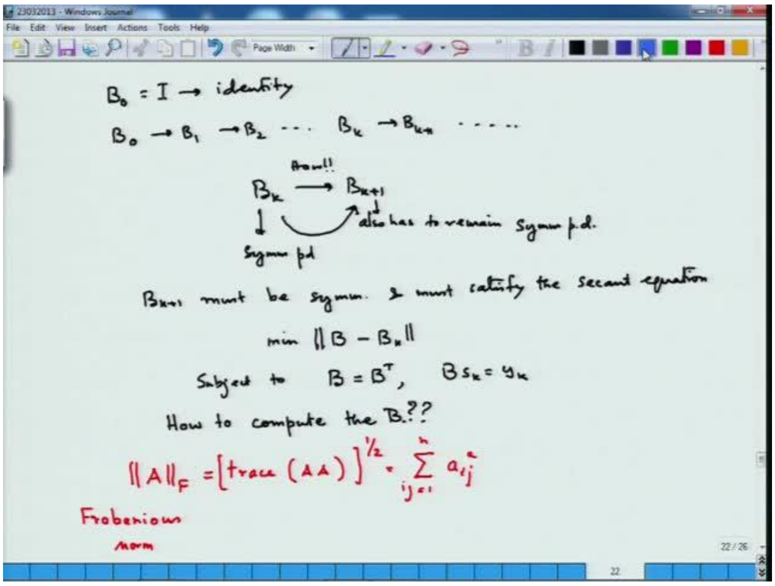
Foundation of Optimization

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

Lecture - 24

(Refer Slide Time: 00:31)


23032013 - Windows Journal
File Edit View Insert Actions Tools Help
Page Width
22/26
22

$B_0 = I \rightarrow$ identity

$B_0 \rightarrow B_1 \rightarrow B_2 \cdots B_k \rightarrow B_{k+1} \cdots\cdots$

$$\begin{tikzcd} B_k \arrow[r, "\text{How!!}"] \arrow[d] \arrow[r, bend right] & B_{k+1} \arrow[d] \\ \text{Symm pd} & \text{also has to remain Symm p.d.} \end{tikzcd}$$

B_{k+1} must be symm. & must satisfy the secant equation

$$\min \|B - B_k\|$$

Subject to $B = B^T$, $Bs_k = y_k$

How to compute the B.??

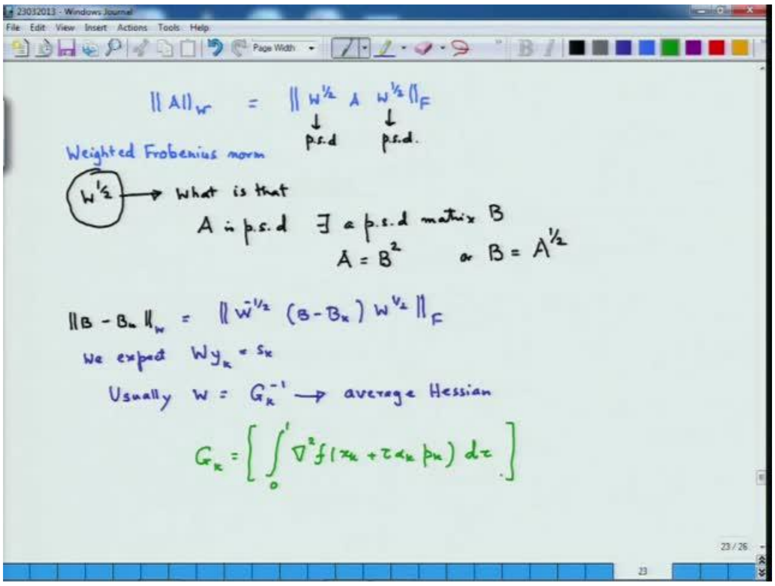
$$\|A\|_F = [\text{trace}(AA)]^{1/2} = \sum_{i,j=1}^{n} a_{ij}^2$$

Frobenious norm

Welcome once again to this course in optimization. Today we are going back to our study of Quasi Newton method which are by the way very effective methods. Now, this is the problem that we really intended to solve, because we had a matrix B k. We now want to improve from B k to B k plus 1, such that that B k is B k plus 1 would be symmetric satisfy the second equation as well as it has to be a positive definite at the end. So, how do I do it? How do you choose the norm of a matrix? Now, because I am taking a symmetric matrices, in the case of symmetric matrices there is an important norm.

Say, if you have A matrix a there is an important class of norm called the Frobenious norm. This Frobenious norm can be actually written as the trace of A A and you take half of that right. Now, of course this simply means that you can just write it down right. Now, because if you have an (( )) matrix, you can write down this as summation if all the elements of these are a i j. So, you can write them as a i j square where i and j are varying from 1 to n.

(Refer Slide Time: 02:29)



23032013 - Windows Journal
File Edit View Insert Actions Tools Help
Page Width
23 / 26
23

$$\|A\|_W = \|W^{1/2} \, A \, W^{1/2}\|_F$$

$$\underset{\text{p.s.d}}{\downarrow} \qquad \underset{\text{p.s.d.}}{\downarrow}$$

Weighted Frobenius norm

$\boxed{W^{1/2}} \rightarrow$ What is that

A is p.s.d $\exists$ a p.s.d matrix B

$A = B^2$ or $B = A^{1/2}$

$$\|B - B_k\|_W = \|W^{-1/2} (B - B_k) W^{1/2}\|_F$$

We expect $Wy_k = s_k$

Usually $W = G_k^{-1} \rightarrow$ average Hessian

$$G_k = \left[\int_0^1 \nabla^2 f(x_k + \tau \alpha_k p_k) \, d\tau \right]$$

Now, usually in practice people have used what is called a weighted Frobenious norm of a matrix. So, it is called the weighted Frobenious norm. Now, the weighted Frobenious norm of the matrix is defined as the matrix a multiplied by a weight, now I would like to remind you why we use this symbol w to the power half. So, what is that square root of A matrix precisely? So, these matrices are p s d matrices right usually these are positive semi definite matrices.

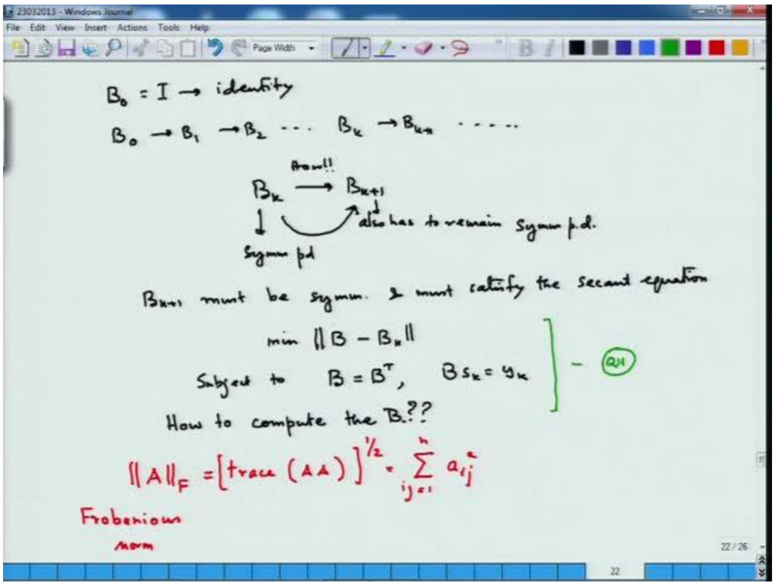
So, what it means that if you give me a positive semi definite matrix, there is a famous result if a is positive semi definite. Then there exists a positive semi definite matrix p s d matrix symmetric A is a n cross n percent symmetric p d positive semi definite matrix, then there will exist a symmetric positive definite matrix B, such that a is equal to B square or B is usually marked as or note denoted as a to the power half.

So, with this that is that is the meaning of this w, now usually w is, so chosen that you know if you take a matrix like this basically, then if you look at this matrix B minus B bar or B minus B k. So, you really have to choose you want the weighted thing, so it should be W to the more minus half, so this is not the only way to do it. It is, but this is one of the ways used in practice, so we are just showing one of you this approach, but it might be quite difficult for a beginners.

So, we will not just go into too much details give you an idea about what would go on and then take a simpler version of the problem and give a solution which Guler has given which is the simple and beautiful solution. Once you see that you can really appreciate and understand what has happened, so if you take A w like this we expect w to satisfy W y k should be is equal to s k, which is nothing but the reverse of it is nothing but the reverse of the if W y k is s k, then of course B s B k s k is y k this can be y k and so and so forth.

Sorry, B k, so this is what you would expect. We we expect this to happen and how do you know what is your W? So, usually W is given as the inverse of the matrix, G k inverse, which is called the average Hessian. This has a particular formula which says you are a basically integrating out the Hessian matrix. So, you are taking the matrix and integrating it out term by term every term of the matrix is integrated and the new matrix means. This is now a matrix G is a matrix whose every term is some sort an integral of this you take the Hessian matrix term by term you integrate right this is s k minus s k plus 1 in to d 2. If you choose something like this and try to optimize this problem get a solution to this problem.

(Refer Slide Time: 08:04)



23032013 - Windows Journal
File Edit View Insert Actions Tools Help
Page Width
22 / 26
22

$B_0 = I \rightarrow$ identity

$B_0 \rightarrow B_1 \rightarrow B_2 \cdots B_k \rightarrow B_{k+1} \cdots\cdots$

$B_k \xrightarrow{\text{How!!}} B_{k+1}$

$\downarrow$ symm pd

also has to remain symm p.d.

B_{k+1} must be symm. & must satisfy the secant equation

$$\left.\begin{array}{l} \min \|B - B_k\| \\ \text{Subject to} \quad B = B^T, \quad B s_k = y_k \end{array}\right] - \text{QN}$$

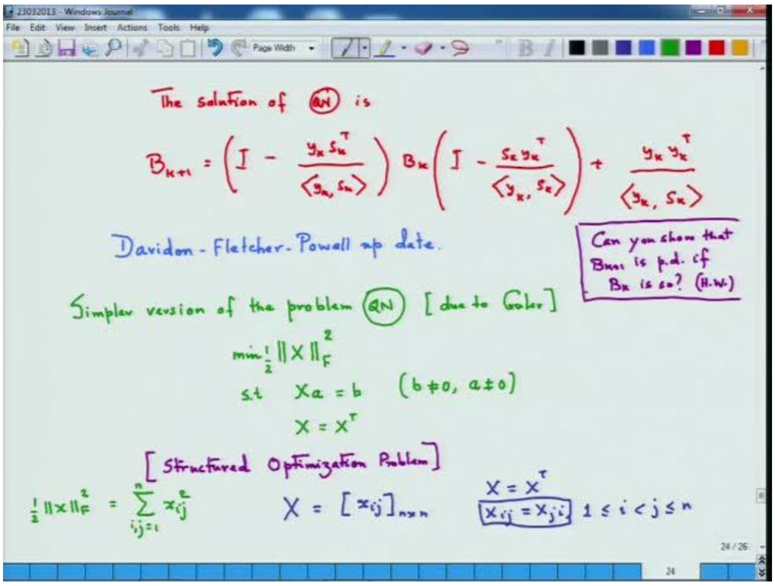
How to compute the B.??

$$\|A\|_F = [\text{trace}(AA)]^{1/2} = \sum_{i,j=1}^{n} a_{ij}^2$$

Frobenious norm

Then your solution becomes, so if I call this problem as Q N or the Quasi Newton problem.

(Refer Slide Time: 08:13)



The solution of $\boxed{QN}$ is

$$B_{k+1} = \left(I - \frac{y_k s_k^T}{\langle y_k, s_k\rangle}\right) B_k \left(I - \frac{s_k y_k^T}{\langle y_k, s_k\rangle}\right) + \frac{y_k y_k^T}{\langle y_k, s_k\rangle}$$

Davidon - Fletcher - Powell up date.

Can you show that B_{k+1} is p.d. if B_k is so? (H.W.)

Simpler version of the problem $\boxed{QN}$ [due to Guler]

$$\min \frac{1}{2}\|X\|_F^2$$
$$\text{s.t} \quad Xa = b \quad (b \neq 0,\ a \neq 0)$$
$$X = X^T$$

[Structured Optimization Problem]

$$\frac{1}{2}\|X\|_F^2 = \sum_{i,j=1}^{n} x_{ij}^2 \qquad X = [x_{ij}]_{n \times n} \qquad X = X^T \quad \boxed{x_{ij} = x_{ji}} \quad 1 \leq i < j \leq n$$

23032013 - Windows Journal
File Edit View Insert Actions Tools Help
Page Width
24 / 26
24

So, the solution of Q N is I minus y k s k transpose divided by y k s k, looks elegant that such a complicated looking thing would give such a nice result. So, these are matrices, so if the starting matrix B naught is I, then B 1 is product of this, right? Product of this 2 into this plus, this now this sort of updating moving from B k to B k plus 1 is called the Davidon Fletcher Powell update or the Davidon Fletcher Powell method. So, a the upper one is called… Now, the question is how do you actually solve this problem? The weighted a graph problem ha how do you actually solve that problem?

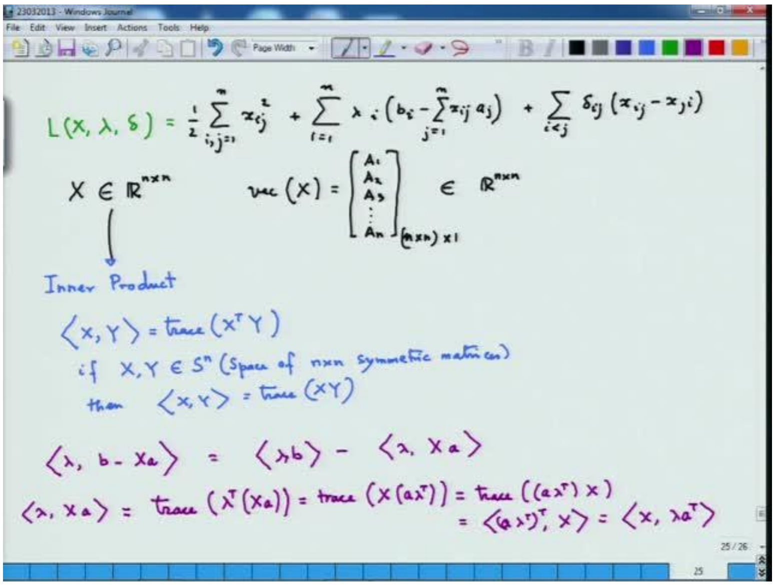
That is a very, very important question and if I really want to solve that kind of problem it is quite difficult and in the sense that involved. It might be the scope of it might be beyond the scope of this audience which is a mixed 1, I guess so I will look into much simpler version of this problem which i take from Guler, Osman Guler Foundations of Optimization and with that simple problem I will try to solve that simple problem leaving a difficult job to you of course, but let us look at that simple problem, so simpler version, a problem Q N due to Guler. Now, let us look at this problem, see we will just look into this problem, so we will take an x let us, so let us.

So, we will consider now the problem is on the space S N of symmetric matrices. So, we will minimize a matrix n cross n matrix X using the Frobenius norm, in fact I I will minimize half of the square of that such that x times a is b, a is an n cross n matrix and b is an n cross n matrix and of course, we do not expect b to be 0, b of course we will expect b to be non zero and a to be non zero too. So, let us just take for the time being, so X is my… So, this is a structured optimization problem as Guler calls it.

So, we will just write it down here as this is a structured optimization, problem. In fact I have a query to give you here as a home work can you take this? Can you show that B k plus 1 is positive definite, if B k is so. Take it take take this as a home work. Now, we are going to look into this problem and try to solve this problem. If I write down the I can write down everything in terms of the variables x i j, where X I can choose, X the matrix X can be written as, so these x i j can be my variables.

So, they are n square variables known and these thing, X equal to x transpose this can be setup for for… So, I can write the, this constraint can be written like this and this constraint this 1 as as I have shown you earlier can be written as. Now, so how do I construct the Lagrangian of this. So, that I can write down the Karush Kuhn Tucker conditions the Lagrangian which is of course, given in terms of X lambda, right? Delta and this is now given as half summation i j from 1 to n x i j square plus.

(Refer Slide Time: 15:25)


23032013 - Windows Journal
File Edit View Insert Actions Tools Help
Page Width
25 / 26
25

$$L(X, \lambda, \delta) = \frac{1}{2}\sum_{i,j=1}^{n} x_{ij}^2 + \sum_{i=1}^{n} \lambda_i \left(b_i - \sum_{j=1}^{n} x_{ij} a_j\right) + \sum_{i<j} \delta_{ij}(x_{ij} - x_{ji})$$

$$X \in \mathbb{R}^{n\times n} \qquad \text{vec}(X) = \begin{bmatrix} A_1 \\ A_2 \\ A_3 \\ \vdots \\ A_n \end{bmatrix}_{(n\times n)\times 1} \in \mathbb{R}^{n\times n}$$

$\downarrow$

Inner Product

$\langle X, Y \rangle = \text{trace}(X^T Y)$

if $X, Y \in S^n$ (Space of $n\times n$ symmetric matrices)

then $\langle X, Y \rangle = \text{trace}(XY)$

$$\langle \lambda, b - Xa \rangle = \langle \lambda, b \rangle - \langle \lambda, Xa \rangle$$

$$\langle \lambda, Xa \rangle = \text{trace}(\lambda^T(Xa)) = \text{trace}(X(a\lambda^T)) = \text{trace}((a\lambda^T)X)$$
$$= \langle (a\lambda^T)^T, X \rangle = \langle X, \lambda a^T \rangle$$

Now, this x a equal to b can be written as summation x i j a j minus b j, so you can write this very basic matrix multiplication. So, it is i is equal to 1 to n lambda i b i minus summation x i j a j. Now, if you look at this j, j is equal to 1 to n and then you of course have, so this is your Lagrangian, but this you really try to take the derivative and all those things this will become very, very complicated. So, in order to make the solution look simpler to do the solution simply what we do is, we a actually put in this whole thing in form of a matrix, in terms of the matrix X.

So, X will now play a role, you see the complicacies that will arise, if you are now putting a weighted thing and getting this, all right? Now, let me write down the stuff, so you know in general if you have a matrix X in the space R n cross n, so any X is a member of R n cross n, it can be represented because if I I can represent a matrix x, if I by using a fake operation if it is if it is column is first column is a 1 second column is a 2, then I stack up the columns like this. And I get this n, n, n, n, n times. So, this is a n cross n into one vector.

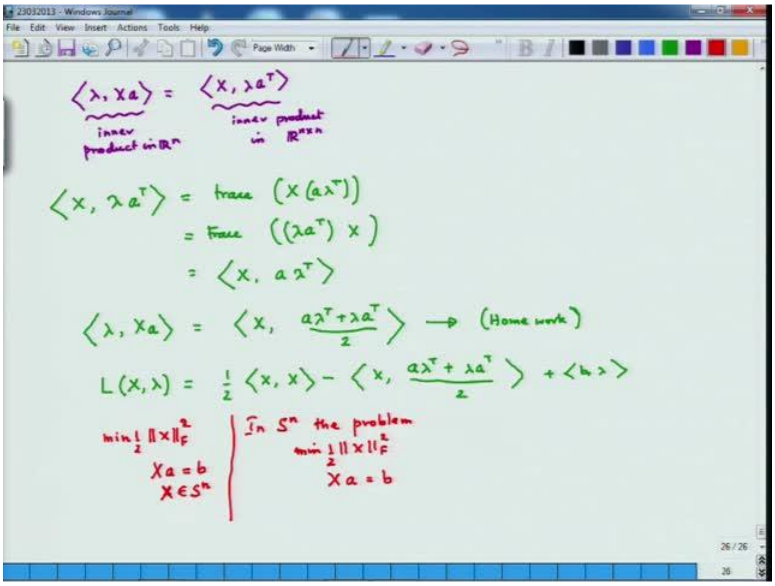
So, it is an n square vector and this vector is a member of R n cross n. So, that is why it is always written like this X is R n cross n. Then the inner product if this is the case, then inner product, then the inner product is given as follows. The inner product of two members of R n cross n is usually given as trace of X transpose Y. If x and y belong to S n, which space of symmetric matrices, space of… Then x y can also be written as just trace of, now what happened is as follows.

Now, it is very important how do I express this thing, this part of course, I can write this part is nothing but half x x in a in terms of the inner product. So, I am writing every term thing in terms of the matrix. Now, how do I translate this one? The second part is translated as follows, is written as lambda times b minus X a, right? This means lambda b minus lambda times X a. Now, look at this expression lambda times X a. So, this inner product, if you look at this inner product this inner product is nothing but trace of lambda transpose X a.

So, what I am trying to mean is as follows. So, what is the trace of lambda transpose X a because lambda transpose X a is just a number. So, it is a 1 cross 1 matrix. So, it traces that number, so that is a trace of this. So, this can be in written as using the associative law matrices the trace of X a lambda transpose. See this is not a symmetric matrix a lambda transpose, this a lambda transpose is not a symmetric matrix. So, when I write the, so it will be if X into y means, so it means trace of X so basically this would become trace of a b is trace of b a.

So, here what had happened? They have taken x a lambda transpose, then by associative law it become x a lambda transpose. Now, this is I can write this as trace of, right? So, this simply means transpose of, the transpose if I take, transpose of the transpose, of the transpose. So, basically this actually means a lambda transpose transpose X, which is same as writing as X lambda a transpose.

(Refer Slide Time: 23:32)


23032013 - Windows Journal
File Edit View Insert Actions Tools Help
Page Width
26 / 26
26

$$\underbrace{\langle \lambda, Xa \rangle}_{\text{inner product in } \mathbb{R}^n} = \underbrace{\langle X, \lambda a^T \rangle}_{\text{inner product in } \mathbb{R}^{n\times n}}$$

$$\langle X, \lambda a^T \rangle = \text{trace}\,(X(a\lambda^T))$$
$$= \text{Trace}\,((\lambda a^T)\,X)$$
$$= \langle X, a\lambda^T \rangle$$

$$\langle \lambda, Xa \rangle = \left\langle X, \frac{a\lambda^T + \lambda a^T}{2} \right\rangle \rightarrow \text{(Home work)}$$

$$L(X,\lambda) = \frac{1}{2}\langle X, X\rangle - \left\langle X, \frac{a\lambda^T + \lambda a^T}{2} \right\rangle + \langle b, \lambda \rangle$$

$$\min \frac{1}{2}\|X\|_F^2$$
$$Xa = b$$
$$X \in S^n$$

In S^n the problem
$$\min \frac{1}{2}\|X\|_F^2$$
$$Xa = b$$

Lambda times x of a is equal to X times lambda of a transpose. So, this is the inner product in R n and this is the inner product in R n cross n, but there is a interesting relation you see, that is the beauty of mathematics, better one. So, once I know this then I look into it bit more deeply. It is same as, as you have written in the last page trace of X a lambda transpose and this this is a, this is a sorry, just this step. Let me have a look at it the samething which can be a lambda transpose X.

So, which can be written as follows this can be written as taking taking up from here it can be also written as lambda. So, this is already lambda a transpose transpose. So, it is all alright. So, this is nothing but X of a lambda transpose. So, you see there is a lot of links, so once you have seen this, I can also write lambda X a can be also written as X of a lambda transpose plus lambda a transpose. These are matrices by 2, so this I leave as, so this is now a symmetric matrix.

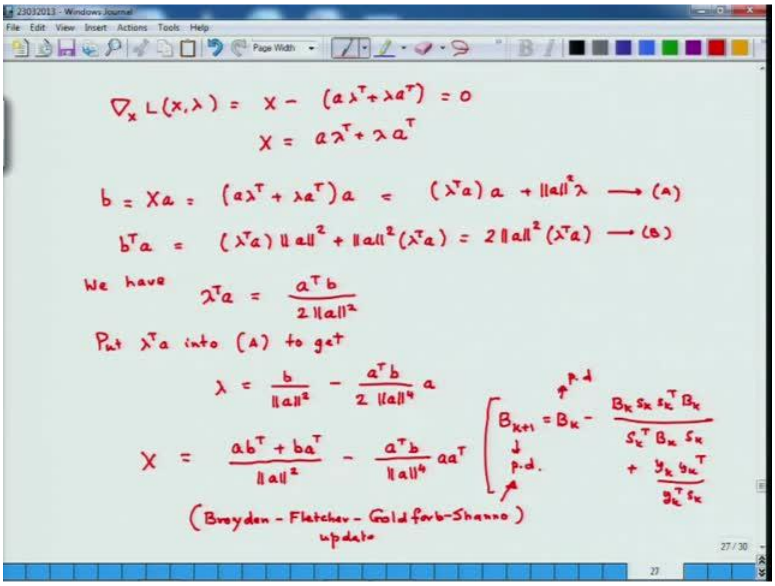
So, this is homework this is very simple you have to observe these things that is all half of this plus half of this, that is all half of this plus half of this is this. Now, once you do that, then I can now write the Lagrangian function l X lambda delta as, now what about this part? This part is that we are already taking this whole thing to be true, because this part, now X is equal to X transpose. So, we are now taking only symmetric matrices. Now, my problem is of this form, that I have to only bother about symmetric matrices, right?

So, X is so I have to now I now, what I do is, I make not this Lagrangian. I I now just look at this Lagrangian like this, right? I actually now construct this Lagrangian, the first part and the second part because now I am just working in S n I will tell you what is the issue. Now, you let me just tell you something what happens here is this, you might tell me what what about this constant, where does it vanish to? So, I can because here what I have done, I have already taken X equal to X transpose and I have set up my problem as this, my the Gulers problem is now this.

Now, S n is my whole space S n is a space of all symmetric matrices. Now, then what I am doing is my optimality conditions should tell my something my optimality conditions are the Karush Kuhn Tucker condition is to note that basically my my S n is just like R n the whole thing that is the whole space. So, the normal cone to that whole space is actually 0, so basically we are now minimizing some inequality constraint problem. So, if I am putting whole thing in S n I do not need to bother to put this this thing, I know that it is in S n only, my problem is in S n.

So, I need not bother to put this thing, so I am looking at the whole thing in from the from the matrix perspective, from the completely, from the matrix perspective. So, my S n is my whole space. Now, so basically in in in s n the problem is, so in S n, I am doing this problem, that is my problem. On this I can apply the Karush Kuhn Tucker condition by taking the gradient of the Lagrangian.

(Refer Slide Time: 30:31)



$$\nabla_x L(X,\lambda) = X - (a\lambda^T + \lambda a^T) = 0$$

$$X = a\lambda^T + \lambda a^T$$

$$b = Xa = (a\lambda^T + \lambda a^T)a = (\lambda^T a)a + \|a\|^2\lambda \longrightarrow (A)$$

$$b^T a = (\lambda^T a)\|a\|^2 + \|a\|^2(\lambda^T a) = 2\|a\|^2(\lambda^T a) \longrightarrow (B)$$

We have $\lambda^T a = \dfrac{a^T b}{2\|a\|^2}$

Put $\lambda^T a$ into (A) to get

$$\lambda = \frac{b}{\|a\|^2} - \frac{a^T b}{2\|a\|^4}a$$

$$X = \frac{ab^T + ba^T}{\|a\|^2} - \frac{a^T b}{\|a\|^4}aa^T$$

$$\left[\underset{\overset{\downarrow}{\text{p.d.}}}{B_{k+1}} = \overset{\overset{\text{p.d}}{\uparrow}}{B_k} - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k} \right.$$

(Broyden - Fletcher - Goldfarb-Shanno) update

So, now what we do is a you take the gradient of the Lagrangian and equate it to 0, you see the Lagrangian is now written like this and then that would give you X minus a lambda transpose plus lambda a transpose, So, this is 0, so this simply tells you x must itself be of the form a lambda transpose plus lambda a transpose. So, a is given to you what you have to know is lambda the Lagrange multiplier. So, here you see the importance of the Lagrange multiplier of the Karush Kuhn Tucker multipliers.

Now, you might ask me, how, why are you applying Karush Kuhn tucker multipliers? How do you know the KKT condition will hold? How do you know that there is a constant qualification which is satisfied. Now, in the setting of S n X a equal to b is a linear inequality and hence there is no requirement of any satisfaction of a constant qualification and KKT, there would be a KKT multiplier, right? We have shown that there would exist a KKT multiplier. I hope you remember that we have shown that every Fritz John point can be thought as a KKT point because there would exist a KKT multiplier.

I can always show that I can directly reach the KKT conditions, if I have linear constraints, right? So, if that is the case then here, we have so this would be this and this would give me this is an inner product. Of course, real number in to this vector this real number into this vector. Now, b transpose a, which is b inner product a can be written like this. I want you to check the calculation that I am doing now. So, the last equation from this this one, we have lambda transpose a is equal to a transpose b, divided by 2 norm a square.

So, lambda transpose a can now be put in here. So, if I put this here, I can get out lambda. So, then this is a, and b put lambda transpose a into put lambda transpose a into the equation a to get. So, you can understand, this will become bit complicated when you have that particular structure s k y k you can a is your s k and b is your y k here, that is all. Now, putting this into X, so X would be a b transpose plus b a transpose. So, this is what is usually called the Broyden this updating is called the Broyden Broyden Fletcherm, Gold Farb Shanno, we will write it in terms of S k y k etcetera tomorrow.

So we will write this down in terms of so in terms of the b k 1 plus b k, but there they may know a norm is taken in a very different way. So, here that will give you, just try to use this technique and just take the Frobenius norm of b norm b minus b k. Try to see whether you get this formula, so here you see basically b k minus one norm b k, that that is what you are basically looking at. Here you see the minus 1 the minus 1 is here, so b k s k is your y k right that is your norm here plus so minus part is here plus part is here so x is actually your b minus b k b k plus one minus b k so he here comes you are a, a transpose a a transpose, right?

So, this is what is up this called this is your BFG's update. So, we will write down the BFG's formula algorithm tomorrow and go ahead a bit into our study of the. So, this when BBK is p d b k plus 1 is p d. So, we can go ahead into our study of the problem the Quasi Newton method, I will take once after I finish the Quasi Newton method. Then I want to go into the issues of constraint programming. I want to talk about penalty function method I would take a very special lecture in on the KKT conditions for linear programming problem.

See it is usually felt and it written in every book that I have seen is that every FJ point is a KKT point, but that actually is a misinformation in a sense that what we can show that there would exist a Fritz John multiplier give me any local global minimum of a linear programming problem there would exist 1 Fritz John multiplier which is normal you cannot say that every Fritz John multiplier associated with a linear programming problem is normal.

I think in from my earlier lecture that wouldn’t have been clear. So, I would take a complete step by step lecture to study the linear programming problem KKT conditions for linear programming problem of Fritz John conditions for linear programming problem. So, this is that would be a very special lecture and I think that everybody should listen very, very carefully that if a constraint qualification fails, the Mangarsarian Fromovitz constraint qualification fails, even if your problem is linear there would exist an abnormal multiplier, this is something which is very important in your understanding of optimization. So, tomorrow we will continue our journey on the Quasi Newton method.