Optimization & Operational Research : First Part

Guillaume Metzler

Laboratoire Hubert Curien
UMR CNRS 5516
Université Jean Monnet Saint-Etienne

January 2018 - Semester II

Topic of the course

Headline

- Mathematical bakcground : Convex sets and derivatives.
- Convex function and their properties.
- What is a convex optimization problem ?
- Algorithm for convex optimization.

Some references

Linear Algebra

- K.B Petersen, M.S Pedersen, *The Matrix Cookbook*,2012.
Available at : http ://matrixcookbook.com

Convex Optimization

- Stephen Boyd & Lieven Vandenberghe, *Convex Optimization*,
Cambridge University Press, 2014

Mathematical Background

Norms

Given $x, y \in \mathbb{R}^n$, the inner product is given by :

$$\langle x, y \rangle = x^T y = \sum_{i=1}^{n} x_i y_i .$$

The inner product of x with itself is called the square of the norms of x

$$\langle x, x \rangle = \|x\|^2 .$$

Definition

Let E be a $\mathbb{R}$-vectorial space, then the application $\|.\|$ is said to be a norm if for all $u, v \in E$ and $\lambda \in \mathbb{R}$

1. (positive) $\|u\| \geq 0$,
2. (definite) $\|u\| = 0 \iff u = 0$,
3. (scalability) $\|\lambda u\| = |\lambda| \|u\|$,
4. (triangle inequality) $\|u + v\| \leq \|u\| + \|v\|$.

Norms

- The norm can be seen as distance between two vectors x, y in the same vectorial space

$$dist(x, y) = \|x - y\|.$$

Example of usual norms :

- $\|x\|_1 = \sum_{i=1}^{n} |x_i|$ (Manhattan)
- $\|x\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2}$ (Euclidean)
- $\|x\|_\infty = max\left(|x_1|, \ldots, |x_n|\right)$
- More generally we define the norm $\|.\|_p$ for all integers p as

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}.$$

Example 1/2

We will show that the Euclidean norm is effectively a norm. We have to check each point of the definition. Let us consider $x, y \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$

- It is obvisous that $\|x\|_2 = \sqrt{\sum_{i=1}^n |x_i|^2}$ is positive.
- Each $|x_i|^2$ is a positive number and a sum of positive number is equal to zero if and only if all of the numbers are equal to zero $\Rightarrow x = 0$
- $$\begin{aligned} \|\lambda x\|_2 &= \sqrt{\sum_{i=1}^n |\lambda x_i|^2} \\ &= \sqrt{\sum_{i=1}^n |\lambda|^2 |x_i|^2} \\ &= |\lambda| \sqrt{\sum_{i=1}^n |x_i|^2}. \end{aligned}$$

Example 2/2

To prove the last point we will use the Cauchy-Schwartz Inequality :

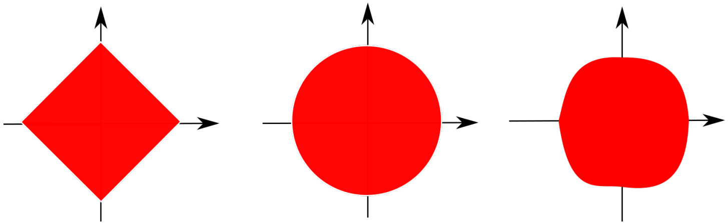
$$\langle x, y\rangle \leq \|x\| \|y\|.$$

- $$\begin{aligned} \|x+y\|_2^2 &= \|x\|_2^2 + 2\langle x, y\rangle + \|y\|_2^2 \\ &\leq \|x\|_2^2 + 2\|x\|_2\|y\|_2 + \|y\|_2^2 \\ &\leq \left(\|x\|_2 + \|y\|_2\right)^2 . \end{aligned}$$

By taking the square root, which is an increasing function, we get the result.

Norms and Unit Ball

Unit ball for the norms $\|\|_p$ for $p = 1, 2$ and $p > 2$



Exercise

- Represent the unit ball for the norm $\|.\|_\infty$.
- Show that the application $f : \mathbb{R}^n \to \mathbb{R}$ defined by $f(x) = \sum_{i=1}^{n} |x_i|$ is a norm.

Correction

- The Unit Ball using the $\|.\|_\infty$ is a full square.
- We have to check the four points of the definition.

1. $\|x\|_1 = \sum_{i=1}^n |x_i| \geq 0$ by definition of the absolute value.
2. $\|x\|_1 = \sum_{i=1}^n |x_i| \geq 0 \implies x = 0$ because the sum of positive numbers is equal to zero if and only if all the terms are equal to zero.
3. $\|\lambda x\|_1 = \sum_{i=1}^n |\lambda x_i| = |\lambda| \sum_{i=1}^n |x_i| = |\lambda| \|x\|_1$.
4. $\|x + y\|_1 = \sum_{i=1}^n |x_i + y_i| \leq \sum_{i=1}^n |x_i| + \sum_{i=1}^n |y_i| = \|x\|_1 + \|y\|_1$.

Norms on matrices

It is also to define an inner product and a norms on matrices :

- Given two matrices $X, Y \in \mathbb{R}^{m \times n}$ the inner product is defined by :

$$\langle X, Y \rangle = Tr\left(X^T Y\right) = \sum_{i=1}^{m} \sum_{j=1}^{n} x_{ij} y_{ij}.$$

- A classical norm used with matrices is the Frobenius norm :

$$\|X\|_F = \sqrt{Tr\left(X^T X\right)} = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} x_{ij}^2 \right)^{1/2}.$$

What is the inner product of the matrices $X, Y \in \mathcal{S}^n(\mathbb{R})$?

Convex Sets

Definition

A set C is said to be convex if, for every $(u, v) \in C$ and for all $t \in [0, 1]$ we have :

$$tu + (1-t)v \in C.$$

In other words, C is said to be convex if every point on the segment connecting u and v is in the set.

Proposition

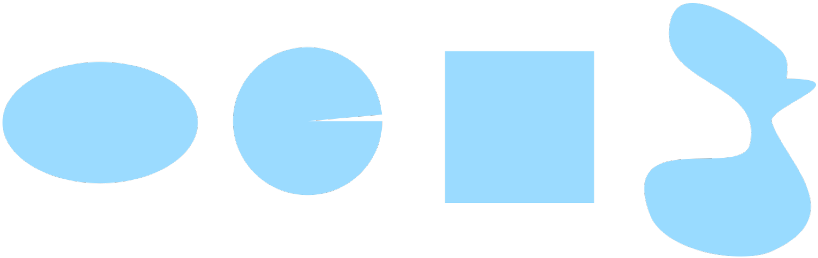
Let $(u_1, u_2, \dots, u_n)$ be a set of n points belonging to a convex set C. Then for every reel numbers $\lambda_1, \lambda_2, \dots, \lambda_n$ such that $\sum_{i=1}^{n} \lambda_i = 1$:

$$v = \sum_{i=1}^{n} \lambda_i u_i \in C.$$

It means that every convex combination of points belonging in a convex set, belongs in the convex set.

Convex Sets

Which of the sets are convex ?



Examples of Convex Sets

- $\mathcal{B} = \{u \in \mathbb{R}^n \mid \|u\| \leq 1\}$ is convex.
- Every segment in $\mathbb{R}$ is convex.
- Every hyperplane $\{x \in \mathbb{R}^n \mid a^T x = b\}$ is convex.
- If C_1 and C_2 are two convex sets, then the intersection $C = C_1 \cap C_2$ is also convex.

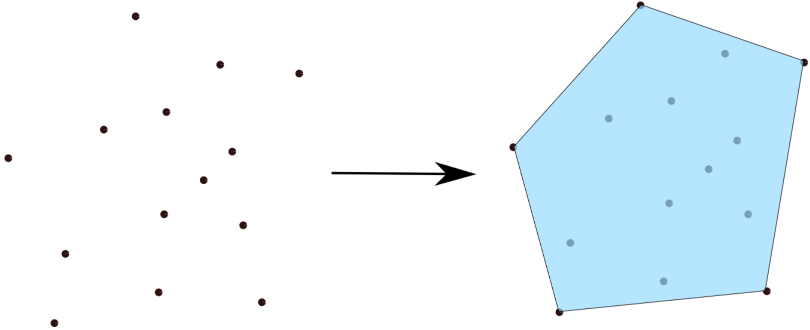
Exercise

- Proove that the Euclidean Unit Ball is convex.
- (At home) Proove that a set A is convex if and only if its intersection with any line is convex.

Build a convex set

Given the definition of a convex set and a set of point $x_1, \ldots x_n$, it is possible to build a convex set. This new set is called the convex hull $\mathcal{H}$ of a set of points

$$\mathcal{H} = \{y = \sum_{i=1}^{n} \lambda_i x_i \mid \sum_{i=1}^{n} \lambda_i = 1\}.$$



Derivative for real functions

Recall

Let $f : \mathbb{R} \to \mathbb{R}$ be continuous and $x_0 \in \mathbb{R}$. We say that f is differentiable at x_0 if the limit :

$$\lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h},$$

exists and is finite.

If f is continuously differentiable at x_0, so for $h \simeq 0$ we have

$$f(x_0 + h) = f(x_0) + hf'(x_0) + \varepsilon(h).$$

This formula (Taylor's Formula) can be generalized to a function g n-times continuously diffenrentiable :

$$f(x_0 + h) = f(x_0) + \sum_{i=1}^{n} \frac{h^{(i)}}{i!} f^{(n)}(x_0) + \varepsilon(h^n).$$

First order derivative

Definition

Let $f : \mathbb{R}^m \to \mathbb{R}^n$ be a C^0 application and $x \in \mathbb{R}^m$. Then f is differentiable at x_0 if it exists $J \in \mathbb{R}^{m \times n}$ such that :

$$\lim_{x \to x_0} \frac{\|f(x) - f(x_0) - Jf(x_0)(x - x_0)\|}{\|x - x_0\|} = 0.$$

D is called the Jacobian of the application f.

For all i, j the elements of the matrix J are given by :

$$J_{ij} f(x_0) = \left.\frac{\partial f_i(x)}{\partial x_j}\right|_{x = x_0}$$

First order derivative

Remark

Usually $f : \mathbb{R}^m \to \mathbb{R}$ so the Jacobian of the application f (also called the gradient) will be a **vector** $\nabla f(x_0)$

The gradient gives the possibilty to approximate the function near the point its gradient is calculated. For all $x \in V(x_0)$ we have

$$f(x) \simeq f(x_0) + \nabla f(x_0)(x - x_0)$$

This affine approximation of the function f will help us to characterize convex functions.

First order derivative : example

Let us consider the application $f : \mathbb{R}^3 \to \mathbb{R}$ defined by

$$f(x, y, z) = 3x^2 + 2xyz + 6z + 5yz + 9xz.$$

We want to calculate the Jacobian of this function. To do so, we need to calculate : $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$. The Jacobian of f at (x, y, z) is given by :

$$J_{f(x,y,z)} = \begin{pmatrix} 6x + 2yz + 9z, & 2xz + 5z & 2xy + 6 + 5y + 9x \end{pmatrix}$$

First order derivative

Exercise

• On other function :
Let $x, y, z \in \mathbb{R}^n$.
Calculate the Jacobian of the function
$f(x, y, z) = exp(xyz) + x^2 + y + log(z)$.

• Linear Regression :
Let $Y \in \mathbb{R}^n, X \in \mathbb{R}^{n\times d}$ and $\beta \in \mathbb{R}^d$.
Calculate the derivative of the function $f(\beta) = \|Y - X\beta\|_2^2$

• Log-Sum-Exp :
Let $x, b \in \mathbb{R}^n$.
Calculate the derivative of the function $f(x) = \log \sum_{i=1}^n \exp(x_i + b_i)$

Correction

• You simply have to apply the definition as it wa done in the previous example and you will have :

$$\nabla f(x,y,z) = \left(yz\exp(xyz) + 2x, xz\exp(xyz) + 1, xy\exp(xyz) + \frac{1}{z} \right).$$

• Here, you have to use the face that : $\|x\|^2 = \langle x, x \rangle$. Then you compute the derivative using the fact that f is defined as a product of to functions of β.

$$\nabla f(\beta) = -X^T(Y - X\beta) + (Y - X\beta)(-X))^T = -2X^T(Y - X\beta).$$

• Remember that the Jacobian $\nabla_f = J_f$ is a vector where each entry i is equal to :

$$\nabla f(x)_i = \frac{\exp(x_i + b_i)}{\sum_{i=1}^n \exp(x_i + b_i)}.$$

Second order derivative

Definition

Let $f : \mathbb{R}^m \to \mathbb{R}$ be a real function. Provided that this function is twice diffentiable, the second derivative H, (also called the Hessian)of f at x_0 is given by :

$$H_{ij}f(x_0) = \left.\frac{\partial^2 f(x)}{\partial x_i \partial x_j}\right|_{x=x_0},$$

and $H \in \mathbb{R}^{m\times m}$

This matrix is usefull to prove that a function f is convex or not and also to build efficient algorithms.

Second order derivative : example

Let us consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f(x,y) = 4x^2 + 6y^2 + 3xy + 2\left(\cos(x) + \sin(y)\right)$$

and calculate the Hessian of this function. We first have to calculate the Jacobian of the matrix and then the Hessian.

$$J_{f(x,y)} = \begin{pmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \end{pmatrix} = \begin{pmatrix} 8x + 3y - 2\sin(x) & 12y + 3x + 2\cos(y) \end{pmatrix}$$

$$H_{f(x,y)} = \begin{pmatrix} \frac{\partial^2 f}{\partial^2 x} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial^2 y} \end{pmatrix} = \begin{pmatrix} 8 - 2\cos(x) & 3 \\ 3 & 12 - 2\sin(y) \end{pmatrix}$$

Second order derivative : example

Exercise

Calculate the second order derivative of the following functions :

- $f(x, y) = \log(x + y) + x^2 + 2y + 4$
- $f(x, y, z) = \dfrac{6x}{1 + y} + \exp(xy) + z$

Correction

The process is similar as in the previous example, so I only give the results.

$$H_{f(x,y)} = \begin{pmatrix} \dfrac{\partial^2 f}{\partial^2 x} & \dfrac{\partial^2 f}{\partial x \partial y} \\ \dfrac{\partial^2 f}{\partial y \partial x} & \dfrac{\partial^2 f}{\partial^2 y} \end{pmatrix} = \begin{pmatrix} 2 - \dfrac{1}{(x+y)^2} & -\dfrac{1}{(x+y)^2} \\ -\dfrac{1}{(x+y)^2} & -\dfrac{1}{(x+y)^2} \end{pmatrix}$$

$$H_{f(x,y)} = \begin{pmatrix} y^2 \exp(xy) & -\dfrac{6}{(1+y)^2} + (xy+1)\exp(xy) & 0 \\ -\dfrac{6}{(1+y)^2} + (xy+1)\exp(xy) & \dfrac{12x}{(1+y)^3} + x^2 \exp(xy) & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

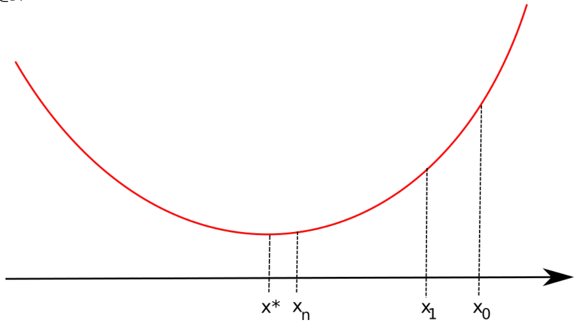
Convexity

What is a convex optimization problem ?

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

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

The aim of the course is to introduce some algorithms to build a serie of reals $(x_n)_{n\in\mathbb{N}}$ which converges to $\hat{x}$.



Optimization

It exists several type of optimization problem :

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

$\rightarrow$ In this part, we'll only focus on convex optimization problem.
As presented before, we need first to study convex functions.

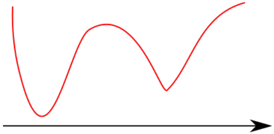
Why do we study them

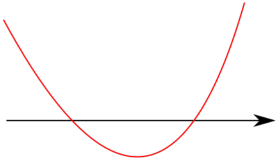
• They are very important in Machine Learning. As we'll see later in the course, their properties guaranties the uniquiness of any convex optimization problem.

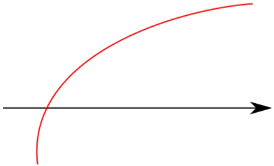
• Studying convex function gives also the possibility to build better algorithm that converge faster than classical methods (**Gradient Descent** vs **Newton's Method**.)

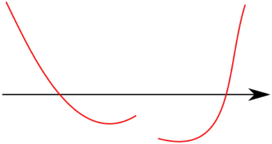
Convex Functions

Which of the following functions are convex graphically ?









Convex Functions

Definition

Let $\mathcal{U}$ be an on empty set of a vectorial 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,
- $f : \mathbb{R} \to \mathbb{R}, \quad f(x) = x^2,$
- $f : \mathbb{R} \to \mathbb{R}, \quad f(x) = \exp(x).$

Convex Functions and line segment

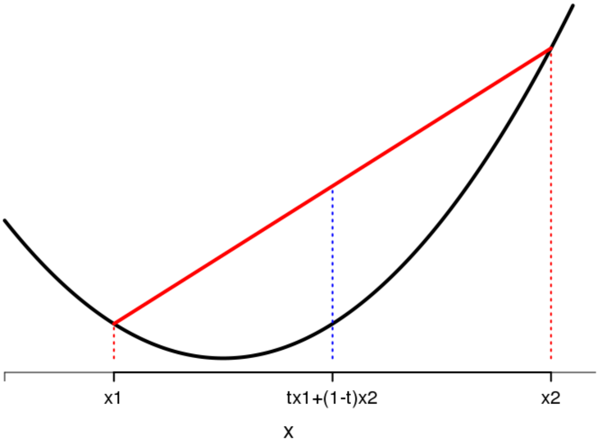
Proposition

A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if any only if the restriction to a line is always convex, i.e. if the function $g : \mathbb{R} \to \mathbb{R}$ defined by $g(t) = f(x + tv)$ is convex for all x and v such that $x + tv$ belongs to the domain of definition of f (f is concave if and only if g is concave).

Exercise

Try to proove this result at home. Hint : you just have to apply (write) the definition of convex function see ealier.

Example : show that the function $f : \mathbb{S}^n_{++} \to \mathbb{S}$ with $f(X) = log(det(X))$ is concave



A convex function and its chord

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 $(tx + (1-t)y)^2 \leq tx^2 + (1-t)y^2$.

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

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

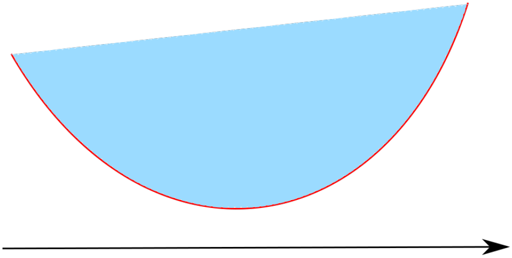
$$\iff t(t-1)x^2 - 2t(t-1)xy + t(t-1)y^2 \leq 0,$$

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

Convex functions

Equivalent definition

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



The epigraph is the blue domain, which is convex. Consider this result as an exercise.

Concavity

Remark

Let $\mathcal{U}$ be an on empty set of a vectorial 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 $h(u) = max\,(f(u), g(u))$ is also convex

Exercise

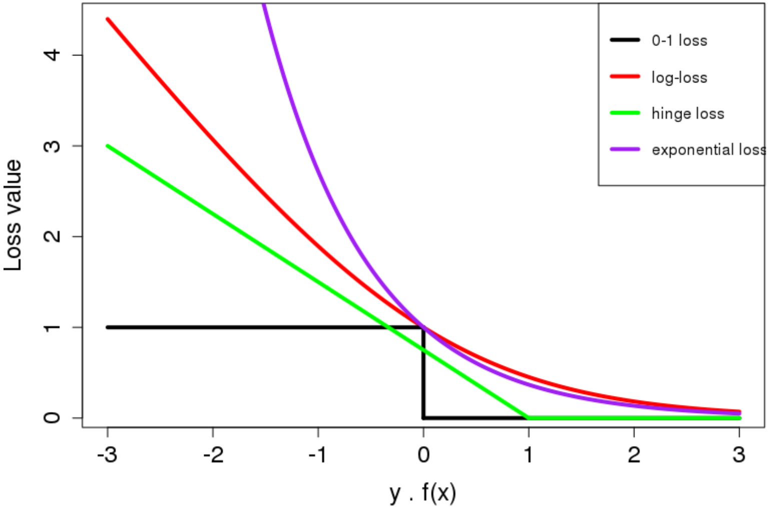
Proove the two first points using the definition of convexity.

Correction

1. For this one, you have to notice that $(f+g)(x) = f(x) + g(x)$ and apply the definition of convexity
2. $$\begin{aligned} g(tx+(1-t)y)) &\leq tg(x)+(1-t)g(y) \\ f(g(tx+(1-t)y))) &\leq f(tg(x)+(1-t)g(y)) \\ f(g(tx+(1-t)y))) &\leq tf(g(x))+(1-t)f(g(y)) \\ f\circ g(tx+(1-t)y) &\leq tf\circ g(x)+(1-t)f\circ g(y) \end{aligned}$$

   On applique successivement (1) la convexité de g puis (2) utilise la croissance de la fonction f et enfin (3) utilise la convexité de la fonction f

Convex Loss Functions



Convexity and differentiability

Proposition

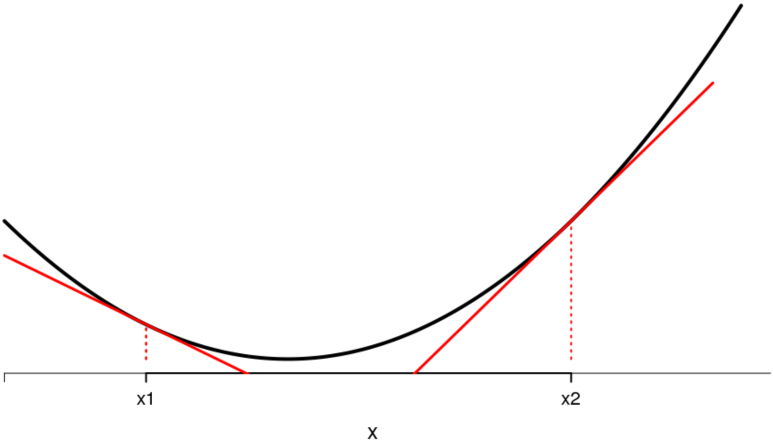
Let f be a continuously differentiable function (C^1) on $\mathcal{U}$. Then f is convex if and only if, for all $(u, v) \in \mathcal{U}$, we have :

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

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

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

Try to proove the previous proposition using the definition of convexity.


x1
x2
x

Convexity and differentiability

Definition

Let f be an application of class C^2 on $\mathcal{U}$ and let H be the matrix of the application $\nabla^2 f$ (The Hessian of f). Then f is said to convex if :

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

$$^t uHu \geq 0.$$

Recall

A matrix H is said to be PSD if and only if all of it's eigenvalues are non-negative.

Convexity and differentiability

Interpretation

Having positive eigenvalues means that the gradient is an increasing function along each directions of the space.
Furthermore, a function is convex if and only if its gradient is an increasing function along each direction.

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 polynom $det(A - XI_2)$).

Convexity and differentiability

We'll show why, for a 2×2 matrix, we have the following equivalence :

$$\text{A is PSD} \iff Tr(A) \geq 0 \textbf{ and } det(A) \geq 0.$$

• We have $det(A - XI_2)) = x^2 - (a+c)x + ad - bc$. The roots of this polynom are exactly the eigenvalues of the matrix A (by definition), so

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

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

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

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

Convexity and differentiability

- ($\Rightarrow$) If the eigenvalues are positive, we immediatly see that both :

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

($\Leftarrow$) Conversly, 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.

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 $i = 1, \dots, n, \quad \lambda_i \geq 0$, then $H = diag(\lambda_i)$ is PSD.
- The function $f : \mathbb{R}^n \to \mathbb{R}^n$ defined by $f(x_1, \dots, x_n) = \sum_{i=1}^n x_i^2$ 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 polynom 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 $z = (z_1, ..., z_n)$.

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

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

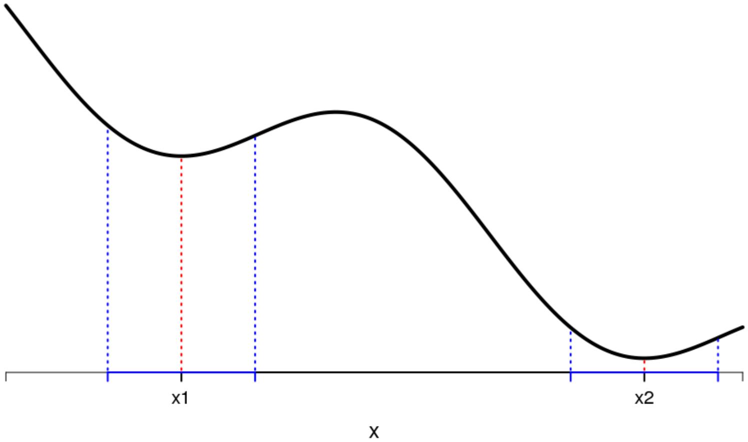
Definition

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 is 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 : $\nabla f(u) = 0$.

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.

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

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 Inequation -

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 ? (without calculous).
  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 :

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

• 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 postive 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 vectorial 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.

• The function f can represent the cost for a company to store a series of products (represented by the parameter x). It can also represent a risk that is taken by making decisions.

• Most of times, we want to minimize the function f under some contraints.

Linear Regression 1/3

Let us first consider the linear regression :

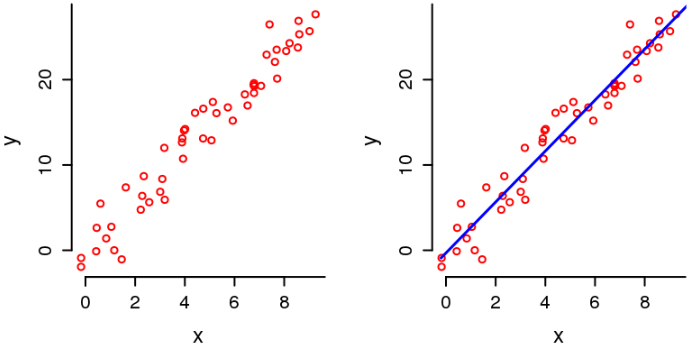
• Given a vector response $Y \in \mathbb{R}^n$ and feature vectors $X = (x_1, \dots, x_n)^T, x_i \in \mathcal{R}^m$ where $m + 1 < n$.
We'd like to find a vector β that explain the value of Y using X with the following model

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

• 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 easilly check that is problem is convex : as we have seen before

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

and

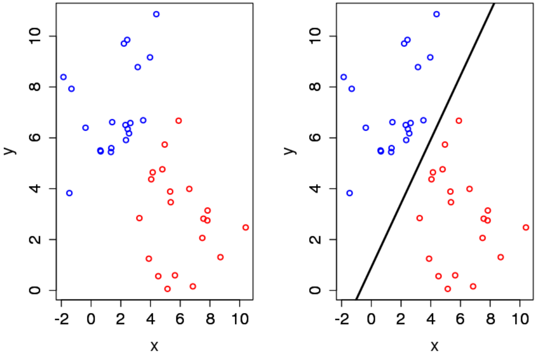
$$\nabla^2_\beta = 2X^TX,$$

which is positive semi definite.
The solution given by $\beta = (X^TX)^{-1}X^TY.$

$\rightarrow$ In this case, an analytical solution exist to solve this problem of minimization. Unfortunately this is not always the case.

Logistic regression 1/2

Let us consider now the logistic regression problem which is quite similar as the previous one.
We want to find a model that predict the class of our data.



→ An example of straight line that separate the two classes using logistic regression

Logistic Regression 2/2

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

$$g(x,a) = \log\left(\frac{\mathbb{P}(X \mid Y=1)}{1-\mathbb{P}(X \mid Y=1)}\right) = a_0 + a_1x_1 + \ldots + a_mx_m.$$

- The parameters of the model are astimated by maximizing the (log-)likelihood of our data.

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

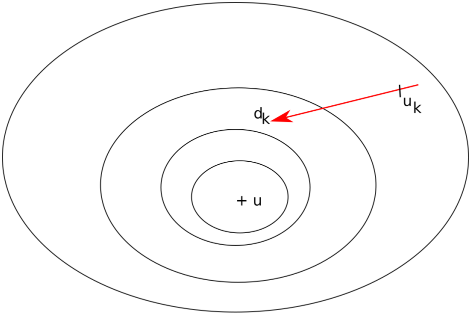
$\rightarrow$ There is no analytical solution to this problem. To solve it, we need a way to approximate it step by step.

Algorithms

Generality

Given a function f and a non empty set $\mathcal{U}$ and knowing there is a solution to the problem : $f(u) = \underset{v \in \mathcal{U}}{Inf}\ f(v)$.

Idea : build a serie $(u_k)_{k\in\mathbb{N}}$ which converges to u.

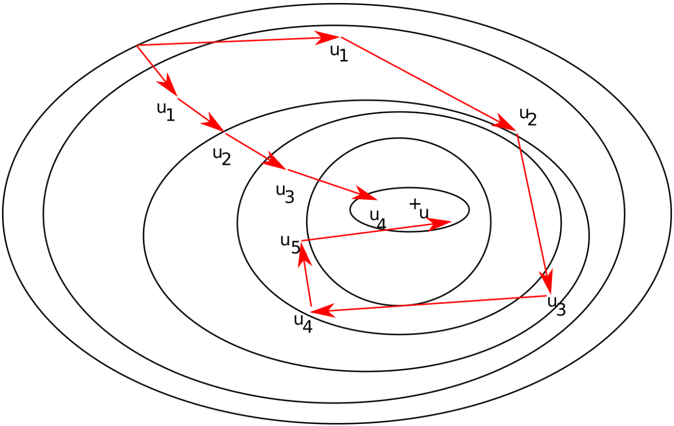


Algorithm :

- Take an initial value u_0.
- $u_k \to u_{k+1}$: Choose a direction d_k and minimize the function f along this direction.
- Solve $\underset{\rho>0}{Argmin}\ f(u_k - \rho d_k) = \rho_k$
- $u_{k+1} = u_k - \rho_k d_k$

Generality : direction of descent

How to choose the direction d_k ?



$\rightarrow$ Some ways seem to be faster than others to reach the solution

Generality : direction of descent

Recall that $f(u_k - \rho d_k) = f(u_k) - \rho\langle \nabla f(u_k), d_k\rangle + \rho\varepsilon(\rho)$ when ρ is closed to 0.
To minimize f we have to choose the direction d_k that maximize the scalar product $\langle \nabla f(u_k), d_k\rangle$ (we suppose, without loss of generality, $\|d_k\| = 1$).
Due to Cauchy-Scwhartz Inequality we have $d_k = \nabla f(u_k)$. So the previous algorithm become :

- Choose u_0 to initialize the algorithm,
- set $u_{k+1} = u_k - \rho_k \nabla f(u_k)$ for $\rho_k > 0$
- till $\|\nabla f(u_k)\| \leq \varepsilon$.

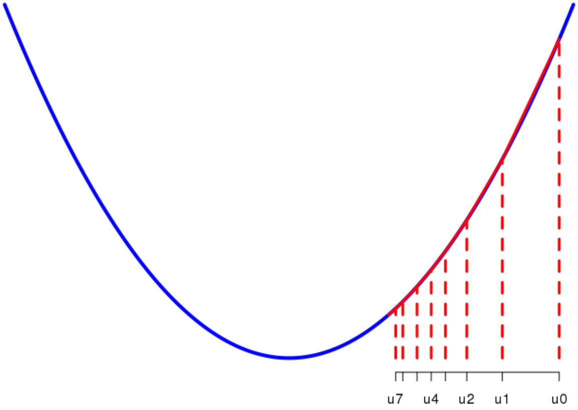
Generality

• It exists several ways to reach the solution using the gradient (or more) of the function we want to optimize.
• We will focus on gradient descent algorithms and their variants. (**Gradient Descent, Line Search, Newton's Method,...**)

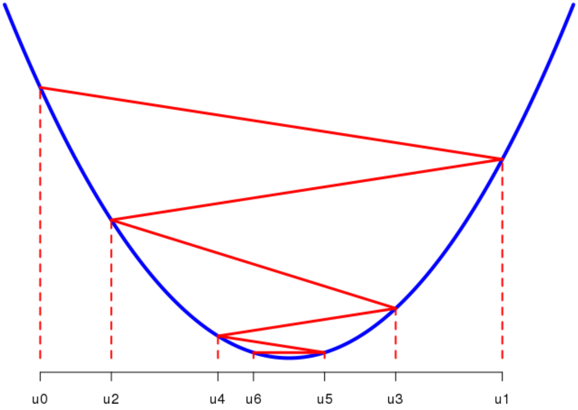
Remark

It exists other algorithms which are able to solve that kind of problem without using the derivatives of the function.

Gradient descent : choose the step 1/3



Gradient descent : choose the step 2/3



Gradient descent : choose the step 3/3

• If the step is too large, the sequence of iterates will oscillates near the global optimum.

• If the step is too small, the algorithm will need a large number of iterations to reach the solution.

$\rightarrow$ We will see different algorithm to choose the step for the gradient descents method.

Gradient descent : with optimal step

A basic idea is, at each iteration, to choose the step in order to minimize the objective function along a given direction. The following is presented bellow :

- Choose u_0 to initialize the algorithm,
- for $k = 0, 1, \dots$ solve $\underset{\rho>0}{Argmin}\, f(u_k - \rho\nabla f(u_k))$,
- set $u_{k+1} = u_k - \rho_k \nabla f(u_k)$
- till $\|\nabla f(u_k)\| \leq \varepsilon$.

This algorithm is called the Gradient Descent with optimal step.

Gradient descent : with optimal step

Definition

Let f be a convex and continuously differentiable function on $\mathbb{R}^n$. We say that f is strongly convex or α-elliptical if it exists $\alpha > 0$ such that

$$\langle \nabla f(v) - \nabla f(u), v - u \rangle \geq \alpha \|v - u\|, \ \forall u, v \in \mathbb{R}^n$$

Proposition

If f is a strongly convex function with respect to the above definition so the Gradient Descent with optimal step converge.

$\rightarrow$ What can we say about $\langle \nabla f(u_{k+1}), \nabla f(u_k) \rangle$ using
$\rho_k = \underset{\rho>0}{Argmin}\, f(u_k - \rho d_k)$?

Gradient descent : with optimal step

If ρ_k minimize $f(u_k - \rho_k d_k)$ we have :

$$\frac{\partial}{\partial \rho} f(u_k - \rho \nabla f(u_k))\Big|_{\rho=\rho_k} = 0,$$

$$\iff \langle \nabla f(u_k - \rho_k \nabla f(u_k), \nabla f(u_k) \rangle = 0,$$

$$\iff \langle \nabla f(u_{k+1}), \nabla f(u_k) \rangle = 0.$$

The last equality is called the optimality condition.

Let us show how it works on an example.

Gradient descent : with optimal step

Let A be a symetric and positive definite matrix (so all of it's eigenvalues are positive and $A^T = A$) and $b \in \mathbb{R}^n$. We want to write the previous algorithm at the iteration $k > 0$ for the function f defined by

$$f(v) = \frac{1}{2}\langle Av, v\rangle - \langle b, v\rangle$$

- Calculate the gradient : $\nabla f(u_k) = Au_k - b$
- We then have to solve : $\rho_k = \underset{\rho>0}{Argmin}\, f(u_k - \rho d_k)$. The optimality condition gives us : $\langle \nabla f(u_k), \nabla f(u_{k+1})\rangle = 0$

$$\begin{aligned} \nabla f(u_{k+1}) &= Au_{k+1} - b \\ &= A(u_k - \rho_k(Au_k - b) - b \\ &= Au_k - b - \rho_k A(Au_k - b) \end{aligned}$$

Gradient descent : with optimal step

$$
\begin{aligned}
\Rightarrow \langle Au_k - b, Au_k - b - \rho_k A(Au_k - b)\rangle &= 0 \\
\Rightarrow \langle Au_k - b, Au_k - b\rangle &= \langle Au_k - b, \rho_k A(Au_k - b)\rangle \\
\Rightarrow \rho_k &= \frac{\langle Au_k - b, Au_k - b\rangle}{\langle Au_k - b, A(Au_k - b)\rangle}
\end{aligned}
$$

We finally have the following algorithm :

- Initialize $u_0 \in \mathbb{R}^n$
- At each step, calculate $\rho_k = \dfrac{\|Au_k - b\|^2}{\|Au_k - b\|_A^2}$
- Set $u_{k+1} = u_k - \rho_k(Au_k - b)$
- Stop if $\|\nabla J(u_{k+1})\| = \|Au_{k+1} - b\| \leq \epsilon$

Gradient descent : with optimal step

Exercise

Consider the matrices $A = \begin{pmatrix} 6 & 2 \\ 2 & 4 \end{pmatrix}$ and $b = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$ and the application f defined by $f(v) = \langle Av, v \rangle + \langle b, v \rangle$

1. Explain why f in convex.
2. Solve the problem $u = \underset{v \in \mathbb{R}^2}{Argmin}\, f(v)$.
3. For a given vector u_k, calculate ∇f_{u_k} and ρ_k.
4. Implement the presented method to solve this problem.

- f is defined as a quadratic function, so f is convex.
- We have to solve :

$$J_{f(x,y)} = \begin{pmatrix} 12x + 4y + 2, & 4x + 8y + 3 \end{pmatrix} = (0, 0).$$

The solution is $(\frac{-1}{20}, \frac{-7}{20})$.

- Let set $u_k = (v_1, v_2)$ then :

$$\nabla f_{u_k} = \begin{pmatrix} 12v_1 + 4v_2 + 2, & 4v_1 + 8v_2 + 3 \end{pmatrix},$$

and $\rho_k = \dfrac{\|2Au_k - b\|_2^2}{\|2Au_k - b\|_{2A}}$

Exercise

Let f be the function defined by : $f(x, y) = 4x^2 - 4xy + 2y^2$.

1. Is the function f convex ?
2. Apply the gradient descent with optimal step to calculate the three first step of the algorithm using $(x_0, y_0) = (1, 1)$.

Correction 1/2

- The function f can be rewritten as : $f(u) = \dfrac{1}{2}u^T Au - b^T u$, where $b = (0,0)^T$ and $A = \begin{pmatrix} 8 & -4 \\ -4 & 4 \end{pmatrix}$. The function f is a quadratic function, furthermore the matrix A is PSD so the function f is convex.
- The optimal learning rate is given by :

$$\rho_k \frac{\|Au_k - b\|_2^2}{\|Au_k - b\|_A^2},$$

where the matrix A and the vector b were previously introduced.

Correction 1/2

- The function f can be rewritten as : $f(u) = \frac{1}{2}u^T Au - b^T u$, where $b = (0\ 0)^T$ and $A = \begin{pmatrix} 8 & -4 \\ -4 & 4 \end{pmatrix}$. The function f is a quadratic function, furthermore the matrix A is PSD so the function f is convex.
- The optimal learning rate is given by :

$$\rho_k = \frac{\|Au_k - b\|_2^2}{\|Au_k - b\|_A^2},$$

where the matrix A and the vector b were previously introduced.
Recall that the process is defined by :

$$u_{k+1} = u_k - \rho_k \nabla f(u_k).$$

We will now apply this process to compute the three first iterations.

Correction 2/2

1. For the first iteration : $\rho_0 = \frac{\|Au_0\|_2^2}{\|Au_0\|_A^2} = \frac{16}{128} = \frac{1}{8}$. And $\nabla f(u_0) = Au_0 = (4\ 0)^T$.

$$u_1 = (1\ 1)^T - \frac{1}{8}(4\ 0)^T = (0.5\ 1)^T.$$

2. For the second iteration : $\nabla f(u_1) = Au_1 = (0\ 2)^T$. The learning rate is given by : $\rho_1 = \frac{\|Au_1\|_2^2}{\|Au_1\|_A^2} = \frac{4}{16} = \frac{1}{4}$. Thus u_2 is given by :

$$u_2 = (0.5\ 1)^T - \frac{1}{4}(0\ 2)^T = (0.5\ 0.5)^T.$$

3. For the third iteration : $\nabla f(u_2) = Au_2 = (2\ 0)^T$. The learning rate is given by : $\rho_2 = \frac{\|Au_2\|_2^2}{\|Au_2\|_A^2} = \frac{4}{32} = \frac{1}{8}$. Thus u_3 is given by :

$$u_3 = (0.5\ 0.5)^T - \frac{1}{8}(2\ 0)^T = (0.25\ 0.5)^T.$$

Gradient Descent : Armijo Criterium

An other method to find the best **learning rate** is to use a linear search process. The idea is, given a $\theta \in]0,1[$, choose the greatest ρ such that :

$$f(u_k - \rho \nabla f(u_k)) \leq f(u_k) - \theta\rho\|\nabla f(u_k)\|^2.$$

It means : at each step, we reduce the function's value of at least $\theta\|\nabla f(u_k)\|$.

Armijo's condition :

- Choose $\alpha_0 > 0$ and $0 < \theta < 1$,
- Choose the greatest $s \in \mathbb{Z}$ such that :

$$f(u_k - \alpha_0 2^s \nabla f(u_k)) \leq f(u_k) - 2^s \alpha_0 \theta \|\nabla f(u_k)\|^2.$$

- Set $u_{k+1} \leftarrow u_k - \alpha_0 2^s \nabla f(u_k)$.

Gradient Descent : Armijo Criterium and Wolfe's Criteria

Theorem

If the function f is strictly convex and if its gradient ∇f is Lipschitz, then the Armijo's algorithm converge.

If we add the following condition to the previous one, given $0 < \theta < \eta < 1$:

$$\langle \nabla f(u_k), \nabla f(u_k - \rho \nabla f(u_k)) \rangle \geq \eta \|\nabla f(u_k)\|^2,$$

we get the Wolfe's Criteria

Conjugate Gradient

Definition

Let A be a symetric positive and definite matrix and u, v two vectors.
u, v are conjugate with respect to A if $\langle Au, v\rangle = 0$.
If $A = I$ we say that they are orthogonal.

Let A be a symetric positive and definite matrix and f the function defined by

$$f(v) = \frac{1}{2}\langle Av, v\rangle - \langle b, v\rangle.$$

$\rightarrow$ The objective of this algorithm is to build a serie of conjugate descent direction. Let us see how it is built in pratice.

Conjugate Gradient

• Let $u_0 \in \mathbb{R}^n$. We define a first direction of descent $d_0 = \nabla f(u_0)$ for instance and we minimize f along this direction, i.e we solve :

$$\underset{\alpha_0}{Argmin}\, f(u_0 - \alpha_0 d_0).$$

Solving this problem we get :

$$\alpha_0 = \frac{\langle \nabla f(u_0), d_0 \rangle}{\langle Ad_0, d_0, \rangle}.$$

And we set $u_1 = u_0 - \alpha_0 d_0$
•To build $d_1 = \nabla f(u_1) + \beta_0 d_0$ we have to calculate the value of $\beta_0 \in \mathbb{R}$ such that

$$\langle Ad_1, d_0 \rangle = 0.$$

Conjugate Gradient

• We then have to solve $\langle A\nabla f(u_1), d_0\rangle + \langle A\beta_0 d_0, d_0\rangle = 0$. The solution is given by

$$\beta_0 = -\frac{\langle A\nabla f(u_1), d_0\rangle}{\langle Ad_0, d_0\rangle}.$$

Once it's done, you'll do as before.

You set $\alpha_1 = \underset{\alpha}{Argmin}\, f(u_1 - \alpha d_1)$.

Set $u_2 = u_1 - \alpha_1 d_1$. And so on ...

Conjugate Gradient : Summary

Algorithm :

- Choose $u_0 \in \mathbb{R}^n$ and $d_0 = \nabla f(u_0)$.
- Set $\alpha_0 = \dfrac{\langle \nabla f(u_0), d_0 \rangle}{\langle A d_0, d_0, \rangle}$ and $u_1 = u_0 - \alpha_0 d_0$.
- $\beta_0 = -\dfrac{\langle A \nabla f(u_1), d_0 \rangle}{\langle A d_0, d_0 \rangle}$.

For $k \geq 1$ do,

- Set $d_k = \nabla f(u_k) + \beta_{k-1} d_{k-1}$.
- Set $\alpha_k = \dfrac{\langle \nabla f(u_k), d_k \rangle}{\langle A d_k, d_k, \rangle}$ and $u_{k+1} = u_k - \alpha_k d_k$.
- Set $\beta_k = \dfrac{\langle A \nabla f(u_{k+1}), d_k \rangle}{\langle A d_k, d_k \rangle}$

Untill $\|\nabla f(u_{k+1})\| \leq \varepsilon$.

Conjugate Gradient : Results

Proposition

For all $1 \leq k \leq n$ such that $\nabla f(u_0), \dots, \nabla f(u_n)$ are non equal to zero, we have the following relations for all $0 \leq l \leq k-1$:

$$\langle \nabla f(u_k), \nabla f(u_l) \rangle = 0$$

and

$$\langle Ad_k, d_l \rangle = 0.$$

Theorem

If A is a symetric positive and definite matrix, then the conjugate gradient method converges with at most n steps.

$\rightarrow$ You can try to prove the proposition by induction. (At home)

Newton's Method

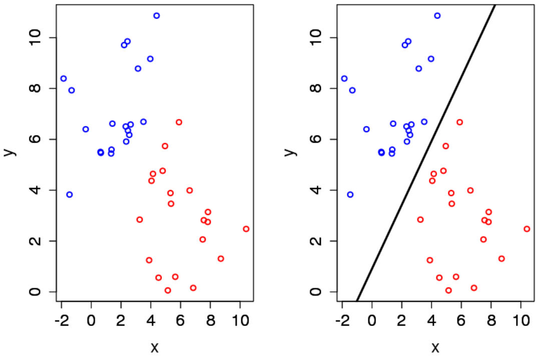
The Newton's Method is also an algorithm of gradient descent. It uses the second derivative to refine the direction of the descent as follow :

$$u_{k+1} \leftarrow u_k - \left(\nabla^2 f(u_k)\right)^{-1} \cdot \nabla f(u_k).$$

- It requires less iteration to converge to the solution compared to the other gradient descent methods.
- Harder to implement, requires the inverse of the Hessian of the function we want to optimize ($\Theta(n^3)$).
- The Hessian is not always invertible at a given point.

Newton's Method

Let's come back to the logistic regression.
We want to find a model that predict the class of our data.



→ An example of straight line that separate the two classes using logistic regression.

Newton's Method

In the case of Logistic Regression, we want to maximize the Log-Likelihood of our data, noted $l(x, a)$ where x refers to the data and a the parameters of our model.

Remember that a **possible** solution is given by solving the equation :

$$\nabla_a l(x, a) = \nabla_a \left(\sum_{i=1}^{n} y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \right) = 0,$$

where $p = \left(1 + \exp(-a^T x)\right)^{-1}$.

$\rightarrow$ Explain why the log-likelihood is concave. Calculate the first and second derivatives of the function l.

Newton's Method

If we apply the Newton's Method to the logistic regression we have

$$\nabla_a l(x,a) = \sum_{i=1}^{n} (y_i - p_i)\, x_i, \quad \nabla_a^2 l(x,a) = -\sum_{i=1}^{n} p_i(1-p_i) x_i x_i^T$$

We can then write the algorithm :

- Choose a_0,
- Calculate $\nabla_a l(x,a)$ and $\left(\nabla_a^2 l(x,a)\right)^{-1}$
- Set $a_{k+1} \leftarrow a_k - \left(\nabla_a^2 l(x,a)\right)^{-1} \nabla_a l(x,a)$
- Stop when $\|\nabla_a l(x,a)\| \leq \varepsilon$.

Quasi-Newton's Method : Motivation

The main drawback of the Newton's Method is calculous of the inverse of the Hessian matrix H_k^{-1}. To avoid it, an other process is proposed as follow

$$
\begin{aligned}
u_{k+1} &= u_k - M_k \nabla f(u_k), \\
M_{k+1} &= M_k + C_k.
\end{aligned}
$$

The idea is to approximate the H_k^{-1} by matrix M_k at which, we add a matrix of correction C_k at each step.

Quasi-Newton's Method : Motivation

Recall that :

$$\nabla f(u_k) = \nabla f(u_{k+1} + (u_k - u_{k+1})) \sim \nabla f(u_{k+1}) + \nabla^2 f(u_{k+1})(u_k - u_{k+1}),$$

we then have :

$$\left(\nabla^2 f(u_{k+1})\right)^{-1} (\nabla f(u_{k+1}) - \nabla f(u_k)) \sim u_{k+1} - u_k.$$

If we set : $M_{k+1} = \left(\nabla^2 f(u_{k+1})\right)^{-1}$, $\gamma_k = \nabla f(u_{k+1}) - \nabla f(u_k)$ and $\delta_k = u_{k+1} - u_k$, we get the Quasi Newton's Condition :

$$\boxed{M_{k+1}\gamma_k = \delta_k}$$

Quasi-Newton's Method : Davidon-Fletcher-Powell

• The matrix of correction C_k is supposed to be of rank 1. So we can rewrite C_k as $v_k v_k^T$ where $v_k \in \mathbb{R}^n$.
• The update become : $M_{k+1} = M_k + v_k v_k^T$.
• The Quasi Newton's Condition gives :

$$\begin{aligned}
(M_k + v_k v_k^T)\gamma_k &= \delta_k, \\
M_k\gamma_k + v_k v_k^T \gamma_k &= \delta_k, \\
v_k v_k^T \gamma_k &= \delta_k - M_k\gamma_k, \\
v_k &= \frac{\delta_k - M_k\gamma_k}{v_k^T \gamma_k}.
\end{aligned}$$

And the second line gives us : $v_k^T \gamma_k = (\gamma_k\delta_k - \gamma_k M_k \gamma_k)^{1/2}$.

Quasi-Newton's Method : Broyden Algorithm

Broyden Algorithm

Algorithm

- Initialize $u_0 \in \mathbb{R}^n$ and M_0 (usually $M_0 = Id$),
- for $k \geq 0$ do
  - set $\rho_k = \underset{\rho \in \mathbb{R}}{Argmin}\, f(u_k - \rho M_k \nabla f(u_k))$,
  - set $u_{k+1} = u_k - \rho_k M_k \nabla f(u_k)$,
  - set $M_{k+1} = M_k + \dfrac{(\delta_k - M_k\gamma_k)(\delta_k - M_k\gamma_k)^T}{(\delta_k - M_k\gamma_k)^T\gamma_k}$,

  Untill $\|\nabla f(u_{k+1})\| \leq \varepsilon$.

Quasi-Newton's Method : BFGS

This method was proposed by Broyden-Fletcher-Goldfarb-Shanno.
Instead of taking a matrix of correction C_k of rank 1. They suppose this matrix is of rank 2.
The inverse of the Hessian, at each step, is then approximate by :

$$M_{k+1} = M_k + \left[1 + \frac{\langle M_k \gamma_k, \gamma_k \rangle}{\langle \delta_k, \gamma_k \rangle}\right] \frac{\delta_k \delta_k^T}{\langle \delta_k, \gamma_k \rangle} - \frac{\langle \delta_k, \gamma_k \rangle M_k + M_k \gamma_k \delta_k^T}{\langle \delta_k, \gamma_k \rangle}.$$

The algorithm is the same as the previous one.

Conclusion

- **Gradient descent with a constant learning rate :** Easy to implement and the algorithm works fastly (time). Its convergence depends on the value of the learning rate. It can take several iterations to converge (if the learning rate is to small).
- **Gradient descent with an optimal step :** Need less iterations to converge, but an iteration takes more time to be done (time to calculate the optimal learning rate). Very fast for quadratic function !
- **Newton's Method :** Faster than the two others. Requires less iterations. But we need to compute the Hessian matrix and to invert it (which is not always possible). In practice we use in low dimension problem and to solve a logistic regression problem.

To go further

It exists several other algorithms for convex optimization which consists of a variant the gradient descent one. One of them is the **Adam Gradient Descent** which is commonly used in neural networks. Indeed, this last seems to be really fast in practice.
You can also found other algorithms for **constrained convex optimization** to solve problem as follows :

$$\begin{array}{ll} \underset{x \in \mathbb{R}^n}{min} & f(x) \\ s.t. & 0 \leq x_i \leq b_i, \forall i = 1, ..., n, \end{array}$$

using **Projected gradient descent** or Uzawa Algorithm, but for this last, you need the notion of duality that you will see later in this course.