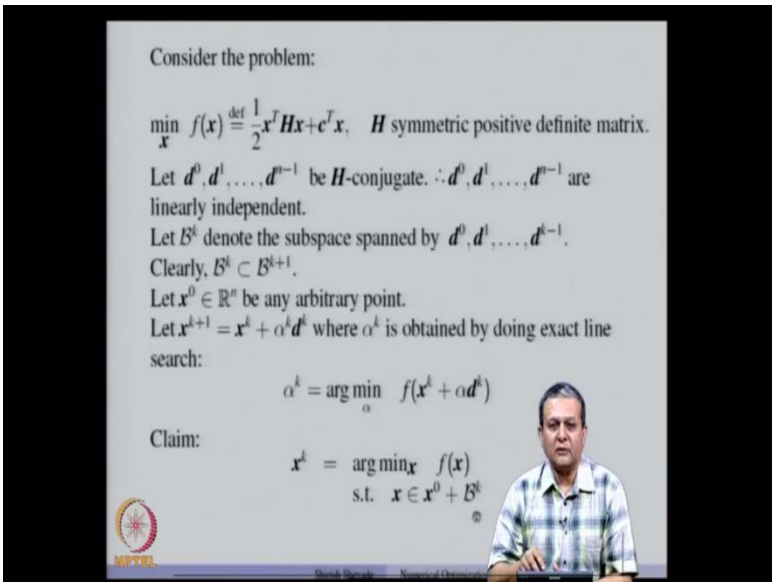
Numerical Optimization
Prof. Shirish K. Shevade
Department of Computer Science and Automation
Indian Institute of Science, Bangalore

Lecture - 19
Conjugate Gradient Method

(Refer Slide Time: 00:53)



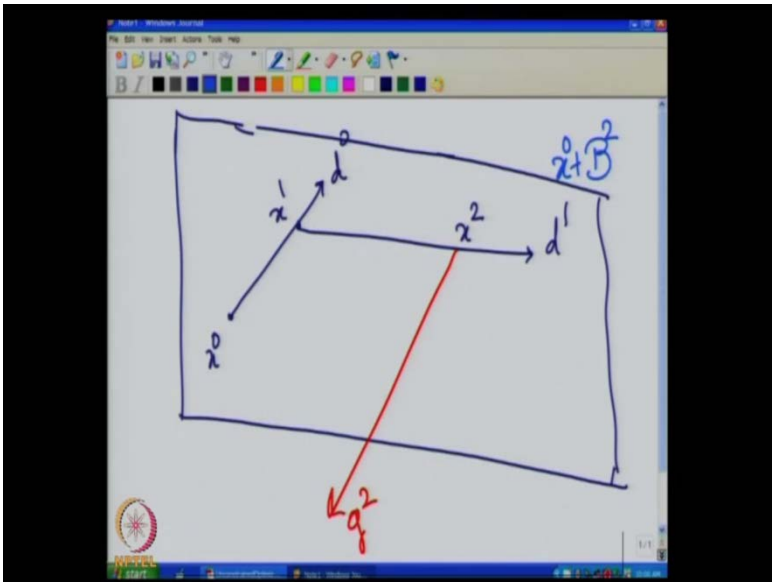
Welcome back to this series of lecturers on optimization. So, in the last class, we started studying about conjugate directions and we saw some of the properties of conjugate directions. So in today's class, we will disuss about expanding subspace theorem, and the use of conjugate gradient method to quadratic as well as non-quadratic functions. So let us first consider the quadratic function, which is to be optimized. So, let the function to be optimize be half x transpose H x plus c transpose x, where H is a symmetric positive definite matrix. And let us assume that we have n conjugate directions d 0 to d n minus 1, and we already saw in the last class that these directions which are H conjugate are linearly independent. So they form, basis for n-dimensional space.

Now, let us take the first k vectors d 0 to d k minus 1, and since they are independent they will form a basis of a k-dimensional space, which is a subspace of array. So, let us denote by B k that subspace with of R a which is spanned by d 0 to d k minus 1. Now, vector d k is independent of this, so clearly the space spanned by the k vectors d 0 to d k

minus 1 is a subset of the subspace spanned by k plus 1 vectors d 0 to d k, because d 0 to d k are linearly independent.

Now, let us take any arbitrary point in the n-dimensional space, and will denote by denote x k plus 1 by x k plus alpha k d k asusual, where alpha k is obtained by doing the exact line search that is alpha k is optimum minimizing the function f of x k plus alpha d k with respect to alpha. Now, x k and d k are known and alpha is a variable here and the value of alpha that gives the minimum will be denoted by alpha k. Now, under these conditions that d 0 to d n minus 1 are independent, now the claim is that then we find x k as by using the conjugate direction method, the minimum value of f of x in the space, spanned by x 0 plus d k, that is nothing but x k or in other words.

(Refer Slide Time: 03:42)

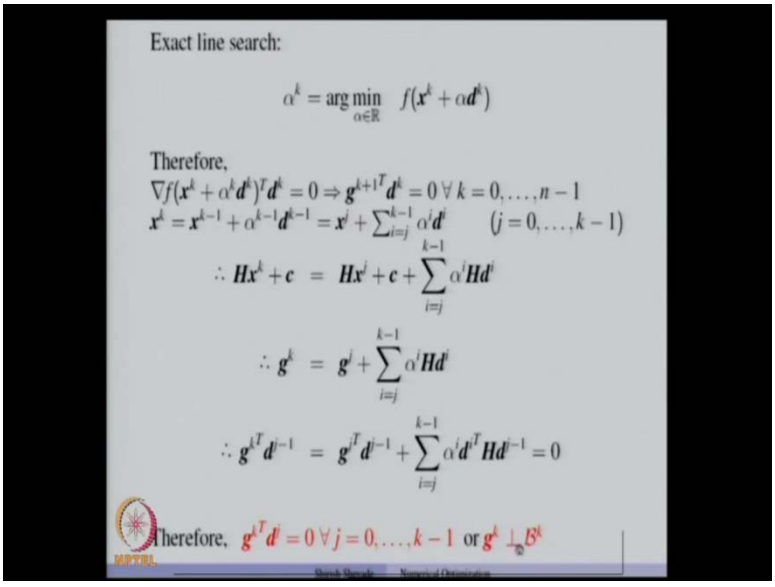


So, suppose, we have a point x 0 from which we start, we get a direction which is say d 0 and then we get a point which is x 1 on this direction by doing the line search. And then we have a direction d 1 along which we do again line search, to get a point x 2. Now, the important point is that, d 0 and d 1 are independent, so they will span a two-dimensional space. So this is two-dimensional space spanned by d 0 and d 1, and now if we considered g 2, the gradient direction at x 2. Now, g 2 will be perpendicular to the space spanned by d 0 and d 1 and containing x 0, so this is the space which contains x 0 plus B 2.

So, the space containing x 0, and the space spanned by d 0 and d 1 which are linearly independent, now the expanding subspace theorem says that the direction g 2 that you get at x 2 is orthogonal to this space. Now, if we choose the direction d 2, which is in the which is along the or which is the combination of g 2 and the previous directions. Then what happens is that when we do the minimization of f of x in the now three-dimensional space spanned by d 0 d 1 and B 2. Then the claim is that, we will have got the minimum of f x with respect to the three-dimensional space d 0 d 1 and d 2, and if we keep on repeating this, for n steps. Then we will have found the minimum of f of x with respect to the n-dimensional space and that is what we will see now.

So, the claim is that, x k that we get is the minimum of f of x subject to the constrained that x belongs to the space spanned by d 0 to d k minus 1, and the space containing x naught. Now, remember that x naught is any arbitrary point, so it really does not matter what our x naught is as long as we get the conjugate directions, then we can say that the x k that we get is the minimum of f of x subject to this space. And if we keep on repeating this then finally, what happenses that at the end of any tritions we will have found x n which will be equal to x star.

(Refer Slide Time: 07:32)



So, let us true this thing first, now we know that alpha k is obtained using exact line search. So, this is the one-dimensional optimization problem, and we know that the way to solve it is that take the gradient of this function and equated to 0, because it is one-

dimensional optimation problem. And note that f is also convex quadratic functions in this case. So at the new point, the gradient of, of the function at a new point is perpendicular to d k.

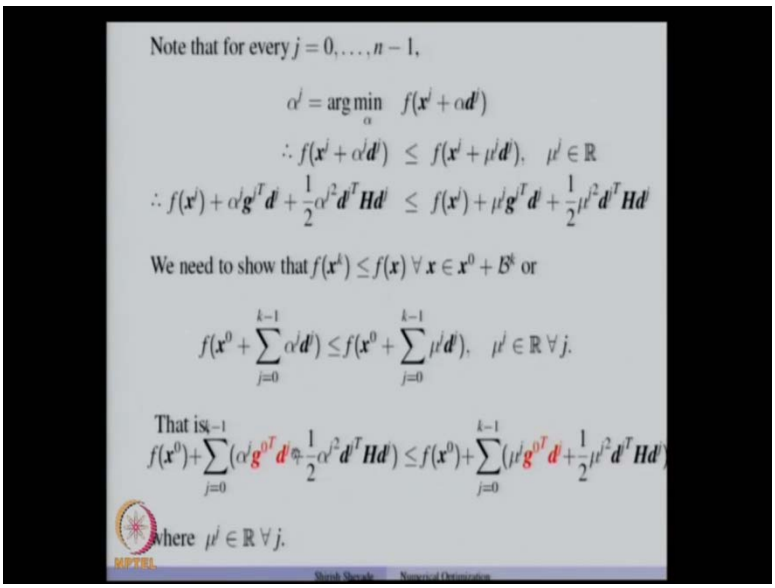
And therefore, we can say that g k plus 1 transpose d k is equal to 0, for all k going from 0 to n minus 1, so or in other words g k plus 1 is perpendicular to d k. So in our previous example g 2 will be perpendicular to d 1, but the claim is that g 2 will be not only perpendicular to d 1. But also, d 0 or in general g k will be perpendicular to d 0 to d k minus 1 that is the claim, and let us see how to show that. So, now, let us look at a general step where x k is equal to x k minus 1, plus alpha k minus 1 d k minus 1 and x k. Therefore, can be written as some of x j and sigma alpha i d i, i going from j to k minus 1. So x k can be written in terms of x j and the linear sum of the remaining conjugate vectors d i. Now, now if we premultiply both sides by H and add c, so what we get is H x k plus c, that will be equal to H x j plus c plus sigma i going from j to k minus 1 alpha i H d i.

So note that, we want to find out what happens to g k plus 1 transpose d j for any j which is less than k. We know the g k plus 1 transpose d k is 0, or g k plus 1 transpose d j is equal to 0 when j is equal to k. But what happens when j is less than k, and that is why we do this multiplication so that now we can treat this as a gradient of the objective function. Note that, we are working with the quadratic function of the type f of x is equal to half x transpose H x plus c transpose x. So, these quantity denotes the gradient of the objective function at x k, and this quantity denotes the gradient of the objective function at x j, so we can write g k is equal to g j plus sigma alpha i H d i. And therefore, now we want to c what happens to g k transpose d j, so let us multiply both sides by take the dot product of both sides with respect to d j minus 1. So g k transpose d j minus 1 is equal to g j transpose d j minus 1 plus sigma alpha i d i transpose H d j minus 1.

Now, g j transpose d j minus 1 is equal to 0, because of this result. Here we have shown that g k plus 1 transpose d k is equal to 0. So similarly, g j transpose d j minus 1 is equal to 0, so this quantity is 0 and now what about this quantity? So you will see that the ds are the conjugate directions so d i transpose H d j minus 1, where i goes from j to k minus 1. So certainly, none of the i is going to take the valve g j minus 1, so all these quantities with respect to all i is d i transpose H d j minus 1 will be 0.

So together this whole quantity will be 0, and this will be true for any j. So therefore, what we have is g k transpose d j is equal to 0 for all j going from 0 to k minus 1. So, here were shown that g k transpose d j minus 1 is equal to 0 for all j, and from this we know that g k transpose d k minus 1 also equal to 0. So, in other words, what we have shown is that g k is perpendicular to all the d j j going from 0 to k minus 1, or in other words g k is perpendicular to the space spanned by d 0 to d k minus 1 and that space is nothing but B k. And if we consider the, space containing B k as space spanned by d 0 to d k minus 1 that is B k that contains x naught, then g k will be perpendicular to that entire space. So this is a interesting observation, and therefore we can say that g k is perpendicular to B k.

(Refer Slide Time: 13:04)



Now, again let us go back to the our earlier alpha determination of alpha j using exact line search. And that it shown here where we want to find alpha j by minimizing f of x j plus alpha d j with respect to alpha. Now, what we want to show is that, f of x what we want to show is that, now if we minimize f of x subject to the consraint that x belongs to this space, then that is nothing but x k. Or in other words, so x 2 here is the minimum of f of x, subject to the this space which is given by x 0 plus B 2. And now suppose if you find the direction d 3 using g 2 and other the d 0 and d 1, such that d 3 is d 2 is H conjugate to d 0 and H conjugate to d 1.

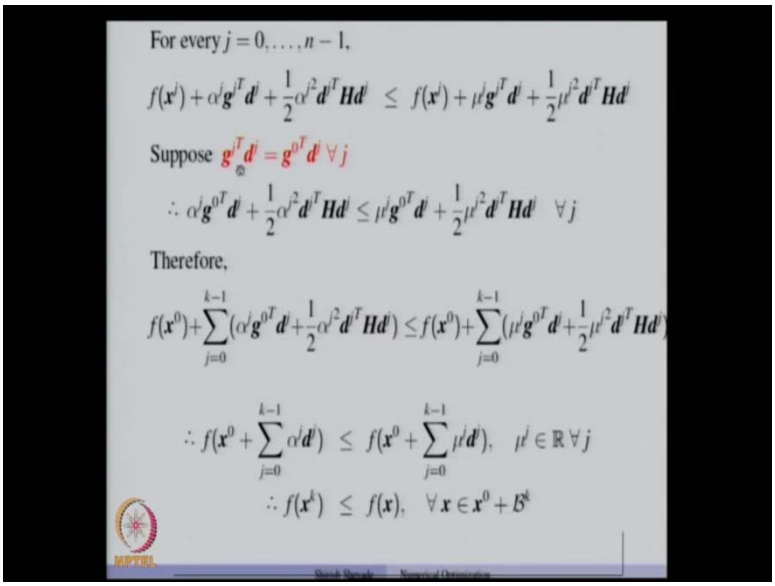
Then, when we find the point x 3 that point x 3 will be the minimum of f of x in the space, which contains x 0 and spanned by d 0 d 1 and d 2, so that is going to be our next claim. So let us see how to do that, now because of the exact line search we know that f of x j plus alpha j d j is less than or equal to f of x j plus mu j d j when mu j belongs to R. This is because of the exact line search that along the line, the value of the function at x j plus alpha j d j will be always less than or equal to the value of the function at any point on the line x j plus mu j d j where mu j belongs to R. And since the function f is quadratic, we can use the Talyor series to write f of x as to write the or to expand f of x j plus alpha j 2 gradient of f of j transpose d j plus half alpha j square d j transpose H d j. And that will be less then is equal to similar expansion of f, then using the Taylor series, so we can write this for all mu js belong to R.

So, what we know is that, along one direction, the function value is minimum at a particular point. So, if we consider if we start from x 0 and take d 0 as the direction x 1 is the minimum is x 1 is the point at which the function attends the minimum value. So as far as this direction is concerned no other point in this direction has a function value which is greater than which is strictly less than f of x 1. Now, the same thing holds here also, that if we start from x 1 and take this direction now no other point in this direction has a function value which is strictly less than f of x 2. So, this is true for one-dimensional case, but what is the guarantee that the x 2 that we here is actually the minimum of f in the space spanned by d 0 and d 1, and that is what we want to show. Now, what we need to show is that f of x k is less than or equal to f of x for all x in the space spanned by d 0 to d k minus 1 and that space containing x naught.

So, this is what we want to prove because that is our claim, so that means what we want to show if we write x k as x 0 plus sigma alpha j d j j going from 0 to k minus 1. Remember that these alpha j is are obtained using the line search in the respective directions d j and that quantity, we want to show that that quantity is less than or equal to f of x. Now, any general x in this space can be written as x 0 plus sigma mu j d j, where d j is the basis of this k-dimensional space j going from 0 to k minus 1 and muj is a real number so this is what we want to show. Now, if we expand this using Taylor series, because this is a quadratic function, what we can do is that we can write this as, f of the x 0 plus sigma alpha j g 0 transpose d j plus half alpha j squire d j transpose H d j.

Now, let us compare these 2 quantities, so suppose here, we write, we combine all the j is accmulated so far. So, we take the sum as, sigma j going from 0 to k minus 1 here and similarly, here and then combine these 2 quantities then we see that there is a difference of these 2 quantities, so here we have d j transpose d j and here we have g 0 transpose d j. Now, suppose we show that d j transpose d j is nothing but g 0 transpose d j then we can replace this quantity by g 0 transpose d j and then cancel f of x j from both sides and add f of x 0 on both sides and then what we get is this quantity? So what we need to show is that g 0 transpose d j should be equal to d j transpose d j. So, in other words, the dot product of g j and d j should be same as dot product of g 0 and d j, and if that holds let us see what happens. So suppose g j transpose d j is nothing but g 0 transpose d j remember that we had got this inquality earlier.

(Refer Slide Time: 19:33)

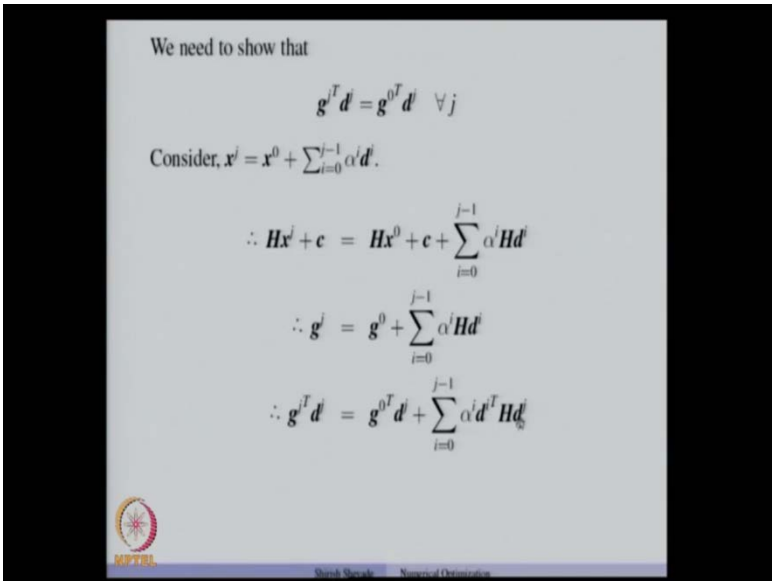


So, this is the inequality that we had got by doing line search along the direction d j. Now, suppose this holds, then what we can do is that we can write this quantity as so the f of x j gets cancelled. And we can write this as alpha j g 0 transpose d j this quantity remains the same is less than or equal to mu j g 0 transposed d j so the g j transpose d j in either side is replaced by g 0 transpose d j, and the reason for doing it is that then it we are more closed to the the desired thing.

Now, we just add f of x 0 on both sides, and then some over all j is. So if we add f of x 0 on both sides and some over j going from 0 to k minus, because for each j this inquality

holds. So if we sum them up, then what we get is f of x 0 plus sigma j going from to k minus 1 alpha j d j less than or equal to f of x 0 plus sigma j going from 0 to k minus 1 mu j d j. So, in other words, what we have shown is that at the end of k iterations, we will have reach the point x k and that f of x k is less than or equal to f of x for any x in that k-dimensional space containing x naught and spanned by d 0 to d k minus 1. So this is a very important observation, that f of x k is less than or equal to f of x for all x 0 for all x belong in the space x 0 plus B k. So all these was possible, because of the fact that g j transpose d j equal to g 0 transpose d j, now we want to see whether this really holds or not.

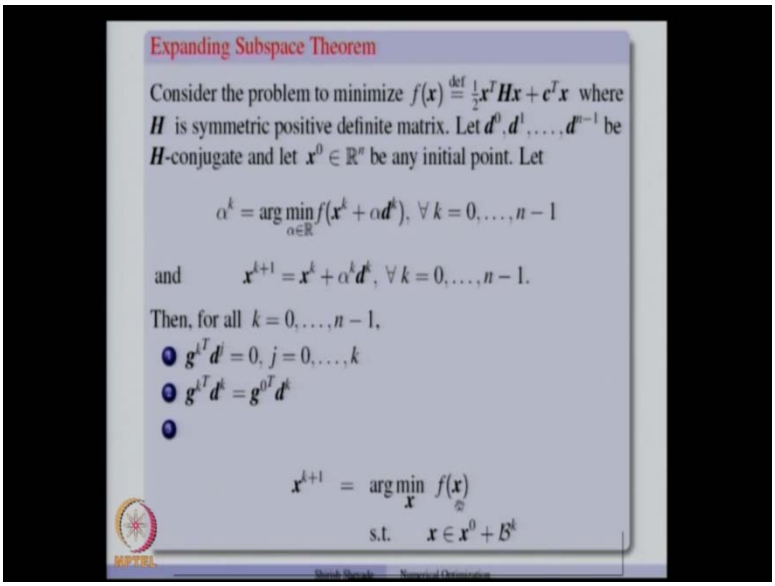
(Refer Slide Time: 21:43)



So, let us now show that g j transposed d j is equal to g 0 transposed d j, now consider any point x j which can be written as x 0 plus sigma alpha i d I, so which can be written as a point in j minus j-dimensional space spanned by g 0 to d j minus 1 and containing x 0. Now, again we want to show the relationship between g j transpose g j d j and g 0. So, we we need a gradient information, so we premultiply troughout by H and add c. So, the so what we will get is the gradient of f at x j on the left hand side, and this is the gradient of f at x 0, so that means g j is equal to g 0 plus sigma alpha i H d i. Now, so what we want to do is the that want to show what happens g g j transpose d j so g j transpose d j is equal to g 0 transpose d j plus sigma alpha i d i transpose H d j.

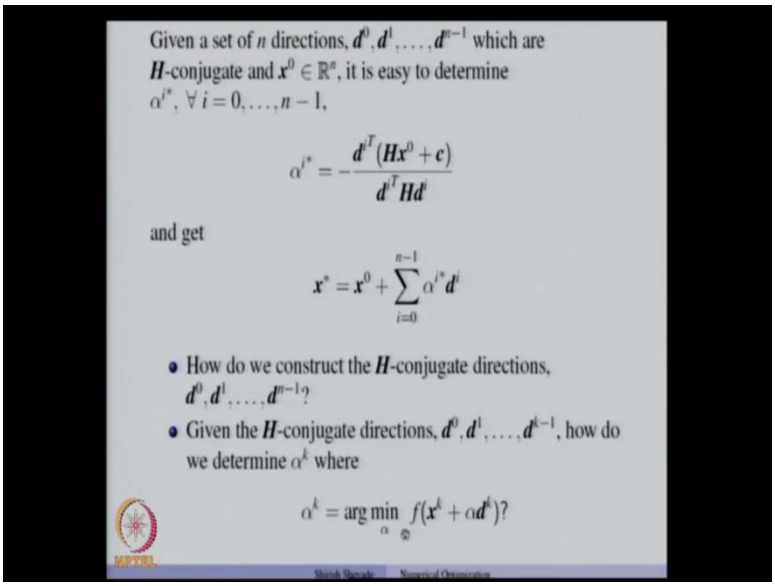
So, now you will see that, the i where is from 0 to j minus 1, so d j transpose H d i or d i transpose H d j will be equal to 0, because these are H conjugate directions. So, this entire quantity becomes 0 and therefore, what we get is g j transpose d j is equal to g 0 transpose d j. And because of that, we can say that if this holds then we can write g j transpose d j to be g 0 transpose d j and then this holds for all j so sum them up. Then we get this quantity then add f of x 0 on both sides the inequality does not change the direction. And therefore, we can say that f of x k is clearly less than or equal to f of x, for all x in the space x 0 plus B k. Now, if you repeat this step n times, at the end of n iterations we will have got the n conjugate directions d 0 to d n minus 1. And we know that they are linearly independent so at the end of n iterations whatever x n that we get will be, the minimum of the objective function.

(Refer Slide Time: 24:18)



So, we have this expanding subspace theorem, if we consider the problem to minimize f of x to be half x transpose H x plus c transpose x, where H is a symmetric positive definite matrix. And suppose d 0 to d n minus 1 H conjugate directions, so which clearly from the basis R n and x 0 is any initial point then and if alpha k is chosen using exact line search and x k plus 1 is said to x k plus alpha k d k. Then what we have is, g k transpose d j is equal to 0, so g k is perpendicular to the space spanned by d 0 to, this should be k minus 1. So, g k transpose d j is equal to 0 for all j going from 0 to k minus 1 g k transpose d k is equal to g 0 transpose d k. And the important thing is that x k plus 1 is nothing but arg min of f of x, x 0 plus B k.

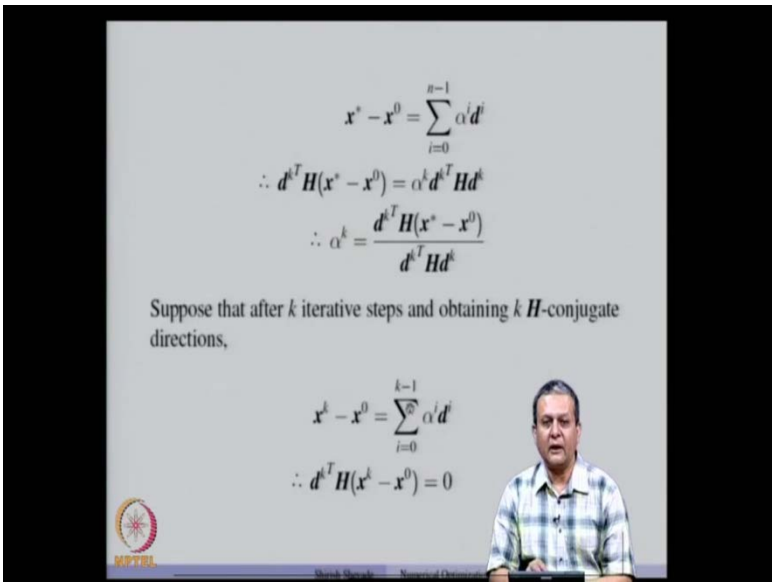
(Refer Slide Time: 25:36)



So, this is the important theorem, now we are given a set of n directions d 0 to d n minus 1, which are H conjugate, and x 0 is in R n. Now, we have seen that it is easy to determine alpha i star to be, minus d i transpose H x naught plus c divided by d i transpose H d I, and then we know that x star can be obtained using this all this is possible provided d 0 to d n minus 1 or H conjugate.

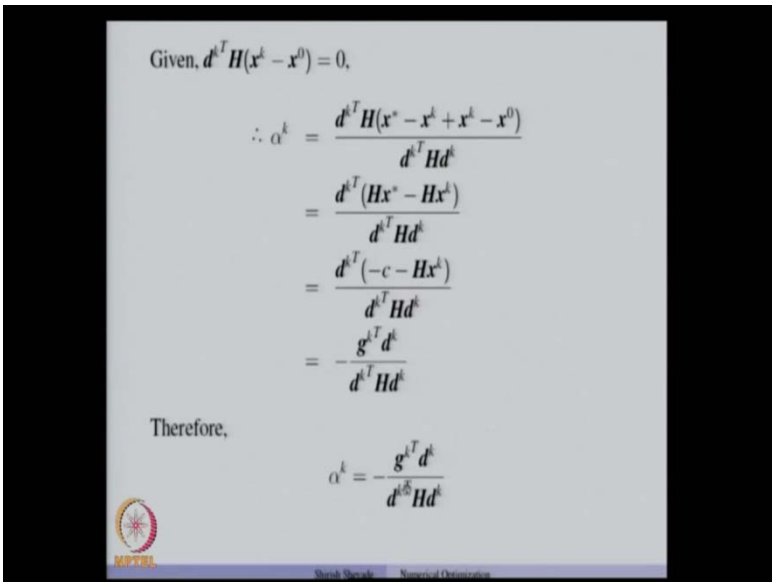
Now, the important question is that, how do we construct the H conjugate directions d 0 to d n minus 1, and then we will come up with a iterative procedure, which will construct d 0 d 1 d 2 and so on. So, at the end of k iterations, how do you find out alpha k star, based on the point x k, because this alpha i star depends on H x naught plus c and all the previous conjugate directions. So, so how do we get alpha k at a current point x k, so given the H conjugate directions d 0 to d k minus 1 how do we determine alpha k which basically the minimizer of this so does there exist a close form expression perpending alpha k. So, let us first look at this and then go to the other question where we want to construct the H conjugate directions.

(Refer Slide Time: 27:16)



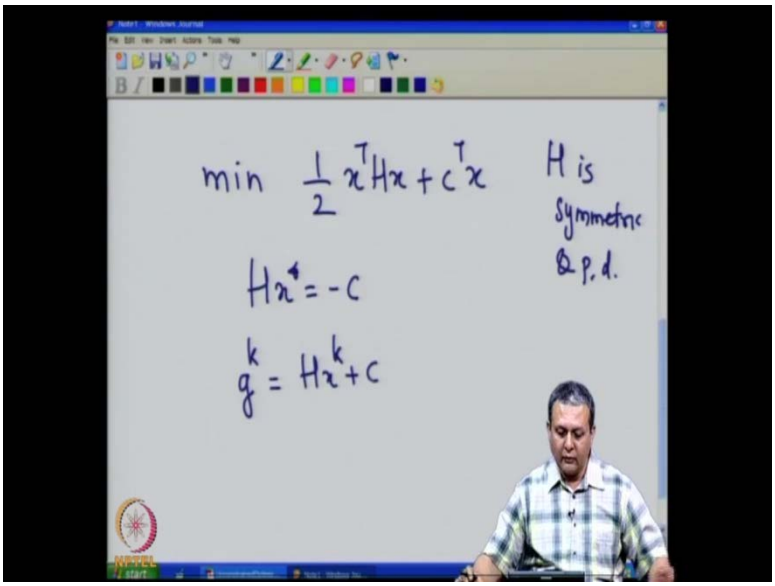
Now, note that x star can, can be written as a combination as a sum of x 0 and the linear combination of d i is where i going from 0 to n minus 1. So, if we premultiply throughout by d k transpose H then only 1 quantity on the right side remains the rest of the quantity become 0 and the quantity which remains is alpha k d k transpose H d k. And therefore, alpha k is equal to d k transpose H into x star minus x naught divided by d k transpose H d k. Now, this alpha k is still dependent on x star and x naught and we want to get the dependence of alpha k only on x k, so let us see how to do that. Now, after k iterative steps, and obtaining k H conjugate directions what we have is x k is nothing, but x naught plus the linear combination of the k minus 1 vectors k vectors d 0 to d k minus 1. Now, clearly, we can say that d k transpose H into x k minus x naught will be 0, because d k will be conjugate to all d i is going from 0 to k minus 1.

(Refer Slide Time: 28:46)



Now, we let us use this fact in this formula, so given that d k transpose H into x k minus x naught equal to 0,. We can write alpha k is in the previous formula the previously we had written that alpha k is equal to d k transpose H into x star minus x naught by d k transpose H d k., so we just add subtract and add x k in this formula. Now, if you expand this, so take the first 2 terms and combine them with this and then the take the last 2 terms, and combine them with this so d k transpose H into x k minus x naught that is equal to 0. So, the quantity involving the the last 2 terms vanishes, so what we are left with is d k transpose H into x star minus x k. Now, we know that the function that we want to optimize is half x transpose H x plus c transpose x.

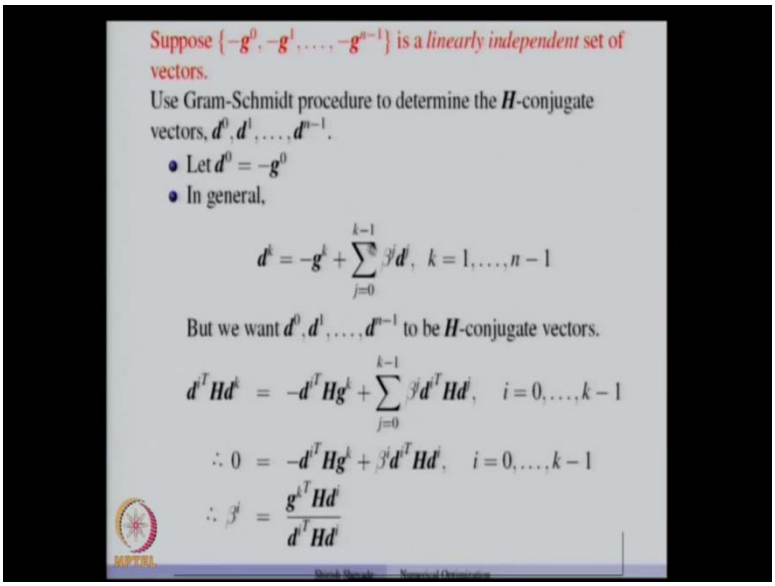
(Refer Slide Time: 29:53)



So, the function that we want to minimize is half x transpose H x plus c transpose x, H is symmetric and positive definite. So, taking the so we take the gradient equated to 0, so H x star should be equal to minus c. So, we use this fact, so this H x star is nothing but minus c. And so we replace H x star by minus c and again if we look at the gradient of the function so g k will be H x k plus c. So, if we look at the the formula here so what we have is negative of H x k plus c. So, which is nothing but negative of g k, so therefore, we have got alpha k using g k and d k and H.

So, we have got alpha k using g k d k and H, so you will see that this alpha k that we have got now does not depend on x naught, but it just depends on the the current gradient value. So, compare this alpha k with the alpha k that we have got earlier, that d k transpose H into x star minus x naught or even before that we had got alpha k which depended on x naught so we got the formula for alpha k based only on g k and d k.

(Refer Slide Time: 31:51)



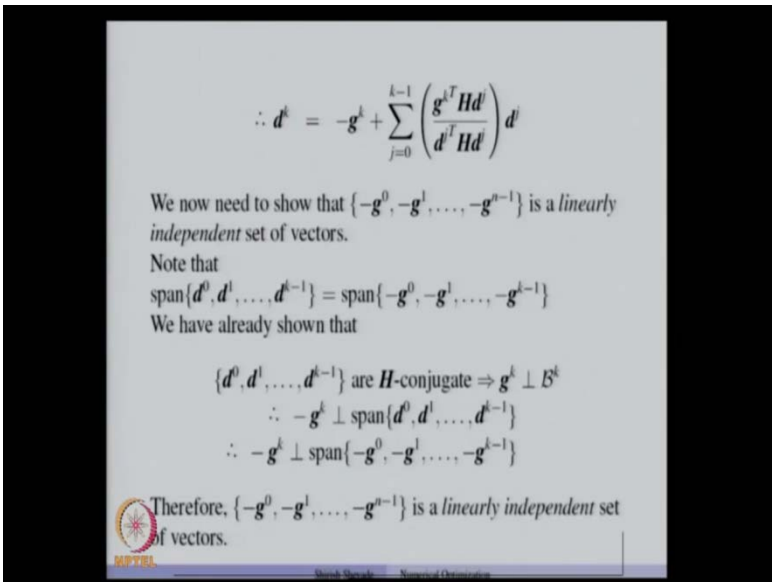
Now, the important question of finding the conjugate directions, H conjugate directions d 0 to d n minus 1. Now, let us assume that we have got n linearly independent set of vectors and those we are going to denote by minus g 0 minus g 1 to g n minus 1. Now, these are the negatives of the gradient vectors that we have got at every iteration, and let us assume that they are linearly independent. If they are linearly dependent then at some point of time we have the gradient is 0, so gradient is 0 for quadratic function means that we have from from the solution. So, this is the case where in n minus 1 iterations we have not found the solution. And finally, at the at the end of last iteration, what we have is minus minus g 0 to minus g n minus 1 forming a linearly independent set of vectors. Now, we have already seen that Gram-Schmidt procedure can be used to find out, the orthonormal basis given a set of linearly independent vectors.

Now, that procedure can be extended to find out the H conjugate directions d 0 to d n minus 1, and this is the basis that we are going to use. Rremember that, we are assuming that minus g 0 to minus g n minus 1 are linearly independent later on we will show that this is indeed case. So, to begin with, we start with d 0 to be the negative of the gradient at x 0 and in general we will write the Gram-Schmidt procedure as d k to be minus g k plus, linear combination of k vectors d 0 to d k minus 1. Now, this beta j is obtained using Gram-Schmidt procedure, and what we want is that d 0 to d n minus 1 to be H conjugate. So, if we premultiply this with d i transpose H, there i varies from 0 to k minus 1.

So, then what we get is d i transpose H d k is nothing but minus d i transpose H g k plus this quantity now d i transpose H d k is 0 because i is going from 0 to k minus 1 so this holds for every i going from 0 to k minus 1. So, we get 0 equal to minus d i transpose H g k plus, only 1 term in this expression remains and that corresponds to the vector d i. And therefore what we get is the expression for beta i and which is nothing, but g k transpose H d i divided by d i transpose H d i.

So, this is the usual Gram-Schmidt procedure, at i mean to get d k using minus g k which is the linearly independent vector in from this set. So, at the k iteration we get 1 vector from this set, and we use the combination of all the previous vectors to get d k now when we talk about H conjugate directions, this formula becomes much simpler and that is what we will see now. So, we will see that the beta j will be 0 for all j going from 0 to k minus 2, and only the k minus 1 direction comes into picture.

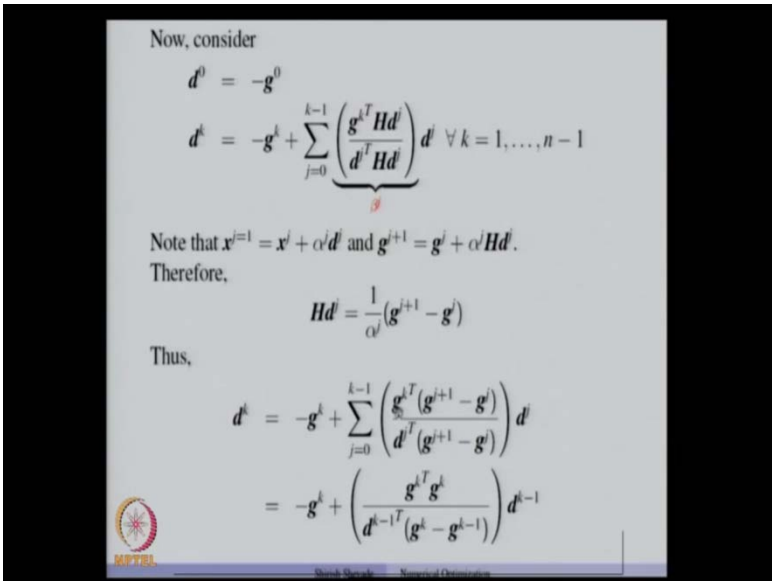
(Refer Slide Time: 35:43)



So, let us see how to get that, so if we rewrite that formula d k equal to minus g k plus sigma beta j d j wherebeta j is actually replace by this. Now, first let us show that minus g 0 to minus g n minus 1, they from a linearly independent set of vectors. Now, note that, the span of d 0 to d k minus 1 is same as span of minus g 0 to minus g k minus 1, because the way d k is form is using minus g k and all the previous vectors. So, if all g j is are independent, up to k minus 1 then span of d 0 to d k minus 1 is same as span of minusg g 0 to minus g k minus 1.

And we have already shown the d 0 to d k minus 1 are H conjugate that is g k perpendicular to, and if they are H conjugate then g k is perpendicular B k, this is what we have already shown. So, the new direction that we get is perpendicular to the space spanned by d 0 to d k minus 1, and that essentially means that g k will be perpendicular to the space spanned by minus g 0 to minus g k minus 1. So, which means that, g k is independent of minus g 0 to g k minus 1, so therefore minus g 0 to minus g n minus 1 forms a linearly independent set of a vectors.

(Refer Slide Time: 37:30)



Now, we have got d k to be minus g k plus sigma beta j d j where beta j is this quantity, and we know that x j plus 1 it should be j plus 1 equal to x j plus alpha j d j and g j plus 1 is equal to g j plus alpha j H d j. So, so let us see what happens to the x quantity in the numerator here, so H d j, H d j is nothing but H into x k, x j plus 1 minus x j by alpha j. And that is nothing but this quantity, and therefore, what we get is d k to be minus g k plus sigma j going from 0 to k minus 1 into this quantity.

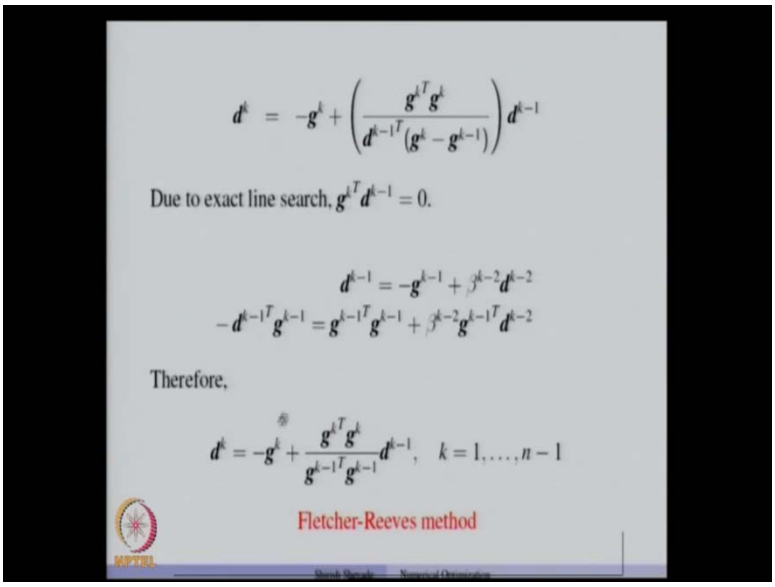
So, we have g j plus 1 equal to g j plus alpha j into H d j, and we use this quantity to write this and this H d j in the numerator is now replaced by both in the numerator and denominator is replace by this quantity. So, this alpha j k get canceled and what we get is g j plus 1 minus g j transpose g k divided by g j plus 1 minus g j transpose d j. Now, let us look at this quantity, so you will see that if you do the exact line search, then g k

transpose g j j goes from 0 to k minus 1 will be 0. And what we are what we are left with this quantity g k transpose g k by d k minus 1 transpose g k minus 1 g k minus 1.

So, the important quantity, that that is left here is this quantity related to only j equal to k minus 1, the rest of the quantities become 0, because g k is perpendicular to all g j is that we have already shown. So, g k is perpendicular to all g j, j going from 0 to k minus 1, so this quantity vanishes and because of the line search, d j transpose g j j going from 0 to k minus 1 becomes 0. And we are left with only quantity corresponding to j equal to k minus 1.

So, you will see that, as compare to the usual gramschmidt procedure which used all the previous directions and the minus g k direction to find d k, here we use only minus g k and only the previous direction d k minus 1. So, because of the H conjugacy of this vectors. formula becomes very simple, so we do not have to worry about all the previous directions to find out d k v. Or it is sufficient for us to determine d k using minus g k and d k minus 1, and this is possible because g k is perpendicular to all the previous g j's and we do the exact line search.

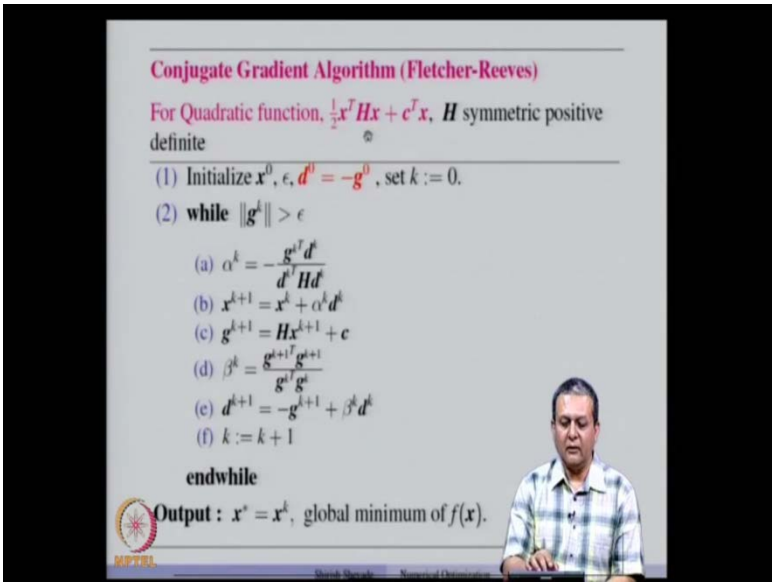
(Refer Slide Time: 41:12)



So, due to exact line search, g k transpose d k minus 1 is equal to 0 and, and if we expand this d k minus 1 transpose g k minus 1, then what we get is this quantity becomes 0 and only this quantity remains. And therefore, what we get is formula for finding the conjugate directions and which is call the Fletcher-Reeves formula or Fletcher-Reeves

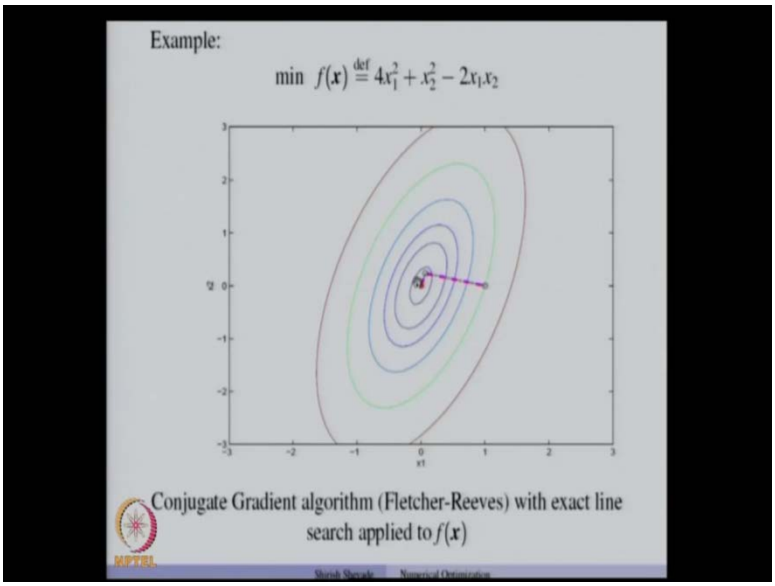
method. So, the direction d k is found using minus g k plus g k transpose g k by g k minus 1 transpose g k minus 1 into d k minus 1, so it is a combination of minus g k and d k minus 1. Now, here we have assume that d 0 to be minus g 0, so we can get the all the n conjugate directions by making use of the gradient at the respective points, and then the previous direction.

(Refer Slide Time: 42:15)

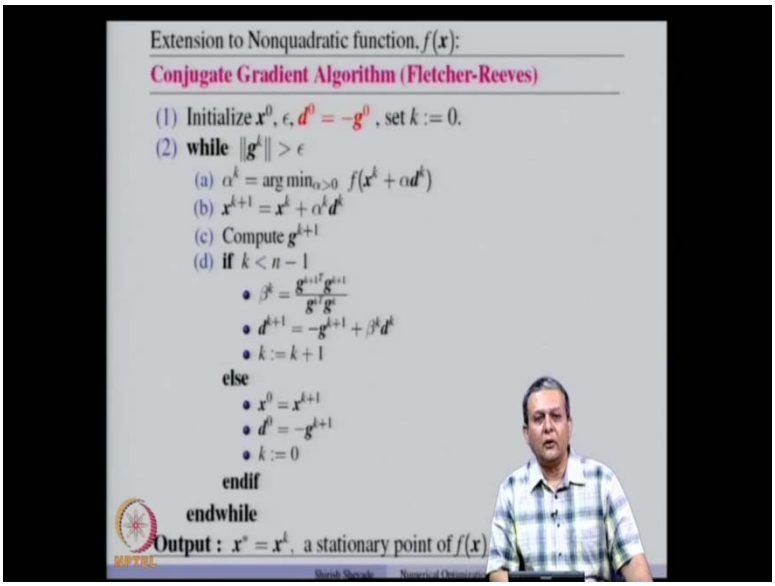


So, here is the Conjugate Gradient Algorithm, so we start with d 0 to be minus d 0, initialize x 0 epsilon. And remember that initially we, we have given the algorithm only for the quadratic function half x transpose H x plus c transpose x, where H is a symmetric positive definite matrix. So, the first step is to find alpha k we have a close from expression for this if if we do the exact line search then x k plus 1 is equal to x k plus alpha k d k and g k plus 1 is H x k plus 1 plus c, that is the gradient. We determine beta k and then d k plus 1 the new direction is determine minus g k pus 1 plus beta k d k, and the k is incremented by 1. And we check whether the the gradient of the function, and the new point is less than or equal to epsilon less than or equal to epsilon we stop. And we get a global minimum of f, because let this a global minimum H is a symmetric positive definite matrix for this quadratic function.

(Refer Slide Time: 43:31)



(Refer Slide Time: 43:58)

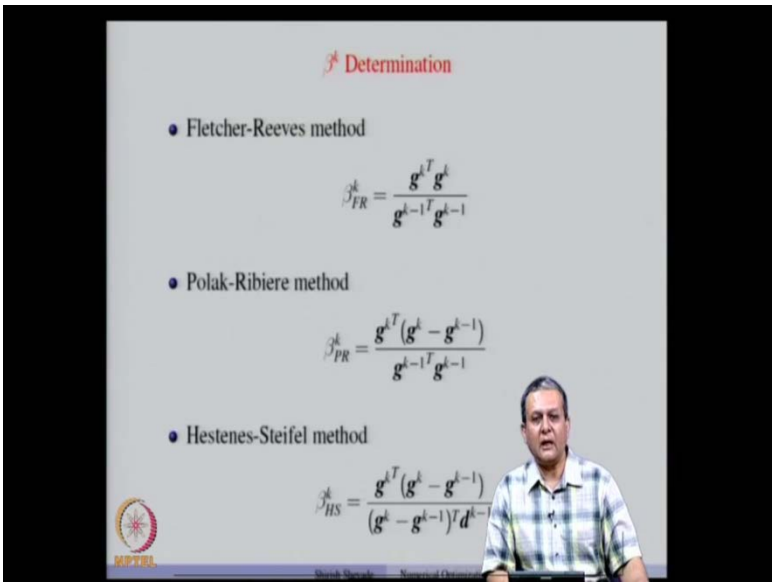


Now, when we apply this to, the function that we had seen earlier. So, the use of conjugate gradient method to this function results in this steps so we start from here, so this is x 0 x 1 and x star so within 2 steps we reach the solution. Now, if you start from a different point, initial point even then you will see that we reach the solution in 2 steps. Now, how do we extend this formula to non-quadratic functions, so this initial step is still the same. But then after every any iterations there is a need to restart the method. And the reason for this is that, when we talk about non-quadratic function we have to do this line search, and sometimes that line search also may not be exact.

And more over this directions that we get in the conjugate gradient method, where we use the Fletcher-Reeves formula ,these directions may not be descent. Now, to make sure that in the n retractions, at least 1 direction is negative or at least 1 direction is a descent direction there is a need to restart the conjugate gradient method for non-quartaratic functions after every n iterations, and this is the restart. That the new point is again set as x 0 d 0 is set as the negative of the gradient at that point so which will make sure, that we have the descent direction at least one descent direction in n iterations remember start with d 0 equal to minus g 0.

So, at least 1 direction in the n iterations will be a descent direction, and the procedure is repeated till norm of g k becomes less than equal to epsilon. So, this is the difference between the application of Conjugate Gradient method to Quadratic functions and non-quartaratic functions. For quarararic functions, we reach the solution in the at most n iterations and we did not have to worry about the restart there. And finally, what we get here is a stationary point of f of x, Now, if we look at the algorithm we use Fletcher-Reeves method to find beta k. Now, there exist different ways to determine beta k and the different Conjugate Gradient Algorithms vary with respect to this calculation of beta k.

(Refer Slide Time: 46:23)

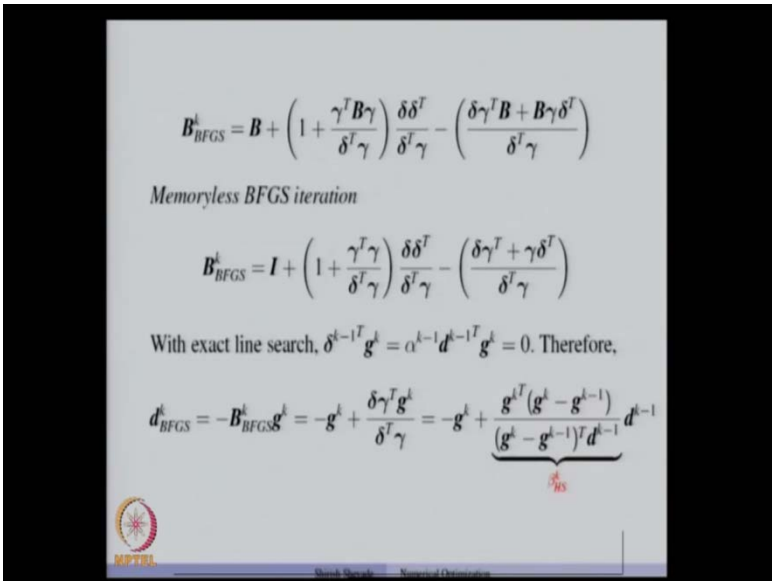


So, so the formula that we have seen so far is in the Fletcher-Reeves formula, which uses beta k to be g k transpose g k by g k 1 g k minus 1 transpose g k minus 1. This is another

formula for Polak-Ribiere formula, and that uses g k transpose into g k minus g k minus 1 by g k minus 1 transpose g k minus 1.

Now, the only difference between these 2 formulas is that, there is a extra quantity here. Now, if we use quadratic functions with exact line search, then we know that g k is perpendicular to all the previous gradients. So this quantity is 0, so the 2 formulas are same, but for non-quadratic function it was empirically observe that, this formula the Polak-Rebiere formula performs better compare to the Fletcher-Reeves formula. So, for Quadratic functions with exact line search these 2 formulas will give us the same steps, there is another formula got Hestenes-Steifel formula, and that is given here now the motivation for this formula was different, but we can look it from the B F G S method view point for Quasi-Newton directions that we saw earlier.

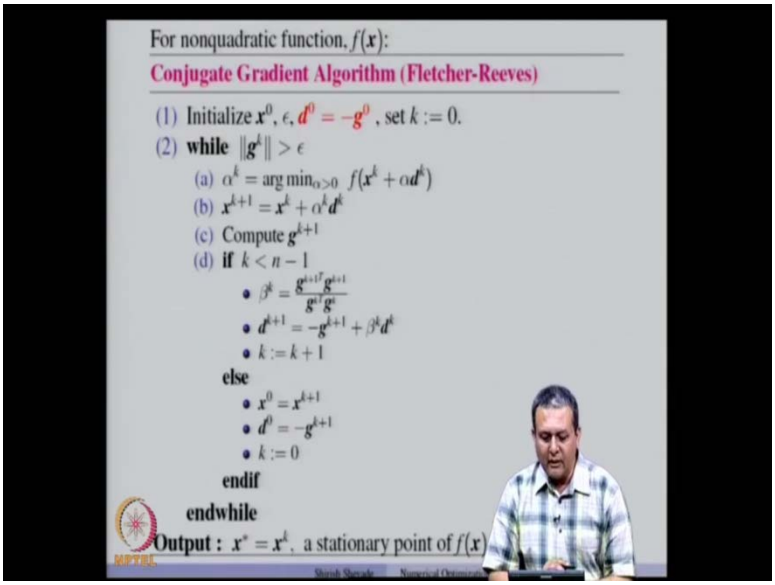
(Refer Slide Time: 47:50)



So, recall that the update formula for B F G S method is given here, and there is a variant of this B F G S method which is call the Memoryless B F G S method. And in the memoryless variant this B is replace by identity matrix, so which means that at every x k we do not have to use B k. But use in this case this will be B k minus 1, so we do not have to use B k minus 1, but instead replace B k minus 1 by identity matrix. So, if we do that, then what we get is called a memoryless Quasi-Newton direction. And if we do the exact line search, the delta k minus 1 transpose g k equal to 0, so remember that this delta is a function of x k and x k minus 1 and g k is a function of g k.

So, g k is the gradient and therefore, so if we use this update, the Quasi-Netwon update B F G S which uses the Memoryless iteration. So, this quantity will be nothing, but the quantity which is given here, and syou will see that this formula this update of d k will be same as minus g k plus some beta k into d k minus 1. And that beta k is nothing but the formula given by Hestenes-Steifel, so this B F G S method. The D F P method these can be thought of other way to generate conjugate directions.

(Refer Slide Time: 49:40)



Now, now let us look at the the difference between the D F P and the B F G S method. D F P or the Quasi-Netwon method and the Conjugate Gradient method. Now, this is a algorithm for minimizing ageneral non-quadratic function f of x using Conjugate Gradient method. Now, this does not require any second order information, so or it does not use any matrix operations. So, the main difference between the Conjugate Gradient method and Quasi-Newton method, is that there are no matrix operations involved. Secondly the Conjugate Gradient for the Quasi-Netwon method like the D F P method or the Quasi-Netwon, the L B F G S method both of them in addition to generating conjugate generations.

They also update the matrix B k, and whenever we are near the solution that B k will be a good approximation of the Hessian Inverse, and therefore, the directions that we obtained using D F P or B F G S methods near the solution will be a close to the Netwon directions, and then the convergence will be faster. So, in addition to maintaining

conjugate directions strategy, they also find the directions which are more close to the netwon directions near the solution. And therefore, they are fast and that does not happen in the Conjugate Gradient method.

Conjugated Gradient method, since it does not work with any matrices there is no guaranty for general non-quadratic function with in exact line search, that the direction that we get is a descent direction. And therefore, there is a need to restart, while that is not the case, this restarting is not necessary for the Quasi-Netwon methods like D F P or B F G S. So that is a important point and more over that Conjugate Direction methods or Conjugate Gradient methods they are very sensitive to the line search.

While Quasi-Netwon methods are more robust, now on the other hand the Quasi-Newton methods require, the multiplication of a matrix by a vector. And that require order n square computations as well as order n square storage, because the matrix the enter matrix needs to be store. If we do not use the limited memory B F G S kind of updates. Now, that is not the case here, if you look at the storage required here will, will require order n storage we need to store some n-dimensional vectors in the Conjugate Gradient method. So, the storage wise Conjugate Gradient method has the advantage as compare to the Quasi-Netwon method.

But, both these methods Quasi-Netwon method as well as the Conjugate Gradient method, they have better computational complexity as compare to Netwon method. Because Netwon method requires inversion of a matrix to get a Netwon direction and that is computationally expensive operation. Well, here neither conjugat gradient method nor quasi-netwon method require any inversion of a matrix, further Conjugated Gradient methods do not even require to store any matrix.

So, storage wise these methods clearly have a advantage, so if one wants to get a robust method one has to for Quasi-Netwon methods. Because they were found to be more robust compare to conjugate gradient methods and they are less sensitive to the line search, and of the 2 methods D F P and B F G S, B F G S method of finding Quasi-Netwon direction, what is found to be empirically superior to D F P method. So, based on the need, one has to decide the algorithm that can be applied to minimize given function. So, for quadratic functions methods like conjugate gradient method are quite

good, for general non-quadratic case one depending upon the requirement, one can go for the Quasi-Netwon kind of update especially the B F G S kind of update.

So, this completes our discussion on unconstrained and optimization. So, we saw Hestenes-Steifel method, then the Newton method, then the variants of Newton method the Quasi-Newton method and conjugate gradient method, Now, all these methods are useful in different ways to solve given optimization or a given minimization problem. And in the next class, we will study the methods for constrained optimization, and find the properties related to those methods.

Thank you.