Distributed Optimization and Machine Learning

Prof. Mayank Baranwal

Computer Science & Engineering, Electrical Engineering, Mathematics

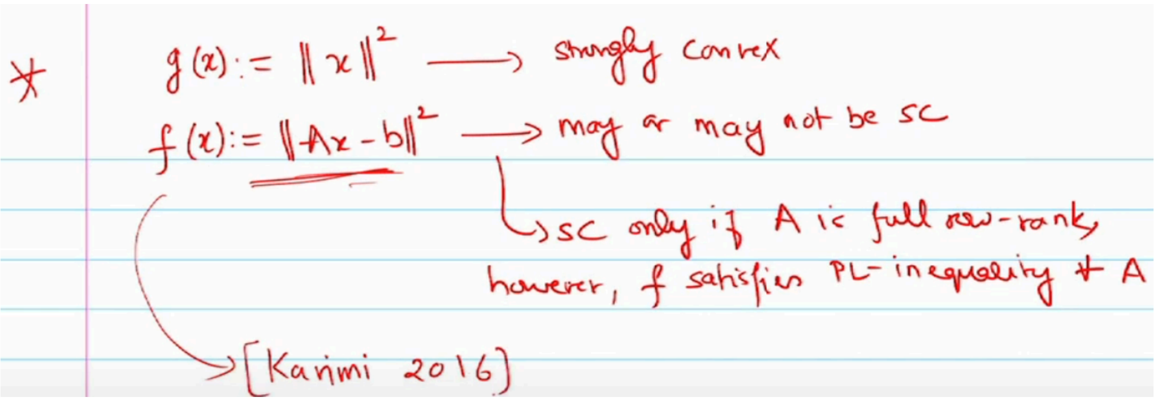
Indian Institute of Technology Bombay

Week-6

Lecture - 20: Advanced Results on PL inequality: Part 2

So, another thing about PL inequality and why do we stress so much about PL inequality is. So, if I look at least square problems. So, for instance ordinary. So, we know that a function g of x which is like norm x square. So, this is strongly convex, but the moment we look at the affine transformation of this function. So, if I define f of x like this I mean the b part is not that important, but if I look at a function of this form.

So, this may or may not be strongly convex. So this is strongly convex only if A is full row rank. So however, f satisfies PL inequality for every A. So, this function as I said like even though the original x norm x square this is strongly convex and a fine transformation of it need not be strongly convex, it is strongly convex only if a is full row rank otherwise it would not be strongly convex, but it will always satisfy pair inequality for any a.



*

$g(x) := \|x\|^2 \longrightarrow$ strongly convex

$f(x) := \|\underline{Ax - b}\|^2 \longrightarrow$ may or may not be SC

↳ SC only if A is full row-rank

however, f satisfies PL-inequality $\forall A$

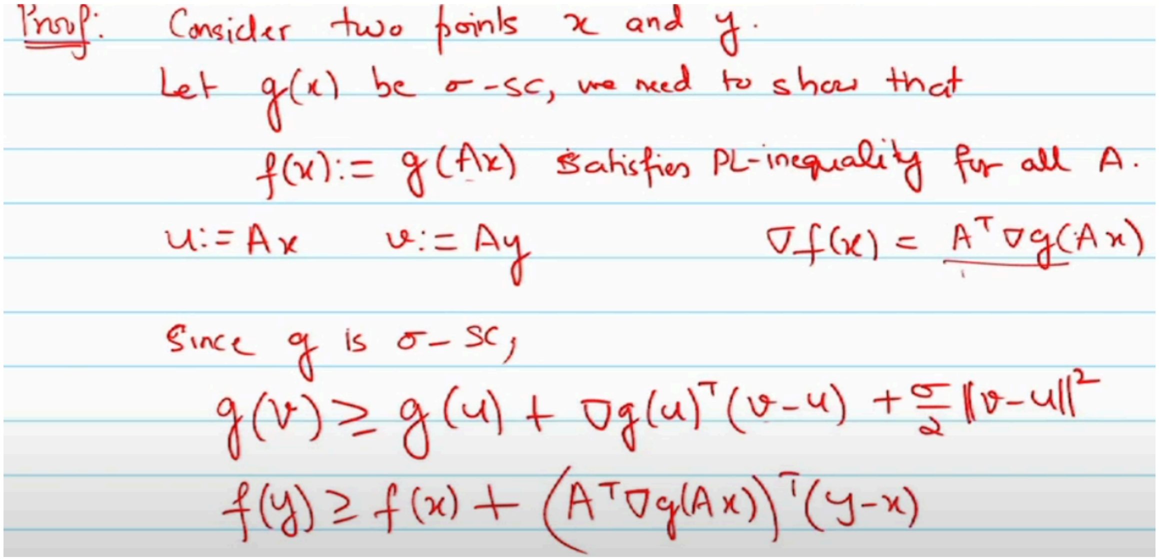
→ [Karimi 2016]

And, this kind of function is very common in least squares problem right like this I mean finding an x such that a x minus the distance between a x and b is minimized. So, this is very common right. So, again this was actually shown in by Karimi in his the same paper 2016 paper that this function always satisfies real inequality. So, let us look at this. So, let us consider two points x and y.

ok and let the question. So, the what Karimi showed was if you have a function g which is strongly convex to start with and you robustness the affine transformation for any a this will always satisfy PLInequality ok. So, if I have a g of x which is sigma strongly convex. So, we need to show that f of x defined as g of a x is satisfies PL Inequality for every for all. ok and this is not just in the context of regression problem, linear regression problem like this, but I mean the same thing you can translate it to logistic regression as well right, where you have 1 over 1 plus e to the negative Ax minus v kind of thing and original I mean if.

So, as I said I mean like it is not a very restrictive class of function that we are looking at. In fact, it if you can show something for PL inequality function satisfying PL inequality a lot more simple like optimization problems can be mapped to this particular class of functions. So, let us define u to be a x and v to be a y. So, since g is strongly convex Yeah, any matrix A. So, this need not be just for this particular choice of strongly convex function, any g can be any strongly convex function.

Now, if I consider the affine transformation of that particular function like particular x as A x minus B or A x let us say. So, this would be this need not be strongly convex for any a. In fact, if a is not full row rank it would not be strongly convex, but this will still satisfy PL inequality. So, that is the statement. So, since g sigma is strongly convex because we somehow want to get f of x.



Proof: Consider two points x and y.

Let $g(x)$ be σ-sc, we need to show that

$f(x) := g(Ax)$ satisfies PL-inequality for all A.

$u := Ax \quad v := Ay \qquad \nabla f(x) = \underline{A^T \nabla g(Ax)}$

Since g is σ-sc;

$$g(v) \geq g(u) + \nabla g(u)^T (v-u) + \frac{\sigma}{2} \|v-u\|^2$$

$$f(y) \geq f(x) + \left(A^T \nabla g(Ax)\right)^T (y-x)$$

So, we should write g of v because it holds for any x and y. So, it also holds for any u and v. So, what is g of v which is g of a x that becomes f of x by definition is greater than equal to f of x plus. So, what is gradient of g of u in terms of gradient of f? So, A inverse or A transpose, ok. So, this becomes, so g of A x, ok.

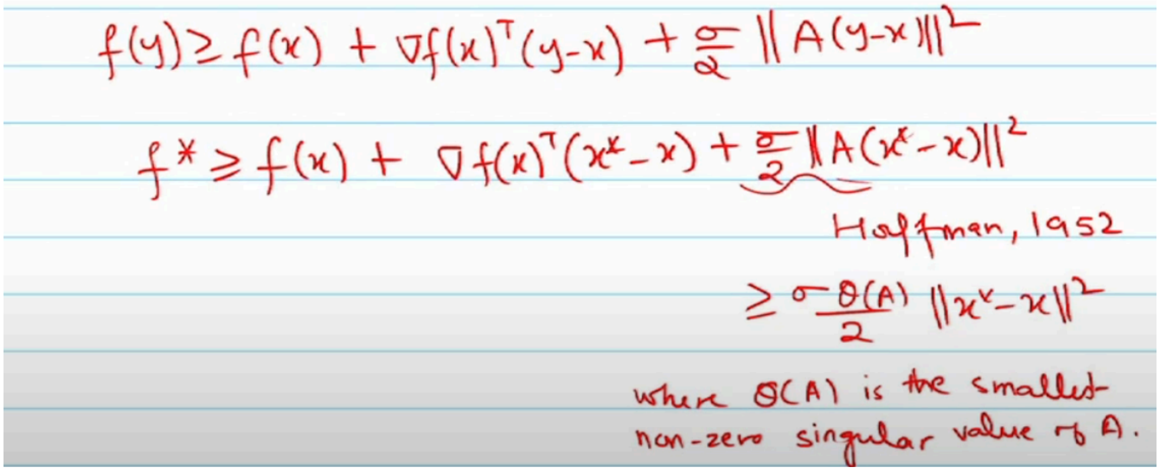
g of A x might not No, no. So, v and for v and u it satisfies right. So, for f it does not satisfy a strong convexity. This function for f it does not, but let us say a x is another point in space right v. I mean x is like a x is u and a y is v.

So, for those points u and v g satisfies I mean there is a I mean g is g is always strongly convex right for any two points. So, u and v are any two points defined through x and y. So, we would not have this directly in terms of let me write this first yes. So, this is what it is. So, right because A transpose g A x is gradient of f of x.

So, this is v minus u is A y minus A x. So, you can I can write this as A transpose this particular term and that is what it comes down to A transpose So, sorry about being sloppy on this one, but let me rewrite this. So, this is A transpose and this is nothing but gradient of f of x. Yeah, so that A is A transpose. I have taken this inside right.

So, trans. So, this is no, no. So, this is still g here right and this is a y minus a x is what I have written this like this and this is a times y minus x also. So, and by definition I mean this is nothing but gradient of f of x. So, what we get is f y. So, if I choose y to be x star or the optimal solution.

So, this turns out to be f star is greater than equal to f of x plus gradient of f of x transpose x star minus x. So, there is a result by Hoffman in I think it in his book in 1952. So, that shows that for this optimal point x star. So, this particular term is actually greater than equal to or the term inside this thing is actually greater than equal to the smallest nonzero singular value of. So, let me first write this of matrix A.



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

$$f^* \geq f(x) + \nabla f(x)^T (x^* - x) + \underbrace{\frac{\sigma}{2} \|A(x^* - x)\|^2}$$

Hoffman, 1952

$$\geq \frac{\sigma\,\theta(A)}{2} \|x^* - x\|^2$$

where $\theta(A)$ is the smallest non-zero singular value of A.

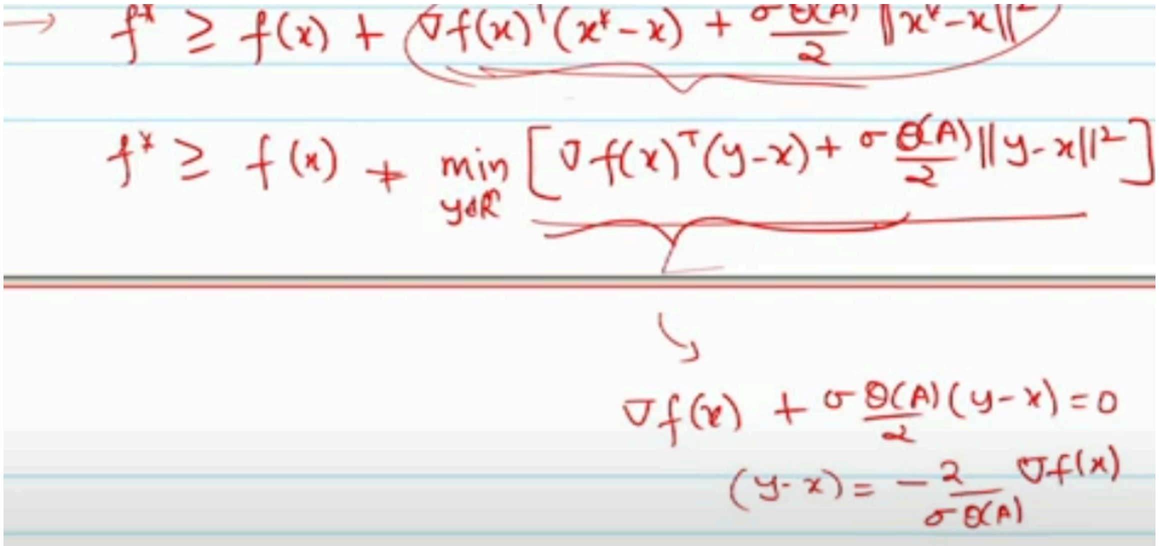
So, where theta A is the smallest nonzero singular value of A. So, essentially, why do we see a singular value of 'a' here? This is because it is \(x^* - x \), right? So, you will see that the terms are of the form \(A^T A \), and the singular values are simply the eigenvalues of this \(A^T A \) matrix or its square, right? So, this is what it is. (The sentence is already grammatically correct.) So, basically, what you get is that f star is greater than or equal to f of x plus the gradient of f of x transposed multiplied by x star minus x. No, this result is valid only for y equal to x*.

I mean, if that had been the case, then this would become a strongly convex setting, right? If that result had been valid for any y, it would be different. However, this result is valid only for x star, yes. No, but it is not true for any y; it is only true when y is equal to x star. No, that is the result from Hoffman '95 that I am talking about. So, for this type of function and this type of inequality—usually for a strongly convex function—you have that f(y) is greater than or equal to f(x), but if you only have it in terms of the optimal x*.

So, functions of this type are called weakly strongly convex if they satisfy this condition. So, if \(f \) satisfies the \(g \) pale inequality, then the first key point to show here is that \(f \) is weakly strongly convex, and we will then demonstrate that it also satisfies the PL inequality. So, this means that f satisfies only this inequality without any y, right? So, we call it weakly strongly convex; otherwise, with \(y \), it would have been strongly convex right away, correct? So, f* is greater than or equal to this. So \(f^* \) must also be greater, which means the right-hand side is true specifically for \(x^* \). So if I try to minimize this with respect to \(y \), let's say in \(\mathbb{R}^n \), the entire right-hand side would still be true because this minimum value will be smaller than when you choose \(y \) to be exactly \(x^* \).

So, this will always be true. (The sentence is already grammatically correct.) So, if I try to minimize this, the minimum value can be x star or something else, but the value you obtain from this particular function will always be less than or equal to the value obtained when you choose y equal to x star. So, that is what we wrote. (The sentence is already grammatically correct.

) So, f star is greater than or equal to this particular term. Now, the unconstrained minimizer of this expression should have, with respect to \(y \), the gradient of \(f(x) \) plus \(\frac{\sigma \theta a}{2}(y - x) \) equal to 0, correct? Let us see, as we are attempting this as an unconstrained minimization on y. This means that y minus x is equal to minus 2 over sigma theta, which is the gradient of f of x. So, let us now rewrite this to mean that \(f^* \) is greater than or equal to \(f(x) \) plus the gradient of \(f(x) \) transposed.



$$\to \quad f^* \geq f(x) + \underbrace{\nabla f(x)^T(x^*-x) + \frac{\sigma\theta(A)}{2}\|x^*-x\|^2}$$

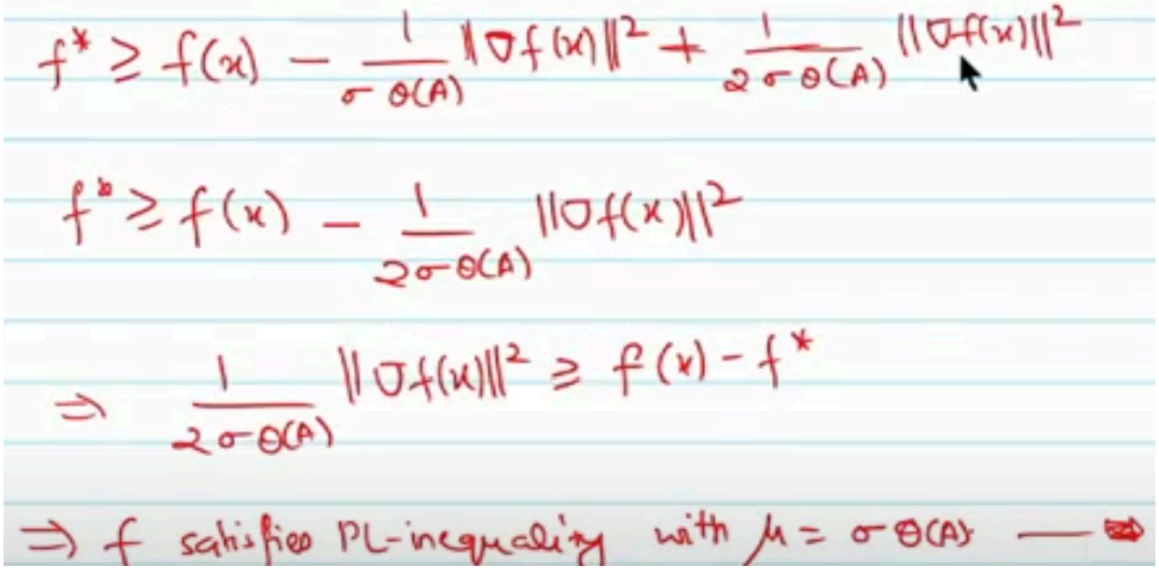
$$f^* \geq f(x) + \min_{y\in\mathbb{R}} \underbrace{\left[\nabla f(x)^T(y-x) + \sigma\frac{\theta(A)}{2}\|y-x\|^2\right]}$$

$$\nabla f(x) + \sigma\frac{\theta(A)}{2}(y-x) = 0$$

$$(y-x) = -\frac{2}{\sigma\theta(A)}\nabla f(x)$$

Let me just write this term. Plus sigma theta a over 2, y minus x squared.   Corrected: Plus \(\sigma \theta a / 2\), \(y - x^2\). So, sigma theta is equal to a divided by 2, and then you have 4 sigma squared theta squared. So, you get \(f^* \geq f(x) - \frac{1}{2} \sigma \theta a \), which is the same as \(\frac{1}{2} \sigma \theta a \) times the square of the gradient. So, this is greater than or equal to f of x minus f*.

So, this implies that f satisfies the pair inequality with μ equal to σ multiplied by θ. So, this is proof.   (Note: The original sentence is already grammatically correct.) So, essentially, I mean that this again shows that if you have a function that is σ-strongly convex, and you consider an affine transformation of that function with respect to the variable x, you can define a new function like this. This would satisfy—I mean, it need not be strongly convex, but it would always satisfy the PL inequality with exponent sigma times theta a.



$$f^* \geq f(x) - \frac{1}{\sigma\theta(A)}\|\nabla f(x)\|^2 + \frac{1}{2\sigma\theta(A)}\|\nabla f(x)\|^2$$

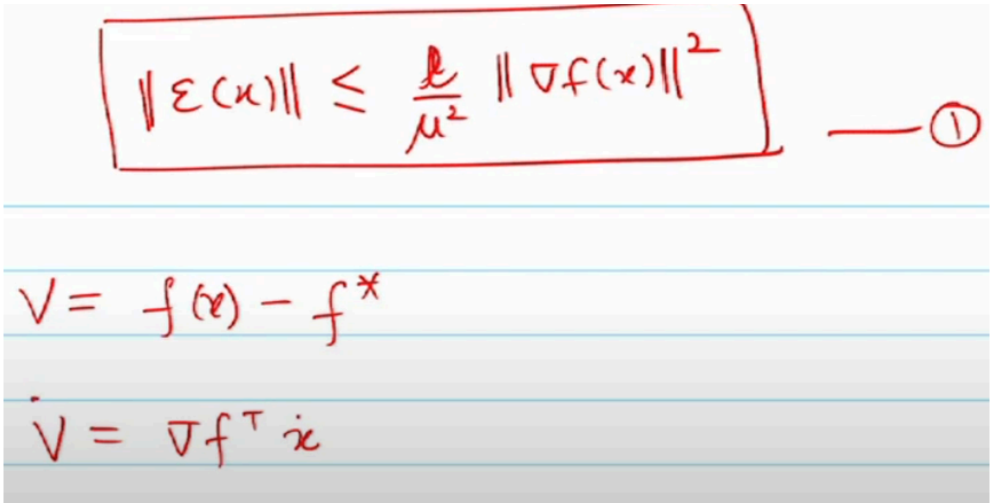
$$f^* \geq f(x) - \frac{1}{2\sigma\theta(A)}\|\nabla f(x)\|^2$$

$$\Rightarrow \quad \frac{1}{2\sigma\theta(A)}\|\nabla f(x)\|^2 \geq f(x) - f^*$$

$\Rightarrow$ f satisfies PL-inequality with $\mu = \sigma\theta(A)$ →

So, we are now going to examine the robustness of f(x, t) = d f, and I will explain what we mean by robustness. So, typically, when we compute—at least in the data-driven regime—we try to... So, we need not have the analytical expression for the gradient of f correct.

So, we consider the sample as different examples or sample points, and based on those, we try to estimate the gradient of f(x) as, let us say, i equals 1 through n, with the gradient evaluated at different points, right? That is how we attempt to approximate the gradient. So, what I am trying to suggest is that the gradient computation need not be exact. So, if you have a dynamical system that looks something like this: \(\frac{p - 2}{p - q - 2} \) or \(q - 1 \), So, the gradient of f(x) need not be exact in most cases. So, let's say we have some inexactness in the gradient, and we are going to capture this inexactness through an additive disturbance, like this. So, for now, think of it purely as a control issue.



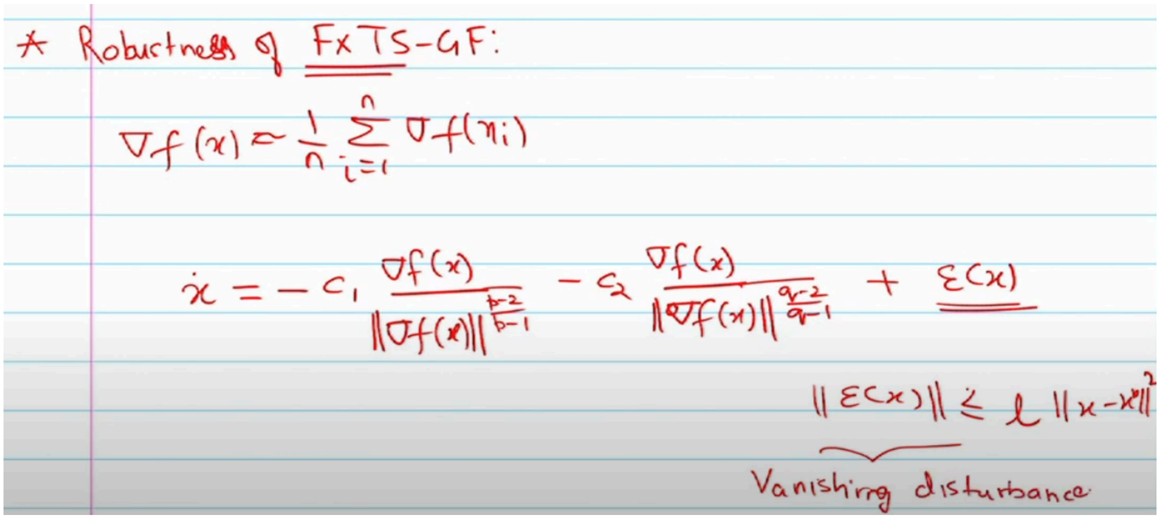
$$\boxed{\|\varepsilon(x)\| \le \frac{\ell}{\mu^2}\|\nabla f(x)\|^2} \text{ — } ①$$

$$V = f(x) - f^*$$

$$\dot{V} = \nabla f^T \dot{x}$$

So, you have an additive disturbance that is being added to your system. The word "equilibrium" is a noun and does not contain any grammatical errors by itself. If you intended to provide a complete sentence or need a specific context, please share that, and I will help you correct it! To guarantee that the equilibrium remains at x star, or the optimal solution, you must assume a certain structure concerning this disturbance. So, this disturbance should be a temporary disturbance. So, what we assume is that the norm of this should be less than or equal to some constant \(l \) times the square of \(x - x^* \), something like that.

So, that x star is still the equilibrium for this situation. Why does it mean that, even if we do not have this kind of assumption, we are at x star if epsilon f of x is non-zero? So, you would still be oscillating around the equilibrium, right? So, we assume this kind of vanishing perturbation or vanishing disturbance, and what we are going to show is that if we choose. So, if we choose c1 and c2 to be sufficiently large, we will still converge in a fixed amount of time, even in the presence of disturbance. So, if you have a vanishing disturbance like this, you are still guaranteed to converge in a fixed amount of time, as long as you choose C1 and C2 to be sufficiently large. This also answers your question regarding the ISS type of situation.



* Robustness of FxTS-GF:

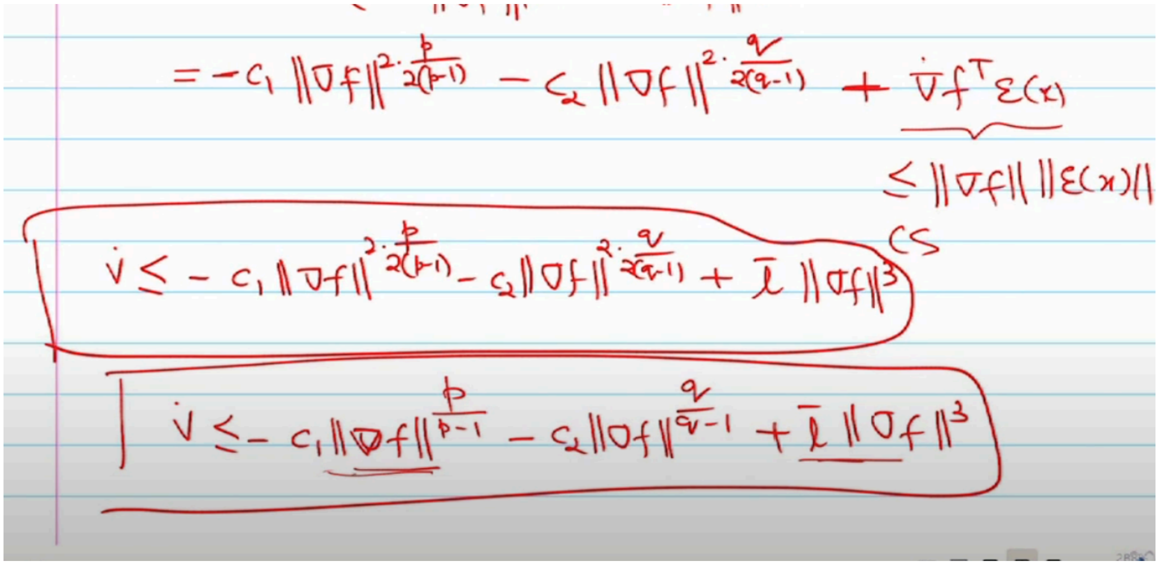
$$\nabla f(x) \approx \frac{1}{n}\sum_{i=1}^{n} \nabla f(x_i)$$

$$\dot{x} = -c_1 \frac{\nabla f(x)}{\|\nabla f(x)\|^{\frac{p-2}{p-1}}} - c_2 \frac{\nabla f(x)}{\|\nabla f(x)\|^{\frac{q-2}{q-1}}} + \underline{\varepsilon(x)}$$

$$\underbrace{\|\varepsilon(x)\| \le \ell}_{\text{Vanishing disturbance}} \|x-x^*\|^2$$

Well, I mean you do not really need ISS, but it is largely, I mean, a significant gain. If you choose C1 and C2 to be sufficiently large, you can actually subsume this disturbance. Let us see how. We are going to use the result that we just derived for this purpose. If a function \(f \) satisfies the PL inequality, then it has at least one quadratic root.

So, we are going to assume that \(f \) in this case satisfies the PL inequality with some exponent \(\mu \) greater than 0. So, if \(f \) satisfies the PL inequality, what do we know about it? That function must have at least quadratic growth, right? So, this is from a function that has at least quadratic growth. F satisfies the PL inequality. The definition of PL inequality is as follows. The sentence "Is this clear?" is indeed already grammatically correct.



$$= -c_1\|\nabla f\|^{2\cdot\frac{p}{2(p-1)}} - c_2\|\nabla f\|^{2\cdot\frac{q}{2(q-1)}} + \underbrace{\dot{\nabla} f^T\varepsilon(x)}_{\le \|\nabla f\|\,\|\varepsilon(x)\|}$$

CS

$$\boxed{\dot{V} \le -c_1\|\nabla f\|^{2\cdot\frac{p}{2(p-1)}} - c_2\|\nabla f\|^{2\cdot\frac{q}{2(q-1)}} + \bar{\ell}\,\|\nabla f\|^3}$$

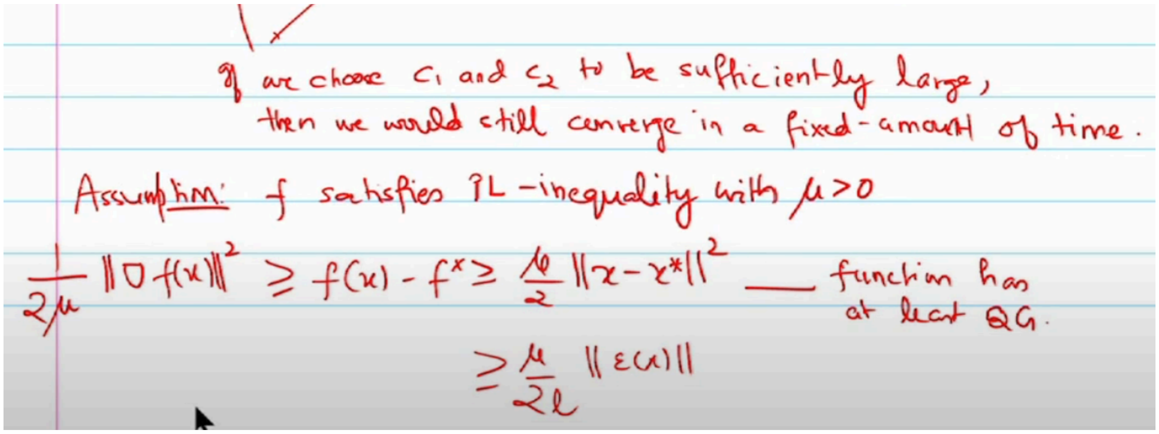
$$\boxed{\dot{V} \le -c_1\underline{\|\nabla f\|^{\frac{p}{p-1}}} - c_2\|\nabla f\|^{\frac{q}{q-1}} + \underline{\bar{\ell}}\,\|\nabla f\|^3}$$

No changes are needed. With this assumption, x minus x star is greater than or equal to epsilon x over L, which is greater than or equal to mu over 2L. The sentence is already grammatically correct. No changes are needed. So, from here, we know that this value is less than or equal to L divided by mu squared. The sentence is indeed grammatically correct as it stands.

No corrections are necessary. The sentence is already grammatically correct. So, we have an upper bound on this vanishing disturbance in terms of the gradient of \(f(x) \), and this will be useful later. So, we know that \(f \) satisfies the PL Inequality. So, if we want to demonstrate convergence to the optimal solution, what could be a good Lyapunov candidate? f of x minus f star, right? Here, we are not going to use the half gradient norm squared because we do not have any information about the Hessian of \(f \). We have this information for the strongly convex case, and if we are using a Newton-type flow, we can also work with the strictly convex case.

But then, since functions that satisfy this condition appear only in equality and need not necessarily be strongly convex, I mean, this part is a good choice for the Lyapunov candidate, right? \(f(x) - f^* \). So, v dot turns out to be okay, and x dot represents this particular dynamical system over here. So, let's write this down. So, the gradient of f transpose minus c times (1 plus epsilon x) is acceptable. So, this expression is equal to minus c1 times p, just as before, minus c2 times this particular term, and the last term is the transpose of the gradient of f times epsilon.

So, how do we eliminate epsilon from here using this method, given that we only have a constraint on the norm of epsilon? So, for that, you can use the Cauchy-Schwarz inequality here, right? This expression will be less than or equal to the product of the norms of these two quantities, using Cauchy-Schwarz. The sentence "Is this clear?" is already correct. So, that means \(\dot{v} \) is less than or equal to \(-c_1 - c_2\), and if I use this particular condition here, let us call it \(\bar{l} \) or something. This is going to be less than or equal to \(\bar{l} \) times \(q \), okay? So, let me rephrase this: we should not do this particular thing right now. So, let us consider \(\frac{p}{p - 1} \times \frac{q}{q - 1} \) and label this as \(\bar{l} \).



If we choose c_1 and c_2 to be sufficiently large, then we would still converge in a fixed-amount of time.

Assumption: f satisfies PL-inequality with $\mu > 0$

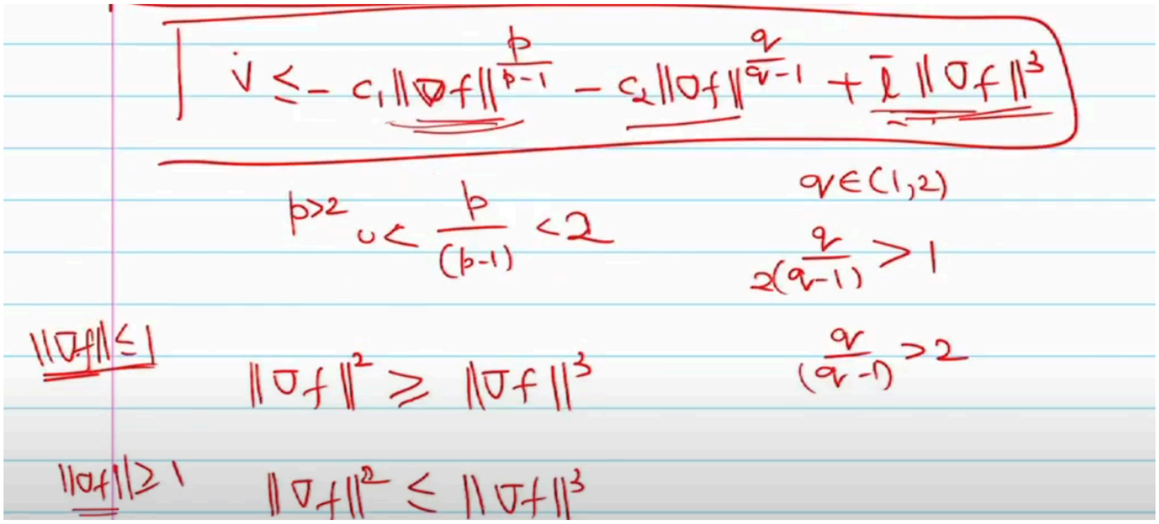
$$\frac{1}{2\mu}\|\nabla f(x)\|^2 \geq f(x) - f^* \geq \frac{\mu}{2}\|x - x^*\|^2$$ — function has at least QG.

$$\geq \frac{\mu}{2L}\|\varepsilon(x)\|$$

Here, I mean that you are getting—almost without this particular term—the derivation would have been exactly the same, right? Now, you have a positive term that conveys a meaning, and it has an opposite polarity, right? So, we want to subsume this in some way. So, for p greater than 2, what is this exponent $\frac{p}{p - 1}$? (The original sentence is already grammatically correct.) So, for p greater than 2, $\frac{p}{2p - 1}$ is a number between 0 and 1, correct? So, the square of this will be a number between 0 and 2. So, if it is essentially okay for q to be a number between 1 and 2. So, we know that $\frac{q}{q - 2q - 1}$ is a number greater than 1.

So, q over (q minus 1) is basically going to be greater than 2. Now, we have two regimes: one with an exponent less than 2 and the other with an exponent greater than 2. Now, this positive term is over here. When the norm of the gradient of f is less than 1, what kind of inequality would we have between something like this and something larger? When the norm of the gradient of f is greater than or equal to 1, it implies that this quantity is also greater than or equal to that, right? When the norm of the gradient of f is less than or equal to 1, no, sorry, that was my mistake; I meant the other one. The sentence can be corrected as follows: "It is less than or equal to 1; this is the inequality you have.

When the norm of the gradient of f is greater than or equal to 1, you have the other inequality, which is valid." So, that means depending on which regime you are in, you can either use this to subsume this particular additive term or use it to subsume this additive term. So, effectively, if I choose c1 and c2, whose values are greater than L bar. I can write this as $v \cdot = c_1 - \frac{\bar{L} p}{p - 1} - c_2 - \bar{L}$ because it really depends on which regime you are in. So, at least one of the terms will be subsumed, and then you can also decrease the other exponent by this.



$$\boxed{\dot{v} \leq -c_1\|\nabla f\|^{\frac{p}{p-1}} - c_2\|\nabla f\|^{\frac{q}{q-1}} + \bar{l}\,\|\nabla f\|^3}$$

$p>2$ $\quad 0<\frac{p}{(p-1)}<2$ $\qquad q\in(1,2)$

$\frac{q}{2(q-1)}>1$

$\|\nabla f\|\leq 1$ $\quad \|\nabla f\|^2 \geq \|\nabla f\|^3$ $\qquad \frac{q}{(q-1)}>2$

$\|\nabla f\|\geq 1$ $\quad \|\nabla f\|^2 \leq \|\nabla f\|^3$

So, as long as you choose c_1 and c_2 to be sufficiently large, in this case, the definition of "sufficiently large" is that they should be greater than $\bar{l}$. So, this holds true, right? Therefore, the rest of the proof follows similarly. So, I can write this as \(\dot{v} \leq -c_1 \bar{l} \). Now, this represents the gradient of the squared norm \(\frac{1}{2} p - 1 - c_2 - \bar{l} q^{2} - 1 \).

Now, this is nothing, but it is 2 times v. Well, it's not exactly 2 times v, but we can further use the peer inequality here, right? So, this is nothing less than or equal to minus c, 1 minus l bar. The sentence is already grammatically correct. However, if you are looking for a slight rephrase, you could say, "What does this term mean?" It is less than or equal to 2μv, right? The sentence is already correct. No changes are needed. No, if we choose \(p \) to be a number greater than 2 and \(q \) to be a number between 1 and 2, right? For these choices, we know that this holds true.

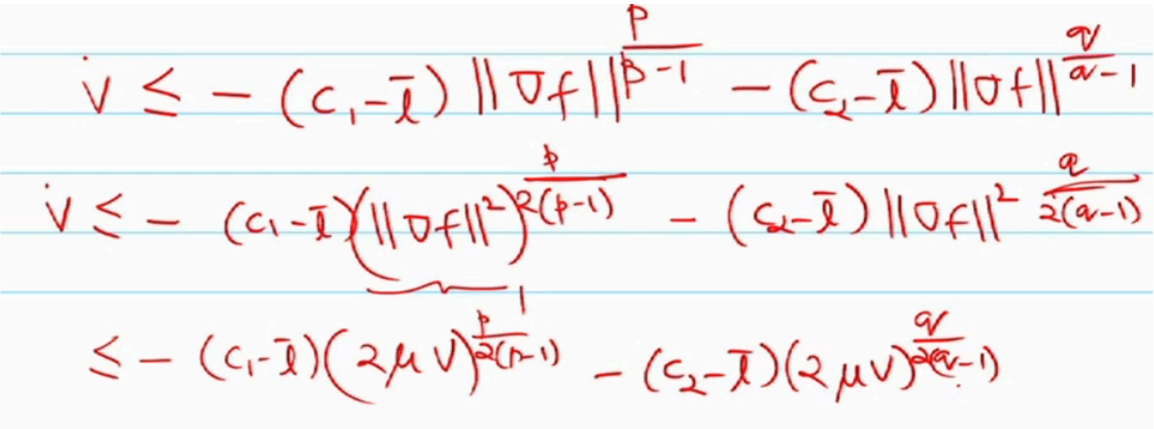
No, that is \(\frac{p}{2p - 1} \) between 0 and 1, correct? So, p divided by p minus 1 is the result. The original sentence "Between 1 and 2." is grammatically correct, but it is indeed incomplete. A complete sentence could be: "The numbers fall between 1 and 2.

" or "The value lies between 1 and 2." Between 1 and 2, yes, sure. (This sentence is already grammatically correct.) I mean, this is also true, but you can always write it as 1 and 2; that's fine. I mean, it is always, yeah, dividing it by a smaller number. So, it is always going to be a number greater than 1, which is fine; however, the point is that you will get an exponent that is less than 2 and an exponent that is greater than 2. So, using those exponents, you can always subsume this, right? (Note: The original sentence is already grammatically correct.

) So, if you want to subsume this particular item, So you won't be able to subsume it in this way. You need to choose a value of q such that you will actually get an exponent greater than 3. So it won't work for any \(q \) between 1 and 2. But let's say I choose a number \(q \), like \(\frac{3}{2} \). It is also a number between 1 and 2, but if I choose q to be 3 divided by 2, then q over (q minus 1) is 1.

5 divided by 0.5, which equals 3, right? So, for \(q \) greater than \(\frac{3}{2} \), this statement would be true, right? It would also encompass this particular case, okay? So, C1 and C2 should be greater than L̅, and q should be greater than 3/2. Yeah, I mean, if you don't know the information about L bar or L, you can just choose them to be sufficiently large. It's not upper-bounded, right? It's lower-bounded. So, you can choose them to be large.

(Note: The sentence is already grammatically correct. If you would like a different phrasing, you could say: "Therefore, you can choose them to be large.") So, eventually, you would subsume that L-bar. But Q, you need it to be greater than \(\frac{3}{2} \). And this is P over 2P minus 1. "Minus c² minus l̅² mu v, this is q over 2q minus 1, right?" Corrected: "Minus c² minus l̅² mu v; this is q over 2q minus 1, right?" So, now you have it in a form where you can simply use Polykov's condition for fixed-time stability, and you can show that \(x \) will converge to \(x^* \) in a fixed time, okay? Is this clear? (The sentence is already correct.



$$\dot{V} \leq -(c_1-\bar{l})\|\nabla f\|^{\frac{p}{p-1}} - (c_2-\bar{l})\|\nabla f\|^{\frac{q}{q-1}}$$

$$\dot{V} \leq -(c_1-\bar{l})\underbrace{(\|\nabla f\|^2)}^{\frac{p}{2(p-1)}} - (c_2-\bar{l})\|\nabla f\|^{2\,\frac{q}{2(q-1)}}$$

$$\leq -(c_1-\bar{l})(2\mu V)^{\frac{p}{2(p-1)}} - (c_2-\bar{l})(2\mu V)^{\frac{q}{2(q-1)}}$$

) It seems you haven't provided a sentence for correction. Please share the sentence you would like me to help with! The sentence is grammatically correct as is. However, if you're looking for a more formal version, you could say, "Yes, yes." So, the conditions are: let me rewrite them.

This works. So, we need c1 and c2 to be greater than l-bar. Again, even if you do not know lbar, you can choose c1 and c2 to be large enough, correct? So, it will eventually be greater than lbar, and q will be greater than three halves. Obviously, q is a number between 1 and 2, but you want it to be greater than 3 by 2. So, why do we need \(q \) to be greater than \(\frac{3}{2} \)? (This sentence is already grammatically correct.) To ensure that this particular exponent, either q or q minus 1, is greater than 3, q must be less than 3/2. If I use q as, let us say, 2, then q should indeed be less than 3/2, right? So, if q is less than 3 by 2, then the exponent is q divided by (q minus 1).

So, that is \(\frac{q}{q - 1} \), which is greater than 3, okay? So, you need \(q \) to be less than \(\frac{3}{2} \). This implies that \(\frac{q}{q - 1} \) is greater than 3, and in that case, when the gradient of \(f \) is greater than 1, you can actually use this particular term to subsume this positive term. When the gradient of \(f \) is less than 1, you can use this term to encompass it. Because there is a minus sign here, right? Again, today's lecture was a bit math-heavy, mainly because we are looking at concepts that are relatively more advanced. But I just wanted to cover those for people who are generally interested in this area.

From the next lecture onwards, we will revisit the discretized optimization algorithm, and we may start exploring something called the augmented Lagrangian method or the method of multipliers. And then that would eventually connect us with how we approach solving constraints of the form \(ax = b \), where different agents know different parts of the matrix \(A \). How do they collaborate to work under a common constraint and minimize a shared objective function? And then, starting from that lecture onwards, we would eventually begin to diverge more towards distributed optimization, which is key.

This is a part of the course. The sentence is already correct as it is. However, if you're looking for a more formal way to express agreement, you could say, "Yes." It seems you haven't provided a sentence for correction. Please share the sentence you'd like me to help with! Yeah, I mean at least from this, if you choose C1 and C2 to be sufficiently large. Then let \(q \) be an exponent that is less than \(\frac{3}{2} \).

This will work in continuous time. Now, let's move on to the discretized implementation. The discretized implementation would have exactly the same considerations. For instance, if I look at a discretized implementation like this, As I said, the only guarantee you have is that if the continuous-time dynamical system has a certain property, the discretized implementation will have that property only for a sufficiently small eta. So, eta cannot be too large, and that also makes sense because you are actually scaling the gradients up, right? Whether you are closer to the optimal solution or even farther away, you are still scaling the gradients up. So, if you choose very large step sizes, you are probably making this discretized implementation diverge.

No eta wall is going to exist; in general, eta means there is an upper bound. However, we do not know what that upper bound is; this is an open problem.