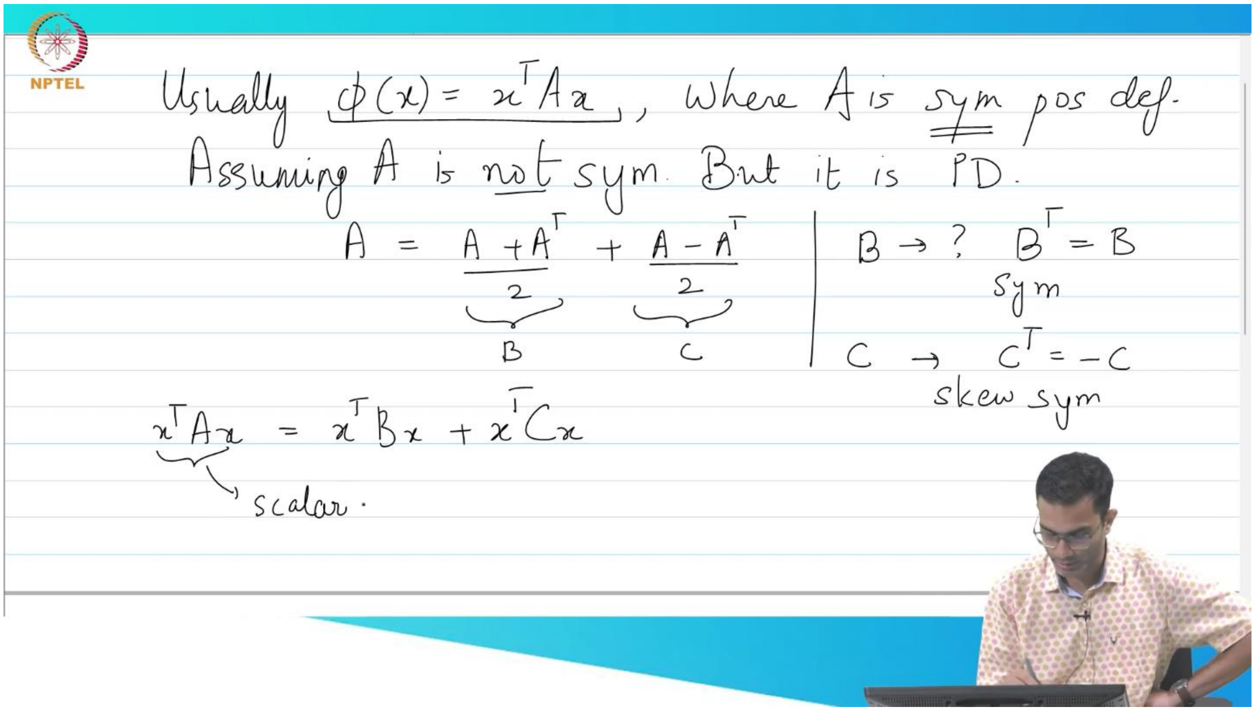
Course Name: Optimization Theory and Algorithms
Professor Name: Dr. Uday K. Khankhoje
Department Name: Electrical Engineering
Institute Name: Indian Institute of Technology Madras
Week - 05
Lecture - 33

Conjugate Gradient Methods - Introduction and Proof

So, I am going to talk about one aspect of very simple linear algebra which I probably should have told you early on, but you know I felt it was kind of very easy for you to work out, but I will mention it nevertheless. In whatever we have been doing so far whenever we talk about a matrix we have what quality or property of it have we specified? Symmetric positive definite right. Usually when I have a quadratic form like this I say $x^T Ax$ where A is symmetric positive definite. Now, a very natural question which someone asked me a couple of classes ago was why do we need this? Why do we need it to be symmetric? Supposing I, so what is the requirement for a positive definite matrix? How do I, I mean what is the signature of a positive definite matrix? Eigenvalues are positive. Does that automatically imply that the matrix is symmetric? No, no reason whatsoever, but yet we say symmetric positive definite. In fact, all the books you read will say symmetric positive definite.

Any idea why? Eigen decomposition will not change whether it is symmetric or asymmetric. It will be of use in analysis that is one way of that is one general hope. Any other? Eigenvectors are orthogonal that does not really matter. So, the real answer is that, pardon me.



Usually $\underline{\phi(x) = x^T A x}$, where A is $\underset{\neq}{\underline{\text{sym}}}$ pos def.

Assuming A is $\underline{\text{not}}$ sym. But it is PD.

$$A = \underbrace{\frac{A+A^T}{2}}_{B} + \underbrace{\frac{A-A^T}{2}}_{C}$$

$B \rightarrow ?\quad B^T = B$
sym

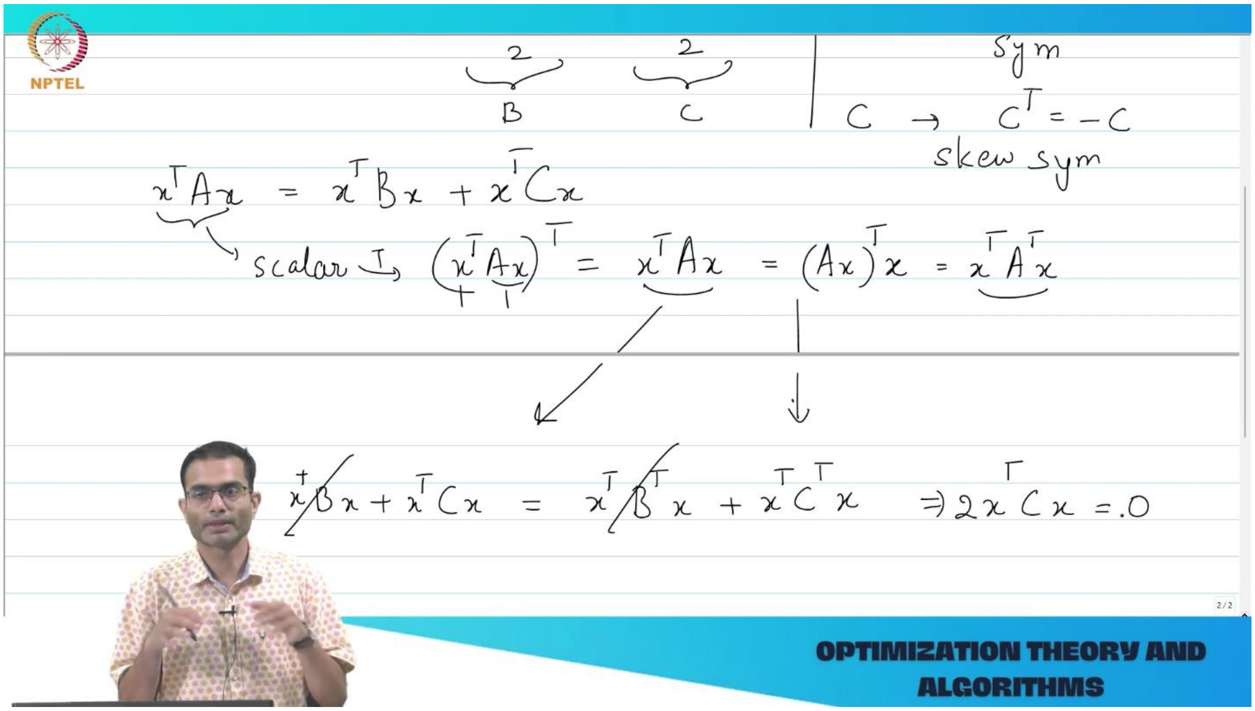
$C \rightarrow \quad C^T = -C$
skew sym

$$\underbrace{x^T A x}_{\text{scalar.}} = x^T B x + x^T C x$$

Yeah, we assumed A was symmetric, A which means $A^T = A$. Yeah, am I saying why are we assuming that? So, the gradient is more elegant, ok. We will come to matters of elegance later, first let us get the things here. Ok, so the answer is actually very simple. You can take A to be positive definite and not insist on it being symmetric.

You know why? It does not matter. So, let us see why it does not matter. So, now let us start by assuming A is not symmetric, but we will assume that it is positive definite. So, what follows is very very simple linear algebra. Can I write A in this form? Obvious I just added and subtracted half of A^T by I mean half of A^T.

But let us call this let us say B, let us call this C. So, can you say something about B? B obviously is symmetric because $B^T = B$, right. You can easily check that $B^T = B$. So, this is called the symmetric part of A. What about C? Antisymmetric or skew symmetric because $C^T = -C$.



$\underbrace{}_{B}^{2}$ $\quad \underbrace{}_{C}^{2}$ $\quad C \rightarrow C^T = -C$

Sym

skew sym

$$\underbrace{x^T A x}_{\text{scalar}} = x^T B x + x^T C x$$

$$\text{scalar} \rightarrow (x^T A x)^T = x^T A x = (Ax)^T x = x^T A^T x$$

$$\cancel{x^T B x} + x^T C x = \cancel{x^T B^T x} + x^T C^T x \Rightarrow 2x^T C x = 0$$

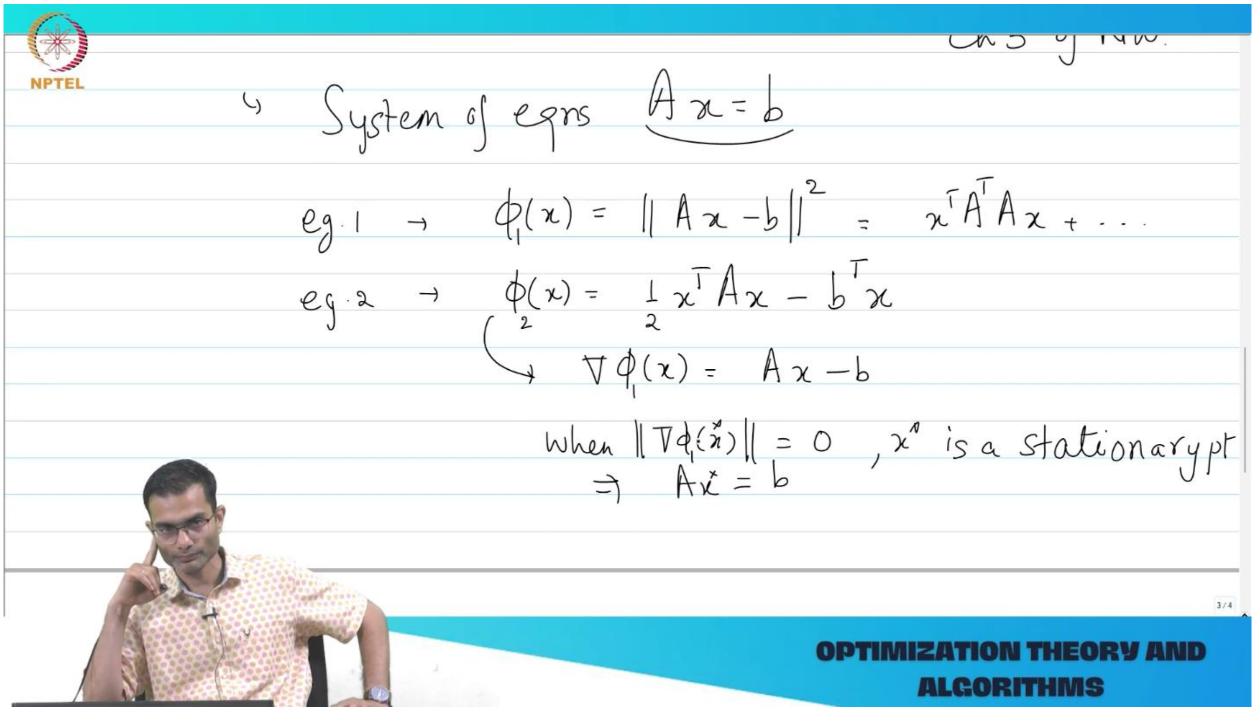
So, let us call this skew symmetric. So, before we go ahead what is the simple conclusion of this? Any matrix can be written as part symmetric, part skew symmetric, ok. Now, let us look at the action of let us take up our quadratic form. Our quadratic form is $x^T Ax$. So, $x^T Ax$.

Obviously, I am going to split that in terms of B and C, fine. Now, $x^T Ax$ is it a scalar vector matrix? It is a scalar, right. So, for a scalar if I put the transpose operation what will happen? I will get the same number back, right. So, if I do a transpose that means that $x^T Ax$ should be equal to $x^T Ax$. So, I am going to group these two guys together so that we can open up the transpose operation.

So, therefore, this is going to be $Ax^T x$ right. $AB^T = B^T A^T$. So, x^T this is my first guy and Ax I flip the two. So, $Ax^T x^T$ transpose becomes x. So, if I open this, this becomes $x^T A^T x$ right.

So, this is your hint in which direction it is going right. So, now I have this is equal to this. Let us substitute over here for A. So, what am I going to get? On this side when I substitute $x^T Ax$ is going to give me $x^TBx + x^TCx$ is equal to $x^TB^Tx + x^TC^Tx$, right. So, what else can we do? Is there any simplification that follows? What about x^TBx? Can I cancel it, right? Because $B^T = B$.

What does this imply? And $C^T = -C$. So, that leaves you with what? $x^TCx = 0$ right. So, the action of the quadratic form what does it do? It eliminates the skew symmetric component. So, if $x^TCx = 0$ this implies that x^TAx is essentially x^TBx because $x^TCx = 0$.



Ch 3 of NW.

↳ System of eqns $\underbrace{Ax = b}$

eg.1 → $\phi_1(x) = \|Ax - b\|^2 = x^TA^TAx + \ldots$

eg.2 → $\phi_2(x) = \frac{1}{2}x^TAx - b^Tx$

$\nabla\phi_1(x) = Ax - b$

when $\|\nabla\phi_1(\hat{x})\| = 0$, $\hat{x}$ is a stationary pt

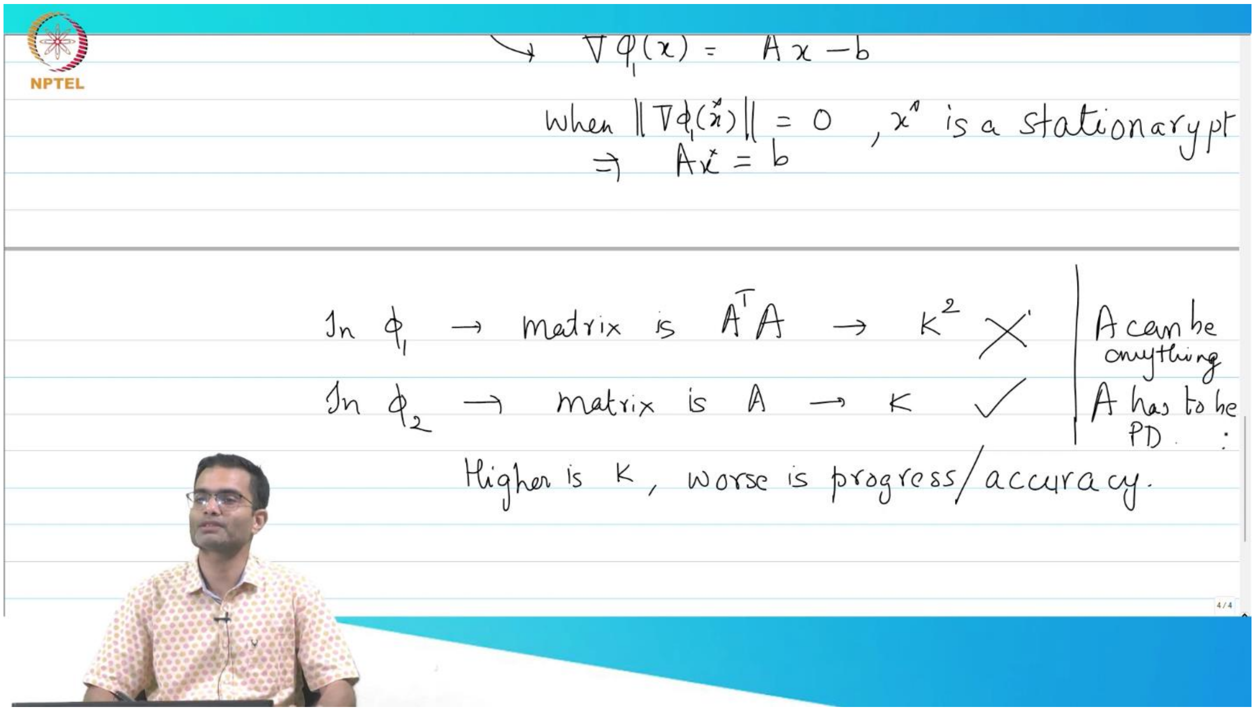
$\Rightarrow A\hat{x} = b$

Right. So, this is why I mean you see there is absolutely nothing fancy we did not even do an eigenvalue decomposition over here, right. So, what is it telling you that it does not matter if you add in the skew symmetric part when I stick a x^Tx on both sides from left and right it is going to knock off the skew symmetric part, right. So, now we can come to the part of elegance which was stated earlier that if you know this to be the case may as well start working with a symmetric matrix in the first place itself right because it is not going to matter. Is that clear? Very simple linear algebra ok. So the next part of we are going to begin the next module of this course.

We have finished line search methods in but not we have not finished line search methods we have finished the gradient descent method which is like your first algorithm for unconstrained optimization. And what was the basic premise in the gradient descent method was to walk in a direction of descent. Now this you in a first introduction to optimization you would imagine that this is the most obvious thing. Go in the direction of descent you are guaranteed to reduce the function value. And as we will see further on in this course there are two non-intuitive ways of working with this.

You say ok I my search direction need not be a descent direction I will come up with some other clever criteria for it. And that gives rise to what is called the conjugate gradient method. ok. There is another version of this which gives rise to what is called as the accelerated gradient method. So, we will look at that later.

Both of them question this idea of should I go in the direction of descent. Turns out you need not go in the direction of descent, but you come up with some other criteria and you are able to get good convergence ok. So, that is why it is a very interesting way of looking at this problem ok. So, conjugate So, this is roughly chapter 5 of Nocedal and Wright in case you are following along ok. Now, what we will do is we will introduce this again from a very simple case and then we will begin to generalize it.



$\rightarrow \nabla \phi_1(x) = Ax - b$

when $\|\nabla \phi_1(\overset{*}{x})\| = 0$, x^* is a stationary pt

$\Rightarrow A\overset{*}{x} = b$

In ϕ_1 → matrix is A^TA → K^2 ✗ | A can be anything

In ϕ_2 → matrix is A → K ✓ | A has to be PD

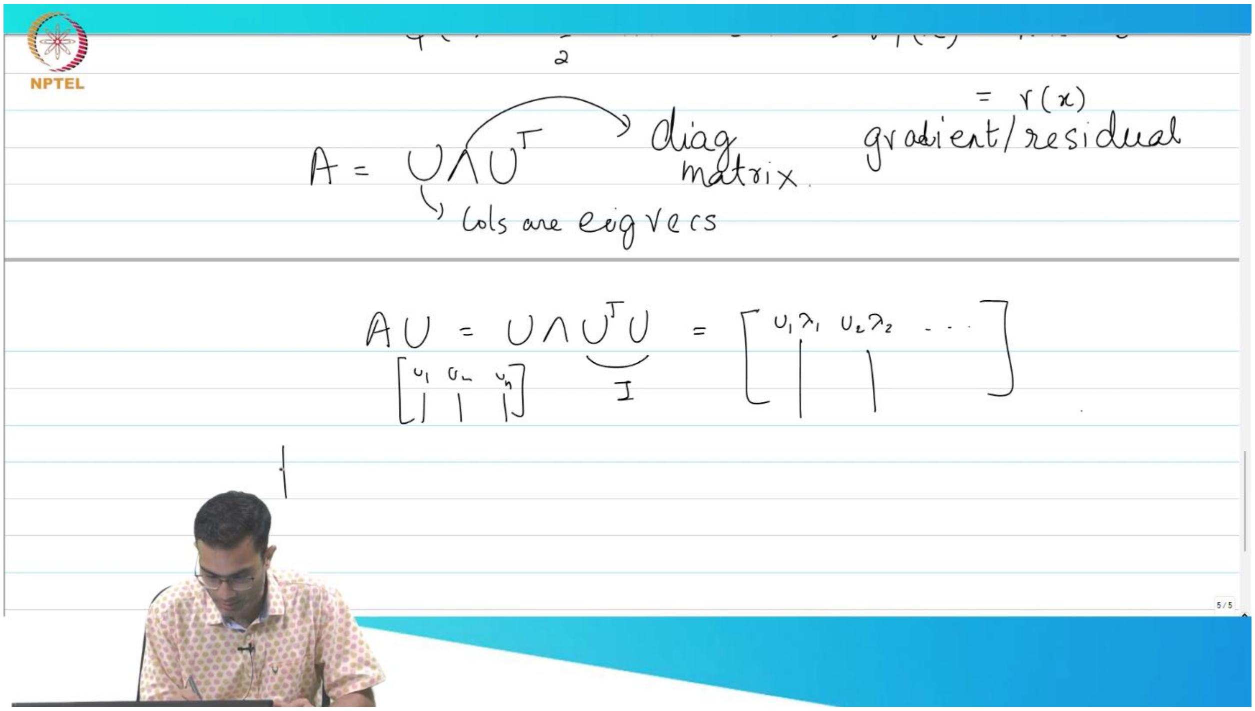
Higher is K, worse is progress/accuracy.

In fact, this method has such a nice generalization that it gives you a way to work with non-linear problems immediately ok, it need not be a quadratic form. But the motivation for this is as follows, in a lot of problems of engineering whether it is mechanical, aerospace, electrical, you end up with a linear system of equations to solve, right. You will basically get system of equations of the form $Ax = b$ and you want to solve this. A and b can usually be extremely large, hundreds of thousands, million by million size right. So, your usual way of solving this when you took linear algebra would be what like LU decomposition right, which is called a direct method which has a complexity of order n^3 right.

So, sometimes you are not willing to pay that price and you want a faster way of doing it. Now, having learnt optimization how would you link it to solving this problem? This so far doesn't look like an optimization problem. I haven't defined an objective function for you but can I apply the idea of optimization to solving this problem? And if so, how should I think about it? Okay, so one idea example one would be if I define my objective function as this. Seems like a good idea right because when this is minimized Ax is going to be equal to b ok. Any other idea? That is example number 2.

So, he is saying $\phi(x) = \frac{1}{2}x^TAx - b^Tx$. So, that is also correct. Why is this correct? Right when I take the gradient of this expression what will I get right. So, if I do $\nabla\phi(x)$ I am going to get

$Ax - b$ and we know that when we are in an optimization problem we are looking for a stationary point. A stationary point is defined as a point where $\nabla\phi(x) = 0$.



$= r(x)$

$A = U \Lambda U^T$ diag matrix. gradient/residual

Cols are eigvecs

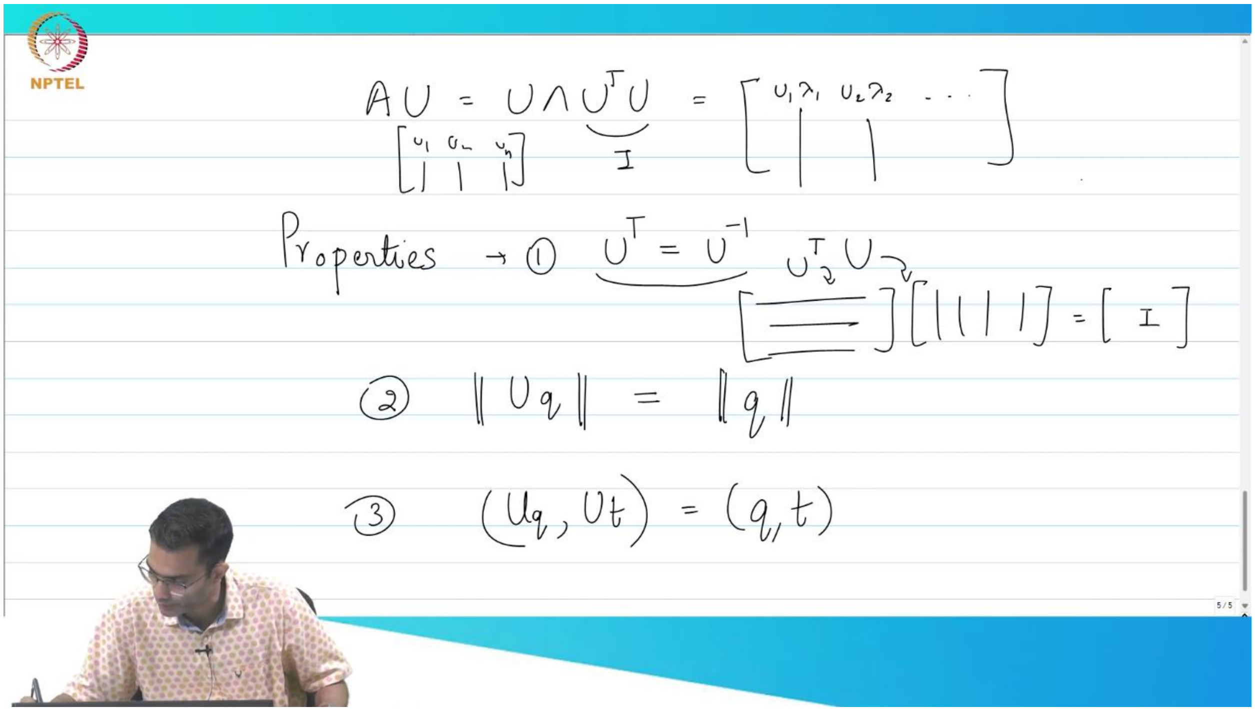
$$\underset{\begin{bmatrix} u_1 & u_2 & u_n \\ | & | & | \end{bmatrix}}{AU} = U\Lambda \underbrace{U^T U}_{I} = \begin{bmatrix} u_1\lambda_1 & u_2\lambda_2 & \cdots \\ | & | & \end{bmatrix}$$

So this is the, so which of these two ideas is gonna be better? They both on the face of it are perfectly legitimate ideas, right? Now which should we choose, example one or example two? In example one, if I open this up, what am I gonna get? Norm something squared for example, would be $(Ax - b)^T(Ax - b)$. What is going to be the term that involves x^T I mean the quadratic term what will it have? What will be the expression? $x^T A^T Ax$, it is going to be $x^T A^T Ax$ and three more terms ok. I am focusing only on the quadratic term. Why do you think I am doing that? $A^T A$ is symmetric, it is positive semi-definite at least we do not know about the eigenvalues. So, from this point of view can someone say which is better? So, let us call this $\phi_1(x)$ and $\phi_2(x)$.

Let us assume A is invertible ok. One answer is that it does not matter both are the same. So, the hint is to look back at the rate of convergence that we had discussed in a previous module. What is that rate of we did not derive the full expression, but we give an expression. What was the essential quality of this the matrix that is sitting inside that comes into picture? Condition number, right.

So, in case of ϕ_1 the matrix is $A^T A$, in ϕ_2 the matrix was that I am working with was simply A, ok. Now, better the condition higher the condition number is it good for us or worse for us in terms of convergence? Worse, right. So, higher is κ worse is progress. Progress is one thing, another thing is the higher the condition number as you saw from the tutorial problem, accuracy is also questionable, right. If there is an error in the b vector your solution is less accurate.

So, worse is the progress and the accuracy. Now, let us say the matrix A has condition number κ. What is the condition number of $A^T A$? It is κ^2, right. It is simple eigenvalue decomposition you stick $A^T A$ together you will get the square of the two diagonal matrices will multiply with each other all the eigenvalues will get squared. Therefore, the condition number is largest eigenvalue to smallest eigenvalue.



$$A\underset{\begin{bmatrix} u_1 & u_2 & u_n \\ | & | & | \end{bmatrix}}{U} = U\Lambda \underbrace{U^T U}_{I} = \begin{bmatrix} u_1\lambda_1 & u_2\lambda_2 & \cdots \\ | & | & \end{bmatrix}$$

Properties → ① $\underline{U^T = U^{-1}}$ $U^T_2 U \rightarrow \begin{bmatrix} \text{—} \\ \text{—} \\ \text{—} \end{bmatrix} \begin{bmatrix} | & | & | & | \end{bmatrix} = \begin{bmatrix} I \end{bmatrix}$

② $\| Uq \| = \| q \|$

③ $(Uq, Ut) = (q, t)$

So, this becomes κ^2. Did everyone follow this? Write down the eigenvalue decomposition for A then A^T multiplied to get squared. κ is a number that is greater than 1 or less than 1? Greater than 1, right. So, that is why this method is less preferable. It will give the correct answer, but the moment you enter the real world where there is error and noise, the noise is going to get amplified by a square of the condition number rather than the second approach where you are working just with condition.

Yes, yes correct. Correct. Excellent right. So, that is a very good point you have beaten me to it. Let us put over here. So, A can be anything whereas in the second approach A has to be positive definite or as we have learnt symmetric positive. So, if you were dealing with a very very general purpose problem you may have no choice, you may have to go for approach 1 and pay the price of condition number.

But as we will see there are lots of tricks around these things right, but if your matrix was given to be positive definite you should definitely go for approach 2 because your condition number is better. So, in what follows we will now make because we want to study the essential core idea of conjugate gradient we are going to make our life easy we will say assume A is symmetric positive definite ok. So, here on assume A is symmetric positive definite ok. and as we have already seen the cost function that we are going to assume is half of $x^T Ax - b^T x$ which leads to a common shorthand instead of $\nabla f(x)$ a common way to call this term $Ax - b$ is $r(x)$ which simply stands for the remainder right. Ideally we want the remainder or residual to go to 0.

So, another way in which this variable is kept track of is $r(x)$ ok which is the residue. Now, given the fact that we are working with a symmetric positive definite matrix there are some nice properties of this matrix that we can that we know of. So, we will just quickly summarize those two three properties we will use them throughout the method they are very crucial. So, also again like a small revision of linear algebra ok. So, A is symmetric positive definite therefore, if I write its eigenvalue decomposition that is how it is going to look right.

The columns of U are what? Eigenvectors right. So, the columns are eigenvectors and λ is a diagonal matrix of eigenvalues right. Sometimes we tend to forget is it $U\lambda U^T$ or $U^T\lambda U$, what is the quick way of figuring it out? Which one is which if you don't want to rely on memory. These are small, small hacks which you should build in. Correct. So, you know when anyone says eigenvalue problem, what is the first thing that even in class 11 you study? $Ax = \lambda x$, everyone knows that, right.

So, if I were to multiply from the right by U, what will I get? I will get $U, U\lambda U^T U$. Now U consists of orthogonal columns, this is a property of positive definite matrices that the eigenvectors are orthogonal to each other right and in fact you can orthonormalize them also. So, this becomes what? Identity. Now U right multiplied by a diagonal matrix what will happen? The lambdas will multiply each of the columns right. So, λ_1 this will become like $U_1\lambda_1 U_2 2$ ... this is how it becomes.

And so, U is already there a into AU. So, you can see that if I take $U_1 U_2 U_n$. So, you can see now this each of there are n eigenvalue equations here right $AU_1 = \lambda_1 U_1$ that is your eigenvalue problem right. So, that is a very quick way of just getting it right that it is $U\lambda U^T$ and not the other way around. All right, this has given us one property I have already stated which was that $U^T = U^{-1}$, ok. Is it easy to see this? Right, U's if I take U the columns are all orthogonal, right.

So, if I take U like this as columns And if I take U^T this is U^T over here all those guys become rows. Now, when I multiply U^T into U what will I get? Since they are orthonormal to each other right I am only going to get diagonal terms in this right. So, this is going to give me diagonal. In fact, since they are orthonormal what will this be? Identity right. So, this shows you and it is a square matrix of course right we are working only with square matrices.

So, this gives you $U^T = U^{-1}$ quite straight forward. This U matrix has a few more you know nice interpretations. For example, what does it do to the length of a vector? Supposing I multiply U multiplied by some vector q ok and take the length of that vector and compare it to the length of the original vector itself. What is what happens? is equal right.

So, actually equal. So, that gives us the reason I state this is because it gives us a nice geometrical interpretation. In n dimensions what is going on? It is a rotation. I take a vector multiply by U it just rotates it. Rotation does not change length right. So, you can write this algebraically like this does not have much intuition, but now you see ok it is actually just a rotation right.

So, it is length preserving. It's length preserving, but now what about if I have two vectors, let us say q and t and I apply U to both of them, will it change the angle between them? If it's a rotation, it should not change the angle between them, right? So that's the second property. If I do Uq and Ut, so this is the inner product. It should be the same as the inner product of q and t. you would have seen the proofs of all of these in linear algebra they are quite straightforward.