Course Name: Optimization Theory and Algorithms
Professor Name: Dr. Uday K. Khankhoje
Department Name: Electrical Engineering
Institute Name: Indian Institute of Technology Madras
Week - 04
Lecture - 28

Line Search - Convergence and Rate - 1

Okay, so this is just sort of the geometry of how a gradient descent algorithm would proceed. You would have this exact behavior and orthogonal set. Now, yes, question. Yeah. Correct. This proof is not telling us; this proof in fact requires me to take two steps only then I can talk about orthogonality, but what we are saying is that there are some special directions where I just reach in one shot that is not covered in this only we will, but we will prove that later we will figure out when does it happen. Can I say anything about it? No, this is in general what we are saying is that if the contours are circular, I will reach in one shot. If the contours are anything but circular, I will go in a zigzag way with the property that every subsequent step is at 90°.

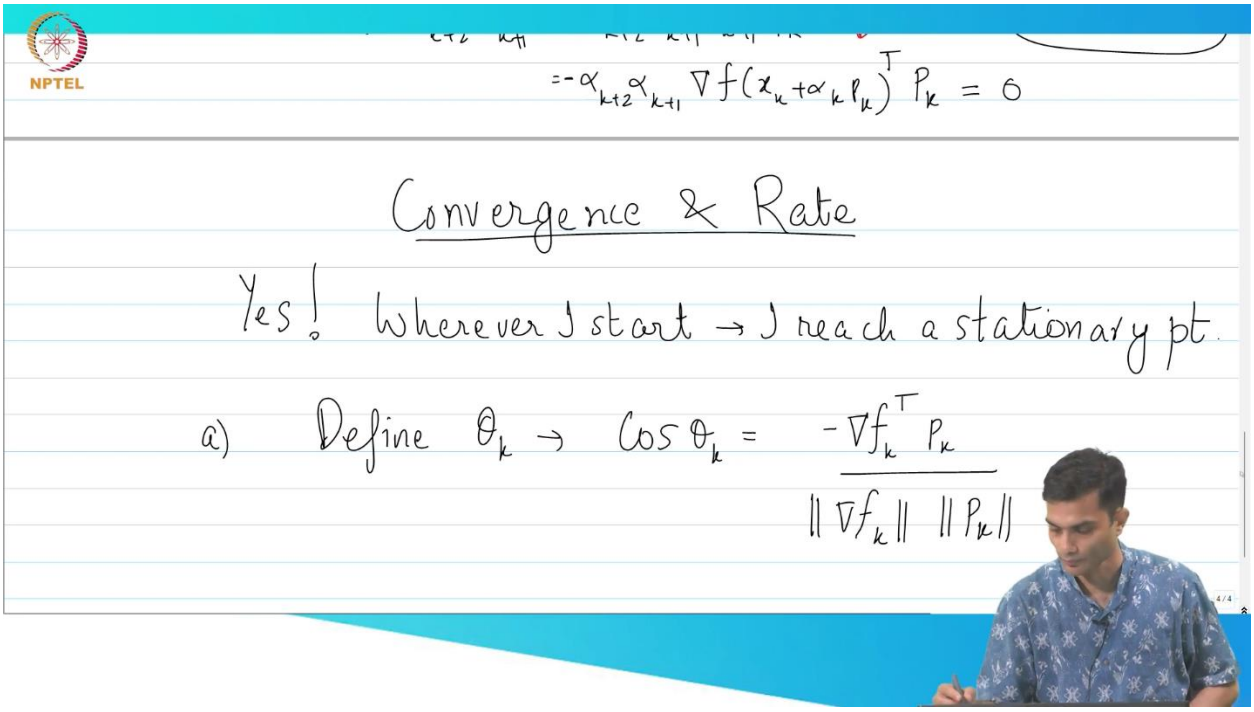
If it so happens, if all the stars align that the gradient. No, it will actually have to be circular because at each point you are, you know, you should not; the function value should not increase your algorithm will end there. So, if you have a circular contour on the outside and some funny contour inside, then if you are going along that line, it is possible I can construct some kind of a landscape where at some point you will begin a ascent. The contours being circular is a sufficient condition, but this is—is it necessary?

In this case yes, and again as I said we will come to it. We will come to it when we do the next thing which we are going to do. We have spoken about descent directions; we have spoken about Wolfe conditions—what are the conditions that the step length should satisfy—but so to speak the elephant in the room; no one has asked the question: is there any guarantee that gradient descent converges? Does there exist a solution which is hoping that α and praying that a solution exists, but people would like to know does a solution exist? Can you prove it to me that this algorithm will converge, right? So, this is really important: two questions—convergence and the convergence is kind of does it converge yes or no? If it converges that means it exists; the second is rate at what rate do I converge.

So the good news is that yes, there is convergence. In the sense that wherever I start, let us write that down, I reach a stationary point. In other words, wherever I start, I will reach the nearest local minima. So, this is not a proof that it will reach the global minima. So this is saying that wherever, whichever, so to speak, whichever valley I start in, I will reach the bottom of that valley. Whether or not that valley is good enough for you is—I mean your call. So, I am going to go about the proof of this now.

So, that we get particularly for this algorithm, I am going to do the proof because this algorithm—the steepest descent or the gradient descent algorithm—is used so extensively in AI and ML, which is why I think it is important that you know the proof: how does it go through? There are like 10,000 variants of this proof depending on the latest algorithm (Adam or this, that, and the other), which people use to calculate the step length and all of that. The skeleton of this proof is used by all other guys, so we will prove this over here.

Ok, so let us write down a few things that we will need. One thing is the angle. Remember we spoke about when we spoke about whether or not a direction is a descent direction? We spoke about the angle between p_k and the negative gradient; that angle should always be in a cone of 90°. So, to quantify that a little bit better, I am going to define this angle, right? So, I am going to define θ_k; how will I define it? Simply by an inner product. So, what will be for example, $\cos(\theta_k)$ equal to?

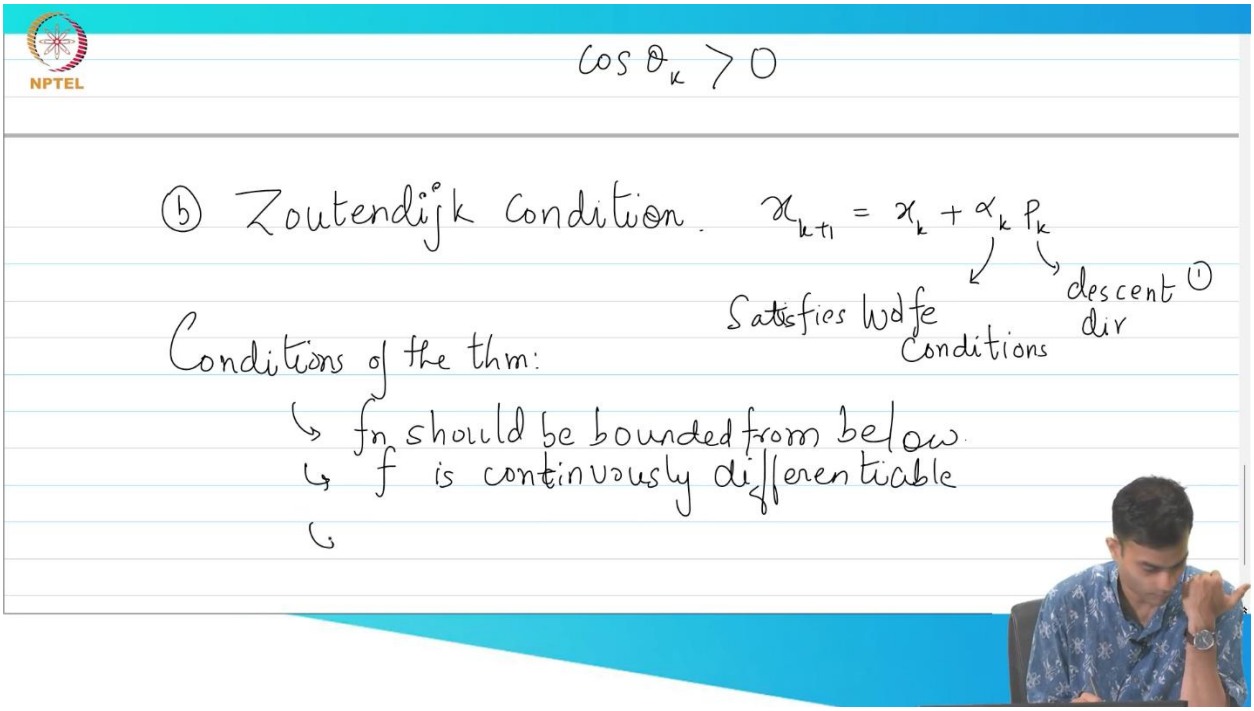


We just use the definition of the inner product. What is the inner product of—if I write, if I say $A^T B$—what is it equal to? $\| A \| \| B \| \cos$. That is the same thing; I am going to use here and what are the two characters I am interested in? The descent direction p_k and the negative gradient. So, this is going to be $-\nabla f_k^T p_k$, and since it is cosine, I need to normalize. Obviously, $\theta_k = 0$ corresponds to what? If I were to choose $\theta_k = 0$, I am picking which p_k? The negative gradient. The negative gradient, right. So, this is steepest descent.

But in general, I can have $-90 < \theta_k < 90$. Ok, and when this is the case, can I say, can I put down any inequality for $\cos(\theta_k)$? If the angle is always between -90 and 90, what value does cos take? It is always positive, right? So, this is fairly simple trigonometry that we need to keep in mind. So, this is just some definition that we will use; we have not started the theorem yet. So, let us note this down.

So, now point B is actually—point B, I am going to state the theorem and then we will sort of get some intuition and then try to prove it. It is a difficult spelling also to write this person's name; I do not know which country; probably Dutch. Zoutendijk's condition is called. Remember we are doing line search; it implies we are doing something like this, right? So, what are we assuming? We are going to assume that this is a descent direction. This is the first thing that we are assuming.

We need to assume something about α_k. What is the most reasonable thing that you would say α should satisfy? Based on what you know so far, supposing you were in charge of this algorithm, you would not be ok with some random α, right? You would want α to satisfy something; what is that something? That is exact line search; that is too strict. Something lesser than that we have studied it. Wolfe conditions, right? Wolfe conditions is sufficient decrease and the curvature condition, right?



Satisfies. With these two ingredients in place, I am all set for doing gradient descent. Insisting on exact line search may be too strict. So what are the conditions of the theorem? They are all very, what should I say, reasonable conditions. The first condition is the function should be bounded from below. I call this the idiot check.

Why should we call this the idiot check? Supposing I say find me the minima of $\frac{1}{|x|}$, it does not exist, right? I mean minus infinity is not a number or whatever, right? So, the function has to be bounded from below; only then does it even make sense for you to do optimization, right? So, this is very, very much common sense. The next two conditions are basically to ensure I can do calculus. So, they are going to be about continuity and differentiability, right?

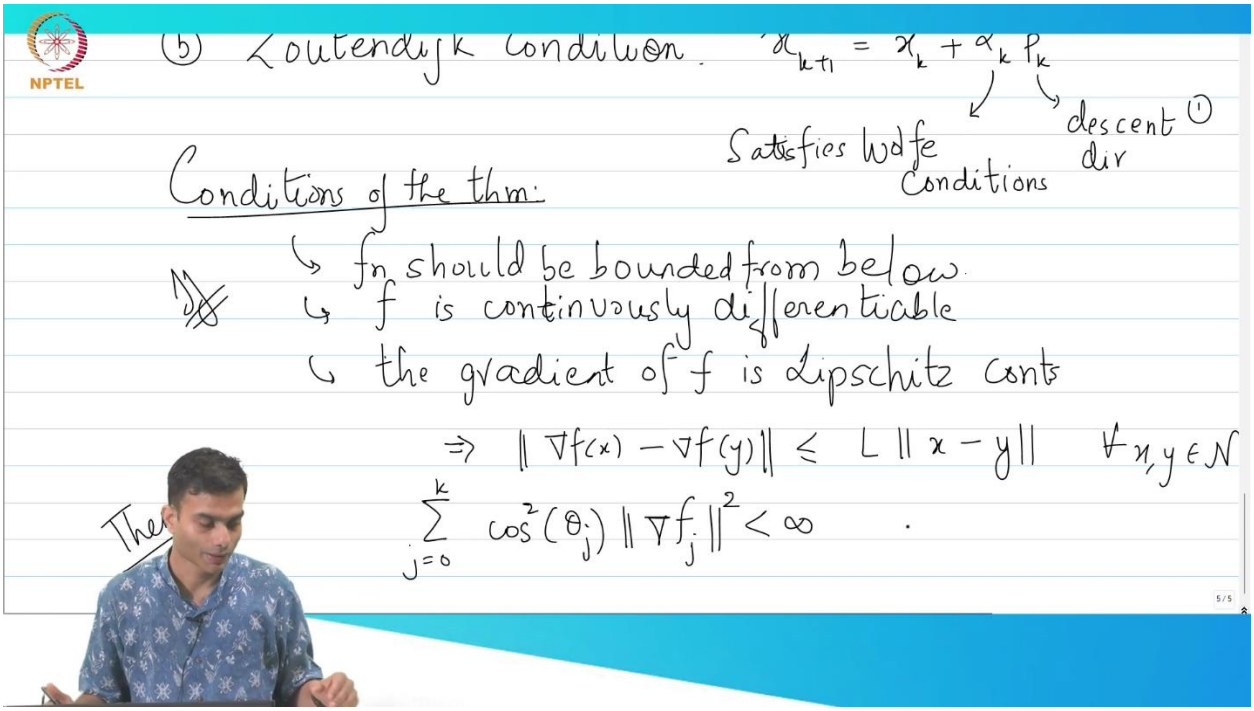
So, we will just note that down: f is continuously differentiable and the second is that the gradient of f is Lipschitz continuous. We have all had a review of Lipschitz continuity simply, and I am saying that the gradient should be Lipschitz continuous. That means if I take $|\nabla f(x) - \nabla f(y)|$, what should this be? $\leq L \parallel x - y \parallel$ for all x and y in some open neighborhood, ok. So, this is what I need; these are the conditions of the theorem, right. Now, if all of these conditions are satisfied, now I will write down what is this Zoutendijk condition, right.

So, this is you can think of this as the if part and then I have the then part. Ok, then looks a little grungy, does not look like what you are expecting; it is saying that if these fairly straightforward

conditions hold true, then this big summation is what? Not blowing up, basically, that is what it is saying. It is less than infinity means it is a bounded number. k is left to us; that means K can be as large as I want. So, it is saying that take k to be whatever you want as large as you want; this summation somehow is always going to be less than infinity means it is a number, it is a finite number, right.

So, from here before we sketch out the proof, can you kind of guess how does this imply that gradient descent converges? Just intuitively, because it is better to get an intuition before we go into the grungy details. Let us assume that you know the proof and now you have the statement in front of you; this summation is less than infinity; what can we say? That means it is the summation; I keep adding terms; are each of the terms positive? $\cos^2$ is positive; $\| f_k \|^2$ is positive. I keep adding positive terms, keep adding, keep adding.

So, the only way that this can happen is if the sequence must have to approach 0. So, if that is true, I have that if $\| \nabla f_k \|^2$ must approach 0; that is one term and the other term is this term, right? So, both of these terms must approach 0; this must mean that either α_k approaches 0 or ∇f_k approaches 0. So, in either case, we will land up concluding that the algorithm converges, and you can see this goes back to your original idea of when you are walking down a valley; you are either very close to the edge or you are very close to the bottom of the valley, right. So, let us sketch out this proof.



So, just to recap: the function has to be bounded from below, the function is continuously differentiable, the gradient is Lipschitz continuous, and I am asserting that this must hold true. So, now I will just sketch out the proof before I run out of time. So, we will write down from what we just spoke about; we will write down k arbitrary points and write down the value of the function at those points at the minimum of $f(x)$ at the minimum of $f(x)$ let us write that f^* which is a bounded number; we know that this number is finite.

So, this is just a point. So, for every k I can compare this at $k = 0$. So, we are going to write down a couple of important inequalities. $f_k - f^*$ is greater than $f_k - f_{k+1}$. What is happening here is I am moving in a descent direction, right, so this is a downward step for us. The next inequality we can obtain from Wolfe conditions because we insisted on Wolfe conditions; I am free to assert that $f_k - f_{k+1}$ is greater than or equal to $c\alpha_k \| \nabla f_k \|^2$, where c is one of the Wolfe coefficients we spoke about, right. So, combining these two inequalities, this must hold because we are doing descent; as long as we are doing decent direction at all times, this should be true.

So, $f_k - f^*$ is greater than or equal to $c\alpha_k \| \nabla f_k \|^2$. So, we have $\| \nabla f_k \|^2$ is less than or equal to this quantity. Now note that α_k is bounded away from zero; otherwise, I cannot use Wolfe conditions, right. Hence, $\| \nabla f_k \|^2$ must approach 0. Now I need to say what happens with that summation. So, now we go back to the condition. We will analyze this condition.

We said that if I keep going in this direction for each k, there is a condition we know; this should hold true. Hence, let us see what happens with that quantity. If I take a summation of the squared norms up to some K, this becomes $\sum_{k=0}^{K} \| \nabla f_k \|^2 < \infty$ because we are bounded in this case. So, we have this summation and it is less than infinity.

If it is less than infinity and we have also proved that $\| \nabla f_k \|^2$ must approach 0. Hence, for large K, $\| \nabla f_k \|^2$ must go to zero; K can go as large as we want. So, we have $K \to \infty$. So, therefore, $\| \nabla f_k \|^2$ must approach 0. So, this gives us a valid way of proving that we are converging. Hence, we can write down $\lim_{k \to \infty} \| \nabla f_k \|^2 = 0$; so either I am converging to a stationary point or you can think of it as a local minima.

So, before we wrap up, does that kind of all make sense? Is that intuitive? Because you are proving something simple. We have a nice theorem on convergence and you can ask the converse question. If the converse was true, the convergence was guaranteed, and if it was not satisfied, what would happen? If $\sum_{k=0}^{\infty} \| \nabla f_k \|^2 = \infty$, what does it tell you? I am stuck in a loop. You may not be able to find a minimum, right? You may be oscillating or you may be going away from the solution, right. So, one of the things that we have seen so far, so let us wrap up, and I will make sure we summarize the conditions before we dive into the next chapter.