18.03 Differential equations

Fall 2018 lecture notes

Instructor: Jörn Dunkel

This PDF is a combination of MITx material and my own notes. Credit for course design and content should go to the authors of the MITx 18.03 course pages. Responsibility for typos and errors lies with me.

Contents

1 Basic math review

This class deals with differential equations (DEs). These are equation that contain functions and their derivatives. To solve DEs, we will need a few basic concepts such as vectors and complex numbers, most of which should be familiar from 18.01 and 18.02.

1.1 Vectors and matrices

For simplicity, we will focus on the case of the 2D plane $\mathbb{R}^2$. Let's assume we have fixed a coordinate frame spanned by two orthonormal unit vectors $\{\boldsymbol{e}_1, \boldsymbol{e}_2\}$, which we may represent as column vectors

$$\boldsymbol{e}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} , \qquad \boldsymbol{e}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix} .$$

Orthonormal means that $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ are *orthogonal* with respect to the standard Euklidean scalar product,

$$\boldsymbol{e}_1 \cdot \boldsymbol{e}_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = 1 \cdot 0 + 0 \cdot 1 = 0,$$

and *normalized*

$$\boldsymbol{e}_1 \cdot \boldsymbol{e}_1 = \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = 1 , \qquad \boldsymbol{e}_2 \cdot \boldsymbol{e}_2 = \begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = 1.$$

The two vectors $\{\boldsymbol{e}_1, \boldsymbol{e}_2\}$ form a *basis* of $\mathbb{R}^2$ because any other two dimensional vector $\boldsymbol{r}$ in the plane $\mathbb{R}^2$ can be represented as a *superposition*

$$\boldsymbol{r} = x\boldsymbol{e}_1 + y\boldsymbol{e}_2 = \begin{pmatrix} x \\ y \end{pmatrix} , \tag{1}$$

where (x, y) are the *Cartesian* coordinates of $\boldsymbol{r}$. The *length* $|\boldsymbol{r}|$ of $\boldsymbol{r}$ is given by the square root of the Euklidean scalar product of $\boldsymbol{r}$ with itself

$$|\boldsymbol{r}| = (\boldsymbol{r} \cdot \boldsymbol{r})^{1/2} = \left[\begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \right]^{1/2} = \left(x^2 + y^2\right)^{1/2} .$$

The vectors constructed from $\{\boldsymbol{e}_1, \boldsymbol{e}_2\}$ form a *linear vector space*, which means that the sum $\boldsymbol{s}$ of any pair of 2D vectors $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$ is again a 2D vector

$$\boldsymbol{s} = \boldsymbol{r}_1 + \boldsymbol{r}_2 = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} + \begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = \begin{pmatrix} x_1 + x_2 \\ y_1 + y_2 \end{pmatrix} = (x_1 + x_2)\boldsymbol{e}_1 + (y_1 + y_2)\boldsymbol{e}_2$$

and that multiplying a 2D vector $\boldsymbol{r}$ by a scalar λ gives a new 2D vector

$$\boldsymbol{w} = \lambda\boldsymbol{r} = \lambda(x\boldsymbol{e}_1 + y\boldsymbol{e}_2) = \lambda x\boldsymbol{e}_1 + \lambda y\boldsymbol{e}_2 = \begin{pmatrix} \lambda x \\ \lambda y \end{pmatrix} .$$

That is, the vector space $\mathbb{R}^2$ is closed under vector addition and scalar multiplication.

Furthermore, recall that we can operate on a 2D vector with a 2 × 2-matrix

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

by applying the usual rules of matrix multiplication,

$$A\boldsymbol{r} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} := \begin{pmatrix} A_{11}x + A_{12}y \\ A_{21}x + A_{22}y \end{pmatrix},$$

where the symbol := means that the rhs defines the lhs. The matrix operation is *linear*, since[1]

$$A(\lambda_1 \boldsymbol{r}_1 + \lambda_2 \boldsymbol{r}_2) = \lambda_1 A\boldsymbol{r}_1 + \lambda_2 A\boldsymbol{r}_2. \tag{2}$$

That is, it does not matter in which order the scalar multiplication, vector addition and matrix operation are performed. The *trace* $\mathrm{tr}A$ and the *determinant* $\det A$ of a 2×2-matrix A are defined by

$$\mathrm{tr}A = A_{11} + A_{22}\ , \qquad \det A = A_{11}A_{22} - A_{12}A_{21}$$

Stretching a vector $\boldsymbol{r}$ by a factor λ is achieved through multiplication with the matrix

$$\Lambda = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix} = \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \lambda I$$

where I denotes the identity matrix that leaves all vectors unchanged. Another important example is the rotation matrix

$$R(\phi) = \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix}$$

which rotates a vector counter-clockwise by an angle ϕ. Trace and determinant of $R(\phi)$ are given by

$$\mathrm{tr}R = 2\cos\phi\ , \qquad \det R = (\cos\phi)^2 + (\sin\phi)^2 = 1.$$

For example, consider the vector $\hat{\boldsymbol{r}}$ obtained by first rotating the unit vector $\boldsymbol{e}_1$ by ϕ and then stretching by λ, i.e.,

$$\hat{\boldsymbol{r}} = \Lambda R(\phi)\boldsymbol{e}_1 = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix} \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \lambda \begin{pmatrix} \cos\phi \\ \sin\phi \end{pmatrix} = \begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix}.$$

The last equality relates the Cartesian coordinates $(\hat{x}, \hat{y})$ of $\hat{\boldsymbol{r}}$ to its polar coordinates (λ, ϕ).

If the above concepts appear unfamiliar, you may practice by trying to generalize them to 3D.

[1] It's a good exercise to check this explicitly.

1.2 Complex numbers

Some polynomial equations, like

$$x^2 = 1,$$

can be solved within the set of real numbers $\mathbb{R}$, which can be identified with the 1D vector space $\mathbb{R}^1$ spanned by $\boldsymbol{e}_1$. For others, like

$$z^2 = -1 \tag{3}$$

this is not possible and one needs to extend the solution space. A minimal extension of the real numbers, which allows to find solutions of (3) are the complex numbers $\mathbb{C}$, which can be thought of as 2D generalizations of the real numbers. A complex number[2] $z \in \mathbb{C}$ can be written as

$$z = x + iy \in \mathbb{C} \ , \qquad i^2 = -1$$

with real part $\Re z = x \in \mathbb{R}$ and imaginary part $\Im z = y \in \mathbb{R}$ representing the Cartesian coordinates of z. Real numbers are complex numbers with $y = 0$. The conjugate of a complex number $z = x + iy$ is given by

$$\bar{z} = x - iy$$

and corresponds to a reflection at the real axis or, equivalently, at the line $\Im(z) = 0$.

Addition of complex numbers is linear

$$z = z_1 + z_2 = (x_1 + iy_1) + (x_2 + iy_2) = (x_1 + x_2) + i(y_1 + y_2) = x + iy$$

equivalent to the addition of the two 2D vectors (x_1, y_1) and (x_2, y_2). In contrast, complex multiplication mixes real and imaginary parts

$$z = z_1 z_2 = (x_1 + iy_1)(x_2 + iy_2) = (x_1 x_2 - y_1 y_2) + i(x_1 y_2 + y_1 x_2) = x + iy.$$

Euler's formula

$$e^{i\phi} = \cos\phi + i \sin\phi \ , \qquad \phi \in \mathbb{R} \tag{4}$$

relates the exponential function

$$\exp(i\phi) := 1 + \frac{i\phi}{1!} + \frac{(i\phi)^2}{2!} + \frac{(i\phi)^3}{3!} + \ldots = \sum_{k=0}^{\infty} \frac{(i\phi)^k}{k!}$$

to the trigonometric functions

$$\begin{aligned} \sin(\phi) &:= \frac{\phi}{1!} - \frac{\phi^3}{3!} + \frac{\phi^5}{5!} - \frac{\phi^7}{7!} + \ldots \\ \cos(\phi) &:= 1 - \frac{\phi^2}{2!} + \frac{\phi^4}{4!} - \frac{\phi^6}{6!} + \ldots \end{aligned}$$

[2] The symbol $\in$ means 'an element of' or 'in'.

It is often advantageous to use the polar representation of a complex number

$$z = re^{i\phi} = r(\cos\phi + i\sin\phi) = x + iy \tag{5a}$$

with

$$r = |z| = \sqrt{z\bar{z}} \geq 0 \ , \qquad \phi = \arctan 2(y, x) \in [0, 2\pi). \tag{5b}$$

Useful examples are

$$1 = 1e^{i0}, \qquad i = 1e^{i\pi/2}, \qquad -1 = 1e^{i\pi}, \qquad -i = 1e^{i3\pi/2} = e^{-i\pi/2}.$$

In the last case, we have used that, for any integer n,

$$e^{i\phi} = e^{i(\phi+2\pi n)},$$

which follows directly from the periodicity of trigonometric functions in Eq. (5a).

From the properties of the exp-function, one sees directly that the multiplication of complex numbers

$$z = z_1 z_2 = r_1 e^{i\phi_1} r_2 e^{i\phi_2} = r_1 r_2 e^{i(\phi_1+\phi_2)}$$

corresponds to a combined rotation and dilation. As a general rule, the Cartesian representation of a complex number is advantageous for performing additions, while the polar (exponential) representation is more convenient for multiplications.

To illustrate the usefulness of the polar representation, let's consider the polynomial equation

$$1 = z^3,$$

which we would like to solve for z. Writing $1 = e^{i0} = e^{i2\pi n}$ and $z = |z|e^{i\phi}$ with $|z| = 1$, this equation becomes

$$e^{i2\pi n} = (e^{i\phi})^3 = e^{i3\phi}$$

so we find

$$\phi = \frac{2\pi}{3} n \ , \qquad n = 0, \pm 1, \pm 2, \ldots,$$

implying three distinct solutions[3]

$$z_0 = 1 \ , \qquad z_\pm = e^{\pm i2\pi/3}.$$

This is an illustration of the fundamental theorem of algebra which asserts that a polynomial of order k with real coefficients has exactly k complex zeros, when counted with their multiplicity. The non-real roots of polynomials with real coefficients come in conjugate pairs. For example, the fourth-order polynomial

$$p(z) = (z-1)^2(z^2+1)$$

has the four zeros $z = 1$ (multiplicity 2) and $z = \pm i$. One of the main goals when dealing with differential equations is to transform them into polynomial equations, which means that we will see at lot of the above throughout this course.

[3] The other $n = \pm 2, \pm 3, \ldots$ just repeat one of the three.

2 Introduction to modeling and differential equations

In this course, we will mostly deal with ordinary differential equations (ODEs) and, in the latter parts, also with simple partial differential equations (PDEs).

2.1 ODEs *vs.* PDEs

ODEs are equations that contain derivatives of scalar functions [such as planar curves $y(x)$ or particle traces $x(t)$ in one space dimension], or vector-valued functions [such as particle traces $\boldsymbol{X}(t) = (X_1(t), X_2(t))$ in the plane] that only depend on a single position variable x or time-variable t. Depending on context, we will denote derivatives of such functions by

$$\frac{dy}{dx} = y'(x) \;, \qquad \frac{d^2y}{dx^2} = y''(x) \;, \qquad \frac{d^ny}{dx^n} = y^{(n)}(x) \;, \qquad \frac{dx}{dt} = \dot{x}(t) \;, \qquad \frac{d^2x}{dt^2} = \ddot{x}(t)$$

often omitting the arguments, simply writing y' or $\ddot{x}$.

The perhaps simplest non-trivial ODE is

$$y' = \alpha y \tag{6a}$$

which for $\alpha \neq 0$ is solved by

$$y(x) = y_0 e^{\alpha x}. \tag{6b}$$

Another important ODE is the 1D harmonic oscillator equation

$$\ddot{x} = -\omega^2 x, \tag{7a}$$

where ω is the oscillator frequency, which is solved by

$$x(t) = A \sin(\omega t) + B \cos(\omega t). \tag{7b}$$

Note the structural similarity with vectors, when we interpret $\{\sin(\omega t), \cos(\omega t)\}$ as basis functions and (A, B) as coordinates. As we will see later this analogy is not merely formal but in fact crucial for our ability to solve arbitrary linear ODEs.

PDEs are equations that contain derivatives of scalar or vector-valued functions that depend on more than one variable. Their properties are discussed in detail in 18.02. An example of a scalar function that depends on two variables, is the time- and position-dependent temperature along a thin rod, $T(t, x)$, whose partial derivatives will be denoted by

$$\frac{\partial T}{\partial x} \;, \qquad \frac{\partial^2 T}{\partial x^2} \;, \qquad \frac{\partial T}{\partial t}$$

An important PDE, discussed (much) later in class, is the diffusion equation

$$\frac{\partial T}{\partial t} = D \frac{\partial^2 T}{\partial x^2},$$

which relates the local time-change of $T(t, x)$ to the spatial temperature variations (the material parameter D is the temperature diffusion constant).

2.2 Building ODE models

From now on, until stated otherwise, we consider ODEs for scalar functions, such as $y(x)$ or $x(t)$. There exist many ways of constructing ODE models for natural and artificial systems. In this course, we will only discuss a few basic ones that will help you to understand the construction principles underlying more complex mathematical models that you will encounter in your physics, engineering or biology classes.

2.2.1 Population growth

A key problem in biology is to understand how cell populations grow and compete under different conditions. The most basic ODE model, which describes for example the growth of bacterial colonies quite well, can be constructed as follows:

Denote by

$$\Delta N(t) = N(t + \Delta t) - N(t)$$

the net growth of the population during the small time-interval Δt. To first approximation, it is then plausible to assume that the ΔN is proportional to Δt and the number of cells $N(t)$ present at time time t, i.e.

$$\Delta N(t) = \alpha N(t) \Delta t$$

where $\alpha > 0$ is a constant of proportionality which is expected to depend on the external growth conditions (nutrient concentrations, temperature, etc.). Dividing by Δt and letting $\Delta t \to 0$, to replace discrete difference quotients by continuous derivatives, we find the ODE growth model

$$\dot{N}(t) = \alpha N(t). \tag{8}$$

We mentioned earlier that, in order to solve ODEs, one typically tries to transform them into simpler algebraic equations by guessing a suitable trial solution, also called *ansatz*. To illustrate this general recipe here, let's rewrite Eq. (8) in standard form

$$\dot{N}(t) - \alpha N(t) = 0 \tag{9}$$

and try the exponential ansatz

$$N_h(t) = Ce^{\lambda t} \tag{10}$$

which contains two free parameters C and λ. Noting that $\dot{N}_h(t) = \lambda C e^{\lambda t}$, insertion of Eq. (10) into (9) and subsequent division by $Ce^{\lambda t}$ gives the algebraic equation

$$p(\lambda) = \lambda - \alpha = 0, \tag{11}$$

which determines one of the free parameter, $\lambda = \alpha$. The function $p(\lambda)$ is the *characteristic polynomial* of Eq. (9). The solutions of Eq. (8) therefore describe exponential growth

$$N_h(t) = Ce^{\alpha t}. \tag{12}$$

In particular, we have infinitely many of them, parameterized by C, unless we specify an additional condition that fixes C. This is a generic feature of DE models: *In order to specify a unique solution, one needs to add extra information in the form of initial or boundary conditions to the DE.* In our population growth example, it is natural to specify the initial cell population $N(0)$ at time $t = 0$, $N(0) = N_0$, which fixes C by

$$N_0 = N(0) = Ce^{\alpha 0} = C.$$

The minimal model (13) can be made more realistic by also accounting for cell deaths. This can be achieved by adding a constant rate term on the rhs. of Eq. (13),

$$\dot{N}(t) = \alpha N(t) - q \ , \qquad q > 0. \tag{13}$$

The death *rate* q has dimensions[4] of 1/Time. Equation (13) is linear in the sought-after function $N(t)$. As we shall see below, this means that the general solution of Eq. (13) can be constructed by taking the general (homogeneous) solution for $q = 0$ and adding just one (particular) solution of Eq. (13). This is an example of what's called the *superposition principle for linear ODEs.* A particular solution to Eq. (13) is readily found in the form of the constant function

$$N_p(t) = \frac{q}{\alpha}.$$

which has $\dot{N}_p = 0$. Hence, the general solution of Eq. (13) is given by

$$N(t) = N_h(t) + N_p = Ce^{\alpha t} + \frac{q}{\alpha}.$$

Moreover, if we again specify an initial condition $N(0) = N_0$, then

$$N_0 = N(0) = N_h(0) + N_p = C + \frac{q}{\alpha} \qquad \Rightarrow \qquad C = N_0 - \frac{q}{\alpha}, \tag{14}$$

and the solution takes the form

$$N(t) = \left(N_0 - \frac{q}{\alpha}\right) e^{\alpha t} + \frac{q}{\alpha}. \tag{15}$$

Try to check by direct insertion that (15) does indeed solve (13).

2.2.2 Newton's force law

The above example illustrates how to obtain an ODE model by thinking about the growth and decay rates of certain biological, physical or chemical quantities. Another way of constructing ODE models starts from Newton's force law connecting force and acceleration, which in 1D can be written as

$$F = ma \tag{16}$$

where F the force acting on a particle of mass m at position $x(t)$, and

$$a(t) = \dot{v}(t) = \ddot{x}(t) \tag{17}$$

[4] More generally, the production or depletion rate of a quantity x has the units of $\dot{x}$.

is the particle's acceleration (v is the velocity).

For example, if the particle is attached to a linear spring with rest-position at $x(t) = 0$, then according to Hooke's law

$$F_s = -kx \tag{18a}$$

where $k > 0$ is the spring constant. If the particle is moving in a medium that exerts an approximately linear velocity-dependent friction, then there will be an additional force contribution

$$F_f = -\gamma v \tag{18b}$$

where γ is the Stokes friction coefficient. We may also consider the case where we apply an additional (e.g. electric) oscillatory force field

$$F_e = A \sin(\Omega t) \tag{18c}$$

of amplitude A and frequency Ω.

Combining all these contributions, we obtain the ODE

$$m\ddot{x} = -\gamma \dot{x} - kx + A \sin(\Omega t). \tag{19}$$

Dividing by m and moving x-dependent terms to the lhs., we can rewrite this in the standard form

$$\ddot{x} + b\dot{x} + \omega^2 x = \epsilon \sin(\Omega t). \tag{20}$$

where $b = \gamma/m$ is the mass-rescaled friction coefficent, $\omega = \sqrt{k/m}$ is the intrinsic spring frequency, and $\epsilon = A/m$.

These ideas generalize in a straightforward manner to higher-dimensional forces. In fact, you have seen already one example of a three-dimensional ODE system on the introductory slides, when we glanced over gravitational systems.

2.3 Classification of ODEs: Linear *vs.* nonlinear equations

It's time to introduce some important nomenclature that will allow us to classify ODEs systematically. The *order* n of a DE is the highest derivative $y^{(n)}(x)$ or $x^{(n)}(t)$ appearing in it. For example, the equation

$$e^t \ddot{x} + 5\dot{x} + t^9 x = 0 \tag{21}$$

is a second order ODE, as is the damped harmonic oscillator equation (20).

From a theoretical and practical perspective, the perhaps most important classification of ODEs is their separation into *linear* and *nonlinear* equations. While linear equations are (relatively) straightforward to solve, nonlinear equations are notoriously hard and typically require computer simulations and/or qualitative analysis. Unfortunately, the most relevant real-world problems are described by nonlinear ODEs. Fortunately, however, they can sometimes be approximated by linear ODEs.

Intuitively, a linear ODE does *not* contain any products of the sought-after function and its derivatives (we will also give a more formal definition below). The example in Eq. (21) is linear in $x(t)$, as is the harmonic oscillator equation (7a). By contrast,

$$(y')^2 - y = 0$$

or

$$y' - y^3 = 0$$

are examples of nonlinear differential equations in $y(x)$.

3 Solving first-order ODEs

First-order ODEs for scalar functions $y(x)$ or $x(t)$ are somewhat special in that there exist a number of technical tricks that work even for nonlinear equations.

3.1 Separation of variables

In general, nonlinear DEs are very difficult or impossible to solve analytically. However, certain types of *first-order nonlinear* ODEs can be solved by a technique called separation of variables. Let's assume we can rewrite a given ODE in the form[5]

$$f(x)\dot{x} = q(t), \tag{22}$$

where $f(x)$ can be linear or nonlinear. We can solve Eq. (22) systematically as follows:

1. Rewrite the equation in differential form as

$$f(x)\frac{dx}{dt} = q(t). \tag{23a}$$

2. Multiply by dt to obtain

$$f(x)dx = q(t)dt. \tag{23b}$$

This step separates x and t, hence the name of the procedure.

3. Integrate both sides to get

$$F(x) = Q(t) + C. \tag{23c}$$

These are implicit equations for the solutions $x(t)$, in terms of a parameter C.

4. Solve for x if possible and desired. Optional: Check by insertion if $x(t)$ solves the original ODE.

[5] When divisions are necessary, ensure that you don't divide by 0.

Example. Solve

$$\dot{x} - 2tx = 0 \tag{24}$$

1. Rewrite in differential form as

$$\frac{1}{x}\frac{dx}{dt} = 2t \tag{25a}$$

2. Multiply by dt to obtain

$$\frac{dx}{x} = 2tdt. \tag{25b}$$

3. Integrate both sides to get

$$\ln|x| = t^2 + C. \tag{25c}$$

4. Solving for x gives

$$x(t) = \pm e^C e^{t^2} = ce^{t^2} \tag{25d}$$

   for some positive or negative constant $c = \pm e^C$.

You can check by insertion that $x(t) = ce^{t^2}$ solves Eq. (24).

3.2 Variation of parameters for inhomogeneous linear first-order ODEs

The most general form of a first-order linear ODE is

$$p_1(t)\dot{x} + p_0(t)x = q(t). \tag{26}$$

Assuming $p_1(t) \neq 0$, we can rewrite this as

$$\dot{x} + P(t)x = Q(t) \tag{27}$$

where

$$P(t) = \frac{p_0(t)}{p_1(t)}\,, \qquad Q(t) = \frac{q(t)}{p_1(t)}.$$

We can solve Eq. (27) systematically through a method called *variation of parameters*, which works as follows:

1. Use separation of variables to find a solution $x_h(t)$ to the homogenous problem with $Q(t) \equiv 0$; that is, a function $x_h(t)$ satisfying

$$\dot{x}_h + P(t)x_h = 0. \tag{28}$$

2. Insert the ansatz $x(t) = u(t)\,x_h(t)$ into (27) and determine the unknown function $u(t)$ by using that by product rule

$$\begin{aligned} \dot{x} + P(t)x &= \dot{u}x_h + u\dot{x}_h + P(t)ux_h \\ &= \dot{u}x_h + u[\dot{x}_h + P(t)x_h] \\ &\overset{(28)}{=} \dot{u}x_h \end{aligned} \tag{29a}$$

   Thus

$$\dot{u}(t)x_h(t) = Q(t). \tag{29b}$$

   For $x_h(t) \neq 0$, we can find $u(t)$ by integrating both sides of

$$\dot{u}(t) = \frac{Q(t)}{x_h(t)}. \tag{29c}$$

Example. Let's solve

$$t\dot{x} + 2x = t^5 \ , \qquad t > 0.$$

First, we rewrite this equation as

$$\dot{x} + \frac{2}{t}x = t^4 = Q(t)$$

Next we solve homogenous problem with $Q(t) = 0$ using variation of variables

$$\frac{dx_h}{x_h} = -\frac{2dt}{t}$$

yielding

$$\ln|x_h| = -2\ln|t| + \tilde{C}$$

and therefore

$$x_h(t) = Ct^{-2}$$

for some constant C. We may pick any solution x_h, so let's set $C = 1$. Then Eq. (29c) gives

$$u(t) = c + \int t^6 dt = c + \frac{t^7}{7}$$

Hence, the general solution $x(t) = u(t)x_h(t)$ is obtained as

$$x(t) = \left(c + \frac{t^7}{7}\right)t^{-2} = ct^{-2} + \frac{t^5}{7}.$$

3.3 Integrating factor

Another approach to solving the first order, linear, inhomogeneous ODE

$$\dot{x} + p(t)x = q(t) \tag{30}$$

is to use an *integrating factor*. This method works as follows.

1. Find an antiderivative $P(t)$ of $p(t)$.
2. Multiply both sides of the ODE by the integrating factor $e^{P(t)}$ in order to make the left side of Eq. (30) the derivative of something:

$$\begin{aligned} & e^{P(t)}\dot{x} + e^{P(t)}p(t)x &=& \; e^{P(t)}q(t) \\ \Leftrightarrow \quad & \frac{d}{dt}\left[e^{P(t)}x\right] &=& \; e^{P(t)}q(t) \end{aligned}$$

3. Integrate both sides

$$\begin{aligned} & e^{P(t)}x = \int dt\, e^{P(t)}q(t) \\ \Leftrightarrow \quad & x(t) = e^{-P(t)} \int dt\, e^{P(t)}q(t) \end{aligned}$$

Here $\int dt\, e^{P(t)}q(t)$ represents all possible antiderivatives of $e^{P(t)}q(t)$, so there are infinitely many solutions. If you fix one antiderivative, say $R(t)$, then the others are $R(t) + c$ for a constant c, so the general solution is

$$x(t) = R(t)e^{-P(t)} + ce^{-P(t)}. \tag{31}$$

4 Complex exponential function

When solving linear ODEs, we will need complex exponential functions of the form

$$f(t) = e^{zt} \tag{32}$$

where t is any real number and z any complex number. We can define this complex exponential function as unique solution of the ODE

$$\frac{d}{dt}f(t) = zf(t)\ , \qquad f(0) = 1, \tag{33}$$

where the second equation specifies the initial condition. This definition in terms of the ODE (33) is consistent with other definitions you may have seen, such as for example

$$e^{zt} = \sum_{n=0}^{\infty} \frac{(zt)^n}{n!} = 1 + (zt) + \frac{(zt)^2}{2!} + \frac{(zt)^3}{3!} + \frac{(zt)^4}{4!} + \dots \tag{34}$$

where

$$n! = n \cdot (n-1) \cdot (n-2) \cdots 2 \cdot 1.$$

You can verify this by inserting the series expansion (34) into (33); by differentiating each term of series you will find that it indeed solves the ODE. The complex exponential function inherits useful properties of the real exponential function. For example, we have

$$\begin{aligned} e^{z+w} &= e^z e^w && (35a) \\ (e^z)^n &= e^{zn} && (35b) \end{aligned}$$

for all complex numbers z, w and integer numbers n. The special cases $n = 0$ and $n = 1$ yield

$$(e^z)^0 = e^{0+i0} = \cos 0 + i \sin 0 = 1 \tag{36a}$$

and

$$\frac{1}{e^z} = (e^z)^{-1} = e^{-z} \tag{36b}$$

Moreover, using the standard Euler formula (4), we find for $z = x + iy$ the generalized Euler formula

$$e^{x+iy} = e^x(\cos y + i \sin y). \tag{37}$$

4.1 Computing integrals

Complex exponentials are useful for computing integrals of the form

$$I = \int dt\ e^{\alpha t} \cos t \quad = \quad \Re \int dt\ e^{\alpha t}(\cos t + i \sin t)$$

where α is a real parameter and $\Re$ denotes the real part. We will encounter integrals of this type when dealing with damped oscillations. Using Euler's formula

$$I = \Re \int dt\ e^{\alpha t} e^{it} = \Re \int dt\ e^{(\alpha+i)t} = \Re \left[\frac{e^{(\alpha+i)t}}{\alpha + i} \right] = e^{\alpha t} \Re \left[\frac{e^{it}}{\alpha + i} \right]$$

To compute the real part note that

$$\frac{e^{it}}{\alpha + i} = \frac{\alpha - i}{\alpha - i} \frac{1}{\alpha + i} (\cos t + i \sin t) = \frac{\alpha - i}{\alpha^2 + 1} (\cos t + i \sin t)$$

Hence

$$I = e^{\alpha t} \Re \left[\frac{e^{it}}{\alpha + i} \right] = \frac{e^{\alpha t}}{\alpha^2 + 1} (\alpha \cos t + \sin t) \tag{38}$$

4.2 Roots of complex polynomials

Another important application concerns the finding of roots of a polynomial. The **Fundamental Theorem of Algebra** states:

Every degree n complex polynomial $p(z)$ has exactly n complex roots when counted with multiplicity.

A consequence is that we can rewrite a polynomial

$$p(z) = a_n z^n + a_{n-1} z^{n-1} + \ldots + a_1 z + a_0 \tag{39a}$$

as a product

$$p(z) = a_n (z - z_n)(z - z_{n-1}) \cdots (z - z_1) \tag{39b}$$

where $\{z_1, z_2, \ldots, z_n\}$ are the roots.

We will make use of this fact when computing zeros of characteristic polynomial. As an example, consider

$$p(z) = z^5 - 32 \tag{40}$$

which is the characteristic polynomial of the ODE

$$x^{(5)}(t) - 32x(t) = 0, \tag{41}$$

as one can see by inserting $x = e^{zt}$. To find the five roots, use polar coordinates

$$z = re^{i\phi}$$

Then

$$p(z) = 0 \qquad \Leftrightarrow \qquad r^5 e^{i5\phi} = 32.$$

From this we find that the five distinct complex zeros are given by

$$r = 32^{1/5} = 2 \ , \qquad\qquad 5\phi_k = 2\pi k \qquad \text{where} \quad k = 0, 1, 2, 3, 4. \tag{42}$$

5 Second-order ODEs with constant coefficients and their characteristic polynomials

We had seen in Sec. 2.2.2 that Newton's force law for a damped oscillator can be written in the form of a linear second-order ODE

$$m\ddot{x} = -\gamma\dot{x} - kx + F_e(t), \tag{43}$$

where γ is the friction coefficient, k the spring constant and $F_e(t)$ a position-independent external force. Dividing by m and moving x-dependent terms to the lhs., we can rewrite this in the standard form[6]

$$\ddot{x} + b\dot{x} + \omega^2 x = q(t), \tag{44}$$

[6] See Eq. (20).

where $b = \gamma/m$ is the mass-rescaled friction coefficient and $\omega = \sqrt{k/m}$ is the intrinsic spring frequency.

We would like to solve this equation for the homogeneous case $q(t) \equiv 0$, when there are no external forces.

$$\ddot{x} + b\dot{x} + \omega^2 x = 0, \tag{45}$$

The inhomogeneous case $q(t) \neq 0$ will be discussed in detail later.

5.1 General superposition principle for homogeneous linear equations

Equation (45) is a special case of the general linear nth-order homogeneous ODE

$$p_n(t)x^{(n)} + p_{n-1}(t)x^{(n-1)} + \ldots + p_1(t)\dot{x} + p_0(t)x = 0 \tag{46}$$

For an ODE of the form (46), the following statements are true:

- The zero function $x(t) \equiv 0$ is a solution.
- If $x(t)$ is a solution of Eq. (46), then any scalar multiple $\alpha x(t)$ is also a solution.
- If $x_1(t)$ and $x_2(t)$ both solve Eq. (46), then their sum $x_1(t) + x_2(t)$ also a solution.

Thus, in short, all linear combinations of homogeneous solutions are homogeneous solutions. We demonstrate this important and useful fact for the dashpot example.

5.2 Homogeneous linear 2nd-order ODE

5.2.1 Dashpot: Frictionless case

We first consider the frictionless case $b = 0$, when Eq. (45) reduces to

$$\ddot{x} + \omega^2 x = 0. \tag{47}$$

We plug in the exponential trial function

$$x_\lambda(t) = e^{\lambda t}. \tag{48}$$

This is what you should *always* do when facing a linear homogenous nth order ODE with constant coefficients. Using

$$\ddot{x}_\lambda(t) = \lambda^2 e^{\lambda t}$$

and dividing by $e^{\lambda t}$, we find that Eq. (47) reduces to the algebraic equation

$$p(\lambda) := \lambda^2 + \omega^2 = 0. \tag{49}$$

The quadratic function $p(\lambda)$ is the characteristic polynomial of Eq. (47). Generally, for any nth order linear ODE with constant coefficients, the characteristic polynomial will be of degree n [we had already seen a first order example in Eq. (11)].

Solving (49) for λ, we find the two complex roots

$$\lambda_\pm = \pm i\omega, \tag{50a}$$

yielding the two solutions

$$x_+(t) = e^{i\omega t}\ , \qquad x_-(t) = e^{-i\omega t}. \tag{50b}$$

As we had seen above in Sec. 5.1, these may be combined by linear superposition to find the general solution of the undamped oscillator Eq. (47):

$$x(t) = C_+ e^{i\omega t} + C_- e^{-i\omega t}, \tag{51}$$

where $C_\pm$ are complex parameters. The physical requirement that the solution $x(t)$ must be real-valued, $x(t) = \bar{x}(t)$, implies that

$$C_+ e^{i\omega t} + C_- e^{-i\omega t} = \overline{(C_+ e^{i\omega t} + C_- e^{-i\omega t})} = \bar{C}_+ e^{-i\omega t} + \bar{C}_- e^{i\omega t}.$$

Comparing the coefficients in front of $e^{\pm i\omega t}$, we see that

$$C_+ = \bar{C}_-.$$

This makes sense since a complex number corresponds to two real numbers, and we only have to fix two real parameters specify a unique solution for a linear second-order ODE. We can write the real solution of Eq. (47) in the complex form

$$x(t) = \bar{C}_- e^{i\omega t} + C_- e^{-i\omega t}. \tag{52}$$

It is convenient to reexpress this solution in terms of sin and cos. To this end, we write

$$\bar{C}_- = (c_1 - ic_2)/2$$

with real numbers c_1 and c_2, and use Euler's formula to expand the exponentials

$$x(t) = \frac{1}{2}(c_1 - ic_2)(\cos\omega t + i \sin\omega t) + \frac{1}{2}(c_1 + ic_2)(\cos\omega t - i \sin\omega t)$$

After collecting all the terms, we then recover the perhaps more familiar form of the oscillator solution

$$x(t) = c_1 \cos\omega t + c_2 \sin\omega t. \tag{53}$$

Note that this solutions looks structurally similar to the vectors in Eq. (1), if we identify

$$\boldsymbol{e}_1 = \cos\omega t\ , \qquad \boldsymbol{e}_2 = \sin\omega t$$

as basis vectors and interpret c_1 and c_2 as the coordinates with respect to this basis. This is no coincidence - as we will see soon.

5.2.2 Dashpot: Strongly damped case

We next consider the case with damping $b > 0$. To make computations a bit easier, we assume specific values $b = 3$ and $\omega^2 = 2$. In this case, Eq. (45) reduces to

$$\ddot{x} + 3\dot{x} + 2x = 0. \tag{54}$$

Inserting the trial function $x_\lambda(t) = e^{\lambda t}$, we find the algebraic equation

$$p(\lambda) := \lambda^2 + 3\lambda + 2 = 0. \tag{55}$$

We can rewrite the characteristic polynomial (55) as

$$p(\lambda) = (\lambda + 2)(\lambda + 1). \tag{56}$$

Thus, the roots are given by

$$\lambda_1 = -1 \ , \qquad \lambda_2 = -2. \tag{57}$$

The general solution of (54) is therefore given by

$$x(t) = c_1 e^{\lambda_1 t} + c_2 e^{\lambda_2 t} = c_1 e^{-t} + c_2 e^{-2t}. \tag{58}$$

If we want to specify a solution uniquely, we have to fix two initial conditions. As an example, let's assume that the initial position $x(0)$ and the initial velocity $\dot{x}(0)$ are given by

$$x(0) = 4 \ , \qquad \dot{x}(0) = -3.$$

From the solution

$$x(0) = c_1 + c_2 = 4 \qquad \Rightarrow \qquad c_1 = 4 - c_2$$

Furthermore

$$\dot{x}(0) = -c_1 - 2c_2 = -(4 - c_2) - 2c_2 = -4 - c_2 = -3 \qquad \Rightarrow \qquad c_2 = -1$$

and $c_1 = 5$, yielding the final result

$$x(t) = 5e^{-t} - e^{-2t}. \tag{59}$$

Again, this solution looks structurally similar to the vectors in Eq. (1), if we identify

$$\boldsymbol{e}_1 = e^{-t} \ , \qquad \boldsymbol{e}_2 = e^{-2t}$$

as basis vectors and interpret 5 and -2 as the coordinates with respect to this basis.

5.3 Two complex roots

Now let's consider the general case where a linear homogenous 2nd-oder ODE has a characteristic polynomial

$$p(\lambda) = \lambda^2 + b\lambda + \omega^2 \tag{60}$$

with roots

$$\lambda_\pm = -\frac{b}{2} \pm \sqrt{\frac{b^2}{4} - \omega^2}. \tag{61}$$

When $b < 2\omega$, we are in the strongly damped case discussed in the previous section and the roots are real. In this case, the general solution reads

$$x(t) = c_1 e^{\lambda_+ t} + c_2 e^{\lambda_- t} \tag{62}$$

with real constants c_1 and c_2.

When $b > 2\omega$, we have two complex roots [7]

$$\lambda_\pm = \alpha \pm i\beta \ , \qquad \beta \neq 0. \tag{63}$$

where α and β are real numbers. Note that in this case

$$\lambda_+ = \overline{\lambda}_-. \tag{64}$$

Then a pair of basis solutions are

$$e^{\lambda_\pm t} = e^{(\alpha \pm i\beta)t} = e^{\alpha t} e^{\pm i\beta t} = e^{\alpha t}(\cos\beta t \pm i \sin\beta t) \tag{65}$$

Generally, if $e^{\lambda_+ t}$ and its complex conjugate $e^{\lambda_- t}$ are complex-valued basis solutions, then the real and imaginary parts

$$x_1(t) = \Re x = \frac{x + \bar{x}}{2} = e^{\alpha t}\cos\beta t \ , \qquad x_2(t) = \Im x = \frac{x - \bar{x}}{2i} = e^{\alpha t}\sin\beta t \tag{66}$$

form a pair of real-valued basis functions. Taking all real linear combinations

$$x(t) = c_1 e^{\alpha t}\cos\beta t + c_2 e^{\alpha t}\sin\beta t \tag{67}$$

of this real basis then gives all the real solutions to the ODE. Using the complex representation, we can also write

$$x(t) = Ae^{\lambda_+ t} + \overline{A}e^{\lambda_+ t}, \tag{68}$$

where $A = a_1 + ia_2$. The conjugate coefficient pair A and $\bar{A}$ ensures that $x(t) = \bar{x}(t)$, which is required for all real-valued functions (see also the lecture slides). Note that both (67) and (68) depend on two real parameters.

[7] The degenerate case $\beta = 0$, where $\lambda = \alpha$ is a root of multiplicity 2 will be discussed in detail later; we may anticipate that in this case the solution takes the form

$$x(t) = c_1 e^{\alpha t} + c_2 t e^{\alpha t} \ , \qquad \alpha = -\frac{b}{2}$$

with real constants c_1 and c_2.

5.4 Linear independence

We have seen above that we can interpret functions as vectors. The concept of *linear independence* provides an important characterization of geometric dependencies between collections of vectors or functions. As a reminder, let's first consider the two-dimensional plane $\mathbb{R}^2$. Two vectors $\boldsymbol{r}_1 = (x_1, y_1)$ and $\boldsymbol{r}_2 = (x_2, y_2)$ are *linearly dependent* if one is the multiple of the other, i.e., if there exists a real number α such that

$$\boldsymbol{r}_1 = \alpha \boldsymbol{r}_2.$$

We restate this equivalently as follows: Two vectors $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$ are *linearly dependent* if there exists two non-zero real numbers α_1, α_2 such that

$$\alpha_1 \boldsymbol{r}_1 + \alpha_2 \boldsymbol{r}_2 = \alpha_1 \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} + \alpha_2 \begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = \mathbf{0}. \tag{69}$$

Conversely, we can say that $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$ are *linearly independent*, if Eq. (69) can only be fulfilled for $\alpha_1 = \alpha_2 = 0$. This definition can be generalized to collections of n vectors $\{\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_n\}$, which are called *linearly independent* if

$$\alpha_1 \boldsymbol{r}_1 + \alpha_2 \boldsymbol{r}_2 + \ldots + \alpha_n \boldsymbol{r}_n = 0, \tag{70}$$

can only be fulfilled for $\alpha_1 = \alpha_2 = \ldots = \alpha_n = 0$; otherwise, the collection of vectors is linearly dependent. From this definition, it is easy check that any three (or more) vectors in the plane $\mathbb{R}^2$ are linearly dependent. Similarly, any four (or more) vectors in three-dimensional position space $\mathbb{R}^3$ are linearly dependent.

These concepts translate into a straightforward manner to functions by identifying the vectors $\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots$ with functions $f_1(t), f_2(t), \ldots$. For example, let's consider the two functions

$$f_1(t) = 1 \qquad \text{and} \qquad f_2(t) = t.$$

These two functions are linearly *independent* because, for *arbitrary* values of t, the equation

$$\alpha_1 1 + \alpha_2 t = 0 \tag{71}$$

can be satisfied only when $\alpha_1 = \alpha_2 = 0$. We can say the two functions $\{1, t\}$ are *basis vectors* that *span* the space of linear functions $f(t) = \alpha 1 + \beta t$. We write this compactly as

$$\operatorname{span}\{1, t\} = \{f(t) = \alpha 1 + \beta t\}, \qquad \alpha, \beta \in \mathbb{R}$$

By contrast, the functions

$$f(t) = 2t \qquad \text{and} \qquad g(t) = 4t$$

are linearly dependent, since $f = 2g$. Similarly, the functions $f_1(t) = 1$, $f_2(t) = t$ and $h(t) = 2 + 4t$ are linearly dependent since

$$2f_1 + 4f_2 - h = 0;$$

that is, there exist non-zero constants α_1, α_2, and α_3 such that

$$\alpha_1 f_1(t) + \alpha_2 f_2(t) + \alpha_3 h(t) = 0$$

for all values t, namely $(\alpha_1, \alpha_2, \alpha_3) = (2, 4, -1)$.

We can generalize this to higher-order polynomials, e.g.

$$\operatorname{span}\{1, t, t^4\} = \{f(t) = \alpha 1 + \beta t + \gamma t^4\} \ , \qquad \alpha, \beta, \gamma \in \mathbb{R} \tag{72}$$

The three basis functions $\{1, t, t^4\}$ are linearly independent, whereas any collection of four (or more) functions of the form $\alpha 1 + \beta t + \gamma t^4$ is linearly dependent. Instead of polynomial basis functions, we can also consider trigonometric or exponential basis functions. For example, Eq. (53) shows that $\{\cos \omega t, \sin \omega t\}$ span the solution space of the undamped harmonic oscillator of frequency ω. Similarly, Eq. (59) shows that $\{e^{-t}, e^{-2t}\}$ span the solution space of the damped harmonic oscillator described by Eq. (54).

6 Sinusoidal Functions

At the end of the previous section, we have seen that linear nth order ODEs can be ‘complexified’. As we shall find later, it is often easier to solve the complexified version of a linear ODE and subsequently transform back to a real solution space, by projecting on the real or imaginary part. To utilize this technique, we will introduce basic concepts about sinusoidal functions in this section.

6.1 Complex functions of a real variable

It is helpful to recall that a complex number $z = x + iy$ corresponds to a 2D vector in the plane $\mathbb{R}^2$. Now let’s assume that $x = x(t)$ and $y = y(t)$ are continuous functions of time t. Then, accordingly, the complex function

$$z(t) = x(t) + iy(t) \tag{73}$$

describes a curve in the plane. We can differentiate a complex function by differentiating real and imaginary parts separately

$$\dot{z}(t) = \dot{x}(t) + i\dot{y}(t) \ , \qquad \ddot{z}(t) = \ddot{x}(t) + i\ddot{y}(t) \tag{74}$$

If we interpret $z(t)$ as the trace of a particle moving in the plane, then $\dot{z}(t)$ describes particle’s velocity and $\ddot{z}$ its acceleration.

Analogously, we can integrate $z(t)$ by integrating real and imaginary parts separately

$$Z(t) = \int_{t_0}^{t} ds\, z(s) = \int_{t_0}^{t} ds\, x(s) + i \int_{t_0}^{t} ds\, y(s) \tag{75}$$

6.2 Complex exponential, sine and cosine

We had already seen that complex exponential function $\exp(i\phi)$ and the trigonometric functions $\sin\phi$ and $\cos\phi$, can be related by Euler's formula

$$e^{i\phi} = \cos\phi + i\sin\phi\ , \qquad \phi \in \mathbb{R} \tag{76}$$

Let's consider the two cases $\phi = +t$ and $\phi = -t$. From Euler's formula (76), we then find that the functions e^{it} and e^{-it} are linear combinations of the functions $\cos t$ and $\sin t$:

$$e^{it} \;=\; \cos t + i\ \sin t \tag{77a}$$

$$e^{-it} \;=\; \cos t - i\ \sin t. \tag{77b}$$

Which curves do these functions describe? Computing the modulus of $z(t) = e^{it}$ gives

$$z(t)\bar{z}(t) = e^{it}e^{-it} = e^{it-it} = e^0 = 1. \tag{78}$$

Furthermore, let's compute the derivative of $z(t)$ at $t = 0$, noting that

$$\dot{z}(t) = \frac{d}{dt}e^{it} = ie^{it} = i(\cos t + i\ \sin t) = -\sin t + i\cos t \tag{79}$$

and therefore

$$\dot{z}(0) = 0 + i\cdot 1. \tag{80}$$

Thus, $z(t) = e^{it}$ describes a circle of radius 1, which is traversed in counterclockwise direction. Similarly, we find that the curve corresponding to $z(t) = e^{-it}$ describes a clockwise circle.

If we view e^{it} and e^{-it} as known, and $\cos t$ and $\sin t$ as unknown, then this is a system of two linear equations in two unknowns, and can be solved for $\cos t$ and $\sin t$. This gives

$$\cos t = \frac{e^{it} + e^{-it}}{2}, \qquad\qquad \sin t = \frac{e^{it} - e^{-it}}{2i}. \tag{81}$$

Thus $\cos t$ and $\sin t$ are linear combinations of e^{it} and e^{-it}. Explicitly,

$$\sin t = \frac{1}{2i}e^{it} + \frac{-1}{2i}e^{-it}.$$

where $(\frac{1}{2i}, \frac{-1}{2i})$ are the coordinates of $\sin t$ with respect to the basis vectors e^{it} and e^{-it}.

From a practical point of view, it is important that the function e^z has nicer properties than $\cos t$ and $\sin t$, so it is often a good idea to use these formulas to replace $\cos t$ and $\sin t$ by these combinations of e^{it} and e^{-it}, or to view $\cos t$ and $\sin t$ as the real and imaginary parts of e^{it}.

Similarly, replacing $\phi = \pm\omega t$ in the identities above leads to

$$\begin{aligned} e^{i\omega t} &= \cos\omega t + i\sin\omega t \\ e^{-i\omega t} &= \cos\omega t - i\sin\omega t. \end{aligned}$$

and

$$\cos\omega t = \frac{e^{i\omega t} + e^{-i\omega t}}{2}, \qquad\qquad \sin\omega t = \frac{e^{i\omega t} - e^{-i\omega t}}{2i},$$

where ω is the angular frequency.

6.3 Sinusoidal functions

Sinusoidal functions are obtained from sines and cosines through stretching and shifting.

6.3.1 Construction

Start with the curve $y = \cos x$. Then

1. Shift the graph ϕ units to the right (ϕ is *phase lag*, measured in radians). For example, shifting by $\phi = \pi/2$ gives the graph of $\sin x$, which reaches its maximum $\pi/2$ radians after $\cos x$ does.
2. Compress the result horizontally by *dividing* by a scale factor ω, called *angular frequency* and measured in radians/s.
3. Amplify (stretch vertically) by a factor of A (*amplitude*).

Here, we assume that $A, \omega > 0$, but ϕ can be any real number. The graph of the new function $f(t)$, called a *sinusoid function.* What is the formula for $f(t)$? According to the instructions, each point (x, y) on $y = \cos x$ is related to a point $(t, f(t))$ on the graph of f by

$$t = \frac{x + \phi}{\omega}, \quad f = Ay.$$

Solving for x gives $x = \omega t - \phi$; substituting into $f = Ay = A\cos x$ gives

$$f(t) = A\cos(\omega t - \phi). \tag{82a}$$

The period is

$$T = \frac{2\pi}{\omega}. \tag{82b}$$

We can also write

$$f(t) = A\cos[\omega(t - \tau)]\ , \qquad \tau = \phi/\omega. \tag{83}$$

Remark: There is also *frequency* $\nu := 1/T$, measured in Hz = cycles/s. It is the number of complete oscillations per second. To convert from frequency ν to angular frequency ω, multiply by (2π radians)/(1 cycle); thus $\omega = 2\pi\nu = 2\pi/T$, which is consistent with the formula $T = 2\pi/\omega$ above.

6.3.2 Three representations

There are three ways to write a sinusoid function:

(i) *amplitude-phase form*: $A\cos(\omega t - \phi)$

(ii) *complex form*: $\Re\left[ce^{i\omega t}\right]$, where $c = a - ib = Ae^{-i\phi}$ is a complex number

(iii) *linear combination*: $a \cos \omega t + b \sin \omega t$, where a and b are real numbers

We convert between them as follows:

1. (ii)→ (i):

$$\Re\left[ce^{i\omega t}\right] = \Re\left[Ae^{-i\phi}\, e^{i\omega t}\right] = \Re\left[Ae^{i(\omega t - \phi)}\right] = A\cos(\omega t - \phi).$$

2. (ii)→ (iii):

$$\begin{aligned} \Re\left[ce^{i\omega t}\right] &= \Re\left[(a-bi)(\cos\omega t + i \sin\omega t)\right] \\ &= \Re\left[a\cos\omega t + b\sin\omega t + i(\cdots)\right] \\ &= a\cos\omega t + b\sin\omega t. \end{aligned} \tag{84}$$

3. (i)→ (iii): Using

$$\cos(x-y) = \cos x\cos y + \sin x \sin y$$

shows that

$$\begin{aligned} A\cos(\omega t - \phi) &= A\cos\omega t\cos\phi + A\sin\omega t \sin\phi \\ &= a\cos\omega t + b\sin\omega t \end{aligned}$$

since $a = A\cos\phi$ and $b = A\sin\phi$, when A, ϕ are the polar coordinates of (a, b).

6.4 Example: Beats

Beats occur when two very nearby pitches are sounded simultaneously. As an example, consider two sinusoid sound waves of angular frequencies $\omega+\epsilon$ and $\omega-\epsilon$, say $\cos[(\omega+\epsilon)t]$ and $\cos[(\omega-\epsilon)t]$, where ϵ is much smaller than ω. What happens when they are superimposed?

The sum is

$$\begin{aligned} \cos((\omega+\epsilon)t) + \cos((\omega-\epsilon)t) &= \Re[e^{i(\omega+\epsilon)t}] + \Re[e^{i(\omega-\epsilon)t}] \\ &= \Re[e^{i\omega t}(e^{i\epsilon t} + e^{-i\epsilon t})] \\ &= \Re[e^{i\omega t}(2\cos\epsilon t)] \\ &= 2\cos(\epsilon t)\Re[e^{i\omega t}] \\ &= 2\cos(\epsilon t)\ (\cos\omega t). \end{aligned}$$

The function $\cos\omega t$ oscillates rapidly between ±1. Multiplying it by the slowly varying function $2\cos\epsilon t$ produces a rapid oscillation between $\pm 2\cos\epsilon t$, so one hears a sound wave of angular frequency ω whose amplitude is the slowly varying function $|2\cos\epsilon t|$.

7 Some linear algebra

In this section, we will summarize and review essential concepts of about vectors spaces and matrices that will be important in the remainder of this course. A main goal of this part is to provide a rigorous framework for some of the mathematical 'tools' that we have seen and used in the earlier parts.

7.1 Vector spaces

Intuitively, vectors are objects that we can add and/or multiply by scalars (real or complex numbers) to obtain new vectors. Sets of objects that are 'closed' under these operations are called *vector spaces*. Here, *closedness* means that the operations 'vector addition' and 'multiplication by scalars' do not lead out of the set. In previous classes, we have already seen two specific realizations of vector spaces. One frequently encountered example was the plane $\mathbb{R}^2$ spanned by the vectors

$$\boldsymbol{e}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \ , \qquad \boldsymbol{e}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix} . \tag{85}$$

This vector space is given by the set

$$\mathbb{R}^2 = \{\text{all} \quad \boldsymbol{v} = c_1\boldsymbol{e}_1 + c_2\boldsymbol{e}_2 \qquad \text{with} \quad c_1, c_2 \in \mathbb{R}\}. \tag{86}$$

Another vector space example are the solutions of the harmonic oscillator

$$\mathbb{H} = \{\text{all} \quad f(t) = c_1 \cos(\omega t) + c_2 \sin(\omega t) \qquad \text{with} \quad c_1, c_2 \in \mathbb{R}\}. \tag{87}$$

The general definition of a vector space $\mathbb{V}$, which covers these two and many other examples, is as follows:

A vector space over the real (complex) numbers is a set such that

(i) the zero vector $\boldsymbol{0}$ is in $\mathbb{V}$.

(ii) if $\boldsymbol{v}$ is an element of $\mathbb{V}$ then $c\boldsymbol{v}$ is also an element of $\mathbb{V}$ for any real (complex) number c.

(iii) if $\boldsymbol{v}$ and $\boldsymbol{w}$ are elements of $\mathbb{V}$ then $\boldsymbol{u} = \boldsymbol{v} + \boldsymbol{w}$ is also an element of $\mathbb{V}$.

If the set of numbers is chosen to be real (complex), then we say that $\mathbb{V}$ is a real (complex) vector space. The last two conditions (ii) and (iii) ensure that the vector space is *closed* under addition and scalar multiplication.

Note that the set of real numbers itself can be interpreted as a real vector space. The set of polynomials of degree n will real coefficients is a real vector space. $\mathbb{C}^2 = \{(z_1, z_2) : z_1, z_2 \in \mathbb{C}\}$ is complex vector space.

Subvector spaces are subsets of vector spaces that are by themselves vector spaces. For example, any straight line through the origin is a subvector space of $\mathbb{R}^2$. Similarly, straight lines through the origin or any plane including the origin form a subvector space of $\mathbb{R}^2$.

Recall that a *linear combination* of a collection of vectors $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$ is a vector of the form

$$c_1\boldsymbol{v}_1 + \ldots + c_n\boldsymbol{v}_n. \tag{88}$$

Examples are the solutions of ODEs. The *span* of the set $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$, denoted by

$$\text{Span}(\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\})$$

is the set of all linear combinations of $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_n\}$. This set is always a vector space.

A set of vectors $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ is called linearly independent if one of them is a linear combination of the others, or equivalently, if there exist scalars $c_1, \ldots, c_n$ *not all zero* so that

$$c_1\boldsymbol{v}_1 + \ldots + c_n\boldsymbol{v}_n = \mathbf{0}. \tag{89}$$

Vectors that are not linearly dependent are called *linearly independent.* For example, the two vectors $\{\sin x, \cos x\}$ are linearly independent but the three vectors $\{\sin x, \cos x, \sin x + 137\cos x\}$ are not.

The *dimension* $\dim \mathbb{V}$ of a vector space is the largest number of linearly independent elements one can find in $\mathbb{V}$, for example

$$\dim \mathbb{R} = 1\ , \qquad \dim \mathrm{Span}\{\sin x, \cos x, \sin x + 137\cos x\} = 2. \tag{90}$$

Any set of $d = \dim \mathbb{V}$ linearly independent vectors $\{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_d\}$ forms a *basis* of $\mathbb{V}$. The *components* $(c_1, \ldots, c_n)$ of a vector $\boldsymbol{v}$ with respect to a given basis $\{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_d\}$ are the numbers required to express $\boldsymbol{v}$ as superposition of the basis vectors

$$\boldsymbol{v} = c_1\boldsymbol{b}_1 + \ldots + c_n\boldsymbol{b}_n. \tag{91}$$

We often write the components in column form

$$\boldsymbol{v} = \begin{pmatrix} c_1 \\ \vdots \\ c_d \end{pmatrix}. \tag{92}$$

It is important to keep in mind that this representation refers to a specific fixed basis system. That is, if we pick another basis $\{\boldsymbol{b}'_1, \ldots, \boldsymbol{b}'_d\}$ then the same vector would be characterized by a different set of components $(c'_1, \ldots, c'_n)$.

7.2 Matrices as functions between vector spaces

Any given $n \times m$ matrix

$$A = \begin{pmatrix} A_{11} & A_{12} & \cdots A_{1m} \\ \vdots & \vdots & \vdots \\ A_{n1} & A_{12} & \cdots A_{nm} \end{pmatrix} \tag{93}$$

can be naturally interpreted as a map from an m-dimensional vector space $\mathbb{V}$ to an n-dimensional vector space $\hat{\mathbb{V}}$. To see this, let's assume we have fixed a basis $\{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m\}$ for $\mathbb{V}$ and another basis $\{\hat{\boldsymbol{b}}_1, \ldots, \hat{\boldsymbol{b}}_n\}$ for $\hat{\mathbb{V}}$. Then vectors $\boldsymbol{v} \in \mathbb{V}$ and $\hat{\boldsymbol{v}} \in \hat{\mathbb{V}}$ correspond to column vectors

$$\boldsymbol{v} = \begin{pmatrix} c_1 \\ \vdots \\ c_m \end{pmatrix} \qquad \text{and} \qquad \hat{\boldsymbol{v}} = \begin{pmatrix} \hat{c}_1 \\ \vdots \\ \hat{c}_n \end{pmatrix}$$

respectively. The matrix A assigns to each vector $\boldsymbol{v} \in \mathbb{V}$ another vector $\hat{\boldsymbol{v}} = A\boldsymbol{v} \in \hat{\mathbb{V}}$ with components

$$\begin{pmatrix} \hat{c}_1 \\ \vdots \\ \hat{c}_n \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} & \cdots A_{1m} \\ \vdots & \vdots & \vdots \\ A_{n1} & A_{12} & \cdots A_{nm} \end{pmatrix} \begin{pmatrix} c_1 \\ \vdots \\ c_m \end{pmatrix} \tag{94}$$

That is, the matrix defines a linear map from $\mathbb{V}$ to $\hat{\mathbb{V}}$. Conversely, Eq. (94) can be interpreted as a linear equation that formalizes the following inverse problem:

Given the image vector $\hat{\boldsymbol{v}}$, can we find the original vector(s) $\boldsymbol{v}$ that are mapped by A onto $\hat{\boldsymbol{v}}$?

In the next section, we will learn a systematic procedure for answering this question.

Beforehand, let's briefly consider the special case $\mathbb{V} = \hat{\mathbb{V}}$. Then any linear map $\mathbb{V} \to \mathbb{V}$ can be represented by a square matrix A. An example, is the rotation matrix

$$R_x(\phi) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\phi & \sin\phi \\ 0 & -\sin\phi & \cos\phi \end{pmatrix} \tag{95}$$

which keeps vectors along the x-axis fixed, thus describing rotations about this axis. The determinant of this matrix is given by

$$\det R(\phi) = 1(\cos\phi)(\cos\phi) - (\sin\phi)(-\sin\phi) = 1. \tag{96}$$

Linear maps of determinant ± 1 are *volume conserving.* Volume conservation is intuitively obvious for rotations, but in most cases it is not immediately obvious for a given matrix A whether or not the associated map $\hat{\boldsymbol{v}} = A\boldsymbol{v}$ is volume conserving - in these cases, the determinant criterion is very useful.

8 Higher-order linear systems and superposition

Until stated otherwise, we will focus on linear ODEs for scalar functions from now on. In Eq. (20) above, we had seen that Newton's equations for a damped harmonic oscillator can be written in the standard form

$$\ddot{x} + b\dot{x} + \omega^2 x = \epsilon \sin(\Omega t), \tag{97}$$

where the rhs. represents an oscillatory driving signal (input) that determines the system response $x(t)$. We now generalize from this second-order ODE to linear nth order ODEs.

8.1 General case: Homogeneous *vs.* inhomogeneous linear equations

The most general form of an nth order linear ODE for a function $x(t)$ with given t-dependent coefficients $p_k(t)$ is[8]

$$p_n(t)x^{(n)} + p_{n-1}(t)x^{(n-1)} + \ldots + p_2(t)\ddot{x} + p_1(t)\dot{x} + p_0(t)x = q(t). \tag{98}$$

[8] An ODE that cannot be expressed in this form is nonlinear.

From an engineering perspective, it is natural to interpret the function $q(t)$ on the rhs. of Eq. (98) as an *external input signal*, while the sought-after solution $x(t)$ describes the *system response*. If the function $q(t)$ is identically zero, $q(t) \equiv 0$, the linear ODE is called *homogeneous*. For a non-vanishing function $q(t) \not\equiv 0$, the equations is called *inhomogeneous*.

Thus, the example from Eq. (21),

$$e^t \ddot{x} + 5\dot{x} + t^9 x = 0,$$

is *homogeneous*, whereas

$$e^t \ddot{x} + 5\dot{x} + t^9 x = \sin(3t) \tag{99}$$

is *inhomogeneous*.

8.2 Existence and uniqueness of solutions

Using separation of variables (in the homogeneous case) and variation of parameters (in the inhomogeneous case), we showed that every first-order linear ODE has a 1-parameter family of solutions. To nail down a specific solution in this family, we needed *one* initial condition, such as $y(0)$. Similarly, it turns out that every second-order linear ODE has a 2-parameter family of solutions. That is, to nail down a specific solution, we need *two* initial conditions at the same starting time, $y(0)$ and $\dot{y}(0)$. The starting time could also be some number t_0 other than 0. These are consequence of the following general **theorem:**

Let $p_{n-1}(t)$, …, $p_0(t)$, $q(t)$ be continuous functions on an open interval I. Let $t_0 \in I$, and let $b_0, \ldots, b_{n-1}$ be given numbers. Then there exists a unique solution $y(t)$ to the nth order linear ODE

$$y^{(n)} + p_{n-1}(t)\, y^{(n-1)} + \cdots + p_1(t)\, \dot{y} + p_0(t)\, y = q(t)$$

satisfying the n initial conditions

$$y(t_0) = b_0, \quad \dot{y}(t_0) = b_1, \quad \ldots, \quad y^{(n-1)}(t_0) = b_{n-1}.$$

Here, *existence* means that there is *at least* one solution, while *uniqueness* means that there is *only* one solution. For a linear ODE as above, the solution $y(t)$ is defined on the whole interval I where the functions $p_{n-1}(t)$, …, $p_0(t)$, $q(t)$ are continuous. In particular, if $p_{n-1}(t)$, …, $p_0(t)$, $q(t)$ are continuous on all of $\mathbb{R}$, then the solution $y(t)$ will be defined on all of $\mathbb{R}$.

8.3 Superposition principle

During our above discussion of the population growth model in Sec. 2.2.1, we had seen how one can construct the general solution to a linear inhomogeneous first-order ODE by superposition. After looking at a few more examples in Sec. 8.3.1, we will generalize this concept to linear nth order equations in Sec. (8.3.2).

8.3.1 Examples

Let's compare the solutions to a homogeneous equation and some inhomogeneous equations with the same left hand side:

$$\begin{array}{llll} \text{The general solution to} & t\dot{y}+2y=0 & \text{is} & y_h=ct^{-2} \\ \\ \text{A particular solution to} & t\dot{y}+2y=t^5 & \text{is} & y_p= \qquad t^5/7 \\ \text{The general solution to} & t\dot{y}+2y=t^5 & \text{is} & y \ \ =ct^{-2}+t^5/7 \\ \\ \text{A particular solution to} & t\dot{y}+2y=1 & \text{is} & y_p= \qquad 1/2 \\ \text{The general solution to} & t\dot{y}+2y=1 & \text{is} & y \ \ =ct^{-2}+1/2 \end{array}$$

Furthermore, scaler-multiply the particular solutions above to obtain

$$\begin{array}{llll} \text{A particular solution to} & t\dot{y}+2y=9t^5 & \text{is} & y_p=9t^5/7 \\ \text{A particular solution to} & t\dot{y}+2y=3 & \text{is} & y_p= \qquad 3/2 \end{array}$$

and add to find

$$\text{A particular solution to} \quad t\dot{y}+2y=9t^5+3 \quad \text{is} \quad y_p=9t^5/7+3/2$$

8.3.2 General form

The general principle, which works for all *linear* ODEs, is this:

(i) Multiplying a solution to

$$p_n(t)\,y^{(n)}+\cdots+p_0(t)\,y \quad = \quad q(t) \tag{100a}$$

by a number α gives a solution to

$$p_n(t)\,y^{(n)}+\cdots+p_0(t)\,y \quad = \quad \alpha q(t). \tag{100b}$$

(ii) Adding a solution $y_1(t)$ of

$$p_n(t)\,y^{(n)}+\cdots+p_0(t)\,y \quad = \quad q_1(t) \tag{101a}$$

to a solution $y_2(t)$ of

$$p_n(t)\,y^{(n)}+\cdots+p_0(t)\,y \quad = \quad q_2(t) \tag{101b}$$

gives a solution $y=y_1(t)+y_2(t)$ of

$$p_n(t)\,y^{(n)}+\cdots+p_0(t)\,y \quad = \quad q_1(t)+q_2(t). \tag{101c}$$

8.3.3 Outlook: Complexification

The properties (i) and (ii) can be combined to complexify a linear nth order ODE, which means that if $x(t)$ and $y(t)$ solves

$$p_n(t)\,x^{(n)}+\cdots+p_0(t)\,x \quad = \quad u(t) \tag{102a}$$

$$p_n(t)\,y^{(n)}+\cdots+p_0(t)\,y \quad = \quad v(t) \tag{102b}$$

then $z(t)=x(t)+iy(t)$ solves

$$p_n(t)\,z^{(n)}+\cdots+p_0(t)\,z \quad = \quad u(t)+iv(t) \tag{102c}$$

9 Operator notation and matrix analogy

Let us now generalize the approach of the previous section by considering the more general case of an nth order linear ODE for a function $x(t)$ with *constant* coefficients p_k

$$p_n x^{(n)} + p_{n-1}x^{(n-1)} + \ldots + p_2\ddot{x} + p_1\dot{x} + p_0 x = q(t). \tag{103}$$

To write this equation compactly, we introduce the *differential operator*

$$\mathcal{D} := \frac{d}{dt}$$

which acts on a time-dependent function $x(t)$ by producing a new function[9]

$$\mathcal{D}x(t) = \frac{d}{dt}x(t) = \dot{x}(t) = v(t)$$

It follows from the usual rules of differentiation that for any two functions $x(t)$ and $y(t)$ and any two constants α and β

$$\mathcal{D}[\alpha x(t) + \beta y(t)] = \frac{d}{dt}[\alpha x(t) + \beta y(t)] = \alpha\frac{d}{dt}x(t) + \beta\frac{d}{dt}y(t) = \alpha\mathcal{D}x(t) + \beta\mathcal{D}y(t),$$

which means that the operator $\mathcal{D}$ is called *linear*. Please check that the operator $\mathcal{D}^n = d^n/dt^n$ is also linear. Using $\mathcal{D}$ and its powers, we can define a more complicated linear operator

$$\mathcal{L} \quad = \quad p_n D^n + p_{n-1}D^{n-1} + \ldots + p_2 D^2 + p_1 D + p_0 D^0 \tag{104a}$$

which explicitly reads

$$\mathcal{L} \quad = \quad p_n\frac{d^n}{dt^n} + p_{n-1}\frac{d^{n-1}}{dt^{n-1}} + \ldots + p_2\frac{d^2}{dt^2} + p_1\frac{d}{dt} + p_0, \tag{104b}$$

In terms of $\mathcal{L}$, we can rewrite Eq. (103) compactly as

$$\mathcal{L}x(t) = q(t). \tag{105}$$

It is no coincidence that Eq. (105) bears formal resemblance to the linear matrix equation

$$A\boldsymbol{x} = \boldsymbol{b} \tag{106}$$

where A is an $n \times n$ square matrix, and $\boldsymbol{x}$ and $\boldsymbol{b}$ are n-dimensional row vectors

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ \ldots \\ x_n \end{pmatrix}, \qquad \boldsymbol{b} = \begin{pmatrix} b_1 \\ \ldots \\ b_n \end{pmatrix}$$

Matrix multiplication is a linear operation on vectors

$$A(\alpha\boldsymbol{v} + \beta\boldsymbol{u}) = \alpha A\boldsymbol{v} + \beta A\boldsymbol{u} \tag{107}$$

[9] That is, an operator is a map from functions to functions, just as a function is a map from number to numbers.

just as $\mathcal{L}$ is a linear operator on functions.

Given Eq. (105), we would like to find $x(t)$ for given $\mathcal{L}$ and $q(t)$, while in the case of Eq. (106) we would like to find $\boldsymbol{x}$ for given A and $\boldsymbol{b}$. Both Eq. (105) and Eq. (106) are linear, which for example means that their general solutions are given by

$$x(t) = x_h(t) + x_p(t) \ , \qquad\qquad \boldsymbol{x} = \boldsymbol{x}_h + \boldsymbol{x}_p \tag{108}$$

where the subscript h indicates the general solution of the associated homogeneous problems (with $q \equiv 0$ and $\boldsymbol{b} = \boldsymbol{0}$, respectively), and subscript p indicates a particular solution of the full inhomogeneous equation.

In the remainder of this section, we will outline deeper structural similarities between Eqs. (105) and (106).

9.1 Eigenvalues and eigenvectors of matrices

We start by considering the matrix equation $A\boldsymbol{x} = \boldsymbol{b}$. For any square matrix A, we can find a certain number of special vectors $\boldsymbol{v}$, satisfying

$$A\boldsymbol{v} = \lambda\boldsymbol{v} \tag{109}$$

for some real or complex constant λ. These vectors are called the *eigenvectors* of A and λ is called the corresponding *eigenvalue*. The collection of all eigenvalues λ is called the spectrum of the matrix A, and typically denoted by

$$\text{spec}(A).$$

Intuitively, eigenvectors are special because application of A merely stretches or shrinks an eigenvector $\boldsymbol{v}$ by λ; in general, applying A to some non-eigenvector $\boldsymbol{w}$ produces a new vector $\boldsymbol{u} := A\boldsymbol{w}$ that does not point in the same direction as $\boldsymbol{w}$.

To illustrate, how one can find the eigenvectors of a given matrix A, we consider as a specific example the 2×2-matrix

$$A = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$$

and rewrite the eigenvector condition (109) in the equivalent form

$$(A - \lambda I)\boldsymbol{v} = \boldsymbol{0} \tag{110}$$

by subtracting $\lambda I\boldsymbol{v}$ on both sides of Eq. (109), where

$$I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

is the 2×2 identity matrix that leaves every vector unchanged. Equation (110) features the matrix

$$A - I\lambda = \begin{pmatrix} 1-\lambda & 2 \\ 2 & 1-\lambda \end{pmatrix} \tag{111}$$

which is just the matrix A with λ subtracted on the diagonal. Equation (110) is always solved by $\boldsymbol{v} = 0$, but this is a trivial uninteresting solution. To find nontrivial solutions $\boldsymbol{v}$ we must demand that the determinant of the matrix $A - I\lambda$ vanishes[10]

$$p(\lambda) := \det(A - \lambda I) = 0. \tag{112}$$

The function $p(\lambda)$ is the *characteristic polynomial* of A, and its roots are the eigenvalues of A. For our example,

$$p(\lambda) = \det \begin{pmatrix} 1-\lambda & 2 \\ 2 & 1-\lambda \end{pmatrix} = (1-\lambda)\cdot(1-\lambda) - 2\cdot 2 = \lambda^2 - 2\lambda - 3,$$

which has roots

$$\lambda_1 = -1\ , \qquad \lambda_2 = 3.$$

Inserting the eigenvalues into Eq. (110) gives

$$\begin{aligned} \begin{pmatrix} 1-\lambda_1 & 2 \\ 2 & 1-\lambda_1 \end{pmatrix} \boldsymbol{v}_1 &= \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix} \boldsymbol{v}_1 = \mathbf{0} \\ \begin{pmatrix} 1-\lambda_2 & 2 \\ 2 & 1-\lambda_2 \end{pmatrix} \boldsymbol{v}_2 &= \begin{pmatrix} -2 & 2 \\ 2 & -2 \end{pmatrix} \boldsymbol{v}_2 = \mathbf{0} \end{aligned}$$

which is satisfied by the eigenvectors

$$\boldsymbol{v}_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}\ , \qquad \boldsymbol{v}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Note that the eigenvectors are linearly independent and in fact even orthogonal[11], so that any other two-dimensional real vector $\boldsymbol{x}$ can be expressed as

$$\boldsymbol{x} = c_1\boldsymbol{v}_1 + c_2\boldsymbol{v}_2\ , \qquad c_1, c_2 \in \mathbb{R}.$$

In general, the eigenvalues and eigenvectors of a real matrix can be complex (see 18.06 for more on this).

9.2 Eigenvalues and eigenfunctions of linear operators

Let's now compare this to differential operators, by considering the homogeneous nth order ODE

$$\mathcal{L}x(t) = \left(p_n \frac{d^n}{dt^n} + p_{n-1}\frac{d^{n-1}}{dt^{n-1}} + \ldots + p_2\frac{d^2}{dt^2} + p_1\frac{d}{dt} + p_0\right) x(t) = 0 \tag{113}$$

Inserting the trial function $x_\lambda = e^{\lambda t}$, we find

$$p(\lambda) := p_n\lambda^n + p_{n-1}\lambda^{n-1} + \ldots + p_2\lambda^2 + p_1\lambda + p_0 = 0$$

[10] Otherwise, $\boldsymbol{v} = 0$ would be the only possible solution.

[11] For sufficiently nice (e.g., Hermitean) matrices, this is always the case/achievable.

where now $p(\lambda)$ is the now the characteristic polynomial of the ODE. First, let us note that we can express the operator $\mathcal{L}$ in terms of the characteristic polynomial

$$\mathcal{L} = p(\mathcal{D}) \ , \qquad \mathcal{D} = \frac{d}{dt}.$$

Secondly, we see that the differential operator $\mathcal{L}$ plays a role analogous to that of the matrix $A - \lambda I$ in Eq. (110), and that we can interpret the roots $\lambda_1, \ldots, \lambda_n$ of $p(\lambda)$ as 'eigenvalues' and the corresponding solutions $e^{\lambda_1 t}, \ldots, e^{\lambda_n t}$ as 'eigenfunctions'. If all eigenvalues λ_i are distinct, then we say the spectrum of $\mathcal{L}$ is *non-degenerate*. Assuming that all the roots λ_i are distinct and real, the general homogeneous solution of Eq. (113) can then be written as

$$x_h(t) = c_1 e^{\lambda_1 t} + \ldots + c_n e^{\lambda_n t} \ , \qquad c_1, \ldots, c_n \in \mathbb{R}. \tag{114}$$

or, equivalently, in compact sum notation

$$x_h(t) = \sum_{i=1}^{n} e^{\lambda_i t} c_i \ , \qquad c_i \in \mathbb{R}. \tag{115}$$

9.3 Basis solutions for repeated roots

If some of the eigenvalues λ_i are repeated, i.e., have a multiplicity greater than 1, then the spectrum is called *degenerate*. In the degenerate case, we have to modify the basis functions as follows: Let's assume we have $k \leq n$ distinct real roots λ_α of multiplicity m_α, which means that we can write the characteristic polynomial as

$$p(\lambda) = (\lambda - \lambda_1)^{m_1} (\lambda - \lambda_2)^{m_2} \cdots (\lambda - \lambda_k)^{m_k}. \tag{116}$$

For an nth order ODE, there must n roots in total so that

$$m_1 + m_2 + \ldots + m_k = n. \tag{117}$$

In this case, the general solution can be written as

$$\begin{aligned} x_h(t) \ &= \ e^{\lambda_1 t} \left(c_{11} + c_{12} t^1 + \ldots + c_{1m_1} t^{m_1 - 1} \right) + \ldots + \\ & \quad\ e^{\lambda_k t} \left(c_{k1} + c_{k2} t^1 + \ldots + c_{1m_k} t^{m_k - 1} \right) \ , \qquad\qquad c_{\alpha s} \in \mathbb{R}. \end{aligned}$$

or, equivalently, in compact sum notation

$$x_h(t) = \sum_{\alpha=1}^{k} e^{\lambda_\alpha t} \left(\sum_{s=1}^{m_\alpha} c_{\alpha s} t^{s-1} \right) \ , \qquad c_{\alpha s} \in \mathbb{R}. \tag{118}$$

To see how we can construct these solutions, let's consider the case where r is a double-root of $p(\lambda)$. In this case, we can write

$$p(\lambda) = (\lambda - r)^2 g(\lambda). \tag{119}$$

Obviously, e^{rt} is a solution, since

$$p(\mathcal{D}) e^{rt} = p(r) e^{rt} = 0. \tag{120}$$

If we find another nonzero solution $y(t)$ satisfying $y(0) = 0$, then it is definitely not a nonzero multiple of e^{rt}, so it's a viable candidate. We find this solution by moving a root slightly. Let

$$p_s(\lambda) = (\lambda - s)(\lambda - r)g(\lambda)$$

which is related to $p(\lambda)$ by

$$p(\lambda) = p_r(\lambda) = (\lambda - r)^2 g(\lambda)$$

If $s \neq r$, then both $y_s = e^{st}$ and $y_r = e^{rt}$ solve $p_s(\mathcal{D})y = 0$. It follows that

$$y = y_s - y_r = e^{st} - e^{rt}$$

solves $p_s(\mathcal{D})y = 0$ and satisfies $y(0) = 0$. Now this solution does not help in the limit as $s \to r$ because it tends to zero for all values of t. But if we multiply by the constant $1/(s-r)$ we get a solution that has a nonzero limit. Indeed,

$$\lim_{s\to r} \frac{e^{st} - e^{rt}}{s - r} = \frac{d}{ds} e^{st}\bigg|_{s=r} = te^{rt}$$

is the limit as $s \to r$ of solutions to $p_s(\mathcal{D})y = 0$. Hence, in conclusion,

$$p(\mathcal{D})(te^{rt}) = 0.$$

Of course once we know that te^{rt} works, we don't need this derivation any more. We can plug it in to check that it solves the equation. We leave it as an exercise to check that

$$(D - r)^2(te^{rt}) = 0.$$

It then follows that $p(\mathcal{D})(te^{rt}) = 0$ because $p(\mathcal{D}) = g(\mathcal{D})(\mathcal{D} - r)^2$. The same works for higher powers:

$$(\mathcal{D} - r)^k(t^\ell e^{rt}) = 0, \quad \ell = 0, 1, \ldots, k-1.$$

10 Inhomogeneous ODEs and exponential response formula

Thus far, we have mostly focused on *homogeneous* linear nth order ODEs

$$\mathcal{L}x(t) = p(\mathcal{D})x = \left(p_n\mathcal{D}^n + p_{n-1}\mathcal{D}^{n-1} + \ldots p_1\mathcal{D} + p_0\right) x \equiv 0. \tag{121}$$

and we have seen how to find the general solution for these types of problems. In this part, we introduce a systematic procedure for constructing particular solutions for an important class of *inhomogeneous* problems

$$\mathcal{L}x(t) = q(t). \tag{122}$$

As noted earlier, we interpret the *driving function* $q(t)$ as an external *input signal* and $x(t)$ the *system response*.

For a wide range of practically important problems, the driving function $q(t)$ falls in the class of sinusoidal functions. The most basic examples are $q(t) = \sin(\Omega t)$ or $q(t) = \cos(\Omega t)$. The key idea for tackling these types of problems is to complexify Eq. (122), by replacing

$x(t)$ on the lhs. by the complex function $z(t) = x(t) + iy(t)$ and by replacing $q(t)$ on the rhs. by

$$Q(t) = e^{i\Omega t} = \cos(\Omega t) + i\sin(\Omega t). \tag{123}$$

Once we have solved the complexified problem

$$\mathcal{L}z(t) = e^{i\Omega t}, \tag{124}$$

we can recover the sought-after particular solution from the real or imaginary part of the complex solution $z(t)$, depending on whether the driving function $q(t)$ is a cosine or sine function.

10.1 Time Invariance

In the remainder of this course, we will continue to focus on polynomial differential operators with constant coefficients, that is operators of the form

$$p(\mathcal{D}) = p_n\mathcal{D}^n + p_{n-1}\mathcal{D}^{n-1} + \dots p_1\mathcal{D} + p_0,$$

where all of the coefficients p_k are numbers (opposed to functions of t). Clearly all operators of this form are linear. In addition to being linear operators, they are also *time-invariant* operators, which means:

- If $x(t)$ solves $p(\mathcal{D})x = f(t)$, then $y(t) = x(t - t_0)$ solves $p(\mathcal{D})y = f(t - t_0)$.

All that this says is that delaying the input signal $f(t)$ by t_0 seconds delays the output signal $x(t)$ by t_0 seconds. Another way to put this is that if we know that $x(t)$ is a solution to $p(\mathcal{D})x = f(t)$, we can solve $p(\mathcal{D})y = f(t - t_0)$ by replacing t by $t - t_0$ in $x(t)$. This is a useful property because gives us the solutions to many differential equations for free. In particular, it means that it is sufficient to consider the cases $q(t) = \sin(\Omega t)$ or $q(t) = \cos(\Omega t)$, instead of $q(t) = \sin(\Omega t - \phi)$ or $q(t) = \cos(\Omega t - \phi)$ since the phase shift ϕ can be transformed into a time shift $t_0 = \phi/\Omega$.

10.2 Exponential response formula (ERF)

We would like to find a particular solution of

$$\mathcal{L}z(t) = e^{rt}, \tag{125}$$

which includes Eq. (124) when $r = i\Omega$.

10.2.1 Example: Undamped harmonic resonance

Consider the driven harmonic oscillator

$$\mathcal{L}x = \ddot{x} + \omega^2 x = \cos(\Omega t) \tag{126}$$

The complexified equation is

$$\mathcal{L}z = \ddot{z} + \omega^2 z = e^{rt} \ , \qquad r = i\Omega \tag{127}$$

Using $\mathcal{D} = d/dt$, we can write

$$\mathcal{L} = p(\mathcal{D}) \ , \qquad p(\lambda) = \lambda^2 + \omega^2. \tag{128}$$

Try the solution

$$z(t) = f(r)e^{rt} \tag{129}$$

Then

$$p(\mathcal{D})z = p(\mathcal{D})f(r)e^{rt} = \left(\frac{d^2}{dt^2} + \omega^2\right) f(r)e^{rt} = (r^2 + \omega^2)f(r)e^{rt} = e^{rt} \tag{130}$$

Hence, dividing by e^{rt} and inserting $r = i\Omega$

$$(-\Omega^2 + \omega^2)f = 1 \tag{131}$$

If $\omega \neq \Omega$, then

$$f = \frac{1}{\omega^2 - \Omega^2} = \frac{1}{p(i\Omega)} \tag{132}$$

and

$$z(t) = \frac{e^{i\Omega t}}{\omega^2 - \Omega^2}, \tag{133}$$

and the particular real solution of Eq. (126) is given by the real part

$$x_p(t) = \Re\left[\frac{e^{i\Omega t}}{\omega^2 - \Omega^2}\right] = \frac{\cos(\Omega t)}{\omega^2 - \Omega^2} \tag{134}$$

We see that the solution blows up as $\omega \to \Omega$, that is when $r = i\Omega$ approaches a root of the characteristic polynomial $p(\lambda)$. This phenomenon is called *resonance.* Such resonances can be used to amplify weak input signals, and they can have devastating effects in engineering applications. Equation (134) tells you *that* things blow up as $\omega \to \Omega$, but it doesn't tell you *how* things blow up when $\omega = \Omega$.

To see what happens at $\omega = \Omega$, we cannot use the exponential trial function e^{rt} because Eq. (131) becomes unlovable in this case. So let's try

$$z(t) = g(r)te^{rt} \tag{135}$$

instead. Inserting this into Eq. (127), a calculation analogous to that in Eq. (130), gives

$$\left(\frac{d^2}{dt^2} + \omega^2\right) g(r)te^{rt} = g(r)\left(2r + r^2 t + \omega^2 t\right) e^{rt} = e^{rt} \tag{136}$$

Substituting $r = i\Omega = i\omega$ and dividing by e^{rt}, we find

$$g(r)\left[2i\omega + (i\omega)^2 t + \omega^2 t\right] = 1, \tag{137}$$

so

$$g(r) = \frac{1}{2i\omega} = \frac{1}{p'(i\omega)} \tag{138}$$

and

$$z(t) = \frac{te^{i\omega t}}{2i\omega} \tag{139}$$

The sought-after particular solution is then as before given by the real part

$$x_p(t) = \Re\left[\frac{te^{i\omega t}}{i2\omega}\right] = \Re\left[\frac{t(\cos\omega t + i\sin\omega t)}{i2\omega}\right] = \frac{t\sin(\omega t)}{2\omega}. \tag{140}$$

This is an oscillation with linearly growing amplitude, i.e., we have a linear blow-up.

10.2.2 General formulation

Generally, we want to find a particular solution of

$$p(\mathcal{D})z(t) = e^{rt}\ , \qquad \mathcal{D} = \frac{d}{dt} \tag{141}$$

As in the oscillator example, we try

$$z = f(r)e^{rt},$$

yielding

$$p(\mathcal{D})z = p(\mathcal{D})f(r)e^{rt} = p(r)f(r)e^{rt} = e^{rt}.$$

Hence, if $p(r) \neq 0$, then

$$f(r) = \frac{1}{p(r)}$$

and

$$z(t) = \frac{e^{rt}}{p(r)} \tag{142}$$

solves Eq. (141). This is known as the **Exponential Response Formula**, abbreviated as **ERF** from now on.

10.3 Generalized ERF

The ERF (142) relies on the assumption that we are off-resonance, i.e., that the driving parameter r is *not* a root of the characteristic polynomial $p(\lambda)$. If r is a root of multiplicity k of $p(\lambda)$, then the following generalization of Eq. (142) holds

$$z(t) = \frac{t^k e^{rt}}{p^{(k)}(r)}\,, \qquad p^{(k)}(\lambda) = \frac{d^k}{d\lambda^k} p(\lambda). \tag{143}$$

We next derive this **generalized ERF** for simple roots ($k = 1$); the derivation for repeated roots ($k \geq 2$) works similarly. To solve the equation

$$p(\mathcal{D})z = e^{rt}$$

in the exceptional case when the number r is a root of the characteristic equation, that is, $p(r) = 0$, we first note that

$$p(\mathcal{D})e^{rt} = p(r)e^{rt} = 0,$$

so that the exponential can't work. On the other hand, for $s \neq r$, but very near r, we still have $p(s) \neq 0$, so that by the ordinary ERF,

$$z_s = \frac{e^{st}}{p(s)} \quad \text{solves} \quad p(\mathcal{D})z_s = e^{st}.$$

We would like to pass to the limit $s \to r$ and get a solution to $p(\mathcal{D})z = e^{rt}$ using z_s, but as $s \to r$, we find

$$z_s = \frac{e^{st}}{p(s)} \longrightarrow \frac{e^{rt}}{p(r)} = \frac{e^{rt}}{0},$$

which is undefined. To find an acceptable solution, we can make use of the fact that, if we add c times $p(\mathcal{D})e^{rt} = 0$ to the equation $p(\mathcal{D})z_s = e^{st}$, we see that for any c,

$$z = z_s + ce^{rt} \quad \text{solves} \quad p(\mathcal{D})z = e^{st}.$$

Now we choose the constant c, so that we get well defined limit. Namely, if

$$c = -1/p(s),$$

then

$$z_s + ce^{rt} = \frac{e^{st} - e^{rt}}{p(s)} \longrightarrow \frac{e^{rt} - e^{rt}}{p(r)} = \frac{0}{0}$$

which is an indeterminate form, to which L'Hôpital's rule applies. Indeed,

$$\lim_{s\to r} \frac{e^{st} - e^{rt}}{p(s)} = \lim_{s\to r} \frac{(d/s)[e^{st} - e^{rt}]}{(d/ds)p(s)} = \lim_{s\to r} \frac{te^{st}}{p'(s)} = \frac{te^{rt}}{p'(r)}$$

In summary, we have found that $z = (e^{st} - e^{rt})/p(s)$ solves

$$p(\mathcal{D})z = e^{st}$$

Moreover, taking the limit as $s \to r$, we have

$$z(t) \longrightarrow \frac{te^{rt}}{p'(r)}, \quad e^{st} \longrightarrow e^{rt} \quad \Longrightarrow \quad p(\mathcal{D})\frac{te^{rt}}{p'(r)} = e^{rt}.$$

This is the generalized ERF (143) for $k = 1$, sometimes also denoted as **ERF'**: if $p(r) = 0$, but $p'(r) \neq 0$, then

$$z = \frac{te^{rt}}{p'(r)} \quad \text{solves} \quad p(\mathcal{D})z = e^{rt}.$$

10.4 Complex replacements

Complex replacements are helpful also with other real input signals, with any real-valued function that can be written as the real (or imaginary) part of a reasonably simple complex input signal. Here are some examples:

real input signal	complex replacement
$\cos \Omega t$	$e^{i\Omega t}$
$A\cos(\Omega t - \phi)$	$Ae^{-i\phi}e^{i\Omega t}$
$a\cos \Omega t + b \sin \Omega t$	$(a - bi)e^{i\Omega t}$
$e^{at}\cos \Omega t$	$e^{(a+i\Omega)t}$

10.5 Complex gain, (real)gain, and phase lag for an ODE

Our goal is to explain how the amplitude and phase lag of the system response depend on system parameters and the input frequency. To do so, we will use our method of complex replacement and introduce the *complex gain* first. We use the method of complex replacement to solve the ODE

$$\dot{x} + kx = A\cos(\Omega t). \tag{144}$$

The complex replacement ODE is

$$\dot{z} + kz = Ae^{i\Omega t}$$

with input signal $Ae^{i\Omega t}$. The response determined by ERF is

$$z(t) = \frac{A}{i\Omega + k}\, e^{i\Omega t}.$$

The *complex gain* is the ratio of the complex system response to the complexified system input:

$$G := \quad \frac{\text{complexified system response}}{\text{complexified system input}} \quad = \quad \frac{\frac{A}{i\Omega+k}e^{i\Omega t}}{Ae^{i\Omega t}} \quad = \quad \frac{1}{i\Omega + k}.$$

Observe that G is a complex number that depends on the frequency of the input signal Ω, as well as the system parameter k.

The original ODE has output signal

$$x_p(t) = A\,\Re\left[Ge^{i\Omega t}\right].$$

Working this out we find that[12]

$$G = \frac{1}{i\Omega + k} \quad = \quad \frac{k - i\Omega}{\Omega^2 + k^2}$$

and

$$x_p(t) = A\,\Re\left[Ge^{i\Omega t}\right] = A\left[\frac{k}{\Omega^2+k^2}\cos(\Omega t) + \frac{\Omega}{\Omega^2+k^2}\sin(\Omega t)\right] \tag{145}$$

Observe that the amplitude of the response is different than the amplitude of the input. That difference in amplitude is the *(real) gain* g, which is the magnitude of the complex gain

$$g := |G|. \tag{146}$$

The *phase lag* is

$$\phi = -\arg G. \tag{147}$$

For our example system,

$$\text{gain} = |G| = \frac{1}{|k+i\Omega|} = \frac{1}{\sqrt{k^2+\Omega^2}}$$

and

$$\text{phase lag} = -\arg G = -\arg\frac{1}{k+i\Omega}.$$

A few summarizing remarks

- Given an LTI system with system input $f(t)$, the complexified input is the complex valued function $F(t)$ such that $\Re[F(t)] = f(t)$.
- The complex gain depends only on the system and the input angular frequency (that is, on p and Ω), not on the specific sinusoid used as input.
- The gain and phase lag depend only on the system and the input angular frequency. (This is because gain and phase lag are determined by complex gain.)
- Note that in applications, gain is usually defined in terms of how the input physical variable is related to the output physical variable. That is, if I and O are complexified versions of the input and output, and $O = GI$, with G a complex number, then G is the complex gain.

In actual applications the relationship between the equation and the input-output physical variables can be more complicated than this, so that calculating the gain could end up being more complicated than in the simple example above.

[12] In the MITx mathlet, the complex gain is depicted in the Nyquist Plot. The (real) gain and phase lag are depicted in the Bode Plot. Click on the buttons at the bottom to see these plots in the mathlet.

11 RLC circuits

In the previous sections, we frequently referred to the damped harmonic oscillator model as an example to illustrate general solution strategies for linear ODEs. In this part, we will show that a damped mechanical oscillator is, in fact, mathematically equivalent to the ODE that describes a basic electronic circuit. That is, the results derived earlier for the mechanical oscillator can also be used to understand how an electronic circuit functions!

11.1 Simple series circuit

Let's model a RLC circuit with a voltage source, resistor, inductor, and capacitor attached in series, as shown in Fig. 1A. The relevant variables and functions (with units) are:

t	time (s)
R	resistance of the resistor (ohms)
L	inductance of the inductor (henries)
C	capacitance of the capacitor (farads)
$Q(t)$	charge on the capacitor (coulombs)
$I(t)$	current (amperes)
$V(t)$	voltage source (volts)
$V_R(t)$	voltage drop across the resistor (volts)
$V_L(t)$	voltage drop across the inductor (volts)
$V_C(t)$	voltage drop across the capacitor (volts).

The independent variable is t. The quantities R, L, C are constants parameters characterizing the components of the circuit. Everything else is a function of t. The electric current

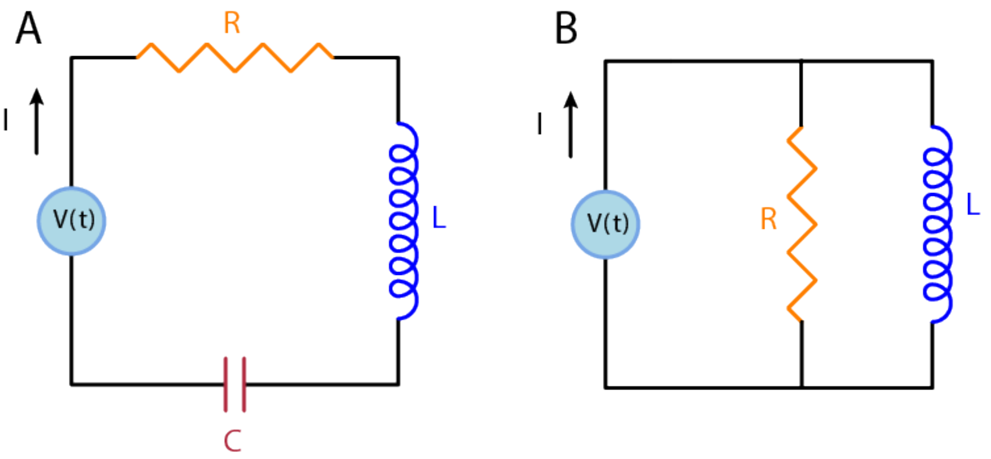

Figure 1: Basic electronic circuits. (A) RLC series circuit: The sum of the voltage drops across the resistor (resistance R), inductor (inductance L) and capacitor (capacitance C) must equal the externally applied voltage $V(t)$, so that $V(t) = V_R(t) + V_L(t) + V_C(t)$. (B) RL parallel circuit: At each junction, ingoing currents must equal outgoing currents $I = I_R + I_L$.

$I(t)$ in a circuit is defined as a time-derivative of the electric charge

$$I(t) := \dot{Q}. \tag{148}$$

From physics, we know that the voltage drops across the different components are given by

$$V_R(t) = R\,I(t) \qquad \text{(Ohm's law)} \tag{149a}$$
$$V_L(t) = L\,\dot{I}(t) \qquad \text{(Faraday's law)} \tag{149b}$$
$$V_C(t) = \frac{1}{C}\,Q(t) \tag{149c}$$

Moreover, *Kirchoff's voltage law* tells us that the sum of the voltage drops across the resistor, inductor and capacitor in Fig. 1A must equal the externally applied voltage $V(t)$,

$$V_R(t) + V_L(t) + V_C(t) = V(t). \tag{150}$$

Inserting the expressions from Eq. (149), the last equation can be written as follows:

$$L\ddot{Q} + R\dot{Q} + \frac{1}{C}Q = V(t). \tag{151}$$

This second-order inhomogeneous linear ODE with unknown function $Q(t)$ is mathematically equivalent to the damped spring-mass-dashpot ODE

$$m\ddot{x} + \gamma\dot{x} + kx = F_{\text{ext}}(t), \tag{152}$$

with the following table of analogies:[13]

Spring-mass-dashpot system		RLC circuit	
displacement	x	Q	charge
velocity	$\dot{x}$	I	current
mass	m	L	inductance
damping constant	γ	R	resistance
spring constant	k	$1/C$	1/capacitance
external force	$F_{\text{ext}}(t)$	$V(t)$	voltage source

To demonstrate the practical usefulness of the ERF, let's consider an applied AC voltage $V = V_0 \cos(\Omega t)$, so that

$$\ddot{Q} + b\dot{Q} + \omega^2 Q = A\cos(\Omega t). \tag{153}$$

where

$$b = \frac{R}{L}\,, \qquad \omega^2 = \frac{1}{LC}\,, \qquad A_0 = \frac{V_0}{L} \tag{154a}$$

[13] Similarly, an undamped driven harmonic oscillator with $b = 0$ is analogous to an LC circuit (no resistor, $R = 0$).

The characteristic polynomial and its roots are

$$p(\lambda) = \lambda^2 + b\lambda + \omega^2 \ , \qquad \lambda_\pm = \frac{1}{2}\left(-b \pm \sqrt{b^2 - 4\omega^2}\right) \tag{155}$$

Lets consider a strongly damped circuit with

$$b > 2\omega \qquad \Leftrightarrow \qquad \frac{R}{L} > \frac{2}{\sqrt{LC}} \qquad \Leftrightarrow \qquad R > 2\sqrt{\frac{L}{C}}.$$

In this case the roots are real and negative, $\lambda_\pm < 0$, and the general solution of Eq. (153) will take the form

$$Q(t) = c_1 e^{\lambda_- t} + c_2 e^{\lambda_+ t} + Q_p(t), \tag{156}$$

where the particular solution is obtained from the ERF formula

$$Q_p(t) = A\Re\left[\frac{e^{i\Omega t}}{p(i\Omega)}\right] = \frac{V_0}{L}\left\{\Re\left[\frac{1}{p(i\Omega)}\right]\cos(\Omega t) - \Im\left[\frac{1}{p(i\Omega)}\right]\sin(\Omega t)\right\}. \tag{157}$$

In particular, we see that for long times $t > \min(\lambda_\pm^{-1})$, *any* solution approaches rapidly the particular solution $Q_p(t)$, because the contributions from the homogeneous solution become exponentially damped. This demonstrates the practical importance of the ERF formula for predicting the asymptotic behavior of driven mechanical and electronic systems!

11.2 Simple parallel circuit

We still briefly discuss a basic parallel circuit, shown in Fig. 1B, with a resistor of resistance R and an inductor of inductance L attached in parallel. A voltage source provides the combination with AC voltage of angular frequency Ω. We would like to find the gain and phase lag of the resistor current relative to the total current through the voltage source.

Let $V(t)$ be the sinusoidal voltage provided. The unknown functions are the resistor current $I_R(t)$, the inductor current $I_L(t)$ and the total current $I(t)$. Physics and *Kirchoff's current law* tell us that

$$V(t) = R\,I_R(t) = L\dot{I}_L(t) \ , \qquad I(t) = I_R(t) + I_L(t). \tag{158}$$

The same relationships hold between the complex replacements,

$$\widetilde{V} = R\,\widetilde{I_R} = L\,\dot{\widetilde{I_L}} \ , \qquad \widetilde{I} = \widetilde{I_R} + \widetilde{I_L}, \tag{159}$$

because taking real parts is compatible with real scalar multiplication and with taking derivatives. Suppose that $\widetilde{V} = e^{i\Omega t}$. (In general, $\widetilde{V} = \gamma e^{i\Omega t}$ for some $\gamma \in \mathbb{C}$, but then everything will be multiplied by γ, so when we take a ratio to get complex gain, γ will disappear.) How do we solve for the other three functions $\widetilde{I_R}$, $\widetilde{I_L}$, $\widetilde{I}$? The ERF suggests steady-state solutions of the form

$$\widetilde{I_L} = \alpha_R e^{i\Omega t} \ , \qquad \widetilde{I_R} = \alpha_L e^{i\Omega t} \ , \qquad \widetilde{I} = \beta e^{i\Omega t}$$

for some unknown complex numbers $\alpha_R, \alpha_L, \beta$. To find $\alpha_R, \alpha_L, \beta$, substitute into the three complex replacement equations (159) and divide by $e^{i\Omega t}$, to get

$$1 = R\alpha_R = L\alpha_L i\Omega \ , \qquad \beta = \alpha_R + \alpha_L.$$

So

$$\alpha_R = \frac{1}{R}\ , \qquad \alpha_L = \frac{1}{iL\Omega}\ , \qquad \beta = \frac{1}{R} + \frac{1}{iL\Omega}.$$

The complex gain of I_R relative to I is the complex constant

$$G = \frac{\widetilde{I_R}}{\widetilde{I}} = \frac{\alpha_R}{\beta} = \frac{1/R}{1/R + 1/(iL\Omega)}. \tag{160}$$

The real gain is $|G|$, and the phase lag is $-\arg G$.

12 Nonlinear first-order ODEs

Thus far, our focus has been primarily on linear ODEs. Many real-world problems, however, are described by nonlinear differential equations. Examples range from gravitational systems, gases and fluids to biological systems and chemical reactions. Nonlinear equations are typically not exactly solvable and therefore must be tackled with approximative analytical or numerical techniques, to extract useful information from them. In this part, we focus on qualitative methods that allow us to infer something about the long-time asymptotic behavior of a nonlinear ODEs. To this end, we consider autonomous first-order equations of the form

$$\dot{x}(t) = f(x(t)) \tag{161}$$

where f is a given nonlinear real-valued function of x. Here, *autonomous* means that f does not explicitly depend on t, and *nonlinear* means that

$$f(\alpha x + \beta y) \neq \alpha f(x) + \beta f(y). \tag{162}$$

Examples of nonlinear functions are x^n with $n \neq 1$, $\exp(\alpha x)$, $\log(x)$, etc. To understand intuitively, why nonlinear ODEs are generally difficult or impossible to solve analytically, let's assume $f(x) \neq 0$ and try separation of variables, which gives

$$\int_{x_0}^{x(t)} \frac{dx}{f(x)} = t - t_0. \tag{163}$$

Clearly, if f is sufficiently complicated, then we cannot solve the integral in closed form and have to resort to other, more qualitative approaches. We will illustrate the key ideas using the logistic equation as an example.

12.1 Logistic equation

The simplest model for a population $x(t)$ assumes that population growth is proportional to the size of the current population. This is modeled by the linear ODE

$$\dot{x} = ax \tag{164}$$

where $a > 0$ is the growth rate. If food or space are scarce, competition will limit the population growth. A simple generalization of Eq. (164) describing such effects is the *logistic model*

$$\dot{x} = ax - bx^2, \tag{165}$$

where the second term on the rhs. accounts for pair interactions. Equation 165 is a nonlinear ODE.

12.2 Solution curves, slope fields and blow-up

The graph $\boldsymbol{\gamma}(t) = (t, x(t))$ of a function $x(t)$ solving Eq. (161), or more generally, a nonlinear non-autonomous ODE

$$\dot{x} = f(t, x) \tag{166}$$

is called a solution curve. The vector field

$$\dot{\boldsymbol{\gamma}} = \begin{pmatrix} 1 \\ \dot{x} \end{pmatrix} = \begin{pmatrix} 1 \\ f(t, x) \end{pmatrix} \tag{167}$$

is called the *slope field*. *Isoclines*[14] are curves in the (t, x) plane that have constant slope $\dot{x} = f = C$, where C is a constant. As an example, let's consider the logistic equation with $a = 0$ and $b = -1$,

$$\dot{x} = x^2. \tag{168}$$

In this case, $f(t, x) = x^2$, and the slope field is

$$\dot{\boldsymbol{\gamma}} = \begin{pmatrix} 1 \\ x^2 \end{pmatrix}, \tag{169}$$

which is depicted in Fig. 2. The nonlinear ODE (168) is in fact still exactly solvable. One solution is the constant function $x(t) \equiv 0$ (yellow curve in Fig. 2). To find non-constant solutions with $x(t) \neq 0$, we use separation of variables [Eq. (163) with $f(x) = x^2$], yielding

$$\int_{x_0}^{x(t)} \frac{dx}{x^2} = \frac{-1}{x(t)} - \left(\frac{-1}{x_0}\right) = t - t_0, \tag{170}$$

where $x_0 = x(t_0)$. Hence,

$$x(t) = \left[\frac{1}{x_0} - (t - t_0)\right]^{-1} \tag{171}$$

[14] Greek: isos=same, klisi=slope.

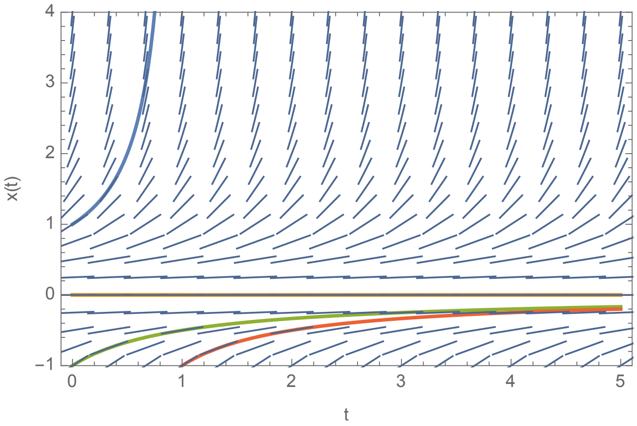

Figure 2: Slope field $\dot{\gamma}$ and solutions for Eq. (168). The solution with $x_0 = x(0) = -1$ (green line) approaches the hyperbola $-1/t$ (orange) at large times $t \to \infty$. The constant solution $x(t) \equiv 0$ (yellow) exists for all times, whereas the solution with $x_0 = 1$ (blue) blows up in finite time at $t = 1$.

Let's assume $t_0 = 0$, then

$$x(t) = \frac{x_0}{1 - x_0 t}. \tag{172}$$

For initial conditions $x_0 < 0$, the denominator is always positive for $t > 0$ and for $t \gg 1/|x_0|$ the solutions approach the same hyperbola (orange curve in Fig. 2),

$$x(t) \simeq \frac{-1}{t}. \tag{173}$$

By contrast, for $x_0 > 0$, the denominator becomes zero when $t = 1/x_0$; that is, the solution *blows up* in finite time.

12.3 Fixed points and linear stability analysis

Fixed points (FPs) of are constant solutions $x(t) \equiv x_*$, which satisfy $\dot{x}_* = 0$ and, hence, the FP criterion[15]

$$f(x_*) = 0. \tag{174}$$

Such FPs, if they exist, correspond to 0-isoclines (curves of slope zero) in the (t, x)-plane. FPs are extremely useful for understanding the asymptotic behavior of solutions $x(t)$ of a given nonlinear ODE. In fact, whenever you encounter a nonlinear ODE, the first thing you should do is to look for FPs and study their *stability*. To see how this is done in practice, let's consider the logistic equation (165) with $a = b = 1$,

$$\dot{x} = x - x^2 = f(x). \tag{175}$$

[15] Given a non-autonomous ODE $\dot{x} = f(t, x)$, we call curves $x_*(t)$ satisfying $f(t, x_*) = 0$ also *critical curves* or *critical points*. According to this definition, fixed points are exactly the critical points of autonomous ODEs.

According to the criterion (174), the FPs of Eq. (175) satisfy

$$x_* - x_*^2 = 0, \tag{176}$$

implying that there are two of them

$$x_0 = 0\ , \qquad x_1 = 1. \tag{177}$$

To understand how solutions behave in the vicinity of a FP x_*, let's consider 'nearby' solutions of the form

$$x(t) = x_* + \epsilon(t), \tag{178}$$

where $\epsilon(t)$ is small. We first note that

$$\dot{x} = \dot{\epsilon}. \tag{179}$$

Moreover, we can Taylor expand

$$\begin{aligned} f(x) = f(x_* + \epsilon) &= f(x_*) + f'(x_*)\epsilon + \frac{1}{2}f''(x_*)\epsilon^2 + \ldots \\ &= 0 + f'(x_*)\epsilon + \ldots \end{aligned} \tag{180}$$

Here, we have used that $f(x_*) = 0$ for FPs x_*. Keeping only terms up to linear order in ϵ, we can approximate the original nonlinear ODE $\dot{x} = f(x)$ by the linear ODE

$$\dot{\epsilon} = f'(x_*)\epsilon. \tag{181a}$$

Since $f'(x_*)$ is a constant, this linearized equation has the exponential solution

$$\epsilon(t) = \epsilon(0)\, e^{f'(x_*)t} \qquad \Rightarrow \qquad x(t) = x_* + \epsilon(0)\, e^{f'(x_*)t}. \tag{181b}$$

This result shows that, for $f'(x_*) > 0$, any solution starting near x_* will diverge exponentially from x_*; in this case, the FP x_* is called *linearly unstable.* On the other hand, when $f'(x_*) < 0$, any solution starting near x_* will converge towards x_*; in this case, the FP x_* is called *linearly stable.*[16] Equations (181) demonstrate why understanding linear ODEs is important for understanding nonlinear ODEs.

Let's check the stability of the FPs (177) of the logistic equation (175). To Taylor expand $f(x) = x - x^2$, we note that

$$f'(x) = 1 - 2x$$

so that

$$f'(x_0) = f'(0) = 1 > 0\ , \qquad f'(x_1) = f'(1) = -1 < 0. \tag{182}$$

Consequently the FP x_0 is linearly unstable, whereas $x_1 = 1$ is linearly stable. Assuming that the initial population size is positive, $x(0) > 0$, this means that any solution will converge to the stable FP $x_1 = 1$; see Fig. 3. Thus, even though we do not know the exact time-dependent solution, the FP analysis reveals the stationary behavior of the biological system described by the nonlinear ODE (175).

[16] If $f'(x_*) = 0$, then the quadratic term of the Taylor expansion has to be considered.

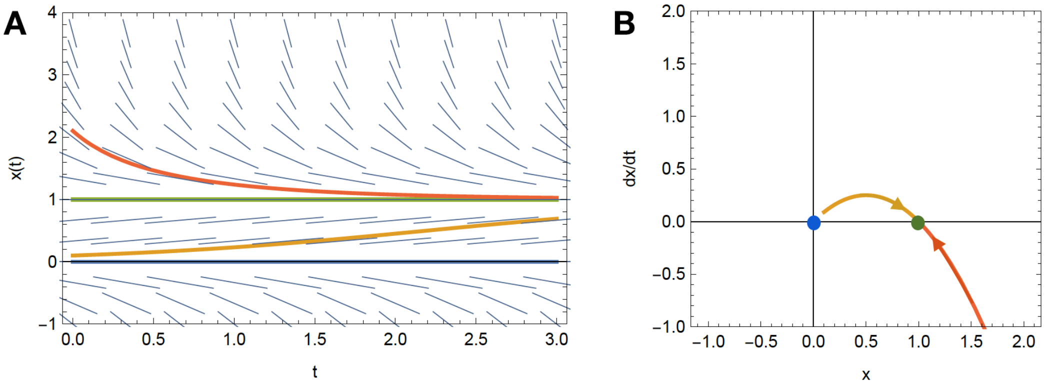

Figure 3: (A) Slope field and numerical solutions of the logistic equation (175) in the (t, x)-plane. Linearly stable and unstable fixed points correspond to the green and blue lines (0-isoclines) respectively. The red and yellow solution curves converge to the stable fixed point $x_1 = 1$. (B) Phase portrait in the $(x, \dot{x})$-plane for the solutions in panel A.

12.4 Bifurcation diagrams

In the previous section, we have seen that linear stability analysis of FPs x_* offers qualitative understanding of nonlinear autonomous ODEs. If the ODE features a number parameters $\boldsymbol{b} = (b_1, b_2, \ldots)$,

$$\dot{x} = f(x; \boldsymbol{b}), \tag{183}$$

then the FPs become functions of these parameters, $\boldsymbol{x}_*(\boldsymbol{b})$ and their stability can change depending on the parameter values. To illustrate this, let's consider the logistic equation with harvesting

$$\dot{x} = 3x - x^2 - h =: f(x; h), \tag{184}$$

where $h > 0$ is the harvesting rate. To find FPs, we have to find the real roots of $f(x; h) = 3x - x^2 - h$. Solving

$$0 = 3x_* - x_*^2 - h \tag{185}$$

gives

$$x_*^{\pm}(h) = \frac{1}{2}\left(3 \pm \sqrt{9 - 4h}\right) \tag{186}$$

Obviously these two FPs only exist when

$$h \leq \frac{9}{4}. \tag{187}$$

At the critical value $h_c = 9/4$, the two FPs meet at the value $x_*(h_c) = 3/2$. To analyze the stability of $x_*^{\pm}(h)$ when $0 < h < 9/4$, we note that

$$f'(x; h) = \frac{df}{dx} = 3 - 2x, \tag{188}$$

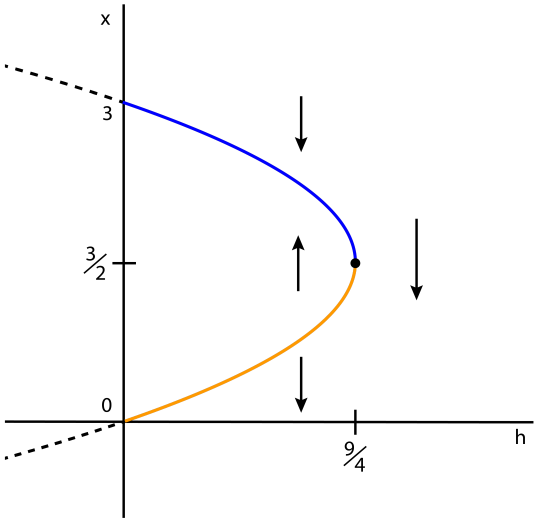

Figure 4: Bifurcation diagram for the logistic equation with harvesting, Eq. (184), in the (h, x)-plane. Colored curves show the fixed points $x_*^{\pm}(h)$ from Eq. (186). Arrows indicate stability through the flow directions of the solutions $x(t; h)$.

so that

$$f'(x_*^-; h) = \sqrt{9 - 4h} > 0 \ , \qquad f'(x_*^+; h) = -\sqrt{9 - 4h} < 0. \tag{189}$$

This means that $x_*^+(h)$ is stable whereas $x_*^-(h)$ is unstable. The critical FP $x_*(h_c) = 3/2$ is called *semi-stable*.

The information about the parameter dependence of FPs is typically summarized in a bifurcation diagram, as shown in Fig. 4. Unstable FPs (orange curve in Fig. 4) is also referred to as *separatrix*, because they separate qualitatively different solutions. In our example, a population described by Eq. (184) with $h < 9/4$ goes extinct only when

$$x(t = 0) < x_*^-(h).$$

For $x(t = 0) < x_*^-(h)$, the population will always stabilize at $x_*^+(h)$ after a sufficiently long time $t \to \infty$. By contrast, when $h > 9/4$, the population goes always extinct.

13 Numerical solution of first-order ODEs

Highly nonlinear ODEs describing real-world problems cannot be solved by hand in most cases. One then has to resort to numerical solution techniques. The basic idea of numerical solvers is to discretize a given ODE efficiently, so that numerical solutions converge rapidly as the time-step is decreased. In general, one can discretize ODEs in many different ways. In this part, we illustrate the general approach for the nonlinear first-order ODE

$$\dot{x}(t) = f(t, x(t)) \tag{190}$$

with initial condition $x(0) = x_0$ The algorithms sketched below can be generalized to more complicated ODEs in a fairly straightforward manner.

13.1 Euler's method

Euler's method is simplest algorithms for solving an ODE by discretizing time t into equidistant intervals Δt by considering discrete time-points

$$t_n = n\,\Delta t\ , \qquad n = 0, 1, \ldots .$$

The goal is to compute the values of the solutions at these time-points, $x_n = x(t_n)$ where $x_0 = x(0)$ is the given initial condition.

The time derivative on the lhs. of Eq. (190) can be approximated by the forward difference quotient

$$\dot{x}(t_n) \approx \frac{x(t_{n+1}) - x(t_n)}{\Delta t} = \frac{x_{n+1} - x_n}{\Delta t}, \tag{191}$$

while the rhs. of Eq. (190) is given by

$$f_n = f(t_n, x(t_n)) = f(t_n, x_n). \tag{192}$$

The discretized version of the ODE (190) thus becomes

$$\frac{x_{n+1} - x_n}{\Delta t} = f(t_n, x_n). \tag{193}$$

This can be solved for x_{n+1}, yielding the first-order recursion relation

$$x_{n+1} = x_n + f(t_n, x_n)\,\Delta t. \tag{194}$$

Let's write this explicitly for the first few steps

$$\begin{array}{lrcl} t_0 = 0\Delta t: & x_0 & = & x(0) \\ t_1 = 1\Delta t: & x_1 & = & x_0 + f(0, x_0)\,\Delta t \\ t_2 = 2\Delta t: & x_2 & = & x_1 + f(t_1, x_1)\,\Delta t \\ t_3 = 3\Delta t: & x_3 & = & x_2 + f(t_2, x_2)\,\Delta t \end{array}$$

and so on. A graph of the approximate numerical solution is then obtained by plotting the points $(0, x_0)$, (t_1, x_1), (t_2, x_2), etc.

13.2 Reliability checks

To test, whether or not a numerical algorithm converges to the correct solution, one can check certain heuristic reliability criteria[17]. These include:

- **Self-consistency:** Solution curves should not cross! If numerically computed solution curves appear to cross, a smaller step size Δt is needed.[18]

[17] 'Heuristic' means that these tests seem to work in practice, but they are not proved to work always.

[18] Try the mathlet *Euler's Method* with $\dot{y} = y^2 - t$, step size 1, and starting points $(0, 0)$ and $(0, 1/2)$.

- **Convergence as $\Delta \to 0$:** The estimate for $y(t)$ at a fixed later time t should converge to the true value as $\Delta \to 0$. If shrinking h causes the estimate to change a lot, then Δt is probably not small enough yet.
- **Structural stability:** If small changes in the ODE's parameters or initial conditions change the outcome completely, the answer probably should not be trusted. However, one should also keep in mind that one reason for this could be a separatrix, a curve such that nearby starting points on different sides lead to qualitatively different outcomes; this is not a fault of the numerical method, but is an instability in the answer nevertheless.

In general, the quality of numerically determined solution depends on the function f and on the details of the numerical method. In the next section, we still look at a slightly more sophisticated algorithm that produces better results than the Euler scheme.

13.3 Runge-Kutta scheme

When computing $\int_a^b f(t)\,dt$ numerically, the most primitive method is to use the left Riemann sum: divide the range of integration into subintervals of width Δt, and estimate the value of $f(t)$ on each subinterval as being the value at the left endpoint. More sophisticated methods are the *trapezoid rule* and *Simpson's rule*, which have smaller errors. There are analogous improvements to Euler's method; see Table 1.

The widely applied Runge-Kutta methods "look ahead" to get a better estimate of what happens to the slope over the course of the interval $[t_0, t_0 + \Delta t]$. Here is how one step of the *second-order Runge-Kutta method (RK2)* goes:

1. Starting from (t_0, x_0), look ahead to see where one step of Euler's method would land, say (t_1, x_1), but do not go there!
2. Instead sample the slope at the *midpoint* $\left(\frac{t_0+t_1}{2}, \frac{x_0+x_1}{2}\right)$.
3. Now move along the segment of *that* slope: the new point is

$$\left(t_0 + \Delta t, x_0 + f\left(\frac{t_0+t_1}{2}, \frac{x_0+x_1}{2}\right)\Delta t\right).$$

Repeat, reassessing the slope after each step. RK2 is also called *midpoint Euler*.

Integration	Differential equation	Error
left Riemann sum	Euler's method	$O(\Delta t)$
trapezoid rule	second-order Runge-Kutta method (RK2)	$O(\Delta t^2)$
Simpson's rule	fourth-order Runge-Kutta method (RK4)	$O(\Delta t^4)$

Table 1: The big-O notation $O(\Delta t^4)$ means that there is a constant C (depending on everything except for Δt) such that the error is at most $C\Delta t^4$, assuming that Δt is small. The error estimates in the table are valid for reasonable functions.

The *fourth-order Runge-Kutta method (RK4)* is similar, but more elaborate, averaging several slopes. It is probably the most commonly used method for solving ODEs numerically. Some people simply call it *the* Runge-Kutta method. The mathlets use RK4 with a small step size to compute the 'actual' solution to an ODE.

14 Introduction to ODE systems

So far we focuses on ODEs for a single scalar function, such as $x(t)$ or $y(x)$. However, any time a quantity and its derivative depend not just on themselves but also on other quantities, we get a system of DEs. Below are some examples.

- **Physics:** To navigate the Juno spacecraft to Jupiter, NASA scientists needed to understand gravity. Like many other important quantities in nature, gravity is governed by ODEs that constrain vector-valued functions; that is, by a system of differential equations.
- **Electrical systems:** We have seen that an RLC circuit can be described by a single second order ODE. More complex circuits with numerous loops, present in computers, cell phones, amplifiers, require systems of DEs to describe.
- **Chemical reactions:** The time derivative of the concentration of a chemical species is a function of the reaction rates and concentrations of other species. These relations give systems of DEs.
- **Engineering:** A simulation of a car crash test is a system that comprises millions of masses, springs, and dashpots. Simulation of cloth uses a 2D networks of springs. Simulations of fire, smoke, water, snow in movies are all systems of DEs.
- **Gene expression network:** The time derivatives of mRNA and protein concentrations are functions of their degradation rates and transcription factors of the proteins. These relations constitute systems of ODEs.
- **Network dynamics:** How a virus spreads in a population, the flow of traffic, and information flow in a social network can all be described by systems of ODEs.

14.1 Brewery example

Let's consider the basic two-tank system in Fig. 5, to demonstrate how we can construct ODE systems for specific physical, chemical or biological systems. Let $Q_{1\to 2}$ be the flow rate (fluid volume per time) moving from tank 1 to tank 2. Then

$$Q_{1\to 2} \propto (h_1 - h_2)$$

or equivalently

$$Q_{1\to 2} = c(h_1 - h_2),$$

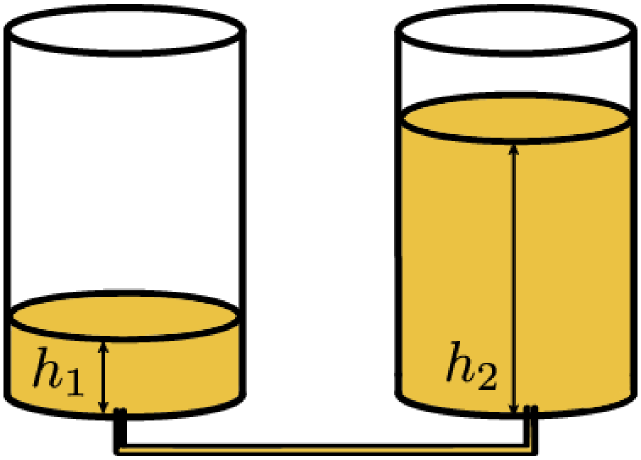

Figure 5: If the pipes have a small diameter relative to the overall size of the tanks, we can model the flow linearly, and the flow rate will be proportional to the difference in fluid height.

where the proportionality constant c encodes pipe geometry (diameter, etc.). The change in fluid volume in tank 1 during the small time-interval Δt is then

$$\Delta V_1 = A\Delta h_1 = Q_{1\to 2}\Delta t.$$

Dividing by Δt and taking the limit $\Delta t \to 0$, we find

$$A\frac{dh_1}{dt} = c(h_2 - h_1). \tag{195a}$$

Analogously, assuming the two tanks have the same geometry, we find for tank 2

$$A\frac{dh_2}{dt} = c(h_1 - h_2). \tag{195b}$$

Defining a new constant $a = c/A$, we have the ODE system

$$\begin{aligned} \dot{h}_1 &= a(h_2 - h_1), && (196a) \\ \dot{h}_2 &= a(h_1 - h_2). && (196b) \end{aligned}$$

Note that this system has a conserved quantity $h(t) = h_1(t) + h_2(t)$. To check this explicitly, let's verify

$$\dot{h} = \dot{h}_1 + \dot{h}_2 = a(h_2 - h_1) + a(h_1 - h_2) = 0. \tag{197a}$$

That is if $h_1(0)$ and $h_2(0)$ are given, then

$$h(t) = h_1(0) + h_2(0) \tag{197b}$$

at all times t.

The ODE system (196) is a *linear* system. To see this, let's define the vector of functions $\boldsymbol{h}(t) = (h_1(t), h_2(t))$ and the constant matrix

$$A = \begin{pmatrix} -a & a \\ a & -a \end{pmatrix}.$$

We can then rewrite the ODE system (196) in the form

$$\dot{\boldsymbol{h}} = \begin{pmatrix} \dot{h}_1 \\ \dot{h}_2 \end{pmatrix} = A\boldsymbol{h}. \tag{198}$$

14.2 Lotka-Volterra system

This model is standard model in ecology and describes a simple predator-prey dynamics. The prey population $u(t)$ and the predator population $v(t)$ are governed by

$$\begin{aligned} \dot{u} &= Au - Buv, && \text{(199a)} \\ \dot{v} &= -Cv + Euv && \text{(199b)} \end{aligned}$$

with positive rate (coupling) parameters $A, B, C, E > 0$. The model is a *nonlinear* ODE system and has two fixed points

$$(u_0, v_0) = (0, 0)\ , \qquad (u_*, v_*) = (C/E, A/B). \tag{200}$$

Later in class, we will learn how to analyze the stability of such fixed points.

14.3 Newton's equations and companion matrix

Any nth order ODE for a function $x(t)$ with can be rewritten as ODE system. The trick is to interpret the derivatives $\dot{x}$, $\ddot{x}$, $\dddot{x}, \ldots$ as independent variables. Let's illustrate this for the damped oscillator system, described by Newton's law

$$m\ddot{x} = -\gamma\dot{x} - kx. \tag{201}$$

Defining *momentum*

$$p = mv = m\dot{x} \tag{202}$$

we can rewrite Eq. (201) as a system for the two-component vector of functions $\boldsymbol{\xi} = (x, p)$,

$$\begin{aligned} \dot{x} &= \frac{1}{m}p, && \text{(203a)} \\ \dot{p} &= -\frac{\gamma}{m}p - k\,x. && \text{(203b)} \end{aligned}$$

Let us consider the function

$$H(x, p) = \frac{p^2}{2m} + \frac{kx^2}{2} = E_{\text{kin}} + E_{\text{pot}} \tag{204}$$

and compute its time-derivative

$$\begin{aligned}\frac{d}{dt}H &= \frac{d}{dt}\left(\frac{p^2}{2m}\right)+\frac{d}{dt}\left(\frac{kx^2}{2}\right)\\ &= \left(\frac{p}{m}\right)\dot p+(kx)\,\dot x\\ &= \left(\frac{p}{m}\right)\left(-\frac{\gamma}{m}\,p-k\,x\right)+(kx)\left(\frac{1}{m}\,p\right) \;=\; -\frac{\gamma}{m^2}\,p^2.\end{aligned}$$

That is, when there is no friction, $\gamma = 0$, then H is conserved in time and fixed by the initial conditions

$$H(x,p)=\frac{p(t)^2}{2m}+\frac{kx(t)^2}{2}=\frac{p(0)^2}{2m}+\frac{kx(0)^2}{2}=:E_0\ ,\qquad \forall\, t>0. \tag{205}$$

H is called the total *energy* (or also the Hamiltonian) of the harmonic oscillator. If there is damping, $\gamma > 0$, then the energy H decreases and the oscillator will eventually come to rest.

Using the two-component vector of functions $\boldsymbol{\xi} = (x, p)$, we can rewrite Eq. (203) in matrix form

$$\dot{\boldsymbol{\xi}} = A\boldsymbol{\xi} \tag{206a}$$

with the *companion matrix*

$$A=\begin{pmatrix}0 & \frac{1}{m}\\ -k & -\frac{\gamma}{m}\end{pmatrix}, \tag{206b}$$

because

$$\begin{pmatrix}\dot x\\ \dot p\end{pmatrix}=\begin{pmatrix}0 & \frac{1}{m}\\ -k & -\frac{\gamma}{m}\end{pmatrix}\begin{pmatrix}x\\ p\end{pmatrix}=\begin{pmatrix}\frac{1}{m}p\\ -kx-\frac{\gamma}{m}p\end{pmatrix}$$

is the same as (203). We rewrote (201) as a system using the momentum variable $p = mv = m\dot x$, which is the common approach in physics, because momentum is more fundamental[19] than velocity. Of course, from a purely mathematical perspective, we could have also used x and v instead – you may try this yourself as an exercise.

14.4 General companion systems

The above procedure allows us to rewrite any linear nth order inhomogeneous ODE of the form

$$a_n x^{(n)} + a_{n-1}x^{(n-1)} + \ldots + a_1\dot x + a_0 x = q(t) \tag{207}$$

as linear first-order system. To see this explicitly, let's define n functions $z_i(t)$ by

$$z_1(t) = x\ ,\qquad z_2(t) = \dot x\ ,\qquad z_3(t) = \ddot x\ ,\qquad \ldots\qquad z_n(t) = x^{(n-1)}.$$

[19] Mostly, because momentum is conserved in elastic collisions whereas velocity isn't.

Noting that

$$z_{i+1}(t) = \dot{z}_i(t) \ , \qquad\qquad i = 1, \ldots, n-1$$

and that

$$\dot{z}_n = -\frac{a_0}{a_n} z_1 - \frac{a_1}{a_n} z_2 - \ldots - \frac{a_{n-2}}{a_n} z_{n-1} - \frac{a_{n-1}}{a_n} z_n + \frac{1}{a_n} q(t), \tag{208}$$

we can rewrite (207) as

$$\begin{pmatrix} \dot{z}_1 \\ \dot{z}_2 \\ \vdots \\ \dot{z}_{n-1} \\ \dot{z}_n \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \\ -\frac{a_0}{a_n} & -\frac{a_1}{a_n} & -\frac{a_2}{a_n} & \ldots & -\frac{a_{n-1}}{a_n} \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_{n-1} \\ z_n \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \frac{q(t)}{a_n} \end{pmatrix} . \tag{209}$$

This can be compactly expressed as

$$\dot{\boldsymbol{z}}(t) = A\boldsymbol{z} + \boldsymbol{b}(t) \tag{210}$$

with A being the companion matrix and vector $\boldsymbol{b}(t)$ representing the inhomogeneity.

More generally, if we are given a nonlinear nth order ODE of the form

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

then we can always rewrite this as nonlinear first-order system

$$\begin{pmatrix} \dot{z}_1 \\ \dot{z}_2 \\ \vdots \\ \dot{z}_{n-1} \\ \dot{z}_n \end{pmatrix} = \begin{pmatrix} z_2 \\ z_3 \\ \vdots \\ z_n \\ F(t, z_1, z_2, \ldots, z_n). \end{pmatrix}$$

This form is often used to represent ODEs in numerical solvers.

15 Solving linear 2×2 homogeneous systems

The two-dimensional examples (the coupled tanks and the damped oscillator) discussed above are of the form

$$\dot{x} = ax + by, \tag{211a}$$
$$\dot{y} = cx + dy. \tag{211b}$$

Defining $\boldsymbol{\xi}(t) = (x(t), y(t))$, we can rewrite this as a matrix equation

$$\dot{\boldsymbol{\xi}} = A\boldsymbol{\xi} \ , \qquad A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} . \tag{212}$$

We would like to use this 2×2 case to demonstrate the general solution procedure for linear first-order systems.

15.1 Exponential ansatz, eigenvectors and eigenvalues

To solve an ODE system of the form (212) for a constant matrix A, we can try the exponential ansatz

$$\boldsymbol{\xi} = e^{\lambda t}\boldsymbol{v}$$

where the constant vector $\boldsymbol{v}$ and λ are to be determined. Using

$$\dot{\boldsymbol{\xi}} = \lambda e^{\lambda t}\boldsymbol{v},$$

inserting into (212) and dividing by $e^{\lambda t}$ gives

$$\lambda \boldsymbol{v} = A\boldsymbol{v} \qquad \Leftrightarrow \qquad \mathbf{0} = (A - \lambda I)\boldsymbol{v} \tag{213}$$

where I is the identity matrix. We thus see that $e^{\lambda t}\boldsymbol{v}$ is a solution if λ is an eigenvalue of A and $\boldsymbol{v} = (v_1, v_2)$ the corresponding eigenvector of A, as introduced in Sec. 9.1. Let's write the last equation explicitly

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} a-\lambda & b \\ c & d-\lambda \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \tag{214}$$

This is a linear system for the two unknown eigenvector components (v_1, v_2). We immediately see that $(v_1, v_2) = (0, 0)$ is a trivial solution. For other non-trivial solutions to exist, the determinant of the coefficient matrix

$$M(\lambda) = \begin{pmatrix} a-\lambda & b \\ c & d-\lambda \end{pmatrix} \tag{215}$$

must vanish – otherwise, as we know from linear algebra, $(v_1, v_2) = (0, 0)$ would be the only solution. The requirement of a zero-determinant means that

$$p(\lambda) := \det M(\lambda) = (a-\lambda)(d-\lambda) - bc \overset{!}{=} 0. \tag{216}$$

Here, $p(\lambda)$ is the characteristic polynomial of A. The eigenvalues are the two roots λ_1 and λ_2 of this polynomial.[20] We list two important facts about eigenvalues:

(i) Let's assume all eigenvalues are distinct. If $\boldsymbol{v}_1$ is an eigenvector of, say, eigenvalue λ_1 then $\boldsymbol{v}_1' = \alpha \boldsymbol{v}_1$ is also an eigenvector of λ_1 for any real α. This means that we can always normalize $\boldsymbol{v}_1$ to length 1 and that the eigenvectors form a 1-dimensional (sub-)vector space.

(ii) Eigenvectors associated with different eigenvalues are linearly independent. This means that the set of two eigenvectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ associated with λ_1 and λ_2 respectively, forms a basis of $\mathbb{R}^2$.

It then follows that, when $\lambda_1 \neq \lambda_2$, the general solution of the linear Eq. (212) can be expressed as a superposition

$$\boldsymbol{\xi}(t) = C_1 e^{\lambda_1 t}\boldsymbol{v}_1 + C_2 e^{\lambda_2 t}\boldsymbol{v}_2, \tag{217}$$

where C_1 and C_2 are determined by the initial conditions. This is pretty much the same solution formula that we had for homogeneous nth order system, with the slight difference that we now need to multiply each term $e^{\lambda_i t}$ by the corresponding eigenvector $\boldsymbol{v}_i$.

[20] For any upper or lower triangular square matrix, the eigenvalues are simply the entries on the diagonal.

15.2 Brewery example continued

To illustrate the solution procedure step-by-step, let's return to the brewery example and assume $a = 1$ for simplicity. Then the dynamics of the height levels in the two tanks is described by the system

$$\begin{pmatrix} \dot{h}_1 \\ \dot{h}_2 \end{pmatrix} = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}. \tag{218}$$

We would like to solve this equation for the initial condition

$$\begin{pmatrix} h_1(0) \\ h_2(0) \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \tag{219}$$

corresponding to the second tank being initially empty.

The characteristic polynomial of the coefficient matrix is

$$p(\lambda) = \det \begin{pmatrix} -1-\lambda & 1 \\ 1 & -1-\lambda \end{pmatrix} = (-1-\lambda)^2 - 1 = \lambda(2+\lambda). \tag{220}$$

The eigenvalues are therefore $\lambda_1 = 0$ and $\lambda_2 = -2$, and the solution can be written as

$$\begin{pmatrix} h_1 \\ h_2 \end{pmatrix} = C_1 e^{0t} \boldsymbol{v}_1 + C_2 e^{-2t} \boldsymbol{v}_2. \tag{221}$$

From this general form we can already see that at large times $t \to \infty$, the stationary solution will be proportional to $\boldsymbol{v}_1$, as the second term decays.

To fully specify the solution, we still need to find the eigenvectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$. To find the components of $\boldsymbol{v}_1$, we must solve

$$\begin{pmatrix} -1-\lambda_1 & 1 \\ 1 & -1-\lambda_1 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \mathbf{0}. \tag{222}$$

It is not difficult to see that

$$\boldsymbol{v}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{223}$$

is an eigenvector of $\lambda_1 = 0$. This make sense and we could have guessed this without calculation, since it is intuitively clear that the tanks should approach equal filling levels in the long-time limit.

To find the components of $\boldsymbol{v}_2$, we must solve

$$\begin{pmatrix} -1-\lambda_2 & 1 \\ 1 & -1-\lambda_2 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \mathbf{0}. \tag{224}$$

We see that

$$\boldsymbol{v}_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \tag{225}$$

is an eigenvector for $\lambda_2 = -2$.

The general solution can therefore be written as

$$\begin{pmatrix} h_1(t) \\ h_2(t) \end{pmatrix} = C_1 \begin{pmatrix} 1 \\ 1 \end{pmatrix} + C_2 \begin{pmatrix} 1 \\ -1 \end{pmatrix} e^{-2t} \tag{226}$$

The constants C_1 and C_2 need to be determined from the initial conditions. At time $t = 0$

$$\begin{pmatrix} h_1(0) \\ h_2(0) \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} = C_1 \begin{pmatrix} 1 \\ 1 \end{pmatrix} + C_2 \begin{pmatrix} 1 \\ -1 \end{pmatrix}. \tag{227}$$

The second line implies $C_1 = C_2$ and the first $C_1 = C_2 = 1/2$, so that the time-dependent solution is given by

$$\begin{pmatrix} h_1(t) \\ h_2(t) \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \frac{1}{2} \begin{pmatrix} 1 \\ -1 \end{pmatrix} e^{-2t} = \frac{1}{2} \begin{pmatrix} 1 + e^{-2t} \\ 1 - e^{-2t} \end{pmatrix}. \tag{228}$$

The same sequence of steps can be applied to find the general solution of $n \times n$ first-order ODE systems with constant coefficients.

15.3 Phase portrait revisited

As an example, we again consider the damped harmonic oscillator

$$\ddot{x} + b\dot{x} + kx = 0, \tag{229}$$

where we have set the mass $m = 1$ to simplify subsequent formulas a bit.[21] Setting $y(t) = \dot{x}(t)$, the companion system is given by

$$\frac{d}{dt} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -k & -b \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \tag{230}$$

The characteristic polynomial of the companion matrix is

$$p(\lambda) = \det \begin{pmatrix} -\lambda & 1 \\ -k & -b - \lambda \end{pmatrix} = \lambda(b + \lambda) + k = \lambda^2 + b\lambda + k \tag{231}$$

This is the same characteristic polynomial that we found earlier in Eq. (60). This illustrates the general fact that the characteristic polynomial of an LTI operator $p(D)$ is the same (up to a constant multiple) as that of its companion matrix.

The eigenvalues (roots of p) are then given by

$$\lambda_{\pm} = -\frac{b}{2} \pm \frac{1}{2}\sqrt{b^2 - 4k}. \tag{232}$$

and, if $\lambda_+ \neq \lambda_-$, the general solution can be written as

$$\begin{pmatrix} x \\ y \end{pmatrix} = C_+ e^{\lambda_+ t} \boldsymbol{v}_+ + C_- e^{\lambda_- t} \boldsymbol{v}_- \tag{233}$$

[21] To obtain the results for arbitrary m, simply replace $b \to b/m$ and $k \to k/m$ everywhere.

where $\boldsymbol{v}_\pm$ are eigenvectors of $\lambda_\pm$, which satisfy

$$\begin{pmatrix} -\lambda_\pm & 1 \\ -k & -b-\lambda_\pm \end{pmatrix} \boldsymbol{v}_\pm = \mathbf{0}. \tag{234}$$

We only need to solve one of the two equations in the system (since otherwise there is only the trivial solution and we know that is not true). We choose the first and easier equation to solve and get

$$\boldsymbol{v}_\pm = \begin{pmatrix} 1 \\ \lambda_\pm \end{pmatrix}. \tag{235}$$

We know from our earlier discussion in Sec. 5.2 that, depending on the values of b and k, the eigenvalues can be real (overdamped case), purely imaginary (undamped case) or complex (underdamped case). We illustrate these three cases again in 'eigenvector language'.

15.3.1 Overdamped case $b^2 > 4k$

In this case, the eigenvalues $\lambda_\pm$ are real and negative so that the solutions asymptotically decay

$$\begin{pmatrix} x \\ y \end{pmatrix} = C_+ e^{\lambda_+ t} \begin{pmatrix} 1 \\ \lambda_+ \end{pmatrix} + C_- e^{\lambda_- t} \begin{pmatrix} 1 \\ \lambda_- \end{pmatrix} \qquad \to \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \text{as} \quad t \to \infty, \tag{236}$$

see blue curve in Fig. 6.

15.3.2 Undamped case $b^2 = 0$

In this case, the eigenvalues are purely imaginary

$$\lambda_\pm = \pm\frac{1}{2}\sqrt{-4k} = \pm i\sqrt{k} =: \pm i\omega, \tag{237}$$

and the solution is given by

$$\begin{pmatrix} x \\ y \end{pmatrix} = c_+ e^{i\omega t} \begin{pmatrix} 1 \\ i\omega \end{pmatrix} + c_- e^{-i\omega t} \begin{pmatrix} 1 \\ -i\omega \end{pmatrix}. \tag{238}$$

The coefficients $c_\pm$ are complex and satisfy $c_+ = \bar{c}_-$ to ensure that the solutions are real. For instance, let $c_+ = \bar{c}_- = 1$, then

$$\begin{pmatrix} x \\ y \end{pmatrix} = e^{i\omega t} \begin{pmatrix} 1 \\ i\omega \end{pmatrix} + e^{-i\omega t} \begin{pmatrix} 1 \\ -i\omega \end{pmatrix} = \begin{pmatrix} e^{i\omega t} + e^{-i\omega t} \\ i\omega(e^{i\omega t} - e^{-i\omega t}) \end{pmatrix} = 2 \begin{pmatrix} \cos(\omega t) \\ -\omega \sin(\omega t) \end{pmatrix} \tag{239}$$

which describes an ellipse in the (x, y)-phase plane (orange curve in Fig. 6).

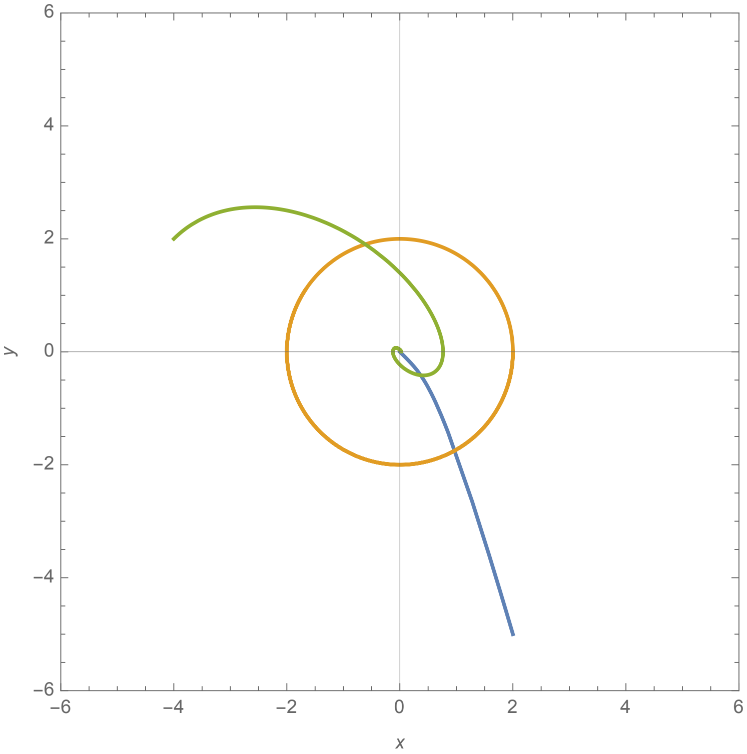

Figure 6: Phase plane trajectories of the damped harmonic oscillator. Blue curve: overdamped case $b = 5, k = 4$. Orange: undamped case $b = 0, k = 1$. Green: underdamped case $b = 1, k = 1$.

15.3.3 Underdamped case $0 < b^2 < 4k$

In this case, the eigenvalues are complex

$$\lambda_\pm = -\frac{b}{2} \pm i\,\frac{1}{2}\sqrt{4k - b^2} = -s \pm i\omega \tag{240}$$

and the solution is given by

$$\begin{aligned}\begin{pmatrix} x \\ y \end{pmatrix} &= c_+ e^{-st+i\omega t}\begin{pmatrix} 1 \\ -s+i\omega \end{pmatrix} + c_- e^{-st-i\omega t}\begin{pmatrix} 1 \\ -s-i\omega \end{pmatrix} \\ &= e^{-st}\left[c_+ e^{i\omega t}\begin{pmatrix} 1 \\ -s+i\omega \end{pmatrix} + c_- e^{-i\omega t}\begin{pmatrix} 1 \\ -s-i\omega \end{pmatrix}\right]\end{aligned}$$

which produces an inward spiral in the (x, y)-phase plane (green curve in Fig. 6).

16 Stability and trace-determinant plane of 2×2 systems

In this section, we will continue to study the stability of the 2×2 matrix ODE

$$\dot{\boldsymbol{\xi}} = A\boldsymbol{\xi} \,, \qquad A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} , \tag{241}$$

where $\boldsymbol{\xi}(t) = (x(t), y(t))$. In particular, we would like to classify the behavior of its solutions in terms of the *matrix invariants* tr$A$ and $\det A$. Here, the word 'invariant' means that the two numbers tr$A$ and $\det A$ do not change under transformations of the form SAS^{-1}, where S is any invertible 2×2 matrix.

16.1 Trace and determinant

Recall that the trace of a $n \times n$ matrix is defined as the sum of its diagonal elements

$$\mathrm{tr}A := A_{ii} := \sum_{i=1}^{n} A_{ii} = A_{11} + A_{22} + A_{33} + \ldots \tag{242}$$

Here, we have introduced Einstein's summation conventions which implies a summation over indices that appear twice. Using this notation, we have

$$SS^{-1} = I \qquad \Leftrightarrow \qquad S_{ij}S^{-1}_{jk} = \delta_{ik},$$

where δ_{ij} is the Kronecker-delta defined by

$$\delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

We then find

$$\mathrm{tr}(SAS^{-1}) = (SAS^{-1})_{ii} = S_{ij}A_{jk}S^{-1}_{ki} = S^{-1}_{ki}S_{ij}A_{jk} = \delta_{kj}A_{jk} = A_{kk} = \mathrm{tr}A, \tag{243}$$

as started above.

Similarly, the determinants of any two matrices satisfy[22]

$$\det(AB) = \det A \, \det B,$$

from which it follows that

$$\begin{aligned} \det(SAS^{-1}) &= \det S \, \det A \, \det S^{-1} \\ &= \det S^{-1} \, \det S \, \det A \\ &= \det(S^{-1}S) \, \det A \\ &= \det I \, \det A \\ &= \det A \end{aligned} \tag{244}$$

These two results demonstrate the usefulness of tr and det for the characterization of matrices.

[22] This will be shown in 18.06

16.2 Trace-determinant plane

Let's consider the matrix A from Eq. (241) with

$$\mathrm{tr}A = a + d\ , \qquad \det A = ad - bc. \tag{245}$$

The eigenvalues λ and eigenvectors $\boldsymbol{v}$ of A satisfy the condition

$$\lambda \boldsymbol{v} = A\boldsymbol{v} \qquad \Leftrightarrow \qquad \mathbf{0} = (A - \lambda I)\boldsymbol{v} \tag{246}$$

where I is the identity matrix. For non-trivial solutions to exist, the eigenvalues λ of A must be roots of the characteristic polynomial

$$\begin{aligned} p(\lambda) = \det(A - \lambda I) = \det \begin{pmatrix} a-\lambda & b \\ c & d-\lambda \end{pmatrix} &= (a-\lambda)(d-\lambda) - bc \\ &= \lambda^2 - a\lambda - d\lambda + ad - bc \\ &= \lambda^2 - (a+d)\lambda + ad - bc \end{aligned}$$

which can be rewritten as

$$p(\lambda) = \lambda^2 - (\mathrm{tr}A)\lambda + \det A \tag{247}$$

On the other hand, we can rewrite $p(\lambda)$ in terms of its roots (the eigenvalues) as

$$p(\lambda) = (\lambda - \lambda_1)(\lambda - \lambda_2) = \lambda^2 - (\lambda_1 + \lambda_2)\lambda + \lambda_1\lambda_2. \tag{248}$$

Comparing the last two equations, we see that

$$\mathrm{tr}A = \lambda_1 + \lambda_2\ , \qquad \det A = \lambda_1 \cdot \lambda_2 \tag{249}$$

These are special cases of the more general formula

$$\mathrm{tr}A = \sum_{i=1}^{n} \lambda_i = \lambda_1 + \lambda_2 + \lambda_3 + \ldots + \lambda_n \tag{250a}$$

$$\det A = \prod_{i=1}^{n} \lambda_i = \lambda_1 \cdot \lambda_2 \cdot \lambda_3 \cdot \ldots \cdot \lambda_n \tag{250b}$$

The roots of (247) can be written as

$$\lambda_{1,2} = \frac{\mathrm{tr}A}{2} \pm \sqrt{\frac{(\mathrm{tr}A)^2}{4} - \det A}. \tag{251}$$

As we have seen in the previous class, the expression under the square root is essential for the qualitative behavior of the solutions. It is therefore often useful to classify solutions of 2×2 systems in terms of the trace-determinant plane (Fig. 7).

An obviously important curve in the trace-determinant plane is the critical parabola (red line Fig. 7)

$$\det A = \frac{(\mathrm{tr}A)^2}{4}. \tag{252}$$

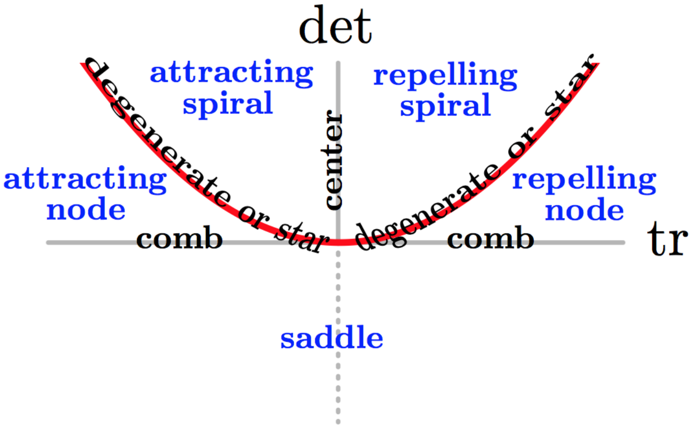

Figure 7: Qualitative different behaviors of 2×2 systems summarized in the trace-determinant plane. The upper left quadrant represents stable solutions, where for all initial conditions the system approaches the fixed point at $(x, y) = (0, 0)$. The structurally stable cases are those corresponding to the large regions in the trace-determinant plane, not the borderline cases.

Above this parabola, we have $\det A > \frac{(\mathrm{tr}A)^2}{4}$ and the system possesses imaginary eigenvalues, leading to spiraling solutions. Below it, we have $\det A < \frac{(\mathrm{tr}A)^2}{4}$ and the two eigenvalues are real, leading to non-spiraling inwards our outwards flowing solutions. Moreover, if $\det A < 0$ then at least on of the eigenvalues will be positive, and the system will have at least one unstable direction. Stable solutions can therefore only be found in the upper left quadrant.

For companion matrices of second-order ODEs from Newtonian dynamics (such as the damped harmonic oscillator), we have

$$A = \begin{pmatrix} 0 & 1 \\ -\det & \mathrm{tr} \end{pmatrix} \tag{253}$$

In this case, the stability regimes can be directly read off the matrix.

16.3 Stability of solutions

Consider a system $\dot{\boldsymbol{\xi}} = A\boldsymbol{\xi}$ with fixed point $\boldsymbol{\xi}_* = \mathbf{0}$. We distinguish three possibilities

(i) If all trajectories tend to $\mathbf{0}$ as $t \to \infty$, the system is called *stable*.

(ii) If some trajectories are unbounded as $t \to \infty$, then the system is called *unstable*.

(iii) In the borderline case in which all solutions are bounded, but do not all tend to $\mathbf{0}$, the system is called *neutrally stable* (example: a center, as found in the undamped oscillator).

Stable systems require that *all* eigenvalues have *negative* real parts. For 2×2 systems, this means that $\mathrm{tr}A < 0$ and $\det A > 0$ (i.e., the characteristic polynomial has only positive coefficients).

16.4 Structural stability

Structural stability concerns the qualitative changes of the ODE system $\dot{\boldsymbol{\xi}} = A\boldsymbol{\xi}$ when the matrix A is modified, and is defined as follows:

- If the phase portrait type is robust in the sense that small perturbations in the entries of A cannot change the type of the phase portrait, then the system is called structurally stable.

The structurally stable cases are those corresponding to the large regions in the trace-determinant plane, not the borderline cases (Fig. 7). For a 2×2-matrix A, the system $\dot{\boldsymbol{\xi}} = A\boldsymbol{\xi}$ is structurally stable if and only if A has either

(i) distinct nonzero real eigenvalues (saddle, repelling node, or attracting node), or

(ii) complex eigenvalues with nonzero real part (spiral).

16.5 Romeo and Juliet

Let's assume the affection dynamics can be described by

$$\begin{pmatrix} \dot{R}(t) \\ \dot{J}(t) \end{pmatrix} = A \begin{pmatrix} R(t) \\ J(t) \end{pmatrix} \tag{254}$$

where R is Romeo's affection for Juliet, and J Juliet's affection for Romeo. Effects from the outside world are neglected, so the system is homogeneous.

As an example, consider the matrix

$$A = \begin{pmatrix} 0 & 4 \\ 1 & 0 \end{pmatrix} \qquad \Rightarrow \qquad \mathrm{tr}A = 0\ , \qquad \det A = -4. \tag{255}$$

In this case, Juliet is responsive to Romeo. If he likes her, then she likes him more. But Romeo is hypersensitive: if Julia likes him, he likes her 4 times more than before. However, this dynamics is generally not stable as it corresponds to a saddle. For example, if Julia starts do dislike Romeo ($J < 0$), then things go downhill. Other relationship scenarios are discussed on MITx.

17 Nonlinear systems

Thus far we have focussed on solving linear systems

$$\dot{\boldsymbol{x}} = A\boldsymbol{x} + \boldsymbol{q}(t). \tag{256}$$

But 'real life' is usually highly nonlinear. Realistic models of complex physical, biological and chemical systems are therefore often based on nonlinear ODE systems

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(t, \boldsymbol{x}), \tag{257}$$

where $\boldsymbol{f}$ is some given vector-valued function.

An important example are gravitational systems[23]

$$\dot{\boldsymbol{r}}_i = \boldsymbol{v}_i \ , \qquad \dot{\boldsymbol{v}}_i = -\sum_{i\neq j} \frac{Gm_j}{|\boldsymbol{r}_i - \boldsymbol{r}_j|^2} \tag{258}$$

where $\boldsymbol{r}_i$ and $\boldsymbol{v}_i$ are the position and velocity vectors of, for example, the stars in our galaxy. If there are N stars then the gravitational system has $6N$ equations, since the collection of all the position and velocity coordinates of the particles forms a $6N$ dimensional vector

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{r}_1 \\ \boldsymbol{v}_1 \\ \boldsymbol{r}_2 \\ \boldsymbol{v}_2 \\ \vdots \\ \boldsymbol{r}_N \\ \boldsymbol{v}_N \end{pmatrix} = \begin{pmatrix} x_1 \\ y_1 \\ z_1 \\ v_{x1} \\ v_{y1} \\ v_{z1} \\ \vdots \\ x_N \\ y_N \\ z_N \\ v_{xN} \\ v_{yN} \\ v_{zN} \end{pmatrix} . \tag{259}$$

Solving such systems for large N numerically, may become a prohibitively ‘expensive’ task even for the fastest super-computers, which is why clever people have invented coarse-grained continuum models (which are then governed by PDEs as in the next classes).

Another much simpler example of a two-dimenionsal ODE system is the Lotka-Volterra population model introduced in Sec. 14.2, which is a standard model in ecology. This model describes a simple predator-prey dynamics, with the prey population $u(t)$ and predator population $v(t)$ governed by

$$\dot{u} = Au - Buv, \tag{260a}$$

$$\dot{v} = -Cv + Euv. \tag{260b}$$

This set of coupled nonlinear ODEs can be rewritten in the form (257) by identifying

$$\boldsymbol{x} = \begin{pmatrix} u \\ v \end{pmatrix} \ , \qquad \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{f}(u,v) = \begin{pmatrix} Au - Buv \\ -Cv + Euv \end{pmatrix} . \tag{260c}$$

In this example $\boldsymbol{f}$ is nonlinear map from $\mathbb{R}^2$ to $\mathbb{R}^2$.

When facing such nonlinear systems, one usually has to resort to computer simulations or linear stability analysis. In this section, we briefly illustrate the main ideas underlying these two approaches. If you are interested in learning more about nonlinear ODE systems, then you should think about taking 18.353.

[23] $|\boldsymbol{x}|$ denotes the Euklidean distance (length) of the vector $\boldsymbol{x} = (x, y, z)$, defined by $|\boldsymbol{x}| = \sqrt{x^2 + y^2 + z^2}$.

17.1 Euler's method for systems

The simplest numerical scheme for solving a nonlinear equation

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(t, \boldsymbol{x}) \tag{261}$$

is the Euler method. The main steps of this algorithm are a straightforward generalization of those for scalar first-order ODEs discussed in Sec. 13.1: We first discretize time t into equidistant intervals Δt by considering discrete time-points

$$t_n = n\,\Delta t\;, \qquad n = 0, 1, \ldots.$$

The goal is to compute the values of the solutions at these time-points, $\boldsymbol{x}_n = \boldsymbol{x}(t_n)$ where $\boldsymbol{x}_0 = \boldsymbol{x}(0)$ is the given vector of initial conditions.

The time derivative $\dot{\boldsymbol{x}} = (\dot{x}_1, \ldots \dot{x}_N)$ is approximated by the forward difference quotient

$$\dot{\boldsymbol{x}}(t_n) \approx \frac{\boldsymbol{x}(t_{n+1}) - \boldsymbol{x}(t_n)}{\Delta t} = \frac{\boldsymbol{x}_{n+1} - \boldsymbol{x}_n}{\Delta t}. \tag{262}$$

The rhs. of Eq. (261) is computed as

$$\boldsymbol{f}_n = \boldsymbol{f}(t_n, \boldsymbol{x}(t_n)) = \boldsymbol{f}(t_n, \boldsymbol{x}_n). \tag{263}$$

The discretized version of Eq. (261) thus becomes

$$\frac{\boldsymbol{x}_{n+1} - \boldsymbol{x}_n}{\Delta t} = \boldsymbol{f}(t_n, \boldsymbol{x}_n). \tag{264}$$

Solving for $\boldsymbol{x}_{n+1}$ yields the first-order recursion relation

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + \boldsymbol{f}(t_n, \boldsymbol{x}_n)\,\Delta t. \tag{265}$$

A graph of the approximate numerical solution is then obtained by plotting and interpolating the points $(0, \boldsymbol{x}_0)$, $(t_1, \boldsymbol{x}_1)$, $(t_2, \boldsymbol{x}_2)$, etc.

The Euler method is not a very efficient numerical scheme, and in practice one usually uses more sophisticated methods such as Runge-Kutta algorithms (see 18.330 or 18.335J for more on this).

17.2 Linear stability analysis

We still demonstrate how linear stability analysis, as introduced in Sec. 12.3 for nonlinear scalar autonomous ODEs $\dot{x} = f(x)$, generalizes to ODE systems. We use the Lotka-Volterra system with $E = B = 1$

$$\dot{u} \;=\; Au - uv \;\;=: f_1(u, v), \tag{266a}$$

$$\dot{v} \;=\; -Cv + uv =: f_2(u, v). \tag{266b}$$

as an example. Just as in the scalar case, we start by looking for fixed points $\boldsymbol{x}_* = (u_*, v_*)$ corresponding to solutions that do not change in time. The fixed points must be simultaneous zeros of the functions f_1 and f_2, i.e., more generally

$$\boldsymbol{f}(\boldsymbol{x}_*) = \boldsymbol{0}. \tag{267}$$

For the Lotka-Volterra system, we find two fixed points

$$(u_0, v_0) = (0, 0)\ , \qquad (u_1, v_1) = (C, A). \tag{268}$$

We then linearize Eq. (266) by Taylor-expanding $\boldsymbol{f}$ near the fixed points $\boldsymbol{x}_*$. Since $\boldsymbol{f}(\boldsymbol{x}_*) = \boldsymbol{0}$, this gives to linear order

$$\boldsymbol{f}(\boldsymbol{x}) \simeq J(\boldsymbol{x}_*)(\boldsymbol{x} - \boldsymbol{x}_*) \tag{269a}$$

where in the 2 × 2 case the Jacobian matrix is given by

$$J(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial f_1}{\partial u} & \frac{\partial f_1}{\partial v} \\ \frac{\partial f_2}{\partial u} & \frac{\partial f_2}{\partial v} \end{pmatrix} \tag{269b}$$

In the general case, abbreviating $\partial_i = \partial / \partial x_i$, we have $J = (J_{ij}) = (\partial_j f_i)$. Writing $\boldsymbol{\epsilon}(t) = \boldsymbol{x}(t) - \boldsymbol{x}_*$, we see that a small perturbation is governed by the linear system

$$\dot{\boldsymbol{\epsilon}} = J(\boldsymbol{x}_*)\boldsymbol{\epsilon}. \tag{270}$$

We can now apply our knowledge of linear systems to understand how the system behaves in the vicinity of the fixed point $\boldsymbol{x}$. In particular, the stability of the FP is determined by the eigenvalues of J. For the Lotka-Volterra system (266), we find

$$J(\boldsymbol{x}) = J(u, v) = \begin{pmatrix} A - v & -u \\ v & u - C \end{pmatrix}, \tag{271}$$

so that

$$J_0 = J(u_0, v_0) = \begin{pmatrix} A & 0 \\ 0 & -C \end{pmatrix}\ , \qquad J_1 = J(u_1, v_1) = \begin{pmatrix} 0 & -C \\ A & 0 \end{pmatrix} \tag{272}$$

The eigenvalues of J_0 are A and $-C$, which means that the fixed point $(0, 0)$, corresponding to both species being extinct, is always unstable since $A > 0$. The eigenvalues of J_1 are $\pm i\sqrt{AC}$, implying that there exist stable oscillatory solutions for the population dynamics!

18 Solving linear equations

In this section, we intro a standard algorithm for solving linear systems

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \qquad \Leftrightarrow \qquad \begin{pmatrix} A_{11} & A_{12} & \cdots A_{1m} \\ \vdots & \vdots & \vdots \\ A_{n1} & A_{n2} & \cdots A_{nm} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} \tag{273}$$

where the matrix A and the vector $\boldsymbol{b}$ are given, and we would like to find *all* vectors $\boldsymbol{x}$ satisfying this equation.

18.1 Balancing a chemical reaction: linear systems in matrix form

Let's consider the simple problem of balancing a chemical reaction. Suppose that we need to find the smallest positive integers a, b, c, and d that balance the reaction

$$a\mathrm{NO}_2 + b\mathrm{H}_2\mathrm{O} \longrightarrow c\mathrm{HNO}_2 + d\mathrm{HNO}_3. \tag{274}$$

The reuirement that the number of nitrogen atoms on the left and right hand sides of the reaction must be equal gives us the mathematical constraint

$$a = c + d, \tag{275}$$

which can be rewritten as $a - c - d = 0$. Similarly, the same constraint on the oxygen and hydrogen atoms yield the equations

$$2a + b - 2c - 3d = 0 \tag{276}$$

and

$$2b - c - d = 0, \tag{277}$$

respectively. These equations constitute a linear system

$$\begin{cases} a \qquad - c - d = 0 \\ 2a + b - 2c - 3d = 0 \\ \quad 2b - c - d = 0 \end{cases} \tag{278}$$

Any linear system can be written in the matrix form (273). For the chemical reaction example, we can write the system as:

$$\underbrace{\begin{pmatrix} 1 & 0 & -1 & -1 \\ 2 & 1 & -2 & -3 \\ 0 & 2 & -1 & -1 \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix}}_{\boldsymbol{x}} = \underbrace{\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}}_{\boldsymbol{b}} \tag{279}$$

and can be represented by the *augmented matrix*

$$(A|\boldsymbol{b}) = \begin{pmatrix} 1 & 0 & -1 & -1 & 0 \\ 2 & 1 & -2 & -3 & 0 \\ 0 & 2 & -1 & -1 & 0 \end{pmatrix} \tag{280}$$

(augmented with an extra column containing the right hand sides). Each row corresponds to an equation. Each column except the last one corresponds to a variable (and contains the coefficients of that variable).

A linear system is *homogeneous* if the right hand sides (the constants) are all zero, and *inhomogeneous* otherwise. So a linear system is homogeneous *if and only if the zero vector is a solution.*

A linear system is called *consistent* if it has at least one solution, and *inconsistent* if there are no solutions.

18.2 Equation operations and row operations

A good way to solve a linear system is to perform the following operations repeatedly, in some order:

- Multiply an equation by a nonzero number.
- Interchange two equations.
- Add a multiple of one equation to another equation.

The solution set is unchanged at each step. The equation operations correspond to operations on the augmented matrix, called *elementary row operations*:

- Multiply a row by a nonzero number.
- Interchange two rows.
- Add a multiple of one row to another row (while leaving the first row as it was).

18.3 Row-echelon form and pivots

To solve the linear systems (273) systematically, we use the following sequence of steps:

1. Use row operations to convert the augmented matrix $(A|\boldsymbol{b})$ to a particularly simple form, called *row-echelon form*.
2. Solve the new system by *back-substitution*.

Before explaining row-echelon form, we need a few preliminary definitions. A *zero row* of a matrix is a row consisting entirely of zeros. A *nonzero row* of a matrix is a row with at least one nonzero entry. In each nonzero row, the first nonzero entry is called the *pivot*. The following 4×5 matrix has one zero row, and three pivots shown in red

$$\begin{pmatrix} 0 & -5 & 4 & 4 & 3 \\ 2 & 0 & 0 & 1 & 7 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 \end{pmatrix}.$$

A matrix is in *row-echelon form* if it satisfies both of the following conditions:

1. All the zero rows (if any) are grouped at the bottom of the matrix.
2. Each pivot lies farther to the right than the pivots of higher rows.

Some books require also that each pivot be a 1. We are not going to require this for row-echelon form, but we will require it for *reduced* row-echelon form later on.

18.4 Gaussian elimination

Gaussian elimination is an algorithm for converting any matrix into row-echelon form by performing row operations. Here are the steps:

1. Find the leftmost nonzero column, and the first nonzero entry in that column (read from the top down).
2. If that entry is not already in the first row, interchange its row with the first row.
3. Make all other entries of the column zero by adding suitable multiples of the first row to the others.
4. At this point, the first row is done, so ignore it, and repeat the steps above for the remaining submatrix (with one fewer row). In each iteration, ignore the rows already taken care of. Eventually the whole matrix will be in row-echelon form.

As an example, let us convert the 4 × 7 matrix

$$\begin{pmatrix} 0 & 0 & 6 & 2 & -4 & -8 & 8 \\ 0 & 0 & 3 & 1 & -2 & -4 & 4 \\ 2 & -3 & 1 & 4 & -7 & 1 & 2 \\ 6 & -9 & 0 & 11 & -19 & 3 & 0 \end{pmatrix}$$

to row-echelon form. [24]
Step 1. The leftmost nonzero column is the first one, and its first nonzero entry is the 2:

$$\begin{pmatrix} 0 & 0 & 6 & 2 & -4 & -8 & 8 \\ 0 & 0 & 3 & 1 & -2 & -4 & 4 \\ 2 & -3 & 1 & 4 & -7 & 1 & 2 \\ 6 & -9 & 0 & 11 & -19 & 3 & 0 \end{pmatrix}.$$

Step 2. The 2 is not in the first row, so interchange its row with the first row:

$$\begin{pmatrix} 2 & -3 & 1 & 4 & -7 & 1 & 2 \\ 0 & 0 & 3 & 1 & -2 & -4 & 4 \\ 0 & 0 & 6 & 2 & -4 & -8 & 8 \\ 6 & -9 & 0 & 11 & -19 & 3 & 0 \end{pmatrix}.$$

Step 3. To make all other entries of the column zero, we need to add −3 times the first row to the last row (the other rows are OK already):

$$\begin{pmatrix} 2 & -3 & 1 & 4 & -7 & 1 & 2 \\ 0 & 0 & 3 & 1 & -2 & -4 & 4 \\ 0 & 0 & 6 & 2 & -4 & -8 & 8 \\ 0 & 0 & -3 & -1 & 2 & 0 & -6 \end{pmatrix}.$$

[24] his example is taken from Hill, *Elementary linear algebra with applications*, p. 17.)

Step 4. Now the first row is done. Start over with the 3×7 submatrix that remains beneath it:

$$\begin{pmatrix} 2 & -3 & 1 & 4 & -7 & 1 & 2 \\ 0 & 0 & 3 & 1 & -2 & -4 & 4 \\ 0 & 0 & 6 & 2 & -4 & -8 & 8 \\ 0 & 0 & -3 & -1 & 2 & 0 & -6 \end{pmatrix}.$$

Step 1. The leftmost nonzero column in that submatrix is now the third column, and its first nonzero entry is the 3:

$$\begin{pmatrix} 2 & -3 & 1 & 4 & -7 & 1 & 2 \\ 0 & 0 & 3 & 1 & -2 & -4 & 4 \\ 0 & 0 & 6 & 2 & -4 & -8 & 8 \\ 0 & 0 & -3 & -1 & 2 & 0 & -6 \end{pmatrix}.$$

Step 2. The 3 is already in the first row of the submatrix (we are ignoring the first row of the whole matrix), so no interchange is necessary.

Step 3. To make all other entries of the column zero, add -2 times the (new) first row to the (new) second row, and 1 times the (new) first row to the (new) third row:

$$\begin{pmatrix} 2 & -3 & 1 & 4 & -7 & 1 & 2 \\ 0 & 0 & 3 & 1 & -2 & -4 & 4 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -4 & -2 \end{pmatrix}.$$

Step 4. Now the first and second row of the original matrix are done. Start over with the 2×7 submatrix beneath them:

$$\begin{pmatrix} 2 & -3 & 1 & 4 & -7 & 1 & 2 \\ 0 & 0 & 3 & 1 & -2 & -4 & 4 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -4 & -2 \end{pmatrix}.$$

Step 1. The leftmost nonzero column in that submatrix is now the penultimate column, and its first nonzero entry is the -4 at the bottom:

$$\begin{pmatrix} 2 & -3 & 1 & 4 & -7 & 1 & 2 \\ 0 & 0 & 3 & 1 & -2 & -4 & 4 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -4 & -2 \end{pmatrix}.$$

Step 2. The -4 is not in the first row of the submatrix, so interchange its row with the first row of the submatrix:

$$\begin{pmatrix} 2 & -3 & 1 & 4 & -7 & 1 & 2 \\ 0 & 0 & 3 & 1 & -2 & -4 & 4 \\ 0 & 0 & 0 & 0 & 0 & -4 & -2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Step 3. The other entry in this column of the submatrix is already 0, so this step is not necessary.

The matrix is now in row-echelon form:

$$\begin{pmatrix} 2 & -3 & 1 & 4 & -7 & 1 & 2 \\ 0 & 0 & 3 & 1 & -2 & -4 & 4 \\ 0 & 0 & 0 & 0 & 0 & -4 & -2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

18.5 Reduced row-echelon form and Gauss-Jordan elimination

With even more row operations, one can simplify a matrix in row-echelon form to an even more special form:

A matrix is in *reduced row-echelon form (RREF)* if it satisfies all of the following conditions:

1. It is in row-echelon form.
2. Each pivot is a 1.
3. In each pivot column, all the entries are 0 except for the pivot itself.

Gauss-Jordan elimination is an algorithm for converting any matrix into RREF by performing row operations. Here are the steps:

1. Use Gaussian elimination to convert the matrix to row echelon form.
2. Divide the last nonzero row by its pivot, to make the pivot 1.
3. Make all entries in that pivot's column 0 by adding suitable multiples of the pivot's row to the rows above.
4. At this point, the row in question (and all rows below it) are done. Ignore them, and go back to Step 2, but now with the remaining submatrix, above the row just completed.

Eventually the whole matrix will be in RREF.

18.6 Back-substitution

Matrices in row-echelon form correspond to systems that are ready to be solved immediately by *back-substitution*: solve for each variable in reverse order, while introducing a parameter for each variable not directly expressed in terms of later variables, and substitute values into earlier equations once they are known. We illustrate this procedure for the chemical reaction example (274), which led to the linear system

$$\begin{aligned} a \qquad - c - d &= 0 \\ 2a + b - 2c - 3d &= 0 \\ 2b - c - d &= 0, \end{aligned}$$

represented by the augmented matrix

$$A = \begin{pmatrix} 1 & 0 & -1 & -1 & 0 \\ 2 & 1 & -2 & -3 & 0 \\ 0 & 2 & -1 & -1 & 0 \end{pmatrix}.$$

We use Gauss-Jordan elimination to put the matrix in reduced row-echelon form

$$\begin{aligned} A &= \begin{pmatrix} 1 & 0 & -1 & -1 & 0 \\ 2 & 1 & -2 & -3 & 0 \\ 0 & 2 & -1 & -1 & 0 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 0 & -1 & -1 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 2 & -1 & -1 & 0 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 0 & -1 & -1 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & -1 & 1 & 0 \end{pmatrix} \\ &\longrightarrow \begin{pmatrix} 1 & 0 & -1 & -1 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 1 & -1 & 0 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 0 & 0 & -2 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 1 & -1 & 0 \end{pmatrix} = \mathrm{RREF}(A). \end{aligned}$$

Back-substituting the original variables, we are left with the equations

$$a - 2d = 0 \qquad (281a)$$

$$b - d = 0 \qquad (281b)$$

$$c - d = 0. \qquad (281c)$$

Setting $d = c_1$, we see that

$$\begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} = c_1 \begin{pmatrix} 2 \\ 1 \\ 1 \\ 1 \end{pmatrix},$$

for some parameter c_1. The smallest positive integers a, b, c, and d that balance the reaction occur when $c_1 = 1$, so the balanced chemical reaction is

$$2\mathrm{NO}_2 + \mathrm{H}_2\mathrm{O} \longrightarrow \mathrm{HNO}_2 + \mathrm{HNO}_3. \qquad (282)$$

This method can be used in general to balance chemical reactions.

Analogy with solutions to differential equations. If $p(D)$ is a linear differential operator, then the general solution to $p(D)x = f$ is $x = x_p + x_h$, where x_p is any particular solution, and x_h is the general solution to the homogeneous equation $p(D)x = 0$. The general solution to a linear system $A\boldsymbol{x} = \boldsymbol{b}$ has the form $\boldsymbol{x} = \boldsymbol{x}_p + \boldsymbol{x}_h$, where $\boldsymbol{x}_p$ is a particular solution and $\boldsymbol{x}_p$ is the general solution to the homogeneous system. In this sense a matrix A can be thought of as the analog of the a differential operator $p(D)$.

19 Null space, column space and determinants

19.1 Nullspace

Here is what happens in general for homogeneous linear systems:

Theorem: For any homogeneous linear system $A\boldsymbol{x} = \boldsymbol{0}$, the set of all solutions is a vector space, called the *nullspace* of the matrix A, and denoted $\mathrm{NS}(A)$.

The analogy to solutions of ODEs is as follows: If $p(D)$ is a linear differential operator, the space of solutions to a homogeneous ODE $p(D)x = 0$ is analogous to the nullspace of a matrix A.

The nullspace of a matrix arises in many physical, chemical or biological scenarios. For example, suppose that we want to balance the chemical reaction

$$a\mathrm{C}_8\mathrm{H}_{18} + b\mathrm{O}_2 \longrightarrow c\mathrm{CO}_2 + d\mathrm{H}_2\mathrm{O},$$

that is, find the smallest positive integers a, b, c, and d that make the number of each atom on both sides of the reaction equal. This constraint make this problem equivalent to finding the vector with the smallest integer coefficients in the nullspace of the matrix

$$S = \begin{pmatrix} 8 & 0 & -1 & 0 \\ 18 & 0 & 0 & -2 \\ 0 & 2 & -2 & -1 \end{pmatrix}.$$

It's easy to see how a matrix like this could be constructed for a general reaction. Note that, just as in the previous section, the system has *one free variable* since there are many possible ways to balance the reaction, but any vector of coefficients a, b, c, and d that make the number of atoms on both sides of the reaction equal is an integer multiple of the vector with the smallest positive integer coefficients.

Suppose that the result of putting a matrix A in row-echelon form is B. Then $\mathrm{NS}(A) = \mathrm{NS}(B)$, since row reductions do not change the solutions, and

$$\begin{aligned} \dim \mathrm{NS}(A) &= \#\text{non-pivot columns of} B \\ &= \#\text{free variables}. \end{aligned} \tag{283}$$

Sometimes $\dim \mathrm{NS}(A)$ is called the *nullity* of A. To summarize, here are the steps to find the dimension of the space of solutions to a homogeneous linear system $A\boldsymbol{x} = \mathbf{0}$:

1. Perform Gaussian elimination on A to convert it to a matrix B in row-echelon form.

2. Identify the pivots of B

3. Count the number of *non-pivot* columns of B; that number is $\dim \mathrm{NS}(A)$.

Warning: You must put the matrix in row-echelon form before counting non-pivot columns!

And here are the steps to find a basis of the space of solutions to a homogeneous linear system $A\boldsymbol{x} = \mathbf{0}$:

1. Perform Gaussian elimination on A to convert it to a matrix B in row-echelon form.

2. Use back-substitution to find the general solution to $B\boldsymbol{x} = \mathbf{0}$.

3. The general solution will be expressed as the general linear combination of a list of vectors; that list is a basis for $\mathrm{NS}(A)$.

19.2 Inhomogeneous linear systems: theory and algorithms

For an inhomogeneous linear system

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

there are two possibilities:

1. There are no solutions.
2. There exists a solution. In this case, if $\boldsymbol{x}_p$ is a particular solution to $A\boldsymbol{x} = \boldsymbol{b}$, and $\boldsymbol{x}_h$ is the *general* solution to the homogeneous system $A\boldsymbol{x} = \boldsymbol{0}$, then $\boldsymbol{x} := \boldsymbol{x}_p + \boldsymbol{x}_h$ is the general solution to $A\boldsymbol{x} = \boldsymbol{b}$.

Here is why: Suppose that a solution exists; let $\boldsymbol{x}_p$ be one, so $A\boldsymbol{x}_p = \boldsymbol{b}$. If $\boldsymbol{x}_h$ satisfies $A\boldsymbol{x}_h = 0$, adding the two equations gives

$$A(\boldsymbol{x}_p + \boldsymbol{x}_h) = \boldsymbol{b}, \tag{284}$$

so adding $\boldsymbol{x}_p$ to $\boldsymbol{x}_h$ produces a solution $\boldsymbol{x}$ to the inhomogeneous equation. Every solution $\boldsymbol{x}$ to $A\boldsymbol{x} = \boldsymbol{b}$ arises this way from some $\boldsymbol{x}_h$. Specifically, from $\boldsymbol{x}_h := \boldsymbol{x} - \boldsymbol{x}_p$, which satisfies

$$A\boldsymbol{x}_h = A\boldsymbol{x} - A\boldsymbol{x}_p = \boldsymbol{b} - \boldsymbol{b} = \boldsymbol{0}.$$

Remark: To *solve* $A\boldsymbol{x} = \boldsymbol{b}$ in practice, however, don't use $\boldsymbol{x} = \boldsymbol{x}_p + \boldsymbol{x}_h$. Instead use Gaussian elimination and back-substitution. The above is just to describe the shape of the solution.

19.3 Column space and rank

The *column space* of a matrix A is the span of its columns. The notation for it is $\mathrm{CS}(A)$. It is also called the *image* of A, and written $\mathrm{im}(A)$; the reason will be clearer when we talk about the geometric interpretation. Since $\mathrm{CS}(A)$ is a span, it is a vector space.

Here is what happens in general for (possibly inhomogeneous) linear systems:

Theorem: The linear system $A\boldsymbol{x} = \boldsymbol{b}$ has a solution if and only if $\boldsymbol{b}$ is in $\mathrm{CS}(A)$.

For example, for the matrix

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \end{pmatrix}$$

we have

$$\mathrm{CS}(A) = \text{the span of } \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 2 \\ 4 \end{pmatrix}, \begin{pmatrix} 3 \\ 6 \end{pmatrix},$$

which is 1-dimensional vector, since the column vectors are all constant multiples of the vector $\begin{pmatrix} 1 \\ 2 \end{pmatrix}$. The span of the vector

$$\begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

is the line $y = 2x$ in $\mathbb{R}^2$. The statement of the theorem becomes obvious, when we rewrite the equation $A\boldsymbol{x} = \boldsymbol{b}$ in component form as

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix} x + \begin{pmatrix} 2 \\ 4 \end{pmatrix} y + \begin{pmatrix} 3 \\ 6 \end{pmatrix} z = \boldsymbol{b}$$

which shows that a solution (x, y, z) exists only if $\boldsymbol{b}$ lies in the span $\mathrm{CS}(A)$ of the columns of A.

In general, the steps to compute a basis for $\mathrm{CS}(A)$ are: [25]

1. Perform Gaussian elimination to convert A to a matrix B in row-echelon form.
2. Identify the pivot columns of B.
3. The corresponding columns of A are a basis for $\mathrm{CS}(A)$.

In particular,

$$\dim \mathrm{CS}(A) = \# \quad \text{pivot columns of } B.$$

Warning: Usually $\mathrm{CS}(A) \neq \mathrm{CS}(B)$.
The *rank* of A is defined by

$$\mathrm{rank}(A) := \dim \mathrm{CS}(A).$$

Theorem: For any $m \times n$ matrix A,

$$\dim \mathrm{NS}(A) + \mathrm{rank}(A) = n. \tag{285}$$

This theorem is sometimes called the *Rank-Nullity Theorem*. The proof is straightforward

$$\begin{aligned} \dim \mathrm{NS}(A) + \mathrm{rank}(A) &= (\#\text{ non-pivot columns of } B) + (\#\text{ pivot columns of } B) \\ &= \#\text{ columns of } B \\ &= n. \end{aligned}$$

[25] **Proof:** Let C be the reduced row-echelon form of A. If

$$\text{fifth column} = 3(\text{first column}) + 7(\text{second column})$$

is true for a matrix, it will remain true after any row operation.

Similarly, any linear relation between columns is preserved by row operations. So the linear relations between columns of A and the same as the linear relations between columns of C. The condition that certain numbered columns (say the first, second, and fourth) of a matrix form a basis is expressible in terms of which linear relations hold. So if certain columns form a basis for $\mathrm{CS}(C)$, the same numbered columns will form a basis for $\mathrm{CS}(A)$.

Also, performing Gauss-Jordan reduction on B to obtain C in reduced row-echelon form does not change the pivot locations. Thus it will be enough to show that the pivot columns of C form a basis of $\mathrm{CS}(C)$. Since C is in reduced row-echelon form, the pivot columns of C are the first r of the m standard basis vectors for $\mathbb{R}^m$, where r is the number of nonzero rows of C. These columns are linearly *independent,* and every other column is a linear combination of them, since the entries of C below the first r rows are all zeros. Thus the pivot columns of C form a basis of $\mathrm{CS}(C)$.

19.4 Computing a basis for a span

The Gauss algorithm is of immense practical importance. Consider, for example, the following common question: Given vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n \in \mathbb{R}^m$, how can one compute a basis of $\operatorname{Span}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$?

This problem can by solved by the following algorithm:

1. Form the matrix A whose columns are $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$.

2. Obtain a new matrix B by putting A in row-echelon form.

3. Find the pivot columns of B by putting A in row-echelon form.

4. The associated columns of A form a basis for $\mathrm{CS}(A)$.

19.5 Determinants

Recall that to each *square* matrix A is associated a number called the *determinant*:

$$\begin{aligned} \det \begin{pmatrix} a \end{pmatrix} &:= a \\ \det \begin{pmatrix} a & b \\ c & d \end{pmatrix} &:= ad - bc \\ \det \begin{pmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{pmatrix} &:= a_1b_2c_3 + a_2b_3c_1 + a_3b_1c_2 - c_1b_2a_3 - c_2b_3a_1 - c_3b_1a_2. \end{aligned}$$

An alternative notation for the determinant is

$$|A| = \begin{vmatrix} a & b \\ c & d \end{vmatrix}.$$

This is a (pseudo-)scalar, not a matrix. Geometrically, the *absolute value* of $\det A$ is the volume spanned by the columns of A.

19.5.1 Laplace expansion for determinants

The *Laplace expansion* (along the first row) for a 3×3 determinant is obtained as follows:

$$\begin{vmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{vmatrix} = +a_1 \begin{vmatrix} b_2 & b_3 \\ c_2 & c_3 \end{vmatrix} - a_2 \begin{vmatrix} b_1 & b_3 \\ c_1 & c_3 \end{vmatrix} + a_3 \begin{vmatrix} b_1 & b_2 \\ c_1 & c_2 \end{vmatrix}.$$

The general rule leading to the formula above is this:

1. Move your finger along the entries in a row.

2. At each position, compute the *minor*, defined as the smaller determinant obtained by crossing out the row and the column through your finger; then multiply the minor by

the number you are pointing at, and adjust the sign according to the checkerboard pattern

$$\begin{array}{ccc} + & - & + \\ - & + & - \\ + & - & + \end{array}$$

The pattern always starts with + in the upper left corner.

3. Add up the results.

There is a similar expansion for a determinant of any size, computed along any row or column.

19.5.2 Properties of Determinants

We summarize some key properties of the determinant.

1. Interchanging two rows changes the sign of $\det A$.
2. Multiplying an entire row by a scalar c multiples $\det A$ by c.
3. Adding a multiple of a row to another row does not change $\det A$.
4. If one of the rows is all 0, then $\det A = 0$.
5. $\det(AB) = \det(A)\det(B)$, assuming A, B are square matrices of the same size.

In particular, row operations multiply $\det A$ by nonzero scalars, but do not change whether $\det A = 0$.

19.5.3 Diagonal matrices

The *diagonal* of a matrix consists of the entries a_{ij} with $i = j$. A *diagonal matrix* is a square matrix that has zeros everywhere outside the diagonal[26]

$$\begin{pmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{pmatrix}$$

An *upper triangular matrix* is a matrix whose entries strictly below the diagonal are all 0:

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{pmatrix}$$

The entries on or above the diagonal may or may not be 0. In particular, any square matrix in row-echelon form is upper triangular.

Theorem: The determinant of an upper triangular matrix equals the product of the diagonal entries.

[26] It may have some zeros along the diagonal too.

For example,

$$\det \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{pmatrix} = a_{11}\, a_{22}\, a_{33}.$$

Why is the theorem true in general? The Laplace expansion along the first column shows that the determinant is a_{11} times a upper triangular minor with diagonal entries $a_{22}, \ldots, a_{nn}$. We can repeat this argument successively for all minors to proof the theorem by induction.

19.6 Identity matrix and matrix inversion

The linear transformation $f\colon \mathbb{R}^3 \to \mathbb{R}^3$ that does nothing to its input, $\boldsymbol{f}(x, y, z) := (x, y, z)$, is called the *identity*. The corresponding matrix, the 3×3 *identity matrix* I, is given by

$$I = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

The $n \times n$ identity matrix is similar, with 1s along the diagonal. It has the property that $AI = A$ and $IA = A$ whenever the matrix multiplication is defined. Thus, the identity matrix I is the analog for matrix multiplication of the number 1 for multiplication of real (or complex) numbers.

The *inverse* of an $n \times n$ matrix A is another $n \times n$ matrix A^{-1} such that

$$AA^{-1} = I \qquad \text{and} \qquad A^{-1}A = I.$$

It exists if and only if $\det A \neq 0$.

Suppose that A represents the linear transformation $\boldsymbol{f}$. Then A^{-1} exists if and only if an inverse function $\boldsymbol{f}^{-1}$ exists; in that case, A^{-1} represents $\boldsymbol{f}^{-1}$. As an example, consider the rotation matrix

$$R := \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

Its determinant is $\det R = 1 \neq 0$ and the inverse linear transformation is rotation by $-\theta$, so

$$R^{-1} = \begin{pmatrix} \cos(-\theta) & -\sin(-\theta) \\ \sin(-\theta) & \cos(-\theta) \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}.$$

As a check, try multiplying R by this matrix, in either order.

Suppose that $\det A \neq 0$. In 18.02, you learned one algorithm to compute A^{-1}, using the cofactor matrix. Now that we know how to compute RREF, we can give a faster algorithm (faster for big matrices, at least):

1. Form the $n \times 2n$ augmented matrix $[A|I]$.

2. Convert to RREF; the result will be $[I|B]$ for some $n \times n$ matrix B.

3. Then $A^{-1} = B$.

This is a special case of solving a matrix equation $AX = B$, since A^{-1} is the solution to $AX = I$.

Finding inverse matrices rapidly is essential for solving complicated ODEs and PDEs numerically. This corroborates the practical importance of the Gauss-Jordan scheme.

20 Eigenthings

In this part, we will review important aspects of eigenvalues and eigenvectors for general $n \times n$ square matrices.

20.1 Characteristic polynomials

The *characteristic polynomial* of a $n \times n$ square matrix A can be defined as

$$p(\lambda) = \det(A - \lambda I) \tag{286}$$

or equivalently as

$$p_+(\lambda) = \det(\lambda I - A) \tag{287}$$

instead.[27] The two definitions are related by

$$p(\lambda) = (-1)^n p_+(\lambda). \tag{288}$$

That is, they coincide when n is even and differ by an overall minus sign when n is odd. Since we are only interested in the zeros of the characteristic polynomial, which are identical for $p(\lambda)$ and $p_+(\lambda)$, you may choose whichever definition you prefer. As we have mentioned earlier, although the characteristic polynomial of matrix *per se* is not the exactly same concept as the characteristic polynomial of a constant-coefficient linear ODE, there exists a close connection, arising when such a DE is converted to a first-order system of linear ODEs.

We have seen that, if A is a 2×2 matrix, then the characteristic polynomial of A is

$$p_+(\lambda) = \lambda^2 - (\mathrm{tr}A)\lambda + (\det A).$$

Suppose that $n > 2$. Then, for an $n \times n$ matrix A, the characteristic polynomial has the form

$$p_+(\lambda) = \lambda^n - (\mathrm{tr}A)\lambda^{n-1} + \cdots \pm \det A$$

where the $\pm$ is $+$ if n is even, and $-$ if n is odd. So knowing $\mathrm{tr}A$ and $\det A$ determines some of the coefficients of the characteristic polynomial, but not all of them.

20.2 Eigenvalues, eigenvectors and eigenspaces

Suppose that A is an $n \times n$ matrix.

- An *eigenvector* of A is a nonzero[28] vector $\boldsymbol{v}$ such that $A\boldsymbol{v} = \lambda\boldsymbol{v}$ for some scalar λ.

- An *eigenvalue* of A is a scalar λ such that $A\boldsymbol{v} = \lambda\boldsymbol{v}$ for some nonzero vector $\boldsymbol{v}$.

[27] A minor technical advantage of $p_+(\lambda)$ is that this is a monic degree n polynomial in the variable λ, which means that the leading coefficient is 1, so the polynomial looks like $\lambda^n + \ldots$.

[28] Warning: Some authors consider $\mathbf{0}$ to be an eigenvector too, but we do not.

Let A be a square matrix, and let λ be a scalar. Then λ is an eigenvalue of A if and only if $\det(\lambda I - A) = 0$.

As an example, let's find the eigenvalues of the upper triangular matrix

$$A := \begin{pmatrix} 2 & 3 & 5 \\ 0 & 2 & 7 \\ 0 & 0 & 6 \end{pmatrix}.$$

The characteristic polynomial is

$$\det(\lambda I - A) = \det \begin{pmatrix} \lambda - 2 & -3 & -5 \\ 0 & \lambda - 2 & -7 \\ 0 & 0 & \lambda - 6 \end{pmatrix} = (\lambda - 2)(\lambda - 2)(\lambda - 6),$$

so the eigenvalues, listed with multiplicity, are 2, 2, 6.

In general, for any upper triangular or lower triangular matrix, the eigenvalues are the diagonal entries.

20.2.1 Complex eigenvalues and eigenvectors

Generally, for any $n \times n$ matrix, the characteristic polynomial is of degree n, so the fundamental theorem of algebra shows that the total number of complex eigenvalues counted with multiplicity is n.

As an example, let's find the eigenvalues and eigenvectors of the 90° counterclockwise rotation matrix

$$R = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

Since $\mathrm{tr} R = 0$ and $\det R = 1$, the characteristic polynomial of R is $\lambda^2 + 1$. Its roots are i and $-i$; these are the eigenvalues.

The eigenspace of $\lambda_1 = i$ is $\mathrm{NS}(R - iI)$. Converting

$$R - iI = \begin{pmatrix} -i & -1 \\ 1 & -i \end{pmatrix}$$

to row-echelon form gives

$$\begin{pmatrix} -i & -1 \\ 0 & 0 \end{pmatrix}, \tag{289}$$

so we solve $-ix - y = 0$ by back-substitution and get the general solution

$$\boldsymbol{v}_1 = c \begin{pmatrix} i \\ 1 \end{pmatrix}. \tag{290}$$

Thus the eigenvectors having eigenvalue i are the nonzero scalar multiples of $(i, 1)$.

Applying complex conjugation to $R - iI$ gives $R + iI$. So the vector in the nullspace of $R + iI$ is the eigenvector with eigenvalue $\lambda_2 = -i$ by definition. But it is also the complex conjugate of the vector in the nullspace of $R - iI$. This shows that the eigenvectors having $-i$ as an eigenvector are the nonzero scalar multiples of $(-i, 1)$.

20.2.2 Zero as an eigenvalue

While the vector $\mathbf{0}$ is not an eigenvector, it is possible for the scalar 0 to be an eigenvalue. The following statements are equivalent for a square matrix A.

- 0 is an eigenvalue.
- There exists a nonzero solution to $A\boldsymbol{x} = \mathbf{0}$.
- $\det A = 0$.

The last one is a consequence of the fact that the determinant can be written as the product of the eigenvalues.

20.2.3 Eigenspaces

The *eigenspace* of an eigenvalue λ of a square matrix A is the set of all eigenvectors having that eigenvalue, together with the zero vector $\mathbf{0}$.

So each eigenspace is a set of vectors. In fact, each eigenspace is a *vector space.* Why? It is the set of all solutions to $A\boldsymbol{x} = \lambda\boldsymbol{x}$ (including $\boldsymbol{x} = \mathbf{0}$), or equivalently to $(A-\lambda I)\boldsymbol{x} = \mathbf{0}$. Thus the eigenspace of λ is the same as $\mathrm{NS}(A - \lambda I)$, which is a vector space.

Assume we have found a certain eigenvalue λ for a given square matrix A. We may then use the Gauss algorithm to find the eigenvectors associated to this eigenvalue as follows:

1. Calculate $A - \lambda I$.
2. Use Gaussian elimination and back-substitution to compute a basis of $\mathrm{NS}(A - \lambda I)$.
3. The eigenvectors having eigenvalue λ are all the linear combinations of those basis vectors (not including the zero vector).

Let λ be an eigenvalue of an $n \times n$ matrix A. Suppose that the multiplicity of λ as a root of the characteristic polynomial is m. Then

$$1 \quad \leq \quad \text{(dimension of eigenspace of } \lambda) \quad \leq \quad m.$$

Given λ, the dimension of the eigenspace of λ is also the maximum number of linearly independent eigenvectors of eigenvalue λ that can be found. This dimension is at least 1 since A has at least one eigenvector of eigenvalue λ (otherwise λ would not have been an eigenvalue). That this dimension is at most m requires more work to prove, and we're not going to do it in this class.

Generally, the eigenspace of λ is called *complete* if its dimension equals the multiplicity m of λ, and *deficient* if its dimension is less than m. If the multiplicity is 1, then the dimension of the eigenspace is sandwiched between 1 and 1, so the eigenspace is not deficient. A matrix is called *deficient* if one of its eigenspaces is deficient.

For example, consider 9×9 matrix has characteristic polynomial $(\lambda-2)^3(\lambda-5)^6$. What are the possibilities for the dimension of the eigenspace of 2? In this case, $m = 3$, so the dimension is either 1, 2, or 3.

20.3 Linear independence of eigenvectors

For the application to solving linear systems of ODEs, given an $n \times n$ matrix A we will want to find as many linearly independent eigenvectors as possible. To do this, we choose a basis of each eigenspace, and concatenate these lists of eigenvectors; it turns out that the resulting list is linearly independent. How many eigenvectors are in this list? There are two possibilities:

- If all the eigenspaces are complete, then the number of linearly independent eigenvectors from each eigenspace is the multiplicity of the eigenvalue, so the total number of eigenvectors in our list is the total number of eigenvalues counted with multiplicity, which is n. In this case, the n eigenvectors form a basis of $\mathbb{C}^n$, since their span is n-dimensional.
- If instead A is deficient, then the number of linearly independent eigenvectors is less than n.

Why does concatenating the bases produce a linearly independent list? The vectors within each basis are linearly independent, and there are no linear relations involving eigenvectors from different eigenspaces because of the following:

Theorem: Fix a square matrix A. Eigenvectors corresponding to distinct eigenvalues are linearly independent.

Proof: Suppose that $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ are eigenvectors corresponding to distinct eigenvalues $\lambda_1, \ldots, \lambda_n$. Suppose that there were a linear relation

$$c_1\boldsymbol{v}_1 + \cdots c_n\boldsymbol{v}_n = \mathbf{0}.$$

Apply $A - \lambda_1 I$ to both sides; this gets rid of the first summand on the left. Next apply $A - \lambda_2 I$, and so on, up to $A - \lambda_{n-1} I$. This shows that some nonzero number times $c_n\boldsymbol{v}_n$ equals $\mathbf{0}$. But $\boldsymbol{v}_n \neq \mathbf{0}$, so $c_n = 0$. Similarly each c_i must be 0. Thus only the trivial relation between $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ exists, so they are linearly independent.

21 Matrix exponentials and linear systems

We know that the first order equation

$$\dot{x} = ax \tag{291a}$$

with constant rate parameter a and initial condition $x(0) = c$ is solved by

$$x(t) = e^{at}c. \tag{291b}$$

Equation (291a) is the 1×1 case of the matrix equation

$$\dot{\boldsymbol{x}} = A\boldsymbol{x}, \tag{292}$$

where A is an $n \times n$ square matrix with constant entries. Our goal in this section is to construct the matrix generalization of the exponential solution formula (291b).

21.1 Matrix diagonalization

Consider an $n \times n$ square matrix A with eigenvalues

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n.$$

Assume the corresponding eigenvectors

$$\boldsymbol{v}_1 = \begin{pmatrix} v_{11} \\ v_{21} \\ \vdots \\ v_{n1} \end{pmatrix}, \qquad \boldsymbol{v}_2 = \begin{pmatrix} v_{12} \\ v_{22} \\ \vdots \\ v_{n2} \end{pmatrix}, \qquad \ldots, \qquad \boldsymbol{v}_n = \begin{pmatrix} v_{1n} \\ v_{2n} \\ \vdots \\ v_{nn} \end{pmatrix} \tag{293}$$

form a linearly independent system. This is always the case if all the eigenvalues λ_i are distinct (i.e., have multiplicity 1), as we showed at the end of the previous section. If some eigenvalues have multiplicities and eigenspace dimensions $m > 1$, then one can always select m linearly independent vectors from that eigenspace.

We next define the matrix $\Lambda = \mathrm{diag}(\lambda_1, \ldots \lambda_n)$, explicitly given by

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_n \end{pmatrix} \tag{294}$$

and the matrix V which has the eigenvectors as columns

$$V = \begin{pmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \ldots & \boldsymbol{v}_n \end{pmatrix} = \begin{pmatrix} v_{11} & v_{12} & \cdots & v_{1n} \\ v_{21} & v_{22} & \cdots & v_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ v_{n1} & v_{n2} & \cdots & v_{nn} \end{pmatrix}. \tag{295}$$

It will be useful to write the inverse V^{-1} in terms of its row vectors

$$V^{-1} = \begin{pmatrix} \boldsymbol{w}_1^\top \\ \boldsymbol{w}_2^\top \\ \vdots \\ \boldsymbol{w}_n^\top \end{pmatrix} \tag{296}$$

Since $V^{-1}V = I$, where I denotes the identity matrix, the rows $\boldsymbol{w}_k^\top$ of V^{-1} must be orthonormal to the columns $\boldsymbol{v}_j$ of V, i.e.,

$$\boldsymbol{w}_k \cdot \boldsymbol{v}_j = \delta_{kj} = \begin{cases} 1, & k = j, \\ 0, & \text{otherwise.} \end{cases} \tag{297}$$

Moreover, since the n eigenvectors are linearly independent, any other vector $\boldsymbol{x} \in \mathbb{R}^n$ can be expressed as

$$\boldsymbol{x} = c_1 \boldsymbol{v}_1 + \cdots + c_n \boldsymbol{v}_n = \sum_{i=1}^{n} c_i \boldsymbol{v}_n \tag{298}$$

for some real constants $c_1, \dots, c_n$. Acting with A on $\boldsymbol{x}$ gives

$$A\boldsymbol{x} = A(c_1\boldsymbol{v}_1 + \cdots + c_n\boldsymbol{v}_n) = c_1\lambda_1\boldsymbol{v}_1 + \cdots + c_n\lambda_n\boldsymbol{v}_n = \sum_{i=1}^{n} c_i\lambda_i\boldsymbol{v}_n \tag{299}$$

Now consider the matrix $V\Lambda V^{-1}$ and let it act on $\boldsymbol{x}$

$$(V\Lambda V^{-1})\boldsymbol{x} = (V\Lambda)\begin{pmatrix}\boldsymbol{w}_1^\top \\ \boldsymbol{w}_2^\top \\ \vdots \\ \boldsymbol{w}_n^\top\end{pmatrix}(c_1\boldsymbol{v}_1 + \cdots + c_n\boldsymbol{v}_n) \overset{(297)}{=} (V\Lambda)\begin{pmatrix}c_1 \\ c_2 \\ \vdots \\ c_n\end{pmatrix} = V\begin{pmatrix}\lambda_1 c_1 \\ \lambda_2 c_2 \\ \vdots \\ \lambda_n c_n\end{pmatrix}$$

Still working out the last product on the rhs., we find

$$(V\Lambda V^{-1})\boldsymbol{x} = \begin{pmatrix}\boldsymbol{v}_1 & \boldsymbol{v}_2 & \dots & \boldsymbol{v}_n\end{pmatrix}\begin{pmatrix}\lambda_1 c_1 \\ \lambda_2 c_2 \\ \vdots \\ \lambda_n c_n\end{pmatrix} = \sum_{i=1}^{n} c_i\lambda_i\boldsymbol{v}_i. \tag{300}$$

Comparing this result with Eq. (299), we see that is A and $V\Lambda V^{-1}$ map $\boldsymbol{x}$ to the same vector! Since this is true for all vectors $\boldsymbol{x}$, we can conclude that

$$A = V\Lambda V^{-1} \tag{301}$$

This important and practically useful result is called *spectral decomposition*[29] of A. For example, the spectral decomposition can used for image compression by approximating an image by keeping only the dominant eigenvalue contributions. The usefulness of Eq. (301) becomes also evident if one considers matrix functions, as we shall see now.

21.2 Matrix exponentials

The exponential function of scalar a can be defined by the infinite series

$$\exp(a) := 1 + \frac{a}{1!} + \frac{a^2}{2!} + \frac{a^3}{3!} + \ldots = \sum_{k=0}^{\infty} \frac{a^k}{k!} \tag{302}$$

This definition generalizes naturally to $n \times n$ matrices

$$\exp(B) := I + \frac{B}{1!} + \frac{B^2}{2!} + \frac{B^3}{3!} + \ldots = \sum_{k=0}^{\infty} \frac{B^k}{k!} \tag{303}$$

where $B^k = BB \cdot B$ is the $n \times n$ matrix obtained by multiplying B with itself k times. It is then obvious that $\exp(B)$ also is a $n \times n$ matrix. Warning: Note that $\exp(B + C) = \exp(B)\exp(C)$ only if $BC = CB$.

[29] There exists an analogous result for non-square matrices called *singular value decomposition*, which plays a key role in modern data analysis.

In general, it is difficult to compute $\exp(B)$ directly for an arbitrary matrix B. Analytical calculations become possible by using (301) and noting that

$$B^k = BB \cdot B = (V\Lambda V^{-1})(V\Lambda V^{-1})\ldots(V\Lambda V^{-1}) = V\Lambda^k V^{-1}, \tag{304}$$

where

$$\Lambda^k = \begin{pmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_n \end{pmatrix}^k = \begin{pmatrix} \lambda_1^k & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2^k & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_n^k \end{pmatrix}. \tag{305a}$$

We thus find that

$$\exp(B) = V\left(I + \frac{\Lambda}{1!} + \frac{\Lambda^2}{2!} + \frac{\Lambda^3}{3!} + \ldots\right)V^{-1} = V\exp(\Lambda)V^{-1}$$

or, equivalently,

$$\exp(B) = V \begin{pmatrix} \exp(\lambda_1) & 0 & 0 & \cdots & 0 \\ 0 & \exp(\lambda_2) & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \exp(\lambda_n) \end{pmatrix} V^{-1} \tag{306}$$

The same reasoning can be applied to any matrix function $F(B)$ that is defined in terms of a power series.

As an example, let's exponentiate the matrix

$$P = \begin{pmatrix} 0 & \theta \\ -\theta & 0 \end{pmatrix} \tag{307}$$

which has eigenvalues and eigenvectors

$$\lambda_\pm = \pm i\theta\ , \qquad \boldsymbol{v}_\pm = \begin{pmatrix} \mp i \\ 1 \end{pmatrix} \tag{308}$$

We then have

$$\Lambda = \begin{pmatrix} i\theta & 0 \\ 0 & -i\theta \end{pmatrix}\ , \qquad V = \begin{pmatrix} -i & i \\ 1 & 1 \end{pmatrix}\ , \qquad V^{-1} = \frac{1}{2}\begin{pmatrix} i & 1 \\ -i & 1 \end{pmatrix} \tag{309}$$

Then

$$\begin{aligned} \exp(P) = V\exp(\Lambda)V^{-1} &= \begin{pmatrix} -i & i \\ 1 & 1 \end{pmatrix}\begin{pmatrix} e^{i\theta} & 0 \\ 0 & e^{-i\theta} \end{pmatrix}\frac{1}{2}\begin{pmatrix} i & 1 \\ -i & 1 \end{pmatrix} \\ &= \frac{1}{2}\begin{pmatrix} -ie^{i\theta} & ie^{-i\theta} \\ e^{i\theta} & e^{-i\theta} \end{pmatrix}\begin{pmatrix} i & 1 \\ -i & 1 \end{pmatrix} \\ &= \frac{1}{2}\begin{pmatrix} -ie^{i\theta} & ie^{-i\theta} \\ e^{i\theta} & e^{-i\theta} \end{pmatrix}\begin{pmatrix} i & 1 \\ -i & 1 \end{pmatrix} \\ &= \frac{1}{2}\begin{pmatrix} e^{i\theta} + e^{-i\theta} & -i(e^{i\theta} - e^{-i\theta}) \\ i(e^{i\theta} - e^{-i\theta}) & e^{i\theta} + e^{-i\theta} \end{pmatrix}. \end{aligned}$$

Now, using Euler's formula, we see that

$$\exp(P) = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \tag{310}$$

is a rotation matrix!

21.3 Fundamental matrix

Let's now return to our original problem of solving the matrix differential equation $\dot{\boldsymbol{x}} = A\boldsymbol{x}$. Setting $B = tA$, we find that

$$\begin{aligned} \frac{d}{dt}\exp(tA) = \frac{d}{dt}\sum_{k=0}^{\infty}\frac{(tA)^k}{k!} = \frac{d}{dt}\sum_{k=0}^{\infty}\frac{t^kA^k}{k!} &= \sum_{k=1}^{\infty}\frac{kt^{k-1}A^k}{k!} \\ &= \sum_{k=1}^{\infty}\frac{t^{k-1}A^k}{(k-1)!} \\ &= A\sum_{k=1}^{\infty}\frac{t^{k-1}A^{k-1}}{(k-1)!} \\ &= A\sum_{k=0}^{\infty}\frac{(tA)^k}{k!} = A\exp(tA) \end{aligned}$$

We thus see that

$$\boldsymbol{x}(t) = \exp(tA)\boldsymbol{c} \tag{311}$$

solves Eq. (292) for the initial condition $\boldsymbol{x}(0) = \boldsymbol{c}$. This is the matrix generalization of the scalar exponential solution formula (291b). The matrix $\exp(tA)$ is often called the *fundamental matrix* of the system $\dot{\boldsymbol{x}} = A\boldsymbol{x}$.

To relate Eq. (311) to the eigenvector representation of system solutions found earlier, we express $\boldsymbol{c}$ in the eigenvector basis

$$\boldsymbol{c} = \sum_{i=1}^{n} c_i\boldsymbol{v}_i$$

and use Eq. (306) to write

$$
\begin{aligned}
\boldsymbol{x}(t) = V\exp(t\Lambda)V^{-1}\boldsymbol{c} &= V\begin{pmatrix} e^{t\lambda_1} & 0 & 0 & \cdots & 0 \\ 0 & e^{t\lambda_2} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & e^{t\lambda_n} \end{pmatrix} V^{-1}\sum_{i=1}^{n} c_i \boldsymbol{v}_i \\
&= V\begin{pmatrix} e^{t\lambda_1} & 0 & 0 & \cdots & 0 \\ 0 & e^{t\lambda_2} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & e^{t\lambda_n} \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix} \\
&= \begin{pmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \cdots & \boldsymbol{v}_n \end{pmatrix} \begin{pmatrix} e^{t\lambda_1}c_1 \\ e^{t\lambda_2}c_2 \\ \vdots \\ e^{t\lambda_n}c_n \end{pmatrix} \\
&= \sum_{i=1}^{n} c_i e^{t\lambda_i}\boldsymbol{v}_i. \qquad (312)
\end{aligned}
$$

This is indeed the n-dimensional generalization of our earlier result (217) for 2×2 matrices.

22 Solving inhomogeneous systems

Thus far we have focussed on solving homogeneous systems

$$\dot{\boldsymbol{x}} = A\boldsymbol{x} \qquad (313)$$

where A is a constant $n \times n$-matrix. We now turn our attention to *inhomogeneous systems* of the form

$$\dot{\boldsymbol{x}} = A\boldsymbol{x} + \boldsymbol{q}(t), \qquad (314)$$

where $\boldsymbol{q}(t)$ is a n-dimensional vector of (possibly constant) functions

$$\boldsymbol{q}(t) = \begin{pmatrix} q_1(t) \\ \vdots \\ q_n(t) \end{pmatrix}.$$

In general, the matrix A will couple the components of $\boldsymbol{x} = (x_1(t), \ldots, x_n(t))$, which makes it often more difficult to solve Eq. (314) directly. However, as we shall see shortly, if the matrix A is diagonalizable, so that

$$A = V\Lambda V^{-1} \qquad (315)$$

with a diagonal matrix Λ, then we can rewrite (314) in terms of new variables $\boldsymbol{y} = (y_1(t), \ldots, y_n(t))$ such that the equations for $\dot{\boldsymbol{y}}$ becomes *decoupled*; that is, $\dot{y}_1(t)$ only depends on $y_1(t)$, and $\dot{y}_2(t)$ only depends on $y_2(t)$, and so on. How can we tell if a matrix A

is diagonalizable? A sufficient condition is that the matrix A is *symmetric*, which means that A equals its transpose $A^\top$

$$A = A^\top \qquad \Leftrightarrow \qquad A_{ij} = A_{ji}. \tag{316}$$

In this case, we have the following

Theorem.[30] For a symmetric matrix with real number entries, the eigenvalues are real numbers and it's possible to choose a complete set of eigenvectors that are perpendicular (or even orthonormal).

More generally, if A has n independent eigenvectors, and we can construct V from the eigenvectors as discussed in the previous section. If A is symmetric, then we can write $A = V\Lambda V^{-1}$, and by constructing V from the *normalized* eigenvectors of A, we have $V^{-1} = V^\top$ (we call such a matrix with orthonormal columns *orthogonal*[31]). Hence, it follows in this case that

$$A = V\Lambda V^\top. \tag{317}$$

This demonstrates that symmetric matrices A are particularly 'nice' – instead of computing the inverse V^{-1} by Gauss-Jordan elimination, we merely have to take the transpose.

[30] This theorem will be proved in 18.06.

[31] A matrix Q whose columns are orthonormal is called an *orthogonal matrix.* If Q is square, then $Q^\top Q = I$ tells us that $Q^\top = Q^{-1}$. This is one of the reasons orthonormal matrices are so important, because they are very easy to invert!

22.1 Decoupling by diagonalization

We can now summarize the steps to solve $\dot{\boldsymbol{x}} = A\boldsymbol{x} + \boldsymbol{q}(t)$ by decoupling:

1. Find the eigenvalues of A (with multiplicity), and put them in a diagonal matrix Λ.
2. Find a basis of each eigenspace.[32] Put the eigenvectors as columns of a matrix V.
3. Substitute $\boldsymbol{x} = V\boldsymbol{y}$ to get

   $$V\dot{\boldsymbol{y}} = AV\boldsymbol{y} + \boldsymbol{q}(t) \qquad \Leftrightarrow \qquad V\dot{\boldsymbol{y}} = (V\Lambda V^{-1})V\boldsymbol{y} + \boldsymbol{q}(t)$$

   and therefore (after multiplication by V^{-1})

   $$\dot{\boldsymbol{y}} \;\; = \;\; \Lambda\boldsymbol{y} + V^{-1}\boldsymbol{q}(t). \tag{318}$$

   This is a *decoupled* system of inhomogeneous linear ODEs for $\boldsymbol{y}$.
4. Solve for each coordinate function of $\boldsymbol{y}$.
5. Compute $V\boldsymbol{y}$; the result is the solution $\boldsymbol{x}$.

22.2 Variation of parameters

We still discuss an alternative method for solving inhomogeneous systems. Long ago we learned how to use variation of parameters to solve inhomogeneous linear ODEs

$$\dot{y} + p(t)y = q(t).$$

Now we're going to use the same idea to solve an inhomogeneous linear *system* of ODEs such as

$$\dot{\boldsymbol{x}} = A\boldsymbol{x} + \boldsymbol{q}, \tag{319}$$

where $\boldsymbol{q}$ is a vector-valued function of t. To this end, we first write the solution of the corresponding homogeneous system

$$\dot{\boldsymbol{x}} = A\boldsymbol{x},$$

in terms of the fundamental[33] matrix $X = e^{At}$:

$$\boldsymbol{x}(t) = e^{At}\boldsymbol{c}. \tag{320}$$

Analogous to the variation-of-parameters method in the the scalar, case, we now try to find a particular solution of the inhomogeneous system' (319) by replacing the constant vector $\boldsymbol{c}$

[32] If the total number of independent eigenvectors found is less than n, then a more complicated method (not discussed here) is required.

[33] In general, the fundamental matrix is not unique as one could in principle pick any basis of solutions and combine them into a fundamental matrix X to express the homogeneous solutions as $\boldsymbol{x}(t) = X\boldsymbol{c}$ for some constant vector c encoding initial conditions. For our purposes, however, it is sufficient to consider $X = e^{At}$ as a specific choice for the fundamental matrix.

through a vector-valued function $\boldsymbol{u}(t)$. Inserting the ansatz $\boldsymbol{x}(t) = e^{At}\boldsymbol{u}(t)$ into the original system. Noting that

$$\dot{\boldsymbol{x}}(t) = Ae^{At}\boldsymbol{u}(t) + e^{At}\dot{\boldsymbol{u}}(t) = A\boldsymbol{x} + e^{At}\dot{\boldsymbol{u}}(t)$$

we have

$$e^{At}\dot{\boldsymbol{u}}(t) = \boldsymbol{q}(t) \qquad \Rightarrow \qquad \dot{\boldsymbol{u}}(t) = e^{-At}\boldsymbol{q}(t).$$

Integrating this with respect to t, we find

$$\boldsymbol{x}(t) = e^{At}\boldsymbol{u}(t) = e^{At}\int e^{-At}\boldsymbol{q}(t)\,dt. \tag{321}$$

Note that the indefinite integral on the rhs. gives a constant vector $\boldsymbol{c}$, so that we have enough free parameters to satisfy initial conditions.

22.3 Example: Inhomogeneous heated rod

Let us practice variation of parameters using the heated rod example. In this problem, we consider the four thermometers, labeled by $i = 0, 1, 2, 3$ placed at equal distances along a thin insulated metal rod. One end is held at 0 degrees Celsius and the other is held at 100 degrees Celsius, so that the first and last thermometer always show

$$T_0 = 0\ , \qquad T_3 = 100.$$

We are interested in the temperatures $T_1(t)$ and $T_2(t)$ of the two middle thermometers. These are determined by

$$\dot{T}_1 = \kappa(T_0 - 2T_1 + T_2)\ , \qquad \dot{T}_2 = \kappa(T_1 - 2T_2 + T_3). \tag{322}$$

where κ is the thermal conductivity. To keep formulas simple, let us assume $\kappa = 1$ min^{-1} from now on. Equations (322) formalize the physical law that, in good approximation, the temperature change of the thermometers is proportional to the (discretized) second space derivative of the temperature. We can rewrite these equations as a system

$$\frac{d}{dt}\begin{pmatrix} T_1 \\ T_2 \end{pmatrix} = \begin{pmatrix} -2 & 1 \\ 1 & -2 \end{pmatrix}\begin{pmatrix} T_1 \\ T_2 \end{pmatrix} + \begin{pmatrix} T_0 \\ T_3 \end{pmatrix} \tag{323}$$

The coefficient matrix A has eigenvalues $\lambda_1 = -1$ and $\lambda_3 = -3$ with normalized eigenvectors

$$\boldsymbol{v}_1 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \end{pmatrix}\ , \qquad \boldsymbol{v}_2 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & -1 \end{pmatrix},$$

which we can combine into the 2×2 matrix $V = (\boldsymbol{v}_1 \quad \boldsymbol{v}_2)$. Since the coefficient matrix A is symmetric, we have $V^{-1} = V^\top$, and we can write the matrix exponential as

$$e^{At} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} e^{-t} & 0 \\ 0 & e^{-3t} \end{pmatrix}\frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = e^{-2t}\begin{pmatrix} \cosh t & \sinh t \\ \sinh t & \cosh t \end{pmatrix}. \tag{324}$$

The inverse is given by

$$e^{-At} = e^{2t} \begin{pmatrix} \cosh t & -\sinh t \\ -\sinh t & \cosh t \end{pmatrix}. \tag{325}$$

According to Eq. (321), the solution is given by

$$\boldsymbol{T}(t) \;=\; e^{At} \int e^{-At} \begin{pmatrix} T_0 \\ T_3 \end{pmatrix} dt \tag{326}$$

Noting that

$$\int e^{-At} \begin{pmatrix} T_0 \\ T_3 \end{pmatrix} dt = \int \begin{pmatrix} -T_3 e^{2t} \sinh t \\ T_3 e^{2t} \cosh t \end{pmatrix} dt = T_3 \frac{e^t}{6} \begin{pmatrix} 3 - e^{2t} \\ 3 + e^{2t} \end{pmatrix} + \boldsymbol{c} \tag{327}$$

We have

$$\boldsymbol{T}(t) = e^{-2t} \begin{pmatrix} \cosh t & \sinh t \\ \sinh t & \cosh t \end{pmatrix} \left[T_3 \frac{e^t}{6} \begin{pmatrix} 3 - e^{2t} \\ 3 + e^{2t} \end{pmatrix} + \boldsymbol{c} \right], \tag{328}$$

which for large times $t \to \infty$ approaches

$$\boldsymbol{T}(t) = T_3 \begin{pmatrix} 1/3 \\ 2/3 \end{pmatrix}. \tag{329}$$

This is exactly as we could have predicted. The temperature diffuses no matter what the initial condition until it is linearly distributed across the bar from $T_0 = 0$ to $T_3 = 100$. The temperatures at the two internal thermometers measuring being 1/3 and 2/3 of the difference in temperature between the two end points.

23 Introduction to Fourier series

23.1 Periodic functions

Because

$$\sin(t + 2\pi) = \sin t, \quad \cos(t + 2\pi) = \cos t,$$

hold for all t, the functions $\sin t$ and $\cos t$ are called periodic with period 2π. In general a function $f(t)$ defined for all real t is *periodic of period P* if

$$f(t + P) = f(t) \qquad\qquad \forall\, t. \tag{330}$$

There are many such functions beyond the sinusoidal functions. To construct one, divide the real line into intervals of length P, start with any function defined on one such interval $[t_0, t_0 + P)$, and then copy its values in the other intervals. The entire graph consists of horizontally shifted copies of the width P graph. For the remainder of this section we assume $P = 2\pi$ and consider the interval $t \in [-\pi, \pi)$.

Is $\sin 3t$ periodic of period 2π? The *shortest* period is $2\pi/3$, but $\sin 3t$ is also periodic with period any positive integer multiple of $2\pi/3$, including $3(2\pi/3) = 2\pi$:

$$\sin(3(t + 2\pi)) = \sin(3t + 6\pi) = \sin 3t.$$

So the answer is yes.

Another example of a periodic function, which we will frequently refer to in this section is the *square wave*, defined by

$$\mathrm{Sq}(t) := \begin{cases} -1 & \text{if } -\pi < t < 0, \\ 1, & \text{if } 0 < t < \pi \end{cases}$$

and extended to a periodic function of period 2π. The function $\mathrm{Sq}(t)$ has jump discontinuities,[34] for example at $t = 0$. The graph is usually drawn with vertical segments at the jumps (even though this violates the vertical line test). Below, we will see that

$$\mathrm{Sq}(t) = \frac{4}{\pi}\left(\sin t + \frac{\sin 3t}{3} + \frac{\sin 5t}{5} + \cdots\right).$$

23.2 Fourier series

A *Fourier series* is a linear combination of the infinitely many functions $\cos nt$ and $\sin nt$ as n ranges over integers:

$$\begin{aligned} f(t) \quad = \quad & \frac{a_0}{2} + a_1 \cos t + a_2 \cos 2t + a_3 \cos 3t + \cdots && (331) \\ & + b_1 \sin t + b_2 \sin 2t + b_3 \sin 3t + \cdots && (332) \end{aligned}$$

Terms like $\cos(-2t)$ are redundant since $\cos(-2t) = \cos 2t$. Also $\sin 0t = 0$ produces nothing new. But $\cos 0t = 1$ is included. We'll explain later why we write a_0 times $1/2$ instead of times 1. In sum-notation, we can write the Fourier series as

$$f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos nt + \sum_{n=1}^{\infty} b_n \sin nt.$$

Recall that, for example, $\sum_{n=1}^{\infty} b_n \sin nt$ means the sum of the series whose nth term is obtained by plugging in the positive integer n into the expression $b_n \sin nt$, so

$$\sum_{n\geq 1} b_n \sin nt = b_1 \sin t + b_2 \sin 2t + b_3 \sin 3t + \cdots .$$

Any Fourier series as above is, by construction, periodic of period 2π. Later we will extend to the definition of Fourier series to include functions of other periods. The numbers a_n and b_n are called the *Fourier coefficients* of f. Each summand, such as $a_0/2$, $a_n \cos nt$, or $b_n \sin nt$ is called a *Fourier component* of f.

A famous and practically very important theorem is

Fourier's Theorem. 'Every' periodic function f of period 2π 'is' a Fourier series, and the Fourier coefficients are uniquely determined by f.

The word 'Every' has to be taken with a grain of salt: The function has to be 'reasonable'.

[34] If you must define $\mathrm{Sq}(0)$, compromise between the upper and lower values by setting $\mathrm{Sq}(0) := 0$.

Piecewise differentiable functions with jump discontinuities are reasonable, as are virtually all other functions that arise in physical applications. Similarly, the word 'is' has to be taken with a grain of salt: If f has a jump discontinuity at τ, then the Fourier series might disagree with f there; the value of the Fourier series at τ is always the average of the left limit $f(\tau^-)$ and the right limit $f(\tau^+)$, regardless of the actual value of $f(\tau)$.

The statement of the theorem means that the functions

$$1, \cos t, \cos 2t, \cos 3t, \ldots, \sin t, \sin 2t, \sin 3t, \ldots$$

form a basis for the vector space of 'all' periodic functions of period 2π. The Fourier coefficients are then the coordinates of the function f with respect to this basis.

23.3 Scalar product for real-valued functions

Given f, how do we find the Fourier coefficients a_n and b_n? That is, how do we find the coordinates of f with respect to the basis of cosines and sines? Let's recall how this works for ordinary vectors. If $\boldsymbol{v}$ and $\boldsymbol{w}$ are vectors in $\mathbb{R}^n$, then their scalar product is defined by

$$\langle \boldsymbol{v}, \boldsymbol{w} \rangle := \boldsymbol{v} \cdot \boldsymbol{w} := \sum_{i=1}^{n} v_i w_i.$$

The coordinates w_i of the vector $\boldsymbol{w}$ with respect to some orthonormal basis $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\}$ of $\mathbb{R}^n$ are then obtained by forming the scalar products of $\boldsymbol{w}$ with all the $\boldsymbol{e}_i$

$$w_i = \langle \boldsymbol{w}, \boldsymbol{e}_i \rangle. \tag{333}$$

This concept translates directly to functions: If f and g are real-valued periodic functions with period 2π, then their *scalar (or inner) product* is

$$\langle f, g \rangle := \int_{-\pi}^{\pi} f(t) g(t)\, dt.$$

For example, by definition,

$$\langle 1, \cos t \rangle = \int_{-\pi}^{\pi} \cos t\, dt = 0.$$

Thus, the functions 1 and $\cos t$ are orthogonal. In fact, calculating all the inner products shows that

$$1, \cos t, \cos 2t, \cos 3t, \ldots, \sin t, \sin 2t, \sin 3t, \ldots \tag{334}$$

is an *orthogonal* basis! Is it an orthonormal basis? The answer is no, since

$$\langle 1, 1 \rangle = \int_{-\pi}^{\pi} 1\, dt = 2\pi \neq 1. \tag{335}$$

Let's try to compute

$$\langle \sin t, \sin t \rangle = \int_{-\pi}^{\pi} \sin^2 t\, dt \tag{336a}$$

and

$$\langle \cos t, \cos t\rangle = \int_{-\pi}^{\pi} \cos^2 t\, dt \tag{336b}$$

Since $\cos t$ is just a shift of $\sin t$, the integrals are going to be the same. Also, the two integrals can be added up to give

$$\int_{-\pi}^{\pi} (\sin^2 t + \cos^2 t) dt = \int_{-\pi}^{\pi} dt = 2\pi, \tag{337}$$

so each is π. The same idea works to show that

$$\langle \cos nt, \cos nt\rangle = \pi \ , \qquad \langle \sin nt, \sin nt\rangle = \pi \tag{338}$$

for each positive integer n.

23.4 Fourier coefficient formulas

We can now provide formulas for the coefficients a_n and b_n such that

$$\begin{aligned} f(t) \quad = \quad & \frac{a_0}{2} + a_1 \cos t + a_2 \cos 2t + a_3 \cos 3t + \cdots \\ & + b_1 \sin t + b_2 \sin 2t + b_3 \sin 3t + \cdots \end{aligned} \tag{339}$$

By the shortcut formulas (338),

$$a_n = \frac{\langle f, \cos nt\rangle}{\langle \cos nt, \cos nt\rangle} = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t) \cos nt\, dt,$$

and the coefficient of 1 is

$$\frac{a_0}{2} = \frac{\langle f, 1\rangle}{\langle 1, 1\rangle} = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(t)\, dt.$$

so

$$a_0 = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t) \cos 0t\, dt.$$

Using $a_0/2$ in the series ensures that the formula for a_n for $n > 0$ works also for $n = 0$. A similar formula holds for b_n.

To summarize: Given f, its Fourier coefficients can be calculated as follows:

$$a_n \quad = \quad \frac{1}{\pi}\int_{-\pi}^{\pi} f(t) \cos nt\, dt \ , \qquad \forall\, n \geq 0 \tag{340a}$$

$$b_n \quad = \quad \frac{1}{\pi}\int_{-\pi}^{\pi} f(t) \sin nt\, dt \ , \qquad \forall\, n \geq 1 \tag{340b}$$

In particular, the constant term of the Fourier series of f is

$$\frac{a_0}{2} = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(t)\, dt,$$

which is the average value of f on $(-\pi, \pi)$.

23.5 Even and odd symmetry

Before starting to compute Fourier coefficients for a function $f(t)$, you should always check whether f is symmetric. Recall that a function $f(t)$ is *even* if $f(-t) = f(t)$ for all t and *odd* if $f(-t) = -f(t)$ for all t. If

$$f(t) \quad = \quad \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos nt + \sum_{n=1}^{\infty} b_n \sin nt,$$

then substituting $-t$ for t gives

$$f(-t) \quad = \quad \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos nt + \sum_{n=1}^{\infty} (-b_n) \sin nt.$$

The right hand sides match if and only if $b_n = 0$ for all n. This means that Fourier series of an even function f has only cosine terms (including the constant term):

$$f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos nt. \tag{341a}$$

Similarly, the Fourier series of an odd function f has only sine terms:

$$f(t) = \sum_{n=1}^{\infty} b_n \sin nt. \tag{341b}$$

For example, the square wave $\mathrm{Sq}(t)$ is an odd function, so

$$\mathrm{Sq}(t) = \sum_{n=1}^{\infty} b_n \sin nt$$

for some numbers b_n. The Fourier coefficient formula says

$$\begin{aligned}
b_n &= \frac{1}{\pi} \int_{-\pi}^{\pi} \underbrace{\mathrm{Sq}(t) \sin nt}_{\text{even}} \, dt \\
&= \frac{2}{\pi} \int_0^{\pi} \mathrm{Sq}(t) \sin nt \, dt \qquad \text{(by symmetry)} \\
&= \frac{2}{\pi} \int_0^{\pi} \sin nt \, dt \qquad (\text{since } \ \mathrm{Sq}(t) = 1 \text{ whenever } \ 0 < t < \pi) \\
&= \left. \frac{2(-\cos nt)}{\pi n} \right|_0^{\pi} \\
&= \frac{2}{\pi n}(-\cos n\pi + \cos 0) \\
&= \begin{cases} \dfrac{4}{\pi n}, & \text{if } n \text{ is odd,} \\ 0, & \text{if } n \text{ is even.} \end{cases}
\end{aligned}$$

Thus

$$b_1 = \frac{4}{\pi}, \quad b_3 = \frac{4}{3\pi}, \quad b_5 = \frac{4}{5\pi}, \ldots$$

and all other Fourier coefficients are 0.

$$\mathrm{Sq}(t) = \frac{4}{\pi}\left(\sin t + \frac{\sin 3t}{3} + \frac{\sin 5t}{5} + \cdots\right).$$

23.6 Gibbs' Phenomenon

You might be wondering how we can use Fourier series, series composed of continuous functions, to approximate a discontinuous functions. While it's true that we can get a Fourier series that converges to a function with discontinuities, *the convergence is very slow* (everywhere, but particularly near the discontinuities). Something very curious happens near the points where the target function is discontinuous. Namely, the *partial sums* of the Fourier series overshoot and undershoot the values of the function at these points. This is known as the *Gibbs' Phenomenon.*

We illustrate this phenomenon using the example of the square wave of height $\pi/4$ — which we denote by $S(t) = \frac{\pi}{4}\mathrm{Sq}(t)$. In this case the discontinuities are the integer multiples of π. The Fourier series for $S(t)$ is

$$\begin{aligned} S(t) &= \frac{\pi}{4}\mathrm{Sq}(t) = \sin t + \frac{1}{3}\sin 3t + \frac{1}{5}\sin 5t + \cdots \\ &= \sum_{n \mathrm{odd}}^{\infty} \frac{1}{n}\sin nt. \end{aligned} \tag{342}$$

For an odd positive integer N, write $S_N(t)$ for the N^{th} *partial sum*

$$S_N(t) := \sum_{n \mathrm{odd}}^{N} \frac{1}{n}\sin nt. \tag{343}$$

Then $S(t)$ is the limit of the functions $S_N(t)$ as $N \to \infty$. It turns out that it is possible to do a calculation that's a little bit beyond the scope of this class that for N very large, S_N overshoots $\pi/4$ (or undershoots $-\pi/4$) at a point of discontinuity by about $\frac{\pi}{2}\cdot(0.089490\ldots)$, or about 9%. This overshooting and undershooting can easily be seen in the figures below.

This overshoot/undershoot by about 9% happens *always*, not just for the case of $S(t)$, but for any function with discontinuities. In practice, one can only generate approximate Fourier series (with only a finite number of Fourier coefficients, known with some error). Thus the poor convergence of Fourier series for functions with discontinuities – and, particularly the Gibb's phenomenon, creates (serious) difficulties in areas such as signal processing.

The main mathematical intuition behind this is that approximating a discontinuous function with a *finite* sum of continuous functions, such as each sucessive S_N, always gives a continuous function, so that continuous function will act poorly at the point of discontinuity (if the discontinuity isn't a removable discontinuity). A confusing point is that even though for N arbitrarily large, the function S_N acts poorly at the discontinuity, the series $\sum_{n \mathrm{odd}}^{\infty} \frac{1}{n}\sin nt$ doesn't. This illustrates how limits can act strangely sometimes, as well as the fact that the convergence of $S_N(t)$ to $S(t)$ 'doesn't happen in the best possible way'. If you're interested in questions like this, you might find 18.311 or 18.100 interesting.

24 Fourier series of arbitrary period

24.1 Functions of arbitrary period

Everything we did with periodic functions of period 2π can be generalized to periodic functions of other periods. Let's for example consider the 'stretched' square wave, defined by

$$f(t) := \begin{cases} 1, & 0 < t < L, \\ -1 & -L < t < 0. \end{cases} \tag{344}$$

and extended to a periodic function of period $2L$. We can express this new square wave $f(t)$ in terms of the original square wave function $\mathrm{Sq}(u)$ by setting

$$u = \frac{\pi}{L} t$$

so that $u = \pi$ corresponds to $t = L$. Then $f(t) = \mathrm{Sq}(u)$ or explicitly

$$f(t) = \mathrm{Sq}\left(\frac{\pi t}{L}\right). \tag{345}$$

Similarly, we can stretch any function of period 2π to get a function of different period. Let L be a positive real number. Start with "any" periodic function

$$g(u) = \frac{a_0}{2} + \sum_{n\geq 1} a_n \cos nu + \sum_{n\geq 1} b_n \sin nu,$$

of period 2π. Stretching horizontally by a factor L/π gives a periodic function $f(t)$ of period $2L$, and 'every' f of period $2L$ arises this way. By the same calculation as above,

$$\begin{aligned} f(t) &= g\left(\frac{\pi t}{L}\right) \\ &= \frac{a_0}{2} + \sum_{n\geq 1} a_n \cos \frac{n\pi t}{L} + \sum_{n\geq 1} b_n \sin \frac{n\pi t}{L}. \end{aligned} \tag{346}$$

The substitution $u = \pi t/L$, which implies $du = (\pi/L)\, dt$, also leads to Fourier coefficient formulas for period $2L$:

$$\begin{aligned} a_n &= \frac{1}{\pi} \int_{-\pi}^{\pi} g(u) \cos nu \, du \\ &= \frac{1}{\pi} \int_{-L}^{L} g\left(\frac{\pi t}{L}\right) \cos\left(\frac{n\pi t}{L}\right) \frac{\pi}{L}\, dt \\ &= \frac{1}{L} \int_{-L}^{L} f(t) \cos\left(\frac{n\pi t}{L}\right) dt. \end{aligned} \tag{347}$$

Similarly, we find

$$b_n = \frac{1}{L} \int_{-L}^{L} f(t) \sin\left(\frac{n\pi t}{L}\right) dt. \tag{348}$$

24.2 Inner product for periodic functions of period $2L$

We can adapt the definition of the inner product to the case of functions f and g of period $2L$ by defining

$$\langle f, g \rangle := \int_{-L}^{L} f(t)g(t)\,dt, \tag{349}$$

which reduces to our our earlier definition for 2π-periodic functions, when we set $L = \pi$. The same calculations as before show that the functions

$$1, \cos\frac{\pi t}{L}, \cos\frac{2\pi t}{L}, \cos\frac{3\pi t}{L}, \dots, \sin\frac{\pi t}{L}, \sin\frac{2\pi t}{L}, \sin\frac{3\pi t}{L}, \dots$$

form an orthogonal basis for the vector space of 'all' periodic functions of period $2L$, with[35]

$$\langle 1, 1\rangle = 2L\ , \qquad \left\langle \cos\frac{n\pi t}{L}, \cos\frac{n\pi t}{L} \right\rangle = L\ , \qquad \left\langle \sin\frac{n\pi t}{L}, \sin\frac{n\pi t}{L} \right\rangle = L. \tag{350}$$

As another example of how to work with scalar (inner) products of functions, let $T(t)$ be the *even* periodic function of period 2 such that $T(t) = |t|$ for $-1 \le t \le 1$; this is called a *triangle wave*. Let $\mathrm{Sq}(x)$ denote the *odd* square wave of period 2. We then find that

$$\langle T, \mathrm{Sq}\rangle \;=\; \int_{-1}^{1} |x|\mathrm{Sq}(x)dx \;=\; 0.$$

Hence, the triangle wave and square wave are orthogonal on the interval $[-1, 1]$.

24.3 Summary

Fourier's theorem states that every sufficiently regular periodic function f of period $2L$ is a Fourier series

$$f(t) = \frac{a_0}{2} + \sum_{n\ge1} a_n \cos\frac{n\pi t}{L} + \sum_{n\ge1} b_n \sin\frac{n\pi t}{L}. \tag{351}$$

Given f, the Fourier coefficients a_n and b_n can be computed using:

$$a_n \;=\; \frac{1}{L}\int_{-L}^{L} f(t)\cos\frac{n\pi t}{L}\,dt \qquad \text{for all }\ n \ge 0, \tag{352a}$$

$$b_n \;=\; \frac{1}{L}\int_{-L}^{L} f(t)\sin\frac{n\pi t}{L}\,dt \qquad \text{for all }\ n \ge 1. \tag{352b}$$

It is important to keep in mind that

- If f is even, then only the cosine terms (including the $a_0/2$ term) appear.
- If f is odd, then only the sine terms appear.

[35] The average value of $\cos^2\omega t$ is 1/2 for any ω, and the average value of $\sin^2\omega t$ is 1/2 too. This gives another way to derive the Fourier coefficient formulas for functions of period $2L$.

24.4 Convergence of a Fourier series

A periodic function f of period $2L$ is called *piecewise differentiable* if

(i) there are at most finitely many points in $[-L, L)$ where $f'(t)$ does not exist,

(ii) at points where the derivative exists it is bounded, (that is $|f'(t)| \leq M < \infty$), and

(iii) at each such point τ, the left limit $f(\tau^-) := \lim_{t\to\tau^-} f(t)$ and right limit $f(\tau^+) := \lim_{t\to\tau^+} f(t)$ exist (although they might be unequal, in which case we say that f has a *jump discontinuity* at τ).

Theorem. If f is a piecewise differentiable periodic function, then the Fourier series of f (with the a_n and b_n defined by the Fourier coefficient formulas)

- converges to $f(t)$ at values of t where f is continuous, and
- converges to $\dfrac{f(t^-)+f(t^+)}{2}$ where f has a jump discontinuity.

For example, for the square wave we find that the left limit $\mathrm{Sq}(0^-) = -1$ and right limit $\mathrm{Sq}(0^+) = 1$ average to 0. The Fourier series

$$\frac{4}{\pi}\left(\sin t + \frac{\sin 3t}{3} + \frac{\sin 5t}{5} + \cdots\right)$$

evaluated at $t = 0$ converges to 0 too.

24.5 Differentiating a Fourier series

If a function is differentiable, you can simply differentiate its Fourier series term by term to obtain the Fourier series for the derivative. For example, the Fourier Series for the period 2π triangle wave is:

$$g(t) = \frac{\pi}{2} - \frac{4}{\pi}\left(\cos t + \frac{\cos 3t}{3^2} + \frac{\cos 5t}{5^2} + \cdots\right).$$

Differentiating the Fourier series $g(t)$ term-by-term gives

$$g'(t) = \frac{4}{\pi}\left(\sin t + \frac{\sin 3t}{3} + \frac{\sin 5t}{5} + \ldots\right),$$

which is the Fourier series of the 2π–periodic square wave!

There are many subtle, but important, questions in Fourier series that we will not cover here (but which courses such as 18.100 do, at least partially). For example: If I write some arbitrary Fourier series, how do I know if it comes from a differentiable function? If I differentiate term by term the Fourier series for a function that is not differentiable (like the square wave function), is the result the Fourier series for something?

24.6 Antiderivative of a Fourier series

Suppose that f is a piecewise differentiable periodic function, and that F is an antiderivative[36] of f. The function F might not be periodic. For example, if f is a function of period 2 such that

$$f(t) := \begin{cases} 2, & \text{if } 0 < t < 1, \\ -1 & \text{if } -1 < t < 0, \end{cases}$$

then $F(t)$ creeps upward over time. An even easier example: if $f(t) = 1$, then $F(t) = t + C$ for some C, so $F(t)$ is not periodic. But if the constant term $a_0/2$ in the Fourier series of f is 0, then F *is* periodic, and its Fourier series can be obtained by taking the simplest antiderivative of each cosine and sine term, and adding an overall $+C$, where C is the average value of F.

As another example, let $T(t)$ be the triangle wave of period 2 and amplitude 1: so that $T(t) = |t|$ for $-1 \leq t \leq 1$. To find the Fourier series of $T(t)$, we could use the Fourier coefficient formula. But instead, notice that $T(t)$ has slope -1 on $(-1, 0)$ and slope 1 on $(0, 1)$, so $T(t)$ is an antiderivative of the period 2 square wave

$$\mathrm{Sq}(\pi t) = \sum_{n \geq 1 \mathrm{odd}} \frac{4}{n\pi} \sin n\pi t.$$

Taking an antiderivative termwise (and using that the average value of $T(t)$ is $1/2$) gives

$$\begin{aligned} T(t) &= \frac{1}{2} + \sum_{n \geq 1 \text{ odd}} \frac{4}{n\pi} \left(\frac{-\cos n\pi t}{n\pi} \right) \\ &= \frac{1}{2} - \sum_{n \geq 1 \text{ odd}} \frac{4}{n^2\pi^2} \cos n\pi t. \end{aligned} \tag{353}$$

Warning: If a periodic function f is not continuous, it will not be an antiderivative of any piecewise differentiable function, so you cannot find the Fourier series of f by integration.

Remark. The Fourier series for a function with discontinuities can (formally) be differentiated term by term, but *the result will not converge.* For example, the termwise derivative of the Fourier series for $\mathrm{Sq}(t)$ is

$$\frac{4}{\pi} \sum_{n \text{ odd}} \cos(nt).$$

This does not converge anywhere (since the nth term does not even vanish as $n \to \infty$). However, note that it is possible to make sense of this series, and of the anti-derivative of the square wave function, in terms of Dirac's delta functions and the theory of distributions — seen in more advanced courses than this one.

24.7 Generalizing Fourier series: The Fourier transform

You may have heard about the Fourier transform and perhaps want to know what it is. In this class, we only work with periodic functions and signals. In the real world, signals may

[36] If f has jump discontinuities, one can still define $F(t) := \int_0^t f(\tau)\, d\tau + C$, but at the jump discontinuities F will be only continuous, not differentiable.

not be periodic. Below let's generalize the method of Fourier series to non-periodic signals. This is known as the Fourier transform.

Suppose you have an electric field at a point — radio signal or pressure wave due to sound coming from a speaker. This signal changes in time. Assume the signal has some period T. Then we can write the signal in the form

$$f(t) = \sum_{n=-\infty}^{\infty} c_n e^{2\pi i n t/T},$$

where the c_n are the coefficients determined by

$$c_n = \frac{2}{T}\int_{-T/2}^{T/2} f(t) e^{-2\pi i n t/T} dt.$$

We can think of a non-periodic signal as the limit as T goes to infinity of a periodic signal of period T. As T increases, the spacing between the frequncies in our sum are approaching zero. This turns the sum into an integral in the limit, and we have the equations:

$$f(t) = \int_{-\infty}^{\infty} \widehat{f}(n) e^{2\pi i n t} dn \tag{354a}$$

$$\widehat{f}(n) = \int_{-\infty}^{\infty} f(t) e^{-2\pi i n t} dt \tag{354b}$$

We call $\widehat{f}(n)$ the *Fourier Transform* of f.

Note that the continuous function $\widehat{f}(n)$ replaces the discrete coefficients c_n. So now $f(t)$ can be composed of a continuous infinite sum (an integral) of complex sinusoids $e^{2\pi i n t}$ with the weights being given by the $\widehat{f}(n)$ function.

In practice, the above Fourier transform is often a starting point for analysis of signals. However, discrete Fourier series are sufficient approximations in many cases. However, you'll likely encounter a Fourier transform in other courses. See for example 18.103, 18.303, 18.311, 18.353, and 18.354.

25 Solving ODEs with Fourier series

25.1 Reminder: ERF for sinusoidal input

Suppose that $f(t)$ is an odd periodic function of period 2π, and we would like to find the periodic function $x(t)$ of period 2π that is a solution to

$$\ddot{x} + 50x = f(t) \tag{355}$$

We may think of $f(t)$ as the input signal, and the solution $x(t)$ as the system response (output signal). Let's consider the special case is $f(t) = \sin nt$ and try to find the particular solution $x(t)$ to

$$\ddot{x} + 50x = \sin nt$$

that has the same (smallest) period as $\sin nt$. To tackle this problem, we first complexify to find the response to e^{int}, and then take the imaginary part. The complexified problem is

$$\ddot{z} + 50z = e^{int},$$

and has the characteristic polynomial is $p(r) = r^2 + 50$. So, by ERF, the system response to e^{int} is

$$z = \frac{e^{int}}{p(in)} = \frac{e^{int}}{50 - n^2}. \tag{356}$$

This is indeed the solution we were looking for as it has the right period. The complex gain is $1/(50 - n^2)$, and

$$x(t) = \Im\left(\frac{e^{int}}{50 - n^2}\right) = \frac{\sin nt}{50 - n^2} \tag{357}$$

is the system response to $\sin nt$. This explains all the rows of the table below except the last row.

Input signal	System response
e^{int}	$\frac{1}{50-n^2} e^{int}$
$\sin nt$	$\frac{1}{50-n^2} \sin nt$
$\sin t$	$\frac{1}{49} \sin t$
$\sin 2t$	$\frac{1}{46} \sin 2t$
$\sin 3t$	$\frac{1}{41} \sin 3t$
$\vdots$	$\vdots$
$\sum_{n\geq 1} b_n \sin nt$	$\sum_{n\geq 1} \frac{b_n}{50 - n^2} \sin nt$

25.2 System response to Fourier series input

Now let's return to the original problem in Eq. (355). Suppose that the input signal f is an odd periodic function of period 2π. Since f is odd, the Fourier series of f is a linear combination of the shape

$$f(t) = b_1 \sin t + b_2 \sin 2t + b_3 \sin 3t + \cdots. \tag{358}$$

By the superposition principle, the system response to $f(t)$ is

$$x(t) = b_1 \frac{1}{49} \sin t + b_2 \frac{1}{46} \sin 2t + b_3 \frac{1}{41} \sin 3t + \cdots.$$

Note that each Fourier component $\sin nt$ has a different gain: the gain depends on the frequency. One could write the answer using sum

$$x(t) = \sum_{n\geq 1} \frac{b_n}{50 - n^2} \sin nt. \tag{359}$$

This is better since it shows precisely what every term in the series is. For which input signal $\sin nt$ is the gain the largest? The gain is $\left|1/(50-n^2)\right|$, which is largest when $|50-n^2|$ is smallest. This happens for $n=7$.

The gain for $\sin 7t$ is 1, and the next largest gain, occurring for $\sin 6t$ and $\sin 8t$, is $1/14$. Thus the system approximately filters out all the Fourier components of $f(t)$ except for the $\sin 7t$ term.

As related example, let's consider

$$\ddot{x}+50x=\frac{\pi}{4}\mathrm{Sq}(t). \tag{360}$$

As we have seen in the last section, the input signal can be written as

$$\frac{\pi}{4}\mathrm{Sq}(t)=\sum_{n\geq 1,\text{ odd}}\frac{\sin nt}{n} \tag{361}$$

which is a special case of Eq. (358). The system response to this signal is

$$\begin{aligned} x(t) &= \sum_{n\geq 1,\text{ odd}}\left(\frac{1}{50-n^2}\right)\frac{\sin nt}{n} \\ &\approx 0.020\sin t+0.008\sin 3t+0.008\sin 5t+ \\ &\qquad 0.143\sin 7t-0.003\sin 9t-\text{even smaller terms,} \end{aligned} \tag{362}$$

so the coefficient of $\sin 7t$ is largest, and the coefficient of $\sin t$ is second largest. This makes sense since the Fourier coefficient $\dfrac{1}{(50-n^2)n}$ is large only when one of n or $50-n^2$ is small.

This example illustrates a practically important fact: *Even though the system response is a complicated Fourier series, with infinitely many terms, only one or two are significant, and the rest are negligible.*

25.3 Pure resonance

Let's now consider what happens if we change 50 to 49 in the above ODE (360). To solve

$$\ddot{x}+49x=\frac{\pi}{4}\mathrm{Sq}(t) \tag{363}$$

we have to distinguish the cases $n\neq 7$ and $n=7$. For $n\neq 7$, we can solve $\ddot{x}+49x=\sin nt$ using complex replacement and ERF since in is not a root of r^2+49. For $n=7$, we can still solve $\ddot{x}+49x=\sin 7t$ (the existence and uniqueness theorem guarantees this), but the solution requires the generalized ERF, and involves t, and hence is not periodic: it turns out that one solution is $-(t/14)\cos 7t$. For the input signal $\mathrm{Sq}(t)$, we can find a solution x_p by superposition: most of the terms will be periodic, but one of them will be

$$\frac{1}{7}\left(-\frac{t}{14}\cos 7t\right)$$

and this makes the whole solution x_p non-periodic. Moreover, there are infinitely many other solutions, namely $x_p+c_1\cos 7t+c_2\sin 7t$ for any c_1 and c_2, but these solutions still include the term $\frac{1}{7}\left(-\frac{t}{14}\cos 7t\right)$, and hence are not periodic. If the ODE had been

$$\ddot{x}+36x=\frac{\pi}{4}\mathrm{Sq}(t)$$

then all solutions would have been periodic, because $\frac{\pi}{4}\operatorname{Sq}(t)$ has no $\sin 6t$ term in its Fourier series.

In general, for a periodic function f, the ODE $p(D)x = f(t)$ has a periodic solution if and only if for each term $\cos \omega t$ or $\sin \omega t$ appearing with a nonzero coefficient in the Fourier series of f, the number $i\omega$ is not a root of $p(r)$.

25.4 Resonance with damping

In real life, there is always damping, and this prevents the runaway growth in the pure resonance scenario of the previous section. So let's try to find the steady-state solution to

$$\ddot{x} + 0.1\dot{x} + 49x = \frac{\pi}{4}\operatorname{Sq}(t),$$

where $0.1\dot{x}$ is the linear damping term. Recall that the steady-state solution is the periodic solution. (Other solutions will be a sum of the steady-state solution with a transient solution solving the homogeneous ODE

$$\ddot{x} + 0.1\dot{x} + 49x = 0;$$

these transient solutions tend to 0 as $t \to \infty$, because the coefficients of the characteristic polynomial are positive (in fact, this is an underdamped system). First let's solve

$$\ddot{x} + 0.1\dot{x} + 49x = \sin nt.$$

Before doing that, solve the complex replacement ODE

$$\ddot{z} + 0.1\dot{z} + 49z = e^{int}.$$

The characteristic polynomial is $p(r) = r^2 + 0.1r + 49$, so ERF gives

$$z = \frac{1}{p(in)}e^{int} = \frac{e^{int}}{(49 - n^2) + (0.1n)i},$$

with complex gain $\dfrac{1}{(49 - n^2) + (0.1n)i}$ and gain

$$g_n := \frac{1}{|(49 - n^2) + (0.1n)i|}.$$

Thus

$$x = \Im\left(\frac{e^{int}}{(49 - n^2) + (0.1n)i}\right),$$

which is a sinusoid of amplitude g_n, so $x = g_n \cos(nt - \phi_n)$ for some ϕ_n. The input signal

$$\frac{\pi}{4}\operatorname{Sq}(t) = \sum_{n\geq 1,\text{ odd}} \frac{\sin nt}{n}$$

elicits the system response

$$
\begin{aligned}
x(t) = \sum_{n\geq 1,\ \text{odd}} g_n \frac{\cos(nt-\phi_n)}{n} \\
&\approx 0.020\cos(t-\phi_1) + 0.008\cos(3t-\phi_3) + 0.008\cos(5t-\phi_5) \\
&\qquad + 0.204\cos(7t-\phi_7) + 0.003\cos(9t-\phi_9) + \text{even smaller terms.}
\end{aligned}
$$

We conclude that the system response is almost indistinguishable from a pure sinusoid of angular frequency 7.

25.5 Complex Fourier Series

Euler's formula

$$e^{it} = \cos t + i \sin t$$

tells us that complex exponentials can be written as a sum of a sine and a cosine function. This suggests that we might be able to write a Fourier series

$$f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty}(a_n \cos nt + b_n \sin nt) \tag{364}$$

as a series of complex exponentials

$$f(t) = \sum_{n=-\infty}^{\infty} c_n e^{int} \tag{365}$$

for some *complex* coefficients c_n. As it turns out, this is true, that is, we can always write a Fourier series in terms of complex exponentials. Since the two series turn out to be equal, we'll also call the series in terms of complex exponentials a Fourier series. So let's walk through the process of converting a series from the real form (364) to the complex form (365).

Step 1: Rewriting the sum. Using Euler's formula and the fact that $\sin t$ is an odd function and $\cos t$ is an even function, we notice that

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

Then we see that given any Fourier series f, we can write

$$
\begin{aligned}
f(t) &= \frac{a_0}{2} + \sum_{n=1}^{\infty}(a_n \cos nt + b_n \sin nt) \\
&= \frac{a_0}{2} + \sum_{n=1}^{\infty}\left[\frac{a_n}{2}\left(e^{int} + e^{-int}\right) + \frac{ib_n}{2}\left(e^{-int} - e^{int}\right)\right] \\
&= \frac{a_0}{2} + \frac{1}{2}\sum_{n=1}^{\infty}\left[(a_n - ib_n)\, e^{int} + (a_n + ib_n)\, e^{-int}\right].
\end{aligned}
$$

Step 2: Defining coefficients. We can see that f can be written as a sum of complex exponentials. Let's write the coefficients of these exponentials nicely so that we can easily convert back and forth between the two forms. Define $c_0 := a_0/2$. For $n > 0$, define

$$c_n = \frac{1}{2}\left(a_n - ib_n\right) , \qquad c_{-n} := \bar{c}_n = \frac{1}{2}\left(a_n + ib_n\right).$$

Then we can write f compactly as

$$f(t) = \sum_{n=-\infty}^{\infty} c_n e^{int}. \tag{366}$$

This complex Fourier series is useful for two reasons: First, integrals involving complex exponentials are often much easier to compute than integrals involving sines and cosines. Second, we can also make sense of Fourier series of complex-valued functions more easily in this setting.

25.6 Listening to Fourier series

Your ear is capable of decomposing a sound wave into its Fourier components of different frequencies. Each frequency corresponds to a certain pitch. Increasing the frequency produces a higher pitch. More precisely, multiplying the frequency by a number greater than 1 increases the pitch by what in music theory is called an *interval*. For example, multiplying the frequency by 2 raises the pitch by an octave, and multiplying by 3 raises the pitch an octave plus a perfect fifth. Can our ear detect phase shifts? No – try it on the mathlet if you don't believe it.

When an instrument plays a note, it is producing a periodic sound wave in which typically many of the Fourier coefficients are nonzero. In a general Fourier series, the combination of the first two non-constant terms ($a_1 \cos t + b_1 \sin t$, if the period is 2π) is a sinusoid of some frequency ν, and the next combination (e.g., $a_2 \cos 2t + b_2 \sin 2t$) has frequency 2ν, and so on: the frequencies are the positive integer multiples of the lowest frequency ν. The note corresponding to the frequency ν is called the *fundamental*, and the notes corresponding to frequencies 2ν, 3ν, $\ldots$ are called the *overtones*.

26 Boundary value problems

26.1 Failure of existence and uniqueness

Our goal is to use boundary conditions for PDEs, but ODEs can have boundary conditions as well. In general, ODEs are much easier to solve than PDEs, so we now move our discussion to ODEs with boundary conditions to get a sense of how to solve these *boundary value problems* in a more comfortable setting. As a warm-up let's try to find all nonzero functions $v(x)$ on $[0, \pi]$ satisfying

$$v''(x) = \lambda\, v(x) \tag{367a}$$

for a constant λ and satisfying the *boundary conditions*

$$v(0) = 0 , \qquad v(\pi) = 0 \tag{367b}$$

If $\lambda = -n^2$, then any multiple of $\sin(nt)$ is a solution. However, for other λ, there is no solution! This reflects the fact that the existence and uniqueness theorem is valid *only* for initial value problems. As we will see in what follows, linear boundary value problems can have either *no* solutions at all, or infinitely many. The situation for nonlinear boundary value problems is even more complicated.

26.2 Solving a boundary value problem

Let's discuss how to to solve the boundary value problem (367) explicitly. The equation $v''(x) = \lambda\, v(x)$ is a homogeneous linear ODE with characteristic polynomial $r^2 - \lambda$.

Case 1: $\lambda > 0$. Then the general solution is $ae^{\sqrt{\lambda}x} + be^{-\sqrt{\lambda}x}$, and the boundary conditions say

$$a + b = 0\ , \qquad ae^{\sqrt{\lambda}\pi} + be^{-\sqrt{\lambda}\pi} = 0, \tag{368}$$

which is a linear system for (a, b). Since

$$\det \begin{pmatrix} 1 & 1 \\ e^{\sqrt{\lambda}\pi} & e^{-\sqrt{\lambda}\pi} \end{pmatrix} \neq 0, \tag{369}$$

the only solution to this linear system is $(a, b) = (0, 0)$. Thus there are no nonzero solutions v.

Case 2: $\lambda = 0$. Then the general solution is $a + bx$, and the boundary conditions say

$$a = 0\ , \qquad a + b\pi = 0. \tag{370}$$

Again the only solution to this linear system is $(a, b) = (0, 0)$. Thus there are no nonzero solutions v.

Case 3: $\lambda < 0$. We can write $\lambda = -\omega^2$ for some $\omega > 0$. Then the roots of the characteristic polynomial are $\pm i\omega$, and the general solution is $a \cos \omega x + b \sin \omega x$. The first boundary condition says $a = 0$, so $v = b \sin \omega x$. The second boundary condition then says $b \sin \omega\pi = 0$. We are looking for nonzero solutions v, so we can assume that $b \neq 0$. Then $\sin \omega\pi = 0$, so ω is an integer n; also $n > 0$, since $\omega > 0$.

We may thus conclude that there exist nonzero solutions if and only if $\lambda = -n^2$ for some positive integer n; in that case, all solutions are of the form $b \sin nx$. Below, we will use this conclusion as one key step in the solution of the heat qquation.

26.3 Analogy with eigenvalue-eigenvector problems

To describe a function $v(x)$, one needs to give infinitely many numbers, namely its values at all the different input x-values. Thus $v(x)$ is like a vector of infinite length. The linear differential operator $\dfrac{d^2}{dx^2}$ maps each function to a function, just as a 2×2 matrix defines a linear transformation mapping each vector in $\mathbb{R}^2$ to another vector in $\mathbb{R}^2$. Thus $\dfrac{d^2}{dx^2}$ is like an $\infty \times \infty$ matrix.

The ODE $v'' = \lambda v$ (with boundary conditions) amounts to an infinite system of equations: the ODE consists of one equality of numbers at each $x \in (0, \pi)$, and boundary

Eigenvector problem	**Eigenfunction problem**
vector $\boldsymbol{v}$	function $v(x)$
A	the linear operator $\dfrac{d^2}{dx^2}$
eigenvalue-eigenvector problem $A\boldsymbol{v} = \lambda\boldsymbol{v}$	boundary value problem $\dfrac{d^2}{dx^2}v = \lambda v$, $v(0) = 0$, $v(\pi) = 0$
eigenvalues λ	eigenvalues $\lambda = -1, -4, -9, \dots$
eigenvectors $\boldsymbol{v}$	eigenfunctions $v(x) = \sin nx$ with $n = \sqrt{-\lambda}$

Table 2: Summary of the analogies between eigenvector problems and eigenfunction problems.

conditions are equalities at the endpoints. Thus the ODE with boundary conditions is like a system of equations $A\boldsymbol{v} = \lambda\boldsymbol{v}$. Nonzero solutions $v(x)$ to

$$\frac{d^2}{dx^2}v = \lambda v$$

exist only for special values of λ, namely

$$\lambda = -1,\ -4,\ -9,\ \dots,$$

just as $A\boldsymbol{v} = \lambda\boldsymbol{v}$ has a nonzero solution $\boldsymbol{v}$ only for special values of λ, namely the eigenvalues of λ. But the differential operator $\triangle := d^2/dx^2$ has infinitely many eigenvalues, as one would expect for an $\infty \times \infty$ matrix.

The nonzero solutions $v(x)$ to $\triangle v = \lambda v$ satisfying the boundary conditions are called *eigenfunctions*, since they act like eigenvectors.

26.4 Introduction to the heat equation

In this section we meet our first partial differential equation (PDE)

$$\frac{\partial u}{\partial t} = \nu \frac{\partial^2 u}{\partial x^2}$$

This is the equation satisfied by the temperature $u(x, t)$ at position x and time t of a bar depicted as a line segment,

$$0 \le x \le L, \quad t \ge 0$$

The constant ν is the heat diffusion coefficient, which depends on the material of the bar. Let's focus on a specific physical experiment. Suppose that the initial temperature is 1, and then the ends of the bar are put in ice. We write this as

$$u(x,0) = 1, \quad 0 \le x \le L \qquad u(0,t) = 0, \quad u(L,t) = 0, \quad t > 0.$$

The value(s) of $u = 1$ at $t = 0$ are called *initial conditions.* The values at the ends are called *endpoint or boundary conditions.* We think of the initial and endpoint values of u as the input, and the temperature $u(x,t)$ for $t > 0$, $0 < x < L$ as the response. For simplicity, we assume that only the ends are exposed to the fixed external temperature. The rest of the bar is insulated, not subject to any external change in temperature. Fourier's techniques also yield answers even when there is heat input over time at other points along the bar.

As time passes, the temperature decreases as cooling from the ends spreads toward the middle. At the midpoint, $L/2$, one finds Newton's law of cooling,

$$u(L/2,t) \approx ce^{-t/\tau}, \quad t > \tau$$

The so-called characteristic time τ is inversely proportional to the conductivity of the material. If we choose units so that $\tau = 1$ for copper, then according to Wikipedia,

$$\tau \sim 7 \quad \text{(cast iron);} \quad \tau \sim 7000 \quad \text{(dry snow)}$$

The constant c, on the other hand, is *universal*:

$$c \approx 1.3$$

It depends only on the fact that the shape is a bar (modeled as a line segment). Fourier figured out not only how to explain c using differential equations, but the whole temperature profile:

$$u(x,t) \approx e^{-t/\tau} h(x)\ , \qquad h(x) = \frac{4}{\pi} \sin\left(\frac{\pi}{L}x\right), \quad t > \tau.$$

The shape of h reflects how much faster the temperature drops near the ends than in the middle. It's natural that h should be some kind of hump, symmetric around $L/2$.

26.5 Deriving the heat equation

To explain the heat equation, we start with a thought experiment. If we fix the temperature at the ends, $u(0,t) = 0$ and $u(L,t) = T$, what will happen in the long term as $t \to \infty$? The answer is that

$$u(x,t) \to U_{\text{steady}}(x), \quad t \to \infty$$

where U_{steady} is the steady, or equilibrium, temperature given by a linear profile

$$U_{\text{steady}}(x) = T\frac{x}{L}.$$

The temperature $u(L/2,t)$ at the midpoint $L/2$ tends to the average of 0 and T, namely $T/2$. At the point $L/4$, half way between 0 and $L/2$, the temperature tends to the average of the temperature at 0 and $T/2$, and so forth.

At a very small scale, this same mechanism, the tendency of the temperature profile toward a straight line equilibrium means that if u is concave down then the temperature in the middle should decrease (so the profile becomes closer to being straight). If u is concave up, then the temperature in the middle should increase (so that, once again, the profile becomes closer to being straight). We write this as

$$\frac{\partial^2 u}{\partial x^2} < 0 \qquad \Rightarrow \qquad \frac{\partial u}{\partial t} < 0 \tag{371a}$$

$$\frac{\partial^2 u}{\partial x^2} > 0 \qquad \Rightarrow \qquad \frac{\partial u}{\partial t} > 0. \tag{371b}$$

The simplest relationship that reflects this is a linear (proportional) relationship,

$$\frac{\partial u}{\partial t} = \nu \frac{\partial^2 u}{\partial x^2} \ , \qquad \nu > 0 \tag{372}$$

Another 'more macroscopic' way of deriving the heat equation starts from the general conservation law

$$\frac{\partial u}{\partial t} = -\nabla \cdot \boldsymbol{J}, \tag{373}$$

which states that the time-change of u at position x is determined by the net differences of the heat fluxes into and out of x. In 1D, we have only an current along the x-axis, so that the last equation simplifies to

$$\frac{\partial u}{\partial t} = -\frac{\partial}{\partial x} J_x. \tag{374}$$

It is reasonable to assume that the heat flux J_x is negatively proportional to the temperature gradient $\boldsymbol{J} = -\nu \nabla u$, which in 1D becomes

$$J_x(x,t) = -\nu \frac{\partial u}{\partial x}, \tag{375}$$

where ν is a material parameter. Substituting this into Eq. (374), we recover the 1D heat diffusion equation. Equations (373) and Eq. (374) are conservation laws. To see this, let's integrate (374) from $x = 0$ to $x = L$,

$$\begin{aligned} \frac{d}{dt} \int_0^L dx\, u &= \int_0^L dx\, \frac{\partial}{\partial t} u \\ &= -\int_0^L dx\, \frac{\partial}{\partial x} J_x = -[J(L,t) - J(0,t)]. \end{aligned}$$

This shows that the time change of the integral of the quantity u is only determined by the flux at the endpoints. In particular, if the endpoints are insulated, so that

$$J(0,t) = -\nu \frac{\partial u}{\partial x}(0,t) = 0 \ , \qquad J(L,t) = -\nu \frac{\partial u}{\partial x}(L,t) = 0. \tag{376}$$

Then the integral of u is conserved (constant in time).

26.6 Separation of variables and normal modes

Let's now try to solve the PDE. For simplicity, suppose that $L = \pi$ and $\nu = 1$. The general case is similar. In fact, one could reduce to this special case by changes of variable. So now we are solving

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \tag{377a}$$

$$u(0,t) = 0 \,, \qquad u(\pi,t) = 0 \quad \text{for} \quad t \geq 0 \tag{377b}$$

$$u(x,0) = u_0(x) \quad \text{for} \quad x \in (0,\pi), \tag{377c}$$

where $u_0(x)$ is a given initial temperature profile. We now use an important technique called *separation of variables.* Forgetting about the initial condition $u(x,0) = u_0(x)$ for now, we look for nonzero solutions of the form

$$u(x,t) = v(x)\,w(t)$$

Substituting into the PDE gives

$$\dot{w}(t)v(x) \;=\; v''(x)w(t) \tag{378a}$$

$$\frac{\dot{w}(t)}{w(t)} \;=\; \frac{v''(x)}{v(x)} \tag{378b}$$

at least where $w(t)$ and $v(x)$ are nonzero. Whenever a function of x is equal to a function of t there is a constant λ such that

$$\frac{v''(x)}{v(x)} = \lambda \quad \text{and} \quad \frac{\dot{w}(t)}{w(t)} = \lambda,$$

or in other words,

$$v''(x) = \lambda\, v(x) \quad \text{and} \quad \dot{w}(t) = \lambda\, w(t).$$

Substituting $u(x,t) = w(t)v(x)$ into the first boundary condition $u(0,t) = 0$ gives $w(t)v(0) = 0$ for all t, but $w(t)$ is not the zero function, so this translates into $v(0) = 0$. Similarly, the second boundary condition $u(\pi,t) = 0$ translates into $v(\pi) = 0$. At the beginning of this section, we already solved $v''(x) = \lambda\, v(x)$ subject to the boundary conditions $v(0) = 0$ and $v(\pi) = 0$: nonzero solutions $v(x)$ exist only if $\lambda = -n^2$ for some positive integer n, and in that case $v(x)$ is a scalar times $\sin nx$.

For $\lambda = -n^2$, what is a matching possibility for w? Since $\dot{w} = -n^2 w$, the function w is a scalar times $e^{-n^2 t}$. This gives rise to one solution

$$u_n(x,t) = b_n e^{-n^2 t} \sin nx \tag{379}$$

for each positive integer n to the PDE with boundary conditions. Each such solution u_n is called a *normal mode.*

The PDE and boundary conditions are homogeneous, so we can get other solutions by taking linear combinations:

$$u(x,t) = b_1 e^{-t} \sin x + b_2 e^{-4t} \sin 2x + b_3 e^{-9t} \sin 3x + \cdots \,. \tag{380}$$

This turns out to be the general solution to the PDE with the boundary conditions. We can also write this as

$$u(x,t) = \sum_{n=1}^{\infty} b_n e^{-n^2 t} \sin nx. \tag{381}$$

The one thing we haven't used so far is the initial condition $u_0(x) = u(x,0)$. This condition determines the coefficients b_n, as we we can see by setting $t = 0$ in the solution formula, which gives

$$u(x,0) = \sum_{n=1}^{\infty} b_n \sin nx = u_0(x). \tag{382}$$

That is, the coefficients b_n are simply the coefficients of the Fourier-sine series of $u_0(x)$, and we know how determine those from the scalar product of $u_0(x)$ and the basis functions $\sin nx$.

27 Heat equation with insulated ends: von Neumann boundary conditions

In the last class, we considered an insulated metal rod of length $L = \pi$ and thermal conductivity $\nu = 1$ with exposed ends held at 0°C. The temperature profile $u(x,t)$ along the rod was governed by the heat equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \tag{383a}$$

which is a linear PDE. The heat equation is first order in time t and second order in the space coordinate x, so we have to specify one initial condition

$$u(x,0) = u_0(x) \tag{383b}$$

and two boundary conditions, which chose as

$$u(0,t) = u(\pi,t) = 0\ , \qquad \forall\, t > 0. \tag{383c}$$

Keep in mind that all three Eqs. (383) are required to uniquely specify the temperature field $u(x,t)$. When the value of the of the sought-after function (in this case u) described by a PDE are prescribed, then one speaks of *Dirichlet boundary conditions.*

Trying $u = w(t)v(x)$ we found separate ODEs for v and w, leading to solutions $e^{-n^2 t} \sin nx$ for $n = 1, 2, \ldots$ to the PDE with boundary conditions. The condition of discrete n was imposed by the boundary conditions. Since the PDE is linear, we then took linear combinations to get the general solution

$$u(x,t) = b_1 e^{-t} \sin x + b_2 e^{-4t} \sin 2x + b_3 e^{-9t} \sin 3x + \cdots \tag{384}$$

To fully specify the solutions, we still need to determine the coefficients b_n by making use of the initial condition. For simplicity, we specialize to the case where the initial temperature profile is homogeneous $u(x,0) = u_0(x) = 1$. Setting $t = 0$, the general solution becomes

$$u(x,0) = b_1 e^{-t} \sin x + b_2 e^{-4t} \sin 2x + b_3 e^{-9t} \sin 3x + \cdots \tag{385}$$

System of ODEs	**Heat Equation**
vector $\boldsymbol{v}$	function $v(x)$
A	the linear operator d^2/dx^2
eigenvalue-eigenvector problem $A\boldsymbol{v} = \lambda\boldsymbol{v}$	boundary value problem $d^2v/dx^2 = \lambda v$, $v(0) = 0$, $v(\pi) = 0$
eigenvalues λ	eigenvalues $\lambda = -1, -4, -9, \ldots$
eigenvectors $\boldsymbol{v}$	eigenfunctions $v(x) = \sin nx$
linear system of ODEs $\dot{\boldsymbol{x}} = A\boldsymbol{x}$	Heat Equation with boundary conditions $\partial u/\partial t = \partial^2 u/\partial x^2, \quad u(0,t) = 0, \quad u(\pi,t) = 0$
normal modes: $e^{\lambda t}\boldsymbol{v}$ for an eigenvector $\boldsymbol{v}$ with eigenvalue λ	normal modes: $e^{\lambda t}v(x) = e^{-n^2 t}\sin nx$ for eigenfunction $v(x) = \sin nx$, eigenvalue $\lambda = -n^2$
General solution: $\boldsymbol{u} = \sum c_n e^{\lambda_n t}\boldsymbol{v}_n$	General solution: $u = \sum b_n e^{-n^2 t}\sin nx$
Solve $\boldsymbol{u}(0) = \sum c_n \boldsymbol{v}_n$ to get the c_n	Solve $u(x,0) = \sum b_n \sin nx$ to get the b_n

Table 3: Analogy between a linear system of ODEs and the Heat Equation.

Inserting the initial condition on the left, we get

$$1 = b_1 \sin x + b_2 \sin 2x + b_3 \sin 3x + \cdots, \qquad \forall\, x \in (0, \pi), \tag{386}$$

which must be solved for $b_1, b_2, \ldots$. We already showed how to find such b_i: the left hand side extends to an odd period 2π function, namely $\mathrm{Sq}(x)$, so we need to solve

$$\mathrm{Sq}(x) = b_1 \sin x + b_2 \sin 2x + b_3 \sin 3x + \cdots, \qquad \forall\, x \in \mathbb{R}.$$

We already know the answer:

$$\mathrm{Sq}(x) = \frac{4}{\pi}\sin x + \frac{4}{3\pi}\sin 3x + \frac{4}{5\pi}\sin 5x + \cdots. \tag{387}$$

In other words $b_n = 0$ for even n, and $b_n = \dfrac{4}{n\pi}$ for odd n. Substituting these b_n back into the general solution to the heat equation gives

$$u(x,t) = \frac{4}{\pi}e^{-t}\sin x + \frac{4}{3\pi}e^{-9t}\sin 3x + \frac{4}{5\pi}e^{-25t}\sin 5x + \cdots.$$

What does the temperature profile look like when t is large? All the Fourier components are decaying, so $u(x,t) \to 0$ as $t \to +\infty$ at every position. Thus the temperature profile approaches a horizontal segment, the graph of the zero function. But the Fourier components of higher frequency decay much faster than the first Fourier component, so when t is large, the formula

$$u(x,t) \approx \frac{4}{\pi} e^{-t} \sin x$$

is a very good approximation. Eventually, the temperature profile is indistinguishable from a sinusoid of angular frequency 1 whose amplitude is decaying to 0. This can be observed in the *Heat Equation* mathlet.

27.1 Inhomogeneous boundary conditions

The steps to solve a linear PDE with *inhomogeneous* boundary conditions are

1. Find a particular solution u_p to the PDE with the inhomogeneous boundary conditions (but without initial conditions). If the boundary conditions do not depend on t, try to find the steady-state solution $u_p(x,t)$, i.e., the solution that does not depend on t.

2. Find the general solution u_h to the PDE with the homogeneous boundary conditions.

3. Then $u := u_p + u_h$ is the general solution to the PDE with the inhomogeneous boundary conditions.

4. If initial conditions are given, use them to find the specific solution to the PDE with the inhomogeneous boundary conditions. (This often involves finding Fourier coefficients.)

As an example, let's consider the same insulated uniform metal rod as before ($\nu = 1$, length π, initial temperature 1°C), but now suppose that the left end is held at 0°C while the right end is held at 20°C. What is $u(x,t)$ in this case? To find the answer, we proceed through the list of steps above:

1. Forgetting the initial condition for now, we look for a solution $u = u(x)$ that does not depend on t. Plugging this into the Heat Equation PDE gives $0 = \frac{\partial^2 u}{\partial x^2}$. The general solution to this simplified DE is $u(x) = ax + b$ Imposing the boundary conditions $u(0) = 0$ and $u(\pi) = 20$ leads to $b = 0$ and $a = 20/\pi$, so $u_p = \frac{20}{\pi} x$.

2. The PDE with the homogeneous boundary conditions is what we solved earlier; the general solution is

$$u_h = b_1 e^{-t} \sin x + b_2 e^{-4t} \sin 2x + b_3 e^{-9t} \sin 3x + \cdots .$$

3. The general solution to the PDE with inhomogeneous boundary conditions is then

$$u(x,t) = u_p + u_h = \frac{20}{\pi} x + b_1 e^{-t} \sin x + b_2 e^{-4t} \sin 2x + b_3 e^{-9t} \sin 3x + \cdots . \qquad (388)$$

4. To find the b_n, set $t = 0$ and use the initial condition on the left:

$$1 \quad = \quad \frac{20}{\pi}x + b_1 \sin x + b_2 \sin 2x + b_3 \sin 3x + \cdots \ , \qquad \forall\, x \in (0, \pi). \tag{389}$$

Bringing time-independent terms to the right, we have

$$1 - \frac{20}{\pi}x \quad = b_1 \sin x + b_2 \sin 2x + b_3 \sin 3x + \cdots \ , \qquad \forall\, x \in (0, \pi). \tag{390}$$

Extend $1 - \frac{20}{\pi}x$ on $(0, \pi)$ to an odd periodic function $f(x)$ of period 2π. Then use the Fourier coefficient formulas to find the b_n such that

$$f(x) = b_1 \sin x + b_2 \sin 2x + b_3 \sin 3x + \ldots$$

Alternatively, find the Fourier series for the odd periodic extensions of 1 and x separately, and take a linear combination to get $1 - \frac{20}{\pi}x$. Once the b_n are found, plug them back into the general solution for the heat equation with inhomogeneous boundary conditions.

27.2 Insulated ends

Consider the same insulated uniform metal rod as before ($\nu = 1$, length π) but now assume that the ends are insulated too (instead of exposed and held in ice), and that the initial temperature is given by $u(x, 0) = x$ for $x \in (0, \pi)$. Now what is $u(x, t)$? As before, we temporarily forget the initial condition and account for them in the finals step.

'Insulated ends' means that there is zero heat flow through the ends. We showed above that the heat flux density function

$$J_x \propto -\frac{\partial u}{\partial x},$$

so this quantity must vanish at the endpoints $x = 0$ or $x = \pi$

$$\frac{\partial u}{\partial x}(0, t) = 0, \qquad \frac{\partial u}{\partial x}(\pi, t) = 0 \ , \qquad \forall\, t > 0, \tag{391}$$

These conditions on the first order spatial derivatives are called von *Neumann boundary condition*. They describe a different physical condition than the Dirichlet boundary conditions $u(0, t) = 0$ and $u(\pi, t) = 0$ used above. So we need to solve the heat equation

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

with the boundary conditions for insulated ends. Separation of variables $u(x, t) = v(x)\, w(t)$ leads to

$$v''(x) = \lambda\, v(x) \ , \qquad v'(0) = 0 \ , \qquad v'(\pi) = 0 \ , \qquad \dot{w}(t) = \lambda\, w(t). \tag{392}$$

for a constant λ. Looking at the cases $\lambda > 0$, $\lambda = 0$, $\lambda < 0$, we find that

$$\lambda = -n^2 \quad \text{and} \quad v(x) = \cos nx$$

where n is one of $0, 1, 2, \ldots$. This time n starts at 0 since $\cos 0x$ is a nonzero function. For each such $v(x)$, the corresponding w is $w(t) = e^{-n^2 t}$ (times a scalar), and the normal mode is

$$u = e^{-n^2 t} \cos nx.$$

The case $n = 0$ is the constant function 1, so the general solution is

$$u(x,t) = \frac{a_0}{2} + a_1 e^{-t} \cos x + a_2 e^{-4t} \cos 2x + a_3 e^{-9t} \cos 3x + \cdots .$$

Finally, we bring back the initial condition: substitute $t = 0$ and use the initial condition on the left to get

$$x = \frac{a_0}{2} + a_1 \cos x + a_2 \cos 2x + a_3 \cos 3x + \cdots$$

for all $x \in (0, \pi)$. The right hand side is a period 2π even function, so extend the left hand side to a period 2π even function $T(x)$, a triangle wave, which is an antiderivative of

$$\mathrm{Sq}(x) = \frac{4}{\pi} \left(\sin x + \frac{\sin 3x}{3} + \frac{\sin 5x}{5} + \cdots \right).$$

Integration gives

$$T(x) = \frac{a_0}{2} - \frac{4}{\pi} \left(\cos x + \frac{\cos 3x}{9} + \frac{\cos 5x}{25} + \cdots \right),$$

and the constant term $a_0/2$ is the average value of $T(x)$, which is $\pi/2$. Thus

$$T(x) = \frac{\pi}{2} - \frac{4}{\pi} \left(\cos x + \frac{\cos 3x}{9} + \frac{\cos 5x}{25} + \cdots \right) \tag{393}$$

yielding the full time-dependent solution

$$u(x,t) = \frac{\pi}{2} - \frac{4}{\pi} \left(e^{-t} \cos x + e^{-9t} \frac{\cos 3x}{9} + e^{-25t} \frac{\cos 5x}{25} + \cdots \right). \tag{394}$$

This answer makes physical sense: when the entire bar is insulated, its temperature tends to a constant equal to the average of the initial temperature.

27.3 Other boundary conditions

Besides the two simple boundary conditions we described above, there are a few others that can be useful. One other boundary condition, not used as often but still important , is what is know as the *Robin* boundary condition. This condition has the form on the boundary

$$u + a \frac{\partial u}{\partial x} = b \tag{395}$$

where a and b are constants. Such a condition is usually used to represent some sort of convective transport occurring at the boundaries. Imagine a glass of beer or soda with the top open to the atmosphere, and a wind is blowing over it. CO_2 naturally diffuses into the air above the beverage, and the wind will tend to carry it away. The above boundary condition deals with this case.

28 Wave equation

The wave equation in 1+1 dimensions is the following PDE:

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2} \tag{396}$$

You can think of $u(x,t)$ as describing the vertical displacement of a guitar string or elastic fibre from its uncurved equilibrium configuration, which we can assume to be given by a straight horizontal line segment $[0,L]$. Comparing units of both sides of the wave equation shows that the units for c are m/s, and we interpret c as the wave speed. While the rhs. of the wave equation looks similar to a diffusion equation, the lhs. features a second order derivative; that is, the wave equation is a linear second-order equation in both space and time. Hence, to uniquely determine a solution, we will have to specify 4 conditions: 2 initial conditions and 2 boundary conditions.

The 3+1-dimensional generalization of Eq. (396)

$$\frac{\partial^2 u}{\partial t^2} = c^2 \left[\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}\right], \tag{397}$$

where $u(x,y,z)$ can also be a scalar (or a component of a vector or a matrix), describes the propagation of sound waves in a solid (with c the sound speed), the propagation of electromagnetic waves (with c the speed of light), and also weak gravitational waves (with c the speed of light).

28.1 Separation of variables in PDEs and normal modes

Let's find solutions for Eq. (396) assuming that the endpoints of the string are fixed. For simplicity, suppose that $c = 1$ and $L = \pi$. So now we are solving the PDE with boundary conditions

$$\begin{aligned} \frac{\partial^2 u}{\partial t^2} &= \frac{\partial^2 u}{\partial x^2} && \text{(398a)} \\ u(0,t) &= 0 && \text{(398b)} \\ u(\pi,t) &= 0. && \text{(398c)} \end{aligned}$$

As with the heat equation, we try separation of variables. In other words, we try to find normal modes of the form

$$u(x,t) = v(x)w(t), \tag{399}$$

for some nonzero functions $v(x)$ and $w(t)$. Substituting this into the PDE gives

$$v(x)\ddot{w}(t) = v''(x)w(t)\ , \qquad \frac{\ddot{w}(t)}{w(t)} = \frac{v''(x)}{v(x)}. \tag{400}$$

As usual, a function of t can equal a function of x only if both are equal to the same constant, say λ, so this breaks into two ODEs:

$$\ddot{w}(t) = \lambda\, w(t), \qquad v''(x) = \lambda\, v(x).$$

Moreover, the boundary conditions become $v(0) = 0$ and $v(\pi) = 0$.

We already solved the eigenfunction equation $v''(x) = \lambda\, v(x)$ with the boundary conditions $v(0) = 0$ and $v(\pi) = 0$: nonzero solutions exist only when $\lambda = -n^2$ for some positive integer n, and in this case

$$v(x) = \sin nx. \tag{401}$$

What is different this time is that w satisfies a *second*-order ODE

$$\ddot{w}(t) = -n^2\, w(t). \tag{402}$$

The characteristic polynomial is $r^2 + n^2$, which has roots $\pm in$, so

$$w(t) := \cos nt \quad \text{and} \quad w(t) := \sin nt \tag{403}$$

are possibilities (and all the others are linear combinations). Multiplying each by the $v(x)$ with the matching λ gives the normal modes

$$\cos nt \sin nx, \qquad \sin nt \sin nx. \tag{404}$$

Any linear combination

$$u(x,t) = \sum_{n\geq 1} a_n \cos nt \sin nx + \sum_{n\geq 1} b_n \sin nt \sin nx \tag{405}$$

is a solution to the PDE with boundary conditions, and this turns out to be the general solution.

Initial conditions. To obtain a unique solution, we have to specify two initial conditions: not only the initial position $u(x,0)$, but also the initial velocity $\dfrac{\partial u}{\partial t}(x,0)$, at each position of the string.

For a plucked string, it is reasonable to assume that the initial velocity is 0, so one initial condition is

$$\frac{\partial u}{\partial t}(x,0) = 0. \tag{406}$$

What condition does this impose on the a_n and b_n? Well, for the general solution above,

$$\frac{\partial u}{\partial t}(x,t) \quad = \quad \sum_{n\geq 1} -na_n \sin nt \sin nx + \sum_{n\geq 1} nb_n \cos nt \sin nx \tag{407}$$

so that at time $t = 0$

$$\frac{\partial u}{\partial t}(x,0) \quad = \quad \sum_{n\geq 1} nb_n \sin nx, \tag{408}$$

so the initial condition (406) says that $b_n = 0$ for every n; and the solution simplifies to

$$u(x,t) = \sum_{n\geq 1} a_n \cos nt \sin nx. \tag{409}$$

If we also knew the initial position $u(x,0)$, we could solve for the a_n by extending to an odd, period 2π function of x and using the Fourier coefficient formula.

28.2 D'Alembert's solution: traveling waves

D'Alembert figured out another way to write down solutions, in the case when $u(x,t)$ is defined for all real numbers x instead of just $x \in [0,L]$. Then, for any reasonable function f,

$$u(x,t) := f(x-ct) \tag{410}$$

is a solution to the PDE, as shown by the following calculations:

$$\begin{aligned} \frac{\partial u}{\partial t} &= (-c)f'(x-ct) & \qquad \frac{\partial u}{\partial x} &= f'(x-ct) \\ \frac{\partial^2 u}{\partial t^2} &= (-c)^2 f''(x-ct) & \qquad \frac{\partial^2 u}{\partial x^2} &= f''(x-ct), \end{aligned}$$

so

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}. \tag{411}$$

What is the physical meaning of this solution? At $t=0$, we have

$$u(x,0) = f(x), \tag{412}$$

so $f(x)$ is the initial position. For any number t, the position of the wave at time t is the graph of $f(x-ct)$, which is the graph of f shifted ct units to the right. Thus the wave travels at constant speed c to the right, maintaining its shape.

It is easy to check, by going through the same calculation, that the function

$$u(x,t) := g(x+ct), \tag{413}$$

for any reasonable function $g(x)$, is a solution too. This describes wave moving to the left. It turns out that the general solution is a superposition

$$u(x,t) = f(x+ct) + g(x-ct).$$

There is a tiny bit of redundancy: one can add a constant to f and subtract the same constant from g without changing u.

28.3 Wave fronts

Define the step function

$$s(x) := \begin{cases} 1, & \text{if } x < 0 \\ 0, & \text{if } x > 0, \end{cases}$$

and consider the solution $u(x,t) = s(x-t)$. This is a 'cliff-shaped' wave traveling to the right. You would be right to complain that this function is not differentiable and therefore cannot satisfy the PDE in the usual sense, but you can imagine replacing $s(x)$ with a smooth approximation, a function with very steep slope. The smooth approximation also makes more sense physically: a physical wave would not actually have a jump discontinuity.

Another way to plot the behavior is to use a *space-time diagram*, in a plane with axes x (space) and t (time). (Usually one draws only the part with $t \geq 0$.) Divide the (x,t)-plane into regions according to the value of u. The boundary between the regions is called the *wave front*.

In the example above, $u(x,t) = 1$ for points to the left of the line $x - t = 0$, and $u(x,t) = 0$ for points to the right of the line $x - t = 0$. So the wave front is the line $x = t$.

28.4 Real-life waves

In real life, there is always damping. This introduces a new term into the wave equation, giving the damped wave equation (aka the *telegrapher's equation*)

$$\frac{\partial^2 u}{\partial t^2} + \frac{1}{\tau}\frac{\partial u}{\partial t} = c^2\,\frac{\partial^2 u}{\partial x^2}, \tag{414}$$

where τ is a damping time scale. In this case, separation of variables still works, but in each normal mode, the $w(t)$ is a damped sinusoid involving a factor $e^{-t/(2\tau)}$ in the underdamped case, where τ is sufficiently large (if τ becomes small, we recover the diffusion equation). This equation was derived by Lord Kelvin in the 1850s, when he did calculations to estimate the signal transduction for the first transatlantic cable.