Optimization and Data Science

Lecture 22: Lagrange Multiplier Rule for Inequality Constraints

Prof. Dr. Thomas Slawig

Kiel University - CAU Kiel
Dep. of Computer Science

Summer 2020

Contents

Lagrange function and Lagrange mulitipliers

Definition

For the problem

$$\min_{x \in \mathbb{R}^n} f(x) \;\text{ subject to } g(x) \leq 0, h(x) = 0, \quad \text{where } g : \mathbb{R}^n \to \mathbb{R}^m, h : \mathbb{R}^n \to \mathbb{R}^p,$$

we call the function $L : \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^m \to \mathbb{R}$, defined by

$$L(x, \lambda, \mu) := f(x) + \lambda^\top h(x) + \mu^\top g(x) = f(x) + \sum_{i=1}^{p} \lambda_i h_i(x) + \sum_{j=1}^{m} \mu_j g_j(x),$$

the **Lagrange function** or **Lagrangian**. The λ_i, μ_j are called **Lagrange multipliers** corresponding to the constraints $h_i(x) = 0, g_j(x) \leq 0$, respectively.

Contents

Lagrange multiplier rule for equality and inequality constraints

Theorem

Let x^* be a regular point and a local mimimizer (or maximizer) of f subject to the constraints $h(x) = 0, g(x) \leq 0$. Then, there exist **Lagrange multipliers** $\lambda \in \mathbb{R}^p, \mu \in \mathbb{R}^m_{\geq 0}$ *with*

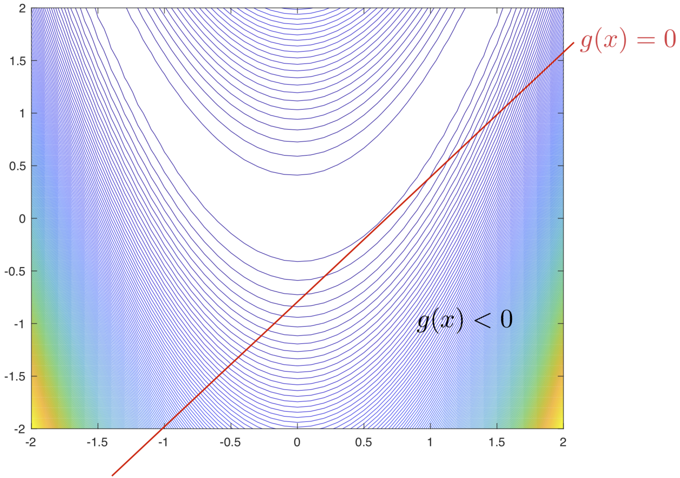
$$\nabla f(x^*) + \sum_{i=1}^{p} \lambda_i \nabla h_i(x^*) + \sum_{j=1}^{m} \mu_j \nabla g_j(x^*) = 0,$$

$$\mu^\top g(x^*) = 0.$$

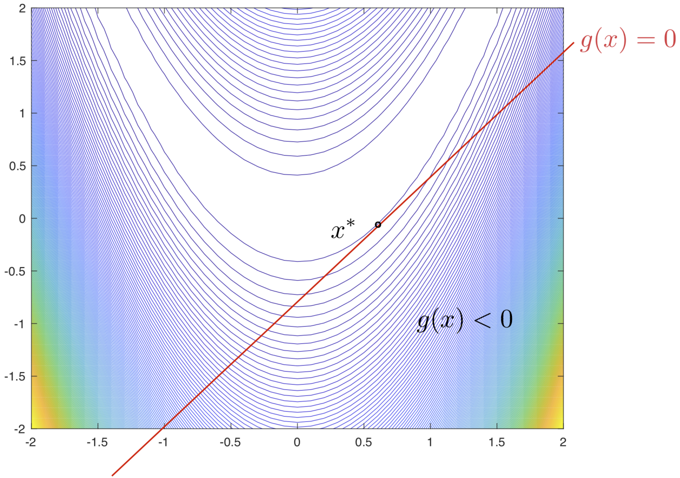
Definition

At a point $x \in X_{ad}$, the constraints g_j where $g_j(x) = 0$ are called **active**.
A point $x \in X_{ad}$ is called **regular** (w.r.t. the constraints $h(x) = 0, g(x) \leq 0$), if the set of gradients $\{\nabla h_i(x), \nabla g_j(x) : i = 1, \dots, p, g_j \text{ is active}\}$ is linear independent.

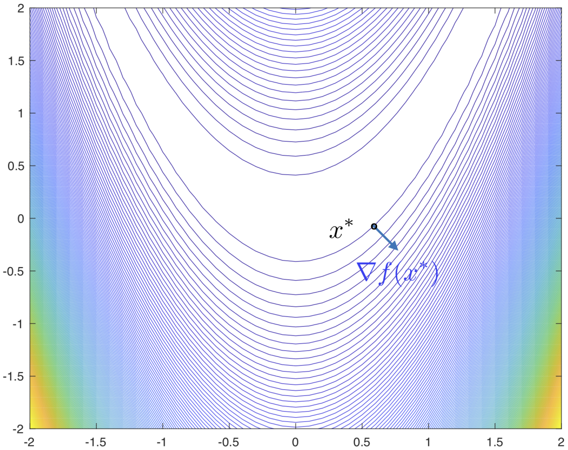
Example: Rosenbrock function with inequality constraint



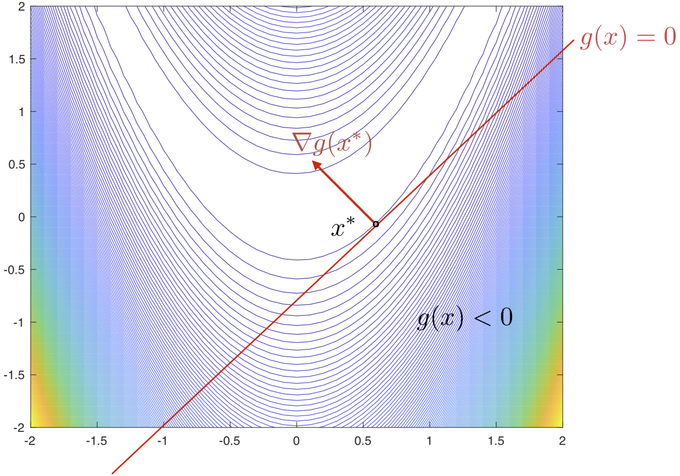
Minimizer x^* of the constrained problem



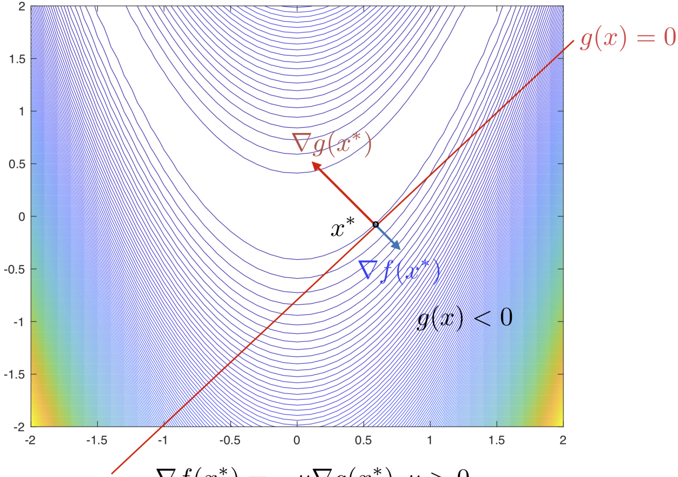
Gradient of cost function at the minimizer



Gradient of constraint at the minimizer



Both gradients are linear dependent, multiplier is positive



$$\nabla f(x^*) = -\mu \nabla g(x^*), \mu > 0$$

Karush-Kuhn-Tucker (KKT) system

- Summarizing, we look for $(x, \lambda, \mu) \in \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^m_{\geq 0}$ that satisfies

$$\begin{aligned} \nabla f(x) + \sum_{i=1}^{p} \lambda_i \nabla h_i(x) + \sum_{j=1}^{m} \mu_j \nabla g_j(x) &= 0 \\ \mu^\top g(x) &= 0, \qquad (1) \\ h(x) &= 0, \\ g(x) &\leq 0. \end{aligned}$$

- This system is also called **Karush-Kuhn-Tucker conditions** or **KKT system**.
- (1) is called **complementarity condition**. It can be equivalently written as

$$\sum_{j=1}^{m} \underbrace{\mu_j}_{\geq 0} \underbrace{g_j(x)}_{\leq 0} = 0 \quad \text{or} \quad \mu_j g_j(x) = 0 \text{ for all } j \quad \text{or} \quad \begin{cases} \mu_j > 0 \Rightarrow g_j(x) = 0 & \text{(active)} \\ g_j(x) < 0 \Rightarrow \mu_j = 0 & \text{(inactive)}. \end{cases}$$

Exploiting the complementarity condition

- Complementarity condition:

$$\begin{aligned} \text{active constraints:} \quad & \mu_j > 0 \quad \Rightarrow \quad g_j(x) = 0, \quad j \in \mathcal{A}(x) \subset \{1, \ldots, m\}, \\ \text{inactive constraints:} \quad & g_j(x) < 0 \quad \Rightarrow \quad \mu_j = 0, \quad j \in \mathcal{I}(x) \subset \{1, \ldots, m\}, \end{aligned}$$

  where we split the indices according to active and inactive constraints at fixed x.
- KKT system re-written (inactive constraints can be omitted in first equation):

$$\begin{aligned} \nabla f(x) + \sum_{i=1}^{p} \lambda_i \nabla h_i(x) + \sum_{j \in \mathcal{A}(x)} \mu_j \nabla g_j(x) &= 0, \\ h(x) &= 0, \\ g_j(x) &= 0, \quad j \in \mathcal{A}(x) \\ g_j(x) &< 0, \quad j \in \mathcal{I}(x). \end{aligned}$$

Example (see last lecture): Lagrange multiplier rule

- Compute a rectangle with maximal area and (small change!) given perimeter $\leq c$:

$$\min_{x \in \mathbb{R}^2} (-x_1 x_2) \;\text{ s.t. }\; g(x) = 2(x_1 + x_2) - c \leq 0.$$

  This is now an inequality constraint.

- Lagrange function:

$$L(x, \mu) = f(x) + \mu g(x) = -x_1 x_2 + \mu(2(x_1 + x_2) - c).$$

- KKT system:

$$\begin{aligned} \nabla f(x) + \mu \nabla g(x) &= 0 \\ g(x) &\leq 0 \\ + \text{ complementarity condition } \rightarrow \qquad \mu g(x) &= 0. \end{aligned}$$

Example: Lagrange multiplier rule

- KKT system:

$$\left.\begin{aligned} \nabla f(x) + \mu \nabla g(x) &= 0 \\ g(x) &\leq 0 \\ \mu g(x) &= 0 \end{aligned}\right\} \Rightarrow \left\{ \begin{aligned} -x_2 + 2\mu &= 0 \\ -x_1 + 2\mu &= 0 \\ 2(x_1 + x_2) - c &\leq 0 \\ \mu\left(2(x_1 + x_2) - c\right) &= 0 \end{aligned}\right.$$

- Because of the inequality, a direct computation is now not possible.
- From the first two equations, we deduce:

$$x_1 = x_2 = 2\mu.$$

- From the complementarity condition, we obtain:

$$\text{constraint inactive: } 2(x_1 + x_2) < c \Rightarrow \mu = 0 \Rightarrow x_1 = x_2 = 0.$$

$\rightsquigarrow$ a non-zero solution we only get if the constraint is active, i.e., if

$$2(x_1 + x_2) = 8\mu = c \;\Rightarrow\; \mu = \frac{c}{8} \;\Rightarrow\; x_1 = x_2 = \frac{c}{4} \text{ (as above).}$$

Second example: Lagrange multiplier rule

- Consider the problem

$$\min_{x\in\mathbb{R}^2} (x_1 - 1)^2 + (x_2 + 1)^2 \quad \text{s.t.} \quad x \in X_{ad} := [2,3] \times [0,1].$$

- Standard notation for the constraints:

$$\begin{aligned} g_1(x) &= 2 - x_1 \le 0, & g_2(x) &= x_1 - 3 \le 0 \\ g_3(x) &= -x_2 \le 0, & g_4(x) &= x_2 - 1 \le 0. \end{aligned}$$

- Gradients:

$$\nabla f(x) = \begin{pmatrix} 2(x_1 - 1) \\ 2(x_2 + 1) \end{pmatrix}, \nabla g_1(x) = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \nabla g_2(x) = \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

$$\nabla g_3(x) = \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \nabla g_4(x) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

- Compute the KKT system!

Example: KKT System

$$
\begin{aligned}
\nabla f(x) + \sum_{j=1}^{4} \mu_j \nabla g_j(x) &= \begin{pmatrix} 2(x_1-1) \\ 2(x_2+1) \end{pmatrix} + \mu_1 \begin{pmatrix} -1 \\ 0 \end{pmatrix} + \mu_2 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \mu_3 \begin{pmatrix} 0 \\ -1 \end{pmatrix} + \mu_4 \begin{pmatrix} 0 \\ 1 \end{pmatrix} \\
&= \begin{pmatrix} 2x_1 - 2 - \mu_1 + \mu_2 \\ 2x_2 + 2 - \mu_3 + \mu_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},
\end{aligned}
$$

$$
\begin{aligned}
g_1(x) &= 2 - x_1 \le 0, \quad & g_2(x) &= x_1 - 3 \le 0, \\
g_3(x) &= -x_2 \le 0, \quad & g_4(x) &= x_2 - 1 \le 0,
\end{aligned}
$$

$$
\sum_{j=1}^{4} \mu_j g_j(x) = 0, \quad \mu_j \ge 0, \quad j = 1, \ldots, 4.
$$

- More complicated, direct solution usually difficult because of inequality $\rightsquigarrow$ need different approach/algorithm.

Contents

Second order sufficient condition for constrained problems

The result is analogous to the one for pure equality constraints. We only have to consider the active inequality constraints.

Theorem (Second order sufficient optimality condition)

Let

- f, g, h *be twice continuously differentiable,*
- x^* *be regular,*
- $(x^*, \lambda, \mu) \in \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^m_{\geq 0}$ *satisfy the KKT system,*
- *the Hessian matrix* $\nabla^2_{xx} L(x^*, \lambda, \mu)$ *be positive definite on the set*

$$S := \left\{ d \in \mathbb{R}^n : \nabla h_i(x^*)^\top d = 0, i = 1, \ldots, p, \nabla g_j(x^*)^\top d = 0, j \in \mathcal{A}(x^*), \mu_j > 0 \right\},$$

where $\mathcal{A}(x^*) \subset \{1, \ldots, m\}$ *is the index set of active inequality constraints in* x^**.*

Then x^* *is a strict local minimizer of* f *in* X_{ad}*.*

Contents

Interpretation of Lagrange Multipliers

- Up to now, the Lagrange multipliers were just additional unknowns that we used to compute the solution x^* of the optimality system, i.e., a candidate for a minimizer.
- But the multipliers have a meaning:

Theorem (Sensitivity)

Let the second order sufficient condition be satisfied in (x^*, λ, μ). We denote by $x^*(\delta, \varepsilon)$ the minimizer of the perturbed problem

$$\min_{x \in \mathbb{R}^n} f(x) \text{ s.t. } g(x) \leq \delta, h(x) = \varepsilon.$$

Then the multipliers λ, μ satisfy

$$\mu = -\frac{\partial}{\partial \delta} f(x^*(\delta = 0, \varepsilon)), \quad \lambda = -\frac{\partial}{\partial \varepsilon} f(x^*(\delta, \varepsilon = 0)).$$

- In economics, the Lagrange mulitipliers are also referred to as **shadow prices**.

Example: Lagrange mulitiplier and sensitivity

- Compute a rectangle with maximal area and given perimeter c (equality constraint).
- Solution was: $(x^*, \lambda) = (\frac{c}{4}, \frac{c}{4}, \frac{c}{8}), f(x^*) = -\frac{c^2}{16}$.
- Perturbed problem for $\varepsilon > 0$: increased perimeter $c \to c + \varepsilon$:

$$\min_{x \in \mathbb{R}^2} (-x_1 x_2) \text{ s.t. } h(x) = 2(x_1 + x_2) - c = \varepsilon.$$

- Solution of perturbed problem: $x^*(\varepsilon) = (\frac{c+\varepsilon}{4}, \frac{c+\varepsilon}{4}), f(x^*(\varepsilon)) = -\frac{(c+\varepsilon)^2}{16}$.
- Sensitivity:

$$\frac{d}{d\varepsilon} f(x^*(\varepsilon)) = \frac{d}{d\varepsilon}\left(-\frac{(c+\varepsilon)^2}{16}\right) = -2\frac{c+\varepsilon}{16}.$$

- This gives:

$$-\frac{d}{d\varepsilon} f(x^*(\varepsilon = 0)) = 2\frac{c}{16} = \frac{c}{8} = \lambda.$$

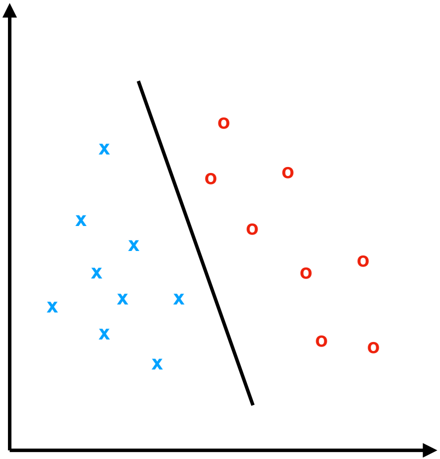
- Lagrange multiplier λ at unperturbed solution $=$ (negative) sensitivity.

Contents

Support vector machine (SVM): Tool for classification

- Given: Data $z_j = (z_{ji})_{i=1}^n \in \mathbb{R}^n, j = 1, \ldots, m$.
- They are separated in (here) two classes ...
- ... by a function f satisfying $f(z_j) = \pm 1$.
- Aim: Find separating hyperplane H for

$$\{z_j : j = 1, \ldots, m, f(z_j) = 1\},$$
$$\{z_j : j = 1, \ldots, m, f(z_j) = -1\}.$$



Support Vector Machine: Hyperplanes

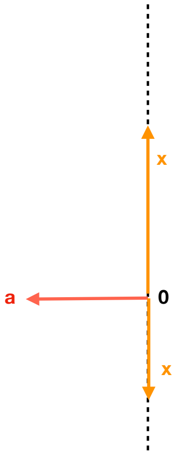
- Take any $a \in \mathbb{R}^n \setminus \{0\}$.
- All points $x \in \mathbb{R}^n$ satisfying

$$a \perp x, \text{i.e.,} \quad a^\top x = 0$$

lie on a linear hyperplane

$$H = \left\{x \in \mathbb{R}^n : a^\top x = 0\right\}$$

going through the origin.



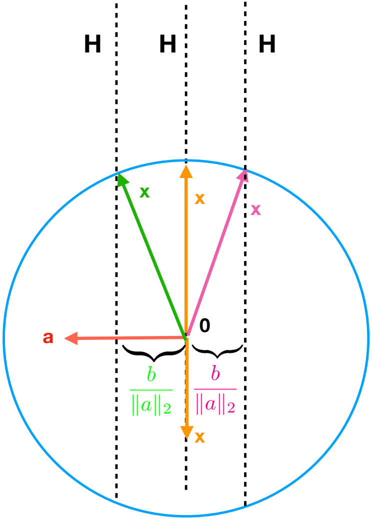
Support Vector Machine: Hyperplanes

- Generalizing:
  Take any $a \in \mathbb{R}^n \setminus \{0\}$ and $b \in \mathbb{R}$.
- All points $x \in \mathbb{R}^n$ satisfying

$$a^\top x = b$$

  lie on a affine-linear (shifted) hyperplane

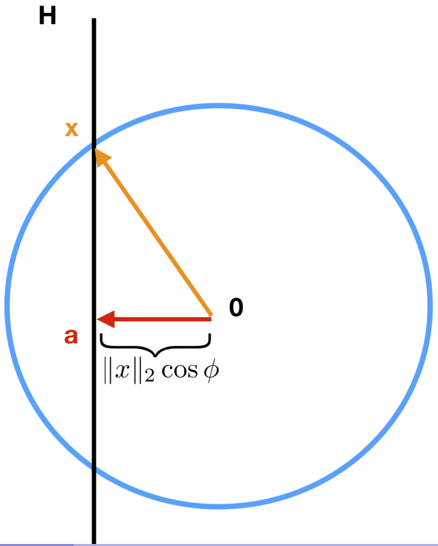
$$H = \left\{x \in \mathbb{R}^n : a^\top x = b\right\}.$$



Recall some definitions from geometry

- Cosine in right-angled triangle
  = projection $P_a(x)$ of $x \in \mathbb{R}^n$ onto $a \in \mathbb{R}^n \setminus \{0\}$:
$$\|P_a(x)\|_2 = \|x\|_2 \cos\phi.$$
- Special case $\|x\|_2 = 1$:
$$\|P_a(x)\|_2 = \cos\phi.$$
- Relation scalar product $\leftrightarrow$ angle $\phi = \angle(a, x)$:
$$\cos\phi = \frac{a^\top x}{\|a\|_2\,\|x\|_2}.$$

$\rightsquigarrow$ Length of projection of x onto a:
$$\|P_a(x)\|_2 = \frac{a^\top x}{\|a\|_2}.$$



Support vector machine

- Length of projection of arbitrary x onto a:

$$\|P_a(x)\|_2 = \|x\|_2 \cos\phi = \frac{a^\top x}{\|a\|_2}.$$

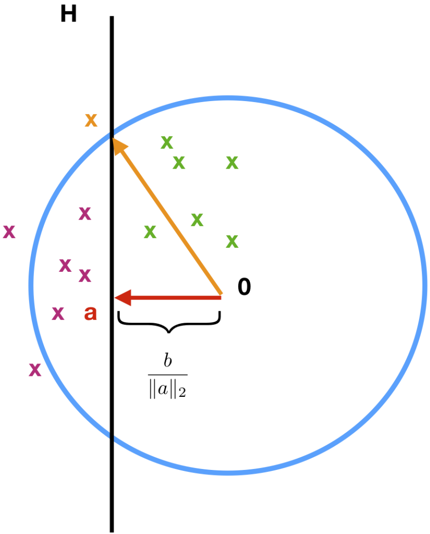
- Points x on the hyperplane satisfy:

$$\|P_a(x)\|_2 = \frac{a^\top x}{\|a\|_2} = \frac{b}{\|a\|_2} \Leftrightarrow a^\top x = b.$$

- Points x on one side of the hyperplane satisfy:

$$\|P_a(x)\|_2 = \frac{a^\top x}{\|a\|_2} > \frac{b}{\|a\|_2} \Leftrightarrow a^\top x - b > 0$$

or $$\|P_a(x)\|_2 = \frac{a^\top x}{\|a\|_2} < \frac{b}{\|a\|_2} \Leftrightarrow a^\top x - b < 0.$$



Support vector machine

- Points x on one side of the hyperplane satisfy:

$$\|P_a(x)\|_2 = \frac{a^\top x}{\|a\|_2} > \frac{b}{\|a\|_2} \Leftrightarrow a^\top x - b > 0,$$

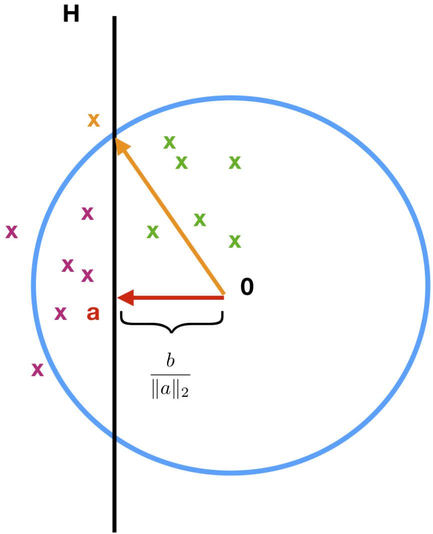
or
$$\|P_a(x)\|_2 = \frac{a^\top x}{\|a\|_2} < \frac{b}{\|a\|_2} \Leftrightarrow a^\top x - b < 0.$$

- More distinct classification: define margin $\Delta > 0$:

$$a^\top x - b \geq \Delta \quad \text{for points with } f(x) = +1$$
$$a^\top x - b \leq -\Delta \quad \text{for points with } f(x) = -1.$$

- Can be written as

$$(a^\top x - b) f(x) \geq \Delta \quad \text{for all points } x.$$



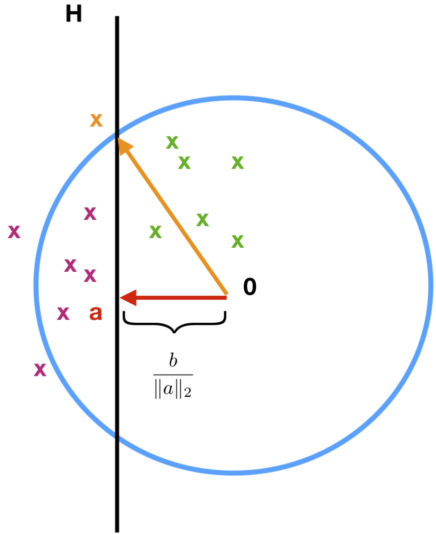
Support vector machine

- (Signed) Distance of a point x to hyperplane H:

$$\operatorname{dist}(x,H) = \frac{a^\top x}{\|a\|_2} - \frac{b}{\|a\|_2} = \frac{1}{\|a\|_2}(a^\top x - b)$$

$$\Rightarrow \quad \operatorname{dist}(x,H) \begin{cases} \geq \quad \dfrac{\Delta}{\|a\|_2}, & f(x) = +1 \\ \leq -\dfrac{\Delta}{\|a\|_2}, & f(x) = -1 \end{cases}$$

- Separation better, if $\|a\|_2$ (or $\|a\|_2^2$) is small.



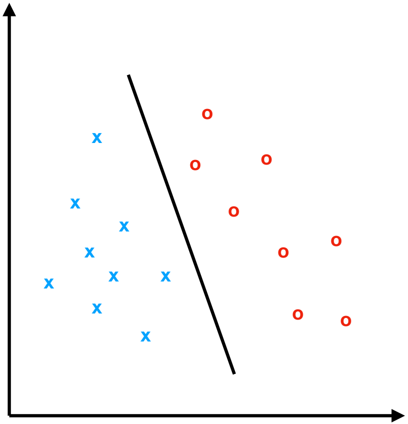
Support vector machine

- Optimization problem for given data $z_j = (z_{ji})_{i=1}^n \in \mathbb{R}^n, j = 1, \ldots, m$:

$$\min_{(a,b)\in\mathbb{R}^n\times\mathbb{R}} \|a\|_2^2$$
$$\text{s.t. } \left(a^\top z_j - b\right) f(z_j) \geq \Delta, \quad j = 1, \ldots, m.$$

$\rightsquigarrow$ quadratic cost with linear inequality constraints.

- Position of separation line just depends on points (vectors) z_i closest to it ("support vectors").
- This procedure works if data are linearly separable.
- If not: "kernel trick", transform data to higher-dimensional space where they are linearly separable.



Support vector machine

- Optimization problem:

$$\min_{(a,b)\in\mathbb{R}^n\times\mathbb{R}} \|a\|_2^2 \text{ s.t. } (a^\top z_j - b)f(z_j) \geq \Delta, \quad j = 1,\ldots,m.$$

- Constraint rewritten as:

$$g_j(a,b) = \Delta - (a^\top z_j - b)f(z_j) \leq 0, \quad j = 1,\ldots,m.$$

- Lagrangian with $\mu \in \mathbb{R}^m$:

$$L(a,b,\mu) = \sum_{i=1}^n a_i^2 + \sum_{j=1}^m \mu_j \left(\Delta - \left(\sum_{i=1}^n a_i z_{ji} - b\right) f(z_j)\right)$$

- Need solution algorithm for problems with inequality constraints.

What is important

- The Lagrange multiplier rule can be extended for problems with both equality and inequality constraints.
- The rule has basically the same form as for pure equality constraints, with two differences:
- The Lagrange multipliers for the inequality constraints are always non-negative and satisfy a complementarity condition.
- The resulting system is called KKT system.
- Its numerical solution is different, since the optimality system now includes inequalities.
- The complementarity condition helps to identify active and inactive inequality constraints.
- Support vector machines are a method for data classification.
- The problem behind is an optimization problem with inequality constraints.