jacobi's method.

So, regarding the convergence whatever you can derive for the Jacobi's method in a in some conservative way that can be applied even to the Gauss Seidel method now let us say that we have a rod of some length one where we are specifying the heat flux boundary condition at both the sides and they are equal. Obviously, if you have a 1 dimensional heat transfer then, if it is at steady state and no heat is generated inside. Whatever is the heart flux enter from the left face should leave the right face now we number the grid points one two and three and write the corresponding discredited equation. So, for the grid point two we have done it a few times earlier. So, let us quickly try to do it.

The governing equation is d 2 d d x 2 equal to 0. So, if you use a central difference then t 3 plus T 1 minus 2 T 2 by if you call this as delta x this as delta x by delta x square that is zero. So, you have T 2 is equal to T 1 plus T 3 by 2 then for the grid points 1 q is equal to k into T 2 minus T 1 by delta x.

So, we have T 1 is equal to T 2 sorry T 1 minus T 2 by delta x not we can write a minus sign here minus k into d t d x. So, T 1 is equal to T 2 plus q delta x by k just for symbolic notation let us call it as c sum constant similarly at the right boundary you have q is equal to k into T 2 minus T 3 by delta x; that means, T 3 is equal to T 2 minus c.

(Refer Slide Time: 26:56)



Now, let us try to arrange or organize these equations. So, for what is our what are our equations? Equation 1 you have T 1 is equal to T 2 plus c; equation 2: 2 T 2 is equal to T 1 plus T 3; equation 3: T 3 is equal to T 2 minus c. So, what is the corresponding summation of sigma summation of mod of a n b by a p for equation 1? How many naives are there? one naive is there and the corresponding naive ring coefficient is 1 and a p is a 1 that is 1. So, 1 by 1 this is equal to 1.

This one, so, 1 plus 1 2 by 2 a p is 2. So, that is 1; this 1, again 1 and we know that this problem is a real pose problem. We have seen by earlier experience that this is a real pose problem because for such a problem you require to specify at least at 1 boundary that value of the temperature because otherwise you are your specification of the heat flux in equal to heat flux is already inbuilt with the governing differential equation it

gives no additional or in the terms of in terminology of mathematics no linearly independent boundary condition if you say that this heat flux and this heat flux they are the same and whatever is entering is whatever is leaving.

If you say that these 2 heat fluxes are different then of course, you are violating the basic physics of the problem of the heat the steady state heat transfer. So, what you can see is that this is real pose problem that we have already figured out now if you see that if you just satisfy the mathematical criteria all of these are less than equal to one one is of course, less than equal to one, but none of these are less than 1. So, how can you make it less than 1.

Let us say that you specify the value the temperature instead of the heat flux at the left boundary. So, you specify in place of q given you have t given if you have t given let us say t equal to some t star then this boundary condition will not be the boundary condition for grid point one for grid point one the boundary condition is t star.

For grid point two now you can write T 2 is equal to. So, there is there is this equation 1 is not pertinent for grid point 2 you have 2 T 2 equal to t 1 plus t 3, but T 1 is not a variable any more it is now t star which is known and for equation 3 whatever it is it remains.

So, what now gets modified is for this equation 2 which is new equation 1 what is now mod of sigma of mod of a n b by a p it is 1 by 2 because now T 1 does no more represent a naive in a sense that naive is represented conceptually by something which is an unknown now you have already substituted a known here. So, it is not a coefficient it is of course, physically a navies, but mathematically it is no more a coefficient. So, mathematically it is no more a naive mathematically the only naive is the point 3.

So, it is 1 by 2 and that means, it is less than 1. So, now, you can see that with with redefinition of this problem it satisfies the Scarborough criteria. So, the whole understanding we can derive out of this that how nicely or beautifully physics follows, mathematics or mathematics follows physics. So, what we can see here is that there is some criterion which from a mathematician's point of view will come from some linear algebra which does not understand that whether there is 1 dimensional heat transfer steady state unsteady state heat source is there whether isothermal or heat flux boundary condition these are irrelevant for this criterion

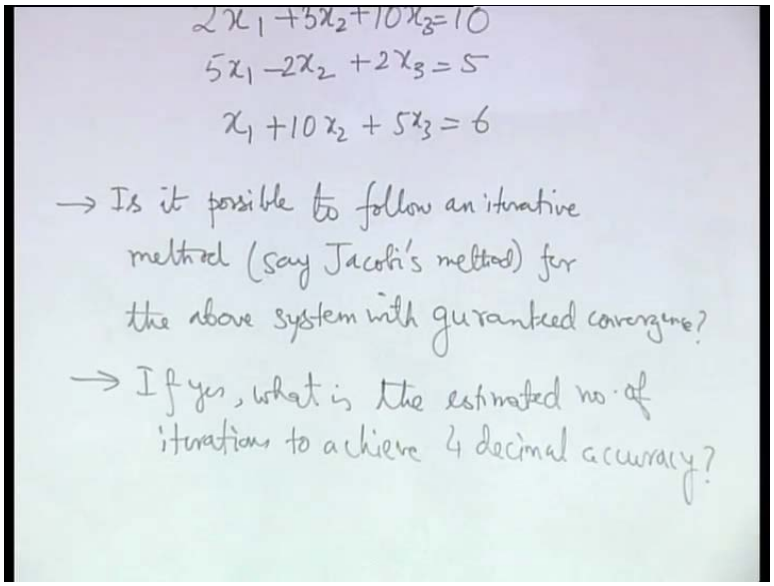
This is not a criterion made for solving a heat transfer problem in particular, but if you look at the physics of the problem you can see that how by imposing the physics you disturb this criteria or you violate this criteria and by we posing the physics you can make the criterion satisfied and that is very important. So, whenever we solve a problem using the Gauss Seidel method say we need to check this criterion and if this criterion is violated it most of the times we mean not an error in the discretization, but error in the problem formulation or specification itself in terms of giving the appropriate boundary conditions.

That is where see in c f d most of the times people fail in solving the problem because of the error in correctly giving boundary condition because most of the times you are solving the same equation. So, either it is a diffusion type of equation then we will see in our next chapter that it could be convection diffusion type of equation and then we will see that it could be the naive's strokes equation the fluid flow equation.

So, these are some of the equations which hardly get changed as you go from one problem to the other of the similar type, but what gets changed from one problem to the other is a correct specification of the boundary condition and if you violate certain rules or certain physical considerations in specifying the boundary condition that is where you come up with trouble.

Sometimes, if you are not physically that intuitive you may not be able to detect that in the formulation stage, but if you look into the system of algebraic equations that you get at the end and try to assess those mathematically in terms of their consis10cy there itself you are you will be able to detect that there must have been something that has gone wrong in some of your earlier stages of the formulation

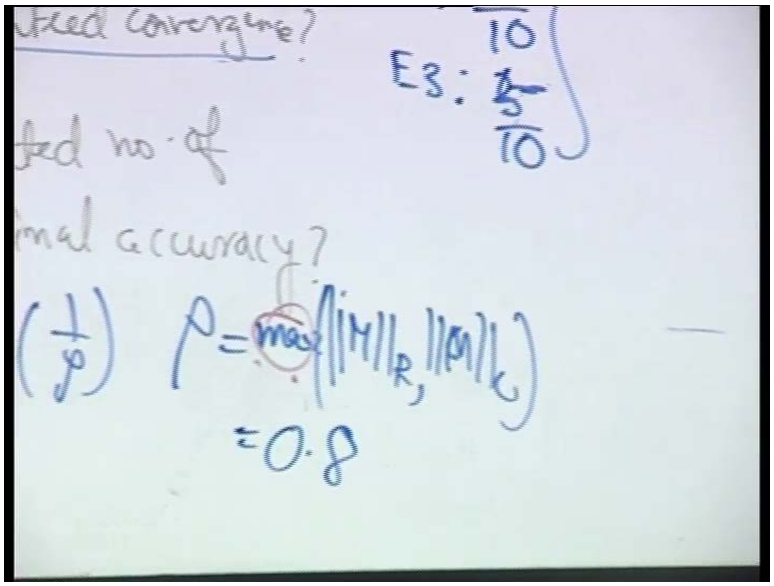
(Refer Slide Time: 35:01)



So, with this understanding let us try to work out a couple of examples to illustrate what we have learnt. So far in terms of sufficient terms for convergence for the iterative methods, consider the following system 2 x 1 plus 3 x 2 plus 10 x 3 equal to 10 5 x 1 minus 2 x 2 plus 2 x 3 equal to 5 x 1 plus 10 x 2 plus 5 x 3 equal to 6.

Question is it possible to follow an iterative method say Jacobi iteration for the above system with guaranteed convergence if yes what is the estimated number of iterations to achieve 4 decimal accuracy now what is the first thing that we should look for before assessing the method in more details that whether it satisfies the sufficient conditions for convergence then the question comes what is rate of convergence and all those things.

(Refer Slide Time: 38:12)



So, to satisfy the to assess whether it satisfies the sufficient condition requirement of sufficient condition for convergence what we have to assess we have to assess the norm of m; that means, essentially we can take as an example a row sum or a column sum norm. So, if we take an example of a row sum norm what we expect is that summation of a i j where j is not equal to i mod of that by a i i that is less than 1 less than equal to one of course, less than 1 equal to one equation, but in general less than equal to 1.

Now, what does that it mean? Now, if you are given two choices one is a small diagonal coefficient another is a large diagonal coefficient which one would you prefer for satisfying this large diagonal coefficient. So, you can see if it is the large diagonal coefficient then that can over way sum of the magnitude of the off diagonal coefficients. So, that this is constrained to be less than 1.

So, you can see the this is this is very beautiful if you if you appreciate that it is in a different way in a different words talking about the diagonal dominance see when we were studying the elimination method we were talking about the diagonal dominance how to re organize the system of equations to make it diagonally dominant. So, that the diagonal element is much more dominant in that particular pivotal row that you are considering for Gaussian elimination with pivoting now you can see that sense has come back maybe through different mathematical roots, but even for the iteration method.

This is essentially what is diagonal dominance, that the diagonal term is dominating over the sum of the magnitude of the off diagonal terms. So, the diagonal term is. So, dominating that even if you add the magnitude of all the off diagonal terms that cannot supersede the diagonal term magnitude by itself. So, that is what is absolute dominance of the diagonal term or the diagonal dominance.

So, let us try to figure out that what is the situation here for the first row for the row one what is the case. So, the diagonal the sum of the off diagonal terms is two plus 3 plus 10 that is thirteen and the diagonal term is two. So, straight away this does not satisfy the sufficient condition for convergence. So, as I have said earlier that that does not tell you that it will not converge because may be you are. So, lucky that you have already got a clue of the solution by magic or whatever and you know the solution and you start iterating that will satisfy the equation automatically, but we are not talking about such fortunate cases.

So, we are talking about cases when that will not occur and you are you are dealing with an arbitrary initial case not a guess which is the solution. So, then this will not give you guaranteed convergence, but one may attempt and see that whether a reordering of this equations can do that. So, what is the rationale behind the reordering.

You first out that out of these 3 equations which has the first coefficient highest that is the second equation as the first 5 which is the highest and sum of the off diagonal terms in that case is two plus two which is 4 which is less than 5. So, let us make this as equation one which has the second coefficient as highest the third one. So, let us make it equation two and the first one let us make it equation 3.

If you make it in this way you see that this criterion is satisfied by all the 3 equations. So, the answer to the first question is yes it is possible to follow an iterative method for the above system with guaranteed convergence, but after reordering the equations not in the original order.

Then the next part is very straight forward let us not move into it more details we will just use the formula k greater than m by the rate of convergence. So, for the rate of convergence that is log 10 one by rho and for rho you can take as maximum of either the row sum norm or the column sum norm.

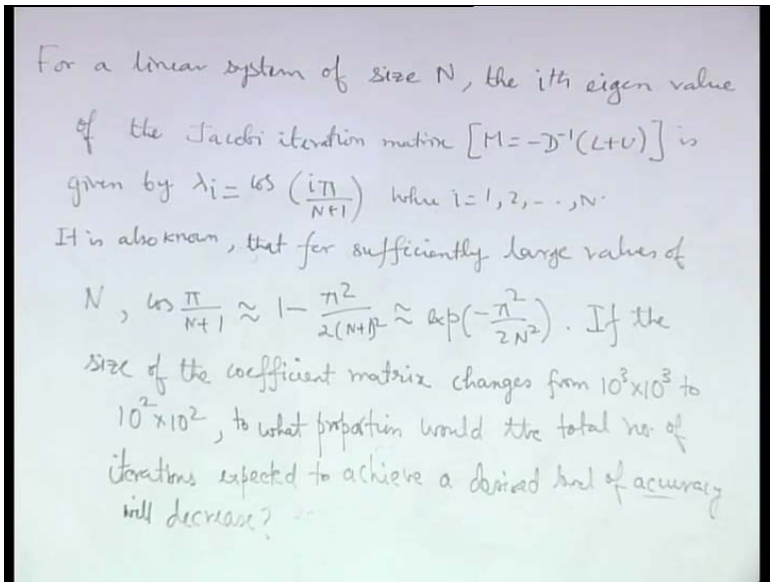
Thus let us see how to calculate these norms of course, we have we have seen here that how to calculate the norms through non numerical examples, but just through this norm numerical examples. So, out of all the rows. So, the row sum what will be the row sum norm for the first row it is. So, the first row is corresponding to this equation one. So, that is two plus two 4 divided by 5.

Sum of the magnitude of the off diagonal divided by the diagonal for the second 1 it is 5 plus one that is 6 by 10 for the third 1 2 plus 3 that is 5 by 10. So, out of this maximum is 4 by 5. So, that is point eight. So, the row sum norm is point eight. Similarly, you can calculate the column sum norm also and if you want to make a conservative estimate out of actual if you want to make a conservative estimate of the number of iterations instead of using the max here you can use the minimum here if you want to be very conservative because or not no sorry.

If want to be conservative if you want to make row as large as possible not the small one because small row is better any way. So, you you want to see what is the largest row that you can take. So, out of these two. So, you can also consider the column sum norm what is the column sum norm. So, for the first column first column what is the. So, let us just try to organize the columns a bit. So, 5 minus two two that is the first row second row 1 10 5 third row two 3 10

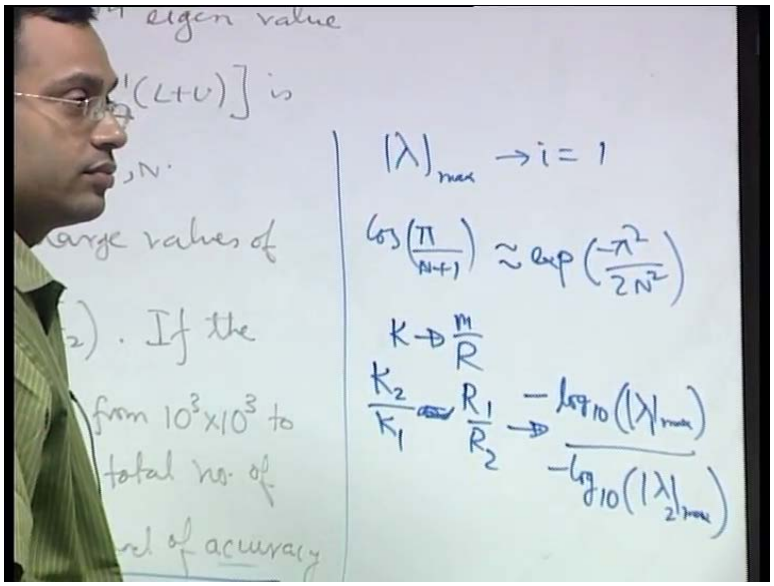
So, for the first it is 3 by 5 column one it is 3 by 5 column two 5 by 10 column 3 7 by 10. So, the max is point 7. So, if you take the maximum of the row sum and the column sum norm that comes out to be point eight and based on that you can calculate k that is an estimate of the number of iterations to achieve the 4 decimal accuracy

(Refer Slide Time: 47:08)



Let us consider one more example for a linear system of size n the ith Eigen value of the Jacobi iteration matrix that is m is equal to minus d inverse l plus u is given by lambda i is equal to cos i pi by n plus 1 where I is equal 1 2 up to n it is also known that for sufficiently large values of n cos pi by n plus 1 is roughly equal to 1 minus pi square by 2 into n plus 1 whole square is again roughly equal to e to the power of minus pi square by 2 n square if the size of coefficient matrix changes from 10 to the power 3 by 10 to the power 2 by 10 to the power 2 to what proportion would the total number of iterations expected to achieve a desired level of accuracy will decrease.

(Refer Slide Time: 50:56)



So, this in this problem you are directly given the information on Eigen value and because it is a cosine function the maximum will be when the argument of the cosine function is the minimum out of all possible ones. So, you can estimate mod lambda max when you consider i equal to one. So, that is cos pi by n plus one magnitude of that

Remember n is the size of the coefficient matrix and for the problem it is given that the size of the coefficient matrix in one case is 10 to the power 3 by 10 to the power 3 and another case 10 to the power two by 10 to the power 2. Both these may be considered to be large enough. So, that this approximation formula for cos can be used. So, this becomes approximately exponential minus pi square by two n square now what is how do you estimate the total number of iterations to achieve a desired level of accuracy you estimate it through the rate of convergence.

So, your k estimation is m by r estimation remember that as you are decreasing the size of the coefficient matrix from 10 to the power 3 by 10 to the power 3 to 10 to the power two by 10 to the power two the accuracy level expectation is not changing to achieve a desired level of accuracy. So; that means, the number of iterations k 2 by k 1 for the two cases will be inversely proportional to r that is it will become R 1 by R 2 where r is the rate of convergence remember this is not exactly equal this is just an estimate what is the rate of convergence for the first case. So, log 10. So, minus log 10 of mod lambda max in case of one divided by minus log 10 of mod lambda 2 max 2 for case 2.

So, for case 1 you have to substitute n equal to 10 to the power 3 and for case 2 you have to substitute n equal to 10 to the power 2 in the formula for mod lambda max. if you do that then, you will get this approximately as 10 to the power minus 2; this you can verify later on through calculations.

But important thing is the understanding. So, what does it give you? It gives you an estimate of what it gives you. An estimate of the reduced number of calculations that you need to do. As you decrease the size of your coefficient matrix to achieve the same accuracy, many times one needs to plan for the computational time. So, many times one needs to plan how many computational hours that you need to devote for solving a problem. That depends on the size of the matrix because computational time requirement in an iterative scheme depends on number of iterations each iteration will have a fixed number of calculations.

So, if you can figure out that what would be the modified number of iterations with a change in the coefficient matrix size then, you can scale the computational time with the size of the coefficient matrix and this simple understanding allows you to do that. We will, our next agenda will be to work out one or two more examples related to the use of iteration method and then we will consider the use of the gradient search method for solving the system of algebraic equations. So, that we will take up in the next lecture; thank you.