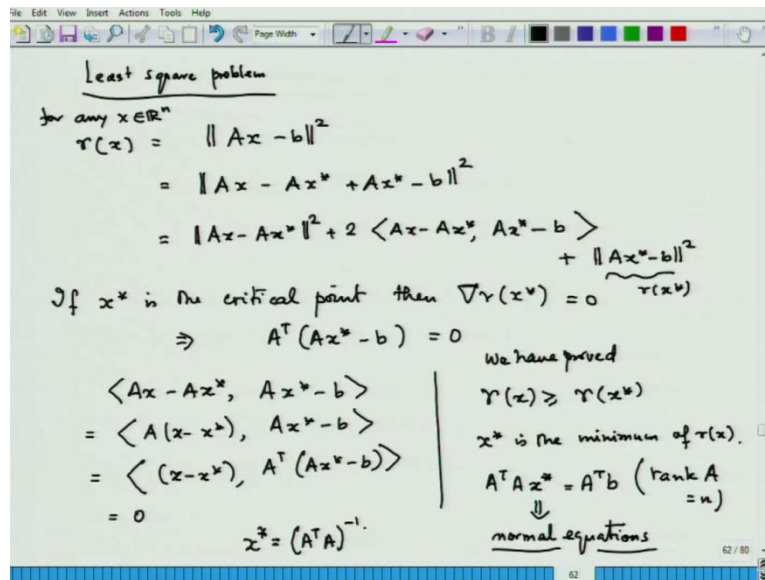


Foundation of Optimization
Prof. Dr. Joydeep Dutta
Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur

Lecture – 13

(Refer Slide Time: 00:24)



So we were discussing this least square problem, basically using this technique of optimization to try to find the solution to $Ax = b$. Now, if you remember we have shown that any critical point of this function is a solution, is a minimum to this function and hence is, and hence is a solution to the original problem.

So any critical point of this problem is a solution to this problem, so the in fact it has only critical, and it has all critical points should satisfy this equation, so these are sometimes called normal equations. And if you remember what we had shown that any x^* which satisfies this, this one will satisfy $Ax = b$ that is what we had proved in the board so basically to solve to find an x^* , which satisfy $Ax^* = b$. We basically have to solve an equation like this, that is we should have $x^* = (A^T A)^{-1} A^T b$ of course, this is only true when rank of A is n is a full column rank, m has the row has to number of rows has to be more than a n , so it would it should have so basically my x should have this expression.

(Refer Slide Time: 02:06)

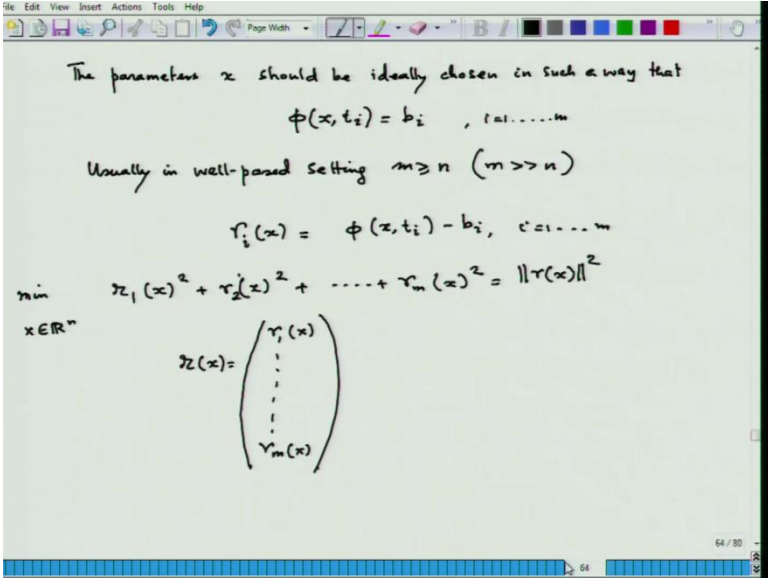
The image shows a whiteboard with handwritten mathematical notes. At the top, it says $(A^T A)^{-1} A^T \rightarrow$ pseudo-inverse of A . Below that, it defines the "Least square problem" as $\min_{x \in \mathbb{R}^n} f(x) = \sum_{i=1}^m (r_i(x))^2$. A bullet point follows: "Regression Analysis in statistics." To the left of this is a graph with a curve and a point labeled (t_i, b_i) . To the right, it says "Curve fitting (important applications)" and gives the equation $b = \phi(x, t)$ with "parameter" written below ϕ . Below that, it shows $b = at + c = \phi((a, c); t)$ with x written below (a, c) . The whiteboard has a toolbar at the top and a status bar at the bottom showing "63 / 80".

So this expression $(A^T A)^{-1} A^T$ is sometimes called the generalized inverse of A or pseudo inverse of A . Now what sort of an algorithm would actually work for the least square problem, in general let me write you let me tell you the least square problem has the following form that this discussion that we were doing in the last class where just an demonstration of how optimization in the form of minimizing a square. So the least square minimum square can be used to really solve a problems of solving linear system of linear equations. So, least square problem is essentially trying to minimize a function $f(x)$ which itself is expressed as the sum of squares, an explanation of the use so I want to minimize over x element of \mathbb{R}^n . So, an example of such thing comes possibly in regression, in regression analysis in statistics, this is the least square method is a quite chosen method.

So, what happens? One of the problem is that you are given certain points, say of the form t_i, b_i , time and positions of a person. Now, what sort of curve will it fit, shall I can I fit a straight line to explain the relationship between t_i and b_i , that is whether t_i and b_i have certain relation or there are some curve like this, that will be a better fit. So this, this there is a whole subject called curve fitting, and that is also useful in statistics what does is exactly what it do in regression analysis, you try first the basic problem is to give first try to fit a straight line. And then you really want to minimize the errors, square of the error that you would get if you use the straight line instead.

And if you just assume that these points actually should have a linear relationship, that is they would lie in some straight line. So, curve fitting is a very important area, the curve fitting has important applications. Now, so suppose I have all these points of the form t_i, b_i is obtained by some experiment, and of course some physical experiment or some statistical experiment. So, we want to see how b is related to t and x here is essentially the parameter if you want to fit a straight line, that is if you want to say b is a t plus c , suppose you want a relationship like this then this is your $\phi(x, t_i)$, in fact a c here plays the role of x .

(Refer Slide Time: 06:37)



The parameters x should be ideally chosen in such a way that

$$\phi(x, t_i) = b_i, \quad i=1, \dots, m$$

Usually in well-posed setting $m \geq n$ ($m \gg n$)

$$r_i(x) = \phi(x, t_i) - b_i, \quad i=1, \dots, m$$

min $r_1(x)^2 + r_2(x)^2 + \dots + r_m(x)^2 = \|r(x)\|^2$
 $x \in \mathbb{R}^n$

$$r(x) = \begin{pmatrix} r_1(x) \\ \vdots \\ r_m(x) \end{pmatrix}$$

Now, once that is done these n set of parameters x should be ideally chosen the parameter x should be ideally chosen in such a way, in such a way that at every t_i , the function value should be b_i . So, suppose I have just generated m points here here, but the number of points I generate for example, in two dimensional, so he has two d i, suppose, I am in two dimension and the number of points I have generated is as two.

And of course, you can say they are laying on a straight line may be actually the relationships are not that, when we run the experiments. So, experiments has to be run much more times than the dimension of the decision variables, so the number of parameters here, whatever be the number of parameters. Suppose a n is a number of parameters usually in a well posed setting m should be much larger than m , actually I should say m , this is way of telling m is much larger than see due to experimental errors

or whatever this condition need not match exactly, so what I conclude is that I compute some reschedules, so these are so if I find an x which will minimize these residuals not really these residuals, but the sum of the squares of this residuals. So, I will put so for each i , I am having one residuals so and then I want to minimize this over x in \mathbb{R}^n and this is nothing but if you take a vector r x . So, if you take r x that is nothing but sorry, an u square of the Euclidian norm of r , so basically you have to minimize this function over \mathbb{R}^n , right.

So this is a very important example where least square techniques are of useful, so statistics. For example, is one of the very, very important areas where least square techniques are indeed very useful, now we are going to see what sort of a algorithm one might use, when one tries to do a least square method ok.

(Refer Slide Time: 10:42)

GAUSS-NEWTON METHOD

$$f(x) = \sum_{i=1}^m r_i(x)^2$$

$$\nabla f(x) = \sum_{i=1}^m 2r_i(x) \nabla r_i(x) = 2J(x)^T r(x)$$

$$J(x) = \begin{pmatrix} \nabla r_1(x)^T \\ \nabla r_2(x)^T \\ \vdots \\ \nabla r_m(x)^T \end{pmatrix}$$

$$\nabla^2 f(x) = H(x)$$

$$= 2J(x)^T J(x) + 2 \sum_{i=1}^m r_i(x) \nabla^2 r_i(x)$$

↓
Homework

$$J(x)^T = [\nabla r_1(x), \dots, \nabla r_m(x)]$$

$$H(x) \approx 2J(x)^T J(x) \quad (\text{Heuristic step})$$

So this algorithm is essentially what we call the gauss Newton method, so it is some sort of Newton method, but modified by gauss, so it is called the gauss Newton method. Now, if I want to minimize a sum of squares that is now f x , first I really have to find a critical point, so I need to compute the grad gradient of grad effects, which is this is what you will have these are very simple thing this is an application of chain rule at this point. Now, you observe that here what one needs to do is that, you can express this in slightly better way you know what is a Jacobean matrix, so Jacobean matrix j x of the vector r x , so if you so basically taking the grad of x grad of r i x .

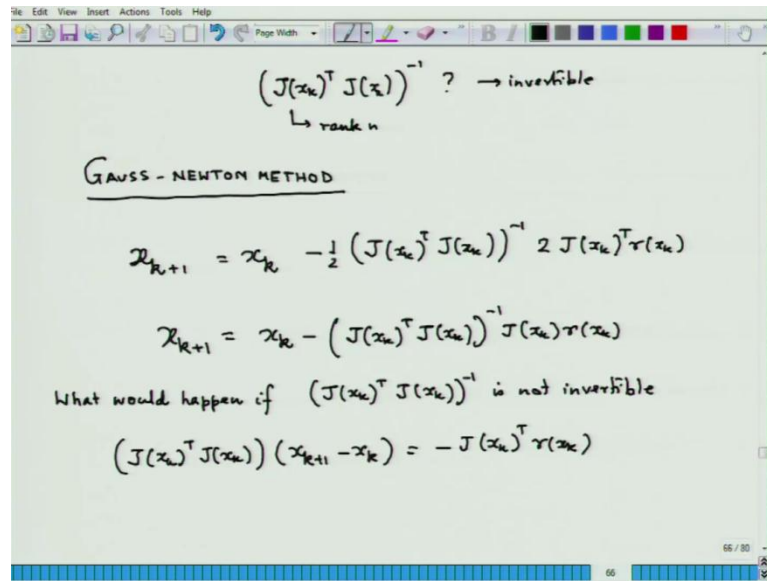
So, if you look at this like, let me write down the Jacobean matrix; Jacobean matrix is a matrix whose first row is, so is a gradient vector. Now, written as a row vector now, once you have this if you know simple matrix multiplication, then you would realize that are a grad r i x. So j x transpose is a matrix whose column is first column is gradient of r 1 x, see we are always writing things as column vectors. Now, this simply means I can write this as twice of j x transpose, if you know a simple multiplication r x.

So, once that is done I would leave you as home work the following computation is a computation of the Jacobean matrix a hessian matrix of f, so which we usually write like this or... Now for simplicity you can also write as h x which is equal to twice of, so this will be home work for you. Now, if I want to use a Newton's method then I really have to compute all these things basically a, I have to compute the hessian of each of the r i x at every point, so that will be too much of hazard not a hazard I would say but I would say that it is too much of computation effort, but instead of doing. So I can use some sort of heuristics not heuristics I do some little bit of tuning here, and then I say ok.

Let me do not not take h x the hessian matrix itself, but some sort of approximation awaits, so I take the h x is almost this. So I do not do this again derivative at all, because this J x depend on the first derivatives, so what I am trying to do I do I do not intend to use second order information in a place, where I should have used second order information, but use first order information to force in an algorithm which is as effective.

As the one which you would have if you had taken in the second order information, so this is some sort of a heuristic step, the term heuristic is very common in optimization. Now, a days many people would know, for example I have heard about genetic algorithms which is also heuristics, but I would not go immediately for algorithms which are not supported by mathematics, because mathematics gives you the strength and tells you how an algorithm would actually behave, and so when you really see a problem of a particular type, you can know that which algorithm would actually fit this scenario. So, this is let us say this is something called a heuristic step, so we are basically ignoring this part. Now, the second order part to he has with only this information, we are going to construct a technique which would actually give me you would lead me to the solution.

(Refer Slide Time: 16:57)



Now the Gauss-Newton method says, I will just do the following you should be able to put two here, but two is of course, does not make much of a difference just, so this is actually acting as some sort of an approximation, this one as some sort of an approximation to your, so it is acting as an approximation to your Hessian matrix. Now, how do I know that such an inverse would actually exist, so you have to know that $J^T J$ has how many rows m and it has n columns n $1 \times n$ is from $n \times n$ to n , so this is clear.

Of course, from the expression you are not getting to all this every time, so which means that there are n rows, now if all these n rows form a nearly independent set of vectors, right then then only the factor two here actually does not matter, because here you have a factor two. And then here if you take the inverse of this, so factor two if you take an inverse of this, then you have to take inverse of two which is half. So, that will cancel out so does not matter, so do not bother about the two here the two will cancel out. So, here in the Gauss-Newton step what we should have is that we should also have that for each x $J^T J$ is a rank n that would guarantee, if J is of rank.

Now here we have written on this heuristic step, and then will take the Gauss-Newton will write down the Gauss-Newton iteration, which is $x_{k+1} = x_k - (J(x_k)^T J(x_k))^{-1} J(x_k)^T r(x_k)$. Now, instead of a gradient squared x inverse, I have here x_{k+1} that iteration. Now of course, what we need to

do is to assume that this is each $x^T J x$ is of rank n , then this inversion formula, so the fact that rank of $J x$ is n would guarantee the following.

So in order to have the fact that this made $f(x)$ becomes invertible, so we write some sort of a Newton step, we x have we have taken that the rank is n . So, you want this to be invertible, so the required condition is that this would be of rank n which we already have mentioned in the last page. So, this leads to the Gauss-Newton method which is basically some sort of Heuristical Newton method specifically done for this least square problem, so will let us write down the Gauss-Newton method, in the Gauss-Newton method the interesting part is the following interesting part is that...

Here we write everything just like a Newton method, and we expect the Hessian matrix to be this one, if think that that is a Hessian matrix, if that was Hessian matrix. So we would expect algorithmic iterative scheme of this form, so the Hessian is twice of this. Now, if I take the inverse that will become half, so into the derivative that is twice of $J x^T$ $r \times k$, and this, this cancels out to give me $x^T k$ plus 1 is $x^T k$ minus, this 1 into that 1, 2 2 being canceling out.

So, this is the Gauss-Newton iteration. Now, in general you might think that what would happen, if this is not invertible what can I do, if then we can create what we what is called the damped Gauss-Newton's scheme, which is as follows basically you can write this one. If you look at it very carefully I can write $J x^T J x^T$, this matrix operating on the vector $x^T k$ plus 1 minus 1 $x^T k$ is equal to minus of $J x^T r \times k$, so we can tell that this difference is nothing but $d k$ the direction of or it is some α times $d k$, so in general the idea would be the following.

(Refer Slide Time: 23:38)

Damped Gauss-Newton Method

Solve $J(x_k)^T J(x_k) d_k = -J(x_k)^T r(x_k) \rightarrow (A)$

- Find λ_k such that
$$f(x_k + \lambda_k d_k) < f(x_k)$$

$$x_{k+1} = x_k + \lambda_k d_k$$

Question: Is d_k the solution of eqn. (A) a descent direction??
(Homework:)

So that let us write down what is called the damped gauss Newton method; damped gauss Newton method the idea is now to take solve this equation, this is a first step. Now find lambda k such that, so I get a complete decent f for x k plus lambda k d k, that is my x k plus 1 must be strictly less than f x k. And then basically you set x k plus 1 to x k plus lambda k d k. Now, the question is will this be a decent direction question is will this be a descent direction, so you can take this as home work. So, if I call this as equation a here, for example so is d k the solution of equation a, a decent direction that is very, very important to know whether of course, if you write d k as inverse of this, then that will become a descent direction that is the material, but in general can you show it to a descent direction kindly take this down as an home work.

(Refer Slide Time: 26:17)

Homework (Gauss-Newton Method)

$$f(x_1, x_2) = 2(x_1 - 2)^4 + (2x_1 - 3x_2)^2$$
$$= r_1(x_1, x_2)^2 + r_2(x_1, x_2)^2$$
$$r_1(x_1, x_2) = \sqrt{2}(x_1 - 2)$$
$$r_2(x_1, x_2) = 2x_1 - 3x_2$$

Now Jacobian is given as

$$J(x) = J(x_1, x_2) = \begin{pmatrix} 2\sqrt{2}(x_1 - 2) & 0 \\ 2 & -3 \end{pmatrix}$$

$J(x_1, x_2)$ is of rank 2 $x_1 \neq 2$

- Quasi-Newton Method
- Trust region method

- Karush-Kuhn-Tucker Conditions (10 Lectures)

How do I characterize a point if I know it is a local minimum of a constrained optimization problem.

The text for KKT
The Foundations of Optimization
Osman Güler
Springer
GTM - 2011

So, once you know this, I will give you an additional example to work on example, homework example to try out the Gauss-Newton method, you can even run it on your computer writing programs, so consider f of... So, this is my least square problem I have to minimize this of course, r_1 \times 1, and x_2 , in this case is root 2. And now of course, the Jacobian, which is the Jacobian of these two vector function Jacobian or vector function are whose components are these r_1 , and r_2 . So, the Jacobian is given as this is $J \times x$ basically in this case x is x_1, x_2 , so this is of rank 2, this is $J \times x_1 \times x_2$ is of rank 2, when x_1 is not equal to 2.

If x_1 is equal to two this will become 0 so the column vector would be 0 0, so that could that 2, 3, and 0 0 would be linearly dependent. And so you cannot have this, you cannot have a 0 vector in the set of linear independent vectors, so hence if x_1 is not equal to 2, then you can obviously have, so you start with points x_1, x_2 never take x_1, x_2 be 2. And if you start with those points, then if your starting point in that then Jacobian is invertible you have to make sure that your Jacobian is never x_1 is never 2, then your Jacobian has stops been invertible. Now try out this procedure using the Gauss-Newton method, and you can and also try out the damp Gauss-Newton, and see what happens? So we can talk about something later on but we can now look at our program ahead means, what are we going to learn and discuss ahead, so in the case of solving an unconstrained optimization problem.

We have two important methods left rather three important method, but first we are going to just do the 2, 1; the two more important once not more important rather very popular once quasi Newton method, number two trust region method the surprising thing about these two methods is that even if they are talking about un constrained optimization problem in order to develop this methods we need constraint optimization. So, we need very special types of constraint optimization problems, so without a better understanding of constraint optimization. And the Karush Kuhn tucker conditions which are as the necessary our sufficient conditions for differentiable optimization problems with with constraints, it is not possible to get a correct idea of these methods. So, the idea is the following that we in the coming lectures study in detail the Karush Kuhn tucker conditions, and the Quasi Newton method. And the trust region method would be done as an example of a as of the application of the Karush Kuhn tucker conditions or the ideas of constant optimization.

So this is what is very, very example; very, very important. And so we will start tomorrow's starting from central issue of optimization theory, that is optimality itself how do I characterize a point. If I know that it is a local optimal to a constant optimization problem, so that that is the first question how do I characterize a point. If I know it is a local minimum of a constrained optimization problem, the text that we are going to largely follow here is the following is a fabulous book called the foundations of optimization.

Only you will talk about the differentiable case, we will not go into the non differentiable case at all, and that will cover at least ten lectures would be needed to complete Karush Kuhn tucker conditions. I would just say possibly for the next ten lectures we will really be bothered about knowing about the Karush Kuhn tucker conditions, we will solve a examples and those examples will be very, very important; and very, very important examples as far as optimization goes, so this is the book by Osman guler a very famous optimizer works in the US. Of course, in in university of paul mary landed baltimore Baltimore, and this book was published by springer under the g t m series or the graduate text in mathematics series in 2011.

I suppose so this is the any anybody the who is looking at optimization from a certainly higher point of view, from the graduate perspective should really go for this book I would rather say that in, if you look at the NPTEL web site, I find that there are larger

courses catering to optimization operations research, those who would be just bother about knowing some techniques of how to compute a problem in various situations, I am not really knowing about the deep issues involved the mathematical issues involved in optimization. Then I would rather tell you to concentrate on those courses rather than concentrating on this course, because this course is given slightly at a graduate level or rather, you can say quite quite or bit of stuff would be at the graduate level.

So we would really like you to get involved, and know the mathematics behind optimization. So, this course is essentially telling you the math behind optimization not just you know telling these are the problem, you do this, do this do this what you will get call it a solution the point is that, in most algorithms what you will get? You can never call it a solution the art of optimization, you should know that how good is your solution, how can you estimate the goodness of your solution. So, how do you do that that is also a story, which you will tell you in the form of a section called error bounds which will come later on ok.

So here we stop, and here we ah from tomorrow we will start Karush Kuhn tucker condition, we talk about the history, first we start with the (()) condition inequality constraints the types of problems are which come in those sort of cases. We can talk about then we talk about both equality inequality constants how to really get those conditions you might think that. I have learnt multiplier in calculus whether that is doing something with constant optimization, but there you never learn that you really have to guarantee the existence of such a multiplier, so that the at the actual solution of the problem is nothing but a critical point of the Lagrangian of the function such a thing has to be guaranteed, and here we show such things right.

So we stop here, and I would rather say that this course would be given from this point onwards it was earlier, if you look at the other things, it was given at quite a simple level, it would not be I understand, but I wanted to keep this course very simple. But in order to give a much more in depth view of optimization to really tell you what optimization is all about what the hell is actually going on, you need to know the math deeply you understand the math properly. I am afraid possibly many of you would like to away from mathematics, but I am afraid optimization is a mathematical subject. And a deeper idea about optimization is not possible without a understanding of its mathematical principles. Thank you very much.