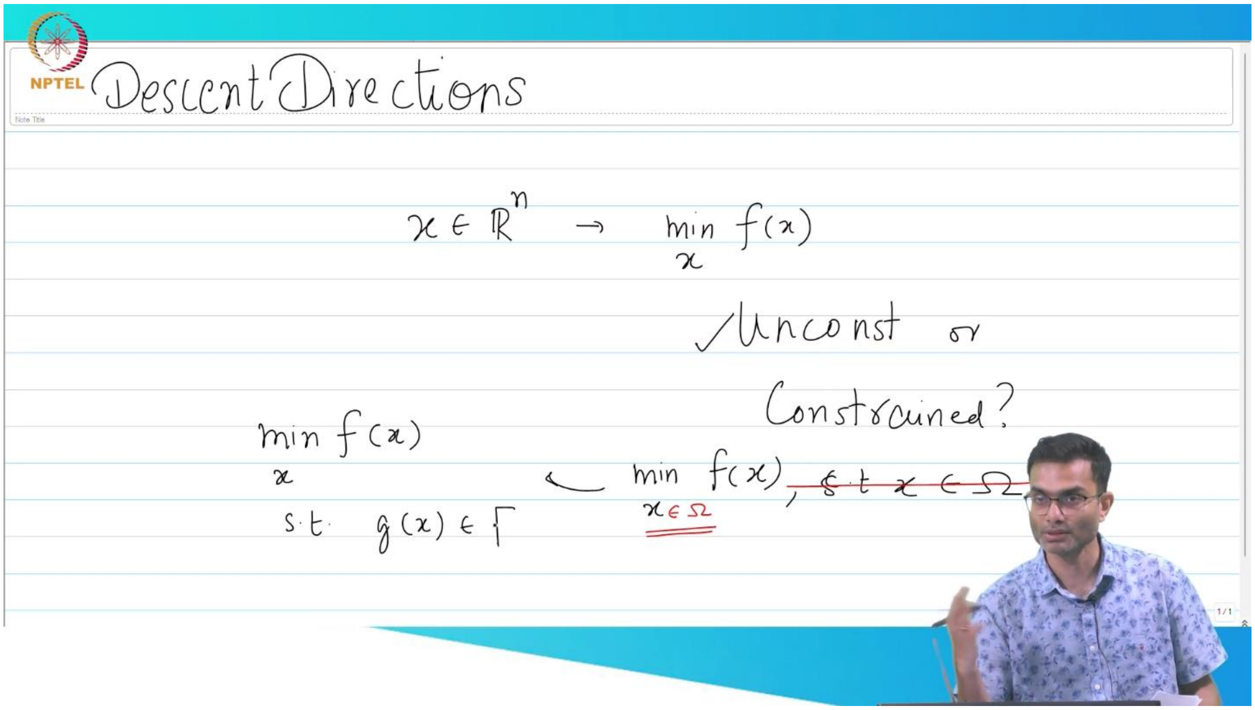
Course Name: Optimization Theory and Algorithms
Professor Name: Dr. Uday K. Khankhoje
Department Name: Electrical Engineering
Institute Name: Indian Institute of Technology Madras
Week - 03
Lecture - 19

Unconstrained Optimization - 5 - Properties of Descent Directions Steepest Descent Direction

So, let us start with the doubt sheets. So, there is a student who has misheard and heard the complete opposite of what I said. Sir, you said that we will be dealing with convex functions only, the other way right. We will not be dealing with only convex functions; we will be dealing with functions in general. They may be convex; they may not be convex. Now do we need to check whether the optimization function is convex before solving the problem for exams and assignments?

I do not think you should make any assumptions about whether the function is convex or not. For convex functions, a local minima is a global minima, but are the majority of the objective functions convex? If not, is finding a local minima sufficient? So, we discussed this in the previous class, right? For convex functions, a local minima is a global minima. So, if you have found a minima, that is it; your search is over. But that said, I would say that most real-life engineering problems are not convex, ok.

People try really hard to somehow make a convex approximation of their problem, and the reason is simple: once you do that, you just need to find one solution. So, this right in the beginning of the course we had spoken about the complexity of modeling the problem versus how simple it is to solve the problem. So, if you make it really simple and make it convex and solve it, then you have to see how realistic that model is, right? So, it is never an easy answer. Could you share some open research problems and papers on choosing the initial points and the other things? So, before we get to research problems and papers, let us go clear our textbook material first, right, which is what we are doing now.

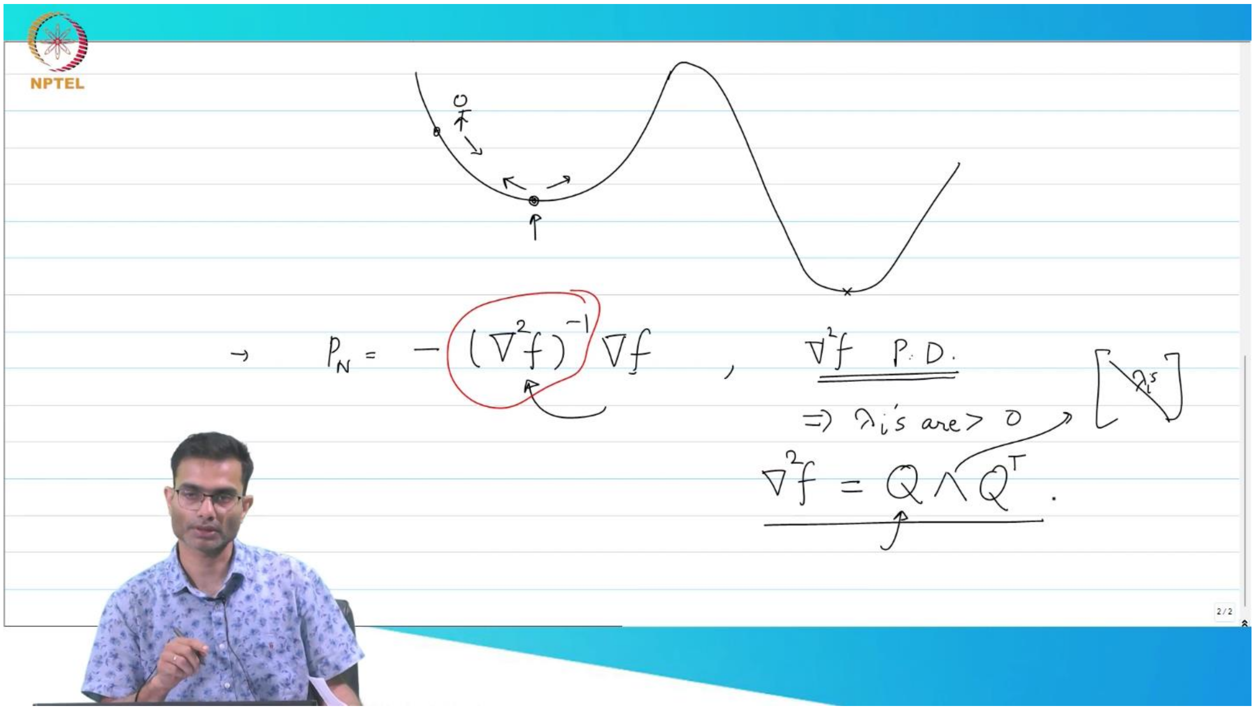


Can we also have a lecture on semi-definite programming on the formalism and methods? So, semi-definite programming is a very nice technique; I hope that towards the end of the semester we will have some space for this, ok. Could you explain the geometric meaning of continuity of Hessian? I think we spoke about this briefly last time. The Hessian has four functions; I mean, the Hessian has all second-order mixed partial derivatives inside it. So, whatever idea you have of continuity, you apply to each one of those guys. I cannot think of any special geometric meaning of continuity of Hessian.

The geometric interpretation of Hessian is there that it talks about whether the bowl is upward-opening, downward-opening, saddle point, or none of these. That is the geometric interpretation of Hessian. But continuity of Hessian, I do not think is anything special. So, I am happy that there are some simple questions being asked because if this is the doubt, then it is better to get it clarified. So, let me put the question to the class.

So, let $x \in \mathbb{R}^n$ and if I write a function like this, this is the optimization problem: is it unconstrained or is it constrained? This is clearly unconstrained. What would make it constrained? Right. So, I could either write it like this such that $x \in$ some set or the simpler way is just to, you know, get rid of this and you write. It conveys the same thing. So, here you can see its constraint because x is not free to live anywhere; it is forced to be in some set Ω (capital Omega), and this is a very general way of writing it. You can also have variations of this; you could have, for example, minimum like this and x such that $g(x) \in$ some set.

So, this is also a type of constrained optimization that you are now seeing; the function of the optimization variable is constrained. So, there are all sorts of flavors available. Is the quasi-Newton method faster than Newton, and if so, what is the rate of convergence? So, we will look at Newton and quasi-Newton rates; I mean algorithms in detail as we go. Newton is as fast as it gets with quadratic; quasi-Newton obviously will not be as fast, but it gets close. Which direction is considered good in the case of a non-convex function? What is actually meant by being stuck at a local minima? So, these are things that we will spend a lot of time talking about, but again, this is a basic question.

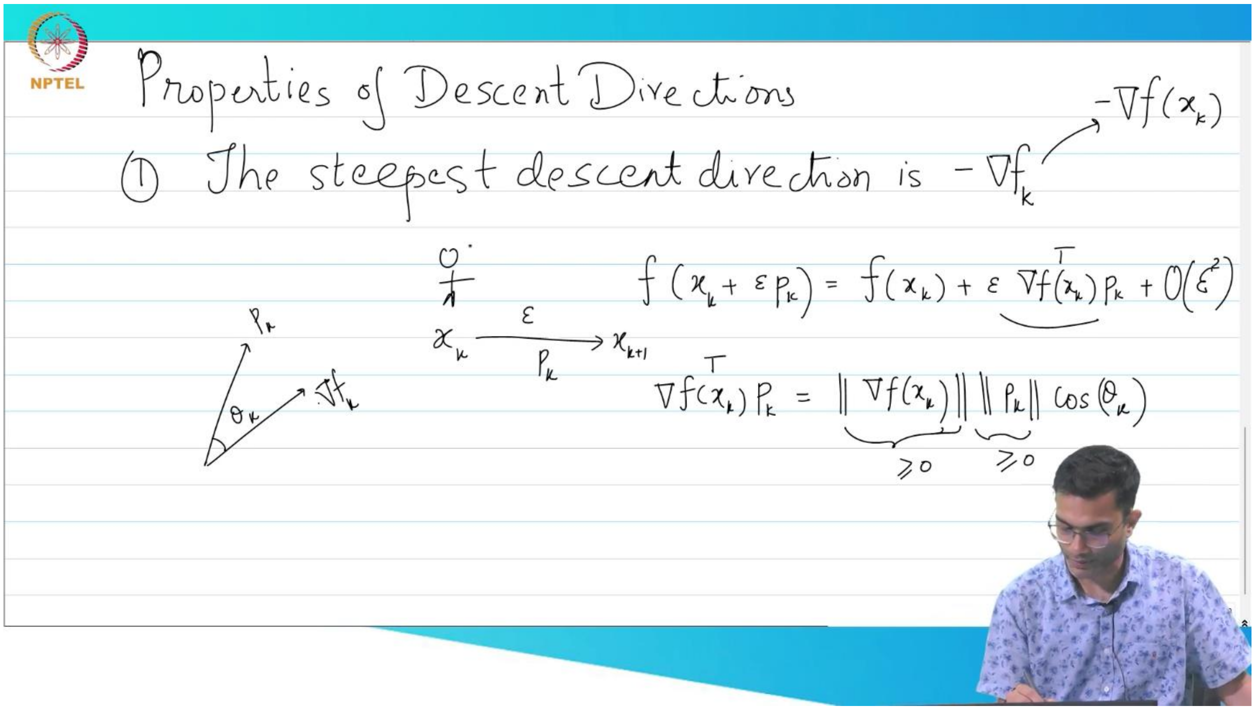


So, let me clarify it. What is meant by saying being stuck in a local minima? It just means that if I am, let us say, I start my optimization from here; I am over here, and at this point, I ask, should I go up or down if I want to get to a minima of the function? Obviously, the answer is to go down. I have to take a step to the right. So, if I start walking over here and I come over here, now any move in this direction or in this direction from this point will increase the function value. So, I will not take it right now. If this function, on the other hand, happens to be something like this, then once I have reached here, I will never get out of here. Even though the true solution is here, there is no mechanism for me, in the way that we have spoken about so far, to get there.

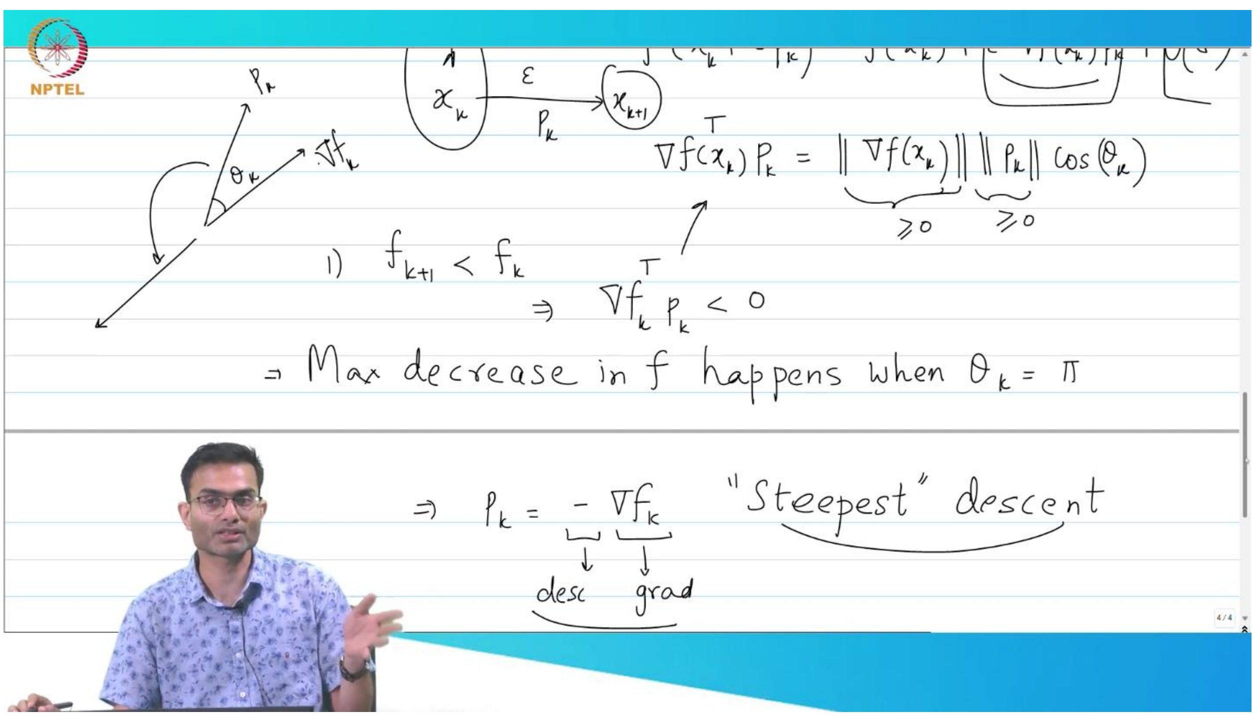
So, this, you would then say that the algorithm is now stuck in a local minima, ok. And remember, I am showing you a one-dimensional plot; imagine this in one million dimensions; you cannot plot diagrams and see where you are, ok. So, that is the intuition of being stuck in a local minima. Is there something similar to the Hessian for a function defined from $\mathbb{R}^n$ to $\mathbb{R}^m$? So, you could always define a Hessian for each of the output function values. It is a mess to keep track of it, but that is how you have to do it.

Here is a question that deals with the revision of your linear algebra. Does positive definiteness always infer the invertibility of the Hessian? So, let us take it step by step. We wrote the Newton direction; can someone remind me what the Newton direction was? Minus. Right. This was the definition of the Newton direction, which we will derive; you do not have to remember it, right?

Now, this in, and we said that the Hessian should be positive definite. We just stated it; we are here to prove it, and the question is that, remember, I made a statement that the Hessian should be positive definite, but the question is: does positive definiteness always imply invertibility of the Hessian? That is the question. Is it enough for me to say that the Hessian is positive definite



for this guy over here, right, for this guy over here to exist? Is there a yes, no, or depends? Yes. Because? So, the explanation from linear algebra is correct that if this is positive definite, it implies that all the eigenvalues λ_i are strictly greater than 0, and remember, for a positive definite matrix, how did I write the eigen decomposition of the matrix? Can I write this Hessian in terms of the product of three matrices? What was it? V (Eigen decomposition, not singular decomposition). So, let us use the standard symbol, let us say Q, next would be Λ (capital lambda).



Next would be Q^T (transpose). Are you sure? So, there is a transpose that we have to get correct over here. Let us assume it is real-valued. So, if I put the transpose over here, right.

So, this is correct. Now, what was the property of Q? What was in the columns of Q? Eigenvectors. Eigenvectors, and what is the property of eigenvectors? They are all orthogonal, right? That means, each of these, I mean. So, therefore, if all the columns are orthogonal, that means it is full rank and invertible. That means Q is invertible; obviously, Q^T is invertible.

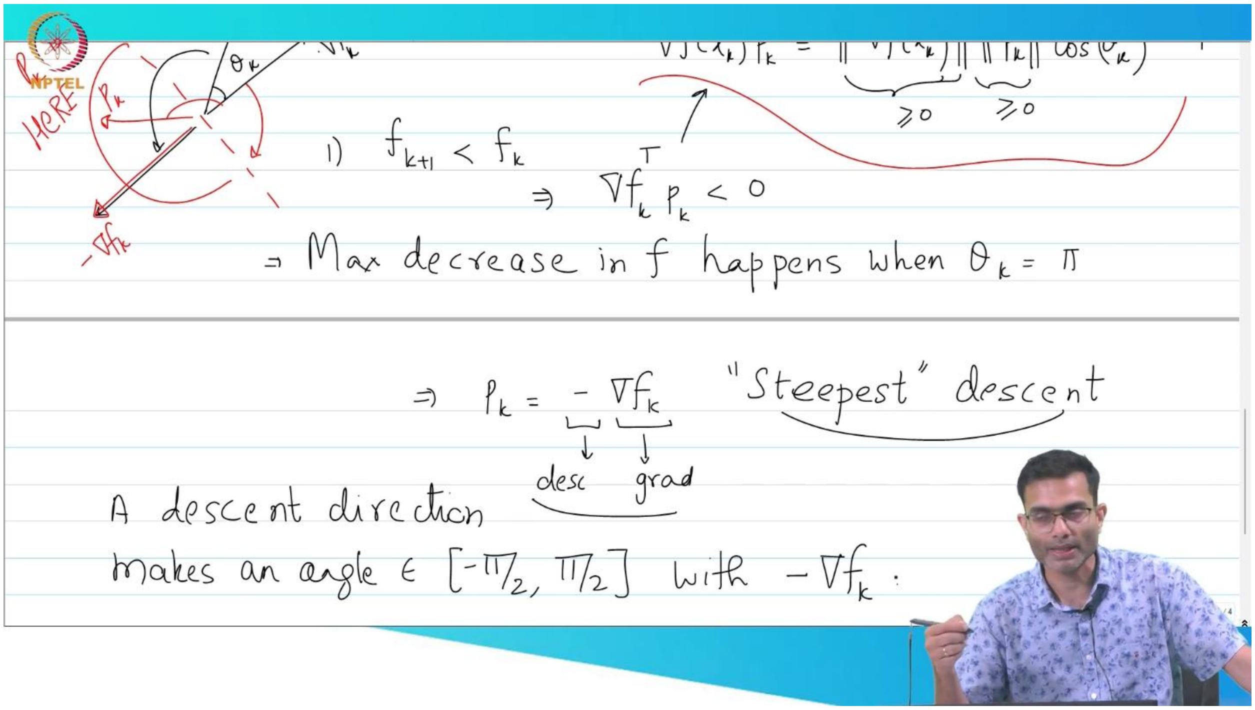
Eigenvalues are positive; therefore, this guy is a diagonal matrix with λ_i's over here. A diagonal matrix with greater than 0 entries can trivially be inverted. So, if it is enough for me to say that the Hessian is positive definite because now you look at this expression, I can take the inverse on both sides; it is guaranteed to exist, which is why we said positive definite, not positive semi-definite, because if positive semi-definite, one eigenvalue being 0 means I cannot invert it, right? So, that is the requirement; this is fine, ok.

Alright, so we have been talking about descent directions. I think we have just stated the descent directions; I mean the property of the descent directions, right? What is the first property? Did we state it? I think that was the second property. What would be the first property? It is the direction; the steepest descent direction is $-\nabla f$. I am going to put a k here.

The moment I put a k here, you know what that means, right? It means ∇f evaluated at x_k. So, that means what do I mean by this d_k? I want to evaluate the function f at some point. So, if I think about this, I want to make it a little bit clearer. I am going to write the Taylor expansion at some point x_k.

So, I will just say that x_k is the point where I am currently at. The function value f at this point is given by $f(x_k)$. That is straightforward. The first thing is that the function at the point x_k is just $f(x_k)$. I am going to add t here, where t is the step length; we will get to that later.

Now, I want to evaluate the Taylor expansion at this point $x_k + td_k$ where d_k is the direction. What is the first-order Taylor expansion? I know there is a Taylor expansion; let us just write down the first few terms; I will write it as $f(x_k + td_k)$.



I need to evaluate it as

$$f(x_k) + td_k^T \nabla f(x_k) + \frac{1}{2} t^2 d_k^T H_k d_k + o(t^2)$$

I hope everyone is clear about this notation; is it correct? So, where H_k is the Hessian. What happens if I try to minimize this expression? So, I am going to put the next line as minimize. So, I want to minimize this over the choice of t. Therefore, I can take the gradient with respect to t, right? I will set it to 0; I will ignore the higher-order terms for a moment.

What do I get? So, the gradient with respect to t, if I set this to 0, what do I have? I have

$$d_k^T \nabla f(x_k) + td_k^T H_k d_k = 0.$$

This gives you the t optimal at this step length for this particular direction. If I do the algebra correctly, I will get

$$t = -\frac{d_k^T \nabla f(x_k)}{d_k^T H_k d_k}.$$

I think there is an issue here that I need to take into account; this should be negative, but I am guaranteed that d_k and the Hessian must be such that the $d_k^T H_k d_k$ must be greater than 0; otherwise, I am not going to get the step.

This will give me a minimum value, and if I substitute this back into this $f(x_k + td_k)$ expression, I will get that is greater than or equal to $f(x_k)$ for all k that is taken. Therefore, this is a descent direction. It shows that if I move in the direction of $-\nabla f$, that gives me the steepest descent.

Now, we have already said that the steepest descent direction exists, and there are many methods of calculating it. So, how do I find it? It is a linear approximation of the function at that point. This is a good place to conclude the properties of descent directions. So, let us now summarize this section of descent directions before we take a break.