Numerical Optimization

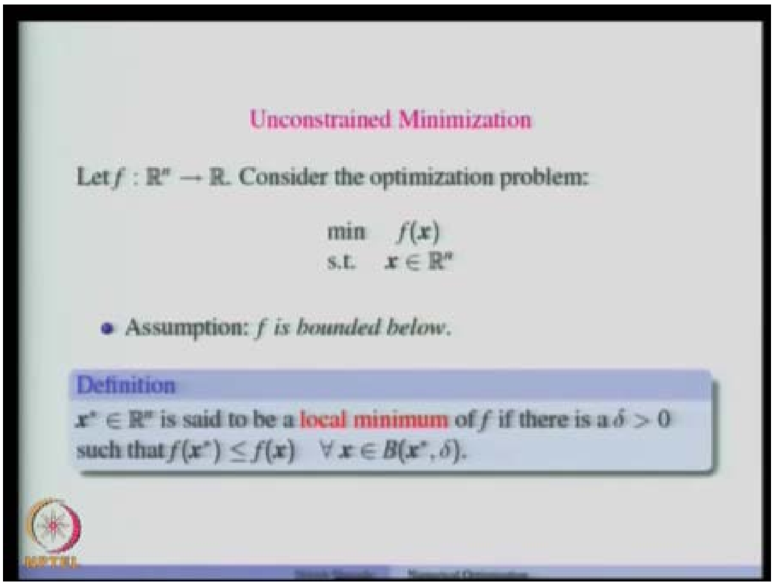
Prof. Shirish K. Shevade
Department of Computer Science and Automation
Indian Institute of Science, Bangalore

Lecture - 10

Multi Dimensional Optimization -
Optimality Conditions, Conceptual Algorithm

Hello, welcome back to the series of lectures on numerical optimization. New topic now which is the unconstrained optimization. So, in some of the earlier classes we started looking at unconstrained optimization, but that optimization was mainly related to one dimensional unconstrained optimization. Now, next few lectures will spend some time studying about n dimensional unconstrained optimization. So, some of the ideas that we discuss for one dimensional unconstrained optimization are very useful in studying this material. So, let us consider a unconstrained minimization problem.

(Refer Slide Time: 01:02)



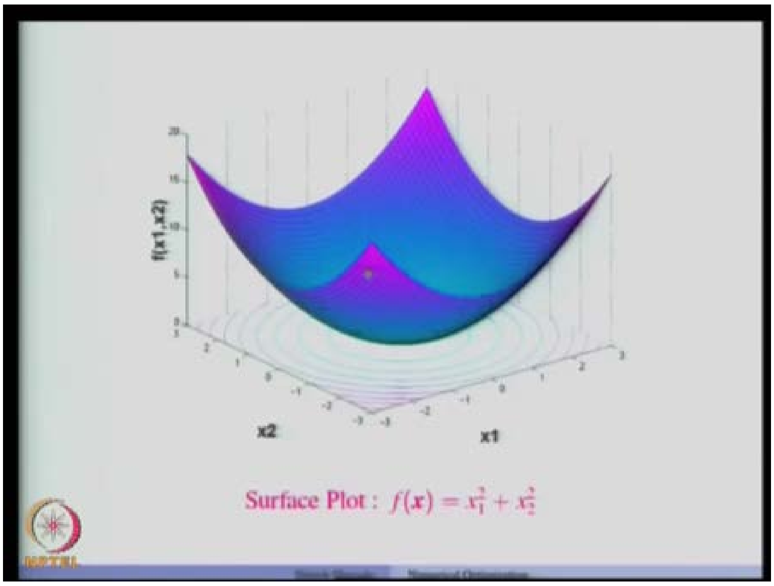
Let us define f m of R and consider the optimization problem minimize f x subject to x belongs to R n. So, this is the unconstrained optimization problem or minimization problem, because x takes the value from the space of numbers. Remember that the theory that we are going to study is equally applicable to the unconstrained maximization problems. So, I will not spend any time discussing about unconstrained maximization, because the ideas can the ideas discussed in well be extended to maximization problems.

Now, one of the assumptions that we make throughout this lectures in unconstrained minimization is that the function we bounded below. The function that we are trying to minimize is bounded below.

This is a reasonable assumption, because we have no interest in minimizing the function which is not bounded below, so will use this assumption throughout our discussion on unconstrained minimization. Now, recall that it will be difficult to find a global minimum of unconstrained minimization problem. So, in (( )) we will be interested in finding the local minimum. So, recall the definition of a local minimum of unconstrained minimization problem.

So, x star belong to R n is said to be a local minimum of f if there exists some delta greater than 0 such that f of x star is less than or equal to f of x in the delta neighborhood of x star. So, if you take a ball of radius delta around x star, open ball of radius delta around x star, then in that open ball the value of the function is at least f of x star. So, such a point is called a local minimum and it is this local minimum that we are interested in finding out.

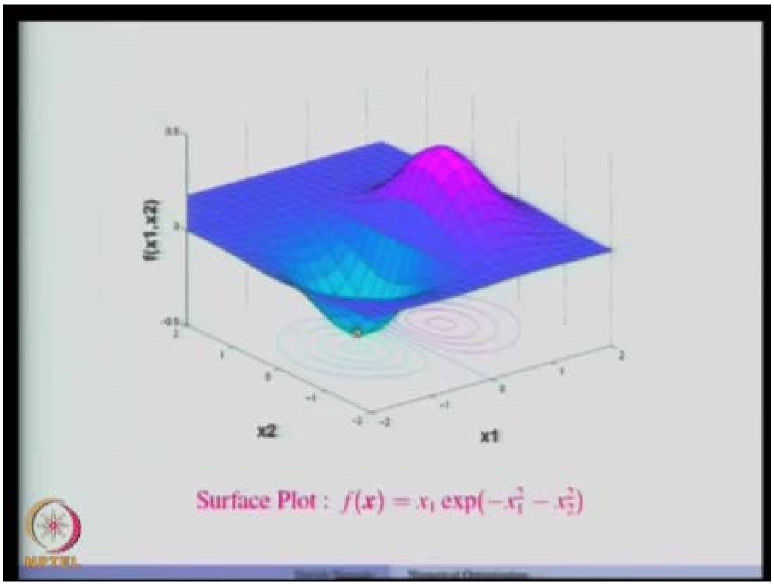
(Refer Slide Time: 03:11)



So, let us see some functions, so here is a function which is f of x is x square x 1 square plus x 2 square. This is these are the two variables x 1 and x 2 and this are this plot of the function, which we have seen earlier also. Now, you will see that the minimum of this function occurs somewhere here. So, this is a cock shaped object and minimum occur.

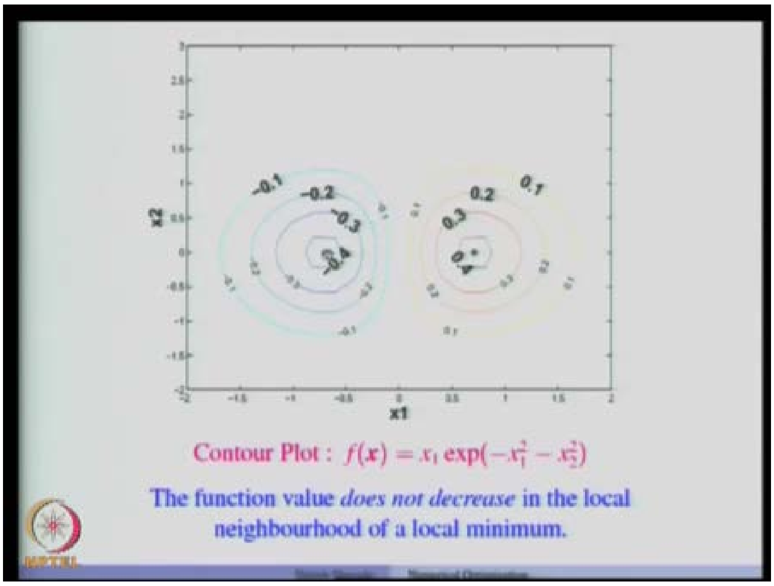
Now, surface plot we will see what are called the contour plots of this function. So, we are mainly interested in looking at the contour plots.

(Refer Slide Time: 03:59)



Let us take another example; so let us look at the surface plot of a function f x equal to x 1 into e to the power minus x 1 square minus x 2 square. This function has a local minimum at this point and a local maximum at this point. Now, beneath the surface plot of the contours of back functions on…

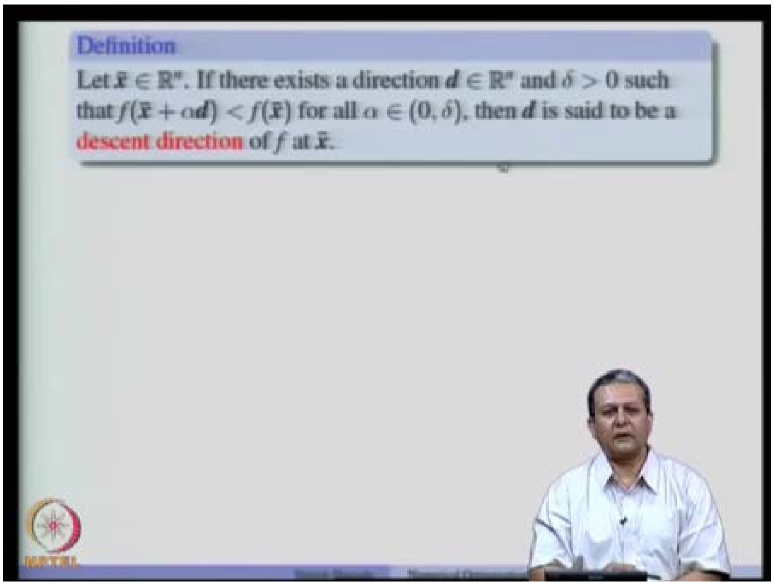
(Refer Slide Time: 04:33)



Now, you will see that what is the property of a local minimum? So, let us look at the contour plots of this function f x equal to x 1 into e to the power of x 1 square minus x 2 square. Well let us take point here, now if we look at this direction you will see that the value of the function keeps on decreasing. So, the value of the function on this different contours is shown here, so at on this on this contour plot, the value of the function is minus 0.1. So, as you move in the interior it becomes minus 2 minus comma minus 0.2, minus 0.3 and so on and then if you move further then the value starts increasing. So, let us take a point here. Now, if you move in this direction the value of the function increases. So, you will see that from 0.1 it goes to 0.2, 0.3.

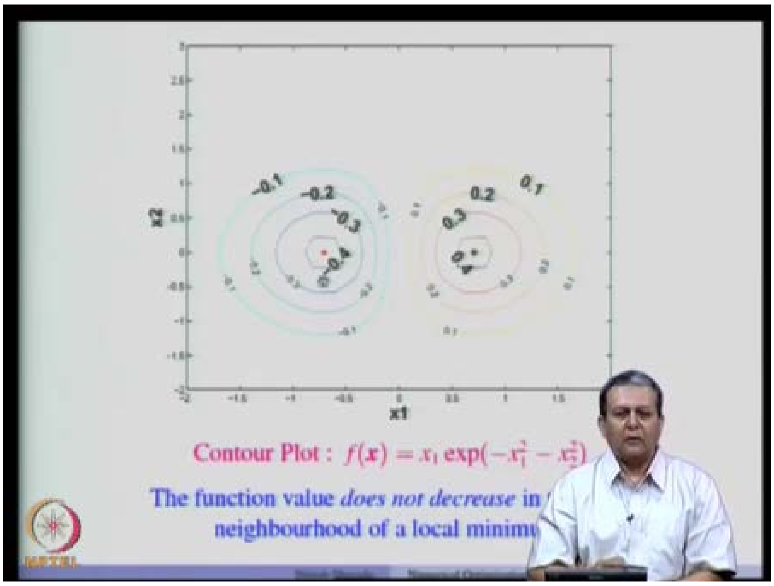
Now, let us see this point is a local minimum and this point is a local maximum, as you can see from the surface plot. Now, you will see that if you look at this point which is a local minimum, now in any direction, if you move the value of the function increases. So, there is at least no decrease in the objective function in the neighborhood of this minimum. So, this is a characterization of a local minimum that in the neighborhood the value of the function does not decrease it may remain constant, but it does not decrease. So, this is the important observation that we make is that, the function value does not decrease in the local number hood of a local minimum. That you can see clearly at this point. That if you move in any direction from this local minimum the value of the function does not decrease in the local network.

(Refer Slide Time: 06:35)



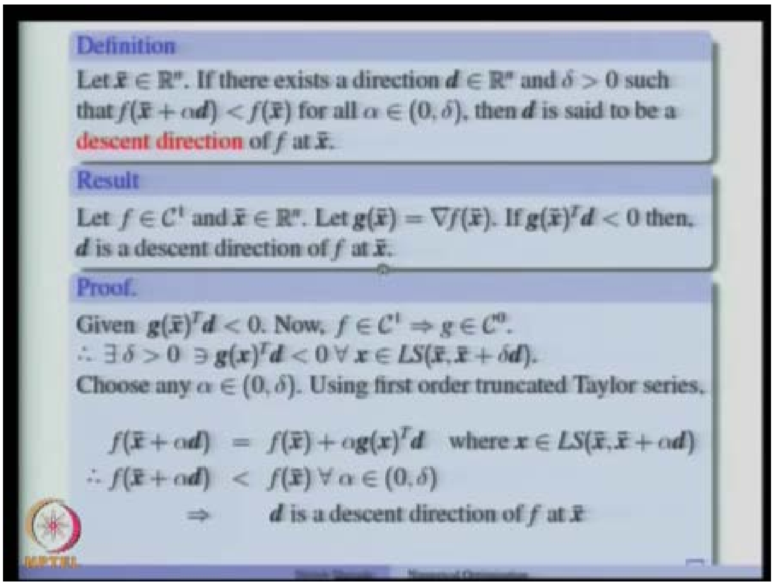
So, let us define the descent direction, let us take a point x bar in R n. now, if there exists a direction d in R n and a constant delta greater than 0 such that the value of the function at x bar plus alpha d is less than the value of the function at x bar. For all alpha in the range 0 to delta, then d is said to be a descent direction of f at x bar.

(Refer Slide Time: 07:08)



So, if you look at this point suppose if you take this point and move start moving in this direction and so this is a descent direction from this point. If you take another point again this is a descent direction.

(Refer Slide Time: 07:25)



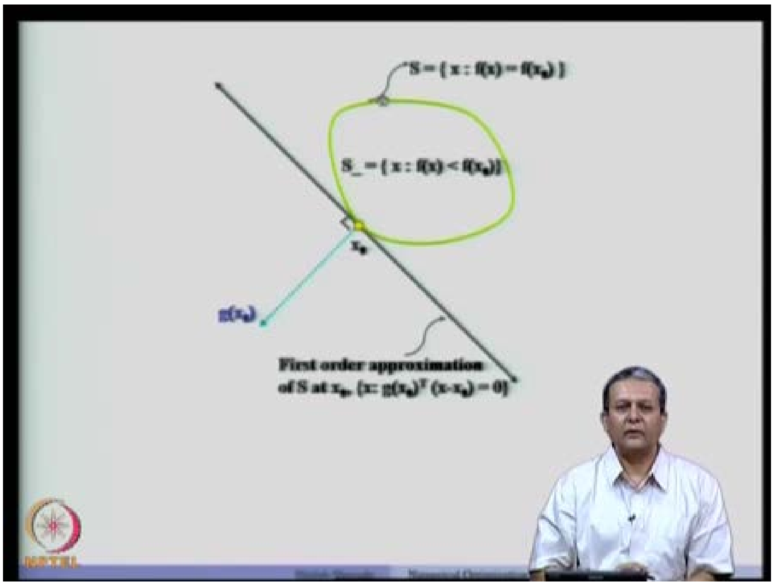
So, these are some descent directions at different points. Now, how do we check that whether the given direction d is a descent direction, so can we algebraically verify this. Now, for that purpose the differentiability of the function is very important. So, let us assume that the function is differentiable, continuously differentiable. Let us take a point x bar in R n and let us denote the gradient of the function at x bar as g of x bar, then if g x bar transpose d is less than 0, then d is a descent direction of f at x bar.

So, what it means is that direction d should make an obtuse angle with the gradient direction at that point x bar in that case d is certainly a descent direction. So, if this happens then d is certainly a descent direction of f of x bar. So, let us prove this result, now what we are given is that we are given a point x bar in R n and a function which is continuously differentiable and we also know that g x bar transpose d is less than 0. Now, remember that f is continuously differentiable, so the derivative of f is continuous.

So therefore, there exists some delta, which is greater than 0 such that g x transpose d is less than 0, for a x and the open line segment joining x bar and x bar plus delta d. So, let us choose any alpha in the open interval 0 to delta and therefore, f of x if we use first order truncated Taylor series then f of x bar plus alpha d is nothing but f of x bar plus alpha g x transpose d, where x is a point on the line segment joining open line segment joining x bar and x bar plus alpha d.

Because alpha belongs to 0 delta and x it belongs to this open line segment, we can see that g x transpose d is less than 0, because of this fact that g x bar transpose d is less than 0 and g is a continuous function. And therefore, what we have is f of x bar plus alpha d is less than f of x bar for all alpha in the close interval 0 to delta, which means that d is a descent direction of f at x bar as per the definition given here. So, this is the one way to characterize a descent direction that if the dot product of the direction with the gradient vector at the given point x bar is less than 0 or the direction makes an obtuse angle with the gradient at x bar, then d is a descent direction of f at x bar.

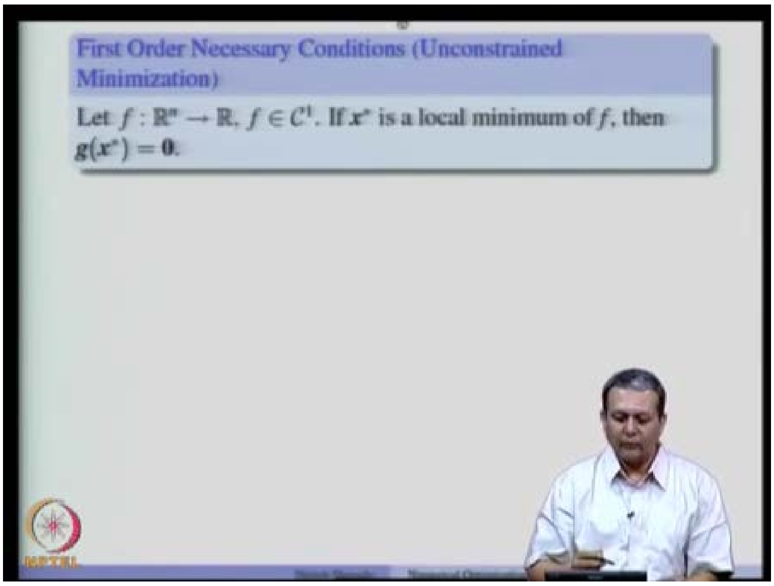
(Reference Slide Time: 10:30)



So, here in the figure we have shown one contour of a function related to particular function value, so let this point be x 0, so s is the set of all x, such that f of x equal to f of x naught. So, the value of the function on this contour is f of x naught. Now, in the interior of the function the value of the function is less than f of x, in the interior of this contour the value of the function is less than f of x naught. And g of x not is the gradient direction at x naught, gradient of f at x naught, so this is pointing in this direction.

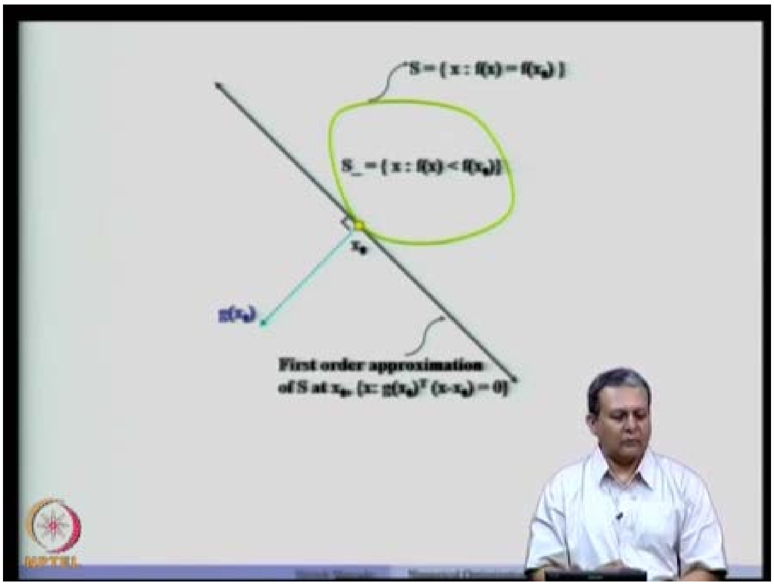
So, the gradient points in this direction means that the value of the function as you move along the gradient direction is going to increase. Now, this is the first order approximation of the function f at x naught and that is the hyper plane as we seen it earlier. Now, if we take any direction, which makes an obtuse angle with g of x naught. Then suppose we take this direction then this direction is a descent direction, because there exists some delta which is greater than 0, such that if you take a small step along the direction up to delta the function value decreases. Then beyond that it is possible that the function value increases.

(Refer Slide Time: 12:05)



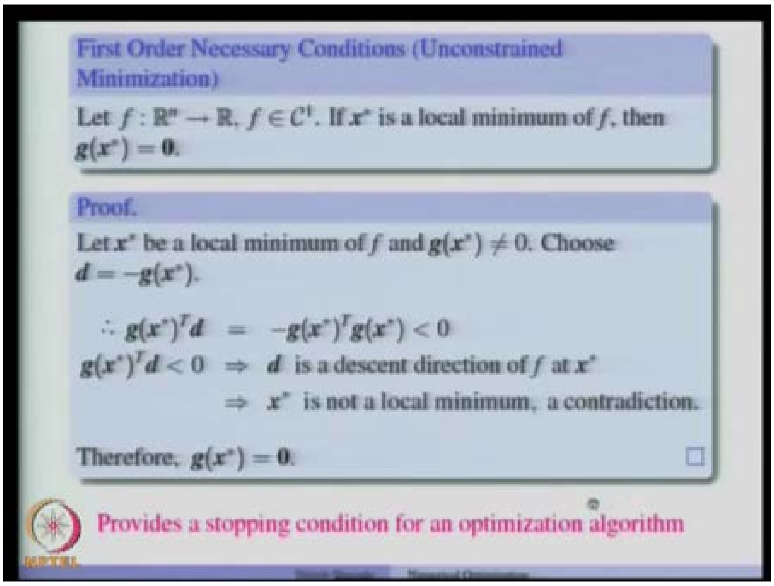
So, similarly one can if you take a direction minus g of x naught that also makes an that also makes a 180 degrees angle with g of x naught. But you can see that g of x naught transpose minus g of x naught will be less than 0, because g of x naught is not equal to 0.

(Refer Slide Time: 12:13)



So, you will see that that there exists many directions along which if we make a movement the value of the function would decrease. And we will see a result which now states that a local minimum is a point where there does not exist a descent direction in the local network.

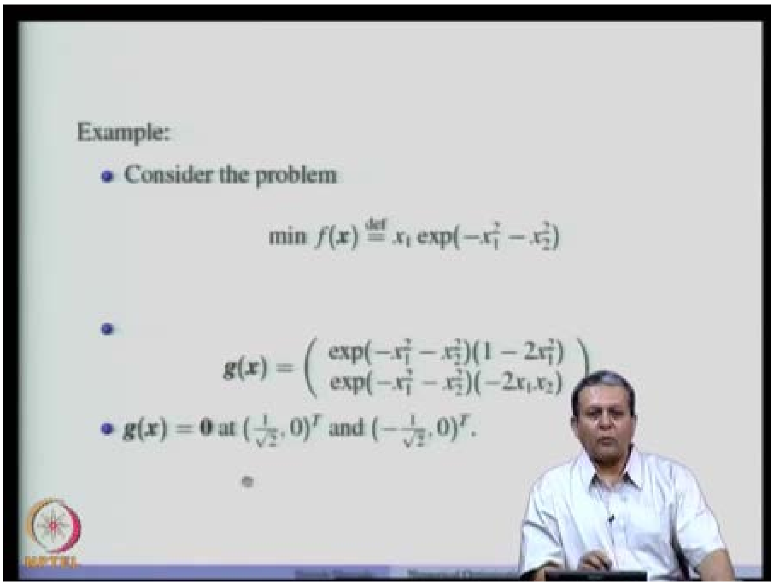
(Refer Slide Time: 13:06)



So, that is called a first order necessary condition for a unconstrained minimization problem. So, let us consider a function from R into R and f is continuously differentiable. Now, if x star is a local minimum of f, then the claim is that g of x star is 0. So, let us prove this result, let us assume that x star is a local minimum of f and g of x star is not equal to 0. Now, let us suppose choose d to be minus g of x star. Now, what we can do is that we can find out what happens to g x star transpose d, where d is minus g x star. So, g x star transpose d is nothing but minus g x star transpose g x star. Now, remember that g of x star not equal to 0, so this quantity is strictly less than 0. Therefore, we have got a direction d such that g of x bar x star transpose d is less than 0.

So, which means that d is a descent direction of f x star. So, if d is the descent direction of f x star, so that means that from x star it is possible to move to make a small movement along the direction d and decrease the function f. So, which means that x star is not a local minimum and that contradicts the assumption that x star is a local minimum of f. So, we are able to find a direction d such that g transpose d is less than 0, which means that d is descent direction, which means that x star is not a local minimum and that has resulted in the contradiction. So, our assumption that x star is the local minimum of f and g of x star not equal to 0 is not correct. So, if x star is the local minimum of f, g of x star has to be 0. So, one good thing about this first order necessary condition is that it provides stopping condition for an optimization algorithm, so will see more about this sometime later.

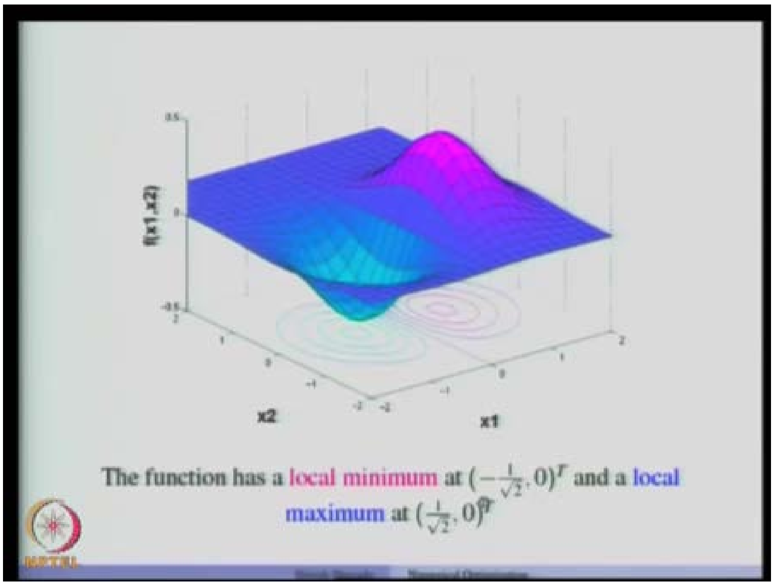
(Refer Slide Time: 15:21)



So, let us consider a example; consider the problem to minimize x 1 into e to the power minus x 1 square minus x 2 square. So, we have seen the contour plot as well as the surface plots of this function earlier. Now, let us write it down the gradient vector of this function and that expression is given here. Now, we are interested in finding out the point x star such that g of x star is a 0 vector. So, both the components of this vector should be 0.

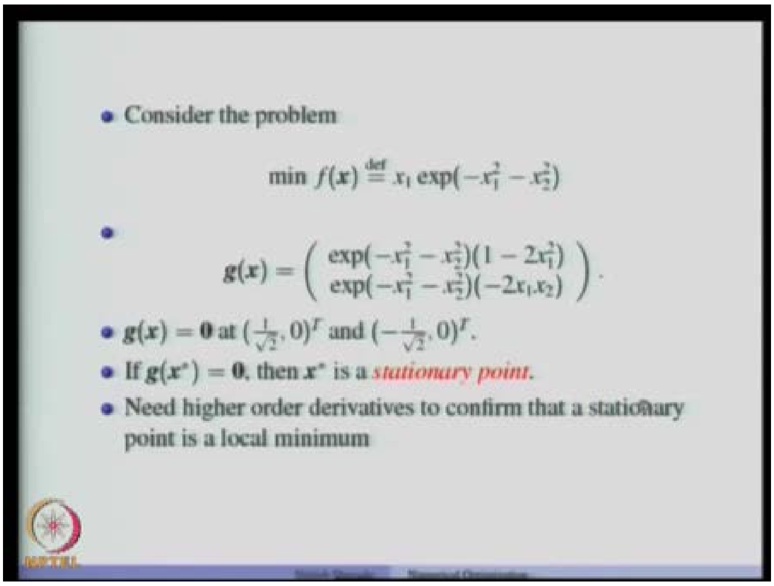
Now, you will see that both the components of this vector will be 0 only when x 1 is 0 and x 2 is 0, I am sorry. So, both the components of this function will be 0 when either at 1 by root 2 0 and minus 1 by root 2 0. So, these two points are the candidates for local minimum. Now, you will see that if you plug in these values here you will get the gradient vector to be 0. So now, among these two points, which are the local minima either this or this or both, so that we have to find out.

(Refer Slide Time: 16:57)



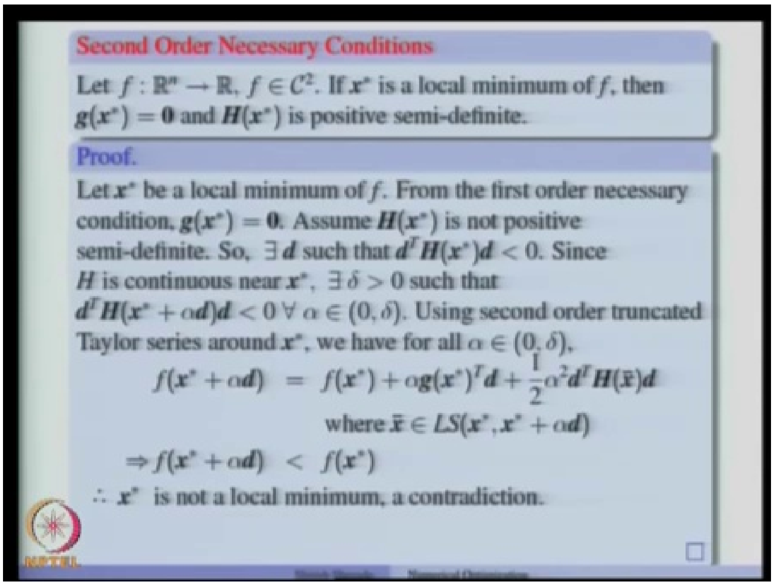
So, let us again look at the surface plot of this function. Now, minus 1 by root 2 0 is a point, which is somewhere here and 1 by root 2 0 is a point somewhere here. So, you can see that minus 1 by root 2 0 is a point which corresponds to the local minimum and minus 1 by root 2 is a point which corresponds to the local maximum. So, the function has a local minimum at minus 1 by root 2 0 and local maximum at 1 by root 2 0, but we have seen that the gradient of the function is the 0 vector at both these points.

(Refer Slide Time: 17:44)



So, certainly the information that we get from the gradient vector is not going to be enough. Now, the points at which the gradient of the function becomes 0 vector, such points are called stationary points. So, in this case we have two stationary points where the gradient of the function becomes a 0 vector. Now, among these two points how do we choose which one is a local minimum or a local maximum or neither. So, we need some higher order derivative information to confirm that a stationary point is indeed a local minimum.

(Refer Slide Time: 18:30)



For that purpose we have to look at the second order necessary condition. Now, this second order necessary conditions are the necessary conditions, which use a second order derivative information of the function. So, for that purpose we have to assume that the function is twice continuously differentiable. So, let us consider a function from R n to R n, which is twice continuously differentiable, where is x star is a local minimum of f, then the gradient at x star is a 0 vector and hessian matrix at x star is positive semi definite at x. Let us prove this result, so let us take a x star to be a local minimum of f.

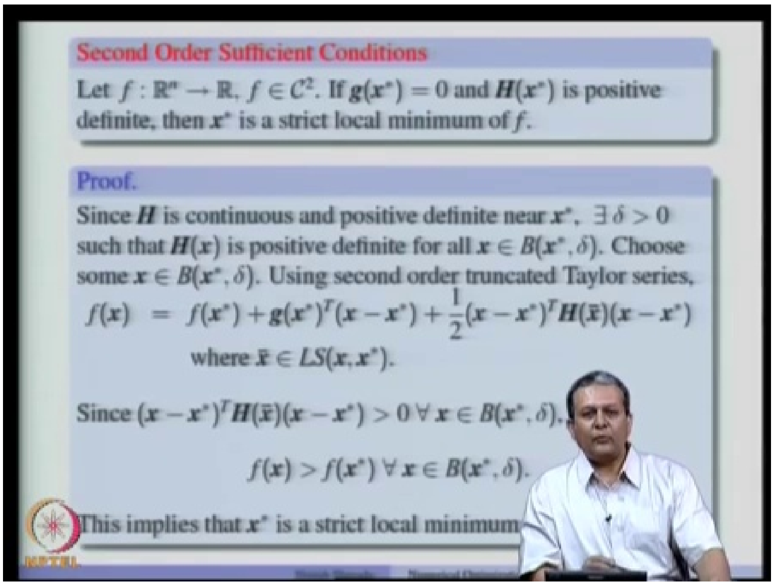
Now, from the first order necessary condition we have g of x star equal to 0 that we have already proved. Now, we have to show that if x star is a local minimum then H of x star is positive semi definite. So, let us assume that h of x star is not positive semi definite. So, there exists some direction d such that d transpose H x star d is less than 0. Now, remember that x is twice continuously differentiable, so h is continuous near x star.

Therefore, there exists some delta such that d transpose H of x star plus alpha d into d is less than 0 for all alpha in the range 0 to delta.

Now, we can use the second order truncated Taylor series to write the expansion of f of x star plus alpha d as f of x star plus alpha g x star transpose d plus half alpha square d transpose h x bar d. Now, may I remember that this quantity is less than 0, because we have taken a point x bar, which is on the line segment joining x bar and x bar to alpha d and because of the continuity of h we have this d transpose h of x star plus alpha d into d is less than 0 for all alpha in the range 0 to delta.

So, this quantity is less than 0 and g of x star is 0, therefore what we have is f of x star plus alpha d is less than f of x star. So, which means that we are able to find a direction d and alpha, which is a positive number such that if you make a movement of length alpha along the direction d the value of the function at that new point will be less than the value of the function at x star. So, some small local movement along x star along the direction d enabled us to decrease the objective function f further in which contradicts the fact that x star is a local minimum. Therefore, if x star is a local minimum of twice differentiable function then the gradient of the function should be 0 vector and the hessian of the function at that point x star is positive semi definite at x.

(Refer Slide Time: 21:42)



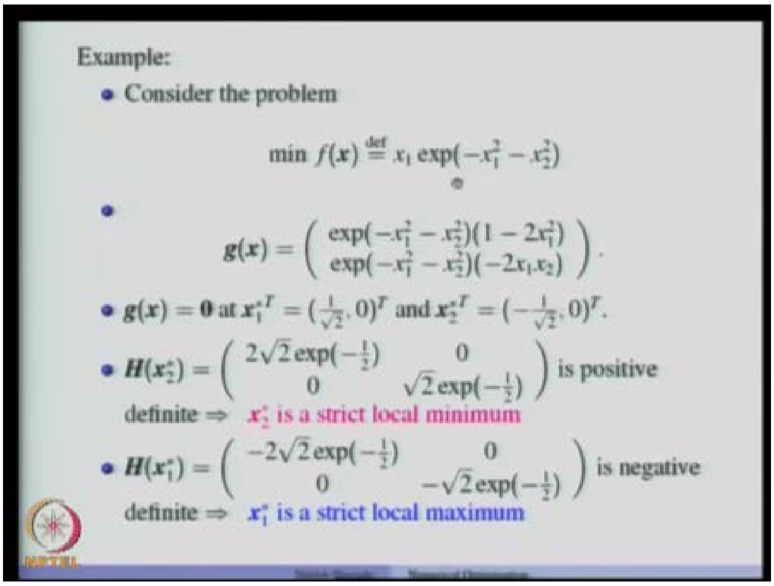
Now, this second order necessary conditions, they are not sufficient, so this is another necessary condition. So, to get sufficient conditions let us assume that at a point x star in

R n the gradient of the function is 0 and the hessian is positive definite, then the claim is that x star is the strict local minimum of f, remember that we are assuming the twice differentiability of f. So, to prove this result we again have to use second order truncated Taylor series, so let us see how to do that. Now, we are given that f is twice continuously differentiable, so h is continuous and positive definite near x star. And therefore, there exists some delta such that h x is positive definite in that delta neighborhood of x star. Now, let us use this fact and write the Taylor series

So, let us first choose some x in that delta neighborhood of x star and use the second order truncated Taylor series to write f of x as f of x star plus. The second quantity involved in the gradient and the third quantity involved in the hessian. Remember that the hessian is calculated a point x bar on the open line segment joining x and x star. Now, clearly x bar belongs to this open ball of radius delta around x star. So, which means that this quantity is always positive and we know that g of x star is equal to 0.

So therefore, what we have is f of x will be strictly greater than f of x star for all x in the delta neighborhood of x star. So, which means that x star is a strict local minimum of f. So, remember that for sufficient conditions we need h of x star to be positive definite, then the result is also more stronger, in the sense that it says that, in that case x star is the strict local minimum of f. So, the necessary conditions were related to a local minimum while here the sufficient conditions are related to strict local minimum of f.

(Refer Slide Time: 24:10)

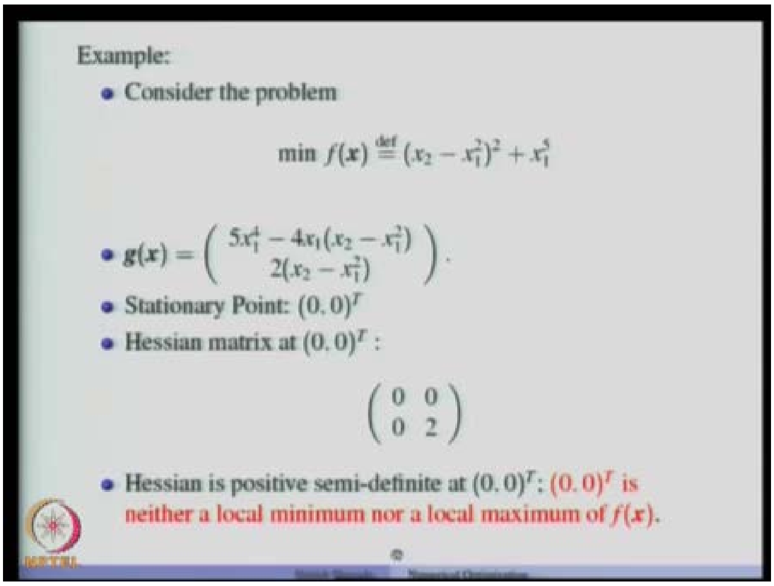


Now, will see some examples to illustrate this necessary, and sufficient conditions. So, let us consider the same problem that we have considered earlier and that is to minimize the function x 1 into e to the power minus x 1 square minus x 2 square. So, the gradient of this function is given here we have seen this earlier and the gradient vanishes at 1 by root 2 0 and minus 1 by root 2 0 and we have seen that these are stationary points. But among these two points, which one is a local minimum and which one is a local maximum and which one is neither that is what we are interested in finding out.

Now, to find out whichever stationary point is a local minimum or not we need to go for a higher order derivative information. So, let us take the hessian matrix. So, let us first take the point x 2 star and the hessian matrix evaluated at x 2 star is given here. Now, you will see that which term is positive this term is positive, so the all the principle minus of this matrix are positive, so which means that the matrix is positive definite. Now, remember that the half diagonal elements are 0 here.

So since, this is a positive definite matrix we can say that because of the second order sufficient conditions we can say that x 2 star is a strict local minimum. Now, what happens at x 1 star? So, let us write down the hessian matrix at x 1 star, the hessian matrix at x 1 star will look like this and you will clearly see that this matrix is related to it and which means that x 1 star is a strict local maximum. So, if you recall the plot of the function we saw one local minimum and one local maximum. So, in fact x 2 star is a strict local minimum and x 1 star is a strict local maximum for this function.

(Refer Slide Time: 26:23)

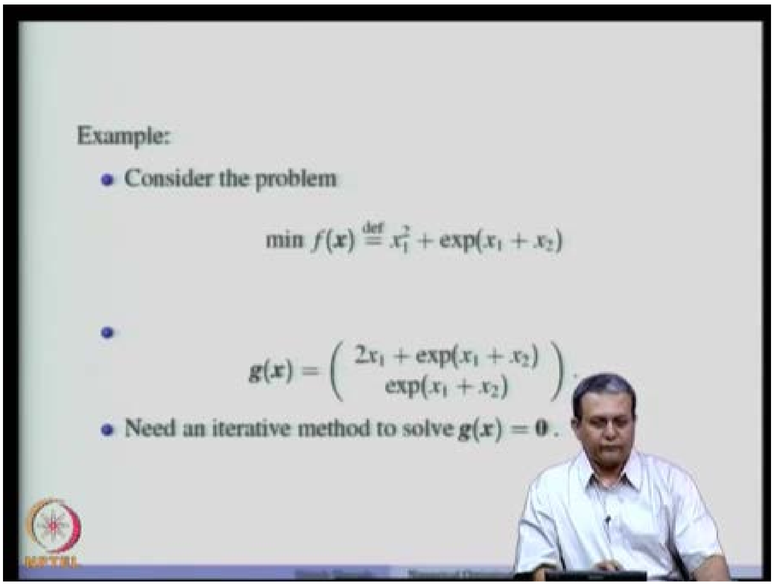


Let us consider another problem. Now, we again this is a two dimensional optimization problem, we are trying to minimize a function x 2 minus x 1 square whole square plus x 1 to the power 5. Now, to find out the stationary points what we have to do is that first we have to write the gradient and equate it to 0 to get the stationary points and then look at the hessian matrix of those points to infer whether the, which to infer whether which of the stationary points are local minima maxima or neither. So, let us write down the gradient vector of the function, so which is shown here. Now, you will see that this gradient vector will be 0 vector at 0.00 or the origin. So, the stationary point of this function is the origin, now this is the only stationary point and we know how to find out whether it is a local minimum or local maximum or neither.

Now, for that purpose let us look at the hessian matrix of this function. Now, hessian matrix of this function at the origin is given here. Now, you will see that this matrix is positive semi definite it is not a positive definite matrix. So, second order conditions or second order sufficient conditions are not satisfied. So, we really cannot conclude anything much from this second order information. Now, if you look at the function, so at 0 0 the function value is so at the origin the function value is 0. Now, if we increase x 1 then the function value increases, and if we increase if we decrease x 1 then the function value decreases.

So, 0 0 is a point and along one direction if we move by keeping x 2 constant function value increases and along the other direction if x 1 decreases then the function value decreases, so 0 0 comes out to be a saddle point. So, this example illustrates that the hessian matrix is positive semi definite, but this stationary point is a saddle point. And this does not have a this point is neither a local maximum nor a local minimum of f of x

(Reference Slide Time: 29:20)

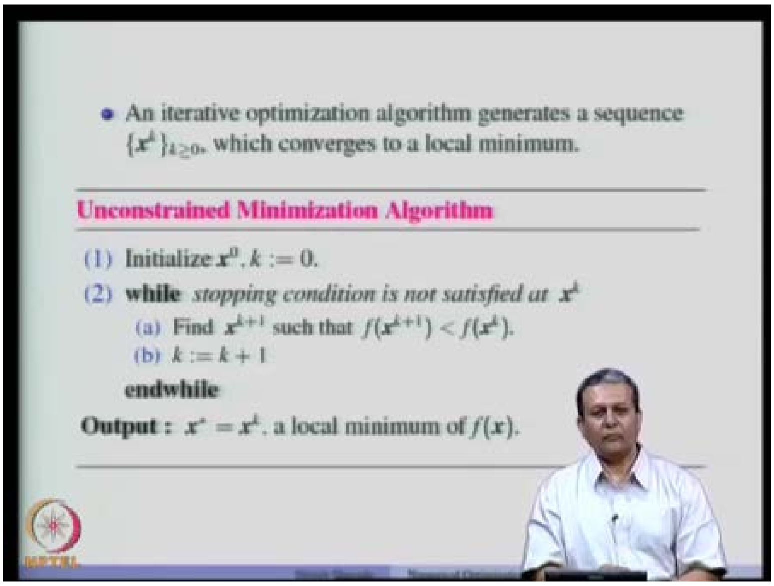


Now, let us consider another problem where the again a two dimensional problem, where we are trying to minimize the function x 1 square plus e to the power x 1 plus x 2. Now, as I said earlier first step is to get the stationary points. So, let us write the gradient vector. Now, equating this to 0 does not directly give us the stationary points of this, so we need an iterative method to solve g x equal to 0. So, finding stationary points of a function is difficult in this case, because there are no close form expressions to get the stationary points, unlike in the previous cases that we saw again.

So, we need an iterative method or an algorithm to find out the stationary point and this is going to be the discussion point for the next few lectures, that how to design an iterative algorithm for a unconstrained minimization problem. Now, remember that all those optimization algorithms that we are going to study will give us a stationary point. And we have to find out the behavior of those stationary points to check whether that point is a local minimum or a local maximum or a saddle point.

Now, the way those algorithms are designed will eliminate the possibility that will end appear a local maximum. So, the possibility of a local minimum or a stationary or a saddle point exist and finally, it is of the user to find out whether the given point is the stationary point that is given by the algorithm is indeed a local minimum or not.

(Refer Slide Time: 31:34)

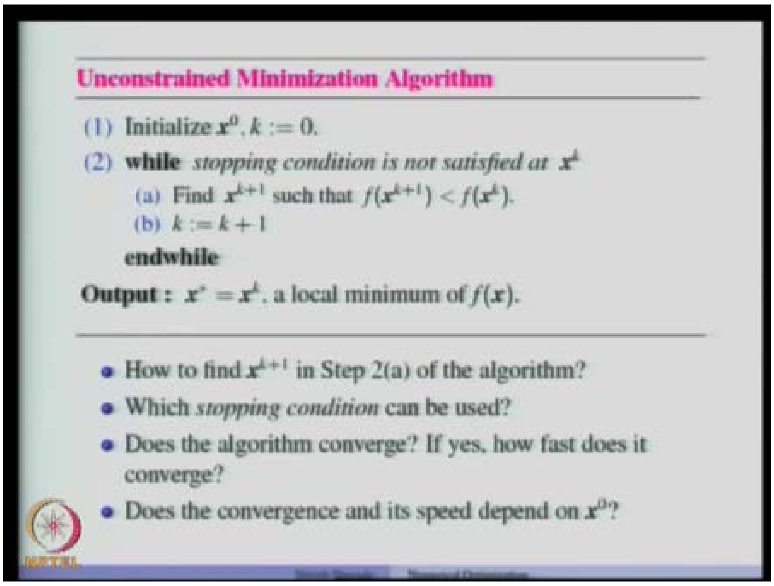


So, the typical iterative optimization algorithm would generate the sequence of points, so let us denote that sequence by x k. So, remember that this the super script k is used to denote the element of the sequence and it is not the power. And the optimization algorithm when it generates the sequence x k we want that sequence to converge to a local minimum. So, conceptual unconstrained minimization algorithm is given here. So, the algorithm typically starts with some initial point x naught. Now, the index k denotes the iteration number, so at every iteration this index is incremented by 1.

So, the algorithm is very simple that while some stopping condition is not satisfied at x k, the algorithm finds a new point x k plus 1 such that the value of the function is going to decrease. So, value of the function at x k plus 1 is less than the value of the function at x k, then the iteration counter is incremented by 1 and the process is repeated. So, this whole process is repeated till some stopping condition is satisfied at x k and when the algorithm terminates what we get is a x star, which is nothing but the x k, the x k at which the algorithm terminated that x k will be our x star.

That x star is the local minimum of f of x. So, this is the conceptual unconstrained minimization algorithm. Now, along similar lines one can write the algorithm for unconstrained maximization problem.

(Refer Slide Time: 33:34)



Now, there are few questions that need to be answered and that will answer as part of this course. Those questions are related to this unconstrained minimization algorithm. So, one of the important questions is that, How do we find x k plus 1 in this state of this algorithm? Such that f of x k plus 1 less than f of x k. Is there any systematic way to find this x k plus 1, such that the objective functional value decreases at a new point. So, this is a very important question and different optimization methods use different ways to generate this x k plus 1 given the point x k. So, we will study those methods some of those methods in the due course.

Then the next question we would like to ask, Is that which stopping condition can be used? As we will see later, there exists different stopping conditions to terminate an algorithm, remember that we cannot afford to generate an infinite sequence of numbers so we want some stopping condition where the algorithm terminates. Now, which stopping condition is more appropriate those we will study that in the next lecture. Then suppose we have identified a method to generate x k plus 1, we also found out the stopping condition. Now, has the algorithm converged to a minimum and if it has converted to a minimum what was its speed?

So, this is also an important question, because sometimes the speed of an algorithm is an important issue and care has to be taken to ensure that the algorithm does converge fast to the solution. So, in the other question that we want to ask is that how about the initial point x naught? Does the convergence of the algorithm depend on the initial point x naught? And does the speed of the algorithm depend on this initial point x naught?

Now, we will start answering this questions some of this questions in the next class. And then move on to some of the methods, which define or which get a point x k plus 1 in this type of algorithms such that the value of the function decreases with respect to the current point. There exist different methods, so we will study those methods some time later in this course.

Thank you.