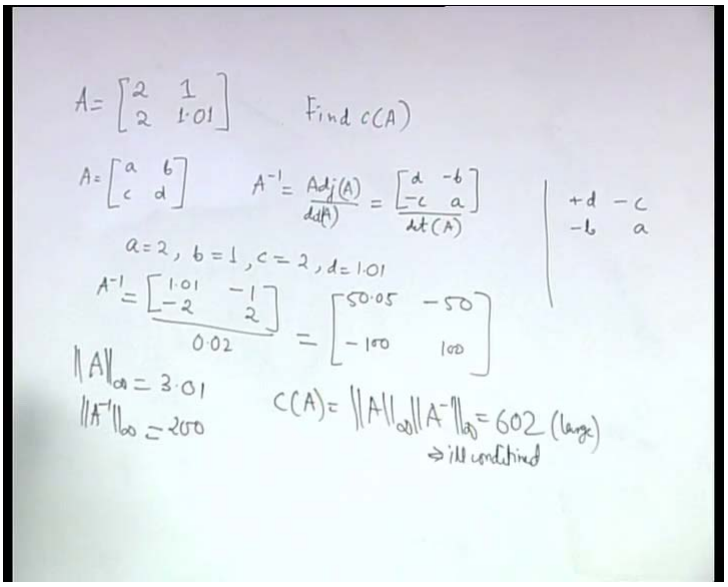
Computational Fluid Dynamics
Prof. Dr. Suman Chakraborty
Department of Mechanical Engineering
Indian Institute of Technology, Kharagpur

Lecture No. # 25
Iterative Methods for Numerical Solution of Systems of Linear Algebraic Equations

In the previous lecture, we were discussing about the solution of algebraic equations using elimination methods and some of their important features. Before going into the next topic, we will like to work out one example, related to what we have discussed in the previous lecture.

(Refer Slide Time: 00:42)



So, example is like this. You are given a co-efficient matrix A. Find the condition number of A. So, if you recall for the condition number of A, we require the norm of A and norm of A inverse. So, we need to calculate what is A inverse. So, if you have A as a collection of 4 numbers, this matrix, then what is A inverse. Adjoint of A by the determinant of A.

So, for adjoint of A, what you do? First, you find out the co-factors corresponding to each. So, let us do that separately. So, for the first term, the co-factor is minus d. So, it is minus 1 to the power i plus j. So, minus 1 to the power. So, only d minus 1 to the power

1 plus 1. So, only d. Then, for the next, 1 minus c. Then, the third, 1 minus b and the 4th 1 a.

Now, what we do, we find out the transpose of this. So, we interchange the rows and the columns. So, we have this as d minus b minus c A by the determinant of A. So, here, what is given a is 2, b is 1, c is 2 and d is 1.01. So, A inverse is equal to 1.01. Then, minus 1, then minus 2, then 2 divided by the determinant of A, that is 0.02.

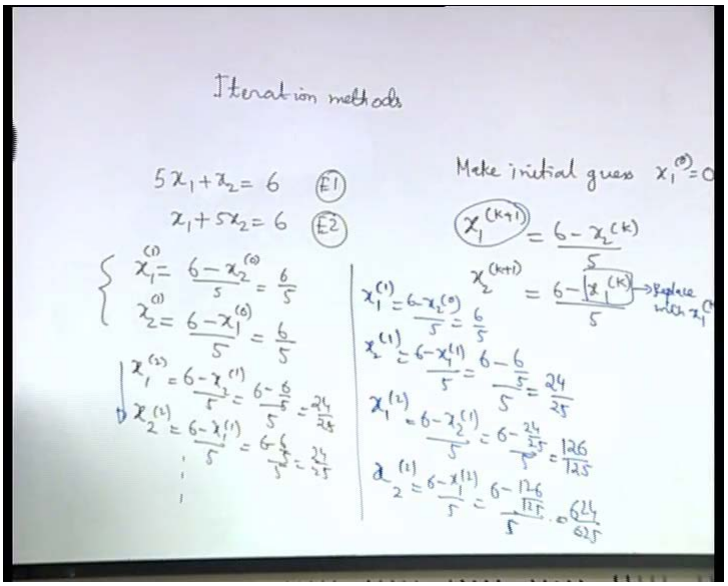
So, if you calculate these terms, this will become 50.05. Then, minus 50, then minus 100 and 100. So, for finding out the condition number, you can calculate different types of norms, but one can use the infinity norm as one of the standard reference norms that people commonly use. So, the infinity norm for A. What is that infinity norm? It is maximum row sum norm.

So, out of the 2 rows, which row gives the maximum sum of the magnitude of the individual numbers? So, the second row, that is 3.01 infinity norm for A inverse. Here also, the second row gives the maximum. It is 100 plus 100, that is 200. So, the condition number, if you evaluate in terms of the infinity norm, this is 602. So, this is quite large, which means that it is an ill-conditioned system. You can figure out the source of this largeness. The source of this largeness is the smallness of the determinan. It is divided by the determinant because the determinant is small. A inverse is blown up and therefore, its norm are blown up.

So, smaller the determinant there, it is greater vulnerability of the matrix of becoming more ill-conditioned. That is one of the important inferences that we can draw from this particular example. Next, we will move on to the iteration methods. So, the condition number. What you are saying is that what is the condition number? Which we can call large, the condition number that we have as the minimum is 1.

So, if you now see, that what can you call as large. It is about two orders of magnitude higher than this smallest one that you can consider as large. So, of course, you can have any possible limit as your large number because the determinant of the matrix can go very close to 0. Therefore, one by that can become a very large number, but at least 1 or 2 or at least 1 order of magnitude higher sometimes is still 2 orders of magnitude higher is considered to be a large 1. So, this typically, the number that we encounter with, that is typically a large number.

(Refer Slide Time: 08:05)



Now, next we will go on to the iteration methods. So, iteration methods. What is the basic philosophy of the iteration methods? So, you start with an initial guess for the solution and iterate on it till you get a final solution which is a converging solution.

So, you expect that final solution will converge. We will see whether the method itself guarantees. Whether the final solution will converge or there are certain conditions that need to be fulfilled. So, we will go into that subsequently, but let us try to understand the iteration methods through some examples.

Let us say, that you have the system of equations. Let us say, we have these 2 equations. So, what we will do is, we will write an explicit formula for x 1 from equation 1. So, x 1 is 6 minus x 2 by 5 and we will write an explicit formula for x 2 from equation 2. So, x 2 will be 6 minus x 1 by 5 .

Now, to begin with, you do not know, what is x 1 or what is x 2. So, let us say, that you make an initial guess. Some initial guess, say x 1 0 equal to 0 and x 2 0 equal to 0. So, the super slip indicates the level of iteration 0, means, that is the initial guess.

Now, you try to update on this initial guess. So, what you do? You make some iterative formula based on these equations that you have got from your given system. So, let us say, that you have x 1 k or k plus 1 is equal to 6 minus x 2 k by 5 and x 2 k plus 1 is equal to 6 minus x 1 k by 5.

If we write the corresponding formula in this way, then this is known as Jacobi's iteration scheme or Jacobi's method. So, let us find out what is x 1 1. So, x 1 1 is 6 minus x 2 0 by 5. So, 6 by 5x 2 0 is 0 then x 2 1 is 6 minus x 1 0 by 5. So, that is 6 by 5. So, that is step 1.

Then, step 2, x 1 2 is equal to 6 minus x 2 1 by 5. So, 6 minus 6 by 5 by 5 24 by 20 5x 2 1, sorry x 2 2 is 6 minus x 1 1 by 5. So, 6 minus 6 by 5 by 5, that is 24 by 20 5. In this way, if you do a few steps, you will see that x 1 and x 2, they are converging to 1 and 1. When you say, when you declare convergence, you declare convergence when the difference in results between the previous step and the present step does not vary substantially. Now, what is that substantial? It all depends on the precession that you expect. So, if you give a 3 decimal accuracy, 4 decimal accuracy, 5 decimal accuracy, 10 decimal, whatever is the number of decimal accuracy that you yourself specify. If you find that, that difference in successive iteration falls within that tolerance, then you accept that as a converge solution.

Now, what we see in the Jacobi's method is something very interesting and which is perhaps, one of the issues that tends to slow down the method a bit. So, when you are calculating x 1 k plus 1, you can use x 2 k. When you are calculating x 2 k plus 1 x 1 k plus 1 is already available to you, but you are still using x 1 k. You have the opportunity of using that updated x 1, which has been calculated just in the previous step.

So, if you do that, you may have your solution having the information of the more updated iterated value. Then, what was possible if you had used the older value. So, if you want to update your iterations at a more first rate, then you may replace this with x 1 k plus 1 instead of x 1 k because x 1 k plus 1 is already available. So, if you make this modification to the Jacobi's method, then that is known as Gauss-Siedel method.

So, let us try to use the Gauss-Siedel method for this particular example and see that what is the difference that we get. So, x 1 1 is equal to 6 minus x 2 0 by 5, that is 6 by 5 x 2 1. Now, you are using 6 minus x 1 1 by 5. So, 6 minus 6 by 5 by 5. So, this is 24 by 25.

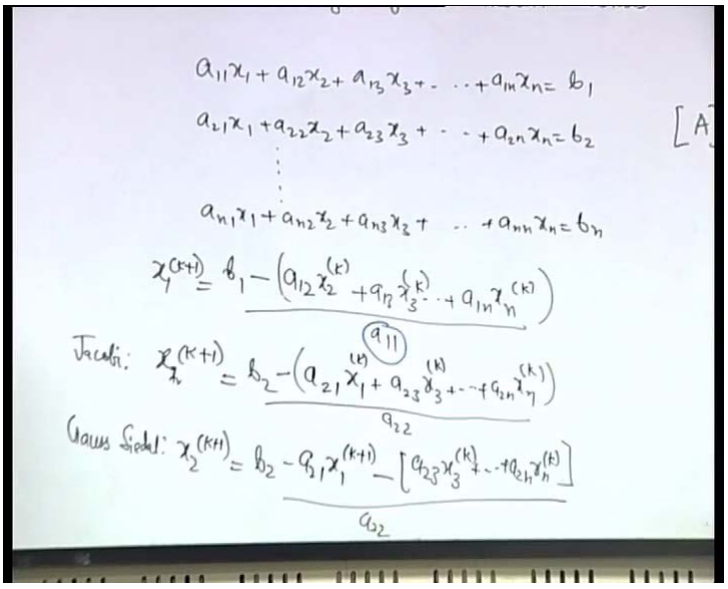
You can see that, what you have got in one more step here, you have already now got it one step before. So, next you calculate one more step, just for example,. So, x 1 2 is equal to 6 minus x 2 1 by 5. So, 6 minus 24 by 25 by 5. So, that is 126 by 125 and x 2 2

is equal to 6 minus x 1 2 by 5. So, 6 minus 126 by 125 by 5, whatever it will be a number close to 624 by 625.

So, you can see, that the numbers in the subsequent steps are going faster close to 1, which is the correct solution here. Also, the numbers are converging to 1, but at a much slower rate here. Within 2 steps, it is converging much more rapidly towards the solution. So, no matter whether, you are using the Jacobi's method or the Gauss-Siedel method, you have a hope that you will converge to the solution and it does. So, the Gauss-Siedel will do it in a faster way. Now, we have to understand that, we have just a hope because we have not derive this formula from a very rigorous consideration.

So, at this stage, we have just a hope that it will converge to the correct solution, but where is the guarantee that it will finally converge to the correct solution. There is no guarantee as such and we have to establish that, what can be the condition which should dictate, whether it is expected to converge or not. So, we will try to have a generalized analysis of the Jacobi's method and the Gauss-Siedel method.

(Refer Slide Time: 19:43)



So, generalized analysis of the iterative methods. Let us say, that we have the system of equations a 11 x 1 plus a 12 x 2 plus a 13 x 3 in this way upto a 1n x n is equal to b 1. Then, a 21 x 1 plus a 22 x 2 plus a 23 x 3. In this way, plus a 2n x n is equal to b 2. Then, a n1 x 1 plus a n2 x 2 plus a n3 x 3 plus a nn x n is equal to b n. So, let us try to write this particular expression in the matrix form a x equal to b. To do that, we keep in mind the

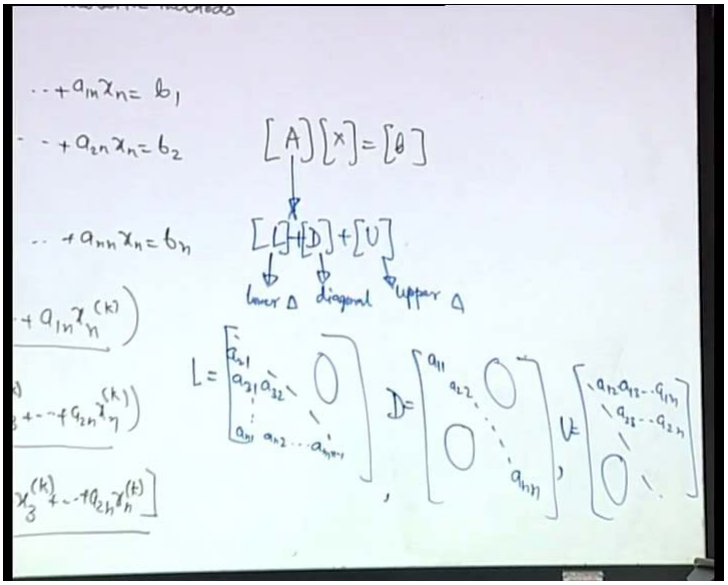
following iterative formula. So, x 1 k plus 1 is equal to b 1 minus a 12 x 2 k minus a 13 x 3 k. In this way, a 1n x n k divided by a 11.

So, you can see that in the formula, the diagonal element a 11 gets multiplied with the variable that you are interested to solve and the off diagonal terms are distributed in a certain way. So, in the first step, all the off diagonal terms there, they are together. Next step. The next step will deform, based on whether, it is Jacobi's method or Gauss-Siedel method. So, let us write the next step by considering one example Jacobi, another example Gauss-Siedel.

So, Jacobi's method x 2 k plus 1 is equal to b 2 minus a 21 x 1 minus, then plus a 23 x 3. Then, in this way, plus a 2n x n all with k divided by a 22. Whereas, in the Gauss-Siedel, it is x 2 k plus 1 is equal to b 2 minus a 21 x 1 k plus 1 and the remaining are with k. So, a 23 x 3. In this way, a 2n x n k by a 22.

Now, the question is that, can we write the Jacobi and the Gauss-Siedel formula in a generic universal way. So, if you see, you can have 3 different types of terms in the most general case. In the numerator, one is the right hand side, then the term on one side of the diagonal. In the second row, a 21 is on one side of the diagonal and a 23 a 24, these are in the other side of the diagonal. The diagonal term comes in the denominator. So, if you multiply the diagonal with the variable and take all the terms in one side together.

(Refer Slide Time: 25:43)

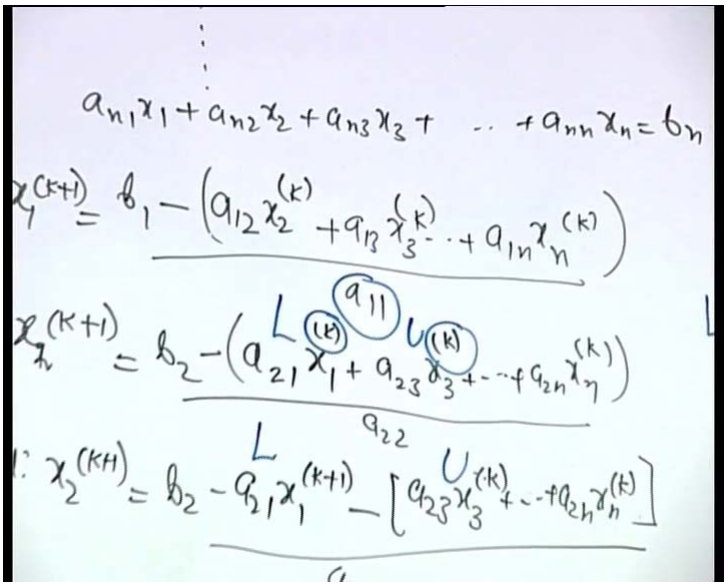


Then, as if you have now decomposed A as a summation of 3 different matrixes L plus D plus U, where L is a lower triangular matrix D is a diagonal matrix and U is an upper triangular matrix. Do not confuse it with LU factorization. It is simply addition of 3 different matrixes. So, that the final result is the matrix A. So, what are the contents of L?

So, L will have only the lower diagonal term. So, a 21, then a 31, a 32, then a n1 a n2. In this way, a nn minus 1. This is L. What is D? D is a 11, a 22, a nn just a diagonal matrix and 0s on both sides of the diagonal. What is U? U is a 12, a 13. In this way, a 1n, then a 23. In this way, a 2n and so on. So, if you add L, D and U, you will get the coefficient matrix A.

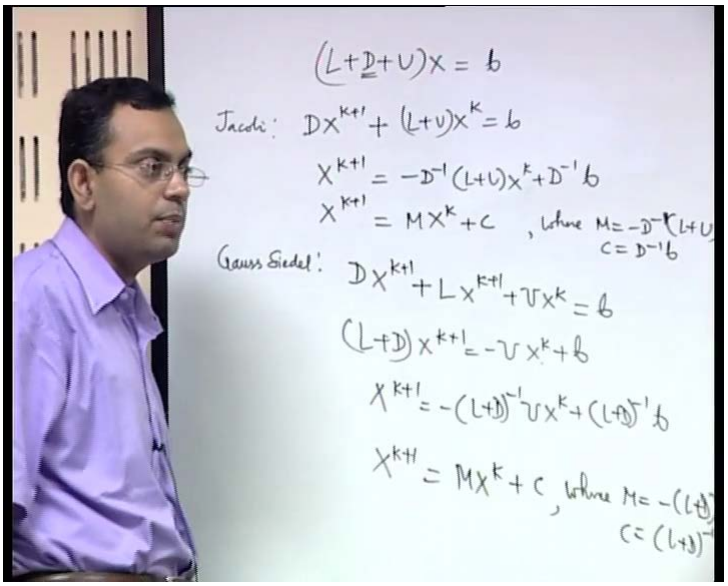
So, let us try to identify that, where do we get this L, D and U in the iteration formula. So, where do we get D? It is straight forward here in the denominator. What is appearing, that is the contribution from D in the numerator. You can see in the Jacobi's formula, you have a 21 a 23, all these. So, this is what. So, you have a 21 corresponding to lower a 23. These are corresponding to upper.

(Refer Slide Time: 28:34)



So, you have the lower corresponding to x k and upper also corresponding to x k. In the Gauss-Siedel formula, you have the lower triangular matrix corresponding to x k plus 1 and the upper triangular matrix corresponding to x k.

(Refer Slide Time: 29:20)

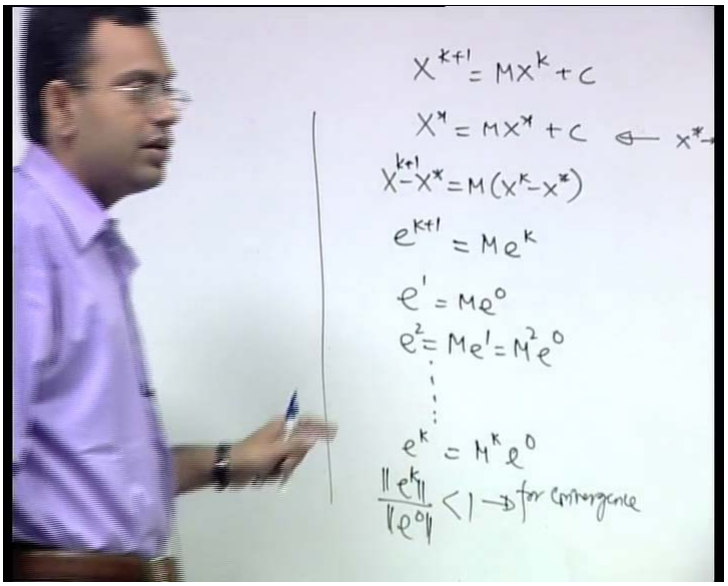


So, let us write the formula in a more abridged manner. So, in place of A x equal to b, we will write L plus D plus U x equal to b. In the Jacobi's method, you have D x k plus 1. D will go with x k plus 1 and L plus U will go with x k is equal to b. So, x k plus 1 is equal to minus D inverse L plus U x k plus D inverse b.

Remember, that the super scripts are the iteration stages. So, these are not powers, just iteration stages. So, this in a short form, you can write x k plus 1 is equal to M X k plus c, where M is equal to minus D inverse L plus U and c is equal to D inverse b. Let us now try to do it for the Gauss-Siedel. In the Gauss-Siedel, D was with x k plus 1. L was also with x k plus 1. U was with x k. So, D x k plus 1 is equal to minus L, sorry D plus L x k plus 1 or L plus D. Whatever L plus D x k plus 1 is equal to minus U x k plus b. This means x k plus 1 is equal to minus L plus D inverse U x k plus L plus D inverse b.

So, x k plus 1 is equal to M X k plus c also corresponds to the form of the Gauss-Siedel iteration scheme, where M is minus L plus D inverse U and c is L plus D inverse b. So, why we have done this unification is that, now if we want to make an analysis of these methods, then we can make an analysis using the generic form x k plus 1 equal to M X k plus c, where the matrix M just differs from one method to the other, but the generic form remains unaltered.

(Refer Slide Time: 34:00)



So, what we will do is, we will try to make an analysis of these methods accordingly. So, X k plus 1 is equal to MX k plus C. Now, let us say that X star is the actual solution. So**,** **X** star is equal to MX star plus C. So, actual solution should satisfy the iteration, where there is no change in the value of the variable with new iteration, that is when it has converged. So, X k plus 1 and X k both will be X star.

So, if you subtract these 2, then X minus X star is equal to MX k minus X star. This is X k plus 1. So, this is the error in the k plus 1th step. E k plus 1, this is equal to M e k, where e represents the difference between X at a particular iteration level and the actual solution X. So, from this, we can write certain important points. One point is, we can first write, e 1 is equal to M e 0. So, e 2 is M e 1, that is equal to M square e 0.

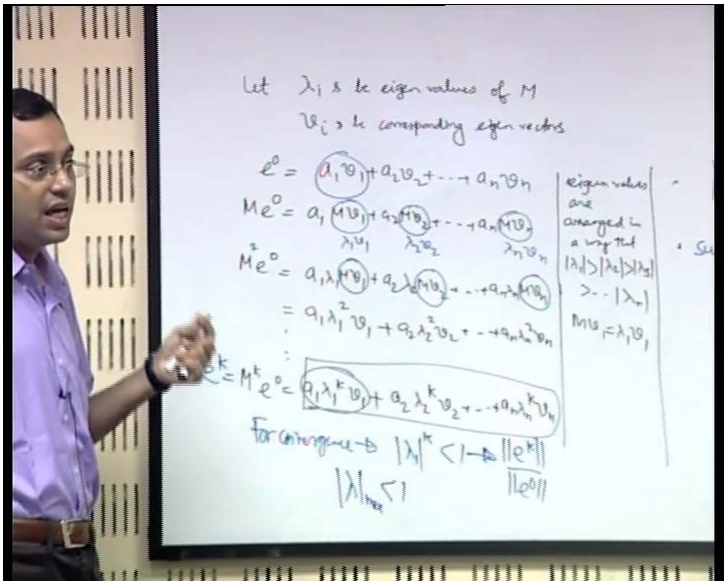
In this way, if you come down, E k is equal to M to the power k e 0. Sorry, yes e k is equal to M to the power k e 0 e 2 is M to the power 2 e 0. So, e k is M to the power k e 0. So, when do you expect the solution to converse? You expect the solution to converge, when in the subsequent steps, the error goes down. So, you must have error in the kth step, possibly the norm of that because this error is a vector divided by the error in the 0th step, that is to begin with this ratio must be less than 1 at least. If that is not the case, then your error is getting magnified. Whatever was the error because of an initial guess that gets magnified with iteration, then it is diverging.

So, this is the requirement for convergence. So, that will require that the norm of M to the power k must be less than 1. Now, it is very difficult to use that as a method of assessment. Why? Because, it is not very computationally inexpensive mechanism to calculate the power of a matrix. So, first of all, you have to determine the matrix M. Remember, M is not the co-efficient matrix A. It is derived from the co-efficient matrix. It involves some L, D, U all these.

So, M requires some matrix inversion in certain cases. So, you can see for example, if you want to calculate M for the Gauss-Siedel method, you require L plus D inverse. So, matrix inversion itself is computationally expensive. So, from that, you find M. Then, put that to the power k. So, to have an estimate of whether, it will converge or not, the kind of steps that you need to go through is more than what it deserves for.

So, the objective now, will be to keep this particular inference or conclusion in mind, but based on that, try to achieve some more simple way of assessing the possibility of convergence. That we will try to do in the subsequent steps. Now, we know that power of a matrix may be equivalently expressed in terms of scalars by using the Eigen values and Eigen vectors. That is what we will try to exploit here.

(Refer Slide Time: 40:06)



So, what we will consider, that let lambda i's be Eigen values of M and v i's be the corresponding Eigen vectors. Now, let us form the vector e 0 as a linear combination of the Eigen vectors corresponding to m. So, we can form this as a 0. May be, we use start

from super from subscript 1. So, a 1 v 1 plus a 2 v 2 plus, in this way a n v n. So, we make e 0 as a linear combination of the Eigen vectors of the matrix M. We can make it in this way, by using the co-efficients a 1 a 2 upto a n judiciously. So, these are according to our particular choice. So, that we can write it in the way.

While, writing it in this way, we are keeping one thing in mind, that the Eigen values are arranged in a way that, mod of lambda 1 greater than mod of lambda 2 greater than mod of lambda 3 greater than mod of lambda n. So, lambda 1 is the Eigen value with the maximum magnitude. Out of all possible Eigen values, lambda 1 is the maximum. So, we are arranging the Eigen values in the descending order of magnitude.

Now, M e 0 is equal to a 1. So, we are interested to calculate M to the power k e 0 because that is e k. So, M e 0 is a 1, M v 1 plus a 2 M v 2. In this way, plus a n M v n. So, now, because M is an Eigen value, sorry lambda 1 is an Eigen value of M and b 1 is the corresponding Eigen vector. We can write, M v 1 is equal to lambda 1 v 1 because lambda is the Eigen value and v is the corresponding Eigen vector of M.

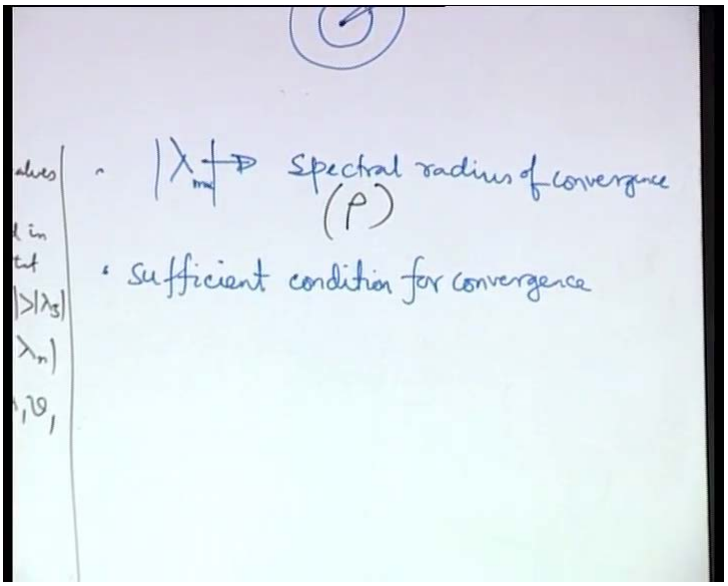
So, in place of M v 1, we will write lambda 1 v 1. In place of M v 2, we will write lambda 2 v 2. In place of M v n, we will write lambda n v n, the next step that into M. So, M square e 0 is equal to a 1 lambda 1 M v 1 plus a 2 lambda 2 M v 2. In this way, a n lambda n M v n again in place of M v 1, we can write lambda 1 v 1 in place of M v 2 lambda 2 v 2 M v n lambda n v n.

So, this becomes a 1 lambda 1 square v 1 plus a 2 lambda 2 square v 2. In this way, plus a n lambda n square v n. Proceeding in this way, M to the power k e 0 is equal to a 1 lambda 1 to the power k v 1 plus a 2 lambda 2 to the power k v 2. In this way, plus a n lambda n to the power k v n.

Now, if we compare M to the power k e 0 with e 0 as if we are dividing this entire expression with e 0 in terms of the norm. So, when we make such a division by taking the norms, then what will be the dominant terms? See, lambda 1 is the maximum Eigen value and you are making it to the power k, where k is large. So, the first term will be the leading order or the dominant term and that you can compare with the leading order term or the dominant term of e 0. If you do that, you will see that what you require for convergence is that mod of lambda 1 to the power k, that is less than 1.

So, this is e k. So, this is nothing, but an estimate of norm of e k by norm of e 0. So, in other words, that means, what you require is norm of lambda. Sorry, mod of lambda max, that is less than 1. There are certain important things that need to be noted in this context.

(Refer Slide Time: 48:12)



First of all, this lambda is called as spectral radius or lambda max. Mod of that it is called as spectral radius of convergence. When we say spectral, it implicates a distribution of frequencies. Why it is given such a name is possibly because of the fact, that in many physical problems, Eigen values represent natural frequencies of the system. So, it has in certain physical problems. It has something to do with the frequencies. In general, what does this implicate. It implicates that whatever are the Eigen values of the matrix M, the maximum of that should be less than 1.

So, if you are considering to draw a circle with the center at origin and the radius as the magnitude of different Eigen values, you can draw different circles. The maximum radius of all these possible circles should be less than 1 because this lambda max mod of that must be less than 1 for convergence. The second important thing is, that it is a sufficient condition for convergence, but it is not a necessary condition for convergence. We have to remember this. So, there is a distinction between necessary and sufficient condition. So, if you say, that something is a necessary condition for convergence, what

does it mean? It means, that without satisfying that condition, never there will be convergence.

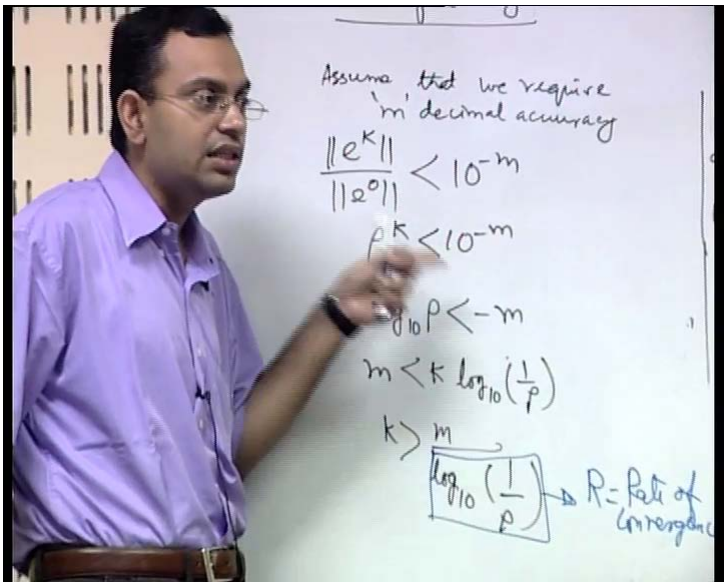
If you say something is sufficient condition for convergence, what you mean? You mean, that if that condition is satisfied, you will definitely have convergence, but you may also have convergence without satisfying that condition. Now, in iterative methods, there are many ways in which you can achieve convergence without satisfying any consideration. For example, say you are so lucky, that your initial guess is the actual solution. May be, from some source by some intuition or whatever you have guessed the correct solution.

So, in the iterative method, what you do? You substitute to your guess in the equation and see, whether it is satisfied or not in a systematic way. So, if you have the guess as the actual solution and put that in the solution, obviously, put that in the equation. Obviously, equation will be satisfied. So, in 1 step, you will get convergence without going through all these regards. So, that means, you can at least, we have identified one limiting case in which, you can have convergence without bothering for these types of condition or constraints. So, that is why, we say that it is a sufficient condition. So, we have to be very careful about using these terms, necessary and sufficient condition for converging.

Let us, may be take one simple example. Common life example to understand this. So, let us say, that let us make a statement that somebody, if he or she wants to become a very famous singer, should be very strong in classical music. So, if we say, that this is a sufficient condition, it can be accepted that if somebody is very good in classical music, definitely he or she will be a very good singer. May be, on the other hand, if we say that it is a necessary condition for becoming a very well established singer, that is not true. Well known example is Kishore Kumar.

So, what we can say is, that you can still become a very prolific singer without going through the roots of the classical music. So, that is not a necessary condition, but having a very strong base in classical music, may be a sufficient condition. So, we have to be very careful and particular about the terminology. Here, use of the terminologies, sufficient condition for convergence, but not necessary condition for convergence. The next terminology, that we will try to understand now, is called as the rate of convergence.

(Refer Slide Time: 53:30)



So, the rate of convergence talks about the following. If we know that, it is converging, the next question will come that at what rate it is converging or in other words, how many iterations that may be necessary to achieve the convergence? Minimum how many number of iterations that may be necessary? So, let us say, that we require a m decimal accuracy, then what we can say? Then, we can say, that in terms of order of magnitude, this is less than 10 to the power minus m. This mod of lambda max which we call as spectral radius of convergence. This we give a name rho.

So, we say that, this means that this is essentially mod of lambda max to the power k. So, rho to the power k less than 10 to the power minus m, where rho is nothing, but mod of lambda max, that is the spectral radius of convergence. We give it a name rho. So, if we take the log of both sides with respect to the base 10, then k log 10 with base rho is less than minus m.

So, we have m greater than k log 10 1 by rho, sorry less than. So, what we have done? So, we have brought m on this side and this on this side. Just check, whether this is all. Then, k greater than m by log 10 1 by rho. So, this log 1 by rho, with respect to base 10, this we call as r or rate of convergence. Why we call this as rate of convergence? We call this as a rate of convergence because you can see k is the number of steps necessary.

So, more is the rate of convergence, less is the number of steps that becomes necessary to achieve the convergence. So, we will stop here today. We will take it up from here in the next class and illustrate this through an example. Thank you.