

Course Name: Optimization Theory and Algorithms
Professor Name: Dr. Uday K. Khankhoje
Department Name: Electrical Engineering
Institute Name: Indian Institute of Technology Madras
Week - 06
Lecture - 40

Expanding Subspace Theorem

All right. So, what we have done so far is that we have shown, we have got the geometric intuition. We have also had a proof which told us that at most n steps are required. Now, there is again one very powerful theorem in linear algebra which helps to formalize this. The proof is a slightly longer proof. I am going to do half the proof to give you the flavor of it.

The remaining half, if you are interested, you can read it in the book. Okay. Because it will take a long time to drive it. But let me state the theorem.

It is a very powerful theorem and it is used in all the analysis of this conjugate gradient method, ok. It has a very cool name also. It is called the expanding subspace theorem. And what is, before we get into the grungy details, what is the why am I telling you this? This is a way to formalize all the intuition that we have had about the method so far, ok. So, if you have a more theoretical bent of mind, you will appreciate formalizing it, formalizing all the hand-waving intuition that we have had so far, ok.

So, I am just going to state the result to begin with so that we can appreciate it, ok. We, so not we say using the CDM. So when I say using the CDM, immediately what does it imply? There were two requirements for CDM, which were there? As well, okay. Asymmetric positive definite, fine. After that, there were two requirements, which were? Give me a set of conjugate directions and then do exact line search along each one of those directions, those alphas. Those, that is what makes a CDM, okay.

Expanding Subspace Theorem.

Result: Using the CDM $\{x_k\}$, starting from x_0
minimizing ϕ

(1) $r_k^T p_i = 0$ for $i \in [0, k-1]$

(2) In an affine space $\{x \mid x_0 + \text{span}\{p_0, \dots, p_{k-1}\}\}$
 x_k is the minimizer of $\phi(x)$.



So that gave me a sequence x_k starting from some arbitrary x_0 . We were minimizing ϕ , my objective function ϕ . So, the first there are two consequences of the expanding subspace theorem. The first is $r_k^T p_i = 0$ for i belonging to 0 to 1. That is fine.

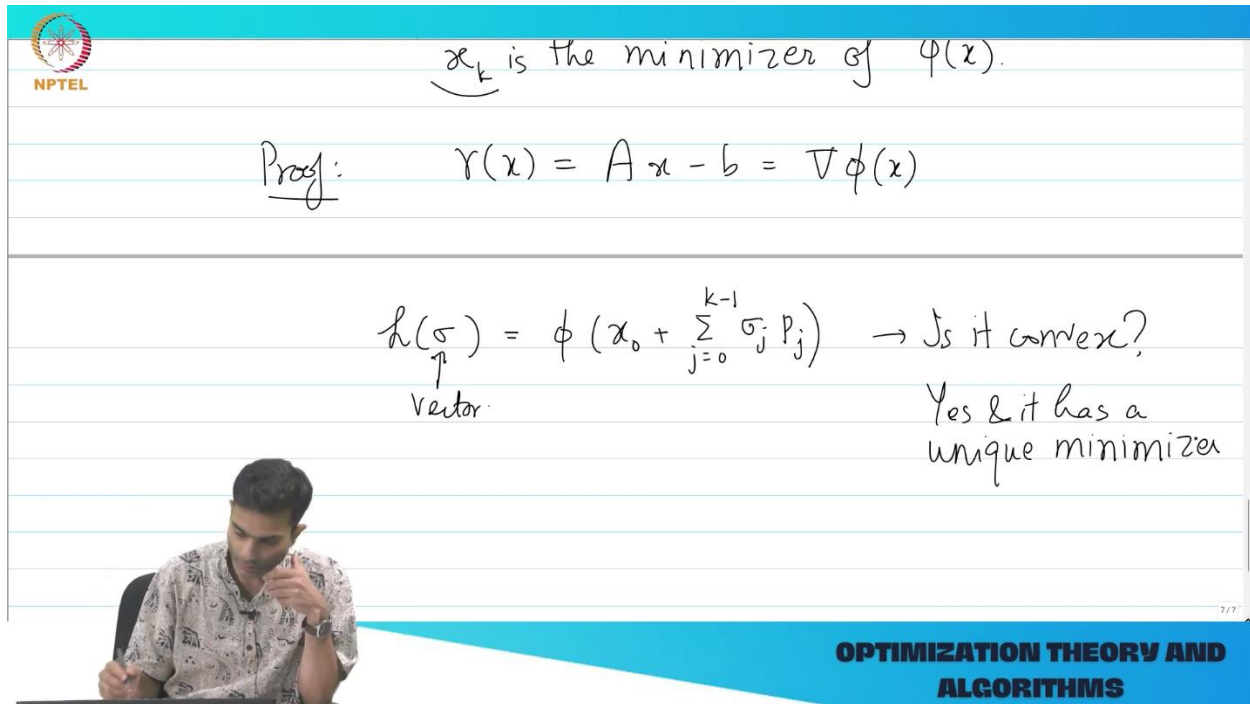
Let us just note it down and then we will give some intuition to it, ok. And affine space. So, these are the two consequences of the theorem. Consequences in the sense that the relation between these two statements is actually an "if and only if" condition, ok. So, let us, this it looks a little tedious, let us just interpret this in words. So, what is the first thing saying? Remember, keep in mind what we are doing is we are formalizing the intuition that we have had.

So intuitively we should immediately be able to appreciate what is going on. So, the first statement is saying the residual at the k 'th step, that is r_k , is orthogonal to what? To all the previous conjugate directions. We kind of expected that to happen, right, because every time we are going along one direction, we are never revisiting it. So, whatever is the balance in that direction is perpendicular to the new directions it goes into, right. So, you can see how this is a formalization of that intuitive idea.

That is step one. Is that clear what is being said? The residual at the k 'th step is orthogonal to all previous directions in which I have walked. That means I never revisit a direction, right. Second, we all know what a vector subspace is, right. I take a bunch of vectors and I say the span of that forms a vector subspace because their linear combination stays within that vector subspace. That is a vector subspace.

To a vector subspace if I add a constant term, it no longer remains a vector subspace because the origin is excluded. When I say span of vectors means linear combination of those vectors. If I set the coefficients of linear combination to be 0, what do I get? The origin. So, the origin always belongs to vector subspace. The moment I add x_0 and without any α_0 behind it, that means the origin is not guaranteed to lie inside.

So, this makes it instead of a vector subspace, it makes it an affine subspace. Fancy way of saying something very simple. So, in this affine space, so what is this affine space? Take the k conjugate directions so far. I am at step k , take the previous, I mean take the k direction so far. Out of that create some affine space, ok fine.



NPTEL

x_k is the minimizer of $\phi(x)$.

Proof: $r(x) = Ax - b = -\nabla \phi(x)$

$h(\underset{\text{vector}}{\sigma}) = \phi(x_0 + \sum_{j=0}^{k-1} \sigma_j p_j) \rightarrow$ Is it convex?
Yes & it has a unique minimizer

OPTIMIZATION THEORY AND ALGORITHMS

Constraining x to live only in this smaller space, try to minimize the objective function. It is saying that the x_k , the k 'th iterate, so x_k is not any x_k ; it is the k 'th iterate of my conjugate direction method, is actually the minimizer of ϕ . I am interested in a global minimizer of ϕ , that means x_1 to x_n all n coordinates I should, I want the best thing, right. So, if I were looking at the total solution I would say linear combination of p_0 to p_{n-1} , that is where my solution lives because I have n linearly independent vectors, the solution has to be a linear combination of the basis function, right. That is what I want. But this is saying something a little bit different.

What is it saying? If I restrict myself to a k -dimensional subspace. Why k -dimensional? Because it is being described by k basis vectors, not n . Now, when I am restricting myself to k basis vectors, what is this result saying? That by even though you are remaining constrained to a k -dimensional subspace, the solution that you get x_k is the best in terms of minimizing my objective function. If I want to minimize any more, I have to expand the subspace. That is why this name comes, expanding subspace.

So, at every step I am doing the best possible. So, in the first, so let us take it very simple. In the first step, where am I? $x_0 + \text{span of } p_0$. What is that? $x_0 + \alpha \times p_0$, that is an affine space. What am I doing? Finding the best α . I have a closed form expression for it, I get the value of α_0 .

Can I do any better for ϕ ? I cannot do any better. This is the minimizer along, why? I am constrained to go along only p_0 , fine, I got it right. So step 0, that is why. When I add now go to step 2, I have now p_1 also. What is the best I could do? So like this, I keep adding.

So I am growing the subspace in which my solution lives till finally in at most n steps I reach the final solution. Okay, so getting this intuitive understanding is far more important than following the nitty-gritty details, right? You will forget, even I forget the details of the proof after some time. But if you keep this in mind, how do you interpret this, right? Five minutes ago when you saw this, it just looked like mathematical symbols, but now it has come alive a little bit more, you know what is going on over here. So that's the key thing to keep in mind. And in general, this is a skill all of you should develop that when you read a theorem, the first step is to get intimidated by it.

The second step is to take literally every symbol and convert it into plain English. You see a transpose, orthogonal. You see r_k , you say a residual. You say p , you say conjugate direction. Residuals are orthogonal to conjugate direction.

So that begins to make a kind of a geometric picture in your mind. Okay. Any questions about interpretation here? I am going to give you half the proof so as to, you know, just to give you a flavor of it, ok. We already know one thing that this residual is $Ax - b$, okay, just to remind you, which is also $\nabla\phi$, ok.

NPTEL

Proof: $r(x) = Ax - b = \nabla\phi(x)$

$h(\sigma) = \phi(x_0 + \sum_{j=0}^{k-1} \sigma_j p_j) \rightarrow$ Is it convex?
 Vector Yes & it has a unique minimizer

Chain rule: $\frac{\partial h}{\partial \sigma_i} = 0 \quad \forall i \in [0, k-1]$
 because these $\sigma_i^* \equiv$ minimizer

$\nabla\phi(x_0 + \sum \sigma_j p_j)^T p_i = 0$
 $r(x_k)^T p_i = 0 \quad \forall i \in [0, k-1]$

I am going to define a function h , which is a function of σ 's, which is something like this: $\phi(x_0) + \sum_{j=0}^{k-1} \sigma_j p_j$. So how many σ 's do I have? The σ over here is a vector. I have k σ 's. So what is this function telling me? This is the function that I get by restricting my x to be in a k -dimensional space or an affine space.

Is this function, is it convex? How did I define ϕ ? $\frac{1}{2}x^T Ax - b$. So, is this also affine, I mean is this also convex? Obviously, this is also convex. If it is a convex function, does it have a minimizer? As a minimizer, we have seen that, right. So, therefore, the answer is yes, and it has, in fact, the minimizer is, is it unique? For a convex function, is a minimizer unique? Okay.

Right. So if it has a unique minimizer, that means the rate of change of this function with respect to each of the σ 's, what can we say? So I'm asking. Close. I can write, I can, so we can expand this by chain rule. But what do we expect the answer to be at the optimal? 0. If this function has a minimizer, it will be expressed as some values of the σ 's. So, the rate of change of this, therefore, is going to be 0 for all i belonging to 0 to $k - 1$.

Because why? It is the best, because it is the minimizer and here is where the chain rule comes in. Immediately I can write this partial derivative, how do I write it? Gradient of this whole thing $x_0 + \sum_{j=0}^{k-1} \sigma_j p_j$ transpose, multiplied by what? If I just apply right $p_i = 0$, right. This is just use of chain rule, I did not do anything special over here. I did chain rule over here and we are almost home. Why? Because now when I look at the expression for the residual, the $\nabla\phi$ is also the residual.

So, this implies that this is the residual now at right, or let me put it like this: x_k right transpose $p_i = 0$ for all i belonging. Notice I am constraining x_k to be only in the affine space, it is not in the entire space, it is only in this affine space and I am showing you that the residual is orthogonal to all the previous conjugate directions, ok. So, I mean it is not surprising, I expected this to happen, it is there is nothing very complicated over here, ok. So, actually what have we proved over here? We have said, we have assumed 2 to be true and we have proven 1, that is what we did right. We assumed that I pick an x from this affine space, ok, which is the minimizer of ϕ_k . If it is the minimizer of ϕ_k , what follows? $\frac{dh}{d\sigma} = 0$, chain rule gives me this which leads to the residual at the k 'th step is orthogonal to all previous search directions.

So, I got my given 2, assuming 2 I got 1, ok. I can, as you can imagine, I can also repeat the process where I start from 1 and go to 2, ok, but we will not do it here, it is not very difficult. If you see the proof in the textbook, they do it by induction, it is a nice refresher on induction. Okay, so this is what this is, you know, all the background that I wanted to give you until we start in the next class with the conjugate gradient method, okay. So any doubts in what we have done so far? You notice there is nothing very complicated.

We are basically this is like applied linear algebra, we are using properties of conjugacy and our calculus over here. So, we will do the conjugate gradient method and then as soon as we have done the basic version of the conjugate gradient method, we are going to have a race between steepest descent and conjugate gradient method and we will see how the two of them perform head to head.