Distributed Optimization and Machine Learning

Prof. Mayank Baranwal

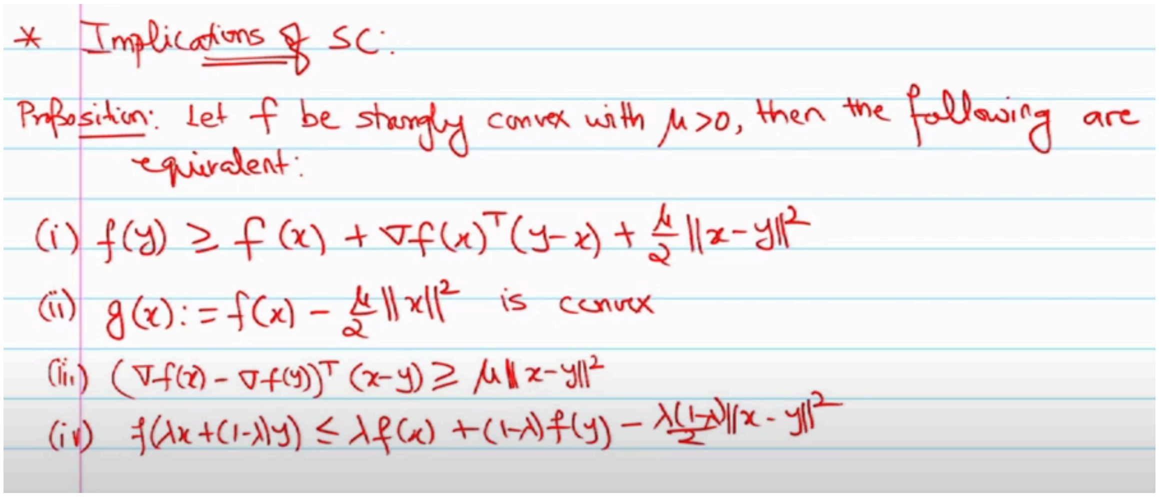
Computer Science & Engineering, Electrical Engineering, Mathematics

Indian Institute of Technology Bombay

Week-2

Lecture - 7: Implications of strong convexity

So, let us now look at implications few more implications of strong convexity. So, let f be strongly convex with mu greater than zero. So, let us look at this proposition, then the following are equivalent. The first condition is the usual definition of strong convexity which is f of y greater than f of x plus gradient of f of x transpose u minus x plus mu over 2 norm x minus y square. So, if I consider g of function g of x defined to be f of x minus mu over 2 norm x square. So, if f is strongly convex then g of x is going to be convex.



* Implications of SC:

Proposition: Let f be strongly convex with $\mu > 0$, then the following are equivalent:

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

(ii) $g(x) := f(x) - \frac{\mu}{2}\|x\|^2$ is convex

(iii) $(\nabla f(x) - \nabla f(y))^T (x-y) \geq \mu \|x-y\|^2$

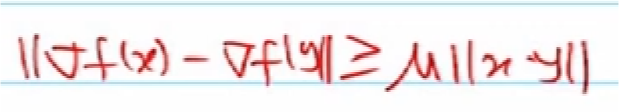
(iv) $f(\lambda x + (1-\lambda) y) \leq \lambda f(x) + (1-\lambda) f(y) - \frac{\lambda(1-\lambda)}{2}\|x-y\|^2$

So, this is the third sort of equivalent definition is gradient of f of x minus gradient of f of y transpose y minus x is greater than mu norm x minus y square. and the fourth equivalent definition is f of lambda x plus one minus lambda y is less than or equal to lambda times f of x plus one minus lambda times f of y minus lambda times one minus lambda mu by two norm x minus y square. So, let us let us revisit the statement. So, let us say the function f is strongly convex with modulus mu greater than 0, then all these conditions are equivalent.

So, all these are equivalent definitions of strong convexity. I mean this is something that we already know and what I mean what this first sort of characterization says is that the function sort of the difference between the I mean when we write the first order condition the difference I mean in in some sense the function sort of moves upward by at least this much amount right that is the first condition. The second condition is if you have an f like a function which is strongly convex and even if you subtract a quadratic kind of function from that basically subtract mu over 2 norm x square that function even though it need not be strongly convex it is going to be convex at least. So, when we think of strong convexity that means the function has a like always as I said always think of a quadratic function when you think of strong convexity. So, even you if you subtract this quadratic type of function from this original function you are still going to be left with a convex function.

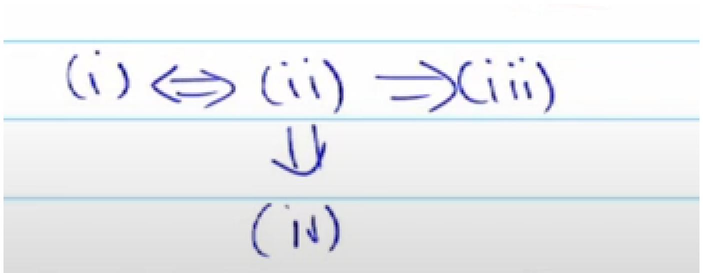
So, that is the second sort of yeah. Strictly convex what? I mean that need not be true. I mean it is. So, all we know that the Hessian of f is greater than equal to mu times identity right and for the other one it is also going to be mu times identity. So, we cannot say it is strictly greater than 0 right.

So, had this been strictly greater than mu times identity over here then you could have said that. So, we are going to prove this as well, but the third condition what did this says is suppose I apply Cauchy Schwarz on this right on the third condition. So, A transpose B is going to be less than or equal to norm A times norm B that is Cauchy Schwarz and using Cauchy Schwarz. and one of the norms would be norm of x minus y. So, what we say is that difference between the norm of the great difference of the gradients is going to be greater than equal to norm of the like basically the difference of x basically x and y right.



$$\|\nabla f(x) - \nabla f(y)\| \geq \mu \|x - y\|$$

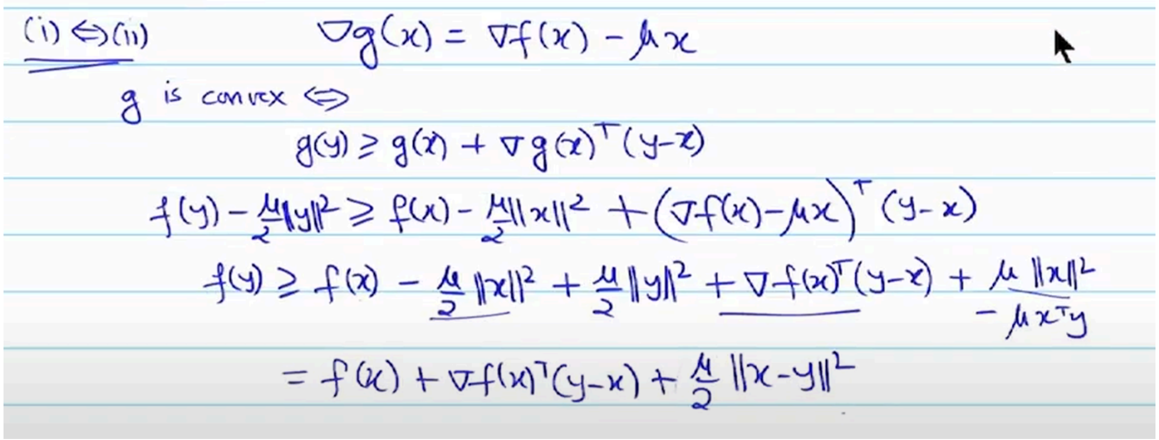
So, an implication of this would be right. So, this is this is one of the conditions which is I mean basically the implication of this third one and fourth one is again something that we had looked at where the right-hand side we know that exceeds the left-hand side at least by this much amount. So, all these are equivalent definitions and we are going to prove this. So, if we if we need to show that these are equivalent sort of proof strategy would imply that if let us say let us say I show that 1 is equivalent like to 2 and 2 implies 3. and 2 also implies 4, then we are done right because we have shown that all of these are going to be I mean you or I think one thing that you possibly need to show is that 4 is also equivalent to 1, but I will probably leave this as an exercise, but then you will be done right.



$$\begin{tikzcd}
(i) \arrow[r, Leftrightarrow] & (ii) \arrow[r, Rightarrow] \arrow[d, Rightarrow] & (iii) \\
 & (iv) &
\end{tikzcd}$$

So, this particular result is going to be important in fact you will see that in most of the proofs. So, let us try and show that 1 is equivalent to 2. So, what is the gradient of G here? what is gradient of g of x gradient of f minus mu x right. So, g is convex you have g of y is greater than or equal to g of x plus gradient of g of x transpose times y minus x ok. So, g of y by definition is f y minus mu over 2 norm y square g of x is f of x minus mu over 2 norm x square and the gradient is gradient of f of x minus mu x transpose times y minus x.

So, we are we are almost there. So, what we have received is what we have now is f of y is greater than f of x minus mu over 2 norm x square plus mu over 2 norm y square plus gradient of f of x transpose times y minus x and then you have minus minus plus mu times norm x square. So, plus mu minus mu by 2 gives you again plus mu by 2 right and then you can do the square completion and this turns out to be f of x plus this term minus sorry plus mu over 2 times norm x minus y square and this is the first sort of characterization of the strongly convex function. Is this clear? So, again minus mu by 2 x norm x square plus mu x norm x square this gives you plus mu by 2 you take mu by 2 common. So, that this norm x square plus norm y square and then you have wait I think there is another yeah there you have another term here sorry my bad minus mu x transpose y.



$\underline{(i) \Leftrightarrow (ii)}$ $\qquad \nabla g(x) = \nabla f(x) - \mu x$

g is convex $\Leftrightarrow$

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

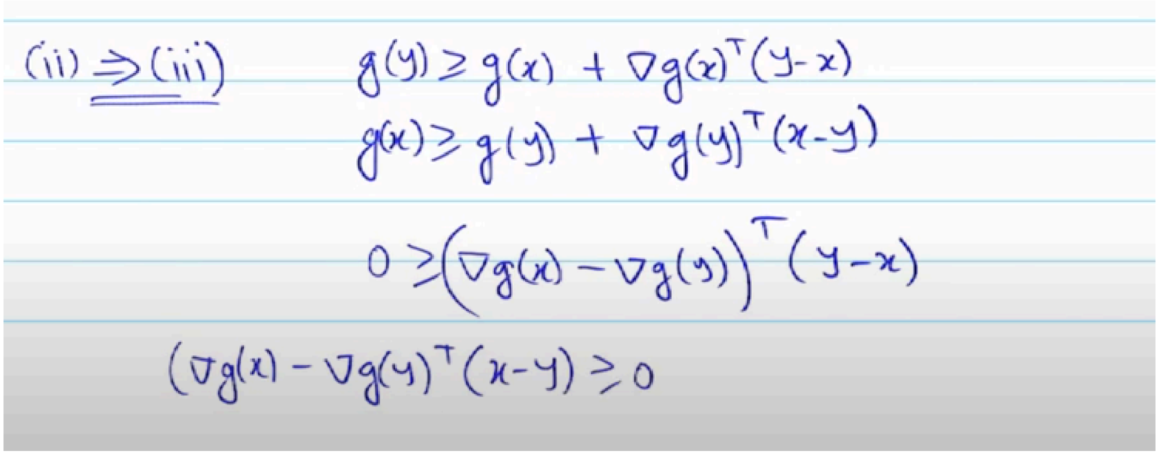
$$f(y) - \frac{\mu}{2}\|y\|^2 \geq f(x) - \frac{\mu}{2}\|x\|^2 + \left(\nabla f(x) - \mu x\right)^T (y-x)$$

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

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

So, if I expand this you have norm x square plus norm y square minus 2 times norm x transpose y and this is any questions on this right. So, if g is convex if and only if this is true. So, g is g is convex if and only if f is strongly convex or this 2 is true if and only if x is 1 is true ok right? So, how do we show 3? say we want to show that 2 implies 3. Again this gives you an idea as to how we can sort of and when you want to prove the results of this form what kind of simple tricks that you can use. So, remember like here we have to collect terms of the form gradient f of x and gradient f of y and from if I want to go from 2 to 3.

then I probably need a gradient of g as well right. So, that means I would have to use the condition for convexity for g first order condition for convexity of g we know that g is convex. So, if g is convex then g y is greater than equal to g of x plus gradient of g at x transpose times y minus x ok. I can equivalently write this as g of x greater than equal to g of y plus gradient g y transpose x minus y right because I mean x and y I can simply interchange and if I add these two equations now this basically gives me 0 greater than equal to gradient g x minus gradient g y transpose times y minus x and I mean taking this to the left-hand side this basically tells you that gradient g of x minus gradient g of y transpose x minus y is greater than equal to 0 and you just use a gradient definition of g of x and you get this inequality, ok. And the final thing is something that I mean we need to show this particular result, again it follows sort of naturally from like if we use this.



(ii) $\Rightarrow$ (iii)

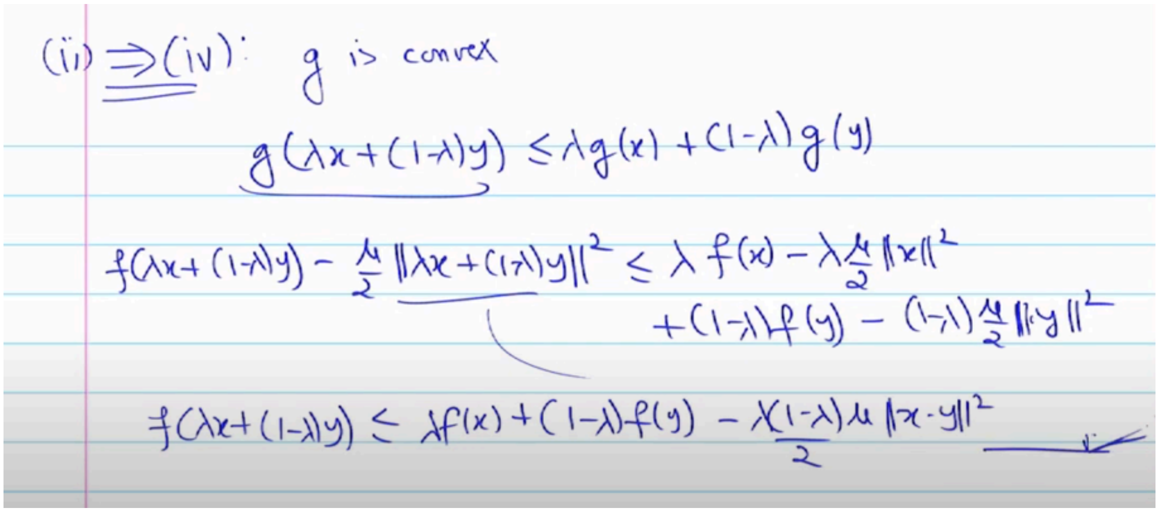
$$g(y) \geq g(x) + \nabla g(x)^T (y-x)$$

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

$$0 \geq \left(\nabla g(x) - \nabla g(y)\right)^T (y-x)$$

$$(\nabla g(x) - \nabla g(y)^T (x-y) \geq 0$$

So, this particular result is often used in proving other results, and more than anything this result is often used in proving other results. And the way you can if you want to show that 2 implies 4, we basically need to use the other definition of convexity which is f g of lambda like if g is convex then g of lambda x plus 1 minus lambda y is less than equal to lambda of g of x plus 1 minus lambda of g of y right. And by definition, this is nothing, but f of lambda x plus one minus lambda of y minus mu over 2 norm lambda x plus one minus lambda of y square And again if I expand it this in terms of definition of square, if you take this term to the right-hand side and do some simple algebraic manipulation, you can simply show that this is less than equal to lambda f of x. plus one minus lambda f of y minus lambda times one minus lambda times mu by two norm x minus y square. So, I mean so we are done ok.



(ii) $\Rightarrow$ (iv): g is convex

$$\underline{g(\lambda x+(1-\lambda)y)} \le \lambda g(x)+(1-\lambda)g(y)$$

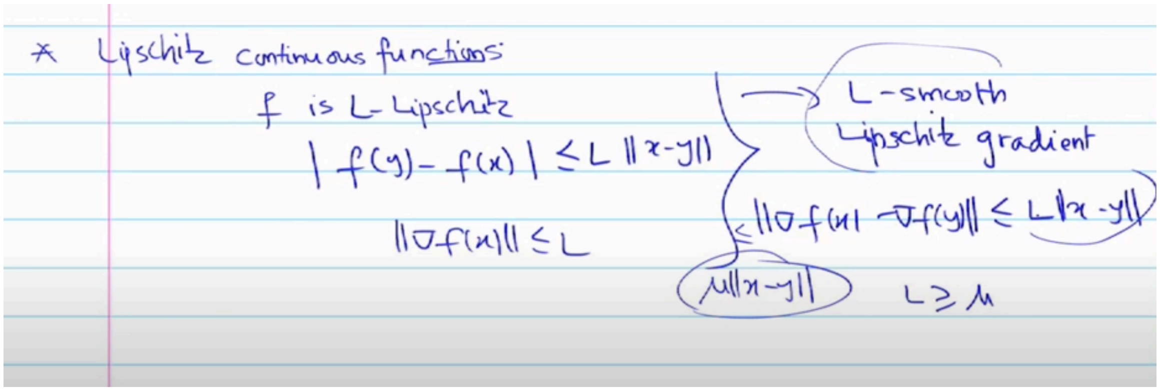
$$f(\lambda x+(1-\lambda)y)-\frac{\mu}{2}\underline{\|\lambda x+(1-\lambda)y\|^2} \le \lambda f(x)-\lambda\frac{\mu}{2}\|x\|^2 + (1-\lambda)f(y)-(1-\lambda)\frac{\mu}{2}\|y\|^2$$

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

So, as I said like you can so this particular result this particular definition of strong convexity it is often used in proving other results about strong convexity. Any questions on this? this to the here yeah Lipschitz yeah yeah yeah not should not be I mean you can still upper bounded by some L, but you can also Lipschitz continuity is about the upper bound on the gradient or upper yeah this is about the lower bound only. Well, Lipschitz continuity is not upper like so again we really need to specify what Lipschitz continuity is defined on. So does everyone know what Lipschitz continuous functions are? So usually we define Lipschitz continuous function So, we say that f is L Lipschitz if mode of f of x minus f of y is less than or equal to L times x minus y or y minus however you want to write it. This is the definition of Lipschitz continuity of a function right.

So, in some sense what you want to say is that the norm of the gradient is upper bounded by L. we also define something called L Lipschitzness of the gradient and not of the function itself. So we say that we call this L smooth or functions with Lipschitz gradient and in fact we are going to look at this condition as well which says that norm of the gradient of f of x minus f of y that is less than equal to L times x minus y. So, in this case the function is not Lipschitz continuous, it is a gradient which is Lipschitz continuous. The gradient of the function which is Lipschitz continuous.

So, such functions are called L smooth functions. And if you look at strong convexity, this is the opposite of L smoothness. So, in some sense you say that it is upper bounded by this, but it is also lower bounded by mu x minus y. So, it is kind of sandwiched between the two. And if you look at a function like x square, and mu have turned out to be the same.



* Lipschitz continuous functions:

f is L-Lipschitz

$|f(y) - f(x)| \leq L \|x-y\|$

$\|\nabla f(x)\| \leq L$

$\to$ L-smooth

Lipschitz gradient

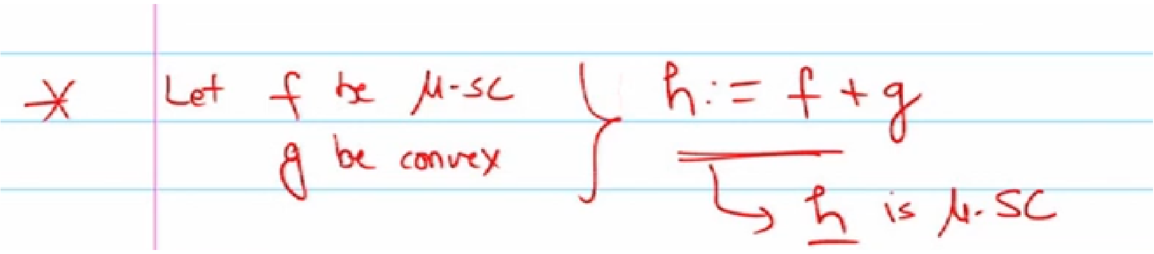
$\mu\|x-y\| \leq \|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\|$

$L \geq \mu$

So, if you if the function is l smooth and mu strongly convex we know for sure that l is greater than equal to mu that is one of the implications of this thing right. Because we know that this is lower bounded by from the definition of strong convexity this is lower bounded by this from Lipschitz gradient or L smoothness it is upper bounded by this. So, L has to be greater than equal to mu. For quadratic functions L is exactly equal to mu, but in general, L is always greater than equal to mu. and this is what and if you look at x square, it is both L smooth and mu are strongly convex with L and mu both being equal to 2, right.

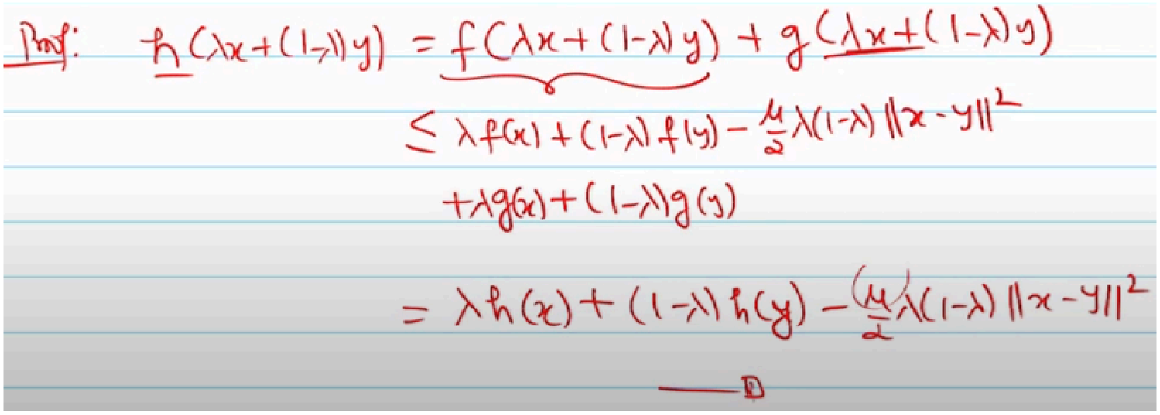
So, that is the distinction between Lipschitz smoothness and strong convexity. Is everyone able to follow? So, as I said one of the strongly convex functions I mean other than the different mathematical characterization that we looked at for strongly convex functions. One of the things that strongly convex functions are really useful in is this particular fact. So, let us say you have two functions let f be mu strongly convex ok and another function is g which is simply convex. So, you have a strongly convex function f and a convex function g and you define a new function h which is basically the sum of these two functions f and g.

So, what you can say is that not only h is strongly convex, in fact h is mu strongly convex. So, suppose I want to optimize a function g. and in some way I am able to come up with a function f, which seems to share the same optimal solutions as g, but behaves like a strongly convex function, then I would rather work with this function h, right and which because this function is going to behave like in a strongly convex function. So, again the idea is, so I mean a quick simple proof. So, we want to show that h is also mu strongly convex, we know that f is mu strongly convex and g is convex.



$$\left.\begin{array}{l}* \quad \text{Let } f \text{ be } \mu\text{-SC} \\ \qquad g \text{ be convex}\end{array}\right\} \; h := f + g \;\; \hookrightarrow \underline{h} \text{ is } \mu\text{-SC}$$

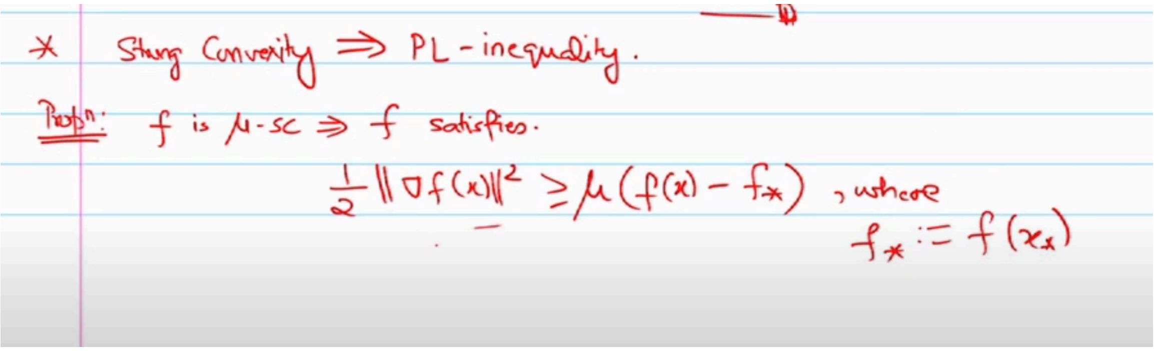
So, what is h of lambda x plus 1 minus lambda y? By definition this is f of lambda x plus 1 minus lambda y plus g of lambda x plus 1 minus lambda y right and we know that f is strongly convex. So, this would be because if it is mu strongly convex this would be less than equal to lambda f x plus one minus lambda f y minus mu over 2 lambda 1 minus lambda norm x minus y square, and because g is convex we know that this particular term is less than equal to lambda g of x plus 1 minus lambda g of y. I can add these numbers add these functions lambda f x plus lambda g x which is equivalent to which is equal to lambda f h x. So, that means this thing is less than equal to your left-hand side is less than equal to lambda of h of x plus 1 minus 1 minus lambda h of y minus mu over 2 lambda 1 minus lambda norm x minus y square. and this is the definition of strong convexity for h.



$$\begin{aligned} \underline{\text{Prof}}: \quad \underline{h}(\lambda x+(1-\lambda)y) &= \underbrace{f(\lambda x+(1-\lambda)y)} + g\underline{(\lambda x+}(1-\lambda)y) \\ &\le \lambda f(x) + (1-\lambda)f(y) - \frac{\mu}{2}\lambda(1-\lambda)\|x-y\|^2 \\ &\quad + \lambda g(x) + (1-\lambda)g(y) \\ &= \lambda h(x) + (1-\lambda)h(y) - \left(\frac{\mu}{2}\right)\lambda(1-\lambda)\|x-y\|^2 \\ &\qquad\qquad — \square \end{aligned}$$

So, not only h is strongly convex in fact h is mu strongly convex ok. So, that is that completes the proof. So, one of the implications of strongly convex function if you add a basically if to a convex function if you add a mu strongly convex function the resultant function is also mu strongly convex and that also kind of proves this particular part right. We know that this function is mu strongly convex If g is g of x is convex or if g of x is convex. So, mu strongly convex plus convex would imply that f of x is also strongly convex ok.

So, that is one. Yes. Yeah, yeah. So, yeah 2 mu is strongly convex yes. So, another implication something that we already discussed is that strongly convex function. So, a strong convexity implies PL inequality ok. So, another proposition. So, we assume that f is mu strongly convex, this implies f satisfies half norm gradient of f at x square is greater than or equal to mu times f of x minus f star where f star is defined to be the optimal f of x star ok.



* Strong Convexity ⇒ PL-inequality.

Propn: f is μ-sc ⇒ f satisfies.

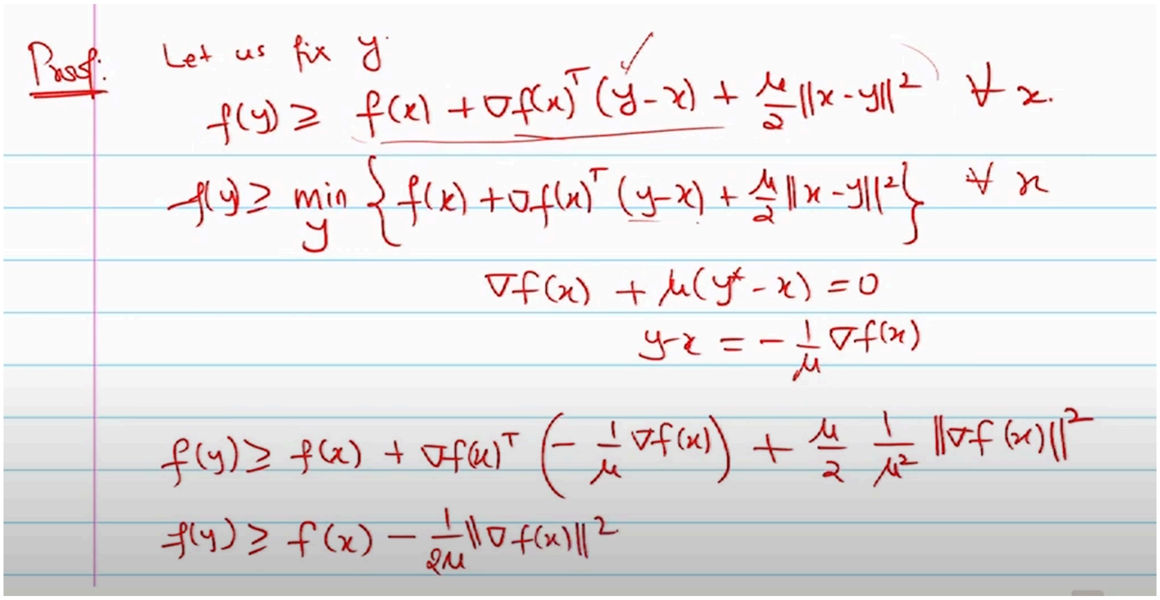
$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu(f(x) - f_*) \text{, where } f_* := f(x_*)$$

So, let us let us look take a look at the proof all right. So, again if I were to show something like this it is in terms of the gradient of f right. So, that means, we would have to use a definition of strong convexity where gradient comes in and that is we know that let us fix y right, let us fix. So, f of y we know is greater than equal to f of x plus gradient of f at x transpose times y minus x plus mu over 2 norm x minus y square for all x ok. So, we want the result in terms of gradient of f of x.

So, that is why we are using this particular definition of strong convexity. So far so good. So, this holds, so the left-hand side exceeds the right-hand side for a specific value of y. So, that means if I try to minimize the right-hand side, it would also hold true for that because right now it holds true for this particular specific choice of y. Now, if I look at all possible y's for which that right-hand side gets minimized, it would also hold true.

So, that means f y is greater than equal to minimum over y f of x plus gradient of f at x transpose times y minus x plus mu over 2 norm x minus y square. ok for all x that is true right. So, if we want to minimize it with respect to y we have to set the gradient with respect to y to be 0 and this gives us the condition the optimal y must satisfy this right. because if this needs to be minimized the optimal y must satisfy this for the right-hand side ok.

So, which basically tells you that. So, y at like the optimal solution this y minus x is nothing but. So, y minus x is negative 1 over mu gradient of f at x this thing right. So, from here if I substitute this I get f of y is greater than equal to f of x plus gradient of f of x transpose y minus x is minus 1 over mu gradient of f at x plus mu over 2 and then again minus 1 over mu square norm gradient of f at x square ok. So, this from here you get negative 1 over mu norm like norm square norm gradient of f norm square and this is 1 over 2 mu right plus 1 over 2 mu. So, basically you get f of y is greater than equal to f of x ok just adding these two terms alright or what do we have 1 over 2 mu norm or in fact you can directly argue from here.



Proof: Let us fix y.

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

$$f(y) \geq \min_y \left\{ f(x) + \nabla f(x)^T (y-x) + \frac{\mu}{2}\|x-y\|^2 \right\} \quad \forall x$$

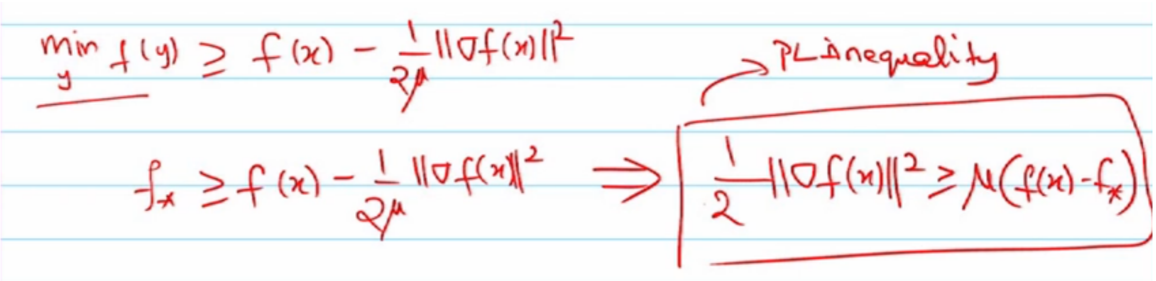
$$\nabla f(x) + \mu(y^* - x) = 0$$

$$y - x = -\frac{1}{\mu}\nabla f(x)$$

$$f(y) \geq f(x) + \nabla f(x)^T \left(-\frac{1}{\mu}\nabla f(x)\right) + \frac{\mu}{2}\frac{1}{\mu^2}\|\nabla f(x)\|^2$$

$$f(y) \geq f(x) - \frac{1}{2\mu}\|\nabla f(x)\|^2$$

So, this holds true for any y right like we had chosen as like we had fixed y in the beginning. So, this holds true for any y. So, this would also true hold true for if I try to minimize the left hand side with respect to y. So, it would hold true for that as well. because this holds true for any y and this is nothing but f star your optimal value right? So, f star is greater than equal to f of x minus 1 1 over mu 2 mu r which is your PL inequality.



$$\underline{\min_y f(y)} \geq f(x) - \frac{1}{2\mu}\|\nabla f(x)\|^2$$

PL inequality

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

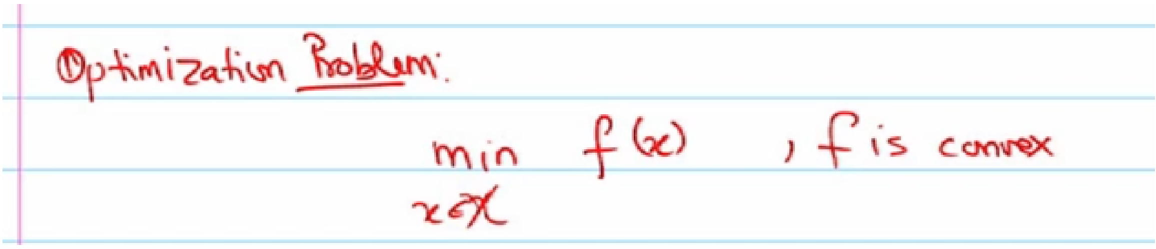
Is this clear? This one here. So, this is I am trying to characterize the value find the value of y for which this gets minimized right. So, that means I need to take the derivative with respect to y and set it to 0. So, if I take the derivative of this with respect to y, the first term is gradient of f of x, the second term is y mu times y minus x and this this needs to be set to be equal to 0 and you get this particular optimality condition for. Yeah. So, it holds for a specific y as I said like you fixed y and then you change your x right.

So, this entire inequality holds for one specific y. Yeah. But if I choose a y for which this whole thing is minimized. So, it would hold for even that one as well right.

If I am able to now let us say choose another y. We could also input that same y on the left hand side right. No. So, this inequality would be you are right. So, again the question is does I mean this inequality holds for a specific value of y yes. So, let us say I have chosen y to be equal to 5 and we know that for y equal to 5 this inequality holds true.

So, this particular term has a value for y equal to 5, but if I have chosen y something else I may even make it even smaller right and that is what I am doing. Okay so that so I am like I mean I am still keeping the right left-hand side to be fixed, but I am trying to choose another value of y which I mean which I mean so that it becomes even smaller. So, in some sense I mean if you search over like I mean this y belongs to a feasible set and if you search over a larger set you can find a better minimum right and that is what we are trying to do. Any other questions on this? and using this particular PL inequality I mean you can you can show that functions like x square plus 3 sin square x this satisfies PL inequality, but as we looked at the plot of this particular function I mean this is not even convex function to start with forget about it being strongly convex ok. This is a very special class of function as I said because lot of I mean you can provide certain convergence guarantees for these class of functions even though these functions are not strongly convex and only may I think starting on maybe it is 2016 onwards people have started to explore more on the PL inequality side because I mean whatever you can guarantee for these class of functions you can in general generalize it I mean you can generalize it for strongly convex functions for sure, but you are also covering a broader class of functions.

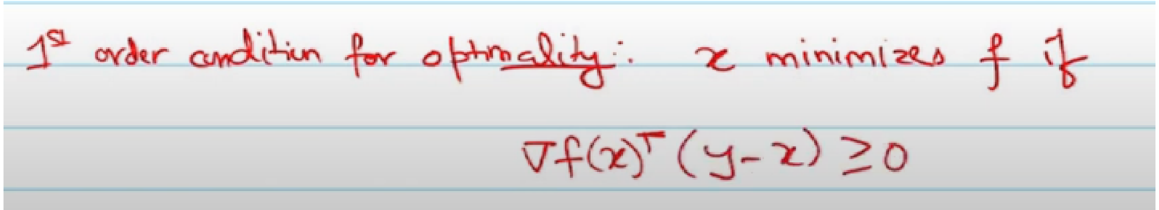
So, that way it is Alright so, so this pretty much sums up the discussion on strongly convex functions and its implications and we will revisit all these implications later when we try and design algorithms. But for the remainder of the class let us try and look at the first-order condition for optimality for convex function. So, let us revisit the mathematical optimization problem so or optimization problem. Suppose, we are looking at optimization problem of this form. Your x is some feasible set, some capital like this cursive x is some feasible set and we want to minimize this function f of x.



Optimization Problem:

$$\min_{x \in \mathcal{X}} f(x) \quad , f \text{ is convex}$$

So, what is the first-order condition for optimality? Let us also assume that f is convex. So, so let us say x minimizes f if, so what is the first order condition gradient of f at x transpose times y minus x is greater than or equal to zero ok. And so, let us look at the geometrical interpretation of this particular condition. Now, suppose let us say you have a constraint set which looks something like this, this is your x and you have a function f and let us look at the level curves of that function.



1st order condition for optimality: x minimizes f if

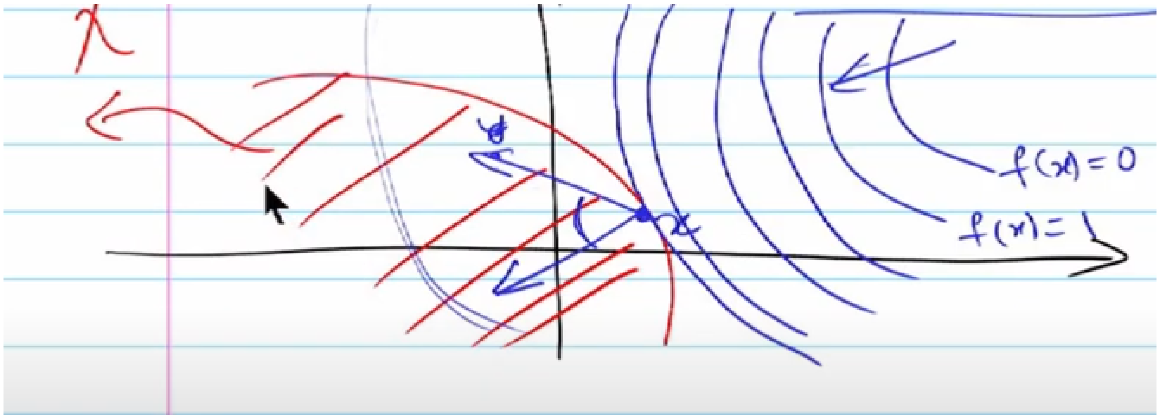
$$\nabla f(x)^T (y-x) \geq 0$$

Let us say this is f of x equal to 0, f of x equal to 1. So, just look at these level curves. So when is this function minimized subject to x belonging to this particular constraint set? When would that get minimized? When it just touches this constraint set right or like feasible solution set at this point. Let me also make this statement complete and this is true for every y in your feasible set. So what this condition, so function is increasing in this direction right.

So the function is increasing in this direction. That means the gradient of this function would be pointing in this direction. So if x is your optimal point and I choose any y in this constraint set. So this is your y. So the angle between these two vectors is less than 90 degree and that is what this particular condition is saying. So, that is the geometrical sort of interpretation of this particular statement.

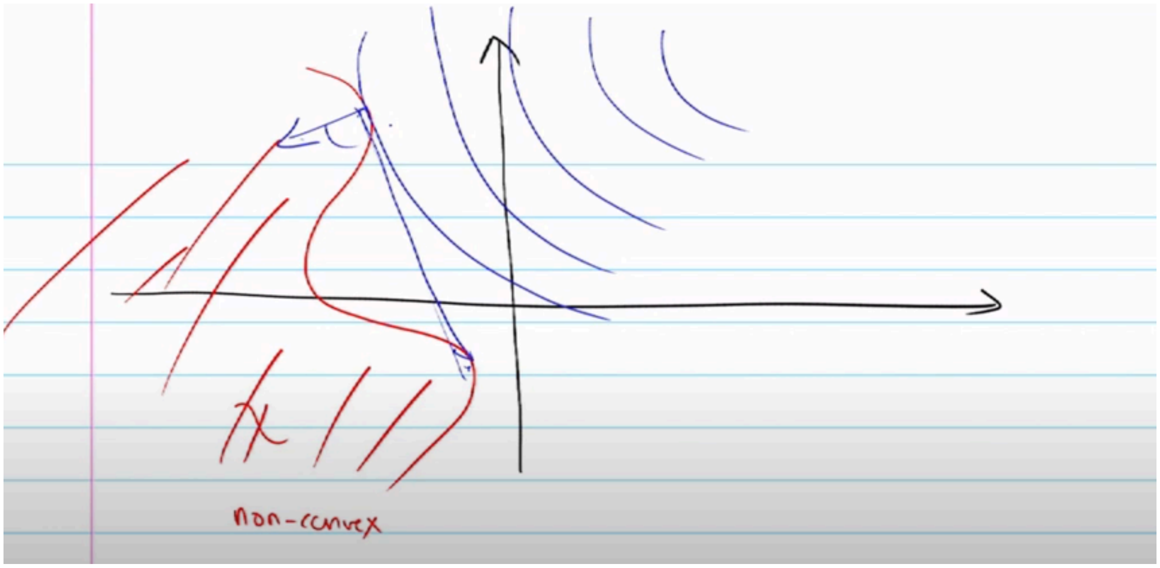
Is this clear? So, the inner product between two vectors, these two vectors is greater than equal to 0. That means the angle between them is, it is an acute angle, the angle between them is less than 90 degree and that is what this particular condition says. Yeah. So, this is the optimal solution right. So, it is when I mean when this particular level curve touches the constraint set I mean you can keep increasing even this would be I mean in this case for instance f of x is 0, 1, 2, 3 and so on.

So, you want to minimize f of x and it gets minimized when this level curve touches the constraint set and because the function is increasing in this direction. So, gradient will be pointing in this direction and you have another vector y in the constraint set. and then the angle between these two vectors is less than 90 degree less than equal to 90 degree and that is the first-order condition for optimality.



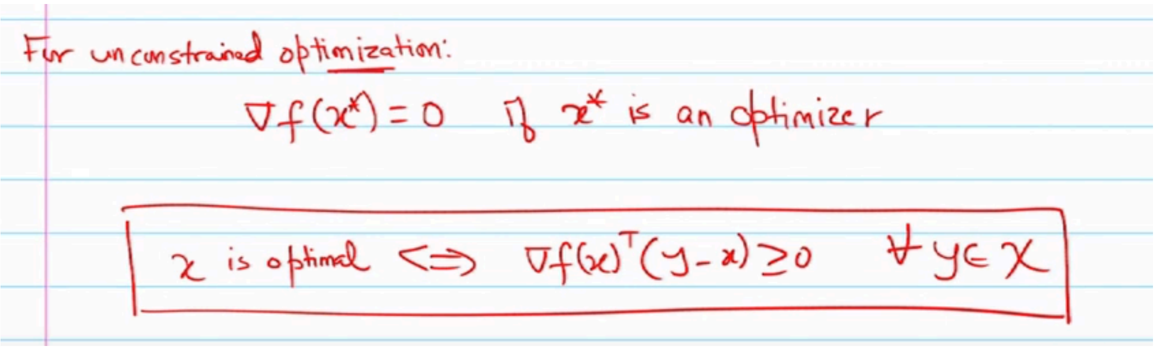
So, we are looking at our constraint set is also convex. Yes. So, we are looking at convex optimization. So, that is a good point. So, constraint set let us also make it clear that x is a convex set ok. Why convexity of x is important? Well, if I look at a similar sort of picture, let us say x is your constraint set x is not convex.

So, it may look some if some it is not convex. So, it may look something like this. So, this would be an example where the function your constraint set x is not convex. And if I look at the level curves right. So, this would be the optimal solution let us say, but if I choose and the gradient is pointing in this direction if I choose a point somewhere over here this angle can potentially be more than 90 degree right. So, this would not be the first order necessary and sufficient condition for optimality.



So, for convex functions when you are trying to minimize convex function over a convex set the first-order condition for optimality is this particular condition over here. Is this clear? So, if you are working with unconstrained optimization let us say x turns out to be this convex set x turns out to be Rn. So, what is the condition for optimality for unconstrained optimization? Gradient is 0 right and you can also see it from here right. So, if x is a whole of Rn then the vector also making positive I mean acute angle and also making an obtuse angle right I mean because you can move in every direction. and for this to be I mean if you just replicate this condition I mean you will have y minus x greater than equal to 0 and also less than equal to 0.

So, the other only way this is true is when gradient of f of x is equal to 0. So, for unconstrained optimization, the first-order condition for optimality is gradient of f of x star is equal to 0 , if x star is an optimizer. ok. Is this clear? And that is precisely because I mean the point will be strictly interior point in this set because x is whole of Rn. it will be an interior point and you will I mean basically it will make all types of angles.

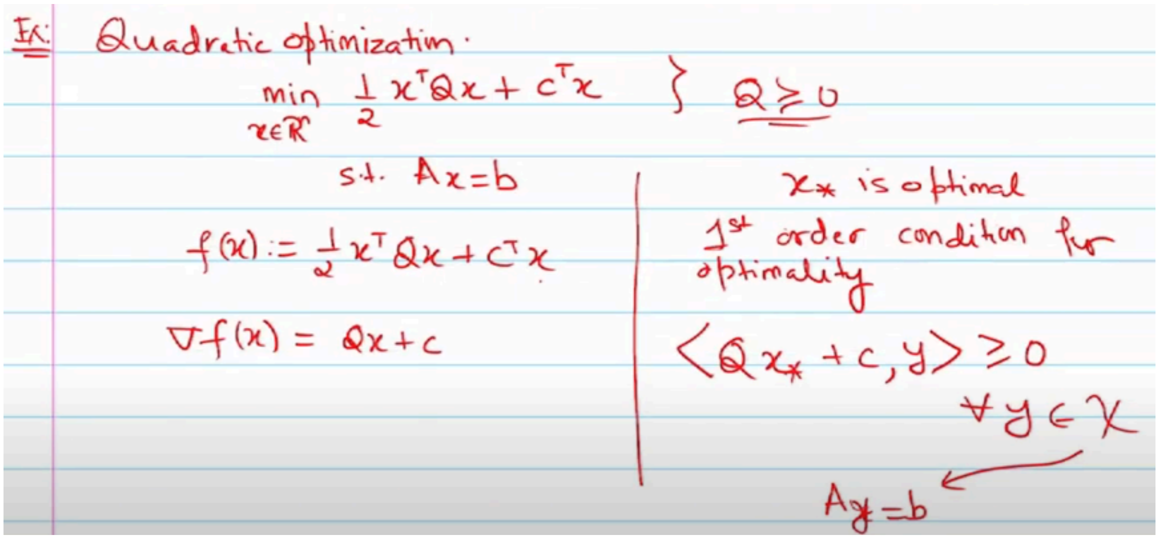


For unconstrained optimization:

$\nabla f(x^*) = 0$ if x^* is an optimizer

$$\boxed{x \text{ is optimal} \iff \nabla f(x)^T (y - x) \geq 0 \quad \forall y \in \mathcal{X}}$$

So, this the only way this is true is that the gradient of f of x vanishes ok. So, the condition the general condition is x is optimal if and only if gradient of f at x transpose times y minus x is greater than or equal to zero for all y in the feasible set okay and let us take a look at an example which is on quadratic optimization. You want to minimize half x transpose Q x plus c transpose x subject to this linear equality constraint A x equal to b. So, what is the first-order condition for optimality for this particular quadratic optimization problem? So, what is the gradient? So, in this case f of x is half x transpose Q x plus c transpose x. So, we also assume that Q is I mean this would make sense when Q is positive semi-definite right otherwise it is not even a convex problem to start with.



Ex: Quadratic optimization.

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} x^T Q x + c^T x \quad \Big\} \quad \underline{Q \succeq 0}$$

s.t. $Ax = b$

$f(x) := \frac{1}{2} x^T Q x + c^T x$

$\nabla f(x) = Qx + c$

x_* is optimal

1st order condition for optimality

$\langle Q x_* + c, y \rangle \geq 0$

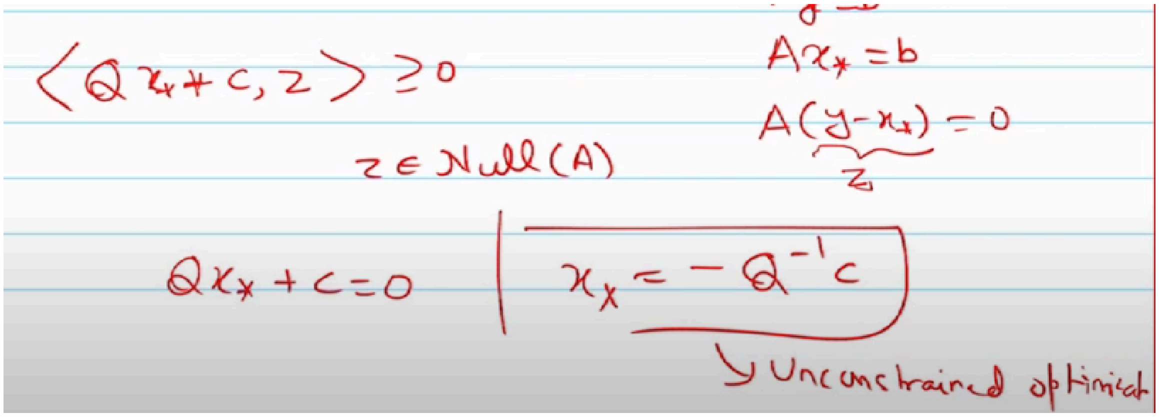
$\forall y \in \mathcal{X}$

$Ay = b$

Why is why do we need q to be positive semi-definite? We want this function to be convex right and what is the second-order condition for convexity? The Hessian should be positive semi-definite and in this case Hessian is exactly Q. So, we want this Q to be positive semi-definite and that is one of the requirements that we have for convex optimization. So, what is the gradient of f? Qx plus c okay? So the first-order condition for optimality is, let us say x star is optimal. So gradient of f of x evaluated at x star, so that means Q x star plus c y it should be greater than equal to 0 for all y in the constraint set. The constraint set is defined by the inequality the equality constraint Ax is equal to b right or Ay equal to b rather ok.

So, this is the first-order condition. Another way to look at this is we know that Ay is equal to b because y is a feasible point in a constraint set. every optimal solution or optimizer is also a feasible point. So, a x star should also be equal to b. So, that means a y minus x star this is equal to 0 right and if we call this z.

So, in some sense or you can basically. So, x star yeah right. So, what we can do is z lies in the null space of A ok, because this is true for any any y right. So, you can also equivalently write mean if equivalently write it like this and if the null space is vacuous then that means I mean it is an unconstrained optimization. So, we know that Q x star plus c is equal to going to be 0 and the x the optimal solution x star is simply going to be negative Q inverse c this is for unconstrained unconstrained optimization ok. Is this clear? Thank you for pointing this out.



$\langle Qx_* + c, z\rangle \geq 0$

$Ax_* = b$

$A(\underbrace{y - x_*}_{z}) = 0$

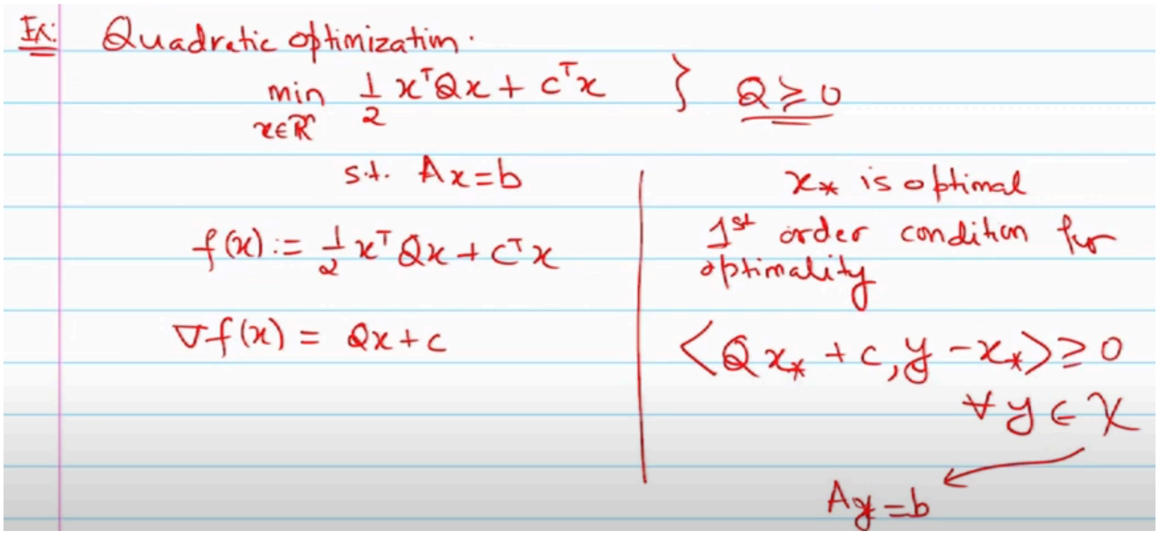
$z \in \mathrm{Null}(A)$

$Qx_* + c = 0$

$\boxed{x_* = -Q^{-1}c}$

Unconstrained optimizat

This should be y minus x star greater than equal to 0 and yeah. Because y minus x star we know it belongs to null space of A right.



Ex: Quadratic optimization:

$\min_{x \in \mathbb{R}^n} \frac{1}{2}x^TQx + c^Tx \Big\} \; Q \succeq 0$

s.t. $Ax = b$

$f(x) := \frac{1}{2}x^TQx + c^Tx$

$\nabla f(x) = Qx + c$

x_* is optimal

1st order condition for optimality

$\langle Qx_* + c, y - x_*\rangle \geq 0$

$\forall y \in \mathcal{X}$

$Ay = b$

So, this is the condition equivalent condition. So, for every z belonging to null space of A. this should be greater than equal to 0. No, so we know that x star x star A x star equal to b, A y is equal to b because both I mean both optimal point and is optimal point is also a feasible point right? So, that means A y minus x star is equal to 0.

So, y minus x star they should belong to the null space of A. and if I consider in because this is true for any y. So, I basically I can pretty much look at the null space of A right. So, for any z in the null space of A if this is greater than equal to 0 then this is another characterization of the like optimal x star. So, if Q if this is going to be Q x star plus c. So, in a product between Q x star plus c and the nulls like any vector in the null space of A if that is greater than equal to 0.

So, that that would be the first-order condition for optimality and if the null space is vacuous. So, that means that is you are in unconstrained optimization case and then in that case x star turns out to be simply negative Q inverse c. Then you just have one feasible point which is only going to be one optimal which is the optimal. No, it will be constraint optimization I mean the just that your constraints that contains just one point. Yeah, no I am saying that if then I am saying that the null space is vacuous like as in like let us say A is all zeros, A is all rows I mean basically A is I mean something which is trivially satisfied right.

So, then it becomes an unconstrained optimization. So, if the basically the if this equality constraint is not even there to start with yeah. I am saying that if you are in an like if it is an if it is a constraint that is that is what you can say, but if let us say if you are I mean if you are solving this problem in an unconstrained like without this equality constraint then the optimal solution turns out to be this if Q is invertible. So, in in fact let us let me also make it Q not greater than Q is strictly like Q is strictly positive definite. So, that we can consider inverse of Q.

So, for unconstrained optimization we know that the gradient vanishes. So, that means Q x star plus c is going to. So, this is true only for unconstrained optimization. So, I mean the solution in general it is I mean really depends. So, this is the solution right this is how you characterize.

No, we find x star such that for any vector z in the null space of A, this is satisfied. So, we are going to look at like in the starting next lecture we are going to look at ways to at least I mean if you have a constraint optimization problem convert them into unconstrained using something called Lagrangian methods and then try to find the optimal solution that way. So, we are going to look at basically being like we are going to look at ways to solve these optimization problem both analytically and then in using. like whenever the functions are nicely behaved or either like using numerical methods. As of now this is just the characterization of the optimal solution. Thank you.