

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 - 37

Discussion on doubts

All right, so just wanted to go through the first one was the first it's not really a doubt but it's more like a observation which I want to talk about out loud because a lot of students go through this and very few ask this question out loud. As a student I've also had this question. So the question is I find it difficult or a bit stressful to enjoy learning a course which I innately like after making mistakes or not performing well in exams. Any suggestions on outlooks so that I don't doubt myself or question my capability? Does this sound familiar? To whom does it not sound familiar? Anyone wants to raise their hand? Okay, so I mean, it's familiar to everyone including me, right? As a student, I've also had some courses where I've just, my head is just above the water. Okay and you know so what is it that I can suggest is there is no short term answer for it because I tell you something it will not really make sense but I can tell you with the benefit of some time having passed after let us say one, two or three years have passed from this course you will forget the bad experience of mistakes and poor marks you forget but if you found the course interesting you remember that part.

You remember being, you know, feeling curious in class, finding the content interesting. And that is the part that stays with you. And I can tell you this by personal experience. So many courses during my PhD I found very difficult because I tried to switch out of electrical engineering into physics.

So I tried to cover up all the physics courses, taking the courses with other students who had had a bachelor's in physics. So it was very hard, right? So every problem, and there we used to have every week one problem set for every course. So it was very hard. So there were some courses where I barely made it through. But after having graduated or even a few years after that, I felt, okay, I'm so glad I did not drop these courses.

One is that you remember the good stuff. The bad stuff has a shorter time constant. The second thing is that, let us say you're interested in making a technical career. And later on, courses like optimization and probability, linear algebra, they're very basic courses. No matter what you do, the next 30 years, you're going to use these tools.

Now what happens if you drop a course as important as, let us say this or any other basic courses. The next time you come across the next new fad, today it's machine learning, tomorrow it's, I don't know what, machine unlearning or something like that. Every time you go to learn that course, you will face one big mental block. Oh, I couldn't understand optimization. There is no chance I'm gonna understand this new area.

And it's just a mental block. So you are making it harder for yourself in the future. The other thing is that, I mean, this is a practice for real life. In real life when you step into the real world of either a job or whatever else, you are going to have multiple deadlines that you cannot meet, multiple expectations of you that you cannot meet. This is relatively much, much simpler.

So it's more like training ground. It's difficult, okay. It's difficult for everyone. Try to make it work. And I can assure you, you will not regret it later.

It's easier. So every time there's a difficult lecture, let us say, or a difficult quiz, I see the drop requests that come in my workflow. Okay, and what is it saying that the first sign of difficulty you say let's drop but that's bad practice in general for life right it's never going to this is the easiest you will get it it's only going to get tougher so may as well you know get on with it. Anyone wants to share anything related to this, anything that helps them to not drop a course or not feel lousy about themselves? Anyone? That you tried and it worked and you felt this is useful, someone else may also learn from it. Yeah. You are taking the course for the second time.

Okay. So, what made it different for you this time? So, let me do it this time. So, if your drop request comes, I am going to reject it. All right, good, good, good. That's a good spirit. Anybody else? Problems.

Okay. Well, from now you are going to have a steady stream of tutorial questions and so on. So no worries on that at all. Okay, let us continue. Let us see. So there was, there were a couple of doubt sheets with an unusual naming convention.

So it says in today's class I got an introduction on conjugate gradient descent. This method is not conjugate gradient descent. This is conjugate gradient method. There's no descent in the name of the word. If you use the word descent, you're gonna confuse yourself a lot.


So is this conjugate gradient method? The trouble is if we think, if we put the word descent in here, first of all, it's wrong. Second of all, you're going to think that somehow there are descent directions involved, but this is a different view or a different approach to this optimization problem. Can the starting point x_0 be any point in the conjugate gradient method? Answer is yes. It can be any point, it doesn't really matter. When can this method fail to converge? Any example, right? So when the method, as long as, so what are the prerequisites we had for this method? Did we have any requirement on A ? Symmetric positive definite.

As long as that's the case, this method is going to converge in how many steps? At most n steps. That is a guarantee and there is a theoretical guarantee behind it. We have seen one proof that shows it to us in that way. How do we get the p_k 's for conjugacy? Right, in whatever we did so far. Where did it go? Visualizing quadratic forms.

Yes. I just introduced the idea of conjugacy. I never told you how we are going to get the p_k 's, but I mentioned that the devil lies in the details and how we will get it. I will talk about it in this class, okay. Given the matrix A , how can we tell whether conjugate vectors exist? So, supposing I give you a matrix A , how can you be sure that any conjugate vectors will exist or not? It is positive definite, so it is full rank, or not they exist again is something that we will see, okay.

Maybe we can have a look at this right away, okay. Ok. So, let us as practice say A is symmetric positive definite.

Yeah. Correct. I am coming to that and there is one question exactly without doubt. Ok. So, this is given to you, do conjugate vectors exist? What is our intuitive feeling? Intuitive feeling is yes.




\hookrightarrow A is sym pos. def. Do conj vects exist?
 Yes. $p_i^T A p_j = 0 \quad i \neq j$

$A = U \Lambda U^T$ u_i 's are orthogonal.

$AU = U\Lambda$

$Au_i = \lambda u_i \xrightarrow[\text{by } u_j^T]{\text{LM}}$ $u_j^T A u_i = 0 \quad i \neq j$



OPTIMIZATION THEORY AND ALGORITHMS

It is sufficient for us to come up with one example when it happens. So, what is the definition of conjugacy? $p_i^T A p_j = 0 \quad i \neq j$.

So, how do we, I mean this definition is fine, but this definition is it going to help me or can it help me construct a set of conjugate vectors? So, let us start with our road into this problem starts with the property of symmetric positive definite matrix. What is it for sure we can say about a symmetric positive definite matrix? It is invertible, something stronger, it is diagonalizable or in other words it has an eigenvector decomposition, right. That means I can write $A = U \Lambda U^T$, right. What happens if I multiply on the right with U ? $U^T U = I$ because the eigenvectors are orthogonal.

So, we did this last time $AU = U\Lambda$, right. Now, if I just take one, take this guy multiplied by one guy, what am I going to get? $U_i = \lambda U_i$, this is basically the what? Eigenvalue problem right. Now, what if I left multiply by U_j^T ? Right. So, left multiply by U_j^T , I am going to get $U_j^T A U_i = 0 \quad i \neq j$. So, what is that in plain English? What does that mean? The eigenvectors are a candidate for conjugate directions, right.

They satisfy this property that is all I need for conjugacy. So, eigenvectors are... So, that means it exists and will there be n of them? Yes, because there are n eigenvectors even if the eigenvalue is repeated the eigenvectors are still distinct. This is a result from linear algebra. So, I am going to have n eigenvectors. So, there is no problem I will give me a symmetric positive definite matrix I can find you at least one set of conjugate vectors. There's no guesswork involved here.

Does this mean that this is the only candidate? No, I mean, I just took one example. This does not in any way tell you that you are limited only to eigenvectors, okay? So this is how you get a system. I mean, this is one way of getting conjugate direction given the matrix A . Any doubts on this? Okay. Can you please explain once more how the step lengths, exact step lengths can be calculated for convex quadratic functions? So this was also a tutorial and also a quiz problem,

right? So if I give you a, no, this is f rather, right? So, who has trouble with this problem still? Should I, ok, let me put it the other way.

Who would like me to derive this? How do we derive the exact step length given a quadratic cross function? Anyone? Everyone is fine with it? Something yes, no? Ok, who wants it raise your hand? Waste of time. So, everyone is familiar with it. In fact, the solutions were released and that also has it, right? It is what is the calculus concept being used? Essentially, chain rule to calculate $\frac{d}{d\alpha}$.

That is what gives me the expression. And in the case of a quadratic function if I had my $f(\mathbf{x})$ was like this

$$\frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x} - \mathbf{b}^T \mathbf{x}$$

what was ∇f ? We could write it analytically it was $\mathbf{A} \mathbf{x} - \mathbf{b}$, right? So, when I substituted this into the expression for $\frac{d}{d\alpha}$, I straight away get the exact step length, right? Ok. Yeah, so this was another student who said in the conjugate gradient descent method as I mentioned it is not gradient descent method, why are the \mathbf{p}_k 's valid descent directions, right? So this question really does not arise because we are not going at it at this problem from the point of view of descent. It will turn out that descent is also a quality and property of it, but that is not how we are looking at it and how do we calculate the \mathbf{p}_i 's? I give an example. Ok, Omkar asks will performing Gram-Schmidt on the conjugate \mathbf{p}_k 's give us any results which are advantageous? Ok.

Well once you get a set of conjugate directions there is if I do Gram-Schmidt the standard Gram-Schmidt on it what will happen to them? They were already linearly independent, they will become orthogonal. Will they continue to be conjugate? Is it necessary? I mean, it is not obvious

to me. Take a bunch of vectors, they are linearly independent and A -conjugate. Now, I do Gram-Schmidt to them. I do not see any reason why they will continue to be conjugate with respect to A .

So, that is not necessary. So, it is not going to give us any advantage. However, he has inadvertently asked a good question. A modified Gram-Schmidt method is a second way by which you can arrive at a set of conjugate, ok. So, this will be one of the problems we will work out. Start with a set of linearly independent vectors.

So, you all know Gram-Schmidt, the recipe for Gram-Schmidt. Start with a set of linearly independent vectors, apply Gram-Schmidt and that results into what? A set of orthogonal or orthonormal vectors. You modify that process a little bit. At the end of it, you get a set of vectors which are A -conjugate rather than orthogonal.

It is straightforward, we will work it out. So, that is the second way of generating conjugate directions, but do not worry we will discuss it. So, this question seems to be by someone who knows a little bit more about optimization. Suppose instead of finding $A\mathbf{x} = \mathbf{b}$ we want to add a regularization term as well, can CG method be used for such a problem? Ok. So, the question is that let's take for simplicity. So we said for when we started the conjugate direction method, we wrote our cost function like this, right? And we said that this is a better way than this approach, right? And the reason was this guy, this second method had what? Had a condition number of κ^2 , this had a condition number of κ^2 , this had a condition number of κ .

So, numerically the first method is going to give us better results and the second method will not give us as good a result because of the squared condition number. Now, the question is can I add a regularization term? So, what this you will come across this word regularization very very often. What does it mean? It means adding some term over, say for example something like this. Now, what does this mean? We will study this towards the end of the course, but let us just get a sense of it because you will come across it in papers that you read and all. What is the meaning of this? Let us just intuitively understand this cost function.

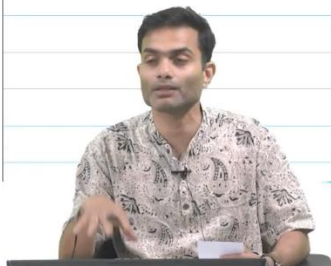


\Rightarrow Eigenvectors are conjugate directions.

$$f(x) = \frac{1}{2} x^T A x - b^T x \quad \text{--- } \textcircled{1}$$

$$\nabla f = (Ax - b)$$

$$\phi(x) = \left[\begin{array}{l} \frac{1}{2} \|Ax - b\|^2 \quad \text{--- } k^2 \textcircled{2} \\ + \lambda \|x - x_0\|^2 \end{array} \right]$$



What is it saying? When will you be happy? In the real ideal world when will you be happy when the cost function goes to zero, that's the best case scenario. For the best case scenario to happen in this case, what should happen? Ax should be equal to b and x should be equal to x_0 , that's the only possibility. Is that always gonna happen? Times is not going to happen. Why? Because x_0 need not be the solution to $Ax = b$. Which means if I go to the solution $Ax = b$, this first term is 0, second term has some non-zero magnitude.

Agreed? Come to the other side. Supposing I said $X = X_0$, second term is zero, first term need not be zero. So in both cases, there is some weight on either side. But now you say, I want you to optimize this. So the solution that you come across will neither satisfy this, the first term nor the second term, but the value of $\Phi(X)$ would be lesser than the first two cases. Now, why would you be interested in such a weird solution? So this is a way of adding a constraint to the problem.

You want to solve your original problem $Ax = b$ but you also want your solution to be close to some vector which you know has some nice properties. All those details are not important. You want the solution to be close to x_0 , right? How close to x_0 is controlled by this? This empirical parameter λ , which is going to be problem-dependent. So, you can at $\lambda = 0$ you are driving the solution to $x = b$ at $\lambda \rightarrow \infty$ you are driving it to $x = x_0$. So, this is a very very powerful and common technique in image processing, machine learning, and so on.

Add all the constraints to your problem in this way, right? This is simply here I have written the norm of $x - x_0$. Another way could be, for example, what could it be? Ok, a common example in compressive sensing is something like this, ok. So, let us ignore this and write this. How do you interpret this? I want a solution to be as close to $Ax = b$, but what else do I want to minimize? The one-norm of x .

So, one-norm of x is something which if you study signal processing, is something which tends to promote solutions that have very few non-zero entries. So, you know, there are image

processing applications where this gets used, right? So, there are lots of tricks you can play with this. Then the next question is how do you get λ , all of those things, right? So we'll study constraint optimization in the second half of this course. So this term which I've added over here is called a regularization term.

Okay. So the question is, can the CG method be used for this problem as well? It depends on what is the regularization I add. This whole thing I can rewrite as a convex function. If I can write it as a convex function, then I can use my standard CG. If I cannot write it as a standard convex function, I can still use CG, I may have to go to a non-linear CG method which we will study after we finish the linear CG.

So, the CG method is a very very powerful method. Right now it looks like we are using it to solve just $A\mathbf{x} = \mathbf{B}$, but soon you will see you can generalize it to a wide variety of problems, ok. Any questions on this? Has anyone heard this word "regularization term"? Right, you have seen it everywhere. So, at least you have a basic idea of what it means over here, ok? Okay, the question is how are we, so let me go to the, okay, the question is I wrote $\mathbf{X}^* - \mathbf{x}_0$ is the linear sum, is a linear combination of the \mathbf{p}_i 's. How are we writing this to start with? So, the motivation is just coming from linear algebra.

If I am in an n -dimensional space and I have n linearly independent vectors, can we say they form a basis? They form a basis. The meaning of forming a basis, I can express any vector as a linear combination of the basis. So, the right-hand side is simply a linear combination of the basis vectors and that is any vector. Is $\mathbf{x}^* - \mathbf{x}_0$ any vector? It is. So, I just use simple linear algebra to say that the difference from initial to final point is some linear combination and then I showed the nice part of this at the end of it was that the coefficients of the linear combination ended up being exactly the step lengths that I chose in my method.

So, if I start with \mathbf{x}_0 and if I make take these k values of α , I land up exactly at \mathbf{x}^* , the solution, right. Is this clear, whoever asked this question? Yeah, the other question was about descent direction. So, if you look at, yeah, if you look at the expression for α , let us just for sake of argument assume that my \mathbf{p}_k was not a descent direction. We are not interested in this descent direction analysis, but let us just take it for sake of argument. Supposing \mathbf{p}_k is not a descent direction, what will be the sign of $\mathbf{r}_k^T \mathbf{p}_k$? It will be positive.



① The p_i 's must be conjugate w.r.t. A ← however

② Step length is an exact minimizer of $\phi(x)$ along the p_k direction, $\frac{d}{d\alpha} \phi(x_k + \alpha p_k) = 0$



$$\alpha_k = - \frac{r_k^T p_k}{p_k^T A p_k}, \quad r_k = Ax_k - b.$$

Starting from x_0 the sequence $\{x_k\}$

OPTIMIZATION THEORY AND ALGORITHMS

If it is not a descent direction, it is going to be positive. Why? Because r_k was what? ∇f , right? So, for gradient descent, I needed to be in the quadrant of $-\nabla f$. So if I, if it is not a descent direction it will be positive. That means what will be the sign of α ? The sign of α will be negative, right. Assuming of course the denominator is positive, okay.

If that is the case, look at here, right. So I have $x_k + \alpha_k p_k$. α_k is negative and p_k is what it is. Now if I define a new vector $q_k = -p_k$, it looks just the same? It will look like a step in the direction q_k , which is now $-p_k$, therefore is it a descent direction? It is a descent direction.

Is the step length positive? Step length is also positive. So, we are back to square one. So, we need not worry too much about descent direction over here in the conjugate gradient scheme of things. It is a different approach and today's class I will try to give you a geometric interpretation of how it is a different approach, ok. What was the motivation behind this method? Okay, this when we do the graphical example I will tell you what is the. So, there was one question I think does linear independence imply conjugacy? So, if two vectors, if a bunch of vectors are linearly independent are they going to be conjugate as we have discussed there is no real compelling reason by which you can say immediately yes they are going to be conjugate and it is not necessary. The question goes on and all of these will be answered graphically, okay.