Mathematical modelling

Lecture notes, version April 5th, 2022

Faculty of Computer and Information Science
University of Ljubljana

2021/22

Chapter 1:

What is Mathematical Modelling?

- Types of models
- Modelling cycle
- Numerical errors

Introduction

Tha task of mathematical modelling is to find and evaluate solutions to real world problems with the use of mathematical concepts and tools.

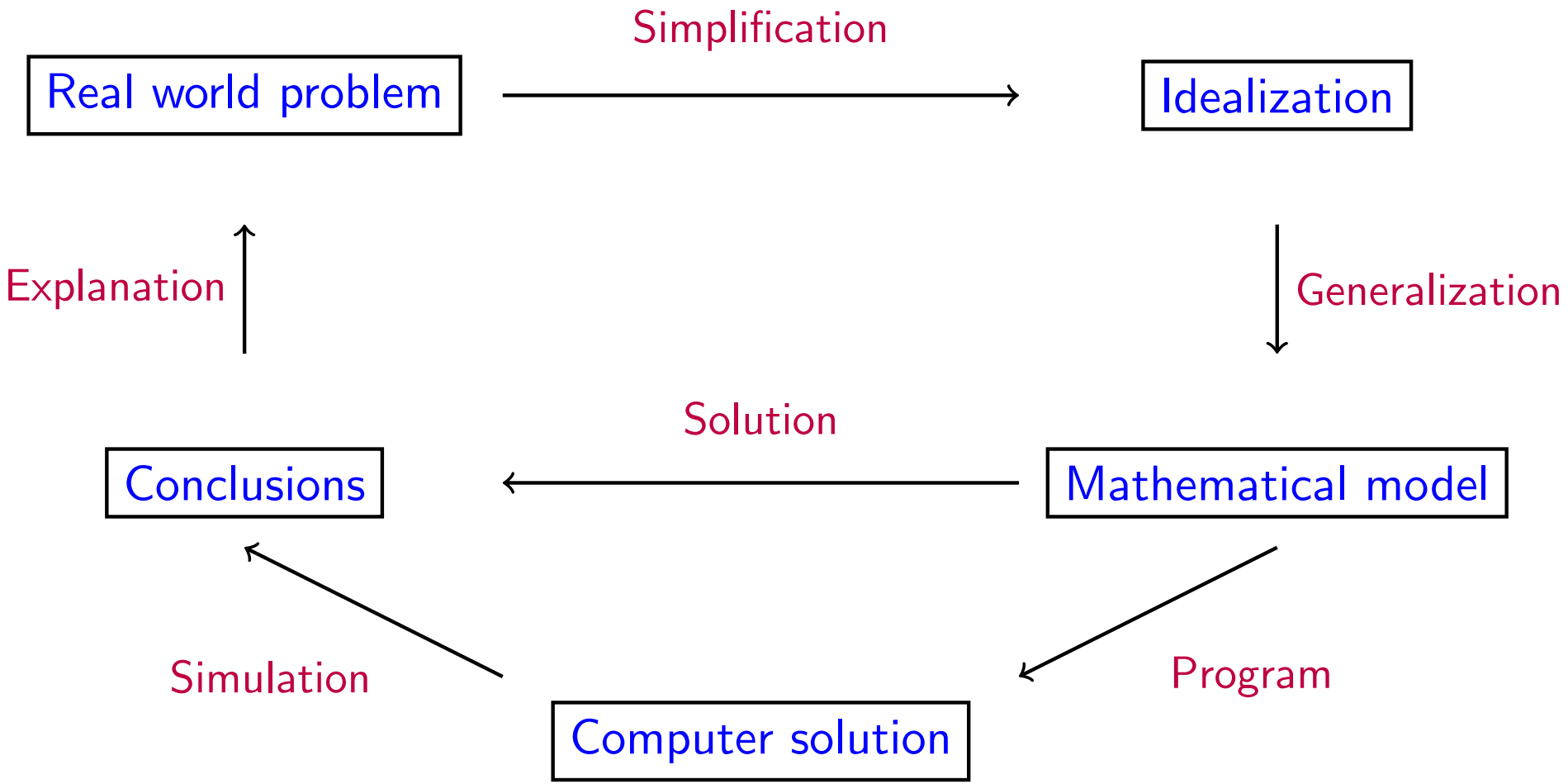
In this course we will introduce some (by far not all) mathematical tools that are used in setting up and solving mathematical models.

We will (together) also solve specific problems, study examples and work on projects.

Contents

- Introduction
- Linear models: systems of linear equations, matrix inverses, SVD decomposition, PCA
- Nonlinear models: vector functions, linear approximation, solving systems of nonlinear equations
- Geometric models: curves and surfaces
- Dynamical models: differential equations, dynamical systems

Modelling cycle



What should we pay attention to?

- Simplification: relevant assumptions of the model (distinguish important features from irrelevant)
- Generalization: choice of mathematical representations and tools (for example: how to represent an object - as a point, a geometric shape, ...)
- Solution: as simple as possible and well documented
- Conclusions: are the results within the expected range, do they correspond to "facts" and experimantal results?

A mathematical model is not universal, it is an approximation of the real world that works only within a certain scale where the assumptions are at least approximately realistic.

Example

An object (ball) with mass m is thrown vertically into the air. What should we pay attention to when modelling its motion?

- The assumptions of the model: relevant forces and parameters (gravitation, friction, wind, ...), how to model the object (a point, a homogeneous or nonhomeogeneous geometric object, angle and rotation in the initial thrust, ...)
- Choice of mathematical model: differential equation, discrete model, ...
- Computation: analytic or numeric, choice of method,...
- Do the results make sense?

Errors

An important part of modelling is estimating the errors!

Errors are an integral part of every model.

Errors come from: assumptions of the model, imprecise data, mistakes in the model, computational precision, errors in numerical and computational methods, mistakes in the computations, mistakes in the programs, ...

Absolute error = Approximate value - Correct value

$$\Delta x = \bar{x} - x$$

$$\text{Relative error} = \frac{\text{Absolute error}}{\text{Correct value}}$$

$$\delta_x = \frac{\Delta x}{x}$$

Example: quadratic equation

$$x^2 + 2a^2x - q = 0$$

Analytic solutions are

$$x_1 = -a^2 - \sqrt{a^4 + q} \quad \text{and} \quad x_2 = -a^2 + \sqrt{a^4 + q}.$$

What happens if $a^2 = 10000$, $q = 1$? Problem with stability in calculating x_2.

More stable way for computing x_2 (so that we do not subtract numbers which are nearly the same) is

$$x_2 = -a^2 + \sqrt{a^4 + q} = \frac{(-a^2 + \sqrt{a^4 + q})(a^2 + \sqrt{a^4 + q})}{a^2 + \sqrt{a^4 + q}}$$

$$= \frac{q}{a^2 + \sqrt{a^4 + q}}.$$

Example of real life disasters

- Disasters caused because of numerical errors: (`http://www-users.math.umn.edu/~arnold//disasters/`)
  - **The Patriot Missile failure, Dharan, Saudi Arabia, February 25 1991**, 28 deaths: bad analysis of rounding errors.
  - **The explosiong of the Ariane 5 rocket, French Guiana, June 4, 1996**: the consequence of overflow in the horizontal velocity.
    `https://www.youtube.com/watch?v=PK_yguLapgA`
    `https://www.youtube.com/watch?v=W3YJeoYgozw`
    `https://www.arianespace.com/vehicle/ariane-5/`
  - **The sinking of the Sleipner offshore platform, Stavanger, Norway, August 12, 1991**, billions of dollars of the loss: inaccurate finite element analysis, i.e., the method for solving partial differential equations.
    `https://www.youtube.com/watch?v=eGdiPs4THW8`

Chapter 2:

Linear model

- Definition
- Systems of linear equations
- Generalized inverses
- The Moore-Penrose (MP) inverse
- Singular value decomposition
- Principal component analysis
- MP inverse and solving linear systems

1. Linear mathematical models

Given points

$$\{(x_1, y_1), \ldots, (x_m, y_m)\}, \quad x_i \in \mathbb{R}^n, \ y_i \in \mathbb{R},$$

the task is to find a function $F(x, a_1, \ldots, a_p)$ that is a good fit for the data.

The values of the parameters $a_1, \ldots, a_p$ should be chosen so that the equations

$$y_i = F(x, a_1, \ldots a_p), \quad i = 1, \ldots, m,$$

are satisfied or, if this is not possible, that the error is as small as possible.

Least squares method: the parameters are determined so that the sum of squared errors

$$\sum_{i=1}^{m} (F(x_i, a_1, \ldots a_p) - y_i)^2$$

is as small as possible.

The mathematical model is linear, when the function F is a linear function of the parameters:

$$F(x, a_1, \ldots, a_p) = a_1\varphi_1(x) + \varphi_2(x) + \cdots + a_p\varphi_p(x),$$

where $\varphi_1, \varphi_2, \ldots \varphi_p$ are functions of a specific type.

Examples of linear models:

1. linear regression: $x, y \in \mathbb{R}$, $\varphi_1(x) = 1, \varphi_2(x) = x$,
2. polynomial regression: $x, y \in \mathbb{R}$, $\varphi_1(x) = 1, \ldots, \varphi_p(x) = x^{p-1}$,
3. multivariate linear regression: $x = (x_1, \ldots, x_n) \in \mathbb{R}^n, y \in \mathbb{R}$,

$$\varphi_1(x) = 1, \varphi_2(x) = x_1, \ldots, \varphi_n(x) = x_n,$$

4. frequency or spectral analysis:

$$\varphi_1(x) = 1, \varphi_2(x) = \cos\omega x, \varphi_3(x) = \sin\omega x, \varphi_4(x) = \cos 2\omega x, \ldots$$

(there can be infinitely many functions $\varphi_i(x)$ in this case)

Examples of nonlinear models: $F(x, a, b) = ae^{bx}$ and $F(x, a, b, c) = \dfrac{a + bx}{c + x}$.

Given the data points $\{(x_1, y_1), \ldots, (x_m, y_m)\}$, $x_i \in \mathbb{R}^n$, $y_i \in \mathbb{R}$, the parameters of a linear model

$$y = a_1\varphi_1(x) + a_2\varphi_2(x) + \cdots + a_p\varphi_p(x)$$

should satisfy the system of linear equations

$$y_i = a_1\varphi_1(x_i) + a_2\varphi_2(x_i) + \cdots + a_p\varphi_p(x_i), \quad i = 1, \ldots, m,$$

or, in a matrix form,

$$\begin{bmatrix} \varphi_1(x_1) & \varphi_2(x_1) & \ldots & \varphi_p(x_1) \\ \varphi_1(x_2) & \varphi_2(x_2) & \ldots & \varphi_p(x_2) \\ \ldots & \ldots & \ldots & \ldots \\ \varphi_1(x_m) & \varphi_2(x_m) & \ldots & \varphi_p(x_m) \end{bmatrix} \begin{bmatrix} a_1 \\ a_1 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} y_1 \\ y_1 \\ \vdots \\ y_p \end{bmatrix}.$$

1.1 Systems of linear equations and generalized inverses

A system of linear equations in the matrix form is given by

$$Ax = b,$$

where

- A is the matrix of coefficients of order $m \times n$ where m is the number of equations and n is the number of unknowns,
- x is the vector of unknowns and
- b is the right side vector.

Existence of solutions:

Let $A = [a_1, \ldots, a_n]$, where a_i are vectors representing the columns of A.

For any vector $x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$ the product Ax is a linear combination

$$Ax = \sum_i x_i a_i.$$

The system is **solvable** if and only if the vector b can be expressed as a linear combination of the columns of A, that is, it is in the column space of A, $b \in \mathcal{C}(A)$.

By adding b to the columns of A we obtain the extended matrix of the system

$$[A \mid b] = [a_1, \ldots, a_n \mid b],$$

Theorem

The system $Ax = b$ is solvable if and only if the rank of A equals the rank of the extended matrix $[A \mid b]$, i.e.,

$$\text{rank } A = \text{rank } [A \mid b] =: r.$$

The solution is unique if the rank of the two matrices equals the number of unknowns, i.e., $r = n$.

An especially nice case is the following:

If A is a square matrix ($n = m$) that has an inverse matrix A^{-1}, the system has a unique solution

$$x = A^{-1}b.$$

Let $A \in \mathbb{R}^{n \times n}$ be a square matrix. The following conditions are equivalent and characterize when a matrix A is invertible or nonsingular:

- The matrix A has an inverse.
- The rank of A equals n.
- $\det(A) \neq 0$.
- The null space $N(A) = \{x : Ax = 0\}$ is trivial.
- All eigenvalues of A are nonzero.
- For each b the system of equations $Ax = b$ has precisely one solution.

A square matrix that does not satisfy the above conditions does not have an inverse.

Example

$$A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \\ 1 & 1 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \\ 1 & 1 & 0 \end{bmatrix}$$

A is invertible and is of rank 3, B is not invertible and is of rank 2.

For a rectangular matrix A of dimension $m \times n$, $m \neq n$, its inverse is not defined (at least in the above sense...).

Definition

A generalized inverse of a matrix $A \in \mathbb{R}^{n \times m}$ is a matrix $G \in \mathbb{R}^{m \times n}$ such that

$$AGA = A. \tag{1}$$

Remark

Note that the dimension of A and its generalized inverse are transposed to each other. This is the only way which enables the multiplication $A \cdot * \cdot A$.

Proposition

If A is invertible, it has a unique generalized inverse, which is equal to A^{-1}.

Proof.

Let G be a generalized inverse of A, i.e., (1) holds. Multiplying (1) with A^{-1} from the left and the right side we obtain:

$$\text{Left hand side (LHS):} \quad A^{-1}AGAA^{-1} = IGI = G,$$
$$\text{Right hand side (RHS):} \quad A^{-1}AA^{-1} = IA^{-1} = A^{-1},$$

where I is the identity matrix. The equality LHS=RHS implies that $G = A^{-1}$.

Theorem

Every matrix $A \in \mathbb{R}^{n\times m}$ has a generalized inverse.

Proof.

Let r be the rank of A.

Case 1. $\operatorname{rank} A = \operatorname{rank} A_{11}$, where

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

and $A_{11} \in \mathbb{R}^{r\times r}, A_{12} \in \mathbb{R}^{r\times(m-r)}, A_{21} \in \mathbb{R}^{(n-r)\times r}, A_{22} \in \mathbb{R}^{(n-r)\times(m-r)}$.
We claim that

$$G = \begin{bmatrix} A_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix},$$

where 0s denote zero matrices of appropriate sizes, is the generalized inverse of A. To prove this claim we need to check that

$$AGA = A.$$

$$AGA = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} A_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} I & 0 \\ A_{21}A_{11}^{-1} & 0 \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$
$$= \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{21}A_{11}^{-1}A_{12} \end{bmatrix}.$$

For AGA to be equal to A we must have

$$A_{21}A_{11}^{-1}A_{12} = A_{22}. \tag{2}$$

It remains to prove (2). Since we are in Case 1, it follows that every column of $\begin{bmatrix} A_{12} \\ A_{22} \end{bmatrix}$ is in the column space of $\begin{bmatrix} A_{11} \\ A_{21} \end{bmatrix}$. Hence, there is a cofficient matrix $W \in \mathbb{R}^{r\times(m-r)}$ such that

$$\begin{bmatrix} A_{12} \\ A_{22} \end{bmatrix} = \begin{bmatrix} A_{11} \\ A_{21} \end{bmatrix} W = \begin{bmatrix} A_{11}W \\ A_{21}W \end{bmatrix}.$$

We obtain the equations $A_{11}W = A_{12}$ and $A_{21}W = A_{22}$. Since A_{11} is invertible, we get $W = A_{11}^{-1}A_{12}$ and hence $A_{21}A_{11}^{-1}A_{12} = A_{22}$, which is (2).

Case 2. *The upper left $r \times r$ submatrix of A is not invertible.*

One way to handle this case is to use permutation matrices P and Q, such that $PAQ = \begin{bmatrix} \widetilde{A}_{11} & \widetilde{A}_{12} \\ \widetilde{A}_{21} & \widetilde{A}_{22} \end{bmatrix}$, $\widetilde{A}_{11} \in \mathbb{R}^{r \times r}$ and $\operatorname{rank} \widetilde{A}_{11} = r$. By Case 1 we have that the generalized inverse $(PAQ)^g$ of PAQ equals to $\begin{bmatrix} \widetilde{A}_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix}$.

Thus,

$$(PAQ) \begin{bmatrix} \widetilde{A}_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix} (PAQ) = PAQ. \tag{3}$$

Multiplying (3) from the left by P^{-1} and from the right by Q^{-1} we get

$$A \left(Q \begin{bmatrix} \widetilde{A}_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix} P \right) A = A.$$

So, $Q \begin{bmatrix} \widetilde{A}_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix} P = \left(P^T \begin{bmatrix} \left(\widetilde{A}_{11}^{-1}\right)^T & 0 \\ 0 & 0 \end{bmatrix} Q^T \right)^T$ is a generalized inverse of A. □

Algorithm for computing a generalized inverse of A

Let r be the rank of A.

1. Find any nonsingular submatrix B in A of order $r \times r$,
2. in A substitute
   - elements of the submatrix B for corresponding elements of $(B^{-1})^T$,
   - all other elements with 0,
3. the transpose of the obtained matrix is a generalized inverse G.

Example

Compute at least one generalized inverse of

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

- Note that $\operatorname{rank} A = 2$. For B from the algorithm one of the possibilities is

$$B = \begin{bmatrix} 1 & 0 \\ 1 & 4 \end{bmatrix},$$

  i.e., the submatrix in the right lower corner.

- Computing B^{-1} we get $B^{-1} = \begin{bmatrix} 1 & 0 \\ -\frac{1}{4} & \frac{1}{4} \end{bmatrix}$ and hence

$$\left(B^{-1}\right)^T = \begin{bmatrix} 1 & -\frac{1}{4} \\ 0 & \frac{1}{4} \end{bmatrix}.$$

- A generalized inverse of A is then

$$G = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -\frac{1}{4} \\ 0 & 0 & 0 & \frac{1}{4} \end{bmatrix}^T = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -\frac{1}{4} & \frac{1}{4} \end{bmatrix}.$$

Generalized inverses of a matrix A play a similar role as the usual inverse (when it exists) in solving a linear system $Ax = b$.

Theorem

Let $A \in \mathbb{R}^{n\times m}$ and $b \in \mathbb{R}^m$. If the system

$$Ax = b \tag{4}$$

is solvable (that is, $b \in \mathcal{C}(A)$) and G is a generalized inverse of A, then

$$x = Gb \tag{5}$$

is a solution of the system (4).

Moreover, all solutions of the system (4) are exaclty vectors of the form

$$x_z = Gb + (GA - I)z, \tag{6}$$

where z varies over all vectors from $\mathbb{R}^m$.

Proof.

We write A in the column form

$$A = \begin{bmatrix} a_1 & a_2 & \dots & a_m \end{bmatrix},$$

where a_i are column vectors of A. Since the system (4) is solvable, there exist real numbers $\alpha_1, \dots, \alpha_m \in \mathbb{R}$ such that

$$\sum_{i=1}^{m} \alpha_i a_i = b. \tag{7}$$

First we will prove that Gb also solves (4). Multiplying (7) with G we get

$$Gb = \sum_{i=1}^{m} \alpha_i G a_i. \tag{8}$$

Multiplying (9) with A the left side becomes $A(Gb)$, so we have to check that

$$\sum_{i=1}^{m} \alpha_i A G a_i = b. \tag{9}$$

Since G is a generalized inverse of A, we have that $AGA = A$ or restricting to columns of the left hand side we get

$$AGa_i = a_i \quad \text{for every } i = 1, \ldots, m.$$

Plugging this into the left side of (9) we get exactly (**??**), which holds and proves (9).

For the moreover part we have to prove two facts:

(i) Any x_z of the form (6) solves (4).

(ii) If $A\tilde{x} = b$, then $\tilde{x}$ is of the form x_z for some $z \in \mathbb{R}^m$.

(i) is easy to check:

$$\begin{aligned} Ax_z &= A\left(Gb + (GA - I)z\right) = AGb + A(GA - I)z \\ &= b + (AGA - A)z = b. \end{aligned}$$

To prove (ii) note that

$$A(\tilde{x} - Gb) = 0,$$

which implies that

$$\tilde{x} - Gb \in \ker A.$$

It remains to check that

$$\ker A = \{(GA - I)z \colon z \in \mathbb{R}^m\}. \tag{10}$$

The inclusion ($\supseteq$) of (10) is straightforward:

$$A((GA - I)z) = (AGA - A)z = 0.$$

For the inclusion ($\subseteq$) of (10) we have to notice that any $v \in \ker A$ is equal to $(GA - I)z$ for $z = -v$:

$$(GA - I)(-v) = -GAv + v = 0 + v = v. \quad \square$$

Example

Find all solutions of the system

$$Ax = b,$$

where $A = \begin{bmatrix} 0 & 0 & 2 & 0 \\ 0 & 0 & 1 & 0 \\ 2 & 0 & 1 & 4 \end{bmatrix}$ and $b = \begin{bmatrix} 2 \\ 1 \\ 4 \end{bmatrix}$.

- Recall from the example a few slides above that $G = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -\frac{1}{4} & \frac{1}{4} \end{bmatrix}$.
- Calculating Gb and $GA - I$ we get

$$Gb = \begin{bmatrix} 0 \\ 0 \\ 1 \\ \frac{3}{4} \end{bmatrix} \quad \text{and} \quad A = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 \end{bmatrix}.$$

- Hence,

$$x_z = \begin{bmatrix} -z_1 & -z_2 & 1 & \frac{3}{4} + \frac{1}{2}z_1 \end{bmatrix}^T$$

where z_1, z_2 vary over $\mathbb{R}$.

1.2 The Moore-Penrose generalized inverse

Among all generalized inverses of a matrix A, one has especially nice properties.

Definition

The Moore-Penrose generalized inverse, or shortly the MP inverse of $A \in \mathbb{R}^{n\times m}$ is any matrix $A^+ \in \mathbb{R}^{m\times n}$ satifying the following four conditions:

1. A^+ is a generalized inverse of A: $AA^+A = A$.
2. A is a generalized inverse of A^+: $A^+AA^+ = A^+$.
3. The square matrix $AA^+ \in \mathbb{R}^{n\times n}$ is symmetric: $(AA^+)^T = AA^+$.
4. The square matrix $A^+A \in \mathbb{R}^{m\times m}$ is symmetric: $(A^+A)^T = A^+A$.

Remark

There are two natural questions arising after defining the MP inverse:

- *Does every matrix admit a MP inverse? Yes.*
- *Is the MP inverse unique? Yes.*

Theorem

The MP inverse A^+ of a matrix A is unique.

Proof.

Assume that there are two matrices M_1 and M_2 that satisfy the four conditions in the definition of MP inverse of A. Then,

$$\begin{aligned} AM_1 &= (AM_2A)M_1 && \text{by property (1)} \\ &= (AM_2)(AM_1) = (AM_2)^T(AM_1)^T && \text{by property (3)} \\ &= M_2^T(AM_1A)^T = M_2^TA^T && \text{by property (1)} \\ &= (AM_2)^T = AM_2 && \text{by property (3)} \end{aligned}$$

A similar argument involving properties (2) and (4) shows that

$$M_1A = M_2A,$$

and so

$$M_1 = M_1AM_1 = M_1AM_2 = M_2AM_2 = M_2.$$

Remark

Let us assume that A^+ exists (we will shortly prove this fact). Then the following properties are true:

- *If A is a square invertible matrix, then it $A^+ = A^{-1}$.*
- $(A^+)^+ = A$.
- $(A^T)^+ = (A^+)^T$.

In the rest of this chapter we will be interested in two obvious questions:

- How do we compute A^+?
- Why would we want to compute A^+?

To answer the first question, we will begin by three special cases.

Construction of the MP inverse of $A \in \mathbb{R}^{n\times m}$:

Case 1: $A^TA \in \mathbb{R}^{m\times m}$ *is an invertible matrix. (In particular,* $m \leq n$*.)*

In this case $A^+ = (A^TA)^{-1}A^T$.

To see this, we have to show that the matrix $(A^TA)^{-1}A^T$ satisfies properties (1) to (4):

1. $AMA = A(A^TA)^{-1}A^TA = A(A^TA)^{-1}(A^TA) = A.$
2. $MAM = (A^TA)^{-1}A^TA(A^TA)^{-1}A^T = (A^TA)^{-1}A^T = M.$
3. $$(AM)^T = \left(A(A^TA)^{-1}A^T\right)^T = A\left(\left(A^TA\right)^{-1}\right)^T A^T =$$ $$= A\left(\left(A^TA\right)^T\right)^{-1} A^T = A(A^TA)^{-1}A^T = AM.$$
4. Analoguous to the previous fact.

Case 2: *AA^T is an invertible matrix. (In particular, $n \leq m$.)*

In this case A^T satisfies the condition for Case 1, so $(A^T)^+ = (AA^T)^{-1}A$.

Since $(A^T)^+ = (A^+)^T$ it follows that

$$\begin{aligned} A^+ &= \left((A^+)^T\right)^T = \left((AA^T)^{-1}A\right)^T = A^T\left((AA^T)^{-1}\right)^T \\ &= A^T\left((AA^T)^{-T}\right)^{-1} = A^T(AA^T)^{-1}. \end{aligned}$$

Hence, $\underline{A^+ = A^T(AA^T)^{-1}}$.

Case 3: $\Sigma \in \mathbb{R}^{n\times m}$ *is a diagonal matrix of the form*

$$\Sigma = \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_n \end{bmatrix} \quad \text{or} \quad \widetilde{\Sigma} = \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_m \end{bmatrix}.$$

The MP inverse is

$$\Sigma^+ = \begin{bmatrix} \sigma_1^+ & & & \\ & \sigma_2^+ & & \\ & & \ddots & \\ & & & \sigma_n^+ \end{bmatrix} \quad \text{or} \quad \widetilde{\Sigma}^+ = \begin{bmatrix} \sigma_1^+ & & & \\ & \sigma_2^+ & & \\ & & \ddots & \\ & & & \sigma_m^+ \end{bmatrix},$$

where $\sigma_i^+ = \begin{cases} \frac{1}{\sigma_i}, & \sigma_i \neq 0, \\ 0, & \sigma_i = 0. \end{cases}$

Case 4: *A general matrix A. (using SVD)*

Theorem (Singular value decomposition - SVD)

Let $A \in \mathbb{R}^{n\times m}$ be a matrix. Then it can be expressed as a product

$$A = U\Sigma V^T,$$

where

- *$U \in \mathbb{R}^{n\times n}$ is an orthogonal matrix with left singular vectors u_i as its columns,*
- *$V \in \mathbb{R}^{m\times m}$ is an orthogonal matrix with right singular vectors v_i as its columns,*
- $\Sigma = \left[\begin{array}{ccc|c} \sigma_1 & & & 0 \\ & \ddots & & \vdots \\ & & \sigma_r & 0 \\ \hline 0 & & & 0 \end{array}\right] = \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{n\times m}$ *is a diagonal matrix*

with singular values

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$$

on the diagonal.

Derivations for computing SVD

If $A = U\Sigma V^T$, then

$$A^TA = (V\Sigma^TU^T)(U\Sigma V^T) = V\Sigma^T\Sigma V^T = V\begin{bmatrix} S^2 & 0 \\ 0 & 0 \end{bmatrix}V^T \in \mathbb{R}^{m\times m},$$

$$AA^T = (U\Sigma V^T)(U\Sigma V^T)^T = U\Sigma\Sigma^TU^T = U\begin{bmatrix} S^2 & 0 \\ 0 & 0 \end{bmatrix}U^T \in \mathbb{R}^{n\times n}.$$

Let

$$V = \begin{bmatrix} v_1 & v_2 & \cdots & v_m \end{bmatrix} \quad \text{and} \quad U = \begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix}$$

be the column decompositions of V and U.

Let $e_1, \ldots, e_m \in \mathbb{R}^m$ and $f_1, \ldots, f_n \in \mathbb{R}^n$ be the standard coordinate vectors of $\mathbb{R}^m$ and $\mathbb{R}^n$, i.e., the only nonzero component of e_i (resp. f_j) is the i-th one (resp. j-th one), which is 1. Then

$$A^TAv_i = V\Sigma^T\Sigma V^Tv_i = V\Sigma^T\Sigma e_i = \begin{cases} \sigma_i^2 v_i, & \text{if } i \le r, \\ 0, & \text{if } i > r, \end{cases}$$

$$AA^Tu_j = U\Sigma\Sigma^TU^Tu_j = U\Sigma\Sigma^Tf_j = \begin{cases} \sigma_i^2 u_j, & \text{if } j \le r, \\ 0, & \text{if } j > r. \end{cases}$$

Further on,

$$(AA^T)(Av_i) = A(A^TA)v_i = \begin{cases} \sigma_i^2 Av_i, & \text{if } i \leq r, \\ 0, & \text{if } i > r, \end{cases}$$

$$(A^TA)(A^Tu_j) = A^T(AA^T)u_j = \begin{cases} \sigma_j^2 A^Tu_j, & \text{if } j \leq r, \\ 0, & \text{if } j > r. \end{cases}$$

It follows that:

- $\Sigma^T\Sigma = \begin{bmatrix} S^2 & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{m\times m}$ (resp. $\Sigma\Sigma^T = \begin{bmatrix} S^2 & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{n\times n}$) is the diagonal matrix with eigenvalues σ_i^2 of A^TA (resp. AA^T) on its diagonal, so the singular values σ_i are their square roots.
- V has the corresponding eigenvectors (normalized and pairwise orthogonal) of A^TA as its columns, so the right singular vectors are eigenvectors of A^TA.
- U has the corresponding eigenvectors (normalized and pairwise orthogonal) of AA^T as its columns, so the left singular vectors are eigenvectors of AA^T.

- Av_i is an eigenvector of AA^T corresponding to σ_i^2 and so

$$u_i = \frac{Av_i}{\|Av_i\|} = \frac{Av_i}{\sigma_i}$$

is a left singular vector corresponding to σ_i, where in the second equality we used that

$$\|Av_i\| = \sqrt{(Av_i)^T(Av_i)} = \sqrt{v_i^T A^T A v_i} = \sqrt{\sigma_i^2 v_i^T v_i} = \sigma_i \|v_i\| = \sigma_i.$$

- $A^T u_j$ is an eigenvector of $A^T A$ corresponding to σ_j^2 and so

$$v_j = \frac{A^T u_j}{\|A^T u_j\|} = \frac{A^T u_j}{\sigma_j}$$

is a right singular vector corresponding to σ_j, where in the second equality we used that

$$\|A^T u_j\| = \sqrt{(A^T u_j)^T(A^T u_j)} = \sqrt{u_j^T A A^T u_j} = \sqrt{\sigma_j^2 u_j^T u_j} = \sigma_j \|u_j\| = \sigma_j.$$

Algorithm for SVD computation

- Compute the eigenvalues and an orthonormal basis consisting of eigenvectors of the symmetric matrix $A^T A$ or AA^T (depending on which is of them is of smaller size).
- The singular values of the matrix $A \in \mathbb{R}^{n \times m}$ are equal to $\sigma_i = \sqrt{\lambda_i}$, where λ_i are the nonzero eigenvalues of $A^T A$ (resp. AA^T).
- The left singular vectors are the corresponding orthonormal eigenvectors of AA^T.
- The right singular vector are the corresponding orthonormal eigenvectors of $A^T A$.
- If u (resp. v) is a left (resp. right) singular vector corresponding to the singular value σ_i, then $v = Au$ (resp. $u = A^T v$) is a right (resp. left) singular vector corresponding to the same singular value.
- The remaining columns of U (resp. V) consist of an orthonormal basis of the kernel (i.e., the eigenspace of $\lambda = 0$) of AA^T (resp. $A^T A$).

General algorithm for computation of A^+ (long version)

1. For A^TA compute its eigenvalues

$$\lambda_1 \geq \lambda_2 \geq \cdots, \geq \lambda_r > \lambda_{r+1} = \ldots = \lambda_m = 0$$

and the corresponding orthonormal eigenvectors

$$v_1, \ldots, v_r, v_{r+1}, \ldots, v_m,$$

and form the matrices

$$\Sigma = \operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_m}) \in \mathbb{R}^{n \times m},$$

$$V_1 = \begin{bmatrix} v_1 & \cdots & v_r \end{bmatrix}, \quad V_2 = \begin{bmatrix} v_{r+1} & \cdots & v_m \end{bmatrix} \quad \text{and} \quad V = \begin{bmatrix} V_1 & V_2 \end{bmatrix}.$$

2. Let

$$u_1 = \frac{Av_1}{\sigma_1}, \quad u_2 = \frac{Av_2}{\sigma_2}, \quad \ldots \quad , \quad u_r = \frac{Av_r}{\sigma_r},$$

and $u_{r+1}, \ldots, u_n$ vectors, such that $\{u_1, \ldots, u_n\}$ is an ortonormal basis for $\mathbb{R}^n$. Form the matrices

$$U_1 = \begin{bmatrix} u_1 & \cdots & u_r \end{bmatrix}, \quad U_2 = \begin{bmatrix} u_{r+1} & \cdots & u_n \end{bmatrix} \quad \text{and} \quad U = \begin{bmatrix} U_1 & U_2 \end{bmatrix}.$$

3. Then

$$A^+ = V\Sigma^+U^T.$$

General algorithm for computation of A^+ (short version)

1. For $A^T A$ compute its **nonzero** eigenvalues

$$\lambda_1 \geq \lambda_2 \geq \cdots, \geq \lambda_r > 0$$

and the corresponding orthonormal eigenvectors

$$v_1, \ldots, v_r,$$

and form the matrices

$$S = \operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_r}) \in \mathbb{R}^{r \times r},$$
$$V_1 = \begin{bmatrix} v_1 & \cdots & v_r \end{bmatrix} \in \mathbb{R}^{m \times r}.$$

2. Put the vectors

$$u_1 = \frac{Av_1}{\sigma_1}, \quad u_2 = \frac{Av_2}{\sigma_2}, \quad \ldots \quad , \quad u_r = \frac{Av_r}{\sigma_r}$$

in the matrix

$$U_1 = \begin{bmatrix} u_1 & \cdots & u_r \end{bmatrix}.$$

3. Then

$$A^+ = V_1 \Sigma^+ U_1^T.$$

Correctness of the computation of A^+

Step 1. $V\Sigma^+U^T$ is equal to A^+.

(i) $AA^+A = A$:

$$\begin{aligned} AA^+A &= (U\Sigma V^T)(V\Sigma^+U^T)(U\Sigma V^T) = U\Sigma(V^TV)\Sigma^+(U^TU)\Sigma V^T \\ &= U\Sigma\Sigma^+\Sigma V^T = U\Sigma V^T = A. \end{aligned}$$

(ii) $A^+AA^+ = A^+$: Analoguous to (i).

(iii) $(AA^+)^T = AA^+$:

$$\begin{aligned} (AA^+)^T &= \Big((U\Sigma V^T)(V\Sigma^+U^T)\Big)^T = \Big(U\Sigma\Sigma^+U^T\Big)^T \\ &= \left(U\begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}U^T\right)^T = U\begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}U^T \\ &= (U\Sigma V^T)(V\Sigma^+U^T) = A^+. \end{aligned}$$

(iv) $(A^+A)^T = A^+A$: Analoguous to (iii).

Step 2. $V\Sigma^+U^T$ is equal to $V_1\Sigma^+U_1^T$.

$$V\Sigma U^T = \begin{bmatrix} V_1 & V_2 \end{bmatrix} \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix} = \begin{bmatrix} V_1 S & 0 \end{bmatrix} \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix} = V_1 S U_1^T.$$

Example

Compute the SVD and A^+ of the matrix $A = \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{bmatrix}$.

- $AA^T = \begin{bmatrix} 17 & 8 \\ 8 & 17 \end{bmatrix}$ has eigenvalues 25 and 9.
- The eigenvectors of AA^T corresponding to the eigenvalues 25, 9 are

$$u_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}^T, \qquad u_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}^T.$$

- The left singular vectors of A are

$$v_1 = \frac{A^T u_1}{\sigma_1} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \end{bmatrix}^T, \qquad v_2 = \frac{A^T u_2}{\sigma_2} = \begin{bmatrix} \frac{1}{3\sqrt{2}} & -\frac{1}{3\sqrt{2}} & \frac{4}{3\sqrt{2}} \end{bmatrix}^T.$$

$$v_3 = v_1 \times v_2 = \begin{bmatrix} \frac{2}{\sqrt{3}} & -\frac{2}{3} & -\frac{1}{3} \end{bmatrix}^T.$$

- $$A = U\Sigma V^T = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 5 & 0 & 0 \\ 0 & 3 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{3\sqrt{2}} & -\frac{1}{3\sqrt{2}} & \frac{4}{3\sqrt{2}} \\ \frac{2}{\sqrt{3}} & -\frac{2}{3} & -\frac{1}{3} \end{bmatrix}.$$
- $$A^+ = V\Sigma^+ U^T = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{3\sqrt{2}} & \frac{2}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{3\sqrt{2}} & -\frac{2}{3} \\ 0 & \frac{4}{3\sqrt{2}} & -\frac{1}{3} \end{bmatrix} \begin{bmatrix} \frac{1}{5} & 0 \\ 0 & \frac{1}{3} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}$$
$$= \begin{bmatrix} \frac{7}{45} & \frac{2}{45} \\ \frac{2}{45} & \frac{7}{45} \\ \frac{2}{9} & -\frac{2}{9} \end{bmatrix}.$$

1.3 The MP inverse and systems of linear equations

Let $A \in \mathbb{R}^{n\times m}$, where $m > n$. A system of equations $Ax = b$ that has more variables than constraints. Typically such system has infinitely many solutions, but it may happen that it has no solutions. We call such system an underdetermined system.

Theorem

1. *An underdetermined system of linear equations*

$$Ax = b \tag{11}$$

   is solvable if and only if $AA^+b = b$.

2. *If there are infinitely many solutions, the solution* A^+b *is the one with the smallest norm, i.e.,*

$$\|A^+b\| = \min\{\|x\| : Ax = b\}.$$

   Moreover, it is the unique solution of smallest norm.

Proof of Theorem.

We already know that $Ax = b$ is solvable iff Gb is a solution, where G is any generalized inverse of A. Since A^+ is one of the generalized inverses, this proves the first part of the theorem.

To prove the second part of the theorem, first recall that all the solutions of the system are precisely the set

$$\{A^+b + (A^+A - I)z : z \in \mathbb{R}^m\}.$$

So we have to prove that for every $z \in \mathbb{R}^m$,

$$\|A^+b\| \leq \|A^+b + (A^+A - I)z\|.$$

We have that:

$$\begin{aligned}
&\|A^+b + (A^+A - I)z\|^2 = \\
&= \left(A^+b + (A^+A - I)z\right)^T \left(A^+b + (A^+A - I)z\right) \\
&= \left(A^+b\right)^T \left(A^+b\right) + 2\left(A^+b\right)^T (A^+A - I)z + \left((A^+A - I)z\right)^T \left((A^+A - I)z\right) \\
&= \|A^+b\|^2 + 2\left(A^+b\right)^T (A^+A - I)z + \|(A^+A - I)z\|^2
\end{aligned}$$

Now,

$$\begin{aligned}\left(A^+b\right)^T(A^+A-I)z &= b^T(A^+)^T(A^+A-I)z\\ &= b^T(A^+)^T(A^+A)^Tz - b^T(A^+)^Tz\\ &= b^T\left((A^+A)A^+\right)^Tz - b^T(A^+)^Tz\\ &= b^T\left(A^+AA^+\right)^Tz - b^T(A^+)^Tz\\ &= b^T(A^+)^Tz - b^T(A^+)^Tz = 0,\end{aligned}$$

where we used the fact $(A^+A)^T = A^+A$ in the second equality.

Thus,

$$\|A^+b + (A^+A-I)z\|^2 = \|A^+b\|^2 + \|(A^+A-I)z\|^2 \geq \|A^+b\|^2,$$

with the equality iff $(A^+A-I)z = 0$. This proves the second part of the theorem. □

Example

- The solutions of the underdetermined system $x + y = 1$ geometrically represent an affine line. Matricially, $A = \begin{bmatrix} 1 & 1 \end{bmatrix}$, $b = 1$. Hence, $A^+ b = A^+ 1$ is the point on the line, which is the nearest to the origin. Thus, the vector of this point is perpendicular to the line.
- The solutions of the underdetermined system $x + 2y + 3z = 5$ geometrically represent an affine hyperplane. Matricially, $A = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}$, $b = 5$. Hence, $A^+ b = A^+ 5$ is the point on the hyperplane, which is the nearest to the origin. Thus, the vector of this point is normal to the hyperplane.
- The solutions of the underdetermined system $x + y + z = 1$ and $x + 2y + 3z = 5$ geometrically represent an affine line in $\mathbb{R}^3$. Matricially, $A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix}$, $b = \begin{bmatrix} 1 \\ 5 \end{bmatrix}$. Hence, $A^+ b$ is the point on the line, which is the nearest to the origin. Thus, the vector of this point is perpendicular to the line.

Example

Find the point on the plane $3x + y + z = 2$ closest to the origin.

- In this case,
$$A = \begin{bmatrix} 3 & 1 & 1 \end{bmatrix} \quad \text{and} \quad b = [2].$$
- We have that $AA^T = [11]$ and hence its only eigenvalue is $\lambda = 11$ with eigenvector $u = [1]$, implying that
$$U = [1] \quad \text{and} \quad \Sigma = \begin{bmatrix} \sqrt{11} & 0 & 0 \end{bmatrix}.$$
- Hence,
$$v_1 = \frac{A^T u}{\|A^T u\|} = \frac{A^T u}{\sigma_1} = \frac{1}{\sqrt{11}} \begin{bmatrix} 3 & 1 & 1 \end{bmatrix}^T.$$
- $$A^+ = V\Sigma^+ U^T = \frac{1}{\sqrt{11}} \begin{bmatrix} 3 \\ 1 \\ 1 \end{bmatrix} \frac{1}{\sqrt{11}} [1] = \begin{bmatrix} \frac{3}{11} \\ \frac{1}{11} \\ \frac{1}{11} \end{bmatrix}.$$
- $$x^+ = A^+ b = \begin{bmatrix} \frac{6}{11} & \frac{2}{11} & \frac{2}{11} \end{bmatrix}^T.$$

Overdetermined systems

Let $A \in \mathbb{R}^{n \times m}$, where $n > m$. This system is called overdetermined, since here are more constraints than variables. Such a system typically has no solutions, but it might have one or even infinitely many solutions.

Least squares approximation problem: if the system $Ax = b$ has no solutions, then a best fit for the solution is a vector x such that the error $||Ax - b||$ or, equivalently in the row decomposition

$$A = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix},$$

its square

$$||Ax - b||^2 = \sum_{i=1}^{n} (\alpha_i x - b_i)^2,$$

is the smallest possible.

Theorem

If the system $Ax = b$ has no solutions, then $x^+ = A^+ b$ is the unique solution to the least squares approximation problem:

$$||Ax^+ - b|| = \min\{\|Ax - b\| : x \in \mathbb{R}^n\}.$$

Proof.

Let $A = U\Sigma V^T$ be the SVD decomposition of A. We have that

$$\|Ax - b\| = \|U\Sigma V^T - b\| = \|\Sigma V^T - U^T b\|,$$

where we used that

$$\|U^T v\| = \|v\|$$

in the second equality (which holds since U^T is an orthogonal matrix). Let

$$\Sigma = \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix}, \quad U = \begin{bmatrix} U_1 & U_2 \end{bmatrix}, \quad V = \begin{bmatrix} V_1 & V_2 \end{bmatrix}, \quad \text{where}$$

$S \in \mathbb{R}^{r\times r}$, $U_1 \in \mathbb{R}^{n\times r}$, $U_2 \in \mathbb{R}^{n\times(n-r)}$, $V_1 \in \mathbb{R}^{m\times r}$, $V_2 \in \mathbb{R}^{m\times(m-r)}$.

Thus,

$$\|\Sigma V^T - U^T b\| = \left\| \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix} x - \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix} b \right\|$$
$$= \left\| \begin{bmatrix} SV_1^T x - U_1^T b \\ U_2^T b \end{bmatrix} \right\|.$$

But this norm is minimal iff

$$SV_1^T x - U_1^T b = 0$$

or equivalently

$$x = V_1 S^{-1} U_1^T b = A^+ b.$$

□

Remark

The closest vector to b in the column space $C(A) = \{Ax : x \in \mathbb{R}^m\}$ of A is the orthogonal projection of b onto $C(A)$. It follows that $A^+ b$ is this projection. Equivalently, $b - (A^+ b)$ is orthogonal to any vector Ax, $x \in \mathbb{R}^m$, which can be proved also directly.

Example

Given points $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ in the plane, we are looking for the line $ax + b = y$ which is the least squares best fit.

If $n > 2$, we obtain an overdetermined system

$$\begin{bmatrix} x_1 & 1 \\ \vdots & \\ x_n & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}.$$

The solution of the least squares approximation problem is given by

$$\begin{bmatrix} a \\ b \end{bmatrix} = A^{+} \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}.$$

The line $y = ax + b$ in the regression line.

An application of SVD: principal component analysis or PCA

PCA is a very well-known and efficient method for data compression, dimension reduction, ...

Due to its importance in different fields, it has many other names: discrete Karhunen-Loève transform (KLT), Hotelling transform, empirical orthogonal functions (EOF), ...

Let $\{X_1, \ldots, X_m\}$ be a sample of vectors from $\mathbb{R}^n$.

In applications, often $m << n$, where n is very large, for example, $X_1, \ldots, X_m$ can be

- vectors of gene expressions in m tissue samples or
- vectors of grayscale in images
- bag of words vectors, with components corresponding to the numbers of certain words from some dictionary in specific texts, ...,

or $n << m$ for example if the data represents a point cloud in a low dimensional space $\mathbb{R}^n$ (for example in the plane).

We will assume that $m << n$. Also assume that the data is centralized, i.e., the centeroid is in the origin

$$\mu = \frac{1}{m}\sum_{i=1}^{m} X_i = 0 \in \mathbb{R}^n.$$

If not, we substract μ from all vectors in the data set.

A matrix norm $\|\cdot\| : \mathbb{R}^{n\times m} \to \mathbb{R}$ is a function, which generalizes the notion of the absolute value for numbers to matrices. It is used to measure a distance between matrices. In contrast with the absolute value, which is unique up to multiplication with a positive constant, there are many different matrix norms.

Two important matrix norms are the following:

1. Spectral norm $\|\cdot\|_2$:

$$\|A\|_2 := \max_{\|x\|_2=1} \|Ax\|_2 = \max_{j=1,\ldots,\min(n,m)} \sigma_j(A).$$

2. Frobenius norm $\|\cdot\|_F$:

$$\|A\|_F := \sqrt{\sum_{i,j} a_{i,j}^2} = \sqrt{\sum_{j=1,\ldots,\min(n,m)} \sigma_j(A)^2}.$$

Let

$$X = \begin{bmatrix} X_1 & X_2 & \cdots & X_m \end{bmatrix}^T$$

be the matrix of dimension $m \times n$ with data in the rows.

Let $X^T X \in \mathbb{R}^{m \times m}$ and $XX^T \in \mathbb{R}^{n \times n}$ be the covariance matrices of the data.

- The principal values of the data set $\{X_1, \ldots, X_r\}$ are the nonzero eigenvalues $\lambda_i = \sigma_i^2$ of the covariance matrices (where σ_i are the singular values of X).
- The principal directions in $\mathbb{R}^n$ are corresponding eigenvectors $v_1, \ldots, v_r$, i.e. the columns of the matrix V from the SVD of X. The remaining clolumns of V (i.e. the eigenvectors correspondong to 0) form a basis of the null space of X.
- The first column v_1, the first principal direction, corresponds to the direction in $\mathbb{R}^n$ with the largest variance in the data X_i, that is, the most informative direction for the data set, the second the second most important, ...
- The principal directions in $\mathbb{R}^m$ are the columns $u_1, \ldots, u_r$ of the matrix U and represent the coefficients in the linear decomposition of the vectors $X_1, \ldots, X_m$ along the orthonormal basis $v_1, \ldots v_n$ of $\mathbb{R}^n$.

PCA provides a linear dimension reduction method based on a projection of the data from the space $\mathbb{R}^n$ into a lower dimensional subspace spanned by the first few principal vectors $v_1, \ldots, v_k$ in $\mathbb{R}^n$.

The idea is to approximate

$$X_i = \sigma_1 u_{1,i} v_1 + \cdots + \sigma_m u_{m,i} v_m \cong \sigma_1 u_{1,i} v_1 + \cdots + \sigma_k u_{k,i} v_k$$

with the first k most informative directions in $\mathbb{R}^n$ and supress the last $m - k$.

PCA has the following amazing property:

Theorem

Among all possible projections of $p : \mathbb{R}^n \to \mathbb{R}^k$ onto a k-dimensional subspace, PCA provides the best in the sense that the errors

$$\|X - p(X)\|_F^2 \qquad \textit{and} \qquad \|X - p(X)\|_2^2,$$

where $p(X) = \begin{bmatrix} p(X_1) & \cdots & p(X_m) \end{bmatrix}^T$, are the smallest possible.

Chapter 3:

Nonlinear models

- Definition and examples
- Systems of nonlinear equations
- Vector functions of vector variables
  - Derivative and Jacobian matrix
  - Linear approximation
- Newton's method for square systems
  - Univariate case: Tangent method
  - Use in optimization
- Gauss-Newton's method for rectangular systems

3. Nonlinear models

General formulation

Given is a sample of points $\{(x_1, y_1), \ldots, (x_m, y_m)\}$, $x_i \in \mathbb{R}^n$, $y_i \in \mathbb{R}$.

The mathematical model is nonlinear if the function

$$y = F(x, a_1, \ldots, a_p) \tag{12}$$

is a nonlinear function of the parameters a_i. This means it cannot be written in the form

$$y = a_1 f_1(x) + a_2 f_2(x) + \ldots + a_p f_p(x),$$

where each $f_i : \mathbb{R}^n \to \mathbb{R}$ is some function.

Plugging each data points into (12) we obtain a system of nonlinear equations

$$\begin{aligned} y_1 &= F(x_1, a_1, \ldots, a_p), \\ &\vdots \\ y_m &= F(x_m, a_1, \ldots, a_p), \end{aligned} \tag{13}$$

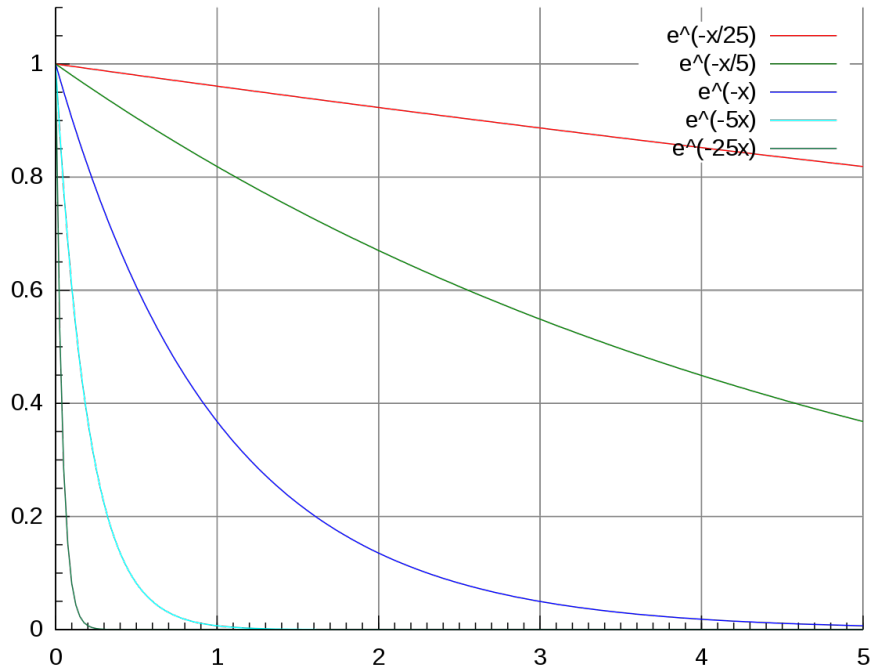
in the parameters $a_1, \ldots, a_p \in \mathbb{R}$.

1. Exponential decay or growth: $F(x, a, k) = ae^{kx}$, a and k are parameters.

   A quantity y changes at a rate proportional to its current value, which can be described by the differential equation

   $$\frac{dy}{dx} = ky.$$

   The solution to this equation (obtained by the use of separation of variables) is $y = F(x, a, k)$.

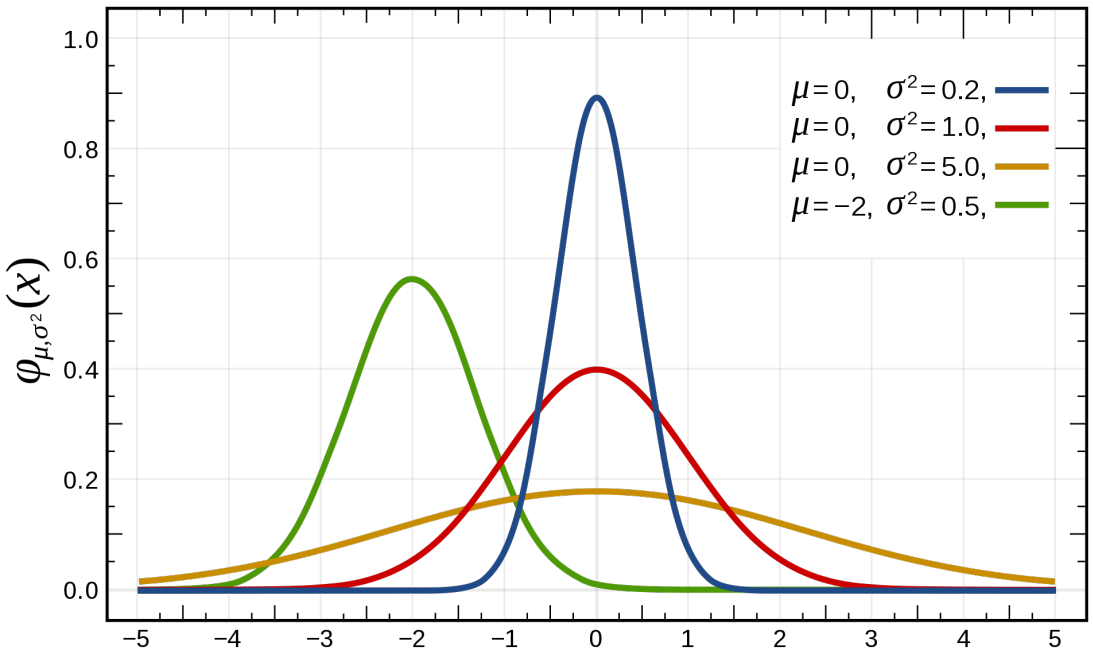


2. Gaussian model: $F(x, a, b, c) = ae^{-\left(\frac{x-b}{c}\right)^2}$, $a, b, c \in \mathbb{R}$ parameters.

   a is the value of the maximum obtained at $x = b$ and c determines the width of the curve.

   It is used in statistics to describe the normal distribution, but also in signal and image processing.

   In statistics $a = \frac{1}{\sigma\sqrt{2\pi}}$, $b = \mu$, $c = \sqrt{2\sigma}$, where μ, σ are the expected value and the standard deviation of a normally distributed random variable.

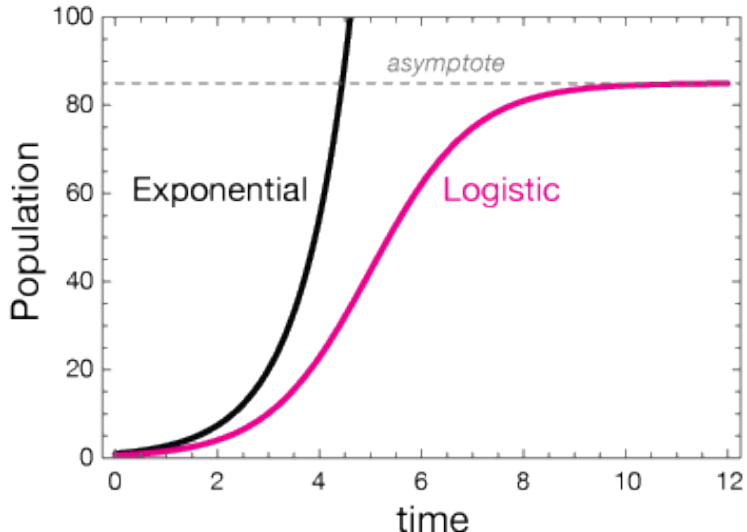


3. Logistic model: $F(x, a, b, k) = \frac{a}{(1+be^{-kx})}, k > 0$

The logistic function was devised as a model of population size by adjusting the exponential model which also considers the saturation of the environment, hence the growth first changes to linear and then stops.

The logistic function $F(x, a, b, k)$ is a solution of the first order non-linear differential equation

$$\frac{dy(x)}{dx} = ky(x)\left(1 - \frac{y(x)}{a}\right).$$



4. In the area around a radiotelescope the use of microwave ovens is forbidden, since the radiation interferes with the telescope. We are looking for the location (a, b) of a microwave oven that is causing problems.

   The radiation intensity decreases with the distance r from the source according to $u(r) = \frac{\alpha}{1+r}$. In cartesian coordinates:

$$u(x, y) = \frac{\alpha}{1 + \sqrt{(x-a)^2 + (y-b)^2}},$$

   where (a, b) is a position of the microwave.

   Task: Find the position of the microwave, if the measured values of the signal at three locations are $u(0, 0) = 0.27$, $u(1, 1) = 0.36$ in $u(0, 2) = 0.3$.

   This gives the following system of equations for the parameters α, a, b:

$$\begin{aligned} \frac{\alpha}{1 + \sqrt{a^2 + b^2}} &= 0.27 \\ \frac{\alpha}{1 + \sqrt{(1-a)^2 + (1-b)^2}} &= 0.36 \\ \frac{\alpha}{1 + \sqrt{a^2 + (2-b)^2}} &= 0.3 \end{aligned}$$

- Our goal is to fit the data points

$$\{(x_1, y_1), \ldots, (x_m, y_m)\}, \quad x_i \in \mathbb{R}^n, \ y_i \in \mathbb{R}.$$

- We choose a fitting function

$$F(x, a_1, \ldots, a_p)$$

which depends on the unknown parameters $a_1, \ldots, a_p$.

- Equivalent formulation of the system (13) (which will be more suitable for solving with numerical algorithms) is:
  1. For $i = 1, \ldots, m$ define the functions

$$g_i : \mathbb{R}^p \to \mathbb{R} \qquad \text{by the rule} \qquad g_i(a_1, \ldots, a_p) = y_i - F(x_i, a_1, \ldots, a_p).$$

  2. Solve or approximate the following system by the least squares method

$$\begin{aligned} g_1(a_1, \ldots, a_p) &= 0, \\ &\vdots \\ g_m(a_1, \ldots, a_p) &= 0. \end{aligned} \tag{14}$$

In a compact way (14) can be expressed by introducing a vector function

$$G : \mathbb{R}^p \to \mathbb{R}^m, \quad G(a_1, \ldots, a_p) = (g_1(a_1, \ldots, a_p), \ldots, g_m(a_1 \ldots, a_p)), \tag{15}$$

and search for the tuples $(a_1, \ldots, a_p)$ that solve the system (or minimize the norm of the left-hand side)

$$G(a_1, \ldots, a_p) = (0, \ldots, 0). \tag{16}$$

Remark

Solving (16) *is a difficult problem. Even if the exact solution exists, it is not easy (or even impossible) to compute. For example, there does not even exist an analytic formula to determine roots of a general polynomial of degree 5 or more.*

But we will learn some numerical algortihms to *approximate* the solutions of (16).

3.1 Vector functions of a vector variable

Neccessary terminology to achieve our plan

G from (15) is an example of

- a vector function: since it maps into $\mathbb{R}^m$, where m might be bigger than 1.
- a vector variable: since it maps from $\mathbb{R}^p$, where p might be bigger than 1.

Remark

- *If $m = 1$ and $p > 1$, then G is a usual multivariate function.*
- *If $m = 1$ and $p = 1$, then G is a usual (univariate) function.*

For easier reference in the continuation we call $g_1, \ldots, g_m$ from (15) the component (or coordinate) functions of G.

Examples

1. A linear vector function $G : \mathbb{R}^n \to \mathbb{R}^m$ is such that all the component functions g_i are linear:

$$g_i(x_1, \ldots, x_n) = a_{i1} \cdot x_1 + a_{i2} \cdot x_2 + \ldots + a_{in} \cdot x_n, \quad \text{where } a_{ij} \in \mathbb{R}. \quad (17)$$

In this case

$$G(x) = Ax,$$

where

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{bmatrix}.$$

2. Adding constants $b_i \in \mathbb{R}$ to the left side of (17) we get the definition of an affine linear vector function,

$$g_i(x_1, \ldots, x_n) = a_{i1}x_1 + a_{i2}x_2 + \ldots a_{in}x_n + b_i,$$

and then

$$G(x) = Ax + b, \qquad \text{where} \quad b = \begin{bmatrix} b_1 & b_2 & \ldots & b_n \end{bmatrix}^T.$$

3. Most of the (vector) functions are nonlinear, e.g.,

$$
\begin{aligned}
f &: \mathbb{R}^3 \to \mathbb{R}^2, \quad f(x,y,z) = (x^2+y^2+z^2-1, x+y+z),\\
g &: \mathbb{R}^2 \to \mathbb{R}^3, \quad g(z,w) = (zw, \cos z + w^2 - 2, e^{2z}),\\
h &: \mathbb{R} \to \mathbb{R}^2, \quad h(t) = (t+3, e^{-3t}).
\end{aligned}
$$

Derivative of a vector function - is needed in the algorithms we will use

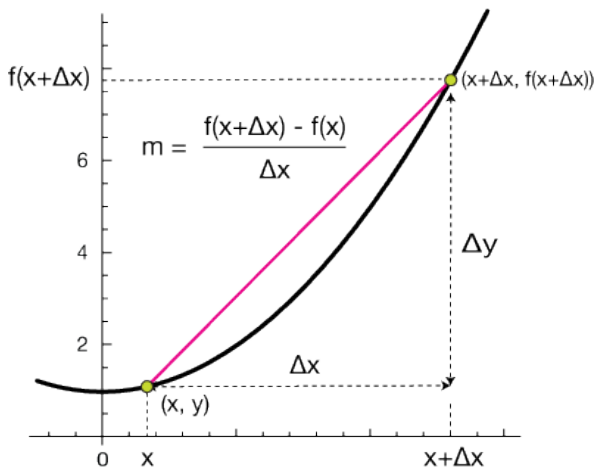
The derivative of a vector function $F : \mathbb{R}^n \to \mathbb{R}^m$ in the point

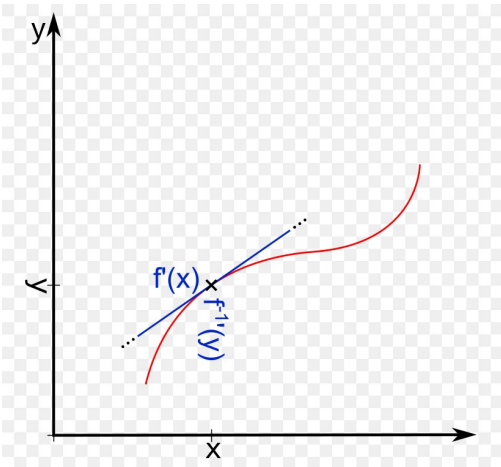
$$a := (a_1, \ldots, a_n) \in \mathbb{R}^n$$

is called the Jacobian matrix of F in a:

$$J_F(a) = DF(a) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(a) & \cdots & \frac{\partial f_1}{\partial x_n}(a) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(a) & \cdots & \frac{\partial f_m}{\partial x_n}(a) \end{bmatrix}.$$

- If $n = m = 1$, the $Df(x) = f'(x)$ is the usual derivative.

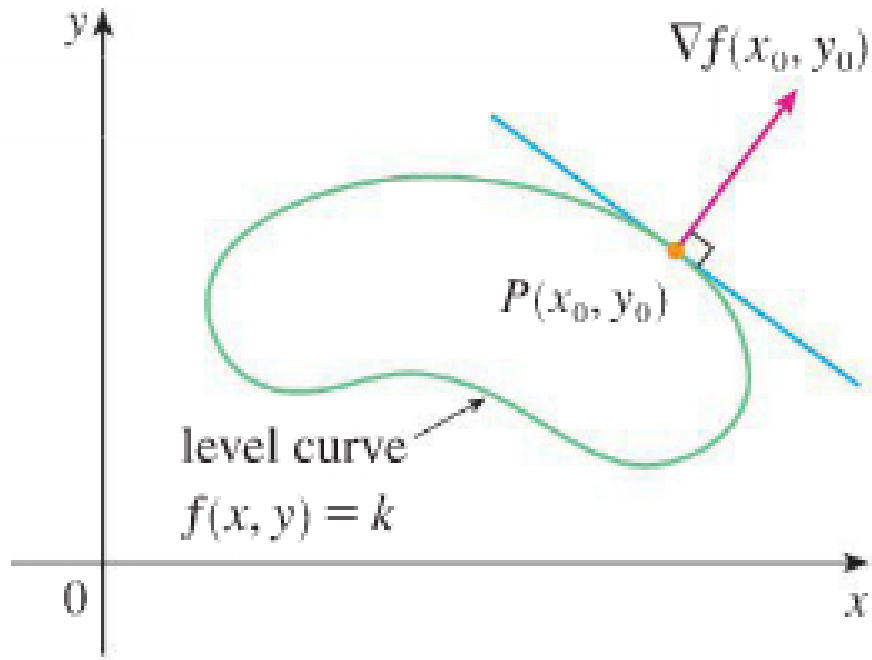




- For general n and $m = 1$, f is a function of n variables and

$$Df(x) = \text{grad } f(x)$$

is its gradient.



- For general m and n, $Df(x) = \begin{bmatrix} \text{grad } f_1 \\ \vdots \\ \text{grad } f_m \end{bmatrix}$ is a vector of gradients of component functions.

1. For an affine linear function $f : \mathbb{R}^n \to \mathbb{R}^m$, given by $f(x) = Ax + b$, it is easy to check that
$$Df(x) = A.$$
2. For a vector function $f : \mathbb{R}^3 \to \mathbb{R}^2$, given by
$$f(x, y, z) = (x^2 + y^2 + z^2 - 1, x + y + z),$$
then
$$Df(x) = \begin{bmatrix} 2x & 2y & 2z \\ 1 & 1 & 1 \end{bmatrix}.$$

A linear approximation of the vector function $f : \mathbb{R}^n \to \mathbb{R}^m$ at the point $a \in \mathbb{R}^n$ is the affine linear function

$$L_a : \mathbb{R}^n \to \mathbb{R}^m, \quad L_a(x) = Ax + b$$

that satisfies the following conditions:

1. It has the **same value** as f in a: $L_a(a) = f(a)$.
2. It has the **same derivative** as f at a: $DL_a(a) = Df(a)$.

It is easy to check that

$$L_a(x) = f(a) + Df(a)(x - a).$$

- $n = m = 1$:

$$L_a(x) = f(a) + f'(a)(x - a)$$

The graph $y = L_a(x)$ is the tangent to the graph $y = f(x)$ at the point a.

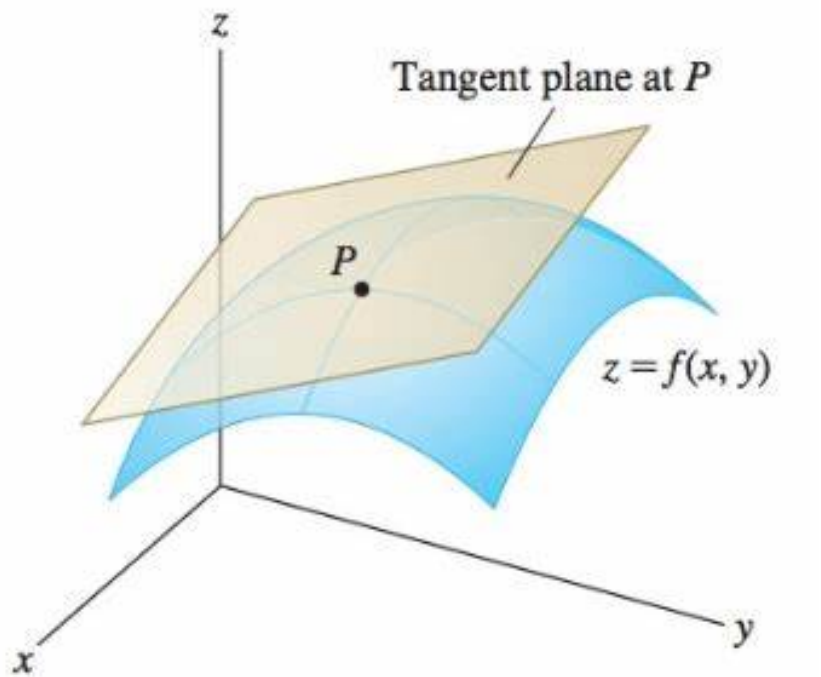
- If $n = 2$ and $m = 1$, then

$$L_{(a,b)}(x, y) = f(a, b) + \text{grad} f(a, b) \begin{bmatrix} x - a \\ y - b \end{bmatrix}.$$

The graph

$$z = L_{(a,b)}(x, y)$$

is the tangent plane to the surface $z = f(x, y)$ at the point (a, b).



Example

The linear approximation of the function

$$f : \mathbb{R}^3 \to \mathbb{R}^2, \qquad f(x, y, z) = (x^2 + y^2 + z^2 - 1, x + y + z)$$

at $a = (1, -1, 1)$ is the affine linear function

$$\begin{aligned} L_a(x, y, z) &= f(1, -1, 1) + Df(1, -1, 1) \begin{bmatrix} x - 1 \\ y + 1 \\ z - 1 \end{bmatrix} \\ &= \begin{bmatrix} 2 \\ 1 \end{bmatrix} + \begin{bmatrix} 2 & -2 & 2 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} x - 1 \\ y + 1 \\ z - 1 \end{bmatrix} \\ &= \begin{bmatrix} 2 + 2(x - 1) - 2(y + 1) + 2(z - 1) \\ 1 + (x - 1) + (y + 1) + (z - 2) \end{bmatrix} \\ &= \begin{bmatrix} 2 & -2 & 2 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} + \begin{bmatrix} -4 \\ 0 \end{bmatrix}. \end{aligned}$$

3.2 Solving systems of nonlinear equations

Let $f : D \to \mathbb{R}^m$ be a vector function, defined on some set $D \subset \mathbb{R}^n$.

We will study the Gauss-Newton method to solve the system $f(x) = 0$ in terms of least squares. This is one of the numerical methods for searching approximate solution of this system. It is based on linear approximations of f.

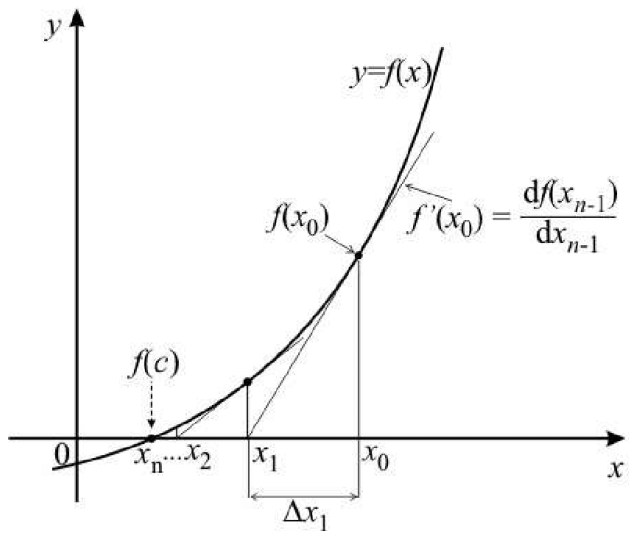
We are searching zeroes of the function $f : D \to \mathbb{R}$, $D \subseteq \mathbb{R}$, i.e., we are solving $f(x) = 0$.

Newton's or tangent method:

We construct a recursive sequence with:

- x_0 is an initial term,
- x_{k+1} is a solution of

$$L_{x_k}(x) = f(x_k) + f'(x_k)(x - x_k) = 0, \text{ so } x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}.$$



Theorem

The sequence x_i converges to a solution α, $f(\alpha) = 0$, if:

(1) $0 \neq |f'(x)|$ *for all $x \in I$, where I is some interval containing α,*

(2) x_0 *is sufficiently close to α.*

Under these assumptions the convergence is quadratic, meaning that:

$$\text{If we denote by } \varepsilon_j = |x_j - \alpha|, \text{ then } \varepsilon_{i+1} \leq M\varepsilon_i^2,$$

where M is some constant. If f is twice differentiable, then

$$M \leq \max_{x \in I} |f''(x)| / \min_{x \in I} |f'(x)|.$$

Proof.

Condition (1) implies in particular that α is a simple zero of f. Plugging α in the Taylor expansion of f around x_i we get

$$\begin{aligned} 0 = f(\alpha) &= f(x_i) + f'(x_i)(\alpha - x_i) + \frac{f''(\eta)}{2}(\alpha - x_i)^2 \\ &= f(x_i) + f'(x_i)(\alpha - x_i) + \frac{f''(\eta)}{2}(\alpha - x_i)^2 \end{aligned} \tag{18}$$

where η is between α and x_i. Dividing (18) with $f'(x_i)$ we get

$$0 = \frac{f(x_i)}{f'(x_i)} - (\alpha - x_i) + \frac{f''(\eta)}{2f'(x_i)} e_i^2$$

and hence

$$\left(x_i - \frac{f(x_i)}{f'(x_i)} \right) - \alpha = x_{i+1} - \alpha = \frac{f''(\eta)}{2f'(x_i)} e_i^2.$$

Thus,

$$e_{i+1} = \left| \frac{f''(\eta)}{2f'(x_i)} \right| e_i^2$$

Now

$$\left|\frac{f''(\eta)}{2f'(x_i)}\right| \leq \frac{\max_{x\in I}|f''(x)|}{\min_{x\in I}|f'(x)|}.$$

To prove that the sequence converges note that there exists $\delta_0 > 0$ such that

$$M\delta_0 < \frac{1}{2}.$$

Hence, if $e_i \leq \delta_0$, then

$$e_{i+1} = \left|\frac{f''(\eta)}{2f'(x_i)}\right| e_i^2 = \frac{1}{2}e_i.$$

Therefore

$$\lim_{n\to\infty} e_n = \lim_{n\to\infty} \frac{1}{2^n} \cdot e_0 = 0.$$

□

Newton's method generalizes to systems of n nonlinear equations in n unknowns:

- x_0 – initial approximation,
- x_{k+1} – solution of

$$L_{x_k}(x) = f(x_k) + Df(x_k)(x - x_k) = 0,$$

so

$$x_{k+1} = x_k - Df(x_k)^{-1} f(x_k).$$

In practice inverses are difficult to calculate (require to many operations) and the linear system for $\Delta x_k = x_{k+1} - x_k$

$$Df(x_k)\Delta x_k = -f(x_k)$$

is solved at each step (using LU decomposition of $Df(x_k)$) and hence

$$x_{k+1} = x_k + \Delta x_k.$$

Example

Derive Newton's method for solving the system of quadratic equations:

$$x^2 + y^2 - 10x + y = 1,$$
$$x^2 - y^2 - x + 10y = 25.$$

We are searching for the zero of the vector function

$$F : \mathbb{R}^2 \to \mathbb{R}^2, \quad F(x, y) = (x^2 + y^2 - 10x + y - 1, x^2 - y^2 - x + 10y - 25).$$

The Jacobian of F in (x, y) is

$$DF(x, y) = \begin{bmatrix} 2x - 10 & 2x - 1 \\ 2y + 1 & -2y + 10 \end{bmatrix}.$$

Using Newton's metod we:

- Choose an initial term (x_0, y_0).
- Calculate $x_{r+1} = x_r + \Delta x_r$, where $DF(x_r, y_r)\Delta x_r = -F(x_r, y_r)^T$.

Newton optimization method:

We would like to find the extrema of the function $F : \mathbb{R}^n \to \mathbb{R}$.

Since the extrema are *critical (or stationary) points*, the candidates are zeroes of the gradient, i.e.,

$$G(x) := \text{grad } F(x) = \begin{bmatrix} F_{x_1}(x) & \cdots & F_{x_n}(x) \end{bmatrix} = 0. \tag{19}$$

(19) is a system of n equations for n variables, the Jacobian of the vector function G is the so called Hessian of F:

$$DG(x) = H(x) = \begin{bmatrix} F_{x_1x_1} & \cdots & F_{x_1x_n} \\ \vdots & \ddots & \vdots \\ F_{x_nx_1} & \cdots & F_{x_nx_n} \end{bmatrix}.$$

If the sequence of iterates

$$x_0, \quad x_{k+1} = x_k - H^{-1}(x_k)G(x_k)$$

converges, the limit is a critical point of F, i.e., a candidate for the minimum (or maximum).

Gradient descent

Optimization methods can also be used to ensure a sufficiently accurate starting approximation for the Newton-based techniques. (Like bisection does for a single one-variable equation.)

Finding solutions of the system $F(x) = 0$, where

$$F = [F_1, \ldots, F_n]^T : \mathbb{R}^n \to \mathbb{R}^n$$

is equivalent to finding global minima of

$$g(x) := \|F\|^2 = F_1(x)^2 + \ldots + F_n(x)^2 : \mathbb{R}^n \to \mathbb{R}.$$

We search for the local minima (which are not necessarily global minima!) of g as follows:

1. Choose x_0.
2. Determine the constant α in $x_r - \alpha \cdot \text{grad}(g(x_r))$ which mimimizes

$$h(\alpha) = g(x_r - \alpha \cdot \text{grad}(g(x_r))).$$

   (Or is significantly smaller than $h(0) = g(x_r)$.)
3. $x_{r+1} = x_r - \alpha \cdot \text{grad}(g(x_r))$.

Quasi-Newtonov methods: Broyden's method

- For large n, the Newton's method is very expensive, since we need to evaluate n^2 partial derivatives at each step and use $\mathcal{O}(n^3)$ flops $(+, -, \cdot, :)$ to solve the linear system.
- Broyden's method avoids computing derivatives. For $n = m = 1$ it replaces the tangent by a secant throught the last two iterates. It mimicks this idea also for larger $n = m$.

Let B_r be an approximate for $J_f(x_r)$. Broyden's method works as follows:

1. Solve $B_r \Delta x_r = -f(x_r)$,
2. $x_{r+1} = x_r + \Delta x_r$,
3. Determine B_{r+1}.

The last step searches for a matrix B_{r+1}, which fulfils the secant condition:

$$B_{r+1}(x_{r+1} - x_r) = f(x_{r+1}) - f(x_r)$$

and is the closest to B_r in the spectral norm $\|\cdot\|_2$.

It turns out that

$$B_{r+1} = B_r + \frac{f(x_{r+1})(\Delta x_r)^T}{\|\Delta x_r\|_2^2}.$$

Recall from above the microwave oven example. The system of equations for the parameters α, a, b is:

$$\begin{aligned}
\frac{\alpha}{1+\sqrt{a^2+b^2}} - 0.27 &= 0 \\
\frac{\alpha}{1+\sqrt{(1-a)^2+(1-b)^2}} - 0.36 &= 0 \\
\frac{\alpha}{1+\sqrt{a^2+(2-b)^2}} - 0.3 &= 0.
\end{aligned}$$

https://zalara.github.io/Algoritmi/newtonsys.m
https://zalara.github.io/Algoritmi/broyden.m
https://zalara.github.io/Algoritmi/gradient_descent.m
https://zalara.github.io/Algoritmi/test_newtonsys_2.m

We have an overdetermined system

$$f : \mathbb{R}^n \to \mathbb{R}^m, \quad f(x) = (0, \ldots, 0) \tag{20}$$

of m nonlinear equations for n unknowns, where $m > n$.

The system (20) generally does not have a solution, so we are looking for a solution of (20) by the least squares method, i.e., $\alpha \in \mathbb{R}^n$ such that the distance of $f(\alpha)$ from the origin is the smallest possible:

$$\|f(\alpha)\|^2 = \min\{\|f(x)\|^2\}.$$

The Gauss-Newton method is a generalization of the Newton's method, where instead of the inverse of the Jacobian its MP inverse is used at each step:

$$x_0 \ldots \text{ initial approximation}, \quad x_{k+1} = x_k - Df(x_k)^+ f(x_k),$$

where $Df(x_k)^+$ is the MP inverse of $Df(x_k)$. If the matrix

$(Df(x_k)^T Df(x_k))$ is nonsingular at each step k, then

$$x_{k+1} = x_k - (Df(x_k)^T Df(x_k))^{-1} Df(x_k)^T f(x_k).$$

At each step x_{k+1} is the least squares approximation to the solution of the overdetermined linear system $L_{x_k}(x) = 0$, that is,

$$\|L_{x_k}(x_{k+1})\|^2 = \min\{\|L_{x_k}(x)\|^2, x \in \mathbb{R}^n\}.$$

Convergence is not guaranteed, but:

- if the sequence x_k converges, the limit $x = \lim_k x_k$ is a local (but not necessarily global) minimum of $\|f(x)\|^2$.

It follows that the Gauss-Newton method is an algorithm for the local minimum of $\|f(x)\|^2$.

Example

We are given point $(x_i, y_i) \in \mathbb{R}^2$ for $i = 1, \ldots, m$ and are searching for the function

$$f(x, a, b) = ae^{bx}$$

which fits this data best by the method of least squares.

So we have the overdetermined system $F(a, b) = 0$, where

$$F : \mathbb{R}^2 \to \mathbb{R}^m, \quad F(a, b) = (y_1 - ae^{bx_1}, \ldots, y_m - ae^{bx_m}).$$

The Jacobian of F is

$$DF(a, b) = \begin{bmatrix} -e^{bx_1} & ax_1 e^{bx_1} \\ \vdots & \\ -e^{bx_m} & ax_m e^{bx_m} \end{bmatrix}.$$

Using the Gauss-Newton method:

- We choose initianl approximation (a_0, b_0),
- Calculate iterates

$$\begin{bmatrix} a_{r+1} \\ b_{r+1} \end{bmatrix} = \begin{bmatrix} a_r \\ b_r \end{bmatrix} - DF(a_r, b_r)^+ F(a_r, b_r)^T.$$

Chapter 4:

Curves and surfaces

- Curves
  - Definition and examples
  - Derivative
  - Arc length and the natural parametrization
  - Curvature
  - Plotting plane curves
  - Area bounded by plane curves
  - Curves in the polar form
  - Motion in $\mathbb{R}^3$
- Surfaces
  - Definition and examples
  - Cartesian, cylindrical and spherical coordinates
  - Surface of revolution
  - Tangent plane

Curves - definition and examples

A parametric curve (or parametrized curve) in $\mathbb{R}^m$ is a vector function

$$f : I \to \mathbb{R}^m, \qquad f(t) = \begin{bmatrix} f_1(t) \\ \vdots \\ f_m(t) \end{bmatrix},$$

where $I \subset \mathbb{R}$ is a bounded or unbounded interval.

The independent variable (in this case t) is the parameter of the curve.

For every value $t \in I$, $f(t)$ represents a point in $\mathbb{R}^m$.

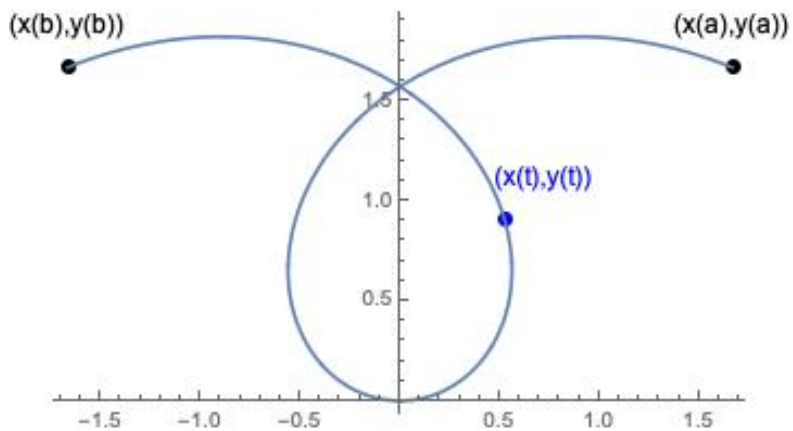
As t runs through I, $f(t)$ traces a path, or a curve, in $\mathbb{R}^m$.

If $m = 2$, then for every $t \in I$,

$$f(t) = \begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \mathbf{r}(t)$$

is the position vector of a point in the plane $\mathbb{R}^2$.

All points $\{f(t), t \in I\}$ form a plane curve:

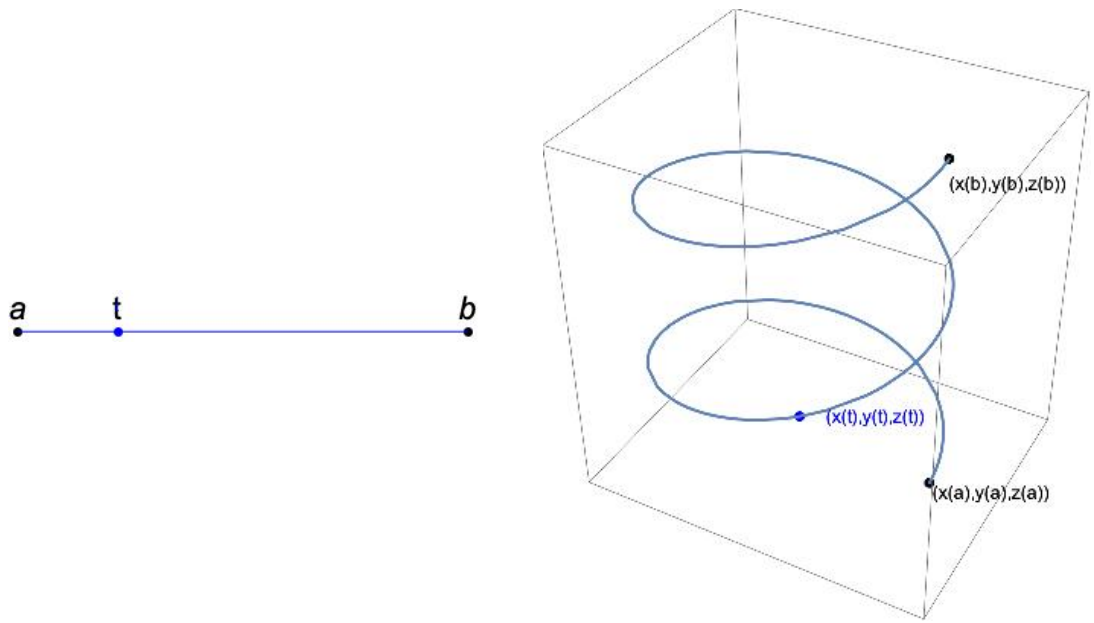


In this example $x(t) = t\cos t, y(t) = t\sin t, t \in [-3\pi/4, 3\pi/4]$

If $m = 3$, then

$$f(t) = \begin{bmatrix} x(t) \\ y(t) \\ z(t) \end{bmatrix} = \mathbf{r}(t)$$

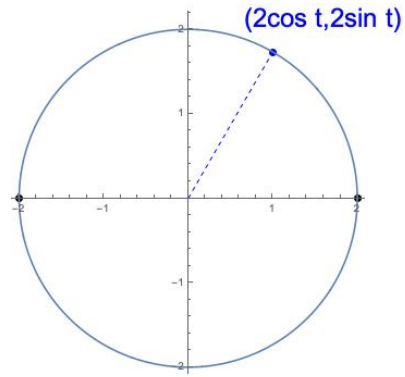
is the position vector of a point in $\mathbb{R}^3$ for every t, and $\{f(t), t \in I\}$ is a space curve:



In this example $x(t) = \cos t, y(t) = \sin t, z(t) = t/5, t \in [0, 4\pi]$

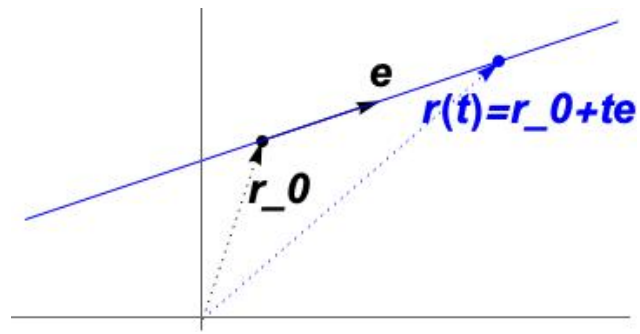
Example

$$f(t) = \begin{bmatrix} 2\cos t \\ 2\sin t \end{bmatrix}, t \in [0, 2\pi]$$



a circle with radius 2 and center $(0, 0)$

$f(t) = \mathbf{r}_0 + t\mathbf{e}, t \in \mathbb{R}$,
$\mathbf{r}_0, \mathbf{e} \in \mathbb{R}^m, \mathbf{e} \neq \mathbf{0}$

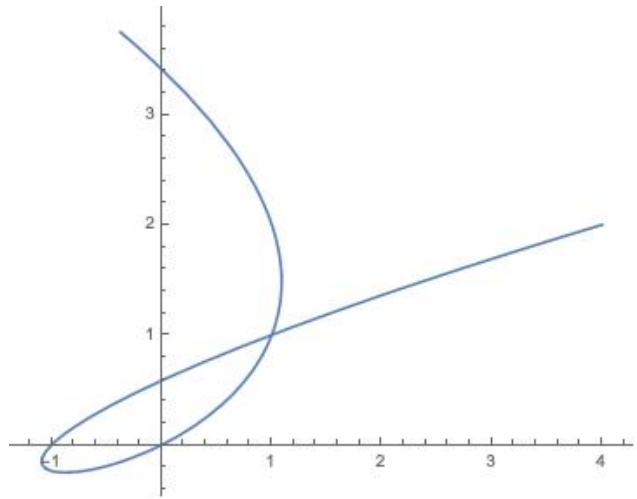


line through $\mathbf{r}_0$ in the direction of $\mathbf{e}$ in $\mathbb{R}^m$

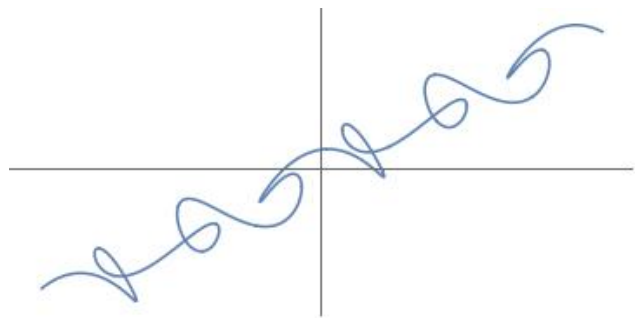
m=2:
slope $k = e_2/e_1$ if $e_1 \neq 0$
vertical if $\mathbf{e} = (0, e_2)$
horizontal if $\mathbf{e} = (e_1, 0)$

Example

$$f(t) = \begin{bmatrix} t^3 - 2t \\ t^2 - t \end{bmatrix}, t \in \mathbb{R}$$



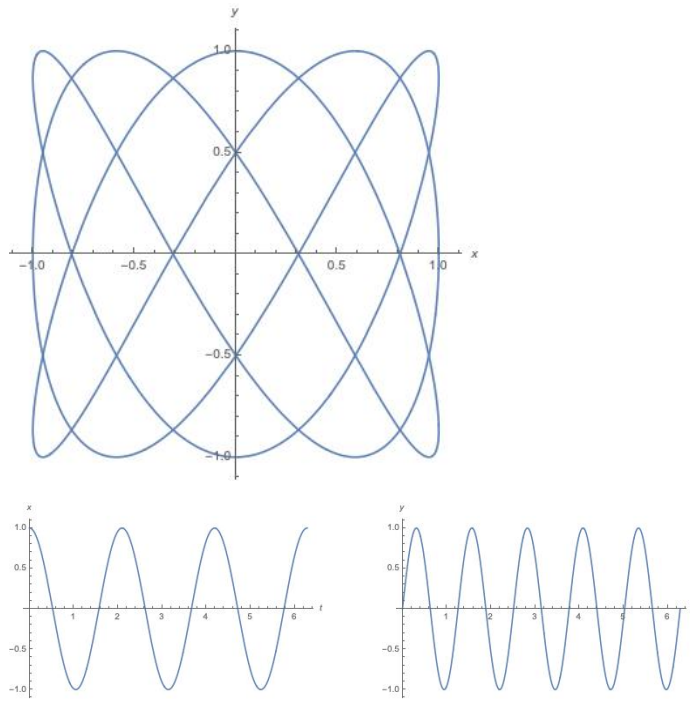
$$f(t) = \begin{bmatrix} t + \sin(3t) \\ t + \cos(5t) \end{bmatrix}, t \in \mathbb{R}$$



A parametric curve $f(t), t \in [a, b]$ is closed if $f(a) = f(b)$.

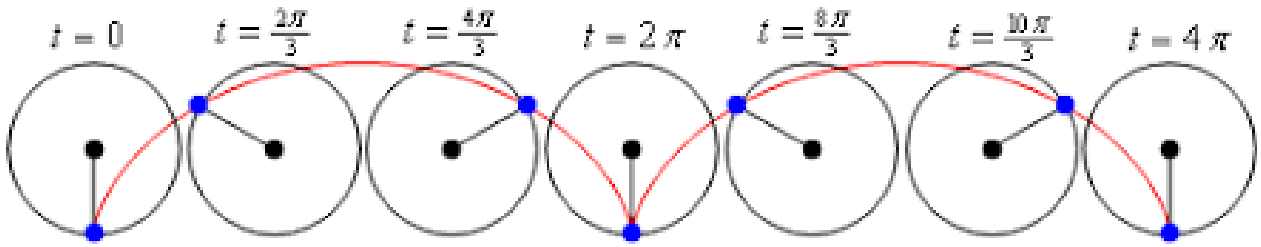
Example

$$f(t) = \begin{bmatrix} \cos 3t \\ \sin 5t \end{bmatrix}, t \in [0, 2\pi]$$

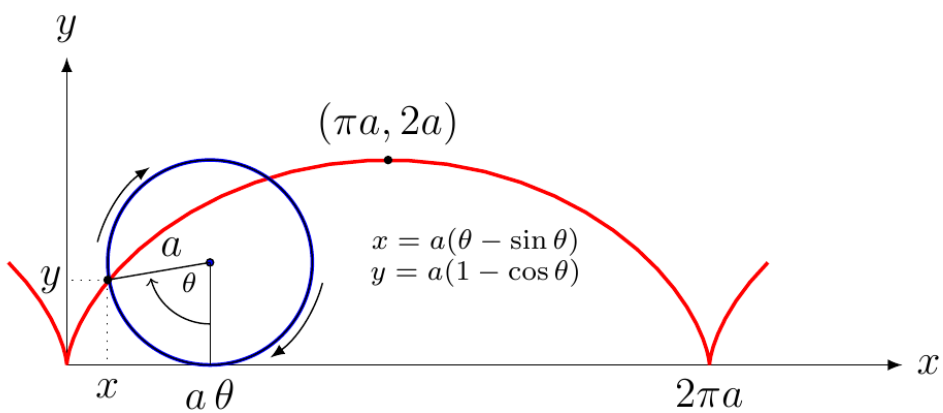


Problem: What path does the valve on your bicycle wheel trace as you bike along a straight road?

Represent the wheel as a circle of radius a rolling along the x-axis, the valve as a fixed point on the circle, the parameter is the angle of rotation:

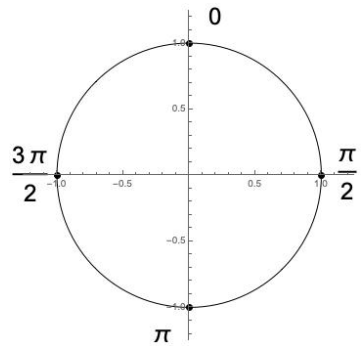


The curve is a cycloid: $x(\theta) = a\theta - a\sin\theta, y(\theta) = a - a\cos\theta$.

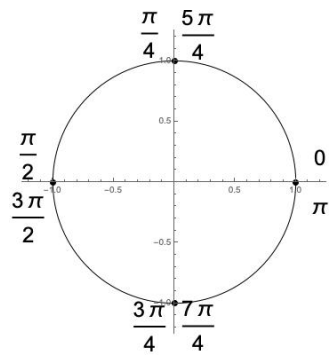


The following parametric curves all describe the circle with radius a around the origin (as well as many others):

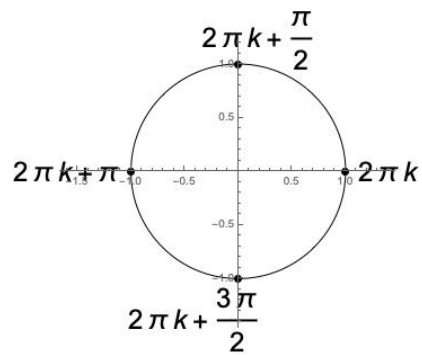
$$f_1(t) = \begin{bmatrix} a\sin t \\ a\cos t \end{bmatrix},\ t \in [0, 2\pi]$$



$$f_2(t) = \begin{bmatrix} a\cos 2t \\ a\sin 2t \end{bmatrix},\ t \in [0, 2\pi]$$



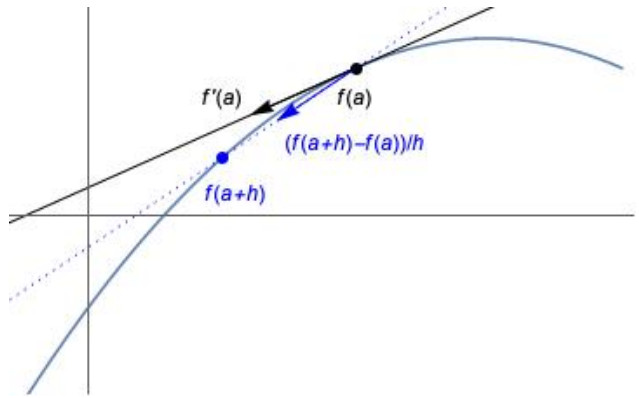
$$f_3(t) = \begin{bmatrix} a\cos t \\ a\sin t \end{bmatrix},\ t \in \mathbb{R}$$



Derivative, linear approximation, tangent

The derivative of the vector function $f(t) = \begin{bmatrix} x_1(t) \\ \vdots \\ x_m(t) \end{bmatrix}$ at the point a is the vector:

$$Df(a) = \begin{bmatrix} x_1'(a) \\ \vdots \\ x_m'(a) \end{bmatrix} = f'(a) = \lim_{h \to 0} \frac{1}{h}(f(a+h) - f(a))$$

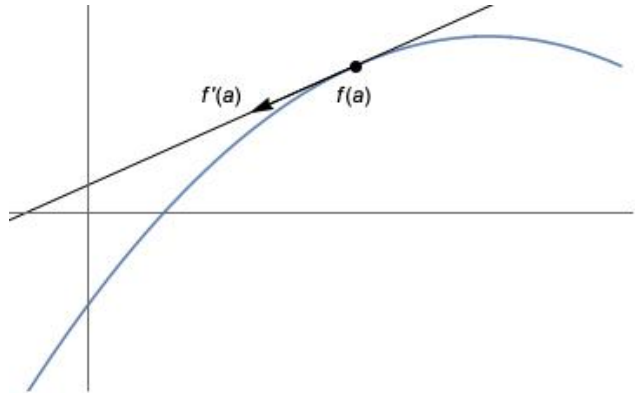


The vector $f'(a)$ (if it exists) represents the velocity vector of a point moving along the curve at the point $t = a$.

If $f'(a) \neq 0$ it points in the direction of the tangent at $t = a$.

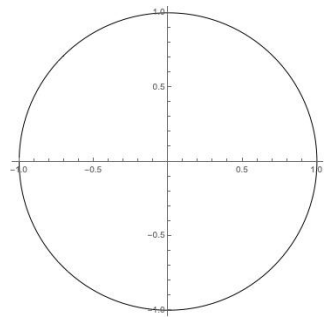
The linear approximation of the function f at $t = a$ is

$$L_a(t) = f(a) + (t - a)f'(a)$$



- If $f'(a) \neq \mathbf{0}$, this is a parametric line corresponding to the tangent line to the curve $f(t)$ at $t = a$. In this case $f(a)$ is a regular point of the parametrization.
- If $f'(a) = \mathbf{0}$ (or if it does not exist), the parametrization of the curve is singular in the point $f(a)$.
- A curve $C \in \mathbb{R}^m$ is smooth at a point P on C if there exists a parametrization $f(t)$ of C, such that $f(a) = P$ and $f'(a) \neq 0$.
- A smooth curve has a tangent at every point $P \in C$.

Problem: Is the curve $C = \{f(t), t \in [0, \sqrt{2\pi}]\}$, $f(t) = \begin{bmatrix} \cos(t^2) \\ \sin(t^2) \end{bmatrix}$, smooth?



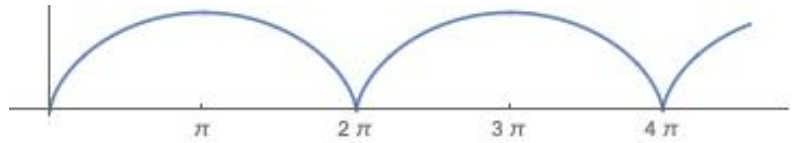
Since $x^2 + y^2 = 1$, $f(t)$ is a parametrization of the unit circle which is a smooth curve (it has a tangent at every point).

Since $f'(0) = \mathbf{0}$ the parametrization f is singular in the point $(1, 0)$.

However, a smooth parametrization exists. Can you find it?

Problem: Is the cycloid a smooth curve?

Our parametrization



$$f(t) = \begin{bmatrix} t - \sin t \\ 1 - \cos t \end{bmatrix}, \quad f'(t) = \begin{bmatrix} 1 - \cos t \\ \sin t \end{bmatrix}$$

is not smooth at $t = 2k\pi$ since $f'(2k\pi) = \mathbf{0}$.

Does a tangent exist?
The slope of the tangent line at a point $f(t)$ is:

$$k_t = \frac{y'(t)}{x'(t)} = \frac{\sin t}{1 - \cos t}$$

The left and right limits as $t \to 2k\pi$ are

$$\lim_{t \nearrow 2k\pi} k_t = \lim_{t \nearrow 2k\pi} \frac{\cos t}{\sin t} = -\infty, \quad lim_{t \searrow 2k\pi} k_t = \lim_{t \searrow 2k\pi} \frac{\cos t}{\sin t} = \infty,$$

so at these points the curve forms a sharp spike (a cusp) and a tangent does not exist.

So, the cycloid is not smooth at the points where it touches the x axis.

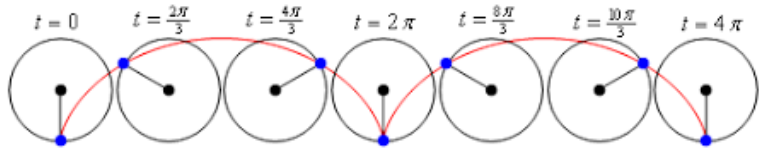
(l'Hospital's rule was used to compute the limits.)

Arc length and the natural parametrization

The arc length s of a parametric curve $f(t)$, $t \in [a, b]$, in $\mathbb{R}^m$ is the length of the curve between the points $t = a$ in $t = b$, i.e. the distance covered by a point moving along the curve between these two points.

Example

For example, what distance does a point on the circle cover when the circle makes one full turn?



Proposition

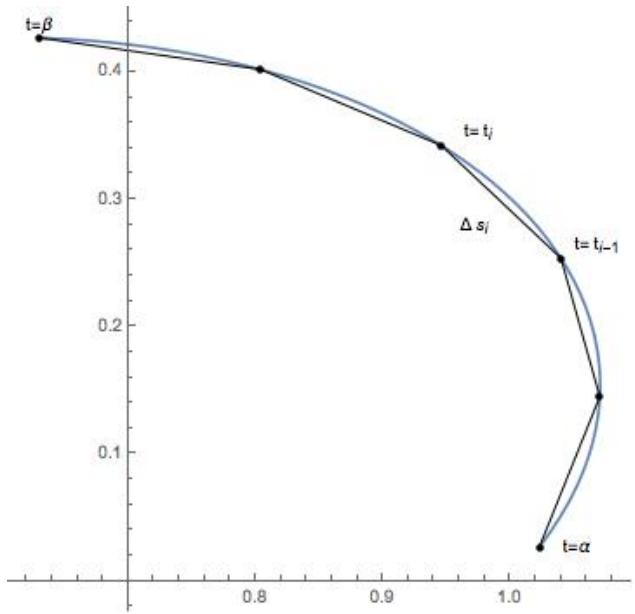
The arc length s of a parametric curve $f(t)$ between the points $t = a$ and $t = b$ is given by

$$s = \int_a^b \|f'(t)\|\, dt.$$

Proof of the Proposition

An aproximate value for s is the length of a polygonal curve connecting close enough points on the curve:

$$
\begin{aligned}
s_n &= \sum_{i=1}^{n} \|f(t_i) - f(t_{i-1})\| \\
&= \sum_{i=1}^{n} \|f'(t_{i-1})\| \Delta t \\
&\to_{n\to\infty} \int_a^b \|f'(t)\|\, dt
\end{aligned}
$$



where:

- The value $f(t_i) = f(t_{i-1} + \Delta t)$, where $\Delta t = t_i - t_{i-1}$, was approximated as $f(t_i) = f(t_{i-1}) + f'(t_{i-1})\Delta t$ and hence $f(t_i) = f(t_{i-1}) + f'(t_{i-1})\Delta t$. (Under the assumption that f' is continuous.
- In the last line we used that the sum represents a Riemannian sum of the function $\|f'(t)\|$.
- For n big enough, s_n is a practical approximation for s.

Problem: The length of the path traced by a point on the circle after a full turn?

A parametrization is $f(t) = \begin{bmatrix} t - \sin t \\ 1 - \cos t \end{bmatrix}$ and hence:

$$\begin{aligned} s &= \int_0^{2\pi} \sqrt{(1-\cos t)^2 + \sin^2 t}\, dt = \int_0^{2\pi} \sqrt{2 - 2\cos t}\, dt = \int_0^{2\pi} \sqrt{4\sin^2(t/2)}\, dt \\ &= \int_0^{2\pi} 2\sin(t/2)\, dt = -4(\cos(\pi) - \cos(0)) = 8. \end{aligned}$$

Problem: What is the arc length of the helix $f(t) = \begin{bmatrix} a\cos t \\ a\sin t \\ bt \end{bmatrix}$, $0 \le t \le 2\pi$?

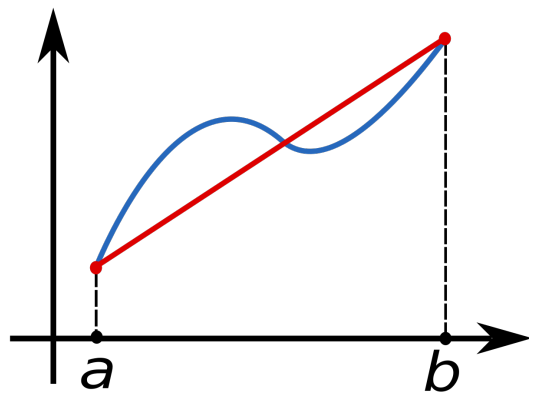
Problem: The circumference of the elipse $\begin{bmatrix} a\cos t \\ b\sin t \end{bmatrix}$, $a \neq b$?

$$\int_0^{2\pi} \sqrt{a^2\sin^2 t + b^2\cos^2 t}\, dt = 4a\int_0^{\pi/2} \sqrt{1 - e^2\sin^2 t}\, dt = 4a\mathrm{E}(e)$$

where $e = \sqrt{1-(b/a)^2}$ is its eccentricity and the function E is the nonelementary elliptic integral of 2nd kind. It can be computed numerically, which is briefly explained in the next few slides.

Numerical integration

The integral $\int_a^b f(x)\,dx$ can be approximated by a linear approximation of f over the interval $[a, b]$ and computing the area of the trapezoid formed.



$$\int_a^b f(x)\,dx \approx f(a) + \frac{f(b) - f(a)}{b - a}(x - a) =: T(b - a)$$

Of course the error of this approximation is usually large and we are not satisfied. How do we estimate how good is this approximation?

Adaptive trapezoid rule (*integral*$(\cdots)$ in Matlab)

1. $T(b-a) = f(a) + \frac{f(b)-f(a)}{b-a}(x-a)$.

2. We add another point in the middle of the interval, i.e., $x = \frac{a+b}{2}$ and compute the sum of the areas of two trapezoids formed:

$$T((b-a)/2) = \frac{1}{2}T(b-a) + \frac{b-a}{2} \cdot f((a+b)/2).$$

3. If $e := |T(b-a) - T((b-a)/2)|$ is smaller than the tolerance *tol*, we are satisfied and return $T((b-a)/2)$.

4. Otherwise we have to repeat the procedure on each of the subintervals $[a, (a+b)/2]$ and $[(a+b)/2, b]$, where the tolerance on each of them must be smaller than $tol/2$.

5. We can implement this recursively, obtaining the so called adaptive trapezoid rule, where on different subintervals of $[a, b]$ different number of recursions is needed (this depends on the behaviour of the function f).

Natural parametrization

The arc length from the initial $t = a$ to an arbitrary $t = T$

$$s(t) = \int_a^t \|f'(u)\| \, du$$

is an increasing function of t if f is a smooth parametrization, so it has an inverse

$$t(s) : [0, s(T)] \to [a, T].$$

So, the original parameter t can be expressed as a funcion of the arc length s.

Inserting this into the parametrization gives the same curve with a different parametrization:

$$g(s) = f(t(s)).$$

The arc length s is called the natural parameter of the curve.

Proposition

A curve C is parametrized with the natural parameter s satisfies

$$\|g'(s)\| = 1, \tag{21}$$

i.e., the length of the velocity vector is 1 at every point and so a parametrization with the natural parameter is the unit speed parametrization.

Proof. Indeed,

$$g'(s) = \frac{dg}{ds}(s) = \frac{d(f \circ t)}{ds}(s) = \frac{df}{dt}(t(s)) \cdot \frac{dt}{ds}(s) = f'(t(s))t'(s). \tag{22}$$

Now note that by the fundamental theorem of calculus we have that

$$s'(t) = \|f'(t)\|$$

and hence

$$t'(s) = \frac{1}{\|f'(t(s))\|}.$$

Plugging this into (22) we get

$$g'(s) = \frac{f'(t(s))}{\|f'(t(s))\|},$$

which is equivalent to (21).

Example

The standard parametrization of the circle

$$f(t) = \begin{bmatrix} a\cos t \\ a\sin t \end{bmatrix}$$

is not the natural parametrization if $a \neq 1$, since

$$\|f'(t)\| = \sqrt{a^2\cos^2 t + a^2\sin^2 t} = a \neq 1.$$

Since

$$s(t) = \int_0^t a\,dt = at,$$

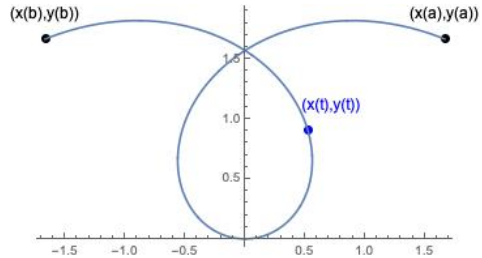
it follows that $t = s/a$ and the natural parametrization is

$$g(s) = \begin{bmatrix} a\cos(s/a) \\ a\sin(s/a) \end{bmatrix}.$$

Remember:

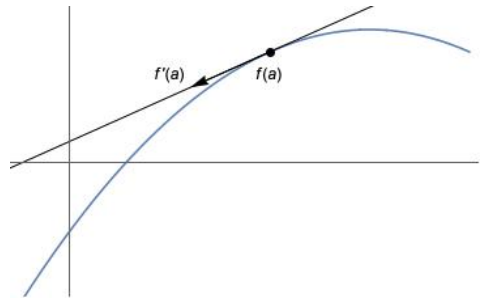
A parametric curve: $f(t) = \begin{bmatrix} x_1(t) \\ \vdots \\ x_m(t) \end{bmatrix}$,

$t \in I \subset \mathbb{R}$,



The derivative $f'(t) = \begin{bmatrix} x_1'(t) \\ \vdots \\ x_m'(t) \end{bmatrix}$

is the velocity vector or
tangent vector if $f'(t) \neq 0$,

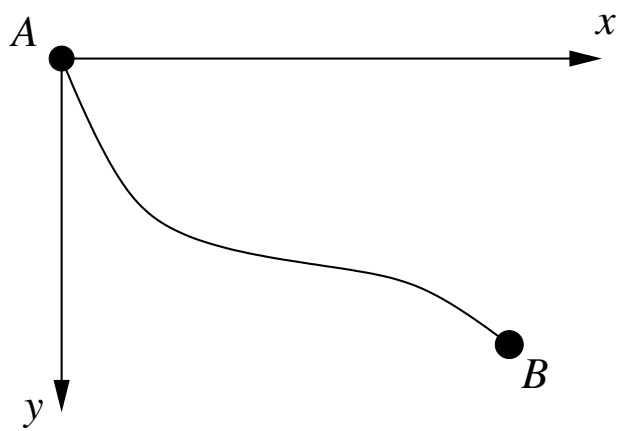


The image $C = \{f(t), t \in I\}$: a (geometric) curve in $\mathbb{R}^m$.A curve C has many parametrizations.

The arc length parametrization or natural parametrization $f(s)$:
s is the length of the chord from $f(a)$ to $f(s)$, $\|f'(s)\| = 1$.

Brachistochrone problem

Problem: Given points A and B what is the fastest path of a mass starting in A and ending in B, being accelerated only by gravity? We assume no friction is present.



- We denote $A = (0, 0)$ and $B = (b, 0)$. We are searching for a curve

$$y(x) : [0, b] \to \mathbb{R}.$$

- Law of conservation of energy:

$$\text{Potential energy} + \text{Kinetic energy} = \text{constant}$$

$$\frac{1}{2} m v(x)^2 = mgy(x) \quad \Rightarrow \quad v(x) = \sqrt{2gy(x)}.$$

- Let $s(x)$ be the arc length of the curve from A to $(x, y(x))$. We have:

$$s(x) = \int_0^x \sqrt{1 + y'(x)^2}\, dx$$

and hence

$$s'(x) = \frac{ds}{dx} = \sqrt{1 + y'(x)^2}.$$

- Let $T(y)$ be the travel time along the curve $\{(x, y(x)) : x \in [0, b]\}$. We have:

$$T(y) = \int_0^{T(y)} dt = \int_0^{s(b)} \frac{ds}{v(s)} = \int_0^b \frac{\sqrt{1 + y'(x)}}{\sqrt{2gy(x)}}\, dx.$$

- We need to minimize the functional $T(y) : C[0, b] \to \mathbb{R}$ on the vector space of continuous functions on $[0, b]$.

Theorem (Euler-Lagrange equation)

If y^* is the solution of the minimization problem $\min_{y \in C[0,b]} T(y)$, then it satisfies the equation

$$\frac{\partial}{\partial y} f(x, y(x), y'(x)) = \frac{d}{dx} \frac{\partial}{\partial y'} f(x, y(x), y'(x)).$$

Applying Euler-Lagrange equations for the brachistochrone problem, we come to the differential equation

$$y' = \sqrt{\frac{C - y}{y}} \quad \text{for some constant } C.$$

Separation of variables:

$$\sqrt{\frac{y}{C - y}}\, dy = dx.$$

Integrating both sides and using the substitution $y = C \sin^2(t)$ we get

$$x(t) = C\left(t - \frac{1}{2}\sin 2t\right),\ y(t) = C\left(\frac{1}{2} - \frac{1}{2}\cos 2t\right),$$

which is the cycloide.

For those who want to know more:

https://wiki.math.ntnu.no/_media/tma4180/2015v/calcvar.pdf

https://www.youtube.com/watch?v=Cld0p3a43fU

Curvature

1. Intuitively we would like to measure for what amount does the curve deviate from being the straight line.
2. For the circle of radius R we would like that the curvature is proportional to $1/R$.

The curvature $\kappa(t)$ of a smooth curve $f(t)$ at a point $t = a$ is the rate of change of the unit tangent vector $T(t) = \dfrac{f'(t)}{\|f'(t)\|}$:

$$\kappa(t) = \left\| \frac{1}{ds/dt} T'(t) \right\|.$$

If the curve is parametrized by the arc length s, i.e., $\|f'(s)\| = 1$, then this is simply

$$\kappa(s) = \|f''(s)\|$$

Problem: what is the curvature of a circle with radius a?

The natural parametrization of the circle is $f(s) = \begin{bmatrix} a\cos(s/a) \\ a\sin(s/a) \end{bmatrix}$, so

$$f'(s) = \begin{bmatrix} -\sin(s/a) \\ \cos(s/a) \end{bmatrix} \qquad \text{and} \qquad f''(s) = \begin{bmatrix} -\cos(s/a)/a \\ -\sin(s/a)/a \end{bmatrix}.$$

The curvature

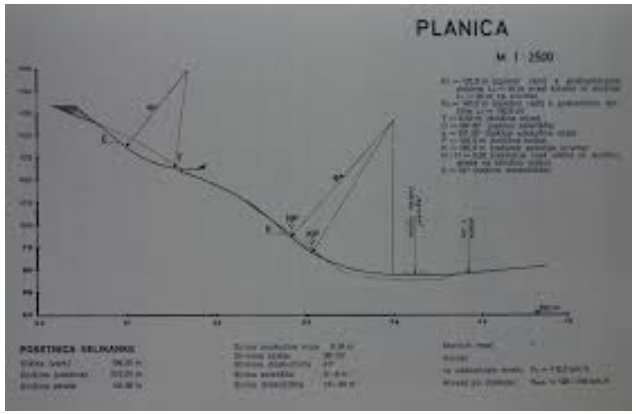
$$\kappa(s) = \|f''(s)\| = 1/a$$

is constant along the circle.

- As $a \to \infty$, the circle goes towards a line and $\kappa \to 0$.
- On the other hand, as $a \to 0$, the circle goes towards a point and $\kappa \to \infty$.

Problem: designing roads and railways

Roads, railway bends, roller coaster loops, the ski jump in Planica . . . are designed so that the transitions from the straight to the circular parts are as smooth as possible.

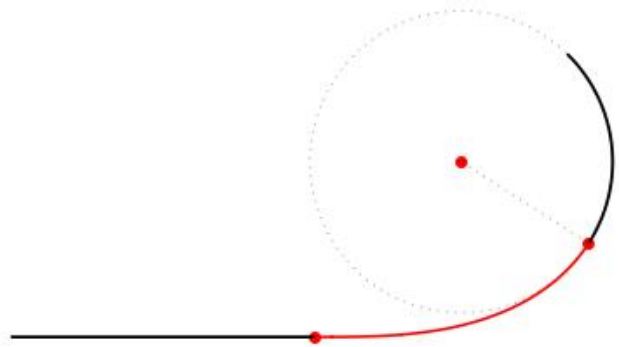


The force acting on a moving point on the curve (car, train, ski jumper,. . .) increases and decreases as evenly as possible.

The transition curve from

- the straight part (with curvature 0) to
- the circular part (with curvature $a > 0$)

has several names: clotoid, Euler spiral, Cornu spiral . . .



Its characteristic property is that the curvature $\kappa(s)$ is a linear function of arc length s.

Let us find its arc length parametrization $f(s)$. Assume that $\kappa(s) = \|f''(s)\| = 2s$.

Remember that the arc length parametrization is the unit speed parametrization, so $\|f'(s)\| = 1$ and so $f'(s)$ can be written in the form

$$f'(s) = \begin{bmatrix} x'(s) \\ y'(s) \end{bmatrix} = \begin{bmatrix} \cos\varphi(s) \\ \sin\varphi(s) \end{bmatrix}.$$

This gives

$$\kappa(s) = \sqrt{x''(s)^2 + y''(s)^2} = \varphi'(s) = 2s, \quad \varphi(s) = s^2,$$

$$x'(s) = \cos(s^2), \quad y'(s) = \sin(s^2),$$

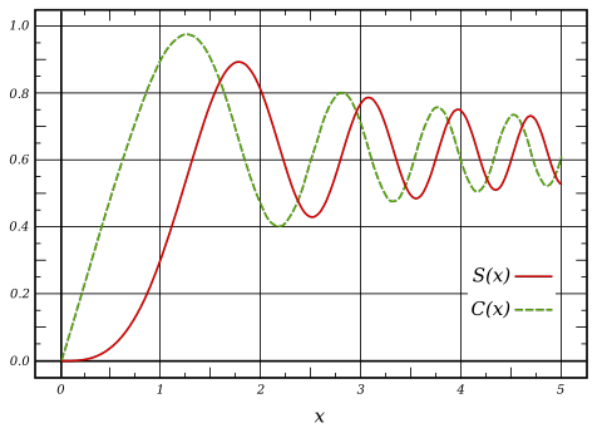
so

$$x(s) = \int_0^s \cos(u^2)\,du, \quad y(s) = \int_0^s \sin(u^2)\,du$$

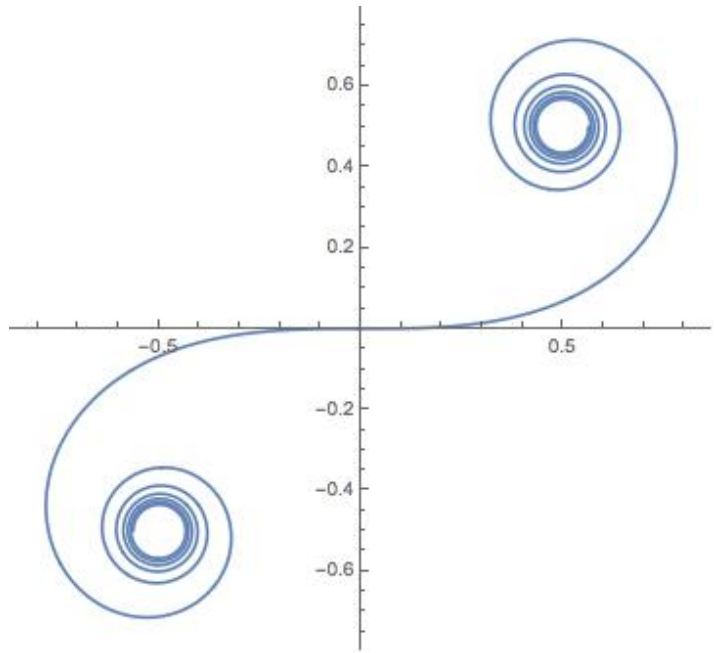
The functions

$$x(s) = \int_0^s \cos(u^2)\,du = C(s), \quad y(s) = \int_0^t \sin(u^2)\,du = S(s)$$

are nonelementary functions called the Fresnel integrals



Fresnel integrals



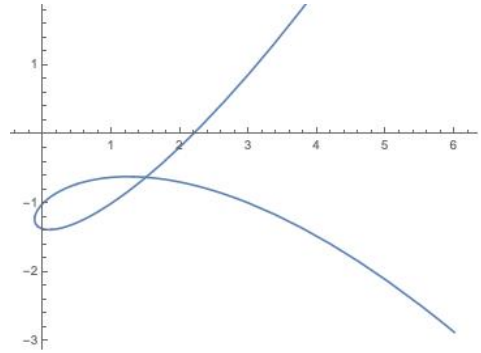
clotoid

Plane curves

For a plane curve $f(t) = \begin{bmatrix} x(t) \\ y(t) \end{bmatrix}$ the tangent at a regular point $f(a)$ is

- the vertical line
$$x = x(a)$$
if $x'(a) = 0$ and $y'(a) \neq 0$,
- the line
$$y - y(a) = \frac{y'(a)}{x'(a)}(x - x(a))$$
if $x'(a) \neq 0$,
- the horizontal line
$$y = y(a)$$
if $y'(a) = 0$ and $x'(a) \neq 0$.

Plotting a parametric plane curve

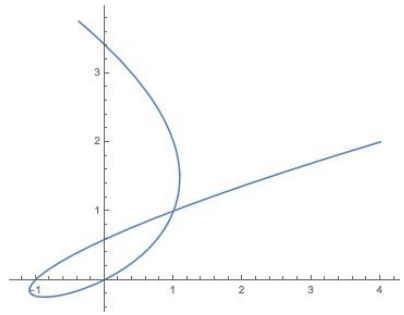


Here is a general strategy:

- find the asymptotic behaviour: $\lim_{t\to\infty} f(t)$, $\lim_{t\to-\infty} f(t)$
- find intersections with coordinate axes: solve $y(t) = 0$ and $x(t) = 0$
- find points where the tangent is vertical or horizontal: solve $x'(t) = 0$ and $y'(t) = 0$
- find self-intersections: solve $f(t) = f(s),\ t \neq s$
  - and the two tangents there
- look for other helpful features ...
- connect points $\mathbf{r}(t) = f(t)$ by increasing t

Problem: find the self-intersection (if there is one) of a parametric curve

Let $f(t) = \begin{bmatrix} t^3 - 2t \\ t^2 - t \end{bmatrix}$



A self-intersection is at a point $f(t) = f(s)$, with $t \neq s$, so:

$$t^3 - 2t = s^3 - 2s \quad \text{and} \quad t^2 - t = s^2 - s$$
$$\Rightarrow \quad t^3 - s^3 = 2t - 2s \quad \text{and} \quad t^2 - s^2 = t - s$$

Since $t \neq s$ we can divide by $t - s$:

$$t^2 + ts + s^2 = 2 \quad \text{and} \quad t + s = 1$$
$$\Rightarrow \quad t = 1 - s \quad \text{and} \quad (1-s)^2 + s(1-s) + s^2 = 2.$$

The self-intersection (where s and t can be interchanged) is at

$$s = (1+\sqrt{5})/2,\ t = (1-\sqrt{5})/2, \quad f(t) = f(s) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Problem: do two parametric curves intersect. Imagine two cars speeding along the two curves. Do they crash?

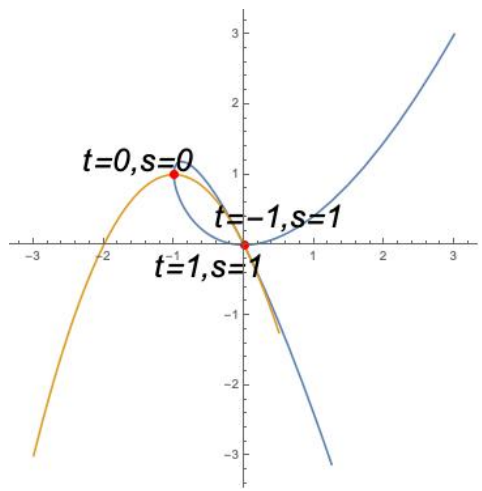
Let $f_1(t) = \begin{bmatrix} t^2 - 1 \\ -t^3 - t^2 + t + 1 \end{bmatrix}$, $f_2(s) = \begin{bmatrix} s - 1 \\ 1 - s^2 \end{bmatrix}$.

To find the intersections, solve the system

$$t^2 - 1 = s - 1 \quad \text{and} \quad -t^3 - t^2 + t + 1 = 1 - s^2$$
$$\Rightarrow \quad s = t^2 \quad \text{and} \quad -s^6 - s^4 + s^2 + 1 = 1 - s^2$$

There are three solutions:

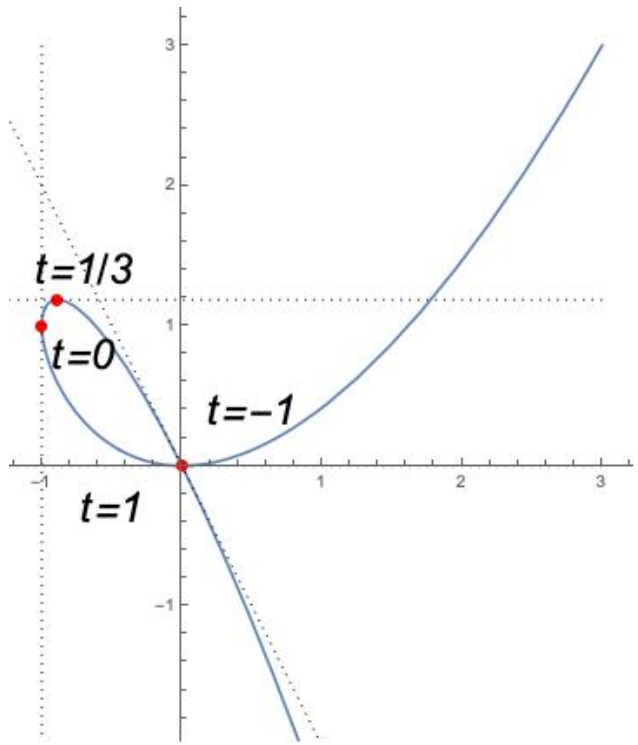
$$t = -1, s = 1 \quad \Rightarrow \quad x = 0, y = 0$$
$$t = 0,\ s = 0 \quad \Rightarrow \quad x = -1,\ y = 1$$
$$t = 1,\ s = 1 \quad \Rightarrow \quad x = 0,\ y = 0$$



The cars meet at $t = 0, s = 0$ at the point $(-1, 1)$ and at $t = 1, s = 1$ at the point $(0, 0)$.

Problem: plot $f(t) = \begin{bmatrix} t^2 - 1 \\ -t^3 - t^2 + t + 1 \end{bmatrix}$, $f'(t) = \begin{bmatrix} 2t \\ -3t^2 - 2t + 1 \end{bmatrix}$

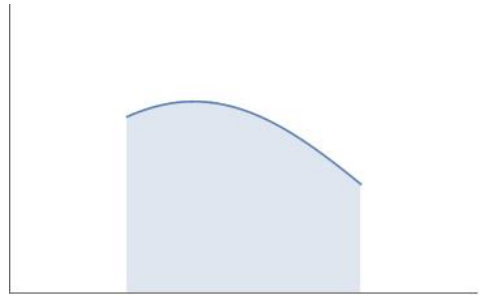
- Asymptotic behaviour: $\lim_{t\to\infty} f(t) = \begin{bmatrix} \infty \\ -\infty \end{bmatrix}$, $\lim_{t\to-\infty} f(t) = \begin{bmatrix} \infty \\ \infty \end{bmatrix}$,
  - intersections with axes: $t = \pm 1$, at $(0, 0)$
    this is also a self-intersection
  - the two tangent lines at $(0, 0)$
    - at $t = -1$: $y = 0$,
    - at $t = 1$: $y = -2x$
  - vertical tangent: $t = 0$ at $(-1, 1)$
  - horizontal tangent
    - at $t_1 = -1$, $y = 0$,
    - at $t_2 = 1/3$, $y = 32/27$



Areas bounded by plane curve

I. Let $f(t) = \begin{bmatrix} x(t) \\ y(t) \end{bmatrix}, \quad t \in [a, b]$

$x'(t) > 0$



The area of the quadrilateral bounded by the curve and the x-axis is

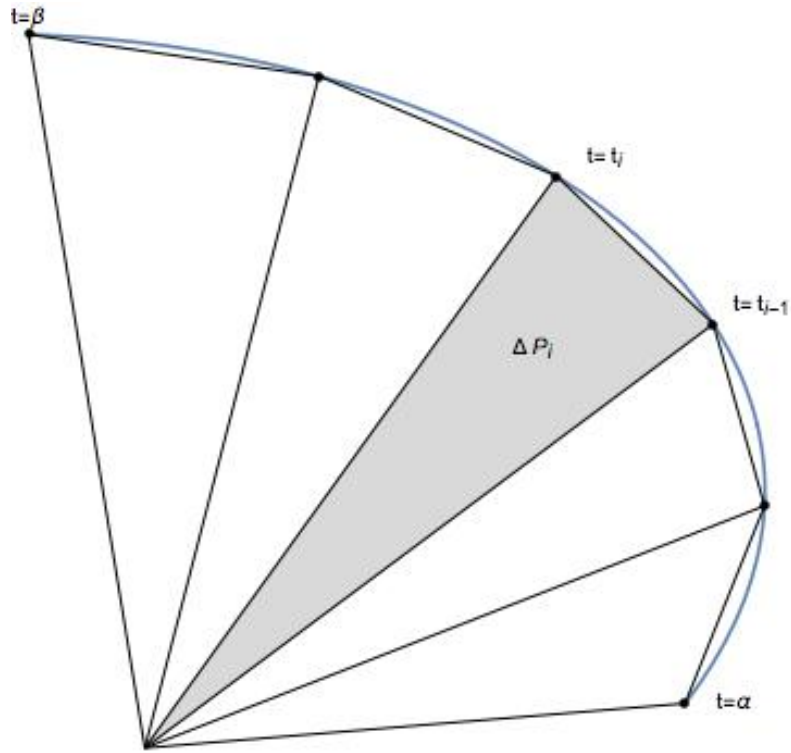
$$P = \int_{x(a)}^{x(b)} |y(x)|\, dx = \int_a^b |y(t)| x'(t)\, dt$$

Problem: the area under one arc of the cycloid:

$$x(t) = at - a\sin t, \quad y(t) = a - a\cos t,$$

$$P = \int_0^{2\pi} a^2(1-\cos t)^2\, dt = a^2 \int_0^{2\pi} \left(\frac{3}{2} - 2\cos t + \frac{1}{2}\cos(2t)\right) dt = 3a^2\pi.$$

II. The area of the triangular region bounded by the curve $f(t), t \in [a, b]$, and the two end-point position vectors $f(a)$ and $f(b)$:



$$P = \frac{1}{2} \int_a^b |x(t)y'(t) - y(t)x'(t)| \, dt.$$

Proof of the area formula

An approximate value of the area is the sum of areas of triangles obtained by subdividing the interval $[a, b]$ into n intervals of length $\Delta t = (b-a)/n$.

The area of a triangle with vertices $(0,0)$, $f(t_i)$, $f(t_{i+1})$ is

$$\Delta P_i = \frac{1}{2}\|f(t_{i+1}) \times f(t_i)\| \doteq \frac{1}{2}\|(f(t_i) + f'(t_i)\Delta t) \times f(t_i)\|$$

$$= \frac{1}{2}\|f'(t_i) \times f(t_i)\|\Delta t = \frac{1}{2}|y'(t_i)x(t_i) - x'(t_i)y(t_i)|\Delta t,$$

where the last equlatiy follows from the calculation

$$\begin{aligned} f'(t_i) \times f(t_i) &= (x'(t_i), y'(t_i), 0) \times (x(t_i), y(t_i), 0) \\ &= (x'(t_i)y(t_i) - y'(t_i)x(t_i), 0, 0). \end{aligned}$$

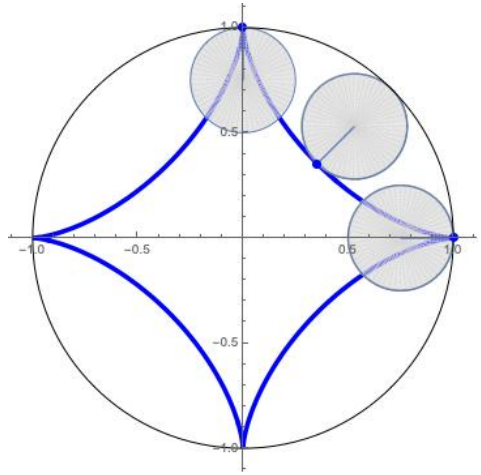
The area is obtained by adding these and letting $n \to \infty$:

$$P = \lim_{n\to\infty} \frac{1}{2} \sum_{i=0}^{n-1} |y'(t_i)x(t_i) - x'(t_i)y(t_i)|\Delta t$$

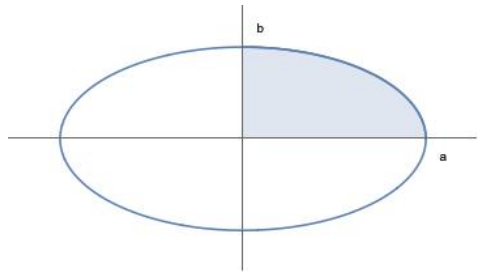
$$= \frac{1}{2} \int_a^b |x(t)y'(t) - y(t)x'(t)|\, dt.$$

Problem: the area bounded by

1. the asteroid $x(t) = \cos^3 t, y(t) = \sin^3 t, t \in [0, 2\pi]$ is



2. the elipse $x = a\cos t,\ y = b\sin t, t \in [0, 2\pi]$ is



Hint. In both problems use the identities

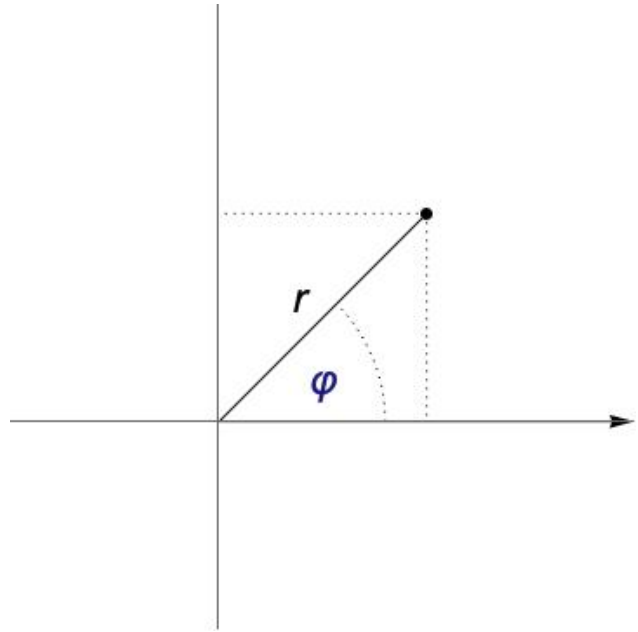
$$\sin^2 t = \frac{1}{2}(1 - \cos(2t)), \qquad \cos^2 t = \frac{1}{2}(1 + \cos(2t)).$$

In the first problem all you have to really integrate after subtractions of some terms is $1 - \cos^2(2t)$. The results are $\frac{3\pi}{8}$ for the first and $ab\pi$ for the second problem.

Curves in the polar plane

Polar coordinates of a point in the plane are

- distance to the origin r, $r \geq 0$, and
- polar angle φ, determined up to a multiple of 2π, defined for $r \neq 0$.



Usually the polar axis corresponds to the positive part of the x-axis, so

- $x = r\cos\varphi$, $y = r\sin\varphi$
- $r = \sqrt{x^2 + y^2}$, $\tan\varphi = \frac{y}{x}$

A curve in polar coordinates is given by $r = r(\varphi), \quad \varphi \in I \subset \mathbb{R}$.

Rule. If $r(\varphi) < 0$, then the point on the curve at an angle φ is equal to

$$(x(\varphi), y(\varphi)) := |r(\varphi)|(\cos\varphi, \sin\varphi) \cdot e^{i\pi}.$$

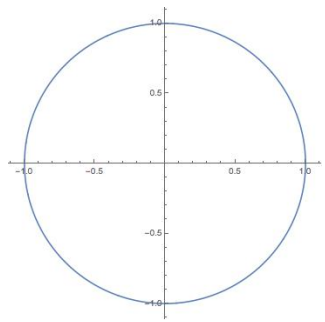
In other words, we reflect the point

$$|r(\varphi)|(\cos\varphi, \sin\varphi)$$

over the origin.

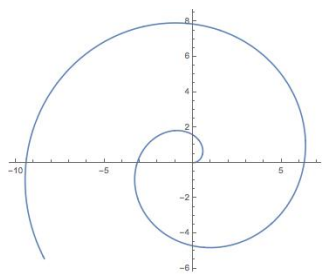
Example

$r = 1$



unit circle

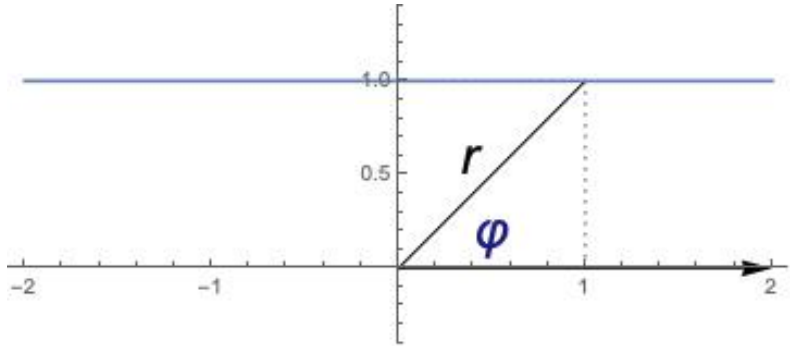
$r = \varphi$



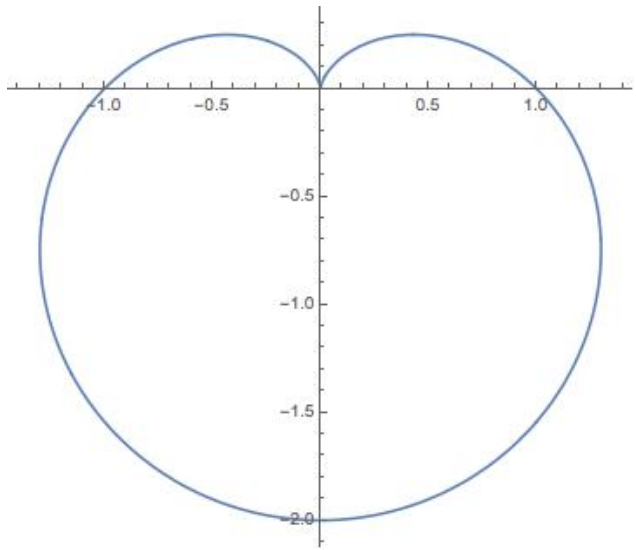
Arhimedean spiral

Example

line $y = 1$, $\quad r = \frac{1}{\sin\varphi}$



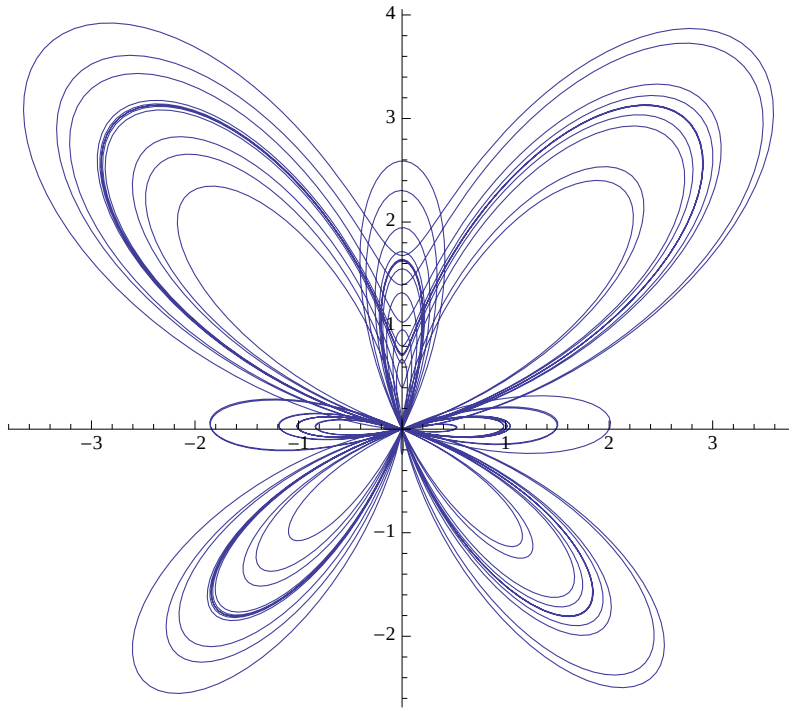
cardioid, $\quad r = 1 - \sin\varphi$



Example

a butterfly

$$r = \sin^5\left(\frac{\varphi-\pi}{12}\right) + e^{\sin\varphi} - 2\cos(4\varphi)$$



Matlab files:

`https://zalara.github.io/Algoritmi/curves_polar.m`

A parametrization of the curve with parameter being the polar angle is:

$$f(\varphi) = \begin{bmatrix} r(\varphi)\cos(\varphi) \\ r(\varphi)\sin(\varphi) \end{bmatrix}, \varphi \in I.$$

Example

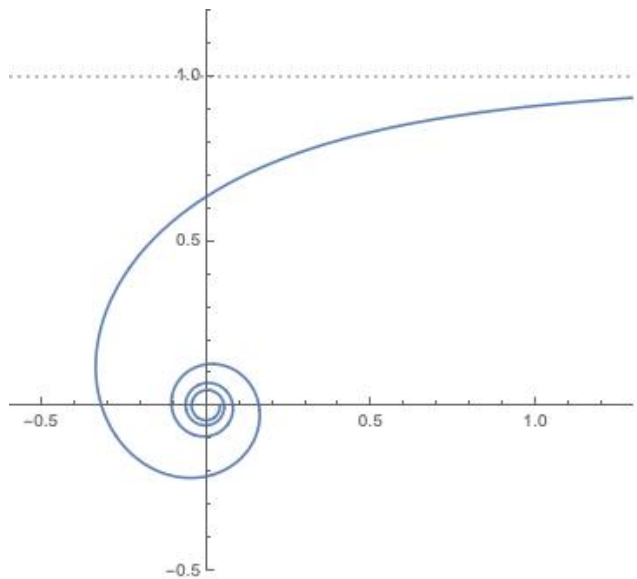
The hyperbolic spiral $r = \dfrac{1}{\varphi}$ is parametrized by $f(t) = \begin{bmatrix} \frac{\cos\varphi}{\varphi} \\ \frac{\sin\varphi}{\varphi} \end{bmatrix}$,

as $\varphi \to 0$, $\quad r(\varphi) \to \infty$

$$x(\varphi) = \frac{\cos\varphi}{\varphi} \to \infty$$

$$y(\varphi) = \frac{\sin\varphi}{\varphi} \to 1$$

as $\varphi \to \infty$, $\quad r(\varphi) \to 0$



The tangent vector to the curve at a point $r(\varphi)$ is given by

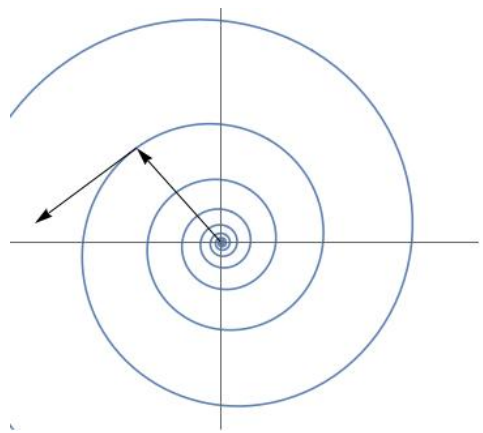
$$f'(\varphi) = \begin{bmatrix} r'(\varphi)\cos(\varphi) - r(\varphi)\sin(\varphi) \\ r'(\varphi)\sin(\varphi) + r(\varphi)\cos(\varphi) \end{bmatrix}$$

Problem: compute the angle between the coordinate vector of a point on the **logarithmic spiral** $r(\varphi) = be^{a\varphi}$ and the tangent vector at that point.

coordinate vector: $f(\varphi) = \begin{bmatrix} be^{a\varphi}\cos(\varphi) \\ be^{a\varphi}\sin(\varphi) \end{bmatrix}$,

tangent vector: $f'(\varphi) = \begin{bmatrix} be^{a\varphi}(a\cos\varphi - \sin\varphi) \\ be^{a\varphi}(a\sin\varphi + \cos\varphi) \end{bmatrix}$,

angle: $\cos\alpha = \dfrac{f(t)\cdot f'(t)}{\|f(t)\|\,\|f'(t)\|} = \dfrac{a}{\sqrt{1+a^2}}$,



so the angle is independent of φ so it is the same at every point on the curve.

Area in polar coordinates

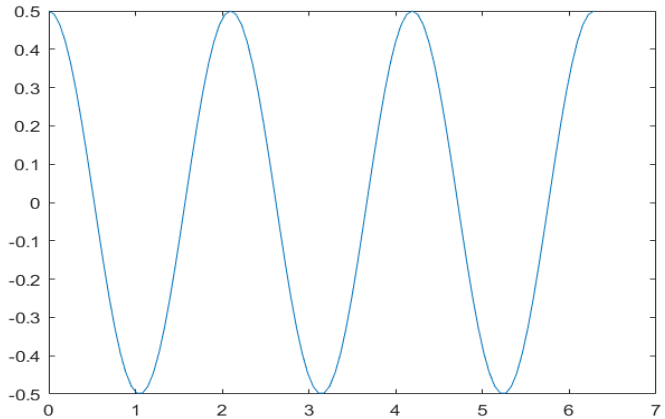
$$P = \frac{1}{2}\int_{\alpha}^{\beta} |xy' - x'y|\, d\varphi = \frac{1}{2}\int_{\alpha}^{\beta} r^2\, d\varphi$$

Indeed:

$$\begin{aligned} xy' - x'y &= r\cos\varphi(r'\sin\varphi + r\cos\varphi) - r\sin\varphi(r'\cos\varphi - r\sin\varphi) \\ &= r^2(\cos^2\varphi + \sin^2\varphi) = r \end{aligned}$$

Problem: what is the area of one petal of the clover $r(\varphi) = \dfrac{\cos(3\varphi)}{2}$?

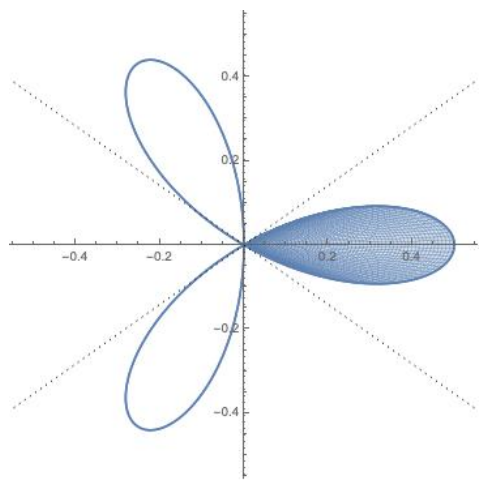
To plot the clover it is convenient to sketch the function $r(\varphi)$ first.



Useful angles are

φ	0	$\frac{\pi}{6}$	$\frac{2\pi}{6}$	$\frac{3\pi}{6}$	$\frac{4\pi}{6}$	$\frac{5\pi}{6}$	$\frac{6\pi}{6}$	$\frac{7\pi}{6}$	$\frac{8\pi}{6}$	$\frac{9\pi}{6}$	$\frac{10\pi}{6}$	$\frac{11\pi}{6}$	$\frac{12\pi}{6}$
$r(\varphi)$	$\frac{1}{2}$	0	$-\frac{1}{2}$	0	$\frac{1}{2}$	0	$-\frac{1}{2}$	0	$\frac{1}{2}$	0	$-\frac{1}{2}$	0	$\frac{1}{2}$

$$P = 2\int_0^{\pi/6} \frac{\cos^2(3\varphi)}{4}\, d\varphi = \frac{\pi}{12}$$



Motion in $\mathbb{R}^3$

Let $\mathbf{r}(t) = f(t)$ be the position vector of a particle in space at time t, $1 \leq t \leq 2$.

Then $\mathbf{v}(t) = \mathbf{r}'(t)$ is its velocity and $\mathbf{a}(t) = \mathbf{r}''(t)$ is its acceleration at time t.

Problem: Let $\mathbf{r}(t) = \begin{bmatrix} t^2 \\ 2t \\ \log t \end{bmatrix}$.

1. Compute its position, velocity and acceleration at time $t = 1$, and the length of its path between $t = 1$ and $t = 2$.
2. If at time $t = 2$ the particle leaves its path and goes off in the tangential direction with constant velocity, where will it be at time $t = 3$? What is the length of its path from $t = 1$ to $t = 3$?

1. Since $\mathbf{r}'(t) = \begin{bmatrix} 2t \\ 2 \\ 1/t \end{bmatrix}$ and $\mathbf{r}''(t) = \begin{bmatrix} 2 \\ 0 \\ -1/t^2 \end{bmatrix}$, the position, velocity and acceleration at $t = 1$ are

$$\mathbf{r}(1) = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}, \quad \mathbf{v}(1) = \begin{bmatrix} 2 \\ 2 \\ 1 \end{bmatrix} \quad \mathbf{a}(1) = \begin{bmatrix} 2 \\ 0 \\ -1 \end{bmatrix}$$

and the length of path

$$\int_1^2 \|\mathbf{r}'(t)\|\, dt = \int_1^2 \sqrt{4t^2 + 4 + (1/t)^2}\, dt = \int_1^2 (2t + 1/t)\, dt =$$

$$[2t^2/2 + \log t]_1^2 = 3 + \log 2$$

2. The tangent line at $t = 2$, and the position at $t = 3$ are:

$$L_2(t) = \begin{bmatrix} 4 \\ 4 \\ \log 2 \end{bmatrix} + (t-2) \begin{bmatrix} 4 \\ 2 \\ 1/2 \end{bmatrix}, L_2(3) = \begin{bmatrix} 8 \\ 6 \\ \log 2 + 1/2 \end{bmatrix}$$

and length of the path along the tangent from $t = 2$ to $t = 3$ is

$$\int_2^3 \|\mathbf{v}(2)\|\, dt = 9/2,$$

so the total length is $\log 2 + 7 + \frac{1}{2}$.

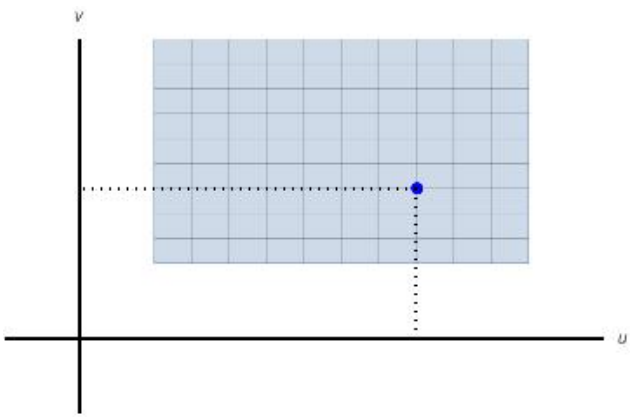
Parametric surfaces

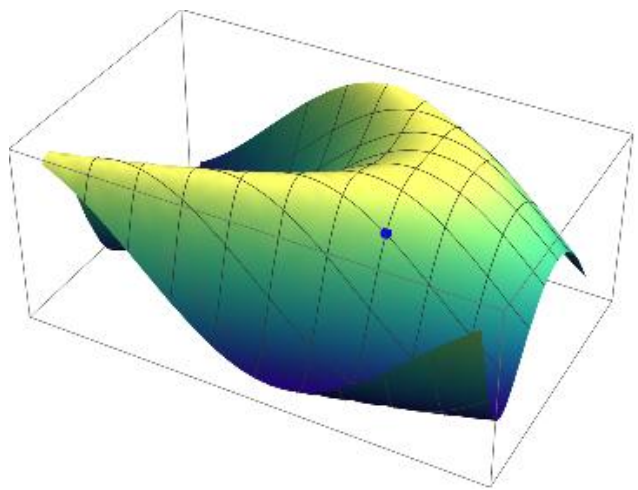
A parametric surface in $\mathbb{R}^m$ is given by a continuous vector function

$$f : D \to \mathbb{R}^m, \qquad D \subset \mathbb{R}^2.$$

We will consider the case $m = 3$:

$$\begin{bmatrix} u \\ v \end{bmatrix} \in D \qquad f(u, v) = \begin{bmatrix} x(u, v) \\ y(u, v) \\ z(u, v) \end{bmatrix} \in \mathbb{R}^3$$

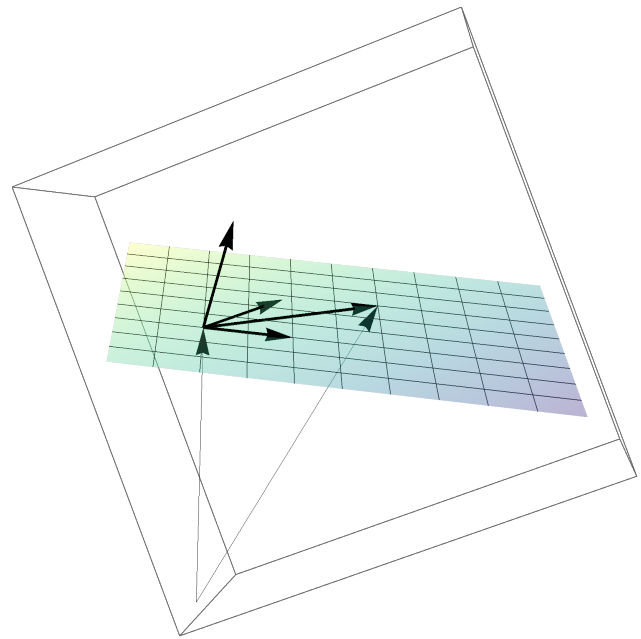




Example

1. A parametric plane through a given point $\mathbf{r}_0 \in \mathbb{R}^3$ with given (noncolinear) vectors $\mathbf{e}_1$ and $\mathbf{e}_2$:

$$f(u, v) = \mathbf{r}_0 + u\mathbf{e}_1 + v\mathbf{e}_2, \quad u, v \in \mathbb{R},$$



The normal to the plane is $\mathbf{n} = \mathbf{e}_1 \times \mathbf{e}_2 \neq 0$.

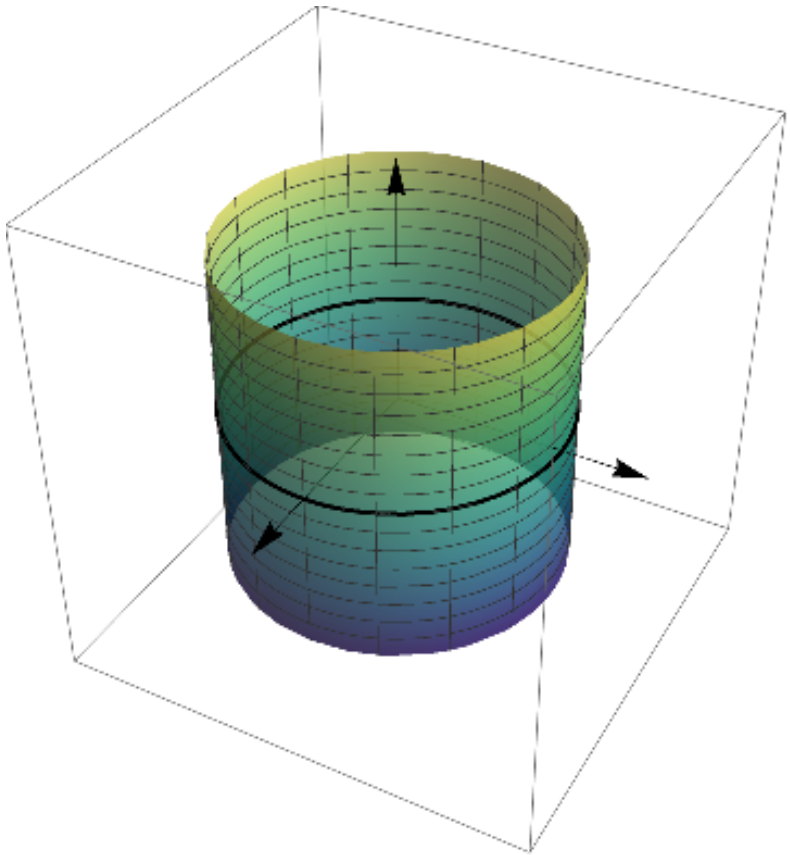
The equation the plane: $(\mathbf{r} - \mathbf{r}_0) \cdot \mathbf{n} = 0$

Matlab file:

https://zalara.github.io/Algoritmi/plane.m

2.

$$f(u,v) = \begin{bmatrix} \cos u \\ \sin u \\ v \end{bmatrix}, \quad u \in [0, 2\pi], v \in [0, 1]$$



a cylinder with radius 1 and axis the z-axis

Matlab file:

https://zalara.github.io/Algoritmi/cylinder.m

For every point $f(u_0, v_0)$ on the surface there are two coordinate curves through it:

- $f(u_0, v)$,
- $f(u, v_0)$,

both lie on the surface.

Example

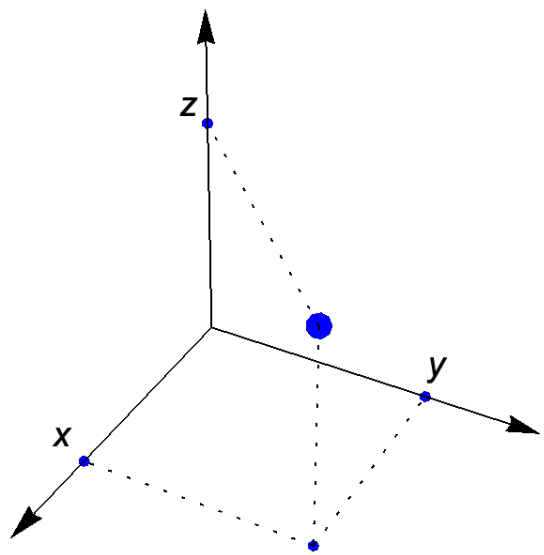
1. In the parametrized plane $f(u, v) = r_0 + ue_1 + ve_2$, $e_1 \times e_2 \neq 0$, coordinate curves are lines parallel to e_2 for a fixed $u = u_0$ and to e_1 for a fixed $v = v_0$.
2. In the cylinder, coordinate curves $u = u_0$ are vertical lines, and $v = v_0$ are circles.

Coordinate systems in $\mathbb{R}^3$

The parameters u and v in surface parametrizations often have a geometric meaning.

For example, they could be two coordinates from one of the standard coordinate systems in $\mathbb{R}^3$:

Cartesian coordinates x, y, z (we know these well)



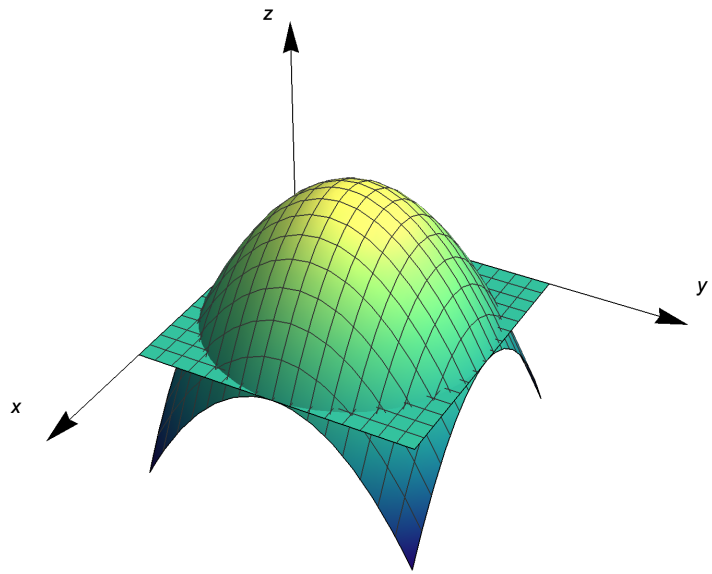
Example

$$f(x,y) = \begin{bmatrix} x \\ y \\ 1-(x-1)^2-(y-1)^2 \end{bmatrix}, \; 0 \le x, y \le 2$$

The surface is the graph
$z = 1-(x-1)^2-(y-1)^2$,

Coordinate curves:
intersection with planes

$x = x_0$ and $y = y_0$



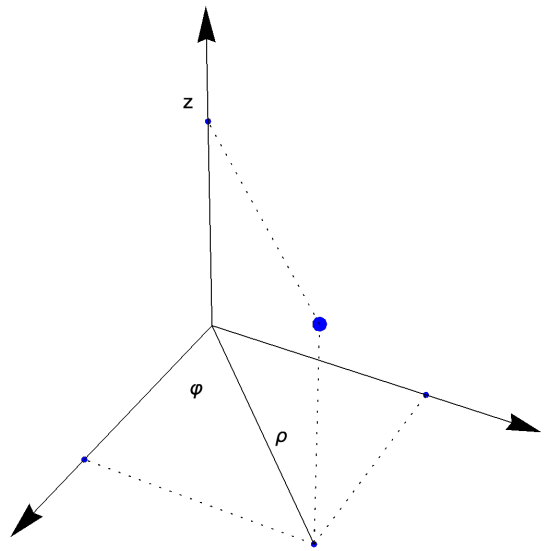
Matlab file:

https://zalara.github.io/Algoritmi/surfaces_coordinate_curves.m

Cylindrical coordinates:

$\rho \geq 0$ distance from z axis, polar radius in plane $z = 0$

φ polar angle in plane $z = 0$



Conversion to cartesian coordinates: $x = \rho \cos \varphi, y = \rho \sin \varphi, z = z$

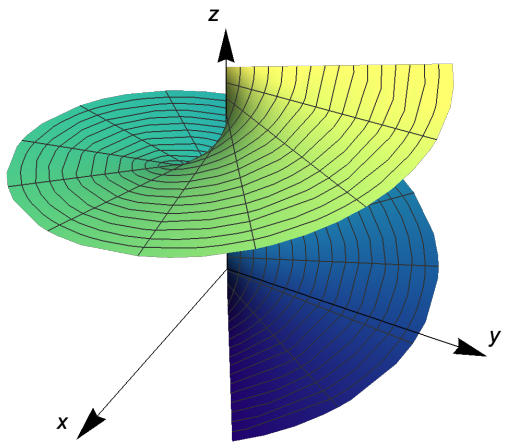
Example

$$f(u, v) = \begin{bmatrix} u \cos v \\ u \sin v \\ v \end{bmatrix}$$

Coordinate curves:

$u = u_0$: helix with radius u_0

$v = v_0$: ray from z-axis with polar angle and height v_0



Matlab file:

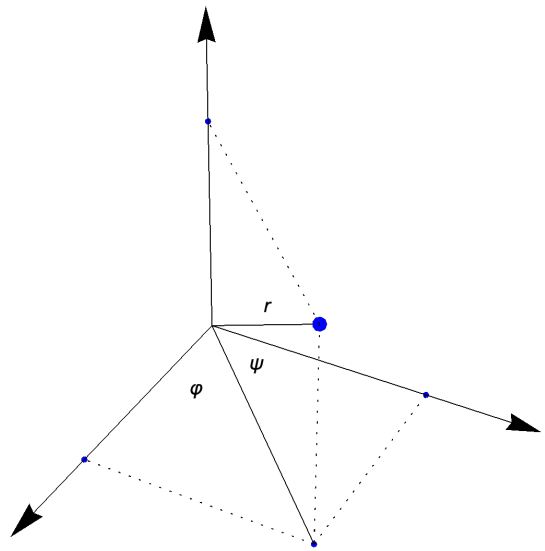
https://zalara.github.io/Algoritmi/cylindrical_coordinates_helix.m

Spherical coordinates: r, φ, ψ, where

r, $r \geq 0$: distance to the origin,

φ: polar angle in plane $z = 0$

ψ, $-\pi/2 \leq \psi \leq \pi/2$: azimuthal angle between the coordinate vector and plane $z = 0$,



$\psi = \pi/2$: positive part of z axis
$\psi = 0$: plane $z = 0$
$\psi = -\pi/2$ negative part of z-axis

Conversion to cartesian coordinates: $x = r\cos\varphi\cos\psi$, $y = r\sin\varphi\cos\psi$, $z = r\sin\psi$

Conversion to cylindrical coordinates: $\rho = r\cos\psi$, $z = r\sin\psi$

Example

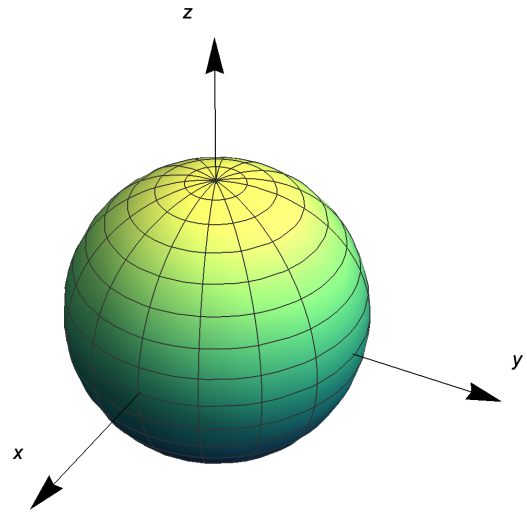
$$f(u,v) = \begin{bmatrix} \cos u \cos v \\ \sin u \cos v \\ sinv \end{bmatrix}, \; 0 \leq u \leq 2\pi, -\pi/2 \leq v \leq \pi/2$$

The surface is the unit sphere $r = 1$

Coordinate curves:

$u = u_0$: latitude $u = u_0$

$v = v_0$: longitude $v = v_0$



Matlab file:

https://zalara.github.io/Algoritmi/spherical_coordinates.m

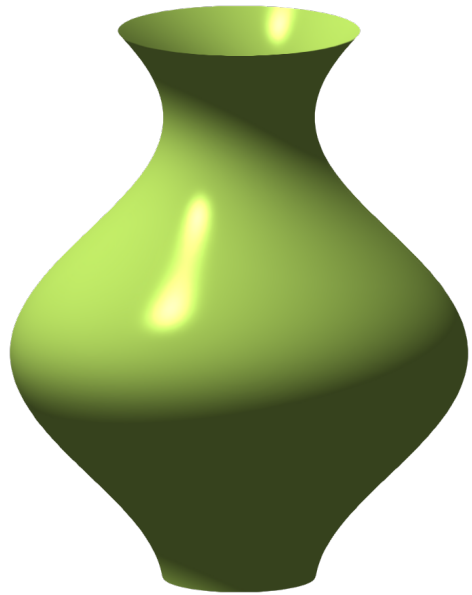
Surfaces of revolution

A surface of revolution is obtained by revolving a curve $x = x(u), z = z(u)$ in the (x, z)-plane around the z axis:

$$f(u, v) = \begin{bmatrix} x(u)\cos v \\ x(u)\sin v \\ z(u) \end{bmatrix}$$

$u \in [a, b]$

$v \in [0, 2\pi]$,



$x = 2 + \cos t, z = t$, from wikipedia

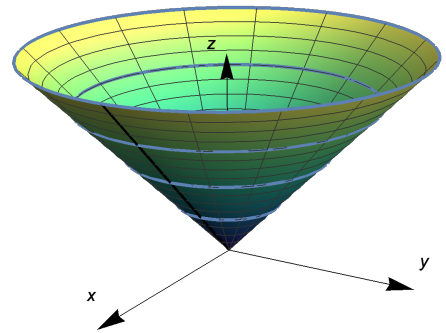
Coordinate curves:

- $u = u_0$, horizonal circle $\begin{bmatrix} x(u_0)\cos v \\ x(u_0)\sin v \\ z(u_0) \end{bmatrix}$,
- $v = v_0$, original curve rotated by the angle v_0

Example

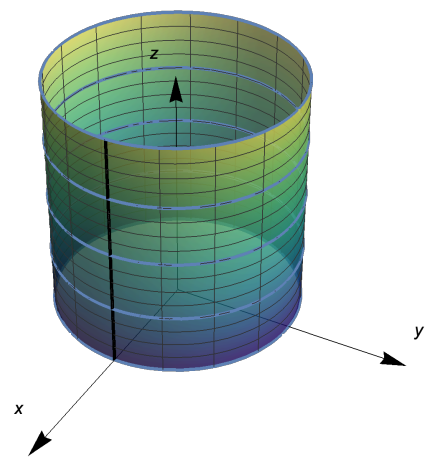
Revolving the line $z = x = u$: a cone

$$\begin{bmatrix} u\cos v \\ u\sin v \\ u \end{bmatrix}$$



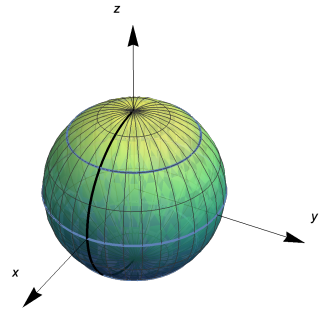
Revolving the line $x = a$: a cylinder

$$f(u,v) = \begin{bmatrix} a\cos v \\ a\sin v \\ u \end{bmatrix}$$

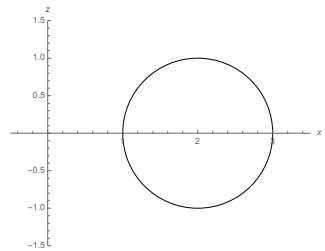


Revolving the half-circle
$x = \cos t, z = \sin t, -\pi/2 \leq t \leq \pi/2$: a sphere

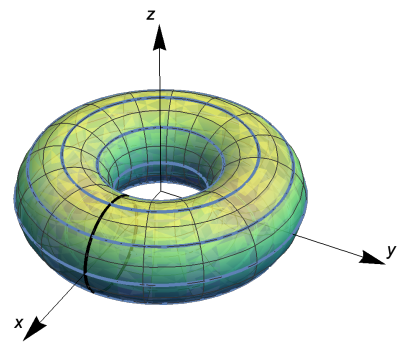
$$f(t, v) = \begin{bmatrix} \cos t \cos v \\ \cos t \sin v \\ \sin t \end{bmatrix}$$



Revolving the circle $x = 2 + \cos t, z = \sin t$,
$0 \leq t \leq 2\pi$: a torus



$$f(u, v) = \begin{bmatrix} (2 + \cos t) \cos v \\ (2 + \cos t) \sin v \\ \sin t \end{bmatrix}$$



Matlab file: https://zalara.github.io/Algoritmi/surface_of_revolution.m

Smooth surfaces

Let $\mathbf{r}_0 = f(u_0, v_0)$ be a point on the surface.

Coordinate curves through this point:

- $f(u_0, v)$ with parameter v and tangent vector $f_v(u_0, v_0)$,
- $f(u, v_0)$ with parameter u and tangent vector $f_u(u_0, v_0)$.

The parametric surface is smooth at the point $f(u_0, v_0)$, if both tangent vectors exist and

$$f_u(u_0, v_0) \times f_v(u_0, v_0) \neq 0.$$

The vector $\mathbf{n}_0 = f_u(u_0, v_0) \times f_v(u_0, v_0)$ is the normal vector to the surface at the point $\mathbf{r}_0$.

Tangent plane

If the surface is smooth at a point $\mathbf{r}_0 = f(u_0, v_0)$ then it has a tangent plane at this point that is given:

- in implicit form by $(\mathbf{r} - \mathbf{r}_0) \cdot \mathbf{n}_0 = 0$
- in parametric form by

$$\mathbf{r}(u, v) = \mathbf{r}_0 + uf_u(u_0, v_0) + vf_v(u_0, v_0) = L_{(u_0,v_0)}(u, v)$$

where $L_{(u_0,v_0)}(u, v) = f(u_0, v_0) + Df(u_0, v_0) \begin{bmatrix} u \\ v \end{bmatrix}$ is the linear approximation and

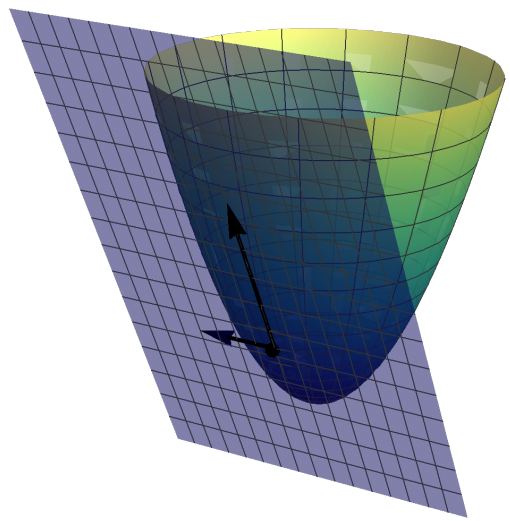
$$Df(u_0, v_0) = \begin{bmatrix} x_u(u_0, v_0) & x_v(u_0, v_0) \\ y_u(u_0, v_0) & y_v(u_0, v_0) \\ z_u(u_0, v_0) & z_v(u_0, v_0) \end{bmatrix}$$

is the Jacobian.

Problem: find the tangent plane to the surface $f(u,v) = \begin{bmatrix} u\cos v \\ u\sin v \\ u^2 \end{bmatrix}$ at $u = 1, v = \pi/2$.

Since $f_u(u,v) = \begin{bmatrix} \cos v \\ \sin v \\ 2u \end{bmatrix}$ and $f_v(u,v) = \begin{bmatrix} -u\sin v \\ u\cos v \\ 0 \end{bmatrix}$ the tangent plane in parametric form is

$$\mathbf{r}(u,v) = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} + u \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix} + v \begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix}.$$



In implicit form:

$$\mathbf{n} = f_u(u_0,v_0) \times f_v(u_0,v_0) = \begin{bmatrix} 0 \\ 1 \\ -2 \end{bmatrix} \times \begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix}$$

so:

$$\mathbf{n} \cdot (\mathbf{r} - \mathbf{r}_0) = -2(y-1) + (z-1) = 0.$$

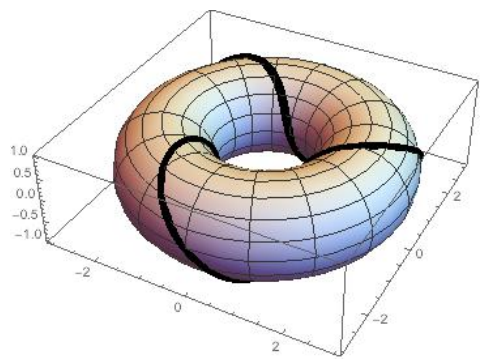
A curve $\alpha(t) = \begin{bmatrix} u(t) \\ v(t)) \end{bmatrix}$ in the (u, v)-plane corresponds to a curve on the parametric surface:

$$f(\alpha(t)) = \begin{bmatrix} x(u(t), v(t)) \\ y(u(t), v(t)) \\ z(u(t), v(t)) \end{bmatrix}$$

Example

The line $\alpha(t) = \begin{bmatrix} 3t \\ t \end{bmatrix}$ corresponds to a curve on the torus

$$(f \circ \alpha)(t) = \begin{bmatrix} (2 + \sin 3t) \cos t \\ (2 + \sin 3t) \sin t \\ \cos 3t \end{bmatrix}$$



An application: the configuration space of a robot

A robot, or a mechanical device, is described by its

- work space : the space of points reached by the end effector
- configuration space: the space of parameter values that determine the position of the robot

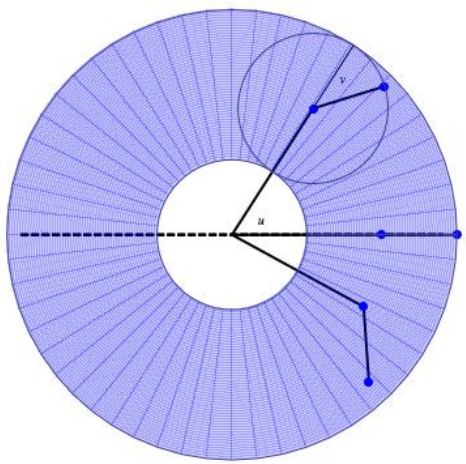
The number of parameters is the degrees of freedom, (DOF), this determines the dimension of the configuration space.

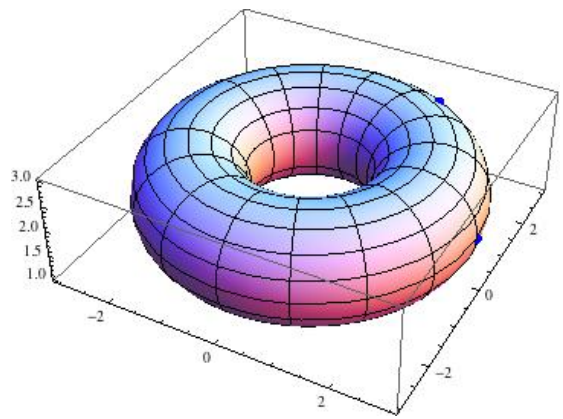
If DOF=2, thet the configuration space is (often) a parametric surface.

Classical example: A robotic arm with two links of lengths l_1 and l_2, $l_2 < l_1$ and two rotational joints.

The work space is a ring with interior circle of radius $l_1 - l_2$ and exterior circle of radius $l_1 + l_2$,

The configuration space is parametrized by the angles u and v in the joints, the two independent rotations can be represented by a torus: $\begin{bmatrix} (a + b\cos u)\cos v \\ (a + b\cos u)\sin v \\ b\sin u \end{bmatrix}$, $b < a$, $u \in [0, 2\pi]$, $v \in [0, 2\pi]$





In robot motion planning, the motion of the robotic arm from point T_0 to T_1 in the work space is directed by a curve, or path, in the configuration space.

Chapter 5:

Differential equations and dynamic models

- Ordinary differential equation (ODE)
  - Definition and examples
  - Solving first order ODEs
    - Separable ODEs
    - First order linear ODEs
    - Homogeneous ODEs
  - Orthogonal trajectories
  - Exact ODEs
  - Geometric picture of ODEs
- Systems of first order ODEs
- Numerical methods for solving ODEs
- Autonomous system of ODEs
- Dynamics of systems of 2 linear ODEs
- Linear ODEs of order n
- Application - vibrating systems

Differential equations and dynamic models

Ordinary differential equation, ODE, is an equation of an unknown function and an independent variable. ODE relates the independent variable with the function and its derivatives.

If t is an independent variable, $x(t)$ is a function of t, then the ODE is of the form:

$$F(t, x, \dot{x}, \ddot{x}, \ldots, x^{(n)}) = 0.$$

Similarly if x is an independent variable, $y(x)$ a function of x, then the ODE is of the form:

$$F(x, y, y', y'', \ldots, y^{(n)}) = 0.$$

The order of a differential equation is the order of the highest derivative.

Examples of ODEs

- $\dot{x} - 3t^2 = 0.$

  So,

  $$\frac{dx}{dt} = 3t^2 \;\Rightarrow\; x(t) = t^3 + C, \quad \text{where } C \text{ is a constant.}$$

  If we want to determine C, we need an additional condition, e.g., initial condition $x(0) = x_0$, $x_0 \in \mathbb{R}$, or any other condition $x(t_0) = x_0$, $x_0 \in \mathbb{R}$.

- $y''(x) + 2y'(x) = 3y(x).$

  We will learn how to solve such an ODE, but right now let us only check that $y(x) = Ce^{-3x}$, $C \in \mathbb{R}$ a constant, is a solution:

  - Calculate $y''(x)$, $y'(x)$:

    $$y'(x) = -3Ce^{-3x}, \quad y''(x) = 9Ce^{-3x}.$$

  - Plug into the given ODE:

    $$9Ce^{-3x} - 6Ce^{-3x} = 3Ce^{-3x}.$$

- $\cos t \cdot \ddot{x} - 3t^4 \cdot \dot{x} + 5e^t = 0.$

  Such ODE's cannot be solved analytically (or are at least hard to solve). We will learn how to solve such ODE's by using numerical methods.

Partial differential equation, PDE, is an equation for an unknown function u of $n \geq 2$ independent variables, e.g., for $n = 2$ we have

$$F(x, y, u_x, u_y, u_{xx}, \ldots) = 0,$$

where x, y are the independent variables.

We will not consider PDE's, from now on DE means an ODE.

Applications of DEs

Differential equations are used for modelling a deterministic process: a law relating a certain quantity depending on some independent variable (for example time) with its rate of change, and higher derivatives.

1. Newton's law of cooling:

$$\dot{T} = k(T - T_\infty), \tag{23}$$

where $T(t)$ is the temperature of a homogeneous body (can of beer) at time t, T_0 is the initial temperature at time $t_0 = 0$, T_∞ is the temperature of the environment, k is a constant (heat transfer coefficient).

(23) is an example of a separable ODE and also the first order linear ODE. We will see shortly how to solve such types of ODE's. For now you can check easily by yourself that the solution is

$$T(t) = (T_0 - T_\infty)e^{kt}.$$

2. Radioactive decay:

$$\dot{y}(t) = -ky(t), \quad k = \frac{\log 2}{t_{1/2}},$$

where $y(t)$ is the remaining quantity of a radioactive isotope at time t, $t_{1/2}$ is the half-life and k is the decay constant. The solution is

$$y(t) = Ce^{-kt}, \quad \text{where } C \text{ is a constant.}$$

Let's verify, that $t_{1/2}$ really represents the time in which the amount of the isotope decreases to half of its current amount. At time $t = 0$ the amount is $y(0) = Ce^0 = C$. We have to check that $y(t_{1/2}) = \frac{C}{2}$:

$$y(t_{1/2}) = Ce^{-\frac{k \log 2}{k}} = Ce^{-\log 2} = Ce^{\log 1/2} = \frac{C}{2}.$$

3. Simple harmonic oscillator:

$$\ddot{x} + \omega x = 0.$$

Solution of a DE

The function $x(t)$ is a solution of a DE

$$F(t, x, \dot{x}, \ddot{x}, \ldots, x^{(n)}) = 0$$

on an interval I if it is at least n times differentiable and satisfies the identity

$$F(t, x(t), \dot{x}(t), \ddot{x}(t), \ldots, x^{(n)}(t)) = 0$$

for all $t \in I$.

Analytically solving a DE is typically very difficult, very often impossible.

To find approximate solutions we use different simplifications and numerical methods.

First order ODEs

We will (mostly) consider first order ODEs in the form

$$\dot{x} = f(t, x).$$

- The general solution is a one-parametric family of solutions $x = x(t, C)$.
- A particular solution is a specific function from the general solution, that usually satisfies some initial condition $x(t_0) = x_0$.
- A singular solution is an exceptional solution that is not part of the general solution.

We will first look at some simple types of 1.-st order DEs that are analytically solvable.

Separable DE

A separable DE is of the form

$$\dot{x} = f(t)g(x). \tag{24}$$

This can be solved by:

- Inserting $\dot{x} = \dfrac{dx}{dt}$ into (24):

$$\frac{dx}{dt} = f(t)g(x). \tag{25}$$

- Separating variables in (25):

$$\frac{dx}{g(x)} = f(t)\,dt. \tag{26}$$

- Integrating both sides of (25):

$$\int \frac{1}{g(x)}\,dx = \int f(t)\,dt + C$$

Example 1 of a separable DE

$$\boxed{\dot{x} = kx} \quad \text{where } k \in \mathbb{R} \text{ is a fixed real number} \tag{27}$$

- $$\frac{dx}{dt} = kx,$$
- $$\frac{dx}{x} = k dt,$$
- $$\log|x| = \int \frac{dx}{x} = \int k\, dt = kt + C,$$

where C is a constant and so

$$|x| = e^{kt+C}$$

is a general solution to (27). Clearly, $x(t) = 0$ is also a solution of the equation. By introducing a new constant e^C which, by abuse of notation, we again denote by C, this is equivalent to

$$x(t) = Ce^{kt}, C \in \mathbb{R}.$$

Example 2 of a separable DE

$$\boxed{\dot{x} = kx(1-x)} \quad \text{where } k \in \mathbb{R} \text{ is a fixed real number} \tag{28}$$

- $$\frac{dx}{dt} = kx(1-x),$$
- $$\frac{dx}{x(1-x)} = k\,dt,$$
- By the method of partial fractions we get

$$\log\left|\frac{x}{1-x}\right| = \log|x| - \log|1-x| = \int \frac{dx}{x} - \int \frac{dx}{1-x} = \int k\,dt = kt + C,$$

where C is a constant and so

$$\frac{x}{1-x} = Ce^{kt}.$$

Expressing $x(t)$ we get

$$x(t) = \frac{1}{Ce^{-kt}+1} \tag{29}$$

is a general solution to (28). $x(t)$ from (29) is called a logistic function.

Example 3 of a separable DE

$$\boxed{y' = \frac{-x}{ye^{x^2}},} \quad y(0) = 1. \tag{30}$$

- $$\frac{dy}{dx} = \frac{-x}{ye^{x^2}},$$
- $$ydy = -xe^{-x^2}dx,$$
- Integrating:

$$\frac{y^2}{2} = \int ydy = \int (-xe^{-x^2})dx = \frac{1}{2}e^{-x^2} + C,$$

where C is a constant.

- $\frac{1}{2} = \frac{y^2(0)}{2} = \frac{1}{2} + C \ \Rightarrow \ C = 0.$

Expressing $y(x)$ we get $y(x) = \pm\sqrt{e^{-x^2}}$ and since $y(0) > 0$ we have

$$y(x) = \sqrt{e^{-x^2}}.$$

Real life DE example: population growth

Let $x(t)$ be the size of a population (bacteria, trees, people, ...) at time t. The most common models for population growth are:

- exponential growth: the growth rate is proportional to the size, modelled by $\dot{x} = kx$, with the solution the exponential function $x(t) = x_0 e^{kt}$, where $x_0 = x(0)$ is the initial population size.
- logistic growth: the growth rate is proportional to the size and the resources, modelled by $\dot{x} = kx(1 - x/x_{max})$, where x_{max} is the capacity of the environment, i.e., maximal population size that it still supports, with the solution is the logistic function.
- general model: the growth rate is proportional to the size, but the proportionality factor depends on time and size, modelled by $\dot{x} = k(x, t)f(x)$; the equation is not separable and is analytically solvable only in very specific cases.

Real life DE example: information spreading

$x(t)$ is the ratio of people in a given group that at time t knows a certain piece of information.

Let $x_0 = x(t_0)$ be the 'informed' ratio at time $t = t_0$.

Consider two possible models:

- spreading through an external source: the rate of change is proportional to the uninformed ratio $\dot{x} = k(1 - x)$ with $x_0 = 0$,

- spreading through "word of mouth" the rate of change is proportional to the number of encounters between informed and uninformed members $\dot{x} = kx(1 - x)$ logistic law, again, with $x_0 > 0$.

First order linear ODE

A first order linear DE is of the form

$$\dot{x} + f(t)x = g(t) \tag{31}$$

The equation is homogeneous if $g(t) = 0$ and nonhomogenous if $g(t) \neq 0$.

A homogeneous part of (31),

$$\dot{x} + f(t)x = 0, \tag{32}$$

has a general solution of the form

$$Cx_h(t), \tag{33}$$

where $C \in \mathbb{R}$ is a constant and $x_h(t)$ is a particular solution. Indeed:

- Every $x(t)$ of the form (33) is a solution of (32):

$$\begin{aligned} x'(t) + f(t)x(t) &= (Cx_h)'(t) + f(t)Cx_h(t) \\ &= Cx_h'(t) + f(t)Cx_h(t) \\ &= C(x_h'(t) + f(t)x_h(t)) \\ &= 0 \end{aligned}$$

- If $x(t)$ is a solution of (32), then it must be of the form (33). Indeed, since $x(t)$ and $x_h(t)$ both solve (32),

$$\begin{aligned}\left(\frac{x(t)}{x_h(t)}\right)' &= \frac{x'(t)x_h(t) - x(t)x_h'(t)}{x_h^2(t)} \\ &= \frac{-f(t)x(t)x_h(t) + f(t)x(t)x_h(t)}{x_h^2(t)} \\ &= 0.\end{aligned}$$

Hence, $\frac{x(t)}{x_h(t)} = C$ for some constant C and $x(t)$ is of the form (33).

Let $x_p(t)$ be any particular solution of (31):

$$x_p'(t) + f(t)x_p(t) = g(t). \tag{34}$$

The general solution of (31) is a sum

$$x(t) = Cx_h(t) + x_p(t). \tag{35}$$

Indeed:

- Every $x(t)$ of the form (35) is a solution of (31):

$$\begin{aligned} x'(t) + f(t)x(t) &= (Cx_h(t) + x_p(t))' + f(t)(Cx_h(t) + x_p(t)) \\ &= Cx_h'(t) + x_p'(t) + f(t)Cx_h(t) + f(t)x_p(t) \\ &= (Cx_h'(t) + f(t)Cx_h(t)) + (x_p'(t) + f(t)x_p(t)) \\ &= 0 + g(t), \end{aligned}$$

  where we used (34) in the last equality.

- If $x(t)$ is a solution of (31), then it must be of the form (35). Indeed, since $x(t)$ and $x_p(t)$ both solve (31), $x(t) - x_p(t)$ solves the homogenous part (32) of (31). Hence, $x(t) - x_p(t) = Cx_h(t)$ for some C and $x(t) = Cx_h(t) + x_p(t)$.

The particular solution x_p can be obtained by variation of the constant, that is, by substituting the constant C is the homogenous solution by an unknown function $C(t)$ which is then determined from the equation.

Example of a linear ODEs

$$\boxed{t^2\dot{x} + tx = 1}, \quad \boxed{x(1) = 2}. \tag{36}$$

1. The homogenous part is

$$t^2\dot{x} + tx = 0. \tag{37}$$

So the solution x_h to (37) is

$$t^2\,dx = -tx\,dt \;\Rightarrow\; \frac{dx}{x} = -\frac{dt}{t} \;\Rightarrow\; \log|x| = -\log|t| + \log C = \log\frac{C}{|t|}$$

$$\Rightarrow\; x_h = \frac{C}{t}.$$

2. A particular solution of the nonhomogenous equation is obtained by variation of the constant:

$$x = \frac{C(t)}{t}, \quad \dot{x} = \frac{C'(t)t - C(t)}{t^2}$$

by inserting into (36) we obtain

$$C'(t)t - C(t) + C(t) = 1 \;\Rightarrow\; C'(t) = \frac{1}{t} \;\Rightarrow\; C(t) = \log|t|.$$

3. So the general solution of the nonhomogenous equation is

$$x(t) = \frac{C}{t} + \frac{\log|t|}{t}. \tag{38}$$

4. Finally, since $x(1) = 2$, we get by plugging $t = 1$ into (38)

$$2 = x(1) = C$$

and hence the solution of (36) is

$$x(t) = \frac{2 + \log|t|}{t}.$$

General solution of a linear DE

$$\boxed{y'(x) = f(x)y(x) + g(x)}. \tag{39}$$

1. The homogenous part is

$$y'(x) = f(x)y(x). \tag{40}$$

So the solution $y(x)$ to (40) is

$$\log|y| = \int \frac{dy}{y} = \int f(x)dx + C \;\Rightarrow\; y(x) = C \cdot e^{\int f(x)dx}$$

2. A particular solution of the nonhomogenous equation is obtained by the variation of the constant:

$$y(x) = C(x) \cdot e^{\int f(x)dx}. \tag{41}$$

$$y'(x) = C'(x) \cdot e^{\int f(x)dx} + C(x)f(x)e^{\int f(x)dx}. \tag{42}$$

Using that (39)=(42) and by inserting the RHS of (41) instead of $y(x)$ in (39), we obtain

$$C'(x) \cdot e^{\int f(x)dx} + C(x)f(x)e^{\int f(x)dx} = f(x)C(x) \cdot e^{\int f(x)dx} + g(x)$$

Hence

$$C'(x) \cdot e^{\int f(x)dx} = g(x),$$

and so

$$C(x) = \int (g(x)e^{-\int f(x)dx})dx.$$

Proposition

The solution of (39) *is*

$$y(x) = e^{\int f(x)dx}(C + \int (g(x)e^{-\int f(x)dx})dx).$$

In the example $t^2\dot{x} + tx = 1$ (or $\dot{x} = -\frac{1}{t}x + \frac{1}{t^2}$) above we get

$$\begin{aligned} x(t) &= e^{\int -\frac{1}{t}dt}\left(C + \int \left(\frac{1}{t^2}e^{\int \frac{1}{t}dt}\right)dt\right) \\ &= e^{\log|\frac{1}{t}|}\left(C + \int \left(\frac{1}{t^2}t\right)dt\right) \\ &= \frac{1}{t}(C + \log|t|). \end{aligned}$$

Real life example: Newton's second law

A ball of mass m kg is thrown vertically into the air with initial velocity $v_0 = 10$ m/s. We follow its trajectory. By Newton's second law of motion,

$$F = ma,$$

where m is the mass, $a = \dot{v} = \ddot{x}$ is acceleration and v velocity, and F is the sum of forces acting on the ball.

- Assuming no air friction the model is

  $$m\dot{v} = -mg,$$

  where g is the gravitational constant. The solution is

  $$v = -gt + C \quad \text{where } C \text{ is a constant.}$$

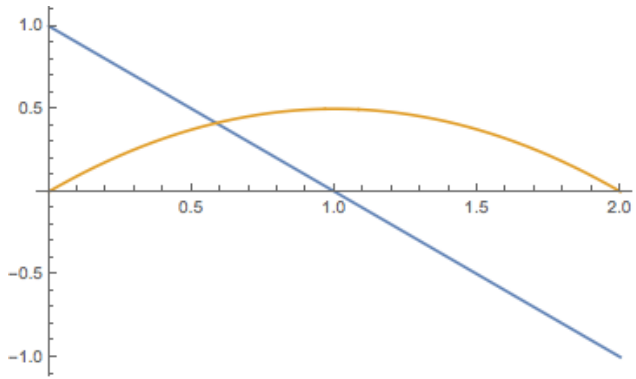
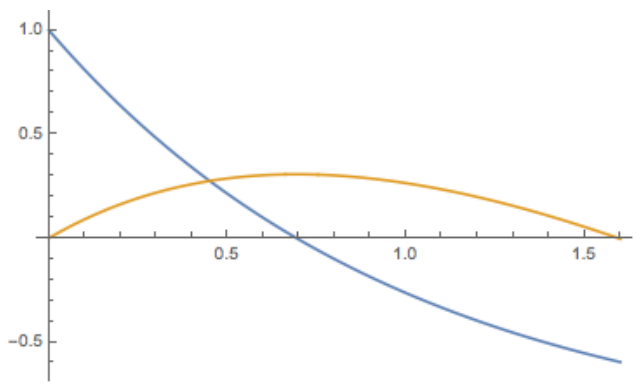
- Assuming the linear law of resistance (drag) $F_u = -kv$ the model is

  $$m\dot{v} = -mg - kv.$$

  The solution is $v = v_h + v_p$ where

  $$v_h = Ce^{-kt/m} \quad \text{and} \quad v_p = -mg/k.$$

Motion of ball in the case $m = 1$, $k = 1$ and approximating $g \doteq 10$ (we will omit units)

Model	Velocity and position	Solution
$ma = -mg$ $\dot{v} = -10$		$v(t) = -10t + 10$ $x(t) = -5t^2 + 10t$
$ma = -mg - kv$ $\dot{v} = -v - 10$		$v(t) = 20e^{-t} - 10$ $x(t) = 20 - 20e^{-t} - 10t$

The ball reaches the top at time t where $v(t) = 0$ and the ground at time t where $x(t) = 0$.

- Assuming no friction, the ball is at the top at $t = 10$.

  At time $t = 1$, $x(t) = 0$, so it takes the same time going up and falling down.

- Assuming linear friction, the ball reaches the top at $t = \log 2$.

  At time $2 \log 2$, $x(2 \log 2) = 20 - 5 - 20 \log 2 > 0$ so it takes longer falling down than going up.

Homogeneous DE

A homogeneous (nonlinear) DE is of the form

$$\dot{x} = f\left(\frac{x}{t}\right). \tag{43}$$

The solution is obtained by introducing a new dependent variable

$$u = \frac{x}{t}.$$

Hence $x = ut$ and differentiating with respect to t we get

$$\dot{x} = \dot{u}t + u. \tag{44}$$

Plugging (44) into (43) we get

$$\dot{u}t + u = f(u). \tag{45}$$

Rearranging (45) we obtain

$$t\dot{u} = f(u) - u,$$

which is a separable DE.

Example (Homogeneous DE)

$$\boxed{y' = \frac{y - x}{x}}$$

can be written as

$$y' = \frac{y}{x} - 1. \tag{46}$$

Introducing a new dependent variable

$$u = \frac{y}{x},$$

plugging in (46), we get

$$u'x + u = u - 1. \tag{47}$$

This is equivalent to

$$u'x = -1$$

and hence

$$u = \frac{y}{x} = \log\left(\frac{C}{x}\right).$$

Orthogonal trajectories

Given a 1-parametric family of curves

$$F(x, y, a) = 0 \quad \text{where} \quad a \in \mathbb{R},$$

an orthogonal trajectory is a curve

$$G(x, y) = 0$$

that intersects each curve from the given family at a right angle.

Algorithm to obtain orthogonal trajectories:

1. The family $F(x, y, a) = 0$ is the general solution of a 1st order DE, that is obtained by differentiating the equation with respect to the independent variable (using implicit differentiation) and eliminating the parameter a.
2. By substituting y' for $-1/y'$ in the DE for the original family, we obtain a DE for curves with orthogonal tangents at every point of intersection.
3. The general solution to this equation is the family of orthogonal trajectories to the original equation.

Example (Orthogonal trajectories to the family of circles)

Let us find the orthogonal trajectories to the family of circles through the origin with centers on the y axis:

$$x^2 + y^2 - 2ay = 0. \tag{48}$$

Differentiating (48) w.r.t. the independent variable gives

$$2x + 2yy' - 2ay' = 0. \tag{49}$$

Expressing a from (49) gives

$$a = \frac{x}{y'} + y. \tag{50}$$

Inserting (50) into (48) we obtain the DE for the given family

$$x^2 - y^2 - \frac{2xy}{y'} = 0. \tag{51}$$

Next we express y' from (51) and obtain

$$y' = \frac{2xy}{x^2 - y^2}. \tag{52}$$

The DE for orthogonal trajectories is obtained by substituting y' for $-1/y'$ in (52) to obtain

$$-\frac{1}{y'} = \frac{2xy}{x^2 - y^2}, \tag{53}$$

which is equivalent to

$$y' = -\frac{x^2 - y^2}{2xy}. \tag{54}$$

(54) is a homogeneous DE:

$$y' = -\frac{x^2 - y^2}{2xy} = -\frac{x}{2y} + \frac{y}{2x}$$

By introducing $y = ux$ we obtain

$$u'x + u = -\frac{1}{2u} + \frac{u}{2} \quad \Rightarrow \quad u'x = -\frac{1 + u^2}{2u} \quad \Rightarrow \quad \frac{2udu}{1 + u^2} = -\frac{dx}{x}$$

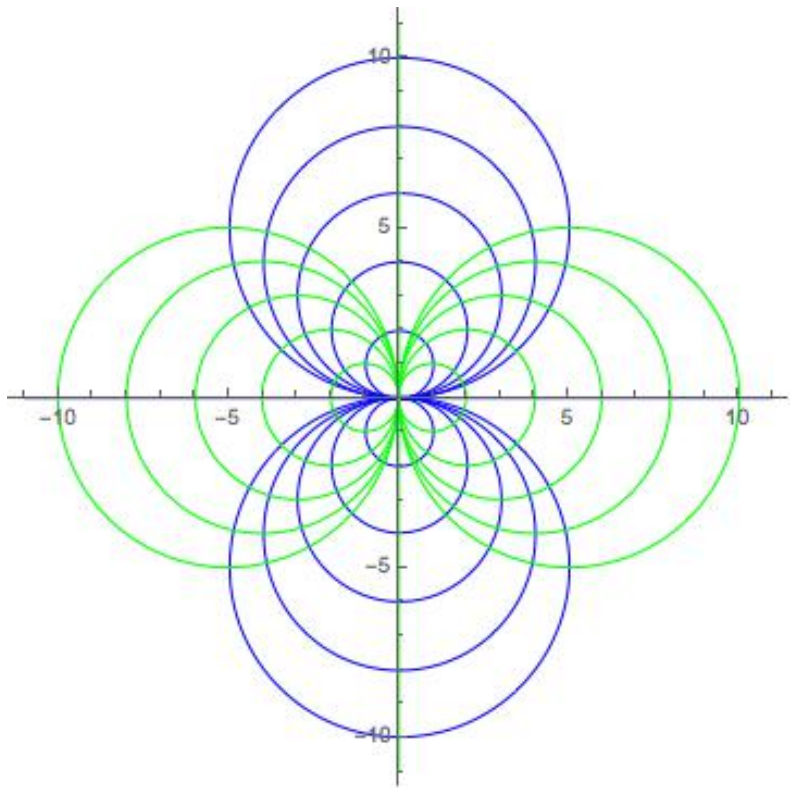
$$\Rightarrow \quad \log(1 + u^2) = -\log x + \log C \quad \Rightarrow \quad 1 + u^2 = \frac{C}{x},$$

Plugging in $u = \frac{y}{x}$ again gives the general solution

$$x^2 + y^2 = Cx.$$

Orthogonal trajectories to circles through the origin with centers on the y axis are circles through the origin with centers on the x axis.

Both families together form an orthogonal net:



Exact ODEs

Notice first that a 1st order DE

$$\dot{x} = f(t, x)$$

can be rewritten in the form

$$M(t, x)dt + N(t, x)dx = 0. \tag{55}$$

Recall that the differential of a function $u(t, x)$ is equal to

$$du = \frac{\partial u}{\partial t}dt + \frac{\partial u}{\partial x}dx = \left(\frac{\partial u}{\partial t}, \frac{\partial u}{\partial x}\right) \cdot (dt, dx),$$

where $\cdot$ denotes the usual inner product in $\mathbb{R}^2$.

DE (55) is exact if there exists a differentiable function $u(t, x)$ such that

$$\frac{\partial u}{\partial t} = M(t, x) \quad \text{and} \quad \frac{\partial u}{\partial x} = N(t, x).$$

Proposition

If the DE (55) *is exact, then the solutions are level curves of the function* u*:*

$$u(t,x) = C, \qquad \text{where} \quad C \in \mathbb{R}.$$

Recall from Calculus that if u has continuous second order partial derivatives then

$$\frac{\partial u}{\partial x \partial t} = \frac{\partial u}{\partial t \partial x}.$$

Proposition

The necessary condition for the DE (55) *to be exact is*

$$\frac{\partial M}{\partial x} = \frac{\partial N}{\partial t}. \tag{56}$$

Moreover, if M *and* N *are differentiable for every* $(t,x) \in \mathbb{R}^2$*, the condition* (56) *is also sufficient.*

A potential function u can be determined from the following equality

$$u(x,t) = \int M(t,x)\,dt + C(x) = \int N(t,x)\,dx + D(t),$$

where $C(x)$ and $D(t)$ are some functions.

Example. The DE

$$x + ye^{2xy} + xe^{2xy}y' = 0$$

can be rewritten as

$$(x + ye^{2xy})dx + xe^{2xy}\,dy = 0.$$

The equation is exact since

$$\frac{\partial(x + ye^{2xy})}{\partial y} = \frac{\partial(xe^{2xy})}{\partial x} = (e^{2xy} + 2xye^{2xy}).$$

A potential function is equal to

$$\begin{aligned} u(x,y) &= \int (x + ye^{2xy})\,dx = \frac{x^2}{2} + \frac{1}{2}e^{2xy} + C(y) \\ &= \int (xe^{2xy})\,dy = \frac{1}{2}e^{2xy} + D(x), \end{aligned}$$

Defining $C(y) = 0$ and $D(x) = x^2/2$, we get $u(x,y) = \frac{x^2}{2} + \frac{1}{2}e^{2xy}$. The general solution is the family of level curves $u(x,y) = E$, where $E \in \mathbb{R}$.

Geometric picture of ODEs

Let $D \subset \mathbb{R}^2$ be the domain of the function $f(x, y)$. For each point $(x, y) \in D$ the DE

$$y' = f(x, y)$$

gives the value y' of the coefficient of the tangent to the solution $y(x)$ through this specific point, that is, the direction in which the solution passes through the point.

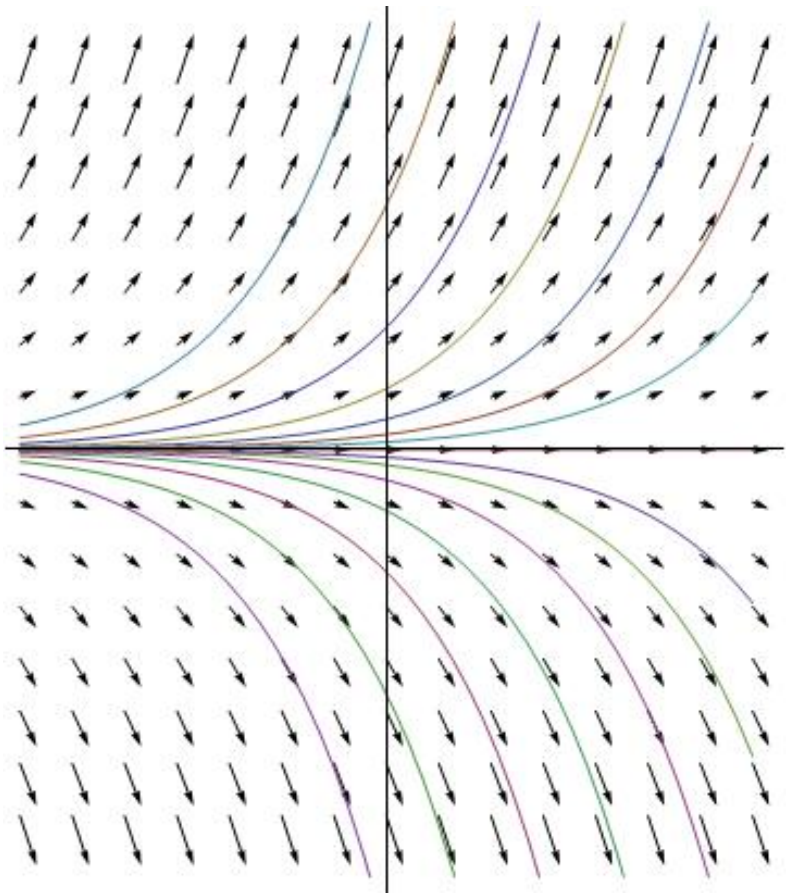
All these directions together form the directional field of the equation.

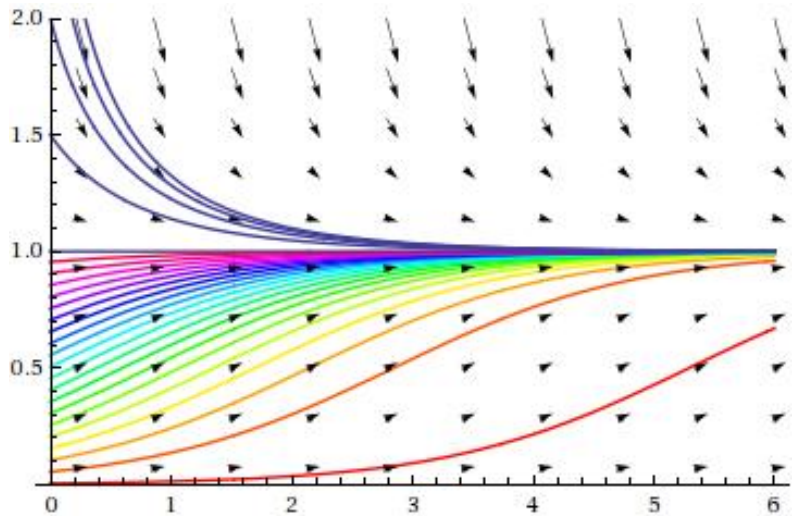
A solution of the equation is represented by a curve $y = y(x)$ that follows the given directions at every point x, i.e., the coefficient of the tangent corresponds to the value $f(x, y(x))$.

The general solution to the equation is a family of curves, such that each of them follows the given direction.

Directional fields and solutions of

$$y' = ky \qquad\qquad y' = ky(1-y)$$





Examples: https://zalara.github.io/Algoritmi/example_direction_fields.m

Theorem (Existence and uniqueness of solutions)

If $f(x, y)$ is continuous and differentiable with respect to y on the rectangle

$$D = [x_0 - a, x_0 + a] \times [y_0 - b, y_0 + b], \quad a, b > 0$$

then the DE with initial condition

$$y' = f(x, y), \quad y(x_0) = y_0,$$

has a unique solution $y(x)$ defined at least on the interval

$$[x_0 - \alpha, x_0 + \alpha], \quad \alpha = \min\left\{a, \frac{b}{M}, \frac{1}{N}\right\},$$

where

$$M = \max\{f(x, y) \colon (x, y) \in D\} \text{ and } N = \max\left\{\frac{\partial f(x, y)}{\partial y} \colon (x, y) \in D\right\}.$$

Proof: https://en.wikipedia.org/wiki/Picard%E2%80%93Lindel%C3%B6f_theorem.

Numerical methods for solving DE's

We are given the DE with the initial condition

$$y'(x) = f(y, x), \quad y(x_0) = y_0.$$

Instead of analytically finding the solution $y(x)$, we construct a recursive sequence of points

$$x_i = x_0 + ih, \quad y_i \doteq y(x_i), \quad i \geq 0$$

where y_i is an approximation to the value of the exact solution $y(x_i)$, and h is the step size.

A number of numerical methods exists, the choice depends on the type of equation, desired accuracy, computational time,...

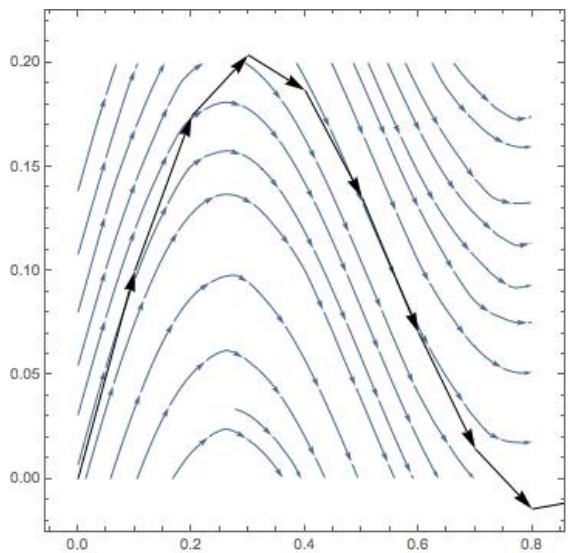
We will first look at the simplest and best known method and then a more practical improvement.

Euler's method

Euler's method is the simplest and most intuitive approach to numerically solve a DE.

At each step the value y_{i+1} is obtained as the point on the tangent to the solution through (x_i, y_i) at $x_{i+1} = x_i + h$:

- initial condition: (x_0, y_0)
- for each i: $x_{i+1} = x_i + h$, $y_{i+1} = y_i + hf(x_i, y_i)$.



The point (x_{i+1}, y_{i+1}) typically lies on a different particular solution than (x_i, y_i), at each step, the error at each step is of order $\mathcal{O}(h^2)$. The cumulative error is of order $\mathcal{O}(h)$.

Runge-Kutta methods

The idea of those methods is to approximate the derivative on the interval $[x_n, x_{n+1}]$ not only based on the derivative in the point x_n, but using a weighted average of more different derivatives on the interval $[x_n, x_{n+1}]$.

Example (Runge-Kutta of order 2 (RK2))

We approximate the derivate using the derivatives in the points x_n and $x_n + ch \in [x_n, x_{n+1}]$, where $h = x_{n+1} - x_n$ and $c \in [0, 1]$. The approximation y_{n+1} is computed using the weighted average of linear approximations in the points x_n and $x_n + ch$:

$$y_{n+1} = y_n + \underbrace{b_1}_{\text{weight}} \cdot \underbrace{(h \cdot f(x_n, y_n))}_{\substack{\text{move along} \\ \text{the tangent in } x_n}} + \underbrace{b_2}_{\text{weight}} \cdot \underbrace{(h \cdot f(x_n + ch, y(x_n + ch)))}_{\substack{\text{move along} \\ \text{the tangent in } x_n + ch}} \quad (57)$$

We use a linear approximation

$$y(x_n + ch) \approx y_n + chy'(x_n) = y_n + chf(x_n, y_n) \approx y_n + ahf(x_n, y_n), \quad (58)$$

where a is a new parameter.

Using (58) in (57) we obtain

$$y_{n+1} = y_n + b_1 \cdot \underbrace{(h \cdot f(x_n, y_n))}_{k_1} + b_2 \cdot \underbrace{(h \cdot f(x_n + ch, y_n + a \cdot k_1))}_{k_2}. \quad (59)$$

Using Taylor series' of $y(x_n + h)$, $f(x_n + ch, y_n + ak_1)$ and comparing the coefficients at h and h^2 in (59) we get a system of equations

$$\begin{aligned} 1 &= b_1 + b_2, \\ \frac{1}{2}(f_x + f_y f)_n &= b_2 c (f_x)_n + b_2 a (f f_y)_n, \end{aligned} \quad (60)$$

where f_n, $(f_x)_n$, $(f_y)_n$ stands for $f(x_n, y_n)$, $f_x(x_n, y_n)$, $f_y(x_n, y_n)$. The system (60) has many different solutions, e.g.:

- $b_1 = b_2 = \frac{1}{2}$ and $c = a = 1$. RK method is:

$$\begin{aligned} y_{n+1} &= y_n + \frac{1}{2}(k_1 + k_2), \\ k_1 &= hf(x_n, y_n), \\ k_2 &= hf(x_n + h, y_n + k_1). \end{aligned}$$

- $b_1 = 1$, $b_2 = 0$ in $c = a = \frac{1}{2}$. RK method is:

$$\begin{aligned} y_{n+1} &= y_n + k_2, \\ k_1 &= hf(x_n, y_n), \\ k_2 &= hf(x_n + \frac{1}{2}h, y_n + \frac{1}{2}k_1). \end{aligned}$$

A general RK method is of the form

$$\begin{aligned} y_{n+1} &= y_n + b_1 k_1 + b_2 k_2 + \ldots + b_s k_s, \\ k_1 &= hf(x_n, y_n), \\ k_2 &= hf(x_n + c_2 h, y_n + a_{2,1} k_1), \\ k_3 &= hf(x_n + c_3 h, y_n + a_{3,1} k_1 + a_{3,2} k_2), \\ &\vdots, \\ k_s &= hf(x_n + c_s h, y_n + a_{s,1} k_1 + \ldots + a_{s,s-1} k_{s-1}). \end{aligned} \tag{61}$$

Butcher tableau

In a compact form the RK method (61) is given in the form of a Butcher tableau:

$$
\begin{array}{c|cccccc}
0 & 0 & & & & & \\
c_2 & a_{2,1} & 0 & & & & \\
c_3 & a_{3,1} & a_{3,2} & 0 & & & \\
\vdots & \vdots & & & & & \\
c_s & a_{s_1} & a_{s_2} & a_{s,3} & \cdots & a_{s,s-1} & 0 \\
\hline
 & b_1 & b_2 & b_3 & \cdots & b_{s-1} & b_s,
\end{array}
$$

where

$$
\begin{aligned}
c_2 &= a_{2,1}, \\
c_3 &= a_{3,1} + a_{3,2}, \\
&\vdots \\
c_s &= a_{s,1} + a_{s,2} + \ldots + a_{s,s-1}.
\end{aligned}
$$

Runge-Kutta mehod of order 4

Butcher tableau:

$$\begin{array}{c|cccc} 0 & 0 & & & \\ \frac{1}{2} & \frac{1}{2} & 0 & & \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & \\ 1 & 0 & 0 & 1 & 0 \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array}$$

The method is

$$\begin{aligned} y_{n+1} &= y_n + \frac{1}{6}k_1 + \frac{1}{3}k_2 + \frac{1}{3}k_3 + \frac{1}{6}k_4, \\ k_1 &= hf(x_n, y_n), \\ k_2 &= hf(x_n + \frac{1}{2}h, y_n + \frac{1}{2}k_1), \\ k_3 &= hf(x_n + \frac{1}{2}h, y_n + \frac{1}{2}k_2), \\ k_4 &= hf(x_n + h, y_n + k_3). \end{aligned}$$

The error at each step is of order $\mathcal{O}(h^5)$. The cumulative error is of order $\mathcal{O}(h^4)$.

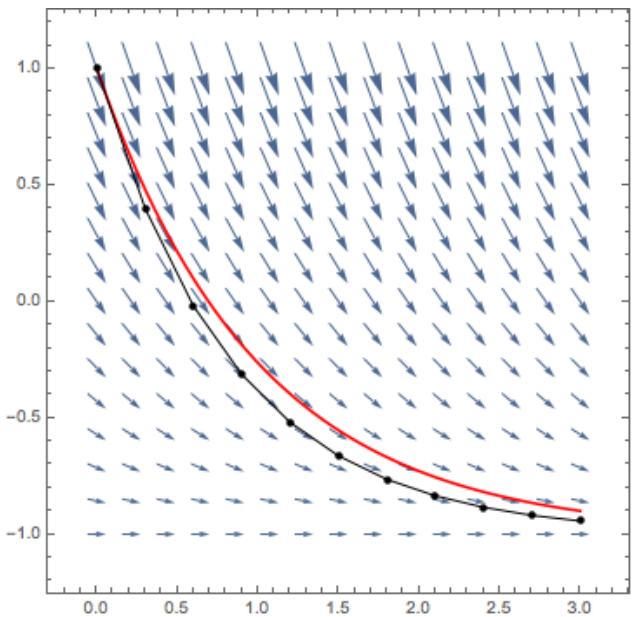
Euler vs RK4

Below is a comparison of Euler's and Rk4 methods for the DE

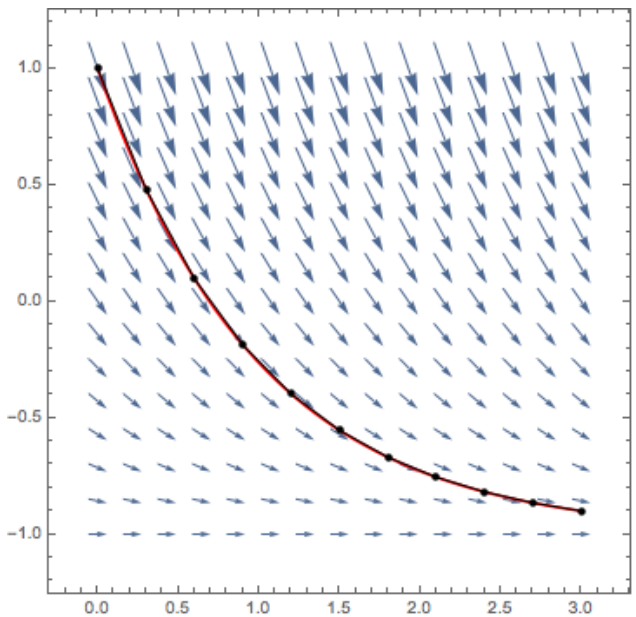
$$y' = -y - 1, \quad y(0) = 1 \quad \text{with step size } h = 0.3 :$$

The red curve is the exact solution $y = 2e^{-x} - 1$.

Euler's method



RK4



Algorithms and example:

https://zalara.github.io/Algoritmi/euler_eng.m

https://zalara.github.io/Algoritmi/RK4_eng.m

https://zalara.github.io/Algoritmi/Euler_vs_RK4.m

Adaptive Runge Kutta methods

Let M_1, M_2 be two RK methods with the same matrices of coefficients $a_{i,j}$ (and hence also c_i), but different vectors of weights b_i and b_i^*. Let M_1 be of order p (global error $\mathcal{O}(h^p)$), while the other of order $p+1$ (global error $\mathcal{O}(h^{p+1})$).

Example: We use the adaptive method for the Butcher tableaus:

$$\begin{array}{c|cc} 0 & 0 & \\ 1 & 1 & 0 \\ \hline & 1 & 0 \\ & \frac{1}{2} & \frac{1}{2} \end{array}.$$

The first is Euler's method and has order 1, while the other is RK method of order 2:

$$y_{n+1} = y_n + k_1,$$
$$y_{n+1}^* = y_n + \frac{1}{2}(k_1 + k_2).$$

The approximation of the local error:

$$\ell_{n+1} \approx y_{n+1}^* - y_{n+1} = (-k_1 + k_2)/2.$$

If ℓ_{n+1} is small enough (we choose what this means in our problem), we accept y_{n+1} and continue, otherwise we decrease the step size and repeat the computations.

DOPRI5, Fehlberg, Cash-Karp

Very useful methods for practical computations are DOPRI5 (1980, authors Dormand in Prince), Fehlberg (1969), Cash-Karp, which are adaptive methods combining two RK methods, one of order 4 and one of order 5.

- `https://en.wikipedia.org/wiki/Dormand%E2%80%93Prince_method`
- `https://en.wikipedia.org/wiki/Runge%E2%80%93Kutta%E2%80%93Fehlberg_method`
- `https://en.wikipedia.org/wiki/Cash%E2%80%93Karp_method`

Algorithm:

`https://zalara.github.io/Algoritmi/DOPRI5_eng.m`

`https://zalara.github.io/Algoritmi/DOPRI5_example.m`

Systems of first order ODE's

Let

$$f := (f_1, \ldots, f_n) : \mathbb{R}^{n+1} \to \mathbb{R}^n,$$

$$f(x_1, \ldots, x_{n+1}) = (f_1(x_1, \ldots, x_{n+1}), \ldots, f_n(x_1, \ldots, x_{n+1})).$$

be a vector function. A system of first order DE's is an equation

$$\dot{x}(t) = f(x(t), t), \tag{62}$$

where

$$x(t) := (x_1(t), \ldots, x_n(t)) : I \to \mathbb{R}^n$$

is an unknown vector function and $I \subset \mathbb{R}$ is some interval. Coordinate-wise the system (62) is equal to

$$\begin{aligned} \dot{x}_1(t) &= f_1(x_1(t), \ldots, x_n(t), t), \\ &\vdots \\ \dot{x}_n(t) &= f_n(x_1(t), \ldots, x_n(t), t). \end{aligned}$$

Solution of the system of DE's

For every $(x, t) \in \mathbb{R}^{n+1}$ in the domain of f, the value $f(x, t)$ is the tangent vector $\dot{x}(t)$ to the solution $x(t)$ at the given t.

The general solution is a family of parametric curves

$$x(t, C_1, \ldots, C_n),$$

where $C_1, C_2, \ldots, C_n \in \mathbb{R}$ are parameters, with the given tangent vectors.

An initial condition

$$x(t_0) = x_0 \in \mathbb{R}^n$$

gives a particular solution, that is, a specific parametric curve from the general solution that goes through the point x_0 at time t_0.

Linear systems of 1st order ODEs

A linear system of DEs is of the form

$$\begin{bmatrix} \dot{x}_1(t) \\ \vdots \\ \dot{x}_n(t) \end{bmatrix} = \begin{bmatrix} a_{11}(t) & \dots & a_{1n}(t) \\ \vdots & \ddots & \vdots \\ a_{n1}(t) & \dots & a_{nn}(t) \end{bmatrix} \begin{bmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{bmatrix} + \begin{bmatrix} g_1(t) \\ \vdots \\ g_n(t) \end{bmatrix}, \qquad (63)$$

where

$$x_i : I \to \mathbb{R}, \quad a_{ij} : I \to \mathbb{R} \quad \text{and} \quad g_i : I \to \mathbb{R}$$

are functions of t and $I \subseteq \mathbb{R}$ is an interval. In a compact form (63) can be written as

$$\dot{x}(t) = A(t)x + g(t), \qquad (64)$$

where

$$A(t) = [a_{ij}(t)]_{i,j=1}^{n}$$

is a $n \times n$ matrix function and

$$g(t) = \begin{bmatrix} g_1(t) & \dots & g_n(t) \end{bmatrix}^T$$

is a $n \times 1$ vector function.

The system (64)

- is homogeneous if for every t in the domain I we have $g(t) = \mathbf{0}$.
- has constant coefficients, if the matrix A is constant, i.e., independent of t.
- is autonomous, if it is homogeneous and has constant coefficients.

An autonomous linear system

$$\dot{x} = Ax \tag{65}$$

of 1st order DEs can be solved analytically, using methods from linear algebra. Recall that such a system can be written in coordinates as:

$$\begin{aligned}
\dot{x}_1 &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n, \\
\dot{x}_2 &= a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n, \\
&\vdots \\
\dot{x}_n &= a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n.
\end{aligned}$$

Autonomous system: diagonal matrix A

Assume first that the matrix A in (65) is diagonal. Then (65) is the following:

$$\begin{bmatrix} \dot{x}_1 \\ \vdots \\ \dot{x}_n \end{bmatrix} = \begin{bmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_n \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}.$$

Or equivalently,

$$\dot{x}_1 = \lambda_1 x_1, \qquad \dot{x}_2 = \lambda_2 x_2, \qquad \dots \qquad , \qquad \dot{x}_n = \lambda_n x_n.$$

In this (simple) case the general solution is easily determined:

$$x(t) = \begin{bmatrix} C_1 e^{\lambda_1 t} \\ C_2 e^{\lambda_2 t} \\ \vdots \\ C_n e^{\lambda_n t} \end{bmatrix} = C_1 e^{\lambda_1 t} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + C_2 e^{\lambda_2 t} \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} + \cdots + C_n e^{\lambda_n t} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}.$$

Autonomous system: n linearly independent eigenvectors

Assume next, that A in (65) has n linearly independent eigenvectors $v_1, \dots v_n$ with the corresponding eigenvalues $\lambda_1, \dots, \lambda_n$.

- For every fixed t, the vector $x(t)$ can be expressed as a linear combination
$$x(t) = \varphi_1(t)v_1 + \cdots + \varphi_n(t)v_n.$$
- Hence, the coefficients
$$\varphi_i(t) : I \to \mathbb{R}, \qquad i = 1, \dots, n,$$
are functions of t.
- Since $v_1, \dots v_n$ are eigenvectors it follows from $\dot{x} = Ax$, that
$$\sum_{i=1}^{n} \dot{\varphi}_i(t)v_i = \sum_{i=1}^{n} \varphi_i(t)Av_i = \sum_{i=1}^{n} \varphi_i(t)\lambda_i v_i.$$
- Since $v_1, \dots v_n$ are linearly independent, it follows that for every i we have
$$\dot{\varphi}_i(t) = \lambda_i\varphi_i(t) \quad \Rightarrow \quad \varphi_i(t) = C_i e^{\lambda_i t}, \quad C_i \in \mathbb{R}.$$
- Hence the general solution of the system is
$$x(t) = C_1 e^{\lambda_1 t}v_1 + \cdots + C_n e^{\lambda_n t}v_n.$$

Example

Find the general solution of the system

$$\begin{aligned} \dot{x}_1 &= x_1 + x_2, \\ \dot{x}_2 &= 4x_1 - 2x_2. \end{aligned}$$

The matrix of the system is $A = \begin{bmatrix} 1 & 1 \\ 4 & -2 \end{bmatrix}$. Its eigenvalues are the solutions of

$$\det(A - \lambda I) = (1 - \lambda)(-2 - \lambda) - 4 = \lambda^2 + \lambda - 6 = 0,$$

so $\lambda_1 = -3$ and $\lambda_2 = 2$, and the corresponding eigenvectors are

$$v_1 = \begin{bmatrix} 1 & -4 \end{bmatrix}^T \qquad \text{and} \qquad v_2 = \begin{bmatrix} 1 & 1 \end{bmatrix}^T.$$

The general solution of the system is

$$x(t) = C_1 e^{-3t} \begin{bmatrix} 1 \\ -4 \end{bmatrix} + C_2 e^{2t} \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Example

Find the general solution of

$$\dot{x}_1 = x_2,$$
$$\dot{x}_2 = -4x_1.$$

The matrix of the system is $A = \begin{bmatrix} 0 & 1 \\ -4 & 0 \end{bmatrix}$. It has a conjugate pair of complex eigenvalues and a corresponding conjugate pair of eigenvectors:

$$\lambda_{1,2} = \pm 2i, \quad v_{1,2} = \begin{bmatrix} 1 & \pm 2i \end{bmatrix}^T.$$

The general solution is a family of complex valued functions

$$x(t) = C_1 e^{2it} \begin{bmatrix} 1 \\ 2i \end{bmatrix} + C_2 e^{-2it} \begin{bmatrix} 1 \\ -2i \end{bmatrix}$$

(which is not very useful in modelling real-valued phenomena).

Autonomous system: complex conjugate eigenvalues

Assume that the matrix of the system A has a complex pair of eigenvalues $\lambda_{1,2} = \alpha \pm i\beta$ and corresponding eigenvectors $v_{1,2} = u \pm iw$.

The real and imaginary parts of the two complex valued solutions are:

$$\begin{aligned} & e^{(\alpha \pm i\beta)t}(u \pm iw) \\ = \; & e^{\alpha t}(\cos(\beta t) \pm i \sin(\beta t))(u \pm iw) \\ = \; & e^{\alpha t}\left[\cos(\beta t)u - \sin(\beta t)w \pm i(\sin(\beta t)u + \cos(\beta t)w)\right]. \end{aligned}$$

Any linear combination (with coefficients $C_1, C_2 \in \mathbb{R}$) of these is a real-valued solution, so the real-valued general solution is

$$x(t) = e^{\alpha t}\left[C_1(\cos(\beta t)u - \sin(\beta t)w) + C_2(\sin(\beta t)u + \cos(\beta t)w)\right].$$

Autonomous system: complex conjugate eigenvalues

Example

In the case of the previous example, $\lambda_{1,2} = \pm 2i$, i.e. $\alpha = 0$ and $\beta = 2$, and

$$v_{1,2} = \begin{bmatrix} 1 \\ \pm 2i \end{bmatrix} \quad \Rightarrow \quad u = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \text{and} \quad w = \begin{bmatrix} 0 \\ 2 \end{bmatrix}.$$

Hence, the general solution is

$$x(t) = C_1 \left(\cos(2t) \begin{bmatrix} 1 \\ 0 \end{bmatrix} - \sin(2t) \begin{bmatrix} 0 \\ 2 \end{bmatrix} \right) + C_2 \left(\sin(2t) \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \cos(2t) \begin{bmatrix} 0 \\ 2 \end{bmatrix} \right).$$

Autonomous system: less than n eigenvectors

If A has less than n linearly independent eigenvectors, additional solutions can also be obtained (e.g., with the use of Jordan form of A), but we will not consider this case here.

The general solution of a system $\dot{x} = Ax$ of n equations is of the form

$$x(t) = C_1 x^{(1)}(t) + \ldots + C_n x^{(n)}(t),$$

where $x^{(1)}(t), \ldots, x^{(n)}(t)$ are specific, linearly independent solutions.

For every eigenvalue $\lambda \in \mathbb{R}$ or a pair of eigenvalues $\lambda = \alpha \pm i\beta$ we obtain as many solutions as there are corresponding linearly independent eigenvectors.

Adding initial conditions to an autonomous system

An initial condition $x(t_0) = x^{(0)}$ gives a nonsingular system (if the vectors $x_1(t_0), \ldots, x_n(t_0)$ are linearly independent) of n linear equations for the constants $C_1, \ldots, C_n$.

$$x^{(0)} = C_1 x_1(t_0) + \ldots + C_n x_n(t_0).$$

This implies that a problem

$$\dot{x} = Ax, \quad x(t_0) = x^{(0)}$$

has a unique solution for any $x^{(0)}$.

Example

The initial condition $x^{(0)} = x(0) = \begin{bmatrix} 0 & 5 \end{bmatrix}^T$ for the system in the first example above gives the following system of equations for C_1 and C_2:

$$C_1 + C_2 = 0, \quad -4C_1 + C_2 = 5,$$

so $C_1 = -1$ and $C_2 = 1$.