Optimization & Operational Research : Second Part

Antoine Gourru

Slides built by Guillaume Metzler, updated by Ievgen Redko

January 2024 - Semester II

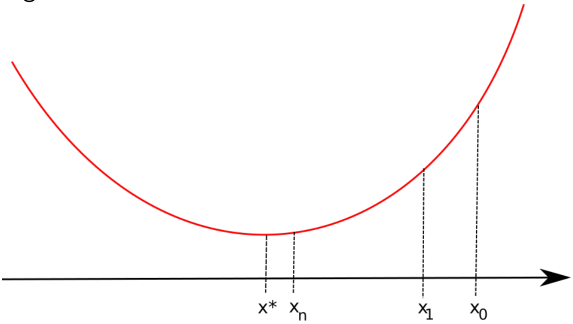
Convexity

What is a convex optimization problem ?

Given a convex function $f : \mathbb{R}^n \to \mathbb{R}$ we would solve the problem :

$$\hat{x} = \underset{x \in \mathbb{R}^n}{\arg\min} \quad f(x).$$

The aim of this part is to introduce algorithms building a series $(x_n)_{n \in \mathbb{N}}$ which converges to $\hat{x}$.



Optimization

There exists several type of optimization problems :

- convex optimization as presented before
- constrained optimization problem,
- non convex optimization problem,
- non differentiable convex optimization problem
- ...

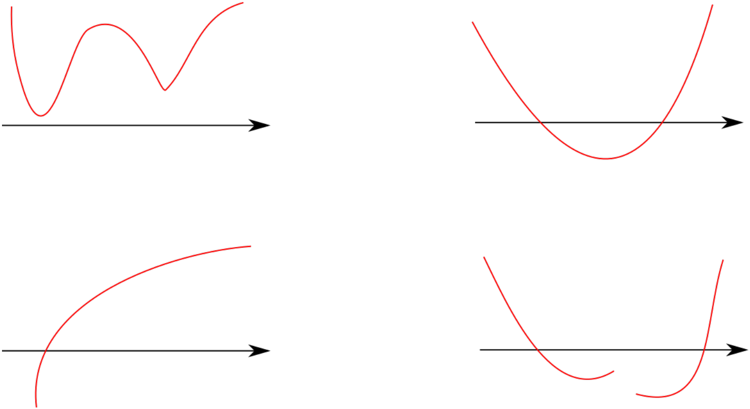
we only focus on convex optimization problem !

Why do we study them

1. **Cornerstone** in modern Machine Learning.
2. Convex function can be optimized easier (**Gradient Descent** vs **Newton's Method**.)

Convex Functions

Which of the following functions are convex graphically ?



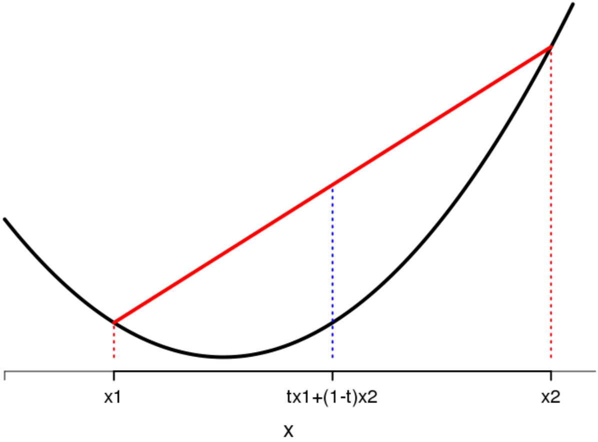
Convex Functions

Definition

Let $\mathcal{U}$ be an on empty set of a vector space ($\mathcal{U} = \mathbb{R}^n$). A function $f : \mathcal{U} \to \mathbb{R}$ is said to be convex if, for every $(u, v) \in \mathcal{U}$ and for all $t \in [0, 1]$, we have :

$$f(tu + (1-t)v) \leq tf(u) + (1-t)f(v).$$

- A linear function is convex,
- $f : \mathbb{R} \to \mathbb{R}, \quad f(x) = x^2$,
- $f : \mathbb{R} \to \mathbb{R}, \quad f(x) = \exp(x)$.



A convex function and its chord

Convex Functions and line segment

Proposition

A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is **convex** if and only the function
$g(t) = f(x + tv)$ is convex
for all x, v such that $x + tv$ belongs to the domain of definition of f
(f is concave if and only if g is concave).

Convex Functions

Exercise

Show that the function $F : \mathbb{R} \to \mathbb{R}, \quad f(x) = x^2$ is convex.

Solution : we need to show $(tx + (1-t)y)^2 \leq tx^2 + (1-t)y^2$.

$$\iff t^2x^2 + 2t(1-t)xy + (1-t)^2y^2 \leq tx^2 + (1-t)y^2,$$

$$\iff (t^2 - t)x^2 + 2t(1-t)xy + ((1-t)^2 - (1-t))y^2 \leq 0,$$

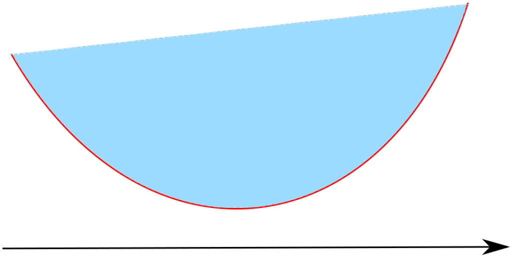
$$\iff t(t-1)x^2 - 2t(t-1)xy + t(t-1)y^2 \leq 0,$$

$$\iff t(t-1)(x-y)^2 \leq 0,$$

Convex functions

Equivalent definition

A function f is convex on $\mathcal{U}$ if and only if its epigraph E is convex, where $E = \{(x, y) \in \mathcal{U} \mid f(x) \leq y\}$.



Epigraph is the blue domain, which is convex

Concavity

Remark

Let $\mathcal{U}$ be an on empty set of a vector space ($\mathcal{U} = \mathbb{R}^n$). A function $f : \mathcal{U} \to \mathbb{R}$ is said to be concave if, for every $(u, v) \in \mathcal{U}$ and for all $t \in [0, 1]$, we have :

$$f(tu + (1-t)v) \geq tf(u) + (1-t)f(v).$$

If f is concave, then $-f$ is a convex function.

The function f defined by $f(x) = \ln(x)$ is concave.

Convex Functions

1. Given two convex functions f and g defined on $\mathcal{U}$, the sum $f+g$ is also a convex function.
2. If f is an increasing and convex function, g a convex function, then $f \circ g(x)$ is convex.
3. If f and g are convex functions, then h defined by $h(u) = \max\left(f(u), g(u)\right)$ is also convex

Exercise

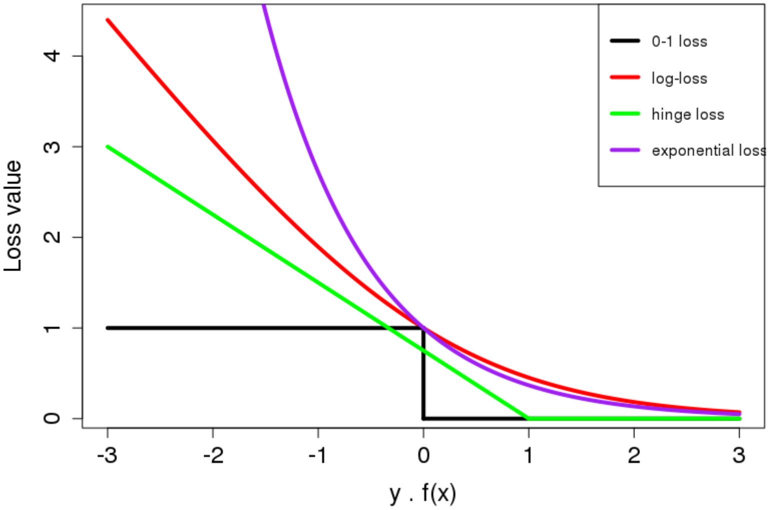
Prove the two first points using the definition of convexity.

Correction

1. For this one, you have to notice that $(f+g)(x) = f(x) + g(x)$ and apply the definition of convexity
2.

$$\begin{array}{rcl} g(tx+(1-t)y)) & \leq & tg(x)+(1-t)g(y) \\ f(g(tx+(1-t)y))) & \leq & f(tg(x)+(1-t)g(y)) \\ f(g(tx+(1-t)y))) & \leq & tf(g(x))+(1-t)f(g(y)) \\ f\circ g(tx+(1-t)y) & \leq & tf\circ g(x)+(1-t)f\circ g(y) \end{array}$$

Convex Loss Functions



Convexity and differentiability

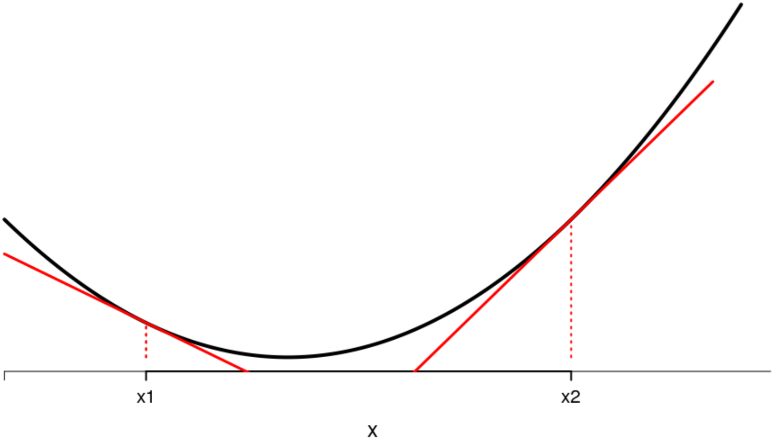
Proposition

Let f be a continuously differentiable function (C^1) on $\mathcal{U}$. Then f is convex if and only if, for all $(u, v) \in \mathcal{U}$, we have :

$$f(v) \geq f(u) + \nabla f(u)(v - u).$$

Equivalently if and only if, for all $(u, v) \in \mathcal{U}$, we have :

$$(\nabla f(v) - \nabla f(u))(v - u) \geq 0$$


x1
x2
x

Convexity and differentiability

Definition

Let f be a function of class C^2 on $\mathcal{U}$ and let H be its Hessian. Then f is **convex** if :

- $\nabla^2 f(u) \geq 0$ for all $u \in \mathcal{U}$.
- H is a positive semi definite (PSD), i.e, $\forall u \in \mathcal{U}$

$$u^T H u \geq 0.$$

Recall

A matrix H is PSD if and only if all of it's eigenvalues are **non-negative**

Convexity and differentiability

Interpretation

Positive eigenvalues imply that the **gradient is an increasing function** along each direction of the space

We consider a 2×2 matrix A :

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

where a, b, c, d are real numbers. We denote by λ_1, λ_2 the eigenvalues of this matrix (roots of the polynomial $\det(XI_2 - A)$).

Convexity and differentiability

1. We'll show why, for a 2×2 matrix, we have the following equivalence :
   A is PSD $\iff Tr(A) \geq 0$ **and** $\det(A) \geq 0$.
2. We have $\det(XI_2 - A)) = x^2 - (a+d)x + ad - bc$. The roots of this polynomial are exactly the eigenvalues of the matrix A (by definition), so

   $$\det(XI_2 - A) = (x - \lambda_1)(x - \lambda_2) = x^2 - (\lambda_1 + \lambda_2)x + \lambda_1\lambda_2.$$

   So we have, for all $x \in \mathbb{R}$:

   $$x^2 - (a+d)x + ad - bc = x^2 - (\lambda_1 + \lambda_2)x + \lambda_1\lambda_2.$$

3. It implies : $\lambda_1 + \lambda_2 = a + d = Tr(A)$ and $\lambda_1\lambda_2 = ad - bc = \det(A)$.

Convexity and differentiability

1. We'll show why, for a 2×2 matrix, we have the following equivalence :
   A is PSD $\iff$ $Tr(A) \geq 0$ **and** $\det(A) \geq 0$.
2. We have $\det(XI_2 - A)) = x^2 - (a+d)x + ad - bc$. The roots of this polynomial are exactly the eigenvalues of the matrix A (by definition), so

   $$\det(XI_2 - A) = (x - \lambda_1)(x - \lambda_2) = x^2 - (\lambda_1 + \lambda_2)x + \lambda_1\lambda_2.$$

   So we have, for all $x \in \mathbb{R}$:

   $$x^2 - (a+d)x + ad - bc = x^2 - (\lambda_1 + \lambda_2)x + \lambda_1\lambda_2.$$

3. It implies : $\lambda_1 + \lambda_2 = a + d = Tr(A)$ and $\lambda_1\lambda_2 = ad - bc = \det(A)$.

Convexity and differentiability

1. We'll show why, for a 2×2 matrix, we have the following equivalence :
   A is PSD $\iff Tr(A) \geq 0$ **and** $\det(A) \geq 0$.
2. We have $\det(XI_2 - A)) = x^2 - (a+d)x + ad - bc$. The roots of this polynomial are exactly the eigenvalues of the matrix A (by definition), so

   $$\det(XI_2 - A) = (x - \lambda_1)(x - \lambda_2) = x^2 - (\lambda_1 + \lambda_2)x + \lambda_1\lambda_2.$$

   So we have, for all $x \in \mathbb{R}$:

   $$x^2 - (a+d)x + ad - bc = x^2 - (\lambda_1 + \lambda_2)x + \lambda_1\lambda_2.$$

3. It implies : $\lambda_1 + \lambda_2 = a + d = Tr(A)$ and $\lambda_1\lambda_2 = ad - bc = \det(A)$.

Convexity and differentiability

1. ($\Rightarrow$) If the eigenvalues are positive, we immediately see that both :

$$Tr(A) > 0 \quad and \quad \det(A) \geq 0.$$

2. ($\Leftarrow$) Conversely, if $\det(A) \geq 0$ it means that the two eigenvalues have the same sign. Moreover, if the trace is positive then the two eigenvalues are positive.

Convexity and differentiability

Remark

A matrix A is said to be NSD (Negative Semi-Definite) if its eigenvalues are non-positive. A 2×2 matrix A is NSD if we have :

$$Tr(A) < 0 \quad and \quad \det(A) \geq 0.$$

Examples

- If for all $i = 1, \ldots, n, \quad \lambda_i \geq 0$, then $H = \text{diag}(\lambda_i)$ is PSD.
- The function $f : \mathbb{R}^n \to \mathbb{R}$ defined by $f(x_1, \ldots, x_n) = \sum_{i=1}^{n} x_i^2$ is convex.

Examples

- If for all $i = 1, \ldots, n, \quad \lambda_i \geq 0$, then $H = \text{diag}(\lambda_i)$ is PSD.
- The function $f : \mathbb{R}^n \to \mathbb{R}$ defined by $f(x_1, \ldots, x_n) = \sum_{i=1}^n x_i^2$ is convex.

Exercises

- Show that the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by $f(x, y) = 2x^2 + 2xy + 2y^2$ is convex.
- Show that the function $f : \mathbb{R}^3 \to \mathbb{R}$ defined by $f(x, y, z) = 5x^2 + 2\sqrt{2}xy + 6y^2 + 3z^2$ is convex.
- Show that the function $f : \mathbb{R}^n \to \mathbb{R}$ defined by $f(x) = \log\left(\sum_{i=1}^N e^{x_i}\right)$ is convex.

Correction 1/6

For the two first functions, you have to check that all the eigenvalues of the Hessian Matrix are non-negative. So you need : 1) to compute the Hessian Matrix of the given function and 2) to compute the eigenvalues of this last. Remember that the eigenvalues of a given matrix H are given by finding the roots of the following polynomial in λ :

$$det(H - \lambda I_d)$$

Correction 2/6

- For the first function, the Hessian Matrix is given by :

$$H_f(x,y) = \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix},$$

The eigenvalues are then given by finding the roots of the polynom :

$$det\,(H_f(x,y) - \lambda I_2) = det\begin{pmatrix} 4-\lambda & 2 \\ 2 & 4-\lambda \end{pmatrix} = (4-\lambda)^2 - 2^2 = (\lambda-2)(\lambda-6).$$

The eigenvalues are 2 and 6, they are non-negative so the function f is convex.

Correction 3/6

- For the second function, the Hessian Matrix is given by :

$$H_f(x,y) = \begin{pmatrix} 10 & 2\sqrt{2} & 0 \\ 2\sqrt{2} & 12 & 0 \\ 0 & 0 & 6 \end{pmatrix},$$

The eigenvalues are then given by finding the roots of the polynom :

$$det\left(H_f(x,y) - \lambda I_3\right) = det\begin{pmatrix} 10-\lambda & 2\sqrt{2} & 0 \\ 2\sqrt{2} & 12-\lambda & 0 \\ 0 & 0 & 6-\lambda \end{pmatrix}.$$

$$det\left(H_f(x,y) - \lambda I_3\right) = (6-\lambda)[(10-\lambda)(12-\lambda)-8] = (6-\lambda)(\lambda-8)(\lambda-14).$$

The eigenvalues are $6, 8$ and 14, they are non-negative so the function f is convex.

Correction 4/6

- For this last function, we will use the expression of the Jacobian previously computed :

$$J_f(x) = \frac{1}{\sum_{i=1}^{n} \exp(x_i)} (\exp(x_1, ..., \exp(x_n))$$

Then we compute the Hessian, we will seperate the diagonal terms with the non-diagonal one. For convenience, we will set $z_i = \exp(x_i)$, $Z = \sum_{i=1}^{n} \exp(x_i)$ and $z = (z_1, ..., z_n)$.

$$H_f(x, y)_{(i,j)} = \begin{cases} \frac{z_i Z - z_i^2}{Z^2} \quad if \quad i = j \\ \\ -\frac{z_i z_j}{Z^2} \quad if \quad i \neq j \end{cases}$$

Correction 5/6

Using the previous notations, we can write :

$$H_f(x,y)_{(i,j)} = \frac{1}{Z} diag(z) - \frac{1}{Z^2} zz^T.$$

To proove that this function is convex, we will show that for vector $u \in \mathbb{R}^n$ we have $u^T H_f u \geq 0$.

$$u^T H_f u = \frac{1}{Z^2} \left(\left(\sum_{i=1}^{n} u_i^2 z_i \right) \left(\sum_{i=1}^{n} z_i \right) - \left(\sum_{i=1}^{n} u_i z_i \right)^2 \right).$$

We need to show that is expression is non-negative. For that, we use the **Cauchy-Schwarz Inequality**. So we will introduce inner product and norms.

Correction 6/6

Note that : $\sum_{i=1}^n u_i^2 z_i = \|u_i\sqrt{z_i}\|_2^2$, $\sum_{i=1}^n z_i = \|\sqrt{z_i}\|_2^2$ and $\left(\sum_{i=1}^n u_i z_i\right)^2 = \|u_i z_i\|_2^2$. So that :

$$u^T H_f u = \frac{1}{Z^2}\left(\|u\sqrt{z}\|\|\sqrt{z}\| - \langle u\sqrt{z}, \sqrt{z}\rangle^2\right).$$

We can bound the inner product as follow :

$$\langle u\sqrt{z}, \sqrt{z}\rangle^2 \leq \|u\sqrt{z}\|\|\sqrt{z}\|.$$

We conclude that :

$$u^T H_f u \geq 0.$$

Convex Optimization

Condition of Optimality

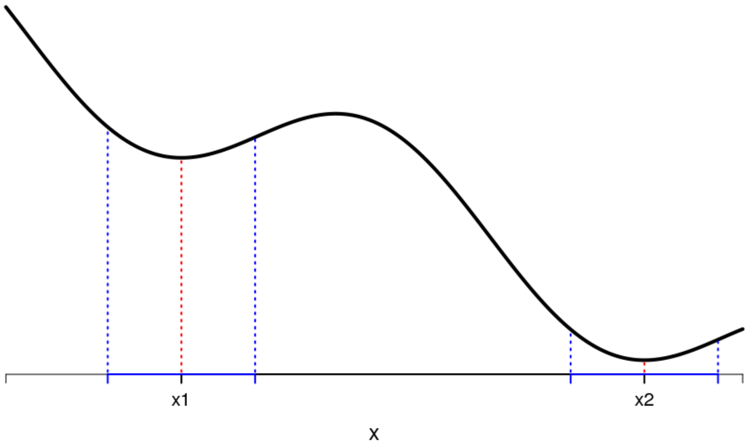
Definition

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous function. We say that $u \in \mathbb{R}^n$ is a local minimum of f if it exists a neighborhood $V \subset \mathbb{R}^n$ of u such that :

$$f(u) \leq f(v), \quad \forall v \in V.$$

u is a global minimum of the function f if and only if :

$$f(u) \leq f(v), \quad \forall v \in \mathbb{R}^n.$$



- x_1 and x_2 are two local minima of f.
- x_2 is the global minimum of the function f

Condition of Optimality

Proposition : - Euler's Equation -

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous function and differentiable at $u \in \mathbb{R}^n$. If u is a local minimum then we have : $\nabla f(u) = 0.$

Condition of Optimality

Proposition : - Euler's Equation -

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous function and differentiable at $u \in \mathbb{R}^n$. If u is a local minimum then we have : $\nabla f(u) = 0$.

Proof : In fact, using the definition : $\forall v \in \mathbb{R}^n, \exists t > 0$ such that $u + tv \in V$ a neighborhood of u.

$$\begin{array}{rcl} & f(u) & \leq f(u+tv) = f(u) + \nabla f(u)(tv) + tv\,\varepsilon(tv), \quad t \ll 1 \\ \Longleftrightarrow & 0 & \leq \nabla f(u)(tv) + tv\,\varepsilon(tv) \end{array}$$

Dividing by $t > 0$ and taking the limit $t \to 0$ we have : $0 \leq \nabla f(u)v$.
Same thing by replacing $v \to -v$ we have $0 \leq -\nabla f(u)v$.
So $\forall v \in \mathbb{R}^n, \quad \nabla f(u)v = 0 \Rightarrow \nabla f(u) = 0$.

Condition of Optimality

The solution of *Euler's Equation* gives us the points where the function f reaches a local extremum (a minimum or maximum (local or global)).

Given a solution u of $\nabla f(u) = 0$, we can say that :

- u is local minimum if $\nabla^2 f(u) = H_f(u) \geq 0$, i.e. the Hessian matrix evaluated at the point u is PSD. This point is a global minimum if the function is convex everywhere or if for all $v \neq u$ we have $f(u) \leq f(v)$.
- u is local maximum if $\nabla^2 f(u) = H_f(u) \leq 0$, i.e. the Hessian matrix evaluated at the point u is NSD. This point is a global maximum if the function is concave everywhere or if for all $v \neq u$ we have $f(u) \geq f(v)$.

Condition of Optimality

Definition

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous function and $\mathcal{U}$ a non empty set. We say that f has a relative minimum u if

$$f(u) \leq f(v), \quad \forall v \in \mathcal{U}.$$

Proposition : - Euler's Inequality -

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous function and $\mathcal{U}$ a non empty and convex set. Furthermore, let $u \in \mathcal{U}$ be a relative minimum of f. If f is differentiable at u we have : $\nabla f(u)(v - u) \geq 0$ $\forall v \in \mathcal{U}$.

Exercise

- Let f defined by $f(x,y) = (4-2y)^2 + 5x^2 + x + 3y + 4xy$
  1. Is the function f convex ?
  2. What is the global minimum of f ?
- Let f defined by $f(x,y) = 2x^2 + 4(y-2)^2 + 4x + 6y - 2xy + 2y^3$.
  1. Is f convex ?
  2. Give a condition on y so that f is convex.
  3. (Optional) For the previous condition on y, find the local minimum of f

1. The function f is convex. In fact, we have :

$$H_{f(x,y)} = \begin{pmatrix} \frac{\partial^2 f}{\partial^2 x} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial^2 y} \end{pmatrix} = \begin{pmatrix} 10 & 4 \\ 4 & 8 \end{pmatrix}.$$

Because f is convex, if we find (x, y) such that $\nabla f(x, y) = 0$ then (x, y) is the Argmin of f.

$$J_{f(x,y)} = \begin{pmatrix} 10x + 4y + 1, & 4x + 8y - 13 \end{pmatrix} = (0, 0).$$

The solution is $(x, y) = (-\frac{15}{16}, \frac{67}{32})$.

2) Same as before, we calculate the Hessian matrix :

$$H_{f(x,y)} = \begin{pmatrix} 4 & -2 \\ -2 & 12y+8 \end{pmatrix}.$$

We have $Tr(H) = 12y + 12$ and $det(H) = 48y + 28$. These quantities are both positie if and only if $y \geq -\frac{28}{48} = -\frac{7}{12}$.
So the function is not convex on $\mathbb{R}^2$, but it is on $\mathbb{R} \times [-\frac{7}{12}, \infty[$.

- You have to solve the following system :

$$\begin{aligned} 4x + 4 - 2y &= 0, \\ 6y^2 + 8y - 2x - 10 &= 0. \end{aligned}$$

$$\begin{aligned} 4x + 4 - 2y &= 0, \\ 6y^2 + 7y - 8 &= 0. \end{aligned}$$

You solve the following system, keeping the appropriate value of y and then you calculate x.

Convex Problems

The basic formulation

Given a vector space E and a function $f : E \to \mathbb{R}$, an optimization problem consists of solving the following problem :

$$\min_{x \in E} f(x).$$

- The function f is sometimes called the cost function (ie, cost for a company to store goods).

- Most of times, we want to minimize the function f under some constraints.

Linear Regression 1/3

Let us first consider the linear regression :

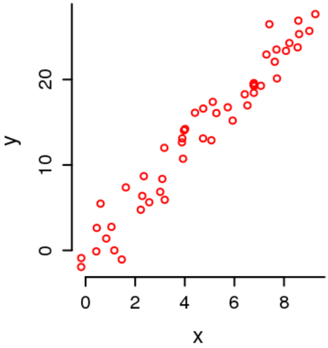
- Given a response vector $Y \in \mathbb{R}^n$ and feature vectors $X = (x_1, \ldots, x_n)^T, x_i \in \mathcal{R}^m$ where $m + 1 < n$.
We'd like to find a vector β that explain the value of Y using X with the following model

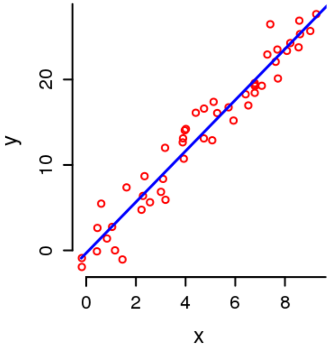
$$Y = X\beta + \varepsilon, \quad where \quad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

- ε represent the error due to the model. To find the best vector β we have to minimize this error, i.e. to solve :

$$\min_{\beta \in \mathbb{R}^{m+1}} \varepsilon \|Y - X\beta\|^2$$

Linear Regression 2/3





Linear Regression 3/3

We easily check that is problem is convex :

$$\nabla_\beta \, \varepsilon = -2X^T(Y - X\beta),$$

and

$$\nabla^2_\beta = 2X^T X,$$

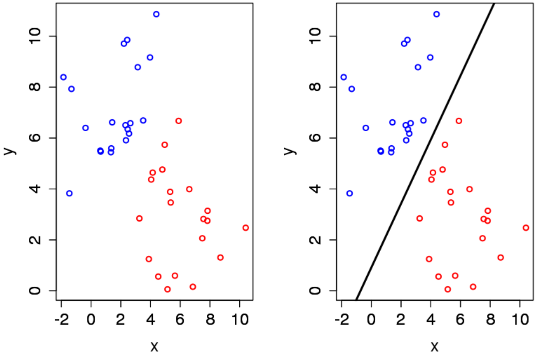
which is positive semi definite.
The solution given by

$$\beta = (X^T X)^{-1} X^T Y.$$

Analytic solution exists but this is not always the case

Logistic regression 1/2

We want to find a model that predict the class of our data.



Logistic Regression 2/2

- To predict the class of the individual we use a model of the form :

$$g(x, a) = \log\left(\frac{\mathbb{P}(X \mid Y = 1)}{1 - \mathbb{P}(X \mid Y = 1)}\right) = a_0 + a_1 x_1 + \ldots + a_m x_m.$$

- Solved by maximizing the (log-)likelihood of our data :

$$l(x, a) = \sum_{i=1}^{n} y_i \log(p_i) + (1 - y_i) \log(1 - p_i), \; p_i = \frac{1}{1 + \exp(-\sum_{j=1}^{m} a_j x_{ij})}.$$

No analytic solution, we need a way to **approximate it** step by step.