Matrix Solvers

Prof. Somnath Roy
Department of Mechanical Engineering
Indian Institute of Technology, Kharagpur

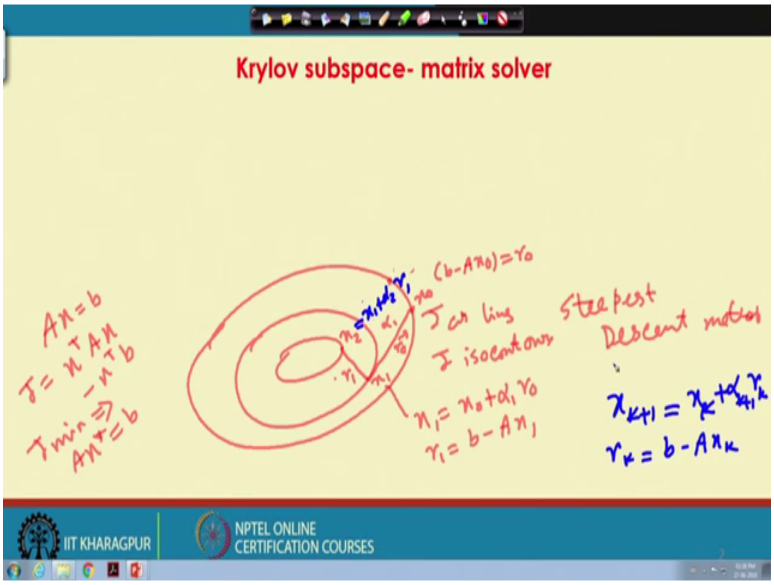
Lecture – 46
Conjugate Gradient Methods

Welcome we are discussing about Krylov subspace methods for solving Matrix equations iteratively. So, today we will discuss about Conjugate Gradient Method which is a very first Krylov subspace method for solving symmetric matrix equations. The classes before we have discussed about Arnoldi's method with through which using something like a modified Gram Schmidt method we can get orthogonal basis for Krylov subspace. And then we have also seen how to apply this for solving matrix equation using full orthogonal method or FOM.

And we also looked about different variants of FOM and then further looked into Lanczos algorithm in which, FOM can be converted for symmetric matrix equation which essentially gives something like a tri diagonal matrix system in which, direct solution like TDMA type algorithm can be utilized for faster solution. And now, we will continue this discussion for conjugate gradient method.

Before going into that, I will again quickly review few of the important things in Krylov subspace method, as well as full orthogonal method or and Arnoldi's method, which will give a base for continuing this discussion to conjugate gradient method.

(Refer Slide Time: 01:49)



So, if we look into the matrix solver solution equations for example, when we discussed about steepest descent algorithm, which is the building block of any projection method. And Krylov subspace method is just an extension of general projection method whether, the where the projection space is not one dimensional rather multi dimensional.

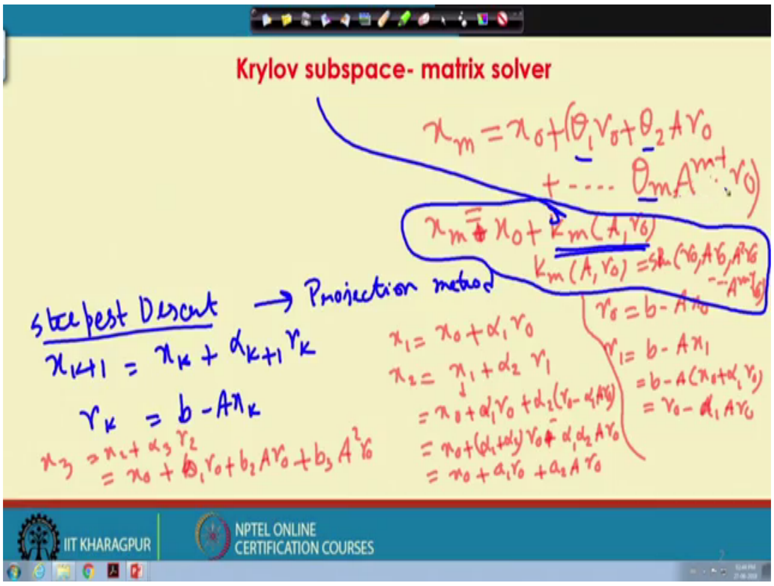
So, if you look into the steepest descent algorithm for example, we have a we have to solve Ax is equal to b and we define a function J is equal to x transpose A x minus x transpose b, which has to be minimized J min will imply, A x star is equal to b. This has to be solved so, we get the exact solution of A x is equal to b.

So, we start with the iso contours of J and this is basically, steepest descent what I am discussing here. So, these are the J constant lines or we call them J isocontours, these are the lines over which J is constant. So, we will start from any value x 0 and evaluate what is b minus A x 0 that is equal to r 0. And then I will move along r 0, I will keep on moving along r 0, till just one second this r 0 has to be orthogonal to it.

So, let us select the x 0 here. This is just for convenience of explaining that this x 0 here and I will keep on moving along r 0, say up to a distance alpha 1. Here, the r 0 vector will be tangential to the new iso contour and I have to change the search direction. I have to evaluate so, this is x 1 where, x 1 is equal to x 0 plus alpha 1 r 0 for example. And I will evaluate r 1 is equal to b minus A x 1 and then I will move along r 1 to find out x 2 where, x 2 is equal to I will write x 2 is equal to b minus sorry, x 2 is equal to x 1 plus alpha 2 r 1 something like that.

So, at each step x will be updated as a function as xk plus 1 is equal to xk plus alpha k plus 1 rk something like this. And my rk will be defined as b minus A x k. So, now if we look into this update into little detail or just work it, work out the updates for first 2-3 steps.

(Refer Slide Time: 05:19)



So, I rather let us this I keep with the relations that, x k plus 1 is equal to xk plus alpha k plus 1 r k and r k is equal to b minus A x k. Exactly, this is what we have done in steepest descent, which is building block of projection method. So, this later we will give as what are the general projection methods. Now, I can write that let us use another pen, x 1 is equal to x 0 plus alpha 1 r 0. Then, x 2 will be x 1 plus alpha 2 r 1.

Now, what is x 1? X 1 is nothing but x 0 plus alpha 1 r 0. And what is r 1? R 0 is equal to b minus A x 0, r 1 is equal to b minus A x 1 is equal to b minus A into x 0 plus alpha 1 r 0, which is b minus A x 0 is r 0 minus alpha 1 A r 0. So, I can substitute here that x x 1 is x 0 plus alpha r 0 and this is plus alpha 2 into r 0 minus alpha 1 A r 0. So, this is x 0 plus alpha 1 plus alpha 2 r 0 plus minus alpha 1 alpha 2 A r 0.

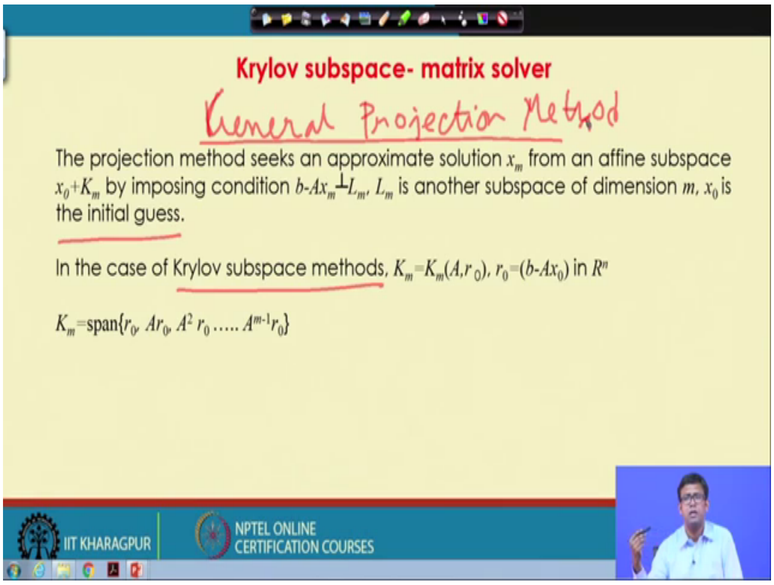
Now, can we see that this can be written as also x 0 plus say; A 1 r 0 plus A 2 A r 0. Similarly, if I write expression for x 3, this will be x 2 plus alpha 3 r 2. And now, I will substitute calculate the values of r 2 and x 2 and substitute it, here I will again get x 0 plus A say, b 1 r 0 plus b 2 A r 0 plus b 3 A square r 0 so on. So, I will get a general expression for mth order or mth iteration of x as, let us let me wipe this part out so that, or let me write down the general expression that, x m after mth iteration is equal to x 0 plus say, theta 1 r 0 plus theta 2 A r 0 plus and it goes up to theta m plus 1 A to the power x 3 is related with beta 3 so, theta rather theta 3 theta 3 A to the power m minus 1 r 0.

So, the entire update of x 0 becomes a function of r 0 plus A r 0 plus A to the power m minus 1 r 0. And precisely that gives us the idea of Krylov subspace that, x will be updated in a sub in a affine space so that, x m is equal to x 0, x m is equal to x 0 plus k Krylov subspace of A and r 0 where, k m A r 0 is the space spanned by is span of r 0 A r 0 is square r 0 so on, A to the power m minus 1 r 0.

So, this is: what is the basic definition of Krylov subspace; what we have used earlier. So, now, if we look into the definition how x is updated and when updating x for example, when we looked into steepest descent x is always updated along r. And in that way, we made it also a point that the new residual vector r should be orthogonal to the previous residual vector or also to the functional subspace iso contour of J.

So, there is a constraint on how x will be updated and based on which the values of alpha are calculated. So, x is a x is any way updated in this particular plane, which is space or which is Krylov subspace, this is the Krylov subspace, x is always updated in along on Krylov subspace. And what are the coefficients through which, these updates will be there, what will be theta 1 theta 2 theta m, that will come from the projection constraint that the update should be orthogonal to or the residual should be orthogonal to a particular plane.

(Refer Slide Time: 11:20)

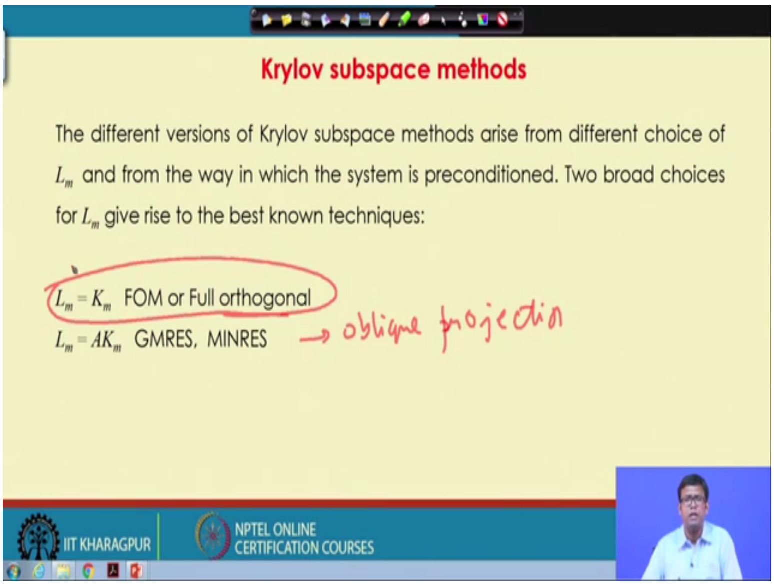


So, if we now look into the recap what we have seen in Krylov space, which is we will also write that this is a general projection method. And, we have also seen that, this update is updating x in Krylov subspace we will make it converge into the exact solution x star and the residual as we update x and (Refer Time: 11:50) Krylov subspace also the residual will also converge to 0.

The projection methods seeks an approximate solution x m from an affine subspace of x 0 plus K m, I have discussed: what is the definition of affine subspace in last few classes. By imposing the condition that b minus x m is a this particular perpendicular to L m where, L m is another subspace of dimension f and x0 is the initial guess in case of so, this is the statement for general projection method. In case of Krylov subspace methods, K m is given by K m of a r 0 where r 0 is b minus a x 0 and K m is a space spanned by r0 A r0 8 A square r 0 up to the power A to the power m minus 1 r 0.

So, what we are thinking as multiple iterative steps or multiple search directions in a steepest descent algorithm is basically, the multiple bases vectors through which x m is updated x0 is updated as x m and these are the multiple base several bases vectors of the Krylov subspace. And we have seen what is the grade of why with respect to A what can be the maximum dimension of Krylov subspace etcetera in last few classes and have shown that, this is a convergent method, this the in general Krylov subspace method should converge to the right solution based on any (Refer Time: 13:33) value of x0 or any starting residual r 0. So, our goal is to develop efficient solvers using Krylov subspace method.

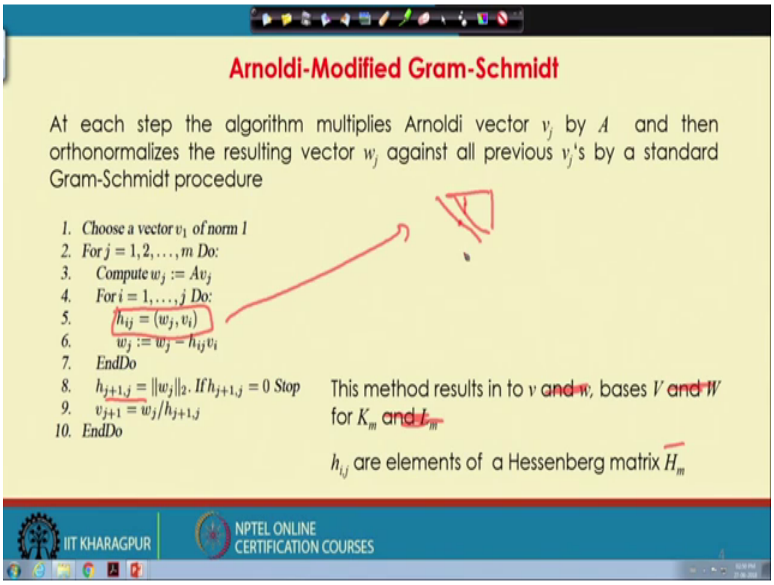
(Refer Slide Time: 13:36)



The different values of versions of Krylov subspace methods arise from different choices of L m; that means, the space on which to which the residual should be orthogonal and from the way in which the system is preconditioned. This is a recapitulation slide of previous lectures. There are 2 broad choices for L m which, gives based on techniques: one is L m is equal to K m, which is an orthogonal method and another is L m is equal to AK m, which is an oblique projection.

And now, in this particular session we are concentrating on the full orthogonal method. We have earlier discussed about full orthogonal method also and we will see how for symmetric matrix we can generate very efficient solver from full orthogonal method.

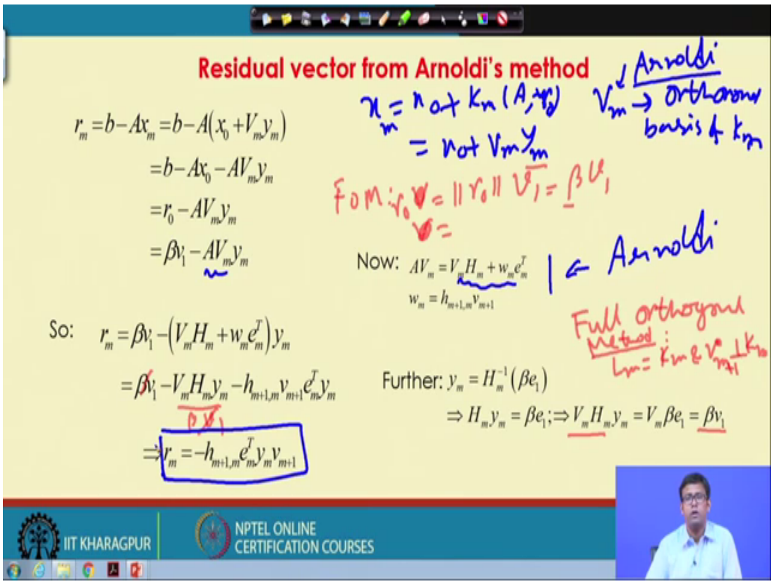
(Refer Slide Time: 14:42)



And Arnoldi modified Gram Schmidt is an algorithm through which, we can get the bases vectors of the this is bases vectors of v, the bases vectors of the Krylov subspace which are the base v's are the bases vectors of Krylov subspace. We can generate it through Arnoldi modified Gram Schmidt method that, you start with any vector v 1 and which is basically, your when you are using full orthogonal method this will be r 0. And then, take a product of v with respect to the A matrix A v, from v you subtract A v, with v is the bases of Krylov subspace A v is the next bases of Krylov subspace so, you orthogonalize v and A v using something like, Gram Schmidt error in find out the orthonormal bases vectors.

In that way you get another matrix which is h i j; which is the product between the new bases of Krylov subspace space and the older bases of Krylov subspace, before orthogonalization of the new bases. And this h i j forms a Hessenberg matrix h m. And this Hessenberg matrix means, sorry this is h bar m forms a Hessenberg matrix h bar m. The Hessenberg matrix is basically, an upper triangular matrix plus 1 sub diagonal term into it. And this is how we get h j because, it goes up to h j plus 1 j so, it goes up to the diagonal term and plus one sub diagonal term in it.

(Refer Slide Time: 16:39)



Now, we look into the residual vector form Arnoldi's method, what is a residual vector at one particular iterative step on Arnoldi's method. And this is r m is equal to b minus A x m and x m is x belongs to, so we have seen it earlier that x is equal to x 0 x m is equal to x 0 plus K m A y and V m is orthonormal bases of Kn. So, we can write x m is equal to x 0 plus V m multiplied by sorry they K m of x r 0 multiplied by y. So, x m is x0 plus the Krylov subspace of A and r0 and the orthonormal bases of this Krylov subspace is obtained as V m from Arnoldi method. So, this comes from Arnoldi, from Arnoldi's method. So, you can write x m is equal to x0 plus V m into y.

So, if substituted x m is equal to x0 plus V m into y V m into y m (Refer Time: 18:03) and y m's are the coefficients which are multiplied with different bases vectors of Arnoldi's of Krylov subspace to get a general expression of a general vector in Krylov subspace. So, x m is equal to x0 plus a general vector in Krylov subspace and we have to find out what y m is that so that, we can satisfy the equation come to it later.

And so this is b minus A x0 plus V m y m b minus A x0 is equal to r0. So, r 0 minus A V m y m, which is r0 is again we have if we go to the maybe to the previous slide. So, we started with not here, but if we if we not in the previous slide if we see, in full orthogonal method right in FOM we started with the first bases of Krylov subspace is capital V is equal to mod of r0 into v 1 and which I have written as beta into v 1.

So, beta is mod of r 0 which is the first this magnitude of the unit vector first case of first residual in Krylov subspace are 0 in Krylov space. So, beta v 1 is equal to and we can write and this is so, we can write that V is equal to or V is equal rather V not should not write that we star put r0 is mod r0 into v 1, v 1 is an unit bases you first unit bases first bases vectors of the Krylov subspace which is an unit vector this is beta v 1.

So, r0 is equal to beta v 1 I substitute r0 is equal to beta v 1 and the mth ordered residual is beta v1 minus A V m y m. Now, from Arnoldi's method we got these identities, this comes again from Arnoldi. If we look into the Arnoldi's algorithm we have just shown in the last slide, this these relations are evident from there that A V m is equal to V m H m plus w m e m transpose, w m is h m plus 1 h m v m. So, now in A V m, I will substitute A V m by this particular quantity. So, r m is equal to beta v 1 minus V m H m plus w m e m transpose y m. And this is beta v 1 minus V m H m y m minus h m plus 1 v m e m transpose.

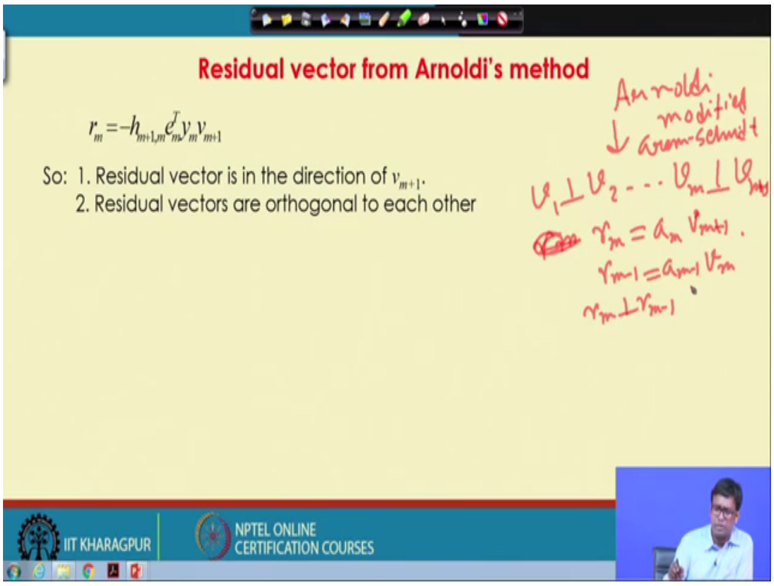
Now, what is y m and that is what we are trying to find out what is the value of y m. Because, if we know y m we have defined the Krylov subspace, Arnoldi's method will give us the orthonormal bases of Krylov subspace. If we can find y m, we will multiply it with different bases vectors and find out what is the update in Krylov subspace which is to be added with initial guess x0 so that I get the final solution. So, now this update is obtained, if we can remember steepest descent method, alpha is the amount of distance that I should go along one particular direction vector r. And, this is obtained from the relation that the residual is orthogonal to particular subspace.

Similarly, full orthogonal method uses the fact that the residual is orthogonal to L m which is same as K m, the residual is also orthogonal to the prior surface only. And using that we can get, so this is using the fact that full orthogonal method or Arnoldi's method for solving linear equation that gives L m is equal to K m and r m is orthogonal to r m plus 1 rather is orthogonal to K m.

So, using that we can find out we have shown it in few lectures before, y m is H m inverse beta e 1. So, H m Y m, H m is the Hessenberg matrix the a part of the Hessenberg matrix which we have obtained in the last shown in the last slide. So, H m y m is beta e 1 multiply both side by V m, V m H m y m which is the this term is V m into beta e 1, e 1 is an unit first unit vector 1 0 0 which is multiplied with V m will give the first vector in the of the orthonormal bases (Refer Time: 23:26) beta v 1. So, I can write that, this is equal to beta V 1 utilizing this factor and they will cancel out. So, I will get that r m is minus h m plus 1 e m transpose y m v m plus 1.

This is a very important identity from the Arnoldi's method or from full orthogonal method that, the residual is orthogonal to the next bases in the Krylov subspace, residual at one particular m level is orthogonal to the next bases function of Krylov subspace. Sorry, residual is not orthogonal, residual is parallel, the residual is along the next bases of Krylov subspace. So, residual will be orthogonal to the previous bases I am sorry. Residual r m is along v m plus 1 where, v m plus 1 is the m plus 1th bases of Krylov subspace; mth residual is in the direction of m plus 1th bases of Krylov subspace that is the finding.

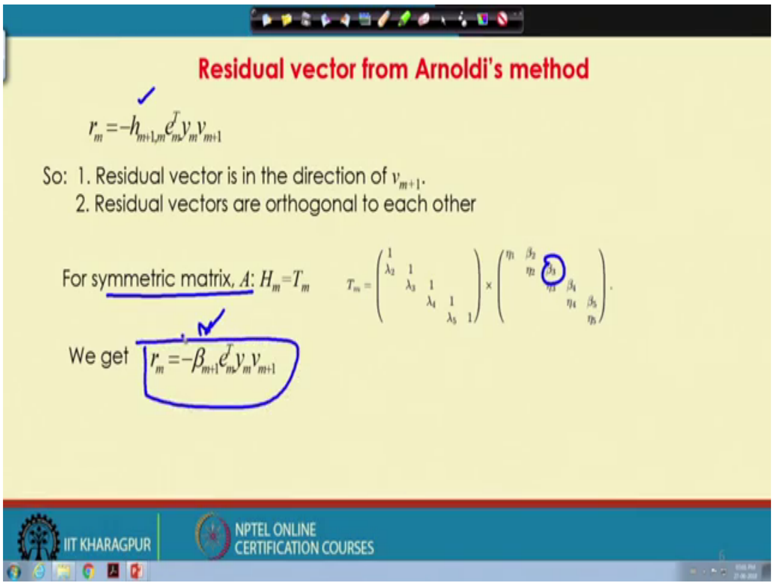
(Refer Slide Time: 24:45)



So, we start with r m is equal to minus h m plus 1 m e m transpose y m v m plus 1. So, the residual vector is in the direction of v m plus 1 and residual vectors are orthogonal to each other. Why? This is simply due to the fact that v 1 is orthogonal to v 2 is v m is orthogonal to v m plus 1, which comes due to Arnoldi method is modified by Gram Schmidt.

Gram Schmidt method will give a set of orthonormal vectors. So, these v vectors are orthogonal to each other. Therefore, once I see residual r m or r m is equal to something say a m v m plus 1 and r m minus 1 is equal to a m minus 1 v m right, that is the founding here; v m plus 1 and v m are orthogonal to each other. So, r m and r m minus 1; that means, r m is orthogonal to r m minus 1 and so on. So, residual vectors are orthogonal to each other.

(Refer Slide Time: 26:08)



For a now for a symmetric matrix A: H m or the part of that Hessenberg matrix is a tri diagonal matrix. We have shown it earlier if A becomes symmetric H m is a Hessenberg matrix, so it has a it is a upper triangular matrix with one sub diagonal if A is symmetric, so H m is also symmetric. So, only the sub diagonal and super diagonal exists and it becomes a tri diagonal matrix. And we can express this tri diagonal matrix as product of two bi diagonal matrices; one is a lower triangular form, another is an upper triangular form. And if we use this form these betas are comes in terms of the h here.

So, r m becomes minus beta m plus 1 e m y m v m plus 1 for symmetric matrix A. If we can tri diagonalize the Hessenberg matrix h m for symmetric matrix A, we can directly calculate from that form of the Hessenberg matrix, we can directly calculate what is the coefficient which will be multiplied with the m transpose y m v m plus 1 to get the residual vector.

So, there is a relationship between the residual vector and the tridiagonal matrix which directly comes here. So, with will I will stop here in this session and with this particular idea that our residual vectors are at mth level is along the base m plus 1 in bases of Krylov subspace and residual vectors are orthogonal to each other. We will see how this we got an expression for the residual vector for symmetric matrices, how this can be substituted into d Lanczos algorithm or direct Lanczos algorithm where, we are doing this TDMA type of calculation and we can get a faster solution method.

Thank you.