

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 - 73

Proof of convergence - Part 2

This last term on the right-hand side, can we see something is getting clubbed together? For example, do I know if $\alpha \nabla f_k - x_k$ is anything that we already know? It is y , it is the negative of y_{k+1} . And so this term will simply become, what will it become? It will become $x^* - y_{k+1}$. So, this is, this is good because I have now brought y_{k+1} into the picture. Bringing y_{k+1} into the picture is good because that will be the next step once I get y_{k+1} , what is the next step from y_{k+1} ? Projection. So, x_{k+1} will enter the picture. So, I want x_k and x_{k+1} both to enter the picture.

$$p^T q = \frac{1}{2} [\|p\|^2 + \|q\|^2 - \|p - q\|^2]$$

$$f_k - f^* \leq \frac{1}{2\alpha} \left[\|\alpha \nabla f_k\|^2 + \|x_k - x^*\|^2 - \underbrace{\|\alpha \nabla f_k - (x_k - x^*)\|^2}_{\|x^* - y_{k+1}\|^2} \right]$$

$$\leq \frac{1}{2\alpha} \left[\|x_k - x^*\|^2 - \|y_{k+1} - x^*\|^2 \right] + \frac{\alpha \|\nabla f_k\|^2}{2}$$

Since $\|y_{k+1} - x^*\| \geq \|x_{k+1} - x^*\|$, Subst:

$$f_k - f^* \leq \frac{1}{2\alpha} \left[\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 \right] + \frac{\alpha \|\nabla f_k\|^2}{2}$$

I want this intermediate guy y to be gone so that I do not have too many of these variables. So I have got y_{k+1} into the picture. So let us just rewrite this over here.

So, $\frac{1}{2\alpha}$, okay. I have $x_k - x^*$. I have $y_{k+1} - x^*$, okay. And this is a positive term. So this is going to be $\frac{\alpha \|\nabla f_k\|^2}{2}$, right? Are we in a position now to telescope the series? Not yet.

I still have my y_{k+1} sitting there. So, let us look at this geometry over here. I have y_{k+1} and x_{k+1} . Let us connect y_{k+1} and x^* also, right. So I am going to connect all the guys that are there in the expression.


Oh, there we go. All of these guys are connected. So $x_k - x^*$, this guy, is it here? It is this guy, right. And then I have $y_{k+1} - x^*$, that is, let me use another color, that is here, that is here, okay. So these two, these are the guys over here, okay.

Now, because y_{k+1} projects to x_{k+1} , can we say something about the distance of y to x and x_k to x ? Will there be some kind of an inequality? Which will be closer to x^* , y_{k+1} or x_{k+1} ? x_{k+1} because it is the projection, right. So, let us, we are assuming it is a convex set. Yeah, yeah, all of this is based on a convex set. So, $y_{k+1} - x^*$, this distance is obviously greater than $x_{k+1} - x^*$. That is clear from the geometry.

So, I am not writing convexity because it is true for, I mean we wrote it at the beginning. So, just one moment where you have to be a little bit careful. $y_{k+1} - x^*$ is greater than $x_{k+1} - x^*$. If I substitute this into the above expression, will the inequality remain the way it is? Yes, because there is a negative sign associated with it, right. So, this can be, so let us substitute.

So, this will be $f_k - f^* \leq \frac{1}{2\alpha}$, this term remains as it is and this term will become $x_{k+1} - x^*{}^2$.

So, we are now, you can see that I mean what I had in mind of telescoping will actually work now. I managed to eliminate y out of the picture and I got, so it has the current iterate, it has the next iterate and very nicely what they have opposite signs. One has a plus, the other has a minus, okay. Is everyone clear at this point what we did? We have simply used this one property over here that the projected point is closer to the solution point than the point which is farther off and of course this is only true for on the convex set. So, now let us.



Telescope the series:

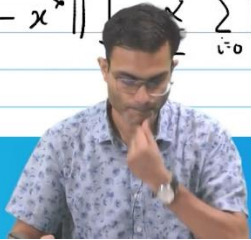
$$f_0 - f^* \leq \frac{1}{2\alpha} \left[\|x_0 - x^*\|^2 - \|x_1 - x^*\|^2 \right] + \frac{\alpha}{2} \|\nabla f_0\|^2$$

$$\vdots$$

$$f_k - f^* \leq \frac{1}{2\alpha} \left[\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 \right] + \dots$$

$$\sum_{i=0}^k f_i - (k+1)f^* \leq \frac{1}{2\alpha} \left[\|x_0 - x^*\|^2 - \|x_{k+1} - x^*\|^2 \right] \leq \sum_{i=0}^k \|\nabla f_i\|^2$$

OPTIMIZATION THEORY AND ALGORITHMS



So, I will just write down a few terms. What will be the very first term on the left-hand side? I had look at your notes. First term would be for in terms of f , $f_0 - f^*$, less than equal to. First term will be $\|x_0 - x^*\|^2$, then $\|x_1 - x^*\|^2$ plus what? $\frac{\alpha}{2} \|\nabla f_0\|^2$, right. That is the first term.

Everyone agree? $\frac{1}{2\alpha}$. Now, if I keep going like this, I am going to get finally $f_k - f^*$, supposing I write it like this. This $\frac{1}{2\alpha}$ is there. What am I going to get? This is going to be $\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2$. And now if I telescope the series means to sum the left-hand side and right-hand side.

What am I left with on the, what am I left with on the left-hand side? Does anything cancel? No, right? Life's like that. These terms don't cancel. So, I have a summation f_i , i going from 0 to k , okay? And what else do I have? How many times has this guy come? $k + 1$ times. So, $k + 1$ times f^* . What is left over here? $\frac{1}{2\alpha}$, the first term and the last term.

Those are the guys that do not cancel out. So, I have $\|x_0 - x^*\|^2 - \|x_{k+1} - x^*\|^2$. Does this other final term cancel? Negative. So this guy is going to be there as fine. Now this is the part where you have to get a little creative.

The image shows a handwritten derivation on lined paper. In the top left corner, there is a circular logo with a star-like pattern and the text "NPTEL" below it. The main text is written in black ink and includes several equations and a note. At the bottom right, there is a small video inset showing a man with glasses and a patterned shirt, who appears to be the lecturer.

$$f_k - f^* \leq \frac{1}{2\alpha} \left[\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 \right] + \dots$$

$$\sum_{i=0}^k f_i - (k+1)f^* \leq \frac{1}{2\alpha} \left[\|x_0 - x^*\|^2 - \|x_{k+1} - x^*\|^2 \right] + \frac{\alpha}{2} \sum_{i=0}^k \|\nabla f_i\|^2$$

$$k+1 \left[\frac{\sum_{i=0}^k f_i}{k+1} - f^* \right] \leq \frac{1}{2\alpha} \left[\|x_0 - x^*\|^2 \right] + \frac{\alpha}{2} \sum \|\nabla f_i\|^2$$

(2nd term is -ve)

↳ Jensen's inequality: (for a convex f_n)

$$f\left(\frac{\sum x_k}{k+1}\right) \leq \frac{\sum f(x_k)}{k+1}$$

So, I have, if I have any negative terms on the right-hand side, if I knock them off, will the inequality change? No, right. So, I can as well write this as $\frac{1}{2\alpha} \|x_0 - x^*\|^2 + \frac{\alpha}{2} \sum \nabla f_i^2$, right, because the second term does not affect the inequality. So, we are getting there. This term over here I am going to pull out $k + 1$ common from here, $k + 1$ if I take common from here, what do I have? Still not looking very good, right? I have this entire summation of function values on the left-hand side.

Now, this is the time when we pull one more, see there are only a few inequalities that are used. Can you name some inequalities that you know of? Cauchy-Schwarz is one, AM-GM is one, there is another one which is called Jensen's inequality. So, this is the time to pull the Jensen's rabbit out of the hat. So, if I look at Jensen's inequality and Jensen's inequality for a convex function. Does anyone know what it is? This actually you can appreciate from the figure.

I will write the inequality and then you can. In words, what is this saying? Function of the average point is below the average of the function values. So, when we drew this figure over here. Yeah, this is assuming f is convex. I mean it is the same as convexity.


Yeah, yeah, right. So here you can see function, so function of average. So the average point is for example over here, this green point over here and the function value is here. This is below the average of the function value which is the dashed line. The dashed line is $\frac{f(x)+f(z)}{2}$. So, that is this pink point over here.

So, you can see this is, you can call it Jensen's inequality or you can say this is basically the definition of a convex function. So, let us scroll back over here. Because that's the only term with, why did I pull out the term $x_{k+1} - x^*$? Because that's the only term with a negative sign. I want to simplify this expression as much as possible.

It's a negative sign. So imagine for example, this is, let's, okay. So imagine that the left-hand side has a value of three, okay. This is say five and this is one and this is 0.5. So, what is the left-hand side? 3 What is the right-hand side? 5 minus 1, 4.

4 plus half, 4.5. Now, if I remove this term, the inequality still holds, right? So, any negative term is not adding to it. So, my goal is to simplify this expression as much. You can see, did it affect, I got rid of a minus 1, did it affect the inequality? It did not affect the inequality. That is what I am interested in. If it were an equality, I would not be able to do this.

But it is not an equality, it is an inequality. So I have that freedom to play around with these terms. Did everyone follow? It does not matter, right. I am making a true mathematical statement, right.



$$\hookrightarrow f\left(\frac{\sum x_k}{k+1}\right) - f^* \leq \frac{\|x_0 - x^*\|^2}{2\alpha(k+1)} + \frac{\alpha}{2(k+1)} \sum \|\nabla f_i\|^2$$

Avg of iterates converges. $\rightarrow f(x_k) - f^* < \frac{M}{k}$

① $\|\nabla f_k\|^2$ must not grow faster than $O(k)$

OPTIMIZATION THEORY AND ALGORITHMS

So we are almost there. We can substitute this Jensen's inequality. We are, I will, the next step will show you what we are trying to get at. So, I will write down the final expression after using Jensen's inequality, right. So, $f\left(\frac{x_k}{k+1}\right) - f^* \leq \frac{1}{2\alpha} \|x_0 - x^*\|^2 + \frac{\alpha}{2} \sum_{i=0}^k \nabla f_i^2$. So this is actually the final step of our very, not very, but somewhat complicated looking convergence proof and now is where the sort of creative part is in interpreting what I have got. So let us interpret the result. What is the left-hand side saying? What is the left-hand side, so now we will switch to plain English. What is the left-hand side saying? There is an optimum function value which is given by f^* .

Everyone agrees? I have f of the average of iterates. So, what is it saying? The distance between f of the average of iterates and the optimum function value, it is less than or equal to some expression. What is that some expression? There is a starting error $x_0 - x^*$, so it is like a constant. There is α , it is a constant. $k + 1$, what is it telling me? As I increase my iterations, what is happening to this term? It is reducing, reducing.

Let us say $k \rightarrow \infty$, what will happen to this term? This term will tend to 0, right? If it tends to 0, what will happen to this guy on the left-hand side? It will say that in the average sense, the average of the iterates is tending in function value to the optimum f value, ok. And this guy on the, the second term over here on the right, what about him? We have to make sure what? So there is a $k + 1$ also in the denominator. So ∇f should not blow up. It can grow, but it should not grow what? Faster than k .

If it does not grow faster than k , if its growth is less than order k , what will happen to this term? This term will also die as $k \rightarrow \infty$. So, I am adding smaller and smaller things. The net result is the right-hand side is going to be some finite number which keeps decreasing, right. So, this is, this proof, it is relatively simple if you look at the other proofs in the literature.

It is saying that the average of the iterates is converging to the correct function value, ok. So, for that to happen, we will just make a quick note of what we discussed. ∇f_k right, must not grow faster than order k , must not grow faster than order k , ok. And so, this can be satisfied by you know things like Lipschitz etcetera, right.

So, this is not very hard to ensure. So, what is this saying? That the average of the iterates is what it is saying, right. Now, what would have been really nice if instead we had $f(x_k) - f^* \leq$ or maybe something like this, right. This would have been great, yeah, okay. ∇f squared should not go faster than. Ideally, I would have liked this that as my iterations go $f(x_k)$ and f^* should come as close to each other as possible.

This is what we would like. So this proof exists in the literature. It is much more complicated than what we did. So I am going to, I am not going to do it. What we have done is that is why when I started the proof I said we are going to do a slightly weaker version. The weaker version is we are going to prove that the average of the iterates converges.

Average of the iterates in a long series of k , what you started out in the beginning is not so important, right. You may have 10 points that are far away, but as the operation goes on you may come and converge close to the, right. So, this is also true, supposed to be look up. So, the average of the iterates is not converging, I am saying in the f value of the average of the iterates matches the optimum function value. You can imagine a downward like a cup right, the points

are getting closer and closer and closer as you get to the optimum point over there right, that is one way of doing it.

Okay, so this is your we could say the last proof of the course which at least now when you say that you are using the projected gradient descent method you know at least why it works. Otherwise it is like you know downloading an algorithm from Wikipedia and implementing it without understanding why it works. So at least you have some idea of why it works and you can see that in the proof what are the key points? We have basically used properties of convex sets, right? So there are courses where you have, you study only convexity and convex functions for the entire semester, prove each of those properties very, very laboriously in detail. So we have taken a little different approach. Let's just use those properties to build on and give you some interesting results.

So I'm going to show you just one example of this projection operation. In subsequent classes, we'll come up with a little bit more complicated examples. So the simplest example that you can think of is projection onto a feasible set, which is a ball. That's the first projection example that we rotate. So the L_2 ball, yeah, projection is giving me that point in the feasible set closest to the point that I am asking.

So, this is my feasible set. You can see that this is a sphere, right, the sphere in n dimensions, and I am saying this is my feasible set, x must live inside this. Now, so let us just draw a two-dimensional version of it and I am, here is my point x_0 , okay. And what I want to work out is, what is, okay. This again looks simple enough that you could solve this problem without any calculus, without any Lagrangian business. So when, let us take two cases, when x_0 is in the feasible set, projection operation obviously is what? x_0 .

Now, when x_0 is not in this, how do, what is the solution to this problem? It is very simple, anyone? Smaller circle okay, but can we draw this with a pencil and ruler? Correct origin, exactly. Take the origin, connect it over here and the point over here, right. This is actually what we have discussed later, draw circles of bigger and bigger radius from x_0 is going to touch exactly over here, right. Now, have a look at this new, can I write this analytically? So, this is $\frac{x_0}{\|x_0\|}$, which norm? ℓ_2 -norm obviously. Why? Because this gives me a unit vector in that direction and that is exactly what this sphere has, radius 1.

So, this is our projection operation. Is there a compact way of writing this? No, the two cases, Is there a compact way of condensing this? So, basically if I use the max operation, supposing I write this as $\frac{x_0}{\max(1, \|x_0\|)}$, does this work? When x_0 is, when $\|x_0\|$ is greater than 1, I am outside, then this will give me the unit vector. When I am inside, 1 is greater than the norm of x_0 , it gives me back x_0 . This is your first very very simple example of a projection operation. This was so simple that we did not have to actually solve any optimization problem.

And this is a very useful object in itself. You will come across this in many, many examples. So we are done for today. Anything that you would like me to go over once again? So what we did was we started with the brief statement of the PGD, right? Some properties of the projection operation, how to think of it graphically or geometrically and then we started this proof, right. The proof was trying to get a telescoping series that I could sum and along the way I had obviously I got y and I tried to get rid of y , I got x and then I used my Jensen's inequality or

property of convex set and I got that the final expression here that the average of the iterates is shown to converge.

And that is a weaker version, but there is also stronger version. Yeah. The tangent cone, tangent cone, the set of all tangents at y 's and y 's. Okay. Any, anything that I should clarify once again? M , some number. I mean, you mean, where it go? This guy, some number, some finite number which is growing slower than k .

NPTEL

Avg of iterates converges. $\left[f(x_k) - f^* < \frac{M}{k} \right]$
 also true, look up.

① Projection. eg. L_2 ball: $\Omega = \|x\|_2^2 \leq 1$. $P_\Omega(x_0) = ?$

When $x_0 \in \Omega$, $P_\Omega(x_0) = x_0$
 When $x_0 \notin \Omega$, $P_\Omega(x_0) = \frac{x_0}{\|x_0\|_2}$

Then as I take the limit $k \rightarrow \infty$, $f(x_k)$ will tend to f^* , that is what I want, right. So that is the proof which we have not done, but that proof also exists in the literature, you can find it. No, it does not imply, this is not the running average because I started from 0. If you wanted to you could start from somewhere else. Instead of starting from x_0 , you can start from x_{100} , it is your choice.

Then x_{100} will appear on the right-hand side. It is up to you where you want to start the iteration counter. The form of the expression will be the same, right. Will PGD work for non-convex where the function is itself non-convex or the constraint set is convex? There are two things over here to ask about. So which one are you talking about? See the function could be non-convex, that is one thing.

The constraint could be convex, could be non-convex. So this is a, it is a good question. How do you define a convex, a convex optimization problem when it is a constraint optimization? Both f and the constraint set. So the constraint set has to be a convex set and the function, objective function has to be a convex function. Only when that is the case, this PGD has this nice proof of convergence and so on. But if it is a non-convex optimization problem, then you do not have all of these nice guarantees.

Then you have to go case by case and see whether you can make it, whether you can show convergence. There is proof for everything. I am not, I have not done it. Of course, there is proof for everything. All of these statements as I mentioned if you had a full semester course on this convexity you would prove all of these things to go really deep, it is very much possible.

I think Stephen Boyd's book which is titled *Convex Optimization* has all of these proofs, ok. Which statement? It is at the top, right? Second statement or second slide. No, it is remember the tangent is defined as the limiting direction. Not, I do not take any $z_k - x$, I have to take the limiting as $k \rightarrow \infty$. So, Yeah, $x_1 - y$ you mean? Yeah, but it is not, I mean it need not be radially inward, but it is perpendicular to the tangent in this case.

I can draw some funny diagram, but at that point where the projection point is I draw a tangent and these two lines are perpendicular.

I am stating without proof. Correct. Yes, it is. That is what I am saying. I am writing this without proof. It is a true statement. You will have to take it on faith and we can, I can refer the proof of that.

No. So, here it is a slightly different route we have taken. The final term which is $\frac{\alpha}{2} \nabla f_k^2$, there is no, it is a positive term, right. And as long as we said ∇f_k^2 does not grow faster than k , it is going to be a bounded sum. So, we are not following the authentic proof exactly like that. So, $\cos\theta$. The only thing I mean the only reason I remind you of Zoutendijk's condition was because of we had used the idea of telescoping series that's all, fine.