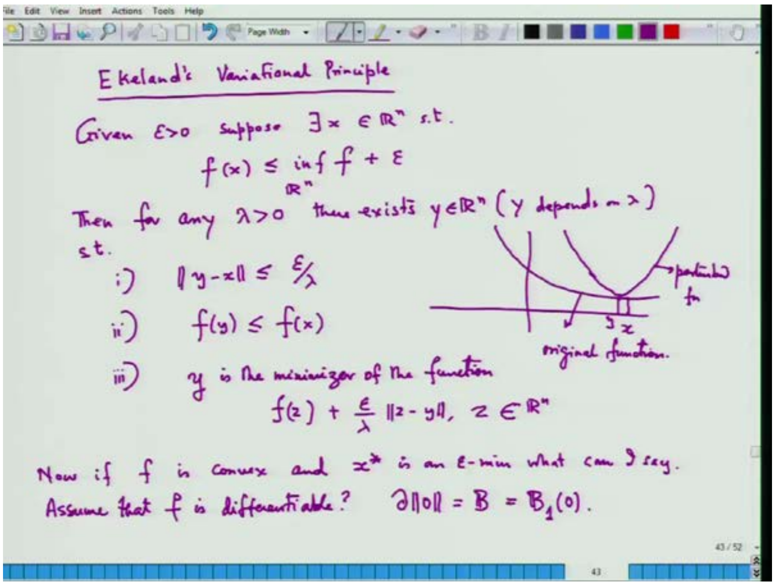
Convex Optimization

Prof. Joydeep Dutta
Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur

Lecture No. # 36

So, as we progressed towards the end of our course, here we as we told we had taking a little bit of break from a semi definite programming and conic programming and going into little bit of different stuff, where we are trying to analyze approximate solution; approximate solutions are exactly what you basically see in real computations. So, it is good to have a little bit of analysis of them.

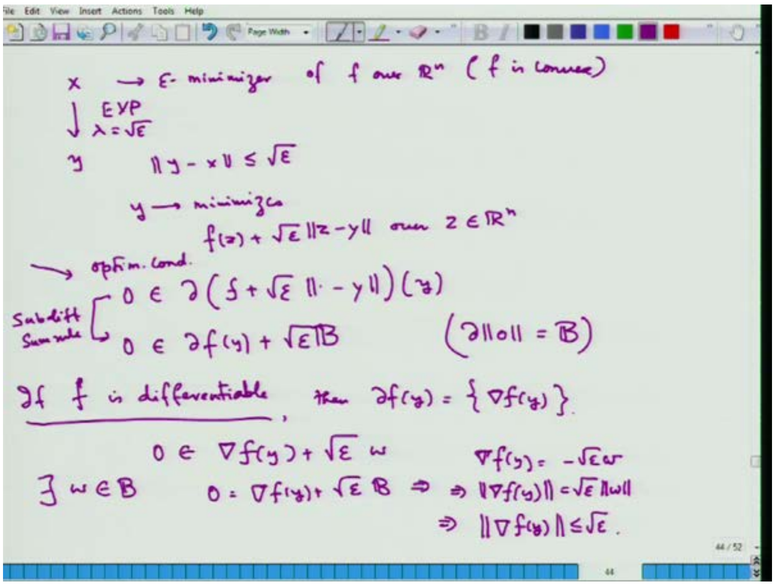
(Refer Slide Time: 00:37)



And one of the main tools that we will use in that is Ekeland's Variational principle, which I had outline in the last lecture, and you observe that here just just now, before I started the lecture, I change here this I had written this as x, this is nothing but a dummy variable, but you might possibly confuse it with this x, so I change it to z. So what it says is that if you have an aproxi… if the function is bounded below, and if there is an approximate minimizer, I can find another approximate minimizer near the current one, such that that approximate minimizer is an exact minimizer of the part of function. The problem with this part of function business is that this part of function is a non differentiable function, because addition of the norm norm term, and so that is what would disturb us; but that we need not bother much.

Here, the example is that of a convex function, so here what happens is that if say like E to the power minus such a function like this, has a this x is a global is a epsilon minimize, then y is a another epsilon minimizer near x, such that this minimizer is a part of function, which is also convex function, because you are adding a convex function with the convex function. So in our case, we are just bothered about convex function, so we I am just or diagrams etcetera all represented through convexity. So what happens if I have a… So what what can we do, what can we say more?

(Refer Slide Time: 02:27)



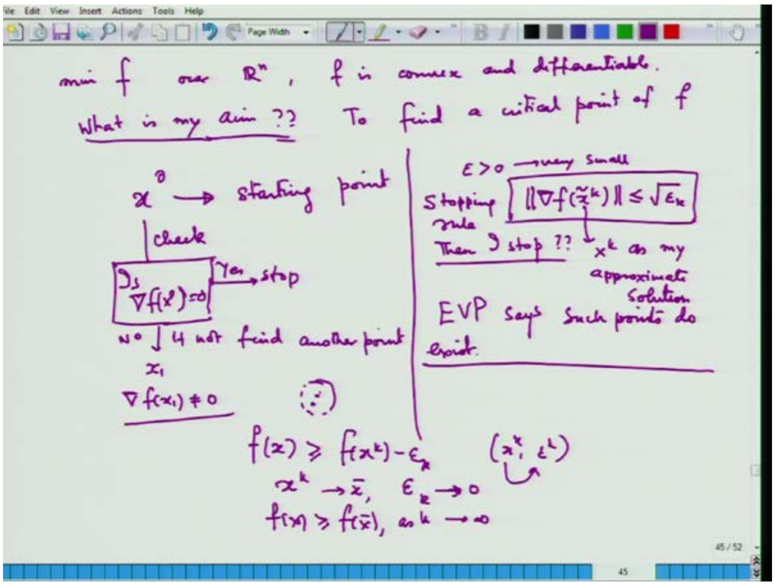
So, I have asked the question what would happen if f is differentiable; in this case observe that if f is differentiable, what I am having is that I will have a y, suppose x is an epsilon minimizer. So, corresponding to x is an epsilon minimizer of f over R n, where f it goes without saying is convex, and then what you have to realize is the fact that now, what I can figure out using the Ekeland's Variational Principle, which I write EVP that I will have y such that y minus x, so instead of y, let me just take lambda equal to root epsilon, so I will have y minus x is less than equal to root epsilon, epsilon by root epsilon, and y minimizes other strictly.

So, I am taking lambda equal to root epsilon in the in the EVP and that will give me, so this is what we already have known from EVP. Now if that is the case, I will apply the optimality condition, which would say that 0 element of del of f plus root epsilon at y, here because we are basically getting the norm value at 0, the sub differential would be nothing but the unit ball, so by applying the standard sub differential sum rule, we will have 0 element of del f of y plus root epsilon B, but because let us if we have taken f to be differentiable again, if f is differentiable, then we have del f of y equal to…

So, this condition becomes 0 element of… so there exist w, element of B, this is the way, it is written in literature B, such that now that del… again I am here, I am using the fact that del of non-zero is (( )), from here to here, I have applied the sub differential sum rule; here I have applied optimality condition, and you have applied sub differential sum rule, and then what would you have what would happen is once you look at this condition, would immediately realize I can write, so there must be a w in this, such that 0 is equal to grad f (y) plus root epsilon B, which would imply that minus grad f (y) or grad f (y) just is equal to minus put epsilon w.

So, it should imply that norm of grad f (y) is equal to is equal to norm of w, so we should imply that norm of grad f (y) is less than equal to root epsilon; epsilon is arbitrary small. So, what it says? There is a very, very important thing, it says a lot about algorithms; see when we are trying to minimize it differentiable convex function, we are essentially trying to find a a point y, so if I trying to find a point y such that grad of f (y) equal to 0.

(Refer Slide Time: 06:50)



So, if I am just minimizing f over R n, where f is convex and differentiable, so then what is my aim? Because every critical point is a global minimum, I have to find a critical point; sorry what is my aim not name; what is my aim? My aim is the following to find the critical point of f.

(No audio from 07:26 to 07:36)

So, I have to find a critical point of f, because I know that critical point of f is global minimize f, because f is convex and differentiable. So, the question would be how do find the… how do I find a critical point of f, what I do is, I try to start with the sequence of point x naught, a vector x naught, so this is a arbitrary starting point; this is how an algorithms are done; and then you check is if not, find another point, but you cannot go on like this, you might not even find a point x k, for which, you might not even find a point x k, for which f of x k would become 0. So the idea of the algorithm is that you fix up an very small epsilon greater than 0, and let me say my stopping criteria is following, that if I find an x k, such that the gradient of f of x k is less than root epsilon stopping rule, then I stop; and choose x k as my approximate solution, as my solution approximation or approximates solution it is not the exact solution, what I as my approximate solution, which has made me happy that is all.

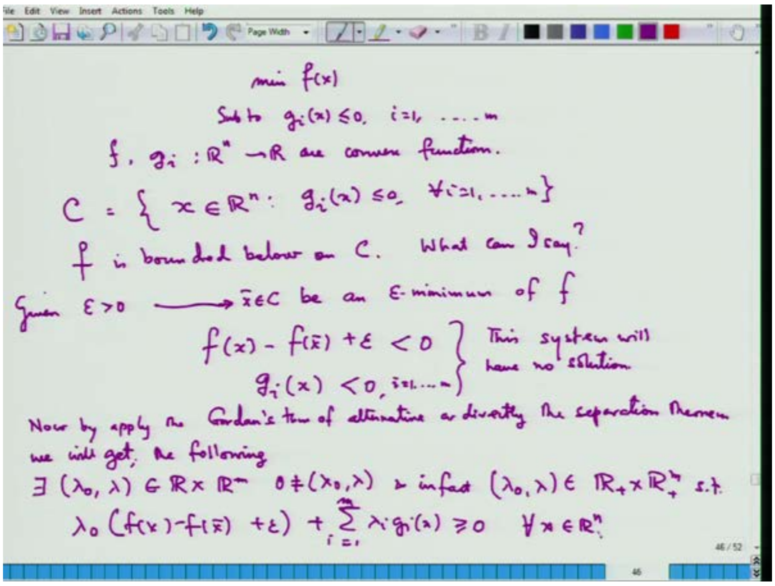
So, we are basically setting up a threshold barrier, and if something is going beyond the threshold barrier, we are stopping the algorithm; and this is algorithms are done, because you cannot keep on finding another point for which grad f of x k say there will be x 1 which is you are trying to find another point x 1, for which the grad of f (x 1) you keep on checking, but that is not really fair; if it is not 0, again you check; if it is 0, fine you stop; if it is not yes - stop; no - you have to do something and you have to continue this process, but how long? You might not even find any x 1, which is giving you 0, so that is why keep take epsilon greater than 0 very small, and make this as a stopping rule.

So what does what does… now the question is will I get a point x k which will actually satisfy this? That is what if my problem is bounded, because that is why the Ekeland's Variational principle says; EVP says such points do exists, that is this topic rule makes sense that is you would be able to find possibly a point, for which this would be less than root epsilon. Now that point, you might say that okay, you have generated one sequence of points how do you know that such a point, which would satisfy this would be on that sequence; form EVP you cannot say that there that would lie on a sequence.

But EVP says that if you take any approximates solution, then needs never would, there is another approximate solution, where such such a result would hold. So basically what would happen is that if I start with an epsilon solution, then what it shows, you can easily show that if I start with an epsilon case solution for example, I have a solution x k going to x bar, and each x k is an epsilon solution of f - epsilon k solution. So, I have this pair x k, epsilon k, so x k is epsilon k solution, so what an x k, suppose x k goes to x bar, and epsilon k goes to 0; then f of x would become greater than f (x bar) for all as k goes to infinity; so which means that if I choose points like this, then if we did x k is going to some x bar that x bar would be the actual solution, but now this point x k is an epsilon k solution.

So what would happen that for a k sufficiently large, x k would be very near f (x bar), and then for that particular epsilon k, I will get some other x tilde k, for which norm of f of x tilde k would be less than root epsilon, root epsilon k sorry here I should write epsilon k; see my indexes are all in the superscript, when I am writing the vector in sub script, when I am writing a scalar, so that is exactly what I have. So, I will be able to find a point where it will be stop, so there is a point around; so it is a quite obvious that may be because the function is continuous, if the grad f is also continuous, because it is the convex function f is continuous, and f is differentiable, grad f is also continuous, there will be hardly much difference between the two, the two norms. So, possibly that you can norm of grade f (x) could be also less than root epsilon k, so that is why EVP actually says that this stopping rules that we have designed or not really bad stopping rules, they are quite well on a… they can do pretty well in practice. Now the interesting part is that here we have assumed differentiability or sub differentiability etcetera, etcetera, ecetera.

(Refer Slide Time: 14:06)



Now what would happen we make the problem slightly complex that is instead of this, I am talking about minimize f (x) subject to… now where f and g i, all from R n to R are convex functions; and I have no information, what I know that this convex function f has is bounded below on the set on the set of feasible points. So, C set of feasible points given as set of all X in R n a such that g i x is less than equal to 0 for all x, for all i; suppose I am only given that fact that f is bounded below on C, the question is what can I say out? Now if such a situation arises, it means if it is bounded below on C, there is an x bar, so if I take any epsilon greater than 0, then correspondingly, there is an epsilon, there is an epsilon solution, let x bar B and epsilon mean, because it is bounded below, there will be on always an epsilon minimum of f; see whenever a function is bounded below, there is an epsilon x bar, which is in C; so I have to write more clearly that x bar element of C be…

Now, once you know this, so given epsilon greater than equal to 0, there would exist; there would exist x bar element of C, which is an epsilon minimum of f. So anyway x bar would exist, and we are just assuming that that x bar that there is an x bar, which is the epsilon minimum of f. Now what does this show? If x bar is an epsilon solution, then the following system sorry less than 0, this is greater than equal to 0, then x bar is an epsilon minimum. This… So this system has no solution, because if there is an x, which satisfies this for all i equal to 1 to m, and for which this happens, then that x breaks the fact that x bar is an epsilon minimum; so that would lead to a contradiction to the hypothesis, which says that x bar is an epsilon minimum of f, so this system will have no solution.

Then applying either the Gordon's Alternative theorem or Directly Separation theorem, we would have some important thing to do. So now, so these are convex functions all this, this; now by applying the Gordon's theorem of alternative or directly the separation theorem…

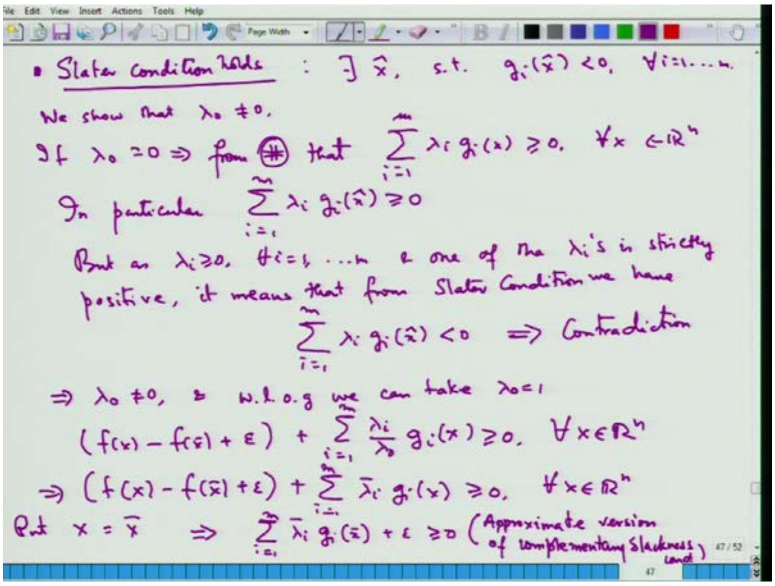
(No audio from 17:50 to 18:00)

A separation theorem, we will get the following.

(No audio from 18:16 to 18:28)

Now, we will get the following; so what would be the following that there would exist (lambda naught, lambda) element of R cross R m, and in fact not only R cross R m, if 0 naught equal to lambda naught and lambda, and they cannot the whole vector cannot become 0, and in fact lambda naught, lambda is element of R plus cross R m plus, such that it would mean at lambda naught into f (x) minus f (x bar) plus epsilon plus summation lambda i g i (x), i is equal to 1 to m, so this is true for all x in R m; so once this is done, let us see what can we say more.

(Refer Slide Time: 20:05)



Now, like anything in convex programming, one important condition that we always expect to hold is the interiority of the feasible set that the feasible set has an interior, so for that we assume that there exist an x star, for which g i (x star) is strictly less than 0, the Slater condition; so let us let this be an assumption for us, Slater condition holds. See, if this Slater condition holds, so there exist x hat, such that for all i. So once you know this, what would happen, which means what we will prove that we show that lambda naught is not equal to 0, because if lambda naught is equal to 0, then summation… So if lambda naught is equal to 0, it would imply from this previous equation, which I can call as star or hash, so it implies from hash that…

(No audio from 21:13 to 21:28)

Now you see one of these lambda is cannot be 0, because the whole vector lambda, lambda naught lambda, naught lambda cannot be 0, so one of this lambda is cannot be 0. Now this is true for every x, take this for this must be true for g i (x hat), so in particular…

(No audio from 21:46 to 21:56)

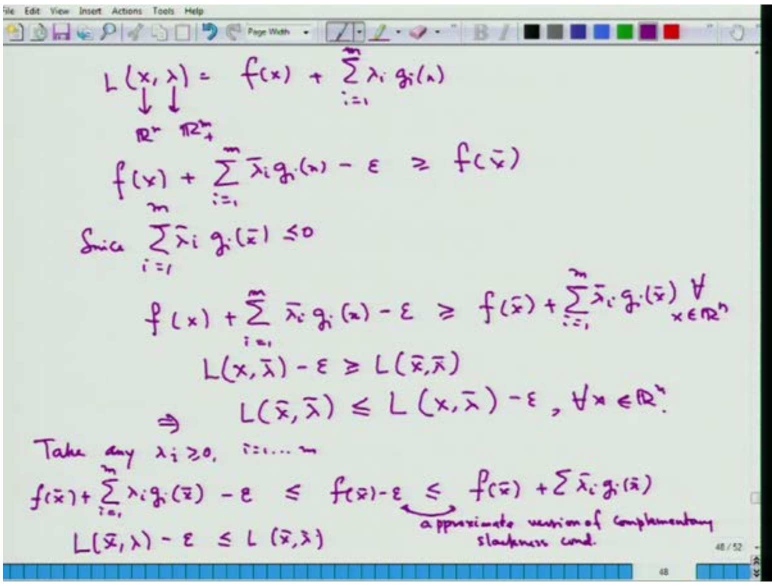
But as lambda is greater than equal to 0 for all i, equal to 1 to m, and one of the lambda i's is strictly positive, it means that one of the lambda is the strictly positive, it means that from Slater's condition Slater condition, we have… So these two are contradicting to each other, so what we have here is a contradiction. So, this implies that lambda naught is not equal to 0, and without loss of generality, we can take lambda naught equal to 1, just divide by lambda naught on both sides, that is all. So, what would happen is that so what what what would happen, so I will divide by lambda naught on both sides to have f (x) minus f (x bar) plus epsilon plus summation lambda i by lambda naught… Now, I will call this as lambda i bar, so I will have this would imply f (x) minus f (x bar) not x bar sorry this is x.

(No audio from 24:03 to 24:25)

This is what I have. Again, if I put put x equal to x bar, so this would imply immediately that summation i equal to 1 to m lambda i bar g i (x bar) plus epsilon is greater than equal to 0, this is an approximate version of complementarily slackness condition, we can say this as an approximate version of here, this is not 0. So it could be negative, but if I add epsilon to it, this thing, this becomes greater than equal to 0. So epsilon is such that in general, this has to be non non-zero, it should have been if epsilon was 0, it would be greater than equal to 0, and this is because x bar is positive, this would actually combined to give this equal to 0, but here why know that this is only strictly greater than equal to 0, what when x is equal to x bar, what I can say that adding this same epsilon I will get this whole thing to be greater than equal to 0. So it is called an approximate version of complementary slackness condition.

(No audio from 25:36 to 25:49)

(Refer Slide Time: 25:58)



Now keeping this aside, if I notice that the Lagrangian function is written like this, where this is in R n and this is in R m plus, then from the previous equation here, what would I get that is something one has to see, what would I get? I would get the following; from here I will get f (x) plus summation i is equal to 1 to m lambda i bar g i (x) minus epsilon is greater than equal to f of x bar, so it is, this inequality will lead meet to this. So, this will show what this will now since summation lambda i bar g i (x bar) is less than equal to 0. I can always add this to this, if I add a negative quantity, it will drop, the value would drop from f (x bar).

So, f (x) plus summation i is equal to 1 to m lambda i bar g i x minus epsilon is greater than equal to f (x bar) plus summation lambda i bar g i x bar. Once you know this, you know what you have got; you have got L (x,lambda bar) minus epsilon x greater than equal to L (x bar,lambda bar), so or in other words L(x bar,lambda bar) is less than equal to L (x,lambda bar) minus epsilon, this is true for all x in R n, this is also true for all x in R n.

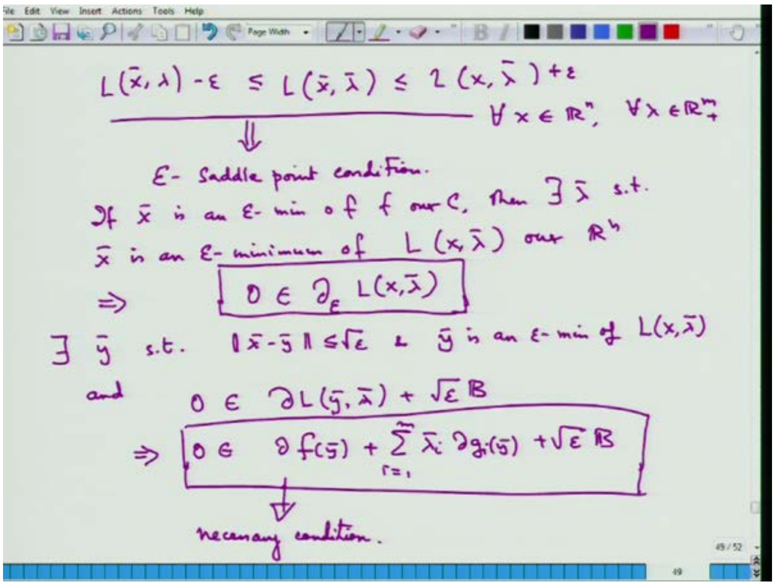
Now, so you have got something quite interesting from here, some interesting feature you have got, without any assumption of differentiability nothing; looks what you can call and some sort of one, some sort of strange looking saddle point thing; now again let me write down this equation; take any lambda i greater than equal to 0, then f (x) plus summation i equal to 1 to m, lambda i g i (x) sorry f (x bar) plus lambda i g i (x bar), because g i (x bar) is solution; so is an approximate solution, so x bar is in C, so this is less than equal to 0, so this whole thing is less than equal to 0, minus epsilon is less than equal to now this thing is less than f (x bar), f (x bar), because this is a negative quantity. So it will become f (x bar) minus epsilon, which is less than equal to f (x bar) plus summation lambda i bar g i (x bar).

Here is where we have applied, this point we have applied the approximate here to here, we have applied the approximate version of complementarily slackness condition.

(No audio from 29:54 to 30:07)

So, this is the point, we have applied the version of complementary slackness condition, and as a result of which, this will immediately show as the following, it will show that L (x bar,lambda) minus epsilon is less than equal to L (x bar,lambda bar).

(Refer Slide Time: 30:30)



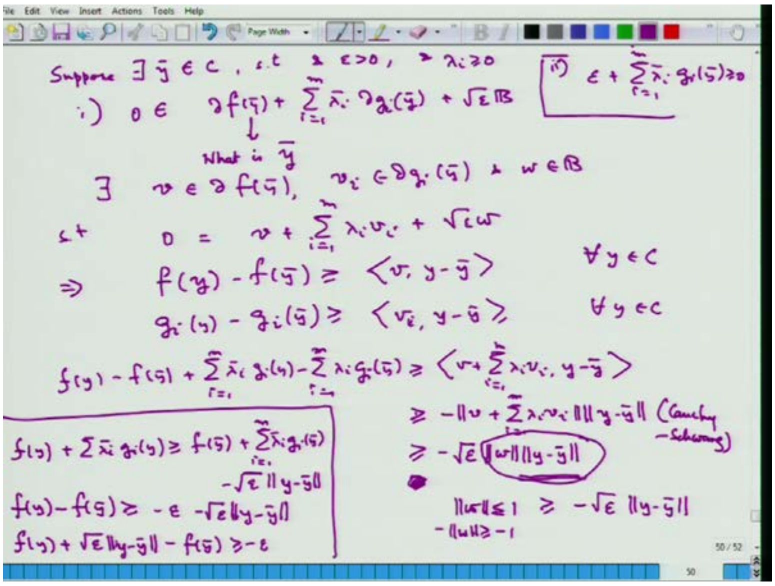
So some sort of saddle point result we have, what we have is that that whenever x bar is in epsilon minima, I could show the existence of a lambda bar in R m plus such that sorry here it was a plus epsilon, I made a mistake here plus epsilon, this was minus f (x), which went there sorry, so here it was a plus epsilon, so this would be a plus epsilon; so here it would be a minus epsilon, so kindly check this thing, because if you look at this thing, there is a plus epsilon here.

So now, what I got is something like this; for all this is true, for all x element of R n, and for all lambda element of R m plus, so this is what is called the epsilon saddle point conditions. Let me tell you this is a necessary condition, this does not give you sufficiency; anything which satisfies the epsilon saddle point condition is not necessarily an epsilon optimal point. We will come to that slightly later, it is pretty slightly interesting this part is, and so it is just the necessary condition that if I have a epsilon solution, this is what I will get.

Now, if look at this condition, what does it says that so if x bar is an epsilon mean of f over C, then there exist lambda bar such that x bar is an epsilon mean minimum of L (x,lambda bar) over R n; this would simply mean that 0 would element of which, so this is this is an optimality condition; this is a necessary optimality condition. Now, can you write something better from here is not really possible, because we have not studied sum rules for this sort of sub differentials, but what we can say is that there exist another y bar, so such that norm x bar minus y bar is less than equal to root epsilon, y bar is an epsilon mean of lambda x of L (x, lambda bar), and 0 is element of del L (x, lambda bar) del is with respect to x sorry del at y bar plus root epsilon norm B.

So this implies again by applying the sum rule 0 element of del of f (y bar) plus summation lambda i bar del of g i y bar plus root epsilon norm B. So what I have written down here is actually a necessary condition; now if there is an y bar or x bar, which satisfies this, the question is will that y bar be an epsilon solution, that is quite a natural question to ask so that is something, we should also like to investigate. We would like to remind that if this condition, this condition is a necessary condition, it is really not a sufficient condition; so if you figure out something like this, you look at not only at the sub differentials themselves, but you look at some at some flattened set of this particular sub differential sum, and then you found 0 to be there, 0 may not be exactly in here, but 0 is there, then what would happen; so, then can you tell something so that would be an important thing that let us try to figure out what would happen, if I have this.

(Refer Slide Time: 35:34)



So what I know is suppose there exist a y bar, element of C, I have missed something here, I have missed something; here you cannot say right, I did not missed anything here, it is fine. Suppose I have something like this, I have written the g so sorry I thought that I missed the g here, so now I have written this now, I suppose have an y bar element of C, which is such that an epsilon greater than 0, such that I know that 0 is not exactly in this set, that is del f (y bar) plus summation lambda i bar, there would exist y bar element of C epsilon greater than 0, and lambda i bar greater than equal to 0 such that…

So it is not exactly in this set, but in this set, the flattened set; then what can I say about f; what is what is y bar right; then if you try to figure this out, you really have to apply the convexity of these things; so let me try to figure it out; so what what does this mean, so there exist v element of del f (y bar) v i element of del g i (y bar) and w element of B bar, such that 0 is equal v plus summation lambda i v i plus root epsilon w.

Now this would imply by very definition of convexities or in the sub gradient, that for any other y in C f (y) minus f (y bar)

(No audio from 37:36 to 37:58)

Now you see this condition that we have derived, we have derived this condition from the saddle point results, but the saddle point results is not just this, the saddle point result also has this additional fact that we also get that summation i is equal to 1 to m, lambda i g i (x) bar this one, lambda i bar greater than equal to 0, this is something we always have, so this is something we have to keep in mind; so we not only take this condition, we also add this condition, so this is one condition we have plus, this condition is condition one, and also let us take this approximate version of complementary slackness, this is greater than equal to 0. I think this line is getting so let me write it properly; so it is epsilon plus summation lambda i bar g i (y bar) is greater than equal to 0.

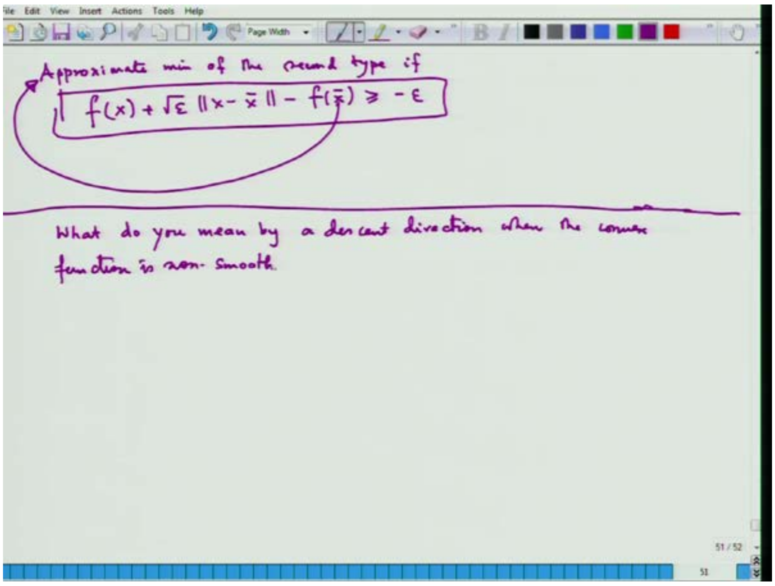
So, now what I do is do the same sampling, so I have f (y) minus f (y bar) plus summation lambda i bar g i (y) i equal to 1 to m minus summation lambda i g i (y bar), i is equal to 1 to m is greater than v plus summation lambda i v i, i equal to 1 to m y minus y bar; but this thing is by Cauchy Schwarz inequalities greater than norm v plus summation i is equal to 1 to m by Cauchy Schwarz, which I expect all of you to know, those who are viewing it

So this is nothing but root epsilon w, so it will become minus of right this is what you will finally get; so basically, because norm w is greater than equal to 1; here what we if we had written here as this is equal to minus root epsilon w, so it will become minus of root epsilon norm w norm y y bar, so from here you will see that I am unable to here this norm y minus y bar term comes in, and then I am unable to adjust the thing to get something that as to make y bar becoming an epsilon solution

So, because of this term, I am unable to make an adjustment, because norm w I know only is less than equal to 1, I do not know what would happen if it is something else right. so what minus norm of w is greater than equal to minus 1; so from here what I can do at the maximum is this is greater than minus root epsilon of norm y minus y bar, this is what I can have. So what I can have from here is as follows that f (y) plus summation lambda i bar g i (y), this is greater than equal to f (y bar), because summation lambda i g i (y bar) lambda i bar, g i y bar plus root epsilon y minus y bar, minus root epsilon y minus y bar.

Now this is of course, again this is this whole thing is less than f (y), so I can write f (y) minus f (y bar) is bigger than what did I have, what what did I have was the following, I had that epsilon, because of that result, I had this thing is greater than minus epsilon, this whole thing; so I will have minus epsilon minus root epsilon norm y minus y bar. So I can write this as f (y) plus root epsilon norm y minus y bar, f (y bar) is greater than minus epsilon, but here if you are in case of y, put y equal to y bar, this part will vanish; so I can also write here root epsilon norm plus root epsilon norm y minus y bar. So what it shows is that this is some sort of a different sort of a minima, one can also define and approximate minima of this form, so we can say an approximate minima of the second type.

(Refer Slide Time: 44:09)



If you have f of x plus root epsilon x minus x bar minus f (x bar) is bigger than minus epsilon, then x bar is an approximate minima of the… so x bar is an approximate minima of the second type. So the condition that I got here here is not really a sufficient condition, it is a necessary condition; so we will not discuss much more into this, we will go into other aspects; so tomorrow we will discuss a very interesting topic of what is called what do you mean by a descent direction, when the convex function is non smooth, so this will lead to algorithms using sub differentials. Thank you.