UNIVERSITY OF LEEDS

Class: Machine Learning

Support Vector Machines – part 3

Instructor: Matteo Leonetti

Learning outcomes

UNIVERSITY OF LEEDS

- Derive the dual formulation of a constrained optimisation problem

Substitution

UNIVERSITY OF LEEDS

minimise: $\frac{1}{2}\|\boldsymbol{w}\|^2$

Subject to the constraints: $t_i(\boldsymbol{w}^T\Phi(\boldsymbol{x}_i)+w_0)\geq 1$

This way we would have a higher dimensional problem, which is also more difficult to solve.

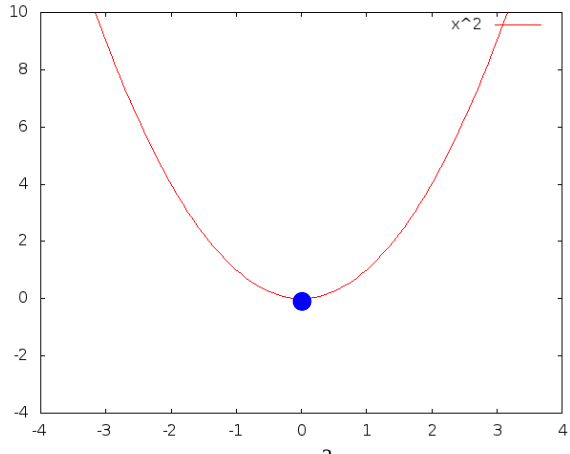
Is there a better formulation?

We are now on a quest to find a different formulation of the SVM problem, that allows us to use a high-dimensional decision boundary without having to solve a high-dimensional problem.

Duality Theory

Example - unconstrained

UNIVERSITY OF LEEDS



$$\min x^2$$

$$\nabla_x f(x) = 2x = 0 \quad \Rightarrow x = 0$$

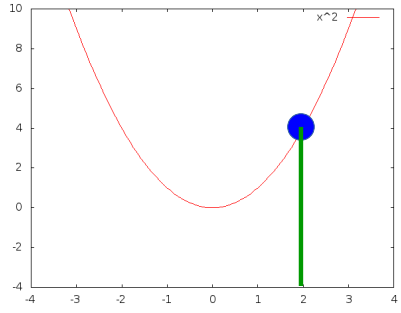
We know that, in unconstrained optimisation, for a point to be stationary (that is, either a minimum, maximum, or saddle), the derivative has to be zero on that point.

Therefore, solving for the derivative equal to zero gives the stationary points.

In the simple example above, the function x^2 has a single minimum in x=0.

Example - constrained

UNIVERSITY OF LEEDS



$$\min\ x^2$$

$$\text{s.t.}\quad x=2$$

What if we added a constraint? A constraint is a condition over the input of a function, which determines which points are *feasible*, that is, can be valid solutions.

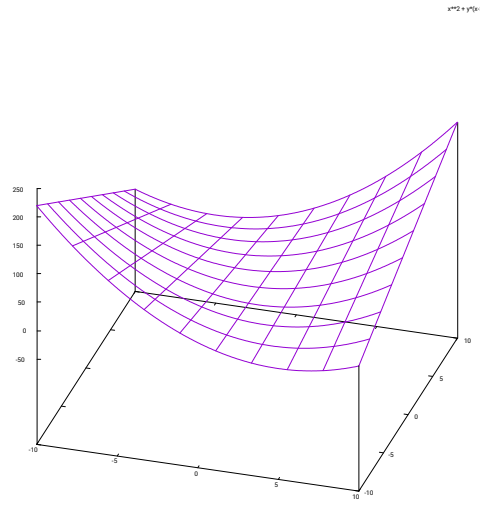
Let's say that we have the constraint x=2. In 1D unfortunately an equality constraint is a single point, which doesn't leave much space to optimisation: since there is a single feasible point, that is also the minimum.

In more dimensions equality constraints force the solution to be on a particular line or surface. For instance, if the input of a function is in 2D, the unconstrained input is the whole real plane. However, a constraint might reduce the feasibility region to a circle (or any other line).

Anyway, you can see that at the constrained optimum the gradient is not zero, so we need to find a different condition to identify this point.

The Lagrangian

UNIVERSITY OF LEEDS



$$\min_x \max_\lambda \; x^2+\lambda(x-2)$$

$$\begin{cases} \nabla_x f(x,\lambda)=2x+\lambda=0 \\ \nabla_\lambda f(x,\lambda)=x-2=0 \end{cases}$$

$$\begin{aligned} x&=2 \\ \lambda&=-4 \end{aligned}$$

$$L(x,\lambda)=x^2+\lambda(x-2)$$

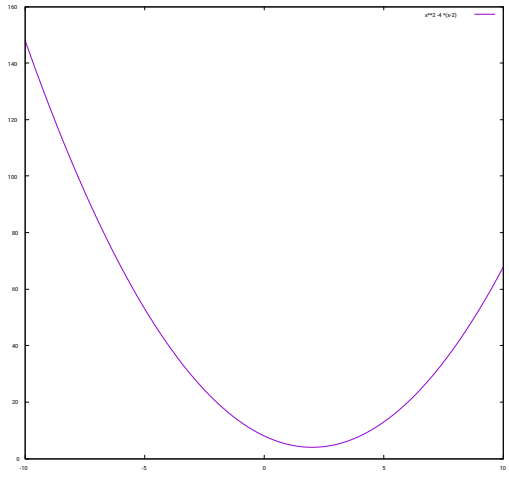
It is possible to incorporate the constraint into a new function, by adding one variable per constraint. The function is called Lagrangian, and the variables are called Lagrangian multipliers.

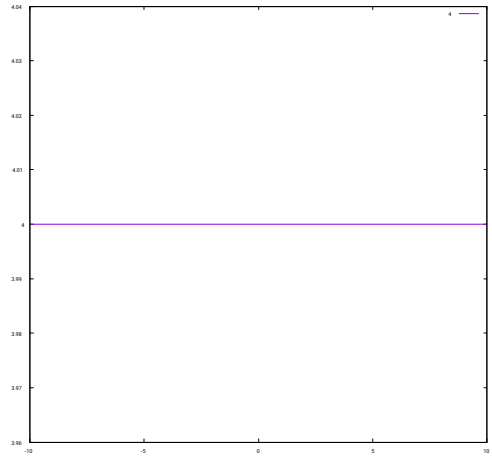
The new function has more dimensions than the original one, but has an interesting property: the saddle point (minimum with respect to the original variables, and maximum with respect to the Lagrange multipliers) of the Lagrangian is also a minimum for the original, constrained, problem.

Therefore, by optimising the Lagrangian we can solve an unconstrained problem whose solution is the same as the solution of the original constrained problem.

The Lagrangian - sliced

UNIVERSITY OF LEEDS





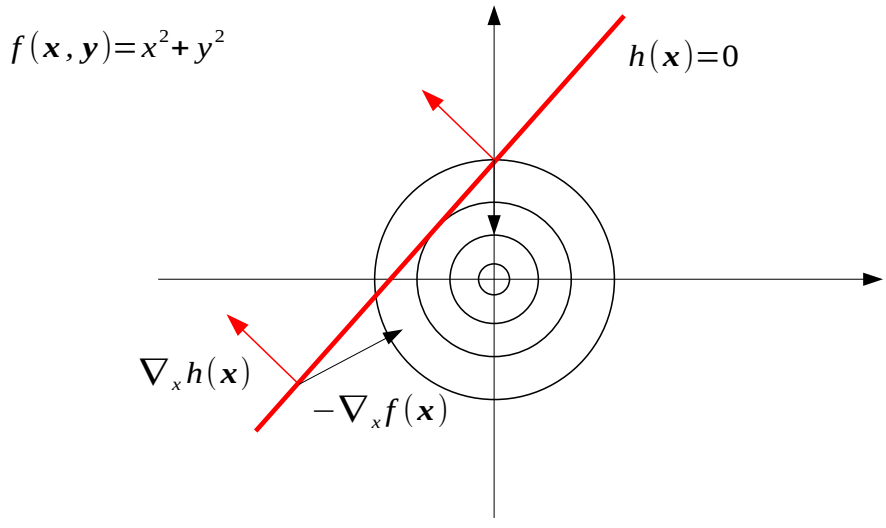
$$L(x,-4)=x^2-4(x-2)$$

$$L(2,\lambda)=4$$

This is a view of the Lagrangian sliced along x and λ. It is clear the minimum along x, while along λ the function is constant at x=2 so it also a maximum.

Lagrange Multipliers

UNIVERSITY OF LEEDS



By looking at the gradients, can you tell when a point is a local minimum for the constrained problem?

To understand the role of the first part of the gradient, we need to look at an example with more dimensions. This is a function of two variables and its unconstrained minimum is in <0,0>.

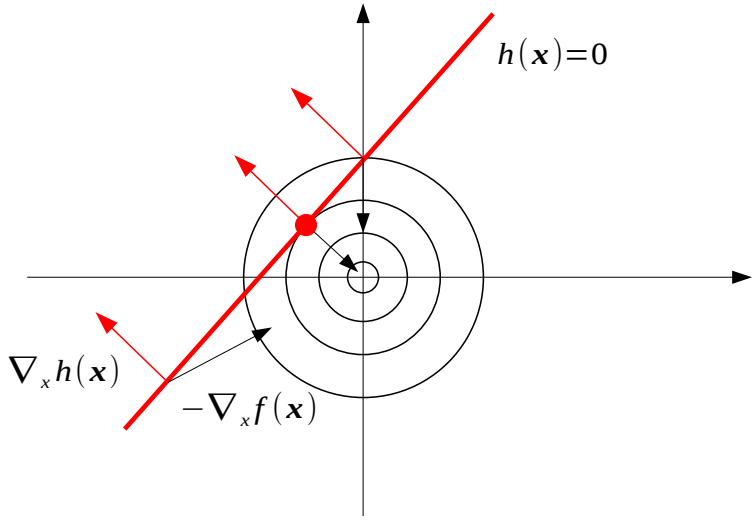
I am plotting the x,y plane, and the function lives in a third dimension above this. The circles are *contour* lines, that is, lines over which the function has the same value. So this function is a parabola in x and y, and grows from the origin towards the outer values of the axes.

Again, if we were solving the unconstrained problem it would be easy, because there is a single minimum in <0,0>. However, we are solving a constrained problem, where we impose the solution to be on some line with equation h(x) = 0.

What point one the line corresponds to the minimum value of the function?

Lagrange Multipliers

UNIVERSITY OF LEEDS



When the gradients are parallel! $-\nabla_x f(\boldsymbol{x})=\lambda \nabla_x h(\boldsymbol{x})$

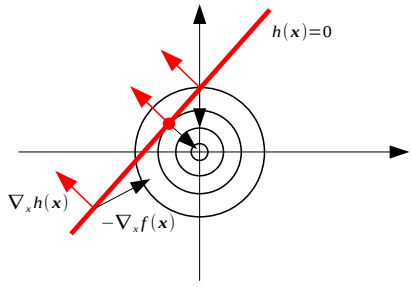
For a point on the line to be a local minimum, it must not be possible to move from that point along the line and improve the function. Therefore, the antigradient (remember we are minimising…) of the function must be **orthogonal** to the line.

Since the gradient of a surface is always orthogonal to it, the gradient of the constraints and the gradient of the objective function must be **parallel**.

This condition is expressed by the equation at the bottom, which says that the vector of the antigradient and the gradient of the constraint only differ in a multiplication by a constant. Therefore they are parallel.

Lagrange Multipliers

UNIVERSITY OF LEEDS



When the gradients are parallel! $-\nabla_x f(\boldsymbol{x})=\lambda\nabla_x h(\boldsymbol{x})$

This is achieved by: $\nabla_x L(\boldsymbol{x},\boldsymbol{\lambda})=0$

since:

$$\nabla_x L(\boldsymbol{x},\boldsymbol{\lambda})=\nabla_x f(\boldsymbol{x})+\lambda\nabla_x h(\boldsymbol{x})=0$$

This condition of parallelism between the antigradient of the objective function and the gradient of the constraint is indeed enforced by the first part of the gradient of the Lagrangian. If we derive the Lagrangian with respect to x, we obtain the equation imposing our parallelism.

Lagrange multipliers

UNIVERSITY OF LEEDS

$$\min \quad f(\boldsymbol{x})$$

Subject to

$$h_i(\boldsymbol{x}) = 0 \quad \forall\, i = 1, \ldots, m$$

It is possible to form a function such that its stationary points are optimal solutions to the original problem:

$$L(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i h_i(\boldsymbol{x})$$

$\nabla_x L(\boldsymbol{x}, \boldsymbol{\lambda}) = \nabla_x f(\boldsymbol{x}) + \lambda \nabla_x h(\boldsymbol{x}) = 0$ Ensures that the gradients are parallel

$\nabla_\lambda L(\boldsymbol{x}, \boldsymbol{\lambda}) = h(\boldsymbol{x}) = 0$ Ensures that the solution satisfies the constraints

To summarise, the Lagrangian is a very special function.

It is a function of the original variables + the Lagrange multipliers. We add one multiplier per constraint.

Because of how the function is built, when we solve for the gradient to be zero, we are implicitly enforcing two things: the gradient of the original objective function is parallel to the gradient of the constraint, and the solution must satisfy every constraint.

This implies that a stationary point (that is, where the gradient is zero) for the Lagrangian is also a solution of the original constrained problem.

Awesome! What do we do with this beautiful result?

Question

UNIVERSITY OF LEEDS

How many variables lambda did I add?

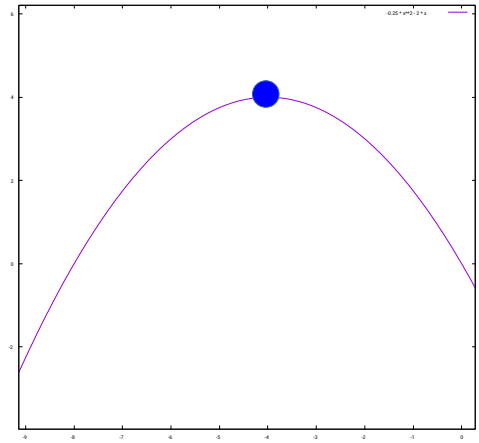
One per constraint of the original problem!

The Dual Problem

UNIVERSITY OF LEEDS

$$L(x,\lambda)=x^2+\lambda(x-2)$$

$$\nabla_x f(x,\lambda)=2x+\lambda=0 \qquad x=-\frac{1}{2}\lambda$$



Substitute x:

$$q(\lambda)=\left(-\frac{1}{2}\lambda\right)^2+\lambda\left(-\frac{1}{2}\lambda-2\right)=\frac{1}{4}\lambda^2-\frac{1}{2}\lambda^2-2\lambda$$

$$=-\frac{1}{4}\lambda^2-2\lambda$$

$$\nabla_\lambda q=-\frac{1}{2}\lambda-2=0 \qquad \lambda=-4$$

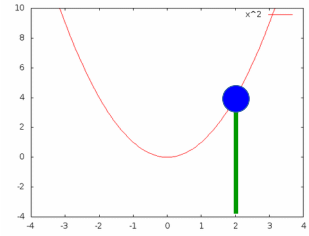
The Lagrangian can be used to derive a new function of the sole Lagrangian multipliers (in this example, λ), with a very special property: maximsing this new function yields the same solution as minimising the original constrained problem.

This new function is called the *dual* formulation of the original problem, which is called the *primal* problem. The dual formulation is, in practice, another way to look at the same problem, since the solution of both the dual and primal problems are the same.

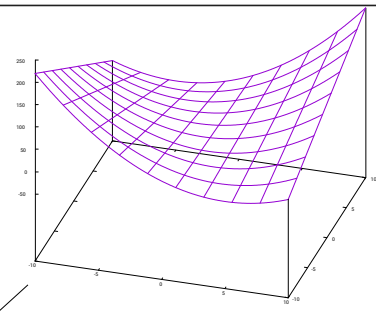
The dual formulation, however, may have advantages over the primal one, as we will see is the case for SVMs.

The Dual Problem

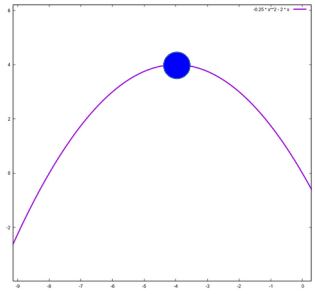
UNIVERSITY OF LEEDS



$$\min f(x)=x^2 \quad \text{s.t.} \quad x=2$$



$$L(x,\lambda)=x^2+\lambda(x-2)$$



$$\max q(\lambda)=-\frac{1}{4}\lambda^2-2\lambda$$

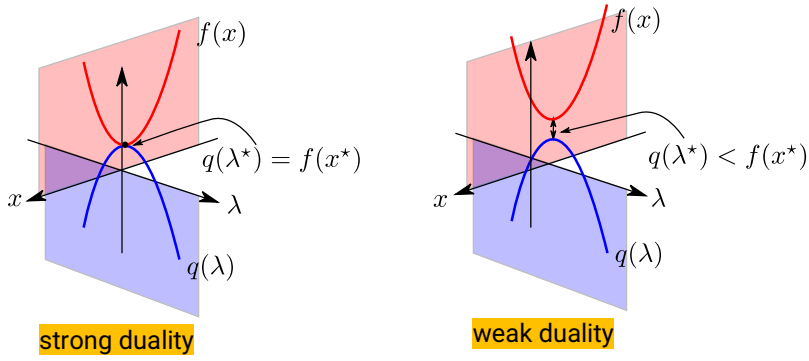
$$f(2)=q(-4)=4$$

So to recap, given a constrained optimization problem, we can compile the constrains into a new function, by adding one additional variable per constraint. This new function has a saddle for the same values of x where the original constrained problem has a minimum.

We can then solve for the optimal x, that is where the derivative with respect to x is zero, and obtain an expression of x in terms of λ, which is valid when x is optimal. We substitute this back into the Lagrangian, and obtain a function of the Lagrange multipliers only: the objective function of the dual formulation of the original problem.

Duality

UNIVERSITY OF LEEDS

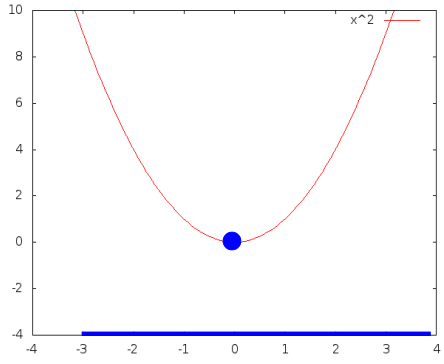


This is a representation of duality. The primal problem, a function of x, lives in a different space with respect to the dual problem, which is a function of λ. However, their solutions (a minimum in the first case and a maximum in the second) coincide.

For this reason, the two problems are equivalent, and solving one also gives a solution of the other.

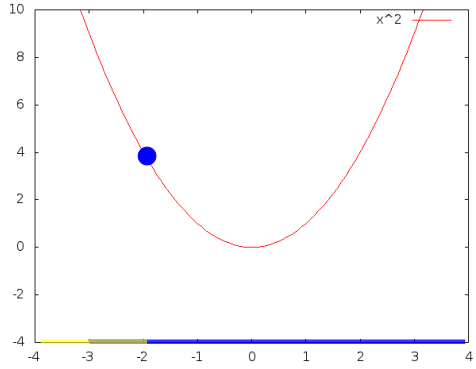
Inequality Constraints

UNIVERSITY OF LEEDS



$$\min f(x)=x^2$$
$$\text{s.t. } x\geq -3$$

Minimum inside the constraint



$$\min f(x)=x^2$$
$$\text{s.t. } x\geq -3$$
$$x\leq -2$$

Minimum on the border

In the SVM formulation, however, we don't have equality constraints, but inequalities. The main idea stays the same: we can derive an equivalent dual formulation. With inequality constraints, though, we need to do a little more work.

With inequality constraints, the minimum may or may not change as a consequence of the constraints. For instance the first constraint (x>= -3) has the unconstrained minimum in its feasibility region, and therefore the constrained and unconstrained minimum are the same.
However, by adding the constraint (x <= -2) the minimum is not feasible any more (because it's > -2). Therefore, the constrained minimum becomes -2.

The first type of constraint is said to be "inactive", while the second one is "active". Active constraints cause the minimum to change, and are satisfied on the *edge*, that is, at the equality (note how the new minimum is at -2, which is indeed the edge of the constraint).

Does that ring a bell? Think about how support vectors satisfy their constraint at the equality… Support vectors correspond to active constraints, and indeed they are the ones that determine the separating boundary, just like active constraints here determine where the (constrained) minimum is.

Inactive constraints can be ignored, since they do not affect the solution (but we don't know which ones are inactive until we solve the problem…).

Karush–Kuhn–Tucker conditions

UNIVERSITY OF LEEDS

Extend lagrangian multipliers to inequality constraints

$$\min f(\vec{x})$$

Subject to

$$h_i(\vec{x}) \leq 0 \quad \forall i = 1, \ldots, n$$

$$L(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i h_i(\boldsymbol{x})$$

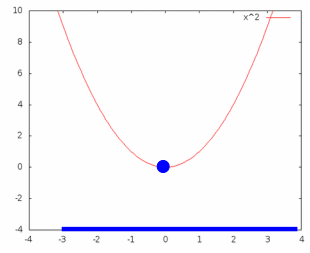
$\nabla_x L(\boldsymbol{x}, \boldsymbol{\lambda}) = 0$ What else?

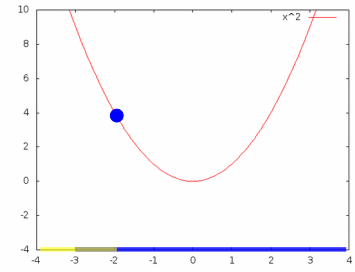
We have seen how the solution is on the edge of an active constraint. Because of that, the condition on the parallelism of the gradients must be retained. However, now we need to distinguish between active and inactive constraints, because the solution is on the edge the active ones only.

Somehow the inactive constraints must disappear from this equation. How do we achieve that?

Complementary Slackness

UNIVERSITY OF LEEDS





$$\min \quad L(\boldsymbol{x},\boldsymbol{\lambda}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i h_i(\boldsymbol{x})$$

$$\text{s.t.} \quad h_i(\boldsymbol{x}) \leq 0$$

For Inactive constraints:

$\lambda_j = 0 \rightarrow \lambda_j h_j(\boldsymbol{x}) = 0$

For active constraints

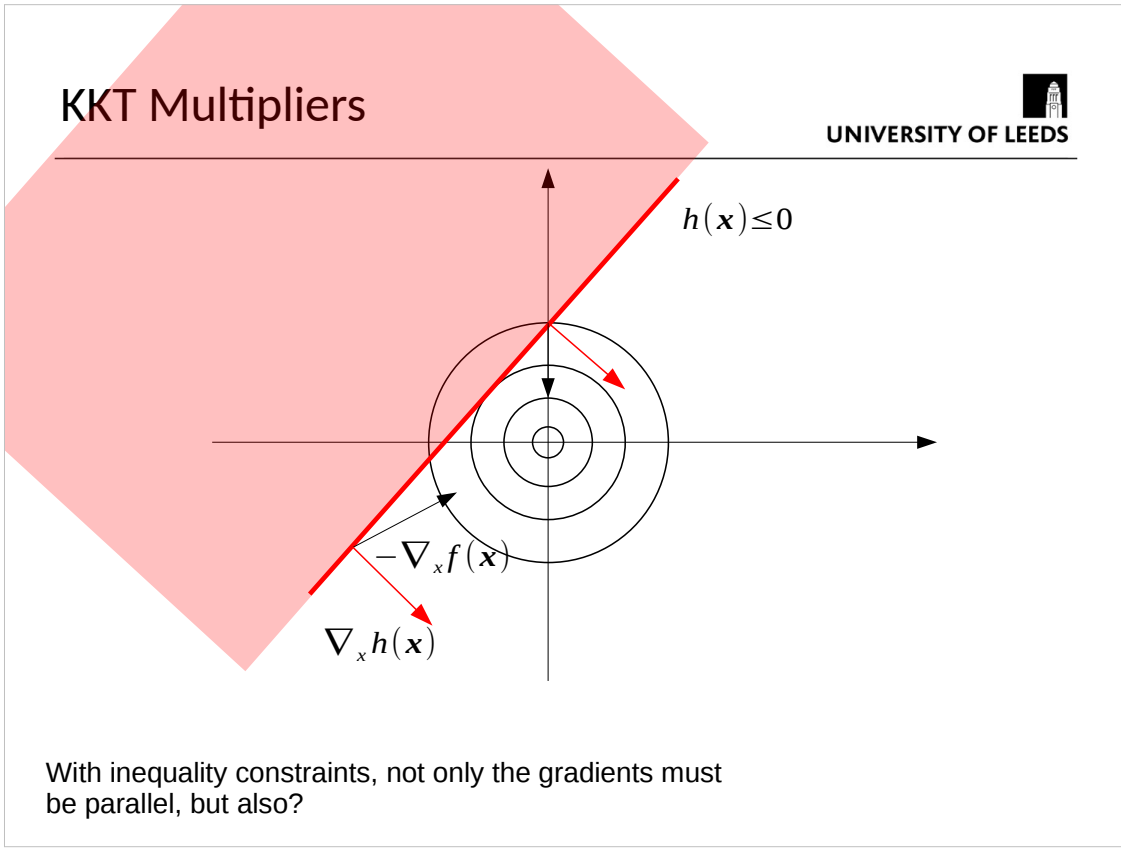
$h_j(\boldsymbol{x}) = 0 \rightarrow \lambda_j h_j(\boldsymbol{x}) = 0$

If we enforce that the Lagrange multiplier times the constraint is zero, then either the constraint or the multiplier must be zero.

This means that if the constraint is inactive, that is h(x) < 0, then the Lagrange multiplier will be forced to be 0, and the inactive constraint will disappear from the Lagrangian.

On the other hand, if the constraint is active, that is h(x) = 0, then the Lagrange multiplier is free to be what it wants to be.

The equation (one per constraint!) achieving this is called *complementary slackness*.



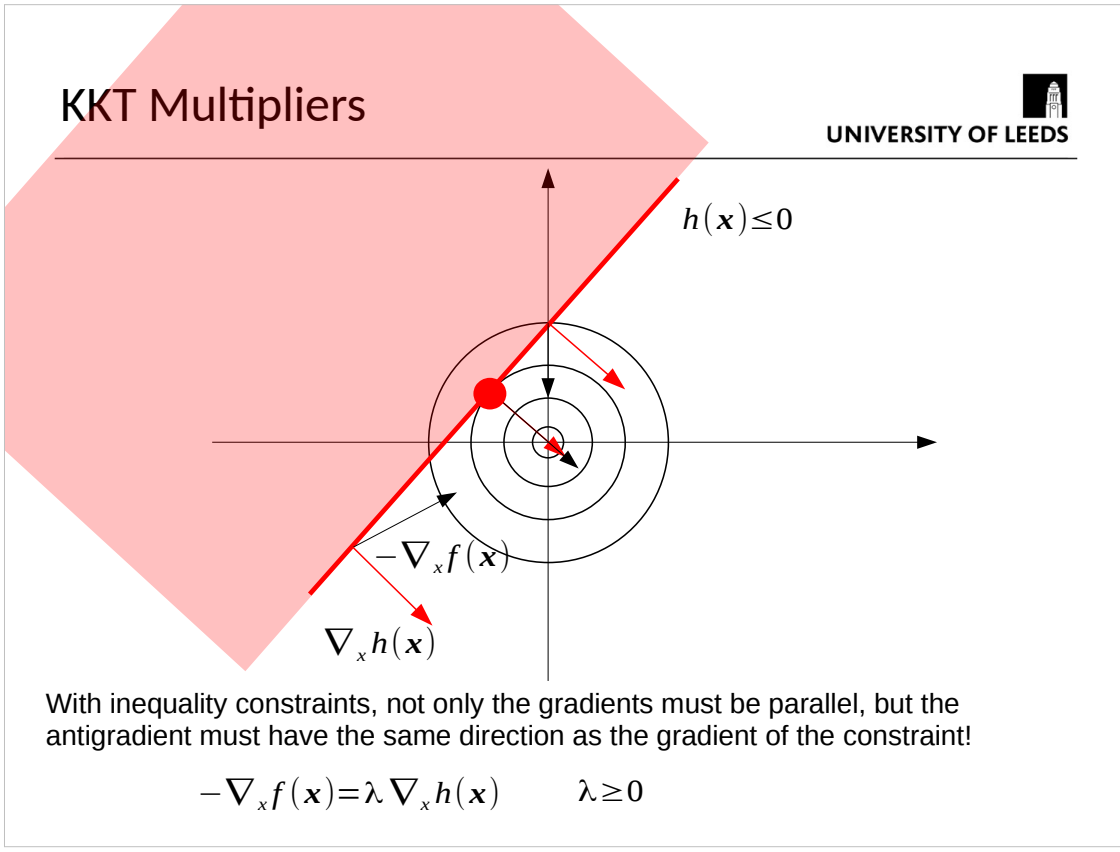
With inequality constraints, the solution must not lie only on the constraint, but can be on the half plane either above or below the edge.

This half we care about matters, because makes the difference between an active and an inactive constraint.

Recall that the gradient of the constraint h(x) points in the direction of growth for the function h(x). Therefore, the gradient always points in the direction of the half-plan where h(x) > 0.

However, our constraints have the form h(x) <=0, which means we want the solution to be in the half-plane opposite to the gradient.

If you look at the example above, you can see that the gradient of the constraint is opposite to the half-plane that satisfies the constraint.



For inequality constraints, therefore, not only we are looking for the point in which the gradient of the function and the gradient of the constraint are parallel.

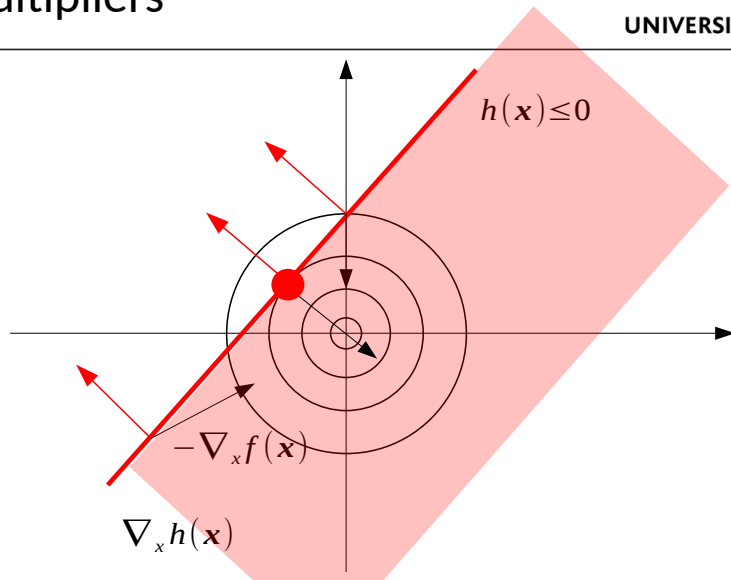
Also, we need the antigradient of the function and the gradient of the constraint to the the **same direction**.

To impose this property, we need to enforce, in addition to the equation on the bottom left about parallelism, that the lambdas are positive. This way, the antigradient and the gradient of the constraint must have the same direction.

This last property we are enforcing is called *dual feasibility* (the reason for this name will become apparent shortly).

KKT Multipliers

UNIVERSITY OF LEEDS



Otherwise, I could improve in the direction of the gradient, while satisfying the constraint

$$-\nabla_x f(\boldsymbol{x}) = \lambda \nabla_x h(\boldsymbol{x}) \qquad \lambda \geq 0$$

This image shows that If the antigradient of the objective function and the constraint have opposite directions, then the constraint is inactive, because we could move along the gradient and still satisfy the constraint.

KKT Conditions

UNIVERSITY OF LEEDS

$\min f(\boldsymbol{x})$

$\min f(\boldsymbol{x})$

Subject to

$h_i(\boldsymbol{x}) = 0 \quad \forall i = 1, \ldots, m$

$\min f(\boldsymbol{x})$

Subject to

$h_i(\boldsymbol{x}) \leq 0 \quad \forall i = 1, \ldots, m$

Corresponding system of equations

$\nabla_x f(\boldsymbol{x}) = 0$

$\nabla_x L(\boldsymbol{x}, \boldsymbol{\lambda}) = 0$

$\nabla_\lambda L(\boldsymbol{x}, \boldsymbol{\lambda}) = 0$

$\nabla_x L(\boldsymbol{x}, \boldsymbol{\lambda}) = 0$

$\lambda_i g_i(\boldsymbol{x}) = 0 \quad \forall i = 1, \ldots, n$

$\lambda_i \geq 0 \quad \forall i = 1, \ldots, n$

To summarise, if the problem is unconstrained, the gradient being zero is a necessary condition for a point to be a minimum.

If the problem has only equality constraints, the gradient **of the Lagrangian** being zero is a necessary condition for the point to be a minimum of the constrained problem. These equations (the two parts of the gradient being zero) are called the *Lagrangian conditions*.

If the problem has only inequality constraints, the following properties are necessary conditions: the derivative with respect to the original variable is zero (the vectors are parallel) + complementary slackness (either the constraint is active, or it disappears) + dual feasibility (the multipliers have to be positive, so that the gradients of the objective and the active constraints have the same direction). These are called the KKT conditions, from the names of Karush, Khun, and Tucker.

Example

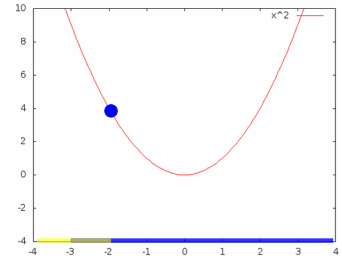
UNIVERSITY OF LEEDS

Lagrangian: $L(x, \boldsymbol{\lambda}) = x^2 + \lambda_1(-x-3) + \lambda_2(x+2)$

s.t. $\lambda_1, \lambda_2 \geq 0$

$\lambda_1(-x-3) = 0$

$\lambda_2(x+2) = 0$



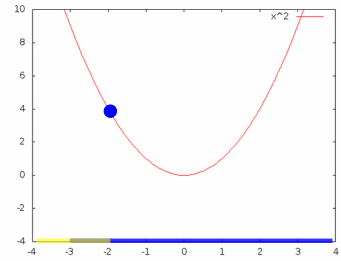
Let's apply the KKT conditions to our example problem, and see what happens.

Stationary point for the Lagrangian

UNIVERSITY OF LEEDS

Lagrangian: $L(x,\boldsymbol{\lambda})=x^2+\lambda_1(-x-3)+\lambda_2(x+2)$

s.t. $\lambda_1,\lambda_2\geq 0$



$$\begin{cases}\nabla_x L(x,\boldsymbol{\lambda})=2x-\lambda_1+\lambda_2=0\\ \lambda_1(-x-3)=0\\ \lambda_2(x+2)=0\end{cases}$$

Let's assume that the first constraint, -x -3, is active and x = -3

$$x=-3$$
$$\downarrow$$
$$\lambda_2(-3+2)=0 \quad \lambda_2=0$$
$$\downarrow$$
$$-6-\lambda_1+0=0 \quad \boxed{\lambda_1=-6}$$

It would violate the constraint on λ_1

So the system of equations that we have to solve to obtain the solution of the original problem is composed of: the derivative of the Lagrangian with respect to x, and the two equations of complementary slackness, subject to a constrain on the positivity of the multipliers.

This system can have more than one solution (it is not a linear system! Indeed the variables appear multiplied by each other) and we can only accept the solutions that satisfy the original constraints and the constraints on the positivity of the multipliers.

Now we proceed by solving the system, and then we will check whether the solution satisfies all the constraints or not.

We look at the last two equations, which allow for a total of 4 options:

1) $\lambda_1 = 0$ & $\lambda_2 = 0$.
2) λ_1 != 0 & $\lambda_2 = 0$.
3) $\lambda_1 = 0$ & λ_2 != 0.
4) λ_1 != 0 & λ_2 != 0.

The first case makes both constraints disappear, and we are back with the unconstrained problem. Obviously the solution of the unconstrained problem does not satisfy -3 <= x <= -2, so this one is not a valid solution.

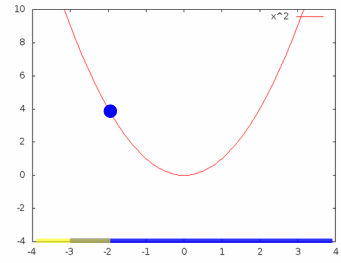
In this slide we look at the second case. If λ_1 != 0 then -x -3 = 0, that is x= -3. We can start substituting into the other equations and see what we get to. We obtain $\lambda_2 = 0$ (which indeed belongs to case 2) and $\lambda_1 = -6$. Unfortunately this one violates the constraint on the positivity of the multipliers, and must be discarded. In the next slide we look at case 3.

Stationary point for the Lagrangian

UNIVERSITY OF LEEDS

Lagrangian: $L(x,\boldsymbol{\lambda})=x^2+\lambda_1(-x-3)+\lambda_2(x+2)$

s.t. $\lambda_1,\lambda_2\geq 0$ $\quad x+2\leq 0$ $\quad -x-3\leq 0$



$$\begin{cases}\nabla_x L(x,\boldsymbol{\lambda})=2x-\lambda_1+\lambda_2=0\\ \lambda_1(-x-3)=0\\ \lambda_2(x+2)=0\end{cases}$$

Let's now assume that the second constraint is active and x = -2

$x=-2$ ↓

$\lambda_1(2-3)=0 \quad \lambda_1=0$ ↓

$-4-0+\lambda_2=0 \quad \lambda_2=4$ OK!

$x=-2$, $\lambda_1=0$, $\lambda_2=4$

This is a stationary point of the Lagrangian AND the solution of the original constrained problem

I'll write again the four cases for reference:

1) $\lambda_1 = 0$ & $\lambda_2 = 0$.
2) λ_1 != 0 & $\lambda_2 = 0$.
3) $\lambda_1 = 0$ & λ_2 != 0.
4) λ_1 != 0 & λ_2 != 0.

We are now looking at case 3. If λ_2 != 0 then x+2 = 0, that is x=-2. Again we substitute this into the other equations and see what happens.

We obtain $\lambda_1 = 0$ which is OK, and indeed was part of case 3, and $\lambda_2 = 4$, which satisfies the positivity constrained. To recap: x=-2 is in the feasible region (-3<=x<=2), and both multipliers are positive. This is a valid solution! Indeed, x=-2 is the solution of the original constrained problem.

Lastly, we need to check case 4. We have actually already done that, because if both multipliers are non-zero then we have that both -x -3 = 0 and x+2 = 0 which has no solutions.

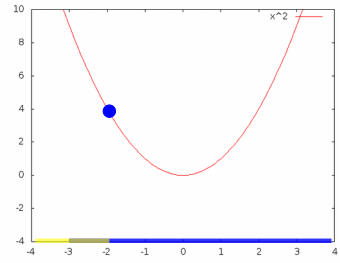
So the only minimum of the original problem is, indeed, at x=-2.

Dual Problem

UNIVERSITY OF LEEDS

Lagrangian: $L(x,\boldsymbol{\lambda}) = x^2 + \lambda_1(-x-3) + \lambda_2(x+2)$

s.t. $\lambda_1, \lambda_2 \geq 0$



Let's build the dual formulation!

$$\nabla_x L(x,\boldsymbol{\lambda}) = 2x - \lambda_1 + \lambda_2 = 0 \qquad x = \frac{\lambda_1 - \lambda_2}{2}$$

$$q(\boldsymbol{\lambda}) = \left(\frac{\lambda_1 - \lambda_2}{2}\right)^2 + \lambda_1\left(-\frac{\lambda_1 - \lambda_2}{2} - 3\right) + \lambda_2\left(\frac{\lambda_1 - \lambda_2}{2} + 2\right)$$

$$= \frac{1}{4}(\lambda_1^2 + \lambda_2^2 - 2\lambda_1\lambda_2) - \frac{1}{2}\lambda_1^2 + \frac{1}{2}\lambda_1\lambda_2 - 3\lambda_1 - \frac{1}{2}\lambda_2^2 + \frac{1}{2}\lambda_1\lambda_2 + 2\lambda_2$$

$$= -\frac{1}{4}\lambda_1^2 - \frac{1}{4}\lambda_2^2 - 3\lambda_1 + 2\lambda_2 + \frac{1}{2}\lambda_1\lambda_2$$

Now we look at the dual problem of our example.

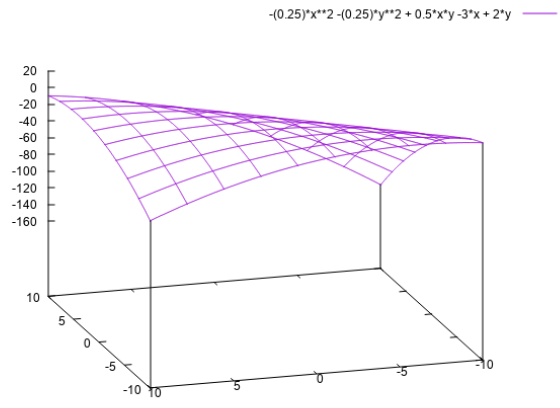
We start by deriving with respect to the original variables x, and solving for x. We get a function of only lambdas.

Dual Problem

UNIVERSITY OF LEEDS

$$q(\boldsymbol{\lambda}) = -\frac{1}{4}\lambda_1^2 - \frac{1}{4}\lambda_2^2 - 3\lambda_1 + 2\lambda_2 + \frac{1}{2}\lambda_1\lambda_2$$

s.t.: $\lambda_1, \lambda_2 \geq 0$



Of all the KKT conditions, the only ones that are solely a function of lambdas are the positivity of the multipliers, and must be retained. So the positivity of the multipliers are the constraints of the dual problem. That's why they are called dual feasibility! They determine the feasible region of the dual problem.

If the original problem, which we will call the *primal* problem, was a minimisation problem, this new one, which we will call the *dual* problem, is a maximisation problem.

These constraints are easier though. The constrained minimum is at $\lambda_1 = 0$ $\lambda_2 = 4$, and the first constraint is inactive.

Question

UNIVERSITY OF LEEDS

Why did we do all this, again?

We can now build the dual problem of a given primal problem, which we are about to do for SVMs. We'll see that the dual problem has an interesting property, which will allow to project our dataset in very high-dimensional spaces at very little cost.

Substitution

UNIVERSITY OF LEEDS

What is the dual formulation of this?

minimise: $\frac{1}{2}\|\boldsymbol{w}\|^2$

Subject to the constraints: $t_i(\boldsymbol{w}^T \Phi(\boldsymbol{x}_i) + w_0) \geq 1$