# Ordinary Differential Equations

(preliminary draft updated Jan 4 2024)



李思 *(Si Li)*

Tsinghua University

Thanks very much for your support of this note! It is greatly appreciated if you are willing to help improve it by sending your comments such as typo, mistake or suggestion at your tea time. You are welcome to submit your comment via either the website

https://www.wjx.cn/vm/wfNWML4.aspx

or the barcode below

You can also contact me at sili@mail.tsinghua.edu.cn. The draft will be updated on my homepage: https://sili-math.github.io/. Thank you.

# Contents

# Preface

This note provides an elementary introduction to the theory of ordinary differential equations. It is based on the undergraduate course "Ordinary Differential Equations" that I lectured at Tsinghua University in 2023. Chapter 1 and 2 start with concepts in ordinary differential equations and discuss a few cases where solutions can be found explicitly. Chapter 3 and 4 cover fundamental theory on the well-posedness of initial value problems as well as analytic properties of solutions. Chapter 5 explains boundary value problems and their solutions via Green's functions. Chapter 6 discusses the action principle and variation method, with focus on the Euler-Lagrange equation and its application to a few historical examples. Chapter 7 introduces basic ideas on numerical solutions and their convergence. These topics serve for a one-semester lecturing on an introductory course in the subject of ordinary differential equations. There are many great resources about these topics in the literature. We listed a few that we have consulted at the end of this note.

I would like to thank 顾坪昕 and 汤乐琪, who have done amazing jobs of teaching assistant for this course. An early version of this note was typed by 顾坪昕, including all those beautiful figures. I want to thank 曲仟仟, 刘汉 and 任子逸 for their help on careful proofreading of this note, as well as their important roles of being excellent students for the whole semester. Special thanks go to a few friends who have been asking me various questions in differential equations during the semester and have been pushing me to finish this note.

# Chapter 1 Introduction

## 1.1 Basic Concepts

### 1.1.1 ODE and PDE

A differential equation is a relation between a set of unknown functions and their derivatives. To find a solution is to figure out a set of functions that satisfy the corresponding relation.

For example

$$\frac{dy}{dt} = 2y$$

is a differential equation describing a function $y(t)$ with variable $t$. One solution is given by

$$y = e^{2t}.$$

It is straightforward to check that this function indeed satisfies the above differential equation:

$$\frac{d}{dt}(e^{2t}) = 2(e^{2t}).$$

Solution of a differential equation may not be unique. For example, $y = Ce^{2t}$ is a solution to the above equation $\frac{dy}{dt} = 2y$ for any constant $C$. Also, solution may not exist. For example, the differential equation

$$\left(\frac{dy}{dt}\right)^2 = -1 - t^2 - y^2$$

does not admit a solution for real valued function $y(t)$. The existence and uniqueness problems for solution under certain circumstances play a majoy role in the study of differential equations.

Ordinary differential equations (ODE) are about unknown functions that depend on a single independent variable. The example above is an ODE. In general, an ODE could involve several unknown functions that all depend on the same single variable. For example, consider a particle of mass $m$ moving in the space in the presence of a force $\vec{\mathbf{F}} = (F_x, F_y, F_z)$ which depend on the position $(x, y, z)$ in the space. Then Newton's law says that the trajectories of the particle are solutions to

$$\begin{cases} m\dfrac{d^2x}{dt^2} = F_x \\ m\dfrac{d^2y}{dt^2} = F_y \\ m\dfrac{d^2z}{dt^2} = F_z \end{cases}$$

which is an ODE with three functions $\{x(t), y(t), z(t)\}$ in terms of the single time variable $t$.

On the other hand, partial differential equations (PDE) are about unknown functions that depend on several variables. For example, the heat equation

$$\frac{\partial}{\partial t}u(x,t) = \frac{\partial^2}{\partial x^2}u(x,t)$$

is a partial differential equation in two variables $t$ and $x$. It describes the conduction of heat in a solid body distributed on the line at position $x$ and time $t$.

In this note, we will focus on ordinary differential equations.

### 1.1.2 System of Differential Equations

If there is only one unknown function, then one equation is sufficient. If there are two or more unknown functions, then a system of equations is required. We have seen one example of moving particle above. As another example, the Lotka-Volterra (or predator-prey) equations

$$\begin{cases} \dfrac{dx}{dt} = \alpha x - \beta xy \\ \dfrac{dy}{dt} = \delta xy - \gamma y \end{cases} \qquad \alpha, \beta, \delta, \gamma \text{ are constants}$$

describe the dynamics of biological system of a predator and a prey. The function $x(t)$ describes the population density of prey, and $y(t)$ describes the population density of predator.

Such equations are called a system of differential equations.

### 1.1.3 Linear and Nonlinear Equations

Let us consider an ordinary differential equation

$$F(t, y, y', \cdots, y^{(n)}) = 0$$

for a unknown function $y$ with variable $t$. Here $y' = \frac{dy}{dt}$ and $y^{(n)} = (\frac{d}{dt})^n y$ denote the corresponding derivatives of $y$. The above function is called linear if $F$ is linear in the functions $y, y', \cdots, y^{(n)}$. Otherwise it is called nonlinear. For example

$$y'' - y' + ty + t^2 = 0$$

is a linear differential equation. The following

$$y' + y^2 = 0$$

is a nonlinear differential equation.

Linear equations are easier to study and explicit solutions are usually available. Nonlinear equations are more complicated and exhibit further exotic phenomenons. We will start to study linear equations in Chapter 2, and the rest of this note is mainly devoted to nonlinear equations.

### 1.1.4 Order

The oder of a differential equation is the order of the highest derivative that appears in the equation. For example,

$$(y')^{\,2} + ty = 0$$

is a 1st-order nonlinear equation, while

$$y''' + ty' + y = 0$$

is a 3rd-order linear equation.

A general linear ODE of order $n$ is of the form

$$a_0(t)y^{(n)} + a_1(t)y^{(n-1)} + \cdots + a_n(t)y + b(t) = 0$$

where $a_i(t)$ and $b(t)$ are known functions of $t$. It is called homogeneous if $b(t) = 0$. It is said to have constant coefficients if the functions $a_0(t), \cdots, a_n(t)$ do not depend on $t$, *i.e.* are constant functions. For exmaple,

$$y'' + y = 0$$

is a 2nd-order homogeneous linear equation with constant coefficients.

Let us point out that all ODEs can be equivalently described by a system of 1st-order ODE. This is useful since it will often reduce our work to study 1st-order ODE only.

To illustrate the basic idea of this reduction, consider the following ODE of order $n$

$$F(t, y, y', \cdots, y^{(n)}) = 0.$$

This can be recast as the following 1st-order system

$$\begin{cases} F(t, y, y_1, \cdots, y_{n-1}, y'_{n-1}) = 0 \\ y' = y_1 \\ y'_1 = y_2 \\ \quad \vdots \\ y'_{n-2} = y_{n-1} \end{cases}$$

It is clear that solving this 1st-order system is equivalent to solving the original order $n$ equation.

As an exapmle, let us consider a homogeneous linear system of order $n$

$$y^{(n)} = a_0 y + a_1 y' + \cdots + a_{n-1} y^{(n-1)}.$$

This can be reduced to a linear system

$$\frac{d}{dt}\vec{y} = A \cdot \vec{y}$$

where the column vector $\vec{y}$ is

$$\vec{y} = \begin{pmatrix} y \\ y_1 \\ y_2 \\ \vdots \\ y_{n-1} \end{pmatrix}$$

and $A$ is the $n \times n$ matrix

$$A = \begin{pmatrix} 0 & 1 & \cdots & \cdots & 0 \\ \vdots & 0 & 1 & \cdots & \vdots \\ \vdots & \cdots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 & 1 \\ a_0 & a_1 & \cdots & a_{n-2} & a_{n-1} \end{pmatrix}$$

In general, any linear equation can be put into a 1st-order linear system of form

$$\frac{d}{dt}\vec{y} = A(t)\vec{y} + B(t)$$

where $\vec{y}$ is the column of unknown functions. $A(t)$ and $B(t)$ are a matrix and a column vector respectively that could depend on the variable $t$.

### 1.1.5 Integral Curve

We illustrate some basic geometric idea which is useful to keep in mind along our study of differential equations. Consider a 1st-order ODE of the form

$$\frac{dy}{dt} = f(t, y).$$

We can draw the vector $(1, f)$ at each point $(t, y)$. This will be called the direction field.

Now we can draw any solution $y(t)$ as a curve in the $(t, y)$-plane parametrized by

$$(t, y(t)).$$

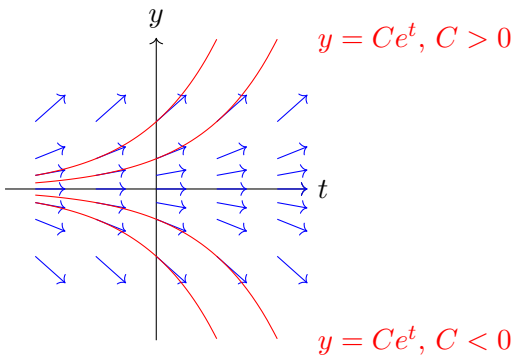
In other words, we consider the graph of the solution $y(t)$ in the $(t, y)$-plane. This is called the integral curve. Being a solution, the integral curve has the property that it is tangent to the direction vector $(1, f)$ since

$$\frac{d}{dt}(t, y(t)) = (1, f(t, y)).$$

**Example 1.1.1.** Consider the equation

$$\frac{dy}{dt} = y$$

The direction field is plotted as blue and the integral curves are drawn red.




Geometrically, the integral curve is obtained by "following the direction field".

## 1.2 Examples of Solutions

Very few differential equations can be solved in closed form. However, when explicit solving techniques are available, they provide some insights about behavious of differential equations. Such solving tricks for certain types of equations are, in some sense, a kind of art. We discuss a few of them in this section and study systematically for linear equations in Chapter 2.

### 1.2.1 Integrating Factor

Consider a differential equation for unknown function $y(t)$ which you can turn equivalently into the form

$$\frac{d}{dt}F(t,y) = 0$$

for some expression $F(t,y)$. Then solutions can be obtained via the relation defined by

$$F(t,y) = C$$

where $C$ is a constant. Sounds cheap, right? Yes! This is the most lucky situation you can have. Let us look at some examples.

**Example 1.2.1.** Consider the differential equation

$$\frac{dy}{dt} - ay = 0, \qquad a \text{ is a constant.}$$

We can multiply both sides by $e^{-at}$

$$e^{-at}\frac{dy}{dt} - ae^{-at}y = 0.$$

Then this is equivalent to

$$\frac{d}{dt}(e^{-at}y) = 0$$

which is solved by

$$e^{-at}y = C$$

*i.e.* ,

$$y = Ce^{at}$$

for some constant $C$. □

This example looks simple, but here you need to guess the factor $e^{-at}$ by observation. Things can be more tricky.

**Example 1.2.2.** Consider the equation

$$t\frac{dy}{dt} = y^2 + y + t^2.$$

Let us divide $t^2$ on both sides and write

$$\begin{aligned}
& \frac{1}{t}\frac{dy}{dt} - \frac{1}{t^2}y = \left(\frac{y}{t}\right)^2 + 1 \\
\Rightarrow \quad & \frac{d}{dt}\left(\frac{y}{t}\right) = \left(\frac{y}{t}\right)^2 + 1 \\
\Rightarrow \quad & \frac{\frac{d}{dt}\left(\frac{y}{t}\right)}{\left(\frac{y}{t}\right)^2 + 1} - 1 = 0 \\
\Rightarrow \quad & \frac{d}{dt}\left(\arctan\left(\frac{y}{t}\right) - t\right) = 0 \\
\Rightarrow \quad & \arctan\left(\frac{y}{t}\right) - t = C \qquad C \text{ is some constant}
\end{aligned}$$

from which we can solve $y$ for

$$y = t\tan(t + C).$$

□

In general, the equation

$$\frac{d}{dt}F(t, y) = 0$$

takes the form

$$\frac{\partial}{\partial t}F(t, y) + \frac{\partial F(t, y)}{\partial y}\frac{dy}{dt} = 0.$$

Suppose we have an equation of the form

$$Q(t, y) + P(t, y)\frac{dy}{dt} = 0.$$

Then a necessary condition for this to be a total derivative of the above form is

$$\frac{\partial}{\partial t}P(t, y) = \frac{\partial}{\partial y}Q(t, y).$$

In fact, if we can write

$$\begin{cases} P(t, y) = \dfrac{\partial F(t, y)}{\partial y} \\ Q(t, y) = \dfrac{\partial F(t, y)}{\partial t} \end{cases}$$

for some $F$, then

$$\begin{aligned}
\frac{\partial}{\partial t}P &= \frac{\partial^2 F(t, y)}{\partial t \partial y} \\
\frac{\partial}{\partial y}Q &= \frac{\partial^2 F(t, y)}{\partial y \partial t}
\end{aligned}$$

and therefore $\frac{\partial P}{\partial t}$ has to be the same as $\frac{\partial Q}{\partial y}$. Thus if you observe

$$\frac{\partial P}{\partial t} = \frac{\partial Q}{\partial y}$$

holds, then you are likely to find such $F$.

If you find $\frac{\partial P}{\partial t} \neq \frac{\partial Q}{\partial y}$, you can still try to multiply by a function $A(t, y)$

$$A(t, y)Q(t, y) + A(t, y)P(t, y)\frac{dy}{dt} = 0$$

and see whether

$$\frac{\partial}{\partial t}(AP) = \frac{\partial}{\partial y}(AQ)$$

holds. Such $A$ is called an integrating factor. If you can guess or find such an integrating factor, then you are done with good luck.

In the example

$$\frac{dy}{dt} - ay = 0$$

we have $P = 1$ and $Q = -ay$. We check

$$\frac{\partial P}{\partial t} = 0 \neq \frac{\partial Q}{\partial y} = -a.$$

The integrating factor is $A = e^{-at}$, which helps us to solve the equation in this case.

It is a very special situation to be able to find an integrating factor. But somehow this is the first thing that you would try and guess.

### 1.2.2 Seperation of Variables

Consider a differential equation of the form

$$\frac{dy}{dt} = \varphi(t, y).$$

If $\varphi$ can be written as a product of the form

$$\varphi(t, y) = f(y)g(t),$$

we say this equation is separable. For separable equation

$$\frac{dy}{dt} = f(y)g(t)$$

we can write this as (need extra care about the process of dividing, see below)

$$\frac{dy}{f(y)} = g(t)dt.$$

If we integrate both sides, we get

$$\int \frac{dy}{f(y)} = \int g(t)dt$$

which gives the relation between $y$ and $t$.

Note that in this process, we may miss some special solutions due to the process of dividing. For example, if $f(a) = 0$, then the constant function $y = a$ is a solution. Such special solution can be added by hand at the end.

**Example 1.2.3.** Let us again look at

$$\frac{dy}{dt} = ay.$$

This time we treat it as a separable equation, with $f(y) = y$ and $g(t) = a$. Then

$$\begin{aligned} & \frac{dy}{y} = a\ dt \\ \Rightarrow \qquad & \ln|y| = at + C' \\ \Rightarrow \qquad & y = Ce^{at}, \qquad C = \pm e^{C'} \end{aligned}$$

Here $C = \pm e^{C'} \neq 0$. The missing special solution is $y = 0$, which corresponds to the case $C = 0$. Adding this back, we get the same set of solutions as before.

**Example 1.2.4.**

$$\frac{dy}{dt} = e^y \sin t$$

This is a separable equation. We have

$$\begin{aligned} & e^{-y}dy = \sin t dt \\ \Rightarrow \qquad & \int e^{-y}dy = \int \sin t dt \\ \Rightarrow \qquad & -e^{-y} = -\cos t + C \\ \Rightarrow \qquad & y = -\ln(\cos t - C). \end{aligned}$$

This expression illustrates the following interesting phenomenon about different solutions.

- If $C < -1$, then $\cos t - C > 0$ for all $t$. The solution $y(t)$ exists for all $t$.
- If $-1 \leq C \leq 1$, the solution exists only for a finite time interval and then blow up.

We will discuss this phenomenon systematically in Section 3.2.

### 1.2.3 Change of Variable

As we study in calculus, change of variable is a useful method to perform integrals. It is also a standard trick to solve differential equations. We illustrate by a few examples.

#### Linear Change

Consider a differential equation of the form

$$\frac{dy}{dt} = f(at + by)$$

where $a, b$ are constants. This can be solved by introducing

$$\begin{aligned} & z(t) = at + by(t) \\ \Rightarrow \qquad & \frac{dz}{dt} = a + bf(z) \end{aligned}$$

which becomes a separable equation.

**Example 1.2.5.**

$$\frac{dy}{dt} = \frac{1}{t+3y}$$

Let $z = t + 3y$. Then

$$\begin{aligned}
& \frac{dz}{dt} = 1 + \frac{3}{z} = \frac{z+3}{z} \\
\Rightarrow \quad & \frac{z}{z+3} dz = dt \\
\Rightarrow \quad & \int (1 - \frac{3}{z+3}) dz = \int dt \\
\Rightarrow \quad & z - 3\ln|z+3| = t + C \\
\overset{z=t+3y}{\Rightarrow} \quad & 3y - 3\ln|t+3y+3| = C \\
\Rightarrow \quad & t + 3y + 3 = C' e^y, \qquad C' = e^{-C/3}
\end{aligned}$$

The special solution is $t + 3y + 3 = 0$, *i.e.* ,

$$y = -\frac{t+3}{3}.$$

So all solutions can be obtained from the relation

$$t + 3y + 3 = Ce^y$$

where now $C$ can be an arbitrary constant.

**Homogeneous Equation**

This is the case for

$$\frac{dy}{dt} = f\left(\frac{y}{t}\right).$$

We can rewrite this equation in terms of

$$\begin{aligned}
& u(t) = \frac{y(t)}{t} \\
\Rightarrow \quad & \frac{du}{dt} = -\frac{y}{t^2} + \frac{1}{t}\frac{dy}{dt} \\
\Rightarrow \quad & \frac{du}{dt} = \frac{f(u) - u}{t}
\end{aligned}$$

which becomes a separable equation.

**Example 1.2.6.**

$$\frac{dy}{dt} = \frac{y}{t} - \frac{t}{y}$$

Let $u = \frac{y}{t}$. Then

$$\begin{aligned}
& \frac{du}{dt} = \frac{u - \frac{1}{u} - u}{t} = -\frac{1}{ut} \\
\Rightarrow \quad & u du = -\frac{dt}{t} \\
\Rightarrow \quad & \frac{1}{2}u^2 = -\ln|t| + C_1 \\
\Rightarrow \quad & u = \pm\sqrt{C - 2\ln|t|}
\end{aligned}$$

# Chapter 2 Linear Equations

In this chapter we study linear ordinary differential equations. Linear equations appear frequently in applications, and they can be solved and analyzed explicitly. We introduce the method of exponential matrices for solving linear systems with constant coefficients, and generalize it to the path-ordered exponential for nonautonomous linear systems (linear systems with varying coefficients). We also discuss the method of characteristic polynomials and variation of parameters for solving linear differential equations.

## 2.1 Linear Systems with Constant Coefficients

We will start our study from linear systems with constant coefficients. As we have discussed in Section 1.1.4, such a system can always be reduced to a 1st-order system of the form

$$\frac{d}{dt}\vec{y} = A \cdot \vec{y} + \vec{b}(t)$$

where $\vec{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ is the column of unknown functions,

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}$$

is a $n \times n$ matrix with constant entries, and

$$\vec{b}(t) = \begin{pmatrix} b_1(t) \\ b_2(t) \\ \vdots \\ b_n(t) \end{pmatrix}$$

is a column that in general depends on the variable $t$.

### 2.1.1 1st Order Homogeneous System

Let us first consider the homogeneous case when $\vec{b}(t) = 0$. The equation

$$\frac{d}{dt}\vec{y} = A \cdot \vec{y}$$

looks very much like the scalar equation $\frac{dy}{dt} = ay$. In fact, the followig theorem shows that it can be solved in a similar way.

**Theorem 2.1.1.** *Given any column vector $\vec{y}_0 \in \mathbb{R}^n$, there exists a unique solution to the equation*

$$\frac{d}{dt}\vec{y} = A\vec{y}$$

*that satisfies the initial condition $\vec{y}(0) = \vec{y}_0$. The solution is explicitly given by*

$$\vec{y}(t) = e^{tA}\vec{y}_0.$$

We will explain the meaning of the exponential matrix $e^{tA}$ in a minute. Nevertheless, the expression of the above solution is intuitive. Based on our experience on the exp function, with a bit of brave, we can check

$$\frac{d}{dt}(e^{tA}\vec{y}_0) = Ae^{tA}\vec{y}_0$$

which indeed satisfies the required equation. The initial condition is also manifest:

$$e^{tA}\vec{y}_0\big|_{t=0} = e^0 \cdot \vec{y}_0 = \vec{y}_0.$$

### Exponential Matrix

Let us denote

$$M_n = \{\text{real } n \times n \text{ matrices}\}.$$

We will define a norm, hence a distance function, on the space $M_n$. This will allow us to talk about limit and convergence for matrices.

Recall that for any vector

$$\vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n$$

we have a Euclidean norm defined by

$$|\vec{x}| := \sqrt{x_1^2 + x_2^2 + \cdots x_n^2}.$$

Let $A \in M_n$ be a $n \times n$ matrix. It defines a linear map

$$\begin{aligned} A: \qquad & \mathbb{R}^n \to \mathbb{R}^n \\ & \vec{x} \mapsto A \cdot \vec{x}. \end{aligned}$$

This allows us to define a norm, which is called operator norm, by

$$\|A\| := \sup_{\substack{|\vec{x}|=1 \\ \vec{x}\in\mathbb{R}^n}} |A\vec{x}|.$$

The norm $\|A\|$ measures the "size" of $A$. Note that there are many different kinds of norms that can be defined on $M_n$. We will use the above operator norm which is convenient for our discussions (in fact, this is also defined for linear operators on infinite dimensional spaces). Note that the sup can be achieved since $\{\vec{x} \in \mathbb{R}^n \mid |\vec{x}| = 1\}$ is compact.

**Proposition 2.1.2.** *Assume* $A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}$. *Then*

$$\max_{i,j} |a_{ij}| \leq \|A\| \leq \sum_{i,j} |a_{ij}|.$$

*Proof:* Let us write

$$A = \begin{pmatrix} \vec{u}_1^T \\ \vdots \\ \vec{u}_n^T \end{pmatrix} \qquad \text{where} \qquad \vec{u}_k = \begin{pmatrix} a_{k1} \\ \vdots \\ a_{kn} \end{pmatrix} \in \mathbb{R}^n$$

and $T$ refers to transpose. Then for any $\vec{x} \in \mathbb{R}^n$ with $|\vec{x}| = 1$, we have

$$A\vec{x} = \begin{pmatrix} \vec{u}_1^T \cdot \vec{x} \\ \vdots \\ \vec{u}_n^T \cdot \vec{x} \end{pmatrix} = \begin{pmatrix} \langle \vec{u}_1, \vec{x} \rangle \\ \vdots \\ \langle \vec{u}_n, \vec{x} \rangle \end{pmatrix}$$

where $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product on $\mathbb{R}^n$. Using the Cauchy-Schwarz inequality

$$|\langle \vec{u}, \vec{v} \rangle| \leq |\vec{u}||\vec{v}| \qquad \forall \vec{u}, \vec{v} \in \mathbb{R}^n,$$

we have

$$|A\vec{x}| \leq \sum_i |\langle \vec{u}_i, \vec{x} \rangle| \leq \sum_i |\vec{u}_i||\vec{x}| = \sum_i |\vec{u}_i| \leq \sum_i \sum_j |a_{ij}|.$$

This proves the inequality on the right hand side.

Consider the unit vector

$$\vec{e}_j = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \leftarrow j$$

we have

$$A\vec{e}_j = \begin{pmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{nj} \end{pmatrix}$$

By definition

$$\|A\| \geq |A\vec{e}_j| \geq |a_{ij}| \qquad \text{for any } i, j.$$

This proves the inequality on the left hand side. □

**Proposition 2.1.3.** *For matrices $A, B \in M_n$ and $\vec{x} \in \mathbb{R}^n$*

① $\|A\| \geq 0$. $\|A\| = 0$ *if and only if* $A = 0$

② $\|\lambda A\| = |\lambda|\|A\|$ *for* $\lambda \in \mathbb{R}$

③ $\|A + B\| \leq \|A\| + \|B\|$

④ $|A\vec{x}| \leq \|A\||\vec{x}|$

⑤ $\|AB\| \leq \|A\|\|B\|$

*Proof:* Exercise □

In particular, properties ①②③ simply that $\|\cdot\|$ defines a norm on $M_n$. This allows us to talk about limit and convergence. We say a sequence of matrices $\{A_n\}$ converges to $B$ if

$$\lim_{n\to\infty} \|A_n - B\| = 0.$$

In this case, we write $B$ as the limit

$$\lim_{n\to\infty} A_n = B.$$

Note that by Proposition 2.1.3, $\lim_{n\to\infty} A_n = B$ is the same as saying that the limit of each entry of $A_n$ is the corresponding entry of $B$.

Let $A \in M_n$ be any square matrix. By ⑤ of Proposition 2.1.3, we have

$$\left\| \frac{A^n}{n!} \right\| \leq \frac{\|A\|^n}{n!}$$

which decays to zero very fast as $n \to \infty$. This implies that the limit

$$\lim_{N\to\infty} \sum_{k=0}^{N} \frac{A^k}{k!}$$

exists, and we denote this limit matrix by

$$e^A := \sum_{k=0}^{\infty} \frac{A^k}{k!}.$$

So the power series for exponential function works for matrices.

Similarly, the matrix $e^{tA}$ that depends on the variable $t$ is

$$e^{tA} = \sum_{k=0}^{\infty} \frac{t^k A^k}{k!}.$$

By a bit of further analysis, you can show that $e^{tA}$ depends smoothly on $t$. Its growth with $t$ is bounded by

$$\|e^{tA}\| = \left\| \sum_{k=0}^{\infty} \frac{t^k A^k}{k!} \right\| \leq \sum_{k=0}^{\infty} \left\| \frac{t^k A^k}{k!} \right\| \leq \sum_{k=0}^{\infty} \frac{|t|^k \|A\|^k}{k!} = e^{|t|\|A\|}$$

when $|t| \to \infty$.

The convergence property for the series defining $e^A$ is as good as that for the series defining the exponential function $e^x$. This allows us to do many calculations for $e^A$ as that for $e^x$.

**Proposition 2.1.4.**

$$\frac{d}{dt}e^{tA} = Ae^{tA}.$$

*Proof:* For each $k > 0$, we have

$$\frac{d}{dt}\left(\frac{t^k A^k}{k!}\right) = \frac{t^{k-1}A^k}{(k-1)!}.$$

The proposition follows by sum over $k$. □

**Proposition 2.1.5.** *If square matrices $A$ and $B$ commute with each other,* i.e. $AB = BA$, *then*

$$e^A \cdot e^B = e^{A+B}$$

*Proof:*

$$\begin{aligned} e^{A+B} &= \sum_{n=0}^{\infty} \frac{(A+B)^n}{n!} \\ &\overset{\text{Using } AB=BA}{=} \sum_{n=0}^{\infty} \frac{1}{n!} \sum_{k=0}^{n} \binom{n}{k} A^k B^{n-k} \\ &= \sum_{n=0}^{\infty} \sum_{k+m=n} \frac{A^k B^m}{k!m!} \\ &= \left(\sum_{k=0}^{\infty} \frac{A^k}{k!}\right)\left(\sum_{m=0}^{\infty} \frac{B^m}{m!}\right) = e^A e^B \end{aligned}$$

There is a slight missing analysis justifying that the above power series calculation is allowed. We leave it to the reader. □

*Remark* 2.1.6. In general, if $AB \neq BA$, there is the beautiful Baker-Campbell-Hausdorff formula

$$e^A \cdot e^B = e^{A+B+\frac{1}{2}[A,B]+\frac{1}{12}[A,[A,B]]-\frac{1}{12}[B,[A,B]]+\cdots}$$

where "$\cdots$" indicates explicit higher commutator expressions.

**Proposition 2.1.7.** *For square matrix $A$, the exponential matrix $e^A$ is invertible and*

$$(e^A)^{-1} = e^{-A}$$

*Proof:* $A$ and $-A$ clearly commute. Then

$$e^A \cdot e^{-A} = e^{A+(-A)} = e^0 = I$$

□

We are now ready to prove Theorem 2.1.1.

*Proof of Theorem 2.1.1.* It is clear that

$$\vec{y}(t) = e^{tA}\vec{y}_0$$

solves the differential equation

$$\frac{d\vec{y}}{dt} = A\vec{y}$$

with the initial condition $\vec{y}(0) = \vec{y}_0$.

To show uniqueness, let $\tilde{\vec{y}}(t)$ be another solution that also satisfies $\tilde{\vec{y}}(0) = \vec{y}_0$. Then

$$\frac{d}{dt}\left(e^{-tA}\tilde{\vec{y}}(t)\right) = -e^{-tA}A\tilde{y} + e^{-tA}\frac{d}{dt}\tilde{y} = -e^{-tA}A\tilde{y} + e^{-tA}A\tilde{y} = 0$$

which implies that $e^{-tA\tilde{y}}$ is constant vector. At $t = 0$,

$$e^{-tA}\tilde{y}\Big|_{t=0} = \tilde{y}(0) = y_0,$$

from which we conclude $\tilde{y} = e^{tA}y_0$. This proves uniqueness. □

### 2.1.2 1st Order Inhomogeneous System

Let us move on to consider inhomogeneous linear equation with constant coefficients. The general form, after reduction to order one, is

$$\frac{d\vec{y}}{dt} = A \cdot \vec{y} + \vec{b}(t)$$

where $A$ is a constant $n \times n$ matrix and

$$\vec{b}(t) = \begin{pmatrix} b_1(t) \\ b_2(t) \\ \vdots \\ b_n(t) \end{pmatrix}$$

is a column vector that could vary with $t$.

We can use similar strategy to solve the above equation. Let us multiply $e^{-tA}$ on both sides. Then the equation becomes

$$\frac{d}{dt}(e^{-tA}\vec{y}) = e^{-tA}\vec{b}(t).$$

Integrating both sides, we find

$$e^{-tA}\vec{y} - \vec{y}_0 = \int_0^t e^{-sA}\vec{b}(s)ds.$$

Here $\vec{y}_0 = \vec{y}(0)$ is the initial value of $\vec{y}$ at $t = 0$. It follows that

$$\vec{y} = e^{tA}\vec{y}_0 + \int_0^t e^{(t-s)A}\vec{b}(s)ds.$$

This immediately leads to the following result

**Theorem 2.1.8.** *Given any column $\vec{y}_0 \in \mathbb{R}^n$, there exists a unique solution to the equation*

$$\frac{d\vec{y}}{dt} = A\vec{y} + \vec{b}(t)$$

*that satisfies the initial condition $\vec{y}(0) = \vec{y}_0$. The solution is explicitly given by*

$$\vec{y}(t) = e^{tA}\vec{y}_0 + \int_0^t e^{(t-s)A}\vec{b}(s)ds.$$

### 2.1.3 n-th Order Linear Equation

Let us now discuss how to solve an $n$-th order linear equation with constant coefficients. The equation is of the form

$$y^{(n)} + a_1 y^{(n-1)} + \cdots + a_{n-1} y' + a_n y = b(t)$$

where $a_i$'s are constants.

As we have discussed, this equation can be reduced to the 1st-order system

$$\frac{d\vec{y}}{dt} = A \cdot \vec{y} + \vec{b}(t)$$

where

$$\vec{y} = \begin{pmatrix} y \\ y_1 \\ y_2 \\ \vdots \\ y_{n-1} \end{pmatrix} \qquad A = \begin{pmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & & \\ & & & 1 & \\ & & & 0 & 1 \\ -a_n & -a_{n-1} & \cdots & -a_2 & -a_1 \end{pmatrix} \qquad \vec{b} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ b(t) \end{pmatrix}$$

**n-th order homogeneous equation**

Let us first focus on the homogeneous case when $b(t) = 0$. By Theorem 2.1.1, it can be shown (see Section 2.2.1) that solutions will be expressed as certain linear combination of functions of the form $t^m e^{\lambda_i t}$ where $\lambda_i$ is an eigenvalue of $A$. Instead, we will present a more direct way to show this using the method of characteristic polynomial.

The characteristic polynomial of $A$ is

$$\det(\lambda I - A) = \det \begin{pmatrix} \lambda & -1 & & & 0 \\ & \lambda & -1 & & \\ & & \ddots & \ddots & \\ 0 & & & \lambda & -1 \\ a_n & a_{n-1} & \cdots & a_2 & \lambda + a_1 \end{pmatrix} = \lambda^n + a_1 \lambda^{n-1} + a_2 \lambda^{n-2} + \cdots + a_n.$$

Eigenvalues of $A$ are roots of this polynomial.

**Definition 2.1.9.** The characteristic polynomial of the equation

$$y^{(n)} + a_1 y^{(n-1)} + a_2 y^{(n-2)} + \cdots + a_n y = 0$$

is defined to be

$$\lambda^n + a_1 \lambda^{n-1} + a_2 \lambda^{n-2} + \cdots + a_n = 0.$$

We explain how to use the characteristic polynomial to solve the equation. Let

$$P(\lambda) = \lambda^n + a_1 \lambda^{n-1} + a_2 \lambda^{n-2} + \cdots + a_n$$

denote the characteristic polynomial. It can be factorized as

$$P(\lambda) = (\lambda - \lambda_1)^{m_1}(\lambda - \lambda_2)^{m_2} \cdots (\lambda - \lambda_k)^{m_k}$$

where $\{\lambda_1, \lambda_2, \cdots, \lambda_k\}$ are all different roots of $P(\lambda)$ with multiplicity $m_1, m_2, \cdots, m_k$. The differential equation can be written as

$$P\left(\frac{d}{dt}\right) y = 0$$

or equivalently

$$\left(\frac{d}{dt} - \lambda_1\right)^{m_1} \left(\frac{d}{dt} - \lambda_2\right)^{m_2} \cdots \left(\frac{d}{dt} - \lambda_k\right)^{m_k} y = 0.$$

**Proposition 2.1.10.** *The following functions*

$$\{t^{i_1} e^{\lambda_1 t}\}_{0 \le i_1 < m_1}, \quad \{t^{i_2} e^{\lambda_2 t}\}_{0 \le i_2 < m_2}, \quad \cdots, \quad \{t^{i_k} e^{\lambda_k t}\}_{0 \le i_k < m_k}$$

*are $n$ linearly independent solutions to*

$$y^{(n)} + a_1 y^{(n-1)} + a_2 y^{(n-2)} + \cdots + a_n y = 0.$$

*Proof:* We check that the above functions are all solutions. Let us consider $t^{i_1} e^{\lambda_1 t}$ for $0 \le i_1 < m_1$. Since the differential operators $\frac{d}{dt} - \lambda_1, \cdots, \frac{d}{dt} - \lambda_k$ all commute with each other, we have

$$P\left(\frac{d}{dt}\right)\left(t^{i_1} e^{\lambda_1 t}\right) = \left(\frac{d}{dt} - \lambda_2\right)^{m_2} \cdots \left(\frac{d}{dt} - \lambda_k\right)^{m_k} \left(\frac{d}{dt} - \lambda_1\right)^{m_1} \left(t^{i_1} e^{\lambda_1 t}\right).$$

Thus it suffices to show $\left(\frac{d}{dt} - \lambda_1\right)^{m_1} \left(t^{i_1} e^{\lambda_1 t}\right) = 0$. Using

$$\left(\frac{d}{dt} - \lambda_1\right)(t f(t)) = t\left(\frac{d}{dt} - \lambda_1\right) f(t) + f(t)$$

and $i_1 < m_1$, we see that

$$\left(\frac{d}{dt} - \lambda_1\right)^{m_1} \left(t^{i_1} e^{\lambda_1 t}\right) = \sum_{j=0}^{i_1} \binom{j}{i_1} t^{i_1 - j} \left(\frac{d}{dt} - \lambda_1\right)^{m_1 - j} e^{\lambda_1 t} = 0$$

since $\left(\frac{d}{dt} - \lambda_1\right) e^{\lambda_1 t} = 0$. □

A general solution can be obtained via a linear combination of the above $n$ solutions. In the case when we have complex root $\alpha \pm i\beta$, with multiplicity $m$, the above proposition leads to

$$\{e^{(\alpha + i\beta)t}, t e^{(\alpha + i\beta)t}, \cdots, t^{m-1} e^{(\alpha + i\beta)t}\}$$

and

$$\{e^{(\alpha - i\beta)t}, t e^{(\alpha - i\beta)t}, \cdots, t^{m-1} e^{(\alpha - i\beta)t}\}.$$

If we want solutions expressed in terms of real functions, we can equivalently rewrite the above solutions via a different basis

$$\{\cos(\beta t) e^{\alpha t}, t\cos(\beta t) e^{\alpha t}, \cdots, t^{m-1}\cos(\beta t) e^{\alpha t}\}$$

and

$$\{\sin(\beta t)e^{\alpha t}, t\sin(\beta t)e^{\alpha t}, \cdots, t^{m-1}\sin(\beta t)e^{\alpha t}\}.$$

To uniquely determine a solution, we need to impose an initial condition, say at $t = 0$. Based on our general discussion on 1-st order system, the initial condition (at $t = 0$) is to specify the value of $\vec{y}$ at $t = 0$. Since

$$\vec{y} = \begin{pmatrix} y \\ y_1 \\ \vdots \\ y_{n-1} \end{pmatrix} = \begin{pmatrix} y \\ y' \\ \vdots \\ y^{(n-1)} \end{pmatrix},$$

the initial condition to be imposed is

$$y(0) = c_0, \qquad y'(0) = c_1, \qquad \cdots, \qquad y^{(n-1)}(0) = c_{n-1}$$

for some constants $c_i$. In other word, the initial condition for determining a solution of $n$-th order linear system is to specify the value of $y, y', \cdots, y^{(n-1)}$ at $t = 0$. This leads to $n$ equations that can be used to determine the $n$ linear coefficients for the $n$ solutions in Proposition 2.1.10.

**Example 2.1.11.** Solve the equation

$$y''' - 3y' + 2y = 0$$

with the initial condition $y(0) = 3$, $y'(0) = -4$, $y''(0) = 7$.

*Solution.* The characteristic polynomial is

$$\lambda^3 - 3\lambda + 2 = (\lambda - 1)^2(\lambda + 2).$$

By Proposition 2.1.10, the genenral solution is given by

$$a_1e^t + a_2te^t + a_3e^{-2t}.$$

The initial condition leads to linear equations for $a_i$'s

$$\begin{cases} a_1 + 0 + a_3 = 3 \\ a_1 + a_2 - 2a_3 = -4 \\ a_1 + 2a_2 + 4a_3 = 7 \end{cases}$$

which gives $a_1 = 1$, $a_2 = -1$, $a_3 = 2$. Hence the solution with required initial condition is

$$y(t) = (1 - t)e^t + 2e^{-2t}.$$

□

**Example 2.1.12.** Solve the equation

$$y'' - 2y' + 5y = 0$$

with the initial condition $y(0) = 2$, $y'(0) = 0$.

*Solution.* The characteristic polynomial

$$\lambda^2 - 2\lambda + 5$$

has complex roots $\lambda = 1 \pm 2i$. The general solution is

$$y = a_1 e^t \cos 2t + a_2 e^t \sin 2t.$$

The initial condition requires

$$\begin{cases} a_1 & & = & 2 \\ a_1 & + \quad 2a_2 & = & 0 \end{cases}$$

which gives $a_1 = 2$, $a_2 = -1$. The required solution is

$$y(t) = (2\cos 2t - \sin 2t)e^t.$$

□

**n-th order inhomogeneous equation**

Now we discuss the inhomogeneous case

$$y^{(n)} + a_1 y^{(n-1)} + \cdots + a_n y = b(t)$$

where $a_i$'s are constants. We first observe that if $y_1$ and $y_2$ are two solutions, then their difference $\tilde{y} = y_1 - y_2$ satisfies the homogeneous equation

$$\tilde{y}^{(n)} + a_1 \tilde{y}^{(n-1)} + \cdots + a_n \tilde{y} = 0.$$

Equivalently, let $u(t)$ be any special solution of the inhomogeneous equation, *i.e.* satisfying

$$u^{(n)} + a_1 u^{(n-1)} + \cdots + a_n u = b(t).$$

Then a general solution can be written as

$$y = u + \tilde{y}$$

where $\tilde{y}$ is a general solution to the homogeneous equation

$$\tilde{y}^{(n)} + a_1 \tilde{y}^{(n-1)} + \cdots + a_n \tilde{y} = 0.$$

As we have learned, such $\tilde{y}$ can be solved using the characteristic polynomial.

To find a special solution, let us write the equation in the 1st-order form

$$\frac{d\vec{y}}{dt} = A \cdot \vec{y} + \vec{b}(t)$$

where

$$\vec{y} = \begin{pmatrix} y \\ y_1 \\ y_2 \\ \vdots \\ y_{n-1} \end{pmatrix} \qquad A = \begin{pmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & & \\ & & & 1 & \\ & & & 0 & 1 \\ -a_n & -a_{n-1} & \cdots & -a_2 & -a_1 \end{pmatrix} \qquad \vec{b} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ b(t) \end{pmatrix}$$

By Theorem 2.1.8, we see that a special solution can be found by

$$\vec{u}(t) = \int_0^t e^{(t-s)A}\vec{b}(s)ds.$$

The matrix manipulation could be complicated in general. Nevertheless, it can be written as

$$u(t) = \int_0^t G(t-s)b(s)ds$$

for some function $G$. Such $G$ is an example of Green's function, which we will discuss in detail in Chapter 5. There is another method, called "variation of parameters", to write down a special solution once we know a basis of solutions for the homogeneous case. This method works for general linear system with varying coefficients, and will be discussed in Section 2.3.2.

In practice, we can guess a special solution via certain ansatz from the shape of the equation.

**Example 2.1.13.**

$$y'' + y = \cos 2t.$$

Let us consider the complex form of the above equation

$$z'' + z = e^{2it}, \qquad y = \mathrm{Re}(z).$$

It is natural to try the ansatz of the form

$$z(t) = ae^{2it}.$$

Plug into the equation, we find

$$(-3a)e^{2it} = e^{2it} \qquad \Rightarrow \qquad a = -\frac{1}{3}.$$

Thus we find a special solution $y = \mathrm{Re}\left(-\frac{1}{3}e^{2it}\right) = -\frac{1}{3}\cos 2t$.

**Example 2.1.14.**

$$y'' + y = \cos t$$

We again look at the complex equation

$$z'' + z = e^{it}.$$

Then we try the same ansatz as before

$$z = ae^{it}.$$

This time it does not work: if we plug into the above equation,

$$(ae^{it})'' + (ae^{it}) = 0.$$

The reason for the failure is that $i$ is a root of the characteristic polynomial. Thus we next try

$$z = ate^{it}.$$

This time we find $a = -\frac{i}{2}$ works. Thus

$$y = \frac{1}{2}\mathrm{Re}\left(-ite^{it}\right) = \frac{1}{2}t\sin t$$

is a special solution.

## 2.2 Long-term Behavior

We discuss examples of long-term behavior of solutions to homogeneous linear system with constant coefficients in order to gain some idea on the limiting behavior of solutions.

### 2.2.1 Jordan Canonical Form

We consider the behavior of solutions to

$$\frac{d\vec{y}}{dt} = A\vec{y}$$

when $t$ becomes large. The solution with initial condition $\vec{y}(0) = \vec{y}_0$ is given by

$$\vec{y}(t) = e^{tA}\vec{y}_0.$$

Therefore we need to understand $e^{tA}$ more explicitly. The key is the following

**Proposition 2.2.1.** *Let $A, B$ be two square matrices and $A = PBP^{-1}$. Then*

$$e^A = Pe^BP^{-1}.$$

*Proof:* This follows from

$$A^k = (PBP^{-1})^k = PB\underbrace{P^{-1}P}_{=1}B\underbrace{P^{-1}}\cdots PBP^{-1} = PB^kP^{-1}.$$

Then

$$e^A = \sum_{k\geq 0}\frac{A^k}{k!} = \sum_{k\geq 0}P\left(\frac{B^k}{k!}\right)P^{-1} = P\left(\sum_{k\geq 0}\frac{B^k}{k!}\right)P^{-1} = Pe^BP^{-1}.$$

□

Given any square matrix $A$, we can write it in terms of Jordan canonical form

$$A = P \begin{pmatrix} \begin{pmatrix} \lambda_1 & 1 & & 0 \\ & \lambda_1 & 1 & \\ & & \ddots & 1 \\ 0 & & & \lambda_1 \end{pmatrix} & & & 0 \\ & \begin{pmatrix} \lambda_2 & 1 & & 0 \\ & \lambda_2 & 1 & \\ & & \ddots & 1 \\ 0 & & & \lambda_2 \end{pmatrix} & & \\ & & \ddots & \\ 0 & & & \begin{pmatrix} \lambda_l & 1 & & 0 \\ & \lambda_l & 1 & \\ & & \ddots & 1 \\ 0 & & & \lambda_l \end{pmatrix} \end{pmatrix} P^{-1}$$

Here $\lambda_i$'s are eigenvalues of $A$ (could be complex numbers) and $P$ is an invertible matrix (complex valued in case of complex eigenvalues).

*Remark* 2.2.2. Before we move on, we simply remark that the exponential matrix works well the same for complex valued matrices, with the same defining power series

$$e^A = \sum_{k=0}^{\infty} \frac{A^k}{k!} \qquad A\text{: complex } n \times n \text{ matrix.}$$

This is very convenient for many applications.

Using the Jordan canonical form, we can write

$$A = P(D + N)P^{-1}$$

where $D = \begin{pmatrix} \lambda_1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & \lambda_1 & & & & & & \\ & & & \lambda_2 & & & & & \\ & & & & \ddots & & & & \\ & & & & & \lambda_2 & & & \\ & & & & & & \ddots & & \\ & & & & & & & \lambda_l & \\ & & & & & & & & \ddots \\ & & & & & & & & & \lambda_l \end{pmatrix}$ is the diagonal part of Jordan form,

and

$$N = \begin{pmatrix} \begin{pmatrix} 0 & 1 & & 0 \\ & 0 & 1 & \\ & & \ddots & 1 \\ 0 & & & 0 \end{pmatrix} & & & & 0 \\ & \begin{pmatrix} 0 & 1 & & 0 \\ & 0 & 1 & \\ & & \ddots & 1 \\ 0 & & & 0 \end{pmatrix} & & & \\ & & \ddots & & \\ & & & \begin{pmatrix} 0 & 1 & & 0 \\ & 0 & 1 & \\ & & \ddots & 1 \\ 0 & & & 0 \end{pmatrix} & \\ 0 & & & & \end{pmatrix}$$

is the off diagonal part. It is clear that $N$ is nilpotent

$$N^n = 0, \qquad n\text{: size of } A$$

and $D, N$ commute

$$DN = ND.$$

It follows that

$$e^{tA} = Pe^{t(D+N)}P^{-1} = Pe^{tD}e^{tN}P^{-1}.$$

Let us consider the two terms $e^{tD}$ and $e^{tN}$ in the middle. Since $D$ is diagonal, $e^{tD}$ is simply

$$e^{tD} = \begin{pmatrix} e^{t\lambda_1} & & & & & & & & \\ & \ddots & & & & & & & \\ & & e^{t\lambda_1} & & & & & & \\ & & & e^{t\lambda_2} & & & & & \\ & & & & \ddots & & & & \\ & & & & & e^{t\lambda_2} & & & \\ & & & & & & \ddots & & \\ & & & & & & & e^{t\lambda_l} & \\ & & & & & & & & \ddots \\ & & & & & & & & & e^{t\lambda_l} \end{pmatrix}$$

For the nilpotent matrix $N$, we have

$$N^p = 0$$

where $p$ is the maximal size of the Jordan blocks. Nevertheless, $N^n = 0$ always hold where $n$ is the size of $A$. It follows that

$$e^{tN} = \sum_{i=0}^{n} \frac{N^i}{i!} t^i$$

which is a polynomial in $t$. This allows for explicit calculation of $e^{tD}e^{tN}$, hence for $e^{tA}$.

In the case when $A$ is diagonalizable, *i.e.* $N = 0$, we have

$$A = P \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix} P^{-1}.$$

Then

$$e^{tA} = P \begin{pmatrix} e^{t\lambda_1} & & & \\ & e^{t\lambda_2} & & \\ & & \ddots & \\ & & & e^{t\lambda_n} \end{pmatrix} P^{-1}.$$

As an application, we have the following result

**Proposition 2.2.3.** *Assume all eigenvalues of $A$ have negative real part. Then for any solution $\vec{y}(t)$ of*

$$\frac{d\vec{y}}{dt} = A\vec{y}$$

*we have*

$$\lim_{t\to+\infty} \vec{y}(t) = 0.$$

*In other words, all solutions will go to the origin eventually. In this case, the origin is called a sink or attractor.*

*Proof:* We can write $\vec{y}$ as

$$\vec{y}(t) = e^{tA}\vec{y}_0.$$

Using the Jordan decomposition as above, we have

$$e^{tA} = Pe^{tD}\left(1 + tN + \cdots + \frac{t^n N^n}{n!}\right) P^{-1}.$$

Here

$$e^{tD} = \begin{pmatrix} e^{t\lambda_1} & & & \\ & e^{t\lambda_2} & & \\ & & \ddots & \\ & & & e^{t\lambda_n} \end{pmatrix}$$

where $\lambda_1, \lambda_2, \cdots, \lambda_n$ are all eigenvalues (could repeat) of $A$. By assumption,

$$\mathrm{Re}(\lambda_i) < 0 \qquad \text{for all } \lambda_i.$$

So $e^{t\lambda_i}$'s exponentially decay to zero when $t \to +\infty$. Since

$$\lim_{t\to+\infty} e^{-at}t^m = 0 \qquad \text{for all } a > 0,\ m \in \mathbb{Z},$$

it follows that

$$\lim_{t\to+\infty} e^{tD}\left(1+tN+\cdots+\frac{t^nN^n}{n!}\right)=0.$$

Thus

$$\lim_{t\to+\infty} e^{tA}=0, \qquad \text{hence} \qquad \lim_{t\to+\infty} e^{tA}\vec{y}_0=0.$$

□

In general, when eigenvalues of $A$ have both positive and negative real parts, things will be more complicated and the initial condition will play an important role. Before we move to this, let us discuss the real Jordan canonical form.

Recall that if we want to stay in the realm of real matrices, then $A$ can be put into the real Jordan canonical form in terms of real matrix $P$

$$A = P \begin{pmatrix} \square & & & \\ & \square & & \\ & & \ddots & \\ & & & \square \end{pmatrix} P^{-1}$$

Here each Jordan block $\square$ is of the form

- for real eigenvalue $\lambda$ of $A$,

$$\square = \begin{pmatrix} \lambda & 1 & & O \\ & \lambda & \ddots & \\ & & \ddots & 1 \\ O & & & \lambda \end{pmatrix}$$

- for complex eigenvalue $\lambda=\alpha\pm i\beta$,

$$\square = \begin{pmatrix} \begin{pmatrix}\alpha & -\beta\\ \beta & \alpha\end{pmatrix} & \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix} & & 0 \\ & \begin{pmatrix}\alpha & -\beta\\ \beta & \alpha\end{pmatrix} & \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix} & \\ & & \ddots & \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix} \\ 0 & & & \begin{pmatrix}\alpha & -\beta\\ \beta & \alpha\end{pmatrix} \end{pmatrix}$$

In particular, if $A$ is diagonalizable, then there exists an invertible real matrix $P$ such that

$$A = P \begin{pmatrix} \lambda_1 & & & & & & & & 0 \\ & \lambda_2 & & & & & & & \\ & & \ddots & & & & & & \\ & & & \lambda_k & & & & & \\ & & & & \begin{pmatrix} \alpha_1 & -\beta_1 \\ \beta_1 & \alpha_1 \end{pmatrix} & & & & \\ & & & & & \begin{pmatrix} \alpha_2 & -\beta_2 \\ \beta_2 & \alpha_2 \end{pmatrix} & & & \\ & & & & & & & & \ddots \\ 0 & & & & & & & & \end{pmatrix} P^{-1}$$

### 2.2.2 Examples of Two-dim System

To illustrate the main phenomenon, let us focus on the two-dim system

$$\frac{d\vec{y}}{dt} = A\vec{y}$$

where $\vec{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$ and $A$ is a $2 \times 2$ matrix. Let

$$A = PJP^{-1}$$

where $J$ is the real Jordan canonical form. We can use a linear transformation to redefine

$$\vec{\hat{y}} = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \end{pmatrix} = P^{-1} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

Then the system becomes

$$\frac{d\vec{\hat{y}}}{dt} = J\vec{\hat{y}}.$$

So without loss of generality, we can assume $A$ is in real Jordan canonical form from the start.

Note that $\vec{y} = 0$ is always a solution, and so the origin is called an equilibrium point. We discuss below in detail for cases when $A$ is diagonalizable.

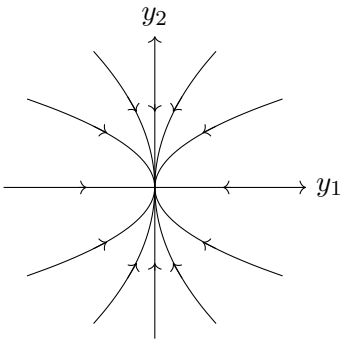
• Case I : $A$ is diagonal with real eigenvalues

$$A = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}, \qquad \lambda_i \in \mathbb{R}.$$

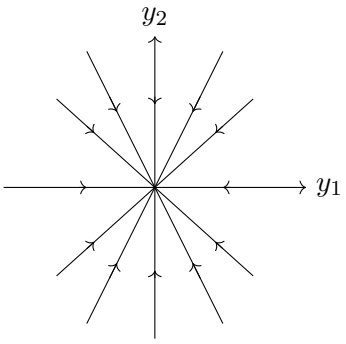
Then solutions take the form

$$\begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix} = \begin{pmatrix} y_1(0)e^{\lambda_1 t} \\ y_2(0)e^{\lambda_2 t} \end{pmatrix}.$$

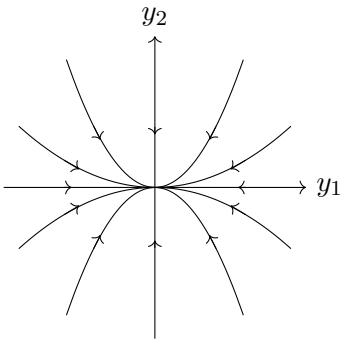
Let us assume both $\lambda_i \neq 0$, so both $y_i$'s can flow. We draw the solutions on the $y_1 - y_2$ plane with arrow pointing to the $t$ increasing direction. Then we have the following situations




$\lambda_1 < \lambda_2 < 0$

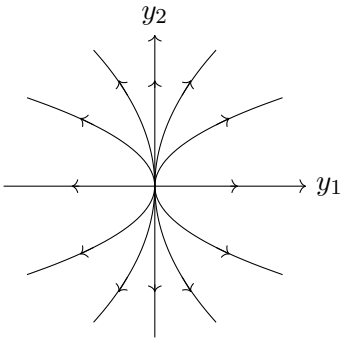



$\lambda_1 = \lambda_2 < 0$

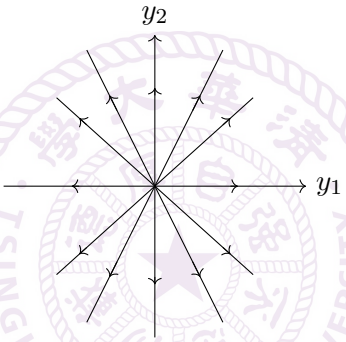



$\lambda_2 < \lambda_1 < 0$

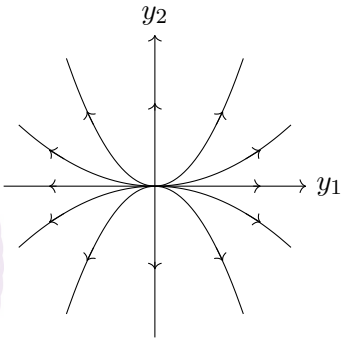
In the above cases, the origin is called the "sink": all solutions will flow eventually to the origin at $t \to +\infty$.




$\lambda_1 > \lambda_2 > 0$

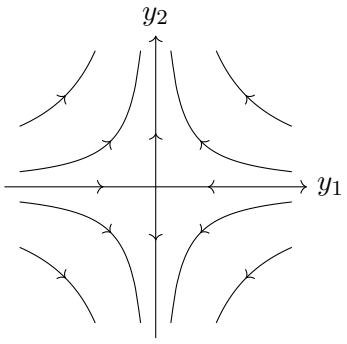



$\lambda_1 = \lambda_2 > 0$

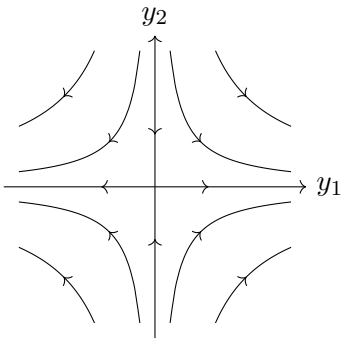



$\lambda_2 > \lambda_1 > 0$

In the above cases, the origin is called the "source": all solutions will back-flow to the origin at $t \to -\infty$.




$\lambda_1 < 0 < \lambda_2$




$\lambda_2 < 0 < \lambda_1$

In the above cases, the origin is called the "saddle". For example, in the case $\lambda_1 < 0 < \lambda_2$, the solution

$$\begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix} = \begin{pmatrix} y_1(0)e^{\lambda_1 t} \\ y_2(0)e^{\lambda_2 t} \end{pmatrix}$$

- will flow to the origin along $y_1$-axis when $y_2(0) = 0$,
- will flow from the origin along $y_2$-axis when $y_1(0) = 0$,
- when $y_1(0) \neq 0$, $y_2(0) \neq 0$,

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \to \begin{pmatrix} 0 \\ \pm\infty \end{pmatrix} \qquad \text{when} \qquad t \to +\infty$$

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \to \begin{pmatrix} \pm\infty \\ 0 \end{pmatrix} \qquad \text{when} \qquad t \to -\infty$$

• Case II : $A$ has complex eigenvalues $\lambda = \alpha + i\beta$,

$$A = \begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix}$$

Let us collect $y_1, y_2$ into a complex function

$$z = y_1 + iy_2.$$

Then the matrix equation

$$\frac{d\vec{y}}{dt} = A\vec{y}$$

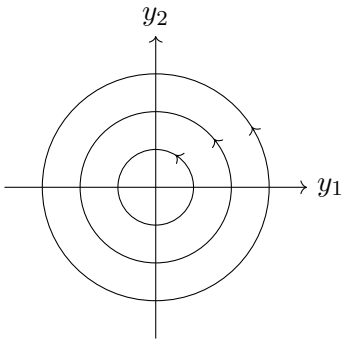
becomes a single equation for complex valued equation

$$\frac{dz}{dt} = \lambda z$$

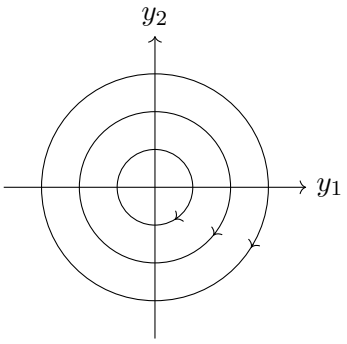
which is easily solved by

$$z(t) = e^{t\lambda} z(0).$$

Viewing $(y_1, y_2)$ as the complex plane, we have the following cases

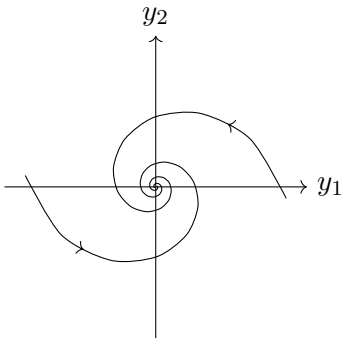



$\alpha = 0$, $\beta > 0$

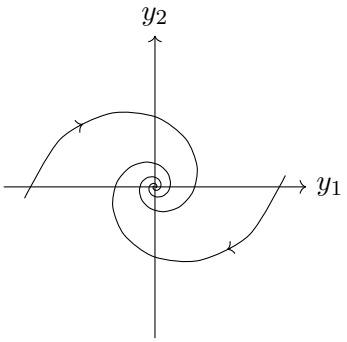



$\alpha = 0$, $\beta < 0$

In the above cases, the equilibrium of the origin is called the "center": the trajectories of other solutions are concentric circles.

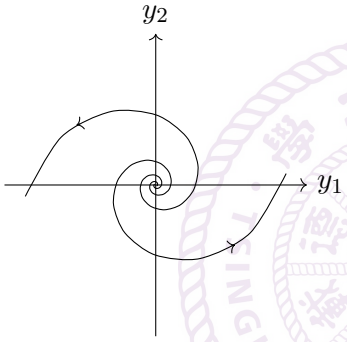



$\alpha < 0, \beta > 0$

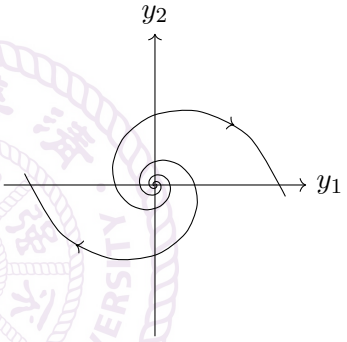



$\alpha < 0, \beta < 0$

In the above cases, the equilibrium of the origin is called "stable focus".




$\alpha > 0, \beta > 0$




$\alpha > 0, \beta < 0$

In the above cases, the equilibrium of the origin is called "unstable focus".

## 2.3 Linear Systems with Varying Coefficients

### 2.3.1 Path-ordered Exponential

We now discuss nonautonomous linear system (linear systemsswith varying coefficients)

$$\frac{d\vec{y}}{dt} = A(t)\vec{y} + \vec{b}(t)$$

where the matrix $A(t)$ is no longer constant, but could vary continuously with $t$.

Let us start with the homogeneous case

$$\frac{d\vec{y}}{dt} = A(t)\vec{y}.$$

Since $A(t)$ could vary with $t$, the exponential $e^{tA}$ no longer works

$$\frac{d}{dt}(e^{tA(t)}) \neq A(t)e^{tA(t)}.$$

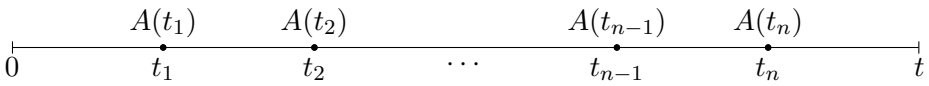
Nevertheless, we can still find an analogue of the exponential that works for this case.

Let us first consider the following expression

$$P_n(t) = \int_0^t dt_n \int_0^{t_n} dt_{n-1} \cdots \int_0^{t_2} dt_1 A(t_n)A(t_{n-1}) \cdots A(t_1).$$

Equivalently, we can write

$$P_n(t) = \int_{0 \le t_1 \le t_2 \le \cdots \le t_n \le t} dt_1 \cdots dt_n A(t_n)A(t_{n-1}) \cdots A(t_1).$$




The operation of matrix multiplication by $P_n(t)$ can be viewed as applying $A(t_1)$ first at time $t_1$, then applying $A(t_2)$ at a later time $t_2$, etc, until finally applying $A(t_n)$ at last. This application of matrix multiplication is ordered in time $t$, and integrated over all such ordered configurations. Observe that

$$\frac{d}{dt}P_n(t) = A(t)P_{n-1}(t).$$

**Definition 2.3.1.** We define the path-ordered exponential

$$\mathcal{P}\left(e^{\int_0^t A}\right) := \sum_{n=0}^{\infty} \int_0^t dt_n \int_0^{t_n} dt_{n-1} \cdots \int_0^{t_2} dt_1 A(t_n) \cdots A(t_1)$$

In terms of the notation above, this is

$$\mathcal{P}\left(e^{\int_0^t A}\right) = \sum_{n=0}^{\infty} P_n(t).$$

To see the convergence of this series, assume

$$\|A(s)\| \le M, \quad \forall \quad 0 \le s \le t.$$

Then

$$\|P_n(t)\| \le \int_{0 \le t_1 \le t_2 \le \cdots \le t_n \le t} dt_1 \cdots dt_n \|A(t_n)\| \cdots \|A(t_1))\| \le \frac{t^n M^n}{n!}.$$

This implies the convergence of the series sum for the path-ordered exponential.

**Proposition 2.3.2.** *The path-ordered exponential satisfies*

① *If $A(t) = A$ is a constant matrix, then*

$$\mathcal{P}\left(e^{\int_0^t A}\right) = e^{tA}.$$

② $\mathcal{P}\left(e^{\int_0^t A}\right)\Big|_{t=0} = 1$

③ *The following differential equation holds*

$$\frac{d}{dt}\mathcal{P}\left(e^{\int_0^t A}\right) = A(t)\mathcal{P}\left(e^{\int_0^t A}\right).$$

④ *The path-ordered exponential* $\mathcal{P}\left(e^{\int_0^t A}\right)$ *is invertible. Its inverse is given by the sum (note for reversed ordered in time)*

$$\mathcal{P}\left(e^{\int_0^t A}\right)^{-1} = \sum_{n=0}^{\infty}(-1)^n \int_0^t dt_n \int_0^{t_n} dt_{n-1} \cdots \int_0^{t_2} dt_1 A(t_1)\cdots A(t_n).$$

*Proof:* ① When $A(t) = A$ is a constant matrix,

$$P_n(t) = \int_{0\le t_1\le\cdots\le t_n\le t} dt_1\cdots dt_n A^n = \frac{t^n}{n!}A^n$$

$$\Rightarrow \mathcal{P}(e^{\int_0^t A}) = \sum_{n=0}^{\infty}\frac{t^n}{n!}A^n = e^{tA}.$$

② is clear.

③ follows from $\frac{d}{dt}P_n(t) = A(t)P_{n-1}(t)$. Then

$$\frac{d}{dt}\mathcal{P}\left(e^{\int_0^t A}\right) = \frac{d}{dt}\sum_{n\ge 0}P_n(t) = \sum_{n\ge 1}\frac{d}{dt}P_n(t) = \quad A(t)\sum_{n\ge 1}P_{n-1}(t) = A(t)\mathcal{P}\left(e^{\int_0^t A}\right).$$

④ Let $\mathcal{P}(t) = \mathcal{P}\left(e^{\int_0^t A}\right)$ and let

$$\mathcal{Q}(t) = \sum_{n=0}^{\infty}(-1)^n \int_0^t dt_n \int_0^{t_n} dt_{n-1} \cdots \int_0^{t_2} dt_1 A(t_1)\cdots A(t_n).$$

By the same calculation, taking care of the reversed order, $\mathcal{Q}(t)$ satisfies the equation

$$\frac{d\mathcal{Q}(t)}{dt} = -\mathcal{Q}(t)A(t), \quad \mathcal{Q}(0) = 1.$$

Then

$$\frac{d}{dt}(\mathcal{Q}(t)\mathcal{P}(t)) = -(\mathcal{Q}(t)A(t))\mathcal{P}(t) + \mathcal{Q}(t)(A(t)\mathcal{P}(t)) = 0.$$

Thus $\mathcal{Q}(t)\mathcal{P}(t)$ is independent of $t$. Since $\mathcal{Q}(0)\mathcal{P}(0) = 1$, we have $\mathcal{Q}(t) = \mathcal{P}(t)^{-1}$ for all $t$.

□

This proposition says that the path-ordered exponential can indeed be viewed as a generalization of the exponential $e^{tA}$ to the case when $A$ depends on $t$.

**Theorem 2.3.3.** *Given any column vector* $\vec{y}_0 \in \mathbb{R}^n$*, there exists a unique solution to the equation*

$$\frac{d}{dt}\vec{y} = A(t)\vec{y} + \vec{b}(t)$$

*that satisfies the initial condition* $\vec{y}(0) = \vec{y}_0$*. The solution is explicitly given by*

$$\vec{y} = \mathcal{P}(t)\vec{y}_0 + \mathcal{P}(t)\int_0^t \mathcal{P}^{-1}(s)\vec{b}(s)ds.$$

*Here* $\mathcal{P}(t) = \mathcal{P}\left(e^{\int_0^t A}\right)$.

*Proof:* By Proposition 2.3.2

$$\frac{d}{dt}\mathcal{P}(t) = A(t)\mathcal{P}(t), \quad \frac{d}{dt}\mathcal{P}(t)^{-1} = -\mathcal{P}(t)^{-1}A(t).$$

We can solve the above equation via the same strategy as before. Multiplying $\mathcal{P}^{-1}(t)$,

$$\begin{aligned} \Rightarrow \quad & \frac{d}{dt}(\mathcal{P}^{-1}(t)\vec{y}) = \mathcal{P}^{-1}(t)\vec{b}(t) \\ \Rightarrow \quad & \mathcal{P}^{-1}(t)\vec{y} = \vec{y_0} + \int_0^t \mathcal{P}^{-1}(s)\vec{b}(s)ds \\ \Rightarrow \quad & \vec{y} = \mathcal{P}(t)\vec{y_0} + \mathcal{P}(t)\int_0^t \mathcal{P}^{-1}(s)\vec{b}(s)ds. \end{aligned}$$

Here $\vec{y_0} = \vec{y}\big|_{t=0}$ is the initial value of $\vec{y}$ at $t = 0$. □

The long-term behavior of this solution will become complicated in general, and depend on how $A(t)$ varies with $t$.

*Remark* 2.3.4. If we write $\mathcal{P}(t)$ as $n$ column vectors

$$\mathcal{P}(t) = \begin{pmatrix} \mathbf{y}_1(t) & \mathbf{y}_2(t) & \cdots & \mathbf{y}_n(t) \end{pmatrix},$$

then $\{\mathbf{y}_1, \cdots, \mathbf{y}_n\}$ form $n$ linearly independent solutions of the homogeneous equation

$$\frac{d\mathbf{y}}{dt} = A(t)\mathbf{y}.$$

The linear independency follows from the invertibility of $\mathcal{P}(t)$. A basis of $n$ linearly independent solutions is also called a fundamental solution.

### 2.3.2 Variation of Parameters

Let us discuss the case for $n$-th order linear equation

$$y^{(n)} + a_1(t)y^{(n-1)} + \cdots + a_{n-1}(t)y' + a_n(t)y = b(t)$$

when the coefficients are no longer constants. Again, the general solution is given by a sum of a special solution and a general solution to the homogeneous equation

$$\tilde{y}^{(n)} + a_1(t)\tilde{y}^{(n-1)} + \cdots + a_n(t)\tilde{y} = 0.$$

We can obtain a general solution to the homogeneous equation, say using the path-ordered exponential. Here we explain a trick to find a special solution, called the method of variation of parameters.

Let $\tilde{y}_1(t), \cdots, \tilde{y}_n(t)$ be $n$ linearly independent solutions to the homogeneous equation

$$\tilde{y}^{(n)} + a_1(t)\tilde{y}^{(n-1)} + \cdots + a_n(t)\tilde{y} = 0.$$

The method of variation of parameters is to seek for a special solution of the inhomogeneous equation in terms of a form

$$y(t) = c_1(t)\tilde{y}_1(t) + \cdots + c_n(t)\tilde{y}_n(t)$$

where $c_i(t)$'s are functions of $t$ that are assumed to satisfy the conditions

$$\sum_{i=1}^{n} c_i'(t)\tilde{y}_i^{(m)}(t) = 0, \qquad \text{for} \quad 0 \le m \le n-2.$$

Assume this holds. Then

$$y^{(m)}(t) = \sum_{i=1}^{n} c_i(t)\tilde{y}_i^{(m)}(t), \qquad 0 \le m \le n-1$$

$$y^{(n)}(t) = \sum_{i=1}^{n} c_i(t)\tilde{y}_i^{(n)}(t) + \sum_{i=1}^{n} c_i'(t)\tilde{y}_i^{(n-1)}(t).$$

Substituting this into the original equation, we find

$$\sum_{i=1}^{n} c_i'(t)\tilde{y}_i^{(n-1)}(t) = b(t).$$

Thus a special solution can be obtained by finding $c_i(t)$'s satisfying

$$\begin{cases} \sum_{i=1}^{n} c_i'(t)\tilde{y}_i^{(m)}(t) = 0, & 0 \le m \le n-2 \\ \sum_{i=1}^{n} c_i'(t)\tilde{y}_i^{(n-1)}(t) = b(t) \end{cases} \tag{$*$}$$

$(*)$ can be written in matrix form as

$$\begin{pmatrix} \tilde{y}_1 & \cdots & \tilde{y}_n \\ \tilde{y}_1' & \cdots & \tilde{y}_n' \\ \vdots & & \vdots \\ \tilde{y}_1^{(n-1)} & \cdots & \tilde{y}_n^{(n-1)} \end{pmatrix} \begin{pmatrix} c_1' \\ c_2' \\ \vdots \\ c_n' \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ b(t) \end{pmatrix}$$

This can be solved using Cramer's rule by

$$c_i'(t) = \frac{W_i(t)}{W(t)}$$

Here

$$W(t) = \det \begin{pmatrix} \tilde{y}_1 & \cdots & \tilde{y}_n \\ \tilde{y}_1' & \cdots & \tilde{y}_n' \\ \vdots & & \vdots \\ \tilde{y}_1^{(n-1)} & \cdots & \tilde{y}_n^{(n-1)} \end{pmatrix}$$

and

$$W_i(t) = \det \begin{pmatrix} \tilde{y}_1 & \cdots & \tilde{y}_{i-1} & 0 & \tilde{y}_{i+1} & \cdots & \tilde{y}_n \\ \tilde{y}_1' & \cdots & \tilde{y}_{i-1}' & 0 & \tilde{y}_{i+1}' & \cdots & \tilde{y}_n' \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ \tilde{y}_1^{(n-1)} & \cdots & \tilde{y}_{i-1}^{(n-1)} & b(t) & \tilde{y}_{i+1}^{(n-1)} & \cdots & \tilde{y}_n^{(n-1)} \end{pmatrix}$$

*Remark* 2.3.5. $W(t)$ is called the Wronskian determinant of the functions $\tilde{y}_1, \tilde{y}_2, \cdots, \tilde{y}_n$.

Therefore a special solution of the inhomogeneous equation can be found by

$$\sum_{i=1}^{n} \tilde{y}_i(t) \int^t \frac{W_i(s)}{W(s)} ds.$$

**Example 2.3.6.** Consider the inhomogeneous equation

$$y'' + y = f(t).$$

This equation has constant coefficients, so the general solution to the homogeneous equation can be found from the characteristic polynomial

$$\lambda^2 + 1 = 0 \qquad \Rightarrow \qquad \lambda = \pm i$$

We obtain two independent solutions to the homogeneous equation $\tilde{y}'' + \tilde{y} = 0$ by

$$\tilde{y}_1 = \cos t, \qquad \tilde{y}_2 = \sin t.$$

The Wronskian of them is

$$W = \det \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix} = 1$$

and

$$\begin{cases} W_1 = \det \begin{pmatrix} 0 & \sin t \\ f(t) & \cos t \end{pmatrix} = -\sin t f(t) \\ W_2 = \det \begin{pmatrix} \cos t & 0 \\ -\sin t & f \end{pmatrix} = \cos t f(t) \end{cases}$$

The method of variation of parameters leads to

$$\begin{cases} c_1(t) = \displaystyle\int_0^t \frac{W_1}{W} = -\int_0^t \sin(s) f(s) ds \\ c_2(t) = \displaystyle\int_0^t \frac{W_2}{W} = \int_0^t \cos(s) f(s) ds \end{cases}$$

A special solution is found by

$$y(t) = c_1(t)\tilde{y}_1(t) + c_2(t)\tilde{y}_2(t) = \int_0^t \sin(t-s) f(s) ds.$$

**Example 2.3.7.** Consider the equation

$$ty'' - (t+1)y' + y = t^2.$$

We divide both sides by $t$ to arrive at the standard form

$$y'' - \frac{t+1}{t} y' + \frac{1}{t} y = t.$$

Two independent solutions of the homogeneous equations are

$$\tilde{y}_1(t) = e^t, \qquad \tilde{y}_2(t) = t + 1.$$

Their Wronskian is

$$W = \det \begin{pmatrix} e^t & t+1 \\ e^t & 1 \end{pmatrix} = -te^t$$

and

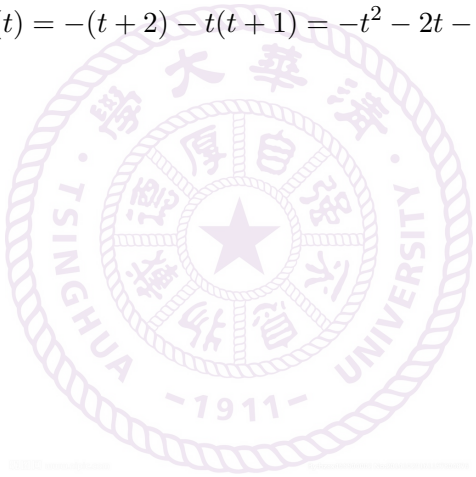
$$\begin{cases} W_1 = \det \begin{pmatrix} 0 & t+1 \\ t & 1 \end{pmatrix} = -t^2 - t \\ W_2 = \det \begin{pmatrix} e^t & 0 \\ e^t & t \end{pmatrix} = te^t. \end{cases}$$

Therefore we can choose

$$c_1(t) = \int_{-2}^{t} \frac{W_1}{W} = \int_{-2}^{t} (s+1)e^{-s} ds = -(t+2)e^{-t}$$
$$c_2(t) = \int_{0}^{t} \frac{W_2}{W} = \int_{0}^{t} (-1) ds = -t.$$

A special solution is found by

$$y(t) = -(t+2) - t(t+1) = -t^2 - 2t - 2.$$



# Chapter 3 Initial Value Problem

We move on to study non-linear differential equations. By reduction to first order, a general form of non-linear equations can be expressed by a system

$$\frac{d\mathbf{y}}{dt} = \mathbf{F}(\mathbf{y}, t)$$

where $\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ is the column of unknown functions, and

$$\mathbf{F} = \begin{pmatrix} F_1(\mathbf{y}, t) \\ F_2(\mathbf{y}, t) \\ \vdots \\ F_n(\mathbf{y}, t) \end{pmatrix} : U \to \mathbb{R}^n$$

is a function from some open subset $U \subset \mathbb{R}^n \times \mathbb{R}$ to $\mathbb{R}^n$.

In this chapter, we will focus on the initial value problem which amounts to solve

$$\frac{d\mathbf{y}}{dt} = \mathbf{F}(\mathbf{y}, t), \qquad \mathbf{y}(t_0) = \boldsymbol{\xi} \in \mathbb{R}^n$$

where the initial value of $\mathbf{y}$ at a specified time $t_0$ is chosen. Another important situation is the boundary value problem, which we will study in Chapter 5.

## 3.1 Local Solutions

Solutions of non-linear system may blow up in a finite time $t$ even though all coefficients of $\mathbf{F}$ behave very well. For example, consider the following non-linear equation

$$\frac{dy}{dt} = y^2.$$

One solution is $y = \frac{1}{1-t}$, which blows up when $t$ goes from $t = 0$ to $t = 1$.

Nevertheless, local solutions with specified initial condition are guaranteed for a large class of non-linear systems. We will establish such local theory in this subsection.

### 3.1.1 Integral Equation

We will always assume $\mathbf{F}$ is continuous. The nonlinear system with initial condition

$$\frac{d\mathbf{y}}{dt} = \mathbf{F}(\mathbf{y}, t), \qquad \mathbf{y}(t_0) = \boldsymbol{\xi} \in \mathbb{R}^n$$

can be equivalently formulated as the integral equation

$$\mathbf{y}(t) = \boldsymbol{\xi} + \int_{t_0}^{t} \mathbf{F}(\mathbf{y}(s), s)ds.$$

Let us introduce the operator $T$ on the space of column functions $\mathbf{u}(t) = \begin{pmatrix} u_1(t) \\ \vdots \\ u_n(t) \end{pmatrix}$ in $t$ by

$$(T\mathbf{u})(t) = \boldsymbol{\xi} + \int_{t_0}^{t} \mathbf{F}(\mathbf{u}(s), s)ds.$$

Then the integral equation can be expressed as

$$\mathbf{y} = T\mathbf{y}.$$

In other word, $\mathbf{y}(t)$ is a fixed point of the operator $T$. It reduces the problem on solving the equation to the study of fixed points of $T$. This can be analyzed in terms of the Banach Fixed-point Theorem, aslo known as the Contraction Mapping Theorem.

### 3.1.2 The Contraction Mapping Theorem

Let $(X, d)$ be a metric space. A mapping

$$T : X \to X$$

is called a contraction mapping if there exists a constant $\lambda$ with $0 \leq \lambda < 1$ such that

$$d(T(x), T(y)) \leq \lambda d(x, y), \qquad \forall x, y \in X.$$

Here $d(x, y)$ is the distance between $x$ and $y$. Thus a contraction mapping brings points closer. Note that $T$ is necessarily continuous.

In the next, we will assume $(X, d)$ is a complete metric space, *i.e.* , every Cauchy sequence in $X$ has a limit. When $X$ is a vector space, and $d(x, y) = \|x - y\|$ comes from a norm $\|\cdot\|$, a complete normed linear space is also called a Banach space.

A point $x \in X$ is called a fixed point of $T$ if

$$Tx = x.$$

In other words, $x$ will remain the same after the mapping by $T$.

**Theorem 3.1.1** (Contraction Mapping)**.** *Let $T : X \to X$ be a contraction mapping on a complete metric space $(X, d)$. Then there exists a unique point $x \in X$ such that*

$$Tx = x,$$

*i.e., $T$ has a unique fixed point.*

*Proof:* Let $x_0$ be any point in $X$. We define a sequence $\{x_n\}$ in $X$ by

$$x_{n+1} = Tx_n, \qquad \text{for } n \geq 0.$$

Equivalently, $x_n = T^n x_0$ is the $n$th iteration of $T$ on $x_0$.

For any $n > m \geq 1$, the contraction property implies

$$\begin{aligned} d(x_n, x_m) \leq & d(x_n, x_{n-1}) + d(x_{n-1}, x_{n-2}) + \cdots + d(x_{m+1}, x_m) \\ = & d(T^{n-1}x_1, T^{n-1}x_0) + d(T^{n-2}x_1, T^{n-2}x_0) + \cdots + d(T^m x_1, T^m x_0) \\ \leq & (\lambda^{n-1} + \lambda^{n-2} + \cdots + \lambda^m) d(x_1, x_0) \\ \leq & \frac{\lambda^m}{1-\lambda} d(x_1, x_0). \end{aligned}$$

Thus $\{x_n\}$ is a Cauchy sequence. Since $X$ is complete, $\{x_n\}$ converges to a point $x \in X$. Then

$$Tx = T \lim_{n\to\infty} x_n = \lim_{n\to\infty} Tx_n = \lim_{n\to\infty} x_{n+1} = x$$

*i.e.* , $x$ is a fixed point of $T$. This proves the existence.

Assume $x$ and $y$ are two fixed points of $T$. Then

$$0 \leq d(x, y) = d(Tx, Ty) \leq \lambda d(x, y)$$

$$\Rightarrow d(x, y) = 0.$$

So $x = y$. This proves the uniqueness. □

### 3.1.3 Lipschitz Condition

To apply the Contraction Mapping Theorem to obtain a solution of the integral equation, we need some control on $\mathbf{F}$. Let $U$ be an open subset of $\mathbb{R}^n \times \mathbb{R}$ and

$$\mathbf{F}(\mathbf{y}, t) : U \mapsto \mathbb{R}^m$$

be a continuous function, where $\mathbf{y} \in \mathbb{R}^n$ and $t \in \mathbb{R}$. We say $\mathbf{F}$ satisfies the Lipschitz condition in $U$ with respect to $\mathbf{y}$ if there exists a constant $L \geq 0$ such that

$$|\mathbf{F}(\mathbf{y}_1, t) - \mathbf{F}(\mathbf{y}_2, t)| \leq L|\mathbf{y}_1 - \mathbf{y}_2|$$

for all $(\mathbf{y}_1, t), (\mathbf{y}_2, t) \in U$. Here $|\cdot|$ for a vector is the Euclidean norm.

A typical example is when $U$ is convex, $\mathbf{F}(\mathbf{y}, t)$ is $C^1$ in $\mathbf{y}$, and all partial derivatives

$$|\partial_{y^i} \mathbf{F}| \leq M, \qquad \mathbf{y} = \{y^1, \cdots, y^n\}$$

are uniformly bounded. Then $\mathbf{F}(\mathbf{y}, t)$ is Lipschitz in $U$ with respect to $\mathbf{y}$. In fact, using

$$\begin{aligned}
&\mathbf{F}(\mathbf{y}_1, t) - \mathbf{F}(\mathbf{y}_2, t) \\
=&\int_0^1 \frac{\partial}{\partial s}\mathbf{F}(\mathbf{y}_2 + s(\mathbf{y}_1 - \mathbf{y}_2), t)ds \\
=&\int_0^1 ((\mathbf{y}_1 - \mathbf{y}_2)\cdot\nabla)\mathbf{F}(\mathbf{y}_2 + s(\mathbf{y}_1 - \mathbf{y}_2), t)ds
\end{aligned}$$

we find

$$|\mathbf{F}(\mathbf{y}_1, t) - \mathbf{F}(\mathbf{y}_2, t)| \leq M|\mathbf{y}_1 - \mathbf{y}_2|$$

*i.e.* , the Lipschitz condition holds with $L = M$.

### 3.1.4 Existence and Uniqueness

We now analyze the contraction property of the operator $T$ arising from the integral equation with the help of the Lipschitz condition. Let

$$C([t_0 - \varepsilon, t_0 + \varepsilon], \mathbb{R}^n)$$

be the space of continuous functions from the closed interval $[t_0 - \varepsilon, t_0 + \varepsilon]$ to $\mathbb{R}^n$. We equip it with the norm $\|\cdot\|_\infty$ defined by

$$\|\mathbf{x}(t)\|_\infty = \max_{t_0-\varepsilon\leq t\leq t_0+\varepsilon} |\mathbf{x}(t)|$$

The norm $\|\cdot\|_\infty$ is complete. This follows from the fact that the uniform limit of a sequence of continuous function is itself continuous. Therefore $(C([t_0 - \varepsilon, t_0 + \varepsilon], \mathbb{R}^n), \|\cdot\|_\infty)$ is a Banach space.

Now consider the initial value problem

$$\frac{d\mathbf{y}}{dt} = \mathbf{F}(\mathbf{y}, t), \qquad \mathbf{y}(t_0) = \boldsymbol{\xi}.$$

We assume $\mathbf{F}$ satisfies the Lipschitz condition in $U \subset \mathbb{R}^n \times \mathbb{R}$ with respect to $\mathbf{y}$. For the initial condition, we require

$$(\boldsymbol{\xi}, t_0) \in U.$$

We choose sufficiently small numbers $\varepsilon > 0, \delta > 0$ such that

$$\overline{B(\boldsymbol{\xi}, \delta)} \times [t_0 - \varepsilon, t_0 + \varepsilon] \subset U.$$

Here $\overline{B(\boldsymbol{\xi}, \delta)} \subset \mathbb{R}^n$ is the closed ball of radius $\delta$ centered at $\boldsymbol{\xi}$. Let

$$X_{\varepsilon,\delta} = \{\mathbf{x}(t) \in C([t_0 - \varepsilon, t_0 + \varepsilon], \mathbb{R}^n) \mid \mathbf{x}(t) \in \overline{B(\boldsymbol{\xi}, \delta)}, \forall t \in [t_0 - \varepsilon, t_0 + \varepsilon]\}$$

If we treat $\boldsymbol{\xi}$ as an constant function element of $C([t_0 - \varepsilon, t_0 + \varepsilon], \mathbb{R}^n)$, then $X_{\varepsilon,\delta}$ is the closed ball in $C([t_0 - \varepsilon, t_0 + \varepsilon], \mathbb{R}^n)$ of radius $\delta$ centered at $\boldsymbol{\xi}$

$$X_{\varepsilon,\delta} = \{\mathbf{x} \in C([t_0 - \varepsilon, t_0 + \varepsilon], \mathbb{R}^n) \mid \|\mathbf{x}(t) - \boldsymbol{\xi}\|_\infty \leq \delta\}.$$

Thus $X_{\varepsilon,\delta}$ is also a complete metric space.

Recall the operator $T$ from the integral equation

$$(T\mathbf{x})(t) = \boldsymbol{\xi} + \int_{t_0}^{t} \mathbf{F}(\mathbf{x}(s), s)ds.$$

Assume $\mathbf{x}(t) \in X_{\varepsilon,\delta}$. We have

$$|(T\mathbf{x})(t) - \boldsymbol{\xi}| \leq \int_{t_0}^{t} |\mathbf{F}(\mathbf{x}(s), s)|ds.$$

Since $\mathbf{F}$ is continuous, let

$$K = \max_{(\mathbf{y},t)\in\overline{B(\boldsymbol{\xi},\delta)}\times[t_0-\varepsilon,t_0+\varepsilon]} |\mathbf{F}(\mathbf{y}, t)| < +\infty.$$

Then

$$|(T\mathbf{x})(t) - \boldsymbol{\xi}| \leq |t - t_0|K.$$

By Shrinking $\varepsilon$ if necessary, say $K\varepsilon < \delta$ holds, then $T\mathbf{x}$ will lie in $X_{\varepsilon,\delta}$. Thus we will choose a sufficiently small $\varepsilon$ such that

$$T : X_{\varepsilon,\delta} \to X_{\varepsilon,\delta}$$

defines a map from $X_{\varepsilon,\delta}$ to itself.

Let us now analyze the contraction property. Let $\mathbf{x}_1(t), \mathbf{x}_2(t)$ be two elements of $X_{\varepsilon,\delta}$. Then

$$\begin{aligned}
&|(T\mathbf{x}_1)(t) - (T\mathbf{x}_2)(t)| \\
=&\left|\int_{t_0}^{t} (\mathbf{F}(\mathbf{x}_1(s), s) - \mathbf{F}(\mathbf{x}_2(s), s))ds\right| \\
\leq&\int_{t_0}^{t} |\mathbf{F}(\mathbf{x}_1(s), s) - \mathbf{F}(\mathbf{x}_2(s), s)|ds \\
\leq&L\int_{t_0}^{t} |\mathbf{x}_1(s) - \mathbf{x}_2(s)|ds \\
\leq&L\varepsilon\|\mathbf{x}_1 - \mathbf{x}_2\|_\infty
\end{aligned}$$

holds for any $t \in [t_0 - \varepsilon, t_0 + \varepsilon]$. Here $L$ is the Lipschitz constant. This implies

$$\|T\mathbf{x}_1 - T\mathbf{x}_2\|_\infty \leq L\varepsilon\|\mathbf{x}_1 - \mathbf{x}_2\|_\infty.$$

Therefore by further shrinking $\varepsilon$ if necessary, say $L\varepsilon < 1$ holds, then $T$ is a contraction mapping on $X_{\varepsilon,\delta}$. Now we are ready to prove the local existence and uniqueness theorem.

**Theorem 3.1.2** (Picard–Lindelöf)**.** *Let $\mathbf{F}(\mathbf{y}, t) : U \to \mathbb{R}^n$ be a continuous function from an open subset $U \subset \mathbb{R}^n \times \mathbb{R}$ to $\mathbb{R}^n$, which is Lipschitz with respect to $\mathbf{y}$. Assume $U$ contains the point $(\boldsymbol{\xi}, t_0)$. Then for sufficiently small $\varepsilon > 0$, there exists a unique function $\mathbf{y}(t)$ on the interval $(t_0 - \varepsilon, t_0 + \varepsilon)$ solving the following initial value problem*

$$\begin{cases} \dfrac{d}{dt}\mathbf{y} = \mathbf{F}(\mathbf{y}, t), & t \in (t_0 - \varepsilon, t_0 + \varepsilon) \\ \mathbf{y}(t_0) = \boldsymbol{\xi} \end{cases}$$

*Proof:* Since the question is for sufficiently small $\varepsilon$, the statement is the same for either $(t_0 - \varepsilon, t_0 + \varepsilon)$ or $[t_0 - \varepsilon, t_0 + \varepsilon]$. By choosing sufficiently small $\delta > 0, \varepsilon > 0$ (as we discussed above), the operator $T$

$$(T\mathbf{x})(t) = \boldsymbol{\xi} + \int_{t_0}^{t} \mathbf{F}(\mathbf{x}(s), s)ds$$

defines a contraction mapping on $X_{\varepsilon,\delta}$. Since $X_{\varepsilon,\delta}$ is a complete metric space, we obtain a unique fixed point $\mathbf{y}(t)$ of $T$ by the Contraction Mapping Theorem. The relation $T\mathbf{y} = \mathbf{y}$, *i.e.*

$$\mathbf{y}(t) = \boldsymbol{\xi} + \int_{t_0}^{t} \mathbf{F}(\mathbf{y}(s), s)ds$$

implies that $\mathbf{y}(t)$ solves the required initial value problem. □

*Remark* 3.1.3. The proof is in fact constructive. It shows that the solution can be found by the limit of the iteration process of $T$. This is called Picard iteration.

*Remark* 3.1.4. If we only assume $\mathbf{F}$ is continuous but not require Lipschitz condition, then local existence still holds. This is the Peano Existence Theorem. We will not prove this general existence theorem here. The basic idea is that the iterated sequence by $T$ may not converge now, but will have a convergent subsequence by using the Arzelà–Ascoli Theorem. However, different choices of convergent subsequence may lead to different limits, so uniqueness may fail.

For example, consider

$$y' = \sqrt{|y|}, \qquad y(0) = 0.$$

It is clear that both $y = 0$ and

$$y(t) = \begin{cases} 0, & t \le 0, \\ \frac{t^2}{4}, & t \ge 0. \end{cases}$$

solve the equation, but they are different in any small neighborhood of $t = 0$.

Finally, we observe that the local existence and uniqueness only requires Lipschitz condition locally. We make this precisely by introducing the notion of local Lipschitz condition.

**Definition 3.1.5.** Let $U$ be an open subset of $\mathbb{R}^n \times \mathbb{R}$ and

$$\mathbf{F}(\mathbf{y}, t) : U \to \mathbb{R}$$

be a continuous function, where $\mathbf{y} \in \mathbb{R}^n, t \in \mathbb{R}$. We say $\mathbf{F}(\mathbf{y}, t)$ is locally Lipschitz in $U$ with respect to $\mathbf{y}$ if for any point $p \in U$, there exists an open neighborhood $V \subset U$ containing $p$ such that $\mathbf{F}$ is Lipschitz in $V$ with respect to $\mathbf{y}$.

**Theorem 3.1.6.** *Let $\mathbf{F}(\mathbf{y}, t) : U \to \mathbb{R}^n$ be a continuous function from an open subset $U \subset \mathbb{R}^n \times \mathbb{R}$ to $\mathbb{R}^n$, which is locally Lipschitz with respect to $\mathbf{y}$. Then for any point $(\boldsymbol{\xi}, t_0) \in U$, there exists a unique solution $\mathbf{y}(t)$ to the initial value problem*

$$\frac{d\mathbf{y}}{dt} = \mathbf{F}(\mathbf{y}, t), \qquad \mathbf{y}(t_0) = \boldsymbol{\xi}$$

*on the interval $(t_0 - \varepsilon, t_0 + \varepsilon)$, for sufficiently small $\varepsilon > 0$.*

*Proof:* This follows directly from Theorem 3.1.2. □

## 3.2 Extension of solutions

In the previous subsection, we have established the existence and uniqueness of local solutions to the initial value problem

$$\frac{d\mathbf{y}}{dt} = \mathbf{F}(\mathbf{y}, t), \qquad \mathbf{y}(t_0) = \boldsymbol{\xi}$$

for $\mathbf{F}$ Lipschitz with respect to $\mathbf{y}$. We now consider how far the solution can be extended.

### 3.2.1 Maximal Interval of Existence

We first observe that once two solutions coincide at some point, they will be the same on their defining domains.

**Proposition 3.2.1.** *Let $\mathbf{F}(\mathbf{y}, t) : U \to \mathbb{R}^n$ be a continuous function from an open subset $U \subset \mathbb{R}^n \times \mathbb{R}$ to $\mathbb{R}^n$, which is locally Lipschitz with respect to $\mathbf{y}$. Let $(\boldsymbol{\xi}, t_0) \in U$. Let $\mathbf{y}_i(t)$ be a function on the interval $(\alpha_i, \beta_i)$ containing $t_0$, for $i = 1, 2$, which both solve the initial value problem*

$$\frac{d\mathbf{y}}{dt} = \mathbf{F}(\mathbf{y}, t), \qquad \mathbf{y}(t_0) = \boldsymbol{\xi}.$$

*Then $\mathbf{y}_1(t) = \mathbf{y}_2(t)$ for $\max\{\alpha_1, \alpha_2\} < t < \min\{\beta_1, \beta_2\}$.*

*Proof:* Let $\beta = \min\{\beta_1, \beta_2\}$. We consider the forward time part for $t \in [t_0, \beta)$. The other direction is similar. Assume $\mathbf{y}_1(t)$ and $\mathbf{y}_2(t)$ are not the same on $[t_0, \beta)$. Let

$$\tau = \inf\{t \in [t_0, \beta) \mid \mathbf{y}_1(t) \neq \mathbf{y}_2(t)\}.$$

We have $t_0 \leq \tau < \beta$ and $\mathbf{y}_1(\tau) = \mathbf{y}_2(\tau)$. Let $\boldsymbol{\eta} = \mathbf{y}_1(\tau)$.

Applying Theorem 3.1.6 to the point $(\boldsymbol{\eta}, \tau)$, we find a sufficiently small $\varepsilon > 0$ such that $\mathbf{y}_1(t) = \mathbf{y}_2(t)$ on $t \in (\tau - \varepsilon, \tau + \varepsilon)$. This contradicts the definition of $\tau$. □

Now we come to the maximal interval of the solution of an initial value problem.

**Theorem 3.2.2.** *Let $\mathbf{F}(\mathbf{y}, t) : U \to \mathbb{R}^n$ be a continuous function from an open subset $U \subset \mathbb{R}^n \times \mathbb{R}$ to $\mathbb{R}^n$, which is locally Lipschitz with respect to $\mathbf{y}$. Given $(\boldsymbol{\xi}, t_0) \in U$, there exists $t_- < t_0 < t_+$ and a function $\mathbf{y}(t)$ on $(t_-, t_+)$ solving the initial value problem*

$$\frac{d\mathbf{y}}{dt} = \mathbf{F}(\mathbf{y}, t), \qquad \mathbf{y}(t_0) = \boldsymbol{\xi} \tag{$*$}$$

*which is maximal in the following sense: if $\tilde{\mathbf{y}}(t)$ is another solution of $(*)$ on some interval $I \ni t_0$, then*

$$I \subset (t_-, t_+) \qquad \text{and} \qquad \tilde{\mathbf{y}}(t) = \mathbf{y}(t) \quad \text{for } t \in I.$$

*Here $t_\pm$ could be $\pm\infty$.*

*Proof:* Define

$$t_- = \inf\{\alpha : (*) \text{ is solvable for } \alpha < t \le t_0\}$$

$$t_+ = \sup\{\beta : (*) \text{ is solvable for } t_0 \le t < \beta\}$$

Note that if $\mathbf{y}_+(t)$ solves $(*)$ for $t \in [t_0, \beta)$ and $\mathbf{y}_-(t)$ solves $(*)$ for $t \in (\alpha, t_0]$, then by the uniqueness of local solution, $\mathbf{y}_\pm(t)$ glue to define a solution for $t \in (\alpha, \beta)$.

Let us choose intervals

$$t_0 \in (\alpha_1, \beta_1) \subset (\alpha_2, \beta_2) \subset \cdots \subset (\alpha_n, \beta_n) \subset \cdots \subset (t_-, t_+)$$

such that $\alpha_n \to t_-$, $\beta_n \to t_+$. Let $\mathbf{y}_n(t)$ solves $(*)$ on the interval $(\alpha_n, \beta_n)$. By Proposition 3.2.1,

$$\mathbf{y}_{n+1}(t) = \mathbf{y}_n(t), \qquad \text{for} \quad t \in (\alpha_n, \beta_n).$$

Then we define the solution $\mathbf{y}(t)$ on $(t_-, t_+)$ by

$$\mathbf{y}(t) = \mathbf{y}_n(t) \qquad \text{if } t \in (\alpha_n, \beta_n).$$

It is readily checked (Using Proposition 3.2.1 again) that such $\mathbf{y}(t)$ has the required property. □

The next proposition provides a useful tool to analyze global existence of solutions.

**Proposition 3.2.3.** *Let $\mathbf{y}(t)$ be a maximal solution defined on $(t_-, t_+)$ in the sense of Theorem 3.2.2. Assume $t_+ < +\infty$. Then for any compact set $K \subset U$, there exists $\varepsilon > 0$ such that $(\mathbf{y}(t), t) \notin K$ for $t_+ - \varepsilon < t < t_+$. In other words, the solution will run out of $K$ eventually after certain time. There is a similar result in the negative time direction if $t_- > -\infty$.*

*Proof:* Let us concentrate on $t \in [t_0, t_+)$. First, we can find a larger compact set $\tilde{K} \subset U$ and $\delta > 0, \varepsilon > 0$ such that

$$\overline{B(\boldsymbol{\xi}, \delta)} \times [t - \varepsilon, t + \varepsilon] \subset \tilde{K} \qquad \forall (\boldsymbol{\xi}, t) \in K.$$

Since $\tilde{K}$ is compact and $\mathbf{F}$ is locally Lipschitz, we can find a constant $L$ such that (Exercise 1)

$$|\mathbf{F}(\mathbf{y}_1, t) - \mathbf{F}(\mathbf{y}_2, t)| \le L|\mathbf{y}_1 - \mathbf{y}_2|, \qquad \forall (\mathbf{y}_1, t), (\mathbf{y}_2, t) \in \tilde{K}.$$

Let $M > 0$ be chosen such that

$$\sup_{(\mathbf{y}, t) \in \tilde{K}} |\mathbf{F}(\mathbf{y}, t)| < M.$$

This can be achieved since $\tilde{K}$ is compact. Shrinking $\varepsilon$ if necessary ($M$ can be fixed), we can assume $\varepsilon < \min\{\frac{\delta}{M}, \frac{1}{L}\}$. We claim that the maximal solution $\mathbf{y}(t)$ has the property that

$$(\mathbf{y}(t), t) \notin K \qquad \text{when} \quad t_+ - \varepsilon < t < t_+.$$

In fact, assume $(\mathbf{y}(\tau) = \boldsymbol{\xi}, \tau) \in K$ for $\tau \in (t_+ - \varepsilon, t_+)$. By the proof in Theorem 3.1.2, our choice of $\varepsilon$ implies that the contraction property of $T$ holds, leading to a solution of

$$\frac{d\mathbf{y}}{dt} = \mathbf{F}(\mathbf{y}, t), \qquad \mathbf{y}(\tau) = \boldsymbol{\xi}$$

on the interval $(\tau - \varepsilon, \tau + \varepsilon)$. This implies that the solution $\mathbf{y}(t)$ can be extended to $[t_0, \tau + \varepsilon)$. But $\tau + \varepsilon > t_+$, contradicting the definition of $t_+$. □

### 3.2.2 Grönwall's Inequality

**Proposition 3.2.4** (Grönwall's Inequality)**.** *Let $f : [a, b] \to \mathbb{R}$ be continuous and satisfies*

$$f(t) \le C + k \int_a^t f(s)ds \qquad a \le t \le b$$

*for some constants $C, k$ with $k \ge 0$. Then*

$$f(t) \le Ce^{k(t-a)}, \qquad a \le t \le b.$$

*Proof:* Define a function $g(t)$ by

$$g(t) = C + k \int_a^t f(s)ds.$$

By assumption

$$f(t) \le g(t), \quad g'(t) = kf(t), \qquad a \le t \le b.$$

It follows that

$$\frac{d}{dt}(e^{-kt}g(t)) = e^{-kt}(g'(t) - kg(t)) \le 0.$$

Thus

$$e^{-kt}g(t) \le e^{-ka}g(a) = Ce^{-ka}$$
$$\Rightarrow \qquad f(t) \le g(t) \le Ce^{k(t-a)}.$$

□

As an application of Proposition 3.2.3 and Grönwall's inequality, we show the global existence of solution for all time when $\mathbf{F}$ is at most linear growth.

**Proposition 3.2.5.** *Let $\mathbf{F}(\mathbf{y}, t) : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ be a continuous function which is locally Lipschitz with respect to $\mathbf{y}$. Assume there exists positive constants $k$ and $C$ such that*

$$|\mathbf{F}(\mathbf{y}, t)| \le k|\mathbf{y}| + C.$$

*Then the solution $\mathbf{y}(t)$ of the initial value problem*

$$\frac{d\mathbf{y}}{dt} = \mathbf{F}(\mathbf{y}, t), \qquad \mathbf{y}(0) = \boldsymbol{\xi}$$

*exists for all time $t \in (-\infty, +\infty)$. Moreover,*

$$|\mathbf{y}(t)| \le |\boldsymbol{\xi}|e^{k|t|} + \frac{C}{k}(e^{k|t|} - 1) \qquad \forall t.$$

*Proof:* Let $(t_-, t_+)$ be the maximal interval for the solution as in Theorem 3.2.2. We show $t_+ = +\infty$. The proof for $t_- = -\infty$ is similar. Suppose $t_+ < +\infty$. Define

$$g(t) = |\mathbf{y}(t)|.$$

Using the integral equation

$$\mathbf{y}(t) = \boldsymbol{\xi} + \int_0^t \mathbf{F}(\mathbf{y}(s), s)ds$$

and the Linear growth condition, we have

$$g(t) \le |\boldsymbol{\xi}| + \int_0^t (kg(s) + C)ds$$

i.e.,

$$\left(g(t) + \frac{C}{k}\right) \le \left(|\boldsymbol{\xi}| + \frac{C}{k}\right) + \int_0^t k\left(g(s) + \frac{C}{k}\right) ds.$$

Applying Grönwall's inequality to $f(t) = g(t) + \frac{C}{k}$

$$\Rightarrow \qquad g(t) + \frac{C}{k} \le \left(|\boldsymbol{\xi}| + \frac{C}{k}\right) e^{kt}, \qquad 0 \le t < t_+.$$

Therefore

$$|\mathbf{y}(t)| \le \frac{C}{k}(e^{kt} - 1) + |\boldsymbol{\xi}|e^{kt}, \qquad 0 \le t < t_+.$$

Assume $t_+ < +\infty$. Then $\mathbf{y}(t)$ will always stay in the following compact region for $0 \le t < t_+$

$$K = \left\{ (\mathbf{y}, t) \in \mathbb{R}^n \times \mathbb{R} \,\middle|\, |\mathbf{y}| \le \frac{C}{k}(e^{kt_+} - 1) + |\boldsymbol{\xi}|e^{kt_+}, 0 \le t \le t_+ \right\}.$$

This contradicts Proposition 3.2.3. □

Global solution at all time may or may not exist when $\mathbf{F}$ goes beyond linear growth. For example, the solution for

$$\frac{dy}{dt} = y^2, \qquad y(0) = 1$$

is $y(t) = \frac{1}{1-t}$, whose maximal interval of existence is $(-\infty, 1)$. For another example, consider

$$\frac{dy}{dt} = -2ty^2, \qquad y(0) = 1.$$

This is not linear growth. But the solution

$$y(t) = \frac{1}{1 + t^2}$$

exists at all time.

## 3.3 Dependence on Initial Data

A mathematical problem modeling a reasonable physical system is called well-posed if it satisfies the following requirements

① Existence: the problem has at least one solution

② Uniqueness: the problem has no more than one solution

③ Continuous dependence: the solution depends continuously on the data given.

So far we have established the local existence and uniqueness of the initial value problem of the non-linear system

$$\frac{d\mathbf{y}}{dt} = \mathbf{F}(\mathbf{y}, t), \qquad \mathbf{y}(t_0) = \boldsymbol{\xi}$$

for $\mathbf{F}$ with local Lipschitz condition. In this subsection we will prove the continuous dependence of the solution on the initial data. This establishes its well-posedness.

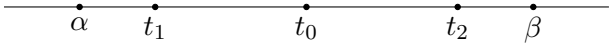
### 3.3.1 Continuous Dependence on Initial Value

**Theorem 3.3.1.** *Let $\mathbf{F}(\mathbf{y}, t) : U \to \mathbb{R}^n$ be a continuous function from an open subset $U \subset \mathbb{R}^n \times \mathbb{R}$ to $\mathbb{R}^n$, which is locally Lipschitz with respect to $\mathbf{y}$. Let $(\boldsymbol{\xi}_0, t_0) \in U$. Let $\mathbf{y}_0(t)$ be a solution of $\frac{d\mathbf{y}_0}{dt} = \mathbf{F}(\mathbf{y}_0, t)$ on the interval $(\alpha, \beta) \ni t_0$ with initial condition $\mathbf{y}_0(t_0) = \boldsymbol{\xi}_0$. Then*

*(i) For every $\alpha < t_1 < t_0 < t_2 < \beta$, there exists a neighborhood $V \subset \mathbb{R}^n$ of $\boldsymbol{\xi}_0$ such that the initial value problem*

$$\frac{d\mathbf{y}}{dt} = \mathbf{F}(\mathbf{y}, t), \qquad \mathbf{y}(t_0) = \boldsymbol{\xi}$$

*has a solution on $t \in (t_1, t_2)$ for any $\boldsymbol{\xi} \in V$.*




*(ii) Let $\mathbf{y}(t, \boldsymbol{\xi})$ denote the solution with initial condition $\mathbf{y}(t_0) = \boldsymbol{\xi}$ for $\boldsymbol{\xi} \in V$ as in (i). Then there exists nonnegative constants $L, M$ such that*

$$|\mathbf{y}(t', \boldsymbol{\xi}') - \mathbf{y}(t, \boldsymbol{\xi})| \le M|t' - t| + |\boldsymbol{\xi}' - \boldsymbol{\xi}|e^{L|t-t_0|}$$

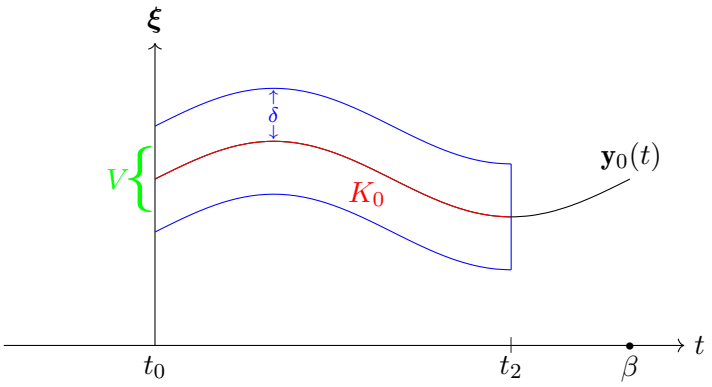
*for any $\boldsymbol{\xi}, \boldsymbol{\xi}' \in V$ and $t, t' \in (t_1, t_2)$. In particular, $\mathbf{y}(t, \boldsymbol{\xi})$ is Lipschitz continuous as a function of $(t, \boldsymbol{\xi})$ on the domain $(t_1, t_2) \times V$.*

*Proof:* We consider the forward time part on $[t_0, t_2)$. The discussion on $(t_1, t_0]$ is similar. They combine to prove the theorem. Consider

$$K_0 = \{(\mathbf{y}_0(t), t) \mid t \in [t_0, t_2]\} \subset U$$

which is a compact subset of $U$. Let $\delta > 0$ be sufficiently small such that

$$K = \bigcup_{t \in [t_0, t_2]} \overline{B(\mathbf{y}_0(t), \delta)} \times \{t\} \subset U.$$




Since $K$ is compact and $\mathbf{F}$ is locally Lipschitz, there exists $L > 0$ such that (Exercise 1)

$$|\mathbf{F}(\mathbf{y}_1, t) - \mathbf{F}(\mathbf{y}_2, t)| \le L|\mathbf{y}_1 - \mathbf{y}_2|, \qquad \forall (\mathbf{y}_1, t), (\mathbf{y}_2, t) \in K.$$

Let $V$ be the open ball

$$V = B(\boldsymbol{\xi}_0, e^{-L(t_2 - t_0)}\delta)$$

We show that for each $\boldsymbol{\xi} \in V$, the initial value problem

$$\frac{d\mathbf{y}}{dt} = \mathbf{F}(\mathbf{y}, t), \qquad \mathbf{y}(t_0) = \boldsymbol{\xi}$$

has a solution on $[t_0, t_2)$ with the required property.

In fact, for $\boldsymbol{\xi} \in V$, let $[t_0, t_+)$ be the maximal interval of the solution $\mathbf{y}$ with $\mathbf{y}(t_0) = \boldsymbol{\xi}$ in the forward time. By the choice of $V$, $\mathbf{y}(t)$ will lie inside $K$ at least for $t$ close to $t_0$. Let

$$g(t) = |\mathbf{y}(t) - \mathbf{y}_0(t)|, \qquad t_0 \le t < \min(t_+, t_2).$$

Using the integral equation

$$\begin{cases} \mathbf{y}(t) = \boldsymbol{\xi} + \displaystyle\int_{t_0}^{t} \mathbf{F}(\mathbf{y}(s), s)ds \\ \mathbf{y}_0(t) = \boldsymbol{\xi}_0 + \displaystyle\int_{t_0}^{t} \mathbf{F}(\mathbf{y}_0(s), s)ds \end{cases}$$

and the Lipschitz condition, we have (for $t$ close to $t_0$)

$$\begin{aligned} g(t) \le & |\boldsymbol{\xi} - \boldsymbol{\xi}_0| + \int_{t_0}^{t} |\mathbf{F}(\mathbf{y}(s), s) - \mathbf{F}(\mathbf{y}_0(s), s)|ds \\ \le & |\xi - \xi_0| + L\int_{t_0}^{t} g(s)ds. \end{aligned}$$

Hence by Grönwall's Lemma

$$g(t) \le |\boldsymbol{\xi} - \boldsymbol{\xi}_0| e^{L(t-t_0)}.$$

Since $\boldsymbol{\xi} \in V$, $|\boldsymbol{\xi} - \boldsymbol{\xi}_0| \le \delta e^{-L(t_2 - t_0)}$, thus

$$g(t) \le \delta.$$

This means that the solution $\mathbf{y}(t)$ will stay inside the compact space $K$ for $t_0 \le t < \min(t_+, t_2)$. By Proposition 3.2.3, this implies $t_+ \ge t_2$. Thus the solution $\mathbf{y}(t)$ exists on $[t_0, t_2)$, and the above application of Grönwall's Lemma gives

$$|\mathbf{y}(t) - \mathbf{y}_0(t)| \le |\boldsymbol{\xi} - \boldsymbol{\xi}_0| e^{L(t-t_0)}$$

for $\boldsymbol{\xi} \in V$ and $t \in [t_0, t_2)$. This proves (i).

Let $\mathbf{y}(t, \boldsymbol{\xi})$ denote the solution with initial condition $\mathbf{y}(t_0, \xi) = \xi$ for $\xi \in V$ in forward time as above. It satisfies the integral equation

$$\mathbf{y}(t, \boldsymbol{\xi}) = \boldsymbol{\xi} + \int_{t_0}^{t} \mathbf{F}(\mathbf{y}(s, \boldsymbol{\xi}), s)ds$$

and

$$(\mathbf{y}(t, \boldsymbol{\xi}), t) \in K, \qquad \forall\ \boldsymbol{\xi} \in V, \quad t \in [t_0, t_2).$$

The same argument as above via Grönwall's Lemma implies

$$|\mathbf{y}(t, \boldsymbol{\xi}') - \mathbf{y}(t, \boldsymbol{\xi})| \le |\boldsymbol{\xi}' - \boldsymbol{\xi}| e^{L(t-t_0)}, \qquad \forall\ \boldsymbol{\xi}', \boldsymbol{\xi} \in V,\ t \in [t_0, t_2).$$

Since $K$ is compact, there exists $M > 0$ such that

$$|\mathbf{F}(\mathbf{y}, t)| \leq M, \qquad \forall (\mathbf{y}, t) \in K.$$

Then for any $\boldsymbol{\xi}, \boldsymbol{\xi}' \in V$ and $t, t' \in [t_0, t_2)$, we have

$$\begin{aligned} |\mathbf{y}(t', \boldsymbol{\xi}') - \mathbf{y}(t, \boldsymbol{\xi})| &\leq |\mathbf{y}(t', \boldsymbol{\xi}') - \mathbf{y}(t, \boldsymbol{\xi}')| + |\mathbf{y}(t, \boldsymbol{\xi}') - \mathbf{y}(t, \boldsymbol{\xi})| \\ &\leq \int_t^{t'} |\mathbf{F}(\mathbf{y}(s, \boldsymbol{\xi}'), s)| ds + |\mathbf{y}(t, \boldsymbol{\xi}') - \mathbf{y}(t, \boldsymbol{\xi})| \\ &\leq M|t' - t| + |\boldsymbol{\xi}' - \boldsymbol{\xi}| e^{L|t - t_0|}. \end{aligned}$$

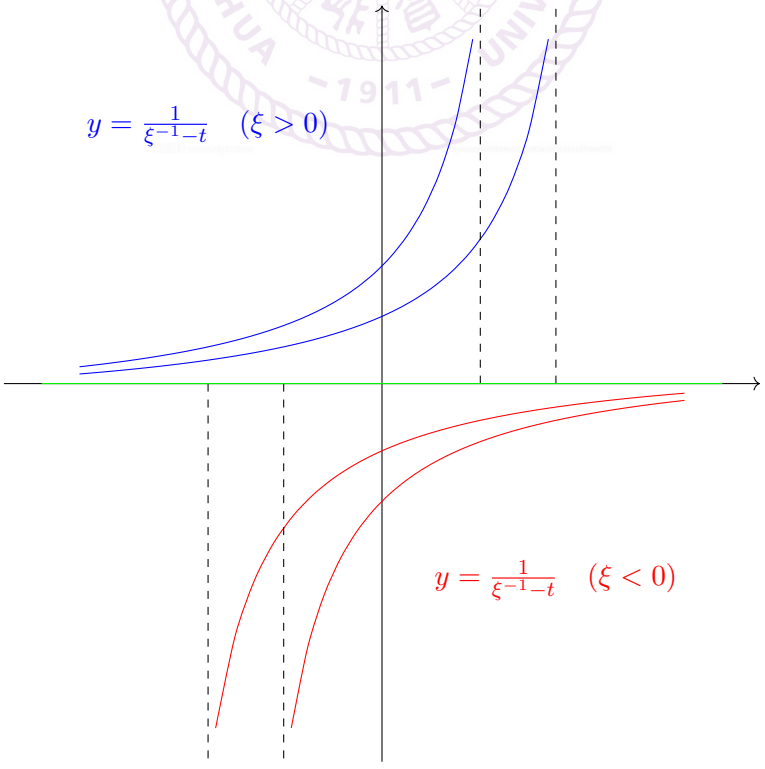
This proves (ii).

□

*Remark* 3.3.2. From the proof, we see that the case for either $\alpha = -\infty$ or $\beta = +\infty$ is allowed. However, the condition $t_1 > \alpha, t_2 < \beta$ asks that $(t_1, t_2)$ is a finite interval. Such shrinking of intervals for nearby solutions is necessary. For example, consider

$$y' = y^2, \qquad y(0) = \xi.$$

For $\xi = 0$, the solution $y(t) = 0$ exists at all time. For $\xi \neq 0$, the solution is $y(t) = \frac{1}{\xi^{-1} - t}$, where the maximal interval of solution is $(-\infty, \xi^{-1})$ for $\xi > 0$ and $(\xi^{-1}, +\infty)$ for $\xi < 0$. It is also clear from this example that the bigger the finite interval $(t_1, t_2)$, the smaller the neighborhood of the initial condition that we have to restrict to.



### 3.3.2 Continuous Dependence on Parameters

Now we consider the case when the equation itself can vary with some parameters. Let us write $\mathbf{F} = \mathbf{F}(\mathbf{y}, t, \boldsymbol{\lambda})$ where $\boldsymbol{\lambda}$ collects the parameters. We consider the initial value problem

$$\frac{d\mathbf{y}}{dt} = \mathbf{F}(\mathbf{y}, t, \boldsymbol{\lambda}), \qquad \mathbf{y}(t_0) = \boldsymbol{\xi}.$$

Then the solution $\mathbf{y}(t, \boldsymbol{\xi}, \boldsymbol{\lambda})$ will depend both on the initial value $\boldsymbol{\xi}$ and on the parameter $\boldsymbol{\lambda}$.

This problem can be reduced to what we have studied in the previous Section 3.3.1. In fact, solving $\mathbf{y}(t, \boldsymbol{\xi}, \boldsymbol{\lambda})$ is the same as solving the initial value problem of the enlarged system

$$\begin{cases} \frac{d\mathbf{y}}{dt} = \mathbf{F}(\mathbf{y}, t, \boldsymbol{z}) \\ \frac{d\boldsymbol{z}}{dt} = 0 \end{cases} \quad \text{with} \quad \begin{cases} \mathbf{y}(t_0) = \boldsymbol{\xi} \\ \boldsymbol{z}(t_0) = \boldsymbol{\lambda} \end{cases}$$

This reformulation, together with Theorem 3.3.1, immediately leads to the following theorem.

**Theorem 3.3.3.** *Let $\mathbf{F}(\mathbf{y}, t, \boldsymbol{\lambda}) : U \to \mathbb{R}^n$ be a continuous function from an open subset $U \subset \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^m$ to $\mathbb{R}^n$, which is locally Lipschitz with respect to $(\mathbf{y}, \boldsymbol{\lambda})$. Let $(\boldsymbol{\xi}_0, t_0, \boldsymbol{\lambda}_0) \in U$. Let $\mathbf{y}_0(t)$ be a solution of $\frac{d\mathbf{y}_0}{dt} = \mathbf{F}(\mathbf{y}_0, t, \boldsymbol{\lambda}_0)$ on the interval $(\alpha, \beta) \ni t_0$ with initial condition $\mathbf{y}_0(t_0) = \boldsymbol{\xi}_0$. Then*

*(i) For every $\alpha < t_1 < t_0 < t_2 < \beta$, there exists a neighborhood $V \subset \mathbb{R}^n \times \mathbb{R}^m$ of $(\boldsymbol{\xi}_0, \boldsymbol{\lambda}_0)$ such that the initial value problem*

$$\frac{d\mathbf{y}}{dt} = \mathbf{F}(\mathbf{y}, t, \boldsymbol{\lambda}), \qquad \mathbf{y}(t_0) = \boldsymbol{\xi}$$

*has a solution on $t \in (t_1, t_2)$ for any $(\boldsymbol{\xi}, \boldsymbol{\lambda}) \in V$.*

*(ii) Let $\mathbf{y}(t, \boldsymbol{\xi}, \boldsymbol{\lambda})$ denote the solution with initial condition $\mathbf{y}(t_0, \boldsymbol{\xi}) = \boldsymbol{\xi}$ for $(\boldsymbol{\xi}, \boldsymbol{\lambda}) \in V$ as in (i). Then there exists nonnegative constant $L, M$ such that*

$$|\mathbf{y}(t', \boldsymbol{\xi}', \boldsymbol{\lambda}') - \mathbf{y}(t, \boldsymbol{\xi}, \boldsymbol{\lambda})| \le M|t' - t| + (|\boldsymbol{\xi}' - \boldsymbol{\xi}| + |\boldsymbol{\lambda}' - \boldsymbol{\lambda}|)e^{L|t - t_0|}$$

*for any $(\boldsymbol{\xi}, \boldsymbol{\lambda}), (\boldsymbol{\xi}', \boldsymbol{\lambda}') \in V$ and $t, t' \in (t_1, t_2)$. In particular, $\mathbf{y}(t, \boldsymbol{\xi}, \boldsymbol{\lambda})$ is Lipschitz continuous as a function of $(t, \boldsymbol{\xi}, \boldsymbol{\lambda})$ on the domain $(t_1, t_2) \times V$.*

### 3.3.3 Differentiability

In this section, we show that the solution is differentiable with respect to the initial data if $\mathbf{F}$ is $C^1$. Before we move into technical details, let us first discuss intuitively what the derivative with respect to the initial data should look like.

Let $\mathbf{y}(t, \boldsymbol{\xi})$ be the solution to the initial value problem

$$\mathbf{y}' = \mathbf{F}(\mathbf{y}, t), \qquad \mathbf{y}(t_0) = \boldsymbol{\xi} \tag{$*$}$$

Here the parameter $\boldsymbol{\xi}$ represents initial data at $t_0$. Assume $\mathbf{y}(t, \boldsymbol{\xi})$ is also differentiable in $\boldsymbol{\xi}$. Let us compute the derivative in the $\boldsymbol{\xi}_1$-direction at $\boldsymbol{\xi} = \boldsymbol{\xi}_0$

$$D_{\boldsymbol{\xi}_1}\mathbf{y}(t, \boldsymbol{\xi}_0) := \lim_{\varepsilon \to 0} \frac{\mathbf{y}(t, \boldsymbol{\xi}_0 + \varepsilon\boldsymbol{\xi}_1) - \mathbf{y}(t, \boldsymbol{\xi}_0)}{\varepsilon}.$$

To simplify notation, let

$$\mathbf{y}_0(t) = \mathbf{y}(t, \boldsymbol{\xi}_0)$$

which solves

$$\frac{d\mathbf{y}_0}{dt} = \mathbf{F}(\mathbf{y}_0, t), \qquad \mathbf{y}_0(t_0) = \boldsymbol{\xi}_0.$$

Let

$$\mathbf{y}_1(t) = D_{\boldsymbol{\xi}_1}\mathbf{y}(t, \boldsymbol{\xi}_0).$$

Applying the $\boldsymbol{\xi}_1$-direction derivative on both sides of $(*)$

$$\begin{aligned}
\frac{d}{dt}\mathbf{y}_1(t) =& \frac{d}{dt} D_{\boldsymbol{\xi}_1}\mathbf{y}(t, \boldsymbol{\xi}_0) \\
=& D_{\boldsymbol{\xi}_1}\left(\frac{\partial}{\partial t}\mathbf{y}(t, \boldsymbol{\xi})\right)\Big|_{\xi=\xi_0} \\
=& D_{\boldsymbol{\xi}_1}(\mathbf{F}(\mathbf{y}(t, \boldsymbol{\xi}), t))|_{\xi=\xi_0} \\
=& \sum_j \partial_{y^j}\mathbf{F}(\mathbf{y}(t, \boldsymbol{\xi}_0), t) D_{\boldsymbol{\xi}_1} y^j(t, \boldsymbol{\xi}_0) \\
=& D\mathbf{F}(\mathbf{y}_0(t), t)\mathbf{y}_1(t)
\end{aligned}$$

where $\mathbf{y} = (y^j)$ is the column vector and $D\mathbf{F}$ is the matrix

$$(D\mathbf{F})_{ij} = \partial_{y^j} F^i, \qquad \mathbf{F} = (F^i).$$

Moreover, the initial condition gives

$$\mathbf{y}_1(t_0) = D_{\boldsymbol{\xi}_1}\, \mathbf{y}(t_0, \boldsymbol{\xi})|_{\xi=\xi_0} = \boldsymbol{\xi}_1.$$

In other words, $\mathbf{y}_1(t)$ solves

$$\begin{cases} \dfrac{d\mathbf{y}_1}{dt} = D\mathbf{F}(\mathbf{y}_0(t), t)\mathbf{y}_1(t) \\ \mathbf{y}_1(t_0) = \boldsymbol{\xi}_1 \end{cases}$$

This linear system has a unique solution for given $\mathbf{y}_0$. This solution $\mathbf{y}_1(t)$ is reasonable to be the candidate for the $\boldsymbol{\xi}_1$-direction derivative of $\mathbf{y}(t, \boldsymbol{\xi})$ at $\boldsymbol{\xi} = \boldsymbol{\xi}_0$, as illustrated by the above calculation. This is indeed the case.

**Theorem 3.3.4.** *Assume $\mathbf{F}(\mathbf{y}, t)$ is $C^1$ on $U \subset \mathbb{R}^n \times \mathbb{R}$. Let $(\boldsymbol{\xi}_0, t_0) \in U$ and $\mathbf{y}_0(t)$ be the solution of $\frac{d\mathbf{y}_0}{dt} = \mathbf{F}(\mathbf{y}_0, t)$, $\mathbf{y}_0(t_0) = \boldsymbol{\xi}_0$, on the interval $(\alpha, \beta) \ni t_0$. For any $\alpha < t_1 < t_0 < t_2 < \beta$, and $\boldsymbol{\xi}_1 \in \mathbb{R}^n$, let $\mathbf{y}(t, \varepsilon)$ be the solution of*

$$\begin{cases} \dfrac{\partial}{\partial t}\mathbf{y}(t, \varepsilon) = \mathbf{F}(\mathbf{y}(t, \varepsilon), t) \\ \mathbf{y}(t_0, \varepsilon) = \boldsymbol{\xi}_0 + \varepsilon\boldsymbol{\xi}_1 \end{cases}$$

*on the interval $(t_1, t_2)$ and for $\varepsilon$ sufficiently small, as guaranteed by Theorem 3.3.1. Let $\mathbf{y}_1(t)$ be the solution of*

$$\begin{cases} \dfrac{d\mathbf{y}_1}{dt} = D\mathbf{F}(\mathbf{y}_0(t), t)\mathbf{y}_1, \qquad \alpha < t < \beta, \\ \mathbf{y}_1(t_0) = \boldsymbol{\xi}_1 \end{cases}$$

*Here* $D\mathbf{F}$ *is the matrix* $(D\mathbf{F})_{ij} = \partial_{y^j} F^i$ *where* $\mathbf{F} = (F^i)$. *Then*

$$\lim_{\varepsilon \to 0} \frac{|\mathbf{y}(t,\varepsilon) - \mathbf{y}_0(t) - \varepsilon \mathbf{y}_1(t)|}{\varepsilon} = 0$$

*uniformly for* $t \in (t_1, t_2)$. *In particular,* $\mathbf{y}(t,\xi)$ *is differentiable in* $\xi$ *for* $\xi$ *near* $\xi_0$ *and*

$$D_{\xi_1} \mathbf{y}(t, \xi_0) = \mathbf{y}_1(t), \qquad t_1 < t < t_2.$$

*Proof:* We consider only the forward time $t \in [t_0, t_2)$. We have the integral equation

$$\begin{cases} \mathbf{y}(t,\varepsilon) = \boldsymbol{\xi}_0 + \varepsilon \boldsymbol{\xi}_1 + \displaystyle\int_{t_0}^{t} \mathbf{F}(\mathbf{y}(s,\varepsilon), s) ds \\ \mathbf{y}_0(t) = \boldsymbol{\xi}_0 + \displaystyle\int_{t_0}^{t} \mathbf{F}(\mathbf{y}_0(s), s) ds \\ \mathbf{y}_1(t) = \boldsymbol{\xi}_1 + \displaystyle\int_{t_0}^{t} D\mathbf{F}(\mathbf{y}_0(s), s) \mathbf{y}_1(s) ds \end{cases}$$

The solutions exist on $t \in [t_0, t_2]$ for $\varepsilon$ sufficiently small by Theorem 3.3.1.

Let $g(t,\varepsilon) = |\mathbf{y}(t,\varepsilon) - \mathbf{y}_0(t) - \varepsilon \mathbf{y}_1(t)|$. Then

$$\begin{aligned} g(t,\varepsilon) =& \left| \int_{t_0}^{t} (\mathbf{F}(\mathbf{y}(s,\varepsilon), s) - \mathbf{F}(\mathbf{y}_0(s), s) - \varepsilon D\mathbf{F}(\mathbf{y}_0(s), s) \mathbf{y}_1(s)) ds \right| \\ \leq & I(t,\varepsilon) + \int_{t_0}^{t} \|D\mathbf{F}(\mathbf{y}_0(s), s)\| \, g(s,\varepsilon) ds \end{aligned}$$

where

$$I(t,\varepsilon) = \left| \int_{t_0}^{t} [\mathbf{F}(\mathbf{y}(s,\varepsilon), s) - \mathbf{F}(\mathbf{y}_0(s), s) - D\mathbf{F}(\mathbf{y}_0(s), s)(\mathbf{y}(s,\varepsilon) - \mathbf{y}_0(s))] \, ds \right|.$$

Since $\mathbf{F}$ is $C^1$, there exists $M > 0$ such that

$$\begin{aligned} &\|D\mathbf{F}(\mathbf{y}_0(s), s)\| \leq M, \qquad t_0 \leq s \leq t_2 \\ &\Rightarrow g(t,\varepsilon) \leq I(t,\varepsilon) + M \int_{t_0}^{t} g(s,\varepsilon) ds. \end{aligned}$$

Again using $\mathbf{F}$ is $C^1$,

$$\begin{aligned} &\mathbf{F}(\mathbf{y}(s,\varepsilon), s) - \mathbf{F}(\mathbf{y}_0(s), s) - D\mathbf{F}(\mathbf{y}_0(s), s)(\mathbf{y}(s,\varepsilon) - \mathbf{y}_0(s)) \\ &= \int_0^1 \frac{\partial}{\partial x} \mathbf{F}(x\mathbf{y}(s,\varepsilon) + (1-x)\mathbf{y}_0(s), s) dx - D\mathbf{F}(\mathbf{y}_0(s), s)(\mathbf{y}(s,\varepsilon) - \mathbf{y}_0(s)) \\ &= \int_0^1 [D\mathbf{F}(x\mathbf{y}(s,\varepsilon) + (1-x)\mathbf{y}_0(s), s) - D\mathbf{F}(\mathbf{y}_0(s), s)] \cdot (\mathbf{y}(s,\varepsilon) - \mathbf{y}_0(s)) dx. \end{aligned}$$

Since $\mathbf{F}$ is $C^1$, it is locally Lipschitz. By Grönwall's inequality, there exists $L > 0$ such that for $\varepsilon$ sufficiently small

$$|\mathbf{y}(s,\varepsilon) - \mathbf{y}_0(s)| \leq \varepsilon |\boldsymbol{\xi}_1| e^{L(s-t_0)} \leq \varepsilon |\boldsymbol{\xi}_1| e^{L(t_2-t_0)}, \qquad t_0 \leq s \leq t_2$$

and

$$\lim_{\varepsilon \to 0} (D\mathbf{F}(x\mathbf{y}(s,\varepsilon) + (1-x)\mathbf{y}_0(s), s) - D\mathbf{F}(\mathbf{y}_0(s), s)) = 0$$

uniformly for $t_0 \le s \le t_2$. In follows that

$$\lim_{\varepsilon \to 0} \frac{I(t,\varepsilon)}{\varepsilon} = 0 \qquad \text{uniformly for } t_0 \le t \le t_2.$$

Applying Gronwall's inequality again to $g(t,\varepsilon)$, we have

$$g(t,\varepsilon) \le \max_{t_0 \le t \le t_2} I(t,\varepsilon) e^{M(t-t_0)}, \qquad t_0 \le t \le t_2$$

$$\Longrightarrow \quad \max_{t_0 \le t \le t_2} g(t,\varepsilon) \le \max_{t_0 \le t \le t_2} I(t,\varepsilon) e^{M(t_2-t_0)}$$

It follows that

$$\lim_{\varepsilon \to 0} \frac{g(t,\varepsilon)}{\varepsilon} = 0, \qquad \text{uniformly on } t_0 \le t \le t_2.$$

□

*Remark* 3.3.5. We could also consider the case when the equation varies with parameters. As explained in Section 3.3.2, the parameter-dependence problem can be reduced to the initial value problem. This allows us to show the differentiability of the solutions with respect to the parameters by Theorem 3.3.4. We leave the details to the reader.

**Example 3.3.6.** Consider the nonlinear system

$$\begin{cases} x' = x + y^2 \\ y' = -y \end{cases}$$

The solution with initial condition $x(0) = 0, y(0) = 0$ is the equilibruim one

$$\begin{cases} x_0(t) = 0 \\ y_0(t) = 0 \end{cases}$$

The matrix $A(t) = D\mathbf{F}$ evaluated at $(x_0(t), y_0(t))$ is

$$A(t) = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

Thus the variation of the solution with respect to the initial condition at $t = 0$ along the direction $\begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix}$ is

$$\begin{pmatrix} x_1(t) \\ y_1(t) \end{pmatrix} = e^{tA} \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} = \begin{pmatrix} \lambda_1 e^t \\ \lambda_2 e^{-t} \end{pmatrix}$$

In other words, let $x(t,\lambda_i), y(t,\lambda_i)$ be the solution with initial condition $x(0) = \lambda_1, y(0) = \lambda_2$. Then the 1st order approximation is

$$x(t,\lambda_i) \simeq \lambda_1 e^t, \quad y(t,\lambda_i) \simeq \lambda_2 e^{-t}.$$

In fact, we can solve the exact solution and find

$$\begin{cases} x(t,\lambda_i) = \lambda_1 e^t + \frac{\lambda_2^2}{3}(e^t - e^{-2t}) \\ y(t,\lambda_i) = \lambda_2 e^{-t} \end{cases}$$

## 3.4 Analyticity

We have established the well-posedness of the initial value problem for the non-linear system

$$\frac{d\mathbf{y}}{dt} = \mathbf{F}(\mathbf{y}, t), \qquad \mathbf{y}(t_0) = \boldsymbol{\xi}.$$

In this section we deal with the issue of analyticity of the solutions, which is established by the celebrated Cauchy-Kovalevskaya Theorem. This is extremely useful since it allows us to study the solutions in terms of power series, in other words, order by order. We will study power series solutions in Chapter 4.

### 3.4.1 Analytic Function

We will mainly deal with real analytic functions, though the discussion extends easily to the complex analytic case. Let us first recall the notion of real analyticity.

**Definition 3.4.1.** A function $f(x)$ is real analytic on an open subset $U \subset \mathbb{R}$ if it is a smooth function and the Taylor series at any point $x_0 \in U$

$$\sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n$$

converges to $f(x)$ for $x$ in a neighborhood of $x_0$.

*Remark* 3.4.2. The definition of complex analytic function $f(z)$ is obtained by replacing, in the definition above, "$U \subset \mathbb{R}$" with an open subset "$U \subset \mathbb{C}$". It turns out that a function is complex analytic if and only if it is holomorphic.

**Example 3.4.3.** Polynomials, $e^x, \sin(x), \cos(x)$ are real analytic on $\mathbb{R}$.

**Example 3.4.4.**

$$f(x) = \begin{cases} e^{-\frac{1}{x}}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

is smooth but not analytic in any neighborhood of 0. In fact, the Taylor series of $f(x)$ at $x_0 = 0$ is identically zero.

The following is an useful equivalent description for analyticity, which we omit its proof.

**Proposition 3.4.5.** *$f$ is real analytic in $U$ if and only if $f$ is smooth and for every compact set $K \subset U$ there exist positive constants $C, r$ such that*

$$|f^{(n)}(x)| \leq C\frac{n!}{r^n}, \qquad \forall x \in K, n \geq 0.$$

### 3.4.2 Cauchy-Kovalevskaya Theorem

We first consider scalar autonomous differential equation

$$\frac{dy}{dt} = f(y), \qquad y(0) = y_0 \in \mathbb{R}.$$

**Theorem 3.4.6** (Cauchy-Kovalevskaya, ODE version)**.** *Assume $f$ is real analytic on a neighborhood $U$ of $y_0$. Let $y(t)$ be the unique solution of the initial value problem*

$$\frac{dy}{dt} = f(y), \qquad y(0) = y_0$$

*on some interval $(-\varepsilon, \varepsilon)$ for sufficiently small $\varepsilon > 0$. Then $y(t)$ is real analytic on $(-\varepsilon, \varepsilon)$.*

*Proof:* We prove via the "method of majorants". Without lost of generality, we can assume $y_0 = 0$. Otherwise we can consider the shift $y(t) \to y(t) - y_0$ and $f(y) \to f(y + y_0)$.

By repeatedly differentiating $\frac{dy}{dt} = f(y)$, we find

$$\begin{aligned}
y^{(1)} =& f(y) \\
y^{(2)} =& f^{(1)}(y)y^{(1)} = f^{(1)}(y)f(y) \\
y^{(3)} =& f^{(2)}(y)f(y)^2 + f^{(1)}(y)^2 f(y) \\
&\vdots \\
y^{(n)} =& \left(f(y)\frac{d}{dy}\right)^{n-1} f(y)
\end{aligned}$$

It is clear that there are universal (independent of $f$) polynomials $P_n$ with non-negative integer coefficients such that

$$y^{(n)} = P_n(f(y), f^{(1)}(y), \cdots, f^{(n-1)}(y)).$$

We look for a function $g$ with non-negative derivatives at zero such that

$$g^{(n)}(0) \geq |f^{(n)}(0)|, \qquad \forall n \geq 0.$$

Such a function $g$ is called a majorant function of $f$. Assume such a $g$ is chosen. We will come back to its construction at the end. Let $u(t)$ solve the initial value problem

$$\frac{du}{dt} = g(u), \qquad u(0) = 0.$$

Then we also have

$$u^{(n)} = P_n(g(u), g^{(1)}(u), \cdots, g^{(n-1)}(u)).$$

By the choice of $g$ and the fact about the nonnegative coefficients of $P_n$, we have

$$|P_n(f(0), f^{(1)}(0), \cdots, f^{(n-1)}(0))| \leq P_n(g(0), g^{(1)}(0), \cdots, g^{(n-1)}(0))$$

Thus

$$|y^{(n)}(0)| \leq |u^{(n)}(0)|, \qquad \forall n \geq 0.$$

Let us assume $u$ is analytic near zero. We will show this below for the constructed majorant function $g$. Then the power series

$$\sum_{n\geq 0} u^{(n)}(0)\frac{t^n}{n!}$$

has a positive radius of convergence. It follows that the power series

$$T(t) = \sum_{n\geq 0} y^{(n)}(0)\frac{t^n}{n!}$$

will also have a positive radius of convergence and analytic near zero. The function $T(t)$ satisfies

$$T^{(n)}(0) = y^{(n)}(0), \qquad \forall n \geq 0 \tag{$*$}$$

We next show $T(t) = y(t)$ near zero. First, we observe that $(*)$ implies (via the chain rule)

$$\left(\frac{d}{dt}\right)^n\Big|_{t=0} f(T(t)) = \left(\frac{d}{dt}\right)^n\Big|_{t=0} f(y(t)), \qquad \forall n \geq 0$$

Thus

$$\left(\frac{d}{dt}\right)^n\Big|_{t=0} f(T(t)) = \left(\frac{d}{dt}\right)^n\Big|_{t=0} y'(t) = y^{(n+1)}(0).$$

Consider the function

$$e(t) = T'(t) - f(T(t)).$$

The above calculation shows

$$e^{(n)}(0) = T^{(n+1)}(0) - \left(\frac{d}{dt}\right)^n\Big|_{t=0} f(T(t)) = y^{(n+1)}(0) - y^{(n+1)}(0) = 0, \qquad \forall n \geq 0.$$

On the other hand, since $f$ and $T$ are analytic, $e(t)$ is also analytic near zero. Then $e^{(n)}(0) = 0, \forall n$, implies $e(t) = 0$ identically near zero. Thus

$$\frac{dT}{dt} = f(T), \qquad T(0) = 0.$$

By the uniqueness, $y(t) = T(t)$ near zero, hence analytic near zero.

It remains to construct a majorant function $g$ and show that the solution $u(t)$ of

$$\frac{du}{dt} = g(u), \qquad u(0) = 0$$

is analytic near zero. From the analyticity of $f$, there exist constant $C, r > 0$, such that

$$|f^{(n)}(0)| \leq C\frac{n!}{r^n}, \qquad \forall n \geq 0.$$

Then we can simply choose

$$g(x) = C\sum_{n=0}^{\infty}\frac{x^n}{r^n} = \frac{C}{1-\frac{x}{r}}.$$

Clearly, $g^{(n)}(0) = \frac{Cn!}{r^n}$, hence

$$|f^{(n)}(0)| \leq g^{(n)}(0) \quad \text{holds} \quad \forall n \geq 0.$$

Now let us consider the solution $u(t)$ for

$$\frac{du}{dt} = g(u) = \frac{C}{1-\frac{u}{r}}, \quad u(0) = 0.$$

We can use separation of variable to solve this

$$\begin{aligned}
& (r-u)du = Crdt \\
\overset{u(0)=0}{\Rightarrow} \quad & ru - \frac{1}{2}u^2 = Crt \\
\Rightarrow \quad & u^2 - 2ru + 2Crt = 0 \\
\Rightarrow \quad & u = r \pm \sqrt{r^2 - 2rCt} \\
\overset{u(0)=0}{\Rightarrow} \quad & u = r - r\sqrt{1 - \frac{2Ct}{r}}
\end{aligned}$$

It is clear that $u(t)$ is analytic near $t = 0$. □

Now we discuss the generalization to a system

**Theorem 3.4.7** (Cauchy-Kovalevskaya, ODE version'). *Assume* $\mathbf{F}(\mathbf{y})$ *is analytic in a neighborhood of* $\boldsymbol{\xi}$. *Let* $\mathbf{y}(t)$ *be the unique solution of the initial value problem*

$$\frac{d\mathbf{y}}{dt} = \mathbf{F}(\mathbf{y}), \qquad \mathbf{y}(0) = \boldsymbol{\xi}$$

*on some interval* $(-\varepsilon, \varepsilon)$ *for sufficiently small* $\varepsilon > 0$. *Then* $\mathbf{y}(t)$ *is analytic on* $(-\varepsilon, \varepsilon)$.

*Proof:* The above proof of Theorem 3.4.6 can be modified slightly and adapted to this case. We leave this to the reader. □

*Remark* 3.4.8. In general, for the initial value problem of a nonautonomous system

$$\frac{d\mathbf{y}}{dt} = \mathbf{F}(\mathbf{y}, t), \qquad \mathbf{y}(0) = \boldsymbol{\xi}$$

we can describe it equivalently by

$$\begin{cases} \frac{d\mathbf{y}}{dt} = \mathbf{F}(\mathbf{y}, z) \\ \frac{dz}{dt} = 1 \end{cases} \qquad \mathbf{y}(0) = \boldsymbol{\xi}, z(0) = 0.$$

This will be reduced to Theorem 3.4.7.

*Remark* 3.4.9. For the ODE version of Cauchy-Kovalevskaya Theorem, the above proof via the method of majonants is not the simplest one. However, this method can be used in the PDE version as well. Nevertheless, the proof via method of majonants is clearly beautiful.

# Chapter 4 Power Series Solutions

In this chapter we study a general linear ODE

$$a_0(t)y^{(n)} + a_1(t)y^{(n-1)} + \cdots + a_n(t)y + b(t) = 0$$

where the coefficients $a_i(t), b(t)$ are analytic on certain domain of our interest. We can write the above equation as

$$y^{(n)} + p_1(t)y^{(n-1)} + p_2(t)y^{(n-2)} + \cdots + p_n(t)y + q(t) = 0$$

where $p_i(t) = \frac{a_i(t)}{a_0(t)}$ and $q(t) = \frac{b(t)}{a_0(t)}$. We say

- $t_0$ is an ordinary point of the equation if all $p_i(t)$ and $q(t)$ are analytic near $t = t_0$.
- otherwise, $t_0$ is a singular point of the equation.

Similarly, we will consider a linear system

$$\frac{d\mathbf{y}}{dt} = A(t) \cdot \mathbf{y} + \mathbf{b}(t)$$

where entries of the matrix $A(t)$ and the vector $\mathbf{b}(t)$ are quotients of analytic functions on certain domain of our interest. We say

- $t_0$ is an ordinary point of the system if all entries of $A(t)$ and $\mathbf{b}(t)$ are analytic near $t = t_0$.
- otherwise, $t_0$ is a singular point of the system.

The above two definitions are consistent. It is clear that $t_0$ is an ordinary point of a linear ODE if and only if $t_0$ is an ordinary point of the associated 1st-order linear system.

The goal of this chapter is to understand solutions of linear ODE around certain points of interest (ordinary or singular) in terms of power series.

## 4.1 Ordinary Point

Let us start with a linear ODE

$$y^{(n)} + p_1(t)y^{(n-1)} + p_2(t)y^{(n-2)} + \cdots + p_n(t)y + q(t) = 0$$

around an ordinary point $t = t_0$. By assumption, $p_i(t)$ and $q(t)$ are analytic near $t = t_0$. By Cauchy-Kovalevskaya Theorem (Remark 3.4.8), solutions are analytic near $t = t_0$, and therefore can be written as a convergent power series

$$y(t) = \sum_{n=0}^{\infty} c_n (t - t_0)^n.$$

Substituting this back to the equation will allow us to solve the coefficients $\{c_n\}$ recursively order by order. We look at some examples to illustrate.

**Example 4.1.1.** Consider the initial value problem

$$(1 + t)y' = py, \qquad y(0) = 1$$

where $p$ is an arbitrary constant.

This equation has analytic coefficients, and $t = 0$ is an ordinary point. The unique solution is analytic near $t = 0$ and has a power series expansion

$$y(t) = c_0 + c_1 t + c_2 t^2 + \cdots + c_n t^n + \cdots$$

The initial condition sets $c_0 = 1$. Substituting the power series into the equation, we find

$$\sum_{n=0}^{\infty} ((n + 1)c_{n+1} + nc_n)t^n = \sum_{n=0}^{\infty} pc_n t^n$$

Comparing the two sides, we find

$$(n + 1)c_{n+1} + nc_n = pc_n, \qquad n = 0, 1, \cdots$$
$$\Rightarrow \qquad c_{n+1} = \frac{p - n}{n + 1} c_n.$$

Recursively,

$$c_n = \frac{p(p - 1)(p - 2) \cdots (p - n + 1)}{n!}, \qquad \forall n \geq 1.$$

Therefore the solution is given by

$$y(t) = 1 + pt + \frac{p(p - 1)}{2!} t^2 + \cdots + \frac{p(p - 1) \cdots (p - n + 1)}{n!} t^n + \cdots$$

Note if $p$ is a nonnegative integer, then the series becomes a polynomial of degree $p$.

On the other hand, we can explicitly solve the equation and find

$$y(t) = (1 + t)^p.$$

It is clear that the above found series is precisely the Taylor series of $(1 + t)^p$ centered at $t = 0$.

**Example 4.1.2 (Legendre's Equation).**

$$(1 - t^2)y'' - 2ty' + p(p + 1)y = 0$$

where $p$ is a constant.

We can write the equation as

$$y'' - \frac{2t}{1-t^2}y' + \frac{p(p+1)}{1-t^2}y = 0.$$

It is clear that the coefficients $\frac{2t}{1-t^2}$ and $\frac{p(p+1)}{1-t^2}$ are analytic near $t = 0$. The origin is an ordinary point and therefore we expect power series solutions

$$y(t) = \sum_{n=0}^{\infty} a_n t^n \qquad \text{near } t = 0.$$

Substituting this power series into the Legendre's equation, we find

$$(n+1)(n+2)a_{n+2} - n(n-1)a_n - 2na_n + p(p+1)a_n = 0, \qquad \forall n \geq 0$$

$$a_{n+2} = -\frac{(p-n)(p+n+1)}{(n+1)(n+2)}a_n$$

We have two free parameters $a_0, a_1$, which are to determined by the initial condition. In fact,

$$a_0 = y(0), \qquad a_1 = y'(0).$$

All other $a_n$'s are expressed via $a_0$ and $a_1$ in terms of the above recursion formula

$$\begin{aligned}
a_2 =& -\frac{p(p+1)}{1\cdot 2}a_0 \\
a_3 =& -\frac{(p-1)(p+2)}{2\cdot 3}a_1 \\
a_4 =& \frac{p(p-2)(p+1)(p+3)}{4!}a_0 \\
a_5 =& \frac{(p-1)(p-3)(p+2)(p+4)}{5!}a_1 \\
&\vdots
\end{aligned}$$

The power series solution is given by

$$\begin{aligned}
y(t) =& a_0\left[1 - \frac{p(p+1)}{2!}t^2 + \frac{p(p-2)(p+1)(p+3)}{4!}t^4 - \frac{p(p-2)(p-4)(p+1)(p+3)(p+5)}{6!}t^6 + \cdots\right] \\
&+ a_1\left[t - \frac{(p-1)(p+2)}{3!}t^3 + \frac{(p-1)(p-3)(p+2)(p+4)}{5!}t^5 \right. \\
&\left. - \frac{(p-1)(p-3)(p-5)(p+2)(p+4)(p+6)}{7!}t^7 + \cdots\right]
\end{aligned}$$

- When $p$ is not an integer. Each series in brackets has radius of convergence $R = 1$. This can be proved by ratio test (show this). The functions defined by the above power series are called Legendre functions. One important feature is that these functions are in general not elementary functions, and so can not be expressed via finite compositions of rational, trigonometric, hyperbolic and exponential function.

- When $p = n$ is an nonnegative integer, one of the series in the bracket terminates and is thus a polynomial. The polynomial $P_n(t)$ of degree $n$ satisfying the Legendre equation

$$(1-t^2)y'' - 2ty' + n(n+1)y = 0$$

with $P_n(1) = 1$ is called Legendre polynomial. Explicity they are given by

$$P_n(t) = \frac{1}{2^n n!}\frac{d^n}{dt^n}(t^2-1)^n, \qquad n = 0, 1, 2, \cdots$$

**Example 4.1.3** (**Airy Equation**).

$$y'' - ty = 0$$

The origin is an ordinary point, and we look for power series solutions

$$y(t) = \sum_{n=0}^{\infty} c_n t^n \qquad \text{near } t = 0.$$

Substituting into the Airy equation

$$\sum_{n=0}^{\infty}(n+2)(n+1)c_{n+2}t^n - \sum_{n=1}^{\infty} c_{n-1}t^n = 0,$$

we obtain the recursion relation

$$c_2 = 0, \qquad c_{n+3} = \frac{c_n}{(n+3)(n+2)}, n \geq 0.$$

The general solution of Airy's Equation is

$$\begin{aligned} y(t) =& c_0\left[1 + \frac{t^3}{2\cdot 3} + \frac{t^6}{2\cdot 3\cdot 5\cdot 6} + \cdots + \frac{t^{3n}}{2\cdot 3\cdot\cdots\cdot(3n-1)(3n)} + \cdots\right] \\ &+ c_1\left[t + \frac{t^4}{3\cdot 4} + \frac{t^7}{3\cdot 4\cdot 6\cdot 7} + \cdots + \frac{t^{3n+1}}{3\cdot 4\cdot\cdots\cdot(3n)(3n+1)} + \cdots\right] \end{aligned}$$

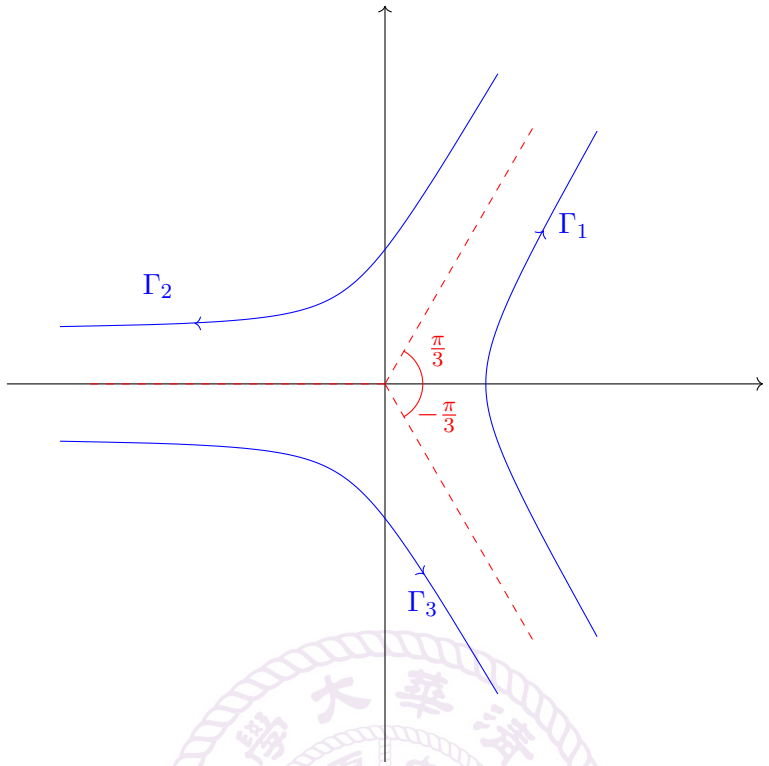
The series in each bracket converges for all $t$, which can be shown via the ratio test. These functions are called Airy functions.

One important expression for Airy function is through the integral in the complex plane

$$y(t) = \frac{1}{2\pi i}\int_{\Gamma} e^{\left(\frac{z^3}{3} - tz\right)} dz$$

where $\Gamma$ is a contour in the complex plane starting and ending at $\infty$ along directions such that

$\mathrm{Re}(\frac{z^3}{3} - tz) \to -\infty$. Here below are three possible choices of $\Gamma$




Each $\Gamma_i$ will lead to a solution to the Airy equation

$$y_i(t) = \frac{1}{2\pi i}\int_{\Gamma_i} e^{\left(\frac{z^3}{3}-tz\right)}dz.$$

However, they are not independent. We have

$$y_1(t) + y_2(t) + y_3(t) = \frac{1}{2\pi i}\int_{\Gamma_1+\Gamma_2+\Gamma_3} e^{\left(\frac{z^3}{3}-tz\right)}dz = 0$$

since $\Gamma_1+\Gamma_2+\Gamma_3$ essentially forms a closed loop in the complex plane and hence can be shrinked to zero without varying the value of the complex integral, by Cauchy integral formula.

It turns out that any two of $y_1(t), y_2(t), y_3(t)$ form the two linearly independent solutions of the Airy function. Let us check that Airy's equation holds.

$$y_i''(t) - ty_i(t) = \frac{1}{2\pi i}\int_{\Gamma_i}(z^2 - t)e^{\frac{z^3}{3}-tz}dz = \frac{1}{2\pi i}\int_{\Gamma_i} d\left(e^{\frac{z^3}{3}-tz}\right) = 0.$$

Here the choice of $\Gamma_i$ guarantees that the boundary in the above integral does not contribute.

## 4.2 Linear System with Regular Singularity

### 4.2.1 Regular Singular Point

We consider the linear system

$$\frac{d\mathbf{y}}{dt} = A(t)\mathbf{y}$$

in the region where the matrix $A(t)$ is continuous. This equation can be explicitly solved via the path-ordered exponential. Moreover, when $A(t)$ is analytic, solutions will also be analytic.

In this section, we study the case when $A(t)$ could have singularities. We consider the simplest type of analytic singularity defined as follows.

**Definition 4.2.1.** The point $t = t_0$ is called a regular singular point or a singularity of the first kind for the linear system $\frac{d\mathbf{y}}{dt} = A(t)\mathbf{y}$ if

$$A(t) = \frac{B(t)}{t - t_0}$$

where $B(t)$ is analytic near $t = t_0$.

So at a regular singular point, the coefficient $A(t)$ has a pole of at most order $= 1$. Without lost of generality, we assume the regular singularity is at $t_0 = 0$. In this case, we can write

$$A(t) = \sum_{k \geq -1} A_k t^k$$

where $A_k$'s are $n \times n$ matrices and $A_{-1}$ represents the coefficient of the pole. Our aim is to understand how the solutions behave near such a regular singularity.

Before we discuss the general theory, let us first look at a few examples.

**Example 4.2.2.**

$$y' = \frac{b}{t} y, \qquad b \in \mathbb{R} \text{ is a constant.}$$

This equation can be solved using separation of variables

$$\Rightarrow \qquad y(t) = ct^b$$

where $c$ is a constant.

**Example 4.2.3.**

$$\mathbf{y}' = \frac{1}{t} B\mathbf{y}, \qquad B \text{ is a constant } n \times n \text{ matrix.}$$

This can be solved in a similar fashion formally by

$$\mathbf{y}(t) = t^B \cdot \boldsymbol{\xi}$$

where $\boldsymbol{\xi}$ is a constant column vector.

Since $B$ is a matrix, we need to clarify the meaning of $t^B$. Precisely, we can define $t^B$ via the exponential matrix

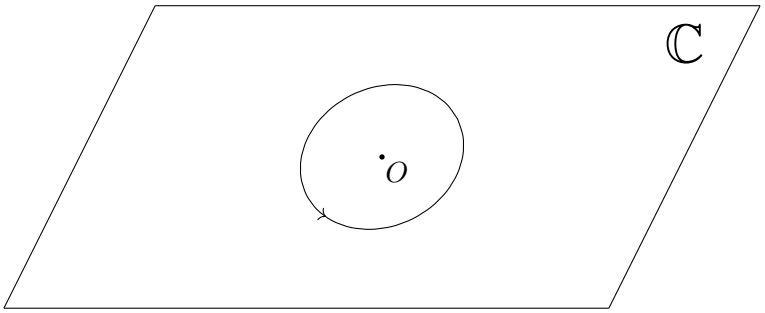
$$t^B := e^{B \ln t}.$$

This expression is well-defined for $t > 0$. In the region $t > 0$, we have

$$\frac{d}{dt} e^{B \ln t} = \frac{1}{t} B e^{B \ln t}$$

thus $\mathbf{y}(t) = e^{B \ln t} \cdot \boldsymbol{\xi}$ are indeed solutions.

In general, it is better to think about $t$ as a variable in the complex plane $\mathbb{C}$. Then these solutions will have branches of their defining domain, and analytic continuation around the origin will lead to a transformation of solutions.




$$\ln t \mapsto \ln t + 2\pi i$$

Indeed, when $t$ goes around the origin, the function $\ln t$ undergoes a transformation to $\ln t + 2\pi i$. Hence the matrix $e^{B \ln t}$ will change by

$$e^{B \ln t} \mapsto e^{B \ln t} e^{2\pi i B}$$

The transformation $e^{2\pi i B}$ is called the monodromy.

We will not go much into the complex analytic perspective, but instead work with the real domain for $t > 0$ in the forward time in our current context. Let us look closely at the matrix function $e^{B \ln t}$. There are essentially two main cases of building blocks.

① Assume $B$ is diagonalizable and

$$B = P \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} P^{-1}$$

Then

$$e^{B \ln t} = P \begin{pmatrix} e^{\lambda_1 \ln t} & & \\ & \ddots & \\ & & e^{\lambda_n \ln t} \end{pmatrix} P^{-1} = P \begin{pmatrix} t^{\lambda_1} & & \\ & \ddots & \\ & & t^{\lambda_n} \end{pmatrix} P^{-1}.$$

Thus the solutions are of the form

$$\mathbf{y}(t) = P \begin{pmatrix} t^{\lambda_1} & & \\ & \ddots & \\ & & t^{\lambda_n} \end{pmatrix} P^{-1} \boldsymbol{\xi}.$$

Equivalently, if we change variables and define

$$\mathbf{z}(t) = P^{-1} \mathbf{y}(t)$$

then the equation becomes

$$\frac{d}{dt}\mathbf{z}(t) = \frac{1}{t}\begin{pmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_n\end{pmatrix}\mathbf{z}(t)$$

which is reduced to $n$ scalar equations

$$z_i'(t) = \frac{\lambda_i}{t}z_i(t), \qquad \mathbf{z}(t) = \begin{pmatrix}z_1(t)\\ \vdots \\ z_n(t)\end{pmatrix}.$$

② Assume $B$ is of the form of a Jordan block

$$B = \begin{pmatrix}\lambda & 1 & & \\ & \lambda & \ddots & \\ & & \ddots & 1 \\ & & & \lambda\end{pmatrix} = \lambda I_n + N, \qquad N = \begin{pmatrix}0 & 1 & & \\ & 0 & \ddots & \\ & & \ddots & 1 \\ & & & 0\end{pmatrix}.$$

Then

$$\begin{aligned} e^{B\ln t} =& e^{\lambda I_n \ln t + N\ln t} = t^\lambda e^{N\ln t} = t^\lambda \sum_{k=0}^{n-1}\frac{N^k}{k!}(\ln t)^k \\ =& \begin{pmatrix} t^\lambda & t^\lambda \ln t & \cdots & t^\lambda\frac{(\ln t)^k}{k!} & \cdots & t^\lambda\frac{(\ln t)^{n-1}}{(n-1)!} \\ & t^\lambda & \cdots & t^\lambda\frac{(\ln t)^{k-1}}{(k-1)!} & \cdots & t^\lambda\frac{(\ln t)^{n-2}}{(n-2)!} \\ & & \ddots & & & \\ & & & \ddots & & \\ & & & & t^\lambda & t^\lambda \ln t \\ 0 & & & & & t^\lambda \end{pmatrix}. \end{aligned}$$

This matrix will give the explicit form of the solution $\mathbf{y}(t) = e^{B\ln t}\boldsymbol{\xi}$. Note that

$$\lim_{t\to 0^+} e^{B\ln t} = \begin{cases} 0 & \text{if } \lambda > 0 \\ \text{blow-up} & \text{if } \lambda \le 0 \end{cases}$$

The above two examples illustrate the main features of solutions near a regular singular point. The general case can be essentially reduced to the above as we next show.

### 4.2.2 Gauge Transformation

Let us consider a change of variables

$$\mathbf{y}(t) \to \mathbf{y}(t) = P(t)\mathbf{z}(t)$$

where entries of the matrix $P(t)$ are analytic functions near $t = 0$ and $\det P(t) \neq 0$ near $t = 0$. Thus $P(t)$ is invertible near $t = 0$ and its inverse $P(t)^{-1}$ also consists of analytic entries.

Under this change of variables, the equation becomes

$$\begin{aligned} & \frac{d}{dt}(P(t)\mathbf{z}(t)) = A(t)P(t)\mathbf{z}(t) \\ \Leftrightarrow \qquad & \frac{d}{dt}\mathbf{z}(t) = \left(P^{-1}(t)A(t)P(t) - P^{-1}(t)\frac{d}{dt}P(t)\right)\mathbf{z}(t). \end{aligned}$$

This new equation is of the same form as before but changes

$$A(t) \mapsto P^{-1}(t)A(t)P(t) - P^{-1}(t)\frac{d}{dt}P(t)$$

This will be called a gauge transformation.

Two linear systems, whose coefficient matrices are related by gauge transformations, are equivalent under the corresponding change of variables as above.

**Proposition 4.2.4.** *Assume $tA(t)$ is analytic near $t = 0$. Let $A(t) = \sum_{k\geq -1} A_k t^k$ and assume no eigenvalues of $A_{-1}$ differ by positive integers. Then there exists a gauge transformation $P(t)$ such that*

$$P^{-1}(t)A(t)P(t) - P^{-1}(t)\frac{d}{dt}P(t) = \frac{A_{-1}}{t}.$$

*Proof:* Let us first look for a formal power series

$$P(t) = \sum_{k\geq 0} P_k t^k$$

that will do the job. Then we show its convergence. Plugging the above series into the equation

$$A(t)P(t) - \frac{d}{dt}P(t) = \frac{P(t)A_{-1}}{t}$$

and comparing $t$-orders of both sides, we need to solve

$$A_{-1}P_0 = P_0A_{-1}$$

$$(A_{-1} - kI_n)P_k - P_kA_{-1} = -\sum_{i=0}^{k-1} A_iP_{k-1-i}, \qquad k > 0.$$

We can solve the first equation by simply choosing

$$P_0 = I_n.$$

By assumption, $A_{-1} - kI_n$ and $A_{-1}$ have no common eigenvalues for any $k > 0$. Then by Lemma 4.2.5 below, $P_k$ can be uniquely determined for $k > 0$. Thus we have found the required formal series $P(t) = \sum_{k\geq 0} P_k t^k$.

We next show $P(t)$ is analytic near $t = 0$. By the above recursive relation

$$kP_k = A_{-1}P_k - P_kA_{-1} + \sum_{i=0}^{k-1} A_iP_{k-1-i}.$$

Taking the operator norm of both sides, we have

$$k\|P_k\| \leq 2\|A_{-1}\|\|P_k\| + \sum_{i=0}^{k-1} \|A_i\|\|P_{k-1-i}\|.$$

Let $N$ be a positive integer such that

$$2\|A_{-1}\| \leq N - 1.$$

Then for $k \geq N$, we have

$$\|P_k\| \leq \sum_{i=0}^{k-1} \|A_i\|\|P_{k-1-i}\|.$$

Let us define

$$u_k = \begin{cases} \|P_k\| & k < N \\ \sum_{i=0}^{k-1} \|A_i\| u_{k-1-i} & k \geq N \end{cases}$$

Then we have $\|P_k\| \leq u_k$ for any $k \geq 0$. It is enough to show that the power series

$$u(t) = \sum_{k=0}^{\infty} u_k t^k$$

is convergent near $t = 0$. Let

$$a(t) = \sum_{k \geq 0} \|A_k\| t^k$$

By assumption, $a(t)$ is convergent near $t = 0$. By construction, we have

$$u(t) = ta(t)u(t) + f(t)$$

where $f(t)$ is a polynomial of deg $\leq N - 1$. Therefore

$$u(t) = \frac{f(t)}{1 - ta(t)}$$

which is clearly analytic near $t = 0$. □

**Lemma 4.2.5.** *Let $M_n(\mathbb{C})$ denote the space of $n \times n$ complex matrices. Let $U, V \in M_n(\mathbb{C})$ without common eigenvalue. Then the linear map*

$$\begin{aligned} M_n(\mathbb{C}) &\to M_n(\mathbb{C}) \\ X &\mapsto XU - VX \end{aligned}$$

*is an isomorphism*

*Proof:* Exercise. □

### 4.2.3 Solutions in General

Now we discuss how to solve

$$\frac{d\mathbf{y}}{dt} = A(t)\mathbf{y}$$

in general around the regular singular point $t = 0$. Let

$$A(t) = \sum_{k\geq -1} A_k t^k$$

① If no eigenvalues of $A_{-1}$ differ by positive integers. Then by Proposition 4.2.4, we can find a gauge transformation $\mathbf{y}(t) = P(t)\mathbf{z}(t)$ such that $\mathbf{z}(t)$ satisfies

$$\frac{d\mathbf{z}}{dt} = \frac{A_{-1}}{t}\mathbf{z}(t).$$

Then the solutions can be found by Example 4.2.3 and

$$\mathbf{y}(t) = P(t)e^{A_{-1}\ln t}\boldsymbol{\xi}$$

where $\boldsymbol{\xi}$ is some constant vector.

② If there are eigenvalues of $A_{-1}$ which differ by positive integers. Assume $A_{-1}$ is of the Jordan form

$$\begin{pmatrix} \begin{pmatrix} \lambda & * & & \\ & \lambda & \ddots & \\ & & \ddots & * \\ & & & \lambda \end{pmatrix} & & \\ & \begin{pmatrix} \lambda+k & * & & \\ & \lambda+k & \ddots & \\ & & \ddots & * \\ & & & \lambda+k \end{pmatrix} & \\ & & \ddots \end{pmatrix}$$

where $k > 0$ is an integer. We can consider the change of variables

$$\mathbf{y}(t) = Q(t)\mathbf{z}(t)$$

where

$$Q(t) = \begin{pmatrix} \begin{pmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{pmatrix} & & \\ & \begin{pmatrix} t & & \\ & \ddots & \\ & & t \end{pmatrix} & \\ & & \ddots \end{pmatrix}$$

Note that $Q(t)$ is not a gauge transformation since $Q(t)^{-1}$ will be singular at $t = 0$. Nevertheless, we can use this to derive an equation for $\mathbf{z}(t)$

$$\frac{d\mathbf{z}(t)}{dt} = B(t)\mathbf{z}(t)$$

where

$$B(t) = Q^{-1}(t)A(t)Q(t) - Q^{-1}(t)\frac{d}{dt}Q(t).$$

Observe $B(t)$ is also regular singular. Let $B(t) = \sum_{k\geq -1} B_k t^k$. Then $B_{-1}$ is of the form

$$B_{-1} = \begin{pmatrix} \begin{pmatrix} \lambda & * & & \\ & \lambda & \ddots & \\ & & \ddots & * \\ & & & \lambda \end{pmatrix} & & 0 & & 0 \\ & & & & \\ & * & & & \\ & & & & \\ 0 & & 0 & & \ddots \end{pmatrix}$$

$$\begin{pmatrix} \lambda+k-1 & * & & \\ & \lambda+k-1 & \ddots & \\ & & \ddots & * \\ & & & \lambda+k-1 \end{pmatrix}$$

The difference of the corresponding eigenvalues decreases by one. Now we can transform $B_{-1}$ into a Jordan form and repeat the above process. Eventually we will arrive at the situation in ①.

## 4.3 Scalar Equation with Regular Singularity

### 4.3.1 Regular Singular Point

We next consider the scalar linear equation of order $n$

$$y^{(n)} + p_1(t)y^{(n-1)} + \cdots + p_n(t)y = 0 \qquad (*)$$

near a singular point.

**Definition 4.3.1.** We say $t = t_0$ is a regular singularity of $(*)$ if $t = t_0$ is a singular point and

$$(t - t_0)^k p_k(t)$$

is analytic near $t = t_0$ for any $k$. A singular point that is not regular is called irregular.

In other word, at the regular singular point, $p_k(t)$ may develop a pole but the pole order is at most $k$ for $p_k(t)$. This definition is related to the regular singularity of 1st-order linear system as follows.

Assume $t = 0$ is a regular singular point of $(*)$. Define

$$z_i(t) = t^{i-1}y^{(i-1)}, \qquad i = 1, 2, \cdots, n.$$

We have

$$tz_i' = (i-1)z_i + z_{i+1}, \qquad i = 1, 2, \cdots, n-1.$$

The above scalar linear equation then becomes the 1st-order linear system

$$\frac{d}{dt}\begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix} = \frac{1}{t}\begin{pmatrix} 0 & 1 & & & & \\ & 1 & 1 & & & \\ & & 2 & 1 & & \\ & & & \ddots & \ddots & \\ & & & & n-2 & 1 \\ -t^n p_n(t) & \cdots & \cdots & \cdots & -t^2 p_2(t) & n-1-tp_1(t) \end{pmatrix}\begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix}$$

Now it is clear that $t = 0$ is a regular singular point of $(*)$ if and only if $t = 0$ is a regular singular point of the above 1st-order linear system. Therefore these two notions of regular singularity are consistent. This relation will also help us to understand the solution near a regular singular point of $(*)$ via the result of Section 4.2. Let

$$A(t) = \frac{1}{t}\begin{pmatrix} 0 & 1 & & & & \\ & 1 & 1 & & & \\ & & 2 & 1 & & \\ & & & \ddots & \ddots & \\ & & & & n-2 & 1 \\ -t^n p_n(t) & \cdots & \cdots & \cdots & -t^2 p_2(t) & n-1-tp_1(t) \end{pmatrix} = \sum_{k \geq -1} A_k t^k$$

The singular part of $A(t)$ corresponds to

$$A_{-1} = \begin{pmatrix} 0 & 1 & & & & \\ & 1 & 1 & & & \\ & & 2 & 1 & & \\ & & & \ddots & \ddots & \\ & & & & n-2 & 1 \\ -a_n & \cdots & \cdots & \cdots & -a_2 & n-1-a_1 \end{pmatrix}$$

where

$$a_k = \lim_{t \to 0} t^k p_k(t), \qquad k = 1, 2, \cdots, n.$$

The characteristic equation of $A_{-1}$ is

$$\det(\lambda I_n - A_{-1}) = 0$$

*i.e.*

$$\lambda(\lambda-1)\cdots(\lambda-n+1) + \lambda(\lambda-1)\cdots(\lambda-n+2)a_1$$
$$+ \lambda(\lambda-1)\cdots(\lambda-n+3)a_2 + \cdots + \lambda(\lambda-1)a_{n-2} + \lambda a_{n-1} + a_n = 0$$

This equation is called the indicial equation of $(*)$ at a regular singular point $t = 0$.

If roots of the indicial equation do not differ by positive integers, then our analysis of linear system with a regular singularity immediately yields the structure of the solutions of the linear equation $(*)$ near a regular singularity.

**Example 4.3.2.** Legendre's equation

$$(1-t^2)y'' - 2ty' + p(p+1)y = 0, \qquad p \text{ is a constant}$$

has regular singularities at $t = \pm 1$.

**Example 4.3.3.** Euler's equation

$$t^2y'' + \alpha ty' + \beta y = 0, \qquad \alpha, \beta \text{ are constants}$$

has a regular singularity at $t = 0$.

**Example 4.3.4.** Bessel's Equation

$$t^2y'' + ty' + (t^2 - \alpha^2)y = 0, \qquad \alpha \text{ is a constant}$$

has a regular singularity at $t = 0$.

**Example 4.3.5.** Hypergeometric Equation

$$t(1-t)y'' + [c - (a+b+1)t]y' - aby = 0, \qquad a, b, c \text{ are constants}$$

has regular singularities at $t = 0$ and $t = 1$.

### 4.3.2 Method of Frobenius

Let us focus on linear equations of order two

$$y'' + p(t)y' + q(t)y = 0$$

with a regular singularity at $t = 0$. Let

$$tp(t) = \sum_{n=0}^{\infty} p_n t^n, \qquad t^2 q(t) = \sum_{n=0}^{\infty} q_n t^n$$

which are analytic near $t = 0$ by assumption.

The method of Frobenius looks for a series solution of the form

$$y = t^m \sum_{n=0}^{\infty} a_n t^n = \sum_{n=0}^{\infty} a_n t^{n+m}$$

where we require $a_0 \neq 0$. Plugging the above series ansatz for $y$ into the equation, we find

$$(n+m)(n+m-1)a_n + \sum_{k=0}^{n}((m+k)a_k p_{n-k} + a_k q_{n-k}) = 0, \qquad \forall n \geq 0.$$

Equivalently, we can write this as a recursion relation

$$((n+m)(n+m-1) + (n+m)p_0 + q_0)a_n = -\sum_{k=0}^{n-1} a_k[(m+k)p_{n-k} + q_{n-k}].$$

Recall that the indicial polynomial of our second order equation is

$$f(\lambda) = \lambda(\lambda - 1) + \lambda p_0 + q_0.$$

Then the above recursion relation becomes

$$f(n+m)a_n = -\sum_{k=0}^{n-1} a_k[(m+k)p_{n-k} + q_{n-k}].$$

For $n = 0$, this relation gives

$$f(m)a_0 = 0 \qquad \Rightarrow \qquad f(m) = 0.$$

In other words, $m$ is a root of $f(\lambda)$, *i.e.* , a solution of the indicial equation. This is compatible with our discussion in Section 4.2.

For $n > 0$, assume $f(n+m) \neq 0$ for any $n > 0$, *i.e.* , the other root of $f(\lambda)$ is not of the form $m+n$ for a positive integer $n$. Then we can solve $a_n$'s recursively for all $n > 0$, and obtain a series solution

$$y(t) = t^m \sum_{n=0}^{\infty} a_n t^n.$$

By our general discussion in linear system, the series $\sum_{n=0}^{\infty} a_n t^n$ will be analytic near $t = 0$. This is the series solution in Frobenius form, or Frobenius series.

Now we can summarize the above discussion as follows.

① Assume $f(\lambda)$ has two distinct roots $m_1, m_2$ such that $m_1 - m_2 \notin \mathbb{Z}$. Then we find two Frobenius series solutions of the form

$$t^{m_1} \sum_{n=0}^{\infty} a_n t^n, \qquad t^{m_2} \sum_{n=0}^{\infty} b_n t^n.$$

② Assume $f(\lambda)$ has two distinct roots $m_1 < m_2$ and $m_2 - m_1$ is a positive integer. Then we have at least one Frobenius series solution of the form

$$t^{m_2} \sum_{n=0}^{\infty} b_n t^n.$$

We may or may not have another Frobenius series solution $t^{m_1} \sum_{n=0}^{\infty} a_n t^n$. It depends on the solvability of $a_{m_2-m_1}$. If in the recursive relation

$$\sum_{k=0}^{m_2-m_1-1} a_k[(m_1+k)p_{m_2-m_1-k} + q_{m_2-m_1-k}] = 0$$

then we can set $a_{m_2-m_1} = 0$ and continue to find a second Frobenius series solution. If the above is not zero, then there can not exist a second Frobenius series solution.

③ Assume $f(\lambda)$ has two roots $m_1 = m_2$. Then we have only one Frobenius series solution.

### 4.3.3 Hypergeometric Series

We consider the example of hypergeometric equation

$$t(1-t)y'' + [c-(a+b+1)t]y' - aby = 0$$

where $a, b, c$ are constants. We write it as

$$y'' + p(t)y' + q(t)y = 0$$

where

$$p(t) = \frac{c-(a+b+1)t}{t(1-t)}, \qquad q(t) = -\frac{ab}{t(1-t)}.$$

The points $t = 0$ and $t = 1$ are regular singular points. We consider near $t = 0$. Then

$$tp(t) = \frac{c-(a+b+1)t}{1-t} = \sum_{n=0}^{\infty} p_n t^n$$

$$t^2 q(t) = -\frac{abt}{1-t} = \sum_{n=0}^{\infty} q_n t^n$$

where

$$p_n = c-(a+b+1) \quad n \geq 1, \qquad and \qquad p_0 = c,$$

$$q_n = -ab \quad n \geq 1, \qquad and \qquad q_0 = 0.$$

The indicial equation is

$$\lambda(\lambda-1) + \lambda c = 0$$

whose roots are

$$m_1 = 0, \qquad m_2 = 1-c.$$

If $1-c$ is not a positive integer, then we have a Frobenius series solution of the form

$$y = t^{m_1} \sum_{n=0}^{\infty} a_n t^n = \sum_{n=0}^{\infty} a_n t^n.$$

Plugging this into the hypergeometric equation

$$t(1-t)\left(\sum_{n=0}^{\infty} a_n t^n\right)'' + [c-(a+b+1)t]\left(\sum_{n=0}^{\infty} a_n t^n\right)' - ab\left(\sum_{n=0}^{\infty} a_n t^n\right) = 0$$

and comparing with the coefficient of $t^n$, we find

$$[-n(n-1)a_n + (n+1)na_{n+1}] + [c(n+1)a_{n+1} - (a+b+1)na_n] - aba_n = 0$$

$$\Rightarrow \qquad a_{n+1} = \frac{(n+a)(n+b)}{(n+1)(n+c)} a_n.$$

We simply set $a_0 = 1$ and find the Frobenius series solution

$$y(t) = 1 + \sum_{n=1}^{\infty} \frac{a(a+1)\cdots(a+n-1)b(b+1)\cdots(b+n-1)}{n!c(c+1)\cdots(c+n-1)} t^n$$

This is known as the hypergeometric series and we denote it by $F(a, b, c, t)$. It generalizes the familar geometric series which corresponds to

$$F(1, b, b, t) = \frac{1}{1-t}.$$

If $1 - c$ is not an integer, then we will have a second Frobenius series solution of the form

$$y = t^{1-c} \sum_{n=0}^{\infty} b_n t^n.$$

We can also see this by a change of variable

$$y = t^{1-c} z$$

Plugging this into the hypergeometric equation, we find

$$t(1-t)z'' + [(2-c) - ((a-c+1) + (b-c+1) + 1)t]z' - (a-c+1)(b-c+1)z = 0.$$

This is again of the form of hypergeometric equation. We have a Frobenius series solution by

$$z = F(a-c+1, b-c+1, 2-c, t).$$

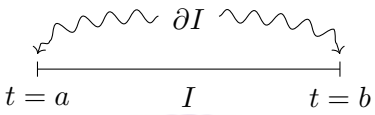
Therefore when $c$ is not an integer, a general solution of the hypergeometric equation is

$$y = \alpha F(a, b, c, t) + \beta t^{1-c} F(a-c+1, b-c+1, 2-c, t)$$

where $\alpha, \beta$ are constants.

# Chapter 5 Boundary Value Problem

We have so far focused on initial value problems for solving differential equations with specified data at a single point. Now we move on to discuss the boundary value problem. This is to study solutions of differential equations defined on the interval $I = [a, b]$ with prescribed data at both boundary endpoints $t = a$ and $t = b$.




## 5.1 Boundary Value Problem for Second Order Equations

We start with boundary value problem for second order linear equations

$$a_0(t)y'' + a_1(t)y' + a_2(t)y = g(t) \qquad (*)$$

with continuous coefficients on the interval $I = [a, b]$ such that $a_0(t) \neq 0$ for $t \in I$. These include a large class of important equations with numerous applications in science and engineering.

### 5.1.1 Boundary Conditions

These are three most common types of boundary conditions for the equation $(*)$

(i) Dirichlet boundary conditions (boundary conditions of the first kind)

$$y(a) = \xi_1, \qquad y(b) = \xi_2$$

(ii) Neumann boundary conditions (boundary conditions of the second kind)

$$y'(a) = \xi_1, \qquad y'(b) = \xi_2$$

(iii) Robin boundary conditions (boundary conditions of the third kind)

$$\alpha_1 y(a) + \beta_1 y'(a) = \xi_1, \qquad \alpha_2 y(b) + \beta_2 y'(b) = \xi_2$$

In contrast to the initial value problem where general existence and uniqueness are established, solutions of boundary value problem may not exist or may not be unique.

Consider the simple example

$$y'' = 0.$$

The general solution is $y(t) = c_1 + c_2 t$ where $c_i$ are constants. The Dirichlet boundary condition can be always uniquely solved by $y = \xi_1 \frac{(t-b)}{a-b} + \xi_2 \frac{(t-a)}{b-a}$. The Neumann boundary condition has no solution if $\xi_1 \neq \xi_2$ and infinitely many solutions if $\xi_1 = \xi_2$ by $y = c_1 + \xi_1 t$.

### 5.1.2 Sturm-Liouville Form

We can always rewrite equation $(*)$ into the form

$$\frac{d}{dt}\left(p(t)\frac{d}{dt}y\right) + q(t)y = f(t) \qquad (**)$$

This equation $(**)$ is called in the Sturm-Liouville form, or the self-adjoint form. The operator

$$L = \frac{d}{dt}\left(p(t)\frac{d}{dt}\right) + q(t)$$

is called the Sturm-Liouville operator.

To see how to turn $(*)$ into the Sturm-Liouville form, we can multiply $(*)$ by

$$\frac{1}{a_0(t)} e^{\int_a^t \frac{a_1(s)}{a_0(s)} ds}.$$

Then $(*)$ becomes

$$\begin{aligned} & e^{\int_a^t \frac{a_1(s)}{a_0(s)} ds}\left(y'' + \frac{a_1(t)}{a_0(t)}y' + \frac{a_2(t)}{a_0(t)}y\right) = e^{\int_a^t \frac{a_1(s)}{a_0(s)} ds}\frac{g(t)}{a_0(t)} \\ \iff \quad & (p(t)y')' + q(t)y = f(t) \end{aligned}$$

where

$$\begin{cases} p(t) = e^{\int_a^t \frac{a_1(s)}{a_0(s)} ds} \\ q(t) = \dfrac{a_2(t)}{a_0(t)} e^{\int_a^t \frac{a_1(s)}{a_0(s)} ds} \\ f(t) = \dfrac{g(t)}{a_0(t)} e^{\int_a^t \frac{a_1(s)}{a_0(s)} ds} \end{cases}$$

From this computation, we also see that $p(t)$ is continuous differentiable and positive on $I$. Thus we will focus on the boundary value problem

$$\begin{cases} Ly := (p(t)y')' + q(t)y = f(t) & \text{on } I = [a, b] \\ B_1 y := \alpha_1 y(a) + \beta_1 y'(a) = \xi_1 & \text{at } t = a \\ B_2 y := \alpha_2 y(b) + \beta_2 y'(b) = \xi_2 & \text{at } t = b \end{cases} \qquad (S)$$

where $p(t) \in C^1(I)$ with $p(t) > 0$ and $q(t), f(t) \in C^0(I)$. $\{\alpha_i, \beta_i, \xi_i\}$ are real numbers and

$$(\alpha_1, \beta_1) \neq (0, 0), \qquad (\alpha_2, \beta_2) \neq (0, 0).$$

The problem $(S)$ is called the boundary value problem of Sturmian type.

For any two functions $u(t), v(t) \in C^2(I)$, the following identity holds

$$vLu - uLv = (p(u'v - v'u))'.$$

Here $L$ is the Sturm-Liouville operator as above. This identity is called the Lagrange identity.

One important consequence of Lagrange identity is

**Proposition 5.1.1.** *Assume $B_i u = B_i v = 0$, $(i = 1, 2)$. Then*

$$\int_a^b (vLu - uLv)dt = 0.$$

*Proof:* By Lagrange identity

$$\int_a^b (vLu - uLv)dt = p(u'v - v'u)\Big|_a^b.$$

Let us consider the point $a$. By assumption

$$\begin{pmatrix} u(a) & u'(a) \\ v(a) & v'(a) \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} B_1 u \\ B_1 v \end{pmatrix} = 0.$$

Since $(\alpha_1, \beta_1) \neq (0, 0)$, we have

$$\det \begin{pmatrix} u(a) & u'(a) \\ v(a) & v'(a) \end{pmatrix} = uv' - u'v\Big|_{t=a} = 0.$$

Similarly, $(uv' - u'v)\big|_{t=b} = 0$. Therefore the boundary term vanishes

$$p(u'v - v'u)\big|_a^b = 0.$$

□

This proposition says $L$ is self-adjoint on functions with required boundary conditions.

### 5.1.3 Homogeneous Problem

Given the boundary value problem $(S)$, we will consider the corresponding homogeneous boundary value problem

$$\begin{cases} Lu = 0 & \text{on } I = [a, b] \\ B_1 u = 0 & \text{at } t = a \\ B_2 u = 0 & \text{at } t = b \end{cases} \tag{H}$$

Since the equation is linear, a general solution $y$ of the inhomogeneous problem $(S)$ can be written in the form

$$y = y^* + u$$

where $y^*$ is a special solution of the inhomogeneous problem $(S)$ and $u$ is a general solution of the homogeneous problem $(H)$. As a consequence, if the homogeneous problem $(H)$ has only the trivial solution $u = 0$, then the inhomogeneous problem $(S)$ has at most one solution. Actually, we shall show that there exists exactly one solution in this case, i.e., the triviality of the homogeneous problem $(H)$ also implies the solvability of the inhomogeneous problem $(S)$.

**Theorem 5.1.2.** *The inhomogeneous boundary value problem $(S)$ is uniquely solvable if and only if the homogeneous boundary value problem $(H)$ has only the zero solution.*

*Proof:* Assume the homogeneous problem $(H)$ has only the zero solution. We are left to show that the inhomogeneous problem $(S)$ can be solved. Let us consider the linear equation

$$Lu = 0$$

without requiring any boundary conditions. By the general theory of linear ODE, there exist two linearly independent solutions $u_1(t)$ and $u_2(t)$ such that a general solution of $Lu = 0$ can be written as a linear combination

$$u(t) = c_1u_1(t) + c_2u_2(t), \qquad c_i \in \mathbb{R}.$$

For $u(t)$ to satisfy the boundary condition $B_1u = B_2u = 0$, we need to find $c_1, c_2$ such that

$$\begin{pmatrix} B_1u_1 & B_1u_2 \\ B_2u_1 & B_2u_2 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = 0.$$

By assumption, this can have only trivial solution $c_1 = c_2 = 0$. This is the same as saying that the matrix $\begin{pmatrix} B_1u_1 & B_1u_2 \\ B_2u_1 & B_2u_2 \end{pmatrix}$ is invertible.

Now let $v(t)$ be any solution of the equation

$$Lv = f$$

without requiring any boundary condition. Such $v$ always exists. A general solution of $Ly = f$ can be written as

$$y = v + c_1u_1 + c_2u_2.$$

For $y$ to solve the inhomogeneous problem $(S)$, we need to find $c_1, c_2$ such that the boundary condition $B_1y = \xi_1, B_2y = \xi_2$ hold. This is equivalent to solve

$$\begin{pmatrix} B_1u_1 & B_1u_2 \\ B_2u_1 & B_2u_2 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} \xi_1 - B_1v \\ \xi_2 - B_2v \end{pmatrix}.$$

Since $\begin{pmatrix} B_1u_1 & B_1u_2 \\ B_2u_1 & B_2u_2 \end{pmatrix}$ is invertible, this matrix equation has a unique solution. □

*Remark* 5.1.3. The proof of the theorem implies that the unique solvability of the inhomogeneous problem $(S)$ is equivalent to

$$\det \begin{pmatrix} B_1u_1 & B_1u_2 \\ B_2u_1 & B_2u_2 \end{pmatrix} \neq 0.$$

Here $u_1, u_2$ are two linearly independent solutions of $Lu = 0$.

**Example 5.1.4.** Consider the boundary value problem

$$\begin{cases} y'' + y = f(t), & 0 \le t \le \pi \\ B_1 y = y(0) + y'(0) = \xi_1 \\ B_2 y = y(\pi) = \xi_2 \end{cases}$$

The corresponding homogeneous equation

$$u'' + u = 0$$

has two linearly independent solutions

$$u_1(t) = \cos t, \qquad u_2(t) = \sin t.$$

Since

$$\det \begin{pmatrix} B_1 u_1 & B_1 u_2 \\ B_2 u_1 & B_2 u_2 \end{pmatrix} = \det \begin{pmatrix} 1 & 1 \\ -1 & 0 \end{pmatrix} = 1$$

the above boundary value problem has a unique solution.

If we change the boundary conditions to consider instead

$$\begin{cases} y'' + y = f(t), & 0 \le t \le \pi \\ B_1 y = y(0) = \xi_1 \\ B_2 y = y(\pi) = \xi_2 \end{cases}$$

Then

$$\det \begin{pmatrix} B_1 u_1 & B_1 u_2 \\ B_2 u_1 & B_2 u_2 \end{pmatrix} = \det \begin{pmatrix} 1 & 0 \\ -1 & 0 \end{pmatrix} = 0$$

The homogeneous problem has infinitely many solutions

$$u(t) = c \sin t, \qquad c \in \mathbb{R}.$$

The following inhomogeneous boundary value problem

$$\begin{cases} y'' + y = 0, & 0 \le t \le \pi \\ y(0) = 0, & y(\pi) = 1 \end{cases}$$

has no solution.

## 5.2 Green's Function for Second Order Equations

### 5.2.1 Idea of Green's Function

We are interested in solving the equation

$$Ly = f, \qquad t \in I = [a, b]$$

with certain boundary conditions $B = (B_1, B_2)$ on $\partial I = \{a, b\}$. We have shown in Theorem 5.1.2 that if the homogeneous boundary value problem

$$Lu = 0, \qquad \text{with} \qquad B_1 u = B_2 u = 0$$

has only the trivial solution $u = 0$, then the above inhomogeneous problem is uniquely solvable.

The situation is very similar to the problem in linear algebr. Let $A$ be an $n \times n$ matrix. Then the linear equation

$$A \cdot \mathbf{x} = \mathbf{b}$$

is uniquely solvable if and only if the homogeneous equation

$$A \cdot \mathbf{u} = 0$$

has only the trivial solution $\mathbf{u} = 0$. In this case, the matrix $A$ is invertible, and the above inhomogeneous equation is uniquely solved by

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

Back to our problem, we hope to solve the equation $Ly = f$ in a similar fashion by

$$y = L^{-1} f.$$

This turns out to be the case, and the inverse $L^{-1}$ is called the Green's operator. More precisely, we will construct a function $G(t, s)$ such that the expression

$$y(t) = \int_a^b G(t, s) f(s) ds$$

solves the equation $Ly = f$. Comparing with the matrix case $\mathbf{x} = A^{-1}\mathbf{b}$ where in components

$$x_i = \sum_j A_{ij}^{-1} b_j,$$

the expression

$$y(t) = \int_a^b G(t, s) f(s) ds$$

can be viewed as an infinite matrix multiplication by replacing

$$\begin{aligned} i &\longrightarrow t \\ \sum_j &\longrightarrow \int ds \\ A_{ij}^{-1} &\longrightarrow G(t, s). \end{aligned}$$

Thus the function $G(t, s)$ can be viewed as representing the inverse of $L$

$$L^{-1} \Longrightarrow G(t, s).$$

This function $G(t, s)$ to be constructed is called the Green's function.

### 5.2.2 Construction of Green's Function

We assume that the homogeneous boundary value problem $(H)$

$$Lu = 0 \quad \text{on} \quad I, \quad B_1 u = B_2 u = 0$$

has only the trivial solution $u = 0$.

Let $u_1$ be a nonzero solution of

$$Lu_1 = 0 \quad \text{on} \quad I, \quad B_1 u_1 = 0.$$

This can be obtained by solving the initial value problem

$$Lu_1 = 0, \qquad u_1(a) = \xi_1, \qquad u_1'(a) = \eta_1$$

where $(\xi_1, \eta_1)$ is chosen such that

$$B_1 u_1 = \alpha_1 \xi_1 + \beta_1 \eta_1 = 0.$$

Let $u_2$ be a nonzero solution of

$$Lu_2 = 0 \quad \text{on} \quad I, \quad B_2 u_2 = 0.$$

This can be obtained by solving the initial value problem in the backward direction

$$Lu_2 = 0, \qquad u_2(b) = \xi_2, \qquad u_2'(b) = \eta_2$$

where $(\xi_2, \eta_2)$ is chosen such that

$$B_2 u_2 = \alpha_2 \xi_2 + \beta_2 \eta_2 = 0.$$

Such $u_1, u_2$ are uniquely determined up to a rescaling constant by the linearity of the problem.

Now we claim that these two functions $u_1$ and $u_2$ are two linearly independent solutions of $Lu = 0$. In fact, if $u_2 = \lambda u_1$ for some constant $\lambda \neq 0$. Then

$$B_1 u_2 = \lambda B_1 u_1 = 0.$$

Thus $u_2$ solves the homogeneous boundary value problem

$$Lu_2 = 0, \qquad B_1 u_2 = B_2 u_2 = 0.$$

By assumption, $u_2 = 0$ and this is a contradiction. Thus $u_1, u_2$ are linearly independent.

By Lagrange's identity,

$$\big(p(u_1 u_2' - u_1' u_2)\big)' = u_1 L u_2 - u_2 L u_1 = 0.$$

Therefore the quantity $p(u_1 u_2' - u_1' u_2)$ is independent of $t$. Let this constant be

$$C = p(u_1 u_2' - u_1' u_2).$$

Note that this constant $C \neq 0$. Otherwise

$$\det \begin{pmatrix} u_1(a) & u_2(a) \\ u_1'(a) & u_2'(a) \end{pmatrix} = \frac{C}{p(a)} = 0.$$

Then the initial data $(u_2(a), u_2'(a))$ will be proportional to $(u_1(a), u_1'(a))$, hence $u_1$ and $u_2$ will be linearly dependent by the linearity of the equation.

**Definition 5.2.1.** We define the Green's function $G(t,s)$ by

$$G(t,s) := \begin{cases} \frac{1}{C} u_1(t) u_2(s), & a \leq t \leq s \leq b \\ \frac{1}{C} u_1(s) u_2(t), & a \leq s \leq t \leq b. \end{cases}$$

**Proposition 5.2.2.** *The above defined $G(t,s)$ satisfies the following properties*

① *$G(t,s)$ is continuous on $(t,s) \in I \times I$.*

② *$G(t,s) = G(s,t)$ is symmetric.*

③ *The derivatives $\partial_t G, \partial_t^2 G, \partial_s G, \partial_s^2 G$ exist and is continuous away from the diagonal $t = s$.*

④ *On the diagonal $t = s$, the one sided limit*

$$\partial_t G(t^+, t) := \lim_{x \to t^+} \partial_x G(x,t)$$

$$\partial_t G(t^-, t) := \lim_{x \to t^-} \partial_x G(x,t)$$

*exist and they differ by*

$$\partial_t G(t^+, t) - \partial_t G(t^-, t) = \frac{1}{p(t)}, \qquad a < t < b.$$

⑤ *Let*

$$L_t = \frac{d}{dt}\left(p(t)\frac{d}{dt}\right) + q(t)$$

$$L_s = \frac{d}{ds}\left(p(s)\frac{d}{ds}\right) + q(s)$$

*denote the Sturm-Liouville operators in the variable $t$ and $s$ respectively. Then*

$$L_t G(t,s) = L_s G(t,s) = 0, \qquad at \quad t \neq s.$$

*Proof:* ①②③⑤ are obvious. We prove ④. By construction

$$\partial_t G(t^+, t) := \lim_{x \to t^+} \frac{1}{C} u_1(t) u_2'(x) = \frac{1}{C} u_1(t) u_2'(t)$$

$$\partial_t G(t^-, t) := \lim_{x \to t^-} \frac{1}{C} u_1'(x) u_2(t) = \frac{1}{C} u_1'(t) u_2(t)$$

Thus

$$\partial_t G(t^+, t) - \partial_t G(t^-, t) = \frac{1}{C}(u_1(t)u_2'(t) - u_1'(t)u_2(t)) = \frac{1}{p(t)}.$$

□

### 5.2.3 Solution via Green's Function

**Theorem 5.2.3.** *Assume the homogeneous boundary value problem $(H)$ has only the zero solution. Then the following semi-homogeneous boundary value problem*

$$Ly = f \quad on \quad I, \qquad B_1 y = B_2 y = 0$$

*is uniquely solved by*

$$y(t) = \int_a^b G(t,s) f(s) ds.$$

*Here $G(t,s)$ is the Green's function in Definition 5.2.1.*

*Proof:* We show the function $y(t)$ defined by $y(t) = \int_a^b G(t,s) f(s) ds$ solves the required boundary value problem. Let us first check the boundary conditions. By construction

$$B_1 y = \frac{B_1 u_1}{C} \int_a^b u_2(s) f(s) ds = 0$$

$$B_2 y = \frac{B_2 u_2}{C} \int_a^b u_1(s) f(s) ds = 0$$

as required.

Let us now consider the differential equation. We can write $y(t)$ as two contributions

$$\begin{aligned} y(t) =& \int_a^t G(t,s) f(s) ds + \int_t^b G(t,s) f(s) ds \\ =& \frac{1}{C} u_2(t) \int_a^t u_1(s) f(s) ds + \frac{1}{C} u_1(t) \int_t^b u_2(s) f(s) ds. \end{aligned}$$

This allows us to compute its derivatives by

$$\begin{aligned} y'(t) =& \frac{1}{C} u_2'(t) \int_a^t u_1(s) f(s) ds + \frac{1}{C} u_1'(t) \int_t^b u_2(s) f(s) ds \\ y''(t) =& \frac{1}{C} u_2''(t) \int_a^t u_1(s) f(s) ds + \frac{1}{C} u_1''(t) \int_t^b u_2(s) f(s) ds \\ & + \frac{1}{C} u_2'(t) u_1(t) f(t) - \frac{1}{C} u_1'(t) u_2(t) f(t) \\ =& \frac{1}{C} u_2''(t) \int_a^t u_1(s) f(s) ds + \frac{1}{C} u_1''(t) \int_t^b u_2(s) f(s) ds + \frac{f(t)}{p(t)}. \end{aligned}$$

It follows that

$$Ly = \frac{1}{C}(Lu_2(t)) \int_a^t u_1(s) f(s) ds + \frac{1}{C}(Lu_1(t)) \int_t^b u_2(s) f(s) ds + f(t) = f(t).$$

□

*Remark* 5.2.4. For those who are familar with distributions, the properties in Proposition 5.2.2

$$\begin{cases} L_t G(t,s) = 0 & t \neq s \\ \partial_t G(t^+, t) - \partial_t G(t^-, t) = \dfrac{1}{p(t)} \end{cases}$$

implies the distributional identity

$$L_t G(t,s) = \delta(t-s).$$

This says that $G(t,s)$ is the solution of the semi-homogeneous boundary value problem with a $\delta$-function source. Therefore the general source case is solved by the superposition

$$\int_a^b G(t,s)f(s)ds.$$

Then

$$L_t \int_a^b G(t,s)f(s)ds = \int_a^b L_t G(t,s)f(s)ds = \int_a^b \delta(t-s)f(s)ds = f(t).$$

The equation $L_t G(t,s) = \delta(t-s)$ is another way to express $G(t,s)$ as the inverse of $L$. $\delta(t,s)$ can be viewed as the identity matrix in the infinite dimension case.

*Remark* 5.2.5. For general inhomogeneous boundary value problem

$$Ly = f \quad \text{on} \quad I, \qquad B_1 y = \xi_1, \quad B_2 y = \xi_2$$

we can first find the unique solution $u = c_1 u_1 + c_2 u_2$ of

$$Lu = 0 \quad \text{on} \quad I, \qquad B_1 u = \xi_1, \quad B_2 u = \xi_2$$

by solving

$$\begin{pmatrix} B_1 u_1 & B_1 u_2 \\ B_2 u_1 & B_2 u_2 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}.$$

Then the above equation is reduced to the semi-homogeneous case

$$L(y-u) = f - Lu = f, \qquad B_1(y-u) = B_2(y-u) = 0$$

which is solved by

$$y(t) = u(t) + \int_a^b G(t,s)f(s)ds.$$

**Example 5.2.6.** Consider the boundary value problem on $I = [0,1]$

$$y'' = f(t), \qquad y(0) = y(1) = 0.$$

The required solutions $u_1, u_2$ of

$$\begin{cases} u_1'' = 0 & u_1(0) = 0 \\ u_2'' = 0 & u_2(1) = 0 \end{cases}$$

are found by

$$u_1(t) = t, \qquad u_2(t) = t - 1.$$

The constant $C = u_1 u_2' - u_1' u_2 = 1$. The Green's function is

$$G(t,s) = \begin{cases} t(s-1) & 0 \le t \le s \le 1 \\ s(t-1) & 0 \le s \le t \le 1 \end{cases}$$

The corresponding Dirichlet boundary value problem is solved by

$$y(t) = (t-1)\int_0^t sf(s)ds + t\int_t^1 (s-1)f(s)ds.$$

## 5.3 Boundary Value Problem in General

### 5.3.1 Linear System and Green's Matrix

We consider the boundary value problem for first order linear system

$$\begin{cases} \mathbf{y}' = A(t)\mathbf{y} + \mathbf{f}(t) & \text{on } I = [a,b] \\ B\mathbf{y} := \Gamma_1 \mathbf{y}(a) + \Gamma_2 \mathbf{y}(b) = \boldsymbol{\xi} \end{cases}$$

Here $\mathbf{y}(t)$ is the column of $n$ unknown functions. $A(t)$ is an $n \times n$ matrix varying continuously with respect to $t$ on $I$. $\Gamma_1, \Gamma_2$ are constant $n \times n$ matrices, and $\boldsymbol{\xi}$ is a constant column vector.

We will write the above equation as

$$L\mathbf{y} = \mathbf{f} \qquad \text{where} \qquad L = \frac{d}{dt} - A(t).$$

**Example 5.3.1.** The initial value problem corresponds to

$$\Gamma_1 = I_n, \qquad \Gamma_2 = 0.$$

**Example 5.3.2.** Consider the boundary value problem of Sturmian type $(S)$

$$\begin{cases} (py')' + qy = f \\ B_1 y = \alpha_1 y(a) + \beta_1 y'(a) = \xi_1 \\ B_2 y = \alpha_2 y(b) + \beta_2 y'(b) = \xi_2 \end{cases}$$

We can turn this into the first order system by introducing $\mathbf{y} = \begin{pmatrix} y \\ py' \end{pmatrix}$. Then the above boundary value problem becomes

$$\begin{cases} \mathbf{y}' = \begin{pmatrix} 0 & \frac{1}{p} \\ -q & 0 \end{pmatrix} \mathbf{y} + \begin{pmatrix} 0 \\ f \end{pmatrix} \\ \begin{pmatrix} \alpha_1 & \frac{\beta_1}{p(a)} \\ 0 & 0 \end{pmatrix} \mathbf{y}(a) + \begin{pmatrix} 0 & 0 \\ \alpha_2 & \frac{\beta_2}{p(b)} \end{pmatrix} \mathbf{y}(b) = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} \end{cases}$$

This corresponds to the boundary value problem of first order linear system with

$$A(t) = \begin{pmatrix} 0 & \frac{1}{p} \\ -q & 0 \end{pmatrix}, \qquad \Gamma_1 = \begin{pmatrix} \alpha_1 & \frac{\beta_1}{p(a)} \\ 0 & 0 \end{pmatrix}, \qquad \Gamma_2 = \begin{pmatrix} 0 & 0 \\ \alpha_2 & \frac{\beta_2}{p(b)} \end{pmatrix}$$

.

Let

$$\mathcal{P}(t) = \mathcal{P}\left(e^{\int_a^t A}\right), \qquad a \le t \le b$$

be the path-ordered exponential (Definition 2.3.1). This $n \times n$ invertible matrix function satisfies

$$\begin{cases} \dfrac{d}{dt}\mathcal{P}(t) = A(t)\mathcal{P}(t) \\ \mathcal{P}(a) = I_n \end{cases}$$

If we write $\mathcal{P}(t)$ as $n$ column vectors

$$\mathcal{P}(t) = \begin{pmatrix} \mathbf{u}_1(t) & \mathbf{u}_2(t) & \cdots & \mathbf{u}_n(t) \end{pmatrix},$$

then $\{\mathbf{u}_1, \cdots, \mathbf{u}_n\}$ form $n$ linearly independent solutions of the homogeneous equation

$$\frac{d\mathbf{u}}{dt} = A(t)\mathbf{u}.$$

The linear independency follows from the invertibility of $\mathcal{P}(t)$.

A general solution of the inhomogeneous linear system

$$L\mathbf{y} = \mathbf{f}$$

can be expressed as

$$\mathbf{y} = \mathbf{v} + c_1\mathbf{u}_1 + \cdots + c_n\mathbf{u}_n$$

where $\mathbf{v}$ is a special solution of $L\mathbf{v} = \mathbf{f}$. In matrix form, this is

$$\mathbf{y} = \mathbf{v} + \mathcal{P}\mathbf{c}, \qquad \text{where } \mathbf{c} = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix}.$$

To solve the boundary condition, we need to find constants $\mathbf{c}$ such that

$$B(\mathbf{v} + \mathcal{P}\mathbf{c}) = \boldsymbol{\xi}$$

*i.e.*

$$(\Gamma_1\mathcal{P}(a) + \Gamma_2\mathcal{P}(b))\mathbf{c} = \boldsymbol{\xi} - (\Gamma_1\mathbf{v}(a) + \Gamma_2\mathbf{v}(b)).$$

The situation is very similar to the boundary value problem of Sturmian type. The unique solvability of $\mathbf{c}$ is equivalent to the invertibility of the matrix $\Gamma_1\mathcal{P}(a)+\Gamma_2\mathcal{P}(b)$, which is equivalent to that the homogeneous matrix equation

$$(\Gamma_1\mathcal{P}(a) + \Gamma_2\mathcal{P}(b))\mathbf{c} = 0$$

has only the trivial solution. This is equivalent to that the homogeneous boundary value problem

$$\begin{cases} L\mathbf{u} = 0 \\ B\mathbf{u} = \Gamma_1\mathbf{u}(a) + \Gamma_2\mathbf{u}(b) = 0 \end{cases}$$

has only the trivial solution $u = 0$. Thus we have proved

**Theorem 5.3.3.** *The inhomogeneous boundary value problem $L\mathbf{y} = \mathbf{f}$, $B\mathbf{y} = \boldsymbol{\xi}$ on the interval $I$ is uniquely solvable if and only if the homogeneous boundary value problem $L\mathbf{u} = 0$, $B\mathbf{u} = 0$ has only the zero solution on $I$.*

Let us now assume the homogeneous boundary value problem $L\mathbf{u} = 0$, $B\mathbf{u} = 0$ on $I$ has only the zero solution. We look for an $n \times n$ matrix function $\mathbf{G}(t, s)$ such that the inhomogeneous boundary value problem

$$L\mathbf{y} = \mathbf{f}, \qquad B\mathbf{y} = 0$$

is solved by

$$\mathbf{y}(t) = \int_a^b \mathbf{G}(t, s)\mathbf{f}(s)ds.$$

Such $\mathbf{G}(t, s)$ is called the Green's matrix.

To construct $\mathbf{G}(t, s)$, let $\mathbf{v}$ be the special solution of $L\mathbf{v} = \mathbf{f}$ by (Theorem 2.3.3)

$$\mathbf{v}(t) = \mathcal{P}(t) \int_a^t \mathcal{P}^{-1}(s)\mathbf{f}(s)ds.$$

Let

$$R = \Gamma_1\mathcal{P}(a) + \Gamma_2\mathcal{P}(b) = \Gamma_1 + \Gamma_2\mathcal{P}(b)$$

which is invertible by assumption. Let $\mathbf{y} = \mathbf{v} + \mathcal{P}(t)\mathbf{c}$. We need to solve

$$\begin{aligned} R\mathbf{c} =& -(\Gamma_1\mathbf{v}(a) + \Gamma_2\mathbf{v}(b)) = -\Gamma_2\mathbf{v}(b) \\ \Rightarrow \qquad \mathbf{c} =& -R^{-1}\Gamma_2\mathbf{v}(b) = -R^{-1}\Gamma_2\mathcal{P}(b) \int_a^b \mathcal{P}^{-1}(s)\mathbf{f}(s)ds. \end{aligned}$$

It follows that

$$\begin{aligned} \mathbf{y}(t) =& \mathbf{v}(t) + \mathcal{P}(t)\mathbf{c} \\ =& \mathcal{P}(t) \int_a^t \mathcal{P}^{-1}(s)\mathbf{f}(s)ds - \mathcal{P}(t)R^{-1}\Gamma_2\mathcal{P}(b) \int_a^b \mathcal{P}^{-1}(s)\mathbf{f}(s)ds \\ =& \int_a^b \mathcal{P}(t)[H(t-s) - R^{-1}\Gamma_2\mathcal{P}(b)]\mathcal{P}^{-1}(s)\mathbf{f}(s)ds \end{aligned}$$

where $H(x)$ is the Heaviside step function

$$H(x) = \begin{cases} 1, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

We find the Green's matrix to be

$$\mathbf{G}(t, s) = \mathcal{P}(t)(H(t-s) - R^{-1}\Gamma_2\mathcal{P}(b))\mathcal{P}^{-1}(s).$$

### 5.3.2 Nonlinear Equation

We discuss some basic ideas for solving nonlinear boundary value problems in terms of Green's functions. Consider the boundary value problem

$$\begin{cases} Ly = f(y, t) & \text{on} \quad I = [a, b] \\ B_1 y = 0 & \text{at} \quad t = a \\ B_2 y = 0 & \text{at} \quad t = b \end{cases}$$

For example, we can consider the Sturm-Liouville operator $L$ with the boundary condition $B_1 y, B_2 y$ as before. We can use Green's function to transform this boundary value problem into an integral equation

$$y(t) = \int_a^b G(t,s) f(y(s), s) ds.$$

The equivalence of this integral equation with the boundary value problem is proved in the same way as in Section 5.2.3. The quickest way to see this is to use the distributional identity

$$LG(t,s) = \delta(t-s)$$

Then

$$\begin{aligned} Ly(t) &= \int_a^b LG(t,s) f(u(s), s) ds \\ &= \int_a^b \delta(t-s) f(u(s), s) ds = f(u(t), t). \end{aligned}$$

**Example 5.3.4.** As an illustration, we consider the following Dirichlet boundary value problem

$$\begin{cases} y'' = f(y,t) & \text{on } I = [0,1] \\ y(0) = y(1) = 0 \end{cases}$$

The corresponding Green's function is computed in Example 5.2.6

$$G(t,s) = \begin{cases} t(s-1) & 0 \le t \le s \le 1 \\ s(t-1) & 0 \le s \le t \le 1 \end{cases}$$

Let us consider the integral equation

$$y(t) = \int_0^1 G(t,s) f(y(s), s) ds.$$

Assume $f$ is continuous and satisfies the following Lipschitz condition with respect to $y$

$$|f(y_1, t) - f(y_2, t)| \le L|y_1 - y_2|, \quad L > 0.$$

We can try to solve the above integral equation using contraction mapping as in Section 3.1.4. In fact, define the transformation

$$(Ty)(t) = \int_0^1 G(t,s) f(y(s), s) ds.$$

The integral equation is the same as the fixed point equation

$$y = Ty.$$

Under the above Lipschitz condition, we obtain

$$\begin{aligned} |(Ty_1 - Ty_2)(t)| &= \left| \int_0^1 G(t,s)(f(y_1(s), s) - f(y_2(s), s)) ds \right| \\ &\le \left( \int_0^1 |G(t,s)| ds \right) L \|y_1 - y_2\|_\infty \\ &= \left( \int_0^t s(1-t) ds + \int_t^1 t(1-s) ds \right) L \|y_1 - y_2\|_\infty \\ &= (\frac{1}{2} t^2 (1-t) + \frac{1}{2} t (1-t)^2) L \|y_1 - y_2\|_\infty \\ &\le \frac{1}{8} L \|y_1 - y_2\|_\infty \end{aligned}$$

Taking the $\max_{0\le t\le 1}$ of the left hand side, we have

$$\|Ty_1 - Ty_2\|_\infty \le \frac{L}{8}\|y_1 - y_2\|_\infty.$$

Therefore if the Lipschitz constant $L < 8$, the transformation $T$ is a contraction map and we obtain a unique solution of this nonlinear boundary value problem.

A more careful estimate shows that if $L < \pi^2$, then this nonlinear Dirichlet boundary value problem is uniquely solvable, and this constant $\pi^2$ is sharp. Consider the case

$$\begin{cases} y'' = -\pi^2 y & 0 \le t \le 1 \\ y(0) = y(1) = 0 \end{cases}$$

We will treat the linear term on the right as $f(y) = -\pi^2 y$ which is Lipschitz with $L = \pi^2$. There are infinitely many solutions

$$y(t) = c \sin \pi t, \qquad c \in \mathbb{R}$$

For another example, consider

$$\begin{cases} y'' = -\pi^2 (y+1) & 0 \le t \le 1 \\ y(0) = y(1) = 0 \end{cases}$$

where $f(y) = -\pi^2(y+1)$ which is Lipschitz with $L = \pi^2$. A general solution of $y'' = -\pi^2(y+1)$ takes the form

$$y = c_1 \cos \pi t + c_2 \sin \pi t - 1.$$

There are no $c_1, c_2$ such that the boundary condition $y(0) = y(1) = 0$ holds. Thus there is no solution in this case.

## 5.4 Compact Self-adjoint Operators

We will digress for a moment to recall some basic properties for compact self-adjoint operators in preparation for the Sturm-Liouville eigenvalue problem in Section 5.5.

### 5.4.1 Inner Product Space

**Definition 5.4.1.** A $\mathbb{R}$(or $\mathbb{C}$) inner product space is a $\mathbb{R}$(or $\mathbb{C}$)-linear space $H$ equipped with a mapping (called inner product)

$$(\cdot, \cdot) : H \times H \to \mathbb{R}(\text{or } \mathbb{C})$$

such that the following hold

① Symmetry: $(f, g) = \overline{(g, f)}$

② Linearity: $(\alpha f + \beta g, h) = \alpha(f, h) + \beta(g, h)$

③ Positivity: $(f, f) > 0$ for $f \neq 0$.

Here $f, g, h \in H$, $\alpha, \beta \in \mathbb{R}$(or $\mathbb{C}$).

An inner product space is a normed space. The norm of a vector $f \in H$ is defined to be

$$\|f\| = \sqrt{(f, f)}.$$

*Remark* 5.4.2. An inner product space is called a Hilbert space if it is complete as a normed space, *i.e.* , a Banach space.

**Example 5.4.3.** Let $I = [a, b]$ and

$$C(I) = \{\text{continuous functions on } I\}.$$

This can be equipped with an $L^2$-inner product

$$(f, g) = \int_a^b f(x)\overline{g(x)}dx.$$

The induced norm is denoted by

$$\|f\|_2 := \sqrt{\int_a^b |f(x)|^2 dx}.$$

This inner product space $C(I)$ is not a Hilbert space. It can be completed to a Hilbert space $L^2(I)$ which consists of measurable functions $f(x)$ such that

$$\int_a^b |f(x)|^2 dx < +\infty.$$

Then $C(I)$ becomes a dense linear subspace of $L^2(I)$.

### 5.4.2 Compact Self-adjoint Operators

**Definition 5.4.4.** Let $H$ be a $\mathbb{R}$(or $\mathbb{C}$) inner product space and $T : H \to H$ be a linear operator. $T$ is called

① bounded if the norm of $T$ defined by

$$\|T\| := \sup_{\substack{\|f\|=1 \\ f \in H}} \|Tf\|$$

is finite.

② self-adjoint if

$$(Tf, g) = (f, Tg), \qquad \forall f, g \in H.$$

③ compact if for every bounded sequence $\{f_n\}$ in $H$, the sequence $\{Tf_n\}$ has a convergent subsequence with limit in $H$.

A compact operator is bounded. For a bounded linear operator $T$, we have

$$\|Tf\| \leq \|T\|\|f\| \qquad \forall f \in H.$$

**Proposition 5.4.5.** *If $T$ is bounded and self-adjoint, then*

$$(Tf, f) \in \mathbb{R}, \qquad \forall f \in H$$

*and*

$$\|T\| = \sup_{\substack{\|f\|=1 \\ f\in H}} |(Tf, f)|.$$

*Proof:*

$$\overline{(Tf, f)} = (f, Tf) = (Tf, f)$$

$$\Rightarrow (Tf, f) \in \mathbb{R}.$$

Let us denote

$$M = \sup_{\substack{\|f\|=1 \\ f\in H}} |(Tf, f)|.$$

Using the Cauchy-Schwartz inequality

$$|(Tf, f)| \leq \|Tf\|\|f\| \leq \|T\|\|f\|^2$$

we have $M \leq \|T\|$.

On the other hand, for any $f \in H$ with $\|f\| = 1$. Let $\lambda = \|Tf\|$. Then using

$$\begin{aligned}&(T(Tf + \lambda f), Tf + \lambda f) - (T(Tf - \lambda f), Tf - \lambda f)\\ =&2\lambda(Tf, Tf) + 2\lambda(T^2 f, f) = 4\lambda\|Tf\|^2 = 4\lambda^3\end{aligned}$$

we find

$$\begin{aligned}4\lambda^3 \leq& M\|Tf + \lambda f\|^2 + M\|Tf - \lambda f\|^2\\ =&2M(\|Tf\|^2 + \lambda^2\|f\|^2) = 4M\lambda^2\\ \Rightarrow \qquad& \lambda \leq M\end{aligned}$$

Since $f$ is arbitrary $\Rightarrow \|T\| \leq M$. Thus $\|T\| = M$. □

**Definition 5.4.6.** If $Tf = \lambda f$ for $f \neq 0$, then $\lambda$ is called an eigenvalue of $T$ and $f$ is called an eigenvector.

**Proposition 5.4.7.** *Let $T$ be a compact self-adjoint operator in the inner product space $H$. Then any eigenvalue $\lambda$ of $T$ satisfies*

$$|\lambda| \leq \|T\|$$

*and there exists an eigenvalue $\lambda_0$ such that $|\lambda_0| = \|T\|$.*

*Proof:* Let $\lambda$ be an eigenvalue of $T$ with nonzero eigenvector $f \in H$. We can assume $\|f\| = 1$. Then $Tf = \lambda f$ implies

$$|\lambda| = |(Tf, f)| \leq \|T\|.$$

To determine an eigenvalue $\lambda_0$ with $|\lambda_0| = \|T\|$, consider a sequence $\{f_n\}$ in $H$ with $\|f_n\| = 1$ such that

$$|(Tf_n, f_n)| \to \|T\| \qquad \text{as } n \to +\infty.$$

Since $T$ is compact, $\{Tf_n\}$ has a convergent subsequence with limit in $H$. Moreover, since $(Tf_n, f_n)$ is bounded, we find a further subsequence such that $(Tf_n, f_n)$ has a limit. Replacing $\{f_n\}$ by appropriate subsequence, we can assume both the following limits exist

$$(Tf_n, f_n) \to \lambda_0 \in \mathbb{R} \qquad n \to +\infty$$
$$Tf_n \to \lambda_0 g \in H \qquad n \to +\infty$$

Here by construction $|\lambda_0| = \|T\|$ which we assume $\neq 0$. Since

$$\begin{aligned} 0 \leq & \|Tf_n - \lambda_0 f_n\|^2 = \|Tf_n\|^2 + \lambda_0^2 \|f_n\|^2 - 2\lambda_0 (Tf_n, f_n) \\ \leq & \|T\|^2 + \lambda_0^2 - 2\lambda_0 (Tf_n, f_n) \\ = & 2\lambda_0^2 - 2\lambda_0 (Tf_n, f_n) \to 0 \qquad \text{as } n \to +\infty \end{aligned}$$

it follows that

$$\lambda_0^2 \|f_n - g\|^2 \leq \|Tf_n - \lambda_0 f_n\|^2 + \|Tf_n - \lambda_0 g\|^2 \to 0 \qquad \text{as } n \to \infty.$$

Thus $f_n \to g$ as $n \to +\infty$, and

$$\|Tg - \lambda_0 g\|^2 = \lim_{n \to +\infty} \|Tf_n - \lambda_0 f_n\|^2 = 0,$$

*i.e.* , $Tg = \lambda_0 g$. So $\lambda_0$ is indeed an eigenvalue of $T$. □

### 5.4.3 Orthonormal Sequence

**Definition 5.4.8.** Let $H$ be an $\infty$-dim inner product space. A sequence $\{\phi_n\}_{n=0}^{\infty}$ of elements in $H$ is called an orthonormal sequence if

$$(\phi_n, \phi_m) = \delta_{nm}, \qquad \forall n, m \geq 0.$$

Let $\{\phi_n\}_{n=0}^{\infty}$ be an orthonormal sequence. Given $f \in H$, the series

$$\sum_{k=0}^{\infty} c_k \phi_k \qquad \text{with } c_k = (f, \phi_k)$$

is called the Fourier series of $f$ with respect to $\{\phi_n\}$, and $c_k$'s are called the Fourier coefficients.

**Proposition 5.4.9** (Bessel's Inequality)**.** *Let* $\sum_{k=0}^{\infty} c_k \phi_k$ *be the Fourier series of* $f \in H$*. Then*

$$\sum_{k=0}^{\infty} |c_k|^2 \leq \|f\|^2.$$

*Equality holds if and only if* $f = \sum_{k=0}^{\infty} c_k \phi_k$.

*Proof:* Let $s_n = \sum_{k=0}^{n} c_k \phi_k$ be the partial sum. By construction

$$(f - s_n, \phi_k) = 0, \qquad \text{for } 0 \le k \le n.$$

Then we have

$$\begin{aligned} 0 \le & (f - s_n, f - s_n) = (f, f) - (f - s_n, s_n) - (s_n, f - s_n) - (s_n, s_n) \\ = & (f, f) - (s_n, s_n) = \|f\|^2 - \sum_{k=0}^{n} |c_k|^2. \end{aligned}$$

It follows that $\sum_{k=0}^{\infty} |c_k|^2 \le \|f\|^2$. Equality holds if and only if $\lim_{n\to\infty} \|f - s_n\|^2 = 0$, *i.e.*

$$f = \lim_{n\to\infty} s_n = \sum_{k=0}^{\infty} c_k \phi_k.$$

□

**Theorem 5.4.10.** *Let $H$ be an $\infty$-dim inner product space, and $T : H \to H$ be a compact self-adjoint operator. Then there exists countably many real eigenvalues $\lambda_0, \lambda_1, \cdots$ of $T$, with*

$$|\lambda_0| \ge |\lambda_1| \ge |\lambda_2| \ge \cdots \qquad \text{and} \qquad \lambda_n \to 0 \text{ as } n \to +\infty.$$

*The corresponding eigenvectors $\{\phi_n\}_{n=0}^{\infty}$, where*

$$T\phi_n = \lambda_n \phi_n, \qquad \|\phi_n\| = 1$$

*form an orthonormal sequence.*

*Each element in the image of $T$ is represented by its Fourier series*

$$Tf = \sum_{k=0}^{\infty} (Tf, \phi_k)\phi_k, \qquad \forall f \in H.$$

*Furthermore, any nonzero eigenvalue of $T$ equals to some $\lambda_i$ above.*

*Proof:* Let $\lambda_0$ be an eigenvalue of $T$ with $|\lambda_0| = \|T\|$, as promised by Proposition 5.4.7. Let $\phi_0$ be a corresponding eigenvector with $\|\phi_0\| = 1$. By construction

$$|\lambda_0| = |(T\phi_0, \phi_0)| = \sup_{\substack{f \in H \\ \|f\|=1}} |(Tf, f)| = \|T\|.$$

Consider

$$H_1 = \{f \in H \mid (f, \phi_0) = 0\}.$$

$H_1$ is a closed subspace of $H$ and

$$T : H_1 \to H_1.$$

Indeed, let $f \in H_1$. Then

$$(Tf, \phi_0) = (f, T\phi_0) = \lambda_0 (f, \phi_0) = 0$$

which implies $Tf \in H_1$.

It is easy to check that $T : H_1 \to H_1$ is again a compact self-adjoint operator. It is clear that the norm of $T$ on $H_1$ is no larger than its norm on $H$. Thus by Proposition 5.4.7, there exists an eigenvalue $\lambda_1$ with eigenvector $\phi_1$ such that

$$|\lambda_0| \geq |\lambda_1|, \qquad (\phi_1, \phi_0) = 0, \qquad \|\phi_1\| = 1.$$

Next we can consider

$$H_2 = \{f \in H \mid (f, \phi_0) = (f, \phi_1) = 0\}$$

and similarly find eigenvalue $\lambda_2$ and eigenvector $\phi_2$ such that

$$|\lambda_0| \geq |\lambda_1| \geq |\lambda_2|, \qquad (\phi_2, \phi_1) = (\phi_2, \phi_0) = 0, \qquad \|\phi_2\| = 1.$$

We can repeat this process and find eigenvalues $\lambda_0, \lambda_1, \cdots$, with

$$|\lambda_0| \geq |\lambda_1| \geq \cdots$$

and orthonormal sequence of eigenvectors $\{\phi_n\}$.

We claim that $\lambda_n \to 0$ as $n \to +\infty$. Otherwise the sequence $\{\frac{1}{\lambda_n}\phi_n\}$ would be bounded. By compactness of $T$, the sequence

$$\{T(\frac{1}{\lambda_n}\phi_n)\} = \{\phi_n\}$$

would have convergence subsequence. But this is impossible, since $\|\phi_n - \phi_m\|^2 = 2$ for any $n \neq m$.

Now for any $f \in H$, let

$$s_n = \sum_{k=0}^{n} c_k \phi_k, \qquad c_k = (f, \phi_k)$$

be the partial sum of the Fourier series of $f$. Then

$$(f - s_n, \phi_k) = 0, \qquad k = 0, 1, \cdots, n.$$

Since $|\lambda_{n+1}|$ equals to the norm of $T$ restricting to the subspace that is orthogonal to $\phi_0, \cdots, \phi_n$, we have

$$\|T(f - s_n)\| \leq |\lambda_{n+1}|\|f - s_n\| \leq |\lambda_{n+1}|\|f\| \to 0 \qquad \text{as } n \to +\infty.$$

Thus

$$Tf = \lim_{n\to\infty} Ts_n = \sum_{k=0}^{\infty} \lambda_k c_k \phi_k.$$

Finally, let $\lambda \neq 0$ be an eigenvalue of $T$ with eigenvector $\phi$. Then $\phi = T(\frac{1}{\lambda}\phi)$ lies in the image of $T$ and therefore

$$\phi = \sum_{k=0}^{\infty} (\phi, \phi_k)\phi_k.$$

If $\lambda \neq \lambda_k$ for any $k$, then

$$\begin{aligned}\lambda(\phi, \phi_k) =& (T\phi, \phi_k) = (\phi, T\phi_k) = \lambda_k(\phi, \phi_k) \\ \Rightarrow \quad & (\phi, \phi_k) = 0 \\ \Rightarrow \quad & \phi = 0\end{aligned}$$

Contradiction. It follows that $\lambda = \lambda_k$ for some $k$ and

$$\phi = \sum_{\lambda_k = \lambda} (\phi, \phi_k)\phi_k.$$

□

## 5.5 Sturm-Liouville Eigenvalue Problem

### 5.5.1 Eigenvalue Problem

Recall Theorem 5.2.3 for the following boundary value problem

$$\begin{cases} Ly = f \qquad \text{on } I = [a, b] \\ B_1 y := \alpha_1 y(a) + \beta_1 y'(a) = 0 \\ B_2 y := \alpha_2 y(b) + \beta_2 y'(b) = 0 \end{cases}$$

Assume the homogeneous boundary value problem

$$\begin{cases} Lu = 0 \qquad \text{on } I = [a, b] \\ B_1 u = 0 \\ B_2 u = 0 \end{cases}$$

has only the trivial solution $u = 0$. Then there exists the Green's function $G(t, s)$ such that the above inhomogeneous boundary value problem is uniquely solved by

$$y(t) = \int_a^b G(t, s) f(s) ds.$$

Another related problem is the Sturm-Liouville eigenvalue problem:

$$\begin{cases} Ly + \lambda y = 0 \qquad \text{on } I \\ B_1 y = B_2 y = 0 \end{cases}$$

which depends on a real parameter $\lambda$. This problem is interested in finding certain value of $\lambda$ such that the boundary value problem has a nontrivial solution $y \neq 0$. Such $\lambda$ is called the eigenvalue of the problem. If the solution space has $\dim = m > 0$, we say the eigenvalue $\lambda$ has multiplicity $m$. We can think about $\lambda$ as the eigenvalue of the operator $-L$

$$-Ly = \lambda y.$$

We will always assume the homogeneous boundary value problem has only the trivial solution. This is the same as saying that $\lambda = 0$ is not an eigenvalue.

**Example 5.5.1.** Consider the eigenvalue problem

$$\begin{cases} y'' + \lambda y = 0 & \text{on } I = [0, \pi] \\ y(0) = y(\pi) = 0 \end{cases}$$

This boundary value problem has nontrivial solution only when

$$\lambda = \lambda_n = n^2, \qquad n = 1, 2, 3, \cdots$$

Each eigenvalue $\lambda_n$ has multiplicity 1 with solution

$$y_n(t) = \sin(nt).$$

We observe that any function $\phi \in C^1(I)$ with $\phi(0) = \phi(\pi)$ can be expanded

$$\phi(t) = \sum_{n=1}^{\infty} a_n \sin(nt)$$

by the eigenfunctions $\{\sin(nt)\}$. This is the well-known Fourier series expansion.

### 5.5.2 Green's function as Compact Self-adjoint Operator

It turns out that Fourier expansion such as that in Example 5.5.1 exists for Sturm-Liouville eigenvalue problem in general. To set things up, we will consider the inner product space $C(I)$ of continuous real functions on $I = [a, b]$. The inner product is

$$(f, g) = \int_a^b f(t)g(t)dt.$$

We assume that the homogeneous problem

$$\begin{cases} Lu = 0 & \text{on } I \\ B_1 u = B_2 u = 0 \end{cases}$$

has only the trivial solution $u = 0$. Thus the Green's function $G(t, s)$ exists. Then the eigenvalue boundary value problem

$$\begin{cases} Ly = -\lambda y \\ B_1 y = B_2 y = 0 \end{cases} \qquad (\lambda \neq 0)$$

is equivalent to the integral equation

$$y(t) = -\lambda \int_a^b G(t, s)y(s)ds.$$

We define the following linear operator

$$T : C(I) \to C(I)$$

by

$$(Tf)(t) = -\int_a^b G(t, s)f(s)ds.$$

Then the above eigenvalue boundary value problem is equivalent to find $y \in C(I)$ satisfying

$$Ty = \frac{1}{\lambda} y.$$

The corresponding solutions become the eigenvectors of the operator $T$.

Now the key is to realize that $T$ is a compact self-adjoint operator. This will allow us to apply the results in Section 5.4.

**Proposition 5.5.2.** *$T : C(I) \to C(I)$ defines a compact self-adjoint operator on $C(I)$.*

*Proof:* Let us first consider self-adjointness. Let $f, g \in C(I)$. Then

$$\begin{aligned}(f, Tg) &= \int_a^b f(t)(Tg)(t)dt \\ &= \int_a^b dt \int_a^b ds f(t)G(t,s)g(s)\end{aligned}$$

Similarly,

$$(Tf, g) = \int_a^b dt \int_a^b ds f(s)G(t,s)g(t)$$

Since $G(t,s) = G(s,t)$ is symmetric, we have

$$(Tf, g) = (f, Tg)$$

So $T$ is self-adjoint.

Next we consider the compactness. Let $\{f_n\}$ be a bounded sequence in $C(I)$ with $\|f_n\| \leq M$. Let

$$g_n(t) = (Tf_n)(t) = -\int_a^b G(t,s) f_n(s) ds.$$

Since $G(t,s)$ is continuous hence bounded on $I \times I$, there exists $A > 0$ such that

$$|G(t,s)| \leq A, \qquad t, s \in I.$$

Thus

$$\begin{aligned}|g_n(t)| \leq & A \int_a^b |f_n(s)| ds \\ \leq & A(b-a)^{\frac{1}{2}} \left( \int_a^b f_n^2(s) ds \right)^{\frac{1}{2}} \\ \leq & AM(b-a)^{\frac{1}{2}}.\end{aligned}$$

So the sequence $\{g_n\}$ in $C(I)$ is uniformly bounded.

Again by the continuity of $G(t,s)$, for any $\varepsilon > 0$, there exists $\delta > 0$ such that

$$|G(t_1, s) - G(t_2, s)| < \varepsilon, \qquad \text{for } |t_1 - t_2| < \delta, \quad (t_i, s) \in I \times I.$$

Therefore for $|t_1 - t_2| < \delta$, we have

$$\begin{aligned}
&|g_n(t_1) - g_n(t_2)| \\
\leq &\int_a^b |G(t_1, s) - G(t_2, s)||f_n(s)|ds \\
\leq &\varepsilon \int_a^b |f_n(s)|ds \\
\leq &\varepsilon (b-a)^{\frac{1}{2}} \left( \int_a^b f_n^2(s)ds \right)^{\frac{1}{2}} \\
\leq &\varepsilon (b-a)^{\frac{1}{2}} M.
\end{aligned}$$

So the sequence $\{g_n\}$ in $C(I)$ is also uniformly equicontinuous.

Thus the sequence $\{g_n\}$ in $C(I)$ is uniformly bounded and uniformly equicontinuous. By Arzelà-Ascoli Theorem, there exists a subsequence that converges uniformly to a function $g$ in $C(I)$. This proves the compactness of $T$. □

### 5.5.3 Eigenfunctions and Fourier Series

Since $T : C(I) \to C(I)$ is compact self-adjoint, by Theorem 5.4.10, there exists eigenvalues $\mu_0, \mu_1, \cdots$ and eigenvectors $\phi_0, \phi_1, \cdots$ of $T$ such that

$$|\mu_0| \geq |\mu_1| \geq \cdots \qquad \mu_n \to 0, \quad n \to +\infty$$

and $\{\phi_n\}$ form an orthonormal sequence. Any function in the image of $T$ can be expressed as a Fourier series of $\{\phi_n\}$. Equivalently,

$$\lambda_0 = \frac{1}{\mu_0}, \qquad \lambda_1 = \frac{1}{\mu_1}, \qquad \cdots \qquad \lambda_n = \frac{1}{\mu_n}, \qquad \cdots$$

are eigenvalues of the Sturm-Liouville boundary value problem whose solutions are given by $\phi_0, \phi_1, \cdots$.

Given any $u \in C^2(I)$ with $B_1 u = B_2 u = 0$, we have $f := Lu \in C(I)$. We can view $u$ as a function solving the boundary value problem

$$\begin{cases} Lu = f \\ B_1 u = B_2 u = 0 \end{cases}$$

Thus

$$u = \int_a^b G(t, s) f(s) ds = -T(f).$$

So $u$ lies in the image of $T$. It follows that $u$ has a Fourier series expression by

$$u = \sum_{k=0}^{\infty} c_k \phi_k, \qquad c_k = (u, \phi_k) = \int_a^b u(t)\phi_k(t)dt.$$

*Remark* 5.5.3. Using Lebesgue integral, one can actually show that

$$T : L^2(I) \to L^2(I)$$

defines a compact self-adjoint operator on the Hilbert space $L^2(I)$. The eigenfunctions $\{\phi_n\}$ of the Sturm-Liouville problem form an orthonormal basis of $L^2(I)$.

From the series expansion,

$$\begin{aligned} u(t) &= \sum_{k=0}^{\infty} c_k \phi_k(t) \\ &= \sum_{k=0}^{\infty} (u, \phi_k)\phi_k(t) \\ &= -\sum_{k=0}^{\infty} \frac{1}{\lambda_k}(u, L\phi_k)\phi_k(t) \\ &= -\sum_{k=0}^{\infty} \frac{1}{\lambda_k}\phi_k(t) \int_a^b \phi_k(s)(Lu)(s)ds \\ &= \int_a^b \left( -\sum_{k=0}^{\infty} \frac{\phi_k(t)\phi_k(s)}{\lambda_k} \right) (Lu)(s)ds. \end{aligned}$$

Comparing with the Green's function formula

$$u(t) = \int_a^b G(t,s)(Lu)(s)ds$$

we find a very useful expression of Green's function in terms of eigenvalue problem

$$G(t,s) = -\sum_{k=0}^{\infty} \frac{\phi_k(t)\phi_k(s)}{\lambda_k}.$$

**Example 5.5.4.** Consider the Dirichlet boundary value problem

$$\begin{cases} y'' + 4y = t^2 & \text{on } I = [0,1] \\ y(0) = y(1) = 0 \end{cases}$$

The eigenvalue problem is

$$\phi'' + (4+\lambda)\phi = 0 \qquad \phi(0) = \phi(1) = 0.$$

It is easy to find the eigenvalues

$$\lambda_n = n^2\pi^2 - 4, \qquad n = 1, 2, \cdots$$

with normalized eigenfunctions

$$\phi_n(t) = \sqrt{2}\sin(n\pi t).$$

Thus the Green's function of $L = \left(\frac{d}{dt}\right)^2 + 4$ for the Dirichlet boundary condition is

$$G(t,s) = -\sum_{n=1}^{\infty} \frac{\phi_n(t)\phi_n(s)}{\lambda_n} = 2\sum_{n=1}^{\infty} \frac{\sin(n\pi t)\sin(n\pi s)}{4 - n^2\pi^2}$$

This allows us to solve the above Dirichlet boundary value problem by

$$\begin{aligned} y(t) =& \int_0^1 G(t,s)s^2 ds \\ =& 2\sum_{n=1}^{\infty} \frac{\sin(n\pi t)}{4-n^2\pi^2} \int_0^1 \sin(n\pi s)s^2 ds \\ =& 2\sum_{n=1}^{\infty} \frac{\sin(n\pi t)}{4-n^2\pi^2} \frac{(-1)^n(2-n^2\pi^2)-2}{n^3\pi^3}. \end{aligned}$$

# Chapter 6 Calculus of Variations

## 6.1 Euler-Lagrange Equation

### 6.1.1 Principle of Least Action

In classical Newtonian mechanics, the trajectory

$$\mathbf{q}(t) : \mathbb{R} \to \mathbb{R}^n$$

of a particle of mass $m$ moving in the space $\mathbb{R}^n$ obeys Newton's 2nd Law

$$\mathbf{F} = m\ddot{\mathbf{q}}(t).$$

Here $\mathbf{F} = \mathbf{F}(\mathbf{q}, t)$ is the force. This is a second order ODE, whose solutions are completely determined by specifying the initial condition $\mathbf{q}(t_0)$ and $\dot{\mathbf{q}}(t_0)$ at some time $t_0$.

We will mainly consider conservative forces, in which case we can write

$$\mathbf{F}(\mathbf{q}) = -\nabla V(\mathbf{q})$$

for some function $V : \mathbb{R}^n \to \mathbb{R}$ called the potential. Define the Kinetic energy

$$K = \frac{1}{2} m\dot{\mathbf{q}}^2$$

and the total Energy

$$E = K + V.$$

**Proposition 6.1.1.** *The total energy is conserved along the motion.*

*Proof:*

$$\begin{aligned}\frac{dE}{dt} =& \frac{d}{dt}(\frac{1}{2}m\dot{\mathbf{q}}^2) + \frac{d}{dt}V(\mathbf{q}(t)) \\ =& m\ddot{\mathbf{q}} \cdot \dot{\mathbf{q}} + \dot{\mathbf{q}}\nabla V \\ =& (\mathbf{F} + \nabla V) \cdot \dot{\mathbf{q}} = 0.\end{aligned}$$

□

**Definition 6.1.2.** Define the Lagrangian of the motion by

$$\mathcal{L} := K - V.$$

For any path

$$\mathbf{q}(t) : [t_0, t_1] \to \mathbb{R}^n, \qquad \mathbf{q}(t) = (q_1(t), \cdots, q_n(t))$$

we define its action functional by

$$S[\mathbf{q}(t)] := \int_{t_0}^{t_1} \mathcal{L}(q_i, \dot{q}_i) dt$$

Principle of Least Action: Trajectories of classical particles are extremal points of the system's action functional on the path space.

Often though not always, the action is minimized for classical trajectories, then this is the least action. We will show this for the above system in a minute. This turns out to be a basic principle governing classical mechanical problems.

### 6.1.2 Euler-Lagrange Equation

For simplicity, let us first consider the 1-dim case

$$q(t) : [t_0, t_1] \to \mathbb{R}.$$

**Theorem 6.1.3.** *Assume $x(t) : [t_0, t_1] \to \mathbb{R}$ is a smooth path that extremizes the action functional of the form*

$$S[q(t)] = \int_{t_0}^{t_1} \mathcal{L}(q, \dot{q}, t) dt$$

*for all possible smooth paths $q(t) : [t_0, t_1] \to \mathbb{R}$ with the fixed endpoints $q(t_0) = x(t_0)$, $q(t_1) = x(t_1)$. Then $x(t)$ satisfies the* ***Euler-Lagrange Equation***

$$\frac{d}{dt}\left(\frac{\partial \mathcal{L}}{\partial \dot{x}}\right) = \frac{\partial \mathcal{L}}{\partial x}.$$

*Proof:* Let $x(t)$ be such an extremizer. For any smooth map

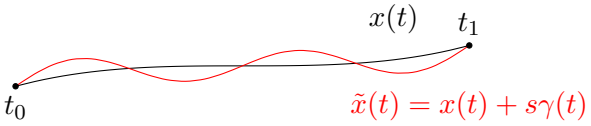
$$\gamma : [t_0, t_1] \to \mathbb{R}, \qquad \gamma(t_0) = \gamma(t_1) = 0,$$

and any small number $s$, the path

$$\tilde{x}(t) = x(t) + s\gamma(t)$$

has the same endpoints with $x(t)$:

$$\tilde{x}(t_i) = x(t_i) + s\gamma(t_i) = x(t_i), \qquad i = 0, 1.$$




Let us consider the function of $s$ defined by

$$f(s) := S[x(t) + s\gamma(t)]$$

By assumption, $f$ takes an extremal value at $s = 0$

$$\implies \quad f'(0) = 0.$$

On the other hand,

$$f(s) = \int_{t_0}^{t_1} \mathcal{L}(x + s\gamma, \dot{x} + s\dot{\gamma}, t)dt$$

By Chain rule

$$\begin{aligned} f'(0) &= \int_{t_0}^{t_1} \left( \frac{\partial \mathcal{L}}{\partial x}\gamma + \frac{\partial \mathcal{L}}{\partial \dot{x}}\dot{\gamma} \right) dt \\ &= \int_{t_0}^{t_1} \left( \frac{\partial \mathcal{L}}{\partial x}\gamma - \frac{d}{dt}\left( \frac{\partial \mathcal{L}}{\partial \dot{x}} \right)\gamma \right) dt + \frac{\partial \mathcal{L}}{\partial \dot{x}}\gamma \Big|_{t_0}^{t_1} \\ &= \int_{t_0}^{t_1} \left( \frac{\partial \mathcal{L}}{\partial x} - \frac{d}{dt}\left( \frac{\partial \mathcal{L}}{\partial \dot{x}} \right) \right) \gamma dt. \end{aligned}$$

Here the boundary term from integration by parts vanishes since $\gamma(t_0) = \gamma(t_1) = 0$. Thus

$$\int_{t_0}^{t_1} \left( \frac{\partial \mathcal{L}}{\partial x} - \frac{d}{dt}\left( \frac{\partial \mathcal{L}}{\partial \dot{x}} \right) \right) \gamma dt = 0.$$

Since the choice of $\gamma$ is arbitrary, it follows that

$$\frac{\partial \mathcal{L}}{\partial x} - \frac{d}{dt}\left( \frac{\partial \mathcal{L}}{\partial \dot{x}} \right) = 0$$

□

*Remark* 6.1.4. The above calculation generalizes to the $n$-dim case

$$\mathbf{q}(t) : [t_0, t_1] \to \mathbb{R}^n, \qquad \mathbf{q}(t) = (q_1(t), \cdots, q_n(t)).$$

The corresponding extremal path satisfies the **Euler-Lagrange Equation**

$$\frac{d}{dt}\left( \frac{\partial \mathcal{L}}{\partial \dot{q}_i} \right) = \frac{\partial \mathcal{L}}{\partial q_i}, \qquad i = 1, 2, \cdots, n.$$

**Example 6.1.5.** Let us consider a particle moving in the potential $V$. The Lagrangian is

$$\mathcal{L} = K - V = \frac{1}{2}m\dot{\mathbf{q}}^2 - V(\mathbf{q}).$$

We calculate

$$\frac{\partial \mathcal{L}}{\partial \dot{q}_i} = m\dot{q}_i, \qquad \frac{\partial \mathcal{L}}{\partial q_i} = -\frac{\partial V}{\partial q_i}$$

The Euler-Lagrange equation reads

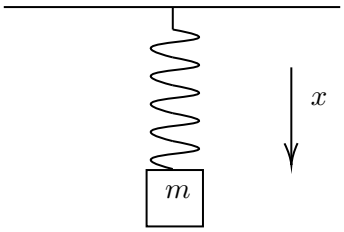
$$m\ddot{q}_i = -\frac{\partial V}{\partial q_i}$$

or in vector notation

$$m\ddot{\mathbf{q}} = -\nabla V.$$

This is precisely the motion via Newton's 2nd Law.

**Example 6.1.6** (Spring with gravity)**.** Consider a massless spring with elastic coefficient $k$.




We have

$$K = \frac{1}{2}m\dot{x}^2, \qquad V = \frac{1}{2}kx^2 + mgh = \frac{1}{2}kx^2 - mgx.$$

The Lagrangian is

$$\mathcal{L} = \frac{1}{2}m\dot{x}^2 - \frac{1}{2}kx^2 + mgx.$$

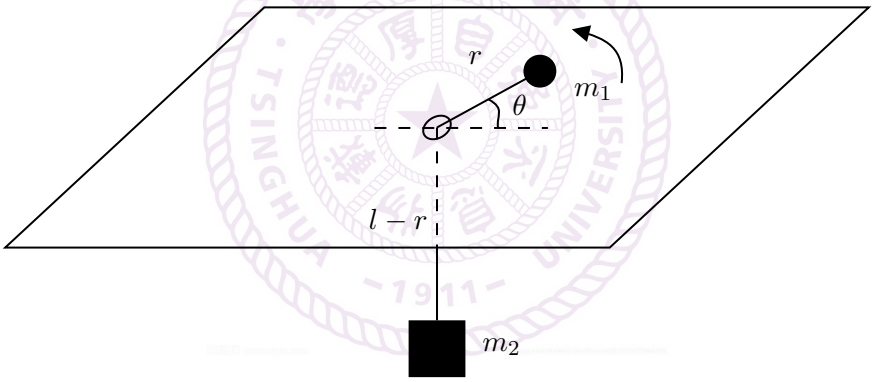
The Euler-Lagrange Equation

$$\frac{d}{dt}\frac{\partial \mathcal{L}}{\partial \dot{x}} = \frac{\partial \mathcal{L}}{\partial x}$$

reads

$$m\ddot{x} = mg - kx.$$

**Example 6.1.7** (Disk pulled by falling mass)**.** Consider a disk of mass $m_1$ pulled across a table by a falling object of mass $m_2$. Assume there is no friction.




We parametrize the position of $m_1$ via radial coordinate by

$$x(t) = r(t)\cos\theta(t), \quad y(t) = r(t)\sin\theta(t).$$

The kinetic energy of $m_1$ is

$$\frac{1}{2}m_1(\dot{x}^2 + \dot{y}^2) = \frac{1}{2}m_1(\dot{r}^2 + r^2\dot{\theta}^2).$$

The kinetic energy of $m_2$ is

$$\frac{1}{2}m_2(\dot{l - r})^2 = \frac{1}{2}m_2\dot{r}^2.$$

The gravitational potential is

$$V = -m_2g(l - r) = m_2g(r - l).$$

Therefore the Lagrangian is

$$\mathcal{L} = \frac{1}{2}m_1(\dot{r}^2 + r^2\dot{\theta}^2) + \frac{1}{2}m_2\dot{r}^2 - m_2g(r - l).$$

The Euler-Lagrange equation

$$\begin{cases} \dfrac{d}{dt}\left(\dfrac{\partial \mathcal{L}}{\partial \dot{r}}\right) = \dfrac{\partial \mathcal{L}}{\partial r} \\ \dfrac{d}{dt}\left(\dfrac{\partial \mathcal{L}}{\partial \dot{\theta}}\right) = \dfrac{\partial \mathcal{L}}{\partial \theta} \end{cases}$$

reads

$$\begin{cases} (m_1 + m_2)\ddot{r} = m_1 r\dot{\theta}^2 - m_2 g \\ \dfrac{d}{dt}(m_1 r^2 \dot{\theta}) = 0 \end{cases}$$

From the second equation we get

$$J := m_1 r^2 \dot{\theta} = \text{constant}.$$

The constant $J$ is precisely the angular momentum. Plugging

$$\dot{\theta} = \frac{J}{m_1 r^2}$$

into the first equation, we find

$$(m_1 + m_2)\ddot{r} = \frac{J^2}{m_1 r^3} - m_2 g.$$

Effectively, this can be viewed as a 1-dim problem along $r$, where the disk feels a force $\frac{J^2}{m_1 r^3} - m_2 g$ with potential (by integrating)

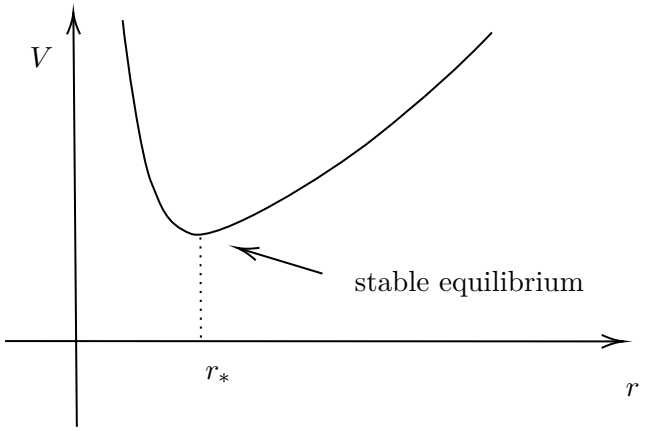
$$\frac{J^2}{2m_1 r^2} + m_2 g r.$$

The stable equilibrium is at

$$\frac{J^2}{m_1 r_*^3} - m_2 g = 0$$

which is solved by

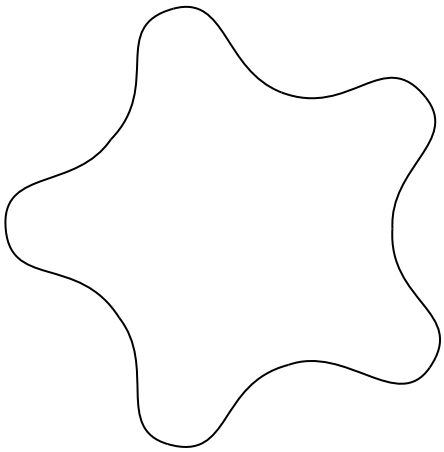
$$r_* = \left(\frac{J^2}{m_1 m_2 g}\right)^{1/3}.$$




At this point, the disk rotates with frequency

$$\dot{\theta} = \frac{J}{m_1 r_*^2} = \left(\frac{m_2^2 g^2}{m_1 J}\right)^{1/3}.$$

Otherwise we would find orbits like



## 6.2 Kepler Problem

The classical Kepler problem describes the motion in $\mathbb{R}^3$ in the potential of the form

$$V(r) = -\frac{K}{r}, \qquad K \text{ is constant}$$

We use $\mathbf{r} = (x_1, x_2, x_3)$ for the position and $r = \sqrt{x_1^2 + x_2^2 + x_3^2}$ for the length. The force is

$$\mathbf{F} = -\nabla V = -\frac{K}{r^2}\frac{\mathbf{r}}{r}$$

which is a central conservative force with inverse square growth of $r$.

- If $K > 0$, the force is attractive. Gravitational force and attractive electrostatic force are such examples.
- If $K < 0$, the force is repulsive. Repulsive electrostatic force is such an example.

### 6.2.1 Solutions of Motion

We will next focus on the attractive force so $K > 0$. The equation of motion is

$$m\ddot{\mathbf{r}} = -\frac{K}{r^2}\frac{\mathbf{r}}{r}.$$

In components, we have a system of ODE

$$\begin{cases} m\ddot{x}_1 = -\dfrac{Kx_1}{r^3} \\ m\ddot{x}_2 = -\dfrac{Kx_2}{r^3} \\ m\ddot{x}_3 = -\dfrac{Kx_3}{r^3} \end{cases}$$

Our goal is to solve the above equations.

Recall the angular momentum is defined by

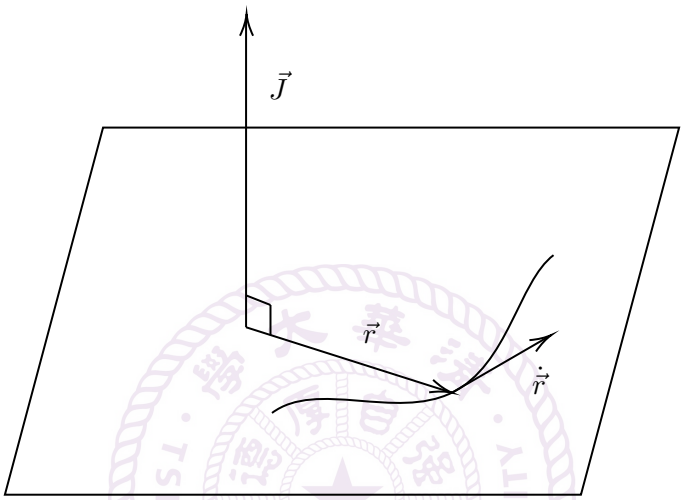
$$\mathbf{J} = m\mathbf{r} \times \dot{\mathbf{r}}.$$

We first observe that $\mathbf{J}$ is conserved along the motion (this in fact follows from the rotational symmetry of the problem via Noether's principle). Indeed

$$\begin{aligned}\frac{d}{dt}\mathbf{J} =& m\dot{\mathbf{r}}\times\dot{\mathbf{r}} + m\mathbf{r}\times\ddot{\mathbf{r}}\\ =& \mathbf{r}\times(m\ddot{\mathbf{r}})\\ \overset{\text{Eqn of Motion}}{=}& \mathbf{r}\times(-\frac{K}{r^2}\frac{\mathbf{r}}{r}) = 0.\end{aligned}$$

Note that

$$\mathbf{J}\cdot\mathbf{r} = 0, \qquad \mathbf{J}\cdot\dot{\mathbf{r}} = 0$$




Since the direction of $\mathbf{J}$ is fixed, the motion is confined to the plane containing the initial position and velocity. This reduces the problem to a plane motion with central force.

There is another hidden conserved vector called Laplace-Runge-Lenz vector

$$\mathbf{A} = \frac{\dot{\mathbf{r}}\times\mathbf{J}}{K} - \frac{\mathbf{r}}{r}.$$

Let us check that $\mathbf{A}$ is conserved along the motion. Using the identity for vectors in $\mathbb{R}^3$

$$\vec{a}\times(\vec{b}\times\vec{c}) = (\vec{a}\cdot\vec{c})\vec{b} - (\vec{a}\cdot\vec{b})\vec{c}$$

and the motion equation $m\ddot{\mathbf{r}} = -\frac{K\mathbf{r}}{r^3}$, we have (using conservation of $\mathbf{J}$)

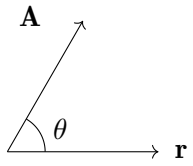
$$\begin{aligned}\frac{d\mathbf{A}}{dt} =& \frac{1}{K}\ddot{\mathbf{r}}\times(m\mathbf{r}\times\dot{\mathbf{r}}) - \frac{d}{dt}\left(\frac{\mathbf{r}}{r}\right)\\ =& -\frac{1}{r^3}(\mathbf{r}\times(\mathbf{r}\times\dot{\mathbf{r}})) - \frac{d}{dt}\left(\frac{\mathbf{r}}{r}\right)\\ =& -\frac{\mathbf{r}\cdot\dot{\mathbf{r}}}{r^3}\mathbf{r} + \frac{\dot{\mathbf{r}}}{r} - \frac{d}{dt}\left(\frac{\mathbf{r}}{r}\right)\\ =& 0\end{aligned}$$

So $\mathbf{A}$ is indeed conserved.

Let us now consider the inner product $\mathbf{A}\cdot\mathbf{r}$. Using $(\vec{a}\times\vec{b})\cdot\vec{c}=(\vec{c}\times\vec{a})\cdot\vec{b}$, we have

$$\begin{aligned}\mathbf{A}\cdot\mathbf{r}=&\frac{1}{K}(\dot{\mathbf{r}}\times\mathbf{J})\cdot\mathbf{r}-r\\=&\frac{1}{K}(\mathbf{r}\times\dot{\mathbf{r}})\cdot\mathbf{J}-r\\=&\frac{J^2}{Km}-r.\end{aligned}$$

Let $\mathbf{A}\cdot\mathbf{r}=Ar\cos\theta$, where $A$ is the length of $\mathbf{A}$ and $\theta$ is the angle between $\mathbf{A}$ and $\mathbf{r}$.



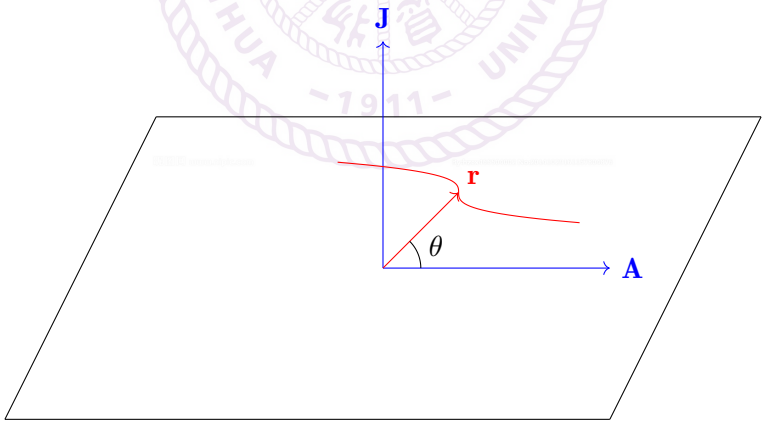

Then

$$\begin{aligned}Ar\cos\theta=&\frac{J^2}{Km}-r\\\Longrightarrow\quad r=&\frac{J^2}{Km}\frac{1}{1+A\cos\theta}.\end{aligned}$$

Now we know the motion is on the plane and $J^2$, $A$ are conserved constants. Since

$$\mathbf{A}\cdot\mathbf{J}=0,$$

we have the following picture about the motion




The above equation becomes an equation in polar coordinate. It follows immediately that

- If $A<1$, the motion orbit is an ellipse
- If $A=1$, the motion orbit is a parabola
- If $A>1$, the motion orbit is a hyperbola

### 6.2.2 Kepler's Laws

#### Kepler's First Law

Let us take a closer look at the quantity $A$.

$$\begin{aligned}
\mathbf{A}\cdot\mathbf{A} =& \left(\frac{\dot{\mathbf{r}}\times\mathbf{J}}{K}-\frac{\mathbf{r}}{r}\right)\cdot\left(\frac{\dot{\mathbf{r}}\times\mathbf{J}}{K}-\frac{\mathbf{r}}{r}\right)\\
=&\frac{(\dot{\mathbf{r}}\times\mathbf{J})\cdot(\dot{\mathbf{r}}\times\mathbf{J})}{K^2}-2\frac{(\dot{\mathbf{r}}\times\mathbf{J})\cdot\mathbf{r}}{Kr}+1\\
=&\frac{((\dot{\mathbf{r}}\times\mathbf{J})\times\dot{\mathbf{r}})\cdot\mathbf{J}}{K^2}-2\frac{(\mathbf{r}\times\dot{\mathbf{r}})\cdot\mathbf{J}}{Kr}+1\\
=&\frac{(\dot{\mathbf{r}}\cdot\dot{\mathbf{r}})\mathbf{J}\cdot\mathbf{J}}{K^2}-2\frac{\mathbf{J}\cdot\mathbf{J}}{mKr}+1\\
=&1+\frac{2J^2}{mK^2}\left(\frac{1}{2}m\dot{\mathbf{r}}^2-\frac{K}{r}\right)\\
=&1+\frac{2J^2E}{mK^2}
\end{aligned}$$

where $E=\frac{1}{2}m\dot{\mathbf{r}}^2-\frac{K}{r}$ is the total energy. So we have the equivalent descriptions

- If $E<0$, the motion orbit is an ellipse
- If $E=0$, the motion orbit is a parabola
- If $E>0$, the motion orbit is a hyperbola

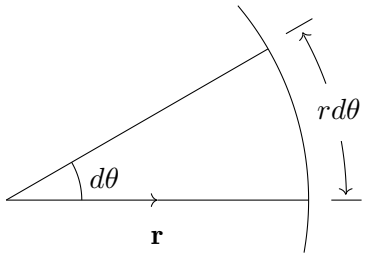
This is Kepler's first law on the orbit shapes.

#### Kepler's Second Law

There is another interesting property about the angular momentain conservation. The rate of sweeping out area is

$$d\text{Area}=\frac{1}{2}r\cdot rd\theta=\frac{1}{2}r^2d\theta$$

$$\Longrightarrow\quad \frac{d\text{Area}}{dt}=\frac{1}{2}r^2\dot{\theta}=\frac{J}{2m}\qquad \text{is a constant.}$$




This is Kepler's second law about the constant rate of sweeping out area.

**Kepler's Third Law**

Consider the elliptic orbit case $E < 0$. The equation is

$$r = \frac{l}{1 + A\cos\theta}, \qquad \text{where} \quad l = \frac{J^2}{Km}, \qquad A < 1.$$

Let us rewrite this into plane coordinate $(x, y)$

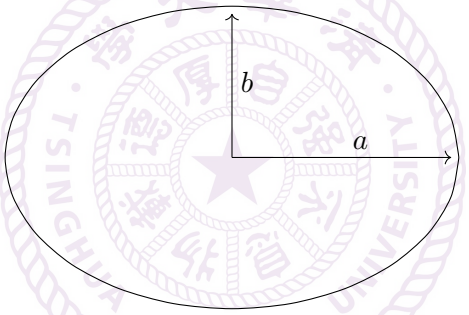
$$x = r\cos\theta, \qquad y = r\sin\theta$$

Then

$$\begin{aligned} & r = l - Ar\cos\theta \\ \Rightarrow \qquad & \sqrt{x^2 + y^2} = l - Ax \\ \Rightarrow \qquad & (x + \frac{lA}{1 - A^2})^2 + \frac{y^2}{1 - A^2} = \frac{l^2}{(1 - A^2)^2} \end{aligned}$$

Let $a, b$ denote the semi-axes of the ellipse. From the above equation, we have

$$a = \frac{l}{1 - A^2}, \qquad b = \frac{l}{\sqrt{1 - A^2}} = \sqrt{la}$$




Let $\tau$ denote the period of the motion. Using

$$\frac{d\text{Area}}{dt} = \frac{J}{2m}$$

and the total area of the ellipse is $\pi ab$, we have

$$\pi ab = \frac{J}{2m}\tau$$

$$\Rightarrow \quad \tau = \frac{2m\pi ab}{J}$$

So

$$\left(\frac{\tau}{2\pi}\right)^2 = \frac{m^2a^2b^2}{J^2} = \frac{m^2a^3l}{J^2} = \frac{ma^3}{K}$$

For Gravitational force, $K = GMm$ where $M$ is the mass of the center. Then

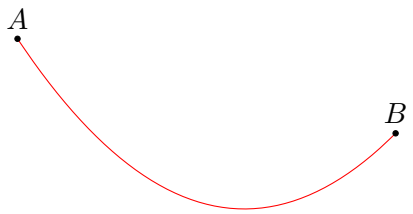
$$\left(\frac{\tau}{2\pi}\right)^2 = \frac{a^3}{GM}$$

So $\frac{\tau^2}{a^3}$ is the same for all trajectories orbiting around the center. This is Kepler's third law.

## 6.3 Brachistochrone Problem

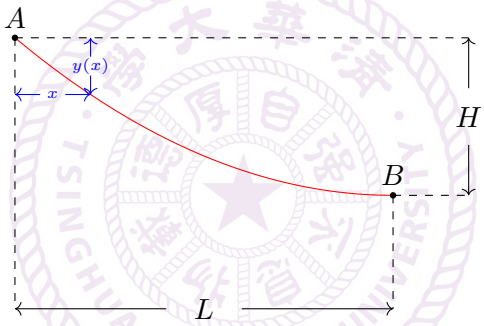
### 6.3.1 Brachistochrone Curve

In 1696, Johann Bernoulli posed the following problem of the brachistochrone: Given two points $A$ and $B$ in a verticle plane, what is the curve traced out by a point acted on only by gravity, which starts at $A$ and reaches $B$ in the shortest time.




The problem of finding the brachistochrone curve, or curve of quickest descent, is the cornerstone and one of the earliest problem for calculus of variations.

Let us pick an arbitrary curve $\gamma$ from $A$ to $B$. We parametrize the curve as follows




The kinetic energy is

$$K = \frac{1}{2}mv^2$$

and the gravitational potential is

$$V(x) = -mgy(x).$$

By energy conservation and the initial condition $v = 0$ at $x = 0$, we have $K + V = 0$

$$\implies \quad v = \sqrt{2gy}$$

The total time of descent along $\gamma$ is

$$T = \int_A^B dt = \int_A^B \frac{ds}{v}$$

where $ds$ is the arclength element

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

Therefore

$$T[y(x)] = \int_0^L \frac{\sqrt{1 + y'(x)^2}}{\sqrt{2gy(x)}}dx$$

can be viewed as a functional for $y(x)$. The curve of quickest descent is such $y(x)$ that extremizes this functional. We treat $T$ as the action functional with Lagrangian

$$\mathcal{L}(y, y') = \frac{1}{\sqrt{2g}} \frac{\sqrt{1 + y'(x)^2}}{\sqrt{y(x)}}.$$

The equation for $y$ can be found by the Euler-Lagrangian Equation

$$\frac{d}{dx}\left(\frac{\partial \mathcal{L}}{\partial y'}\right) = \frac{\partial \mathcal{L}}{\partial y}.$$

To solve this equaiton, consider

$$\mathcal{H} = \frac{\partial \mathcal{L}}{\partial y'} y' - \mathcal{L}.$$

Euler-Lagrangian Equation implies

$$\begin{aligned}
\frac{d}{dx}\mathcal{H} =& \frac{d}{dx}\left[\frac{\partial \mathcal{L}}{\partial y'} y' - \mathcal{L}\right] \\
=& \left(\frac{d}{dx}\left(\frac{\partial \mathcal{L}}{\partial y'}\right) y' + \frac{\partial \mathcal{L}}{\partial y'} y''\right) - \left(y'' \frac{\partial \mathcal{L}}{\partial y'} + y' \frac{\partial \mathcal{L}}{\partial y}\right) \\
=& \left(\frac{d}{dx}\left(\frac{\partial \mathcal{L}}{\partial y'}\right) - \frac{\partial \mathcal{L}}{\partial y}\right) y' = 0.
\end{aligned}$$

Thus $\mathcal{H} = C_1$ is a constant.

*Remark* 6.3.1. In classical mechanics, $\mathcal{L} \to \mathcal{H}$ is the Legendre transform and $\mathcal{H}$ is the Hamiltonian. The above calculation is basically the following well-known statement: Hamiltonian is conserved in the motion if the Lagrangian does not have explicit dependence on time.

Now the integrated equation

$$\frac{\partial \mathcal{L}}{\partial y'} y' - \mathcal{L} = C_1$$

reads

$$\begin{aligned}
& \frac{(y')^2}{\sqrt{y}\sqrt{1 + (y')^2}} - \frac{\sqrt{1 + (y')^2}}{\sqrt{y}} = C_1 \sqrt{2g} \\
\implies \qquad & y(1 + (y')^2) = c, \qquad c \text{ is some constant.}
\end{aligned}$$

We can solve this equation via separation of variables

$$y' = \left(\frac{c - y}{y}\right)^{\frac{1}{2}}$$

$$\implies \qquad \left(\frac{y}{c - y}\right)^{\frac{1}{2}} dy = dx.$$

Consider a change of variable by

$$y = c \sin^2 \phi$$

Then

$$\left(\frac{y}{c - y}\right)^{\frac{1}{2}} = \tan \phi, \qquad dy = 2c \sin \phi \cos \phi d\phi.$$

The equation

$$dx = \left(\frac{y}{c-y}\right)^{\frac{1}{2}} dy = 2c\sin^2\phi d\phi$$

is solved by (using the initial condition $y(0) = 0$)

$$\begin{aligned} x =& \int_0^{\phi} 2c\sin^2 u du \\ =& c\int_0^{\phi}(1-\cos 2u)du \\ =& c(\phi - \frac{1}{2}\sin 2\phi). \end{aligned}$$

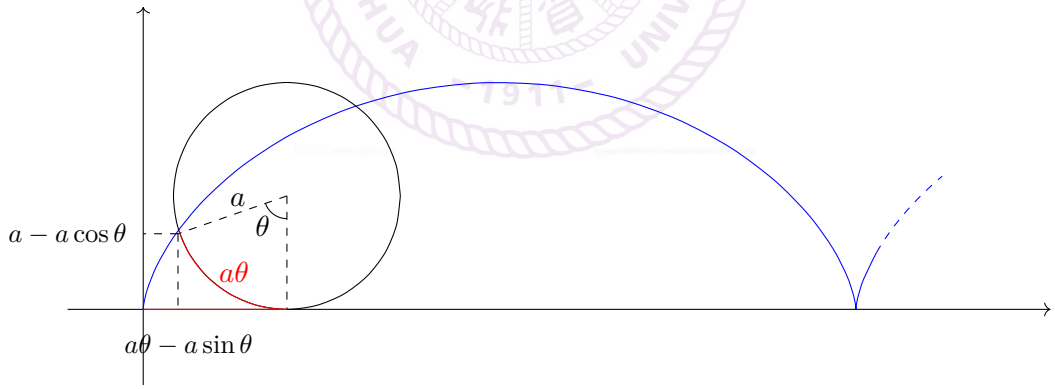
Thus we have found the solution parametrized by

$$\begin{cases} x(\phi) = \frac{c}{2}(2\phi - \sin 2\phi) \\ y(\phi) = \frac{c}{2}(1-\cos 2\phi) \end{cases}$$

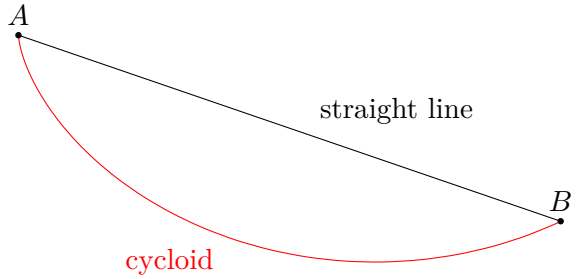
We can reparametrize using $\theta = 2\phi$, $a = \frac{c}{2}$ and get

$$\begin{cases} x(\theta) = a(\theta - \sin\theta) \\ y(\theta) = a(1-\cos\theta) \end{cases}$$

These are the standard parametric equations of the cycloid generated by the rolling of a circle of radius $a$ along the $x$-axis.




Now if turn this cycloid upside-down (since the $y$-coordinate is directed downward as we start with), then it gives the figure for the curve of quickest descent

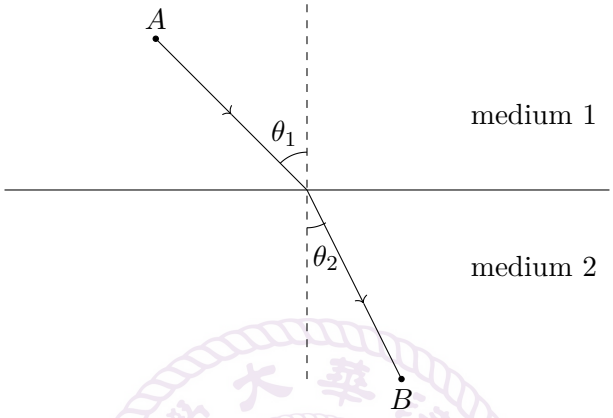


### 6.3.2 Fermat's Principle

Fermat's Principle: light travels along the path that takes the least time. Johann Bernoulli found a beautiful interpretation of the brachistochrone curve via Fermat's principle as follows.

Consider a beam of light traveling in a medium with refraction index $n$. In general, $n$ may vary along the medium, and the speed of light is $\frac{c}{n}$ where $c$ is the speed of light in the vacuum.

**Example 6.3.2.** Consider a light traveling from medium 1 with constant refraction index $n_1$ to medium 2 with constant refraction index $n_2$. Then the path obey's the Snell's law




$$n_1 \sin\theta_1 = n_2 \sin\theta_2$$

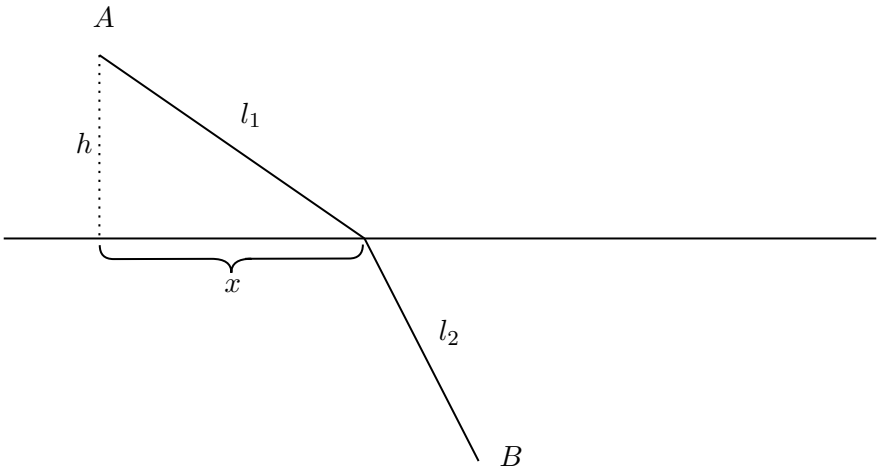
or equivalently

$$\frac{\sin\theta_1}{v_1} = \frac{\sin\theta_2}{v_2}, \qquad v_i = \frac{c}{n_i}.$$

Let us see how this follows from the principle of least time. The travel time is

$$T = \frac{n_1 l_1}{c} + \frac{n_2 l_2}{c}$$

where $l_i$ is the travel distance in medium $i$.




Let $h$ be the distance of $A$ to the interface, and $x$ be the coordinate on the interface as in the figure. Then $l_1 = \sqrt{h^2 + x^2}$. We have

$$\frac{dl_1}{dx} = \frac{x}{\sqrt{h^2 + x^2}} = \sin\theta_1.$$

Similarly consider $l_2$ as a function of $x$ and find $\frac{dl_2}{dx} = -\sin\theta_2$. Minimizing the time $T$ asks for

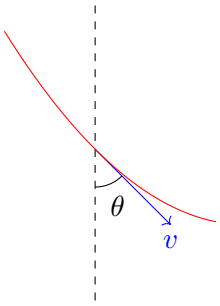
$$\frac{d}{dx}T = n_1\frac{dl_1}{dx} + n_2\frac{dl_2}{dx} = 0,$$

from which we conclude

$$n_1 \sin\theta_1 = n_2 \sin\theta_2.$$

In general, when the light travels in the medium whose refraction index varies vertically (but constant horizontally), then the Snell's law shows that along the trajectory of light

$$\frac{\sin\theta}{v} = \text{constant}$$




Johann Bernoulli noticed that we can think about the brachristochrone curve as a trajectory of a beam of light in a medium where the speed of light increases vertically by gravity ($v = \sqrt{2gy}$ only depends on $y$). Then Snell's law leads to

$$\frac{\sin\theta}{v} = \frac{\frac{dx}{ds}}{v} = \text{constant}$$

Plugging

$$v = \sqrt{2gy}, \qquad \frac{ds}{dx} = \sqrt{1 + (y'(x))^2},$$

we find

$$y(1 + (y')^2) = \text{constant}$$

This precisely leads to the cycloid as we find above.

## 6.4 Isoperimetric Problem

The isoperimetric problem is to determine the closed plane curve of given length that encloses the largest area. This problem was proposed by the ancient Greeks and the answer is the obvious one – a circle. The problem has been extended in many different situations.

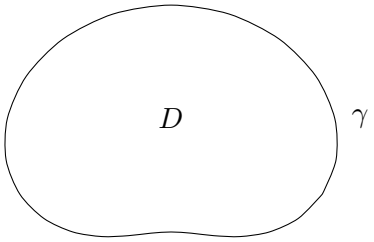
We consider curves on the plane parametrized by differentiable functions

$$\gamma(t) = (x(t), y(t)), \qquad 0 \le t \le 1.$$

The length of the curve is

$$L = \int_0^1 \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}\, dt.$$

Assume $\gamma(0) = \gamma(1)$, so $\gamma$ is a closed curve. Let $D$ be the region enclosed by $\gamma$.




Then the area of $D$ is

$$\begin{aligned} A =& \int_D dx \wedge dy = \frac{1}{2}\int_D d(xdy - ydx) \\ =& \frac{1}{2}\int_\gamma xdy - ydx \\ =& \frac{1}{2}\int_0^1 \left(x\frac{dy}{dt} - y\frac{dx}{dt}\right) dt. \end{aligned}$$

Thus the isoperimetric problem in the case amounts to maximize $A = \frac{1}{2}\int_0^1 \left(x\frac{dy}{dt} - y\frac{dx}{dt}\right) dt$ subject to the condition that $L = \int_0^1 \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2} dt$ is fixed. So this is an example of finding extremals of a functional subject to a constraint.

### 6.4.1 Action Principle with Constraint

One commonly used strategy to find extremals subject to constraints is the method of Lagrange multipliers. We briefly review this. Suppose we want to find extremals of a function

$$f(x, y)$$

subject to the constraint

$$g(x, y) = 0.$$

The constraint leads to a relation between $x$ and $y$, say suppose $x$ is an independent variable and $y$ is expressed as a function of $x$. Then the extremal of $f$ subject to the constraint is at

$$\begin{aligned} & \frac{d}{dx} f(x, y(x)) = 0 \\ \Longrightarrow \quad & \frac{\partial f}{\partial x} + \frac{dy}{dx}\frac{\partial f}{\partial y} = 0. \end{aligned}$$

On the other hand, the constraint $g$ gives

$$\begin{aligned} & \frac{\partial g}{\partial x} + \frac{dy}{dx}\frac{\partial g}{\partial y} = 0 \\ \Longrightarrow \quad & \frac{dy}{dx} = -\frac{\partial_x g}{\partial_y g}. \end{aligned}$$

Thus the extremals of $f$ subject to constraint $g = 0$ are described by the locus

$$\begin{cases} \dfrac{\partial f}{\partial x} - \dfrac{\partial_x g}{\partial_y g}\dfrac{\partial f}{\partial y} = 0 \\ g(x, y) = 0. \end{cases}$$

The method of Lagrange multiplier provides an elegant way to organize the above computations. We introduce a new variable $\lambda$ and form the function

$$F(x, y, \lambda) = f(x, y) + \lambda g(x, y).$$

Then we consider extremals of $F$ as a function of $(x, y, \lambda)$ without any constraint. They are

$$\begin{cases} \dfrac{\partial F}{\partial x} = \dfrac{\partial f}{\partial x} + \lambda \dfrac{\partial g}{\partial x} = 0 \\ \dfrac{\partial F}{\partial y} = \dfrac{\partial f}{\partial y} + \lambda \dfrac{\partial g}{\partial y} = 0 \\ \dfrac{\partial F}{\partial \lambda} = g(x, y) = 0 \end{cases}$$

If we eliminate $\lambda$ from the first two equations, they are reduced to

$$\begin{cases} \dfrac{\partial f}{\partial x} - \dfrac{\partial_x g}{\partial_y g} \dfrac{\partial f}{\partial y} = 0 \\ g(x, y) = 0 \end{cases}$$

which are precisely the equations we found above. One advantage of using $F$ is that it treats $(x, y)$ symmetrically and does not depend on which one is an independent variable. The parameter $\lambda$ is called the Lagrange multiplier.

Now we consider the problem of finding extremals of the action functional

$$S[q(t)] = \int_{t_0}^{t_1} \mathcal{L}(q, \dot{q}, t) dt$$

subject to the condition

$$G[q(t)] = \int_{t_0}^{t_1} R(q, \dot{q}, t) dt = 0$$

in the space of paths $q(t) : [t_0, t_1] \to \mathbb{R}$ with endpoints $q(t_0) = q_0$ and $q(t_1) = q_1$ fixed.

Let $x(t)$ be such an extremal path. We can explore the extremal condition by comparing $x(t)$ with an arbitrary nearby path satisfying the constraint. To deal with finding nearby path subject to the constraint, let $\gamma_1(t)$ and $\gamma_2(t)$ are two arbitrary paths with endpoints

$$\gamma_i(t_0) = \gamma_i(t_1) = 0, \qquad i = 1, 2.$$

Consider nearby paths with two parameters $s_1, s_2$

$$\tilde{x}(t) = x(t) + s_1 \gamma_1(t) + s_2 \gamma_2(t)$$

The two parameters $s_1, s_2$ are not independent, but subject to the condition

$$G[\tilde{x}(t)] = \int_{t_0}^{t_1} R(\tilde{x}, \dot{\tilde{x}}, t) dt = 0.$$

This ensures that $\tilde{x}(t)$ satisfies the constraint.

Now consider the following two functions of $s_1, s_2$

$$f(s_1, s_2) = S[\tilde{x}(t)]$$

$$g(s_1, s_2) = G[\tilde{x}(t)]$$

By assumption, $(s_1, s_2) = (0, 0)$ is an extremal point of $f$ subject to the constraint $g = 0$. We introduce the Lagrangian multiplier $\lambda$ and define

$$F(s_1, s_2, \lambda) = f(s_1, s_2) + \lambda g(s_1, s_2).$$

Then the following equations hold

$$\begin{cases} \dfrac{\partial F}{\partial s_1}(0, 0, \lambda) = \dfrac{\partial f}{\partial s_1}(0, 0) + \lambda \dfrac{\partial g}{\partial s_1}(0, 0) = 0 \\ \dfrac{\partial F}{\partial s_2}(0, 0, \lambda) = \dfrac{\partial f}{\partial s_2}(0, 0) + \lambda \dfrac{\partial g}{\partial s_2}(0, 0) = 0 \\ \dfrac{\partial F}{\partial \lambda}(0, 0, \lambda) = g(0, 0) = 0 \end{cases}$$

Define the new action functional with Lagrange multiplier

$$S_\lambda[q(t)] = \int_{t_0}^{t_1} \mathcal{L}_\lambda(q, \dot{q}, t)dt$$

where

$$\mathcal{L}_\lambda = \mathcal{L} + \lambda R.$$

Then the above equations lead to

$$\begin{cases} \displaystyle\int_{t_0}^{t_1} \left[ \frac{\partial \mathcal{L}_\lambda}{\partial x} - \frac{d}{dt}\left( \frac{\partial \mathcal{L}_\lambda}{\partial \dot{x}} \right) \right] \gamma_i dt = 0, \qquad i = 1, 2 \\ \displaystyle\int_{t_0}^{t_1} R(x, \dot{x}, t)dt = 0 \end{cases}$$

Since $\gamma_1, \gamma_2$ are arbitrary, we conclude that the extremal path $x(t)$ of $S$ subject to the constraint $G = 0$ satisfies the following Euler-Lagrange Equations with Lagrange multiplier

$$\begin{cases} \dfrac{d}{dt}\left( \dfrac{\partial \mathcal{L}_\lambda}{\partial \dot{x}} \right) = \dfrac{\partial \mathcal{L}_\lambda}{\partial x} \\ G[x(t)] = \displaystyle\int_{t_0}^{t_1} R(x, \dot{x}, t)dt = 0 \end{cases}$$

where $\mathcal{L}_\lambda = \mathcal{L} + \lambda R$.

The case with more variables $\{x_i\}$ is similar, and we will have an equation for each independent variable $x_i$

$$\begin{cases} \dfrac{d}{dt}\left( \dfrac{\partial \mathcal{L}_\lambda}{\partial \dot{x}_i} \right) = \dfrac{\partial \mathcal{L}_\lambda}{\partial x_i} \qquad i = 1, 2, \cdots \\ G[x_i(t)] = \displaystyle\int_{t_0}^{t_1} R(x_i, \dot{x}_i, t)dt = 0 \end{cases}$$

### 6.4.2 Isoperimetric Problem

Now we turn back to the original isoperimetric problem. We want to maximize

$$A[x(t), y(t)] = \frac{1}{2}\int_0^1 (x\dot{y} - y\dot{x})dt$$

subject to the constraint

$$G[x(t), y(t)] := \int_0^1 \sqrt{\dot{x}^2 + \dot{y}^2} dt - L = 0.$$

The Lagrangian with multiplier is

$$\mathcal{L}_\lambda = \frac{1}{2}(x\dot{y} - y\dot{x}) + \lambda(\sqrt{\dot{x}^2 + \dot{y}^2} - L).$$

The Euler-Lagrange equation

$$\begin{cases} \dfrac{d}{dt}\left(\dfrac{\partial \mathcal{L}_\lambda}{\partial \dot{x}}\right) = \dfrac{\partial \mathcal{L}_\lambda}{\partial x} \\ \dfrac{d}{dt}\left(\dfrac{\partial \mathcal{L}_\lambda}{\partial \dot{y}}\right) = \dfrac{\partial \mathcal{L}_\lambda}{\partial y} \end{cases}$$

reads

$$\begin{cases} \dfrac{d}{dt}\left(-\dfrac{1}{2}y + \dfrac{\lambda \dot{x}}{\sqrt{\dot{x}^2 + \dot{y}^2}}\right) = \dfrac{1}{2}\dot{y} \\ \dfrac{d}{dt}\left(\dfrac{1}{2}x + \dfrac{\lambda \dot{y}}{\sqrt{\dot{x}^2 + \dot{y}^2}}\right) = -\dfrac{1}{2}\dot{x} \end{cases}$$

These two equations can be integrated

$$\begin{cases} \dfrac{\lambda \dot{x}}{\sqrt{\dot{x}^2 + \dot{y}^2}} = y - c_2 \\ \dfrac{\lambda \dot{y}}{\sqrt{\dot{x}^2 + \dot{y}^2}} = -(x - c_1) \end{cases}$$

$$\Longrightarrow \quad (x - c_1)^2 + (y - c_2)^2 = \lambda^2.$$

So the maximizing curve is a circle. The Lagrangian multiplier has the interpretation of the radius and is solved by the constraint: $\lambda = \frac{L}{2\pi}$. The extremal values of $A$ are $\frac{L^2}{4\pi}$ for maximal and $-\frac{L^2}{4\pi}$ for minimal, where both curves are circles but with opposite orientations.

# Chapter 7 Numerical Solutions

Finding exact solutions of differential equations is extremely difficulty and unrealistic in general. In practice, we would like to find approximate solutions with explicit algorithm that can be implemented in computer programs and can approximate the real solutions to a controlled accuracy. This will be extremely useful for applications in real life problems. In this chapter, we explain some basic ideas about numerical methods of solving ordinary differential equations.

## 7.1 Euler's Method

### 7.1.1 Difference Equation

Consider the initial value problem

$$\begin{cases} y'(t) = F(y,t) & \text{on } t \in [t_0, T] \\ y(t_0) = y_0 \end{cases}$$

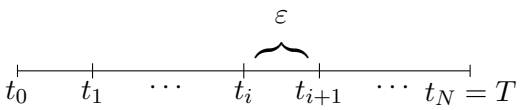
As we discussed in Chapter 3, with appropriate assumptions on $F$, this initial value problem has a unique solution on $[t_0, T]$, but the explicit form of the solution is difficult to find in general.

The simplest numerical method for solving the initial value problem is Euler's method. Though Euler's method is not an effective algorithm, it illustrates many aspects of key ideas for numerical solutions in general.

To start with, we first subdivide the interval $[t_0, T]$ by the mesh-points

$$t_i = t_0 + i\varepsilon, \qquad i = 0, \cdots, N$$

with step-size $\varepsilon = \frac{T - t_0}{N}$.




A numerical solution is to assign a value $y_i$ for each mesh-point $t_i$ such that $y_i$ approximates the value $y(t_i)$ of the true solution at $t = t_i$.

Euler's method is to approximate the differential equation

$$y'(t) = F(y,t)$$

by the difference equation

$$\frac{y_{i+1} - y_i}{\varepsilon} = F(y_i, t_i)$$

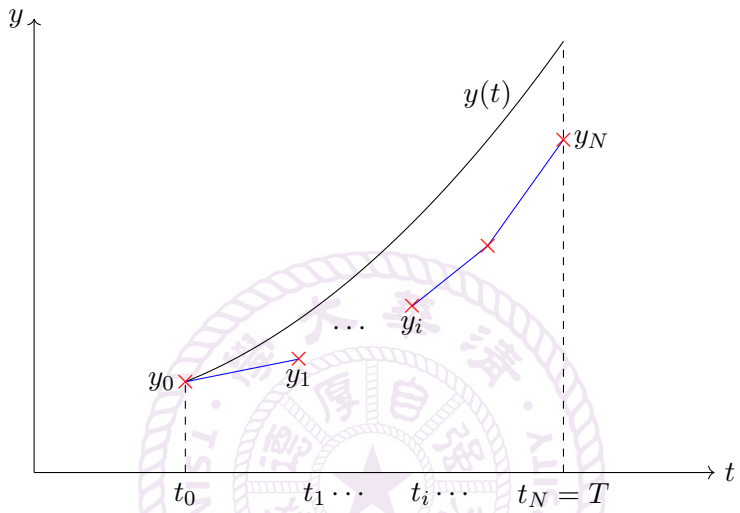
*i.e.*

$$y_{i+1} = y_i + \varepsilon F(y_i, t_i).$$

Starting with $y_0$ as given, this formula gives $y_1, y_2, \cdots, y_N$ iteratively. Since

$$\lim_{\varepsilon \to 0} \frac{y(t+\varepsilon) - y(t)}{\varepsilon} = y'(t)$$

we would expect that the difference equation will approximate the true function as $\varepsilon \to 0$.




Throughout this chapter, we will use $y(t_i)$ for the value of true solution $y(t)$ at $t_i$ and use $y_i$ for the approximate value at $t_i$ obtained from numerical methods.

### 7.1.2 Error Analysis

The notion of local truncation error (LTE) $\tau_i$ describes the difference between the value of true solution at step $t_i$ and the approximate value obtained via one-step iteration from the previous steps using true values. It is "local" in the sense that it uses the true values of solution and does not include errors created in precious steps.

The local truncation error (LTE) $\tau_i$ of Euler's method at time $t_i$ is the quantity

$$\tau_i := y(t_i) - (y(t_{i-1}) + \varepsilon F(y(t_{i-1}), t_{i-1}))$$

It uses the true solution $y(t)$ at the mesh-point to describe the error incurred by one-step application of Euler's method.

The notion of global truncation error $e_i$ describes the difference between the value $y(t_i)$ of true solution and the approximate value $y_i$ from a numerical method

$$e_i := y(t_i) - y_i.$$

We say the numerical method is convergent if

$$\lim_{\varepsilon \to 0} \max_{0 \le i \le N} |e_i| = 0.$$

In numerical method, it is important to establish an estimate of the global truncation error from the local truncation error. In Euler's method, we have

**Lemma 7.1.1.** *Suppose $F(y,t)$ is Lipschitz in $y$ with Lipschitz constant $L$. Then*

$$|e_i| \le \frac{\max\limits_{1 \le k \le i} |\tau_k|}{\varepsilon} \left( \frac{e^{L(t_i - t_0)} - 1}{L} \right).$$

*In particular, we have*

$$\max_{0 \le i \le N} |e_i| \le \frac{\max\limits_{1 \le i \le N} |\tau_i|}{\varepsilon} \left( \frac{e^{L(T - t_0)} - 1}{L} \right).$$

*Proof:* Recall

$$\begin{aligned} y(t_i) =& y(t_{i-1}) + \varepsilon F(y(t_{i-1}), t_{i-1}) + \tau_i \\ y_i =& y_{i-1} + \varepsilon F(y_{i-1}, t_{i-1}) \end{aligned}$$

Subtracting these two equations, we find

$$e_i = e_{i-1} + \varepsilon (F(y(t_{i-1}), t_{i-1}) - F(y_{i-1}, t_{i-1})) + \tau_i.$$

Using the Lipschitz condition for $F$

$$\begin{aligned} |e_i| \le& |e_{i-1}| + \varepsilon |F(y(t_{i-1}), t_{i-1}) - F(y_{i-1}, t_{i-1})| + |\tau_i| \\ \le& (1 + \varepsilon L) |e_{i-1}| + |\tau_i| \end{aligned}$$

Using the initial condition $e_0 = 0$ (we choose the true value at $t = t_0$ from given)

$$\begin{aligned} |e_i| \le& (1 + \varepsilon L) \left( (1 + \varepsilon L) |e_{i-2}| + |\tau_{i-1}| \right) + |\tau_i| \\ \le& \cdots \\ \le& \sum_{k=1}^{i} (1 + \varepsilon L)^{i-k} |\tau_k| \\ \le& \max_{1 \le k \le i} |\tau_k| \left( \frac{(1 + \varepsilon L)^i - 1}{\varepsilon L} \right) \\ \le& \max_{1 \le k \le i} |\tau_k| \left( \frac{e^{\varepsilon i L} - 1}{\varepsilon L} \right) \\ =& \max_{1 \le k \le i} |\tau_k| \left( \frac{e^{L(t_i - t_0)} - 1}{\varepsilon L} \right). \end{aligned}$$

□

*Remark* 7.1.2. If we have an error $e_0$ at the initial $t = t_0$, then a similar argument leads to

$$|e_i| \le e^{L(t_i - t_0)} |e_0| + \frac{\max\limits_{1 \le k \le i} |\tau_k|}{\varepsilon} \left( \frac{e^{L(t_i - t_0)} - 1}{L} \right).$$

**Theorem 7.1.3** (Convergence of Euler's Method)**.** *Suppose the true solution $y(t)$ satisfies*

$$\max_{t\in[t_0,T]} |y''(t)| \le M$$

*and $F(y,t)$ is Lipschitz in $y$ with Lipschitz constant $L$. Then*

$$|e_i| \le \frac{M\varepsilon}{2L}(e^{L(t_i-t_0)} - 1) \qquad \text{for} \quad 0 \le i \le N$$

*In particular, we have*

$$\max_{0\le i\le N} |e_i| \le \frac{M\varepsilon}{2L}(e^{L(T-t_0)} - 1)$$

*which goes to zero as $\varepsilon \to 0$.*

*Proof:* Using the Taylor series expansion

$$y(t_i) = y(t_{i-1}) + \varepsilon y'(t_{i-1}) + \frac{\varepsilon^2}{2}y''(\xi)$$

for some $\xi \in [t_{i-1}, t_i]$, we have the following estimate of local truncation error

$$\begin{aligned}|\tau_i| =& |y(t_i) - (y(t_{i-1}) + \varepsilon F(y(t_{i-1}), t_{i-1}))| \\ =& |y(t_i) - y(t_{i-1}) - \varepsilon y'(t_{i-1})| \\ =& \frac{\varepsilon^2}{2}|y''(\xi)| \le \frac{\varepsilon^2}{2}M.\end{aligned}$$

The theorem now follows from Lemma 7.1.1. □

### 7.1.3 Backward Euler's Method

We can modify Euler's method in several different ways. One modification is the backward Euler's method which has iteration

$$y_{i+1} = y_i + \varepsilon F(y_{i+1}, t_{i+1}).$$

Here the function $F$ is evaluated at $(y_{i+1}, t_{i+1})$ rather than $(y_i, t_i)$ as in the previous Euler's method (also called forward Euler's method).

In constract to the forward Euler's method, the above iteration gives an implicit relation between $y_i$ and $y_{i+1}$. This is the simplest example of an implicit method. To solve $y_{i+1}$, we can iterate Newton's method until convergence. This seems more complicated than the forward Euler's method. The reason we want to use the backward method lies in the fact that it has better stability properties. Similar to Theorem 7.1.3, the backward Euler's method has the same convergence property as the forward one.

### 7.1.4 Trapezoidal Method

The trapezoidal method is a mixing of forward and backward Euler's method with iteration

$$y_{i+1} = y_i + \frac{\varepsilon}{2}(F(y_i, t_i) + F(y_{i+1}, t_{i+1}))$$

which is also an implicit method. The advantage of trapezoidal method is that it converges to the true solution faster than both the forward and backward Euler's method.

To see this, we compare the iteration with the true solution

$$\begin{aligned} y_i =& y_{i-1} + \frac{\varepsilon}{2}(F(y_{i-1}, t_{i-1}) + F(y_i, t_i)) \\ y(t_i) =& y(t_{i-1}) + \frac{\varepsilon}{2}(F(y(t_{i-1}), t_{i-1}) + F(y(t_i), t_i)) + \tau_i \end{aligned}$$

where such defined $\tau_i$ is the local truncation error for the trapezoidal method. Let

$$e_i = y(t_i) - y_i$$

be the global truncation error. Subtracting the above two equations and assume Lipschitz constant $L$ for $F$ in the variable $y$, we get

$$\begin{aligned} |e_i| &\leq |e_{i-1}| + \frac{\varepsilon}{2}L(|e_{i-1}| + |e_i|) + |\tau_i| \\ \Rightarrow |e_i| &\leq \frac{1 + \frac{\varepsilon}{2}L}{1 - \frac{\varepsilon}{2}L}|e_{i-1}| + \frac{|\tau_i|}{1 - \frac{\varepsilon}{2}L} \end{aligned}$$

From this we can get a similar estimate from local error to global error as in Lemma 7.1.1.

Now let us focus on the local truncation error

$$\begin{aligned} \tau_i =& y(t_i) - y(t_{i-1}) - \frac{\varepsilon}{2}(F(y(t_{i-1}), t_{i-1}) + F(y(t_i), t_i)) \\ =& y(t_i) - y(t_{i-1}) - \frac{\varepsilon}{2}(y'(t_{i-1}) + y'(t_i)). \end{aligned}$$

Recall the following Trapezoidal rule

$$\int_a^b f(t)dt = \frac{1}{2}(b-a)(f(a) + f(b)) - \frac{1}{12}(b-a)^3 f''(\xi)$$

for $\xi \in [a, b]$. Applying this to $f(t) = y'(t)$, we find

$$\tau_i = -\frac{1}{12}\varepsilon^3 y'''(\xi) \qquad \text{for some } \xi \in [t_{i-1}, t_i].$$

Thus if $y'''(t)$ is bounded on $[0, T]$, the local truncation error $\tau_i$ has order $\varepsilon^3$, which is better than that $\varepsilon^2$ for Euler's method.

## 7.2 Higher-Order Methods

Euler's method uses the linear Taylor approximation. It is natural to consider higher-order Taylor expansions to achieve approximations of higher accuracy.

### 7.2.1 Taylor Method

Consider the solution of

$$y'(t) = F(y, t).$$

The equation allows us to obtain an expression for higher derivatives $y^{(n)}(t)$ in terms of $F$ by taking further derivatives. For example

$$y^{(2)}(t) = \frac{d}{dt}y^{(1)}(t) \overset{\text{Using Eqn}}{=} \frac{d}{dt}F(y,t) = y'(t)\partial_y F + \partial_t F \overset{\text{Using Eqn}}{=} F\partial_y F + \partial_t F.$$

In general, a similar argument leads to the pattern

$$y^{(k)}(t) = \left(F\frac{\partial}{\partial y} + \frac{\partial}{\partial t}\right)^{k-1} F$$

which are expressed by $F$ and its derivatives. For example,

$$\begin{aligned} y^{(3)}(t) &= \left(F\frac{\partial}{\partial y} + \frac{\partial}{\partial t}\right)(F\partial_y F + \partial_t F) \\ &= F^2\partial_y^2 F + F(\partial_y F)^2 + 2F\partial_y\partial_t F + \partial_t F\partial_y F + \partial_t^2 F. \end{aligned}$$

Let us denote the expression

$$P_k[F] := \left(F\frac{\partial}{\partial y} + \frac{\partial}{\partial t}\right)^{k-1} F.$$

Then we can obtain a numerical iteration method by Taylor approximation up to order $n$

$$y_{i+1} = y_i + \sum_{k=1}^{n} \frac{\varepsilon^n}{k!} P_k[F(y_i, t_i)].$$

For the case $n = 1$, this is Euler's method. For $n > 1$, this is generally called Taylor's method.

**Example 7.2.1** (Taylor's method for $n = 2$)**.** The iteration is

$$y_{i+1} = y_i + \varepsilon F(y_i, t_i) + \frac{\varepsilon^2}{2}[F(y_i, t_i)\partial_y F(y_i, t_i) + \partial_t F(y_i, t_i)].$$

### 7.2.2 Runge-Kutta Method

The Taylor method is conceptually easier to work with but time-consuming to calculate the higher-order derivatives. The more effecitve Runge-Kutta method allows to retain the accuracy of higher order Taylor approximation by evaluating $F$ at more intermediate points.

To illustrate the basic idea, let us add one intermediate point at

$$t_i^* = t_i + \frac{\varepsilon}{2}$$

and consider the iteration of the form

$$y_{i+1} = y_i + \varepsilon(\omega_1\kappa_1 + \omega_2\kappa_2)$$

where

$$\begin{aligned} \kappa_1 &= F(y_i, t_i) \\ \kappa_2 &= F(y_i + \varepsilon\beta\kappa_1, t_i + \frac{\varepsilon}{2}) \end{aligned}$$

and $\omega_1, \omega_2, \beta$ are constants to be found.

Let us consider the local truncation error defined by

$$\tau_{i+1} = y(t_{i+1}) - y(t_i) - \varepsilon(\omega_1 F(y(t_i), t_i) + \omega_2 F(y(t_i) + \varepsilon\beta F(y(t_i), t_i), t_i + \frac{\varepsilon}{2}))$$

The term $y(t_{i+1}) - y(t_i)$ can be expanded by Taylor series

$$\begin{aligned} y(t_{i+1}) - y(t_i) =& \varepsilon y'(t_i) + \frac{\varepsilon^2}{2} y''(t_i) + O(\varepsilon^3) \\ =& \varepsilon F(y(t_i), t_i) + \frac{\varepsilon^2}{2}[F\partial_y F + \partial_t F]|_{y(t_i),t_i} + O(\varepsilon^3) \end{aligned}$$

On the other hand,

$$\begin{aligned} &\omega_1 F(y(t_i), t_i) + \omega_2 F(y(t_i) + \varepsilon\beta F(y(t_i), t_i), t_i + \frac{\varepsilon}{2}) \\ =&[\omega_1 F + \omega_2 (F + \varepsilon\beta F\partial_y F + \frac{\varepsilon}{2}\partial_t F)]|_{y(t_i),t_i} + O(\varepsilon^2). \end{aligned}$$

Combining the above two computations, we find

$$\tau_{i+1} = \varepsilon[1 - \omega_1 - \omega_2]F(y(t_i), t_i) + \frac{\varepsilon^2}{2}[(1 - 2\omega_2\beta)F\partial_y F + (1 - \omega_2)\partial_t F]|_{y(t_i),t_i} + O(\varepsilon^3).$$

Therefore we can achieve accuracy for $\tau_{i+1} = O(\varepsilon^3)$ by choosing

$$\begin{cases} 1 - \omega_1 - \omega_2 = 0 \\ 1 - 2\omega_2\beta = 0 \\ 1 - \omega_2 = 0 \end{cases}$$

$$\Rightarrow \qquad \omega_1 = 0, \qquad \omega_2 = 1, \qquad \beta = \frac{1}{2}.$$

Thus we have arrived at the following two-stage Runge-Kutta iteration method

$$y_{i+1} = y_i + \varepsilon F(y_i + \frac{\varepsilon}{2} F(y_i, t_i), t_i + \frac{\varepsilon}{2})$$

which is also called modified Euler's method.

Now the general idea is similar, we can add more intermediate points to achieve higher order accuracy. The famous four-stage Runge-Kutta method is the iteration

$$y_{i+1} = y_i + \frac{\varepsilon}{6}(\kappa_1 + 2\kappa_2 + 2\kappa_3 + \kappa_4)$$

where

$$\begin{cases} \kappa_1 =& F(y_i, t_i) \\ \kappa_2 =& F(y_i + \frac{\varepsilon}{2}\kappa_1, t_i + \frac{\varepsilon}{2}) \\ \kappa_3 =& F(y_i + \frac{\varepsilon}{2}\kappa_2, t_i + \frac{\varepsilon}{2}) \\ \kappa_4 =& F(y_i + \varepsilon\kappa_3, t_i + \varepsilon) \end{cases}$$

The local truncation error will have order $\tau_i = O(\varepsilon^5)$.

### 7.2.3 Linear Multi-Step Method

Taylor methods and Runge-Kutta methods are known as one-step methods, since the iteration for computing $y_{i+1}$ is determined solely from $y_i$. In general, we can consider multi-step methods to improve the approximation in which $y_{i+1}$ is determined from previous several steps.

A general linear multi-step method has the form of iteration

$$\begin{aligned} y_{i+1} =& \alpha_1 y_i + \alpha_2 y_{i-1} + \cdots + \alpha_p y_{i-p+1} \\ &+ \varepsilon[\beta_0 F(y_{i+1}, t_{i+1}) + \beta_1 F(y_i, t_i) + \cdots + \beta_p F(y_{i-p+1}, t_{i-p+1})], \qquad i \geq p-1 \end{aligned}$$

where we assume $|\alpha_p| + |\beta_p| \neq 0$. This is considered as the $p$-step method since $p$-previous solution values are being used to compute the next one.

**Example 7.2.2.** The $p$-step Adams-Bashforth method is the iteration of the form

$$y_{i+1} = y_i + \varepsilon(\beta_1 F(y_i, t_i) + \cdots + \beta_p F(y_{i-p+1}, t_{i-p+1}))$$

where the constants $\{\beta_1, \cdots, \beta_p\}$ are chosen to give the highest order accuracy.

Let us consider the local truncation error defined by

$$\begin{aligned} \tau_{i+1} =& y(t_{i+1}) - y(t_i) - \varepsilon(\beta_1 F(y(t_i), t_i) + \cdots + \beta_p F(y(t_{i-p+1}), t_{i-p+1})) \\ =& y(t_{i+1}) - y(t_i) - \varepsilon(\beta_1 y'(t_i) + \cdots + \beta_p y'(t_{i-p+1})) \end{aligned}$$

Then we can Taylor expand all terms at the point $t_{i+1}$ and choose $\beta$'s to achieve the highest order approximation. Let us consider the 3-step example for $p = 3$

$$\begin{aligned} \tau_{i+1} =& y(t_{i+1}) - y(t_i) - \varepsilon(\beta_1 y'(t_i) + \beta_2 y'(t_{i-1}) + \beta_3 y'(t_{i-2})) \\ =& y(t_{i+1}) - [y(t_{i+1}) - \varepsilon y'(t_{i+1}) + \frac{\varepsilon^2}{2} y''(t_{i+1}) - \frac{\varepsilon^3}{6} y'''(t_{i+1}) + O(\varepsilon^4)] \\ & - \varepsilon\beta_1[y'(t_{i+1}) - \varepsilon y''(t_{i+1}) + \frac{\varepsilon^2}{2} y'''(t_{i+1}) + O(\varepsilon^3)] \\ & - \varepsilon\beta_2[y'(t_{i+1}) - 2\varepsilon y''(t_{i+1}) + 2\varepsilon^2 y'''(t_{i+1}) + O(\varepsilon^3)] \\ & - \varepsilon\beta_3[y'(t_{i+1}) - 3\varepsilon y''(t_{i+1}) + \frac{9\varepsilon^2}{2} y'''(t_{i+1}) + O(\varepsilon^3)] \\ =& c_1\varepsilon y'(t_{i+1}) + c_2\varepsilon^2 y''(t_{i+1}) + c_3\varepsilon^3 y'''(t_{i+1}) + O(\varepsilon^4) \end{aligned}$$

where

$$\begin{cases} c_1 = 1 - \beta_1 - \beta_2 - \beta_3 \\ c_2 = -\frac{1}{2} + \beta_1 + 2\beta_2 + 3\beta_3 \\ c_3 = \frac{1}{6} - \frac{1}{2}\beta_1 - 2\beta_2 - \frac{9}{2}\beta_3 \end{cases}$$

To achieve the best accuracy, it is preferred to choose $\beta_1, \beta_2, \beta_3$ such that $c_1 = c_2 = c_3 = 0$.

$$\Rightarrow \qquad \beta_1 = \frac{23}{12} \qquad \beta_2 = -\frac{4}{3} \qquad \beta_3 = \frac{5}{12}$$

Thus the 3-step Adam-Bashforth method is explicitly given by the iteration

$$y_{i+1} = y_i + \varepsilon\left[\frac{23}{12}F(y_i, t_i) - \frac{4}{3}F(y_{i-1}, t_{i-1}) + \frac{5}{12}F(y_{i-2}, t_{i-2})\right].$$

**Example 7.2.3.** The $p$-step Adams-Moulton method is the iteration of the form

$$y_{i+1} = y_i + \varepsilon[\beta_0 F(y_{i+1}, t_{i+1}) + \beta_1 F(y_i, t_i) + \cdots + \beta_p F(y_{i-p+1}, t_{i-p+1})]$$

where the constants $\{\beta_0, \beta_1, \cdots, \beta_p\}$ are chosen to given the highest order accuracy. Note that $\beta_0 \neq 0$ in this case, thus Adams-Moulton is an implicit method in contrast to the explicit Adams-Bashforth. Let us again consider the 3-step case. The local truncation error is

$$\begin{aligned}\tau_{i+1} =& y(t_{i+1}) - y(t_i) - \varepsilon[\beta_0 y'(t_{i+1}) + \beta_1 y'(t_i) + \beta_2 y'(t_{i-1}) + \beta_3 y'(t_{i-2})] \\ =& c_1\varepsilon y'(t_{i+1}) + c_2\varepsilon^2 y''(t_{i+1}) + c_3\varepsilon^3 y'''(t_{i+1}) + c_4\varepsilon^4 y''''(t_{i+1}) + O(\varepsilon^5)\end{aligned}$$

where

$$\begin{aligned}c_1 =& 1 - \beta_0 - \beta_1 - \beta_2 - \beta_3 \\ c_2 =& -\frac{1}{2} + \beta_1 + 2\beta_2 + 3\beta_3 \\ c_3 =& \frac{1}{6} - \frac{1}{2}\beta_1 - 2\beta_2 - \frac{9}{2}\beta_3 \\ c_4 =& -\frac{1}{24} + \frac{1}{6}\beta_1 + \frac{4}{3}\beta_2 + \frac{9}{2}\beta_3\end{aligned}$$

Asking $c_1 = c_2 = c_3 = c_4 = 0$ solves

$$\beta_0 = \frac{3}{8} \qquad \beta_1 = \frac{19}{24} \qquad \beta_2 = -\frac{5}{24} \qquad \beta_3 = \frac{1}{24}$$

## 7.3 Stability and Convergence

We focus on the linear p-step method

$$\begin{aligned}y_{i+1} =& \alpha_1 y_i + \alpha_2 y_{i-1} + \cdots + \alpha_p y_{i-p+1} \\ &+ \varepsilon[\beta_0 F(y_{i+1}, t_{i+1}) + \beta_1 F(y_i, t_i) + \cdots + \beta_p F(y_{i-p+1}, t_{i-p+1})], \qquad i \geq p-1.\end{aligned}$$

To apply this iteration, we need $p$ starting values

$$y_0, y_1, \cdots, y_{p-1}$$

The $y_0$ is given by the initial data. The others $y_1, \cdots, y_{p-1}$ have to be computed first, say by using a Runge-Kutta method. It is thus important to understand the approximation error of the numerical solutions arising from errors in the starting values and errors from the iterations.

### 7.3.1 Zero-Stability

**Definition 7.3.1.** A linear $p$-step method is said to be zero-stable if there exists a constant $K$ such that for any two sequences $\{y_i\}$ and $\{\tilde{y}_i\}$ generated by the iteration from the starting values $\{y_0, \cdots, y_{p-1}\}$ and $\{\tilde{y}_0, \cdots, \tilde{y}_{p-1}\}$ respectively, we have

$$|y_i - \tilde{y}_i| \leq K \sup_{0 \leq j \leq p} |y_j - \tilde{y}_j|$$

for all $t_i \leq T$ and as $\varepsilon \to 0$.

Thus "zero-stability" can be understood intuitively as saying "small perturbation at starting data gives rise to small perturbation at output". It turns out that to check the zero-stability of the method, we only need to check it for the trivial differential equation

$$y' = 0$$

This explains the name "zero-stability".

Let us apply the above $p$-step method to the trivial equation $y' = 0$. The iteration becomes

$$y_{i+1} = \alpha_1 y_i + \alpha_2 y_{i-1} + \cdots + \alpha_p y_{i-p+1} \qquad (*)$$

Let us analyze the stability of this iteration.

We consider its characteristic polynomial defined by

$$\rho(z) = z^p - \alpha_1 z^{p-1} - \alpha_2 z^{p-2} - \cdots - \alpha_p.$$

Note that if $\lambda$ is a root of $\rho$, then $\rho(\lambda) = 0$ implies that the sequence

$$y_i = \lambda^i \qquad \text{solves } (*)$$

This observation allows us to find the general solution of $(*)$ as follows

Case 1: Assume $\rho(z) = 0$ has $p$ distinct roots

$$\lambda_1, \lambda_2, \cdots, \lambda_p.$$

Then $(*)$ can be solved by

$$y_i = c_1\lambda_1^i + c_2\lambda_2^i + \cdots + c_p\lambda_p^i \qquad i \geq 0.$$

Here $c_k$'s are constants. These constants can be determined from the starting values $y_0, y_1, \cdots, y_{p-1}$. In fact, the starting values give the relation

$$\underbrace{\begin{pmatrix} 1 & 1 & \cdots & 1 \\ \lambda_1 & \lambda_2 & \cdots & \lambda_p \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_1^{p-1} & \lambda_2^{p-1} & \cdots & \lambda_p^{p-1} \end{pmatrix}}_{=A} \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_p \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_{p-1} \end{pmatrix}$$

The determinant of the marix $A$ is known as the Vandermonde determinant

$$\det A = \prod_{i<j} (\lambda_i - \lambda_j) \neq 0$$

Thus $A$ is invertible and we can solve $c_k$'s by

$$\begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_p \end{pmatrix} = A^{-1} \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_{p-1} \end{pmatrix}$$

Case 2: Assume $\rho(z) = 0$ has repeated roots

$$\rho(z) = (z - \lambda_1)^{m_1} \cdots (z - \lambda_l)^{m_l}$$

where $m_k$ is the multiplicity of the root $\lambda_k$.

Consider the root $\lambda_1$ with multiplicity $m_1 > 1$. Then for any $k < m_1$, $i \geq p$

$$\left(\frac{d}{dz}\right)^k (\rho(z) z^{i-p})|_{z=\lambda_1} = 0.$$

This implies that the following sequence

$$y_i = \lambda_1^k \left(\frac{d}{dz}\right)^k z^i \Big|_{z=\lambda_1} = i(i-1)\cdots(i-k+1)\lambda_1^i$$

solves the equation $(*)$. Thus $(*)$ can be solved in general by

$$\begin{aligned} y_i =& (c_{11} + c_{12} i + \cdots + c_{1m_1} i^{m_1 - 1})\lambda_1^i \\ &+ (c_{21} + c_{22} i + \cdots + c_{2m_2} i^{m_2 - 1})\lambda_2^i \\ &+ \cdots \\ &+ (c_{l1} + c_{l2} i + \cdots + c_{lm_l} i^{m_l - 1})\lambda_l^i. \end{aligned}$$

Here $c_{**}$'s are constants, which again can be determined from the starting values $y_0, y_1, \cdots, y_{p-1}$.

**Theorem 7.3.2** (Root Condition)**.** *A linear k-step method is zero-stable if and only if* $\rho(z) = (z - \lambda_1)^{m_1} \cdots (z - \lambda_l)^{m_l}$ *satisfies the following conditions*

- $|\lambda_k| \leq 1$ *for* $k = 1, \cdots, l$
- *If* $|\lambda_k| = 1$*, then* $\lambda_k$ *is a simple root,* i.e. $m_k = 1$.

*Proof:* As we mentioned above (without proof), it is enough to check the stability for the trivial equation $y' = 0$ where the iteration is $(*)$. Let us assume this fact.

The general solution of $(*)$ takes the form

$$y_i = c_1(i)\lambda_1^i + c_2(i)\lambda_2^i + \cdots + c_l(i)\lambda_l^i$$

where $c_k(i)$ is a polynomial in $i$ of degree $< m_k$. The zero-stability in this case is equivalent to saying that for any choice of $c_1(i), \cdots, c_l(i)$ (which is linearly equivalent to choice of starting values of $y_0, \cdots, y_{p-1}$), the sequence $\{y_i\}$ should be bounded. Thus all $\lambda_k$'s should satisfy $|\lambda_k| \leq 1$. And for $|\lambda_k| = 1$, the polynomial $c_k(i)$ can not depend on $i$, *i.e.* $m_k = 1$. □

*Remark* 7.3.3. Another way to see this is to express the iteration as a matrix relation

$$\mathbf{Y}_{i+1} = B\mathbf{Y}_i, \qquad i \geq p - 1$$

where $\mathbf{Y}_i$ is the column vector

$$\mathbf{Y}_i = \begin{pmatrix} y_i \\ y_{i-1} \\ \vdots \\ y_{i-p+1} \end{pmatrix}, \qquad i \geq p-1$$

and $B$ is the matrix

$$\begin{pmatrix} \alpha_1 & \alpha_2 & \cdots & \cdots & \alpha_p \\ 1 & 0 & & & \\ & 1 & 0 & & \\ & & \ddots & \ddots & \\ & & & 1 & 0 \end{pmatrix}$$

The initial vector $\mathbf{Y}_{p-1}$ collects the starting values. Then the stability asks whether the sequence of vectors

$$B^k \mathbf{Y}_{p-1}$$

will be bounded as $k \to +\infty$. The characteristic polynomial of $B$ is precisely

$$\det(z - B) = \rho(z) = (z - \lambda_1)^{m_1} \cdots (z - \lambda_l)^{m_l}.$$

For any eigenvalue $\lambda$ of $B$, there exists only one eigenvector. In fact, the eigenvalue equation

$$B\mathbf{u} = \lambda \mathbf{u} \qquad \mathbf{u}^T = (u_1, \cdots, u_p)$$

reads

$$\begin{cases} \alpha_1 u_1 + \alpha_2 u_2 + \cdots + \alpha_p u_p = \lambda u_1 \\ u_1 = \lambda u_2 \\ \vdots \\ u_{p-1} = \lambda u_p \end{cases}$$

This clearly has only one solution up to a rescaling constant. Therefore each eigenvalue of $B$ has only one Jordan block and there exists an invertible matrix $P$ such that

$$P^{-1}BP = \begin{pmatrix} \begin{pmatrix} \lambda_1 & 1 & \\ & \ddots & 1 \\ & & \lambda_1 \end{pmatrix} & & & \\ & \begin{pmatrix} \lambda_2 & 1 & \\ & \ddots & 1 \\ & & \lambda_2 \end{pmatrix} & & \\ & & \ddots & \\ & & & \begin{pmatrix} \lambda_l & 1 & \\ & \ddots & 1 \\ & & \lambda_l \end{pmatrix} \end{pmatrix}$$

It is clear that $\lim_{k \to +\infty} B^k$ is bounded if and only if the root condition in Theorem 7.3.2 holds.

### 7.3.2 Convergence

The local truncation error of the linear $p$-step method is

$$\begin{aligned}\tau_{i+1} =& y(t_{i+1}) - (\alpha_1 y(t_i) + \alpha_2 y(t_{i-1}) + \cdots + \alpha_p y(t_{i-p+1})) \\ & - \varepsilon(\beta_0 F(y(t_{i+1}), t_{i+1}) + \beta_1 F(y(t_i), t_i) + \cdots + \beta_p F(y(t_{i-p+1}), t_{i-p+1})) \\ =& y(t_{i+1}) - (\alpha_1 y(t_i) + \alpha_2 y(t_{i-1}) + \cdots + \alpha_p y(t_{i-p+1})) \\ & - \varepsilon(\beta_0 y'(t_{i+1}) + \beta_1 y'(t_i) + \cdots + \beta_p y'(t_{i-p+1})).\end{aligned}$$

**Definition 7.3.4.** The linear $p$-step method is called consistent if

$$\lim_{\varepsilon \to 0} \frac{\tau_i}{\varepsilon} = 0 \qquad \text{for all } i.$$

We can Taylor expand the above local truncation error $\tau_{i+1}$ at the point $t_{i+1}$ and find

$$\tau_{i+1} = y(t_{i+1})(1-(\alpha_1+\alpha_2+\cdots+\alpha_p))+\varepsilon y'(t_{i+1})(\alpha_1+2\alpha_2+\cdots+p\alpha_p-\beta_0-\beta_1-\cdots-\beta_p)+O(\varepsilon^2)$$

Thus the method is consistent if and only if

$$\begin{cases} \alpha_1 + \alpha_2 + \cdots + \alpha_p = 1 \\ \alpha_1 + 2\alpha_2 + \cdots + p\alpha_p = \beta_0 + \beta_1 + \cdots + \beta_p \end{cases}$$

The global error is defined to be the difference

$$e_i = y(t_i) - y_i$$

between the value of the true solution and the approximate solution at $t_i$. The method is called convergent if

$$\lim_{\varepsilon \to 0} \max_{0 \le i \le N} |e_i| = 0$$

for any starting values $y_0, y_1, \cdots, y_{p-1}$ such that

$$\lim_{\varepsilon \to 0} y_k = y_0 \qquad k = 0, 1, \cdots, p-1.$$

We state without proof the following remarkable result on the convergence property.

**Theorem 7.3.5** (Dahlquist's Equivalence Theorem)**.** *A linear multi-step method is convergent if and only if it is zero-stable and consistent.*

**Example 7.3.6.** Consider the 3-step Adams-Bashforth method

$$y_{i+1} = y_i + \varepsilon \left[ \frac{23}{12} F(y_i, t_i) - \frac{4}{3} F(y_{i-1}, t_{i-1}) + \frac{5}{12} F(y_{i-2}, t_{i-2}) \right].$$

The characteristic polynomial is

$$\rho(z) = z^3 - z^2 = z^2(z-1)$$

which has double root $z = 0$ and single root $z = 1$. So the method is zero-stable. In this case

$$\begin{aligned} \alpha_1 = 1 \qquad & \alpha_2 = 0 \qquad & \alpha_3 = 0 \\ \beta_1 = \frac{23}{12} \qquad & \beta_2 = -\frac{4}{3} \qquad & \beta_3 = \frac{5}{12} \end{aligned}$$

The consistency condition

$$\alpha_1 = \beta_1 + \beta_2 + \beta_3$$

is satisfies. Therefore this method is convergent by Dahlquist's Theorem.

## 7.4 Boundary Value Problem

We illustrate some basic ideas and features of numerical method for solving boundary value problems through the following example of Dirichlet boundary value problem

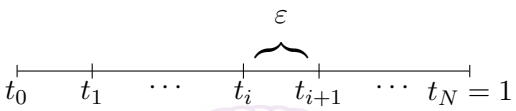
$$\begin{cases} y''(t) = f(t) & \text{on } I = [0,1] \\ y(0) = \alpha \qquad y(1) = \beta \end{cases}$$

This problem is itself simple since we can solve it explicitly by integrating $f(t)$ twice. Nevertheless we will look for a numerical solution.

We again subdivide the interval $[0,1]$ by the mesh-points

$$t_i = t_0 + i\varepsilon$$

with step-size $\varepsilon = \frac{1}{N}$




We look for a function valued on the mesh-points

$$y_0, y_1, \cdots, y_N$$

Such that $y_i$ will approximate the value $y(t_i)$ of the true solution at $t_i$. The situation is different from the initial value problem we discussed before: the endpoint values are fixed

$$y_0 = \alpha, \qquad y_N = \beta$$

and we need to interpolate the interior points from the equation.

### 7.4.1 Difference Equation

The first idea is that we can approximate the differential equation by a difference equation. Consider the following 2nd order centered approximation of a function $u(t)$

$$D^2u(t) := \frac{1}{\varepsilon^2}(u(t+\varepsilon) - 2u(t) + u(t-\varepsilon)).$$

We can use Taylor expansion at $t$

$$u(t+\varepsilon) = u(t) + \varepsilon u'(t) + \frac{\varepsilon^2}{2}u''(t) + \cdots$$

to find

$$D^2u(t) = u''(t) + \frac{\varepsilon^2}{12}u^{(4)}(t) + O(\varepsilon^4).$$

Thus $D^2u(t)$ can be used to approximate the function $u''(t)$. Apply this to our problem, the differential equation becomes a set of algebraic equations

$$\frac{1}{\varepsilon^2}(y_{i+1} - 2y_i + y_{i-1}) = f_i \qquad i = 1, 2, \cdots, N-1$$

where $f_i := f(t_i)$. Explicitly, this is (using the boundary condition $y_0 = \alpha, y_N = \beta$)

$$\begin{cases} \frac{1}{\varepsilon^2}(y_2 - 2y_1) = f_1 - \frac{\alpha}{\varepsilon^2} \\ \frac{1}{\varepsilon^2}(y_3 - 2y_2 + y_1) = f_2 \\ \quad \vdots \\ \frac{1}{\varepsilon^2}(y_{N-1} - 2y_{N-2} + y_{N-3}) = f_{N-2} \\ \frac{1}{\varepsilon^2}(-2y_{N-1} + y_{N-2}) = f_{N-1} - \frac{\beta}{\varepsilon^2} \end{cases}$$

If we denote

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{N-1} \end{pmatrix}, \qquad \mathbf{f} = \begin{pmatrix} f_1 - \frac{\alpha}{\varepsilon^2} \\ f_2 \\ \vdots \\ f_{N-2} \\ f_{N-1} - \frac{\beta}{\varepsilon^2} \end{pmatrix}$$

then the above equations can be written as

$$A\mathbf{y} = \mathbf{f}$$

where $A$ is the $(N-1) \times (N-1)$ tridiagonal matrix

$$A = \frac{1}{\varepsilon^2} \begin{pmatrix} -2 & 1 & & & 0 \\ 1 & -2 & 1 & & \\ & 1 & -2 & \ddots & \\ & & \ddots & \ddots & 1 \\ 0 & & & 1 & -2 \end{pmatrix}$$

This will allow us to solve the approximate values from the difference equation by

$$\mathbf{y} = A^{-1}\mathbf{f}.$$

### 7.4.2 Error Analysis

Let us consider the local truncation error expressed via the values of true solution by

$$\tau_i = \frac{1}{\varepsilon^2}(y(t_{i+1}) - 2y(t_i) + y(t_{i-1})) - f(t_i), \qquad 1 \le i \le N-1.$$

The global error is the difference between the true value and the approximate value produced by the algorithm. Precisely, it is

$$e_i = y(t_i) - y_i, \qquad 1 \le i \le N-1.$$

Observe that by construction

$$A\mathbf{e} = \boldsymbol{\tau}$$

where

$$\mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_{N-1} \end{pmatrix} \qquad \boldsymbol{\tau} = \begin{pmatrix} \tau_1 \\ \tau_2 \\ \vdots \\ \tau_{N-1} \end{pmatrix}$$

This gives a direct relation between the local and global truncation error

$$\mathbf{e} = A^{-1}\boldsymbol{\tau}.$$

Using Taylor expansion, the local truncation error has the behavior

$$\begin{aligned} \tau_i =& y''(t_i) + \frac{\varepsilon^2}{12} y^{(4)}(t_i) - f(t_i) + O(\varepsilon^4) \\ =& \frac{\varepsilon^2}{12} y^{(4)}(t_i) + O(\varepsilon^4) \\ =& O(\varepsilon^2). \end{aligned}$$

Let $\|\cdot\|$ denote the Euclidean norm. Then (using $N = \frac{1}{\varepsilon}$)

$$\|\boldsymbol{\tau}\| = \left( \sum_{i=1}^{N-1} \tau_i^2 \right)^{\frac{1}{2}} = \left( \frac{1}{\varepsilon} O(\varepsilon^4) \right)^{\frac{1}{2}} = O(\varepsilon^{\frac{3}{2}}).$$

It follows that

$$\|\mathbf{e}\| \leq \|A^{-1}\boldsymbol{\tau}\| \leq \|A^{-1}\| \|\boldsymbol{\tau}\| = O(\varepsilon^{\frac{3}{2}}) \|A^{-1}\|$$

Here $\|A^{-1}\|$ is the operator norm.

**Proposition 7.4.1.** *$\|A^{-1}\| \leq C$ is bounded in the limite $\varepsilon \to 0$. As a result, we find*

$$\|\mathbf{e}\| = O(\varepsilon^{\frac{3}{2}})$$

*This implies that the method is convergent.*

*Proof:* The difficulty lies in the fact that the size of $A^{-1}$ (which is $N - 1 = \frac{1}{\varepsilon} - 1$) is also increasing in the limit $\varepsilon \to 0$ ($N \to +\infty$).

Since $A$ is a symmetric matrix, $A^{-1}$ is also symmetric. Recall that for a symmetrix matrix $M$, we have

$$\|M\| = \max_{\lambda:\text{eigenvalue of } M} |\lambda|.$$

Let $\lambda_1, \cdots, \lambda_{N-1}$ be eigenvalues of $A$. Then $\lambda_1^{-1}, \lambda_2^{-1}, \cdots, \lambda_{N-1}^{-1}$ are eigenvalues of $A^{-1}$ and

$$\|A^{-1}\| = \max_{1 \leq i \leq N-1} |\lambda_i^{-1}| = \left( \min_{1 \leq i \leq N-1} |\lambda_i| \right)^{-1}.$$

Therefore we only need to show that the eigenvalues of $A$ are bounded away from zero as $\varepsilon \to 0$ ($N \to +\infty$). Let

$$B = \underbrace{\left.\begin{pmatrix} -2 & 1 & & & 0 \\ 1 & -2 & 1 & & \\ & 1 & -2 & \ddots & \\ & & \ddots & \ddots & 1 \\ 0 & & & 1 & -2 \end{pmatrix}\right\} N-1}_{N-1}$$

So $A = \frac{1}{\varepsilon^2} B$. Let us consider the eigenvalue equation

$$B \begin{pmatrix} u_1 \\ \vdots \\ u_{N-1} \end{pmatrix} = \mu \begin{pmatrix} u_1 \\ \vdots \\ u_{N-1} \end{pmatrix}.$$

In components, this reads

$$\begin{cases} u_0 + u_2 = (2+\mu)u_1 \\ u_1 + u_3 = (2+\mu)u_2 \\ \quad \vdots \\ u_{N-2} + u_N = (2+\mu)u_N \end{cases} \qquad \text{where} \quad u_0 := 0, u_N := 0$$

Observe the following relation

$$\sin((i-1)\theta) + \sin((i+1)\theta) = 2\cos\theta \sin i\theta.$$

We can find the solution of the above eigenvalue equation by

$$\begin{cases} u_i = \sin i\theta, & i = 0, \cdots, N \\ 2 + \mu = 2\cos\theta \end{cases}$$

For $u_0 = u_N = 0$ hold, we need

$$\begin{aligned} & \sin(N\theta) = 0 \\ \Rightarrow \qquad & \theta = \frac{\pi}{N}, \frac{2\pi}{N}, \cdots, \frac{(N-1)\pi}{N}. \end{aligned}$$

Thus we find all eigenvectors of $B$ by

$$\begin{pmatrix} \sin\frac{p\pi}{N} \\ \sin\frac{2p\pi}{N} \\ \vdots \\ \sin\frac{(N-1)p\pi}{N} \end{pmatrix} \qquad 1 \le p \le N-1$$

with eigenvalue $\mu_p = 2\cos\frac{p\pi}{N} - 2$. Therefore the eigenvalues of $A$ are given by

$$\begin{aligned} \lambda_p =& \frac{2}{\varepsilon^2}(\cos\frac{p\pi}{N} - 1) \\ =& \frac{2}{\varepsilon^2}(\cos p\pi\varepsilon - 1), \qquad p = 1, \cdots, N-1. \end{aligned}$$

The eigenvalue with smallest magnitude is

$$\begin{aligned}\lambda_1 =& \frac{2}{\varepsilon^2}(\cos \pi\varepsilon - 1)\\ =& \frac{2}{\varepsilon^2}(-\frac{1}{2}\pi^2\varepsilon^2 + \frac{1}{24}\pi^4\varepsilon^4 + O(\varepsilon^6))\\ =& -\pi^2 + O(\varepsilon^2).\end{aligned}$$

This is clearly bounded away from zero in the limit $\varepsilon \to 0$. This proves the proposition. □

# Bibliography


[1] Arnold, Vladimir I. *Ordinary differential equations*. Springer Science & Business Media, 1992.

[2] Boyce, William E., and Richard C. DiPrima. *Elementary differential equations and boundary value problems*. Wiley, 2020.

[3] Coddington, Earl A., and Norman Levinson. *Theory of ordinary differential equations*. McGraw-Hill, 1955.

[4] Butcher, John Charles. *Numerical methods for ordinary differential equations*. John Wiley & Sons, 2016.

[5] Kythe, Prem K. *Green's functions and linear differential equations: theory, applications, and computation*. CRC Press, 2011.

[6] Li, Si. *Classical Mechanics and Geometry*. International Press of Boston, 2023.

[7] Stakgold, Ivar, and Michael J. Holst. *Green's functions and boundary value problems*. John Wiley & Sons, 2011.

[8] Süli, Endre, and David F. Mayers. *An introduction to numerical analysis*. Cambridge university press, 2003.

[9] Teschl, Gerald. *Ordinary differential equations and dynamical systems*. Vol. 140. American Mathematical Soc., 2012.

[10] Walter, Wolfgang. *Ordinary Differential Equations*. Volume 182 of Graduate Texts in Mathematics, 1998.

[11] Wasow, Wolfgang. *Asymptotic expansions for ordinary differential equations*. Dover Publications, Inc. 1987.