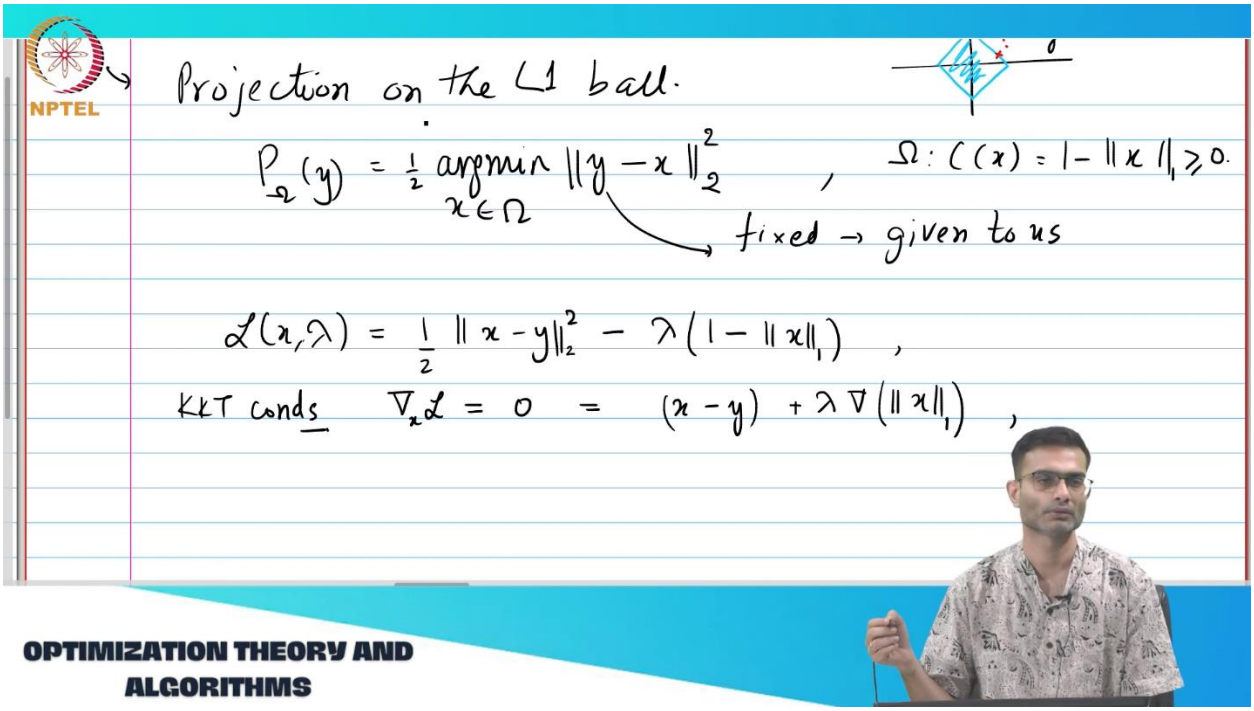
Course Name: Optimization Theory and Algorithms
Professor Name: Dr. Uday K. Khankhoje
Department Name: Electrical Engineering
Institute Name: Indian Institute of Technology Madras
Week - 11
Lecture - 75

Projection onto l1 ball

Right, okay. So, let us now proceed on to sort of warmed up with the whole concept of subgradients. Let us actually now apply it to calculating the projection operator on the L_1 ball. So, you notice that when right in the beginning here this is where I ran into trouble when I went to take ∇ of the Lagrangian I ran into ∇C, that is where the problem was and so now I have built up one tool which is the subdifferential which is going to help me to actually work on this ∇C. This is the whole reason why I did this business, okay. So, now let us go back to applying it. So, maybe I will just start from a new page. Okay.

Okay. So, in this case, why is the point that is fixed? Or I can say it's given to us. Okay. And that is my constraint $C(x)$, right? I am just rewriting what we have.



The Lagrangian of this problem, what is the Lagrangian of this problem? I want to minimize this function $\| y - x \|^2$, right? The distance of my given point from any point in the feasible set, right, is the geometry is clear. So, if I want to plot it over here, this is my feasible set, right? This is my feasible set and now I am giving you a point, let us say here supposing this is my point y. I am asking you from this blue box, tell me the point which has the smallest distance and here the smallest distance is the Euclidean distance (2-norm) because the definition of the projection operator does not change, that remains 2, right? So, of all of these guys, tell me who is, you

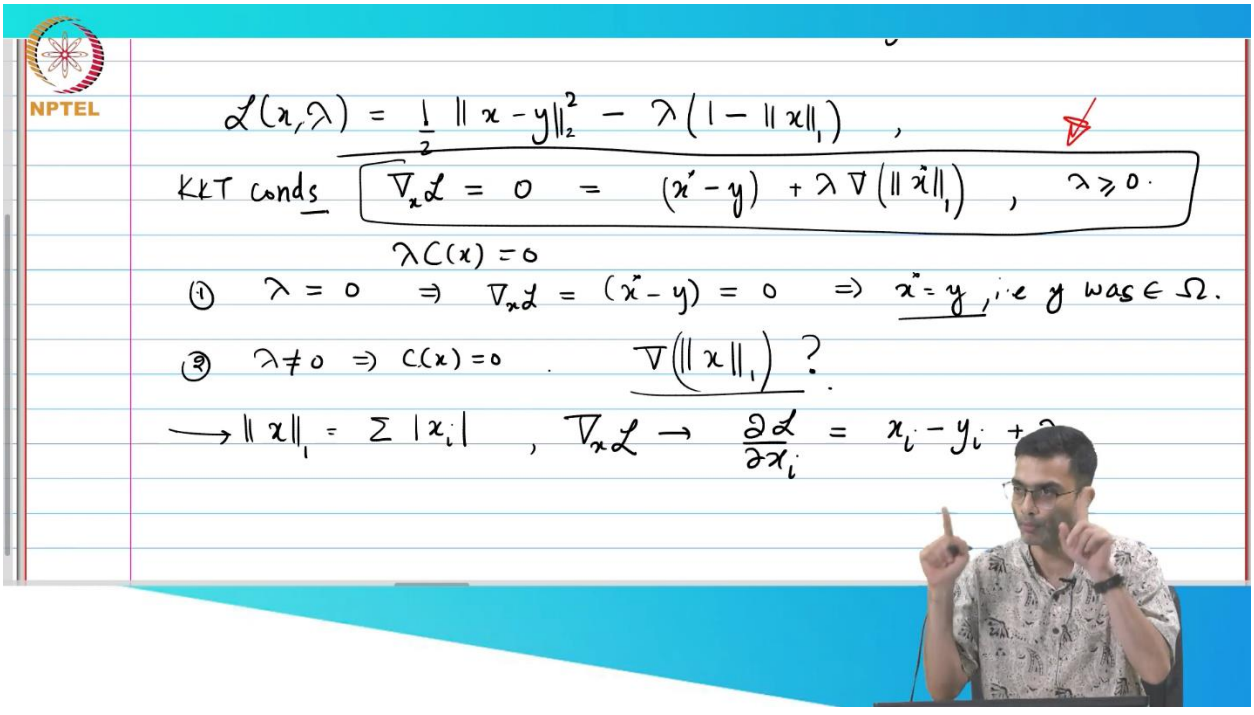
know, for example, it might be some point over here. So, tell me what this point is, that is what I want.

Therefore, the Lagrangian, what is the objective function? No, why mod? Mod, we are not working with scalars anymore. L2 norm. L2 norm, right? Half because there is a $\frac{1}{2}$ here: $\frac{1}{2} \| x - y \|^2 - \lambda \| x \|_1$. This is my Lagrangian, okay. So, now to apply KKT, I need to look at, that is how I wrote it: $f(x) - \lambda_i C_i$. We are relaxing that now. The condition was that f and C should be continuously differentiable, if you remember in the KKT condition.

Now, we are going to relax that a little bit and see can it still work, right? Obviously, if it were differentiable then this problem would not be applicable over here, okay. So, first term derivative gradient rather with respect to x. Gradient of $\frac{1}{2} \| x - y \|^2$ with respect to x is? Go on guys. You can just open it up for example, right? $x_1 - y_1^2$ till x_2 and take the derivative and you will get this, sorry not the norm, whatever I think you will just get this.

Right, it is a vector, okay. Then I have my issue over here which is. So, this becomes a plus λ and I have, right. Is there any condition on the Lagrange multiplier as per KKT? $\lambda \geq 0$? So, we also have the complementarity conditions which will say what? $C(x) \cdot \lambda = 0$, right?

So, if let's take the simple case if $\lambda = 0$, that is a possibility. If $\lambda = 0$, my gradient of the Lagrangian becomes very simple: that λ term goes away. What do I have left? $x - y = 0$ implies what? $x = y$. So, this actually I should write this as x^*, right? The best x, the solution to the problem: $x^* = y$. What does this correspond to? That means, so very simply it means that $C(x) \geq 0$ or $C(x) = 0$, which means where is the point? Where is the point y? It is already inside the blue box. So, if I am asking you to project that point, I will get back that same point. So, that is this point, right.

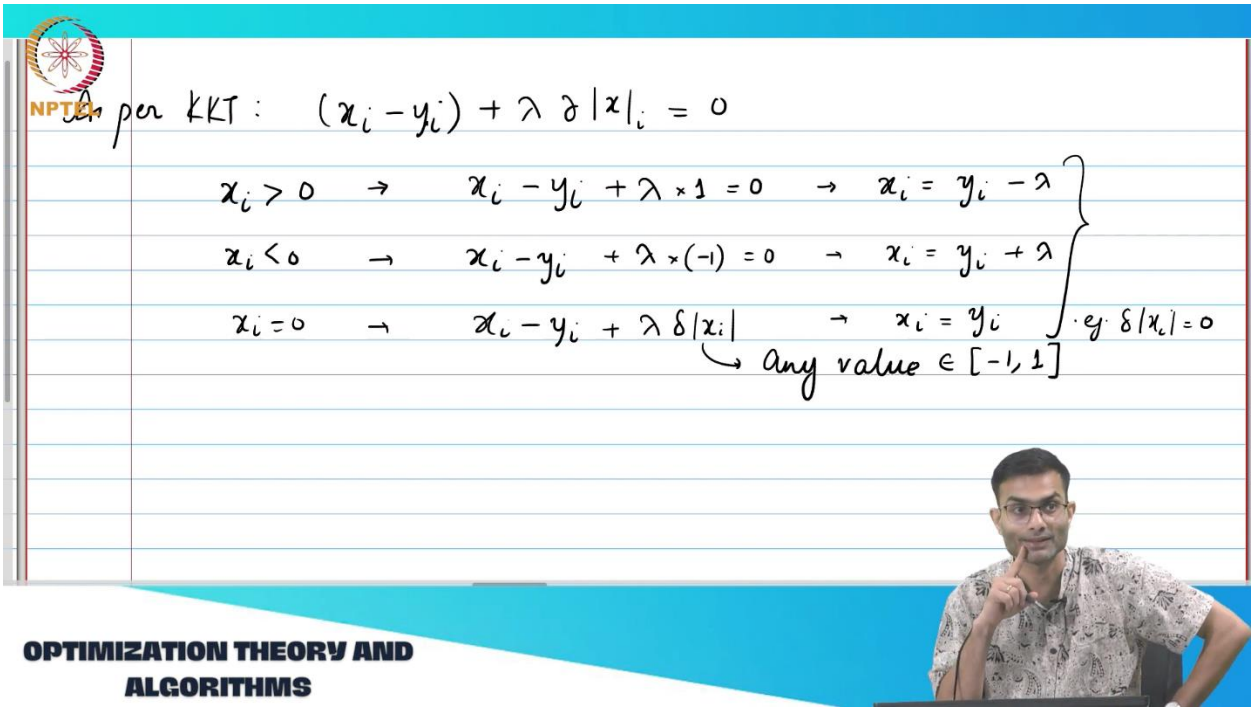


i.e., y was already in the feasible set. $\lambda = 0$ means I am using this over here. Substituted blindly into this, right? The second term vanished, and I said $\nabla L = 0$. I am simply left with $x^* - y = 0$. Now, I do not even need the subgradient over here. Now, the second case is $\lambda \neq 0$, therefore, what does it imply? Complementarity says that $C(x) = 0$, okay. So, I have now.

So, I need to worry about this guy, this is the term I need to worry about because now that term is not 0 in the Lagrangian, right? I need to worry about this term. So, now I can open up this guy, right? I know this from the definition, okay. Now, when I start taking the derivative of this with respect to, when I take for example, ∇ of the Lagrangian, let us go back to first principles. What is gradient? When I take gradient, I am taking partial derivatives with respect to x_1, with respect to x_2, with respect to each one of those guys. So, let me take just the i-th component of this. So, I can write this as this will be the i-th component of ∇ Lagrangian, right? What all will it have?

So, I am asking you to write the i-th component of this expression over here.

The first term, what will it be? 1 or 0 multiplied by $(x_i - y_i)$ obviously you know why it will be 1 or 0, plus λ. Okay, remember I am going to expand this $\| x \|_1$ as $\sum_i |x_i|$. Now, our x_1, x_2 and I want, I am taking the derivative with respect to x_i. Which term will survive? Only $|x_i|$ will survive, right? So, this I am going to write as whatever it is, it is going to be this, right? So, it is a nice thing because all the other x_j for $j \neq i$ do not appear in this feature because I am taking the derivative with respect to x_i. So, x_j for $j \neq i$ do not appear in this expression, and what will I do? I am going to set this equal to 0. That is the meaning.



So, what have I done? I have applied KKT, I have applied KKT with only one tiny difference: instead of writing it as gradient, I am writing it now as subdifferential. That is the only difference that is happening. Go to the next page, okay. I will just rewrite this again on the top:

$$x_i - y_i + \lambda\, \partial x_i = 0$$

So, as per KKT, okay. So, we can split this now, right, because $|x_i|$ has trouble only at $x_i = 0$.

So, let us start with the easier cases. Supposing $x_i > 0$, then what happens? I am going to write this expression:

$$x_i - y_i + \lambda \cdot 1 = 0 \quad \text{when} \quad x_i > 0$$

The gradient of $|x_i|$ is 1, so it is equal to 0. For $x_i < 0$, this will still be:

$$x_i - y_i + \lambda \cdot (-1) = 0$$

And for $x_i = 0$, the same thing:

$$x_i - y_i + \lambda \quad \text{and this is going to still be the subdifferential}$$

What values can it take? Any value, right? Any value is in my hand, okay. So, I can just, it looks a little messy, I can rewrite this in a nicer way.

So,

$$x_i = y_i - \lambda$$

in this case, and

$$x_i = y_i + \lambda$$

in this case, and x_i will be equal to zero, correct me, this is an example, right? So, I can take $\Delta x_i = 0$, 0 is a candidate function, right? So, is there again a compact way to write these three things? The signum of something, right?

So, $x_i = y_i - \lambda \operatorname{sign}(x_i)$ works, right? When $x_i > 0$, I have $+1$; when $x_i < 0$, I have -1; and when $x_i = 0$, I have 0. This is a candidate for a, right? Candidate because I have taken a particular value of 0. It is perfectly valid, and I can just...

So, this is what I get. Let us quickly recall in this problem what is given to me, what do I have in hand, what do I know?

There are three terms in this expression: x, y, λ. What is λ known to me? Other than the fact that it is, in this case, strictly greater than 0, I do not know what it is. Do I know what x_i is? No. Do I know what y_i is? So, I only know y_i. So, we only know y_i and we have to somehow determine both x_i and, you can see from this expression, I need to know λ_i or λ. Also, there is no λ, I also need to know λ, okay.

So, now let us try to make a little bit more progress. Since we only know y_i, what I will do is sketch a number line with y_i on it, okay. I mark this as λ and $-\lambda$ over here, okay.

So, this thing can be rewritten as:

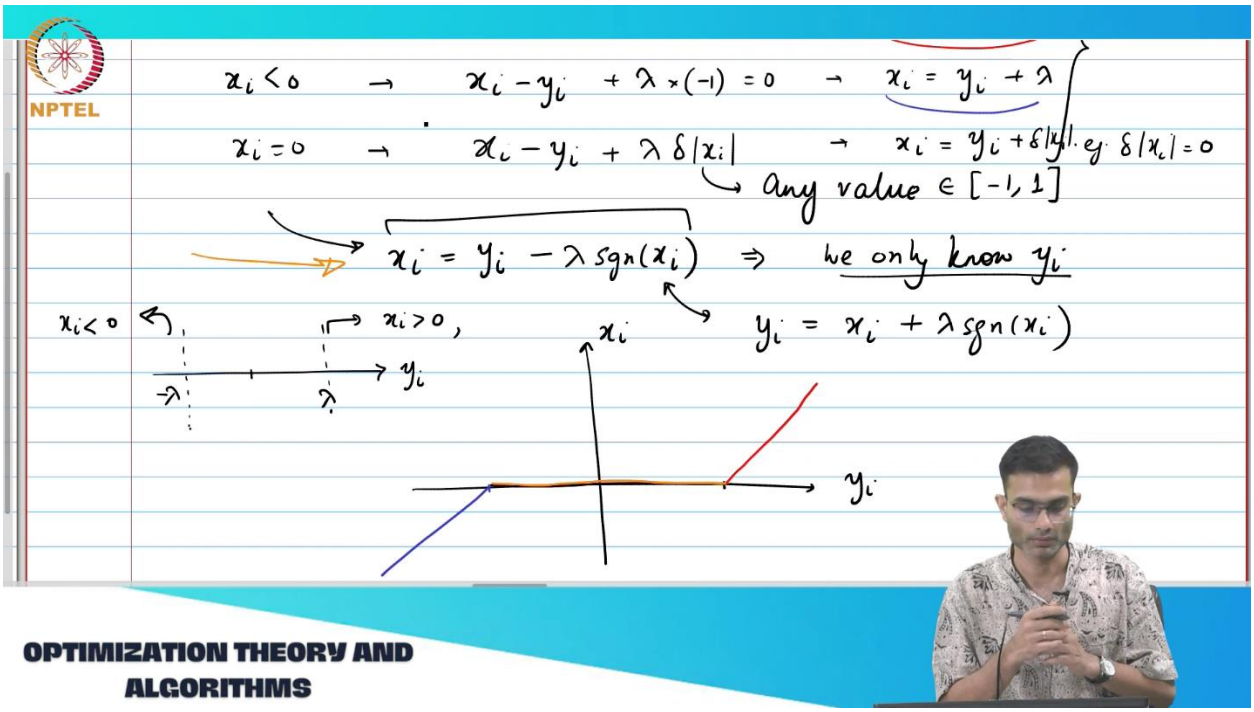
$$y_i = x_i + \lambda$$

So, if $y_i > \lambda$, okay, look at this expression over here. If $y_i > \lambda$, it is compatible with what set of numbers for x_i? Positive, right? So, over here, this is $x_i > 0$, right? If I am over here, where am I? $y_i < -\lambda$, x_i is actually negative. We are just playing around with this expression by

substituting different values, right? You can take either of these two expressions when. Let us take this expression, right? You can substitute when you start with a value of $y_i = 0$, right? When $y_i = 0$, what value will I get for x_i? 0, right? When $y_i = 0$, then I am going to get...

Correct, it is consistent. If I substitute $x_i = 0$, fine. Now, when I start increasing y_i, when y_i goes past λ, what happens? Then x_i becomes $x_i = 0$ at the border of $y_i = \lambda$. Now, as I start increasing x from here, right? It is consistent. So, this is kind of the operation that is happening between x and y. It is interesting to plot it even further. Here we only plotted y_i, now let us actually do a simpler thing: let us plot x_i on one axis and y_i on the other axis. This will give the full picture. So, y_i continues to be here and I am going to plot x_i over here, okay?

So, when y_i is... and let us keep these equations in mind, right? These are the equations to be kept in mind. When $x_i > 0$, we have $x_i = y_i - \lambda$, that is the equation of a line. How does this line look like? The slope is going to be negative, positive or what? Positive, right? So, it is actually going to be, right, because in this equation, what is the intercept? $-\lambda$. $-\lambda$, right? $y = mx + c$, $m = 1$, $c = -\lambda$. So, it is down. Okay, then let us take this equation. This has... what is the slope of this guy? 1. The slope is still 1, but intercept is $+\lambda$, right? So, this is actually going to be like... I mean, drawing is not to scale, but basically something like this, okay? Actually, this I am going to leave as this, okay?



In these equations, if x_i is... and if y_i is less than λ, then $x_i = 0$ is consistent. Right? Do you see this in this expression? If y_i is less than λ, what happens? What is the value of x that I can choose? 0 is the only thing that will work. Between λ and $-\lambda$, only one possible point, meaning... correct? So, this... so you are saying it is not the whole... yeah, it will be like that. Because we chose signum. Okay, so this is how this is the relation between x_i and y_i, okay.

So, let us continue building on this. I will try to write this in an even more compact way, okay, if it helps. Yes, I took the subdifferential as 0, that is why the signum function is floating around

everywhere, okay. It is not undefined; I have chosen the subdifferential to be 0. No, x_i is y_i according to that relation, right?

If y_i is less than λ, yeah. So, let us look at this expression, right. So, if y_i is less than λ, as per this relation, what do I get? We cannot determine x_i from this. That is the point, right? That is precisely the point: I cannot determine x_i from this expression if y_i is less than λ. If y_i is greater than λ, then... yeah, but that is okay; we want to solve this problem. So, we are picking the signum of x to proceed with solving this problem, okay?

Now, this expression that I have written over here for x_i, finally, remember, we know y_i and we want to solve for x_i and λ. That is what we want to do. So, I want to rewrite this in a way that will help me a little bit more. So, I am going to write this as... tell me if you agree. Basically, look at this graph, and we are trying to write this graph in a compact form, okay.

When $y_i > \lambda$, what is the value of x_i? I am reducing y_i by λ, right? And what am I doing? The sign is positive. For example, when y_i is greater than this, when y_i is less than λ, what happens? Again, the value of x is being chopped by λ, but it is getting multiplied by a negative sign. So, this expression over here for x_i is:

$$x_i = \text{sign}(y_i) \cdot \max(|y_i| - \lambda, 0)$$

This will either be plus or minus depending on whether y_i is greater than or less than λ, and the $\max(|y_i| - \lambda, 0)$ is showing me the "chopping" that is happening over here. Graphically, I can see there is a chopping happening over here, but notice I have defined this $+$ operator. This is to take care of what? This issue when y_i is between $-\lambda$ and λ.

So, this is a well-known operator. This $+$ operator is simply:

$$\text{ReLU}(x) = \max(x, 0)$$

If I write a over here, this is the meaning of the $+$ operator. So, if the argument is greater than or equal to 0, then the operator spits out the same number. If it is less than 0, it spits out 0. So, in layman's words, what would you call this? Thresholding. You just threshold if a number is greater than 0 or clipping. Right, you are clipping a signal to not go negative.

So, this expression explains the graph.

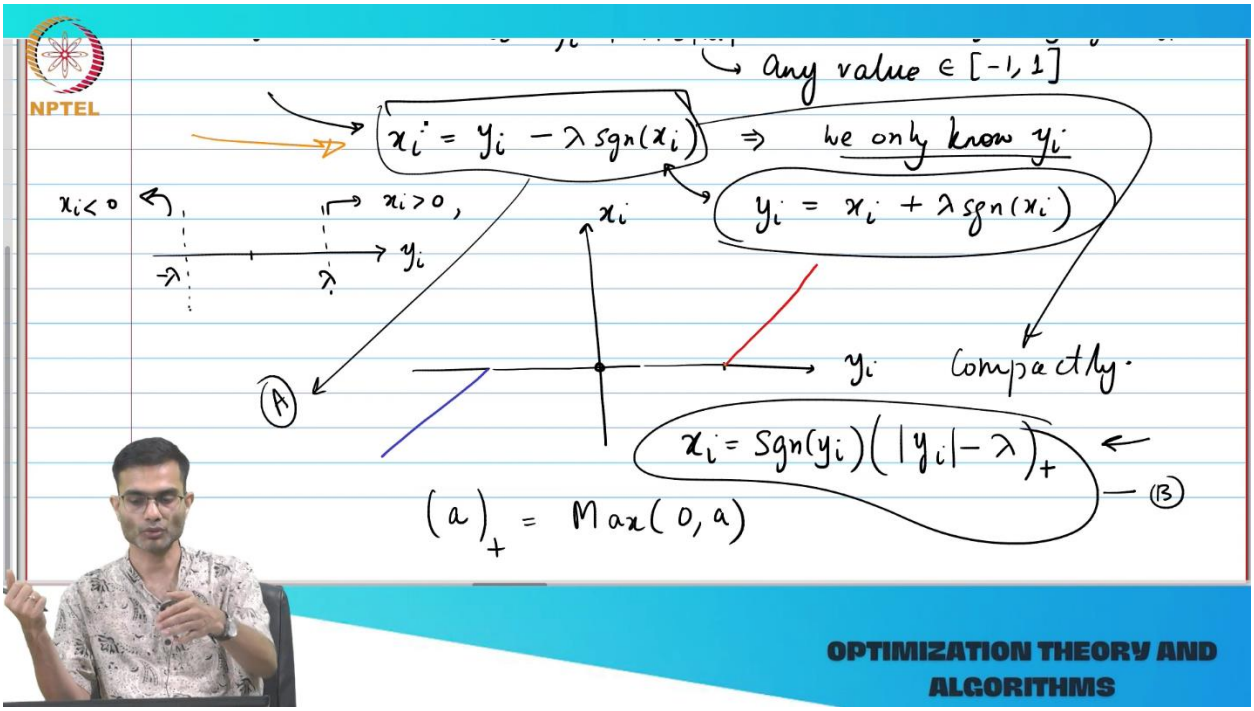
I have got this expression over here. I am still not out of the woods, right, because I want to be given x_i. The projection operation should give me x_i, right, but I do not know x_i and I do not know λ. Okay. I have got a convenient expression. You notice that between this expression, let us call this expression b and this expression, let us call this expression a.

What is the advantage of expression b over expression a? Or the disadvantage of expression a over expression b? Exactly, right? In expression b, it is very neat because everything that I know is appearing on the right-hand side. I can work with this. On the other hand, in expression a, I have the signum of x_i but I do not know x_i to start with, right? And the unknown is appearing both on the left-hand side and right-hand side. That is, x_i is appearing on the left, it is appearing on the right. So, it is a clumsy thing.

So, all I did from that expression is to plot it, and then I rewrote it in this form, right? So, what is happening in going from expression a to expression b is not algebra, it is sketching, or in other words, geometry. That is what has allowed us to go from expression a to expression b. So, do not always rely on brute force algebraic symbol manipulation, as they call it, right? Sometimes just plotting it out shows you this is what it actually is.

Now, let us get... yeah, question. When y_i is between $-\lambda$ and λ, what happens to the $+$ operator? What will it spit out? 0, right? That is what we have.

So, when I have... one sec, so $x_i = 0$ is happening for the entire $-\lambda$ to λ. For when y_i is between $-\lambda$ to λ, all I only get one value. That is like the signum function, right? But I do not want all the points. I am fixing it to be 0.



Which is 1? Which is 0? No, okay. You look at the signum function. For the entire range of values, I am only picking one number: 0, right? It is not continuously varying.

We can work it out, but you see in the signum function, it is picking 1, right? So, for the entire range, I am forcing the function to take only one value of 0 rather than allowing the function to vary. That is the difference. We can work it out by varying the subdifferential, but if I do not vary the subdifferential, correct. So, if I do not vary it, supposing I fix it at 0...

Yes. Oh, for all the values. No, right? Because I will only get $x_i = 0$ no matter what value of y_i is between $-\lambda$ and λ, which is true for all values of subdifferential. Is that true?

So, let us look at the general expression that we had, right? What was the general expression? Yeah, this, right. So, we are looking at this expression, right. So, this expression was written only at $x_i = 0$, right? At $x_i = 0$, you are right, that would happen only if the subdifferential basically mimics the value of y_i and there is a λ missing, correct. To maintain $x_i = 0$, the subdifferential has to basically take the value from -1 to 1 as y is going from -1 to 1, yeah.

No, there is a λ multiplying there. No, as y goes from $-\lambda$ to λ, the subdifferential goes from -1 to 1. Therefore, $x_i = 0$ remains $x_i = 0$, correct? You are right. So, the subdifferential tracks the value. So, this graph is also fine then, I mean, right? The graph is fine. x_i remains at 0 as y_i goes from $-\lambda$ to λ. That is an interesting construction. So, that is why to show it, it was that is why I erased the orange line connecting, right? It is just one point over here for the entire range from $-\lambda$ to λ, which is unusual, kind of. You cannot even draw it, right? Can you draw it? It is a line actually. It is correct then. Yeah, actually.

So, no matter what value y_i takes between $-\lambda$ and λ, x_i remains 0.

That is a proper... So, then that is a proper line. So, I was right to begin with. That is good. Yeah, okay, right? So, you notice this way of writing this graphical relation between x_i and y_i. Particularly for those of you working on the course project, you may come across this operator many times. This is also called a soft thresholding operator, okay.