Distributed Optimization and Machine Learning

Prof. Mayank Baranwal

Computer Science & Engineering, Electrical Engineering, Mathematics

Indian Institute of Technology Bombay

Week-2

Lecture - 6: Strictly and strongly convex functions

So, in the last class we kind of briefly introduced what convex for sets convex functions are and what is one particular one specific implication of working with convex functions something that we iterated upon multiple times. So, every local minima is also a global minima right. So, with convex functions every local minima is also a global minima. But just the function being convex alone, it does not sort of eliminate the fact that you can have I mean you can still have multiple optimizers right. So, let us say I have a function which looks like this. So every point here is a minimizer of this function.

So I can still have multiple minimizers, but all of them are going to be globally optimal. So is there a class of functions where we are guaranteed, for instance, we are guaranteed to have just one unique minimizer? And the answer is yes. So if the minimizer exists, so today we are going to look at something called strictly convex functions and strongly convex functions. So for convex functions what did we have? That if I evaluate, so this is f of lambda x plus 1 minus lambda y.

For convex functions what should this look like? So less than or equal to lambda of f of x plus 1 minus lambda times f of y right and this is true for convex functions. So, this inequality which is less than or equal to I mean it does not eliminate the fact that you can have multiple local minimizers all of them which are I mean also going to be global minimizers, but I mean you can still have multiple optimal solutions right with strictly convex function as a term suggests its strict right. So, that means this inequality is also going to be strict. So, for strictly convex functions we have f of lambda x plus 1 minus lambda y. this is strictly less than lambda of f of x plus 1 minus lambda f of y this is true for every lambda in open interval 0 to 1 and x not equal to y and this is the definition of strictly convex function or strict convexity ok.

$$f(\lambda x+(1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$$

[Convex functions]

$$f(\lambda x+(1-\lambda)y) < \lambda f(x)+(1-\lambda) f(y) \ , \ \forall \lambda \in (0,1)$$

$$x \neq y$$

[strictly convex]

So, what does this strict inequality ensure? This ensures that every time you change a point. So, one of the things that we notice with convex functions is at least geometrically what is happening is suppose you have a function which looks something like this. So, for any points x and y, we know that the function value, like a convex combination of it, it is always below or at least not more than the convex combination of the function values f of x and f of y right. But because of this less than equal to this may still be equal right and that is when you have these flat regions in your functions. With this strict inequality we are guaranteed that this is going to be strictly smaller.

So, this difference is going to be non zero. right and that is why strictly I mean with strictly convex functions and if you have this difference to be non-zero one of the things that you can clearly notice from this picture is or this plot is that it is guaranteed to have a unique minimizer because every time as you as you move you are going to be changing the function value. So, you are always going to be evaluating something which is going to be smaller than the convex combination of f of x and f of y and one of the implications of straight convexity is if the minimizer exists minimizer exists it will be it is unique right. So, this is one of this is the implication of working with strictly convex functions ok all right. So, how do we identify if a particular function is strictly convex? So, sufficient condition I mean one way is to basically validate this particular definition of straight convexity, but like in the like with convex functions we had a second order condition for convexity right where the Hessian is positive semi definite.

So, can we have an equivalent sort of condition? So, sufficient condition for straight convexity is. So, let us assume f is twice differentiable ok and h n of f is positive definite ok not positive semi definite, but positive definite then f is strictly convex. So this is to say like in the scalar case, that means f double prime is greater than 0. Then you have for every x, then you have f to be strictly convex. So what about, let us say if I look at a function like x to the 4, what is the Hessian of x to the 4? So Hessian of f is, so this is let us say f of x.

So let us say if I choose, first of all is this strictly convex? Is x to the 4 strictly convex? Does it have a unique minimizer, right? So x to the 4, this is strictly convex, okay. What about the Hessian of f, this particular function? will be x square is this greater than 0 is this not always right. So, this is greater than equal to 0 not strictly greater than 0 and that is why we call it a sufficient condition. So, if this condition holds then you are guaranteed that the function is going to be strictly convex, but if the condition does not hold then you cannot guarantee otherwise that it is not going to be strictly convex right. So, it is a sufficient condition not a necessary condition ok.

eg: $f(x) = x^4 \longrightarrow$ strictly convex

$\nabla^2 f = 12x^2 \geq 0$

$f(x) = x^2$

$\nabla^2 f = 2 > 0$

But x to the 4 it is strictly convex because it almost looks like this where I mean you have a unique minimizer here. Is this clear? What about function like x square? Is this strictly convex? Right? Function is strictly convex and what about the Hessian of this function? 2 which is strictly greater than 0. So, I mean this is I mean we can using the second order condition we can quickly validate that this function is strictly convex. So, both these functions are strictly convex, but because this is a sufficient condition we can only sort of implicate one way. So, if the if f if the Hessian of f is positive definite then function is going to be strictly convex, but the other way around need not be true ok.

Is every strictly convex function a convex function? Just by definition every strictly convex function is also a convex function. So, a strict convexity implies convexity other way around need not be true right. If a function is convex it need not be strictly convex. So, for instance a function which looks something like this. this is a convex function, but not a strictly convex function or a constant function, it is not a strictly convex function, it is just a convex function.

And similarly, the sufficient condition, so hessian f for every x, this implies strict convexity So, the reverse implication need not be true ok, but why do we care about straight convexity? So, what is happening with convex functions? Like if I look at simple I mean we have multiple minima which is fine I mean because every minima is also a global minima. you can see that something that happens with these kind of functions that the gradients start vanishing right for this to happen you have very small gradients around this particular value right and the gradient start vanishing and think of it from from the point of view of designing an optimization algorithm where you're using the gradient information so if you're updating your x based on the past x and the current like some step size and the current gradient value So, you are making smaller and smaller steps towards your optimal solution right whereas from the point of your strictly convex function I mean you do not really have that kind of condition where I mean you may still have larger gradient values around the optimal solution and you can converge to the optimal solution faster. So, studying these class of functions has its own significance. the other other thing is if hessian of f turns out to be positive definite I mean it need not be the case for all strictly convex function, but let us say for a particular strictly convex function hessian of f is positive definite then that means you can also invert that hessian right. that that is also invertible and Hessian inverse is something as we are going to look at it later, we can use Newton type methods to get even faster convergence.

So, when Hessians are going to be invertible. So, a strictly convex functions in that sense has have a significant role to play. So, in some sense convex functions are more general strict convex I mean a strict convexity in some sense is a sort of stronger assumption on the function. There is even more like stronger assumption which is called strongly convex function or strong convexity. So, let us look at it.

ok what is so so far for a for a simple convex function we know that f of lambda x plus 1 minus lambda y this is less than equal to lambda times f of x. So, this is for a convex function for a strictly convex function this basically holds with strict inequality as long as lambda is between 0 and 1 and x and y are different points this becomes strictly convex function. So, what you are saying is that the function value. So, the function evaluates to a smaller value than the convex combination of the function values, but it does not tell you by how much. So, in a strongly convex function you also provide values.

So, it is strongly convex would look something like this. So first of all this function is going to, so this is your convexity part. Additionally you can guarantee that, so in convexity you have this right hand side. So this particular right hand side is going to exceed the left hand side at least by this much amount. So there exists some mu greater than 0 and this mu is called modulus of strong convexity.

So, for strictly convex function we say that the right hand side exceeds the left hand side. For strongly convex function we also provide the amount by which the right hand side exceeds the left hand side. So, we say that basically it exceeds by this much amount ok. So, that is sort of more restrictive assumption on the kind of function that you can have where you also say that no matter where what your x and y are there exists some mu greater than 0 such that the right hand side is going to exceed the left hand side by this much amount and as y is very close to x let us say y is equal to x then this quantity is anyway 0, but then as y is close to x I mean basically the amount by which it exceeds is going to get smaller, but you specify a rate by which like the right hand side exceeds the left hand side. So that means when you are doing gradient descent at you will always have some definitive rate coming from this particular this additional term whereas that may not be the case with with your convex function or even strictly convex function right.

* Strongly Convex functions:

$$f(\lambda x+(1-\lambda)y) \leq \lambda f(x)+(1-\lambda) f(y) \quad \text{[Convex]}$$

$$f(\lambda x+(1-\lambda)y) < \lambda f(x)+(1-\lambda) f(y) \quad \text{[Strictly Convex]}$$

$$f(\lambda x+(1-\lambda)y) \leq \lambda f(x)+(1-\lambda)f(y) - \frac{\lambda(1-\lambda)\mu}{2}\|x-y\|^2$$

$\mu > 0$

↳ Modulus of Strong Convexity

So, one of the things that we saw at the beginning of the this particular course is optimizing x to the 4 versus x optimizing x square right with x square. So, x square basically is a strict is strongly convex function it is not just strictly convex it is also strongly convex ok. So, whereas x to the 4 is just strictly convex So, we know that closer to the optimal solution with x to the 4 it sort of varies slowly whereas, x to x square it I mean it still has a definitive gradient and therefore, those kind of functions we can optimize much faster and that is why it is called strongly convex function strongly because you are making stronger assumptions on the convexity of the function ok. So, in terms of implication we have strong convexity which implies strict convexity which implies convexity ok. So, this is how the implication goes.

Is this clear? So, an example would be x square. So, a second order necessary and sufficient condition was strongly convex functions is. So, assume that f is twice differentiable then let me write it this way, f is mu strongly convex. So, from now on I am I am going to denote strong convexity with SC ok. So, f is mu strongly convex if and only if h in of f looks something like this.

Yeah, twice continuously differentiable you can I mean in this case you probably do not need the continuity as long as it is twice differentiable, but that is fine. Usually with first order definition you need the continuity because you write f of y in terms of f of x in, but yeah. So, this is if and only if kind of condition that f is mu strongly convex if and only if hessian of f is. sort of not just positive definite, but it is basically lower bounded by this. So, you can see the model of strong convexity also appearing in over here ok.

So, if f of x is x square, what is mu in this case? 2 right. So, f of x is strongly convex with mu equal to 2. Yeah . because if you look at this condition right.

$f(x) = x^2$, $\Rightarrow$ $f(x)$ is sc with $\mu = 2$.

So, that means hessian of f is strictly greater than 0 because it is lower bounded by mu greater than 0 right. So, then if you if I use this second order sufficient condition then it is ok. No this is if and only if. So, we had the first order condition for convexity and so just to briefly recall. So, what was the first order condition for convexity? f of y is greater than equal to f of x transpose y minus x right and pictorially speaking what did that imply suppose you have a function which looks something like this and if I compute a gradient at one particular point let us call it x and this is another point y take this point is y.

1st order condn for convexity:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x)$$

So, this is f of y So, essentially the function in some sense the function would have moved upwards, if I keep moving in the direction of the gradient the function would have moved upwards right. So, this is what this particular condition captures. In case of strong convexity, so we have an equivalent definition of first like first order characterization of a strongly convex function. So, first order condition for a strong convexity So, here we say that the function not only moves upward, but it moves by this much amount, at least by this much amount. So, this gap is going to be at least this much.

So, in some sense you are trying to sort of impose a condition on the gradients that the gradients are not going to be like vanishingly small, because you all the time you are specifying that you are going to be moving upwards by at least this much. Is this clear?

So, this is pretty much the gap and it says that like for instance if this is this particular first order condition holds with equality. So, that means the function is moving like if I move in the direction of the tangent the function is pretty much moving along that line right. Whereas, here I mean the function is moving upwards. So, we are getting larger and larger gradients every time and not only it moves upwards, but it moves upwards at least by this much amount.

and this basically helps when we are closer to the optimal solution. So, we know that there is always going to be gradient because of this particular condition like non vanishing kind of gradient because of this condition ok. So, all these definitions are the first order condition for convexity I mean. So, these make sense when f is continuously differentiable let us say right, but what if f is not differentiable what do we do? if the gradient of f does not exist can we still use these first order conditions for convexity or strong convexity? The answer is yes and so let me just make a quick remark. So, when f is not differentiable, so then gradient of f need not exist right.

1st order condition for SC:

$$f(y) \geq f(x) + \nabla f(x)^T (y-x) + \frac{\mu}{2}|x-y|^2$$

Remark: When f is not differentiable, we use "Subgradients."

So, we use something called subgradients. So, we will look at this sub gradient part later in the course, but just to tell you what subgradients are. So, consider a function which looks. So, basically f of x is equal to mod x right. So what is the gradient for any x greater than 0 1 right for x less than 0 the gradient is negative 1 and what is the gradient at x equal to 0 not defined right.

But when we can define something called sub gradient sub gradient would be any line. such that it is sort of is below the curve right. So, again one of the definitions of convexity is if we move along that particular line the function would have moved upwards right. So, sub gradient for this function is defined to be plus 1 if x is greater than 0 minus 1 if x is less than 0 and any number if x is equal to 0 right. and we can write down the equivalent definition as f y is greater than equal to f of x plus g transpose y minus x.

Instead of using the gradient of f of x transpose y minus x, we can define the equivalent like we can write down the equivalent sort of characterization of convexity in terms of something in terms of subgradients. And just to verify this for x equal to 0, so for x equal to 0, f y is greater than equal to f 0 is simply 0 g transpose y right. In this case g is just a scalar and as long as let us say y is positive and if I choose g to be between minus 1 and 1 then we know that this function sort of exceeds for positive y and so on right. So, this constraint is satisfied for example, for this particular choice of sub gradient. So, whenever gradients are not defined we we write down the equivalent characterization because the function is convex.

So, it must follow some kind of first order convexity condition and when gradient is not defined we use we basically define it using subgradients. And we can use the same sub gradient even like same sub gradient definition even for strongly convex functions. Is this clear? Any questions so far? Yeah. x to the 4 right. So, the question is there a function which is strictly convex, but not strongly convex and that is x to the 4.

If you do if you if I if you take the Hessian of x to the 4 that is 12 x square and you cannot lower bounded by some mu some mu right like because it will vary with x. So, Whenever I say strongly convex, the picture that you should have in your mind is something like x square, something equivalent of x square. I mean there are other functions which are also strongly convex, but x square is the most commonly sort of used strongly convex function. Something which is strictly convex, but not as strongly convex would be x to the 4 and anything which is like I mean all these functions are anyway going to be convex. So, we are going to look at a few implications of strong convexity and how they sort of.

So, one of the things that strongly convex functions satisfy is by the way do strongly convex functions have unique minimizer. Yes right, because strongly convex functions are by definition also strictly convex and strictly convex functions have unique minimizers. So, they would also have unique minimizers. So, strongly convex functions satisfy something called PL Inequality.

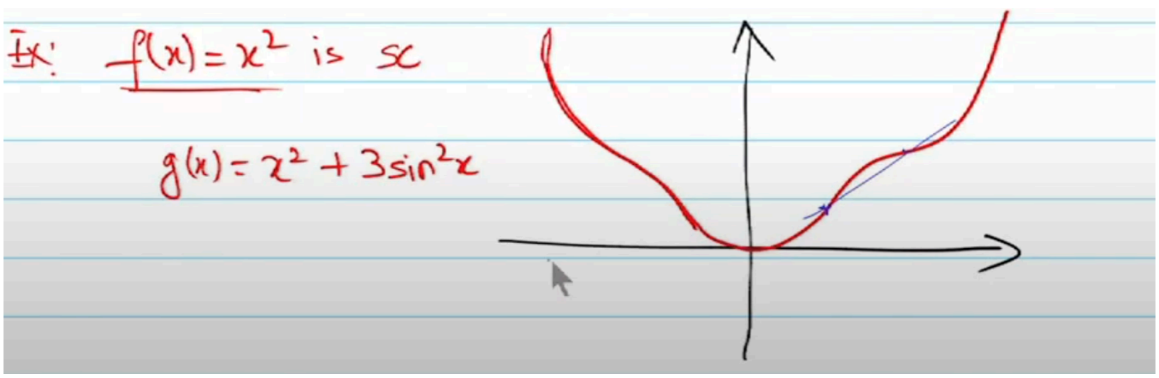
PL-inequality:

$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu\left(f(x) - f_*\right)$$

optimal value of the function

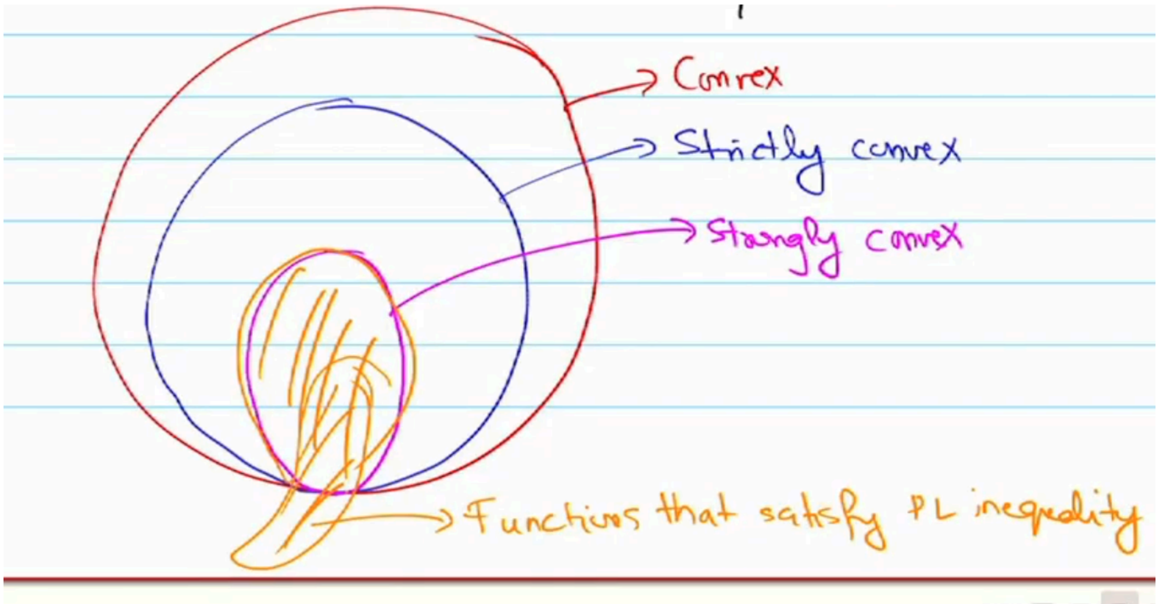
So, PL stands for Polyak Lojasiewicz Inequality. So, this condition is also called gradient dominance condition and this PL inequality where f star is the optimal value of the function. In some sense what this definition is trying to say is that function at least has a quadratic ; quadratic type of growth and the reason why PL inequality is particularly important is now if you remember that in the first lecture we looked at a function of the form. So, we know that f of x equal to half x square or x square this is strongly convex right and you can verify this definition is I mean trivially satisfied right. So, gradient of f of x is 2 x right.

So, that will be 4 x square divided by 2. So, 2 x square is going to dominate and in this case mu is also 2. So, I mean in fact this is satisfied with kind of equality, but there is a class. So, if I look at a function of this form g of x is x square plus 3 sin square x, the plot of this function looks something like this. something like this. So, is this strongly convex? Is this function strongly convex? It is not even convex right, I can always I can draw a line and the function sort of evaluates to a value which is more than the convex combination of the function like evaluations at these points right.



So, this function is not even convex to start with forget about it being strongly convex, but this function still satisfies this PL inequality. right and so in terms of hierarchy, so every we know that the implication goes this way right. So, strongly convex functions, strong convexity implies straight convexity implies convexity. So, you have a larger sort of let us say this is all convex functions. Within this class you have functions that satisfy are strictly convex, then within this class we have functions which are strongly convex.

and not within this class entirely, but you have another class of functions which satisfy PL Inequality which also includes strongly convex function, but so these are the functions that satisfy PL Inequality. So, this is not a convex function, but it still has a unique minimizer right. So, this for instance this particular function is very different from a function which looks something like this. which is which we know that has multiple local minima.



In this case you still have a unique minimizer. So, near the optimal solution it behaves like a strongly convex function, but far from the optimal solution it does not behave like an strongly convex function. So, this function these kind of functions are called Invex functions. The PL inequality here. No, but then there are functions which are not even strongly convex, but also satisfy peer inequality, right.

Yeah, yeah, you can extend it this way. Thanks, thanks for pointing this out. Yeah. Okay. All right. So, we are going to look at few implications of strong convexity and yeah.

Invex function yes I mean, but then you can also you can generalize this more and you can say it is going to be even like I mean it is also going to be convex. In this case I mean it is it is not convex right. Yes. I mean they have their own like I mean there is a particular definition for in mathematical definition for invex function. What I just wanted to I mean without getting into the details what I just wanted you to look at is or.

Eg: $\frac{1}{2}\|x\|^2 \rightarrow$ is SC.

$\frac{1}{2}\|Ax-b\|^2$ is not SC. always.

$\hookrightarrow$ when A is full-row-rank, then this function is SC.

So, this function is not even convex. Yeah. And and in more more often than not you will come across functions. So, for instance when you. So, something like for instance. So, this this function is strongly convex right, but then if I do an affine transformation of this you can show that this is not strongly convex for any a. So, only when a is full row rank then this function is strongly convex.

but you can show that this function still satisfies the PL inequality. So, by studying the class of functions which satisfy PL inequality you are I mean and if you can guarantee let us say you are designing an optimization algorithm and you say my optimization algorithm works for functions that satisfy PL inequality. So, by being able to do so you know for sure that if your function if your optimization algorithm works for strongly functions that satisfy PL inequality it will work by definition for function that satisfy strongly like that are strongly convex, but that are also potentially not convex something like this. So, this function is not convex to start with, but you can still guarantee convergence with the optimal solution for this class of functions. So, and we are in this course we are going to design algorithms and guarantee that those algorithms would converge faster for functions that satisfy PL inequality alone without even without those functions even being convex to start with. Thank you.