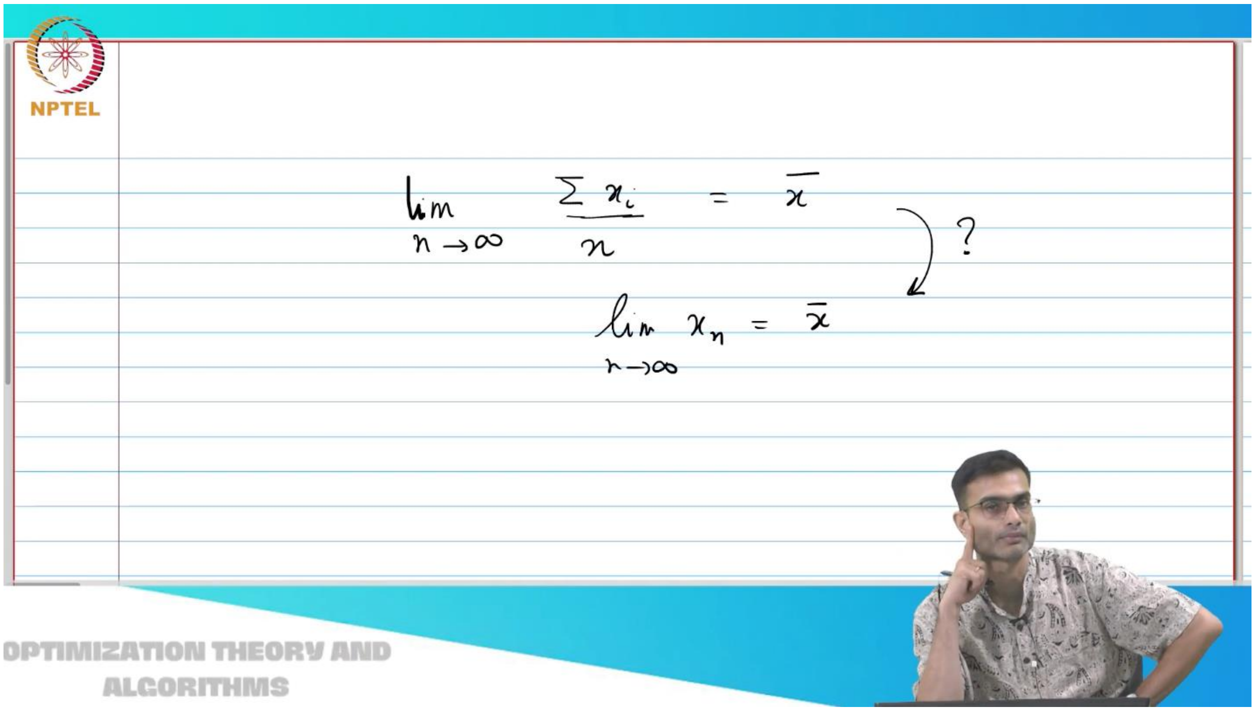
Course Name: Optimization Theory and Algorithms
Professor Name: Dr. Uday K. Khankhoje
Department Name: Electrical Engineering
Institute Name: Indian Institute of Technology Madras
Week - 11
Lecture - 74

Subgradients and Subdifferential

Ok, though we have sparse attendance now, I will start with the one and only interesting doubt sheet question. So, this question is coming from the proof that we did of convergence of PGD about the limit of an average versus the limit of the sequence. You remember that discussion? So, here is the question, there are two statements over here. Ok. If the limit of an average converges to some number, does it imply this statement? That is the question. Right? In our proof that we did, we basically looked at this something like this: the function average of the function values was converging to something, but the more interesting thing was, does $f(x_n)$ itself converge to something? That is what we wanted to know. So, do you think this follows, and do you think this follows? $2 \Longrightarrow 1$ is what you are saying, what about $1 \Longrightarrow 2$? No, because yeah, but ok.

So, let us take your argument. She is saying that at large values of n, each one of those has a small weight. So, divided by n, it goes away, but if they have converged to a certain value, then I have a large number of them getting divided by a large number, right? x_{1000}, x_{1001}, all of these guys also have the same value. So, I have a large number summing up, and I am dividing by a large number. So, that ok.



$$\lim_{n \to \infty} \frac{\sum x_i}{n} = \bar{x} \quad ?$$

$$\lim_{n \to \infty} x_n = \bar{x}$$

Right. So, exactly. So, to prove something is harder to disprove. All I need is a counterexample, right? So, if I take, uh, wait, let us look at that. This is the sequence, right? So, it is $-1,1,-1,1$.

What is the limit of the average? Why is it 0? I mean, supposing I have an odd number of terms, then the sum is $\frac{1}{n}$ tends to 0. That is why this converges to 0, right? But the sequence itself does not converge.

So, this is a counterexample which shows that 1 does not imply 2, but 2 does imply 1. If I have 2, how would I show this? That is the intuitive way. What we could do is simply this:

$$\frac{1}{n}\sum_{i=1}^{n} x_i$$

This is the term that I have to see, right? What is the limit of this? Is this equal to 0? That is what we are asking, right? So, if I subtract, I can just write this as

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_1 - \bar{x} + x_2 - \bar{x} + \cdots + x_n - \bar{x}).$$

Now, if n is very very large, what can I say about each of these terms? So, we are saying that you are given that x_n is given this, right?

So, for larger terms, what will happen over here? This is almost ϵ, you can say. What about these terms? They need not be 0. So, they are finite, right? But finite divided by large n is 0. So, this is actually equal to 0. Right. So, $2 \Longrightarrow 1$, 1 does not imply 2. Ok.

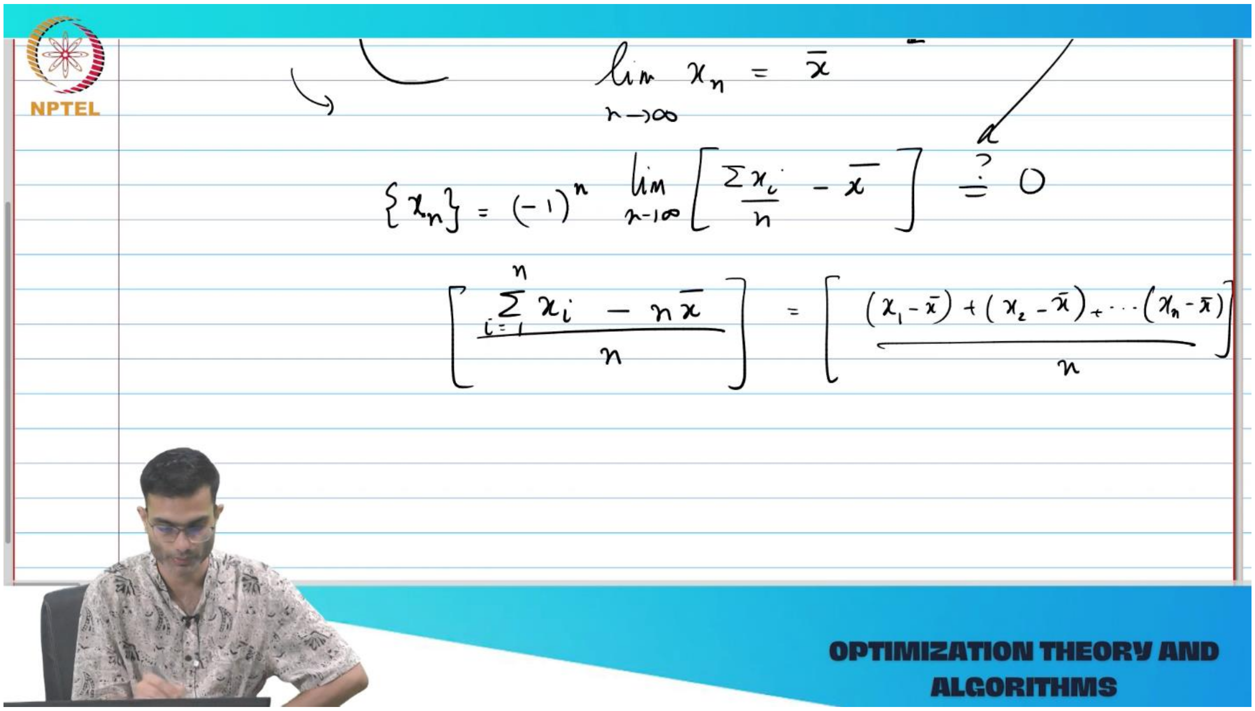
So, this was one doubt sheet question. Any questions on this? Quite intuitive, right? So, the proof that we had for convergence was a weaker kind of a proof, but as I said, there are stronger proofs that can be used. So, now let us proceed with projection operations. So, what example did we take so far? We took the projection on the L_2 ball, right?

How did I write that? So, the projection operation was: Find the x that belongs to Ω which is closest to x_0. That would be x. Ok. So, we have already done this when we did this, and that is fine. This is the easy part. Let us say that there is some problem where instead of this 2-norm, it becomes the 1-norm. So, supposing let us say σ is such that x. Ok. Now, this, on one end of it, you can interpret this as just some mathematical exercise.

So, instead of the 2-norm, I took the 1-norm, right? But they can be several applications or you know, scenarios where this is a very realistic problem. For example, you have a city block kind of a thing. Your pizza source is over here, the delivery is somewhere over here, and the pizza guy obviously cannot just go like this, right? The pizza guy has to go like this, like this, in this kind of a way, right? So, what is the cost incurred? It is going to be proportional to how much the delivery person has to walk, and how much the delivery person has to walk is going to get captured by the 1-norm of x, not the 2-norm of x, because that is the crow flies. According to the 2-norm, the delivery guy has to follow the block, right?

So, I may have a constraint that if assuming the delivery guy has, let us say, a motorcycle or walks at a certain speed, and I want to ensure delivery within, like, Domino’s, right? 30 minutes has to be done.

So, I may want to put a constraint that the 1-norm of x has to be less than or equal to 30 minutes in whatever units you choose, right, and solve this problem for me. The problem could be to minimize the cost, minimize how cold the pizza gets, whatever, right? So, these kinds of constraints are fairly regular, and they appear on the face of it. Is there any problem with this constraint? Like, how would I define this constraint in my usual language? So, $C(x)$, what will I write $C(x)$ as in the normal notation that we have been following in this course? So, is it equality or inequality? Inequality, inequality constraints are written as $C(x) \geq 0$, right?



$$\lim_{n\to\infty} x_n = \bar{x}$$

$$\{x_n\} = (-1)^n \quad \lim_{n\to\infty}\left[\frac{\Sigma x_i}{n} - \bar{x}\right] \overset{?}{=} 0$$

$$\left[\frac{\sum_{i=1}^{n} x_i - n\bar{x}}{n}\right] = \left[\frac{(x_1-\bar{x})+(x_2-\bar{x})+\cdots(x_n-\bar{x})}{n}\right]$$

So, what is $C(x)$? Right? It is $1-\| x \|_1 \geq 0$. Ok. Now, this looks like a fine enough constraint. All that I have to do is to solve. What would I do? In constrained optimization, I have to, the first step would be to form what? I have a constraint, I have some objective function, I am giving you the objective functions of the feasible set is given to me from $C(x)$, that is formed. What is the next step? Form the Lagrangian. So, I would write the Lagrangian as this. Ok. Let us say all the theory, all the if conditions of the KKT theorem hold true, very good, I go to the next step.

What is the first in statement of the then part? ∇ Lagrangian. Right, supposing I am being nice to you, I tell you that f is differentiable. So, no problem. What about $C\nabla C$? Do I have a problem? It is not differentiable whenever any of the components. So, what is what is x_1 norm of x? Right. Now, $\| x \|_1$ we just take the scalar function $\| x \|_1$ which looks like this. It is differentiable everywhere except 0.

So, I have a problem in taking gradients over here. So, what do I do? So, here is where we have to upgrade our definition of gradients, okay, because we will come across so many problems where I need to differentiate a function that is not differentiable. Okay, so let us go to the next page. So, we are going to learn a new concept of a derivative, which generalizes the idea of a derivative. So, it is a pretty fundamental concept, okay.

So, I will start, and we will restrict ourselves to convex functions, okay. So, let us start with something very basic. So, for take a very simple case, if I take give you a differentiable function for which I can draw a very convenient graph like this. Let us say this is my $f(x)$, okay. I am at some point over here, and this is my x-axis, okay. At this point x, let us say this is what the gradient looks like.

So, and let us say now I am at some point y over here, okay. So, there are a few important points over here, there are these three points marked in the black circle, okay. So, if I form the linear approximation of f at x, we did that in the last class. What would it be? So, this is the first order Taylor approximation of f at x. What all would I write in it? The first term would be $f(x)$, that is the first term of Taylor's theorem. What is the second term? You are going to add $f(x)$ at what point, x next transpose, next $y - x$. This is the first order Taylor theorem for this function, and I have taken the starting point as x over here, okay.

Now, given that this is a convex differentiable function means ∇f is defined, I am in my usual world where gradients are defined, no problem. If it is a convex function, what is the relation between these two quantities? $\geq$ right. Okay. So, we already know this, we did this in the last class. This is almost like the definition of many people will consider write this actually as a definition of convexity, okay. Now, we are going to give this another interpretation.

Let us focus on the right-hand side, okay. Now, this has been drawn on a 2D piece of screen. So, it looks like a line. Now, imagine x and y they are all living in n dimensions, okay. So, in n dimensions, this object on the right-hand side, right, it is a function x is fixed, y is the guy that is varying.

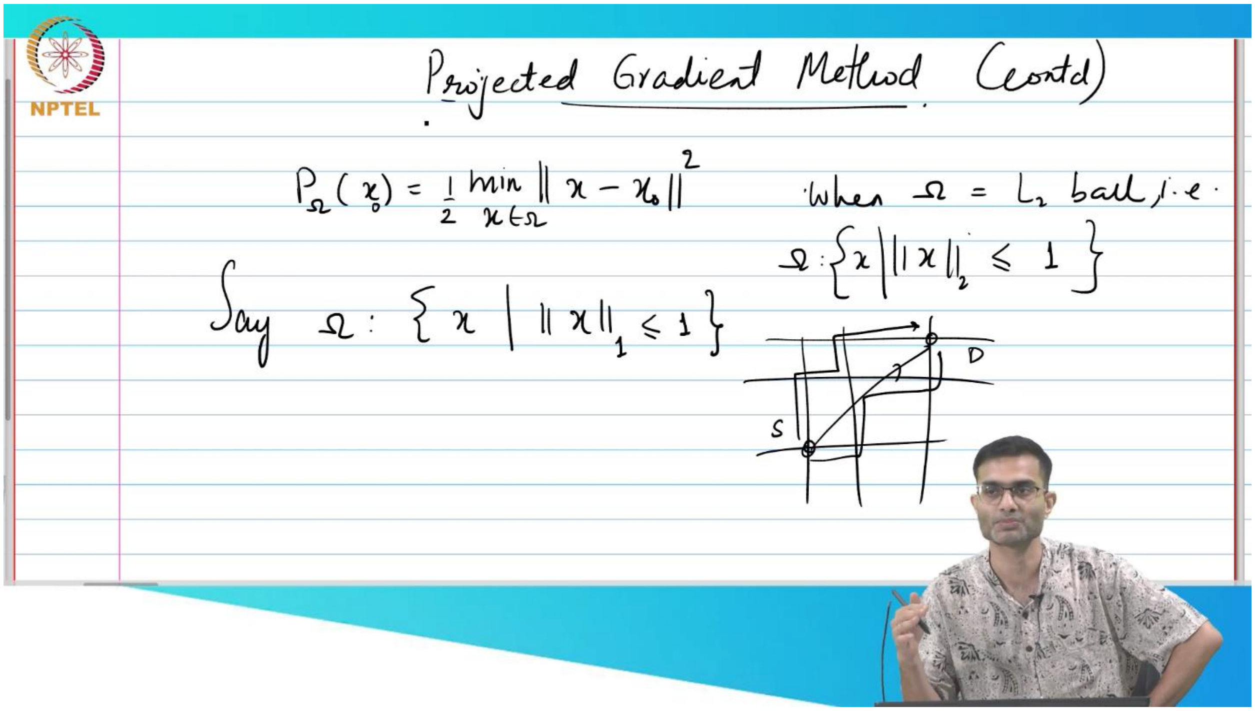
What kind of a function is this? Is it, what is the exponential, trigonometric, what is it? I want something more precise than linear. If you have it, it is an affine function because there is an intercept which does not need to go through the origin. That is the difference between linear and affine. Linear always goes through the origin; affine need not, can have an offset. So, this is an affine function. Another word for an affine function like this is a people will call this a hyperplane, right. Why hyper? Because the dimensions may be more than 3.

So, instead of just calling it a plane, which we are used to in 3D Euclidean space, you call this a hyperplane, okay. So, this, because of the way it kind of supports the function f, this is also called a supporting hyperplane. Okay. So, you can almost think of a gradient for a convex function as something that supports the function that is opening up above it, okay. So, hence the words supporting hyperplane, and this gives us a very convenient way to try to come up with a little bit more general definition of the gradient, okay.

So, let us see what that means. So, now I am going to define a subgradient, and I am going to stay on convex functions. So, I have kind of just replaced, does not look like I have done much, I have just removed this ∇f and put g inside it, right. So, what is it saying? A subgradient g of a convex function is such that this property of the supporting hyperplane is still maintained. And we know that for a convex function which is differentiable, if this g, if I replace it by ∇f, this ∇f will always support the function like a supporting hyperplane everywhere. Where does it get interesting? Is if the function is not differentiable.

So, let us have a look at that, okay. So, imagine a function that looks like this. So, it is like this, this looks like a convex function. It has a kink, right. So, this is my function $f(x)$, right. So, at this point over here, do you think it looks differentiable at this point? No, right, because the derivative from the left and the derivative from the right are not matching up, right.

So, if I draw the tangents over here, if I am coming from the left, this is the tangent, right. If I am coming from the right, this is the tangent. So, this is like a textbook example of a function that is not differentiable, okay. Now, that said, each of these guys is like an affine function, these tangents which I have drawn, right. So, if they are affine functions, what is different between these two affine functions? So, at this point x, I can write all sorts of different affine functions that pass through x. What will their form be?



Projected Gradient Method (contd)

$P_\Omega(x_0) = \frac{1}{2}\min_{x \in \Omega} \| x - x_0 \|^2$

When $\Omega = L_2$ ball, i.e.

$\Omega : \{ x \mid \|x\|_2 \leq 1 \}$

Say $\Omega : \{ x \mid \|x\|_1 \leq 1 \}$

So, this thing on the right-hand side, you all agree is an affine function or a supporting hyperplane. Now, I tell you that at this point x, I want to write a general form of an affine function which passes through x and has some random slope. How would I write that expression? What would change in this expression? Union? No, give me just one way of describing it before you come to union, give me one single change.

Okay, let me make it even simpler: in this expression, what should I change so that I can go from the blue slope to the blue line to the pink line? g is the only thing I need to change, right. So, in fact, I can write this as

$$f(x) + g_1^\top (y - x)$$

that is the blue line and the pink line I can write as

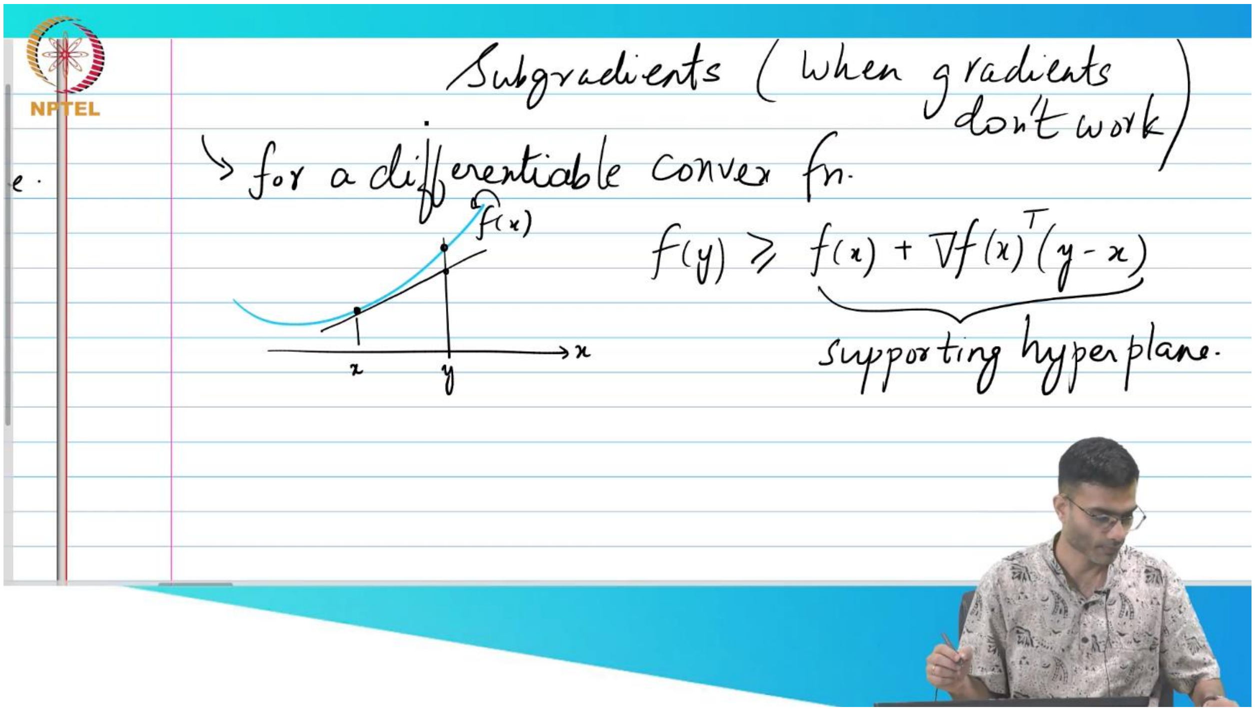
$$f(x) + g_2^\top (y - x)$$

right. And both of these guys have the property: can you say that both of these lines are they supporting hyperplanes for this function at x? Yes, right. So, they are both okay. There is another way of identifying a supporting hyperplane: the plane never cuts the function. You can see that neither the blue line nor the red line ever cuts the function, all right.

So, you can say alternatively, okay. So, in technical language, you will say that g_1 and g_2 are the subgradients of f at x. That is how we would call g_1, g_2. So, this is of course something new. We are used to, you know, if you give me a function $f(x)$, I calculate the derivative, $\frac{d}{dx}$, now after that I do ∇f. We are used to it, it is very hardwired on us into thinking that ∇f is a unique function. But now, we have generalized the idea of a gradient to a subgradient, and I am telling you that both g_1 and g_2 are subgradients of f at x, right. So, that is how this idea is generalizing: that we are replacing something that is unique by something that is non-unique. However, it seems to satisfy the same or have the same properties as the gradient did. In particular, the property was that of being a supporting hyperplane. That is how I am generalizing it in this way. So, let us see what that bias is, but is this clear? Properties of the supporting hyperplane, there is no argument that both g_1 and g_2 seem to be supporting it, right. Okay.

Let us go to the next page. So, as I just said, since there are so many different possibilities of g_1, g_2, in fact, let us just go back there. Do you think g_1 and g_2 are unique, or can there be some other supporting hyperplane as well? They can be, right. I can, for example, if I take a line like this, if in other words, if the slope is between g_1 and g_2, it is also a supporting hyperplane, right.

So, it is also a subgradient, right. So, it is definitely not unique. It is not that it has two values, it has an infinity of values, right. So, when I have a whole bunch of things which are possible, you remember we had this situation happen when we were defining feasible sequences. Feasible sequences gave us tangents, and then I collected all of those tangents, and what would I call it? The tangent cone. So, the same concept is happening over here. I have collected, I have a lot of subgradients, I collect them all together, and what do I call it? I call it a subdifferential, okay.



Subgradients (when gradients don't work)

↳ for a differentiable convex fn.

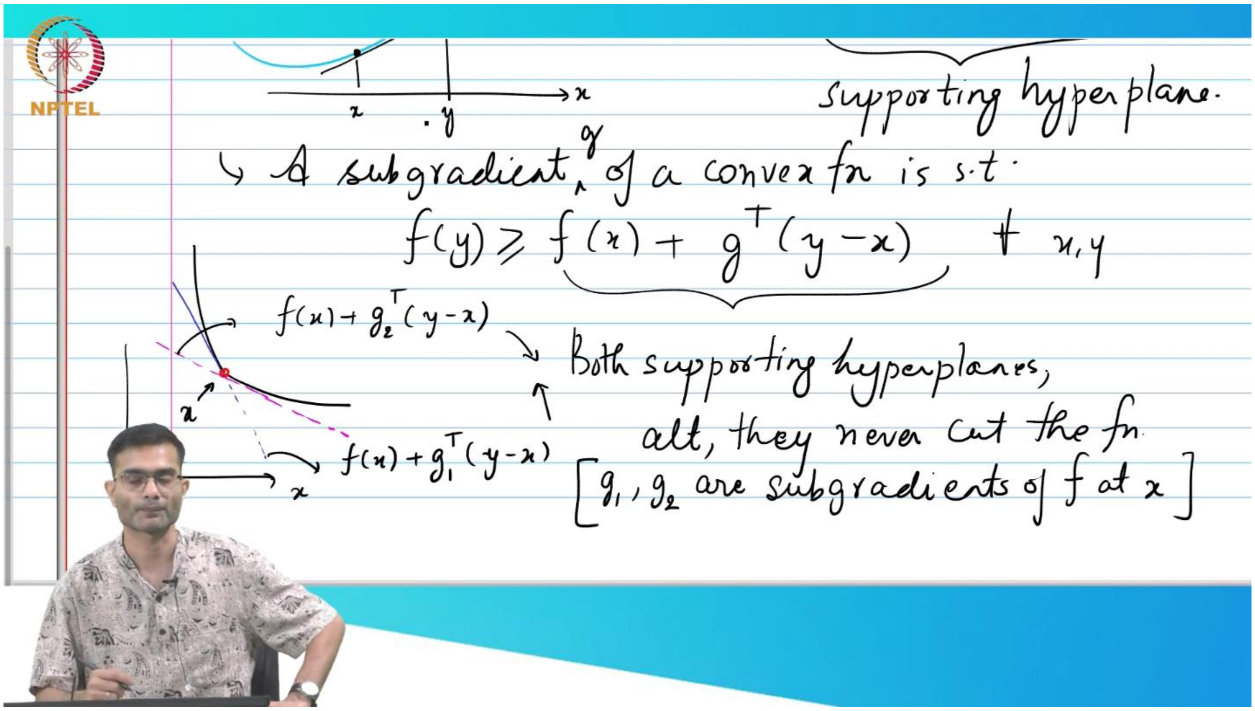
$$f(y) \geqslant \underbrace{f(x) + \nabla f(x)^T (y-x)}_{\text{supporting hyperplane.}}$$

So, that is the next concept. The symbol for it is this: ∂f (instead of ∇f now, I have this small ∂f), and this is defined as g and the same definition that I had for the subgradient is just over here. So, any g which satisfies the supporting hyperplane property works okay. So, this is the subdifferential, okay. It happens to be a closed convex set, right.

So, let us take an example for the question. Yeah, it is defined at a point x. Ok. So, let me just write it like this.

In the previous... You mean this line over here? This is in general true if you take any I mean this is just the property of convexity, that's all right. So, you pick any point x and pick any point y, this will always hold true. So, this is the gradient correspondingly at that point x at that point x.

So, let us take some simple example to get familiar with this concept.



Supporting hyperplane.

↳ A subgradient $\overset{g}{\wedge}$ of a convex fn is s.t.

$$f(y) \geq \underbrace{f(x) + g^T(y-x)} \quad \forall\ x, y$$

$f(x) + g_2^T(y-x)$

Both supporting hyperplanes, alt, they never cut the fn.

$f(x) + g_1^T(y-x)$

[g_1, g_2 are subgradients of f at x]

It does not depend on y, you can take y to be. So, yeah, you can, the definition of convexity is that no matter where you go, the function is always above this, that does not depend on how far you go, right? It does not depend on that, okay. So, I am going to take an example where... (draw this properly) this is, let us say $f_1(x)$. Ok, and I have a second function that looks like this. I am going to call this $f_2(x)$, okay? And I am going to define my function f in a slightly irritating way. I am going to say that

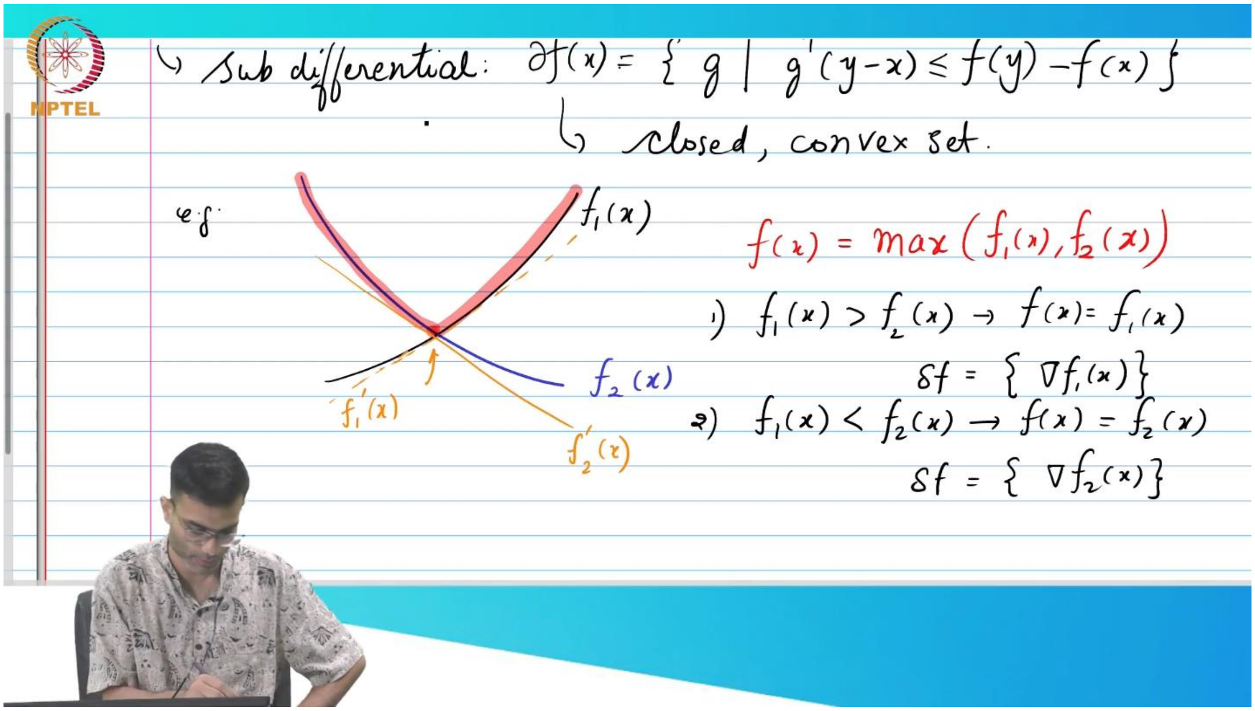
$$f(x) = \max\big(f_1(x), f_2(x)\big)$$

So, if this is my function definition, how is it going to look? If I start from the left, I am going along the blue line, I go like this, and then I hit this point, what happens? I jump up, right. So, this becomes my function $f(x)$, okay? So, even though f_1 and f_2 are nice and smooth and differentiable, etc., the resulting function f wherever you have a max or min operation, these kinds of things can happen, and there is clearly a point where there is a kink.

Right. So, now you have the situation where you can have this, the gradient at this point is right. So, this is $f_1{}'(x)$, and then I can also have this, which is my $f_2{}'(x)$, right? So, far, we have not yet come to the subgradient. So now, let us look at the cases when I need to. So, you should also know when to worry about gradients and subgradients, right? Like in the case of $\|x\|_1$ or, sorry, the 1-norm of x, or the function $f(x) = |x|$, I needed to worry about differentiability only at the origin, right?

So, if I am far away from the origin, there is no problem, the derivative is well-defined, right? So, the three cases that I will need to worry about are: I am on the left of the kink, the right of the kink, or at the kink, right? Those are the three situations, right. So, if $f_1(x)$ is greater than $f_2(x)$, then what is my f? $f_1(x)$, right? And what is the subgradient then? There is no controversy here; the subgradient is simply the gradient. So, if the function is differentiable, the subgradient becomes the gradient, right? So, it becomes basically $\nabla f_1(x)$ without any controversy.

The second situation: well, let us take the reverse situation, $f_1(x)$ is less than $f_2(x)$. So, $f(x)$ simply becomes $f_2(x)$, and the subdifferential is simply $\nabla f_2(x)$.



↳ Sub differential: $\partial f(x) = \{ g \mid g'(y-x) \leq f(y) - f(x) \}$

↳ closed, convex set.

e.g.

$f(x) = \max(f_1(x), f_2(x))$

1) $f_1(x) > f_2(x) \rightarrow f(x) = f_1(x)$

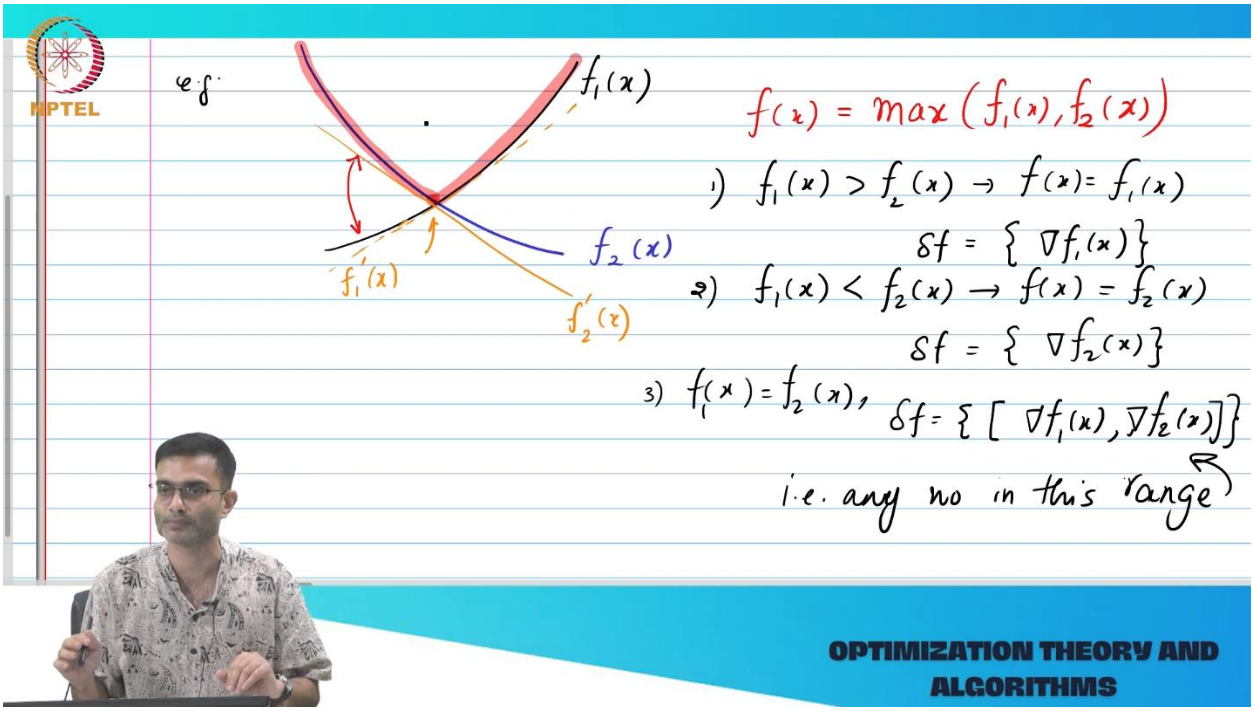
$\delta f = \{ \nabla f_1(x) \}$

2) $f_1(x) < f_2(x) \rightarrow f(x) = f_2(x)$

$\delta f = \{ \nabla f_2(x) \}$

Now, let us take the case where $f(x)$ is equal to $f_1(x)$ and $f_2(x)$ at the same point. Now, at this point, we have the issue that the function is not differentiable, okay, but now we know that we have to take the idea of the supporting hyperplane. So, I have $f_1{}'(x)$ at this point, I have $f_2{}'(x)$ at this point, and of course, any of these guys will also work. Right, so look back at the definition of the subdifferential. It is all those g's which satisfy the supporting hyperplane condition, right. So, in this case, the subdifferential becomes what? Correct, so I can take any convex combination of ∇f_1 and ∇f_2, right. Another way of writing it is simply as the interval between ∇f_1 and ∇f_2, right? In this case, I am saying it is a number because I am drawing a slope on paper.

So, the slope is just going to be a scalar number, right. If it were, but you can see this generalizes to a vector also, right. So, it is any number between this and this that is permissible because it corresponds to a slope of a line with this red double arrowed structure. Any line with a slope between ∇f_1 and ∇f_2 is captured by this, that is the meaning of this, right. Just to put it into simple words, any number in this range is how you would say it in English. Of course, what I am making the very simple assumption that numerically ∇f_1 is less than ∇f_2, right.



e.g.

$f(x) = \max(f_1(x), f_2(x))$

1) $f_1(x) > f_2(x) \rightarrow f(x) = f_1(x)$

$\delta f = \{\nabla f_1(x)\}$

2) $f_1(x) < f_2(x) \rightarrow f(x) = f_2(x)$

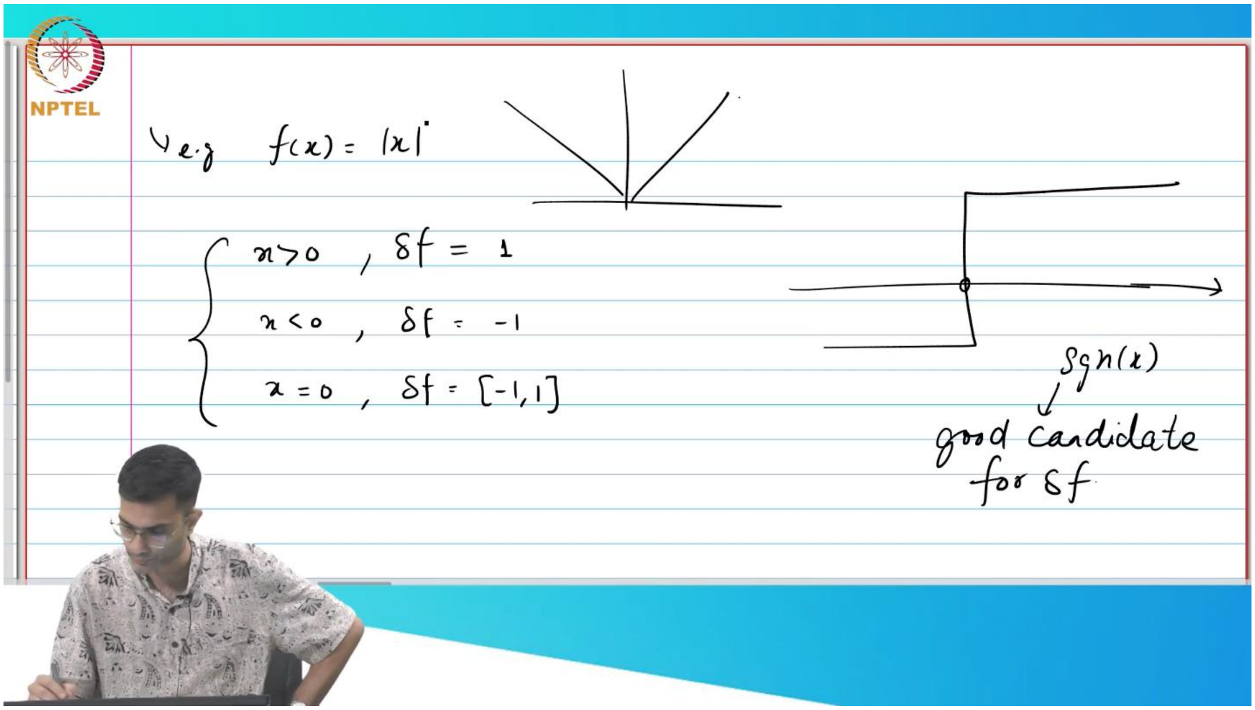
$\delta f = \{\nabla f_2(x)\}$

3) $f_1(x) = f_2(x)$, $\delta f = \{[\nabla f_1(x), \nabla f_2(x)]\}$

i.e. any no in this range

So, the set $[\nabla f_1, \nabla f_2]$ makes sense; $[\nabla f_2, \nabla f_1]$ does not make sense. I mean, the closed interval. So, the beauty of it is that in numerical applications, because the subdifferential has a full range, you can actually choose the range according to what you want. It is not fixed; it is open, it is up for you to do it, okay. So, let us... is this fine? So, let us go to that example which we were sort of motivating our study by, which was $|x|$. Right, $f(x) = |x|$, and we all know how that looks, right. So, for $x > 0$, what is ∇f? I mean $\partial f = 1$, for $x < 0$, what is $\partial f = -1$, for $x = 0$, what is $\partial f = [-1,1]$, any number in this range, right.

So, this is what I do, right, and this function should look somewhat familiar to you from signals and systems. So, I plot it: what does it look like? Right, it looks a little bit like the signum function of x. However, there is... well, okay, the signum function, how is it defined? -1, 1, and at 0 you define it to be 0, right? Something like this. So, is the signum function a subdifferential? Yes, because I can pick any value. 0 is any value, right, but there can be... if I... it is up to me, I could have chosen $\partial f = -0.1765$.



e.g. $f(x) = |x|$

$$\begin{cases} x>0, & \delta f = 1 \\ x<0, & \delta f = -1 \\ x=0, & \delta f = [-1,1] \end{cases}$$

sgn(x)

good candidate for δf

So, it is not the signum function, but it is a subdifferential. So, this is a good candidate for a subdifferential.