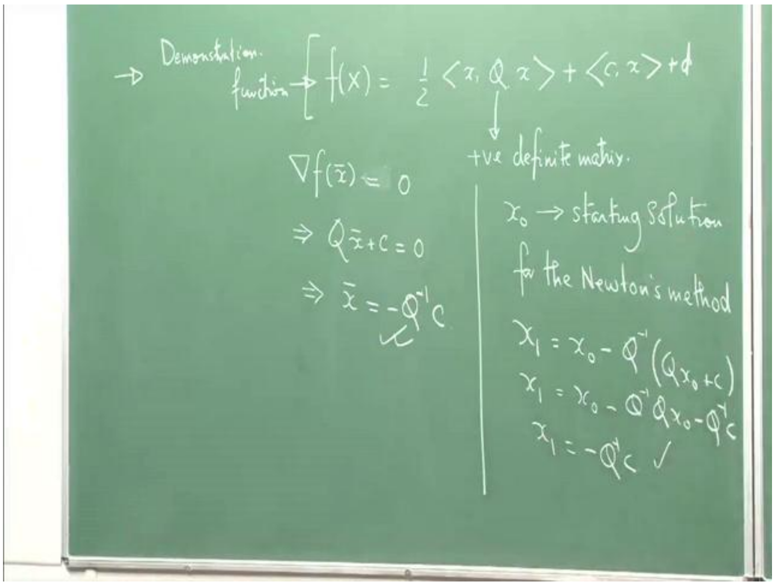
Foundation of Optimization

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

Lecture – 10

Ok, today we continue our discussion of conjugate directions method. And, later on conjugate gradient methods, which are important class of algorithms for solving un constant so, called non-linear optimization problems.

(Refer Slide Time: 00:46)



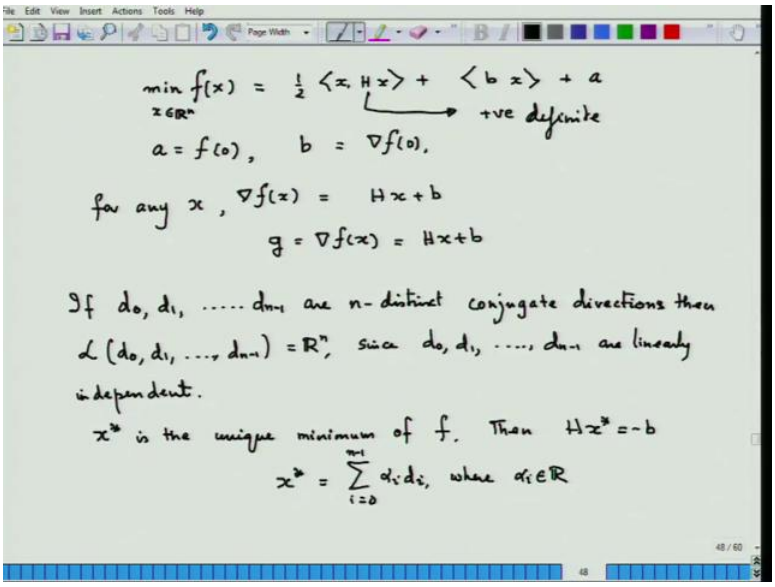
But, I begin with trying to give you a solution for this problem, where I said that; if you have a function f (x) given like this, where it is a quadratic function with Q being a positive definite matrix. Then, we said that; if you apply Newton's method to solve this problem then, you can solve this problem in just 2 steps. So, we shall use this idea of the Newton method to see that we can actually solve this problem in 2 steps. This was one of the home works, which I had given in just in last lecture. So, let me just try to solve it for you, but it does not mean that all the home works would be solved, because you need to try yourself and have the confident about whatever answer you are giving.

So, if you look at this; if this is a convex function so, it is critical point is the global minimum and in this particular case this is a strongly convex function and hence, has a unique minimum. Now, if x bar is really the minimum then, it must be satisfying. I am sorry, grad f x bar is equal to 0, and that would imply in this particular case because the gradient at x bar is this. Or the solution is now, let me start with a guest solution x naught, x naught is a guest solution right, starting solution for the Newton's method.

Now, once you have taken a guest solution; your next solution assumes that this does not give me the solution. Then, the next solution is so, this first step is choosing x naught the next step is getting x 1 which is x naught minus the hessian at x naught is Q so, inverse of Q into the gradient at x naught. So, that will lead to x 1 is equal to x naught minus Q inverse Q x naught minus Q inverse c. So, this Q inverse Q is identity so, this will cancel so, this will become x naught and so, x naught minus x naught. So, it is x 1 is equal to minus Q inverse c and that is exactly what the solution should be and hence in just 2 steps you have actually solve the quadratic problem.

Now, this quadratic this sort of problem would continue to be very useful as a demonstration tool, this is usually used as a very important demonstration function to demonstrate things or prototype function. So, this; a large number of properties are first checked of this class of functions to see whether algorithms are working well, with this class of functions. If they start working well with this class of functions then, they are working well with many other class of they could possibly be doing well with some other class of functions. So, let us go into the conjugate gradient methods or conjugate directions method again; and, we again follow this book by ((Refer Time: 05:14)) practical optimization whose reference I had given to you in the last class.

(Refer Slide Time: 05:24)

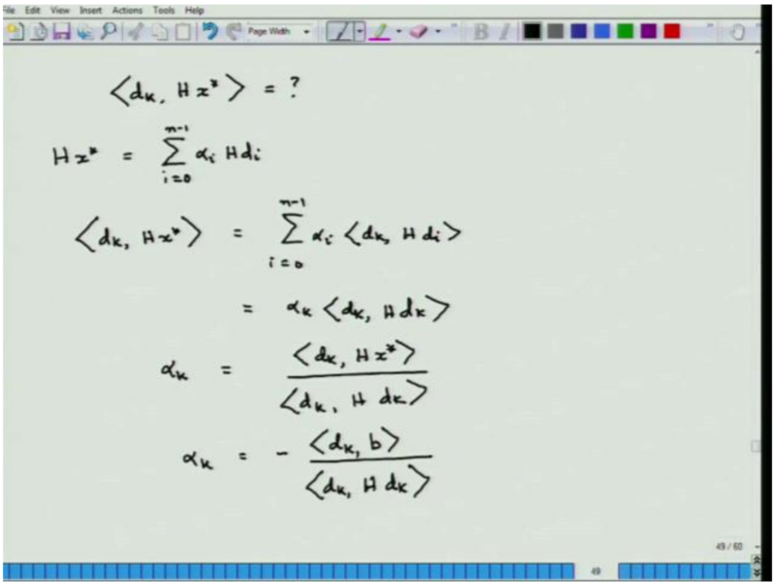


Now, our problem would be to minimize this function f x over whole x in R n, where I will follow their notations so, a g is actually the hessian matrix of f so, we just write H. Now, of course; a is nothing but f (0), and b is nothing but the gradient of f at 0, because for any x f of x is sorry, the gradient of f at x is nothing but H x plus b. So, you put x equal to 0, the grad f 0 is b. So, in general we for short hand we will write g is equal to grad f x which is same as H x plus b. Now, if d 0, d 1, d n minus 1 are n distinct of course, we are only concerned about the case, where H this function this mapping this hessian matrix is positive definite.

This is a very important thing we are only bothered about positive definite. So, if n distinct conjugate directions then, n distinct conjugate directions then the linear span of there is a subspace generated by is R n. Because you see there are n linearly independent vectors in R n, and their span would naturally generate R n, right. You cannot have n linearly independent vectors in R n and whose span is not generating R n because so, they will form a basis since; that is what we proved for the specific case when H is positive definite. So, any x star that you take may be the solution so, any solution of this problem. So, if x star is a solution which is the unique solution; will exist for this problem. I have not told you a detail as to why is solution would exist for this problem, but just accept for the time being that the solution exists for this problem.

Because take taking details would push me into much more deeper details. So, we are now talk on done all these things. So, of course; you can always find a minima because if you take the derivative and put equal to 0 you will have x star equal to this. And, because it is a convex function, and of course; there is a unique critical point and that would be the global minima, because the function is convex. So, let x star is a unique minimum of f say assume that. Then H of x star is equal to minus b because grad g x star is 0. So, grad f x star is 0 so, but whatever x star is an element of R and n and we can write always x star is summation, where alpha i are some scalars some real numbers.

(Refer Slide Time: 10:17).

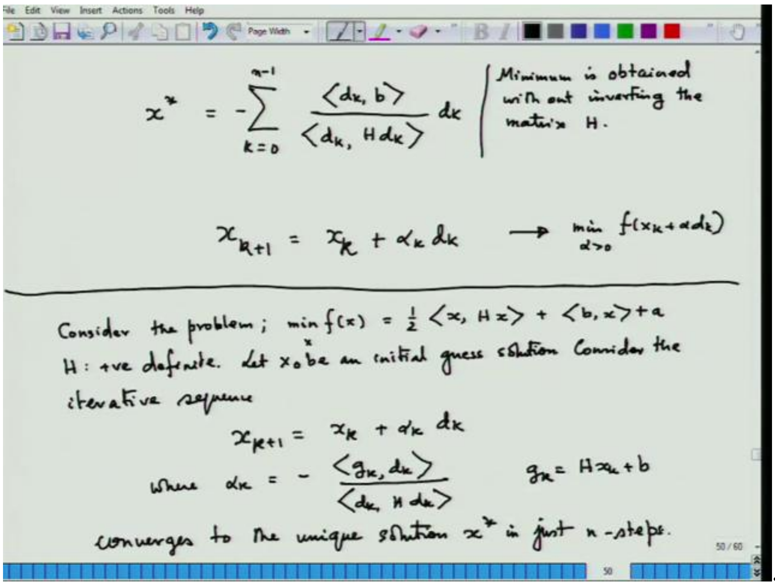


Now, let us do one thing look at inner product for any k among this d 1, d 2, d k. So, we will ask our question what is this? Now, H of x star by rules of simple rules of matrices or linear operators. If you want to say is nothing but H of that I hope everyone would agree, you could put i equal to 1 2 n also. I am just following the methodology of this book so, I am just trying to maintain the symbol so, that they are I mean if you read this book you will not get into much trouble. Now, once you do this then, I would have d k H x star would be summation alpha i d k H d i.

Now because there are conjugate directions except the d k everything else would be 0. When k is not equal to i this is 0 so, what I finally, get is alpha k d k H d k that is what we will get. So, you see what I am doing is I am trying to compute the alpha i's or alpha k's whatever, which is nothing but the coordinates of the vector x star the solution in terms of the basis d naught d 1 d n minus 1. Now, alpha k can be written as d k H x star now, because this is positive definite and d k's are not equal to 0, because there part of part of a set of linearly independent vectors. So, this is strictly bigger than 0 so, i can write this as d k H d k basically I can divide it both sides.

Now you also know that H x star is equal to minus b because I have assumed that x star is the unique minimum. So, what I would have here is alpha k is equal to see once I know alpha k i basically know the x, because that is it for each k equal to 0 to n minus 1 if I know alpha k i basically know the x. So, x star can now be easily computed once you know the from the problem data, you can easily compute because you have to compute this k's for n alpha k's you have to compute. So, basically in n steps essentially you really know the solution, right.

(Refer Slide Time: 14:00)



So, then x star can be now written as summation there will be a minus sign i equal to 0 to n minus 1 d k inner product b divided by the same d k H or I should write k equal to 0 to n minus to be more precise d H d k into d k. This is exactly the expression of the so, what have you done, what have I achieved here? I have been able to find the minimum of that problem minimum is obtained without inverting H. So, if H is very large then, we really do not want to invert it because the cost of inverting a matrix is pretty high. And, you see this problem data once b is known, H is known these two data are known c is not really in nothing to be bothered about this a, b is known H is known and d's are known. So, I can easily compute out the extra and the interesting thing that we have not inverted the matrix H, this is one way of doing it.

And, another way of doing it is that; how to use more iterative approach here, what I have done that I have I know that x star is a solution. So, I know what is the optimality condition which satisfy the, for that unique solution. And, then I have used the fact that d is a linearly independent and went in and solved it. The question would be that can I develop some sort of iterative scheme, any sort of iterative scheme. And that iterative scheme would lead to the solution in n steps of course; here you see we have solved it in n steps, because for each of alpha 0 alpha 1 dot dot dot alpha n minus 1 I have to compute the alpha naught, right.

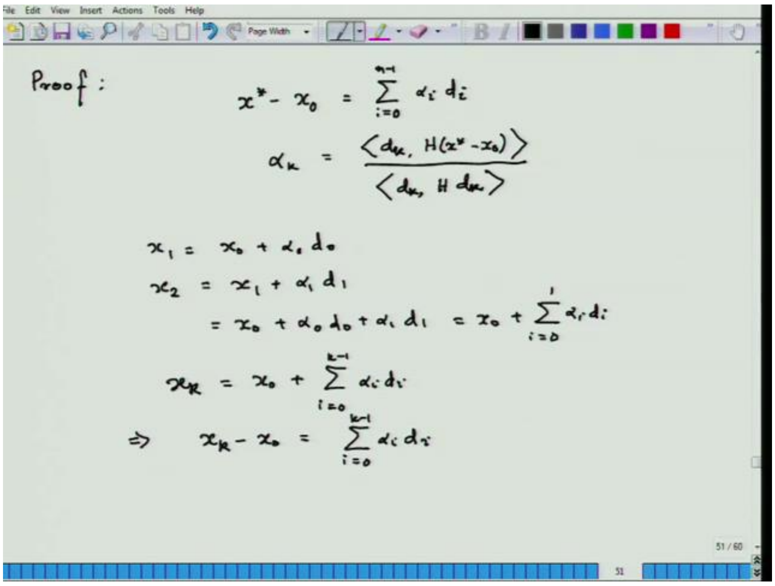
So, basically I have to change my d k's for every alpha's so, for every alpha 0, I have to take d 0 and so, and so. If my conjugate directions are known, what we can also do is that we can generate iteration. See here what we have done is a direct approach, we know that this x star would be there is a unique minima x star and that can be expressed as a linear combination of summation alpha a d i. And, we know that if i should be able to using the fact that these are conjugate directions, I should be able to calculate all these coordinates alpha i's. Once I know that I can easily write down x star so, that is essentially the process that we have used, but this is not a iterative process. That it is more computationally extensive that is you are doing in n different steps.

So, what we can write down is that using these conjugate directions, which are actually not descent directions. We can still write down an iterative line search process where you search along these conjugate directions. So, you have x naught from there you use d naught to get go to x 1; x 1 is x naught plus alpha naught d naught so, you at every step basically find that alpha naught. So, what we are going to now interpret and show is that this scaling factors that is if you want, if you are in k and you want to go to k plus 1 for this particular problem that we have studied with x k. We want to show that this alpha k is actually the coordinate associated with x k, this alpha k this length, right.

This alpha k that you see here, can actually be computed out so, basically you are you are at x k and you have to go to d k, you use d k as your direction along with which you move. And, you compute x k plus 1 and then, you find out a alpha k, right. So, to what extent you will go so, that you maintain a drop in the, or you minimize the function f of x k plus alpha k d k. And, you find an alpha k which minimizes this function basically what you do? You find the alpha over alpha greater than 0 you minimize the function f of x k plus alpha d k. But very well you do not know that this d k is a decent direction, you know only that it is a conjugate direction, but these sequence in n steps x naught x 1 x 2 x 3 x 4 dot dot dot x n minus 1 x n minus 1 will be the solution of the problem. So, this is a very good approach by which you can actually program in to solve this particular type of problem and very useful in the case of large number of variables.

So, here is our next result, but this again concerns the same function f x so, consider min f x over x. H positive definite, let x naught be an initial guest point, initial guest solution. Then, consider the sequence, consider the iterative sequence x k plus 1 is equal to x k plus alpha k d k, where alpha k is equal to minus g k, d k, d k H d k of course; g k is nothing but H x k plus b. And, this iterative sequence, where alpha k is this converges to the unique solution x star. Here, what we have done we have computed every alpha i's and then computed the x star, which is quite a heavy computation. If you really want here you are generating just sequences just getting a new taking a new starting with d z then going to d when you come to x 1 from x 1 to go to x 2 you use d 1 from x 2 to x 3 you use d 2 and so, on. And, you keep on generating these points and these points in finally, n steps x n minus 1 would be the solution.

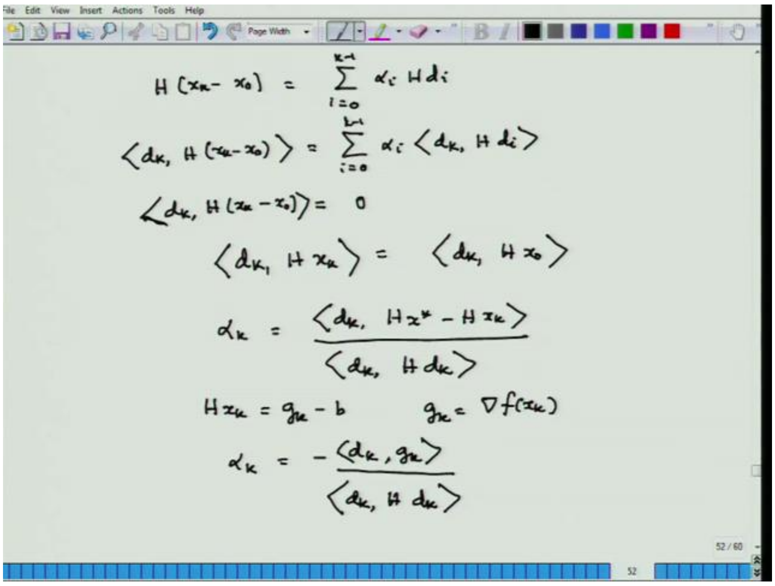
(Refer Slide Time: 23:36)



Let us try to give a proof of this fact so; there is a unique solution to this problem which we know. So, what we can do and x naught is my starting solution so, considered x star minus x naught so, I do not know this thing, but I know this solution x naught. So, this can be expressed as summation i equal to 0 to n minus 1 alpha i d i by the same way that we have evaluated. We can now compute for every k alpha k is d k H x star minus x naught divided by d k H d k in the same process that we have used. Now, if you observe what I have done, you have x 1 is x naught plus alpha naught d naught. While x 2 is nothing, but x 1 plus alpha 1 d 1 right, which is nothing, but x naught plus alpha naught d naught plus alpha 1 d 1.

That is nothing, but x naught plus summation i is equal to 0 to 1 alpha i d i. so, basically the iteration x k can be written as x naught plus summation i is equal to 0 to k minus 1 alpha i d i. So, this would immediately imply x k minus x naught is summation i is equal to 0 to k minus 1 alpha i d i. So, this is the way things are calculated so, you know the iterations are also very similar. So, what we will discuss more about this iteration in a short while after we finish doing what we were discussing.

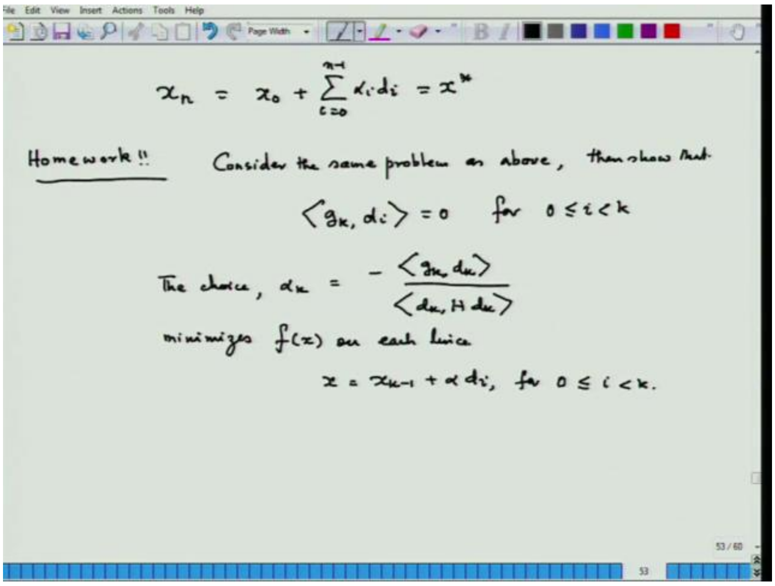
(Refer Slide Time: 25:41)



So, again observe that H of x k minus x naught is nothing, but summation i is equal to 0 to k minus 1 alpha i H of d I, just by properties of matrices. Then, d k H x k minus x naught is equal to summation i is equal to 1 to k minus 1, I am possibly skipping some steps. So, again in the same way using the conjugate that these are conjugate directions because these are conjugate direction for all k not equal to i this is 0. So, this would finally, give me that d k H x k minus sorry, so; what it should give me here I have all i from 0 to k minus 1 so, k is not here so, for all i from 0 to k minus 1. So, none of the i's are k's so, this would be all 0 so, this is 0 so, this gives me d k H x k is same as d k H x naught. So, alpha k need not now depend on x naught so, alpha k which we had computed out in this page, this one this alpha k can now be written as d k. Because d k of H x naught is H x k so, it can be written as d k H x star minus H x k d k H d k.

Now, we already know that H x k is nothing but g x g k minus b that this is nothing, but the g k is what? g k is nothing, but the gradient of f at x k. So, I can write here that as g k minus b, but what is H x star? H x star is nothing but minus b because H x star is the unique minimum. So, ultimately alpha k can be now written as minus d k g k so, you have to put here H x k as g k minus b. And, so, it will become minus g k plus b while H x star is minus b so, that minus b plus b cancels and minus d k g k would be the thing left on the top. And, so this alpha k is now so, that is what that is exactly the alpha k that we wanted.

(Refer Slide Time: 28:50)

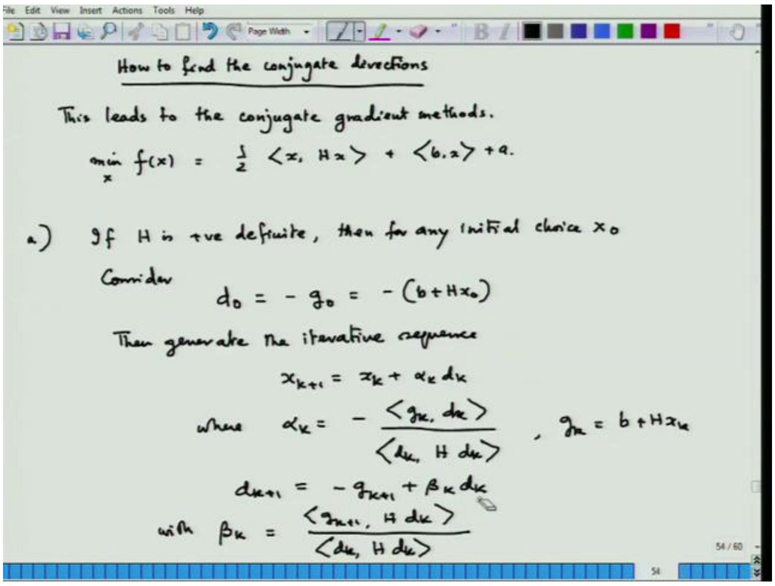


Now, x n the nth step is x naught plus i is equal to 0 to n minus 1 alpha i d i, but x naught plus this, and the very beginning what we have written? This is x star is equal to x naught plus summation alpha i d i. And, now from our iteration scheme we have x n equal to x naught plus i equal to 0 to n minus 1 alpha i d i and this is nothing but x star. So, the x n is nothing, but x star. So, after iteration from 0 to n minus 1, we have done in n minus 1 step in n steps, we have reached the solution. So, this is a very beautiful thing about conjugate rate directions method. So, that shows that gives us a possible hope that we can use this method flexibly for certain classes of problems, which need not be so, nice that it need not be a quadratic function with positive definite hessian; that is having a unique solution. So, if there even if there is non-unique solution there must be so, me ways to handle such problems using this trick. So, we would tell a little bit more about this iteration, if you do not mind; if I was a student in your place I would first ask a question that how do I get all those conjugate directions.

How do I know, how do you I cannot just arbitrarily try finding this is a conjugate direction take the H and try to do this straight. I have been given the same function same problem as before minimize f x like the same one a quadratic function with positive definite hessian, but that is not the real question the question is how do I know that how to generate my conjugate directions. Now, before I tell you that how do I start generating conjugate directions. So, I cannot generate it from the blue, I cannot just take arbitrary in vectors and then, try all those things with H that is stupidity and waste of time. So, numerical optimization the important thing is that you must tell me how to generate each of the objects that I need. And, then before I tell you how to do that; let me now give you a few, homework actually, this is not very difficult.

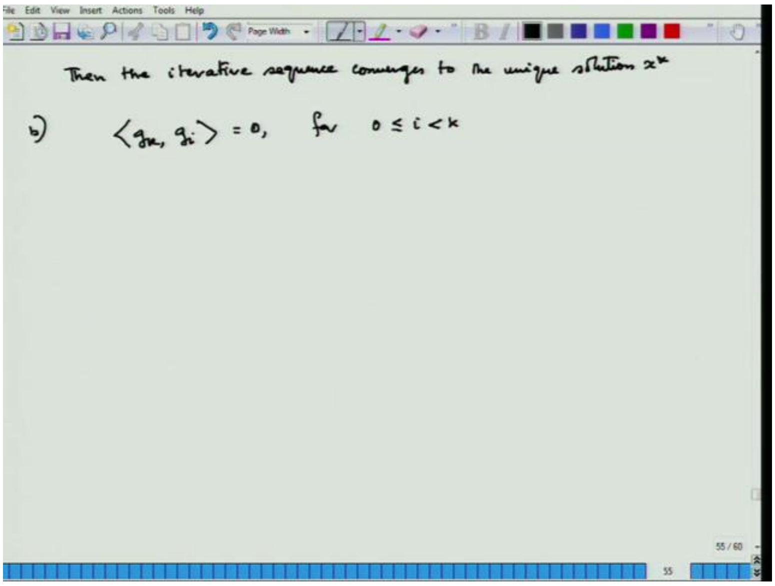
So, you can try it out if you are stuck maybe we can try it out later on. So consider the all the hypothesis of consider a problem in the above result, same problem in the above result. Then, show that g k d i is equal to 0 for all i till k not k to k minus 1 so, basically for all i till k the gradient that you obtain is actually perpendicular to the conjugate directions till k, there you have not reached k. Here, they are strictly less than 0, they are strictly less than k and the choice of alpha k is equal to choice of alpha k with this alpha k. The choice alpha k minimizes effects on each line x equal to x k minus 1 plus alpha d i for. So, this would be home work which I am not going to explain in detail.

(Refer Slide Time: 33:30)



Now so, our next attempt is how to find the conjugate directions? Now, these attempt this leads to the conjugate gradient method so; consider the same problem, which I am writing again, but I do not think I would like to repeat it every time. So, there are 2 results, which I want to write down, if H is positive definite then, for any initial x naught initial choice x naught. So, for any initial choice x naught consider; my initial direction d naught as minus g naught which is nothing but. Now, this then generate the iterative sequence, where alpha k as before and of course; you have where g k of course, is the gradient at x k, which is d plus H x k and d k plus 1 minus g k plus 1 plus beta k d k with beta k, this is a very important scale factor now. Because it allows you to find from so, if I have d 0 so, d 1 is minus g 1 plus beta 0 d 0. So, that is how I am generating the orthogonal and the conjugate directions, we really have to prove that these are conjugate directions. Then only we can apply the previous results.

(Refer Slide Time: 37:40)



g k plus 1 inner product H d k these are very nice. So, this is a very standard denominator in conjugate gradient method. So, this is what you will have and so, then generate the iterative sequence then, the iterative sequence then, the iterative sequence converges to the unique solution converges to the unique solution x star in n steps. Of course, b we have g k g i is equal to 0 for so, it is for g k orthogonal to all the other g i's whenever i is strictly less then k. So, this is what we are going to prove in the next class, and once we prove this we will write down the conjugate gradient algorithm. Now, there is a there are various ways of choosing this beta k, and that leads to various types of a conjugate gradient algorithms specially; when you are talking about non convex or non minimization of non quadratic functions, you have very different types of that leads to methods like Fletcher reeves method pawls method and so and so, forth.

So, we will for example; do in detail the pawls we will write down the Fletcher reeves algorithm, but we will write down also in details. We will study in detail the pawls algorithm including the proofs. So, that would be a very good introduction to the conjugate gradient methods and then, we will switch over to quadratic Newton methods. So, that is our plan for the few coming lectures; and then once that is over we will try to understand a bit about trust region methods which are very modern techniques, which have been used at present and studied and understood at present. And, then go over to the convergent to the theory of non-linear optimization largely of in constraint optimization we will start talking about the Karsh Kuhn tucker conditions, and related issues.

Thank you very much, for the attention.