Math 285
Introduction to Differential Equations

Thomas Honold

ZJU-UIUC Institute

Spring Semester 2024

Outline

Today's Lecture: Introduction

Teaching Staff

Lecturer

Prof. Dr. Thomas Honold
ZJU-UIUC Institute, Zhejiang University
International Campus, Haining
Office: Room C415, ZJUI Building
Office hours: by appointment
Email: honold@zju.edu.cn

Teaching Assistants

Li Pengyu pengyu.20@intl.zju.edu.cn
Hu Kejia kejia.20@intl.zju.edu.cn
Yu Jiarui jiarui.20@intl.zju.edu.cn
Ren Hao haor.20@intl.zju.edu.cn UG
Chen Yang yangc.20@intl.zju.edu.cn
Liang Yilai yilai.20@intl.zju.edu.cn
Ge Yilei yilei.20@intl.zju.edu.cn
Liang Weijie Weijie weijie.20@intl.zju.edu.cn

Weekly Timetable

Lecture A

Wed 15-16, Thu/Fri 13-14
RC-2 Multifunctional Hall

Lecture B

Wed 16-17, Thu/Fri 14-15
RC-2 Multifunctional Hall

Discussion with TA's

There will be informal weekly discussion groups on Thu at different timeslots.

Details will be announced later.

Homework

Homework is assigned on Thursdays and must be handed in in the following week before the discussion session. Late homework will not be accepted.

Textbook

[BDM17] William E. Boyce, Richard C. DiPrima, Douglas B. Meade, *Elementary Differential Equations and Boundary Value Problems*, 11th global edition, Wiley, 2017.

Teaching Calendar (tentative)

Adapted for Math 285

Week	Topics	[BDM17] Sections
1	Introduction to ODE's	Ch. 1
2–4	1st Order ODE's	Ch. 2
5,6	2nd-Order ODE's	Ch. 3
7,8	Higher Order ODE's	Ch. 4
9,10	Series Solutions	Ch. 5
11	Laplace Transform	Ch. 6
12,13	1st Order ODE Systems	Ch. 7
14	PDE's	Ch. 10

Course material

Textbook + Lecture Slides + Exercises

The Weekly schedule is only approximately true.
The lecture won't follow the textbook strictly.

Examination Regulations

Calculation of the final score

- 45 % final exam (3 hours, closed book)
- 15 % 1 midterm exam (1 hour, closed book)
- 25 % homework
- 15 % lab project
- 5 % extra credit for presenting solutions of exercises

Exam dates will be announced in due course.

The lab projects will be assigned after the midterm, and details will be fixed at this time.

Course Website

As usual, lecture slides, homework assignments, and other accompanying material will be made available through Blackboard `https://learn.intl.zju.edu.cn`

Further details regarding homework submission, TA office hours, etc., will be announced later.

Some Advice Before We Start

- Attend each class!
- Solve (well, at least try hard to solve) each exercise!
- Don’t hesitate to ask (stupid) questions!

Attendance Control

As in Fall 2023, lecture attendance will be checked electronically, and multiple unauthorized absence can be penalized through score deduction or by other means.

Students can be exempted from class attendance, but only for very important reasons and with prior authorization.

Attending discussion sessions in full is not required.

The Subject of the Course

Informally, an *ordinary differential equation* (or ODE, for short) is an equation for a one-variable function $f(t)$ and its derivatives $f'(t)$, $f''(t)$, etc.
This is in contrast to *partial differential equations* (or PDE, for short), which involve multi-variable functions $f(x_1, \ldots, x_n)$ and their partial derivatives $\frac{\partial f}{\partial x_i}$, $\frac{\partial^2 f}{\partial x_i \partial x_j}$, etc.

Definition (Ordinary Differential Equation, ODE)

An ODE of order n has the form

$$F(t, \mathbf{y}, \mathbf{y}', \mathbf{y}'', \ldots, \mathbf{y}^{(n)}) = 0, \qquad (\star)$$

where F has domain $D \subseteq \mathbb{R} \times \underbrace{\mathbb{R}^m \times \cdots \times \mathbb{R}^m}_{n+1}$ and depends on the last variable (otherwise the order is $< n$).

A *solution* to $(\star)$ is a function (curve) $f\colon I \to \mathbb{R}^m$, defined on an interval $I \subseteq \mathbb{R}$, which is n times differentiable and satisfies $F\big(t, f(t), f'(t), f''(t), \ldots, f^{(n)}(t)\big) = 0$ for all $t \in I$.

Definition (Initial Value Problem, IVP)

Suppose that an ODE as above is given and $t_0 \in \mathbb{R}$, $\mathbf{y}_0, \ldots, \mathbf{y}_n \in \mathbb{R}^m$ are such that $F(t_0, \mathbf{y}_0, \ldots, \mathbf{y}_n) = 0$. A solution to the *initial value problem*

$$F(t, \mathbf{y}, \mathbf{y}', \mathbf{y}'', \ldots, \mathbf{y}^{(n)}) = 0, \quad \mathbf{y}^{(i)}(t_0) = \mathbf{y}_i \ \textit{ for } 0 \le i \le n$$

is any function (curve) $f\colon I \to \mathbb{R}^m$ solving $(\star)$ on the previous slide and satisfying $f(t_0) = \mathbf{y}_0$, $f'(t_0) = \mathbf{y}_1$, $\ldots$, $f^{(n)}(t_0) = \mathbf{y}_n$.

Notes

- It is custom to use "no-name notation" $y = y(t)$ if $m = 1$ (resp., $\mathbf{y} = \mathbf{y}(t)$ if $m > 1$) for solutions of ODE's.
- Denoting the "independent" variable by t reflects the virtually zillions of applications in Physics, where $\mathbf{y}(t)$ models the state of a physical system at time t. Be prepared, however, that many texts on ODE's use x in place of t, i.e., $y(x)$ or $\mathbf{y}(x)$ for the solution function of an ODE.

Notes cont'd

- While our definition of ODE's and IVP's is the most general, the following *explicit* form of an n-th order ODE occurs most frequently:

$$\mathbf{y}^{(n)} = G(t, \mathbf{y}, \mathbf{y}', \ldots, \mathbf{y}^{(n-1)}).$$

A corresponding implicit form is $F(t, \mathbf{y}, \mathbf{y}', \ldots, \mathbf{y}^{(n)}) = 0$ with $F(t, \mathbf{y}_0, \ldots, \mathbf{y}_n) = \mathbf{y}_n - G(t, \mathbf{y}_0, \ldots, \mathbf{y}_{n-1})$.

An IVP in explicit form needs to specify only $\mathbf{y}^{(i)}(t_0) = \mathbf{y}_i$ for $0 \le i \le n-1$, since the last condition

$$\begin{aligned}\mathbf{y}^{(n)}(t_0) = \mathbf{y}_n &= G(t_0, \mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_{n-1}) \\ &= G(t_0, \mathbf{y}(t_0), \mathbf{y}'(t_0), \ldots, \mathbf{y}^{(n-1)}(t_0))\end{aligned}$$

is a particular case of the explicit ODE.

- Sometimes it is easy to find solutions (or a family of solutions) to a given ODE, and the question arises whether there are further solutions. Since, trivially, restricting a solution $y\colon I \to \mathbb{R}$ to a subinterval $J \subset I$ yields another solution, we are only interested in **maximal solutions**, i.e., those which do not arise by (proper) restriction from another solution.

Ten Examples

1. $y' = a(t)$; more generally, $\mathbf{y}' = \mathbf{a}(t)$
   $F(t, y_0, y_1) = a(t) - y_1$ resp. $F(t, \mathbf{y}_0, \mathbf{y}_1) = \mathbf{a}(t) - \mathbf{y}_1$
2. $y' = y$; more generally, $\mathbf{y}' = \mathbf{y}$
   $F(t, y_0, y_1) = y_1 - y_0$ resp. $F(t, \mathbf{y}_0, \mathbf{y}_1) = \mathbf{y}_1 - \mathbf{y}_0$
3. $y' = ay$ with $a \in \mathbb{R}$; more generally, $\mathbf{y}' = \mathbf{A}\mathbf{y}$ with $\mathbf{A} \in \mathbb{R}^{n\times n}$
   $F(t, y_0, y_1) = y_1 - ay_0$ resp. $F(t, \mathbf{y}_0, \mathbf{y}_1) = \mathbf{y}_1 - \mathbf{A}\mathbf{y}_0$
4. $y' = ay + b$ with $a, b \in \mathbb{R}$; more generally, $\mathbf{y}' = \mathbf{A}\mathbf{y} + \mathbf{b}$
   $F(t, y_0, y_1) = y_1 - ay_0 - b$
5. $y' = 2ty$; more generally, $\mathbf{y}' = 2t\,\mathbf{y}$
   $F(t, y_0, y_1) = y_1 - 2ty_0$
6. $y' = y^2$
   $F(t, y_0, y_1) = y_1 - y_0^2$
7. $y' = \sqrt{|y|}$
   $F(t, y_0, y_1) = y_1 - \sqrt{|y_0|}$

Ten Examples Cont'd

8. $(e^{x-y} + 2x - e)\,dx - (e^{x-y} + 2y)\,dy = 0$
   $F(t, x_0, y_0, x_1, y_1) = (e^{x_0 - y_0} + 2x_0 - e)x_1 - (e^{x_0 - y_0} + 2y_0)y_1$
9. $y' = -x/y$
   $F(x, y_0, y_1) = y_1 + x/y_0$
10. $y'' + y = 0$ (or, in explicit form, $y'' = -y$)
    $F(t, y_0, y_1, y_2) = y_2 + y_0$ (resp., $G(t, y_0, y_1) = -y_0$)

The first nine examples have order 1. The last example has order 2.

Examples 8 and 10 are implicit ODE's. The remaining examples are explicit ODE's.

Reading assignment for Week 1

[BDM17], Chapter 1

Ten Examples Cont'd

Solutions

1. $y' = a(t)$: Assuming that $t \mapsto a(t)$ is continuous, the general solution is $y(t) = \int a(t)\,\mathrm{d}t$ (by the Fundamental Theorem of Calculus).

   The solution of the IVP $y' = a(t)$, $y(t_0) = y_0$ is $y(t) = y_0 + \int_{t_0}^{t} a(\tau)\,\mathrm{d}\tau$.

   The solution of the IVP $\mathbf{y}' = \mathbf{a}(t)$, $\mathbf{y}(t_0) = \mathbf{y}^{(0)}$ is

$$\mathbf{y}(t) = \begin{pmatrix} y_1^{(0)} + \int_{t_0}^{t} a_1(\tau)\,\mathrm{d}\tau \\ \vdots \\ y_m^{(0)} + \int_{t_0}^{t} a_m(\tau)\,\mathrm{d}\tau \end{pmatrix},$$

   where $\mathbf{a}(t) = \big(a_1(t), \ldots, a_m(t)\big)^{\mathsf{T}}$.

Solutions cont'd

2. $y' = y$: A solution is $y(t) = \mathrm{e}^t$, or more generally $y(t) = c\,\mathrm{e}^t$ with $c \in \mathbb{R}$ (all with maximal domain $\mathbb{R}$).

   Using this family of solutions, we can solve any IVP $y' = y \wedge y(t_0) = y_0$: Just solve $c\,\mathrm{e}^{t_0} = y_0$ for c, i.e., $c = y_0\mathrm{e}^{-t_0}$ and $y(t) = y_0\mathrm{e}^{t-t_0}$.

   Are there other (maximal) solutions?
   No there aren't: A solution has the form $y(t) = c\,\mathrm{e}^t$ iff $t \mapsto y(t)\mathrm{e}^{-t}$ is constant. We have

$$(y(t)\mathrm{e}^{-t})' = y'(t)\mathrm{e}^{-t} + y(t)(-\mathrm{e}^{-t}) = \big((y'(t) - y(t)\big)\mathrm{e}^{-t} = 0,$$

   since $y = y'$. $\quad\Longrightarrow y(t)\mathrm{e}^{-t} = c \in \mathbb{R}$ is indeed a constant.

   These results imply that through any point (t_0, y_0) in the plane $\mathbb{R}^2$ there passes exactly one solution of $y' = y$. In other words, the graphs of the family of functions $y(t) = c\,\mathrm{e}^t$, $c \in \mathbb{R}$ (with domain $\mathbb{R}$) partition the plane.

   It also follows that the general solution of $\mathbf{y}' = \mathbf{y}$ is $\mathbf{y}(t) = (c_1\mathrm{e}^t, \ldots, c_m\mathrm{e}^t) = \mathrm{e}^t\mathbf{c}$ with $\mathbf{c} = (c_1, \ldots, c_m) \in \mathbb{R}^m$ or, linking it to the corresponding IVP, $\mathbf{y}(t) = \mathrm{e}^{t-t_0}\mathbf{y}_0$ with $\mathbf{y}_0 \in \mathbb{R}^m$.

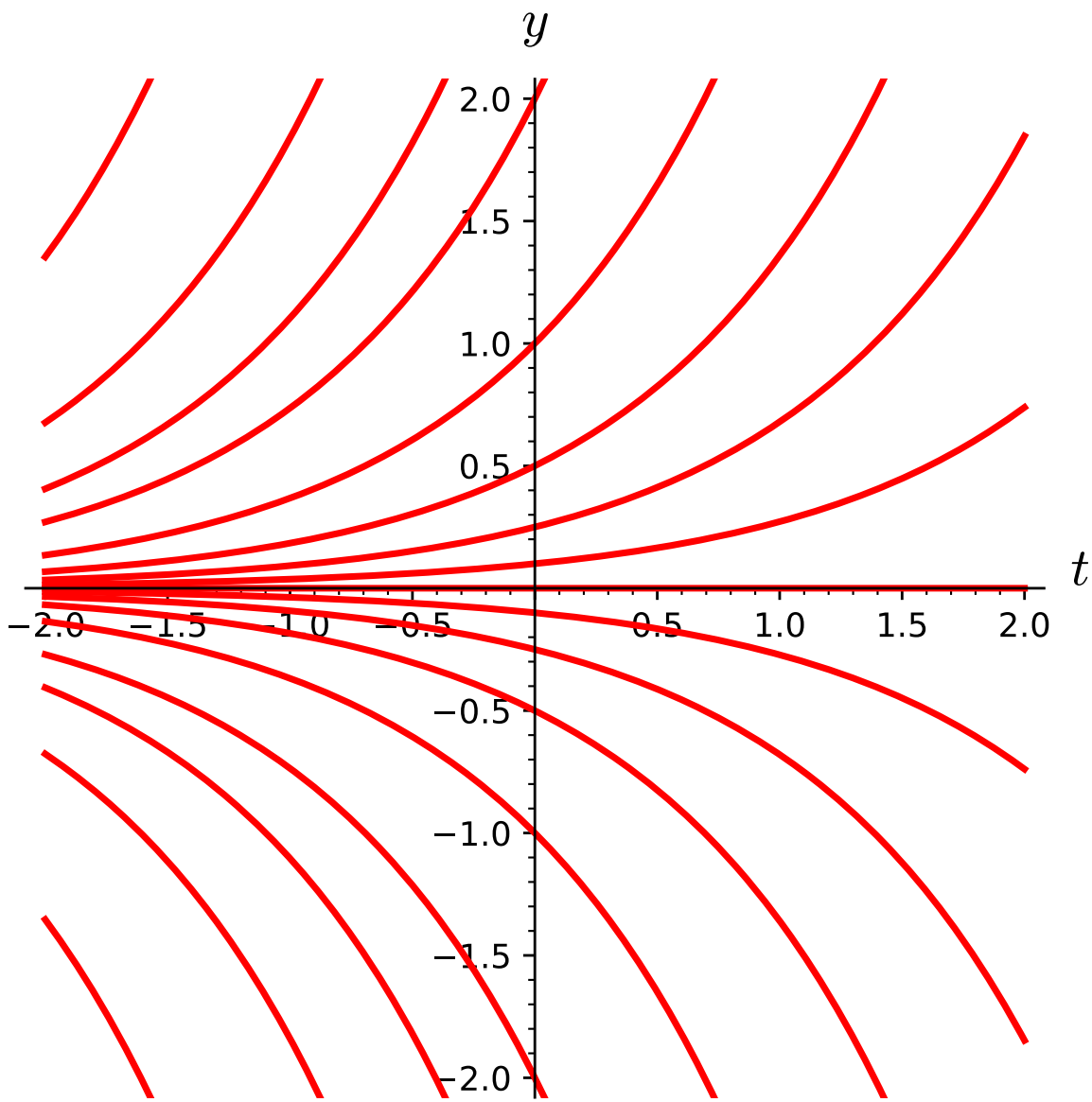


Figure: The solutions $y(t) = c\,\mathrm{e}^{t}$ for various constants c

Solutions cont'd

3. $y' = ay$: This is almost the same as Equation 2. The solutions are $y(t) = c\,\mathrm{e}^{at}$, $c \in \mathbb{R}$, and the corresponding IVP has a unique solution for each (t_0, y_0). That there are no more solutions is proved in the same way, working with $t \mapsto y(t)\mathrm{e}^{-at}$.

   For the vectorized form a keen guess is that the solution has the form $\mathbf{y}(t) = \mathrm{e}^{\mathbf{A}t}\mathbf{c}$ with $\mathbf{c} \in \mathbb{R}^n$. This is in fact true, as we will see later. For now let us only note that in order to make sense of this $\mathrm{e}^{\mathbf{A}t}$ should be an $n \times n$ matrix as well.

4. $y' = ay + b$: If y_1 and y_2 are solutions of $y' = ay + b$ then $y_1 - y_2$ is a solution of the associated *homogeneous* linear ODE $y' = ay$, since

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\big(y_1(t) - y_2(t)\big) &= \frac{\mathrm{d}y_1(t)}{\mathrm{d}t} - \frac{\mathrm{d}y_2(t)}{\mathrm{d}t} = a\,y_1(t) + b - a\,y_2(t) - b \\ &= a\big(y_1(t) - y_2(t)\big).\end{aligned}$$

   A particular solution of $y' = ay + b$ is $y(t) \equiv -b/a$, and hence the general solution is $y(t) = -b/a + c\,\mathrm{e}^{at}$, $c \in \mathbb{R}$ (with domain $\mathbb{R}$). From this we see as before that every IVP $y' = ay + b \wedge y(t_0) = y_0$, has a unique solution.

Solutions cont'd

5. $y' = 2ty$: The solutions are $y(t) = c\,\mathrm{e}^{t^2}$, $c \in \mathbb{R}$, and the observed general picture continues to prevail. In particular, we can use the function $t \mapsto y(t)\mathrm{e}^{-t^2}$ to show that there are no more solutions. (If $y(t)$ is a solution, this function must again be constant.) You should also compare this with the solution to Exercise H64 in Homework 12 of Calculus III.

   Clearly something more general works behind the scene in the examples discussed so far. All these are instances of 1st-order linear ODE's, and we shall present a unified treatment of this class of ODE's later.

Solutions cont'd

6. $y' = y^2$: It is not hard to guess a solution. One solution is $y(t) = -1/t$, since $(-1/t)' = 1/t^2 = (-1/t)^2$.
An even more obvious one is $y(t) \equiv 0$.

   Since $y' = G(y)$ with G not depending on t (such ODE's are called *autonomous*), we can make a translation in the argument t to obtain further solutions:

$$\frac{\mathrm{d}}{\mathrm{d}t}\, y(t-c) = y'(t-c) = y(t-c)^2,$$

   i.e., $t \mapsto y(t-c)$ is a solution for any $c \in \mathbb{R}$ (provided that $y(t)$ is). In the case under consideration this gives the family of solutions

$$y(t) = -\frac{1}{t-c} = \frac{1}{c-t}, \quad c \in \mathbb{R}.$$

   Since we can solve $\frac{1}{c-t_0} = y_0$ uniquely for c if (t_0, y_0) is not on the t-axis (i.e., $y_0 \neq 0$), the solutions found so far (including $y \equiv 0$) partition the plane $\mathbb{R}^2$, and any IVP $y' = y^2 \wedge y(t_0) = y_0$ has a unique solution within this family.

Solutions cont'd

6. (cont'd)
In contrast with the preceding examples, $y(t) = 1/(c-t)$ is not defined for all $t \in \mathbb{R}$. In fact, according to our definition of "solution of an ODE" we rather have two maximal solutions corresponding to a fixed c,

$$y_1(t) = 1/(c-t), \quad t \in (-\infty, c),$$
$$y_2(t) = 1/(c-t), \quad t \in (c, +\infty).$$

Unique solvabilty of the corresponding IVP is not affected by this change of viewpoint. (The solutions $y_1(t)$ are obtained for initial values $y(t_0) > 0$, the solutions $y_2(t)$ for $y(t_0) < 0$.)

Now we show that there are no further (maximal) solutions.

Firstly, suppose $y : I \to \mathbb{R}$ is a solution with $y(t) \neq 0$ for $t \in I$. Then $y'(t)/y^2(t) = 1$ on I. Fixing some $t_0 \in I$ and integrating, we get

$$t - t_0 = \int_{t_0}^{t} \frac{y'(\tau)}{y(\tau)^2}\,\mathrm{d}\tau = \int_{y(t_0)}^{y(t)} \frac{\mathrm{d}\eta}{\eta^2} = \left[-\frac{1}{\eta}\right]_{y(t_0)}^{y(t)} = \frac{1}{y(t_0)} - \frac{1}{y(t)}.$$

Solutions cont'd

6. (cont'd)
This holds for $t \in I$ and can be solved for $y(t)$:

$$y(t) = \frac{1}{t_0 + y(t_0)^{-1} - t} = \frac{1}{c - t} \quad \text{with } c = t_0 + y(t_0)^{-1}.$$

Hence I is contained in either $(-\infty, c)$ or $(c, +\infty)$, and $y(t)$ coincides with one of the corresponding solutions $y_1(t)$ or $y_2(t)$ on I.

Secondly suppose there exists $t_0, t_1 \in I$ such that $y(t_0) \neq 0$, $y(t_1) = 0$. We may assume $t_1 < t_0$ (the case that $y(t)$ "branches from $y \equiv 0$", the other case being similar) and $y(t) > 0$ for $t \in (t_1, t_0]$ (since $y(t)$ is continuous, it has a largest zero in I, which we can choose as t_1).
On one hand we now have $\lim_{t \downarrow t_1} y(t) = y(t_1) = 0$. On the other hand, the first case applies to $t \in (t_1, t_0]$ and yields

$$y(t) = \frac{1}{t_0 + y(t_0)^{-1} - t} \geq \frac{1}{t_0 + y(t_0)^{-1} - t_1} > 0 \quad \text{for } t \in (t_1, t_0].$$

This contradiction shows that the second case does not occur.

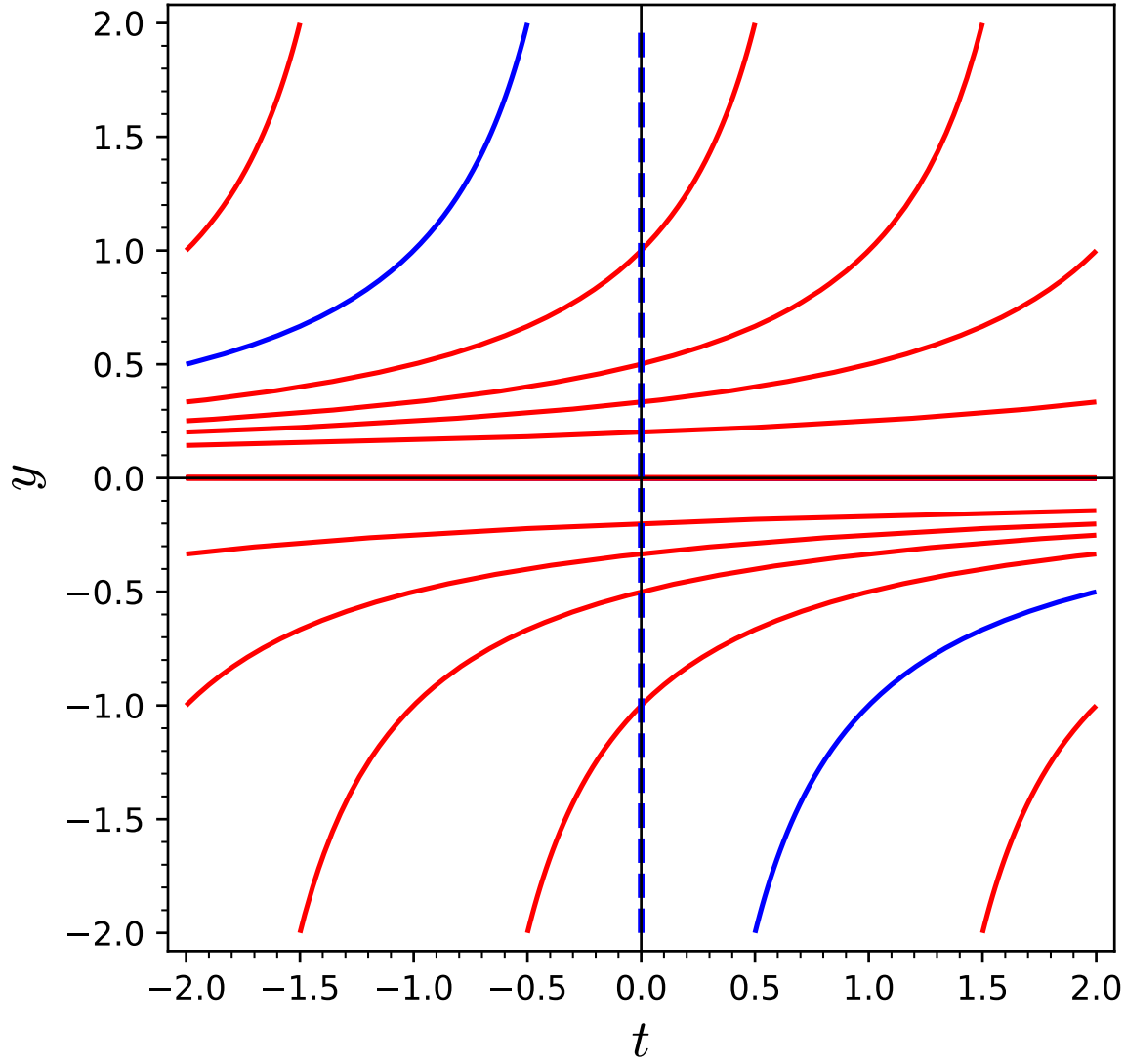


Figure: The solutions $y(t) = 1/(c - t)$ for various constants c (including ∞), with the one for $c = 0$ and its asymptote colored blue

Remark on Notation

No-name notation $y(t)$ for solutions of ODE's can easily cause confusion, when more than one solution is considered, e.g., when we say "all solutions are horizontal shifts of a particular solution". At least in such cases we should follow good mathematical practice and specify solutions, which are mathematical functions, in full—like this:

Three particular maximal solutions of $y' = y^2$ are

$$f_1 : (-\infty, 0) \to \mathbb{R},\ t \mapsto -1/t,$$
$$f_2 : (0, +\infty) \to \mathbb{R},\ t \mapsto -1/t,$$
$$f_3 : \mathbb{R} \to \mathbb{R},\ t \mapsto 0.$$

Every further solution arises from one of these solutions by a horizontal shift and/or restriction to a subinterval, i.e., it is zero or has the form

$$g : I \to \mathbb{R},\ t \mapsto -1/(t - c)$$

for some $c \in \mathbb{R}$ and some interval I contained in either $(-\infty, c)$ or $(c, +\infty)$.

Solutions cont'd

7. $y' = \sqrt{|y|}$: Again $y \equiv 0$ is an obvious solution, and further solutions may be guessed: Since $y(t) = t^2$ satisfies $y'(t) = 2t = 2\sqrt{y}$ on $[0, +\infty)$, we can scale this "approximate solution" by an appropriate constant, viz. $\frac{1}{4}$, to obtain the real solution $y(t) = \frac{1}{4}t^2$, $t \in [0, +\infty)$. Note that both $y(t) \equiv 0$ and $y(t) = \frac{1}{4}t^2$ solve the IVP $y' = \sqrt{|y|}$, $y(0) = 0$.

   Since $y' = \sqrt{|y|}$ is autonomous, $y(t) = \frac{1}{4}(t - c)^2$, $t \in [c, +\infty)$, is a solution for any $c \in \mathbb{R}$, and so is $y(t) = -\frac{1}{4}(t - c)^2$, $t \in (-\infty, c]$.

   Since on the t-axis all solutions have horizontal tangents ($y = 0 \Longrightarrow y' = \sqrt{|y|} = 0$), we can "glue together" the different types of solutions and obtain the following 2-parameter family of solutions with domain $\mathbb{R}$:

$$y(t) = \begin{cases} -\frac{1}{4}(t - c_1)^2 & \text{if } t \le c_1, \\ 0 & \text{if } c_1 \le t \le c_2, \\ \frac{1}{4}(t - c_2)^2 & \text{if } t \ge c_2. \end{cases}$$

   Here $-\infty \le c_1 \le c_2 \le +\infty$; equality in either of these indicates that the corresponding section is omitted.

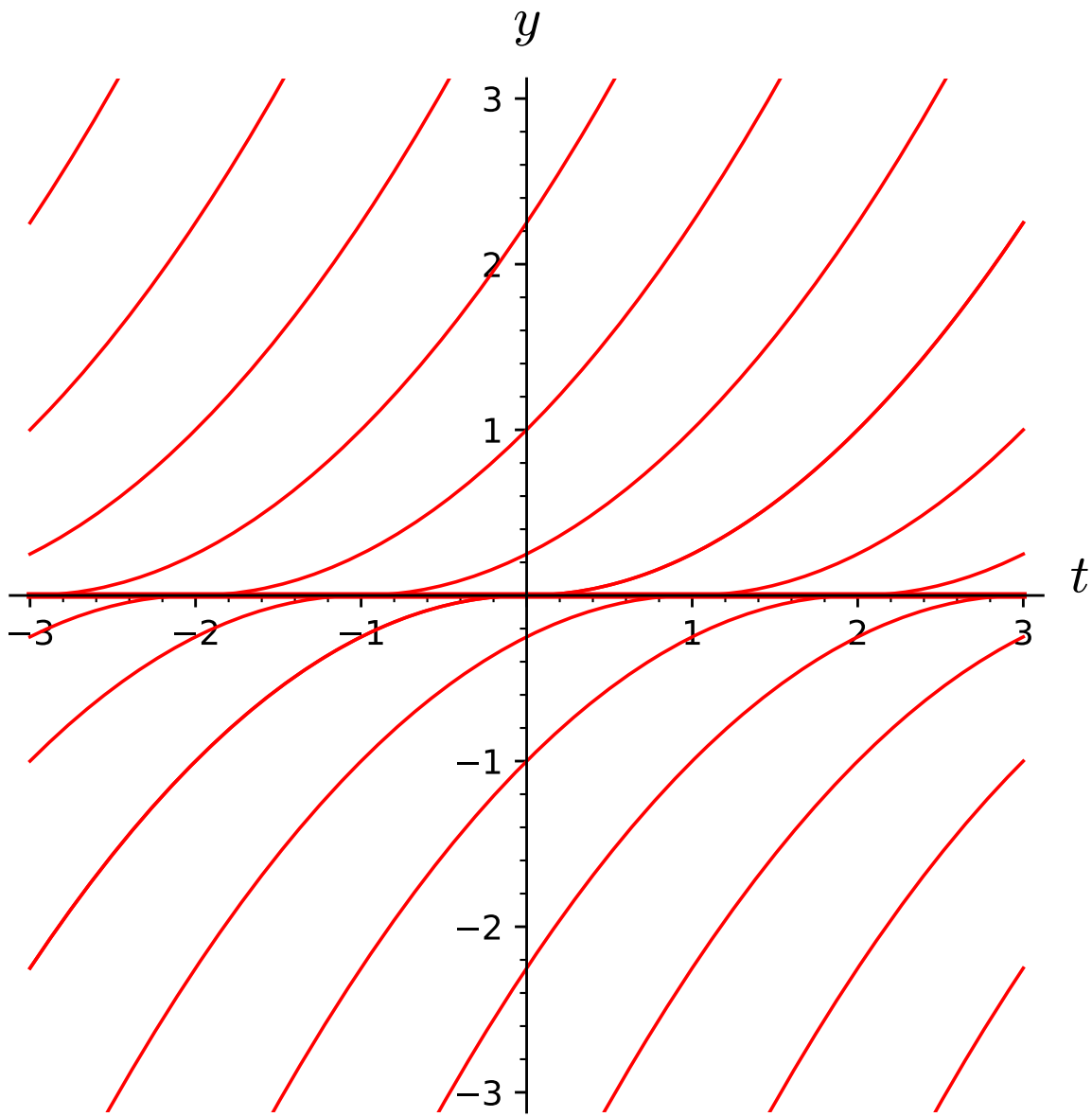


Figure: Solutions of $y' = \sqrt{|y|}$

Solutions cont'd

7. (cont'd)
   It can be shown that these are all maximal solutions.

   Brief sketch: For solutions not meeting the t-axis use a similar argument as for $y' = y^2$. A solution meeting the t-axis must meet it in an interval, since it can only flow in from below and branch off above, hence never return. Since the solutions are continuous, the interval must be closed. Denote it by $[c_1, c_2]$ and show that the solution is as stated on the previous slide.

Solutions cont'd

8. $(e^{x-y} + 2x - e)\,dx - (e^{x-y} + 2y)\,dy = 0$

   By a solution of $M(x,y)\,dx + N(x,y)\,dy = 0$ we mean a smooth plane curve $\gamma(t) = \big(x(t), y(t)\big)$, $t \in I$, satisfying

   $$M\big(x(t),y(t)\big)x'(t) + N\big(x(t),y(t)\big)y'(t) = 0 \quad \text{for all } t \in I;$$

   equivalently,

   $$\begin{pmatrix} M\big(\gamma(t)\big) \\ N\big(\gamma(t)\big) \end{pmatrix} \cdot \gamma'(t) = 0 \quad \text{for all } t \in I.$$

   Geometrically this says that the tangents to γ at every regular point should be orthogonal to the vector field (M,N) (the vector field corresponding to $\omega = M\,dx + N\,dy$) at this point.

   In the case under consideration ω is exact, $\omega = df$ for $f(x,y) = e^{x-y} + x^2 - ex - y^2$, and hence the parametrized contours $e^{x-y} + x^2 - ex - y^2 = c$, $c \in \mathbb{R}$, of f provide solutions to $(e^{x-y} + 2x - e)\,dx - (e^{x-y} + 2y)\,dy = 0$.

   Accordingly, ODE's of the special form $f_x\,dx + f_y\,dy = 0$ are said to be *exact*.

Solutions cont'd

8. The Implicit Function Theorem gives that every IVP

$$(e^{x-y} + 2x - e)\,dx - (e^{x-y} + 2y)\,dy = 0,$$
$$\gamma(t_0) = (x_0, y_0) \neq \left(\frac{e - e^{e/2}}{2}, -\frac{e^{e/2}}{2}\right)$$

(the unique critical point of f) has a solution. The point curve $\gamma(t) \equiv \left(\frac{e-e^{e/2}}{2}, -\frac{e^{e/2}}{2}\right)$, as well as any other point curve $\gamma(t) \equiv (x_0, y_0)$, trivially satisfies the ODE but isn't counted as a solution, since it is not smooth.

Solutions are highly non-unique, since we can choose the parametrization of the contours freely.

Further, the Implicit Function Theorem gives that at every point (x_0, y_0) with $f_y(x_0, y_0) = -(e^{x_0-y_0} + 2y_0) \neq 0$ the corresponding contour admits locally a parametrization $y(x)$, which must be a solution of the ODE

$$y' = \frac{dy}{dx} = -\frac{f_x}{f_y} = \frac{e^{x-y} + 2x - e}{e^{x-y} + 2y}.$$

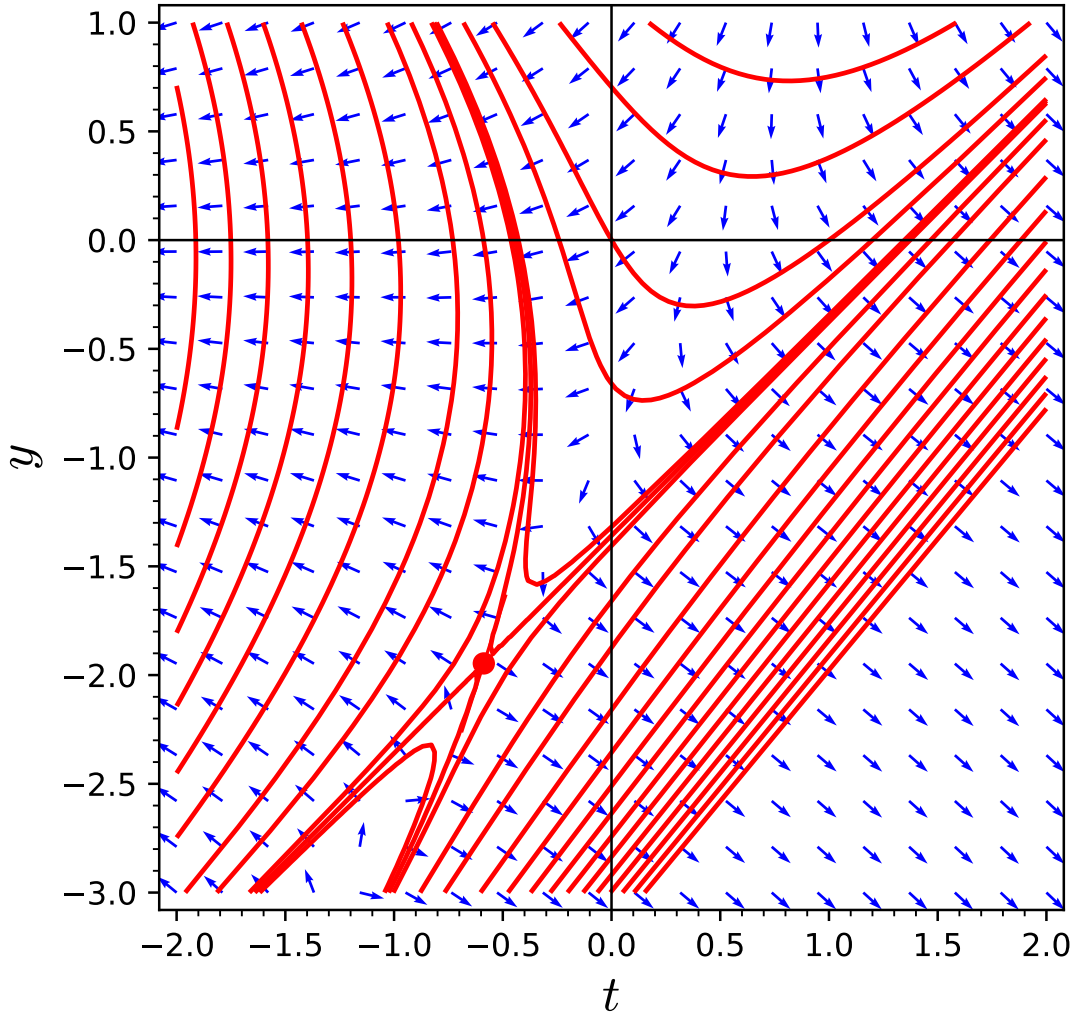


Figure: Solutions of $(e^{x-y} + 2x - e)\,dx - (e^{x-y} + 2y)\,dy = 0$ in implicit form (red), and the gradient field of $f(x, y) = e^{x-y} + x^2 - ex - y^2$ normalized to unit length (blue) The critical point $\left(\frac{e-e^{e/2}}{2}, -\frac{e^{e/2}}{2}\right) \approx (-0.59, -1.95)$ is on two intersecting solution curves.

Remark on the plot

If you look carefully at the two solution curves through the critical point $\mathbf{p}_0 = (x_0, y_0)$, you can see that they have been approximated by line segments, the reason being that the contour plot function of SageMath doesn't draw the correct picture near $\mathbf{p}_0$.

The tangent directions at $\mathbf{p}_0$ of the two curves can be found from the Taylor approximation

$$f(\mathbf{p}_0 + \mathbf{h}) = f(\mathbf{p}_0) + \tfrac{1}{2}\mathbf{h}^{\mathsf{T}}\mathbf{H}_f(\mathbf{p}_0)\mathbf{h} + o(|\mathbf{h}|^2) \quad \text{for } \mathbf{h} \to \mathbf{0}.$$

From this one sees using *Morse's Lemma* that the contour of f through $\mathbf{p}_0$ is, after translation of $\mathbf{p}_0$ into the origin, locally well-approximated by the 0-contour of the Hesse quadratic form

$$\begin{aligned} q(\mathbf{h}) = \mathbf{h}^{\mathsf{T}}\mathbf{H}_f(\mathbf{p}_0)\mathbf{h} &= (\mathrm{e}^{e/2} + 2)h_1^2 - 2\mathrm{e}^{e/2}h_1h_2 + (\mathrm{e}^{e/2} - 2)h_2^2 \\ &= (\mathrm{e}^{e/2} - 2)(h_2 - h_1)\left(h_2 - \tfrac{\mathrm{e}^{e/2}+2}{\mathrm{e}^{e/2}-2}h_1\right), \end{aligned}$$

which is the union of two lines (expressing the fact that $\mathbf{p}_0$ is a saddle point of f). The said tangent directions are the directions of these lines, i.e., have slopes 1 and $\frac{\mathrm{e}^{e/2}+2}{\mathrm{e}^{e/2}-2} \approx 3.11$.

Solutions cont'd

9. $y' = -x/y$

   This ODE can be written as $y'y + x = 0$ and integrated to yield $\frac{y^2}{2} + \frac{x^2}{2} = C = \frac{C'}{2} \in \mathbb{R}$. The solutions are therefore the half-circle parametrizations

$$y(x) = \pm\sqrt{r^2 - x^2}, \quad x \in (-r, r) \quad \text{(with } r > 0\text{)}.$$

   Every IVP $y' = -x/y \wedge y(x_0) = y_0 \neq 0$ has a unique solution, as is easily seen.

   Alternatively, we can rewrite

$$y' = \frac{\mathrm{d}y}{dx} = -\frac{x}{y} \quad \text{as} \quad x\,\mathrm{d}x + y\,\mathrm{d}y = 0,$$

   which is exact with anti-derivative $f(x, y) = \frac{x^2+y^2}{2}$. This gives whole-circles centered as $(0, 0)$, which are the contours of f, as implicit solutions. The exceptional role of the x-axis, visible in the original explicit ODE, has gone away.

Solutions cont'd

10. $y'' + y = 0$

Two particular solutions of $y'' = -y$ are $y_1(t) = \cos t$ and $y_2(t) = \sin t$ (with domain $\mathbb{R}$). For $A, B \in \mathbb{R}$, since

$$\begin{aligned}(A\cos t + B\sin t)'' &= A(\cos t)'' + B(\sin t)'' = -A\cos t - B\sin t \\ &= -(A\cos t + B\sin t),\end{aligned}$$

we obtain further solutions $y(t) = A\cos t + B\sin t$ (also with domain $\mathbb{R}$).

Since $y(0) = A\cos 0 + B\sin 0 = A$,
$y'(0) = -A\sin 0 + B\cos 0 = B$, there is exactly one solution of any IVP $\quad y'' = -y \wedge y(0) = A \wedge y'(0) = B$ $(A, B \in \mathbb{R})$.

Similarly, one can show that the IVP

$$y'' = -y, \quad y(t_0) = y_0,\ y'(t_0) = y_1 \quad (t_0, y_0, y_1 \in \mathbb{R}).$$

has exactly one solution of the said form
$y(t) = A\cos t + B\sin t$ (Exercise). This means that the graphs of the maps $t \mapsto \big(y(t), y'(t)\big)$, or traces of the curves $t \mapsto \big(t, y(t), y'(t)\big)$, with $y(t) = A\cos t + B\sin t$, $A, B \in \mathbb{R}$, partition the space $\mathbb{R}^3$. (Can you imagine that?).

Solutions cont'd

10 (con't)

Now we show that every solution $y(t)$ of $y'' = -y$ has the form $A\cos t + B\sin t$.

A trick that will do the job is considering the function $t \mapsto y(t)^2 + y'(t)^2$. Since

$$(y^2 + y'^2)' = 2yy' + 2y'y'' = 2y'(y + y'') = 0,$$

this function must be constant for any solution of $y'' = -y$.

Now let $y(t)$ be any solution and set $A = y(0)$, $B = y'(0)$. We have already a solution in our family with these initial values, viz. $z(t) = A\cos t + B\sin t$. The difference $d(t) = y(t) - z(t)$ is then also a solution and satisfies $d(0) = d'(0) = 0$. Since $d^2 + d'^2$ is constant, we have

$$d(t)^2 + d'(t)^2 = d(0)^2 + d'(0)^2 = 0^2 + 0^2 = 0.$$

This obviously implies $d(t) \equiv 0$, i.e., $y(t) = z(t) = A\cos t + B\sin t$.

Exercise

We have considered the ODE $y' = -x/y$ as an example. Actually there are four ODE's $y' = \pm x/y$ and $y' = \pm y/x$, which look very similar. Draw direction fields for the other three ODE's and determine their solutions in both implicit and explicit form (if possible).

Exercise

Let $t_0, y_0, y_1 \in \mathbb{R}$. Show that the IVP

$$y'' = -y, \quad y(t_0) = y_0,\ y'(t_0) = y_1$$

has a unique solution.

Supplementary Remarks on the Material in [BDM17], Chapter 1

Direction fields

Let $f\colon D \to \mathbb{R}$, $D \subseteq \mathbb{R}^2$, be a function

Solving the 1st-order ODE $y' = f(t,y)$ amounts to finding a function $y = y(t)$, defined on some interval $I \subseteq \mathbb{R}$, and such that for all $t \in I$

1. $\big(t, y(t)\big) \in D$;
2. the slope of the graph of y at the point $\big(t, y(t)\big)$ equals $f\big(t, y(t)\big)$. Alternatively, the tangent direction to the graph at $\big(t, y(t)\big)$ is represented by the vector $\big(1, f(t, y(t))\big)$.

We can illustrate this by attaching to sample points $(t,y) \in D$ a small line segment with direction $\big(1, f(t,y)\big)$ (or a positive multiple of this vector). This is called a *slope field* or *direction field* of $y' = f(t,y)$. Solving $y' = f(t,y)$ graphically then amounts to finding a function $y = y(t)$ such that the tangent direction of the graph of y at every sample point encountered is given by the corresponding line segment.

Examples

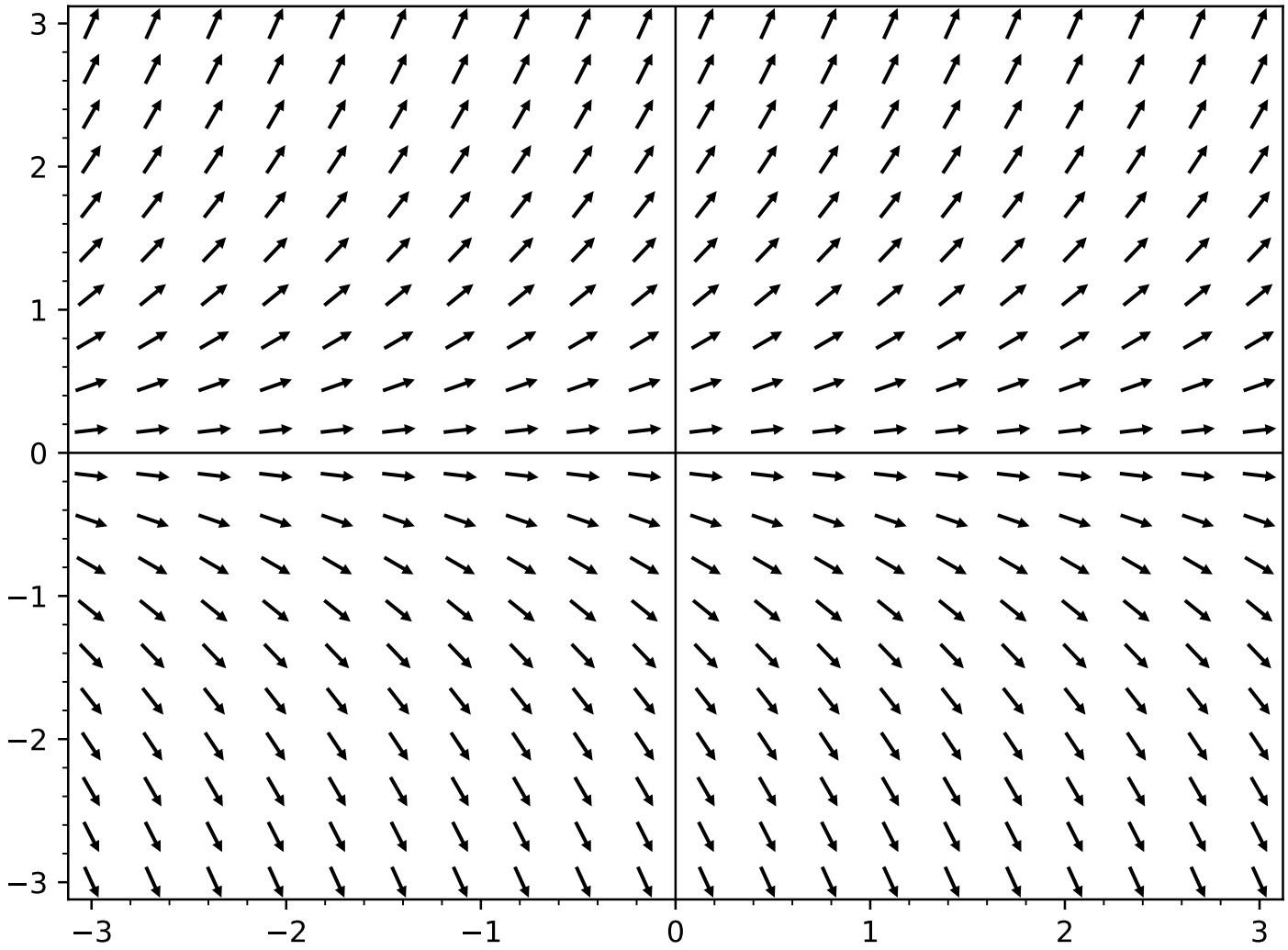


Figure: Direction field of $y' = y$

Examples Cont'd

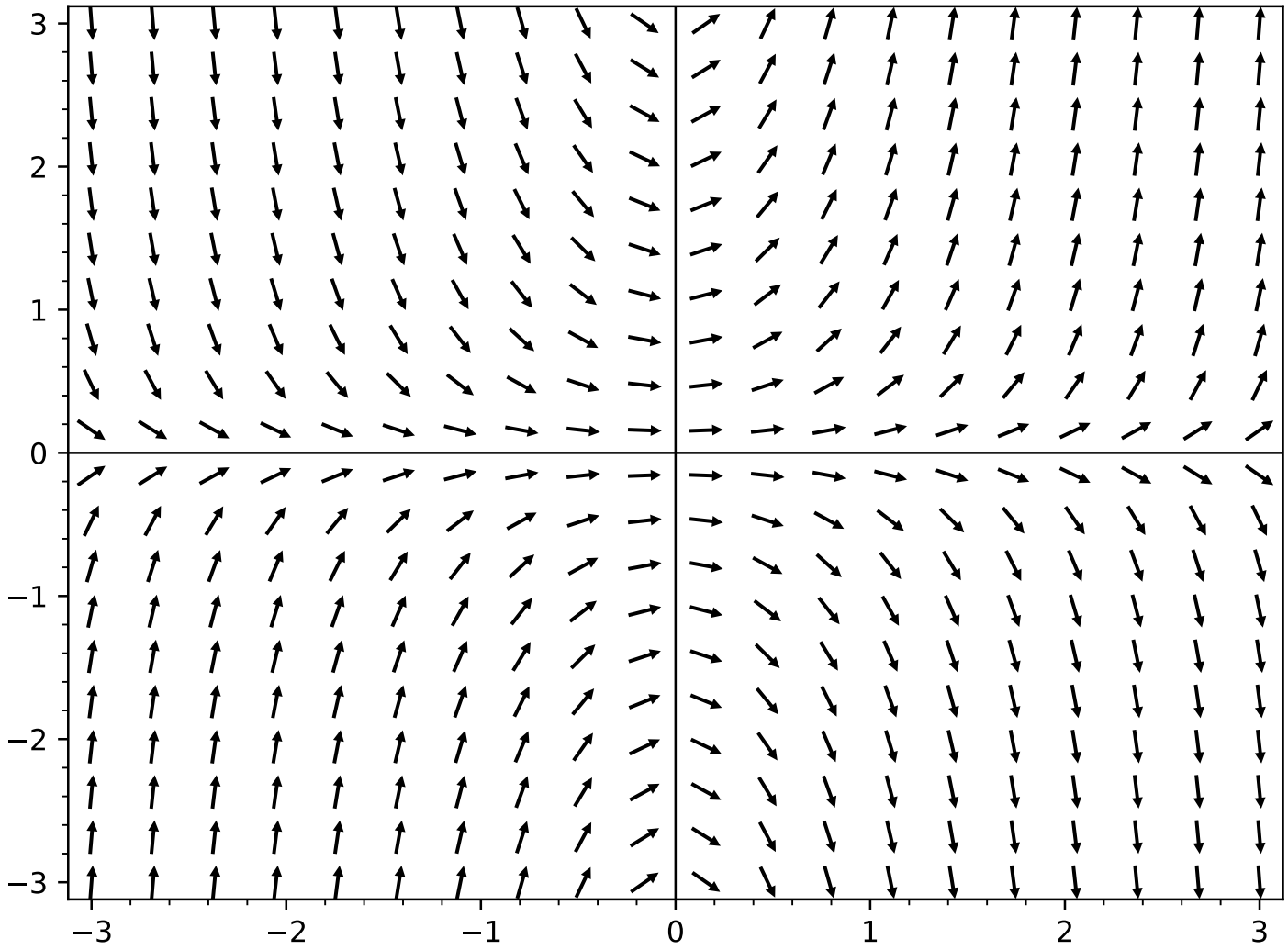


Figure: Direction field of $y' = 2ty$

Examples Cont'd

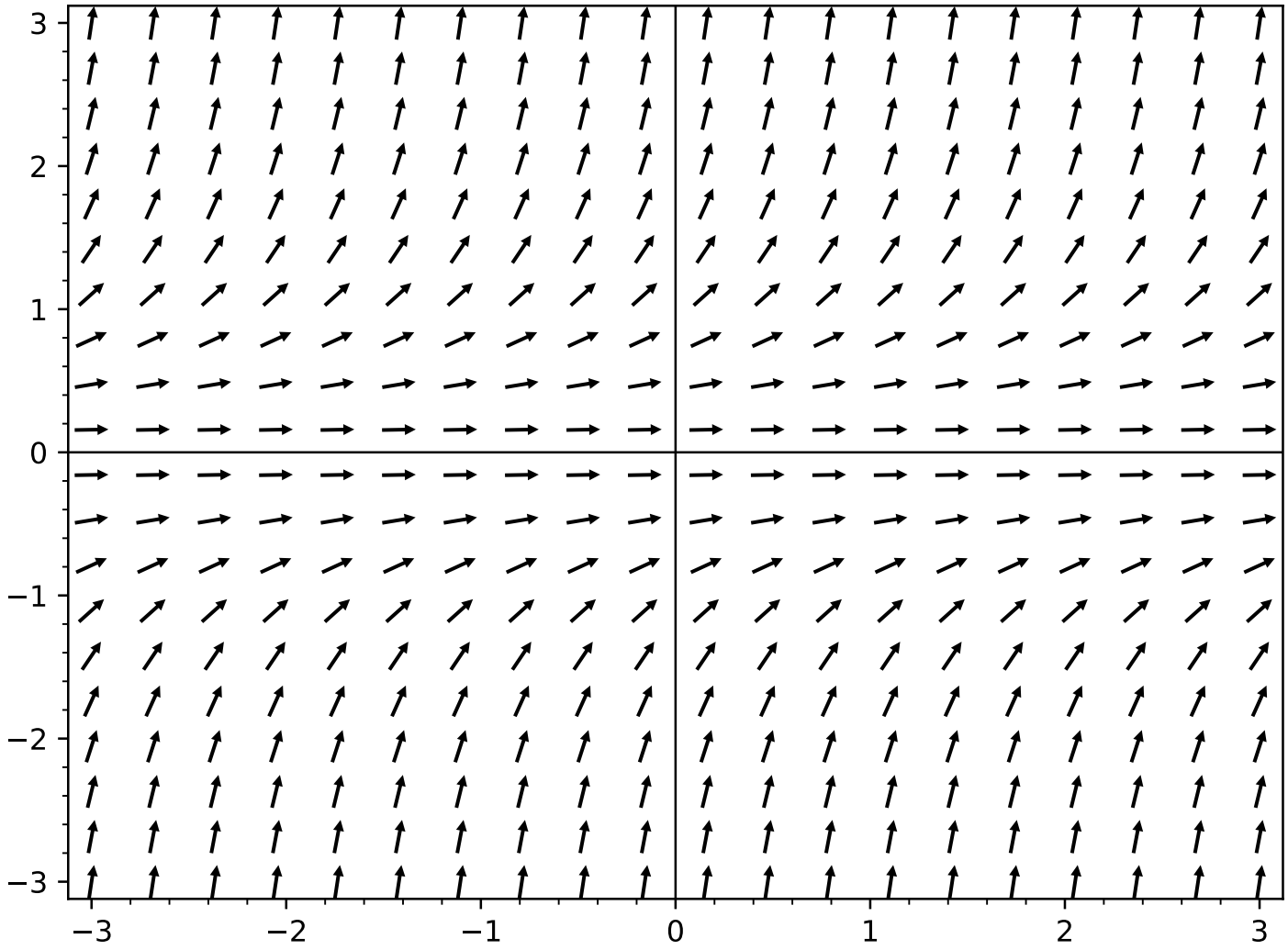


Figure: Direction field of $y' = y^2$

Examples Cont'd

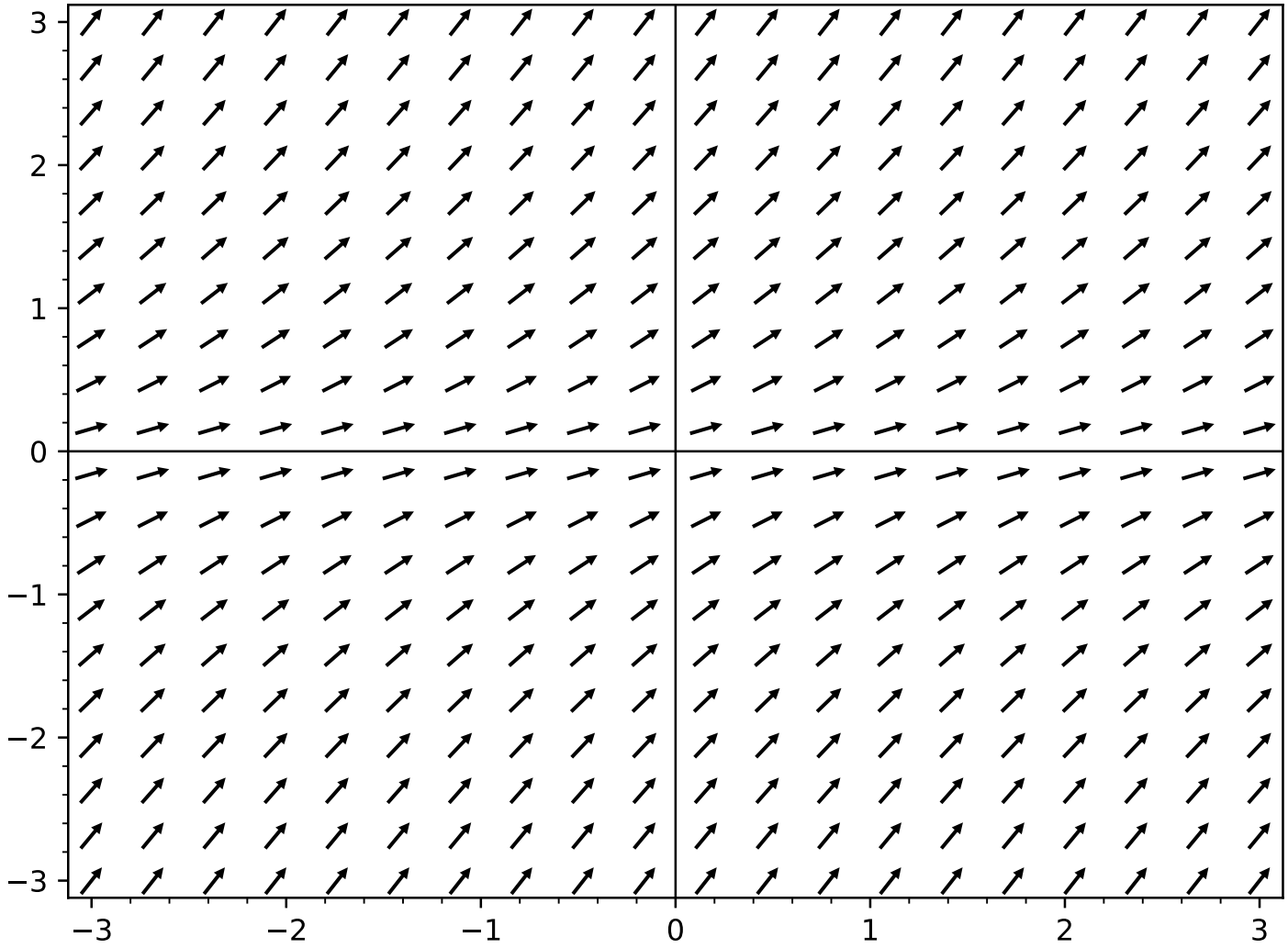


Figure: Direction field of $y' = \sqrt{|y|}$

Examples Cont'd

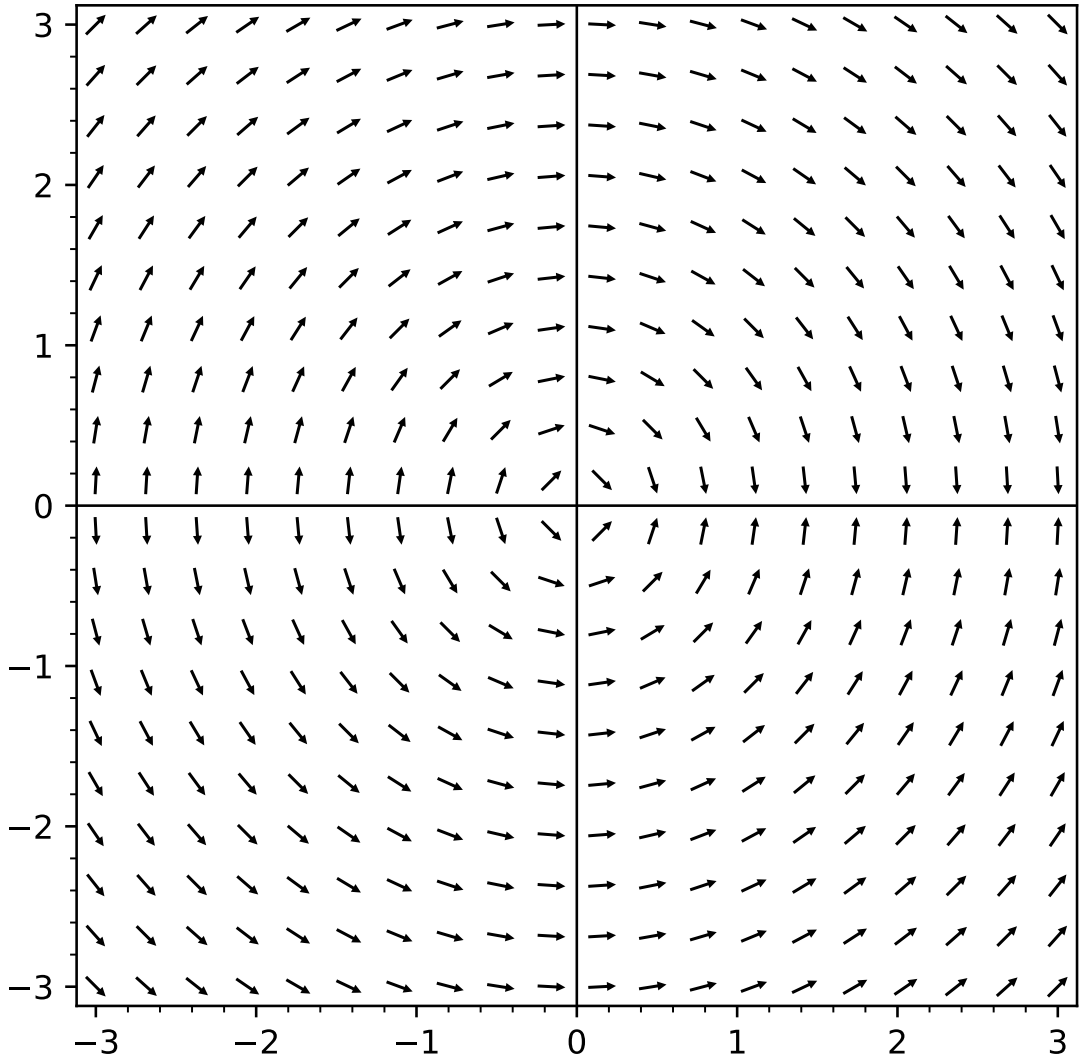


Figure: Direction field of $y' = -t/y$

Examples Cont'd

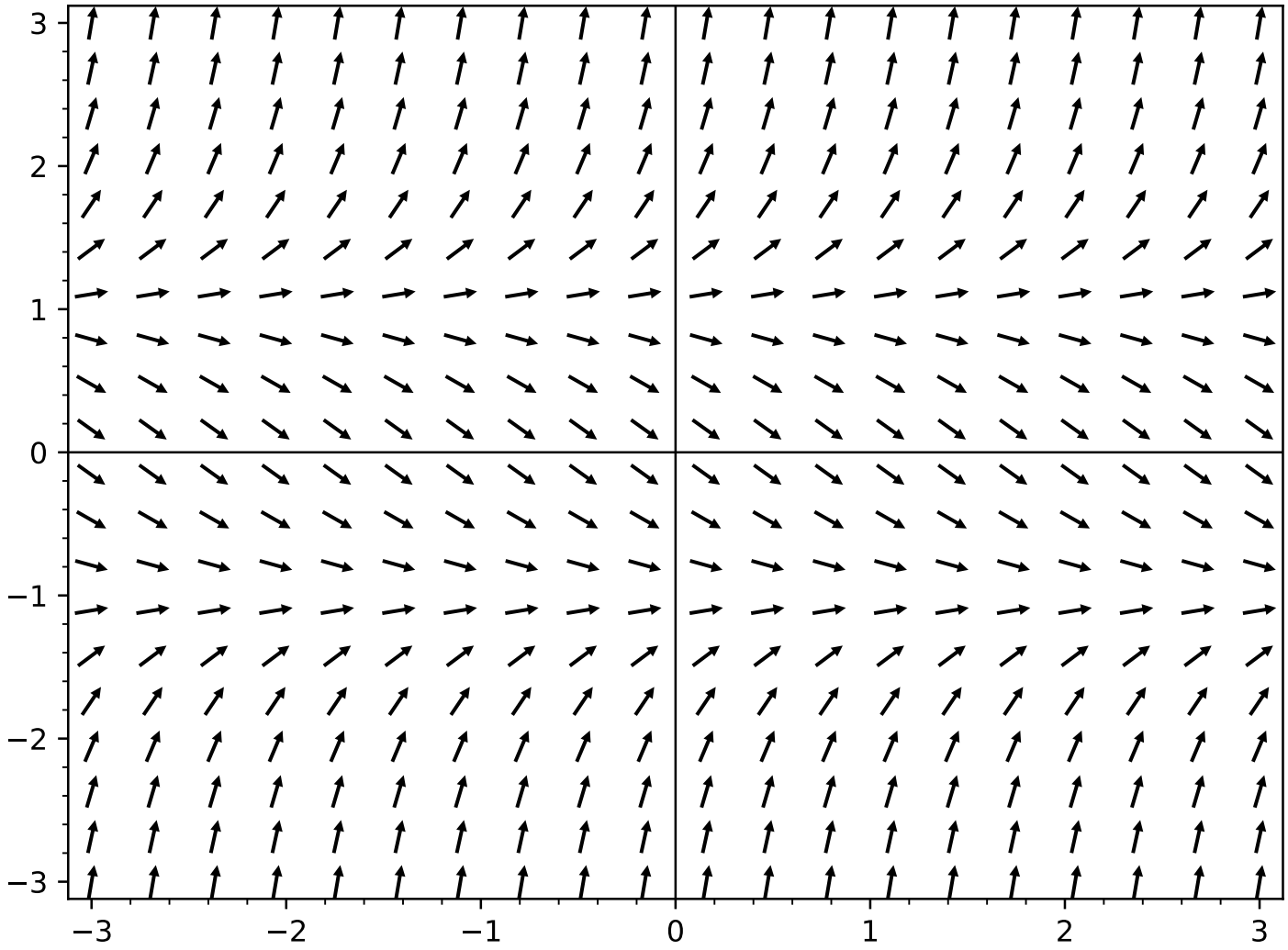


Figure: Direction field of $y' = y^2 - 1$

Mathematical Modelling with ODE's and PDE's

Differential equations are ubiquitous in Physics and Engineering, because fundamental laws of Physics can be expressed in terms of differential equations, and hence physicists and engineers need to solve such equations in order to describe the quantities involved in a physical process or system.

We consider, without actually solving the differential equations, a few examples.

Modeling with Initial Value Problems

Falling objects

$v(t)$ denotes the speed of the falling object;
initial condition $v(0) = 0$ (object is released at time $t = 0$);
Newton's 2nd Law $F(t) = m\,a(t) = m\,\mathrm{d}v/\mathrm{d}t$ and assumptions on the *drag force* F_D due to air resistance give the following models:

$$m\frac{\mathrm{d}v}{\mathrm{d}t} = mg \qquad F_D = 0 \qquad \text{(no air resistance)}$$

$$m\frac{\mathrm{d}v}{\mathrm{d}t} = mg - k_1 v \qquad F_D = k_1 v \qquad \text{(very small objects)}$$

$$m\frac{\mathrm{d}v}{\mathrm{d}t} = mg - k_2 v^2 \qquad F_D = k_2 v^2 \qquad \text{(most common case)}$$

The first model is realistic only for short distances, the second for very small objects like dust particels, and the third applies in all other cases (assuming that air density and gravitational acceleration are approximately the same as on the surface of the earth).

Modeling with Initial Value Problems Cont'd

Oscillating pendulum

$\theta(t)$ denotes the angle between the rod at time t makes with the vertical direction and L the length of the rod;
initial condition $\theta(0) = \theta_0$ (angle when the pendulum is released);
Newton's 2nd Law gives the following ODE for $\theta(t)$:

$$mL\frac{\mathrm{d}^2\theta}{\mathrm{d}t^2} = -mg\sin\theta$$

Modeling with Initial Value Problems Cont'd

Predator-Prey Models

$x(t)$, $y(t)$ denote the population sizes of two species. We assume that the second species (the *predator*) preys on the first species (the *prey*), while the prey lives on a different source of food;
initial population size $x(0) = x_0$, $y(0) = y_0$;
reasonable assumptions on the reproduction rates of the two species and their interaction lead to the following system of ODE's:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = ax - \alpha xy,$$
$$\frac{\mathrm{d}y}{\mathrm{d}t} = -cy + \gamma xy, \quad \text{with constants } a, \alpha, c, \gamma > 0.$$

These are known as the *Lotka-Volterra equations*.

Modeling with Boundary Value Problems

Heat Conduction

$u(x,t)$ denotes the temperature in a thin solid bar of length L at distance x from one end and at time $t \geq 0$.
The temperature variation is subject to

$$\alpha^2 u_{xx}(x,t) = u_t(x,t) \quad \text{for } 0 < x < L,\ t > 0.$$

(Heat conduction equation)

Boundary conditions:

$$u(x,0) = f(x), \quad u(0,t) = T_1, \quad u(L,t) = T_2.$$

These express the requirements that the initial ($t = 0$) temperature distribution in the bar is some known function $f(x)$, and the ends of the bar are kept at constant temperatures T_1 resp. T_2.

Modeling with Boundary Value Problems Cont'd

Vibrating String

$u(x,t)$ denotes the vertical displacement of an elastic string of length L from its horizontal equilibrium position at distance x from one end and at time $t \geq 0$.
The string's vibration is subject to

$$a^2\, u_{xx}(x,t) = u_{tt}(x,t) \quad \text{for } 0 < x < L,\ t > 0.$$

(Wave equation)

Boundary conditions:

$$u(x,0) = f(x), \quad u(0,t) = u(L,t) = 0.$$

These express the requirements that the string is released at time $t = 0$ from some known position $f(x)$ (i.e., a plucked guitar string) and is fixed at both ends.

Further Notes on Modeling with ODE's and PDE's

- As in the case of falling objects, it is often not clear which mathematical model is appropriate. Solving the model equations and making predictions based on the results must be checked against real-world data.
- Even when using a well-established model there is the problem of estimating numerically the corresponding physical parameters involved. If the model is very sensitive in this regard, small inaccuracies in the input data may lead to completely false predictions by the model.
- ODE's and PDE's used in modeling have at least one undetermined parameter, because physical quantities are relative to the unit of measurement used. For example, the ODE $v'(t) = g$ describing a falling object over a short time has the parameter g, which can take different real values depending on the choice of units, e.g., $g = 9.81\ [\mathrm{m/s^2}]$ vs. $g = 127\,000\ [\mathrm{km/h^2}]$.

Math 285
Introduction to Differential Equations

Thomas Honold

ZJU-UIUC Institute

Spring Semester 2024

Outline

Today's Lecture: First-Order Linear Equations

The Linear Case

Definition

An (explicit) first-order *linear* ODE has the form

$$y' = a(t)y + b(t).$$

If $b(t) \equiv 0$, the linear ODE is called *homogeneous*; if $b(t) \neq 0$ for at least one t, it is called *inhomogeneous*.

Compare this with the theory of linear recurring sequences.

Theorem (homogeneous case)

If $a(t)$ is continuous, the general solution of $y' = a(t)y$ is given by

$$y(t) = c\,\mathrm{e}^{\int_{t_0}^{t} a(s)\,\mathrm{d}s} = y(t_0)\mathrm{e}^{\int_{t_0}^{t} a(s)\,\mathrm{d}s}, \quad c \in \mathbb{R}.$$

The domain of $y(t)$ is that of $a(t)$. (If the domain T of $a(t)$ is not an interval, there exists a solution of the stated form on every connected component of T.)

Proof.

The chain rule and the Fundamental Theorem of Calculus give

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(c\,\mathrm{e}^{\int_{t_0}^{t} a(s)\,\mathrm{d}s}\right) = c\,\mathrm{e}^{\int_{t_0}^{t} a(s)\,\mathrm{d}s}\cdot\frac{\mathrm{d}}{\mathrm{d}t}\int_{t_0}^{t} a(s)\,\mathrm{d}s = c\,\mathrm{e}^{\int_{t_0}^{t} a(s)\,\mathrm{d}s}\cdot a(t),$$

showing that $t \mapsto c\,\mathrm{e}^{\int_{t_0}^{t} a(s)\,\mathrm{d}s}$ is a solution of $y' = a(t)y$.

Now let $y(t)$ be any solution and consider the function $f(t) = y(t)\mathrm{e}^{-A(t)}$, where $A(t)$ is an antiderivative of $a(t)$, say $A(t) = \int_{t_0}^{t} a(s)\,\mathrm{d}s$.

$$f'(t) = y'(t)\mathrm{e}^{-A(t)} + y(t)\mathrm{e}^{-A(t)}\big(-a(t)\big) = \big(y'(t) - a(t)y(t)\big)\mathrm{e}^{-A(t)} \equiv 0$$

$\Longrightarrow f(t) = c$ is constant, and hence $y(t) = c\,\mathrm{e}^{A(t)}$ as claimed. (The choice of $A(t)$ does not matter, since the additive constant K involved in the choice turns into a positive constant e^{-K}, which is "eaten up" by c.) □

The Inhomogeneous Case

Solved by variation of parameters

Variation of parameters

Idea: The homogeneous ODE $y' = a(t)y$ is solved by $y(t) = c\,\mathrm{e}^{A(t)}$. In order to solve $y' = a(t)y + b(t)$, make $c = c(t)$ variable; that is we set $y_p(t) = c(t)\mathrm{e}^{A(t)} = c(t)y_h(t)$, where $y_h(t)$ denotes a solution of the homogeneous ODE.

$$y_p' = (cy_h)' = c'y_h + cy_h' = c'y_h + cay_h = a(cy_h) + b \iff c' = by_h^{-1}$$

Theorem

Suppose $a(t)$ and $b(t)$ are continuous.

1. *A particular solution of $y' = a(t)y + b(t)$ is*

$$y_p(t) = \mathrm{e}^{A(t)} \int_{t_0}^{t} b(s)\mathrm{e}^{-A(s)}\,\mathrm{d}s, \quad \text{where} \quad A(t) = \int_{t_0}^{t} a(s)\,\mathrm{d}s\,.$$

2. *The general solution of $y' = a(t)y + b(t)$ is*

$$y(t) = c\,\mathrm{e}^{A(t)} + y_p(t) = y(t_0)\mathrm{e}^{A(t)} + y_p(t), \quad c \in \mathbb{R}.$$

The remark about the domain of solutions made in the homogeneous case applies, except that now the maximal domain is the intersection of the domains of $a(t)$ and $b(t)$.

Proof.

(1) should be clear from the preceding consideration. Continuity of $b(t)$ is needed for $\frac{\mathrm{d}}{\mathrm{d}t}\int_{t_0}^{t} b(s)\mathrm{e}^{-A(s)}\,\mathrm{d}s = b(t)\mathrm{e}^{-A(t)}$; cf. the proof of the Fundamental Theorem of Calculus.

(2) One needs to show that the difference $t \mapsto y_1(t) - y_2(t)$ of two solutions of $y' = a(t)y + b(t)$ is a solution of $y' = a(t)y$, which is straightforward. $\quad\Longrightarrow y(t) = \underbrace{y(t) - y_p(t)}_{\text{solves } y' = a(t)y} + y_p(t).$ □

Further Notes

- W.l.o.g. we could have assumed that $t = 0$. This assumption is justified, since the "time shift" $z(t) = y(t - t_0)$ transforms the IVP $y' = a(t)y + b(t) \wedge y(0) = y_0$ into $z'(t) = a(t - t_0)z(t) + b(t - t_0) \wedge z(t_0) = y_0$, which is also 1st-order linear with slightly changed coefficient functions.
- The preceding considerations apply, more generally, to functions $a(t)$, $b(t)$ with finitely many discontinuities of the first kind (i.e., the one-sided limits exist but are different).

Further Notes (cont'd)

- (cont'd)
  In this case the preceding formula gives all continuous functions $y\colon I \to \mathbb{R}$ that satisfy $y'(t) = a(t)y(t) + b(t)$ at every point t at which $a(t)$ and $b(t)$ are continuous.

  This follows from a more general version of the Fundamental Theorem of Calculus, which states that $F(t) = \int_a^t f(s)\,\mathrm{d}s$ satisfies $F'(t) = f(t)$ at every t at which f is continuous and has one-sided derivatives equal to $\lim_{s\uparrow t} f(s)$, $\lim_{s\downarrow t} f(s)$ at discontinuities of f of the first kind.

- The following alternative representation of $y_p(t)$ is sometimes useful: Since $A(t) - A(s) = \int_s^t a(\tau)\,\mathrm{d}\tau$, we have

$$y_p(t) = \int_{t_0}^t b(s)\mathrm{e}^{A(t)-A(s)}\,\mathrm{d}s = \int_{t_0}^t G(s,t)b(s)\,\mathrm{d}s$$

  with $G(s,t) = \exp\left(\int_s^t a(\tau)\,\mathrm{d}\tau\right)$.

Examples

1. $y' = 2y + 3$

   In this case $a(t) = 2$, $b(t) = 3$ are constant, and the general solution is

   $$y(t) = -\tfrac{3}{2} + c\,\mathrm{e}^{2t}, \quad c \in \mathbb{R},$$

   because the associated homogeneous ODE $y' = 2y$ is solved by $y_h(t) = c\,\mathrm{e}^{2t}$ and $y' = 2y + 3$ has the constant solution $y_p(t) \equiv -\frac{3}{2}$.

   Solving $y(t_0) = -\frac{3}{2} + c\,\mathrm{e}^{2t_0}$ for c gives the solution of any corresponding IVP:

   $$c = \big(y(t_0) + \tfrac{3}{2}\big)\mathrm{e}^{-2t_0} \implies \boxed{y(t) = \big(y(t_0) + \tfrac{3}{2}\big)\mathrm{e}^{2(t-t_0)} - \tfrac{3}{2}}.$$

   We can also solve it by variation of parameters:

   $$y_p(t) = \mathrm{e}^{2t} \int_{t_0}^{t} \mathrm{e}^{-2s} \cdot 3 \,\mathrm{d}s = \mathrm{e}^{2t} \left[-\frac{3}{2}\mathrm{e}^{-2s} \right]_{t_0}^{t} = -\tfrac{3}{2} + \frac{3\,\mathrm{e}^{-2t_0}}{2}\,\mathrm{e}^{2t},$$

   which is the constant $-\frac{3}{2}$ plus a solution of $y' = 2y$.

Examples Cont'd

1. (cont'd)
   Note that any solution $y(t)$ with $y(t_0) \neq -\frac{3}{2}$ grows exponentially for $t \to +\infty$.
2. $y' = -2y + 3$
   Here the general solution is

$$\boxed{y(t) = (y(t_0) - \tfrac{3}{2})e^{-2(t-t_0)} + \tfrac{3}{2}},$$

   and every solution (regardless of the initial value $y(t_0)$) converges for $t \to +\infty$ towards the constant (equilibrium, steady-state) solution $y(t) \equiv \frac{3}{2}$.

Examples Cont'd

3. $y' = -2y + t$
The associated homogeneous ODE remains the same, and a particular solution is

$$y_p(t) = \mathrm{e}^{-2t} \int t\,\mathrm{e}^{2t}\,\mathrm{d}t$$
$$= \mathrm{e}^{-2t}\left(\tfrac{1}{2}t\,\mathrm{e}^{2t} - \tfrac{1}{2}\int \mathrm{e}^{2t}\,\mathrm{d}t\right) = \tfrac{1}{2}t - \tfrac{1}{4}.$$

$\Longrightarrow$ The general solution is

$$y(t) = \tfrac{1}{2}t - \tfrac{1}{4} + c\,\mathrm{e}^{-2t} \quad (c \in \mathbb{R})$$
$$= \tfrac{1}{2}t - \tfrac{1}{4} + \left(y(t_0) - \tfrac{1}{2}t_0 + \tfrac{1}{4}\right)\mathrm{e}^{-2(t-t_0)}.$$

For $t \to +\infty$ every solution quickly approaches the particular solution $y_p(t) = \frac{1}{2}t - \frac{1}{4}$.

Examples Cont'd

4. $y' = -ty + 1$
   The associated homogeneous ODE $y' = -ty$ has the solution $y(t) = c\,\mathrm{e}^{-t^2/2}$, $c \in \mathbb{R}$.
   A particular solution of $y' = -ty + 1$ is

$$y_p(t) = \mathrm{e}^{-t^2/2} \int_{t_0}^{t} \mathrm{e}^{s^2/2}\,\mathrm{d}s,$$

   and the general solution of $y' = -ty + 1$ is

$$y(t) = \mathrm{e}^{-t^2/2}\left(c + \int_{t_0}^{t} \mathrm{e}^{s^2/2}\,\mathrm{d}s\right), \quad c = y(t_0)\mathrm{e}^{t_0^2/2}.$$

Examples Cont'd

5. $y' = -ty + t$
The associated homogeneous ODE remains unchanged, so that we only need to find one particular solution.
Using variation of parameters we get

$$y_p(t) = \mathrm{e}^{-t^2/2} \int_{t_0}^{t} s\,\mathrm{e}^{s^2/2}\,\mathrm{d}s = \mathrm{e}^{-t^2/2}\left[\mathrm{e}^{s^2/2}\right]_{t_0}^{t}$$
$$= \mathrm{e}^{-t^2/2}\left(\mathrm{e}^{t^2/2} - \mathrm{e}^{t_0^2/2}\right) = 1 - e^{t_0^2/2}\mathrm{e}^{-t^2/2},$$

so that the general solution is

$$y(t) = 1 - e^{t_0^2/2}\mathrm{e}^{-t^2/2} + c\,\mathrm{e}^{-t^2/2} = 1 + c'\mathrm{e}^{-t^2/2} \quad (c, c' \in \mathbb{R}).$$

Surprise?
Not really, because $y' = -ty + t = -t(y-1)$ has the solution $y(t) \equiv 1$.

Examples Cont'd

6. $ty' + 2y = 4t^2$ (cf. [BDM17])
This is an example of an implicit 1st-order linear ODE.
First we rewrite it in explicit form:

$$y' = -\frac{2}{t}\,y + 4t.$$

Note that this splits the original domain $\mathbb{R}$ (for t) into the two subintervals $I_1 = (-\infty, 0)$ and $I_2 = (0, +\infty)$. In what follows we consider only I_2 and choose $t_0 = 1$.
The usual method yields

$$y_h(t) = \exp\left(\int_1^t (-2/s)\,\mathrm{d}s\right) = \mathrm{e}^{-2\ln t} = \frac{1}{t^2},$$

$$y_p(t) = \frac{1}{t^2}\int_1^t s^2 \cdot 4s\,\mathrm{d}s = \frac{1}{t^2}\left[s^4\right]_1^t = t^2 - \frac{1}{t^2}.$$

It follows that another particular solution is $y_p(t) = t^2$, and the general solution is

$$y(t) = t^2 + \frac{c}{t^2} \text{ for } t > 0, \quad \text{with parameter } c \in \mathbb{R}.$$

Examples Cont'd

6. (cont'd) Note that exactly one of these solutions, viz. $y(t) = t^2$, is defined also for $t = 0$.

   The solutions on I_1 are the mirror images w.r.t. to the y-axis of the solutions on I_1, and $y(t) = t^2$ is the only solution of the original ODE $ty' + 2y = 4t^2$ that is defined in a neighborhood of $t = 0$.
   In other words, the IVP $ty' + 2y = 4t^2 \wedge y(0) = y_0$ has a solution precisely for $y_0 = 0$ (and this solution is defined on all of $\mathbb{R}$).

   Plotting solutions:
   Rewriting $y = t^2 + c/t^2$ as $t^2 y - t^4 = c$, we see that the integral curves (solution curves) of $ty' + 2y = 4t^2$ are the contours of $F(t, y) = t^2 y - t^4$.
   $\Longrightarrow$ We can use the contour plot facilities, e.g., of SageMath to plot the solutions.

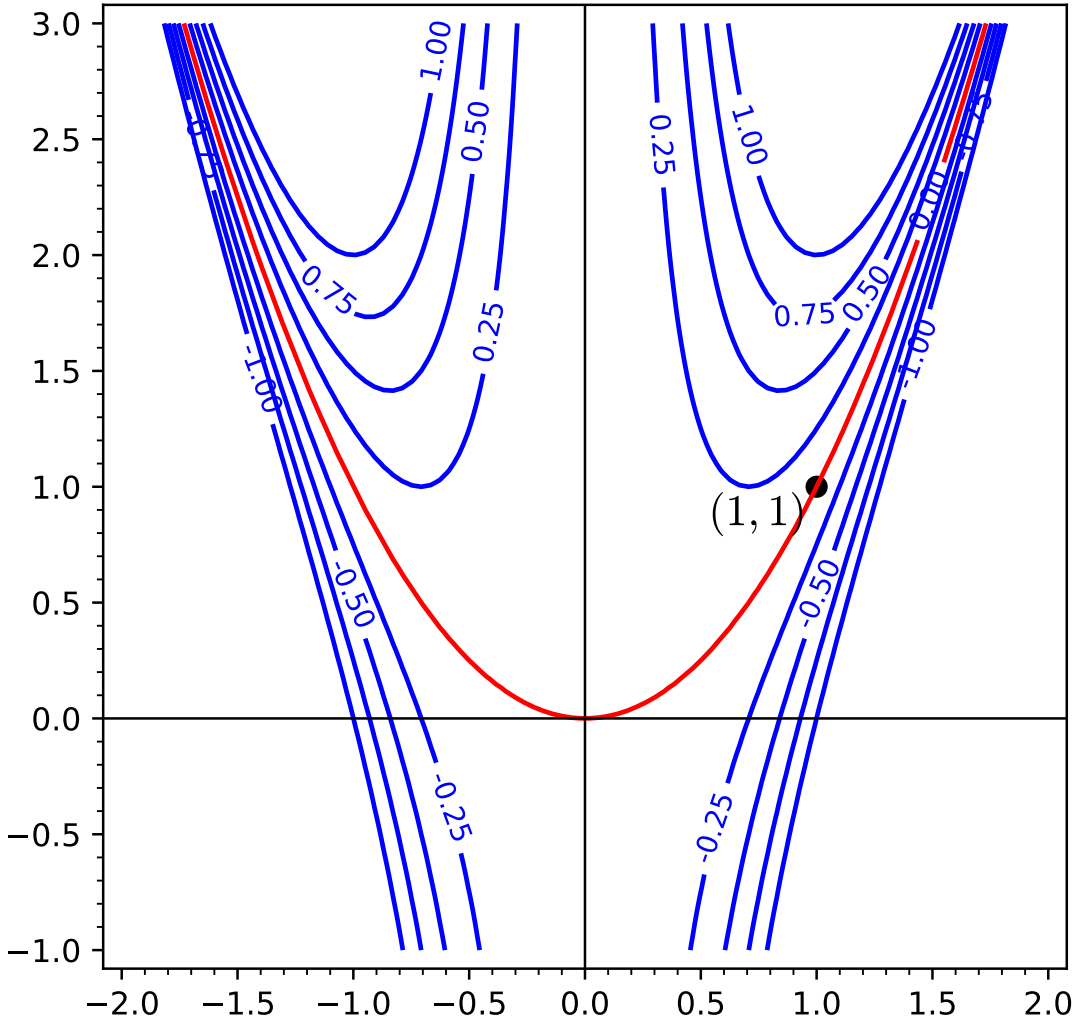


Figure: Graphs of $y_c(t) = t^2 + c/t^2$ for various values of c (including the branches for $t < 0$)

Note on the picture

The empty regions in the plot (due to laziness of your professor) are misleading. Since we can solve $t_0^2 + c/t_0^2 = y_0$ for c provided only that $t_0 \neq 0$, there passes a solution curve through any point of the (t, y)-plane that is not on the y-axis.

Afternote

Our derivation of the general solution of $y' = (-2/t)y + 4t$ illustrates another important point: For solving an inhomogeneous linear 1st-order ODE it suffices to compute 1 nonzero solution $y_h(t)$ of the associated homogeneous ODE and 1 solution $y_p(t)$ of the given inhomogeneous ODE, because the general solution is then $y(t) = y_p(t) + c\,y_h(t)$, $c \in \mathbb{R}$. For the determination of $y_h(t)$, $y_p(t)$ one may integrate from any $t_0 \in I$.

However, it is not necessarily true that varying t_0 over all of I yields all solutions $y_p(t)$. (In the homogeneous case it never does.) In our example this produces $(1/t^2)\int_{t_0}^{t} s^2 \cdot 4s\,\mathrm{d}s = t^2 - t_0^4/t^2$, missing all solutions $t^2 + c/t^2$ with $c \geq 0$.

The correct way to obtain all solutions by integration is to fix t_0 and add an arbitrary constant to the factor $c(t)$ in $y_p(t) = c(t)\mathrm{e}^{A(t)}$, i.e., $y_p(t) = (1/t^2)\left(c_0 + \int_1^t s^2 \cdot 4s\,\mathrm{d}s\right)$, $c_0 \in \mathbb{R}$.

Examples Cont'd

7. $mv' = mg - kv$ (2nd model for a falling object)
This ODE has the form $v' = av + b$ with $a = -k/m$, $b = g$.
The general solution is $v(t) = mg/k + ce^{-kt/m}$, $c \in \mathbb{R}$.
Suppose the object is released at time $t = 0$.
$\Longrightarrow v(0) = 0 \Longrightarrow c = -mg/k$

$$\Longrightarrow v(t) = \frac{mg}{k}\left(1 - e^{-kt/m}\right) \quad \text{for } 0 \le t \le T,$$

where T is the time when the object hits the ground.
The "limiting velocity" is $v_\infty = mg/k$.

Suppose the object is released at height x_0 above ground.
For the distance traveled by the object we obtain by integrating and using $x(0) = 0$

$$x(t) = \frac{mg}{k}\left(t + \frac{m}{k}e^{-kt/m}\right) + C, \quad C = -\frac{m^2 g}{k^2}.$$

$\Longrightarrow T$ can be found by (numerically) solving the equation

$$\frac{mg}{k}\left(T + \frac{m}{k}e^{-kT/m}\right) - \frac{m^2 g}{k^2} = x_0.$$

Integrating Factors

There is an alternative way to solve $y' = a(t)y + b(t)$ using a so-called integrating factor. We can rewrite the ODE as

$$y'(t) - a(t)y(t) = b(t).$$

This equation can be multiplied by any function $m(t)$ with domain I to yield the equivalent form

$$m(t)y'(t) - a(t)m(t)y(t) = m(t)b(t), \qquad (\star)$$

provided that $m(t) \neq 0$ for all $t \in I$.
The goal is to choose $m(t)$ in such a way that the left-hand side can be integrated to yield $y(t)$ ($\rightarrow$ *integrating factor*).
Here $m(t) = \mathrm{e}^{-A(t)}$, $A(t) = \int a(t)\,\mathrm{d}t$, does the job, since $m'(t) = -a(t)m(t)$ and hence the left-hand side of $(\star)$ is $m(t)y'(t) + m'(t)y(t) = \frac{\mathrm{d}}{\mathrm{d}t}\left(m(t)y(t)\right)$:

$$\mathrm{e}^{-A(t)}y'(t) - a(t)\mathrm{e}^{-A(t)}y(t) = \frac{\mathrm{d}}{\mathrm{d}t}\left(\mathrm{e}^{-A(t)}y(t)\right).$$

$$\Longrightarrow \mathrm{e}^{-A(t)}y(t) = \int \mathrm{e}^{-A(t)}b(t)\,\mathrm{d}t \Longrightarrow y(t) = \mathrm{e}^{A(t)}\int \mathrm{e}^{-A(t)}b(t)\,\mathrm{d}t$$

The Linear Algebra Aspect

The set of real-valued functions on a given domain I (i.e., maps $f\colon I \to \mathbb{R}$) is often denoted by $\mathbb{R}^I$. It forms a vector space over $\mathbb{R}$ with respect to the "point-wise" operations

$$(f+g)(t) = f(t) + g(t) \quad \text{for } f, g \in \mathbb{R}^I,$$
$$(cf)(t) = c\,f(t) \quad \text{for } f \in \mathbb{R}^I,\ c \in \mathbb{R}.$$

The general definition of subspaces of an abstract vector space specializes to:

Definition

A set of functions $S \subseteq \mathbb{R}^I$ is called a *subspace* if $S \neq \emptyset$ and S is closed w.r.t. the vector space operations, i.e., $f, g \in S$ implies $f + g \in S$ and $f \in S$ implies $cf \in S$ for all $c \in \mathbb{R}$.

Linear independence, generating set (spanning set), basis, and dimension of subspaces of $\mathbb{R}^I$ are defined in the same way as for $\mathbb{R}^n$ (and are special cases of the corresponding definitions for abstract vector spaces).

From now on we assume that I is an interval of positive length (and thus in particular an infinite set).

Remark

The most important difference between $\mathbb{R}^n$ and $\mathbb{R}^I$ is that $\mathbb{R}^I$ is infinite-dimensional (i.e., does not have a finite basis). For $I = \mathbb{R}$ this can be inferred from the following exercise.

Exercise

Let $f_\lambda(t) = \mathrm{e}^{\lambda t}$ for $\lambda \in \mathbb{R}$. Show that $\{f_\lambda; \lambda \in \mathbb{R}\}$ is linearly independent in $\mathbb{R}^{\mathbb{R}}$.

Hint: Suppose there exists $r \in \mathbb{Z}^+$ and distinct numbers $\lambda_1, \ldots, \lambda_r, c_1, \ldots, c_r \in \mathbb{R}$ such that

$$c_1\mathrm{e}^{\lambda_1 t} + c_2\mathrm{e}^{\lambda_2 t} + \cdots + c_r\mathrm{e}^{\lambda_r t} = 0 \quad \text{for all } t \in \mathbb{R}.$$

Assuming $\lambda_1 < \lambda_2 < \cdots < \lambda_r$ and $c_r \neq 0$, divide this equation by $\mathrm{e}^{\lambda_r t}$ and let $t \to +\infty$ to obtain a contradiction.

Proposition

Assume that $t \mapsto a(t)$ is continuous on I. Then the solution set S of $y' = a(t)y$ forms a 1-dimensional subspace of $\mathbb{R}^I$ and, for any choice of $t_0 \in I$, is generated by the function $I \to \mathbb{R}$,
$t \mapsto \exp\left(\int_{t_0}^{t} a(s)\,\mathrm{d}s\right)$.

Note that we assume that all solutions have maximal domain I.

Proof.

We have $S \neq \emptyset$, since the all-zero function $I \to \mathbb{R}$, $t \mapsto 0$ is a solution of $y' = a(t)y$. Further, it is easy to verify that sums and scalar multiples of solutions of $y' = a(t)y$ are again solutions.
$\Longrightarrow$ S is a subspace of $\mathbb{R}^n$.
The fact that S is 1-dimensional is less trivial; it follows from our theorem on solutions of homogeneous linear 1st-order ODE's, which says that every solution is a scalar multiple of
$t \mapsto \exp\left(\int_{t_0}^{t} a(s)\,\mathrm{d}s\right)$. □

Note

In a way it is surprising that $\dim(S) = 1$, because S is defined by a single linear differential equation. Looking at the case of $\mathbb{R}^n$, where solution spaces of single (nontrivial) linear equations have dimension $n-1$, one would rather expect $\dim(S) = \infty - 1 = \infty$.

Further Notes

- The theorem also gives that the solution set of an inhomogeneous ODE $y' = a(t)y + b(t)$ forms a *line* in the corresponding space $\mathbb{R}^I$, which does not pass through the origin (the all-zero function $I \to \mathbb{R}$). As in our Linear Algebra crash course you may check that any affine combination $t \mapsto \lambda y_1(t) + (1-\lambda)y_2(t)$, $\lambda \in \mathbb{R}$, of two solutions y_1, y_2 of $y' = a(t)y + b(t)$ is again a solution.
- In Example 10 of the introduction we found that the solutions of $y'' + y = 0$ on $\mathbb{R}$ form a 2-dimensional subspace of $\mathbb{R}^{\mathbb{R}}$ with basis $\sin t$, $\cos t$. (We had proved that every solution has the form $A\cos t + B\sin t$, i.e., is in the span of $\{\cos t, \sin t\}$, and it only remains to observe that $\cos t$, $\sin t$ are not constant multiples of each other.)

  The evaluation map $S \to \mathbb{R}^2$, $y \mapsto \big(y(0), y'(0)\big)$, which sends a solution to the corresponding initial values at $t = 0$, is a linear bijection with inverse map $\mathbb{R}^2 \to S$, $(A, B) \mapsto A\cos t + B\sin t$.
- Linear Algebra will play a much more prominent role when we analyze higher-order linear ODE's and 1st-order ODE systems later.

Special Cases of $y' = a(t)y + b(t)$

1. $a(t) = a$ and $b(t) = b$ are constants.

   In this case we have

$$y_p(t) = e^{at}\int_0^t be^{-as}\,ds = be^{at}\left[-\frac{1}{a}e^{-as}\right]_0^t = \frac{b}{a}e^{at}(1-e^{-at})$$
$$= \frac{b}{a}(e^{at}-1)$$

   and hence as solution of the IVP $y' = ay + b \wedge y(t_0) = y_0$ the function

$$\boxed{y(t) = y(t_0)e^{a(t-t_0)} + \frac{b}{a}\left(e^{a(t-t_0)} - 1\right)}$$

   Setting $y(t_0) = -b/a$ gives the particular solution $y_p(t) \equiv -b/a$ noted earlier.

   For $a < 0$ we have $\lim_{t\to+\infty} y(t) = -b/a = y_\infty$, say, and

$$\boxed{y(t) = y_\infty + (y_0 - y_\infty)e^{a(t-t_0)}, \quad y_0 = y(t_0)}.$$

   In other words, every solution $y(t)$ approaches the *steady-state* y_∞ exponentially fast.

Special Cases (cont'd)

2. $a(t) = a$, $b(t) = \mathrm{e}^{ct}$.
   In this case we have

$$y_p(t) = e^{at} \int_{t_0}^{t} \mathrm{e}^{(c-a)s}\,\mathrm{d}s$$

$$= \begin{cases} \mathrm{e}^{at}\left[\frac{1}{c-a}\mathrm{e}^{(c-a)s}\right]_{t_0}^{t} = \frac{\mathrm{e}^{ct}-\mathrm{e}^{at+(c-a)t_0}}{c-a} & \text{if } c \neq a, \\ (t-t_0)\mathrm{e}^{at} & \text{if } c = a. \end{cases}$$

and hence

$$\boxed{y(t) = \begin{cases} y(t_0)\mathrm{e}^{a(t-t_0)} + \mathrm{e}^{ct_0} \cdot \frac{\mathrm{e}^{c(t-t_0)}-\mathrm{e}^{a(t-t_0)}}{c-a} & \text{if } c \neq a, \\ y(t_0)\mathrm{e}^{a(t-t_0)} + \mathrm{e}^{at_0} \cdot (t-t_0)\mathrm{e}^{a(t-t_0)} & \text{if } c = a. \end{cases}}$$

In the second case (a type of *resonance*) the solution may grow initially (i.e., for $t \downarrow t_0$) even if $a < 0$. This happens precisely for $y'(t_0) = ay(t_0) + \mathrm{e}^{at_0} > 0$, i.e., $y(t_0) < -\frac{1}{a}\mathrm{e}^{at_0}$.

Special Cases (cont'd)

3. $a(t) = a$, $b(t) = \begin{cases} 0 & \text{if } t < T, \\ b & \text{if } t \geq T. \end{cases}$

Assuming that $t_0 = T$, we have $y_p(t) = 0$ for $t \leq T$ and

$$y_p(t) = \mathrm{e}^{at} \int_T^t b\mathrm{e}^{-as}\,\mathrm{d}s = b\mathrm{e}^{at}\left[-\frac{1}{a}\mathrm{e}^{-as}\right]_T^t = \frac{b}{a}\left(e^{a(t-T)} - 1\right)$$

for $t \geq T$. This gives the general solution as

$$\boxed{y(t) = \begin{cases} y(T)\mathrm{e}^{a(t-T)} & \text{for } t \leq T, \\ y(T)\mathrm{e}^{a(t-T)} + \frac{b}{a}\left(e^{a(t-T)} - 1\right) & \text{for } t \geq T. \end{cases}}$$

We can verify that

$$\lim_{h\uparrow 0} \frac{y(T+h) - y(T)}{h} = a\,y(T), \quad \lim_{h\downarrow 0} \frac{y(T+h) - y(T)}{h} = a\,y(T) + b,$$

in accordance with the preceding note about discontinuities of $b(t)$. The solutions $y(t)$ simply arise by continuously gluing a solution of $y'(t) = a\,y(t)$ for $t \leq T$ with the corresponding solution of $y'(t) = a\,y(t) + b$ for $t \geq T$.

Special Cases (cont'd)

4. $a(t) = a$, $b(t) = \begin{cases} +\infty & \text{if } t = T, \\ 0 & \text{if } t \neq T. \end{cases}$

In the special case $t = 0$ this function is called *delta function* and denoted by $\delta(t)$, so that in general $b(t) = \delta(t - T)$.

$\delta(t)$ is not an ordinary function but represents a so-called *distribution*, which acts by integration on functions. The precise definition is

$$\int_{\mathbb{R}} f(t)\delta(t)\,\mathrm{d}t = \lim_{h \downarrow 0} \int_{\mathbb{R}} f(t)\delta_h(t),$$

where $\delta_h(t) = \frac{1}{2h} \times$ characteristic function of $[-h, h]$. If f is continuous at $t = 0$, this definition gives $\int_{\mathbb{R}} f(t)\delta(t)\,\mathrm{d}t = f(0)$ and in particular $\int_{\mathbb{R}} \delta(t)\,\mathrm{d}t = 1$.

Substituting $\delta(t - T)$ into the solution formula gives $y_p(t) = 0$ for $t < T$ and, assuming $t_0 < T$,

$$y_p(t) = \mathrm{e}^{at} \int_{t_0}^{t} \delta(T - s)\mathrm{e}^{-as}\,\mathrm{d}s = \mathrm{e}^{at}\mathrm{e}^{-aT} = \mathrm{e}^{a(t-T)} \quad \text{for } t > T.$$

Special Cases (cont'd)

4. (cont'd)
The general solution of $y'(t) = a\,y(t) + \delta(t-T)$ is then

$$\boxed{y(t) = \begin{cases} y(t_0)\mathrm{e}^{a(t-t_0)} & \text{if } t < T, \\ y(t_0)\mathrm{e}^{a(t-t_0)} + \mathrm{e}^{a(t-T)} & \text{if } t \geq T. \end{cases}}$$

We have $\lim_{t\uparrow T} y(t) = y(t_0)\mathrm{e}^{a(t-t_0)}$,
$\lim_{t\downarrow T} y(t) = y(T) = y(t_0)\mathrm{e}^{a(t-t_0)} + 1$, $y'(t) = a\,y(t)$ for $t \neq T$,
and $y(t)$ arises from gluing together solutions of $y' = ay$ on $(-\infty, T)$ and $(T, +\infty)$ which differ by a unit step at $t = T$.

A Brief Introduction to Complex Numbers

For a more gentle introduction see [Ste16], Appendix H

A complex number is a point in the Euclidean plane $\mathbb{R}^2$. Complex numbers are added and multiplied according to the rules

$$(a,b)+(c,d) := (a+c, b+d), \qquad \text{(Vector addition)}$$
$$(a,b)(c,d) := (ac-bd, ad+bc). \qquad \text{(Well, just fancy)}$$

In particular we have

$$(a,0)+(c,0) = (a+c,0),\ (a,0)(c,0) = (ac,0) \quad \text{for } a,c \in \mathbb{R}$$

(the numbers on the real axis are multiplied as usual), and

$$(0,1)^2 = (0,1)(0,1) = (-1,0)$$

(the square of the "imaginary unit" $\mathrm{i} = (0,1)$ is the point on the real axis corresponding to -1).

Making the identification $(a,0) \triangleq a$, we obtain

$$(a,b) = (a,0)+(0,b) = (a,0)+(b,0)(0,1) = a+b\,\mathrm{i}, \quad \mathrm{i}^2 = -1.$$

The complex numbers form a field, i.e., their addition/multiplication follows the usual laws of arithmetic. Thus it suffices to memorize only $\mathrm{i}^2 = -1$: Any complex number z has the form $z = a + b\,\mathrm{i}$ for some unique real numbers a, b, and

$$\begin{aligned} z + w &= (a + b\,\mathrm{i}) + (c + d\,\mathrm{i}) = a + c + (b + d)\mathrm{i}, \\ zw &= (a + b\,\mathrm{i})(c + d\,\mathrm{i}) = ac + ad\,\mathrm{i} + bc\,\mathrm{i} + bd\,\mathrm{i}^2 \\ &= ac - bd + (ad + bc)\mathrm{i}. \qquad \text{(Using } \mathrm{i}^2 = -1\text{)} \end{aligned}$$

Complex variables are commonly denoted by $z, w, \ldots$ (cp. with $x, y, \ldots$ for real variables), and the field of complex numbers is denoted by $\mathbb{C}$. (But keep in mind that $a + b\,\mathrm{i}$ is just the point (a, b), i.e., $\mathbb{C}$ is just $\mathbb{R}^2$ equipped with a fancy multiplication.)

The key property that distinguishes fields from commutative rings such as $\mathbb{Z}$ is that every element $z \neq 0$ has a "multiplicative inverse w" satisfying $zw = 1$. One writes z^{-1} or $1/z$ for w and z_1/z_2 for $z_1 z_2^{-1}$.

For a complex number $z = a + b\,\mathrm{i} \neq 0$ (i.e., at least one of a, b is nonzero) the multiplicative inverse is easily obtained by rationalizing the denominator:

$$\frac{1}{z} = \frac{1}{a + b\,\mathrm{i}} = \frac{a - b\,\mathrm{i}}{(a + b\,\mathrm{i})(a - b\,\mathrm{i})} = \frac{a - b\,\mathrm{i}}{a^2 + b^2} = \frac{a}{a^2 + b^2} + \frac{-b}{a^2 + b^2}\,\mathrm{i}.$$

The analogy with $\mathbb{R}^2$ goes further: The *absolute value* $|z|$ of a complex number z is its Euclidean length, i.e.,

$$|z| = |a + b\,\mathrm{i}| := \sqrt{a^2 + b^2}.$$

It satisfies $|z + w| \le |z| + |w|$ (triangle inequality for the Euclidean length/distance) and $|zw| = |z|\,|w|$. For the latter check that

$$(a^2 + b^2)(c^2 + d^2) = (ac - bd)^2 + (ad + bc)^2.$$

The *complex conjugate* of $z = a + b\,\mathrm{i} \in \mathbb{C}$ is $\overline{z} = a - b\,\mathrm{i}$. Geometrically, the map $z = (a, b) \mapsto \overline{z} = (a, -b)$ is reflection at the x-axis ("real axis"). Algebraically, it satisfies $\overline{z + w} = \overline{z} + \overline{w}$, $\overline{zw} = \overline{z}\,\overline{w}$, i.e., forms an automorphism of $\mathbb{C}$.

The coordinates a, b of $z = (a, b) = a + b\,\mathrm{i} \in \mathbb{C}$ are called *real part* resp. *imaginary part* of z, notation $a = \mathrm{Re}(z)$, $b = \mathrm{Im}(z)$.

Since $\mathbb{C} = \mathbb{R}^2$ as a set, we can do analysis in $\mathbb{C}$ as you have learned in Calculus III. For example, a sequence (z_n) of complex numbers converges to $z \in \mathbb{C}$ if for every $\epsilon > 0$ there exists an $N \in \mathbb{N}$ such that $|z_n - z| < \epsilon$ for all $n > N$. Writing $z_n = a_n + b_n\,\mathrm{i}$, $z = a + b\,\mathrm{i}$ ($a_n, b_n, a, b \in \mathbb{R}$), the convergence $z_n \to z$ is equivalent to $a_n \to a \wedge b_n \to b$ (coordinate-wise convergence).

A series $\sum_{n=0}^{\infty} z_n$ of complex numbers converges (i.e., the associated sequence $s_n = z_1 + z_2 + \cdots + z_n$ of partial sums converges), provided it *converges absolutely*, i.e., the $\sum_{n=0}^{\infty} |z_n|$ (an ordinary series of real numbers) converges. This follows by applying the absolute convergence test for real series [Ste21, Ch. 11.5, Th. 3] to $\sum_{n=0}^{\infty} \operatorname{Re}(z_n)$, $\sum_{n=0}^{\infty} \operatorname{Im}(z_n)$. The details are left as an exercise. (One should note that there is nothing special about complex numbers here. The analogous statement holds for series of points in $\mathbb{R}^n$.)

Polar Form for complex numbers

Using polar coordinates in $\mathbb{R}^2$ we can write every nonzero complex number z in the form

$$z = (r\cos\phi, r\sin\phi) = r\cos\phi + r\sin\phi\,\mathrm{i} = r(\cos\phi + \mathrm{i}\sin\phi).$$

Here $r = |z|$ and $\phi \in [0, 2\pi)$ are uniquely determined by z.The complex exponential function $\exp\colon \mathbb{C} \to \mathbb{C}$ is defined by the same power series as in the real case (and extends $x \mapsto \mathrm{e}^x$ to $\mathbb{C}$):

$$\mathrm{e}^z := \exp(z) := \sum_{n=0}^{\infty} \frac{z^n}{n!}.$$

Polar Form for complex numbers cont'd

That the exponential series converges for all $z \in \mathbb{C}$, can be proved using the absolute convergence test mentioned above.

The functional equation $e^{z+w} = e^z e^w$ holds for all $z, w \in \mathbb{C}$. This can be proved by rearranging the double series representing $e^z e^w$ according to $z^i w^j$ with $i + j$ fixed and using the Binomial Theorem; cf. exercise.

Finally, extracting real and imaginary part of $e^{i\phi} = \sum_{n=0}^{\infty} (i\phi)^n/n!$ and using the known Taylor series of cos, sin, one arrives at *Euler's Identity*

$$e^{i\phi} = \cos\phi + i\sin\phi, \quad \phi \in \mathbb{R}.$$

Combining this with polar coordinates in $\mathbb{R}^2$, we see that every $z \in \mathbb{C} \setminus \{0\}$ admits a unique representation

$$z = r(\cos\phi + i\sin\phi) = re^{i\phi} \quad \text{with } r = |z| > 0,\ \phi \in [0, 2\pi).$$

This is the so-called *polar form* of z. The angle ϕ is called the *argument* of z, notation $\phi = Arg(z)$. Analytically, for $z = x + yi$ we have $\phi = \arctan(y/x)$ if $x > 0$, $\phi = \arctan(y/x) + \pi$ if $x < 0$, and $\phi = \pm\pi/2$ if $x = 0 \wedge y \gtrless 0$.

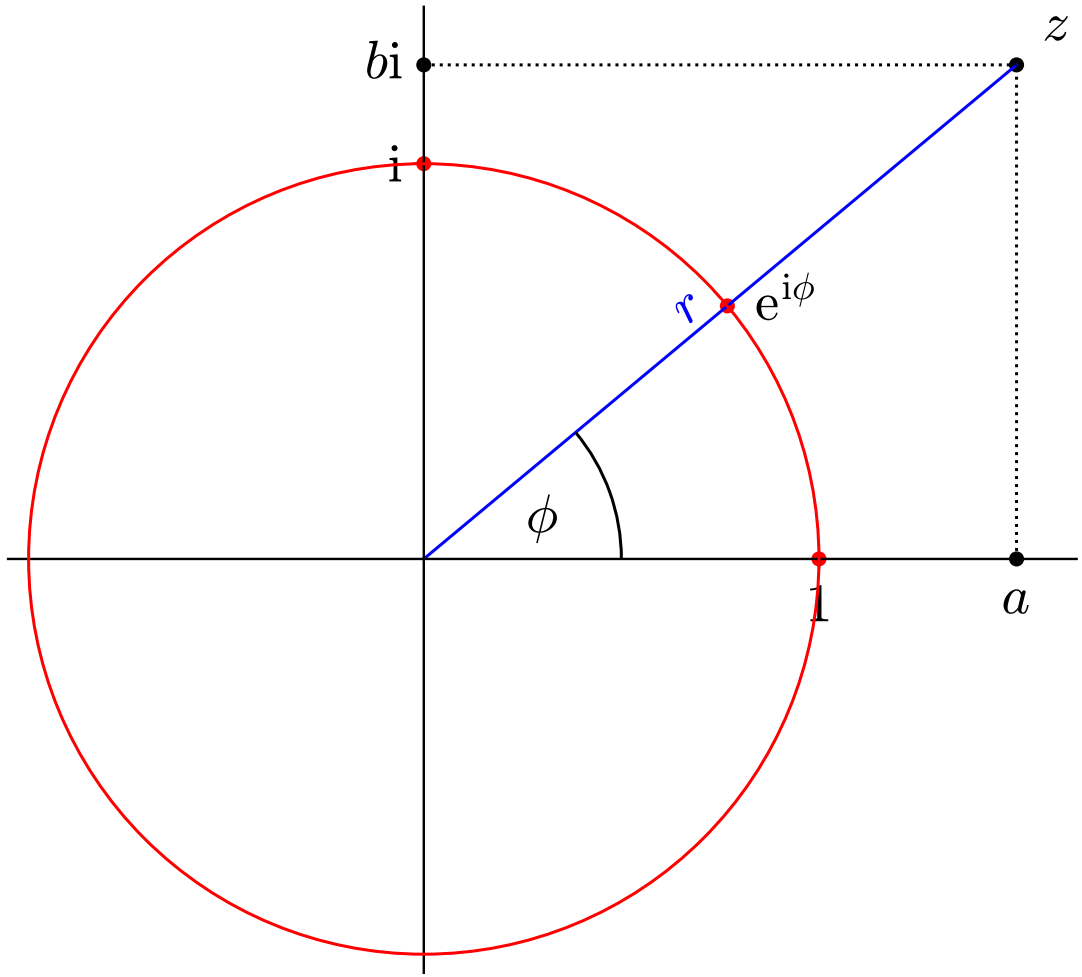


Figure: The polar form $z = r\,\mathrm{e}^{\mathrm{i}\phi}$ of $z = a + b\mathrm{i}$

r, ϕ are computed from a, b as $r = |z| = \sqrt{a^2 + b^2}$, $\phi = \arctan(b/a)$.

The polar form easily shows the geometric meaning of complex multiplication. For $z = re^{i\phi}$, $w = se^{i\psi}$ in polar form, we have

$$zw = rs\,e^{i\phi}e^{i\psi} = rs\,e^{i(\phi+\psi)}$$

(using the functional equation for $z \mapsto e^z$). This is the polar form of zw, except that $\phi + \psi$ is not necessarily reduced modulo 2π. From it we see that multiplication by z is composed of a rotation with angle $\phi = Arg(z)$ (the map $w \mapsto e^{i\phi}w$ and a scaling map (the map $w \mapsto |z|\,w$.

For example, multiplication by the imaginary unit $i = e^{i\pi/2}$ just rotates every $w \in \mathbb{C} \setminus \{0\}$ around the origin by 90°, and multiplication by $1 + i = \sqrt{2}\,e^{i\pi/4}$ rotates $w \in \mathbb{C} \setminus \{0\}$ by 45° and scales it by the factor $\sqrt{2}$.

Roots of Unity

A complex number z is said to be an *n-th root of unity* if $z^n = 1$. Writing this equation in polar form, $z^n = r^n e^{in\phi} = 1\,e^{0\,\phi}$, shows that the n-th roots of unity are precisely the n numbers $e^{2\pi ik/n}$ with $k \in \{0, 1, \ldots, n-1\}$. These form the vertices of the regular n-gon centered at 0 and with one vertex at 1, which is inscribed into the unit circle. Writing $\zeta_n = e^{2\pi i/n}$, the n-th roots of unity are $1, \zeta_n, \zeta_n^2, \ldots, \zeta_n^{n-1}$.

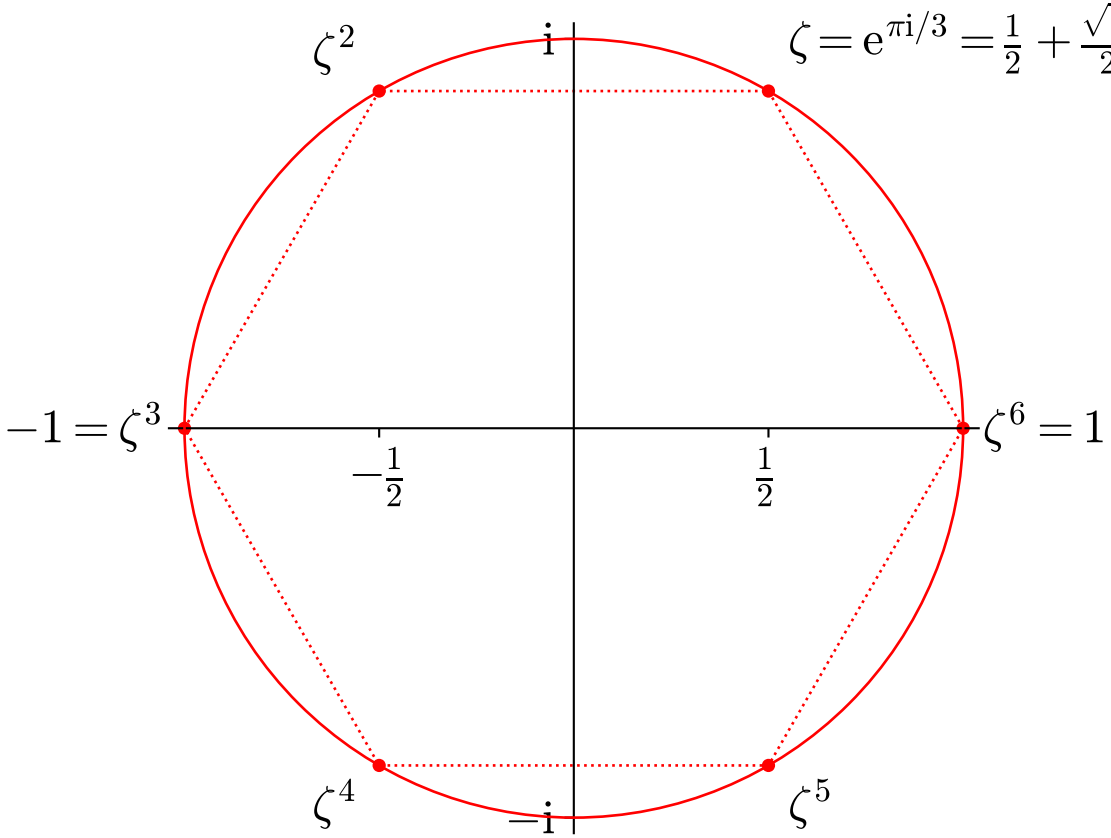


Figure: The 6th roots of unity in $\mathbb{C}$ form a regular hexagon inscribed into the unit circle $|z| = 1$

One can also verify directly that $\zeta = \frac{1+i\sqrt{3}}{2}$ satisfies $\zeta^2 = \frac{-1+i\sqrt{3}}{2}$, $\zeta^3 = -1$, $\zeta^4 = \frac{-1-i\sqrt{3}}{2}$, $\zeta^5 = \frac{1-i\sqrt{3}}{2}$, and $\zeta^6 = 1$.

The Fundamental Theorem of Algebra

That the polynomial $X^n - 1$ has n distinct roots in $\mathbb{C}$ and hence splits in $\mathbb{C}[X]$ (the polynomial ring in one indeterminate over $\mathbb{C}$) into linear factors, viz.

$$X^n - 1 = \prod_{k=0}^{n-1} \left(X - \mathrm{e}^{2\pi \mathrm{i} k/n} \right),$$

is a special case of the so-called *Fundamental Theorem of Algebra*:

> *Every polynomial* $p(X) = p_0 + p_1 X + \cdots + p_d X^d$ *with coefficients* $p_i \in \mathbb{C}$ *and degree* $d \geq 1$ *(i.e.,* $p_d \neq 0$*) has at least one root in* $\mathbb{C}$*.*

Since $p(c) = 0$ implies $p(X) = (X - c)q(X)$ for some polynomial $q(X)$of degree $d - 1$, it follows by induction that $p(X)$ splits into linear factors in $\mathbb{C}[X]$.

No easy proof of the Fundamental Theorem of Algebra is known. A rather elementary, but still quite intricate proof due to ARGAND (1814) is within the scope of a Calculus III course. One assumes, by contradiction, that $p(X)$ has no root in $\mathbb{C}$. Then $f\colon \mathbb{C} \to \mathbb{C}$, $z \mapsto \frac{1}{|p(z)|}$ is well defined, and one can easily show that f attains a maximum at some point $z_0 \in \mathbb{C}$. Algebraic properties of $\mathbb{C}$ are then used to derive a contradiction from this.

Exercises on Complex Numbers

1. Show $\overline{zw} = \overline{z}\,\overline{w}$ and $|zw| = |z|\,|w|$ for $z, w \in \mathbb{C}$.
2. Show $z\overline{z} = |z|^2$ for $z \in \mathbb{C}$, and give a geometric interpretation of the inversion map $\mathbb{C} \setminus \{0\} \to \mathbb{C} \setminus \{0\}$, $z \mapsto \frac{1}{z} = \frac{\overline{z}}{z\overline{z}}$.
3. Show that a series $\sum_{n=0}^{\infty} z_n$ of complex numbers converges if it converges absolutely, i.e., $\sum_{n=0}^{\infty} |z_n|$ converges in $\mathbb{R}$.
4. The complex exponential function is defined by
$$\mathrm{e}^z := \exp(z) := \sum_{n=0}^{\infty} \frac{z^n}{n!} \quad \text{for } z = x + \mathrm{i}\,y \in \mathbb{C}.$$
Show that this series converges for all $z \in \mathbb{C}$.
5. Evaluate $\sum_{n=0}^{\infty} \left(\frac{1+\mathrm{i}}{2}\right)^n$ and graph the first few partial sums of this series in the complex plane (i.e., in $\mathbb{R}^2$).

Exercises on Complex Numbers Cont'd

6. Prove Euler's identity $\mathrm{e}^{\mathrm{i}\phi} + \cos\phi + \mathrm{i}\sin\phi$.
   Hint: $\mathrm{i}^2 = 1$, $\mathrm{i}^3 = -\mathrm{i}$, $\mathrm{i}^4 = 1$, $\mathrm{i}^5 = \mathrm{i}$, etc.
7. Prove the functional equation for the complex exponential function: $\mathrm{e}^{z+w} = \mathrm{e}^z\mathrm{e}^w$ for $z, w \in \mathbb{C}$.
   Hint: For two absolutely convergent series $\sum_{k=0}^{\infty} c_k$, $\sum_{l=0}^{\infty} d_l$ the identity
   $\left(\sum_{k=0}^{\infty} c_k\right)\left(\sum_{l=0}^{\infty} d_l\right) = \sum_{n=0}^{\infty}(c_0 d_n + c_1 d_{n-1} + \cdots + c_n d_0)$ holds.
8. For $z = x + \mathrm{i}y$ show that $\mathrm{Re}(\mathrm{e}^z) = \mathrm{e}^x\cos y$, $\mathrm{Im}(\mathrm{e}^z) = \mathrm{e}^x\sin y$.
9. Show that the range of the complex exponential function is $\mathbb{C}\setminus\{0\}$ and that $\mathrm{e}^{z+2\pi\mathrm{i}} = \mathrm{e}^z$ for $z \in \mathbb{C}$.

Exercises on Complex Numbers Cont'd

10. Suppose $c = a + \mathrm{i}b \in \mathbb{C}$ is nonzero. Show without recourse to Euler's Identity (cf. previous exercise) that the equation $z^2 = c$ has exactly two solutions in $\mathbb{C}$.

    Hint: For $z = x + \mathrm{i}y$ the equation $z^2 = c$ is equivalent to $x^2 - y^2 = a \wedge 2xy = b$. Express $x^2 + y^2$ in terms of a, b.

11. Show (e.g., by completing the square) that a quadratic equation $Az^2 + Bz + C = 0$, $A, B, C \in \mathbb{C}$, $A \neq 0$, has (exactly) 2 solutions in $\mathbb{C}$ if $B^2 - 4AC \neq 0$ and 1 solution if $B^2 - 4AC = 0$.

12. Euler's Identity and the functional equation for $z \mapsto \mathrm{e}^z$ (cf. previous exercise) imply that the solutions of $z^n = 1$ in $\mathbb{C}$ (*n-th roots of unity*) have the form $\mathrm{e}^{2\pi\mathrm{i}k/n} = \zeta_n^k$ with $k \in \{0, 1, \ldots, n-1\}$, $\zeta_n = \mathrm{e}^{2\pi\mathrm{i}/n}$, and form the vertices of a regular n-gon inscribed in the unit circle. Using the result of a), determine ζ_{24} in the form $u + \mathrm{i}v$ and sketch the solutions of $z^{24} = 1$ that are contained in the 1st quadrant of the plane.

Complex 1st-Order Linear ODE's

Definition

An (explicit) first-order linear ODE with (non-constant) complex coefficients has the form

$$z'(t) = a(t)z(t) + b(t) \quad \text{with } a, b\colon D \to \mathbb{C}.$$

A solution of such a complex ODE is a complex-valued function $z(t) = x(t) + \mathrm{i}y(t)$, defined on an interval $I \subseteq D$ and satisfying $z'(t) = x'(t) + \mathrm{i}y'(t) = a(t)z(t) + b(t)$ for all $t \in I$.

Writing $a(t) = a_1(t) + \mathrm{i}a_2(t)$, $b(t) = b_1(t) + \mathrm{i}b_2(t)$, the complex ODE is equivalent to

$$\begin{aligned} x'(t) &= a_1(t)x(t) - a_2(t)y(t) + b_1(t), \\ y'(t) &= a_2(t)x(t) + a_1(t)y(t) + b_2(t); \end{aligned}$$

in matrix form:

$$\begin{pmatrix} x'(t) \\ y'(t) \end{pmatrix} = \begin{pmatrix} a_1(t) & -a_2(t) \\ a_2(t) & a_1(t) \end{pmatrix} \begin{pmatrix} x(t) \\ y(t) \end{pmatrix} + \begin{pmatrix} b_1(t) \\ b_2(t) \end{pmatrix}.$$

General solution

The general solution of $z'(t) = a(t)z(t) + b(t)$ is $z(t) = z_0 e^{A(t)} + z_p(t)$ with $z_0 \in \mathbb{C}$ and $A, z_p \colon I \to \mathbb{C}$ defined by

$$A(t) = \int_{t_0}^{t} a(s)\,\mathrm{d}s, \quad z_p(t) = e^{A(t)} \int_{t_0}^{t} b(s) e^{-A(s)}\,\mathrm{d}s\,.$$

The proof given in the real case carries over—essentially because differentiation/integration of complex-valued functions of a real variable is done component-wise and the formula $\frac{\mathrm{d}}{\mathrm{d}t} e^{A(t)} = A'(t) e^{A(t)}$ also holds for complex-valued functions $A(t)$.

The chosen normalization of $A(t)$, $z_p(t)$ implies $A(t_0) = z_p(t_0) = 0$, showing that $z(t) = z_0 e^{A(t)} + z_p(t)$ is the unique solution of the corresponding IVP $z'(t) = a(t)z(t) + b(t) \wedge z(t_0) = z_0$.

Complexification of real ODE's

In order to solve a real ODE $y'(t) = a(t)y(t) + b(t)$, write $b(t) = \operatorname{Im} B(t)$ and solve the complex ODE $z'(t) = a(t)z(t) + B(t)$.

$$\begin{aligned} z'(t) = x'(t) + \mathrm{i}y'(t) &= a(t)\big(x(t) + \mathrm{i}y(t)\big) + \operatorname{Re} B(t) + \mathrm{i} \operatorname{Im} B(t) \\ &= a(t)x(t) + \operatorname{Re} B(t) + \mathrm{i}\big(a(t)y(t) + b(t)\big), \end{aligned}$$

$\Longrightarrow y(t) = \operatorname{Im} z(t)$ will then be a solution of $y'(t) = a(t)y(t) + b(t)$.

Even though it adds additional complexity, complexification can be useful since complex functions are sometimes easier to evaluate/differentiate/integrate than real functions. As an example, recall the computation of $\int_0^{2\pi} \cos(mt)\cos(nt)\,\mathrm{d}t$ by using $\mathrm{e}^{\mathrm{i}x}$ in place of $\cos x$.

Example

We solve $y' = ay + \sin(\omega t)$ with $a, \omega \in \mathbb{R}$.

Complexifying this ODE leads to $z' = az + \mathrm{e}^{\mathrm{i}\omega t}$, which is a complex analogue of $y' = ay + \mathrm{e}^{ct}$ (with $c = \mathrm{i}\omega$).
Now we could recall the corresponding formula derived by variation of parameters, but it is also instructive to solve the complex ODE ad hoc.

Since $(\mathrm{e}^{\mathrm{i}\omega t})' = \mathrm{i}\omega\,\mathrm{e}^{\mathrm{i}\omega t}$, it is reasonable to guess that there exists a particular solution of the form $z(t) = A\,\mathrm{e}^{\mathrm{i}\omega t}$ with $A \in \mathbb{C}$.

$$z'(t) = A\mathrm{i}\omega\,\mathrm{e}^{\mathrm{i}\omega t} = a(A\,\mathrm{e}^{\mathrm{i}\omega t}) + \mathrm{e}^{\mathrm{i}\omega t} \iff A\mathrm{i}\omega = aA + 1 \iff A = \frac{1}{\mathrm{i}\omega - a}.$$

$$\Longrightarrow z(t) = \frac{1}{-a + \mathrm{i}\omega}\,\mathrm{e}^{\mathrm{i}\omega t} = \frac{-a - \mathrm{i}\omega}{a^2 + \omega^2}\big(\cos(\omega t) + \mathrm{i}\,\sin(\omega t)\big)$$

Example (cont'd)

$$\Longrightarrow y(t) = \operatorname{Im} z(t) = -\frac{\omega}{a^2+\omega^2}\cos(\omega t) - \frac{a}{a^2+\omega^2}\sin(\omega t)$$

This function is indeed a solution of $y' = ay + \sin(\omega t)$, as the following double-check shows:

$$\begin{aligned} y'(t) &= \frac{\omega^2}{a^2+\omega^2}\sin(\omega t) - \frac{a\omega}{a^2+\omega^2}\cos(\omega t) \\ &= \sin(\omega t) - \frac{a^2}{a^2+\omega^2}\sin(\omega t) - \frac{a\omega}{a^2+\omega^2}\cos(\omega t) \\ &= \sin(\omega t) + ay(t) \end{aligned}$$

Notes

- Of course we can also complexify using $y(t) = \operatorname{Re} z(t)$.
- Using the polar form $A = R\,\mathrm{e}^{\mathrm{i}\phi}$, the solution of the preceding example can also be expressed as

$$y(t) = \operatorname{Im}\left(R\,\mathrm{e}^{\mathrm{i}\phi}\mathrm{e}^{\mathrm{i}\omega t}\right) = \operatorname{Im}\left(R\,\mathrm{e}^{\mathrm{i}(\omega t+\phi)}\right) = R\sin(\omega t+\phi).$$

Example (cont'd)

Since

$$\left|\frac{-a-\mathrm{i}\omega}{a^2+\omega^2}\right| = \left|\frac{1}{-a+\mathrm{i}\omega}\right| = \frac{1}{|-a+\mathrm{i}\omega|} = \frac{1}{\sqrt{a^2+\omega^2}},$$

our previously found particular solution of $y' = ay + \sin(\omega t)$ admits the two alternative representations

$$\begin{aligned} y(t) &= -\frac{\omega}{a^2+\omega^2}\cos(\omega t) - \frac{a}{a^2+\omega^2}\sin(\omega t) \\ &= \frac{1}{\sqrt{a^2+\omega^2}}\sin(\omega t+\phi) \end{aligned}$$

with

$$\phi = \begin{cases} \arctan(\omega/a) & \text{if } a<0, \\ \arctan(\omega/a)+\pi & \text{if } a>0. \end{cases}$$

In fact any linear combination $A\cos(\omega t) + B\sin(\omega t)$ ($A, B \in \mathbb{R}$) can be brought into such a form (with cos or sin), since

$$A\cos(\omega t) + B\sin(\omega t) = \mathrm{Re}\big((A-\mathrm{i}B)\mathrm{e}^{\mathrm{i}\omega t}\big) = \mathrm{Im}\big((B+\mathrm{i}A)\mathrm{e}^{\mathrm{i}\omega t}\big).$$

Pure Mathematicians would ...

cf. the previous set of exercises

- start with the series representation $e^z = \exp z = \sum_{n=0}^{\infty} \frac{z^n}{n!}$, $z \in \mathbb{C}$, of the exponential function.
- Derive the functional equation $\exp(z+w) = (\exp z)(\exp w)$ $(z, w \in \mathbb{C})$ from this using the binomial theorem in the form $\frac{(z+w)^n}{n!} = \sum_{k=0}^{n} \frac{z^k}{k!} \frac{w^{n-k}}{(n-k)!}$.
- Define $\cos(x) = \mathrm{Re}(e^{ix})$ and $\sin(x) = \mathrm{Im}(e^{ix})$, making Euler's Identity a trivial fact.
- Derive the powers series representations

$$\cos x = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k}}{(2k)!}, \quad \sin x = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{(2k+1)!}$$

  by separating $\exp(ix) = \sum_{n=0}^{\infty} \frac{(ix)^n}{n!}$ into real and imaginary part.
- Derive all the well-known properties of cos, sin from their power series representations and the functional equation for the exponential function. As an example, we have $\cos^2 x + \sin^2 x = \left|e^{ix}\right|^2 = e^{ix} \cdot \overline{e^{ix}} = e^{ix} \cdot e^{-ix} = 1$.

Pure Mathematicians would ... (cont'd)

- Define the famous numbers e and π by $\mathrm{e} = \exp(1)$ and
$$\pi = 2 \times \text{smallest positive zero of } x \mapsto \cos x.$$
That this zero is well-defined, follows from continuity of $\cos$ (which requires its own proof, of course) and the intermediate value theorem on account of $\cos(0) = 1 > 0$,
$$\cos(2) = 1 - \frac{2^2}{2!} + \frac{2^4}{4!} - < 1 - 2 + 16/24 = -1/3 < 0,$$
where we have used the alternating series test for convergence and the corresponding limit estimation. As a by-product, we obtain $0 < \pi/2 < 2$ or $0 < \pi < 4$ (a rather weak estimate, which can be easily improved using, e.g., Newton's Iteration).
- Use $\cos(\pi/2) = 0$, $\sin(\pi/2)^2 + \cos(\pi/2)^2 = 1$ and $\sin' x = \cos x > 0$ for $x \in [0, \pi/2)$ to conclude that $\sin(\pi/2) = 1$, $\mathrm{e}^{\mathrm{i}\pi/2} = \cos(\pi/2) + \mathrm{i}\sin(\pi/2) = \mathrm{i}$, and $\mathrm{e}^{z+w} + \mathrm{e}^z\mathrm{e}^w$ to conclude further that $\mathrm{e}^{z+\mathrm{i}\pi/2} = \mathrm{e}^z\mathrm{e}^{\mathrm{i}\pi/2} = \mathrm{i}\,\mathrm{e}^z$, $\mathrm{e}^{z+\mathrm{i}\pi} = -\mathrm{e}^z$ and $\mathrm{e}^{z+2\pi\mathrm{i}} = \mathrm{e}^z$ (hence $\mathrm{e}^{\pi\mathrm{i}} = -1$, $\mathrm{e}^{2\pi\mathrm{i}} = 1$).

Exercise

Suppose $A\colon I \to \mathbb{C}$, $t \mapsto A_1(t) + \mathrm{i}\, A_2(t)$ is differentiable (i.e., $A_1 = \operatorname{Re} A$ and $A_2 = \operatorname{Im} A$ are differentiable). Show that $I \to \mathbb{C}$, $t \mapsto \mathrm{e}^{A(t)}$ is differentiable as well, and

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathrm{e}^{A(t)} = A'(t)\mathrm{e}^{A(t)}.$$

Hint: Start with

$$\mathrm{e}^{A(t)} = \mathrm{e}^{A_1(t)+\mathrm{i}\,A_2(t)} = \mathrm{e}^{A_1(t)}\mathrm{e}^{\mathrm{i}\,A_2(t)} = \mathrm{e}^{A_1(t)}\cos A_2(t) + \mathrm{i}\,\mathrm{e}^{A_1(t)}\sin A_2(t).$$

The Analogy with Linear Recurring Sequences

We consider only the case $y' = ay + b$ with constant coefficients a, b, since for the discussion of linear recurring sequences in Discrete Mathematics the same assumption was made.

The discrete analog of $y' = ay + b$ is the 1st-order linear recurrence relation $x_{n+1} = ax_n + b$ (equivalently, $x_n = ax_{n-1} + b$).

Three ways to solve $x_n = ax_{n-1} + b$

1. *Direct solution.*

$$
\begin{aligned}
x_1 &= ax_0 + b,\\
x_2 &= a(ax_0 + b) + b = a^2x_0 + (1+a)b,\\
x_3 &= a\big(a^2x_0 + (1+a)b\big) + b = a^3x_0 + (1+a+a^2)b,\\
&\vdots\\
x_n &= a^nx_0 + (1+a+\cdots+a^{n-1})b.
\end{aligned}
$$

Three ways to solve $x_n = ax_{n-1} + b$ cont'd

2. *Use the theory developed in Discrete Mathematics.*
The associated homogeneous linear recurrence relation $x_n = ax_{n-1}$ has the solution $x_n = c\,a^n$, $c \in \mathbb{R}$.
A particular solution $x_n^{(p)}$ can be found by trying a constant $x_n^{(p)} = x$ and solving the resulting equation $x = ax + b$.
This gives $x = \frac{b}{1-a}$, and the general solution is therefore

$$x_n = c\,a^n + \frac{b}{1-a} \quad (c \in \mathbb{R}), \quad \text{provided that } a \neq 1.$$

If $a = 1$ (the "resonance case"), we have $x_n = nb + x_0$.

3. *Use variation of parameters.*
Setting $x_n = c(n)a^n = c_n a^n$, we have

$$x_n = c_n a^n = a \cdot c_{n-1} a^{n-1} + b \iff c_n - c_{n-1} = ba^{-n}.$$

This gives

$$c_n = \sum_{k=1}^{n} ba^{-k} + c_0 \quad \text{and} \quad x_n = c_0\,a^n + \sum_{k=1}^{n} ba^{n-k}.$$

Exercise

Verify that Methods 1 and 2 for solving $x_n = ax_{n-1} + b$ actually yield the same solution (although this is not directly visible in the formulas).

Hint: In the formula derived in Method 2, determine c in terms of x_0.

Math 285
Introduction to Differential Equations

Thomas Honold

ZJU-UIUC Institute

Spring Semester 2024

Outline

Today's Lecture:

Separable Equations

Definition

An (explicit) first-order ODE $y' = f(x, y)$ is said to be *separable* if $f(x, y)$ factors as $f(x, y) = f_1(x)f_2(y)$

We assume that the domains I, J of f_1, resp., f_2 are open intervals and that f_2 has no zero in J. Then $N(y) = 1/f_2(y)$ is well-defined and has no zero in J as well.

Writing $M = f_1$, we can rewrite $y' = f_1(x)f_2(y)$ as

$$y' = \frac{\mathrm{d}y}{\mathrm{d}x} = \frac{M(x)}{N(y)} \quad \text{or} \quad M(x)\,\mathrm{d}x - N(y)\,\mathrm{d}y = 0.$$

Theorem

Suppose $M\colon I \to \mathbb{R}$ and $N\colon J \to \mathbb{R}$ are continuous and N has no zero in J. Let $(x_0, y_0) \in I \times J$, and define $H_1\colon I \to \mathbb{R}$, $H_2\colon J \to \mathbb{R}$ by

$$H_1(x) = \int_{x_0}^{x} M(\xi)\mathrm{d}\xi, \qquad H_2(y) = \int_{y_0}^{y} N(\eta)\mathrm{d}\eta.$$

Let further $I' \subseteq I$ be an interval with $x_0 \in I'$ and $H_1(I') \subseteq H_2(J)$. Then there exists a unique solution $y\colon I' \to \mathbb{R}$ of the IVP $y' = M(x)/N(y) \wedge y(x_0) = y_0$, viz. $y(x) = H_2^{-1}\big(H_1(x)\big)$ for $x \in I'$.

Remark

The subsequent proof (cf. the notes thereafter) shows that for sufficiently small $\delta > 0$ the interval $I' = (x_0 - \delta, x_0 + \delta)$ has the required property and hence that the IVP $y' = M(x)/N(y) \wedge y(x_0) = y_0$ has locally near (x_0, y_0) a (unique) solution.

Proof of the theorem.

Since N is continuous and has no zero in J, we have either $N > 0$ or $N < 0$ on J and hence that H_2 is either strictly increasing or strictly decreasing on J. In particular, $H_2 \colon J \to H_2(J)$ is bijective and $y \colon I' \to \mathbb{R}$, $x \mapsto H_2^{-1}\big(H_1(x)\big)$ is well-defined.

$$y'(x) = (H_2^{-1})'\big(H_1(x)\big) \cdot H_1'(x) = \frac{H_1'(x)}{H_2'\big(H_2^{-1}(H_1(x))\big)} = \frac{M(x)}{N\big(y(x)\big)},$$

i.e., $y(x)$ satisfies $y' = M(x)/N(y)$
$H_1(x_0) = 0 = H_2(y_0) \Longrightarrow y(x_0) = H_2^{-1}\big(H_1(x_0)\big) = y_0$
It remains to show that any solution $y \colon I' \to \mathbb{R}$ of the IVP must satisfy $H_2\big(y(x)\big) = H_1(x)$ for $x \in I'$.

Proof cont'd.

To this end we write the ODE in the form $y'(x)N\big(y(x)\big) = M(x)$ and integrate:

$$\int_{x_0}^{x} y'(\xi)N\big(y(\xi)\big)\mathrm{d}\xi = \int_{x_0}^{x} M(\xi)\mathrm{d}\xi = H_1(x)$$

Making the substitution $\eta = y(\xi)$, $\mathrm{d}\eta = y'(\xi)\mathrm{d}\xi$ on the left-hand side gives

$$\int_{y(x_0)}^{y(x)} N(\eta)\mathrm{d}\eta = \int_{y_0}^{y(x)} N(\eta)\mathrm{d}\eta = H_2\big(y(x)\big),$$

as desired. □

Notes

- The proof has shown that $H_2(J)$ is an open interval containing $0 = H_2(y_0)$. Since H_1 is continuous and $H_1(x_0) = 0$, there exists $\delta > 0$ such that $H_1(x) \in H_2(J)$ for $x_0 - \delta < x < x_0 + \delta$, justifying the remark made before the proof.

Notes cont'd

- If the integrals in $H_2(y) = \int_{y_0}^{y} N(\eta)\mathrm{d}\eta = \int_{x_0}^{x} M(\xi)\mathrm{d}\xi = H_1(x)$ can be evaluated, we obtain $y = y(x)$ in impicit form $H_2(y) = H_1(x)$. The condition $H_1(I') \subseteq H_2(J)$ guarantees that this equation has a solution $y \in J$ for each $x \in I'$. If we are lucky, we may be able to solve for y and obtain an explicit formula for $y(x)$.
- The notation used in [BDM17], Ch. 2.2 is the same except that N, H_2 are replaced by $-N$, $-H_2$ to put the implicit ODE into the more symmetric form $M(x)\,\mathrm{d}x + N(y)\,\mathrm{d}y = 0$.

Example

We determine all solutions of the ODE $y' = dy/dt = t\,y^2$, which is separable with $f_1(t) = t$, $f_2(y) = y^2$.

One solution is the steady-state solution $y \equiv 0$.

For f_1 there is no restriction, and hence $I = \mathbb{R}$ in the theorem. Since $f_2(y) = 0$ is not allowed in the theorem, we split the domain of f_2 into the intervals $J_1 = (-\infty, 0)$, $J_2 = (0, +\infty)$. This corresponds to initial values $y(t_0) < 0$ and $y(t_0) > 0$, respectively.

Example (cont'd)

Rewriting the ODE formally as $\mathrm{d}y\,/y^2 = t\,\mathrm{d}t$ and integrating gives

$$\int_{y_0}^{y} \frac{\mathrm{d}\eta}{\eta^2} = \int_{t_0}^{t} \tau \mathrm{d}\tau$$

$$\frac{1}{y_0} - \frac{1}{y} = \left[-\frac{1}{\eta}\right]_{y_0}^{y} = \left[\tfrac{1}{2}\tau^2\right]_{t_0}^{t} = \tfrac{1}{2}(t^2 - t_0^2)$$

$$\Longrightarrow y(t) = \frac{1}{1/y_0 - \frac{1}{2}(t^2 - t_0^2)} = \frac{2}{2/y_0 + t_0^2 - t^2}$$

The non-constant solutions of $y' = t\,y^2$ are therefore $y(t) = y_C(t) = 2/(C - t^2)$, $C \in \mathbb{R}$. The solution y_C

- is defined for all $t \in \mathbb{R}$ if $C < 0$ or, equivalently, $-2/t_0^2 < y_0 < 0$ ($y_0 < 0$ for $t_0 = 0$);
- is defined only on the finite interval $(-\sqrt{C}, \sqrt{C})$ if $C > 0 \wedge |t_0| < \sqrt{C}$ or, equivalently, $y_0 > 0$;
- is defined only on $(-\infty, -\sqrt{C})$ if $C \geq 0 \wedge t_0 < -\sqrt{C}$ or, equivalently, $t_0 < 0 \wedge y_0 \leq -2/t_0^2$;
- is defined only on $(\sqrt{C}, +\infty)$ if $C \geq 0 \wedge t_0 > \sqrt{C}$ or, equivalently, $t_0 > 0 \wedge y_0 \leq -2/t_0^2$.

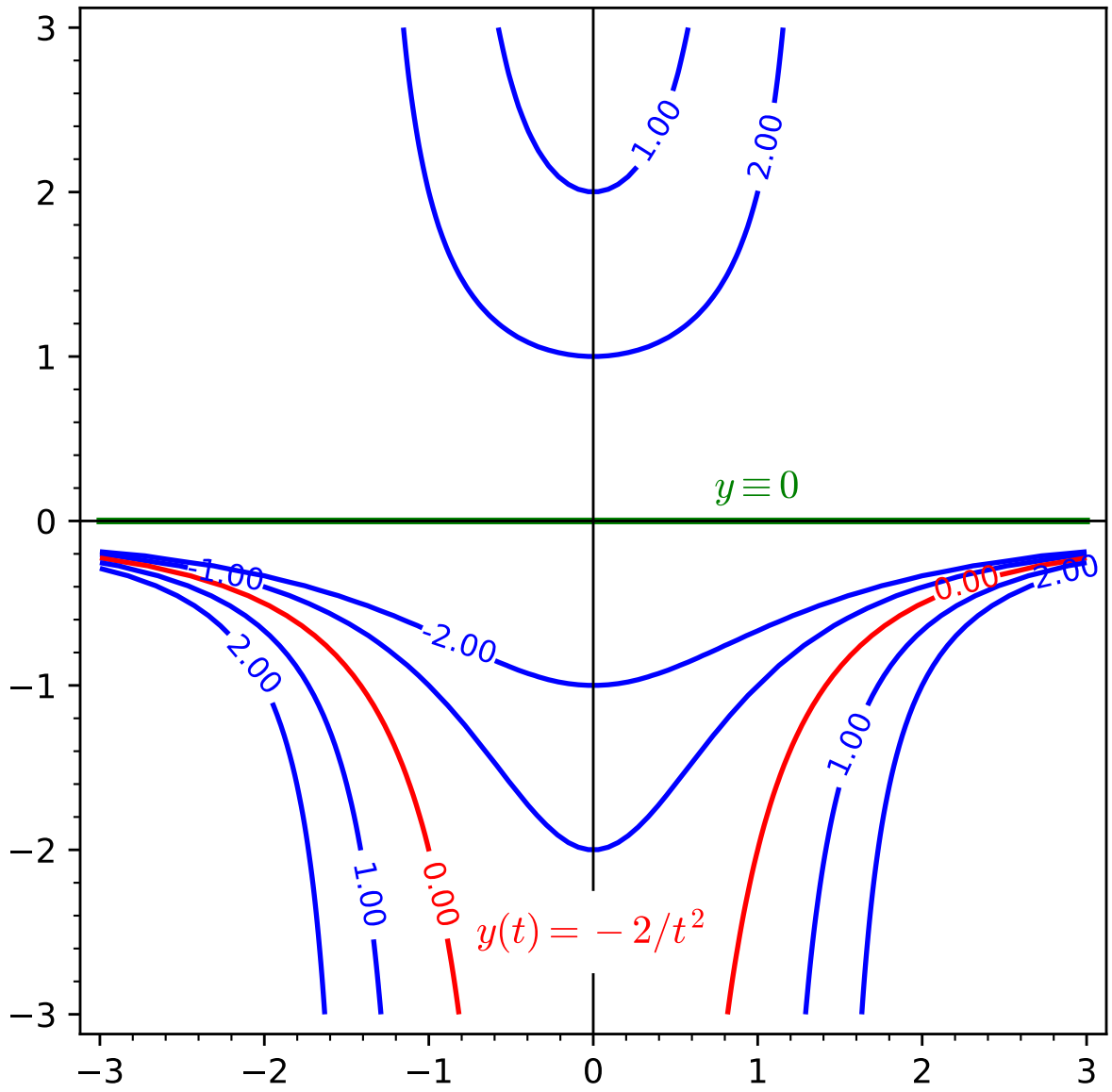


Figure: Solution curves $y_C(t) = 2/(C - t^2)$ of $y' = t\,y^2$

Simplified Variant (but keep the derivation in mind!)

It is often easier to use indefinite integration to determine the general solution of a separable 1st-order ODE as a 1-parameter family and then adapt the constant to satisfy a given initial condition:

$$y'(t)N(y(t)) = M(t)$$

$$\Longrightarrow \int y'(t)N(y(t))\,\mathrm{d}t = \int M(t)\,\mathrm{d}t + C, \quad C \in \mathbb{R}$$

$$\Longrightarrow \int N(y)\,\mathrm{d}y = \int M(t)\,\mathrm{d}t + C, \quad C \in \mathbb{R},\ y = y(t)$$

Memorizing the ODE as $\mathrm{d}y\,/\,\mathrm{d}t = M(t)/N(y)$ and formally rewriting it as $N(y)\,\mathrm{d}y = N(t)\,\mathrm{d}t$, we can directly short-circuit to the previous line:

$$N(y)\,\mathrm{d}y = M(t)\,\mathrm{d}t$$

$$\Longrightarrow \int N(y)\,\mathrm{d}y = \int M(t)\,\mathrm{d}t + C$$

In our present example: "$\mathrm{d}y\,/y^2 = t\,\mathrm{d}t \Longrightarrow -1/y = t^2/2 + C$", leading again to $y = -\frac{1}{C+t^2/2} = \frac{2}{-2C-t^2} = \frac{2}{C'-t^2}$, $C' \in \mathbb{R}$.

Example

We determine the solution of the IVP $mv' = mg - kv^2 \wedge v(0) = 0$ (best of the three models for a falling object).

Separating the variables gives

$$\frac{v'}{g - (k/m)v^2} = 1$$

$$\Longrightarrow \int_0^v \frac{\mathrm{d}\eta}{g - (k/m)\eta^2} = \int_0^t \mathrm{d}\tau = t$$

Since $\int \frac{\mathrm{d}x}{1-x^2} = \operatorname{artanh}(x) + C$, using the substitution $x = \sqrt{k/(mg)}\,\eta$ and $\tanh(y) = \frac{\mathrm{e}^y - \mathrm{e}^{-y}}{\mathrm{e}^y + \mathrm{e}^{-y}} = 1 - \frac{2}{\mathrm{e}^{2y}+1}$, we obtain

$$v(t) = \sqrt{\frac{mg}{k}} \tanh\left(t\sqrt{\frac{gk}{m}}\right) = \sqrt{\frac{mg}{k}}\left(1 - \frac{2}{\mathrm{e}^{2t\sqrt{gk/m}} + 1}\right),$$

for $0 \le t \le T$ (the time when the object hits the ground).

Setting $v_\infty = \sqrt{\frac{mg}{k}}$ (*limiting velocity*), this can also be written as

$$v(t) = v_\infty\left(1 - \frac{2}{\mathrm{e}^{2t\,g/v_\infty} + 1}\right), \quad 0 \le t \le T.$$

Example (cont'd)

Reasonable values for a skydiver (S) and a parachutist (P) with round canopy are $v_\infty = 50\,\mathrm{m/s}$ and $v_\infty = 5\,\mathrm{m/s}$, respectively, which gives

$$v_S(t) = 50\left(1 - \frac{2}{e^{0.4\,t} + 1}\right) \quad [\mathrm{m/s}]\,,$$

$$v_P(t) = 5\left(1 - \frac{2}{e^{4\,t} + 1}\right) \quad [\mathrm{m/s}]\,,$$

when time is measured in seconds.
This agrees well with experimentally found data.

Remark

Here, in contrast with the 2nd model, we can compute T in closed form: Denoting by $s(t)$ the distance traveled at time t, we have

$$s(t) = \frac{m}{k}\log\cosh\left(t\sqrt{\frac{gk}{m}}\right),$$

$$t(s) = \sqrt{\frac{m}{gk}}\,\operatorname{arcosh}\left(e^{sk/m}\right),$$

and $T = t(s_0)$ if the object is released at height s_0.

Exercise

a) Show that in the 3rd model for a falling object released at height s_0 the terminal velocity v_T of the object at time of impact is given by

$$v_T = \sqrt{\frac{mg}{k}} \cdot \sqrt{1 - \mathrm{e}^{-2ks_0/m}}.$$

Hint: Consider the velocity as a function $v(s)$ of the distance s traveled. Show that $y(s) = v(s)^2$ satisfies the ODE $my' = 2mg - 2ky$.

b) The limiting velocity of a falling basketball ($m = 620\,\mathrm{g}$) has been estimated at $20\,\mathrm{m/s}$. Using this data, graph v_T as a function of s_0. For which heights s_0 does the basketball reach 50 %, 90 %, and 99 % of its limiting velocity?

Example

We solve $y' = y^2$, which is autonomous and hence separable with $f_1(t) = 1$, $f_2(y) = y^2$, using the simplified solution method.

There is the constant solution $y \equiv 0$.
Otherwise we can rewrite $\mathrm{d}y \,/\, \mathrm{d}t = y^2$ as $\mathrm{d}y \,/ y^2 = \mathrm{d}t$ and obtain

$$\frac{\mathrm{d}y}{y^2} = \mathrm{d}t$$

$$\Longrightarrow \; -\frac{1}{y} = \int \frac{\mathrm{d}y}{y^2} = \int \mathrm{d}t = \int 1 \,\mathrm{d}t = t + C$$

$$\Longrightarrow \; y = \frac{1}{-C - t} = \frac{1}{C' - t}, \quad C, C' \in \mathbb{R}.$$

This recovers the already known general solution.

But don't forget: The informal computation is justified by rewriting it in terms of $y(t)$ and using the subsitution $\eta = y(t)$:

$$\frac{y'(t)}{y(t)^2} = 1$$

$$\Longleftrightarrow \; \frac{-1}{y(t)} = \frac{-1}{\eta} = \int \frac{\mathrm{d}\eta}{\eta^2} = \int \frac{y'(t)}{y(t)^2} \,\mathrm{d}t = \int \mathrm{d}t = t + C$$

Example (cont'd)

This tells us that the solutions $y\colon I \to \mathbb{R}$ of $y' = y^2$ with $y(t) \neq 0$ for all $t \in I$ are precisely the functions whose graph is contained in a contour of $F(t, y) = -1/y - t$, i.e., satisfy $F\big(t, y(t)\big) = C$ for some $C \in \mathbb{R}$.

It doesn't tell us whether such functions actually exist.

However, in this particular case we can solve for y to show that precisely the functions $y(t) = 1/(C - t)$, $C \in \mathbb{R}$ (defined on an appropriate interval I) have this property.

In the case of a general separable ODE we can't solve for y and must invoke the theorem on separable ODE's to conclude the local existence and uniqueness of solutions for any prescribed initial value $y(t_0) \neq 0$. (The Implicit Function Theorem also yield this, cf. subsequent remark.)

Example

The ODE $y' = \sqrt{|y|}$ can of course also be solved by the new method:

A solution $y : I \to \mathbb{R}$ with $y(t) \neq 0$ for all $t \in I$ must satisfy either $y > 0$ on I or $y < 0$ on I.

$\underline{y > 0}$: In this case $y' = \sqrt{y}$ and we get

$$\frac{dy}{\sqrt{y}} = 1\,\mathrm{d}t \iff 2\sqrt{y} = t + C \iff y = \frac{(t+C)^2}{4}, \quad C \in \mathbb{R}.$$

Because of the middle equation, we must have $t > -C$, i.e., $I \subseteq (-C, +\infty)$.

$\underline{y < 0}$: Here $y' = \sqrt{-y}$ and we get

$$\frac{dy}{\sqrt{-y}} = 1\,\mathrm{d}t \iff -2\sqrt{-y} = t+C \iff y = -\frac{(t+C)^2}{4}, \; C \in \mathbb{R},$$

and $t < -C$, i.e., $I \subseteq (-\infty, -C)$.

The guaranteed uniqueness of solutions applies only to the regions $y > 0$ and $y < 0$ in the (t, y)-plane and doesn't exlude the observed branching of solutions on the t-axis.

General Remarks on $y' = f_1(x)f_2(y)$

Extracted from the previous examples

We assume that $f_1 : I \to \mathbb{R}$, $f_2 : J \to \mathbb{R}$ are continuous functions on open intervals $I, J \subseteq \mathbb{R}$. Thus $I \times J$ is an open rectangle with possibly infinite sides.

1. The zeros of f_2 (if any) partition J into open subintervals on which f_2 has no zeros. If J' is such a subinterval then on the rectangle $I \times J'$ we have local existence and uniqueness of solutions of IVP's $y' = f_1(x)f_2(y) \wedge y(x_0) = y_0$ at any point $(x_0, y_0) \in I \times J'$.
2. Rewriting $y' = f_1(x)f_2(y)$ as $y' = M(x)/N(y)$ and denoting by $F : I \times J' \to \mathbb{R}$ an antiderivative of $M(x)\,\mathrm{d}x - N(y)\,\mathrm{d}y$ (i.e., $\partial F/\partial x = M \wedge \partial F/\partial y = -N$), the solutions $y(x)$ with graph $\mathrm{G}_y \subset I \times J'$ are given in implicit form as $F(x, y) = C$, $C \in \mathbb{R}$.

The function F in (2) can be chosen as

$$F(x, y) = \int_{x_0}^{x} M(\xi)\mathrm{d}\xi - \int_{y_0}^{y} N(\eta)\mathrm{d}\eta, \quad (x_0, y_0) \in I \times J'.$$

In particular the differential 1-form $M(x)\,\mathrm{d}x - N(y)\,\mathrm{d}y$ is exact on $I \times J'$ (which also follows form $M_y = N_x = 0$ and the shape of $I \times J'$).

Remarks on $y' = f_1(x)f_2(y)$ Cont'd

3. For any zero y_0 of f_2 there is the steady-state solution $y(x) \equiv y_0$ on I. Together with (1) this shows that all IVP's $y' = f_1(x)f_2(y) \wedge y(x_0) = y_0$ with $(x_0, y_0) \in I \times J$ are solvable.

Linear Versus Separable 1st-Order ODE's

Note the following important differences between the two cases.

1. Domains of $y' = a(x)y + b(x)$ are of the form $I \times \mathbb{R}$; domains of $y' = f_1(x)f_2(y)$ are of the form $I \times J$, where J may be a proper subinterval of $\mathbb{R}$.
2. Solutions of $y' = a(x)y + b(x)$ can be extended to I (i.e., maximal solutions have domain I); solutions of $y' = f_1(x)f_2(y)$ may be defined only on proper subintervals $I' \subset I$, which depend on the solution and are not visible in the ODE.
3. Solutions of IVP's $y' = a(x)y + b(x) \wedge y(x_0) = y_0$ are unique in the sense that if $y_1 : I_1 \to \mathbb{R}$, $y_2 : I_2 \to \mathbb{R}$, solve the IVP then $y_1(x) = y_2(x)$ for all $x \in I_1 \cap I_2$; solutions of IVP's $y' = f_1(x)f_2(y) \wedge y(x_0) = y_0$ are unique only at points (x_0, y_0) with $f_2(y_0) \neq 0$, and only if their ranges don't contain zeros of f_2.

The Logistic Equation

Definition

The ODE $y' = ay - by^2$ with constants $a, b > 0$ is called *logistic equation*.

The logistic equation was introduced by the Belgian mathematician P. VERHULST (1804–1849) in 1837 as a mathematical model for population growth. It provides a more accurate model of population growth than the exponential model $y' = ay$, adding a term $-by^2$, which accounts for the competition between individuals if resources are limited.

The logistic equation has the form $y' = f_1(t)f_2(y)$ with $f_1(t) = 1$, $f_2(y) = ay - by^2$, and hence is separable (even autonomous).

Since $ay - by^2 = y(a - by)$ the steady-state solutions are $y \equiv 0$ and $y \equiv a/b$.

We determine the general solution by the usual method. Since

$$\frac{1}{y(a-by)} = \frac{1/a}{y} + \frac{b/a}{a-by},$$

e.g., by the method of partial fractions, we obtain

$$\begin{aligned}
\int \frac{1}{y} + \frac{b}{a-by}\,\mathrm{d}y &= \int a\,\mathrm{d}t + C \\
\ln|y| - \ln|a-by| &= at + C \\
\ln\left|\frac{y}{a-by}\right| &= at + C \\
\pm\frac{y}{a-by} &= \mathrm{e}^{at+C} \\
\pm y &= \mathrm{e}^{at+C}(a-by) \\
y &= \frac{a\,\mathrm{e}^{at+C}}{\pm 1 + b\,\mathrm{e}^{at+C}} = \frac{a}{\pm\mathrm{e}^{-C}\mathrm{e}^{-at} + b}.
\end{aligned}$$

Setting $d = \pm\mathrm{e}^{-C}$, we obtain the solution $\boxed{y(t) = \dfrac{a}{d\mathrm{e}^{-at} + b},\ \ d \in \mathbb{R}.}$

$d = 0$ gives the steady-state $y \equiv a/b$ (and $d = \infty$ gives $y \equiv 0$).

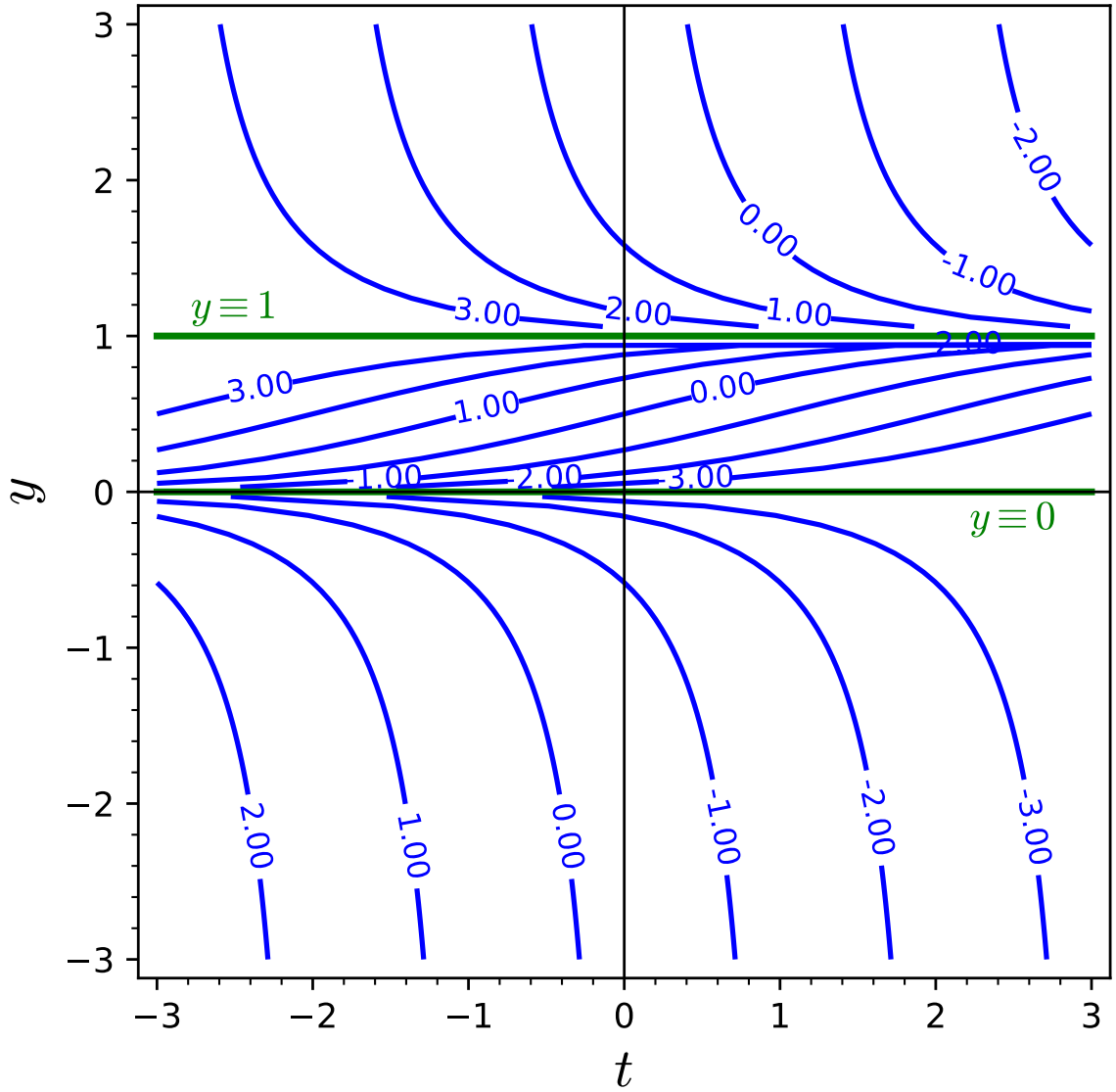


Figure: Solution curves of $y' = y - y^2$, represented as level sets $F(t, y) = \ln\left|\frac{y}{1-y}\right| - t = C$

Asymptotic behaviour

Observation

For every $d \in \mathbb{R}$ we have

$$\lim_{t\to+\infty} \frac{a}{de^{-at}+b} = \frac{a}{b}, \qquad \lim_{t\to-\infty} \frac{a}{de^{-at}+b} = 0.$$

Caution

This does not imply that all solutions $y(t)$ to the logistic equation exist at any time t and have the indicated limits for $t \to \pm\infty$.

The precise asymptotics are given on the next slide.

Since $y' = ay - by^2$ is autonomous, horizontal shifts $t \mapsto y(t - t_0)$ of solutions $y(t)$ are again solutions and we can assume w.l.o.g. that $y(t)$ is defined at $t_0 = 0$. As usual, we set $y(0) = y_0$.

In terms of y_0, the parameter d is given by

$$\frac{a}{d+b} = y_0, \quad \text{i.e.,} \quad d = \frac{a}{y_0} - b.$$

The solution with $d = -b$ is not defined at $t = 0$.

Asymptotic behaviour cont'd

1. Solutions $y(t)$ with $d > 0$ or, equivalently, $0 < y_0 < a/b$ exist at any time t (i.e., have maximal domain $\mathbb{R}$) and for $t \to \pm\infty$ have the limits indicated on the previous slide.
2. Solutions $y(t)$ with $d < 0$ have two branches and a vertical asymptote at $t_\infty = \big(\ln(-d) - \ln b\big)/a$, which is the solution of $de^{-at} + b = 0$.

   (2.1) If $-b < d < 0$, we have $t_\infty < 0$ and the branch defined at $t = 0$ has domain $(t_\infty, +\infty)$; moreover, $\lim_{t \downarrow t_\infty} = +\infty$, $\lim_{t \to +\infty} y(t) = a/b$. All solutions satisfying $y_0 > a/b$ arise in this way (with $d = a/y_0 - b$).

   (2.2) If $d < -b$, we have $t_\infty > 0$ and the branch defined at $t = 0$ has domain $(-\infty, t_\infty)$; moreover, $\lim_{t \to -\infty} y(t) = 0$, $\lim_{t \uparrow t_\infty} y(t) = -\infty$. All solutions satisfying $y_0 < 0$ arise in this way (with $d = a/y_0 - b$).

The remaining solutions defined at $t = 0$ are the two steady-state solutions $y(t) \equiv 0$ ($d = \infty$) and $y(t) \equiv a/b$ ($d = 0$).

$\Longrightarrow$ The solutions (single branches!) defined at $t = 0$ are in 1-1 correspondence with $d \in \mathbb{R} \setminus \{-b\} \cup \{\infty\}$.

But there are further solutions (the 2nd branches of the solutions for $d < 0$, $d \neq -b$, and both branches for $d = -b$).

Up to horizontal shifts, there are only 3 essentially different solutions:

$$y_1(t) = \frac{a}{b(1 + \mathrm{e}^{-at})}, \qquad t \in \mathbb{R},$$

$$y_2(t) = \frac{a}{b(1 - \mathrm{e}^{-at})}, \qquad t \in (-\infty, 0),$$

$$y_3(t) = \frac{a}{b(1 - \mathrm{e}^{-at})}, \qquad t \in (0, +\infty).$$

We also see that the corresponding graphs ("integral curves") depend only on the quotient a/b.

The Case $d > 0$

For applications to population growth only Cases 1 and 2 are interesting. Information about the solution graphs can easily be obtained from the logistic equation:

$$y' = ay - by^2 = y(a - by),$$
$$y'' = ay' - 2byy' = y'(a - 2by)$$

1. Solutions $y(t)$ with $0 < y(0) < a/b$ are strictly increasing (since they satisfy $0 < y(t) < a/b$ for all $t \in \mathbb{R}$). Denoting by t_h the unique solution of $de^{-at} = b$, i.e. $t_h = (\ln d - \ln b)/a$, we have $y(t_h) = \frac{a}{de^{-at_h}+b} = a/2b$ and further that $y(t)$ is convex in $[-\infty, t_h]$ (since $0 < y(t) < a/2b$ in this interval) and concave in $[t_h, +\infty]$. In particular $y(t)$ has a (unique) inflection point in $(t_h, a/2b)$.
2. Solutions $y(t)$ with $y(0) > a/b$ are strictly decreasing and convex in their domain $[t_\infty, +\infty)$.

Since the logistic equation is autonomous, in Case 1 every solution arises from the solution with $t_h = 0$ (i.e., $d = b$) by a time shift. This is visible in $y(t) = a/(de^{-at} + b) = a/(be^{-a(t-t_h)} + b)$.

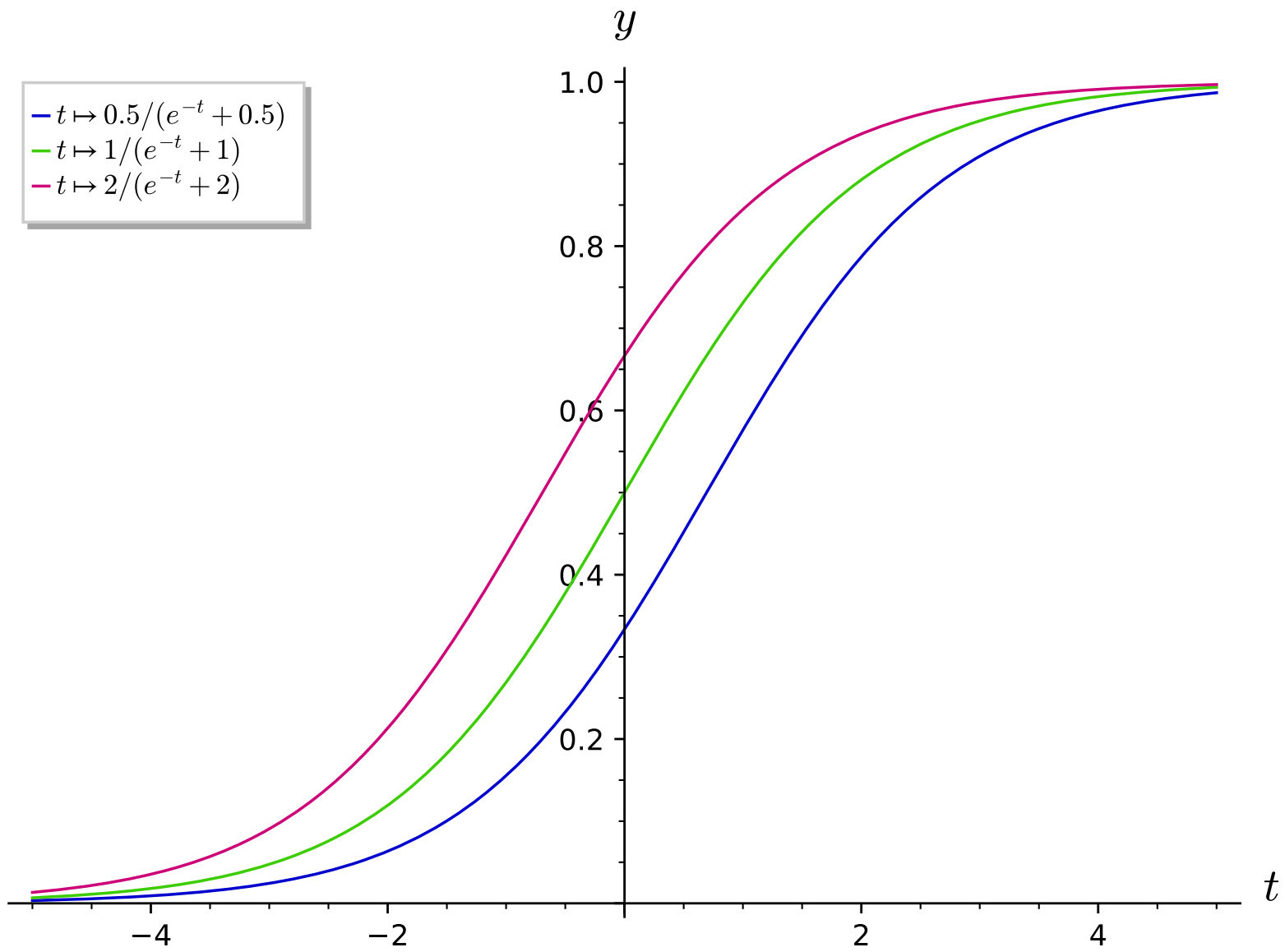


Figure: Three S-curves following the Logistic Law with $a/b = 1$ and $d > 0$

Population of the Earth

The US Department of Commerce estimated in 1965 the world's population at 3.34 billion people, with an annual increase of 2 % per year. Using the exponential model $y' = ay$, this gives

$$y(t) = 3.34 \cdot 10^9 \times \mathrm{e}^{0.02(t-1965)}.$$

In this model the population would double every $\frac{\ln 2}{0.02} \approx 34.6$ years.

The logistic model $y' = ay - by^2$ with the reasonable parameter $a = 0.029$ (natural reproduction rate, if unlimited resources are available) and b, d' computed from

$$\frac{y'(1965)}{y(1965)} = a - b\,y(1965) = a - b \times 3.34 \cdot 10^9 = 0.02,$$

$$y(1965) = \left.\frac{a}{d'\mathrm{e}^{-a(t-1965)} + b}\right|_{t=1965} = \frac{a}{d' + b}$$

i.e. $b = 2.695 \cdot 10^{-12}$, $d' = 5.988 \cdot 10^{-12}$, gives

$$y(t) = \frac{0.029 \cdot 10^{12}}{5.988\,\mathrm{e}^{-0.029(t-1965)} + 2.695}, \quad y(2020) = 7.42 \cdot 10^9, \quad \frac{a}{b} = 10.76 \cdot 10^9.$$

Uniqueness of Solutions

So far we have proved uniqueness of solutions of initial value problems $y' = G(t, y) \wedge y(t_0) = y_0$ in the following two ways:

1. Derive the general solution of $y' = G(t, y)$ and observe that it is a 1-parameter family of functions $y_C(t)$ depending on a constant C; plug in $y_C(t_0) = y_0$ to determine C, and hence the solution, uniquely.
2. If the solution to $y' = G(t, y)$ involves more than one parameter, show additionally that an initial condition $y(t_0) = y_0$ cannot be satisfied by solutions corresponding to different parameters.

Way (1) applies to 1st-order linear ODE's (homogeneous or inhomogeneous) and to separable ODE's without steady-state solutions.

Way (2) applies to separable ODE's with steady-state solutions, such as $y' = y^2$, $y' = ty^2$, $y' = ay - by^2$.
"Different parameters" refers to both continuous 1-parameter families of solutions and "exceptional" steady-state solutions.

Neither way applies to $y' = \sqrt{|y|}$.

Uniqueness of Solutions Cont'd

Example

The logistic equation $y' = ay - by^2$ has the solutions $y_\infty(t) \equiv 0$ and $y_d(t) = \frac{a}{de^{-at}+b}$, $d \in \mathbb{R}$. We assume that the solutions are maximal, i.e., the domains are $\mathbb{R}$ for $d \geq 0$ and $\mathbb{R} \setminus \{t_\infty\}$ for $d < 0$. For $d < 0$ we count the two branches $y_d^{\pm}(t)$ as different solutions, according to our requirement that domains of ODE solutions should be intervals.

For $t_0 \in \mathbb{R}$ and $y_0 \neq 0$ we can solve $\frac{a}{de^{-at_0}+b} = y_0$ uniquely for d, showing that (t_0, y_0) is on precisely one solution curve (graph) $y_d(t)$, $d \in \mathbb{R}$. Moreover, since $\frac{a}{de^{-at}+b} \neq 0$, these solution curves don't intersect the steady-state solution $y_\infty(t) \equiv 0$. This implies that the solution curves $y_d(t)$, $d \in \mathbb{R} \cup \{\infty\}$, partition the (t, y)-plane, which is equivalent to the unique solvability of all IVP's $y' = ay - by^2 \wedge y(t_0) = y_0$ within the given class of functions.

However, this doesn't exclude the existence of further solutions. In fact there are no further solutions, and a rigorous proof is given on the next slide.

Example (cont'd)

The theorem on separable ODE's implies that there can't be two distinct solutions through a point (t_0, y_0) with $y_0 \notin \{0, a/b\}$, and hence all solutions not intersecting the lines $y = 0$, $y = a/b$ are known.

Now suppose there is a non-constant solution $y(t)$ satisfying $y(t_0) = 0$, say, for some $t_0 \in \mathbb{R}$. (The case $y(t_0) = a/b$ is done in the same way.)

W.l.o.g. we can assume that $0 < y(t) < a/b$ for $t_0 < t < t_0 + \delta$, where δ is some positive number. (By symmetry, we can assume that there exists $t_1 > t_0$ satisfying $y(t_1) > 0$. Since $y(t)$ is continuous, there exists a largest zero t^* of $y(t)$ in $[t_0, t_1]$. Then $y(t) > 0$ for $t^* < t < t_1$, and hence our assumption is satisfied if we replace t_0 by t^* and set $\delta = t_1 - t^*$.)

Now, by continuity, we must have $\lim_{t \downarrow t_0} y(t) = 0$, but none of the solutions that are defined for $t \in (t_0, t_0 + \delta)$ and attain small positive values there (these must be of the form $y_d(t)$ with $d > 0$) has this property, since $y_d(t) = \frac{a}{de^{-at} + b} \to \frac{a}{de^{-at_0} + b} \neq 0$ for $t \downarrow t_0$.

This contradiction completes the proof.

Example

The equation $y = \sqrt{|y|}$ has the steady-state solution $y(t) \equiv 0$ and the two 1-parameter families

$$y_c^-(t) = -\tfrac{1}{4}(t-c)^2, \qquad t \in (-\infty, c),$$
$$y_c^+(t) = \tfrac{1}{4}(t-c)^2, \qquad t \in (c, +\infty),$$

as solutions, where $c \in \mathbb{R}$ is arbitrary.
Collectively, these solutions partition the (t, y)-plane, so that every point $(t_0, y_0) \in \mathbb{R}^2$ is on exactly one solution curve of this kind. (This follows, e.g., from the theorem on separable ODE's.)
However, there are further (maximal) solutions obtained by glueing together $y_c^{\pm}(t)$ at $t = c$ (and other combinations as well), which leads to non-uniqueness of solutions of all IVP's $y = \sqrt{|y|} \wedge y(t_0) = y_0$. (The indicated combination shows this only for the points $(c, 0)$ on the t-axis, through which we have the solution combined from $y_c^{\pm}(t)$ and also the steady-state solution $y(t) \equiv 0$.)

Remark

The general Existence and Uniqueness Theorem for solutions of 1st-order ODE's (to be proved later) will explain the observed fundamental difference between the two examples and give a more conceptual proof of the uniqueness of solutions of all IVP's
$y' = ay - by^2 \wedge y(t_0) = y_0$ (and, similarly, of the uniqueness of solutions of all IVP's corresponding to the harvesting equation discussed subsequently).

The Harvesting Equation

Suppose a population follows the logistic law of growth but additionally individuals are removed ("harvested") at a constant rate $h > 0$.

Definition

The ODE $y' = ay - by^2 - h$ $(a, b, h > 0)$ is called *harvesting equation*.

Changes

- For $h < a^2/4b$ the quadratic $-by^2 + ay - h = 0$, whose discriminant is $\Delta = a^2 - 4bh$, still has two zeros, viz.

  $$y_1 = \big(a - \sqrt{a^2 - 4bh}\big)/2b, \quad y_2 = \big(a + \sqrt{a^2 - 4bh}\big)/2b,$$

  which satisfy $0 < y_1 < y_2$ and provide two steady-state solutions.
- For $h = a^2/4b$ the quadratic has a double root, which provides one steady-state solution $y \equiv a/2b$.
- $h > a^2/4b$ the quadratic has no real zeros, and the harvesting equation has no steady-state solutions.

For a more detailed analysis we transform the harvesting equation into canonical form.

Lemma

We can transform the harvesting equation by means of a substitution $y(t) = u\,z(mt) + v$ *with* $u, v, m \in \mathbb{R}$ *and* $u, m > 0$ *into one of the three canonical forms*

$$z' = -z^2 + 1, \quad z' = -z^2, \quad z' = -z^2 - 1.$$

Proof.

Writing $s = mt$ we have $y'(t) = mu\,z'(mt) = mu\,z'(s)$ and hence, using the usual shorthands

$$\begin{aligned} z' = \frac{y'}{mu} &= \frac{-b(uz+v)^2 + a(uz+v) - h}{mu} \\ &= -\frac{bu}{m}\,z^2 + \frac{a-2bv}{m}\,z + \frac{-bv^2 + av - h}{mu}. \end{aligned}$$

With $m = bu$, $v = a/2b$ this becomes

$$z' = -z^2 + \frac{\frac{-a^2}{4b} + \frac{a^2}{2b} - h}{bu^2} = -z^2 + \frac{a^2 - 4bh}{4b^2u^2} = -z^2 + \frac{\Delta}{4b^2u^2}.$$

Proof cont'd.

If $\Delta > 0$ ($\Delta < 0$) then $u = \sqrt{\Delta}/(2b)$ (resp., $u = \sqrt{-\Delta}/(2b)$) gives $z' = -z^2 + 1$ (resp., $z' = -z^2 - 1$). □

Notes

- Substitutions of the form $y(t) = u\,z(mt) + v$ ($u, v, m \in \mathbb{R}$, $u, v > 0$) arise from changing the units of measurement on both the t-axis and the y-axis and an additional vertical shift of the graph of $t \mapsto y(t)$. They do not change the overall shape of the solution graphs.
- Substitutions of this form do not change the number of steady-state solutions, and hence the corresponding canonical form is also determined by the number of zeros of $-by^2 + ay - h = 0$.
- The logistic equation $y' = ay - by^2$ has canonical form $z' = -z^2 + 1$ (regardless of the particular choice of $a, b > 0$).

Analysis of the Canonical Forms

In the following we will test "stability" of solutions $y(t)$ of the harvesting equation—a concept that describes their asymptotic behaviour for $t \to +\infty$.

Definition (Stability)

A steady-state solution $y \equiv y_0$ of an autonomous first-order ODE $y' = f(y)$ (i.e., $f(y_0) = 0$) is said to be *(asymptotically) stable* if there exists $\delta > 0$ such that every solution $y(t)$ of $y' = f(y)$ with initial value $y(0) \in [y_0 - \delta, y_0 + \delta]$ is defined for sufficiently large t and satisfies $\lim_{t\to+\infty} y(t) = y_0$, and *unstable* otherwise.

1. $z' = -z^2 + 1$.
   This is the logistic equation (without harvesting), with steady states $z \equiv \pm 1$.
   Our previous analysis shows that

$$\lim_{s\to+\infty} z(s) = \begin{cases} 1 & \text{if } z_0 > -1, \\ \text{undefined} & \text{if } z_0 < -1. \end{cases}$$

   Thus $z \equiv 1$ is stable and $z \equiv -1$ is unstable.

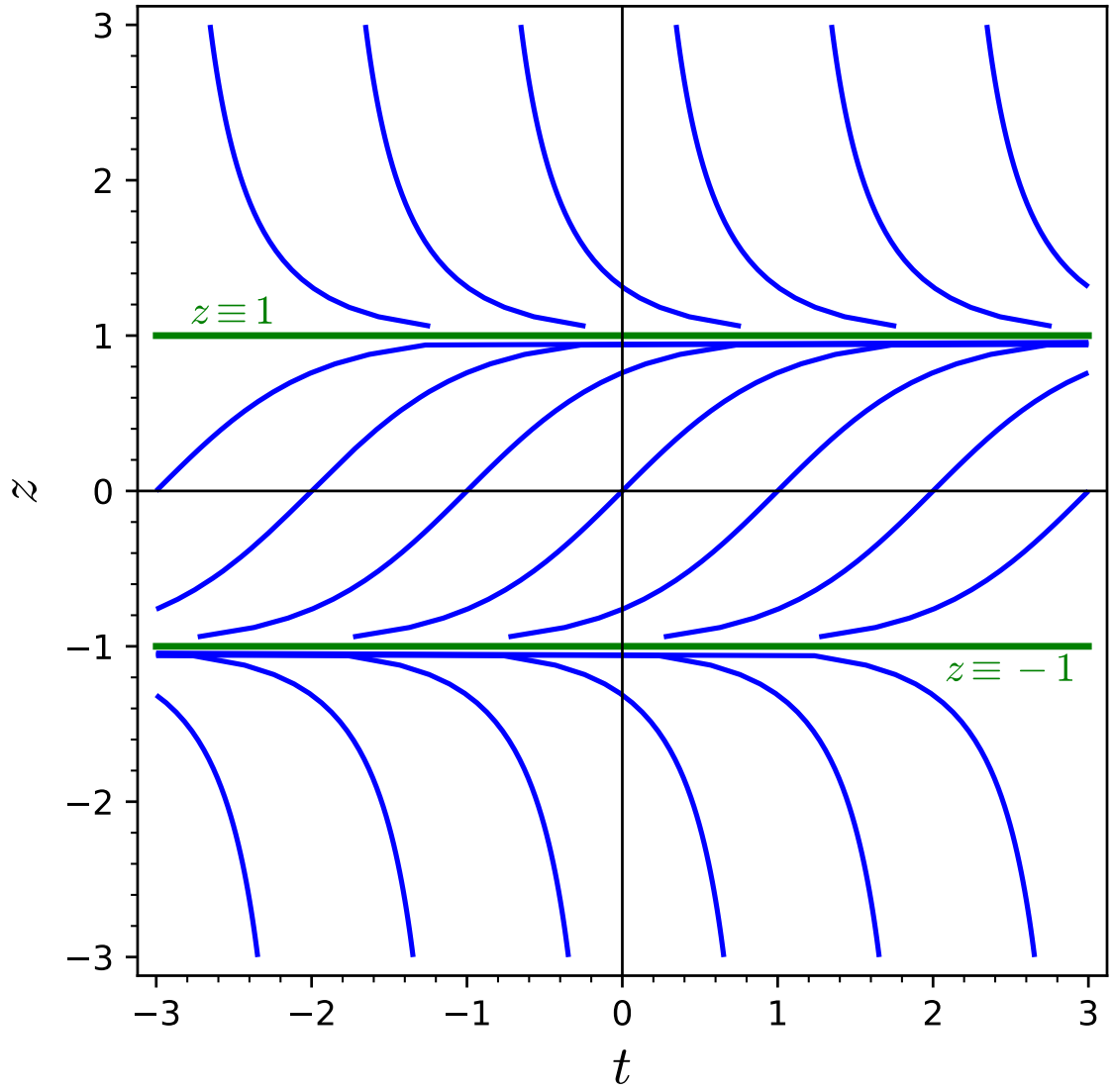


Figure: Solution curves of $z' = 1 - z^2$

Analysis of the Canonical Forms Cont'd

2. $z' = -z^2$.
   The standard solution method gives

$$\frac{1}{z} - \frac{1}{z_0} = \int_{z_0}^{z} -\frac{\mathrm{d}\zeta}{\zeta^2} = \int_{s_0}^{s} \mathrm{d}\sigma = s - s_0,$$

   i.e., $z(s) = 1/(s - C)$ with $C = s_0 - 1/z_0$.

   This tells us:

   Solutions $z(s)$ with $z(s_0) = z_0 > 0$ (equivalently, $s_0 > C$) exist forever and satisfy $\lim_{s\to+\infty} z(s) = 0$.

   Solutions $z(s)$ with $z(s_0) = z_0 < 0$ (equivalently, $s_0 < C$) exist only on $(-\infty, C)$ and satisfy $\lim_{s\uparrow C} z(s) = -\infty$.

   In other words, the steady-state solution $z \equiv 0$ is *one-sided stable* (*stable from above*).

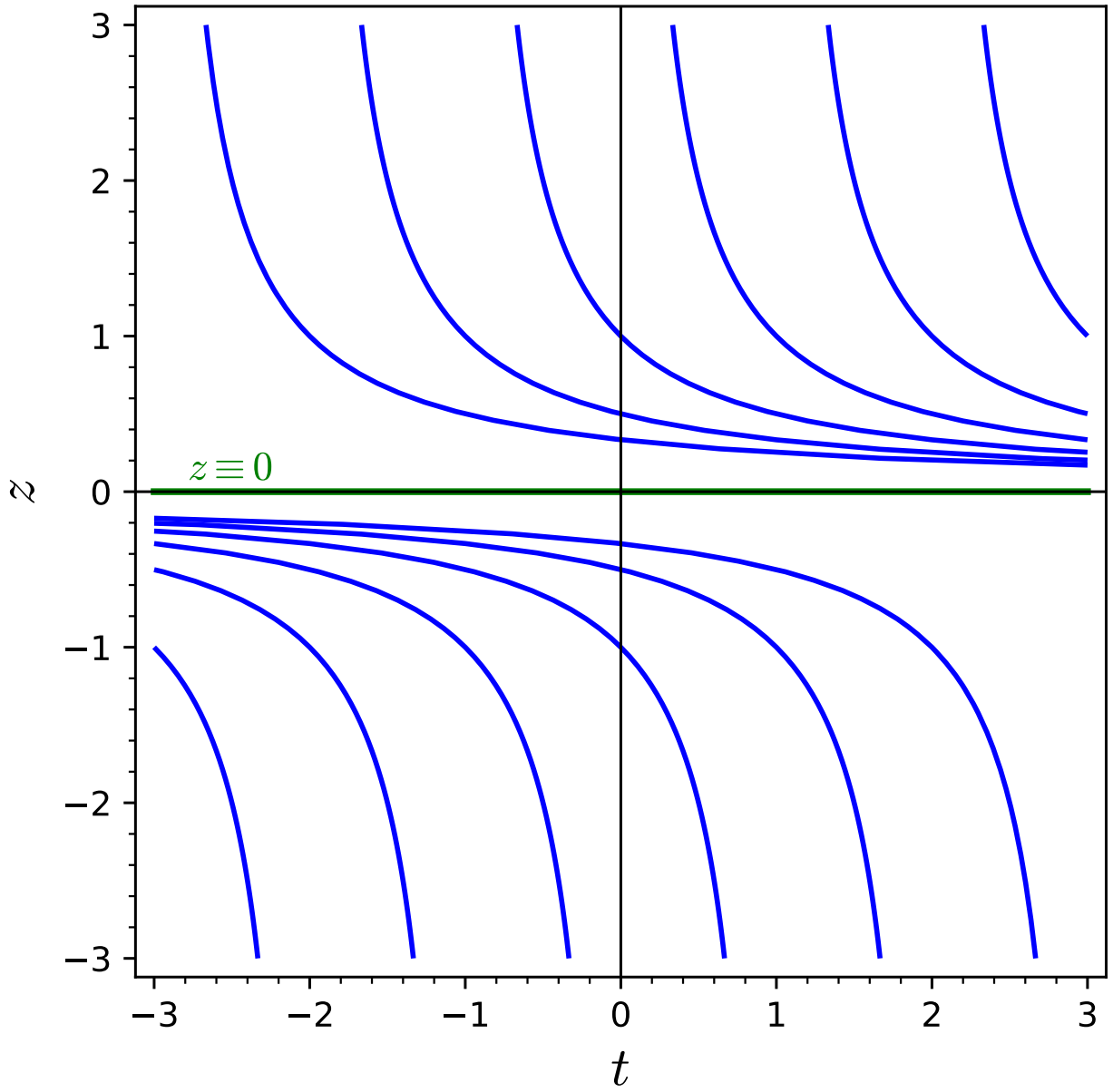


Figure: Solution curves of $z' = -z^2$

Analysis of the Canonical Forms Cont'd

3. $z' = -z^2 - 1$.
   Here the standard solution method gives

$$\arctan z_0 - \arctan z = \int_{z_0}^{z} -\frac{\mathrm{d}\zeta}{\zeta^2+1} = \int_{s_0}^{s} \mathrm{d}\sigma = s - s_0,$$

   i.e., $z(s) = \tan(C - s)$ with $C = s_0 + \arctan z_0$.

   This tells us:

   Solutions $z(s)$ with $z(s_0) = z_0$ exist only on $(C - \pi/2, C + \pi/2)$ and satisfy $\lim_{s \uparrow C+\pi/2} z(s) = -\infty$.

   The solutions with $z_0 > 0$ have $C > s_0$ and hence exist for a period larger than $\pi/2$, while those with $z_0 < 0$ have $C < s_0$ and exist for a period less than $\pi/2$.

   Since there are no steady-state solutions, the question of stability does not arise.

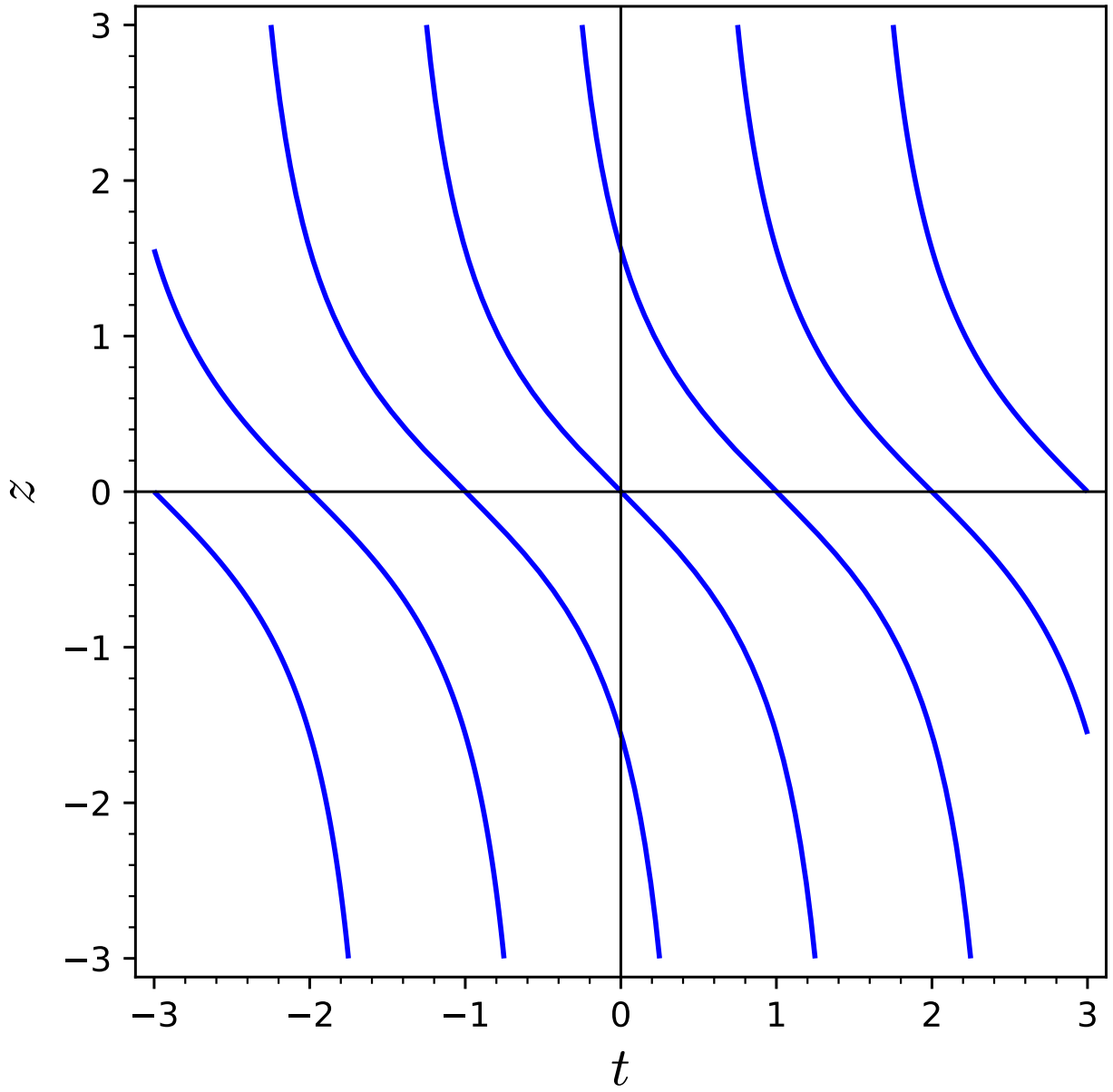


Figure: Solution curves of $z' = -1 - z^2$

Remark

It is instructive to represent the solution curves (except the steady-state solutions) in the preceding examples as function $t(y)$ resp. $s(z)$. This makes sense for any autonomous ODE and (provided the ODE can be integrated in closed form) often yields a simpler formula for the solution curves which better explains their shape.

Exercise

Show that the graph of $y(t) = a/(d\,\mathrm{e}^{-at} + b)$ $(a, b, d > 0)$ is point-symmetric to its inflection point.

Hint: A superb way to solve this exercise is to observe that the mirror image of a solution curve w.r.t. its inflection point represents a solution as well and use the uniqueness of solutions of associated IVP's.

Remark: For $d < 0$ graphs have a similar symmetry, but the meaning of the center of symmetry is different.

Finally we translate the results on the asymptotic behaviour back into the original harvesting equation $y' = ay - by^2 - h$ ($a, b, h > 0$). Recall that for $\Delta = a^2 - 4bh \geq 0$ there are the steady-sate solutions $y \equiv y_{1/2}$ with $y_1 = \left(a - \sqrt{a^2 - 4bh}\right)/2b$, $y_2 = \left(a + \sqrt{a^2 - 4bh}\right)/2b$, which satisfy $0 < y_1 \leq y_2$.

Analysis of the Harvesting Equation

$h < a^2/4b$ If the initial population $y(t_0)$ satisfies $y_1 < y(t_0) < y_2$ then the population $y(t)$ increases and $\lim_{t\to+\infty} y(t) = y_2$. If $y(t_0) > y_2$ then $y(t)$ decreases and $\lim_{t\to\infty} y(t) = y_2$. If $y(t_0) < y_1$ then $y(t)$ decreases and $y(t_1) = 0$ for some $t_1 > t_0$, i.e., the population dies out.

$h = a^2/4b$ If $y(t_0) > a/2b$, the population decreases and $\lim_{t\to\infty} y(t) = a/2b$. If $y(t_0) < a/2b$, the population decreases and dies out at some time $t_1 > t_0$.

$h > a^2/4b$ Regardless of the initial population, the population dies out in finite time.

Exact First-Order Equations

Definition

A first-order ODE of the form

$$M(x,y)\,\mathrm{d}x + N(x,y)\,\mathrm{d}y = 0 \tag{D}$$

with $M, N\colon D \to \mathbb{R}$, $D \subseteq \mathbb{R}^2$ open, is said to be *exact* if there exists a function $f\colon D \to \mathbb{R}$ satisfying $\mathrm{d}f = M(x,y)\,\mathrm{d}x + N(x,y)\,\mathrm{d}y$ or, equivalently, $\nabla f = (f_x, f_y) = (M, N)$.

Notes

- Criteria for exactness have been developed in Calculus III. Recall that for C^1-functions $M, N\colon D \to \mathbb{R}$ a necessary condition for exactness is $M_y = N_x$, which is also sufficient if D is simply connected.
- As explained on the next two slides, the "differential-like" form (D) of a first-order ODE is essentially equivalent to the explicit form

$$y' = \frac{\mathrm{d}y}{\mathrm{d}x} = -\frac{M(x,y)}{N(x,y)}.$$

obtained from (D) by pretending that $\mathrm{d}x$, $\mathrm{d}y$ are numbers.

Solutions of (D)

By a *solution curve* ("*integral curve*", *parametrized solution*) of (D) we mean a smooth differentiable curve $\gamma\colon I \to D$, $t \mapsto \big(x(t), y(t)\big)$ satisfying

$$M\big(x(t), y(t)\big)x'(t) + N\big(x(t), y(t)\big)y'(t) = 0 \quad \text{for } t \in I. \qquad \text{(O)}$$

Geometrically, the tangent to the curve at any point must be orthogonal (perpendicular) to the vector of the vector field (M, N) at that point (since (D) is equivalent to $(M, N) \cdot \gamma' = 0$).

By an *(explicit) solution* $y = y(x)$ (resp., $x = x(y)$) we mean a function $y\colon I \to \mathbb{R}$ (resp., $x\colon J \to \mathbb{R}$) with graph contained in D and satisfying

$$M\big(x, y(x)\big) + N\big(x, y(x)\big)y'(x) = 0 \quad \text{for } x \in I, \quad \text{resp.,}$$
$$M\big(x(y), y\big)x'(y) + N\big(x(y), y\big)\big) = 0 \quad \text{for } y \in J.$$

Notes

- These concepts make sense for any (not necessarily exact) 1st-order ODE in differential-like form.

Notes cont'd

- A point $(x_0, y_0) \in D$ is said to be a *singular point* of the ODE $M(x,y)\,\mathrm{d}x + N(x,y)\,\mathrm{d}y = 0$ if $M(x_0,y_0) = N(x_0,y_0) = 0$. The orthogonality condition (O) is trivially satisfied in any singular point.
- Suppose (x_0, y_0) is a non-singular point of $M(x,y)\,\mathrm{d}x + N(x,y)\,\mathrm{d}y = 0$ and satisfies $N(x_0,y_0) = 0$.
  $\Longrightarrow$ Any solution curve $\gamma = (x,y)$ passing through (x_0,y_0) must have $x' = 0$ at (x_0,y_0).
  This says that γ has a vertical tangent at (x_0,y_0) and clearly forms an obstruction to representing it as a function $y(x)$.
  Conversely, if γ satisfies $\gamma(t_0) = (x_0,y_0)$ and $x'(t_0) = 0$ then $y'(t_0) \neq 0$ (since solution curves are smooth) and hence $N(x_0,y_0) = 0$.

The last note helps to clarify the correspondence between solution curves of $M(x,y)\,\mathrm{d}x + N(x,y)\,\mathrm{d}y = 0$ and solutions of $y' = -M(x,y)/N(x,y)$; cf. next slide.

Correspondence

Solution curves of $M\,\mathrm{d}x + N\,\mathrm{d}y = 0$ and explicit solutions correspond to each other in the following way:

1. Given a solution curve γ, smoothness implies that at each non-singular point $(x_0, y_0) \in \gamma(I)$ we can write the curve locally as graph of a C^1-function $y(x)$ or $x(y)$ (or both), and these functions satisfy

$$y' = \frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\mathrm{d}y\,/\,\mathrm{d}t}{\mathrm{d}x\,/\,\mathrm{d}t} = -\frac{M(x,y)}{N(x,y)}, \quad \text{resp.,}$$

$$x' = \frac{\mathrm{d}x}{\mathrm{d}y} = \frac{\mathrm{d}x\,/\,\mathrm{d}t}{\mathrm{d}y\,/\,\mathrm{d}t} = -\frac{N(x,y)}{M(x,y)},$$

   i.e., are explicit solutions. Note that, e.g., the representation $y(x)$ implies $x'(t) \neq 0$ and hence $N\big(x(t), y(t)\big) \neq 0$, as remarked in the previous note.

2. Conversely, given an explicit solution $y(x)$, we can use, e.g., $x(t) = t$ as parameter to define a curve $\gamma(t) = \big(t, y(t)\big)$, and this curve γ is a solution curve on account of

$$M(x,y)x' + N(x,y)y' = M(x,y) \cdot 1 + N(x,y)y' = 0.$$

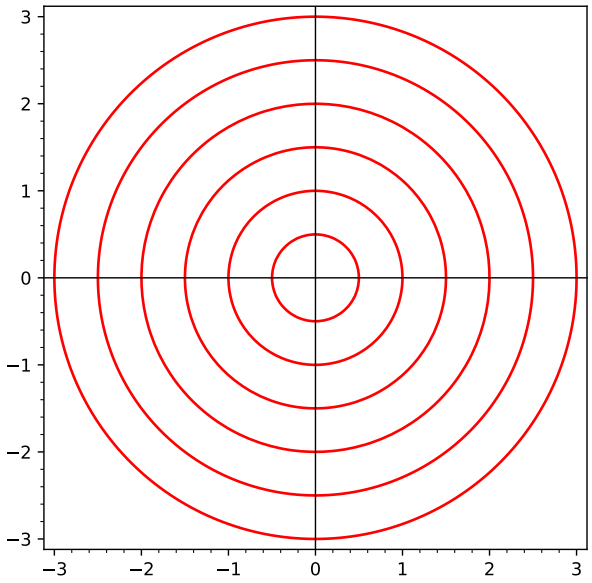
Thus, if we remove from D all points (x, y) with $N(x, y) = 0$, which form a closed set, and call the resulting domain D', we get a 1-1 correspondence between non-parametric solution curves of $M\,\mathrm{d}x + N\,\mathrm{d}y = 0$ (or classes of parametric solution curves under the equivalence relation of smooth reparametrization) and explicit solutions of $y' = -M(x, y)/N(x, y)$; and similarly for the case $M(x, y) = 0$.

Example

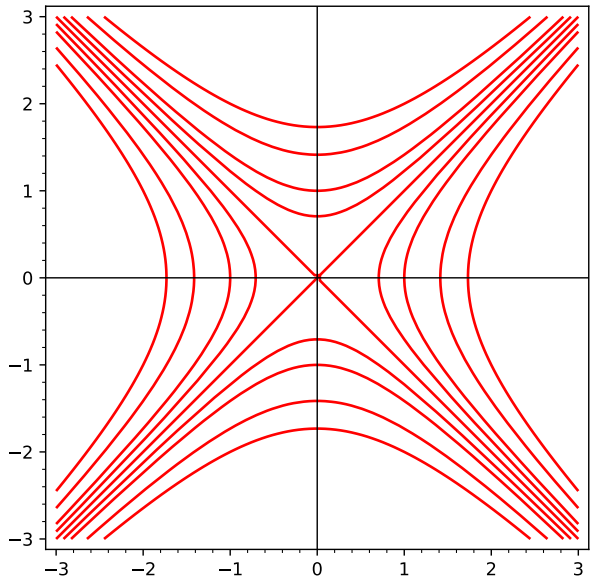
In the lecture and an exercise we have considered the four ODE's $y' = \pm x/y$, $y' = \pm y/x$. Associated differential-like forms are

1. $y' = -x/y \triangleq x\,\mathrm{d}x + y\,\mathrm{d}y = 0$;
2. $y' = x/y \triangleq x\,\mathrm{d}x - y\,\mathrm{d}y = 0$;
3. $y' = -y/x \triangleq x\,\mathrm{d}y + y\,\mathrm{d}x = 0$;
4. $y' = y/x \triangleq x\,\mathrm{d}y - y\,\mathrm{d}x = 0$.

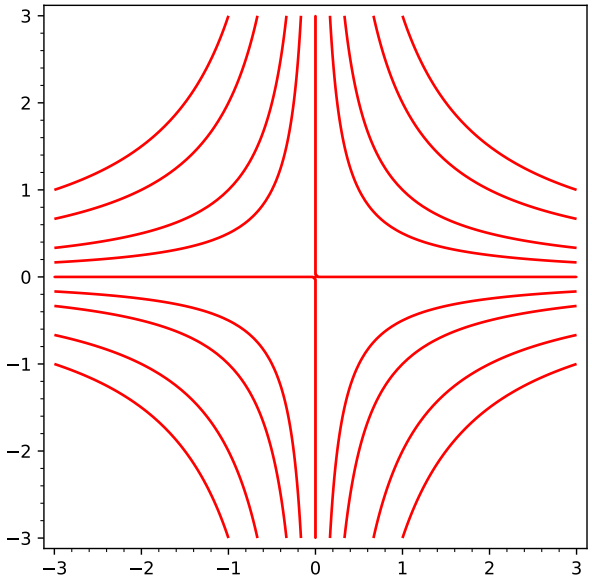
All four differential-like ODE's have exactly one singular point, viz. $(0, 0)$, and we need to remove either the x-axis (1st and 2nd ODE) or the y-axis (3rd and 4th ODE) in order to get a 1-1 correspondence of their solution curves with the solutions of the original explicit ODE. Solution curves are shown on the next slide.



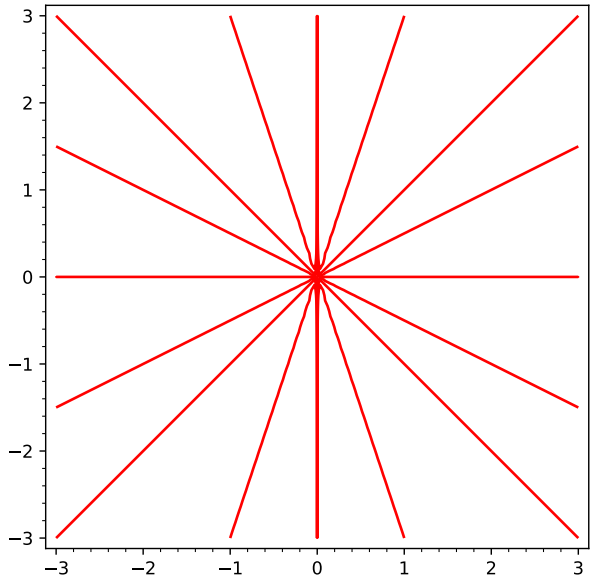
(1) $x\,\mathrm{d}x + y\,\mathrm{d}y = 0$



(2) $x\,\mathrm{d}x - y\,\mathrm{d}y = 0$



(3) $x\,\mathrm{d}y + y\,\mathrm{d}x = 0$



(4) $x\,\mathrm{d}y - y\,\mathrm{d}x = 0$

Theorem

Suppose $M(x,y)\,\mathrm{d}x + N(x,y)\,\mathrm{d}y = 0$ *is exact with antiderivative (potential function)* F*. Then the solution curves of* $M(x,y)\,\mathrm{d}x + N(x,y)\,\mathrm{d}y = 0$ *are precisely the parametrized level sets (contours)* $F(x,y) = C$*,* $C \in \mathbb{R}$*, or (sub-)branches thereof.*

Proof.

It suffices to show that any solution $\gamma(t) = \big(x(t), y(t)\big)$, $t \in I$, of $M(x,y)\,\mathrm{d}x + N(x,y)\,\mathrm{d}y = 0$ is contained in a level set of F.
We have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} F\big(\gamma(t)\big) &= \nabla F\big(\gamma(t)\big) \cdot \gamma'(t) \\ &= \begin{pmatrix} M\big(x(t), y(t)\big) \\ N\big(x(t), y(t)\big) \end{pmatrix} \cdot \begin{pmatrix} x'(t) \\ y'(t) \end{pmatrix} \\ &= M\big(x(t), y(t)\big)x'(t) + N\big(x(t), y(t)\big)y'(t) \\ &= 0,\end{aligned}$$

because $\gamma(t)$ is a solution of $M(x,y)\,\mathrm{d}x + N(x,y)\,\mathrm{d}y = 0$.
This shows that $t \mapsto F\big(\gamma(t)\big)$ is constant on I, i.e., $\big\{\gamma(t); t \in I\big\}$ is contained in a level set of F. □

Example

Consider the ODE

$$(x-y)\,\mathrm{d}x + \left(\frac{1}{y^2} - x\right)\mathrm{d}y = 0$$

The domain consists of all $(x,y) \in \mathbb{R}^2$ with $y \neq 0$ and has two simply-connected components (upper half plane and lower half plane).

Since $\frac{\mathrm{d}}{\mathrm{d}y}(x-y) = -1 = \frac{\mathrm{d}}{\mathrm{d}x}(y^{-2} - x)$, the ODE is exact.

An antiderivative, determined as usual by partial integration, is

$$f(x,y) = \frac{x^2}{2} - xy - \frac{1}{y}.$$

The general solution in implicit form is therefore

$$x^2y - 2y^2x - 2 - Cy = 0, \quad C \in \mathbb{R}.$$

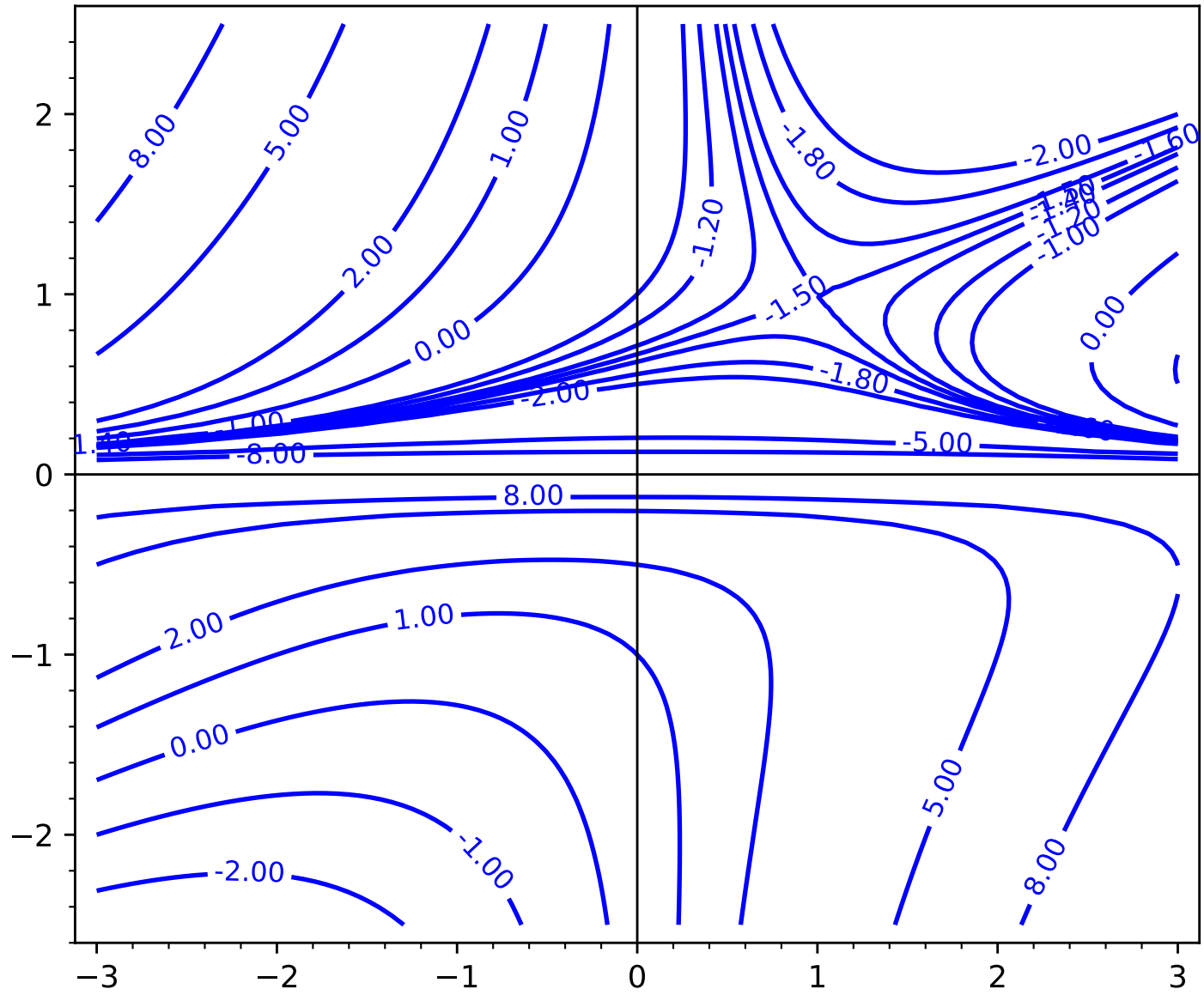


Figure: Solution curves of $(x-y)\,\mathrm{d}x+(y^{-2}-x)\,\mathrm{d}y=0$, represented as contours of $F(x,y)=x^2/2-xy-1/y$

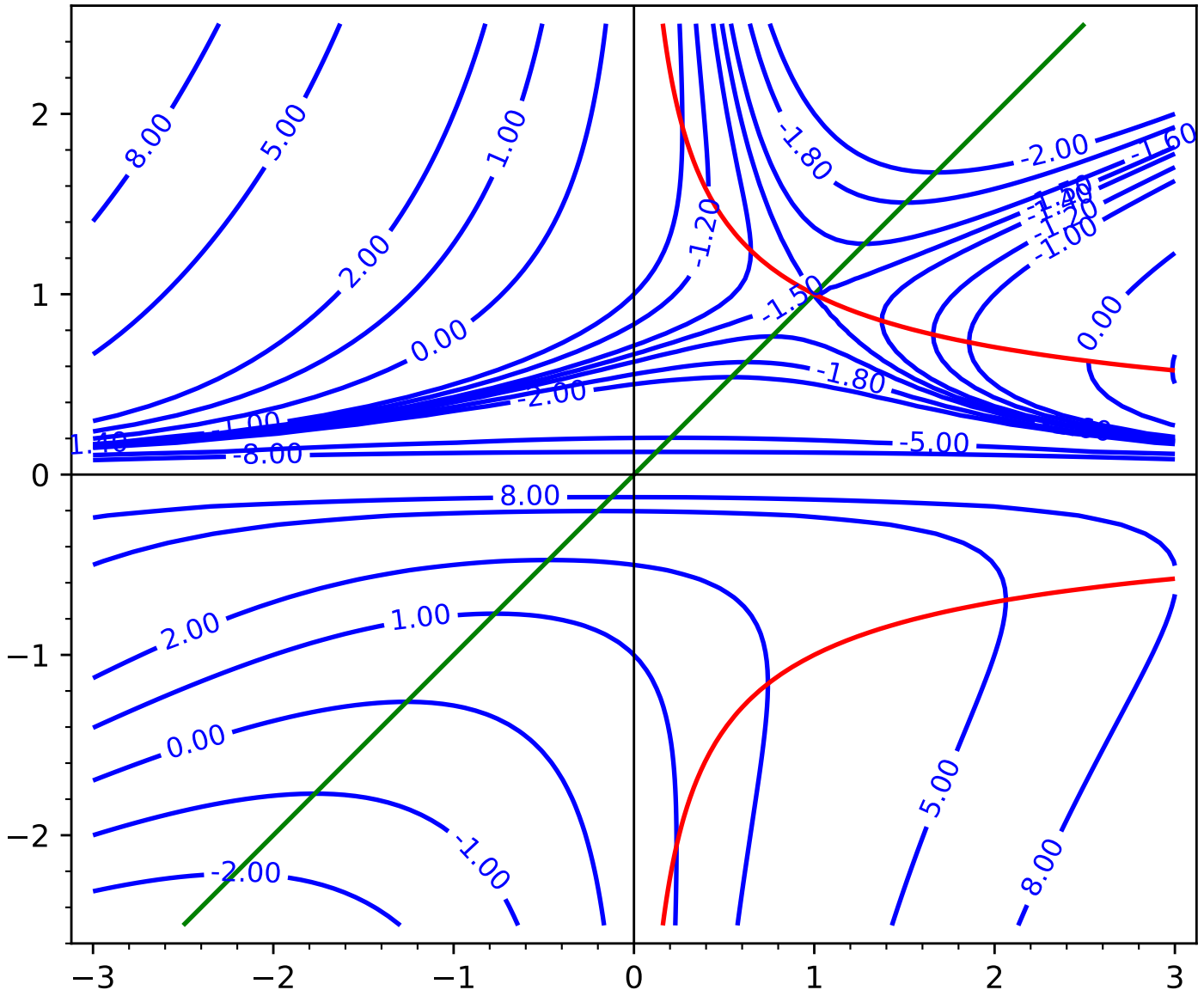


Figure: The same with all points highlighted that satisfy $M(x, y) = 0$ or $N(x, y) = 0$; removing the red (green) curve leaves solutions $y(x)$ of $y' = \frac{x-y}{x-y^{-2}}$ (resp., solutions $x(y)$ of $x' = \frac{x-y^{-2}}{x-y}$)

Example

Of the four ODE's $y' = \pm x/y$, $y' = \pm y/x$ three are exact, viz.

1. $x\,\mathrm{d}x + y\,\mathrm{d}y = \mathrm{d}F$ for $F(x,y) = \frac{1}{2}(x^2+y^2)$;
2. $x\,\mathrm{d}x - y\,\mathrm{d}y = \mathrm{d}F$ for $F(x,y) = \frac{1}{2}(x^2-y^2)$;
3. $x\,\mathrm{d}y + y\,\mathrm{d}x = \mathrm{d}F$ for $F(x,y) = xy$.

This shows that the corresponding solution curves are

1. circles centered at the origin (contours of $(x,y) \mapsto x^2+y^2$);
2. hyperbolas centered at the origin with asymptotes $y = \pm x$ (contours of $(x,y) \mapsto x^2 - y^2$);
3. hyperbolas centered at the origin with asymptotes $x = 0$ and $y = 0$ (contours of $(x,y) \mapsto xy$).

The 4th ODE $x\,\mathrm{d}y - y\,\mathrm{d}x = 0$ (corresponding to the winding form/field) is not exact.

But it can be multiplied by $1/(xy)$ to yield the exact (even separable) ODE $y^{-1}\,\mathrm{d}y - x^{-1}\,\mathrm{d}x = 0$ ($\rightarrow$ integrating factors), which has solution curves $\ln|y| - \ln|x| = C$ or, equivalently, $y/x = \pm \mathrm{e}^C$; compare with the previous plot of these curves.

Integrating Factors

The ODE $y\,\mathrm{d}x + (x^2y - x)\,\mathrm{d}y = 0$ is not exact, since $\frac{\partial M}{\partial y} = 1$ and $\frac{\partial N}{\partial x} = 2xy - 1$.

But we can multiply the equation by $1/x^2$, turning it into the exact ODE

$$\frac{y}{x^2}\,\mathrm{d}x + \left(y - \frac{1}{x}\right)\mathrm{d}y = 0$$

with potential $f(x,y) = -y/x + y^2/2$ and general solution $xy^2 - 2y - Cx = 0$.

Since the exact ODE has a strictly smaller domain, viz. $\mathbb{R}^2$ without the y-axis, we also need to check whether the parametrized y-axis $\gamma(t) = (0,t)$ is a solution of $y\,\mathrm{d}x + (x^2y - x)\,\mathrm{d}y = 0$, and indeed it is ($x(t) = x'(t) = 0$). But it is missing in the implicit solution.

Definition

A function $\mu(x,y)$ with domain $D' \subseteq D$ is called an *integrating factor* (or *Euler multiplier*) of $M(x,y)\,\mathrm{d}x + N(x,y)\,\mathrm{d}y = 0$, if

1. $\mu(x,y) \neq 0$ for all $(x,y) \in D'$;
2. $\mu(x,y)M(x,y)\,\mathrm{d}x + \mu(x,y)N(x,y)\,\mathrm{d}y = 0$ is exact on D'.

Lemma

If an ODE $M\,\mathrm{d}x + N\,\mathrm{d}y = 0$ has a general solution of the form $f(x,y) = C$ then it has an integrating factor.

Proof.

Differentiating $f(x,y) = C$ gives $\frac{\partial f}{\partial x}\,\mathrm{d}x + \frac{\partial f}{\partial y}\,\mathrm{d}y = 0$.

$$\Longrightarrow \frac{\mathrm{d}y}{dx} = -\frac{M}{N} = -\frac{\partial f/\partial x}{\partial f/\partial y},$$

which can be rewritten as

$$\frac{\partial f/\partial x}{M} = \frac{\partial f/\partial y}{N} = \mu(x,y), \quad \text{say.}$$

This says that $\mu M\,\mathrm{d}x + \mu N\,\mathrm{d}y = \frac{\partial f}{\partial x}\,\mathrm{d}x + \frac{\partial f}{\partial y}\,\mathrm{d}y = 0$ is exact. □

Remark

We can multiply an integrating factor μ by any continuous function $F(f)$ of the antiderivative f of the resulting exact equation, thereby obtaining another integrating factor $\mu F(f)$. (Check that a suitable antiderivative is $G(f)$, where $G' = F$.) Hence integrating factors are highly non-unique.

How to Find an Integrating Factor?

The (local) exactness condition for an integrating factor μ is $\partial(\mu M)/\partial y = \partial(\mu N)/\partial x$. This gives

$$\mu \frac{\partial M}{\partial y} + M \frac{\partial \mu}{\partial y} = \mu \frac{\partial N}{\partial x} + N \frac{\partial \mu}{\partial x}, \qquad \text{or}$$

$$\mu \left(\frac{\partial M}{\partial y} - \frac{\partial N}{\partial x} \right) = N \frac{\partial \mu}{\partial x} - M \frac{\partial \mu}{\partial y}.$$

This partial differential equation (PDE) for μ is not easy to solve in general, but frequently one can make a particular „Ansatz“ for μ and solve it in this special case.

Example

Suppose $\frac{1}{N}\left(\frac{\partial M}{\partial y} - \frac{\partial N}{\partial x}\right) = g(x)$ depends only on x but not on y.

Then $M\,\mathrm{d}x + N\,\mathrm{d}y = 0$ has the integrating factor $\mu(x) = \mathrm{e}^{\int g(x)\,\mathrm{d}x}$.

Reason: In this case the PDE for $\mu(x, y) = \mu(x)$ is equivalent to $\mu'(x) = g(x)\mu(x)$.

Example (cont'd)

As a concrete example we reconsider $y\,\mathrm{d}x + (x^2y - x)\,\mathrm{d}y = 0$.

Here we have

$$\frac{1}{N}\left(\frac{\partial M}{\partial y} - \frac{\partial N}{\partial x}\right) = \frac{M_y - N_x}{N} = \frac{1 - (2xy - 1)}{x^2y - x} = \frac{2(1 - xy)}{x(xy - 1)} = -\frac{2}{x}.$$

An integrating factor is therefore

$$\mu(x) = \mathrm{e}^{\int (-2/x)\,\mathrm{d}x} = \mathrm{e}^{-2\ln x} = \frac{1}{x^2},$$

as we have seen before.

Remark

In particular we can solve the PDE for μ if all of μ, $M_y - N_x$, N depend only on x. But we only need the weaker condition "$(M_y - N_x)/N$ depends only on x". In the example above both $M_y - N_x$ and N depend on both x and y but $(M_y - N_x)/N$ depends only on x.

Theorem

The ODE $M\,\mathrm{d}x + N\,\mathrm{d}y = 0$ *has an integrating factor of the form*

1. $\mu(x)$ *if* $\dfrac{M_y - N_x}{N} = g(x)$;
2. $\mu(y)$ *if* $\dfrac{M_y - N_x}{M} = g(y)$;
3. $\mu(xy)$ *if* $\dfrac{M_y - N_x}{Ny - Mx} = g(xy)$;
4. $\mu(y/x)$ *if* $\dfrac{x^2(M_y - N_x)}{Ny + Mx} = g(y/x)$.

Proof.

In each case the PDE $(M_y - N_x)\mu = N\mu_x - M\mu_y$ derived for $\mu(x, y)$ becomes a homogeneous linear 1st-order ODE for the one-variable function $\mu(s)$ (note the slight abuse of notation in the last two cases!), which can be solved using the standard method. The resulting ODE for $\mu(s)$ is $\mu'(s) = g(s)\mu(s)$ in Cases (1) and (3), and $\mu'(s) = -g(s)\mu(s)$ in Cases (2) and (4).

Proof cont'd.

We do this explicitly for the last case:

$$\frac{\partial}{\partial x}\,\mu\left(\frac{y}{x}\right) = \mu'\left(\frac{y}{x}\right)\left(-\frac{y}{x^2}\right),$$
$$\frac{\partial}{\partial y}\,\mu\left(\frac{y}{x}\right) = \mu'\left(\frac{y}{x}\right)\frac{1}{x}.$$

Hence $(M_y - N_x)\mu = N\mu_x - M\mu_y$ becomes

$$(M_y - N_x)\,\mu\left(\frac{y}{x}\right) = N\,\mu'\left(\frac{y}{x}\right)\left(-\frac{y}{x^2}\right) - M\,\mu'\left(\frac{y}{x}\right)\frac{1}{x}$$
$$\iff \quad x^2(M_y - N_x)\,\mu\left(\frac{y}{x}\right) = -(Ny + Mx)\,\mu'\left(\frac{y}{x}\right)$$
$$\iff \quad \mu'\left(\frac{y}{x}\right) = -\frac{x^2(M_y - N_x)}{Ny + Mx}\,\mu\left(\frac{y}{x}\right).$$

If $\frac{x^2(M_y - N_x)}{Ny + Mx} = g(y/x)$ depends only on y/x, we can substitute $s = y/x$ and obtain the equivalent ODE $\mu'(s) = -g(s)\mu(s)$. □

Final Remarks

- In some texts the case of an integrating factor of the form $\mu(x/y)$ is listed as well. But this reduces to Case (4) if we consider $\tilde{\mu}(s) = \mu(1/s)$.
- The PDE $(M_y - N_x)\mu = N\mu_x - M\mu_y$ only guarantees local exactness of $(\mu M)\,\mathrm{d}x + (\mu N)\,\mathrm{d}y$ on D'. To obtain an anti-derivative, it may be necessary to restrict the domain further to simply-connected subsets of D', on which $(\mu M)\,\mathrm{d}x + (\mu N)\,\mathrm{d}y$ then must be exact, and determine solutions there.

  For example, $x\,\mathrm{d}y - y\,\mathrm{d}x = 0$ has the integrating factor $1/(xy)$, as we have seen, whose domain $\mathbb{R}^2$ with the coordinate axes removed consists of 4 simply connected regions (the 4 open quadrants). On each quadrant, an antiderivative of $(xy)^{-1}(x\,\mathrm{d}y - y\,\mathrm{d}x) = y^{-1}\,\mathrm{d}y - x^{-1}\,\mathrm{d}x$ exists and can be taken as $f(x,y) = \ln|y| - \ln|x|$, amounting to 4 different choices of signs of x, y for the 4 regions.

Orthogonal Trajectories

Problem

Given a family of smooth (non-parametric) plane curves, does there exist another such family with the following property: All angles of intersection between members of the first family and members of the second family are right angles (90°).

In this situation we say that the members of the second family are the *orthogonal trajectories* of the first family (and vice versa).

Observation

If a family of plane curves arises as solution of an explicit first-order ODE $y' = f(x,y)$ ($M(x,y)\,\mathrm{d}x + N(x,y)\,\mathrm{d}y = 0$), its orthogonal trajectories can be obtained by solving

$$y' = -\frac{1}{f(x,y)}, \qquad \text{resp.,} \qquad -N(x,y)\,\mathrm{d}x + M(x,y)\,\mathrm{d}y = 0.$$

Reason: $y' = f(x,y)$ prescribes the slope $m = f(x_0,y_0)$ for (the tangent of) a solution passing through (x_0,y_0). It is a general fact that the lines through (x_0,y_0) with slopes m and $-1/m$ are orthogonal.

Reason (cont'd): In the second case solution curves are orthogonal to $\begin{pmatrix} M(x_0,y_0) \\ N(x_0,y_0) \end{pmatrix}$, which is the direction in (x_0, y_0) prescribed for a solution of $-N(x,y)\,\mathrm{d}x + M(x,y)\,\mathrm{d}y = 0$.

Example

Determine the orthogonal trajectories of the family of parabolas $y = C\,x^2$, $C \in \mathbb{R}$.

Solution: First we determine a differential equation for $y = C\,x^2$. Since $y = C\,x^2 \iff y\,x^{-2} = C$, the parabolas are the contours of $F(x,y) = y\,x^{-2}$.

$$\implies -2y\,x^{-3}\,\mathrm{d}x + x^{-2}\,\mathrm{d}y = 0.$$

The orthogonal trajectories must then solve $x^{-2}\,\mathrm{d}x + 2y\,x^{-3}\,\mathrm{d}y = 0$, which simplifies to $x\,\mathrm{d}x + 2y\,\mathrm{d}y = 0$. This ODE is exact with solution $x^2/2 + y^2 = C$, or $x^2 + 2\,y^2 = C'$ (ellipses with center $(0,0)$ and semiaxes $a = \sqrt{C'}$, $b = \sqrt{C'/2}$; C' must be positive).

Math 285
Introduction to Differential Equations

Thomas Honold

ZJU-UIUC Institute

Spring Semester 2024

Outline

Today's Lecture: Uniform Convergence

The concept of uniform convergence arises from the question whether the limit function of a sequence or series of functions inherits properties like continuity or differentiability from the terms of the sequence/series. For ordinary (point-wise) convergence the answer is notably false.

Uniform convergence was not discussed in Calculus I/II/III, but is needed to understand the existence theorem for solutions of ODE's, the theory of Fourier series, and many other important topics in Real Analysis.

As background reference for the material on uniform convergence I recommend the respective chapters in

[Bre07] David Bressoud, *A Radical Approach to Real Analysis*, 2nd edition, Mathematical Association of America 2007;

[Ru76] Walter Rudin, *Principles of Mathematical Analysis*, 3rd edition, McGraw-Hill 1976.

[Bre07] is very accessible and retraces the historical development of Calculus in the 19th century and the mathematicians involved in it. [Ru76] is a good reference also for other important concepts not covered by Stewart's book (e.g., the Implicit Function Theorem), but be warned that it is pretty advanced.

The following 3 slides show plots of 3 sequences of functions to be defined and discussed later. These function sequences converge point-wise but not uniformly, and serve as counterexamples to the naive belief, prevalent until the beginning of the 19th century, that point-wise limits of sequences of continuous/differentiable/integrable functions inherit the respective property.

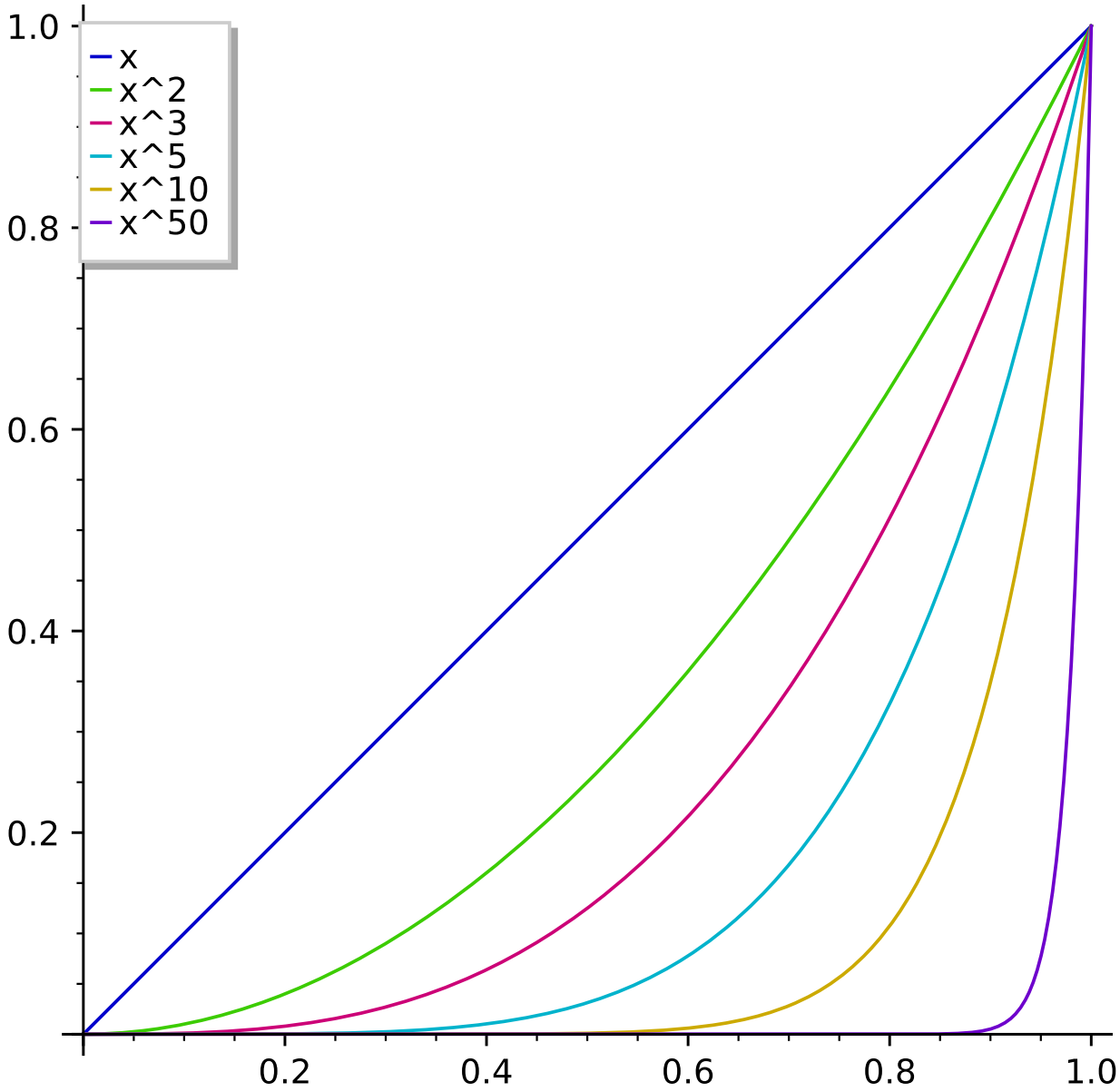


Figure: $f_n(x) = x^n$ for $n = 1, 2, 3, 5, 10, 50$

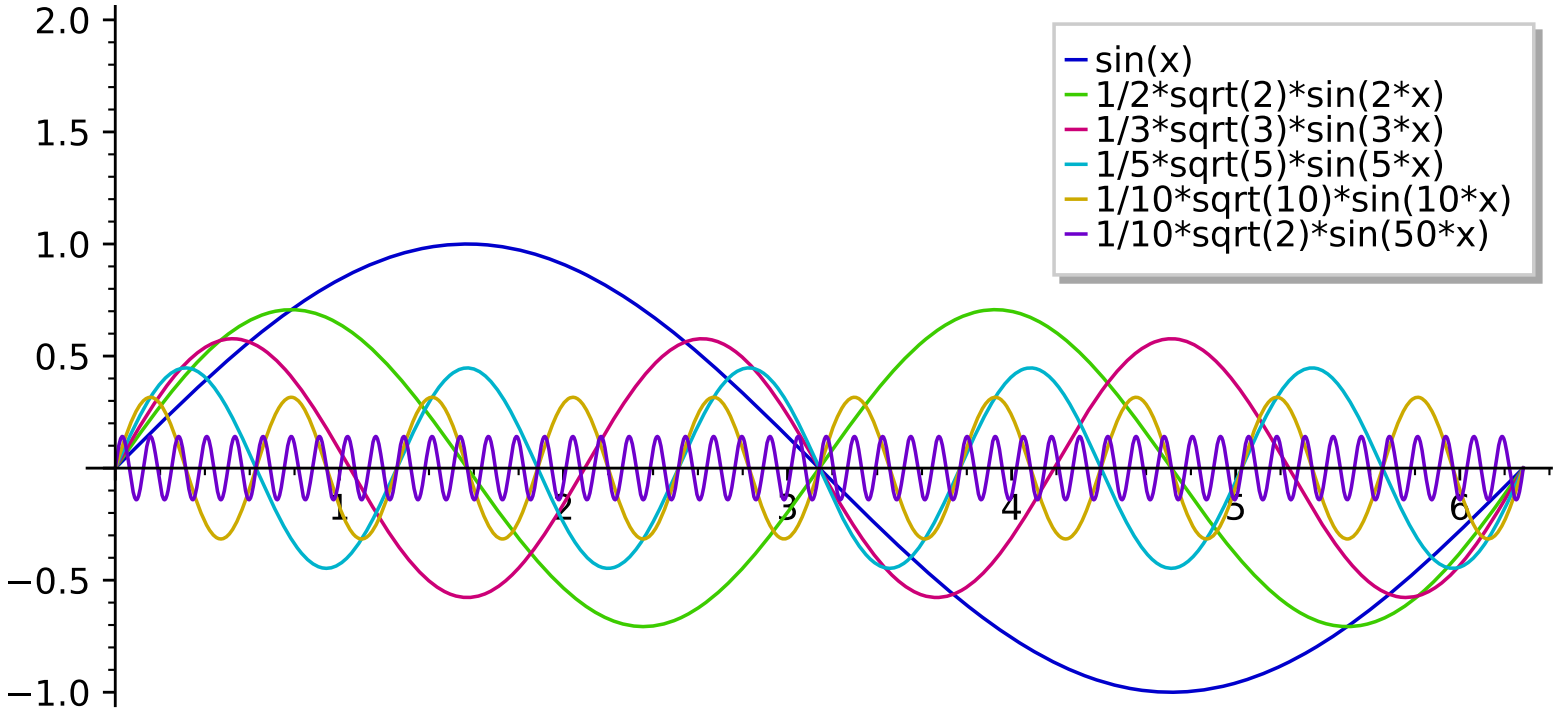


Figure: $g_n(x) = \frac{\sin(nx)}{\sqrt{n}}$ for $n = 1, 2, 3, 5, 10, 50$

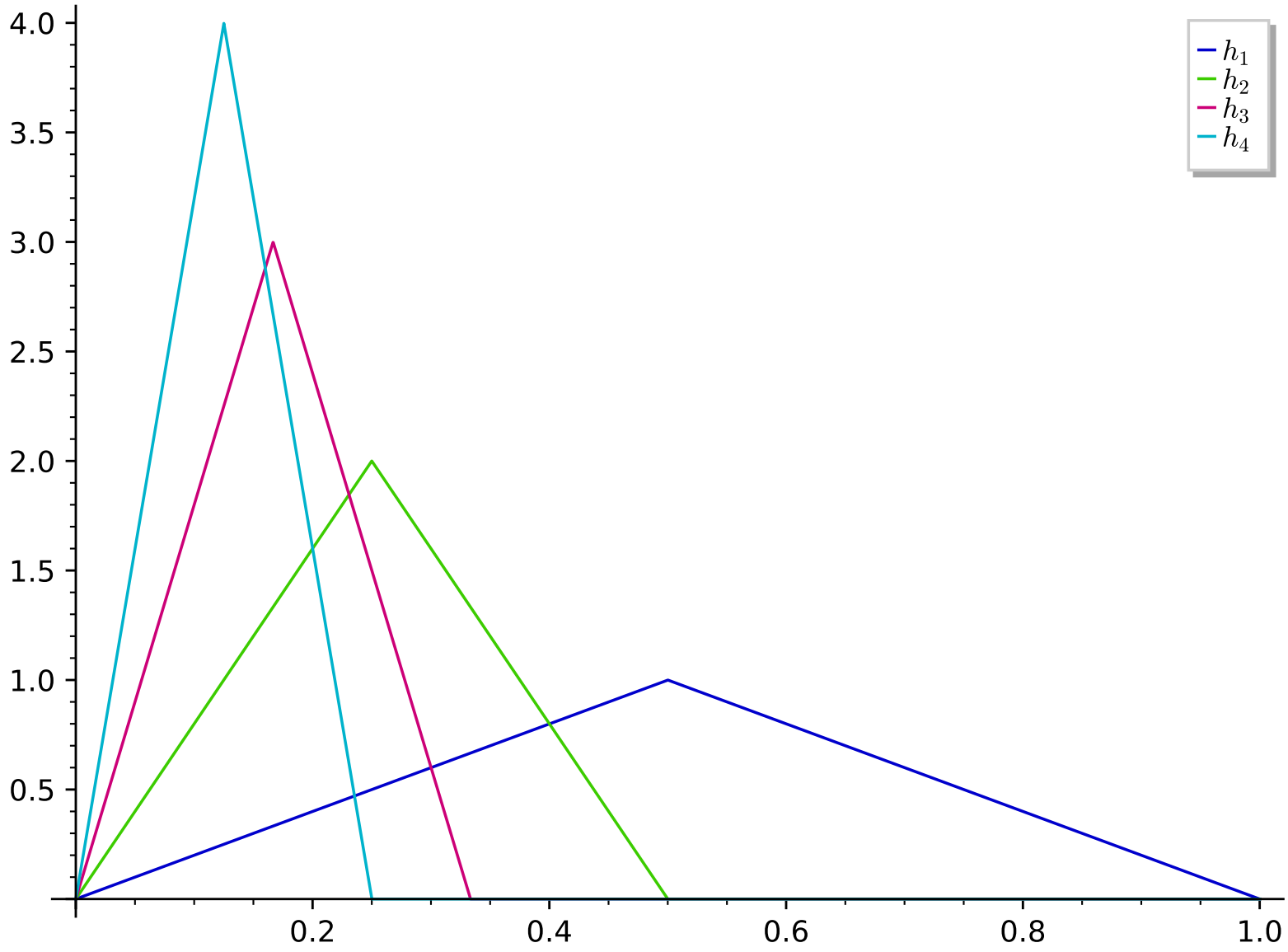


Figure: $h_n(x)$ for $n = 1, 2, 3, 4$

Point-wise vs Uniform Convergence

Definition

Let $I \subseteq \mathbb{R}$ be an interval and $(f_n)_{n=0}^{\infty}$ a sequence of functions $f_n \colon I \to \mathbb{R}$.

1. (f_n) *converges point-wise* (on I) if for every $x \in I$ the sequence $\big(f_n(x)\big)$, an ordinary sequence of real numbers, converges. If this is the case then $f(x) = \lim_{n\to\infty} f_n(x)$ defines a function $f \colon I \to \mathbb{R}$, called "limit function" or "point-wise limit" of the sequence (f_n).
2. (f_n) *converges uniformly* (on I) if it converges point-wise and the limit function $f \colon I \to \mathbb{R}$ has the following property: For every $\epsilon > 0$ there is a "uniform" response $N \in \mathbb{N}$ such that $|f(x) - f_n(x)| < \epsilon$ for all $n > N$ and all $x \in I$.

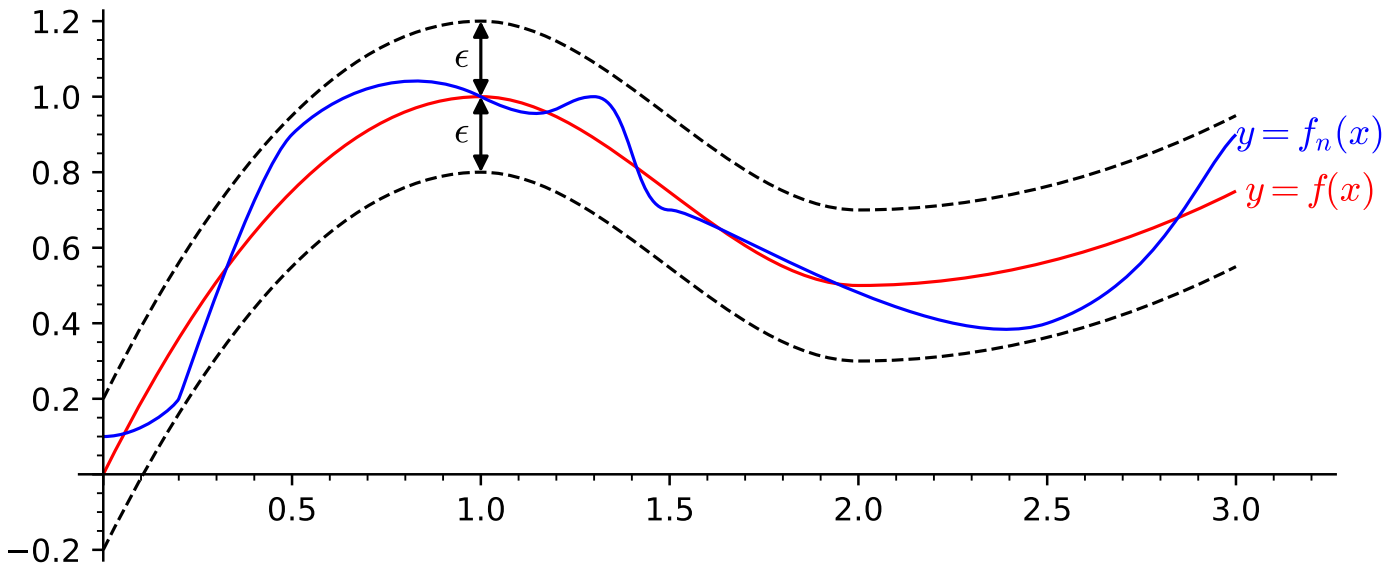
If (1), resp., (2) hold, we also say that (f_n) converges to f point-wise, resp., uniformly.

Notes

- Uniform convergence requires the response $N = N_\epsilon$ to be independent of $x \in I$, while point-wise convergence allows $N = N_{\epsilon,x}$ to depend on x (and ϵ).

Notes cont'd

- Geometrically speaking, (f_n) converges uniformly to f iff for every $\epsilon > 0$ all except finitely many of the graphs of f_n are contained in the strip of vertical width 2ϵ around the graph of f; see picture.



Looking at the preceding plots, you can see that the sequences (f_n) and (h_n) fail to have this property for any $\epsilon < 1$, while (g_n) seems to have it. (At least we can see that the graph of g_{50} is within 0.2 of the graph of the point-wise limit, viz., the x-axis.)

Notes cont'd

- The definition generalizes to functions $f_n\colon X \to \mathbb{R}$ with arbitrary domain X. Further we can replace the codomain $\mathbb{R}$ by $\mathbb{R}^k$, because the concept of convergence for the corresponding vectorial sequences $f_0(x), f_1(x), f_2(x), \ldots$ is well-defined (cf. Calculus III) and $|f(x) - f_n(x)| < \epsilon$ can be read as an inequality for the Euclidean length of the vector $f(x) - f_n(x) \in \mathbb{R}^k$. Even more generally, we can take the codomain of f_n as any set M with a distance function $d\colon M \times M \to \mathbb{R}$ (replacing, e.g., "$|f(x) - f_n(x)| < \epsilon$" by "$d\big(f(x), f_n(x)\big) < \epsilon$"), i.e., by a (generalized) metric space (M, d).
- In the definition of convergence it does not matter whether $<$ or $\leq$ is used. Using the latter has the advantage that the condition "$|f(x) - f_n(x)| \leq \epsilon$ for all $x \in I$" can be succintly stated as $\sup\big\{|f(x) - f_n(x)|\,; x \in I\big\} \leq \epsilon$. We can view the left-hand side of this inequality as a measure for the distance between the functions f and f_n. More precisely, if we define $\mathrm{d}_\infty(f, g) = \sup\big\{|f(x) - g(x)|\,; x \in I\big\}$ for $f, g \in \mathbb{R}^I$ (referred to as *metric of uniform convergence* or L^∞*-metric*) then uniform convergence amounts to ordinary convergence in the generalized metric space $(\mathbb{R}^I, \mathrm{d}_\infty)$.

Questions

1. *Is the limit function $f(x) = \lim_{n\to\infty} f_n(x)$ of a point-wise/uniformly convergent sequence of continuous functions itself continuous?*

   A suggestive reformulation of this property is obtained by recalling that a function g is continuous at x iff $x_k \to x$ implies $g(x_k) \to g(x)$. Applying this to f and f_n above gives that f is continuous iff

$$\lim_{k\to\infty} \lim_{n\to\infty} f_n(x_k) = \lim_{k\to\infty} f(x_k) = f(x) = \lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} \lim_{k\to\infty} f_n(x_k).$$

2. *How about the related problem of interchanging limits with differentiation? Under which conditions does $f'(x) = (\lim f_n)'(x) = \lim f_n'(x)$ hold?*
3. *How about integration in this regard?*

Three Counterexamples

The following examples show that point-wise convergence is not sufficient for any of the three properties.

Example (continuity)

Consider the sequence of functions $f_n(x) = x^n$, $x \in [0,1]$. The functions f_n are continuous and converge point-wise to

$$f(x) = \lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} x^n = \begin{cases} 0 & \text{if } 0 \le x < 1, \\ 1 & \text{if } x = 1, \end{cases}$$

but f has a discontinuity at $x = 1$.

Example (differentiation)

Consider the sequence of functions $g_n(x) = \frac{\sin(nx)}{\sqrt{n}}$, $x \in [0, 2\pi]$. We have $g_n(x) \to g(x) \equiv 0$ (the all-zero function on $[0, 2\pi]$) point-wise (even uniformly!), and g is differentiable with $g'(x) \equiv 0$. But

$$g_n'(x) = \sqrt{n}\cos(nx), \quad x \in [0, 2\pi],$$

and $\lim_{n\to\infty} g_n'(x)$ doesn't exist for $0 < x < 2\pi$.

Example (integration)

Consider the sequence of functions $h_n\colon [0,1] \to \mathbb{R}$ defined by

$$h_n(x) = \begin{cases} 2n^2x & \text{if } 0 \le x \le 1/2n, \\ 2n - 2n^2x & \text{if } 1/2n \le x \le 1/n, \\ 0 & \text{if } 1/n \le x \le 1. \end{cases}$$

The graph of h_n and the x-axis determine an (isosceles) triangle with vertices $(0,0)$, $(1/2n, n)$, $(1/n, 0)$, and h_n vanishes on $[1/n, 1]$.

It follows that $h_n(x) \to h(x) \equiv 0$ (the all-zero function on $[0,1]$) point-wise, and that the area under the graph of h_n is $1/2 \cdot 1/n \cdot n = 1/2$ for all n. This gives

$$\lim_{n\to\infty} \int_0^1 h_n(x)\,\mathrm{d}x = \frac{1}{2} \neq 0 = \int_0^1 \lim_{n\to\infty} h_n(x)\,\mathrm{d}x\,.$$

Three Theorems

The purpose of introducing the concept of uniform convergence is to prevent such "counterexamples". The answer to all three questions will be positive, provided we require the sequence of functions (f_n) and/or the sequence of its derivatives (f_n') to be uniformly convergent.

Theorem (continuity)

If all functions f_n are continuous at $x_0 \in I$ and (f_n) converges uniformly on I then $f(x) = \lim f_n(x)$, $x \in I$, is continuous at x_0 as well. In particular, the limit function of a uniformly convergent sequence of continuous functions is itself continuous.

Proof.

Let $\epsilon > 0$ be given. Then there exists $N \in \mathbb{N}$ such that $|f(x) - f_n(x)| < \epsilon/3$ for $n > N$ and $x \in I$. Further, since f_{N+1} is continuous at x_0, there exists $\delta > 0$ such that $|f_{N+1}(x) - f_{N+1}(x_0)| < \epsilon/3$ for $x \in I$ with $|x - x_0| < \delta$. For such x we then have

$$|f(x) - f(x_0)| \le |f(x) - f_{N+1}(x)| + |f_{N+1}(x) - f_{N+1}(x_0)| + |f_{N+1}(x_0) - f(x_0)|$$
$$< \tfrac{\epsilon}{3} + \tfrac{\epsilon}{3} + \tfrac{\epsilon}{3} = \epsilon.$$

□

Theorem (differentiation)

If all functions f_n are C^1-functions, (f_n) converges point-wise on I, and (f_n') converges uniformly on I, then $f(x) = \lim_{n\to\infty} f_n(x)$, $x \in I$, is a C^1-function as well and satisfies $f'(x) = \lim_{n\to\infty} f_n'(x)$.

Proof.

Choosing an arbitary point $a \in I$, the Fundamental Theorem of Calculus gives

$$f_n(x) = f_n(a) + \int_a^x f_n'(t)\,\mathrm{d}t \quad \text{for } x \in I.$$

Since (f_n') converges uniformly to $g\colon I \to \mathbb{R}$, say, we can find an $N \in \mathbb{N}$ such that $|f_n'(t) - g(t)| < 1$ for all $n > N$ and $t \in I$.
By the preceding theorem, g is continuous, and the inequality implies

$$|f_n'(t)| \le 1 + |g(t)| \quad \text{for } n > N \text{ and } t \in I.$$

Thus $\Phi(t) = 1 + |g(t)|$ is an integrable bound for $(f_n')_{n>N}$ on $[a, x]$, and we can apply Lebesgue's Bounded Convergence Theorem to conclude that $\lim_{n\to\infty} \int_a^x f_n'(t)\,\mathrm{d}t = \int_a^x g(t)\,\mathrm{d}t$. □

Proof cont'd.

Hence, letting $n \to \infty$ in the first identity we obtain

$$f(x) = f(a) + \int_a^x g(t)\,\mathrm{d}t \quad \text{for } x \in I.$$

Finally, applying the Fundamental Theorem of Calculus a second time gives that f is differentiable with $f'(x) = g(x) = \lim f_n'(x)$. □

Notes

1. The proof also shows that

$$f(x) - f_n(x) = f(a) - f_n(a) + \int_a^x [g(t) - f_n'(t)]\,\mathrm{d}t\,.$$

Since $f_n' \to g$ uniformly, given $\epsilon > 0$, we can find a response N such that

$$|f(x) - f_n(x)| \le \epsilon\,(1 + |x - a|) \quad \text{for all } n > N \text{ and } x \in I.$$

This shows that (f_n) converges not only point-wise to f but uniformly on every bounded subinterval of I. (If I is unbounded, however, we don't get uniform convergence of (f_n) on I, as the example $I = \mathbb{R}$, $f_n(x) = x/n$ shows.)

Notes cont'd

2. The key assumption in the Differentiation Theorem is that the *sequence of derivatives* (f_n') converges uniformly (and not, as one might think in the first place, the sequence (f_n)). For (f_n) the weaker assumption of point-wise convergence is enough. (In fact it would even be sufficient to require only that $\big(f_n(a)\big)$ converges.) But at least some assumption on (f_n) is clearly necessary, because we can add arbitrary constants to f_n without affecting f_n'.

3. One can use a variant of this theorem to prove that analytic functions of a complex variable, i.e., functions $f\colon D \to \mathbb{C}$ defined on some open disk $D = \mathrm{B}_R(a) \subseteq \mathbb{C}$ ($a \in \mathbb{C}$, $R > 0$) by a convergent power series $f(z) = \sum_{n=0}^{\infty} a_n(z-a)^n$, are holomorphic. For this the following two key observations are needed: (1) Power series converge uniformly on any closed disk $\overline{\mathrm{B}_{R'}(a)}$, $R' < R$, where R denotes the radius of convergence. (2) The series $\sum_{n=1}^{\infty} na_n(z-a)^{n-1}$ of derivatives has the same radius of convergence as the original series and hence converges uniformly on $\overline{\mathrm{B}_{R'}(a)}$ as well.

Theorem (integration)

If I is a bounded interval, all functions f_n are (Lebesgue) integrable over I and (f_n) converges uniformly on I then the limit function $f(x) = \lim_{n\to\infty} f_n(x)$ is integrable as well, and we have $\int_I f(x)\,\mathrm{d}x = \lim_{n\to\infty} \int_I f_n(x)\,\mathrm{d}x$.

Proof.

This follows by using Lebesgue's Theorem in a similar way as in the preceding proof:

There exists $N \in \mathbb{N}$ such that $|f(x) - f_n(x)| < 1$ for all $n > N$ and $x \in I$.
$\Longrightarrow |f_n(x) - f_{N+1}(x)| < 2$ for $n > N$ and $x \in I$.
$\Longrightarrow$ An integrable bound for $(f_n)_{n>N}$ is $\Phi(x) := |f_{N+1}(x)| + 2$. □

Note on the proof

If you wonder where the assumption "I is bounded" is needed in the proof: It is hidden in the definition of $\Phi(x)$: The function f_{N+1} (and hence $|f_{N+1}|$) is integrable by assumption, but the constant function 2 is integrable only if I is bounded. Therefore, Φ is integrable only if I is bounded.

Notes on the Integration Theorem

1. For a sequence of continuous functions on a compact intervall $I = [a, b]$ (or any other sequence of functions for which it is known in advance that the limit function is integrable) we can alternatively argue as follows:

$$\left|\int_a^b f(x)\,\mathrm{d}x - \int_a^b f_n(x)\,\mathrm{d}x\right| = \left|\int_a^b f(x) - f_n(x)\,\mathrm{d}x\right|$$
$$\leq \int_a^b |f(x) - f_n(x)|\,\mathrm{d}x$$
$$\leq (b-a)\sup\left\{|f(x) - f_n(x)|\,;\, a \leq x \leq b\right\}$$

   Hence, if $f_n \to f$ uniformly on $[a, b]$ then $\int_a^b f_n(x)\,\mathrm{d}x \to \int_a^b f(x)\,\mathrm{d}x$.

2. The assumption that I is bounded is essential. Without this assumption, the conclusion generally fails to hold. For example, define $f_n\colon \mathbb{R} \to \mathbb{R}$ by $f_n(x) = 1/n$ if $0 \leq x \leq n$ and $f_n(x) = 0$ otherwise. Then $f_n \to 0$ uniformly, but $\int_{\mathbb{R}} f_n(x)\,\mathrm{d}x = 1 \not\to 0 = \int_{\mathbb{R}} 0\,\mathrm{d}x$.

Notes cont'd

3. Using the integration theorem, we can give a simpler proof of the differentiation theorem, which avoids reference to the rather deep theory of Lebesgue integration:

   In the previous proof the key step is the implication

$$f_n(x) = f_n(a) + \int_a^x f_n'(t)\,\mathrm{d}t \quad (n \in \mathbb{N},\ x \in I)$$

$$\Longrightarrow\ f(x) = f(a) + \int_a^x g(t)\,\mathrm{d}t \quad (x \in I),$$

   where g denotes the uniform limit of the sequence of derivatives (f_n').

   Since $f_n \to f$ point-wise, we have $f_n(x) \to f(x)$ and $f_n(a) \to f(a)$. Since $I = [a, x]$ is bounded, f_n' is integrable over I (since it is continuous), and $f_n' \to g$ uniformly, we can apply the integration theorem to conclude $\int_a^x f_n'(t)\,\mathrm{d}t \to \int_a^x g(t)\,\mathrm{d}t$. This provides an alternative proof of the key step.

   In fact the special case of the integration theorem considered in Note 1, valid also for the Riemann integral, is sufficient.

Weierstrass's Criterion

A handy test for the uniform convergence of function series

Theorem (Weierstrass's Criterion)

Suppose $f_n\colon D \to \mathbb{R}$ $(n = 0, 1, 2, \dots)$, are functions with common domain D and there exist "uniform" bounds $M_n \in \mathbb{R}$ such that $|f_n(x)| \le M_n$ for all $n \in \mathbb{N}$ and $x \in D$. If the series $\sum_{n=0}^{\infty} M_n$ converges in $\mathbb{R}$ (i.e., $\sum_{n=0}^{\infty} M_n < \infty$) then the function series $\sum_{n=0}^{\infty} f_n$ converges uniformly.

Proof.

First we show that $\sum_{n=0}^{\infty} f_n$ converges point-wise.
Fix $x \in D$. Since $\sum_{n=0}^{\infty} M_n$ is convergent and $|f_n(x)| \le M_n$, the comparison test yields that $\sum_{n=0}^{\infty} f_n(x)$ is absolutely convergent and hence convergent.
Thus $\sum_{n=0}^{\infty} f_n$ converges point-wise and has a limit function $F\colon D \to \mathbb{R}$, $x \mapsto \sum_{n=0}^{\infty} f_n(x)$.

That the convergence is uniform is shown on the next slide. First recall that $\sum_{n=0}^{\infty} f_n$ refers to the sequence of partial sums $F_n = \sum_{k=0}^{n} f_k$, i.e., $F_n\colon D \to \mathbb{R}$, $x \mapsto \sum_{k=0}^{n} f_k(x)$.

Proof cont'd.

We estimate as follows:

$$|F(x) - F_n(x)| = \left| \sum_{k=n+1}^{\infty} f_k(x) \right| \leq \sum_{k=n+1}^{\infty} |f_k(x)| \leq \sum_{k=n+1}^{\infty} M_k$$

Since $\sum_{n=0}^{\infty} M_n$ converges, we can find, for every $\epsilon > 0$, an index N such that $\sum_{k=N+1}^{\infty} M_k < \epsilon$. Using the above estimate and $M_k \geq 0$ then shows $|F(x) - F_n(x)| < \epsilon$ for all $n \geq N$ and $x \in D$. This completes the proof. □

Application to trigonometric series

The function series

$$\sum_{n=1}^{\infty} \frac{\cos(nx)}{n^2}, \qquad \sum_{n=1}^{\infty} \frac{\sin(nx)}{n^2}$$

converge uniformly on $\mathbb{R}$ (and hence represent continuous functions of x with domain $\mathbb{R}$).

To prove this, e.g., for the first series, use the estimate $\left| \frac{\cos(nx)}{n^2} \right| \leq \frac{1}{n^2}$. Since the series $\sum_{n=1}^{\infty} 1/n^2$ is convergent, Weierstrass's Criterion can be applied with $M_n = 1/n^2$.

Application to Power Series

A (complex) power series $\sum_{n=0}^{\infty} a_n(z-a)^n$ with radius of convergence $R > 0$ (including the possibility $R = \infty$) represents a differentiable (holomorphic) function $f(z) = \sum_{n=0}^{\infty} a_n(z-a)^n$ on the open disk $\mathrm{B}_R(a) = \{z \in \mathbb{C}; |z-a| < R\}$ (respectively, on $\mathbb{C}$ if $R = \infty$) and can be differentiated term-wise:

$$f'(z) = \sum_{n=1}^{\infty} n a_n (z-a)^{n-1} = \sum_{n=0}^{\infty} (n+1) a_{n+1} (z-a)^n.$$

Moreover, the radius of convergence of the derived series is again R.
$\Longrightarrow$ We can iterate the argument, showing that f has derivatives of all orders explicitly given by

$$\begin{aligned} f^{(k)}(z) &= \sum_{n=k}^{\infty} n(n-1)\cdots(n-k+1) a_n (z-a)^{n-k} \\ &= \sum_{n=0}^{\infty} (n+1)(n+2)\cdots(n+k) a_{n+k} (z-a)^n, \quad k \in \mathbb{N}. \end{aligned}$$

These facts are proved on the next slides. The key step is to show that power series converge uniformly on all strictly smaller disks $\mathrm{B}_{R'}(a)$, $R' < R$ (but not necessarily on $\mathrm{B}_R(a)$).

Power Series cont'd

For the proof of the key step we choose $z_1 = a + (R' + R)/2$, so that $R' < |z_1 - a| < R$. (In the case $R = \infty$, in which R' may be any positive radius, we can take $z_1 = a + 2R'$.)

For $z \in \mathrm{B}_{R'}(a)$ (in fact $|z-a| \le R'$ suffices) we then have

$$|a_n(z-a)^n| = |a_n(z_1-a)^n| \left|\frac{z-a}{z_1-a}\right|^n \le |a_n(z_1-a)^n| \left(\frac{2R'}{R'+R}\right)^n$$
$$= |a_n(z_1-a)^n|\,\theta^n$$

with $\theta := \frac{2R'}{R'+R} < 1$.

Since $|z_1 - a| < R$, the series $\sum_{n=0}^{\infty} a_n(z_1-a)^n$ converges. Hence we have $|a_n(z_1-a)^n| \le M$ for some constant M and $|a_n(z-a)^n| \le M\theta^n$ on $\mathrm{B}_{R'}(a)$. Since $\sum_{n=0}^{\infty} M\theta^n = \frac{M}{1-\theta}$ converges, we can apply Weierstrass's Criterion to conclude that $\sum_{n=0}^{\infty} a_n(z-a)^n$ converges uniformly on $\mathrm{B}_{R'}(a)$.

Note

If a power series $\sum_{n=0}^{\infty} a_n(z-a)^n$ converges for some $z_1 \in \mathbb{C}$, it necessarily converges for all $z \in \mathbb{C}$ with $|z-a| < |z_1-a|$ (i.e., in the open disk with center a and z_1 on its boundary). For the proof we can use the same estimate as above with $\theta := \frac{|z-a|}{|z_1-a|}$.

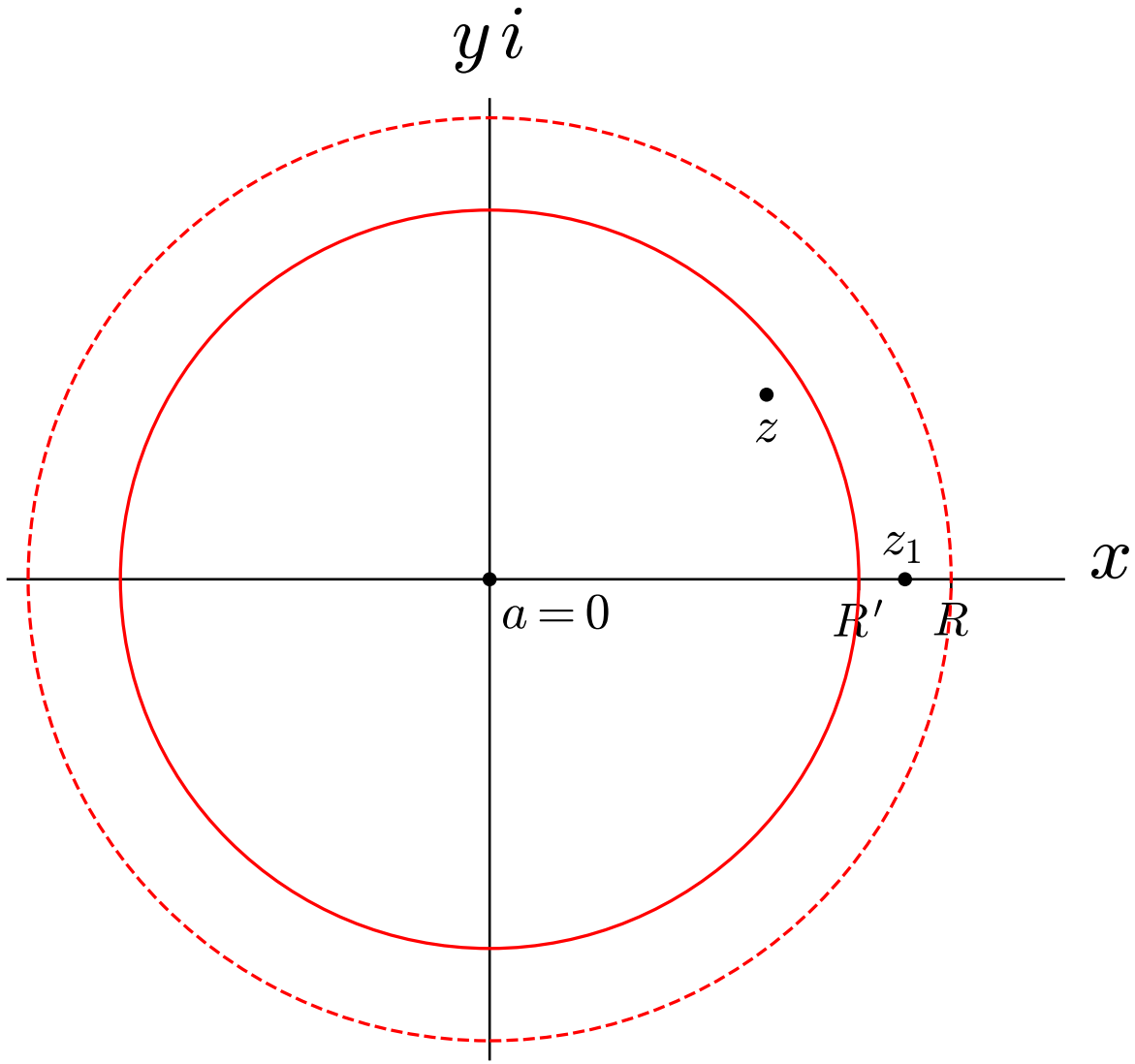


Figure: The geometry behind the proof of the key step (assuming $a = 0$): If $|z - a| \leq R'$ then $\frac{|z-a|}{|z_1-a|} \leq \frac{R'}{(R+R')/2} = \frac{2R'}{R+R'} < 1$.

Note cont'd

This observation implies that the number

$$R := \sup\{r \in \mathbb{R}; \sum_{n=0}^{\infty} a_n r^n \text{ converges in } \mathbb{C}\}$$

has the property that $\sum_{n=0}^{\infty} a_n(z-a)^n$ converges for $|z-a| < R$ and diverges for $|z-a| > R$. (For, if $|z-a| < R$ then there exists $r > |z-a|$ for which $\sum_{n=0}^{\infty} a_n r^n$ converges, and hence the observation with $z_1 := a + r$ yields that $\sum_{n=0}^{\infty} a_n(z-a)^n$ converges. Similarly, if the power series would converge for some z_1 with $|z_1 - a| > R$, it would necessarily converge for all $z = a + r$ with $R < r < |z_1 - a|$, contradicting the definition of R.)

Thus we have proved that a complex power series has a *radius of convergence* in the first place; cf. our Calculus textbook [Ste21], Theorem 11.8.4. (The proof of Th. 11.8.4 given in Appendix F generalizes to complex numbers x and is essentially the same as our argument.)

Power Series cont'd

The radius of convergence of $\sum_{n=0}^{\infty} a_n(z-a)^n$ is given by the CAUCHY-HADAMARD formula

$$R = \frac{1}{L}, \quad \text{where} \quad L = \limsup \sqrt[n]{|a_n|}. \tag{CH}$$

Here "lim sup" (limit superior) refers to the largest accumulation point of a sequence (including the possibilities $\pm\infty$ for sequences which are unbounded from above/below) and coincides with the ordinary limit if the limit exists. It is necessary to use "lim sup", because for lacunary power series (power series with "gaps") such as $\sum_{k=0}^{\infty} z^{2k}$ or $\sum_{k=1}^{\infty} z^{k^2}$ the ordinary limit $\lim_{n\to\infty} \sqrt[n]{|a_n|}$ doesn't exist, but the limit superior is 1 and gives the correct value $R = 1$. For a proof of (CH) see HW3, Ex. H19, and for the said special case see also [Ste21], Ch. 11.8, Ex. 43.

Since $\sqrt[n]{n} \to 1$ for $n \to \infty$, its follows that a power series $\sum_{n=0}^{\infty} a_n(z-a)^n$ and its derived series $\sum_{n=1}^{\infty} na_n(z-a)^{n-1}$ have the same radius of convergence, as asserted earlier. Hence the derived series converges uniformly for $|z-a| \le R' < R$ as well.

Power Series cont'd

A different formula for the radius of convergence can be obtained from the ratio test for ordinary series: $R = \lim_{n\to\infty} \frac{|a_n|}{|a_{n+1}|}$, provided that this limit exists; cf. [Ste21], Ch. 11.8, Ex. 44. This formula, often called *ratio test for power series*, also fails for lacunary power series.

Sometimes one can apply the ordinary ratio test to the series obtained by omitting the gaps. For example, in the case of $\sum_{k=1}^{\infty} z^{k^2}$ we can set $b_k = z^{k^2}$ and obtain

$$\frac{|b_{k+1}|}{|b_k|} = |z|^{(k+1)^2-k^2} = |z|^{2k+1} \to \begin{cases} 0 & \text{if } |z| < 1, \\ \infty & \text{if } |z| > 1, \end{cases}$$

showing together with the ratio test for ordinary series that $\sum_{k=1}^{\infty} z^{k^2}$ has radius of convergence $R = 1$.
But even in this modified form the ratio test is weaker than the Cauchy-Hadamard formula, since, e.g., it can't be applied to

$$z + 2\,z^2 + z^3 + 2\,z^4 + z^5 + 2\,z^6 + \cdots,$$

which clearly satisfies $\sqrt[n]{a_n} \to 1$ for $n \to \infty$ and hence has $R = 1$.

Power Series cont'd

The term-wise differentiabilty of complex power series can be proved by generalizing the Differentiation Theorem to complex derivatives. This requires the concept of complex line integrals and will be discussed on the next two slides. Here is a direct proof of this fact (w.l.o.g. we can assume $a = 0$):

$$\begin{aligned}\frac{f(z)-f(z_0)}{z-z_0} &= \sum_{n=0}^{\infty} a_n \frac{z^n - z_0^n}{z-z_0} \\ &= \sum_{n=1}^{\infty} a_n (z^{n-1} + z^{n-2}z_0 + \cdots + z_0^{n-1}) \\ &\to \sum_{n=1}^{\infty} n a_n z_0^{n-1} \quad \text{for } z \to z_0,\end{aligned}$$

provided we can interchange the two limits. This is precisely what the Continuity Theorem asserts. So we have to prove uniform convergence of the above series in some neighborhood of z_0, which can be done using the Weierstrass criterion and the estimate

$$\left| a_n (z^{n-1} + z^{n-2}z_0 + \cdots + z_0^{n-1} \right| \le n\,|a_n| \left(\max\{|z|, |z_0|\}\right)^{n-1}.$$

Power Series cont'd

Given z_0 with $|z|_0 < R$, we set $R' := (R + |z_0|)/2$. Then $z_0 \in \mathrm{B}_{R'}(0)$, and for $z \in \mathrm{B}_{R'}(0)$ the series terms are bounded in absolute value by $M_n = n\,|a_n|\,(R')^{n-1}$. Since $R' < R$, the series $\sum_{n=1}^{\infty} M_n$ converges, and hence the above series converges uniformly on $\mathrm{B}_{R'}(0)$. Thus $\mathrm{B}_{R'}(0)$ provides the desired neighborhood of z_0, and the proof is complete.

Note

The series representing $\frac{f(z)-f(z_0)}{z-z_0}$ is not a power series, but like $\sum_{n=0}^{\infty} a_n z^n$ and $\sum_{n=1}^{\infty} n a_n z^{n-1}$ converges uniformly on every disk $\mathrm{B}_{R'}(0)$ with $R' < R$. Whereas uniform convergence of $\sum_{n=0}^{\infty} a_n z^n$ yields only the continuity of f and that of $\sum_{n=1}^{\infty} n a_n z^{n-1}$ requires reasoning beyond the ordinary Differentation Theorem to yield the differentiability of f (see below), the present argument yields both properties (recall that differentiable functions are automatically continuous) in the most economic way.

Power Series cont'd

Finally, we transfer our proof of the Differentiation Theorem to the present complex setting. Writing $f_n(z) = \sum_{k=0}^{n} a_k (z-a)^k$, we have

$$f_n(z) = f_n(a) + \int_a^z f_n'(w)\,\mathrm{d}w\,.$$

Here $\int_a^z f_n'(w)\,\mathrm{d}w$ is the (path-independent) complex line integral of the (closed) differential 1-form $f_n'(z)\,\mathrm{d}z$ from a to z, which can be computed using the straight line path $\gamma(t) = a + t(z-a)$, $t \in [0,1]$, as $\int_0^1 f_n'\big(\gamma(t)\big)\gamma'(t)\,\mathrm{d}t$.

From the preceding slide we know that (f_n') converges uniformly on $[a,z]$ (the line segment joining a zu z, which is contained in a suitable disk $\mathrm{B}_{R'}(a)$) to $g(z) = \sum_{n=1}^{\infty} n a_n (z-a)^{n-1}$. Together with the estimate

$$\left|\int_a^z f_n'(w)\,\mathrm{d}w - \int_a^z g(w)\,\mathrm{d}w\right| = \left|\int_a^z f_n'(w) - g(w)\,\mathrm{d}w\right|$$
$$\leq \left(\max_{w\in[a,z]} |f_n'(w) - g(w)|\right)|z-a|$$

this shows $\lim_{n\to\infty} \int_a^z f_n'(w)\,\mathrm{d}w = \int_a^z g(w)\,\mathrm{d}w$ und further,

Power Series cont'd

letting $n \to \infty$ in the above identity,

$$f(z) = f(a) + \int_a^z g(w)\,\mathrm{d}w \quad \text{for } z \in \mathrm{B}_R(a).$$

Finally, as in the proof of a theorem in Calculus III ("independence of path of $\int_\gamma \omega$ implies exactness of ω") we obtain from this

$$\frac{f(z) - f(z_0)}{z - z_0} - g(z_0) = \frac{1}{z - z_0} \int_{z_0}^z g(w) - g(z_0)\,\mathrm{d}w$$

for $z_0, z \in \mathrm{B}_R(a)$ with $z \neq z_0$, which for $z \to z_0$ tends to zero on account of the continuity of g; cp. the estimate for $\int_a^z f_n'(w) - g(w)\,\mathrm{d}w$ on the previous slide.
This shows that f is differentiable with $f' = g$.

Remarks

- The Diffentiation Theorem holds more generally for sequences of uniformly convergent holomorphic functions $f_n\colon D \to \mathbb{C}$, $D \subseteq \mathbb{C}$; it is even sufficient that every point $z \in D$ has a neighborhood on which the convergence is uniform.

  If you think we should rather have required uniform convergence of the sequence (f_n')—true, but surprisingly this is equivalent to uniform convergence of (f_n) in the complex case!

- The uniform convergence of power series on proper subdisks (with the same center) of their open disk $\mathrm{B}_R(a)$ of convergence also implies that power series may be integrated term-wise along any path γ contained in $\mathrm{B}_R(a)$. This follows from the analogue of the Integration Theorem for line integrals in the plane, which can be deduced from the Integration Theorem and (assuming the parameter intervall of γ is $[0,1]$) the explicit formula $\int_\gamma f(z)\,\mathrm{d}z = \int_0^1 f\big(\gamma(t)\big)\gamma'(t)\,\mathrm{d}t$. The factor $\gamma'(t)$ doesn't affect uniform convergence, since it is bounded.

  In particular this holds for ordinary integrals of real power series $\sum_{n=0}^\infty a_n(x-a)^n$, $a_n, a \in \mathbb{R}$, over compact intervals $[\alpha,\beta]$ that are contained in $\mathrm{B}_R(a)$; cp. subsequent example.

Remarks (cont'd)

- As an interesting fact, note that the formula for the derivatives of a power series, viz. $f^{(k)}(z) = \sum_{n=0}^{\infty}(n+1)\dots(n+k)a_{n+k}(z-a)^n$, implies $f^{(k)}(a) = k!\,a_k$ for $k = 0, 1, 2, \dots$, i.e., a power series is its own Taylor series, and knowledge of f in an arbitrarily small neighborhood of a (and hence of its derivatives $f^{(k)}(a)$) determines the coefficients $a_k = f^{(k)}(a)/k!$ and hence f uniquely.
- Power series with coefficients $a_n \in \mathbb{R}$, center $a \in \mathbb{R}$ and radius of convergence $R > 0$ define an ordinary real function $f\colon (a-R, a+R) \to \mathbb{R}$, $x \mapsto \sum_{n=0}^{\infty} a_n(x-a)^n$. Since these are discussed in our Calculus textbook [Ste21], Ch. 11.8–11.10, I suppose you are at least familiar with this more restricted view of power series, which doesn't reveal some important aspects of the theory, though, for example why are we saying "radius of convergence"? For understanding Math 285 the restricted view will be mostly enough, because power series solutions of ODE's, to be discussed later, will only involve real power series. Holomorphic functions, complex differential forms $f(z)\,\mathrm{d}z$ and their properties, and complex line integrals $\int_\gamma f(z)\,\mathrm{d}z$ won't be needed in the sequel.

Examples

Example

In a mathematically rigorous development of Calculus the trigonometric functions sin, cos are defined by their power series expansions, which amounts to taking real and imaginary part in the expansion $e^{ix} = \sum_{n=0}^{\infty} (ix)^n/n!$ (thus giving $e^{ix} = \cos x + i \sin x$):

$$\cos x = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k)!} x^{2k}, \quad \sin x = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)!} x^{2k+1}.$$

Both series have radius of convergence $R = \infty$, and therefore can be differentiated (and integrated) term-wise for every x. This gives the known relations $\sin' = \cos$, $\cos' = -\sin$; e.g.,

$$\frac{d}{dx} \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)!} x^{2k+1} = \sum_{k=0}^{\infty} \frac{d}{dx} \left(\frac{(-1)^k}{(2k+1)!} x^{2k+1} \right)$$

$$= \sum_{k=0}^{\infty} \frac{(-1)^k (2k+1)}{(2k+1)!} x^{2k} = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k)!} x^{2k} = \cos x.$$

Example (cont'd)

The preceding dicussion holds more generally for complex arguments (replace $x \in \mathbb{R}$ by $z \in \mathbb{C}$ throughout). In the complex world there is no need to distinhuish between trigonometric and hyperbolic functions:

$$\cos z = \tfrac{1}{2}\left(e^{iz} + e^{-iz}\right) = \cosh(iz),$$
$$\sin z = \tfrac{1}{2i}\left(e^{iz} - e^{-iz}\right) = -i\sinh(iz),$$

and hence sin, cos are obtained from sinh, cosh by 90° rotations in the domain/codomain. Thus, e.g., $\cos(iy) = \cosh(-y) = \cosh y$, revealing that the complex cosine function on the imaginary axis looks like the real hyperbolic cosine, and $\sin(iy) = i\sinh y$, having a similar geometric interpretation.

Exercise

Using $e^{x+iy} = e^x \cos y + i\, e^x \sin y$, show that the complex cosine and sine functions have the following explicit representation:

$$\cos z = \cos(x+iy) = \cos x \cosh y - i \sin x \sinh y,$$
$$\sin z = \sin(x+iy) = \sin x \cosh y + i \cos x \sinh y.$$

The next example is for the integration theorem, since this is the one most widely applicable (due to the fact that integration "smoothes" functions, while differentiation "roughens" them).

Example

The *sine integral* (function) is defined as

$$\operatorname{Si}(x) = \int_0^x \frac{\sin t}{t}\,\mathrm{d}t = \int_0^x \sum_{n=0}^{\infty} (-1)^n \frac{t^{2n}}{(2n+1)!}\,\mathrm{d}t \quad \text{for } x \in \mathbb{R}.$$

Since the power series defining $\sin x$ and $\sin(x)/x$ have radius of convergence ∞, the function series in the definition of $\operatorname{Si}(x)$ converges uniformly on every interval $[0, x]$ and hence can be integrated termwise:

$$\begin{aligned}\operatorname{Si}(x) &= \sum_{n=0}^{\infty} \int_0^x (-1)^n \frac{t^{2n}}{(2n+1)!}\,\mathrm{d}t = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} \left[\frac{t^{2n+1}}{2n+1}\right]_0^x \\ &= \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)!(2n+1)} = x - \frac{x^3}{18} + \frac{x^5}{600} - \frac{x^7}{35280} \pm \cdots.\end{aligned}$$

Of course, the sine series itself can also be integrated term-wise over $[0, x]$, producing the power series of $1 - \cos x$.

Power series are very useful in combinatorial enumeration. Here is one of my favorite examples in this regard. (Students of Discrete Mathematics may have seen it earlier.)

Example

Find a closed formula for $s_n = 1^2 + 2^2 + \cdots + n^2$. In high school you may have seen the formula already and been asked to prove it, but how to discover it in the first place?

Using power series this can be done as follows. Start with the geometric series and differentiate it term-wise (valid for $|x| < 1$):

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x},$$

$$\Longrightarrow \sum_{n=1}^{\infty} n x^{n-1} = \frac{\mathrm{d}}{\mathrm{d}x}\frac{1}{1-x} = \frac{1}{(1-x)^2},$$

$$\Longrightarrow \sum_{n=1}^{\infty} n x^{n} = x\frac{\mathrm{d}}{\mathrm{d}x}\frac{1}{1-x} = \frac{x}{(1-x)^2}.$$

From this we see that the operator $x(\mathrm{d}/\,\mathrm{d}x)$ ("first differentiate, then mutliply by x") effects the transformation $(a_n) \mapsto (na_n)$ on the corresponding coefficient sequence of a power series.

Example (cont'd)

Hence, applying the operator twice we obtain

$$\sum_{n=1}^{\infty} n^2 x^n = x \frac{\mathrm{d}}{\mathrm{d}x} \frac{x}{(1-x)^2} = \frac{x+x^2}{(1-x)^3}.$$

Further we have

$$\begin{aligned} \frac{1}{1-x} \sum_{n=1}^{\infty} n^2 x^n &= \left(\sum_{n=0}^{\infty} x^n \right) \left(\sum_{n=1}^{\infty} n^2 x^n \right) \\ &= 1^2 x + (1^2 + 2^2) x^2 + (1^2 + 2^2 + 3^2) x^3 + \cdots + \\ &= \sum_{n=1}^{\infty} s_n x^n. \end{aligned}$$

In general, the operator "$\times \frac{1}{1-x}$" effects on the corresponding coefficient sequence of a power series the transformation $(a_0, a_1, a_2, \dots) \mapsto (a_0, a_0 + a_1, a_0 + a_1 + a_2, \dots)$, i.e., taking the partial sums of the sequence.

Example (cont'd)

Putting both computations together, we get

$$\sum_{n=1}^{\infty} s_n x^n = \frac{x+x^2}{(1-x)^4}.$$

This tells us that the so-called *generating function* of the sequence (s_n) is the rational function $\frac{x+x^2}{(1-x)^4}$. A closed formula for s_n may then be obtained by expanding $\frac{x+x^2}{(1-x)^4}$ into a power series and comparing coefficients. This can be done using partial fractions or, if you happen to know the power series expansion $(1-x)^{-s} = \sum_{n=0}^{\infty} \binom{n+s-1}{s-1} x^n$, $|x| < 1$, $s \in \mathbb{N}$, quickly as follows:

$$\begin{aligned} \Longrightarrow \sum_{n=0}^{\infty} s_n x^n = \frac{x+x^2}{(1-x)^4} &= (x+x^2) \sum_{n=0}^{\infty} \binom{n+3}{3} x^n \\ &= \sum_{n=0}^{\infty} \left(\binom{n+2}{3} + \binom{n+1}{3} \right) x^n, \end{aligned}$$

and hence $s_n = \binom{n+2}{3} + \binom{n+1}{3} = \frac{(2n+1)(n+1)n}{6}$.

Example (geometry of the complex geometric series)

The geometric series evaluation

$$\sum_{n=0}^{\infty} z^n = 1 + z + z^2 + z^3 + \cdots = \frac{1}{1-z}$$

is valid for all complex numbers z with $|z| < 1$. This follows from $1 + z + \cdots + z^n = \frac{1-z^{n+1}}{1-z}$ and $z^{n+1} \to 0$ for $n \to \infty$.

For example, since $\left|\frac{\mathrm{i}}{2}\right| = \frac{1}{2}$, $\left|\frac{1+\mathrm{i}}{2}\right| = \frac{1}{2}\sqrt{2}$, we have

$$\sum_{n=0}^{\infty} \left(\frac{\mathrm{i}}{2}\right)^n = \frac{1}{1-\mathrm{i}/2} = \frac{2}{2-\mathrm{i}} = \frac{4}{5} + \frac{2}{5}\,\mathrm{i},$$

$$\sum_{n=0}^{\infty} \left(\frac{1+\mathrm{i}}{2}\right)^n = \frac{1}{1-(1+\mathrm{i})/2} = \frac{2}{1-\mathrm{i}} = 1 + \mathrm{i}.$$

Since complex numbers are just vectors in the plane (which also can be multiplied), these limits have nice geometric illustrations; cf. next slide.

For the snakes' shapes note that multiplication by $\mathrm{i}/2$ amounts to a 90° rotation and a scaling by 0.5, and similarly for $(1+\mathrm{i})/2$.

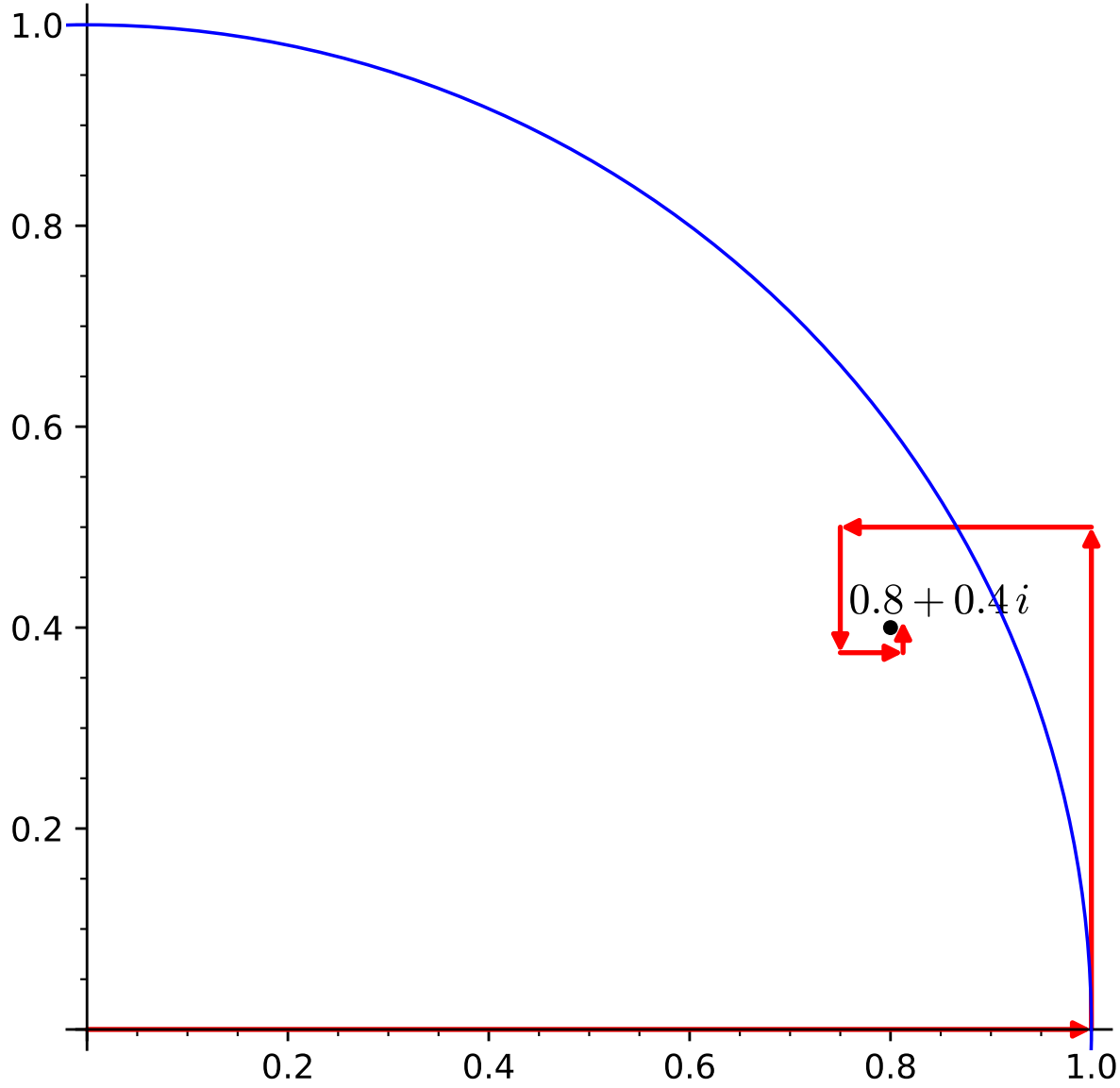


Figure: Illustration of $\displaystyle\sum_{k=0}^{5} \left(\frac{\mathrm{i}}{2}\right)^k \approx \frac{4}{5} + \frac{2}{5}\,\mathrm{i}$

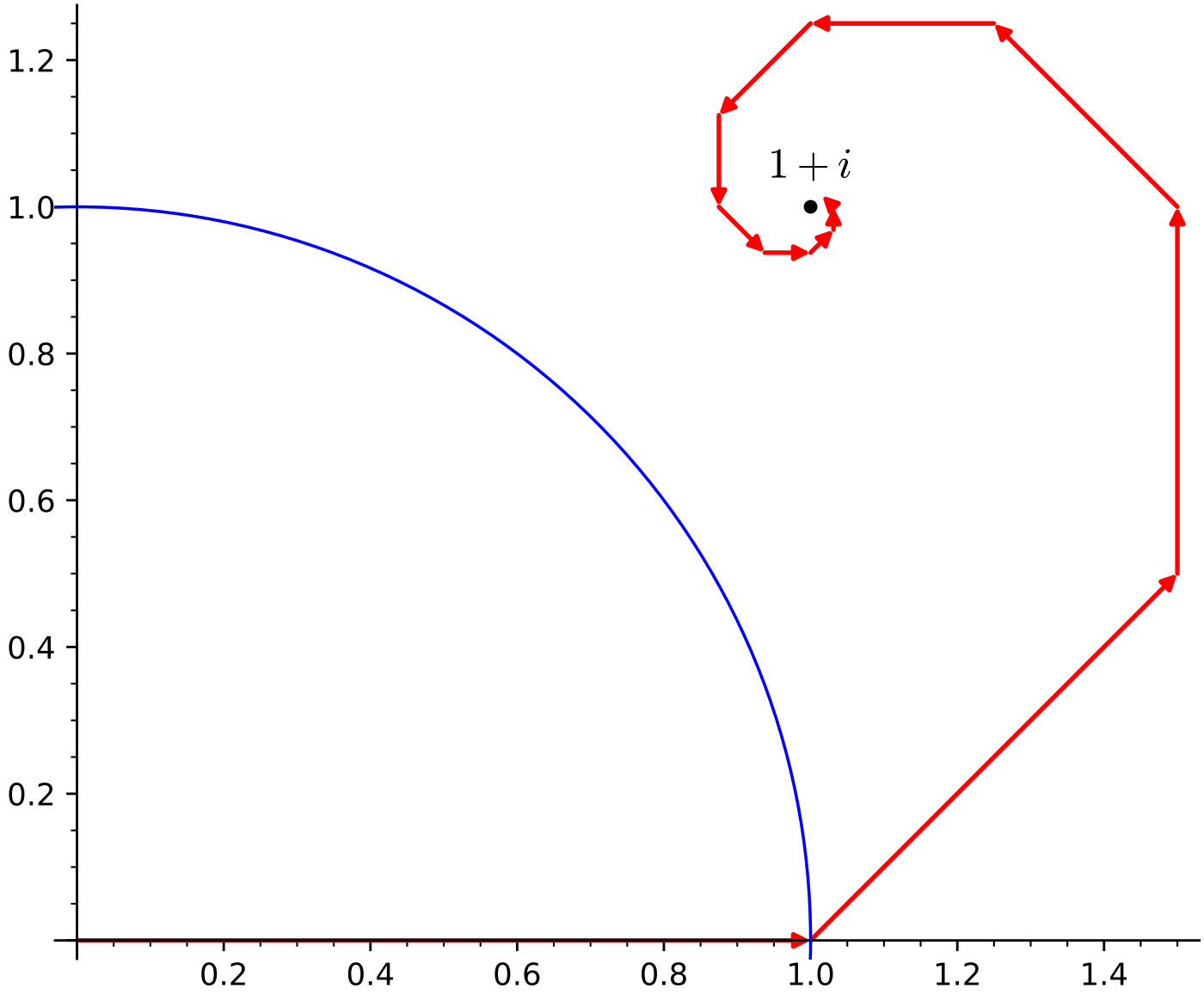


Figure: Illustration of $\sum_{k=0}^{11} \left(\frac{1+\mathrm{i}}{2}\right)^k \approx 1+\mathrm{i}$

Complex vs Real Differentiability

Since complex numbers $z = (x, y) = x + y\mathrm{i}$ are just points in the plane, a function $f\colon D \to \mathbb{C}$, $D \subseteq \mathbb{C}$, corresponds to a pair of real 2-variable functions $u, v\colon D \to \mathbb{R}$ via

$$f(z) = f(x, y) = u(x, y) + \mathrm{i}\, v(x, y), \quad \text{resp.,} \quad u = \operatorname{Re} f,\ v = \operatorname{Im} f.$$

Complex differentiabililty of f in $z = (x, y) \in D$ (which must be an inner point of D), i.e., the existence of the limit

$$f'(z) = \lim_{\substack{h \to 0 \\ h \in \mathbb{C}}} \frac{f(z+h) - f(z)}{h} = \lim_{(h_1, h_2) \to (0,0)} \frac{f(x + h_1, y + h_2) - f(x, y)}{h_1 + h_2\mathrm{i}},$$

has a nice characterization in terms of real (total) differentiability of u, v in (x, y) and certain conditions on the partial derivatives u_x, u_y, v_x, v_y in (x, y). For this note that it is equivalent to the existence of $c \in \mathbb{C}$ such that $\lim_{h \to 0} \big(f(z+h) - f(z) - ch\big)/h = 0$. (If applicable, we have $f'(z) = c$.) Real differentiability of f in (x, y) in turn means the existence of $a_{11}, a_{12}, a_{21}, a_{22} \in \mathbb{R}$ such that $\lim_{h \to 0} \big(u(x + h_1, y + h_2) - u(x, y) - a_{11}h_1 - a_{12}h_2\big)/|h| = 0$ and $\lim_{h \to 0} \big(v(x + h_1, y + h_2) - v(x, y) - a_{21}h_1 - a_{22}h_2\big)/|h| = 0$.

Theorem

f is complex differentiable in $z = (x, y)$ iff f is real differentiable in (x, y) (i.e., u, v are differentiable in (x, y)) and

$$u_x(x, y) = v_y(x, y), \quad u_y(x, y) = -v_x(x, y).$$

In particular, f is complex differentiable per se (i.e., in every point of D, which requires D to be open) iff f is real differentiable and u, v satisfy the so-called *Cauchy-Riemann PDE's*
$u_x = v_y \wedge u_y = -v_x$.

Proof.

If f is complex differentiable in z with $f'(z) = a + b\mathrm{i}$ then

$$\begin{aligned} f(z+h) - f(z) &= f'(z)h + \mathrm{o}(h) = (a + b\mathrm{i})(h_1 + h_2\mathrm{i}) + \mathrm{o}(h) \\ &= ah_1 - bh_2 + (ah_2 + bh_1)\mathrm{i} + \mathrm{o}(h), \end{aligned}$$

where $g(h) = \mathrm{o}(h)$ means $g(h)/|h| \to 0$ for $h \to 0$.
Extracting real and imaginary part we obtain

$$\begin{aligned} u(x + h_1, y + h_2) - u(x, y) &= ah_1 - bh_2 + \mathrm{o}(h), \\ v(x + h_1, y + h_2) - v(x, y) &= bh_1 + ah_2 + \mathrm{o}(h), \end{aligned}$$

Proof cont'd.

... which says that u, v are differentiable in (x, y) with $u_x(x, y) = a$, $u_y(x, y) = -b$, $v_x(x, y) = b$, $v_y(x, y) = a$; in particular we have $u_x(x, y) = v_y(x, y)$, $u_y(x, y) = -v_x(x, y)$.

Conversely, if u, v satisfy the conditions of the theorem, it is equally easy to see that f is complex differentiable in z with $f'(z) = u_x(x, y) + \mathrm{i}\, v_x(x, y)$. □

Note

The Cauchy-Riemann PDE's say that the Jacobi matrix $\mathbf{J}_f(x, y)$ has the special form

$$\mathbf{J}_f = \begin{pmatrix} u_x & u_y \\ v_x & v_y \end{pmatrix} = \begin{pmatrix} u_x & -v_x \\ v_x & u_x \end{pmatrix},$$

i.e., it is at every point (x, y) a scaled rotation matrix.

Principal Branch of the Complex Logarithm

The equation

$$w = e^z = e^{x+iy} = e^x e^{iy} = e^x \cos y + i\, e^x \sin y$$

is solvable for each nonzero $w \in \mathbb{C}$, and the solutions are

$$z_k = \ln|w| + i(\arg w + 2k\pi), \quad k \in \mathbb{Z},$$

where $\arg w \in (-\pi, \pi]$ is the angle of w in polar coordinates. (To see this, write $w = re^{i\phi}$ and compare with $e^x e^{iy}$.)

Considering the "principal" solution z_0 as a function of w and swapping notation, we obtain the *principal branch of the complex logarithm*

$$\ln z = \ln|z| + i \arg z = \ln\sqrt{x^2+y^2} + i \arctan(y/x), \quad x = \operatorname{Re} z > 0.$$

By definition, the logarithm satisfies $e^{\ln z} = z$ in its domain, which can be extended to the "slotted" plane $\mathbb{C} \setminus \{(x,0); x \le 0\}$ (with a different expression for the imaginary part) without affecting the truth of the following theorem.

Theorem

$f(z) = \ln z$ is complex differentiable with $f'(z) = 1/z$.

Proof.

We assume $\operatorname{Re} z > 0$, so that $v(x,y) = \arctan(y/x)$ can be used.

$$f(x,y) = \left(\ln\sqrt{x^2+y^2}, \arctan(y/x)\right),$$

$$u_x = \frac{\mathrm{d}}{\mathrm{d}x}\ln\sqrt{x^2+y^2} = \frac{x}{x^2+y^2},$$

$$u_y = \frac{\mathrm{d}}{\mathrm{d}y}\ln\sqrt{x^2+y^2} = \frac{y}{x^2+y^2},$$

$$v_x = \frac{\mathrm{d}}{\mathrm{d}x}\arctan(y/x) = \frac{-y/x^2}{1+(y/x)^2} = -\frac{y}{x^2+y^2},$$

$$v_y = \frac{\mathrm{d}}{\mathrm{d}y}\arctan(y/x) = \frac{1/x}{1+(y/x)^2} = \frac{x}{x^2+y^2}.$$

Evidently the Cauchy-Riemann PDE's are satisfied, and hence f is complex differentiable with

$$f'(z) = u_x(x,y) + \mathrm{i}\, v_x(x,y) = \frac{x-y\mathrm{i}}{x^2+y^2} = \frac{\overline{z}}{|z|^2} = 1/z. \qquad \square$$

For $|z-1|<1$ we have

$$\begin{aligned}\ln z &= \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n}(z-1)^n \\ &= z-1-\frac{(z-1)^2}{2}+\frac{(z-1)^3}{3}-\frac{(z-1)^4}{4}\pm\cdots\end{aligned}$$

This can be seen as follows. The power series has radius of convergence 1, and hence for $|z-1|<1$ may be differentiated term-wise to yield

$$1-(z-1)+(z-1)^2-(z-1)^3\pm\cdots=\frac{1}{1+z-1}=\frac{1}{z},$$

the same derivative as $\ln z$.
$\Longrightarrow$ $\ln z$ differs from the power series by an additive constant. The constant must be zero, since both $\ln z$ and the power series vanish at $z=1$.

The following example plays an important role in the theory of Fourier series.

Example

The power series $\displaystyle\sum_{n=1}^{\infty}\frac{z^n}{n}$ has radius of convergence $R=1$ and hence defines an analytic (holomorphic) function $f(z)$ on the open unit disk $\mathrm{B}_1(0)=\{z\in\mathbb{C};|z|<1\}$, whose derivative can be obtained by termwise differentiation:

$$f'(z)=\sum_{n=1}^{\infty}\frac{n\,z^{n-1}}{n}=\sum_{n=1}^{\infty}z^{n-1}=\frac{1}{1-z}.$$

Together with $f(0)=0$ it follows that

$$f(z)=-\ln(1-z)=-\ln|1-z|-\mathrm{i}\arg(1-z)\quad\text{with}$$

$$\ln|1-z|=\ln\sqrt{(1-x)^2+y^2}=\ln\sqrt{1-2x+|z|^2},$$

$$\arg(1-z)=\arg(1-x-\mathrm{i}y)=-\arctan\left(\frac{y}{1-x}\right),$$

where we have written $z=x+\mathrm{i}y$; cf. the preceding discussion (or Calculus III) for the principal branch of the complex logarithm.

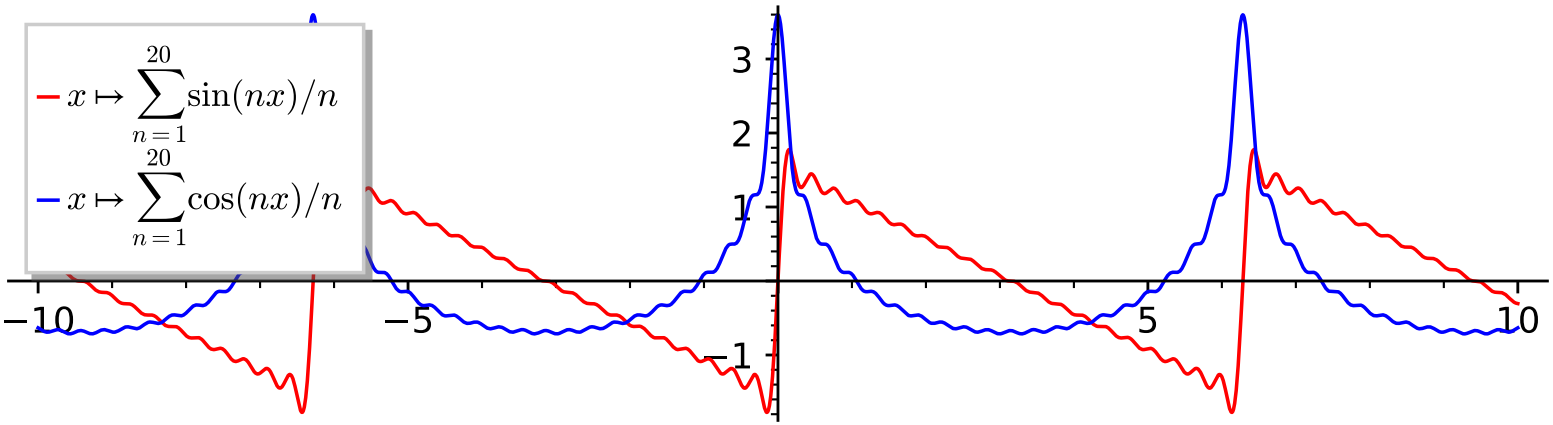
Example (cont'd)

Question: What happens on the boundary $S^1 = \{z \in \mathbb{C}; |z| = 1\}$?

A point $z \in S^1$ has the form $z = e^{ix}$ with $x \in [0, 2\pi)$ (with a different meaning of x !), and we are asking for the convergence of the series

$$f(e^{ix}) = \sum_{n=1}^{\infty} \frac{e^{inx}}{n} = \sum_{n=1}^{\infty} \frac{\cos(nx)}{n} + i \sum_{n=1}^{\infty} \frac{\sin(nx)}{n}.$$

From Calculus I we know already that the series diverges for $z = 1$ ($x = 0$), because $f(1) = \sum_{n=1}^{\infty} \frac{1^n}{n} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots$ is just the harmonic series, and converges for $z = -1$ ($x = \pi$), because $f(-1) = \sum_{n=1}^{\infty} \frac{(-1)^n}{n} = -\left(1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} \pm \cdots\right) = -\ln(2)$ (alternating harmonic series).



Example (cont'd)

From the plot it appears that:

- The series of cosines converges for $0 < x < 2\pi$ and represents a continuous (and maybe differentiable) function on $(0, 2\pi)$.
- The series of sines converges for all $x \in \mathbb{R}$ and represents the 2π-periodic function h defined by

$$h(x) = \begin{cases} (\pi - x)/2 & \text{for } 0 < x < 2\pi, \\ 0 & \text{for } x = 0, \end{cases}$$

and 2π-periodic extension to $\mathbb{R}$.

Since $h(0+) = \pi/2$, $h(0-) = h(2\pi-) = -\pi/2$, the function h has discontinuities at $x \in 2\pi\mathbb{Z}$. The value at any discontinuity x satisfies $h(x) = \frac{h(x+)+h(x-)}{2}$.

Example (cont'd)

As key step towards the proof of these assertions we now show that $f(z) = \sum_{n=1}^{\infty} \frac{z^n}{n}$ converges for all $z \in \mathbb{C}$ with $|z| \leq 1$ except for $z = 1$, and the convergence is uniform on every subset D of $\overline{B_1(0)} \setminus \{1\}$ that excludes a (small) circle around $z = 1$.

For this we use a technique called "Abel summation" or "partial summation" (a discrete analogue of integration by parts). Setting $s_n(z) = \sum_{k=1}^{n} z^k$, we have for $m, n \in \mathbb{N}$ with $m < n$

$$\sum_{k=m}^{n} \frac{z^k}{k} = \sum_{k=m}^{n} \frac{s_k(z) - s_{k-1}(z)}{k}$$

$$= -\frac{s_{m-1}(z)}{m} + \sum_{k=m}^{n-1} s_k(z) \left(\frac{1}{k} - \frac{1}{k+1} \right) + \frac{s_n(z)}{n}$$

Now suppose that $s_n(z)$ is uniformly bounded on D, i.e., there exists $M > 0$ such that $|s_n(z)| \leq M$ for all $z \in D$ and all $n \in \mathbb{N}$. Then we obtain the estimate

$$\left| \sum_{k=m}^{n} \frac{z^k}{k} \right| \leq \frac{M}{m} + M \sum_{k=m}^{n-1} \left(\frac{1}{k} - \frac{1}{k+1} \right) + \frac{M}{n} = \frac{2M}{m}.$$

Example (cont'd)

Hence, given $\epsilon > 0$, we have for the partial sums $f_n(z) = \sum_{k=1}^{n} z^k/k$ of our series the estimate

$$d_\infty(f_m, f_n) = \max\left\{\left|\sum_{k=m+1}^{n} \frac{z^k}{k}\right|; z \in D\right\} < \epsilon \quad \text{if } m, n > N_\epsilon = \lceil 2M/\epsilon \rceil.$$

This shows that the series satisfies the Cauchy-Criterion for uniform convergence on D and hence that it converges uniformly on D; cf. subsequent lecture for more details.

It yet remains to derive the bound M. This is easy, however, since

$$s_n(z) = \sum_{k=1}^{n} z^k = \frac{z^{n+1} - z}{z - 1}.$$

Hence, setting $D = D_r = \{z \in \mathbb{C}; |z| \le 1, |z - 1| \ge r\}$ for $r > 0$, we have have $|s_n(z)| \le 2/r$ for $z \in D_r$ and $n \in \mathbb{N}$, so that we can take $M = 2/r$.

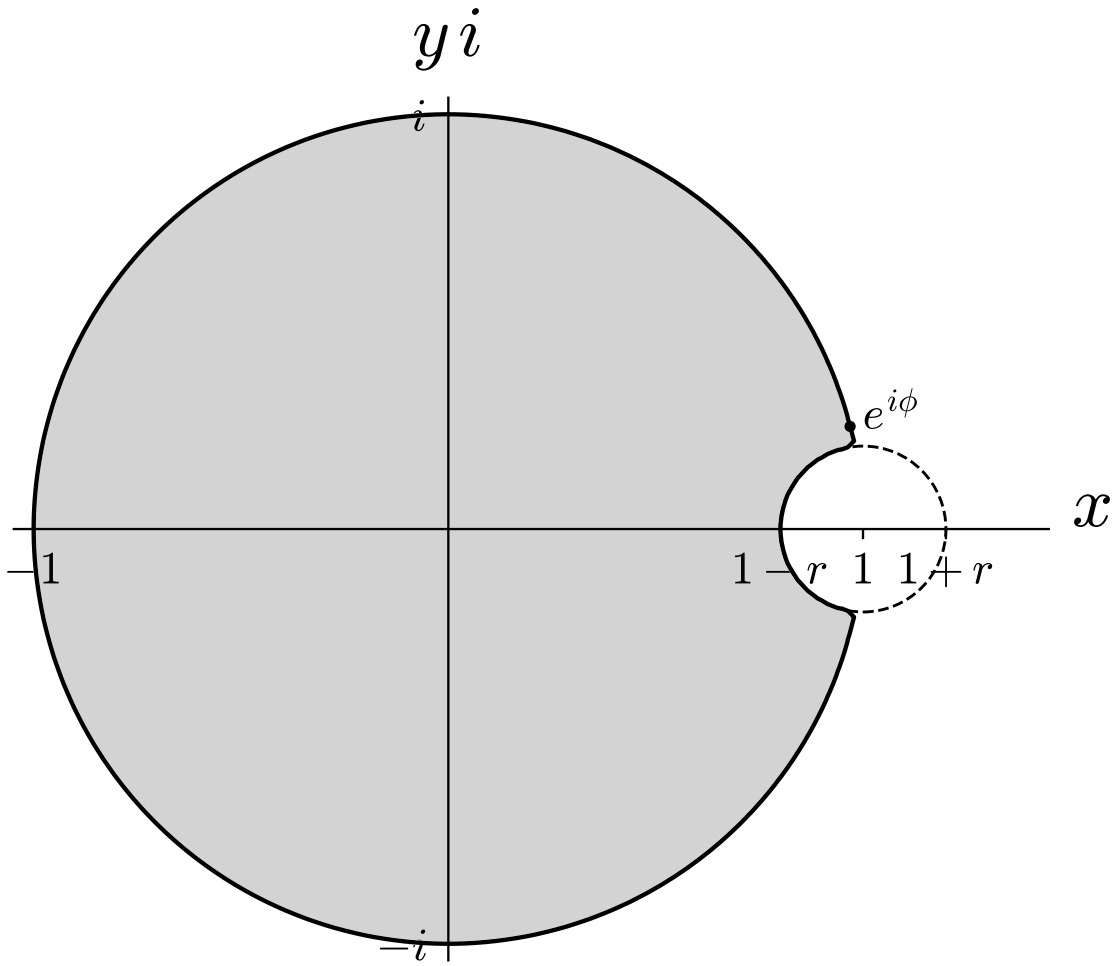


Figure: The region D_r

You may think of a map of China with Hangzhou at $1 - r$ and Shanghai at $\mathrm{e}^{\mathrm{i}\phi}$. (Question: Where is Ningbo?)

Example (cont'd)

Now the continuity theorem gives that $f(z) = \sum_{n=1}^{\infty} z^n/n$ defines a continuous function on $\overline{\mathrm{B}_1(0)} \setminus \{1\}$: To prove continuity at a particular point z_0, let $r = \frac{1}{2}|z_0 - 1|$ and use the uniform convergence on D_r.
In particular the series $\sum_{n=1}^{\infty} \cos(nx)/n$ and $\sum_{n=1}^{\infty} \sin(nx)/n$ converge for every $x \in (0, 2\pi)$ (and the second series trivially converges also for $x = 0$).

Knowing that f is continuous in $z = \mathrm{e}^{\mathrm{i}\phi} \in \mathrm{S}^1 \setminus \{1\}$, we can compute $f(\mathrm{e}^{\mathrm{i}\phi})$ from the explicit representation of f in $\mathrm{B}_1(0)$ as the limit

$$
\begin{aligned}
f(\mathrm{e}^{\mathrm{i}\phi}) &= \lim_{r\uparrow 1} f(r\mathrm{e}^{\mathrm{i}\phi}) \\
&= \lim_{r\uparrow 1} \left[-\ln\sqrt{1 - 2r\cos\phi + r^2} + \mathrm{i}\arctan\left(\frac{r\sin\phi}{1 - r\cos\phi}\right)\right] \\
&= -\ln\sqrt{2(1-\cos\phi)} + \mathrm{i}\arctan\left(\frac{\sin\phi}{1-\cos\phi}\right) \\
&= -\ln\left(2\sin\frac{\phi}{2}\right) + \mathrm{i}\,\frac{\pi - \phi}{2}, \quad \ldots
\end{aligned}
$$

Example (cont'd)

. . . where we have used

$$1-\cos\phi = 1-\left(\cos^2\frac{\phi}{2}-\sin^2\frac{\phi}{2}\right) = 2\sin^2\frac{\phi}{2},$$

$$\frac{\sin\phi}{1-\cos\phi} = \frac{2\sin\frac{\phi}{2}\cos\frac{\phi}{2}}{2\sin^2\frac{\phi}{2}} = \cot\frac{\phi}{2} = \tan\frac{\pi-\phi}{2}.$$

As a corollary we have the trigonometric series evaluations

$$\sum_{n=1}^{\infty}\frac{\cos(nx)}{n} = -\ln\left(2\sin\frac{x}{2}\right), \quad \sum_{n=1}^{\infty}\frac{\sin(nx)}{n} = \frac{\pi-x}{2} \quad (0<x<2\pi).$$

In particular, setting $x=\pi/2$ (or $z=\mathrm{e}^{\mathrm{i}x}=\mathrm{i}$) this gives

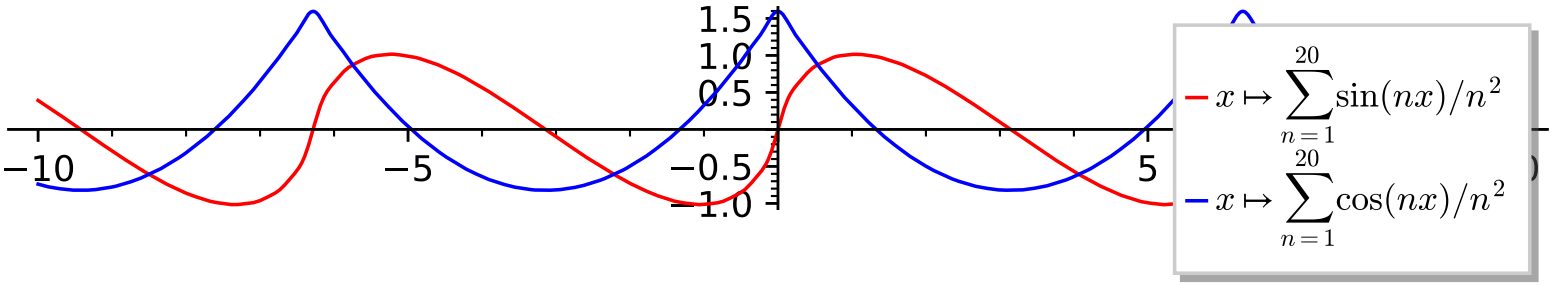
$$1-\tfrac{1}{2}+\tfrac{1}{3}-\tfrac{1}{4}\pm\cdots = \ln 2, \quad 1-\tfrac{1}{3}+\tfrac{1}{5}+\tfrac{1}{7}\pm\cdots = \pi/4.$$

In the next subsection the criterion we have used for proving the uniform convergence of $\sum_{n=1}^{\infty} z^n/n$ is stated in more generality. It is known as "Dirichlet's test for uniform convergence".

Example

We compute the related series $\displaystyle\sum_{n=1}^{\infty} \frac{\cos(nx)}{n^2}$.

Since $\left|\cos(nx)/n^2\right| \le 1/n^2$ and $\sum_{n=1}^{\infty} 1/n^2$ converges, this series converges uniformly (and absolutely) on $\mathbb{R}$ and represents a continuous, 2π-periodic function $g\colon \mathbb{R} \to \mathbb{R}$.



The series of derivatives is

$$\sum_{n=1}^{\infty} \frac{-\sin(nx)n}{n^2} = -\sum_{n=1}^{\infty} \frac{\sin(nx)}{n} = -\operatorname{Im}\left(\sum_{n=1}^{\infty} \frac{(e^{ix})^n}{n}\right).$$

From the preceding example we know that the series of derivatives converges uniformly on every interval of the form $[\delta, 2\pi - \delta]$ with $\delta > 0$ (and δ sufficiently small).

Example (cont'd)

$\Longrightarrow$ The differentiation theorem can be applied and gives

$$g'(x) = -\sum_{n=1}^{\infty} \frac{\sin(nx)}{n} = \frac{x-\pi}{2} \quad \text{for } 0 < x < 2\pi.$$

$$\Longrightarrow g(x) = \frac{(x-\pi)^2}{4} + C \quad \text{for } 0 \le x \le 2\pi,$$

where C is some constant. (Note that $g(0) = g(2\pi) = \pi^2/4 + C$, so that the 2π-periodic extension to $\mathbb{R}$ will be automatically continuous.)

The constant can be determined by evaluating the integral $\int_0^{2\pi} g(x)\,\mathrm{d}x$ in two ways:

1. Applying the integration theorem to the series defining g, we obtain

$$\int_0^{2\pi} g(x)\,\mathrm{d}x = \int_0^{2\pi} \sum_{n=1}^{\infty} \frac{\cos(nx)}{n^2}\,\mathrm{d}x = \sum_{n=1}^{\infty} \int_0^{2\pi} \frac{\cos(nx)}{n^2}\,\mathrm{d}x$$

$$= \sum_{n=1}^{\infty} \left[\frac{\sin(nx)}{n^3}\right]_0^{2\pi} = \sum_{n=1}^{\infty} \frac{\sin(2n\pi) - \sin(0)}{n^3} = 0.$$

Example (cont'd)

2. Using the expression $g(x) = (x-\pi)^2/4 + C$, we obtain

$$\int_0^{2\pi} g(x)\,\mathrm{d}x = \left[\frac{(x-\pi)^3}{12} + Cx\right]_0^{2\pi} = \frac{\pi^3}{12} + 2\pi C - \frac{(-\pi)^3}{12}$$
$$= \frac{\pi^3}{6} + 2\pi C.$$

Since this is equal to zero, we conclude $C = -\pi^2/12$, and finally

$$g(x) = \sum_{n=1}^{\infty} \frac{\cos(nx)}{n^2} = \frac{(x-\pi)^2}{4} - \frac{\pi^2}{12} \quad \text{for } 0 \le x \le 2\pi.$$

As a by-product, setting $x = 0$, resp., $x = \pi$, we obtain from this the series evaluations

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}, \qquad \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n^2} = \frac{\pi^2}{12}.$$

Exercise

a) Assuming that $\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$, show that $\sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n^2} = \frac{\pi^2}{12}$ without resorting to the evaluation of $\sum_{n=1}^{\infty} \frac{\cos(nx)}{n^2}$ on the previous slide.

Hint: Add the two series.

b) Show that $\frac{1}{1^2} + \frac{1}{3^2} + \frac{1}{5^2} + \frac{1}{7^2} + \cdots = \frac{\pi^2}{8}$.

Exercise

Determine the two series

$$\sum_{n=1}^{\infty} \frac{\sin(nx)}{n^3} \quad \text{and} \quad \sum_{n=1}^{\infty} \frac{\cos(nx)}{n^4} \quad \text{for } x \in \mathbb{R},$$

and use the results to evaluate in turn $\sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)^3}$ and $\sum_{n=1}^{\infty} \frac{1}{n^4}$.

Exercise

Riemann's Zeta function is defined for complex arguments $s = \sigma + it$ with $\sigma = \mathrm{Re}(s) > 1$ by

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}, \quad \text{where} \quad n^s = e^{\ln(n)s}.$$

1. Show that the above series converges uniformly on every closed half plane $H_\delta = \{s \in \mathbb{C}; \mathrm{Re}(s) \geq 1 + \delta, \delta > 0$, and conclude from this that ζ is continuous.
2. Using a variant of the Differentiation Theorem, show in a similar fashion that ζ is complex differentiable (in fact infinitely often) and give a series representation for $\zeta'(s)$.
3. Using properties of the prime factorization of integers, show
$$\zeta(s) = \frac{1}{(1-2^{-s})(1-3^{-s})(1-5^{-s})(1-7^{-s})(1-11^{-s})\cdots}.$$
4. Show that $(2^{1-s}-1)\zeta(s)$ has a series representation of the form $\sum_{n=1}^{\infty} a_n n^{-s}$, which converges for $\mathrm{Re}(s) > 0$ and uniformly for $\mathrm{Re}(s) \geq \delta > 0$. Conclude that $\zeta(s)$ is holomorphic in $\{s \in \mathbb{C}; \mathrm{Re}(s) > 0, s \neq 1\}$ and satisfies $\lim_{s\to 1}(s-1)\zeta(s) = 1$.

According to my experience, many students find it hard to understand and master the concept of uniform convergence. Sometimes even basic things go wrong, such as confusing function sequences and series. The following exercise, which is derived from an earlier midterm question, addresses this.

Problem

Suppose $f_n : \mathbb{R} \to \mathbb{R}$ *is defined by*

$$f_n(x) = \frac{x}{1 + n x^2}, \quad n = 0, 1, 2, \ldots$$

1. *Does the sequence* (f_n) *converge uniformly?*
2. *Does the series* $\sum_{n=0}^{\infty} f_n$ *converge uniformly?*
3. *Investigate uniform convergence of the sequence* (f_n) *by converting it to a series and applying the Weierstrass test.*

Solution

(1) First we check point-wise convergence: For $x = 0$ we have $f_n(0) = 0$, which trivially converges to 0 for $n \to \infty$. For $x \neq 0$ we have $1 + nx^2 \to +\infty$ and hence $\frac{x}{1+nx^2} \to 0$ for $n \to \infty$. (Note that x is treated as a constant here.)
$\Longrightarrow$ f_n converges point-wise to the all-zero function $\mathbb{R} \to \mathbb{R}$, $x \mapsto 0$.

Since $\lim_{x\to\pm\infty} f_n(x) = 0$ and f_n is continuous, the function f_n attains a minimum and a maximum. By symmetry, if the maximum is at $x_n > 0$ with value M_n, the minimum must be at $-x_n$ with the opposite value $-M_n$.) Uniform convergence to the all-zero function requires precisely that the resulting sequence of maxima converge to zero, because given $\epsilon > 0$ we need to find a response N such that

$$-\epsilon < f_n(x) < \epsilon \quad \text{for } n > N \text{ and all } x \in \mathbb{R},$$

which is equivalent to $M_n < \epsilon$ for $n > N$.

Since $f_n'(x) = \frac{1+n\,x^2-(2nx)x}{(1+nx^2)^2} = \frac{1-nx^2}{(1+nx^2)^2} = 0 \iff x = \pm\frac{1}{\sqrt{n}}$, the (unique) maximum of f_n is $f_n\left(\frac{1}{\sqrt{n}}\right) = \frac{1}{2\sqrt{n}}$. Hence, since $\frac{1}{2\sqrt{n}} \to 0$, we can conclude that $f_n \to 0$ uniformly.

Solution cont'd

Once can also use the inequality $1 + nx^2 \geq 2\sqrt{n}\,x$ (an instance of $a^2 + b^2 \geq 2ab$) to conclude that $|f_n(x)| \leq \frac{1}{2\sqrt{n}}$, which is sufficient for the conclusion.

(2) Here the answer is "No", because the function series doesn't even converge point-wise (except for $x = 0$). In order to see this, rewrite it for $x \neq 0$ like this:

$$\sum_{n=0}^{\infty} \frac{x}{1 + nx^2} = \sum_{n=0}^{\infty} \frac{\frac{1}{x}}{\frac{1}{x^2} + n}.$$

Since for point-wise convergence $1/x$ is treated as a constant, this shows that the series behaves like the harmonic series and hence diverges. (More precisely, for $n > 1/x^2$ the n-th summand is lower-bounded by $\frac{1}{x}\,\frac{1}{2n} = \frac{1}{2x}\,\frac{1}{n}$; since finitely many summands of a series don't affect convergence/divergence, the series diverges just like $\sum_{n=1}^{\infty} \frac{1}{2x}\frac{1}{n} = \frac{1}{2x}\sum_{n=1}^{\infty} \frac{1}{n}$.)

Solution cont'd

(3) This is the most advanced part, but please be warned that this example is artificial and included solely for teaching purposes: In reality nobody would want to replace the rather straightforward proof in (1) that $f_n \to 0$ uniformly by a complicated proof, which doesn't even yield the limit function.

First we make the correspondence between (function) sequences and (function) series precise: As usual, with a series $\sum_{n=0}^{\infty} f_n$ we associate the sequence (F_n) of partial sums $F_n = \sum_{k=0}^{n} f_k$, $n \geq 0$, and you should recall from Calculus II that convergence/sum of $\sum_{n=0}^{\infty} f_n$ means convergence/limit of (F_n).

Conversely, given a sequence $(f_n)_{n\geq 0}$ we can write

$$\begin{aligned} f_n &= f_n - f_{n-1} + f_{n-1} - f_{n-2} + \cdots + f_1 - f_0 + f_0 \\ &= g_n + g_{n-1} + \cdots + g_1 + g_0 \end{aligned}$$

with $g_0 = f_0$ and $g_n = f_n - f_{n-1}$ for $n \geq 1$. The sequence (f_n) is then the sequence of partial sums of the series $\sum_{n=0}^{\infty} g_n$.

This correspondence between sequences and series is evidently one-to-one.

Solution cont'd

In our case we obtain $g_0(x) = f_0(x) = x$ and

$$g_n(x) = \frac{x}{1+nx^2} - \frac{x}{1+(n-1)x^2} = \frac{-x^3}{(1+nx^2)(1+(n-1)x^2)}$$

for $n \geq 1$. By construction, the n-th partial sum of the function series

$$x + \sum_{n=1}^{\infty} \frac{-x^3}{(1+nx^2)(1+(n-1)x^2)}$$

is then equal to $f_n(x) = \frac{x}{1+nx^2}$.

Now we will prove, using the Weierstrass test, that the function series, and hence the sequence (f_n), converges uniformly on $\mathbb{R}$. For this we must solve two problems.

Problem 1: The first summand of the series, viz. $g_0(x) = x$ is unbounded.

Hence we can apply the Weierstrass test only to the function series $\sum_{n=1}^{\infty} g_n$. But it is sufficient to show uniform convergence of this series, because adding finitely many summands to a function series doesn't affect uniform convergence.

Solution cont'd

For the sequence of partial sums this means "add the same function to every sequence term" and clearly has no effect on uniform (or point-wise) convergence. (In fact, since we already know that $f_n \to 0$ uniformly, it must be that $\sum_{n=1}^{\infty} g_n$ converges uniformly to $-x$.)

Problem 2: Find a suitable uniform bound M_n for $|g_n(x)|$.

Here we can argue as follows: For small x (maybe "small in comparison with n") we don't lose too much if we use $\left|1 + n\,x^2\right| \geq 1$, and similarly for the 2nd factor in the denominator. W.l.o.g restricting to positive x, this gives

$$|g_n(x)| = \frac{x^3}{(1 + n\,x^2)\,(1 + (n-1)x^2)} \leq x^3 \leq \frac{1}{n\sqrt{n}},$$

provided that $x \leq \frac{1}{\sqrt{n}}$. (If you can't see the idea behind the last estimate, note at least that $M_n = \frac{1}{n\sqrt{n}}$ would work in the Weierstrass test because the series $\sum_{n=1}^{\infty} \frac{1}{n\sqrt{n}} = \sum_{n=1}^{\infty} \frac{1}{n^{3/2}}$ converges.)

Solution cont'd

It remains to estimate $|g_n(x)|$ for $x > \frac{1}{\sqrt{n}}$. Here we can use

$$\frac{x^3}{(1+nx^2)(1+(n-1)x^2)} \le \frac{x^3}{nx^2(n-1)x^2} = \frac{1}{n(n-1)x} < \frac{1}{(n-1)\sqrt{n}}.$$

For $n = 1$ this estimate doesn't work but, as remarked before, we may as well consider the series $\sum_{n=2}^{\infty} g_n$ and prove its uniform convergence.

In all we have shown

$$|g_n(x)| \le \frac{1}{(n-1)\sqrt{n}} \quad \text{for } n \ge 2 \text{ and all } x \in \mathbb{R}.$$

Applying the Weierstrass test to the series $\sum_{n=2}^{\infty} g_n$ with constants $M_n = \frac{1}{(n-1)\sqrt{n}}$ then finishes the proof. (If you have doubts that the series $\sum_{n=2}^{\infty} \frac{1}{(n-1)\sqrt{n}}$ converges, observe that $(n-1)\sqrt{n} \ge n\sqrt{n}/2$ and apply the comparison test with $\sum_{n=2}^{\infty} \frac{2}{n\sqrt{n}}$ as upper bound.)

Further Tests for Uniform Convergence

Theorem

Let (f_n) be a monotonically decreasing sequence of real-valued functions on D (i.e., $f_1(x) \geq f_2(x) \geq f_3(x) \geq \cdots$ for $x \in D$) and (g_n) a sequence of complex-valued functions on D. The function series $\sum_{n=1}^{\infty} f_n g_n$ converges uniformly on D if one of the following two criteria is satisfied:

1. DIRICHLET*'s test for uniform convergence*
   (f_n) converges to 0 uniformly on D and the series $\sum_{n=1}^{\infty} g_n$ (i.e., its partial sums $G_n = g_1 + \cdots + g_n$) is uniformly bounded on D.
2. ABEL*'s test for uniform convergence*
   (f_n) is uniformly bounded on D and the series $\sum_{n=1}^{\infty} g_n$ converges uniformly on D.

The domain D in the theorem is completely arbitrary (i.e., any set).

Dirichlet's test in particular includes the case where f_n is constant, i.e., (f_n) is ordinary sequence of real numbers satisfying $f_n \downarrow 0$.

Cauchy's Test for Uniform Convergence

The mother of all such tests

Before proving the theorem, we state the analogue of Cauchy's convergence test for uniformly convergent function sequences/series, which was behind the scene also in Weierstrass' test.

Lemma

1. *A sequence (f_n) of (real-/complex-/vector-valued) functions on D converges uniformly iff for every $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that $|f_m(x) - f_n(x)| < \epsilon$ for all $m, n > N$ and all $x \in D$.*
2. *A series $\sum_{n=1}^{\infty} f_n$ of functions on D converges uniformly iff for every $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that $\left|\sum_{k=m}^{n} f_k(x)\right| < \epsilon$ for all $n \geq m > N$ and all $x \in D$.*

The proof is more or less the same as the earlier proof (discussed in Calculus III) that a sequence of real or complex numbers (or vectors) converges iff it is a Cauchy sequence, except that now all responses N must be uniform. We omit the details.

Proof of the theorem.

In both cases, given $\epsilon > 0$, we must find a response N such that $\left|\sum_{k=m}^{n} f_k(x)g_k(x)\right| < \epsilon$ for $n \geq m > N$. This can be done with the help of Abel summation:

$$\sum_{k=m}^{n} f_k(x)g_k(x) = \sum_{k=m}^{n} f_k(x)\left[G_k(x) - G_{k-1}(x)\right]$$

$$= -f_m(x)G_{m-1}(x) + f_n(x)G_n(x) + \sum_{k=m}^{n-1}\left[f_k(x) - f_{k+1}(x)\right]G_k(x)$$

Case 1 (Dirichlet): By assumption, there exists $M > 0$ such that $|G_k(x)| \leq M$ for all $k \in \mathbb{N}$ and $x \in D$, and since $f_n(x) \downarrow 0$ we must have $f_n(x) \geq 0$. This allows us to estimate as follows:

$$\left|\sum_{k=m}^{n} f_k(x)g_k(x)\right| \leq f_m(x)M + f_n(x)M + \sum_{k=m}^{n-1}\left[f_k(x) - f_{k+1}(x)\right]M$$

$$= 2M\,f_m(x).$$ (The sum "telescopes".)

Since $f_n \downarrow 0$ uniformly, given $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that $f_m(x) < \epsilon/(2M)$ for $m > N$ and $x \in D$, showing that $\sum_{n=1}^{\infty} f_n g_n$ satisfies the assumption of the Cauchy test for uniform convergence.

Proof cont'd.

Case 2 (Abel): Here the key observation is that in the expression obtained for $\sum_{k=m}^{n} f_k(x)g_k(x)$ by Abel summation the coefficient sum of the functions $G_k(x)$ is

$$f_n(x) - f_m(x) + f_m(x) - f_{m+1}(x) + \cdots + f_{n-1}(x) - f_n(x) = 0.$$

$\Longrightarrow$ We can add subtract $G(x) = \lim_{n\to\infty} G_n(x)$ from every summand without affecting the sum, i.e., we have (suppressing arguments)

$$\sum_{k=m}^{n} f_k g_k = -f_m(G_{m-1} - G) + f_n(G_n - G) + \sum_{k=m}^{n-1}(f_k - f_{k+1})(G_k - G).$$

Since $G_n \to G$ uniformly, given $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that $|G_n - G| < \epsilon$ for $n > N$. For $n \geq m > N + 1$ we then obtain

$$\left|\sum_{k=m}^{n} f_k g_k\right| \leq |f_m|\,\epsilon + |f_n|\,\epsilon + \sum_{k=m}^{n-1}(f_k - f_{k+1})\epsilon = (|f_m| + |f_n| + f_m - f_n)\,\epsilon.$$

If M is a uniform bound for f_n, i.e., $|f_n(x)| \leq M$ for all $n \in \mathbb{N}$ and $x \in D$, then $\left|\sum_{k=m}^{n} f_k g_k\right| \leq 4M\epsilon$ for $n \geq m > N + 1$, so that again the Cauchy test for uniform convergence can be applied. □

Note

Uniform convergence of the series $\sum_{n=1}^{\infty} f_n g_n$ is not affected by deleting a finite number of summands from it. This can be helpful if some of the functions f_n or g_n are unbounded. For example, if g_1 is unbounded and all of $g_2, g_3, \ldots$ are bounded then all partial sums $G_n = g_1 + \cdots + g_n$ are unbounded as well, and hence Dirichlet's test cannot be applied directly, but nevertheless it may be possible to obtain a uniform bound for $G'_n = g_2 + \cdots + g_n$ and apply it to the series $\sum_{n=2}^{\infty} f_n g_n$.

As discussed earlier, Dirichlet's test gives the uniform convergence of $\sum_{n=1}^{\infty} z^n/n$ on the regions D_r, $0 < r < 1$. Here $f_n(z) = 1/n$, $g_n(z) = z^n$, and the key observation is that $G_n(z) = z + z^2 + \cdots + z^n$ is uniformly bounded on D_r.

As an application of Abel's Test we prove the following

Theorem (Abel's Limit Theorem)

Suppose $\sum_{n=0}^{\infty} a_n(z-a)^n$ *has radius of convergence* $0 < R < \infty$. *If the power series converges for a point* $z_1 = a + R\,\mathrm{e}^{\mathrm{i}\phi}$ *on the boundary of its disk of convergence, it converges uniformly on the line segment* $[a, z_1] = \{a + r\,\mathrm{e}^{\mathrm{i}\phi}; 0 \le r \le R\}$, *and hence represents a continuous function on* $[a, z_1]$.

That the so-defined function f is continuous in z_1 means

$$\lim_{r\uparrow R} f(a + r\,\mathrm{e}^{\mathrm{i}\phi}) = \lim_{r\uparrow R} \sum_{n=0}^{\infty} a_n r^n \mathrm{e}^{\mathrm{i}n\phi} = \sum_{n=0}^{\infty} a_n R^n \mathrm{e}^{\mathrm{i}n\phi} = f(a + R\,\mathrm{e}^{\mathrm{i}\phi}),$$

and explains the name "limit theorem".

Proof.

Writing $f(z) = \sum_{n=0}^{\infty} a_n (z-a)^n$, we have

$$f(a + r\,\mathrm{e}^{\mathrm{i}\phi}) = \sum_{n=0}^{\infty} a_n r^n \mathrm{e}^{\mathrm{i}n\phi} = \sum_{n=0}^{\infty} \frac{r^n}{R^n}\, a_n R^n \mathrm{e}^{\mathrm{i}n\phi}.$$

Now define $f_n(r) = r^n/R^n = (r/R)^n$, $g_n(r) = a_n R^n \mathrm{e}^{\mathrm{i}n\phi}$ for $n \in \mathbb{N}_0$ and $r \in [0, R]$. Then $0 \le f_{n+1}(r) \le f_n(r) \le 1$ for all n, so that f_n has the properties required in Abel's test for uniform convergence.

The series $\sum_{n=0} g_n$ converges uniformly on $[0, R]$, since it converges by assumption and doesn't depend on r.

$\Longrightarrow$ Abel's test can be applied and yields the uniform convergence of $\sum_{n=0}^{\infty} f_n g_n$ on $[0, R]$, i.e., the uniform convergence of $\sum_{n=0}^{\infty} a_n (z-a)^n$ on $[a, z_1]$. □

Example

The binomial series

$$\sum_{n=0}^{\infty}\binom{s}{n}z^n=\sum_{n=0}^{\infty}\frac{s(s-1)\cdots(s-n+1)}{1\cdot 2\cdots n}\,z^n,\quad s\notin\{0,1,2,\dots\},$$

has radius of convergence $R=1$ (by the ratio test) and represents the function $(1+z)^s=\mathrm{e}^{s\log(1+z)}$ in $\mathrm{B}_1(0)$; cf. Homework 4, Exercise H25. (Here, as usual, $\log$ denotes the principal branch of the complex logarithm. For $s=0,1,2,\dots$ the series terminates, and the identity reduces to the ordinary binomial theorem $(1+z)^s=\sum_{n=0}^{s}\binom{s}{n}z^n$.

Claim: For $s>-1$ the series $\sum_{n=0}^{\infty}\binom{s}{n}$ satisfies the assumptions of the alternating series test, and hence converges.
Proof: Since $s+1>0$ and

$$\binom{s}{n}=-\frac{n-s-1}{n}\binom{s}{n-1},$$

for large n the sequence $a_n=\binom{s}{n}$ will be alternating in sign and $|a_n|<|a_{n-1}|$.

Example (cont'd)

In order to show $\binom{s}{n} \to 0$ for $n \to \infty$, we rewrite the binomial coefficient as

$$\binom{s}{n} = \pm \frac{s}{n}\left(1-\frac{s}{1}\right)\left(1-\frac{s}{2}\right)\cdots\left(1-\frac{s}{n-1}\right)$$

For $s > 0$ we have $0 < 1 - s/k < 1$ except for $k = 1, 2, \ldots, \lfloor s \rfloor$, and hence for $n > \lfloor s \rfloor$ the estimate $\left|\binom{s}{n}\right| \le sP/n$, where $P = \prod_{k=1}^{\lfloor s \rfloor}(s/k - 1)$. This shows $\binom{s}{n} \to 0$ for $n \to \infty$.

For $-1 < s < 0$ we have

$$\ln\left|\binom{s}{n}\right| = \ln(-s) - \ln(n) + \sum_{k=1}^{n-1} \ln\left(1-\frac{s}{k}\right)$$

$$\le \ln(-s) - \ln(n) - \sum_{k=1}^{n-1}\frac{s}{k} = \ln(-s) - \ln(n) - s\sum_{k=1}^{n-1}\frac{1}{k} \to -\infty,$$

since $\sum_{k=1}^{n-1} 1/k = \ln(n) + \mathrm{O}(1)$ and $-s < 1$. This shows $\binom{s}{n} \to 0$ for $n \to \infty$ also in this case and completes the proof of our claim.

Example (cont'd)

Now Abel's Limit Theorem gives that $z \mapsto \sum_{n=0}^{\infty} \binom{s}{n} z^n$ defines a continuous function on $(-1, 1]$ for $s > 1$ and hence represents $z \mapsto (1+z)^s$ also for $z = 1$.

$$\Longrightarrow \sum_{n=0}^{\infty} \binom{s}{n} = 2^s \quad \text{for } s > -1.$$

For $s = -1/2$ we have $\binom{-1/2}{n} = (-1)\,\frac{1\cdot 3\cdot 5\cdots(2n-1)}{2\cdot 4\cdot 6\cdots 2n}$ and obtain the series evaluation

$$1 - \frac{1}{2} + \frac{1\cdot 3}{2\cdot 4} - \frac{1\cdot 3\cdot 5}{2\cdot 4\cdot 6} + \frac{1\cdot 3\cdot 5\cdot 7}{2\cdot 4\cdot 6\cdot 8} \mp \cdots = \frac{\sqrt{2}}{2}.$$

Similarly, one can show that for $s > 0$ the binomial series converges at $z = -1$ and hence represents $z \mapsto (1+z)^s$ also for $z = -1$.

$$\Longrightarrow \sum_{n=0}^{\infty} (-1)^n \binom{s}{n} = 0 \quad \text{for } s > 0.$$

The Multivariable Case

Usually uniform convergence is covered in textbooks as part of Calculus I or II, when the differential calculus of functions of several variables is not yet available. For this reason the main theorems about uniform convergence are usually stated for one-variable functions, as we have done.

But, of course, these theorems have multivariable generalizations, which are no less important. We consider only the case of the Differentiation Theorem.

Theorem (differentiation, multivariable case)

Suppose $f_k : D \to \mathbb{R}$, $D \subseteq \mathbb{R}^n$ ($k \in \mathbb{N}$) *are* C^1*-functions,* $(f_k)_{k\in\mathbb{N}}$ *converges point-wise on* D*, and the* n *sequences of partial derivatives* $(\partial f_k/\partial x_i)_{k\in\mathbb{N}}$, $1 \le i \le n$, *converge uniformly on* D*. Then* $f(\mathbf{x}) = \lim_{k\to\infty} f_k(\mathbf{x})$, $\mathbf{x} \in D$, *is a* C^1*-function as well and satisfies*

$$\frac{\partial f}{\partial x_i}(\mathbf{x}) = \lim_{k\to\infty} \frac{\partial f_k}{\partial x_i}(\mathbf{x}) \quad \text{for } 1 \le i \le n \text{ and } \mathbf{x} \in D. \qquad (\star)$$

Proof.

Recall from Calculus III that a function $g\colon D \to \mathbb{R}$ is C^1 iff the partial derivatives $\partial g/\partial x_i$ exist on D and are continuous as multivariable functions.

First consider a sequence of partial derivatives, say $(\partial f_k/\partial x_1)_{k\in\mathbb{N}}$. Since $x_1 \mapsto \frac{\partial f_k}{\partial x_1}(x_1, x_2, \ldots, x_n)$ is the ordinary derivative of $x_1 \mapsto f_k(x_1, x_2, \ldots, x_n)$ and since the uniform convergence of $(\partial f_k/\partial x_1)_{k\in\mathbb{N}}$ on D implies the uniform convergence of the one-variable functions $x_1 \mapsto \frac{\partial f_k}{\partial x_1}(x_1, x_2, \ldots, x_n)$ on the set of all $x \in \mathbb{R}$ for which there exists an $(x_1, \ldots, x_n) \in D$ such that $x = x_1$, we can apply the one-variable Differentation Theorem to conclude that $\partial f/\partial x_1$ exists on D and

$$\lim_{k\to\infty} \frac{\partial f_k}{\partial x_1}(x_1, \ldots, x_n) = \frac{\partial f}{\partial x_1}(x_1, \ldots, x_n) \quad \text{for } (x_1, \ldots, x_n) \in D.$$

This shows that f is partially differentiable and satisfies $(\star)$. The proof is then finished by applying the Continuity Theorem (whose generalization to n-variable functions is straightforward) to the sequence $(\partial f_k/\partial x_1)_{k\in\mathbb{N}}$, say, which yields the continuity of $\frac{\partial f}{\partial x_1}$ as a multivariable function. □

Note

Since differentiability is a local property, the conclusions of the Differentiation Theorem remain valid under the weaker assumption that every point $\mathbf{x} \in D$ has a neighborhood $D_{\mathbf{x}}$ on which the sequences $(\partial f_k/\partial x_i)_{k\in\mathbb{N}}$, $1 \leq i \leq n$, converge uniformly. This is also true for the Continuity Theorem (since continuity is a local property as well) and for the Integration Theorem (since the interval of integration is compact and locally uniform convergence on D implies uniform convergence on every compact subset of D by the Heine-Borel covering property; cf. Calculus III).

Application

Suppose $c_0, c_1, c_2, \ldots$ is a sequence of real numbers which grows at most polynomially, i.e., there exists $d \in \mathbb{Z}^+$ such that $c_k = \mathrm{O}(k^d)$ for $k \to \infty$. We show that

$$f(x,y) = \sum_{k=0}^{\infty} c_k \mathrm{e}^{-ky} \cos(kx), \quad (x,y) \in \mathbb{R} \times \mathbb{R}^+,$$

solves Laplace's Equation $\partial^2 f/\partial x^2 + \partial^2 f/\partial y^2 = 0$.

First let us assume that f is well-defined, is a C^2-function, and that partial differentiation can be done termwise.

$$\begin{aligned}
\Longrightarrow \quad \frac{\partial f}{\partial x}(x,y) &= \sum\nolimits_{k=0}^{\infty} -k c_k \mathrm{e}^{-ky} \sin(kx), \\
\frac{\partial^2 f}{\partial x^2}(x,y) &= \sum\nolimits_{k=0}^{\infty} -k^2 c_k \mathrm{e}^{-ky} \cos(kx), \\
\frac{\partial f}{\partial y}(x,y) &= \sum\nolimits_{k=0}^{\infty} -k c_k \mathrm{e}^{-ky} \cos(kx), \\
\frac{\partial^2 f}{\partial y^2}(x,y) &= \sum\nolimits_{k=0}^{\infty} k^2 c_k \mathrm{e}^{-ky} \cos(kx),
\end{aligned}$$

and clearly $\frac{\partial^2 f}{\partial x^2}(x,y) + \frac{\partial^2 f}{\partial y^2}(x,y) = 0$.

Application cont'd

In order to justify the preceding computation, it suffices to show that all series involved (including that defining f) converge uniformly on some neighborhood of a given point (x_0, y_0) in the upper half-plane $\mathbb{R} \times \mathbb{R}^+$. (Be sure to check in detail how this and the Differentiation Theorem imply all assumptions made.)
We can take the neighborhoods as

$$H_\delta = \{(x, y) \in \mathbb{R}^2; y \geq \delta\}, \quad \delta > 0.$$

For the coefficients of the series representing $\partial^2 f/\partial x^2$ we have

$$\left|-k^2 c_k \mathrm{e}^{-ky} \cos(kx)\right| \leq k^2 \left|c_k\right| \mathrm{e}^{-ky} \leq M k^{d+2} \mathrm{e}^{-k\delta}$$

for $(x, y) \in H_\delta$ and k sufficiently large, where M is some constant.
The series $\sum_{k=0}^{\infty} k^{d+2} \mathrm{e}^{-k\delta}$ converges, because the rapid growth of $x \mapsto \mathrm{e}^x$ implies that $\mathrm{e}^{-k\delta} \leq k^{-d-4}$, and hence $k^{d+2} \mathrm{e}^{-k\delta} \leq 1/k^2$, for sufficiently large k.
$\Longrightarrow$ We can apply Weierstrass's Criterion to conclude the uniform convergence of $\sum_{k=0}^{\infty} -k^2 c_k \mathrm{e}^{-ky} \cos(kx)$ on H_δ.
The other series are treated similarly.

Uniform Convergence of Improper Parameter Integrals

So far we have considered uniform convergence of function sequences/series, the "discrete case" so-to-speak. The continuous analog of (function) series are improper (parameter) integrals, and accordingly it also makes sense to speak of uniform convergence of parameter integrals:

Definition

Suppose f is a real-valued function with domain $D \times [0,\infty)$ and such that $\int_0^R f(x,t)\,\mathrm{d}t$ is defined for every $R \in [0,\infty)$ and $x \in D$.

1. $\int_0^\infty f(x,t)\,\mathrm{d}t$ is said to converge point-wise on D if $\lim_{R\to\infty} \int_0^R f(x,t)\,\mathrm{d}t$ exists for every $x \in D$. If this is the case, $F(x) := \int_0^\infty f(x,t)\,\mathrm{d}t := \lim_{R\to\infty} \int_0^R f(x,t)\,\mathrm{d}t$ defines a real-valued function on D ("limit function").
2. $\int_0^\infty f(x,t)\,\mathrm{d}t$ is said to converge uniformly on D if it converges point-wise and for every $\epsilon > 0$ there exists a "uniform" response $R_0 \in [0,\infty)$ such that $\left|F(x) - \int_0^R f(x,t)\,\mathrm{d}t\right| = \left|\int_R^\infty f(x,t)\,\mathrm{d}t\right| < \epsilon$ for all $R > R_0$ and all $x \in D$.

Notes

- The definition is easily extended to improper integrals with other domains of integration, such as $[a, \infty)$, $(-\infty, b]$, $(-\infty, \infty)$, (a, b), etc. The subsequent discussion applies mutatis mutandis to all these cases.
- Uniform convergence makes sense for any limit involving a further parameter, e.g., under the assumptions of the definition it makes sense to define "$f(x,t) \to F(x)$ uniformly for $t \to \infty$" if $\lim_{t\to\infty} f(x,t) = F(x)$ for every $x \in D$ and in a proof of this responses $R_0 = R_0(\epsilon)$ can be found that do not depend on x.
- The theory we shall now develop overlaps with that of the Lebesgue integral, but is not contained in it, because it also applies to improper integrals that don't converge absolutely, e.g., $\int_0^\infty \sin(t)/t \, \mathrm{d}t$.

Our first goal is to prove analogues of the Continuity and Differentiation Theorems for proper parameter integrals with continuous integrands. These are special cases of theorems for the Lebesgue integral stated earlier in Calculus III. "Elementary" proofs that are independent of the Lebesgue theory will be given.

Lemma (continuity)

Suppose $I \subseteq \mathbb{R}$ is an interval and $f\colon I \times [a,b] \to \mathbb{R}$ is a continuous two-variable function. Then $F\colon I \to \mathbb{R}$, $x \mapsto \int_a^b f(x,t)\,\mathrm{d}t$ is continuous.

Proof.

Since all functions $t \mapsto f(x,t)$, $x \in I$, are continuous, existence of all (Riemann) integrals involved is trivial. For $x, x_0 \in I$ we have

$$F(x) - F(x_0) = \int_a^b [f(x,t) - f(x_0,t)]\,\mathrm{d}t,$$

and, given $\epsilon > 0$, need to find a response δ such $|f(x,t) - f(x_0,t)| < \epsilon$ for $x \in [x_0 - \delta, x_0 + \delta] \cap I$ and all $t \in [a,b]$. (For such x we then have $|F(x) - F(x_0)| \le \epsilon(b-a)$, showing that F is continuous in x_0.)

Now we use the uniform continuity of f on compact rectangles $K = [\alpha,\beta] \times [a,b]$ with $[\alpha,\beta] \subseteq I$. If x_0 is an inner point of I, we can choose $\alpha < x < \beta$, and if x_0 is the left end point of I, say, we can choose $\alpha = x_0 < \beta$.

Given $\epsilon > 0$, there exists $\delta > 0$ such that $|f(x_1,t_1) - f(x_2,t_2)| < \epsilon$ if $(x_1,t_1), (x_2,t_2) \in K$ and $|x_1 - x_2| < \delta$, $|t_1 - t_2| < \delta$. Specializing to $x_1 = x$, $x_2 = x_0$, $t_1 = t_2 = t$ shows that this δ can serve as the desired response. □

Lemma (differentiation)

Suppose $I \subseteq \mathbb{R}$ is an interval, $f : I \times [a,b] \to \mathbb{R}$ is a continuous two-variable function, and the partial derivative $f_x = \frac{\partial f}{\partial x} : I \times [a,b] \to \mathbb{R}$ exists and is a continuous two-variable function as well. Then $F : I \to \mathbb{R}$, $x \mapsto \int_a^b f(x,t)\,\mathrm{d}t$ is differentiable with

$$F'(x) = \int_a^b f_x(x,t)\,\mathrm{d}t,$$

i.e., we can differentiate F under the integral sign.

Proof.

For $x_0 \in I$ and $h \neq 0$ such that $x_0 + h \in I$ we have

$$\frac{F(x_0+h)-F(x_0)}{h} - \int_a^b f_x(x_0,t)\,\mathrm{d}t = \int_a^b \left[\frac{f(x_0+h,t)-f(x_0,t)}{h} - f_x(x_0,t)\right]\mathrm{d}t,$$

$$\left|\frac{F(x_0+h)-F(x_0)}{h} - \int_a^b f_x(x_0,t)\,\mathrm{d}t\right| \leq \int_a^b \left|\frac{f(x_0+h,t)-f(x_0,t)}{h} - f_x(x_0,t)\right|\mathrm{d}t,$$

and all integrals involved exist because of the continuity assumptions on f and f_x.

Proof cont'd.

Using the Mean Value Theorem, the last integrand can also be expressed as

$$\left|\frac{f(x_0+h,t)-f(x_0,t)}{h}-f_x(x_0,t)\right| = \left|f_x(\xi_{h,t},t)-f_x(x_0,t)\right|$$

with $\xi_{h,t}$ between x_0 and x_0+h. (Considering x_0 as fixed, there is no dependence of $\xi_{h,t}$ on x_0.)

Now the proof can be finished as in the Continuity Lemma, this time using the uniform continuity of f_x on compact rectangles $K=[\alpha,\beta]\times[a,b]$: If $\delta=\delta(\epsilon)$ is such that $|f_x(x_1,t_1)-f_x(x_2,t_2)|<\epsilon$ for all (x_1,t_1), $(x_2,t_2)\in K$ with $|x_1-x_2|<\delta$, $|t_1-t_2|<\delta$, the previous estimates imply that

$$\left|\frac{F(x_0+h)-F(x_0)}{h}-\int_a^b f_x(x_0,t)\,\mathrm{d}t\right| \le \epsilon(b-a)$$

for all nonzero $h\in(-\delta,\delta)$ such that $x_0+h\in I$.

$\Longrightarrow \lim_{h\to 0}\frac{F(x_0+h)-F(x_0)}{h}=\int_a^b f_x(x_0,t)\,\mathrm{d}t$, as desired. □

Now we are ready to state and prove analogues of our Continuity and Differentiation Theorems for uniformly convergent improper parameter integrals.

Theorem (continuity)

Suppose $I \subseteq \mathbb{R}$ is an interval, $f\colon I \times [a,\infty) \to \mathbb{R}$ is a continuous two-variable function, and $\int_a^\infty f(x,t)\,\mathrm{d}t$ converges uniformly on I. Then $F\colon I \to \mathbb{R}$, $x \mapsto \int_a^\infty f(x,t)\,\mathrm{d}t$ is continuous.

Proof.

For $n \in \mathbb{N}$ define $F_n\colon I \to \mathbb{R}$ by $F_n(x) = \int_a^{a+n} f(x,t)\,\mathrm{d}t$. The functions F_n are continuous by the Continuity Lemma, and converge uniformly to $F(x) = \int_a^\infty f(x,t)\,\mathrm{d}t$, $x \in I$. (If R_0 is such that $\left|\int_R^\infty f(x,t)\,\mathrm{d}t\right| < \epsilon$ for all $R > R_0$ then $|F(x) - F_n(x)| = \int_{a+n}^\infty f(x,t)\,\mathrm{d}t < \epsilon$ for all $n > R_0 - a$, so that we can take $N = \lceil R_0 - a \rceil$ as corresponding response.)
$\Longrightarrow$ By the Continuity Theorem for function sequences, F is continuous. $\square$

Theorem (differentiation)

Suppose $I \subseteq \mathbb{R}$ is an interval, $f : I \times [a, \infty) \to \mathbb{R}$ is a continuous two-variable function, the partial derivative $f_x = \frac{\partial f}{\partial x} : I \times [a, \infty) \to \mathbb{R}$ exists and is a continuous two-variable function as well, $\int_a^\infty f(x,t)\,\mathrm{d}t$ converges point-wise on I, and $\int_a^\infty f_x(x,t)\,\mathrm{d}t$ converges uniformly on I. Then $F : I \to \mathbb{R}$, $x \mapsto \int_a^\infty f(x,t)\,\mathrm{d}t$ is differentiable with

$$F'(x) = \int_a^\infty f_x(x,t)\,\mathrm{d}t,$$

i.e., we can differentiate F under the integral sign.

Proof.

As before we set $F_n(x) = \int_a^{a+n} f(x,t)\,\mathrm{d}t$ for $x \in I$, which converges point-wise on I to F by assumption. The Differentiation Lemma gives that F_n is differentiable with $F_n'(x) = \int_a^{a+n} f_x(x,t)\,\mathrm{d}t$, and the uniform convergence of $\int_a^\infty f_x(x,t)\,\mathrm{d}t$ that (F_n') converges uniformly on I to $x \mapsto \int_a^\infty f_x(x,t)\,\mathrm{d}t$. Hence the Differentiation Theorem for function sequences can be used to finish the proof. □

In a way the deepest result used in the foregoing "elementary" proofs is that continuous functions on compact subsets of $\mathbb{R}^n$, here rectangles $K = [\alpha, \beta] \times [a, b] \subset \mathbb{R}^2$, are uniformly continuous. In the Calculus III lecture slides this was shown for the 1-dimensional case $K = [a, b]$ on two different occasions with two different proofs, one based on the Bolzano-Weierstrass Theorem and the other on the Heine-Borel covering property.

Exercise

Translate in detail one of the two proofs mentioned above into the present 2-dimensional setting.

Exercise

Suppose $U \subseteq \mathbb{R}^2$ is open and contains a compact segment $\{(0, y); a \leq y \leq b\}$ of the y-axis. Show that there exists $\delta > 0$ such that $[-\delta, \delta] \times [a, b] \subseteq U$.

A proof of the Continuity Lemma can also be based on this property.

Criteria for Uniform Convergence

The tests for uniform convergence of function series have continuous analogues, which we now discuss. We state these in the original setting for a function $f\colon D \times [0,\infty) \to \mathbb{R}$ and tacitly assume that the Riemann integrals $\int_0^R f(x,t)\,\mathrm{d}t$, and hence also $\int_0^R |f(x,t)|\,\mathrm{d}t$, exist for all $R \in [0,\infty)$ and $x \in D$. (In particular this is the case if D is an interval in $\mathbb{R}$ and f is continuous as a two-variable function.) As usual, the Cauchy test is the basis for all others.

Cauchy test

$\int_0^\infty f(x,t)\,\mathrm{d}t$ converges uniformly on D iff for every $\epsilon > 0$ there exists $R_0 > 0$ such that $\left|\int_R^{R'} f(x,t)\,\mathrm{d}t\right| < \epsilon$ for all $R' > R > R_0$ and $x \in D$.

Proof.

We only prove "$\Longleftarrow$", which is more difficult. Define a sequence of functions $F_n\colon D \to \mathbb{R}$ by $F_n(x) = \int_0^n f(x,t)\,\mathrm{d}t$, $n = 0, 1, 2, \ldots$. Under the given assumption F_n clearly satisfies the Cauchy test for uniform convergence of function sequences, and hence converges uniformly to a function $F\colon D \to \mathbb{R}$.

Proof cont'd.

Further, given $\epsilon > 0$, let R_0 be the corresponding response as stated in the Cauchy test. For $n > R > R_0$ we then have

$$\left| F_n(x) - \int_0^R f(x,t)\,\mathrm{d}t \right| = \left| \int_R^n f(x,t)\,\mathrm{d}t \right| < \epsilon \quad \text{for } x \in D.$$

Letting $n \to \infty$ we obtain $\left| F(x) - \int_0^R f(x,t)\,\mathrm{d}t \right| \le \epsilon$ for all $R > R_0$ and $x \in D$, which shows that $\int_0^\infty f(x,t)\,\mathrm{d}t$ converges uniformly on D to $F(x)$. □

Weierstrass's Test

Suppose there exists a function $\Phi\colon [0,\infty) \to \mathbb{R}$ such that $|f(x,t)| \le \Phi(t)$ for all $(x,t) \in D \times [0,\infty)$ and $\int_0^\infty \Phi(t)\,\mathrm{d}t$ converges in $\mathbb{R}$. Then $\int_0^\infty f(x,t)\,\mathrm{d}t$ converges uniformly and (absolutely) on D.

Since necessarily $\Phi \ge 0$, this is actually a special case of Lebesgue's Dominated Convergence Theorem, but it has a simple proof based on the Cauchy test: One needs only observe that $\left| \int_R^{R'} f(x,t)\,\mathrm{d}t \right| \le \int_R^{R'} |f(x,t)|\,\mathrm{d}t \le \int_R^{R'} \Phi(t)\,\mathrm{d}t$ independently of $x \in D$, and use the (reverse) Cauchy test for the ordinary improper integral $\int_0^\infty \Phi(t)\,\mathrm{d}t$.

Example

The Gamma function

$$\Gamma(x) = \int_0^\infty t^{x-1}\mathrm{e}^{-t}\,\mathrm{d}t, \quad x \in (0,\infty)$$

can be discussed in the present framework without recourse to Lebesgue integration theory. Since for $x < 1$ the integral is improper on both ends, an appropriate definition is

$$\Gamma(x) = \int_0^\infty t^{x-1}\mathrm{e}^{-t}\,\mathrm{d}t := \int_0^1 t^{x-1}\mathrm{e}^{-t}\,\mathrm{d}t + \int_1^\infty t^{x-1}\mathrm{e}^{-t}\,\mathrm{d}t$$

reducing the discussion to the one-sided improper integrals

$$\Gamma_0(x) = \int_0^1 t^{x-1}\mathrm{e}^{-t}\,\mathrm{d}t = \lim_{r\downarrow 0}\int_r^1 t^{x-1}\mathrm{e}^{-t}\,\mathrm{d}t,$$

$$\Gamma_1(x) = \int_1^\infty t^{x-1}\mathrm{e}^{-t}\,\mathrm{d}t = \lim_{R\to\infty}\int_1^R t^{x-1}\mathrm{e}^{-t}\,\mathrm{d}t\,.$$

Provided we are allowed to differentiate k-times under the integral sign, the corresponding derivatives are

Example (cont'd)

$$\Gamma_0^{(k)}(x) = \int_0^1 (\ln t)^k t^{x-1} e^{-t}\,dt, \quad \Gamma_1^{(k)}(x) = \int_1^\infty (\ln t)^k t^{x-1} e^{-t}\,dt.$$

To justify the differentiations, it suffices to show uniform convergence of these integrals on intervals of the form $[a, b]$ with $0 < a < b$. This can be done with the Weierstrass test:

$$\begin{aligned} \left|(\ln t)^k t^{x-1} e^{-t}\right| &\le |\ln t|^k\, t^{a-1}, && t \in (0,1],\ x \ge a, \\ \left|(\ln t)^k t^{x-1} e^{-t}\right| &\le (\ln t)^k t^{b-1} e^{-t}, && t \in [1,\infty)\ x \le b. \end{aligned}$$

The integrals $\int_0^1 |\ln t|^k\, t^{a-1}\,dt$, $\int_1^\infty (\ln t)^k t^{b-1} e^{-t}\,dt$ exist, as is easily shown using the growth behavior of log, exp. Thus the Weierstrass can be applied as claimed.

Putting things together, we obtain that Γ has derivatives of all orders, given by

$$\Gamma^{(k)}(x) = \Gamma_0^{(k)}(x) + \Gamma_1^{(k)}(x) = \int_0^\infty (\ln t)^k t^{x-1} e^{-t}\,dt \quad \text{for } k = 1, 2, \ldots$$

Example (cont'd)

Recall that we had derived these results using the theory of the Lebesgue integral. In particular, by applying the Monotone Convergence Theorem to the sequence of functions $f_n\colon \mathbb{R} \to \mathbb{R}$ defined by

$$f_n(t) = \begin{cases} t^{x-1}\mathrm{e}^{-t} & \text{if } x \in \left[\frac{1}{n}, n\right], \\ 0 & \text{otherwise} \end{cases}$$

(considering x as fixed), we had established the existence of $\int_0^\infty t^{x-1}\mathrm{e}^{-t}\,\mathrm{d}t$ as a Lebesgue integral.

The argument used in an essential way the non-negativity of the integrand $t^{x-1}\mathrm{e}^{-t}$, which implies that (f_n) is monotonically increasing. The Monotone Convergence Theorem then gives

$$\int_0^\infty t^{x-1}\mathrm{e}^{-t}\,\mathrm{d}t = \lim_{n\to\infty}\int_{\mathbb{R}} f_n(t)\,\mathrm{d}t = \lim_{n\to\infty}\int_{1/n}^{n} t^{x-1}\mathrm{e}^{-t}\,\mathrm{d}t\,.$$

For general functions, however, the existence of such very special limits do not imply the existence of the corresponding "infinite" (improper Riemann or Lebesgue) integral.

Example (cont'd)

As an example consider

$$\int_{1/n}^{n} \frac{\ln x}{x}\,dx = \left[\tfrac{1}{2}\ln(x)^2\right]_{1/n}^{n} = \ln(n)^2 - \ln(1/n)^2 = 0.$$

Thus we also have

$$\lim_{n\to\infty} \int_{1/n}^{n} \frac{\ln x}{x}\,dx = 0,$$

but the improper integral

$$\int_0^{\infty} \frac{\ln x}{x}\,dx = \int_0^{1} \frac{\ln x}{x}\,dx + \int_1^{\infty} \frac{\ln x}{x}\,dx = -\infty + \infty$$

doesn't exist.

Exercise

1. Show that in the definition of a two-sided improper integral over $(0,\infty)$ in terms of two one-sided improper integrals we can take any number $a>0$ instead of $a=1$ as intermediate point, i.e., the value of

$$\int_0^a f(x)\,\mathrm{d}x+\int_a^\infty f(x)\,\mathrm{d}x=\lim_{r\downarrow 0}\int_r^a f(x)\,\mathrm{d}x+\lim_{R\to\infty}\int_a^R f(x)\,\mathrm{d}x$$

doesn't depend on a.

2. In stark contrast with this, prove the following fact regarding the previous example. Given any numbers $0<r<R$ and $C\in\mathbb{R}$ show that there exists $a\in(0,r)$ and $b\in(R,\infty)$ such that

$$\int_a^b \frac{\ln x}{x}\,\mathrm{d}x=C.$$

In other words, by letting $a\downarrow 0$ and $b\to\infty$ in a certain "dependent" way, we can achieve that $\int_0^\infty \frac{\ln x}{x}\,\mathrm{d}x$ converges to any prescribed value.

Dirichlet's and Abel's Tests

Suppose that for every $x \in D$ the function $t \mapsto f(x,t)$ is continuously differentiable and monotonically decreasing, and $t \mapsto g(x,t)$ is continuous. Then $\int_0^\infty f(x,t)g(x,t)\,\mathrm{d}t$ converges uniformly on D under each of the following assumptions:

1. $f(x,t)$ converges uniformly to zero for $t \to \infty$, and there exists a "uniform bound" $M > 0$ such that $\left|\int_0^R g(x,t)\,\mathrm{d}t\right| \le M$ for all $R \in [0,\infty)$ and $x \in D$.
2. There exists $M > 0$ such that $|f(x,t)| \le M$ for all $t \in [0,\infty)$ and $x \in D$, and $\int_0^\infty g(x,t)\,\mathrm{d}t$ converges uniformly on D.

Proof.

The functions $t \mapsto f(x,t)$ and $t \mapsto g(x,t)$ satisfy the assumptions of the Second Mean Value Theorem for integrals; cf. subsequent slide. Hence, given $x \in D$ and $0 < R < R'$, there exists $\tau \in [R,R']$ such that

$$\int_R^{R'} f(x,t)g(x,t)\,\mathrm{d}t = f(x,R)\int_R^{\tau} g(x,t)\,\mathrm{d}t + f(x,R')\int_\tau^{R'} g(x,t)\,\mathrm{d}t\,.$$

Proof cont'd.

Case 1 (Dirichlet): We have

$$\left|\int_R^\tau g(x,t)\,\mathrm{d}t\right| = \left|\int_0^\tau g(x,t)\,\mathrm{d}t - \int_0^R g(x,t)\,\mathrm{d}t\right| \le 2M,$$

and similarly $\left|\int_\tau^{R'} g(x,t)\,\mathrm{d}t\right| \le 2M$. Since $f(x,t) \to 0$ uniformly for $t \to \infty$, there exists $R_0 \in [0,\infty)$ such that $|f(x,t)| < \epsilon/(4M)$ for all $t > R_0$ and $x \in D$. For $R' > R > R_0$ we then get

$$\left|\int_R^{R'} f(x,t)g(x,t)\,\mathrm{d}t\right| < \frac{\epsilon}{4M}\cdot 2M + \frac{\epsilon}{4M}\cdot 2M = \epsilon,$$

independently of $x \in D$. Applying the Cauchy test for uniform convergence finishes the proof.

Case 2 (Abel): Here we can apply the reverse Cauchy test to $\int_0^\infty g(x,t)\,\mathrm{d}t$. If R_0 denotes a response to $\epsilon/(2M)$ in this test and $R' > R > R_0$, we can upper-bound $\left|\int_R^\tau g(x,t)\,\mathrm{d}t\right|$ and $\left|\int_\tau^{R'} g(x,t)\,\mathrm{d}t\right|$ by $\epsilon/(2M)$, and then finish the proof in the same way. □

Exercise

The following facts are commonly called the First and Second Mean Value Theorem for (Riemann) integrals. Prove these facts.

1. Suppose $f\colon [a,b] \to \mathbb{R}$ is non-negative and integrable and $g\colon [a,b] \to \mathbb{R}$ is continuous. Then there exists $\tau \in [a,b]$ such that
$$\int_a^b f(t)g(t)\,\mathrm{d}t = g(\tau)\int_a^b f(t)\,\mathrm{d}t\,.$$
Hint: Let $m = \min\{g(t); t \in [a,b]\}$, $M = \max\{g(t); t \in [a,b]\}$. Then $m \le g(t) \le M$ for $t \in [a,b]$. Multiply these inequalities by $f(t)$ and integrate over $[a,b]$.

2. Suppose $f\colon [a,b] \to \mathbb{R}$ is continuously differentiable and monotonic and $g\colon [a,b] \to \mathbb{R}$ is continuous. Then there exists $\tau \in [a,b]$ such that
$$\int_a^b f(t)g(t)\,\mathrm{d}t = f(a)\int_a^\tau g(t) + f(b)\int_\tau^b g(t)\,\mathrm{d}t\,.$$
Hint: Use integration by parts with $G(t) = \int_a^t g(s)\,\mathrm{d}s$, and apply the First Mean Value Theorem to the resulting integral.

Example

We consider the function

$$F(x) = \int_0^\infty \frac{\sin t}{t}\,\mathrm{e}^{-xt}\,\mathrm{d}t, \quad x \in [0,\infty). \tag{FE}$$

For $x \geq \delta > 0$ we can estimate the absolute values of the integrand $f(x,t) = \frac{\sin t}{t}\,\mathrm{e}^{-xt}$ and its partial derivative $f_x(x,t) = -\sin t\,\mathrm{e}^{-xt}$ by $\Phi(t) = \mathrm{e}^{-\delta t}$, and apply Weierstrass's test to show the uniform convergence of the integrals $\int_0^\infty \frac{\sin t}{t}\,\mathrm{e}^{-xt}\,\mathrm{d}t$ and $\int_0^\infty \sin t\,\mathrm{e}^{-xt}\,\mathrm{d}t$ on $[\delta,\infty)$ for any $\delta > 0$. The Differentiation Theorem then implies that F is differentiable on $(0,\infty)$ with $F'(x) = -\int_0^\infty \sin t\,\mathrm{e}^{-xt}\,\mathrm{d}t$.

Using integration by parts on the expression for $F'(x)$ two times (either way), one finds that $F'(x) = -1/(1+x^2)$ and hence $F(x) = -\arctan x + C$ for $x > 0$, where C is some constant. In fact $C = \pi/2$ and $F(x) = \pi/2 - \arctan x = \operatorname{arccot} x$, as follows from $\lim_{x\to\infty} F(x) = 0$. The latter can be proved by (with justification!) interchanging the order of limit and integral in (FE), but is also easy to see directly:

$$|F(x)| \leq \int_0^\infty \left|\frac{\sin t}{t}\,\mathrm{e}^{-xt}\right| \mathrm{d}t \leq \int_0^\infty \mathrm{e}^{-xt}\,\mathrm{d}t = \frac{1}{x} \quad \text{for } x > 0.$$

Example (cont'd)

The actual motivation to consider the function $F(x)$ is that it leads to an evaluation of the famous *Dirichlet integral*

$$F(0) = \int_0^\infty \frac{\sin t}{t}\,\mathrm{d}t = \frac{\pi}{2}. \tag{D}$$

For this we need to show that F is continuous in $x = 0$, because then we can use
$F(0) = \lim_{x\downarrow 0} F(x) = \lim_{x\downarrow 0} [\pi/2 - \arctan x] = \pi/2$. However, since (D) doesn't converge absolutely, it is impossible to argue with the Weierstrass test or Lebesgue's Bounded Convergence Theorem, making this step the most difficult in the evaluation of (D).

But with the Dirichlet test for uniform convergence at hand, it is easy to do: For $x \geq 0$, $t \geq 1$ let $f(x,t) = \mathrm{e}^{-xt}/t$, $g(x,t) = \sin t$. All assumptions of the test are satisfied (e.g., $f(x,t) \to 0$ uniformly for $t \to \infty$ follows from $0 < f(x,t) \leq 1/t$).
$\Longrightarrow \int_1^\infty \frac{\sin t}{t}\,\mathrm{e}^{-xt}\,\mathrm{d}t$ converges uniformly on $[0,\infty)$ and hence represents a continuous function $F_1(x)$ (by the Continuity Theorem).
But $F(x) = F_1(x) + \int_0^1 \frac{\sin t}{t}\,\mathrm{e}^{-xt}\,\mathrm{d}t$, and the latter is also continuous (by the Continuity Lemma).

Example

Similar reasoning can be used to evaluate the integral $\displaystyle\int_0^\infty \frac{\cos t}{t^2+1}\,\mathrm{d}t$.

This is the subject of an accompanying exercise (H18 of Homework 3). Here, as a preparation for the exercise, we show that $F\colon \mathbb{R} \to \mathbb{R}$ defined by

$$F(x) = \int_0^\infty \frac{\cos(xt)}{t^2+1}\,\mathrm{d}t$$

is continuous on $\mathbb{R}$ and differentiable on $\mathbb{R}\setminus\{0\}$ with derivative

$$F'(x) = -\int_0^\infty \frac{t\sin(xt)}{t^2+1}\,\mathrm{d}t, \quad x \neq 0.$$

Since $\left|\frac{\cos(xt)}{t^2+1}\right| \leq \frac{1}{t^2+1}$ and $\int_0^\infty \frac{\mathrm{d}t}{t^2+1} = [\arctan(t)]_0^\infty = \pi/2$ is finite, $\int_0^\infty \frac{\cos(xt)}{t^2+1}\,\mathrm{d}t$ converges uniformly on $\mathbb{R}$ (Weierstrass's test).
$\Longrightarrow$ F is continuous on $\mathbb{R}$ (Continuity Theorem).

Justifying the differentiation under the integral sign is more complicated, since the corresponding integrand $t \mapsto \frac{t\sin(xt)}{t^2+1}$ is not absolutely integrable.

Example (cont'd)

It suffices to show that $\int_0^\infty \frac{t\sin(xt)}{t^2+1}\,\mathrm{d}t$ converges uniformly on $[\delta,\infty)$ for every $\delta > 0$. Then the Differentiation Theorem gives $F'(x) = -\int_0^\infty \frac{t\sin(xt)}{t^2+1}\,\mathrm{d}t$ for $x > 0$ and, since F is even, this also holds for $x < 0$.

For a proof we can apply Dirichlet's test with $f(x,t) = \frac{t}{t^2+1}$, $g(x,t) = \sin(xt)$. Since

$$\int_0^R \sin(xt)\,\mathrm{d}t = \left[-\frac{\cos(xt)}{x}\right]_0^R = \frac{1-\cos(xR)}{x} \le \frac{2}{\delta},$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\frac{t}{t^2+1} = \frac{1-t^2}{(t^2+1)^2} < 0 \quad \text{for } t > 1,$$

and of course $\lim_{t\to\infty} \frac{t}{t^2+1} = 0$, the assumptions of Dirichlet's test are satisfied (strictly speaking, only for $\int_1^\infty \frac{t\sin(xt)}{t^2+1}\,\mathrm{d}t$, but uniform convergence of $\int_0^\infty \frac{t\sin(xt)}{t^2+1}\,\mathrm{d}t$ is equivalent to that of $\int_1^\infty \frac{t\sin(xt)}{t^2+1}\,\mathrm{d}t$).

Example (cont'd)

The following, more direct proof using integration by parts is also instructive:

$$\int_0^\infty \frac{t\sin(xt)}{t^2+1}\,\mathrm{d}t = \left[-\frac{\cos(xt)}{x}\,\frac{t}{t^2+1}\right]_0^\infty - \int_0^\infty -\frac{\cos(xt)}{x}\,\frac{1-t^2}{(t^2+1)^2}\,\mathrm{d}t$$
$$= \frac{1}{x}\int_0^\infty \cos(xt)\,\frac{1-t^2}{(t^2+1)^2}\,\mathrm{d}t\,.$$

Since $\frac{1-t^2}{(t^2+1)^2} = \mathrm{O}(t^{-2})$ for $t \to \infty$, the last integral converges uniformly for $x \in [0,\infty)$ (Weierstrass's test), and hence $\int_0^\infty \frac{t\sin(xt)}{t^2+1}\,\mathrm{d}t$ converges uniformly on each interval $[\delta,\infty)$, $\delta > 0$.

The example nicely illustrates what can go wrong if you blindly interchange limits and integration without thinking about proper justification: From the formula for $F'(x)$ one is tempted to conclude that $F'(0) = -\int_0^\infty \frac{t\sin(0\,t)}{t^2+1}\,\mathrm{d}t = 0$, but this is wrong! When you solve Exercise H18 you will see that F is not differentiable at $x = 0$.

Math 285
Introduction to Differential Equations

Thomas Honold

ZJU-UIUC Institute

Spring Semester 2024

Outline

Today's Lecture: Preparations for the Existence and Uniqueness Theorem

Problem Restatement

Consider an explicit first-order ODE $y' = f(t, y)$ with a continuous function $f\colon D \to \mathbb{R}$, $D \subseteq \mathbb{R}^2$ open, and the corresponding initial value problems $y' = f(t, y) \wedge y(t_0) = y_0$ for $(t_0, y_0) \in D$.

Observation

$\phi\colon I \to \mathbb{R}$ is a solution of $y' = f(t, y) \wedge y(t_0) = y_0$ if and only if the graph $\mathrm{G}_\phi = \{(t, \phi(t)); t \in I\}$ of ϕ is contained in D and

$$\phi(t) = y_0 + \int_{t_0}^{t} \phi'(s)\mathrm{d}s = y_0 + \int_{t_0}^{t} f\big(s, \phi(s)\big)\,\mathrm{d}s$$

for all $t \in I$. Here $I \subseteq \mathbb{R}$ is an interval containing t_0 in its interior.

Equivalently, $\phi(t)$ is a fixed point ("fixed function") of the operator $\phi \mapsto T\phi$ defined by

$$(T\phi)(t) = y_0 + \int_{t_0}^{t} f\big(s, \phi(s)\big)\,\mathrm{d}s, \quad t \in I.$$

As domain of T we can take the set of continuous functions $\phi\colon I \to \mathbb{R}$ with $\mathrm{G}_\phi = \{(t, \phi(t)); t \in I\} \subseteq D$.

Thus the (local) existence of solutions of the IVP $y' = f(t, y) \wedge y(t_0) = y_0$ reduces to the following

Problem

Given $(t_0, y_0) \in D$, show that there exists an interval $I = (t_0 - \delta, t_0 + \delta)$, $\delta > 0$, such that the corresponding operator T (which depends on I) has a fixed point.

The Existence Theorem for solutions of 1st-order ODE's (and ODE systems) will be proved in this way, but the proof can be given only after several further preparations.

The Uniqueness Theorem is easier to prove and essentially requires only to find the correct condition on the function $f(t, y)$ that implies the uniqueness of solutions. But the proof is also far from being trivial, as you will see.

Order Reduction

Now consider an explicit n-th order ODE $y^{(n)} = f(t, y, y', \ldots, y^{(n-1)})$ with a continuous function $f\colon D \to \mathbb{R}$, $D \subseteq \mathbb{R}^{n+1}$ open, and the corresponding initial value problems obtained by prescribing $y^{(i)}(t_0) = y_i$ for $0 \le i \le n-1$ for some $(t_0, y_0, \ldots, y_{n-1}) \in D$.

Observation

Writing the ODE in the vectorial form

$$\begin{pmatrix} y \\ y' \\ \vdots \\ y^{(n-1)} \end{pmatrix}' = \begin{pmatrix} y' \\ y'' \\ \vdots \\ y^{(n)} \end{pmatrix} = \begin{pmatrix} y' \\ y'' \\ \vdots \\ f(t, y, y', \ldots, y^{(n-1)}) \end{pmatrix},$$

we see that it is equivalent to the first-order ODE system $\mathbf{y}' = \mathbf{f}(t, \mathbf{y})$ with $\mathbf{f}\colon D \to \mathbb{R}^n$ defined by

$$\mathbf{f}(t, y_0, \ldots, y_{n-1}) = \begin{pmatrix} y_1 \\ \vdots \\ y_{n-1} \\ f(t, y_0, y_1, \ldots, y_{n-1}) \end{pmatrix}.$$

Order Reduction Cont'd

This is so because $\mathbf{y}' = \mathbf{f}(t, \mathbf{y})$ written out in full means

$$\begin{pmatrix} y_0' \\ y_1' \\ \vdots \\ y_{n-2}' \\ y_{n-1}' \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \\ f(t, y_0, y_1, \ldots, y_{n-1}) \end{pmatrix},$$

and a solution n-tuple $(y_0, y_1, \ldots, y_{n-1})$ must satisfy

$$\begin{aligned} y_1 &= y_0', \\ y_2 &= y_1' = y_0'', \\ &\vdots \\ y_{n-1} &= y_0^{(n-1)}, \quad \text{and hence} \\ y_0^{(n)} &= y_{n-1}' = f(t, y_0, y_1, \ldots, y_{n-1}) = f(t, y_0, y_0', \ldots, y_0^{(n-1)}). \end{aligned}$$

In other words, the first coordinate function is a solution of the n-th order ODE and the remaining coordinate functions are its derivatives up to order $n-1$.

Order Reduction Cont'd

For the corresponding IVP's the same reduction applies:
A solution of the vectorial IVP

$$\mathbf{y}' = \mathbf{f}(t, \mathbf{y}), \quad \mathbf{y}(t_0) = \mathbf{y}^0 = (y_0^0, y_1^0, \ldots, y_{n-1}^0)$$

has as its first component function $y_0(t)$ a solution of the n-th order IVP

$$y^{(n)} = f(t, y, y', \ldots, y^{(n-1)}), \quad y^{(i)}(t_0) = y_i^0 \text{ for } 0 \le i \le n-1.$$

Conclusion

Extending the scope to systems of ODE's allows us to restrict attention to first-order systems only.

The operator view applies also to this case and shows that a solution of $\mathbf{y}' = \mathbf{f}(t, \mathbf{y}) \wedge \mathbf{y}(t_0) = \mathbf{y}^0$ satisfies

$$\mathbf{y}(t) = \mathbf{y}^0 + \int_{t_0}^{t} \mathbf{f}\big(s, \mathbf{y}(s)\big)\, \mathrm{d}s\,.$$

and hence is a fixed point of the operator T defined by $(T\phi)(t) = \mathbf{y}^0 + \int_{t_0}^{t} \mathbf{f}\big(s, \phi(s)\big)\, \mathrm{d}s$.

For the subsequent development it is instructive to recall a similar setting from Calculus I, where solving a fixed-point equation for a certain "operator" (map) was also required:

Newton's Method for finding roots (cf. [Ste21], Ch. 4.8)

Suppose we want to compute a solution of an univariate equation like $\sin(x) = 1/3$. This equation can be rewritten as $f(x) = 0$ with $f(x) = \sin x - 1/3$ and solved as follows.

Suppose we know already a good approximation x_n to the unknown root x^* of f. It is then reasonable to replace $f(x)$ by its linear approximation $\ell(x)$ in x_n and take the root of ℓ as new (hopefully better) approximation to x^*.

$$\ell(x) = f(x_n) + f'(x_n)(x - x_n) = 0 \iff x = x_n - \frac{f(x_n)}{f'(x_n)} =: x_{n+1}$$

Repeating this step gives a sequence $x_0, x_1, x_2, \ldots$, which is determined by the recurrence relation $x_{n+1} = x_n - f(x_n)/f'(x_n)$ and the initial value x_0.

After introducing the operator $T(x) = x - f(x)/f'(x)$, the recurrence relation becomes $x_{n+1} = T(x_n)$.

Newton's Method cont'd

We are interested in the case where the sequence (x_n) converges in $\mathbb{R}$, say to x^*. Passing to the limit in $x_{n+1} = T(x_n)$ and using continuity of T (which requires that f is C^1 and f' has no zero "nearby"), we obtain

$$x^* = \lim_{n\to\infty} x_{n+1} = \lim_{n\to\infty} T(x_n) = T\left(\lim_{n\to\infty} x_n\right) = T(x^*).$$

$\Longrightarrow$ x^* is a fixed point of T.
$\Longrightarrow$ x^* is a root of f, since $T(x) = x - f(x)/f'(x) = x$ is equivalent to $f(x) = 0$.

Thus (x_n) can only converge to a root of f. But how can we be sure that the sequence actually converges (or, rather, how to choose the starting value x_0, so that the sequence must converge?).

First answer: Suppose we know already that f has a root x^*. (For example, if $a < b$ are such that $f(a) < 0$, $f(b) > 0$ then the Intermediate Value Theorem implies $f(x^*) = 0$ for some $x^* \in (a, b)$.)

$$\Longrightarrow\ x_{n+1} - x^* = T(x_n) - T(x^*) = T'(\xi_n)(x_n - x^*)$$

for some ξ_n between x^* and x_n.

Newton's Method cont'd

Since

$$T'(x) = 1 - \frac{f'(x)^2 - f(x)f''(x)}{f'(x)^2} = \frac{f(x)f''(x)}{f'(x)^2},$$

we have $T'(x^*) = 0$, and hence (provided T' is continuous, which requires f to be of class C^2) $|T'(x)|$ is very small near x^*.
$\Longrightarrow$ (x_n) will converge rapidly to x^* if the starting value x_0 is sufficiently close to x^*.
For example, suppose we know that $x^* \in (a, b)$, $|T'(x)| \leq \frac{1}{2}$ for $x \in [a, b]$, and $T(a) \leq b$. Then the iteration with initial value $x_0 = a$ will converge to x^*. (For the proof consider the sign of $T'(\xi_n)$.)

Speed of convergence: 2nd-order Talor approximation of T in x^* gives, using $T'(x^*) = 0$,

$$x_{n+1} - x^* = \frac{T''(\xi_n)}{2}(x_n - x^*)^2,$$

again with ξ_n between x^* and x_n; moreover, if f is of class C^3 then $T''(\xi_n) \to T''(x^*) = f''(x^*)/f'(x^*)$.
This is called *quadratic convergence* and says that the number of correct digits in the decimal expansion of x_n essentially doubles at every iteration.

Newton's Method cont'd

Second answer: Suppose we don't yet know that f has a root. In this case we consider the difference

$$x_{n+1} - x_n = T(x_n) - T(x_{n-1}) = T'(\xi_n)(x_n - x_{n-1})$$

with ξ_n between x_{n-1} and x_n.
Further we suppose that there exists a constant $C < 1$ such that $|T'(\xi_n)| \le C$ for all n. (For example, this holds if $|T'(x)| \le C < 1$ on $[a, b]$ and $x_n \in [a, b]$ for all n.) Then, using induction, we obtain

$$\begin{aligned}
|x_{n+1} - x_n| &\le C^n |x_1 - x_0|, \\
|x_{n+k} - x_n| &\le \sum_{i=1}^{k} |x_{n+i} - x_{n+i-1}| \\
&\le (C^n + C^{n+1} + \cdots + C^{n+k-1}) |x_1 - x_0| \\
&\le \left(\sum_{i=n}^{\infty} C^i \right) |x_1 - x_0| = \frac{C^n}{1 - C} |x_1 - x_0|.
\end{aligned}$$

Since $\lim_{n\to\infty} C^n = 0$, given $\epsilon > 0$ we can find a response $N \in \mathbb{N}$ such that $|x_m - x_n| < \epsilon$ whenever $m, n > N$. This says that the sequence (x_n) is a Cauchy sequence and hence converges in $\mathbb{R}$.

Definition

A real-valued sequence (a_n) is said to be a *Cauchy sequence* (or to satisfy the *Cauchy criterion* for convergence) if for every $\epsilon > 0$ there exists $N = N_\epsilon \in \mathbb{N}$ such that $|x_m - x_n| < \epsilon$ whenever $m, n > N$.

Theorem

Every Cauchy sequence in $\mathbb{R}$ converges.

Proof.

We have stated and proved this theorem in Calculus III. Here is a different proof: Given a Cauchy sequence (a_n), define two further sequences (ℓ_n), (u_n) by

$$\ell_n = \inf\{a_n, a_{n+1}, a_{n+2}, \dots\},$$
$$u_n = \sup\{a_n, a_{n+1}, a_{n+2}, \dots\}.$$

$$\Longrightarrow\ \ell_1 \le \ell_2 \le \cdots \le \ell_n \le a_n \le u_n \le u_{n-1} \le \dots u_1.$$

Since (ℓ_n) is non-decreasing and bounded from above by u_1, say, the limit $\ell = \lim_{n\to\infty} \ell_n$ exists in $\mathbb{R}$. Similarly, $u = \lim_{n\to\infty} u_n$ exists in $\mathbb{R}$. We claim that $\ell = u$.

Proof cont'd.

Consider $\epsilon > 0$. Then for $n = N_\epsilon + 1$ and $m > n$ we have

$$\begin{aligned} & a_n - \epsilon < a_m < a_n + \epsilon; \\ \Longrightarrow\ & a_n - \epsilon \le \ell_n \le u_n \le a_n + \epsilon. \end{aligned}$$

Hence $u_n - \ell_n \le 2\epsilon$, which (since $\epsilon > 0$ is arbitrary) implies $\ell = u$.

Finally we can apply the squeezing theorem to conclude from $\ell_n \le a_n \le u_n$ that $\lim_{n\to\infty} a_n = \ell = u$ as well. □

Remark

The definition of Cauchy sequences makes also sense for the Euclidean spaces $\mathbb{R}^d$, $d > 1$, and in particular for $\mathbb{C} \triangleq \mathbb{R}^2$. An easy adaption of the previous proof shows that Cauchy sequences in $\mathbb{R}^d$ converge as well; cf. next exercise.

Exercise

Show that every Cauchy sequence $(\mathbf{x}^{(n)})$ in $\mathbb{R}^d$, $d > 1$, converges.

Hint: Show first that for $1 \le i \le d$ the i-th coordinate sequence of $(\mathbf{x}^{(n)})$, which is defined as $x_i^{(1)}, x_i^{(2)}, x_i^{(3)}, \ldots$ where $\mathbf{x}^{(n)} = (x_1^{(1)}, \ldots, x_d^{(n)})$, is a Cauchy sequence in $\mathbb{R}$.

Review of Differentiable Multivariate Functions

Recall that a function $f\colon D \to \mathbb{R}^m$, $D \subseteq \mathbb{R}^n$, is differentiable in a point $\mathbf{x}_0 \in D$ (which must be an inner point of D) if $f(\mathbf{x})$ can be linearly approximated near $\mathbf{x}_0$ with an error $\mathrm{o}(\mathbf{x} - \mathbf{x}_0)$; more precisely, if there exists a linear map $L\colon \mathbb{R}^n \to \mathbb{R}^m$ such that

$$f(\mathbf{x}_0 + \mathbf{h}) = f(\mathbf{x}_0) + L(\mathbf{h}) + \mathrm{o}(\mathbf{h}) \quad \text{for } \mathbf{h} \to \mathbf{0}$$

or, equivalently, $\lim_{\mathbf{h}\to\mathbf{0}} |f(\mathbf{x}_0 + \mathbf{h}) - f(\mathbf{x}_0) - L(\mathbf{h})| \,/\, |\mathbf{h}| = 0$.

If applicable, the linear map L is uniquely determined by this condition. It is called the *differential* of f at the point $\mathbf{x}_0$ and usually denoted by $\mathrm{d}f(\mathbf{x}_0)$. In terms of the differential, the above condition takes the form (rewritten in terms of $\mathbf{x} = \mathbf{x}_0 + \mathbf{h}$)

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \mathrm{d}f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + \mathrm{o}(\mathbf{x} - \mathbf{x}_0) \quad \text{for } \mathbf{x} \to \mathbf{x}_0.$$

The matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ representing $L = \mathrm{d}f(\mathbf{x}_0)$ (i.e., $L(\mathbf{h}) = \mathbf{A}\mathbf{h}$ for $\mathbf{h} \in \mathbb{R}^n$) is called *Jacobi matrix* (or *functional matrix*) of f at $\mathbf{x}_0$ and denoted by $\mathbf{J}_f(\mathbf{x}_0)$. The entries a_{ij} of $\mathbf{A}$ turn out to be the partial derivatives of f at $\mathbf{x}_0$: Writing $f = (f_1, \ldots, f_m)$, we have $a_{ij} = \frac{\partial f_i}{\partial x_j}(\mathbf{x}_0)$.

Example (Squaring map in $\mathbb{C}$)

Since $z^2 = (x+y\mathrm{i})^2 = x^2 + 2xy\mathrm{i} + \mathrm{i}^2y^2 = x^2 - y^2 + 2xy\,\mathrm{i}$, it is natural to call the map $s\colon \mathbb{R}^2 \to \mathbb{R}^2$ defined by $s(x,y) = (x^2 - y^2, 2xy)$ the complex squaring map.

$$\begin{aligned} s(x+h_1, y+h_2) &= \begin{pmatrix} (x+h_1)^2 - (y+h_2)^2 \\ 2(x+h_1)(y+h_2) \end{pmatrix} \\ &= \begin{pmatrix} x^2 - y^2 + 2xh_1 - 2yh_2 + h_1^2 - h_2^2 \\ 2xy + yh_1 + xh_2 + h_1h_2 \end{pmatrix} \\ &= \begin{pmatrix} x^2 - y^2 \\ 2xy \end{pmatrix} + \begin{pmatrix} 2x & -2y \\ 2y & 2x \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} + \begin{pmatrix} h_1^2 - h_2^2 \\ 2h_1h_2 \end{pmatrix} \\ &= s(x,y) + \mathbf{J}_s(x,y)\mathbf{h} + R(\mathbf{h}) \end{aligned}$$

with $R(\mathbf{h}) = \mathrm{o}(\mathbf{h})$.

You can verify that the entries of $\mathbf{J}_s(x,y)$ are the partial derivatives $(s_1)_x$, $(s_1)_y$, $(s_2)_x$, $(s_2)_y$ of $s_1(x,y) = x^2 - y^2$ and $s_2(x,y) = 2xy$.

Newton's Method cont'd

Newton's Method can also be used to solve vectorial equations numerically, e.g.,

$$\begin{array}{rcrcr} 5x & + & e^y & = & -4, \\ x^2 & - & xy & = & 2. \end{array}$$

Setting $f(x,y) = (5x + e^y + 4, x^2 - xy - 2)$ and $\mathbf{x} = (x,y)$, the system becomes $f(\mathbf{x}) = \mathbf{0}$, and we can use the same idea as in the 1-dimensional case (writing $\mathbf{x}^{(n)} = (x_n, y_n)$):

$$\begin{aligned} & \ell(\mathbf{x}) = f(\mathbf{x}^{(n)}) + \mathbf{J}_f(\mathbf{x}^{(n)})(\mathbf{x} - \mathbf{x}^{(n)}) = \mathbf{0} \\ \iff \quad & \mathbf{x} = \mathbf{x}^{(n)} - \mathbf{J}_f(\mathbf{x}^{(n)})^{-1} f(\mathbf{x}^{(n)}) =: \mathbf{x}^{(n+1)}, \end{aligned}$$

provided that $\mathbf{J}_f(\mathbf{x}^{(n)})$ is invertible.

Choosing $\mathbf{x}^{(0)} = (x_0, y_0)$ suitably and assuming that during the execution only invertible matrices $\mathbf{J}_f(\mathbf{x}^{(n)})$, $n = 0, 1, 2, \ldots$, are encountered, the iteration $\mathbf{x}^{(n+1)} = T(\mathbf{x}^{(n)})$, $T(\mathbf{x}) = \mathbf{x} - \mathbf{J}_f(\mathbf{x})^{-1} f(\mathbf{x})$, defines a sequence $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots$, of points in $\mathbb{R}^2$, which converges to the unique solution of $f(\mathbf{x}) = \mathbf{0}$; see the subsequent example. (You can verify that the system has a unique solution, e.g., by eliminating y and applying standard Calculus techniques to the resulting equation for x.)

Newton's Method cont'd

The method just outlined generalizes to systems of n equations in n unknowns.

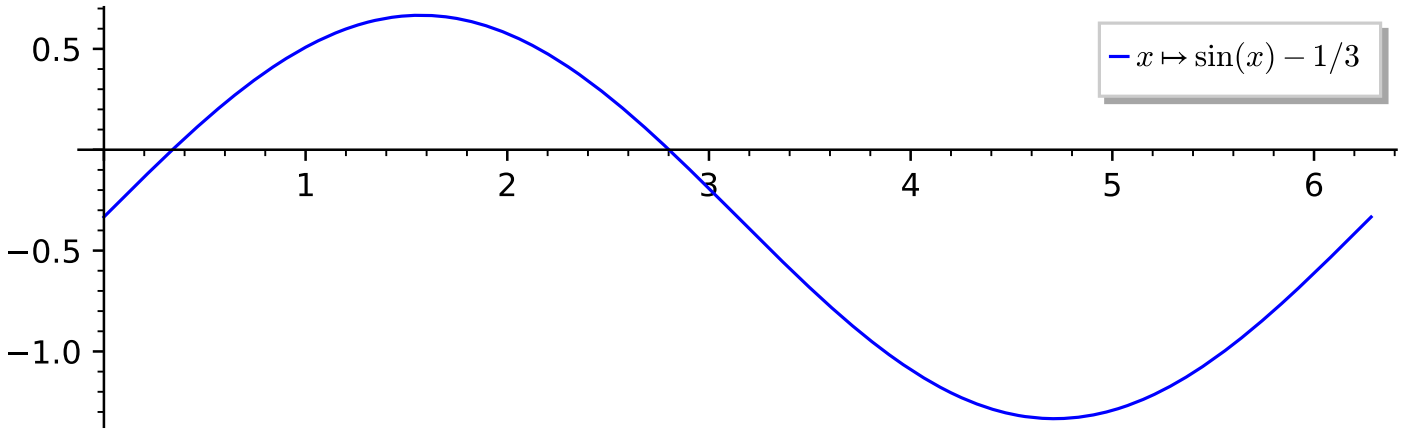
In general, however, the convergence analysis of this higher-dimensional Newton iteration is much more involved than that of the 1-dimensional iteration.
Since there is no analog of the Intermediate Value Theorem for $\mathbb{R}^d$, $d > 1$, we can only use the second method ("second answer") to prove convergence. The "contraction" property $\left|\mathbf{x}^{(n+1)} - \mathbf{x}^{(n)}\right| \leq C \left|\mathbf{x}^{(n)} - \mathbf{x}^{(n-1)}\right|$ for all n, where $C < 1$ is a fixed constant, turns out to work for $d > 1$ as well. A few more details on the method will be provided when we discuss matrix norms and in the exercises.

Example ($f(x) = \sin(x) - 1/3$)

For $f(x) = \sin(x) - 1/3$ we have $T(x) = x - \frac{\sin(x)-1/3}{\cos(x)}$ and the recurrence $x_{n+1} = x_n - \frac{\sin(x_n)-1/3}{\cos(x_n)}$. The following lists the Newton iterates for the starting values $x_0 = 1$ and $x_0 = 2$.

n	x_n
0	1.000000000000000
1	0.059530847905406 3
2	0.333854436356604 0
3	0.339830667137674 8
4	0.339836909447233 6
5	0.339836909454121 9
6	0.339836909454121 9

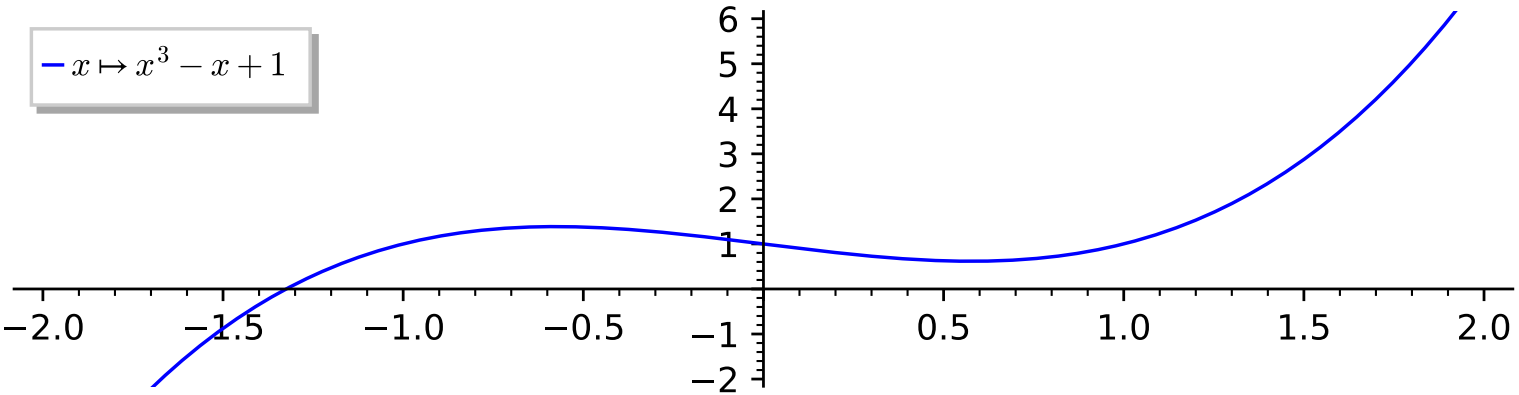
n	x_n
0	2.000000000000000
1	3.384040542687392 0
2	2.793351812048839 0
3	2.801768465049102 4
4	2.801755744164277 0
5	2.801755744135671 3
6	2.801755744135671 3



Example ($f(x) = x^3 - x + 1$)

Here it takes quite a while until quadratic convergence sets in.

n	x_n	n	x_n
0	1.000000000000000	13	−0.742494298720700**9**
1	0.500000000000000	14	−2.781295940677608**3**
2	3.000000000000000	15	−1.982725247043830**6**
3	2.038461538461538**3**	16	−1.536927379758256**3**
4	1.390282147216736**2**	17	−1.35726248318773**25**
5	0.911611897717927**0**	18	−1.32566309442886**79**
6	0.345028496748169**2**	19	−1.32471878861525**72**
7	1.427750704027270**3**	20	−1.32471795724539**02**
8	0.942417912509482**9**	21	−1.32471795724474**60**
9	0.404949357199379**6**	22	−1.32471795724474**60**
10	1.706904645182851**6**	23	−1.32471795724474**60**
11	1.155756361074813**4**	24	−1.32471795724474**60**
12	0.694191813329546**9**	25	−1.32471795724474**60**

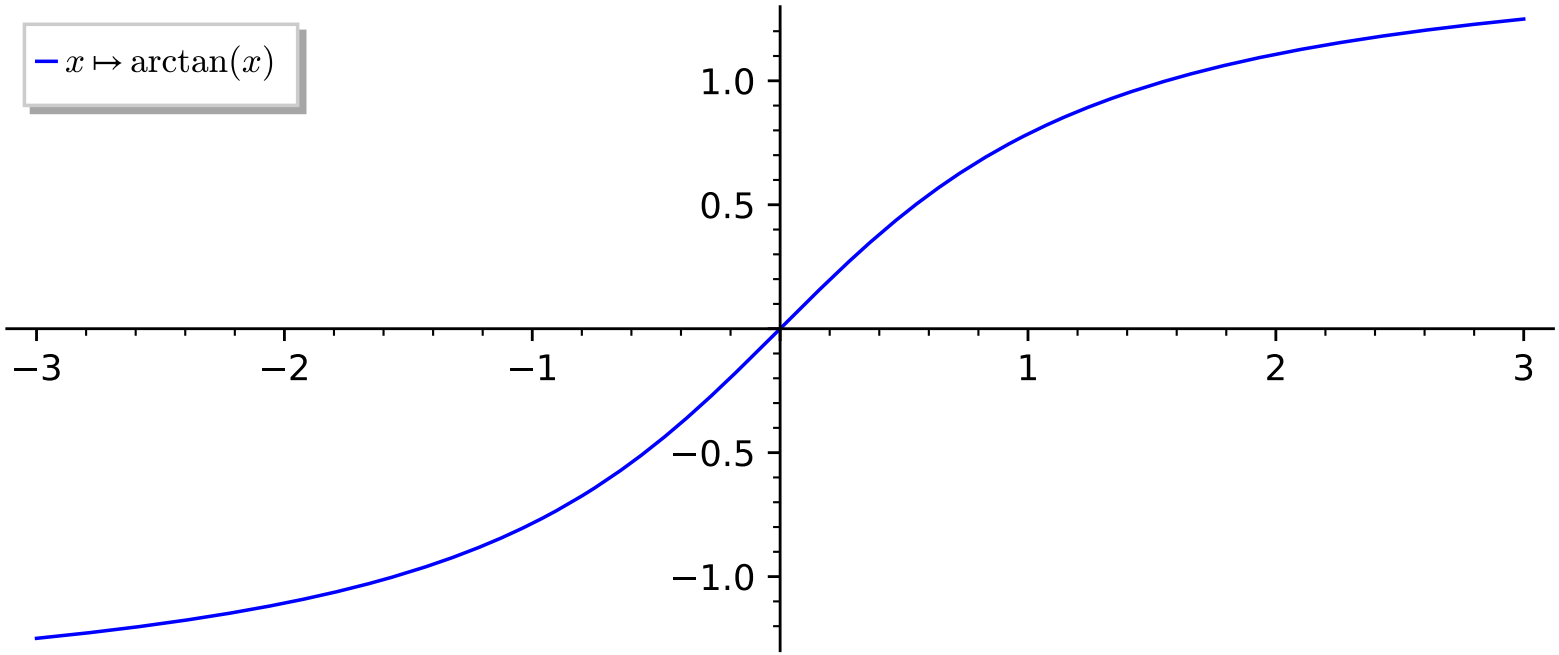


Example ($f(x) = \arctan x$)

Here convergence/divergence of the Newton iteration $x_{n+1} = T(x_n) = x_n - \arctan(x_n)(1 + x_n^2)$ depends on the choice of the initial value x_0.

n	x_n
0	1.000000000000000
1	−0.5707963267948966
2	0.1168599039989130
3	−0.0010610221170447
4	0.000000007963096
5	0.000000000000000

n	x_n
0	2.000000000000000
1	−3.5357435889704525
2	13.950959086927493
3	−279.34406653361738
4	122016.99891795458
5	−23386004197.933886



Example ($f(x,y) = (5x + e^y + 4, x^2 - xy - 2)$)

Here we have

$$T\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} 5 & e^y \\ 2x - y & -x \end{pmatrix}^{-1} \begin{pmatrix} 5x + e^y + 4 \\ x^2 - xy - 2 \end{pmatrix}$$
$$= \begin{pmatrix} x \\ y \end{pmatrix} + \frac{1}{5x + (2x - y)e^y} \begin{pmatrix} -x & -e^y \\ y - 2x & 5 \end{pmatrix} \begin{pmatrix} 5x + e^y + 4 \\ x^2 - xy - 2 \end{pmatrix}.$$

Starting with the "approximate" solution $(x_0, y_0) = (-1, 0)$ (well, rather it solves the first equation exactly), we obtain the sequence

n	x_n	y_n
0	-1.00000000000000	0.00000000000000
1	-1.14285714285714	$0.71428571428571\underline{4}$
2	-1.15343194160013	$0.57938401044252\underline{5}$
3	-1.15552495201267	$0.57528648658840\underline{1}$
4	-1.15552722080764	$0.57528445025160\underline{2}$
5	-1.15552722080795	$0.57528445025038\underline{5}$
6	-1.15552722080795	$0.57528445025038\underline{5}$

You can check that (x_5, y_5) is indeed very close to being a root of f (the entries of $f(x_5, y_5)$ have absolute value $< 10^{-14}$).

Metric Spaces

Definition

A *metric space* (M, d) consists of a set M and a map $d\colon M \times M \to \mathbb{R}$ ("distance function") satisfying the following for all $x, y, z \in M$:

(M1) $d(x,y) \geq 0$; $d(x,y) = 0 \iff x = y$; (*non-negativity*)

(M2) $d(x,y) = d(y,x)$; (*symmetry*)

(M3) $d(x,y) \leq d(x,z) + d(z,y)$. (*triangle inequality*)

Examples

1. $(\mathbb{R}^n, \mathrm{d_E})$ with $\mathrm{d_E}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$ (*Euclidean distance*);
   includes $\mathbb{R}$ and $\mathbb{C}$ with $\mathrm{d_E}(x,y) = |x-y|$, resp.,
   $\mathrm{d_E}(z,w) = |z-w| = \sqrt{(\operatorname{Re} z - \operatorname{Re} w)^2 + (\operatorname{Im} z - \operatorname{Im} w)^2}$ as special cases.

Examples (cont'd)

2. $(\mathbb{R}^n, \mathrm{d}_1)$ and $(\mathbb{R}^n, \mathrm{d}_\infty)$ with the metrics

$$\mathrm{d}_1(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n} |x_i - y_i|\,, \quad (\ell^1\text{-distance, "Manhattan distance"})$$

$$\mathrm{d}_\infty(\mathbf{x},\mathbf{y}) = \max\bigl\{|x_i - y_i|\,; 1 \le i \le n\bigr\}. \qquad (\ell^\infty\text{-distance})$$

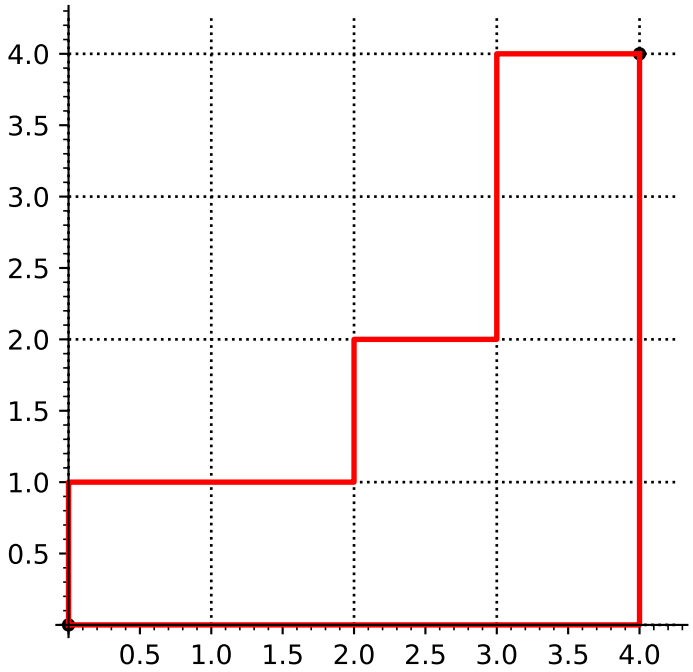
3. $(\mathbb{R}^n, \mathrm{d}_\mathrm{F})$ with $\mathrm{d}_\mathrm{F}(\mathbf{x},\mathbf{y}) = \begin{cases} \mathrm{d}_\mathrm{E}(\mathbf{x},\mathbf{y}) & \text{if } \mathbb{R}\mathbf{x} = \mathbb{R}\mathbf{y}, \\ \mathrm{d}_\mathrm{E}(\mathbf{x},\mathbf{0}) + \mathrm{d}_\mathrm{E}(\mathbf{0},\mathbf{y}) & \text{if } \mathbb{R}\mathbf{x} \ne \mathbb{R}\mathbf{y}. \end{cases}$

   d_F is sometimes called *"French distance"* or, more accurately, *metric of the French railway network*.

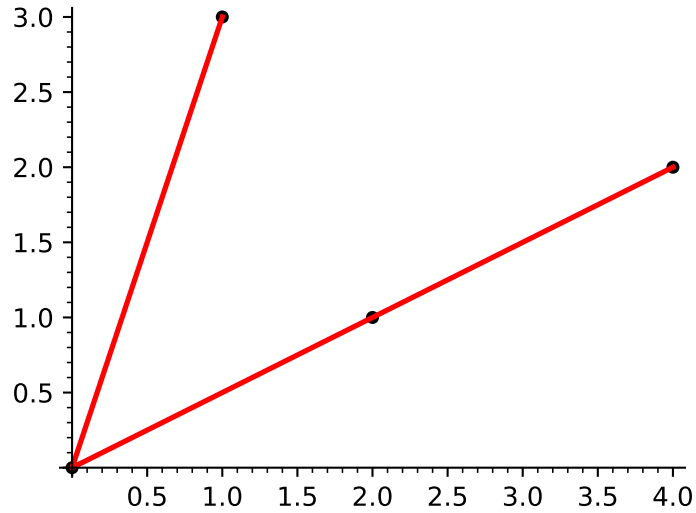
4. The set of all complex-valued, infinite sequences $(a_n)_{n=0}^{\infty}$ satisfying $\sum_{n=0}^{\infty} |a_n|^2 < \infty$ with distance function

$$d\bigl((a_n),(b_n)\bigr) = \sqrt{\sum_{n=0}^{\infty} |a_n - b_n|^2}.$$

   This metric space is known as *Hilbert's Cube* and usually denoted by ℓ^2.



(a) Manhattan distance



(b) French distance

(a) The Manhattan distance (ℓ^1-distance) from $(0,0)$ to $(4,4)$ is 8, equal to the length of any path from $(0,0)$ to $(4,4)$ that uses only northward and eastward unit steps.

(b) The French distance of the points $(2,1)$ and $(4,2)$ equals their Euclidean distance, viz. $\mathrm{d_E}\big((2,1),(4,2)\big) = \sqrt{5}$, while that of $(4,2)$ and $(1,3)$ is $\mathrm{d_E}\big((4,2),(0,0)\big) + \mathrm{d_E}\big((0,0),(1,3)\big) = \sqrt{20} + \sqrt{10}$.

Examples (cont'd)

5. Any set M (for example, $M = \mathbb{R}^n$) with distance

$$d(x,y) = \begin{cases} 0 & \text{if } x = y, \\ 1 & \text{if } x \neq y. \end{cases} \qquad (\textit{discrete metric})$$

6. Weighted connected simple graphs (V, E, w) with the shortest path distance. Here $w\colon E \to \mathbb{R}^+$ is a weight function on the edge set E, the weight or length of a path is the sum of the weights of its edges, the underlying set is the vertex set V, and $d(v,w)$ is defined as the length of a shortest path (i.e., path of smallest weight) between v and w. This example includes unweighted graphs with the shortest path distance if we assign weight 1 to all edges.

7. The set of all bit strings of length n with

$$\mathrm{d}_{\mathrm{Ham}}(\mathbf{s},\mathbf{t}) = \left|\{1 \le i \le n; s_i \neq t_i\}\right|. \qquad (\textit{Hamming distance})$$

This is a special case of Example 6, because $\mathrm{d}_{\mathrm{Ham}}(\mathbf{s},\mathbf{t})$ is equal to the length of a shortest path between $\mathbf{s}$ and $\mathbf{t}$ in the hypercube Q_n.

Examples (cont'd)

8. Any subset $M' \subseteq M$ of a metric space (M, d) forms a metric space (M', d') of its own by defining $d'(x, y) = d(x, y)$ for $x, y \in N$ (i.e., the distance on N is the induced distance).

9. The set $\mathrm{C}([a, b])$ of all continuous functions $f : [a, b] \to \mathbb{R}$ on a compact interval $[a, b] \subset \mathbb{R}$ with distance

$$d_\infty(f, g) = \max\{|f(x) - g(x)| ; a \le x \le b\}.$$

d_∞ is also referred to as *metric of uniform convergence*, since $f_n \to g$ in this metric, i.e., $\lim_{n\to\infty} d_\infty(f_n, g) = 0$, is equivalent to $f_n \to g$ uniformly.

This example admits various generalizations, e.g., the domain $[a, b]$ can be replaced by a compact set $K \subset \mathbb{R}^n$, the codomain $\mathbb{R}$ can be replaced by $\mathbb{R}^m$ if we change "absolute value" to "Euclidean length", we could work more generally with bounded functions if we change "maximum" to "supremum", or restrict to C^1-functions and use the maximum of $|f(x) - g(x)|$ and $|f'(x) - g'(x)|$ in the definition of $d_\infty(f, g)$, etc.

Examples

10. A non-connected (weighted) simple graph with the shortest-path distance forms an example of a *generalized metric space*, in which distances are allowed to take the value $\infty = +\infty$. Axioms (M1)–(M3) must still be satisfied—for example, if $d(x,y) = \infty$ then $d(x,z) < \infty \wedge d(y,z) < \infty$ is impossible.

    Further examples of generalized metric spaces are the extended real line $\overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$ with $d = \mathrm{d_E}$ on $\mathbb{R} \times \mathbb{R}$ and

    $$d(-\infty, +\infty) = d(\pm\infty, x) = \infty \quad \text{for all } x \in \mathbb{R},$$

    and the set of all functions $f\colon X \to \mathbb{R}$ on an arbitrary (but fixed) domain X with distance

    $$d_\infty(f,g) = \sup\bigl\{\,|f(x) - g(x)|\,;\, x \in X\bigr\}.$$

    Similar to the case of $\mathrm{C}\bigl([a,b]\bigr)$, d_∞ captures uniform convergence in the sense that $f_n \to g$ uniformly iff $\lim_{n\to\infty} d_\infty(f_n, g) = 0$ (which requires $d_\infty(f_n, g) < \infty$ for all but finitely many n).

Exercise

A metric space (M, d) (or just the metric d) is said to be *translation-invariant* or *norm-induced* if an addition $(x, y) \to x + y$ is defined on M and $d(x, y) = d(x + z, y + z)$ holds for all $x, y, z \in M$.

1. Which of the preceding examples of metric spaces are translation-invariant?
2. Show that a translation-invariant metric $d\colon M \times M \to \mathbb{R}$ is determined by the corresponding *norm* $n\colon M \to \mathbb{R}$ defined by $n(x) = d(x, 0)$. (The zero element $0 \in M$ is distinguished by $x + 0 = 0 + x = x$ for all $x \in M$.)
3. Which properties should a function $n\colon M \to \mathbb{R}$ satisfy in order to determine a metric on M as in b) ?

Exercise (Product metric spaces)

Suppose (M_1, d_1) and (M_2, d_2) are metric spaces and $M = M_1 \times M_2$. For $p \in \mathbb{R}^+$ define $d_p \colon M \times M \to \mathbb{R}$ by

$$d_p(\mathbf{x}, \mathbf{y}) = d_p\big((x_1, x_2), (y_1, y_2)\big) = \sqrt[p]{|x_1 - y_1|^p + |x_2 - y_2|^p}.$$

1. For which $p \in \mathbb{R}^+$ is (M, d_p) a metric space?
2. Which of our 10 examples fall under this product construction?
3. Show that $d_\infty(\mathbf{x}, \mathbf{y}) = \lim_{p \to +\infty} d_p(\mathbf{x}, \mathbf{y})$ also defines a metric on M. To which of our examples does it correspond?

Exercise

Let $d \colon M \times M \to \mathbb{R}$ be a function satisfying $d(a, a) = 0$ for $a \in M$, $d(a, b) \neq 0$ for $a, b \in M$ with $a \neq b$, and $d(a, b) \leq d(b, c)) + d(c, a)$ for $a, b, c \in M$.

1. Show that d is a metric.
2. Does this conclusion also hold if $d(a, b) \leq d(b, c)) + d(c, a)$ is replaced by the ordinary triangle inequality $d(a, b) \leq d(a, c)) + d(c, b)$?

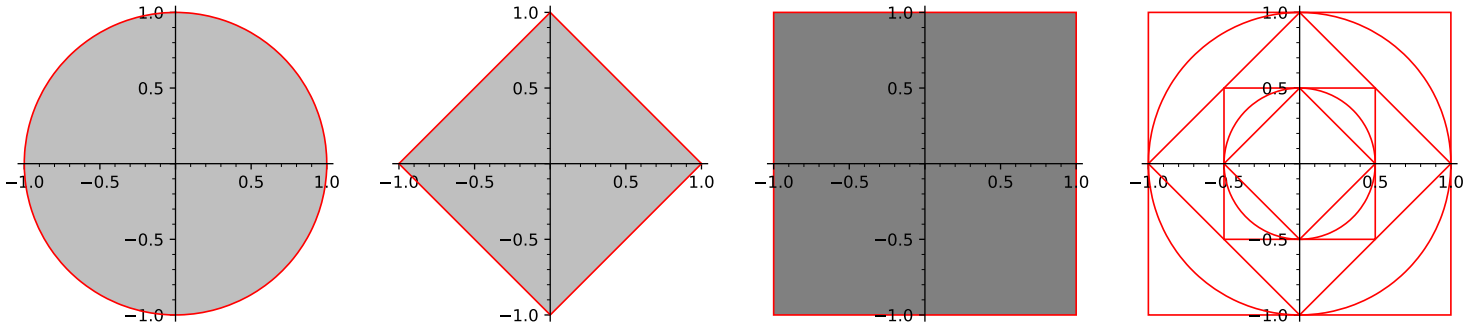
Analysis on Metric Spaces

Observations

- In any metric space we can define balls (and spheres) just as we did in $(\mathbb{R}^n, \mathrm{d_E})$. For example, the *open ball with center* $a \in M$ *and radius* $r \in \mathbb{R}^+$ is defined as $\mathrm{B}_r(a) = \{x \in M; d(x,a) < r\}$.
- Using balls, we can define open sets, closed sets, inner points, boundary points, accumulation points, limits of sequences, and continuity of maps (but not differentiability!) for arbitrary metric spaces. For example, a sequence $(a_n)_{n=0}^{\infty}$ of points $a_n \in M$ *converges to a point* $a \in M$, notation $\lim_{n\to\infty} a_n = a$, if for every $\epsilon > 0$ there exists $N = N_\epsilon \in \mathbb{N}$ such that $a_n \in \mathrm{B}_\epsilon(a)$ for all $n > N$; equivalently, $d(a_n, a) < \epsilon$ for all $n > N$.
- Care must be taken, however, when generalizing some of the less obvious (but important) properties of $(\mathbb{R}^n, \mathrm{d_E})$ to arbitrary metric spaces. An example is the Bolzano-Weierstrass Theorem, which fails to hold in a general metric space; another example is completeness.

Example

The following three figures show the unit balls with respect to the three metrics $\mathrm{d_E}$, $\mathrm{d_1}$, $\mathrm{d_\infty}$ on $\mathbb{R}^2$. (For $\mathrm{d_1}$ the closed unit ball is given by $|x| + |y| \leq 1$, and for $\mathrm{d_\infty}$ by $\max\{|x|, |y|\} \leq 1$.)



(a) $\mathrm{d_E}$ (b) $\mathrm{d_1}$ (c) $\mathrm{d_\infty}$ (d) all

The 4th figure shows the nested structure of the balls of the three metrics: Every ball of one metric contains balls of the other two metrics, possibly of smaller radius. This implies that the three metric spaces have the same convergent sequences, open sets, etc.; they are topologically indistinguishable; cf. also the exercise on strongly equivalent metrics. On the other hand, the French distance $\mathrm{d_F}$ is essentially different from these three. For example, the sequence of points $\big(\cos(1/n), \sin(1/n)\big)$, $n \in \mathbb{N}$, converges to $(1, 0)$ in $\mathrm{d_E}$, $\mathrm{d_1}$, $\mathrm{d_\infty}$ but not in $\mathrm{d_F}$, where all these points have distance 2.

Example (cont'd)

(Unit) Balls of general metric spaces can look quite weird. For the French metric this is discussed in a subsequent exercise. For a discrete metric space (M, d), the closed balls of radius $0 < r < 1$ contain only 1 element (the center) and those of radius $r \geq 1$ are all equal to M. For $\mathrm{C}([a,b])$ equipped with the metric d_∞ of uniform convergence, the unit ball centered at f consists of all continuous functions whose graph is contained in the strip of width 2 symmetrically around the graph of f; see picture.

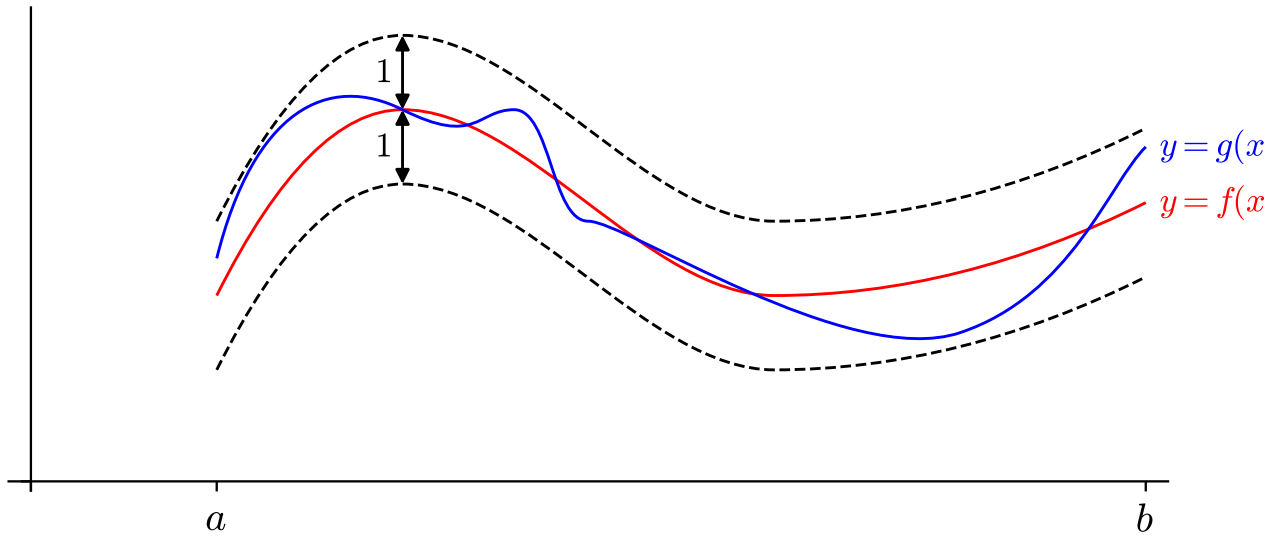


Figure: Illustration of a ball $\mathrm{B}_1(f)$ in $\mathrm{C}([a,b])$ with respect to the uniform metric and a particular function $g \in \mathrm{B}_1(f)$

It is rather obvious that the Bolzano-Weierstrass Theorem fails for discrete metric spaces, but it also fails for the Hilbert cube, which otherwise very much looks like $(\mathbb{R}^n, \mathrm{d_E})$. This is the subject of the following

Exercise

1. Give an example of a set M that when equipped with the discrete metric does not obey the Bolzano-Weierstrass Theorem "Every bounded sequence has a convergent subsequence".
2. Show that the Bolzano-Weierstrass Theorem also fails for the Hilbert cube ℓ^2.

Exercise

Determine the (closed, say) unit balls in the French metric $\mathrm{d_F}$ around each point $(x, y) \in \mathbb{R}^2$.

Exercise (Continuity of a metric)

Let (M, d) be a metric space and $(a, b) \in M \times M$. Show that for every $\epsilon > 0$ there exists a $\delta > 0$ such that $d(x, a) < \delta \wedge d(y, b) < \delta$ implies $|d(x, y) - d(a, b)| < \epsilon$.

Hint: First derive the so-called *quadrangle inequality* $|d(x, y) - d(a, b)| \leq d(x, a) + d(y, b)$.

Complete Metric Spaces

We have defined completeness of $\mathbb{R}$ using the natural ordering $\leq$. This does not generalize to arbitrary metric spaces, but there is a reformulation of the completeness property which does:

Definition

Let (M, d) be a metric space.

1. A sequence $(a_n)_{n=0}^{\infty}$ of points $a_n \in M$ is said to be a *Cauchy sequence* (or satisfy the *Cauchy criterion*) if for every $\epsilon > 0$ there exists $N = N_\epsilon \in \mathbb{N}$ such that $d(a_m, a_n) < \epsilon$ for all $m, n > N$.
2. (M, d) is said to be *complete* if every Cauchy sequence in M converges (i.e., has a limit $a \in M$).

Note

When dealing with series $\sum_{n=1}^{\infty} a_n$ rather than sequences, we must check the Cauchy criterion for the sequence of partial sums $s_n = \sum_{k=1}^{n} a_k$. This requires bounding

$$s_n - s_m = \sum_{k=m+1}^{n} a_k \quad \text{for } n > m > N.$$

Examples/Counterexamples

- We have proved that $\mathbb{R}$ is complete according to the new definition and mentioned that, more generally, the Euclidean spaces $(\mathbb{R}^d, \mathrm{d_E})$, $d = 1, 2, 3, \ldots$, are complete. In particular the field $\mathbb{C}$ of complex numbers is complete (the case $d = 2$). Here we are tacitly assuming that the underlying metric is the Euclidean metric $\mathrm{d_E}$. (Otherwise the assertion could be false.)
- Any subset M of $\mathbb{R}^n$ forms a metric space of its own with the metric induced by $\mathrm{d_E}$ (i.e., distances between points in M are the same as in $\mathbb{R}^n$). Such a metric subspace is complete iff M is a closed subset of $\mathbb{R}^n$; cf. subsequent exercise. (Recall that M is closed if the boundary ∂M is contained in M or, equivalently, M contains with any convergent sequence also its limit).

Examples/Counterexamples Cont'd

- The "punctured" real line $\mathbb{R} \setminus \{0\}$ forms an incomplete metric space (since it is not closed in $\mathbb{R}$). We can prove this directly as follows: Consider the sequence $x_n = 1/n \in \mathbb{R} \setminus \{0\}$. This sequence is a Cauchy sequence, since it has a limit in $\mathbb{R}$, viz. $lim_{n\to\infty} 1/n = 0$, and the definition of "Cauchy sequence" makes no reference to the ambient metric space M (we could even take $M = \{1/n; n \in \mathbb{N}\}$). But it has no limit in $\mathbb{R} \setminus \{0\}$, and hence $\mathbb{R} \setminus \{0\}$ is incomplete.

  On the other hand, $\mathbb{R} \setminus (0,1) = (-\infty, 0] \cup [1, +\infty)$ is complete since it is closed in $\mathbb{R}$. The analogous incompleteness "proof" using the sequence $x_n = 1/2 + 1/n$ is invalid (can you see where the argument breaks down?).

- The Hilbert cube H is complete. The proof of this is a bit technical, since the elements of H are itself sequences and hence a Cauchy sequence in H is sort of an infinite matrix of real numbers with a particular property.

Examples/Counterexamples Cont'd

- The metric spaces $\mathrm{C}\big([a,b]\big)$ of Example 9 are complete. This can be seen as follows:
  (f_n) is a Cauchy sequence w.r.t. d_∞ if for every $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that

$$|f_n(x) - f_m(x)| < \epsilon \quad \text{for all } m, n > N \text{ and } x \in [a,b]. \qquad \text{(C)}$$

  $\Longrightarrow$ All sequences $\big(f_n(x)\big)$, $x \in [a,b]$, are Cauchy sequences in $\mathbb{R}$ and hence convergent, showing that (f_n) has a point-wise limit function $f\colon [a,b] \to \mathbb{R}$.
  Letting $n \to \infty$ in (C) gives $|f(x) - f_m(x)| \le \epsilon$ for all $m > N$ and $x \in [a,b]$, showing that $f_n \to f$ uniformly.
  Now the Continuity Theorem can be applied to conclude that f is continuous, i.e., $f \in \mathrm{C}\big([a,b]\big)$. Thus every Cauchy sequence in $\mathrm{C}\big([a,b]\big)$ converges.

  For continuous functions on unbounded intervals (and on other domains such as $\mathbb{R}^n$) similar assertions hold: A sequence (f_n) of continuous functions that forms a Cauchy sequence w.r.t. to the generalized metric d_∞, converges uniformly and hence has a continuous limit function.

Examples/Counterexamples Cont'd

The subsequent exercises contain still further examples and counterexamples. Discrete metric spaces (see Example 5 and a subsequent exercise) are complete, and so are the metric spaces arising from graphs (Example 6). It is possible to change the Euclidean metric $\mathrm{d_E}$ on $\mathbb{R}$ in such a way that convergence of sequences is not affected but the new metric space $(\mathbb{R}, d)$ is bounded and incomplete; see subsequent exercises.

Exercise

Two metrics d_1, d_2 on a set M are said to be *strongly equivalent* if there exist constants $\alpha, \beta > 0$ such that

$$\alpha d_1(x,y) \le d_2(x,y) \le \beta d_1(x,y) \quad \text{for all } x, y \in M.$$

a) Show that the metric spaces (M, d_1), (M, d_2) have the same open (closed) sets, the same set of convergent sequences (Cauchy sequences), and are either both complete or both incomplete.

b) Show that the Euclidean metric $\mathrm{d_E}$ and the metrics d_1, d_∞ in Example 2 are strongly equivalent.

Exercise

For $x, y \in \mathbb{R}$ set

$$d(x,y) = \frac{\mathrm{d_E}(x,y)}{1+\mathrm{d_E}(x,y)} = \frac{|x-y|}{1+|x-y|}.$$

Show that d defines a metric on $\mathbb{R}$, which is not strongly equivalent to $\mathrm{d_E}$, but that nevertheless the conclusions in Part (1) of the previous Exercise hold for $d_1 = d$ and $d_2 = \mathrm{d_E}$.

Exercise

For $x, y \in \mathbb{R}$ set

$$d(x,y) = \mathrm{d_E}\left(\frac{x}{1+|x|}, \frac{y}{1+|y|}\right) = \left|\frac{x}{1+|x|} - \frac{y}{1+|y|}\right|.$$

a) Show that d defines a metric on $\mathbb{R}$.

b) Show that $(\mathbb{R}, d)$ has the same open sets and the same convergent sequences as $(\mathbb{R}, \mathrm{d_E})$.

c) Show that $(\mathbb{R}, d)$ is incomplete.

Hint: Consider the sequence $a_n = n$.

Exercise

Let (M, d) be a discrete metric space; cf. Example 5. Describe convergent sequences and Cauchy sequences in (M, d) in an alternative way (without using ϵ), and conclude that (M, d) is complete.

Exercise

a) Show that a closed subset N of a complete metric space (M, d) is complete in the induced metric $N \times N \to \mathbb{R}$, $(x, y) \mapsto d(x, y)$.

b) Conversely, show that a subset of a metric space that is complete in the induced metric must be closed.

Exercise

A metric space (M, d) is said to be an *ultrametric* space if it satisfies the following sharper variant of the triangle inequality:

$$d(a,b) \le \max\{d(a,c), d(c,b)\} \quad \text{for } a,b,c \in M.$$

a) Which of our ten introductory examples are ultrametric spaces?

b) For a prime number p the *p-adic absolute value* on $\mathbb{Q}$ is defined by $|0|_p = 0$ and

$$|x|_p = p^{-m} \quad \text{if } x = p^m \frac{a}{b} \text{ with } m \in \mathbb{Z} \text{ and } p \nmid ab.$$

Show that $d_p(x,y) = |x-y|_p$ turns $\mathbb{Q}$ into an ultrametric space.

c) Show that an infinite series $\sum_{n=0}^{\infty} x_n$ in $(\mathbb{Q}, d_p)$ satisfies the Cauchy criterion for convergence iff $x_n \to 0$ for $n \to \infty$.

d) Show that the metric spaces $(\mathbb{Q}, d_p)$, p prime, are not complete.

BANACH's Fixed-Point Theorem

Also called "Contraction Mapping Theorem"

Definition

Let (M, d) be a metric space. A map ("transformation") $T\colon M \to M$ is said to be a *contraction* if there exists a constant $0 \le C < 1$ such that

$$d\big(T(x), T(y)\big) \le C \cdot d(x, y) \quad \text{for all } x, y \in M.$$

Note

The condition in the definition is stronger than $d\big(T(x), T(y)\big) < d(x, y)$ for all $x, y \in M$ with $x \neq y$. For example, the transformation $T(x) = x + 1/x$ of $[1, +\infty)$ (equipped with the Euclidean metric) has this property, since

$$\left|x + \frac{1}{x} - y - \frac{1}{y}\right| = \left|x - y + \frac{y - x}{xy}\right| = \left(1 - \frac{1}{xy}\right)|x - y|$$

and $0 \le 1 - \frac{1}{xy} < 1$ for all $x, y \ge 1$. But T is not a contraction since, given $0 \le C < 1$, the numbers x, y can be chosen to satisfy $1 - \frac{1}{xy} > C$ (take, e.g., $x = 1$ and $y > (1 - C)^{-1}$).

Banach's Fixed-Point Theorem applies to contractions of complete metric spaces.

Theorem (BANACH, 1922)

Suppose (M, d) is a complete metric space and $T\colon M \to M$ a contraction.

1. *T has a unique fixed point, i.e., there exists precisely one element $x^* \in M$ satisfying $T(x^*) = x^*$.*
2. *For every point $x_0 \in M$ the sequence $x_0, x_1, x_2, \ldots$ defined recursively by $x_{n+1} = T(x_n)$ converges to x^*, and we have the error estimates*

$$d(x_n, x^*) \leq \begin{cases} \dfrac{C^n}{1-C}\, d(x_1, x_0), \\ \dfrac{C}{1-C}\, d(x_n, x_{n-1}). \end{cases}$$

(The constant C has the same meaning as on the previous slide.)

Proof.

(1) Choose $x_0 \in M$ and define the sequence (x_n) in M recursively by $x_n = T(x_{n-1}) = T^2(x_{n-2}) = \cdots = T^n(x_0)$. (Here $T^2 = T \circ T$, $T^3 = T \circ T \circ T$, etc.) For $m < n$ the triangle inequality (used successively) and the contraction property of T give

$$\begin{aligned} d(x_m, x_n) &= d\big(T(x_{m-1}), T(x_{n-1})\big) \le C\, d(x_{m-1}, x_{n-1}) \\ &\le C^2\, d(x_{m-2}, x_{n-2}) \le \cdots \le C^m\, d(x_0, x_{n-m}) \\ &\le C^m \big[d(x_0, x_1) + d(x_1, x_2) + \cdots + d(x_{n-m-1}, x_{n-m})\big] \\ &\le C^m \big[1 + C + C^2 + \cdots + C^{n-m-1}\big]\, d(x_0, x_1) \\ &= \frac{C^m - C^n}{1 - C}\, d(x_0, x_1) \\ &\le \frac{C^m}{1 - C}\, d(x_0, x_1). \end{aligned}$$

Since $0 \le C < 1$ we have $\lim_{m\to\infty} C^m = 0$.
$\Longrightarrow$ For given $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that $d(x_m, x_n) < \epsilon$ for all $n > m > N$. This means that (x_n) is a Cauchy sequence in the complete metric space (M, d) and hence converges.

Proof cont'd.

Let $x^* = \lim_{n\to\infty} x_n$.
From $d\big(T(x), T(y)\big) \le C\, d(x,y) \le d(x,y)$ it is clear that T is continuous ($\delta = \epsilon$ works).

$$\Longrightarrow T(x^*) = T\left(\lim_{n\to\infty} x_n\right) = \lim_{n\to\infty} T(x_n) = \lim_{n\to\infty} x_{n+1} = x^*.$$

Suppose that also $T(x') = x'$.

$$d(x^*, x') = d\big(T(x^*), T(x')\big) \le C\, d(x^*, x')$$

$\Longrightarrow d(x^*, x') = 0 \Longrightarrow x^* = x'$.

(2) The first assertion is clear from the proof of Part (1).
Since metrics are continuous, we get from
$d(x_m, x_n) \le \frac{C^m}{1-C}\, d(x_0, x_1)$ by passing to the limit:

$$d(x_m, x^*) = d\left(x_m, \lim_{n\to\infty} x_n\right) = \lim_{n\to\infty} d(x_m, x_n) \le \frac{C^m}{1-C}\, d(x_0, x_1).$$

The second inequality follows by applying the first inequality to the shifted sequence $x_{n-1}, x_n, x_{n+1}, \ldots$, which also has the limit x^*. □

Notes

- The "weak contraction" property $d\big(T(x), T(y)\big) < d(x, y)$ is not sufficient for the existence of a fixed point of T. A counterexample is the previously considered map $T(x) = x + 1/x$ on $[1, +\infty)$. (For this note that closed intervals in $\mathbb{R}$ form complete metric spaces of their own.)
- The 2nd error estimate $d(x_n, x^*) \leq \frac{C}{1-C}\, d(x_n, x_{n-1})$ is particularly useful, since $d(x_n, x_{n-1})$ can be read off by looking at the last two iterates. (In fact, this is what in the examples allowed us to conclude from equality of the floating-point representations of x_n and x_{n-1} that x^* has the same floating point representation.)
- In many applications, e.g. Newton's method, the map T becomes a contraction only when restricted to a suitable complete subspace M of its domain. In this case the condition $T(M) \subseteq M$, which is often difficult to verify, can be relaxed to "$x_n = T^n(x_0) \in M$ for all $n = 0, 1, 2, \ldots$"; that is, we are now looking at particular sequences.
  Specifically, if $M = \overline{\mathrm{B}_r(a)}$ is a ball and $x_0 = a$, it suffices to check the single condition $d\big(a, T(a)\big) = d(x_0, x_1) \leq (1 - C)r$. This is proved on the next slide.

Notes cont'd

- (cont'd)
  As in the proof of Banach's Theorem, this condition gives $d(x_n, a) \leq (C^{n-1} + C^{n-2} + \cdots + 1)d(x_1, x_0) \leq (1 - C^n)r \leq r$, i.e., $x_n \in \overline{\mathrm{B}_r(a)}$, and the proof of Part (1) of Banach's Theorem goes through.
  $\Longrightarrow$ (x_n) converges to $x^* \in \overline{\mathrm{B}_r(a)}$, and x^* is the unique fixed point of T in $\overline{\mathrm{B}_r(a)}$.
  Part (2), however, is not necessarily true in this setting, since for a different sequence (y_n), $y_0 \in \overline{\mathrm{B}_r(a)} \setminus \{a\}$, the contraction property of T on $\overline{\mathrm{B}_r(a)}$ doesn't exclude the possibility that some iterate y_n falls outside $\overline{\mathrm{B}_r(a)}$.
  In fact, if T doesn't map $\overline{\mathrm{B}_r(a)}$ into itself, there exists $y_0 \in \overline{\mathrm{B}_r(a)}$ such that $y_1 = T(y_0) \notin \overline{\mathrm{B}_r(a)}$.
- For the analysis of iterations on subsets of $\mathbb{R}^n$ we can use any metric on $\mathbb{R}^n$ that is strongly equivalent to the Euclidean metric $\mathrm{d_E}$ (cf. previous exercise), e.g., also d_1 or d_∞. Convergence proofs may become easier by choosing a metric different from $\mathrm{d_E}$.

Example

Consider the squaring map $T\colon \mathbb{C} \to \mathbb{C}$, $z \mapsto z^2$. (The metric on $\mathbb{C}$ is taken as the usual Euclidean one.)

$\mathbb{C} \triangleq \mathbb{R}^2$ is complete, but T is not a contraction since

$$|T(z) - T(w)| = \left|z^2 - w^2\right| = |z + w|\,|z - w|$$

and $z + w$ can have arbitarily large absolute value.
Hence we cannot use Banach's Theorem to find the fixed points of T, which are 0 and 1.

However, we can can restrict the domain of T suitably and then apply Banach's Theorem:

Suppose $0 < r < 1/2$ and let $M = \overline{\mathrm{B}_r(0)} = \{z \in \mathbb{C}; |z| \le r\}$.

- M is complete, since it is a closed subset of $\mathbb{C}$.
- For $z \in M$ we have $\left|z^2\right| = |z|^2 \le r^2 \le r$ and hence $T(M) \subseteq M$.
- For $z, w \in M$ we have $|z + w| \le |z| + |w| \le 2r < 1$.
  $\Longrightarrow T\colon M \to M$ is a contraction (take $C = 2r$).

Example (cont'd)

Hence Banach's Theorem gives that any sequence $z_n = z_{n-1}^2 = z_{n-2}^4 = \cdots = z_0^{2^n}$, $z_0 \in M$, converges to the unique fixed point of T in M, which is 0.

This is of course rather trivial and true for all $z_0 \in \mathbb{C}$ with $|z_0| < 1$.

Definition

Suppose (M, d) is a metric space, $T : M \to M$ a map and x^* a fixed point of T.

1. x^* is said to be *attracting* if there exists a neighborhood U of x^* such that any sequence $x_n = T^n(x_0)$ $(n \in \mathbb{N})$ with initial point $x_0 \in U$ converges to x^*;
2. x^* is said to be *repelling* if there exists a neighborhood U of x^* such that any sequence $x_n = T^n(x_0)$ $(n \in \mathbb{N})$ with initial point $x_0 \in U \setminus \{x^*\}$ eventually leaves U (i.e., $x_n \notin U$ for some n).

Exercise

a) Decide whether the fixed points 0 and 1 of $T : \mathbb{C} \to \mathbb{C}$, $z \to z^2$ are attracting or repelling, and prove your assertions.

b) For which $z_0 \in \mathbb{C}$ does $z_n = T^n(z_0) = z_0^{2^n}$ converge to $z^* = 1$?

Exercise

The system of equations

$$\begin{aligned} x &= 0,01\,x^2 + \sin(y) \\ y &= \cos(x) + 0,01\,y^2 \end{aligned}$$

has a unique solution (x^*, y^*) with $0,5 \leq x^* \leq 1$, $\pi/6 \leq y^* \leq 1$. Prove this statement and compute (x^*, y^*)

a) with simple fixed-point iteration;

b) with Newton Iteration.

Matrix Norms—Motivation

The quest for the norm ("absolute value") of a square matrix arises naturally during the convergence analysis of higher-dimensional Newton iteration.

Recall that this iteration has the form $\mathbf{x}^{(k+1)} = T\big(\mathbf{x}^{(k)}\big)$ with

$$T(\mathbf{x}) = \mathbf{x} - \mathrm{d}f(\mathbf{x})^{-1}\big(f(\mathbf{x})\big) = \mathbf{x} - \mathbf{J}_f(\mathbf{x})^{-1} f(\mathbf{x})$$

As in the 1-dimensional case we have $f(\mathbf{x}^*) = \mathbf{0} \iff T(\mathbf{x}^*) = \mathbf{x}^*$ (clear from the definition of T) and $f(\mathbf{x}^*) = \mathbf{0} \Longrightarrow \mathrm{d}T(\mathbf{x}^*) = 0$ (i.e., $\mathbf{J}_T(\mathbf{x}^*) = \mathbf{0} \in \mathbb{R}^{n\times n}$), as one can show with some effort.
Now we would like to show that near a zero $\mathbf{x}^*$ of f the map T defines a contraction, because then Banach's Fixed-Point Theorem implies $\lim_{k\to\infty} \mathbf{x}^{(k)} = \mathbf{x}^*$, provided only that $\mathbf{x}^{(0)}$ (or some other iterate) is sufficiently close to $\mathbf{x}^*$.

The Mean Value Theorem of n-variable calculus gives

$$\begin{aligned} T(\mathbf{x}) - T(\mathbf{y}) &= \left(\int_0^1 \mathbf{J}_T\big(\mathbf{x} + t(\mathbf{y}-\mathbf{x})\big)\,\mathrm{d}t\right)(\mathbf{x}-\mathbf{y}) \\ &= \mathbf{A}(\mathbf{x}-\mathbf{y}) \quad \text{for some matrix } \mathbf{A} = \mathbf{A}(\mathbf{x},\mathbf{y}). \end{aligned}$$

Motivation Cont'd

In the 1-dimensional case one continues with taking absolute values, viz. $|T(x) - T(y)| = |T'(\xi)|\,|x - y|$, and using continuity of T' to conclude $|T'(\xi)| \le C < 1$ provided x, y are near x^*.

Here we postulate the existence of a real number $\|\mathbf{A}\|$ such that "taking Euclidean lengths" yields the inequality

$$|T(\mathbf{x}) - T(\mathbf{y})| = |\mathbf{A}(\mathbf{x} - \mathbf{y})| \le \|\mathbf{A}\|\,|\mathbf{x} - \mathbf{y}|\,.$$

If such a norm ("absolute value") of $\mathbf{A} = \mathbf{A}(\mathbf{x}, \mathbf{y})$ exists and satisfies $\|\mathbf{A}(\mathbf{x}, \mathbf{y})\| \le C < 1$ for $\mathbf{x}, \mathbf{y}$ near $\mathbf{x}^*$ then the analysis in the 1-dimensional case carries over to the n-dimensional case.

Replacing the particular matrices $\mathbf{A}(\mathbf{x}, \mathbf{y})$ by an arbitrary matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ and setting $\mathbf{v} = \mathbf{x} - \mathbf{y}$ turns the inequality into

$$|\mathbf{Av}| \le \|\mathbf{A}\|\,|\mathbf{v}| \ \text{ for } \mathbf{v} \in \mathbb{R}^n \quad\Longleftrightarrow\quad \|\mathbf{A}\| \ge \frac{|\mathbf{Av}|}{|\mathbf{v}|} \ \text{ for } \mathbf{v} \in \mathbb{R}^n \setminus \{\mathbf{0}\}.$$

It turns out that the set $\left\{ |\mathbf{Av}| \,/\, |\mathbf{v}| \,;\, \mathbf{v} \in \mathbb{R}^n, \mathbf{v} \ne \mathbf{0} \right\}$ contains a maximum, which then clearly provides the best definition of $\|\mathbf{A}\|$.

Definition (norm of a matrix or linear map)

The *norm* of $\mathbf{A} \in \mathbb{R}^{n\times n}$ (more precisely, the *matrix norm subordinate to the Euclidean length on* $\mathbb{R}^n$) is defined as

$$\|\mathbf{A}\| = \max\left\{\frac{|\mathbf{Ax}|}{|\mathbf{x}|}; \mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}\right\} = \max\{|\mathbf{Ax}|\,; \mathbf{x} \in \mathbb{R}^n, |\mathbf{x}| = 1\}.$$

The norm of a linear map $L\colon \mathbb{R}^n \to \mathbb{R}^n$ is defined as the norm of its representing matrix, i.e., if $L(\mathbf{x}) = \mathbf{Ax}$ then

$$\|L\| = \|\mathbf{A}\| = \max\left\{\frac{|L(\mathbf{x})|}{|\mathbf{x}|}; \mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}\right\}.$$

Notes

- The second equality in the definition follows from the linearity of L:

$$\frac{|L(\mathbf{x})|}{|\mathbf{x}|} = \left|\frac{1}{|\mathbf{x}|}L(\mathbf{x})\right| = \left|L\left(\frac{\mathbf{x}}{|\mathbf{x}|}\right)\right|, \quad \text{with } \frac{\mathbf{x}}{|\mathbf{x}|} \text{ of length 1.}$$

  Since linear maps are continuous and the unit sphere in $\mathbb{R}^n$ is compact, the maximum is attained.

Notes cont'd

- The definition of $\|\mathbf{A}\|$ trivially implies $|\mathbf{Ax}| \leq \|\mathbf{A}\|\,|\mathbf{x}|$ for all $\mathbf{x} \in \mathbb{R}^n$, and similarly for the corresponding linear map L, as desired.
- The function $\mathbb{R}^{n\times n} \to \mathbb{R}$, $\mathbf{A} \mapsto \|\mathbf{A}\|$ satisfies the same axioms as the Euclidean length function:

  (N1) $\|\mathbf{A}\| \geq 0$ with equality iff $\mathbf{A} = \mathbf{0}$;
  (N2) $\|c\mathbf{A}\| = |c|\,\|\mathbf{A}\|$ for $c \in \mathbb{R}$;
  (N3) $\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$.

  Hence the definition $d(\mathbf{A}, \mathbf{B}) = \|\mathbf{A} - \mathbf{B}\|$ turns $\mathbb{R}^{n\times n}$ into a (translation-invariant) metric space.
- A further important property of $\|\ \|$ is

  (N4) $\|\mathbf{AB}\| \leq \|\mathbf{A}\|\,\|\mathbf{B}\|$. (*submultiplicativity*).
- If you wonder how to actually compute matrix norms—the answer is not easy. It uses the *Spectral Theorem for Symmetric Matrices* (cf. our Linear Algebra course), and is the subject of the next theorem. A few particular examples, which have an adhoc solution, are discussed in the exercises.

Notes cont'd

- Functions on $\mathbb{R}^n$ satisfying the same axioms as the Euclidean length are called *vector norms* and denoted in the same way. Examples are

$$\|\mathbf{x}\|_1 = |\mathbf{x}|_1 = |x_1| + |x_2| + \cdots + |x_n|,$$
$$\|\mathbf{x}\|_\infty = |\mathbf{x}|_\infty = \max\{|x_1|, |x_2|, \ldots, |x_n|\}.$$

  For the Euclidean length the notation $|\mathbf{x}|_2$ or $\|\mathbf{x}\|_2$ is frequently used in place of $|\mathbf{x}|$. With any vector norm one may associate a subordinate matrix norm in the same way as for the Euclidean length, for example $\|\mathbf{A}\|_1 = \max\{|\mathbf{A}\mathbf{x}|_1 ; \mathbf{x} \in \mathbb{R}^n, |\mathbf{x}|_1 = 1\}$ for $\mathbf{A} \in \mathbb{R}^{n\times n}$.

- Reading $\mathbf{A} \in \mathbb{R}^{n\times n}$ as an n^2-dimensional vector with entries a_{ij}, it is quite natural to consider the Euclidean length of this vector. This quantity is called *Frobenius norm* of $\mathbf{A}$ and denoted by $\|\mathbf{A}\|_{\mathrm{F}}$, i.e., one defines

$$\|\mathbf{A}\|_{\mathrm{F}} = \sqrt{\sum_{i,j=1}^{n} a_{ij}^2} \quad \text{for } \mathbf{A} \in \mathbb{R}^{n\times n}.$$

  One can show that $\mathbf{A} \mapsto \|\mathbf{A}\|_{\mathrm{F}}$ satisfies Axioms (N1)–(N4).

Example

For the 2×2 identity matrix

$$\mathbf{I}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

we have $\|\mathbf{I}_2\| = 1$, since $\mathbf{I}_2\mathbf{x} = \mathbf{x}$ and hence in particular $|\mathbf{I}_2\mathbf{x}| = |\mathbf{x}|$ for all $\mathbf{x} \in \mathbb{R}^2$. The same argument shows that $\|\mathbf{I}_n\| = 1$ in general.

Since $\mathbf{I}_2$ corresponds to the vector $(1, 0, 0, 1) \in \mathbb{R}^4$ (provided we arrange the entries of a 2×2 matrix in the order $a_{11}, a_{12}, a_{21}, a_{22}$), the Frobenius norm of $\mathbf{I}_2$ is $\|\mathbf{I}_2\|_{\mathrm{F}} = |(1, 0, 0, 1)| = \sqrt{2}$ (and, more generally, $\|\mathbf{I}_n\|_{\mathrm{F}} = \sqrt{n}$).

Exercise

Prove that $\mathbb{R}^{n\times n} \to \mathbb{R}$, $\mathbf{A} \mapsto \|\mathbf{A}\|$ satisfies (N1)–(N4).

Exercise

Prove that $\mathbb{R}^{n\times n} \to \mathbb{R}$, $\mathbf{A} \mapsto \|\mathbf{A}\|_{\mathrm{F}}$ satisfies (N1)–(N4).

Exercise

Compute the norms $\|\mathbf{A}\|$ of the following matrices $\mathbf{A} \in \mathbb{R}^{2\times 2}$ and compare them with their Frobenius norms $\|\mathbf{A}\|_{\mathrm{F}}$:

$$\begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix}, \quad \begin{pmatrix} 2 & 0 \\ 0 & -3 \end{pmatrix}, \quad \begin{pmatrix} \frac{1}{2} & \pm 1 \\ 0 & \frac{1}{2} \end{pmatrix}.$$

Exercise

Show that the norm $\|\mathbf{D}\|$ of a diagonal matrix $\mathbf{D} \in \mathbb{R}^{n\times n}$ is the largest absolute value of an entry on the diagonal.

Exercise

Show that $\|\mathbf{A}\| \le \|\mathbf{A}\|_{\mathrm{F}}$ for all matrices $\mathbf{A} \in \mathbb{R}^{n\times n}$ or, equivalently, $|\mathbf{A}\mathbf{x}| \le \|\mathbf{A}\|_{\mathrm{F}}\,|\mathbf{x}|$ for all $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{x} \in \mathbb{R}^n$.

Hint: Use $\|\mathbf{A}\| = \max\{|\mathbf{A}\mathbf{x}|\,;\mathbf{x} \in \mathbb{R}^n, |\mathbf{x}| = 1\}$ and the Cauchy-Schwarz Inequality for vectors in $\mathbb{R}^n$.

Spectral Radius and (Spectral) Norm

Definition

Suppose $\mathbf{A} \in \mathbb{C}^{n\times n}$ has eigenvalues $\lambda_1, \ldots, \lambda_n$, ordered such that $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$. The real number $|\lambda_1|$ is called *spectral radius* of $\mathbf{A}$ and denoted by $\rho(\mathbf{A})$.

Thus $\rho(\mathbf{A})$ is the radius of the smallest circle centered at $0 \in \mathbb{C}$ that contains all the eigenvalues of $\mathbf{A}$.

Theorem

For $\mathbf{A} \in \mathbb{R}^{n\times n}$ we have $\|\mathbf{A}\| = \sqrt{\rho\left(\mathbf{A}^{\mathsf{T}}\mathbf{A}\right)}$.

Notes

1. Because of this relation the matrix norm subordinate to the Euclidean length is also known as *spectral norm*.
2. The eigenvalues of $\mathbf{B} = \mathbf{A}^{\mathsf{T}}\mathbf{A}$ are nonnegative, since $\mathbf{Bx} = \lambda\mathbf{x}$, $\mathbf{x} \neq \mathbf{0}$, implies $\mathbf{x}^{\mathsf{T}}\mathbf{Bx} = \lambda\mathbf{x}^{\mathsf{T}}\mathbf{x}$,

$$\lambda = \frac{\mathbf{x}^{\mathsf{T}}\mathbf{Bx}}{\mathbf{x}^{\mathsf{T}}\mathbf{x}} = \frac{\mathbf{x}^{\mathsf{T}}\mathbf{A}^{\mathsf{T}}\mathbf{Ax}}{\mathbf{x}^{\mathsf{T}}\mathbf{x}} = \frac{(\mathbf{Ax})^{\mathsf{T}}\mathbf{Ax}}{\mathbf{x}^{\mathsf{T}}\mathbf{x}} = \frac{|\mathbf{Ax}|^2}{|\mathbf{x}|^2}.$$

Thus $\rho\left(\mathbf{A}^{\mathsf{T}}\mathbf{A}\right)$ is the largest eigenvalue of $\mathbf{A}^{\mathsf{T}}\mathbf{A}$.

Proof of the theorem.

Note 2 on the preceding slide shows

$$\|\mathbf{A}\|^2 = \max\left\{|\mathbf{Ax}|^2 ; \mathbf{x} \in \mathbb{R}^n, |\mathbf{x}| = 1\right\} \geq \lambda_1 = \rho\left(\mathbf{A}^\mathsf{T}\mathbf{A}\right).$$

(Normalizing $\mathbf{x}$ to unit length, the note reads $\lambda = |\mathbf{Ax}|^2$.)

Since $\mathbf{B} = \mathbf{A}^\mathsf{T}\mathbf{A}$ is symmetric, there exists an orthonormal basis $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ of $\mathbb{R}^n$ consisting of eigenvectors of $\mathbf{B}$; cf. the Spectral Theorem. We may assume $\mathbf{Bu}_j = \lambda_j \mathbf{u}_j$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Then for $\mathbf{x} = \sum_{i=1}^n \alpha_i \mathbf{u}_i$, $\alpha_i \in \mathbb{R}$, we get

$$|\mathbf{x}|^2 = \mathbf{x}^\mathsf{T}\mathbf{x} = \sum_{i,j=1}^n \alpha_i \alpha_j \mathbf{u}_i^\mathsf{T}\mathbf{u}_j = \sum_{i=1}^n \alpha_i^2,$$

$$|\mathbf{Ax}|^2 = \mathbf{x}^\mathsf{T}\mathbf{Bx} = \sum_{i,j=1}^n \alpha_i \alpha_j \mathbf{u}_i^\mathsf{T}\mathbf{Bu}_j = \sum_{i,j=1}^n \alpha_i \alpha_j \lambda_j \mathbf{u}_i^\mathsf{T}\mathbf{u}_j = \sum_{i=1}^n \alpha_i^2 \lambda_i.$$

If $|\mathbf{x}| = 1$ then $|\mathbf{Ax}|^2 \leq \sum_{i=1}^n \alpha_i^2 \lambda_1 = \left(\alpha_1^2 + \cdots + \alpha_n^2\right)\lambda_1 = \lambda_1$.

$\Longrightarrow \|\mathbf{A}\|^2 \leq \lambda_1 = \rho\left(\mathbf{A}^\mathsf{T}\mathbf{A}\right)$

In all we have shown $\|\mathbf{A}\|^2 = \rho\left(\mathbf{A}^\mathsf{T}\mathbf{A}\right)$, as claimed. □

Math 285
Introduction to Differential Equations

Thomas Honold

ZJU-UIUC Institute

Spring Semester 2024

Outline

Today's Lecture:

The Uniqueness Theorem

We will prove the Existence and Uniqueness Theorem for solutions of 1st-order ODE's in a more general form than [BDM17, Theorem 2.8.1]. The generalization to n-dimensional ODE systems $\mathbf{y}' = f(t, \mathbf{y})$ will enable us to conclude from it a corresponding theorem for higher-order scalar ODE's. Further, we will relax the assumption "$f(t, \mathbf{y})$ has continuous partial derivatives with respect to the variables in $\mathbf{y}$" to "$f(t, \mathbf{y})$ satisfies locally a so-called Lipschitz condition with respect to $\mathbf{y}$". This will allow us to apply the Existence and Uniqueness Theorem to certain ODE's that are not covered by [BDM17, Theorem 2.8.1] but still important in Engineering Mathematics.

Definition

Suppose $f\colon D \to \mathbb{R}^n$, $D \subseteq \mathbb{R} \times \mathbb{R}^n$, is a map.

1. We say that $f = f(t, \mathbf{y})$ satisfies a *Lipschitz condition with respect to* $\mathbf{y}$ if there exists a constant $L \geq 0$ (the corresponding *Lipschitz constant*) such that

$$|f(t, \mathbf{y}_1) - f(t, \mathbf{y}_2)| \leq L\,|\mathbf{y}_1 - \mathbf{y}_2| \quad \text{for all } (t, \mathbf{y}_1), (t, \mathbf{y}_2) \in D.$$

2. We say that f satisfies *locally* a Lipschitz condition with respect to $\mathbf{y}$ if every point $(t, \mathbf{y}) \in D$ has a neighborhood $D' \subseteq D$ in which (1) holds with a constant L (which may depend on the particular point).

Condition (2), together with continuity of f as a function of $n+1$ variables, will be taken as the premise of the Existence and Uniqueness Theorem. The next proposition shows that these conditions are implied by those used in [BDM17, Theorem 2.8.1], so that our Existence and Uniqueness Theorem covers (i.e., implies) that in the textbook.

Proposition

Suppose $D \subseteq \mathbb{R} \times \mathbb{R}^n$ is an open set and $f : D \to \mathbb{R}^n$ has continuous (as $(n+1)$-variable functions!) partial derivatives with respect to the variables $\mathbf{y} = (y_1, \ldots, y_n)$. Then f satisfies locally a Lipschitz condition with respect to $\mathbf{y}$.

Proof.

Let $(a, \mathbf{b}) \in D$. Since D is open, there exists $r > 0$ such that

$$V = \{(t, \mathbf{y}); |t - a| \leq r \wedge |\mathbf{y} - \mathbf{b}| \leq r\} \subseteq D.$$

V is a compact subset of D.

The Mean Value Theorem (integral form) of Calculus III, applied to $\mathbf{y} \mapsto f(t, \mathbf{y})$, gives for $(t, \mathbf{y}_1), (t, \mathbf{y}_2) \in V$ the identity

$$f(t, \mathbf{y}_1) - f(t, \mathbf{y}_2) = \left(\int_0^1 \mathbf{J}_{f,\mathbf{y}}\big(t, \mathbf{y}_1 + s(\mathbf{y}_2 - \mathbf{y}_1)\big)\, \mathrm{d}s \right) (\mathbf{y}_1 - \mathbf{y}_2)$$

with $\mathbf{J}_{f,\mathbf{y}}(t, \mathbf{y}) = \left(\frac{\partial f_i}{\partial y_j}(t, \mathbf{y}) \right)_{1 \leq i,j \leq n}$ ("partial Jacobi matrix" of f).

Proof cont'd.

Since the entries of the $n \times n$ matrix $\left(\frac{\partial f_i}{\partial y_j}(t, \mathbf{y})\right)$ are continuous functions of $(t, \mathbf{y})$, there exists a constant M such that $\left|\frac{\partial f_i}{\partial y_j}(t, \mathbf{y})\right| \le M$ for all $(t, \mathbf{y}) \in V$ and all i, j. This implies $\left\|\mathbf{J}_{f,\mathbf{y}}(t, \mathbf{y})\right\|_{\mathrm{F}} \le nM$ for all $(t, \mathbf{y}) \in V$.

The matrix $\mathbf{A} = \mathbf{A}(t, \mathbf{y}_1, \mathbf{y}_2)$ appearing in the Mean Value Theorem is obtained by averaging $\mathbf{J}_{f,\mathbf{y}}(t, \mathbf{y})$ over the line segment $[\mathbf{y}_1, \mathbf{y}_2]$ and is subject to the same bound. More precisely, we have

$$\begin{aligned} |f(t, \mathbf{y}_1) - f(t, \mathbf{y}_2)| &\le \|\mathbf{A}\| \, |\mathbf{y}_1 - \mathbf{y}_2| \\ &\le \|\mathbf{A}\|_{\mathrm{F}} \, |\mathbf{y}_1 - \mathbf{y}_2| \qquad \text{(cf. exercise)} \\ &\le \sqrt{n^2 M^2} \, |\mathbf{y}_1 - \mathbf{y}_2| = nM \, |\mathbf{y}_1 - \mathbf{y}_2| \,, \end{aligned}$$

since $|a_{ij}| = \left|\int_0^1 \frac{\partial f_i}{\partial y_j}\big(t, \mathbf{y}_1 + s(\mathbf{y}_2 - \mathbf{y}_1)\big) \, \mathrm{d}s\right| \le (1 - 0) \cdot M = M$.

Thus we can take $L = nM$ as the desired local Lipschitz constant and the corresponding neighborhood of $(a, \mathbf{b})$ as V.

□

Remark (LIPSCHITZ-continuity of 1-variable functions)

$f\colon [a,b] \to \mathbb{R}$ is said to be *Lipschitz-continuous* or to satisfy a *Lipschitz condition* with *Lipschitz constant* L if there exists $L > 0$ such that

$$|f(x) - f(y)| \leq L\,|x - y| \quad \text{for all } x, y \in [a,b].$$

The name "Lipschitz-continuity" comes from the fact that this property implies that f is uniformly continuous (take $\delta = \epsilon/L$ as response to ϵ).

The preceding Proposition is a multi-variable generalization of the following fact:

> *Every* C^1*-function on a compact intervall* $[a,b] \subset \mathbb{R}$ *is Lipschitz-continuous.*

For the (much easier) proof define $L = \max\{|f'(x)|\,; a \leq x \leq b\}$ and use the Mean Value Theorem:

$$|f(x) - f(y)| = |f'(\xi)(x-y)| \leq L\,|x-y|\,.$$

Remark (cont'd)

A C^1-function f on an arbitrary interval $I \subseteq \mathbb{R}$ need not be Lipschitz-continuous, but we still get Lipschitz-continuity on every compact subinterval $[a, b] \subseteq I$. Equivalently, every point $x \in I$ has a neighborhood $(x - \delta, x + \delta)$, $\delta > 0$, such that f is Lipschitz-continuous on $I \cap (x - \delta, x + \delta)$.

The property of uniform continuity is weaker than Lipschitz-continuity. For example, $x \mapsto \sqrt{x}$ is uniformly continuous on $[0, \infty)$, but not Lipschitz-continuous.

Remark (Metric spaces and Lipschitz-continuity)

The concept of Lipschitz-continuity makes sense for maps between arbitrary metric spaces. If (M, d) and (M', d') are metric spaces and $T\colon (M, d) \to (M', d')$ is a map, we call T *Lipschitz-continuous* if there exists $L > 0$ such that

$$d'\big(T(x), T(y)\big) \leq L\, d(x, y) \quad \text{for all } x, y \in M.$$

In fact much of the preceding discussion, including Banach's Fixed Point Theorem, is related to this concept. As an example, observe that $T\colon M \to M$ is a contraction iff it satisfies a Lipschitz condition with Lipschitz constant $L < 1$.

Now we can state and prove the Existence and Uniqueness Theorems for explicit 1st-order ODE systems $\mathbf{y}' = f(t, \mathbf{y})$.
For implicit ODE's there are no such Theorems, and hence an implicit ODE must first be converted to explicit form before drawing any conclusions about existence/uniqueness of solutions.

A crucial ingredient for both proofs will be the observation made earlier, that solutions of the differential equation $\mathbf{y}' = f(t, \mathbf{y})$ can be characterized as solutions of a related integral equation:

Observation (recalled)

Suppose $D \subseteq \mathbb{R} \times \mathbb{R}^n$ is open, $f\colon D \to \mathbb{R}^n$ continuous and $(t_0, \mathbf{y}_0) \in D$. A continuous function (curve) $\phi\colon I \to \mathbb{R}^n$ with $\big(t, \phi(t)\big) \in D$ for $t \in I$ solves the IVP $\mathbf{y}' = f(t, \mathbf{y})) \wedge \mathbf{y}(t_0) = \mathbf{y}_0$ iff

$$\phi(t) = \mathbf{y}_0 + \int_{t_0}^{t} f\big(\tau, \phi(\tau)\big)\mathrm{d}\tau \quad \text{for } t \in I.$$

By the Fundamental Theorem of Calculus, this *integral equation* implies that ϕ is differentiable with $\phi'(t) = f\big(t, \phi(t)\big)$, and of course $\phi(t_0) = \mathbf{y}_0$. For the converse integrate $\phi'(t) = f\big(t, \phi(t)\big)$.

Theorem (Uniqueness Theorem)

Suppose that $D \subseteq \mathbb{R} \times \mathbb{R}^n$ is open and that $f\colon D \to \mathbb{R}^n$ is continuous and satisfies locally a Lipschitz condition with respect to $\mathbf{y}$*. If $\phi, \psi\colon I \to \mathbb{R}^n$ are solutions of an IVP*

$$\mathbf{y}' = f(t, \mathbf{y}) \wedge \mathbf{y}(t_0) = \mathbf{y}_0, \qquad (t_0, \mathbf{y}_0) \in D,$$

then $\phi(t) = \psi(t)$ for all $t \in I$.

The key step in the proof is the following lemma, which says that the set $A \subseteq I$ of arguments t where ϕ and ψ agree is open in I.

Lemma

Suppose $a \in I$ is such that $\phi(a) = \psi(a)$ (e.g., take $a = t_0$). Then there exists $\epsilon > 0$ such that $\phi(t) = \psi(t)$ for all $t \in I \cap [a - \epsilon, a + \epsilon]$.

Proof.

Integrating the two equations $\phi'(t) = f\big(t, \phi(t)\big)$, $\psi'(t) = f\big(t, \psi(t)\big)$ and using $\phi(a) = \psi(a) = \mathbf{b}$, say, we obtain

$$\begin{aligned}\phi(t) - \psi(t) &= \mathbf{b} + \int_a^t f\big(\tau, \phi(\tau)\big)\mathrm{d}\tau - \mathbf{b} - \int_a^t f\big(\tau, \psi(\tau)\big)\mathrm{d}\tau \\ &= \int_a^t f\big(\tau, \phi(\tau)\big) - f\big(\tau, \psi(\tau)\big)\mathrm{d}\tau.\end{aligned}$$

By assumption, there exists a neighborhood V of $(a, \mathbf{b})$ on which f satisfies a Lipschitz condition with respect to $\mathbf{y}$. Further, since ϕ and ψ are continuous in a, there exists $\delta > 0$ such that $\big(\tau, \phi(\tau)\big) \in V$ and $\big(\tau, \psi(\tau)\big) \in V$ for all $\tau \in I \cap [a - \delta, a + \delta]$. Thus we have, denoting the Lipschitz constant by L as usual,

$$\big|f\big(\tau, \phi(\tau)\big) - f\big(\tau, \psi(\tau)\big)\big| \leq L\,|\phi(\tau) - \psi(\tau)| \quad \text{for } \tau \in I \cap [a - \delta, a + \delta].$$

Proof cont'd.

$$\Longrightarrow\ |\phi(t)-\psi(t)| \le \begin{cases} L\int_a^t |\phi(\tau)-\psi(\tau)|\,\mathrm{d}\tau & \text{for } t\in I\cap[a,a+\delta], \\ L\int_t^a |\phi(\tau)-\psi(\tau)|\,\mathrm{d}\tau & \text{for } t\in I\cap[a-\delta,a]. \end{cases}$$

Now we set $M(t) := \sup\{|\phi(\tau)-\psi(\tau)|\,;\tau \text{ between } a \text{ and } t\}$ for $t\in I\cap[a-\delta,a+\delta]$.

$$\Longrightarrow\ |\phi(t)-\psi(t)| \le L\,|t-a|\,M(t)$$

for all such t. Replacing t by any t' between a and t, we also get

$$|\phi(t')-\psi(t')| \le L\,|t'-a|\,M(t') \le L\,|t-a|\,M(t).$$

Taking the supremum over all t' between a and t gives

$$M(t) \le L\,|t-a|\,M(t) \quad \text{for } t\in I\cap[a-\delta,a+\delta].$$

Setting $\epsilon = \min\left\{\delta, \frac{1}{2L}\right\}$, this implies $M(t)\le \frac{1}{2}M(t)$ for all $t\in I\cap[a-\epsilon,a+\epsilon]$. Clearly this can hold only if $M(t)=0$ for all $t\in I\cap[a-\epsilon,a+\epsilon]$, and the proof of the lemma is complete. □

Note

It was necessary to use "$t \in I \cap [a-\delta, a+\delta]$", etc., throughout the proof, because a may be an endpoint of I (left or right endpoint). In such a case solutions are defined only on one of the intervals $[a-\delta, a]$, $[a, a+\delta]$, etc.

Proof of the Uniqueness Theorem.

We prove the theorem by contradiction.
Let $A = \{t \in I; \phi(t) = \psi(t)\}$, $N = I \setminus A$, and suppose that $N \neq \emptyset$.
Since $\phi(t_0) = \psi(t_0)$, we have $t_0 \notin N$. Hence there are the following two cases to consider.

Case 1: There exists $t_1 \in N$ with $t_1 > t_0$.
Define t_2 as the infimum of the (non-empty and bounded from below) set $N \cap [t_0, +\infty)$. Then obviously $t_2 \in [t_0, t_1] \subseteq I$.
We claim that $\phi(t_2) = \psi(t_2)$. If $t_2 = t_0$ this is trivial. Otherwise $t_2 > t_0$ and $\phi(t) = \psi(t)$ for all $t \in [t_0, t_2)$. Continuity of ϕ, ψ then implies $\phi(t_2) = \lim_{t \uparrow t_2} \phi(t) = \lim_{t \uparrow t_2} \psi(t) = \psi(t_2)$.
Now the lemma yields $\epsilon > 0$ such that $[t_2, t_2 + \epsilon] \subseteq A$. This obviously contradicts the definition of t_2.

Case 2: There exists $t_1 \in N$ with $t_1 < t_0$.
For this case a contradiction is derived in a similar way. □

Example

Consider the ODE $y' = \sqrt{|y|}$.

We have seen earlier that this ODE has, among others, the solutions $y_1 \equiv 0$ and

$$y_2(t) = \begin{cases} \frac{1}{4}(t - t_0)^2 & \text{if } t \geq t_0, \\ -\frac{1}{4}(t - t_0)^2 & \text{if } t \leq t_0, \end{cases}$$

where $t_0 \in \mathbb{R}$ is arbitrary. We have $y_1(t_0) = y_2(t_0) = 0$, but $y_1(t) \neq y_2(t)$ for $t \neq t_0$.

This doesn't contradict the Uniqueness Theorem, since $f(t, y) = \sqrt{|y|}$ has partial derivative

$$\frac{\partial f}{\partial y}(t, y) = \begin{cases} \frac{1}{2\sqrt{y}} & \text{if } y > 0, \\ -\frac{1}{2\sqrt{-y}} & \text{if } y < 0, \end{cases}$$

and hence doesn't satisfy locally a Lipschitz condition at any point $(t_0, 0)$.

On the other hand, f satisfies locally a Lipschitz condition at every point (t_0, y_0) with $y_0 \neq 0$, and hence solutions of the IVP $y' = \sqrt{|y|} \wedge y(t_0) = y_0 \neq 0$ are unique as long as they stay away from the t-axis $y = 0$.

The next example, which a student contributed, shows that $f(t, \mathbf{y})$ need not satisfy a local Lipschitz condition with respect to $\mathbf{y}$ in order for solutions of IVP's $\mathbf{y}' = f(t, \mathbf{y}) \wedge \mathbf{y}(t_0) = \mathbf{y}_0$ to be unique.

Example

Consider the ODE $y' = \begin{cases} y \ln|y| & \text{if } y \neq 0, \\ 0 & \text{if } y = 0. \end{cases}$

The solutions are $y(t) \equiv 0$ and $y(t) = \pm e^{c\,e^t}$, $c \in \mathbb{R}$, as is easily derived using the standard machinery for autonomous/separable equations and observing that no non-constant solution can attain a value 0, ± 1 (the constant solutions). If it did, there would exist $t_1 \in \mathbb{R}$ and $c \in \mathbb{R} \setminus \{0\}$ such that $\lim_{t \to t_1} e^{c\,e^t} = e^{c\,e^{t_1}} \in \{0, \pm 1\}$, which is impossible. Thus all associated IVP's have a unique solution.

But $f(t, y) = y \ln|y|$ doesn't satisfy a local Lipschitz condition at any point on the t-axis $y = 0$, because, e.g., for $0 < y_1 < y_2$ we have

$$f(t, y_2) - f(t, y_1) = \frac{\partial f}{\partial y}(t, \eta)(y_2 - y_1) = (1 + \ln \eta)(y_2 - y_1)$$

for some $\eta \in (y_1, y_2)$ by the Mean Value Theorem of Calculus I, and $1 + \ln \eta \to -\infty$ for $\eta \downarrow 0$.

Remark

"Continuity of $f(t, \mathbf{y})$" and "local Lipschitz condition with respect to $\mathbf{y}$" has been adopted as premise in both the Uniqueness Theorem and the Existence Theorem (cf. subsequent slide), because under these assumptions the theorems are fairly easy to prove and the assumptions are sufficiently general to cover most applications. At the cost of more difficult proofs, the assumptions can be relaxed. For example, the conclusion of the Existence Theorem remains true if one merely stipulates that $f(t, \mathbf{y})$ is continuous (PEANO*'s Existence Theorem*), and the conclusion of the Uniqueness Theorem remains true if the Lipschitz condition is relaxed to

$$|f(t, \mathbf{y}_1) - f(t, \mathbf{y}_2)| \leq L\,|\mathbf{y}_1 - \mathbf{y}_2|\,\ln|\mathbf{y}_1 - \mathbf{y}_2|$$

(a consequence of OSGOOD*'s Condition*; cf. the literature).

Exercise

Does the Uniqueness Theorem apply to the ODE $y' = |y|$?
If you are unsure, solve the ODE directly.

The Existence Theorem

Theorem (Picard-Lindelöf)

Suppose $D \subseteq \mathbb{R} \times \mathbb{R}^n$ is open and $f\colon D \to \mathbb{R}^n$, $(t, \mathbf{y}) \mapsto f(t, \mathbf{y})$ is a continuous function which satisfies on D locally a Lipschitz condition with respect to $\mathbf{y}$. Then for every $(t_0, \mathbf{y}_0) \in D$ there exists an interval I containing t_0 as an inner point and a solution $\phi\colon I \to \mathbb{R}^n$ of the IVP $\mathbf{y}' = f(t, \mathbf{y}) \wedge \mathbf{y}(t_0) = \mathbf{y}_0$.

Proof.

By our previous observation it suffices to construct a continuous function $\phi^* : [t_0 - \epsilon, t_0 + \epsilon] \to \mathbb{R}^n$ satisfying $T\phi^* = \phi^*$, where T is the "operator"

$$(T\phi)(t) = \mathbf{y}_0 + \int_{t_0}^{t} f\big(\tau, \phi(\tau)\big)\mathrm{d}\tau.$$

Right now T is not well-defined, because we haven't yet specified a suitable domain from which the function ϕ is taken. But this will be cured in a moment.

Proof cont'd.

Our goal is to apply Banach's Fixed Point Theorem to T.

By assumption there exists $r > 0$ such that the compact set

$$V = \{(t, \mathbf{y}) \in \mathbb{R} \times \mathbb{R}^n; |t - t_0| \le r, |\mathbf{y} - \mathbf{y}_0| \le r\}$$

is contained in D and f satisfies a Lipschitz condition with respect to $\mathbf{y}$ on V. Denote the corresponding Lipschitz constant by L. Further, since f is continuous, there exists $M > 0$ such that $|f(t, \mathbf{y})| \le M$ on V.

Now let $\epsilon = \min\{r, r/M, 1/(2L)\}$ and define $\mathscr{M}$ as the set of all continuous functions $\phi\colon [t_0 - \epsilon, t_0 + \epsilon] \to \mathbb{R}^n$ satisfying $|\phi(t) - \mathbf{y}_0| \le r$ for all $t \in [t_0 - \epsilon, t_0 + \epsilon]$), and hence $\big(t, \phi(t)\big) \in V$ for such t.

$\mathscr{M}$ is equipped with the metric of uniform convergence, i.e.

$$d_\infty(\phi, \psi) = \max\big\{|\phi(t) - \psi(t)|\,; t_0 - \epsilon \le t \le t_0 + \epsilon\big\} = \|\phi - \psi\|_\infty\,,$$

where $\|\phi\|_\infty = \max\big\{|\phi(t)|\,; t_0 - \epsilon \le t \le t_0 + \epsilon\big\}$.
The metric space $(\mathscr{M}, d_\infty)$ is complete, since it is a closed subspace of $\mathrm{C}\big([t_0 - \epsilon, t_0 + \epsilon]\big)$ (in fact the closed ball $\overline{\mathrm{B}_r(\mathbf{y}_0)}$).

Proof cont'd.

In order to apply Banach's Theorem, it remains to show $T(\mathscr{M}) \subseteq \mathscr{M}$ and that T defines a contraction of $(\mathscr{M}, d_\infty)$.

Let $\phi \in \mathscr{M}$ and $\psi = T\phi$. For $t_0 - \epsilon \le t \le t_0 + \epsilon$ we have

$$|\psi(t) - \mathbf{y}_0| = \left| \int_{t_0}^{t} f\big(\tau, \phi(\tau)\big) \mathrm{d}\tau \right| \le \pm \int_{t_0}^{t} \big| f\big(\tau, \phi(\tau)\big) \big| \, \mathrm{d}\tau$$
$$\le \epsilon M \le r.$$

(The minus sign is necessary to account for the case $t < t_0$, in which we rather mean $\int_t^{t_0} f\big(\tau, \phi(\tau)\big) \mathrm{d}\tau$.) This shows $T(\mathscr{M}) \subseteq \mathscr{M}$.

Let $\phi_1, \phi_2 \in \mathscr{M}$ and $\psi_1 = T\phi_1$, $\psi_2 = T\phi_2$.

$$|\psi_1(t) - \psi_2(t)| = \left| \int_{t_0}^{t} f\big(\tau, \phi_1(\tau)\big) - f\big(\tau, \phi_2(\tau)\big) \mathrm{d}\tau \right|$$
$$\le \pm \int_{t_0}^{t} \big| f\big(\tau, \phi_1(\tau)\big) - f\big(\tau, \phi_2(\tau)\big) \big| \, \mathrm{d}\tau$$
$$\le \pm \int_{t_0}^{t} L \, |\phi_1(\tau) - \phi_2(\tau)| \, \mathrm{d}\tau \le \epsilon L \, \|\phi_1 - \phi_2\|_\infty$$
$$\le \tfrac{1}{2} \, \|\phi_1 - \phi_2\|_\infty$$

Proof cont'd.

Maximizing over $t \in [t_0 - \epsilon, t_0 + \epsilon]$ gives

$$\|\psi_1 - \psi_2\|_\infty \leq \tfrac{1}{2} \|\phi_1 - \phi_2\|_\infty, \quad \text{i.e.,} \quad d_\infty(T\phi_1, T\phi_2) \leq \tfrac{1}{2} d_\infty(\phi_1, \phi_2).$$

This shows that $T : \mathscr{M} \to \mathscr{M}$ is a contraction with $C = 1/2$. Now Banach's Theorem can be applied and yields $\phi^* \in \mathscr{M}$ with $T\phi^* = \phi^*$. This function ϕ^* is the desired solution of the given IVP. □

Notes

- In the proof of the Existence Theorem (and similarly in the proof of the key lemma to the Uniqueness Theorem) we have used estimates of the form $\cdots \leq \pm \int_{t_0}^{t} |\ldots| \,\mathrm{d}\tau$, where the minus sign is chosen in the case $t < t_0$ to make the right-hand side non-negative. This rather awkward notation, or the even more awkward $\cdots \leq \left| \int_{t_0}^{t} |\ldots| \,\mathrm{d}\tau \right|$ used in some books, can be avoided if we interpret $\int_{t_0}^{t}$ in these cases as the Lebesgue integral over the interval with endpoints t_0 and t (which can be either $[t_0, t]$ or $[t, t_0]$).
- Likewise, in the proof of both theorems we have used estimates of the form $\left| \int_a^b \phi(t)\,\mathrm{d}t \right| \leq \int_a^b |\phi(t)|\,\mathrm{d}t$ with $\phi\colon [a,b] \to \mathbb{R}^n$ continuous. For $n > 1$ the vertical bars refer to the Euclidean length in $\mathbb{R}^n$ rather than the absolute value on $\mathbb{R}$, and the inequality does not follow from the 1-dimensional integration theory developed in Calculus II. A proof can be found in Exercise H17 of Homework 3, Calculus III (Fall 2022). Alternatively, approximate ϕ by vector-valued step functions and check that for such functions the inequality reduces to the triangle inequality for the Euclidean length in $\mathbb{R}^n$.

Notes cont'd

- Banach's Theorem also gives that a solution of the IVP can be obtained as the limit function $\phi(t) = \lim_{k\to\infty} \phi_k(t)$ of the "Picard-Lindelöf iteration"

$$\phi_0(t) \equiv \mathbf{y}_0, \quad \phi_{k+1}(t) = \mathbf{y}_0 + \int_{t_0}^{t} f\big(\tau, \phi_k(\tau)\big)\,\mathrm{d}\tau, \quad k = 0, 1, 2, \ldots,$$

  because certainly the constant function $\phi_0(t) \equiv \mathbf{y}_0$ is in $\mathscr{M}$ (whatever the chosen domain $[t_0 - \epsilon, t_0 + \epsilon]$ is). This is illustrated in the following example.

Example

We apply Picard-Lindelöf iteration to construct a solution of the IVP $y' = 2ty \wedge y(0) = y_0$; cf. our introductory Example **??**.

Here the iteration takes the form

$$\phi_{k+1}(t) = y_0 + 2\int_0^t \tau\phi_k(\tau)\mathrm{d}\tau.$$

We obtain

$$\phi_1(t) = y_0 + 2\int_0^t \tau y_0 \mathrm{d}\tau = y_0 + 2y_0\int_0^t \tau\mathrm{d}\tau = y_0(1+t^2),$$

$$\phi_2(t) = y_0 + 2\int_0^t \tau y_0(1+\tau^2)\mathrm{d}\tau = y_0(1+t^2+t^4/2)$$

and in general, using induction,

$$\phi_k(t) = y_0\left(1 + t^2 + \frac{t^4}{2!} + \frac{t^6}{3!} + \cdots + \frac{t^{2k}}{k!}\right).$$

The limit function is $\phi(t) = y_0\sum_{k=0}^{\infty}\frac{t^{2k}}{k!} = y_0\mathrm{e}^{t^2}$, the already known solution.

Example (cont'd)

Since the functions ϕ_k have maximal domain $\mathbb{R}$ and converge uniformly to $\phi(t) = y_0 \mathrm{e}^{t^2}$ on every compact subinterval of $\mathbb{R}$ (known from the theory of power series), we can conclude without checking the assumptions of the Existence Theorem, or direct verification, that ϕ solves the given IVP on $\mathbb{R}$. This is done as follows:

For any $R > 0$ and any continuous functions $\psi_1, \psi_2 \colon [-R, R] \to \mathbb{R}$ we have

$$|(T\psi_1)(t) - (T\psi_2)(t)| = \left| 2\int_0^t \tau\big(\psi_1(\tau) - \psi_2(\tau)\big)\mathrm{d}\tau \right| \leq R^2 \|\psi_1 - \psi_2\|_\infty$$

on $[-R, R]$ and hence $\|T\psi_1 - T\psi_2\|_\infty \leq R^2 \|\psi_1 - \psi_2\|_\infty$, where $\|\psi\|_\infty = \|\psi\|_{\infty,R} = \max\big\{|\psi(t)|\,; -R \leq t \leq R\big\}$. This shows that T defines a continuous operator on $\mathrm{C}\big([-R, R]\big)$ and implies

$$T\phi = T\left(\lim_{k\to\infty} \phi_k\right) = \lim_{k\to\infty} T\phi_k = \lim_{k\to\infty} \phi_{k+1} = \phi,$$

which in turn implies that ϕ solves the given IVP on $I = [-R, R]$, as we have seen. Letting $R \to +\infty$ then shows the same for $I = \mathbb{R}$.

Exercise

Use Picard-Lindelöf iteration to compute the solution $\phi = (\phi_1, \phi_2)^T$ of the system

$$\begin{pmatrix} y_1' \\ y_2' \end{pmatrix} = \begin{pmatrix} -y_2 \\ y_1 \end{pmatrix}$$

with initial condition $\phi(0) = (1, 0)^T$.

Exercise

Suppose that $f\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is continuous and satisfies locally a Lipschitz condition, and that

$$f(-t, y) = -f(t, y) \quad \text{for all } (t, y) \in \mathbb{R}^2.$$

Show that any solution $\phi\colon [-r, r] \to \mathbb{R}$, $r > 0$, of $y' = f(t, y)$ is its own mirror image with respect to the y-axis.

Exercise (hard)

Suppose $I \subseteq \mathbb{R}$ is an interval and $f\colon I \times \mathbb{R}^n \to \mathbb{R}^n$ is continuous and satisfies (globally) a Lipschitz condition with Lipschitz constant L. For any two solutions $\phi, \psi\colon I \to \mathbb{R}^n$ of $y' = f(t, y)$ and $t_0 \in I$ show that $|\phi(t) - \psi(t)| \leq \delta e^{L|t-t_0|}$ on I, where $\delta = |\phi(t_0) - \psi(t_0)|$.

Maximal Solutions

Recall that domains of solutions of ODE's must be intervals in $\mathbb{R}$ of positive (possibly infinite) length and may or may not contain their boundary point(s).

Definition

A solution $\phi\colon I \to \mathbb{R}$ of $y' = f(t, y)$ is said to be *maximal* (or *non-extendable*) if there is no solution $\psi\colon J \to \mathbb{R}$ with $J \supsetneq I$ and $\psi(t) = \phi(t)$ for $t \in I$.

This definition extends in the obvious way to higher-order ODE's, ODE systems (both explicit and implicit ones).

Corollary

Under the assumptions of the Existence and Uniqueness Theorem,

1. *for every $(t_0, \mathbf{y}_0) \in D$ there exists a unique maximal solution $\phi_0\colon I_0 \to \mathbb{R}^n$ of the IVP $\mathbf{y}' = f(t, \mathbf{y}) \wedge \mathbf{y}(t_0) = \mathbf{y}_0$;*
2. *I_0 is open in $\mathbb{R}$, and for every end point e of I_0 (if any) the solution curve $\{(t, \phi_0(t)); t \in I_0\}$ comes arbitrarily close to the boundary of D when $t \to e$.*

Maximal Solutions Cont'd

Note

The precise mathematical definition of "comes arbitrarily close to the boundary" is that, e.g., if a is the left end point of I_0 and $t_0 \in I_0$ then $\{(t, \phi_0(t)); a < t \le t_0\}$ is not contained in a compact subset of D. This means that there exists a sequence $t_k \downarrow a$ such that either $|\phi_0(t_k)| \to \infty$ or there exists $\mathbf{b} \in \mathbb{R}^n$ such that $(a, \mathbf{b}) \notin D$ and $\lim_{k\to\infty} \phi_0(t_k) = \mathbf{b}$.

Proof of the corollary.

(1) Let $I_0 = \bigcup I$ be the union of all domains of solutions $\phi\colon I \to \mathbb{R}^n$ of the given IVP and define $\phi_0\colon I_0 \to \mathbb{R}^n$ by $\phi_0(t) = \phi(t)$ if t is contained in the domain of ϕ. Clearly I_0 is an interval containing t_0. If $t \in I_0$ and ϕ_1, ϕ_2 are solutions of the IVP defined at t, we must have $\phi_1(t) = \phi_2(t)$ (apply the Uniqueness Theorem with $I = [t_0, t]$ or $[t, t_0]$). Hence ϕ_0 is well-defined, and it cleary solves the IVP. Since the domain of ϕ_0 contains the domains of all solutions, ϕ_0 is maximal. Finally the Uniqueness Theorem gives that there cannot be another maximal solution (whose domain would necessarily be I_0).

Proof cont'd.

(2) First we show that I_0 is open.
Suppose by contradiction, e.g., that I_0 contains its left end point a. Then $\big(a, \phi_0(a)\big) \in D$, and the Existence Theorem provides us with a solution $\phi\colon [a-\epsilon, a+\epsilon] \to \mathbb{R}^n$ of the IVP $\mathbf{y}' = f(t, \mathbf{y}) \wedge y(a) = \phi_0(a)$ for some $\epsilon > 0$. By the Uniqueness Theorem, $\phi(t) = \phi_0(t)$ for $t \in [a, a+\epsilon]$. Hence, using the definition in terms of ϕ on $[a-\epsilon, a)$, we can prolong ϕ_0 to a solution on $[a-\epsilon, a) \cup I_0$, which is an interval strictly containing I_0; contradiction.

For a proof of the remaining assertion, assume $e = a$ and by contradiction that $\big\{\big(t, \phi_0(t)\big); a < t \le t_0\big\}$ is contained in a compact subset $C \subset D$. To derive the desired contradiction, it then suffices to show that ϕ_0 admits an extension to a solution on $\{a\} \cup I_0$. The integral equation

$$\phi_0(t) = \mathbf{y}_0 - \int_t^{t_0} f\big(\tau, \phi_0(\tau)\big)\,\mathrm{d}\tau.$$

holds for $t \in (a, t_0]$. Since f is continuous, it is bounded on C and hence $\big|f\big(\tau, \phi_0(\tau)\big)\big| \le M$ for $\tau \in (a, t_0]$. Using this, it is easy to see that $\phi_0(a) := \mathbf{y}_0 - \int_a^{t_0} f\big(\tau, \phi_0(\tau)\big)\,\mathrm{d}\tau$ provides the desired extension. □

Higher-Order ODE's

In order to adapt the Existence and Uniqueness Theorems for 1st-order ODE systems to explicit n-th order (scalar) ODE's

$$y^{(n)} = f(t, y, y', \ldots, y^{(n-1)})$$

with $f\colon D \to \mathbb{R}$, $D \subseteq \mathbb{R} \times \mathbb{R}^n$ open (again there is no analogue for implicit n-th order ODE's), we need to relate the property "f satisfies locally a Lipschitz condition w.r.t. $\mathbf{y}$" to that of the corresponding 1st-order system $\mathbf{y}' = \mathbf{f}(t, \mathbf{y})$. (Here, and only here, we are using bold type to distinguish scalar and vectorial functions.)

Inspecting the explicit formula for $\mathbf{f}$ ("order reduction") and writing $\mathbf{y} = (y_0, \ldots, y_{n-1})$, $\mathbf{z} = (z_0, \ldots, z_{n-1})$, we have

$$\mathbf{f}(t, \mathbf{y}) - \mathbf{f}(t, \mathbf{z}) = \begin{pmatrix} y_1 - z_1 \\ \vdots \\ y_{n-1} - z_{n-1} \\ f(t, \mathbf{y}) - f(t, \mathbf{z}) \end{pmatrix}$$

Now suppose that $|f(t, \mathbf{y}) - f(t, \mathbf{z})| \leq L\,|\mathbf{y} - \mathbf{z}|$. For the squared Euclidean length of $\mathbf{f}(t, \mathbf{y}) - \mathbf{f}(t, \mathbf{z})$ we then obtain the estimate

$$
\begin{aligned}
|\mathbf{f}(t,\mathbf{y}) - \mathbf{f}(t,\mathbf{z})|^2 &= \sum_{i=1}^{n-1}(y_i - z_i)^2 + |f(t,\mathbf{y}) - f(t,\mathbf{z})|^2 \\
&\le \sum_{i=0}^{n-1}(y_i - z_i)^2 + L^2\,|\mathbf{y} - \mathbf{z}|^2 \\
&= (1 + L^2)\,|\mathbf{y} - \mathbf{z}|^2
\end{aligned}
$$

This says that $\mathbf{f}(t,\mathbf{y})$ satisfies a Lipschitz condition w.r.t. $\mathbf{y}$ with Lipschitz constant $\sqrt{1+L^2}$.

Conclusion: If f satisfies on D locally a Lipschitz condition w.r.t. $\mathbf{y}$ then so does $\mathbf{f}$ (with slightly larger Lipschitz constants).

Of course we also have: If f has continuous partial derivatives $\frac{\partial f}{\mathrm{d}y_0}(t,\mathbf{y}), \ldots, \frac{\partial f}{\mathrm{d}y_{n-1}}(t,\mathbf{y})$ then f satisfies on D locally a Lipschitz condition w.r.t. $\mathbf{y}$, and hence so does $\mathbf{f}$.

Corollary (Existence and Uniqueness Theorem for n-th order ODE's)

Suppose $f\colon D \to \mathbb{R}$, $D \subseteq \mathbb{R} \times \mathbb{R}^n$, is continuous and satisfies on D locally a Lipschitz condition w.r.t. $\mathbf{y}$. Further, let $(a, \mathbf{b}) = (a, b_0, \ldots, b_{n-1}) \in D$.

1. *If $\phi, \psi\colon I \to \mathbb{R}$ are solutions of the IVP*

$$y^{(n)} = f(t, y, y', \ldots, y^{(n-1)}) \wedge y^{(i)}(a) = b_i \text{ for } 0 \leq i \leq n-1, \qquad (\star)$$

   then $\phi(t) = \psi(t)$ for all $t \in I$.

2. *There exists $\epsilon > 0$ and a solution $\phi\colon [a-\epsilon, a+\epsilon] \to \mathbb{R}$ of the IVP $(\star)$.*

As remarked before, continuity of f, $\frac{\partial f}{y_0}, \ldots, \frac{\partial f}{y_{n-1}}$ is sufficient for the assumptions of the corollary to hold.

Proof.

Use the reduction to a 1st-order system $\mathbf{y}' = \mathbf{f}(t, \mathbf{y})$ as discussed earlier (setting $y_0 = y$, $y_1 = y'$, etc.), and apply the Existence and Uniqueness Theorem to $\mathbf{y}' = \mathbf{f}(t, \mathbf{y})$. As we have shown on the previous slide, this system satisfies the necessary assumptions (continuity is clear). □

Example

We have seen in the introduction that $y'' + y = 0$ has the general (real) solution $y(t) = A\cos t + B\sin t$ with constants $A, B \in \mathbb{R}$.

Here $f(t, y_0, y_1) = -y_0$, which even satisfies a global Lipschitz condition with $L = 1$.
$\Longrightarrow$ The Existence and Uniqueness Theorem applies.

Since $y \mapsto \big(y(0), y'(0)\big) = (A, B)$ produces every vector in $\mathbb{R}^2$ exactly once, the Existence and Uniqueness Theorem gives without any previous knowledge (except, of course, that $t \mapsto A\cos t + B\sin t$ is a solution of $y'' + y = 0$) that locally at $t = 0$ all solutions have this form for unique constants A, B. Since these solutions are defined on the whole of $\mathbb{R}$, one can then conclude that this remains true globally.

Using the addition theorems for $\cos t$, $\sin t$, one can show that an alternative representation of the general nonzero solution of $y'' + y = 0$ is, e.g., $y(t) = A\sin(t - t_0)$ with $A > 0$ and $0 \le t_0 < 2\pi$. This follows from the Existence and Uniqueness Theorem as well, since $y \mapsto \big(y(0), y'(0)\big) = (-A\sin t_0, A\cos t_0)$ produces every nonzero vector in $\mathbb{R}^2$ exactly once (by the polar coordinate representation of points in $\mathbb{R}^2$).

Example

Consider the (rather fancy) 3rd-order ODE

$$y''' = \begin{cases} \sin(e^y - y') & \text{if } t \le 0, \\ \sin(e^y - y' + ty'') & \text{if } t > 0. \end{cases} \qquad (\star)$$

Here we have $y''' = f(t, y, y', y'')$ with $f\colon \mathbb{R}^4 \to \mathbb{R}$ defined by

$$f(t, y_0, y_1, y_2) = \begin{cases} \sin(e^{y_0} - y_1) & \text{if } t \le 0, \\ \sin(e^{y_0} - y_1 + ty_2) & \text{if } t > 0. \end{cases}$$

f is continuous (check the behaviour near $t = 0$) and partially differentiable w.r.t. y_0, y_1, y_2, and $\frac{\partial f}{\partial y_0}$, $\frac{\partial f}{\partial y_1}$, $\frac{\partial f}{\partial y_2}$ are continuous; e.g.,

$$\frac{\partial f}{\partial y_2}(t, y_0, y_1, y_2) = \begin{cases} 0 & \text{if } t \le 0, \\ t\cos(e^{y_0} - y_1 + ty_2) & \text{if } t > 0. \end{cases}$$

$\Longrightarrow$ f satisfies on $\mathbb{R}^4$ locally a Lipschitz condition with respect to $\mathbf{y} = (y_0, y_1, y_2)$ (it doesn't matter that $\frac{\partial f}{\partial t}$ doesn't exist at some points).
$\Longrightarrow$ The Existence and Uniqueness Theorem applies, giving unique solvability of $(\star)$ for any initial values $y^{(i)}(a) = b_i$, $i = 0, 1, 2$.

Exercise

Determine all maximal solutions of the 2nd order ODE $y'' = |y|$.

Integral Curves

As with the concept of a maximal solution, we first have to make precise what we mean by "integral curve". For simplicity we consider only 1st-order, scalar valued ODE's, including those in "differential-like" form $M(x,y)\,\mathrm{d}x + N(x,y)\,\mathrm{d}y = 0$.

Definition

1. By an *integral curve* of $y' = f(t,y)$, or the more general implicit form $f(t,y,y') = 0$, we mean the graph $\{(t,\phi(t)); t \in I\}$ of a maximal solution $\phi\colon I \to \mathbb{R}$.
2. By an *integral curve* of $M(x,y)\,\mathrm{d}x + N(x,y)\,\mathrm{d}y = 0$ we mean the range $\gamma(I) \subseteq \mathbb{R}^2$ of a solution $\gamma\colon I \to \mathbb{R}^2$, i.e., $\gamma(t) = \big(x(t), y(t)\big)$ should be smooth and satisfy $M\big(x(t),y(t)\big)x'(t) + N\big(x(t),y(t)\big)y'(t) = 0$ for all $t \in I$, which is maximal with respect to this property, i.e., there must not be a solution with range strictly containing $\gamma(I)$.

Thus integral curves are smooth non-parametric plane curves describing/representing the solutions of 1st-order, scalar valued ODE's.

Example

Consider the ODE $x\,\mathrm{d}x + y\,\mathrm{d}y = 0$ and the corresponding explicit form $y' = \mathrm{d}y\,/\,\mathrm{d}x = -x/y$.

Parametric solutions $\gamma(t) = \big(x(t), y(t)\big)$ of the differential-like ODE must satisfy $x(t)x'(t) + y(t)y'(t) \equiv 0$, which after multiplication by 2 becomes

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(x(t)^2 + y(t)^2\right) \equiv 0.$$

Thus $x(t)^2 + y(t)^2 = C$ must be constant, showing that the integral curves of $x\,\mathrm{d}x + y\,\mathrm{d}y = 0$ are precisely the circles $x^2 + y^2 = R^2$, $R > 0$. (For this note that smoothness of γ excludes the case $R = 0$, and that the maximality condition excludes proper pieces of circles.)

The explicit ODE $y' = -x/y$ is not defined at $y = 0$. Its integral curves are the half-circles $x^2 + y^2 = R^2$, $y \gtrless 0$ (again excluding $R = 0$), which represent the graphs of its solutions $y(x) = \pm\sqrt{R^2 - x^2}$, $x \in (-R, R)$.

We see from this that integral curves of a differential-like ODE $M(x, y)\,\mathrm{d}x + N(x, y)\,\mathrm{d}y = 0$ may split into several pieces forming integral curves of the explicit ODE $y' = \mathrm{d}y\,/dx = -M(x, y)/N(x, y)$. This happens at zeros of N (singular points or points with a vertical tangent).

Corollary (Uniqueness of Integral Curves)

Suppose $N, M\colon D \to \mathbb{R}$, $D \subseteq \mathbb{R}^2$ open, are C^1-functions. If $M(x,y)\,\mathrm{d}x + N(x,y)\,\mathrm{d}y = 0$ has no singular points then through every point of D there passes exactly one integral curve ("solution curve").

Proof.

Let $(x_0, y_0) \in D$. By assumption (x_0, y_0) is non-singular, i.e. $M(x_0, y_0) \neq 0$ or $N(x_0, y_0) \neq 0$. Then the tangent direction of an integral curve in (x_0, y_0) is uniquely determined as the direction orthogonal to the vector $\big(M(x_0, y_0), N(x_0, y_0)\big)$. Since the tangent cannot be horizontal and vertical at the same time, we can parametrize γ locally either as $y(x)$ or as $x(y)$, which then must solve the explicit ODE

$$\frac{\mathrm{d}y}{\mathrm{d}x} = -\frac{M(x,y)}{N(x,y)} \quad \text{or} \quad \frac{\mathrm{d}x}{\mathrm{d}y} = -\frac{N(x,y)}{M(x,y)}, \quad \text{respectively.}$$

$\Longrightarrow$ The Existence and Uniqueness Theorem can be applied and yields that an integral curve through (x_0, y_0) exists. Uniqueness follows from the maximality condition (cf. the uniqueness proof for maximal solutions of IVP's). $\square$

Remark

Since M and N are continuous, the set S of singular points of $M(x,y)\,\mathrm{d}x + N(x,y)\,\mathrm{d}y$ is closed in D. Hence $D' = D \setminus S$ is open and satisfies all assumptions of the corollary. $\Longrightarrow$ If two integral curves intersect in one point, this point must be singular.

Afternote

It is not true in general that through every non-singular point of $M(x,y)\,\mathrm{d}x + N(x,y)\,\mathrm{d}y = 0$ (where M, N are C^1-functions on some open set $D \subseteq \mathbb{R}^2$) there passes exactly one integral curve.

As a counterexample consider the family of curves $y = Cx^2$, $C \in \mathbb{R}$. Since all these curves have a horizontal tangent in $(0,0)$, it is clear that we can glue branches with different C together at $(0,0)$ to form differentiable functions $y(x)$ on $\mathbb{R}$ other than $y(x) = Cx^2$ (e.g., $y(x) \equiv 0$ for $x \leq 0$ and $y(x) = x^2$ for $x \geq 0$). On the other hand, solving $y = Cx^2$ for C and taking partial derivatives gives the ODE $-2yx^{-3}\,\mathrm{d}x + x^{-2}\,\mathrm{d}y = 0$ or, clearing denominators,

$$2y\,\mathrm{d}x - x\,\mathrm{d}y = 0.$$

At the singular point $(0,0)$ there is no condition for parametric solutions (except differentiability), and hence all curves described above solve the ODE.

Afternote con't

(You can also check directly that, e.g., $\gamma(t) = (t, 0)$ for $t \leq 0$ and $\gamma(t) = (t, t^2)$ for $t \geq 0$ is differentiable at $t = 0$ and solves the ODE.)

$\Longrightarrow$ Through any point (x_0, y_0) with $x_0 \neq 0$ there are infinitely many integral curves—follow the curve with $C = y_0/x_0^2$ to the origin and from there proceed to the other side of the y-axis using any choice for C. These curves have the half-parabola $y = Cx^2$, $xx_0 \geq 0$ in common. If we remove $(0, 0)$ from the domain of $2y\,\mathrm{d}x - x\,\mathrm{d}y = 0$, the half-parabola becomes an integral curve of its own.

The picture is completed by the curve $x = x(y) = 0$ (the y-axis), which is the only solution through a point $(0, y_0)$ with $y_0 \neq 0$.

Thus all points in $\mathbb{R}^2 \setminus \{(0, 0)\}$ are non-singular, but only through some of these points passes a unique integral curve.

The situation is similar to that for the ODE $y' = \sqrt{|y|}$, or $\mathrm{d}y - \sqrt{|y|}\,\mathrm{d}x = 0$, which has $M(x, y) = -\sqrt{|y|}$ non-differentiable.

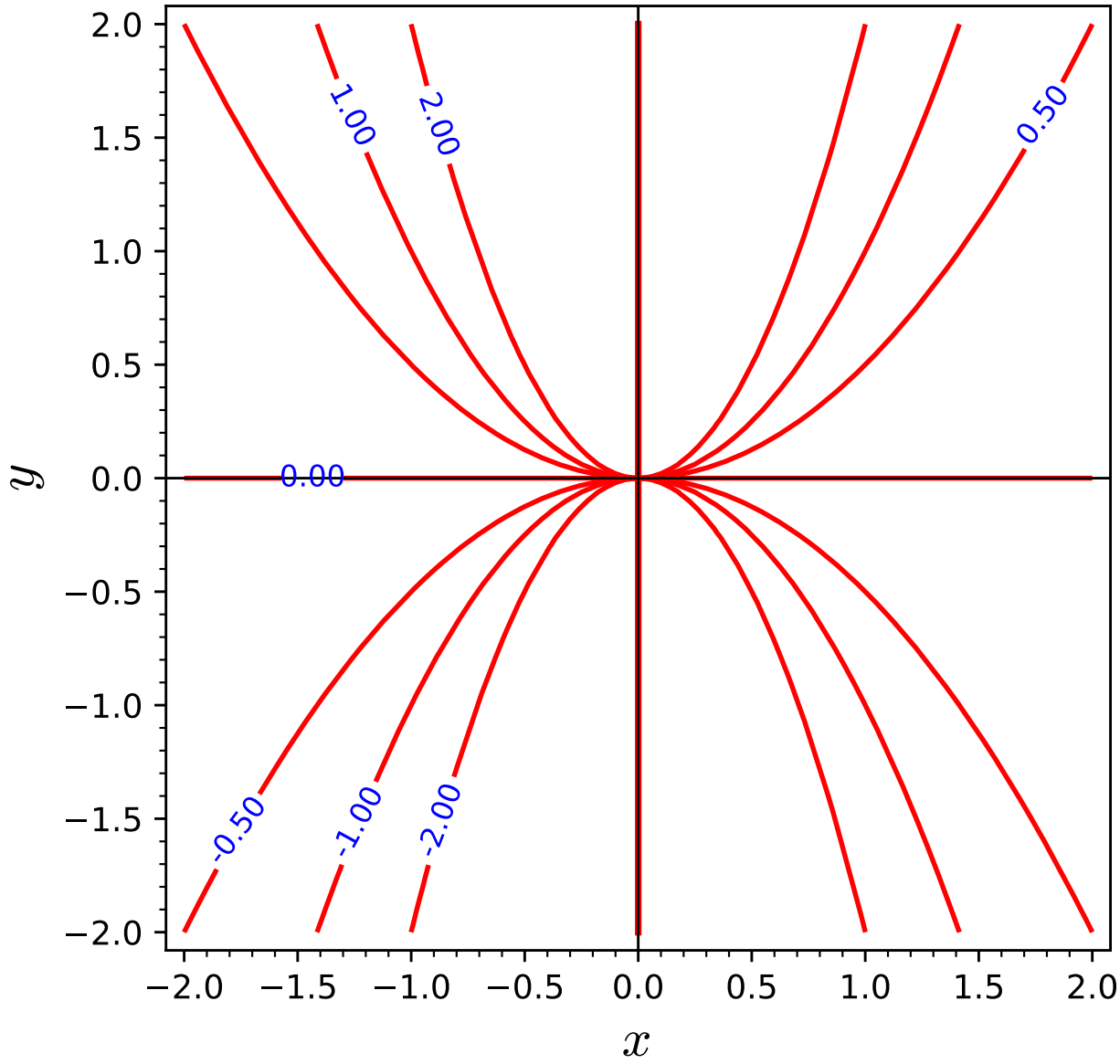


Figure: Integral curves of $2y\,\mathrm{d}x - x\,\mathrm{d}y = 0$, represented (except for $x = 0$) as contours of $F(x, y) = y/x^2$

Math 285
Introduction to Differential Equations

Thomas Honold

ZJU-UIUC Institute

Spring Semester 2024

Outline

Today's Lecture:

Phase Space

We consider an $n \times n$ ODE system

$$\mathbf{y}' = f(\mathbf{y}) \quad \text{with } f\colon D \to \mathbb{R}^n,\ D \subseteq \mathbb{R}^n \text{ open.} \tag{A}$$

Such a system is said to be *autonomous*, because f doesn't depend on t.

Observations

1. Solutions of (A) are parametric curves $\mathbf{y}(t) = \big(y_1(t), \ldots, y_n(t)\big)$, $t \in I$, contained in D. (More precisely, the range (or trace) of the associated non-parametric curve is contained in D.)
2. $\mathbf{y}(t)$, $t \in I$ is a solution iff $t \mapsto \mathbf{y}(t - t_0)$, $t \in I + t_0$ is a solution, where $I + t_0 = \{t + t_0; t \in I\}$. This holds for all $t_0 \in \mathbb{R}$.
3. If f is continuous and satisfies on D locally a Lipschitz condition, then for any point $\mathbf{y}^{(0)} \in D$ there exists precisely one maximal solution of the IVP $\mathbf{y}' = f(\mathbf{y}) \wedge \mathbf{y}(0) = \mathbf{y}^{(0)}$, and this solution is defined on a certain open interval I containing $t = 0$ as an inner point (by the Existence and Uniqueness Theorem).

Definition

1. The ambient space $\mathbb{R}^n$ containing D and the solution curves $\mathbf{y}(t)$ is called *phase space* of the autonomous system $\mathbf{y}' = f(\mathbf{y})$.
2. The non-parametric maximal solution curves $\{\mathbf{y}(t), t \in I\}$ (ranges/traces of $t \mapsto \mathbf{y}(t)$) are called *trajectories* or *orbits* of $\mathbf{y}' = f(\mathbf{y})$.

Corollary

Suppose f is continuous and satisfies on D locally a Lipschitz condition. Then every point of D is contained in a unique orbit of $\mathbf{y}' = f(\mathbf{y})$. In other words, the orbits form a partition of D.

Proof.

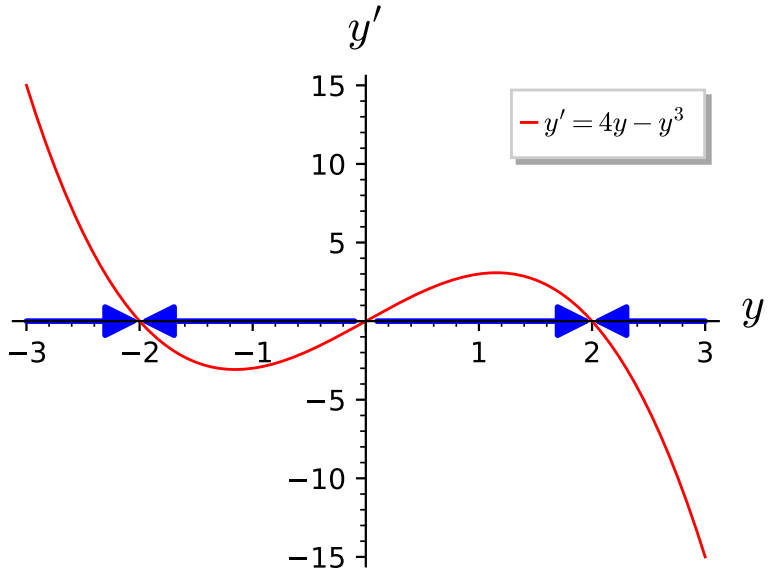
Let $\mathbf{y}^{(0)} \in D$. As already observed, $\mathbf{y}^{(0)} \in D$ is contained in an orbit of a maximal solution curve $\mathbf{y}(t)$, $t \in I$ that is defined at $t = 0$. Now suppose $\mathbf{z}(t)$, $t \in J$ is another maximal solution satisfying $\mathbf{z}(t_0) = \mathbf{y}^{(0)}$. Replacing $\mathbf{z}(t)$ by $t \mapsto \mathbf{z}(t + t_0)$, $t \in J - t_0$, which is a maximal solution as well and has the same orbit as $\mathbf{z}(t)$, we may assume $0 \in J$ and $\mathbf{z}(0) = \mathbf{y}^{(0)}$. But then the Uniqueness Theorem gives $\mathbf{y}(t) = \mathbf{z}(t)$ for $t \in I \cap J$, and maximality forces $I = J$ (because the two curves have a common extension to $I \cup J$). Thus the parametric curves and in particular their orbits are equal. $\square$

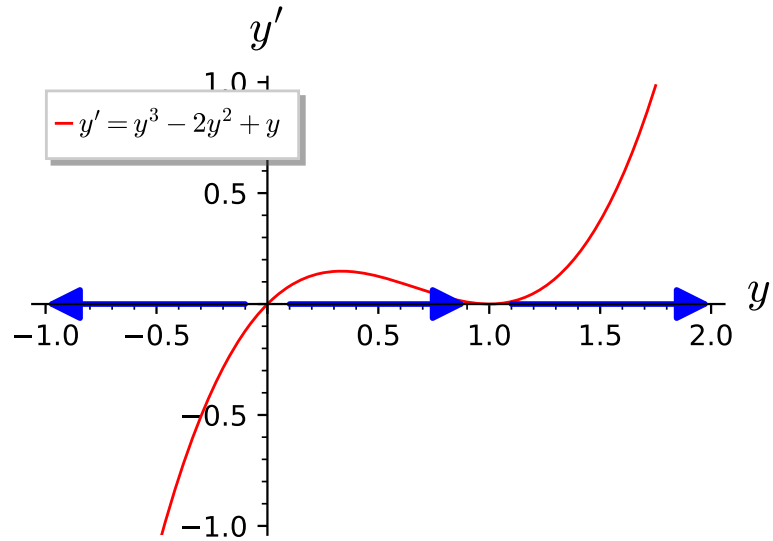
Note

Actually the proof shows more: Suppose we know a family of (parametric) maximal solutions whose associated orbits partition D. Then every further maximal solution has the form $t \mapsto \mathbf{y}(t - t_0)$ for some solution $\mathbf{y}(t)$ in the known family and some $t_0 \in \mathbb{R}$.

The Case $n = 1$

In this case $y' = f(y)$ for some one-variable function f. It is convenient to graph y' versus y, i.e., the function f.





The *phase line* is the y-axis (horizontal axis). The blue arrows indicate whether $y(t)$ is increasing/decreasing in the respective interval. *Caution:* This property depends on $y(t)$ rather than t !

Theorem

Suppose f is analytic on D, i.e., $f(y)$ is a polynomial ($D = \mathbb{R}$) or a power series in $y - y_0$ ($D = (y_0 - R, y_0 + R)$ for some $y_0 \in \mathbb{R}$ and $0 < R \leq \infty$), and $Z \subset D$ denotes the (discrete) set of zeros of f.

1. *The orbits of $y' = f(y)$ are the singleton sets $\{z\}$ for $z \in Z$ and the connected components of $D \setminus Z$, which in the polynomial case are the open intervals determined by adjacent zeros and intervals of the form $(-\infty, z)$, $(z, +\infty)$).*
2. *For $z \in Z$, $y' = f(y)$ has the equilibrium solution $y(t) \equiv z$.*
3. *If $f'(z) < 0$ then $y(t) \equiv z$ is asymptotically stable.*
   More generally, if f has a zero of odd multiplicity $m = 2k + 1$ at z and $f^{(2k+1)}(z) < 0$ then $y(t) \equiv z$ is asymptotically stable.
4. *If $f'(z) > 0$ then $y(t) \equiv z$ is unstable.*
   More generally, if f has a zero of odd multiplicity $m = 2k + 1$ at z and $f^{(2k+1)}(z) > 0$ then $y(t) \equiv z$ is unstable.
5. *If f has a zero of even multiplicity $m = 2k$ at z then $y(t) \equiv z$ is semistable (asymptotically stable from below if $f^{(2k)}(z) > 0$, respectively, from above if $f^{(2k)}(z) < 0$).*

Sketch of proof.

(2) is by now well-known and implies that for $z \in Z$ the set $\{z\}$ forms an orbit (arising from $y(t) \equiv z$).
Regarding (1), we prove only that if $z_1 < z_2$ are adjacent zeros of f and $f(y) > 0$ for $z_1 < y < z_2$ then (z_1, z_2) forms an orbit of $y' = f(y)$. (The other cases are similar.)

If suffices to show that a maximal solution $y(t)$ of $y' = f(y)$ with $y(0) = y_0 \in (z_1, z_2)$ exists for all $t \in \mathbb{R}$, is strictly increasing, and satisfies

$$\lim_{t \to -\infty} y(t) = z_1, \qquad \lim_{t \to +\infty} y(t) = z_2,$$

because then clearly $y(\mathbb{R}) = (z_1, z_2)$.

Let I be the (open) interval on which $y(t)$ is defined. We can write $I = (a, b)$, where $a = -\infty$ and/or $b = +\infty$ is possible.

First we show that $y(t) \in (z_1, z_2)$ for all $t \in I$.
This is true for $t = 0$ and can fail for some t only if there exists t_0 such that $y(t_0) = z_1$ or $y(t_0) = z_2$ (by the Intermediate Value Theorem). This, however, would contradict the Uniqueness Theorem, because we also have the constant solutions $y(t_0) \equiv z_1$ and $y(t_0) \equiv z_2$.

$\Longrightarrow$ $y(t)$ is strictly increasing on I and bounded from above by z_2.
$\Longrightarrow$ $y_2 := \lim_{t \uparrow b} y(t)$ exists and satisfies $y_0 < y_2 \leq z_2$.

Proof cont'd.

Now we distinguish two cases:

Case 1: $b \in \mathbb{R}$

In this case $y(t)$ can be extended to $(a, b]$ by setting $y(b) = y_2$, and one verifies easily that the extension solves $y' = f(y)$ also in $t = b$. This contradicts the maximality of $y(t)$.

Case 2: $b = +\infty$

Here we use that the limit

$$\lim_{t\to+\infty} y'(t) = \lim_{t\to+\infty} f\big(y(t)\big) = f(y_2)$$

exists. Since $\lim_{t\to+\infty}\big(y(t+1) - y(t)\big) = y_2 - y_2 = 0$, for sufficiently large t the quantity

$$0 < y(t+1) - y(t) = y'(\tau), \quad \tau \in (t, t+1),$$

is smaller than any given $\epsilon > 0$. Together with the existence of $\lim_{t\to+\infty} y'(t)$ this implies $\lim_{t\to+\infty} y'(t) = 0$, i.e., $f(y_2) = 0$ and hence $y_2 = \lim_{t\to+\infty} y(t) = z_2$.

In the same way one proves $a = -\infty$ and $\lim_{t\to-\infty} y(t) = z_1$.

Proof cont'd.

(3), (4), (5) follow from (2) and the known characterization of sign changes/non-changes at zeros of f in terms of the first non-vanishing derivative. □

Example ($y' = 4y - y^3$)

The preceding theorem gives immediately that the equilibrium solutions $y(t) \equiv \pm 2$ are asymptotically stable and $y(t) \equiv 0$ is unstable; cf. picture.

Example ($y' = y^3 - 2y^2 + y$)

$y(t) \equiv 0$ is unstable and $y(t) \equiv 1$ is semistable (more precisely, asymptotically stable from below and unstable from above); cf. picture.

Example ($y' = y - y^2$)

This is the logistic equation with $a = b = 1$. The graph of $f(y) = y - y^2 = -(y - 1/2)^2 + 1/4$ is the standard parabola upside down. It has zeros 0 and 1.
$\Longrightarrow$ $y(t) \equiv 0$ (corresponding to the left zero) is unstable, and $y(t) \equiv 1$ (corresponding to the right zero) is asymptotically stable.

Remark

Solutions of scalar autonomous ODE's are best viewed as functions $t(y)$.

$$\begin{aligned} y' &= f(y) \\ \mathrm{d}y\,/f(y) &= \mathrm{d}t \\ \int \frac{\mathrm{d}y}{f(y)} &= t = t(y). \end{aligned}$$

$\Longrightarrow$ $y' = f(y)$ can be solved by a single integration (just like $y' = f(t)$, only the roles of t and y are interchanged.

For example, in the case of $y' = y - y^2$ we obtain

$$t(y) = \int \frac{\mathrm{d}y}{y - y^2} = \int \left(\frac{1}{y} + \frac{1}{1-y}\right) \mathrm{d}y = \ln\left|\frac{y}{1-y}\right| + C.$$

The plot on the next slide shows 5 particular representative solutions for the 5 orbits of $y' = y - y^2$. The 3 branches of $y \mapsto \ln\left|\frac{y}{1-y}\right|$ represent the non-constant solutions. They can be independently shifted vertically to produce the remaining solutions.

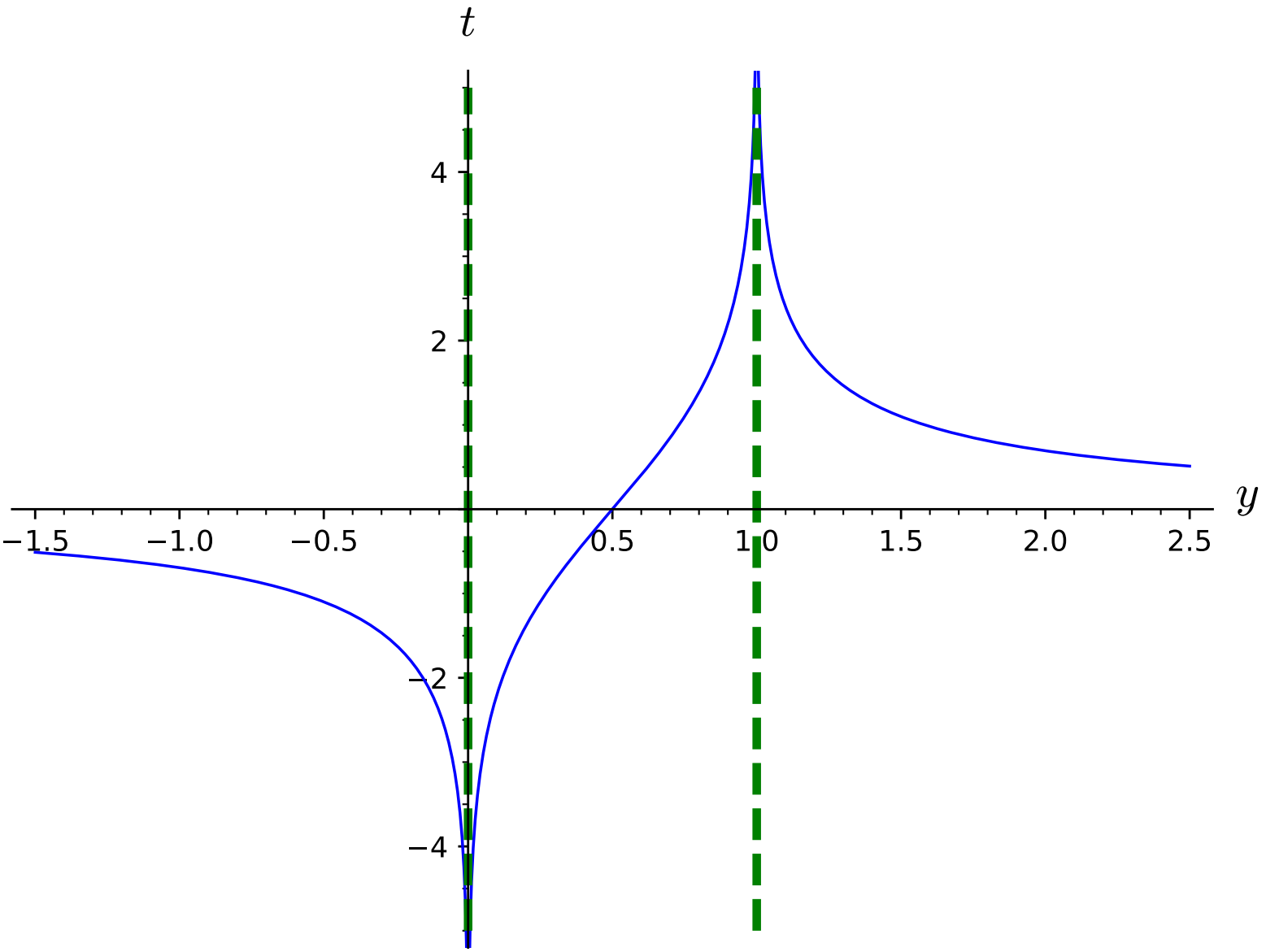


Figure: $t(y) = \ln\left|\frac{y}{1-y}\right|$

Exercise

In the proof of the theorem we have seen that maximal solutions representing orbits of $y' = f(y)$ of the form (z_1, z_2) have domain $\mathbb{R}$. How can we determine from properties of f the domain of solutions representing orbits of the form $(-\infty, z)$ or $(z, +\infty)$? In particular, answer this question for the case of a polynomial $f(y)$.

Exercise (cf. [BDM17], Sect. 2.5, p. 61)

The phase line can also be used to determine the curvature (i.e., whether it is convex or concave) of solutions of $y' = f(y)$. Show that solutions $y(t)$ are strictly convex (concave) in regions of the (t, y)-plane where $f(y)f'(y) > 0$ (respectively, $f(y)f'(y) < 0$). In particular, the inflection points of solutions (if any) are located on lines $y = y_0$ with $f(y_0) \neq 0 \wedge f'(y_0) = 0$ (e.g., for $y' = y - y^2$ on the line $y = 1/2$). What can be said about the number of inflection points of a non-constant solution with domain of the form $(-\infty, a)$, (a, b), or (b, ∞) with $a, b \in \mathbb{R}$?

The Case $n = 2$

Phase planes and planar trajectories/orbits are associated to 2×2 autonomous ODE systems

$$\begin{aligned} y_1' &= f_1(y_1, y_2), \\ y_2' &= f_2(y_1, y_2). \end{aligned}$$

Every maximal solution $\mathbf{y}(t) = \big(y_1(t), y_2(t)\big)$, $t \in I$ of such a system is a parametric plane curve. The orbit of $y(t)$, viz. $\big\{\big(y_1(t), y_2(t)\big); t \in I\big\}$, is the corresponding non-parametric curve.

Here we consider only one important example.

Example (Phase portrait of $y'' + y = 0$)

Order reduction $y_1 = y$, $y_2 = y'$ transforms this 2nd-order ODE into

$$\begin{pmatrix} y_1' \\ y_2' \end{pmatrix} = \begin{pmatrix} y_2 \\ -y_1 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}.$$

The orbit of a nonzero solution

$$\begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix} = \begin{pmatrix} y(t) \\ y'(t) \end{pmatrix} = \begin{pmatrix} A\cos t + B\sin t \\ -A\sin t + B\cos t \end{pmatrix} = \begin{pmatrix} A & B \\ B & -A \end{pmatrix} \begin{pmatrix} \cos t \\ \sin t \end{pmatrix}$$

is a circle of radius $\sqrt{A^2 + B^2}$ with center $(0, 0)$.

Example (cont'd)

Geometrically this says, that the solution with initial values $y(0) = A$, $y'(0) = B$ has the property that all "state vectors" $\big(y(t), y'(t)\big)$, describing the displacement from the equilibrium position $y = 0$ and its velocity of change at an arbitrary time t, are located on the circle $X^2 + Y^2 = A^2 + B^2$. (Recall that when we first determined the solutions of $y'' + y = 0$ we used this property, viz. $y(t)^2 + y'(t)^2 = A^2 + B^2 = y(0)^2 + y'(0)^2$, as a key fact.)

As predicted by the corollary, the orbits partition the plane if we also include $\big\{(0,0)\big\}$, the orbit of the constant solution $y(t) \equiv 0$.

We have also seen that solutions $y(t)$, $z(t)$ with the same orbit, i.e., the same $\sqrt{A^2 + B^2}$, differ only by a time shift (phase shift) $z(t) = y(t - t_0)$, $t_0 \in \mathbb{R}$. This is visible in the alternative representation

$$y(t) = A\cos t + B\sin t = \operatorname{Im}\left[\mathrm{e}^{\mathrm{i}t}(B + A\,\mathrm{i})\right] = \sqrt{A^2 + B^2}\,\sin(t + \phi),$$

in which ϕ is determined from $\cos\phi = \frac{B}{\sqrt{A^2+B^2}}$, $\sin\phi = \frac{A}{\sqrt{A^2+B^2}}$.

The collection of all orbits (or a good representative selection of orbits) of a given autonomous ODE system is referred to as a *phase portrait*.

Math 285
Introduction to Differential Equations

Thomas Honold

ZJU-UIUC Institute

Spring Semester 2024

Outline

Today's Lecture: Higher-Order Linear ODE's with Constant Coefficients

General (time-dependent) linear ODE's

An n-th order linear ODE has the form

$$a_n(t)y^{(n)} + a_{n-1}(t)y^{(n-1)} + \cdots + a_1(t)y' + a_0(t)y = b(t) \qquad (\star)$$

with coefficient functions $a_0(t), \ldots, a_n(t)$, $b(t)$, and $a_n(t) \neq 0$ for at least one t (i.e., $a_n(t)$ is not the all-zero function).

Solutions of $(\star)$ are n-times differentiable functions $y\colon I \to \mathbb{R}$ (or $y\colon I \to \mathbb{C}$), where I is an interval on which all coefficient functions are defined, satisfying
$a_n(t)y^{(n)}(t) + a_{n-1}(t)y^{(n-1)}(t) + \cdots + a_1(t)y'(t) + a_0(t)y(t) = b(t)$
for all $t \in I$.

As usual, $(\star)$ is said to be *homogeneous* if $b(t) \equiv 0$ (and inhomogeneous if $b(t) \neq 0$ for at least one t).

Notes

- In the homogeneous case $b(t) \equiv 0$, the real solutions of $(\star)$ with fixed domain J form a subspace of $\mathbb{R}^J = \{\phi; \phi\colon J \to \mathbb{R}\}$, and similarly for the complex solutions. This is easily shown using the subspace test: The all-zero function on J is a solution, and linear combinations (with constant coefficients) of solutions are again solutions.

Notes cont'd

- We can divide $(\star)$ by $a_n(t)$ and turn it into an explicit ODE, to which the Existence and Uniqueness Theorem can be applied. If $a_n(t)$ has zeros, this will generally split I into two or more subintervals for which $(\star)$ must be solved separately.
- In theory linear ODE's are well understood. There exists a sharpened version of the Existence and Uniqueness Theorem, asserting that an n-th order homogeneous linear ODE has an n-dimensional solution space (which is a subspace of $\mathbb{R}^I$ resp. $\mathbb{C}^I$) and that there are no obstructions to taking I as large as possible. Further, a particular solution of an inhomogeneous linear ODE can be found using "variation of parameters", and its general solution can be expressed in the usual way in terms of one particular solution and the general solution of the associated homogeneous ODE.
- In practice, however, it is difficult to solve time-dependent linear ODE's of orders $n \geq 2$. There are no known formulas for computing a basis of the associated solution space. For time-independent homogeneous linear ODE's, on the other hand, a basis of the solution space can be computed "algebraically".

The time-independent (autonomous) case

An n-th order linear ODE with constant coefficients (time-independent/autonomous linear ODE) has the form

$$y^{(n)} + a_{n-1}y^{(n-1)} + \cdots + a_1 y' + a_0 y = b \qquad \text{(DE)}$$

with coefficients $a_0, \ldots, a_{n-1}, b \in \mathbb{C}$.

It is not necessary to consider the more general form $a_n y^{(n)} + \cdots + a_0 y = b$, $a_n \neq 0$, since we can always divide by a_n to obtain the "monic" form (DE). This doesn't change anything, and neither does rewriting the ODE in explicit form $y^{(n)} = b - a_{n-1}y^{(n-1)} - \cdots - a_0 y$.

Several important physical quantities/systems can be described using ODE's of the form (DE), especially 2nd-order equations. In such applications, the left-hand side of (DE) is usually time-independent, expressing internal characteristics of the system. But $b = b(t)$ may be time-dependent, modeling the influence of an external source that changes over time. Since the basic theory of time-independent linear ODE's applies to this case as well, we will generally allow $b = b(t)$ in (DE).

The analogy with linear recurrence relations

Recall from Discrete Mathematics (or just take it as a definition) that a linear recurrence relation of order n with constant coefficients has the form

$$y_{i+n} = a_{n-1}y_{i+n-1} + \cdots + a_1 y_{i+1} + a_0 y_i + b_i, \quad i = 0, 1, 2, \ldots, \quad \text{(RR)}$$

and that a solution of (RR) is a sequence $(y_0, y_1, y_2, \ldots)$ satisfying (RR) for all $i \geq 0$.

In order to make the analogy with linear ODE's more visible, we replace a_i by $-a_i$, write $y(i)$ in place of y_i, (after all, a real sequence is just a function $y\colon \mathbb{N} \to \mathbb{R}$, $i \mapsto y_i$), and rename the variable i as t. Then (RR) becomes

$$y(t+n) + a_{n-1}y(t+n-1) + \cdots + a_1 y(t+1) + a_0 y(t) = b(t), \quad t \in \mathbb{N}.$$

Thus, compared with (DE), solutions of the recurrence relation (RR) have the "discrete" domain $\mathbb{N}$ (not a "continuous" interval I), and the differentiation operator $\mathrm{D}\colon y \mapsto y'$ has been replaced by the shift operator (truncation operator)

$$\mathrm{S}\colon (y_0, y_1, y_2, y_3, \ldots) \mapsto (y_1, y_2, y_3, \ldots).$$

Because of this striking analogy, it comes as no suprise that the methods for solving linear recurrence relations with constant coefficients and (higher-order) linear ODE's with constant coefficients are very closely related. If you know how to do one of these tasks, you will find it easy to do the other.

Example (Fibonacci numbers)

The Fibonacci numbers are defined by the order-two recurrence relation

$$f_{i+2} = f_{i+1} + f_i, \quad f_0 = 0,\ f_1 = 1.$$

Of course we can use our brain (or another computer) to compute the Fibonacci numbers successively from this:

i	0	1	2	3	4	5	6	7	8	9	10	11	12
f_i	0	1	1	2	3	5	8	13	21	34	55	89	144

But this leaves several questions open, e.g.

1. How fast do the Fibonacci numbers grow?
2. If we change the initial values f_0, f_1, how does the Fibonacci sequence change?

These questions can be answered by developing some theory, which yields a closed formula for the Fibonacci numbers as by-product.

Example (Fibonacci numbers cont'd)

We start with the following

Question: How does the collection of all solutions of the recurrence relation $y_{i+2} = y_{i+1} + y_i$ (without specifying initial values) look like?

Answer: Since the recurrence relation is homogeneous, sums of solutions and constant multiples of solutions will be again solutions. Moreover, there exist solutions, e.g., the Fibonacci sequence and the all-zero sequence $(0, 0, 0, \dots)$.
$\Longrightarrow$ The solutions form a subspace S of the vector space $\mathbb{R}^{\mathbb{N}}$ of all real sequences (with term-wise addition/scalar multiplication).

Question: What is the dimension of S?

Answer: When specifying a solution, we can choose $y_0 = A$, $y_1 = B$ freely, determining the rest of the sequence. Thus there are two degrees of freedom, which suggests that the dimension is 2. (But this is not a proof, of course.)
Consider the special solutions $f = (f_i)$, $g = (g_i)$ defined as follows:

n	0	1	2	3	4	5	6	7	8	9	10	11	12
f_n	0	1	1	2	3	5	8	13	21	34	55	89	144
g_n	1	0	1	1	2	3	5	8	13	21	34	55	89

Example (Fibonacci numbers cont'd)

Then for $A, B \in \mathbb{R}$ the sequence $y = Af + Bg$, i.e. $y_i = A f_i + B g_i$, is also a solution and satisfies

$$y = (Af_0 + Bg_0, Af_1 + Bg_1, \dots) = (B, A, \dots)$$

$\Longrightarrow$ Every solution is uniquely a linear combination of f and g.
$\Longrightarrow$ f, g form a basis of S; in particular we have $\dim S = 2$.

Remark: Since $g_i = f_{i-1}$ for $i \geq 1$, we can express the general solution of $y_{i+2} = y_{i+1} + y_i$ also as $y_i = Af_i + Bf_{i-1}$, using the convention that $f_{-1} = 0$.

The answer obtained so far is not really satisfying—for example we still have no information on the growth of (f_i) and other solutions of $y_{i+2} = y_{i+1} + y_i$, and how these relate to solutions of other linear recurrence relations.

Key idea

Every homogeneous linear recurrence relation with constant coefficients in $\mathbb{R}$ has solutions of the special form $(1, r, r^2, r^3, \dots)$, i.e., $y_i = r^i$, for some $r \in \mathbb{C}$.

Example (Fibonacci numbers cont'd)

Using the „Ansatz" $y_i = r^i$, we get

$$y_{i+2} = y_{i+1} + y_i \iff r^{i+2} = r^{i+1} + r^i \iff r^i(r^2 - r - 1) = 0.$$

$\Longrightarrow (y_i) = (r^i)$ satisfies the recurrence relation iff r is a root of the polynomial $X^2 - X - 1$. This polynomial is called *characteristic polynomial* of the recurrence relation $y_{i+2} = y_{i+1} + y_i$, and $r^2 - r - 1 = 0$ is called *characteristic equation*.

The solutions of $r^2 - r - 1 = 0$ are $r_1 = \frac{1+\sqrt{5}}{2}$, $r_2 = \frac{1-\sqrt{5}}{2}$, giving the two solutions

$$\mathbf{y}^{(1)} = \left(1, \tfrac{1+\sqrt{5}}{2}, \left(\tfrac{1+\sqrt{5}}{2}\right)^2, \left(\tfrac{1+\sqrt{5}}{2}\right)^3, \dots\right),$$

$$\mathbf{y}^{(2)} = \left(1, \tfrac{1-\sqrt{5}}{2}, \left(\tfrac{1-\sqrt{5}}{2}\right)^2, \left(\tfrac{1-\sqrt{5}}{2}\right)^3, \dots\right).$$

Since $r_1 \neq r_2$, the solutions are linearly independent (look at the first two terms of both sequences!) and form a basis of S.
$\Longrightarrow$ The general solution is $y_i = c_1 r_1^i + c_2 r_2^i$ with $c_1, c_2 \in \mathbb{R}$.
Plugging in $y_0 = 0$, $y_1 = 1$ gives $c_1 + c_2 = 0$, $c_1 r_1 + c_2 r_2 = 1$ (a linear system of equations for c_1, c_2), with solution $c_1 = \frac{1}{\sqrt{5}}$, $c_2 = -\frac{1}{\sqrt{5}}$.

Example (Fibonacci numbers cont'd)

$\Longrightarrow$The Fibonacci numbers have the closed-form representation

$$f_n = \frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^n - \left(\frac{1-\sqrt{5}}{2}\right)^n\right], \qquad n = 0, 1, 2, \ldots.$$

Since $\frac{1-\sqrt{5}}{2} \approx -0.62$ has absolute value < 1, we have

$$f_n \simeq \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^n \approx \frac{1}{\sqrt{5}} \times 1.62^n \quad \text{for large } n,$$

showing that the Fibonacci numbers grow exponentially.

Example (The "Fibonacci IVP")

By this we mean the IVP

$$y'' = y' + y, \quad y(0) = 0,\ y'(0) = 1.$$

Here we don't have an easy method at hand to compute a nontrivial solution, but we can observe at least the following analogy to the Fibonacci recurrence relation: For any interval $I \subseteq \mathbb{R}$ the real solutions $y\colon I \to \mathbb{R}$ of $y'' = y' + y$ form a subspace of $\mathbb{R}^I$.

Reason: Rewriting the ODE in the form $y'' - y' - y = 0$ shows that it is homogeneous.
$\Longrightarrow$ The all-zero function on I is a solution, and for solutions $y, z\colon I \to \mathbb{R}$ the sum $y + z\colon I \to \mathbb{R}$, $t \mapsto y(t) + z(t)$, as well as any scalar multiple $cy\colon I \to \mathbb{R}$, $t \mapsto c\,y(t)$ ($c \in \mathbb{R}$) are again solutions:

$$\begin{aligned}(y+z)'' - (y+z)' - (y+z) &= y'' + z'' - y' - z' - y - z \\ &= y'' - y' - y + (z'' - z' - z) \\ &= 0 + 0 = 0,\end{aligned}$$

$$\begin{aligned}(cy)'' - (cy)' - cy &= c\,y'' - c\,y' - cy \\ &= c(y'' - y' - y) = c \cdot 0 = 0.\end{aligned}$$

Example (The "Fibonacci IVP" cont'd)

Question: What is the dimension of the solution space S of $y'' - y' - y = 0$, and how to compute a basis of S?

Key idea

Try functions of the form $y(t) = \mathrm{e}^{rt}$.

Because differentiation "preserves exponentials", this might work. Indeed, for such functions $y(t)$, defined on $\mathbb{R}$, say, we have

$$\begin{aligned} y''(t) - y'(t) - y(t) &= r^2\,\mathrm{e}^{rt} - r\,\mathrm{e}^{rt} - \mathrm{e}^{rt} \\ &= (r^2 - r - 1)\mathrm{e}^{rt}, \end{aligned}$$

which is zero if $r^2 - r - 1 = 0$.
$\Longrightarrow$ $y'' - y' - y = 0$ has the same *characteristic equation*/*characteristic polynomial* as the Fibonacci recurrence relation (in the sense that this data fully characterizes the solution), and all functions

$$y(t) = c_1\mathrm{e}^{\frac{1+\sqrt{5}}{2}\,t} + c_2\mathrm{e}^{\frac{1-\sqrt{5}}{2}\,t}, \quad c_1, c_2 \in \mathbb{R},$$

are solutions of $y'' - y' - y = 0$.

Example (The "Fibonacci IVP" cont'd)

Next we fit the initial conditions:

$$y(0) = c_1 + c_2 = 0,$$
$$y'(0) = c_1 \tfrac{1+\sqrt{5}}{2} + c_2 \tfrac{1-\sqrt{5}}{2} = 1.$$

This system is the same as for the Fibonacci recurrence relation and was solved before.

$$\Longrightarrow\ y(t) = \frac{1}{\sqrt{5}}\left(\mathrm{e}^{\frac{1+\sqrt{5}}{2}t} - \mathrm{e}^{\frac{1-\sqrt{5}}{2}t}\right), \quad t \in \mathbb{R}$$

solves the Fibonacci IVP.

The solution is unique according to the Uniqueness Theorem.

But we can say more: Since for any $A, B \in \mathbb{R}$ the (linear) system

$$\begin{aligned} c_1 &+ c_2 &&= A, \\ r_1 c_1 &+ r_2 c_2 &&= B, \end{aligned}$$

with $r_1 = \frac{1+\sqrt{5}}{2}$, $r_2 = \frac{1-\sqrt{5}}{2}$ can be solved for c_1, c_2, we can fit solutions in S to any prescribed initial values $y(0) = A$, $y'(0) = B$.

Example (The "Fibonacci IVP" cont'd)

$\Longrightarrow$ There are no further solutions (by the Uniqueness Theorem), and hence the solution space S of $y'' - y' - y = 0$ is spanned by $\mathrm{e}^{\frac{1+\sqrt{5}}{2}t}$, $\mathrm{e}^{\frac{1-\sqrt{5}}{2}t}$ and has dimension 2.

$\Longrightarrow$ $\mathrm{e}^{\frac{1+\sqrt{5}}{2}t}$, $\mathrm{e}^{\frac{1-\sqrt{5}}{2}t}$ form a basis of S, since they generate S and are linearly independent.

We also say that $\mathrm{e}^{\frac{1+\sqrt{5}}{2}t}$, $\mathrm{e}^{\frac{1-\sqrt{5}}{2}t}$ form a *fundamental system* of solutions of $y'' - y' - y = 0$, according to the following

Definition

A basis of the solution space of a homogeneous linear ODE (or a homogeneous linear ODE system) is called a *fundamental system* of solutions.

Example

Determine the general solution of

$$y_i = 4y_{i-1} - 4y_{i-2} \quad \text{and}$$
$$y'' = 4y' - 4y.$$

First we rewrite the equations in standard form:

$$y_{i+2} - 4y_{i+1} + 4y_i = 0,$$
$$y'' - 4y' + 4y = 0.$$

The characteristic polynomial is $X^2 - 4X + 4 = (X-2)^2$ and has only one root, viz. $r = 2$. This gives the solutions $y_i = c\,2^i$ in the discrete case and $y(t) = c\,\mathrm{e}^{2t}$ in the continuous case, but these are obviously not enough to fit all possible initial conditions.

Question: How to obtain further solutions?

Answer: Try the sequence $y_i = i\,2^i$ (for a root r of multiplicity 2 in general $y_i = i\,r^i$, cf. Discrete Mathematics), respectively, the function $y(t) = t\,\mathrm{e}^{2t}$ (in general $y(t) = t\,\mathrm{e}^{rt}$). At least in the continuous case this is reasonable, because $t\,\mathrm{e}^{rt}$ when differentiated also reproduces in a way itself.

Example (cont'd)

$$\begin{aligned} y(t) &= t\,\mathrm{e}^{2t}, \\ y'(t) &= \mathrm{e}^{2t} + 2t\,\mathrm{e}^{2t} = (1+2t)\mathrm{e}^{2t}, \\ y''(t) &= 2\,\mathrm{e}^{2t} + 2(1+2t)\mathrm{e}^{2t} = (4+4t)\mathrm{e}^{2t}, \\ \Longrightarrow\ y'' - 4y' + 4y &= (4+4t)\mathrm{e}^{2t} - 4(1+2t)\mathrm{e}^{2t} + 4t\,\mathrm{e}^{2t} \\ &= (4+4t-4-8t+4t)\mathrm{e}^{2t} = 0. \end{aligned}$$

It works!

One can check that the solutions

$$\begin{aligned} y_i &= c_1 2^i + c_2 i\,2^i, \\ y(t) &= c_1\mathrm{e}^{2t} + c_2 t\,\mathrm{e}^{2t} \end{aligned}$$

can be used to fit arbitrary initial conditions ($y_0 = A$, $y_1 = B$ in the discrete case, $y(0) = A$, $y'(0) = B$ in the continuous case). Moreover, the two sequences (2^i), $(i\,2^i)$, respectively, the two functions e^{2t}, $t\,\mathrm{e}^{2t}$ are linearly independent.
$\Longrightarrow$ They form a basis of the solution space in both cases.

Example

Solve the inhomogenous recurrence relation

$$g_i = g_{i-1} + g_{i-2} + 1, \quad g_0 = g_1 = 1,$$

and the corresponding IVP $y'' = y' + y + 1$, $y(0) = y'(0) = 1$.

We do the continuous case first.
If we have two solutions y, z of $y'' - y' - y = 1$ then

$$\begin{aligned}(y-z)'' - (y-z)' - (y-z) &= y'' - y' - y - (z'' - z' - z) \\ &= 1 - 1 = 0,\end{aligned}$$

so that $y - z$ solves the associated homogeneous ODE
$y'' - y' - y = 0$, which is just the Fibonacci ODE.
This tells us that one particular solution y_p is enough to determine the general solution:

$$y = y - y_p + y_p = \text{sol. of the hom. ODE} + y_p.$$

Question: How to find a particular solution?

Example (cont'd)

Answer: The constant function $y(t) \equiv -1$ is clearly a solution.
$\Longrightarrow$ The general solution of $y'' - y' - y = 1$ is

$$y(t) = -1 + c_1 \, \mathrm{e}^{\frac{1+\sqrt{5}}{2} t} + c_2 \, \mathrm{e}^{\frac{1-\sqrt{5}}{2} t}, \quad c_1, c_2 \in \mathbb{R}.$$

Fitting the initial conditions gives the linear system

$$\begin{array}{rcrcrcrcl} y(0) & = & -1 & + & c_1 & + & c_2 & = & 1, \\ y'(0) & = & & & r_1 c_1 & + & r_2 c_2 & = & 1, \end{array}$$

which is solved by $c_1 = c_2 = 1$.

$$\Longrightarrow \; y(t) = -1 + \mathrm{e}^{\frac{1+\sqrt{5}}{2} t} + \mathrm{e}^{\frac{1-\sqrt{5}}{2} t}, \quad t \in \mathbb{R}$$

(uniquely) solves the given IVP.

Example (cont'd)

In the discrete case we can make a table of the numbers g_n and compare with the Fibonacci numbers.

n	0	1	2	3	4	5	6	7	8	9	10	11	12
f_n	0	1	1	2	3	5	8	13	21	34	55	89	144
g_n	1	1	3	5	9	15	25	41	67	109	177	287	465

A particular solution of $y_{i+2} - y_{i+1} - y_i = 1$ is $y_i = -1$, i.e. the sequence $y = (-1, -1, -1, \dots)$, and the general solution is therefore

$$y_i = -1 + c_1 \left(\frac{1+\sqrt{5}}{2}\right)^i + c_2 \left(\frac{1-\sqrt{5}}{2}\right)^i, \quad c_1, c_2 \in \mathbb{R}.$$

The initial conditions $y_0 = y_1 = 1$ yield the system

$$\begin{aligned} c_1 &+ c_2 &&= 2, \\ r_1 c_1 &+ r_2 c_2 &&= 2, \end{aligned}$$

which is solved by $c_1 = \frac{1+\sqrt{5}}{\sqrt{5}}$, $c_2 = \frac{1-\sqrt{5}}{-\sqrt{5}}$.

Example (cont'd)

$$\begin{aligned} \Longrightarrow\ g_i &= -1 + \frac{2}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^{i+1} - \left(\frac{1-\sqrt{5}}{2}\right)^{i+1}\right], \\ &= 2f_{i+1} - 1, \qquad i = 0, 1, 2, \ldots \end{aligned}$$

The relation $g_i = 2f_{i+1} - 1$ is also visible in the table (well, with some effort).
Using the alternative representation $y_i = -1 + Af_i + Bf_{i-1}$ (with $f_{-1} = 0$), we could have found it more quickly: $y_0 = y_1 = 1$ give $A = B = 2$ and hence $y_i = -1 + 2(f_i + 2f_{i-1}) = -1 + 2f_{i+1}$.

Exercise

In the example we have found that the ODE $y'' - y' - y = 1$ and its discrete "analogue" $y_{i+2} - y_{i+1} - y_i = 1$ both have the constant function $y(t) \equiv -1$ as a solution (of course, with different domains $\mathbb{R}$ resp. $\mathbb{N}$). Is this a pure coincidence or an instance of a more general correspondence between the continuous and discrete case?

Hint: It may help to identify the discrete analogue of the exponential function e^t first.

Complex Roots

Example

Determine the general solution of $y'' + y = 0$ and $y_{i+2} + y_i = 0$ from the characteristic equation.

Here the characteristic polynomial is $X^2 + 1 = (X - \mathrm{i})(X + \mathrm{i})$, so that the general complex solutions are

$$y(t) = c_1 \mathrm{e}^{\mathrm{i}t} + c_2 \mathrm{e}^{-\mathrm{i}t}, \quad t \in \mathbb{R},$$
$$y_n = c_1 \mathrm{i}^n + c_2 (-\mathrm{i})^n, \quad n = 0, 1, 2, \ldots.$$

with $c_1, c_2 \in \mathbb{C}$.

Question: How can we find the corresponding real solutions?

Answer: In the discrete case direct inspection gives that the solution can also be written as

$$y = (A, B, -A, -B, A, B, -A, -B, \ldots), \quad A, B \in \mathbb{C},$$

Here we simply need to restrict A, B to real numbers to obtain the general real solution.

Example (cont'd)

In the continuous case we can argue as follows:

$$\overline{y(t)} = \overline{c}_1 \mathrm{e}^{-\mathrm{i}t} + \overline{c}_2 \mathrm{e}^{\mathrm{i}t} = y(t) \quad \text{iff} \quad \overline{c}_1 = c_2$$

(since $\mathrm{e}^{\mathrm{i}t}$ and $\mathrm{e}^{-\mathrm{i}t}$ are linearly independent).
$\Longrightarrow$ The general real solution is

$$y(t) = c\,\mathrm{e}^{\mathrm{i}t} + \overline{c}\mathrm{e}^{-\mathrm{i}t} = 2\,\mathrm{Re}(c\,\mathrm{e}^{\mathrm{i}t}), \quad c \in \mathbb{C}.$$

Setting $2c = a - b\mathrm{i}$, $a, b \in \mathbb{R}$, we see that the general real solution can also be represented as

$$\begin{aligned} y(t) &= a\,\mathrm{Re}(\mathrm{e}^{\mathrm{i}t}) - b\,\mathrm{Re}(\mathrm{i}\mathrm{e}^{\mathrm{i}t}) = a\,\mathrm{Re}(\mathrm{e}^{\mathrm{i}t}) + b\,\mathrm{Im}(\mathrm{e}^{\mathrm{i}t}) \\ &= a\cos t + b\sin t, \quad a, b \in \mathbb{R}. \end{aligned}$$

A different argument to prove this uses the observation that for a linear ODE with real coefficients the real and imaginary part of any complex solution must be solutions as well.

A Stronger Link between the Continuous and Discrete Case

If you were already familiar with the solution methods for linear ODE's in the examples discussed so far, but not with their discrete analogues, you may have wondered where the key idea "try sequences of the form $(1, r, r^2, r^3, \dots)$" in the discrete case comes from. The correct explanation uses the concept of "eigenvectors/eigenvalues" of an endomorphism (linear operator) of a vector space.

Definition (recalled)

Suppose V is a vector space over a field K and $f\colon V \to V$ a linear map from V into itself (a so-called *endomorphism* of V). A nonzero vector $\mathbf{v} \in V$ is said to be an *eigenvector* of f if f maps $\mathbf{v}$ to a scalar multiple of itself, i.e.,

$$f(\mathbf{v}) = \lambda\mathbf{v} \quad \text{for some } \lambda \in K.$$

The scalar λ is called the corresponding *eigenvalue*.

Observation

The function e^{rt}, with $r \in \mathbb{C}$ arbitrary, is an eigenvector ("eigenfunction") of the differentiation operator $\mathrm{D}\colon y \mapsto y'$. Likewise, the sequence $(1, r, r^2, r^3, \dots)$ is an eigenvector ("eigensequence") of the shift operator $\mathrm{S}\colon (y_i) \mapsto (y_{i+1})$. In both cases the corresponding eigenvalue is r.

Of course in the continuous case you know this already: $\mathrm{D}(\mathrm{e}^{rt}) = r\,\mathrm{e}^{rt}$. In the discrete case we have likewise

$$\mathrm{S}(1, r, r^2, r^3, \dots) = (r, r^2, r^3, \dots) = r(1, r, r^2, \dots).$$

Both D and S can be iterated. Writing $\mathrm{D} \circ \mathrm{D} = \mathrm{D}^2$ ("differentiating twice"), $\mathrm{D} \circ \mathrm{D} \circ \mathrm{D} = \mathrm{D}^3$, etc., and similarly for S, we have, e.g.,

$$\begin{aligned}(y_{i+2} - y_{i+1} - y_i) &= (y_{i+2}) - (y_{i+1}) - (y_i)\\ &= \mathrm{S}^2(y_i) - \mathrm{S}(y_i) - (y_i) = (\mathrm{S}^2 - \mathrm{S} - 1)(y_i),\\ (\mathrm{S}^2 - \mathrm{S} - 1)(r^i) &= \mathrm{S}^2(r^i) - \mathrm{S}(r^i) - (r^i) = r^2(r^i) - r(r^i) - (r^i)\\ &= (r^2 - r - 1)(r^i),\end{aligned}$$

and similarly

$$y'' - y' - y = \mathrm{D}^2 y - \mathrm{D}y - y = (\mathrm{D}^2 - \mathrm{D} - 1)y,$$
$$(\mathrm{D}^2 - \mathrm{D} - 1)\mathrm{e}^{rt} = \mathrm{D}^2\mathrm{e}^{rt} - \mathrm{D}\mathrm{e}^{rt} - \mathrm{e}^{rt} = r^2\,\mathrm{e}^{rt} - r\,\mathrm{e}^{rt} - \mathrm{e}^{rt}$$
$$= (r^2 - r - 1)\mathrm{e}^{rt}.$$

One sees that iterating both operators and taking linear combinations, which corresponds to applying a polynomial in D or S (such as $p(\mathrm{D}) = \mathrm{D}^2 - \mathrm{D} - 1$ or $p(\mathrm{S}) = \mathrm{S}^2 - \mathrm{S} - 1$ in the case of $p(X) = X^2 - X - 1$) to the function/sequence and can produce the left-hand side of any higher-order linear ODE/linear recurrence relation, for their eigenfunctions/eigensequences effectively reduces the computation to a scalar multiplication with $p(r)$, where r is the corresponding eigenvalue; cf. also the discussion of matrix polynomials in Math 257.

This property will be crucial in the theoretical analysis of general higher-order linear ODE's (or linear recurrence relations) with constant coefficients.

While S acts on the vector space of (complex) sequences, there is a subtlety involved in finding a suitable domain for D. In order to be able to apply D repeatedly, we should define its domain as $\mathrm{C}^\infty(\mathbb{R})$ (complex-valued functions on $\mathbb{R}$, which have derivatives of all orders).

The Homogeneous Case

cf. also [BDM17], Ch. 4.2

We work over $\mathbb{C}$ (because generality does not hurt at this point) and denote by $\mathbb{C}[X]$ the ring of all polynomials in one indeterminate over $\mathbb{C}$. I assume that you know how to add and multiply polynomials, and that equality in $\mathbb{C}[X]$ means coefficient-wise equality.

We will also use the fact that every nonzero polynomial $a(X) \in \mathbb{C}[X]$ splits into linear factors in $\mathbb{C}[X]$, i.e.,

$$a(X) = a_d \prod_{i=1}^{r} (X - \lambda_i)^{m_i},$$

where $d \geq 0$ is the degree of $a(X)$, $a_d \neq 0$ is the leading coefficient of $a(X)$, $\lambda_1, \ldots, \lambda_r$ are the distinct roots (zeros) of $a(X)$ in $\mathbb{C}$ and $m_i \geq 1$ the corresponding multiplicities. This "prime factorization" is clearly unique, and its existence follows from the Fundamental Theorem of Algebra; cf. Calculus II.

The Fundamental Theorem of Algebra, asserting that every complex polynomial of degree $d \geq 1$ has a root in $\mathbb{C}$, is a mere existence theorem and doesn't say anything about how to actually compute the roots.

In fact, according to the ABEL-RUFFINI Theorem, the roots of a general polynomial of degree ≥ 5 (even with integer coefficients) cannot be computed algebraically, and one needs to use numerical approximations instead.

Definition

For an n-th order linear ODE

$$y^{(n)} + a_{n-1}y^{(n-1)} + \cdots + a_1 y' + a_0 y = b(t).$$

with constant coefficients $a_0, \ldots, a_{n-1} \in \mathbb{C}$ (but possibly non-constant right-hand side $b(t)$), the polynomial $a(X) = X^n + a_{n-1}X^{n-1} + \cdots + a_1 X + a_0 \in \mathbb{C}[X]$ is called its *characteristic polynomial* (and $r^n + a_{n-1}r^{n-1} + \cdots + a_1 r + a_0 = 0$ the corresponding *characteristic equation*).

Caution: In what follows, the letter "r" will have a different meaning (number of distinct roots of $a(X)$).

We first consider the homogeneous case $b(t) \equiv 0$. In this case the solutions $y : \mathbb{R} \to \mathbb{C}$ form a subspace of $\mathbb{C}^{\mathbb{R}}$ (since sums of solutions and linear combinations of solutions with coefficients in $\mathbb{C}$ are again solutions), and it is reasonable to conjecture that this subspace has dimension n (on the basis of our examples and the Existence and Uniqueness Theorem).

Theorem

Suppose the characteristic polynomial $a(X)$ of

$$y^{(n)} + a_{n-1}y^{(n-1)} + \cdots + a_1 y' + a_0 y = 0 \qquad \text{(H)}$$

has prime factorization $a(X) = \prod_{i=1}^{r}(X-\lambda_i)^{m_i}$. Then the functions

$$\mathbb{R} \to \mathbb{C},\ t \mapsto t^j e^{\lambda_i t}, \quad 1 \le i \le r,\ 0 \le j \le m_i - 1,$$

form a basis of the complex solution space S of (H) (a so-called fundamental system of solutions); in particular $\dim S = n$.

Proof.

First we note that the number of such functions is $\sum_{i=1}^{r} m_i = \deg a(X) = n$, which equals the conjectured dimension of the solution space S.

Hence it suffices to prove the following

1. S has dimension at most n.
2. The functions $t^j e^{\lambda_i t}$ actually solve (H), and
3. they are linearly independent.

Proof cont'd.

(1) Suppose, by contradiction, that $\dim S > n$. Then there exist $n+1$ linearly independent solutions $\phi_1, \ldots, \phi_{n+1} : \mathbb{R} \to \mathbb{C}$ of (H). Now we try to fit the initial conditions $y(0) = y'(0) = \cdots = y^{(n-1)}(0) = 0$ for a linear combination $y = \sum_{j=1}^{n+1} c_j \phi_j \in S$.

$$\begin{array}{rcccccccc} y(0) & = & c_1\phi_1(0) & + & \ldots & + & c_{n+1}\phi_{n+1}(0) & = & 0 \\ y'(0) & = & c_1\phi_1'(0) & + & \ldots & + & c_{n+1}\phi_{n+1}'(0) & = & 0 \\ & & \vdots & & & & \vdots & & \\ y^{(n-1)}(0) & = & c_1\phi_1^{(n-1)}(0) & + & \ldots & + & c_{n+1}\phi_{n+1}^{(n-1)}(0) & = & 0 \end{array}$$

This is a linear system of n equations for the $n+1$ unknowns c_j. From Linear Algebra we know that such a system must have a solution $\mathbf{c} = (c_1, \ldots, c_{n+1}) \neq \mathbf{0}$. The corresponding function $y = c_1\phi_1 + \cdots + c_{n+1}\phi_{n+1}$ is not the all-zero function (since the ϕ_j are linearly independent), but satisfies the same initial conditions as the all-zero function. This contradicts the Uniqueness Theorem and proves (1).

Proof cont'd.

(2) This is the most technical step. As discussed earlier, we can work with polynomial differential operators

$$p(\mathrm{D}) = p_0\,\mathrm{id} + p_1\mathrm{D} + \cdots + p_d\mathrm{D}^d, \quad \mathrm{D} = \frac{\mathrm{d}}{\mathrm{d}t},\ p_i \in \mathbb{C}.$$

Such an operator acts on $y(t)$ via

$$p(\mathrm{D})y = p_0 y + p_1 \mathrm{D}y + \cdots + p_d \mathrm{D}^d y = p_0 y + p_1 y' + \cdots + p_d y^{(d)},$$

and the ODE can be concisely written as $\boxed{a(\mathrm{D})y = 0.}$

The action is compatible with polynomial addition/multiplication in the following sense:

$$(p_1 + p_2)(\mathrm{D})y = \big(p_1(\mathrm{D}) + p_2(\mathrm{D})\big)y = p_1(\mathrm{D})y + p_2(\mathrm{D})y,$$
$$(p_1 p_2)(\mathrm{D})y = \big(p_1(\mathrm{D})p_2(\mathrm{D})\big)y = p_1(\mathrm{D})\big(p_2(\mathrm{D})y\big).$$

In other words, we can treat D like an indeterminate (it is also true that $p(\mathrm{D}) = 0$ iff $p_0 = p_1 = \cdots = p_d = 0$; cf. subsequent note), and polynomial addition/multiplication corresponds to addition/composition of the corresponding linear operators $p(\mathrm{D})$.

Proof cont'd.

In particular we have $p_1(\mathrm{D})p_2(\mathrm{D}) = p_2(\mathrm{D})p_1(\mathrm{D})$ for any two polynomials $p_1(X), p_2(X) \in \mathbb{C}[X]$.

This commuting relation is quite useful. For example, it shows easily that the derivative of a solution is itself a solution:

$$a(\mathrm{D})y = 0 \implies a(\mathrm{D})y' = a(\mathrm{D})\mathrm{D}y = \mathrm{D}a(\mathrm{D})y = \mathrm{D}0 = 0.$$

Keep in mind that $p_1(\mathrm{D})p_2(\mathrm{D})$ is an abbreviation for the composition $p_1(\mathrm{D}) \circ p_2(\mathrm{D})$ ("first apply $p_2(\mathrm{D})$ then $p_1(\mathrm{D})$"), just like D^i is for $\underbrace{\mathrm{D} \circ \mathrm{D} \circ \cdots \circ \mathrm{D}}_{i \text{ times}}$. The notation $p_1(\mathrm{D})p_2(\mathrm{D})$ makes the analogy with polynomials even more visible.

Also note that composition of differential operators makes only sense after specifying a suitable domain from which y is taken, which in this case is $\mathrm{C}^\infty(\mathbb{R})$, the set of all complex-valued functions f on $\mathbb{R}$ that have derivatives of all orders. The somewhat sloppy notation "id" refers to the identity map with this domain, viz., $\mathrm{C}^\infty(\mathbb{R}) \to \mathrm{C}^\infty(\mathbb{R})$, $y \mapsto y$.

Next we generalize our observation that the exponentials $\mathrm{e}^{\lambda t}$ form eigenfunctions of D to arbitrary polynomials $p(\mathrm{D})$.

Proof cont'd.

$$\mathrm{D}e^{\lambda t} = \lambda e^{\lambda t} \implies \mathrm{D}^i e^{\lambda t} = \lambda^i e^{\lambda t} \implies p(\mathrm{D})e^{\lambda t} = p(\lambda)e^{\lambda t}$$

This is the second useful relation and implies that $y(t) = e^{\lambda_i t}$ satisfies $a(\mathrm{D})y = a(\lambda_i)y = 0$, i.e., solves the ODE.

Further we have $\mathrm{D}(f(t)e^{\lambda t}) = f'(t)e^{\lambda t} + \lambda f(t)e^{\lambda t}$, giving

$$(\mathrm{D} - \mu\,\mathrm{id})\big(f(t)e^{\lambda t}\big) = \begin{cases} f'(t)e^{\lambda t} & \text{if } \mu = \lambda, \\ \big[(\lambda - \mu)f(t) + f'(t)\big]e^{\lambda t} & \text{if } \mu \neq \lambda. \end{cases}$$

This is the 3rd useful relation, which we will apply to polynomials $f(t)$.

By induction, we get $(\mathrm{D} - \lambda\ \mathrm{id})^m\big(f(t)e^{\lambda t}\big) = f^{(m)}(t)e^{\lambda t}$, and hence

$$(\mathrm{D} - \lambda\,\mathrm{id})^m t^j e^{\lambda t} = \begin{cases} m!\,e^{\lambda t} & \text{if } m = j, \\ 0 & \text{if } m > j. \end{cases}$$

In particular $(\mathrm{D} - \lambda_i\,\mathrm{id})^{m_i}(t^j e^{\lambda_i t}) = 0$ for $0 \le j \le m_i - 1$.
$\implies a(\mathrm{D})(t^j e^{\lambda_i t}) = 0$, since $a(\mathrm{D})$ is a multiple of $(\mathrm{D} - \lambda_i\,\mathrm{id})^{m_i}$.
Thus the functions $t \mapsto t^j e^{\lambda_i t}$, $1 \le i \le r$, $0 \le j \le m_i - 1$, solve the ODE, completing the proof of (2).

Note on this part of the proof

The idea behind it is that

$$a(\mathrm{D})y = (\mathrm{D} - \lambda_1 \,\mathrm{id})^{m_1}(\mathrm{D} - \lambda_2 \,\mathrm{id})^{m_2} \cdots (\mathrm{D} - \lambda_r \,\mathrm{id})^{m_r} y$$

can be computed by applying n-times an operator of the simple form $\mathrm{D} - \mu \,\mathrm{id}$ with $\mu \in \mathbb{C}$, which acts like this: $(\mathrm{D} - \mu \,\mathrm{id})y = \mathrm{D}y - \mu y = y' - \mu y$. The order in which these operators are applied does not matter, because polynomial multiplication is commutative.

In the following example we write $\mathrm{D} - \mu$ for $\mathrm{D} - \mu \,\mathrm{id}$ (i.e., the identity map is simply denoted by 1). We have used this abbreviation before (when discussing D and S together), and it makes the formulas look a little less cluttered.

$y'' + y = (\mathrm{D}^2 + 1)y = (\mathrm{D} + \mathrm{i})(\mathrm{D} - \mathrm{i})y$ can be computed as the composition of $\mathrm{D} + \mathrm{i}$ and $\mathrm{D} - \mathrm{i}$ in either order. Here is one:

$$\begin{aligned} (\mathrm{D} - \mathrm{i})y &= y' - \mathrm{i}y, \\ (\mathrm{D} + \mathrm{i})(y' - \mathrm{i}y) &= (y' - \mathrm{i}y)' + \mathrm{i}(y' - \mathrm{i}y) \\ &= y'' - \mathrm{i}y' + \mathrm{i}y' - \mathrm{i}^2 y = y'' + y, \quad \text{as asserted.} \end{aligned}$$

Proof cont'd.

(3) A linear dependency relation among the functions $t^j e^{\lambda_i t}$ amounts to the existence of polynomials $f_i(X) \in \mathbb{C}[X]$ with $\deg f_i(X) \leq m_i - 1$, not all zero, and such that

$$f_1(t)e^{\lambda_1 t} + f_2(t)e^{\lambda_2 t} + \cdots + f_r(t)e^{\lambda_r t} = 0 \quad \text{for all } t \in \mathbb{R}.$$

Write $a(X) = (X - \lambda_1)^{m_1} A_1(X)$, i.e., $A_1(X)$ is the product of all polynomials $(X - \lambda_i)^{m_i}$ with $i \geq 2$.
Since $(\mathrm{D} - \lambda_i \,\mathrm{id})^{m_i}\big(f_i(t)e^{\lambda_i t}\big) = 0$, we have $A_1(\mathrm{D})\big(f_i(t)e^{\lambda_i t}\big) = 0$ for $i \geq 2$ and hence

$$A_1(\mathrm{D})\big(f_1(t)e^{\lambda_1 t}\big) = 0 \quad \text{for all } t \in \mathbb{R}.$$

But each factor $\mathrm{D} - \lambda_i \,\mathrm{id}$ of $A_1(\mathrm{D})$ preserves the degree of $f_1(X)$ in the product $f_1(t)e^{\lambda_1 t}$ and hence acts as a bijection on the space consisting of all such functions.

$$\Longrightarrow\ f_1(t)e^{\lambda_1 t} = 0\ (t \in \mathbb{R}) \Longrightarrow\ f_1(t) = 0\ (t \in \mathbb{R})\ \Longrightarrow\ f_1(X) = 0.$$

The last implication uses the fact that polynomials have only finitely many zeros.
In the same way one shows that $f_2(X) = \cdots = f_r(X) = 0$. □

Notes

- An analogous theorem holds in the discrete case. The solution space of a homogeneous linear recurrence relation of order n with constant coefficients and characteristic polynomial $a(X) = \prod_{i=1}^{r}(X-\lambda_i)^{m_i}$ has a basis consisting of the sequences

$$(k^j\lambda_i^k)_{k\in\mathbb{N}}, \quad 1 \le i \le r,\ 0 \le j \le m_i - 1.$$

  The proof given in the continuous case remains valid—just replace everywhere D by S (and functions by sequences, of course).
- The relation $\mathrm{D}f = \lambda f$ is equivalent to $(\mathrm{D}-\lambda\,\mathrm{id})f = 0$, i.e., to the ODE $y' - \lambda y = 0$ for f, whose solution is $y(t) = c\,\mathrm{e}^{\lambda t}$, $c \in \mathbb{C}$. Thus the eigenspace E_λ of D (acting on $\mathrm{C}^\infty(\mathbb{R})$, viewed as a complex vector space) is 1-dimensional and spanned by $t \mapsto \mathrm{e}^{\lambda t}$. The corresponding generalized eigenspace is $\mathrm{G}_\lambda = \{f; (\mathrm{D}-\lambda\,\mathrm{id})^m f = 0 \text{ for some } m \in \mathbb{N}\}$. In the proof of the theorem we have seen that G_λ consists precisely of the polynomial multiples $p(t)\mathrm{e}^{\lambda t}$, $p(X) \in \mathbb{C}[X]$. Thus, in contrast with the matrix case, the generalized eigenspaces of D have infinite (countable) dimension.

Notes cont'd

- The functions in the span of $\{t^k e^{\lambda t}; k \in \mathbb{N}, \lambda \in \mathbb{C}\}$ are called *exponential polynomials*.
  The theorem implies in particular that any solution of a homogeneous linear ODE with constant coefficients is an exponential polynomial.
  Conversely, every exponential polynomial solves a nontrivial homogeneous linear ODE with constant coefficients. For $t^k e^{\lambda t}$ the corresponding ODE can be taken as $(\mathrm{D} - \lambda \,\mathrm{id})^{k+1} y = 0$, and for a linear combination $\sum_{i=1}^{r} c_i t^{k_i} e^{\lambda_i t}$ we can then take the ODE as
  $$\left[\prod_{i=1}^{r} (\mathrm{D} - \lambda_i \,\mathrm{id})^{k_i+1}\right] y = 0.$$
  It should be noted here that a fixed exponential polynomial $y(t)$ satisfies many different such ODE's, since $a(\mathrm{D})y = 0$ implies $b(\mathrm{D})a(\mathrm{D})y = 0$ for any polynomial $b(X) \in \mathbb{C}[X]$. However, it can be shown that there exists a unique monic polynomial $a(X) \in \mathbb{C}[X]$ of smallest degree such that $a(\mathrm{D})y = 0$ and hence a unique "monic" linear homogeneous ODE of smallest order satisfied by $y(t)$; cf. exercises.

Notes cont'd

- The argument used in Part (3) of the proof actually shows that the sum of the generalized eigenspaces of D is a direct sum. (Using the previous note, convince yourself that the proof more generally shows: If $\lambda_1, \ldots, \lambda_r \in \mathbb{C}$ are distinct and $f_1(X), \ldots, f_r(X)$ are polynomials in $\mathbb{C}[X]$ satisfying $f_1(t)\mathrm{e}^{\lambda_1 t} + \cdots + f_r(t)\mathrm{e}^{\lambda_r t} = 0$ for $t \in \mathbb{R}$ then $f_1(X) = \cdots = f_r(X) = 0$.)

  In fact the argument is the same as that used in Math 257 to show that the sum of the generalized eigenspaces of $\mathbf{A} \in \mathbb{C}^{n\times n}$ is direct. It relies solely on the fact that members of different generalized eigenspaces G_λ and G_μ are annihilated by relatively prime polynomials (in this case $a(\mathrm{D}) = (\mathrm{D} - \lambda\,\mathrm{id})^k$, $b(\mathrm{D}) = (\mathrm{D} - \mu\,\mathrm{id})^l$ for some $k, l \in \mathbb{N}$).

  Also note that, in contrast with the matrix case, the sum of the generalized eigenspaces of D, viz. the space of exponential polynomials, is a proper subspace of the domain $\mathrm{C}^\infty(\mathbb{R})$.

Notes cont'd

Here we supply the yet missing precise definition of polynomial differential operators $p(\mathrm{D}) = p_0 + p_1\mathrm{D} + p_2\mathrm{D}^2 + \cdots + p_d\mathrm{D}^d$ corresponding to polynomials $p(X) \in \mathbb{C}[X]$.

We have defined $p(\mathrm{D})$ as the map $y \mapsto p_0 y + p_1 \mathrm{D}y + \cdots + p_d \mathrm{D}^d)y$, but this is incomplete without specifying the domain and codomain of $p(\mathrm{D})$. Since we want to compose differential operators as maps, domain and codomain should be equal.
Now care must be taken to avoid the following problem: If $f\colon \mathbb{R} \to \mathbb{C}$ is differentiable but f' is not, $\mathrm{D}(\mathrm{D}f) = \mathrm{D}f'$ is undefined. (In other words, D doesn't map the space of differentiable functions into itself.)
The problem can be cured by taking as domain of D the set of functions $f \in \mathbb{C}^{\mathbb{R}}$ that have derivatives of all orders. This set is commonly denoted by $\mathrm{C}^\infty(\mathbb{R})$ and forms a subspace of $\mathbb{C}^{\mathbb{R}}$. For $f \in \mathrm{C}^\infty(\mathbb{R})$ we have $f' \in \mathrm{C}^\infty(\mathbb{R})$ as well (check it!), and hence $\mathrm{D}\colon \mathrm{C}^\infty(\mathbb{R}) \to \mathrm{C}^\infty(\mathbb{R})$, $f \mapsto f'$ is well-defined.
Finally, we check that solutions of $a(\mathrm{D})y = 0$ are in fact in $\mathrm{C}^\infty(\mathbb{R})$. Writing the ODE in the form $y^{(n)} = a_0 y + a_1 y' + \cdots + a_{n-1} y^{(n-1)}$ and differentiating gives $y^{(n+1)} = a_0 y' + a_1 y'' + \cdots + a_{n-1} y^{(n)}$, showing that $y^{(n+1)}$ exists. Iterating this argument gives $y \in \mathrm{C}^\infty(\mathbb{R})$.

Notes cont'd

- In the proof of the theorem we have tacitly used that the Existence and Uniqueness Theorem holds also for complex ODE systems and higher-order ODE's. This can be seen as follows: Apply reduction of order $z_1 = z$, $z_2 = z'$, …, $z_n = z^{(n-1)}$ to reduce a complex n-th order ODE $z^{(n)} = f(t, z, z', \ldots, z^{(n-1)})$ to a complex 1st-order system $z_k' = f_k(t, z_1, \ldots, z_n)$, $1 \le k \le n$. Then, writing $z_k = x_k + \mathrm{i}y_k$ and using $z_k' = x_k' + \mathrm{i}y_k'$, we see that this system is equivalent to

$$\begin{aligned} x_k' &= \operatorname{Re} f_k(t, x_1 + \mathrm{i}y_1, \ldots, x_n + \mathrm{i}y_n), && 1 \le k \le n, \\ y_k' &= \operatorname{Im} f_k(t, x_1 + \mathrm{i}y_1, \ldots, x_n + \mathrm{i}y_n) && 1 \le k \le n, \end{aligned}$$

  which is a $2n$-dimensional real system. Corresponding IVP's are also equivalent—a vectorial initial condition $\mathbf{z}(t_0) = \mathbf{z}^{(0)} \in \mathbb{C}^n$ translates into $\mathbf{x}(t_0) = \operatorname{Re}\mathbf{z}^{(0)} \wedge \mathbf{y}(t_0) = \operatorname{Im}\mathbf{z}^{(0)}$, which gives $2n$ real initial conditions, matching the dimension of the real system. Finally the real version of the Existence and Uniqueness Theorem can be applied and gives the truth of the corresponding complex version.

The Real Case

Corollary

If $a_0, a_1, \ldots, a_{n-1} \in \mathbb{R}$ then the complex solution space of

$$y^{(n)} + a_{n-1}y^{(n-1)} + \cdots + a_1 y' + a_0 y = 0 \qquad \text{(H)}$$

has a basis consisting of n real solutions, and these form a basis of the real solution space as well. In particular the real solution space has dimension n as well.

The subsequent proof shows how to actually obtain a basis of real solutions. One simply takes the real and imaginary parts of the (possibly complex) solutions $t^j e^{\lambda_i t}$, discarding repetitions.

Proof.

Writing a complex solution as $z(t) = x(t) + \mathrm{i}y(t)$, or $z = x + \mathrm{i}y$ for short, we have

$$\begin{aligned} 0 &= z^{(n)} + a_{n-1}z^{(n-1)} + \cdots + a_1 z' + a_0 z \\ &= x^{(n)} + \mathrm{i}y^{(n)} + \cdots + a_1(x' + \mathrm{i}y') + a_0(x + \mathrm{i}y) \\ &= x^{(n)} + \cdots + a_1 x' + a_0 x + \mathrm{i}\left(y^{(n)} + \cdots + a_1 y' + a_0 y\right). \end{aligned}$$

Proof cont'd.

By assumption, $x^{(n)} + \cdots + a_1 x' + a_0 x$ and $y^{(n)} + \cdots + a_1 y' + a_0 y$ are real for each $t \in \mathbb{R}$, and hence both must be zero.
$\Longrightarrow$ The real and imaginary parts of a complex solution are itself solutions.

Applying this to a basis of the complex solution space S, we obtain $2n$ real solutions, which generate S and from which we can then select n linearly independent real solutions $\phi_1, \ldots, \phi_n$ forming a basis of S. Now suppose $c_1, \ldots, c_n \in \mathbb{C}$ are such that

$$\begin{aligned} & y(t) = c_1\phi_1(t) + \cdots + c_n\phi_n(t) \in \mathbb{R} \quad \text{for all } t \in \mathbb{R}. \\ \Longrightarrow\ & \overline{y(t)} = \overline{c}_1\phi_1(t) + \cdots + \overline{c}_n\phi_n(t) = y(t) \end{aligned}$$

for all t, and hence $c_i = \overline{c}_i \in \mathbb{R}$ for $1 \leq i \leq n$ by the linear independency of ϕ_i. This shows that the real solution space is generated by $\phi_1, \ldots, \phi_n$.
Moreover, since these functions are linearly independent over $\mathbb{C}$ and $\mathbb{R} \subset \mathbb{C}$, they must also be linearly independent over $\mathbb{R}$. Hence they form a basis of the real solution space. $\square$

Notes

- Without the condition $a_0, a_1, \ldots, a_{n-1} \in \mathbb{R}$ the conclusion in the corollary is false, as the example $y' - \mathrm{i}y = 0$ shows. The complex solutions are $c\,\mathrm{e}^{\mathrm{i}t}$, $c \in \mathbb{C}$, and the only real solution among these is the all-zero function.

- The relation between the real and complex solution space of $y^{(n)} + a_{n-1}y^{(n-1)} + \cdots + a_1 y' + a_0 y = 0$ just described actually holds in more generality and can be formulated in a pure Linear Algebra setting (and for arbitrary fields E, F with $E \supset F$ in place of $\mathbb{C}, \mathbb{R}$). If S is a subspace of some function space $\mathbb{C}^I$ (considered as a vector space over $\mathbb{C}$), we can consider the subset $S_{\mathbb{R}} = S \cap \mathbb{R}^I$ consisting of all real-valued functions in S (*field reduction*). The set $S_{\mathbb{R}}$ forms a vector space over $\mathbb{R}$ (a subspace of $\mathbb{R}^I$). Conversely, starting with a subspace T of $\mathbb{R}^I$, we can consider the span $T_{\mathbb{C}}$ of T over $\mathbb{C}$, which is a subspace of $\mathbb{C}^I$ (*field extension*).
  Then $\dim(T_{\mathbb{C}}) = \dim(T)$ holds in general, and any basis of T over $\mathbb{R}$ forms a basis of $T_{\mathbb{C}}$ over $\mathbb{C}$; moreover, $T = (T_{\mathbb{C}})_{\mathbb{R}}$. But $\dim(S_{\mathbb{R}}) = \dim(S)$ (equivalently, $S = (S_{\mathbb{R}})_{\mathbb{C}}$ iff S has a basis consisting of real-valued functions, and $\dim(S_{\mathbb{R}}) < \dim(S)$ otherwise; the latter case actually occurs if $|I| > 1$.

Exercise

a) Prove the assertions about field extension $T \mapsto T_{\mathbb{C}}$ in the previous note.
Hint: The key fact to be established is that functions $f_1, \ldots, f_r : I \to \mathbb{R}$ that are linearly independent over $\mathbb{R}$ remain linearly independent over the larger field $\mathbb{C}$.

b) Prove the assertions about field reduction $S \mapsto S_{\mathbb{R}}$ in the previous note, including for $|I| > 1$ an example of a subspace S of $\mathbb{C}^I$ for which $\dim(S_{\mathbb{R}}) < \dim(S)$.
Hint: For the example it suffices to consider the case $|I| = 2$, i.e., $\mathbb{C}^I \cong \mathbb{C}^2$.

c) Show that $T_{\mathbb{C}}$ forms a vector space of dimension $2\dim(T)$ over $\mathbb{R}$.

Exercise

Show that for any matrix $\mathbf{A} \in \mathbb{C}^{m\times n}$ the following are equivalent.

1. The real solution space and the complex solution space of $\mathbf{Ax} = \mathbf{0}$ have the same dimension.
2. The row space of $\mathbf{A}$ has a basis consisting of vectors in $\mathbb{R}^n$.

How about the column space of $\mathbf{A}$ in this regard?

Example

We solve the 3rd-order ODE

$$y''' - y'' - 2y' = 0.$$

The characteristic polynomial is

$$a(X) = X^3 - X^2 - 2X = X(X+1)(X-2)$$

with roots $\lambda_1 = 0$, $\lambda_2 = -1$, $\lambda_3 = 2$ and all multiplicities equal to 1.

$$\Longrightarrow \mathrm{e}^{0\,t} = 1,\ \mathrm{e}^{-t},\ \mathrm{e}^{2t}$$

form a fundamental system of solutions.

$\Longrightarrow$ The general solution is

$$y(t) = c_1 + c_2\mathrm{e}^{-t} + c_3\mathrm{e}^{2t}, \quad c_1, c_2, c_3 \in \mathbb{C}$$

(or "$c_1, c_2, c_3 \in \mathbb{R}$" if only real solutions are considered).

Example (cont'd)

It is worth recalling the argument why 1, e^{-t}, e^{2t} solve the ODE $y''' - y'' - 2y' = 0$:

In polynomial differential operator notation the ODE is

$$a(\mathrm{D})y = (\mathrm{D}^3 - \mathrm{D}^2 - 2\mathrm{D})y = \mathrm{D}(\mathrm{D}+1)(\mathrm{D}-2)y = 0,$$

where $\mathrm{D}+1 = \mathrm{D} + \mathrm{id}$ and $\mathrm{D}-2 = \mathrm{D} - 2\ \mathrm{id}$.

- $y_1(t) = 1$ is a solution, since $\mathrm{D}y_1 = 0$ and hence $(\mathrm{D}+1)(\mathrm{D}-2)\mathrm{D}y_1 = 0$. (The order of the factors in $a(\mathrm{D})$ doesn't matter.
- $y_2(t) = \mathrm{e}^{-t}$ is a solution, since $(\mathrm{D}+1)y_2 = y_2' + y_2 = 0$ and hence $\mathrm{D}(\mathrm{D}-2)(\mathrm{D}+1)y_2 = 0$.
- $y_3(t) = \mathrm{e}^{2t}$ is a solution, since $(\mathrm{D}-2)y_3 = y_3' - 2\,y_3 = 0$ and hence $\mathrm{D}(\mathrm{D}+1)(\mathrm{D}-2)y_3 = 0$.

Example

We solve the homogeneous 4th-order ODE

$$y^{(4)} + 8y'' + 16y = 0.$$

The characteristic polynomial is

$$X^4 + 8X^2 + 16 = (X^2 + 4)^2 = (X - 2\mathrm{i})^2(X + 2\mathrm{i})^2.$$

$\Longrightarrow \lambda_1 = 2\mathrm{i},\ \lambda_2 = -2\mathrm{i}$ with multiplicities $m_1 = m_2 = 2$.

$$\Longrightarrow \quad \mathrm{e}^{2\mathrm{i}t},\ t\,\mathrm{e}^{2\mathrm{i}t},\ \mathrm{e}^{-2\mathrm{i}t},\ t\,\mathrm{e}^{-2\mathrm{i}t}$$

form a complex fundamental system.

A real fundamental system is then obtained by taking the real and imaginary parts of one function from each complex conjugate pair, i.e.,

$$\cos(2t),\ t\cos(2t),\ \sin(2t),\ t\sin(2t).$$

The general real solution of $y^{(4)} + 8y'' + 16y = 0$ is therefore $y(t) = c_1\cos(2t) + c_2 t\cos(2t) + c_3\sin(2t) + c_4 t\sin(2t)$ with $c_1, c_2, c_3, c_4 \in \mathbb{R}$ (and the general complex solution is of the same form with $c_1, c_2, c_3, c_4 \in \mathbb{C}$).

Example (cont'd)

Again let us recall why this works:

In polynomial differential operator notation the ODE is

$$a(\mathrm{D})y = (\mathrm{D}^4 + 8\mathrm{D}^2 + 16)y = (\mathrm{D}^2 + 4)^2 y = 0.$$

- $\mathrm{e}^{2\mathrm{i}t}$, $\mathrm{e}^{-2\mathrm{i}t}$ (or $\cos(2t)$, $\sin(2t)$) are solutions, since they solve $(\mathrm{D}^2 + 4)y = y'' + 4\,y = 0$, and hence also $(\mathrm{D}^2 + 4)^2 y = 0$.
- $t\,\mathrm{e}^{2\mathrm{i}t}$ is a solution, since $(\mathrm{D} - 2\mathrm{i})\left[t\,\mathrm{e}^{2\mathrm{i}t}\right] = \mathrm{e}^{2\mathrm{i}t}$, hence $(\mathrm{D} - 2\mathrm{i})^2)\left[t\,\mathrm{e}^{2\mathrm{i}t}\right] = (\mathrm{D} - 2\mathrm{i})\left[\mathrm{e}^{2\mathrm{i}t}\right] = 0$, and then $(\mathrm{D}^2 + 4)^2\left[t\,\mathrm{e}^{2\mathrm{i}t}\right] = (\mathrm{D} + 2\mathrm{i})^2)(\mathrm{D} - 2\mathrm{i})^2)\left[t\,\mathrm{e}^{2\mathrm{i}t}\right] = 0$ as well.

Note

If you are wondering why the real and imaginary parts of any complex fundamental system form a real fundamental system—in particular why the number of functions in both systems is the same, here is the argument in more detail:

If $a(X)$ is real, its non-real roots (if any) come in complex-conjugate pairs $\mu, \overline{\mu}$, which must have the same multiplicity, say m. The corresponding $2m$ functions in the complex fundamental system are $\{t^k e^{\mu t}, t^k e^{\overline{\mu} t}; 0 \le k \le m-1\}$. Writing $\mu = \alpha + \beta \mathrm{i}$, $\overline{\mu} = \alpha - \beta \mathrm{i}$, we have

$$t^k e^{\mu t} = t^k e^{\alpha t} \cos(\beta t) + \mathrm{i}\, t^k e^{\alpha t} \sin(\beta t),$$
$$t^k e^{\overline{\mu} t} = t^k e^{\alpha t} \cos(\beta t) - \mathrm{i}\, t^k e^{\alpha t} \sin(\beta t).$$

$\Longrightarrow$ The $2m$ real and imaginary parts of both kinds of functions are the same (except for a sign change in the imaginary parts). Discarding these "repetitions", we obtain the correct number $2m$ of real fundamental solutions, viz.
$\{t^k e^{\alpha t} \cos(\beta t), t^k e^{\alpha t} \sin(\beta t); 0 \le k \le m-1\}$.

Example (Harmonic oszillator)

The corresponding ODE is

$$\frac{\mathrm{d}^2 y}{\mathrm{d}t^2} + \omega^2 y = 0, \quad \omega > 0.$$

Here the characteristic polynomial is $X^2 + \omega^2 = (X - \mathrm{i}\omega)(X + \mathrm{i}\omega)$ and a fundamental system is $\{\mathrm{e}^{\mathrm{i}\omega t}, \mathrm{e}^{-\mathrm{i}\omega t}\}$.

The general real solution may be written in either of the two forms

1. $y(t) = c_1 \cos(\omega t) + c_2 \sin(\omega t), \quad c_1, c_2 \in \mathbb{R}$;
2. $y(t) = A \cos(\omega t + \alpha), \quad A \geq 0, \alpha \in [0, 2\pi)$.

The second form arises from the general complex solution $y(t) = c_1 \mathrm{e}^{\mathrm{i}\omega t} + c_2 \mathrm{e}^{-\mathrm{i}\omega t}$ by observing that $y(t)$ is real iff $c_2 = \overline{c}_1$, and setting $2c_1 = A\mathrm{e}^{\mathrm{i}\alpha}$.

Example (Harmonic oszillator with damping)

The corresponding ODE is

$$\frac{\mathrm{d}^2y}{\mathrm{d}t^2} + 2\mu\frac{\mathrm{d}y}{\mathrm{d}t} + \omega_0^2 y = 0.$$

The quantity $2\mu > 0$ is the damping factor, and $\omega_0 > 0$ is the (suitably normalized) characteristic frequency of the undamped system, whose solutions are generated by $\cos(\omega_0 t)$, $\sin(\omega_0 t)$.

This is a time-independent 2nd-order linear ODE with characteristic polynomial $X^2 + 2\mu X + \omega_0^2$, whose roots are

$$\lambda_{1,2} = -\mu \pm \sqrt{\mu^2 - \omega_0^2}.$$

Case 1: $\mu < \omega_0$.

In this case we have $\lambda_{1,2} = -\mu \pm \mathrm{i}\sqrt{\omega_0^2 - \mu^2}$, and a real fundamental system of solutions is

$$\mathrm{e}^{-\mu t}\cos(\omega t),\ \mathrm{e}^{-\mu t}\sin(\omega t), \quad \omega = \sqrt{\omega_0^2 - \mu^2} < \omega_0.$$

Example (cont'd)

Solutions form periodic oszillations with lower frequency and exponentially decreasing amplitude.

Case 2: $\mu = \omega_0$.
In this case $\lambda_1 = \lambda_2 = -\mu$, and a (real) fundamental system of solutions is

$$e^{-\mu t},\ te^{-\mu t}.$$

Solutions ultimately approach zero exponentially, but may have one maximum or minimum.

Case 3: $\mu > \omega_0$.
In this case λ_1 and λ_2 are distinct negative real numbers, and a fundamental system of solutions is

$$e^{-\mu_1 t},\ e^{-\mu_2 t}, \quad \mu_{1,2} = \mu \pm \sqrt{\mu^2 - \omega_0^2} > 0.$$

All solutions approach zero exponentially.

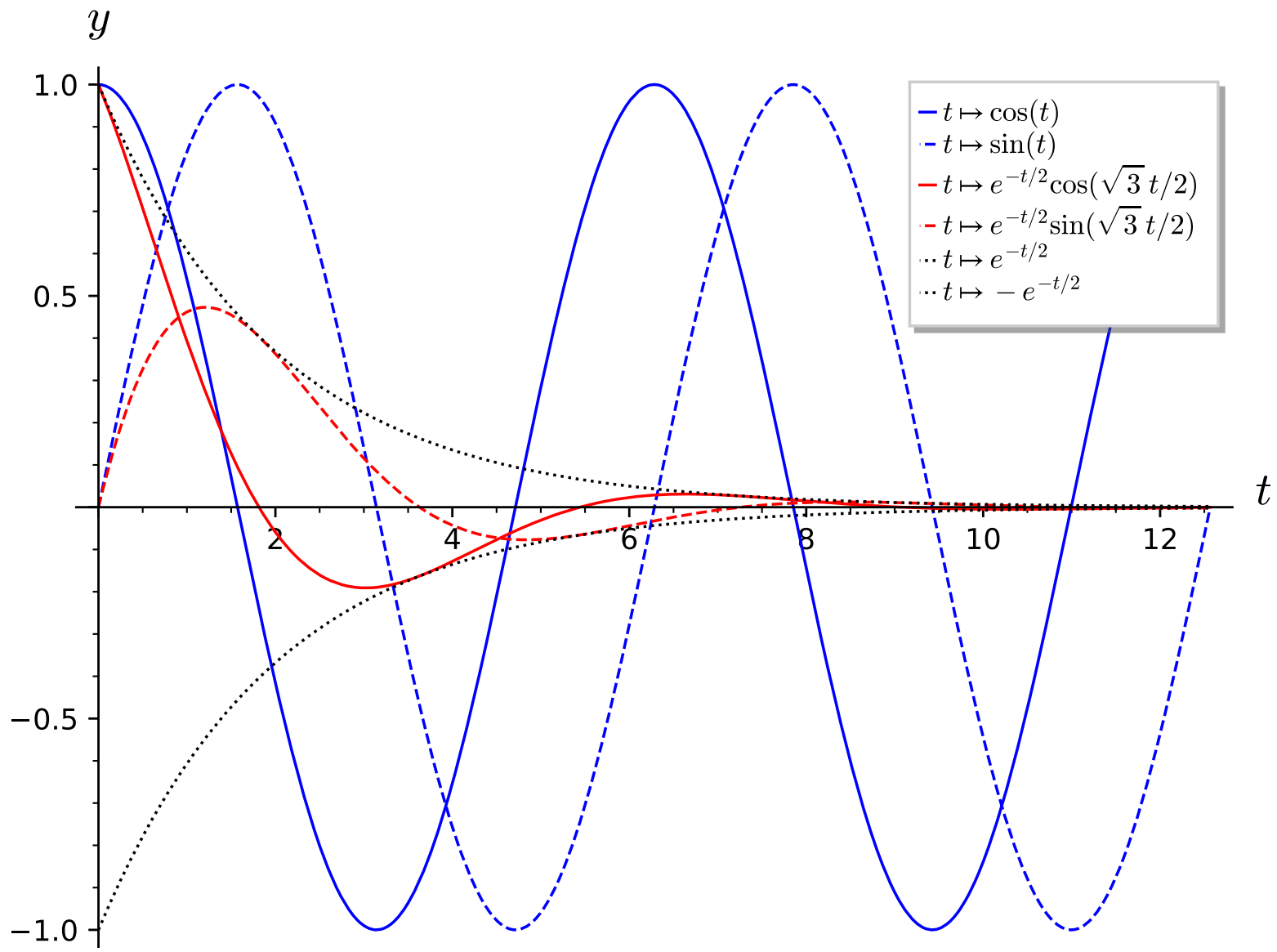


Figure: Fundamental system (in red) for $\omega_0 = 1$, $\mu = 1/2$

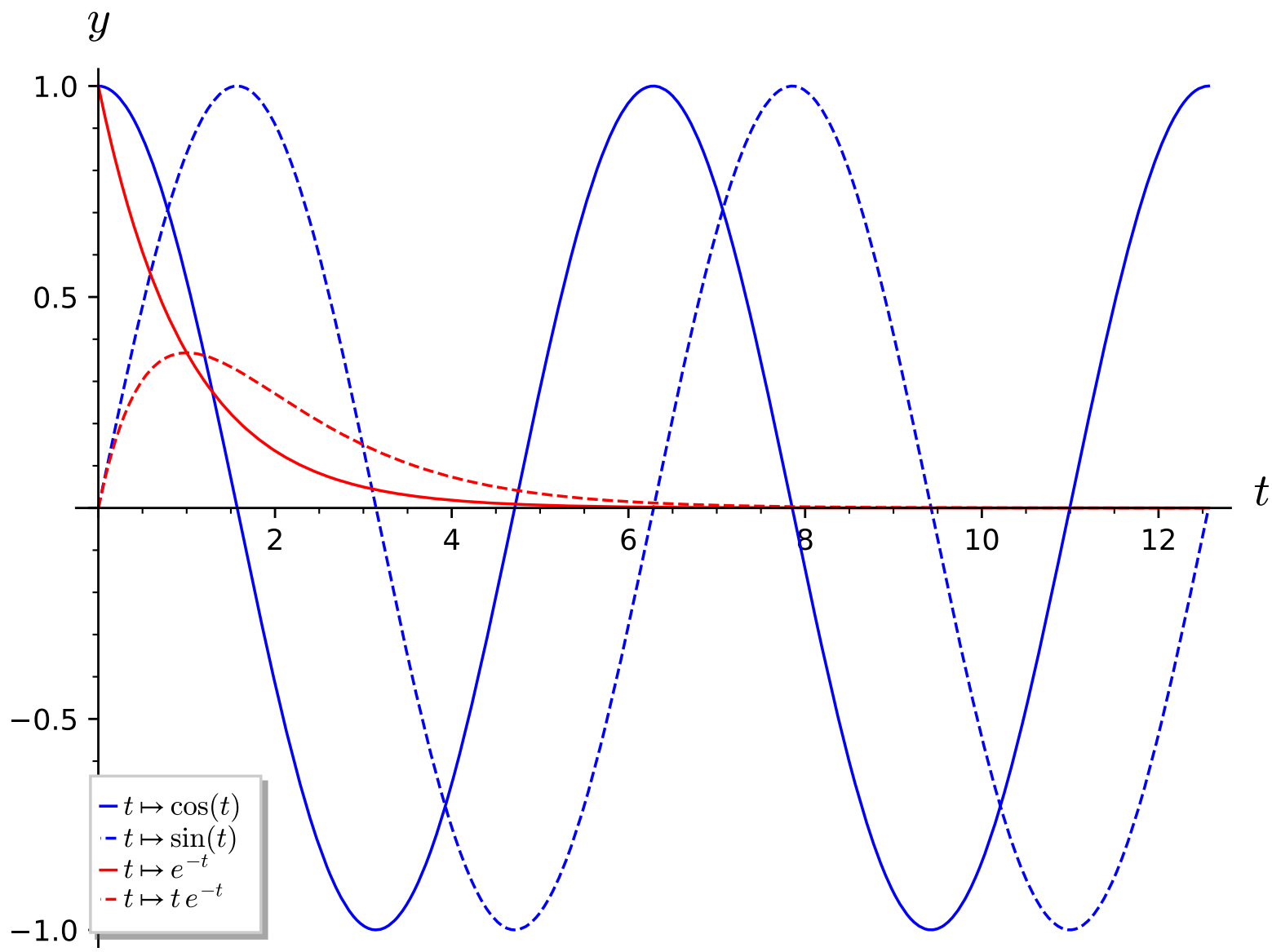


Figure: Fundamental system (in red) for $\omega_0 = 1$, $\mu = 1$

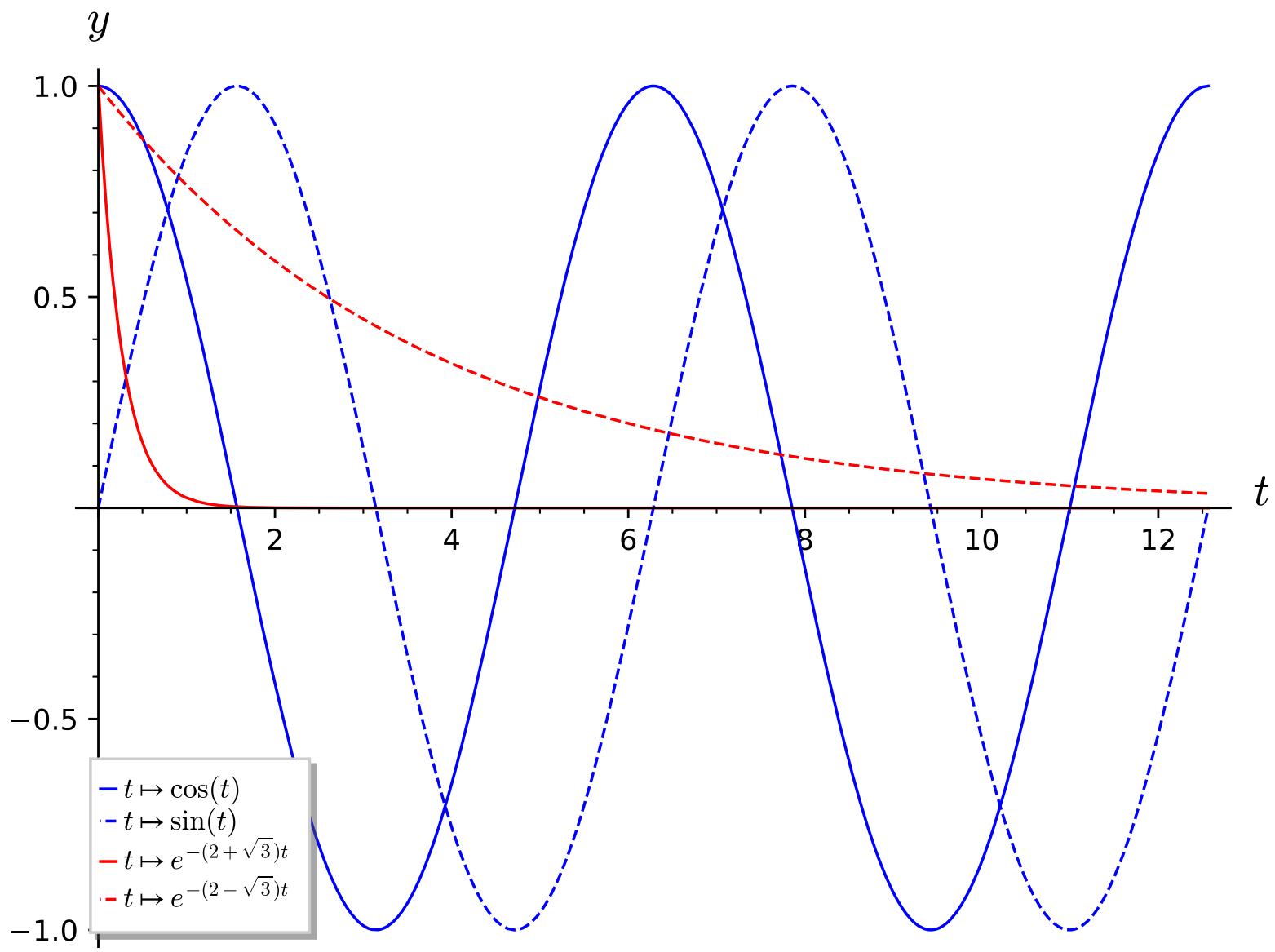


Figure: Fundamental system (in red) for $\omega_0 = 1$, $\mu = 2$

Notes on the preceding figures

- In the first figure you can see that the "period" (obtained by neglecting the decay factor $\mathrm{e}^{-t/2}$) of the two fundamental solutions is larger than for the corresponding undamped system, whose solutions are $\cos t$, $\sin t$ ($4\pi/\sqrt{3} \approx 7.255$ vs. 2π). Every nonzero solution $y(t)$ inherits the "period" and the decay factor, as can be seen by writing it in the form $y(t) = \mathrm{e}^{-t/2}\big(c_1 \cos(\omega t) + c_2 \sin(\omega t)\big)$, $\omega = \sqrt{3}/2$.
- The figures show that, contrary to the case of 1st-order ODE's, solution graphs of 2nd-order ODE's may intersect but can't touch. In fact, the Existence and Uniqueness Theorem tells us that in the cases under consideration for any point $(t_0, y_0) \in \mathbb{R}^2$ and any $m_0 \in \mathbb{R}$ there is exactly one solution passing through this point and having slope m_0 there. For the case $\mu = \omega_0 = 1$ this is illustrated on the next slide.

Exercise

It appears that in the first figure the first fundamental solution $y_1(t)$ (that involving $\cos$) doesn't have a maximum at $t = 0$. Verify this property, and describe the extrema of $y_1(t)$ in terms of those of $\cos$.

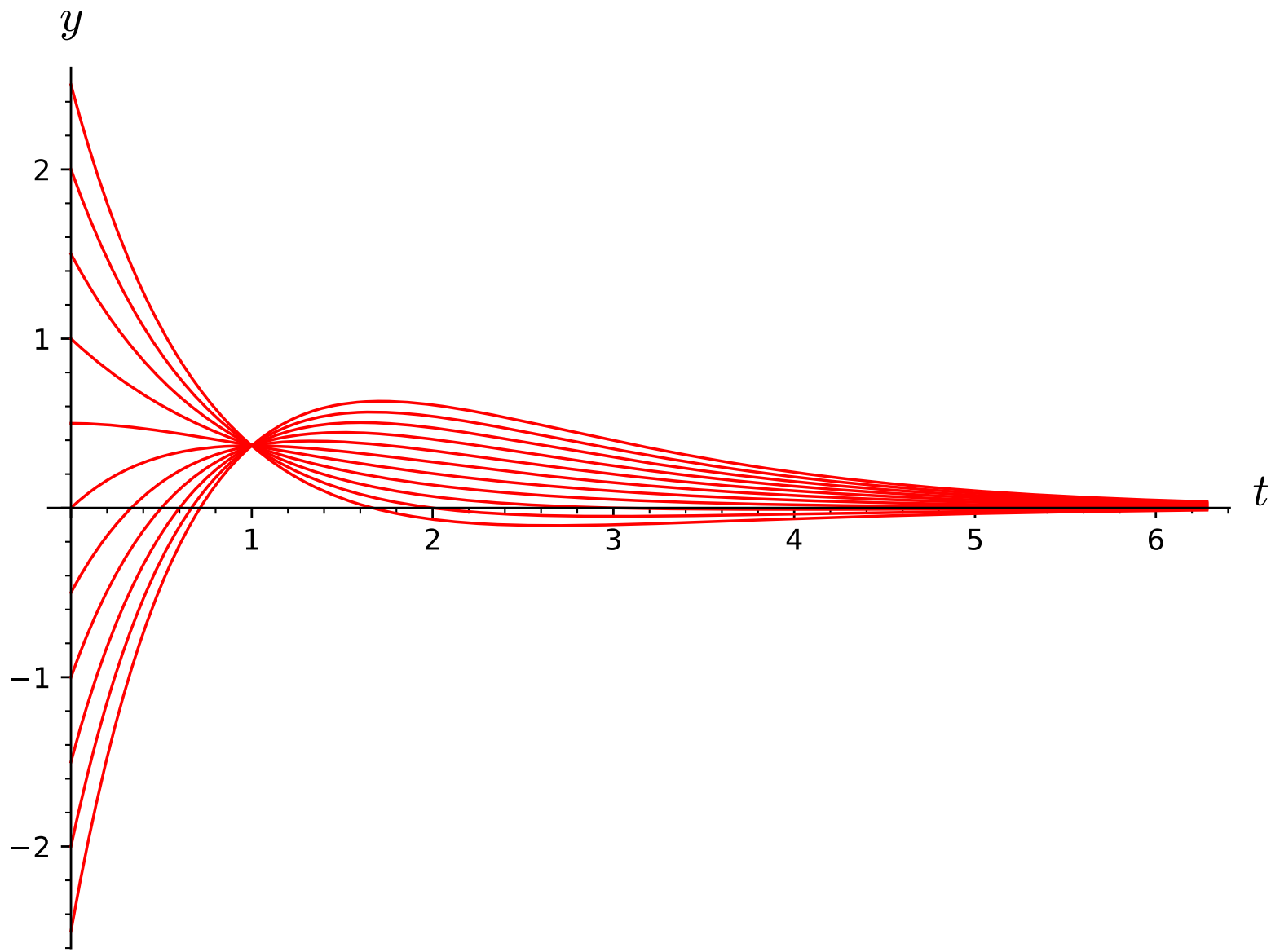


Figure: The solutions of $y'' + 2y' + y = 0$ satisfying $y(1) = 1/e \approx 0.368$ form a 1-parameter family, viz. $y(t) = c\,\mathrm{e}^{-t} + (1-c)t\,\mathrm{e}^{-t}$, $c \in \mathbb{R}$.

The Inhomogeneous Case

cf. also [BDM17], Ch. 4.3

The general solution of

$$y^{(n)} + a_{n-1}y^{(n-1)} + \cdots + a_1 y' + a_0 y = b(t) \qquad \text{(I)}$$

($a_0, \ldots, a_{n-1} \in \mathbb{C}$, $b\colon I \to \mathbb{C}$) has the form $y(t) = y_h(t) + y_p(t)$, where y_p denotes one particular solution and y_h the general solution of the associated homogeneous ODE (H). This is proved in the same way as in a previously considered example case.

For a general continuous "source" $b(t)$ order reduction and variation of parameters in the general solution of the resulting 1st-order system provide a method for finding a particular solution. This will be discussed later. Here we consider only the case

$$b(t) = f(t)\mathrm{e}^{\mu t} \quad \text{with } \mu \in \mathbb{C},\ f(X) \in \mathbb{C}[X].$$

The solution of this special case allows us to solve $a(\mathrm{D})y = b(t)$ for any exponential polynomial $b(t) = \sum_{i=1}^{s} f_i(t)\mathrm{e}^{\mu_i t}$ according to the

Superposition principle

If $y_1(t)$ solves $a(\mathrm{D})y = b_1(t)$ and $y_2(t)$ solves $a(\mathrm{D})y = b_2(t)$ then $y(t) = c_1 y_1(t) + c_2 y_2(t)$ solves $a(\mathrm{D})y = c_1 b_1(t) + c_2 b_2(t)$ ($c_1, c_2 \in \mathbb{C}$).

Theorem

Suppose μ is a root of $a(X)$ of multiplicity m ("$m = 0$" means "not a root of $a(X)$") and $f(X)$ has degree k. Then $a(\mathrm{D})y = f(t)\mathrm{e}^{\mu t}$ has a (unique) solution $y(t)$ of the form

$$y(t) = t^m(c_0 + c_1 t + \cdots + c_k t^k)\mathrm{e}^{\mu t}.$$

Proof.

We may assume $a(X) = \prod_{i=1}^{r}(X - \lambda_i)^{m_i}$ with $\mu = \lambda_1$, $m_1 = m$ (provided $m \geq 1$). We have seen that $\mathrm{D} - \lambda_i \,\mathrm{id}$ acts on the spaces

$$\mathrm{G}_\mu^{(k)} = \big\{t \mapsto g(t)\mathrm{e}^{\mu t}; g(X) \in \mathbb{C}[X], \deg g(X) \leq k\big\}, \quad t = 0, 1, 2, \ldots,$$

bijectively if $i \geq 2$ and maps $\mathrm{G}_\mu^{(k)}$ onto $\mathrm{G}_\mu^{(k-1)}$ (with the convention $\mathrm{G}_\mu^{(-1)} = \{0\}$) if $i = 1$.

Since $a(\mathrm{D})$ is the composition of such operators, with exactly m of them equal to $\mathrm{D} - \mu \,\mathrm{id}$, it is clear that $a(\mathrm{D})$ maps $\mathrm{G}_\mu^{(m+k)}$ surjectively onto $\mathrm{G}_\mu^{(k)}$.

In other words, there exists $g(X) \in \mathbb{C}[X]$ of degree $\leq m + k$ such that $a(\mathrm{D})\big(g(t)\mathrm{e}^{\mu t}\big) = f(t)\mathrm{e}^{\mu t}$.

Moreover, since $a(\mathrm{D})$ annihilates $t^j\mathrm{e}^{\mu t}$ for $0 \leq j \leq m - 1$, we can choose $g(X)$ of the form $g(X) = X^m(c_0 + c_1 X + \cdots + c_k X^k)$. □

Notes

- The proof remains valid for $m = 0$, if we omit the normalization $\mu = \lambda_1$.
- In the case $m = k = 0$, i.e. $b(t) = \mathrm{e}^{\mu t}$ with μ not a root of $a(X)$, we have $a(\mu) \neq 0$ and we can solve $a(\mathrm{D})y = \mathrm{e}^{\mu t}$ directly as follows:

$$a(\mathrm{D})\mathrm{e}^{\mu t} = a(\mu)\mathrm{e}^{\mu t} \implies a(\mathrm{D})\left(\frac{1}{a(\mu)}\,\mathrm{e}^{\mu t}\right) = \mathrm{e}^{\mu t}.$$

- If you have difficulties to understand the argument using the differential operators $\mathrm{D} - \mu\,\mathrm{id}$, consider first the special case $\mu = 0$, in which the operator is just $\mathrm{D}\colon y \to y'$ and $\mathrm{G}_0^{(k)}$ is the space of polynomials of degree $\leq k$. From Calculus I we know that differentiation decreases the degree of a polynomial by 1 (except for the constant case) and hence maps $\mathrm{G}_0^{(k)}$ onto $\mathrm{G}_0^{(k-1)}$. On the other hand, if $\lambda \neq 0$ then

$$\mathrm{D}(t^k\mathrm{e}^{\lambda t}) = kt^{k-1}\mathrm{e}^{\lambda t} + t^k\lambda\mathrm{e}^{\lambda t} = (\lambda t^k + kt^{k-1})\mathrm{e}^{\lambda t},$$

  showing that in this case the degree of any non-constant polynomial factor is preserved and $\{g(t)\mathrm{e}^{\lambda t}; \deg g(X) \leq k\}$ is mapped bijectively onto itself.

Example

We determine the general solution of the 3rd-order ODE

$$y''' + 2y'' + y' = t + 2\mathrm{e}^{-t}.$$

The characteristic polynomial is
$a(X) = X^3 + 2X^2 + X = X(X+1)^2$.
$\Longrightarrow$ 1, e^{-t}, $t\mathrm{e}^{-t}$ form a fundamental system of solutions of $y''' + 2y'' + y' = 0$.

A particular solution of the inhomogeneous ODE can be obtained by solving

1. $y''' + 2y'' + y' = t$,
2. $y''' + 2y'' + y' = \mathrm{e}^{-t}$,

and applying superposition.

(1) Here $\mu = 0$, which is a root of $a(X)$ of multiplicity 1.
$\Longrightarrow$ The „Ansatz“ $y_1(t) = c_1 t + c_2 t^2$ yields a solution.
Substituting this in the ODE (1) gives

$$2(2c_2) + (c_1 + 2c_2 t) = c_1 + 4c_2 + 2c_2 t = t.$$

Example (cont'd)

The solution is $c_2 = \frac{1}{2}$, $c_1 = -2$, i.e., $y_1(t) = -2t + \frac{1}{2}t^2$.

(2) Here $\mu = -1$, which is a root of $a(X)$ of multiplicity 2.

$\Longrightarrow$ The „Ansatz“ $y_2(t) = c\,t^2\mathrm{e}^{-t}$ yields a solution.

$$a(\mathrm{D})y_2(t) = c\,\mathrm{D}(\mathrm{D}+\mathrm{id})^2(t^2\mathrm{e}^{-t}) = 2c\,\mathrm{D}(\mathrm{D}+\mathrm{id})(t\mathrm{e}^{-t}) = 2c\,\mathrm{D}\mathrm{e}^{-t} = -2c\,\mathrm{e}^{-t}$$

$\Longrightarrow$ $y_2(t) = -\frac{1}{2}t^2\mathrm{e}^{-t}$ solves the ODE (2).

Finally, superposition gives that

$$y(t) = y_1(t) + 2y_2(t) = -2t + \frac{1}{2}t^2 - t^2\mathrm{e}^{-t}$$

solves the original ODE $a(\mathrm{D})y = t + 2\mathrm{e}^{-t}$.

The general complex (real) solution of the original ODE is therefore

$$y(t) = c_1 + c_2\mathrm{e}^{-t} + c_3 t\mathrm{e}^{-t} - 2t + \frac{1}{2}t^2 - t^2\mathrm{e}^{-t}$$

with constants $c_1, c_2, c_3 \in \mathbb{C}$ (respectively, $c_1, c_2, c_3 \in \mathbb{R}$).

Further Notes

- Pay attention to the fact that in the monomial case $b(t) = t^k \mathrm{e}^{\mu t}$ the correct „Ansatz“ is $y(t) = t^m(c_0 + c_1 t + \cdots + c_k t^k)\mathrm{e}^{\mu t}$ (i.e., a full exponential polynomial).
- Before applying superposition, collect monomials with the same factors $\mathrm{e}^{\mu t}$. For example, when solving $a(\mathrm{D})y = 1 + t + 2\,\mathrm{e}^{-t}$, use $b_1(t) = (1 + t)\mathrm{e}^{0\,t}$, $b_2(t) = \mathrm{e}^{-t}$ (and not a superposition of three solutions corresponding to 1, t, e^{-t}). This saves computation time.

Example (Harmonic oszillator with periodic source)

The corresponding ODE is

$$\frac{\mathrm{d}^2 y}{\mathrm{d}t^2} + \omega_0^2 y = A\cos(\omega t), \quad \omega_0, \omega, A > 0.$$

ω_0 denotes the characteristic frequency of the oszillator and ω the frequency of the external source.

In order to apply the machinery developed, we consider the "complexified" ODE

$$y'' + \omega_0^2 y = A\,\mathrm{e}^{\mathrm{i}\omega t}.$$

The real part of any particular solution of the complex ODE will then solve the real ODE.

The characteristic polynomial $a(X) = X^2 + \omega_0^2 = (X - \mathrm{i}\omega_0)(X + \mathrm{i}\omega_0)$ has roots $\lambda_{1,2} = \pm\mathrm{i}\omega_0$ (the same as in the homogeneous case $b(t) = 0$).

Hence we need to distinguish the cases $\omega = \omega_0$ (the so-called *resonance* case) and $\omega \neq \omega_0$.

Example (cont'd)

Case 1: $\omega \neq \omega_0$.

In this case the „Ansatz" $y(t) = c\,\mathrm{e}^{\mathrm{i}\omega t}$ yields a solution.

$$a(\mathrm{D})y(t) = c\,a(\mathrm{i}\omega)\mathrm{e}^{\mathrm{i}\omega t} = c(\omega_0^2 - \omega^2)\mathrm{e}^{\mathrm{i}\omega t} = A\,\mathrm{e}^{\mathrm{i}\omega t}$$

$$\Longrightarrow c = \frac{A}{a(\mathrm{i}\omega)} = \frac{A}{\omega_0^2 - \omega^2}, \quad \text{and a real particular solution is}$$

$$y(t) = \frac{A}{\omega_0^2 - \omega^2}\cos(\omega t).$$

Case 2: $\omega = \omega_0$.

In this case the „Ansatz" $y(t) = c\,t\mathrm{e}^{\mathrm{i}\omega_0 t}$ yields a solution.

$$\begin{aligned} a(\mathrm{D})y(t) &= c(\mathrm{D} + \mathrm{i}\omega_0\ \mathrm{id})(\mathrm{D} - \mathrm{i}\omega_0\ \mathrm{id})(t\,\mathrm{e}^{\mathrm{i}\omega_0 t}) = c(\mathrm{D} + \mathrm{i}\omega_0\ \mathrm{id})(\mathrm{e}^{\mathrm{i}\omega_0 t}) \\ &= c(2\mathrm{i}\omega_0)\mathrm{e}^{\mathrm{i}\omega_0 t} = A\,\mathrm{e}^{\mathrm{i}\omega_0 t} \end{aligned}$$

$$\Longrightarrow c = \frac{A}{2\mathrm{i}\omega_0}\left(= \frac{A}{a'(\mathrm{i}\omega_0)}\right), \quad \text{and a real particular solution is}$$

$$y(t) = \frac{A}{2\omega_0}t\sin(\omega_0 t).$$

$\Longrightarrow$ The general real solution of $y'' + \omega_0^2 y = A\cos(\omega t)$ is

$$y(t) = c_1\cos(\omega_0 t) + c_2\sin(\omega_0 t) + \begin{cases} \frac{A}{\omega_0^2-\omega^2}\cos(\omega t) & \text{if } \omega \neq \omega_0, \\ \frac{A}{2\omega_0} t\sin(\omega_0 t) & \text{if } \omega = \omega_0 \end{cases}$$

with constants $c_1, c_2 \in \mathbb{R}$.

Notes

- The resonance phenomenon must, e.g., be taken into account when constructing bridges, which are subject to vertical vibrations caused by the airflow around the bridge. If the frequency of the external force (which is periodic) matches the natural frequency of the bridge's material (steel), vibrations are amplified—leading ultimately to disaster ($\longrightarrow$ *Tacoma Narrows Bridge*).
- In the preceding examples we have sometimes used the factorization of $a(D)$, which we knew from solving the associated homogeneous ODE $a(D)y = 0$, to speed up the computation of $a(D)y$ for certain functions y. The standard method for obtaining $a(D)y$ is of course to compute all derivatives $y', y'', \ldots, y^{(n)}$ and then form the linear combination $y^{(n)} + \cdots + a_2 y'' + a_1 y' + a_0 y$.

Notes cont'd

- Higher-order linear ODE's with constant coefficients are discussed in [BDM17], Ch. 4.2–4.4. Our theorem for the inhomogeneous case can be viewed as a generalization of the "method of undetermined coefficients" in Ch. 4.3; cf. also Ex. 14 in this chapter. The "method of annihilators", discussed in exercises for the same chapter, provides an alternative but equivalent approach to our result. The "method of variation of parameters" in Ch. 4.4, which is more powerful, will be discussed within the framework of linear ODE systems later in the course.
- The explicit formula $y(t) = \frac{1}{a(\mu)}\,\mathrm{e}^{\mu t}$ for the solution of $a(\mathrm{D})y = \mathrm{e}^{\mu t}$, which requires μ to be a nonzero of the characteristic polynomial $a(X)$, admits the generalization $y(t) = \frac{1}{a^{(m)}(\mu)}t^m\mathrm{e}^{\mu t}$ in the case where μ is a root of multiplicity m of $a(X)$; cp. the solution in Case 2 of the example. For the proof write $a(X) = (X-\mu)^m A(X)$ and compute $a(\mathrm{D})[t^m\mathrm{e}^{\mu t}] = A(\mathrm{D})(\mathrm{D}-\mu)^m[t^m\mathrm{e}^{\mu t}] = A(\mathrm{D})[m!\mathrm{e}^{\mu t}] = m!A(\mu)\mathrm{e}^{\mu t}$. This yields the solution $y(t) = \frac{1}{m!A(\mu)}t^m\mathrm{e}^{\mu t}$, and the Leibniz formula for the mth derivative $\mathrm{D}^m(fg)$ of a product of two functions can then be used to show that $a^{(m)}(\mu) = m!A(\mu)$.

Exercise

Determine a real fundamental system of solutions for the following ODE's:

a) $y'' - 4y' + 4y = 0$;

b) $y''' - 2y'' - 5y' + 6y = 0$;

c) $y''' - 2y'' + 2y' - y = 0$;

d) $y''' - y = 0$;

e) $y^{(4)} + y = 0$;

f) $y^{(8)} + 4y^{(6)} + 6y^{(4)} + 4y'' + y = 0$.

Exercise

Determine the general real solution of

a) $y'' + 3y' + 2y = 2$;

b) $y'' + y' - 12y = 1 + t^2$;

c) $y'' - 5y' + 6y = 4te^t - \sin t$;

d) $y''' - 2y'' + y' = 1 + e^t \cos(2t)$;

e) $y^{(4)} + 2y'' + y = 25e^{2t}$;

f) $y^{(n)} = te^t$, $n \in \mathbb{N}$.

Exercise

For $a, b \in \mathbb{C}$ consider the ODE

$$y'' + \frac{a}{t}\,y' + \frac{b}{t^2}\,y = 0 \qquad (t > 0). \tag{1}$$

a) Show that $\phi\colon \mathbb{R}^+ \to \mathbb{C}$ is a solution of (1) iff $\psi\colon \mathbb{R} \to \mathbb{C}$ defined by $\psi(s) = \phi(\mathrm{e}^s)$ is a solution of

$$y'' + (a-1)y' + by = 0. \tag{2}$$

b) Determine the general solution of (1) for $(a, b) = (6, 4)$ and $(a, b) = (3, 1)$.

Exercise

Solve the initial value problem

$$\mathbf{y}' = \begin{pmatrix} 1 & 2 \\ 3 & 6 \end{pmatrix}\mathbf{y} + \begin{pmatrix} t \\ \sin t \end{pmatrix}, \quad \mathbf{y}(0) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Definition

Suppose $\mathbf{y} = (y_0, y_1, y_2, \dots)$ is a sequence of (complex) numbers. The (formal) power series

$$f(t) = \sum_{n=0}^{\infty} \frac{y_n}{n!}\, t^n = \frac{y_0}{0!} + \frac{y_1}{1!}\, t + \frac{y_2}{2!}\, t^2 + \frac{y_3}{3!}\, t^3 + \cdots$$

is called *exponential generating function* of $\mathbf{y}$ and denoted by $\mathrm{egf}(\mathbf{y})$.

Notes

- $\mathrm{egf}(\mathbf{y}) = f(t)$ contains all information about the sequence $\mathbf{y}$. This is true at least in a formal sense, but in the case where $f(t)$ has radius of convergence $R > 0$ can also be seen by term-wise differentiation: $y_n = f^{(n)}(0)$ is determined by f.
- Exponential generating functions (and likewise their ordinary counterparts $g(t) = \sum_{n=0}^{\infty} y_n t^n$) are used with great success in Enumerative Combinatorics. The most famous example is the sequence $\mathbf{d} = (d_0, d_1, d_2, \dots) = (1, 0, 1, 2, 9, \dots)$ of fixed-point free permutations of n letters (so-called *derangements*), whose exponential generating function turns out to be $\frac{\mathrm{e}^{-t}}{1-t}$, showing that $d_n = n! \sum_{k=0}^{n} \frac{(-1)^k}{k!}$.

Notes cont'd

- If the sequence $\mathbf{y} = (y_0, y_1, y_2, \dots)$ grows at most exponentially, i.e., there exists a constant $C > 1$ such that $|y_n| \le C^n$ for sufficiently large n, then $\operatorname{egf}(\mathbf{y})$ has radius of convergence $R = \infty$. This follows from the elementary estimate $n! \ge (n/\mathrm{e})^n$, which implies $\sqrt[n]{|y_n|/n!} \le C\mathrm{e}/n \to 0$ for $n \to \infty$. All homogeneous linear recurring sequences with constant coefficients have this property (as we know from Discrete Mathematics). The same is true in the inhomogeneous case, provided that the right-hand side $\mathbf{b} = (b_0, b_1, b_2, \dots)$ grows at most exponentially, as is easily proved.

Theorem

1. *For all sequences* $\mathbf{y} \in \mathbb{C}^{\mathbb{N}}$ *and polynomials* $p(X) \in \mathbb{C}[X]$ *we have*
$$p(\mathrm{D})\,\mathrm{egf}(\mathbf{y}) = \mathrm{egf}\big(p(\mathrm{S})\mathbf{y}\big).$$

2. *Now suppose* $p(X)$ *is monic of degree* n*. The sequence* $\mathbf{y} = (y_0, y_1, y_2, \dots)$ *solves the linear recurrence relation* $p(\mathrm{S})\mathbf{y} = \mathbf{b}$ *iff the function* $y(t) = \mathrm{egf}(\mathbf{y})$ *solves the IVP* $p(\mathrm{D})y = \mathrm{egf}(\mathbf{b})$, $y^{(i)}(0) = y_i$ *for* $0 \le i \le n-1$.

Proof.

(1) We have

$$\mathrm{D}[\mathrm{egf}(\mathbf{y})] = \frac{\mathrm{d}}{\mathrm{d}t}\sum_{n=0}^{\infty}\frac{y_n}{n!}\,t^n = \sum_{n=1}^{\infty}\frac{n\,y_n}{n!}\,t^{n-1} = \sum_{n=0}^{\infty}\frac{y_{n+1}}{n!}\,t^n = \mathrm{egf}(\mathrm{S}\mathbf{y})$$

This implies $\mathrm{D}^k\,\mathrm{egf}(\mathbf{y}) = \mathrm{egf}(\mathrm{S}^k\mathbf{y})$ for $k \in \mathbb{N}$ and, since egf is $\mathbb{C}$-linear, further $p(\mathrm{D})\,\mathrm{egf}(\mathbf{y}) = \sum_{k=0}^{d} p_k\,\mathrm{D}^k\,\mathrm{egf}(\mathbf{y}) = \sum_{k=0}^{d} p_k\,\mathrm{egf}(\mathrm{S}^k\mathbf{y}) = \mathrm{egf}\left(\sum_{k=0}^{d} p_k\mathrm{S}^k\mathbf{y}\right) = \mathrm{egf}\big(p(\mathrm{S})\mathbf{y}\big)$.

Proof cont'd.

(2) If $p(\mathrm{S})\mathbf{y} = \mathbf{b}$ then $\operatorname{egf}\big(p(\mathrm{S})\mathbf{y}\big) = \operatorname{egf}(\mathbf{b})$, which on account of Part (1) means $p(\mathrm{D})\operatorname{egf}(\mathbf{y}) = \operatorname{egf}(\mathbf{b})$. Thus $y(t) := \operatorname{egf}(\mathbf{y})$ solves $p(\mathrm{D})y = \operatorname{egf}(\mathbf{b})$, and as remarked after the definition of $\operatorname{egf}(\mathbf{y})$ we have $y^{(i)}(0) = y_i$.

Conversely, suppose $y(t) = \sum_{k=0}^{\infty} \frac{y_k}{k!}\, t^k$ solves $p(\mathrm{D})y = \operatorname{egf}(\mathbf{b})$. Then $p(\mathrm{D})\operatorname{egf}(\mathbf{y}) = \operatorname{egf}(\mathbf{b})$ and hence $\operatorname{egf}\big(p(\mathrm{S})\mathbf{y}\big) = \operatorname{egf}(\mathbf{b})$ by Part (1). Since egf maps sequences bijectively onto formal power series, this implies $p(\mathrm{S})\mathbf{y} = \mathbf{b}$. □

Note

The theorem merely expresses the fact that term-wise differentiation of an exponential generating function amounts to shifting and truncating the corresponding sequence and that its derivatives evaluated at zero are just the entries of the sequence. If $\mathbf{b}$ grows at most exponentially then $b(t) = \operatorname{egf}(\mathbf{b})$ represents a function with domain $\mathbb{R}$ and $y(t) = \operatorname{egf}(\mathbf{y})$, which has also domain $\mathbb{R}$, forms a solution of the "real" ODE $p(\mathrm{D})y = b(t)$.

Example

The solution of the Fibonacci IVP $y'' = y' + y$, $y(0) = 0$, $y'(0) = 1$ is

$$y(t) = \sum_{n=0}^{\infty} \frac{f_n}{n!}\, t^n, \quad t \in \mathbb{R}.$$

If you have difficulties to see this, use the following argument (which also works in general):

Differentiating $y'' = y' + y$ repeatedly gives $y^{(n+2)} = y^{(n+1)} + y^{(n)}$, i.e., $(y, y', y'', \dots)$ satisfies the Fibonacci recurrence relation.
$\Longrightarrow$ $(y(0), y'(0), y''(0), \dots)$ is the Fibonacci sequence, because the initial conditions are the same.

$$\Longrightarrow\ y(t) = \sum_{n=0}^{\infty} \frac{y^{(n)}(0)}{n!}\, t^n = \sum_{n=0}^{\infty} \frac{f_n}{n!}\, t^n$$

Of course, one must also provide an argument that the solution is analytic (i.e., represented by its Taylor series). The machinery developed gives that solutions are exponential polynomials (hence analytic). One could also use the ODE to bound $y^{(n)}$ recursively, and use this bound in turn to show that the remainder in the Taylor expansion of y converges to zero.

Example

We consider again $y'' = 4y' - 4y$, this time with particular initial values $y(0) = y'(0) = 1$. The corresponding characteristic polynomial is $X^2 - 4X + 4 = (X-2)^2$, so that the general solution of $y'' = 4y' - 4y$ has the form $y(t) = c_1 \mathrm{e}^{2t} + c_2 t \mathrm{e}^{2t}$.
Since $y'(t) = 2c_1 \mathrm{e}^{2t} + c_2(1+2t)\mathrm{e}^{2t}$, we obtain the system $c_1 = 2c_1 + c_2 = 1$, $c_2 = -1$, and hence $y(t) = \mathrm{e}^{2t} - t\mathrm{e}^{2t}$.

The corresponding discrete IVP is $y_{k+2} = 4y_{k+1} - 4y_k$, $y_0 = y_1 = 1$. Here we have $y_i = c_1 2^k + c_2 k 2^k$, and the initial conditions give $c_1 = 1$, $c_2 = -1/2$, so that $y_k = 2^k - k2^{k-1}$.

By the theorem the solutions must be related by $y(t) = \mathrm{egf}(\mathbf{y})$. Indeed, we have

$$y(t) = \sum_{k=0}^{\infty} \frac{2^k t^k}{k!} - \sum_{k=0}^{\infty} \frac{2^k t^{k+1}}{k!} = \sum_{k=0}^{\infty} \frac{2^k}{k!} t^k - \sum_{k=1}^{\infty} \frac{2^{k-1}}{(k-1)!} t^k$$
$$= \sum_{k=0}^{\infty} \frac{2^k - k2^{k-1}}{k!} t^k = \mathrm{egf}(\mathbf{y}).$$

Example (cont'd)

The coefficients c_1, c_2 in the discrete and continuous case are not the same. This is due to the fact that the chosen fundamental systems are not mapped onto each other by $\mathbf{y} \mapsto \mathrm{egf}(\mathbf{y})$. Rather, an exponential monomial

$$\begin{aligned} t^k \mathrm{e}^{\lambda t} &= \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} t^{n+k} = \sum_{n=k}^{\infty} \frac{\lambda^{n-k}}{(n-k)!} t^n \\ &= \sum_{n=0}^{\infty} \frac{n(n-1)\cdots(n-k+1)\lambda^{n-k}}{n!} t^n \\ &= \lambda^{-k} \sum_{n=0}^{\infty} \frac{n(n-1)\cdots(n-k+1)\lambda^{n}}{n!} t^n \end{aligned}$$

is (up to a constant factor) the egf of the sequence $y_n = n(n-1)\cdots(n-k+1)\lambda^n$, which involves falling factorials instead of the powers n^k. In our example the difference is not really visible, but it becomes apparent if there are fundamental solutions with $k = 2$ (i.e., the characteristic polynomial has a zero of multiplicity ≥ 3).

Math 285
Introduction to Differential Equations

Thomas Honold

ZJU-UIUC Institute

Spring Semester 2024

Outline

Today's Lecture:

First-Order Linear Systems

Definition

A (possibly *time-dependent*) *first-order linear system of ODE's* has the form

$$\mathbf{y}' = \mathbf{A}(t)\mathbf{y} + \mathbf{b}(t), \qquad \text{(LS)}$$

where $\mathbf{A}\colon I \to \mathbb{C}^{n\times n}$, $t \mapsto \mathbf{A}(t) = \big(a_{ij}(t)\big)$ and $\mathbf{b}\colon I \to \mathbb{C}^n$, $t \mapsto \mathbf{b}(t) = \big(b_i(t)\big)$ are continuous (i.e., all component functions of $\mathbf{A}$ and $\mathbf{b}$ are continuous).

The domain I must be an interval contained in the domains of all component functions. The cases $I = \emptyset$ and $I = \{a\}$ are excluded.

As usual, the system (LS) is said to be *homogeneous* if $\mathbf{b}(t) \equiv 0$ and *inhomogeneous* otherwise.

A solution of (LS) is a differentiable map $\mathbf{y}\colon J \to \mathbb{C}^n$ (i.e., a parametric curve) defined on some subinterval $J \subseteq I$ and satisfying $\mathbf{y}'(t) = \mathbf{A}(t)\mathbf{y}(t) + \mathbf{b}(t)$ for all $t \in J$.

Existence and Uniqueness of Solutions

$f(t, \mathbf{y}) = \mathbf{A}(t)\mathbf{y} + \mathbf{b}(t)$ satisfies

$$|f(t, \mathbf{y}_1) - f(t, \mathbf{y}_2)| = |\mathbf{A}(t)(\mathbf{y}_1 - \mathbf{y}_2)| \le \|\mathbf{A}(t)\| \, |\mathbf{y}_1 - \mathbf{y}_2| \le L \, |\mathbf{y}_1 - \mathbf{y}_2| \,,$$

provided we restrict t to compact (closed and bounded) subintervals of I, and hence in particular a local Lipschitz condition with respect to $\mathbf{y}$.

$\Longrightarrow$ The Existence and Uniqueness Theorem applies and gives the local solvability and uniqueness of solutions of any IVP $\mathbf{y}' = \mathbf{A}(t)\mathbf{y} + \mathbf{b}(t) \wedge \mathbf{y}(t_0) = \mathbf{y}_0$ ($t_0 \in I$, $\mathbf{y}_0 \in \mathbb{C}^n$).

Remark

If you are uncomfortable with complex-valued linear ODE systems, note that any such system is equivalent to a real valued system with twice as many equations/component functions in the following sense: $\mathbf{y}(t)$ solves the complex $n \times n$ system $\mathbf{y}' = \mathbf{A}(t)\mathbf{y} + \mathbf{b}(t)$ iff $\begin{pmatrix} \operatorname{Re}\mathbf{y}(t) \\ \operatorname{Im}\mathbf{y}(t) \end{pmatrix}$ solves $\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix}' = \begin{pmatrix} \operatorname{Re}\mathbf{A}(t) & -\operatorname{Im}\mathbf{A}(t) \\ \operatorname{Im}\mathbf{A}(t) & \operatorname{Re}\mathbf{A}(t) \end{pmatrix} \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} + \begin{pmatrix} \operatorname{Re}\mathbf{b}(t) \\ \operatorname{Im}\mathbf{b}(t) \end{pmatrix}$.

But More Is True

Theorem

Solutions to any IVP $\mathbf{y}' = \mathbf{A}(t)\mathbf{y} + \mathbf{b}(t) \wedge \mathbf{y}(t_0) = \mathbf{y}_0$ *exist on the whole domain I (and are unique).*

Compare this with the nonlinear case, e.g., $y' = y^2$ whose solutions $y = 1/(C - x)$ do not exist on all of $\mathbb{R}$.

Proof.

We estimate the difference of successive Picard-Lindelöf iterates $\mathbf{y}_k(t) = \mathbf{y}_0 + \int_{t_0}^{t} \mathbf{A}(s)\mathbf{y}_{k-1}(s) + \mathbf{b}(s)\,\mathrm{d}s$, $k = 1, 2, 3, \ldots$

$$|\mathbf{y}_{k+1}(t) - \mathbf{y}_k(t)| = \left| \int_{t_0}^{t} \mathbf{A}(s)\big(\mathbf{y}_k(s) - \mathbf{y}_{k-1}(s)\big)\,\mathrm{d}s \right|$$

$$\leq \pm L \int_{t_0}^{t} |\mathbf{y}_k(s) - \mathbf{y}_{k-1}(s)|\,\mathrm{d}s,$$

provided t is restricted to a compact subinterval $J \subseteq I$ with $t_0 \in J$ and $L = \max\{\|\mathbf{A}(t)\|\,; t \in J\}$. (Recall that "$\pm$" is necessary to include the case $t < t_0$.)

Proof cont'd.

$$|\mathbf{y}_1(t) - \mathbf{y}_0| \le K := \max\{|\mathbf{y}_1(t) - \mathbf{y}_0|\,; t \in J\},$$

$$|\mathbf{y}_2(t) - \mathbf{y}_1(t)| \le \pm L \int_{t_0}^{t} K\,\mathrm{d}s = LK\,|t - t_0|\,,$$

$$|\mathbf{y}_3(t) - \mathbf{y}_2(t)| \le \pm L \int_{t_0}^{t} LK\,|s - t_0|\,\mathrm{d}s = L^2K\,\frac{|t - t_0|^2}{2!},$$

and in general, using mathematical induction,

$$|\mathbf{y}_{k+1}(t) - \mathbf{y}_k(t)| \le L^kK\,\frac{|t - t_0|^k}{k!}.$$

Setting $J = [a, b]$ we can bound the (vectorial) function series $\sum_{k=0}^{\infty}(\mathbf{y}_{k+1} - \mathbf{y}_k)$ independently of t by the convergent series

$$\sum_{k=0}^{\infty} K\,\frac{L^k(b-a)^k}{k!} = K\mathrm{e}^{L(b-a)}.$$

Proof cont'd.

By Weierstrass's Criterion, this implies that the function sequence $(\mathbf{y}_k(t))$ converges uniformly on J, the limit function $\mathbf{y}_\infty(t) = \lim_{k\to\infty} \mathbf{y}(t)$ is continuous on J and satisfies the integral equation

$$\mathbf{y}_\infty(t) = \mathbf{y}_0 + \int_{t_0}^{t} \mathbf{A}(s)\mathbf{y}_\infty(s) + \mathbf{b}(s)\,\mathrm{d}s, \quad t \in J.$$

(The fixed-point property $T\mathbf{y}_\infty = \mathbf{y}_\infty$ requires continuity of the operator $(T\phi)(t) = \mathbf{y}_0 + \int_{t_0}^{t} \mathbf{A}(s)\phi(s) + \mathbf{b}(s)\,\mathrm{d}s$ in the metric of uniform convergence on $[a, b]$. From the previous estimate we can infer $\|T\phi_1 - T\phi_2\|_\infty \le L(b-a)\,\|\phi_1 - \phi_2\|_\infty$, i.e., T is Lipschitz-continuous; cf. also the 1-dimensional example $y' = 2t\,y$ discussed after the Existence Theorem.)

$\Longrightarrow$ The Fundamental Theorem of Calculus gives $\mathbf{y}'_\infty(t) = \mathbf{A}(t)\mathbf{y}_\infty(t) + \mathbf{b}(t)$ for $t \in J$, and of course $\mathbf{y}_\infty(t_0) = \mathbf{y}_0$.

Finally, since an arbitrary interval I can be exhausted by compact intervals, i.e., $I = \bigcup_{m=1}^{\infty} J_m$ with $J_m = [a_m, b_m]$ and $J_1 \subseteq J_2 \subseteq J_3 \subseteq \cdots$, we obtain that $(\mathbf{y}_k(t))$ converges on I, and the limit function $\mathbf{y}_\infty : I \to \mathbb{C}^n$ solves the IVP as well. □

Note on the proof

You can see what goes wrong with the proof in the nonlinear case by computing the Picard-Lindelöf iterates for the IVP $y' = y^2 \wedge y(0) = 1$, whose solution is $y(t) = 1/(1-t)$.

$$\phi_0(t) = 1,$$

$$\phi_1(t) = 1 + \int_0^t 1^2 \,\mathrm{d}s = t + 1,$$

$$\phi_2(t) = 1 + \int_0^t (s+1)^2 \,\mathrm{d}s = 1 + \left[\tfrac{1}{3}s^3 + s^2 + s\right]_0^t = \tfrac{1}{3}t^3 + t^2 + t + 1,$$

$$\phi_3(t) = 1 + \int_0^t \left(\tfrac{1}{3}s^3 + s^2 + s + 1\right)^2 \mathrm{d}s = \cdots$$

$$= 1 + t + t^2 + t^3 + \tfrac{2}{3}t^4 + \tfrac{1}{3}t^5 + \tfrac{1}{9}t^6 + \tfrac{1}{63}t^7$$

With some effort one can show in general that $\phi_k(t)$ is a polynomial in t which starts with $1 + t + t^2 + \cdots + t^k$ and has the remaining coefficients in $[0, 1)$. It follows that $\sum_{j=0}^{k} t^j \le \phi_k(t) \le \sum_{j=0}^{\infty} t^j = \frac{1}{1-t}$ and $\phi_k(t) \to 1/(1-t)$ for $k \to \infty$ if $0 \le t < 1$. For $t \ge 1$ the sequence $\big(\phi_k(t)\big)$ does not converge, and hence nothing prevents the solution from blowing up at $t = 1$.

The Link with Linear Algebra

Theorem

The solutions of a homogeneous linear system $\mathbf{y}' = \mathbf{A}(t)\mathbf{y}$ *with* $\mathbf{A}\colon I \to \mathbb{C}^{n\times n}$ *form an n-dimensional vector space over* $\mathbb{C}$*. For solutions* $\mathbf{y}_1, \dots, \mathbf{y}_k\colon I \to \mathbb{C}^n$ *of* $\mathbf{y}' = \mathbf{A}(t)\mathbf{y}$ *the following are equivalent:*

1. *The functions* $\mathbf{y}_1, \dots, \mathbf{y}_k \in (\mathbb{C}^n)^I$ *are linearly independent.*
2. *For some* $t_0 \in I$ *the vectors* $\mathbf{y}_1(t_0), \dots, \mathbf{y}_k(t_0) \in \mathbb{C}^n$ *are linearly independent.*
3. *For all* $t_0 \in I$ *the vectors* $\mathbf{y}_1(t_0), \dots, \mathbf{y}_k(t_0)$ *are linearly independent.*

Proof.

If $\mathbf{y}_1, \mathbf{y}_2\colon I \to \mathbb{C}^n$ are solutions then
$(\mathbf{y}_1 + \mathbf{y}_2)' = \mathbf{y}_1' + \mathbf{y}_2' = \mathbf{A}(t)\mathbf{y}_1 + \mathbf{A}(t)\mathbf{y}_2 = \mathbf{A}(t)(\mathbf{y}_1 + \mathbf{y}_2)$,
i.e., $\mathbf{y}_1 + \mathbf{y}_2$ is a solution as well. Similarly, scalar multiples of solutions are again solutions, and of course the all-zero function is a solution. This proves that the solutions form a vector space over $\mathbb{C}$ (subspace of the vectorial function space $(\mathbb{C}^n)^I$).

Proof cont'd.

Next we prove the equivalences. The implications (3)$\Longrightarrow$(2)$\Longrightarrow$(1) are trivial and it remains to show (1)$\Longrightarrow$(3).

Suppose $\mathbf{y}_1, \ldots, \mathbf{y}_k$ are linearly independent and that $c_1\mathbf{y}_1(t_0) + \cdots + c_k\mathbf{y}_k(t_0) = \mathbf{0} \in \mathbb{C}^n$. Then the two solutions $c_1\mathbf{y}_1 + \cdots + c_k\mathbf{y}_k$ and $\mathbf{y} \equiv 0$ agree at $t = t_0$.

$\Longrightarrow$ By the Existence and Uniqueness Theorem, they must agree everywhere, i.e., $c_1\mathbf{y}_1 + \cdots + c_k\mathbf{y}_k = 0$ in $(\mathbb{C}^n)^I$.

$\Longrightarrow$ $c_1 = \cdots = c_k = 0$, since $\mathbf{y}_1, \ldots, \mathbf{y}_k$ are linearly independent.

This proves (1)$\Longrightarrow$(3).

Finally we show that the solution space V has dimension n.

Fix $t_0 \in I$ and consider the evaluation map $V \to \mathbb{C}^n$, $\mathbf{y} \mapsto \mathbf{y}(t_0)$, which is obviously linear.

Since (1)$\Longrightarrow$(2), this map is injective.

Since every IVP $\mathbf{y}' = \mathbf{A}(t)\mathbf{y} \wedge \mathbf{y}(t_0) = \mathbf{y}_0 \in \mathbb{C}^n$ is solvable, the map is surjective.

$\Longrightarrow$ The map is a vector space isomorphism and $\dim V = \dim \mathbb{C}^n = n$.

Remarks

1. The same Theorem holds, mutatis mutandis, for real-valued solutions of "real" linear systems $\mathbf{y}' = \mathbf{A}(t)\mathbf{y}$ with $\mathbf{A}\colon I \to \mathbb{R}^{n\times n}$. This can be proved in the same way or inferred from the complex case.
2. A basis $\mathbf{y}_1, \dots, \mathbf{y}_n$ of the solution space of $\mathbf{y}' = \mathbf{A}(t)\mathbf{y}$ is called a *fundamental system of solutions*. The equivalence (1)$\Longrightarrow$(2) yields the following handy test for fundamental systems:

   Writing $\mathbf{y}_j(t) = \big(y_{1j}(t), \dots, y_{nj}(t)\big)^{\mathsf{T}}$ as columns of a matrix $\mathbf{\Phi}(t)$ (so-called *fundamental matrix*), we have that $\{\mathbf{y}_1, \dots, \mathbf{y}_n\}$ is a fundamental system of solutions iff $\mathbf{\Phi}(t_0)$ has rank n for some (and hence all) $t_0 \in I$.
3. With $\mathbf{\Phi}(t)$ as in (2) we have, using matrix-vector multiplication for functions, the matrix version $\mathbf{\Phi}'(t) = \mathbf{A}(t)\mathbf{\Phi}(t)$ of the homogeneous ODE system and the representation

$$\mathbf{y}(t) = c_1\mathbf{y}_1(t) + \cdots + c_n\mathbf{y}_n(t) = \mathbf{\Phi}(t)\mathbf{c}$$

   with $\mathbf{c} = (c_1, \dots, c_n)^{\mathsf{T}} \in \mathbb{C}^n$ for the general solution of $\mathbf{y}' = \mathbf{A}(t)\mathbf{y}$.

The Inhomogeneous Case

Theorem

1. *Every inhomogeneous linear system* $\mathbf{y}' = \mathbf{A}(t)\mathbf{y} + \mathbf{b}(t)$ *is solvable. A particular solution is* $\mathbf{y}_\mathrm{p} : I \to \mathbb{C}^n$ *defined by*

$$\mathbf{y}_\mathrm{p}(t) = \mathbf{\Phi}(t)\mathbf{c}(t) \quad \textit{with} \quad \mathbf{c}(t) = \int_{t_0}^{t} \mathbf{\Phi}(s)^{-1}\mathbf{b}(s)\,\mathrm{d}s,$$

   where $t_0 \in I$ *can be arbitrarily chosen.*

2. *The general solution of* $\mathbf{y}' = \mathbf{A}(t)\mathbf{y} + \mathbf{b}(t)$ *is obtained by adding to the particular solution* $\mathbf{y}_\mathrm{p}$ *from (1) the general solution of the associated homogeneous linear system* $\mathbf{y}' = \mathbf{A}(t)\mathbf{y}$*, i.e.,*

$$\mathbf{y}(t) = \mathbf{\Phi}(t)\mathbf{c}(t) + \mathbf{\Phi}(t)\mathbf{c}_0$$

   with $\mathbf{c}(t)$ *as in (1) and* $\mathbf{c}_0 \in \mathbb{C}^n$*.*

Note that $\mathbf{c}_0$ and $\mathbf{y}(t_0) = \mathbf{y}_0$ determine each other via $\mathbf{y}_0 = \mathbf{\Phi}(t_0)\mathbf{c}_0$, and that the general solution can also be obtained by using $\mathbf{c}(t) = \int \mathbf{\Phi}(t)^{-1}\mathbf{b}(t)\,\mathrm{d}t = \mathbf{c}_0 + \int_{t_0}^{t} \mathbf{\Phi}(s)^{-1}\mathbf{b}(s)\,\mathrm{d}s$ in (1).

Proof.

(1) is proved by a higher-dimensional analogue of "variation of parameters". Any fundamental matrix $\mathbf{\Phi}(t)$ satisfies $\mathbf{\Phi}' = \mathbf{A}\mathbf{\Phi}$ and hence

$$(\mathbf{\Phi c})' = \mathbf{\Phi}'\mathbf{c} + \mathbf{\Phi c}' = \mathbf{A\Phi c} + \mathbf{\Phi c}'$$

$\Longrightarrow (\mathbf{\Phi c})' = \mathbf{A\Phi c} + \mathbf{b}$ is equivalent to $\mathbf{\Phi c}' = \mathbf{b}$, i.e., to $\mathbf{c}' = \mathbf{\Phi}^{-1}\mathbf{b}$. Since $t \mapsto \mathbf{\Phi}(t)^{-1}\mathbf{b}(t)$ is continuous, the Fundamental Theorem of Calculus applies and $\mathbf{c}(t) = \int_{t_0}^{t} \mathbf{\Phi}(s)^{-1}\mathbf{b}(s)\,\mathrm{d}s$ solves $\mathbf{c}' = \mathbf{\Phi}^{-1}\mathbf{b}$.

(2) is proved as in the one-dimensional case: If $\mathbf{y}_1, \mathbf{y}_2$ are solutions of $\mathbf{y}' = \mathbf{A}(t)\mathbf{y} + \mathbf{b}(t)$ then $\mathbf{y}_1 - \mathbf{y}_2$ solves the associated homogeneous system $\mathbf{y}' = \mathbf{A}(t)\mathbf{y}$. □

Note

An important step in the proof is the observation that $t \mapsto \mathbf{\Phi}(t)^{-1}$ is continuous. Why is this true?

Reason: The entries of $\mathbf{\Phi}(t)^{-1}$ are obtained from the entries of $\mathbf{\Phi}(t)$ (which are continuous) by applying the four basic arithmetic operations. This follows from $\mathbf{\Phi}(t)^{-1} = \frac{1}{\det \mathbf{\Phi}(t)} \operatorname{Adj} \mathbf{\Phi}(t)$ (see Linear Algebra course), which expresses the entries of $\mathbf{\Phi}(t)^{-1}$ in terms of certain subdeterminants of $\mathbf{\Phi}(t)$.

Example

We consider the 1st-order system

$$y_1' = y_1 + t\,y_2 + 1,$$
$$y_2' = t\,y_1 + y_2.$$

This is a time-dependent inhomogeneous linear system, with standard form

$$\begin{pmatrix} y_1' \\ y_2' \end{pmatrix} = \begin{pmatrix} 1 & t \\ t & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

i.e., $\mathbf{A}(t) = \left(\begin{smallmatrix} 1 & t \\ t & 1 \end{smallmatrix}\right)$, $\mathbf{b}(t) = \mathbf{b} = \left(\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right)$.
The associated homogeneous system $y_1' = y_1 + t\,y_2$, $y_2' = t\,y_1 + y_2$ can be solved using the observation that $s = y_1 + y_2$, $d = y_1 - y_2$ satisfy the "decoupled" system

$$s' = y_1' + y_2' = y_1 + t\,y_2 + t\,y_1 + y_2 = (1+t)(y_1+y_2) = (1+t)s,$$
$$d' = y_1' - y_2' = y_1 + t\,y_2 - t\,y_1 - y_2 = (1-t)(y_1-y_2) = (1-t)d.$$

The solution is $s(t) = c_1\,\mathrm{e}^{t+t^2/2}$, $d(t) = c_2\,\mathrm{e}^{t-t^2/2}$ with $c_1, c_2 \in \mathbb{R}$, say.

Example (cont'd)

$$\begin{aligned}\Longrightarrow\quad y_1(t) &= \frac{s(t)+d(t)}{2} = \frac{1}{2}\left(c_1\,\mathrm{e}^{t+t^2/2} + c_2\,\mathrm{e}^{t-t^2/2}\right),\\ y_2(t) &= \frac{s(t)-d(t)}{2} = \frac{1}{2}\left(c_1\,\mathrm{e}^{t+t^2/2} - c_2\,\mathrm{e}^{t-t^2/2}\right).\end{aligned}$$

The corresponding matrix-vector form is

$$\mathbf{y}(t) = \begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix} = \frac{1}{2}\underbrace{\begin{pmatrix} \mathrm{e}^{t+t^2/2} & \mathrm{e}^{t-t^2/2} \\ \mathrm{e}^{t+t^2/2} & -\mathrm{e}^{t-t^2/2} \end{pmatrix}}_{\mathbf{\Phi}(t)} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}.$$

(The factor $1/2$ doesn't matter for the fundamental matrix.)
Thus every solution of $\mathbf{y}' = \mathbf{A}(t)\mathbf{y}$ is a linear combination of

$$\mathbf{y}_1(t) = \begin{pmatrix} \mathrm{e}^{t+t^2/2} \\ e^{t+t^2/2} \end{pmatrix} = e^{t+t^2/2}\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \mathbf{y}_2(t) = e^{t-t^2/2}\begin{pmatrix} 1 \\ -1 \end{pmatrix},$$

which therefore form a fundamental system of solutions.

Example (cont'd)

In order to solve the original inhomogeneous system, we compute

$$\mathbf{\Phi}(t)^{-1}\mathbf{b}(t) = \frac{1}{-2e^{2t}}\begin{pmatrix} -e^{t-t^2/2} & -e^{t-t^2/2} \\ -e^{t+t^2/2} & e^{t+t^2/2} \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} e^{-t-t^2/2} \\ e^{-t+t^2/2} \end{pmatrix},$$

$$\mathbf{c}(t) = \int_0^t \frac{1}{2}\begin{pmatrix} e^{-s-s^2/2} \\ e^{-s+s^2/2} \end{pmatrix} ds = \frac{1}{2}\begin{pmatrix} \int_0^t e^{-s-s^2/2}\,ds \\ \int_0^t e^{-s+s^2/2}\,ds \end{pmatrix}.$$

A particular solution of $\mathbf{y}' = \mathbf{A}(t)\mathbf{y} + \mathbf{b}$ is therefore

$$\mathbf{y}_p(t) = \mathbf{\Phi}(t)\mathbf{c}(t) = \frac{1}{2}\begin{pmatrix} e^{t+t^2/2} & e^{t-t^2/2} \\ e^{t+t^2/2} & -e^{t-t^2/2} \end{pmatrix}\begin{pmatrix} \int_0^t e^{-s-s^2/2}\,ds \\ \int_0^t e^{-s+s^2/2}\,ds \end{pmatrix}$$

$$= \frac{1}{2}\begin{pmatrix} e^{t+t^2/2}\int_0^t e^{-s-s^2/2}\,ds + e^{t-t^2/2}\int_0^t e^{-s+s^2/2}\,ds \\ e^{t+t^2/2}\int_0^t e^{-s-s^2/2}\,ds - e^{t-t^2/2}\int_0^t e^{-s+s^2/2}\,ds \end{pmatrix}.$$

Of course this is a toy example. You can check that $s(t)$ and $d(t)$, defined as in the homogeneous case, solve the decoupled system $s' = (1+t)s + 1$, $d' = (1-t)d + 1$, and that ordinary variation of parameters for these two ODE's and the backwards substitution $\mathbf{y}_p = \frac{1}{2}\begin{pmatrix} s_p + d_p \\ s_p - d_p \end{pmatrix}$ leads to the same result.

The Exponential of a Matrix

Providing a solution in the time-independent case $\mathbf{y}' = \mathbf{A}\mathbf{y}$

Definition

The *matrix exponential function* $\exp\colon \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$ is defined by

$$\exp \mathbf{A} := \mathrm{e}^{\mathbf{A}} := \sum_{k=0}^{\infty} \frac{1}{k!}\mathbf{A}^k = \mathbf{I}_n + \mathbf{A} + \frac{1}{2}\mathbf{A}^2 + \frac{1}{6}\mathbf{A}^3 + \cdots .$$

Since convergence of a sequence/series in $\mathbb{C}^{n\times n}$ is equivalent to entry-wise convergence, this limit is well-defined if the (i,j)-entries of the partial sums $\sum_{k=0}^{K} \frac{1}{k!}\mathbf{A}^k$, $K \in \mathbb{N}$, which are $\sum_{k=0}^{K} \frac{1}{k!}(\mathbf{A}^k)_{ij}$, converge in $\mathbb{C}$ for $K \to \infty$.

Let $a = \max\{|a_{ij}|\,; 1 \le i,j \le n\}$.

$\Longrightarrow$ The entries of $\mathbf{A}^k$ are bounded by $n^{k-1}a^k$.

$$\Longrightarrow \quad \sum_{k=0}^{K} \left|\frac{1}{k!}(\mathbf{A}^k)_{ij}\right| \le 1 + a + \frac{na^2}{2!} + \frac{n^2a^3}{3!} + \cdots + \frac{n^{K-1}a^K}{K!} \le \mathrm{e}^{na} < \infty$$

By the comparison test, the series formed by the (i,j)-entries of the partial sums converges (even absolutely!), and hence the limit defining $\mathrm{e}^{\mathbf{A}}$ is well defined.

Lemma

If $\mathbf{AB} = \mathbf{BA}$ *then* $\mathrm{e}^{\mathbf{A}+\mathbf{B}} = \mathrm{e}^{\mathbf{A}}\mathrm{e}^{\mathbf{B}}$.

Proof.

The preceding argument shows that the series defining $\mathrm{e}^{\mathbf{A}}$ converges absolutely. Hence we can freely rearrange the summands in the following double series:

$$\mathrm{e}^{\mathbf{A}}\mathrm{e}^{\mathbf{B}} = \left(\sum_{k=0}^{\infty} \frac{1}{k!}\mathbf{A}^k\right)\left(\sum_{l=0}^{\infty} \frac{1}{l!}\mathbf{B}^l\right) = \sum_{m=0}^{\infty} \sum_{\substack{k,l \\ k+l=m}} \frac{1}{k!l!}\mathbf{A}^k\mathbf{B}^l$$

$$= \sum_{m=0}^{\infty} \frac{1}{m!} \sum_{k=0}^{m} \binom{m}{k} \mathbf{A}^k\mathbf{B}^{m-k}$$

$$= \sum_{m=0}^{\infty} \frac{1}{m!}(\mathbf{A}+\mathbf{B})^m = \mathrm{e}^{\mathbf{A}+\mathbf{B}} \qquad \text{(Binomial Theorem)}$$

For the Binomial Theorem to hold we need the assumption $\mathbf{AB} = \mathbf{BA}$. As an example consider the case $m = 2$: $(\mathbf{A}+\mathbf{B})^2 = \mathbf{A}^2 + \mathbf{AB} + \mathbf{BA} + \mathbf{B}^2 = \mathbf{A}^2 + 2\mathbf{AB} + \mathbf{B}^2$ iff $\mathbf{AB} = \mathbf{BA}$. □

Note

Since $\pm\mathbf{A}$ commute with each other, the lemma gives in particular $e^{\mathbf{A}}e^{-\mathbf{A}} = e^{\mathbf{A}-\mathbf{A}} = e^{\mathbf{0}} = \mathbf{I}_n$, i.e., $e^{\mathbf{A}}$ is invertible with $(e^{\mathbf{A}})^{-1} = e^{-\mathbf{A}}$.

Example

$\mathbf{A} = \mathbf{E}_{11} = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$, $\mathbf{B} = \mathbf{E}_{12} = \left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right)$; $\mathbf{AB} = \left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right)$,
$\mathbf{BA} = \left(\begin{smallmatrix} 0 & 0 \\ 0 & 0 \end{smallmatrix}\right) \neq \mathbf{AB}$;

$e^{\mathbf{A}} = \left(\begin{smallmatrix} e & 0 \\ 0 & 1 \end{smallmatrix}\right)$, $e^{\mathbf{B}} = \mathbf{I}_2 + \mathbf{B} = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ (since $\mathbf{B}^2 = \mathbf{0}$),
$\mathbf{A} + \mathbf{B} = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 0 \end{smallmatrix}\right)$, $(\mathbf{A}+\mathbf{B})^2 = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 0 \end{smallmatrix}\right)\left(\begin{smallmatrix} 1 & 1 \\ 0 & 0 \end{smallmatrix}\right) = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 0 \end{smallmatrix}\right) = \mathbf{A} + \mathbf{B}$;

$$e^{\mathbf{A}}e^{\mathbf{B}} = \begin{pmatrix} e & e \\ 0 & 1 \end{pmatrix},$$

$$e^{\mathbf{B}}e^{\mathbf{A}} = \begin{pmatrix} e & 1 \\ 0 & 1 \end{pmatrix},$$

$$e^{\mathbf{A}+\mathbf{B}} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \sum_{k=1}^{\infty} \frac{1}{k!} \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} e-1 & e-1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} e & e-1 \\ 0 & 1 \end{pmatrix}.$$

Exercise

Show that for any diagonal matrix

$$\mathbf{D} = \begin{pmatrix} d_1 & & & \\ & d_2 & & \\ & & \ddots & \\ & & & d_n \end{pmatrix} \quad \text{we have} \quad \mathrm{e}^{\mathbf{D}} = \begin{pmatrix} \mathrm{e}^{d_1} & & & \\ & \mathrm{e}^{d_2} & & \\ & & \ddots & \\ & & & \mathrm{e}^{d_n} \end{pmatrix}.$$

Exercise

Which condition should a matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ satisfy in order to conclude that $\mathrm{e}^{\mathbf{A}}$ is

a) symmetric;

b) orthogonal.

Exercise

1. Does $\mathrm{e}^{\mathbf{A}} = \mathrm{e}^{\mathbf{B}}$ imply $\mathbf{A} = \mathbf{B}$?
2. Is every invertible matrix $\mathbf{B} \in \mathbb{R}^{n\times n}$ in the range of $\mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$, $\mathbf{A} \mapsto \mathrm{e}^{\mathbf{A}}$? What if $\mathbb{R}$ is replaced by $\mathbb{C}$ in this problem?

Now consider for a fixed $n \times n$ matrix $\mathbf{A}$ the matrix function

$$t \mapsto \mathrm{e}^{\mathbf{A}t} = \sum_{k=0}^{\infty} \frac{t^k}{k!}\,\mathbf{A}^k = \mathbf{I}_n + t\,\mathbf{A} + \frac{t^2}{2}\,\mathbf{A}^2 + \frac{t^3}{6}\,\mathbf{A}^3 + \cdots.$$

The (i,j)-entry of $\mathrm{e}^{\mathbf{A}t}$, viz. $\sum_{k=0}^{\infty} \frac{(\mathbf{A}^k)_{ij}}{k!}\,t^k$, is a power series, which converges for all $t \in \mathbb{R}$. Termwise differentiation yields

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathrm{e}^{\mathbf{A}t} = \frac{\mathrm{d}}{\mathrm{d}t}\left(\sum_{k=0}^{\infty} \frac{t^k}{k!}\,\mathbf{A}^k\right) = \sum_{k=1}^{\infty} \frac{k\,t^{k-1}}{k!}\,\mathbf{A}^k = \mathbf{A}\,\mathrm{e}^{\mathbf{A}t}.$$

Theorem

The columns of $\mathrm{e}^{\mathbf{A}t}$ form a fundamental system of solutions of $\mathbf{y}' = \mathbf{A}\mathbf{y}$, and the general solution of $\mathbf{y}' = \mathbf{A}\mathbf{y}$ is $\mathbf{y}(t) = \mathrm{e}^{\mathbf{A}t}\,\mathbf{y}(0)$.

Proof.

Since $\mathbf{\Phi}(t) := \mathrm{e}^{\mathbf{A}t}$ satisfies the matrix ODE $\mathbf{\Phi}'(t) = \mathbf{A}\mathbf{\Phi}(t)$, its columns solve $\mathbf{y}' = \mathbf{A}\mathbf{y}$.

$$\mathbf{\Phi}(0) = \mathbf{I}_n + 0\,\mathbf{A} + \frac{0^2}{2!}\,\mathbf{A}^2 + \cdots = \mathbf{I}_n$$

In particular $\mathbf{\Phi}(0)$ is invertible and the assertion follows. □

Example

Consider the system $y_1' = y_2$, $y_2' = -y_1$, which arises from the 2nd-order ODE $y'' + y = 0$ by setting $(y_1, y_2) = (y, y')$. We are interested in the solution with initial values $y_1(0) = 6$, $y_2(0) = 2$, (or $y(0) = 6$, $y'(0) = 2$ for the 2nd-order ODE).

Of course this solution is $y(t) = 6\cos t + 2\sin t$, which we can use to verify our computation.

The system has the form $\mathbf{y}' = \mathbf{A}\mathbf{y}$ with $\mathbf{A} = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$.
Since $\mathbf{A}^2 = \left(\begin{smallmatrix} -1 & 0 \\ 0 & -1 \end{smallmatrix}\right) = -\mathbf{I}_2$, $\mathbf{A}^3 = -\mathbf{A}$, $\mathbf{A}^4 = \mathbf{I}_2$, we get

$$\begin{aligned} e^{\mathbf{A}t} &= \begin{pmatrix} 1 & \\ & 1 \end{pmatrix} + \begin{pmatrix} & t \\ -t & \end{pmatrix} + \begin{pmatrix} -\frac{t^2}{2!} & \\ & -\frac{t^2}{2!} \end{pmatrix} + \begin{pmatrix} & -\frac{t^3}{3!} \\ +\frac{t^3}{3!} & \end{pmatrix} + \begin{pmatrix} \frac{t^4}{4!} & \\ & \frac{t^4}{4!} \end{pmatrix} + \cdots \\ &= \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix}. \end{aligned}$$

Hence the solution of our IVP is

$$\mathbf{y}(t) = e^{\mathbf{A}t}\mathbf{y}(0) = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix} \begin{pmatrix} 6 \\ 2 \end{pmatrix} = \begin{pmatrix} 6\cos t + 2\sin t \\ -6\sin t + 2\cos t \end{pmatrix},$$

in accordance with the known solution.

Example (cont'd)

Continuing with the example, we use the opportunity to illustrate the solution method for an inhomogeneous system. The new task is to determine the general solution of $y_1' = y_2$, $y_2' = -y_1 + t$, which arises from the 2nd-order equation $y'' + y = t$.

This inhomogeneous system has the form $\mathbf{y}' = \mathbf{A}\mathbf{y} + \mathbf{b}(t)$ with $\mathbf{A} = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$ and $\mathbf{b}(t) = \left(\begin{smallmatrix} 0 \\ t \end{smallmatrix}\right)$.

A particular solution is $\mathbf{y}_\mathrm{p}(t) = \mathrm{e}^{\mathbf{A}t}\mathbf{c}(t)$ with

$$\begin{aligned}
\mathbf{c}(t) &= \int_0^t \mathrm{e}^{-\mathbf{A}s}\mathbf{b}(s)\,\mathrm{d}s = \int_0^t \begin{pmatrix} \cos s & -\sin s \\ \sin s & \cos s \end{pmatrix} \begin{pmatrix} 0 \\ s \end{pmatrix} \mathrm{d}s \\
&= \int_0^t \begin{pmatrix} -s\sin s \\ s\cos s \end{pmatrix} \mathrm{d}s = \left[\begin{pmatrix} s\cos s - \sin s \\ s\sin s + \cos s \end{pmatrix} \right]_0^t \\
&= \begin{pmatrix} t\cos t - \sin t \\ t\sin t + \cos t - 1 \end{pmatrix},
\end{aligned}$$

so that

$$\mathbf{y}_\mathrm{p}(t) = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix} \begin{pmatrix} t\cos t - \sin t \\ t\sin t + \cos t - 1 \end{pmatrix} = \begin{pmatrix} t - \sin t \\ 1 - \cos t \end{pmatrix}.$$

Example (cont'd)

Since $t \mapsto \begin{pmatrix} \sin t \\ \cos t \end{pmatrix}$ is a solution of the associated homogeneous system, we may also take

$$\mathbf{y}_{\mathrm{p}}(t) = \begin{pmatrix} t \\ 1 \end{pmatrix}.$$

Well, this solution could have been guessed without going through the rather tedious computation!

The general solution of $y_1' = y_2$, $y_2' = -y_1 + t$ is therefore

$$\mathbf{y}(t) = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} + \begin{pmatrix} t \\ 1 \end{pmatrix} = \begin{pmatrix} c_1 \cos t + c_2 \sin t + t \\ -c_1 \sin t + c_2 \cos t + 1 \end{pmatrix}$$

with $c_1, c_2 \in \mathbb{C}$ (or $\mathbb{R}$), and that of $y'' + y = t$ the 1st coordinate function $y_1(t) = c_1 \cos t + c_2 \sin t + t$ (and $y_2(t) = y_1'(t)$, of course).

Note that solving the homogeneous system with the matrix exponential has produced (and for real systems always produces) a real fundamental system of solutions, whereas diagonalizing $\left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$ (cf. Linear Algebra part) gives the complex fundamental system

$$\mathbf{y}_1(t) = \mathrm{e}^{\mathrm{i}t} \begin{pmatrix} 1 \\ \mathrm{i} \end{pmatrix}, \quad \mathbf{y}_2(t) = \mathrm{e}^{-\mathrm{i}t} \begin{pmatrix} 1 \\ -\mathrm{i} \end{pmatrix}.$$

Concept Check

True or False?

Let $a, b, c, d\colon \mathbb{R} \to \mathbb{R}$ be continuous, $\mathbf{A}(t) = \begin{pmatrix} a(t) & b(t) \\ c(t) & d(t) \end{pmatrix}$, and $\mathbf{Y}_0 = \begin{pmatrix} 1 & 2 \\ 3 & 0 \end{pmatrix}$.

1. The matrix IVP $\mathbf{Y}' = \mathbf{A}(t)\mathbf{Y} \wedge \mathbf{Y}(0) = \mathbf{Y}_0$ has a unique solution $\mathbf{Y}(t) = \begin{pmatrix} y_{11}(t) & y_{12}(t) \\ y_{21}(t) & y_{22}(t) \end{pmatrix}$ that is defined for $t \in \mathbb{R}$.
   True. Setting $\mathbf{Y}(t) = \big(\mathbf{y}_1(t) | \mathbf{y}_2(t)\big)$, the matrix IVP is equivalent to the two vector IVP's $\mathbf{y}_1' = \mathbf{A}(t)\mathbf{y}_1 \wedge \mathbf{y}_1(0) = \begin{pmatrix} 1 \\ 3 \end{pmatrix}$, $\mathbf{y}_2' = \mathbf{A}(t)\mathbf{y}_2 \wedge \mathbf{y}_2(0) = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$, to which the EUT (sharpened version in the linear case) applies.
2. The columns of $\mathbf{Y}(t)$, cf. (1), form a fundamental system of solutions of $\mathbf{y}' = \mathbf{A}(t)\mathbf{y}$. *True,* since $\mathbf{Y}_0$ is invertible.
3. The solution of $\mathbf{Y}' = \mathbf{A}(t)\mathbf{Y} \wedge \mathbf{Y}(0) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ is $t \mapsto \mathrm{e}^{\mathbf{B}(t)}$ with $\mathbf{B}(t) = \int_0^t \mathbf{A}(s)\,\mathrm{d}s$. *False (in general),* but true if a, b, c, d are constant, in which case $\mathbf{B} = \int_0^t \mathbf{A}\,\mathrm{d}s = \mathbf{A}t$.
4. The statement in (1) remains true if $\mathbf{Y}' = \mathbf{A}(t)\mathbf{Y}$ is replaced by $\mathbf{Y}' = \mathbf{A}(t)\mathbf{Y} + \mathbf{B}(t)$ with $\mathbf{B}\colon \mathbb{R} \to \mathbb{R}^{2\times 2}$ continuous as well. *True,* since the EUT also applies to the inhomogeneous case.

Higher-Order Linear ODE's

The general time-dependent case

We consider only scalar ODE's, that is

$$y^{(n)} + a_{n-1}(t)y^{(n-1)} + \cdots + a_1(t)y' + a_0(t)y = b(t)$$

with continuous functions $a_0, a_1, \ldots, a_{n-1}, b\colon I \to \mathbb{C}$.

Order reduction $(y_1, y_2, \ldots, y_n) = (y, y', \ldots, y^{(n-1)})$ transforms such an ODE into the 1st-order system

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \\ y_n \end{pmatrix}' = \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & 0 & \ldots & 0 & 1 \\ -a_0(t) & -a_1(t) & \ldots & -a_{n-2}(t) & -a_{n-1}(t) \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \\ y_n \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ b(t) \end{pmatrix}.$$

The coefficient matrix $\mathbf{A}(t)$ is the transposed companion matrix (cf. Linear Algebra part) of the polynomial $X^n + a_{n-1}(t)X^{n-1} + \cdots + a_1(t)X + a_0(t) \in \mathbb{C}[X]$ (when $t \in I$ is considered as fixed).

The sharpened version of the Existence and Uniqueness Theorem for solutions of linear 1st-order ODE systems has the following

Corollary

1. *The solutions of any homogeneous nth-order ODE* $y^{(n)} + a_{n-1}(t)y^{(n-1)} + \cdots + a_1(t)y' + a_0(t)y = 0$ *exist on the whole interval I and form an n-dimensional subspace S of the function space* $\mathbb{C}^I$.

2. *Solutions* $y_1(t), \ldots, y_n(t)$ *form a basis of the solution space S iff for some (and hence all)* $t \in I$ *the matrix*

$$\mathbf{W}(t) = \begin{pmatrix} y_1(t) & y_2(t) & \ldots & y_n(t) \\ y_1'(t) & y_2'(t) & \ldots & y_n'(t) \\ \vdots & \vdots & & \vdots \\ y_1^{(n-1)}(t) & y_2^{(n-1)}(t) & \ldots & y_n^{(n-1)}(t) \end{pmatrix} \quad \textit{is invertible.}$$

3. *Any inhomogeneous nth-order ODE* $y^{(n)} + a_{n-1}(t)y^{(n-1)} + \cdots + a_1(t)y' + a_0(t)y = b(t)$ *is solvable. Solutions exist on the whole interval I, and they form a coset* $\{y_{\mathrm{p}}(t) + y_{\mathrm{h}}(t); y_{\mathrm{h}}(t) \in S\}$ *of S.*

Corollary (cont'd)

4. *Any IVP* $y^{(n)} + a_{n-1}(t)y^{(n-1)} + \cdots + a_1(t)y' + a_0(t)y = b(t) \wedge y^{(i)}(t_0) = c_i$ *for* $0 \le i \le n-1$ *has a unique solution, which is defined on the whole interval I.*

For real *n*th-order ODE's mutatis mutandis the same assertions hold (in particular such an ODE has a fundamental system consisting of real-valued solutions).

Proof of the corollary.

All assertions follow from the said theorem and the observation that solutions $\mathbf{y}(t) = \big(y_1(t), \ldots, y_n(t)\big)^{\mathsf{T}}$ of the reduced 1st-order system must satisfy $y_2(t) = y_1'(t)$, $y_3(t) = y_2'(t) = y_1''(t)$, $\ldots$, $y_n(t) = y_1^{(n-1)}(t)$ and hence $y_n'(t) = y_1^{(n)}(t)$, so that $y_1(t)$ is a solution of the *n*th-order ODE. The map $\mathbf{y}(t) \mapsto y_1(t)$ ("strip off all components of $\mathbf{y}(t)$ except the first") is then a vector space isomorphism from the solution space of the reduced 1st-order system onto S. □

Note

In the statement of the corollary and its proof $y_1(t), \ldots, y_n(t)$ have a different meaning (solutions of the *n*-th order scalar ODE versus coordinate functions of a solution of the 1st-order ODE system).

Example

Let us consider the ODE

$$y''' - y'' - 2y' = \frac{1}{1-t}, \quad t \in (-\infty, 1).$$

For the associated homogeneous ODE $y''' - y'' - 2y' = 0$ we had determined a fundamental system of solutions earlier, viz. $y_1(t) = 1$, $y_2(t) = \mathrm{e}^{-t}$, $y_3(t) = \mathrm{e}^{2t}$.

The corresponding 1st-order system is $y_1' = y_2$, $y_2' = y_3$, $y_3' = y_1''' = y_1'' + 2y_1' = y_3 + 2y_2$ or, in matrix form,

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}' = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 2 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}.$$

Reading the proof of the corollary backwards, we see that

$$\mathbf{W}(t) = \begin{pmatrix} y_1(t) & y_2(t) & y_3(t) \\ y_1'(t) & y_2'(t) & y_3'(t) \\ y_1''(t) & y_2''(t) & y_3''(t) \end{pmatrix} = \begin{pmatrix} 1 & \mathrm{e}^{-t} & \mathrm{e}^{2t} \\ 0 & -\mathrm{e}^{-t} & 2\mathrm{e}^{2t} \\ 0 & \mathrm{e}^{-t} & 4\mathrm{e}^{2t} \end{pmatrix}$$

is a fundamental matrix of this system.

Example (cont'd)

Variation of parameters requires to determine $\mathbf{W}(t)^{-1}$. This is done using Gaussian elimination as usual:

$$\left(\begin{array}{ccc|ccc} 1 & \mathrm{e}^{-t} & \mathrm{e}^{2t} & 1 & 0 & 0 \\ 0 & -\mathrm{e}^{-t} & 2\mathrm{e}^{2t} & 0 & 1 & 0 \\ 0 & \mathrm{e}^{-t} & 4\mathrm{e}^{2t} & 0 & 0 & 1 \end{array}\right) \to \left(\begin{array}{ccc|ccc} 1 & 0 & 3\mathrm{e}^{2t} & 1 & 1 & 0 \\ 0 & -\mathrm{e}^{-t} & 2\mathrm{e}^{2t} & 0 & 1 & 0 \\ 0 & 0 & 6\mathrm{e}^{2t} & 0 & 1 & 1 \end{array}\right)$$

$$\to \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & \frac{1}{2} & -\frac{1}{2} \\ 0 & -\mathrm{e}^{-t} & 0 & 0 & \frac{2}{3} & -\frac{1}{3} \\ 0 & 0 & 6\mathrm{e}^{2t} & 0 & 1 & 1 \end{array}\right) \to \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & \frac{1}{2} & -\frac{1}{2} \\ 0 & 1 & 0 & 0 & -\frac{2}{3}\mathrm{e}^{t} & \frac{1}{3}\mathrm{e}^{t} \\ 0 & 0 & 1 & 0 & \frac{1}{6}\mathrm{e}^{-2t} & \frac{1}{6}\mathrm{e}^{-2t} \end{array}\right)$$

$$\Longrightarrow \mathbf{W}(t)^{-1}\mathbf{b}(t) = \begin{pmatrix} 1 & \frac{1}{2} & -\frac{1}{2} \\ 0 & -\frac{2}{3}\mathrm{e}^{t} & \frac{1}{3}\mathrm{e}^{t} \\ 0 & \frac{1}{6}\mathrm{e}^{-2t} & \frac{1}{6}\mathrm{e}^{-2t} \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ \frac{1}{1-t} \end{pmatrix} = \begin{pmatrix} -\frac{1}{2(1-t)} \\ \frac{\mathrm{e}^{t}}{3(1-t)} \\ \frac{\mathrm{e}^{-2t}}{6(1-t)} \end{pmatrix}$$

$$\Longrightarrow \mathbf{c}(t) = \int_0^t \begin{pmatrix} -\frac{1}{2(1-s)} \\ \frac{\mathrm{e}^{s}}{3(1-s)} \\ \frac{\mathrm{e}^{-2s}}{6(1-s)} \end{pmatrix} \mathrm{d}s = \begin{pmatrix} -\int_0^t \frac{1}{2(1-s)}\,\mathrm{d}s \\ \int_0^t \frac{\mathrm{e}^{s}}{3(1-s)}\,\mathrm{d}s \\ \int_0^t \frac{\mathrm{e}^{-2s}}{6(1-s)}\,\mathrm{d}s \end{pmatrix}$$

Example (cont'd)

$\Longrightarrow$ One particular solution is

$$y_p(t) = c_1(t)y_1(t) + c_2(t)y_2(t) + c_3(t)y_3(t)$$
$$= -\int_0^t \frac{1}{2(1-s)}\,\mathrm{d}s + \left(\int_0^t \frac{\mathrm{e}^s}{3(1-s)}\,\mathrm{d}s\right)\mathrm{e}^{-t} + \left(\int_0^t \frac{\mathrm{e}^{-2s}}{6(1-s)}\,\mathrm{d}s\right)\mathrm{e}^{2t}.$$

$\Longrightarrow$ The general solution is

$$y(t) = y_p(t) + \gamma_1 y_1(t) + \gamma_2 y_2(t) + \gamma_3 y_3(t)$$
$$= \big(c_1(t) + \gamma_1\big)y_1(t) + \big(c_2(t) + \gamma_2\big)y_2(t) + \big(c_3(t) + \gamma_3\big)y_3(t)$$

with constants $\gamma_1, \gamma_2, \gamma_3$.

Now suppose we want to solve the IVP

$$y''' - y'' - 2y' = \frac{1}{1-t}, \quad y(0) = 1,\ y'(0) = y''(0) = 0, \quad \text{say.}$$

It is possible to do this from the general solution by determining the constants γ_i from the given initial conditions.

However, there is a more conceptional approach using the matrix exponential function $\mathrm{e}^{\mathbf{A}t}$.

Example (cont'd)

In terms of the initial conditions $\mathbf{y}(0) = \big(y(0), y'(0), y''(0)\big) = \mathbf{y}_0$, the general solution of the associated 1st-order system is also given by

$$\mathbf{y}(t) = \mathrm{e}^{\mathbf{A}t}(\mathbf{c}(t) + \mathbf{y}_0),$$

where $\mathbf{y}_p(t) = \mathrm{e}^{\mathbf{A}t}\mathbf{c}(t)$ is the particular solution satisfying $\mathbf{c}(0) = \mathbf{0}$. For this note that any solution $\mathbf{y} = (y_1, y_2, y_3)^{\mathsf{T}}$ of the inhomogeneous system

$$\mathbf{y}' = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 2 & 1 \end{pmatrix} \mathbf{y} + \begin{pmatrix} 0 \\ 0 \\ \frac{1}{1-t} \end{pmatrix}$$

still satisfies $y_2 = y_1'$, $y_3 = y_2' = y_1''$.

It is not necessary to compute the matrix exponential $\mathrm{e}^{\mathbf{A}t}$ directly from the series representation. Instead we can use that $\mathbf{\Phi}(t) = \mathrm{e}^{\mathbf{A}t}$ is the unique solution of the matrix IVP $\mathbf{\Phi}'(t) = \mathbf{A}\mathbf{\Phi}(t) \wedge \mathbf{\Phi}(0) = \mathbf{I}_3$.

Claim: $\mathrm{e}^{\mathbf{A}t} = \mathbf{W}(t)\mathbf{W}(0)^{-1}$, where $\mathbf{W}(t)$ denotes the fundamental matrix determined earlier. (In fact, $\mathbf{W}(t)$ can also be any other fundamental matrix of the given system.)

Example (cont'd)

Proof of the claim: An arbitrary fundamental matrix $\mathbf{\Phi}(t) = (\mathbf{y}_1(t)|\mathbf{y}_2(t)|\mathbf{y}_3(t))$ has the form $\mathbf{\Phi}(t) = \mathbf{W}(t)\mathbf{S}$ for some invertible matrix $\mathbf{S}$, since its columns must be linear combinations of $\mathbf{y}_1(t)$, $\mathbf{y}_2(t)$, $\mathbf{y}_3(t)$, and vice versa.
For $\mathbf{\Phi}(t) = \mathrm{e}^{\mathbf{A}t}$ we have $\mathbf{\Phi}(0) = \mathbf{I}_3$ and hence $\mathbf{S} = \mathbf{W}(0)^{-1}$.

$$\Longrightarrow\ \mathrm{e}^{\mathbf{A}t} = \mathbf{W}(t)\mathbf{W}(0)^{-1} = \begin{pmatrix} 1 & \mathrm{e}^{-t} & \mathrm{e}^{2t} \\ 0 & -\mathrm{e}^{-t} & 2\mathrm{e}^{2t} \\ 0 & \mathrm{e}^{-t} & 4\mathrm{e}^{2t} \end{pmatrix} \begin{pmatrix} 1 & 1/2 & -1/2 \\ 0 & -2/3 & 1/3 \\ 0 & 1/6 & 1/6 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & \frac{1}{2} - \frac{2}{3}\mathrm{e}^{-t} + \frac{1}{6}\mathrm{e}^{2t} & -\frac{1}{2} + \frac{1}{3}\mathrm{e}^{-t} + \frac{1}{6}\mathrm{e}^{2t} \\ 0 & \frac{2}{3}\mathrm{e}^{-t} + \frac{1}{3}\mathrm{e}^{2t} & -\frac{1}{3}\mathrm{e}^{-t} + \frac{1}{3}\mathrm{e}^{2t} \\ 0 & -\frac{2}{3}\mathrm{e}^{-t} + \frac{2}{3}\mathrm{e}^{2t} & \frac{1}{3}\mathrm{e}^{-t} + \frac{2}{3}\mathrm{e}^{2t} \end{pmatrix}$$

$$\Longrightarrow\ \mathbf{c}(t) = \int_0^t \mathrm{e}^{-\mathbf{A}s}\mathbf{b}(s)\,\mathrm{d}s = \begin{pmatrix} \int_0^t \frac{-\frac{1}{2} + \frac{1}{3}\mathrm{e}^{s} + \frac{1}{6}\mathrm{e}^{-2s}}{1-s}\,\mathrm{d}s \\ \int_0^t \frac{-\frac{1}{3}\mathrm{e}^{s} + \frac{1}{3}\mathrm{e}^{-2s}}{1-s}\,\mathrm{d}s \\ \int_0^t \frac{\frac{1}{3}\mathrm{e}^{s} + \frac{2}{3}\mathrm{e}^{-2s}}{1-s}\,\mathrm{d}s \end{pmatrix},$$

from which the general solution of any IVP is given as

Example (cont'd)

$$y(t) = 1\left(y(0) + \int_0^t \frac{-\frac{1}{2} + \frac{1}{3}e^s + \frac{1}{6}e^{-2s}}{1-s}\,ds\right)$$
$$+ \left(\frac{1}{2} - \frac{2}{3}e^{-t} + \frac{1}{6}e^{2t}\right)\left(y'(0) + \int_0^t \frac{-\frac{1}{3}e^s + \frac{1}{3}e^{-2s}}{1-s}\,ds\right)$$
$$+ \left(-\frac{1}{2} + \frac{1}{3}e^{-t} + \frac{1}{6}e^{2t}\right)\left(y''(0) + \int_0^t \frac{\frac{1}{3}e^s + \frac{2}{3}e^{-2s}}{1-s}\,ds\right).$$

Example

We determine all solutions of the 2nd-order ODE

$$y'' - \frac{1}{2t}\,y' + \frac{1}{2t^2}\,y = 0 \qquad \text{on } I = (0, +\infty).$$

Since the coefficients are polynomials in t, it is reasonable to guess that there are solutions of the form $y(t) = t^k$. Substituting this into the ODE gives

$$k(k-1)t^{k-2} - \frac{kt^{k-1}}{2t} + \frac{t^k}{2t^2} = \left(k^2 - \frac{3}{2}k + \frac{1}{2}\right) t^{k-2} = 0.$$

The solutions of the quadratic are $k = 1$ and $k = \frac{1}{2}$, giving the solutions

$$y_1(t) = t \quad \text{and} \quad y_2(t) = \sqrt{t}.$$

From the theory we know that the solution space is 2-dimensional. Hence $y_1(t)$, $y_2(t)$ form a basis (fundamental system of solutions) iff they are linearly independent.

$$\begin{vmatrix} y_1(t) & y_2(t) \\ y_1'(t) & y_2'(t) \end{vmatrix} = \begin{vmatrix} t & \sqrt{t} \\ 1 & \frac{1}{2\sqrt{t}} \end{vmatrix} = \frac{1}{2}\sqrt{t} - \sqrt{t} = -\frac{1}{2}\sqrt{t} \neq 0$$

Example (cont'd)

$\Longrightarrow$ $y_1(t)$, $y_2(t)$ form a basis and the general solution is

$$y(t) = c_1 t + c_2\sqrt{t}, \quad c_1, c_2 \in \mathbb{C}.$$

The general real solution is then of course $y(t) = c_1 t + c_2\sqrt{t}$, $c_1, c_2 \in \mathbb{R}$.

Note

Linear independence of t and $\sqrt{t}$ is actually trivial to check—the functions are not scalar multiples of each other—, but for higher-order linear ODE's you will learn to appreciate the test using $\mathrm{W}(t)$, which needs to be evaluated only for one particular number t_0.

The Wronskian

In the preceding example, the determinant $\begin{vmatrix} y_1(t) & y_2(t) \\ y_1'(t) & y_2'(t) \end{vmatrix} = -\frac{1}{2}\sqrt{t}$ is called the Wronskian of $y_1(t)$, $y_2(t)$. More generally we define

Definition

1. Suppose $\mathbf{y}_1(t), \ldots, \mathbf{y}_n(t)$ are solutions of the 1st-order linear ODE system $\mathbf{y}' = \mathbf{A}(t)\mathbf{y}$ with $\mathbf{A}\colon I \to \mathbb{R}^{n\times n}$. The function

$$\mathrm{W}(t) = \det\big(\mathbf{y}_1(t)|\ldots|\mathbf{y}_n(t)\big)$$

   is called the *Wronskian* (*Wronski determinant*) of $\mathbf{y}_1(t), \ldots, \mathbf{y}_n(t)$.

2. Suppose $y_1(t), \ldots, y_n(t)$ are solutions of the n-th order scalar ODE $y^{(n)} + a_{n-1}(t)y^{(n-1)} + \cdots + a_0(t)y = 0$. The function

$$\mathrm{W}(t) = \det \begin{pmatrix} y_1(t) & \ldots & y_n(t) \\ y_1'(t) & \ldots & y_n'(t) \\ \vdots & & \vdots \\ y_1^{(n-1)}(t) & \ldots & y_n^{(n-1)}(t) \end{pmatrix}$$

   is called the *Wronskian* of $y_1(t), \ldots, y_n(t)$.

Note

Earlier we had denoted the matrix in Part 2 of the definition by $\mathbf{W}(t)$, anticipating its name *Wronski matrix*. The matrix appearing in Part 1 is also called a *Wronski matrix* (though less frequently). Note that "Wronski matrix" refers to the matrix formed from any set of n solutions, while "fundamental matrix" requires the solutions to be linearly independent.

Theorem (Abel's Theorem)

$\mathrm{W}(t)$ satisfies a homogeneous 1st-order linear ODE $\mathrm{W}'(t) = a(t)\mathrm{W}(t)$. The function $a(t)$ is the sum of the main diagonal entries of $\mathbf{A}(t)$ in Case (1) and equal to $-a_{n-1}(t)$ in Case (2).

Corollary

There exists a constant $c \in \mathbb{C}$ such that $\mathrm{W}(t) = c\,\mathrm{e}^{\int_{t_0}^{t} a(s)\,\mathrm{d}s}$ for $t \in I$.
In particular $\mathrm{W}(t) \equiv 0$ iff $c = 0$ iff $\mathrm{W}(t_0) = 0$.

Note

This explains in a different way the criterion for solutions $\mathbf{y}_1(t), \ldots, \mathbf{y}_n(t)$ to form a basis of the solution space of $\mathbf{y}' = \mathbf{A}(t)\mathbf{y}$ established earlier.

Proof of Abel's Theorem.

We give the proof only in the case $n = 2$. The proof of the general case requires properties of the determinant we haven't developed yet. Moreover, it suffices to prove Case (1), because Case (2) then follows by inspecting the form of $\mathbf{A}(t)$ in the order reduction formula.

Writing $(\mathbf{y}_1(t)|\mathbf{y}_2(t)) = \mathbf{\Phi}(t) = \begin{pmatrix} \phi_{11}(t) & \phi_{12}(t) \\ \phi_{21}(t) & \phi_{22}(t) \end{pmatrix}$ and using $\mathbf{\Phi}' = \mathbf{A}\mathbf{\Phi} = \begin{pmatrix} a_{11}\phi_{11}+a_{12}\phi_{21} & a_{11}\phi_{12}+a_{12}\phi_{22} \\ a_{21}\phi_{11}+a_{22}\phi_{21} & a_{21}\phi_{12}+a_{22}\phi_{22} \end{pmatrix}$, we have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\,\mathrm{W}(t) = \frac{\mathrm{d}}{\mathrm{d}t}\begin{vmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{vmatrix} &= (\phi_{11}\phi_{22} - \phi_{21}\phi_{12})' \\
&= \phi_{11}'\phi_{22} + \phi_{11}\phi_{22}' - \phi_{21}'\phi_{12} - \phi_{21}\phi_{12}' \\
&= (a_{11}\phi_{11} + a_{12}\phi_{21})\phi_{22} + \phi_{11}(a_{21}\phi_{12} + a_{22}\phi_{22}) \\
&\quad - (a_{21}\phi_{11} + a_{22}\phi_{21})\phi_{12} - \phi_{21}(a_{11}\phi_{12} + a_{12}\phi_{22}) \\
&= (a_{11} + a_{22})(\phi_{11}\phi_{22} - \phi_{21}\phi_{12}) \\
&= (a_{11} + a_{22})\mathrm{W}(t).
\end{aligned}$$

Example (cont'd)

We continue our previous example

$$y'' - \frac{1}{2t}\,y' + \frac{1}{2t^2}\,y = 0 \qquad t \in (0, +\infty).$$

Order reduction reduces this 2nd-order ODE to the 2×2-system

$$\begin{pmatrix} y \\ y' \end{pmatrix}' = \begin{pmatrix} 0 & 1 \\ -\frac{1}{2t^2} & \frac{1}{2t} \end{pmatrix} \begin{pmatrix} y \\ y' \end{pmatrix}$$

$\Longrightarrow \mathrm{W}'(t) = \frac{1}{2t}\,\mathrm{W}(t)$ by Abel's Theorem.
The solution of this ODE is

$$\mathrm{W}(t) = c \exp\left(\int \frac{\mathrm{d}t}{2t}\right) = c\,|t|^{1/2} = c\sqrt{t},$$

since $t > 0$.
For the fundamental system $y_1(t) = t$, $y_2(t) = \sqrt{t}$ the constant is $c = \mathrm{W}(1) = \begin{vmatrix} 1 & 1 \\ 1 & \frac{1}{2} \end{vmatrix} = -\frac{1}{2}$, as determined earlier.

Example (cont'd)

Now we consider the inhomogeneous ODE

$$y'' - \frac{1}{2t}\,y' + \frac{1}{2t^2}\,y = b(t), \quad t \in (0, +\infty).$$

For the case $b(t) = t^\ell$ our previous computation suggests a solution. In terms of the linear differential operator $L = \mathrm{D}^2 - \frac{1}{2t}\mathrm{D} + \frac{1}{2t^2}\,\mathrm{id}$ the ODE can be concisely written as $Ly = b(t)$, and we had found earlier that

$$L[t^k] = \left(k^2 - \frac{3}{2}\,k + \frac{1}{2}\right) t^{k-2}.$$

Since L is linear, it follows that $L\left[\frac{1}{k^2 - \frac{3}{2}\,k + \frac{1}{2}}\,t^k\right] = t^{k-2}$.

Substituting $\ell = k - 2$ gives

$k^2 - \frac{3}{2}\,k + \frac{1}{2} = (\ell+2)^2 - \frac{3}{2}(\ell+2) + \frac{1}{2} = \ell^2 + \frac{5}{2}\ell + \frac{3}{2} = (\ell+1)(\ell+\frac{3}{2})$

and hence

$$L\left[\frac{1}{(\ell+1)(\ell+\frac{3}{2})}\,t^{\ell+2}\right] = t^\ell, \qquad \text{valid for } \ell \notin \{-1, -\tfrac{3}{2}\}.$$

Example (cont'd)

Solutions for $\ell \in \{-1, -\frac{3}{2}\}$ can be determined using variation of parameters. We consider $b(t) = t^{-1}$ as an example.

It suffices to extract the first coordinate function of the vectorial solution (cf. the Theorem on Slide 13 of the handout version) of the corresponding 2×2 system:

$$y_p(t) = c_1(t)y_1(t) + c_2(t)y_2(t) \qquad \text{with}$$

$$\begin{pmatrix} c_1(t) \\ c_2(t) \end{pmatrix} = \int \begin{pmatrix} y_1(t) & y_2(t) \\ y_1'(t) & y_2'(t) \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ b(t) \end{pmatrix} \mathrm{d}t = \int \frac{1}{\mathrm{W}(t)} \begin{pmatrix} -y_2(t)b(t) \\ y_1(t)b(t) \end{pmatrix} \mathrm{d}t$$
$$= \int \frac{1}{-\frac{1}{2}\sqrt{t}} \begin{pmatrix} -\sqrt{t}\, t^{-1} \\ t\, t^{-1} \end{pmatrix} \mathrm{d}t = \int \begin{pmatrix} 2t^{-1} \\ -2t^{-1/2} \end{pmatrix} \mathrm{d}t = \begin{pmatrix} 2\ln t \\ -4\sqrt{t} \end{pmatrix}.$$

$\Longrightarrow y_p(t) = 2t \ln t - 4t$ is a particular solution, and (since $t \mapsto 4t$ solves the homogeneous ODE) $t \mapsto 2t \ln t$ as well.

The same formula works for any continuous right-hand side $b(t)$, except that it may not be integrable in closed form; cp. also [BDM17], Theorem 3.6.1.

Caveats when working with time-dependent differential operators

Polynomial differential operators with coefficients depending on t, like $L = \mathrm{D}^2 - \frac{1}{2t}\mathrm{D} + \frac{1}{2t^2}\,\mathrm{id}$ (or $L = \mathrm{D}^2 - \frac{1}{2t}\mathrm{D} + \frac{1}{2t^2}$, for short) do not satisfy the usual algebraic rules for working with polynomials! In particular we must not transpose time-dependent coefficients from right to left.

Example

$L = \mathrm{D}^2 - \frac{1}{2t}\mathrm{D} + \frac{1}{2t^2} \neq \mathrm{D}^2 - \mathrm{D}\frac{1}{2t} + \frac{1}{2t^2} = L^*$.

You can work out Lf and L^*f for suitable functions f and see that they differ.

But the following easier example also tells you what's going on:

$$(t\mathrm{D})[f] = t(\mathrm{D}f) = tf',$$
$$(\mathrm{D}t)[f] = \mathrm{D}(tf) = f + tf'.$$

Thus $\mathrm{D}t - t\mathrm{D} = \mathrm{id}$ rather than $\mathrm{D}t - t\mathrm{D} = 0$.

Second-Order Linear ODE's

Additional Remarks

First we state three famous (time-dependent) 2nd-order linear ODE's. Actually these are one-parameter families of ODE's providing one ODE for every $n \in \mathbb{N} = \{0, 1, 2, \dots\}$.

1. LEGENDRE's Differential Equation is

$$(1 - t^2)y'' - 2ty' + n(n+1)y = 0 \qquad (\mathrm{Le}_n)$$

with time domain $-1 < t < 1$.

2. HERMITE's Differential Equation is

$$y'' - 2ty' + 2ny = 0 \qquad (\mathrm{He}_n)$$

with time domain $t \in \mathbb{R}$.

3. LAGUERRE's Differential Equation is

$$ty'' + (1 - t)y' + ny = 0 \qquad (\mathrm{La}_n)$$

with time domain $t > 0$.

Theorem

The following polynomials solve these ODE's:

1. *(Le_n) is solved by the Legendre Polynomial $\mathrm{P}_n(X)$ of degree n, which is defined by*

$$\mathrm{P}_n(t) = \frac{1}{2^n n!}\left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^n ((t^2-1)^n).$$

2. *(He_n) is solved by the Hermite Polynomial $\mathrm{H}_n(X)$ of degree n, which is defined by*

$$\mathrm{H}_n(t) = (-1)^n \mathrm{e}^{t^2}\left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^n \mathrm{e}^{-t^2}.$$

3. *(La_n) is solved by the Laguerre Polynomial $\mathrm{L}_n(X)$ of degree n, which is defined by*

$$\mathrm{L}_n(t) = \mathrm{e}^t\left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^n (t^n \mathrm{e}^{-t}).$$

Notes

- Since $|\mathbb{R}| = \infty$, polynomials $a(X) = \sum_{i=0}^{n} a_i X^i \in \mathbb{R}[X]$ and polynomial functions $\mathbb{R} \to \mathbb{R}$, $t \mapsto a(t) = \sum_{i=0}^{n} a_i t^i$ determine each other uniquely, validating the preceding definitions.

  This follows from the *degree bound* for polynomials with coefficients in a field F, which implies that polynomial functions $t \mapsto a(t)$ and $t \mapsto b(t)$ arising from distinct polynomials $a(X), b(X) \in F[X]$ can have at most $\deg\big(a(X) - b(X)\big)$ equal values.

  For a finite field F the corresponding proposition is no longer true. For example, the polynomials $0 \in \mathbb{F}_2[X]$ and $X^2 + X \in \mathbb{F}_2[X]$ both determine the all-zero function $\mathbb{F}_2 \to \mathbb{F}_2$, as follows from the identities $0^2 + 0 = 1^2 + 1 = 0$ in $\mathbb{F}_2$. However, the proposition remains true under the additional assumption that $a(X)$ and $b(X)$ have degree less than $|F|$ (again by the degree bound).

- The normalization factors, $\frac{1}{2^n n!}$ for $\mathrm{P}_n(X)$ and $(-1)^n$ for $\mathrm{H}_n(X)$, $\mathrm{L}_n(X)$ do not matter for the solution of the ODE (since it is linear). We could as well have assumed that all three families consist of monic polynomials, obtained by dividing the non-normalized polynomials by their leading coefficients.

Proof.

We prove the assertion only for the Legendre polynomials (with different normalization factors). Writing as usual $\mathrm{D} = \frac{\mathrm{d}}{\mathrm{d}t}$, we evaluate $\mathrm{D}^{n+1}\left[(t^2-1)\mathrm{D}\big((t^2-1)^n\big)\right]$ in two different ways.

Using Leibniz's Formula $\mathrm{D}^n(fg) = \sum_{i=0}^{n}\binom{n}{i}(\mathrm{D}^i f)(\mathrm{D}^{n-i}g)$ for the n-th derivative of a product with $f = t^2-1$, $g = \mathrm{D}\big((t^2-1)^n\big)$, we have

$$\begin{aligned}
&\mathrm{D}^{n+1}\left[(t^2-1)\mathrm{D}\big((t^2-1)^n\big)\right] = \\
&\quad = (t^2-1)\mathrm{D}^{n+2}\big((t^2-1)^n\big) + (n+1)(2t)\mathrm{D}^{n+1}\big((t^2-1)^n\big) + n(n+1)\mathrm{D}^n\big((t^2-1)^n\big) \\
&\quad = (t^2-1)\mathrm{P}_n''(t) + 2(n+1)t\,\mathrm{P}_n'(t) + n(n+1)\mathrm{P}_n(t).
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\mathrm{D}^{n+1}\left[(t^2-1)\mathrm{D}\big((t^2-1)^n\big)\right] = \\
&\quad = \mathrm{D}^{n+1}\left[(t^2-1)2nt(t^2-1)^{n-1}\right] \\
&\quad = 2n\,\mathrm{D}^{n+1}\left[t(t^2-1)^n\right] \\
&\quad = 2n\left[t\,\mathrm{D}^{n+1}\big((t^2-1)^n\big) + (n+1)\mathrm{D}^n\big((t^2-1)^n\big)\right] \\
&\quad = 2nt\,\mathrm{P}_n'(t) + 2n(n+1)\mathrm{P}_n(t).
\end{aligned}$$

$$\Longrightarrow (1-t^2)\mathrm{P}_n''(t) - 2t\,\mathrm{P}_n'(t) + n(n+1)\mathrm{P}_n(t) = 0$$

□

Order Reduction

A different way

Theorem

Suppose $I \subseteq \mathbb{R}$ is an interval and $a, b\colon I \to \mathbb{C}$ are continuous. Further, suppose that $\phi\colon I \to \mathbb{C}$ is a nonzero solution of

$$y'' + a(t)y' + b(t)y = 0, \tag{$\star$}$$

and $J \subseteq I$ is a subinterval (of length > 0) such that $\phi(t) \neq 0$ for all $t \in J$. Then a second fundamental solution of ($\star$) on J (i.e., linearly independent of the restriction $\phi|_J$), is obtained as $\psi(t) = \phi(t)u(t)$, where $u(t)$ is any non-constant solution of

$$u'' + \left(2\frac{\phi'(t)}{\phi(t)} + a(t)\right)u' = 0. \tag{R}$$

Note

(R) is a 1st-order linear ODE for u', solved as usual by

$$u'(t) = \exp\left(-\int_{t_0}^{t} 2\frac{\phi'(s)}{\phi(s)} + a(s)\,\mathrm{d}s\right) = \frac{1}{\phi(t)^2}\exp\left(-\int_{t_0}^{t} a(s)\,\mathrm{d}s\right).$$

A further integration then yields $u(t)$.

Proof of the theorem.

We have

$$\begin{aligned}\psi &= \phi u,\\ \psi' &= \phi' u + \phi u',\\ \psi'' &= \phi'' u + 2\phi' u' + \phi u''.\end{aligned}$$

$$\begin{aligned}\Longrightarrow\ \psi'' + a\psi' + b\psi &= (\phi'' + a\phi' + b\phi)u + (2\phi' + a\phi)u' + \phi u''\\ &= (2\phi' + a\phi)u' + \phi u'',\end{aligned}$$

since ϕ solves $y'' + ay' + by = 0$.

Hence we have

$$\psi'' + a\psi' + b\psi = 0 \iff u'' + (2\phi'/\phi + a)u' = 0.$$ □

Example

We compute a fundamental system of solutions of Legendre's ODE for $n = 1$, which has the explicit form

$$y'' - \frac{2t}{1-t^2}\,y' + \frac{2}{1-t^2}\,y = 0, \quad -1 < t < 1. \qquad (\mathrm{Le}_1)$$

As we have seen, one solution is $\mathrm{P}_1(t) = \frac{1}{2}\mathrm{D}(t^2-1) = t$. Hence, by the theorem, a second linearly independent (of the first) solution on $J = (0,1)$ is $\psi(t) = t\,u(t)$, where $u'(t)$ solves

$$u''(t) + \left(2\,\frac{\mathrm{P}_1'(t)}{\mathrm{P}_1(t)} - \frac{2t}{1-t^2}\right)u'(t) = u''(t) + \left(\frac{2}{t} - \frac{2t}{1-t^2}\right)u'(t) = 0.$$

A nonzero solution is

$$u'(t) = \exp\left(\int -\frac{2}{t} + \frac{2t}{1-t^2}\,\mathrm{d}t\right) = \exp\bigl(-2\ln t - \ln(1-t^2)\bigr)$$
$$= \frac{1}{t^2(1-t^2)}$$

Example (cont'd)

$$\Longrightarrow\ u(t) = \int \frac{\mathrm{d}t}{t^2(1-t^2)} = \int \frac{1}{t^2} + \frac{\frac{1}{2}}{1+t} + \frac{\frac{1}{2}}{1-t}\,\mathrm{d}t = -\frac{1}{t} + \frac{1}{2}\,\ln\frac{1+t}{1-t}$$

$$\Longrightarrow\ \psi(t) = t\,u(t) = \frac{t}{2}\,\ln\frac{1+t}{1-t} - 1$$

The solution $\psi(t)$ was guaranteed to exist only on $(0,1)$, but clearly it is defined on the whole interval $(-1,1)$ and solves (Le_1).

$\Longrightarrow$ A fundamental system of solutions of (Le_1) is

$$t, \quad \frac{t}{2}\,\ln\frac{1+t}{1-t} - 1.$$

Euler Equations

cf. [BDM17], Ch. 5.4

Definition

The ODE

$$t^2 y'' + \alpha t y' + \beta y = 0 \qquad \text{(E)}$$

is called *Euler equation* with parameters α, β.

We will assume $\alpha, \beta \in \mathbb{R}$ and consider only real solutions. (The complex case is easily reduced to the real case.)

(E) is homogeneous linear, time-dependent, of order 2.

Except for the trivial case $\alpha = \beta = 0$, (E) has a singular point in $t = 0$, where the corresponding explicit equation $y'' + (\alpha/t)y' + (\beta/t^2)y = 0$ is not defined.
$\Longrightarrow$ Solutions exist "independently" on $(-\infty, 0)$, $(0, +\infty)$ and form a 2-dimensional real vector space in both cases.

For Part (2) of the following theorem, recall that solutions of (E) are twice differentiable functions $y : I \to \mathbb{R}$ satisfying (E) for every $t \in I$. For $I = \mathbb{R}$ and $t = 0$ the ODE reduces to $\beta y(0) = 0$, which for $\beta \neq 0$ requires $y(0) = 0$ (in particular associated IVP's with $y(0) = y_0 \neq 0$ are not solvable).

Reflection Principle

1. A solution $\phi\colon (0,+\infty) \to \mathbb{R}$ yields a solution $\psi\colon (-\infty,0) \to \mathbb{R}$ by reflecting the graph of ϕ at the y-axis (and vice versa).
2. ϕ can be extended to a solution on $\mathbb{R}$ whose graph is symmetric w.r.t. the y-axis if $\lim_{t\downarrow 0} \phi'(t) = 0$ and $\lim_{t\downarrow 0} \phi''(t)$ exists in $\mathbb{R}$.

As sketched in the proof, the existence of $\lim_{t\downarrow 0} \phi''(t)$ implies that of $\lim_{t\downarrow 0} \phi'(t)$ and $\lim_{t\downarrow 0} \phi(t)$, and the first condition requires that $\lim_{t\downarrow 0} \phi'(t)$ is zero.

Proof.

(1) For $t < 0$ let $\psi(t) = \phi(-t) = \phi(|t|)$. Then $\psi'(t) = -\phi'(-t)$, $\psi''(t) = \phi''(-t)$, and hence

$$\begin{aligned} t^2\psi''(t) + \alpha t\psi'(t) + \beta\psi(t) &= t^2\phi''(-t) - \alpha t\phi'(-t) + \beta\phi(-t) \\ &= (-t)^2\phi''(-t) + \alpha(-t)\phi'(-t) + \beta\phi(-t) \\ &= 0, \end{aligned}$$

since $s = -t$ runs through $(0,+\infty)$ if t runs through $(-\infty,0)$. This proves Part (1).

Proof cont'd.

(2) The existence of $\lim_{t\downarrow 0}\phi''(t)$ implies that ϕ, which is C^∞ on $(0,+\infty)$, can be extended to a C^2-function on $[0,+\infty)$ (with one-sided derivatives in $t=0$). This follows from

$$\phi'(t) = \phi'(t_0) + \int_{t_0}^{t} \phi''(s)\,\mathrm{d}s,$$

$$\phi(t) = \phi(t_0) + \int_{t_0}^{t} \phi'(s)\,\mathrm{d}s, \qquad t_0, t > 0,$$

which makes also sense for $t=0$ and gives continuous extensions of ϕ', ϕ (first for ϕ', then for ϕ) to $[0,+\infty)$.
It is then not difficult to show that ϕ is twice differentiable at $t=0$ from the right and that $\phi'(0)=\lim_{t\downarrow 0}\phi'(t)$, $\phi''(0)=\lim_{t\downarrow 0}\phi''(t)$.

The reflected function $\psi(t)=\phi(-t)$, $t\in(-\infty,0]$, is C^2 as well and satisfies $\psi(0)=\phi(0)$, $\psi'(0)=-\phi'(0)$, $\psi''(0)=\phi''(0)$.
Hence, provided that $\phi'(0)=0$ (and only then) we obtain a consistent extension of ϕ to $\mathbb{R}$.

Finally, the ODE is satisfied also for $t=0$ (for $\beta=0$ this is trivial, for $\beta\neq 0$ it follows from $y(0)=\lim_{t\to 0}y(t)=\lim_{t\to 0}\frac{-t^2y''(t)-\alpha t\,y'(t)}{\beta}=0$), completing the proof. □

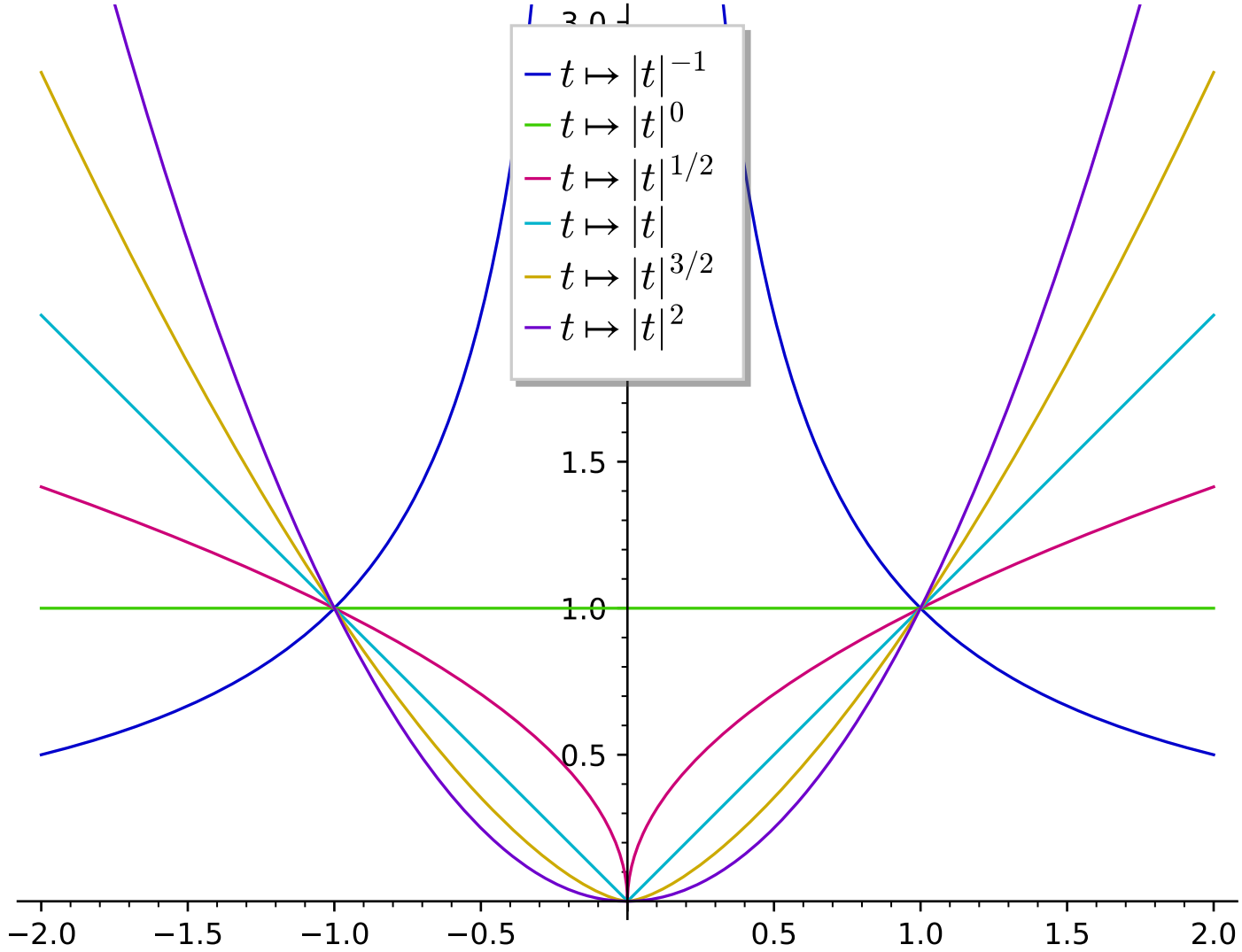


Figure: Illustration of the Reflection Principle: $t \mapsto |t|^r$ defines a C^2-function on $\mathbb{R}$ iff $r = 0 \vee r \geq 2$

Solution of the Euler Equations

Using the reflection principle, we can restrict attention to $t \geq 0$. We have already considered the special case $\alpha = -1/2$, $\beta = 1/2$, in which a fundamental system was given by t and $\sqrt{t}$. The earlier method works with some modifications also in the general case:

Setting $L = t^2\,\mathrm{D}^2 + \alpha t\,\mathrm{D} + \beta\,\mathrm{id}$, the Euler equation becomes $Ly = 0$, and we have

$$L[t^r] = \big(r(r-1) + \alpha r + \beta\big)t^r \equiv 0 \iff r^2 + (\alpha - 1)r + \beta = 0,$$

which is solved by $r_1 = \frac{1}{2}\left(1 - \alpha + \sqrt{(\alpha-1)^2 - 4\beta}\right)$ and $r_2 = \frac{1}{2}\left(1 - \alpha - \sqrt{(\alpha-1)^2 - 4\beta}\right)$.

Case 1: $(\alpha - 1)^2 > 4\beta$

In this case $r_1 > r_2$ are real, so that

$$\phi_1(t) = t^{r_1} \quad \text{and} \quad \phi_2(t) = t^{r_2}$$

(resp., $\phi_1(t) = (-t)^{r_1}$, $\phi_2(t) = (-t)^{r_2}$ for $t < 0$) form a fundamental system of solutions.

Case 1 cont'd

For the following analysis, let S_0 be the solution space with domain $I = \mathbb{R}$, i.e., S_0 consists of all functions $y : \mathbb{R} \to \mathbb{R}$ satisfying (E). Because a solution defined on some interval $(-\delta, \delta)$, $\delta > 0$, can be uniquely extended to $\mathbb{R}$, we can alternatively view S_0 as the local solution space at $t = 0$.

Combinatorial counting gives that there are $\binom{6}{2} - 2 = 13$ (???) cases to consider. We consider only a few of them.

If r_1, r_2 are negative (equivalently $\beta > 0$ and $\alpha > 1 + 2\sqrt{\beta}$) then the only solution defined at $t = 0$ is $y(t) \equiv 0$. In other words, $S_0 = \{0\}$ and the only realizable initial values at $t = 0$ are $y(0) = y'(0) = 0$.

If $r_1 = 0$ (equivalently $\beta = 0$ and $\alpha > 1$) then the solutions defined at $t = 0$ are the constant functions $y(t) \equiv c$, $c \in \mathbb{R}$.
$\Longrightarrow \dim(S_0) = 1$, and the (uniquely) realizable initial values at $t = 0$ are $y(0) = c \in \mathbb{R}$ arbitrary, $y'(0) = 0$.

If $0 < r_1 < 1 \vee 1 < r_1 < 2$ and r_2 is either negative or satisfies the same condition as r_1, then again the only solution defined at $t = 0$ is $y(t) \equiv 0$, and consequently $S_0 = \{0\}$.

If $r_1 = 1$, $r_2 = 0$ (the non-singular case $\alpha = \beta = 0$, in which the general solution is $y(t) = c_1 + c_2 t$) then $\dim(S_0) = 2$ and all initial conditions at $t = 0$ are uniquely realizable.

Case 1 cont'd

If $r_1 > 2$ and $0 < r_2 < 1 \vee 1 < r_2 < 2$, the solutions on $(0, +\infty)$ that can be extended to $[0, +\infty)$ have the form $y(t) = c\,t^{r_1}$ and satisfy $y(0) = y'(0) = y''(0) = 0$. Solutions $y(t) = c_1 t^{r_1}$ on $[0, +\infty)$ and $z(t) = c_2(-t)^{r_1}$ on $(-\infty, 0]$ can be combined freely to yield a solution on $\mathbb{R}$.
$\Longrightarrow \dim(S_0) = 2$, and a basis of S_0 (fundamental system of solutions) is formed by

$$y_1(t) = \begin{cases} t^{r_1} & \text{if } t \geq 0, \\ 0 & \text{if } t < 0, \end{cases} \qquad y_2(t) = \begin{cases} (-t)^{r_1} & \text{if } t \leq 0, \\ 0 & \text{if } t > 0. \end{cases}$$

If $r_1 > r_2 = 2$ then the solutions on $(0, +\infty)$ have the form $y(t) = c_1 t^{r_1} + c_2 t^2$ and can be uniquely extended to $[0, +\infty)$ by setting $y(0) = y'(0) = 0$, $y''(0) = 2c_2$. Solutions $y(t) = c_1 t^{r_1} + c_2 t^2$ on $[0, +\infty)$ and $z(t) = c_3(-t)^{r_1} + c_4(-t)^2$ on $(-\infty, 0]$ can be glued to yield a solution on $\mathbb{R}$ iff $c_2 = c_4$.
$\Longrightarrow \dim(S_0) = 3$, and a basis of S_0 is formed by the functions $y_1(t)$, $y_2(t)$ defined above and $y_3(t) = t^2$.

Case 1 cont'd

If $r_1 > r_2 > 2$ then all solutions $y(t)$ on $(0, +\infty)$ can be uniquely extended to $[0, +\infty)$ by setting $y(0) = y'(0) = y''(0) = 0$. Solutions on $(-\infty, 0]$ and $[0, +\infty)$ can be freely combined to yield solutions on $\mathbb{R}$. $\implies \dim(S_0) = 4$, and a basis of S_0 is formed by

$$y_1(t) = \begin{cases} t^{r_1} & \text{if } t \geq 0, \\ 0 & \text{if } t < 0, \end{cases} \qquad y_2(t) = \begin{cases} (-t)^{r_1} & \text{if } t \leq 0, \\ 0 & \text{if } t > 0, \end{cases}$$

$$y_3(t) = \begin{cases} t^{r_2} & \text{if } t \geq 0, \\ 0 & \text{if } t < 0, \end{cases} \qquad y_4(t) = \begin{cases} (-t)^{r_2} & \text{if } t \leq 0, \\ 0 & \text{if } t > 0. \end{cases}$$

Be sure to understand the precise meaning of "basis" here:

1. Every solution $y(t) \in S_0$ is of the form $y(t) = c_1 y_1(t) + c_2 y_2(t) + c_3 y_3(t) + c_4 y_4(t)$ for some $c_1, c_2, c_3, c_4 \in \mathbb{R}$. This follows from the above discussion and

$$c_1 y_1(t) + c_2 y_2(t) + c_3 y_3(t) + c_4 y_4(t) = \begin{cases} c_1 t^{r_1} + c_3 t^{r_2} & \text{if } t > 0, \\ 0 & \text{if } t = 0, \\ c_2 (-t)^{r_1} + c_4 (-t)^{r_2} & \text{if } t < 0. \end{cases}$$

2. The coefficients c_1, c_2, c_3, c_4 are uniquely determined by $y(t)$.

Case 2: $(\alpha-1)^2 = 4\beta$

Here $r_1 = r_2 = (1-\alpha)/2$, yielding only one fundamental solution $\phi_1(t) = t^{(1-\alpha)/2}$.

A second fundamental solution $\phi_2(t)$ can be determined using order reduction. The cofactor $u(t)$ in $\phi_2(t) = u(t)\phi_1(t)$ satisfies

$$u''(t) + \left(2\,\frac{\phi_1'(t)}{\phi_1(t)} + a(t)\right) u'(t) = u''(t) + \left(2\cdot\frac{1-\alpha}{2t} + \frac{\alpha}{t}\right) u'(t)$$

$$= u''(t) + \frac{1}{t}\,u'(t) = 0.$$

$$\Longrightarrow\ u'(t) = \exp\left(-\int\frac{\mathrm{d}t}{t}\right) = \frac{c_1}{t} \quad\Longrightarrow\quad u(t) = c_1 \ln t + c_2$$

Hence a fundamental system of solutions on $(0,+\infty)$ in this case is

$$\phi_1(t) = t^{(1-\alpha)/2}, \quad \phi_2(t) = (\ln t)t^{(1-\alpha)/2}.$$

Extendability to solutions on $\mathbb{R}$ is discussed in the same way as before. We omit the general discussion, but one case is worth noting: For $\alpha < -3$ (the case in which $\frac{1-\alpha}{2} > 2$) we have $\lim_{t\downarrow 0}\phi_2(t) = \lim_{t\downarrow 0}\phi_2'(t) = \lim_{t\downarrow 0}\phi_2''(t) = 0$. Hence the analysis done for $r_1 > r_2 > 2$ carries over, and it follows that $\dim(S_0) = 4$.

Case 3: $(\alpha-1)^2 < 4\beta$

In this case r_1, r_2 are complex,

$$r_1 = \frac{1}{2}\left(1-\alpha+\mathrm{i}\sqrt{4\beta-(\alpha-1)^2}\right),$$
$$r_2 = \frac{1}{2}\left(1-\alpha-\mathrm{i}\sqrt{4\beta-(\alpha-1)^2}\right) = \overline{r_1}.$$

A complex fundamental system of solutions is again t^{r_1}, t^{r_2}.
A real fundamental system can be obtained by extracting real and imaginary part of one of these. Writing $r_{1/2} = \lambda \pm \mathrm{i}\mu$, we get

$$y_1(t) = \mathrm{Re}(t^{\lambda+\mathrm{i}\mu}) = \mathrm{Re}\left(t^\lambda \mathrm{e}^{\mathrm{i}\mu \ln t}\right) = t^\lambda \cos(\mu \ln t),$$
$$y_2(t) = t^\lambda \sin(\mu \ln t).$$

In this case the determination of S_0 is comparatively easy: Nonzero solutions on $(0,+\infty)$ are extendable to solutions on $\mathbb{R}$ iff $\lambda > 2$. (For $\lambda = 2$ the 2nd derivative of $y_1(t)$, $y_2(t)$ oszillates wildly near $t = 0$; the same is true of any nonzero linear combination of $y_1(t)$, $y_2(t)$.) If $\lambda > 2$ then all solutions on $[0,+\infty)$ satisfy $y(0) = y'(0) = y''(0) = 0$, and hence we have again $\dim(S_0) = 4$.

Math 285
Introduction to Differential Equations

Thomas Honold

ZJU-UIUC Institute

Spring Semester 2024

Outline

Today's Lecture: Series Solutions of Ordinary (mainly 2nd-Order Linear) Differential Equations

Definition

An *analytic solution* (or *power series solution*) of a scalar ODE is a "power series function"

$$y(t) = \sum_{n=0}^{\infty} a_n (t - t_0)^n, \quad t \in I,$$

for some interval $I \subseteq \mathbb{R}$ of positive length and some $t_0 \in I$.

Notes

- The name "analytic" comes from *analytic function*, which refers to a function $f : D \to \mathbb{C}$, $D \subseteq \mathbb{C}$, which locally admits a power series representation $f(z) = \sum_{n=0}^{\infty} a_n (z - z_0)^n$ (either for all $z_0 \in D$, in which case f is said to be holomorphic, or only for a fixed point $z_0 \in D$).
- The (w.l.o.g. open) interval I must be contained in the interval of convergence of $\sum_{n=0}^{\infty} a_n (t - t_0)^n$, which is of the form $(t_0 - \rho, t_0 + \rho) = \mathrm{B}_\rho(t_0) \cap \mathbb{R}$ with $\mathrm{B}_\rho(t_0)$ denoting the open disk of convergence of $\sum_{n=0}^{\infty} a_n (z - t_0)^n$. In particular the radius of convergence ρ of the power series must be positive.

Notes cont'd

- There is an "analytic" version of the Existence and Uniqueness Theorem, which roughly says that if $f(t, y_0, y_1, \ldots, y_{n-1})$ is analytic then the solution of an IVP $y^{(n)} = f(t, y, y', \ldots, y^{(n-1)})$, $y^{(i)}(t_0) = c_i$ for $0 \leq i \leq n-1$ must be analytic at t_0 and solve the ODE wherever it is defined.

As a consequence of this theorem we can solve ODE's in the analytic case by a power series „Ansatz".

In what follows, we will switch notation from $y(t)$ to $y(x)$, because for power series the variable symbol 'x' is more common in view of the link $z = x + y\,\mathrm{i}$ with the complex case.

Example

Determine a solution of the IVP $y' = x^2 + y^2$, $y(0) = 1$.

Because $f(x, y) = x^2 + y^2$ is analytic, the solution must be analytic as well, i.e., of the form $y(x) = \sum_{n=0}^{\infty} a_n x^n$ with $a_n = y^{(n)}(0)/n!$.

Method 1: Determine $y^{(n)}(0)$ from the ODE.

$$
\begin{aligned}
y' &= x^2 + y^2 && \Longrightarrow && y'(0) &= 0^2 + y(0)^2 = 1,\\
y'' &= 2x + 2yy' && \Longrightarrow && y''(0) &= 2y(0)y'(0) = 2,\\
y''' &= 2 + 2y'^2 + 2yy'' && \Longrightarrow && y'''(0) &= 2 + 2 + 4 = 8,\\
y^{(4)} &= 6y'y'' + 2yy''' && \Longrightarrow && y^{(4)}(0) &= 12 + 16 = 28
\end{aligned}
$$

$$
\begin{aligned}
\Longrightarrow\ y(x) &= 1 + x + \tfrac{2}{2!}\,x^2 + \tfrac{8}{3!}\,x^2 + \tfrac{28}{4!}\,x^4 + \cdots\\
&= 1 + x + x^2 + \tfrac{4}{3}x^3 + \tfrac{7}{6}x^4 + \cdots
\end{aligned}
$$

This method is cumbersome, and it does not tell us anything about the radius of convergence of the resulting power series, and hence about the domain of the solution (except that by the general theory it must be an interval of positive length).

Example (cont'd)

Method 2: Substitute $y(x) = \sum_{n=0}^{\infty} a_n x^n$ into the ODE and equate coefficients.

$$y'(x) = \sum_{n=1}^{\infty} n a_n x^{n-1} = \sum_{n=0}^{\infty} (n+1) a_{n+1} x^n,$$

$$x^2 + y(x)^2 = x^2 + \sum_{n=0}^{\infty} \left(\sum_{k=0}^{n} a_k a_{n-k} \right) x^n$$

$$\Longrightarrow \quad (n+1) a_{n+1} = \begin{cases} \sum_{k=0}^{n} a_k a_{n-k} & \text{if } n \neq 2, \\ 1 + 2a_0 a_2 + a_1^2 & \text{if } n = 2 \end{cases}$$

For $n \geq 1$ this determines a_n from a_k, $k < n$, and thus together with the initial value $a_0 = y(0) = 1$ provides a recursion formula for a_n, which can easily be programmed:

$$y(x) = 1 + x + x^2 + \tfrac{4}{3}x^3 + \tfrac{7}{6}x^4 + \tfrac{6}{5}x^5 + \tfrac{37}{30}x^6 + \tfrac{404}{315}x^7 + \tfrac{369}{280}x^8 + \tfrac{428}{315}x^9 + \\ + \tfrac{1961}{1400}x^{10} + \tfrac{75092}{51975}x^{11} + \tfrac{1238759}{831600}x^{12} + \tfrac{9884}{6435}x^{13} + \cdots$$

Example (cont'd)

The radius of convergence:

It is clear that all a_n are positive. Using induction, we can easily show that $a_n \geq 1$ for all n:

$$\Longrightarrow a_n = \frac{a_0 a_{n-1} + a_1 a_{n-2} + \cdots + a_{n-1} a_0}{n} \geq \frac{1^2 + 1^2 + \cdots + 1^2}{n} = 1$$

$$\Longrightarrow \rho \leq 1$$

Conversely, suppose $a_k \leq c^k$ for all $k < n$ and some constant c.

$$a_n \leq \frac{c^0 c^{n-1} + c^1 c^{n-2} + \cdots + c^{n-1} c^0}{n} = c^{n-1} \leq c^n,$$

provided that $c \geq 1$ and $n \geq 4$.

$\Longrightarrow a_N \leq c^N$ for all $N \geq n$ (using induction)

$\Longrightarrow \rho \geq 1/c$.

For example we can take $n = 4$ and $c = \sqrt[3]{\frac{4}{3}}$.

$$\Longrightarrow \rho \geq \sqrt[3]{\frac{3}{4}} = 0.9085\cdots > 0.9$$

Example

Determine the general solution of $y' = y^2$ using the power series „Ansatz".

Since $y' = y^2$ is autonomous, we can restrict ourselves to the case $x_0 = 0$, i.e., make again the „Ansatz" $y(x) = \sum_{n=0}^{\infty} a_n x^n$. The recursion formula of the previous example changes to

$$a_n = \frac{1}{n} \sum_{k=0}^{n-1} a_k a_{n-1-k} \quad \text{for all } n \geq 1.$$

$\Longrightarrow a_1 = a_0^2$, $a_2 = \frac{1}{2}(a_0 a_1 + a_1 a_0) = a_0^3$,
$a_3 = \frac{1}{3}(a_0 a_2 + a_1^2 + a_2 a_0) = a_0^4$, etc., and in general $a_n = a_0^{n+1}$.

$$\Longrightarrow \; y(x) = \sum_{n=0}^{\infty} a_0^{n+1} x^n = a_0 \sum_{n=0}^{\infty} (a_0 x)^n = \frac{a_0}{1 - a_0 x}, \quad |x| < \frac{1}{|a_0|} = \rho$$

Thus $y(x) = 1/(C - x)$ with $C = 1/a_0$, recovering the previously determined general solution of $y' = y^2$.

Example (cont'd)

Notes

- Two solutions, which are not defined at $x_0 = 0$, were missed. These are

$$y_1(x) = -\frac{1}{x} \quad \text{for } x < 0,$$
$$y_2(x) = -\frac{1}{x} \quad \text{for } x > 0.$$

  They can be formally included in $y(x) = a_0/(1 - a_0 x)$ if we permit $a_0 = \infty$.

- The solution $y(x) = a_0/(1 - a_0 x)$ is defined on the unbounded interval containing 0 and having $1/a_0$ as one endpoint (the other endpoint is $+\infty$ or $-\infty$). But the power series representation is valid only on the bounded subinterval $|x| < \frac{1}{|a_0|}$.

Two Euler-Like Equations

Example

We consider simultaneously the ODE's

$$xy'' + y' + y = 0, \tag{E1}$$

$$x^2y'' + y' + y = 0, \tag{E2}$$

which have a singularity at $x = 0$ like the general Euler equation. Note that, in a way, (E2) is more singular at $x = 0$ than (E1).

Making the usual power series „Ansatz" $y(x) = \sum_{n=0}^{\infty} a_n x^n$ and using

$$y'(x) = \sum_{n=1}^{\infty} n a_n x^{n-1} = \sum_{n=0}^{\infty} (n+1) a_{n+1} x^n,$$

$$x\,y''(x) = \sum_{n=2}^{\infty} n(n-1) a_n x^{n-1} = \sum_{n=1}^{\infty} (n+1) n a_{n+1} x^n,$$

$$x^2\,y''(x) = \sum_{n=2}^{\infty} n(n-1) a_n x^n,$$

Example (cont'd)

the ODE's become

$$a_0 + a_1 + \sum_{n=1}^{\infty} \big((n+1)na_{n+1} + (n+1)a_{n+1} + a_n\big)x^n = 0,$$

$$a_0 + a_1 + (a_1 + 2a_2)x + \sum_{n=2}^{\infty} \big(n(n-1)a_n + (n+1)a_{n+1} + a_n\big)x^n = 0.$$

Equating coefficients gives the recursion formulas

$$a_{n+1} = -\frac{1}{(n+1)^2}\, a_n \quad \text{for } n = 0, 1, 2, \ldots, \tag{E1}$$

$$a_{n+1} = -\frac{n^2 - n + 1}{n+1}\, a_n \quad \text{for } n = 0, 1, 2, \ldots \tag{E2}$$

$\Longrightarrow \rho = \infty$ for (E1), giving the analytic solution

$$y(x) = a_0 \sum_{n=0}^{\infty} \frac{(-1)^n}{(n!)^2}\, x^n \quad \text{for } x \in \mathbb{R}.$$

$\Longrightarrow \rho = 0$ for (E2), except in the trivial case $a_0 = 0$, giving no nonzero analytic solution.

Recall that a power series is a series of the form

$$a(z) = \sum_{n=0}^{\infty} a_n (z - z_0)^n, \quad \text{with } z,\, z_0,\, a_n \in \mathbb{C}.$$

The complex number z_0 (*center* of the power series) can often be assumed to be zero, since we can make the translation $z \mapsto z + z_0$.

Definition

1. A function $f\colon D \to \mathbb{C}$, $D \subseteq \mathbb{C}$, is *analytic* in $z_0 \in D$, if z_0 is an inner point of D and there exist $a_n \in \mathbb{C}$ and $\delta > 0$ such that

$$f(z) = \sum_{n=0}^{\infty} a_n (z - z_0)^n \quad \text{for } z \in \mathbb{C} \text{ with } |z - z_0| < \delta,$$

   and *analytic* per se (or *analytic in D*) if f is analytic in every point of D.

2. A function $f\colon D \to \mathbb{R}$, $D \subseteq \mathbb{R}$, is *analytic* in $x_0 \in D$, if x_0 is an inner point of D and there exist $a_n \in \mathbb{R}$ and $\delta > 0$ such that

$$f(x) = \sum_{n=0}^{\infty} a_n (x - x_0)^n \quad \text{for } x \in \mathbb{R} \text{ with } x_0 - \delta < x < x_0 + \delta.$$

Properties of Analytic Functions

1. A real function f satisfying Part (2) of the definition can be extended to a complex function satisfying Part (1) with $z_0 = x_0$ and the same δ, since the radius of convergence of $\sum_{n=0}^{\infty} a_n(z - x_0)^n$ must be $\geq \delta$ and hence the power series converges for all z in the disk $\mathrm{B}_\delta(x_0) = \{z \in \mathbb{C}; |z - x_0| < \delta\}$. For this reason it is hardly necessary to discuss the case of real analytic functions separately. Just consider $\frac{1}{1-z}$, e^z, $\cos z$, $\sin z$, etc. in place of $\frac{1}{1-x}$, e^x, $\cos x$, $\sin x$, etc.

2. For every power series $\sum_{n=0}^{\infty} a_n(z - z_0)^n$ there exists $0 \leq \rho \leq +\infty$ (*radius of convergence*) such that the power series converges for $|z - z_0| < \rho$ and diverges for $|z - z_0| > \rho$. The number ρ is given by

$$\rho = \frac{1}{L}, \quad \text{where} \quad L = \limsup_{n\to\infty} \sqrt[n]{|a_n|}.$$

If $\lim_{n\to\infty} \frac{|a_{n+1}|}{|a_n|}$ exists, it must be equal to L, giving another formula for ρ. But the latter is not directly applicable to power series with gaps such that $\sum_{k=1}^{\infty} z^{k^2}$; cf. our earlier discussion.

Properties cont'd

3. Power series $f(z) = \sum_{n=0}^{\infty} a_n (z - z_0)^n$ can be differentiated termwise within their open disc of convergence, and the power series

$$f'(z) = \sum_{n=1}^{\infty} n a_n (z - z_0)^{n-1} = \sum_{n=0}^{\infty} (n+1) a_{n+1} (z - z_0)^n$$

has the same radius of convergence. Iterating, we obtain

$$f^{(k)}(z) = \sum_{n=0}^{\infty} (n+k)(n+k-1)\cdots(n+1) a_{n+k} (z - z_0)^n,$$

$$f^{(k)}(z_0) = k!\, a_k,$$

so that $a_k = f^{(k)}(z_0)/k!$ and $\quad f(z) = \sum_{n=0}^{\infty} \frac{f^{(n)}(z_0)}{n!} (z - z_0)^n.$

In other words, a function f is analytic at z_0 if all derivatives $f^{(n)}(z_0)$ exist, so that we can form the Taylor series of f in z_0, and the Taylor series converges and represents f in some neighborhood of z_0.

Properties cont'd

4. Suppose $f(z) = \sum_{n=0}^{\infty} a_n (z - z_0)^n$ has radius of convergence $\rho > 0$ and $z_1 \in \mathrm{B}_\rho(z_0)$. Then $f(z)$ can be expanded into a power series with center z_1 in the disk $|z - z_1| < \rho - |z_1 - z_0|$ (the disk inside $\mathrm{B}_\rho(z_0)$ that is centered at z_1 and touches the circle $|z - z_0| = \rho$). In particular f is analytic in the whole disk $\mathrm{B}_\rho(z_0) = \{z \in \mathbb{C}; |z - z_0| < \rho\}$.
   If $\rho = \infty$ then the new power series has radius of convergence ∞ as well, and both series represent f everywhere in $\mathbb{C}$.

Proof.

$$
\begin{aligned}
f(z) &= \sum_{n=0}^{\infty} a_n (z - z_0)^n = \sum_{n=0}^{\infty} a_n (z - z_1 + z_1 - z_0)^n \\
&= \sum_{n=0}^{\infty} \sum_{k=0}^{n} a_n \binom{n}{k} (z - z_1)^k (z_1 - z_0)^{n-k} \\
&= \sum_{k=0}^{\infty} \left(\sum_{n=k}^{\infty} a_n \binom{n}{k} (z_1 - z_0)^{n-k} \right) (z - z_1)^k .
\end{aligned}
$$

Since $\sum_{n=k}^{\infty} a_n \binom{n}{k} (z_1 - z_0)^{n-k} = \frac{f^{(k)}(z_1)}{k!}$, we see from this that the new series is the Taylor series of f in z_1, which comes as no surprise.

Proof cont'd.

The reordering is valid if the double series converges absolutely. Since

$$\sum_{n=0}^{\infty}\sum_{k=0}^{n}\left|a_n\binom{n}{k}(z-z_0)^k(z_1-z_0)^{n-k}\right| = \sum_{n=0}^{\infty}|a_n|\left(|z-z_1|+|z_1-z_0|\right)^n,$$

this is the case provided that $|z-z_1|+|z_1-z_0|) < \rho$ (because $\sum_{n=0}^{\infty} a_n(z-z_0)^n$ converges absolutely in $\mathrm{B}_\rho(z_0)$). □

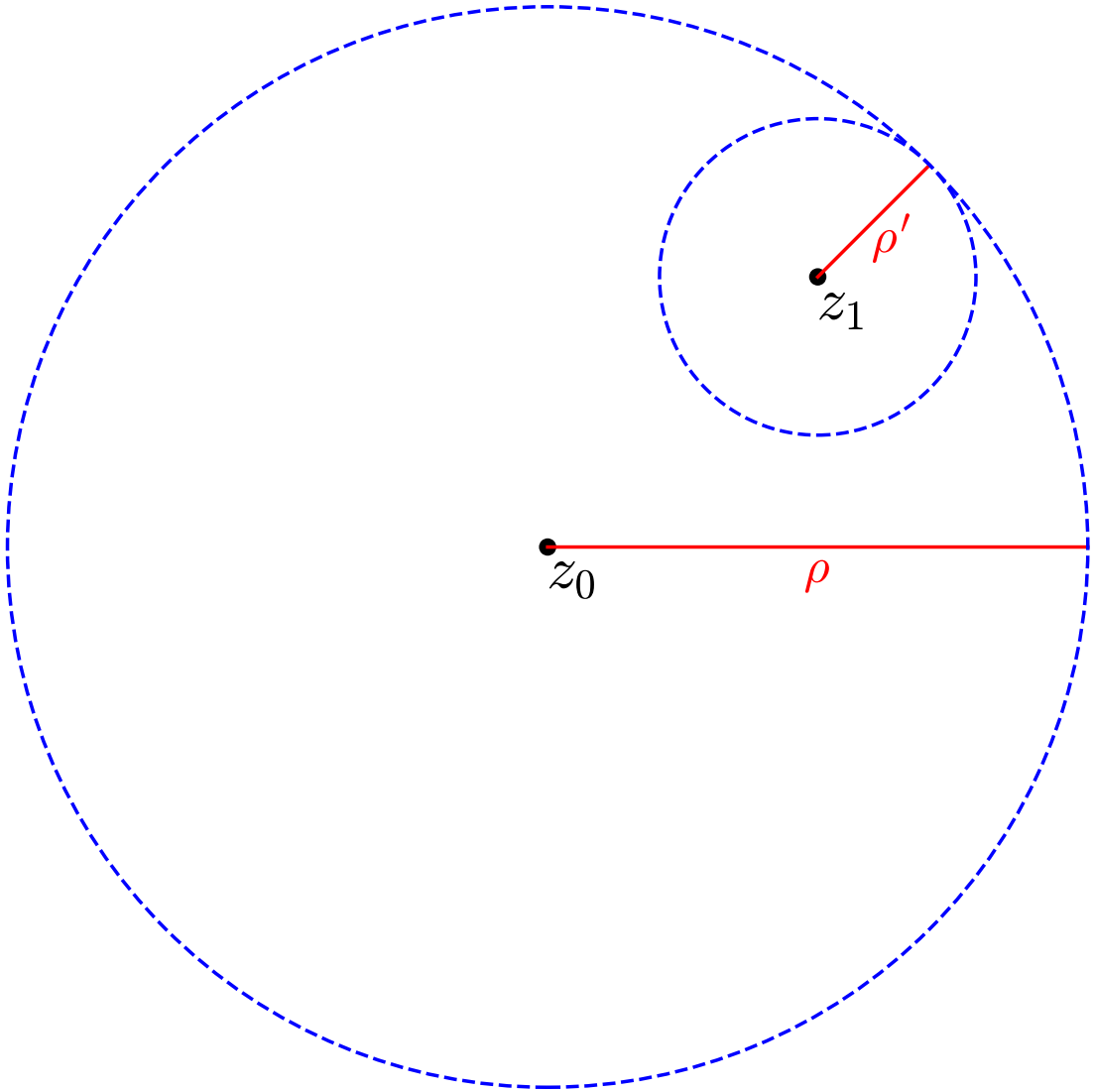


Figure: The Taylor series of f in z_1 converges at least in the open disk $|z - z_1| < \rho'$, $\rho' = \rho - |z_1 - z_0|$, and represents f in this disk.

Interlude on Double Series

Roughly speaking, infinite double series bear to doubly-infinite matrices ("doubly-infinite sequences") the same relation as infinite series do to infinite sequences.

Theorem (Fubini's Theorem for double series)

Suppose $a\colon \mathbb{N}\times\mathbb{N}\to\mathbb{C}$, $(m,n)\mapsto a(m,n)=a_{mn}$ *is any function (called a doubly-infinite matrix), and there exists* $B>0$ *such that* $\sum_{m=0}^{M}\sum_{n=0}^{N}|a_{mn}|\le B$ *for all* $(M,N)\in\mathbb{N}\times\mathbb{N}$. *Then*

$$\sum_{m=0}^{\infty}\left(\sum_{n=0}^{\infty}a_{mn}\right)=\sum_{(m,n)\in\mathbb{N}\times\mathbb{N}}a_{mn}=\sum_{n=0}^{\infty}\left(\sum_{m=0}^{\infty}a_{mn}\right),$$

where it is understood that $\mathbb{N}=\{0,1,2,\dots\}$ *and all series and double series involved converge in* $\mathbb{C}$.

Of course the same theorem holds mutatis mutandis for functions $a\colon\mathbb{N}\times\mathbb{N}\to\mathbb{R}$ and with $\{1,2,3,\dots\}$ in place of $\mathbb{N}$.

The assumption of the theorem implies that $\sum_{(m,n)\in\mathbb{N}\times\mathbb{N}}|a_{mn}|$ converges in $\mathbb{C}$ (resp., $\mathbb{R}$) as well and is often stated as "the double series $\sum_{(m,n)\in\mathbb{N}\times\mathbb{N}}a_{mn}$ converges absolutely".

Interlude on Double Series Cont'd

The theorem says in particular that for a doubly-infinite matrix (a_{mn}) whose "non-negative" partial sums $\sum_{m,n} |a_{mn}|$ over finite rectangles satisfy a uniform bound as stated we can compute the total sum of the matrix either row-wise or column-wise:

$m\backslash n$	0	1	2	...	$\sum$
0	a_{00}	a_{01}	a_{02}	...	r_0
1	a_{10}	a_{11}	a_{12}	...	r_1
2	a_{20}	a_{21}	a_{22}	...	r_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
$\sum$	c_0	c_1	c_2	...	s

If r_m denotes the sum of Row m and c_n the sum of Column n, we have $\sum_{m=0}^{\infty} r_m = \sum_{n=0}^{\infty} c_n$ (denoted by s in the matrix).

For ordinary matrices (i.e., with a finite number of rows and columns) this property is a rather trivial consequence of the commutative and associative laws for addition in $\mathbb{C}$, resp., $\mathbb{R}$.

Interlude on Double Series Cont'd

As a concrete example consider $a_{mn} = \frac{1}{2^m 3^n}$. Here we obtain

$m\backslash n$	0	1	2	...	$\sum$
0	1	1/3	1/9	...	3/2
1	1/2	1/6	1/18	...	3/4
2	1/4	1/12	1/36	...	3/8
⋮	⋮	⋮	⋮		⋮
$\sum$	2	2/3	2/9	...	3

The column sums arise from the geometric series evaluation $1 + 1/2 + 1/4 + \cdots = \frac{1}{1-1/2} = 2$, the row sums from $1 + 1/3 + 1/9 + \cdots = \frac{1}{1-1/3} = 3/2$, and we have indeed

$$2 + \tfrac{2}{3} + \tfrac{2}{9} + \cdots = 3 = \tfrac{3}{2} + \tfrac{3}{4} + \tfrac{3}{8} + \cdots .$$

In fact the identity

$$\sum_{m,n=0}^{\infty} \frac{1}{2^m 3^n} = \left(\sum_{m=0}^{\infty} \frac{1}{2^m}\right)\left(\sum_{n=0}^{\infty} \frac{1}{3^n}\right) = 2 \cdot \tfrac{3}{2} = 3$$

is a discrete analogue of $\int g(x)h(y)\mathrm{d}^2(x,y) = \left(\int g(x)\,\mathrm{d}x\right)\left(\int h(y)\,\mathrm{d}y\right)$.

Interlude on Double Series Cont'd

But what is $\sum_{(m,n)\in\mathbb{N}\times\mathbb{N}} a_{mn}$ in the first place?

It is possible to define "$\sum_{(m,n)\in\mathbb{N}\times\mathbb{N}} a_{mn} = A$" as "for every $\epsilon > 0$ there exist $M_\epsilon, N_\epsilon \in \mathbb{N}$ such that $\left|\left(\sum_{m=1}^{M}\sum_{n=1}^{N} a_{mn}\right) - A\right| < \epsilon$ for all $M > M_\epsilon$ and $N > N_\epsilon$". But this definition doesn't imply that $\sum_{(m,n)\in\mathbb{N}\times\mathbb{N}} a_{mn}$ is preserved under permutations of $\mathbb{N}\times\mathbb{N}$, as the notation "$\sum_{(m,n)\in\mathbb{N}\times\mathbb{N}} a_{mn}$" suggests.

Modern definition: $\sum_{(m,n)\in\mathbb{N}\times\mathbb{N}} a_{mn} = A$ if for every $\epsilon > 0$ there exists a finite set $F \subset \mathbb{N}\times\mathbb{N}$ such that for every finite set E with $F \subset E \subset \mathbb{N}\times\mathbb{N}$ we have $\left|\left(\sum_{(m,n)\in E} a_{mn}\right) - A\right| < \epsilon$.

Note that finite sums $\sum_{(m,n)\in E} a_{mn}$ are well defined, since it doesn't matter in which order we add the elements a_{mn} (by the commutative and associative laws in $\mathbb{C}$).

Interlude on Double Series Cont'd

Notes

- The modern definition applies mutatis mutandis to every complex-valued (or real-valued) function and yields a definition of $\sum_{i\in I} a_i$ for any domain ("index set") I and function $a\colon I \to \mathbb{C}$, $i \mapsto a_i$.
- Only countably infinite domains I are of interest, since $\sum_{i\in I} a_i = A \in \mathbb{C}$ implies that $\{i \in I; a_i \neq 0\}$ is either finite or countably infinite.
- $\sum_{i\in I} a_i$ exists (in $\mathbb{C}$) iff $\sum_{i\in I} |a_i|$ exists (in $\mathbb{R}$). In other words, there is no difference between convergence and absolute convergence; cp. with the Lebesgue integral, of which the modern definition of infinite summation is actually a special case.
- If $\pi\colon I \to I$ is any permutation (i.e., bijection) then $\sum_{i\in I} a_i = \sum_{i\in I} a_{\pi(i)}$, a trivial consequence of the definition.
- If $\sum_{i\in I} a_i$ exists and $J \subset I$, the sum $\sum_{i\in J} a_i$ exists as well.
- If $\sum_{i\in I} a_i$ exists and $\mathcal{P}$ is a partition of I then $\sum_{i\in I} a_i = \sum_{J\in\mathcal{P}} \left(\sum_{i\in J} a_i\right)$. Fubini's Theorem for double series is a special case of this.

Interlude on Double Series Cont'd

More precisely, Fubini's Theorem for double series is the statement

$$\sum_{R\in\mathcal{R}}\left(\sum_{(m,n)\in R} a_{mn}\right) = \sum_{(m,n)\in\mathbb{N}\times\mathbb{N}} a_{mn} = \sum_{C\in\mathcal{C}}\left(\sum_{(m,n)\in C} a_{mn}\right),$$

where $\mathcal{R}$, $\mathcal{C}$ are the partitions of $\mathbb{N}\times\mathbb{N}$ into "rows" resp. "columns"; i.e., the members of $\mathcal{R}$ are $\{0\}\times\mathbb{N}$, $\{1\}\times\mathbb{N}$, $\{2\}\times\mathbb{N}$, ..., and the members of $\mathcal{C}$ are $\mathbb{N}\times\{0\}$, $\mathbb{N}\times\{1\}$, $\mathbb{N}\times\{2\}$, ...

The last property is also meaningful for ordinary series. For example, it tells us that for an absolutely convergent series we have

$$a_1 + a_2 + a_3 + \cdots = (a_1 + a_3 + a_5 + \dots) + (a_2 + a_4 + a_6 + \cdots),$$

and also the rather fancy

$$\sum_{n=1}^{\infty} a_n = a_1 + (a_2 + a_3 + a_5 + a_7 + a_{11} + a_{13} + \cdots) + (a_4 + a_6 + a_9 + a_{10} + a_{14} + \cdots) + (a_8 + a_{12} + \cdots) + \cdots.$$

Inner sums are taken over all n with a fixed number of prime factors.

Interlude on Double Series Cont'd

Proofs of these properties can be found in some texts on Real Analysis. Walter Rudin's Principles of Mathematical Analysis that I have recommended as background reference doesn't include it, but Terence Tao's Analysis I (3rd edition, Springer 2015) has it in Ch. 8.2, for example. The notation there is slightly different from our's, and yet different from the one in my source (a not so well-known German textbook).

Properties cont'd

5. *Equating coefficients*
   Suppose f and g are analytic in some common connected domain D (i.e., analytic at every point of D) and that $E = \{z \in D; f(z) = g(z)\}$ has an accumulation point in D. Then $f(z) = g(z)$ for all $z \in D$, and consequently the power series expansions of f and g at any point $z_0 \in D$ must be the same.

Sketch of proof.

Call the accumulation point z_0 and suppose that $f(z) = \sum_{n=0}^{\infty} a_n(z-z_0)^n$, $g(z) = \sum_{n=0}^{\infty} b_n(z-z_0)^n$ are represented by different power series at z_0, i.e., $a_n \neq b_n$ for some n. If N is the least such n, we have

$$\begin{aligned} f(z) - g(z) &= (a_N - b_N)(z-z_0)^N + (a_{N+1} - b_{N+1})(z-z_0)^{N+1} + \cdots \\ &= (z-z_0)^N \left(a_N - b_N + (a_{N+1} - b_{N+1})(z-z_0) + \cdots\right) \\ &= (z-z_0)^N h(z), \end{aligned}$$

where h is analytic at z_0 and $h(z_0) = a_N - b_N \neq 0$.
$\Longrightarrow h(z) \neq 0$ in some disk $|z-z_0| < \delta$ (since h is continuous)
$\Longrightarrow f(z) \neq g(z)$ in the punctured disk $0 < |z-z_0| < \delta$.
This contradicts the assumption that z_0 is an accumulation point of E.

Proof cont'd.

We have thus proved that the set $E_1 \subseteq E$ consisting of all points $z_0 \in D$ where f and g are represented by the same power series is non-empty.

E_1 is closed in D, since to any limit point z_0 of E_1 in D we can apply the preceding argument to show that $z_0 \in E_1$.

But E_1 is also open, since for $z_0 \in E_1$ the functions f and g are represented by the same power series in some disk $|z - z_0| < \delta$, which must be contained in E_1 by Property 4. (For this note that both f and g are represented by a power series $\sum_{k=0}^{\infty} c_k (z-a)^k$ at any point $a \in \mathrm{B}_\delta(z_0)$; the coefficients c_k can be computed from the power series representation at z_0, viz. $c_k = \sum_{n=k}^{\infty} a_n \binom{n}{k} (a - z_0)^{n-k}$, and hence must be the same for f and g.)

Since D is connected, this implies $D = E_1$ and in particular that $f(z) = g(z)$ for all $z \in D$. □

Remark

Property 5 holds a forteriori for real analytic functions defined on an open interval $D \subseteq \mathbb{R}$. For C^∞-functions on $\mathbb{R}$ it grossly fails: There exists, e.g., a C^∞-function $f\colon \mathbb{R} \to \mathbb{R}$ satisfying $f(x) = 0$ for $x \leq 0$ and $f(x) > 0$ for $x > 0$; cf. also the subsequent example of a "bell-shaped" function.

Example

Consider $f\colon \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} e^{-\frac{1}{1-x^2}} & \text{for } |x| < 1, \\ 0 & \text{for } |x| \geq 1. \end{cases}$$

Then f satisfies $f^{(n)}(\pm 1) = 0$ for all $n \geq 0$, since on $(-1,1)$ all derivatives $f^{(n)}(x)$ exist and have the form $f^{(n)}(x) = R_n(x)e^{-\frac{1}{1-x^2}}$ for some rational function $R_n(x)$. It follows that $\lim_{x\to\pm 1} f^{(n)}(x) = 0$, and this is enough to prove by induction that $f^{(n)}(\pm 1)$ exists and equals 0. Thus f is a C^∞-function on $\mathbb{R}$.

But f is not analytic at $x_0 = \pm 1$, since the Taylor series at ± 1 vanishes but f does not vanish in any neighborhood of ± 1.

Moreover, f vanishes on a large subset of $\mathbb{R}$ but not entirely. This cannot happen for (per se) analytic functions (cf. Property 5): If a real analytic function g vanishes on an interval of positive length, it must vanish entirely. Similarly, if g is analytic, defined at zero, and $g(1/n) = 0$ for all sufficiently large integers n then g must vanish entirely.

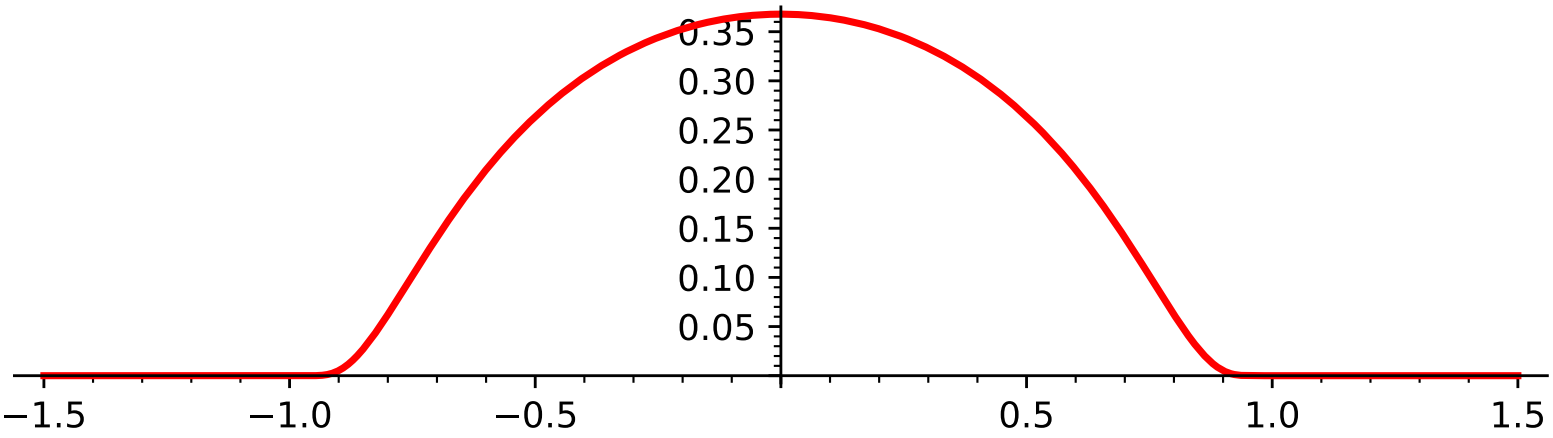


Figure: Graph of $f(x) = \mathrm{e}^{-\frac{1}{1-x^2}}$ for $|x| < 1$, $f(x) = 0$ for $|x| \geq 1$

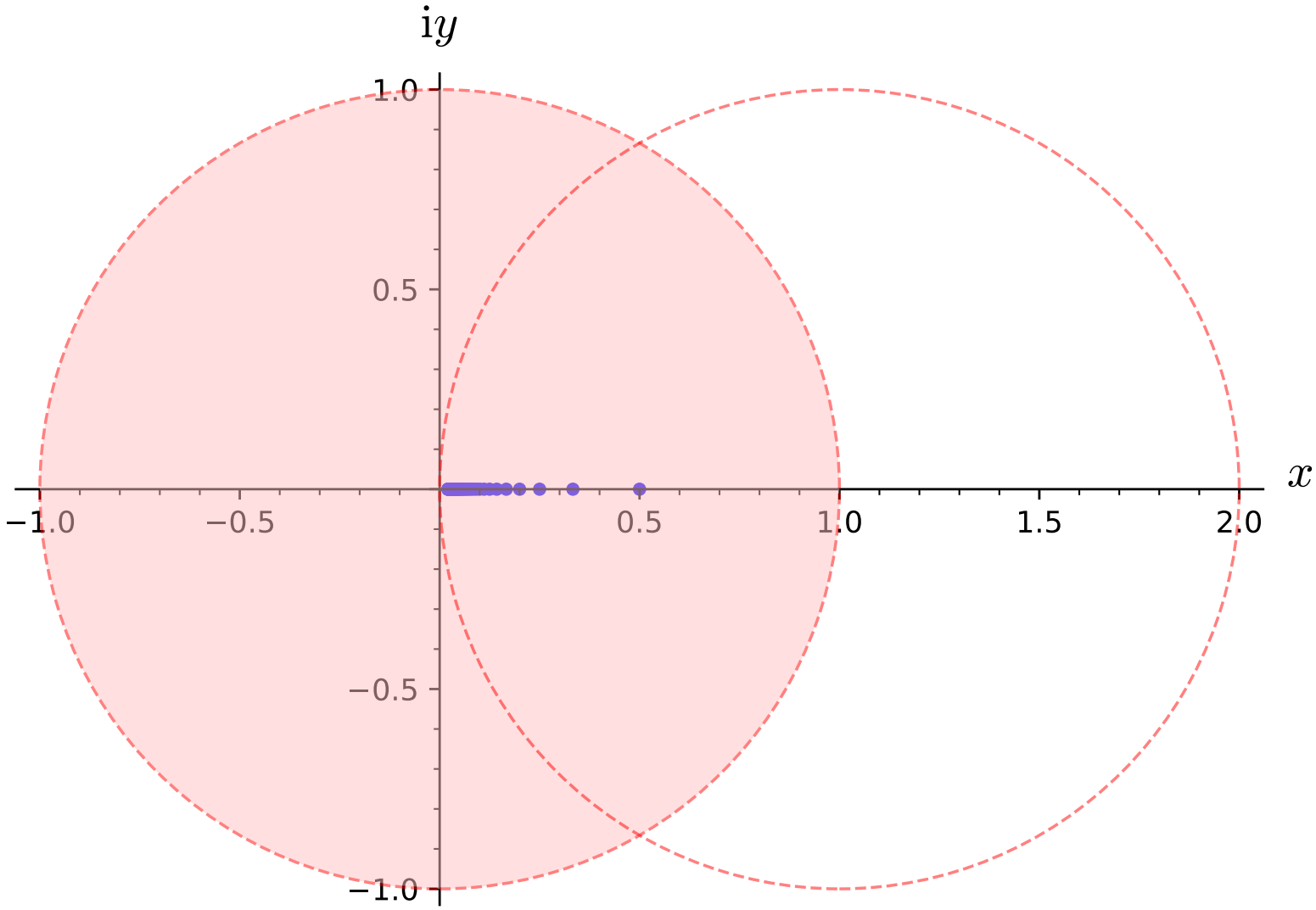


Figure: There is no nonzero analytic function defined on the unit disk $|z| < 1$ and having zeros at $z = 1/k$, $k = 2, 3, 4, \ldots$, but there is such a function defined on the disk $|z - 1| < 1$, e.g., $z \mapsto \sin(\pi/z)$.

Properties cont'd

6. *Algebraic Operations on Power Series*
   We assume for the following w.l.o.g. that the centers of the power series involved are equal to 0.

   Power series functions $f(z) = \sum_{n=0}^{\infty} a_n z^n$, $g(z) = \sum_{n=0}^{\infty} b_n z^n$ can be added/subtracted/multiplied by scalars termwise,

$$f(z) \pm g(z) = \sum_{n=0}^{\infty} (a_n \pm b_n) z^n,$$

$$c\, f(z) = \sum_{n=0}^{\infty} (c a_n) z^n,$$

   and multiplied according to Cauchy's multiplication formula

$$f(z)g(z) = \sum_{n=0}^{\infty} \left(\sum_{k=0}^{n} a_k b_{n-k} \right) z^n.$$

   The radius of convergence of the resulting power series is at least the minimum of the radii of convergence of f and g. In particular, sums and products of analytic functions are again analytic.

Properties cont'd

6. *Algebraic Operations on Power Series cont'd*
   Moreover, if $b_0 = g(0) \neq 0$ then the quotient $h(z) = f(z)/g(z)$ is represented in some neighborhood of 0 by a power series $\sum_{n=0}^{\infty} c_n z^n$ as well, which can be obtained by solving

$$\sum_{n=0}^{\infty} a_n z^n = f(z) = g(z)h(z) = \left(\sum_{n=0}^{\infty} b_n z^n\right)\left(\sum_{n=0}^{\infty} c_n z^n\right)$$
$$= \sum_{n=0}^{\infty}\left(\sum_{k=0}^{n} b_{n-k}c_k\right) z^n,$$

   i.e. $c_0 = a_0/b_0$, $c_1 = (a_1 - b_1 c_0)/b_0$,
   $c_2 = (a_2 - b_1 c_1 - b_2 c_0)/b_0$, etc.

   Thus quotients of per se analytic functions are analytic wherever they are defined.

   Finally, there is a "chain rule" for analytic functions: If f is analytic at z_0 and g is analytic at $w_0 = f(z_0)$ then the composition $g \circ f : z \mapsto g\big(f(z)\big)$ is analytic at z_0. Thus compositions of per se analytic functions are analytic as well.

Properties cont'd

6. *Algebraic Operations on Power Series cont'd*

The power series representation of $g \circ f$ at z_0 can be computed from those of f and g, $f(z) = \sum_{n=0}^{\infty} a_n (z - z_0)^n$ and $g(w) = \sum_{n=0}^{\infty} b_n (w - w_0)^n$, as follows: In

$$g\big(f(z)\big) = \sum_{n=0}^{\infty} b_n \big(f(z) - w_0\big)^n = \sum_{n=0}^{\infty} b_n \left(\sum_{k=1}^{\infty} a_k (z - z_0)^k \right)^n$$

expand for each n the power $\left(\sum_{k=1}^{\infty} a_k (z - z_0)^k\right)^n$ into a power series $\sum_{l=n}^{\infty} A_{nl} (z - z_0)^l$, and rearrange the resulting double series $\sum_{n,l=0}^{\infty} b_n A_{nl} (z - z_0)^l$ into a power series $\sum_{l=0}^{\infty} c_l (z - z_0)^l$, i.e., $c_l = \sum_{n=0}^{l} b_n A_{nl}$. The required absolute convergence of the double series can be shown to hold in a neighboorhod of z_0.

It should be noted that the computation of the coefficients c_l in $g\big(f(z)\big) = \sum_{n=0}^{\infty} c_l (z - z_0)^l$ doesn't require taking any limits but uses only the arithmetic of the base field (which is $\mathbb{Q}$, $\mathbb{R}$, or $\mathbb{C}$ in our case).

Example (The Fibonacci generating function)

Consider the rational function

$$f(z) = \frac{1}{1-z-z^2}, \quad z \in \mathbb{C} \setminus \left\{ \tfrac{-1\pm\sqrt{5}}{2} \right\}.$$

f is analytic at $z_0 = 0$ and hence has a power series expansion $f(z) = \sum_{n=0}^{\infty} f_n z^n$ for small z. Equating coefficents in

$$\begin{aligned} 1 &= \left(f_0 + f_1 z + f_2 z^2 + f_3 z^3 + \cdots\right)(1 - z - z^2) \\ &= f_0 + (f_1 - f_0)z + (f_2 - f_1 - f_0)z^2 + (f_3 - f_2 - f_1)z^3 + \cdots, \end{aligned}$$

we see that $f_0 = f_1 = 1$, $f_n = f_{n-1} + f_{n-2}$ for $n \geq 2$, i.e., f_n is the n-th Fibonacci number (with the convention that $f_0 = f_1 = 1$).

The closed form of f_n can be obtained by $\frac{1}{1-z-z^2}$ into a power series:

$$\begin{aligned} \sum_{n=0}^{\infty} f_n z^n &= \frac{1}{(1-\alpha z)(1-\beta z)} = \frac{1}{\alpha-\beta}\left(\frac{\alpha}{1-\alpha z} - \frac{\beta}{1-\beta z}\right) \\ &= \sum_{n=0}^{\infty} \frac{\alpha^{n+1} - \beta^{n+1}}{\alpha - \beta} z^n \quad \text{with} \quad \alpha = \frac{1+\sqrt{5}}{2},\ \beta = \frac{1-\sqrt{5}}{2}. \end{aligned}$$

Example (cont'd)

As a simple example for the composition of power series we consider the series expansion

$$f(z) = \frac{1}{1-z-z^2} = \sum_{n=0}^{\infty}(z+z^2)^n = \sum_{n=0}^{\infty}\sum_{k=0}^{\infty}\binom{n}{k}z^{n+k}$$
$$= \sum_{k=0}^{\infty}\sum_{n=0}^{\infty}\binom{n}{k}z^{n+k} = \sum_{k=0}^{\infty}\sum_{n=k}^{\infty}\binom{n-k}{k}z^{n}$$
$$= \sum_{n=0}^{\infty}\left(\sum_{k=0}^{n}\binom{n-k}{k}\right)z^n,$$

which is valid for $|z|+|z|^2<1$, i.e, $|z|<\left(\sqrt{5}-1\right)/2$.

Equating coefficients of z^n shows

$$f_n = \sum_{k=0}^{\lfloor n/2\rfloor}\binom{n-k}{k} = \binom{n}{0}+\binom{n-1}{1}+\binom{n-2}{2}+\cdots,$$

evaluating the SW-NO diagonal sums of Pascal's Triangle $\left(\binom{n}{k}\right)_{n,k=0}^{\infty}$.

Example (cont'd)

But the example shows more: Since

$$(z+z^2)^n = \sum_{(i_1,i_2,\ldots,i_n)\in\{1,2\}^n} z^{i_1+i_2+\cdots+i_n},$$

the coefficient of z^n in $\sum_{n=0}^{\infty}(z+z^2)^n$ is equal to the number of ordered partitions of n into one's and two's. The power series identity $\sum_{n=0}^{\infty}(z+z^2)^n = \frac{1}{1-z-z^2}$ shows that these numbers are just the Fibonacci numbers. For example, we have

$$\begin{aligned}
1 &= 1,\\
2 &= 2 = 1+1,\\
3 &= 2+1 = 1+2 = 1+1+1,\\
4 &= 2+2 = 2+1+1 = 1+2+1 = 1+1+2 = 1+1+1+1,\\
5 &= 2+2+1 = 2+1+2 = 1+2+2\\
&= 2+1+1+1 = 1+2+1+1 = 1+1+2+1 = 1+1+1+2\\
&= 1+1+1+1+1.
\end{aligned}$$

Compare this with

n	0	1	2	3	4	5
f_n	1	1	2	3	5	8

Example (EULER Numbers)

The *Euler numbers* (or *secant numbers*) $\mathrm{E}_0, \mathrm{E}_1, \mathrm{E}_2, \ldots$ are defined in the such a way that the corresponding exponential generating function is $1/\cos x = \sec x$:

$$\frac{1}{\cos x} = \sum_{n=0}^{\infty} \frac{\mathrm{E}_n}{n!} x^n.$$

Since $\cos x = \cos(-x)$, we must have $\mathrm{E}_1 = \mathrm{E}_3 = \mathrm{E}_5 = \cdots = 0$ (equate coefficients!), so that we can write the defining equation as

$$\left(\sum_{n=0}^{\infty} \frac{\mathrm{E}_{2n}}{(2n)!} x^{2n}\right) \left(\sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} x^{2n}\right) = 1$$

Expanding the product and equating coefficients gives the recurrence relation

$$\mathrm{E}_0 = 1, \quad \sum_{k=0}^{n} \mathrm{E}_{2k} \frac{(-1)^{n-k}}{(2k)!(2n-2k)!} = 0 \quad \text{for } n \geq 1.$$

Multiplying by $(2n)!$ puts it into the more convenient integral form

Example (cont'd)

$\sum_{k=0}^{n}(-1)^{n-k}\binom{2n}{2k}\mathrm{E}_{2k} = 0$, or

$$\mathrm{E}_{2n} = \binom{2n}{2}\mathrm{E}_{2n-2} - \binom{2n}{4}\mathrm{E}_{2n-4} + \binom{2n}{6}\mathrm{E}_{2n-6} \mp \cdots$$

The first few even Euler numbers are $\mathrm{E}_0 = 1$,

$$\mathrm{E}_2 = \binom{2}{2}\mathrm{E}_0 = 1,$$

$$\mathrm{E}_4 = \binom{4}{2}\mathrm{E}_2 - \binom{4}{4}\mathrm{E}_0 = 5,$$

$$\mathrm{E}_6 = \binom{6}{2}\mathrm{E}_4 - \binom{6}{4}\mathrm{E}_2 + \binom{6}{6}\mathrm{E}_0 = 15\cdot 5 - 15\cdot 1 + 1 = 61,$$

$$\begin{aligned}\mathrm{E}_8 &= \binom{8}{2}\mathrm{E}_6 - \binom{8}{4}\mathrm{E}_4 + \binom{8}{6}\mathrm{E}_2 - \binom{8}{8}\mathrm{E}_0\\ &= 28\cdot 61 - 70\cdot 5 + 28\cdot 1 - 1 = 1385,\end{aligned}$$

$$\mathrm{E}_{10} = 50\,521,$$

$$\mathrm{E}_{12} = 2\,702\,765,$$

$$\mathrm{E}_{14} = 199\,360\,981.$$

Example (BERNOULLI Numbers)

The *Bernoulli numbers* $\mathrm{B}_0, \mathrm{B}_1, \mathrm{B}_2, \ldots$ are defined in the such a way that the corresponding exponential generating function is $x/(\mathrm{e}^x - 1)$:

$$\frac{x}{\mathrm{e}^x - 1} = \frac{1}{1 + \frac{x}{2!} + \frac{x^2}{3!} + \cdots} = \sum_{n=0}^{\infty} \frac{\mathrm{B}_n}{n!} x^n.$$

This gives the recurrence relation

$$\mathrm{B}_0 = 1, \quad \sum_{k=0}^{n} \frac{\mathrm{B}_k}{k!(n-k+1)!} = 0 \quad \text{for } n \geq 1,$$

which can also be written as

$$\mathrm{B}_0 = 1, \quad \sum_{k=0}^{n} \binom{n+1}{k} \mathrm{B}_k = 0 \quad \text{for } n \geq 1.$$

The first few Bernoulli numbers are:

n	0	1	2	3	4	5	6	7	8	9	10
B_n	1	$-\frac{1}{2}$	$\frac{1}{6}$	0	$-\frac{1}{30}$	0	$\frac{1}{42}$	0	$-\frac{1}{30}$	0	$\frac{5}{66}$

Example (cont'd)

It is no coincidence that the odd Bernoulli numbers $\mathrm{B}_3, \mathrm{B}_5, \mathrm{B}_7, \ldots$ are zero:

$$\sum_{n\neq 1} \frac{\mathrm{B}_n}{n!}\, x^n = \frac{x}{\mathrm{e}^x - 1} + \frac{x}{2} = \frac{2x + x(\mathrm{e}^x - 1)}{2(\mathrm{e}^x - 1)} = \frac{x(\mathrm{e}^x + 1)}{2(\mathrm{e}^x - 1)}$$

$$= \frac{x}{2}\,\frac{\mathrm{e}^{x/2} + \mathrm{e}^{-x/2}}{\mathrm{e}^{x/2} - \mathrm{e}^{-x/2}} = \frac{x}{2}\,\frac{\cosh(x/2)}{\sinh(x/2)} = \frac{x}{2}\coth\frac{x}{2}$$

is an even function of x, and hence has all odd coefficients equal to zero.

As a by-product, we obtain the power series expansions at $x_0 = 0$ of $x\coth x$, $x\cot x = \mathrm{i}x\coth(\mathrm{i}x)$, $\tan x = \cot x - 2\cot(2x)$, viz.,

$$x\coth x = \sum_{n=0}^{\infty} \frac{\mathrm{B}_{2n}\, 2^{2n}}{(2n)!}\, x^{2n}, \qquad x\cot x = \sum_{n=0}^{\infty} \frac{(-1)^n\, \mathrm{B}_{2n}\, 2^{2n}}{(2n)!}\, x^{2n},$$

$$\tan x = \sum_{n=1}^{\infty} \frac{(-1)^{n-1}\, \mathrm{B}_{2n}\, 2^{2n}(2^{2n}-1)}{(2n)!}\, x^{2n-1} = x + \tfrac{1}{3}\,x^3 + \tfrac{2}{15}\,x^5 + \tfrac{17}{315}\,x^7 + \cdots.$$

Properties cont'd

7. *Zeros and Poles*

In general a quotient $h = f/g$ of nonzero functions f, g that are analytic at z_0 is only defined in a punctured disk $0 < |z - z_0| < \delta$. We can write

$$f(z) = (z - z_0)^{m_1} f_1(z), \quad g(z) = (z - z_0)^{m_2} g_1(z)$$

with $m_1, m_2 \in \mathbb{N}$, f_1, g_1 analytic at z_0 and $f_1(z_0) \neq 0$, $g_1(z_0) \neq 0$. (The exponents m_1, m_2 are those of the smallest powers $(z - z_0)^n$ appearing with a nonzero coefficient in the power series representation of f and g at z_0; cf. Property 5).

$\Longrightarrow$ h has the representation

$$h(z) = (z - z_0)^m h_1(z)$$

with $m = m_1 - m_2 \in \mathbb{Z}$, $h_1 = f_1/g_1$ analytic at z_0, and $h_1(z_0) = f_1(z_0)/g_1(z_0) \neq 0$.

In this case we say that h has *order* m at z_0. If $m > 0$, we call z_0 a *zero of h of order m*; if $m < 0$, we call z_0 a *pole of h of order* $-m$ (note that $-m > 0$ in this case).

Properties cont'd

7. *Zeros and Poles cont'd*

Thus h has a pole of order m at z_0 iff $h_1(z) = (z - z_0)^m h(z)$ is analytic at z_0 and $h_1(z_0) \neq 0$; equivalently, the "power series expansion" of h at z_0 (valid for $0 < |z - z_0| < \delta$) starts with the negative power $(z - z_0)^{-m}$:

$$h(z) = \sum_{n=-m}^{\infty} c_n (z - z_0)^n, \quad c_{-m} \neq 0.$$

The concept of "pole" applies to any analytic function defined on a punctured disk, e.g., $z \mapsto 1/(e^z - 1)$ has a pole of order 1 in $z_0 = 0$.

In Complex Anaysis it is shown that a bounded analytic function h on a punctured disk $0 < |z - z_0| < \delta$ can be extended to an analytic function on the whole disk (in particular $\lim_{z \to z_0} h(z)$ exists in this case). This gives a characterization of poles of h in terms of boundedness of $z \mapsto (z - z_0)^m h(z)$ and implies that other types of isolated singularities (called *essential* singularities) must be tied to high local fluctuations of the values of the function.

Properties cont'd

We close with two properties whose proofs require the more advanced machinery of Complex Analysis.

8. We have seen that per se analytic functions $f\colon D \to \mathbb{C}$, $D \subseteq \mathbb{C}$, are differentiable, i.e., $\lim\limits_{h\to 0,\, h\in\mathbb{C}} \frac{1}{h}\big(f(z+h)-f(z)\big)$ exists for all $z \in D$, and that the derivative $f'(z)$ is again analytic. Conversely, it can be shown that differentiable functions are analytic (and thus have derivatives of all orders). This is in sharp contrast with the real case: A differentiable function $f\colon D \to \mathbb{R}$, $D \subseteq \mathbb{R}$, can have a derivative which is not differentiable, and f can be C^∞ without being analytic; cf. examples.

   Differentiable per se (in the above sense) complex functions are also called *holomorphic*.

Properties cont'd

9. It can be shown that the power series expansion $f(z) = \sum_{n=0}^{\infty} a_n (z - z_0)^n$ of a holomorphic function is valid in the largest open circle around z_0 on which f is defined. In other words, if f is defined on the whole complex plane (a so-called *entire* function) then every power series representing f has radius of convergence $\rho = \infty$, and if $\rho < \infty$ then the circle $|z - z_0| = \rho$ must contain a singularity of f (i.e., a point where f is not defined).

   The proof in the general case requires Cauchy's Integral Formula from Complex Analysis, but in the special case of a rational function $f = P/Q$ (P, Q polynomials with $\gcd(P, Q) = 1$) we can see it rather quickly using the partial fractions decomposition.

   Suppose $Q(X) = \prod_{i=1}^{r} (X - \lambda_i)^{m_i}$ is the prime factorization of $Q(X)$ in $\mathbb{C}[X]$ and $0 < |\lambda_1| \leq |\lambda_2| \leq \cdots \leq |\lambda_r|$. Then λ_1 is a singularity of f closest to the origin.
   The partial fractions decomposition of f has the form

$$f(z) = R(z) + \sum_{i=1}^{r} \sum_{s=1}^{m_i} \frac{c_{is}}{(z - \lambda_i)^s} \text{ with } R \in \mathbb{C}[X],\ c_{is} \in \mathbb{C},\ c_{i,m_i} \neq 0.$$

Properties cont'd

9. (continued)
The non-polynomial summands can be expanded into a power series using the generalized binomial theorem:

$$\frac{1}{(z-\lambda_i)^s} = \frac{(-1)^s\lambda_i^{-s}}{(1-\lambda_i^{-1}z)^s} = (-1)^s\lambda_i^{-s}\sum_{n=0}^{\infty}\binom{n+s-1}{s-1}\lambda_i^{-n}z^n.$$

The series converges precisely for $\left|\lambda_i^{-1}z\right| < 1$, i.e., for $|z| < |\lambda_i|$.
$\Longrightarrow$ For $|z| < |\lambda_1|$ all such expansions converge, showing that f has a power series expansion $f(z) = \sum_{n=0}^{\infty} a_n z^n$ in the circle $|z| < |\lambda_1|$.
The radius of convergence ρ of $\sum_{n=0}^{\infty} a_n z^n$ cannot be larger than $|\lambda_1|$, because $|z| < \rho$ cannot include a singularity of f.
$\Longrightarrow \rho = |\lambda_1|$
Finally, the change of variables $w = z - z_0$, which transforms $f(z)$ into another rational function, shows that the preceding statement holds for the power series expansion of f at an arbitrary point $z_0 \notin \{\lambda_1, \ldots, \lambda_r\}$.

Example (Geometric series)

The function $f\colon \mathbb{C}\setminus\{1\}\to\mathbb{C}$, $z\mapsto \frac{1}{1-z}$ is holomorphic with $f'(z)=\frac{1}{(1-z)^2}$. At $z_0=0$ it has the well-known series representation

$$1+z+z^2+\cdots=\sum_{n=0}^{\infty} z^n=\frac{1}{1-z}.$$

The radius of convergence of the geometric series is $\rho=1$.
(It cannot be larger since on the circle $|z|=1$ there is a singularity of f.)
We can expand f into a power series at any point $a\in\mathbb{C}\setminus\{1\}$ by the following computational trick. (There is no need to compute the derivatives $f^{(n)}(a)$.)

$$f(z)=\frac{1}{1-z}=\frac{1}{1-a-(z-a)}=\frac{\frac{1}{1-a}}{1-\frac{z-a}{1-a}}$$
$$=\sum_{n=0}^{\infty}\frac{(z-a)^n}{(1-a)^{n+1}}.$$

Example (Geometric series cont'd)

The radius of convergence of the new power series, which except for the factor $(1-a)^{-1}$ is a geometric series as well and converges for $|z-a| < |1-a|$, is $\rho' = |1-a|$ (the distance from a to the singularity of f, as predicted by Property 9).

The derivatives of f can now be read off from the power series expansion:

$$f^{(n)}(a) = \frac{n!}{(1-a)^{n+1}}, \quad a \in \mathbb{C} \setminus \{1\}.$$

Of course you can also prove by induction that $f^{(n)}(z) = \frac{n!}{(1-z)^{n+1}}$.

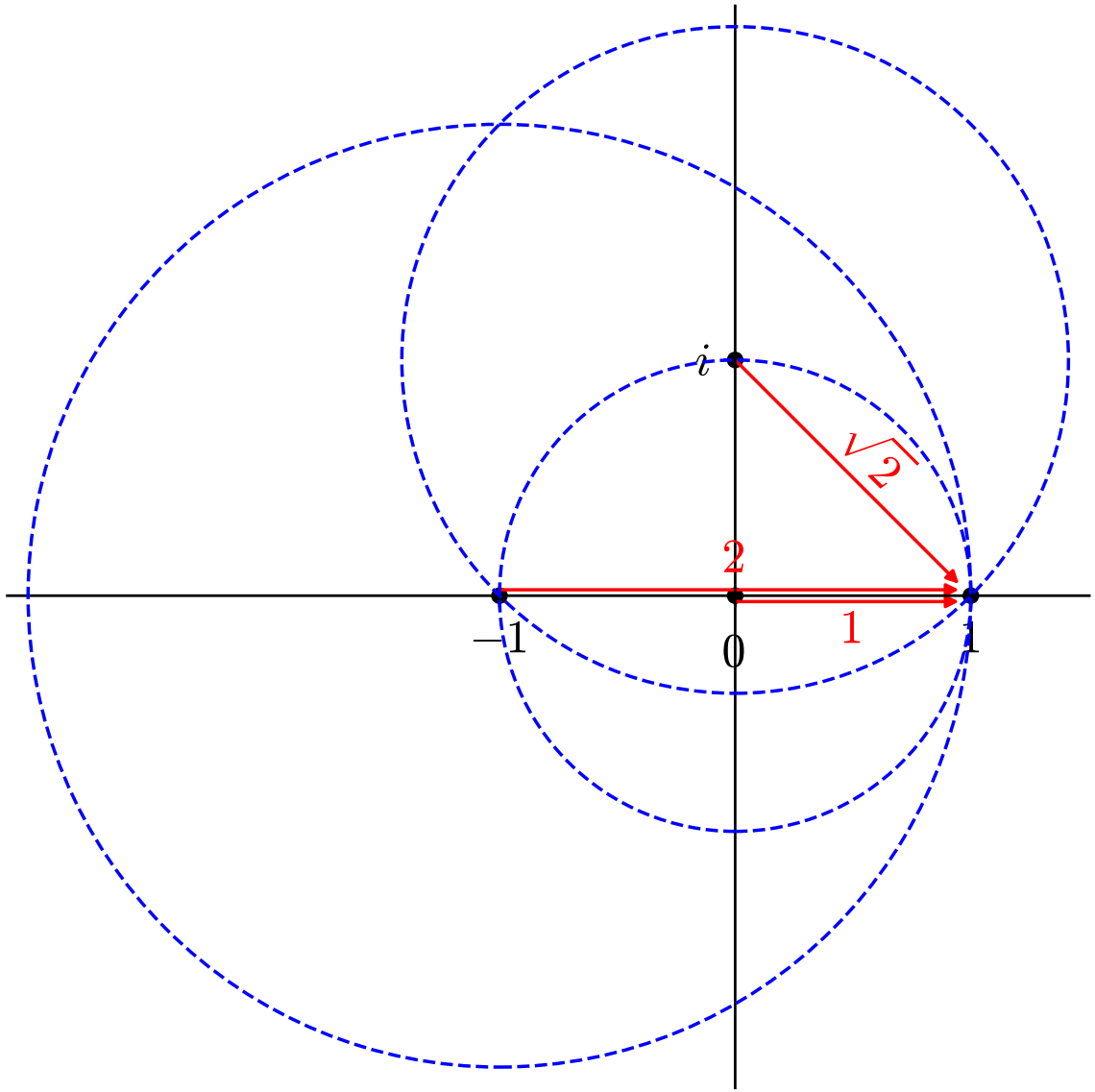


Figure: Disks of convergence of the Taylor series of $z \mapsto \frac{1}{1-z}$ at $a \in \{0, -1, \mathrm{i}\}$

Example

The function $g\colon \mathbb{C} \setminus \{\pm \mathrm{i}\} \to \mathbb{C}$, $z \mapsto \frac{1}{z^2+1}$ is holomorphic with $g'(z) = -\frac{2z}{(z^2+1)^2}$. At $z_0 = 0$ it has the series representation

$$1 - z^2 + z^4 - z^6 \pm \cdots = \frac{1}{1+z^2}, \quad |z| < 1,$$

which is also an instance of the geometric series.

Restricting g to $\mathbb{R}$ gives the function $\mathbb{R} \to \mathbb{R}$, $x \mapsto \frac{1}{x^2+1}$, which is real analytic everywhere. But unlike the exponential series, its Taylor series $1 - x^2 + x^4 - x^6 \pm \cdots$ at $x_0 = 0$ doesn't converge on all of $\mathbb{R}$, but only for $|x| < 1$.

The same is true for $\mathbb{R} \to \mathbb{R}$, $x \mapsto \arctan x$, which has Taylor series $x - x^3/3 + x^5/5 - x^7/7 \pm \cdots$ at $x_0 = 0$ and represents an antiderivative of $x \mapsto \frac{1}{x^2+1}$.

This example vividly explains why we need to look at the complex extensions of real analytic functions to determine their more subtle properties.

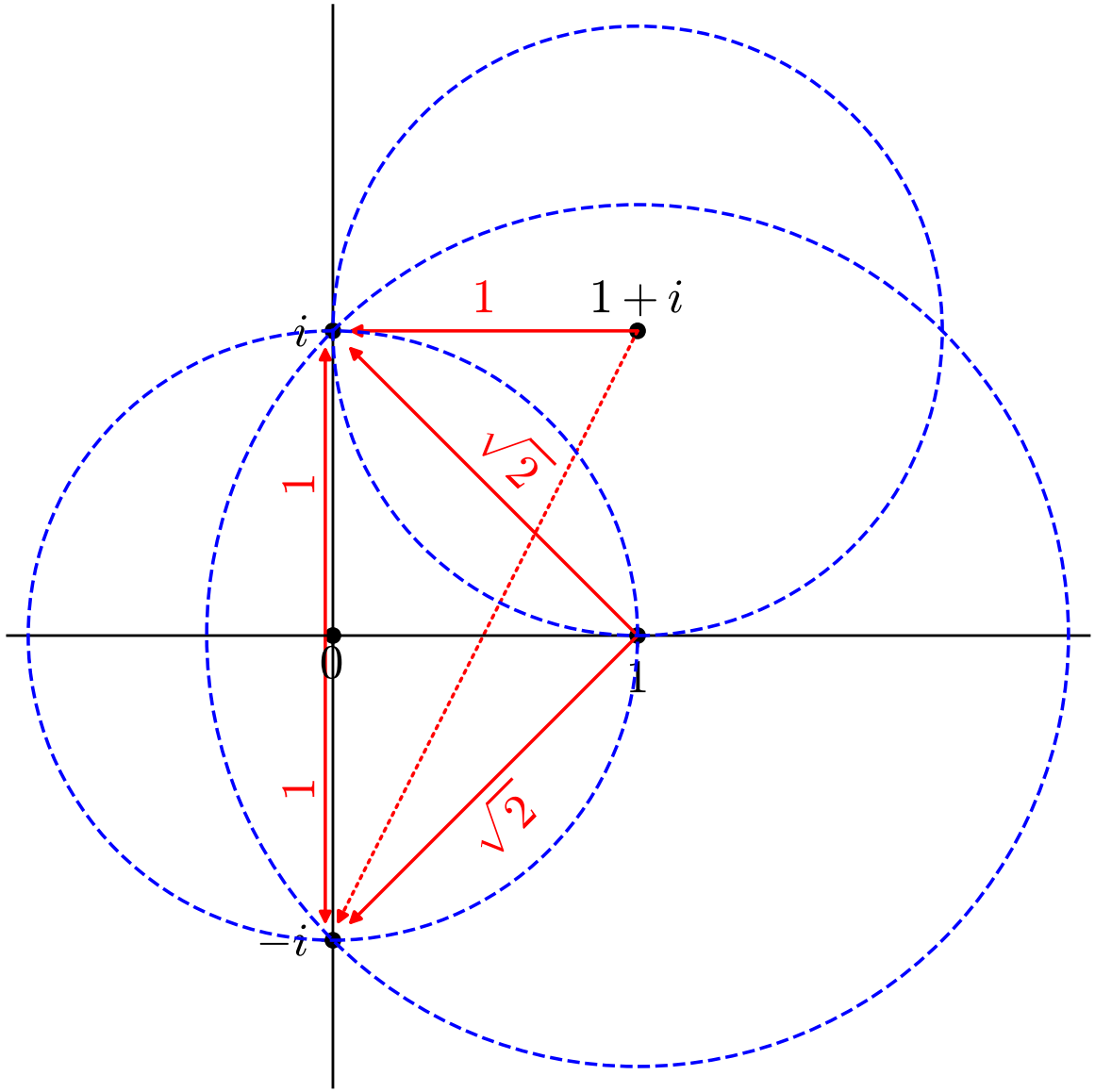


Figure: Disks of convergence of the Taylor series of $z \mapsto \frac{1}{z^2+1}$ at $a \in \{0, 1, 1+\mathrm{i}\}$. For $a = 1+\mathrm{i}$ the nearest singularity is i, and hence $\rho = 1$.

Exercise

Compute the Taylor series of $z \mapsto \frac{1}{z^2+1}$ at $a = 1$ and $a = 1 + \mathrm{i}$.

Hint: Proceed as for $z \mapsto \frac{1}{1-z}$ and then use partial fractions.

Ordinary and Singular Points

We consider implicit 2nd-order homogeneous linear time-dependent ODE's with analytic coefficients, i.e.,

$$P(x)y'' + Q(x)y' + R(x)y = 0 \tag{1}$$

with real analytic functions $P \neq 0$, Q, and R defined on some common interval I.

At points $x_0 \in I$ with $P(x_0) \neq 0$ we can put (1) into the explicit form

$$y'' + p(x)y' + q(x)y = 0, \quad p(x) = \frac{Q(x)}{P(x)}, \; q(x) = \frac{R(x)}{P(x)}, \tag{2}$$

and we know from the discussion of quotients of analytic functions (see Property 6) that p and q are analytic at x_0.

If $P(x_0) = 0$, let $P(x) = P_1(x)(x - x_0)^m$ with P_1 analytic at x_0 and $P_1(x_0) \neq 0$. (The integer $m \geq 1$ is the multiplicity of x_0 as a zero of P or, equivalently, the smallest index of a nonzero coefficient in the power series expansion of P at x_0; it exists in view of $P \neq 0$.) We can assume that one of $Q(x_0)$, $R(x_0)$ is $\neq 0$, since otherwise we can divide (1) by $x - x_0$, which doesn't change solutions (why?) and reduces the multiplicity of x_0 as a zero of P by one.

Then (1) becomes

$$(x-x_0)^m y''+p_1(x)y'+q_1(x)y=0, \quad p_1(x)=\frac{Q(x)}{P_1(x)},\ q_1(x)=\frac{R(x)}{P_1(x)}, \tag{3}$$

and again p_1, q_1 are analytic at x_0.

Finally we can put (3) formally into an "explicit form", which is not defined for $x = x_0$, if we admit the coefficients of y' and y to have poles at x_0:

$$y'' + \underbrace{\frac{p_1(x)}{(x-x_0)^m}}_{p(x)}\, y' + \underbrace{\frac{q_1(x)}{(x-x_0)^m}}_{q(x)}\, y = 0. \tag{4}$$

Since one of $p_1(x_0)$, $q_1(x_0)$ is nonzero, either $p(x)$ or $q(x)$ (or both) have a pole of exact order m at x_0.

Definition

1. $x_0 \in I$ is called a *singular point* of (1) if $P(x_0) = 0$, and an *ordinary point* otherwise.
2. A singular point $x_0 \in I$ of (1) is called a *regular singular point* if $\lim_{x\to x_0}(x-x_0)p(x)$ and $\lim_{x\to x_0}(x-x_0)^2 q(x)$ exist in $\mathbb{R}$; equivalently, $m \in \{1,2\}$ in (4) and $p_1(x_0) = 0$ if $m = 2$.

Notes on the definition

1. x_0 is a singular point iff at least one of $p(x)$, $q(x)$ has a pole at x_0.
2. A singular point x_0 is a regular singular point iff the order of the pole(s) in Note 1 is ≤ 1 for $p(x)$ and ≤ 2 for $q(x)$.
3. The condition for a regular singular point may also be rephrased as: $f(x) := (x - x_0)p(x)$ and $g(x) := (x - x_0)^2 q(x)$ can be made analytic at x_0 by setting $f(x_0) = \lim_{x \to x_0}(x - x_0)p(x)$, $g(x_0) = \lim_{x \to x_0}(x - x_0)^2 q(x)$. (For an ordinary point x_0 the functions f, g are trivially analytic at x_0.)

From now on we will assume w.l.o.g. that $x_0 = 0$.

If $x_0 = 0$ is an ordinary point or a regular singular point of the ODE (1) then (4) can be rewritten as

$$y'' + \underbrace{\left(\frac{p_0}{x} + p_1 + p_2 x + p_3 x^2 + \cdots\right)}_{p(x)} y' + \underbrace{\left(\frac{q_0}{x^2} + \frac{q_1}{x} + q_2 + q_3 x + q_4 x^2 + \cdots\right)}_{q(x)} y = 0.$$

The case of an ordinary point corresponds to $p_0 = q_0 = q_1 = 0$.

After multiplication by x^2 this takes the more convenient "analytic" form

$$x^2y''+x(p_0+p_1x+p_2x^2+\dots)y'+(q_0+q_1x+q_2x^2+\dots)y=0, \quad (5)$$

which is of course still equivalent to (1).
Here $f(x)=p_0+p_1x+p_2x^2+\cdots$, $g(x)=q_0+q_1x+q_2x^2+\cdots$ are the analytic functions defined in the previous Note 3.

Caution: Don't confuse $f(x)$, $g(x)$ with $p(x)$, $q(x)$, which may contain negative powers of x and whose cofficients are indexed in a non-standard way (cf. previous slide). Also don't confuse the real numbers p_1, q_1 in (5) with the analytic functions $p_1(x)$, $q_1(x)$ appearing in (3).

Example

The Euler equation $x^2y''+\alpha xy'+\beta y=0$, $(\alpha,\beta)\neq(0,0)$, has a regular singular point at $x=0$. The corresponding analytic functions are the constant functions $f(x)=\alpha$, $g(x)=\beta$.

We also see that truncating the coefficient functions of y' and y in the general form (5) after the first term yields an Euler equation, viz. $x^2y''+p_0xy'+q_0y=0$. This Euler equation will play an important role when solving (5) in the case of a regular singular point. But first we consider the case of an ordinary point.

The Analytic Case

Theorem

Suppose that x_0 is an ordinary point of

$$P(x)y'' + Q(x)y' + R(x)y = 0, \tag{1}$$

and that the power series representing $p(x) = Q(x)/P(x)$, $q(x) = R(x)/P(x)$ at x_0 converge for $|x - x_0| < \rho$. Then for any pair $(a_0, a_1) \in \mathbb{R}^2$ (or $\mathbb{C}^2$) there exists an analytic solution of (1) of the form $y(x) = \sum_{n=0}^{\infty} a_n(x - x_0)^n$ for $|x - x_0| < \rho$.

Notes

- Since $y(x_0) = a_0$ and $y'(x_0) = a_1$, this says in particular that every IVP associated with (1) locally at x_0 has an analytic solution.
- The best choice of ρ in the theorem is the minimum of the radii of convergence of the power series representing $p(x)$ and $q(x)$ at x_0, and for this choice the theorem says that the radius of convergence of $\sum_{n=0}^{\infty} a_n(x - x_0)^n$ is $\geq \rho$.

Proof of the theorem.

We assume w.l.o.g. $x_0 = 0$ and use the equivalent form (5) of (1) with $p_0 = q_0 = q_1 = 0$ for the proof. Plugging

$$x^2 y'' = x^2 \sum_{n=2}^{\infty} n(n-1)a_n x^{n-2} = \sum_{n=2}^{\infty} n(n-1)a_n x^n,$$

$$xy' = x \sum_{n=1}^{\infty} n a_n x^{n-1} = \sum_{n=1}^{\infty} n a_n x^n$$

into (5) gives

$$\sum_{n=2}^{\infty} n(n-1)a_n x^n + \left(\sum_{n=1}^{\infty} n a_n x^n\right)\left(\sum_{n=1}^{\infty} p_n x^n\right) + \\ + \left(\sum_{n=0}^{\infty} a_n x^n\right)\left(\sum_{n=2}^{\infty} q_n x^n\right) = 0,$$

or, equivalently,

$$n(n-1)a_n + \sum_{k=1}^{n-1} k a_k p_{n-k} + \sum_{k=0}^{n-2} a_k q_{n-k} = 0 \quad \text{for } n = 2, 3, 4, \ldots$$

Proof cont'd.

It is clear that this recurrence relation for the sequence $(a_0, a_1, a_2, \dots)$ has a unique solution for any given a_0, a_1.

It remains to show that the so-defined power series $\sum_{n=0}^{\infty} a_n x^n$ converges for $|x| < \rho$. For this we use the following

Lemma

The radius of convergence of any power series $\sum_{n=0}^{\infty} b_n (z - z_0)^n$ is given by

$$R := \sup\{r \in \mathbb{R};\ \textit{the sequence } (|b_n|\, r^n) \textit{ is bounded}\}.$$

Proof of the lemma.

We show that $\sum_{n=0}^{\infty} b_n (z - z_0)^n$ converges for $|z - z_0| < R$ and diverges for $|z - z_0| > R$. (Notably this also proves the existence of the radius of convergence.) W.l.o.g. we assume $z_0 = 0$.

$\underline{|z| < R}$: There exists $r > |z|$ such that $(|b_n|\, r^n)$ is bounded, say $\underline{|b_n|\, r^n \le M}$ for all n. $\Longrightarrow |b_n z^n| = |b_n|\, r^n (|z|/r)^n \le M q^n$ with $q := |z|/r < 1$. Since $\sum M q^n$ converges, we can apply the comparison test and conclude that $\sum b_n z^n$ converges.

$\underline{|z| > R}$: If $\sum b_n z^n$ converges then $b_n z^n \to 0$ and $|b_n z^n| = |b_n|\, |z|^n$ is bounded. This contradicts the definition of R. □

Remark

The sequence $(|b_n|\,R^n)$ can be bounded or unbounded. For example, $\sum_{n=1}^{\infty} nz^n$ has $R = 1$ and $|b_n|\,R^n = n \to \infty$, whereas $\sum_{n=1}^{\infty}(1/n)z^n$ has $R = 1$ and $|b_n|\,R^n = 1/n \to 0$.

Proof of the theorem cont'd.

Let $r < \rho$ be given. Then, by the lemma, the sequences $(|p_n|\,r^n)$ and $(|q_n|\,r^n)$ are bounded, say by M.
Using the recurrence relation for a_n, we now try to bound $(|a_n|\,r^n)$. Guided by one of our introductory examples, we find

$$n(n-1)\,|a_n|\,r^n = \left|\sum_{k=1}^{n-1} k(a_k r^k)(p_{n-k} r^{n-k}) + \sum_{k=0}^{n-2}(a_k r^k)(q_{n-k} r^{n-k})\right|$$

$$\leq M\left(\sum_{k=1}^{n-1} k\,|a_k|\,r^k + \sum_{k=0}^{n-2}|a_k|\,r^k\right)$$

$$\Longrightarrow\quad |a_n|\,r^n \leq \sum_{k=0}^{n-1}\frac{M(k+1)}{n(n-1)}\,|a_k|\,r^k \quad (n \geq 2)$$

This is a recursive bound for the sequence $(|a_n|\,r^n)$, which unfortunately is not as simple to handle as the former one.

Proof of the theorem cont'd.

Now we proceed as follows: We define an auxiliary sequence (u_n) by $u_0 = |a_0|$, $u_1 = |a_1|\, r$, and

$$u_n = \sum_{k=0}^{n-1} \frac{M(k+1)}{n(n-1)}\, u_k \quad \text{for } n \geq 2.$$

One can show by induction that

$$|a_n|\, r^n \leq u_n \quad \text{for all } n \quad \text{and} \quad \lim_{n\to\infty} \frac{u_{n+1}}{u_n} = 1;$$

cf. exercises. It follows that for any positive $r_1 < r$ we have

$$|a_n|\, r_1^n \leq u_n (r_1/r)^n \quad \text{with } q := r_1/r < 1,$$

and that the series $\sum_{n=0}^{\infty} u_n q^n$ converges (because $\sum_{n=0}^{\infty} u_n x^n$ has radius of convergence 1).

$\Longrightarrow$ We can apply the comparison test to conclude that $\sum_{n=0}^{\infty} |a_n|\, r_1^n$ converges as well (or that $(|a_n|\, r_1^n)$ is bounded, whatever you prefer!).

Finally, since r and r_1 were chosen arbitrarily subject only to $r_1 < r < \rho$, it is clear that $\sum_{n=0}^{\infty} |a_n|\, r_1^n$ converges for all $r_1 < \rho$, and hence $\sum_{n=0}^{\infty} a_n x^n$ converges for $|x| < \rho$. □

Remark

If y_1, y_2 are solutions of (1) satisfying the particular initial conditions $y_1(x_0) = 1$, $y_1'(x_0) = 0$ and $y_2(x_0) = 0$, $y_2'(x_0) = 1$ then the general solution of (1) is $y(x) = a_0 y_1(x) + a_1 y_2(x)$. In particular, y_1, y_2 form a fundamental system of solutions of (1).

Exercise

Suppose (α_n) and (u_n) are sequences of nonnegative real numbers satisfying

$$\alpha_n \le \sum_{k=0}^{n-1} \frac{M(k+1)}{n(n-1)} \alpha_k \quad (n \ge 2),$$

$$u_n = \sum_{k=0}^{n-1} \frac{M(k+1)}{n(n-1)} u_k \quad (n \ge 2),$$

$$u_0 = \alpha_0, \; u_1 = \alpha_1$$

for some constant $M > 0$.

a) Show $\alpha_n \le u_n$ for all n.

b) Show $\lim_{n\to\infty} \frac{u_{n+1}}{u_n} = 1$.
Hint: Express u_{n+1} in terms of u_n.

c) Is the sequence (u_n) (and hence (α_n) as well) necessarily bounded from above?

Example (Airy's Equation)

The ODE $y'' - xy = 0$ is known as *Airy's Equation*. The theorem predicts that its general solution is analytic everywhere.

Making the usual power series „Ansatz" $y(x) = \sum_{n=0}^{\infty} a_n x^n$ and substituting $y(x)$ into Airy's Equation, we obtain

$$y''(x) = \sum_{n=0}^{\infty} (n+2)(n+1) a_{n+2} x^n,$$

$$x\,y(x) = \sum_{n=0}^{\infty} a_n x^{n+1} = \sum_{n=1}^{\infty} a_{n-1} x^n$$

$$2a_2 + \sum_{n=1}^{\infty} \left((n+2)(n+1)a_{n+2} - a_{n-1}\right) x^n = 0$$

$$\Longrightarrow \quad a_2 = 0 \quad \text{and} \quad a_{n+2} = \frac{a_{n-1}}{(n+2)(n+1)} \quad \text{for } n = 1, 2, 3, \ldots$$

$$\Longrightarrow \quad a_{3n+2} = 0, \; a_{3n} = a_0 \prod_{k=1}^{n} \frac{1}{(3k)(3k-1)}, \; a_{3n+1} = a_1 \prod_{k=1}^{n} \frac{1}{(3k+1)(3k)}.$$

Example (cont'd)

$\Longrightarrow$ A fundamental system of solutions of Airy's Equation is

$$y_1(x) = 1 + \frac{x^3}{3\cdot 2} + \frac{x^6}{6\cdot 5\cdot 3\cdot 2} + \frac{x^9}{9\cdot 8\cdot 6\cdot 5\cdot 3\cdot 2} + \cdots,$$

$$y_2(x) = x + \frac{x^4}{4\cdot 3} + \frac{x^7}{7\cdot 6\cdot 4\cdot 3} + \frac{x^{10}}{10\cdot 9\cdot 7\cdot 6\cdot 4\cdot 3} + \cdots.$$

The radius of convergence of these power series is $\rho = \infty$, as you can check by applying the ratio test to the series with the gaps removed (or substitute $z = x^3$).

$\Longrightarrow$ The general solution of Airy's Equation, which is

$$\begin{aligned} y(x) &= a_0 y_1(x) + a_1 y_2(x) \\ &= a_0 + a_1 x + \frac{a_0}{3\cdot 2}x^3 + \frac{a_1}{4\cdot 3}x^4 + \frac{a_0}{6\cdot 5\cdot 3\cdot 2}x^6 + \frac{a_1}{7\cdot 6\cdot 4\cdot 3}x^7 + \cdots, \end{aligned}$$

also has radius of convergence $\rho = \infty$.

This direct proof of $\rho = \infty$ is instructive but not necessary, since the theorem implies $\rho = \infty$ (as for any explicit linear 2nd-order ODE with polynomial coefficients).

Example (Legendre's Equation)

The Legendre equation (or family of equations)

$$(1-x^2)y'' - 2x\,y' + n(n+1)y = 0 \qquad (\mathrm{Le}_n)$$

has regular singular points in $x_0 = \pm 1$, but an ordinary point in $x_0 = 0$.

$\Longrightarrow$ We can solve it in $(-1,1)$ with the usual power series „Ansatz" $y(x) = \sum_{k=0}^{\infty} a_k x^k$ ("k" is necessary, since (Le_n) is indexed by n). Since the coefficients are polynomials, it is better not to rewrite it in explicit form (which would produce the power series $\frac{2x}{1-x^2}$ and $\frac{n(n+1)}{1-x^2}$) but solve it directly.

We obtain

$$(1-x^2)\sum_{k=2}^{\infty} k(k-1)a_k x^{k-2} - 2x\sum_{k=1}^{\infty} k a_k x^{k-1} + n(n+1)\sum_{k=0}^{\infty} a_k x^k$$

$$= \sum_{k=0}^{\infty}(k+2)(k+1)a_{k+2}x^k - \sum_{k=2}^{\infty} k(k-1)a_k x^k - \sum_{k=1}^{\infty} 2k a_k\, x^k + \sum_{k=0}^{\infty} n(n+1)a_k x^k$$

$$= \sum_{k=0}^{\infty}\big((k+2)(k+1)a_{k+2} - (k^2+k-n^2-n)a_k\big)x^k = 0.$$

Example (cont'd)

$$\Longrightarrow\ a_{k+2} = \frac{k^2+k-n^2-n}{(k+2)(k+1)}\,a_k = \frac{(k-n)(k+n+1)}{(k+2)(k+1)}\,a_k \quad (k \in \mathbb{N});$$

$$\Longrightarrow\ a_{2m} = a_0 \frac{(-1)^m}{(2m)!} \prod_{i=0}^{m-1} \left[(n-2i)(n+2i+1)\right],$$

$$a_{2m+1} = a_1 \frac{(-1)^m}{(2m+1)!} \prod_{i=0}^{m-1} \left[(n-2i-1)(n+2i+2)\right]$$

A fundamental system of solutions of the Legendre equation is therefore

$$y_1(x) = \sum_{m=0}^{\infty} (-1)^m \frac{n(n-2)\cdots(n-2m+2)(n+1)(n+3)\cdots(n+2m-1)}{(2m)!}\,x^{2m},$$

$$y_2(x) = \sum_{m=0}^{\infty} (-1)^m \frac{(n-1)(n-3)\cdots(n-2m+1)(n+2)(n+4)\cdots(n+2m)}{(2m+1)!}\,x^{2m+1}.$$

Example (cont'd)

Notes

- One of the two fundamental solutions (y_1 if n is even, y_2 if n is odd) is a polynomial function of degree n and hence analytic everywhere. The other solution is analytic in $(-1,1)$, since $p(x) = \frac{-2x}{1-x^2}$, $q(x) = \frac{n(n+1)}{1-x^2}$ have power series expansions at $x_0 = 0$ with $\rho = 1$. In fact the ratio test applied to the non-polynomial solution shows that its radius of convergence is precisely 1.
- Since the polynomial solutions are scalar multiples of the polynomial fundamental solution, this must also hold for the n-th Legendre polynomial $\mathrm{P}_n(x)$. Hence up to a normalizing factor $\mathrm{P}_n(x)$ is equal to $y_1(x)$ if n is even and to $y_2(x)$ if n is odd. The normalizing factor can be determined from the leading coefficients of $\mathrm{P}_n(x)$ and the polynomial fundamental solution, which are

$$\frac{(2n)(2n-1)\cdots(n+1)}{2^n n!} = \frac{1}{2^n}\binom{2n}{n} \text{ resp. } (-1)^{\lfloor n/2 \rfloor} \frac{\prod_{k=n+1,\, k \text{ odd}}^{2n} k}{\prod_{k=1,\, k \text{ odd}}^{n} k}.$$

Example (cont'd)

Notes cont'd

- (cont'd)
  For the latter observe that for even n the signless coefficient of x^n n $y_1(x)$ is

$$\frac{n(n-2)\cdots 2(n+1)(n+3)\cdots(2n-1)}{n!} = \frac{(n+1)(n+3)\cdots(2n-1)}{(n-1)(n-3)\cdots 3\cdot 1},$$

  and similarly for odd n. It follows that

$$\mathrm{P}_n(x) = \frac{(-1)^{\lfloor n/2\rfloor}\prod_{k=n+1,\,k\text{ even}}^{2n} k}{2^n\prod_{k=1,\,k\text{ even}}^{n} k} \times \begin{cases} y_1(x) & \text{if } n \text{ is even,} \\ y_2(x) & \text{if } n \text{ is odd,} \end{cases}$$

  which together with the formulas for $y_1(x)$, $y_2(x)$ determines the coefficients of $\mathrm{P}_n(x)$.
  Alternatively (and less cumbersome), differentiate $(x^2-1)^n = \sum_{k=0}^{n}(-1)^k\binom{n}{k}x^{2n-2k}$ exactly n times to obtain the coefficients of $\mathrm{P}_n(x) = \frac{1}{2^n n!}\,\mathrm{D}^n\left[(x^2-1)^n\right]$ directly.

Inhomogeneous Equations

x_0 is said to be an ordinary point (resp., a regular singular point) of

$$P(x)y'' + Q(x)y' + R(x)y = S(x), \tag{2}$$

if the same is true of the associated homogeneous equation and $S(x)$ is analytic at x_0 as well.

Corollary

If x_0 is an ordinary point of (2) then solutions $y_p(x)$ of (2) are analytic in x_0, and the power series representation $y_p(x) = \sum_{n=0}^{\infty} a_n(x - x_0)^n$ is valid (at least) for $|x - x_0| < \rho$, where ρ denotes the minimum of the radii of convergence of the three power series representing $p(x) = Q(x)/P(x)$, $q(x) = R(x)/P(x)$, and $r(x) = S(x)/P(x)$ at x_0.

Proof of the corollary.

Let I be an open interval containing x_0 and not containing any zero of P. In terms of a fundamental system $y_1(x), y_2(x)$ of solutions of (1) on I, any particular solution of (2) on I has the form $y_p(x) = c_1(x)y_1(x) + c_2(x)y_2(x)$ with

$$c_1(x) = \gamma_1 + \int_{x_0}^{x} \frac{-y_2(t)r(t)}{\mathrm{W}(t)}\,\mathrm{d}t, \quad c_2(x) = \gamma_2 + \int_{x_0}^{x} \frac{y_1(t)r(t)}{\mathrm{W}(t)}\,\mathrm{d}t,$$

where γ_1, γ_2 are constants and $\mathrm{W}(x) = y_1(x)y_2'(x) - y_1'(x)y_2(x)$ is the Wronskian of $y_1(x), y_2(x)$; cf. the vectorial variation-of-parameters formula. By Abel's Theorem, the Wronskian has the form $\mathrm{W}(x) = \gamma \exp \int_{x_0}^{x} -p(t)\,\mathrm{d}t$ for some constant $\gamma \neq 0$. Since $p(x)$ is analytic in the disk $|z - x_0| < \rho$ and integration doesn't change the radius of convergence of a power series, the function $\mathrm{W}(x)^{-1} = \gamma^{-1} \exp \int_{x_0}^{x} p(t)\,\mathrm{d}t$ is analytic in $|z - x_0| < \rho$ as well. The same is true of $y_1(x)$, $y_2(x)$ (by the theorem) and $r(x)$ (by assumption). Since $y_p(x)$ is obtained from these functions by a finite number of additions, multiplications and integrations, it must also be analytic in $|z - x_0| < \rho$. □

Example

We solve the IVP

$$y'' + y = \frac{1}{1-x} \wedge y(0) = y'(0) = 0 \quad \text{on } (-1,1).$$

Notably, the machinery developed for higher-order linear ODE's with constant coefficients can't be used to solve this ODE (since $\frac{1}{1-x}$ is not an exponential polynomial), but order reduction and vectorial variation of parameters can be, of course.

Expanding the right-hand side into a geometric series and making the usual power series „Ansatz" turns the ODE into $y(x) = \sum_{n=0}^{\infty} a_n x^n$

$$\sum_{n=0}^{\infty} \left[(n+2)(n+1)a_{n+2} + a_n\right] x^n = \sum_{n=0}^{\infty} x^n.$$

$$\Longrightarrow \quad a_{n+2} = \frac{1-a_n}{(n+2)(n+1)} \quad \text{for } n = 0, 1, 2, \ldots$$

Example (cont'd)

Together with $a_0 = a_1 = 0$ this gives

$$y(x) = \frac{1}{2}x^2 + \frac{1}{6}x^3 + \frac{1}{24}x^4 + \frac{1}{24}x^5 + \frac{23}{720}x^6 + \frac{23}{1008}x^7 + \frac{697}{40320}x^8 + \\ + \frac{985}{72576}x^9 + \frac{39623}{3628800}x^{10} + \cdots$$

By the corollary, the series is guaranteed to converge for $|x| < 1$ and solves the ODE in $(-1, 1)$.

In fact, it is not difficult to see that

$$a_n \simeq \frac{1}{n(n-1)} \quad \text{for } n \to \infty,$$

i.e., $\lim_{n\to\infty}\big(n(n-1)a_n\big) = 1$.
This follows, e.g., from $\frac{n-3}{n(n-1)(n-2)} < a_n < \frac{1}{n(n-1)}$ for $n \geq 3$, which can be shown by induction.
$\Longrightarrow$ The radius of convergence of $\sum_{n=0}^{\infty} a_n x^n$ is exactly 1.

From the general theory we know, however, that $y(x)$ has a (unique) extension to $(-\infty, 1)$, which is analytic as well.

Example (cont'd)

Question: How to find the extension of $y(x)$ to $(-\infty, 1)$?

Answer: With power series we cannot do this in one fell swoop, but we can use a different center to enlarge the domain.

Let us consider this for $x_0 = -1$, i.e., we make the powers series „Ansatz“ $y(x) = \sum_{n=0}^{\infty} b_n (x+1)^n$.

$$\Longrightarrow \sum_{n=0}^{\infty} \left[(n+2)(n+1)b_{n+2} + b_n\right](x+1)^n = \frac{1}{2-(x+1)} = \sum_{n=0}^{\infty} \frac{(x+1)^n}{2^{n+1}}$$

$$\Longrightarrow \quad b_{n+2} = \frac{1/2^{n+1} - b_n}{(n+2)(n+1)} \quad \text{for } n = 0, 1, 2, \ldots$$

This gives the general solution as

$$y(x) = b_0 \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!}(x+1)^{2n} + b_1 \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!}(x+1)^{2n+1} +$$
$$+ \tfrac{1}{4}(x+1)^2 + \tfrac{1}{24}(x+1)^3 - \tfrac{1}{96}(x+1)^4 + \tfrac{1}{960}(x+1)^5 +$$
$$+ \tfrac{1}{720}(x+1)^6 + \tfrac{1}{2880}(x+1)^7 + \tfrac{37}{322560}(x+1)^8 + \cdots$$

Example (cont'd)

This expansion is valid (at least) for $-3 < x < 1$, because the distance from $x_0 = -1$ to the singularity of $1/(1-x)$ is 2.

Since we are interested in the extension of the solution on $(-1, 1)$ determined before, we need to solve the same IVP $y(0) = y'(0) = 0$. This gives the two equations

$$b_0 \cos 1 + b_1 \sin 1 + \tfrac{1}{4} + \tfrac{1}{24} - \tfrac{1}{96} + \tfrac{1}{960} + \tfrac{1}{720} + \cdots = 0,$$
$$-b_0 \sin 1 + b_1 \cos 1 + \tfrac{2}{4} + \tfrac{3}{24} - \tfrac{4}{96} + \tfrac{5}{960} + \tfrac{6}{720} + \cdots = 0,$$

from which b_0, b_1 can be determined. (Likely the two series involved can't be evaluated in closed form, but we can use numerical approximations instead.)

Note: Alternatively, one could determine $b_0 = y(-1)$, $b_1 = y'(-1)$ directly from the series representation in $(-1, 1)$. But this is not advisable, since the resulting alternating series converge slowly (because -1 is on the boundary of the disk of convergence).

Similarly, we can obtain power series solutions of $y'' + y = \frac{1}{1-x}$ that are defined for $x > 1$ by choosing a center $x_0 > 1$, but we stop the discussion here.

The Case of a Regular Singular Point

Now suppose that x_0 is a regular singular point of

$$P(x)y'' + Q(x)y' + R(x)y = 0. \tag{1}$$

As discussed before we can assume $x_0 = 0$, in which case the ODE has the equivalent form

$$x^2y'' + x(p_0 + p_1x + p_2x^2 + \dots)y' + (q_0 + q_1x + q_2x^2 + \dots)y = 0 \tag{5}$$

with $x \mapsto f(x) = \sum_{n=0}^{\infty} p_n x^n$ and $x \mapsto g(x) = \sum_{n=0}^{\infty} q_n x^n$ analytic in $\mathrm{B}_\rho(0)$ for some $\rho > 0$.

Since truncating $f(x)$, $g(x)$ after their constant term yields the Euler equation $x^2y'' + xp_0y' + q_0y = 0$, it is reasonable to try the generalized ("fractional") power series „Ansatz"

$$y(x) = \sum_{n=0}^{\infty} a_n x^{r+n} = x^r \sum_{n=0}^{\infty} a_n x^n = x^r \times \text{analytic}$$

for finding a nonzero solution of (1). The number $r \in \mathbb{C}$ is uniquely determined by $y(x)$ if we require $a_0 \neq 0$ (or, yet better, normalize to $a_0 = 1$).

Substituting

$$y'(x) = \sum_{n=0}^{\infty}(r+n)a_n x^{r+n-1},$$

$$y''(x) = \sum_{n=0}^{\infty}(r+n)(r+n-1)a_n x^{r+n-2}$$

into the explicit form of (1), we obtain the power series equation

$$\sum_{n=0}^{\infty}(r+n)(r+n-1)a_n x^{r+n} + \left(\sum_{n=0}^{\infty}(r+n)a_n x^{r+n}\right)\left(\sum_{n=0}^{\infty}p_n x^n\right) + \\ + \left(\sum_{n=0}^{\infty}a_n x^{r+n}\right)\left(\sum_{n=0}^{\infty}q_n x^n\right) = 0,$$

which (take the factor x^r out!) is equivalent to

$$(r+n)(r+n-1)a_n + \sum_{k=0}^{n}(r+k)a_k p_{n-k} + \sum_{k=0}^{n}a_k q_{n-k} = 0,$$

$$n = 0, 1, 2, \ldots$$

This can be rewritten as

$$[(r+n)(r+n-1)+(r+n)p_0+q_0]\,a_n+\sum_{k=0}^{n-1}[(r+k)p_{n-k}+q_{n-k}]\,a_k=0$$

for $n=0,1,2,\ldots$

Observations

- The first equation is

  $$[r(r-1)+rp_0+q_0]\,a_0=0.$$

  It has the form $F(r)a_0=0$ with the quadratic polynomial $F(r)=r(r-1)+rp_0+q_0=r^2+(p_0-1)r+q_0$.

- In each of the subsequent equations, a_n appears with the coefficient

  $$(r+n)(r+n-1)+(r+n)p_0+q_0=F(r+n).$$

  Apart from a_n only numbers a_k with $k<n$ appear in such an equation.

Observations cont'd

- Regarding r as a variable, we can solve the equations for $n = 1, 2, 3, \ldots$ by defining $a_n = a_n(r)$ recursively as

$$a_0(r) = 1,$$

$$a_n(r) = -\frac{1}{F(r+n)} \sum_{k=0}^{n-1} [(r+k)p_{n-k} + q_{n-k}]\, a_k(r) \quad \text{for } n \geq 1.$$

The so-defined $r \mapsto a_n(r) = P_n(r)/Q_n(r)$ is a rational function of r, whose denominator can be taken as the polynomial $Q_n(r) = F(r+1)F(r+2)\cdots F(r+n)$.

For example, we have

$$a_1(r) = -\frac{1}{F(r+1)}[rp_1 + q_1]a_0(r) = -\frac{rp_1 + q_1}{F(r+1)},$$

$$a_2(r) = -\frac{1}{F(r+2)}\big([rp_2 + q_2]a_0(r) + [(r+1)p_1 + q_1]a_1(r)\big)$$

$$= \frac{F(r+1)[rp_2 + q_2] - [(r+1)p_1 + q_1][rp_1 + q_1]}{F(r+1)F(r+2)}.$$

These observations remain mutatis mutandis true for regular singular points $x_0 \neq 0$. In the general case, $Q(x)/P(x)$ and $R(x)/P(x)$ must be expanded into powers of $x - x_0$, viz.,

$$p(x) = \frac{Q(x)}{P(x)} = \frac{p_0}{x - x_0} + p_1 + p_2(x - x_0) + p_3(x - x_0)^2 + \cdots,$$

$$q(x) = \frac{R(x)}{P(x)} = \frac{q_0}{(x - x_0)^2} + \frac{q_1}{x - x_0} + q_2 + q_3(x - x_0) + \cdots$$

and $F(r) = r^2 + (p_0 - 1)r + q_0$ formed from the numbers p_0, q_0 appearing in these expansions.

Definition

The quadratic equation $F(r) = 0$ is called *indicial equation* associated with the regular singular point x_0 of (1). Its roots r_1, r_2 (in the case $r_1, r_2 \in \mathbb{R}$ ordered as $r_1 \geq r_2$) are called *exponents at the singularity* x_0.

Note

$F(r) = 0$ is exactly the equation that r should satisfy in order for $y(x) = x^r$ to form a solution of the Euler equation $x^2y'' + p_0xy' + q_0y = 0$ obtained by truncating (5).

Theorem (cf. [BDM17], Th. 5.6.1)

Suppose that $x_0 = 0$ is a regular singular point of

$$P(x)y'' + Q(x)y' + R(x)y = 0 \tag{1}$$

and that

$$p(x) = \frac{Q(x)}{P(x)} = \frac{p_0}{x} + p_1 + p_2 x + p_3 x^2 + p_4 x^3 + \cdots,$$

$$q(x) = \frac{R(x)}{P(x)} = \frac{q_0}{x^2} + \frac{q_1}{x} + q_2 + q_3 x + q_4 x^2 + \cdots$$

holds for $0 < |x| < \rho$ (i.e., $\sum_{n=0}^{\infty} p_n x^n$, $\sum_{n=0}^{\infty} q_n x^n$ converge for $|x| < \rho$).

1. *If the exponents r_1, r_2 at x_0 are distinct and $r_1 - r_2 \notin \mathbb{Z}$, then (1) has two linearly independent solutions on $(0, \rho)$, viz.*

$$y_i(x) = x^{r_i}\left(1 + \sum_{n=1}^{\infty} a_n(r_i)x^n\right), \quad i = 1, 2.$$

These are obtained by setting $a_0(r_i) = 1$ and for $n \geq 1$ determining $a_n(r_i)$ recursively from

$$a_n(r_i) = -\frac{1}{F(r_i + n)} \sum_{k=0}^{n-1} \left[(r_i + k)p_{n-k} + q_{n-k}\right] a_k(r_i).$$

Theorem (cont'd)

2. *If $r_1 - r_2 \in \mathbb{Z}$, the larger root r_1 (respectively, the double root $r_1 = r_2$) yields one solution of (1) of the form $y_1(x) = x^{r_1}\left(1 + \sum_{n=1}^{\infty} a_n(r_1)x^n\right)$ on $(0, \rho)$. The coefficients $a_n(r_1)$ are determined in the same way as in Case (1).*

3. *If $r_1 = r_2$, a second solution of (1) on $(0, \rho)$ that is linearly independent of $y_1(x)$ is*

$$y_2(x) = y_1(x)\ln x + x^{r_1}\sum_{n=1}^{\infty} b_n(r_1)x^n.$$

with $b_n(r_1) = a_n'(r_1)$.

4. *If $r_1 - r_2 = N \in \mathbb{Z}^+$, a second solution of (1) on $(0, \rho)$ that is linearly independent of $y_1(x)$ is*

$$y_2(x) = a\,y_1(x)\ln x + x^{r_2}\left(1 + \sum_{n=1}^{\infty} c_n(r_2)x^n\right)$$

with $a = \lim_{r\to r_2}(r - r_2)a_N(r)$ *and* $c_n(r_2) = \dfrac{\mathrm{d}}{\mathrm{d}r}\left[(r - r_2)a_n(r)\right]\Big|_{r=r_2}$.

Notes on the theorem

- The theorem also holds for regular singular points $x_0 \neq 0$, provided one replaces x by $x - x_0$ and $(0, \rho)$ by $(x_0, x_0 + \rho)$ everywhere in its statement.
- Solutions on $(-\rho, 0)$ (resp., on $(x_0 - \rho, x_0)$ can be obtained by making the substitution $z(x) = y(-x)$ in (1), which gives $P(-x)z'' - Q(-x)z' + R(-x)z = 0$. The corresponding equation (5) is

  $$x^2 z'' + x(p_0 - p_1 x + p_2 x^2 \mp \cdots)z' + (q_0 - q_1 x + q_2 x^2 \mp \cdots)z = 0,$$

  which has the same indicial equation as (1). Further one can show by induction that the coefficients $a_n(r)$ change to $(-1)^n a_n(r)$ when using the "alternating" sequences $(p_0, -p_1, p_2, \dots)$, $(q_0, -q_1, q_2, \dots)$ instead of $(p_0, p_1, p_2, \dots)$, $(q_0, q_1, q_2, \dots)$. This implies that solutions on $(-\rho, 0)$ have the same form as in the theorem (with the same $a_n(r)$, $b_n(r)$, $c_n(r)$, a, because the change $a_n(r) \to (-1)^n a_n(r)$ is undone by the the back substitution $y(x) = z(-x)$), except that x^{r_i} is replaced by $(-x)^{r_i} = |x|^{r_i}$ and $\ln x$ by $\ln(-x) = \ln|x|$. Thus, if we write $|x|^{r_i}$ and $\ln|x|$ in the formulas then both cases $0 < x < \rho$ and $-\rho < x < 0$ are covered; cf. [BDM17], Th. 5.6.1.

Notes on the theorem cont'd

- It is usually difficult to obtain an explicit formula for the functions $a_n(r)$ from the recurrence relation. Hence, instead of computing $a_n(r)$ and the expressions for $b_n(r)$, $c_n(r)$ and a in terms of $a_n(r)$, it is often better to use the postulated form of the solution as an „Ansatz“ and try to determine the coefficients $a_n(r_i)$, $b_n(r_i)$, $c_n(r_i)$ and a by substituting it into (1).
- Since the roots of the indicial equation are $r_{1,2} = \frac{1}{2}\left(1 - p_0 \pm \sqrt{(p_0-1)^2 - 4q_0}\right)$, Case 4 ($r_1 - r_2 \in \mathbb{Z}^+$) occurs iff $(p_0-1)^2 - 4q_0 = N^2$ is a perfect square, and then $r_1 = (1 - p_0 + N)/2$, $r_2 = (1 - p_0 - N)/2$.
- In Case 1 ($r_1 - r_2 \notin \mathbb{Z}$) it is possible that $r_1, r_2 \in \mathbb{C} \setminus \mathbb{R}$. Then $r_2 = \overline{r_1}$, the two indicated solutions satisfy $y_2(x) = \overline{y_1(x)}$, and the real and imaginary part of one of them provide a real fundamental system.
- The subsequent proof of the theorem shows that the solution $y_2(x)$ in Case 4 is obtained in the same way as in Case 3 except that the exponent r_1 is replaced by r_2 and the rational functions $a_n(r)$ by $\alpha_n(r) = (r - r_2)a_n(r)$.

Notes on the theorem cont'd

- The theorem is essentially due to GEORG FERDINAND FROBENIUS (1849–1917), and the method developed to solve 2nd-order time-dependent linear ODE's at regular singular points is commonly referred to as the *method of Frobenius*.

Proof of the theorem.

(1) For the first equation to be satisfied, we must choose r as a root of the indicial equation., i.e., $r = r_1$ or $r = r_2$.
Since $r_1 - r_2 \notin \mathbb{Z}$, we have $F(r_i + n) \neq 0$ for all $n \in \mathbb{Z}^+$.
$\Longrightarrow$ $a_n(r_i)$ is defined for all $n \in \mathbb{N}$ and yields a solution of (1).
The solutions $y_1(x)$, $y_2(x)$ obtained in this way are linearly independent, since $y_i(x) \simeq x^{r_i}$ for $x \downarrow 0$ and certainly not $x^{r_1} \simeq c\,x^{r_2}$ for $x \downarrow 0$.

(2) Here we have $F(r_1 + n) \neq 0$ for all $n \in \mathbb{Z}^+$.
$\Longrightarrow$ r_1 gives rise to a solution of (1) as in Case 1.

(3) We work with the two-variable function

$$\phi(r, x) = x^r \sum_{n=0}^{\infty} a_n(r) x^n = \sum_{n=0}^{\infty} a_n(r) x^{r+n},$$

which is defined for $|x| < \rho$, $r \notin \{r_1 - 1, r_1 - 2, \dots\}$ (see below), and the differential operator

$$\begin{aligned} L &= x^2\mathrm{D}^2 + x\,f(x)\mathrm{D} + g(x) \\ &= x^2\mathrm{D}^2 + x(p_0 + p_1 x + \cdots)\mathrm{D} + (q_0 + q_1 x + \cdots)\,\mathrm{id}, \end{aligned}$$

which in the following acts like a partial derivative (i.e., $\mathrm{D} \triangleq \frac{\partial}{\partial x}$).

Proof cont'd.

By definition of the coefficients $a_n(r)$, we have

$$L[\phi] = F(r)a_0(r)x^r + \sum_{n=1}^{\infty} 0\, x^{r+n} = (r-r_1)^2 x^r.$$

Since L involves only $\frac{\partial}{\partial x}$, Clairaut's Theorem gives $L \circ \frac{\partial}{\partial r} = \frac{\partial}{\partial r} \circ L$ (provided we apply it to a C^2-function).

$$\Longrightarrow\ L\left[\frac{\partial \phi}{\partial r}\right] = \frac{\partial}{\partial r} L[\phi] = 2(r-r_1)x^r + (r-r_1)^2(\ln x)x^r,$$

$$L\left[\frac{\partial \phi}{\partial r}(r_1, x)\right] = \frac{\partial}{\partial r} L[\phi]\Big|_{r=r_1} = 0.$$

It follows that a second solution of (1) is

$$\frac{\partial \phi}{\partial r}(r_1, x) = (\ln x)x^{r_1} \sum_{n=0}^{\infty} a_n(r_1)x^n + x^{r_1} \sum_{n=0}^{\infty} a_n'(r_1)x^n$$

$$= (\ln x)y_1(x) + x^{r_1} \sum_{n=1}^{\infty} a_n'(r_1)x^n.$$

Clearly this solution is linearly independent of $y_1(x)$.

Proof cont'd.

The proof of (3) is not yet finished, because the termwise differentiation used in the computation needs to be justified. Also we need to show that the generalized power series series solutions in Parts (1)–(4) actually converge for $|x| < \rho$.

The latter can be done by a slight modification of the method used in the analytic case. For $0 < \rho_1 < \rho$ we have a recursive bound

$$|a_n(r)|\,\rho_1^n \le \sum_{k=0}^{n-1} \frac{(|r|+k+1)\,M\,|a_k(r)|\,\rho_1^k}{|F(r+n)|}$$

for the coefficients of $\phi(r,x)$, obtained from the recurrence relation for $a_n(r)$ in the same way as before (and with the same meaning of M). The auxiliary sequence $\big(u_n(r)\big)$ defined by $u_0(r) = 1$ and

$$u_n(r) = \sum_{k=0}^{n-1} \frac{(|r|+k+1)\,M u_k(r)}{|F(r+n)|} \quad \text{for } n \ge 1$$

still satisfies $\lim_{n\to\infty} \frac{u_{n+1}(r)}{u_n(r)} = 1$ (independently of r), so that the proof can be finished in the same way as before.

Proof cont'd.

The preceding argument can be modified to yield uniform convergence of $\phi(r,x) = \sum_{n=0}^{\infty} a_n(r)x^{r+n}$ and its partial derivatives up to certain orders (order 1 for $\frac{\partial}{\partial r}$ and order 2 for $\frac{\partial}{\partial x}$) on compact subsets of their domain, justifying termwise differentiation. The arguments in Parts (1), (2), (4) are similar.

(4) Here we set

$$\phi(r,x) = (r-r_2)x^r \sum_{n=0}^{\infty} a_n(r)x^n = x^r \sum_{n=0}^{\infty}(r-r_2)a_n(r)x^n.$$

The functions $\alpha_n(r) = (r-r_2)a_n(r)$ satisfy the same recurrence relation as $a_n(r)$, but start with $\alpha_0(r) = r - r_2$.

$$\Longrightarrow\ \alpha_N(r) = -\frac{1}{(r+N-r_1)(r+N-r_2)} \sum_{k=0}^{N-1} \left[(r+k)p_{n-k} + q_{n-k}\right](r-r_2)a_k(r)$$

$$= -\frac{1}{r+N-r_2} \sum_{k=0}^{N-1} \left[(r+k)p_{n-k} + q_{n-k}\right] a_k(r),$$

since $r_1 = r_2 + N$.

Proof cont'd.

$\Longrightarrow \alpha_N(r)$ is analytic at r_2.

The recurrence relation then implies that $\alpha_n(r)$ are analytic at r_2 for $n > N$. Clearly this also holds for $n < N$, in which case $\alpha_n(r_2) = 0$.

As in (3) it then follows that $\phi(r, x)$ defined for $|x| < \rho$, $r \notin \{r_1 - 1, r_1 - 2, \ldots, r_1 - N + 1, r_1 - N - 1, r_1 - N - 2, \ldots\}$, and satisfies

$$L[\phi] = (r - r_1)(r - r_2)\alpha_0(r) = (r - r_1)(r - r_2)^2 x^r,$$

$$L\left[\frac{\partial \phi}{\partial r}(r_2, x)\right] = \frac{\partial}{\partial r} L[\phi]\Big|_{r=r_2} = 0.$$

$$\Longrightarrow \frac{\partial \phi}{\partial r}(r_2, x) = (\ln x) x^{r_2} \sum_{n=0}^{\infty} \alpha_n(r_2) x^n + x^{r_2} \sum_{n=0}^{\infty} \alpha_n'(r_2) x^n$$

is a second solution of (1).

It remains to verify that this solution has the form stated in the theorem. For the 2nd summand this is true by definition of $\alpha_n(r)$.

Proof cont'd.

The first summand can be rewritten as

$$(\ln x)\sum_{n=N}^{\infty}\alpha_n(r_2)x^{n+r_2} = (\ln x)\sum_{n=0}^{\infty}\alpha_{n+N}(r_2)x^{n+r_1},$$

which is equal to $a\,y_1(x)\ln x = \alpha_N(r_2)y_1(x)\ln x$ iff $\alpha_{n+N}(r_2) = \alpha_N(r_2)a_n(r_1)$ for $n \in \mathbb{N}_0$. This in turn can be proved by induction on n (the case $n = 0$ being trivial):

$$\begin{aligned}
\alpha_{n+N}(r_2) &= -\frac{1}{F(r_2+n+N)}\sum_{k=0}^{n+N-1}\left[(r_2+k)p_{n+N-k}+q_{n+N-k}\right]\alpha_k(r_2)\\
&= -\frac{1}{F(r_1+n)}\sum_{k=N}^{n+N-1}\left[(r_2+k)p_{n+N-k}+q_{n+N-k}\right]\alpha_k(r_2)\\
&= -\frac{1}{F(r_1+n)}\sum_{k=0}^{n-1}\left[(r_2+k+N)p_{n-k}+q_{n-k}\right]\alpha_{k+N}(r_2)\\
&= -\frac{1}{F(r_1+n)}\sum_{k=0}^{n-1}\left[(r_1+k)p_{n-k}+q_{n-k}\right]\alpha_N(r_2)a_k(r_1)\\
&= \alpha_N(r_2)a_n(r_1).
\end{aligned}$$

□

Example

Find two linearly independent solutions of the ODE

$$2xy'' + y' + xy = 0, \quad x > 0.$$

Rewriting the ODE as

$$y'' + \frac{1}{2x}\,y' + \frac{1}{2}\,y = 0,$$

we see that $x = 0$ is a regular singular point and $p_0 = 1/2$ ($p_1 = p_2 = \cdots = 0$), $q_0 = 0$ ($q_1 = 0$, $q_2 = 1/2$, $q_3 = q_4 = \cdots = 0$).
$\Longrightarrow$ The indicial equation is

$$r^2 + (p_0 - 1)r + q_0 = r^2 - \tfrac{1}{2}\,r = r\left(r - \tfrac{1}{2}\right) = 0.$$

$\Longrightarrow$ The exponents at the singularity $x = 0$ are $r_1 = 1/2$, $r_2 = 0$.
Thus we are in Case (1) of the theorem and there must be solutions $y_1(x)$, $y_2(x)$ of the form

$$y_1(x) = \sqrt{x}\sum_{n=0}^{\infty} a_n(1/2)x^n, \qquad y_2(x) = \sum_{n=0}^{\infty} a_n(0)x^n.$$

Example (cont'd)

$r_1 = 1/2$:

Instead of using the general recurrence relation for $a_n(r)$ at $r = 1/2$, we determine it directly from the ODE, writing a_n in place of $a_n(1/2)$.

$$y(x) = \sum_{n=0}^{\infty} a_n x^{n+1/2}$$

$$2x\,y''(x) = 2x \sum_{n=0}^{\infty} \left(n+\tfrac{1}{2}\right)\left(n-\tfrac{1}{2}\right) a_n x^{n-3/2}$$

$$= \sum_{n=0}^{\infty} 2\left(n+\tfrac{1}{2}\right)\left(n-\tfrac{1}{2}\right) a_n x^{n-1/2}$$

$$y'(x) = \sum_{n=0}^{\infty} \left(n+\tfrac{1}{2}\right) a_n x^{n-1/2}$$

$$x\,y(x) = \sum_{n=0}^{\infty} a_n x^{n+3/2} = \sum_{n=2}^{\infty} a_{n-2} x^{n-1/2}$$

$$\Longrightarrow 0 \cdot a_0 + 3a_1 x + \sum_{n=2}^{\infty} \left[(2n+1)na_n + a_{n-2}\right] x^n = 0$$

Example (cont'd)

$\Longrightarrow a_1 = 0$ and $a_n = -\dfrac{a_{n-2}}{n(2n+1)}$ for $n \geq 2$

$\Longrightarrow$ All odd coefficients a_{2n+1} are zero, and

$$y_1(x) = x^{1/2}\left(1 - \frac{x^2}{2\cdot 5} + \frac{x^4}{2\cdot 4\cdot 5\cdot 9} - \frac{x^6}{2\cdot 4\cdot 6\cdot 5\cdot 9\cdot 13} \pm \cdots\right)$$
$$= \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+\frac{1}{2}}}{2^n n!\, 5\cdot 9\cdot 13\cdots(4n+1)}.$$

The theorem predicts $\rho = \infty$ for the power series (without the factor $x^{1/2}$), which can also be seen with the aid of the ratio test.

Example (cont'd)

$r_2 = 0$: Writing again a_n in place of $a_n(0)$, we obtain

$$y(x) = \sum_{n=0}^{\infty} a_n x^n$$

$$2x\,y''(x) = 2x \sum_{n=2}^{\infty} n(n-1)a_n x^{n-2}$$

$$= \sum_{n=2}^{\infty} 2n(n-1)a_n x^{n-1}$$

$$y'(x) = \sum_{n=1}^{\infty} n a_n x^{n-1}$$

$$x\,y(x) = \sum_{n=0}^{\infty} a_n x^{n+1} = \sum_{n=2}^{\infty} a_{n-2} x^{n-1}$$

$$\Longrightarrow 0 \cdot a_0 + a_1 x + \sum_{n=2}^{\infty} \left[n(2n-1)a_n + a_{n-2}\right] x^n = 0$$

Example (cont'd)

$\Longrightarrow a_1 = 0$ and $a_n = -\dfrac{a_{n-2}}{n(2n-1)}$ for $n \geq 2$

$\Longrightarrow$ Again all odd coefficients a_{2n+1} are zero, and

$$y_2(x) = 1 - \frac{x^2}{2\cdot 3} + \frac{x^4}{2\cdot 4\cdot 3\cdot 7} - \frac{x^6}{2\cdot 4\cdot 6\cdot 3\cdot 7\cdot 11} \pm \cdots$$
$$= \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{2^n n!\, 3\cdot 7\cdot 11\cdots(4n-1)}.$$

The theorem predicts $\rho = \infty$, which again can also be easily found with the ratio test.

In all we have shown that $y_1(x)$, $y_2(x)$ form a fundamental system of solutions of $2xy'' + y' + xy = 0$ on $(0, \infty)$. A fundamental system $y_1^-(x)$, $y_2^-(x)$ of solutions on $(-\infty, 0)$ is then obtained by changing the fractional part of $y_1(x)$ to $\sqrt{-x}$.

In the special case under consideration, since only even powers x^{2n} appear in $y_1(x)$, $y_2(x)$, this is equivalent to setting $y_1^-(x) = y_1(-x)$, $y_2^-(x) = y_2(-x)$ for $x < 0$.

$y_2(x)$ and its constant multiples are defined and solve the ODE on $\mathbb{R}$.

Example (cont'd)

Finally we compare our method of determining the fundamental solutions $y_1(x)$, $y_2(x)$ directly from the ODE with that of employing the rational functions $a_n(r)$. Since the only nonzero coefficients among p_i, q_i are $p_0 = q_2 = 1/2$, the general recurrence relation becomes

$$a_n(r) = -\frac{q_2\, a_{n-2}(r)}{F(r+n)} = -\frac{a_{n-2}(r)}{2(r+n)(r+n-\frac{1}{2})}$$
$$= -\frac{a_{n-2}(r)}{(r+n)(2r+2n-1)},$$

supplemented by $a_0(r) = 1$, $a_1(r) = 0$.
$\Longrightarrow a_{2n+1}(r) = 0$ and

$$a_{2n}(r) = \frac{(-1)^n}{(r+2)(r+4)\cdots(r+2n)(2r+3)(2r+7)\ldots(2r+4n-1)}.$$

For $r \in \{0, \frac{1}{2}\}$ this coincides with the formula determined for the coefficients of $y_1(x)$, $y_2(x)$ earlier. (For $r = 1/2$ the denominator of $a_{2n}(1/2)$ is $2^{-n}\,5\cdot 9\cdots(4n+1)4\cdot 8\cdots 4n = 2^n n!\,5\cdot 9\cdots(4n+1)$.)

Example

We solve the Legendre equation

$$(1-x^2)y'' - 2xy' + n(n+1)y = 0, \qquad n \in \mathbb{N},$$

near the singular point $x_0 = 1$.

Since $x_0 = 1$ is a simple zero of $P(x) = 1 - x^2$ and not a zero of $Q(x) = -2x$ (or not a zero of $R(x) = n(n+1)$), $x_0 = 1$ is a regular singular point.

First we put the ODE into explicit form and rewrite the coefficients in terms of $x - 1$:

$$\begin{aligned} & y'' + \frac{2x}{(x-1)(x+1)}\,y' - \frac{n(n+1)}{(x-1)(x+1)}\,y \\ &= y'' + \left(\frac{1}{x-1} + \frac{1}{2+(x-1)}\right) y' - \frac{n(n+1)}{(x-1)\big(2+(x-1)\big)}\,y \\ &= y'' + \left(\frac{1}{x-1} + \sum_{k=0}^{\infty} \frac{(-1)^k}{2^{k+1}}(x-1)^k\right) y' + n(n+1)\left(\sum_{k=-1}^{\infty} \frac{(-1)^k}{2^{k+2}}(x-1)^k\right) y, \end{aligned}$$

from which we can read off $p_0 = 1$, $q_0 = 0$, and $\rho = 2$.

Example (cont'd)

Remark: Since we do not need the full expansion of $p(x)$ and $q(x)$, it is easier in this case to use the formulas

$$p_0 = \lim_{x\to 1}(x-1)p(x) = \lim_{x\to 1}\frac{2x}{x+1} = 1,$$
$$q_0 = \lim_{x\to 1}(x-1)^2 q(x) = \lim_{x\to 1}\left(-\frac{n(n+1)(x-1)}{x+1}\right) = 0,$$

and by Property (9) of power series the radii of convergence of $\sum p_n(x-1)^n$, $\sum q_n(x-1)^n$ are the distances from $x=1$ to the singularity of $\frac{2x}{x+1}$ resp. $-\frac{n(n+1)}{x+1}$, viz. $1-(-1)=2$.

The indicial equation at $x_0 = 1$ is therefore

$$r^2 + (p_0 - 1)r + q_0 = r^2 = 0. \qquad \Longrightarrow r_1 = r_2 = 0.$$

Hence we are in Cases (2) and (3) of the theorem, and there are fundamental solutions of the form

$$y_1(x) = \sum_{k=0}^{\infty} a_k(x-1)^k, \quad y_2(x) = y_1(x)\ln|x-1| + \sum_{k=1}^{\infty} b_k(x-1)^k.$$

Example (cont'd)

For the computation of $y_1(x)$ we make the substitution $t = x - 1$, i.e., $x = t + 1$, which gives $1 - x^2 = -t^2 - 2t$, $-2x = -2t - 2$ and turns the Legendre equation into

$$-(t^2+2t)y''(t+1) - (2t+2)y'(t+1) + n(n+1)y(t+1) = 0.$$

Substituting $y_1(x) = y_1(t+1) = \sum_{k=0}^{\infty} a_k t^k$ gives

$$-(t^2+2t)\left(\sum_{k=2}^{\infty} k(k-1)a_k t^{k-2}\right) - (2t+2)\left(\sum_{k=1}^{\infty} k a_k t^{k-1}\right) + n(n+1)\sum_{k=0}^{\infty} a_k t^k$$

$$= \sum_{k=0}^{\infty}\Big[-k(k-1)a_k - 2(k+1)k a_{k+1} - 2k a_k - 2(k+1)a_{k+1} + n(n+1)a_k\Big]t^k$$

$$= \sum_{k=0}^{\infty}\Big[\big(n(n+1) - k(k+1)\big)a_k - 2(k+1)^2 a_{k+1}\Big]t^k = 0$$

$$\Longrightarrow a_{k+1} = \frac{n(n+1) - k(k+1)}{2(k+1)^2}\,a_k = \frac{(n-k)(n+k+1)}{2(k+1)^2}\,a_k$$

for $k = 0, 1, 2, \ldots$

Example (cont'd)

Setting $a_0 = 1$ gives

$$a_k = \frac{n(n-1)\cdots(n-k+1)(n+1)(n+2)\cdots(n+k)}{2^k(k!)^2}$$
$$= \frac{1}{2^k}\binom{n}{k}\binom{n+k}{k}.$$

The coefficients a_k with $k > n$ vanish, since in this case $n(n-1)\cdots(n-k+1)$ contains the factor 0.

$$\Longrightarrow\ y_1(x) = \sum_{k=0}^{n} \frac{1}{2^k}\binom{n}{k}\binom{n+k}{k}(x-1)^k.$$

Since $y_1(x)$ is a polynomial, it solves the Legendre equation everywhere, and hence must be a constant multiple of the Legendre polynomial $\mathrm{P}_n(x)$. The leading coefficient of $y_1(x)$ is

$$\frac{1}{2^n}\binom{n}{n}\binom{n+n}{n} = \frac{1}{2^n}\binom{2n}{n},$$

the same as that of $\mathrm{P}_n(x)$!!!

Example (cont'd)

So we have discovered the identity

$$\mathrm{P}_n(x) = \sum_{k=0}^{n} \frac{1}{2^k} \binom{n}{k} \binom{n+k}{k} (x-1)^k$$
$$= 1 + \frac{n(n+1)}{2}(x-1) + \frac{(n-1)n(n+1)(n+2)}{16}(x-1)^2 + \cdots$$

From this we see that $\mathrm{P}_n(1) = 1$, which is not obvious from the original definition of $\mathrm{P}_n(x)$ and explains why the Legendre polynomials are normalized in the strange way $\mathrm{P}_n(x) = \frac{1}{2^n}\binom{2n}{n} x^n +$ smaller powers.

Since $\mathrm{P}_n(x)$ is even (odd) when n is even (resp., odd), this also gives $\mathrm{P}_n(-1) = (-1)^n$, which in turn yields the binomial coefficient identity

$$\sum_{k=0}^{n} (-1)^k \binom{n}{k} \binom{n+k}{k} = (-1)^n, \quad n = 0, 1, 2, \ldots$$

Example (cont'd)

For determining the 2nd fundamental solution $y_2(x)$ it will be convenient to use the associated differential operator $L[y] = (1-x^2)y'' - 2xy' + n(n+1)y$. For $x > 1$ we compute

$$
\begin{aligned}
L\big[y_2(x)\big] &= L\big[\mathrm{P}_n(x)\ln(x-1)\big] + L\left[\sum_{k=1}^{\infty} b_k (x-1)^k\right] \\
&= (1-x^2)\left(\mathrm{P}_n''(x)\ln(x-1) + 2\mathrm{P}_n'(x)\frac{1}{x-1} + \mathrm{P}_n(x)\frac{-1}{(x-1)^2}\right) \\
&\quad - 2x\left(\mathrm{P}_n'(x)\ln(x-1) + \mathrm{P}_n(x)\frac{1}{x-1}\right) \\
&\quad + n(n+1)\mathrm{P}_n(x)\ln(x-1) + L\left[\sum_{k=1}^{\infty} b_k (x-1)^k\right] \\
&= -2(x+1)\mathrm{P}_n'(x) - \mathrm{P}_n(x) + L[\ldots] \\
&= -2(t+2)\mathrm{P}_n'(t+1) - \mathrm{P}_n(t+1) + \\
&\quad + \sum_{k=0}^{\infty}\Big[\big(n(n+1) - k(k+1)\big)b_k - 2(k+1)^2 b_{k+1}\Big]t^k = 0,
\end{aligned}
$$

where we have set $b_0 = 0$.

Example (cont'd)

The resulting inhomogeneous linear recurrence relation for $(b_1, b_2, b_3, \dots)$ clearly has a unique solution. (For $b_0 \neq 0$ it has a unique solution as well, but this amounts to adding $b_0 y_1(x) = b_0 \mathrm{P}_n(x)$ to $y_2(x)$, which gives nothing new.)

If the order n increases, the number of nonzero terms in the inhomogeneous part (which is a polynomial of degree n) will increase as well, making it unlikely that there is a simple formula for b_n in general. For this reason we will consider only the cases $n = 0$ and $n = 1$.

It should be noted here that Frobenius' power series method doesn't provide a convenient way of solving the general Legendre equation completely. A 2nd fundamental solution of the Legendre equation on $(-\infty, -1)$ or $(1, +\infty)$ can be more easily found by other methods; cf. the subsequent remarks.

Example (cont'd)

We consider only the cases $n = 0$ and $n = 1$.

$n = 0$

Since $\mathrm{P}_0(x) = 1$, we obtain $b_1 = -1/2$ and the recurrence relation

$$b_{k+1} = \frac{(0-k)(0+k+1)}{2(k+1)^2}\, b_k = -\frac{k}{2(k+1)}\, b_k.$$

The solution is $b_k = \frac{(-1)^k}{k2^k}$ (obvious from $(k+1)b_{k+1} = -\frac{1}{2}\, kb_k$).

$$\begin{aligned}
\Longrightarrow y_2(x) &= \ln(x-1) - \frac{x-1}{2} + \frac{1}{2}\frac{(x-1)^2}{2^2} - \frac{1}{3}\frac{(x-1)^3}{2^3} \pm \dots \\
&= \ln(x-1) - \ln\left(1 + \frac{x-1}{2}\right) = \ln(x-1) - \ln\left(\frac{x+1}{2}\right) \\
&= \ln\frac{x-1}{x+1} + \ln 2 \qquad \text{for } 1 < x < 3.
\end{aligned}$$

An equivalent choice is

$$\mathrm{Q}_0(x) = \tfrac{1}{2}\ln\left|\frac{1+x}{1-x}\right| = \tfrac{1}{2}\ln\frac{x+1}{x-1} = -\tfrac{1}{2}\big(y_2(x) - \ln 2\big) \quad \text{for } 1 < x < 3.$$

Example (cont'd)

The function $\mathrm{Q}_0(x) = \frac{1}{2}\ln\left|\frac{1+x}{1-x}\right|$ is defined on $\mathbb{R}\setminus\{\pm 1\}$ and solves the Legendre equation with $n=0$, viz. $(1-x^2)y'' - 2xy' = 0$, on the three intervals into which $\mathbb{R}\setminus\{\pm 1\}$ decomposes. On the middle interval $(-1,1)$ it is characterized as the solution satisfying the initial conditions $\mathrm{Q}_0(0)=0$, $\mathrm{Q}_0'(0)=1$.

Moreover, $\mathrm{Q}_0(x)$ coincides on $(-1,1)$ with the non-polynomial series solution obtained earlier (and also with $\tanh^{-1}$).
In fact one easily verifies

$$\mathrm{Q}_0(x) = x + \frac{x^3}{3} + \frac{x^5}{5} + \frac{x^7}{7} + \cdots,$$

$$\mathrm{Q}_0'(x) = 1 + x^2 + x^4 + x^6 + \cdots = \frac{1}{1-x^2}$$

for $|x|<1$. The identity $\mathrm{Q}_0'(x) = \frac{1}{1-x^2}$ holds for all $x\in\mathbb{R}\setminus\{\pm 1\}$.

Finally let us note that $(1-x^2)y'' - 2xy' = 0$ is 1st-order linear in y' and hence can be solved by the standard method
$y'(x) = \exp\left(\int \frac{2x}{1-x^2}\,\mathrm{d}x\right)$ and one further integration.

Example (cont'd)

$n = 1$

Since $\mathrm{P}_1(x) = x$, the condition for y_2 takes the form

$$L\big[y_2(x)\big] = -3t - 5 + \sum_{k=0}^{\infty}\Big[\big(2 - k(k+1)\big)b_k - 2(k+1)^2 b_{k+1}\Big]t^k = 0.$$

$$\Longrightarrow b_1 = -\frac{5}{2},\ b_2 = -\frac{3}{8}, \quad b_{k+1} = -\frac{(k-1)(k+2)}{2(k+1)^2}\,b_k \quad \text{for } k \geq 2$$

The solution is $b_k = \frac{(-1)^{k-1}(k+1)}{k(k-1)2^k}$ for $k \geq 2$, as can be seen by writing the recurrence relation in the form $\frac{b_{k+1}}{k+2} = -\frac{k-1}{2(k+1)}\,\frac{b_k}{k+1}$ or,equivalently, $\frac{b_k}{k+1} = -\frac{k-2}{2k}\,\frac{b_{k-1}}{k}$.

$$\Longrightarrow y_2(x) = x\ln(x-1) - 5\,\frac{x-1}{2} + \sum_{k=2}^{\infty}\frac{k+1}{k(k-1)}\,\frac{(x-1)^k}{2^k},$$

valid for $1 < x < 3$. Replacing $\ln(x-1)$ by $\ln|x-1|$, we can extend the range to $-1 < x < 3$, $x \neq 1$.

Example (cont'd)

With some effort one can derive from this that another equivalent choice is

$$\mathrm{Q}_1(x) = \frac{x}{2}\ln\left|\frac{1+x}{1-x}\right| - 1 = -\tfrac{1}{2}y_2(x) - x.$$

The function $\mathrm{Q}_1(x)$ is defined on $\mathbb{R}\setminus\{\pm 1\}$ and solves the Legendre equation for $n = 1$, viz. $(1-x^2)y'' - 2xy' + 2y = 0$, on all three subintervals. On the middle interval $(-1,1)$ it is the solution characterized by $\mathrm{Q}_1(0) = -1$, $\mathrm{Q}_1'(0) = 0$.

Like $\mathrm{Q}_0(x)$, the solution $\mathrm{Q}_1(x)$ can be found with less effort by other means, for example by using the method of order reduction for linear 2nd-order ODE's (see our earlier example in `lecture27-28`), or by inspecting the non-polynomial series solution obtained earlier for $x \in (-1,1)$, rewriting it in terms of ln, and extending it to $\mathbb{R}\setminus\{\pm 1\}$.

Remark

The Legendre polynomials (or "Legendre P-functions") are determined by $\mathrm{P}_0(x) = 1$, $\mathrm{P}_1(x) = x$, and the recurrence relation

$$\mathrm{P}_{n+1}(x) = \frac{2n+1}{n+1}\, x\, \mathrm{P}_n(x) - \frac{n}{n+1}\, \mathrm{P}_{n-1}(x), \quad n = 1, 2, 3, \ldots.$$

The *Legendre Q-functions* $\mathrm{Q}_n(x)$ are defined for $x \in \mathbb{R} \setminus \{\pm 1\}$ by $\mathrm{Q}_0(x) = \frac{1}{2} \ln \left|\frac{1+x}{1-x}\right|$, $\mathrm{Q}_1(x) = \frac{x}{2} \ln \left|\frac{1+x}{1-x}\right| - 1$ and the same recurrence relation

$$\mathrm{Q}_{n+1}(x) = \frac{2n+1}{n+1}\, x\, \mathrm{Q}_n(x) - \frac{n}{n+1}\, \mathrm{Q}_{n-1}(x), \quad n = 1, 2, 3, \ldots.$$

One can show that for each $n \in \mathbb{N}$ the functions P_n and Q_n form a fundamental system of solutions of Legendre's equation $(1 - x^2)y'' - 2xy' + n(n+1)y = 0$ on $(-\infty, -1)$, $(-1, 1)$, and $(1, +\infty)$.

Exercise

a) Determine $\mathrm{P}_n(0)$ for $n \in \mathbb{N}$.

b) Use a) to derive a binomial coefficient identity along the lines of the previous example.

Math 285
Introduction to Differential Equations

Thomas Honold

ZJU-UIUC Institute

Spring Semester 2024

Outline

Today's Lecture: BESSEL's Differential Equation

Definition (BESSEL's Differential Equation)

The 2nd-order linear ODE

$$x^2y'' + xy' + (x^2 - \nu^2)y = 0, \qquad x > 0,$$

with parameter $\nu \geq 0$ is known as *Bessel's Differential Equation*. For $\nu \in \mathbb{Z}$ solutions are called *cylinder functions of order* ν.

Rewriting Bessel's ODE as

$$y'' + \frac{1}{x}\,y' + \left(1 - \frac{\nu^2}{x^2}\right) y = 0$$

shows that $x_0 = 0$ is a regular singular point with $p_0 = 1$, $q_0 = -\nu^2$, and that the corresponding indicial equation is

$$r^2 + (p_0 - 1)r + q_0 = r^2 - \nu^2 = (r - \nu)(r + \nu).$$

$\Longrightarrow$ The exponents at the singularity $x_0 = 0$ are $r_1 = \nu$, $r_2 = -\nu$. This means we are in Case 1 (for $\nu \notin \{0, \frac{1}{2}, 1, \frac{3}{2}, 2, \dots\}$, Case 3 (for $\nu = 0$), or Case 4 (for $\nu \in \{\frac{1}{2}, 1, \frac{3}{2}, 2, \dots\}$, with $N = 2\nu$) of the theorem.

Part 2 of the theorem guarantees that one solution is obtained by the fractional power series „Ansatz“ $y(x) = \sum_{n=0}^{\infty} a_n x^{n+\nu}$.

$$\begin{aligned}
\Longrightarrow L[y] &= x^2y'' + xy' + (x^2 - \nu^2)y = \\
&= \sum_{n=0}^{\infty}(n+\nu)(n+\nu-1)a_n x^{n+\nu} + \sum_{n=0}^{\infty}(n+\nu)a_n x^{n+\nu} \\
&\quad + \sum_{n=0}^{\infty} a_n x^{n+\nu+2} - \sum_{n=0}^{\infty}\nu^2 a_n x^{n+\nu} \\
&= x^{\nu}\left(0a_0 + (2\nu+1)a_1 x + \sum_{n=2}^{\infty}\big(n(n+2\nu))a_n + a_{n-2}\big)x^n\right),
\end{aligned}$$

since $(n+\nu)(n+\nu-1) + (n+\nu) - \nu^2 = (n+\nu)^2 - \nu^2 = n(n+2\nu)$.
$L[y] = 0$ implies $a_1 = 0$ (since $2\nu + 1$ is ≥ 1 and hence nonzero) and

$$a_n = -\frac{a_{n-2}}{n(n+2\nu)} \quad \text{for } n \geq 2.$$

$$\Longrightarrow a_{2m+1} = 0,$$

$$a_{2m} = -\frac{a_{2(m-1)}}{4m(m+\nu)} = \cdots = \frac{(-1)^m}{m!\,4^m(\nu+1)(\nu+2)\cdots(\nu+m)}\,a_0.$$

Normalizing by $a_0 = 1$ gives the solution

$$y_1(x) = x^\nu \sum_{m=0}^{\infty} \frac{(-1)^m}{m!\, 2^{2m}(\nu+1)(\nu+2)\cdots(\nu+m)}\, x^{2m} \quad \text{on } (0,\infty).$$

For $\nu \in \mathbb{N}_0$ a different normalization, which gives the coefficients a slightly simpler form, is $a_0 = \frac{1}{2^\nu \nu!}$. The corresponding solution is

$$\begin{aligned} \mathrm{J}_\nu(x) &= x^\nu \sum_{m=0}^{\infty} \frac{(-1)^m}{m!\, 2^{2m+\nu}\nu!(\nu+1)(\nu+2)\cdots(\nu+m)}\, x^{2m} \\ &= \sum_{m=0}^{\infty} \frac{(-1)^m}{m!(m+\nu)!} \left(\frac{x}{2}\right)^{\nu+2m}. \end{aligned}$$

This makes also sense for non-integral ν, provided we interpret $(m+\nu)!$ as $\Gamma(m+\nu+1)$ (which is true for $\nu \in \mathbb{N}_0$).
In Exercise H60 of HW10 it is shown that $1/\Gamma$ can be continuously extended to $\mathbb{R}$. $\Longrightarrow$ $1/\Gamma(m+\nu+1)$ is defined for all $m \in \mathbb{N}_0$ and $\nu \in \mathbb{R}$.

Definition

For $\nu \in \mathbb{R}$, the function

$$\mathrm{J}_\nu(x) = \sum_{m=0}^{\infty} \frac{(-1)^m}{m!\,\Gamma(m+\nu+1)} \left(\frac{x}{2}\right)^{\nu+2m}, \quad x \in (0,\infty),$$

is called *Bessel function of order* ν.

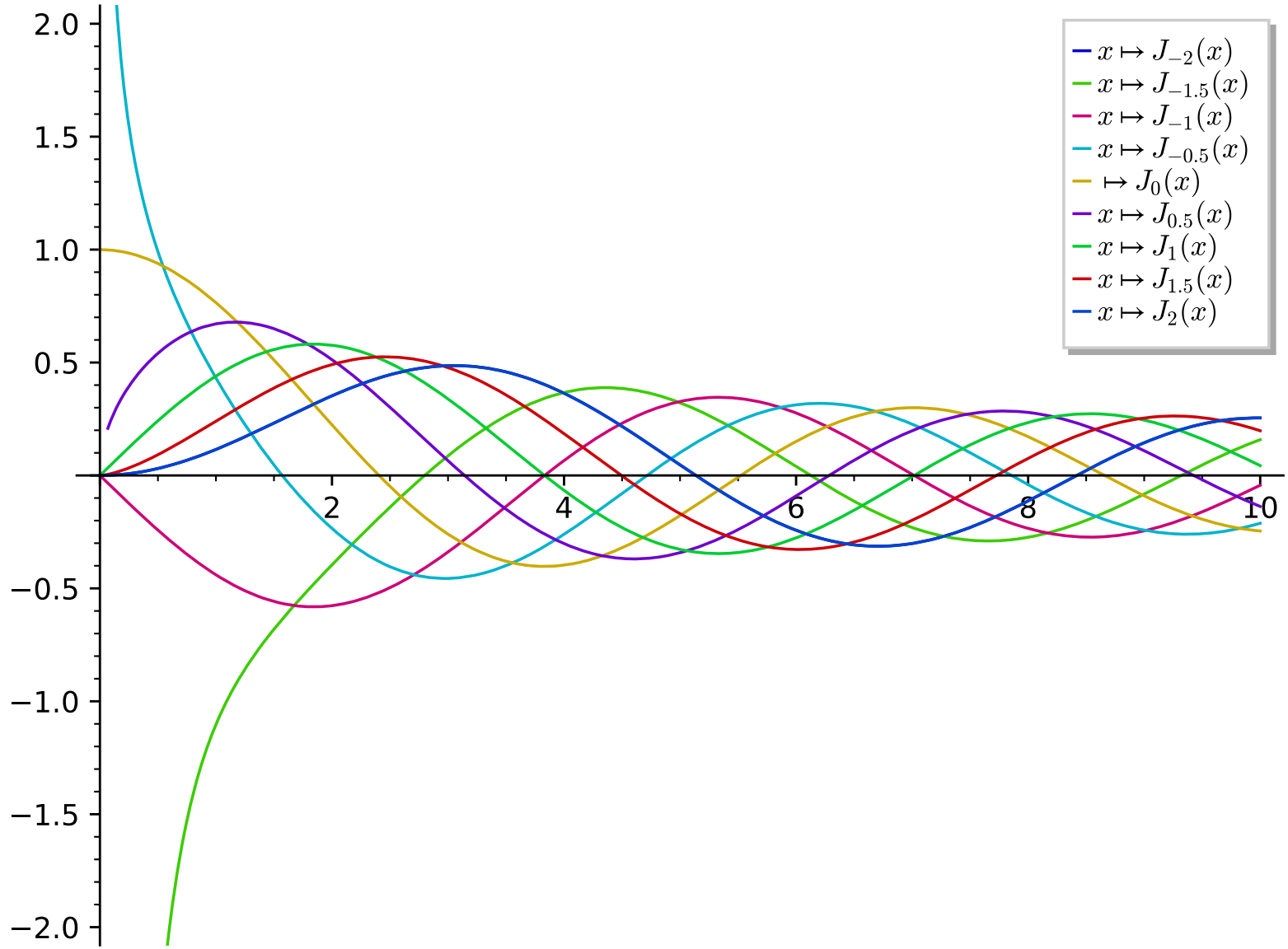


Figure: Bessel functions of orders $\nu = -2, -\frac{3}{2}, -1, -\frac{1}{2}, 0, \frac{1}{2}, 1, \frac{3}{2}, 2$ with domain $(0, \infty)$

For the analysis of the different cases of Bessel's Differential Equation (depending on ν) we switch back to the convention $\nu \geq 0$ adopted earlier.

Though it is not needed for most cases, we first determine the rational functions $a_n(r)$ arising from the condition $L[\phi] = L\left[\sum_{n=0}^{\infty} a_n(r)x^{r+n}\right] = F(r)x^r$.
Since $p(x) = 1/x$, $q(x) = 1 - \nu^2/x^2$, all coefficients p_i, q_i are zero except for $p_0 = 1$, $q_0 = -\nu^2$ and $q_2 = 1$. This gives

$$L[\phi] = \sum_{n=0}^{\infty} \left(F(r+n)a_n(r) + \sum_{k=0}^{n-1} \left[(r+k)p_{n-k} + q_{n-k}\right] a_k(r) \right) x^{r+n}$$

$$= F(r)a_0(r)x^r + F(r+1)a_1(r)x^{r+1} + \sum_{n=2}^{\infty} \left[F(r+n)a_n(r) + a_{n-2}(r)\right] x^{r+n}$$

$$\Longrightarrow a_0(r) = 1,\ a_1(r) = 0,\ a_n(r) = -\frac{a_{n-2}(r)}{F(r+n)} = -\frac{a_{n-2}(r)}{(r+n-\nu)(r+n+\nu)}$$

$$\Longrightarrow a_{2m+1}(r) = 0, \quad a_{2m}(r) = \frac{(-1)^m}{\prod_{i=1}^{m} \left[(r+2i-\nu)(r+2i+\nu)\right]}.$$

(Check that for $r = \nu$ this reduces to the previous formula $a_{2m} = a_{2m}(\nu) = \frac{(-1)^m}{2^{2m} m!(\nu+1)(\nu+2)\cdots(\nu+m)}$.)

The case $\nu \notin \mathbb{Z}$

In this case we claim that there exists a fundamental system of solutions of the form

$$y_1(x) = x^{\nu} \sum_{n=0}^{\infty} a_n(\nu) x^n, \quad y_2(x) = x^{-\nu} \sum_{n=0}^{\infty} a_n(-\nu) x^n$$

with $a_0(\nu) = a_0(-\nu) = 1$.

We have already computed $y_1(x)$ and observed that $\mathrm{J}_\nu(x)$ is a constant multiple of $y_1(x)$.

For $r = -\nu$ we have $a_{2m+1}(-\nu) = 0$,

$$\begin{aligned} a_{2m}(-\nu) &= \frac{(-1)^m}{2 \cdot 4 \cdots 2m(2-2\nu)(4-2\nu)\cdots(2m-2\nu)} \\ &= \frac{(-1)^m}{2^{2m} m!(1-\nu)(2-\nu)\cdots(m-\nu)}, \end{aligned}$$

which is defined for all m, since $\nu \notin \mathbb{Z}$. Since $F(-\nu) = 0$, the function $y_2(x)$ defined in this way must then satisfy $L[y_2] = 0$. Moreover, y_1 and y_2 are linearly independent since $y_1(x) \simeq x^{\nu}$, $y_2(x) \simeq x^{-\nu}$ for $x \downarrow 0$.

The case $\nu \notin \mathbb{Z}$ cont'd

Multiplication of $y_2(x)$ with $\frac{2^\nu}{\Gamma(1-\nu)}$ yields $\mathrm{J}_{-\nu}(x)$ (use the functional equation $\Gamma(x+1) = x\,\Gamma(x)$ repeatedly) and shows that in this case the two Bessel functions J_ν, $\mathrm{J}_{-\nu}$ form a fundamental system of solutions.

Remark

For $\nu \in \left\{\frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \dots\right\}$ the number $N = r_1 - r_2 = 2\nu$ is a nonzero integer and Case 4 of our "big theorem" (Case 3 in [BDM17], Th. 5.6.1) applies. Thus it is rather surprising that there is such a simple formula for $y_2(x)$ (the same as in Case 1 of the theorem).

Explanation: Since $N = 2\nu$ is odd in this case, we have $a_N(r) = 0$ and hence $a = \lim_{r\to r_2}(r - r_2)a_N(r) = 0$. Thus the formula for $y_2(x)$ in Case 4 of the theorem contains no logarithmic term. Moreover, all functions $a_n(r)$ are analytic at $r_2 = -\nu$ and hence

$$\alpha_n'(r) = \frac{\mathrm{d}}{\mathrm{d}r}\left[(r - r_2)a_n(r)\right] = a_n(r) + (r - r_2)a_n'(r),$$
$$\alpha_n'(r_2) = a_n(r_2),$$

reducing the formula for $y_2(x)$ to that in Case 1.

The case $\nu = 0$

In this case

$$\mathrm{J}_0(x) = \sum_{m=0}^{\infty} \frac{(-1)^m}{2^{2m}(m!)^2}\, x^{2m} = \sum_{m=0}^{\infty} \frac{(-1)^m}{(m!)^2} \left(\frac{x}{2}\right)^{2m}$$

is a solution.

$\mathrm{J}_0(x)$ is defined for $x \in \mathbb{R}$, as is easily shown using the ratio test. This is also guaranteed by the theorem, because $p(x) = 1/x$ and $q(x) = 1$ have no singularity except $x_0 = 0$.)

Note

J_0 solves the IVP

$$xy'' + y' + xy = 0,\ y(0) = 1,\ y'(0) = 0.$$

According to Case 3 (the case $r_1 = r_2$) of our theorem, there exists a 2nd fundamental solution (linearly independent of J_0) of the form

$$y_2(x) = \mathrm{J}_0(x)\ln x + \sum_{n=1}^{\infty} b_n x^n, \quad b_n = a_n'(0).$$

For $\nu = 0$ the coefficient functions $a_n(r)$ specialize to $a_{2m+1}(r) = 0$ and

$$a_{2m}(r) = \frac{(-1)^m}{(r+2)^2(r+4)^2\cdots(r+2m)^2}.$$

It follows that $a_n'(0) = 0$ for odd n, so that the 2nd summand in $y_2(x)$ is an even function of x, just like $\mathrm{J}_0(x)$.

For even n we use the fact that the *logarithmic derivative* $\mathrm{ld}(f) = f'/f$ (which in the case $f > 0$ coincides with $ln(f')$) satisfies

$$\frac{(f^a g^b)'}{f^a g^b} = \frac{(af^{a-1}f')g^b + f^a(bg^{b-1}g')}{f^a g^b} = a\frac{f'}{f} + b\frac{g'}{g} \quad \text{for } a, b \in \mathbb{Z}.$$

In particular $\mathrm{ld}(fg) = \mathrm{ld}(f) + \mathrm{ld}(g)$ and $\mathrm{ld}(f^a) = a\,\mathrm{ld}(f)$, relations that resemble those of the logarithm.

$$\Longrightarrow \quad \frac{a_{2m}'(r)}{a_{2m}(r)} = m\,\mathrm{ld}(-1) - 2\,\mathrm{ld}(r+2) - 2\,\mathrm{ld}(r+4) - \cdots - 2\,\mathrm{ld}(r+2m)$$

$$= 0 - \frac{2}{r+2} - \frac{2}{r+4} - \cdots - \frac{2}{r+2m} \quad \text{for } m \geq 1$$

$$\Longrightarrow \quad \frac{a_{2m}'(0)}{a_{2m}(0)} = -\left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{m}\right)$$

The numbers $\mathrm{H}_m = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{m}$ are called *harmonic numbers*, because they form the partial sums of the harmonic series. In all we obtain, using $a_{2m}(0) = \frac{(-1)^m}{2^{2m}(m!)^2}$,

$$y_2(x) = \mathrm{J}_0(x)\ln x + \sum_{m=1}^{\infty} \frac{(-1)^{m+1}\mathrm{H}_m}{2^{2m}(m!)^2}\, x^{2m}$$

Another choice for the 2nd fundamental solution is

$$\mathrm{Y}_0(x) = \frac{2}{\pi}\left(y_2(x) + (\gamma - \ln 2)\mathrm{J}_0(x)\right)$$

$$= \frac{2}{\pi}\left[\left(\ln\frac{x}{2} + \gamma\right)\mathrm{J}_0(x) + \sum_{m=1}^{\infty} \frac{(-1)^{m+1}\mathrm{H}_m}{2^{2m}(m!)^2}\, x^{2m}\right],$$

where $\gamma = \lim\limits_{n\to\infty}(\mathrm{H}_n - \ln n) \approx 0.577$ is the *Euler-Mascheroni constant*.

Definition

Y_0 is called *Neumann function* of order 0.

Other names are *Weber function* or *Bessel function of the 2nd kind* of order 0.

In contrast with J_0, the function Y_0 is not analytic at $x = 0$ (not even defined there) and satisfies

$$\mathrm{Y}_0(x) \simeq \frac{2}{\pi} \ln x \quad \text{for } x \downarrow 0.$$

If you want to learn more about J_0 and Y_0 (as well as about Bessel functions in general and many further so-called *special functions*), look for the *Handbook of Mathematical Functions* edited by M. Abramowitz and I. A. Stegun, the classic reference on this topic.

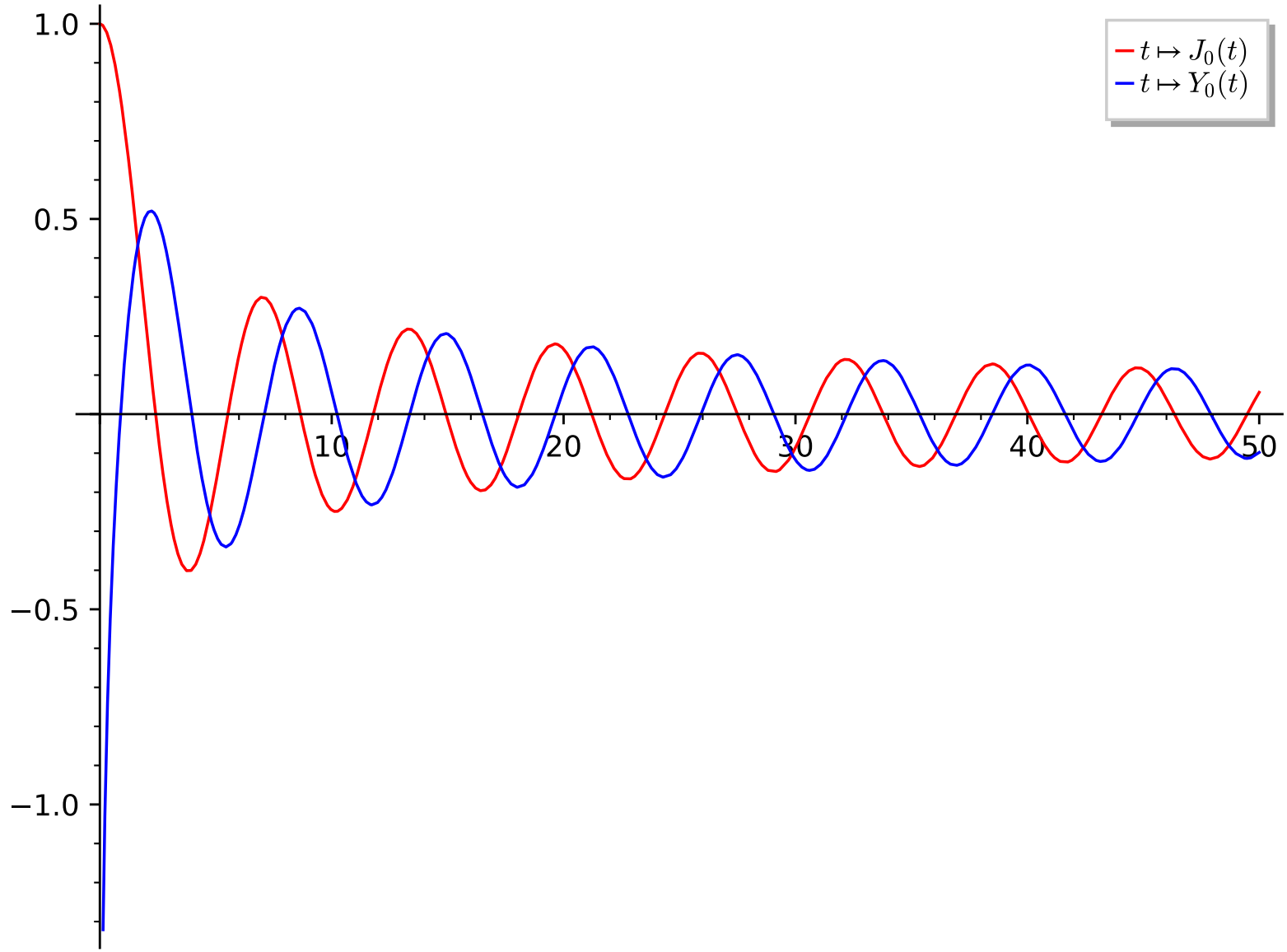


Figure: The Bessel and Neumann functions of order $\nu = 0$ with domain $\mathbb{R}^+$.

The case $\nu \in \mathbb{Z}^+$

In this case the Bessel function J_ν of order ν provides one solution, valid on the whole of $\mathbb{R}$. It is characterized as the unique solution that is analytic at $x_0 = 0$ and has normalization constant $a_0 = \frac{1}{2^\nu \nu!}$.

$$\mathrm{J}_\nu(x) = \sum_{m=0}^{\infty} \frac{(-1)^m}{m!(m+\nu)!} \left(\frac{x}{2}\right)^{\nu+2m} \qquad \text{for } x \in \mathbb{R}.$$

Observe that $\mathrm{J}_\nu(0) = \mathrm{J}'_\nu(0) = \cdots = \mathrm{J}_\nu^{(\nu-1)}(0) = 0$ and $\mathrm{J}_\nu^{(\nu)}(0) = \nu! a_0 = \frac{1}{2^\nu}$.

A second solution $\mathrm{Y}_\nu(x)$, linearly independent of $\mathrm{J}_\nu(x)$, can be obtained in a similar (but increasingly more complicated) way as for $\nu = 0$. Since $N = 2\nu \in \mathbb{Z}^+$, Case 4 of our "big theorem" (Case 3 in [BDM17], Th. 5.6.1) applies, and there is no simplification this time. The case $\nu = 1$ is discussed as part of HW10, Ex. H59.

Remark

The function $\mathrm{J}_{-\nu}$ also solves the Bessel ODE on $\mathbb{R}$, but for $\nu \in \mathbb{Z}^+$ is linearly dependent on J_ν; cf. HW10, Ex. H60 c).

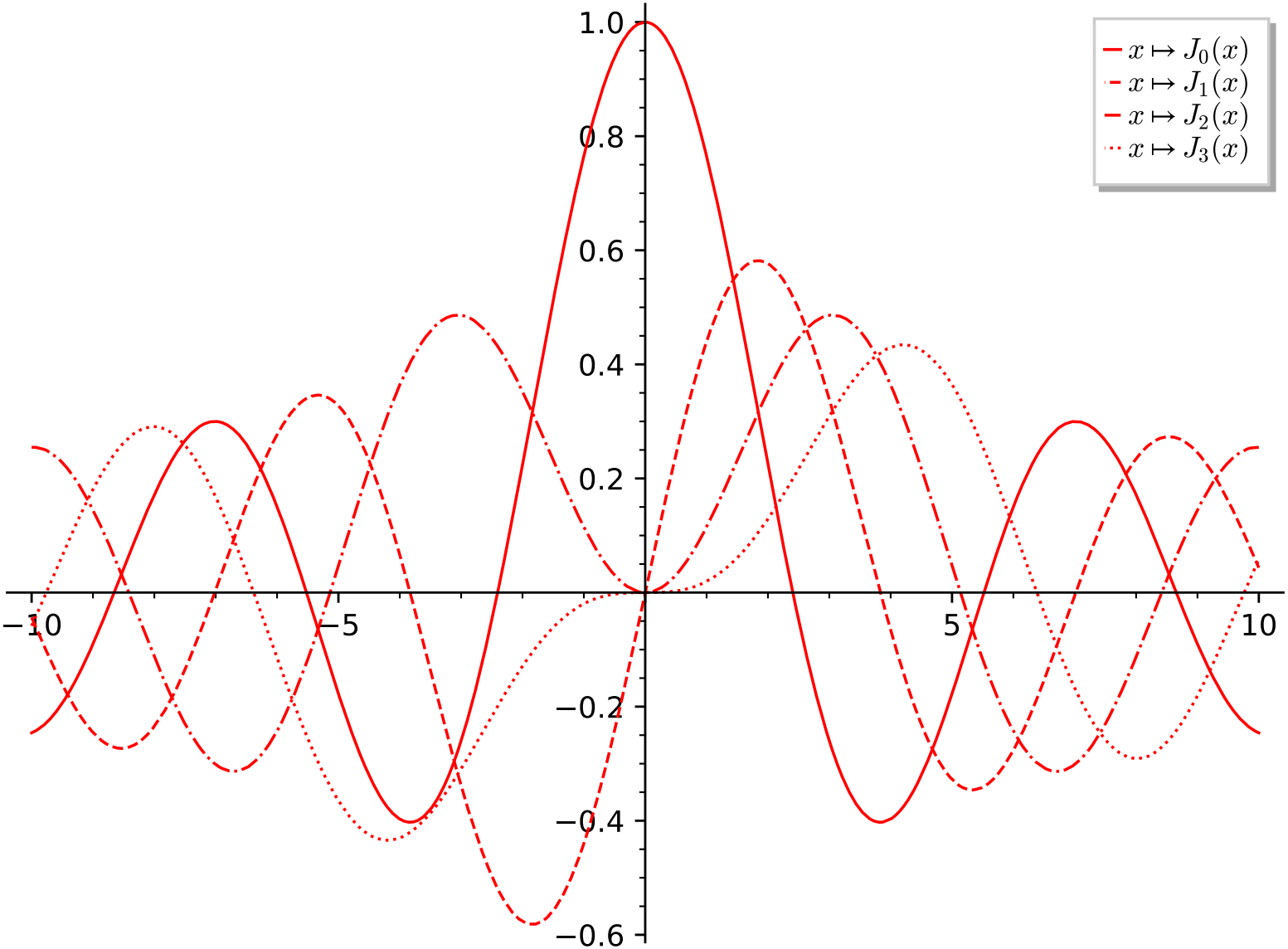


Figure: Bessel functions of various integral orders $\nu \geq 0$ with domain $\mathbb{R}$

The case $\nu = \frac{1}{2}$

This case is in a way special: The fractional power series „Ansatz“ $y(x) = x^{-1/2}\sum_{n=0}^{\infty} a_n x^n$ yields two linearly independent solutions, since a_0 and a_1 can be chosen freely. For this recall that $L\left[\sum_{n=0}^{\infty} a_n x^{n\pm\nu}\right] = x^{\pm\nu}\,(0a_0 + (\pm 2\nu + 1)a_1 x + \cdots)$.
For $(a_0, a_1) = (1, 0)$ the recursion $a_n = -\frac{a_{n-2}}{n(n+2\nu)} = -\frac{a_{n-2}}{n(n-1)}$ yields $a_{2m-1} = 0$, $a_{2m} = \frac{(-1)^m}{(2m)!}$, and hence

$$y(x) = \sum_{m=0}^{\infty} \frac{(-1)^m}{(2m)!} x^{2m-1/2} = \frac{\cos x}{\sqrt{x}}.$$

For $(a_0, a_1) = (0, 1)$ the recursion similarly yields $a_{2m} = 0$, $a_{2m+1} = \frac{(-1)^m}{(2m+1)!}$, and hence

$$y(x) = \sum_{m=0}^{\infty} \frac{(-1)^m}{(2m+1)!} x^{2m+1/2} = \frac{\sin x}{\sqrt{x}}.$$

It follows that $\frac{\cos x}{\sqrt{x}}$, $\frac{\sin x}{\sqrt{x}}$ form a fundamental system of solutions of $x^2y'' + xy' + (x^2 - \frac{1}{4})y = 0$, which can also be verified directly; cf. also HW8, Ex. H49.

The case $\nu = \frac{1}{2}$ cont'd

This case is of course contained in the case $\nu \notin \mathbb{Z}$ considered earlier, which tells us that the Bessel functions $\mathrm{J}_{1/2}$ and $\mathrm{J}_{-1/2}$ form a fundamental system of solutions. The link is best illustrated by computing $\mathrm{J}_{1/2}$ and $\mathrm{J}_{-1/2}$ from the general formula for J_ν :

$$\begin{aligned}
\mathrm{J}_{\frac{1}{2}}(x) &= \sqrt{\frac{x}{2}} \sum_{m=0}^{\infty} \frac{(-1)^m x^{2m}}{m!\,\Gamma(m+\frac{3}{2})2^{2m}} \\
&= \sqrt{\frac{x}{2}} \sum_{m=0}^{\infty} \frac{(-1)^m x^{2m}}{m!\,\Gamma(\frac{1}{2})\frac{3}{2}\frac{5}{2}\cdots\frac{2m+1}{2}2^{2m}} \\
&= \sqrt{\frac{x}{2}} \sum_{m=0}^{\infty} \frac{2(-1)^m x^{2m}}{(2m+1)!\sqrt{\pi}} = \sqrt{\frac{2}{\pi x}}\,\sin x, \\
\mathrm{J}_{-\frac{1}{2}}(x) &= \sqrt{\frac{2}{x}} \sum_{m=0}^{\infty} \frac{(-1)^m x^{2m}}{m!\,\Gamma(m+\frac{1}{2})2^{2m}} = \cdots = \sqrt{\frac{2}{\pi x}}\,\cos x,
\end{aligned}$$

using $\Gamma(x+1) = x\,\Gamma(x)$ and $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

Thus $\mathrm{J}_{1/2}$ and $\mathrm{J}_{-1/2}$ are just scalar multiples of the fundamental solutions previously determined.

An Application of Bessel Functions

Solutions of the 2-dimensional wave equation

Theorem

Suppose $f\colon \mathbb{R}^+ \to \mathbb{C}$ *is a* C^2*-function,* $\nu \in \mathbb{Z}$, $\lambda, c > 0$, $D = \{(x,y,t) \in \mathbb{R}^3; x^2 + y^2 > 0\}$, *and* $u\colon D \to \mathbb{C}$ *is defined by*

$$u(x,y,t) = f(\lambda r)\mathrm{e}^{\mathrm{i}(\nu\phi \pm \lambda c t)}, \quad x = r\cos\phi,\ y = r\sin\phi.$$

Then u *solves the 2-dimensional wave equation,*

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right) u(x,y,t) = 0 \qquad \text{on } D,$$

iff f *solves the Bessel ODE with parameter* ν,

$$s^2 f''(s) + s f'(s) + \left(s^2 - \nu^2\right) f(s) = 0, \qquad s \in \mathbb{R}^+.$$

Notes

1. Solutions having the indicated form arise form the separation ansatz $u(x, y, t) = a(r)b(\phi)c(t)$.
2. The theorem can be used to determine the so-called normal modes of a vibrating circular membrane of radius R, for which u must also be defined and continuous at $(0, 0, t)$, and satisfy the boundary condition

$$u(x, y, t) = 0 \quad \text{if } x^2 + y^2 = R^2.$$

   This is achieved by choosing f as a scalar multiple of J_ν, $\nu = 0, 1, 2, \ldots$, and $\lambda = z_{\nu n}/R$, where $z_{\nu n}$ denotes the n-th positive zero of J_ν. (It can be shown that the positive zeros of J_ν form an infinite sequence $z_{\nu 1} > z_{\nu 2} > z_{\nu 3} > \cdots$.) See https://commons.wikimedia.org/wiki/File:Vibrating_drum_Bessel_function.gif for an animation.
3. The case $\nu = 0$ corresponds to rotation-invariant solutions of the 2-dimensional wave equation. Solutions satisfying the boundary conditions in (2) have the form $u(x, y, t) = \mathrm{J}_0(\lambda r)\left(c_1 \mathrm{e}^{\mathrm{i}\lambda ct} + c_2 \mathrm{e}^{-\mathrm{i}\lambda ct}\right)$, $\lambda = z_{0n}/R$, $c_1, c_2 \in \mathbb{C}$.

Proof of the theorem.

We use the representation of the Laplace operator in polar coordinates (known from an exercise in Calculus III):

$$\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} = \frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial \phi^2}.$$

We have

$$\begin{aligned}\Delta u(x,y,t) &= \mathrm{e}^{\mathrm{i}(\nu\phi \pm \lambda ct)}\left(\frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r}\right) f(\lambda r) + \frac{f(\lambda r)\mathrm{e}^{\pm \mathrm{i}\lambda ct}}{r^2}\frac{\partial^2 \mathrm{e}^{\mathrm{i}\nu\phi}}{\partial \phi^2} \\ &= \mathrm{e}^{\mathrm{i}(\nu\phi \pm \lambda ct)}\left(\lambda^2 f''(\lambda r) + \frac{\lambda}{r} f'(\lambda r) - \frac{\nu^2}{r^2} f(\lambda r)\right),\end{aligned}$$

$$\frac{\partial^2}{\partial t^2} u(x,y,t) = f(\lambda r)\mathrm{e}^{\mathrm{i}\nu\phi}\frac{\partial^2 \mathrm{e}^{\pm \mathrm{i}\lambda ct}}{\partial t^2} = -\lambda^2 c^2 f(\lambda r)\mathrm{e}^{\mathrm{i}(\nu\phi \pm \lambda ct)}.$$

Since $\mathrm{e}^{\mathrm{i}\nu\phi \pm \mathrm{i}\lambda ct} \neq 0$, it follows that $u(x,y,t)$ solves the 2-dimensional wave equation iff

$$\lambda^2 f''(\lambda r) + \frac{\lambda}{r} f'(\lambda r) - \frac{\nu^2}{r^2} f(\lambda r) = -\lambda^2 f(\lambda r).$$

Multiplying this equation by r^2 and setting $s = \lambda r$ gives the Bessel ODE for $f(s)$, as asserted. □

Exercise

Determine a fundamental system of solutions for Bessel's ODE with $\nu = \frac{1}{2}$,

$$y'' + \frac{1}{t}\,y' + \left(1 - \frac{1}{4t^2}\right) y = 0,$$

using the ansatz $z = \sqrt{t}\,y$. Then compare your result with that of the lecture.

Exercise

Determine the general solution of the following ODE's:

a) $(2t+1)y'' + (4t-2)y' - 8y = (6t^2 + t - 3)\mathrm{e}^t, \quad t > -1/2;$

b) $t^2(1-t)y'' + 2t(2-t)y' + 2(1+t)y = t^2, \quad 0 < t < 1.$

Hints: The associated homogeneous ODE in a) has a solution of the form $y(t) = \mathrm{e}^{\alpha t}$ and that in b) a solution of the form $y(t) = t^\beta$ with constants α, β. In both cases a particular solution of the inhomogeneous ODE can be determined by reducing it to a first-order system and using variation of parameters (though this may not be the most economic solution).

Math 285
Introduction to Differential Equations

Thomas Honold

ZJU-UIUC Institute

Spring Semester 2024

Outline

Today's Lecture: The Laplace Transform

Integral Transforms

Integral transforms are maps $f \mapsto F$, which assign to a function f from a certain domain (e.g., the set of integrable functions $f\colon [a,b] \to \mathbb{C}$) another function F of the form

$$F(s) = \int_a^b K(s,t) f(t)\,\mathrm{d}t\,. \tag{IT}$$

Here $K(s,t)$ is a two-variable function called the *kernel* of the integral transform, and $F(s)$ is defined for all $s \in \mathbb{C}$ for which the integral in (IT) exists.

Definition

The *Laplace transform* is the integral transform with $(a,b) = (0,\infty)$ and $K(s,t) = \mathrm{e}^{-st}$, i.e.,

$$F(s) = \int_0^\infty \mathrm{e}^{-st} f(t)\,\mathrm{d}t\,.$$

The Laplace transform will be denoted by $\mathcal{L}$; we will write $F = \mathcal{L}f$ or, making explicit reference to the arguments, $F(s) = \mathcal{L}\{f(t)\}$.

An Appropriate Domain for $\mathcal{L}$

Definition

Let $f\colon [0,\infty) \to \mathbb{C}$ be a function.

1. f is said to be *piecewise continuous* if (i) the set Δ of discontinuities of f is discrete (i.e., has no accumulation point in $\mathbb{R}$), (ii) f is continuous on each connected component of $[0,\infty) \setminus \Delta$ (which must be an interval), and (iii) for every $\alpha \in \Delta$ the one-sided limits $f(\alpha+) = \lim_{t\downarrow\alpha} f(t)$, $f(\alpha-) = \lim_{t\uparrow\alpha} f(t)$ exist (with the obvious adjustment in the case $\alpha = 0$).
2. f is said to be of (at most) *exponential order* for $t \to \infty$ if there exist constants $a \in \mathbb{R}$, $K > 0$, $M > 0$ such that $|f(t)| \leq Ke^{at}$ whenever $t \geq M$.

Piecewise continuous functions of exponential order on $[0,\infty)$ form an appropriate domain for the Laplace transform, i.e., for every such function f the function $\mathcal{L}f$ is well-defined.

Changing the values of f on a discrete subset of $[0,\infty)$ doesn't change $\mathcal{L}f$. It is even sufficient if f is undefined for some $t \in \Delta$.

Notes

- Informally speaking, a piecewise continuous function may have only jump discontinuities; there can be infinitely many jump discontinuities in $[0, \infty)$ (as in the case $t \mapsto \lfloor t \rfloor$), but only finitely many in every bounded subinterval $[0, R]$.
- The condition in Part (2) is equivalent to $f(t) = \mathrm{O}(\mathrm{e}^{at})$ for $t \to \infty$ and should be viewed as a property depending on a. It doesn't necessarily mean "f grows exponentially" (since $a < 0$ is allowed and, moreover, only an upper bound for $|f(t)|$ is given), and becomes stronger if we decrease a. (In fact $a < b$ implies $\mathrm{e}^{at} = \mathrm{o}(\mathrm{e}^{bt})$ for $t \to \infty$.)
- If $f\colon [0, \infty) \to \mathbb{R}$ is of exponential order (i.e., there exists $a \in \mathbb{R}$ such that $f(t) = \mathrm{O}(\mathrm{e}^{at})$ for $t \to \infty$), we can define the *exact exponential order* of f as

$$\mathrm{eo}(f) = \inf\left\{a \in \mathbb{R};\, f(t) = \mathrm{O}(\mathrm{e}^{at}) \text{ for } t \to \infty\right\}.$$

  Using the infimum is necessary, since, e.g., $t^n = \mathrm{O}(\mathrm{e}^{at})$ for $t \to \infty$ whenever $a > 0$, but $t^n \neq \mathrm{O}(1)$. (Thus all nonzero polynomials have exact exponential order 0.)

Examples

1. All nonzero polynomials in $\mathbb{R}[t]$ have exact exponential order 0. The same is true of nonzero rational functions $f(t) = p(t)/q(t)$, $p(t), q(t) \in \mathbb{R}[t] \setminus \{0\}$.
2. $t \mapsto c_1 \mathrm{e}^{a_1 t} + c_2 \mathrm{e}^{a_2 t} + \cdots + c_n \mathrm{e}^{a_n t}$ ($a_1 < a_2 < \cdots < a_n$, $c_i \neq 0$ for $1 \leq i \leq n$) has exact exponential order a_n.
3. $\sin(at)$, $\cos(at)$ for $a \neq 0$ (more generally, non-vanishing trigonometric polynomials) have exact exponential order 0.
4. $t \mapsto \mathrm{e}^{t^2}$ is not of exponential order (or of exact exponential order $+\infty$), because for $t \to \infty$ it increases faster than any exponential function e^{at}.

   On the other hand, the reciprocal function $t \mapsto \mathrm{e}^{-t^2}$ is $\mathrm{O}(\mathrm{e}^{at})$ (and $\mathrm{o}(\mathrm{e}^{at})$ as well) for every $a \in \mathbb{R}$, and hence according to the definition has exact exponential order $-\infty$.

Theorem

Suppose $f : [0, \infty) \to \mathbb{C}$ is piecewise continuous and of exact exponential order $a \in \mathbb{R} \cup \{\pm\infty\}$.

1. *If $a = +\infty$ then $\mathcal{L}(f)$ need not be defined for any $s \in \mathbb{C}$.*
2. *If $a \in \mathbb{R}$ then $\mathcal{L}(f)$ is defined and analytic at least for all s in the open half plane $\mathrm{Re}(s) > a$.*
3. *If $a = -\infty$ then $\mathcal{L}(f)$ is defined and analytic for all $s \in \mathbb{C}$ (a so-called entire function).*

Moreover, in Cases 2 and 3 the Laplace transform $F = \mathcal{L}(f)$ can be differentiated under the integral sign':

$$F'(s) = \int_0^\infty \frac{\mathrm{d}}{\mathrm{d}s} f(t)\mathrm{e}^{-st}\,\mathrm{d}t = -\int_0^\infty t\,f(t)\mathrm{e}^{-st}\,\mathrm{d}t = -\mathcal{L}\{t\,f(t)\}.$$

Notes

- Differentiating repeatedly gives $F^{(n)}(s) = (-1)^n\mathcal{L}\{t^n f(t)\}$ for $n \in \mathbb{N}$.
- In Case 2 it is possible that $\mathcal{L}(f)$ is defined and analytic in a larger region than $\mathrm{Re}(s) > a$; cf. exercises.

Proof.

Since $f = u + \mathrm{i}v$ implies $\mathcal{L}f = \mathcal{L}u + \mathrm{i}\mathcal{L}v$, we may assume w.l.o.g. that f is real-valued.

Suppose $|f(t)| \le K\,\mathrm{e}^{at}$ for $t \ge M$. We claim that $\int_0^\infty f(t)\mathrm{e}^{-st}\,\mathrm{d}t$ converges uniformly (and absolutely) in every closed half plane $\mathrm{Re}(s) \ge a + \delta$, $\delta > 0$. Indeed, for such s and $t \ge M$ we have

$$\left|f(t)\mathrm{e}^{-st}\right| = |f(t)|\,\mathrm{e}^{-\mathrm{Re}(s)t} \le K\,\mathrm{e}^{(a-\mathrm{Re}\,s)t} \le K\,\mathrm{e}^{-\delta t}.$$

Since this bound is independent of s and $\int_M^\infty \mathrm{e}^{-\delta t}\,\mathrm{d}t$ converges, we can apply the Weierstrass test for uniform convergence of improper parameter integrals to conclude that the convergence of $F(s) = \int_0^\infty f(t)\mathrm{e}^{-st}\,\mathrm{d}t$ in $\mathrm{Re}(s) \ge a + \delta$ is uniform. In particular $F(s)$ is defined for all $s \in \mathbb{C}$ with $\mathrm{Re}(s) > a$.

Since $t \mapsto t\,f(t)$ is $\mathrm{O}(\mathrm{e}^{at})$ for $t \to \infty$ as well, the integral $\int_0^\infty \frac{\mathrm{d}}{\mathrm{d}s} f(t)\mathrm{e}^{-st}\,\mathrm{d}t = -\int_0^\infty t\,f(t)\mathrm{e}^{-st}\,\mathrm{d}t$ also converges uniformly in $\mathrm{Re}(s) \ge a + \delta$ for every $\delta > 0$, so that the necessary assumptions for differentiating $F(s)$ under the integral sign (complex version) are satisfied.

$\Longrightarrow$ F is complex differentiable (and hence analytic) in $\mathrm{Re}(s) > a$.

For a proof using only Real Analysis see next slide. □

Note on the proof

Writing $s = x + \mathrm{i}y$ we have

$$\begin{aligned} F(s) = F(x+\mathrm{i}y) &= \int_0^\infty f(t)\mathrm{e}^{-xt-\mathrm{i}yt}\,\mathrm{d}t \\ &= \int_0^\infty f(t)\mathrm{e}^{-xt}\big(\cos(yt) - \mathrm{i}\sin(yt)\big)\,\mathrm{d}t \\ &= \int_0^\infty f(t)\mathrm{e}^{-xt}\cos(yt)\,\mathrm{d}t + \mathrm{i}\int_0^\infty -f(t)\mathrm{e}^{-xt}\sin(yt)\,\mathrm{d}t \\ &= u(x,y) + \mathrm{i}\,v(x,y), \quad \text{say.} \end{aligned}$$

Using this formula and the results on partial differentiation of real-variable functions under the integral sign, one can show that u, v are partially differentiable and satisfy the Cauchy-Riemann equations $u_x = v_y$, $u_y = -v_x$. From this it follows without resort to Complex Analysis that F is complex differentiable; cf. our discussion of real vs. complex differentiability in Calculus III.

Examples

1. $\mathcal{L}\{1\} = \int_0^\infty 1\,\mathrm{e}^{-st}\,\mathrm{d}t = \left[-\frac{1}{s}\mathrm{e}^{-st}\right]_0^\infty = \frac{1}{s}$ for $s > 0$.

   More generally, this holds for $\mathrm{Re}(s) > 0$ since for $s = x + \mathrm{i}y$, $x > 0$, we still have $\mathrm{e}^{-st} = \mathrm{e}^{-xt}\mathrm{e}^{-\mathrm{i}yt} \to 0$ for $t \to \infty$.

2. $\mathcal{L}\{\mathrm{e}^t\} = \int_0^\infty \mathrm{e}^t\,\mathrm{e}^{-st}\,\mathrm{d}t = \int_0^\infty \mathrm{e}^{-(s-1)t}\,\mathrm{d}t = \frac{1}{s-1}$ for $\mathrm{Re}(s) > 1$.

3. $\mathcal{L}\{\cos t\} = \int_0^\infty \frac{1}{2}\left(\mathrm{e}^{\mathrm{i}t} + \mathrm{e}^{-\mathrm{i}t}\right)\mathrm{e}^{-st}\,\mathrm{d}t = \frac{1}{2}\int_0^\infty \mathrm{e}^{-(s-\mathrm{i})t} + \mathrm{e}^{-(s+\mathrm{i})t}\,\mathrm{d}t = \frac{1}{2}\left[\frac{1}{s-\mathrm{i}} + \frac{1}{s+\mathrm{i}}\right] = \frac{1}{2}\,\frac{2s}{(s-\mathrm{i})(s+\mathrm{i})} = \frac{s}{s^2+1}$ for $\mathrm{Re}(s) > 0$.

4. $\mathcal{L}\{\sin t\} = \int_0^\infty \frac{1}{2\mathrm{i}}\left(\mathrm{e}^{\mathrm{i}t} - \mathrm{e}^{-\mathrm{i}t}\right)\mathrm{e}^{-st}\,\mathrm{d}t = \frac{1}{2\mathrm{i}}\left[\frac{1}{s-\mathrm{i}} - \frac{1}{s+\mathrm{i}}\right] = \frac{1}{2\mathrm{i}}\,\frac{2\mathrm{i}}{(s-\mathrm{i})(s+\mathrm{i})} = \frac{1}{s^2+1}$ for $\mathrm{Re}(s) > 0$.

Examples (cont'd)

5

$$\begin{aligned}\mathcal{L}\{t^n\} &= \int_0^\infty t^n \mathrm{e}^{-st}\,\mathrm{d}t \\ &= \int_0^\infty \left(\frac{\tau}{s}\right)^n \mathrm{e}^{-\tau}\frac{\mathrm{d}\tau}{s} \qquad \text{(Subst. } \tau = st,\ \mathrm{d}\tau = s\,\mathrm{d}t\text{)} \\ &= \frac{1}{s^{n+1}}\int_0^\infty \tau^n \mathrm{e}^{-\tau}\,\mathrm{d}\tau = \frac{n!}{s^{n+1}}\end{aligned}$$

for $\mathrm{Re}(s) > 0$.

More generally, we have

$$\mathcal{L}\{t^r\} = \frac{1}{s^{r+1}}\int_0^\infty \tau^r \mathrm{e}^{-\tau}\,\mathrm{d}\tau = \frac{\Gamma(r+1)}{s^{r+1}}$$

for $\mathrm{Re}(s) > 0$, $r > -1$. (It doesn't matter here that $t \mapsto t^r$ isn't defined at $t = 0$ for $-1 < r < 0$.)

In particular, $\mathcal{L}\{t^{-1/2}\} = \Gamma(1/2)s^{-1/2} = \sqrt{\pi}\,s^{-1/2}$, i.e., $t \mapsto 1/\sqrt{t}$ is an eigenfunction of $\mathcal{L}$ for the eigenvalue $\sqrt{\pi}$.

Examples

6. We compute the Laplace transform of the "staircase" function $t \mapsto \lfloor t \rfloor$, which is defined for $\operatorname{Re}(s) > 0$.
   Since $\lfloor t \rfloor = n$ for $t \in [n, n+1)$, we obtain

$$\begin{aligned}
\mathcal{L}\{\lfloor t \rfloor\} &= \int_0^\infty \lfloor t \rfloor \mathrm{e}^{-st}\,\mathrm{d}t = \sum_{n=0}^{\infty} \int_n^{n+1} n\,\mathrm{e}^{-st}\,\mathrm{d}t \\
&= \sum_{n=0}^{\infty} n\left[-\frac{1}{s}\,\mathrm{e}^{-st}\right]_n^{n+1} = \frac{1}{s}\sum_{n=0}^{\infty} n\left(\mathrm{e}^{-ns} - \mathrm{e}^{-(n+1)s}\right) \\
&= \frac{1}{s}\left(\mathrm{e}^{-s} - \mathrm{e}^{-2s} + 2\,\mathrm{e}^{-2s} - 2\,\mathrm{e}^{-3s} + 3\,\mathrm{e}^{-3s} - 3\,\mathrm{e}^{-4s} + \cdots\right) \\
&= \frac{1}{s}\left(\mathrm{e}^{-s} + \mathrm{e}^{-2s} + \mathrm{e}^{-3s} + \cdots\right) \\
&= \frac{1}{s}\,\frac{\mathrm{e}^{-s}}{1-\mathrm{e}^{-s}} = \frac{1}{s\,(\mathrm{e}^{s}-1)}.
\end{aligned}$$

Exercise

Suppose that $f\colon [0,\infty) \to \mathbb{C}$ is piecewise continuous. Show:

1. If $\int_0^\infty f(t)\mathrm{e}^{-st}\,\mathrm{d}t$ converges absolutely for $s = s_0$, it converges absolutely for $\mathrm{Re}(s) \geq \mathrm{Re}(s_0)$.
2. If $\int_0^\infty f(t)\mathrm{e}^{-st}\,\mathrm{d}t$ converges for $s = s_0$, it converges for $\mathrm{Re}(s) > \mathrm{Re}(s_0)$ and for such s satisfies

$$\int_0^\infty f(t)\mathrm{e}^{-st}\,\mathrm{d}t = (s - s_0)\int_0^\infty \phi(t)\mathrm{e}^{-(s-s_0)t}\,\mathrm{d}t$$

   with $\phi(t) = \int_0^t f(\tau)\mathrm{e}^{-s\tau}\mathrm{d}\tau$. Moreover, the integral $\int_0^\infty \phi(t)\mathrm{e}^{-(s-s_0)t}\,\mathrm{d}t$ converges absolutely for $\mathrm{Re}(s) > \mathrm{Re}(s_0)$.
3. There exist numbers $-\infty \leq \beta \leq \alpha \leq \infty$, such that $\int_0^\infty f(t)\mathrm{e}^{-st}\,\mathrm{d}t$ diverges for $\mathrm{Re}(s) < \beta$, converges conditionally (i.e., not absolutely) for $\beta < \mathrm{Re}(s) < \alpha$ and converges absolutely for $\mathrm{Re}(s) > \alpha$. Moreover, on the line $\mathrm{Re}(s) = \alpha$ the Laplace integral converges absolutely either for all s or for no s.
4. $F(s) := \int_0^\infty f(t)\mathrm{e}^{-st}\,\mathrm{d}t$ is analytic in $\mathrm{Re}(s) > \beta$.

The numbers α, β defined in the preceding exercise are called *abscissa of absolute convergence*, resp., *abscissa of convergence* of the Laplace integral $\int_0^\infty f(t)\mathrm{e}^{-st}\,\mathrm{d}t$, and the corresponding lines $\mathrm{Re}(s) = \alpha$, $\mathrm{Re}(s) = \beta$ *line of absolute convergence*, resp., *line of convergence*.

If f has exact exponential order a, we must have $\beta \leq \alpha \leq a$. Both inequalities may be strict. For the second inequality this is shown in the following exercise.

Exercise

Let $f\colon [0,\infty) \to \mathbb{R}$ be defined by

$$f(t) = \begin{cases} \mathrm{e}^n & \text{if } |t-n| < \mathrm{e}^{-2n} \text{ for some } n \in \mathbb{Z}^+, \\ 0 & \text{otherwise.} \end{cases}$$

Show that f has exact exponential order 1, but $\int_0^\infty \mathrm{e}^{-st}\,\mathrm{d}t$ converges (absolutely) for $\mathrm{Re}(s) = 0$.

Linearity

Suppose $\mathcal{L}f_1$ is defined for $\mathrm{Re}(s) > a_1$ and $\mathcal{L}f_2$ for $\mathrm{Re}(s) > a_2$. Then for any $c_1, c_2 \in \mathbb{C}$ the function $\mathcal{L}(c_1 f_1 + c_2 f_2)$ is defined for $\mathrm{Re}(s) > \max\{a_1, a_2\}$ and satisfies

$$\mathcal{L}\{c_1 f_1(t) + c_2 f_2(t)\} = c_1 \mathcal{L}\{f_1(t)\} + c_2 \mathcal{L}\{f_2(t)\}.$$

The proof is trivial.

As an application of linearity, we get from $\mathcal{L}\{t^n\} = n!/s^{n+1}$ the Laplace transform of any polynomial:

$$\mathcal{L}\left\{c_0 + c_1 t + c_2 t^2 + \cdots + c_d t^d\right\} = \frac{c_0}{s} + \frac{c_1}{s^2} + \frac{c_2\, 2!}{s^3} + \cdots + \frac{c_d\, d!}{s^{d+1}}$$

or, writing $a_n = n! c_n$,

$$\mathcal{L}\left\{\frac{a_0}{0!} + \frac{a_1}{1!}\, t + \cdots + \frac{a_d}{d!}\, t^d\right\} = \frac{a_0}{s} + \frac{a_1}{s^2} + \cdots + \frac{a_d}{s^{d+1}},$$

valid in the right half plane $\mathrm{Re}(s) > 0$.

Exercise

1. Suppose $f_i : [0,\infty) \to \mathbb{C}$ are piecewise continuous and of exponential order a_i $(i = 1, 2)$. Show that the product $f_1 f_2 : [0,\infty) \to \mathbb{R}$, $t \mapsto f_1(t) f_2(t)$ is piecewise continuous and of exponential order $a_1 + a_2$ (and hence $\mathcal{L}(f_1 f_2)$ is defined for $\mathrm{Re}(s) > a_1 + a_2$).

2. Suppose $f : [0,\infty) \to \mathbb{C}$ is piecewise continuous and of exponential order a. Show that $g : [0,\infty) \to \mathbb{R}$ defined by

$$g(t) = \int_0^t f(\tau)\mathrm{d}\tau$$

is continuous and of exponential order $\max\{a, 0\}$.

Hint: Show first f satisfies a bound $|f(t)| \leq K\,\mathrm{e}^{at}$ for all $t \geq 0$.

3. Show, by way of a counterexample, that a piecewise continuous function $f : [0,\infty) \to \mathbb{C}$ of exponential order may have a piecewise continuous derivative f' that is not of exponential order.

Hint: Compose a suitable function with $t \mapsto \mathrm{e}^{t^2}$.

Dilations in the argument

Suppose $F(s) = \mathcal{L}\{f(t)\}$ f is defined for $\operatorname{Re}(s) > a$ and $r > 0$. Then $\mathcal{L}\{f(rt)\}$ is defined for $\operatorname{Re}(s) > ra$ and satisfies

$$\mathcal{L}\{f(rt)\} = \frac{1}{r} F\left(\frac{s}{r}\right)$$

Proof.

$$\begin{aligned}
\mathcal{L}\{f(rt)\} &= \int_0^\infty f(rt) \mathrm{e}^{-st}\,\mathrm{d}t \\
&= \frac{1}{r}\int_0^\infty f(\tau)\mathrm{e}^{-s\tau/r}\mathrm{d}\tau \qquad \text{(Subst. } \tau = rt,\ \mathrm{d}\tau = r\,\mathrm{d}t\text{)} \\
&= \frac{1}{r} F\left(\frac{s}{r}\right).
\end{aligned}$$

Example

From $\mathcal{L}\{\cos t\} = \frac{s}{s^2+1}$, $\mathcal{L}\{\sin t\} = \frac{1}{s^2+1}$ we get

$$\mathcal{L}\{\cos(\omega t)\} = \frac{1}{\omega}\,\frac{s/\omega}{(s/\omega)^2+1} = \frac{s}{s^2+\omega^2},$$

$$\mathcal{L}\{\sin(\omega t)\} = \frac{1}{\omega}\,\frac{1}{(s/\omega)^2+1} = \frac{\omega}{s^2+\omega^2}.$$

Remark

The dilation formula can also be stated as $F(rs) = \mathcal{L}\{\frac{1}{r} f(t/r)\}$ for $\mathrm{Re}(s) > a/r$. To see this, multiply the original dilation formula by r, use linearity of $\mathcal{L}$, and replace r by $1/r$.

The next example combines the dilation property with linearity.

Example

Find $\mathcal{L}\{\cos^2 t\}$ and $\mathcal{L}\{\sin^2 t\}$.

Solution: We have $\cos(2t) = \cos^2 t - \sin^2 t = 2\cos^2 t - 1$ and hence $\cos^2 t = \frac{1+\cos(2t)}{2}$.

$$\begin{aligned}
\Longrightarrow \mathcal{L}\{\cos^2 t\} &= \frac{1}{2}\left(\mathcal{L}\{1\} + \mathcal{L}\{\cos(2t)\}\right) = \frac{1}{2}\left(\frac{1}{s} + \frac{s}{s^2+4}\right) \\
&= \frac{s^2+2}{s(s^2+4)}, \\
\mathcal{L}\{\sin^2 t\} &= \mathcal{L}\{1\} - \mathcal{L}\{\cos^2 t\} = \frac{1}{s} - \frac{s^2+2}{s(s^2+4)} \\
&= \frac{2}{s(s^2+4)}.
\end{aligned}$$

The Laplace transform is also well-behaved w.r.t. translations of the argument t, but the corresponding property is more technical to state. For $c \in \mathbb{R}$ define the "unit step function" $\mathrm{u}_c \colon \mathbb{R} \to \mathbb{R}$ by

$$\mathrm{u}_c(t) = \mathrm{u}(t-c) = \begin{cases} 0 & \text{for } t < c, \\ 1 & \text{for } t \geq c \end{cases}$$

($\mathrm{u}_0(t) = \mathrm{u}(t)$ is the familiar *Heaviside function*), and use this to define, for any function $f \colon [0,\infty) \to \mathbb{C}$ a new function $g \colon \mathbb{R} \to \mathbb{C}$ by

$$g(t) = \mathrm{u}_c(t) f(t-c) = \begin{cases} 0 & \text{for } t < c, \\ f(t-c) & \text{for } t \geq c. \end{cases}$$

Here we use the convention "$0 \times$ undefined $= 0$".

Translations in the argument (cf. [BDM17], Th. 6.3.1)

Suppose $F(s) = \mathcal{L}\{f(t)\}$ is defined for $\mathrm{Re}(s) > a$ and $c > 0$. Then

$$\mathcal{L}\{\mathrm{u}_c(t) f(t-c)\} = \mathrm{e}^{-cs} F(s) \quad \text{for } \mathrm{Re}(s) > a.$$

The assumption $c > 0$ guarantees that $g(t) = \mathrm{u}_c(t) f(t-c)$ vanishes on $(-\infty, 0)$, i.e., we can view it as a function on $[0,\infty)$.

Proof.

$$\begin{aligned}\mathcal{L}\{u_c(t)f(t-c)\} &= \int_c^\infty f(t-c)e^{-st}\,dt \\ &= \int_0^\infty f(\tau)e^{-s(\tau+c)}d\tau \\ &\qquad\qquad \text{(Subst. } \tau = t-c,\ d\tau = dt\text{)} \\ &= e^{-sc}\int_0^\infty f(\tau)e^{-s\tau}d\tau = e^{-cs}F(s). \quad \square\end{aligned}$$

Remark

The corresponding translation formula for $F(s)$ is

$$F(s-c) = \mathcal{L}\{e^{ct}f(t)\} \quad \text{for } \mathrm{Re}(s) > a + \mathrm{Re}(c).$$

Here c can be any complex number. This follows immediately from $\mathcal{L}\{e^{ct}f(t)\} = \int_0^\infty e^{ct}f(t)e^{-st}\,dt = \int_0^\infty f(t)e^{-(s-c)t}\,dt$.

Example

For $c > 0$ we have $\mathcal{L}\{u_c(t)\} = e^{-cs}/s$, valid for $\mathrm{Re}(s) > 0$. This follows by taking $f(t) \equiv 1$ in the first translation formula.

Example

The "ceiling" function $f(t) = \lceil t \rceil$ and the "floor" function $g(t) = \lfloor t \rfloor$ are related by $g(t) = u_1(t)f(t-1)$ (picture?). It follows that their Laplace transforms $F(s)$, resp., $G(s)$ are related by $G(s) = \mathrm{e}^{-s}F(s)$. $\Longrightarrow$ The Laplace transform of the "ceiling" function is $F(s) = \mathrm{e}^{s}G(s) = \frac{\mathrm{e}^{s}}{s(\mathrm{e}^{s}-1)}$; cf. previous example.

Example

Earlier we have shown that $\mathcal{L}\{t^n\} = n!/s^{n+1}$ for $n \in \mathbb{N}$. The preceding remark gives, for $\mathrm{Re}(s) > \mathrm{Re}(c)$,

$$\mathcal{L}\{t^n \mathrm{e}^{ct}\} = \frac{n!}{(s-c)^{n+1}} \quad \text{or} \quad \frac{1}{(s-c)^{n+1}} = \mathcal{L}\left\{\frac{t^n}{n!}\,\mathrm{e}^{ct}\right\}.$$

Together with the partial fraction expansion of rational functions and linearity of $\mathcal{L}$ this shows (at least in principle) how to find for any rational function $F(s) = P(s)/Q(s)$ without polynomial part (i.e., $\deg P < \deg Q$) a corresponding function $f(t)$ such that $\mathcal{L}\{f(t)\} = F(s)$. In fact, the function $f(t)$ obtained in this way will be an exponential polynomial (and, conversely, the Laplace transform of any exponential polynomial is a rational function without polynomial part).

Term-wise Integration of Laplace Integrals

From $\mathcal{L}\{t^n\} = n!/s^{n+1}$ and linearity of $\mathcal{L}$ it follows that

$$\mathcal{L}\left\{\frac{a_0}{0!} + \frac{a_1}{1!}\,t + \cdots + \frac{a_d}{d!}\,t^d\right\} = \frac{a_0}{s} + \frac{a_1}{s^2} + \cdots + \frac{a_d}{s^{d+1}}.$$

Under a suitable assumption on the growth of the coefficients, this can be extended to power series (i.e., functions $f(t)$ analytic at $t = 0$). Writing power series $\sum_{n=0}^{\infty} b_n t^n$ as exponential generating functions (i.e., $b_n = a_n/n!$ or $a_n = b_n n!$) makes it apparent that the Laplace transform of a power series in t is a power series in $1/s$.

Theorem

Suppose $\sum_{n=0}^{\infty} a_n z^n$ has radius of convergence $R > 0$. Then $f(t) = \sum_{n=0}^{\infty} (a_n/n!)t^n$ is defined for all $t \geq 0$ and we have

$$\mathcal{L}\{f(t)\} = \mathcal{L}\left\{\sum_{n=0}^{\infty} \frac{a_n}{n!}\,t^n\right\} = \sum_{n=0}^{\infty} \frac{a_n}{s^{n+1}} \quad \textit{for} \quad \mathrm{Re}(s) > 1/R.$$

Note that, by definition of R, the series $\sum_{n=0}^{\infty} \frac{a_n}{s^{n+1}}$ converges for $\mathrm{Re}(s) > 1/R$ (even for all $s \in \mathbb{C}$ with $|s| > 1/R$).

Proof.

Since $R = \sup\{r \geq 0;\, |a_n|\, r^n \text{ is bounded}\} > 0$, there exists for any $r \in (0, R)$ a corresponding constant K such that $|a_n|\, r^n \leq K$ for all n, i.e., $\sqrt[n]{|a_n|} \leq \sqrt[n]{K}/r$ for all n. But then we must have $\sqrt[n]{|a_n|/n!} \to 0$ for $n \to \infty$, so that $\sum_{n=0}^{\infty}(a_n/n!)z^n$ has radius of convergence ∞ and $f(t)$ is defined in particular for all $t \geq 0$.

Moreover,

$$|f(t)| \leq \sum_{n=0}^{\infty} \frac{|a_n|}{n!}\, t^n \leq \sum_{n=0}^{\infty} \frac{K\, r^{-n}}{n!}\, t^n = K\, \mathrm{e}^{t/r} \quad \text{for } 0 < r < R,$$

showing that f has exponential order at most $1/R$, so that $\mathcal{L}\{f(t)\}$ is defined for $\mathrm{Re}(s) > 1/R$.

Writing $f_n(t) = \sum_{k=0}^{n}(a_k/k!)t^k$, the claimed identity takes the form

$$\int_0^{\infty} f(t)\mathrm{e}^{-st}\,\mathrm{d}t = \int_0^{\infty} \lim_{n\to\infty} f_n(t)\mathrm{e}^{-st}\,\mathrm{d}t = \lim_{n\to\infty} \int_0^{\infty} f_n(t)\mathrm{e}^{-st}\,\mathrm{d}t$$

for $\mathrm{Re}(s) > 1/R$.

Proof cont'd.

We prove it directly without resorting to convergence theorems for Lebesgue or improper Riemann integrals. Writing $s = x + \mathrm{i}y$, we have

$$\left|\int_0^\infty f(t)\mathrm{e}^{-st}\,\mathrm{d}t - \int_0^\infty f_n(t)\mathrm{e}^{-st}\,\mathrm{d}t\right| = \left|\int_0^\infty \big(f(t) - f_n(t)\big)\mathrm{e}^{-st}\,\mathrm{d}t\right|$$

$$= \left|\int_0^\infty \sum_{k=n+1}^\infty \frac{a_k}{k!}\,t^k\mathrm{e}^{-st}\,\mathrm{d}t\right| \le \sum_{k=n+1}^\infty \frac{|a_k|}{k!}\int_0^\infty t^k\mathrm{e}^{-xt}\,\mathrm{d}t$$

$$= \sum_{k=n+1}^\infty \frac{|a_k|}{k!}\,\frac{k!}{x^{k+1}} = \sum_{k=n+1}^\infty \frac{|a_k|}{x^{k+1}}.$$

As long as $x = \mathrm{Re}(s) > 1/R$ this converges to zero for $n \to \infty$, because $\sum_{n=0}^\infty |a_n|\,z^n$ has the same radius of convergence as $\sum_{n=0}^\infty a_n z^n$. This completes the proof of the theorem. □

Example

Consider the function $f(t) = \frac{\sin t}{t}$, $t \in [0, \infty)$ (extended continuously to $t = 0$ by defining $f(0) = 1$).

The Laplace transform of f is

$$F(s) = \mathcal{L}\left\{\frac{\sin t}{t}\right\} = \int_0^\infty \frac{\sin t}{t}\, \mathrm{e}^{-st}\, \mathrm{d}t, \quad \mathrm{Re}(s) > 0.$$

We have met $F(s)$ before (in our Calculus III final exam) and evaluated it using integration by parts.

Using the preceding theorem and the Taylor series of $\frac{\sin t}{t}$, we can determine F in a more conceptual way:

$$F(s) = \mathcal{L}\left\{\sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!}\, t^{2n}\right\} = \mathcal{L}\left\{\sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1}\, \frac{t^{2n}}{(2n)!}\right\}$$

$$= \sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1}\, \frac{1}{s^{2n+1}} = \frac{1}{s} - \frac{1}{3\,s^3} + \frac{1}{5\,s^5} \mp \cdots = \arctan(1/s),$$

for $\mathrm{Re}(s) > 1$, since the arctan series has radius of convergence 1.

Example (cont'd)

The extension of the identity $F(s) = \arctan(1/s) = \operatorname{arccot}(s)$ to the whole right half plane $H_0 = \{s \in \mathbb{C}; \operatorname{Re}(s) > 0\}$ is then a consequence of the fact that both $F(s)$ and $\arctan(s)$ are analytic in H_0 and coincide on a subset of H_0, viz.
$H_1 = \{s \in \mathbb{C}; \operatorname{Re}(s) > 1\}$, which has an accumulation point in H_0 (in fact all points of H_1 are accumulation points).

However, the delicate argument required to evaluate $\int_0^\infty \frac{\sin t}{t}\,\mathrm{d}t$ (using continuity of $[0,\infty) \to \mathbb{R}$, $s \mapsto F(s)$ in $s = 0$, which can't be derived from the results on the Laplace transform established so far) is not facilitated in any way by the present discussion.

Exercise

Find the Laplace transform of the Bessel function J_0.

Hint: The power series expansion

$$\frac{1}{\sqrt{1-4x}} = \sum_{n=0}^{\infty} \binom{2n}{n} x^n, \quad \text{valid for } |x| < 1/4,$$

may help (but you should prove it first).

Exercise

1. Show that

$$\int_0^\infty \mathrm{e}^{-t} \ln t \, \mathrm{d}t = -\gamma = -0.577\ldots$$

For this recall that the Euler-Mascheroni constant γ was defined as $\gamma = \lim_{n\to\infty} \left(1 + \frac{1}{2} + \cdots + \frac{1}{n} - \ln n\right)$

Hint: Relate the integral to the Gamma function. Gauss's formula

$$\Gamma(x) = \lim_{n\to\infty} \frac{n!\, n^x}{x(x+1)\cdots(x+n)} \qquad (x \neq 0, -1, -2, \ldots),$$

which you don't need to prove, may help.

2. Use a) to find the Laplace transform of $t \mapsto \ln t$ and the inverse Laplace transform of $s \mapsto \frac{\ln s}{s}$ ($\operatorname{Re} s > 0$).

The Laplace Transform and Differentiation

The formula for $F'(s)$ was already stated and proved:

Differentiation in the Codomain

If $F(s) = \mathcal{L}\{f(t)\}$ is defined for $\operatorname{Re}(s) > a$, it is analytic (complex differentiable) for $\operatorname{Re}(s) > a$ with

$$F'(s) = \mathcal{L}\{-t\,f(t)\}.$$

Example

We use this formula to give an alternative derivation of $\mathcal{L}\{t^k\,\mathrm{e}^{ct}\} = k!/(s-c)^{k+1}$ for $k \in \mathbb{N}$:

$$\frac{1}{s-c} = \mathcal{L}\{\mathrm{e}^{ct}\}, \qquad \text{(from } \mathrm{e}^{ct}\mathrm{e}^{-st} = \mathrm{e}^{-(s-c)t}\text{)}$$

$$\Longrightarrow \quad \frac{1}{(s-c)^2} = -\frac{\mathrm{d}}{\mathrm{d}s}\,\frac{1}{s-c} = -\mathcal{L}\{-t\,\mathrm{e}^{ct}\} = \mathcal{L}\{t\,\mathrm{e}^{ct}\},$$

$$\Longrightarrow \quad \frac{2}{(s-c)^3} = -\frac{\mathrm{d}}{\mathrm{d}s}\,\frac{1}{(s-c)^2} = -\mathcal{L}\{-t^2\,\mathrm{e}^{ct}\} = \mathcal{L}\{t^2\,\mathrm{e}^{ct}\},$$

etc.

The following formula provides the key to applying the Laplace transform to the solution of (time-independent) linear ODE's.

Theorem (Differentiation in the Domain)

Suppose that $f\colon [0,\infty) \to \mathbb{C}$ *is continuous with piece-wise continuous derivative* f'*, and* $F(s) = \mathcal{L}\{f(t)\}$ *is defined for* $\mathrm{Re}(s) > a$*. Then we have*

$$\mathcal{L}\{f'(t)\} = s\,F(s) - f(0) \quad \textit{for } \mathrm{Re}(s) > a.$$

If f is continuous in $(0,\infty)$ but discontinuous in 0, the formula still holds with $f(0)$ replaced by $f(0+)$, i.e., $\mathcal{L}\{f'(t)\} = s\,F(s) - f(0+)$.

Proof.

Assume first that f is a C^1-function. Then integration by parts gives

$$\begin{aligned}\mathcal{L}\{f'(t)\} &= \int_0^\infty f'(t)\mathrm{e}^{-st}\,\mathrm{d}t = \big[f(t)\mathrm{e}^{-st}\big]_0^\infty + s\int_0^\infty f(t)\mathrm{e}^{-st}\,\mathrm{d}t \\ &= s\,F(s) - f(0),\end{aligned}$$

since $|f(t)| \le K\,\mathrm{e}^{at}$ for $t \ge M$ and hence $\big|f(t)\mathrm{e}^{-st}\big| \le K\,\mathrm{e}^{-(\mathrm{Re}\,s-a)t}$, which tends to zero for $t \to \infty$ on account of $\mathrm{Re}(s) - a > 0$.

Proof cont'd.

Next assume that f' has finitely many discontinuities $t_1 < t_2 < \cdots < t_n$. Then we can apply integration by parts to the C^1-functions $f|_{[0,t_1]}$, $f|_{[t_{k-1},t_k]}$ for $2 \le k \le n$, $f_{[t_n,\infty)}$, and obtain

$$\int_0^{t_1} f'(t)e^{-st}\,dt = f(t_1)e^{-st_1} - f(0) + s\int_0^{t_1} f(t)e^{-st}\,dt,$$

$$\int_{t_{k-1}}^{t_k} f'(t)e^{-st}\,dt = f(t_k)e^{-st_k} - f(t_{k-1})e^{-st_{k-1}} + s\int_{t_{k-1}}^{t_k} f(t)e^{-st}\,dt,$$

$$\int_{t_n}^{\infty} f'(t)e^{-st}\,dt = \lim_{t\to\infty} f(t)e^{-st} - f(t_n)e^{-st_n} + s\int_{t_n}^{\infty} f(t)e^{-st}\,dt\,.$$

Since $\lim_{t\to\infty} f(t)e^{-st} = 0$ (as shown above), summing these identities yields again $\mathcal{L}\{f'(t)\} = s\,\mathcal{L}\{f(t)\} - f(0)$.

Finally, if f' has countably many discontinuities $t_1 < t_2 < \ldots$, the preceding argument remains valid (now involving an infinite summation).

The generalization to functions f discontinuous at $t = 0$ follows by changing $f(0)$ to $f(0+)$, which makes f continuous in 0 but doesn't change $\mathcal{L}f$. □

Corollary

Suppose that $f\colon [0,\infty) \to \mathbb{C}$ *is a* C^{n-1}*-function with piece-wise continuous n-th derivative* $f^{(n)}$*, and* $F(s) = \mathcal{L}\{f(t)\}$ *is defined for* $\mathrm{Re}(s) > a$*. Then we have*

$$\mathcal{L}\{f^{(n)}(t)\} = s^n F(s) - s^{n-1} f(0) - s^{n-2} f'(0) - \cdots - s^0 f^{(n-1)}(0)$$

for $\mathrm{Re}(s) > a$*.*

Again the continuity assumption on f, f', $\ldots$, $f^{(n-1)}$ at $t = 0$ can be dropped, if one uses $f^{(k)}(0+)$ in place of $f^{(k)}(0)$ in the formula.

Proof.

Use the theorem and induction on n. □

Remarks

In the theorem and its corollary, the derivatives f' resp. $f^{(n)}$ may be undefined on a discrete subset $\Delta \subset [0,\infty)$; cf. the previous note about this generalization of piecewise continuity. In fact one can show that for a differentiable 1-variable function g the derivative g' cannot have jump discontinuities. Hence if the one-sided limits $g'(t_0\pm)$ exist but are different, $g'(t_0)$ cannot exist. Also, if f is of exponential order a, the derivatives need not be of exponential order, but their Laplace integrals nevertheless exist for $\mathrm{Re}(s) > a$.

The Laplace Transform and Integration

If $f\colon [0,\infty) \to \mathbb{C}$ is piecewise continuous then $g(t) = \int_0^t f(\tau)\mathrm{d}\tau$ is defined for $t \in [0,\infty)$, continuous on $[0,\infty)$, differentiable everywhere except for the discontinuities of f, and at discontinuities t_k of f the one-sided derivatives $g'(t_k\pm)$ still exist.

Integration in the Domain

Suppose $F(s) = \mathcal{L}\{f(t)\}$ is defined for $\mathrm{Re}(s) > a$. Then

$$\mathcal{L}\left\{\int_0^t f(\tau)\mathrm{d}\tau\right\} = \frac{F(s)}{s} \quad \text{for } \mathrm{Re}(s) > \max\{a,0\}.$$

The possible additional singularity of $\mathcal{L}\left\{\int_0^t f(\tau)\mathrm{d}\tau\right\}$ at $s=0$ can be explained as follows: For $f(t) = \mathrm{e}^{-t}$ we have $F(s) = \frac{1}{s+1}$, $\mathrm{Re}(s) > -1$, and $\mathcal{L}\left\{\int_0^t f(\tau)\mathrm{d}\tau\right\} = \mathcal{L}\left\{\int_0^t \mathrm{e}^{-\tau}\mathrm{d}\tau\right\} = \mathcal{L}\left\{1-\mathrm{e}^{-t}\right\} = \frac{1}{s} - \frac{1}{s+1} = \frac{1}{s(s+1)}$, valid only for $\mathrm{Re}(s) > 0$. A new singularity at $s=0$ is introduced, since a constant C of integration has Laplace transform C/s.

Proof.

Let $g(t) = \int_0^t f(\tau)\mathrm{d}\tau$ and $G(s) = \mathcal{L}\{g(t)\}$. The function g is continuous on $[0,\infty)$, and from $|f(t)| \le K\,\mathrm{e}^{at}$ for $t \ge M$ we obtain

$$|g(t)| = \left|\int_0^M f(\tau)\mathrm{d}\tau + \int_M^t f(\tau)\mathrm{d}\tau\right| \le |g(M)| + \int_M^t |f(\tau)|\,\mathrm{d}\tau$$

$$\le |g(M)| + \int_M^t K\,\mathrm{e}^{a\tau}\,\mathrm{d}\tau = |g(M)| + \frac{K}{a}\left(\mathrm{e}^{at} - \mathrm{e}^{aM}\right)$$

$$= |g(M)| - \frac{K}{a}\,\mathrm{e}^{aM} + \frac{K}{a}\,\mathrm{e}^{at}$$

for $t \ge M$. Clearly this implies that $g(t)$ is of exponential order at most $\max\{a, 0\}$, so that $G(s)$ isdefined for $\mathrm{Re}(s) > \max\{a, 0\}$.

Applying differentiation in the domain gives

$$F(s) = \mathcal{L}\{f(t)\} = \mathcal{L}\{g'(t)\} = s\,G(s) - g(0) = s\,G(s),$$

i.e., $G(s) = F(s)/s$. □

Integration in the Codomain

Suppose $F(s) = \mathcal{L}\{f(t)\}$ is defined for $\mathrm{Re}(s) > a$ and $\int_0^1 \frac{f(t)}{t}\,\mathrm{d}t$ exists. Then

$$\mathcal{L}\left\{\frac{f(t)}{t}\right\} = \int_s^\infty F(\sigma)\mathrm{d}\sigma \quad \text{for } s > a.$$

The condition on the existence of $\int_0^1 \frac{f(t)}{t}\,\mathrm{d}t$ is satisfied in particular if $f(0) = 0$ and $f'(0+) = \lim_{t\downarrow 0}\frac{f(t)}{t}$ exists, but also if there exists $r > 0$ such that $f(t) \simeq t^r$ for $t \downarrow 0$.

The formula remains true for complex numbers s with $\mathrm{Re}(s) > a$, provided we replace $\int_s^\infty F(\sigma)\mathrm{d}\sigma$ by $\int_0^\infty F(s+\sigma)\mathrm{d}\sigma$ (or, more generally, as the complex line integral of $F(s)$ along any ray emanating from s and contained in the half plane $\mathrm{Re}(s) > a$).

Proof.

Since $g(t) = f(t)/t$ has the same exponential order as f, the Laplace transform $G(s)$ of g is defined for $\mathrm{Re}(s) > a$ if $\int_0^1 g(t)\mathrm{e}^{-st}\,\mathrm{d}t$ exists for such s. The latter is equivalent to the existence of $\int_0^1 g(t)\,\mathrm{d}t$, which is true by assumption.

The formula an then be proved as follows:

$$G'(s) = -\mathcal{L}\{t\,g(t)\} = -\mathcal{L}\{f(t)\} = -F(s)$$

$$\Longrightarrow\ G(s) = G(s_0) - \int_{s_0}^{s} F(\sigma)\mathrm{d}\sigma = G(s_0) + \int_{s}^{s_0} F(\sigma)\mathrm{d}\sigma$$

for $s_0, s > a$. Letting $s_0 \to \infty$, we obtain $G(s) = \int_s^\infty F(\sigma)\mathrm{d}\sigma$ using the known fact $\lim_{s_0\to\infty} G(s_0) = 0$; cf. exercise. □

Exercise

Suppose $F(s) = \mathcal{L}\{f(t)\}$ is defined for $\mathrm{Re}(s) > a$, $a \in [-\infty, 0)$. Show that $\lim_{s\to\infty} F(s) = 0$; cp. Exercise 24 in [BDM17], Ch. 6.1. This implies, e.g., that no nonzero polynomial can be a Laplace transform.

Hint: Use the uniform convergence of $\int_0^\infty f(t)\mathrm{e}^{-st}$ on $\mathrm{Re}(s) \geq a + 1$ (resp., for $a = -\infty$ on $\mathrm{Re}(s) \geq 0$).

Example

From $\mathcal{L}\{\sin t\} = \frac{1}{s^2+1}$, using integration in the codomain, we find again

$$\mathcal{L}\left\{\frac{\sin t}{t}\right\} = \int_s^\infty \frac{\mathrm{d}\sigma}{\sigma^2+1} = \frac{\pi}{2} - \arctan s = \operatorname{arccot} s \quad \text{for } s > 0.$$

From this in turn, using integration in the domain, we can compute the Laplace transform of the sine integral:

$$\mathcal{L}\{\operatorname{Si} t\} = \mathcal{L}\left\{\int_0^t \frac{\sin\tau}{\tau}\,\mathrm{d}\tau\right\} = \frac{\operatorname{arccot} s}{s} \quad \text{for } s > 0.$$

As remarked before, these formulas also hold for $s \in \mathbb{C}$ with $\operatorname{Re}(s) > 0$.

The Laplace Transform and Convolution

We have seen that the Lapace transform of a sum of two functions is the sum of their Lapace transforms. How about their product? For the product it is not true, since
$1/s = \mathcal{L}\{1\} = \mathcal{L}\{1^2\} \neq \mathcal{L}\{1\}^2 = 1/s^2$ as functions.

But we can try to determine a different product $(f, g) \mapsto f * g$ that satisfies $\mathcal{L}\{f * g\} = \mathcal{L}\{f\} \cdot \mathcal{L}\{g\}$. Suppose $F = \mathcal{L}\{f\}$, $G = \mathcal{L}\{g\}$.

$$
\begin{aligned}
F(s)G(s) &= \left(\int_0^\infty f(t_1)\mathrm{e}^{-st_1}\,\mathrm{d}t_1\right)\left(\int_0^\infty g(t_2)\mathrm{e}^{-st_2}\,\mathrm{d}t_2\right) \\
&= \int_{t_1=0}^\infty \int_{t_2=0}^\infty f(t_1)g(t_2)\mathrm{e}^{-s(t_1+t_2)}\,\mathrm{d}t_2\,\mathrm{d}t_1 \\
&= \int_{t_1=0}^\infty \int_{\tau=t_1}^\infty f(t_1)g(\tau - t_1)\mathrm{e}^{-s\tau}\,\mathrm{d}\tau\,\mathrm{d}t_1 \\
&\qquad\qquad \text{(Subst. } \tau = t_1 + t_2,\ \mathrm{d}\tau = \mathrm{d}t_2\text{)} \\
&= \int_{\tau=0}^\infty \int_{t_1=0}^\tau f(t_1)g(\tau - t_1)\mathrm{e}^{-s\tau}\,\mathrm{d}t_1\,\mathrm{d}\tau \\
&\qquad\qquad \text{(Fubini's Theorem)} \\
&= \mathcal{L}\{h(\tau)\}
\end{aligned}
$$

with $h\colon [0,\infty) \to \mathbb{C}$, $\tau \mapsto \int_0^\tau f(t_1)g(\tau - t_1)\,\mathrm{d}t_1$.

Definition (convolution on $\mathbb{C}^{[0,\infty)}$)

Suppose $f, g\colon [0,\infty) \to \mathbb{C}$ are piecewise continuous. The *convolution (product)* of f and g is the function $f * g\colon [0,\infty) \to \mathbb{C}$ defined by

$$(f * g)(t) = \int_0^t f(\tau)g(t-\tau)\mathrm{d}\tau.$$

Remark

In Real Analysis there are several different types of convolutions in use. The present definition is tailored to the Laplace transform.

Clearly the convolution product is bilinear (i.e., linear in each argument).

Exercise

Show that the convolution product is commutative and associative, i.e. $f * g = g * f$ and $(f * g) * h = f * (g * h)$ hold for all piecewise continuous functions f, g, h on $[0,\infty)$.

Theorem

If $F(s) = \mathcal{L}f$ exists for $\mathrm{Re}(s) > a$ and $G(s) = \mathcal{L}g$ exists for $\mathrm{Re}(s) > b$ then $H(s) = \mathcal{L}(f * g)$ exists for $\mathrm{Re}(s) > \max\{a, b\}$ and satisfies

$$H(s) = F(s)G(s) \quad \textit{for } \mathrm{Re}(s) > \max\{a, b\}.$$

Proof.

The identity $H(s) = F(s)G(s)$, $\mathrm{Re}(s) > \max\{a, b\}$, is true by definition of H, provided we can show the existence of $\mathcal{L}\{f * g\}$ for $\mathrm{Re}(s) > \max\{a, b\}$ and justify the use of Fubini's Theorem.

Clearly $f * g$ is piece-wise continuous as well (even continuous).

Since piecewise continuous functions are bounded on every finite interval $[0, M]$, there exist constants K, L such that $|f(t)| \le K\,\mathrm{e}^{at}$ and $|g(t)| \le L\,\mathrm{e}^{bt}$ for all $t \ge 0$.

$$\begin{aligned}
\Longrightarrow\ |(f * g)(t)| &\le \int_0^t |f(\tau)|\,|g(t-\tau)|\,\mathrm{d}\tau \le \int_0^t KL\,\mathrm{e}^{a\tau}\mathrm{e}^{b(t-\tau)}\mathrm{d}\tau \\
&= KL\,\mathrm{e}^{bt}\int_0^t \mathrm{e}^{(a-b)\tau}\mathrm{d}\tau \\
&= \begin{cases} KL\,t\,\mathrm{e}^{bt} & \text{if } a = b, \\ KL\,\mathrm{e}^{bt}\,\frac{\mathrm{e}^{(a-b)t}-1}{a-b} = KL\,\frac{\mathrm{e}^{at}-\mathrm{e}^{bt}}{a-b} & \text{if } a \ne b. \end{cases}
\end{aligned}$$

From this it is clear that $f * g$ has exponential order at most $\max\{a, b\}$, and hence $H(s) = \mathcal{L}\{f * g\}$ exists for $\mathrm{Re}(s) > \max\{a, b\}$.

Proof cont'd.

Regarding Fubini's Theorem, it suffices to show that the 2-dimensional Lebesgue integral

$$\int_{\mathbb{R}^2} f(t_1)g(t_2)\mathrm{e}^{-s(t_1+t_2)}\mathrm{d}^2(t_1,t_2)$$

exists. (The integral to which we have applied Fubini's Theorem differs only by a change-of-variables from this.) Since f and g are piecewise continuous, the corresponding finite integrals over $[0,R]^2$ exist for every $R>0$, and as shown in Calclus III it then suffices to find a universal bound for

$$\int_{[0,R]^2} |f(t_1)g(t_2)|\,\mathrm{e}^{-s(t_1+t_2)}\mathrm{d}^2(t_1,t_2) = \left(\int_0^R |f(t_1)|\,\mathrm{e}^{-st_1}\,\mathrm{d}t_1\right)\left(\int_0^R |g(t_2)|\,\mathrm{e}^{-st_2}\,\mathrm{d}t_2\right)$$

("integration by exhaustion"). Since the Laplace integrals of f and g converge absolutely for $\mathrm{Re}(s)>\max\{a,b\}$, this is trivial: Just take the product of the corresponding limits for $R\to\infty$. Using $|f(t_1)|\le K\,\mathrm{e}^{at_1}$, $|g(t_2)|\le L\,\mathrm{e}^{bt_2}$ we can also derive the explicit bound $\frac{KL}{(s-a)(s-b)}$.

□

Example

The convolution of exponentials is given by

$$e^{at} * e^{bt} = \begin{cases} t\,e^{at} & \text{if } a = b, \\ \frac{e^{at} - e^{bt}}{a-b} & \text{if } a \neq b. \end{cases}$$

This follows from the preceding computation.

Example

Find the inverse Laplace transform of $F(s) = \dfrac{s}{(s^2+1)^2}$.

Solution: One way to solve this problem is to use the convolution theorem and the known Laplace transforms of sin, cos:

$$\mathcal{L}^{-1}\left\{\frac{s}{(s^2+1)^2}\right\} = \mathcal{L}^{-1}\left\{\frac{1}{s^2+1}\,\frac{s}{s^2+1}\right\} = \sin t * \cos t$$

$$= \int_0^t \sin\tau\cos(t-\tau)\mathrm{d}\tau = \cos t\int_0^t \sin\tau\cos\tau\mathrm{d}\tau + \sin t\int_0^t \sin^2\tau\mathrm{d}\tau$$

$$= \cos t\left[-\frac{1}{4}\cos(2\tau)\right]_0^t + \sin t\left[\frac{\tau}{2} - \frac{1}{4}\sin(2\tau)\right]_0^t = \frac{t\sin t}{2}.$$

Inversion of the Laplace Transform

Changing a function $f\colon [0,\infty) \to \mathbb{C}$ on a discrete subset of $[0,\infty)$, which must be countable, doesn't affect $\mathcal{L}\{f(t)\} = \int_0^\infty f(t)\mathrm{e}^{-st}\,\mathrm{d}t$.
$\Longrightarrow$ The Laplace transform cannot be one-to-one.

However, we have the following

Theorem

Suppose $f_1, f_2\colon [0,\infty) \to \mathbb{C}$ are piecewise continuous and $F_i(s) = \mathcal{L}\{f_i(t)\}$ is defined for $\mathrm{Re}(s) > a_i$ $(i = 1, 2)$. If there exists $s \in \mathbb{C}$ with $\mathrm{Re}(s) > \max\{a_1, a_2\}$ and $x > 0$ such that $F_1(s + kx) = F_2(s + kx)$ for all $k \in \mathbb{N}$, then $f_1(t-) = f_2(t-)$ and $f_1(t+) = f_2(t+)$ for all $t \geq 0$, and hence f_2 arises from f_1 by changing the values on some discrete subset of $[0,\infty)$.

Notes

- The conlusion of the theorem implies $a_1 = a_2$ and $F_1 = F_2$.
- The assumptions of the theorem are satisfied in particular if F_1 and F_2 coincide on their common domain $\mathrm{Re}(s) > \max\{a_1, a_2\}$.
- If f_1, f_2 satisfy the assumptions of the theorem and are continuous, we must have $f_1 = f_2$.

The proof uses the following lemma, whose proof is a bit technical and omitted.

Lemma

If f_1, f_2 satisfy the assumptions of the theorem then

$$\int_0^t f_1(\tau)\mathrm{d}\tau = \int_0^t f_2(\tau)\mathrm{d}\tau \quad \textit{for all } t \in [0,\infty).$$

Proof of the theorem.

The lemma implies that $g(t) := \int_0^t f_1(\tau)\mathrm{d}\tau = \int_0^t f_2(\tau)\mathrm{d}\tau$ for $t \in [0,\infty)$. The function g is continuous, and a straightforward generalization of the Fundamental Theorem of Calculus implies

$$f_1(t+) = \lim_{h\downarrow 0} \frac{g(t+h)-g(t)}{h} = f_2(t+) \quad \text{for } t \geq 0,$$

$$f_1(t-) = \lim_{h\uparrow 0} \frac{g(t+h)-g(t)}{h} = f_2(t-) \quad \text{for } t > 0,$$

completing the proof of the theorem. □

Remark

There is also an explicit inversion formula for the Laplace transform known, but this formula uses complex line integrals and is of practical use only when combined with the residue theorem of Complex Analysis. For now we omit it.

The Basic Idea

The Laplace transform is particularly helpful for solving IVP's corresponding to linear ODE's with constant coefficients and a right-hand side ("forcing function") $f(t)$, whose Laplace transform exists (i.e., f need not be continuous, let alone be an exponential polynomial). If the ODE has order 2 (the most important case for applications in physics/electrotechnics), the IVP looks like

$$y'' + b\,y' + c\,y = f(t), \quad y(0) = y_0,\ y'(0) = y_1,$$

where $b, c, y_0, y_1 \in \mathbb{R}$ are given constants.

The solution method consists of 3 steps:

1. Using differentiation in the domain, translate the IVP for $y(t)$ into an algebraic equation for the Laplace transform $Y(s) = \mathcal{L}\{y(t)\}$.
2. Determine $Y(s)$ by solving this algebraic equation.
3. Use Laplace Transform inversion to find $y(t) = \mathcal{L}^{-1}\{Y(s)\}$.

The solution $y(t)$ is the unique continuous Laplace-inverse of $Y(s)$ and hence well-determined in Step 3.

Example

Solve the IVP $y' + 2y = -1$, $y(0) = 1$ with the Laplace Transform.

Solution: Setting $Y(s) = \mathcal{L}\{y(t)\}$, we have

$$\begin{aligned} \mathcal{L}\{y'(t)\} + 2\,\mathcal{L}\{y(t)\} &= -\mathcal{L}\{1\}, \\ s\,Y(s) - y(0) + 2\,Y(s) &= -1/s, \\ s\,Y(s) - 1 + 2\,Y(s) &= -1/s. \end{aligned}$$

$$\Longrightarrow\ Y(s) = \frac{1 - 1/s}{s+2} = \frac{s-1}{s(s+2)} = \frac{A}{s} + \frac{B}{s+2}$$

with $A = \left.\frac{s-1}{s+2}\right|_{s=0} = -1/2$, $B = \left.\frac{s-1}{s}\right|_{s=-2} = 3/2$.

$$\Longrightarrow\ Y(s) = -\frac{1}{2}\frac{1}{s} + \frac{3}{2}\frac{1}{s+2}$$

$$\Longrightarrow y(t) = -\frac{1}{2} + \frac{3}{2}\,\mathrm{e}^{-2t}.$$

Example

Solve the IVP $y'' + y = \sin(\omega t)$, $y(0) = y'(0) = 1$ with the Laplace Transform.

Solution: Setting $Y(s) = \mathcal{L}\{y(t)\}$, we have

$$s^2\,Y(s) - s\,y(0) - y'(0) + Y(s) = \mathcal{L}\{\sin(\omega t)\} = \frac{\omega}{s^2+\omega^2},$$

$$s^2\,Y(s) - s - 1 + Y(s) = \frac{\omega}{s^2+\omega^2}.$$

$$\Longrightarrow\quad Y(s) = \frac{s+1}{s^2+1} + \frac{\omega}{(s^2+1)(s^2+\omega^2)}$$

Now there are two cases to consider:

$\underline{\omega \neq \pm 1}$:

$$\Longrightarrow\quad Y(s) = \frac{s+1}{s^2+1} + \frac{\omega}{\omega^2-1}\,\frac{1}{s^2+1} - \frac{\omega}{\omega^2-1}\,\frac{1}{s^2+\omega^2}$$

$$\Longrightarrow\quad y(t) = \cos t + \sin t + \frac{\omega}{\omega^2-1}\sin t - \frac{1}{\omega^2-1}\sin(\omega t)$$

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

Example (cont'd)

$\omega = \pm 1$: Here we have

$$Y(s) = \frac{s+1}{s^2+1} \pm \frac{1}{(s^2+1)^2} = \frac{s+1}{s^2+1} \pm \frac{1}{2}\,\frac{1}{s^2+1} \mp \frac{1}{2}\,\frac{s^2-1}{(s^2+1)^2}$$

$$\Longrightarrow\ y(t) = \cos t + \sin t \pm \frac{1}{2}\sin t \mp \frac{1}{2}\,t\cos t$$

$$= \begin{cases} \cos t + \frac{3}{2}\sin t - \frac{1}{2}\,t\cos t & \text{for } \omega = 1, \\ \cos t + \frac{1}{2}\sin t + \frac{1}{2}\,t\cos t & \text{for } \omega = -1. \end{cases}$$

Explanation: The above decomposition of $1/(s^2+1)^2$ and its inverse Laplace transform were found by playing around with the known Laplace transforms $\mathcal{L}\{\cos t\} = s/(s^2+1)$, $\mathcal{L}\{\sin t\} = 1/(s^2+1)$. Use $\mathcal{L}\{t\cos t\} = -\frac{\mathrm{d}}{\mathrm{d}s}\,\frac{s}{s^2+1} = \frac{s^2-1}{(s^2+1)^2} = \frac{1}{s^2+1} - \frac{2}{(s^2+1)^2}$, from which it is obvious.

The standard way to compute $\mathcal{L}\left\{\frac{1}{(s^2+1)^2}\right\}$ is to use complex partial fractions $\frac{1}{(s^2+1)^2} = \frac{1}{(s-\mathrm{i})^2(s+\mathrm{i})^2} = \frac{A}{s-\mathrm{i}} + \frac{B}{(s-\mathrm{i})^2} + \frac{C}{s+\mathrm{i}} + \frac{D}{(s+\mathrm{i})^2}$ together with $\frac{1}{s\mp\mathrm{i}} = \mathcal{L}\left\{\mathrm{e}^{\pm \mathrm{i}t}\right\}$, $\frac{1}{(s\mp\mathrm{i})^2} = \mathcal{L}\left\{t\,\mathrm{e}^{\pm \mathrm{i}t}\right\}$. One obtains $B = D = -1/4$, $A = -\mathrm{i}/4$, $C = +\mathrm{i}/4$, ...

Exercise (advanced)

It appears that we can obtain the solution for $\omega = \pm 1$ by viewing the solution for $\omega \neq \pm 1$ as a two-variable function $y(\omega, t)$ and computing $\lim_{\omega \to \pm 1} y(\omega, t)$ with the aid of L'Hospital's Rule. Can you prove this rigorously? (Compare also with the proof of Part 3 of our big theorem on fractional power series solutions of 2nd-order linear ODE's near regular singular points.)

Continuous Forcing

The preceding two examples had a continuous forcing function $f\colon [0,\infty) \to \mathbb{C}$ (viz. $f(t) = -1$, resp., $f(t) = \sin(\omega t)$). In such a case the sharpened version of the Existence and Uniqueness Theorem for linear ODE's applies and guarantees that there exists a unique solution $y(t)$ on $[0,\infty)$ satisfying any given initial conditions $y(0) = y_0$, $y'(0) = y_1$ (and, similarly, for initial times $t_0 > 0$). As argued before the examples, the solution method using the Laplace transform produces this solution, provided $y(t)$ has a Laplace transform $Y(s)$ and $\mathcal{L}^{-1}\{Y(s)\}$ can be found. For this it is sufficient that $f(t)$ has a Laplace transform; see the subsequent theorem.

If the forcing function f satisfies $f(0) = 0$, its trivial extension to $\mathbb{R}$ (by setting $f(t) < 0$ for $t < 0$) is continuous as well, and hence maximal solutions of corresponding initial value problems are defined on $\mathbb{R}$. Such solutions do not necessarily vanish on $(-\infty, 0)$; this is the case iff the initial values at $t_0 = 0$ are $y_0 = y_1 = 0$.

Theorem

Suppose $f : [0, \infty) \to \mathbb{C}$ *is continuous and of exponential order. Then the same is true of any solution* $y : [0, \infty) \to \mathbb{C}$ *of* $y'' + by' + cy = f(t)$*, and hence* $Y(s) = \mathcal{L}\{y(t)\}$ *is defined in some half plane* $\mathrm{Re}(s) > a$*.*

Proof.

In the homogeneous case $f(t) \equiv 0$ solutions are exponential polynomials and the assertion is obvious. In the general case it follows by inspecting the variation-of-parameters formula for the solution (cf. our earlier discussion of analytic solutions of 2nd-order inhomogeneous linear ODE's) and using that "exponential order" is inherited by products and integrals. Since the ODE's considered here have constant coefficients, the Wronskian appearing in the formula is a nonzero multiple of e^{-bt}.) □

Discontinuous Forcing

Now consider the more general IVP

$$y'' + b\,y' + c\,y = f(t), \quad y(0) = y_0,\ y'(0) = y_1 \qquad (\star)$$

with constants $b, c, y_0, y_1 \in \mathbb{R}$ and forcing function $f\colon [0,\infty) \to \mathbb{C}$, whose Laplace transform $F(s)$ exists in some half plane $\mathrm{Re}(s) > a$. For the following discussion we assume that f is piecewise continuous and of at most exponential order.

Because y'' cannot exist at discontinuities of f (as can be shown), we must adapt the definition of a solution to this more general situation.

Definition

By a *solution* of $(\star)$ we mean a C^1-function $y\colon [0,\infty) \to \mathbb{C}$ with the following properties:

1. $y''(t)$ exists at all points $t \in [0,\infty)$ where f is continuous, and $y''(t) + b\,y'(t) + c\,y(t) = f(t)$ holds for those points t;
2. $y(0) = y_0$, $y'(0) = y_1$.

Notes

- In the definition the values of f at its discontinuities do not matter, and hence need not even be defined.
- For a solution y the derivative y'' must be piecewise continuous (with the same exceptional set as f), as the representation $y''(t) = f(t) - b\,y'(t) - c\,y(t)$ shows together with the assumption that y is a C^1-function.
  $\Longrightarrow \mathcal{L}\{y'' + by' + cy\}$ can be computed using the differentiation-in-the-domain formulas.
- A priori a solution y determines solutions in the original sense only on the open intervals (t_{k-1}, t_k) between adjacent discontinuities of f (including $(0, t_1)$ and, if there is a largest discontinuity t_n, also (t_n, ∞)). However, the endpoints can be included since y'' has one-sided derivatives in the endpoints, viz. $y''_+(t_{k-1}) = \lim_{t \downarrow t_{k-1}} y''(t)$, $y''_-(t_k) = \lim_{t \uparrow t_k} y''(t)$, which satisfy the ODE as well.

Theorem

1. *If the forcing function f is piecewise continuous, the IVP $(\star)$ has a unique solution y.*
2. *If the Laplace transform of f exists then the Laplace transform method can be applied and produces the solution y.*

Proof.

(1) The Existence and Uniqueness Theorem first gives a unique solution y_1 of the IVP $(\star)$ on $[0, t_1]$, then a unique solution y_2 of the ODE on $[t_1, t_2]$ with initial values $y_2(t_1) = y_1(t_1)$, $y_2'(t_1) = y_1'(t_1)$, and so forth. Defining y as y_k on $[t_{k-1}, t_k]$ yields the desired solution of $(\star)$ on $[0, \infty)$. Conversely, the requirement that y be a C^1-function forces the initial conditions of y_k and y_{k+1} at t_k to match and hence determines y uniquely.

(2) Piecewise continuity of y'' ensures that $Y(s) = \mathcal{L}\{y(t)\}$ can be computed as usual from the given data:

$$s^2\,Y(s) - s\,y_0 - y_1 + b\big(s\,Y(s) - y_0\big) + c\,Y(s) = F(s)$$

$$\Longrightarrow\quad Y(s) = \frac{F(s) + s\,y_0 + b\,y_0 + y_1}{s^2 + b\,s + c} = G(s), \quad \text{say.}$$

$\Longrightarrow y(t) = \mathcal{L}^{-1}\{G(s)\}$ (i.e., the unique continuous preimage). $\square$

Remarks

$Y(s) = Y_p(s) + Y_h(s)$ with $Y_p(s) = \frac{F(s)}{s^2+b\,s+c}$, $Y_h(s) = \frac{s\,y_0+b\,y_0+y_1}{s^2+b\,s+c}$.

1. $Y_p(s)$ is the Laplace transform of the solution $y_p(t)$ of $(\star)$ with initial values $y_p(0) = y_p'(0) = 0$.
2. $Y_h(s)$ is the Laplace transform of the solution $y_h(t)$ of the associated homogeneous ODE with initial values y_0, y_1.
3. The denominator of $Y_p(s)$, $Y_h(s)$, viewed as a polynomial in s, is precisely the characteristic polynomial of $(\star)$.
4. $Y_h(s)$ is a rational function of s, and hence $y_h(t) = \mathcal{L}^{-1}\{Y_h(s)\}$ can be determined from the partial fraction decomposition of $Y_h(s)$. This provides an alternative method to determine the general solution in the homogeneous case.
5. If $f(t) = \sum_{i=1}^{r} c_i t^{m_i} e^{\mu_i t}$ is an exponential polynomial, $Y_p(s)$ and $Y(s)$ are rational functions of s as well, so that $y_p(t)$, $y(t)$ can be determined in the same way using partial fractions. If $f(t)$ is not an exponential polynomial, the Laplace-inverse of $Y_p(s)$ may nevertheless be known, providing a method to solve additional instances of such ODE's.

These observations generalize mutatis mutandis to higher-order ODE's.

The Laplace transform of the Heaviside function

$$\mathrm{u}(t) = \begin{cases} 1 & \text{if } t \geq 0, \\ 0 & \text{if } t < 0. \end{cases}$$

is $\mathcal{L}\{\mathrm{u}(t)\} = \mathcal{L}\{1\} = 1/s$.

Here we use the convention that the Laplace transform of a function defined on $\mathbb{R}$ (and vanishing on $(-\infty, 0)$) is that of its restriction to $[0, \infty)$. Conversely, we can view any function $f\colon [0,\infty) \to \mathbb{C}$ as a function on $\mathbb{R}$ by setting $f(t) = 0$ for $t < 0$. The extended function is piecewise continuous and of exponential order a iff the original function is.

Now consider a rectangular forcing function of unit height, i.e.,

$$\mathrm{r}_{a,b}(t) = \begin{cases} 1 & \text{if } a \leq t \leq b, \\ 0 & \text{if } t < a \text{ or } t > b, \end{cases}$$

with $a, b \in \mathbb{R}$ satisfying $0 \leq a < b$.
$\mathrm{r}_{a,b}$ can be expressed in terms of the Heaviside function as

$$\mathrm{r}_{a,b}(t) = \mathrm{u}(t-a) - \mathrm{u}(t-b) = \mathrm{u}_a(t) - \mathrm{u}_b(t)$$

(except for $t = b$, where the right-hand side is $\mathrm{u}(b-a) - \mathrm{u}(0)$ $= 1 - 1 = 0$, but this change doesn't affect the Laplace transform).

Using linearity of the Laplace transform and the translation in the argument formula, we obtain

$$\mathcal{L}\{\mathrm{r}_{a,b}\} = \mathcal{L}\{\mathrm{u}_a\} - \mathcal{L}\{\mathrm{u}_b\} = \mathrm{e}^{-as}\mathcal{L}\{\mathrm{u}\} - \mathrm{e}^{-bs}\mathcal{L}\{\mathrm{u}\} = \frac{\mathrm{e}^{-as} - \mathrm{e}^{-bs}}{s}.$$

In particular, if the upward step is at $t = 0$ ($a = 0$) then $\mathrm{r}_{a,b}(t) = \mathrm{r}_{0,b}(t) = (1 - \mathrm{e}^{-bs})/s$.

Example (discontinuous forcing)

Solve the IVP $y''(t) + y(t) = \begin{cases} 1 & \text{for } 0 < t < 1, \\ 0 & \text{for } t > 1, \end{cases}$ with initial conditions $y(0) = y'(0) = 0$ with the Laplace transform.

Solution: Since $y_0 = y_1 = 0$, the Laplace transform of the left-hand side is $(s^2 + 1)Y(s)$, and that of the right-hand side is $\mathcal{L}\{\mathrm{r}_{0,1}\} = (1 - \mathrm{e}^{-s})/s$.

$$\Longrightarrow \quad Y(s) = \frac{1 - \mathrm{e}^{-s}}{s(s^2 + 1)}.$$

Using partial fractions ($1 = 1 \cdot (s^2 + 1) - s \cdot s$), we obtain

$$Y(s) = (1 - \mathrm{e}^{-s}) \left(\frac{1}{s} - \frac{s}{s^2 + 1} \right).$$

Example (cont'd)

Since $\mathcal{L}^{-1}\left\{\frac{1}{s} - \frac{s}{s^2+1}\right\} = 1 - \cos t$, this gives

$$\begin{aligned} y(t) &= 1 - \cos t - \mathrm{u}_1(t)\,[1 - \cos(t-1)] \\ &= \begin{cases} 1 - \cos t & \text{for } 0 \le t \le 1, \\ \cos(t-1) - \cos t & \text{for } t \ge 1. \end{cases} \end{aligned}$$

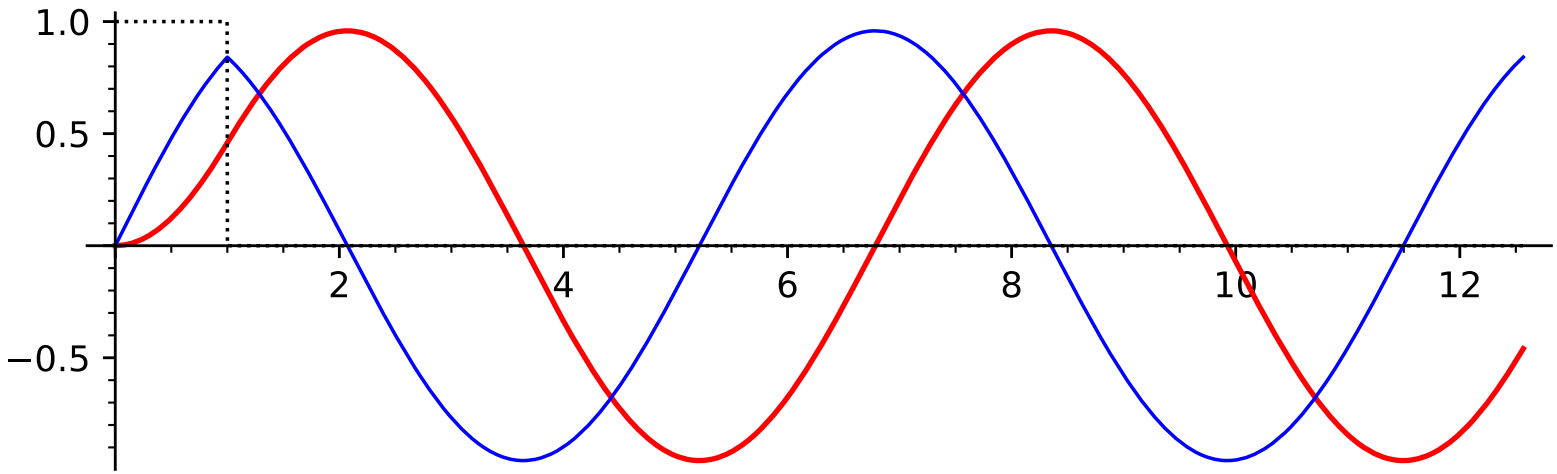


Figure: The solution $y(t)$ (in red), its derivative $y'(t)$ (in blue), and the forcing function $f(t)$ (dotted)

Note that both sections of $y(t)$ are periodic oszillations. For the section on $[1, \infty)$, we alternatively have $\cos(t-1) - \cos t = 2\sin\frac{t+t-1}{2}\sin\frac{t-(t-1)}{2} = 2\sin\left(\frac{1}{2}\right)\sin\left(t - \frac{1}{2}\right) \approx 0.96\sin\left(t - \frac{1}{2}\right)$.

Example (discontinuous forcing)

Solve the IVP $y'' + 3y' + 2y = \begin{cases} 1 & \text{for } t \in [0,1] \cup [2,3] \cup [4,5], \\ 0 & \text{otherwise,} \end{cases}$

$y(0) = y'(0) = 0$ with the Laplace transform.

Solution: Since $y_0 = y_1 = 0$, again $Y(s)$ has the simple form

$$Y(s) = \frac{F(s)}{s^2 + 3s + 2} = \frac{F(s)}{(s+1)(s+2)}$$

with $F(s) = \mathcal{L}\{f(t)\}$, where

$$f(t) = [\mathrm{u}_0(t) - \mathrm{u}_1(t)] + [\mathrm{u}_2(t) - \mathrm{u}_3(t)] + [\mathrm{u}_4(t) - \mathrm{u}_5(t)]\,.$$

$$\Longrightarrow\ F(s) = \frac{1}{s} - \frac{\mathrm{e}^{-s}}{s} + \frac{\mathrm{e}^{-2s}}{s} - \frac{\mathrm{e}^{-3s}}{s} + \frac{\mathrm{e}^{-4s}}{s} - \frac{\mathrm{e}^{-5s}}{s},$$

$$\Longrightarrow\ Y(s) = \frac{1 - \mathrm{e}^{-s} + \mathrm{e}^{-2s} - \mathrm{e}^{-3s} + \mathrm{e}^{-4s} - \mathrm{e}^{-5s}}{s(s+1)(s+2)}.$$

Example (cont')

The partial fractions decomposition of the denominator is

$$\frac{1}{s(s+1)(s+2)} = \frac{1}{2}\frac{1}{s} - \frac{1}{s+1} + \frac{1}{2}\frac{1}{s+2} = \mathcal{L}\left\{\frac{1}{2} - \mathrm{e}^{-t} + \frac{1}{2}\mathrm{e}^{-2t}\right\}.$$

Writing $g(t) = \frac{1}{2} - \mathrm{e}^{-t} + \frac{1}{2}\mathrm{e}^{-2t}$, the solution is

$$y(t) = g(t) - \mathrm{u}_1(t)g(t-1) + \mathrm{u}_2(t)g(t-2) - \mathrm{u}_3(t)g(t-3) + \\ + \mathrm{u}_4(t)g(t-4) - \mathrm{u}_5(t)g(t-5) = \cdots$$

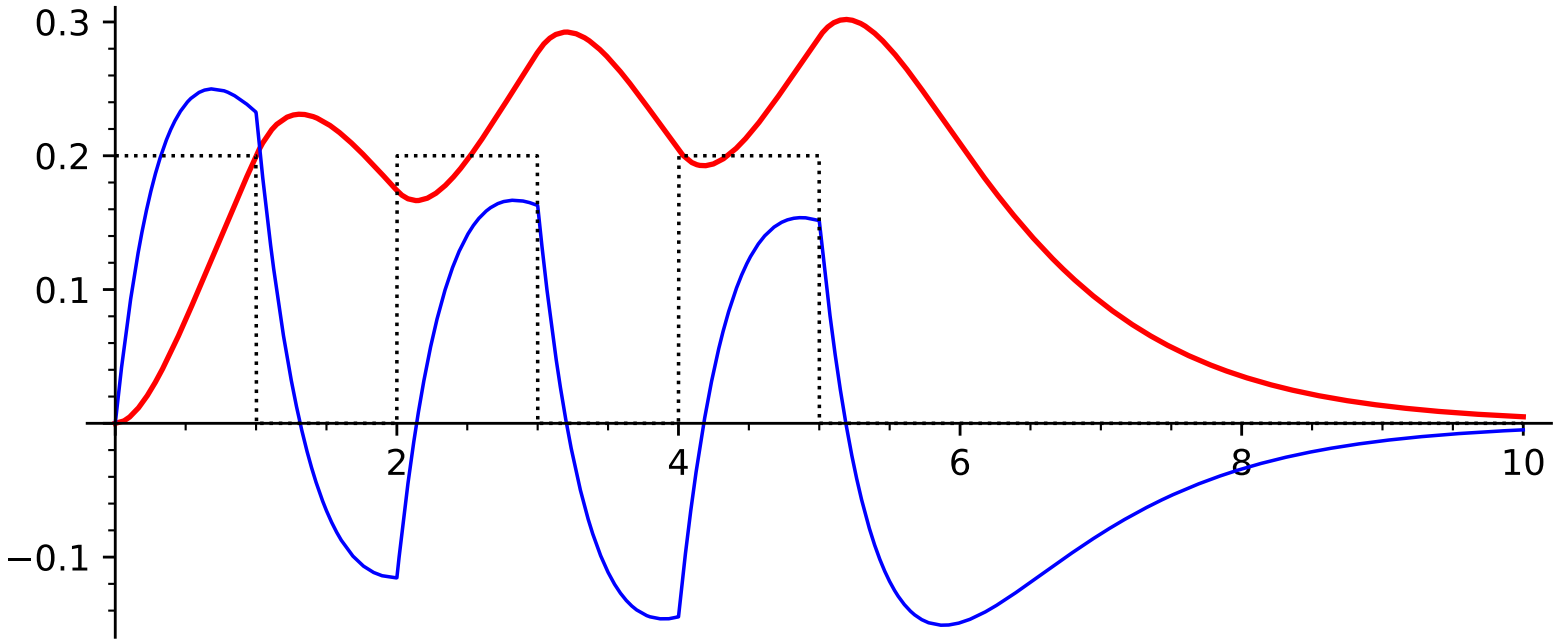


Figure: $y(t)$ (in red), $y'(t)$ (in blue), and $0.2\,f(t)$ (dotted)

Example (continuous forcing)

Solve the IVP $y'' + y = \begin{cases} t & \text{for } t \in [0,1], \\ 2-t & \text{for } t \in [1,2], \\ 0 & \text{otherwise} \end{cases}$

for general initial values $y(0) = y_0$, $y'(0) = y_1$ with the Laplace transform.

Solution: The solution is $y(t) = y_p(t) + y_0 \cos t + y_1 \sin t$, where $y_p(t)$ is the particular solution with $y_p(0) = y_p'(0) = 0$.

As before, $Y(s) = \mathcal{L}\{y_p(t)\}$ has the form $Y(s) = \frac{F(s)}{s^2+1}$ with

$$\begin{aligned} F(s) &= \mathcal{L}\{f(t)\} \\ &= \mathcal{L}\{t(\mathrm{u}(t) - \mathrm{u}(t-1)) + (2-t)(\mathrm{u}(t-1) - \mathrm{u}(t-2))\} \\ &= \mathcal{L}\{t\,\mathrm{u}(t) - 2(t-1)\mathrm{u}(t-1) + (t-2)\mathrm{u}(t-2)\} \\ &= \frac{1}{s^2} - \frac{2\,\mathrm{e}^{-s}}{s^2} + \frac{e^{-2s}}{s^2}. \end{aligned}$$

$$\Longrightarrow \quad Y(s) = \frac{1 - 2\,\mathrm{e}^{-s} + \mathrm{e}^{-2s}}{s^2(s^2+1)} = (1 - 2\,\mathrm{e}^{-s} + \mathrm{e}^{-2s})\left(\frac{1}{s^2} - \frac{1}{s^2+1}\right).$$

Example (cont'd)

Since $\frac{1}{s^2} - \frac{1}{s^2+1} = \mathcal{L}\{t - \sin t\}$, this gives

$$y_p(t) = t - \sin t - 2\,u_1(t)\,[t - 1 - \sin(t-1)] + u_2(t)\,[t - 2 - \sin(t-2)]$$

$$= \begin{cases} t - \sin t & \text{if } 0 \le t \le 1, \\ 2 - t + 2\sin(t-1) - \sin t & \text{if } 1 \le t \le 2, \\ 2\sin(t-1) - \sin t - \sin(t-2) & \text{if } t \ge 2. \end{cases}$$

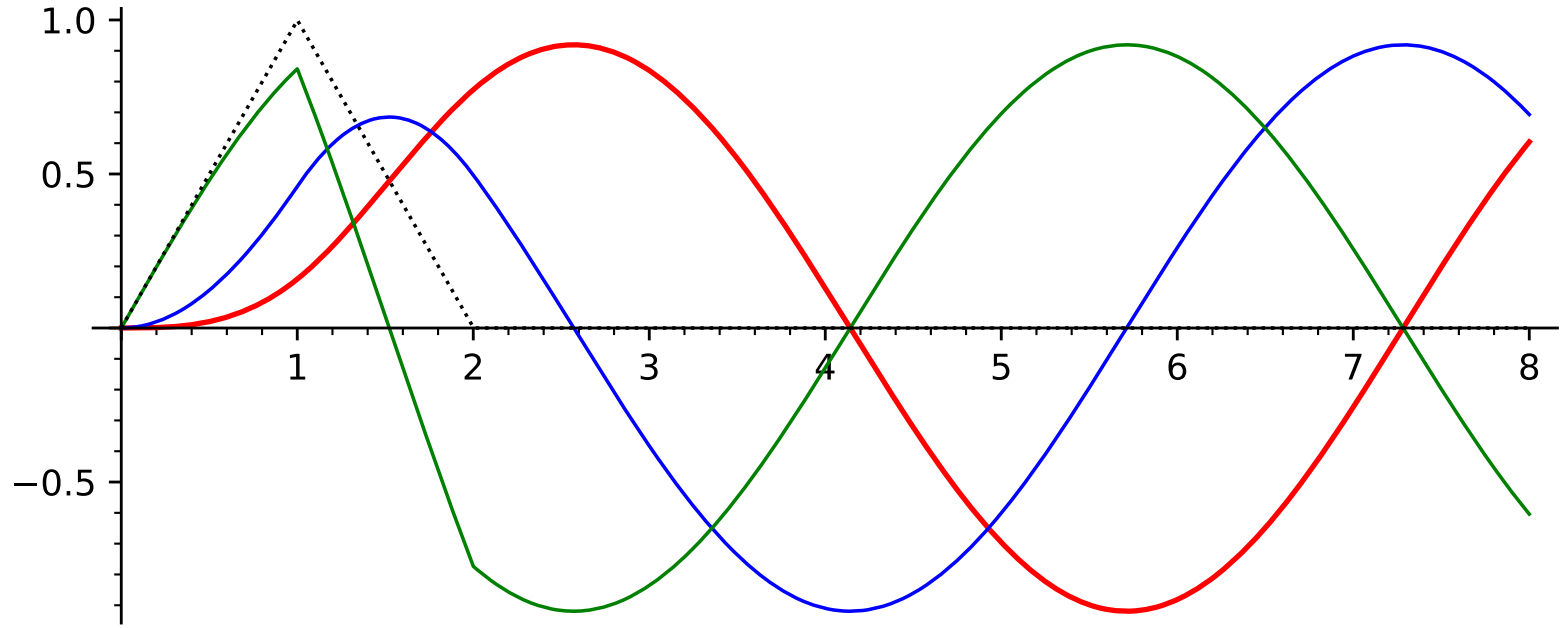


Figure: $y_p(t)$ (in red), $y_p'(t)$ (in blue), $y_p''(t)$ (in green), and $f(t)$ (dotted)

Impulsive forcing

cf. [BDM17], Ch. 6.5

Imagine that in our first example we replace the forcing function $f(t) = \mathrm{r}_{0,1}(t)$ by $f_\epsilon(t) = (1/\epsilon)\mathrm{r}_{0,\epsilon}(t)$, $\epsilon > 0$ (a rectangle with basis ϵ and height $1/\epsilon$, hence still of area 1), and let $\epsilon \downarrow 0$ (or at least consider very small ϵ).

Such forcing functions are important for applications, where they describe time-dependent forces acting over a short period of time and such that the total impulse of the force is constant (for mechanical systems), or electric impulses of high intensity over a short period such that the total voltage of the impulse is constant (for electric circuits).

The solution of the IVP $y'' + y = (1/\epsilon)\mathrm{r}_{0,\epsilon}$, $y(0) = y'(0) = 0$ is

$$y_\epsilon(t) = \begin{cases} \frac{1}{\epsilon}(1 - \cos t) & \text{if } 0 \le t \le \epsilon, \\ \frac{1}{\epsilon}\left[\cos(t-\epsilon) - \cos t\right] & \text{if } t \ge \epsilon. \end{cases}$$

Since $\cos(t-\epsilon) - \cos t = 2\sin\left(t - \frac{\epsilon}{2}\right)\sin\left(\frac{\epsilon}{2}\right)$, the "limiting solution" is

$$y(t) = \lim_{\epsilon\downarrow 0} y_\epsilon(t) = \lim_{\epsilon\downarrow 0} \frac{\sin(t-\epsilon/2)\sin(\epsilon/2)}{\epsilon/2} = \sin t \quad \text{for } t > 0.$$

Since $y_\epsilon(0) = 0$, this also holds at $t = 0$.

Observation

If we assign to the "limit function"

$$\delta(t) = \lim_{\epsilon \downarrow 0} f_\epsilon(t) = \lim_{\epsilon \downarrow 0} (1/\epsilon) \mathrm{r}_{0,\epsilon}(t) = \begin{cases} +\infty & \text{if } t = 0, \\ 0 & \text{if } t \neq 0, \end{cases}$$

the Laplace transform $\mathcal{L}\{\delta(t)\} = 1$, then $y(t) = \sin t = \mathcal{L}^{-1}\left\{\frac{1}{s^2+1}\right\}$ can be obtained directly with the usual solution method.

Of course we know that there is no ordinary function on $[0, \infty)$ with Laplace transform 1. In fact the Laplace integral of $\delta(t)$ is zero for all s, because the single value $\delta(0) = +\infty$ doesn't matter for integration.

But it turns out that we can work with $\delta(t)$ in a meaningful way, provided we leave the definition of $\delta(t)$ as an ordinary function aside, use $f_\epsilon(t)$ in place of $\delta(t)$ in all computations, and obtain the value corresponding to $\delta(t)$ by letting $\epsilon \downarrow 0$. The precise mathematical term for such "generalized functions" is "*distribution*", and the present discussion should be viewed as a simplified (and sometimes non-rigorous) account of Dirac's δ-distribution.

It is custom to use rectangular functions that are symmetric about the origin in the final definition of $\delta(t)$, because then the resulting distribution reflects local properties at zero (on both sides) of the functions it is applied to.

Definition

Dirac's Delta function is the distribution ("generalized function") defined on $\mathbb{R}$ by $\delta(t) = \lim_{\epsilon \downarrow 0} f_\epsilon(t)$ with

$$f_\epsilon(t) = \frac{1}{2\epsilon}\, \mathrm{r}_{-\epsilon,\epsilon}(t) = \frac{\mathrm{u}(t+\epsilon) - \mathrm{u}(t-\epsilon)}{2\epsilon}.$$

As a simple example for the ideas involved in the definition of Dirac's Delta function we prove the following two properties:

1. $\displaystyle\int_{-\infty}^{\infty} \delta(t)\,\mathrm{d}t = 1.$
2. If $f\colon \mathbb{R} \to \mathbb{C}$ is piecewise continuous then

$$\int_{-\infty}^{\infty} f(t)\delta(t-t_0)\,\mathrm{d}t = \frac{f(t_0-) + f(t_0+)}{2};$$

in particular, if f is continuous in t_0 then $\int_{-\infty}^{\infty} f(t)\delta(t-t_0)\,\mathrm{d}t = f(t_0)$.

Proof.

(1) Since $\int_{-\infty}^{\infty} f_\epsilon(t)\,\mathrm{d}t = 1$ for every $\epsilon > 0$, we also have

$$\int_{-\infty}^{\infty} \delta(t)\,\mathrm{d}t = \lim_{\epsilon \downarrow 0} \int_{-\infty}^{\infty} f_\epsilon(t)\,\mathrm{d}t = 1.$$

(2) We have

$$\begin{aligned}\int_{-\infty}^{\infty} f(t) f_\epsilon(t-t_0)\,\mathrm{d}t &= \frac{1}{2\epsilon}\int_{t_0-\epsilon}^{t_0+\epsilon} f(t)\,\mathrm{d}t \\ &= \frac{1}{2\epsilon}\int_{t_0-\epsilon}^{t_0} f(t) - f(t_0-)\,\mathrm{d}t + \frac{1}{2\epsilon}\int_{t_0}^{t_0+\epsilon} f(t) - f(t_0+)\,\mathrm{d}t + \frac{f(t_0-)+f(t_0+)}{2}.\end{aligned}$$

Since $\left|\int_{t_0-\epsilon}^{t_0} f(t) - f(t_0-)\,\mathrm{d}t\right| \le \epsilon \max\{|f(t) - f(t_0-)|\,;\, t_0-\epsilon \le t \le t_0\}$, the first summand tends to zero for $\epsilon \downarrow 0$, and similarly for the 2nd summand.

$\Longrightarrow \int_{-\infty}^{\infty} f(t)\delta(t-t_0)\,\mathrm{d}t = \lim_{\epsilon\downarrow 0}\int_{-\infty}^{\infty} f(t) f_\epsilon(t-t_0)\,\mathrm{d}t = \frac{f(t_0-)+f(t_0+)}{2}$. □

In order to work with $\delta(t)$ symbolically in the context of the Laplace transform, we need further properties:

3. $\mathcal{L}\{\delta(t-t_0)\} = \mathrm{e}^{-st_0}$ for $t_0 > 0$;
4. $\mathcal{L}\{\delta(t)\} = 1$ (the constant function $s \mapsto 1$);
5. $\mathrm{u}'(t) = \delta(t)$.

Proof.

(3) For $\epsilon \le t_0$ the function $t \mapsto f_\epsilon(t-t_0)$ vanishes on $(-\infty, 0)$.

$$\Longrightarrow \int_0^\infty f_\epsilon(t-t_0)\mathrm{e}^{-st}\,\mathrm{d}t = \int_{-\infty}^\infty f_\epsilon(t-t_0)\mathrm{e}^{-st}\,\mathrm{d}t,$$

which for $\epsilon \downarrow 0$ converges to e^{-st_0}, since $t \mapsto \mathrm{e}^{-st}$ is continuous.; cf. Property 2 and its proof.

(4) This follows by letting $t_0 \downarrow 0$ in (3).

(5) We have

$$\int_{-\infty}^t f_\epsilon(\tau)\mathrm{d}\tau = \begin{cases} 0 & \text{if } t \le -\epsilon, \\ \frac{t+\epsilon}{2\epsilon} & \text{if } t \in [-\epsilon, \epsilon], \\ 1 & \text{if } t \ge \epsilon. \end{cases}$$

$\Longrightarrow \int_{-\infty}^t \delta(\tau)\mathrm{d}\tau = \lim_{\epsilon\downarrow 0} \int_{-\infty}^t f_\epsilon(\tau)\mathrm{d}\tau = \mathrm{u}(t)$ (except for $t = 0$). □

Notes

- Since $\lim_{\epsilon \downarrow 0} \mathcal{L}\{f_\epsilon(t)\} = 1/2$, Property 4 cannot be concluded in the usual way. (For this one needs to use the one-sided analog of $f_\epsilon(t)$ as in the example.) If we want the translation-in-the-domain formula also hold for $\delta(t)$, we must define $\mathcal{L}\{\delta(t)\} = 1$.
- Some people define the value of the Heaviside function at $t = 0$ as $\mathrm{u}(0) = 1/2$. With this definition, Property 5 holds also at $t = 0$.

Example

Find the solution of the initial value problem

$$y'' - 4y' + 4y = 3\,\delta(t-1) + \delta(t-2); \quad y(0) = y'(0) = 1.$$

Solution: Applying $\mathcal{L}$ to both sides of the ODE and using Property (3) gives

$$s^2\,Y(s) - s - 1 - 4\big(s\,Y(s) - 1\big) + 4\,Y(s) = 3\,\mathrm{e}^{-s} + \mathrm{e}^{-2s}$$

$$(s^2 - 4s + 4)\,Y(s) = s - 3 + 3\,\mathrm{e}^{-s} + \mathrm{e}^{-2s}$$

Example (cont'd)

$$\begin{aligned}\Longrightarrow Y(s) &= \frac{s-3}{(s-2)^2} + \frac{3\,e^{-s}}{(s-2)^2} + \frac{\mathrm{e}^{-2s}}{(s-2)^2} \\ &= \frac{1}{s-2} - \frac{1}{(s-2)^2} + \frac{3\,e^{-s}}{(s-2)^2} + \frac{\mathrm{e}^{-2s}}{(s-2)^2}\end{aligned}$$

$$y(t) = \mathrm{e}^{2t} - t\,\mathrm{e}^{2t} + 3\,\mathrm{u}_1(t)(t-1)\mathrm{e}^{2(t-1)} + \mathrm{u}_2(t)(t-2)\mathrm{e}^{2(t-2)}.$$

The meaning of this solution is the following: If $y_\epsilon(t)$ denotes the solution of the IVP

$$y'' - 4y' + 4y = 3\,f_\epsilon(t-1) + f_\epsilon(t-2); \quad y(0) = y'(0) = 1, \ (\mathrm{IVP}_\epsilon)$$

we have $\lim_{\epsilon \downarrow 0} y_\epsilon(t) = y(t)$. Hence for small ϵ the solution of (IVP_ϵ) is well approximated by $y(t)$.

The use of the convolution

We may view $Y_p(s) = \frac{F(s)}{s^2+bs+c}$ as a function of $F(s)$, (and hence of the forcing function $f(t)$. This functional relation can be written as

$$Y_p(s) = \mathrm{H}(s)F(s) \quad \text{with} \quad \mathrm{H}(s) = (s^2 + bs + c)^{-1}.$$

The function $\mathrm{H}(s)$ is called *transfer function* of the ODE (or the physical system described by the ODE). The name comes from the fact that we can consider the solution $y(t)$ as "output" of the system when the forcing function $f(t)$ (e.g., a mechanical force/electric impulse) is applied as "input".

The convolution theorem gives

$$y_p(t) = \int_0^t \mathrm{h}(t-\tau)f(\tau)\mathrm{d}\tau$$

with $\mathrm{h}(t) = \mathcal{L}^{-1}\{\mathrm{H}(s)\} = \mathcal{L}^{-1}\left\{\frac{1}{s^2+bs+c}\right\}$. Thus $\mathrm{h}(t)$ (the so-called "impulse response") is the solution for $f(t) = \delta(t)$ ("unit impulse at time $t = 0$"), and the solution $y_p(t)$ in the general case is the convolution of the impulse response and the forcing function.

Math 285
Introduction to Differential Equations

Thomas Honold

ZJU-UIUC Institute

Spring Semester 2024

Outline

Today's Lecture:

Introduction

A first-order autonomous (time-independent) linear ODE system has the form

$$\mathbf{y}' = \mathbf{A}\mathbf{y} + \mathbf{b}, \quad \text{with } \mathbf{A} \in \mathbb{C}^{n\times n},\ \mathbf{b} \in \mathbb{C}^n.$$

As in the case of higher-order scalar ODE's, we will include in the discussion the case of a time-dependent continuous "source" $\mathbf{b}(t)$, i.e., consider more generally $\mathbf{y}' = \mathbf{A}\mathbf{y} + \mathbf{b}(t)$ or, written out in full,

$$\begin{pmatrix} y_1' \\ \vdots \\ y_n' \end{pmatrix} = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nn} \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} + \begin{pmatrix} b_1(t) \\ \vdots \\ b_n(t) \end{pmatrix}.$$

Motivation

ODE systems of the form just described often occur when modeling physical systems with a number of separate but interconnected ("coupled") components. Examples are provided by spring-mass systems and LRC electric circuits. We just reproduce the two introductory examples from [BDM17], Ch. 7.1.

Example ([BDM17], p. 279)

A 1-dimensional *two-mass, three-spring system* under the influence of external forces is described by the 2nd-order ODE system

$$\begin{aligned} m\,x_1''(t) &= -(k_1 + k_2)x_1 + k_2 x_2 + F_1(t), \\ m\,x_2''(t) &= k_2 x_1 - (k_2 + k_3)x_2 + F_2(t), \end{aligned}$$

where x_1, x_2 denote the coordinates of the masses, k_1, k_2, k_3 the spring constants, and $F_1(t)$, $F_2(t)$ the (time-dependent) external forces.

This 2×2 linear system can by reduced to a 4×4 first-order linear system by the the usual method of order reduction, i.e., we introduce two further variables $x_3 = x_1'$, $x_4 = x_2'$.

Example (cont'd)

This gives the four equations

$$\begin{aligned}
x_1' &= x_3, \\
x_2' &= x_4, \\
x_3' &= x_1'' = -(k_1 + k_2)/m \cdot x_1 + (k_2/m)x_2 + F_1(t)/m, \\
x_4' &= x_2'' = (k_2/m)x_1 - (k_2 + k_3)/m \cdot x_2 + F_2(t)/m,
\end{aligned}$$

or, in matrix form,

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}' = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -\frac{k_1+k_2}{m} & \frac{k_2}{m} & 0 & 0 \\ \frac{k_2}{m} & -\frac{k_2+k_3}{m} & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \frac{F_1(t)}{m} \\ \frac{F_2(t)}{m} \end{pmatrix}.$$

Example ([BDM17], p. 280)

The current $I(t)$ and voltage $V(t)$ in a *parallel LRC circuit* satisfy the 2×2 first-order homogeneous linear system

$$I'(t) = \frac{V}{L},$$
$$V'(t) = -\frac{I}{C} - \frac{V}{RC},$$

where L, R, C denote the inductance/resistance/capacitance of the inductor/resistor/capacitor.
In matrix form this system is

$$\begin{pmatrix} I \\ V \end{pmatrix}' = \begin{pmatrix} 0 & \frac{1}{L} \\ -\frac{1}{C} & -\frac{1}{RC} \end{pmatrix} \begin{pmatrix} I \\ V \end{pmatrix}.$$

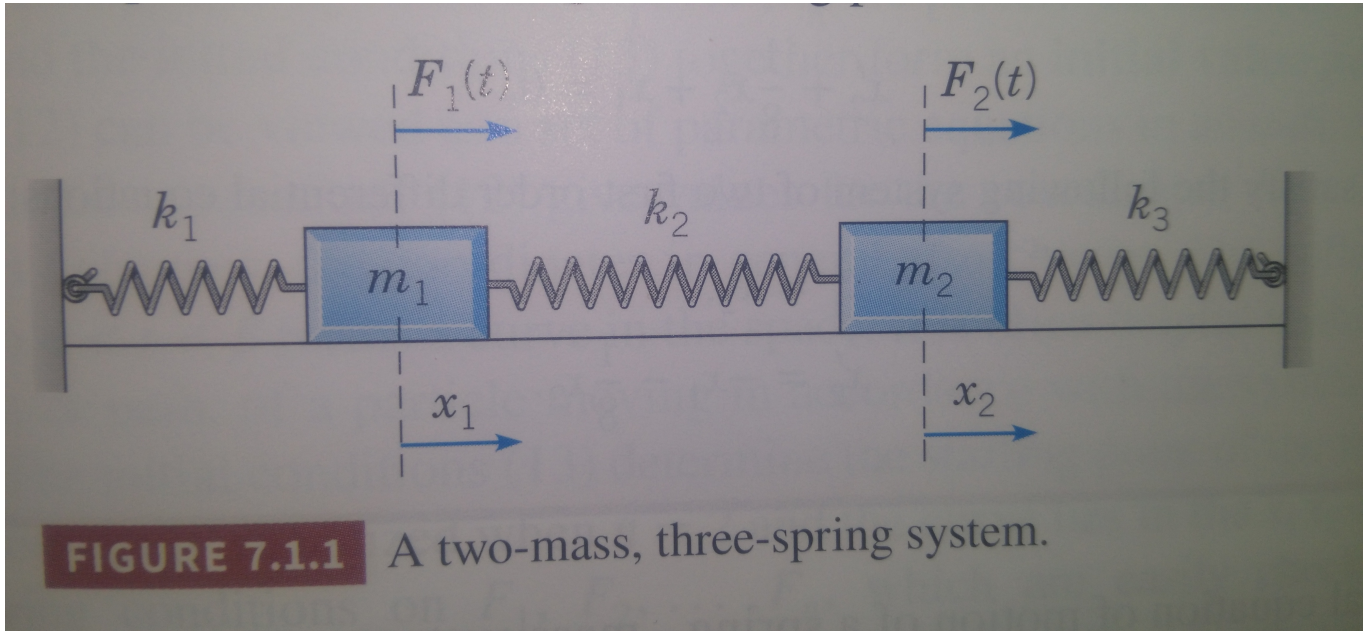


FIGURE 7.1.1 A two-mass, three-spring system.

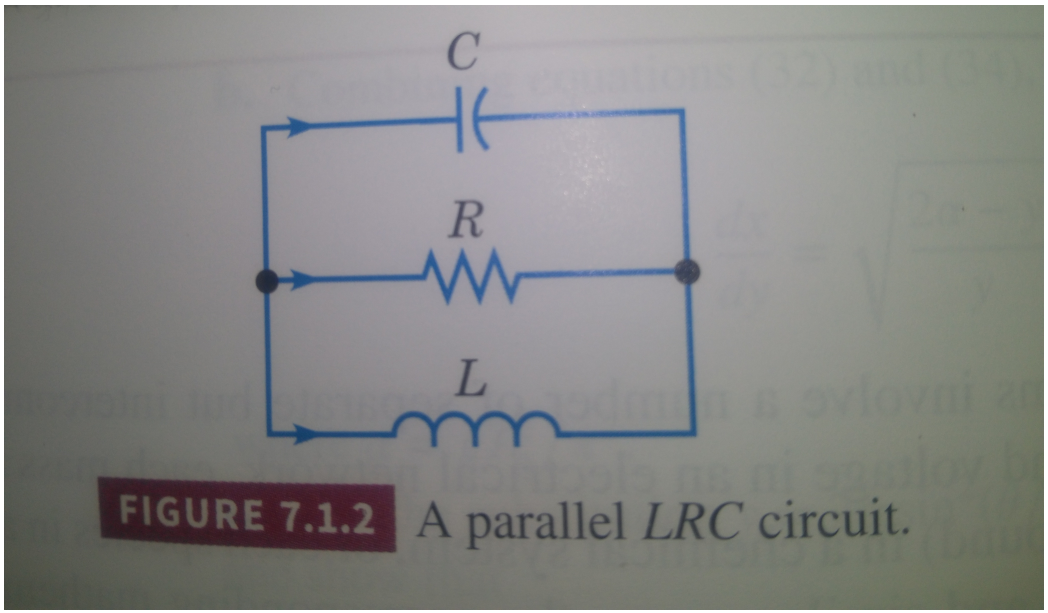


FIGURE 7.1.2 A parallel *LRC* circuit.

Courtesy of our textbook [BDM17]

Facts Already Known

1. Any IVP $\mathbf{y}' = \mathbf{A}\mathbf{y} + \mathbf{b}(t) \wedge \mathbf{y}(t_0) = \mathbf{y}_0$ has a unique maximal solution, which is defined whereever all coordinate functions of $\mathbf{b}(t)$ are defined. In particular, if $\mathbf{b}(t) \equiv \mathbf{b}$ is constant then the solution of the IVP is defined on $\mathbb{R}$.
2. The solutions of any homogeneous system $\mathbf{y}' = \mathbf{A}\mathbf{y}$ form an n-dimensional subspace S of the vectorial function space $(\mathbb{C}^n)^{\mathbb{R}}$ (consisting of all maps $f : \mathbb{R} \to \mathbb{C}^n$).
3. If $\mathbf{\Phi}(t)$ is a fundamental matrix of $\mathbf{y}' = \mathbf{A}\mathbf{y}$ (i.e., the columns of $\mathbf{\Phi}(t)$ form a basis of the solution space S of $\mathbf{y}' = \mathbf{A}\mathbf{y}$), the general solution of an associated inhomogeneous system $\mathbf{y}' = \mathbf{A}\mathbf{y} + \mathbf{b}(t)$ is

$$\mathbf{y}(t) = \mathbf{\Phi}(t)\left(\mathbf{c}_0 + \int_{t_0}^{t} \mathbf{\Phi}(s)^{-1}\mathbf{b}(s)\,\mathrm{d}s\right), \quad \mathbf{c}_0 \in \mathbb{C}^n.$$

Alternatively, if a particular solution $\mathbf{y}_{\mathrm{p}}(t)$ of $\mathbf{y}' = \mathbf{A}\mathbf{y} + \mathbf{b}(t)$ is known, the general solution of $\mathbf{y}' = \mathbf{A}\mathbf{y} + \mathbf{b}(t)$ is $\mathbf{y}(t) = \mathbf{\Phi}(t)\mathbf{c}_0 + \mathbf{y}_{\mathrm{p}}(t)$, $\mathbf{c}_0 \in \mathbb{C}^n$.

Facts Already Known (Cont'd)

While the preceding properties hold more generally for time-dependent systems $\mathbf{y}' = \mathbf{A}(t)\mathbf{y} + \mathbf{b}(t)$ (provided all coefficient functions $a_{ij}(t)$, $b_i(t)$ are considered for Property 1), the next property is a special feature of the time-independent case.

4. The matrix exponential function $t \mapsto \mathrm{e}^{\mathbf{A}t} = \sum_{k=0}^{\infty} \frac{t^k}{k!} \mathbf{A}^k$ satisfies $\mathbf{\Phi}'(t) = \mathbf{A}\mathbf{\Phi}(t)$, $\mathbf{\Phi}(0) = \mathbf{I}_n$, and hence provides a fundamental matrix for the system $\mathbf{y}' = \mathbf{A}\mathbf{y}$.

Problem

How to find an explicit fundamental matrix for $\mathbf{y}' = \mathbf{A}\mathbf{y}$ *?*
Equivalently, how to actually compute $\mathrm{e}^{\mathbf{A}t}$ *?*

Any two fundamental matrices $\mathbf{\Phi}_1, \mathbf{\Phi}_2$ are related by $\mathbf{\Phi}_1(t) = \mathbf{\Phi}_2(t)\mathbf{C}$ for some invertible $\mathbf{C} \in \mathbb{C}^{n\times n}$. The matrix $\mathbf{C}$ is the change-of-basis matrix from the ordered basis of S formed by the columns of $\mathbf{\Phi}_1(t)$ to that formed by the columns of $\mathbf{\Phi}_2(t)$. It is given by $\mathbf{C} = \mathbf{\Phi}_2(t_0)^{-1}\mathbf{\Phi}_1(t_0)$ for any $t_0 \in \mathbb{R}$.

Hence one fundamental matrix is as good as any other, and for any fundamental matrix $\mathbf{\Phi}(t)$ we have $\mathbf{\Phi}(t) = \mathrm{e}^{\mathbf{A}t}\mathbf{\Phi}(0)$, or $\mathbf{\Phi}(t)\mathbf{\Phi}(0)^{-1} = \mathrm{e}^{\mathbf{A}t}$.

A conceptual approach to solve the problem

First we look for instances of $\mathbf{y}' = \mathbf{Ay}$ that are easy to solve directly.

If $\mathbf{A}$ is a diagonal matrix, say with entries $\lambda_1, \ldots, \lambda_n$, then

$$\begin{pmatrix} y_1' \\ \vdots \\ y_n' \end{pmatrix} = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \lambda_1 y_1 \\ \vdots \\ \lambda_n y_n. \end{pmatrix}$$

So the system is "decoupled" into the n scalar ODE's $y_i' = \lambda_i y_i$, $1 \leq i \leq n$.

$\Longrightarrow$ The general solution is

$$\begin{pmatrix} y_1(t) \\ \vdots \\ y_n(t) \end{pmatrix} = \begin{pmatrix} c_1 \mathrm{e}^{\lambda_1 t} \\ \vdots \\ c_n \mathrm{e}^{\lambda_n t} \end{pmatrix} = \begin{pmatrix} \mathrm{e}^{\lambda_1 t} & & \\ & \ddots & \\ & & \mathrm{e}^{\lambda_n t} \end{pmatrix} \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix}.$$

From this we see that the diagonal matrix with i-th entry $\mathrm{e}^{\lambda_i t}$ (considered as a matrix function of $t \in \mathbb{R}$) is a fundamental matrix. Setting $t = 0$ gives the identity matrix. $\Longrightarrow$ This must be $\mathrm{e}^{\mathbf{A}t}$!

Independent verification

$$\exp\left[t\begin{pmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_n\end{pmatrix}\right] = \sum_{k=0}^{\infty}\frac{t^k}{k!}\begin{pmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_n\end{pmatrix}^k$$

$$= \sum_{k=0}^{\infty}\frac{t^k}{k!}\begin{pmatrix}\lambda_1^k & & \\ & \ddots & \\ & & \lambda_n^k\end{pmatrix}$$

$$= \begin{pmatrix}\sum_{k=0}^{\infty}\frac{t^k}{k!}\lambda_1^k & & \\ & \ddots & \\ & & \sum_{k=0}^{\infty}\frac{t^k}{k!}\lambda_n^k\end{pmatrix} = \begin{pmatrix}e^{\lambda_1 t} & & \\ & \ddots & \\ & & e^{\lambda_n t}\end{pmatrix}$$

Thus the problem is solved for diagonal matrices.

A conceptual approach (cont'd)

For a general matrix $\mathbf{A}$ we would like to find a coordinate transformation $\mathbf{y}(t) = \mathbf{S}\,\mathbf{z}(t)$, or $\mathbf{y} = \mathbf{Sz}$ for short, which puts $\mathbf{y}' = \mathbf{Ay}$ into simpler form (diagonal form, if possible). Of course, we must check whether the transformed system has the form $\mathbf{z}' = \mathbf{Bz}$ at all.

The matrix $\mathbf{S}$ must be invertible, i.e., the columns of $\mathbf{S}$ must form an (ordered) basis of $\mathbb{C}^n$. The matrix $\mathbf{S}$ then switches from this basis to the standard basis of $\mathbb{C}^n$.

$$\mathbf{y} = \mathbf{Sz} \implies \mathbf{y}' = \mathbf{Sz}' \implies \mathbf{z}' = \mathbf{S}^{-1}\mathbf{y}' = \mathbf{S}^{-1}\mathbf{Ay} = \mathbf{S}^{-1}\mathbf{ASz}$$

$\implies$ The new system has the desired form $\mathbf{z}' = \mathbf{Bz}$ with $\mathbf{B} = \mathbf{S}^{-1}\mathbf{AS}$.

We have met this situation in Linear Algebra, from which we recall the following:

- $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{n\times n}$ are *similar* if there exists an invertible matrix $\mathbf{S} \in \mathbb{C}^{n\times n}$ such that $\mathbf{B} = \mathbf{S}^{-1}\mathbf{AS}$.
- $\mathbf{A} \in \mathbb{C}^{n\times n}$ is *diagonalisable* if $\mathbf{A}$ is similar to a diagonal matrix.
- $\mathbf{S}^{-1}\mathbf{AS}$ is a diagonal matrix iff the columns of $\mathbf{S}$, which form a basis of $\mathbb{C}^n/\mathbb{C}$, are *eigenvectors* of $\mathbf{A}$. The corresponding eigenvalues, are the diagonal entries of $\mathbf{S}^{-1}\mathbf{AS}$, in order.

A conceptual approach (cont'd)

Conclusion: If $\mathbf{A}$ is similar to a diagonal matrix

$$\mathbf{B} = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}$$

and, secondly, we can compute the corresponding transform matrix $\mathbf{S}$, whose i-th column $\mathbf{v}_i$ must be an eigenvector of $\mathbf{A}$ for the eigenvalue λ_i, then we can solve $\mathbf{y}' = \mathbf{A}\mathbf{y}$ completely.
The general solution will be

$$\mathbf{y}(t) = \mathbf{S} \begin{pmatrix} c_1 \mathrm{e}^{\lambda_1 t} \\ \vdots \\ c_n \mathrm{e}^{\lambda_n t} \end{pmatrix} = c_1 \mathrm{e}^{\lambda_1 t}\mathbf{v}_1 + \cdots + c_n \mathrm{e}^{\lambda_n t}\mathbf{v}_n,$$

where $\mathbf{S} = (\mathbf{v}_1 | \ldots | \mathbf{v}_n)$, and a fundamental system of solutions will be $t \mapsto \mathrm{e}^{\lambda_1 t}\mathbf{v}_1, \ldots, t \mapsto \mathrm{e}^{\lambda_n t}\mathbf{v}_n$.
The vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ form the ordered basis of $\mathbb{C}^n$ corresponding to the coordinate transformation $\mathbf{y} = \mathbf{S}\mathbf{z}$ (which should be viewed as a coordinate transformation of $\mathbb{C}^n$ that gives rise to a corresponding transformation of functions).

A conceptual approach (cont'd)

The fundamental matrix of $\mathbf{y}' = \mathbf{A}\mathbf{y}$ just determined is

$$\mathbf{\Phi}(t) = \left(e^{\lambda_1 t}\mathbf{v}_1 | \ldots | e^{\lambda_n t}\mathbf{v}_n\right) = \mathbf{S}\begin{pmatrix} e^{\lambda_1 t} & & \\ & \ddots & \\ & & e^{\lambda_n t} \end{pmatrix}.$$

It follows that

$$e^{\mathbf{A}t} = \mathbf{\Phi}(t)\mathbf{\Phi}(0)^{-1} = \mathbf{S}\begin{pmatrix} e^{\lambda_1 t} & & \\ & \ddots & \\ & & e^{\lambda_n t} \end{pmatrix}\mathbf{S}^{-1}.$$

This can also be verified directly from the series representation:

$$\mathbf{S}^{-1}\mathbf{A}\mathbf{S} = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} \implies \mathbf{A} = \mathbf{S}\underbrace{\begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}}_{\mathbf{B}}\mathbf{S}^{-1}$$

$$\implies \mathbf{A}^k = (\mathbf{S}\mathbf{B}\mathbf{S}^{-1})(\mathbf{S}\mathbf{B}\mathbf{S}^{-1})\cdots(\mathbf{S}\mathbf{B}\mathbf{S}^{-1}) = \mathbf{S}\mathbf{B}^k\mathbf{S}^{-1}$$

A conceptual approach (cont'd)

$$\begin{aligned}
\Longrightarrow\ \mathrm{e}^{\mathbf{A}t} &= \sum_{k=0}^{\infty} \frac{t^k}{k!} \mathbf{S}\mathbf{B}^k\mathbf{S}^{-1} \\
&= \mathbf{S}\left(\sum_{k=0}^{\infty} \frac{t^k}{k!}\,\mathbf{B}^k\right)\mathbf{S}^{-1} \qquad (\star) \\
&= \mathbf{S}\mathrm{e}^{\mathbf{B}t}\mathbf{S}^{-1} = \mathbf{S}\begin{pmatrix} \mathrm{e}^{\lambda_1 t} & & \\ & \ddots & \\ & & \mathrm{e}^{\lambda_n t} \end{pmatrix}\mathbf{S}^{-1}
\end{aligned}$$

For the step tagged $(\star)$ we have used continuity of matrix multiplication, which implies that for a convergent sequence $\mathbf{X}_k \to \mathbf{X}$ of matrices $\mathbf{X}_k \in \mathbb{C}^{n\times n}$ we have $\mathbf{S}\mathbf{X}_k \to \mathbf{S}\mathbf{X}$, and similarly $\mathbf{X}_k\mathbf{S}^{-1} \to \mathbf{X}\mathbf{S}^{-1}$.

Example

After introducing the matrix exponential function we had seen that the 1st-order system

$$\begin{pmatrix} y_1' \\ y_2' \end{pmatrix} = \begin{pmatrix} y_2 \\ -y_1 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix},$$

which arises from $y'' + y = 0$, has matrix exponential function

$$\exp\left[t \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\right] = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix}.$$

Now we use the present approach to rederive this result.

The characteristic polynomial of $\mathbf{A} = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$ is

$$\chi_{\mathbf{A}}(X) = \begin{vmatrix} -X & 1 \\ -1 & -X \end{vmatrix} = X^2 + 1 = (X - \mathrm{i})(X + \mathrm{i})$$

so that the eigenvalues are $\lambda_1 = \mathrm{i}$, $\lambda_2 = -\mathrm{i}$.

$$\mathbf{A} - \mathrm{i}\,\mathbf{I} = \begin{vmatrix} -\mathrm{i} & 1 \\ -1 & -\mathrm{i} \end{vmatrix} \to \begin{vmatrix} -\mathrm{i} & 1 \\ 0 & 0 \end{vmatrix}, \quad \mathbf{A} + \mathrm{i}\,\mathbf{I} = \begin{vmatrix} \mathrm{i} & 1 \\ -1 & \mathrm{i} \end{vmatrix} \to \begin{vmatrix} \mathrm{i} & 1 \\ 0 & 0 \end{vmatrix}$$

Example (cont'd)

It follows that the eigenspace E_{i} (the right kernel of $\mathbf{A} - \mathrm{i}\mathbf{I}$) is spanned by $\mathbf{v}_1 := (1, \mathrm{i})^{\mathsf{T}}$, and $\mathrm{E}_{-\mathrm{i}}$ by $\mathbf{v}_2 = (1, -\mathrm{i})^{\mathsf{T}}$. The matrix $\mathbf{S} = \left(\begin{smallmatrix} 1 & 1 \\ \mathrm{i} & -\mathrm{i} \end{smallmatrix}\right)$ then diagonalizes $\mathbf{A}$, viz.

$$\begin{pmatrix} 1 & 1 \\ \mathrm{i} & -\mathrm{i} \end{pmatrix}^{-1} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ \mathrm{i} & -\mathrm{i} \end{pmatrix} = \frac{1}{-2\mathrm{i}} \begin{pmatrix} -\mathrm{i} & -1 \\ -\mathrm{i} & 1 \end{pmatrix} \begin{pmatrix} \mathrm{i} & -\mathrm{i} \\ -1 & -1 \end{pmatrix}$$
$$= \frac{\mathrm{i}}{2} \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix} = \begin{pmatrix} \mathrm{i} & 0 \\ 0 & -\mathrm{i} \end{pmatrix}$$

Our previous discussion yields that

$$\mathbf{y}_1(t) = \mathrm{e}^{\mathrm{i}t} \begin{pmatrix} 1 \\ \mathrm{i} \end{pmatrix}, \quad \mathbf{y}_2(t) = \mathrm{e}^{-\mathrm{i}t} \begin{pmatrix} 1 \\ -\mathrm{i} \end{pmatrix}$$

form a fundamental system of solutions of $\mathbf{y}' = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)\mathbf{y}$. The corresponding matrix exponential function is

$$\begin{pmatrix} \mathrm{e}^{\mathrm{i}t} & \mathrm{e}^{-\mathrm{i}t} \\ \mathrm{i}\,\mathrm{e}^{\mathrm{i}t} & -\mathrm{i}\,\mathrm{e}^{-\mathrm{i}t} \end{pmatrix} \begin{pmatrix} 1 & 1 \\ \mathrm{i} & -\mathrm{i} \end{pmatrix}^{-1} = \frac{1}{-2\mathrm{i}} \begin{pmatrix} \mathrm{e}^{\mathrm{i}t} & \mathrm{e}^{-\mathrm{i}t} \\ \mathrm{i}\,\mathrm{e}^{\mathrm{i}t} & -\mathrm{i}\,\mathrm{e}^{-\mathrm{i}t} \end{pmatrix} \begin{pmatrix} -\mathrm{i} & -1 \\ -\mathrm{i} & 1 \end{pmatrix}$$
$$= \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix}.$$

Example (cont'd)

Additional remarks:

- It is not necessary to compute $\mathbf{v}_2$; we can just take $\mathbf{v}_2 = \overline{\mathbf{v}}_1$. More generally, $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$ ($\mathbf{A}$ real, $\lambda, \mathbf{v}$ complex) implies $\mathbf{A}\overline{\mathbf{v}} = \overline{\mathbf{A}}\overline{\mathbf{v}} = \overline{\mathbf{A}\mathbf{v}} = \overline{\lambda\mathbf{v}} = \overline{\lambda}\overline{v}$.
- A real fundamental system of $\mathbf{y}' = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)\mathbf{y}$ can also be obtained by extracting from $\mathbf{y}_1(t)$ the real and imaginary part (more generally, provided $\mathbf{A}$ is real, from each pair of complex conjugate solutions the real and imaginary part of one of them).
- There is the simple matrix identity

$$\exp\left[t\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\right] = \cos t \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \sin t \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = (\cos t)\mathbf{I} + (\sin t)\mathbf{A}.$$

  We will see later that in general $n \times n$ matrix exponentials can be expressed as finite sums $\mathrm{e}^{\mathbf{A}t} = \sum_{k=0}^{n-1} c_k(t)\mathbf{A}^k$. This is surprising at the first glance, since $\mathrm{e}^{\mathbf{A}t}$ was defined by an infinite sum, and it does not mean that the matrix exponential series terminates after a finite number of summands.

Example

We determine a fundamental system of solutions of $\mathbf{y}' = \mathbf{A}\mathbf{y}$ for

$$\mathbf{A} = \begin{pmatrix} 2 & -2 & -16 \\ 0 & 1 & 6 \\ 0 & 0 & -2 \end{pmatrix} \in \mathbb{R}^{3\times 3}.$$

This matrix was one of the examples for eigenvalue/eigenvector computations of Math257 in Fall 2023. The triangular structure of $\mathbf{A}$ greatly facilitates the computations.

$$\chi_{\mathbf{A}}(X) = \begin{vmatrix} X-2 & 2 & 16 \\ 0 & X-1 & -6 \\ 0 & 0 & X+2 \end{vmatrix} = (X-2)(X-1)(X+2) = X^3 - X^2 - 4X + 4.$$

$$\Longrightarrow \lambda_1 = 2,\ \lambda_2 = 1,\ \lambda_3 = -2.$$

Example (cont'd)

Now we determine the corresponding eigenspaces E_{λ_i} :

$\underline{\lambda_1 = 2:}$

$$\mathbf{A} - 2\mathbf{I} = \begin{pmatrix} 0 & -2 & -16 \\ 0 & -1 & 6 \\ 0 & 0 & -4 \end{pmatrix}$$

This matrix has rank 2 and right kernel $\mathbb{R}(1,0,0)^{\mathsf{T}}$.

$\underline{\lambda_2 = 1:}$

$$\mathbf{A} - \mathbf{I} = \begin{pmatrix} 1 & -2 & -16 \\ 0 & 0 & 6 \\ 0 & 0 & -3 \end{pmatrix}$$

This matrix has rank 2 and right kernel $\mathbb{R}(2,1,0)^{\mathsf{T}}$.

$\underline{\lambda_3 = -2:}$

$$\mathbf{A} + 2\mathbf{I} = \begin{pmatrix} 4 & -2 & -16 \\ 0 & 3 & 6 \\ 0 & 0 & 0 \end{pmatrix} \to \begin{pmatrix} 2 & -1 & -8 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{pmatrix}$$

This matrix has rank 2 and right kernel $\mathbb{R}(-3,2,-1)^{\mathsf{T}}$.

In summary, we have shown $\mathrm{E}_2 = \mathbb{R}(1,0,0)^{\mathsf{T}}$, $\mathrm{E}_1 = \mathbb{R}(2,1,0)^{\mathsf{T}}$, and $\mathrm{E}_{-2} = \mathbb{R}(-3,2,-1)^{\mathsf{T}}$.

Example (cont'd)

Since the eigenvalues are distinct, the 3 eigenvectors found must be linearly independent and hence form a basis of $\mathbb{R}^3$. (This is also clear from the triangular form of the corresponding matrix $\mathbf{S}$.)

A fundamental system of solutions of $\mathbf{y}' = \mathbf{A}\mathbf{y}$ is then

$$\mathbf{y}_1(t) = \mathrm{e}^{2t}\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{y}_2(t) = \mathrm{e}^{t}\begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix}, \quad \mathbf{y}_3(t) = \mathrm{e}^{-2t}\begin{pmatrix} -3 \\ 2 \\ -1 \end{pmatrix}.$$

Afternote

It is instructive to show directly that if $\mathbf{v} = (v_1, v_2, v_3)^{\mathsf{T}} \in \mathbb{C}^3$ is an eigenvector of $\mathbf{A} \in \mathbb{C}^{3\times 3}$ with corresponding eigenvalue λ then $\mathbf{y}(t) = \mathrm{e}^{\lambda t}\mathbf{v}$ solves $\mathbf{y}' = \mathbf{A}\mathbf{y}$:

$$\mathbf{y}'(t) = \begin{pmatrix} \mathrm{e}^{\lambda t}v_1 \\ \mathrm{e}^{\lambda t}v_2 \\ \mathrm{e}^{\lambda t}v_3 \end{pmatrix}' = \begin{pmatrix} \lambda\mathrm{e}^{\lambda t}v_1 \\ \lambda\mathrm{e}^{\lambda t}v_2 \\ \lambda\mathrm{e}^{\lambda t}v_3 \end{pmatrix} = \lambda\,\mathbf{y}(t) = \mathbf{A}\,\mathbf{y}(t),$$

because $\mathrm{e}^{\lambda t}\mathbf{v} \in \mathrm{E}_\lambda$ as well.

Example (optional)

This example builds on the discrete analog $\mathbf{y}_{n+1} = \mathbf{A}\mathbf{y}_n$ (arising from the problem to determine the number of bit strings of length n of Hamming weight divisible by 3) considered in Math257 in Fall 2023.

Consider the linear 1st-order ODE system

$$\begin{pmatrix} y_1' \\ y_2' \\ y_3' \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}.$$

The matrix $\mathbf{A} = \left(\begin{smallmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{smallmatrix}\right)$ is the same as in the example from Fall 2023.

Since we had already determined a basis of $\mathbb{C}^3$ consisting of eigenvectors of $\mathbf{A}$, we can write down a fundamental system of solutions immediately.

Example (cont'd)

Inspecting the example, we obtain the general solution of the given system as

$$\begin{pmatrix} y_1'(t) \\ y_2'(t) \\ y_3'(t) \end{pmatrix} = c_1 \mathrm{e}^{2t} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + c_2 \mathrm{e}^{-\omega t} \begin{pmatrix} 1 \\ \omega \\ \omega^2 \end{pmatrix} + c_3 \mathrm{e}^{-\omega^2 t} \begin{pmatrix} 1 \\ \omega^2 \\ \omega \end{pmatrix}$$

with $c_1, c_2, c_3 \in \mathbb{C}$, where $\omega = \mathrm{e}^{2\pi \mathrm{i}/3} = \frac{-1+\mathrm{i}\sqrt{3}}{2}$.

A fundamental matrix is

$$\mathbf{\Phi}(t) = \left(\mathrm{e}^{2t}\mathbf{v}_1 | \mathrm{e}^{-\omega t}\mathbf{v}_2 | \mathrm{e}^{-\omega^2 t}\mathbf{v}_3\right) = \begin{pmatrix} \mathrm{e}^{2t} & \mathrm{e}^{-\omega t} & \mathrm{e}^{-\omega^2 t} \\ \mathrm{e}^{2t} & \omega \mathrm{e}^{-\omega t} & \omega^2 \mathrm{e}^{-\omega^2 t} \\ \mathrm{e}^{2t} & \omega^2 \mathrm{e}^{-\omega t} & \omega \mathrm{e}^{-\omega^2 t} \end{pmatrix},$$

and the canonical fundamental matrix is

$$\mathrm{e}^{\mathbf{A}t} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & \omega & \omega^2 \\ 1 & \omega^2 & \omega \end{pmatrix} \begin{pmatrix} \mathrm{e}^{2t} & & \\ & \mathrm{e}^{-\omega t} & \\ & & \mathrm{e}^{-\omega^2 t} \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ 1 & \omega & \omega^2 \\ 1 & \omega^2 & \omega \end{pmatrix}^{-1}.$$

Example (cont'd)

The matrix

$$\mathbf{S} = (\mathbf{v}_1|\mathbf{v}_2|\mathbf{v}_3) = \begin{pmatrix} 1 & 1 & 1 \\ 1 & \omega & \omega^2 \\ 1 & \omega^2 & \omega \end{pmatrix} = (\mathbf{v}_1|\mathbf{v}_2|\overline{\mathbf{v}}_2)$$

satisfies $\mathbf{S}\overline{\mathbf{S}} = 3\,\mathbf{I}_3$ (check it!), and hence (using $\omega^2 = \omega^{-1} = \overline{\omega}$)

$$\mathbf{S}^{-1} = \frac{1}{3}\overline{\mathbf{S}} = \frac{1}{3}\begin{pmatrix} 1 & 1 & 1 \\ 1 & \omega^2 & \omega \\ 1 & \omega & \omega^2 \end{pmatrix} = \frac{1}{3}\begin{pmatrix} \mathbf{v}_1^{\mathsf{T}} \\ \overline{\mathbf{v}}_2^{\mathsf{T}} \\ \mathbf{v}_2^{\mathsf{T}} \end{pmatrix}.$$

This gives

$$\begin{aligned} \mathrm{e}^{\mathbf{A}t} &= \frac{1}{3}\begin{pmatrix} 1 & 1 & 1 \\ 1 & \omega & \omega^2 \\ 1 & \omega^2 & \omega \end{pmatrix}\begin{pmatrix} \mathrm{e}^{2t} & & \\ & \mathrm{e}^{-\omega t} & \\ & & \mathrm{e}^{-\omega^2 t} \end{pmatrix}\begin{pmatrix} 1 & 1 & 1 \\ 1 & \omega^2 & \omega \\ 1 & \omega & \omega^2 \end{pmatrix} \\ &= \frac{1}{3}\left(\mathrm{e}^{2t}\,\mathbf{v}_1\mathbf{v}_1^{\mathsf{T}} + \mathrm{e}^{-\omega t}\,\mathbf{v}_2\overline{\mathbf{v}}_2^{\mathsf{T}} + \mathrm{e}^{-\omega^2 t}\,\overline{\mathbf{v}}_2\mathbf{v}_2^{\mathsf{T}}\right) \\ &= \frac{1}{3}\left[\mathrm{e}^{2t}\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} + \mathrm{e}^{-\omega t}\begin{pmatrix} 1 & \omega^2 & \omega \\ \omega & 1 & \omega^2 \\ \omega^2 & \omega & 1 \end{pmatrix} + \mathrm{e}^{-\omega^2 t}\begin{pmatrix} 1 & \omega & \omega^2 \\ \omega^2 & 1 & \omega \\ \omega & \omega^2 & 1 \end{pmatrix}\right]. \end{aligned}$$

Example (cont'd)

Alternative representations are

$$e^{\mathbf{A}t} = \frac{1}{3}(e^{2t} + e^{-\omega t} + e^{-\omega^2 t})\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix} + \frac{1}{3}(e^{2t} + \omega e^{-\omega t} + \omega^2 e^{-\omega^2 t})\begin{pmatrix}0&0&1\\1&0&0\\0&1&0\end{pmatrix} + \frac{1}{3}(e^{2t} + \omega^2 e^{-\omega t} + \omega e^{-\omega^2 t})\begin{pmatrix}0&1&0\\0&0&1\\1&0&0\end{pmatrix}$$

$$= \frac{1}{3}(e^{2t} - 2\omega e^{-\omega t} - 2\omega^2 e^{-\omega^2 t})\mathbf{I}_3 + \frac{1}{3}\big(-e^{2t} + (2+3\omega)e^{-\omega t} + (2+3\omega^2)e^{-\omega^2 t}\big)\mathbf{A} + \frac{1}{3}(e^{2t} + \omega^2 e^{-\omega t} + \omega e^{-\omega^2 t})\mathbf{A}^2.$$

Finally, note that the matrix $\mathbf{S}$ simultaneously diagonalizes $\mathbf{A}$ and $e^{\mathbf{A}t}$. So we also have

$$\mathbf{A} = \begin{pmatrix}1&0&1\\1&1&0\\0&1&1\end{pmatrix} = \mathbf{S}\begin{pmatrix}2&&\\&-\omega&\\&&-\omega^2\end{pmatrix}\mathbf{S}^{-1}$$

$$= \frac{1}{3}\left[2\begin{pmatrix}1&1&1\\1&1&1\\1&1&1\end{pmatrix} - \omega\begin{pmatrix}1&\omega^2&\omega\\\omega&1&\omega^2\\\omega^2&\omega&1\end{pmatrix} - \omega^2\begin{pmatrix}1&\omega&\omega^2\\\omega^2&1&\omega\\\omega&\omega^2&1\end{pmatrix}\right].$$

Example (cont'd)

This time (Spring 2022) I made two further notes on the example, which are reproduced here in detail:

1. The eigenvalues and eigenvectors of $\mathbf{A} = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}$ can of course be determined without any knowledge about circulant matrices. The following argument avoids computing $\chi_{\mathbf{A}}(X)$.

   Since $\mathbf{A}$ has constant row sum 2, $\lambda_1 = 2$ is an eigenvalue of $\mathbf{A}$ with associated eigenvector $\mathbf{v}_1 = (1,1,1)^{\mathsf{T}}$. The remaining eigenvalues can be determined from

$$\lambda_1 + \lambda_2 + \lambda_3 = \operatorname{tr}(\mathbf{A}) = 3,$$
$$\lambda_1\lambda_2\lambda_3 = \det(\mathbf{A}) = 1 + 1 + 0 - 0 - 0 - 0 = 2.$$

   This gives $\lambda_2 + \lambda_3 = \lambda_2\lambda_3 = 1$, so that λ_2, λ_3 are the roots of $X^2 - X + 1 = 0$, i.e., $\lambda_{2/3} = \frac{1 \pm \mathrm{i}\sqrt{3}}{2}$.

$$\mathbf{A} - \lambda_2\mathbf{I} = \begin{bmatrix} \frac{1-\mathrm{i}\sqrt{3}}{2} & 0 & 1 \\ 1 & \frac{1-\mathrm{i}\sqrt{3}}{2} & 0 \\ 0 & 1 & \frac{1-\mathrm{i}\sqrt{3}}{2} \end{bmatrix}$$

$$\Longrightarrow \mathbf{v}_2 = \left(1, \tfrac{1+\mathrm{i}\sqrt{3}}{2}, \tfrac{-1+\mathrm{i}\sqrt{3}}{2}\right)^{\mathsf{T}}$$

Example (cont'd)

1. (cont'd)
   For this note that it suffices if $\mathbf{v}_2$ is orthogonal to two rows of $\mathbf{A} - \lambda_2\mathbf{I}$.

   Because $\mathbf{A}$ is real, the eigenvector $\mathbf{v}_3$ associated with $\lambda_3 = \overline{\lambda}_2$ must be $\mathbf{v}_3 = \overline{\mathbf{v}}_2 = \left(1, \frac{1-\mathrm{i}\sqrt{3}}{2}, \frac{-1-\mathrm{i}\sqrt{3}}{2}\right)^{\mathsf{T}}$.

   For this note that $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$ implies $\mathbf{A}\overline{\mathbf{v}} = \overline{\mathbf{A}}\overline{\mathbf{v}} = \overline{\mathbf{A}\mathbf{v}} = \overline{\lambda\mathbf{v}} = \overline{\lambda}\overline{\mathbf{v}}$.

Example (cont'd)

2. A real fundamental system of solutions of $\mathbf{y}' = \mathbf{A}\mathbf{y}$ (different from that formed by the columns of $\mathrm{e}^{\mathbf{A}t}$) can be obtained from the given complex fundamental system by extracting the real and imaginary part of the complex solution:

$$\mathbf{y}_1(t) = \mathrm{e}^{2t}\begin{pmatrix}1\\1\\1\end{pmatrix},$$

$$\mathbf{y}_2(t) = \mathrm{Re}\left[\mathrm{e}^{-\omega t}\begin{pmatrix}1\\\omega\\\omega^2\end{pmatrix}\right] = \mathrm{Re}\left[\mathrm{e}^{t/2}\mathrm{e}^{-\mathrm{i}\sqrt{3}\,t/2}\begin{pmatrix}1\\\mathrm{e}^{2\pi\mathrm{i}/3}\\\mathrm{e}^{4\pi\mathrm{i}/3}\end{pmatrix}\right]$$

$$= \mathrm{e}^{t/2}\begin{pmatrix}\cos(\sqrt{3}\,t/2)\\\cos(\sqrt{3}\,t/2+4\pi/3)\\\cos(\sqrt{3}\,t/2+2\pi/3)\end{pmatrix},$$

$$\mathbf{y}_3(t) = \mathrm{Im}\left[\mathrm{e}^{-\omega t}\begin{pmatrix}1\\\omega\\\omega^2\end{pmatrix}\right] = -\mathrm{e}^{t/2}\begin{pmatrix}\sin(\sqrt{3}\,t/2)\\\sin(\sqrt{3}\,t/2+4\pi/3)\\\sin(\sqrt{3}\,t/2+2\pi/3)\end{pmatrix}.$$

The same works mutatis mutandis for any real $n \times n$ matrix.

General Solution of $\mathbf{y}' = \mathbf{A}\mathbf{y}$

Recall that the solution space of an n-th order scalar homogeneous linear ODE $a(\mathrm{D})y = 0$ (with constant coefficients) is generated by the exponential polynomials $t^k \mathrm{e}^{\lambda t}$ with $\lambda \in \mathbb{C}$ a root of $a(X)$ and k a non-negative integer less than the (algebraic) multiplicity of λ.

Order reduction gives the 1st-order $n \times n$ system

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \\ y_n \end{pmatrix}' = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & 0 & \dots & 0 & 1 \\ -a_0 & -a_1 & \dots & -a_{n-2} & -a_{n-1} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \\ y_n \end{pmatrix}.$$

The coefficient matrix $\mathbf{A}$ is the *companion matrix* of the polynomial $a(X) = X^n + a_{n-1}X^{n-1} + \cdots + a_1 X + a_0$ and has characteristic polynomial equal to $a(X)$; cf. Linear Algebra.

In the special case under consideration there is a fundamental system of $\mathbf{y}' = \mathbf{A}\mathbf{y}$ consisting of vectorial functions of the form

$$\mathbf{y}(t) = \left(t^k \mathrm{e}^{\lambda t}, (t^k \mathrm{e}^{\lambda t})', \ldots, (t^k \mathrm{e}^{\lambda t})^{(n-1)}\right),$$

which have exponential polynomials $p(t)\mathrm{e}^{\lambda t}$ with λ an eigenvalue of $\mathbf{A}$ and polynomial factor $p(t)$ of degree less than the algebraic multiplicity of λ as entries.

This motivates the „Ansatz“

$$\mathbf{y}(t) = \mathrm{e}^{\lambda t}\mathbf{v}_0 + t\,\mathrm{e}^{\lambda t}\mathbf{v}_1 + \cdots + t^{m-1}\,\mathrm{e}^{\lambda t}\mathbf{v}_{m-1}, \quad \mathbf{v}_j \in \mathbb{C}^n,$$

for eigenvalues λ of $\mathbf{A}$ of algebraic multiplicity m to solve $\mathbf{y}' = \mathbf{A}\mathbf{y}$.

$$\begin{aligned}
\mathbf{y}'(t) &= \lambda \mathrm{e}^{\lambda t}\mathbf{v}_0 + (1+\lambda t)\mathrm{e}^{\lambda t}\mathbf{v}_1 + (2t+\lambda t^2)\mathrm{e}^{\lambda t}\mathbf{v}_2 + \cdots + \\
&\quad + \big((m-1)t^{m-2} + \lambda t^{m-1}\big)\mathrm{e}^{\lambda t}\mathbf{v}_{m-1} \\
&= (\lambda\mathbf{v}_0 + \mathbf{v}_1)\mathrm{e}^{\lambda t} + (\lambda\mathbf{v}_1 + 2\mathbf{v}_2)t\,\mathrm{e}^{\lambda t} + \cdots + \\
&\quad + \big(\lambda\mathbf{v}_{m-2} + (m-1)\mathbf{v}_{m-1}\big)t^{m-2}\,\mathrm{e}^{\lambda t} + \lambda\mathbf{v}_{m-1}t^{m-1}\,\mathrm{e}^{\lambda t} \\
\mathbf{A}\mathbf{y}(t) &= \mathrm{e}^{\lambda t}\mathbf{A}\mathbf{v}_0 + t\,\mathrm{e}^{\lambda t}\mathbf{A}\mathbf{v}_1 + \cdots + t^{m-1}\,\mathrm{e}^{\lambda t}\mathbf{A}\mathbf{v}_{m-1}
\end{aligned}$$

$\Longrightarrow$ $\mathbf{y}(t)$ solves $\mathbf{y}' = \mathbf{A}\mathbf{y}$ iff

$$\begin{aligned}
\mathbf{A}\mathbf{v}_0 &= \lambda\mathbf{v}_0 + \mathbf{v}_1,\\
\mathbf{A}\mathbf{v}_1 &= \lambda\mathbf{v}_1 + 2\mathbf{v}_2,\\
&\vdots\\
\mathbf{A}\mathbf{v}_{m-2} &= \lambda\mathbf{v}_{m-2} + (m-1)\mathbf{v}_{m-1},\\
\mathbf{A}\mathbf{v}_{m-1} &= \lambda\mathbf{v}_{m-1}.
\end{aligned}$$

This can be rewritten as $(\mathbf{A} - \lambda\mathbf{I}_n)\mathbf{v}_0 = \mathbf{v}_1$, $(\mathbf{A} - \lambda\mathbf{I}_n)\mathbf{v}_1 = 2\mathbf{v}_2$, $\ldots$, $(\mathbf{A} - \lambda\mathbf{I}_n)\mathbf{v}_{m-2} = (m-1)\mathbf{v}_{m-1}$, $(\mathbf{A} - \lambda\mathbf{I}_n)\mathbf{v}_{m-1} = \mathbf{0}$ and is equivalent to

$$\mathbf{v}_k = \frac{1}{k!}(\mathbf{A} - \lambda\mathbf{I}_n)^k\mathbf{v}_0 \quad \text{for } 1 \le k \le m-1, \quad (\mathbf{A} - \lambda\mathbf{I}_n)^m\mathbf{v}_0 = \mathbf{0}.$$

In particular, $\mathbf{v}_0$ must be taken as a generalized eigenvector of $\mathbf{A}$ for the eigenvalue λ.

On the next slide we recall the most important facts about generalized eigenvectors, which were derived in Linear Algebra.

Generalized Eigenspaces

Suppose $\mathbf{A} \in \mathbb{C}^{n\times n}$ has characteristic polynomial

$$\chi_{\mathbf{A}}(X) = \prod_{i=1}^{r} (X - \lambda_i)^{m_i}$$

with $\lambda_1, \ldots, \lambda_r$ distinct. Thus λ_i, $1 \le i \le r$, are precisely the eigenvalues of $\mathbf{A}$ and m_i the corresponding *algebraic multiplicities*.

- A vector $\mathbf{v} \in \mathbb{C}^n \setminus \{\mathbf{0}\}$ is said to be a *generalized eigenvector* of $\mathbf{A}$ associated to the eigenvalue λ_i if $(\mathbf{A} - \lambda_i \mathbf{I}_n)^{m_i}\mathbf{v} = \mathbf{0}$. The solution space of $(\mathbf{A} - \lambda_i \mathbf{I}_n)^{m_i}\mathbf{x} = \mathbf{0}$ (right kernel of the matrix $(\mathbf{A} - \lambda_i \mathbf{I}_n)^{m_i}$), which also includes $\mathbf{0}$, is called *generalized eigenspace* of $\mathbf{A}$ for λ_i and denoted by G_{λ_i};
- $\dim(\mathrm{G}_{\lambda_i}) = m_i$ for $1 \le i \le r$;
- $\mathbb{C}^n = \mathrm{G}_{\lambda_1} \oplus \mathrm{G}_{\lambda_2} \oplus \cdots \oplus \mathrm{G}_{\lambda_r}$.

The latter means that every vector $\mathbf{v} \in \mathbb{C}^n$ has a unique representation $\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2 + \cdots + \mathbf{v}_r$ with $\mathbf{v}_i \in \mathrm{G}_{\lambda_1}$.

Notes

- $\mathrm{G}_{\lambda_i} \supseteq \mathrm{E}_{\lambda_i}$, and $\mathbf{A}$ is diagonalizable iff $\mathrm{G}_{\lambda_i} = \mathrm{E}_{\lambda_i}$ for all $i \in \{1, 2, \ldots, r\}$.
- If λ_i is a simple root of $\chi_{\mathbf{A}}(X)$ (i.e., $m_i = 1$), eigenvectors and generalized eigenvectors of $\mathbf{A}$ associated to λ_i are the same thing (and thus $\mathrm{G}_{\lambda_i} = \mathrm{E}_{\lambda_i}$).
- In general, writing $\lambda_i = \lambda$ and $m_i = m$, we have the chain of subspaces

$$\mathrm{E}_\lambda = \mathrm{rker}(\mathbf{A} - \lambda \mathbf{I}_n) \subseteq \mathrm{rker}(\mathbf{A} - \lambda \mathbf{I}_n)^2 \subseteq \cdots \subseteq \mathrm{rker}(\mathbf{A} - \lambda \mathbf{I}_n)^m = \mathrm{G}_\lambda.$$

- $\mathbb{C}^n = \mathrm{G}_{\lambda_1} \oplus \mathrm{G}_{\lambda_2} \oplus \cdots \oplus \mathrm{G}_{\lambda_r}$ can be rephrased as follows: $\mathbb{C}^n/\mathbb{C}$ *has a basis consisting of generalized eigenvectors of* $\mathbf{A}$.

Theorem

Suppose $\mathbf{A} \in \mathbb{C}^{n\times n}$ *and* $B = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ *is a basis of* $\mathbb{C}^n/\mathbb{C}$ *consisting of generalized eigenvectors of* $\mathbf{A}$*.*

1. *If* $\mathbf{v}_j \in B$ *is associated to the eigenvalue* λ_i *of* $\mathbf{A}$ *and* $\mathbf{y}_j : \mathbb{R} \to \mathbb{C}^n$ *is defined by*

$$\mathbf{y}_j(t) = \sum_{k=0}^{m_i-1} \frac{1}{k!} t^k \, \mathrm{e}^{\lambda_i t} (\mathbf{A} - \lambda_i \mathbf{I}_n)^k \mathbf{v}_j,$$

   then $\mathbf{y}_1, \ldots, \mathbf{y}_n$ *form a fundamental system of solutions of* $\mathbf{y}' = \mathbf{A}\mathbf{y}$*.*

2. *The matrix exponential function of* $\mathbf{A}$ *is*

$$t \mapsto \mathrm{e}^{\mathbf{A}t} = \big(\mathbf{y}_1(t) | \ldots | \mathbf{y}_n(t)\big) \big(\mathbf{v}_1 | \ldots | \mathbf{v}_n\big)^{-1}.$$

Proof of the theorem.

We have already seen that the functions $\mathbf{y}_j$ solve $\mathbf{y}' = \mathbf{A}\mathbf{y}$.
It remains to show that they are linearly independent. We have

$$\mathbf{y}_j(t) = \mathrm{e}^{\lambda_i t}\mathbf{v}_j + t\,\mathrm{e}^{\lambda_i t}\mathbf{w}_1 + \cdots + t^{m_i-1}\,\mathrm{e}^{\lambda_i t}\mathbf{w}_{m_i-1}$$

for certain vectors $\mathbf{w}_1, \ldots, \mathbf{w}_{m_i-1} \in \mathbb{C}^n$.
$\Longrightarrow \mathbf{y}_j(0) = \mathbf{v}_j$.
Since $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are linearly independent, so are $\mathbf{y}_1, \ldots, \mathbf{y}_n$.
This proves (1);
(2) is an instance of the formula $\mathrm{e}^{\mathbf{A}t} = \mathbf{\Phi}(t)\mathbf{\Phi}(0)^{-1}$. □

Example

We determine a fundamental system of solutions of

$$\mathbf{y}' = \begin{pmatrix} 1 & -1 \\ 1 & 3 \end{pmatrix} \mathbf{y}; \qquad \text{cf. [BDM17], p. 336.}$$

Here the characteristic polynomial is $\chi_{\mathbf{A}}(X) = X^2 - 4X + 4 = (X-2)^2$, so that $\mathbf{A} = \left(\begin{smallmatrix} 1 & -1 \\ 1 & 3 \end{smallmatrix}\right)$ has the single eigenvalue $\lambda = 2$ with algebraic multiplicity $m = 2$.
$\Longrightarrow$ The corresponding generalized eigenspace must be $\mathbb{C}^2$, and, using for B the standard basis of $\mathbb{C}^2$ the theorem gives

$$\mathbf{y}_1(t) = \mathrm{e}^{2t}\begin{pmatrix} 1 \\ 0 \end{pmatrix} + t\,\mathrm{e}^{2t}\begin{pmatrix} -1 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \mathrm{e}^{2t}\begin{pmatrix} 1 \\ 0 \end{pmatrix} + t\,\mathrm{e}^{2t}\begin{pmatrix} -1 \\ 1 \end{pmatrix},$$

$$\mathbf{y}_2(t) = \mathrm{e}^{2t}\begin{pmatrix} 0 \\ 1 \end{pmatrix} + t\,\mathrm{e}^{2t}\begin{pmatrix} -1 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 0 \\ 1 \end{pmatrix} = \mathrm{e}^{2t}\begin{pmatrix} 0 \\ 1 \end{pmatrix} + t\,\mathrm{e}^{2t}\begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

as fundamental system of solutions.

We must have $\mathbf{\Phi}(t) := \big(\mathbf{y}_1(t)|\mathbf{y}_2(t)\big) = \mathrm{e}^{\mathbf{A}t}$, since $\mathbf{\Phi}(0) = \big(\mathbf{y}_1(0)|\mathbf{y}_2(0)\big) = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$, and $\mathrm{e}^{\mathbf{A}t}$ is characterized by this condition among the fundamental matrices of $\mathbf{y}' = \mathbf{A}\mathbf{y}$.

Example (cont'd)

$$\Longrightarrow \exp\left[t\begin{pmatrix} 1 & -1 \\ 1 & 3 \end{pmatrix}\right] = \mathrm{e}^{2t}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + t\,\mathrm{e}^{2t}\begin{pmatrix} -1 & -1 \\ 1 & 1 \end{pmatrix}$$
$$= \begin{pmatrix} \mathrm{e}^{2t} - t\,\mathrm{e}^{2t} & -t\,\mathrm{e}^{2t} \\ t\,\mathrm{e}^{2t} & \mathrm{e}^{2t} + t\,\mathrm{e}^{2t} \end{pmatrix}.$$

The eigenspace E_2 is 1-dimensional and generated by $(1,-1)^{\mathsf{T}}$, so that we can replace one of $\mathbf{y}_1, \mathbf{y}_2$, say $\mathbf{y}_2$, by the "simpler" solution $\mathbf{y}(t) = \mathrm{e}^{2t}(1,-1)^{\mathsf{T}}$. (This amounts to applying the theorem to the basis $(1,0)^{\mathsf{T}}$, $(1,-1)^{\mathsf{T}}$ of $\mathbb{C}^2$ instead.)

The corresponding fundamental matrix is

$$\begin{pmatrix} \mathrm{e}^{2t} - t\,\mathrm{e}^{2t} & -t\,\mathrm{e}^{2t} \\ t\,\mathrm{e}^{2t} & \mathrm{e}^{2t} + t\,\mathrm{e}^{2t} \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} \mathrm{e}^{2t} - t\,\mathrm{e}^{2t} & \mathrm{e}^{2t} \\ t\,\mathrm{e}^{2t} & -\mathrm{e}^{2t} \end{pmatrix}.$$

Notes on the theorem

1. The required basis B can be calculated by determining, for each $i \in \{1, \ldots, r\}$, a basis of the solution space of $(\mathbf{A} - \lambda_i \mathbf{I}_n)^{m_i}\mathbf{x} = \mathbf{0}$. This is done with the usual algorithm based on Gaussian elimination.
2. If the basis vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are indexed in such a way that $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{m_1}$ form a basis of G_{λ_1}, $\mathbf{v}_{m_1+1}, \mathbf{v}_{m_1+2}, \ldots, \mathbf{v}_{m_1+m_2}$ a basis of G_{λ_2}, etc., then $\mathbf{S} = (\mathbf{v}_1 | \ldots | \mathbf{v}_n)$ "block-diagonalizes" $\mathbf{A}$ in the following sense:

$$\mathbf{S}^{-1}\mathbf{A}\mathbf{S} = \begin{pmatrix} \mathbf{A}_1 & & & \\ & \mathbf{A}_2 & & \\ & & \ddots & \\ & & & \mathbf{A}_r \end{pmatrix} \quad \text{with } \mathbf{A}_i \in \mathbb{C}^{m_i \times m_i}.$$

Moreover, the characteristic polynomial of $\mathbf{A}_i$ is $(X - \lambda_i)^{m_i}$. The block-diagonal form expresses the fact that $f_{\mathbf{A}}$ maps the generalized eigenspaces of $\mathbf{A}$ to itself:

$$(\mathbf{A} - \lambda\mathbf{I}_n)^m\mathbf{v} = \mathbf{0} \implies (\mathbf{A} - \lambda\mathbf{I}_n)^m\mathbf{A}\mathbf{v} = \mathbf{A}(\mathbf{A} - \lambda\mathbf{I}_n)^m\mathbf{v} = \mathbf{0}.$$

Notes on the theorem cont'd

3. A different way to calculate $e^{\mathbf{A}t}$ is as follows: From the preceding note we have

$$\mathbf{A} = \mathbf{S}\begin{pmatrix} \mathbf{A}_1 & & \\ & \ddots & \\ & & \mathbf{A}_r \end{pmatrix}\mathbf{S}^{-1} \Longrightarrow e^{\mathbf{A}t} = \mathbf{S}\begin{pmatrix} e^{\mathbf{A}_1 t} & & \\ & \ddots & \\ & & e^{\mathbf{A}_r t} \end{pmatrix}\mathbf{S}^{-1}.$$

Further, since $(\mathbf{A}_i - \lambda_i \mathbf{I})^{m_i} = \mathbf{0}$, where $\mathbf{I} = \mathbf{I}_{m_i}$, we have

$$\begin{aligned} e^{\mathbf{A}_i t} &= e^{\lambda_i t \mathbf{I}} e^{\mathbf{A}_i t - \lambda_i t \mathbf{I}} = (e^{\lambda_i t}\mathbf{I}) e^{t(\mathbf{A}_i - \lambda_i \mathbf{I})} \\ &= e^{\lambda_i t} \sum_{k=0}^{m_i - 1} \frac{1}{k!} t^k (\mathbf{A}_i - \lambda_i \mathbf{I})^k = \sum_{k=0}^{m_i - 1} \frac{1}{k!} t^k e^{\lambda_i t} (\mathbf{A}_i - \lambda_i \mathbf{I})^k. \end{aligned}$$

The exponential series terminates, since $(\mathbf{A}_i - \lambda_i \mathbf{I})^k = \mathbf{0}$ for $k \geq m_i$.

The solutions $\mathbf{y}_j(t)$ in Part 2 of the theorem are in fact the columns of $e^{\mathbf{A}t}\mathbf{S}$, as is clear from Part 3 of the theorem.

Notes on the theorem cont'd

3. (cont'd)
   This can also be seen directly as follows: If $\mathbf{v}_j$ is the j-th column of $\mathbf{S}$ and belongs to the eigenvalue λ_i, we have

$$e^{\mathbf{A}t}\mathbf{v}_j = e^{\lambda_i t \mathbf{I}} e^{(\mathbf{A}-\lambda_i \mathbf{I})t}\mathbf{v}_j = e^{\lambda_i t}\sum_{k=0}^{\infty}\frac{1}{k!}t^k(\mathbf{A}-\lambda_i\mathbf{I})^k\mathbf{v}_j$$

$$= e^{\lambda_i t}\sum_{k=0}^{m_i-1}\frac{1}{k!}t^k(\mathbf{A}-\lambda_i\mathbf{I})^k\mathbf{v}_j. \quad \text{(since } (\mathbf{A}-\lambda_i\mathbf{I})^{m_i}\mathbf{v}_j = \mathbf{0}\text{)}$$

   This is precisely $\mathbf{y}_j(t)$, as defined in Part 2 of the theorem.

Notes on the theorem cont'd

5. The vectors $\mathbf{w}_k = (\mathbf{A} - \lambda_i \mathbf{I}_n)^k \mathbf{v}_j$, $0 \le k \le m_i - 1$ (with $\mathbf{w}_0 = \mathbf{v}_j$), which need to be calculated in order to obtain $\mathbf{y}_j(t)$, are itself generalized eigenvectors associated to λ_i and hence can serve as members of the basis B, provided they are nonzero. (If you find the reasoning circular, change to "serve as members of another basis B' consisting of generalized eigenvectors of $\mathbf{A}$".)

   Suppose that the sum defining $\mathbf{y}_j(t)$ terminates with the summand $\frac{1}{k!} t^k \mathrm{e}^{\lambda_i t} \mathbf{w}_k$, i.e., $\mathbf{w}_k \neq \mathbf{0}$, $\mathbf{w}_{k+1} = \mathbf{0}$. Then the vectors in the *chain* $\mathbf{w}_0, \mathbf{w}_1, \ldots, \mathbf{w}_k$ are linearly independent. This can be seen as follows: Writing $\mathbf{N} = \mathbf{A} - \lambda_i \mathbf{I}_n$, we have $\mathbf{N}\mathbf{w}_s = \mathbf{w}_{s+1}$. If $\sum_{s=0}^{k} c_s \mathbf{w}_s = \mathbf{0}$, we can apply $\mathbf{N}^k$ to this sum and from $\sum_{s=0}^{k} c_s \mathbf{w}_{s+k} = c_0 \mathbf{w}_k = \mathbf{0}$ conclude that $c_0 = 0$. Then we apply $\mathbf{N}^{k-1}$ and obtain $c_1 = 0$, etc.

   If $k = m_i - 1$, the chain forms a basis of G_{λ_i}. If $k < m_i - 1$, this is not the case, but we can use several such chains, starting with other vectors $\mathbf{v}_l \in \mathrm{G}_{\lambda_i}$.

Notes on the theorem cont'd

5. (cont'd)
   The following facts, which are readily proved, provide the key to success of this approach.

   - The last nonzero vector of each chain is an eigenvector corresponding to the eigenvalue λ_i.
   - The vectors in a union of chains (belonging to the same λ_i) are linearly independent iff the corresponding eigenvectors (last vectors of the chains) are linearly independent.
   - There exists a basis of G_{λ_i} that is a union of chains, and the number and lengths of the chains in such a basis are uniquely determined.

   The number of chains is equal to the geometric multiplicity of λ_i, and the lengths of the chains can be determined from the dimensions of $\mathrm{rker}(\mathbf{A} - \lambda_i \mathbf{I})^k$, $1 \leq k \leq m_i$.

   The matrix respresenting $f_{\mathbf{A}}$ w.r.t. such a basis is in Jordan Canonical Form (see subsequent section), with the number/sizes of the Jordan blocks equal to the number/lengths of the chains.

Notes on the theorem cont'd

5. (cont'd)
The preceding observation motivates the following "depth-first" strategy for obtaining a basis of G_{λ_i} :

First determine the smallest non-negative integer k such that $\operatorname{rker}\big((\mathbf{A}-\lambda_i\mathbf{I}_n)^k\big) = \mathrm{G}_{\lambda_i}$ (equivalently, $\operatorname{rker}\big((\mathbf{A}-\lambda_i\mathbf{I}_n)^k\big)$ has dimension m_i), and a vector $\mathbf{w} \in \mathrm{G}_{\lambda_i}$ satisfying $(\mathbf{A}-\lambda_i\mathbf{I}_n)^{k-1}\mathbf{w} \neq \mathbf{0}$. Include the vectors $\mathbf{w}_0 = \mathbf{w}, \mathbf{w}_1, \ldots, \mathbf{w}_{k-1}$ as defined above in the basis. If $k < m_i$, start over and determine the largest non-negative integer k' for which there exists a vector $\mathbf{w}' \in \mathrm{G}_{\lambda_i}$ such that $(\mathbf{A}-\lambda_i\mathbf{I}_n)^{k'-1}\mathbf{w}'$ is linearly independent of $(\mathbf{A}-\lambda_i\mathbf{I}_n)^{k-1}\mathbf{w}$. Include $\mathbf{w}'_0 = \mathbf{w}', \mathbf{w}'_1, \ldots, \mathbf{w}'_{k'-1}$ as defined above in the basis; etc. Clearly the procedure terminates, and it can be shown that it yields a basis of G_{λ_i}. The corresponding fundamental solutions of $\mathbf{y}' = \mathbf{A}\mathbf{y}$ have the simple form

$$\begin{aligned}
\mathbf{y}_0(t) &= \mathrm{e}^{\lambda_i t}\mathbf{w}_0 + t\,\mathrm{e}^{\lambda_i t}\mathbf{w}_1 + \cdots + \tfrac{1}{(k-1)!}\,t^{k-1}\mathrm{e}^{\lambda_i t}\mathbf{w}_{k-1},\\
\mathbf{y}_1(t) &= \mathrm{e}^{\lambda_i t}\mathbf{w}_1 + t\,\mathrm{e}^{\lambda_i t}\mathbf{w}_2 + \cdots + \tfrac{1}{(k-2)!}\,t^{k-2}\mathrm{e}^{\lambda_i t}\mathbf{w}_{k-1},\\
&\vdots\\
\mathbf{y}_{k-1}(t) &= \mathrm{e}^{\lambda_i t}\mathbf{w}_{k-1}, \quad \text{etc.}
\end{aligned}$$

Notes on the theorem cont'd

5. (cont'd)
   Thus a chain of length k gives rise to k fundamental solutions $\mathbf{y}_0(t), \mathbf{y}_1(t), \ldots, \mathbf{y}_{k-1}(t)$ having, in order, k summands, $k-1$ summands, ..., and finally 1 summand. A fundamental system of solutions of $\mathbf{y}' = \mathbf{A}\mathbf{y}$ obtained from a union of such chains is essentially the simplest possible.

Example

We determine a fundamental system of solutions of $\mathbf{y}' = \mathbf{Ay}$ for the matrix

$$\mathbf{A} = \begin{pmatrix} -26 & 49 & 74 \\ -8 & 16 & 25 \\ -4 & 7 & 10 \end{pmatrix}.$$

This matrix was also considered as an example for eigenvalue/eigenvector computations in Math257 of Fall 2023.

$$\begin{aligned}
\chi_{\mathbf{A}}(X) &= \begin{vmatrix} X+26 & -49 & -74 \\ 8 & X-16 & -25 \\ 4 & -7 & X-10 \end{vmatrix} = \begin{vmatrix} X-2 & 0 & -7X-4 \\ 0 & X-2 & -2X-5 \\ 4 & -7 & X-10 \end{vmatrix} \\
&= (X-2)^2(X-10) + 4(X-2)(7X+4) - 7(X-2)(2X+5) \\
&= X^3 - 3X - 2 \\
&= (X-2)(X+1)^2.
\end{aligned}$$

Then, as before we compute the corresponding eigenspaces.

Example (cont'd)

$\Longrightarrow \lambda_1 = 2, \lambda_2 = -1$ (with multiplicity 2).

$\underline{\lambda_1 = 2:}$

$$\mathbf{A} - 2\mathbf{I} = \begin{pmatrix} -28 & 49 & 74 \\ -8 & 14 & 25 \\ -4 & 7 & 8 \end{pmatrix} \to \begin{pmatrix} 4 & -7 & -8 \\ 0 & 0 & 9 \\ 0 & 0 & 18 \end{pmatrix}$$

This matrix has rank 2 and right kernel $\mathbb{R}(7,4,0)^{\mathsf{T}}$.

$\underline{\lambda_2 = -1:}$

$$\mathbf{A}+\mathbf{I} = \begin{pmatrix} -25 & 49 & 74 \\ -8 & 17 & 25 \\ -4 & 7 & 11 \end{pmatrix} \to \begin{pmatrix} 4 & -7 & -11 \\ 0 & 3 & 3 \\ 3 & 0 & -3 \end{pmatrix} \to \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

This matrix has rank 2 and right kernel $\mathbb{R}(-1,1,-1)^{\mathsf{T}}$.

$\Longrightarrow$ The eigenvectors of $\mathbf{A}$ span only a 2-dimensional subspace of $\mathbb{R}^3$, and hence $\mathbf{A}$ is not diagonalizable.

As basis basis of $\mathbb{R}^3$ we can take the two eigenvectors $\mathbf{v}_1 = (7,4,0)^{\mathsf{T}}$, $\mathbf{v}_2 = (-1,1,-1)^{\mathsf{T}}$, and a further vector $\mathbf{v}_3$ solving $(\mathbf{A}+\mathbf{I})\mathbf{v}_3 = \mathbf{v}_2$. Then $(\mathbf{A}+\mathbf{I})^2\mathbf{v}_3 = (\mathbf{A}+\mathbf{I})\mathbf{v}_2 = \mathbf{0}$, so that $\mathbf{v}_3 \in \mathrm{G}_{-1}$.

Example (cont'd)

$$\left(\begin{array}{ccc|c} -25 & 49 & 74 & -1 \\ -8 & 17 & 25 & 1 \\ -4 & 7 & 11 & -1 \end{array}\right) \to \left(\begin{array}{ccc|c} 4 & -7 & -11 & 1 \\ 0 & 3 & 3 & 3 \\ 3 & 0 & -3 & 6 \end{array}\right) \to \left(\begin{array}{ccc|c} 1 & 0 & -1 & 2 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{array}\right)$$

A solution is $\mathbf{v}_3 = (2,1,0)^{\mathsf{T}}$.

A fundamental system of solutions of $\mathbf{y}' = \mathbf{A}\mathbf{y}$ is then

$$\mathbf{y}_1(t) = \mathrm{e}^{2t}\mathbf{v}_1 = \mathrm{e}^{2t}\begin{pmatrix} 7 \\ 4 \\ 0 \end{pmatrix}, \quad \mathbf{y}_2(t) = \mathrm{e}^{-t}\mathbf{v}_2 = \mathrm{e}^{-t}\begin{pmatrix} -1 \\ 1 \\ -1 \end{pmatrix}.$$

(eigenvectors give rise to fundamental solutions in the same way as before), and

$$\begin{aligned} \mathbf{y}_3(t) &= \mathrm{e}^{-t}\mathbf{v}_3 + t\,\mathrm{e}^{-t}(\mathbf{A}+\mathbf{I})\,\mathbf{v}_3 = \mathrm{e}^{-t}\mathbf{v}_3 + t\,\mathrm{e}^{-t}\mathbf{v}_2 \\ &= \mathrm{e}^{-t}\begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} + t\,\mathrm{e}^{-t}\begin{pmatrix} -1 \\ 1 \\ -1 \end{pmatrix}. \end{aligned}$$

The canonical fundamental matrix is (observe that $\mathbf{y}_3(0) = \mathbf{v}_3$ still holds!) $\mathrm{e}^{\mathbf{A}t} = (\mathbf{y}_1(t)|\mathbf{y}_2(t)|\mathbf{y}_3(t))\,\mathbf{S}^{-1} = \cdots$

Example

Determine the general solution of $\mathbf{y}' = \mathbf{A}\mathbf{y}$ for

$$\mathbf{A} = \begin{pmatrix} -2 & 1 & 9 & 0 & 1 & 19 \\ -1 & -2 & -23 & 0 & 0 & -46 \\ 0 & 0 & -8 & 0 & 0 & -12 \\ -6 & 1 & 9 & 4 & 1 & 19 \\ 1 & 0 & -1 & 0 & -2 & -2 \\ 0 & 0 & 6 & 0 & 0 & 10 \end{pmatrix}.$$

$\mathbf{A}$ has the eigenvalue 4, since $\mathbf{A}\mathbf{e}_4 = 4\mathbf{e}_4$. Thus $\chi_{\mathbf{A}}(X)$ is divisible by $X - 4$, which also follows immediately from expanding $\det(X\,\mathbf{I}_6 - \mathbf{A})$ along the 4th column:

$$\chi_{\mathbf{A}}(X) = (4-X)\begin{vmatrix} -2-X & 1 & 9 & 1 & 19 \\ -1 & -2-X & -23 & 0 & -46 \\ 0 & 0 & -8-X & 0 & -12 \\ 1 & 0 & -1 & -2-X & -2 \\ 0 & 0 & 6 & 0 & 10-X \end{vmatrix}$$

Next we add $X + 2$ times first row to the second row in order to obtain a column with only one nonzero entry.

Example (cont'd)

$$\chi_{\mathbf{A}}(X) = (4-X)\begin{vmatrix} -X-2 & 1 & 9 & 1 & 19 \\ -X^2-4X-5 & 0 & 9X-5 & X+2 & 19X-8 \\ 0 & 0 & -X-8 & 0 & -12 \\ 1 & 0 & -1 & -X-2 & -2 \\ 0 & 0 & 6 & 0 & -X+10 \end{vmatrix}$$

$$= (X-4)\begin{vmatrix} -X^2-4X-5 & 9X-5 & X+2 & 19X-8 \\ 0 & -X-8 & 0 & -12 \\ 1 & -1 & -X-2 & -2 \\ 0 & 6 & 0 & -X+10 \end{vmatrix}$$

$$= (X-4)(X+2)\begin{vmatrix} -X^2-4X-5 & 9X-5 & 1 & 19X-8 \\ 0 & -X-8 & 0 & -12 \\ 1 & -1 & -1 & -2 \\ 0 & 6 & 0 & -X+10 \end{vmatrix}$$

$$= (X-4)(X+2)\begin{vmatrix} -X^2-4X-4 & 9X-5 & 1 & 19X-8 \\ 0 & -X-8 & 0 & -12 \\ 0 & -1 & -1 & -2 \\ 0 & 6 & 0 & -X+10 \end{vmatrix}$$

$$= (X-4)(X+2)^3\begin{vmatrix} -X-8 & -12 \\ 6 & -X+10 \end{vmatrix}$$

Example (cont'd)

The final result is

$$\chi_{\mathbf{A}}(X) = (X-4)(X+2)^3(X^2-2X-8) = (X-4)^2(X+2)^4.$$

$\Longrightarrow$ The eigenvalues of $\mathbf{A}$ are $\lambda_1 = 4$ with multiplicity 2 and $\lambda_2 = -2$ with multiplicity 4.

$\underline{\lambda_1 = 4:}$

$$\mathbf{A} - 4\,\mathbf{I} = \begin{pmatrix} -6 & 1 & 9 & 0 & 1 & 19 \\ -1 & -6 & -23 & 0 & 0 & -46 \\ 0 & 0 & -12 & 0 & 0 & -12 \\ -6 & 1 & 9 & 0 & 1 & 19 \\ 1 & 0 & -1 & 0 & -6 & -2 \\ 0 & 0 & 6 & 0 & 0 & 6 \end{pmatrix} \to \begin{pmatrix} 1 & 0 & -1 & 0 & -6 & -2 \\ 0 & 1 & 3 & 0 & -35 & 7 \\ 0 & -6 & -24 & 0 & -6 & -48 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}$$

$$\to \begin{pmatrix} 1 & 0 & -1 & 0 & -6 & -2 \\ 0 & 1 & 3 & 0 & -35 & 7 \\ 0 & 0 & -6 & 0 & -216 & -6 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} \to \begin{pmatrix} 1 & 0 & -1 & 0 & -6 & -2 \\ 0 & 1 & 3 & 0 & -35 & 7 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

$$\Longrightarrow \mathbf{V}_4 = \langle \mathbf{v}_1 = (0,0,0,1,0,0)^{\mathsf{T}},\ \mathbf{v}_2 = (1,-4,-1,0,0,1)^{\mathsf{T}} \rangle$$

Example (cont'd)

$\underline{\lambda_2 = -2:}$

$$\mathbf{A} + 2\mathbf{I} = \begin{pmatrix} 0 & 1 & 9 & 0 & 1 & 19 \\ -1 & 0 & -23 & 0 & 0 & -46 \\ 0 & 0 & -6 & 0 & 0 & -12 \\ -6 & 1 & 9 & 6 & 1 & 19 \\ 1 & 0 & -1 & 0 & 0 & -2 \\ 0 & 0 & 6 & 0 & 0 & 12 \end{pmatrix} \to \begin{pmatrix} 1 & 0 & -1 & 0 & 0 & -2 \\ 0 & 1 & 9 & 0 & 1 & 19 \\ 0 & 1 & 3 & 6 & 1 & 7 \\ 0 & 0 & -24 & 0 & 0 & -48 \\ 0 & 0 & 6 & 0 & 0 & 12 \end{pmatrix}$$

$$\to \begin{pmatrix} 1 & 0 & -1 & 0 & 0 & -2 \\ 0 & 1 & 9 & 0 & 1 & 19 \\ 0 & 0 & -6 & 6 & 0 & -12 \\ 0 & 0 & 1 & 0 & 0 & 2 \end{pmatrix} \to \begin{pmatrix} 1 & 0 & -1 & 0 & 0 & -2 \\ 0 & 1 & 9 & 0 & 1 & 19 \\ 0 & 0 & 1 & 0 & 0 & 2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

$$\Longrightarrow \mathbf{V}_{-2} = \langle \mathbf{v}_3 = (0,-1,0,0,1,0)^T,\ \mathbf{v}_4 = (0,-1,-2,0,0,1)^T \rangle$$

Thus $\mathbf{A}$ has only 4 linearly independent eigenvectors and is not diagonalizable.

The theory developed tells us that $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4$ can be extended to a basis of $\mathbb{C}^6$ by two generalized eigenvectors $\mathbf{v}_5, \mathbf{v}_6$ associated to $\lambda_2 = -2$.

Example (cont'd)

Therefore we compute

$$(\mathbf{A}+2\mathbf{I})^2 = \begin{pmatrix} 0 & 0 & 36 & 0 & 0 & 72 \\ 0 & -1 & -147 & 0 & -1 & -295 \\ 0 & 0 & -36 & 0 & 0 & -72 \\ -36 & 0 & 36 & 36 & 0 & 72 \\ 0 & 1 & 3 & 0 & 1 & 7 \\ 0 & 0 & 36 & 0 & 0 & 72 \end{pmatrix},$$

$$(\mathbf{A}+2\mathbf{I})^3 = \begin{pmatrix} 0 & 0 & 216 & 0 & 0 & 432 \\ 0 & 0 & -864 & 0 & 0 & -1728 \\ 0 & 0 & -216 & 0 & 0 & -432 \\ -216 & 0 & 216 & 216 & 0 & 432 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 216 & 0 & 0 & 432 \end{pmatrix}.$$

We see that $(\mathbf{A}+2\mathbf{I})^3$ has rank 2 and a 4-dimensional right kernel.
$\Longrightarrow W := \operatorname{rker}\left((\mathbf{A}+2\mathbf{I})^3\right)$ is the generalized eigenspace for $\lambda_2 = -2$ and we don't need to compute $(\mathbf{A}+2\mathbf{I})^4$.

Example (cont'd)

A "nice" basis of W is obtained by selecting a vector $\mathbf{w}_1 \in W$ satisfying $(\mathbf{A}+2\mathbf{I})^2\mathbf{w}_1 \neq \mathbf{0}$, e.g., $\mathbf{w}_1 = \mathbf{e}_2 = (0,1,0,0,0,0)^T$.

$$\Longrightarrow \mathbf{w}_1 = \mathbf{e}_2,\ \mathbf{w}_2 = (\mathbf{A}+2\mathbf{I})\mathbf{e}_2 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix},\ \mathbf{w}_3 = (\mathbf{A}+2\mathbf{I})^2\mathbf{e}_2 = \begin{pmatrix} 0 \\ -1 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}$$

can be taken as the first 3 basis vectors.
The 4th vector can be taken as an eigenvector linearly independent from $\mathbf{w}_3$, e.g., $\mathbf{w}_4 = \mathbf{v}_4 = (0,-1,-2,0,0,1)^T$.
For $\mathbf{S} = (\mathbf{v}_1|\mathbf{v}_2|\mathbf{w}_1|\mathbf{w}_2|\mathbf{w}_3|\mathbf{w}_4)$ we then have

$$\mathbf{S}^{-1}\mathbf{A}\mathbf{S} = \left(\begin{array}{cc|ccc|c} 4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & -2 & 0 & 0 & 0 \\ 0 & 0 & 1 & -2 & 0 & 0 \\ 0 & 0 & 0 & 1 & -2 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & -2 \end{array}\right), \ldots$$

Example (cont'd)

... reflecting that $\mathbf{v}_1, \mathbf{v}_2, \mathbf{w}_3, \mathbf{w}_4$ are eigenvectors of $\mathbf{A}$ and $\mathbf{A}\mathbf{w}_1 = -2\mathbf{w}_1 + \mathbf{w}_2$, $\mathbf{A}\mathbf{w}_2 = -2\mathbf{w}_2 + \mathbf{w}_3$.

A fundamental system of solutions of $\mathbf{y}' = \mathbf{A}\mathbf{y}$ is

$$
\begin{aligned}
\mathbf{y}_1(t) &= \mathrm{e}^{4t}\mathbf{v}_1 = \mathrm{e}^{4t}(0,0,0,1,0,0)^{\mathsf{T}},\\
\mathbf{y}_2(t) &= \mathrm{e}^{4t}\mathbf{v}_2 = \mathrm{e}^{4t}(1,-4,-1,0,0,1)^{\mathsf{T}},\\
\mathbf{y}_3(t) &= \mathrm{e}^{-2t}\mathbf{w}_1 + t\,\mathrm{e}^{-2t}\mathbf{w}_2 + \tfrac{1}{2}t^2\,\mathrm{e}^{-2t}\mathbf{w}_3\\
&= \left(t\,\mathrm{e}^{-2t}, \mathrm{e}^{-2t} - \tfrac{1}{2}t^2\mathrm{e}^{-2t}, 0, t\,\mathrm{e}^{-2t}, \tfrac{1}{2}t^2\mathrm{e}^{-2t}, 0\right)^{\mathsf{T}},\\
\mathbf{y}_4(t) &= \mathrm{e}^{-2t}\mathbf{w}_2 + t\,\mathrm{e}^{-2t}\mathbf{w}_3\\
&= \left(\mathrm{e}^{-2t}, -t\,\mathrm{e}^{-2t}, 0, \mathrm{e}^{-2t}, t\,\mathrm{e}^{-2t}, 0\right)^{\mathsf{T}},\\
\mathbf{y}_5(t) &= \mathrm{e}^{-2t}\mathbf{w}_3 = \mathrm{e}^{-2t}(0,-1,0,0,1,0)^{\mathsf{T}},\\
\mathbf{y}_6(t) &= \mathrm{e}^{-2t}\mathbf{w}_4 = \mathrm{e}^{-2t}(0,-1,-2,0,0,1)^{\mathsf{T}}.
\end{aligned}
$$

Note that only the vectors in the basis $\mathbf{v}_1, \mathbf{v}_2, \mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3, \mathbf{w}_4$ or, equivalently, the matrix $\mathbf{S}$ is required to compute the fundamental system.

Example (cont'd)

Finally we determine the canonical fundamental matrix of $\mathbf{y}' = \mathbf{A}\mathbf{y}$ in the usual way (with the help of SageMath):

$$e^{\mathbf{A}t} = \begin{pmatrix} 0 & e^{4t} & te^{-2t} & e^{-2t} & 0 & 0 \\ 0 & -4e^{4t} & e^{-2t} - \frac{1}{2}t^2e^{-2t} & -te^{-2t} & -e^{-2t} & -e^{-2t} \\ 0 & -e^{4t} & 0 & 0 & 0 & -2e^{-2t} \\ e^{4t} & 0 & te^{-2t} & e^{-2t} & 0 & 0 \\ 0 & 0 & \frac{1}{2}t^2e^{-2t} & te^{-2t} & e^{-2t} & 0 \\ 0 & e^{4t} & 0 & 0 & 0 & e^{-2t} \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & -4 & 1 & 0 & -1 & -1 \\ 0 & -1 & 0 & 0 & 0 & -2 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}^{-1}$$

$$= \begin{pmatrix} e^{-2t} & te^{-2t} & 3te^{-2t} + e^{4t} - e^{-2t} & 0 & te^{-2t} & 7te^{-2t} + 2e^{4t} - 2e^{-2t} \\ -te^{-2t} & -\frac{1}{2}t^2e^{-2t} + e^{-2t} & -\frac{3}{2}t^2e^{-2t} + te^{-2t} - 4e^{4t} + 4e^{-2t} & 0 & -\frac{1}{2}t^2e^{-2t} & -\frac{7}{2}t^2e^{-2t} + 2te^{-2t} - 8e^{4t} + 8e^{-2t} \\ 0 & 0 & -e^{4t} + 2e^{-2t} & 0 & 0 & -2e^{4t} + 2e^{-2t} \\ -e^{4t} + e^{-2t} & te^{-2t} & 3te^{-2t} + e^{4t} - e^{-2t} & e^{4t} & te^{-2t} & 7te^{-2t} + 2e^{4t} - 2e^{-2t} \\ te^{-2t} & \frac{1}{2}t^2e^{-2t} & \frac{3}{2}t^2e^{-2t} - te^{-2t} & 0 & \frac{1}{2}t^2e^{-2t} + e^{-2t} & \frac{7}{2}t^2e^{-2t} - 2te^{-2t} \\ 0 & 0 & e^{4t} - e^{-2t} & 0 & 0 & 2e^{4t} - e^{-2t} \end{pmatrix}.$$

Example (cont'd)

For diagonalizable matrices $\mathbf{A}$ there is an alternative method for obtaining a paricular solution of $\mathbf{y}' = \mathbf{A}\mathbf{y} + \mathbf{q}$, which expresses the source $\mathbf{q}(t)$ in terms of eigenvectors of $\mathbf{A}$ and solves the resulting 1-dimensional systems.

In the case under consideration we have

$$\mathbf{q}(t) = \begin{pmatrix} 0 \\ t \end{pmatrix} = \frac{t}{2\mathrm{i}} \begin{pmatrix} 1 \\ \mathrm{i} \end{pmatrix} - \frac{t}{2\mathrm{i}} \begin{pmatrix} 1 \\ -\mathrm{i} \end{pmatrix} = \mathbf{q}_1(t) + \mathbf{q}_2(t),$$

and we can solve $\mathbf{y}' = \mathbf{A}\mathbf{y} + \mathbf{q}_i(t)$ entirely in the corresponding eigenspace.

For $\lambda_1 = \mathrm{i}$ one-dimensional variation of parameters gives a particular solution $\mathbf{z}_1(t) = c(t)\mathbf{y}_1(t)$ with

$$c(t) = \int_0^t \mathrm{e}^{-\mathrm{i}s} \frac{s}{2\mathrm{i}}\,\mathrm{d}s = \frac{1}{2\mathrm{i}} \left[\mathrm{i}s\mathrm{e}^{-\mathrm{i}s} + \mathrm{e}^{-\mathrm{i}s}\right]_0^t = \frac{1}{2\mathrm{i}} \left(\mathrm{i}t\mathrm{e}^{-\mathrm{i}t} + \mathrm{e}^{-\mathrm{i}t} - 1\right),$$

which simplifies (and changes) to $\mathbf{z}_1(t) = \frac{t-\mathrm{i}}{2} \left(\begin{smallmatrix} 1 \\ \mathrm{i} \end{smallmatrix}\right)$. Similarly, for $\lambda_2 = -\mathrm{i}$ we obtain $\mathbf{z}_2(t) = \overline{\mathbf{z}_1(t)} = \frac{t+\mathrm{i}}{2} \left(\begin{smallmatrix} 1 \\ -\mathrm{i} \end{smallmatrix}\right)$. Superposing the individual solutions gives $\mathbf{y}_p(t) = \mathbf{z}_1(t) + \mathbf{z}_2(t) = \left(\begin{smallmatrix} t \\ 1 \end{smallmatrix}\right)$, as before.

How to Compute $e^{\mathbf{A}t}$ in General?

1. **A** *is diagonalizable*.
   If $\mathbf{S}^{-1}\mathbf{A}\mathbf{S} = \mathbf{D}$ then

$$\mathbf{S}^{-1}e^{\mathbf{A}t}\mathbf{S} = \sum_{k=0}^{\infty} \frac{t^k}{k!}\mathbf{S}^{-1}\mathbf{A}^k\mathbf{S}$$

$$= \sum_{k=0}^{\infty} \frac{t^k}{k!}\mathbf{D}^k = \begin{pmatrix} e^{\lambda_1 t} & & \\ & \ddots & \\ & & e^{\lambda_n t} \end{pmatrix} = e^{\mathbf{D}t}.$$

This gives

$$e^{\mathbf{A}t} = \mathbf{S}\begin{pmatrix} e^{\lambda_1 t} & & \\ & \ddots & \\ & & e^{\lambda_n t} \end{pmatrix}\mathbf{S}^{-1},\ e^{\mathbf{A}t}\mathbf{y}(0) = \mathbf{S}\begin{pmatrix} e^{\lambda_1 t} & & \\ & \ddots & \\ & & e^{\lambda_n t} \end{pmatrix}\mathbf{S}^{-1}\mathbf{y}(0).$$

Writing $\mathbf{S} = (\mathbf{v}_1|\ldots|\mathbf{v}_n)$, the right-hand identity says that the general solution of $\mathbf{y}' = \mathbf{A}\mathbf{y}$ is $\mathbf{y}(t) = c_1e^{\lambda_1 t}\mathbf{v}_1 + \cdots + c_ne^{\lambda_n t}\mathbf{v}_n$ with $\mathbf{c}$ determined from $\mathbf{S}\mathbf{c} = c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n = \mathbf{y}(0)$.

1. (cont'd)
   This reaffirms our earlier observation that from a basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$ of $\mathbb{C}^n$ consisting of eigenvectors of $\mathbf{A}$ one obtains a fundamental system of solutions of $\mathbf{y}' = \mathbf{A}\mathbf{y}$ by multiplying each eigenvector $\mathbf{v}_i$ with the corresponding (scalar) exponential function $\mathrm{e}^{\lambda_i t}$.

2. $\mathbf{A}$ has only one eigenvalue λ.
   In this case we have $\chi_{\mathbf{A}}(X) = (X-\lambda)^n$ and $(\mathbf{A}-\lambda\mathbf{I}_n)^k = 0$ for $k \geq n$ by the Cayley-Hamilton Theorem.

$$\Longrightarrow \quad \mathrm{e}^{\mathbf{A}t} = \mathrm{e}^{(\lambda\mathbf{I}+\mathbf{A}-\lambda\mathbf{I})t} = \mathrm{e}^{\lambda\mathbf{I}t}\mathrm{e}^{(\mathbf{A}-\lambda\mathbf{I})t}$$

$$= \mathrm{e}^{\lambda t}\left[\mathbf{I} + t(\mathbf{A}-\lambda\mathbf{I}) + \frac{t^2}{2!}(\mathbf{A}-\lambda\mathbf{I})^2 + \cdots + \frac{t^{n-1}}{(n-1)!}(\mathbf{A}-\lambda\mathbf{I})^{n-1}\right]$$

3. *The general case*
   One possible solution is to compute the Jordan canonical form $\mathbf{J}$ of $\mathbf{A}$ and a matrix $\mathbf{S}$ satisfying $\mathbf{S}^{-1}\mathbf{A}\mathbf{S} = \mathbf{J}$. Then $\mathrm{e}^{\mathbf{A}t} = \mathbf{S}\mathrm{e}^{\mathbf{J}t}\mathbf{S}^{-1}$, and the computation of $\mathrm{e}^{\mathbf{J}t}$ reduces to that of $\mathrm{e}^{\mathbf{J}_i t}$ for the Jordan blocks $\mathbf{J}_i$. The matrices $\mathrm{e}^{\mathbf{J}_i t}$ in turn can be computed by the method in (2).

Example (taken from [Str14])

We solve the two systems

$$\mathbf{y}' = \mathbf{A}\mathbf{y} = \begin{pmatrix} -2 & 1 \\ 1 & -2 \end{pmatrix}\mathbf{y}, \qquad \mathbf{y}' = \mathbf{B}\mathbf{y} = \begin{pmatrix} -2 & 1 \\ -1 & -2 \end{pmatrix}\mathbf{y}$$

and the corresponding IVP's with $\mathbf{y}(0) = (6,2)^{\mathsf{T}}$.

$\mathbf{A}$ has characteristic polynomial
$\chi_{\mathbf{A}}(X) = X^2 + 4X + 3 = (X+1)(X+3)$ and eigenvalues
$\lambda_1 = -1$, $\lambda_2 = -3$.
Corresponding eigenvectors are $\mathbf{v}_1 = (1,1)^{\mathsf{T}}$, $\mathbf{v}_2 = (1,-1)^{\mathsf{T}}$.
$\Longrightarrow$ The general solution of the first system is

$$\mathbf{y}(t) = c_1 \mathrm{e}^{-t}\begin{pmatrix} 1 \\ 1 \end{pmatrix} + c_2 \mathrm{e}^{-3t}\begin{pmatrix} 1 \\ -1 \end{pmatrix}, \quad c_1, c_2 \in \mathbb{C}$$

(and the general real solution is obtained by requiring $c_1, c_2 \in \mathbb{R}$).
The coefficients of the special solution with $\mathbf{y}(0) = (6,2)^{\mathsf{T}}$ are determined by solving

$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 6 \\ 2 \end{pmatrix}, \quad \text{which gives} \quad \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 4 \\ 2 \end{pmatrix}.$$

Example (cont'd)

$$\Longrightarrow \quad \mathbf{y}(t) = 4\mathrm{e}^{-t}\begin{pmatrix}1\\1\end{pmatrix} + 2\mathrm{e}^{-3t}\begin{pmatrix}1\\-1\end{pmatrix} = \begin{pmatrix}4\mathrm{e}^{-t}+2\mathrm{e}^{-3t}\\4\mathrm{e}^{-t}-2\mathrm{e}^{-3t}\end{pmatrix}.$$

B has characteristic polynomial
$\chi_{\mathbf{A}}(X) = X^2 + 4X + 5 = (X + 2 - \mathrm{i})(X + 2 + \mathrm{i})$ and eigenvalues
$\lambda_1 = -2 + \mathrm{i}$, $\lambda_2 = -2 - \mathrm{i}$.
Corresponding eigenvectors are obtained by solving

$$(\mathbf{B} - \lambda_1\mathbf{I})\mathbf{x} = \begin{pmatrix}-\mathrm{i} & 1\\-1 & -\mathrm{i}\end{pmatrix}\begin{pmatrix}x_1\\x_2\end{pmatrix} = \begin{pmatrix}0\\0\end{pmatrix}, \quad \text{e.g.,} \quad \mathbf{v}_1 = \begin{pmatrix}1\\\mathrm{i}\end{pmatrix},$$

and similarly for λ_2, giving $\mathbf{v}_2 = (1, -\mathrm{i})^{\mathsf{T}}$.
(Note that $\mathbf{v}_2 = \overline{\mathbf{v}}_1$, so that no computation is necessary.)
$\Longrightarrow$ The general solution of the second system is

$$\mathbf{y}(t) = c_1\mathrm{e}^{(-2+\mathrm{i})t}\begin{pmatrix}1\\\mathrm{i}\end{pmatrix} + c_2\mathrm{e}^{(-2-\mathrm{i})t}\begin{pmatrix}1\\-\mathrm{i}\end{pmatrix}, \quad c_1, c_2 \in \mathbb{C}.$$

Example (cont'd)

As before, the coefficients of the special solution with $\mathbf{y}(0) = (6, 2)^{\mathsf{T}}$ is determined by solving

$$\begin{pmatrix} 1 & 1 \\ \mathrm{i} & -\mathrm{i} \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 6 \\ 2 \end{pmatrix}, \quad \text{which gives} \quad \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 3-\mathrm{i} \\ 3+\mathrm{i} \end{pmatrix}.$$

$$\begin{aligned} \Longrightarrow \quad \mathbf{y}(t) &= \mathrm{e}^{(-2+\mathrm{i})t} \begin{pmatrix} 3-\mathrm{i} \\ 1+3\mathrm{i} \end{pmatrix} + \mathrm{e}^{(-2-\mathrm{i})t} \begin{pmatrix} 3+\mathrm{i} \\ 1-3\mathrm{i} \end{pmatrix} \\ &= 2\,\mathrm{Re}\left[\mathrm{e}^{(-2+\mathrm{i})t} \begin{pmatrix} 3-\mathrm{i} \\ 1+3\mathrm{i} \end{pmatrix} \right] \\ &= \mathrm{e}^{-2t} \begin{pmatrix} 6\cos t + 2\sin t \\ 2\cos t - 6\sin t \end{pmatrix}. \end{aligned}$$

Note

In these examples it was easier to solve the given ODE system directly without recourse to matrix exponentials. Conversely, we can use the solutions to find the matrix exponentials by means of the formula $\mathrm{e}^{\mathbf{A}t} = \mathbf{\Phi}(t)\mathbf{\Phi}(0)^{-1}$, which switches any known fundamental system of $\mathbf{y}' = \mathbf{A}\mathbf{y}$ into the standard one represented by the matrix exponential; cf. the exercises.

Exercise

a) Show that for $\mathbf{A} \in \mathbb{C}^{n\times n}$ the matrix exponential $\mathrm{e}^{\mathbf{A}t}$ and an arbitrary fundamental matrix $\mathbf{\Phi}(t)$ are related by $\mathbf{\Phi}(t) = \mathrm{e}^{\mathbf{A}t}\mathbf{\Phi}(0)$.

b) For $\mathbf{A} = \begin{pmatrix} -2 & 1 \\ 1 & -2 \end{pmatrix}$ compute $\mathrm{e}^{\mathbf{A}t}$ using a) and the fundamental system determined in the lecture.

c) For the matrix in b), alternatively compute $\mathrm{e}^{\mathbf{A}t}$ using the series representation and the decomposition

$$\begin{pmatrix} -2 & 1 \\ 1 & -2 \end{pmatrix} = -2\,\mathbf{I}_2 + \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Example

As an example of a system with a non-diagonalizable coefficient matrix we consider

$$\mathbf{y}' = \mathbf{C}\mathbf{y} = \begin{pmatrix} -2 & 1 \\ 0 & -2 \end{pmatrix} \mathbf{y}.$$

$\mathbf{C}$ has the eigenvalue $\lambda = -2$ with algebraic multiplicity 2 and geometric multiplicity 1. (The corresponding eigenspace is $\mathbb{C}(1,0)^{\mathsf{T}}$.)

Here we compute $\mathrm{e}^{\mathbf{C}t}$ directly using the method for a single eigenvalue:

$$\mathrm{e}^{\mathbf{C}t} = \mathrm{e}^{-2\mathbf{I}t}\mathrm{e}^{(\mathbf{C}+2\mathbf{I})t} = \mathrm{e}^{-2t}\left[\begin{pmatrix} 1 & \\ & 1 \end{pmatrix} + t\begin{pmatrix} & 1 \\ & \end{pmatrix} + \frac{t^2}{2!}\begin{pmatrix} & 1 \\ & \end{pmatrix}^2 + \cdots\right]$$

$$= \mathrm{e}^{-2t}\begin{pmatrix} 1 & t \\ & 1 \end{pmatrix} = \begin{pmatrix} \mathrm{e}^{-2t} & t\mathrm{e}^{-2t} \\ 0 & \mathrm{e}^{-2t} \end{pmatrix}.$$

It follows that the general solution in this case is

$$\mathbf{y}(t) = \mathrm{e}^{-2t}\begin{pmatrix} c_1 + tc_2 \\ c_2 \end{pmatrix}, \quad c_1, c_2 \in \mathbb{C}.$$

A New Method to Compute $\mathrm{e}^{\mathbf{A}t}$

Theorem

Suppose $\mathbf{A} \in \mathbb{C}^{n\times n}$ *satisfies*

$$a(\mathbf{A}) = a_0\mathbf{I}_n + a_1\mathbf{A} + \cdots + a_d\mathbf{A}^d = \mathbf{0}$$

for some $a(X) = a_0 + a_1X + \cdots + a_dX^d \in \mathbb{C}[X]$.

1. *The entries* $e_{ij}(t)$ *of the matrix exponential* $\mathrm{e}^{\mathbf{A}t} = \big(e_{ij}(t)\big)$ *solve the scalar ODE* $a(\mathrm{D})y = 0$.
2. *The matrix exponential* $\mathrm{e}^{\mathbf{A}t}$ *admits the representation*

$$\mathrm{e}^{\mathbf{A}t} = c_0(t)\mathbf{I}_n + c_1(t)\mathbf{A} + \cdots + c_{d-1}(t)\mathbf{A}^{d-1},$$

where $c_k(t)$ *is the solution of the IVP*

$$a(\mathrm{D})y = 0 \wedge \Big(y(0), y'(0), \ldots, y^{(d-1)}(0)\Big) = \mathbf{e}_{k+1},$$

the standard unit vector in $\mathbb{C}^d$ *of the form* $(\underbrace{0,\ldots,0}_{k},1,0,\ldots,0)$.

In other words, $c_0(t), c_1(t), \ldots, c_{d-1}(t)$ is the special fundamental system of solutions of $a(\mathrm{D})y = 0$ whose Wronski matrix $\mathbf{W}(0)$ at $t = 0$ is the $d \times d$ identity matrix. (This also shows $\mathbf{W}(t) = \mathrm{e}^{\mathbf{C}t}$, where $\mathbf{C}$ is the companion matrix of $a(X)$).

Corollary

Suppose $\mathbf{A}$ *has* r *distinct eigenvalues* $\lambda_1, \ldots, \lambda_r$ *and minimum polynomial* $\mu_{\mathbf{A}}(X) = \prod_{i=1}^{r}(X - \lambda_i)^{m_i} = X^d + \mu_{d-1}X^{d-1} + \cdots + \mu_1 X + \mu_0$. *Then the entries* $\mathrm{e}_{ij}(t)$ *of* $\mathrm{e}^{\mathbf{A}t}$ *and the (uniquely determined) functions* $c_k(t)$ *in the representation* $\mathrm{e}^{\mathbf{A}t} = \sum_{k=0}^{d-1} c_k(t)\mathbf{A}^k$ *have the form*

$$\sum_{i=1}^{r} f_i(t)\mathrm{e}^{\lambda_i t}$$

for some polynomials $f_i(X) \in \mathbb{C}[X]$ *of degree* $\leq m_i - 1$. *The same assertion holds, mutatis mutandis, for* $\chi_{\mathbf{A}}(X)$ *in place of* $\mu_{\mathbf{A}}(X)$ *(except that for* $n > d$ *the functions* $c_k(t)$ *are no longer uniquely determined by the requirement* $\mathrm{e}^{\mathbf{A}t} = \sum_{k=0}^{n-1} c_k(t)\mathbf{A}^k$ *and must be defined as in Part (2) of the theorem).*

Note

If $\mu_{\mathbf{A}}(X)$ properly divides $\chi_{\mathbf{A}}(X)$ then the bound for $\deg f_i(X)$ in terms of $\mu_{\mathbf{A}}(X)$ is stronger.

Proof of the theorem.

(1) Writing $\mathbf{\Phi}(t) = \mathrm{e}^{\mathbf{A}t}$, we infer from $\mathbf{\Phi}'(t) = \mathbf{A}\mathbf{\Phi}(t)$ that

$$a(\mathrm{D})\mathbf{\Phi}(t) = a(\mathbf{A})\mathbf{\Phi}(t) = a(\mathbf{A})\mathrm{e}^{\mathbf{A}t}$$

for all polynomials $a(X) \in \mathbb{C}[X]$.
If $a(\mathbf{A}) = 0$ then $a(\mathrm{D})\mathbf{\Phi}(t) = \mathbf{0}$ and, since differentiation acts entry-wise on $\mathbf{\Phi}(t)$, further $a(\mathrm{D})e_{ij}(t) = 0$ for $1 \le i, j \le n$.

(2) Defining $\mathbf{\Phi}(t)$ as the indicated representation of $\mathrm{e}^{\mathbf{A}t}$, we have

$$\begin{aligned}
\mathbf{\Phi}(t) &= c_0(t)\mathbf{I}_n + c_1(t)\mathbf{A} + \cdots + c_{d-1}(t)\mathbf{A}^{d-1},\\
\mathbf{\Phi}'(t) &= c_0'(t)\mathbf{I}_n + c_1'(t)\mathbf{A} + \cdots + c_{d-1}'(t)\mathbf{A}^{d-1},\\
&\vdots\\
\mathbf{\Phi}^{(d)}(t) &= c_0^{(d)}(t)\mathbf{I}_n + c_1^{(d)}(t)\mathbf{A} + \cdots + c_{d-1}^{(d)}(t)\mathbf{A}^{d-1}.
\end{aligned}$$

If the functions $c_j(t)$ solve the given IVP's then

$$a(\mathrm{D})\mathbf{\Phi}(t) = \mathbf{0} \quad \text{and} \quad \mathbf{\Phi}^{(i)}(0) = \mathbf{A}^i \text{ for } 0 \le i \le d-1.$$

Since $t \mapsto \mathrm{e}^{\mathbf{A}t}$ satisfies these conditions as well, we must have $\mathbf{\Phi}(t) = \mathrm{e}^{\mathbf{A}t}$.

For the last step of the proof note that the matrix IVP $a(\mathrm{D})\mathbf{\Phi}(t) = \mathbf{0} \wedge \mathbf{\Phi}^{(k)}(0) = \mathbf{A}^k$ for $0 \leq k \leq d-1$ amounts to n^2 scalar IVP's for the entries $\mathrm{e}_{ij}(t)$, which are specified in terms of the entries $(\mathbf{A}^k)_{ij}$.

The corollary is an immediate consequence of the theorem in view of $\mu_{\mathbf{A}}(\mathbf{A}) = \mathbf{0}$ and the known structure of the solution space of $\mu_{\mathbf{A}}(\mathrm{D})y = 0$, and similarly for $\chi_{\mathbf{A}}$.

Example

We compute again $\mathrm{e}^{\mathbf{A}t}$ for $\mathbf{A} = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$.

Since $\mathbf{A}^2 = -\mathbf{I}_2$, we can take $a(X) = X^2 + 1$, $d = 2$ in the theorem (in fact $X^2 + 1$ is just the characteristic polynomial of $\mathbf{A}$), which yields the 2nd-order ODE $y'' + y = 0$ for $c_0(t)$ and $c_1(t)$.
Since $\cos t$ and $\sin t$ solve this ODE and satisfy the required initial conditions (i.e, the Wronski matrix of $\cos t, \sin t$ is already $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$), we obtain

$$\mathrm{e}^{\mathbf{A}t} = (\cos t)\mathbf{I}_2 + (\sin t)\mathbf{A} = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix}.$$

Example

Let $\mathbf{P} \in \mathbb{C}^{n\times n}$ be a projection matrix, i.e., $\mathbf{P}^2 = \mathbf{P}$.

Here we can take $a(X) = X^2 - X = X(X-1)$ corresponding to the ODE $y'' - y' = 0$. A fundamental system is $1, \mathrm{e}^t$, and the the required initial conditions are satisfied by $c_0(t) = 1$, $c_1(t) = \mathrm{e}^t - 1$.

$$\Longrightarrow \mathrm{e}^{\mathbf{P}t} = \mathbf{I}_n + (\mathrm{e}^t - 1)\mathbf{P}.$$

This result can also be derived directly from the series representation of $\mathrm{e}^{\mathbf{P}t}$, using the observation that $\mathbf{P}^2 = \mathbf{P}$ imlies $\mathbf{P}^n = \mathbf{P}$ for all $n \geq 1$.

Note

The characteristic polynomial of $\chi_{\mathbf{P}}(X)$ has degree n and leads to a more complicated formula for $\mathrm{e}^{\mathbf{P}t}$ if $n > 2$. For example, projection matrices $\mathbf{P} \in \mathbb{C}^{3\times 3}$ of rank 1 and 2 have characteristic polynomials $X^2(X-1)$ and $X(X-1)^2$, respectively, which lead to representations

$$\mathrm{e}^{\mathbf{P}t} = \mathbf{I}_3 + t\mathbf{P} + (-1 - t + \mathrm{e}^t)\mathbf{P}^2, \quad \text{resp.},$$
$$\mathrm{e}^{\mathbf{P}t} = \mathbf{I}_3 + (-2 + 2\mathrm{e}^t - t\mathrm{e}^t)\mathbf{P} + (1 - \mathrm{e}^t + t\mathrm{e}^t)\mathbf{P}^2.$$

Since $\mathbf{P}^2 = \mathbf{P}$, both representations collapse to $\mathrm{e}^{\mathbf{P}t} = \mathbf{I}_n + (\mathrm{e}^t - 1)\mathbf{P}$.

Example

We compute the matrix eponential of

$$\mathbf{A} = \begin{pmatrix} -26 & 49 & 74 \\ -8 & 16 & 25 \\ -4 & 7 & 10 \end{pmatrix}.$$

This matrix is not diagonalizable, as we have seen earlier, and the example is meant to illustrate the fact that the new method for computing matrix eponentials works just as well for non-diagonalizable matrices.

From the earlier example we use $\chi_{\mathbf{A}}(X) = (X-2)(X+1)^2$ (which happens to coincide with $\mu_{\mathbf{A}}(X)$ in this case, but this fact is not needed for the computation).

A fundamental system of solutions of the corresponding ODE is $y_1(t) = \mathrm{e}^{2t}$, $y_2(t) = \mathrm{e}^{-t}$, $y_3(t) = t\mathrm{e}^{-t}$, which satisfy the initial conditions

$$\mathbf{\Phi}(0) = \begin{pmatrix} y_1(0) & y_2(0) & y_3(0) \\ y_1'(0) & y_2'(0) & y_3'(0) \\ y_1''(0) & y_2''(0) & y_3''(0) \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 2 & -1 & 1 \\ 4 & 1 & -2 \end{pmatrix}.$$

For this observe that $y_3'(t) = (1-t)\mathrm{e}^{-t}$, $y_3''(t) = (t-2)\mathrm{e}^{-t}$.

Example (cont'd)

The required initial conditions $\mathbf{\Psi}(0) = \mathbf{I}_3$ are then satisfied by the transformed system $\mathbf{\Psi}(t) = \mathbf{\Phi}(t)\mathbf{\Phi}(0)^{-1}$, so that we need to invert the matrix $\mathbf{\Phi}(0)$. Applying the standard algorithm gives

$$\mathbf{\Phi}(0)^{-1} = \frac{1}{9}\begin{pmatrix} 1 & 2 & 1 \\ 8 & -2 & -1 \\ 6 & 3 & -3 \end{pmatrix}.$$

This matrix contains the coefficients of $c_0(t)$, $c_1(t)$, $c_2(t)$ with respect to e^{2t}, e^{-t}, $t\mathrm{e}^{-t}$ in the respective column (look at the 1st row of the matrix equation $\mathbf{\Psi}(t) = \mathbf{\Phi}(t)\mathbf{\Phi}(0)^{-1}$, which is $\big(c_0(t), c_1(t), c_2(t)\big) = \big(e^{2t}, \mathrm{e}^{-t}, t\mathrm{e}^{-t}\big)\mathbf{\Phi}(0)^{-1}$), and we finally obtain

$$\begin{aligned}
\mathrm{e}^{\mathbf{A}t} &= \tfrac{1}{9}(\mathrm{e}^{2t} + 8\mathrm{e}^{-t} + 6t\mathrm{e}^{-t})\mathbf{I}_3 + \tfrac{1}{9}(2\mathrm{e}^{2t} - 2\mathrm{e}^{-t} + 3t\mathrm{e}^{-t})\mathbf{A} + \tfrac{1}{9}(\mathrm{e}^{2t} - \mathrm{e}^{-t} - 3t\mathrm{e}^{-t})\mathbf{A}^2 \\
&= \left[\mathrm{e}^{2t}\begin{pmatrix} -7 & 14 & 21 \\ -4 & 8 & 12 \\ 0 & 0 & 0 \end{pmatrix} + \mathrm{e}^{-t}\begin{pmatrix} 8 & -14 & -21 \\ 4 & -7 & -12 \\ 0 & 0 & 1 \end{pmatrix} + t\mathrm{e}^{-t}\begin{pmatrix} -4 & 7 & 11 \\ 4 & -7 & -11 \\ -4 & 7 & 11 \end{pmatrix}\right] \\
&= \begin{pmatrix} -7\mathrm{e}^{2t} + 8\mathrm{e}^{-t} - 4t\mathrm{e}^{-t} & 14\mathrm{e}^{2t} - 14\mathrm{e}^{-t} + 7t\mathrm{e}^{-t} & 21\mathrm{e}^{2t} - 21\mathrm{e}^{-t} + 11t\mathrm{e}^{-t} \\ -4\mathrm{e}^{2t} + 4\mathrm{e}^{-t} + 4t\mathrm{e}^{-t} & 8\mathrm{e}^{2t} - 7\mathrm{e}^{-t} - 7t\mathrm{e}^{-t} & 12\mathrm{e}^{2t} - 12\mathrm{e}^{-t} - 11t\mathrm{e}^{-t} \\ -4t\mathrm{e}^{-t} & 7t\mathrm{e}^{-t} & \mathrm{e}^{-t} + 11t\mathrm{e}^{-t} \end{pmatrix}.
\end{aligned}$$

Example (cont'd)

One should compare the costs of this computation to the one using the JCF. In the earlier example we had computed the JCF

$$\mathbf{J} = \left(\begin{array}{c|cc} 2 & 0 & 0 \\ \hline 0 & -1 & 1 \\ 0 & 0 & -1 \end{array}\right)$$

of $\mathbf{A}$ and $\mathbf{S}$ satisfying $\mathbf{S}^{-1}\mathbf{A}\mathbf{S} = \mathbf{J}$. From this we can continue as follows:

$$\begin{aligned} \mathrm{e}^{\mathbf{A}t} &= \mathbf{S}\mathrm{e}^{\mathbf{J}t}\mathbf{S}^{-1} \\ &= \begin{pmatrix} 7 & -1 & 2 \\ 4 & 1 & 1 \\ 0 & -1 & 0 \end{pmatrix} \left(\begin{array}{c|cc} \mathrm{e}^{2t} & 0 & 0 \\ \hline 0 & \mathrm{e}^{-t} & t\mathrm{e}^{-t} \\ 0 & 0 & \mathrm{e}^{-t} \end{array}\right) \begin{pmatrix} 7 & -1 & 2 \\ 4 & 1 & 1 \\ 0 & -1 & 0 \end{pmatrix}^{-1} \\ &= \begin{pmatrix} 7 & -1 & 2 \\ 4 & 1 & 1 \\ 0 & -1 & 0 \end{pmatrix} \left(\begin{array}{c|cc} \mathrm{e}^{2t} & 0 & 0 \\ \hline 0 & \mathrm{e}^{-t} & t\mathrm{e}^{-t} \\ 0 & 0 & \mathrm{e}^{-t} \end{array}\right) \begin{pmatrix} -1 & 2 & 3 \\ 0 & 0 & -1 \\ 4 & -7 & -11 \end{pmatrix} \\ &= \cdots \end{aligned}$$

The total costs are certainly no less than those of the new method.

What Goes Wrong for $\mathbf{y}' = \mathbf{A}(t)\mathbf{y}$?

The exponential matrix

$$e^{\mathbf{B}(t)} = \sum_{k=0}^{\infty} \frac{1}{k!}\mathbf{B}(t)^k$$

is well-defined but does not satisfy $\frac{d}{dt}e^{\mathbf{B}(t)} = \mathbf{B}'(t)e^{\mathbf{B}(t)}$ in general.
$\Longrightarrow \mathbf{y}(t) = e^{\int \mathbf{A}(t)\,dt}$ does not necessarily solve $\mathbf{y}' = \mathbf{A}(t)\mathbf{y}$.

Reason: When differentiating $e^{\mathbf{B}(t)}$ termwise, we need the relation $\frac{d}{dt}\mathbf{B}(t)^k = k\mathbf{B}(t)^{k-1}\mathbf{B}'(t) = k\mathbf{B}'(t)\mathbf{B}(t)^{k-1}$, but we have only

$$\frac{d}{dt}\mathbf{B}(t)^2 = \mathbf{B}(t)\mathbf{B}'(t) + \mathbf{B}'(t)\mathbf{B}(t),$$
$$\frac{d}{dt}\mathbf{B}(t)^3 = \mathbf{B}'(t)\mathbf{B}(t)^2 + \mathbf{B}(t)\mathbf{B}'(t)\mathbf{B}(t) + \mathbf{B}(t)^2\mathbf{B}'(t), \quad \text{etc.}$$

A special case

If $\mathbf{A}(t)$ and $\mathbf{B}(t) = \mathbf{B}_0 + \int_{t_0}^{t} \mathbf{A}(s)\,ds$ commute then $\mathbf{y}(t) = e^{\mathbf{B}(t)}$ solves $\mathbf{y}' = \mathbf{A}(t)\mathbf{y}$.

Exercise

Suppose $\mathbf{A} \in \mathbb{C}^{n\times n}$ has n distinct eigenvalues $\lambda_1, \dots, \lambda_n$. Show that

$$\mathrm{e}^{\mathbf{A}t} = \sum_{i=1}^{n} \mathrm{e}^{\lambda_i t} \ell_i(A),$$

where $\ell_i(X) = \prod_{j=1, j\neq i}^{n} \frac{X-\lambda_j}{\lambda_i-\lambda_j}$ are the Lagrange polynomials corresponding to $\lambda_1, \dots, \lambda_n$.

Hint: Show that $\mathbf{\Phi}(t) = \sum_{i=1}^{n} \mathrm{e}^{\lambda_i t} \ell_i(A)$ solves the IVP $\mathbf{\Phi}'(t) = \mathbf{A}\mathbf{\Phi}(t) \wedge \mathbf{\Phi}(0) = \mathbf{I}_n$.

Exercise

Consider the two time-dependent linear systems

$$\mathbf{y}' = \mathbf{A}_1(t)\mathbf{y} = \begin{pmatrix} 1 & t \\ t & 1 \end{pmatrix} \mathbf{y} \quad \text{and} \quad \mathbf{y}' = \mathbf{A}_2(t)\mathbf{y} = \begin{pmatrix} 1 & t \\ 0 & 0 \end{pmatrix} \mathbf{y}.$$

Compute the matrix exponentials $\mathbf{E}_i(t) = \exp\left(\int_0^t \mathbf{A}_i(s)\,\mathrm{d}s\right)$, $i = 1, 2$, and show that $\mathbf{E}_1(t)$ forms a fundamental matrix of the corresponding system but $\mathbf{E}_2(t)$ does not.

Math 285
Introduction to Differential Equations

Thomas Honold

ZJU-UIUC Institute

Spring Semester 2024

Outline

Today's Lecture: Introduction to PDE's

Motivation

FOURIER's Problem

Describe the heat flow in a long and thin rectangular plate, when some known temperature function is applied to one of the short sides and the long sides are kept at constant temperature.

For simplicity, the plate is assumed to be infinitely long and thin, and given by the region

$$P = \{(x, y) \in \mathbb{R}^2; -1 \leq x \leq 1, y \geq 0\}.$$

For a stationary solution, the temperature $z(x, y)$ at $(x, y) \in P^\circ$ (i.e., for $-1 < x < 1$, $y > 0$) has to satisfy

$$\frac{\partial^2 z}{\partial x^2} + \frac{\partial^2 z}{\partial y^2} = 0 \qquad \text{(Laplace's Equation)}$$

The boundary conditions will be

$$z(x, 0) = f(x) \quad \text{for } -1 \leq x \leq 1,$$
$$z(-1, y) = z(1, y) = 0 \quad \text{for } y > 0,$$

where $f\colon [-1, 1] \to \mathbb{R}$ gives the temperature on the short side (the other short side is considered as infinitely far away).

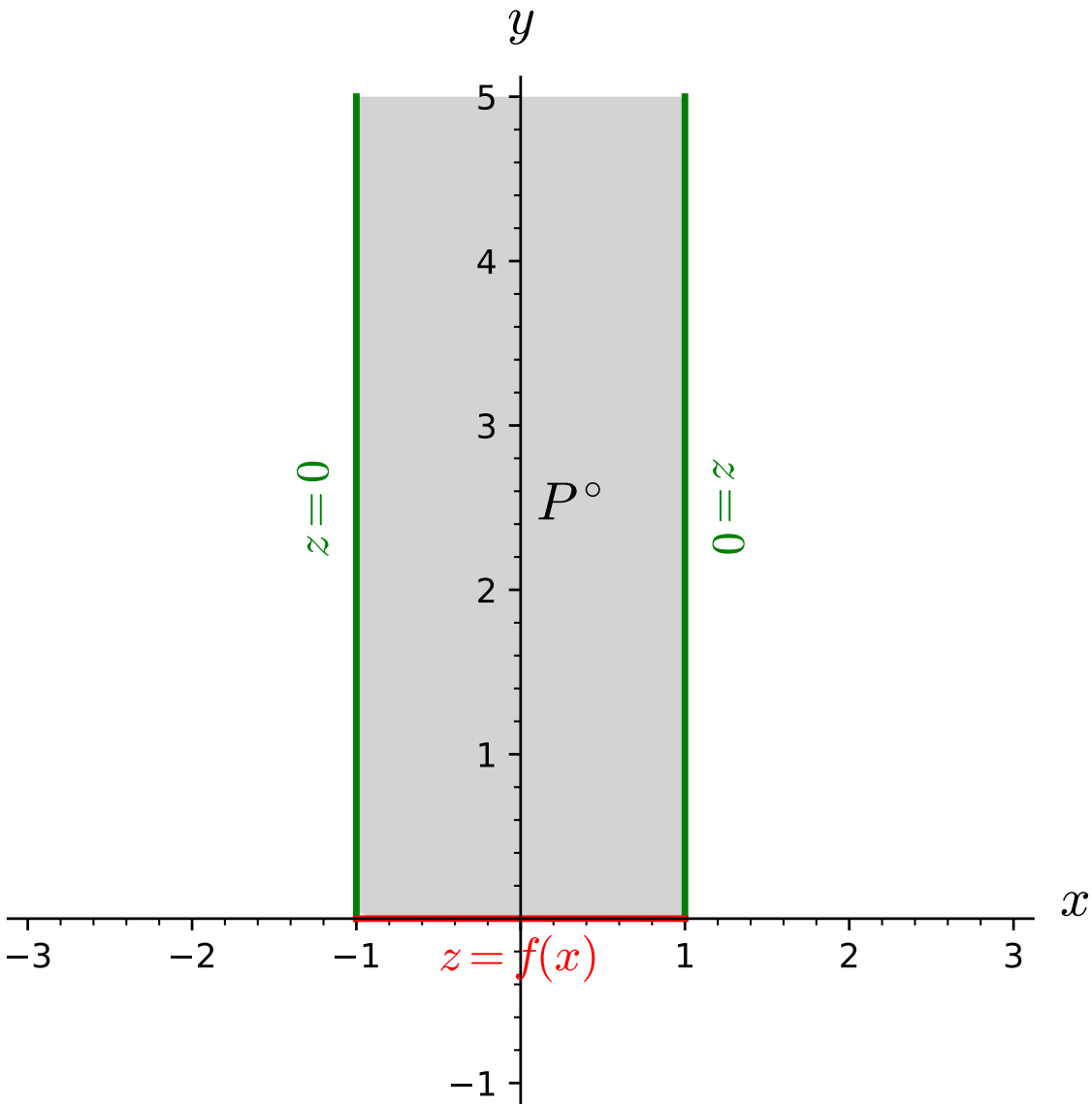


Figure: Fourier's Problem

FOURIER assumed further that $f(x) = f(-x)$.
He was interested especially in the case of a constant temperature > 0 ("heating"), which we can take w.l.o.g. as $f(x) \equiv 1$.

Fourier's Solution

We start with the "separation ansatz" $z(x, y) = a(x)b(y)$.

Plugging this into Laplace's Equation gives

$$\frac{\partial^2 z}{\partial x^2} + \frac{\partial^2 z}{\partial y^2} = a''(x)b(y) + a(x)b''(y) = 0.$$

At points (x, y) with $z(x, y) \neq 0$ we can rewrite this as

$$\frac{a''(x)}{a(x)} = -\frac{b''(y)}{b(y)}.$$

$\Longrightarrow$ Both sides must be constant, i.e., there exists $C \in \mathbb{R}$ such that

$$a''(x) = -C\,a(x), \quad b''(y) = C\,b(y) \quad \text{for all } (x, y) \in P^\circ.$$

Assuming $z(x, y)$ is not identically zero, the first boundary condition implies $a(-1) = a(1) = 0$.

Fourier's Solution cont'd

$\Longrightarrow$ $a(x)$ and $a''(x)$ must have opposite signs
$\Longrightarrow$ $C > 0$, and we can set $C = K^2$ with $K > 0$.

The general real solution of the two resulting ODE's is

$$a(x) = c_1 \cos(Kx) + c_2 \sin(Kx), \qquad c_1, c_2 \in \mathbb{R},$$
$$b(y) = c_3 e^{Ky} + c_4 e^{-Ky}, \qquad c_3, c_4 \in \mathbb{R}.$$

Since $f(x) = z(x,0) = a(x)b(0)$ should be an even function, we must have $c_2 = 0$.
Since the temperature should drop to zero for $y \to +\infty$ (from physics or just common sense) we must have $c_3 = 0$.
Since $a(1) = a(-1) = 0$, K must be an odd multiple of $\pi/2$.

$$\Longrightarrow \quad z(x,y) = a\, e^{-\frac{(2k-1)\pi y}{2}} \cos\left(\frac{(2k-1)\pi x}{2}\right), \quad a \in \mathbb{R},\ k = 1, 2, \ldots.$$

Since superposition preserves solutions, any function of the form

$$z(x,y) = a_1 e^{-\pi y/2} \cos(\pi x/2) + a_2 e^{-3\pi y/2} \cos(3\pi x/2) + \cdots + a_n e^{-(2n-1)\pi y/2} \cos\big((2n-1)\pi x/2\big)$$

Fourier's Solution cont'd

with $a_1, \ldots, a_n \in \mathbb{R}$ will then also be a solution of Laplace's Equation and satisfy the boundary conditions $z(-1, y) = z(1, y) = 0$, as well as

$$f(x) = z(x, 0) = \sum_{k=1}^{n} a_k \cos\left(\frac{(2k-1)\pi x}{2}\right).$$

The function $f(x) \equiv 1$, however, is not of this form, since any (finite) linear combination of the functions $\cos\left(\frac{(2k-1)\pi x}{2}\right)$ vanishes at $x = \pm 1$.

Question: What to do?

FOURIER assumed the existence of an infinite series representation

$$f(x) = 1 = \sum_{k=1}^{\infty} a_k \cos\left(\frac{(2k-1)\pi x}{2}\right) \quad \text{for } -1 < x < 1,$$

and showed how to compute a_k from this and the additional assumption that this series can be integrated termwise.

Fourier's Solution cont'd

Lemma

For $k, l \in \mathbb{Z}^+$ we have

$$\int_{-1}^{1} \cos\left(\frac{(2k-1)\pi x}{2}\right) \cos\left(\frac{(2l-1)\pi x}{2}\right) \mathrm{d}x = \begin{cases} 0 & \text{if } k \neq l, \\ 1 & \text{if } k = l. \end{cases}$$

Proof.

Making the substitution $t = \pi x/2$, $\mathrm{d}t = (\pi/2)\,\mathrm{d}x$, the integral becomes

$$\begin{aligned} &\frac{2}{\pi} \int_{-\pi/2}^{\pi/2} \cos\big((2k-1)t\big) \cos\big((2l-1)t\big)\,\mathrm{d}t \\ &\qquad = \frac{2}{\pi} \int_{0}^{\pi} \cos\big((2k-1)t\big) \cos\big((2l-1)t\big)\,\mathrm{d}t \\ &\qquad = \frac{1}{\pi} \int_{-\pi}^{\pi} \cos\big((2k-1)t\big) \cos\big((2l-1)t\big)\,\mathrm{d}t\,. \end{aligned}$$

The latter integral is well-known to have the value 0 for $k \neq l$ and π for $k = l$; cf. exercises. □

Fourier's Solution cont'd

If $f(x) = \sum_{k=1}^{\infty} a_k \cos\left(\frac{(2k-1)\pi x}{2}\right)$ can be integrated termwise, the lemma implies

$$\begin{aligned}
&\int_{-1}^{1} f(x) \cos\left(\tfrac{(2l-1)\pi x}{2}\right) \mathrm{d}x \\
&\quad = \int_{-1}^{1} \sum_{k=1}^{\infty} a_k \cos\left(\tfrac{(2k-1)\pi x}{2}\right) \cos\left(\tfrac{(2l-1)\pi x}{2}\right) \mathrm{d}x \\
&\quad = \sum_{k=1}^{\infty} a_k \int_{-1}^{1} \cos\left(\tfrac{(2k-1)\pi x}{2}\right) \cos\left(\tfrac{(2l-1)\pi x}{2}\right) \mathrm{d}x = a_l.
\end{aligned}$$

In particular for $f(x) \equiv 1$ we obtain (the values $f(\pm 1)$ do not matter here)

$$a_l = \int_{-1}^{1} \cos\left(\tfrac{(2l-1)\pi x}{2}\right) \mathrm{d}x = \left[\frac{2}{(2l-1)\pi} \sin\left(\tfrac{(2l-1)\pi x}{2}\right)\right]_{-1}^{1} = \frac{4(-1)^{l-1}}{(2l-1)\pi}.$$

Fourier's Solution cont'd

Thus FOURIER arrived at the series representation

$$1 = \frac{4}{\pi}\left(\cos\frac{\pi x}{2} - \frac{1}{3}\cos\frac{3\pi x}{2} + \frac{1}{5}\cos\frac{5\pi x}{2} - \frac{1}{7}\cos\frac{7\pi x}{2} \pm \cdots\right)$$

and concluded from this that

$$\begin{aligned} z(x,y) &= \frac{4}{\pi}\sum_{k=1}^{\infty}\frac{(-1)^{k-1}}{2k-1}\mathrm{e}^{-(2k-1)\pi y/2}\cos\frac{(2k-1)\pi x}{2} \\ &= \frac{4}{\pi}\left(\mathrm{e}^{-\pi y/2}\cos\frac{\pi x}{2} - \frac{1}{3}\mathrm{e}^{-3\pi y/2}\cos\frac{3\pi x}{2}\right. \\ &\qquad \left. + \frac{1}{5}\mathrm{e}^{-5\pi y/2}\cos\frac{5\pi x}{2} - \frac{1}{7}\mathrm{e}^{-7\pi y/2}\cos\frac{7\pi x}{2} + \cdots\right) \end{aligned}$$

solves the Laplace equation in P° (assuming that $z(x,y)$ can be differentiated termwise with respect to x, y) and satisfies the boundary conditions $z(\pm 1, y) = 0$ for $y \geq 0$, $z(x,0) = 1$ for $-1 < x < 1$.

It remains yet to prove:

1. the identity

$$1 = \frac{4}{\pi} \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} \cos \frac{(2k+1)\pi x}{2}, \quad -1 < x < 1;$$

2. the function

$$z(x,y) = \frac{4}{\pi} \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} \mathrm{e}^{-(2k+1)\pi y/2} \cos \frac{(2k+1)\pi x}{2}, \ (x,y) \in P$$

is well-defined for all $(x,y) \in \mathbb{R}^2$ with $y > 0$ and satisfies $\Delta z(x,y) = 0$ for those (x,y).

Property (2) is quite easy and will be proved right now.

Property (1) is more difficult and a proof was only found by DIRICHLET some 20 years after FOURIER had submitted his manuscript *Theory of the Propagation of Heat in Solid Bodies*. It is a consequence of a general theorem on the point-wise convergence of Fourier series, which we will derive later.

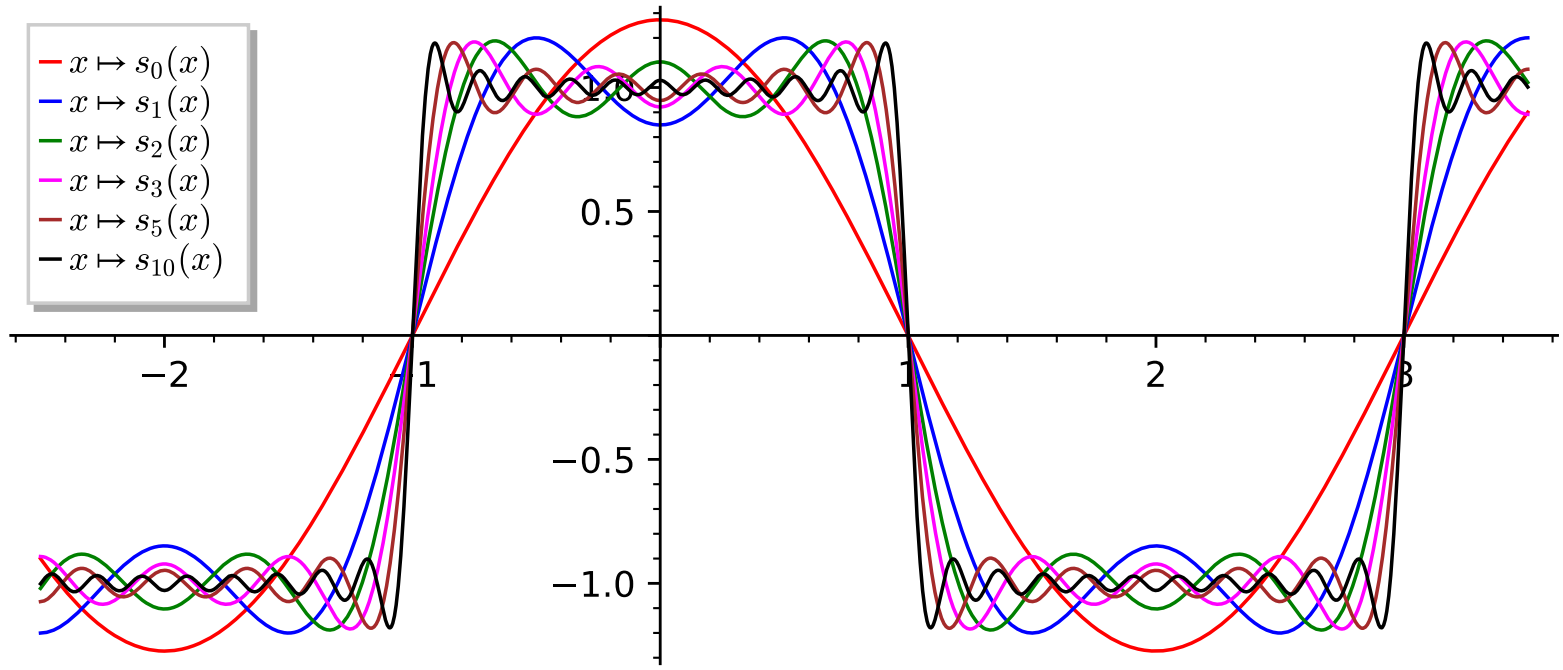


Figure: Partial sums of FOURIER's cosine series

Note that all functions $f_k(x) = \cos \frac{(2k+1)\pi x}{2}$, $x \in \mathbb{R}$ satisfy $f_k(x+2) = -f_k(x)$ and hence $f_k(x+4) = f_k(x)$. Hence the same is true of the limit function f. In particular, the series does not represent the function $f(x) \equiv 1$ outside $[-1, 1]$ (only for about half of the points $x \in \mathbb{R}$).

Proof of Property (2).

It suffices to show that the series defining $z(x,y)$ and the corresponding series of partial derivatives up to order 2 converge uniformly on every subset $H_\delta = \{(x,y) \in \mathbb{R}^2; y \geq \delta\}$, $\delta > 0$, of the domain $H = \{(x,y) \in \mathbb{R}^2; y > 0\}$ (the open "upper half-plane"). This shows that $z(x,y)$ is well-defined, and the Differentiation Theorem gives that $\partial^2 z/\partial x^2$, $\partial^2 z/\partial y^2$ (and the Laplacian as well) can be computed term-wise. Since the terms $z_k(x,y) = \pm\frac{4}{(2k+1)\pi} e^{-(2k+1)\pi y/2} \cos\frac{(2k+1)\pi x}{2}$ satisfy $\Delta z_k(x,y) = 0$ by construction, the same is then true of $z(x,y)$.

We give the proof only for the series representing $\partial^2 z/\partial x^2$. (The remaining proofs are virtually the same.)

$$\sum_{k=0}^{\infty} \frac{\partial^2 z_k(x,y)}{\partial x^2} = \sum_{k=0}^{\infty} \frac{4(-1)^k}{(2k+1)\pi} e^{-(2k+1)\pi y/2} \cos\frac{(2k+1)\pi x}{2} \left(-\frac{(2k+1)^2\pi^2}{4}\right)$$

$$= \sum_{k=0}^{\infty} (-1)^{k+1}(2k+1)\pi\, e^{-(2k+1)\pi y/2} \cos\frac{(2k+1)\pi x}{2}$$

Proof cont'd.

$\Longrightarrow$ For $(x, y) \in H_\delta$ we have

$$\sum_{k=0}^{\infty} \left| \frac{\partial^2 z_k(x,y)}{\partial x^2} \right| \leq \pi \sum_{k=0}^{\infty} (2k+1) e^{-(2k+1)\pi\delta/2} < \infty,$$

since the exponentials decrease faster than $(2k+1)^{-3}$, say. Hence $\sum_{k=0}^{\infty} \frac{\partial^2 z_k(x,y)}{\partial x^2}$ is majorized on H_δ by a convergent series, which is independent of x and y. This implies uniform convergence on H_δ, as asserted (by Weierstrass's Criterion). □

Note

In the case under consideration it would have been sufficient to show that the majorizing series doesn't depend on x, because $\partial^2 z/\partial x^2$ is computed by considering y as a fixed parameter. To make the same argument work for both $\partial^2 z/\partial x^2$ and $\partial^2 z/\partial y^2$, one needs indpendence of x and y.

The Vibrating String Problem

A homogeneous string of length $L > 0$ is stretched along the line segment $0 \leq x \leq L$, $y = 0$ of the (x, y)-plane and fixed at both ends. At time $t = 0$ it is displaced from this "equilibrium position" to an "initial position" $y = f(x)$, $0 \leq x \leq L$, which satisfies $f(0) = f(L) = 0$, and an "initial velocity" $y = g(x)$ in the y-direction is applied to it. The function g should also satisfy $g(0) = g(L) = 0$. From then the string is left at the disposal of the elastic forces acting on it and '"vibrates" around the equilibrium position.

Problem

Determine the "elongation function" $y(x, t)$, $0 \leq x \leq L$, $t \geq 0$ describing the movement of the string over time.

$y(x, t)$ must be a solution of the 1-dimensional wave equation

$$\frac{\partial^2 y}{\partial t^2} = c^2 \frac{\partial^2 y}{\partial x^2},$$

with $c > 0$ a physically determined constant (tension-to-density ratio of the string).

$y(x,t)$ must satisfy the boundary conditions

$$y(0,t) = y(L,t) = 0, \quad t \geq 0,$$
$$y(x,0) = f(x), \qquad y_t(x,0) = g(x).$$

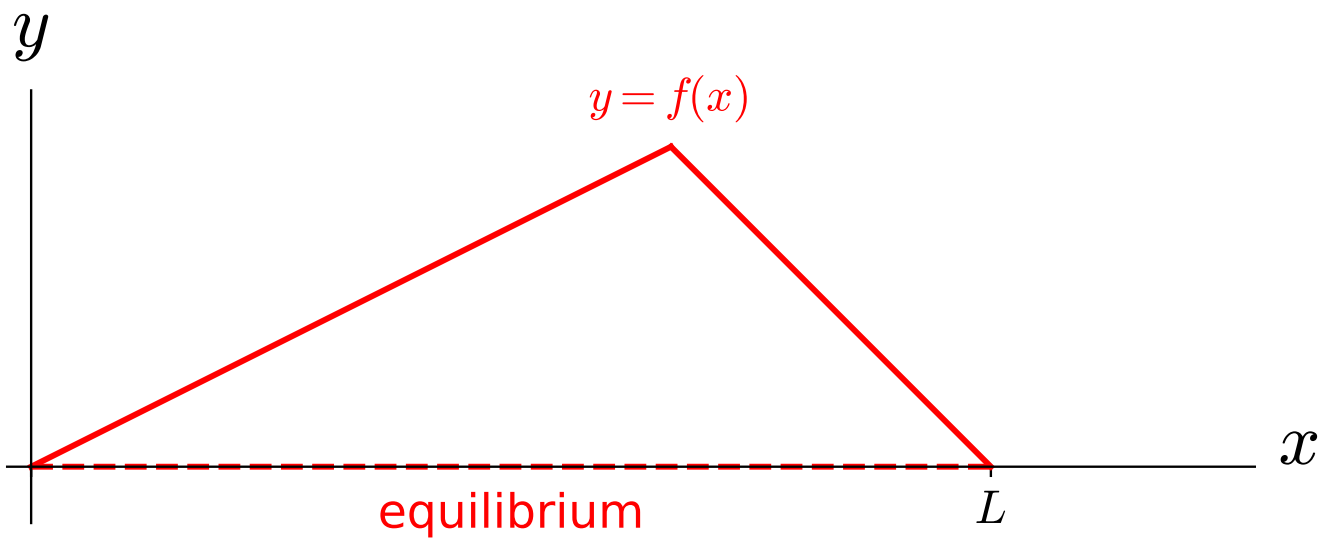


Figure: Vibrating string problem with $g(x) \equiv 0$

D. BERNOULLI's solution

First we determine the solutions of the special form $y(x,t) = a(x)b(t)$.
Proceeding as in FOURIER's solution, we obtain the pair of 2nd-order ODE's

$$\frac{a''(x)}{a(x)} = \frac{1}{c^2}\frac{b''(t)}{b(t)} = -K^2,$$

where $K \geq 0$ is a real constant.

$$\begin{aligned} \Longrightarrow \quad a(x) &= c_1\cos(Kx) + c_2\sin(Kx), \\ b(t) &= c_3\cos(cKt) + c_4\sin(cKt) \quad \text{with } c_1, c_2, c_3, c_4 \in \mathbb{R}. \end{aligned}$$

The boundary condition $y(0,t) = y(L,t) = 0$ translates into $a(0) = a(L) = 0$ and implies

$$c_1 = 0 \quad \text{and} \quad K \in \{\pi/L, 2\pi/L, 3\pi/L, \dots\}.$$

D. BERNOULLI's solution cont'd

Strictly speaking, we only obtain $K \in \mathbb{Z}(\pi/L)$, but the all-zero solution ($K = 0$) can be omitted, and $\pm K$ give scalar multiples of the same solution.

$\Longrightarrow$ All functions of the form

$$y(x,t) = \sum_{k=1}^{n} \sin\frac{k\pi x}{L}\left(a_k \cos\frac{ck\pi t}{L} + b_k \sin\frac{ck\pi t}{L}\right), \quad a_k, b_k \in \mathbb{R},$$

are solutions of $y_{tt} = c^2\, y_{xx}$ satisfying the first boundary condition $y(0,t) = y(L,t) = 0$ and

$$y(x,0) = \sum_{k=1}^{n} a_k \sin\frac{k\pi x}{L},$$

$$y_t(x,0) = \frac{c\pi}{L}\sum_{k=1}^{n} k b_k \sin\frac{k\pi x}{L}.$$

D. BERNOULLI then claimed (without providing the necessary justification for convergence and term-wise differentiability of the series) that the solution to the vibrating string problem is

D. BERNOULLI's solution cont'd

$$y(x,t) = \sum_{k=1}^{\infty} \sin\frac{k\pi x}{L}\left(a_k \cos\frac{ck\pi t}{L} + b_k \sin\frac{ck\pi t}{L}\right),$$

where a_k, b_k are determined from $f(x) = \sum_{k=1}^{\infty} a_k \sin\frac{k\pi x}{L}$ and $g(x) = \frac{c\pi}{L}\sum_{k=1}^{n} k b_k \sin\frac{k\pi x}{L}$, respectively.

Question

Does every continuous function $f\colon [0,L] \to \mathbb{R}$ with $f(0) = f(L) = 0$ have a representation as a sine series of the above form?

Note that the problem for $g(x)$ is the same.

Requiring f and g to be continuous comes from the physical interpretation. Note, however, that at least f need not be differentiable, since we want to model situations like plucking a string, which corresponds to a piece-wise linear function $f(x)$.

Negative answer

There exist continuous functions which are not represented by their Fourier series at every point. Piece-wise C^1-functions, however, and hence virtually all physically meaningful functions, are represented everywhere by their Fourier series.

D'ALEMBERT's solution

This solution is completely different from BERNOULLI's and starts by making the variable substitution

$$\xi = x + ct,\ \eta = x - ct, \quad \text{i.e.,} \quad z(\xi,\eta) = y\left(\tfrac{1}{2}(\xi+\eta), \tfrac{1}{2c}(\xi-\eta)\right).$$

$$\begin{aligned}
\Longrightarrow \quad z_\xi &= \frac{1}{2}y_x + \frac{1}{2c}y_t, \\
z_{\xi\eta} = (z_\xi)_\eta &= \frac{1}{2}z_{\xi x} - \frac{1}{2c}z_{\xi t} \\
&= \frac{1}{2}\left(\frac{1}{2}y_{xx} + \frac{1}{2c}y_{tx}\right) - \frac{1}{2c}\left(\frac{1}{2}y_{xt} + \frac{1}{2c}y_{tt}\right) \\
&= \frac{1}{4}\left(y_{xx} - \frac{1}{c^2}y_{tt}\right),
\end{aligned}$$

provided that y, and hence z, are C^2-functions.
Hence $y(x,t)$ solves the 1-dimensional wave equation iff $z_{\xi\eta}(\xi,\eta) \equiv 0$.

D'ALEMBERT's solution cont'd

The solution of $z_{\xi\eta} = 0$ is

$$\begin{aligned} z_\xi(\xi,\eta) &= \phi(\xi), \\ z(\xi,\eta) &= \Phi(\xi) + \Psi(\eta) && \text{(with } \Phi'(\xi) = \phi(\xi)\text{)} \\ &= \Phi(x+ct) + \Psi(x-ct). \end{aligned}$$

$$\Longrightarrow y(x,t) = \Phi(x+ct) + \Psi(x-ct)$$

with arbitrary C^2-functions $\Phi\colon [0,+\infty) \to \mathbb{R}$ and $\Psi\colon (-\infty, L] \to \mathbb{R}$.

The boundary conditions for $y(x,t)$ translate into

$$\begin{aligned} \Phi(t) + \Psi(-t) = \Phi(L+t) + \Psi(L-t) = 0 \quad &\text{for } t \geq 0, \\ \Phi(x) + \Psi(x) = f(x),\ \Phi'(x) - \Psi'(x) = g(x)/c \quad &\text{for } 0 \leq t \leq L. \end{aligned}$$

The first set of equations imply

$$\begin{aligned} \Phi(t+2L) = \Phi(L+t+L) = -\Psi(L-t-L) = -\Psi(-t) = \Phi(t), \\ \Psi(-t-2L) = -\Phi(t+2L) = -\Phi(t) = \Psi(-t) \qquad \text{for } t \geq 0. \end{aligned}$$

$\Longrightarrow$ We can extend Φ, Ψ to $2L$-periodic functions with domain $\mathbb{R}$.

D'ALEMBERT's solution cont'd

With this definition the first set of equations then hold for all $t \in \mathbb{R}$. On account of periodicity, it suffices to verify this for $t \in [-L, 0)$:

$$\Phi(t) = \Phi(t + 2L) = \Phi(L + t + L) = -\Psi(L - t - L) = -\Psi(-t),$$
$$\Phi(L + t) = -\Psi(-L - t) = -\Psi(L - t),$$

where $t + L = L + t \geq 0$ was used.

Further, since Φ and Ψ are determined only up to an additive constant and $\Phi(0) + \Psi(0) = 0$, we can normalize to $\Phi(0) = \Psi(0) = 0$.

Then the second set of equations gives

$$\Phi(x) - \Psi(x) = \frac{1}{c} \int_0^x g(\xi) \mathrm{d}\xi,$$
$$\Phi(x) = \frac{1}{2} \left(f(x) + \frac{1}{c} \int_0^x g(\xi) \mathrm{d}\xi \right),$$
$$\Psi(x) = \frac{1}{2} \left(f(x) - \frac{1}{c} \int_0^x g(\xi) \mathrm{d}\xi \right) \qquad \text{for } 0 \leq x \leq L.$$

These identities can be made to hold for all $x \in \mathbb{R}$, provided ...

D'ALEMBERT's solution cont'd

... we extend f, g first to odd functions on $[-L, L]$ (possible, since $f(0) = g(0) = 0$) and then $2L$-periodically to the whole of $\mathbb{R}$.
Reason: With this definition we have for $x \in [-L, 0)$

$$\begin{aligned}\Phi(x) = -\Psi(-x) &= -\frac{1}{2}\left(f(-x) - \frac{1}{c}\int_0^{-x} g(\xi)\mathrm{d}\xi\right) \\ &= \frac{f(x)}{2} + \frac{1}{c}\int_0^{x} g(-\eta)(-\mathrm{d}\eta) = \frac{f(x)}{2} + \frac{1}{c}\int_0^{x} g(\eta)\mathrm{d}\eta,\end{aligned}$$

as asserted. For general x the second identity then follows from the $2L$-periodicity of both sides, using $\int_x^{x+2L} g(\xi)\mathrm{d}\xi = \int_{-L}^{L} g(\xi)\mathrm{d}\xi = 0$. The third identity (and hence the first) is proved similarly.

$$\begin{aligned}\Longrightarrow y(x,t) &= \Phi(x+ct) + \Psi(x-ct) \\ &= \frac{1}{2}\left(f(x+ct) + f(x-ct) + \frac{1}{c}\int_{x-ct}^{x+ct} g(\xi)\mathrm{d}\xi\right).\end{aligned}$$

for $0 \le x \le L$, $t \ge 0$. This is D'ALEMBERT's solution.

Notes

- BERNOULLI's solution also has the form $y(x,t) = \Phi(x+ct) + \Psi(x-ct)$, as can be seen by rewriting it using the formulas

$$\sin\phi_1 \cos\phi_2 = \frac{1}{2}\big(\sin(\phi_1+\phi_2) - \sin(\phi_1-\phi_2)\big),$$
$$\sin\phi_1 \sin\phi_2 = -\frac{1}{2}\big(\cos(\phi_1+\phi_2) - \cos(\phi_1-\phi_2)\big)$$

in the following way:

$$\begin{aligned} y(x,t) &= \sum_{k=1}^{\infty} \sin\frac{k\pi x}{L}\left(a_k \cos\frac{ck\pi t}{L} + b_k \sin\frac{ck\pi t}{L}\right) \\ &= \frac{1}{2}\sum_{k=1}^{\infty}\left(a_k \sin\frac{k\pi(x+ct)}{L} - b_k\cos\frac{k\pi(x+ct)}{L}\right) \\ &\quad + \frac{1}{2}\sum_{k=1}^{\infty}\left(a_k \sin\frac{k\pi(x-ct)}{L} + b_k\cos\frac{k\pi(x-ct)}{L}\right). \end{aligned}$$

Notes cont'd

- Although in the derivation of D'ALEMBERT's formula we have implicitly assumed that f is a C^2-function and g is a C^1-function, the formula remains true for functions f, g whose derivatives have jump discontinuities, such as piece-wise linear functions.
- D'ALEMBERT's formula can be physically interpreted as the superposition of two waves with initial states $\Phi(x)$ and $\Psi(x)$ moving at constant speed in opposite directions.
- BERNOULLI's Fourier series solution, although conceptually more complicated than D'ALEMBERT's solution, has the additional benefit of revealing the "harmonic analysis" of the vibrating string.

The Linear Algebra of Fourier Series

Among others, the following two variants of the definition of a Fourier series are most commonly found in the literature.

Definition

1. The *Fourier series* of an absolutely integrable 2π-periodic function $f\colon \mathbb{R} \to \mathbb{R}$ is the series

$$\frac{a_0}{2} + \sum_{k=1}^{\infty} \left(a_k \cos(kt) + b_k \sin(kt)\right),$$

where a_k $(k = 0, 1, 2, \dots)$ and b_k $(k = 1, 2, 3, \dots)$, the so-called *Fourier coefficients* of the series, are defined by

$$a_k = \frac{1}{\pi} \int_0^{2\pi} f(t) \cos(kt)\,\mathrm{d}t, \quad b_k = \frac{1}{\pi} \int_0^{2\pi} f(t) \sin(kt)\,\mathrm{d}t\,.$$

Definition (cont'd)

2. The *Fourier series* of an absolutely integrable 2π-periodic function $f : \mathbb{R} \to \mathbb{C}$ is the series

$$c_0 + \sum_{k=1}^{\infty} \left(c_k e^{ikt} + c_{-k} e^{-ikt} \right),$$

where c_k $(k = 0, \pm 1 \pm 2, \dots)$, the complex *Fourier coefficients* of the series, are defined by

$$c_k = \frac{1}{2\pi} \int_0^{2\pi} f(t) e^{-ikt} \, dt .$$

Notes

- $f\colon \mathbb{R} \to \mathbb{C}$ has *period* L if $f(t+L) = f(t)$ for all $t \in \mathbb{R}$. The substitution $s = 2\pi t/L$, transforming L-periodic functions into 2π-periodic functions, can be used to define Fourier series (and develop the corresponding theory) for complex-valued functions of any period $L > 0$. The corresponding L-periodic cosine, sine and exponential functions are $\cos(2\pi kt/L)$, $\sin(2\pi kt/L)$ and $\mathrm{e}^{2\pi \mathrm{i} kt/L}$, respectively, and the formulas for the Fourier coefficients are

$$a_k = \frac{2}{L}\int_0^L f(t)\cos(2\pi kt/L)\,\mathrm{d}t,$$

$$b_k = \frac{2}{L}\int_0^L f(t)\sin(2\pi kt/L)\,\mathrm{d}t,$$

$$c_k = \frac{1}{L}\int_0^L f(t)\mathrm{e}^{-2\pi \mathrm{i} kt/L}\,\mathrm{d}t\,.$$

- The integral of an L-periodic function over any interval of length L is the same; cf. exercises. Thus, for example, we can obtain the Fourier coefficients a_k, b_k, c_k in the previous definition also by integrating over $[-\pi, \pi]$ instead of $[0, 2\pi]$.

Notes cont'd

- Integrating over $[-\pi, \pi]$ instead of $[0, 2\pi]$ has the advantage that it yields immediately the following result:

$$a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\cos(kt) = 0 \qquad \text{if } f(-t) = -f(t),$$

$$b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\sin(kt) = 0 \qquad \text{if } f(-t) = f(t);$$

  i.e., the Fourier series of an odd periodic function is a pure sine series, and the Fourier series of an even periodic function is a pure cosine series.

Notes cont'd

- Variant (1) of the definition also makes sense for complex valued functions f, and the Fourier series obtained by both definitions are in fact the same!

  In order to see this, recall that a series $\sum_{n=0}^{\infty} f_n$ is defined as the sequence (g_n) of partial sums $g_n = \sum_{k=0}^{n} f_k$. Hence it suffices to verify $c_0 = a_0/2$ and $c_k e^{ikt} + c_{-k} e^{-ikt} = a_k \cos(kt) + b_k \sin(kt)$. We have

$$\begin{aligned} c_k &= \frac{1}{2\pi} \int_0^{2\pi} f(t) e^{-ikt}\, dt = \frac{1}{2\pi} \int_0^{2\pi} f(t)\big(\cos(kt) - i\sin(kt)\big)\, dt \\ &= \frac{1}{2}(a_k - ib_k), \\ c_{-k} &= \frac{1}{2\pi} \int_0^{2\pi} f(t) e^{ikt}\, dt = \frac{1}{2}(a_k + ib_k), \end{aligned}$$

  and hence

$$\begin{aligned} c_k e^{ikt} + c_{-k} e^{-ikt} &= \frac{a_k}{2}(e^{ikt} + e^{-ikt}) - \frac{ib_k}{2}(e^{ikt} - e^{-ikt}) \\ &= a_k \cos(kt) + b_k \sin(kt), \qquad \text{as desired.} \end{aligned}$$

Example

We compute the Fourier series of the putative limit function of FOURIER's cosine series, which has period 4 and is defined on $[0, 4)$ by

$$f(x) = \begin{cases} 1 & \text{if } -1 < x < 1, \\ -1 & \text{if } 1 \le x < 3, \end{cases}$$

and verify that both series coincide (which is a nontrivial fact!). Here $L = 4$, and the Fourier series of f has the form

$$\frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos(k\pi x/2) + b_k \sin(k\pi x/2).$$

Since f is even, we have $b_k = 0$ for all k. For $k \in \mathbb{Z}^+$ we have

$$\begin{aligned} a_k &= \frac{2}{4} \int_{-1}^{3} f(x) \cos(k\pi x/2)\,\mathrm{d}x \\ &= \frac{1}{2} \int_{-1}^{1} \cos(k\pi x/2)\,\mathrm{d}x + \frac{1}{2} \int_{1}^{3} -\cos(k\pi x/2)\,\mathrm{d}x \\ &= \frac{1}{2} \left[\frac{2}{k\pi} \sin(k\pi x/2) \right]_{-1}^{1} - \frac{1}{2} \left[\frac{2}{k\pi} \sin(k\pi x/2) \right]_{1}^{3} \end{aligned}$$

Example (cont'd)

It follows that

$$a_k = \begin{cases} 0 & \text{if } k \equiv 0 \pmod 2, \\ \frac{4}{k\pi} & \text{if } k \equiv 1 \pmod 4, \\ -\frac{4}{k\pi} & \text{if } k \equiv 3 \pmod 4. \end{cases}$$

This also holds for $k = 0$, since $a_0 = \frac{1}{2}\int_{-1}^{3} f(x)\,\mathrm{d}x = 0$.
$\Longrightarrow$ The Fourier series of f is

$$\frac{4}{\pi}\cos\frac{\pi x}{2} - \frac{4}{3\pi}\cos\frac{3\pi x}{2} + \frac{4}{5\pi}\cos\frac{5\pi x}{2} - \frac{4}{7\pi}\cos\frac{7\pi x}{2} \pm \ldots,$$

the same as FOURIER's cosine series.

Exercise

Suppose that $f\colon \mathbb{R} \to \mathbb{C}$ is L-periodic and integrable over $[0, L]$. Show that f is integrable over any interval $[a, a+L]$, $a \in \mathbb{R}$, and

$$\int_a^{a+L} f(x)\,\mathrm{d}x = \int_0^L f(x)\,\mathrm{d}x\,.$$

Exercise

Show that the Fourier coefficients a_k, b_k, c_k of any function f are related by

$$c_0 = \frac{a_0}{2} \quad \text{and} \quad |c_k|^2 + |c_{-k}|^2 = \frac{1}{2}\left(|a_k|^2 + |b_k|^2\right) \quad \text{for } k \geq 1.$$

Exercise

What can you say about the Fourier coefficients a_k, b_k, c_k of an L-periodic function $f\colon \mathbb{R} \to \mathbb{C}$ that satisfies

a) $f(x + L/2) = f(x)$ for all $x \in \mathbb{R}$?

b) $f(x + L/2) = -f(x)$ for all $x \in \mathbb{R}$?

Exercise

Suppose $f : \mathbb{R} \to \mathbb{C}$ is L-periodic and satisfies one of the symmetry properties $f(x_0 - x) = f(x_0 + x)$ or $f(x_0 - x) = -f(x_0 + x)$ for some $x_0 \in \mathbb{R}$. What can you say about x_0 and the Fourier coefficients of f ?

Inner Product Spaces

The dot product $\mathbf{x} \cdot \mathbf{y} = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n$ on $\mathbb{R}^n$ can be generalized in two important ways:

1. Let V be a vector space over $\mathbb{R}$. A map $\sigma\colon V \times V \to \mathbb{R}$ is called *inner product* on V if it satisfies the following axioms:

   (IP1) $\sigma(x, y + y') = \sigma(x, y) + \sigma(x, y')$ and $\sigma(x, cy) = c\,\sigma(x, y)$ for all $x, y, y' \in V$ and $c \in \mathbb{R}$, i.e., σ is linear in the second argument;

   (IP2) $\sigma(y, x) = \sigma(x, y)$ for all $x, y \in V$;

   (IP3) $\sigma(x, x) \geq 0$ for all $x \in V$ with equality iff $x = 0_V$.

2. Let V be a vector space over $\mathbb{C}$. A map $\sigma\colon V \times V \to \mathbb{C}$ is called *inner product* on V if it satisfies Axioms (IP1), (IP3) above and the following replacement for (IP2):

   (IP2') $\sigma(y, x) = \overline{\sigma(x, y)}$ for all $x, y \in V$.

Property (IP2') implies $\sigma(x, x) \in \mathbb{R}$ for $x \in V$, so that (IP3) is meaningful also for complex inner product spaces.

Examples

1. The standard example of a real inner product space is $V = \mathbb{R}^n$ with $\sigma(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot \mathbf{y}$. Another important example is $V = \mathrm{C}\big([a, b]\big)$ (i.e., continuous real-valued functions on $[a, b]$) with $\sigma(f, g) = \langle f, g \rangle = \int_a^b f(x)g(x)\,\mathrm{d}x$.
2. The standard example of a complex inner product space is $V = \mathbb{C}^n$ with

$$\sigma(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot \mathbf{y} := \overline{x}_1 y_1 + \overline{x}_2 y_2 + \cdots + \overline{x}_n y_n.$$

Another important example, especially from the Fourier series perspective, is $V = \mathrm{C}\big([a, b]\big)$ (now this denotes the vector space over $\mathbb{C}$ of all continuous complex-valued functions on $[a, b]$) with

$$\sigma(f, g) = \langle f, g \rangle := \int_a^b \overline{f(x)} g(x)\,\mathrm{d}x\,.$$

Inner Product Spaces–Basic Theory

The basic theory is the same for all real and complex inner product spaces, and mimics that of the standard example $\mathbb{R}^n$ discussed earlier.

1. $|\sigma(x,y)|^2 \leq \sigma(x,x)\sigma(y,y)$ for all $x, y \in V$ with equality iff x and y are linearly dependent (*Cauchy-Schwarz Inequality*).
2. $d(x,y) = \sigma(x-y, x-y)$ defines a metric ("distance") on V. This metric is translation-invariant and arises from the "length" function $|x| = \|x\| := \sqrt{\sigma(x,x)}$ in the same way as the Euclidean metric on $\mathbb{R}^n$ from $|\mathbf{x}| = \sqrt{x_1^2 + \cdots + x_n^2}$.
3. x, y are said to be *orthogonal (perpendicular)* if $\sigma(x,y) = 0$. By (IP2) resp. (IP2'), this relation is symmetric. If $\sigma(x,y) = 0$ then $\sigma(x+y, x+y) = \sigma(x,x) + \sigma(y,y)$ (*Pythagoras' Theorem*).
4. The *orthogonal projection* of $b \in V$ onto the *line* spanned by $a \in V \setminus \{0\}$ ($\mathbb{R}a$ resp. $\mathbb{C}a$) is defined as $\mathrm{proj}_a(b) = \lambda^* a$ where λ^* is the (unique) solution of $\sigma(a, b - \lambda a) = 0$. Since σ is linear in the 2nd argument, we obtain $\sigma(a,b) - \lambda^*\sigma(a,a) = 0$, and hence $\mathrm{proj}_a(b) = \frac{\sigma(a,b)}{\sigma(a,a)}\, a$.

4. (cont'd) In the case of a complex inner product space we have $\operatorname{proj}_a(b) = \frac{\overline{\sigma(b,a)}}{\sigma(a,a)}\, a$, so that the order of a, b in the numerator usually matters!

   $\operatorname{proj}_a(b)$ uniquely minimizes the distance from b to a point on $\mathbb{R}a$ resp. $\mathbb{C}a$.

5. Orthogonal projection generalizes to finite-dimensional subspaces U of V. If $u_1, \ldots, u_r$ is a basis of U, the orthogonal projection of $b \in V$ onto U is defined as $\operatorname{proj}_U(b) = \sum_{j=1}^{r} \lambda_j^* u_j$ where $(\lambda_1^*, \ldots, \lambda_r^*)$ is the (unique) solution of the system

$$\sigma\left(u_i, b - \sum_{j=1}^{r} \lambda_j u_j\right) = 0, \quad 1 \le i \le r.$$

   Again, $\operatorname{proj}_U(b)$ uniquely minimizes the distance from b to a point in U.

6. If $\dim V = n < \infty$, there exists an *orthonormal basis* of V, i.e., a basis $v_1, \ldots, v_n$ satisfying

$$\sigma(v_i, v_j) = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

Approximation by Trigonometric Polynomials

We consider the vector space V (over $\mathbb{C}$) formed by all 2π-periodic functions $f\colon \mathbb{R} \to \mathbb{C}$, which are (Lebesgue-)integrable over $[0, 2\pi]$ (and hence over all intervals of length 2π), and satisfy $\int_0^{2\pi} |f(x)|^2 \,\mathrm{d}x < \infty$ (so-called *square-integrable* periodic functions); e.g., $x \mapsto 1/\sqrt{x}$ is integrable but not square integrable over $[0, 2\pi]$, while $x \mapsto x^{-1/3}$ and $x \mapsto \ln x$ are both integrable and square integrable over $[0, 2\pi]$.

V comes with the "inner product"

$$\langle f, g\rangle = \int_0^{2\pi} \overline{f(x)} g(x) \,\mathrm{d}x\,.$$

Strictly speaking, V is not an inner product space, since there are 2π-periodic functions $f \neq 0$ satisfying $\langle f, f\rangle = 0$, for example the characteristic function of $2\pi\mathbb{Z}$. However, we can identify $f, g \in V$ if $f(x) = g(x)$ almost everywhere and consider the vector space $\overline{V}$ formed by the resulting equivalence classes $[f]$. Setting $\langle [f], [g]\rangle = \langle f, g\rangle$, the space $\overline{V}$ becomes a "real" inner product space, since $\langle f, f\rangle = 0$ implies $f = 0$ almost everywhere and hence $[f] = [0]$. The space $\overline{V}$ is also denoted by $\mathrm{L}^2\big([0, 2\pi]\big)$.

Square-integrability of f, g is needed for showing that $\int_0^{2\pi} \overline{f(x)} g(x)\,\mathrm{d}x$ exists.

We have seen that V (strictly speaking, $\overline{V}$) forms a metric space relative to the mean-square distance (also called L^2-distance) defined by

$$\mathrm{d}_2(f,g) = \|f-g\|_2 = \sqrt{\langle f-g, f-g\rangle} = \sqrt{\int_0^{2\pi} |f(x)-g(x)|^2\,\mathrm{d}x}.$$

Definition

A function $g \in V$ of the form

$$g(x) = \sum_{k=0}^{n} \big(a_k \cos(kx) + b_k \sin(kx)\big) = \sum_{k=-n}^{n} c_k \mathrm{e}^{\mathrm{i}kx}$$

with $a_k, b_k, c_k \in \mathbb{C}$ is called a *trigonometric polynomial* of *degree* at most n (exactly n, if one of a_n, b_n or one of c_n, c_{-n} is nonzero).

Note that the 2π-periodic trigonometric polynomials are precisely the functions in the span TP of $1, \cos x, \sin x, \cos(2x), \sin(2x), \ldots$ or, alternatively, in the span of $\{\mathrm{e}^{\mathrm{i}kx}; k \in \mathbb{Z}\}$. Likewise, the 2π-periodic trigonometric polynomials of degree $\leq n$ form a subspace TP_n of V (and of TP).

For each function $f \in V$, the Fourier coefficients a_k, b_k, c_k, and hence the Fourier series of f, n are well-defined. The partial sums of the Fourier series,

$$\mathrm{S}_n f = \frac{a_0}{2} + \sum_{k=1}^{n} a_k \cos(kx) + b_k \sin(kx) = \sum_{k=-n}^{n} c_k \mathrm{e}^{\mathrm{i}kx} \in \mathrm{TP}_n,$$

are called *Fourier polynomials* and have the following "best-approximation" property:

Theorem

Suppose $f \in V$ and $n \in \mathbb{N}$.

1. *The Fourier polynomial $\mathrm{S}_n f$ is the unique trigonometric polynomial in TP_n minimizing the mean-square distance to f :*

$$\|f - \mathrm{S}_n f\|_2 < \|f - g\|_2 \quad \text{for all } g \in \mathrm{TP}_n \setminus \{\mathrm{S}_n f\}.$$

2. *We have*

$$\|f - \mathrm{S}_n f\|_2^2 = \|f\|_2^2 - \|\mathrm{S}_n f\|_2^2 = \|f\|_2^2 - 2\pi \sum_{k=-n}^{n} |c_k|^2 .$$

Proof.

(1) From the general theory of inner product spaces we know that the distance between f and the functions in the subspace $U = \mathrm{TP}_n$, which has dimension $r = 2n+1$ (cf. Worksheet 2, Exercise H11), is minimized by the orthogonal projection of f onto TP_n, which relative to any basis $g_1, \ldots, g_r$ of TP_n is obtained by solving

$$\left\langle g_i, f - \sum_{j=1}^{r} \lambda_j g_j \right\rangle = 0 \quad \text{for } 1 \le i \le r.$$

The solution is uniquely determined, since the Gram matrix $\big(\langle g_i, g_j\rangle\big)_{1\le i,j\le r}$ is invertible. (This is due to the fact that TP_n consists of continuous functions and hence can be viewed as a subspace of $\overline{V}$. For the special basis $\{\mathrm{e}^{\mathrm{i}kx}; -n \le k \le n\}$ it is directly proved below. For arbitrary subspaces $U \subseteq V$ it fails.)

If the basis functions are orthogonal and have length > 0, we further get

$$\left\langle g_i, f - \sum_{j=1}^{r} \lambda_j g_j \right\rangle = \langle g_i, f\rangle - \sum_{j=1}^{r} \lambda_j \langle g_i, g_j\rangle = \langle g_i, f\rangle - \lambda_i \langle g_i, g_i\rangle,$$

i.e., $\lambda_i = \langle g_i, f\rangle \big/ \langle g_i, g_i\rangle$.

Proof cont'd.

The proof is finished by showing that the Fourier polynomial $S_n f$ is equal to the orthogonal projection of f onto $W = TP_n$.
For the proof we use the basis of TP_n consisting of the exponentials e^{ikx}, $-n \le k \le n$, which turns out to be the most convenient:

$$\langle e^{ikx}, e^{ilx} \rangle = \int_0^{2\pi} e^{i(l-k)x}\,dx = \begin{cases} 0 & \text{for } l \neq k, \\ 2\pi & \text{for } l = k. \end{cases}$$

$\Longrightarrow$ The exponentials are mutually orthogonal, and the coefficient of e^{ikx} in the orthogonal projection of f onto TP_n is equal to

$$\frac{\langle e^{ikx}, f \rangle}{\langle e^{ikx}, e^{ikx} \rangle} = \frac{1}{2\pi} \int_0^{2\pi} e^{-ikx} f(x)\,dx = c_k.$$

Hence the orthogonal projection is $\sum_{k=-n}^{n} c_k e^{ikx} = S_n f$, as asserted.

(2) Since $f - S_n f \perp S_n f$, Pythagoras' Theorem gives
$\|f\|_2^2 = \|f - S_n f\|_2^2 + \|S_n f\|_2^2$ and
$\|S_n f\|_2^2 = \sum_{k,l=-n}^{n} \overline{c}_k c_l \langle e^{ikx}, e^{ilx} \rangle = 2\pi \sum_{k=-n}^{n} |c_k|^2$. □

Corollary (BESSEL's Inequality)

For any $f \in V$ we have

$$\sum_{k\in\mathbb{Z}} |c_k|^2 = \sum_{k=-\infty}^{\infty} |c_k|^2 \le \frac{1}{2\pi}\int_0^{2\pi} |f(x)|^2\,\mathrm{d}x\,.$$

Notes

- Bessel's Inequality implies in particular that $\lim_{k\to+\infty} c_k = \lim_{k\to+\infty} c_{-k} = 0$ (or, equivalently, $\lim_{k\to\infty} a_k = \lim_{k\to\infty} b_k = 0$).
- In fact equality holds in Bessel's Inequality, as we will see subsequently. However, Bessel's Inequality generalizes to any sequence $g_1, g_2, g_3, \ldots$ of mutually orthogonal functions $g_n \in V$ with $\|g_n\|_2^2 = \langle g_n, g_n\rangle > 0$ in the form

$$\sum_{n=1}^{\infty} \frac{\langle g_n, f\rangle\langle f, g_n\rangle}{\langle g_n, g_n\rangle^2} \le \langle f, f\rangle = \int_0^{2\pi} |f(x)|^2\,\mathrm{d}x,$$

and for such sequences equality need no longer hold.

Notes cont'd

- It goes without saying that the theorem and it's corollary (Bessel's Inequality) hold in the more general setting of L-periodic functions. The inner product space view provides in fact the best mnemonic for the various Fourier coefficient formulas. We illustrate this for the orthogonal system of L-periodic functions formed by $c_k(x) = \cos(2k\pi x/L)$, $k \geq 0$, and $s_k(x) = \sin(2k\pi x/L)$, $k \geq 1$.

  The Fourier series of an L-periodic function $f\colon \mathbb{R} \to \mathbb{C}$ is

$$\frac{\langle c_0, f\rangle}{\langle c_0, c_0\rangle} c_0(x) + \sum_{k=1}^{\infty} \left(\frac{\langle c_k, f\rangle}{\langle c_k, c_k\rangle} c_k(x) + \frac{\langle s_k, f\rangle}{\langle s_k, s_k\rangle} s_k(x) \right).$$

$$\Longrightarrow \quad \frac{a_0}{2} = \frac{\langle c_0, f\rangle}{\langle c_0, c_0\rangle} = \frac{\langle 1, f\rangle}{\langle 1, 1\rangle} = \frac{\int_0^L f(x)\,\mathrm{d}x}{\int_0^L 1\,\mathrm{d}x} = \frac{1}{L}\int_0^L f(x)\,\mathrm{d}x,$$

$$a_k = \frac{\langle c_k, f\rangle}{\langle c_k, c_k\rangle} = \frac{\int_0^L f(x)\cos(2k\pi x/L)\,\mathrm{d}x}{\int_0^L \cos^2(2k\pi x/L)\,\mathrm{d}x} = \frac{2}{L}\int_0^L f(x)\cos(2k\pi x/L)\,\mathrm{d}x,$$

$$b_k = \frac{\langle s_k, f\rangle}{\langle s_k, s_k\rangle} = \frac{\int_0^L f(x)\sin(2k\pi x/L)\,\mathrm{d}x}{\int_0^L \sin^2(2k\pi x/L)\,\mathrm{d}x} = \frac{2}{L}\int_0^L f(x)\sin(2k\pi x/L)\,\mathrm{d}x.$$

L^2-Convergence of Fourier Series

The following functions are used in the convergence proofs of Fourier series (for both point-wise convergence and L^2-convergence).

$$\mathrm{D}_n(x) = \sum_{k=-n}^{n} \mathrm{e}^{\mathrm{i}kx} = 1 + 2\sum_{k=1}^{n} \cos(kx)$$

$$= \frac{\sin\left((n+\frac{1}{2})x\right)}{\sin\frac{1}{2}x}, \qquad \text{(DIRICHLET kernel)}$$

$$\mathrm{F}_n(x) = \frac{1}{n}\left(\mathrm{D}_0(x) + \mathrm{D}_1(x) + \cdots + \mathrm{D}_{n-1}(x)\right)$$

$$= \frac{1}{n}\left(\frac{\sin\left(\frac{1}{2}nx\right)}{\sin\frac{1}{2}x}\right)^2. \qquad \text{(FEJÉR kernel)}$$

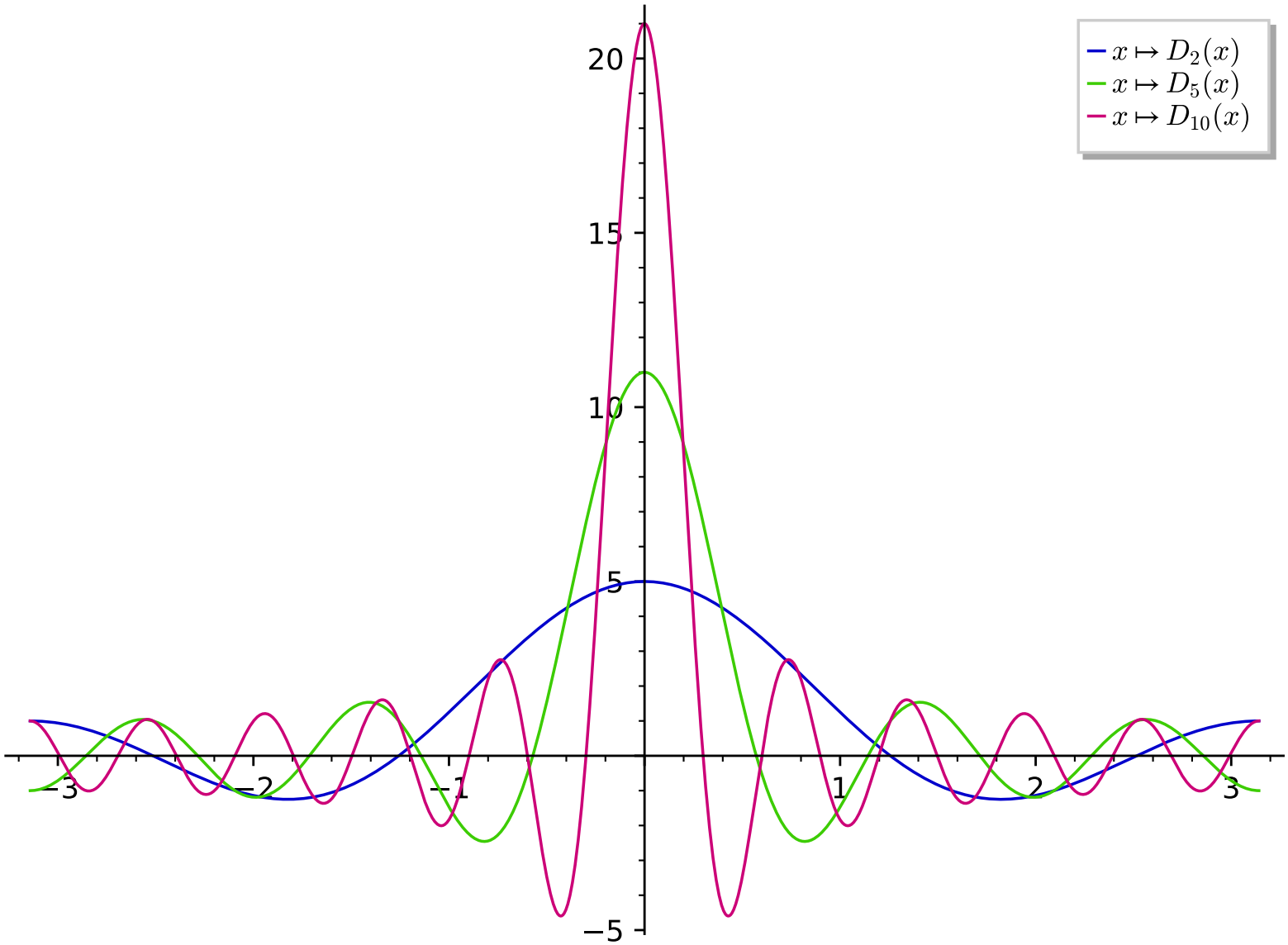


Figure: Some Dirichlet kernels

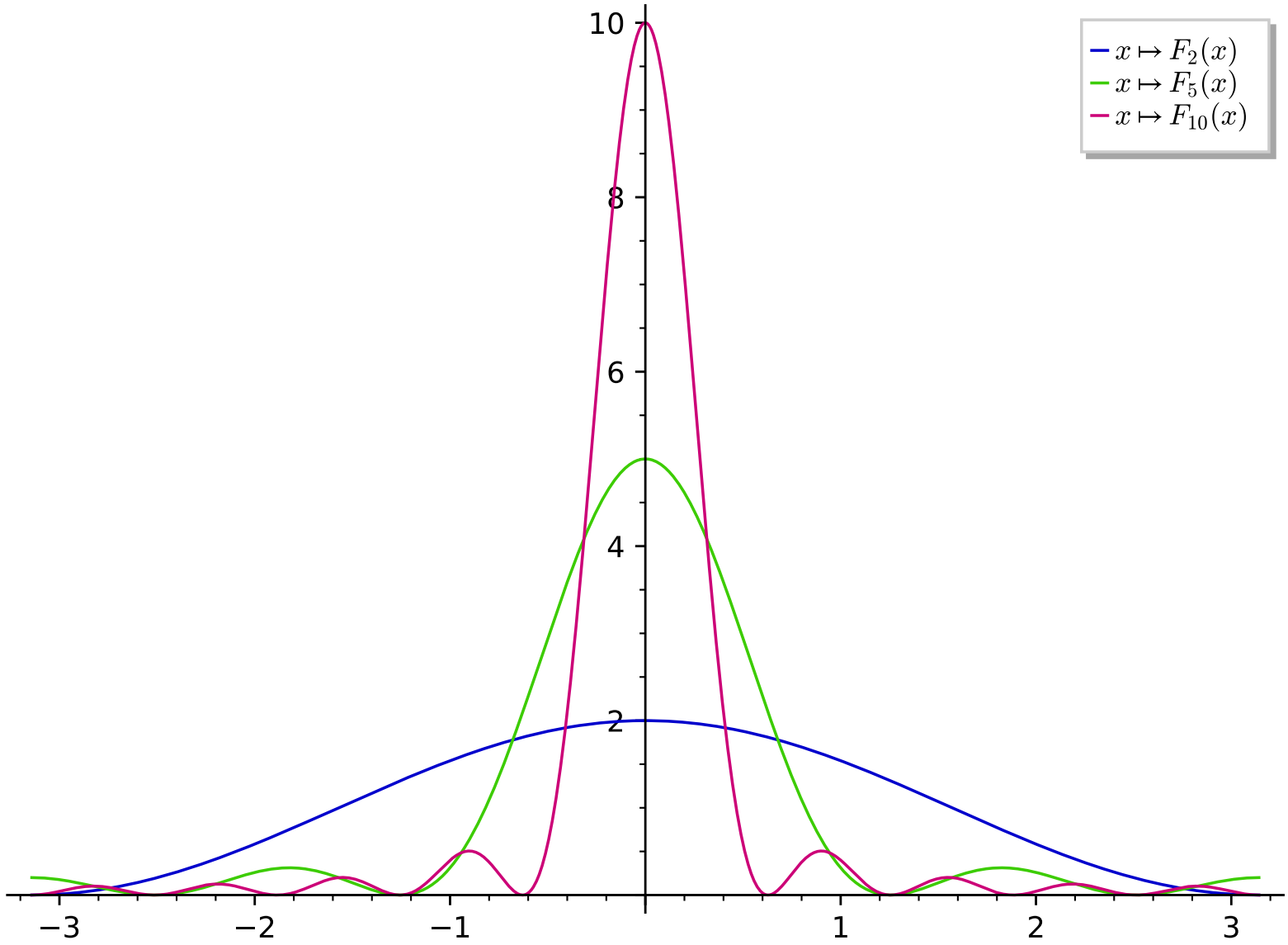


Figure: Some Fejér kernels

Proofs of the formulas

$$D_n(x) = e^{-inx} \sum_{k=0}^{2n} (e^{ix})^k = e^{-inx} \frac{e^{i(2n+1)x} - 1}{e^{ix} - 1} = \frac{e^{i(n+1)x} - e^{-inx}}{e^{ix} - 1}$$
$$= \frac{e^{i(n+1/2)x} - e^{-i(n+1/2)x}}{e^{ix/2} - e^{-ix/2}} = \frac{\sin\left((n + \frac{1}{2})x\right)}{\sin \frac{1}{2}x},$$

$$nF_n(x) = \sum_{k=0}^{n-1} \frac{e^{i(k+1)x} - e^{-ikx}}{e^{ix} - 1} = \frac{e^{ix}(e^{inx} - 1)}{(e^{ix} - 1)^2} - \frac{e^{-i(n-1)x}(e^{inx} - 1)}{(e^{ix} - 1)^2}$$
$$= \frac{(1 - e^{-inx})(e^{inx} - 1)}{(e^{ix/2} - e^{-ix/2})^2} = \frac{e^{inx} + e^{-inx} - 2}{(e^{ix/2} - e^{-ix/2})^2}$$
$$= \frac{(e^{inx/2} - e^{-inx/2})^2}{(e^{ix/2} - e^{-ix/2})^2} = \left(\frac{\sin\left(\frac{1}{2}nx\right)}{\sin \frac{1}{2}x}\right)^2$$

Lemma

The Fejér kernels have the following properties:

1. $\mathrm{F}_n(x) \geq 0$ *for all* $n \in \mathbb{N}$, $x \in \mathbb{R}$;
2. $\int_0^{2\pi} \mathrm{F}_n(x)\,\mathrm{d}x = 2\pi$ *for all* $n \in \mathbb{N}$;
3. $\lim_{n\to\infty} \int_r^{2\pi-r} \mathrm{F}_n(x)\,\mathrm{d}x = 0$ *for all* r *in the range* $0 < r < \pi$.

Property (3) is equivalent to $\lim_{n\to\infty} \int_{-r}^{r} \mathrm{F}_n(x)\,\mathrm{d}x = \int_{-\pi}^{\pi} \mathrm{F}_n(x)\,\mathrm{d}x = 2\pi$ for $0 < r < \pi$ and says that for large n the mass with density function $x \mapsto \mathrm{F}_n(x)\,\mathrm{d}x$ on $[-\pi, \pi]$ is concentrated near $x = 0$.

Proof.

(1) is clear from the closed formula for $\mathrm{F}_n(x)$; (2) follows from

$$\int_0^{2\pi} \mathrm{e}^{\mathrm{i}kx}\,\mathrm{d}x = \begin{cases} 2\pi & \text{if } k = 0, \\ 0 & \text{if } k \neq 0, \end{cases}$$

which shows $\int_0^{2\pi} \mathrm{D}_n(x)\,\mathrm{d}x = 2\pi$ and implies the corresponding result for F_n; (3) follows from the estimate $\mathrm{F}_n(x) \leq \frac{1}{n\sin^2(r/2)}$ for $r \leq x \leq 2\pi - r$ (or $-r < x < r$). □

Theorem

Suppose $f \in V$ (i.e., f is 2π-periodic and square-integrable over $[0, 2\pi]$.

1. *For every $\epsilon > 0$ there exists a trigonometric polynomial $g \in V$ such that $\|f - g\|_2 < \epsilon$.*
2. $\lim_{n\to\infty} \|f - S_n f\|_2 = 0$, *i.e., f is the limit of its Fourier polynomials in the metric space (V, d_2) ("L^2-limit", "mean-square" limit).*
3. *Bessel's Inequality holds with equality, i.e.,*

$$\sum_{k\in\mathbb{Z}} |c_k|^2 = \frac{1}{2\pi} \int_0^{2\pi} |f(x)|^2 \,\mathrm{d}x\,. \qquad \text{(PLANCHEREL's Identity)}$$

Remark

The following more general form of Plancherel's Identity is also true:

$$\sum_{k\in\mathbb{Z}} \overline{c}_k d_k = \frac{1}{2\pi} \int_0^{2\pi} \overline{f(x)} g(x) \,\mathrm{d}x \text{ for } f, g \in V, \quad \text{(PARSEVAL's Identity)}$$

where c_k, d_k denote the Fourier coefficients of f and g, respectively.

Proof.

First we show that (1), (2) and (3) are equivalent. The key to this is the formula

$$\|f-g\|_2^2 \geq \|f-\mathrm{S}_n f\|_2^2 = \|f\|_2^2 - 2\pi \sum_{k=-n}^{n} |c_k|^2 ,$$

which holds for all $g \in \mathrm{TP}_n$. (Recall that $\mathrm{S}_n f$ provides the best approximation to f in TP_n in the L^2-metric.)

If $\|f-g\| < \epsilon$ and $g \in \mathrm{TP}_N$, then $\|f-\mathrm{S}_n f\|_2^2 < \epsilon$ for all $n \geq N$, so that N can be taken as the response to ϵ in the proof of (2). Hence (1) implies (2). The converse is trivial, and that (2) and (3) are equivalent is immediate from the formula.

For the proof of (1) we use the concept of *(periodic) convolution* of two functions, which for $f, g \in V$ is defined by

$$(f*g)(x) = \frac{1}{2\pi}\int_0^{2\pi} f(y)g(x-y)\,\mathrm{d}y .$$

It is easy to see that the function $f*g\colon \mathbb{R} \to \mathbb{C}$ is 2π-periodic and square-integrable as well, and that the convolution operation $V \times V \to V$, $(f,g) \mapsto f*g$ is bilinear, associative, and commutative.

Proof cont'd.

In particular we have

$$f * e^{ikx} = \frac{1}{2\pi}\int_0^{2\pi} f(y)e^{ik(x-y)}\,dy = e^{ikx}\frac{1}{2\pi}\int_0^{2\pi} f(y)e^{-iky}\,dy = c_k\,e^{ikx}.$$

This shows that the Fourier coefficients of f are eigenvalues of the "multiplication map" $V \to V$, $g \mapsto f * g$ and implies

$$f * D_n = \sum_{k=-n}^{n} f * e^{ikx} = \sum_{k=-n}^{n} c_k e^{ikx} = S_n f,$$

$$f * F_n = \frac{1}{n}\sum_{k=0}^{n-1} f * D_k = \frac{1}{n}\sum_{k=0}^{n-1} S_k f$$

The function $\sigma_n f = \frac{1}{n}\sum_{k=0}^{n-1} S_n f$, which is obviously a trigonometric polynomial, is called n-th *Fejér polynomial* of f. The preceding computation shows that the Fejér polynomials have the integral representation

$$\sigma_n f(x) = f * F_n = F_n * f = \frac{1}{2\pi}\int_0^{2\pi} f(x-y)F_n(y)\,dy\,.$$

Proof cont'd.

Since $\frac{1}{2\pi}\int_0^{2\pi} \mathrm{F}_n(y)\,\mathrm{d}y = 1$, this gives

$$f(x) - \sigma_n f(x) = \frac{1}{2\pi}\int_0^{2\pi} \big(f(x) - f(x-y)\big)\mathrm{F}_n(y)\,\mathrm{d}y\,.$$

Now assume first that f is continuous. Then there exists $M > 0$ such that $|f(x)| \le M$ for $x \in [0, 2\pi]$. For $0 < r < \pi$ we then have

$$\begin{aligned}|f(x) - \sigma_n f(x)| &\le \frac{1}{2\pi}\int_0^{2\pi} |f(x) - f(x-y)|\,\mathrm{F}_n(y)\,\mathrm{d}y \\ &\le \frac{1}{2\pi}\int_{-r}^{r} |f(x) - f(x-y)|\,\mathrm{F}_n(y)\,\mathrm{d}y + \frac{2M}{2\pi}\int_r^{2\pi-r} \mathrm{F}_n(y)\,\mathrm{d}y.\end{aligned}$$

(Since the integrand is 2π-periodic, integrating over $[0, r]$ and $[2\pi - r, 2\pi]$ amounts to integrating over $[-r, r]$.)
The 1st summand can be made arbitrarily small (i.e., $< \epsilon/2$) by choosing r sufficiently small, since f is uniformly continuous and $\frac{1}{2\pi}\int_{-r}^{r} \mathrm{F}_n(y)\,\mathrm{d}y \le 1$.
By Property 3 Fejér kernels, the 2nd summand can then be made arbitrarily small by choosing n sufficiently large.

Proof cont'd.

$\Longrightarrow \sigma_n f$ converges uniformly to f, since this estimate holds independently of x.
But then $\sigma_n f$ converges to f also in the L^2-metric, as the estimate

$$\begin{aligned}\|f - \sigma_n f\|_2 &= \sqrt{\int_0^{2\pi} |f(x) - \sigma_n f(x)|^2 \,\mathrm{d}x} \\ &\leq \sqrt{2\pi}\max\{|f(x) - \sigma_n f(x)|\,; 0 \leq x \leq 2\pi\}\end{aligned}$$

shows.

Finally, it can be shown that an arbitrary function $f \in V$ can be approximated in the L^2-metric by continuous 2π-periodic functions, i.e., given $\epsilon > 0$ there exists a continuous function $h \in V$ such $\|f - h\|_2 < \epsilon/2$. The preceding argument then yields $n \in \mathbb{N}$ such that $\|h - \sigma_n h\| < \epsilon/2$, and the triangle inequality for the L^2-metric further $\|f - \sigma_n h\| < \epsilon$. Since $\sigma_n h$ is a trigonometric polynomial, this concludes the proof of (1). □

Example

We apply the Parseval identity to FOURIER's introductory example. Clearly the 4-periodic function f defined by

$$f(x) = \begin{cases} 1 & \text{if } -1 \le x < 1, \\ -1 & \text{if } 1 \le x < 3, \end{cases}$$

(the values at ± 1 don't matter for the mean-square approximation, so we can define them in any way) is square-integrable with $\int_{-1}^{3} f(x)^2 \,\mathrm{d}x = 4$. We have seen that the Fourier coefficients a_k, b_k are zero except for $a_{2k+1} = \frac{4(-1)^k}{\pi(2k+1)}$. Since $|c_k|^2 + |c_{-k}|^2 = \frac{1}{2}\left(|a_k|^2 + |b_k|^2\right)$, the Parseval identity gives

$$\frac{1}{2}\sum_{k=0}^{\infty} \frac{16}{\pi^2(2k+1)^2} = \sum_{k\in\mathbb{Z}} |c_k|^2 = \frac{1}{4}\int_{-1}^{3} f(x)^2 \,\mathrm{d}x = 1.$$

It follows that

$$\sum_{k=0}^{\infty} \frac{1}{(2k+1)^2} = 1 + \frac{1}{3^2} + \frac{1}{5^2} + \frac{1}{7^2} + \cdots = \frac{\pi^2}{8}.$$

Example

We compute the Fourier series of the "repeating ramp" function $g\colon \mathbb{R} \to \mathbb{R}$ defined by $g(x) = |x|$ for $-\pi \le x \le \pi$ and 2π-periodic extension. (Since $|-\pi| = |\pi|$, the 2π-periodic extension is well-defined.)

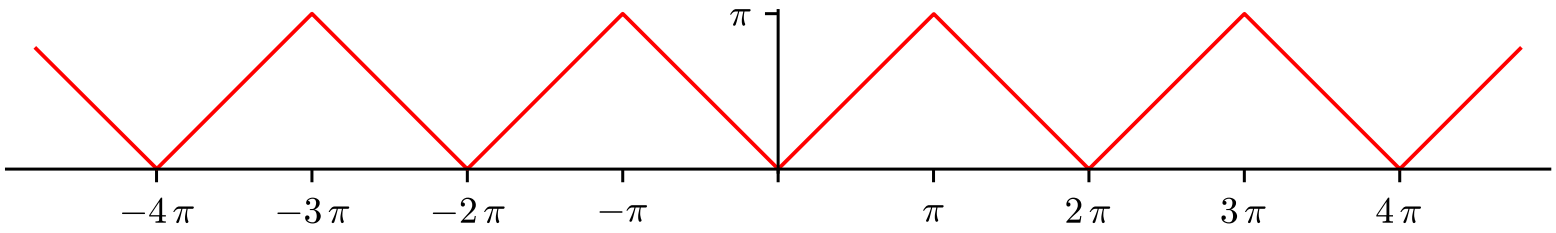


Figure: The repeating-ramp function

Since $g(x) = g(-x)$, the Fourier series of g is likewise a pure cosine series with

$$a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} g(x)\cos(kx)\,\mathrm{d}x = \frac{2}{\pi}\int_0^{\pi} x\cos(kx)$$

$$= \frac{2}{\pi}\left[\frac{x\sin(kx)}{k} + \frac{\cos(kx)}{k^2}\right]_0^{\pi} = \begin{cases} 0 & \text{for } k \ge 2 \text{ even,} \\ -\frac{4}{\pi k^2} & \text{for } k \text{ odd,} \end{cases}$$

Example (cont'd)

Moreover, $a_0 = \frac{2}{\pi}\int_0^\pi x\,\mathrm{d}x = \pi$.

$\Longrightarrow$ The Fourier series of g is

$$\frac{\pi}{2} - \frac{4}{\pi}\left(\cos x + \frac{\cos(3x)}{3^2} + \frac{\cos(5x)}{5^2} + \frac{\cos(7x)}{7^2} + \cdots\right).$$

In this case Parseval's identity gives

$$\frac{\pi^2}{4} + \frac{1}{2}\sum_{k=0}^{\infty}\frac{16}{\pi^2(2k+1)^4} = \frac{1}{2\pi}\int_{-\pi}^{\pi} g(x)^2\,\mathrm{d}x = \frac{1}{\pi}\int_0^\pi x^2\,\mathrm{d}x = \frac{\pi^2}{3}.$$

$$\Longrightarrow \quad \sum_{k=0}^{\infty}\frac{1}{(2k+1)^4} = \frac{\pi^2}{8}\left(\frac{\pi^2}{3} - \frac{\pi^2}{4}\right) = \frac{\pi^4}{96}.$$

From this one can easily derive EULER's formula for the sum of the reciprocals of the 4th powers as follows:

$$\sum_{n=1}^{\infty}\frac{1}{n^4} = \sum_{k=1}^{\infty}\frac{1}{(2k)^4} + \sum_{k=0}^{\infty}\frac{1}{(2k+1)^4} = \frac{1}{16}\sum_{n=1}^{\infty}\frac{1}{n^4} + \sum_{k=0}^{\infty}\frac{1}{(2k+1)^4}$$

Example (cont'd)

It follows that

$$\sum_{n=1}^{\infty} \frac{1}{n^4} = \frac{16}{15} \sum_{k=0}^{\infty} \frac{1}{(2k+1)^4} = \frac{16}{15}\,\frac{\pi^4}{96} = \frac{\pi^4}{90}.$$

This complements $\sum_{n=1}^{\infty} n^{-2} = \pi^2/6$, and similar identities can be derived for the sums $\sum_{n=1}^{\infty} n^{-2r}$, $r = 3, 4, 5, \ldots$.

But more is true: Since g is continuous and piece-wise C^1, the Fourier series of g represents g everywhere, i.e., we have

$$|x| = \frac{\pi}{2} - \frac{4}{\pi}\left(\cos x + \frac{\cos(3x)}{3^2} + \frac{\cos(5x)}{5^2} + \frac{\cos(7x)}{7^2} + \cdots\right)$$

for $x \in [-\pi, \pi]$; cf. the subsequent theorems.

You are invited to substitute a few particular values of x into this series and discover further interesting identities.

Theorem (Fejér)

Suppose $f : \mathbb{R} \to \mathbb{C}$ is 2π-periodic and the one-sided limits

$$f(x+) = \lim_{x' \downarrow x} f(x'), \quad f(x-) = \lim_{x' \uparrow x} f(x')$$

properly exist for all $x \in \mathbb{R}$. (It suffices to require this for $x \in [0, 2\pi)$, of course.)

1. *For every $x \in \mathbb{R}$ we have $\lim_{n\to\infty} \sigma_n f(x) = \frac{f(x+)+f(x-)}{2}$.*
   In particular, if f is continuous at x then $\lim_{n\to\infty} \sigma_n f(x) = f(x)$.
2. *If f is continuous everywhere then $(\sigma_n f)$ converges to f uniformly on $\mathbb{R}$.*

Note

The conditions on f imply that f is integrable (in fact even Riemann integrable) and bounded, hence square-integrable. But it is still too weak to conclude point-wise convergence of the Fourier series of f. However, if the Fourier series converges in x, it must have the limit in (1), i.e.,
$\lim_{n\to\infty} S_n(x) = \lim_{n\to\infty} \sigma_n(x) = \frac{1}{2}\big(f(x+) + f(x-)\big)$; cf. exercises.

Proof of Fejér's Theorem.

We have already shown (2) in the course of the proof of the previous theorem. The argument to prove (1) is similar. For $0 < r < \pi$ we have

$$\sigma_n f(x) = \frac{1}{2\pi}\int_0^{2\pi} f(x-y)\mathrm{F}_n(y)\,\mathrm{d}y$$
$$= \frac{1}{2\pi}\left(\int_0^r + \int_r^{2\pi-r} + \int_{2\pi-r}^{2\pi}\right) = \frac{1}{2\pi}\left(\int_0^r + \int_r^{2\pi-r} + \int_{-r}^{0}\right).$$

Since f is bounded, the middle integral can be made arbitarily small in absolute value by choosing n sufficiently large (possibly depending on r); cf. Porperty 3 of Fejér kernels.

The left integral can be rewritten as

$$\int_0^r f(x-y)\mathrm{F}_n(y)\,\mathrm{d}y = \int_0^r \big(f(x-y) - f(x-)\big)\mathrm{F}_n(y)\,\mathrm{d}y + f(x-)\int_0^r \mathrm{F}_n(y)\,\mathrm{d}y.$$

Here, the 1st summand can be made arbitrarily small in absolute value by choosing r appropriately (since $\int_0^r \mathrm{F}_n(y)\,\mathrm{d}y \le \int_0^\pi \mathrm{F}_n(y)\,\mathrm{d}y = \pi$), and the 2nd summand can be made arbitrarily close to $f(x-)\pi$ by choosing n sufficiently large

Proof of Fejér's Theorem cont'd.

(since $\mathrm{F}_n(x) = \mathrm{F}(-x)$ and hence $\int_0^r \mathrm{F}_n(y)\,\mathrm{d}y = \int_{-r}^0 \mathrm{F}_n(y)\,\mathrm{d}y \to \pi$ for $n \to \infty$).

A similar argument applies to the 3rd integral.

In all it follows that

$$\sigma_n f(x) - \frac{f(x-)}{2} - \frac{f(x+)}{2}$$

can be made arbitrarily small in absolute value by choosing n sufficiently large. This completes the proof of (1).

□

Point-wise Convergence of Fourier Series

If $x \in \mathbb{R}$ is such that $f(x+)$ and $f(x-)$ exist, we can define *one-sided derivatives* of f in x as

$$f'(x+) = \lim_{x' \downarrow x} \frac{f(x') - f(x+)}{x' - x}, \quad f'(x-) = \lim_{x' \uparrow x} \frac{f(x') - f(x-)}{x' - x},$$

provided that these limits exist.

Theorem (Dirichlet)

Suppose $f : \mathbb{R} \to \mathbb{C}$ is 2π-periodic and the one-sided limits $f(x\pm)$ exist for all $x \in \mathbb{R}$.

If $x \in \mathbb{R}$ is such that the one-sided derivatives $f'(x\pm)$ exist as well, then

$$\lim_{n\to\infty} S_n f(x) = \frac{f(x+) + f(x-)}{2}.$$

In particular, if in addition f is continuous at x then $\lim_{n\to\infty} S_n f(x) = f(x)$.

The proof of Dirichlet's Theorem is considerably more involved than that of Fejér's Theorem and will not be given in this lecture. Instead we state and prove a weaker version of Dirichlet's Theorem.

Theorem

Suppose $f : \mathbb{R} \to \mathbb{C}$ is 2π-periodic and continuous, and there exists a subdivision $0 = x_0 < x_1 < \cdots < x_r = 2\pi$ of $[0, 2\pi]$ such that the restriction of f to $[x_{i-1}, x_i]$ is a C^1-function for $1 \le i \le k$. Then the Fourier series of f converges uniformly to f on $\mathbb{R}$.

Proof.

Partial integration over $[x_{i-1}, x_i]$, where the functions involved are continuous, yields for $k \in \mathbb{Z} \setminus \{0\}$

$$\int_{x_{i-1}}^{x_i} f(x)\mathrm{e}^{-\mathrm{i}kx}\,\mathrm{d}x = \frac{1}{-\mathrm{i}k}\left(f(x)\mathrm{e}^{-\mathrm{i}kx}\Big|_{x_{i-1}}^{x_i} - \int_{x_{i-1}}^{x_i} f'(x)\mathrm{e}^{-\mathrm{i}kx}\,\mathrm{d}x\right).$$

Summing these identities for $1 \le i \le r$ and dividing by 2π gives

$$c_k = \frac{1}{2\pi}\int_0^{2\pi} f(x)\mathrm{e}^{-\mathrm{i}kx}\,\mathrm{d}x = \frac{1}{2\pi\mathrm{i}k}\int_0^{2\pi} f'(x)\mathrm{e}^{-\mathrm{i}kx}\,\mathrm{d}x = -\frac{\mathrm{i}}{k}\,c'_k,$$

since $f(2\pi) = f(0)$; here c'_k denote the Fourier coefficients of f'.

Proof cont'd.

The Cauchy-Schwarz Inequality gives further

$$\sum_{k\in\mathbb{Z}\setminus\{0\}} |c_k| = \sum_{k\in\mathbb{Z}\setminus\{0\}} \frac{|c_k'|}{k} \le \left(\sum_{k\in\mathbb{Z}\setminus\{0\}} |c_k'|^2\right)^{\frac{1}{2}} \left(\sum_{k\in\mathbb{Z}\setminus\{0\}} \frac{1}{k^2}\right)^{\frac{1}{2}}$$

The two sums on the right-hand side are $< \infty$, and hence the same is true of $\sum_{k\in\mathbb{Z}\setminus\{0\}} |c_k|$.

$\Longrightarrow$ The Fourier series $\sum_{k\in\mathbb{Z}} c_k e^{ikx}$ converges uniformly (and absolutely) on $\mathbb{R}$, say to $g(x)$, since it is majorized by the convergent series $\sum_{k\in\mathbb{Z}} |c_k|$.

$\Longrightarrow$ We can integrate the Fourier series term-wise and obtain that g has the same Fourier coefficients as f :

$$\int_0^{2\pi} g(x)e^{-ikx}\,dx = \int_0^{2\pi} \sum_{l\in\mathbb{Z}} c_l e^{ilx} e^{-ikx}\,dx$$

$$= \sum_{l\in\mathbb{Z}} c_l \int_0^{2\pi} e^{i(l-k)x}\,dx = 2\pi\, c_k.$$

Moreover, by the Continuity Theorem g is continuous as well.

Proof cont'd.

It remains to show that two continuous 2π-periodic functions which have the same Fourier series must be equal.

By a previous theorem, f and g are equal to the L^2-limit of their common Fourier series, and hence $\|f-g\|_2 = 0$ or, equivalently, $f(x) = g(x)$ almost everywhere.
But for continuous functions this can hold only if $f = g$, because $f(x_0) \neq g(x_0)$ implies $f(x) \neq g(x)$ in some interval $(x_0 - \delta, x_0 + \delta)$ of positive length. □

Exercise

The subject of this exercise is a more down-to-earth proof of the fact used in the last step of the proof of the preceding theorem: *Two continuous, 2π-periodic functions f and g having the same Fourier coefficients must be equal.*

1. Reduce the statement to the following: *A continuous, 2π-periodic function f having all Fourier coefficients equal to zero must be the all-zero function.*
2. Show that all Fejér polynomials $\sigma_n f$ of such a function f are zero.
3. Assume w.l.o.g. $f(x_0) = c > 0$ and hence $f(x) > c/2$ in some interval $(x_0 - r, x_0 + r)$ of positive length. Use the convolution representation

$$\sigma_n f(x_0) = \frac{1}{2\pi} \int_0^{2\pi} f(x_0 - y) \mathrm{F}_n(y)\, \mathrm{d}y$$

and the three properties of Fejér kernels stated earlier to derive a contradiction for large n.

Example

The theorem applies to the repeating-ramp function and gives the identity

$$|x| = \frac{\pi}{2} - \frac{4}{\pi}\left(\cos x + \frac{\cos(3x)}{3^2} + \frac{\cos(5x)}{5^2} + \frac{\cos(7x)}{7^2} + \cdots\right), \quad x \in [-\pi, \pi],$$

announced earlier.

Example (Partial fractions of the cotangent)

As a further example we consider the function $f_a\colon \mathbb{R} \to \mathbb{R}$ defined by $f_a(x) = \cos(ax)$ for $x \in [-\pi, \pi]$ and 2π-periodic extension.

For $a = k \in \mathbb{Z}$ the function $f_k(x) = \cos(kx)$ is its own (one-term) Fourier series and nothing interesting can be concluded.

For $a \in \mathbb{C} \setminus \mathbb{Z}$ the situation is more interesting, because f_a is then a "new" function.

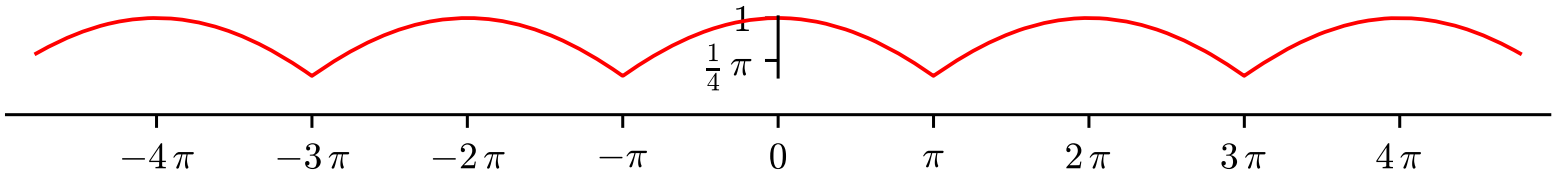


Figure: The function $x \mapsto f_{\frac{1}{4}}(x)$

Example (cont'd)

Since f_a is even, we have $b_k = 0$ for all k.

$$\begin{aligned} a_k &= \frac{2}{\pi}\int_0^\pi \cos(ax)\cos(kx)\,\mathrm{d}x \\ &= \frac{1}{\pi}\int_0^\pi \cos(ax+kx)+\cos(ax-kx)\,\mathrm{d}x \\ &= \frac{1}{\pi}\left[\frac{\sin\big((a+k)x\big)}{a+k}+\frac{\sin\big((a-k)x\big)}{a-k}\right]_0^\pi \\ &= \frac{1}{\pi}\left(\frac{(-1)^k\sin(a\pi)}{a+k}+\frac{(-1)^k\sin(a\pi)}{a-k}\right) \end{aligned}$$

$\Longrightarrow$ The Fourier series of f_a is

$$\frac{\sin(a\pi)}{\pi}\left[\frac{1}{a}+\sum_{k=1}^{\infty}(-1)^k\left(\frac{1}{a+k}-\frac{1}{a-k}\right)\cos(kx)\right].$$

Since $f_a(x)$ is continuous and piece-wise C^1, the theorem gives that this series is equal to $\cos(ax)$ for $x \in [-\pi,\pi]$.

Example (cont'd)

Thus we have for $a \in \mathbb{C} \setminus \mathbb{Z}$ and $x \in [-\pi, \pi]$ the identity

$$\cos(ax) = \frac{\sin(a\pi)}{\pi}\left[\frac{1}{a} + \sum_{k=1}^{\infty}(-1)^k\left(\frac{1}{a+k} + \frac{1}{a-k}\right)\cos(kx)\right].$$

Setting $x = \pi$ gives

$$\pi\cot(a\pi) = \frac{\pi\cos(a\pi)}{\sin(a\pi)} = \frac{1}{a} + \sum_{k=1}^{\infty}(-1)^k\left(\frac{1}{a+k} + \frac{1}{a-k}\right)$$

$$= \frac{1}{a} + 2a\sum_{k=1}^{\infty}\frac{(-1)^k}{a^2-k^2}. \qquad (a \in \mathbb{C} \setminus \mathbb{Z})$$

This famous identity, supposedly due to EULER, is known as the *partial fractions decomposition of the cotangent*.

Example

As an example for Dirichlet's Theorem we compute the Fourier series of the function $h\colon \mathbb{R} \to \mathbb{R}$ defined by $h(x) = (\pi - x)/2$ for $0 \le x < 2\pi$ and extended periodically.

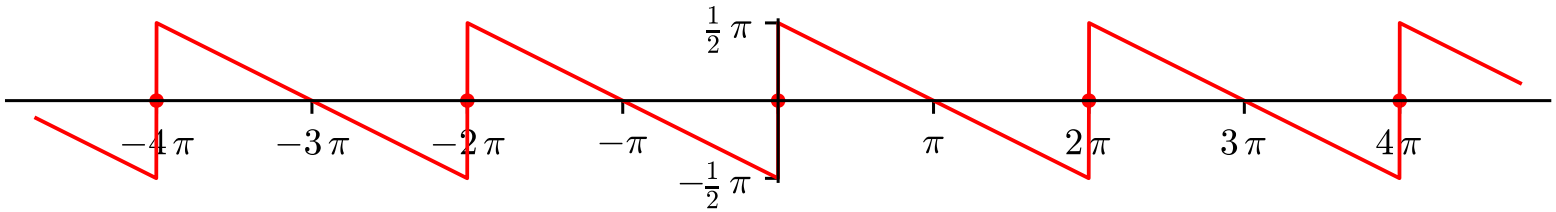


Figure: The function represented by the Fourier series of h

'Since h is odd, we have $a_k = 0$ for all k.

$$
\begin{aligned}
b_k &= \frac{1}{\pi} \int_0^{2\pi} \frac{\pi - x}{2} \sin(kx)\,\mathrm{d}x \\
&= \frac{1}{\pi} \left(-\frac{(\pi - x)\cos(kx)}{2k}\bigg|_0^{2\pi} - \frac{1}{2k} \int_0^{2\pi} \cos(kx)\,\mathrm{d}x \right) \\
&= \frac{1}{\pi} \left(\frac{\pi}{2k} + \frac{\pi}{2k} - 0 \right) = \frac{1}{k}.
\end{aligned}
$$

Example (cont'd)

Hence the Fourier series of h is $\sum_{k=1}^{\infty} \frac{\sin(kx)}{k}$, and we obtain from Dirichlet's Theorem the series representation

$$\sum_{k=1}^{\infty} \frac{\sin(kx)}{k} = \begin{cases} (\pi - x)/2 & \text{if } 0 < x < 2\pi, \\ 0 & \text{if } x = 0 \vee x = 2\pi. \end{cases}$$

Recall that we have derived this result already when discussing uniform convergence.

Exercise

For a sequence (a_n) of complex numbers the associated sequence (c_n) of CESÀRO *means* is defined by

$$c_n = \frac{a_1 + a_2 + \cdots + a_n}{n}.$$

a) Show that $\lim_{n\to\infty} a_n = A \in \mathbb{C}$ implies $\lim_{n\to\infty} c_n = A$.

b) Give an example of a divergent sequence (a_n) for which the sequence of Cesàro means converges.

Since the Fejér polynomials $\sigma_n f$ are Cesàro means of the Fourier polynomials, Part a) shows that that the convergence of the Fourier series of f at x implies $\lim_{n\to\infty} \sigma_n f(x) = \lim_{n\to\infty} S_n f(x)$.

Exercise

Prove the properties of the periodic convolution operation $V \times V \to V$, $(f, g) \mapsto f * g$ mentioned in the lecture.

The End

We wish you every success in the final examination!

Final Examination

Date/Time/Venue

Sun May 26 2023, 14:00–17:00

Instructions to candidates

- This examination paper contains six (6) questions.
- Please answer every question and subquestion, and JUSTIFY your answers.
- For your answers please use the space provided after each question. If this space is insufficient, please continue on the blank sheets provided.
- This is a CLOSED BOOK examination, except that you may bring 1 sheet of A4 paper (hand-written only) and a Chinese-English dictionary (paper copy only) to the examination.