Convex Optimization

Prof. Joydeep Dutta

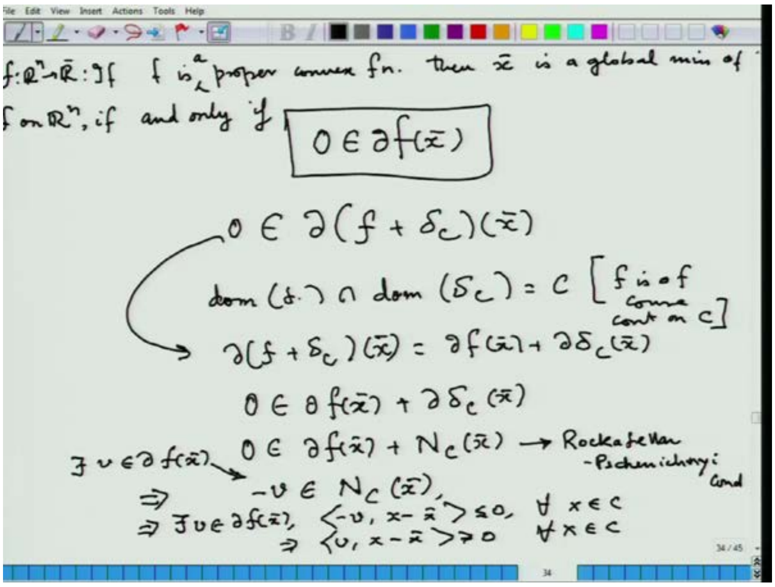
Department of Mathematics and Statistics

Indian Institute of Technology, Kanpur

Lecture No. # 10

Good evening viewers, once again we are back to this course on convex optimization.

(Refer Slide Time: 00:24)



$f:\mathbb{R}^n \to \bar{\mathbb{R}}$: If f is a proper convex fn. then $\bar{x}$ is a global min of f on $\mathbb{R}^n$, if and only if $\boxed{0 \in \partial f(\bar{x})}$

$$0 \in \partial(f + \delta_C)(\bar{x})$$

$$\text{dom}(f) \cap \text{dom}(\delta_C) = C \quad [f \text{ is of conve cont on } C]$$

$$\partial(f + \delta_C)(\bar{x}) = \partial f(\bar{x}) + \partial \delta_C(\bar{x})$$

$$0 \in \partial f(\bar{x}) + \partial \delta_C(\bar{x})$$

$$0 \in \partial f(\bar{x}) + N_C(\bar{x}) \rightarrow \text{Rockafellar-Pschenichnyi: Cond}$$

$\exists v \in \partial f(\bar{x})$

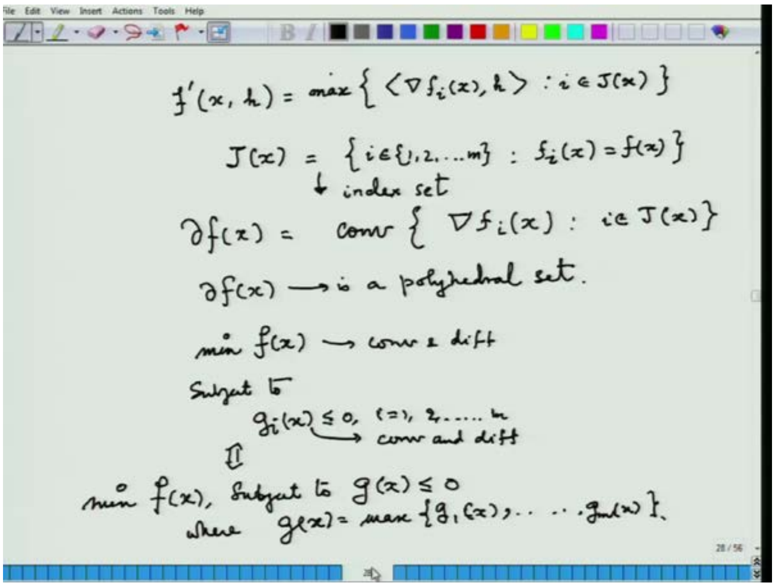
$\Rightarrow \quad -v \in N_C(\bar{x})$,

$\Rightarrow \exists v \in \partial f(\bar{x}), \ \langle -v, x - \bar{x} \rangle \leq 0, \ \forall x \in C$

$\Rightarrow \langle v, x - \bar{x} \rangle \geq 0 \quad \forall x \in C$

Yesterday if you had remembered, we had just spoken about these facts that we have proved, what is called the Rockefeller Pschenichery condition, which is a necessary and sufficient optimality condition for a convex function f to be minimized over convex f c. Of course we have we have proved that this is necessary and the sufficiency is left as a homework to you, which you can do; today we are going to go back a bit more, and use this power of the convex calculus, that is how we will see, how we can exploit the max function. We had spoken yesterday about the max function, if I you can just go back here.

(Refer Slide Time: 01:03)



$$f'(x, h) = \max\{\langle \nabla f_i(x), h\rangle : i \in J(x)\}$$

$$J(x) = \{i \in \{1, 2, \ldots m\} : f_i(x) = f(x)\}$$

↓ index set

$$\partial f(x) = \text{conv}\{\nabla f_i(x) : i \in J(x)\}$$

$\partial f(x) \longrightarrow$ is a polyhedral set.

$\min f(x) \longrightarrow$ conv & diff

Subject to

$g_i(x) \leq 0, \; i = 1, 2 \ldots\ldots m$ → conv and diff

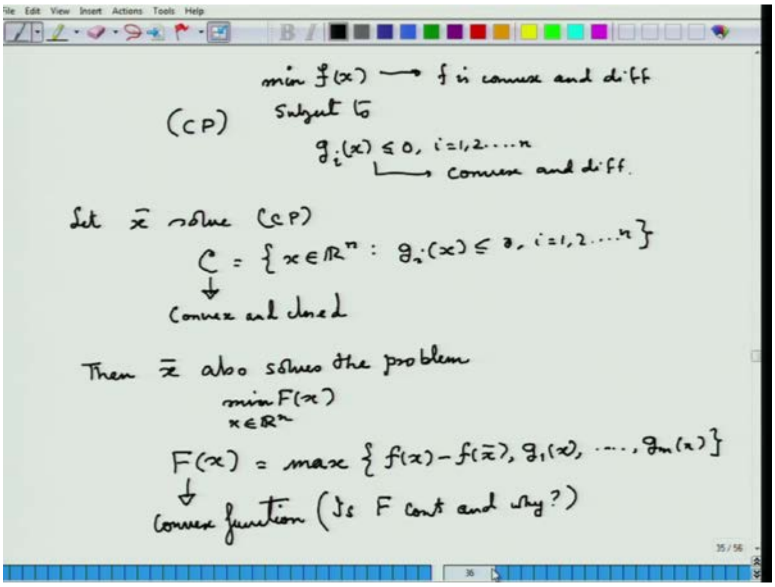
$\Updownarrow$

$\min f(x)$, Subject to $g(x) \leq 0$

where $g(x) = \max\{g_1(x), \ldots\ldots g_m(x)\}$.

Here, we have spoken about the sub differential of a max function, where each of the individual f’s, which make up the max function that is max of f i x, f i x, that is this f 1 x, f 2 x, f m x each are convex and differentiable; and this function f is also convex, which is already well known, which we have spoken earlier. Then the sub differential, just a moment, is this.

(Refer Slide Time: 01:45)



$\min f(x) \longrightarrow f$ is convex and diff

(CP) Subject to

$g_i(x) \leq 0, \; i = 1, 2 \ldots n$ → convex and diff.

Let $\bar{x}$ solve (CP)

$$C = \{x \in \mathbb{R}^n : g_i(x) \leq 0, \; i = 1, 2 \ldots n\}$$

↓

Convex and closed

Then $\bar{x}$ also solves the problem

$$\min_{x \in \mathbb{R}^n} F(x)$$

$$F(x) = \max\{f(x) - f(\bar{x}), g_1(x), \ldots\ldots, g_m(x)\}$$

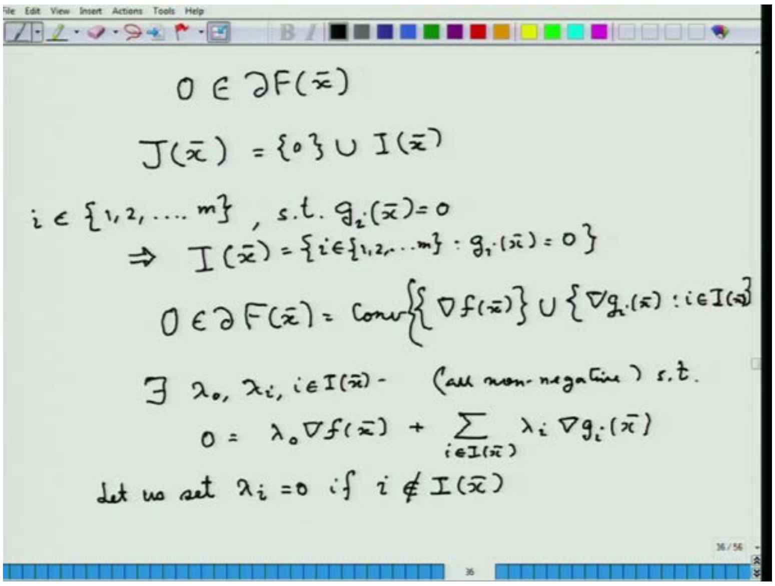
↓

Convex function (Is F cont and why?)

So, today we are going to see that if I have to minimize a convex function f x over x element of c, but this time my c is defined by inequality constraints, and each of this constrains are themselves convex. So, I assume that this is - f is convex and differentiable and g i (x) are also convex and differentiable. Let me assume that this has a minimum, let x bar so this is my convex programming problem CP, and let x bar solve CP. So, x bar is feasible that it satisfies all g i x bar is less than 0 and for every x, which satisfies this constraints f of x is bigger than f x bar that is the meaning of x bar solves CP in the global sense of course, for a convex problem there is no local minimum.

Of course, you can convince yourself that this set C, which is the feasible set of the programs of the programming problem CP is a convex set; once I know that all the function g i is the convex. In fact, it is also a close set, because each of the g i's are defined from R n to R, and any function defined from R n to R is any convex function defined from R n to R is a continuous function. Now, so what information I have about this is convex and closed; now, once I know this facts, then I look into a following problem. So, if x bar is solving this problem CP, then x bar also solves mean f (x), subject to x element of R n, where F capital F of x is the convex function given by max of f (x) minus f x bar g 1 (x) g m (x), you can try out again as a homework to prove this fact that whenever x bar solves this problem x bar will solve this problem. Now, of course, all of these are convex functions. So, capital F is a convex function. So, as homework you figure out is F continuous and why?

(Refer Slide Time: 05:05)



$$0 \in \partial F(\bar{x})$$

$$J(\bar{x}) = \{0\} \cup I(\bar{x})$$

$i \in \{1, 2, \ldots m\}$, s.t. $g_i(\bar{x}) = 0$

$$\Rightarrow I(\bar{x}) = \{i \in \{1, 2, \ldots m\} : g_i(\bar{x}) = 0\}$$

$$0 \in \partial F(\bar{x}) = \text{Conv}\{\{\nabla f(\bar{x})\} \cup \{\nabla g_i(\bar{x}) : i \in I(\bar{x})\}$$

$\exists\ \lambda_0, \lambda_i, i \in I(\bar{x})$ - (all non-negative) s.t.

$$0 = \lambda_0 \nabla f(\bar{x}) + \sum_{i \in I(\bar{x})} \lambda_i \nabla g_i(\bar{x})$$

Let us set $\lambda_i = 0$ if $i \notin I(\bar{x})$

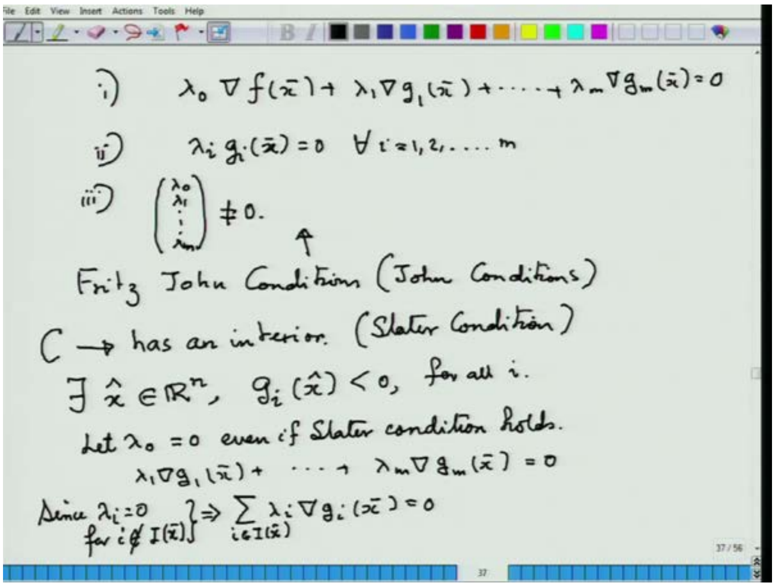
Now once I know this, I can immediately write down the optimality condition that 0 belongs to del of F of x bar; now F is a max function, then I can calculate the optimality conditions by applying the calculus rule for max functions. Now, once I know this, what shall happen; how do I now calculate the J x bar; set corresponding to F, so corresponding to my function F, how do I calculate the index set J x bar; in order to do so, I look into this fact that consider all index i, let us consider all index consider i, which belongs to any one of this indexes such that g i x bar is equal to 0.

So, if you take some i, which for which this is satisfied, these constraints are called constraints, which are activate x bar. So, this would imply one important phenomena, important thing that I will collect corresponding to x bar, all the indexes I belonging to this index set that is this is one among the constraint indexes such that g i x bar. Now, if you look at this function, if I put F x equal to F x bar, this becomes 0; and for all the active indexes that is for all I belonging to the I x bar, I would immediately have g i x bar are also equal to 0.

So, when x is equal to x bar and for all other x is g i (x) is strictly less than 0. So, f of x bar is actually is 0. So, this is something have to be very carefully noting down that F of x bar is actually 0, and this 0 is achieved at this point right, which I can write as g 0 if you want g 0 (x). So, this consists of J (x) consists of the index 0 as well as the active index set corresponding to x bar. So, J x bar consists of the index 0, union I x bar z, which is this. Once I know j x bar and I knowing that all these are differentiable, I know that 0 element of del F x bar, which is equal to the convex (( )) of grad of f x bar, this particular set union grad of g i x bar such that i is belonging to I x bar. So, which means you have to… So, 0 must be an element of the convex (( )) of these elements of these sets. So, there would exists some lambda naught, lambda 1, lambda 2 dot, dot lambda m, all non-negative that is all greater than or equal to 0; such that you have 0 equals to lambda naught grad of f x bar plus summation i belonging to I x bar lambda i not I i I here I should not put lambda 1 lambda 2 lambda m here I should put lambda i with i belonging to I x bar that is my correct sense lambda i grad g i x bar.

Now, let us set lambda i is equal to 0, if i is not in I x bar, that is whenever g i x bar is equal to 0, lambda i would be equal to 0. So, now, I can write bringing this fact in a much simpler way. So, I can now put in lambda is to be 0, I can just put 0, 0, 0 for all the i s, which are not in I x bar, but then I have to account for this behavior of lambda i. So, there must be some additional condition that lambda is such that lambda i into g i x bar would be equal to 0.

(Refer Slide Time: 10:21)



i) $\lambda_0 \nabla f(\bar{x}) + \lambda_1 \nabla g_1(\bar{x}) + \cdots + \lambda_m \nabla g_m(\bar{x}) = 0$

ii) $\lambda_i g_i(\bar{x}) = 0 \quad \forall i = 1, 2, \ldots m$

iii) $\begin{pmatrix} \lambda_0 \\ \lambda_1 \\ \vdots \\ \lambda_m \end{pmatrix} \neq 0.$

$\uparrow$

Fritz John Conditions (John Conditions)

$C \rightarrow$ has an interior. (Slater Condition)

$\exists \hat{x} \in \mathbb{R}^n, \; g_i(\hat{x}) < 0,$ for all i.

Let $\lambda_0 = 0$ even if Slater condition holds.

$\lambda_1 \nabla g_1(\bar{x}) + \cdots + \lambda_m \nabla g_m(\bar{x}) = 0$

Since $\left.\begin{array}{l} \lambda_i = 0 \\ \text{for } i \notin I(\bar{x}) \end{array}\right\} \Rightarrow \sum_{i \in I(\bar{x})} \lambda_i \nabla g_i(\bar{x}) = 0$

So, these would lead to following condition that lambda naught grad f x bar plus lambda 1 grad; so, we have two conditions now grad g 1 x bar. So, I cannot just arbitrary say that all the lambda corresponding to lambda i is corresponding to g i x bar strictly less than 0 is 0; once I have assumed that to give this full expression, then automatically a condition arises and that condition is at lambda i times, g i x bar is equal to 0 that is both of them cannot have strict inequality at the same time that is this cannot be strictly greater than 0 this cannot be strictly less than 0, this fact has to be always maintained by the Lagrangian multipliers or the Kuhn-Tucker multiplier has will soon call them.

But if you observe here, there is also one important thing, which we have not stated; now these lambda naught lambda are elements form a convex (( )). So, an important thing that one should have stated here was that lambda naught plus summation i element of I x bar lambda i is equal to 1. So, basically 0 belongs to the convex (( )) set, which comes from the max function. Now, what does this mean? Each of them are greater than equal to 0, their sum is equal to 1. So, all of them cannot be 0 at the same time. So, for an of course, i not element of I x bar, we have taken everything to be 0. So, what I would also have the third condition is that the vector lambda naught, lambda 1, lambda m is not a 0 vector; this is a very, very important condition and the condition that you get here is called the

Fritz John condition or the John condition; Fritz John conditions or the just the John conditions.

So, Fritz John presented his condition where back in 1948 and he submitted it to the (( )) channel of mathematics, which was rejected and then published in conference proceeding, later on Karush-Kuhn-Tucker condition or the Kuhn-Tucker condition came from here in 1951. Now, let me assume something additional, which I have not assumed earlier, let me assume this fact that this set C, the feasible set has an interior. So, this is my additional restriction, why I want to put this additional restriction? Note that I have said lambda naught, lambda 1, lambda m, the whole vector is not equal to 0; but I did not say that lambda naught need not be 0; lambda naught could be 0; and if lambda naught is 0, then we are in a very, very bad situation that we have a problem where lambda naught is 0 and grad f, which is gradient of the objective function it place no role in the computation, as a result of which things might not turn out as it as you want to want it to.

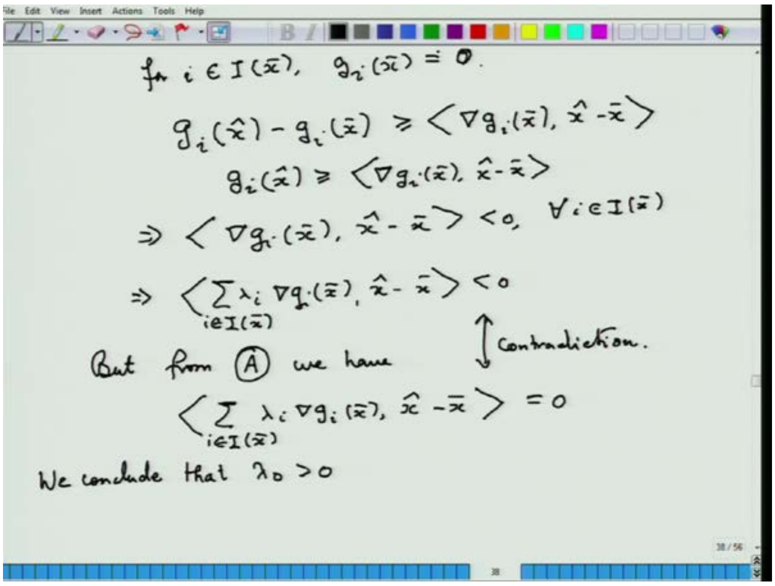
So, the representation of f goes immediately from the optimality condition, which is not a fair thing. So, Kuhn and Tucker in 1951 impose certain conditions, which are not now called the Kuhn-Tucker condition qualification, which we are not going to (( )), but an important assumption was given in 1952 by Slater - Slater condition, which says C has an interior the feasible set, which means there would exists an x hat.

So here, why this condition is imposed; that condition is imposed to stop lambda not from becoming 0. So, in the John conditions, which is this three set of three conditions; lambda naught could become 0, stopping the representation of f from the expression in f from the optimality conditions in order to stop that, that is in order not to make the optimality conditions look abnormal, you have this additional restrictions to stop lambda naught for becoming 0. So, interiority means there exists in x hat in R n such that g i of x hat; now once I know this let me make an assumption like this; and I will then show that lambda naught cannot be 0.

Now, let lambda naught p equal to 0; even if Slater condition holds; now once you do this, when once would lambda naught is 0, you will have lambda 1 grad g 1 x bar plus lambda m grad g m x bar is equal to 0. Now, observe that lambda 1, lambda 2, lambda m cannot be all 0, because lambda naught is 0, so one of the one of the among the, this vector is non-zero. So, if lambda naught is 0, so the non-zero vector, non-zero component must lie among lambda 1 to lambda m. Now, we know that whenever i is not in I x bar, lambda i's are anyways 0.

So, I can write this as lambda i times gradient of g i x bar. So, i element of I x bar you see, because lambda since lambda i is equal to 0, for i not in I x bar, it would imply that this is equal to 0. Now, the non-zero components must lie among these i belonging to I x bar, because when i is not in I x bar anyway lambda is 0. So, we have lambda naught 0 and all these are lambda is 0. So, remaining part the non-zero component must appear. So, there is a nonzero component here.

(Refer Slide Time: 17:45)



for $i \in I(\bar{x})$, $g_i(\bar{x}) = 0$.

$$g_i(\hat{x}) - g_i(\bar{x}) \geq \langle \nabla g_i(\bar{x}), \hat{x} - \bar{x} \rangle$$

$$g_i(\hat{x}) \geq \langle \nabla g_i(\bar{x}), \hat{x} - \bar{x} \rangle$$

$$\Rightarrow \langle \nabla g_i(\bar{x}), \hat{x} - \bar{x} \rangle < 0, \quad \forall i \in I(\bar{x})$$

$$\Rightarrow \Big\langle \sum_{i \in I(\bar{x})} \lambda_i \nabla g_i(\bar{x}), \hat{x} - \bar{x} \Big\rangle < 0$$

$\updownarrow$ Contradiction.

But from $\boxed{A}$ we have

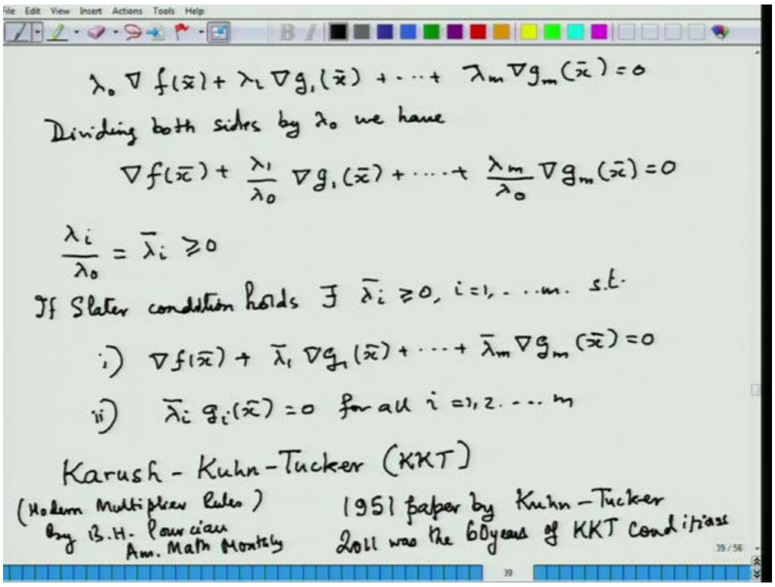
$$\Big\langle \sum_{i \in I(\bar{x})} \lambda_i \nabla g_i(\bar{x}), \hat{x} - \bar{x} \Big\rangle = 0$$

We conclude that $\lambda_0 > 0$

Now, how can I use the Slater condition? Let us see; now for any i element of I x bar, g i x bar is strictly less than 0; now look at the convexity equation, when functions are differential and look at this for this particular pair x hat in x bar, x hat is the point where the Slater condition is actually getting satisfied. So, this is bigger than gradient of g i x bar, x hat minus x bar. Now, g i x bar is for sorry I made a mistake; when g i is i is in I x bar g i x bar is equal to 0. So, g i x bar is 0, g i x hat is… So, what remains here in this expression; now g i x hat is strictly less than 0. So, what I get from the above is gradient of g i x bar x hat minus x bar is strictly less than 0, for all i element of I x bar.

Now, among these elements, there is one lambda, which is non-zero at least to 1. So, if I multiply by that lambda with the corresponding, this corresponding that g i x hat minus x bar that would also remain to be a strictly negative quantity. So, that would imply in general, when a multiply by the lambda for lambda i for each corresponding I in sum up what I will get is summation lambda i gradient of g i x bar i element of I x bar x hat minus x bar is strictly less than 0. This is in contradiction with the fact that here I have this is equal to 0. So, if I take inner product with any vector, if I take the inner product of any vector with the 0 vector that will give me 0; so, but if I put this condition as the condition A, but from condition A we have, so here is a contradiction, here is a contradiction. Now, once you have this contradiction, so you declare that your initial hypothesis that lambda naught, lambda 0 is 0 is wrong and so, we conclude that lambda naught is strictly greater than 0.

(Refer Slide Time: 21:30)



$$\lambda_0 \nabla f(\bar{x}) + \lambda_1 \nabla g_1(\bar{x}) + \cdots + \lambda_m \nabla g_m(\bar{x}) = 0$$

Dividing both sides by λ_0 we have

$$\nabla f(\bar{x}) + \frac{\lambda_1}{\lambda_0} \nabla g_1(\bar{x}) + \cdots + \frac{\lambda_m}{\lambda_0} \nabla g_m(\bar{x}) = 0$$

$$\frac{\lambda_i}{\lambda_0} = \bar{\lambda}_i \geqslant 0$$

If Slater condition holds $\exists\ \bar{\lambda}_i \geqslant 0,\ i = 1, \ldots m.$ s.t.

i) $\nabla f(\bar{x}) + \bar{\lambda}_1 \nabla g_1(\bar{x}) + \cdots + \bar{\lambda}_m \nabla g_m(\bar{x}) = 0$

ii) $\bar{\lambda}_i\, g_i(\bar{x}) = 0$ for all $i = 1, 2 \ldots m$

Karush - Kuhn-Tucker (KKT)

(Modern Multiplier Rules) by B.H. Pourciau Am. Math Monthly

1951 paper by Kuhn-Tucker

2011 was the 60years of KKT Conditions

So, now I have this equation lambda naught, grad f x bar plus lambda 1, this is called the Lagrange equation on the KKT equation, that is I know that lambda naught is strictly greater than 0; so, I can divide both sides by lambda naught.

(No audio from 21:54 to 22:32)

So I will call lambda i by lambda naught has lambda i bar. So, what I have proved that if Slater cq holds and this is of course, greater than equal to 0, because each of them is greater than equal to 0, these greater than 0; if Slater condition holds, there exists lambda i bar greater than equal to 0, i from 1 to m such that number one, grad of f of x bar plus lambda 1 bar grad g 1 x bar lambda m bar grad g m x bar is equal to 0, and number two is lambda i bar g i x bar is equal to 0 for all i.

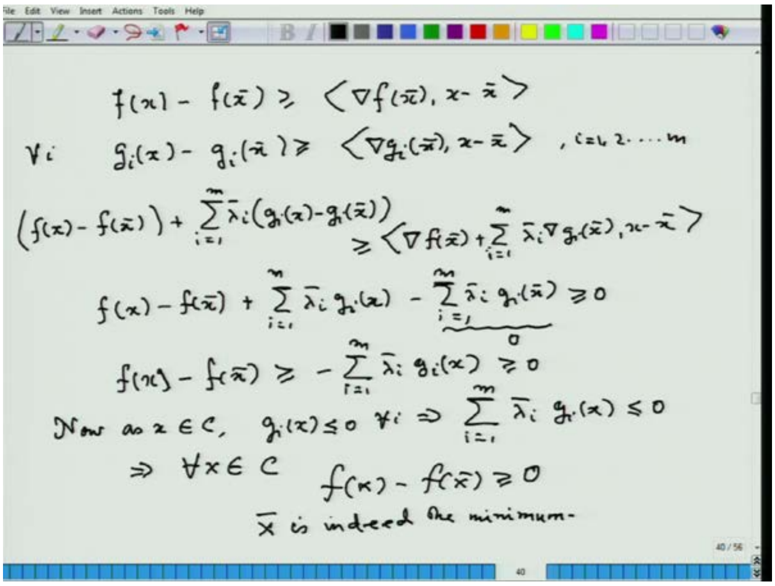
Now, this third condition in the Fritz John one, this is no longer required, because lambda naught I have proved to be strictly greater than 0. So, lambda naught lambda naught is basically one. So, Fritz John KKT condition is Fritz John condition with lambda not equal to 1 under so this holds, if you have some additional condition. So, what we have got here is the famous Karush-Kuhn-Tucker condition, it now goes by the name of Karush-Kuhn-Tucker condition. So, I think one of the first papers who possibly gave this name Karush-Kuhn-Tucker condition was the paper, which was called the modern multiplier rules which appeared in 1980 1980 or 81 in the American main mathematical monthly is called modern multiplier rules; by B. H Pourciao, American Math Monthly, it is beautiful (( )) and its says that though Kuhn and Tucker have independently derived this condition.

So, this was known to Karush, who a back in the mid end end of the 30s, early 40s I guess. So, now it is called the Karush-Kuhn-Tucker conditions or the KKT condition; it became famous in the 1951, seminar 1951 paper by Kuhn and Tucker, Harold W Kuhn and Albert W Tucker both from Springsteen; it is very important to remember that even if Karush had this ideas slightly ahead that does not diminish the value or the worth of the Karush-Kuhn-Tucker of the of the KKT condition on this paper by kuhn and tucker, because kuhn and tucker in that paper demonstrates with example why example where lambda naught become 0, and they did it even for just differentiable function, they found the necessary condition, not as we have done for convex functions, and they did not talk about Slater condition, which came slightly later; they spoke about a general geometrical optimality condition, geometrical constraint qualification or condition on the constraints, which is called the Kuhn-Tucker constraint qualification.

And furthermore this famous paper had also given results for multi objective optimization and that is also very important. So, this paper has a huge worth in optimization community, as Professor Harold Kuhn once told me that once he was inside a conference, trying to he was told to introduce Professor Tucker and he said I am Kuhn and he is Tucker and that is all everybody knows what it is.

So, now what we have got here is a necessary condition that he have given me a solution x bar and I have showed the condition that x bar should satisfied provided the Slater condition holds, we will approach arrive at this condition through various routes. So, there many, many different paths to the Karush-Kuhn-Tucker condition and you can immediately understand that 2011 was the 60 years of KKT condition. (No audio from 27:06 to 27:18) Now, when the functions are convex, the question is this also necessary that is if there is an x bar and if there is some lambda bar, which satisfies all this is my I know, whether that x bar is actually a global minimum, answer turns out to be yes and let us see, how to do it.

(Refer Slide Time: 27:37)



$$f(x) - f(\bar{x}) \geqslant \langle \nabla f(\bar{x}), x - \bar{x} \rangle$$

$$\forall i \quad g_i(x) - g_i(\bar{x}) \geqslant \langle \nabla g_i(\bar{x}), x - \bar{x} \rangle \quad , i = 1, 2 \ldots m$$

$$\left(f(x) - f(\bar{x})\right) + \sum_{i=1}^{m} \bar{\lambda}_i \left(g_i(x) - g_i(\bar{x})\right) \geqslant \langle \nabla f(\bar{x}) + \sum_{i=1}^{m} \bar{\lambda}_i \nabla g_i(\bar{x}), x - \bar{x} \rangle$$

$$f(x) - f(\bar{x}) + \sum_{i=1}^{m} \bar{\lambda}_i g_i(x) - \underbrace{\sum_{i=1}^{m} \bar{\lambda}_i g_i(\bar{x})}_{0} \geqslant 0$$

$$f(x) - f(\bar{x}) \geqslant - \sum_{i=1}^{m} \bar{\lambda}_i g_i(x) \geqslant 0$$

Now as $x \in C$, $g_i(x) \leq 0 \ \forall i \Rightarrow \sum_{i=1}^{m} \bar{\lambda}_i g_i(x) \leq 0$

$$\Rightarrow \forall x \in C \quad f(x) - f(\bar{x}) \geqslant 0$$

$\bar{x}$ is indeed the minimum.

Look at look at the definition of a convex function, when they are differentiable; now we have f x minus f x bar and this is of course, you know take any x in the feasible set, in C; now take for any I, for all i g i x minus say take the same x, g i x bar is again for each i. Now, you can multiply by lambda i, because lambda i is greater than equal to 0. So, there is no harm in multiplying this set by lambda i, and then once you multiply by lambda you add this inequalities what you will get is. (No audio from 28:33 to 28:56) So, I am multiplying by lambda i bar sorry summation lambda i bar grad g i x bar i element sorry i is equal to 1 to m, i is equal to 1 to m x minus x bar.

Now, let us just apply these conditions. So, you know from this condition it is immediate that this right right hand side is 0, because this 0. So, what I have is f x minus f x bar plus summation lambda i bar g i x, so minus summation lambda i bar g i x bar, this is greater than equal to 0, and if you observe this second condition, which is called the complementary slackness condition, this is called the complementary slackness, which says that both of them cannot be slack at the same time, that is lambda i cannot be strictly greater than 0, and g i x also cannot be strictly less than 0 at the same time, both of them can be 0 at the same time, but both of them cannot be having strict inequalities at the same time.

So, now this is 0, from this condition; condition 2, this part is 0. So, what I will have is f x minus f x bar taking this thing to the other side, now as x is in C that it is feasible, g i x is less than equal to 0 for all i; this would imply, because lambda i is bigger than 0, i is equal to 1 to m lambda i bar g i x is less than equal to 0, but I have put a minus sign. So, what I would finally get is that this is greater than equal to 0. So, this would imply for all x in C, because x was an arbitrary element in C, f x minus f x bar is greater than equal to 0, showing that x bar is indeed a minimum.