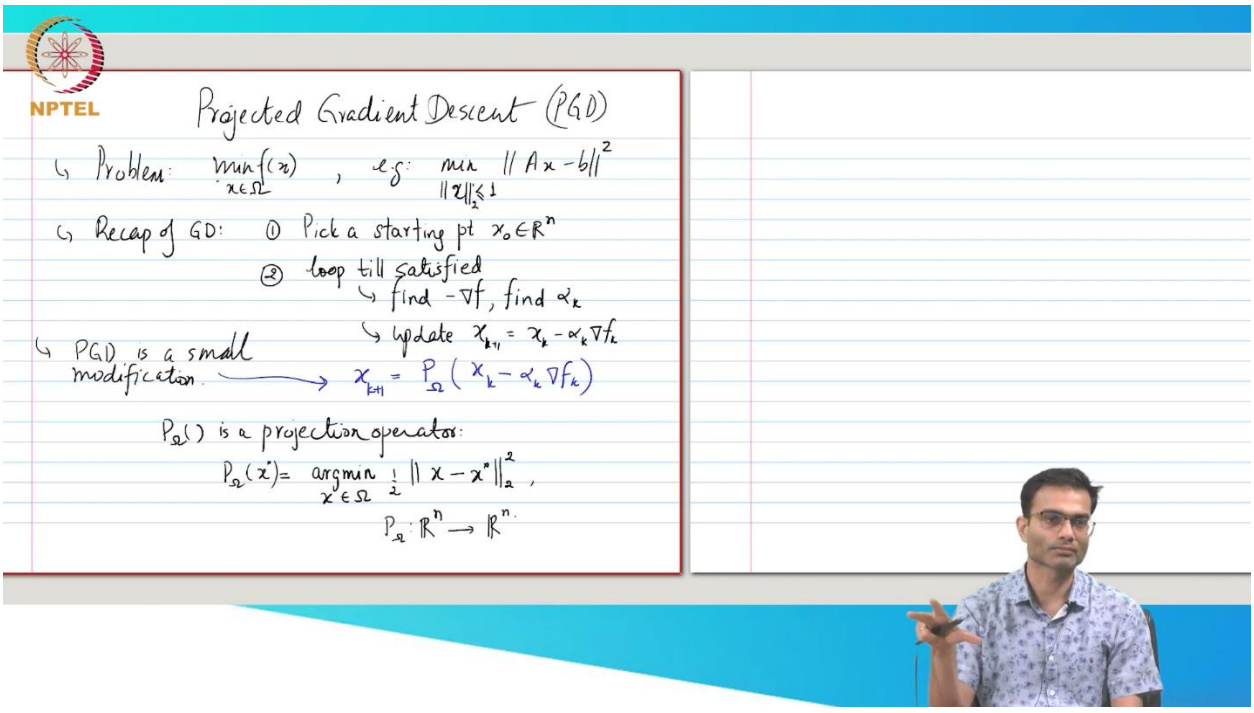
Course Name: Optimization Theory and Algorithms
Professor Name: Dr. Uday K. Khankhoje
Department Name: Electrical Engineering
Institute Name: Indian Institute of Technology Madras
Week - 11
Lecture - 72

Projected gradient descent and proof of convergence

So, what we started talking about yesterday was the first algorithm for constraint optimization, projected gradient descent. So far, we just gave you the basic definition. We saw how you start with gradient descent and do one small update or modification to it, rather, which is that once you do your usual gradient descent update, you also do a projection. And the projection operation is itself an optimization problem. Given that there is some constraint set, feasible set, find me the point in the set which is closest to my target point. That is the projection operation. So we will continue talking about it now.



We have defined, let us say, x^*. So, what does this mean? Projection of the point x^* onto the feasible set Ω. And how would I define it?

$$\underset{x\in\Omega}{\text{argmin}} \| x - x^* \|^2$$

This is the problem that I have to solve.

This problem may or may not be simple depending on σ, right? So we will just make a few notes. PGD is useful only if the projection operation is easy to compute, right. So only if P_σ is easy to compute. It can often happen that the projection operation itself is very, very difficult or resource-intensive, in which case you should not use this algorithm.

There is another very nice property of this projection operation, which is a special case when σ is a convex set. This makes our life very easy, or rather I should say simple. If my constraint set is a convex set, then if I do the projection operation, if I solve this problem correctly, the point that I get is unique. So, this is something that is going to be very useful for me.

So, what does it look like graphically? Let us look at it graphically. So, let us say this is my set, Ω. This is the constraint set, which means x should live in the interior of this, interior or boundary of this. So, let us say this is my point, let us call it x_0. If my point x_0 is over here and I ask you, "What is this? What is the nearest point to x_0 within Ω?" So, that is the trivial very easy case for me to deal with. Now let us take a point over here, x_1, and the question I am asking now is, "What is the nearest point to x_1 within Ω?" That is what I want to know.

So, at least from a graphical perspective, how do you think we should start solving this problem? You see the definition of it, you see the algebraic definition. This simply means find the nearest point in Ω to my target, my candidate point. So that is the geometrical meaning of this. Now I am at x_1, I want you to find me the nearest point from Ω to x_1. So think of yourself as, you know, in high school geometry with protractors and all of those things. How would you solve this problem? I don't think you knew grad C in class nine or ten.

Perpendicular, I mean, so I have a scale and what am I doing? Okay, that is correct. How would I construct it? Would I just keep drawing, take a ruler and draw lines from x_1 until I find a line that is cutting the boundary at 90 degrees, or is there a smarter way of doing it? Exactly, right. I want the smallest distance. What is the locus of points that are at a constant distance from x_1? A circle, right. So, if I take for example something like this.

All the orange points have the same distance from x_1. Now I can, what I can easily start doing is start increasing this. At some point, what will happen? At one very critical point, I will just touch this blue bit. So, I am going to get something like this. So, graphically that is what this operation is doing.

It is finding me this point over here. Let us call it y. So, in our fancy language, we will call this an L_2 ball. So, this is its radius r, then I say that this is an L_2 ball of radius r, obviously centered at x_1, that is the meaning of this guy.

So, I am just going to write down a few properties of this projection operation, which are mostly intuitive. So I would not be stating them by proof.

But the first property, if x_1 is outside, do you expect that the projection point will be in the interior or on the boundary? On the boundary, right. So, y will belong to the boundary. Now, there was one student who mentioned something about orthogonality or perpendicularity, right? So, it is actually correct. So, if you look at this tangent point over here, the line connecting the point x_1 and the projection y turns out to be perpendicular to, what can I draw at this point? Perpendicular to what would you say? The tangent. The tangent at this point.

So, okay. Do we already have a geometric object that captures this property of tangency? The tangent cone, okay. So, this guy actually at y, if I sit and construct feasible sequences, take feasible scalars, get a tangent, what I end up constructing at this point is a tangent cone. Okay, and so again I am going to state this without proof: that the vector $x_1 - y$, this vector, should be

in what relation to the tangent cone? Perpendicular. This is going to be perpendicular to the tangent cone of y. So these are just some simple intuitive properties of this projection operation.

Notice that we are using, I mean, I have written all of these properties for convex sets. If the set is not convex, all sorts of strange things can happen. So you can imagine, let me not draw it, but you can imagine weird shapes where a point approaches, let me draw it. Supposing this is my point and my set is like this. Now strange things can happen, for example, I may have more than one point, right? Various things can happen.

So this discussion is limited to the case of convex sets. And the good news is that many times the feasible set is a convex set. So this, for many, many engineering problems, it works out that using convex sets is good enough. For example, let us say you're optimizing the power of an antenna in an array. Okay, and your options are only, obviously, so what would be the first common sense constraint you would put on the power? It has to be non-negative, it makes no sense to talk about negative power. So, this is the first common sense constraint. What else can you think of? Just think of a real-life engineering problem. I will have some maximum amount of power that I can work with.

I will never have unbounded power, right? So, the second common sense constraint would be less than or equal to some, let us say M, right. Now, what kind of a set is this? If I, if you sketch it, what kind of a set is this? It is a convex set, right? So, if I draw p like this, this is 0, this is M, this is where my, this is my feasible set.

It is clearly a convex set. So, 90 percent of engineering problems will end up having convex constraints. So all the nice properties of the projection operator which I have written will hold true. Now, if you look back here, I have simply defined over here this, this is the projected gradient descent algorithm.

Which is fine. I mean, at this point, it's just like I've given you some information. We haven't proved it. Why should this work is the first question that should come to your mind. This looks almost too simple for it to work. So, what we'll do is I'll give you a proof of why this works.

So it is your last and final convergence proof for this course that gives us faith that the method actually works. So the reason I am saying this is because you may come up with several very intuitive ideas to solve a constraint optimization problem, but they may not work. You have to also prove that it works. So let us look at the convergence analysis of this. Now I am going to make a little bit of a relaxation over here.

I am going to prove a slightly weaker version of it just so that the proof is simple enough to understand. And I will show you what that weaker version is. I am going to make, my starting point is going to be, obviously, gradient descent plus projection. Now in gradient descent, what are the different variants possible for α, the step length? I could do backtracking line search, that is fancy.

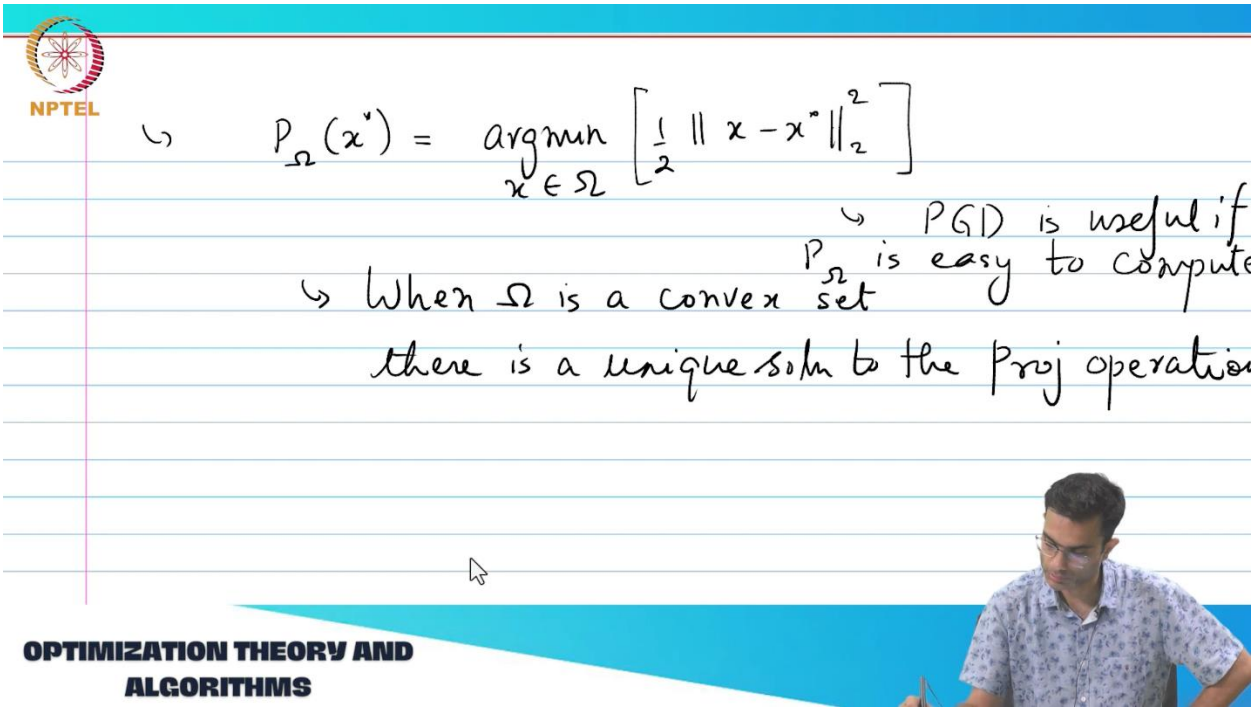
Could I do something less fancy also? Exact line search is even more fancy. I am going in the opposite direction. I want something simple to start with. I could do constant line search also. Constant line search will just make the method converge a little slower.

But it is something that works. So to make this analysis a little bit simpler, I am going to start with constant step length. That is the assumption we will make. By the way, in the literature, you

will find proofs for all flavors of it, you know, backtracking, Wolfe condition, blah blah blah, right. But the proofs are more involved.

Okay. Alright. So again, we are going to make use of graphs for intuition. Okay. So this is f and so I am going to, I am working with a convex function.

Okay, let us say, so let us assume a convex function will look something like this. So, here is, let us say, this is my point x over here and this is some later point over here. And the property of a convex function (not sets) is what? The function will lie above or below this blue line? Below. That is the definition of a convex function.



So, let us say the convex function f is such that, if I were to write Taylor's theorem for f. So, let us assume that x and z are close to each other. So, I want to write $f(z)$. I know all of you are experts in this. First order Taylor's theorem, what should I write? The first term should be $f(x)$ plus what else? $\nabla f(x)^\top (z - x)$.

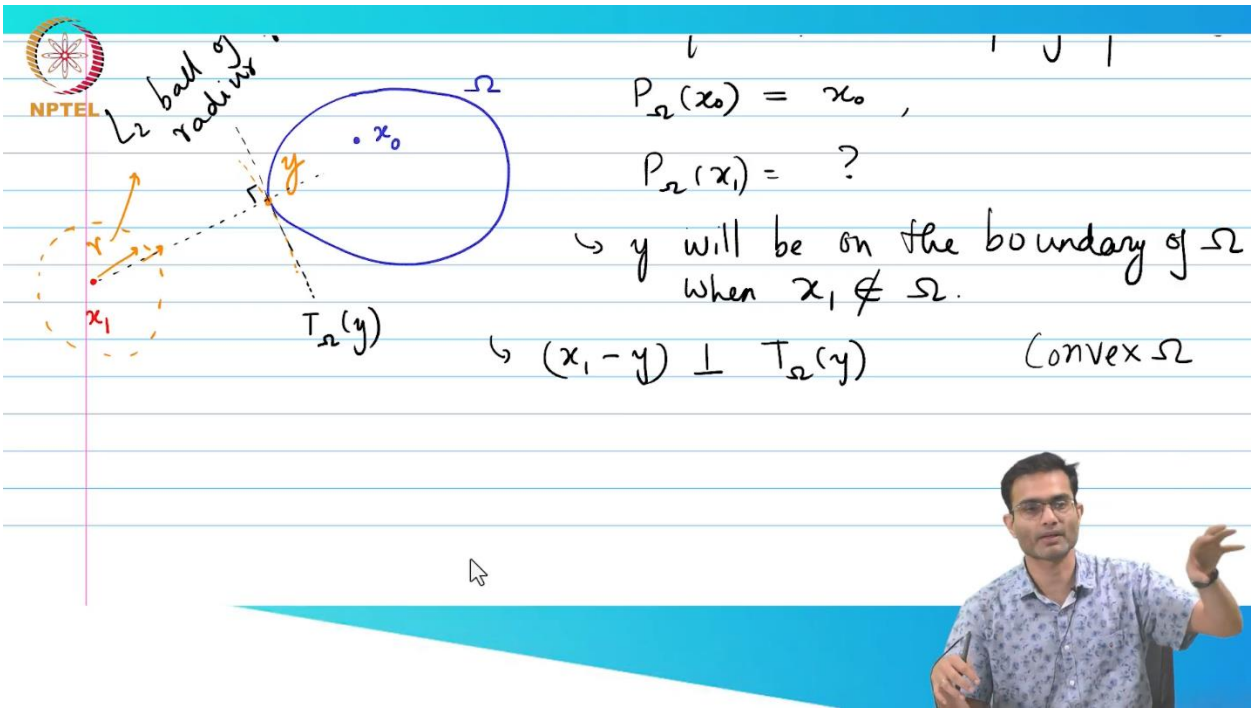
This is first order. What is the error that I am incurring if this distance is, let us say, δ? So, there is a higher-order term over here which is of order δ^2, that is the error that I am incurring. So, assuming that we are using the first order Taylor's theorem, we know that, so $f(z)$ is over here, right? So, what I have drawn is the line connecting $f(x)$ and $f(z)$. Which way will ∇f look? Will it be this dotted line or something else?

You got the difference? This is just a straight line connecting $f(z)$ and $f(x)$. If I wanted to sketch ∇f, because I have written ∇f over here, I want to sketch ∇f. What way would it look? ∇f is what? The tangent, right? So, it should be the tangent at which point? At x, right? So, the tangent is going to be like this. And again, as this is another property of convex functions, the tangent will be above or below the function in the local neighborhood. Below, right? Because it is like an upward opening cup.

So, the tangent is always going to be below. Excellent. Okay. So now I am going to ask you a question. What is this point over here that is marked in orange? In terms of its value, what would you say? That is correct. Anybody else wants to take a guess? How do I arrive at this orange point from what I have already described so far? Yes, $\nabla f(z)$?

No, $\nabla f(z)$ would be the tangent at the point z. $\nabla f(x)$ into δ, are we missing something? Right, so actually this, what, this point that I have written over here is actually this whole thing, right. This tangent over here plus $f(x)$, right, I have to move that tangent up. So, what this is, is the linear approximation of f at x.

That is what it is. So, this is the linear approximation of f at x. This is a graphical meaning of first-order Taylor's theorem. Take a point, take the tangent, go along it, that is your linear approximation. So now that you can see this graphically, can, so let us ignore this error term on the right-hand side. Do you think that we can actually write this now as an inequality? $f(z)$ will be greater than or less than? It is a convex function.



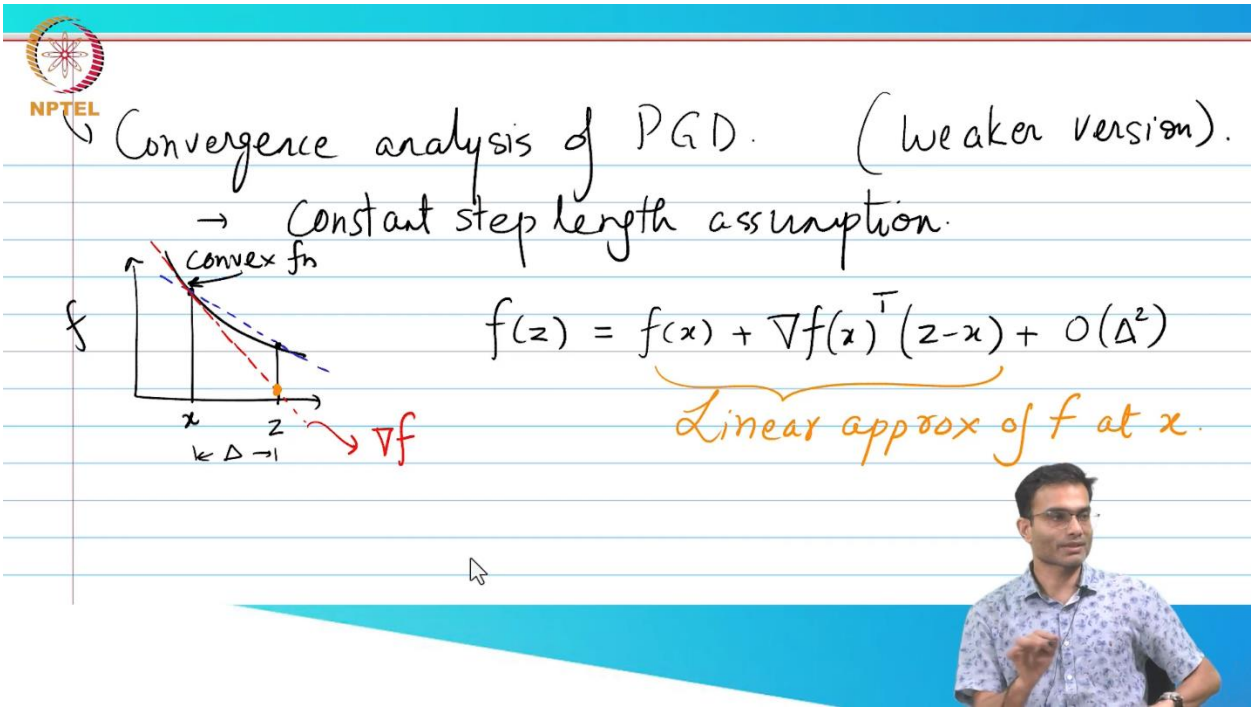
You can see it on the graph, right. This point over here, what is this value? The red dot. This is $f(z)$ clearly, right? So, it is very easy to see that actually $f(z) \geq f(x) + \nabla f(x)^\top (z - x)$, okay. So, I am just going to write this over here.

Assuming or rather assuming convexity of, okay. So, before I go through the entire proof, I want to just give you a sketch of what we are going to do. If you remember, we had that Zoutendijk condition when we had done the previous convergence. So, the trick that we had used there was telescoping of a series in such a way that $f(\text{optimal})$ and $f(\text{initial})$ were all that was there. If I can make this distance as small as possible or bounded in some way, I will arrive.

So that is the strategy that we are going to do, okay. What is the complication over here? The complication is that at each step I have to do a projection, right? Every iterate has a projection

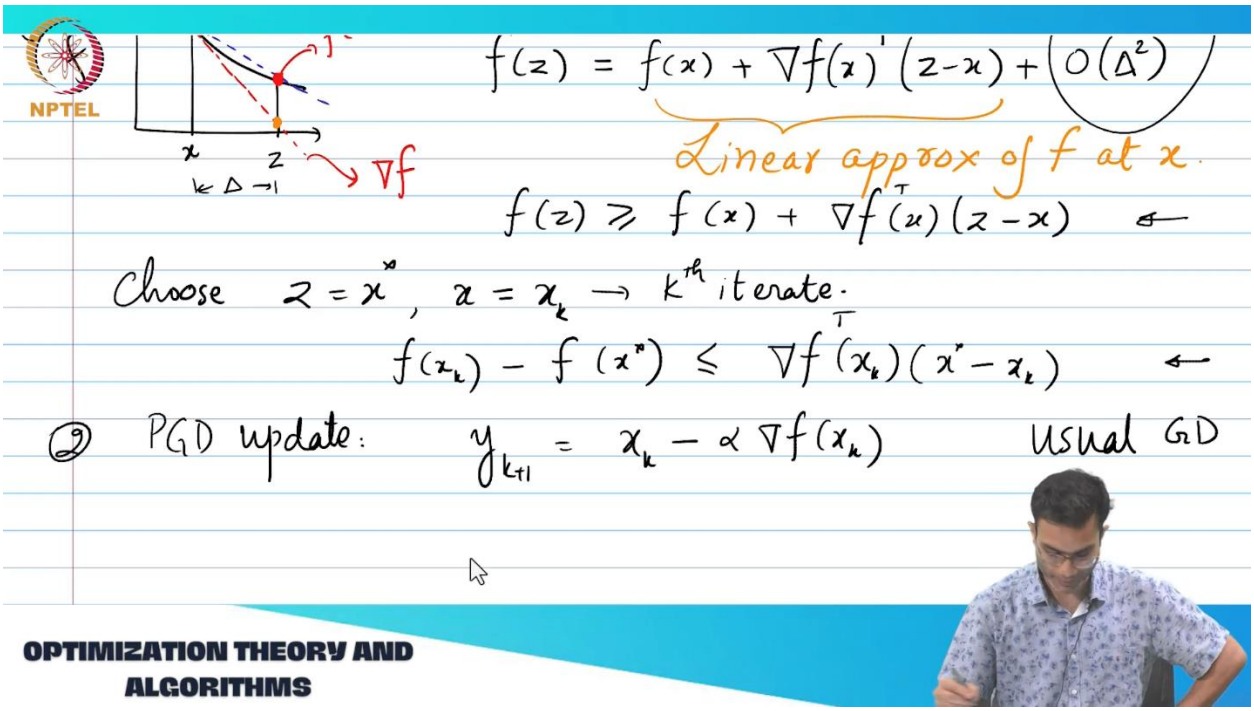
operation and then, I mean, there is an update, but the update is followed by a projection. So, that is different from the usual gradient descent which I had done. So, that is something to keep in mind. So in this inequality, am I free to choose x and z any way I want?

Write it up to me.



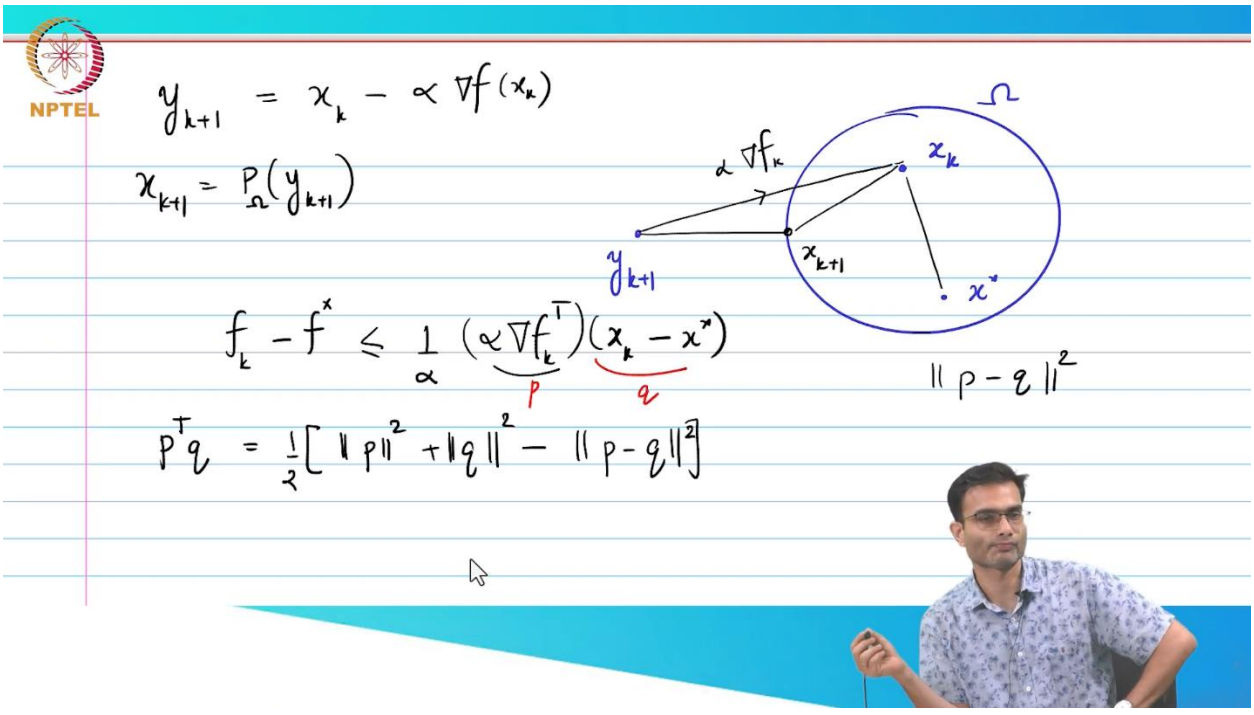
So what I am going to do is I am going to choose, let us assume that this sequence of iterates converges to some point x^*. So, let us call that x^* and I am going to put $x = x_k$. So, this is the k-th iterate. So, then I will get $f(x_k) - f(x^*) \leq \nabla f(x_k)^T (x^* - x_k)$.

So, this is let us call this as this step 1. Step 1 is setting up this inequality and this gave me. So, this is going to be my starting point. So, just you can see how I can think of this telescoping business, right? They have $f(x_k) - f(x^*)$. Just imagine if I had this for many iterations and I summed it up, would stuff cancel out, right? So, that is roughly the direction that we are going in except x_k itself is of not much interest to us.



What is of interest to us? The projection of x_k. So, we have to get in this projection operation somehow. So, that is going to be my step 2, which is the PGD update. So, the step without projection I am just going to use a new symbol for it, y. So, y_{k+1} is simply $x_k - \alpha \nabla f(x_k)$.

It should be? No. I, no, I think it is fine, right. On the right-hand side I put $f(x)$, right-hand side of this expression is $f(x^*)$. Oh, no, it is correct. Does someone want to verify this? It is correct. I have just flipped them on each, on the other side, right.



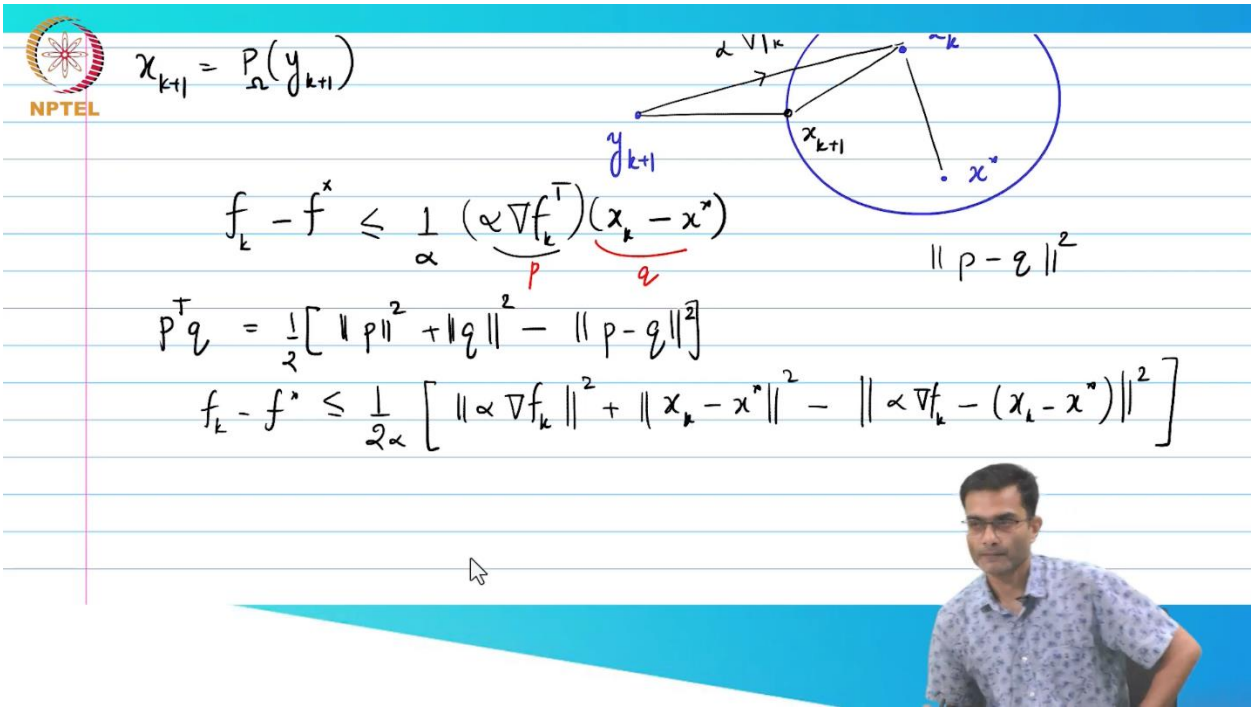
So, this is our usual gradient descent update. Maybe I will go to the next page so that I can draw this. So, let us say that this is my usable set. Let us say this is my x_k, let us say this is my x^*, ok. And let us say that this is my y_{k+1}, ok. Now, after projection, what will happen to y_{k+1}? It will fall on the boundary, right.

Ok, let us look at that once again. So, let us take this guy and let us put the substitute away $f(x^*) \geq f(x_k) + \nabla f(x_k)^\top (x^* - x_k)$, okay. And so I okay, right. So, then this is if I keep it, that is if I want to keep $f(x_k)$ there, then so this is going to be in.

So, you are right. So, there should be the inequality gets flipped. Is that right? What if I do a little bit of correction? Thanks for catching it. Just correcting. This is my projection point.

So, x_{k+1} is obtained by projection of y_{k+1}. This is the definition. We also had y_{k+1} as I just wrote on the previous slide, $x_k - \alpha \nabla f(x_k)$. So, I can draw a vector connecting, basically I can reach y_{k+1} by joining what? x_k and this $-\alpha \nabla f(x_k)$, right. So, in shorthand notation, I am just going to write the previous inequality f_k for $f(x_k)$ and f^* for $f(x^*)$.

This is what I had, $\nabla f_k^\top$ and I had $x_k - x^*$. This is what I had. So, I am going to multiply and divide by α so that all the terms appearing in the graph also appear over here. Okay, everyone with me so far? Now I have this, I have two vectors, $\nabla f^\top \times \alpha$ which I have sketched over here, and the other vector is $x_k - x^*$. And I have the inner product of these two guys.



I want to somehow separate these guys out. Okay. I want to separate them in a way that I can work with them. Because if I just sum this, I am not going to get anything interesting. If I telescope this series, for example, the right-hand side summation is not going to be anything that I can work with. So, I need to somehow split these guys into separate terms.

So, a very simple inequality, not inequality, a way to express $p^\top q$. Do we know how to do this in terms of p and q? So, if I, what is for example $\| p - q \|^2$, what all terms will be there? Think of this. What all terms will be there? $p^\top p$, $q^\top q$, and a negative $p^\top q$ and $q^\top p$ which are the same. So, there is a two factor over there, right. So, that is a nice way to get this. So, $p^\top q$ will be simply $\frac{1}{2} \| p \|^2 + \| q \|^2 - \| p - q \|^2$.

So, I am going to take this guy as p, this guy as q. I am going to apply this over here so that these guys separate out. So,

$$f_k - f^* \leq \frac{1}{2\alpha} (\| \alpha \nabla f_k \|^2 + \| x_k - x^* \|^2 - \| \alpha \nabla f_k - (x_k - x^*) \|^2).$$