Convex Optimization

Prof. Joydeep Dutta

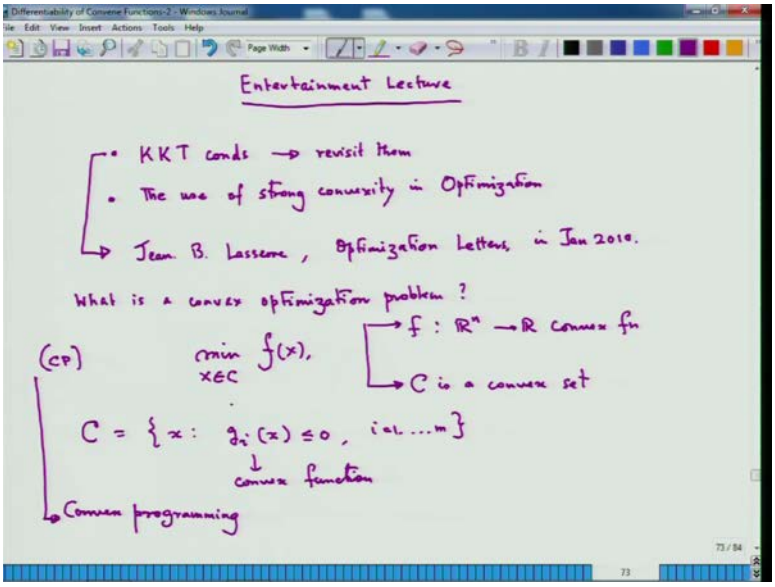
Department of Mathematics and Statistics

Indian Institute of Technology, Kanpur

Lecture No. # 40

So, this is the concluding lecture of the course on convex optimization. The course on convex optimization is really not complete and under graduate course in the true sense of the term. It has a mix of graduate and under graduate flavor. Both having a bit of algorithmic flavor in some sense for very particular class of problems and also having a good amount of theoretical flavor. Now, today's lecture usually the in the last concluding lecture, people summarize what they have done in this course and how could you use this course for your other activities. So, but I I thought that I would be giving you some sort of an entertainment lecture by telling some interesting snippets about convex optimization, and telling you some details about things.

(Refer Slide Time: 01:17)



So, I would let let me just say today's lecture is an entertainment lecture. So, you relax, forget your note books; just listen to what I am saying. You would really look into two aspects - KKT conditions and we will revisit them, and number two - the use of strong convexity.

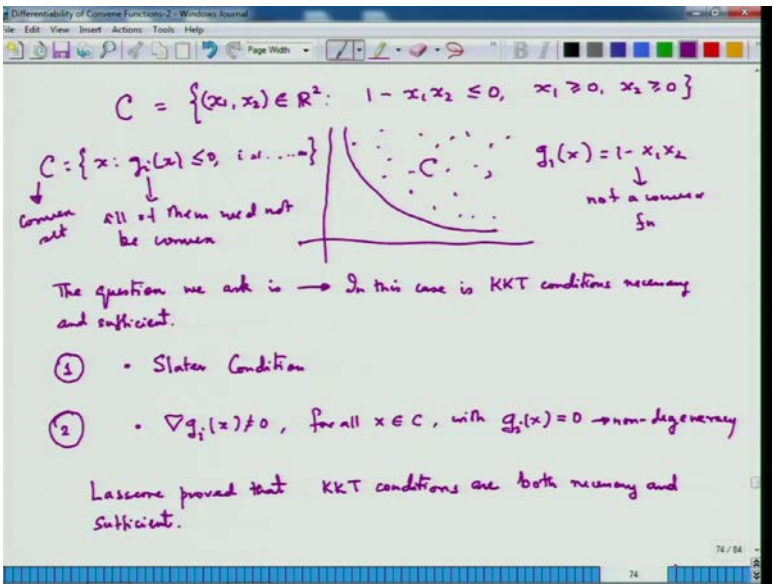
(No audio from 01:52 to 02:10)

Now, this revisiting of KKT conditions will do first and will take the least time. This was due to a work of Jean B. Lasserre which was published in the journal called optimization letters in January of January issue of 2010. This (( )) title called representations of the feasible region in convex optimization. See in general, what is the meaning of a convex optimization problem? So, we come back make a full circle and come back and ask the question…

(No audio from 03:05 to 03:16)

What is a convex optimization problem? This problem simply means you want to minimize a convex function x, f over or more succinctly like this, where f is a convex function and C is a convex set. Now, in general, we have always assumed in this course at this set C in most cases is described by the set of all x is for which g i (x) is less than equal to 0 for every i from 1 to m, and each of these g i(s) are convex functions. If C is described in such a way, then this convex optimization problem is usually called a convex programming problem that is why the CP sign is used in most cases is called convex programming. But usually it means minimization of a convex function over a convex set. Now, the question is, to represent a convex funtion set in terms of inequality constraints, does it mean that I always have to consider functions each of this functions to be convex.

(Refer Slide Time: 05:10)



That need not be the case, because if you take the set C, (( )) R 2, such that… Say if this is my feasible region, then this feasible region depicts a following convex set. And if I consider this function g 1 (x), this is not a convex function, but though. So, there are other functions which are convex, but this is not a convex function. So, I can write this as minus x 1 less than 0 minus x 2 less than 0 which would be convex function, but g 1 (x) is not a convex function, but the set C ultimately is convex. So, here what is happening is a representation, in the representation of the set C, I am force to take in a non-convex function; at least one of the function is non-convex. So, it is really not a convex programming problem in the sense of terminology, but it is a convex optimization problem.

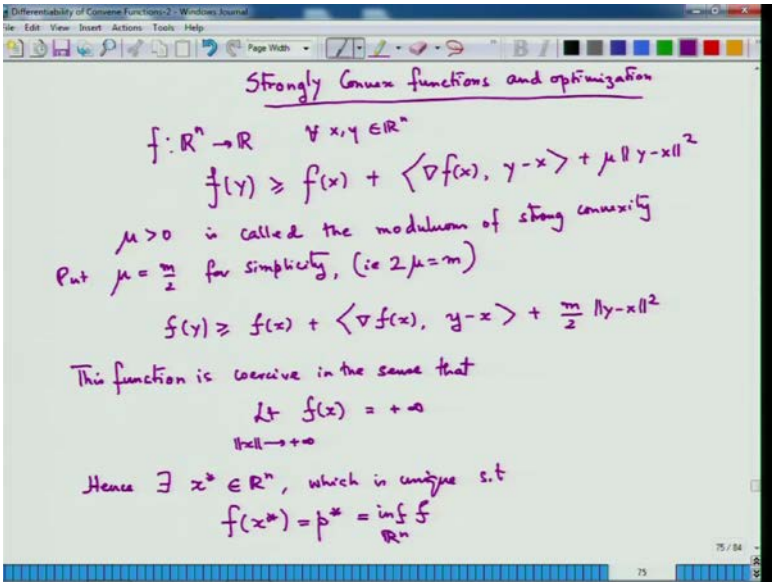
So, what we do is the following that the question we ask is for the following. In this case is KKT conditions necessary and sufficient. So, if that is the question, what constraints qualifications are needed for this to take place? Is Slater condition enough? The answer is no, the Slater condition is not enough. So, apart from the Slater condition, so we will need two conditions for this to happen. And number 2 following, we will assume that all this constraints are differentiable, the g i(s) are differentiable for all x in C with… So, for every point in the boundary this vector is non-zero. That is the idea. You see now you are considering the condition case where C is given in terms of inequality constraints that is C is written in terms of… And this all of them need not be convex.

So, sometimes we if I just list down the constraints, if there is a non-convex constraint we consider it as a non-convex programming problem. For non-convex optimization problem, without realizing there is hidden convexity, because the ultimate feasible set C that that is the convex set. So, for such problems, if these two conditions are satisfied, then Lasserre prove that KKT conditions, so this condition is called a non-degeneracy condition. g i The KKT conditions are both necessary in this situations. So, even if all the elements of the set, it represent function, representing the set C is not convex still we can write down, we can show that the KKT conditions under these two condition, if these two assumption, this is called the non-degeneracy condition. We are not going to too much detail as it is just snippets, a non-degeneracy condition.

So, Lasserre prove that KKT conditions are both necessary and sufficient. So, this is one interesting aspect that even if you representation is not given all by convex functions, you can still have a necessary KKT conditions to be necessary and sufficient, provided

that not only Slater condition holds, but something else additionally also holds. It has also been recently bought to the non non-differentiable situation by colleague or myself, and me and with the colleague.

(Refer Slide Time: 10:56)



So, let us forget that part and let us going to a more interesting part, which is more interesting from an algorithmic point of view, and that is what we will now study strongly convex functions in an optimization. How strongly convex functions effect optimization? So, if you have function from R n to R m strongly convex function is one whose hessian matrix has always got to be positive semi definite sorry positive is not positive semi definite, positive definite. So, for any x, y if the function is differentiable, then this is the definition of… So, for all x, y in R n this is what will happen. Try it out with x square, where mu is greater than 0 is called the modulus of strong convexity.

(No audio from 12:06 to 12:18)

Just for simplicity, we will put mu equal to m by 2 just because we need to differentiate this part also very soon.

(No audio from 12:26 to 12:36)

So, once I do that this expression would now be…

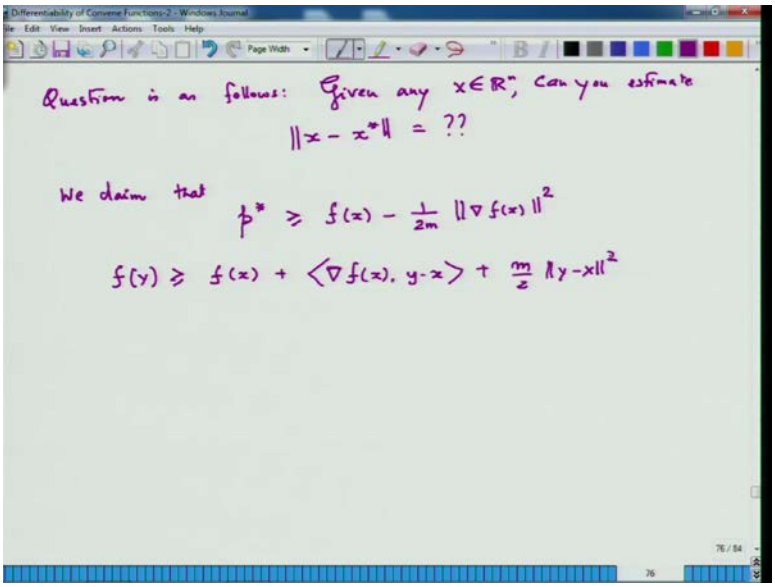
(No audio from 12:42 to 13:00)

Where m is of course some quantity bigger than 0, now m is nothing but twice of mu that is (( ))

(No audio from 13:09 to 13:21)

So now, this function is coercive, I thing which you have heard before in the sense that limit of f(x) is equal to plus infinity, if the norm of x is going to plus infinity. Hence there exist x star in R n, which is unique such that f of x star is equal to p star is equal to the infimum value of f over R m. So, whenever the function is quite shift they would always exist a unique minimizer. The uniqueness comes from function been strongly convex, strongly convex function is a subclass of strictly convex function. So, strongly convex functions are those subclass of strictly convex function for which the hessian matrix at every x is always positive definite. For quadratic problem, strictly and strongly convex classes coincide.

(Refer Slide Time: 15:03)



Now, we are first… So, our question is the following which makes very good sense from the point of view of computation. Of course, if you you want to compute grad f(x) and possibly if put that equal to 0 and get the answer, if you want to solve it. But here if you you would observe very soon that a it is in order to solve this problem, we really again have to go back and apply standard algorithms. Let us doing grad f(x) equal to 0 might not (( )), grad f(x) equal to 0 only you can have approximate answer.

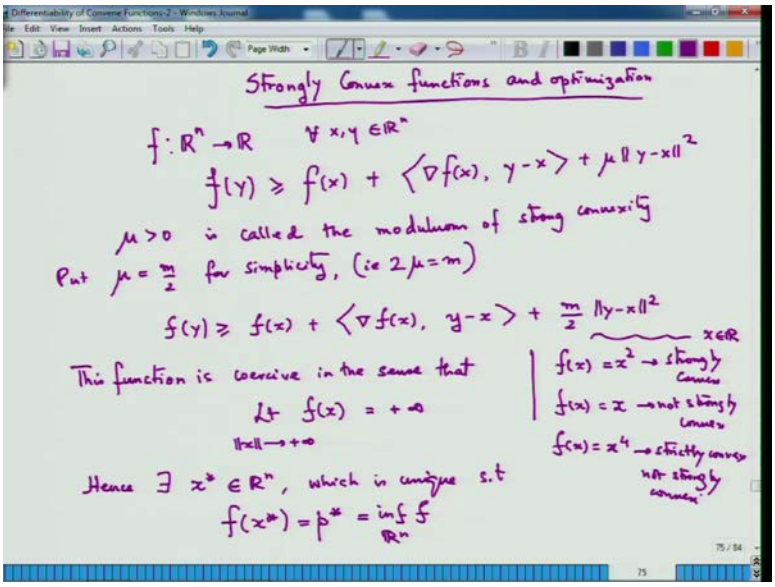
So, given any x, so the natural question is given any x in C, so one of which could be approximate answer; given any x in C any sorry a given any x in R n not C. Can you estimate, the question is, can you estimate this distance. That is the question. That is the absolutely relevant question from the algorithmic point of view. Because if you are stopping, you do not know the exact x star, you just stopping at algorithm there algorithm and taking that x k as your solution; how good is your approximation that is the very important thing from a numerical point of view. And so getting an upper bound on this is a very, very important issue and that is what is the subject of error bound, which are very exciting area of research in convex optimization is all about.

So, let us take a careful look as to how we can do it. We will be our first step; we will claim that p star…

(No audio from 16:58 to 17:10)

We will prove this, for the given x this is p star is always bounded will over this number. For any x, this is always true. So, how do I do it? Now, for a fixed x once I fixed the x for any other y, let us look at the left hand side of the sorry right hand side of the expression for strongly convex function.

(Refer Slide Time: 17:48)



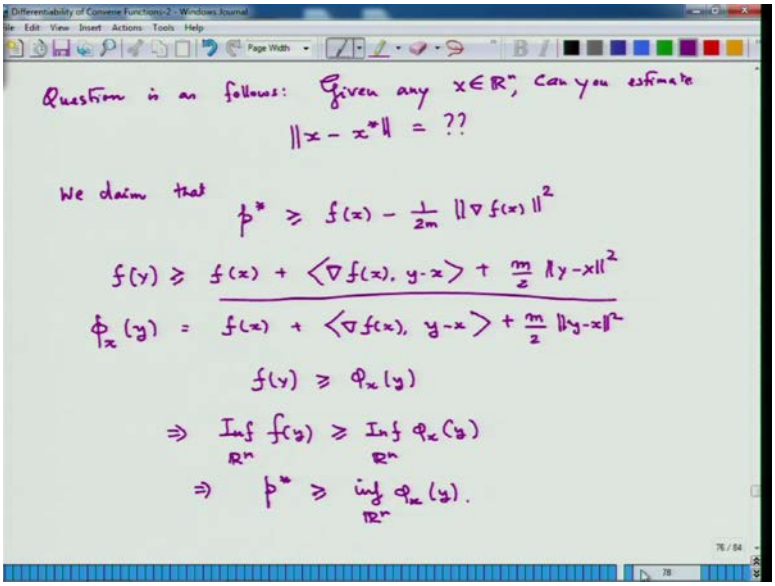
Now, of course, we have started in our definition of strongly convex function, we have started with the case where f is differentiable. I am not getting into the issue of when f is

not differentiable. So, you see the strongly convex function is always convex, because this is greater than equal to 0, so again this whole thing is bigger than this whole thing, but a convex function need not be strongly convex. So, example f(x) equal to x square is strongly convex, f(x) equal to x not strongly convex, f(x) equal to x to the power 4x is in R of course, in this particular cases my x is in R. So, x to the power 4 is strictly convex, but not strongly convex.

(No audio from 18:43 to 18:52)

So, here you see the differences.

(Refer Slide Time: 19:00)

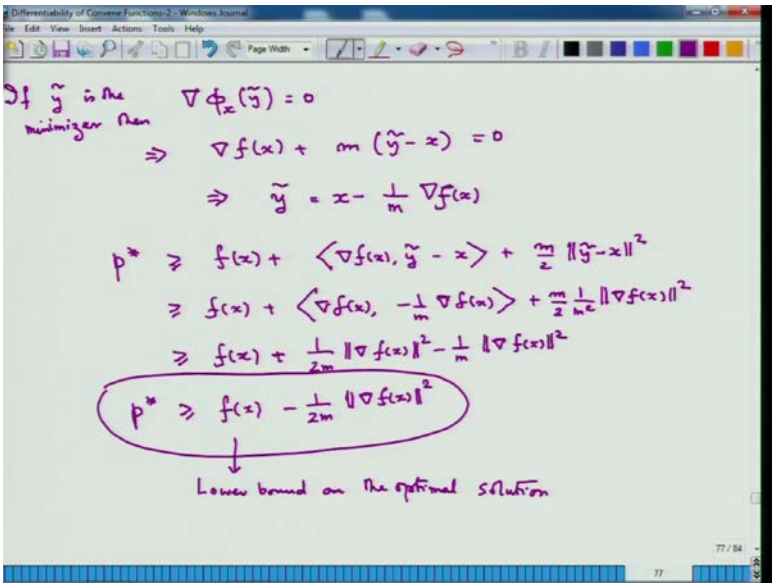


Now, look at this part. This can be viewed as a function of y for a fixed x, putting the reference as x. I am viewing this part as a function of y.

(No audio from 19:12 to 19:28)

So, f of y for a fixed x, this is true, and this would immediately imply infimum of f(y) over R n is infimum of phi x (y) over R n. Now, this is the convex problem. When this is anyway convex in y, it is clear, but this affine in y this part and this strongly convex in y, so this is the strongly convex function in y. So, this is what you will having. This is nothing but p star which we have already seen. So, p star is bigger than inf of R n phi x (y). So, our job now would be to find this infimum.

(Refer Slide Time: 20:25)



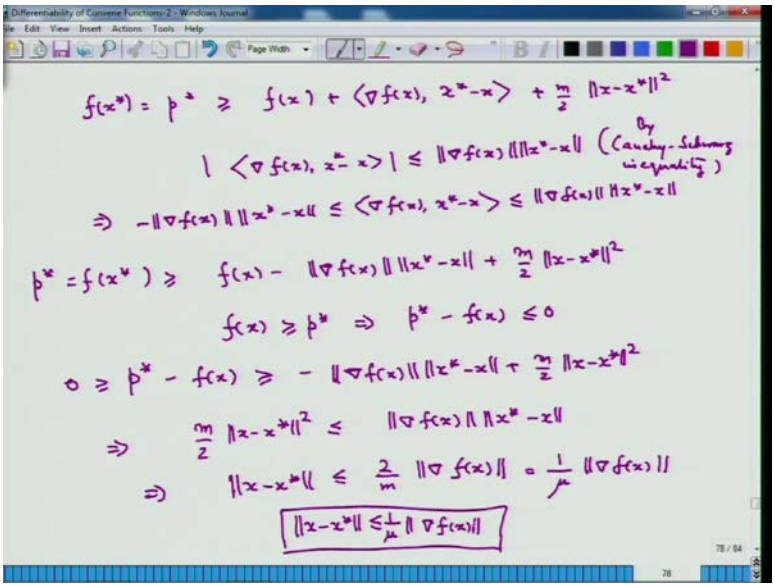
So, how what I would do is to optimize the function phi x (y) by taking gradient with respect to y right. And so if y tilde is the min, and there will be a min because this function is strongly convex, there will be a min just in the same sense. There will be a minimum of this function. Let y bar be that minimum, minimizer rather, than phi by tilde is 0. So, this could imply immediately, you take the gradient. See now the two goals are which makes the calculation look more easy. So, this would simply tell me that y tilde is equal to x minus 1 by m grad f(x). So, how do I do it? Now, observe that f of y bigger than of course, this is the infimum. So, any f of y would be bigger than this rather I would say sorry I make a make a mistake I have already written p star, so p star is bigger than the minimum value of this which is f of x y tilde minus x plus m by 2 y tilde minus x whole square.

If I put down the value of y tilde here, so I will have this bigger than f(x) plus grad f(x), what is this, y tilde is x minus so x will get cancel, so you will have a minus 1 by m grad f(x) plus m by 2 y tilde minus x is nothing but same as, so I will have 1 by m square. So, this would give me f(x) plus here I will from here I will have 1 by 2m and here I will have minus 1 by m. So, this would give me p star to be greater than f(x) minus 1 by 2m. So, you see that is exactly what we had discussed that what is we are what is what what we want to do. So… So, this is providing a lower bound on the optimal value. So, given any x star I can provide you a lower bound on the optimal value, but that into be the

infimum, infimum is p star. These are all lower bound; infimum of the function is p star; lower bound on the optimal value.

So, give me any x star I can give you immediately a rough idea of the… So, if beauty of strong convexity is that give me any f and give me any x, I will tell you here if I put as 2 mu it will become 4 mu basically; this is p star greater than f(x) minus 1 by 4 mu into norm norm of f(x) square. You will simply see that this this thing. See here I have used the fact that this inner product this is same is nothing but norm f(x) square norm grad f(x) square. So, give me any x then I can provide you a lower bound to p star that p star will never go below this value; p star might be much above this value, but it will never go below this value. We use soon be surprise to see that this fact would be used to really figure out how far and given x is from the original x star which will might not even know, we might not be able to figure out.

(Refer Slide Time: 25:02)



So, p star, but f(x star) is p star that is what we know. But p star is greater than by vary definition f of x plus grad f(x) x star minus x plus m by 2 norm x minus x star whole square. Now, here I will apply the Cauchy-Schwarz inequality, the Cauchy-Schwarz inequality says the following. Them… So, this would imply, this is the Cauchy-Schwarz by Cauchy-Schwarz inequality.

(No audio from 26:00 to 26:31)

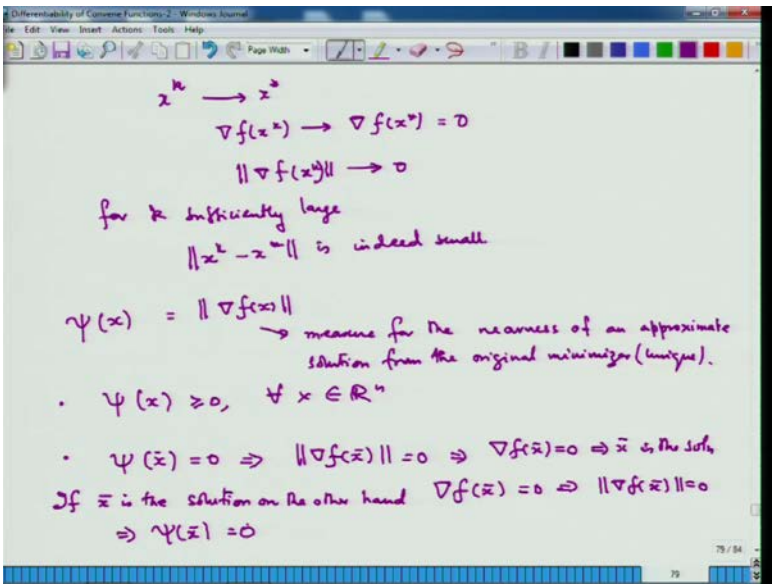
This is what you will get from the Cauchy-Schwarz inequality. And so here you can apply this fact and write…

(No audio from 26:39 to 26:49)

Norm x star minus x, but p star is the infimum. So, f(x) is bigger than equal to p star which would imply p star minus f(x) would be less than equal to 0. Now, from here I will get p star minus f(x) is less than equal to is greater than equal to minus grad f(x) norm x star minus x plus m by 2 norm x minus x star whole square.

Now, what do you get from here? You get the following. Now, this is this is what I know to know to be less than 0. So that would immediately imply that m by 2 norm x minus x star square is less than equal to… So, because x star is not equal to x then the things are obvious, x star is not equal to x, this is not equal to 0 - this norm; so, we can cancel out to write norm x minus x star is less than equal to 2 by m into grad f(x) norm grad f(x). Now, m is equal to twice of mu, so this would be nothing but 1 by mu times norm grad f(x). So, my error bound condition for this, is called the error bound. Sorry 1 by mu. Now, my question is the following; can I what what does this say that if you give me an x if I compute the grad f(x), the grad f(x) - norm of grad f(x) is giving me a measure of how near x is from x star.

(Refer Slide Time: 29:19)



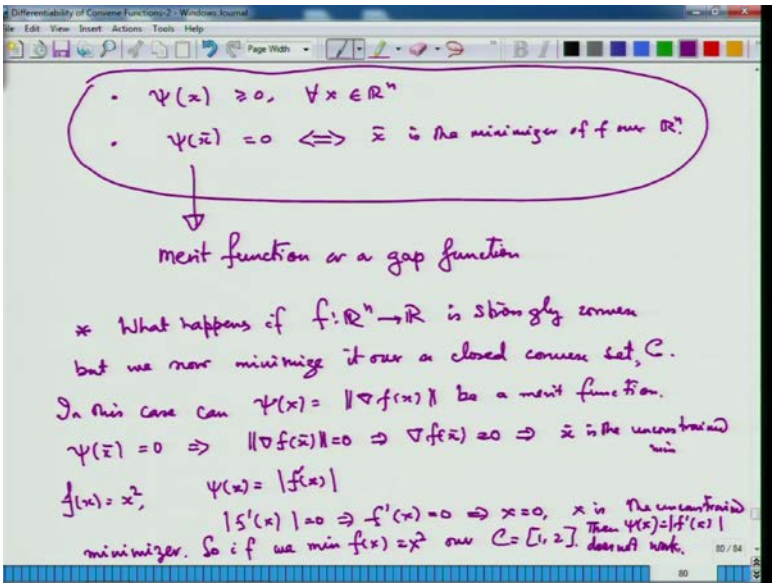
So, suppose I have a sequence x k going to x star. Now, on this sequence, because grad f(x, k) is a continuous function, because if n f is continuous grad f(x) is also continuous. Any convex function is not only differentiable; once it is differentiable it is continuously differentiable. But x star being the solution and its unconstraint case it will be 0. So, which means that norm of grad f(x, k) again by continuity of norm goes to 0, which means that for k sufficiently large is indeed small. So, which means that if you are actually on a sequence of points which is going to the infimum my error bound is actually telling me how close I am coming. So, this norm grad f(x), if I write down this as a function psi x measure for the nearness of an approximate solution measure for the nearness of an approximate solution of a strongly convex function from the original one, from the original unique solution - from the original unique minimizer, unique of course.

Now, this thing has a property; number 1 - psi of x is naturally greater than 0 for all x in R n is obvious this is norm. Number 2 - when psi x is equal to 0, this would imply norm grad of say psi of x bar is equal to 0, so it will be norm grad of x, it will imply grad f of x bar equal to 0 and since f is convex, it will imply that x bar is the solution. So, it would also if x bar is the solution on the other hand…

(No audio from 32:02 to 32:14)

If the x bar is the solution on the other hand then you always have grad of this is equal to 0, and this would immediately imply that norm of grad of f of x bar is equal to 0, and that would imply that psi of x bar is equal to 0. So, what are the properties that we have actually got?

(Refer Slide Time: 32:43)



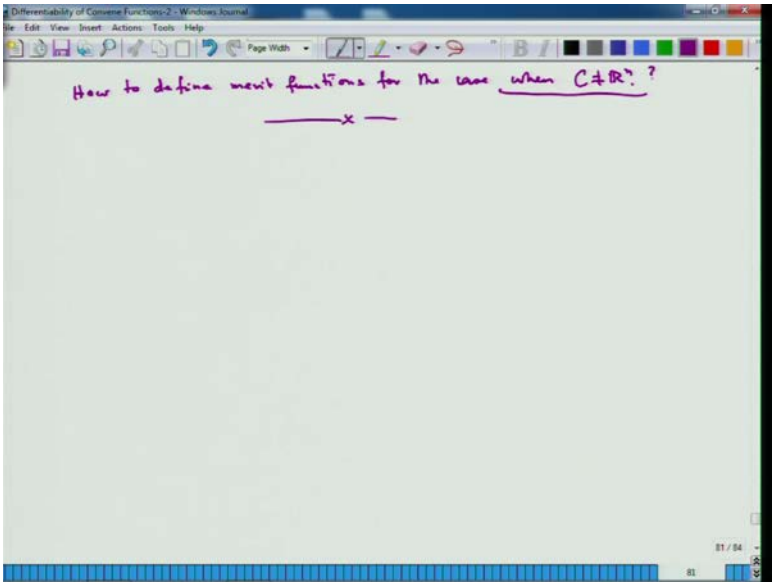
So, for psi we have got the property is that psi of x is greater than equal to 0 for all x in R n. As well as you have the property that psi of x bar is equal to 0 if and only if x bar is a solution of CP with f strongly convex or solution of or x bar is the minimum minimizer of f over R n, obliviously f is strongly convex, is the minimizer is the minimizer of f over R. Such type of functions at least for this, in this particular setting, if you if you can find the function like this which measures actually the distances from the original solution is called a merit function or a gap function.

The important question with which we will end our talk is the following. What happens if f is strongly convex, but we now minimize it over a closed convex set, C; note that now in in this case, can psi be a merit function, psi in the sense is psi is psi x equal to grad f(x) basically, psi x equal to grad f(x) be a merit function be a merit function. The answer is no, there is a difficulty, because if you say that psi x bar is equal to 0. This would imply which would imply which would imply that x bar is the unconstraint minima and not the minima on C.

For example, if you take the function just f(x) equal to x square, and define your psi x in this case would be the absolute value of sorry f dash of x or the grad f(x). So, f dash of x is equal to 0 would imply… The absolute value means f dash x is equal to 0 which would imply x is equal to 0, but x is the unconstraint minimizer. So, if we minimize f(x) equal to x square over C is equal to (1,2), then psi equal to psi x equal to mod f(x), then psi x is

equal to absolute value of f dash x does not work does not work. Because here the minimizer is that obtained at 1, x is equal to 0 is not the minimizer (( )), because x is equal to 0 is not even feasible. So, how do I define merit functions for…

(Refer Slide Time: 37:31)



The interesting question now would be how to define merit functions for the case when C is not equal to R n. I would stop of my talk here, because if I continue it would really go on for at least one half or more, because there is the lot of things, a lot of beautiful things come in when you try to answer this question. (( )) So, you try to create a merit function which will work, and then there lot of issues about that merit function will leads to another improved version of the merit function which really works in practice. So, we will not get into this, so we will end our talk quick, a hope that some of you will try to figure out this particulars, so called simple looking question. Thank you very much.