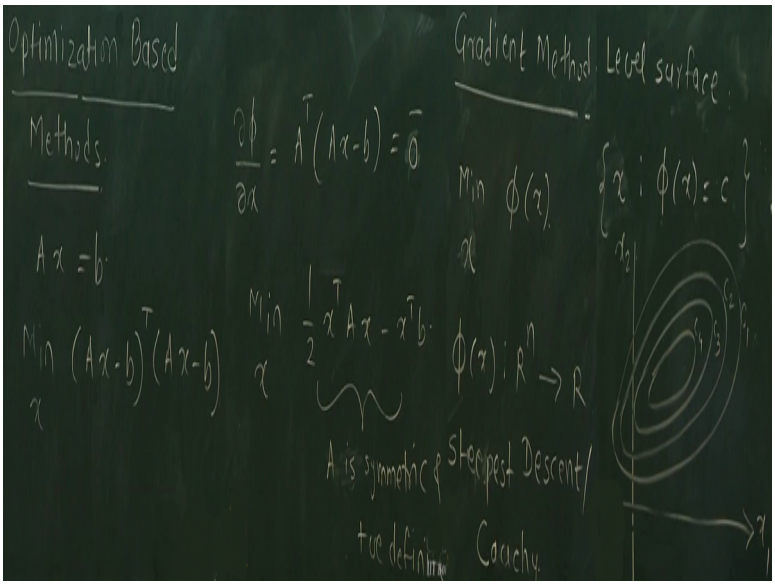
Advanced Numerical Analysis
Prof. Sachin Patwardhan
Department of Chemical Engineering
Indian Institute of Technology - Bombay

Lecture - 32
Optimization Based Methods for Solving Linear Algebraic Equations: Gradient Method

So we have been looking at iterative methods for solving linear algebraic equations, and we have looked at Gauss Seidel Jacobi relaxation methods and its variants, we also have looked at the convergence behaviour, we analyses convergence behaviour and we know how to ensure convergence by modifying the problem and so on. Now there is one more iterative method which is quite popular and which also converges pretty fast. So this is numerical optimization based method.

(Refer Slide Time: 01:05)



So well one of the reasons for covering this is that this will be also useful when we go forward to non-linear algebraic equations, so as I said we want to solve this problem A x=b, and then this can be solved by minimizing with respect to x, A x-b transpose A x-b. If I minimize this with respect to x, then I can reach the solution of A x =b, in fact I reached the solution if I take this as phi, then dou phi/dou x the necessary condition for optimality turns out to be A transpose A x-b=0.

And obviously if A is nonsingular if necessary condition is satisfied you will reach the optimum, the second derivative what is second derivative here, the second derivative will be A transpose A which is symmetric positive definite so you are actually reaching the global minimum okay. Yesterday, somebody had a doubt that iterative methods that we have looked at, will they give you local solutions or global solution?

Jacobi method, Gauss Seidel method, relaxation method, if those methods are converging if you know that they are converging they will converge to the global solution, for linear algebraic equations there is nothing like local and global solutions, if they are converging they will converge to the global solution. Now of course we can a little bit simplify this, if A is symmetric positive definite matrix in that case we can just minimize with respect to x 1/2 x transpose A x-x transpose b.

If A is symmetry positive definite okay A is a special case symmetric positive definite, then minimizing this objective function will give you the optimum. Now I want to do a general method called gradient based optimization method, this is now described in appendix D okay in my notes this is described in appendix D on page 48. I want to solve this using a numerical search, I do not want to use this condition directly I do not want to use this condition and solve it okay.

If I have to use this condition and then solve it for x, it would be either iterative method or it will be a direct method, I do not want to go into that, I want to I do not want to use Gauss Seidel method or anything, I want to use iterative scheme which is based on optimization techniques okay. Optimization techniques in general deal with, so I am going to do this gradient method this is also called as steepest descent method.

So right now, I am going to be worried about developing an iterative method for minimizing with respect to x some objective function phi x, where phi x is from R n to R, phi x is a scalar objective function some scalar objective function, it need not be norm, it need not be it is some objective function that you are defined okay, it need not be always positive I am not worried about that, I am just worried about scalar objective function, so it is R n to R okay.

I want to come up with the iterative scheme to reach a local minimum in this particular case, because in general phi x need not be nicely behaved okay, and then after I derived that I want to apply it to this specific case okay. So I want to it is the purpose is twofold, one is to introduce to you gradient based methods okay and its variants, which are very useful in optimization, and I will show you what are the applications later.

So numerical search which is based on gradient, and then we will of course apply to our specific problem that is solving linear algebraic equations okay. So this method is also known as steepest descent, you may have done this in your undergraduate I am not too sure, the steepest descent it is also called Cauchy method, it is just known by very various names, gradient based method. The basic idea is that if I looked at a level surface, what is the level surface?

Level surface is a set of points x is set of point all point x such that phi x=constant, it is a scalar objective function right the scalar objective function, so phi x=constant, I want to look at level surfaces that is I want to look at locus of x, let us say if it is 2-dimensional object if it is x is a vector which is in 2-dimensions x1, x2 okay. I am actually looking at okay so this is say C1, this is C2, this is C3, this is C4 and so on, so this is my x1, x2 plane.

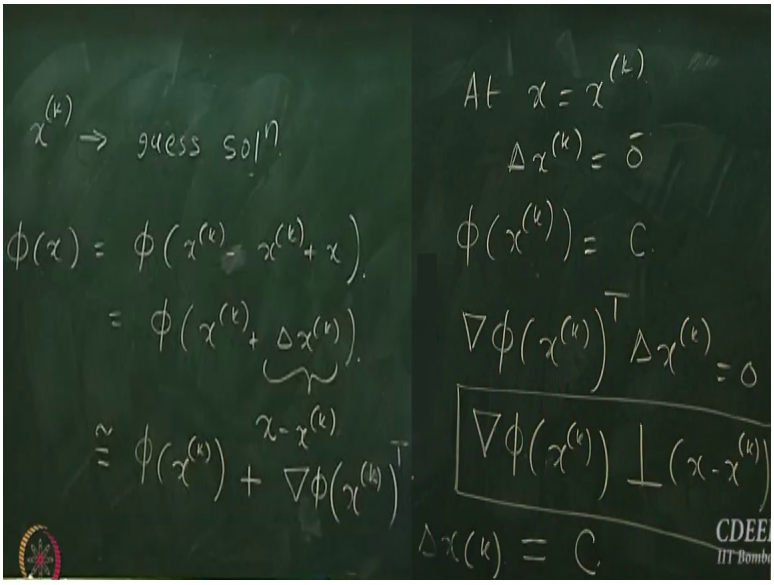
I am plotting all those points in x1, x2 plane for which phi of x1, x2 =constant, so let us say this is 5, this is 4, this is 3, this is 2, I am plotting all the points locus of all the points these are called as level surfaces okay. I am not plotting phi of x in this plot okay, I am plotting. So actually if you do a 3-dimensional plot x1, x2 and phi okay, this will be nothing but the cross-sectional plane projected onto x1, x2, it is set of all points.

See if have you seen mat lab symbol, mat lab symbol is like one speak right, now if you take it as a objective function okay, let us say height above the or you take mountain, height of the mountain above the ground surface is the objective function okay. I am trying to find out set of all points where the height is constant okay, how will you get it? Take a plane horizontal to x y project it onto x and y, you will get the set of all points so these are a set of all points.

What is phi x? View phi x as a height okay, and x1, x2 as ground locations. If you take constant level it is also called level surfaces, probably the reason for level surface is relate it to level okay, they are called as level surfaces okay. Now I am going to use the local behaviour of this level surfaces to come up with iteration scheme for solving this minimization problem, for the time being I am going to forget about solving linear algebra equations.

I am just concentrating on this general problem some phi of x, it need not be this phi of x, any phi of x okay, not as specific one.

(Refer Slide Time: 10:13)



So what I am going to do now is let us say I have some guess solution x k is some guess solution, x k may not unlikely to minimize, but I say it is my guess. What is the philosophy in iterative methods? You start with the guess and then you move onto next guess right, start with one guess move onto the next guess, and then hope that iteration converge okay to the solution. In this case what it will converge to, will be a local solution.

Well, in some cases it will converge to the global solution but that depends, it depends upon the problem a, it depends upon your initial guess, if the problem is highly non-linear with funny shapes, it depends upon the problem is one which has only one peak or one valley, well you know it will reach the global minimum okay. So x k is my guess solution, well our good old friend is Taylor series theorem.

And I am going to use Taylor series theorem to phi x, I am going to write as phi x k+ okay, which is same as phi x k+ delta x k, where delta x k is obviously x-x k, so this is my x-x k this is delta x k okay. If I do Taylor series approximation in the neighborhood of x k okay, so this is approximately =, this phi of x is approximately =phi at x k+ grad phi so let us develop a notation or let us put this grad phi x k.

So gradient of phi evaluated at x k that is what I mean, so this transpose delta x k okay, and there will be higher order terms I am neglecting higher order terms, I am looking locally in the neighborhood of x k, how this function behaves okay how does this function behave in the neighborhood of x k? And then I want to look at the level surface that is phi x=constant okay, in a small neighborhood x k some point x k, I get this approximation of phi x as this okay.

What happens at x=x k, delta x k=0 okay, so which means at x=x k if I am looking at a level surface okay that means phi x k=constant at that point. See suppose let us go back to here let us say this is your x k, this is my x k okay I am trying to model this curve locally okay, you will see that actually I will model it using the tangent okay, I will model it using the tangent plane that will become clear now soon. So what is the simplest approximation? this curve is there.

What is the simplest approximation you can construct? Straight line locally for a small neighborhood you can construct a straight line approximation to the curve that is what I am doing, how do I get the slope of the straight line? Through Taylor series I am getting that so the local slope of this line through Taylor series okay, so Taylor series is my vehicle to construct the local approximation. So now this phi x k is constant, if I substitute here okay what will I get?

See this becomes C so C=C, so what is the local behaviour of the curve? So this implies that gradient of phi at x k transpose delta x k=0, is everyone with me on this? This is a scalar function by the way, this is a vector gradient is a vector okay, this is also vector delta x k is also a vector okay, so this transpose this is 0, geometrically what does it mean? The gradient is perpendicular to delta x k delta x-x k is perpendicular locally to the gradient okay.

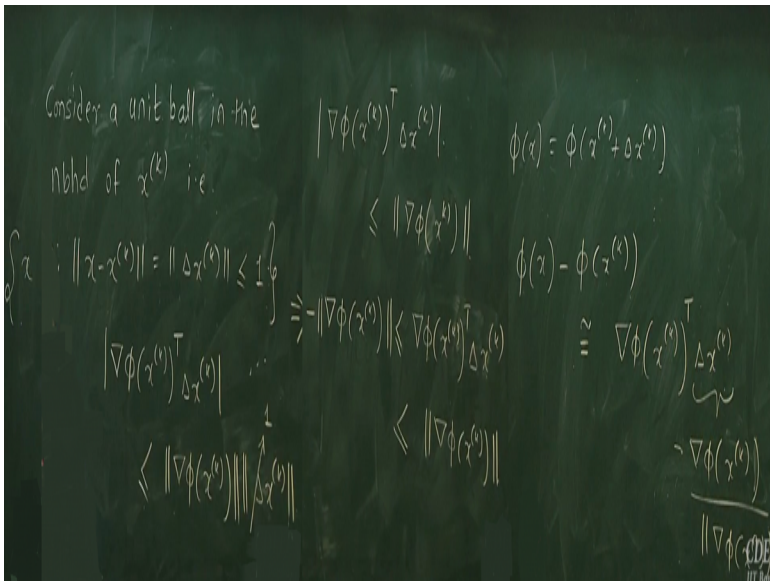
So locally gradient of phi is orthogonal to x-x k, this is what we have found out, actually this gradient transpose delta x k=0 okay is the equation of the tangent plane to the level surface okay, in general I am talking n-dimensions, it is a tangent hyperplane in the n-dimensional space okay. So well what I want to show here is that this local behaviour of the function in the neighborhood of the point x k can be used to find out the direction in which function decreases at the maximum rate.

See if I want to if I am at x k let us go back here, what I am doing? I am minimizing phi okay, so if I want to move from x k to x k+1, which direction I should move? I should move in that direction in which function decreases at a maximum rate. **"Professor - student conversation starts"** why? Question is why is it, (()) (18:21) what is directional derivative? So I want to prove it, angle will be.

So which is the directional derivative here delta x k is the directional derivative or gradient is directional gradient is a directional derivative. **"Professor - student conversation ends."** So I want to show that if delta x k is aligned along the directional derivative that is gradient, then function increases maximum okay, if it is aligned along negative of the gradient direction then the function decreases maximum okay.

So this local gradient actually gives me maximum rate of increase, and negative of that gives me maximum rate of decrease, and I am going to use this local gradient to come up with the new point x k+1 okay. So before I do that I have to show that this is the maximum the direction of maximum decent. First interpretation that we have learnt here is that, this is nothing but equation of the tangent hyperplane, and delta x k is perpendicular to the gradient locally okay.

(Refer Slide Time: 19:48)



Now so I am looking at set of all x, I am looking at a unit ball in the neighborhood of x k okay, I am constructing a small unit ball in the neighborhood of x k okay such that it is set of all points such that magnitude is unity of delta x k okay, so just if you go back to this figure I am constructing a small unit ball here such that you pick up any point okay its distance from x k is <1, is this clear? I am just picking up a set to do the analysis okay.

Now what is going to help me here is something that you probably can guess, what is going to help me here is Cauchy Schwarz inequality okay, this is the inner product of this vector with this vector okay, which is <=by Cauchy Schwarz inequality, what is this? This is norm right. But then I am looking at set of all x in the unit ball okay, so this is 1 maximum value this can take is 1, so which means okay, so if this is 1 so maximum value this can take is 1.

Then I can write that grad phi x k transport delta x k this quantity okay is strictly < norm of this right, this inequality also means that -of is <, I have just expanded this inequality here I had written absolute value. So in a unit ball in the neighborhood of x k, I can say that this quantity is bounded between these 2 numbers, this is a positive number, this is a negative number, this quantity cannot be smaller than this. What is the smallest value this quantity can take?

When will it take this value? When delta x is aligned along which direction gradient direction, when delta x is aligned along the gradient direction, this inequality will be equality smallest

change. Now why I am worried about this okay, let us go back and look here let retain this figure let us go back here. See this phi x which is written as phi x k+ delta x k right, I have written this like this, and actually I am worried about how this function behaves phi x-phi x k.

I want to go to x from x k, I want to go to a new value x from x k okay, this is we say that in small neighborhood this is approximately =gradient of phi okay gradient of phi is given by this okay. So this the behaviour of this quantity actually dictates how locally the, how this function behaves locally, is it clear? This is Taylor series expansion, I just wrote this sometime back okay, I am just rearranging this thing on the right hand side I have taken on the left hand side okay.

See if I move away from x k to some new x okay, if I move away from x k to new x, which direction I should move? If I want to decrease the function which direction, I should move? I should move negative of the gradient direction okay, because what is the smallest value this can take? Using see I am restricting myself to a unit ball around x k, I want to move inside this unit ball, I just want to know where to inside this unit ball.

What is the objective? I want to move in such a way that the function decreases at the maximum rate okay, now I know that from this Cauchy inequality I know that the maximum rate of decrease will be obtained, when delta x k is aligned along the gradient direction, but not along negative of the gradient direction, then I will get this - here okay, I will get - here. This Cauchy inequality when do you get, what is Cauchy inequality can you tell?

We relate Cauchy inequality to cos theta angle okay, so I am talking about 2 special angles, one is angle is 0, other is angle 180 okay, negative and positive directions. If you are maximizing the function you should move along the positive of the gradient direction, if you are minimizing the function you should move along the negative of the gradient direction, because this difference will be smallest negative, when will it be smallest negative?

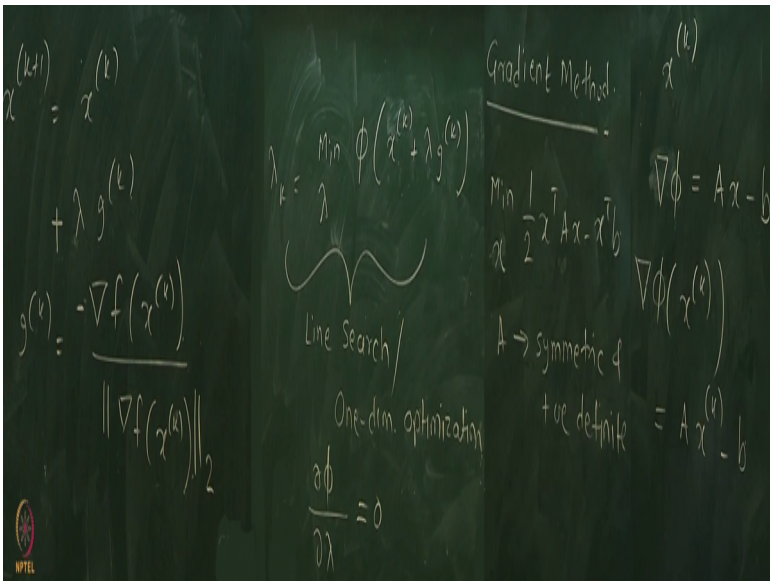
Look here, when will it be smallest negative? Negative of the gradient direction okay, so if I move along the negative of the gradient direction okay, I will decrease the function okay. So way I should choose my next point okay from x k when I go to x k+1, I should choose my next point

okay by moving along the negative of the gradient direction, since I am minimizing the function okay my objective was to minimize phi of x with respect to x k with respect to x okay.

Locally, what I find is that locally the function will decrease maximum if I move along negative of the gradient direction, see what is negative of the gradient direction? If this is grad phi x k/norm right okay, and negative of this, why I am dividing by this because I am looking at unit vectors, so this is a unit vector okay. What will be this transpose this square of the, what will be this transpose this? Inner product, inner product is square of the norm, inner product of vector with itself square of the norm right.

So if you take inner product of this with this you will get square of this divided by this, you will get negative –of, is everyone with me on this, is this clear? You move in the negative of the gradient direction the function well locally decrease at the maximum rate okay. So that is going to be my algorithm for.

(Refer Slide Time: 29:02)



So to find x k+1, I am going to take x k and - negative of the gradient direction, so lambda g k, where g k is nothing but grad f x k/norm okay, is this fine? This is the direction okay or we can put + here and take this - does not matter whichever way you want to look, negative of the gradient direction, I am taking unit direction along the gradient okay, I am well of course I am looking at 2 norm, I am not really right now worried about other norms.

So these are all 2 norms, wherever I am writing norms these are 2 norms. So this is my negative of the gradient direction, and what is this lambda now okay. Now I know that locally if I go along negative of the gradient direction function is decreasing, how much do I move? See I just know that this direction is steepest descent okay, I should move 1 meters, 5 meters, 10 meters, how much should I move okay.

So I am going to put 1 unknown here which is step length, this is my step length, and this is my direction okay. Now how much to move? I am going to do another optimization problem okay, having decided to move in this direction, I am going to now solve for this problem, lambda k is minimization with respect to lambda phi of okay. What is the difference between the original problem and this minimization problem? This is a one-dimensional minimization problem.

Lambda is a scalar, lambda is a step length okay, the direction is fixed, how much to move is given by the step length parameter okay. Now how to solve this problem in some cases this problem can be solved analytically, in some cases this problem can be solved has to be solved numerically okay. Now if you just go back this is called as line search one-dimensional optimization problem, this is called as line search because we know in which direction to move.

We just want to find out how much to move okay, so this phi becomes this x k is known, g k is known, lambda is unknown with respect to 1 scalar I have to find out, of course what I have to do is to solve for dou phi/dou lambda=0, whichever value gives me minimum sorry, whichever value satisfies this optimality condition, I choose that value and use it for my step length. This has to be done in some cases if phi is a highly complex nonlinear function.

This has to be done using non-linear optimization or using iterative process, you guess and then find out the minimum, I have described that but now in the case of solving A x =b we have some nice time we can do this analytically okay. So let me go back and, is this clear, is the ideal clear? The line of argument is like this locally the steepest or the direction in which objective function decreases maximum is negative of the gradient okay.

You do not know how much to move, so you know the direction to move but you do not know how much to move that is quantified by this lambda okay, and then we have obtained lambda by one-dimensional minimization with respect to lambda okay. I am just going to, **"Professor - student conversation starts"** (()) (33:35) maximum value of, so I want to see I want to find out see I am decreasing phi right okay. So now in one shot I would like to decrease when I am taking one step.

I would like to decrease as much as possible, so how do you find out how much is possible, see just imagine that you are going down the slope okay, now let us say the slope is like this and then it flattens out okay, now locally if you go down for 1 meter your height will decrease, but your height might degrees even if you go 5 meters know, so how do you know how much to go, I know that this is local decent, but should I go 1 meters or 3 meters or 5 meters or 9 meters, 9 meters might take me up I do not know.

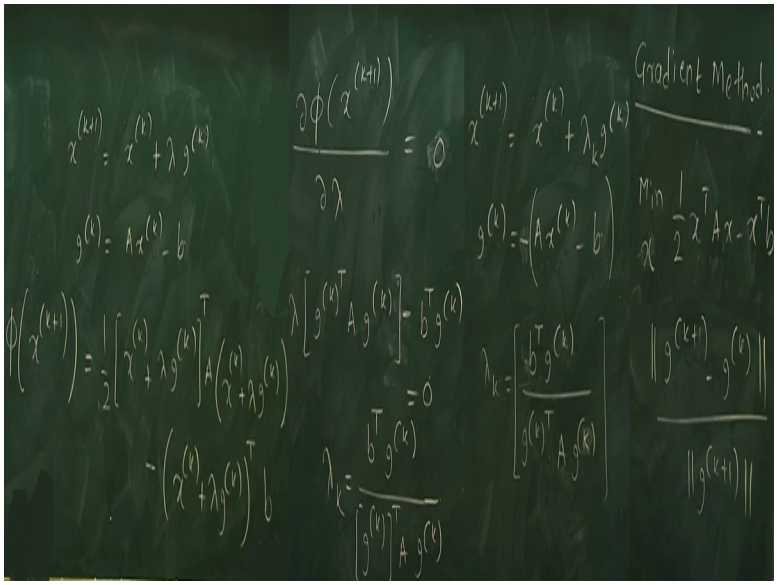
See the contour could be like this and then going up, so I should find out what is best possible step length okay, I should go so that there is a minimization otherwise, see all this just remember one thing you are trying to do a local moment only based on the local derivative, there is limited information one derivative of a function carries okay, so you cannot take too large steps using just local gradient information okay, and then you should not take too small step also right.

So to balance that we actually introduce this lambda, and then we minimize functions with respect to lambda again, and then find out how much to move okay. **"Professor - student conversation ends."** Now let us see this application in solving A x =b okay, so my phi x here is, now I am going to formulate just for the sake of writing simplicity, I am going to say that this is 1/2 x transpose A x-x transpose b okay.

And I am going to solve for the case where A is symmetry positive definite, if your matrix A is not symmetry positive definite what to do you know already, pre-multiply both the sides by A transpose, so you do not have to. So I am just going to look at the case right now for deriving the algorithm for the sake of simplicity of notation, I am going to look at the case where A is symmetric and positive definite okay.

Now let us apply the algorithm this is my phi okay, I have a guess solution my guess solution is x k, what is the local gradient? That is what is grad phi=A x-b, differentiate this with respect this is a vector transpose A x symmetric positive definite vector, differentiate this with respect to x, differentiate with respect to x, derivative of this objective function with respect to x will give you A x-b. What is phi x k? Evaluated at x k, x k is your guess solution okay, A x k-b, everyone with me on this? Okay.

(Refer Slide Time: 38:07)



So what I want to do next, well I do not have to always find unit direction, I wrote the algorithm with unit directions, I can write it with respect to the direction and use lambda, lambda will get scaled accordingly okay. So I can say that I want to move now in the direction which is lambda g k, where g k=A x k-b okay, now you want to do the step length minimization, can you do the step length minimization? Can you solve it? Just write.

What is the step length minimization problem now? What will be phi x? What will be phi x k+1? It will be 1/2+x k+ lambda g k transpose A x k+ lambda g k right -, what are the things which are known here? I know x k, I know g k because g k is function of x k I know g k, I do not know only lambda okay, can you tell me what will be this quantity dou phi x k+1/dou lambda, I want to set this =0, what is this quantity? Just find out, well there is one small problem here.

I want to move in the negative of the gradient direction, so this is make one correction, I want to move in the negative of the gradient direction, the gradient direction is this, negative of the gradient direction is okay, so this is the gradient direction, and my g k direction in which I want to move is negative of the gradient direction, so put a – here. Well, what you have to do a course is expand this, what you will realize is that it terms x k transpose A x k will vanish.

Because they are not functions of lambda, you have to only take those terms in which lambda will appear, there will be crossed terms and there will be lambda square will come out, because lambda square g k transpose A g k okay, here again you can neglect the term x k transpose b, because it is not a function of lambda you can take only this term okay. What you get after you minimize just expand just try, what is this quantity? You do not have to substitute this.

You maintain everything in terms of g k okay, maintaining everything in terms of g k, try to find out what is which value of lambda will give you, what I expect is if you do this scalar optimization problem you should get an equation just check this, you get an equation of the type lambda*g k transpose A g k-b transpose g k=0, you will get an equation of this type just check. If expand this, when you expand this you will get only one variable polynomial lambda square, lambda and the constant.

You will get only one variable polynomial because lambda is a scalar, g is known vector, x is a known vector okay. So actually it turns out that lambda k which minimizes this is nothing but b transpose g k/g k transpose A g k okay. So my algorithm my numerical algorithm becomes, how do you summarize the numerical algorithm? Okay this is my numerical algorithms, how do I go from x k to x k+1? I first compute negative of the gradient direction.

See what is the simplicity here? No matrix inversion is involved okay, I just have to compute the gradient direction gradient direction is nothing but actually error between right hand side and left hand side, this is my guess solution, this is my b. Actually I want this, when will you get the solution? Gradient becomes=0, what is the meaning of gradient becoming=0? You have reached the solution very, very straight forward simple interpretation in this case.

If gradient becomes=0 this is the necessary condition for optimality right, when if the gradient is non 0 okay, you will keep moving how much to move? Lambda k times g k okay, this is the optimum step length, if you move less than this okay, then you are not decreasing the function enough, if you do more than this that will not help okay, using the local gradient you can move only this much okay. This is the optimum value to which you should move every time.

This is the scalar calculation, this is an inner product calculation, A symmetric positive definite is inner product calculation okay, calculating this scalar is very, very easy, calculating this error very, very easy. When will you terminate iterations? When g k is very, very small right, so I could terminate the equations by saying norm g k is <some epsilon, norm g k is very, very small. Or you good also, sometimes it is better to check whether you can put this also g k+1-g k.

This can be time termination criteria if there is no significant change in the derivative okay, if you have very large matrices this is very, very useful. This method can quickly come to the solutions particularly if A is symmetric positive definite, then you can reach the solution, I think there is a specific result about this we will talk about it later. There is a modification of this called as conjugate gradient method.

And we will talk about the conjugate gradient method to very quickly in the next lecture. And then I will move onto well-conditioned and ill conditioned system. So this method actually is very often used for solving large scale problems, and computation involved are very, very simple, we just have to compute the gradient direction and inner products okay, and you can very quickly get approximate solutions of, or you can quickly go very close to the true solution using this method, okay.