NPTEL

NPTEL ONLINE CERTIFICATION COURSE

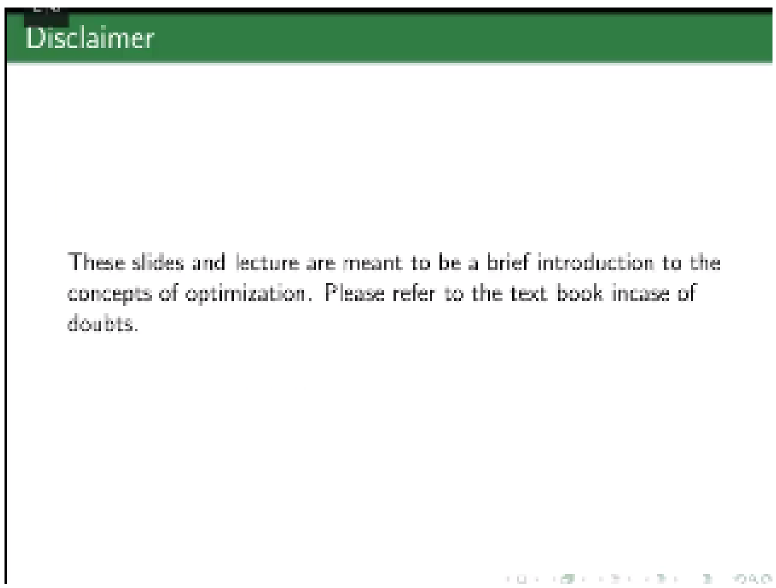
Introduction to Optimization

Abhinav Garlapati

Introduction to Machine Learning
29th Jan 2016

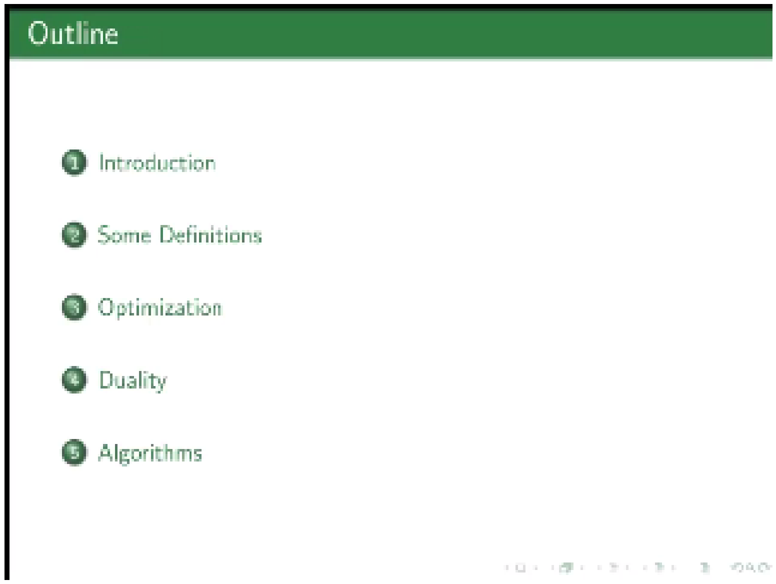
Hello everyone I am Abhinav in this unit we will be covering the basic concept of optimization, which should be useful in this course.

(Refer Slide Time: 00:26)



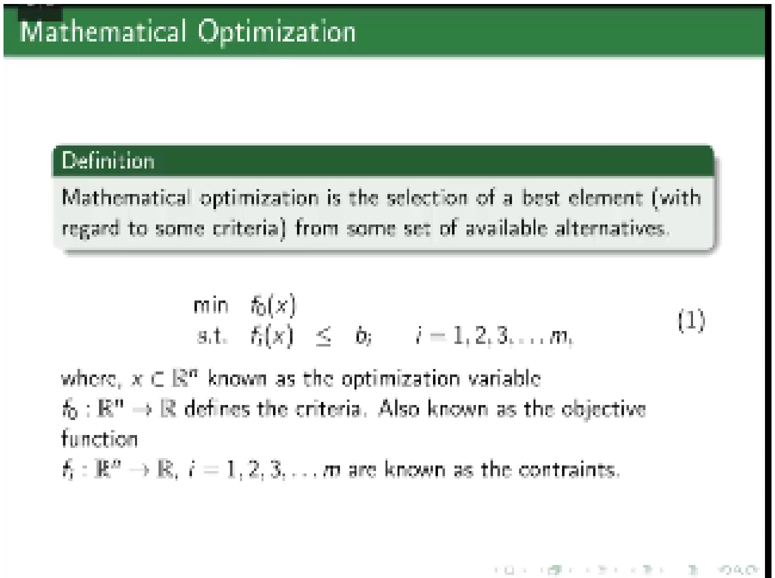
So before going in to detail a small disclaimer this tutorial is meant to be a small introduction for a complete understanding of these concepts please refer to any standard text book.

(Refer Slide Time: 00:43)



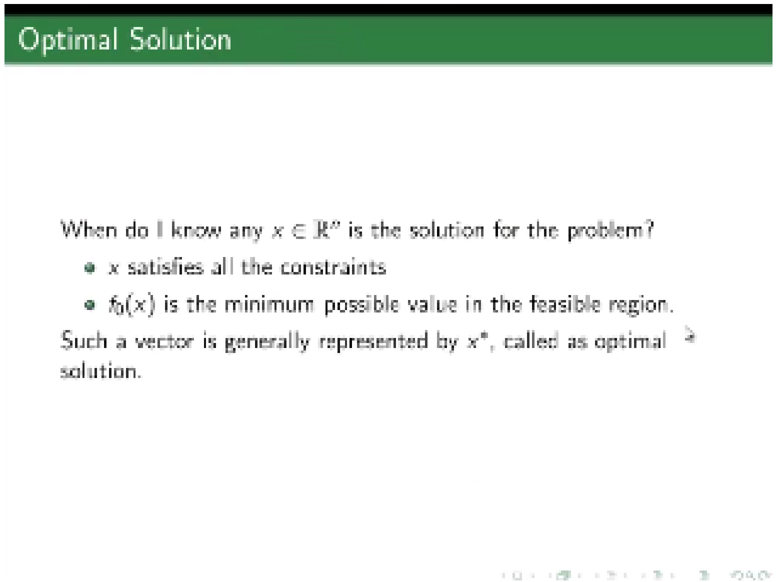
This tutorial is broken in to five chunks first let us start off with the introduction.

(Refer Slide Time: 00:53)



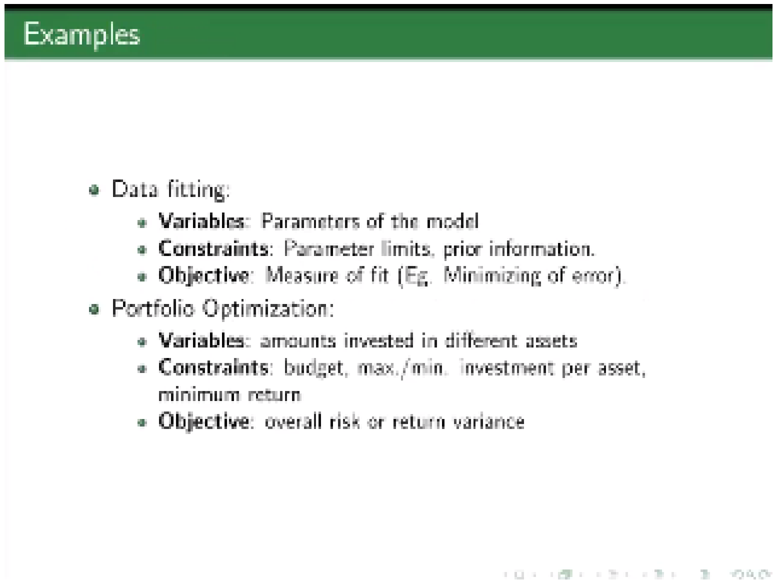
What is mathematical optimization? Mathematical optimization according to keep it here is a selection of a best element with regard to some criteria from some set of available alternatives, now let us look at the mathematical formulation for the same, here we are trying to minimize f_0 (x) subject to m constraints of the form fi of x less than or equal to bi. F_0 is also known as the objective function fi are the constraints and x is known as the optimization variable.

(Refer Slide Time: 01:37)



X is known as the solution for the problem if it is satisfies all the constraints and it minimizes f0 of x such a solution is known as the optimal solution and it is represented by x^* so through all this tutorial whenever you see x* it represents the optimal solution for the optimization problem.

(Refer Slide Time: 02:03)



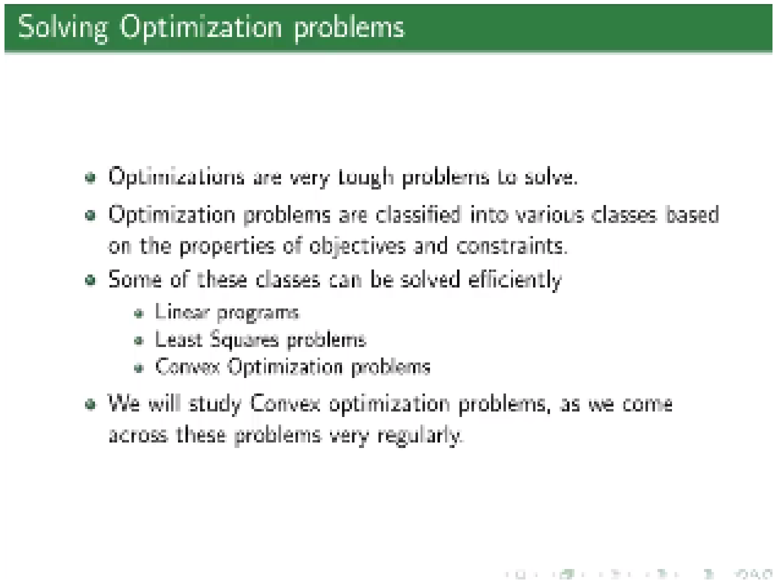
Now let us look at some examples where optimization is use first data fitting, data fitting is a very common problem in the field of machine learning, what I mean by data fitting? Data fitting is fitting of a parametric model given some data. So one's is example both linear regression in linear regression we are trying to fit a linear model whose parameters are β is, so those translate to the optimization variables here.

And constraints, constraints in general include something like parameter limits or prior information which already include, so what is the specific example of linear regression we do not have any constraints. And what would be the objective? You would try to fit get the best fit for the model so one way of doing this would be minimizing this error and in linear regression we have seen how see to minimize this squared error.

So that forms the objective of the optimization problem, another example of application of optimization is portfolio optimization. So by portfolio optimization we mean to optimize the amount of money I can invest in various assets so these assets could be something like shares from different companies or any other investment options, so the variables would be the amount I invest in all the options available the constraints would bead it overall budget the maximum or the minimum investment per asset.

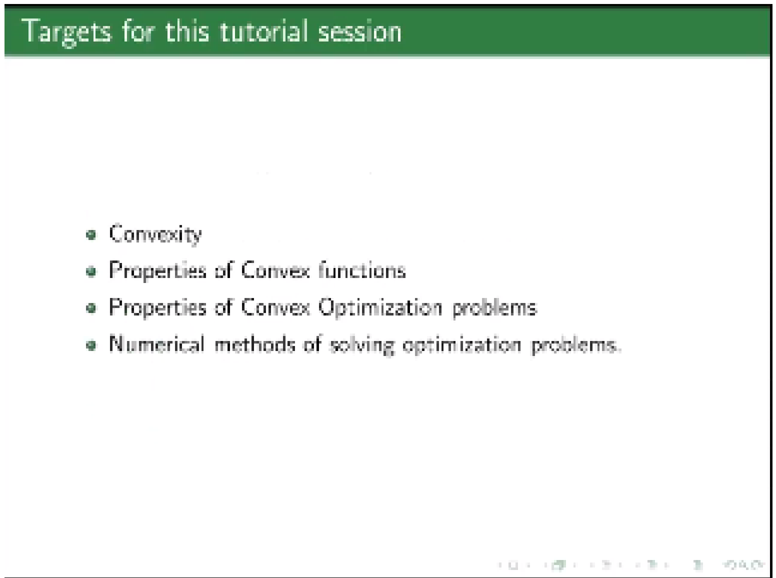
And the minimum return I expect from each asset, objective would be to minimize overall this or minimize the return variants you have seen what optimization problems are and you seen some examples. Now the next big question is how do we solve them.

(Refer Slide Time: 04:08)



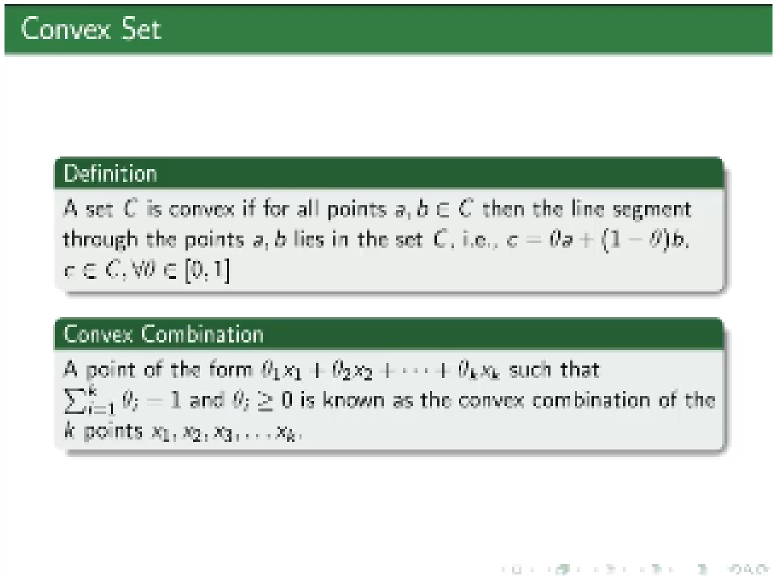
Optimization problems very difficult problems to solve in general optimization problems are classified in to different types based on the properties or objective and constraints the some of the examples are linear programs least square programs and convex optimization problems. These problems are well studied and can we solved efficiently not all class of problems can we solve very efficiently. When this tutorial we will be covering convex of machine problems in detail.

(Refer Slide Time: 04:45)



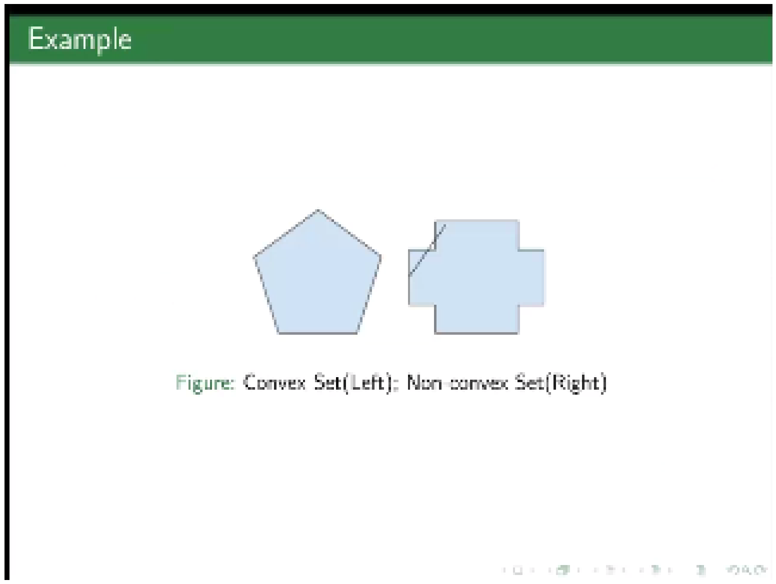
In this tutorial first we will be looking at convexity what convexity means and how do we define it, prop then we will look at properties of convex functions and then we will look at properties of convex optimization problems. And at the end we have briefly cover some numerical methods for solving optimization problems.

(Refer Slide Time: 05:11)



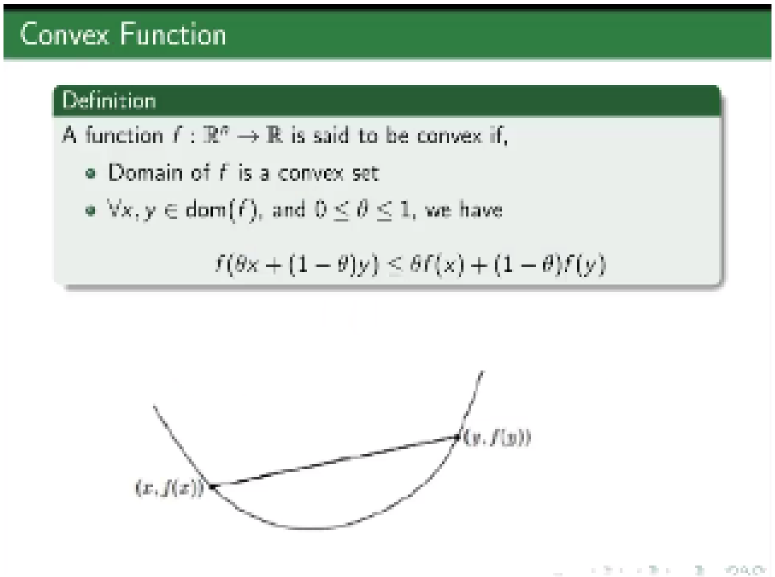
A set C is set to be convex is for all point a, b belong to the set the lines segment passing through this points should also lie inside the set, so mathematically we can see the asset all the points of the forms $\theta a + (1-\theta) b$, when θ lies in the close interval 0 to 1 should also belong to the set C. next let us look at the definition of convex combination. A point of the form $\theta_1 x_1 + \theta_2 x_2$ so what it $\theta_k x_k$ such that the coefficient some of 21 and the coefficients are non negative is known as the convex combination of this k point.

(Refer Slide Time: 06:04)



Now let us look at examples of convex set this pentagon is a convex set because any line joining two points inside the set lies inside the set whereas this set is a non convex set because this lines are going with joints 2 points here passes outside the set, thus theses points do not lie inside the set hence this does not satisfied the definition of convex set right it is not a convex set.

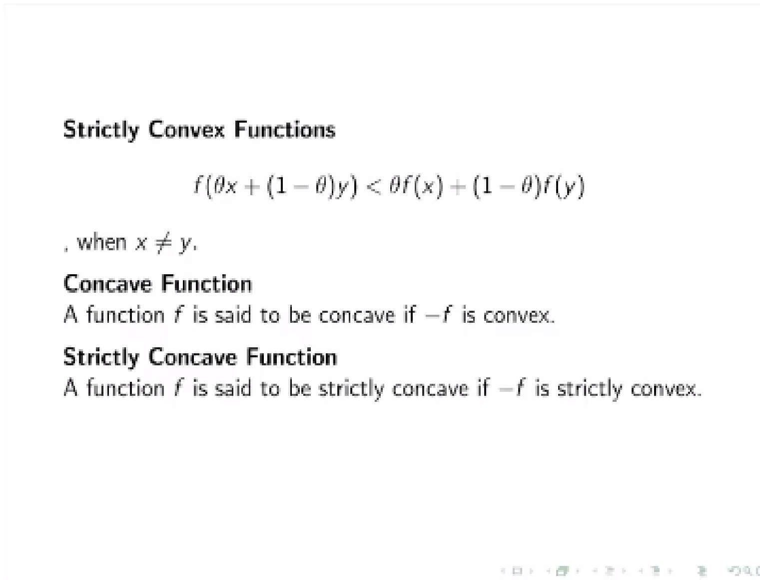
(Refer Slide Time: 06:36)



Let us look at the definition of convex function a function f is set to be convex if the domain is a convex set and if for all x, y which belong to the domain f the convex the value of the convex combination of these two points is less than or equal to the convex combination of the values at these individual points so what I mean is F of θx +1-θy that is the value of the function for the convex combination of these two points should be less than or equal to θf of x+1-θf of y this is the convex combination of the function values at these individual points.

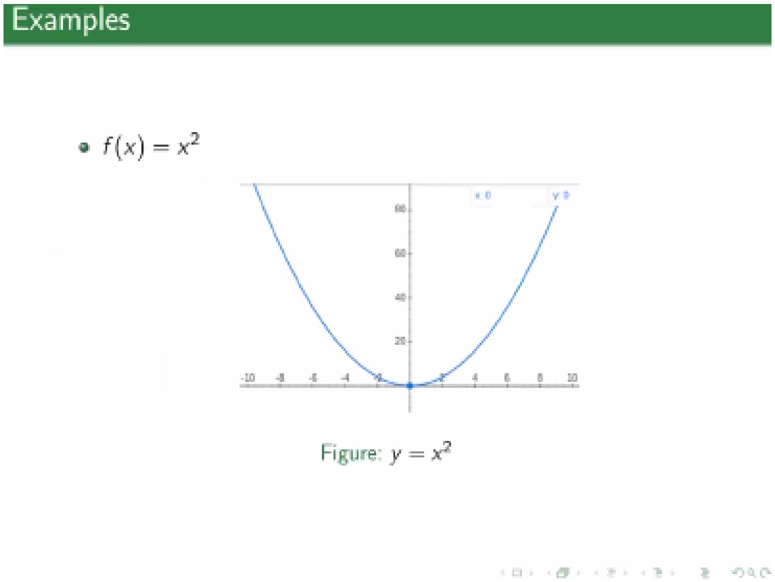
So geometrically you can see that the line joining x, f of x and y, f of y should lie above the curve so if this happens we can see that the value f of θx +1-θy is are the points along the curve and θf of x +1-θ f of y are points along the line segment joining x of x and y of a y so by ensuring that this always above the function we ensure that the inequality holds this making it a convex function.

(Refer Slide Time: 08:03)



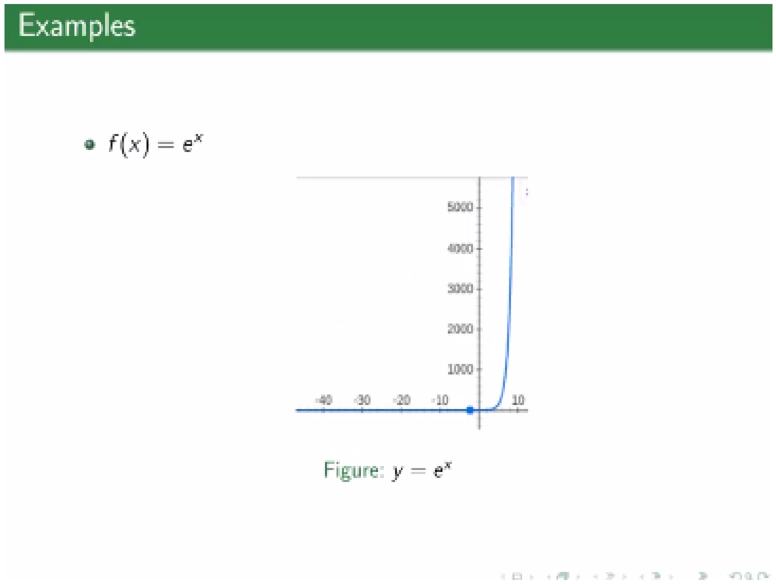
Now let us define what is strictly convex functions is in strictly convex functions the inequality becomes a strong in equality that is f of θx+1-θy is strictly less than θf of x+1-θf of y and now let us define what is concave function is a function f is said to be concave if –f is convex and then similarly we define a strictly concave function a function f is said to be strictly concave function if –f is strictly convex.

(Refer Slide Time: 08:42)



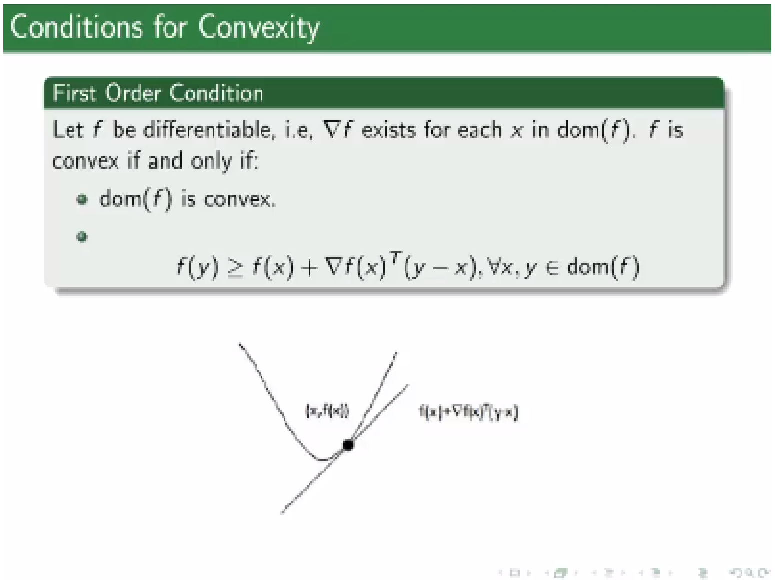
Now let u look some examples first f of x $=x^2$ is a convex function from the graph it is clearly evident that any line joining two points this will lie above the curve between these two points this line can also be verified by using the definition.

(Refer Slide Time: 09:07)



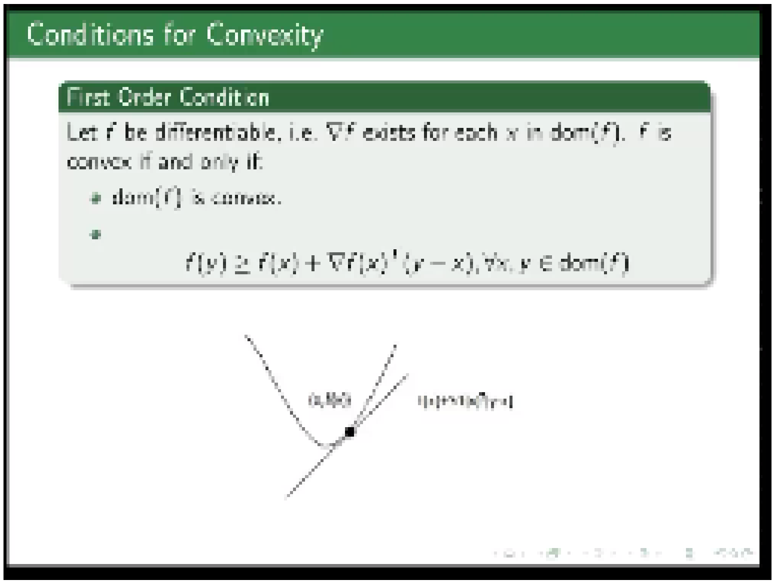
The next example that we see is graph of f of x= e^x again graphically you can clearly see that this is the convex function if you try to prove this according to the definition you can see that this is not tribunal so we would like to see if any other ways to check the convexity of function.

(Refer Slide Time: 09:36)



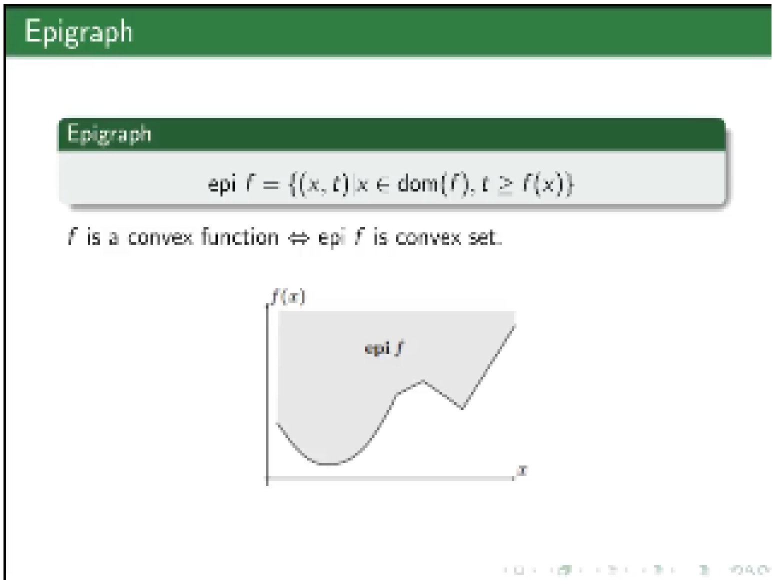
Let us look at the first order condition for the convexity let f be a differentiable function that is grade f exists for all x in a domain of f so our function f is convex if and only if the main of f is convex and this equality satisfy this inequality states that function should always lie above all it is tangents if you look at the right hand side carefully it is i nothing but the equation of the tangent at X / F of X and we expect this value to be less than F of 5 this is nothing but the condition saying the preclude is about the tangent.

(Refer Slide Time: 10:24)



Now let us look te same condition convex sity let F be twice diffrence which isd the function and the F wil bew convex and the diffrent of the layer and the main of the F is convex and the H C F is positive and it is so if ypou look at the second example of the heat power and the X the second derivative is always positive hence it can be prove that in to the convex function

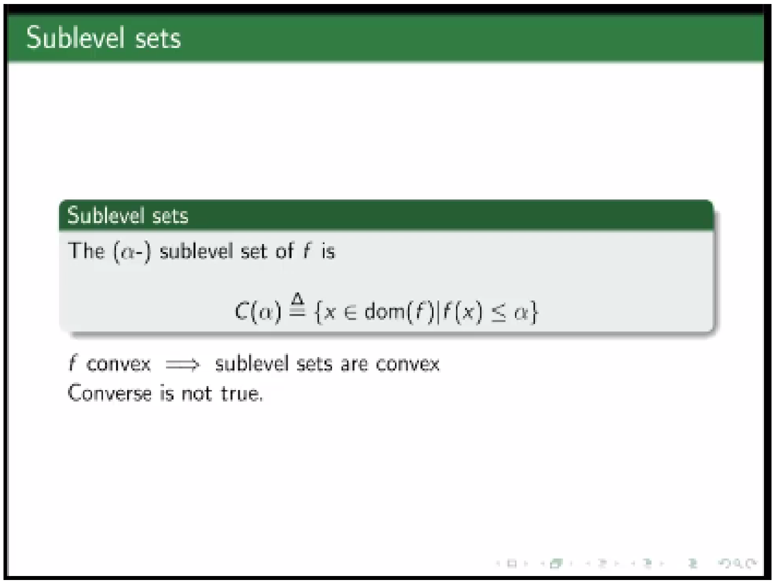
(Refer Slide Time: 11:01)



Now we defined a big graph of function for a given function f epi graph of f is defined as the set of all pairs x committee such that x belong to the domain of f and t is greater than or equal to f of x so if you look at the graph you can see that t area above the curve is belongs to the epi graph of the function. One important property to know is for a convex function the epi graph is always a convex set at the convex also holds statist if for a function the epi graph is a convex set then the function is convex.

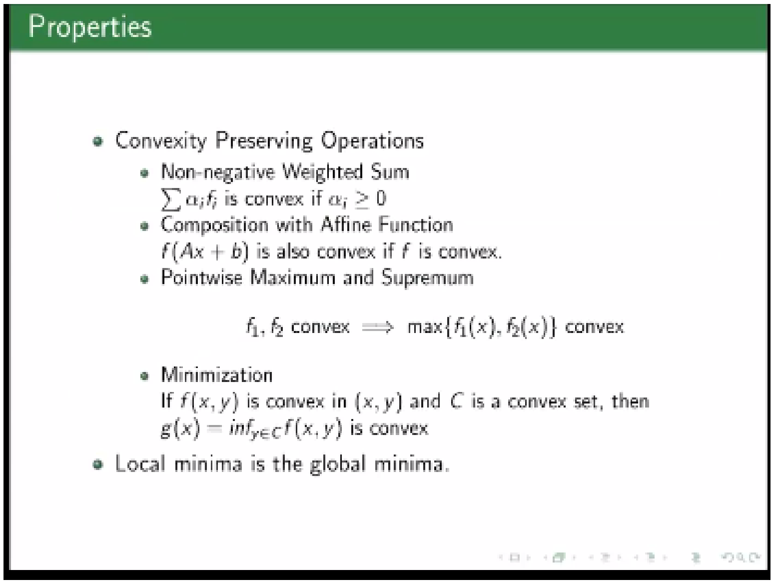
So we till now we have seen three ways of checking for convexity of a function first you can do the first order test or the second order set or you can check for convexity of the epi graph of the function.

(Refer Slide Time: 12:07)



Now let us look at what is sublevel sets of functional in alpha sub level set of a function F is set of all pints x which belong to the domain of F such that the value of the functional least point is <= alpha there is one important property that if the function is convex the sublevel set sets of the function are also convex it is important to note the converse is not true.

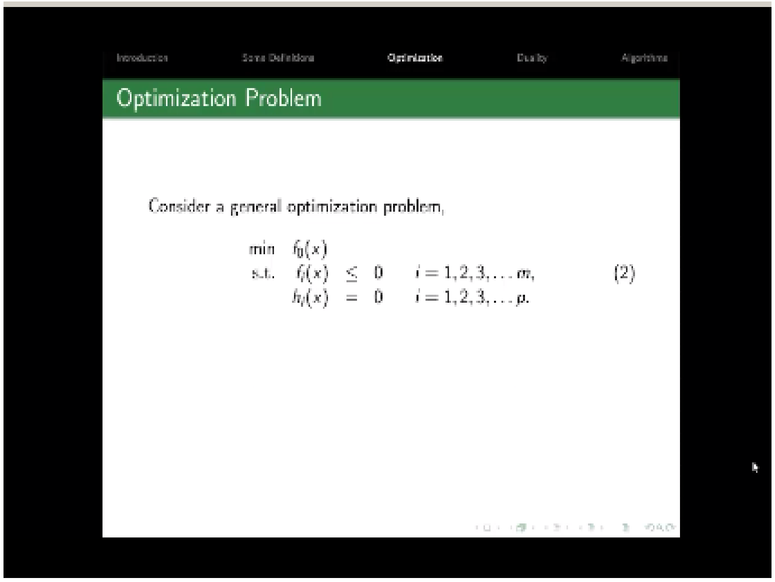
(Refer Slide Time: 12:45)



Now let us look at some other properties of convex functions first we will look at the function operations which preserve the convexity of the function first non negative weighted sum that is a non negative weighted some of various convex functions which still the main of convex function consider fi is the series of convex functions σ α f5 where α is are greater than 0 will also remain a convex function next composition with defined function a fine function is a linear transmission of x so x + b is an fine function if Female Speaker: convex then f of x + b is also convex, point wise maximum and supermum of to convex functions will also remain convex minimization.

If you look at as two variable function f x , y which is convex then if you try to minimize the function allow any one variable in a convex x which resultant function is also convex function the most important property of convex functions extracting local minima is also the global minima is a very powerful result which can proved easily this result grantees that the minima option while searching for the minimum of a convex function is the optimal solution as the important property of convex function is that they satisfy the general in equality which we have seen in the definition of the convex we n points so d value of d convex formation n point is less than equal to the value of the convex combination of n point is less than or equal to the convex combination of the values each of the function at each of the individual points. A local we are saying this is this the value of the average is less than the average of the values. Here by the average I mean a weighted average.

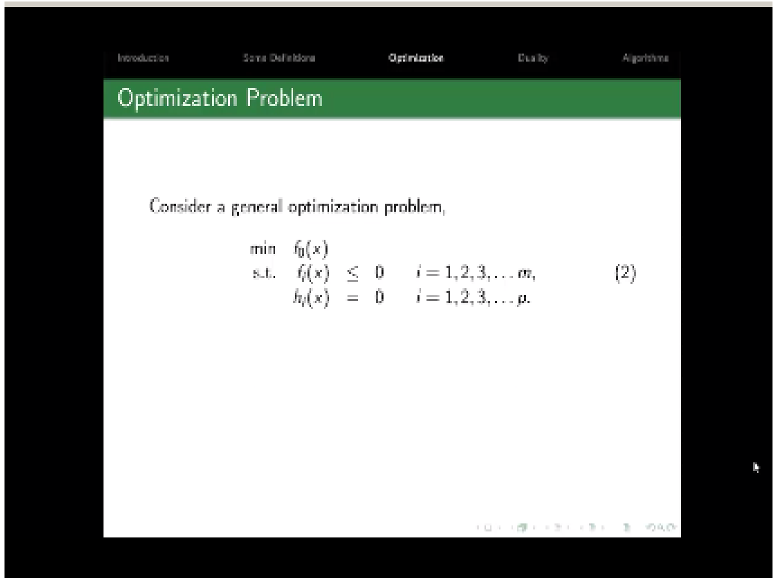
(Refer Slide Time: 15:05)



Now let us look at a general optimization problem, any optimization problem in generally can be reduce to this firm of minimize an objective function subject to few in equality constraints and few equality constraints. So the optimal value P* can also be return as inhume of $f_{0(x)}$ such that $f_i(x)<=0$ for I = 1to m and h_r (x) =0 for i=1to p. now the next question is why did arrived infimum instead of everyone.

In some function the minimum might not be attainable; it might just to 1to minimum value bit not actually attain it. Hence we write infimum instead of minimum.

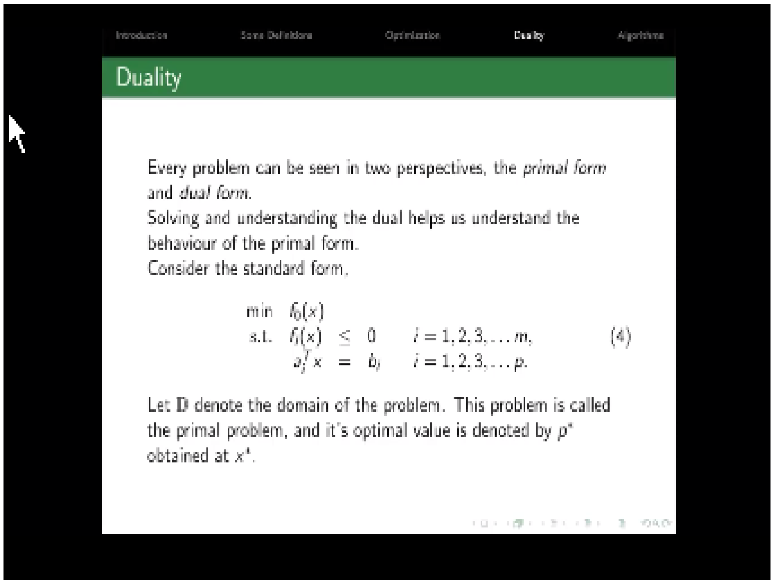
(Refer Slide Time: 16:01)



An optimization problem with satisfies the given three condition is not as a convex optimization problem. So first f_0 the objective function should be convex then the equality constraints f_i should also be convex. And the equality constraints should be a fine. When I say a fine it should be at the form A_i^T x =B_i so one can observe that the domain has become a convex set right now. So these equality constraints represent a sub level of a convex function so it is a convex set.

And a convex set intersection with a find function is a convex set. So why a convex problems are so interesting, so convex representation problems are interesting because with the properties of a convex functions and the convex set so first the most important property which is useful for us, is that the, if there is local minima anywhere. It is guarantee that is the global minimum for the function. So it makes are like very simple and we do not have search a lot for global minimum.

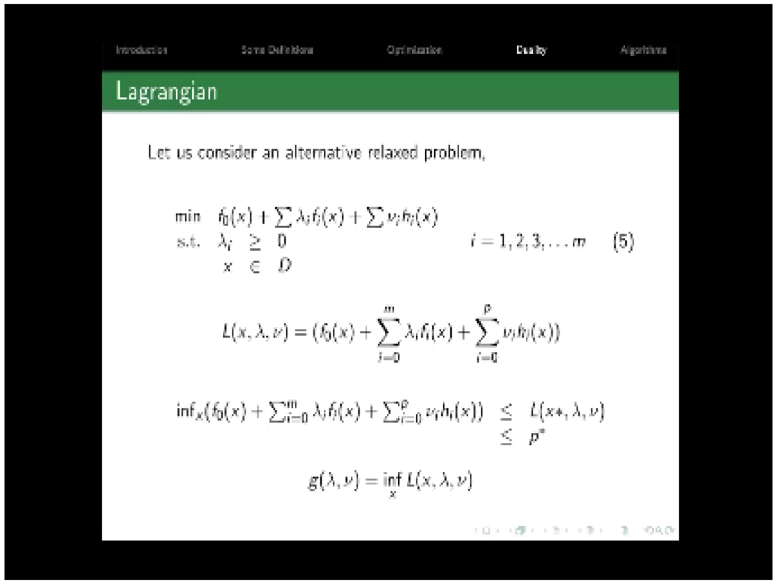
(Refer Slide Time: 17:13)



Every optimization problem can be seen in to perspective, one the prima form and the dual form, so whatever we seen to learn it generally known as the primal form, and we will now develop the dual form. So why do e need another view of the problem, so sometimes the primal form might be very difficult to solve it. So the dual form might be easier to solve and also cases some understanding on how the solution of the primal form may be.

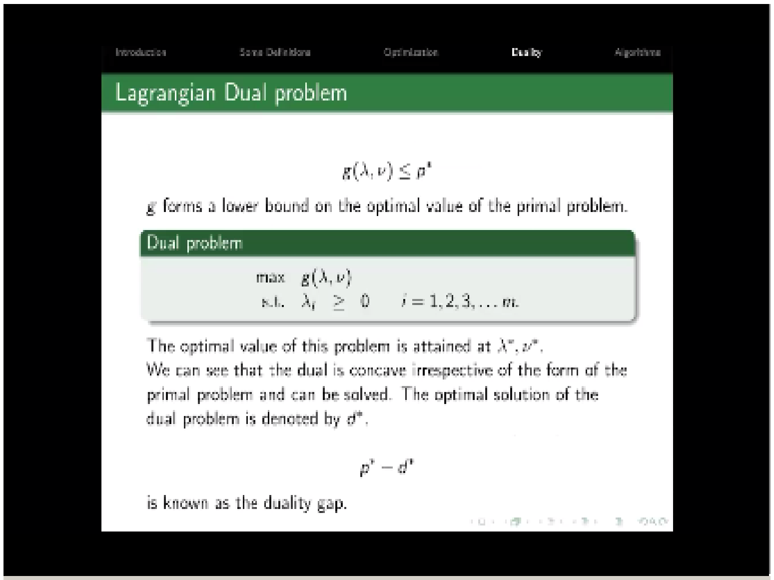
So before going at which is recap the notation which we going to use. So this is the standard optimal convex optimization problem and when I said P* it denotes that the optimal value of this problem and the value of P* is attained at X* which is the solution of a solution. Now let us consider the alternative relax problem. Instruct the minimizing the f_0 will may the weighted some of the objective functions and the constraints.

(Refer Slide Time: 18:19)



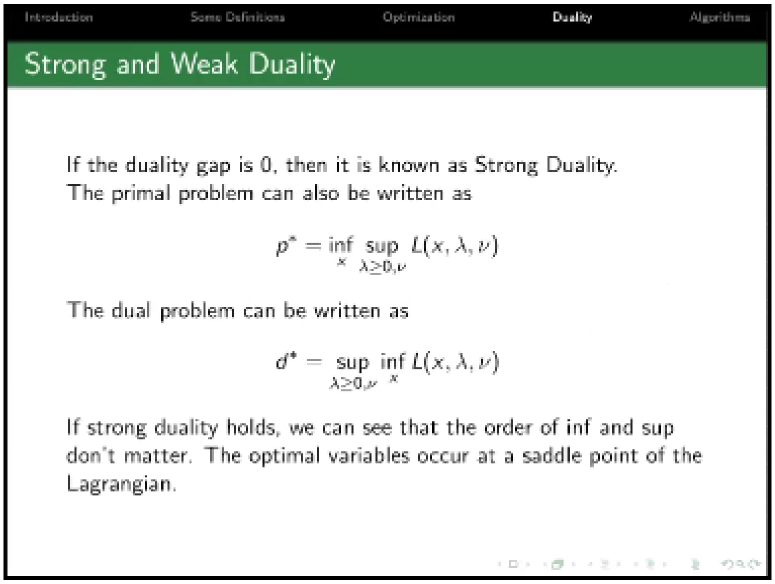
So will minimizing $f_0(x+\sum \lambda f_i +\sum \mu^i h^i)$ here we also have an addition constraint that λ should be greater than or equal to 0. And as usual x should belong to the remain we call the object of this optimization of this optimization problem as the Legrangine so L is a function of x λμ is defined as $f_0+\sum\lambda f_i + \sum\mu_i h_{i.}$ Infimum of the Legrangine over x is less than equal to P* this can be seen very usually, but think to be noted as in equality is valid only one x is feasible. So now to find she as a function of λμ as the infimum of the Legrangine over x.

(Refer Slide Time: 19:14)



So we have seen of a function g which cases lower bound of the optimal value of the primal problem. So if you try to maximize the function g will achieve a very good lower bound of the optimal value. So this is what is may known as the dual problem, so maximizing g of λ column μ such that λ<= 0. The optimal value of the problem is attain it λ* and μ*, we can see that this function g is conquer is respect of the form of the primal problem. So if you go back and see we started with the general form of primal problem and we achieve, when reached with g which is conquer. So g can always be solved the optimum value of the dual problem is denoted by d* so now we would like to see how far is this d* from the actual value p* so p* - d* is know d / d^.

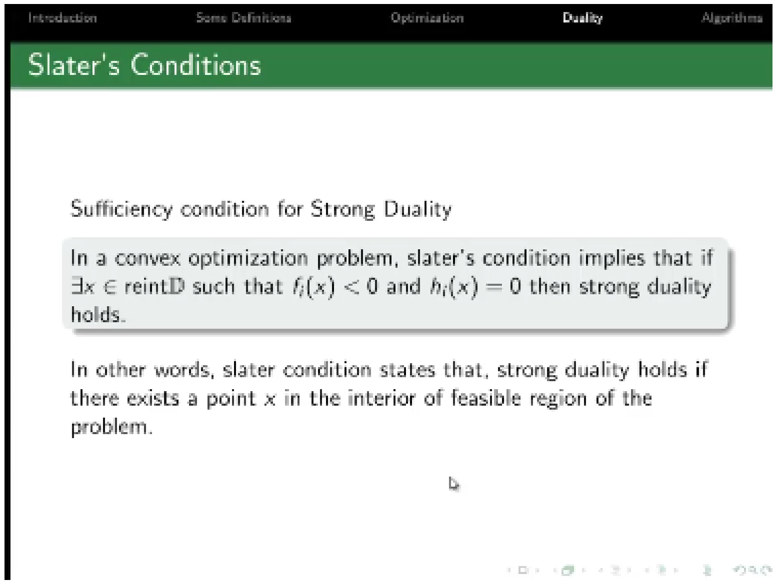
(Refer Slide Time: 20:25)



The next obvious question is to find out when this p* -t* will be 0 and when it is back so whenever it is 0 it is known as the strong duality and when it is not it is known as the weak duality, so next we will try to further characterize when what can occur so first decide we can see that d* can be written as in few know verse of L and d* can be written as supreme over of L or appreciate variables, so when this strong duality holds we know that p* = d* so you can see that the order of the in human supreme can be interchanged and it is equivalent.

So this means that at the same point we have maxima in one direction and the minima in another direction so it is a saddle point, so we have one good result here that is whenever that is strong duality optimal variables occur at the saddle point of the Lagrange.

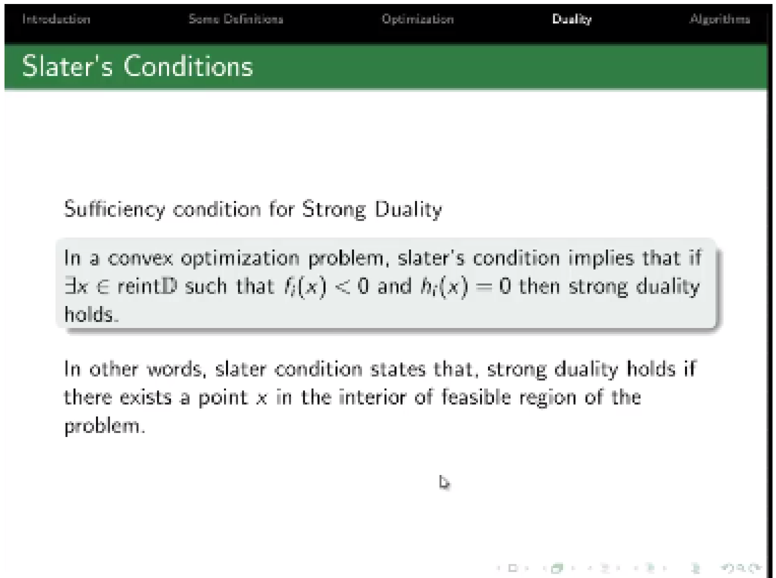
(Refer Slide Time: 21:34)



Now let us look at sufficiency conditions for strong duality so we look at slates conditions which gives us conditions for a convex optimization problem to be strongly dual so Slater's conditions states that for a convex implies volume if the existence at such that it belongs it the relative interior of the domain such that fi(x) is less than 0 and hi(x) = 0 then strong duality so here we require the inequality constrains to be strongly strictly unequal and the function should be the point should belong to the relative integral and not the boundary.

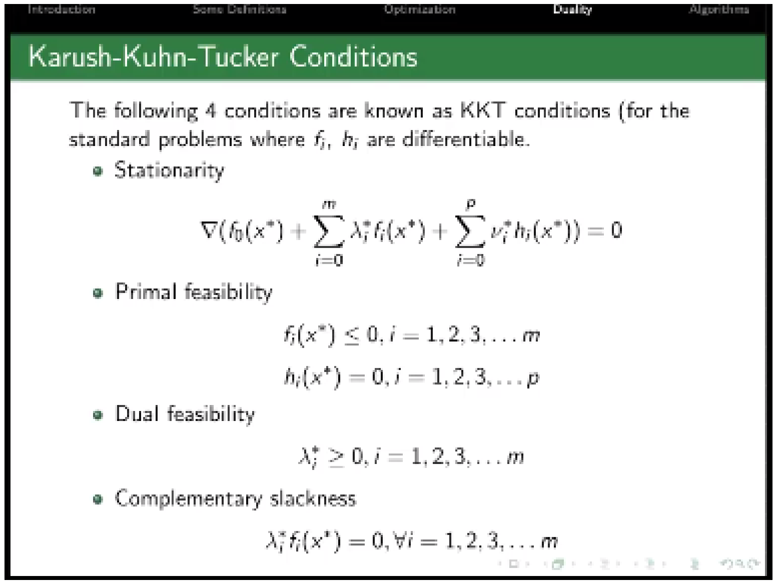
So Slater's conditions state that for any convex optimization problem if that exits a point inside the feasible region then strong duality surly holds so note that this is only for convex of and not a general result.

(Refer Slide Time: 10:22)



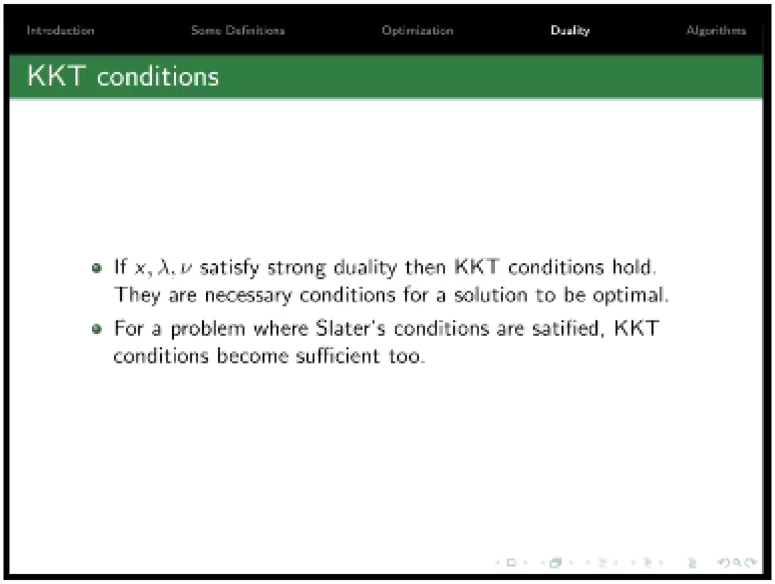
So now we look at complementary slackness assume strong duality holds and x* is the primal variable and λ* also dual variables so when I say strong duality holds we know that f_0 at x* - g at λ* and f* okay so by expanding gr by it is definition and looking at some simple inequities we can reach to a conclusion that for all I λi * fi x* should be 0 okay so basically we know that fi(x*) is ≤ 0 because x* is a feasible value so whenever fi(x*) is not equal to 0 we know that λ it should be = 0, so this is known as complementary slackness that is either λ is x* = 0 or fi(x*) should be = 0 then strong duality holds.

(Refer Slide Time: 23:46)



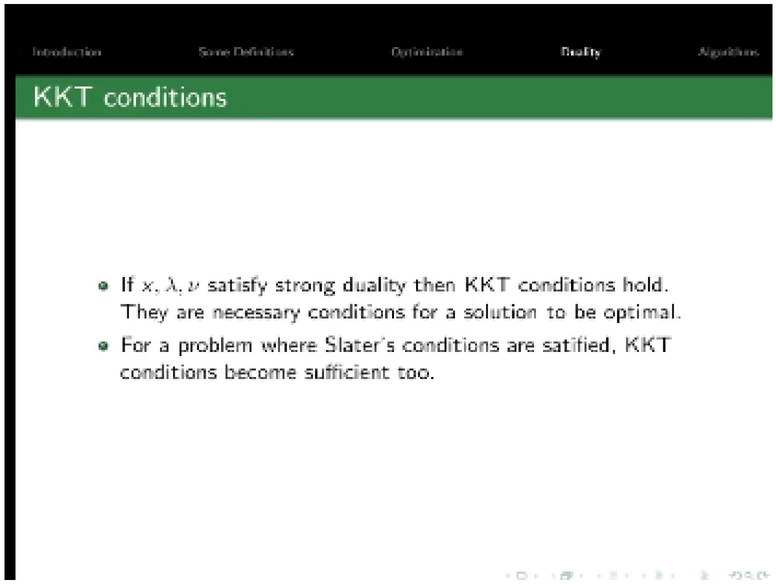
Now we will look at Karush Kuhn Tucker conditions also known as KKT conditions so these provide us the necessary conditions for a point x* λ* μ* to be optimal so consider any point x* λ* μ* if it has to be a optimization these things have to be satisfied so first stationary so since you already seen that at the optimal point l as saddle point so the gradient at that point should be 0 so that is prevail to C n primal feasibility and dual feasibility should hold that is also of a vies and then you have seen complementary Slater's as you seen previously we also be valid at this point.

(Refer Slide Time: 24:39)



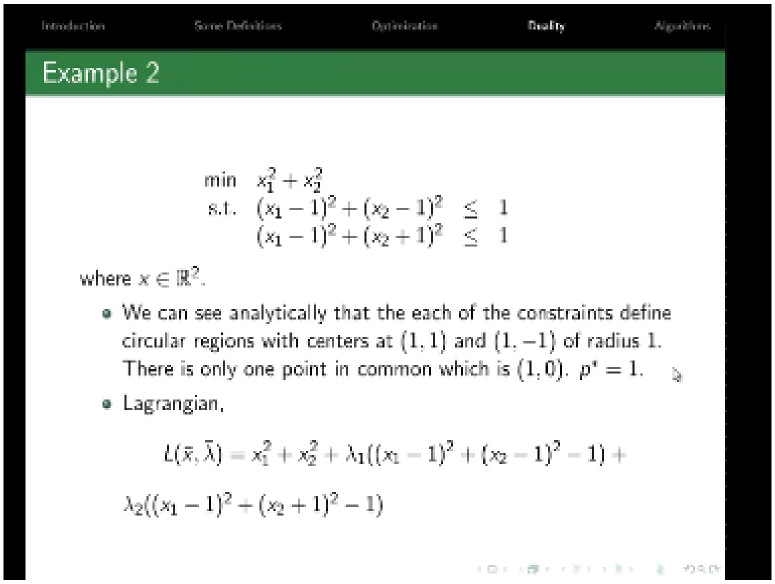
So just to reiterate what you already seen if x, λ, ν satisfy strong duality then KKT conditions hold so these are just necessary conditions and sufficient but further optimization problems when Slater's conditions are satisfied then KKT became sufficient also.

(Refer Slide Time: 25:03)



Now we locate some examples first is the most popular example of least squares, so we are trying to minimize a least square function so we are trying to minimize a least square function so we are trying to minimize this two norm of x – v with new constraints so we can clearly say that this a convex function and there is no constraints and we solved in this thing in while solving linear regression to give x* as A^T A inverse A^T b, so this is a very tribal convex or machine problem which you are able to solve I, but just by differentiating.

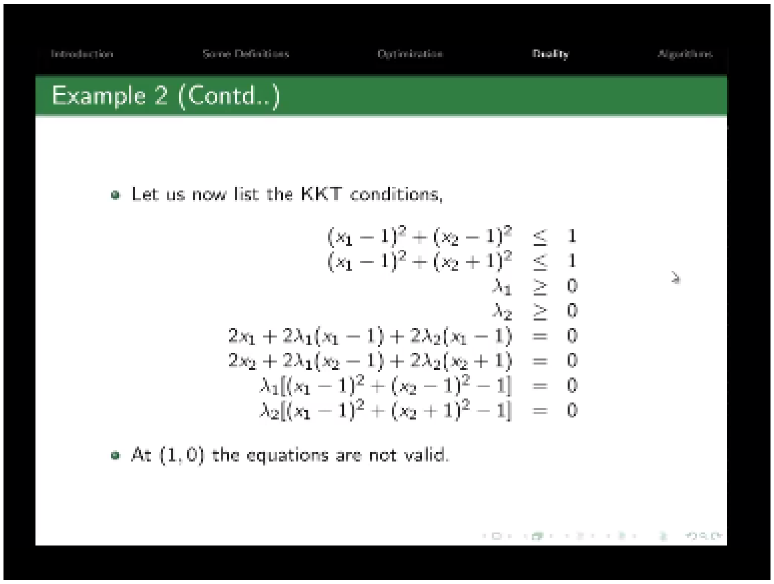
(Refer Slide Time: 25:46)



Now let us look at another example, so here we are trying to minimize $x1^2 + x2^2$ subject to two these two linear quadratic constraints so you look at these constraints carefully both of them are circular regions one centered at (1, 1) and the other centered at (1, -1) each of which is radius at one, so if you just plot them and see that you can see that there is only one feasible point that is (1, 0) so trivially the optimal value will become one.

But now let us do analysis which we have learnt and how to do and then try to analyze in this answer, so first when you have a convex of machine value or for that matter any of machine balance like you have the first thing you do is write that like Lagrangian so here the lagrngian will be $x1^2 + x2^2 + \lambda 1$ times a first constraints next $\lambda 2$ time the second constraint.

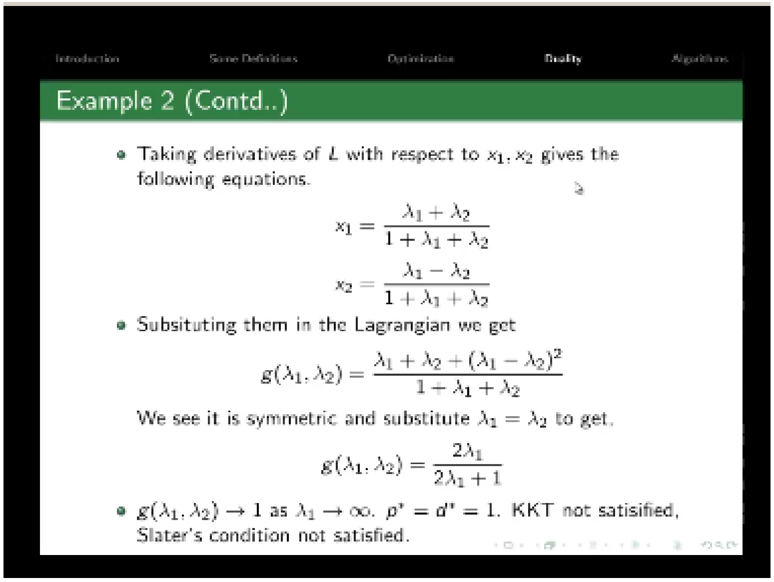
(Refer Slide Time: 26:55)



So now that you see in lagrangian let us try to list out the KKT conditions so here the first two are the primary feasibility conditions second to are the dual feasibility conditions and the next two are obtained by differentiating the lagrangian with x1 and x2 respectively and the next one are obtain by writing the complimentary strategy equations. And we have seen that there is only one feasible point (1, 0) and at that point b these conditions are not valid with you get contradictory answers for λ1 and λ2 and you try to solve.

See that is KKT conditions are not valid but this is tricky so we have already seen that we have an optimal value but KKT conditions are not satisfied we will try to see why this is happening here. Now let us try to investigate what exactly is happening so we will try to solve that your problem now.

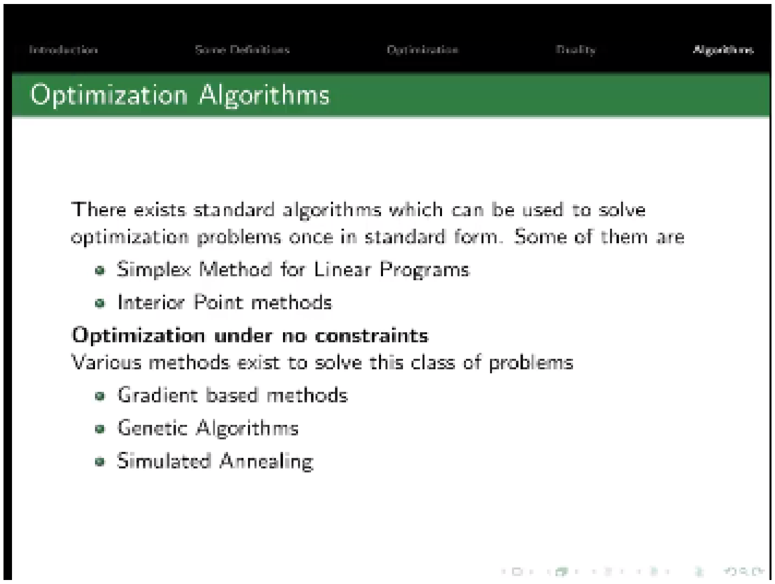
(Refer Slide Time: 28:10)



So for solving the dual problem we have find the maximum of a G, so first lt is find out what the function G is, G is in few form of the lagrangian over x so we will substitute we will try to take the derivative of L with respect to x solve it and then if we arrive at this G function which is the function of λ 1 and λ2 now you can see that this is a concave function which is symmetric λ1 and λ2 so we can substitute this λ1 =λ2= λ1 and the go ahead, so when we do that we get this 2λ1 / 2 λ1 + 1 as that g function. So if you see that under the limit λ1 10 into ∞ g turns to 1 but otherwise there is no maximal sheet.

So under a simple conditions p* = d* = 1 and because this is these points are not been attained at point KKT conditions latest conditions are not satisfied, so this example just to show you that just solving KKT conditions are checking first latest condition is not sufficient we might have to solve sometimes the dual problem and see what exactly is happening.

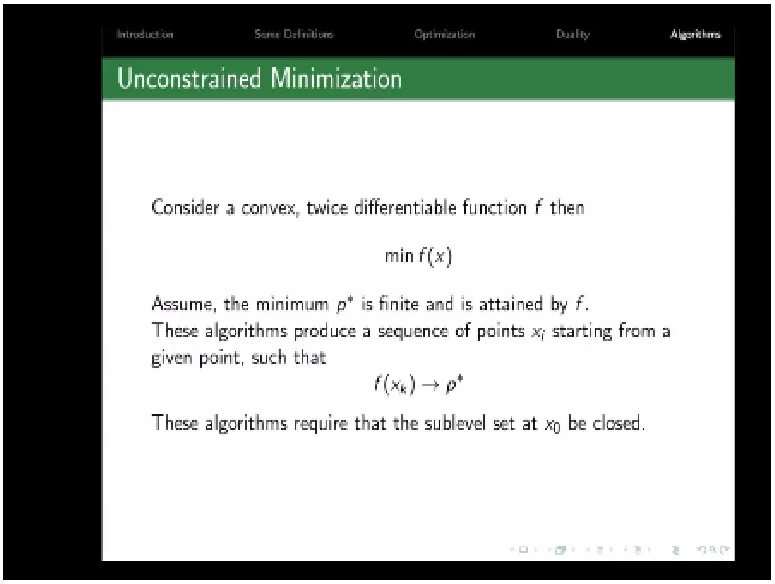
(Refer Slide Time: 29:28)



So do you see the mathematical characterization for optimization problems, now we will try to see how to solve them so there exists very many standard algorithms to solve optimization problems once you taken them to standard form, so for linear programs there is this feel known simplex method and the most popular methods for solving general optimization problems right now are interior point methods, will not be covering in these methods in detail at all will be looking at simpler class problems that is, optimization under no constraints, so that is given an objective function under no constraints, how can we solve this?

We will look at algorithms, so to do this there exist a lot of algorithms; gradient based methods, genetic algorithms and simulated annealing. First we will look at gradient based methods which are very popular used in machine learning.

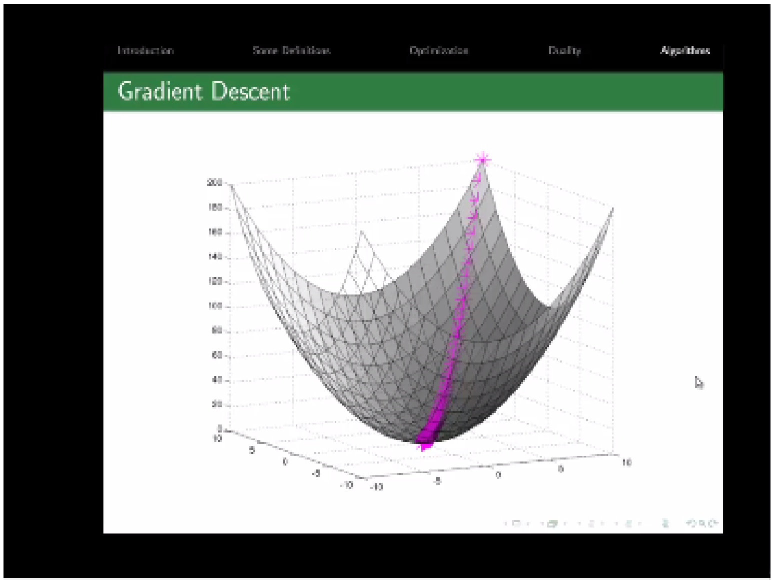
(Refer Slide Time: 30:35)



So first let us look at proper mathematical definition of unconstrained minimization. Consider a convex which twice differentiable function is and we want to find minimum at of this function, so assume there is minimum and it is finite and it is attained by a, so we want algorithms, start from one point and give a series of exercise, such that value of f(xk) tends to this optimal minimum. So these algorithms required one condition that is the sublevel set should be closed.

So what exactly this condition means is, so when I start from x0 and I go to some other point is which is < then so basically each time I am trying to reduce the value of f, so x1,x0 to x1 where f(x) 1 < then f(x) 0. So that is, this belongs to the subset of f at f(x) 0 and this point x 1 should be inside the set, so we just need this condition, so that we get a chain of points, which are in the domain of the function.

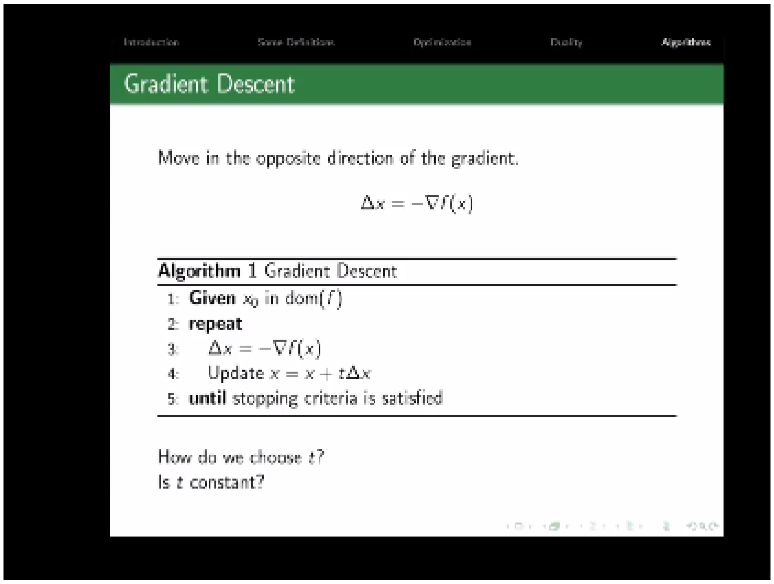
(Refer Slide Time: 32:03)



So now we will look at what are the most popular algorithms gradient descent, so this works in convex problems where there exist in minimum and you start from one point from the top and go down according to the gradient, so if you see this visualization gives you the 3 dimensional surface, which is basically f(fx), x2 say. So if we start of f(x, x2) at the top point and we take the gradient there and move along the negative direction slowly.

As we keep going down we reach the bottom of this, so and the bottom is where the minima exist, at the last point the gradient become 0. So this is the motivation for gradient descent algorithms that is by going along the negative direction of the gradient, we reach the minima in convex functions.

(Refer Slide Time: 33:09)



So let us formally look at the gradient descent, we intuitively seeing that if you move in the direction of the gradient we will reach the minima, so will stay at the algorithm,. So if we start x0 in the domain of f, you can update in every iteration x as x= x+ Δ- f(x). so essentially what we are doing is, we are moving along the negative direction of the function. in some step size of t, basically this t is the multiplication factor, which will magnify or minimize steps that you are taking in the direction.

So the next question is how do we choose t? Should t be constant, so in the ideal case t should be depended on the curvature of the functions? So if you look at the graph in the previous slide carefully, so where ever there is low curvature you could afford to take larger steps, where ever there is high curvature at the bottom especially where there are minima, you should take small steps, so you do not jump over the minima value.

Methods which choose t according this are out of the scope of this tutorial, so but we will just answer this question, is t constant is enough for us. In most conditions a small t if you take a small enough step size it is find and you will very reasonably very close to the minima. So in practice the constant t works. So we will end this tutorial session with this. So the main take home of this tutorial session should be, what accomation problems are? What is the generous form? What are convex formation problems? What is duality? What is strong duality?

So knowing these will be enough for you to navigate, whatever the optimization that come across this course but ideally we can look up other resource o line if you are not clear with these basic still.

IIT Madras Production

Funded by
Department of Higher Education
Ministry of Human Resource Development
Government of India

www.nptel.ac.in