Numerical methods for optimization: support class lectures

Stefano Massei stefano.massei@sns.it

January 4, 2016

Lecture 1: Calculus

Qualitative study of a function in one variable

Suppose that we want to draw the graph of $f : \mathbb{R} \to \mathbb{R}$. In order to get enough information on the function we follow these steps.

(i) Find the domain of f i.e. the set $D \subseteq \mathbb{R}$ where the function is defined.

(ii) Study the sign of f and find its intersection with the axes whether it is possible.

(iii) Study the sign of f'. Where the derivative is positive the function is increasing, instead in the intervals where is negative the function is decreasing.

(iv) Look at the points which verify $f'(x) = 0$ and determine if they are local or global extrema points.

(v) Study the sign of f''. Where the second derivative is positive the function is convex, instead in the intervals where is negative the function is concave.

(vi) Study the behaviour of f at the extremal points of D.

We ignore step (v) because the notion of convexity will be introduced and studied in lecture 3.

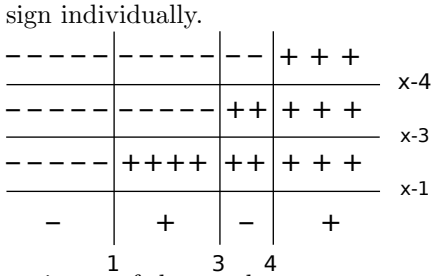
Exercise 1. *Draw the graph of* $f(x) = \frac{x-4}{x^2-4x+3}$.

Observe that since f is defined by a ratio the domain D coincide with the real numbers which does not vanish the denominator. Performing some simple computations we get

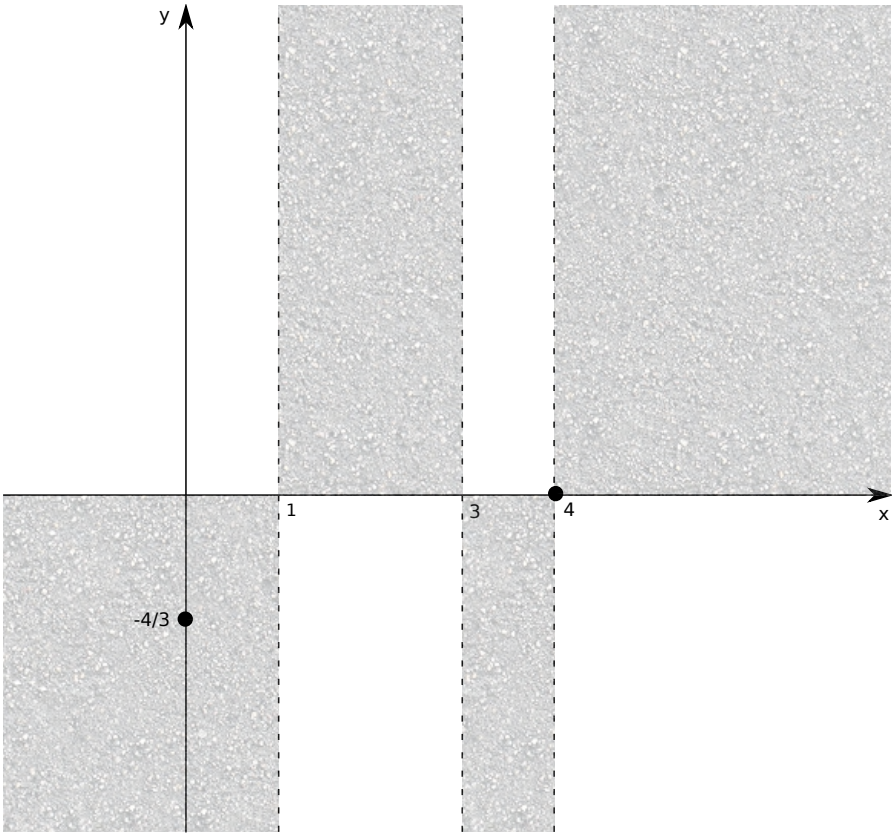
$$x^2 - 4x + 3, \qquad \Delta = 4 \quad \Rightarrow \quad x^2 - 4x + 3 = (x-1)(x-3),$$

therefore $D = \{x \in \mathbb{R} : \ x^2 - 2x + 3 \neq 0\} = \mathbb{R} \setminus \{1, 3\}$. For studying the sign of $f(x)$ it is crucial to write it as a product of factor for which we can study the sign individually.

In our case we have $f(x) = \frac{x-4}{(x-3)(x-1)}$, observing that $x - 4 > 0 \Leftrightarrow x > 4$, $x - 3 > 0 \Leftrightarrow x > 3$, $x - 1 > 0 \Leftrightarrow x > 1$ and using a graphical subdivision of the real line we easily derive that $f(x) > 0$ on $(1, 3) \cup (4, +\infty)$ and $f(x) < 0$ on $(-\infty, 1) \cup (3, 4)$. Moreover $f(0) = -\frac{4}{3}$ and since the numerator is $x-4$ we have $f(x) = 0 \Leftrightarrow x = 4$.



With the results obtained until now we can draw a rough estimate of the graph.



For what concerns the first derivative we have

$$f'(x) = \frac{x^2 - 4x + 3 - (x-4)(2x-4)}{(x^2-4x+3)^2} = \frac{-x^2 + 8x - 13}{(x^2-4x+3)^2} = -\frac{(x-4-\sqrt{3})(x-4+\sqrt{3})}{(x^2-4x+3)^2}.$$

We can use again the graphical representation of the real line for studying the sign of f' getting that

$$\begin{aligned} f' > 0 \quad &\Leftrightarrow \quad x \neq 3 \text{ and } 4-\sqrt{3} < x < 4+\sqrt{3}, \\ f' < 0 \quad &\Leftrightarrow \quad x \neq 1 \text{ and } (x < 4-\sqrt{3} \text{ or } x > 4+\sqrt{3}). \end{aligned}$$

Moreover $f' = 0 \quad \Leftrightarrow \quad x = 4 \pm \sqrt{3}$. Consider $x = 4-\sqrt{3}$, we have that in a left neighobourhood of this point the derivative is negative while in a right neighobourhood is positive. This argument implies that $x = 4-\sqrt{3}$ is a local minimum point and $f(4-\sqrt{3}) = 1 + \frac{\sqrt{3}}{2}$ is a local minimum for f. Analogously one can prove that $x = 4+\sqrt{3}$ is a local maximum point and $f(4+\sqrt{3}) = 1 - \frac{\sqrt{3}}{2}$ is a local maximum for f.

To conclude we look at the behaviour of f as x tends to the extremal points of D. That is we compute the limits:

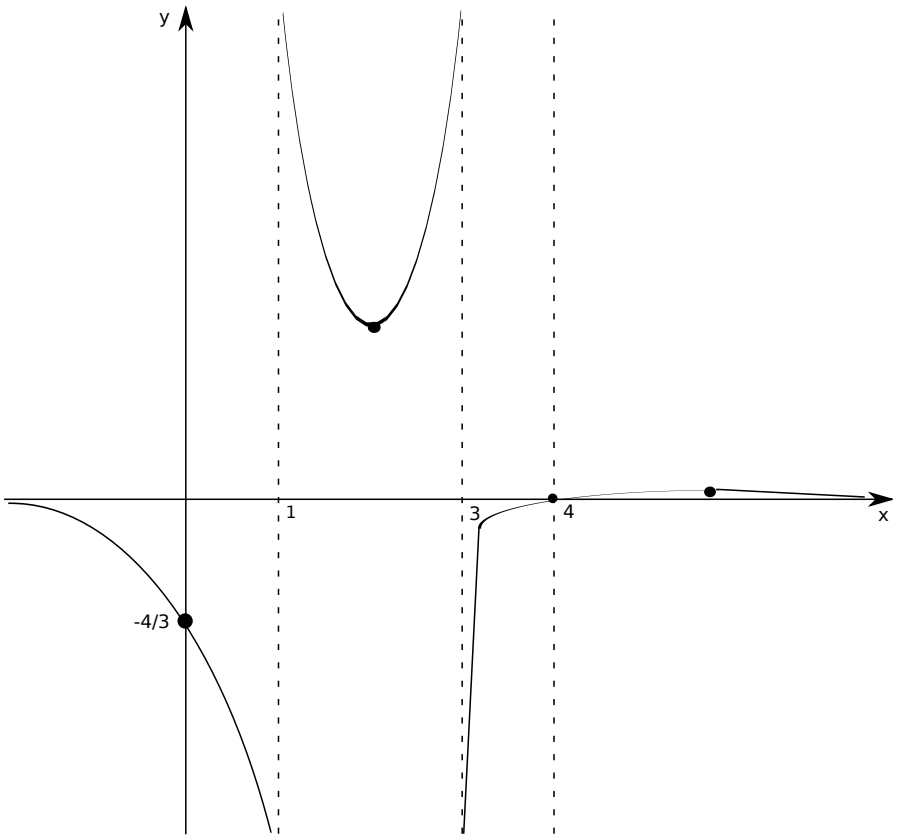
$$\lim_{x\to\pm\infty} \frac{x-4}{x^2-4x+3} = \lim_{x\to\pm\infty} \frac{\frac{1}{x}-\frac{4}{x^2}}{1-\frac{4}{x}+\frac{3}{x^2}} = 0,$$

$$\lim_{x\to 1^-} \frac{x-4}{x^2-4x+3} = \lim_{x\to 1^-} \frac{x-4}{\underbrace{(x-1)}_{0^-}(x-3)} = -\infty,$$

$$\lim_{x\to 1^+} \frac{x-4}{x^2-4x+3} = \lim_{x\to 1^+} \frac{x-4}{\underbrace{(x-1)}_{0^+}(x-3)} = +\infty,$$

$$\lim_{x\to 3^-} \frac{x-4}{x^2-4x+3} = \lim_{x\to 3^-} \frac{x-4}{(x-1)\underbrace{(x-3)}_{0^-}} = +\infty,$$

$$\lim_{x\to 3^+} \frac{x-4}{x^2-4x+3} = \lim_{x\to 3^+} \frac{x-4}{(x-1)\underbrace{(x-3)}_{0^+}} = -\infty.$$



Exercise 2. *Draw the graph of* $f(x) = \frac{1}{1-2^{\frac{1}{x}-1}}$.

In the definition of f there are two ratio so we need to impose the respective denominators to not vanish:

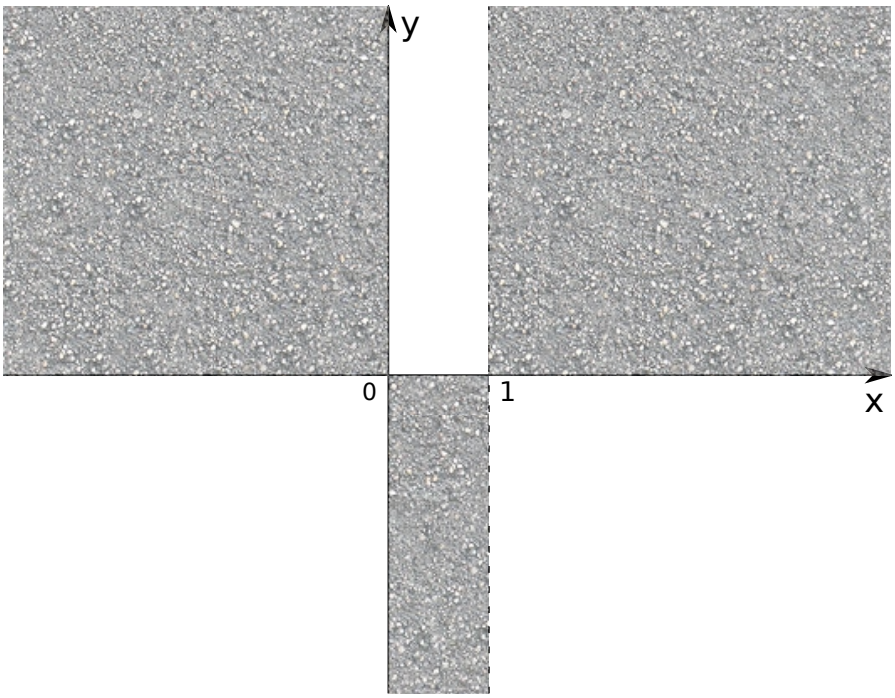
$$\begin{cases} x \neq 0 \\ 1 - 2^{\frac{1}{x}-1} \neq 0 \end{cases} \Leftrightarrow \begin{cases} x \neq 0 \\ 1 - e^{(\frac{1}{x}-1)(\log 2)} \neq 0 \end{cases} \Leftrightarrow \begin{cases} x \neq 0 \\ (\frac{1}{x}-1)(\log 2) \neq 0 \end{cases} \Leftrightarrow \begin{cases} x \neq 0 \\ x \neq 1 \end{cases}.$$

So we have $D = \mathbb{R} \setminus \{0, 1\}$. The sign of f is completely determined by the sign of the factor $1 - 2^{\frac{1}{x}-1}$:

$$1 - 2^{\frac{1}{x}-1} > 0 \Leftrightarrow \frac{1}{x} - 1 > 0 \Leftrightarrow \begin{cases} x > 0 \\ x > 1 \end{cases} \quad \text{or} \quad \begin{cases} x < 0 \\ x < 1 \end{cases} \quad \Leftrightarrow x \in (-\infty, 0) \cup (1, \infty),$$

$$1 - 2^{\frac{1}{x}-1} < 0 \Leftrightarrow x \in (0, 1).$$

Moreover since f is defined by a ratio with a (nonzero) constant numerator we have that $f(x) = 0$ has no solution therefore there are not intersection with the axes.



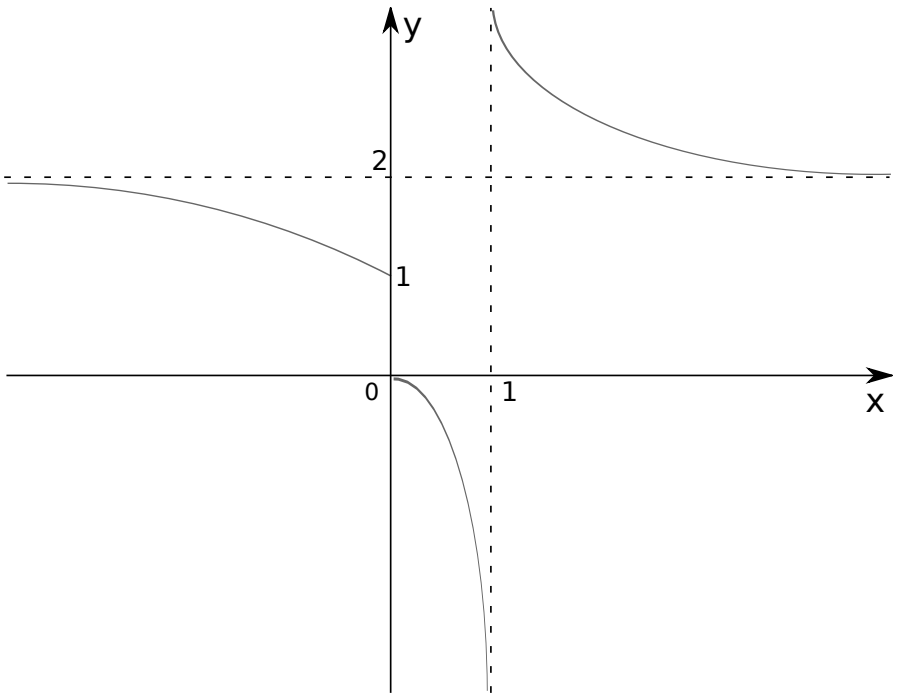
For what concerns the first derivative we have

$$f'(x) = \left(\frac{1}{1 - e^{(\frac{1}{x}-1)\log(2)}}\right)' = \frac{-\frac{1}{x^2}\log(2))e^{(\frac{1}{x}-1)\log(2)}}{(1 - e^{(\frac{1}{x}-1)\log(2)})^2}$$

which is a negative quantity for every $x \in D$. This means that f is decreasing in every interval contained in D and there are not local minima or maxima. To conclude we check the behaviour

of the function close to the extremal points of D.

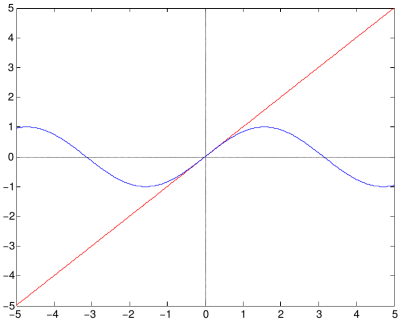
$$\lim_{x \Rightarrow \pm\infty} \frac{1}{1 - e^{(\frac{1}{x} - 1)\log(2)}} = 2,$$
$$\lim_{x \Rightarrow 0^-} \frac{1}{1 - e^{(\frac{1}{x} - 1)\log(2)}} = 1,$$
$$\lim_{x \Rightarrow 0^+} \frac{1}{1 - e^{(\frac{1}{x} - 1)\log(2)}} = 0,$$
$$\lim_{x \Rightarrow 1^-} \frac{1}{1 - e^{(\frac{1}{x} - 1)\log(2)}} = -\infty,$$
$$\lim_{x \Rightarrow 1^+} \frac{1}{1 - e^{(\frac{1}{x} - 1)\log(2)}} = +\infty.$$



One can use these techniques for proving classical inequalities.

Exercise 3. *Prove that* $\sin(x) \leq x \ \forall x \in \mathbb{R}^+$ *and* $\sin(x) \geq x \ \forall x \in \mathbb{R}^-$.

Consider the function $f(x) := \sin(x) - x$. We have that $f(0) = 0$ and $f'(x) = \cos(x) - 1$, in particular $f'(x) \leq 0 \ \forall x \in \mathbb{R}$. Then $f(x)$ is decreasing on $\mathbb{R}$ so we can conclude that $f(x) \leq 0$ on $[0, +\infty)$ and $f(x) \geq 0$ on $(-\infty, 0]$. Replacing the definition of f in the previous inequalities we get the thesis.



Homework 4. *Prove the following inequalities:*

- $\cos(x) \geq 1 - \frac{x^2}{2}\ \forall x \in \mathbb{R}$
- $e^x \geq 1 + x\ \forall x \in \mathbb{R}$
- $\log(1+x) \leq x\ \forall x \geq -1$
- $\arctan(x) \leq x\ \forall x \geq 0$

Taylor expansion

Consider $f : \mathbb{R} \to \mathbb{R}$ and $x_0 \in \mathbb{R}$. We want to approximate f close to x_0 with a polynomial of degree k. If the function is sufficiently regular, precisely if admits derivatives until order k, we can use the Taylor polynomial of order k in x_0:

$$P_{k,x_0}(x) := f(x_0)+f'(x_0)(x-x_0)+\frac{f''(x_0)}{2}(x-x_0)^2+\dots\frac{f^{(k)}(x_0)}{k!}(x-x_0)^k = \sum_{i=0}^{k}\frac{f^{(i)}(x_0)}{i!}(x-x_0)^i.$$

That polynomial is such that if we consider the residual function $R_{k,x_0} := f(x) - P_{k,x_0}$ then it verifies

$$\lim_{x\to x_0}\frac{R_{k,x_0}}{(x-x_0)^k} = 0,$$

which means that going closer to the point x_0 the error goes to 0 faster than $(x-x_0)^k$. If moreover f admits the derivative of order $k+1$ we can explicitly write the residual function with the Lagrange formula:

$$R_{k,x_0} = \frac{f^{(k+1)}(\tau)}{(k+1)!}(x-x_0)^{k+1}$$

where τ is an unknown point which belongs to the segment between x and x_0. This formula allow us to get numerical estimate of the error we make approximating a function using its Taylor polynomial of a certain order.

Example 5. *Let $f(x) = e^x$ and $x_0 = 0$. Then we have*

$$P_{2,x_0}(x) = 1 + x + \frac{x^2}{2}, \qquad R_{2,x_0} = \frac{e^\tau}{6}x^3.$$

If restrict ourself to consider $x \in [-1,1]$ we can claim that in this interval $|R_{2,x_0}| \leq \frac{e}{6}$. In the general case if $x \in [-a,a]$ we can claim that $|R_{2,x_0}| \leq \frac{e}{6}a^3$. When $a < 1$ that estimate can be close to 0.

Example 6. *Let $f(x) = sin(x)$ and $x_0 = \frac{\pi}{2}$ then*

$$P_{2,x_0}(x) = 1 - \frac{(x-\frac{\pi}{2})^2}{2}, \qquad R_{2,x_0}(x) = -\frac{\cos(\tau)}{6}(x-\frac{\pi}{2})^3.$$

Example 7. *Let $f(x) = \log(\cos(x))$ and $x_0 = 0$ then*

$$P_{2,x_0}(x) = -\frac{x}{2}, \qquad R_{2,x_0}(x) = -\frac{\sin(2\tau)}{3\cos^4(\tau)}x^3.$$

This approach can be generalized to functions of several variables. In the case of two variables the Taylor polynomial of the second order in (x_0, y_0) has this expression:

$$P_{1,(x_0,y_0)}(x,y) := f(x_0,y_0) + \nabla f(x_0,y_0)^t \begin{bmatrix} x-x_0 \\ y-y_0 \end{bmatrix}.$$

The residual function $R_{1,(x_0,y_0)}(x,y) := f(x,y) - P_{1,(x_0,y_0)}(x,y)$ has the property

$$\lim_{(x,y)\to(x_0,y_0)} \frac{R_{1,(x_0,y_0)}(x,y)}{\|(x-x_0,y-y_0)\|} = 0.$$

Example 8. *Let* $f(x,y) = y^x = e^{x\log(y)}$ *then*

$$\nabla f(x,y) = \begin{bmatrix} \log(y)e^{x\log(y)} \\ \frac{x}{y}e^{x\log(y)} \end{bmatrix}$$

$$P_{1,(x_0,y_0)}(x,y) = y_0^{x_0} + \log(y_0)e^{x_0\log(y_0)}(x-x_0) + \frac{x_0}{y_0}e^{x_0\log(y_0)}(y-y_0).$$

Example 9. *Let* $f(x,y) = \frac{xy}{x^2+y^2}$ *then*

$$\nabla f(x,y) = \frac{1}{(x^2+y^2)^2}\begin{bmatrix} y^3-x^2y \\ x^3-xy^2 \end{bmatrix}$$

$$P_{1,(x_0,y_0)}(x,y) = \frac{x_0y_0}{x_0^2+y_0^2} + \frac{1}{(x_0^2+y_0^2)^2}[(y_0^3-x_0^2y)(x-x_0) + (x_0^3-x_0y_0^2)(y-y_0)].$$

Example 10. *Let* $f(x,y) = \arctan(\frac{x+y}{x-y})$ *then*

$$\nabla f(x,y) = \frac{1}{(x-y)^2+(x+y)^2}\begin{bmatrix} 2x \\ 2(x-y) \end{bmatrix}$$

$$P_{1,(x_0,y_0)}(x,y) = \arctan(\frac{x_0+y_0}{x_0-y_0}) + \frac{1}{(x_0-y_0)^2+(x_0+y_0)^2}[2x_0(x-x_0) + 2(x_0-y_0)(y-y_0)].$$

Lecture 2: Linear algebra

Gaussian elimination

Exercise 11. *Solve the lynear system*

$$\begin{cases} x + y - 3z = 6 \\ 3x - y + 2z = 3 \\ -x + 2y - z = 1 \end{cases}$$

with the Gaussian elimination method.

$$\left[\begin{array}{ccc|c} 1 & 1 & -3 & 6 \\ 3 & -1 & 2 & 3 \\ -1 & 2 & -1 & 1 \end{array}\right] \to \left[\begin{array}{ccc|c} 1 & 1 & -3 & 6 \\ 0 & -4 & 11 & -15 \\ 0 & 3 & -4 & 7 \end{array}\right] \to \left[\begin{array}{ccc|c} 1 & 1 & -3 & 6 \\ 0 & -4 & 11 & -15 \\ 0 & 0 & \frac{17}{4} & -\frac{17}{4} \end{array}\right] \Rightarrow$$

$$\begin{cases} x + y - 3z = 6 \\ -4y + 11z = -15 \\ z = -1 \end{cases} \Rightarrow \begin{cases} x + 1 + 3 = 6 \\ y = 1 \\ z = -1 \end{cases} \Rightarrow \boxed{\begin{cases} x = 2 \\ y = 1 \\ z = -1 \end{cases}} \text{ unique solution.}$$

Exercise 12. *Solve the lynear system*

$$\begin{cases} -x - y + z = 0 \\ 2z = 1 \\ -x - y = 2 \end{cases}$$

with the Gaussian elimination method.

$$\left[\begin{array}{ccc|c} -1 & -1 & 1 & 0 \\ 0 & 0 & 2 & 1 \\ -1 & -1 & 0 & 2 \end{array}\right] \to \left[\begin{array}{ccc|c} -1 & -1 & 1 & 0 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & -1 & 2 \end{array}\right] \Rightarrow \begin{cases} -x - y + z = 0 \\ 2z = 1 \\ -z = 2 \end{cases} \Rightarrow \text{ no solutions.}$$

Exercise 13. *Solve the lynear system*

$$\begin{cases} x - y + 4z = 10 \\ 3x + y + 5z = 15 \\ x + 3y - 3z = 6 \end{cases}$$

with the Gaussian elimination method.

$$\left[\begin{array}{ccc|c} 1 & -1 & 4 & 10 \\ 3 & 1 & 5 & 15 \\ 1 & 3 & -3 & 6 \end{array}\right] \to \left[\begin{array}{ccc|c} 1 & -1 & 4 & 10 \\ 0 & 4 & -7 & -15 \\ 0 & 4 & -7 & -4 \end{array}\right] \to \left[\begin{array}{ccc|c} 1 & -1 & 4 & 10 \\ 0 & 4 & -7 & -15 \\ 0 & 0 & 0 & 11 \end{array}\right] \Rightarrow \text{ no solutions.}$$

Exercise 14. *Solve the lynear system*

$$\begin{cases} x - 2y - 2z = 0 \\ -2x - y + 4z = 3 \\ x - 2y - 2z = 0 \end{cases}$$

with the Gaussian elimination method.

$$\left[\begin{array}{ccc|c} 1 & -2 & -2 & 0 \\ -1 & -1 & 4 & 3 \\ 1 & -2 & -2 & 0 \end{array}\right] \to \left[\begin{array}{ccc|c} 1 & -2 & -2 & 0 \\ 0 & -5 & 0 & 3 \\ 0 & 0 & 0 & 0 \end{array}\right] \Rightarrow \begin{cases} x - 2y - 2z = 0 \\ -5y = 3 \end{cases} \Rightarrow \begin{cases} x = -\frac{6}{5} - 2z \\ y = -\frac{3}{5} \end{cases}$$

So we have infinite solutions of the form

$$\begin{bmatrix} -\frac{6}{5} \\ -\frac{3}{5} \\ 0 \end{bmatrix} + t \begin{bmatrix} -2 \\ 0 \\ 1 \end{bmatrix}, \qquad t \in \mathbb{R}.$$

Exercise 15. *Solve the lynear system*

$$\begin{cases} x + y + z + w = 1 \\ 2x + \lambda y + \lambda z + \lambda w = \lambda \\ \lambda x + 2(\lambda - 1)y + 2z + 2w = \lambda^2 - 2 \\ x + y + (\lambda - 1)z + w = 1 \end{cases}$$

for the different values of the parameter λ.

$$\left[\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 0 \\ 2 & \lambda & \lambda & \lambda & \lambda \\ \lambda & 2(\lambda-1) & 2 & 2 & \lambda^2 - 1 \\ 1 & 1 & \lambda - 1 & 1 & 1 \end{array}\right] \to \left[\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 0 \\ 0 & \lambda - 2 & \lambda - 2 & \lambda - 2 & \lambda - 2 \\ 0 & \lambda - 2 & 2 - \lambda & 2 - \lambda & (\lambda - 2)(\lambda + 1) \\ 0 & 0 & \lambda - 2 & 0 & 0 \end{array}\right]$$

The second pivot is $\lambda - 2$ so we have to consider the case in which this quantity vanish or not.

$\boxed{\lambda \neq 2}$

$$\left[\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & -1 & -1 & \lambda + 1 \\ 0 & 0 & 1 & 0 & 0 \end{array}\right] \to \left[\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & -2 & -2 & \lambda \\ 0 & 0 & 1 & 0 & 0 \end{array}\right] \to \left[\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & -2 & -2 & \lambda \\ 0 & 0 & 0 & -1 & \frac{\lambda}{2} \end{array}\right] \Rightarrow$$

$$\Rightarrow \boxed{\begin{cases} x = 0 \\ y = \frac{\lambda}{2} + 1 \\ z = 0 \\ w = -\frac{\lambda}{2} \end{cases}} \text{ unique solution.}$$

$\boxed{\lambda = 2}$

$$\left[\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right] \Rightarrow \quad x+y+z+w=1 \quad \Rightarrow \quad x = 1-y-z-w \Rightarrow$$

$$\Rightarrow \text{infinite solutions of the form: } \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + p \begin{bmatrix} -1 \\ 1 \\ 0 \\ 0 \end{bmatrix} s \begin{bmatrix} -1 \\ 0 \\ 1 \\ 0 \end{bmatrix} t \begin{bmatrix} -1 \\ 0 \\ 0 \\ 1 \end{bmatrix} \qquad p, s, t \in \mathbb{R}.$$

Reducible matrices

Definition 16. *Let $\Pi \in \mathbb{R}^{n\times n}$ be a square matrix. Then Π is said to be a permutation matrix if the sequence of its columns (or rows) is a permutation of those of the identity matrix.*

Remark 17. *An example of permutation matrix is*

$$\Pi := \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

To multiply a matrix A on the left by Π is equivalent to apply the inverse permutation on the rows of A. Analogously to multiply A on the right by Π is equivalent to apply the permutation on the columns of A. For example

$$A := \begin{bmatrix} 1 & 2 & 9 & 10 \\ 3 & 4 & 11 & 12 \\ 5 & 6 & 13 & 14 \\ 7 & 8 & 15 & 16 \end{bmatrix}, \quad \Pi A = \begin{bmatrix} 7 & 8 & 15 & 16 \\ 1 & 2 & 9 & 10 \\ 3 & 4 & 11 & 12 \\ 5 & 6 & 13 & 14 \end{bmatrix}, \quad A\Pi = \begin{bmatrix} 2 & 9 & 10 & 1 \\ 4 & 11 & 12 & 3 \\ 6 & 13 & 14 & 5 \\ 8 & 15 & 16 & 7 \end{bmatrix}.$$

Moreover the inverse of a permutation matrix Π is its transpose: $\Pi \cdot \Pi^t = \Pi^t \cdot \Pi = I$.

In particular if A is the coefficient matrix of a linear system the multiplication by a permutation matrix can be interpreted as reordering the equations or as relabelling the variables.

Definition 18. *A matrix $A \in \mathbb{R}^{n\times n}$ with $n \geq 2$ is said reducible if $\exists \Pi$ permutation matrix and an integer $0 < k < n$ such that*

$$\Pi A \Pi^t = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}$$

where $A_{11} \in \mathbb{R}^{k\times k}$ and $A_{22} \in \mathbb{R}^{(n-k)\times(n-k)}$ (are square matrices).

If the coefficient matrix of a linear system is reducible then there exists a way to improve the Gaussian elimination, performing a starting variables relabelling and equations reordering.

Suppose we want to solve $Ax = b$ with A reducible matrix. Then consider Π permutation matrix associated to A and observe that

$$Ax = b \quad \Leftrightarrow \quad \Pi A x = \Pi b \quad \Leftrightarrow \quad \Pi A \Pi^t \cdot \underbrace{\Pi x}_{y} = \underbrace{\Pi b}_{c},$$

where

$$y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \qquad c = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}$$

are partitioned in compatible way to the partitioning of $\Pi A \Pi^t$. In that way

$$\Pi A \Pi^t \cdot y = c \quad \Leftrightarrow \quad \begin{cases} A_{11}\ y_1 + A_{12}\ y_2 = c_1 \\ A_{22}\ y_2 = c_2 \end{cases} .$$

So instead of solving a linear system of dimension n one can solve first the linear system of dimension $n-k$ getting y_2 and then the linear system of dimension k getting y_1. Since the computational complexity of solving a linear system is not linear (is cubic in the general case) this is an efficiency gain.

Example 19. *Let*

$$A = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 2 & 3 & -2 & 1 \\ -1 & 0 & -2 & 0 \\ 1 & -1 & 1 & 4 \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ -2 \\ -1 \\ -2 \end{bmatrix}.$$

The matrix A is reducible infact

$$\Pi = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}, \quad \Pi A \Pi^t = \begin{bmatrix} 4 & -1 & 1 & 1 \\ 1 & 3 & -2 & 2 \\ 0 & 0 & -2 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix}, \quad y = \begin{bmatrix} x_4 1 \\ x_2 \\ x_3 \\ x_1 \end{bmatrix}, \quad c = \begin{bmatrix} -2 \\ -2 \\ -1 \\ 1 \end{bmatrix}.$$

So we can solve

$$\begin{bmatrix} -2 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} x_3 \\ x_1 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

getting $\begin{bmatrix} x_3 \\ x_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. *Then we can solve*

$$\begin{bmatrix} 4 & -1 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} x_4 \\ x_2 \end{bmatrix} = \begin{bmatrix} -2 \\ -2 \end{bmatrix} - \begin{bmatrix} 1 & 1 \\ -2 & 2 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

getting $\begin{bmatrix} x_4 \\ x_2 \end{bmatrix} = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$.

Eigenvalues and eigenvectors

Definition 20. *Let $A \in \mathbb{C}^{n \times n}$ be a square matrix, $\lambda \in \mathbb{C}$ is said to be an* eigenvalue *for A if there exists a non zero vector $x \in \mathbb{C}^n \setminus \{0\}$ such that*

$$Ax = \lambda x.$$

The vector x is said eigenvector *associated with λ and the pair (λ, x) is called* eigenpair.

Remark 21. *Observe that for every eigenvalue there is an infinite number of eigenvectors associated to it. Infact if x is an eiegenvector then θx, with $\theta \in \mathbb{C} \setminus \{0\}$, is an eigenvector associated to same eigenvalue.*

We can characterize the eigenvalues of a matrix A as the roots of particular polynomial associated to the matrix.

Observe that λ is an eigenvalue of $A \in \mathbb{C}^{n\times n}$ if and only if $\exists x \neq 0$ such that

$$Ax = \lambda x \quad \Leftrightarrow \quad (A-\lambda I)x = 0 \quad \Leftrightarrow \quad A - \lambda I \text{ is singular} \quad \Leftrightarrow \quad \det(A-\lambda I) = 0.$$

The object $p_A(\lambda) := \det(A - \lambda I)$ is a polynomial in λ of degree n and its roots correspond to the eigenvalues of the matrix A. $p_A(\lambda)$ is called the characteristic polynomial of the matrix A. Moreover the previous relation implies that the set of the eigenvectors associated to an eigenvalue λ correspond to the kernel of the matrix $A - \lambda I$ minus the zero element.

Exercise 22. *Compute the eigenvalues and eigenvectors of*

$$A = \begin{bmatrix} 1 & 3 \\ 3 & 1 \end{bmatrix}$$

We can proceed by computing the roots of $p_A(\lambda)$:

$$A - \lambda I = \begin{bmatrix} 1-\lambda & 3 \\ 3 & 1-\lambda \end{bmatrix} \quad \Rightarrow \quad p_A(\lambda) = (1-\lambda)^2 - 9 = (\lambda - 4)(\lambda + 2) \quad \Rightarrow \quad \lambda_1 = 4,\ \lambda_2 = -2.$$

In order to find the eigenvectors associated with λ_1 and λ_2 we look for the vectors in the kernel of $A - \lambda_1 I$ and $A - \lambda_2 I$. That is we solve

$$(A-4I)\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Leftrightarrow \quad \begin{cases} -3x + 3y = 0 \\ 3x - 3y = 0 \end{cases} \quad \Leftrightarrow \quad \begin{cases} x = y \\ 0 = 0 \end{cases} \quad \Rightarrow \quad \text{Span}\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}\right) \setminus \{0\},$$

$$(A+2I)\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Leftrightarrow \quad \begin{cases} 3x + 3y = 0 \\ 3x + 3y = 0 \end{cases} \quad \Leftrightarrow \quad \begin{cases} x = -y \\ 0 = 0 \end{cases} \quad \Rightarrow \quad \text{Span}\left(\begin{bmatrix} 1 \\ -1 \end{bmatrix}\right) \setminus \{0\}.$$

The eigenvalues of a certain matrix enjoy a lot of properties and relations that we are going to state without proof. In what follows we assume $A \in \mathbb{C}^{n\times n}$ with entries a_{ij}, $1 \leq i, j, \leq n$ and eigenpairs $(\lambda_1, v_1), \ldots, (\lambda_n, v_n)$.

Properties 23.

- $\det(A) = \prod_{i=1}^n \lambda_i$.
- $Tr(A) := \sum_{i=1}^n a_{ii} = \sum_{i=1}^n \lambda_i$.
- *A and A^t shares the same eigenvalues.*
- *A^h (conjugate transpose) has eigenvalues $\bar{\lambda_i}$.*
- *If A is invertible then (λ_i^{-1}, v_i) are eigenpairs for A^{-1}.*
- *Let $S \in \mathbb{C}^{n\times n}$ be an invertible matrix and call $B := S^{-1}AS$. Then $(\lambda_i, S^{-1}v_i)$ are eigenpairs for B.*
- *Let $q(x) = q_0 + q_1 x + \cdots + q_h x^h$ be a polynomial and define the matrix $q(A) = q_0 I + q_1 A + \cdots + q_h A^h$. Then $(q(\lambda_i), v_i)$ are eigenpairs for $q(A)$.*

Structured matrices

Definition 24. *A matrix $A \in \mathbb{C}^{n\times n}$ ($\mathbb{R}^{n\times n}$) is called* normal *if $A^hA = AA^h$ ($A^tA = AA^t$).*

This class of matrices is important because, as claimed in the *Spectral theorem*, if A is normal then there exists a matrix $V = (v_1|\dots|v_n)$ such that

$$Av_i = \lambda_i v_i, \qquad \underbrace{v_i^h \cdot v_j}_{\text{scalar product}} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}, \qquad V^{-1}AV = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}.$$

That is there exists an orthonormal basis composed of eigenvectors of A which diagonalize the matrix.

Example 25.

$$A = \begin{bmatrix} -i & -i & 0 \\ -i & i & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad A^h = \begin{bmatrix} -i & i & 0 \\ i & i & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad A^hA = AA^h = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Definition 26. *A matrix $A \in \mathbb{C}^{n\times n}$ ($\mathbb{R}^{n\times n}$) is called* hermitian *(*symmetric*) if $A = A^h$ ($A = A^t$).*

Observe that if a matrix is hermitian or symmetric then in particular is normal, so the Spectral theorem holds also in this case. We can actually say something more on the eigenvalues of A. Observe that given an eigenvector v of unitary norm, the corresponding eigenvalue λ is equal to the quantity $v^hAv = v^h\lambda v = \lambda$. In particular

$$\bar{\lambda} = \left(v^hAv\right)^h = v^hA^hv = v^hAv = \lambda,$$

therefore $\lambda \in \mathbb{R}$. So the eigenvalues of an hermitian or symmetric matrix are real. Therefore it makes sense, when we deal with an hermitian or a symmetric matrix, to talk about the sign of its eigenvalues. An hermitian or symmetric matrix with positive eigenvalues is called *positive definite*. If we have the weak condition of nonnegative eigenvalues we call it *positive semidefinite*. It holds that an hermitian matrix is positive definite if and only if $\forall x \in \mathbb{C}^n \setminus \{0\}$ $x^hAx > 0$.

Example 27.

$$A = \begin{bmatrix} 3 & 2+i \\ 2-i & 1 \end{bmatrix} \text{ is hermitian}, \qquad A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 0 & 5 \\ 3 & 5 & 6 \end{bmatrix} \text{ is symmetric.}$$

Definition 28. *A matrix $A \in \mathbb{C}^{n\times n}$ ($\mathbb{R}^{n\times n}$) is called* unitary *(*orthogonal*) if $A^hA = AA^h = I$ ($A^tA = AA^t = I$).*

Again the unitary and orthogonal matrices are subsets of the class of normal matrices, so the Spectral theorem still apply to them. Instead they are not contained and they do not contain the classes of hermitian and symmetric matrices. In particular the eigenvalues of these matrices are not necessarily real but they enjoy another property. Observe that

$$Av = \lambda v \quad \Rightarrow \quad (Av)^h = (\lambda v)^h \quad \Rightarrow \quad v^hA^h = \bar{\lambda}v^h \quad \Rightarrow \quad v^hA^hAv = \bar{\lambda}\lambda v^hv \quad \Rightarrow \quad 1 = |\lambda|^2,$$

therefore the eigenvalues of a unitary or orthogonal matrix have modulus 1.

Example 29. *The matrix*

$$A = \begin{bmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{bmatrix} \in \mathbb{R}^{2\times 2}$$

is orthogonal for any alpha $\in \mathbb{R}$ and its eigenvalues are $\cos(\alpha) \pm \sin(\alpha)$.

Gershgorin circle theorems

We have seen previously that the eigenvalues of a matrix correspond to the roots of its characteristic polynomial. This fact tells us that when n is bigger than 4 there are no explicit formulas to express the eigenvalues. Sometimes we are satisfied to localize them by means of trust regions in order to apply a numerical method which returns an approximation. The cornerstone of the eigenvalue localization is the following Gershgorin Theorem.

Theorem 30. *Let $A \in \mathbb{C}^{n\times n}$ with entries a_{ij} and call Gershgorin circles of A the sets:*

$$G_i := \left\{ z \in \mathbb{C} : \ |z - a_{ii}| \leq \sum_{i\neq j=1}^{n} |a_{ij}| \right\}, \quad i = 1, \dots, n.$$

Then

(i) The eigenvalues of A are all contained in $\bigcup_{i=1}^{n} G_i$.

(ii) Let M_1 be the union of k of Gershgorin circles and M_2 the union of the others $n-k$. If moreover $M_1 \cap M_2 = \emptyset$ then M_1 contains exactly k eigenvalues and M_2 contains exactly $n-k$ eigenvalues.

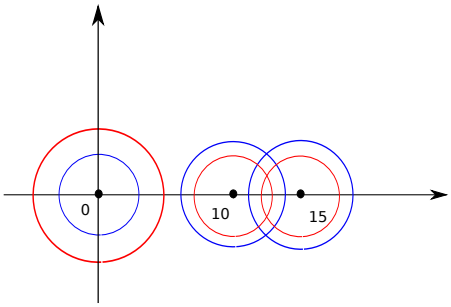
(iii) If the matrix A is irreducible (not reducible) then an eigenvalue belongs to the border of $\bigcup_{i=1}^{n} G_i$ if and only if it belongs to the border of each G_i.

The statements $(i), (ii)$ and (iii) are usually called first second and third theorem of Gershgorin respectively.

Remark 31. *Since the eigenvalues of A and A^t coincide one can define the Gershgorin circles using the columns in place of the rows. Calling them $G_i^{(r)}$ and $G_i^{(c)}$ respectively, as a consequence of the first Gershgorin theorem, the eigenvalues of A are contained in $\left(\bigcup_{i=1}^{n} G_i^{(r)}\right) \cap \left(\bigcup_{i=1}^{n} G_i^{(c)}\right)$.*

Example 32.

$$A = \begin{bmatrix} 15 & -2 & 2 \\ 1 & 10 & -3 \\ -2 & 1 & 0 \end{bmatrix}$$



Gershgorin circles of A: in blue the G_is are defined by rows, in red are defined by columns

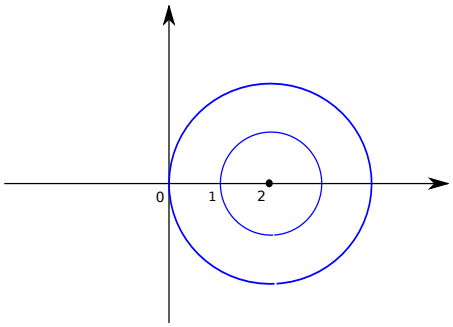
For what seen in the previous remark, the eigenvalues of A are contained in the three smallest circles around $0, 10$ and 15.

Reminder 33. *A matrix is irreducible if and only if its associated graph is strongly connected. The associaceted graph of a matrix is a directed graph having a node with label i for each $i = 1, \dots, n$ and an edge from node i to node j if $a_{ij} \neq 0$. That graph is strongly connected if $\forall i, j$ there is an oriented path from node i to node j.*

Example 34. *The matrix*

$$A = \begin{bmatrix} 2 & 1 & & \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & 2 \end{bmatrix} \in \mathbb{R}^{n \times n}$$

is irreducible, it is easy to see that looking at the graph associated to the matrix. Its Gershgorin circle are:

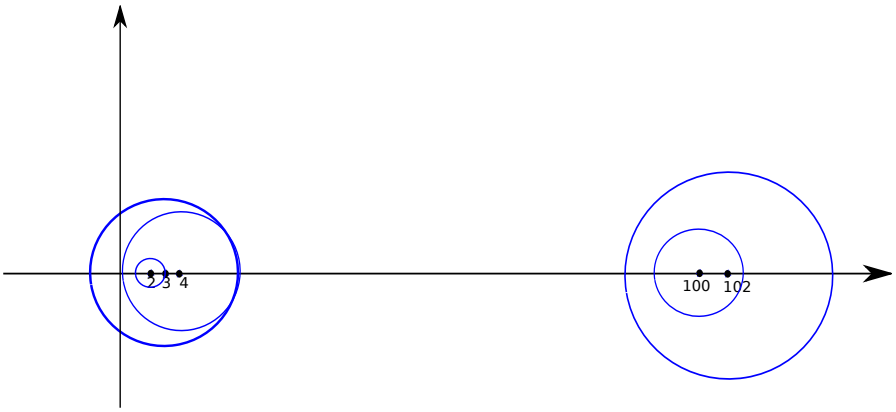


Therefore it is not singular because 0 can not verify the condition in the third Gershgorin theorem.

Example 35. *Given*

$$A = \begin{bmatrix} 2 & 1 & & & \\ -1 & 100 & 2 & & \\ & -2 & 3 & 3 & \\ & & -3 & 102 & 4 \\ & & & -4 & 4 \end{bmatrix}$$

it is possible to separate its Gerschgorin circles getting better estimates for the eigenvalues.



Consider the matrices

$$S = \begin{bmatrix} 1 & & & & \\ & \epsilon^{-1} & & & \\ & & 1 & & \\ & & & \epsilon^{-1} & \\ & & & & 1 \end{bmatrix}, \qquad S^{-1} = \begin{bmatrix} 1 & & & & \\ & \epsilon & & & \\ & & 1 & & \\ & & & \epsilon & \\ & & & & 1 \end{bmatrix},$$

we have that $B := S^{-1}AS$ *has the same eigenvalues of* A *(they are similar) and*

$$B = \begin{bmatrix} 2 & \epsilon^{-1} & & & \\ -\epsilon & 100 & 2\epsilon & & \\ & -2\epsilon^{-1} & 3 & 3\epsilon^{-1} & \\ & & -3\epsilon & 102 & 4\epsilon \\ & & & -4\epsilon^{-1} & 4 \end{bmatrix}.$$

Indicating with $C(x, r)$ *the circle of center* x *and radius* r *the Geshgorin circles defined by rows are*

$$C(2, \epsilon^{-1}), \quad C(3, 5\epsilon^{-1}), \quad C(4, 4\epsilon^{-1}), \quad C(100, 3\epsilon), \quad C(102, 7\epsilon).$$

Therefore we have that letting ϵ *goes to zero I can separate the circles around* 100 *and* 102 *but I enlarge the others. The other way round if I let* ϵ *goes to* $+\infty$*. Observe that if I fix* ϵ *such that one of the circles is separate form the others then I get that this circle contains exactly one eigenvalue (second Gershgorin theorem). Moreover the latter is real because otherwise its conjugate is an eigenvalue too (because the matrix has real entries) and it is contained in the same circle, which is absurd.*

Now note that fixing $\epsilon = \frac{1}{10}$ *we manage to separate the last two circles and with* $\epsilon = 10$ *we separate the first* 3*.*

In particular we get that the eigenvalues are all real and the following estimates hold:

$$\begin{aligned} 2 - \frac{1}{10} \leq \lambda_1 \leq 2 + \frac{1}{10}, \qquad & 3 - \frac{1}{2} \leq \lambda_2 \leq 3 + \frac{1}{2}, \\ 4 - \frac{2}{5} \leq \lambda_3 \leq 4 + \frac{2}{5}, \qquad & 100 - \frac{3}{10} \leq \lambda_4 \leq 100 + \frac{3}{10}, \\ 102 - \frac{7}{10} \leq \lambda_5 \leq 102 + \frac{7}{10}. & \end{aligned}$$

Lecture 3: Calculus

Convexity conditions

Definition 36. *The function $f : D \to \mathbb{R}$ defined on a convex set $D \subseteq \mathbb{R}^n$ is said to be* convex *on D if and only if*

$$\forall x, y \in D, \quad \forall \lambda \in (0,1): \qquad f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y).$$

It is strict convex *if the strictly inequality holds. It is β-*strongly convex *for a constant $\beta > 0$ if the function $f(x) - \frac{\beta}{2}\|x\|_2^2$ is convex.*

Proposition 37. *Let $f : D \to \mathbb{R}$ be a differentiable function, then*

(i) f is convex on D $\quad \Leftrightarrow \quad \forall x, y \in D \qquad f(x) \geq f(y) + \nabla f(y)^t (x - y)$.

(ii) f is strict convex on D $\quad \Leftrightarrow \quad \forall x, y \in D \qquad f(x) > f(y) + \nabla f(y)^t (x - y)$.

(iii) f is β-strongly convex on D $\quad \Leftrightarrow f(x) \geq f(y) + \nabla f(y)^t (x - y) + \frac{\beta}{2}\|x - y\|_2^2$.

Proposition 38. *Let $f : D \to \mathbb{R}$ be a differentiable function, then*

(i) f is convex on D $\quad \Leftrightarrow \quad \forall x \in D \qquad Hf(x)$ is positive semidefinite.

(ii) f is β-strongly convex on D $\quad \Leftrightarrow \forall x \in D \qquad Hf(x)$ is positive definite and $\lambda_{min} \geq \frac{\beta}{2}$.

Remark 39. *Strong convexity $\Longrightarrow$ strict convexity $\Longrightarrow$ convexity.*

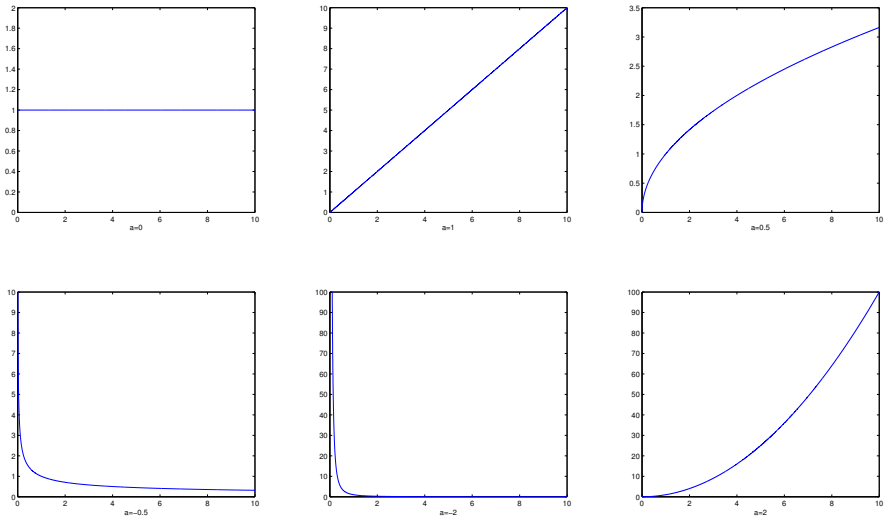
Convex analysis in one variable

In this subsection we consider functions $f : D \to \mathbb{R}$ where $D \subseteq \mathbb{R}$.

Exercise 40. *$D = \mathbb{R}^+$, $f(x) = x^a$. For which values of the parameter a the function f is convex, strictly convex or strongly convex?*

The function is differentiable on D for any value of the parameter a and $f''(x) = a(a-1)x^{a-2}$. Observe that the sign of $f''(x)$ on D is determined by the sign of $a(a-1)$. Therefore we have that the function is not convex for $\alpha \in (0,1)$ and strictly convex for $a \in (-\infty, 0) \cup (1, +\infty)$. If $a = 0$ we have the constant function which is convex but not strictly convex. If $a = 1$ we have a linear function which is convex but not strictly convex.

For what concerns strong convexity we need to study the convexity of $\tilde{f}(x) = f(x) - \frac{\beta}{2}x^2$ for $\beta > 0$. We have that $\tilde{f}''(x) = a(a-1)x^{a-2} - \beta$. Once we fixed the positive β parameter we get that this quantity becomes negative as x goes to 0. Therefore $\tilde{f}$ can not be convex on D for any positive β and consequently $f(x)$ is not strongly convex for any value of a.



Graph of $f(x)$ for different values of a. From left to right on the first row we have $a = 0, 1, \frac{1}{2}$ while on the second row we have $a = -\frac{1}{2}, -2, 2$

Exercise 41. *$D = \mathbb{R}$, $f(x) = x^4$. Is $f(x)$ convex? Is it strictly convex?*

Looking at $f''(x) = 12x^2 \geq 0$ we can claim that f is convex but since $f''(0) = 0$ we can not say anything about the strict convexity. So we use another condition:

$$f \text{ strict convex on D} \quad \Leftrightarrow \quad \forall x, y \in D,\ x \neq y, \quad f(x) > f(y) + f'(y)(x-y).$$

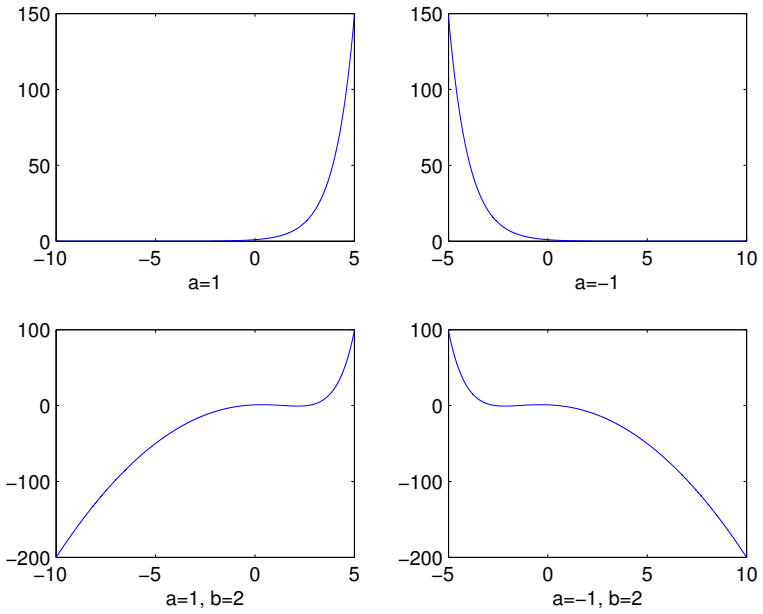
In our case we need to verify that

$$x^4 > y^4 + 4y^3(x-y) = -3y^4 + 4y^3 x \qquad \forall x, y \in \mathbb{R}, \quad x \neq y.$$

We do not need to verify the inequality when $x, y \in \mathbb{R}^+$ because we have just verified the strict convexity of f in that domain in the previous exercise. By symmetry we do not need to verify the case $x, y \in \mathbb{R}^-$ neither. If $x \cdot y = 0$ (one of the two is zero) then by direct verification the inequality holds. The only case left is $x \cdot y < 0$ (opposite signs). In this scenario the right-hand side $-3y^4 + 4y^3x$ is negative while x^4 is positive so the inequality holds. Therefore f is strict convex on the real line.

Exercise 42. *$D = \mathbb{R}$, $f(x) = e^{\alpha x}$ with $\alpha \neq 0$. For which values of the parameter α the function f is convex, strictly convex or strongly convex?*

We compute the second derivative getting $f''(x) = \alpha^2 e^{\alpha x} > 0 \; \forall x \in \mathbb{R}$ so the function is strict convex for any $\alpha \neq 0$. For the strong convexity we consider $\tilde{f}(x) = e^{\alpha x} - \frac{\beta}{2}$ with $\beta > 0$. The second derivative of $\tilde{f}$ is $\tilde{f}''(x) = \alpha^2 e^{\alpha x} - \beta$. Once we fixed β, this quantity becomes negative when $x \to \pm\infty$ (depending on the sign of α) and so $\tilde{f}$ is not convex on D. The latter means that f is not strongly convex on D for all choice of α.



Graphs of $f(x)$ and $\tilde{f}(x)$ for $\alpha = \pm 1$ and $\beta = 2$

Convex analysis in several variables

Exercise 43. *$D = \mathbb{R}^n$, $f(x) = \max_{i=1,\dots,n} x_i$ with $x = (x_1, \dots, x_n) \in \mathbb{R}^n$. Is f convex? Is it strictly convex? Is it differentiable?*

In order to prove the convexity we consider $x, y \in \mathbb{R}^n$, $\lambda \in (0, 1)$ and we observe that

$$\begin{aligned} f(\lambda x + (1-\lambda)y) &= \max_{i=1,\dots n} \lambda x_i + (1-\lambda) y_i \leq \max_{i=1,\dots n} \lambda x_i + \max_{i=1,\dots n} (1-\lambda) y_i \\ &= \lambda \max_{i=1,\dots n} x_i + (1-\lambda) \max_{i=1,\dots n} y_i = \lambda f(x) + (1-\lambda) f(y). \end{aligned}$$

Therefore f is convex.

The function is not strictly convex: consider as a counterexample the choice $x = (x_0, \dots, x_0)$ and $y = (y_0, \dots, y_0)$ with $x_0, y_0 \in \mathbb{R}$. Then we have

$$f(\lambda x + (1-\lambda)y) = \max_{i=1,\dots n} \lambda x_0 + (1-\lambda)y_0 = x_0 + (1-\lambda)y_0 = \lambda f(x) + (1-\lambda)f(y).$$

The function is not differentiable: consider the case $n = 2$, $x = (1, 1)$. We have that

$$\lim_{t\to 0^+} \frac{f(1+t,1) - f(1,1)}{t} = \lim_{t\to 0^+} \frac{1+t-1}{t} = 1, \quad \lim_{t\to 0^-} \frac{f(1+t,1) - f(1,1)}{t} = \lim_{t\to 0^-} \frac{1-1}{t} = 0,$$

therefore the function does not admit partial derivative in the first variable in the point $(1, 1)$. Note that the same argument can be applied for every point of the form (x_0, x_0).

Exercise 44. *$D = \{(x, y) : \ y > 0\}$, $f(x, y) = \frac{x^2}{y}$. Is f convex on D?*

Since f is twice differentiable it is enough to check if the Hessian of f is positive semidefinite on D. Performing some computations we get

$$\nabla f(x,y) = \begin{bmatrix} \frac{2x}{y} \\ -\frac{x^2}{y^2} \end{bmatrix} \qquad Hf(x,y) = \frac{2}{y^3} \begin{bmatrix} y^2 & -xy \\ -xy & x^2 \end{bmatrix}.$$

In particular $Tr(Hf) > 0$ and $\det(Hf) = 0$ on D. This means that one eigenvalue is zero and the other is positive (remember that trace and determinant correspond to sum and product of the eigenvalues), therefore Hf is positive semidefinite on D.

Exercise 45. *$D = \mathbb{R}^2$, $f(x) = x^2 + y^2 + 2xy + 4x - 3y$. Is it convex? Is it strongly convex? Does f admit global extrema?*

Since f is twice differentiable we check the second order conditions:

$$\nabla f(x,y) = \begin{bmatrix} 2x + 2y + 4 \\ 2x + 2y - 3 \end{bmatrix}, \qquad Hf(x,y) = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}.$$

Again $Tr(Hf) > 0, \ \det(Hf) = 0 \ \forall x, y \in \mathbb{R}^2 \quad \Rightarrow \quad Hf$ positive semidefinite on D, therefore f is convex but not strongly convex.

For what concerns global extrema consider a sequence of point of the form $\{(t, -t)\}$, then we have that

$$f(t,-t)) = t^2 + t^2 - 2t^2 + 4t + 3t = 7t.$$

This means that when $t \to \pm\infty$ $f(t, -t))$ tends to $\pm\infty$, hence f have no global extrema.

Exercise 46. *$D = \mathbb{R}^2$, $f(x) = x^2 + y^2 + xy + 4x - 3y$. Is it convex? Is it strongly convex?*

Doing some computations we get

$$\nabla f(x,y) = \begin{bmatrix} 2x + y + 4 \\ x + 2y - 3 \end{bmatrix}, \qquad Hf(x,y) = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}.$$

This time $Tr(Hf) > 0, \ \det(Hf) > 0 \ \forall x, y \in \mathbb{R}^2 \quad \Rightarrow \quad Hf$ positive definite on D, therefore f is strongly convex. The constant of strong convexity coincides with the smallest eigenvalue. Since the product of the eigenvalue is 3 and their sum is 4 it is easy to see that $\lambda_{min} = 1$.

Remark 47. *The latter two exercises concern the special class of quadratic functions, i.e. those functions of the form*

$$f(x,y) = \frac{1}{2}[x,y]Q\begin{bmatrix}x\\y\end{bmatrix} + b^t\begin{bmatrix}x\\y\end{bmatrix} + c, \qquad Q \in \mathbb{R}^{2\times 2},\ b \in \mathbb{R}^2 \text{ and } c \in \mathbb{R}.$$

Computing the second derivative we get that $Hf(x,y) = Q$, so

$$\begin{aligned} f \text{ is convex} &\quad\Leftrightarrow\quad Q \text{ is positive semidefinite} \\ f \text{ is } m\text{-strongly convex} &\quad\Leftrightarrow\quad Q \text{ is positive definite and } \lambda_{min} = m \end{aligned}$$

If Q is positive semidefinite (but not positive definite) its kernel contains $(\bar{x},\bar{y}) \neq (0,0)$. If $b^t(\bar{x},\bar{y}) \neq 0$ then the function does not admit global minima because considering a sequence of point $(t\cdot\bar{x}, t\cdot\bar{y})$ we get

$$f(t\bar{x},t\bar{y}) = \frac{1}{2}t^2[\bar{x},\bar{y}]\underbrace{Q\begin{bmatrix}\bar{x}\\\bar{y}\end{bmatrix}}_{0} + t\cdot\underbrace{b^t\begin{bmatrix}\bar{x}\\\bar{y}\end{bmatrix}}_{\neq 0} + c.$$

So as $t \to \pm\infty$ we can get $f(t\bar{x},t\bar{y}) \to \pm\infty$. In particular holds that

$$f \text{ admits global minima} \quad\Leftrightarrow\quad b^t\begin{bmatrix}\bar{x}\\\bar{y}\end{bmatrix} = 0 \quad \forall(\bar{x},\bar{y}) \in Ker(Q).$$

Exercise 48 (Rosenbrock function)**.** *$D = \mathbb{R}^2$, $f(x) = (a-x)^2 + b(y-x^2)^2$ with $a,b > 0$. Is it convex? Does f admit global minima?*

$$\nabla f(x,y) = \begin{bmatrix}-2(a-x) - 4bx(y-x^2)\\ 2b(y-x^2)\end{bmatrix}, \qquad Hf(x,y) = \begin{bmatrix}2 - 4by + 8bx & -4bx\\ -4bx & 2b\end{bmatrix}$$

If for example we evaluate the hessian in the point $(0,\frac{1}{b})$ we get

$$Hf(0,\frac{1}{b}) = \begin{bmatrix}-2 & 0\\ 0 & 2b\end{bmatrix}$$

which is an indefinite matrix. Therefore f is not convex on D. We look for stationary points getting

$$\nabla f(x,y) = \begin{bmatrix}0\\0\end{bmatrix} \quad\Leftrightarrow\quad \begin{cases}-2(a-x) = 0\\ y = x^2\end{cases} \quad\Leftrightarrow\quad \begin{cases}x = a\\ y = a^2\end{cases}.$$

Since $f(a,a^2) = 0$ and $f(x,y) \geq 0$ on D, because it is a sum of squares, we can conclude that (a,a^2) is a global minima point.

Proposition 49. *Consider $\mathbb{R}^n$ equipped with a norm $\|\cdot\|$ and let $A \subseteq \mathbb{R}^n$ be a convex subset. Then distance function*

$$d_A : \mathbb{R}^n \to \mathbb{R}, \quad d_A = \inf_{y\in A}\|x-y\|,$$

is a convex function.

Proof. Let $x_1, x_2 \in \mathbb{R}^n$ and call $d_1 = d_A(x_1)$ and $d_2 = d_A(x_2)$. So $\forall \epsilon > 0$ there exists $y_1, y_2 \in A$ such that

$$\|x_1 - y_1\| \leq d_1 + \epsilon, \quad \|x_2 - y_2\| \leq d_2 + \epsilon.$$

Moreover, since A is convex, $\forall \lambda \in (0,1)$ $\lambda y_1 + (1-\lambda) y_2 \in A$. Therefore

$$\begin{aligned} d_a(\lambda x_1 + (1-\lambda)x_2) &= \inf_{y \in A} \|\lambda x_1 + (1-\lambda)x_2 - y\| \leq \|\lambda x_1 + (1-\lambda)x_2 - (y_1 + (1-\lambda)y_2)\| \\ &\leq \|\lambda(x_1 - y_1)\| + \|(1-\lambda)(x_2 - y_2)\| \leq \lambda(d_1 + \epsilon) + (1-\lambda)(d_2 + \epsilon) \\ &= \lambda d_1 + (1-\lambda)d_2 + \epsilon = \lambda d_A(x_1) + (1-\lambda) d_A(x_2) + \epsilon. \end{aligned}$$

Since it holds $\forall \epsilon > 0$ we get the thesis. □

Proposition 50 (Composition with an affine function)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function and $g : \mathbb{R}^m \to \mathbb{R}^n$ be an affine function, i.e. there exists $A \in \mathbb{R}^{n \times m}$ and $b \in \mathbb{R}^n$ such that $g(x) = Ax + b$. Then $f \circ g : \mathbb{R}^m \to \mathbb{R}$ is convex.*

Proof. Observe that $(f \circ g)(x) = f(Ax + b)$. Given $x, y \in \mathbb{R}^m$ and $\lambda \in (0,1)$ we have

$$\begin{aligned} c(f \circ g)(\lambda x + (1-\lambda)y) &= f(A(\lambda x + (1-\lambda)y) + b) = f(\lambda Ax + (1-\lambda)Ay + \lambda b + (1-\lambda)b) \\ &= f(\lambda(Ax + b) + (1-\lambda)(Ay + b)) \leq \lambda f(Ax + b) + (1-\lambda) f(Ay + b) \\ &= \lambda (f \circ g)(x) + (1-\lambda)(f \circ g)(y). \end{aligned}$$

□

It is usefull to stress that the convexity property depends only on the behaviour of the function along the straight lines. To be precise

$$f \text{ is convex on } D \quad \Leftrightarrow \quad \forall x \in D,\ \forall v \ \ g(t) = f(x + tv) \text{ is convex on } \tilde{D} = \{t : \ x + tv \in D\}.$$

That reformulation can be useful because moves the problem from a function with domain D (that can be complicated) to a function whose domain is a subset of $\mathbb{R}$.

Exercise 51. *Let $D = \{A \in \mathbb{R}^{n \times n} : \ A \text{ is positive definite}\} \subset \mathcal{S} = \{\text{symmetric } n \times n \text{ matrices}\}$ and $f(X) = -\log(\det(X))$. Is f convex on D?*

First observe that given $A_1, A_2 \in D$ and $\lambda \in (0,1)$

$$\lambda A_1 + (1-\lambda)A_2 \text{ is symmetric}, \qquad \forall x \in \mathbb{R}^n \backslash \{0\} \quad x^t(A_1 + (1-\lambda)A_2)x = \lambda x^t A_1 x + (1-\lambda) x^t A_2 x > 0,$$

therefore D is a convex set. Consider $A \in D$ and $B \in \mathcal{S}$ then $\tilde{D} = \{t \in \mathbb{R} : \ A + tB \in D\}$ is non empty because contains an interval around 0. Moreover we have

$$\begin{aligned} g(t) &= -\log(\det(A + tB)) = -\log(\det(A)\det(I + tA^{-1}B)) = -\log(\det(A)) - \log(\det(I + tA^{-1}B)) \\ &= -\log(\det(A)) - \log(\prod_{i=1}^{n} 1 + t\lambda_i) = -\log(\det(A)) - \sum_{i=1}^{n} \log(1 + t\lambda_i), \end{aligned}$$

where λ_i indicate the eigenvalues of the matrix $A^{-1}B$. Then we compute the higher order derivatives of $g(t)$ getting

$$\begin{aligned} g'(t) &= -\sum_{i=1}^{n} \frac{1 + \lambda_i}{i + t\lambda_i}, \\ g''(t) &= \sum_{i=1}^{n} \frac{(1 + \lambda_i)^2}{(1 + t\lambda_i)^2}. \end{aligned}$$

Since $g''(t) \geq 0$ $\forall t$ and forall $B \in \mathcal{S}$ we can conclude the the function is convex.

Lecture 4: Linear algebra

Perturbation theory and condition number

When we address the numerical solution of a problem we have to deal with the error in the representation of the data and the roundoff error of the computations.

For example suppose that we want to solve the linear system

$$Ax = b$$

with $A \in \mathbb{R}^{n\times n},\ \det(A) \neq 0,\ b \in \mathbb{R}^n$. If the matrix of the coefficients A and the right hand side b come from some measurements then we can have only an approximation of them. It means that we can actually solve

$$(A + \delta A)\tilde{x} = b + \delta b, \qquad \tilde{x} = x + \delta x,$$

where the perturbations δA and δb are due to the finite accuracy of the measurement tools. In such situation one can not hope to avoid these disturbances, what we can hope is that the solution of the perturbed system is not too far away from the real solution. In particular we would like the relative error

$$\frac{\|\delta x\|}{\|x\|}$$

to be small. A theorem that bounds this quantity is the following.

Theorem 52. *Let $A, \delta A \in \mathbb{R}^{n\times n},\ \det(A) \neq 0,\ 0 \neq b \in \mathbb{R}^n$ and let $\|\cdot\|$ an induced matrix norm. If $\|A\|\|\delta A\| < 1$ then*

$$\frac{\|\delta x\|}{\|x\|} \leq \mu(A)\frac{\epsilon_A + \epsilon_b}{1 - \mu(A)\epsilon_A},$$

where $\epsilon_A = \frac{\|\delta A\|}{\|A\|}$, $\epsilon_b = \frac{\|\delta b\|}{\|b\|}$ and $\mu(A) = \|A\|\|A^{-1}\|$.

Remark 53. *The quantity $\mu(A)$ is said* condition number *of the matrix A. Observe that $\mu(A) = \|A\|\|A^{-1}\| \geq \|AA^{-1}\| = \|I\| = 1$.*

If $\mu(A)$ is not big then small perturbations in the intial data cause small relative error in the computed solution.

This notion strictly depends on the norm used so we will write $\mu(A, 1), \mu(A, 2), \mu(A, \infty)$ if we compute the condition number with respect to the 1,2 *of infinity norm.*

We now restrict ourself to the case in which the perturbations affect only the right hand side ($\delta A = 0$):

$$A(x + \delta x) = b + \delta b \qquad \Rightarrow \qquad \frac{\|\delta x\|}{\|x\|} \leq \mu(A) \cdot \epsilon_b = \mu(A)\frac{\|\delta b\|}{\|b\|}.$$

Exercise 54. *Is this bound sharp? If yes, how can I find b of unit norm and δb of norm ϵ such that the equality holds?*

Observe that replacing $Ax = b$ in the perturbed linear system we get

$$\begin{cases} Ax = b \\ A\delta x = \delta b \end{cases} \Rightarrow \begin{cases} x = A^{-1}b \\ \delta x = A^{-1}\delta b \end{cases}$$

The idea is to maximize the relative error and see if it reaches the bound of the theorem. Hence we look for

$$\max_{\|b\|=1, \|\delta b\|=\epsilon} \frac{\|\delta x\|}{\|x\|} = \max_{\|b\|=1, \|\delta b\|=\epsilon} \frac{\|A^{-1}\delta b\|}{\|A^{-1}b\|}.$$

Observe that in order to maximize this ratio we can indipendently look for

$$\max_{\|\delta b\|=\epsilon} \|A^{-1}\delta b\| \qquad \text{and} \qquad \min_{\|b\|=1} \|A^{-1}b\|.$$

For what concerns the first quantity, we have that

$$\max_{\|\delta b\|=\epsilon} \|A^{-1}\delta b\| = \max_{\|\delta b\|=\epsilon} \epsilon \|A^{-1}\frac{\delta b}{\epsilon}\| = \epsilon \max_{\|\delta b\|=1} \|A^{-1}\delta b\| = \epsilon \|A^{-1}\|.$$

where the last equality follows from the fact the the norm is induced.

For the second quantity, let us call

$$r := \min_{\|b\|=1} \|A^{-1}b\|$$

and consider a vector $\tilde{b}$ such that $\|\tilde{b}\| = 1$ and $\|A^{-1}\tilde{b}\| = r$. We also call $\tilde{y} = A^{-1}\tilde{b}$. Oviously $\|A\tilde{y}\| = \|\tilde{b}\| = 1$. Moreover the following property holds:

Proposition 55.

$$\max_{\|y\|=r} \|Ay\| = \|A\tilde{y}\| = 1.$$

Proof. Suppose by contradiction that $\exists y$: $\|y\| = r$ and $\|Ay\| = R > 1$. Then we should have that $\|A\frac{y}{R}\| = 1$ and

$$\|A^{-1}\left(A\frac{y}{R}\right)\| = \|\frac{y}{R}\| = \frac{r}{R} < r.$$

This is a contradiction because of the definition of r. □

So we get

$$1 = \max_{\|y\|=r} \|Ay\| = r \cdot \max_{\|y\|=1} \|Ay\| = r\|A\|,$$

which means

$$r = \min_{\|b\|=1} \|A^{-1}b\| = \|A\|^{-1}$$

Therefore choosing

$$\boxed{\begin{array}{c} \delta\tilde{b} = \epsilon \arg\max_{\|x\|=1} \|A^{-1}x\|, \\ \tilde{b} = \arg\min_{\|x\|=1} \|A^{-1}x\| = \|A\|^{-1} \cdot A \cdot \arg\max_{\|x\|=1} \|Ax\|, \end{array}}$$

the relative error is equal to

$$\frac{\|\delta x\|}{\|x\|} = \frac{\|A^{-1}\delta\tilde{b}\|}{\|A^{-1}\tilde{b}\|} = \frac{\|A^{-1}\|\epsilon}{\|A\|^{-1}} = \mu(A)\epsilon = \epsilon\frac{\|\delta\tilde{b}\|}{\|\tilde{b}\|}.$$

Example 56. *Consider*

$$A = \begin{bmatrix} 4 & 1 \\ 3 & 1 \end{bmatrix}, A^{-1} = \begin{bmatrix} 1 & -1 \\ -3 & 4 \end{bmatrix}.$$

I want to find the right hand side $\tilde{b}$ and the perturbation $\delta\tilde{b}$ of the previous exercise with respect to the infinity and euclidean norm.

$\boxed{\|\cdot\|_\infty}$ *We can easily compute*

$$\|A\|_\infty = 5, \quad \|A^{-1}\|_\infty = 7 \quad \Rightarrow \quad \mu(A,\infty) = 35.$$

Now observe that given a matrix A and indicating with a_{max} the row which determines $\|A\|_\infty$, we have that

$$\arg\max_{\|x\|_\infty=1} \|Ax\|_\infty = sign(a_{max})^t.$$

where the function sign is applied component-wise and

$$sign : \mathbb{R} \to \mathbb{R} \qquad sign(x) = \begin{cases} 1 & if\ x \geq 0 \\ -1 & if\ x < 0. \end{cases}$$

Therefore

$$\delta\tilde{b} = \epsilon \begin{bmatrix} -1 \\ 1 \end{bmatrix} \quad and \quad \tilde{b} = \frac{1}{5}\begin{bmatrix} 4 & 1 \\ 3 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ \frac{4}{5} \end{bmatrix}.$$

With this choice we get

$$\delta x = A^{-1}\delta\tilde{b} = \begin{bmatrix} 1 & -1 \\ -3 & 4 \end{bmatrix}\begin{bmatrix} -\epsilon \\ \epsilon \end{bmatrix} = \begin{bmatrix} -2\epsilon \\ 7\epsilon \end{bmatrix}, \qquad x = A^{-1}\tilde{b} = \begin{bmatrix} 1 & -1 \\ -3 & 4 \end{bmatrix}\begin{bmatrix} 1 \\ \frac{4}{5} \end{bmatrix} = \begin{bmatrix} \frac{1}{5} \\ \frac{1}{5} \end{bmatrix}$$

and the relative error verifies the equality

$$\frac{\|\delta x\|_\infty}{\|x\|_\infty} = 35\epsilon = \mu(A,\infty)\frac{\|\delta b\|_\infty}{\|b\|_\infty}.$$

$\boxed{\|\cdot\|_2}$ *For what concerns the 2-norm analysis observe that*

$$\arg\max_{\|x\|_2=1} \|A^{-1}x\|_2 = \arg\max_{\|x\|_2=1} \|A^{-1}x\|_2^2 = \arg\max_{\|x\|_2=1} (A^{-1}x)^t A^{-1}x = \arg\max_{\|x\|_2=1} x^t (A^{-1})^t A^{-1}x.$$

Since the matrix $(A^{-1})^t A^{-1}$ is symmetric (in particular positive definite) we know the solution of this optimization problem. That is the unitary eigenvector $\tilde{v}_{max}$ of $(A^{-1})^t A^{-1}$ associated to the the biggest eigenvalue λ_{max}.

Analogously

$$\arg\min_{\|x\|_2=1} \|A^{-1}x\|_2 = \arg\min_{\|x\|_2=1} x^t (A^{-1})^t A^{-1}x = \tilde{v}_{min}$$

where $\tilde{v}_{min}$ is the unitary eigenvector of $(A^{-1})^t A^{-1}$ associated to the the smallest eigenvalue λ_{min}. Observing that$(A^{-1})^t A^{-1}$ is the inverse of AA^t and using the relation between the eigenvectors and eigenvalues of a matrix and those of its inverse we get that

$$\tilde{b} = \epsilon \cdot v_{max}, \qquad \delta\tilde{b} = v_{min}$$

where v_{max} and v_{min} are the unitary eigevectors of AA^t associated to the biggest and smallest eigenvalues respectively.

Factorizations

Most of the methods for solving linear systems rely on the factorization of the coefficient matrix A into the product of two matrices

$$A = BC.$$

The solution of $Ax = b$ is then obtained by solving subsequently the two linear systems

$$\begin{cases} By = b \\ Cx = y \end{cases}.$$

This is done to gain in terms of the number of arithmetic operations required, so the matrices B, C must have some structure which enable us to solve the latter two linear systems efficiently.

The three classic factorizations, we consider are:

- **LU factorization**: L is lower triangular with ones on the main diagonal and U is upper triangular. This factorization is associated with the Gauss method.
- **QR factorization**: Q is unitary and R is upper triangular.
- **LL^h factorization:** L is lower triangular with non negative diagonal elements. This applies only to positive semidefinite matrices and it is usually called *Cholesky* factorization.

Exercise 57. *How do we compute the LU factorization?*

The idea is to exploit the closure of the lower (upper) triangular matrices with respect to the matrix multiplication and inversion. This means that if I multiply two lower triangular matrices or if I invert a lower triangular matrix I get again a lower triangular matrix.

So if I manage to transform the matrix A into an upper triangular matrix by multiplying it with lower triangular matrices I get the LU factorization. More explicitly

$$M_{n-1} \cdot \cdots \cdot M_2 \cdot M_1 \cdot A = \underbrace{U}_{\text{upper triangular}} \quad \Rightarrow \quad A = \underbrace{M_1^{-1} \cdot M_2^{-1} \cdot \ldots M_{n-1}^{-1}}_{L} U$$

where M_i is a lower triangular matrix for $i = 1, \ldots, n-1$.

In order to choose the matrices M_i we look back at the Gauss elimination method. At the first step of the algorithm we transform the coefficient matrix in that way

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix} \longrightarrow \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ 0 & a_{22}^{(2)} & \ldots & a_{2n}^{(2)} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(2)} & \ldots & a_{nn}^{(2)} \end{bmatrix}.$$

To obtain this operation is equivalent to multiply the matrix A on the left by the matrix

$$M_1 = \begin{bmatrix} 1 & 0 & \ldots & \ldots & \ldots & 0 \\ -m_{21} & 1 & 0 & \ldots & \ldots & 0 \\ \vdots & 0 & \ddots & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & & \ddots & \ddots & 0 \\ -m_{n1} & 0 & \ldots & \ldots & 0 & 1 \end{bmatrix}, \qquad m_{j1} = \frac{a_{j1}}{a_{11}}.$$

Similarly at the generic step k of the algorithm we perform the transformation

$$\begin{bmatrix} a_{11} & a_{12} & \dots & \dots & \dots & \dots & a_{1n} \\ 0 & a_{22}^{(2)} & \dots & \dots & \dots & \dots & a_{2n}^{(2)} \\ 0 & 0 & \ddots & & & & \vdots \\ \vdots & \vdots & \ddots & \ddots & & & \vdots \\ \vdots & \vdots & & 0 & a_{kk}^{(k)} & \dots & a_{kn}^{(k)} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 & a_{nk^{(k)}} & \dots & a_{nn}^{(k)} \end{bmatrix} \to \begin{bmatrix} a_{11} & a_{12} & \dots & \dots & \dots & \dots & a_{1n} \\ 0 & a_{22}^{(2)} & \dots & \dots & \dots & \dots & a_{2n}^{(2)} \\ 0 & 0 & \ddots & & & & \vdots \\ \vdots & \vdots & \ddots & a_{kk}^{(k)} & & & \vdots \\ \vdots & \vdots & & 0 & a_{k+1k}^{(k+1)} & \dots & a_{k+1n}^{(k+1)} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 & a_{nk+1}^{(k+1)} & \dots & a_{nn}^{(k+1)} \end{bmatrix}$$

which is equivalent to multiply on the left by the matrix

$$M_k = \begin{bmatrix} 1 & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & 1 & 0 & \dots & \dots & \dots & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & & & & \vdots \\ \vdots & & 0 & 1 & \ddots & & & \vdots \\ \vdots & & 0 & -m_{k+1k} & 1 & \ddots & & \vdots \\ \vdots & & \vdots & \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & & \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & -m_{nk} & 0 & \dots & 0 & 1 \end{bmatrix}, \qquad m_{jk} = \frac{a_{jk}^{(k)}}{a_{kk}^{(k)}}.$$

Observe that these kind of matrices can be seen as a rank 1 correction of the identity matrix:

$$M_k = I - u_k e_k^t$$

where e_k is the k-th vector of the canonical basis in $\mathbb{R}^n$ and

$$u_k = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ m_{k+1k} \\ \vdots \\ m_{nk} \end{bmatrix} \in \mathbb{R}^n.$$

The matrices with this structure are called elementary Gauss matrices and enjoy some very nice properties.

Proposition 58. *(i) Let $M = I - ue_k^t$ be an elementary Gauss matrix then*

$$M^{-1} = I + ue_k^t.$$

(i) Let $\tilde{M} = I - \tilde{u}e_j^t$ and $\hat{M} = I - \hat{u}e_i^t$ be two elementary Gauss matrices with $i \neq j$. Then

$$\tilde{M} \cdot \hat{M} = I - \tilde{u}e_j^t - \hat{u}e_i^t$$

Exploiting these properties we can write down the expression of L factor:

$$L = \begin{bmatrix} 1 & 0 & \dots & \dots & \dots & 0 \\ m_{21} & 1 & 0 & \dots & \dots & 0 \\ \vdots & m_{32} & \ddots & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & & \ddots & \ddots & 0 \\ m_{n1} & m_{n2} & \dots & \dots & m_{nn-1} & 1 \end{bmatrix}, \qquad m_{ij} = \frac{a_{ij}^{(j)}}{a_{jj}^{(j)}}.$$

Finally, for the U factor we take the triangular matrix we get at the end of the Gauss elimination method.

Example 59. *Compute the LU factorization of:*

$$A = \begin{bmatrix} 1 & 2 & -1 \\ -1 & -1 & 2 \\ 1 & 1 & 2 \end{bmatrix}.$$

We can already compute

$$m_{21} = -1, \quad m_{31} = 1.$$

Now we apply one step of the Gauss method

$$\begin{bmatrix} 1 & 2 & -1 \\ -1 & -1 & 2 \\ 1 & 1 & 2 \end{bmatrix} \to \begin{bmatrix} 1 & 2 & -1 \\ 0 & 1 & 1 \\ 0 & -1 & 3 \end{bmatrix}$$

getting $m_{31} = -1$. Finally we perform the last step of the Gauss method getting U:

$$\begin{bmatrix} 1 & 2 & -1 \\ 0 & 1 & 1 \\ 0 & -1 & 3 \end{bmatrix} \to \begin{bmatrix} 1 & 2 & -1 \\ 0 & 1 & 1 \\ 0 & 0 & 4 \end{bmatrix} = U.$$

$$A = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 1 & -1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 2 & -1 \\ 0 & 1 & 1 \\ 0 & 0 & 4 \end{bmatrix}.$$

Exercise 60. *What is the computational cost of the LU factorization?*

When we speak about the computational cost we mean the number of arithmetic operations as a function of the dimension n of the linear system. We will use the relation operator $\simeq$ which means "equal neglecting lower order terms". Some formulas which will be usefull are:

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2} \simeq \frac{n^2}{2},$$
$$\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6} \simeq \frac{n^3}{3}.$$

Regarding the algorithm we used for computing the LU factorization we note that at the generic step k the action which requires more arithmetic operations is updating the coefficient matrix

(compute the product $M_k \dot{A}^{(k)}$). It needs $2(n-k)^2$ among multiplications and additions. So the cost of the complete algorithm is

$$2\sum_{k=1}^{n-1}(n-k)^2 = 2\sum_{k=1}^{n-1}k^2 \simeq \frac{2}{3}n^3$$

Exercise 61. *How do we compute the QR factorization?*

The process is the same but we change the elementary matrices that we use. In particular we now want the elementary matrices to be unitary. This because again the class of unitary matrices is closed under inversion and matrix multiplication. In order to get this property we consider elementary matrices of that kind

$$M = I - \beta u u^t, \quad u \in \mathbb{R}^n, \quad \beta = \frac{2}{u^t u} \in \mathbb{R},$$

which are called *Householder transformation*. Since we are interested in getting a triangular matrix we state this result about the Houseldoer transformations.

Proposition 62. *Given $v = (v_1, \dots, v_n)^t \in \mathbb{R}^n$ the unitary matrices $M = I - \beta u u^t$ with*

$$u = v \pm \|v\|_2 e_1 = \begin{bmatrix} v_1 \pm \|v\|_2 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}, \qquad \beta = \frac{2}{u^t u},$$

are such that

$$Mv = \alpha \cdot e_1, \qquad |\alpha| = \|v\|_2.$$

Therefore we can exploit this property for choosing the Householder transformations. For example at the first step we can select the Houselder transformation M_1 which maps the first column of A into a multiple of e_1. So that

$$M_1 A = \left[\begin{array}{c|ccc} \alpha & * & \dots & * \\ \hline 0 & & & \\ \vdots & & A_{n-1} & \\ 0 & & & \end{array}\right].$$

At the second step we select an Householder transformation which maps the first column of A_{n-1} in a multiple of $e_1 \in \mathbb{R}^{n-1}$. Then we complete the matrix in order to get a transformation which leaves unchanged the first row and the first column (of zeros):

$$M_2 = \left[\begin{array}{c|ccc} 1 & 0 & \dots & 0 \\ \hline 0 & & & \\ \vdots & & \tilde{M}_2 & \\ 0 & & & \end{array}\right], \qquad M_2(M_1 A) = \left[\begin{array}{cc|ccc} \alpha & * & * & \dots & * \\ 0 & \gamma & * & \dots & * \\ \hline 0 & 0 & & & \\ \vdots & \vdots & & A_{n-2} & \\ 0 & 0 & & & \end{array}\right].$$

At step k we multiply on the left by

$$M_k = \begin{bmatrix} I_{k-1} & 0 \\ 0 & \tilde{M}_k \end{bmatrix}.$$

After $n-1$ steps the matrix become upper triangular and we have the factor R. For getting the Q factor we compute the product

$$M_1^t \cdot M_2^t \cdot \ldots M_{n-1}^t.$$

Example 63. *Compute the QR factorization of*

$$A = \begin{bmatrix} 12 & -51 & 4 \\ 6 & 167 & -68 \\ -4 & 24 & -41 \end{bmatrix}.$$

$$a_1 = \begin{bmatrix} 12 \\ 6 \\ -4 \end{bmatrix}, \quad \|a_1\|_2 = 14, \quad u_1 = a_1 - \|a_1\|_2 e_1 = \begin{bmatrix} -2 \\ 6 \\ -4 \end{bmatrix} \quad \Rightarrow \quad M_1 = I - \frac{2}{u_1^t u_1} u_1 u_1^t = \begin{bmatrix} \frac{6}{7} & \frac{3}{7} & -\frac{2}{7} \\ \frac{3}{7} & -\frac{2}{7} & \frac{6}{7} \\ -\frac{2}{7} & \frac{6}{7} & \frac{3}{7} \end{bmatrix},$$

$$M_1 A = \begin{bmatrix} 14 & 21 & -14 \\ 0 & -49 & -14 \\ 0 & 168 & -77 \end{bmatrix}.$$

$$a_2 = \begin{bmatrix} -49 \\ 168 \end{bmatrix}, \quad \|a_2\|_2 = 175, \quad u_2 = \begin{bmatrix} 124 \\ 168 \end{bmatrix} \quad \Rightarrow \quad M_2 = \begin{bmatrix} 1 & 0 \\ 0 & I - \frac{2}{u_2^t u_2} u_2 u_2^t \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -\frac{7}{25} & \frac{24}{25} \\ 0 & \frac{24}{25} & \frac{7}{25} \end{bmatrix},$$

$$R = M_2 M_1 A = \begin{bmatrix} 14 & 21 & 14 \\ 0 & 175 & -70 \\ 0 & 0 & -35 \end{bmatrix}, \qquad Q = M_1^t M_2^t = \begin{bmatrix} 0.8571 & -0.3943 & 0.3314 \\ 0.4286 & 0.9029 & -0.0343 \\ -0.2857 & 0.1714 & 0.9429 \end{bmatrix}.$$

Exercise 64. *What is the computational cost of the QR factorization?*

At the genrirc step k the most expensive action is again to compute the product $M_k \cdot A^{(k)}$. This time, since the matrix M_k is not given for free as for the LU decomposition, we need at most $4(n-k+1)^2$ among multiplications and additions. So the total cost is

$$4 \sum_{k=1}^{n-1} (n-k+1)^2 = 4 \sum_{k=2}^{n} k^2 \simeq \frac{4}{3} n^3.$$

Exercise 65. *How do we compute the Cholesky factorization?*

We somply impose the equality $A = LL^t$ component-wise (we are considering the real case so we use the transpose operator) obtaing the relations

$$\begin{cases} a_{jj} = \sum_{k=1}^{j} L_{jk}^2 & j = 1, \ldots, n \\ a_{ij} = \sum_{k=1}^{j} L_{ik} L_{jk} & i = j+1, \ldots, n \quad j < n \end{cases}.$$

Solving these equations for L_{ij} one get

$$\begin{cases} L_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} L_{jk}^2} & j = 1, \ldots, n \\ L_{ij} = \frac{1}{L_{jj}} \left(a_{ij} - \sum_{k=1}^{j-1} L_{ik} L_{jk} \right) & i = j+1, \ldots, n \quad j < n \end{cases}.$$

Observe that L_{ij} depends on the elements $L_{i'j'}$ with $i' \leq i$ and $j' \leq j$. Therefore we can compute all the matrix L following this order:

$$\left[\begin{array}{c|c|c} \boxed{1} & & \\ & \boxed{3} & \\ \boxed{2} & & \\ & \boxed{4} & \ddots \end{array}\right].$$

Example 66. *Compute the Cholesky factorization of*

$$A = \begin{bmatrix} 4 & 12 & -16 \\ 12 & 37 & -43 \\ -16 & -43 & 98 \end{bmatrix}.$$

$$L_{11} = \sqrt{a_{11}} = 2, \quad L_{21} = \frac{1}{L_{11}} a_{21} = 6, \quad L_{13} = \frac{1}{L_{11}} a_{31} = -8,$$

$$L_{22} = \sqrt{a_{22} - L_{21}^2} = 1, \quad L_{32} = \frac{a_{32} - L_{31} L_{21}}{L_{22}} = 5$$

$$L_{33} = \sqrt{a_{33} - L_{31}^2 - L_{32}^2} = 3,$$

$$A = \begin{bmatrix} 2 & 0 & 0 \\ 6 & 1 & 0 \\ 8 & 5 & 3 \end{bmatrix} \cdot \begin{bmatrix} 2 & 6 & 8 \\ 0 & 1 & 5 \\ 0 & 0 & 3 \end{bmatrix}.$$

Exercise 67. *What is the computationale cost of the Cholesky decomposition?*

For the computation of L_{ij} we need at most $2j$ among multiplicative and additive operations. So to compute a row of L we need

$$2 \sum_{j=1}^{i} j \simeq i^2.$$

Therefore the whole algorithm requires

$$\sum_{i=1}^{n} i^2 \simeq \frac{n^3}{3}$$

arithmetic operations. We did not count the square roots which are n.

Convergence of Jacobi and Gauss Seidel methods

Recall that given the linear system $Ax = b$ the methods of Jacobi and Gauss Seidel have iteration matrices defined in this way

$$J = D^{-1}(M + N), \qquad G = (D - M)^{-1} N$$

where

$$D = \begin{bmatrix} a_{11} & & & \\ & \ddots & & \\ & & \ddots & \\ & & & a_{nn} \end{bmatrix}, \quad M = \begin{bmatrix} 0 & & & \\ -a_{21} & \ddots & & \\ \vdots & \ddots & \ddots & \\ -a_{n1} & \dots & -a_{nn-1} & 0 \end{bmatrix}, \quad N = \begin{bmatrix} 0 & -a_{12} & \dots & -a_{1n} \\ & \ddots & \ddots & \vdots \\ & & \ddots & -a_{n-1n} \\ & & & 0 \end{bmatrix}.$$

Theorem 68. *Let $A \in \mathbb{C}^{n \times n}$ and suppose that one of the following conditions hold*

(i) A is strict diagonally dominant

(ii) A^t is strict diagonally dominant

(iii) A is irreducible and diagonally dominant

(iv) A^t *is irreducible and diagonally dominant.*

Then $\rho(J), \rho(G) < 1$ *(Jacobi and Gauss Seidel methods applied to* A *converge).*

Proof. Observe that the hypothesis, so one of the conditions $(i) - (iv)$ implies two important consequences:

- The matrix A is non singular, thanks to the Geshgorin theorems.
- The diagonal elements a_{ii} are non zeros.

$\boxed{\rho(J) < 1}$ Suppose by conttradiction that $\exists\lambda$ eigenvalue of J such that $|\lambda| > 1$. Then

$$0 = \det(\lambda I - J) = \det(\lambda I - D^{-1}(M+N)) = \det(D^{-1}(\lambda D - M - N)) = \det(D^{-1})\det(\lambda D - M - N).$$

Since $\det(D^{-1}) \neq 0$ this would imply that the matrix $\lambda D - M - N$ is singular. But we if we observe the latter

$$\lambda D - M - N = \begin{bmatrix} \lambda a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_{n-1n} \\ a_{n1} & \dots & a_{nn-1} & \lambda a_{11} \end{bmatrix}$$

we note that it coincides with A unless the main diagonal which is rescaled by λ which is of modulus greater than 1. So also this matrix verifies one of the hypothesis $(i) - (iv)$ and therefore must be non singular $\Rightarrow\Leftarrow$.

$\boxed{\rho(G) < 1}$ Analogously, suppose by conttradiction that $\exists\lambda$ eigenvalue of G such that $|\lambda| > 1$. Then

$$\begin{aligned} 0 &= \det(\lambda I - G) = \det(\lambda I - (D-M)^{-1}N) = \det((D-M)^{-1})\det(\lambda(D-M) - N)) \\ &= \det(D-M)^{-1}\lambda^n \det(D - M - \lambda^{-1}N). \end{aligned}$$

Since $D - M$ is non singular and $\lambda \neq 0$ this would imply that the matrix

$$D - M - \lambda^{-1}N = \begin{bmatrix} a_{11} & \lambda^{-1}a_{12} & \dots & \lambda^{-1}a_{1n} \\ a_{21} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \lambda^{-1}a_{n-1n} \\ a_{n1} & \dots & a_{nn-1} & a_{11} \end{bmatrix}$$

is singular. The latter coincides with A unless for the strictly upper triangular part which is divided by λ which is a number of modulus greater than 1. Therefore even this matrix enjoys one of the hypothesis $(i) - (iv)$ and therefore must be non singular $\Rightarrow\Leftarrow$. □

Lecture 5: Calculus

Local maxima and minima in several variables

Suppose $D \subseteq \mathbb{R}^n$ be an open set and let $f : D \to \mathbb{R}$ be a twice differentiable function on D. Then we know that

- $x_0 \in D$ is local minimum point $\Rightarrow \nabla f(x_0) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and $Hf(x_0)$ is positive semidefinite.
- $x_0 \in D$ is local maximum point $\Rightarrow \nabla f(x_0) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and $Hf(x_0)$ is negative semidefinite.
- $\nabla f(x_0) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and $Hf(x_0)$ is positive definite $\Rightarrow x_0 \in D$ is local minimum point.
- $\nabla f(x_0) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and $Hf(x_0)$ is negative definite $\Rightarrow x_0 \in D$ is local maximum point.

So in order to address the issue of finding local maxima and minima points f we follow these steps:

1. We look for the stationary points i.e., the solutions of the system $\nabla f = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.
2. For each point computed at step 1 we evaluate the Hessian of f. If the Hessian is positive or negative definite we can say that the point is a minimum or maximum point respectively. If it is undefined we conclude that it is a saddle point.
3. If the Hessian is semidefinite we need to study the situation locally for example using sequences, changing variables or other techniques (no standard strategy, it depends on the case).

Exercise 69. *Find the stationary points of $f(x, y) = 2x^3 + x^2 + y^2 - 2y^3$. Are there any local maxima or minima points? Are they also global maxima or minima points?*

We compute the gradient and the Hessian of f

$$\nabla f(x, y) = \begin{bmatrix} 6x^2 + 2x \\ -6y^2 + 2y \end{bmatrix}, \qquad Hf(x, y) = \begin{bmatrix} 12x + 2 & 0 \\ 0 & -12y + 2 \end{bmatrix}.$$

We look at the solutions of

$$\nabla f(x_0) = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Leftrightarrow \quad \begin{cases} 2x(3x + 1) = 0 \\ 2y(1 - 3y) = 0 \end{cases} \quad \Leftrightarrow \quad \begin{cases} x = 0, -\frac{1}{3} \\ y = 0, \frac{1}{3} \end{cases}.$$

So the stationary points are

$$P_1 = (0, 0), \qquad P_2 = (0, \frac{1}{3}), \qquad P_3 = (-\frac{1}{3}, 0), \qquad P_4 = (-\frac{1}{3}, \frac{1}{3}).$$

We can now evaluate the Hessian in these points getting

$$Hf(0, 0) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}, \quad Hf(-\frac{1}{3}, 0) = \begin{bmatrix} -2 & 0 \\ 0 & 2 \end{bmatrix}, \quad Hf(0, \frac{1}{3}) = \begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix}, \quad Hf(-\frac{1}{3}, \frac{1}{3}) = \begin{bmatrix} -2 & 0 \\ 0 & -2 \end{bmatrix}.$$

So we can conclude that P_1 is a local minimum point, P_2 and P_3 are saddle points and P_4 is a local maximum point.

P_1 and P_4 can not be also global extremal points because if for example I consider points of the kind $(0,t)$, I get that

$$f(0,t) = t^2 - 2t^3 \qquad \text{and} \qquad \lim_{t\to\pm\infty} f(0,t) = \mp\infty.$$

Exercise 70. *Find the stationary points of $f(x,y) = (x^2+y^2)e^{-(x^2+y^2)}$. Are there any local maxima or minima points? Are they also global maxima or minima points?*

We compute the gradient and the Hessian of f

$$\nabla f(x,y) = \begin{bmatrix} 2x(1-x^2-y^2)e^{-(x^2+y^2)} \\ 2y(1-x^2-y^2)e^{-(x^2+y^2)} \end{bmatrix},$$

$$Hf(x,y) = \begin{bmatrix} \left[(2-4x^2)(1-x^2-y^2)-4x^2\right]e^{-(x^2+y^2)} & -4xy(2-x^2-y^2)e^{-(x^2+y^2)} \\ -4xy(2-x^2-y^2)e^{-(x^2+y^2)} & \left[(2-4y^2)(1-x^2-y^2)-4y^2\right]e^{-(x^2+y^2)} \end{bmatrix}.$$

We look at the solutions of

$$\nabla f(x_0) = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Leftrightarrow \quad \begin{cases} 2x(1-x^2-y^2) = 0 \\ 2y(1-x^2-y^2) = 0 \end{cases} \quad \Leftrightarrow \quad (x,y) = (0,0) \quad \text{or} \quad \{x^2+y^2 = 1\}.$$

So the stationary points are the origin and the points on the unit circle. Evaluating the Hessian in these points we get

$$Hf(0,0) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}, \quad \begin{cases} Hf(x,y) = \begin{bmatrix} -4e^{-1}x^2 & -4e^{-1}xy \\ -4e^{-1}xy & -4e^{-1}y^2 \end{bmatrix} \\ x^2+y^2 = 1 \end{cases}$$

For what concerns the origin we can conclude that it is a local minimum point. Instead the Hessian on the unit circle is negative semidefinite because its trace is negative and its determinant is 0 for every (x,y) in this set. So we can not say anything for the moment. In order to shed some light on this issue we perform a change of variable defining $t := x^2+y^2$. We get

$$f(x,y) = g(t) = t \cdot e^{-t} \qquad \text{for} \qquad t \geq 0.$$

Now we can easily see that to study the function $f(x,y)$ near the unit circle is equivalent to study $g(t)$ near the point 1. Observing that

$$g'(t) = (1-t)e^{-t}, \quad g''(t) = -(2-t)e^{-t} \Rightarrow \begin{cases} g'(1) = 0 \\ g''(1) = -e^{-1} < 0, \end{cases}$$

we can conclude that 1 is a local maximum point for $g(t)$ and so all the points of the unit circle are local maximum points for $f(x,y)$.

We can again use the function $g(t)$ to claim that these points are actually global maximum and minimum points. Infact since $g'(t) > 0$ for $t \in (0,1)$ and $g'(t) < 0$ for $t \in (1,+\infty)$ we can say that 1 is global maximum point for $g(t)$ on $\mathbb{R}^+$ and so the unit circle is composed of global maximum points for $f(x,y)$. Finally observing that $f(x,y) \geq 0 \; \forall (x,y) \in \mathbb{R}^2$ and that $f(0,0) = 0$ we can claim that the origin is a global minimum point.

Contour lines

Suppose $D \subseteq \mathbb{R}^2$ and let $f : D \to \mathbb{R}$ be a differentiable function. We are interested in studying the sets where f assumes constant values. These sets are called contour lines and are formally defined as

$$c \in \mathbb{R}, \qquad f^{-1}(c) = \{(x, y) \in \mathbb{R}^2 : \quad f(x, y) = c\}.$$

The contour lines of a differentiable function enjoy some nice properties.

- $f^{-1}(c)$ can be locally parametrized as a curve i.e., $\forall (x_0, y_0) \in f^{-1}(c)$ there exists a continuos function $\gamma : [a, b] \to \mathbb{R}^2$ such that

$$\gamma(t_0) = (x_0, y_0) \quad \text{for some} \quad t_0 \in (a, b) \qquad \text{and} \quad f(\gamma(t)) = c \quad \forall t \in [a, b].$$

- $c_1, c_2 \in \mathbb{R}$, $c_1 \neq c_2 \Rightarrow f^{-1}(c_1) \cap f^{-1}(c_2) = \emptyset$.

- In every point $(x_0, y_0) \in f^{-1}(c)$ the vector $\nabla f(x_0, y_0)$ is orthogonal to the contour line. Infact if we consider a local parametrization $\gamma(t) = (x(t), y(t))$ we have that

$$f(x(t), y(t)) = c \quad \forall t \in [a, b] \quad \Rightarrow 0 = \frac{\partial f(x(t), y(t))}{\partial t} = \frac{\partial f}{\partial x}(x(t), y(t))x'(t) + \frac{\partial f}{\partial y}(x(t), y(t))y'(t).$$

  Evaluating this formula in t_0 we have

$$0 = \frac{\partial f}{\partial x}(x_0, y_0)x'(t_0) + \frac{\partial f}{\partial y}(x_0, y_0)y'(t_0) = \nabla f(x_0, y_0)^t \cdot \begin{bmatrix} x'(t_0) \\ y'(t_0) \end{bmatrix}.$$

  Observing that $\begin{bmatrix} x'(t_0) \\ y'(t_0) \end{bmatrix}$ is the tangent vector of γ in (x_0, y_0) we prove our claim.

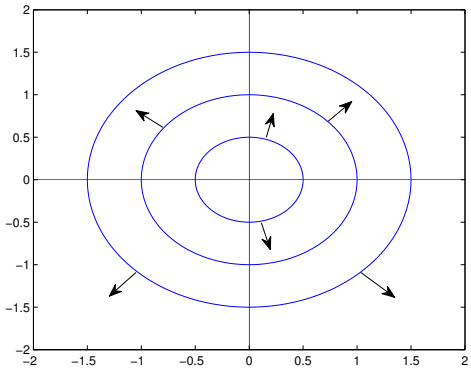
- As a consequence of the implicit function theorem, if the contour line $f^{-1}(c)$ self-intersecates in a point (x_0, y_0) then the latter must be a stationary point ($\nabla f(x_0, y_0) = (0, 0)^t$). If the contour line reaches the intersection with two linearly indipendent tangent vectors then the statement is implied by the previous property.

In the exercises and examples that we will see, it is possible to explicit the contour lines as union of graphics of functions of the form $y = h(x)$ or $x = g(y)$. Unfortunately this kind of analysis is not always applicable.

Example 71.

$$f(x, y) = x^2 + y^2, \qquad \nabla f(x, y) = \begin{bmatrix} 2x \\ 2y \end{bmatrix}, \qquad f^{-1}(c) = \begin{cases} \textit{circle of radius } \sqrt{c} & \textit{if } c \geq 0 \\ \emptyset & \textit{if } c < 0 \end{cases}.$$

I can see $f^{-1}(c)$ as the union of the graphics of $y = \pm\sqrt{c - x^2}$ or as the union of the graphics of $x = \pm\sqrt{c - y^2}$.



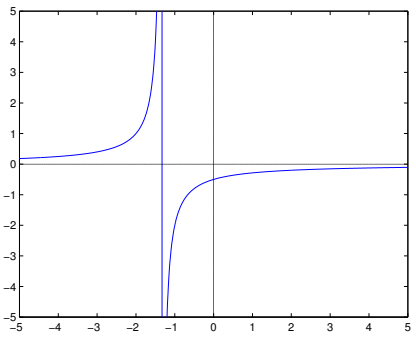
Gradient of $f(x,y) = x^2+y^2$ computed at certain points and contour lines passing through them.

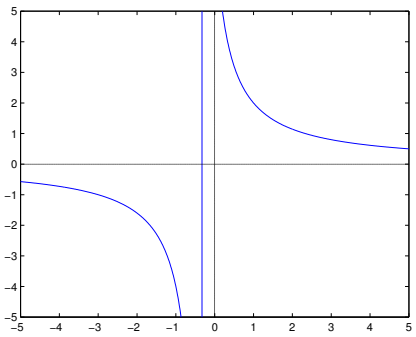
Exercise 72. *Study the contour lines of* $f(x,y) = \frac{3xy+y-4}{y+2}$.

Observe that the domain of f is $y \neq -2$. Since I am interested in $f^{-1}(c)$ I try to solve (in x or y) the equation $f(x,y) = c$, getting

$$\frac{3xy+y-4}{y+2} = c \qquad \Leftrightarrow \qquad \boxed{y = \frac{4+2c}{3x+1-c}}.$$

For every $c \neq -2$ this is an hyperbola with vertical asymptote given by the line $x = \frac{c-1}{3}$. Moreover wheter $c < -2$ or $c > 2$ the branches of the hyperbola are oriented in these ways.





Observe that the property of disjointness between different contour lines is not violated. Infact computing the generic intersection between two contour lines we get

$$c_1 \neq c_2, \qquad \frac{4+2c_1}{3x+1-c_1} = \frac{4+2c_2}{3x+1-c_2} \quad \Leftrightarrow c_1 - c_2 + \underbrace{(c_1-c_2)}_{\neq 0}x = 0 \quad \Leftrightarrow x = -1,$$

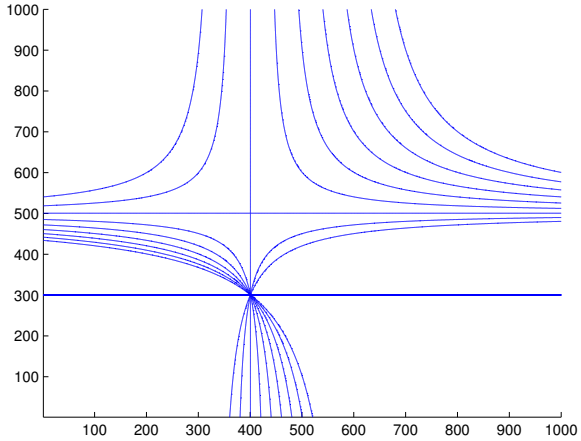
$$y = \frac{4+2c_1}{-3+1-c_1} = -2.$$

So all the contour line for $c \neq -2$ intersecate in the point $(-1,-2)$ which is not in the domain of f.

Instead when $c = -2$ we get

$$\frac{3xy + y - 4}{y + 2} = -2 \qquad \Leftrightarrow \qquad 3y(x + 1) = 0,$$

so the contour line in that case is equal to the union of the line $y = 0$ and $x = -1$.



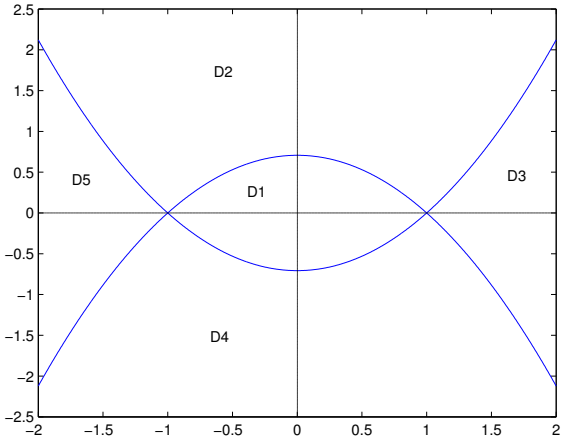
Contour lines for $c \in [-4, 4] \cap \mathbb{Z}$ and the line $y = -2$ where the function is not defined.

Exercise 73. *Given* $f(x, y) = x^2 - \frac{x^4}{4} + y^2$*, draw the contour line* $f^{-1}(\frac{1}{2})$*.*

$$x^2 - \frac{x^4}{4} + y^2 = \frac{1}{2} \qquad \Leftrightarrow \qquad \underbrace{x^4 - 2x^2 + 1}_{(x^2-1)^2} - 2y = 0 \qquad \Leftrightarrow \qquad (x^2 - 1 - \sqrt{2}y)(x^2 - 1 + \sqrt{2}y) = 0$$

This means that $f^{-1}(\frac{1}{2})$ is the union the two parabolas $y = \pm\frac{x^2-1}{\sqrt{2}}$.

Observe that the points $(\pm 1, 0)$ where the contour line self intersecates are stationary point of f. Moreover $f^{-1}(\frac{1}{2})$ separates the plane in the five regions D_i. In each D_i the function f remains under or above the value $\frac{1}{2}$. This is due to the continuity of f. For example observing that $f(0, 0) = 0 < \frac{1}{2}$ we can claim that $f_{|_{D_1}} < \frac{1}{2}$.



Exercise 74 (Bernoulli lemniscate)**.** *Given $f(x,y) = (x^2+y^2)^2 - 2(x^2-y^2)$, study the contour line $f^{-1}(0)$ and draw qualitatively the contour lines $f^{-1}(c)$ for a generic $c \in \mathbb{R}$.*

$$(x^2+y^2)^2 - 2(x^2-y^2) = 0 \qquad \Leftrightarrow \qquad y^4 + 2(x^2+1)y^2 + x^4 - 2x^2 = 0$$

replacing $t = y^2$ we get a quadratic equation in t

$$t^2 + 2(x^2+1)t + x^4 - 2x^2 = 0, \qquad \Delta = 4x^2+1 \qquad \Rightarrow \qquad t = y^2 = -(x^2+1) \pm \sqrt{4x^2+1},$$

therefore

$$y = \pm\sqrt{-(x^2+1) \pm \sqrt{4x^2+1}}.$$

Now observing that we choose the sign minus under the root sign we always get a negative quantity and so an empty set (remember we are in the real field!). So we can conclude that $f^{-1}(0)$ is given by the graphics of the two functions

$$y = g_{\pm}(x) = \pm\sqrt{-(x^2+1) + \sqrt{4x^2+1}}.$$

In order to draw this set we study the two functions g_+ and g_-. Observe that we can restrict ourself to study g_+ because $g_- = -g_+$. The domain of g_+ is determined by the x for which the quantity under the root sign is non negative, so

$$\sqrt{4x^2+1} \geq x^2+1 \qquad \Leftrightarrow \qquad x^4 - x^2 \leq 0 \qquad \Leftrightarrow \qquad -\sqrt{2} \leq x \leq \sqrt{2}.$$

Moreover $g_+ \geq 0$ in this domain and intersect the x-axis when the equality holds in the relation above, so when $x = 0, \pm\sqrt{2}$.

For what concerns the derivative of g_+ we have that

$$g'_+(x) = \frac{x}{\sqrt{-(x^2+1)+\sqrt{4x^2+1}}} \cdot \left(\frac{2}{\sqrt{4x^2+1}} - 1 \right),$$

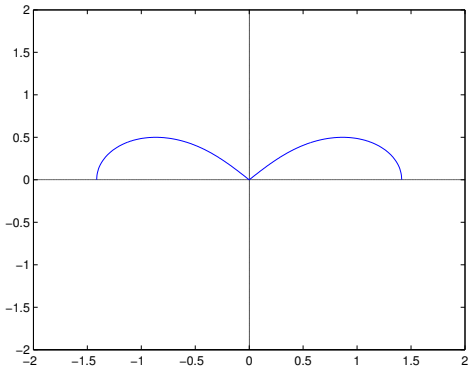
$$\lim_{x\to 0^-} g'_+(x) = -1 \qquad \text{and} \qquad \lim_{x\to 0^+} g'_+(x) = 1,$$

$$g'_+(x) = 0 \qquad \Leftrightarrow \qquad \frac{2}{\sqrt{4x^2+1}} - 1 = 0 \qquad \Leftrightarrow \qquad x = \pm\frac{\sqrt{3}}{2},$$

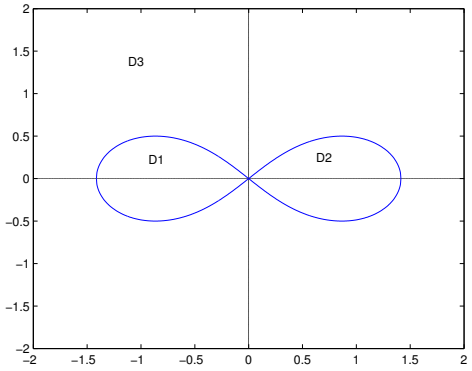
$$g'_+(x) \le 0 \quad \text{for} \quad -\frac{\sqrt{3}}{2} \le x \le 0 \quad \text{and} \quad \frac{\sqrt{3}}{2} \le x \le \sqrt{2},$$

$$g'_+(x) \ge 0 \quad \text{for} \quad -\sqrt{2} \le x \le -\frac{\sqrt{3}}{2} \quad \text{and} \quad 0 \le x \le \frac{\sqrt{3}}{2}.$$

In particular $\pm\frac{\sqrt{3}}{2}$ are maximum points. We can now draw g_+



and g_- by reflection.



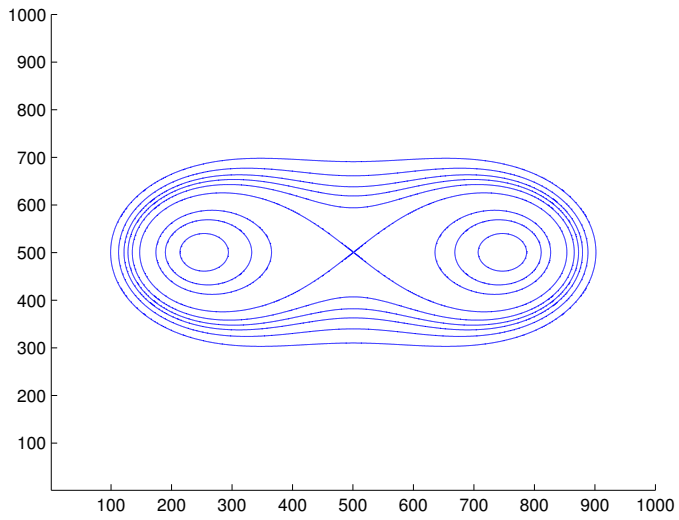
Observing that the plane is partitioned in 3 region D_1, D_2 and D_3 and that

$$f(\pm 1, 0) = -1, \qquad f(0, 2) = 12,$$
$$(1, 0) \in D_1, \qquad (-1, 0) \in D_2, \qquad (0, 2) \in D_3,$$

we can claim that the contour lines $f^{-1}(c)$ for c negative are contained in D_1 and D_2 while for c positive are contained in D_3. Moreover we can say that this curves must be symmetric to the y-axis because $f(x,y) = f(-x,y)$ and are bounded because the function is coercive i.e.,

$$\lim_{\|(x,y)\|_2 \to +\infty} f(x,y) = +\infty.$$

With these informations and keeping in mind that when $c \to 0$ then $f^{-1}(c)$ tends to the Bernoulli lemniscate, we can attempt to draw the generic contour lines qualitatively.



Examples on the Armijo condition

Consider an unconstrained optimization problem of the form $\min_{\mathbb{R}^n} f(x)$, with $f : \mathbb{R}^n \to \mathbb{R}$ differentiable function. A descent iterative method for solving such problem works in the following way.

1. Take the starting point $\tilde{x} \in \mathbb{R}^n$
2. Compute a descent direction i.e., a vector $d \in \mathbb{R}^n$ such that $\nabla f(\tilde{x})^t d \leq 0$.
3. Compute a step lenght $t \in \mathbb{R}$ such that $f(\tilde{x} + t \cdot d) \leq f(\tilde{x})$.
4. Assign $\tilde{x} \leftarrow \tilde{x} + t \cdot d$ and come back to step 2.

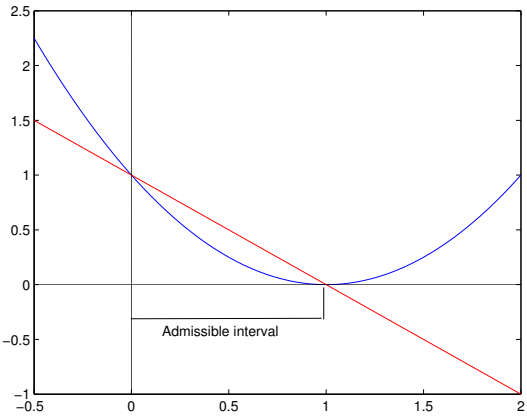
In order to choose the step lenght t we use as a criterion the inequality

$$f(\tilde{x} + t \cdot d) \leq c_1 \cdot t \cdot \nabla f(\tilde{x})^t d + f(\tilde{x}), \qquad c \in (0,1),$$

which is called *Armijo condition.*

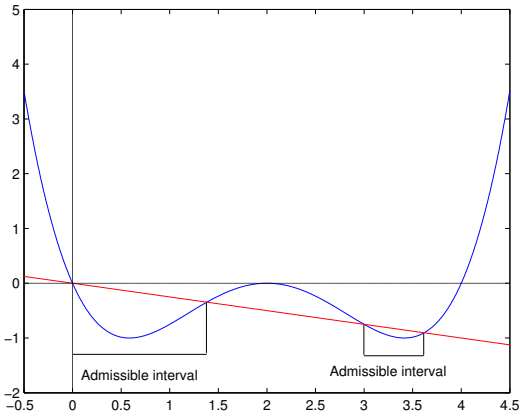
Example 75. *Let* $f(x,y) = \frac{x^2+y^2}{2}$, $\tilde{x} = (1,1)$, $d = -\nabla f(\tilde{x}) = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$ *and* $c_1 = \frac{1}{2}$. *We want to draw the Armijo condition.*

$$f(\tilde{x} + t \cdot d) = (1-t)^2 \qquad \Rightarrow \qquad (1-t)^2 \leq c_1 \cdot t \cdot \nabla f(\tilde{x})^t d + f(\tilde{x}) = -t + 1.$$



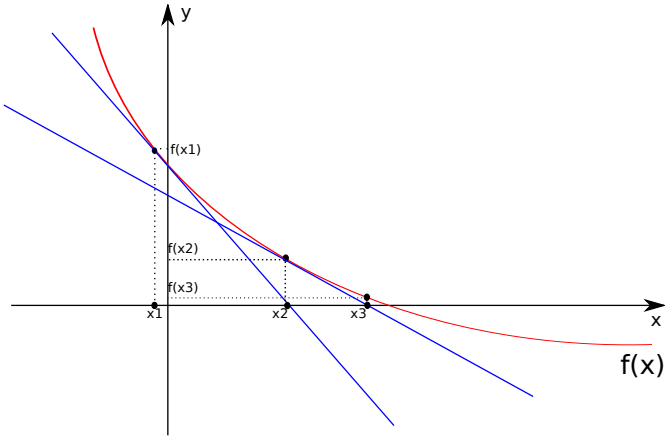
Example 76. *Let* $f(x,y) = y^2 + \frac{x^4}{4} - x^2$, $\tilde{x} = (2,0)$, $d = \begin{bmatrix} -1 \\ 0 \end{bmatrix}$ *and* $c_1 = \frac{1}{16}$. *We want to draw the Armijo condition.*

$$f(\tilde{x} + t \cdot d) = \frac{(2-t)^4}{4} - (2-t)^2 \qquad \Rightarrow \qquad \frac{(2-t)^4}{4} - (2-t)^2 \leq c_1 \cdot t \cdot \nabla f(\tilde{x})^t d + f(\tilde{x}) = -\frac{t}{4}.$$



1 Lecture 6: Linear algebra

Newton's method



For solving $f(x) = 0$ we look for a fixed point of $g(x) = x - \frac{f(x)}{f'(x)}$

$$\begin{cases} x_0 \quad \text{starting guess} \\ x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} \end{cases}$$

Definition 77. *A fixed-point method $\{x_k\}$ which converge to a certain limit α is said to converge*

- ***sublinearly** if*
$$\lim_{k\to+\infty} \frac{x_{k+1} - \alpha}{x_k - \alpha} = 1,$$
- ***linearly** if*
$$\lim_{k\to+\infty} \frac{x_{k+1} - \alpha}{x_k - \alpha} = \gamma \in (0, 1),$$
- ***superlinearly with order** p if*
$$\lim_{k\to+\infty} \frac{x_{k+1} - \alpha}{(x_k - \alpha)^p} = \gamma \in (0, +\infty).$$

Remark 78. *In particular when a fixed-point method $\{x_k\}$ converge superlinearly with order p to α we have that*

$$\lim_{k\to+\infty} \frac{x_{k+1} - \alpha}{(x_k - \alpha)^q} = 0 \quad \forall q < p, \qquad \lim_{k\to+\infty} \frac{x_{k+1} - \alpha}{(x_k - \alpha)^s} = +\infty \quad \forall s > p.$$

Remark 79. *If the iteration function $g(x)$ is differentiable than as a consequence of de l'Hôpital theorem*

$$\lim_{k\to+\infty} \frac{x_{k+1} - \alpha}{x_k - \alpha} = \lim_{k\to+\infty} \frac{g(x_k) - \alpha}{x_k - \alpha} = g'(\alpha).$$

Definition 80. *Let $f : \mathcal{C}^r[a, b]$ be a nonlinear function and $\alpha \in [a, b]$ be a point such that $f(\alpha) = 0$. α is said to be a solution of multiplicity $r \in \mathbb{N}$ if and only if the quantity*

$$\lim_{x\to\alpha} \frac{f(x)}{(x - \alpha)^r}$$

is finite and nonzero. This is equivalent (de l'Hôpital theorem) to ask that

$$f(\alpha) = f'(\alpha) = f''(\alpha) = \cdots = f^{(r-1)}(\alpha) = 0, \quad f^{(r)}(\alpha) \neq 0.$$

Theorem 81. *Let $\alpha \in [a,b]$ be a solution of $f(x)=0$ and suppose $f'(x)\neq 0$ on $[a,b]\setminus\{\alpha\}$, then*

(i) If α has multiplicity 1 and $f\in\mathcal{C}^2[a,b]$ then the Newton's method converge superlinearly with order at least 2. It is exactly 2 if $f'(\alpha)\neq 0$.

(ii) If α has multiplicity $2\leq r<+\infty$ then the Newton's method converge linearly.

Exercise 82. *Compute the inverse of a number $a\neq 0$ without using divisions.*

To compute the inverse of a is equivalent to look for the zeros of $f(x):=x^{-1}-a$. Since

$$x-\frac{f(x)}{f'(x)}=x-\frac{x^{-1}-a}{-x^{-2}}=x+x-a^2=2x-ax^2 \quad \Rightarrow \quad x_{k+1}=2x_k-ax_k^2,$$

we have that the newton method applied to this function does not involve any division. Moreover

$$x_{k+1}-a^{-1}=2x_k-ax_k^2-a^{-1}=-a(x_k-a^{-1})^2$$

therefore

$$\lim_{k\to+\infty}\frac{x_{k+1}-a^{-1}}{(x_k-a^{-1})^2}=-a \quad \Rightarrow \quad \text{Newton converge superlinearly with order 2.}$$

Exercise 83. *Compute the n-th root of $t\in\mathbb{R}^+$ using the Newton's method and analyze the speed of convergence*

$f(x)=x^n-t$, $f'(x)=nx^{n-1}$, $f''(x)=n(n-1)x^{n-2}$,

$$g(x)=x-\frac{=x^n-t}{nx^{n-1}}=\frac{1}{n}[(n-1)x+\frac{t}{x^{n-1}}] \quad \Rightarrow \quad x_{k+1}=\frac{1}{n}[(n-1)x_k+\frac{t}{x_k^{n-1}}].$$

Moreover

$$\lim_{x\to\sqrt[n]{t}}\frac{x^n-t}{x-\sqrt[n]{t}}\underbrace{=}_{Hopital}\lim_{x\to\sqrt[n]{t}}nx^{n-1}=nt^{\frac{n-1}{n}}\neq 0 \quad \Rightarrow \quad \text{superlinear convergence with order 2.}$$

Exercise 84. *Determine the iteration scheme of the Newton's method for $f(x)=x-\sin(x)$ and analyze its speed of convergence.*

$f'(x)=1-\cos(x)\Rightarrow x_{k+1}=x_k-\frac{x_k-\sin(x_k)}{1-\cos(x_k)}$. The method converge to $\alpha=0$ and since

$$\lim_{x\to 0}\frac{x-\sin(x)}{x^3}=\lim_{x\to 0}\frac{1-\cos(x)}{3x^2}=\lim_{x\to 0}\frac{\sin(x)}{6x}=\frac{1}{6}\neq 0$$

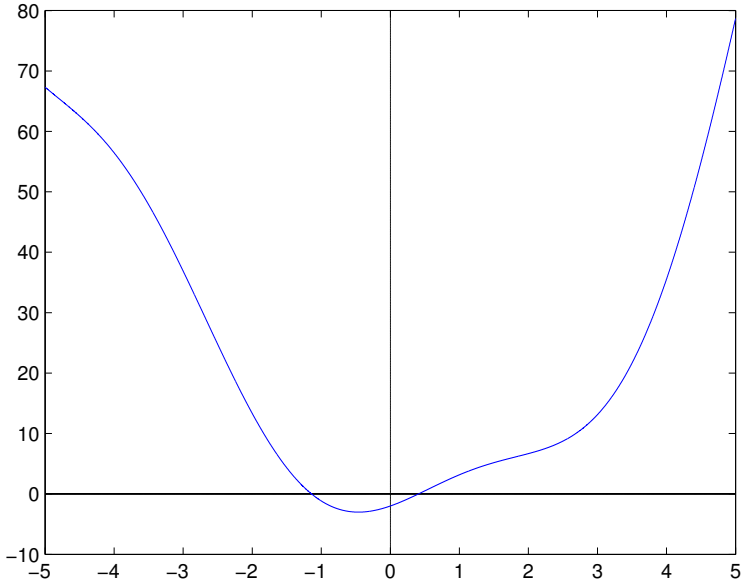
this is a solution of multiplicity 3. Therefore the Newton's method converge linearly.

Exercise 85. *Determine the iteration scheme of the Newton's method for $f(x)=x^3+4x\cos(x)-2$ and analyze its speed of convergence.*

$f'(x)=3x^2+4\cos(x)-4x\sin(x)\Rightarrow x_{k+1}=x_k-\frac{x_k^3+4x_k\cos(x_k)-2}{3x_k^2+4\cos(x_k)-4x_k\sin(x_k)}=\frac{2x_k^3+4x_k^2\sin(x_k)+2}{3x_k^2+4\cos(x_k)-4x_k\sin(x_k)}$.
Moreover observe that if $f(\alpha)=0$ then $\alpha=\frac{2-\alpha^3}{4\cos(\alpha)}$, so

$$\lim_{x\to\alpha}\frac{f(x)}{x-\alpha}=\lim_{x\to\alpha}\frac{x^3+4x\cos(x)-2}{x-\frac{2-\alpha^3}{4\cos(\alpha)}}=\lim_{x\to\alpha}\frac{x^3+4x\cos(x)-2}{4x\cos(\alpha)-2+\alpha^3}4\cos(\alpha)=4\cos(\alpha)\neq 0$$

which means that α is a solution of multiplicity 1. The Newton's method converge superlinearly in particular $f''(x)=6x-8\sin(x)-4x\cos(x)$ and you can prove numerically that $f''(\alpha)\neq 0$ obtaining that the order of superlinear convergence is 2.



$$f(x) = x^3 + 4x\cos(x) - 2$$

Exercise 86 (Brent function)**.** *Determine the iteration scheme of the Newton's method for*

$$f(x) = \begin{cases} 0 & x = 0 \\ xe^{-\frac{1}{x^2}} & x \neq 0 \end{cases}$$

and analyze its speed of convergence to the solution $\alpha = 0$.

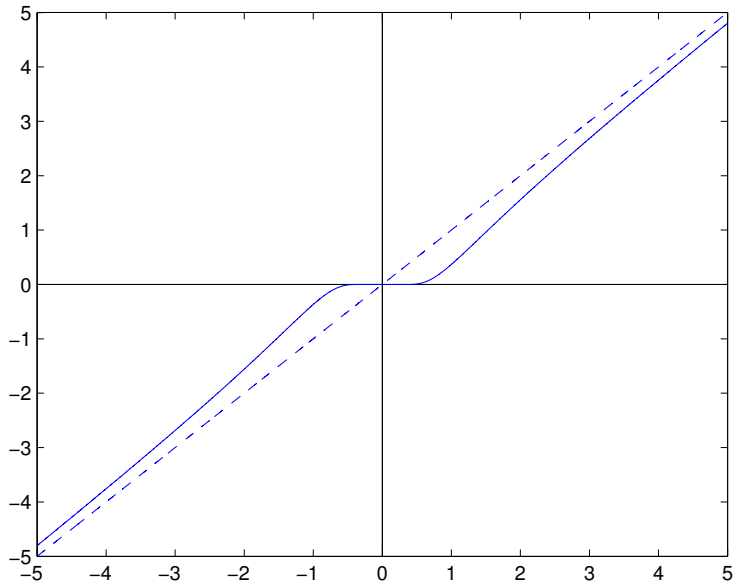
$f'(x) = (1 + \frac{2}{x^2})e^{-\frac{1}{x^2}} \Rightarrow x_{k+1} = x_k - \frac{x_k e^{-\frac{1}{x_k^2}}}{(1+\frac{2}{x_k^2})e^{-\frac{1}{x_k^2}}} = \frac{2x_k}{x_k^2+2}$. Observing the higher order derivatives of f we note that they are of the form $e^{-\frac{1}{x^2}}$ times a rational function, in particular this means that their limit as x goes to 0 is 0. So α is a solution of multiplicity ∞ and we can not apply the theorem for the speed of convergence. We need to study the derivative of $g(x) = \frac{2x}{x^2+2}$:

$$g'(x) = \frac{4 - 2x^2}{(x^2+2)^2} \quad \Rightarrow \quad g'(0) = 1, \quad 0 < g'(x) < 1 \quad \forall 0 \neq x \in (-\sqrt{2}, \sqrt{2}).$$

This means that the method converge sublinearly for all starting point in $(-\sqrt{2}, \sqrt{2})$.

Secants method

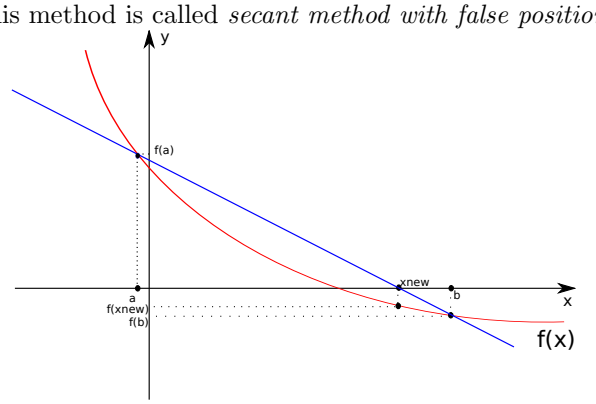
The Newton's method iteration scheme is based on the intersection between the tangent line to the graph of the function and the x-axis. It is possible to build a fixed-point method replacing the tangent lines with secants. This could be profitable for example if do not want to compute derivatives.



Brent function: ∞ multiplicity means a very flat behaviour in the neighbourhood of 0

- Start with two points x_0, x_1 such that $f(x_0)f(x_1) < 0$.
- Build the secant line s passing through $f(x_0)$ and $f(x_1)$
- Consider $x_2 = s \cap \{x = 0\}$
- If $f(x_0)f(x_2) < 0$ then $x_1 = x_0$.
- Restart with x_1 and x_2.

This method is called *secant method with false position*



$$\begin{cases} x_0, x_1 \quad \text{starting points } (f(x_0)f(x_1) < 0) \\ x_{k+1} = x_k - \frac{f(x_k)(x_{k-1} - x_k)}{f(x_k) - f(x_{k-1})} \\ if \quad f(x_{k+1})f(x_{k-1}) < 0 \quad then \quad x_k = x_{k-1} \end{cases}$$

Theorem 87. *Let f be continuos in $[a, b]$, $f(a)f(b) < 0$ and α be the unique solution of $f(x) = 0$ in $[a, b]$. Then the secant method with false position with starting points a and b converges to α.*

Newton-Raphson method

The Newton's method can be generalized in the several variables framework. Let $f : \mathbb{R}^n \to \mathbb{R}^n$ and suppose that I want to solve $f(x) = 0$. I can consider the iteration scheme: $x_{k+1} = g(x_k) = x_k - Jf(x_k)^{-1} f(x_k)$ where

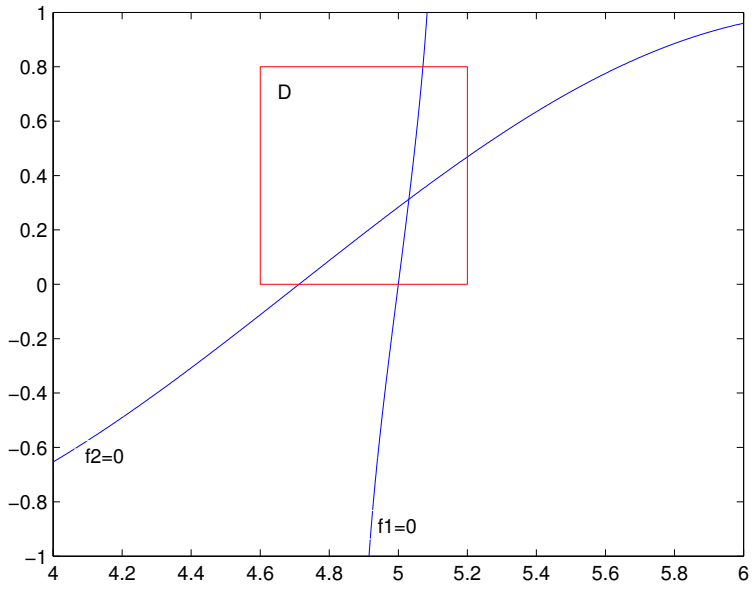
$$f(x) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{bmatrix} \quad f_i : \mathbb{R}^n \to \mathbb{R}, \quad Jf(x) = \begin{bmatrix} \nabla f_1(x)^t \\ \vdots \\ \nabla f_n(x)^t \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x_1} & \dots & \frac{\partial f_1(x)}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_n(x)}{\partial x_1} & \dots & \frac{\partial f_n(x)}{\partial x_n} \end{bmatrix}.$$

In this setting it is possible to state a result of monotone convergence. The following theorem generalize the relation between the sign of $g'(x)$ and the monotone convergence in the one variable case.

Theorem 88. *Suppose $D = [a_1, b_1] \times [a_2, b_2] \subset \mathbb{R}^2$, $f \in \mathcal{C}^1(D)$ convex (each component f_1 and f_2 is convex) and $\alpha \in D$ such that $f(\alpha) = 0$. If $Jf(x)$ is invertible and $Jf(x)^{-1} \geq 0$ (componentwise) $\forall x \in D$ then for every starting point in D Newton Raphson converge monotonically in each component.*

Example 89.

$$f(x_1, x_2) = \begin{cases} f_1(x_1, x_2) = x_1^2 - \sin(x_2) - 25 \\ f_2(x_1, x_2) = -\cos(x_1) + x_2 \end{cases}, \qquad D = [4.6, 5.2] \times [0, 0.8].$$



We want to check if the hypothesis of the previous theorem are satisfied, so we compute

$$Jf(x) = \begin{bmatrix} 2x_1 & -\cos(x_2) \\ \sin(x_1) & 1 \end{bmatrix} \quad \Rightarrow \quad \det(Jf(x)) = \underbrace{2x_1}_{>9 \text{ on } D} + \sin(x_1)\cos(x_2) > 0 \quad \forall x \in D,$$

$$Jf(x)^{-1} = \frac{1}{\det(Jf(x))}\begin{bmatrix} 1 & \cos(x_2) \\ -\sin(x_1) & 2x_1 \end{bmatrix} > 0 \quad \forall x \in D,$$

$$Hf_1(x) = \begin{bmatrix} 2 & 0 \\ 0 & \sin(x_2) \end{bmatrix}, \qquad Hf_2(x) = \begin{bmatrix} \cos(x_1) & 0 \\ 0 & 0 \end{bmatrix} \qquad \textit{are positive semidefinite.}$$

Preconditioned conjugate gradient

Suppose $A \in \mathbb{R}^{n\times n}$ positive definite. We want to solve the linear system $Ax = b$ by means of the conjugate gradient method. We obtain a sequence $\{x_k\}$ which converge to the solution x^*. An estimate for the error at step k $e_k := x_k - x^+$ is given in the following result.

Proposition 90. $\|e_k\|_A \leq \left(\frac{\sqrt{\mu(A)}-1}{\sqrt{\mu(A)}+1}\right)^{2k}$ *where* $\|e_k\|_A = \sqrt{x^t Ax}$ *and* $\mu(A) = \frac{\lambda_{max}}{\lambda_{min}}$ *is the condition number in the* 2*-norm.*

So when the matrix A is ill conditioned ($\mu(A)$ very large) the convergence can be very slow. This does not happen if the eigenvalues are clusterized close to 1. So in order to recover this pattern in the ill conditioned case one can observe that

$$Ax = b \quad \Leftrightarrow \quad B^{-1}Ax = B^{-1}b \quad \Leftrightarrow \quad B^{-1}AB^{-1}\underbrace{Bx}_{y} = \underbrace{B^{-1}b}_{c} \quad \Leftrightarrow \quad \left(B^{-1}AB^{-1}\right)y = c.$$

If B is invertible and symmetric then $B^{-1}AB^{-1}$ is still positive definite ($x^tB^{-1}AB^{-1}x = z^tAz > 0$) but its eigenvalues are different from the eigenvalues of A. So if there is a choice for the invertible symmetric matrix B such that $\mu(B^{-1}AB^{-1}) << \mu(A)$ I can think about applying the conjugate gradient method to the modified linear system getting a fast convergence. Once I get the solution y of the latter I can recover the solution of the original system using the relation $x = B^{-1}y$. I said "think" because in order to have a gain using this procedure we need the matrix B to have some additional structure that enable us to multiply by B and to compute B^{-1} efficiently (less than a cubic cost).

Example 91 (Strang preconditioner)**.** *If A is Töeplitz (constant along its diagonal)*

$$A = \begin{bmatrix} a_0 & a_{-1} & \dots & a_{-n+1} \\ a_1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_{-1} \\ a_{n-1} & \dots & a_1 & a_0 \end{bmatrix}$$

and positive definite (so in particular $a_k = a_{-k}$) then a possible choice for B is the Strang preconditioner *which is again a symmetric Töeplitz matrix defined by*

$$B = \begin{bmatrix} b_0 & b_{-1} & \dots & b_{-n+1} \\ b_1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & b_{-1} \\ b_{n-1} & \dots & b_1 & b_0 \end{bmatrix}, \quad b_k = \begin{cases} a_k & 0 \leq k \leq \lfloor \frac{n}{2} \rfloor \\ a_{k-n} & \lfloor \frac{n}{2} \rfloor < k < n \\ b_{n+k} & 0 < -k < n \quad \textit{symmetric completion} \end{cases}.$$

The Strang preconditioner turns out to be a Circulant matrix and so the arithmetic operations on it can be computed efficiently.

In the case

$$A = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & 2 & -1 & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix}$$

we have

$$B = \begin{bmatrix} 2 & -1 & & & -1 \\ -1 & 2 & -1 & & \\ & -1 & 2 & -1 & \\ & & -1 & 2 & -1 \\ -1 & & & -1 & 2 \end{bmatrix}.$$

Lecture 7: Calculus

Examples on the Wolfe condition

In this subsection we use the notation of "Examples on the Armijo conditions" (Lesson 5). In order to choose a step length t we usually need some criterion that ensures us a significant decrease in the evaluation of the objective function and which is not expensive to verify. We saw the Armijo condition

$$f(\tilde{x} + t \cdot d) \leq c_1 \cdot t \cdot \nabla f(\tilde{x})^t d + f(\tilde{x}), \qquad c \in (0, 1).$$

Observe that this inequality is always verified in an interval of the form $[0, \epsilon]$ because, by construction, the right-hand side decrease in a right neighbour of 0 with a derivative greater (less negative) than the objective function. This could be a problem because it is possible to choose a short step length which does not give us a reasonable progress. For that reason we introduce another requirement called Wolfe condition (or curvature condition):

$$\underbrace{\nabla f(\tilde{x} + td)}_{\varphi(t)} \geq \underbrace{c_2 \nabla f(\tilde{x})^t d}_{\varphi'(0)} \qquad c_1 < c_2 < 1,$$

with c_1 the constant in the Armijo condition.

In that way, requiring Armijo and Wolfe conditions the new point has this property:

the function φ, in a right neighbourhood of t does not decrease as much as in a rightneighbourhood of 0.

Observe that this does not mean that we are close to a minimizer, infact a case in which the Wolfe condition is verified is when $\varphi'(t) \geq 0$.

Example 92. $f(x, y) = x^2 + (x - y)^4$, $\tilde{x} = (1, 1)$

$$f(\tilde{x}) = 1, \quad \nabla f(x, y) = \begin{bmatrix} 2x + 4(x - y)^3 \\ -4(x - y)^3 \end{bmatrix}, \quad \nabla f(\tilde{x}) = \begin{bmatrix} 2 \\ 0 \end{bmatrix}, \quad d = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \quad \varphi'(0) = \nabla f(\tilde{x})^t d = -2.$$

Armijo condition:

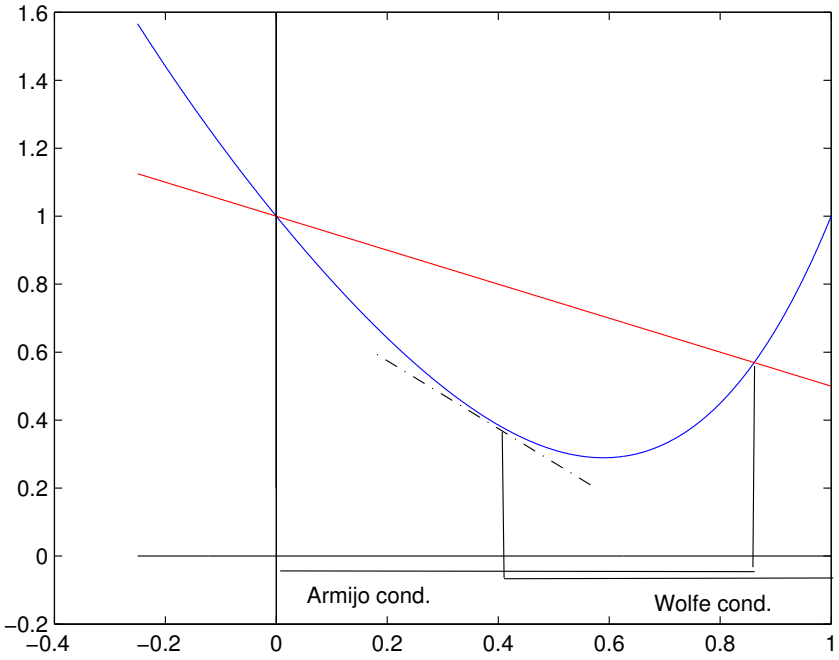
$$(1 - t)^2 + t^4 \leq -2c_1 t + 1.$$

Wolfe condition:

$$4t^3 + t - 2 \geq -2c_2.$$

In order to avoid too positive values for $\varphi'(t)$ one can consider the strong Wolfe condition:

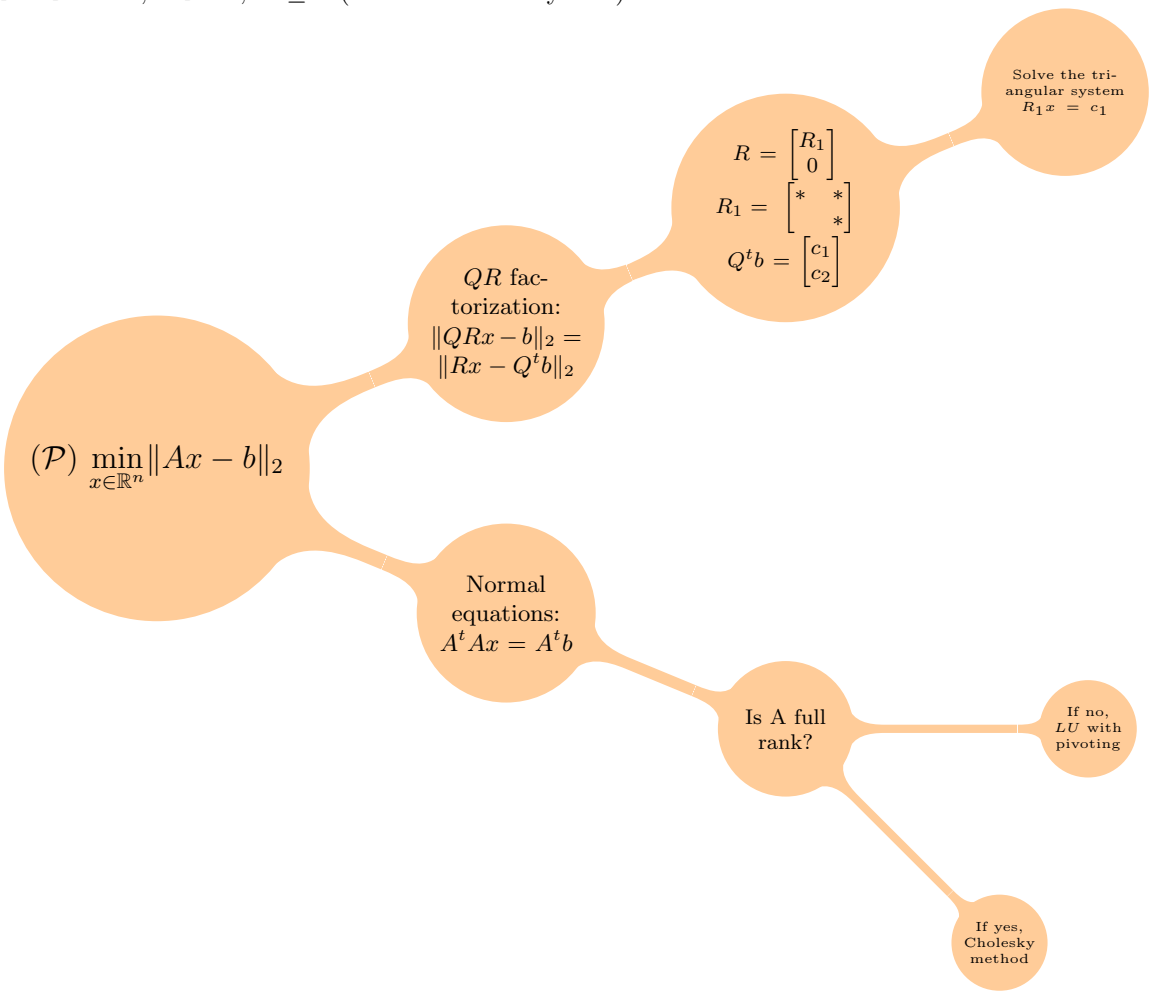
$$\underbrace{|\nabla f(\tilde{x} + td)^t d|}_{|\varphi'(t)|} \leq c_2 \underbrace{|\nabla f(\tilde{x})^t d|}_{|\varphi'(0)|}.$$



Intervals where Armijo and Wolfe conditions hold with $c_1 = \frac{1}{4}$ and $c_1 = \frac{1}{2}$

Linear least squares problem

$A \in \mathbb{R} \in m \times n$, $b \in \mathbb{R}^m$, $m \geq n$ (overdetermined system).



Exercise 93. *Suppose we have the following data coming from an experiment:*

$$(x_i, f(x_i)): \qquad (1, 1.1), \quad (2, 2.5), \quad (3, 2.8)$$

and suppose we want to fit them with a straight line (approximate f with a linear function). Find the best line which describes the dynamics of f in the 2-norm.

Suppose to have a proposoal $ax + b$ for approximating f. The residual in the 2-norm of the proposal is the sum of the squares of the residuals in the given points:

$$\sum_{i=1}^{3} (ax_i + b - f(x_i))^2.$$

The best approximation is the straight line which minimize the latter quantity. Observe that

$$\min_{a,b \in \mathbb{R}} \sum_{i=1}^{3} (ax_i + b - f(x_i))^2 = \min_{a,b \in \mathbb{R}} \left\| \begin{bmatrix} ax_1 + b - f(x_1) \\ ax_2 + b - f(x_2) \\ ax_3 + b - f(x_3) \end{bmatrix} \right\|_2^2 = \min_{a,b \in \mathbb{R}} \left\| A \begin{bmatrix} a \\ b \end{bmatrix} - \begin{bmatrix} f(x_1) \\ f(x_2) \\ f(x_3) \end{bmatrix} \right\|_2^2,$$

with

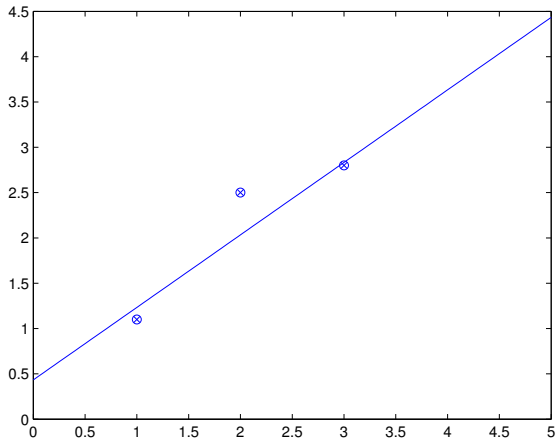
$$A = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ x_3 & 1 \end{bmatrix} \in \mathbb{R}^{3\times 2}.$$

Since applying the square root does not change the minimum point of the previous problem this is actually a linear least squares problem. Since A is full rank we choose the Cholesky method applied to the normal equations, so we compute

$$A^tA = \begin{bmatrix} 14 & 6 \\ 6 & 3 \end{bmatrix}, \qquad A^t \begin{bmatrix} f(x_1) \\ f(x_2) \\ f(x_3) \end{bmatrix} = \begin{bmatrix} 14.5 \\ 6.4 \end{bmatrix}, \qquad A^tA \underbrace{=}_{Cholesky} \begin{bmatrix} \sqrt{14} & 0 \\ \frac{3}{7}\sqrt{14} & \frac{\sqrt{21}}{7} \end{bmatrix} \begin{bmatrix} \sqrt{14} & \frac{3}{7}\sqrt{14} \\ 0 & \frac{\sqrt{21}}{7} \end{bmatrix},$$

$$\begin{bmatrix} \sqrt{14} & 0 \\ \frac{3}{7}\sqrt{14} & \frac{\sqrt{21}}{7} \end{bmatrix} \begin{bmatrix} \sqrt{14} & \frac{3}{7}\sqrt{14} \\ 0 & \frac{\sqrt{21}}{7} \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 14.5 \\ 6.7 \end{bmatrix} \underbrace{\Longrightarrow}_{\text{solve 2 triang. system}} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 0.85 \\ 0.4\bar{3} \end{bmatrix}.$$

We can generalize the latter approach to the case in which we have m evaluations of a function



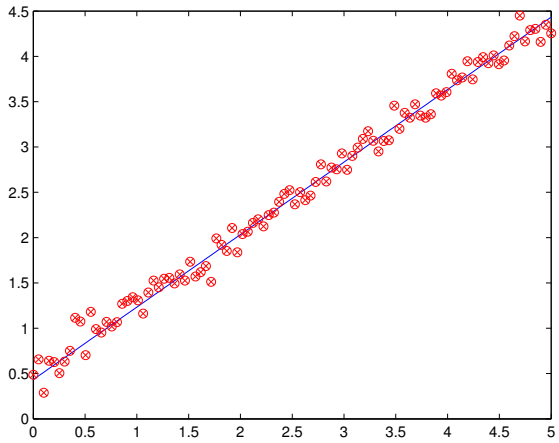
The computed line $0.85x + 0.4\bar{3}$ and the data

$f : \mathbb{R}^n \to \mathbb{R}$. The fitting data are of the form

$$\begin{aligned} \mathbf{x}_1 &= (x_1^{(1)}, \ldots, x_n^{(1)})^t \in \mathbb{R}^n, \quad f(\mathbf{x}_1), \\ &\vdots \\ \mathbf{x}_m &= (x_1^{(m)}, \ldots, x_n^{(m)})^t \in \mathbb{R}^n, \quad f(\mathbf{x}_m). \end{aligned}$$

We look for the hyperplane (linear function form $\mathbb{R}^n$ to $\mathbb{R}$) $a^tx + b$, where $a = (a_1, \ldots, a_n)^t \in \mathbb{R}^n$ and $b \in \mathbb{R}$, which minimize the quantity

$$\left\| \begin{bmatrix} x_1^{(1)} & \ldots & x_n^{(1)} & 1 \\ \vdots & & \vdots & \vdots \\ x_1^{(m)} & \ldots & x_n^{(m)} & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_n \\ b \end{bmatrix} - \begin{bmatrix} f(\mathbf{x}_1) \\ \vdots \\ f(\mathbf{x}_m) \end{bmatrix} \right\|_2$$



Regression line with more fitting points

Another generalization we can make is in the functions we use for fitting. Infact, in order to obtain again a linear least squares problem, is sufficient that the dependence on the parameters we optimize is linear. That is given a finite set of not necessarily linear functions

$$\mathcal{V} = \{\varphi_1, \dots, \varphi_s\}, \qquad \varphi_i : \mathbb{R}^n \to \mathbb{R},$$

I can look for the best linear combination

$$\sum_{i=1}^{s} \alpha_i \varphi_i$$

which approximate f in the 2-norm. This means to solve the linear least squares problem

$$\min_{\alpha \in \mathbb{R}^s} \left\| \begin{bmatrix} \varphi_1(\mathbf{x_1}) & \dots & \varphi_s(\mathbf{x_1}) \\ \vdots & & \vdots \\ \varphi_1(\mathbf{x_m}) & \dots & \varphi_s(\mathbf{x_m}) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_s \end{bmatrix} - \begin{bmatrix} f(\mathbf{x_1}) \\ \vdots \\ f(\mathbf{x_m}) \end{bmatrix} \right\|_2 .$$

Exercise 94. *Solve* $\min_{x \in \mathbb{R}^3} \|Ax - b\|_2$, *where*

$$A = \frac{1}{45} \begin{bmatrix} 14 & 32 & -38 \\ -44 & 58 & 8 \\ -18 & 96 & 51 \\ 63 & -36 & 54 \end{bmatrix}, \qquad b = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

This time we compute the QR factorization of the system getting

$$\frac{1}{45} \left[\begin{array}{ccc|c} 14 & 32 & -38 & 45 \\ -44 & 58 & 8 & 45 \\ -18 & 96 & 51 & 45 \\ 63 & -36 & 54 & 45 \end{array}\right] \xrightarrow[\text{after 3 steps of Householder method}]{} \frac{1}{9\sqrt{257}} \left[\begin{array}{ccc|c} -257 & 244 & -64 & -27 \\ 0 & -306 & -\frac{2124}{27} & -\frac{4221}{17} \\ 0 & 0 & \frac{243\sqrt{257}}{17} & -\frac{9\sqrt{257}}{17} \\ 0 & 0 & 0 & 9\sqrt{257} \end{array}\right].$$

In particular

$$R_1 = \begin{bmatrix} -257 & 244 & -64 \\ 0 & -306 & -\frac{2124}{27} \\ 0 & 0 & \frac{243\sqrt{257}}{17} \end{bmatrix}, \qquad c_1 = \frac{1}{9\sqrt{257}} \begin{bmatrix} 46 \\ 43 \\ -\frac{9\sqrt{257}}{17} \end{bmatrix}.$$

Solving $R_1 x = c_1$ we get $x^* = \frac{1}{56} \begin{bmatrix} -27 \\ -\frac{4221}{17} \\ 2 \end{bmatrix}$ and $\min_{x\in\mathbb{R}^3} \|Ax - b\|_2 = c_2 = 1$.

Exercise 95. *Find the solution with minimum norm of* $\min_{x\in\mathbb{R}^3} \|Ax - b\|_2$ *with*

$$A = \frac{1}{45} \begin{bmatrix} 6 & 12 & -72 \\ -16 & -7 & -8 \\ 58 & 16 & 104 \\ 87 & 24 & 156 \end{bmatrix}, \qquad b = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

The matrix A has rank 2 so we can not apply the Cholesky method. So we proceed considering the normal equations and we apply the LU factorization applying pivoting if necessary.

$$A^tA = \frac{1}{81} \begin{bmatrix} 449 & 128 & 772 \\ 128 & 41 & 184 \\ 772 & 184 & 1616 \end{bmatrix}, \quad A^tb = \begin{bmatrix} 3 \\ 1 \\ 4 \end{bmatrix}, \quad L = \begin{bmatrix} 1 & 0 & 0 \\ \frac{128}{449} & 1 & 0 \\ \frac{772}{449} & -8 & 1 \end{bmatrix}, \quad U = \begin{bmatrix} -449 & 128 & 772 \\ 0 & \frac{2025}{449} & -\frac{16200}{449} \\ 0 & 0 & 0 \end{bmatrix}.$$

We solve $Ly = A^tb$ which is alway solvable (L has ones on its diagonal) and then $Ux = y$ getting infinite solutions of the kind

$$x = \begin{bmatrix} -\frac{1}{5} - 4t \\ \frac{13}{5} + 8t \\ t \end{bmatrix}, \qquad t \in \mathbb{R}.$$

Now we study the function

$$g(t) = \left\| \begin{bmatrix} -\frac{1}{5} - 4t \\ \frac{13}{5} + 8t \\ t \end{bmatrix} \right\|_2^2 = (\frac{1}{5} - 4t)^2 + (\frac{13}{5} + 8t)^2 + t^2.$$

The latter is a convex function and $g'(t) = 0 \quad \Leftrightarrow \quad h = -\frac{4}{15}$ therefore

$$x^* = \frac{1}{15} \begin{bmatrix} 13 \\ 7 \\ -4 \end{bmatrix}, \qquad \min_{x\in\mathbb{R}^3} \|Ax - b\|_2 = \sqrt{2}.$$

It is important to point out that, due to the fact we are working in finite precision, even if A is full rank it could happen that the computed A^tA is not full rank.

Example 96.

$$A = \begin{bmatrix} 3 & 3 \\ 4 & 4 \\ 0 & 10^{-10} \end{bmatrix}, \qquad b = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

$$A^tA = \begin{bmatrix} 25 & 25 \\ 25 & 25 + 10^{-20} \end{bmatrix}, \qquad A^tb = \begin{bmatrix} 7 \\ 7 + 10^{-10}. \end{bmatrix}$$

Since the elements under the machine precision $\approx 10^{-16}$ *are interpreted as* 0 *the computed* A^tA *is* $\begin{bmatrix} 25 & 25 \\ 25 & 25 \end{bmatrix}$ *and the system of normal equations has no solution. In this case is convenient to apply the* QR *factorization approach.*

Lesson 8: Linear algebra

Singular values decomposition

Definition 97. *Let $A \in \mathbb{C}^{m\times n}$ a triple (U, Σ, V) of matrices such that*

$$A = U\Sigma V^h, \qquad U \in \mathbb{C}^{m\times m}, \quad V \in \mathbb{C}^{n\times n}, \quad \Sigma \in \mathbb{C}^{m\times n}, \qquad U^hU = I_m, V^hV = I_n,$$

$$\Sigma = (\sigma_{ij}), \quad \sigma_{ij} = 0 \text{ if } i \neq j \qquad \text{and} \qquad \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min(m,n)} \geq 0,$$

is said a singular values decomposition of A. The columns of U and V are called left and right singular vectors while the elements σ_i are called singular values.

Remark 98. *The SVD of matrix alway exists. Moreover, indicating with u_i and v_i the columns of U and V respectively, we have that*

$$A = U\Sigma V^h = \sum_{i=1}^{\min(m,n)} \sigma_i \underbrace{u_i v_i^h}_{\text{matrix of rank } 1}.$$

Therefore the number of nonzero σ_is is an upper bound for the rank of A. What it turns out is that, due to the orthogonality of the columns of U and V, the latter quantity is exactly the rank of A.

Remark 99.

$$A = U\Sigma V^h \quad \Rightarrow \quad A^hA = V\Sigma^hU^hU\Sigma V^h = V\Sigma^2V^h.$$

Observe that $V\Sigma^2V^h$ is the eigendecomposition of A^hA with decreasing order of the eigenvalues. So we can conclude that the singular values correspond to the square roots of the eigenvalues of A^hA and the matrix V is composed of its eigenvectors. Analogously

$$AA^h = U\Sigma^2U^h,$$

so U is composed of the eigenvectors of AA^h. This give us a method (not efficient) for computing the SVD of A.

Example 100.

$$A = \begin{bmatrix} 2 & 2 \\ -1 & 1 \end{bmatrix} \quad \Rightarrow \quad A^tA = \begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix}, \quad AA^t = \begin{bmatrix} 8 & 0 \\ 0 & 2 \end{bmatrix}.$$

Since

$$\begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 8 & \\ & 2 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}, \qquad \begin{bmatrix} 8 & 0 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 8 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

we have

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \sqrt{8} & 0 \\ 0 & \sqrt{2} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}.$$

Remark 101. *Since the 2-norm is invariant under multiplication of unitary matrices it is easy to see that*

$$\|A\|_2 = \|U\Sigma V^h\|_2 = \|\Sigma\|_2 = \sigma_1.$$

SVD and linear least squares problems

The singular values decomposition allow us give an explicit expression for the solution of linear least squares problem. Suppose $m \geq n$ and that there are $0 \leq k \leq n$ non zero singular values (A is of rank k). Then observe that

$$\|Ax-b\|_2^2 = \|U^hA\Sigma VV^hx - U^hb\|_2^2 \underbrace{=}_{V^hx=y} \|\Sigma y - U^hb\|_2^2 = \sum_{i=1}^{n} |\sigma_i y_i - u_i^h b|^2 + \sum_{i=n+1}^{m} |u_i^h b|^2$$
$$= \sum_{i=1}^{k} |\sigma_i y_i - u_i^h b|^2 + \sum_{i=k+1}^{m} |u_i^h b|^2.$$

So the minimum of this quantity is reached for

$$y_i^* = \begin{cases} \frac{u_i^h b}{\sigma_i} & i = 1, \dots, k \\ 0 & i = k+1, \dots, m \end{cases}$$

Since $x = Vy$ we get $x^* = \sum_{i=1}^{k} \frac{u_i^h b}{\sigma_i} v_i$ and $\min\|Ax-b\| = \sqrt{\sum_{i=k+1}^{m} |u_i^h b|^2}$.

Definition 102 (Moore-Penrose pseudo inverse)**.** *Given $A \in \mathbb{C}^{m\times n}$ with $(U, Sigma, V)$ as SVD, the Moore-Penrose pseudo inverse is defined as*

$$A^+ = V\Sigma^+ U^h, \qquad \textit{where} \qquad \sigma_{ij}^+ = \begin{cases} 0 & i \neq j \\ \sigma_{ii}^{-1} & \sigma_{ii} \neq 0 \\ 0 & \textit{otherwise} \end{cases}.$$

Remark 103. *If $m = n$ and A is invertible then $A^+ = A^{-1}$.*

Using the Moore-Penrose pseudo inverse it is possible to write elegantly the solution of $\min_{x\in\mathbb{R}^n} \|Ax-b\|_2$ as

$$x^* = A^+ x.$$

An application of the SVD: Eigenfaces

Suppose we want to build an automatic procedure for recognizing the face in picture. For example to check if the face in the photo is already contained in a database. The algorithm we are going to see is due to Turk and Pentland [6] and its merit is to be simple and sufficiently effective. Nowdays the proposed techniques has been refined with more advances mathematical tools.

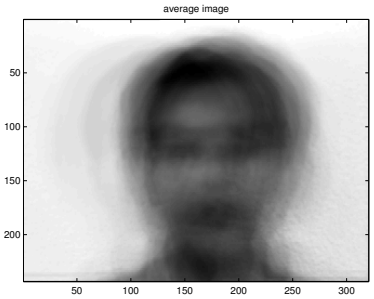
- **Idea:** Use encoding and decoding techniques (truncated SVD) in order to reveal the information contents of pictures and faces. In particular this process should highlight the local and global features of a face. The latters can be realted or not with the physiognomic issues like nose, eyes, lips exc.
- **Purpose:** Encode/represent efficiently the pictures of faces in a database. Recognize faces defining a distance between pictures.

A picture is represent as an $m \times n$-matrix of pixels. We vectorize this representation in a vector of dimension $m \cdot n$ arranging the columns one under the other. Suppose to have a dataset with M images $I_1, \dots, I_M \in \mathbb{R}^{mn}$.

An example of dataset

Instead of starting directly with the images, we consider the difference between each I_i and the average face:

$$\Phi_i = I_i - \frac{1}{M}\sum_{i=1}^{M} I_i, \qquad A = [\Phi_1|\dots|\Phi_M] \in \mathbb{R}^{mn\times M}.$$



The average face

This is not compulsory (everything could work without subtracting the average face) but is done in the paper [6] in order to interpret the matrix AA^t as a covariance matrix. Now we look for an orthonormal basis $\{u_i\}$ of the space spanned by the vectors Φ_is (the faces in the database). Moreover we impose the basis to be the most informative if truncated at a certain

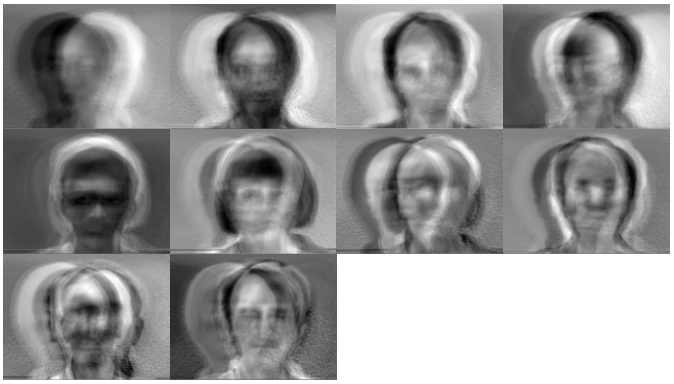
step k. Mathematically this is express by the condition:

$$u_k := \arg\max_{u \in V_k} \sum_{i=1}^{M} (u^t \Phi_i)^2 = \arg\max_{u \in V_k} u^t \underbrace{\sum_{i=1}^{M} \left(\Phi_i \Phi_i^t\right)}_{AA^t} u$$

$$V_k = \{u \in \mathbb{R}^{mn} : \ u_j^t u = 0 \ \forall j = 1, \ldots, k-1\}$$

Note that $u_k^t \Phi_i$ is the k-th coefficient of Φ_i with respect to the basis $\{u_i\}$. What it turns out (Courant-Fisher minimax theorem) is that the optimal value of the previous optimization problem is λ_k, the k-th eigenvalue of AA^t and the maximum point is its associated eigenvector. In particular this correspond to the k-th left singular vector of A.

Since the dataset can be very big I can compute a basis of a subspace with dimension $\bar{k}$ of the space spanned by the vectors Φ_is by computing a truncated SVD of A. I called this generators $u_1, \ldots, u_{\bar{k}}$ eigenfaces.



The $\bar{k} := 10$ eigenfaces

Representation. The picture Φ_i is then represented in the subspace generated by the eigenfaces with the linear combination

$$\Phi_i \approx \sum_{i=1}^{\bar{k}} \omega_j^{(i)} u_j \qquad \text{where } \omega_j^{(i)} = u_j^t \Phi_i, \qquad \Omega_i = (\omega_1^{(i)}, \ldots, \omega_{\bar{k}}^{(i)}).$$

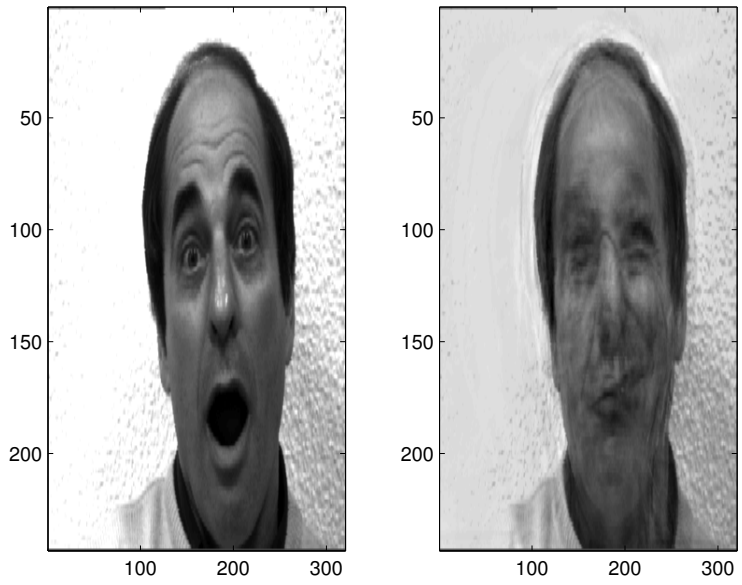
Given an unknown picture I we compute $\Phi = I - \frac{1}{M} \sum_{i=1}^{M} I_i$ and

$$\Phi \approx \sum_{i=1}^{\bar{k}} \omega_j u_j \qquad \text{where } \omega_j = u_j^t \Phi, \qquad \Omega_i = (\omega_1, \ldots, \omega_{\bar{k}}).$$

Recognition. In order to see if the unknown picture matches with some faces in the dataset we compare the vector Ω with the Ω_is by computing

$$\epsilon = \min_{i=1,\ldots,\bar{k}} \|\Omega - \Omega_i\|_2^2.$$

If ϵ is under a certain treshold and i^* is the index where the minimum is attained, then I is recognized as I_i. If not the image is classified as unknow.



An unknown image and its reconstruction on the space generated by the 10 eigenfaces

References

[1] Bevilacqua, R., Bini, D., Capovani, M., & Menchi, O. (1992). Metodi numerici.

[2] Bini, Dario, Milvio Capovani, and Ornella Menchi. "Metodi Numerici Per L'Algebra Lineare." (1997).

[3] Boyd, Stephen, and Lieven Vandenberghe. Convex optimization. Cambridge university press, 2004.

[4] Demmel, James W. Applied numerical linear algebra. Siam, 1997.

[5] Nocedal, Jorge, and Stephen Wright. Numerical optimization. Springer Science & Business Media, 2006.

[6] Turk, M., & Pentland, A. P. (1991, June). Face recognition using eigenfaces. In Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on (pp. 586-591). IEEE.