Mathematical modelling

Neža Mramor Kosta, Polona Oblak, Žiga Virk, Aljaž Zalar

Faculty of Computer and Information Science
University of Ljubljana

2024/25

Chapter 0:

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.

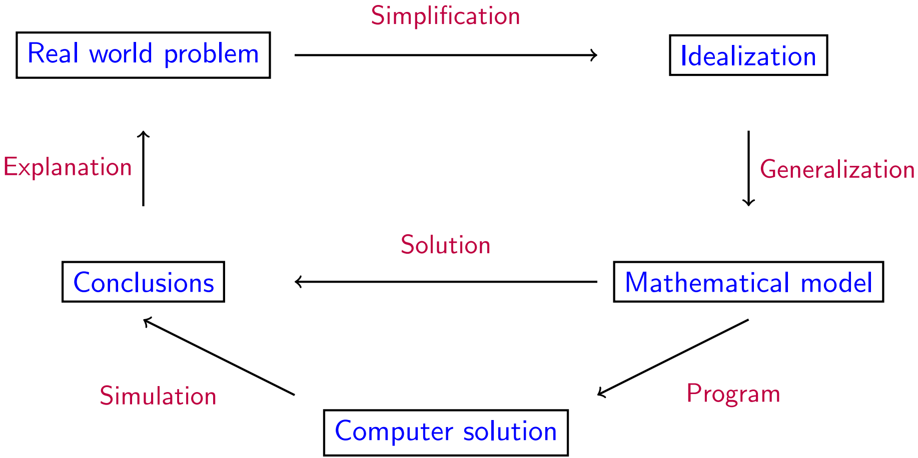
In this course we will introduce some (by far not all) mathematical tools that are used in setting up and solving mathematical models.

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 experimental 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 nonhomogeneous geometric object, angle and rotation in the initial thrust, . . .)
- Choice of the 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

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

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

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

Example: quadratic equation

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

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

$$\begin{aligned} x_2 = -a^2 + \sqrt{a^4 + q} &= \frac{(-a^2 + \sqrt{a^4 + q})(a^2 + \sqrt{a^4 + q})}{a^2 + \sqrt{a^4 + q}} \\ &= \frac{q}{a^2 + \sqrt{a^4 + q}}. \end{aligned}$$

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 1:

Linear model

- Definition
- Systems of linear equations
- The Moore-Penrose (MP) inverse
- Principal component analysis
- MP inverse and solving linear systems

1. Linear mathematical models

Given points

$$\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}, \quad \mathbf{x}_i \in \mathbb{R}^n,\ y_i \in \mathbb{R},$$

the task is to find a function $F(\mathbf{x}, a_1, \ldots, a_p)$ 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

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

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

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

is as small as possible.

The mathematical model is linear, when the function F is a linear function of the parameters $a_1 \ldots, a_p$:

$$F(\mathbf{x}, a_1, \ldots, a_p) = a_1\varphi_1(\mathbf{x}) + a_2\varphi_2(\mathbf{x}) + \cdots + a_p\varphi_p(\mathbf{x}),$$

where $\varphi_1, \varphi_2, \ldots \varphi_p \colon \mathbb{R}^n \to \mathbb{R}$ are functions of a specific type in variable $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$.

Examples of linear models:

1. linear regression: $x, y \in \mathbb{R}$, $\varphi_1(x) = 1, \varphi_2(x) = x$,
2. multivariate linear regression:

$$\varphi_1(\mathbf{x}) = 1, \varphi_2(\mathbf{x}) = x_1, \varphi_3(\mathbf{x}) = x_2, \ldots, \varphi_{n+1}(\mathbf{x}) = x_n,$$

3. polynomial regression: $x, y \in \mathbb{R}$, $\varphi_1(x) = 1, \ldots, \varphi_p(x) = x^{p-1}$,
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 $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$, $\mathbf{x}_i \in \mathbb{R}^n$, $y_i \in \mathbb{R}$, the parameters of a linear model

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

should satisfy the system of linear equations

$$y_i = a_1\varphi_1(\mathbf{x}_i) + a_2\varphi_2(\mathbf{x}_i) + \cdots + a_p\varphi_p(\mathbf{x}_i), \quad i = 1, \ldots, m,$$

In a matrix form, this is equivalent to

$$\begin{bmatrix} \varphi_1(\mathbf{x}_1) & \varphi_2(\mathbf{x}_1) & \ldots & \varphi_p(\mathbf{x}_1) \\ \varphi_1(\mathbf{x}_2) & \varphi_2(\mathbf{x}_2) & \ldots & \varphi_p(\mathbf{x}_2) \\ \ldots & \ldots & \ldots & \ldots \\ \varphi_1(\mathbf{x}_m) & \varphi_2(\mathbf{x}_m) & \ldots & \varphi_p(\mathbf{x}_m) \end{bmatrix} \begin{bmatrix} a_1 \\ a_1 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} y_1 \\ y_1 \\ \vdots \\ y_m \end{bmatrix},$$

or

$$\Phi\vec{a} = \vec{y},$$

where $\vec{a} = [a_1, \ldots, a_p]^{\mathsf{T}}$ is a vector of unknowns (i.e. parameters of our model).

1.1 Systems of linear equations and generalized inverses

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

$$A\vec{x} = \vec{b},$$

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,
- $\vec{x}$ is the vector of unknowns and
- $\vec{b}$ is the right side vector.

Existence of solutions:

Let $A = [\vec{a}_1, \ldots, \vec{a}_n]$, where $\vec{a}_i$ are vectors representing the columns of A.

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

$$A\vec{x} = \sum_i x_i \vec{a}_i.$$

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

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

$$[A \mid \vec{b}] = [\vec{a}_1, \ldots, \vec{a}_n \mid \vec{b}],$$

Theorem

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

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

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

A generic 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

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

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, or A is of full rank.
- $\det(A) \neq 0$.
- The null space $N(A) = \{\vec{x} : A\vec{x} = 0\}$ is trivial.
- All eigenvalues of A are nonzero.
- For each $\vec{b}$ the system of equations $A\vec{x} = \vec{b}$ 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...).

Theorem (Singular value decomposition - SVD)

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

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

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

Let

$$V = \begin{bmatrix} \vec{v}_1 & \vec{v}_2 & \cdots & \vec{v}_m \end{bmatrix} \quad \text{and} \quad U = \begin{bmatrix} \vec{u}_1 & \vec{u}_2 & \cdots & \vec{u}_n \end{bmatrix}$$

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

$$A^TA\vec{v}_i = V\Sigma^T\Sigma V^T\vec{v}_i = V\Sigma^T\Sigma e_i = \begin{cases} \sigma_i^2\vec{v}_i, & \text{if } i \le r, \\ 0, & \text{if } i > r, \end{cases}$$

$$AA^Tu_j = U\Sigma\Sigma^T U^T u_j = U\Sigma\Sigma^T f_j = \begin{cases} \sigma_i^2\vec{u}_j, & \text{if } j \le r, \\ 0, & \text{if } j > r. \end{cases}$$

Further on,

$$(AA^T)(A\vec{v}_i) = A(A^TA)\vec{v}_i = \begin{cases} \sigma_i^2 A\vec{v}_i, & \text{if } i \le r, \\ 0, & \text{if } i > r, \end{cases}$$

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

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.

- $A\vec{v}_i$ is an eigenvector of AA^T corresponding to σ_i^2 and so

$$\vec{u}_i = \frac{A\vec{v}_i}{\|A\vec{v}_i\|} = \frac{A\vec{v}_i}{\sigma_i}$$

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

$$\|A\vec{v}_i\| = \sqrt{(A\vec{v}_i)^T(A\vec{v}_i)} = \sqrt{\vec{v}_i^T A^T A\vec{v}_i} = \sqrt{\sigma_i^2 \vec{v}_i^T \vec{v}_i} = \sigma_i\|\vec{v}_i\| = \sigma_i.$$

- $A^T\vec{u}_j$ is an eigenvector of A^TA corresponding to σ_j^2 and so

$$\vec{v}_j = \frac{A^T\vec{u}_j}{\|A^T\vec{u}_j\|} = \frac{A^T\vec{u}_j}{\sigma_j}$$

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

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

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 $A \in \mathbb{R}^{m \times n}$ 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 $v = A^T u$ (resp. $u = Av$) 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. $A^T A$).

1.2 The Moore-Penrose generalized inverse

Definition

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

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

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

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 $A \in \mathbb{R}^{m\times n}$ is any matrix $A^+ \in \mathbb{R}^{n\times m}$ 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 $AA^+ \in \mathbb{R}^{m\times m}$ is symmetric: $(AA^+)^T = AA^+$.
4. The square matrix $A^+A \in \mathbb{R}^{n\times n}$ 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 $A \in \mathbb{R}^{m \times n}$:

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

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

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

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

Case 2: $AA^T \in \mathbb{R}^{m\times m}$ *is an invertible matrix. (In particular, $m \leq n$.)*

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

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

The MP inverse is

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

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)*

1. For $A^T A$ 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

$$\vec{v}_1, \ldots, \vec{v}_r, \vec{v}_{r+1}, \ldots, \vec{v}_m,$$

   and form the matrices

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

$$V_1 = \begin{bmatrix} \vec{v}_1 & \cdots & \vec{v}_r \end{bmatrix}, \quad V_2 = \begin{bmatrix} \vec{v}_{r+1} & \cdots & \vec{v}_m \end{bmatrix} \quad \text{and} \quad V = \begin{bmatrix} V_1 & V_2 \end{bmatrix}.$$

2. Let

$$\vec{u}_1 = \frac{A\vec{v}_1}{\sigma_1}, \quad \vec{u}_2 = \frac{A\vec{v}_2}{\sigma_2}, \quad \ldots \quad , \quad u_r = \frac{A\vec{v}_r}{\sigma_r},$$

   and $\vec{u}_{r+1}, \ldots, \vec{u}_n$ vectors, such that $\{u_1, \ldots, u_n\}$ is an orthonormal basis for $\mathbb{R}^n$. Form the matrices

$$U_1 = \begin{bmatrix} \vec{u}_1 & \cdots & \vec{u}_r \end{bmatrix}, \quad U_2 = \begin{bmatrix} \vec{u}_{r+1} & \cdots & \vec{u}_n \end{bmatrix} \quad \text{and} \quad U = \begin{bmatrix} U_1 & U_2 \end{bmatrix}.$$

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

$$\vec{v}_1, \ldots, \vec{v}_r,$$

   and form the matrices

$$S = \mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_r}) \in \mathbb{R}^{r \times r},$$
$$V_1 = \begin{bmatrix} \vec{v}_1 & \cdots & \vec{v}_r \end{bmatrix} \in \mathbb{R}^{m \times r}.$$

2. Put the vectors

$$u_1 = \frac{A\vec{v}_1}{\sigma_1}, \quad \vec{u}_2 = \frac{A\vec{v}_2}{\sigma_2}, \quad \ldots \quad , \quad u_r = \frac{A\vec{v}_r}{\sigma_r}$$

   in the matrix

$$U_1 = \begin{bmatrix} \vec{u}_1 & \cdots & \vec{u}_r \end{bmatrix}.$$

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

$$\vec{v}_1 = \frac{A^T u_1}{\sigma_1} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \end{bmatrix}^T, \qquad \vec{v}_2 = \frac{A^T u_2}{\sigma_2} = \begin{bmatrix} \frac{1}{3\sqrt{2}} & -\frac{1}{3\sqrt{2}} & \frac{4}{3\sqrt{2}} \end{bmatrix}^T.$$

$$\vec{v}_3 = \vec{v}_1 \times \vec{v}_2 = \begin{bmatrix} \frac{2}{\sqrt{3}} & -\frac{2}{3} & -\frac{1}{3} \end{bmatrix}^T.$$

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

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

1.3 The MP inverse and systems of linear equations

Let $A \in \mathbb{R}^{m\times n}$, where $m < n$. A system of equations $A\vec{x} = \vec{b}$ 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*

$$A\vec{x} = \vec{b} \tag{2}$$

   is solvable if and only if $AA^{+}\vec{b} = \vec{b}$.

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

$$\|A^{+}\vec{b}\| = \min\left\{\|\vec{x}\| \colon A\vec{x} = \vec{b}\right\}.$$

   Moreover, it is the unique solution of smallest norm.

Proof of Theorem.

1. ($\Rightarrow$) If $A\vec{x}_0 = \vec{b}$, then $\vec{b} = A\vec{x}_0 = AA^+A\vec{x}_0 = AA^\mathsf{T}\vec{b}$.
   ($\Leftarrow$) If $AA^+\vec{b} = \vec{b}$, then clearly $A\vec{x} = \vec{b}$ has solution $\vec{x} = A^+\vec{b}$.

2. Assume $\vec{x} = A^+\vec{b}$ is a solution of $A\vec{x} = \vec{b}$, i.e. $A\vec{x} = \vec{b}$ is solvable. First prove that
   $$\mathcal{S} = \{A^+\vec{b} + (A^+A - I)\vec{z}\colon\ \vec{z} \in \mathbb{R}^m\}$$
   is the set of all solutions of $A\vec{x} = \vec{b}$. Note that
   $(A^+A - I)\vec{z} = C(A^+A - I) = N(A)$.
   - If $\vec{x} \in \mathcal{S}$, then $\vec{x} = A^+\vec{b} + (A^+A - I)\vec{z}$ for some $\vec{z} \in \mathbb{R}^n$. Then
     $$\begin{aligned} A(\vec{x}) = A(A^+\vec{b} + (A^+A - I)\vec{z})) = \\ = AA^+\vec{b} + AA^+A\vec{z} - A\vec{z} = \vec{b}. \end{aligned}$$
   - Every $\vec{x}$ that solves $A\vec{x} = \vec{b}$ can be written as
     $$\vec{x} = \vec{x} - A^+A\vec{x} + A^+A\vec{x} = (A^+A - I)(-\vec{x}) + A^+\vec{b},$$
     and so $\vec{x} \in \mathcal{S}$.

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

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

Simplify:

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

and observe that

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

Thus, $\|A^+\vec{b} + (A^+A - I)\vec{z}\|^2 = \|A^+\vec{b}\|^2 + \|(A^+A - I)\vec{z}\|^2 \geq \|A^+\vec{b}\|^2$, with the equality iff $(A^+A - I)\vec{z} = \vec{0}$. □

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, $A^+\vec{b} = A^+1$ 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, $A^+\vec{b} = A^+5$ 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, $A^+\vec{b}$ 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,
$$\vec{v}_1 = \frac{A^T \vec{u}}{\|A^T \vec{u}\|} = \frac{A^T \vec{u}}{\sigma_1} = \frac{1}{\sqrt{11}} \begin{bmatrix} 3 & 1 & 1 \end{bmatrix}^T.$$
-
$$A^+ = V\Sigma^+ U^T = \frac{1}{\sqrt{11}} \begin{bmatrix} 3 \\ 1 \\ 1 \end{bmatrix} \frac{1}{\sqrt{11}} [1] = \begin{bmatrix} \frac{3}{11} \\ \frac{1}{11} \\ \frac{1}{11} \end{bmatrix}.$$
-
$$A^+ b = \begin{bmatrix} \frac{6}{11} & \frac{2}{11} & \frac{2}{11} \end{bmatrix}^T.$$

Overdetermined systems

Let $A \in \mathbb{R}^{m \times n}$, where $m \geq n$. 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 $A\vec{x} = \vec{b}$ has no solutions, then the best fit for the solution is a vector $\vec{x}$ such that the error $||A\vec{x} - \vec{b}||$ or, equivalently in the row decomposition

$$A = \begin{bmatrix} \alpha_{\mathbf{1}} \\ \vdots \\ \alpha_{\mathbf{m}} \end{bmatrix},$$

its square

$$||A\vec{x} - \vec{b}||^2 = \sum_{i=1}^{m} (\alpha_i x - b_i)^2,$$

is the smallest possible.

Theorem

If the system $A\vec{x} = \vec{b}$ has no solutions, then

$$\vec{x}^{+} = A^{+}\vec{b}$$

is the solution to the least squares approximation problem:

$$\min\{\|A\vec{x} - \vec{b}\| : \vec{x} \in \mathbb{R}^m\}. \tag{3}$$

Moreover, if $\operatorname{rank} A = n$*, then* (3) *has a unique solution. If* $\operatorname{rank} A < n$*, then* $\vec{x}^{+}$ *has the smallest second norm* $\|x^{+}\|_2$ *among all solution to* (3)*.*

Proof.

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

$$\|A\vec{x} - \vec{b}\| = \|U\Sigma V^T\vec{x} - \vec{b}\| = \|\Sigma V^T\vec{x} - U^T\vec{b}\|,$$

where we used that

$$\|U^T\vec{v}\| = \|\vec{v}\|$$

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,

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

But this norm is minimal iff

$$SV_1^T\vec{x} - U_1^T\vec{b} = 0$$

or equivalently

$$V_1^T\vec{x} = S^{-1}U_1^T\vec{b}. \tag{4}$$

Further on,

$$V^TV = \begin{bmatrix} V_1^TV_1 & V_1^TV_2 \\ V_2^TV_1 & V_2^TV_2 \end{bmatrix} = I_n,$$

implies that $V_1^TV_1 = I_r$ and $V_2^TV_1 = 0$, where I_k stands for the $k \times k$ identity matrix.

If $\operatorname{rank} A = m$, then $V_1 \in \mathbb{R}^{m \times m}$ is invertible with the inverse V_1^T and hence,

$$V_1 S^{-1} U_1^T \vec{b} = A^+ \vec{b}$$

is the unique solution to (3).

If $r = \operatorname{rank} A < m$, then all x which solve (4) are of the form $A^+ \vec{b} + \vec{z}$, for $\vec{z} \in \ker V_1^T$. Since $\ker V_1^T = \operatorname{im} V_2$ and $V_2^T V_1 = 0$, it follows that the norm of $A^+ \vec{b} + \vec{z}$ is minimal for $\vec{z} = \vec{0}$. □

Remark

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

Example

Given points $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ 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

$$\begin{bmatrix} \mathbf{x}_1 & 1 \\ \vdots & \\ \mathbf{x}_n & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}.$$

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.

Vector norm is a map $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$, which satisfies:

1. Positive definiteness: $\|\vec{x}\| \geq 0$ for every $\vec{x} \in \mathbb{R}^n$ and $\|\vec{x}\| = 0 \Leftrightarrow \vec{x} = 0$.
2. Homogeneity: $\|\alpha\vec{x}\| = |\alpha| \, \|\vec{x}\|$ for every $\alpha \in \mathbb{R}$ and $\vec{x} \in \mathbb{R}^n$
3. Triangle inequality: $\|\vec{x} + y\| \leq \|\vec{x}\| + \|\vec{y}\|$ for every $x, y \in \mathbb{R}^n$.

Example

Let $\vec{x} = [x_1, \ldots, x_n]^\mathsf{T} \in \mathbb{R}^n$.

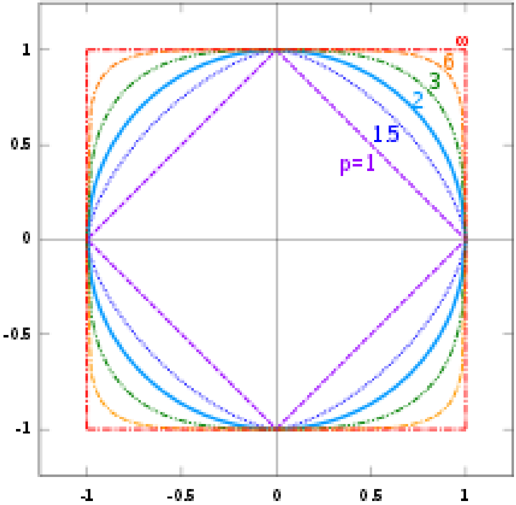
- p–norm for $p \in \mathbb{N}$:

$$\|\vec{x}\|_p := \left(|x_1|^p + \ldots + |x_n|^p\right)^{1/p}.$$

- Supremum norm:

$$\|\vec{x}\|_\infty = \max(|x_1|, \ldots, |x_n|).$$

Unit spheres in various norms



Matrix norm is a map $\|\cdot\| : \mathbb{R}^{n\times n} \to \mathbb{R}$, which satisfies

1. Positive definiteness: $\|A\| \geq 0$ for all $A \in \mathbb{R}^{n\times n}$ and $\|A\| = 0 \Leftrightarrow A = 0$.
2. Homogeneity: $\|\alpha A\| = |\alpha| \, \|A\|$ for all $\alpha \in \mathbb{R}$ in $A \in \mathbb{R}^{n\times n}$.
3. Triangle inequality: $\|A + B\| \leq \|A\| + \|B\|$ for all $A, B \in \mathbb{R}^{n\times n}$.
4. Submultiplicativity: $\|AB\| \leq \|A\| \, \|B\|$ for all $A, B \in \mathbb{R}^{n\times n}$.

Proposition

Let $\|\cdot\|_*$ be a vector norm on $\mathbb{R}^n$. Then

$$\|A\|_* := \max_{\|\vec{x}\|=1} \|A\vec{x}\|_* = \max_{x\neq 0} \frac{\|A\vec{x}\|_*}{\|\vec{x}\|_*}.$$

defines a matrix norm on $\mathcal{C}^{n\times n}$.

Proof: click

Let $A = [a_{ij}]_{i,j=1,\ldots,n} \in \mathbb{R}^{n\times n}$ be a matrix. Some matrix norms are the following:

1. 1–norm:

$$\|A\|_1 = \max_{j=1,\ldots,n} \Big(\sum_{i=1}^{n} |a_{ij}| \Big).$$

Proof: klik

2. Spectral norm: Here $\lambda_j(X)$ stands for the j–th eigenvalue of X.

$$\|A\|_2 = \sqrt{\max_{j=1,\ldots,n} \lambda_j(A^T A)}.$$

3. Frobenius norm:

$$\|A\|_F = \sqrt{\sum_{i,j=1}^{n} |a_{ij}|^2}.$$

4. Supremum norm:

$$\|A\|_\infty = \max_{i=1,\ldots,n} \Big(\sum_{j=1}^{n} |a_{ij}| \Big).$$

1.4 Principal component analysis (PCA)

- SVD is an essential tool for the PCA, which 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), ...

- Suppose we are given n data points $X_1, \ldots, X_n \in \mathbb{R}^d$, viewed as rows of a $n \times d$ matrix X. Each entry $x_{i,j}$ of

  $$X_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,d})$$

  represents the value of some *feature* of X_i, i.e., if X_i represents a person, then $x_{i,j}$'s can represent his/her year of birth, the height, blood sugar level, blood presure, etc. The columns $C_1, \ldots, C_d$ of X are also called **feature vectors**.

- **Basic idea of PCA:** Determine the vectors $Y^{(1)}, \ldots, Y^{(d)} \in \mathbb{R}^n$, called **principal components (PCs)**, which are *uncorrelated* projections of centered data points $X_1, \ldots, X_n$ onto some unit vectors $v^{(1)}, \ldots, v^{(d)} \in \mathbb{R}^d$ such that the variances $\mathrm{Var}(Y^{(1)}), \ldots, \mathrm{Var}(Y^{(d)})$ are maximized.

- **Algorithm for the computation of PCs of X:**

  1. **Centralization of data:**
     For each column C_j compute its mean value

     $$\mu_j := \frac{1}{n}\sum_{i=1}^{n} x_{i,j} = \frac{1}{n}(x_{1,j} + x_{2,j} + \ldots + x_{n,j})$$

     and subtract the **centroid**

     $$\mu := (\mu_1, \mu_2, \ldots, \mu_d)$$

     from each row of X:

     $$X - \mathbf{1}_{n,d}\,\mathrm{diag}(\mu) = [x_{i,j} - \mu_j]_{i,j},$$

     where $\mathbf{1}_{n,d}$ stands for the $n \times d$ matrix with all entries equal to 1 and $\mathrm{diag}(\mu)$ is a diagonal matrix with j-th diagonal entry μ_j.

2. **Computation of the singular value decomposition (SVD) of $X - \mathbf{1}_{n,d}\,\mathrm{diag}(\mu)$:**
   Let
   $$X - \mathbf{1}_{n,d}\,\mathrm{diag}(\mu) = UDV^T$$
   be the SVD of $X - \mathbf{1}_{n,d}\,\mathrm{diag}(\mu)$, where $U \in \mathbb{R}^{n\times n}$, $V \in \mathbb{R}^{d\times d}$ are orthogonal matrices and $D \in \mathbb{R}^{n\times d}$ is a diagonal matrix with the singular values
   $$\sigma_1 \geq \sigma_2 \geq \ldots \sigma_d \geq 0$$
   in decreasing order on the main diagonal.
3. **Computation of the PCs of X:**
   The PCs of X are points $Y^{(1)}, \ldots, Y^{(d)} \in \mathbb{R}^n$ obtained by
   $$Y^{(k)} = (X - \mathbf{1}_{n,d}\,\mathrm{diag}(\mu))v^{(k)} = \sigma_k u^{(k)}, \qquad k = 1, \ldots, d,$$
   where $v^{(k)}$ and $u^{(k)}$ are the k–th columns of V and U, respectively. The vectors $v^{(k)}$ and $u^{(k)}$ are called **right** (resp. **left**) **principal directions**.

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 $\vec{v}_1, \ldots, \vec{v}_k$ in $\mathbb{R}^n$.

The idea is to approximate

$$X_i = \sigma_1 u_{1,i} \vec{v}_1 + \cdots + \sigma_m u_{m,i} \vec{v}_m \cong \sigma_1 u_{1,i} \vec{v}_1 + \cdots + \sigma_k u_{k,i} \vec{v}_k$$

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

PCA has the following amazing property:

Theorem (Eckart-Young)

Among all possible projections of $p \colon \mathbb{R}^n \to \mathbb{R}^k$ 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 2:

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 $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$, $\mathbf{x}_i \in \mathbb{R}^n$, $y_i \in \mathbb{R}$.

The mathematical model is nonlinear if the function

$$y = F(\mathbf{x}, a_1, \ldots, a_p) \tag{5}$$

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

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

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

Plugging each data points into (5) we obtain a system of nonlinear equations

$$\begin{aligned} y_1 &= F(\mathbf{x}_1, a_1, \ldots, a_p), \\ &\vdots \\ y_m &= F(\mathbf{x}_m, a_1, \ldots, a_p), \end{aligned} \tag{6}$$

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

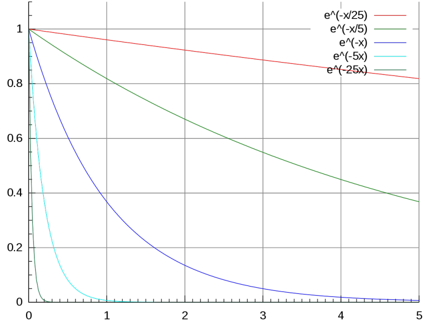
Examples

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 $y = F(x, a, k)$.



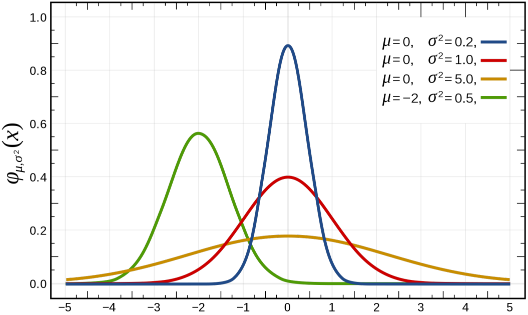
Examples

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.

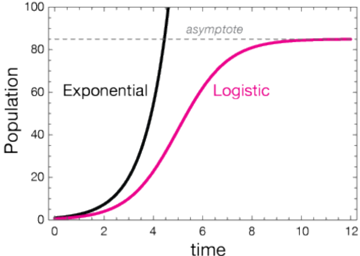


3. Logistic model: $F(x, a, b, k) = \frac{a}{(1+be^{-kx})}, k > 0$

   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

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



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

$$\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}, \quad \mathbf{x}_i \in \mathbb{R}^n, \ y_i \in \mathbb{R}.$$

- We choose a fitting function

$$F(\mathbf{x}, a_1, \ldots, a_p)$$

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

- Equivalent formulation of the system (6) (which will be more suitable for solving with numerical algorithms) is:
  1. For $i = 1, \ldots, m$ define the functions

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

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

$$\begin{aligned} g_1(a_1, \ldots, a_p) &= 0, \\ &\vdots \\ g_m(a_1, \ldots, a_p) &= 0. \end{aligned} \tag{7}$$

In a compact way (7) can be expressed by introducing a vector function

$$G : \mathbb{R}^p \to \mathbb{R}^m, \quad G(a_1, \dots, a_p) = (g_1(a_1, \dots, a_p), \dots, g_m(a_1 \dots, a_p)), \tag{8}$$

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

$$G(a_1, \dots, a_p) = (0, \dots, 0). \tag{9}$$

Remark

Solving (9) *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 (9).

3.1 Vector functions of a vector variable

Neccessary terminology to achieve our plan

G from (8) 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 (8) 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:

$$g_i(x_1, \ldots, x_n) = a_{i1} \cdot x_1 + a_{i2} \cdot x_2 + \ldots + a_{in} \cdot x_n, \quad \text{where } a_{ij} \in \mathbb{R}. \quad (10)$$

   In this case

$$G(\vec{x}) = A\vec{x},$$

   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 (10) we get the definition of an affine linear vector function,

$$g_i(1, \ldots, n) = a_{i1}x_1 + a_{i2}x_2 + \ldots a_{in}x_n + b_i,$$

   and then

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

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

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

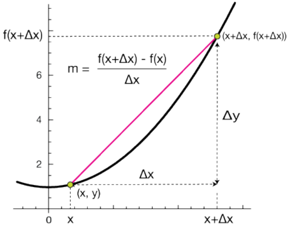
The derivative of a vector function $F : \mathbb{R}^n \to \mathbb{R}^m$ in the point

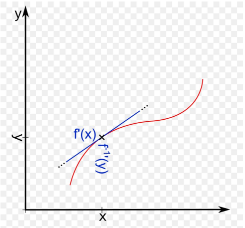
$$\mathbf{a} := (a_1, \ldots, a_n) \in \mathbb{R}^n$$

is called the Jacobian matrix of F in $\vec{a}$:

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

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

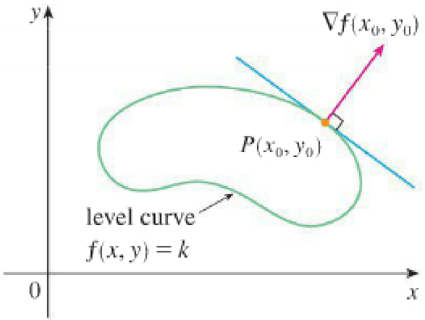




- For general n and $m = 1$, f is a function of n variables and

$$J_f(\vec{x}) = (\text{grad } f(\vec{x}))^{\mathsf{T}}$$

is its gradient.



Proof: Click

1. For an affine linear function $f : \mathbb{R}^n \to \mathbb{R}^m$, given by $f(\vec{x}) = A\vec{x} + \vec{b}$, it is easy to check that
$$J_f(\vec{x}) = A.$$

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
$$J_f(x, y, x) = \begin{bmatrix} 2x & 2y & 2z \\ 1 & 1 & 1 \end{bmatrix}.$$

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

$$L_{\vec{a}} : \mathbb{R}^n \to \mathbb{R}^m, \quad L_{\vec{a}}(\vec{x}) = A\vec{x} + \vec{b}$$

that satisfies the following conditions:

1. It has the **same value** as f in $\vec{a}$: $L_{\vec{a}}(\vec{a}) = f(\vec{a})$.
2. It has the **same derivative** as f at $\vec{a}$: $J_{L_{\vec{a}}}(\vec{a}) = J_f(\vec{a})$.

It is easy to check that

$$L_a(x) = f(\vec{a}) + J_f(\vec{a})(\vec{x} - \vec{a})^{\mathsf{T}}.$$

- $n = m = 1$:

$$L_{\vec{a}}(x) = f(\vec{a}) + f'(\vec{a})(\vec{x} - \vec{a})$$

The graph $y = L_{\vec{a}}(x)$ is the tangent to the graph $y = f(x)$ at the point a.

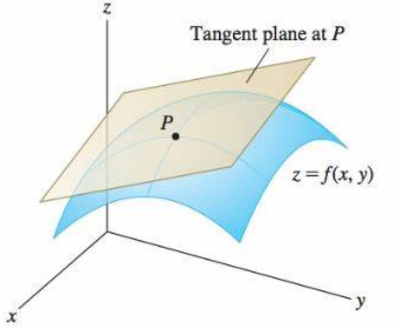
- If $n = 2$ and $m = 1$, then

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

The graph

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

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



Proof: Click

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

$$L_a(x, y, z) = f(1, -1, 1) + Df(1, -1, 1) \begin{bmatrix} x - 1 \\ y + 1 \\ z - 1 \end{bmatrix}$$

$$= \begin{bmatrix} 2 \\ 1 \end{bmatrix} + \begin{bmatrix} 2 & -2 & 2 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} x - 1 \\ y + 1 \\ z - 1 \end{bmatrix}$$

$$= \begin{bmatrix} 2 + 2(x - 1) - 2(y + 1) + 2(z - 1) \\ 1 + (x - 1) + (y + 1) + (z - 2) \end{bmatrix}$$

$$= \begin{bmatrix} 2 & -2 & 2 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} + \begin{bmatrix} -4 \\ 0 \end{bmatrix}.$$

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.

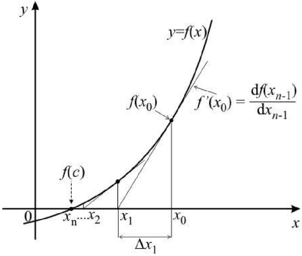
We are searching zeroes of the function $f : D \to \mathbb{R}$, $D \subseteq \mathbb{R}$, i.e., we are solving $f(x) = 0$.

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

$$\begin{aligned} 0 = f(\alpha) &= f(x_i) + f'(x_i)(\alpha - x_i) + \frac{f''(\eta)}{2}(\alpha - x_i)^2 \\ &= f(x_i) + f'(x_i)(\alpha - x_i) + \frac{f''(\eta)}{2}(\alpha - x_i)^2 \end{aligned} \tag{11}$$

where η is between α and x_i. Dividing (11) 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

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

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

$$M\delta_0 < 1.$$

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 for $n = m > 1$

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

- $\vec{x}_0$ – initial approximation,
- $\vec{x}_{k+1}$ – solution of

$$L_{\vec{x}_k}(\vec{x}) = f(\vec{x}_k) + Df(\vec{x}_k)(\vec{x} - \vec{x}_k) = 0,$$

  so

$$\vec{x}_{k+1} = \vec{x}_k - Df(\vec{x}_k)^{-1} f(\vec{x}_k).$$

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

$$Df(\vec{x}_k)\Delta\vec{x}_k = -f(\vec{x}_k)$$

is solved at each step (using LU decomposition of $Df(\vec{x}_k)$) and hence

$$\vec{x}_{k+1} = \vec{x}_k + \Delta\vec{x}_k.$$

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

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

Using Newton's metod we:

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

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 $F(\vec{x}) = \vec{0}$, where

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

is equivalent to finding global minima of

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

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

1. Choose $\vec{x}_0$.
2. Determine the constant α in $\vec{x}_r - \alpha \cdot (\operatorname{grad} g)(\vec{x}_r))$ which mimimises

$$h(\alpha) = g(\vec{x}_r - \alpha \cdot (\operatorname{grad} g)(\vec{x}_r)).$$

   (Or is significantly smaller than $h(0) = g(\vec{x}_r)$.)
3. $\vec{x}_{r+1} = \vec{x}_r - \alpha \cdot (\operatorname{grad} g)(\vec{x}_r)$.

Chapter 3:

Curves

- Definition and examples
- Derivative
- Arc length
- Plotting plane curves
- Area bounded by plane curves

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 $\mathbb{R}^m$.

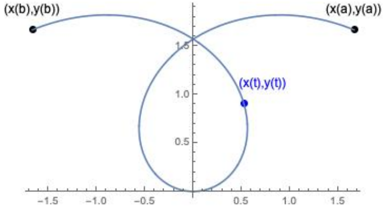
As t runs through I, $f(t)$ traces a path, or a curve, in $\mathbb{R}^m$.

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:

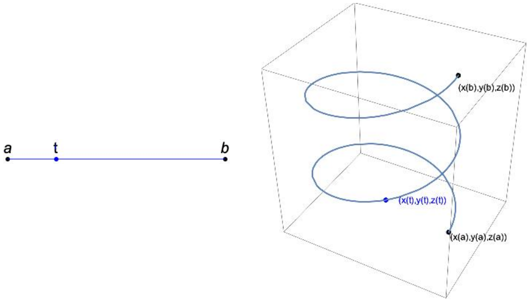


In this example $x(t) = t\cos t, y(t) = t\sin t, t \in [-3\pi/4, 3\pi/4]$

If $m = 3$, then

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

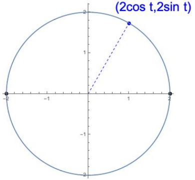
is the position vector of a point in $\mathbb{R}^3$ for every t, and $\{f(t), t \in I\}$ is a space curve:



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

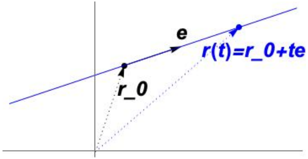
Example

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



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

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

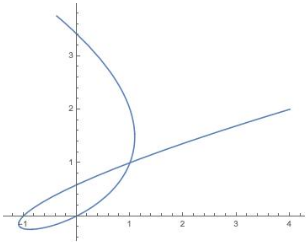


line through $\mathbf{r}_0$ in the direction of $\mathbf{e}$ in $\mathbb{R}^m$

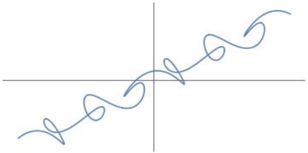
m=2:
slope $k = e_2/e_1$ if $e_1 \neq 0$
vertical if $\mathbf{e} = (0, e_2)$
horizontal if $\mathbf{e} = (e_1, 0)$

Example

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



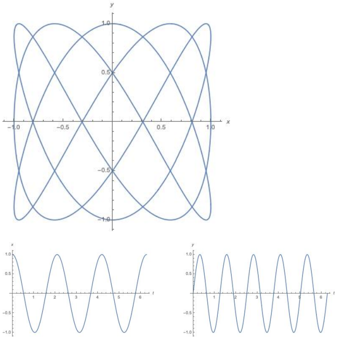
$$f(t) = \begin{bmatrix} t + \sin(3t) \\ t + \cos(5t) \end{bmatrix}, t \in \mathbb{R}$$



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

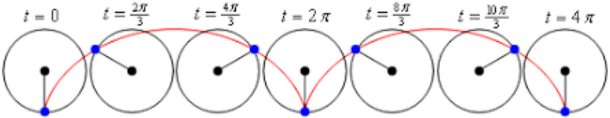
Example

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

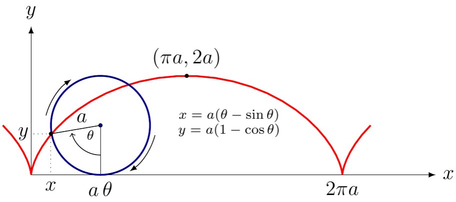


Problem: What path does the valve on your bicycle wheel trace as you bike along a straight road?

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:

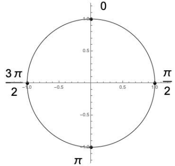


The curve is a cycloid: $x(\theta) = a\theta - a\sin\theta, y(\theta) = a - a\cos\theta$.

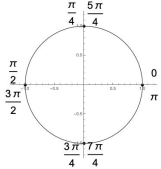


The following parametric curves all describe the circle with radius a around the origin (as well as many others):

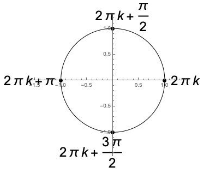
$$f_1(t) = \begin{bmatrix} a \sin t \\ a \cos t \end{bmatrix},\ t \in [0, 2\pi]$$



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



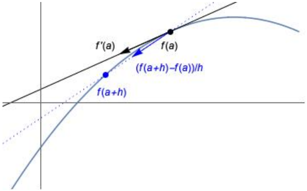
$$f_3(t) = \begin{bmatrix} a \cos t \\ a \sin t \end{bmatrix},\ t \in \mathbb{R}$$



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:

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

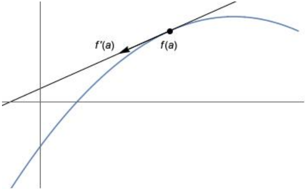


The vector $f'(a)$ (if it exists) represents the velocity vector of a point moving along the curve at the point $t = a$.

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

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

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



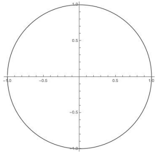
- If $f'(a) \neq \vec{0}$, this is a parametric line corresponding to the tangent line to the curve $f(t)$ at $t = a$. In this case $f(a)$ is a regular point of the parametrization.

- If $f'(a) = \vec{0}$ (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 $f'(a) \neq \vec{0}$.

- 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?



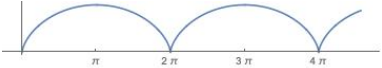
Since $x^2 + y^2 = 1$, $f(t)$ is a parametrization of the unit circle which is a smooth curve (it has a tangent at every point).

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.

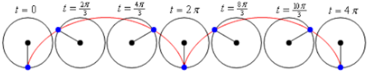
(l'Hospital's rule was used to compute the limits.)

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

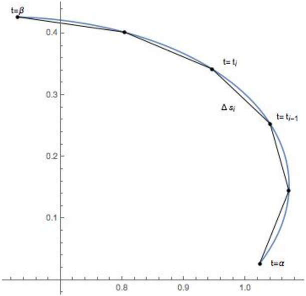
The arc length s of a parametric curve $f(t)$ between the points $t = a$ and $t = b$ is given by

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

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 $\begin{bmatrix} a\cos t \\ b\sin t \end{bmatrix}$, $a \neq b$?

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

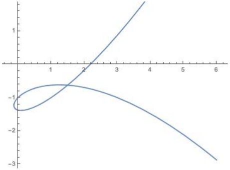
where $e = \sqrt{1-(b/a)^2}$ is its eccentricity and the function E is the nonelementary elliptic integral of 2nd kind. It can be computed numerically, which is briefly explained in the next few slides.

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

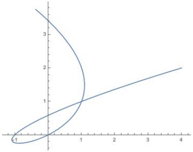


Here is a general strategy:

- 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?

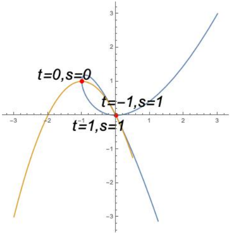
Let $f_1(t) = \begin{bmatrix} t^2 - 1 \\ -t^3 - t^2 + t + 1 \end{bmatrix}, \quad f_2(s) = \begin{bmatrix} s - 1 \\ 1 - s^2 \end{bmatrix}$.

To find the intersections, solve the system

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

There are three solutions:

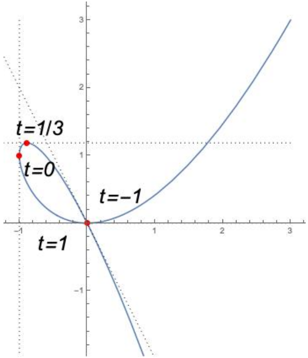
$$\begin{aligned} t = -1, s = 1 \quad &\Rightarrow \quad x = 0, y = 0 \\ t = 0,\ s = 0 \quad &\Rightarrow \quad x = -1,\ y = 1 \\ t = 1,\ s = 1 \quad &\Rightarrow \quad x = 0,\ y = 0 \end{aligned}$$



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

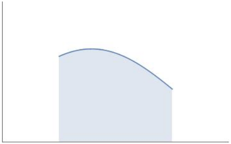
- Asymptotic behaviour: $\lim_{t\to\infty} f(t) = \begin{bmatrix} \infty \\ -\infty \end{bmatrix}$, $\lim_{t\to-\infty} f(t) = \begin{bmatrix} \infty \\ \infty \end{bmatrix}$,
  - intersections with axes:$t = \pm 1$, at $(0,0)$
    this is also a self-intersection
  - the two tangent lines at $(0,0)$
    - 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

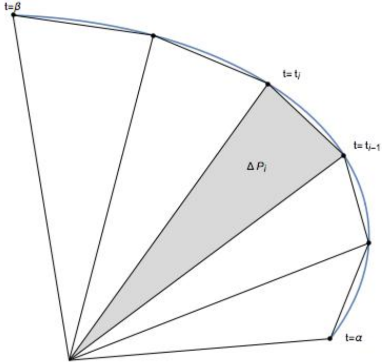
$$P = \int_{x(a)}^{x(b)} |y(x)|\, dx = \int_a^b |y(t)| x'(t)\, dt$$

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

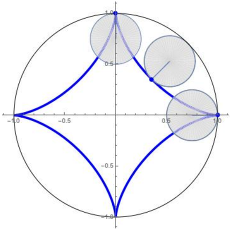
$$\begin{aligned} f'(t_i) \times f(t_i) &= (x'(t_i), y'(t_i), 0) \times (x(t_i), y(t_i), 0) \\ &= (x'(t_i)y(t_i) - y'(t_i)x(t_i), 0, 0). \end{aligned}$$

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

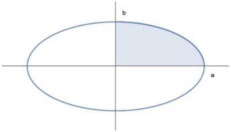
$$\begin{aligned} P &= \lim_{n\to\infty} \frac{1}{2} \sum_{i=0}^{n-1} |y'(t_i)x(t_i) - x'(t_i)y(t_i)|\Delta t \\ &= \frac{1}{2} \int_a^b |x(t)y'(t) - y(t)x'(t)|\, dt. \end{aligned}$$

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.

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:

$$\dot{T} = k(T - T_\infty), \tag{12}$$

where $T(t)$ is the temperature of a homogeneous body 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).

(12) 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

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

which only makes sense if $k < 0$.

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

$$\dot{x} = f(t)g(x). \tag{13}$$

This can be solved by:

- Inserting $\dot{x} = \dfrac{dx}{dt}$ into (13):

$$\frac{dx}{dt} = f(t)g(x). \tag{14}$$

- Separating variables in (14):

$$\frac{dx}{g(x)} = f(t)\,dt. \tag{15}$$

- Integrating both sides of (14):

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

Example 1 of a separable DE

$$\boxed{\dot{x} = kx} \quad \text{where } k \in \mathbb{R} \text{ is a fixed real number} \tag{16}$$

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

where C is a constant and so

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

is a general solution to (16). 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

$$\boxed{\dot{x} = kx(1-x)} \quad \text{where } k \in \mathbb{R} \text{ is a fixed real number} \tag{17}$$

- $$\frac{dx}{dt} = kx(1-x),$$
- $$\frac{dx}{x(1-x)} = kdt,$$
- 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

$$x(t) = \frac{1}{Ce^{-kt}+1} \tag{18}$$

is a general solution to (17). $x(t)$ from (18) is called a logistic function.

Example 3 of a separable DE

$$\boxed{y' = \frac{-x}{ye^{x^2}},} \quad y(0) = 1. \tag{19}$$

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

$$\dot{x} + f(t)x = g(t) \tag{20}$$

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

A homogeneous part of (20),

$$\dot{x} + f(t)x = 0, \tag{21}$$

has a general solution of the form

$$Cx_h(t), \tag{22}$$

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

- Every $x(t)$ of the form (22) is a solution of (21):

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

- If $x(t)$ is a solution of (21), then it must be of the form (22). Indeed, since $x(t)$ and $x_h(t)$ both solve (21),

$$\begin{aligned}\left(\frac{x(t)}{x_h(t)}\right)' &= \frac{x'(t)x_h(t) - x(t)x_h'(t)}{x_h^2(t)} \\ &= \frac{-f(t)x(t)x_h(t) + f(t)x(t)x_h(t)}{x_h^2(t)} \\ &= 0.\end{aligned}$$

Hence, $\frac{x(t)}{x_h(t)} = C$ for some constant C and $x(t)$ is of the form (22).

Let $x_p(t)$ be any particular solution of (20):

$$x_p'(t) + f(t)x_p(t) = g(t). \tag{23}$$

The general solution of (20) is a sum

$$x(t) = Cx_h(t) + x_p(t). \tag{24}$$

Indeed:

- Every $x(t)$ of the form (24) is a solution of (20):

$$
\begin{aligned}
x'(t) + f(t)x(t) &= (Cx_h(t) + x_p(t))' + f(t)(Cx_h(t) + x_p(t)) \\
&= Cx_h'(t) + x_p'(t) + f(t)Cx_h(t) + f(t)x_p(t) \\
&= (Cx_h'(t) + f(t)Cx_h(t)) + (x_p'(t) + f(t)x_p(t)) \\
&= 0 + g(t),
\end{aligned}
$$

  where we used (23) in the last equality.

- If $x(t)$ is a solution of (20), then it must be of the form (24). Indeed, since $x(t)$ and $x_p(t)$ both solve (20), $x(t) - x_p(t)$ solves the homogenous part (21) of (20). 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

$$\boxed{t^2\dot{x} + tx = 1}, \quad \boxed{x(1) = 2}. \tag{25}$$

1. The homogenous part is

$$t^2\dot{x} + tx = 0. \tag{26}$$

So the solution x_h to (26) 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 (25) 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

$$x(t) = \frac{C}{t} + \frac{\log|t|}{t}. \tag{27}$$

4. Finally, since $x(1) = 2$, we get by plugging $t = 1$ into (27)

$$2 = x(1) = C$$

and hence the solution of (25) is

$$x(t) = \frac{2 + \log|t|}{t}.$$

General solution of a linear DE

$$\boxed{y'(x) = f(x)y(x) + g(x)}. \tag{28}$$

1. The homogenous part is

$$y'(x) = f(x)y(x). \tag{29}$$

So the solution $y(x)$ to (29) 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:

$$y(x) = C(x) \cdot e^{\int f(x)dx}. \tag{30}$$

$$y'(x) = C'(x) \cdot e^{\int f(x)dx} + C(x)f(x)e^{\int f(x)dx}. \tag{31}$$

Using that (28)=(31) and by inserting the RHS of (30) instead of $y(x)$ in (28), 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 (28) *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

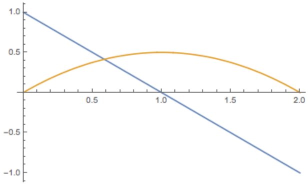
  $$v_h = Ce^{-kt/m} \quad \text{and} \quad v_p = -mg/k.$$

Motion of ball in the case $m = 1$, $k = 1$ and approximating $g \doteq 10$ (we will omit units)

Model

Velocity and position

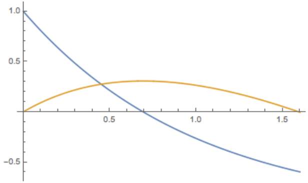
Solution



$ma = -mg$

$\dot{v} = -10$

$v(t) = -10t + 10$
$x(t) = -5t^2 + 10t$



$ma = -mg - kv$

$\dot{v} = -v - 10$

$v(t) = 20e^{-t} - 10$
$x(t) = 20 - 20e^{-t} - 10t$

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

$$\dot{x} = f\left(\frac{x}{t}\right). \tag{32}$$

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

$$\dot{x} = \dot{u}t + u. \tag{33}$$

Plugging (33) into (32) we get

$$\dot{u}t + u = f(u). \tag{34}$$

Rearranging (34) 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

$$y' = \frac{y}{x} - 1. \tag{35}$$

Introducing a new dependent variable

$$u = \frac{y}{x},$$

plugging in (35), we get

$$u'x + u = u - 1. \tag{36}$$

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:

$$x^2 + y^2 - 2ay = 0. \tag{37}$$

Differentiating (37) w.r.t. the independent variable gives

$$2x + 2yy' - 2ay' = 0. \tag{38}$$

Expressing a from (38) gives

$$a = \frac{x}{y'} + y. \tag{39}$$

Inserting (39) into (37) we obtain the DE for the given family

$$x^2 - y^2 - \frac{2xy}{y'} = 0. \tag{40}$$

Next we express y' from (40) and obtain

$$y' = \frac{2xy}{x^2 - y^2}. \tag{41}$$

The DE for orthogonal trajectories is obtained by substituting y' for $-1/y'$ in (41) to obtain

$$-\frac{1}{y'} = \frac{2xy}{x^2 - y^2}, \tag{42}$$

which is equivalent to

$$y' = -\frac{x^2 - y^2}{2xy}. \tag{43}$$

(43) is a homogeneous DE:

$$y' = -\frac{x^2 - y^2}{2xy} = -\frac{x}{2y} + \frac{y}{2x}$$

By introducing $y = ux$ we obtain

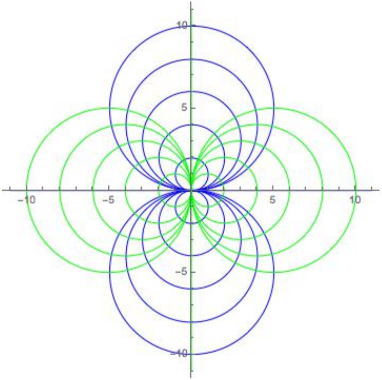
$$u'x + u = -\frac{1}{2u} + \frac{u}{2} \quad \Rightarrow \quad u'x = -\frac{1 + u^2}{2u} \quad \Rightarrow \quad \frac{2udu}{1 + u^2} = -\frac{dx}{x}$$
$$\Rightarrow \quad \log\left(1 + u^2\right) = -\log x + \log C \quad \Rightarrow \quad 1 + u^2 = \frac{C}{x},$$

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

$$M(t, x)dt + N(t, x)dx = 0. \tag{44}$$

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 (44) 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 (44) is exact, then the solutions are level curves of the function u:

$$u(t,x) = C, \qquad \textit{where} \quad C \in \mathbb{R}.$$

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 (44) to be exact is

$$\frac{\partial M}{\partial x} = \frac{\partial N}{\partial t}. \tag{45}$$

Moreover, if M and N are differentiable for every $(t,x) \in \mathbb{R}^2$, the condition (45) 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.

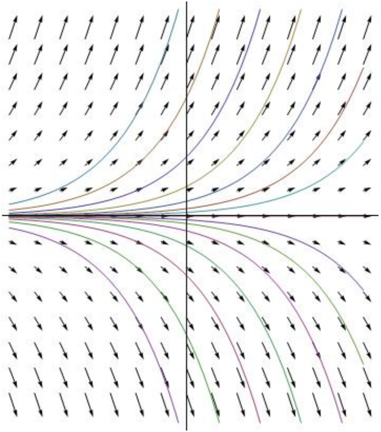
All these directions together form the directional field of the equation.

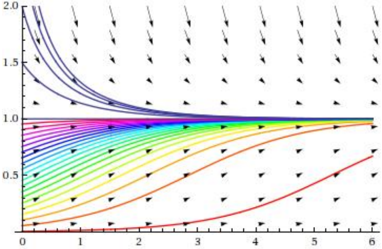
A solution of the equation is represented by a curve $y = y(x)$ that follows the given directions at every point x, i.e., the coefficient of the tangent corresponds to the value $f(x, y(x))$.

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: Click

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,...

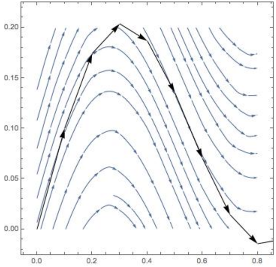
We will first look at the simplest and best known method and then a more practical improvement.

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 $x_n + ch$:

$$y_{n+1} = y_n + \underbrace{b_1}_{\text{weight}} \cdot \underbrace{(h \cdot f(x_n, y_n))}_{\substack{\text{move along} \\ \text{the tangent in } x_n}} + \underbrace{b_2}_{\text{weight}} \cdot \underbrace{(h \cdot f(x_n + ch, y(x_n + ch)))}_{\substack{\text{move along} \\ \text{the tangent in } x_n + ch}} \quad (46)$$

We use a linear approximation

$$y(x_n + ch) \approx y_n + chy'(x_n) = y_n + chf(x_n, y_n) \approx y_n + ahf(x_n, y_n), \quad (47)$$

where a is a new parameter.

Using (47) in (46) we obtain

$$y_{n+1} = y_n + b_1 \cdot \underbrace{(h \cdot f(x_n, y_n))}_{k_1} + b_2 \cdot \underbrace{(h \cdot f(x_n + ch, y_n + a \cdot k_1))}_{k_2}. \tag{48}$$

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 (48) we get a system of equations

$$\begin{aligned} 1 &= b_1 + b_2, \\ \frac{1}{2}(f_x + f_y f)_n &= b_2 c (f_x)_n + b_2 a (f f_y)_n, \end{aligned} \tag{49}$$

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 (49) has many different solutions, e.g.:

- $b_1 = b_2 = \frac{1}{2}$ and $c = a = 1$. RK method is:

$$\begin{aligned} y_{n+1} &= y_n + \frac{1}{2}(k_1 + k_2), \\ k_1 &= hf(x_n, y_n), \\ k_2 &= hf(x_n + h, y_n + k_1). \end{aligned}$$

- $b_1 = 0$, $b_2 = 1$ 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

$$\begin{aligned} y_{n+1} &= y_n + b_1 k_1 + b_2 k_2 + \ldots + b_s k_s, \\ k_1 &= hf(x_n, y_n), \\ k_2 &= hf(x_n + c_2 h, y_n + a_{2,1} k_1), \\ k_3 &= hf(x_n + c_3 h, y_n + a_{3,1} k_1 + a_{3,2} k_2), \\ &\vdots \\ k_s &= hf(x_n + c_s h, y_n + a_{s,1} k_1 + \ldots + a_{s,s-1} k_{s-1}). \end{aligned} \tag{50}$$

Butcher tableau

In a compact form the RK method (50) 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

$$
\begin{aligned}
y_{n+1} &= y_n + \frac{1}{6}k_1 + \frac{1}{3}k_2 + \frac{1}{3}k_3 + \frac{1}{6}k_4, \\
k_1 &= hf(x_n, y_n), \\
k_2 &= hf(x_n + \frac{1}{2}h, y_n + \frac{1}{2}k_1), \\
k_3 &= hf(x_n + \frac{1}{2}h, y_n + \frac{1}{2}k_2), \\
k_4 &= hf(x_n + h, y_n + k_3).
\end{aligned}
$$

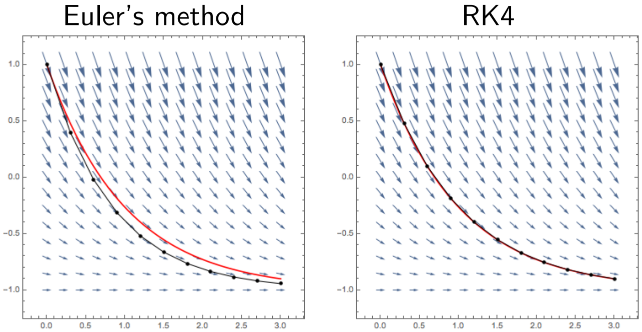
The error at each step is of order $\mathcal{O}(h^5)$. The cumulative error is of order $\mathcal{O}(h^4)$.

Euler vs RK4

Below is a comparison of Euler's and Rk4 methods for the DE

$$y' = -y - 1, \quad y(0) = 1 \qquad \text{with step size } h = 0.3:$$

The red curve is the exact solution $y = 2e^{-x} - 1$.



Algorithms and example: Click Click Click

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:

$$\begin{aligned} y_{n+1} &= y_n + k_1, \\ y_{n+1}^* &= y_n + \frac{1}{2}(k_1 + k_2). \end{aligned}$$

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 and example: Click Click

Systems of first order ODE's

Let

$$f := (f_1, \dots, f_n) : \mathbb{R}^{n+1} \to \mathbb{R}^n,$$

$$f(x_1, \dots, x_{n+1}) = (f_1(x_1, \dots, x_{n+1}), \dots, f_n(x_1, \dots, x_{n+1})).$$

be a vector function. A system of first order DE's is an equation

$$\dot{x}(t) = f(x(t), t), \tag{51}$$

where

$$x(t) := (x_1(t), \dots, x_n(t)) : I \to \mathbb{R}^n$$

is an unknown vector function and $I \subset \mathbb{R}$ is some interval. Coordinate-wise the system (51) is equal to

$$\begin{aligned} \dot{x}_1(t) &= f_1(x_1(t), \dots, x_n(t), t), \\ &\vdots \\ \dot{x}_n(t) &= f_n(x_1(t), \dots, x_n(t), t). \end{aligned}$$

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

$$\begin{bmatrix} \dot{x}_1(t) \\ \vdots \\ \dot{x}_n(t) \end{bmatrix} = \begin{bmatrix} a_{11}(t) & \dots & a_{1n}(t) \\ \vdots & \ddots & \vdots \\ a_{n1}(t) & \dots & a_{nn}(t) \end{bmatrix} \begin{bmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{bmatrix} + \begin{bmatrix} g_1(t) \\ \vdots \\ g_n(t) \end{bmatrix}, \qquad (52)$$

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 (52) can be written as

$$\dot{x}(t) = A(t)x + g(t), \qquad (53)$$

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 (53)

- 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

$$\dot{x} = Ax \tag{54}$$

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 (54) is diagonal. Then (54) 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 (54) 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

$$\dot{x}_1 = x_1 + x_2,$$
$$\dot{x}_2 = 4x_1 - 2x_2.$$

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

$$x(t) = C_1\Big(\cos(2t)\begin{bmatrix} 1 \\ 0 \end{bmatrix} - \sin(2t)\begin{bmatrix} 0 \\ 2 \end{bmatrix}\Big) + C_2\Big(\sin(2t)\begin{bmatrix} 1 \\ 0 \end{bmatrix} + \cos(2t)\begin{bmatrix} 0 \\ 2 \end{bmatrix}\Big).$$

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 $x_1(t_0), \dots, x_n(t_0)$ are linearly independent) of n linear equations for the constants $C_1, \dots, C_n$.

$$x^{(0)} = C_1 x_1(t_0) + \dots + C_n x_n(t_0).$$

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 $C_2 = 1$.

Transformating DEs of higher order into 1st order ODEs

The differential equation of order 2

$$\ddot{x} = f(t, x, \dot{x}) \tag{55}$$

can be transformed into a system of two order 1 DE's by introducing new variables:

$$\begin{aligned} x_1(t) &= x(t), \\ x_2(t) &= \dot{x}(t). \end{aligned}$$

Now DE (55) becomes

$$\begin{aligned} \dot{x}_1(t) &= x_2(t), \\ \dot{x}_2(t) &= f(t, x_1(t), x_2(t)). \end{aligned}$$

An initial condition

$$x(t_0) = \alpha_0, \quad \dot{x}(t_0) = \alpha_1$$

is transformed into an initial condition

$$\begin{bmatrix} x_1(t_0) \\ x_2(t_0) \end{bmatrix} = \begin{bmatrix} \alpha_0 \\ \alpha_1 \end{bmatrix}.$$

In the same way a differential equation of order n

$$x^{(n)} = f(t, x, \dot{x}, \ldots, x^{(n-1)})$$

can be transformed into a system of n differential equations of order 1 by introducing new dependent variables

$$\begin{aligned} x_1 &= x, \\ x_2 &= \dot{x}, \\ &\vdots \\ x_n &= x^{(n-1)}, \end{aligned} \tag{56}$$

and hence (56) becomes:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \vdots \\ \dot{x}_n \end{bmatrix} = \begin{bmatrix} x_2 \\ x_3 \\ \vdots \\ f(t, x_1, x_2, \ldots, x_n) \end{bmatrix}.$$

Example: We are given the differential equation of order 2

$$2\ddot{x} - 5\dot{x} + x = 0, \tag{57}$$

with initial conditions

$$x(3) = 6, \qquad \dot{x}(3) = -1. \tag{58}$$

We introduce new variables:

$$\begin{aligned} x_1(t) &= x(t), \\ x_2(t) &= \dot{x}(t), \end{aligned}$$

and hence (57) becomes the system

$$\begin{aligned} \dot{x_1}(t) &= x_2(t), \\ \dot{x_2}(t) &= \frac{5}{2}x_2 - \frac{1}{2}x_1. \end{aligned}$$

An initial conditions (58) becomes

$$x_1(3) = 6, \qquad x_2(3) = -1.$$

Numerical methods for a system of DEs

Numerical methods for a system of DEs work exactly in the same way as for a single equation, with the exception that the unknown function is a vector function

$$x(t) = \left[\begin{array}{ccc} x_1(t) & \cdots & x_n(t) \end{array} \right]^T .$$

Given the system with initial condition

$$\dot{x} = \left[\begin{array}{c} \dot{x}_1 \\ \vdots \\ \dot{x}_n \end{array} \right] = \left[\begin{array}{c} f_1(t, x_1, \ldots, x_n) \\ \vdots \\ f_n(t, x_1, \ldots, x_n) \end{array} \right], \quad x(t_0) = x^{(0)} = \left[\begin{array}{c} x_1^{(0)} \\ \vdots \\ x_n^{(0)} \end{array} \right],$$

we construct a recursive sequence of points

$$t_i = t_0 + ih, \quad x^{(i)} \doteq x(t_i), i \geq 0$$

where the vector $x^{(i)}$ is an approximation to the value of the exact solution $x(t_i)$, and h is the step size.

Euler's method and RK4

Euler's method:

$$t_{i+1} = t_i + h, \quad x^{(i+1)} = x^{(i)} + hf(t_i, x^{(i)}), \quad i \geq 0.$$

RK4 method:

$$t_{i+1} = t_i + h, \quad x^{(i+1)} = x^{(i)} + (k_1 + 2k_2 + 2k_3 + k_4)/6,$$

where

$$\begin{aligned}
k_1 &= hf(t_i, x^{(i)}), \\
k_2 &= hf(t_i + h/2, x^{(i)} + k_1/2), \\
k_3 &= hf(t_i + h/2, x^{(i)} + k_2/2), \\
k_4 &= hf(t_i + h, x^{(i)} + k_3).
\end{aligned}$$

Autonomous system of DE's - general case

A system of DEs is autonomous if the function $f : \mathbb{R}^n \to \mathbb{R}^n$ does not depend on t:

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

For an autonomous system, the tangent vector to a solution depends only on the point x and is independent of the time t at which the solution reaches a given point. In this case, the tangent vectors can be viewed as a directional field in the space $\mathbb{R}^n$.

In case of an autonomous system of 2 DE's:

$$f = (f_1, f_2) : \mathbb{R}^2 \to \mathbb{R}^2$$

$$\dot{x} = f_1(x, y),$$
$$\dot{y} = f_2(x, y),$$

gives a directional field in the (x, y) plane, which we call the phase plane of the system.

The general solution is a family of parametric curves or trajectories which respect the given directional field at every point (x, y).

The points where $f(x) = 0$ are stationary points or equilibrium points of the system.

At a stationary point $x_0 = x(t_0)$, $\dot{x}(t_0) = 0$, so $x(t) = x_0$ represents a constant, or equilibrium solution of the system.

Real life example of autonomous system

The predator-prey or Volterra-Lotka model is a famous system of DE's proposed by Alfred J. Lotka (1920) for modelling certain chemical reactions, and independently by Vito Volterra (1926) for dynamics of biological systems. It was later applied in economics and is used in a number of domains.

Two populations of species, for example rabbits and foxes, live together and depend on each other.

The number of rabbits (the prey) at time t is $R(t)$ and the number of foxes (the predators) is $F(t)$.

If they live apart, the rabbit, resp. fox, population grows, resp. declines, with the exponential law:

$$\dot{R} = kR, \quad k > 0, \quad \text{resp.} \quad \dot{F} = -rF, \quad r > 0.$$

If they live together, then interactions between rabbits and foxes cause a decline in the rabbit population and a growth of the fox population. This (basic) predator-prey model is the following:

$$\dot{R} = kR - aRF, \quad \dot{F} = -rF + bFR, \quad a, b > 0.$$

The system has two stationary or equilibrium points:

$$kR - aRF = -rF + bFR = 0 \Rightarrow$$

$$\Rightarrow R = F = 0 \qquad \text{or} \qquad R = \frac{r}{b}, F = \frac{k}{a}.$$

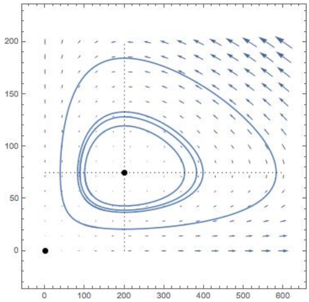
The meaning of these values is that the populations (ideally) coexist peacefully, with no fluctuations in the population sizes.

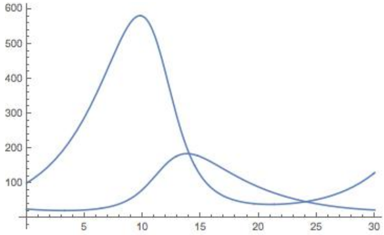
The left figure below shows the directional field and several solutions for the system

$$\dot{R} = 0.3R - 0.004RF \quad \dot{F} = -0.2F + 0.001FR$$

in the (R, F) plane.

The right figure shows dynamics of the population sizes $F(t)$ and $R(t)$ with respect to t:

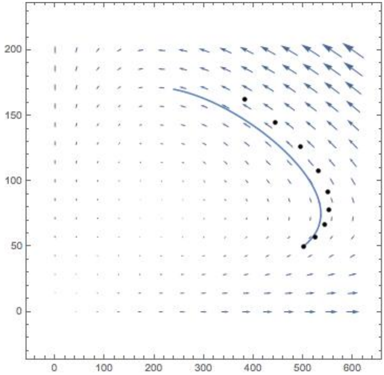




On the figure below, the blue curve is the exact solution and the black dots are approximations for function values for the system with initial condition

$$R(0) = 500, F(0) = 50$$

using Euler's method with step size $h = 0.5$:



Example: Click

The dynamics of systems of 2 equations

For an autonomous linear system

$$\dot{x}_1 = a_{11}x_1 + a_{12}x_2, \quad \dot{x}_2 = a_{21}x_1 + a_{22}x_2,$$

the origin $(0, 0)$ is always a stationary point, i.e., an equilibrium solution.

The eigenvalues of the matrix

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

determine the type of the stationary point $(0, 0)$ and the shape of the phase portrait.

We will assume that $\det A \neq 0$. Let λ_1, λ_2 be the eigenvalues of A. We also assume that there exist two linearly independent vectors v_1, v_2 of A (even if $\lambda_1 = \lambda_2$).

Case 1: $\lambda_1, \lambda_2 \in \mathbb{R}$

The general solution is

$$x(t) = C_1 e^{\lambda_1 t} v_1 + C_2 e^{\lambda_2 t} v_2.$$

- If $C_1 = 0$, the trajectory $x_1(t)$ is a ray in the direction of v_2 if $C_2 > 0$, or $-v_1$ if $C_2 < 0$.
- Similarly, if $C_2 = 0$ the trajectory $x_2(t)$ is a ray in the direction of v_2 or $-v_2$.
- The behaviour of other trajectories depends on the signs of λ_1 and λ_2.

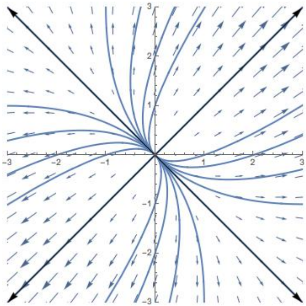
Subcase 1.1: $0 < \lambda_1 < \lambda_2$

- as $t \to \infty$, $x(t)$ asymptotically approaches the solution $\pm e^{\lambda_2 t} v_2$,
- as $t \to -\infty$, $x(t)$ asymptotically approaches the solution $\pm e^{\lambda_1 t} v_1$.

The point $(0, 0)$ is a source.

Example. The general solution of the system $\dot{x}_1 = 3x_1 + x_2, \dot{x}_2 = x_1 + 3x_2$ is

$$x(t) = C_1 e^{4t} \begin{bmatrix} 1 & 1 \end{bmatrix}^T + C_2 e^{2t} \begin{bmatrix} -1 & 1 \end{bmatrix}^T.$$



Example: Click

Subcase 1.2: $\lambda_2 < \lambda_1 < 0$

- as $t \to \infty$, $x(t)$ asymptotically approaches the solution $\pm e^{\lambda_1 t} v_2$,
- as $t \to -\infty$, $x(t)$ asymptotically approaches the solution $\pm e^{\lambda_2 t} v_1$.

The point $(0, 0)$ is a sink.

Example. The general solution of the system $\dot{x}_1 = -3x_1 - x_2, \dot{x}_2 = -x_1 - 3x_2$ is

$$x(t) = C_1 e^{-4t} \begin{bmatrix} 1 & 1 \end{bmatrix}^T + C_2 e^{-2t} \begin{bmatrix} -1 & 1 \end{bmatrix}^T.$$

Example: Click

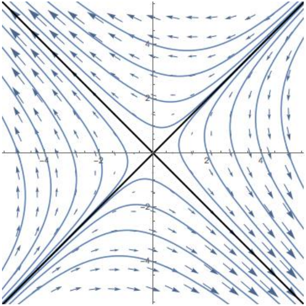
Subcase 1.3: $\lambda_1 < 0 < \lambda_2$

- as $t \to \infty$, $x(t)$ asymptotically approaches the solution $\pm e^{\lambda_2 t} v_2$,
- as $t \to -\infty$, $x(t)$ asymptotically approaches the solution $\pm e^{\lambda_1 t} v_1$.

The point $(0, 0)$ is a saddle.

Example. The general solution of the system $\dot{x}_1 = x_1 - 3x_2, \dot{x}_2 = -3x_1 + x_2$ is

$$x(t) = C_1 e^{-2t} \begin{bmatrix} 1 & 1 \end{bmatrix}^T + C_2 e^{4t} \begin{bmatrix} 1 & -1 \end{bmatrix}^T.$$



Example: Click

Subcase 2.1: $\lambda_{1,2} = \alpha \pm i\beta$, $\alpha \neq 0$

The general solution is

$$x(t) = e^{\alpha t}\left[(C_1 \cos(\beta t) + C_2 \sin(\beta t))u + (-C_1 \sin(\beta t) + C_2 \cos(\beta t))w\right].$$

Hence,

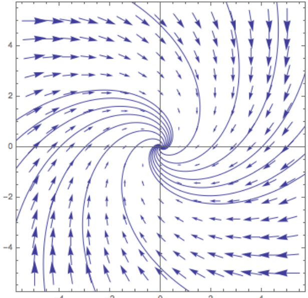
- if $\alpha < 0$, $x(t)$ spirals towards $(0,0)$ as $t \to \infty$, and
- if $\alpha > 0$, $x(t)$ spirals away from $(0,0)$ as $t \to \infty$.

The point $(0,0)$ is a spiral sink in the first case and a spiral source in the second case.

Example

$$\dot{x}_1 = -3x_1 + 2x_2, \dot{x}_2 = -x_1 - x_2$$

$$x(t) = e^{-2t}\cdot \left((C_1 \cos t + C_2 \sin t)\begin{bmatrix} 2 \\ 1 \end{bmatrix} + (-C_1 \sin t + C_2 \cos t)\begin{bmatrix} 0 \\ 1 \end{bmatrix}\right)$$



Subcase 2.2: $\lambda_{1,2} = \pm i\beta$, $\alpha \neq 0$

The trajectories are periodic with period $2\pi/\beta$, i.e. the point $x(t)$ circles around $(0,0)$.

The point $(0,0)$ is a center.

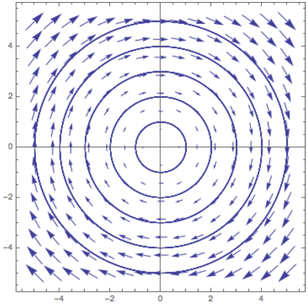
Example

$\dot{x} = v, \quad \dot{v} = -\omega^2 x$

$x(t) =$

$(C_1 \cos(\omega t) + C_2 \sin(\omega t)) \begin{bmatrix} 1 \\ 0 \end{bmatrix} +$

$(-C_1 \sin(\omega t) + C_2 \cos(\omega t)) \begin{bmatrix} 0 \\ 1 \end{bmatrix}$



Examples: Click Click

Nonlinear autonomous systems of equations

$$\dot{x}_1 = f_1(x_1, x_2), \quad \dot{x}_2 = f_2(x_1, x_2)$$

If $a = (a_1, a_2)$ is a critical point, that is,

$$f_1(a_1, a_2) = f_2(a_1, a_2) = 0,$$

then the behaviour of trajectories close to a is approximated by trajectories of the linearization of the system at the point a:

$$\dot{x}_1 \doteq \frac{\partial f_1}{\partial x_1}(x_1 - a_1) + \frac{\partial f_1}{\partial x_2}(x_2 - a_2), \quad \dot{x}_2 \doteq \frac{\partial f_2}{\partial x_1}(x_1 - a_1) + \frac{\partial f_2}{\partial x_2}(x_2 - a_2).$$

This is a linear homogeneous system with coefficient matrix the Jacobian matrix of the vector function $f(x)$:

$$\dot{x} \doteq Df(a)(x - a) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_1}{\partial x_2} \end{bmatrix} (x - a).$$

The critical point is classified as a source, sink, saddle, spiral source, spiral sink or center depending on the eigenvalues of $Df(a)$.

In addition to critical points, that is, equillibrium solutions, a plane nonlinear system (that is, a nonlinear system of two differential equations) can also have limit cycles.

A limit cycle is a periodic solutions $x_\infty(t)$ such that for initial conditions $x(t_0) = x_0$ in a certain domain the corresponding solutions $x(t)$

- either asymptotically tend towards $x_\infty(t)$ as $t \to \infty$ – in this case x_∞ is an attracting limit cycle, or
- $x(t) \to x_\infty(t)$ as $t \to -\infty$ – in this case x_∞ is a repelling limit cycle.

Systems of more than two differential equations can exhibit much more complex, chaotic behaviour.

Example: Click

Differential equations of order 2

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

The general solution is a two-parametric family

$$x = x(t, C_1, C_2).$$

A particular solution is given by specifying

- initial conditions: $x(t_0) = \alpha_0, \quad \dot{x}(t_0) = \alpha_1,$
  where the values of the solution and its derivative are given at some initial time t_0
  or
- boundary conditions: $x(a) = x_0, \quad x(b) = x_1$
  where values of the solution at different times a, b are given (i.e., on the boundary of some interval $[a, b]$)

Differential equations of order n

$$x^{(n)} = f(t, x, \dot{x}, \ldots, x^{(n-1)})$$

The general solution is an n-parametric family

$$x = x(t, C_1, \ldots, C_n).$$

A particular solution is given by

- initial conditions: $x(t_0) = \alpha_0, \ldots x^{(n-1)}(t_0) = \alpha_{n-1}$
  where the values of the solution and its first $(n-1)$ derivatives are given at some initial time t_0
  or
- boundary conditions
  where values of the solution or its derivatives are given in different times.

Linear DE's of order n

A linear DE (LDE) of degree n is of the form

$$x^{(n)} + a_{n-1}(t)x^{(n-1)} + \cdots + a_0(t)x = f(t). \tag{59}$$

The equation is

- homogeneous if $f(t) = 0$, and
- nonhomogeneous if $f(t) \neq 0$.
- The general solution of the homogeneous part is the family of all linear combinations
$$y(t) = C_1 x_1(t) + \cdots + C_n x_n(t)$$
of n linearly independent solutions $x_1(t), \ldots, x_n(t)$.
- If the coefficients $a_1(t), \ldots, a_n(t)$ are continuous functions, then for any $\alpha_0, \ldots, \alpha_n$ there exists exactly one solution satisfying the initial condition
$$x(t_0) = \alpha_0, \quad \dot{x}(t_0) = \alpha_1, \quad \ldots, \quad x^{(n-1)}(t_0) = \alpha_n.$$

LDEs with constant coefficients

Assume that the coefficient functions $a_1(t), \ldots a_n(t)$ in a homogeneous LDE are constant:

$$x^{(n)} + a_{n-1}x^{(n-1)} \cdots + a_0 x = 0, \quad a_1, \ldots a_n \in \mathbb{R} \tag{60}$$

Translating (60) to the system by the usual trick of introducing new variables

$$x_1 = x, \quad x_2 = x_1', \quad x_3 = x_2', \quad \cdots, \quad x_n = x_{n-1}',$$

(60) becomes

$$x_n' = -a_0 x_1 - a_1 x_2 - \ldots - a_{n-1} x_n,$$

or matricially $\vec{x}' = A\vec{x}$:

$$\vec{x}'(t) = \begin{bmatrix} 0 & 1 & 0 & \cdots & & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & & \ddots & \ddots & & \vdots \\ \vdots & & & \ddots & \ddots & \vdots \\ 0 & \cdots & & \cdots & 0 & 1 \\ -a_0 & -a_1 & \cdots & \cdots & -a_{n-2} & -a_{n-1} \end{bmatrix} \vec{x}(t)$$

- The solutions to this system are of the form

$$x(t) = p_k(t)e^{\lambda t}v,$$

  where λ is the eigenvalue of A, $p_k(t)$ is a polynomial of degree k in t and v is the generalized eigenvector. (This follows most easily by the use of the Jordan form of the matrix.)
- In particular, if there are n linearly independent eigenvectors of the matrix A, then all polynomials p_k are constants and generalized eigenvectors are usual eigenvectors.
- By a simple calculation of expressing the determinant of $A - \lambda I$ according to the coefficients and cofactors of the last row, it turns out that the eigenvalues of A are precisely the roots of the characteristic polynomial corresponding to (60):

$$P(\lambda) := \lambda^n + a_{n-1}\lambda^{n-1} + \cdots a_1\lambda + a_0. \quad (61)$$

- A (trivial) fact with a nontrivial proof, called the fundamental theorem of algebra, states that a polynomial of degree n has exactly n roots, counted by multiplicity. In case the matrix A is real, these roots are real or complex conjugate pairs.

- From the roots of the characteristic polynomial (61), n linearly independent solutions of the LDE can be reconstructed.
- For every real root $\lambda \in \mathbb{R}$,

$$x(t) = e^{\lambda t}$$

  is a solution of the homogeneous LDE.
- For a complex conjugate pair of roots $\lambda = \alpha \pm i\beta$, the real and imaginary parts of the complex-valued exponential functions

$$e^{(\alpha \pm i\beta)t} = e^{\alpha t}(\cos(\beta t) + i \sin(\beta t))$$

  are two linearly independent solutions

$$x_1 = e^{\alpha t} \cos(\beta t), \quad x_2 = e^{\alpha t} \sin(\beta t).$$

Proposition

If a root (or a complex pair of roots) λ has multiplicity $k > 1$, then it can be shown that

$$e^{\lambda t}, te^{\lambda t}, \dots t^{k-1} e^{\lambda t}$$

are all linearly independent solutions.

Proof of proposition

Let us prove the last fact by an interesting trick. We introduce the operator

$$L : \mathcal{C}^{(n)}(I) \to \mathcal{C}(I)$$

$$L(u) = u^{(n)} + a_{n-1}u^{(n-1)} \cdots + a_0 u,$$

where $\mathcal{C}^{(n)}(I)$ stands for the vector space of n-times continuously differentiable functions on the interval I and $\mathcal{C}(I)$ stands for the vector space of continuous functions on I.

Let λ_0 be the root of the characteristic polynomial (61) of multiplicity k, i.e.,

$$P(\lambda) = (\lambda - \lambda_0)^k Q(\lambda).$$

Let $0 \leq q \leq k$ by an integer. We will check that $t^q e^{\lambda t}$ solves (60).
Notice that

$$t^q e^{\lambda t} = \frac{d^q}{d\lambda^q} e^{\lambda t}.$$

For ease of notation we define $a_n := 1$. We have that:

$$L(t^q e^{\lambda t}) = \sum_{i=0}^{n} a_i \left(\frac{d^q}{d\lambda^q} e^{\lambda t} \right)^{(i)} = \sum_{i=0}^{n} a_i \frac{d^i}{dt^i} \left(\frac{d^q}{d\lambda^q} e^{\lambda t} \right)$$

$$= \frac{d^q}{d\lambda^q} \left(\sum_{i=0}^{n} a_i \frac{d^i}{dt^i} e^{\lambda t} \right) = \frac{d^q}{d\lambda^q} \left(\sum_{i=0}^{n} a_i \lambda^i e^{\lambda t} \right) = \frac{d^q}{d\lambda^q} (P(\lambda) e^{\lambda t}).$$

Since

$$\frac{d^q}{d\lambda^q}(P(\lambda)e^{\lambda t}) = \sum_{i=1}^{q} \frac{d^i}{d\lambda^i}(P(\lambda)) \cdot Q_i(t,\lambda),$$

where $Q_i(t,\lambda)$ are functions of t, λ and

$$\frac{d^i}{d\lambda^i}(P(\lambda_0)) = 0, \quad \text{for} \quad i = 0, \ldots, q,$$

it follows that

$$L(t^q e^{\lambda_0 t}) = 0.$$

Test for linear independence of solutions

Let $x_1(t), \ldots, x_n(t)$ be the solutions of the homogeneous part of (59) and form a matrix

$$W(x_1(t), \ldots, x_n(t)) := \begin{bmatrix} x_1(t) & \ldots & x_n(t) \\ \dot{x}_1(t) & \ldots & \dot{x}_n(t) \\ \vdots & \ddots & \vdots \\ x_1^{(n-1)}(t) & \ldots & x_n^{(n-1)}(t) \end{bmatrix}$$

We call the determinant

$$\phi(t) = \det W(x_1(t), \ldots, x_n(t)) : I \to \mathbb{R}$$

the Wronskian determinant of $W(x_1(t), \ldots, x_n(t))$, where I is the interval on which t lives.

Theorem (Existence and uniqueness of solutions)

If $x_1(t), \ldots, x_n(t)$ are solutions of a LDE with continuous coefficient functions $a_1(t), \ldots a_n(t)$, then their Wronskian is either identically equal to 0 *or nonzero at every point. In other words, if $W(x_1, \ldots, x_n)$ has a zero at some point t_0, then it is identically equal to* 0*.*

Proof of theorem

Let π_n be the set of all permutations of the set $\{1, \ldots, n\}$. Now we differentiate $\phi(t)$ and obtain

$$
\begin{aligned}
\phi'(t) &= \left(\sum_{\sigma \in \pi_n} x_{\sigma(1)} x^{(1)}_{\sigma(2)} \cdots x^{(n-1)}_{\sigma(n)} \right)'(t) \\
&= \sum_{\sigma \in \pi_n} \Bigg(\left(x_{\sigma(1)}\right)'(t) x^{(1)}_{\sigma(2)}(t) \cdots x^{(n-1)}_{\sigma(n)}(t) \\
&\qquad + x_{\sigma(1)}(t) \left(x^{(1)}_{\sigma(2)} \right)'(t) \cdots x^{(n-1)}_{\sigma(n)}(t) + \cdots \\
&\qquad + x_{\sigma(1)}(t) x^{(1)}_{\sigma(2)}(t) \cdots \left(x^{(n-1)}_{\sigma(n)}(t) \right)' \Bigg) \\
&= \left(\sum_{\sigma \in \pi_n} x^{(1)}_{\sigma(1)}(t) x^{(1)}_{\sigma(2)}(t) \cdots x^{(n-1)}_{\sigma(n)}(t) \right) + \\
&\qquad \left(\sum_{\sigma \in \pi_n} x_{\sigma(1)}(t) x^{(2)}_{\sigma(2)}(t) x^{(2)}_{\sigma(3)}(t) \cdots x^{(n-1)}_{\sigma(n)}(t) \right) + \\
&\qquad \cdots + \left(\sum_{\sigma \in \pi_n} x_{\sigma(1)}(t) x^{(1)}_{\sigma(2)}(t) \cdots x^{(n-2)}_{\sigma(n)}(t) x^{(n)}_{\sigma(n)}(t) \right).
\end{aligned}
$$

Now notice that the first $n-1$ summand are the determinants of the matrices

$$\begin{bmatrix} x_1(t) & \dots & x_n(t) \\ \vdots & \ddots & \vdots \\ x_1^{(i)}(t) & \dots & x_n^{(i)}(t) \\ x_1^{(i)}(t) & \dots & x_n^{(i)}(t) \\ \vdots & \ddots & \vdots \\ x_1^{(n-1)}(t) & \dots & x_n^{(n-1)}(t) \end{bmatrix} \tag{62}$$

and hence are equal to 0.

For the last summand use the initial DE (59) to express

$$x_{\sigma(n)}^{(n)} := -a_{n-1}(t)x_{\sigma(n)}^{(n-1)} - \dots - a_0(t)x_{\sigma(n)}.$$

The summands of the from $-a_i(t)x_{\sigma(n)}^{(i)}$ for $i < n-1$ give 0 terms in the sum $\sum_{\sigma\in\pi_n}$ since the sum is just the $-a_i(t)$ multiple of the determinant of the form (62), while the term $-a_{n-1}(t)x_{\sigma(n)}^{(n-1)}$ gives

$$-a_{n-1}(t)\phi(t).$$

It follows that $\phi(t)$ satisfies the DE

$$\phi'(t) = -a_{n-1}(t)\phi(t).$$

The theorem follows by noticing that the solution of this DE is

$$\phi(t) = ke^{-\int a_{n-1}(t)dt}, \quad \text{where} \quad k \in \mathbb{R}.$$

Second order homogeneous LDE with constant coefficients

We are given a DE

$$a\ddot{x} + b\dot{x} + cx = 0,$$

where $a, b, c \in \mathbb{R}$ are real numbers. We know from the theory above that the general solution is

$$x(t, C_1, C_2) = C_1 x_1(t) + C_2 x_2(t),$$

where $C_1, C_2 \in \mathbb{R}$ are parameters and

1. $x_1(t) = e^{\lambda_1 t}$ and $x_2(t) = e^{\lambda_2 t}$ if the characteristic polynomial has two distinct real roots,
2. $x_1(t) = e^{\alpha t} \cos \beta t$ and $x_2(t) = e^{\alpha t} \sin \beta t$ if the characteristic polynomial has a complex pair $\lambda_{12} = \alpha \pm i\beta$ of roots, and
3. $x_1(t) = e^{\lambda t}$, $x_2(t) = te^{\lambda t}$ if the characteristic polynomial has one double real root.

Nonhomogeneous LDEs

We are given the nonhomogeneous LDE

$$x^{(n)} + a_{n-1}(t)x^{(n-1)} + \cdots + a_0(t)x = f(t),$$

where $f : I \to \mathbb{R}$ is a nonzero function on the interval I. The following holds:

- If x_1 and x_2 are solutions of the nonhomogeneous equation, the difference $x_1 - x_2$ is a solution of the corresponding homogeneous equation.
- The general solution is a sum

  $$x(t, C_1, C_2) = x_p + x_h = x_p + C_1 x_1 + \cdots + C_n x_n,$$

  where x_p is a particular solution of the nonhomogeneous equation and $x_1, \ldots, x_n$ are linearly independent solutions of the homogeneous equation.
- The particular solution can be obtained using the method of "intelligent guessing" or the method of variation of constants.

The method of "intelligent guessing" typically works if the function $f(t)$ belongs to a class of functions that is closed under derivations, like polynomials, exponential functions and sums of these.

Example ($\ddot{x} + \dot{x} + x = t^2$)

We are guessing that the particular solution will be of the form

$$x_p(t) = At^2 + Bt + C.$$

We have that

$$\dot{x}_p(t) = 2At + B, \quad \ddot{x}_p(t) = 2A,$$

and so

$$\begin{aligned} \ddot{x} + \dot{x} + x &= 2A + (2At + B) + (At^2 + Bt + C) \\ &= At^2 + (2A + B)t + (2A + B + C) \end{aligned}$$

The initial DE gives us a linear system in A, B, C :

$$A = 1, \quad 2A + B = 0, \quad 2A + B + C = 0$$

with the solution $A = 1, B = -2, C = 0$. Hence, $x_p(t) = t^2 - 2t$.

Example ($\ddot{x} - 3\dot{x} + 2x = e^{3t}$)

We are guessing that the particular solution will be of the form

$$x_p(t) = Ae^{3t}.$$

We have that

$$\dot{x}_p(t) = 3Ae^{3t}, \quad \ddot{x}_p(t) = 9Ae^{3t},$$

and so

$$\ddot{x} - 3\dot{x} + 2x = 9Ae^{3t} - 3(3Ae^{3t}) + 2Ae^{3t} = 2Ae^{3t}$$

The initial DE gives us an equation $2A = 1$ and hence, $x_p(t) = \frac{1}{2}e^{3t}$.

Example ($\ddot{x} - x = e^t$)

The particular solution will not be of the form $x_p(t) = Ae^t$, since this is a solution of the homogeneous equation, we are guessing that the correct form in this case is

$$x_p(t) = Ate^t.$$

We have that

$$\dot{x}_p(t) = A(e^t + te^t), \quad \ddot{x}_p(t) = A(2e^t + te^t),$$

and so

$$\ddot{x} - x = A(2e^t + te^t) - Ate^t = 2Ae^t.$$

The initial DE gives us an equation $2A = 1$ and hence, $x_p(t) = \frac{1}{2}te^t$.

Example ($\ddot{x} + x = \frac{1}{\cos t}$)

Let us first solve the homogeneous part $\ddot{x} + x = 0$. The characteristic polynomial is $p(\lambda) = \lambda^2 + 1$ with zeroes

$$\lambda_{1,2} = \pm i = \cos t \pm i \sin t.$$

Hence, real solutions of the DE are

$$x_1(t) = \cos t \qquad \text{and} \qquad x_2(t) = \sin t. \tag{63}$$

So the general solution to the homogeneous part is

$$x(t) = C_1 x_1(t) + C_2 x_2(t), \qquad \text{where} \quad C_1, C_2 \in \mathbb{R} \quad \text{are constants.}$$

Now we are searching for the particular solution $x_p(t)$ of the form

$$x_p(t) = C_1(t) x_1(t) + C_2(t) x_2(t).$$

Thus,

$$\dot{x}_p(t) = \dot{C}_1(t)x_1(t) + C_1(t)\dot{x}_1(t) + \dot{C}_2(t)x_2(t) + C_2(t)\dot{x}_2(t). \tag{64}$$

We force an equation

$$\dot{C}_1(t)x_1(t) + \dot{C}_2(t)x_2(t) = 0. \tag{65}$$

Differentiang (64) further under the assumption (65) we get

$$\ddot{x}_p(t) = (\dot{C}_1(t)\dot{x}_1(t) + C_1(t)\ddot{x}_1(t)) + (\dot{C}_2(t)\dot{x}_2(t) + C_2(t)\ddot{x}_2(t)). \tag{66}$$

Plugging this into the initial DE and using that x_1, x_2 are solutions of $\ddot{x} + x = 0$

$$\dot{C}_1(t)\dot{x}_1(t) + \dot{C}_2(t)\dot{x}_2(t) = \frac{1}{\cos t}. \tag{67}$$

Expressing $\dot{C}_2(t)$ from (65) and plugging into (67) we get

$$\dot{C}_1(t)\dot{x}_1(t) - \frac{\dot{C}_1(t)x_1(t)}{x_2(t)}\dot{x}_2(t) = \dot{C}_1(t)\frac{\dot{x}_1(t)x_2(t) - x_1(t)\dot{x}_2(t)}{x_2(t)} = \frac{1}{\cos t}. \tag{68}$$

Using (63) in (68) we get

$$\dot{C}_1(t) = -\frac{\sin t}{\cos t}. \tag{69}$$

Hence,

$$C_1(t) = -\int \frac{\sin t}{\cos t}dt = -\int \frac{1}{u}du = -\log|u| = -\log|\cos t|,$$

where we used the substitution $u = \cos t$.

Using (69) in (65) we get

$$\dot{C}_2(t) = 1. \tag{70}$$

Hence,

$$C_2(t) = t.$$

So,

$$x_p(t) = -\log|\cos t| \cdot \cos t + t \sin t.$$

The complete solution to DE is

$$x(t) = C_1 \cos t + C_2 \sin t - \log|\cos t| \cdot \cos t + t \sin t,$$

where C_1, C_2 are parameters.

Vibrating systems

There are many vibrating systems in many different domains. The mathematical model is always the same, though. We will have in mind a vibrating mass attached to a spring.

Case 1: Free vibrations without damping

Let $x(t)$ denote the displacement of the mass from the equillibrium position.

- According to **Newton's second law of motion**

$$m\ddot{x} = \sum F_i,$$

 where F_i are forces acting on the mass.
- By **Hooke's law**, the only force acting on the mass pulls towards the equilibrium, its size is proportional to the displacement and the direction is opposite

$$F = -kx(t), \quad k > 0.$$

- So the DE in this case is

$$\boxed{m\ddot{x} + kx = 0}.$$

- The characteristic equation

$$m\lambda^2 + k = 0$$

has complex solutions $\lambda = \pm\omega i$, $\omega^2 = k/m$.

- The general solution is

$$x(t) = C_1 \cos \omega t + C_2 \sin \omega t.$$

- So the solutions $x(t)$ are periodic. The equillibrium point $(0, 0)$ in the phase plane (x, v) is a center.

Case 2: Free vibrations with damping

We assume a linear damping force

$$F_d = -\beta \dot{x},$$

so the DE is

$$\boxed{m\ddot{x} + \beta\dot{x} + kx = 0}, \quad \text{where} \quad m, \beta, k > 0.$$

Depending on the solutions of the characteristic equation there are three cases:

- Overdamping when $D = \beta^2 - 4km > 0$ and $x(t) = C_1 e^{\lambda_1 t} + C_2 e^{\lambda_2 t}$, $\lambda_{1,2} < 0$. The mass slides towards the equilibrium. The point $(0, 0)$ in the (x, v) plane is a sink.
- Critical damping when $D = 0$ and $x(t) = C_1 e^{\lambda t} + C_2 t e^{\lambda t}$, $\lambda < 0$. The point mass slides towards the equillibrium after, possibly, one swing. The point $(0, 0)$ in the (x, v) plane is a sink,
- Damped vibration when $D < 0$ and $x(t) = e^{\alpha t}(C_1 \cos \beta t + C_2 \sin \beta t)$. The mass oscillates around the equillibrium with decreasing amplitudes. The point $(0, 0)$ is a spiral sink.

Case 3: Forced vibration without damping

In addition to internal forces of the system there is an additional external force $f(t)$ acting on the system, so

$$\boxed{m\ddot{x} + kx = f(t)}.$$

The general solution is of the form

$$x(t, C_1, C_2) = C_1 \cos(\omega t) + C_2 \sin(\omega t) + x_p(t),$$

where x_p is a particular solution of the nonhomogeneous equations.

Example

Let $f(t) = a \sin \mu t$.

Using the method of intelligent guessing,

- if $\mu \neq \omega$, then $x_p(t) = A \sin \mu t + B \cos \mu t$
- if $\mu = \omega$, then $x_p = t(A \sin \omega t + B \cos \omega t)$, so the solutions of the equation are unbonded and incerase towards ∞ as $t \to \infty$ – the well known phenomenion of resonance occurs.

Case 4: Forced vibration with damping:

$$\boxed{m\ddot{x} + \beta\dot{x} + kx = f(t)}.$$

Example

Let $f(t) = a \sin \mu t$.
The general solution is of the form

$$x(t, C_1, C_2) = x_h + x_p = C_1 x_1(t) + C_2 x_2(t) + x_p(t)$$

where $x_p(t)$ is of the form $A \sin \mu t + B \cos \mu t$, and the two solutions x_1 and x_2 both converge to 0 as $t \to \infty$. For any C_1, C_2 the solution $x(t, C_1, C_2)$ asymptotically tends towards $x_p(t)$.