ES.1803 Topic Notes, Spring 2024

Jeremy Orloff

1 Introduction to differential equations

1.1 Goals

1. Know the definition of a differential equation.
2. Know our first and second most important equations and their solutions.
3. Be able to derive the differential equation modeling a physical or geometric situation.
4. Be able to solve a separable differential equation, including finding lost solutions.
5. Be able to solve an initial value problem (IVP) by solving the differential equation and using the initial condition to find the constant of integration.

1.2 Differential equations and solutions

A differential equation (DE) is an equation with derivatives!

Example 1.1. (DEs modeling physical processes, i.e., rate equations)

1. Newton's law of cooling: $\dfrac{dT}{dt} = -k(T - A)$, where T is the temperature of a body in an environment with ambient temperature A.

2. Gravity near the earth's surface: $m\dfrac{d^2x}{dt^2} = -mg$, where x is the height of a mass m above the surface of the earth.

3. Hooke's law: $m\dfrac{d^2x}{dt^2} = -kx$, where x is the displacement from equilibrium of a spring with spring constant k.

Other examples: Below we will give some examples of differential equations modeling some geometric situations.

A solution to a differential equation is any function that satisfies the DE. Let's focus on what this means by contrasting it with solving an algebraic equation.

The unknown in an algebraic equation, such as

$$y^2 + 2y + 1 = 0$$

is the number y. The equation is solved by finding a numerical value for y that satisfies the equation. You can check by substitution that $y = -1$ is a solution to the equation shown.

The unknown in the differential equation

$$\frac{d^2y}{dx^2} + 2\frac{dy}{dx} + y = 0$$

is the **function** $y(x)$.The equation is solved by finding a function $y(x)$ that satisfies the equation One solution to the equation shown is $y(x) = e^{-x}$. You can check this by substituting $y(x) = e^{-x}$ into the equation. Again, note that the **solution is a function**.

More often we will say that *the* solution is a family of functions, e.g., $y = Ce^{-t}$. The parameter C is like the constant of integration in 18.01. Every value of C gives a different function which solves the DE.

1.3 The most important differential equation in 18.03

Here, in the very first class, we state and give solutions to our most important differential equations. In this case we will check the solutions by substitution. As we proceed in the course we will learn methods that help us discover solutions to equations.

The most important DE we will study is

$$\frac{dy}{dt} = ay, \tag{1}$$

where a is a constant (in units of 1/time). In words the equation says that

the rate of change of y is proportional to y.

Because of its importance we will write down some other ways you might see it:

$$y' = ay; \qquad \frac{dy}{dt} = ay(t); \qquad y' - ay = 0; \qquad \dot{y} - ay = 0.$$

In the last equation, we used the physicist 'dot' notation to indicate the derivative is with respect to time. You should recognize that all of these are the same equation.

The solution to this equation is

$$y(t) = Ce^{at},$$

where C is any constant.

1.3.1 Checking the solution by substitution

The above solution is easily checked by substitution. Because this equation is so important we show the details. Substituting $y(t) = Ce^{at}$ into Equation 1 we have:

$$\begin{aligned} \text{Left side of 1:} \quad y' &= aCe^{at} \\ \text{Right side of 1:} \quad ay &= aCe^{at} \end{aligned}$$

Since after substitution the left side equals the right, we have shown that $y(t) = Ce^{at}$ is indeed a solution of Equation 1.

1.3.2 The physical model of the most important DE

As a physical model this equation says that the quantity y changes at a rate proportional to y.

Because of the form the solution takes we say that Equation 1 models exponential growth or decay.'

In this course we will learn many techniques for solving differential equations. We will test almost all of them on Equation 1. After learning these techniques, you should, of course, understand how to use them to solve 1. However: whenever you see this equation you should remind yourself that it models exponential growth or decay and you should know the solution without computation.

1.4 The second most important differential equation

Our second most important DE is

$$my'' + ky = 0, \tag{2}$$

where m and k are constants. You can easily check that, with $\omega = \sqrt{k/m}$, the function

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

is a solution. Equation 2 models a simple harmonic oscillator. More prosaically, it models a mass m oscillating at the end of a spring with spring constant k.

1.5 Solving differential equations by the method of optimism

In our first and second most important equations above we simply told you the solution. Once you have a possible solution it is easy to check it by substitution into the differential equation. We will call this method, where you guess a solution and check it by plugging your guess into the equation, the method of optimism. In all seriousness, this will be an important method for us. Of course, its utility depends on learning how to make good guesses!

1.6 General form of a differential equation

We can always rearrange a differential equation so that the right hand side is 0. For example, $y' = ay$ can be written as $y' - ay = 0$. With this in mind the most general form for a differential equation is

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

where F is a function. For example,

$$(y')^2 + e^{y'' \sin(t)} - y^{(4)} = 0.$$

The order of a differential equation is the order of the highest derivative that occurs. So the example just above shows a DE of order 4.

1.7 Constructing a differential equation to model a physical situation

We use rate equations, i.e., differential equations, to model systems that undergo change. The following argument using Δt should be somewhat familiar from calculus.

Example 1.2. Suppose a population $P(t)$ has constant birth and death rates:

$$\beta = 2\%/\text{year}, \;\; \delta = 1\%/\text{year}$$

Build a differential equation that models this situation.

Solution: In the interval $[t, t+\Delta t]$, the change in P is given by

$$\Delta P = \text{ number of births - number of deaths.}$$

Over a small time interval Δt the population is roughly constant so:

$$\begin{aligned} \text{Births in the time interval} &\approx P(t)\cdot\beta\cdot\Delta t \\ \text{Deaths in the time interval} &\approx P(t)\cdot\delta\cdot\Delta t \end{aligned}$$

Combining these we have: $\Delta P \approx P(t)\,\beta\,\Delta t - P(t)\,\delta\,\Delta t$. So,

$$\frac{\Delta P}{\Delta t} \approx (\beta-\delta)P(t).$$

Finally, letting Δt go to 0 we have derived the differential equation

$$\frac{dP}{dt} = (\beta-\delta)P.$$

Notice that if $\beta > \delta$ then the population is increasing.

Of course, this DE is our most important DE 1: the equation of exponential growth or decay. We know the solution is $P = P_0 e^{(\beta-\delta)t}$.

Note: Suppose β and δ are more complicated and depend on t, say $\beta = P+2t$ and $\delta = P/t$. The derivation of the DE is the same, i.e.

$$\frac{dP}{dt} = (\beta(t)-\delta(t))P = (P+2t-P/t)P.$$

Because β and δ are no longer constants, this is not a situation of exponential growth and the solution will be more complicated (and probably harder to find).

Example 1.3. Bacteria growth. Suppose a population of bacteria is modeled by the exponential growth equation $P' = kP$. Suppose that the population doubles every 3 hours. Find the growth constant k.

Solution: The equation $P' = kP$ has solution $P(t) = Ce^{kt}$. From the initial condition we have that $P(0) = C$. Since the population doubles every 3 hours we have $P(3) = Ce^{3k} = 2C$.

Solving for k we get $\boxed{k = \frac{1}{3}\ln 2 \quad \text{(in units of 1/hours.)}}$

1.8 Initial value problems

An initial value problem (IVP) is just a differential equation where one value of the solution function is specified. We illustrate with some simple examples.

Example 1.4. Initial value problem. Solve the IVP $\dot{y} = 3y$, $y(0) = 7$.

Solution: We recognize this as an exponential growth equation, so $y(t) = Ce^{3t}$. Using the initial condition we have $y(0) = 7 = C$. Therefore, $\boxed{y(t) = 7e^{3t}.}$

Example 1.5. Initial value problem. Solve the IVP $y' = x^2$, $y(2) = 7$.

Solution: Note, the use of x indicates that the independent variable in this problem is x. This is really an 18.01 problem: integrating we get $y = x^3/3 + C$. Using the initial condition we find $C = 7 - 8/3$.

1.9 Separable Equations

Now it's time to learn our first technique for solving differential equations. A first-order DE is called separable if the variables can be separated from each other. We illustrate with a series of examples.

Example 1.6. Exponential growth. Use separation of variables to solve the exponential growth equation $y' = 4y$.

Solution: We rewrite the equation as $\dfrac{dy}{dt} = 4y$. Next we separate the variables by getting all the y's on one side and the t's on the other.

$$\frac{dy}{y} = 4\,dt.$$

Now we integrate both sides:

$$\int \frac{dy}{y} = \int 4\,dt \qquad \Leftrightarrow \qquad \ln|y| = 4t + C.$$

Now we solve for y by exponentiating both sides:

$$|y| = e^C e^{4t} \quad \textit{or} \quad y = \pm e^C e^{4t}.$$

Since $\pm e^C$ is just a constant we rename it simply K. We now have the solution we knew we'd get:

$$y = Ke^{4t}.$$

Example 1.7. Here is a standard example where the solution goes to infinity in a finite time (i.e., the solutions 'blow up'). One of the fun features of differential equations is how very simple equations can have very surprising behavior.

Solve the initial value problem

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

Solution: We can separate the variables by moving all the y's to one side and the t's to the other

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

Integrating both sides we get: $-\frac{1}{y} = t + C$

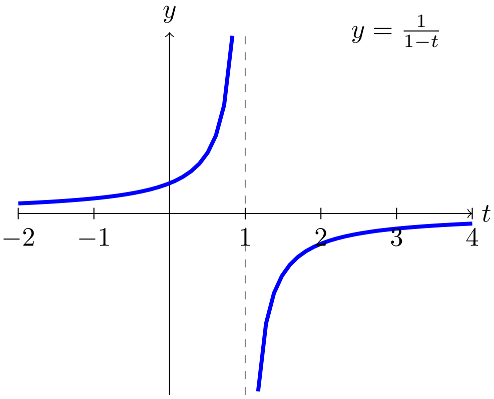
Think: The constant of integration is important, but we only need it on one side.

Solving for y we get the solution:

$$y = -\frac{1}{t+C}.$$

Finally, we use initial condition $y(0) = 1$ to find that $C = -1$. So the solution is: $\boxed{y(t) = \frac{1}{1-t}.}$

We graph this function below. Note that the graph has a vertical asymptote at $t = 1$.



Graph of the function $1/(1-t)$

1.9.1 Technical definition of a solution

Looking at the previous example we see the domain of y consists of two intervals: $(-\infty, 1)$ and $(1, \infty)$. For technical reasons we will require that the domain of a solution consists of exactly one interval. So the above graph really shows **two solutions**:

Solution 1: $y(t) = 1/(1-t)$, where y is in the interval $(-\infty, 1)$
Solution 2: $y(t) = 1/(1-t)$, where y is in the interval $(1, \infty)$

In the example problem, since our IVP had $y(0) = 1$ the solution must have $t = 0$ in its domain. Therefore, solution 1 is the solution to the example's IVP.

1.9.2 Lost solutions

We have to cover one more detail of separable equations. Sometimes solutions get *lost* and have to be recovered. This is a small detail, but you want to pay attention since it's worth 1 easy point on exams and psets.

Example 1.8. In the example $y' = y^2$, we found the solution $y = -\frac{1}{t+C}$. But it is easy to check by substitution that $y(t) = 0$ is also a solution. Since this solution can not be written as $y = -1/(t + C)$ we call it a lost solution.

The simple explanation is that it got lost when we divided by y^2. After all if $y = 0$ it was not legitimate to divide by y^2.

General idea of lost solutions for separable DEs

Suppose we have the differential equation

$$y' = f(x)g(y)$$

If $g(y_0) = 0$ then you can check by substitution that $y(x) = y_0$ is a solution to the DE. It may get lost in when we separate variables because dividing by by $g(y)$ would then mean dividing by 0.

Example 1.9. Find all the (possible) lost solutions of $y' = x(y - 2)(y - 3)$.

Solution: In this case $g(y) = (y-2)(y-3)$. The lost solutions are found by finding all the roots of $g(y)$. That is, the lost solutions are $y(x) = 2$ and $y(x) = 3$.

1.9.3 Implicit solutions

Sometimes solving for y as a function of x is too hard, so we don't!

Example 1.10. Implicit solutions. Solve $y' = \frac{x^3+3x+1}{y^6+y+1}$.

Solution: This is separable and after separating variables and integrating we have

$$\frac{y^7}{7} + \frac{y^2}{2} + y = \frac{x^4}{4} + \frac{3x^2}{2} + x + C.$$

This is too hard to solve for y as a function of x so we leave our answer in this implicit form.

1.9.4 More examples

Example 1.11. Solve $\dfrac{dy}{dx} = xy$.

Solution: Separating variables: $\dfrac{dy}{y} = x\,dx$. Therefore, $\displaystyle\int \frac{dy}{y} = \int x\,dx$, which implies $\ln y = \dfrac{x^2}{2} + C$. Finally after exponentiation and replacing e^C by K we have $\boxed{y = Ke^{x^2/2}.}$

Think: There is a lost solution that was found by some sloppy algebra. Can you spot the solution and the sloppy algebra?

Example 1.12. Solve $\dfrac{dy}{dx} = x^3y^2$.

Solution: Separating variables and integrating gives: $-\frac{1}{y} = \frac{x^4}{4} + C$. Solving for y we have

$$y = -\frac{4}{x^4 + 4C}.$$

There is also a lost solution: $y(x) = 0$.

Example 1.13. Solve $y' + p(x)y = 0$.

Solution: We first rewrite this so that it's clearly separable: $\dfrac{dy}{y} = -p(x)\,dx$. After the usual separation and integration we have

$$\log(|y|) = -\int p(x)\,dx + C$$

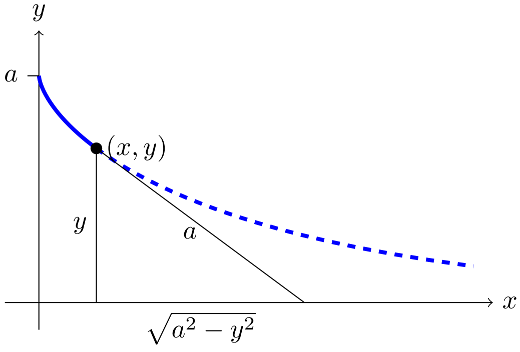
Therefore, $|y(x)| = e^C e^{-\int p(x)\,dx}$ and $y(x) = 0$ is a lost solution.

1.10 Geometric Applications of DEs

Since the slope of a curve is given by its derivatives, we can often use differential equations to describe curves.

Example 1.14. An heavy object is dragged through the sand by rope. Suppose the object starts at $(0, a)$ with the puller at the origin, so the rope has length a. The puller moves along the x-axis so that the rope is always taut and tangent to the curve followed by the object. This curve is called a tractrix. Find an equation for it.

Solution: Since the rope is tangent to the curve, its slope is $\dfrac{dy}{dx}$. Also, computing the slope geometrically as rise/run, the diagram below shows that $\dfrac{dy}{dx} = -\dfrac{y}{\sqrt{a^2-y^2}}$.



The tractrix

Thus, $-\dfrac{\sqrt{a^2-y^2}}{y}\,dy = dx$. Integrating (details below) we get

$$a\ln\left(\frac{a+\sqrt{a^2-y^2}}{y}\right) - \sqrt{a^2-y^2} = x + C.$$

The initial position $(x, y) = (0, a)$ implies $C = 0$. Therefore, $\boxed{x = a\ln\left(\frac{a+\sqrt{a^2-y^2}}{y}\right) - \sqrt{a^2-y^2}.}$

To finish the problem, we show that the integral is what we claimed it was:

Let $I = -\displaystyle\int \frac{\sqrt{a^2-y^2}}{y}\,dy$.

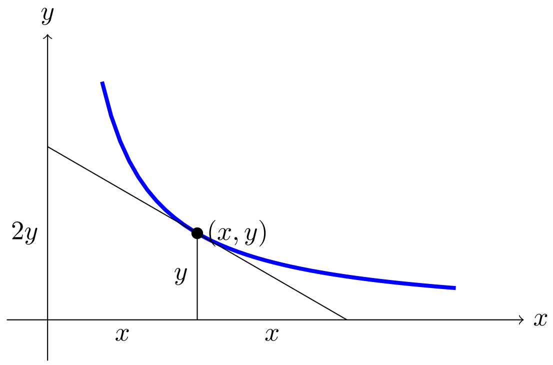
Now use the trig. substitution: $y = a\sin u$:

$$\begin{aligned}\Rightarrow\ I &= -\int \frac{a\cos u}{a\sin u} a\cos u\,du = -a\int \frac{\cos^2 u}{\sin u}\,du \\ &= -a\int \frac{1-\sin^2 u}{\sin u}\,du = -a\int \csc u - \sin u\,du \\ &= a\ln(\csc u + \cot u) - a\cos u\end{aligned}$$

Back substituting we get $I = -\sqrt{a^2-y^2} + a\ln\left(\dfrac{a+\sqrt{a^2-y^2}}{y}\right)$, which is what we claimed above.

Example 1.15. Suppose $y = y(x)$ is a curve in the first quadrant and that the part of the curve's tangent line that lies in the first quadrant is bisected by the point of tangency. Find and solve the DE for this curve.

Solution: The figure shows the piece of the tangent bisected by the point (x, y) on the curve. Thus the slope of the tangent $= \dfrac{dy}{dx} = \dfrac{-y}{x}$. This differential equation is separable and is easily solved: $\boxed{y = C/x.}$

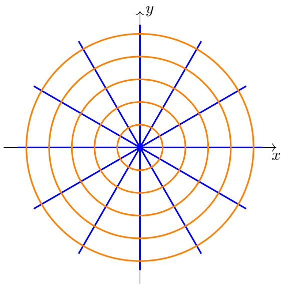


1.11 Orthogonal trajectories

This is mostly taken from the 18.03 Supplementary Notes by Arthur Mattuck.

Given a one-parameter family of plane curves, its orthogonal trajectories are another one-parameter family of curves, each one of which is perpendicular to all the curves in the original family.

Example 1.16. Take the family consisting of all circles having center at the origin, i.e., the one-parameter family of curves $x^2 + y^2 = c^2$. We know that all the rays from the origin are orthogonal to all the circles. That is the orthogonal trajectories to the circles are all the rays (half-lines) starting at the origin.



Blue rays are orthogonal to orange circles wherever they meet.

The examples below will show how to find orthogonal trajectories using differential equations.

Orthogonal trajectories arise in different contexts in applications. For example, if the original family represents the lines of force in a gravitational or electrostatic field, its orthogonal trajectories represent the equipotentials, the curves along which the gravitational or electrostatic potential is constant.

To find the orthogonal trajectories for a one-parameter family:

1. Find the ODE $y' = f(x, y)$ satisfied by the family.

2. The orthogonal family has DE $y' = -\dfrac{1}{f((x, y)}$. That is, the solutions of this DE are the orthogonal trajectories to the original family.

This works because at any point (x, y), the original curve has slope $f(x, y)$, so the orthogonal curve must have slope $-1/f(x, y)$ (negative reciprocal).

Example 1.17. Find the orthogonal trajectories to the family of curves $y = c\,x^n$, where n is a fixed positive integer and c an arbitrary constant.

Solution: First note: If $n = 1$, the curves are lines through the origin, so the orthogonal trajectories should be the circles centered at the origin – this will help check our work.

Step 1 is to find the first-order DE of the family of curves. The parameter c cannot be in this DE – it is the parameter in the solutions.

One common trick is to isolate the c and then differentiate with respect to x. Remember when differentiating that y is a function of x.

$$y = c\,x^n \quad \xrightarrow{\text{isolate } c} \quad yx^{-n} = c \quad \xrightarrow{\text{derivative}} \quad y'x^{-n} - nyx^{-n-1} = 0.$$

Now, solving for y' gives $y' = \frac{ny}{x}$. This is the DE for our family of curves.

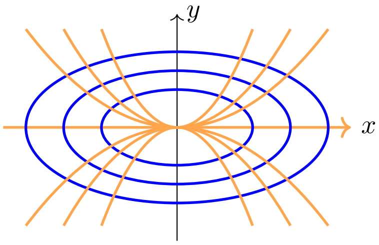
The DE for the orthogonal trajectories is then

$$y' = -\frac{x}{ny}.$$

This is separable. After separating the variables and integrating, we have

$$x^2 + ny^2 = d.$$

We use d as the constant of integration because c was already used. This solution represents a family of ellipses, i.e., for each d we have the equation of an ellipse.



$n = 2$: Orthogonal families $y = cx^2$, $x^2 + 2y^2 = d$.

Note: When $n = 1$, the ellipses are circles centered at the origin, as predicted.

1.12 Definite integral solutions to IVPs

Often we can write the solution to an initial value problem using definite integrals. While *this will not play a major role in 18.03*, it can be quite useful when the integrals are hard to compute or need to be computed numerically. We illustrate with an example.

Example 1.18. Solve $y' = \sin(x^2)\cos(y^2)$, $y(0) = 2$. Give the solution implicitly using definite integrals.

Solution: Seperating variables we have $\dfrac{dy}{\cos(y^2)} = \sin(x^2)\,dx$. We can write the solution as

$$\int_2^y \frac{1}{\cos(u^2)}\,du = \int_0^x \sin(v^2)\,dv.$$

Notes.

1. We used dummy variables in the integrals because x, y are in the limits.

2. The y integral starts at $y = 2$, i.e., the initial y value and the x integral starts at $x = 0$, i.e., at the initial x value.

3. Differentiating both integrals with respect to x, using the fundamental theorem of calculus and the chain rule, we get

$$\frac{1}{\cos(y^2)}\,\frac{dy}{dx} = \sin(x^2).$$

This is equivalent to the original differential equation.

4. The solution is given implicitly, i.e., a function of y = a function of x.

5. Setting $x = 0$ and $y = 2$, the integrals on both sides are 0. That is, the implicit solution satisfies the initial condition.

6. These integrals cannot be computed in terms of our usual elementary functions, but they are easily computed numerically.

2 Linear systems: input-response models

2.1 Goals

1. Be able to classify a first-order differential equation as linear or nonlinear.
2. Be able to put a first-order linear DE into standard form.
3. Be able to use the variation of parameters formula to solve a first-order linear DE.
4. Be able to explain why the superposition principle holds for first-order linear DEs
5. Be able to use the superposition principle to solve a first-order linear DE by breaking the input into pieces.

2.2 Linear first-order differential equations

To start with we will define linear first-order equations by their form. Soon we will understand them by their properties. In particular, you should be on the lookout for the statement of the superposition principle and in later topics for the conceptual definition of linearity.

2.2.1 General and standard form of first-order linear differential equations

Definition. The general first-order linear differential equation has the form

$$A(t)\frac{dy}{dt} + B(t)y(t) = C(t). \tag{3}$$

As long as $A(t) \neq 0$ we can simplify the equation by dividing by $A(t)$. This gives the standard form of a first-order linear differential equation.

$$\frac{dy}{dt} + p(t)y(t) = q(t). \tag{4}$$

Most often when working with linear DEs we will need to put it in the standard form in Equation 4.

2.2.2 Terminology and notation

The functions $A(t)$, $B(t)$ in Equation 3 and $p(t)$ in Equation 4 are called the coefficients of the differential equation. If A and B (or p) are constants, i.e., do not depend on the variable t, then we say the equation is a constant coefficient differential equation.

Notice that the functions $C(t)$ or $q(t)$ on the right-hand side of the equations are not called coefficients and do not have to be constant, even in a constant coefficient DE.

2.2.3 Homogeneous/inhomogeneous

If $C(t) = 0$ in Equation 3 then the resulting equation:

$$A(t)y' + B(t)y = 0$$

is called homogeneous. Otherwise the equation is called inhomogeneous.

Note: Homogeneous is not the same word as homogenous (or homogenized). In homogeneous the syllable 'ge' is pronounced with a long e and is stressed, while the syllable 'mo' is stressed in homogenous.

2.2.4 Identifying first-order linear equations

The Equations 3 and 4 have the form that y' and y occur separately and only as first powers.

Example 2.1. The following differential equations are all linear:

Linear: $y' = ky$; $y' + e^{\sin(t)}y = t^2$; $y' + t^2y = t^3$.

And the following are all non-linear:

Non-linear: $y' + y^2 = t$; $(y')^2 + y = t$; $y'y = t$.

Notice that the coefficient functions in a linear DE are not restricted in any way, but that y and y' never occur in the same term and only have first powers.

Example 2.2. Modeling a population of oryx. A population of oryx has a natural growth rate k in units of 1/year and they are harvested at a constant rate of h oryxes/year. Construct a first-order differential equation modeling the population over time.

An Oryx *gazella*, also known as a Gemsbok

Solution: Let $y(t)$ be the oryx population. By *natural growth rate* we mean that without any outside influences population grows at a rate proportional to itself, i.e., $y' = ky$. The harvesting changes the growth rate by removing oryx at the rate h. Combining the two rates we have

$$y' = ky - h.$$

This is a first-order linear DE. In standard form it reads

$$y' - ky = -h.$$

2.3 Solving first-order linear equations

2.3.1 The variation of parameters formula

We start by giving a formula for the solution to a first-order linear DE in standard form. The differential equation $y' + p(t)y = q(t)$ has solution

$$y(t) = y_h(t) \int \frac{q(t)}{y_h(t)}\, dt + Cy_h(t), \text{ where } y_h(t) = e^{-\int p(t)\, dt}. \tag{5}$$

The function $y_h(t)$ is the solution to the associated homogeneous equation:

$$y_h' + p(t)y_h = 0.$$

Notes: 1. As usual for a first-order DE, the solution is a one parameter family of functions.

2. The formula in Equation 5 is called the variation of parameters formula. The reason for the name comes from the method of deriving it that we give in the last section of this topic's notes.

Warning: The variation of parameters formula is quite beautiful, but don't be seduced into using it in every situation. Because it involves integration it is, generally speaking, our method of **last resort**. When we focus on constant coefficient equations we will learn easier and more informative techniques.

2.3.2 Examples

Example 2.3. Solve $y' + ky = k$, where k is a constant.

Solution: In this case $p(t) = k$ is a constant. The homogeneous solution is

$$y_h(t) = e^{-\int k\, dt} = e^{-kt}.$$

Therefore, the general solution to the DE is

$$\begin{aligned} y(t) = y_h(t) \int q(t)/y_h(t)\, dt + Cy_h(t) &= e^{-kt} \int \frac{k}{e^{-kt}}\, dt + Ce^{-kt} \\ = e^{-kt} \int ke^{kt}\, dt + Ce^{-kt} = e^{-kt} \cdot e^{kt} + Ce^{-kt} &= \boxed{1 + Ce^{-kt}} \end{aligned}$$

(Again: don't get too attached to this technique, later we will learn better techniques for solving constant coefficient equations.)

Example 2.4. Solve $y' + ky = kt$, where k is a constant.

Solution: y_h is the same as in the previous example. Therefore,

$$y(t) = e^{-kt} \int kte^{kt}\, dt + Ce^{-kt} = \left(t - \frac{1}{k}\right) + Ce^{-kt}.$$

(We computed this integral using integration by parts.)

2.4 Superposition principle

We start by defining the terms superposition, input and output.

Superposition is a fancy way of describing adding together multiples of two functions.

Examples: 1. The function $q(t) = 3t + 4t^2$ is a superposition of t and t^2.

2. If $q_1(t)$ and $q_2(t)$ are functions then $q(t) = 3q_1 + 4q_2$ is a superposition of q_1 and q_2.

3. If $q_1(t)$ and $q_2(t)$ are functions and c_1 and c_2 are constants then $q(t) = c_1q_1 + c_2q_2$ is a superposition of q_1 and q_2.

We will also say that $q = c_1q_1 + c_2q_2$ is a **linear combination** of q_1 and q_2.

Suppose we have the first-order linear differential equation

$$y' + p(t)y = q(t). \tag{6}$$

We will often call the $q(t)$ the **input**. We will then call $y(t)$ the **output** of the system to the input q. Of course, $y(t)$ is nothing more than the solution to the DE. In Topic 3 we will expand on the notions of input and output.

The superposition principle is easy but extremely important! It concerns the linear DE in Equation 6 with different inputs $q = q_1$, $q = q_2$ and $q = c_1q_1 + c_2q_2$.

Superposition principle. If

$$y_1 \text{ is a solution of the DE } y' + p(t)y = q_1(t)$$

and

$$y_2 \text{ is a solution of the DE } y' + p(t)y = q_2(t)$$

then for any constants c_1, c_2 we have

$$c_1y_1 + c_2y_2 \text{ is a solution of the DE } y' + p(t)y = c_1q_1(t) + c_2q_2(t).$$

Important note: Notice that the coefficient $p(t)$ is the same for all the DEs.

In words the superposition principle says: For first-order linear DEs

> If the input q_1 has output y_1 and the input q_2 has output y_2 then the input $c_1q_1 + c_2q_2$ has output $c_1y_1 + c_2y_2$

An even simpler formulation is:

> For linear DEs superposition of inputs gives superposition of outputs.

2.4.1 Proof of the superposition principle

First note that saying y_1 is a solution to $y' + py = q_1$ simply means $y_1' + py_1 = q_1$ and likewise for y_2.

To prove the superposition principle we have to verify that $y = c_1y_1 + c_2y_2$ is indeed a solution to $y' + py = c_1q_1 + c_2q_2$. We do this by substitution:

$$\begin{aligned} y' + py &= (c_1y_1 + c_2y_2)' + p(c_1y_1 + c_2y_2) \\ &= c_1y_1' + c_2y_2' + c_1py_1 + c_2py_2 \\ &= c_1(y_1' + py_1) + c_2(y_2' + py_2) \\ &= c_1q_1 + c_2q_2. \end{aligned}$$

The last equality follows because of our assumption that $y_1' + py_1 = q_1$ and the similar assumption for y_2. Now, looking at the first and last terms in this string of equalities we see that we have proved the superposition principle.

Example 2.5. Solve the linear DE $y' + 2y = 2 + 4t$.

Solution: You can easily check that

$$y' + 2y = 1 \text{ has solution } y_1 = 1/2 + C_1 e^{-2t}$$
$$y' + 2y = t \text{ has solution } y_2 = t/2 - \frac{1}{4} + C_2 e^{-2t}$$

The input $2 + 4t$ is a linear combination of the inputs 1 and t, so by the superposition principle the solution to the DE is a linear combination of the outputs y_1, y_2

$$y = 2y_1 + 4y_2 = 1 + 2C_1 e^{-2t} + 2t - 1 + 4C_2 e^{-2t} = 2t + Ce^{-2t}.$$

In the last equality we combined all of the coefficients of e^{-2t} into a single symbol C.

2.5 An extended example

Example 2.6. (Heat diffusion.) I put my root beer in a cooler, but after a while it still gets warm. Let's model its temperature using a differential equation.

Solution: First we need to name the function that measures the temperature:

$$\text{Let } x(t) = \text{root beer temperature at time } t.$$

The simplest model of this situation is Newton's law of cooling. It says that the rate the temperature of the root beer changes is proportional to the difference between the temperatures of the root beer and its environment. In symbols, let $E(t)$ be the temperature of the environment, then (using 'dot' notation)

$$\dot{x}(t) = -k(x(t) - E(t)),$$

where k is the constant of proportionality. Rearranging this equation it becomes

$$\dot{x} + kx = kE(t).$$

This is a first-order linear DE in standard form!

Example 2.7. Suppose the environment in the previous example is $E(t) = 60 + 6t$, where t is the time in hours from 10 AM. (So the temperature is rising linearly.) To be concrete, let's also assume $x(0) = 32$°F and $k = 1/3$. If I want to drink my root beer before it reaches 60°F how much time do I have?

Solution: Our strategy will be to first solve the initial value problem to find $x(t)$ and then use this to determine at what time $x(t)$ will be 60.

From the previous example we know that

$$\dot{x} + kx = kE \qquad \text{so,} \qquad \dot{x} + kx = 60k + 6kt.$$

We could apply the variation of parameters formula directly to this, but the superposition principle will do all the work for us. The input $60k+6kt$ is a superposition of the inputs from

Examples 2.3 and 2.4. Therefore, the solution (output) is a superposition of the outputs from those examples. We know:

From Example 2.3: $\dot{x} + kx = k$ has solution $1 + Ce^{-kt}$.
From Example 2.4: $\dot{x} + kx = kt$ has solution $t - 1/k + Ce^{-kt}$.

Therefore, $\dot{x} + kx = 60k + 6kt$ has solution

$$x(t) = 60(1 + Ce^{-kt}) + 6(t - 1/k + Ce^{-kt}) = 60 + 6t - 6/k + \tilde{C}e^{-kt}.$$

Here we combined all the coefficients of e^{-kt} into one constant $\tilde{C}$. Now we set $k = 1/3$ to get

$$x(t) = 42 + 6t + \tilde{C}e^{-t/3}.$$

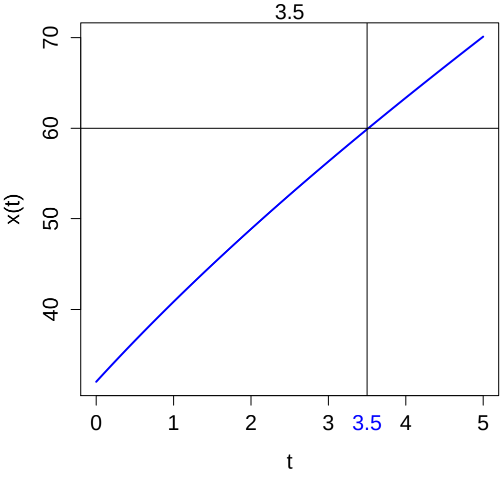
Finally, we use the initial condition to find $\tilde{C}$.

$$x(0) = 42 + \tilde{C} = 32, \text{ so } \tilde{C} = -10.$$

We've found the temperature of the root beer in my cooler is

$$x(t) = 42 + 6t - 10e^{-t/3}.$$

To answer the question we need to compute when $x(t) = 60$. Probably the easiest way to do this is to plot $x(t)$ and see where it crosses $x = 60$. We see this is at about $t = 3.5$. I have until about 1:30 pm to enjoy my drink.



Plot of $x(t)$: $x(t) = 60$ at approximately $t = 3.5$.

Remark redux: We hasten to point out once again that later we will learn faster and nicer techniques for solving equations like this. Techniques involving integration are generally last resorts, to be used when all else has failed.

2.6 Nonlinear equations don't satisfy the superposition principle

The superposition principle is the main reason we focus on linear differential equations. As we have seen in a few examples, it allows us to break the input of a linear equation into pieces and construct the full solution out of the solutions to the pieces. In fact, the superposition principle only holds for linear DEs. We will illustrate this by showing it does not hold for a given nonlinear DE.

Example 2.8. Show that superposition does not hold for the nonlinear equation

$$y' + y^2 = q(t).$$

Solution: We can do this abstractly without actually solving the DE! Suppose $y_1' + y_1^2 = q_1$ and $y_2' + y_2^2 = q_2$. If superposition held for this equation then we would have

$$(y_1 + y_2)' + (y_1 + y_2)^2 = q_1 + q_2.$$

But it's easy to see this equation is not true:

$$\begin{aligned}(y_1 + y_2)' + (y_1 + y_y)^2 &= y_1' + y_2' + y_1^2 + 2y_1y_2 + y_2^2 \\ &= y_1 + y_1^2 + y_2 + y_2^2 + 2y_1y_2 \\ &= q_1 + q_2 + 2y_1y_2 \\ &\neq q_1 + q_2.\end{aligned}$$

What went wrong here? One way to say it is that superposition works for linear equations because the terms in the sum do not really interact. That is, in expressions like $(y_1 + y_2)' = y_1' + y_2'$ and $p(y_1 + y_2) = py_1 + py_2$ the effect on y_1 is exactly what it would be if y_2 was not there. On the other hand in the expression $(y_1 + y_2)^2 = y_1^2 + 2y_1y_2 + y_2^2$ the term $2y_1y_2$ represents an interaction between y_1 and y_2. That is, the effect of squaring on y_1 is affected by the presence of y_2.

2.7 Definite integral solutions to linear initial value problems

Consider the linear IVP

$$y' + p(t)y = q(t); \quad y(0) = y_0.$$

We can solve this equation by two methods.

Method 1: Use the variation of parameters formula in Equation 5 to find the general solution and then use the initial condition to solve for C.

Method 2: Use definite integrals in the variation of parameters formula to give the solution directly. We show how this is done: Take

$$y_h(t) = e^{-\int_{t_0}^t p(u)\,du}.$$

(This is chosen so that $y_h(0) = 1$.) Then

$$y(t) = y_h(t) \int_0^t \frac{q(u)}{y_h(u)}\,du + y_0 \cdot y_h(t).$$

Notes. 1. This form of the solution is well-suited for numerical computation.

2. We stated the problem with initial condition at $t = 0$, but we could have been more general and take $y(t_0) = y_0$.

Here is an example that illustrates both these points. Note that we don't compute the integral exactly, but we can still use the computer to compute approximate values of the solution.

Example 2.9. Solve the initial value problem $x^2y' + xy = \sin x; \quad y(1) = y_0$.

Solution: First we need to convert the DE to standard form:

$$y' + \frac{1}{x}y = \frac{\sin x}{x^2}.$$

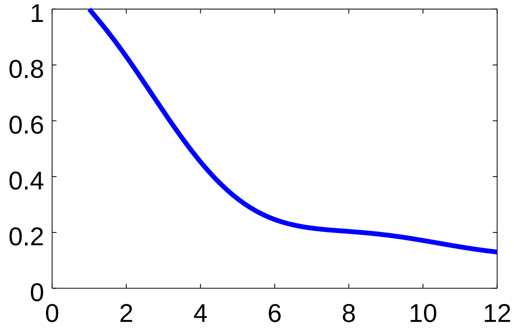
The homogeneous solution is

$$y_h = e^{-\int_1^x \frac{1}{u}\,du} = \frac{1}{x}.$$

So the variation of parameters formula gives

$$y = \frac{1}{x}\int_1^x \frac{\sin u}{u}\,du + \frac{y_0}{x}.$$

There is no closed form for the integral, but we can still use calculus to know a lot about this integral and to compute its value to any desired degree of accuracy. Here is a plot we made in Matlab (actually Octave) using its numerical integration function `quad()`. The initial value is $y_0 = 1$.



Plot of $y = \frac{1}{x}\left(\int_1^x \frac{\sin u}{u}\,du + 1\right)$

2.8 Proof of variation of parameters formula

(You are not responsible for knowing this yet. We will come back to it when we study systems of linear equations.)

One proof that the Equation 5 solves the DE in Equation 4 is by substitution. It's not difficult to plug the formula for $y(t)$ into the differential equation and check that it works. Of course, this is not a very satisfying proof because it fails to answer the question of how we might arrive at such a formula in the first place. Here is another proof that gives more insight.

First we solve the homogeneous equation

$$y' + p(t)y = 0.$$

This equation is separable and easy to solve. We do the algebra quickly: The equation can be written as $y' = -p(t)y$. Separating variables gives: $dy/y = -p(t)dt$. Integrating gives:

$\ln(y) = -\displaystyle\int p(t)\,dt + C$. Now exponentiation gives the general solution to the homogeneous equation:

$$y(t) = Ce^{-\int p(t)\,dt}$$

To avoid writing integrals repeatedly we let $y_h(t) = e^{-\int p(t)\,dt}$. So the general homogeneous solution is $y(t) = Cy_h(t)$.

Now consider the inhomogeneous equation

$$y' + p(t)y = q(t).$$

This is not separable, so we need to do something else. The philosophy behind variation of parameters is to use what we already know. What we know is the homogeneous solution, so we guess that the solution is of the form

$$y(t) = v(t)y_h(t).$$

What we've done is to turn the parameter C in the homogeneous solution into a variable v which depends on t. Hence the name variation of parameters.

Once we've guessed a solution, we substitute it into the inhomogeneous equation to see if we can solve for a $v(t)$ that works. The left-hand side of the inhomogeneous equation is

$$\begin{aligned} y' + p(t)y &= (v(t)y_h(t))' + p(t)v(t)y_h(t) \\ &= v'y_h + vy_h' + pvy_h \\ &= v'y_h + v(y_h' + py_h) \\ &= v'y_h \quad \text{(since } y_h' + py_h = 0). \end{aligned}$$

Equating the left-hand side with the right-hand side we have $v'(t)y_h(t) = q(t)$. This is easy to solve for $v(t)$:

$$v'(t) = q(t)/y_h(t) \;\Rightarrow\; v(t) = \int \frac{q(t)}{y_h(t)}\,dt + C.$$

Now we put this back into our definition of $y(t)$

$$y(t) = v(t)y_h(t) = y_h(t)\left(\int \frac{q(t)}{y_h(t)}\,dt + C\right) = y_h(t)\int \frac{q(t)}{y_h(t)}\,dt + C\,y_h(t).$$

This is the variation of parameters formula we wanted to derive.

3 Input-response models continued

3.1 Goals

1. Be able to use the language of systems and signals.

2. Be familiar with the physical examples in these notes.

3.2 Introduction

In ES.1803 we will use the engineering language of systems and signals. This topic is mostly devoted to learning the vocabulary for this. Our strategy will be to ingrain the words system, signal, input, output (or response) by looking at a series of examples.

It is important to note that these are not mathematical terms and have no formal mathematical definition. Rather, they are engineering terms that will help us organize our thinking when we talk about specific examples. For example, for any given physical model the choice of what to call the input is somewhat arbitrary in the mathematical sense, but usually clear in the engineering sense. What this means in practice is that whenever we need to be mathematically precise, we'll have to say explicitly what we mean by system, input and output. Nonetheless, we'll find the language quite useful. And, in fact, there will be very little confusion when we use these terms.

Another important point is that, in general we will use this language only for **constant coefficient equations** like any of the following:

$$\begin{aligned}&y' + 3y = cos(t),\\ &y' + ky = q(t), \text{ where } k \text{ is a constant},\\ &my'' + by' + ky = F(t), \text{ where } m,\ b,\ k \text{ are constants.}\end{aligned}$$

3.3 Signals

By **signal** we will simply mean any function of time.

Familiar examples are sound, which is a time varying pressure wave; AM radio signals, where the amplitude of the radio wave varies in time; and FM radio signals, where the frequency of the radio wave varies in time. All of these examples agree with the common definition of signal as something conveying information over time.

In ES.1803 two recurring examples will be the position of a mass oscillating at the end of a spring or the temperature of a body over time. Both of these are clearly functions of time, and, if you think about it, both are conveying information.

3.4 System, input, output (response) by example

We'll now give a series of examples to try to draw out how we use these terms. Remember, even though these choices are natural, they are physical and not mathematical. The key point is that in physical setups we can choose the input and response to be what makes the most sense physically. This needs to be fully specified if there is any chance of confusion.

Example 3.1. Recall the example of my root beer from Topic 2. We have the following model.

$$x' + kx = kE(t), \tag{7}$$

where $x(t)$ is the temperature of the root beer over time and $E(t)$ was the temperature of the environment. Let's describe what we'll choose to be the input, output and system for this setup.

Input signal: I'm interested in how the temperature of my root beer is affected by the environment. With this in mind it is natural to consider $E(t)$ to be the input signal. In general we will shorten this by saying that $E(t)$ is the input.

Output signal or response of the system: The temperature of the root beer changes in response to the input $E(t)$. We're interested in the function $x(t)$, so we call it the response of the system to the input. We usually simplify this by calling it the response or output.

System: The system is the 'mechanism' that that converts the input to the output. In this case the mechanism is the root beer together with the insulation quality of the cooler (measured by k) Mathematically the system is modeled by the differential Equation 7.

Think: What happens to k as the insulation quality of the cooler gets better?

Example 3.2. A spring-mass system. Suppose we have a mass on the end of a spring being pushed by an external force $F(t)$. We'll assume there is no damping, so between Newton and Hooke we have the following DE modeling this system $m\dfrac{d^2x}{dt^2} = -kx + F(t)$. In ES.1803 we will typically write this with all the x terms on the left-hand side and $F(t)$ on the right:

$$m\frac{d^2x}{dt^2} + kx = F(t), \tag{8}$$

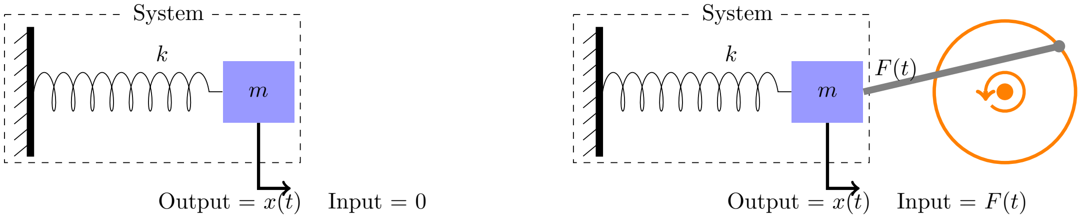
where m=mass, k=spring constant, and $x(t)$=displacement of mass from equilibrium. The following choices seem natural.

System: The spring and mass along with the linkage to the force.

Input: The external force $F(t)$.

Output or response: $x(t)$ the position of the mass over time.

The following figures illustrate this with zero and nonzero input.



Systems with 0 and nonzero input

Example 3.3. Money in the bank. Let $A(t)$ be the amount of money in my retirement account at time t. Suppose also, that interest is paid continuously at the rate r in units of 1/year and that I'm depositing into the account at the rate of $q(t)$ in units of \$/year. While my son was in college $q(t)$ was small. When I retire it will be negative!

Without any deposits or withdrawals $A(t)$ grows exponentially, modeled by $A' = rA$. If we include the deposit rate $q(t)$, we have $A' = rA + q(t)$. We write this with all the A terms on the left and the input $q(t)$ on the right.

$$A' - rA = q(t).$$

Notice that for exponential growth the sign on the rA term is negative. For this situation we will say:

System: Money in the bank earning interest.

Input: The deposit rate $q(t)$.

Output or response: The amount of money in the bank $A(t)$.

3.4.1 Relationship between engineering and mathematical language

We summarize the relationship between the engineering language and the mathematical language as follows:

$$\boxed{\text{Physical: system/input produces a response (output)}} \quad \leftrightarrow \quad \boxed{\text{Model: DE has a solution}}$$

3.4.2 Mathematical input

This is a math class and we may have the differential equation

$$y' + ky = q(t)$$

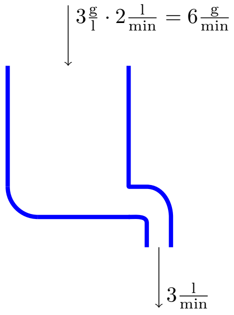
that did not arise from a modeling physical situation. In that case, we will allow ourselves to call the right-hand side $q(t)$ the input and the solution of the DE $y(t)$ the output or response. We will think of $q(t)$ as the mathematical input.

3.5 Worked examples

We'll now work some examples introducing several physical setups that we'll use regularly in this class.

Example 3.4. Mixing tanks. Suppose we have a tank which initially contains 60 liters of pure water. We start adding brine with a concentration of 3 g/liter at the rate of 2 liter/min. While we do this solution leaves the tank at the rate of 3 liter/min. (So the tank will be empty after 60 minutes.)

Assuming instantaneous mixing, find the concentration $C(t)$ of salt in the tank as a function of time.


$3\frac{\text{g}}{\text{l}} \cdot 2\frac{\text{l}}{\text{min}} = 6\frac{\text{g}}{\text{min}}$
$3\frac{\text{l}}{\text{min}}$

Mixing tank with inflow and outflow.

Solution: One key lesson in this example is to work with the amount of salt in the tank not the concentration. This is because when you combine solutions the amounts add, but the concentrations do not. At the end we can go back and compute the concentration from the amount and the volume.

Let t be the time in minutes and let $x(t)$ be the amount of salt in the tank at time t in grams. Since 2 liter/min. is entering the tank and 3 liter/min. is leaving the tank, the tank is emptying at the rate of 1 liter/min., i.e., the volume of solution in the tank is given by $V(t) = 60 - t$ liters. The concentration is $C(t) = x(t)/V(t)$.

We know that

$$x'(t) = \text{ rate salt enters the tank } - \text{ rate salt leaves the tank.}$$

We can easily compute these rates:

$$\begin{aligned} \text{Rate-in } &= 3\,\frac{\text{g}}{\text{liter}} \cdot 2\,\frac{\text{liter}}{\text{min}} = 6\,\frac{\text{g}}{\text{min}} \\ \text{Rate-out } &= 3\,\frac{\text{liter}}{\text{min}} \cdot \frac{x(t)}{V(t)}\,\frac{\text{g}}{\text{liter}} = \frac{3x}{60-t}\,\frac{\text{g}}{\text{min}}, \end{aligned}$$

Putting this together we have the DE $x'(t) = 6 - \dfrac{3x}{60-t}$. As usual we move all the x terms to the left and get the first-order linear initial value problem

$$x'(t) + \frac{3}{60-t}\,x = 6; \qquad x(0) = 0.$$

This is a first-order linear equation and we can solve it using the variation of parameters formula. We could use the method of finding the general solution and then using the initial condition to find C. Instead, we'll practice the definite integral method. First we find the homogenous solution:

$$x_h(t) = e^{-\int_0^t 3/(60-u)\,du} = e^{3\ln(60-u)|_0^t} = \frac{(60-t)^3}{60^3}.$$

The variation of parameters formula (in definite integral form) is

$$\begin{aligned} x(t) &= x_h(t)\left(\int_0^t \frac{q(u)}{x_h(u)}\,du + x_0\right) \\ &= \frac{(60-t)^3}{60^3}\left(\int_0^t \frac{6}{(60-u)^3/60^3}\,du + 0\right) \\ &= 6(60-t)^3\int_0^t \frac{1}{(60-u)^3}\,du \\ &= 6(60-t)^3\left[\frac{1}{2(60-u)^2}\right]_0^t \\ &= 3(60-t) - \frac{3(60-t)^3}{60^2}. \end{aligned}$$

To answer the question asked:

$$\boxed{C(t) = \frac{x(t)}{V(t)} = 3 - \frac{3(60-t)^2}{60^2}.}$$

Of course, this model is only valid until $t = 60$ when the tank will be empty.

Example 3.5. A useful format. Consider the exponential growth equation $x' = 3x$ with initial time $t = 5$ and initial condition $x(5) = 2$. A convenient way to write the solution to initial value problem is

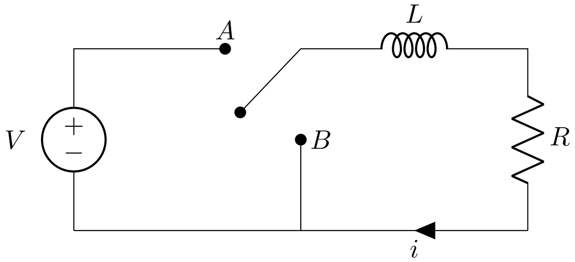
$$x(t) = 2e^{3(t-5)}.$$

This is easy to check by substitution. The point is that since the initial condition is given for $t = 5$ it's easiest to write the solution in terms of $(t - 5)$. This way the coefficient in front of the exponential is just the initial value of x.

Example 3.6. Circuits. (Example of discontinuous input.) An LR circuit is a simple circuit with an inductor L, a resistor R and voltage source V. The differential equation that models the current i is

$$L\frac{di}{dt} + Ri = V.$$

Consider the circuit shown. Assume compatible units and $L = 2$, $R = 4$ and $V = 8$. Also assume that before the switch is closed there is no current in the circuit. At $t = 0$ the switch is moved to position A. Then at $t = 1$ the switch is moved to position B.



Find the current $i(t)$ by writing and solving a differential equation that models this system.

Solution: Each time the switch is moved the input voltage changes. We can write the initial value problem as

$$2i' + 4i = \begin{cases} 8 & \text{for } 0 < t < 1 \\ 0 & \text{for } 1 < t \end{cases}, \qquad \text{with IC } i(t) = 0 \text{ for } t < 0.$$

The format of the input above is called **cases format**. Since the input is given in cases we must solve in cases.

Case (i) For $0 < t < 1$ the DE is: $2i' + 4i = 8$; $i(0) = 0$.
We can solve this using the variation of parameters formula (or by inspection), later we will learn easier techniques: $i(t) = 2 + Ce^{-2t}$. Using the initial condition: $i(0) = 0 = 2 + C$, so $C = -2$. Thus, $\boxed{i(t) = 2 - 2e^{-2t}.}$

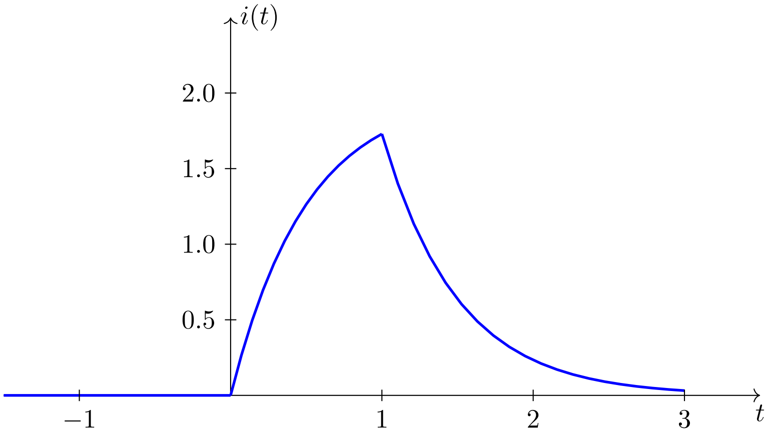
To get the initial condition for the next case we find the value of $i(t)$ at the end of this interval: $i(1) = 2 - 2e^{-2}$.

Case (ii) For $1 < t$ the DE is: $2i' + 4i = 0$; $i(1) = 2 - 2e^{-2}$.
Following the format in Example 3.5 we can write the solution to this as
$\boxed{i(t) = i(1)e^{-2(t-1)} = (2 - 2e^{-2})e^{-2(t-1)}.}$

Writing the full solution in **cases format** we have:

$$i(t) = \begin{cases} 0 & \text{for } t < 0 \\ 2 - 2e^{-2t} & \text{for } 0 < t < 1 \\ (2 - 2e^{-2})e^{-2(t-1)} & \text{for } 1 < t \end{cases}$$

Here's a graph of this solution.



4 Complex numbers and exponentials

4.1 Goals

1. Do arithmetic with complex numbers.
2. Define and compute: magnitude, argument and complex conjugate of a complex number.
3. Be fluent in the use of Euler's formula.
4. Write sine and cosine in terms of complex exponentials ('inverse Euler formulas').
5. Convert complex numbers back and forth between rectangular and polar form.
6. Compute nth roots of complex numbers.

4.2 Motivation

The equation $x^2 = -1$ has no real solutions, yet in ES.1803 we will see that this equation arises naturally and we will want to know its roots. As you may already know, we'll introduce a new symbol for the roots and call it a complex number.

Definition: The symbols $\pm i$ will stand for the solutions to the equation $x^2 = -1$. We will call these new numbers complex numbers. We will also write

$$\sqrt{-1} = \pm i$$

Notes: 1. i is also called an **imaginary number**. This is a historical term. These are perfectly valid numbers that don't happen to lie on the real number line.

2. Our motivation for using complex numbers is not the same as the historical motivation. Mathematicians were willing to say $x^2 = -1$ had no solutions. The problem was in the formula for the roots of cubics. Where square roots of negative numbers appeared even for the real roots of cubics.

3. Engineers typically use j instead of i. We'll follow mathematical custom and use i in ES.1803.

We're going to look at the algebra, geometry and, most important for us, the exponentiation of complex numbers.

Before starting a systematic exposition of complex numbers we'll work a simple example. If the explanation is not immediately clear, it should become clear as we learn more about this topic.

Example 4.1. Solve the equation $r^2 + r + 1 = 0$

Solution: We can apply the quadratic formula to get

$$r = \frac{-1 \pm \sqrt{1-4}}{2} = \frac{-1 \pm \sqrt{-3}}{2} = \frac{-1 \pm \sqrt{3}\sqrt{-1}}{2} = \frac{-1 \pm \sqrt{3}\, i}{2}.$$

Think: Do you know how to solve quadratic equations by completing the square? This is how the quadratic formula is derived and is well worth knowing!

4.2.1 Fundamental theorem of algebra

One of the reasons for using complex numbers is because, by allowing complex roots, every polynomial has exactly the expected number of roots.

Fundamental theorem of algebra. A polynomial of degree n has exactly n complex roots (repeated roots are counted with multiplicity.)

Example 4.2. We'll illustrate what we mean by this with a few examples.

1. The polynomial $r^2 + 3r + 2$ factors as $(r+1)(r+2)$ therefore its roots are $r = -1$ and $r = -2$. It is a second-order polynomial with 2 roots.

2. The polynomial $r^2 + 6r + 9$ factors as $(r+3)(r+3)$. We say it has the roots -3 and -3. That is it has two roots that happen to be the same. We will also say that -3 is a root of this polynomial with **multiplicity** 2.

3. The polynomial $(r+1)(r+2)(r+3)^2(r^2+1)^2$ has degree 8. Its 8 roots are

$$-1,\ -2,\ -3,\ -3,\ i,\ i,\ -i,\ -i.$$

This example illustrates an important point about polynomials: we prefer to have them in factored form. I think you'll agree that you wouldn't want to find the roots of the polynomial

$$r^8 + 9r^7 + 31r^6 + 57r^5 + 77r^4 + 87r^3 + 65r^2 + 39r + 18.$$

Unless you happened to notice that it was the same as the factored polynomial in Example 4.2(3)! Fortunately, computing packages like Matlab or Octave allow us to find these roots numerically for high order polynomials.

4.3 Terminology and basic arithmetic

Definitions.

- **Complex numbers** are defined as the set of all numbers

$$z = x + yi,$$

  where x and y are real numbers.

- We denote the set of all complex numbers by **C**. (On the blackboard we will usually write $\mathbb{C}$ –this font is called *blackboard bold.*)
- We call x the **real part** of z. This is denoted by $x = \text{Re}(z)$.
- We call y the **imaginary part** of z. This is denoted by $y = \text{Im}(z)$.

Note well: The imaginary part of z is a real number. It **DOES NOT** include the i.

The basic arithmetic operations follow the standard rules. All you have to remember is that $i^2 = -1$. We will go through these quickly using some simple examples. For ES.1803 it is essential that you become fluent with these manipulations.

- **Addition:** $(3 + 4i) + (7 + 11i) = 10 + 15i$
- **Subtraction:** $(3 + 4i) - (7 + 11i) = -4 - 7i$
- **Multiplication:** $(3+4i)(7+11i) = 21+28i+33i+44i^2 = -23+61i$. Here we have used the fact that $44i^2 = -44$.

Before talking about division and absolute value we introduce a new operation called conjugation. It will prove useful to have a name and symbol for this, since we will use it frequently.

Complex conjugation is denoted with a bar and defined by

$$\overline{x + iy} = x - iy.$$

If $z = x + iy$ then its conjugate is $\overline{z} = x - iy$ and we read this as "z-bar $= x - iy$".

Example 4.3. $\overline{3 + 5i} = 3 - 5i$.

The following is a very useful property of conjugation. We will use it in the next example to help with division.

Useful property of conjugation: If $z = x + iy$ then $z\overline{z} = (x + iy)(x - iy) = x^2 + y^2$.

Example 4.4. (Division.) Write $\dfrac{3 + 4i}{1 + 2i}$ in the standard form $x + iy$.

Solution: We use the useful property of conjugation to clear the denominator:

$$\frac{3+4i}{1+2i} = \frac{3+4i}{1+2i} \cdot \frac{1-2i}{1-2i} = \frac{11-2i}{5} = \frac{11}{5} - \frac{2}{5}i.$$

In the next section we will discuss the geometry of complex numbers, which give some insight into the meaning of the magnitude of a complex number. For now we just give the definition.

Definition. The **magnitude** of the complex number $x + iy$ is defined as

$$|z| = \sqrt{x^2 + y^2}.$$

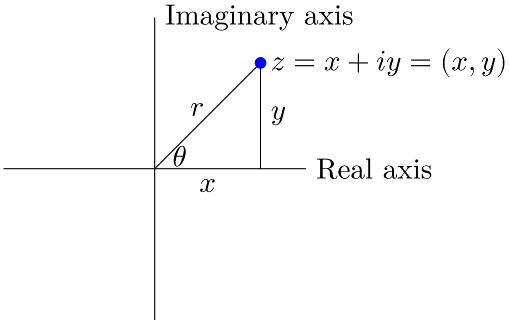
The magnitude is also called the **absolute value** or **norm** or **modulus**.

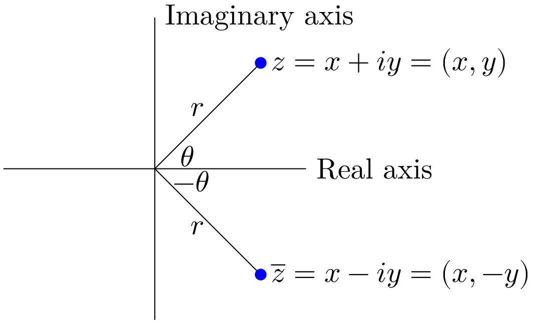
Example 4.5. The norm of $3 + 5i = \sqrt{9 + 25} = \sqrt{34}$.

Note this really well: The norm is the sum of x^2 and y^2 **it does not include the i! Therefore, it is always positive**.

4.4 The complex plane and the geometry of complex numbers

Because it takes two numbers x and y to describe the complex number $z = x + iy$ we can visualize complex numbers as points in the xy-plane. When we do this we call it the complex plane. Since x is the real part of z we call the x-axis the real axis. Likewise, the y-axis is the imaginary axis.





4.5 Polar coordinates

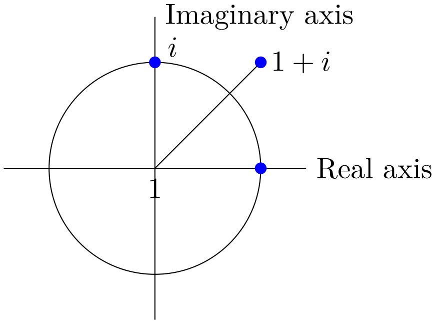
In the figures above we have marked the length r and polar angle θ of the vector from the origin to the point $z = x + iy$. These are the same polar coordinates you saw in 18.02. There are a number of synonyms for both r and θ

$$r = |z| = \text{magnitude} = \text{length} = \text{norm} = \text{absolute value} = \text{modulus}$$
$$\theta = \text{Arg}(z) = \text{ argument of } z = \text{polar angle of } z$$

As in 18.02 you should be able to visualize polar coordinates by thinking about the distance r from the origin and the angle θ with the x-axis.

Example 4.6. In this example we make a table of z, r and θ for some complex numbers. Notice that θ is not uniquely defined since we can always add a multiple of 2π to θ and still be at the same point in the plane.

$z = a + bi$	$r = \lvert z \rvert$	$\theta = \arg(z)$	
1	1	$0,\ 2\pi,\ 4\pi, ...$	Argument $= 0$, means z is along the positive x-axis
i	1	$\pi/2,\ \pi/2 + 2\pi ...$	Argument $= \pi/2$, means z is along the positive y-axis
$1 + i$	$\sqrt{2}$	$\pi/4,\ \pi/4 + 2\pi ...$	Argument $= \pi/4$, means z is along the ray at 45° to the x-axis



4.6 Euler's Formula

Euler's (pronounced 'oilers') formula connects complex exponentials, polar coordinates and sines and cosines. It turns messy trig identities into tidy rules for exponentials. We will use it a lot.

The formula is the following:

$$e^{i\theta} = \cos(\theta) + i\sin(\theta). \tag{9}$$

There are many ways to approach Euler's formula. Our approach is to simply take Equation 9 as the definition of complex exponentials. This is mathematically legal, but does not show that it's a good definition. To do that, we need to show that $e^{i\theta}$ obeys all the rules we expect of an exponential. To do that, we go systematically through the properties of exponentials and check that they hold for complex exponentials.

4.6.1 e^{it} behaves like a true exponential

1. e^{it} differentiates as expected: $\frac{de^{it}}{dt} = ie^{it}$.

Proof. This follows directly from the definition:

$$\frac{de^{it}}{dt} = \frac{d}{dt}(\cos(t) + i\sin(t)) = -\sin(t) + i\cos(t) = i(\cos(t) + i\sin(t)) = ie^{it}.$$

2. $e^{i \cdot 0} = 1$.

Proof. $e^{i \cdot 0} = \cos(0) + i\sin(0) = 1$.

3. The usual rules of exponents hold: $e^{ia}e^{ib} = e^{i(a+b)}$.

Proof. This relies on the cosine and sine addition formulas.

$$\begin{aligned} e^{ia} \cdot e^{ib} &= (\cos(a) + i\sin(a)) \cdot (\cos(b) + i\sin(b)) \\ &= \cos(a)\cos(b) - \sin(a)\sin(b) + i\,(\cos(a)\sin(b) + \sin(a)\cos(b)) \\ &= \cos(a+b) + i\sin(a+b) = e^{i(a+b)}. \end{aligned}$$

4. The definition of $e^{i\theta}$ is consistent with the power series for e^x.

Proof. To see this we have to recall the power series for e^x, $\cos(x)$ and $\sin(x)$. They are

$$\begin{aligned} e^x &= 1 + x + x^2/2! + x^3/3! + x^4/4! + ... \\ \cos(x) &= 1 - x^2/2! + x^4/4! - x^6/6! + ... \\ \sin(x) &= x - x^3/3! + x^5/5! + ... \end{aligned}$$

Now we can write the power series for $e^{i\theta}$ and then split it into the power series for sine and cosine:

$$\begin{aligned} e^{i\theta} &= \sum_0^\infty \frac{(i\theta)^n}{n!} \\ &= \sum_0^\infty (-1)^k \frac{\theta^{2k}}{(2k)!} + i \sum_0^\infty (-1)^k \frac{\theta^{2k+1}}{(2k+1)!} \\ &= \cos(\theta) + i\sin(\theta). \end{aligned}$$

So the Euler formula definition is consistent with the usual power series for e^z.

1-4 should convince you that $e^{i\theta}$ behaves like an exponential.

4.6.2 Complex exponentials and polar form

Now let's turn to the relation between polar coordinates and complex exponentials.

Suppose $z = x + iy$ has polar coordinates r and θ. That is, we have $x = r\cos(\theta)$ and $y = r\sin(\theta)$. Thus we get the important relationship

$$z = x + iy = r\cos(\theta) + ir\sin(\theta) = r(\cos(\theta) + i\sin(\theta)) = re^{i\theta}.$$

This is so important you shouldn't proceed without understanding it. We also record it without the intermediate equation.

$$z = x + iy = re^{i\theta}. \tag{10}$$

Because r and θ are the polar coordinates of (x, y) we call $z = re^{i\theta}$ the polar form of z.

4.6.3 Magnitude, argument, conjugate, multiplication and division are easy in polar form

Magnitude. $|e^{i\theta}| = 1$.

Proof. $|e^{i\theta}| = |\cos(\theta) + i\sin(\theta)| = \sqrt{\cos^2(\theta) + \sin^2(\theta)} = 1$.

In words, this says that $e^{i\theta}$ is always on the unit circle –this is useful to remember!

Likewise, if $z = re^{i\theta}$ then $|z| = r$. You can calculate this, but it should be clear from the definitions: $|z|$ is the distance from z to the origin, which is exactly the same definition as for r.

Argument. If $z = re^{i\theta}$ then $\text{Arg}(z) = \theta$.

Proof. This is again the definition: the argument is the polar angle θ.

Conjugate. $\overline{(re^{i\theta})} = re^{-i\theta}$.

Proof. $\overline{(re^{i\theta})} = \overline{(r(\cos(\theta) + i\sin(\theta)))} = r(\cos(\theta) - i\sin(\theta)) = re^{-i\theta}$.

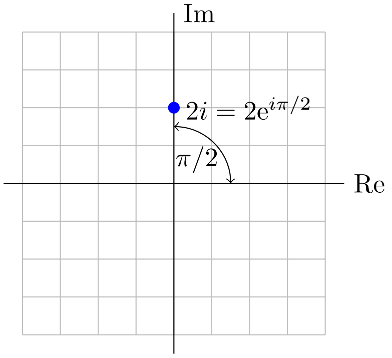
In words: complex conjugation changes the sign of the argument.

Multiplication. If $z_1 = r_1e^{i\theta_1}$ and $z_2 = r_2e^{i\theta_2}$ then $z_1z_2 = r_1r_2e^{i(\theta_1+\theta_2)}$.

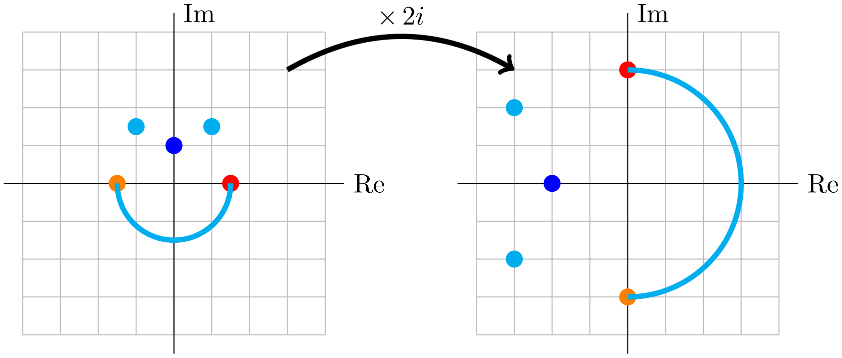
This is what mathematicians call trivial to see, just write the multiplication down. In words, the formula says the for z_1z_2 the magnitudes multiply and the arguments add.

Division. Again it's trivial that $\dfrac{r_1e^{i\theta_1}}{r_2e^{i\theta_2}} = \dfrac{r_1}{r_2}e^{i(\theta_1-\theta_2)}$.

Example 4.7. Multiplication by $2i$. Here's a simple but important example. By looking at the graph we see that the number $2i$ has magnitude 2 and argument $\pi/2$. So, in polar coordinates, it equals $2e^{i\pi/2}$. This means that multiplication by $2i$ multiplies lengths by 2 and adds $\pi/2$ to arguments, i.e., rotates by 90°. The effect is shown in the figures below



$|2i| = 2, \quad \text{Arg}(2i) = \pi/2$



Multiplication by $2i$ rotates by $\pi/2$ and scales by 2

Example 4.8. Raising to a power. Compute (i) $(1+i)^6$; (ii) $\left(\frac{1+i\sqrt{3}}{2}\right)^3$

Solution: (i) $1+i$ has magnitude $|1+i| = \sqrt{2}$ and argument $\text{Arg}(1+i) = \pi/4$, so $1+i = \sqrt{2}e^{i\pi/4}$. Raising to a power is now easy:

$$(1+i)^6 = \left(\sqrt{2}e^{i\pi/4}\right)^6 = 8e^{6i\pi/4} = 8e^{3i\pi/2} = -8i.$$

(ii) $\dfrac{1+i\sqrt{3}}{2} = e^{i\pi/3}$, so $\left(\dfrac{1+i\sqrt{3}}{2}\right)^3 = (1 \cdot e^{i\pi/3})^3 = e^{i\pi} = -1$

4.6.4 Complexification or complex replacement

In the next example we will illustrate the technique of complexification or complex replacement by computing a trigonometric integral. Although, in ES.1803, we are not really concerned with trigonometric integrals, we will use complex replacement regularly in other contexts.

Example 4.9. Use complex replacement to compute $I = \int e^x \cos(2x)\,dx$.

Solution: First we will show the steps for complex replacement. Then, below, we will justify them. We have Euler's formula $e^{2ix} = \cos(2x) + i\sin(2x)$, so $\cos(2x) = \text{Re}(e^{2ix})$. The trick is to replace $\cos(2x)$ by e^{2ix}. We get

$$I_c = \int e^x \cos 2x + ie^x \sin 2x\,dx, \quad \text{where } I = \text{Re}(I_c).$$

Computing I_c is straightforward:

$$I_c = \int e^x e^{i2x}\,dx = \int e^{x(1+2i)}\,dx = \frac{e^{x(1+2i)}}{1+2i}.$$

Now we use polar form to simplify the expression for I_c:

Write $1 + 2i = re^{i\phi}$, where $r = \sqrt{5}$ and $\phi = \text{Arg}(1 + 2i) = \tan^{-1}(2)$ in the first quadrant. Then:

$$I_c = \frac{e^{x(1+2i)}}{\sqrt{5}e^{i\phi}} = \frac{e^x}{\sqrt{5}}e^{i(2x-\phi)} = \frac{e^x}{\sqrt{5}}(\cos(2x-\phi) + i\ \sin(2x-\phi)).$$

Thus, $I = \text{Re}(I_c) = \dfrac{e^x}{\sqrt{5}}\cos(2x-\phi)$.

Justification of complex replacement. The trick comes by cleverly adding a new integral to I as follows. Let $J = \int e^x \sin(2x)\,dx$. Then we let

$$I_c = I + iJ = \int e^x(\cos(2x) + i\sin(2x))\,dx = \int e^x e^{2ix}\,dx.$$

Clearly, $\text{Re}(I_c) = \text{Re}(I + iJ) = I$ as claimed above.

Rectangular coordinates –generally less preferred than polar. We show here the computation in rectangular coordinates –though we hasten to add that in ES.1803 we will almost always prefer polar form because it is easier and gives the answer in a more useable form.

$$\begin{aligned} I_c &= \frac{e^{x(1+2i)}}{1+2i} \cdot \frac{1-2i}{1-2i} = \frac{e^x(\cos(2x) + i\ \sin(2x))(1-2i)}{5} \\ &= \frac{1}{5}e^x(\cos(2x) + 2\sin(2x) + i(-2\cos(2x) + \sin(2x))). \end{aligned}$$

So, $I = \text{Re}(I_c) = \dfrac{1}{5}e^x(\cos(2x) + 2\sin(2x))$.

4.6.5 Nth roots

We are going to need to be able to find the nth roots of complex numbers. The trick is to recall that a complex number has more than one argument, that is we can always add a multiple of 2π to the argument. For example,

$$2 = 2e^{0i} = 2e^{2\pi i} = 2e^{4\pi i} \ldots = 2e^{2n\pi i}$$

Example 4.10. Find all 5 fifth roots of 2.

Solution: In polar form: $\left(2e^{2n\pi i}\right)^{1/5} = 2^{1/5}e^{2n\pi i/5}$. So the fifth roots of 2 are

$$2^{1/5} = 2^{1/5}e^{2n\pi i/5}, \text{ where } n = 0, 1, 2, \ldots$$

The notation is a little strange, because the $2^{1/5}$ on the left side of the equation means the complex roots and the $2^{1/5}$ on the right hand side is a magnitude, so it is the positive real root.

Looking at the right hand side we see that for $n = 5$ we have $2^{1/5}e^{2\pi i}$ which is exactly the same as the root when $n = 0$, i.e., $2^{1/5}e^{0i}$. Likewise $n = 6$ gives exactly the same root as $n = 1$. So we have 5 different roots corresponding to $n = 0, 1, 2, 3, 4$.

$$2^{1/5} = 2^{1/5},\ 2^{1/5}e^{2\pi i/5},\ 2^{1/5}e^{24\pi i/5},\ 2^{1/5}e^{6\pi i/5},\ 2^{1/5}e^{8\pi i/5}.$$

Similarly we can say that, in general, $z = re^{i\theta}$ has N different Nth roots:

$$z^{1/N} = r^{1/N}e^{i\theta/N + i\,2\pi(n/N)} \text{ for } 0, 1, 2, \ldots N-1.$$

Example 4.11. Find the 4 fourth roots of 1.

Solution: $1 = e^{i\,2\pi n}$, so $1^{1/4} = e^{i\,2\pi(n/4)}$. So the 4 different fourth roots are 1, $e^{i\,\pi/2}$, $e^{i\,\pi}$, $e^{i\,3\pi/2}$, $e^{i\,2\pi}$.

When the angles are ones we know about, e.g., 30, 60, 90, 45, etc., we should simplify the complex exponentials. In this case, the roots are 1, i, -1, $-i$.

Example 4.12. Find the 3 cube roots of -1.

Solution: $-1 = e^{i\,\pi + i\,2\pi n}$. So, $(-1)^{1/3} = e^{i\,\pi/3 + i\,2\pi(n/3)}$ and the 3 cube roots are $e^{i\pi/3}$, $e^{i\pi}$, $e^{i5\pi/3}$. Since $\pi/3$ radians is 60° we can simplify:

$$e^{i\pi/3} = \cos(\pi/3) + i\sin(\pi/3) = \frac{1}{2} + i\,\frac{\sqrt{3}}{2} \quad \longrightarrow (-1)^{1/3} = -1,\ \frac{1}{2} \pm \frac{\sqrt{3}}{2}.$$

Example 4.13. Find the 5 fifth roots of $1 + i$.

Solution: $1 + i = \sqrt{2}e^{i(\pi/4 + 2n\pi)}$, for $n = 0, 1, 2, \ldots$. So the 5 fifth roots are

$$(1+i)^{1/5} = 2^{1/10}e^{i\pi/20},\ 2^{1/10}e^{i9\pi/20},\ 2^{1/10}e^{i17\pi/20},\ 2^{1/10}e^{i25\pi/20},\ 2^{1/10}e^{i33\pi/20}.$$

Using a calculator we could write these numerically as $a + bi$, but there is no easy simplification.

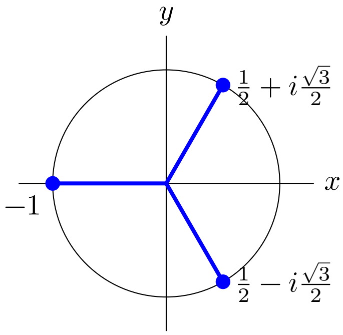
Example 4.14. We should check that our technique works as expected for a simple problem. Find the 2 square roots of 4.

Solution: $4 = 4e^{i\,2\pi n}$. So, $4^{1/2} = 2e^{i\,\pi n}$. So the 2 square roots are $2e^0,\ 2e^{i\pi} = \pm 2$ as expected!

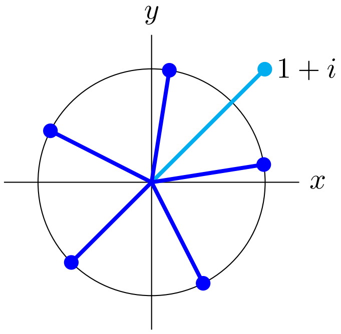
4.6.6 The geometry of nth roots

Looking at the examples above we see that roots are always spaced evenly around a circle centered at the origin. For example, the fifth roots of $1+i$ are spaced at increments of $2\pi/5$ radians around the circle of radius $2^{1/5}$.

Note also that the roots of real numbers always come in conjugate pairs.


y
$\frac{1}{2}+i\frac{\sqrt{3}}{2}$
x
-1
$\frac{1}{2}-i\frac{\sqrt{3}}{2}$

Cube roots of -1



Fifth roots of $1+i$

4.7 Inverse Euler Formula

Euler's formula gives a complex exponential in terms of sines and cosines. We can turn this around to get the inverse Euler formulas.

Euler's formula says:

$$e^{it} = \cos(t) + i\sin(t) \quad \text{and} \quad e^{-it} = \cos(t) - i\sin(t).$$

By adding and subtracting we get:

$$\cos(t) = \frac{e^{it}+e^{-it}}{2} \quad \text{and} \quad \sin(t) = \frac{e^{it}-e^{-it}}{2i}.$$

Warning. We also have the formula $\cos(t) = \text{Re}(e^{it})$ which we used in complex replacement. You want to pay attention to whether this or the inverse Euler formula is appropriate. In general, if you complexified to use complex replacement then at some point you'll need to *decomplexify* by using the formula $\cos(t) = \text{Re}(e^{it})$. If you never complexified then you probably need to use the inverse Euler formula.

5 Homogeneous, linear, constant coefficient differential equations

5.1 Goals

1. Be able to solve homogeneous constant coefficient linear differential equations using the method of the characteristic equation. This includes finding the general real-valued solutions when the roots are complex or repeated.

2. Be able to give the reasoning leading to the method of the characteristic equation.
3. Be able to state and prove the principle of superposition for homogeneous linear equations.
4. For a damped harmonic oscillator be able to map the characteristic roots to the type of damping.
5. Be able to create and interpret pole diagrams.

5.2 Introduction

In this topic we will start our study of constant coefficient differential equations. Most of our examples will look at second-order equations, which can be used to model a rich set of physical situations. Second-order equations are fairly simple computationally, yet feature many of the behaviors that higher order equations display.

5.3 Second-order constant coefficient linear differential equations.

The basic second-order constant coefficient linear differential equation can be written as:

$$mx'' + bx' + kx = f(t), \quad \text{where } m, b,\ k \text{ are constants.}$$

The name says it all:

1. Second-order: obvious.
2. Constant coefficient: because the coefficients m, b, k are constant.
3. Linear: derivatives occur by themselves and to the first power. This is the same rule we had for first-order linear, and, just as in that case, we will see that second-order linear equations follow the superposition principle.
4. Note: the ‘input’ $f(t)$ is not necessarily constant.

Reasons to study second-order linear differential equations:

1. There are a lot of second-order physical systems. For example, for moving particles you need the second derivative to capture acceleration.

2. Many higher order systems are built from second-order components.

3. The computations are easy to do by hand and will help us develop our intuition about second-order equations. This computational and intuitive understanding will guide us when we consider higher order equations.

Remark. For second-order systems we will know how they behave and therefore what the solutions to the DEs should look like. For example, a mass oscillating at the end of a spring is a second-order system and we already have a good sense of what happens when we pull on the mass and let it go. So, in some sense, the math is not telling us that much. However, when you couple together 3 springs you have a sixth-order system and our intuition becomes a bit shakier. If you couple even more springs in a two or three dimensional lattice our intuition is shakier still. The success of our second-order models will give us confidence in our higher-order models. And the techniques used to solve second-order equations will carry over to the higher-order case.

5.4 Second-order homogeneous constant coefficient linear differential equations.

For this topic we will focus on the homogeneous equation (H) given just below.

$$mx'' + bx' + kx = 0. \tag{H}$$

We start with an example which pretty well sums up the general technique. Since this is a first example, we will break the solution into small pieces. In later examples we will give solutions that model what we'll expect in your written work.

Example 5.1. (Solving homogeneous constant coefficient DEs: long form solution.) Solve the DE

$$x'' + 8x' + 7x = 0.$$

Solution: 1. Using the method of optimism we guess a solution of the form $x(t) = e^{rt}$. Note that we have left the r unspecified. Our optimistic hope is that the value of r will come out in the algebra.

2. Substitute our guess (trial solution) into the DE:

$$r^2e^{rt} + 8re^{rt} + 7e^{rt} = 0.$$

Divide by e^{rt} (this is okay, it is never 0) to get the **characteristic equation**

$$r^2 + 8r + 7 = 0.$$

3. This has roots: $r = -7, -1$. Therefore, the method of optimism has found two basic solutions:

$$x_1(t) = e^{-7t}, \qquad x_2 = e^{-t}$$

4. Just below, we will discuss the superposition principle, here we will just apply it to get the **general solution** to the DE:

$$\boxed{x(t) = c_1x_1(t) + c_2x_2(t) = c_1e^{-7t} + c_2e^{-t}.}$$

We remind you that the superposition of x_1 and x_2 is also called a linear combination. We will now explain why it works in this case.

5.5 The principle of superposition for linear homogeneous equations

We will state this as a theorem with a proof. The proof is just a small amount of algebra.

Theorem. The superposition principle part 1. If x_1 and x_2 are solutions to (H) then so are all linear combinations $x = c_1x_1 + c_2x_2$ where c_1, c_2 are constants.

Proof. As we said, the proof is by algebra. Since we are given a supposed solution, we verify it by substitution, i.e., we plug $x = c_1x_1 + c_2x_2$ into (H).

$$\begin{aligned} mx'' + bx' + kx &= m(c_1x_1 + c_2x_2)'' + b(c_1x_1 + c_2x_2)' + k(c_1x_1 + c_2x_2) \\ &= mc_1x_1'' + mc_2x_2'' + bc_1x_1' + bc_2x_2' + kc_1x_1 + kc_2x_2 \\ &= c_1(mx_1'' + bx_1' + kx_1) + c_2(mx_2'' + bx_2' + kx_2) \\ &= c_1 \cdot 0 + c_2 \cdot 0 \\ &\quad (mx_1'' + bx_1' + kx_1 = 0 \text{ by the assumption that } x_1 \text{ solves (H). Likewise for } x_2.) \\ &= 0. \end{aligned}$$

We have verified that $x = c_1x_1 + c_2x_2$ is, in fact, a solution to the homogeneous DE (H).

Superposition = linearity: At this point you should recall the example in Topic 2 where we showed that the nonlinear DE $x' + x^2 = 0$ did not satisfy the superposition principle. It is a general fact that only linear differential equations satisfy the superposition principle.

Example 5.2. (Model solution.) In this example, we suggest a way to give the solutions in your own work. Solve

$$x'' + 4x' + 3x = 0.$$

Solution: Characteristic equation: $r^2 + 4r + 3 = 0$.
Roots: $r = -1, -3$.
Basic solutions: $x_1(t) = e^{-3t}$, $x_2(t) = e^{-t}$.
General solution by superposition: $x(t) = c_1x_1 + c_2x_2 = c_1e^{-3t} + c_2e^{-t}$.

Note. We call the two solutions x_1, x_2 basic or modal solutions.

Suggestion. For the next week or so every time you use this method remind yourself where each step came from (see the solution to Example 5.1.

Every time we learn a new method we want to test it on our favorite DE.

Example 5.3. Test case: exponential decay. Solve $x' + kx = 0$ using the method of the characteristic equation.

Solution: Characteristic equation (try $x = e^{rt}$): $r + k = 0$.
Roots: $r = -k$.
One solution: $x_1(t) = e^{-kt}$
General solution (by superposition): $x(t) = c_1x_1 = c_1e^{-kt}$ (as expected).

In practice, we don't recommend solving this equation with this method. The recommended method is to recognize the DE as the equation of exponential decay and just give the solution.

5.6 Families of solutions

We call $x(t) = c_1e^{2t} + c_2e^{-t}$ a two-parameter family of functions. We will often look for subfamilies with special properties.

Example 5.4. (a) Find all the members in the above family that go to 0 as $t \to \infty$.

(b) Find all the members that go to ∞ as $t \to \infty$.

Solution: (a) All the functions $x(t) = c_2e^{-t}$ (i.e., $c_1 = 0$).

(b) All the functions $x(t) = c_1 e^{2t} + c_2 e^{-t}$, where $c_1 > 0$, c_2 is arbitrary.

5.7 Complex roots

Example 5.5. (Model solution: complex roots) Solve the DE

$$x'' + 2x' + 4x = 0.$$

Solution: 1. Characteristic equation: $r^2 + 2r + 4 = 0$.

2. Roots: $r = (-2 \pm \sqrt{4-16})/2 = -1 \pm \sqrt{3}\, i$.

3. Two basic solutions : $x_1(t) = e^{-t}\cos(\sqrt{3}t)$, $x_2(t) = e^{-t}\sin(\sqrt{3}t)$. Here the exponential e^{-t} uses the real part of the roots and the frequency in the sinusoids $\cos(\sqrt{3}t)$, $\sin(\sqrt{3}t)$ comes from the imaginary part of the roots. All of this will be justified below.

4. General *real-valued* solution by superposition:

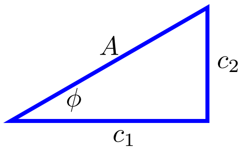
$$x(t) = c_1 x_1(t) + c_2 x_2(t) = c_1 e^{-t}\cos(\sqrt{3}t) + c_2 e^{-t}\sin(\sqrt{3}t).$$

Notes. 1. The damped frequency of oscillation comes from the imaginary part of the roots $\pm\sqrt{3}$.

2. In polar form the solution can be written

$$x(t) = c_1 e^{-t}\cos(\sqrt{3}t) + c_2 e^{-t}\sin(\sqrt{3}t) = Ae^{-t}\cos(\sqrt{3}t - \phi),$$

where A, ϕ, c_1 and c_2 are related by the usual polar triangle with $c_1 = A\cos(\phi)$, $c_2 = A\sin(\phi)$.



Example 5.6. Solve $x'' + 4x = 0$.

Solution: This is the DE for the simple harmonic oscillator a.k.a. a spring-mass system. Using the characteristic equation method:

Characteristic equation: $r^2 + 4 = 0$.
Roots: $r = \pm 2i$.
General real-valued solution: $x = c_1\cos(2t) + c_2\sin(2t)$.

Example 5.7. A fifth-order constant coefficient linear homogeneous DE has roots -2, $1 \pm 7i$, $\pm 3i$. What is the general solution?

Solution: $x = c_1 e^{-2t} + c_2 e^t\cos(7t) + c_3 e^t\sin(7t) + c_4\cos(3t) + c_5\sin(3t)$.

5.7.1 Justification of the model solution

In Example 5.5, the model solution Steps 1, 2 and 4 are the same as in previous examples with real roots. We need to explain the reasoning behind finding the two basic solutions in step 3:

Amazingly, superposition makes this easy to do. We start with a theorem that tells us how to get real-valued solutions from complex-valued ones.

Theorem. If $z(t)$ is a complex-valued solution to a homogeneous linear DE with real coefficients. Then both the real and imaginary parts of z are also solutions.

Proof. The proof is similar to the proofs of all of our other statements about superposition. Consider the linear homogeneous equation

$$mx'' + bx' + kx = 0 \tag{H}$$

and suppose that $z(t) = x(t) + iy(t)$ is a solution, where $x(t)$ and $y(t)$ are respectively the real and imaginary parts of $z(t)$. We have to show that x and y are also solutions of (H).

By assumption $0 = z'' + bz' + kz$. Replacing z by $x + iy$ we get

$$\begin{aligned} 0 + 0\,i &= m(x + iy)'' + b(x + iy)' + k(x + iy) \\ &= (mx'' + bx' + kx) + i(my'' + by'' + ky). \end{aligned}$$

Since both the real and imaginary parts are 0 we have.

$$mx'' + bx' + kx = 0 \qquad \text{and} \qquad my'' + by' + ky = 0.$$

This says exactly that x and y are solutions to (H).

Now, let's apply this to the situation in Example 5.5.

We saw that there were two characteristic roots $-1 \pm i\sqrt{3}$. These roots give two exponential solutions. Of course, since the roots are complex they are complex exponentials:

$$\begin{aligned} z_1 &= e^{(-1+i\sqrt{3})t} = e^{-t}e^{i\sqrt{3}t} = e^{-t}(\cos(\sqrt{3}t) + i\sin\sqrt{(3t)}) \\ z_2 &= e^{(-1-i\sqrt{3})t} = e^{-t}e^{-i\sqrt{3}t} = e^{-t}(\cos(\sqrt{3}t) - i\sin\sqrt{(3t)}) \end{aligned}$$

Now the theorem above says that both the real and imaginary parts of z_1 and z_2 are also solutions. So we have (nominally) four solutions which we'll label u_1, u_2, v_1, v_2 to avoid overusing the letter x.

$$\begin{aligned} z_1 = u_1 + iv_1: &\quad u_1(t) = e^{-t}\cos(\sqrt{3}t), \quad v_1(t) = e^{-t}\sin(\sqrt{3}t) \\ z_2 = u_2 + iv_2: &\quad u_2(t) = e^{-t}\cos(\sqrt{3}t), \quad v_2(t) = -e^{-t}\sin(\sqrt{3}t). \end{aligned}$$

We see that u_1 and u_2 are the same and, except for the minus sign, v_1 and v_2. So we have only two truly different solutions, which is exactly the number we need. These are the basic solutions given in step (3) of Example 5.5 (except that we used the names x_1 and x_2 instead of u_1 and v_1).

5.7.2 Another way to see this

Another way to see that x_1 and x_2 are solutions is to use superposition directly on the two complex exponential solutions. Since z_1 and z_2 are both solutions so are all linear combinations of z_1 and z_2. In particular, x_1 and x_2 are linear combinations of z_1 and z_2 as follows:

$$\begin{aligned}
\frac{1}{2}z_1(t) + \frac{1}{2}z_2(t) &= \left(\frac{e^{-t}}{2}\cos(\sqrt{3}t) + i\frac{e^{-t}}{2}\sin(\sqrt{3}t)\right) + \left(\frac{e^{-t}}{2}\cos(\sqrt{3}t) - i\frac{e^{-t}}{2}\sin(\sqrt{3}t)\right) \\
&= e^{-t}\cos(\sqrt{3}t) \\
&= x_1(t). \\
\frac{1}{2i}z_1(t) - \frac{1}{2i}z_2(t) &= \left(\frac{e^{-t}}{2i}\cos(\sqrt{3}t) + i\frac{e^{-t}}{2i}\sin(\sqrt{3}t)\right) - \left(\frac{e^{-t}}{2i}\cos(\sqrt{3}t) - i\frac{e^{-t}}{2i}\sin(\sqrt{3}t)\right) \\
&= e^{-t}\sin(\sqrt{3}t) \\
&= x_2(t).
\end{aligned}$$

This shows that x_1 and x_2 are both solutions to the DE.

5.7.3 Complex exponential solutions

We have seen that when the roots of the characteristic equation are complex, we get complex exponentials as solutions. But, with a small amount of algebra, we can write our solutions as linear combinations of real-valued functions. We do this because physically meaningful solutions should have real values. In ES.1803 we won't have much need for the general complex-valued solution, but we record it here for posterity.

The **general complex-valued solution** to the equation in Example 5.5 is

$$z = \tilde{c}_1 z_1 + \tilde{c}_2 z_2 = \tilde{c}_1 e^{(-1+i\sqrt{3})t} + \tilde{c}_2 e^{(-1-i\sqrt{3})t},$$

where $\tilde{c}_1$ and $\tilde{c}_2$ are complex constants. You should be aware that many engineers work directly with these complex solutions and don't bother rewriting them in terms of sines and cosines.

5.8 Repeated roots

When the characteristic equation has repeated roots it will not do to use the same solution multiple times. This is because, for example, $c_1e^{2t} + c_2e^{2t}$ is not really a two-parameter family of solutions, since it can be rephrased as ce^{2t}. For now we will simply assert how to find the other solutions. After we have developed some more algebraic machinery we will be able to explain where they come from.

Example 5.8. A constant coefficient linear homogeneous DE has roots 3, 3, 5, 5, 5, 2. Give the general solution to the DE. What is the order of the DE?

Solution: General solution:

$$x(t) = c_1e^{3t} + c_2te^{3t} + c_3e^{5t} + c_4te^{5t} + c_5t^2e^{5t} + c_6e^{2t}.$$

There are 6 roots so the DE has order 6.

In words: every time a root is repeated we get another solution by adding a factor of t to the previous one.

Example 5.9. A constant coefficient linear homogeneous DE has roots $1 \pm 2i$, $1 \pm 2i$, -3. Give the general real-valued solution to the DE. What is the order of the DE?

Solution: The general real-valued solution is

$$x = c_1 e^t \cos(2t) + c_2 e^t \sin(2t) + c_3 t e^t \cos(2t) + c_4 t e^t \sin(2t) + c_5 e^{-3t}.$$

There are 5 roots so the DE has order 5.

5.9 Existence and uniqueness for constant coefficient linear DEs

So far we have rather casually claimed to have found the *general solution* to DEs. Our techniques have guaranteed that these are solutions, but we need a theorem to guarantee that these are all the solutions. There is such a theorem and it is called the existence and uniqueness theorem.

Theorem: Existence and uniqueness. The initial value problem consisting of the DE

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

with initial conditions

$$y(t_0) = b_0, \quad y'(t_0) = b_1, \quad \dots, \quad y^{(n-1)}(t_0) = b_{n-1}$$

has a unique solution.

The proof is beyond the scope of this course. The outline of the proof for a general existence and uniqueness theorem is posted with the class notes.

Here is a short explanation for why this theorem guarantees that what we've called the general solution does indeed include every possible solution: The theorem says that there is exactly one solution for each set of initial conditions. Therefore, all we have to show is that our general solution includes a solution matching every possible set of initial conditions.

Matching a set of n initial conditions means solving for the n coefficients c_1, ..., c_n. That is, it means solving a linear system of n algebraic equations in n unknowns. Once we've done more linear algebra we'll be able to show this without difficulty. Right now we'll just look at a representative example.

Example 5.10. Suppose a linear second-order constant coefficients homogeneous DE has characteristic roots 2 and 3. Show that the resulting general solution can match every possible set of initial conditions.

Solution: Our general solution is the two-parameter family

$$x(t) = c_1 e^{2t} + c_2 e^{3t}.$$

Our initial conditions have the form $x(t_0) = b_0$ and $x'(t_0) = b_1$. To match these conditions we have to solve for c_1 and c_2. That is, we have to solve

$$\begin{aligned} x(t_0) &= c_1 e^{2t_0} + c_2 e^{3t} = b_0 \\ x(t_0) &= 2c_1 e^{2t_0} + 3c_2 e^{3t} = b_1 \end{aligned}$$

Writing these equations in matrix form we have

$$\begin{bmatrix} e^{2t_0} & e^{3t_0} \\ 2e^{2t_0} & 3e^{3t_0} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}.$$

The coefficient matrix has determinant

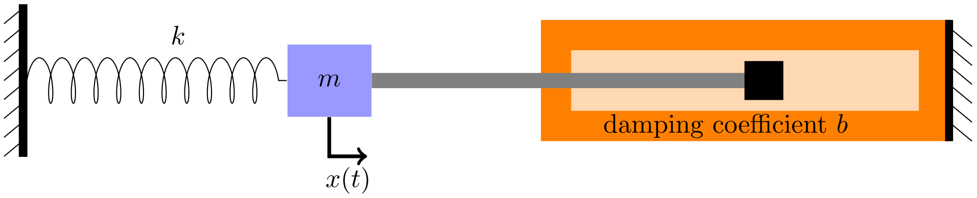
$$\begin{vmatrix} e^{2t_0} & e^{3t_0} \\ 2e^{2t_0} & 3e^{3t_0} \end{vmatrix} = e^{5t_0} \neq 0.$$

Since the determinant is not 0 we know there is a solution giving c_1 and c_2. In fact, we know the solution must be unique.

5.10 Damped harmonic oscillators: the spring-mass-damper

We will use these repeatedly. Please master them.

In ES.1803 one of our main physical examples will be the spring-mass-damper. This is one type of damped harmonic oscillator. (We will encounter others, e.g., an LRC circuit.) In this system we have a mass m attached to a spring with spring constant k. The mass is also attached to a damper that is being dragged through a viscous fluid. The fluid exerts a force on the damper that is proportional to the speed and resists the motion. Let's call the constant in this case the **damping coefficient** b.



Spring-mass-damper with no outside force

For this topic we will assume there is no outside force on the system. So, if $x(t)$ is the displacement of the mass from equilibrium, then Newton's laws tell us

$$mx'' = -kx - bx'.$$

Writing this in our usual fashion, with all the x on the left, we see our standard homogeneous second-order linear constant coefficient DE:

$$mx'' + bx' + kx = 0.$$

A standard notation will be to write $\omega_0 = \sqrt{k/m}$. We'll call ω_0 the **natural frequency** of the system. This term will be explained below.

Simple harmonic oscillator (the undamped spring-mass-dashpot system).

We start with the case of no damping, i.e., $b = 0$. Our equation is then

$$mx'' + kx = 0 \qquad \text{or} \qquad x'' + \omega_0^2 x = 0,$$

where $\omega_0 = \sqrt{k/m}$ = the natural frequency of the oscillator.

Using the characteristic equation method we find:

Characteristic equation: $r^2 + \omega_0^2 = 0$.
Roots: $r = \pm\sqrt{-\omega_0^2} = \pm i\,\omega_0$.
Two solutions: $x_1(t) = \cos(\omega_0 t)$, $x_2(t) = \sin(\omega_0 t)$.
General real-valued solution:

$$x(t) = c_1 x_1(t) + c_2 x_2(t) = c_1 \cos(\omega_0 t) + c_2 \sin(\omega_0 t).$$

We now see why ω_0 is called the natural frequency: it is the angular frequency of the oscillation when the system is undamped and unforced. We will see that damping changes the frequency of oscillation.

Solving the spring-mass-dashpot system: the damped case

Characteristic equation: $mr^2 + br + k = 0$. (Comes from the trial solution $x = e^{rt}$.)

Roots: $r = \dfrac{-b \pm \sqrt{b^2 - 4mk}}{2m}$.

Looking at the formula for the roots we see that there are three cases based on what is under the square root sign. We add a fourth case for when $b = 0$

(i) $b = 0$ **(undamped)**

(ii) $b^2 - 4mk > 0$ (**overdamped;** b large)

(iii) $b^2 - 4mk < 0$ (**underdamped;** b small)

(iv) $b^2 - 4mk = 0$ (**critically damped;** b just right)

We will go through these cases one at a time.

Case (i) (Undamped)

We did this case earlier, The characteristic roots are $\pm\omega_0\, i$. The general real-valued solution is

$$x(t) = c_1 \cos(\omega_0 t) + c_2 \sin(\omega_0 t).$$

The longterm behavior is periodic (sinusoidal) motion.

Case (ii) (Overdamped: real characteristic roots)

To simplify writing we'll name the expression with the square root. Let $B = \sqrt{|b^2 - 4mk|}$, so the roots are

$$r_1 = \frac{-b + B}{2m} \qquad r_2 = \frac{-b - B}{2m}.$$

First we show that **the roots are real and negative**. This follows because B is the square root of something less than b^2. So, in both $-b + B$ and $-b - B$, the B term is not big enough to change the sign of the $-b$ term. Therefore, r_1 and r_2 must both be negative.

The general real-valued solution to the overdamped system is

$$x(t) = c_1 e^{(-b+B)t/(2m)} + c_2 e^{(-b-B)t/(2m)} = c_1 e^{r_1 t} + c_2 e^{r_2 t}.$$

The negative exponents imply that in the longterm as t gets large $x(t)$ goes to 0.

The following claim gives an important feature of overdamped systems.

Claim. If an overdamped system starts from rest at a position away from the equilibrium, then it never crosses the equilibrium position.

Since $x = 0$ is the equilibrium position, the claim says that if $x(0) \neq 0$ and $x'(0) = 0$ then the graph of $x(t)$ does not cross the t-axis for $t > 0$.

Proof. The proof involves some picky algebra: We know that the roots r_1 and r_2 are both negative. We also have $x(t) = c_1 e^{r_1 t} + c_2 e^{r_2 t}$ with initial conditions

$$x(0) = c_1 + c_2 \neq 0, \qquad x'(0) = r_1 c_1 + r_2 c_2 = 0$$

Now we need to show that $x(t) = 0$ never happens for $t > 0$. Let's just do the case $r_1 = -5$ and $r_2 = -2$. The presentation will be simpler and anyone who cares to can redo it for any r_1 and r_2. Using these values of the roots, we have

$$x(t) = c_1 e^{-5t} + c_2 e^{-2t}, \qquad x(0) = c_1 + c_2 \neq 0 \qquad x'(0) = -5c_1 - 2c_2 = 0.$$

The condition $c_1 + c_2 \neq 0$ guarantees that c_1 and c_2 are not both 0. So the other initial condition gives $-c_2/c_1 = 5/2$. Next we'll solve for the times t when $x(t) = 0$.

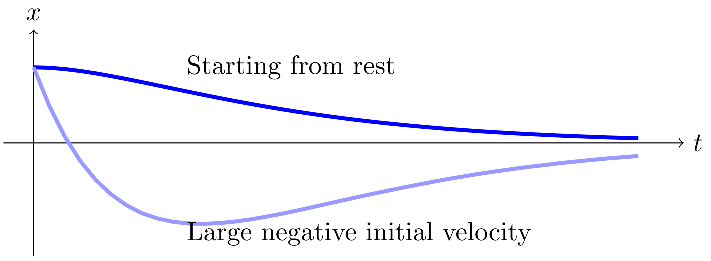
$$x(t) = 0 = c_1 e^{-5t} + c_2 e^{-2t} \quad \text{therefore} \quad -c_2/c_1 = e^{-3t}$$

Combining $-c_2/c_1 = 5/2$ and $-c_2/c_1 = e^{-3t}$, we have $e^{-3t} = 5/2$. Taking the log of both sides we have

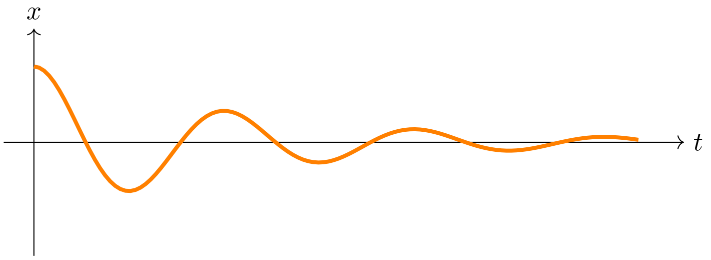
$$-3t = \ln(5/2) > 0, \text{ so, } t < 0.$$

We see that $x(t) = 0$ for exactly one value of t and that value is before $t = 0$. This is exactly what we needed to show!

The proof also showed us that an unforced overdamped harmonic oscillator crosses the equilibrium position at most once.



An overdamped oscillator crosses equilibrium at most once.



An underdamped oscillator crosses equilibrium infinitely many times.

Case (iii) (Underdamping: complex characteristic roots)

Again to simplify the writing we'll name the expression with the square root. Let $B = \sqrt{|b^2 - 4mk|}$. So the characteristic roots are $\dfrac{-b \pm iB}{2m}$. Just as in Example 5.5, the general real-valued solution is

$$x(t) = e^{-bt/(2m)} \left(c_1 \cos\left(\frac{Bt}{2m}\right) + c_2 \sin\left(\frac{Bt}{2m}\right) \right).$$

Longterm behavior: The negative exponent causes $x(t)$ to go to 0 as t goes to ∞. The sine and cosine causes it to **oscillate** back and forth across the equilibrium.

Case (iv) (Critical damping: repeated real characteristic roots)

In this case the expression under the square root is 0, so we have repeated negative characteristic roots $r = -b/(2m)$, $-b/(2m)$. Thus the general solution to the DE is

$$x(t) = c_1 e^{-bt/(2m)} + c_2 t e^{-bt/(2m)}.$$

Qualitatively the picture looks like the overdamped case. Just as in the overdamped case a critically damped oscillator crosses equilibrium at most once.

5.11 Decay rates

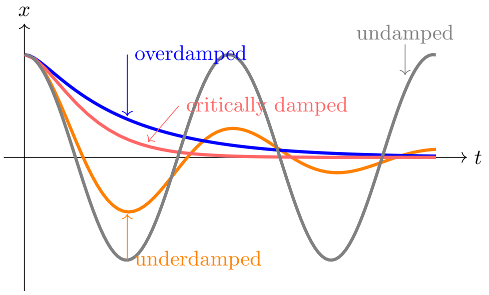
Whether its overdamped, underdamped or critically damped a damped harmonic oscillator goes to 0 as t goes to infinity. We say that $x(t)$ decays to 0. How fast it goes to 0 is its decay rate.

Example 5.11. The rate controlling term. The decay rate of $x(t) = c_1 e^{-3t} + c_2 e^{-5t}$ is the same as that of e^{-3t}. At first glance this might seem surprising because e^{-5t} decays faster than e^{-3t}. But that is exactly the point: the rate of decay is the same as that of the *slowest* term. We might call it the rate controlling term. In this case that is e^{-3t}.

It turns out that critical damping is precisely the level of damping that gives the greatest decay rate. The precise statement is as follows.

Critical damping is optimal. For a fixed mass and spring, i.e., m and k, critical damping is the choice of damping that causes the oscillator to have the greatest decay rate without oscillating.

We will not go through the arithmetic to show this. Anyone interested can ask me about it.



For a fixed m and k critical damping decays the fastest to equilibrium.

5.12 Pole diagrams

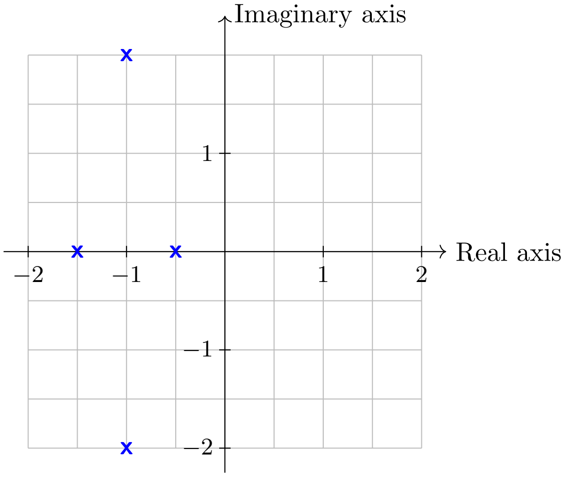
Pole diagrams are a nice way to visualize the characteristic roots of a constant coefficient system $P(D)x = 0$. For these systems the term pole is a synonym for characteristic root. (In general, pole is a mathematical term with a broader meaning.)

The pole diagram is drawn in the complex plane. You construct it by drawing an × at each pole (characteristic root). It is easy to read off information about the system from the diagram.

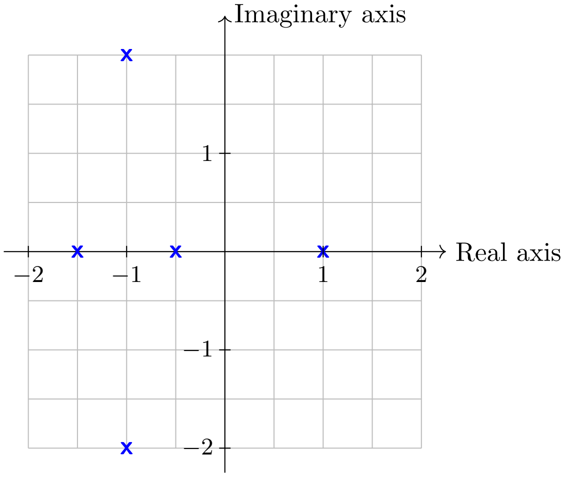
- By counting the poles you can determine the order of the system.
- If all the poles are in the left half-plane then the exponents in the homogeneous solutions all have negative real part. That is, the general homogeneous solution decays to 0, i.e., the system always returns to equilibrium. (We call such a system stable.)
- If there are complex poles then the system is oscillatory.
- For a stable system the exponential rate that the unforced (homogeneous) system returns to equilibrium is determined by the real part of the right-most pole.

Example 5.12. The pole diagram on the left shows 4 poles, all in the left-half plane. Therefore, the system is fourth-order and stable. Since there are complex roots the system is oscillatory. The right-most pole has real part $-1/2$, so the general homogeneous solution decays to 0 like $e^{-t/2}$.

The pole diagram on the right has a pole in the right-half plane at $s = 1$. So the general homogeneous solution grows exponentially, i.e., the system is unstable.



Fourth-order, stable, oscillatory



Fifth-order, unstable, oscillatory

A nice applet showing pole diagrams for second-order systems is the Damped Vibrations applet at `https://mathlets.org/mathlets/damped-vibrations/`. Set $k = .7$, $m = 1$ and let b vary.

6 Operators, inhomogeneous DEs, ERF and SRF

6.1 Goals

1. Be able to define linear differential operators.
2. Be able to define polynomial differential operators and use them to express linear constant coefficient differential equations.
3. Be able to use the Exponential Response Formula to find particular solutions to polynomial differential equations with exponential or sinusoidal input.
4. Be able to derive the Sinusoidal Response Formula.
5. Be able to use the Sinusoidal Response Formula to solve polynomial differential equations with sinusoidal input.
6. Be able to build models of damped harmonic oscillators with input.

6.2 Linear Differential Equations

Linear nth-order differential equations have the form

$$p_0(t)y^{(n)} + p_1(t)y^{(n-1)} + \cdots + p_n(t)y = 0 \qquad \text{(H)}$$

$$p_0(t)y^{(n)} + p_1(t)y^{(n-1)} + \cdots + p_n(t)y = f(t) \qquad \text{(I)}$$

As usual, we call (H) homogeneous and (I) inhomogeneous.

Also as usual, if the coefficients are all constant then we have a **constant coefficient** linear differential equation.

$$a_0y^{(n)} + a_1y^{(n-1)} + \cdots + a_ny = 0 \qquad \text{(H)}$$

$$a_0y^{(n)} + a_1y^{(n-1)} + \cdots + a_ny = f(t) \qquad \text{(I)}$$

In Topic 5 we learned about the characteristic equation

$$a_0r^n + a_1r^{n-1} + \cdots + a_n = 0$$

It will be useful to give a name to the polynomial on the left side of this equation.

$$P(r) = a_0r^n + a_1r^{n-1} + \cdot + a_n.$$

We will call it the **characteristic polynomial**. That is, the characteristic equation can be written $P(r) = 0$.

6.3 Operators

A function is a rule that takes a number as input and returns another number as output.

Example 6.1. (Examples of functions.)
1. $f(t) = t^2$. If $t = 2$ is the input then $f(2) = 4$ is the output.

2. The identity function is $f(t) = t$.

3. The zero function is $f(t) = 0$.

An operator is similar to a function except that it takes as input a function and returns another function as output. We will often use upper case letters like T or L to denote operators. If x is a function when T acts on it we will write

$$T(x) \qquad \text{or} \qquad Tx.$$

We will read this as "T of x" or "T applied to x" or "T acting on x." A few examples will make this clear.

Example 6.2. The **differentiation operator** is $D = \dfrac{d}{dt}$. This takes any function as input and returns its derivative as output. For example,

(i) If $x(t) = t^3$ then $D(x) = 3t^2$. We also write $Dx = 3t^2$.

(ii) If $y(t) = e^{4t}$ then $Dy = 4e^{4t}$.

(iii) $D(t^3 + 2t^2 + 5t + 7) = 3t^2 + 4t + 5$.

(iv) In general, $Dx = x'$.

Example 6.3. The second derivative operator is $D^2 = \dfrac{d^2}{dt^2}$. For example:

(i) $D^2(e^{4t}) = 4^2 e^{4t}$.

In this example we used D^2 to mean first apply D to the function and then apply it again. Writing this out in more detail we get

$$D^2(e^{4t}) = D(D(e^{4t})) = D(4e^{4t}) = 4^2 e^{4t}.$$

(ii) In general, $D^2 x = x''$. Likewise, $D^3 = x'''$.

For obvious reasons we call D, D^2, D^3, ... **differential operators.**

Example 6.4. The **identity operator** I takes any function as input and returns the same function as output. For example:

(i) $I(x) = x$.

(ii) $I(t^2 + 3t + 2) = t^2 + 3t + 2$.

Example 6.5. We can combine these operators. For example we can let

$$T = D^2 + 8D + 7I.$$

To understand what this operator does we have to apply it to a function and see what happens. If we apply T to x we get

$$Tx = (D^2 + 8D + 7I)x = x'' + 8x' + 7x.$$

Example 6.6. The **zero operator** takes any function as input and returns the zero function as output. There is no standard notation for this function, let's call it Z. For example:

(i) $Z(x) = 0$.

(ii) $Z(t^2 + 3t + 2) = 0$.

6.4 Polynomial differential operators

Consider the polynomial $P(r) = r^2+8r+7$. If we replace r by D we have $P(D) = D^2+8D+7$. We will call $P(D)$ a **polynomial differential operator.** We can use it to simplify writing down DEs and to help with algebraic manipulations.

Example 6.7. Consider the constant coefficient differential equation

$$x'' + 8x' + 7x = 0.$$

This has characteristic polynomial $P(r) = r^2+8r+7$. We can rewrite the DE in polynomial notation as

$$(D^2 + 8D + 7I)x = 0 \quad \text{or, even more simply,} \quad P(D)x = 0.$$

One great thing about polynomial operators is how simply we can express constant coefficient differential equations using them. We can rewrite (H) and (I) above as

$$P(D) = 0 \tag{H}$$

$$P(D) = f(t), \tag{I}$$

where $P(D) = D^n + a_1 D^{n-1} + a_2 D^{n-2} + \cdots + a_n I$.

6.5 Linearity/superposition for polynomial differential operators

The superposition principle was awkward to state and prove because it was phrased in terms of equations. Linearity is equivalent to superposition, but easier to discuss because we phrase it in terms of operators.

Important definition. An operator T is called a **linear operator** if for any functions x_1, x_2 and any constants c_1, c_2 we have

$$T(c_1x_1 + c_2x_2) = c_1Tx_1 + c_2Tx_2. \tag{11}$$

Claim. Show that the differential operator D is linear.

Proof. This is easy to check directly from the definition of linearity:

$$D(c_1x_1 + c_2x_2) = (c_1x_1 + c_2x_2)' = c_1x_1' + c_2x_2' = c_1Dx_1 + c_2Dx_2$$

Looking at the first and last terms in this string of equalities we see that Equation 11 holds for the operator D.

Similarly we can show that the operators D^2, D^3 are linear. Likewise, for any polynomial P, the operator $P(D)$ is linear.

Example 6.8. Show directly from the definition that $P(D) = D^2 + 8D + 7I$ is linear.

Solution: We use the same argument as in the proof of the claim just above:

$$\begin{aligned} P(D)(c_1x_1 + c_2x_2) &= (c_1x_1 + c_2x_2)'' + 8(c_1x_1 + c_2x_2)' + 7(c_1x_1 + c_2x_2) \\ &= c_1(x_1'' + 8x_1' + 7x_1) + c_2(x_2'' + 8x_2' + 7x_2) \\ &= c_1P(D)x_1 + c_2P(D)x_2 \end{aligned}$$

I hope the examples have convinced you that the linearity of an operator is **easy to verify**. You might also have noticed how similar the arguments felt to those showing the superposition principle. For completeness we state and show that the two are equivalent.

Equivalence of linearity and the superposition principle. Suppose T is an operator. Then T is linear if and only if the equation $Tx = q(t)$ satisfies the superposition principle.

Proof. This is really just a matter of unwinding the definitions. Suppose $Tx_1 = q_1$ and $Tx_2 = q_2$. Suppose the superposition principle holds, then

$$T(c_1x_1 + c_2x_2) = c_1q_1 + c_2q_2 = c_1Tx_1 + c_2Tx_2.$$

This shows that T is linear. Likewise, if T is linear, then

$$T(c_1x_1 + c_2x_2) = c_1Tx_1 + c_2Tx_2 = c_1q_1 + c_2q_2,$$

which shows that the superposition principle holds.

6.6 The algebra of $P(D)$ applied to exponentials

For this section $P(D)$ will be a polynomial differential operator and a will be a constant. Here are two easy and useful rules concerning $P(D)$ and e^{ax}. We will use them immediately to show why we have factors of t in the solutions to DEs with repeated roots.

6.6.1 Substitution rule

Substitution rule. $P(D)e^{at} = P(a)e^{at}$. This is called the substitution rule because we just substitute a for D.

'Proof' by example. We show the rule holds for $P(r) = r^2 + 8r + 7$:

$$P(D)e^{at} = (e^{at})'' + 8(e^{at})' + 7e^{at} = (a^2 + 8a + 7)e^{at} = P(a)e^{at}.$$

6.6.2 Exponential shift rule

We will call $P(D + aI)$ a shift of $P(D)$ by a. For example, if $P(D) = D^2 + 6D + 9I$ then

$$P(D - 3I) = (D - 3I)^2 + 6(D - 3I) + 9I = D^2 - 6D + 9I + 6D - 18I + 9I = D^2.$$

Exponential shift rule for D. For any function $u(t)$,

$$D(e^{at}u(t)) = e^{at}\,(D + aI)u(t).$$

Proof. The derivation of this is just the product rule for differentiation:

$$D(e^{at}u(t)) = ae^{at}u(t) + e^{at}u'(t) = e^{at}(au(t) + u'(t)) = e^{at}(D + aI)u(t).$$

Exponential shift rule for D^2. For any function $u(t)$,

$$D^2(e^{at}u(t)) = e^{at}\,(D + aI)^2u(t).$$

A similar statememt holds for D^3, D^4, ...

Proof. To derive this for D^2 we just use the rule for D twice. Higher powers are similar.

Now it is clear (by linearity!) that the rule applies to any $P(D)$:

Exponential shift rule for $P(D)$. For any function $u(t)$ and polynmial operator $P(D)$,

$$P(D)(e^{at}u(t)) = e^{at}\, P(D + aI)u(t).$$

6.6.3 Repeated roots

We are now in a positition to explain the rule for solutions with repeated roots. Recall:

Rule for repeated roots. If the characteristic equation $P(r)$ has a repeated root r_1 then both $x_1(t) = e^{r_1 t}$ and $x_2(t) = te^{r_1 t}$ are solutions to the homogeneous DE $P(D)x = 0$.

'Proof' by example. Use the exponential shift rule to show the the equation $x'' - 6x' + 9 = 0$ has general solution $x(t) = c_1 e^{3t} + c_2 te^{3t}$.

Solution: First we rewrite this equation in terms of $P(D)$. The characteristic polynomial is

$$P(r) = r^2 - 6r + 9 = (r - 3)^2.$$

So, $P(D) = (D - 3)^2$ and the differential equation is $P(D)x = 0$.

We know $P(r)$ has repeated roots $r = 3, 3$. So, $x(t) = c_1 e^{3t}$ is a solution. Let's vary the parameters to look for other solutions, i.e., let's try $x(t) = e^{3t}u(t)$. We substitute this into the equation and apply the shift rule:

$$\begin{aligned} P(D)x = 0& \\ &= P(D)(e^{3t}u) \\ &= e^{3t}P(D + 3I)u \\ &= e^{3t}(D + 3I - 3I)^2 u \\ &= e^{3t}D^2 u. \end{aligned}$$

Thus we have the equation $D^2 u = 0$, i.e., $u''(t) = 0$. This is an 18.01 problem and the solution is $u(t) = c_1 + c_2 t$. Putting this back into $x(t)$ we have found

$$x(t) = e^{3t}u(t) = e^{3t}(c_1 + c_2 t),$$

which is exactly what the rule for repeated roots rule said we would find.

6.6.4 Complexification example

Example 6.9. Use complexification to compute $D^3(e^x \sin(x))$.

Solution: We know that $e^x \sin(x) = \mathrm{Im}(e^x e^{ix})$. So, $D^3(e^x \sin(x)) = \mathrm{Im}\left(D^3(e^{x+ix})\right)$. Computing this we have

$$\begin{aligned} \left(D^3(e^{x+ix})\right) &= (1 + i)^3 e^{x+ix} \\ &= (\sqrt{2}e^{i\pi/4})^3 e^x e^{ix} \\ &= 2^{3/2} e^{i3\pi/4} e^x e^{ix} \\ &= 2^{3/2} e^x e^{i(x+3\pi/4)} \end{aligned}$$

Taking the imaginary part we have

$$D^3(e^x\sin(x)) = \text{Im}\left(D^3(e^{x+ix})\right) = \boxed{2^{3/2}e^x\sin(x+3\pi/4)}.$$

6.7 Exponential Response Formula

This is one of our key formulas. We will use throughout the rest of ES.1803.

Exponential Response Formula (ERF). Let $P(D)$ be a polynomial differential operator. The inhomogeneous, constant coefficient, linear DE $P(D)y = e^{at}$ has a particular solution

$$y_p(t) = \begin{cases} e^{at}/P(a) & \text{provided } P(a) \neq 0 \\ te^{at}/P'(a) & \text{if } P(a) = 0 \text{ and } P'(a) \neq 0 \\ t^2e^{at}/P''(a) & \text{if } P(a) = P'(a) = 0 \text{ and } P''(a) \neq 0 \\ \dots & \dots \end{cases}$$

Simple proof: The substitution rule says

$$P(D)e^{at} = P(a)e^{at}. \qquad (12)$$

If $P(a) \neq 0$, then dividing 12 by $P(a)$ proves the theorem in this case.

If $P(a) = 0$, then we differentiate 12 *with respect to a*. This gives

$$P(D)(te^{at}) = P'(a)e^{at} + P(a)te^{at}.$$

Since $P(a) = 0$, the second term on the right is 0 and we have $P(D)(te^{at}) = P'(a)e^{at}$. Dividing by $P'(a)$ proves the theorem in the case $P(a) = 0$ and $P'(a) \neq 0$.

We can continue in this manner for $P(a) = P'(a) = 0$ etc.

Notes:
1. We will call the cases where $P(a) = 0$ the **Extended Exponential Response Formula.**

2. You will need to know how to use the Extended ERF. You will not be asked to know the proof –although doing so is certainly good for you.

Example 6.10. Let $P(D) = D^2 + 4D + 3I$.

(a) Find a solution to $P(D)x = e^{3t}$.

(b) Find a solution to $P(D)x = e^{-3t}$.

Note: The question only asks for one solution, not all of them.

Solution: (a) The equation has exponential input, so we use the exponential response formula:

Compute, $P(3) = 24$, so the ERF gives $x_p(t) = \dfrac{e^{3t}}{P(3)} = \dfrac{e^{3t}}{24}$.

(b) We try the ERF: Since $P(-3) = 0$, we need the extended ERF.

$P(r) = r^2 + 4r + 3$, so $P'(r) = 2r + 4$ and $P'(-3) = -2$. Thus, $x_p(t) = \dfrac{t\,e^{-3t}}{P'(-3)} = -\dfrac{t\,e^{-3t}}{2}$.

In the next example we combine complex replacement and the ERF.

Example 6.11. Let $P(D) = D^2 + 4D + 5I$. Find a solution to $P(D)x = \cos(2t)$.

Solution: (Long form of the solution with explanatory details.)

First we show the details of replacing $\cos(2t)$ by the complex exponential e^{2it}.

Let $y(t)$ satisfy $P(D)y = \sin(2t)$. Then, by linearity, $z(t) = x(t) + iy(t)$ satisfies the DE

$$P(D)z = P(D)x + iP(D)y = \cos(2t) + i\sin(2t) = e^{2ti} \qquad \text{and } x = \text{Re}(z). \qquad (13)$$

Now, in preparation for using the ERF, we compute $P(2i) = 1 + 8i$. Next, we put this in polar form.

$|P(2i)| = |1+8i| = \sqrt{65}$ and $\boxed{\phi = \text{Arg}(P(2i)) = \text{Arg}(1 + 8i) = \tan^{-1}(8) \text{ in quadrant 1.}}$

Thus we have $P(2i) = \sqrt{65}e^{i\phi}$. The ERF gives us complex-valued solution to 13:

$$z_p(t) = \frac{e^{2it}}{P(2i)} = \frac{e^{2it}}{\sqrt{65}e^{i\phi}} = \frac{e^{i(2t-\phi)}}{\sqrt{65}}.$$

All that's left is to take the real part to get a solution to the original DE:

$$x_p(t) = \text{Re}(z_p(t)) = \boxed{\frac{\cos(2t - \phi)}{\sqrt{65}}.}$$

To summarize: $z_p = \dfrac{e^{2i}}{P(2i)}$ and $x_p = \dfrac{1}{|P(2i)|}\cos(2t - \phi)$, where $\phi = \text{Arg}(P(2i))$.

(This example points to the sinusoidal response formula (SRF), which we will look at in the next section.

Example 6.12. Let $P(D) = D^2 + 4D + 5I$. Find a solution to $P(D)x = e^t\cos(2t)$.

Solution: (Short form of solution.) Complexify the DE:

$$P(D)z = e^{-t}e^{2ti} = e^{(-1+2i)t}, \qquad \text{where } x = \text{Re}(z).$$

Side work: $P(-1+2i) = -2+4i = 2\sqrt{5}e^{i\phi}$, where $\boxed{\phi = \text{Arg}(-2 + 4i) = \tan^{-1}(-2), \text{ in Q2}}$.

ERF: $z_p(t) = \dfrac{e^{(-1+2i)t}}{P(-1 + 2i)} = \dfrac{e^{(-1+2i)t}}{-2 + 4i} = \dfrac{e^{-t}e^{2it}}{2\sqrt{5}e^{i\phi}} = \dfrac{e^{-t}e^{i(2t-\phi)}}{2\sqrt{5}}$.

Therefore, $x_p = \text{Re}(z_p) = \boxed{\dfrac{e^{-t}}{2\sqrt{5}}\cos(2t - \phi).}$

Example 6.13. With the same $P(D)$ as in the previous example, find a solution to $P(D)x = e^{-2t}\cos(t)$

Solution: Complexify: $P(D)z = e^{-2t}e^{ti} = e^{(-2+i)t}$ where $x = \text{Re}(z)$.

Side work: $P(-2+i) = 0$, so we'll need $P'(-2+i)$:

$P'(r) = 2r + 4$, So, $P'(-2+i) = 2i = 2e^{i\pi/2}$.

Extended ERF: $z_p(t) = \dfrac{te^{(-2+i)t}}{P'(-2+i)} = \dfrac{te^{(-2+i)t}}{2e^{i\pi/2}} = \dfrac{te^{-2t}e^{i(t-\pi/2)}}{2}$.

Real part: $x_p(t) = \text{Re}(z(t)) = \boxed{\dfrac{te^{-2t}}{2}\cos(t-\pi/2).}$

You want to get good at this, we will do it a lot.

6.8 The Sinusoidal Response Formula

In the examples above we saw a pattern when the input was sinusoidal. We use it so often that we will codify the result as the Sinusoidal Response Formula.

Sinusoidal Response Formula (SRF). Consider the polynomial differential equation

$$P(D)x = \cos(\omega t)$$

If $P(i\omega) \neq 0$ then the DE has a particular solution

$$x_p(t) = \frac{1}{|P(i\omega)|}\cos(\omega t - \phi(\omega)), \text{ where } \phi(\omega) = \text{Arg}(P(i\omega)).$$

If $P(i\omega) = 0$ we have the **Extended SRF**. For example, if $P(i\omega) = 0$ and $P'(i\omega) \neq 0$ then the DE has a particular solution

$$x_p(t) = \frac{t\cos(\omega t - \phi(\omega))}{|P'(i\omega)|}, \text{ where } \phi(\omega) = \text{Arg}(P'(i\omega)).$$

Proof. To prove the extended SRF we just follow the steps from the examples above.

1. Complexify: $P(D)z = e^{i\omega t}$, where $x = \text{Re}(z)$.
2. Write $P'(i\omega)$ in polar coordinates: $P'(i\omega) = |P'(i\omega)|e^{i\phi(\omega)}$, where $\phi(\omega) = \text{Arg}(P'(i\omega))$.
3. Use the extended ERF: $z_p = \dfrac{te^{i\omega t}}{P'(i\omega)} = \dfrac{te^{i(\omega t - \phi(\omega))}}{|P'(i\omega)|}$.
4. Find the real part of z_p:

$$x_p(t) = \text{Re}(z_p(t) = \text{Re}\left(\frac{te^{i(\omega t-\phi(\omega))}}{|P(i\omega)|}\right) = \frac{t\cos(\omega t - \phi(\omega))}{|P(i\omega)|}.$$

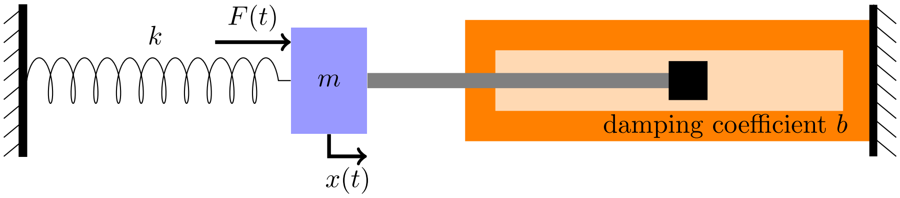
Remember: If in doubt when using the extended SRF, you can always derive it using complexification and the extended ERF.

6.9 Physical models

In this section we will look at three versions of the driven spring-mass-dashpot. These have analogies, which we won't show here, in RLC circuits.

In all three examples, we assume linear damping with damping constant b. That is, if the damper is moving with velocity v through the dashpot, then the force of the dashpot on the damper is $-bv$. This is a reasonable model if the dashpot is filled with a viscous oil.

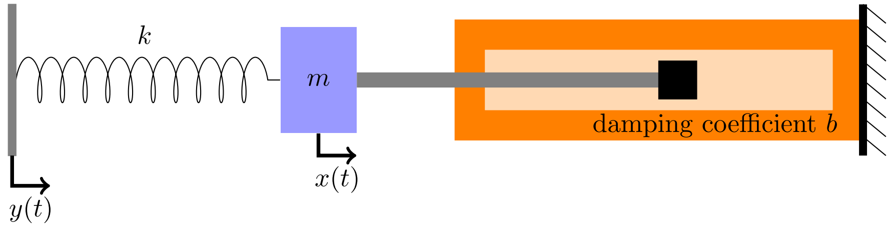
Example 6.14. Driving through the mass. In this version, there is a spring-mass-dashpot which is driven by a variable force applied to the mass as shown. The position of the mass is $x(t)$, with $x = 0$ being the equilibrium position, i.e., the position where the spring is relaxed.



To model this, we consider all the forces on the mass and then use Newton's second law. The spring is stretched by x, so it exerts a restoring force: $-kx$. The velocity of the damper through the dashpot is $\dot{x}$, so it exerts a resisting force: $-b\dot{x}$. Thus Newton's law gives

$$m\ddot{x} = -kx - b\dot{x} + F(t) \quad \Leftrightarrow \quad \boxed{m\ddot{x} + b\dot{x} + kx = F(t)}.$$

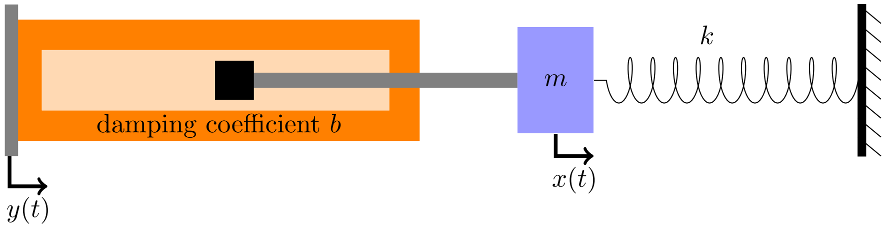
Example 6.15. Driving through the spring. In this version, the spring-mass-dashpot is driven by a mechanism that positions the end of the spring at $y(t)$ as shown. As before, $x(t)$ is position of the mass. We calibrate x and y so that $x = 0$, $y = 0$ is an equilibrium position of the system.



To model this, we must consider all the forces on the mass. At time t, the spring is stretched an amount $x(t) - y(t)$, so the spring force is $-k(x - y)$. Likewise, the velocity of the damper through the dashpot is $\dot{x}$, so the damping force is $-b\dot{x}$. Thus,

$$m\ddot{x} = -k(x - y) - b\dot{x} \quad \Leftrightarrow \quad \boxed{m\ddot{x} + b\dot{x} + kx = ky}.$$

Example 6.16. Driving through the dashpot. In this version, the spring-mass-dashpot is driven by a mechanism that positions the end of the dashpot at $y(t)$ as shown. Again, $x(t)$ is position of the mass and $x = 0$, $y = 0$ is an equilibrium position of the system.



More briefly than the previous examples:
spring force: $-kx$
damping force: $-b(\dot{x}-\dot{y})$.
Model: $m\ddot{x} = -kx - b(\dot{x}-\dot{y}) \ \Leftrightarrow\ \boxed{m\ddot{x} + b\dot{x} + kx = b\dot{y}}$.

7 Solving linear DEs; method of undetermined coefficients

7.1 Goals

1. Be able to solve a linear differential equation by superpositioning a particular solution with the general homogeneous solution.
2. Be able to find a particular solution to a linear constant coefficient differential equation with polynomial input.
3. Be able to work with the operator D, e.g., be able to check if two operators are equal by checking their behavior on test functions.
4. Understand the statement of the existence and uniqueness theorem for second-order linear DEs.

7.2 Linear (not necessarily constant coefficient) DEs

7.2.1 Nice simple operator notation

The general linear differential equation has the form

$$y^{(n)} + p_1(t)y^{(n-1)} + \cdots + p_n(t)y = f(t) \tag{14}$$

We can simplify our notation by defining the differential operator

$$L = D^n + p_1(t)D^{n-1} + \cdots + p_n(t)I.$$

Think: In this case, we used the letter L because it is a linear operator. You should recall what this means from Topic 6.

Remember: to see how an operator behaves we apply it to a function. In this case:

$$Lx = \left[D^n + p_1(t)D^{n-1} + \cdots + p_n(t)I\right]x = x^{(n)} + p_1(t)x^{(n-1)} + \cdots p_n(t)x.$$

So we can rewrite Equation 14 as

$$Ly = f(t) \qquad \text{(pretty simple looking).}$$

7.2.2 General solution to a linear inhomogeneous equation

The superposition principle for a linear differential operator L says the following:

- If y_p is a particular solution to the inhomogeneous equation $Ly = f$
- and y_h is a solution to the homogeneous equation $Ly = 0$
- then $y = y_p + y_h$ is also a solution to $Ly = f$.

The proof of this is a straightforward use of the definition of linearity:

$$Ly = L(y_p + y_h) = Ly_p + Ly_h = f + 0 = f.$$

7.2.3 Strategy for finding the general solution to $Ly = f$

1. Find the general solution to the homogeneous equation $Ly = 0$. Call it y_h.
2. Find any one particular solution to $Ly = f$. Call it y_p.
3. The general solution to $Ly = f$ is $y = y_p + y_h$.

Example 7.1. Let $L = D^2 + 4D + 5$. Solve $Ly = e^{-t}$.

Solution: 1. First we solve the homogeneous equation: $Ly = 0$. Since this is a constant coefficient equation, we can use the method of the characteristic equation.

Characteristic equation: $P(r) = r^2 + 4r + 5 = 0$. This has roots $r = -2 \pm i$.

General real-valued homogeneous solution:

$$\boxed{y_h(t) = c_1 e^{-2t}\cos t + c_2 e^{-2t}\sin t.}$$

2. Find a particular solution using the exponential response formula:

$$\boxed{y_p(t) = \frac{e^{-t}}{P(-1)} = \frac{e^{-t}}{2}.}$$

3. The general real-valued solution to $Ly = e^{-t}$ is a superposition of the particular and the homogeneous solutions:

$$\boxed{y(t) = y_p(t) + y_h(t) = \frac{e^{-t}}{2} + c_1 e^{-2t}\cos t + c_2 e^{-2t}\sin t.}$$

7.3 The method of undetermined coefficients for polynomial input

The method of undetermined coefficients for polynomial input is yet another version of the method of optimism. In this case, we try a polynomial solution and use algebra to find the coefficients.

Example 7.2. Solve $y'' + 5y' + 4y = 2t + 3$.

Solution: We follow these steps:

1. First, to find a particular solution:

(a) We guess a trial solution of the form $y_p(t) = At + B$. Our guess has the same degree as the input.

(b) Substitute the guess into the DE and do the algebra to compute the coefficients. Here is one way to present the calculation:

$$\begin{aligned} y_p &= At + B \\ y_p' &= A \\ y_p'' &= 0 \\ y_p'' + 5y_p' + 4y_p &= 5A + 4(At + B)\ = 4At + (5A + 4B) \end{aligned}$$

Substituting this into the DE we get:

$$4At + (5A + 4B) = 2t + 3.$$

Now we equate the coefficients on both sides to get two equations in two unknowns.

$$\begin{array}{lllll} \text{Coefficients of } t: & 4A & & = 2 \\ \text{Coefficients of } 1: & 5A & +4B & = 3 \end{array}$$

This is called a triangular system of equations. It is easy: $A = 1/2$, $B = 1/8$. So, $\boxed{y_p(t) = \frac{1}{2}t + \frac{1}{8}.}$

2. Next we find solution of homogeneous DE: $y'' + 5y' + 4y = 0$.

Characteristic equation: $r^2 + 5r + 4 = 0$. This has roots $r = -1, -4$.

General homogeneous solution: $\boxed{y_h(t) = c_1e^{-t} + c_2e^{-4t}.}$

3. Finally, we use the superposition principle to write the general solution to our DE:

$$y(t) = y_p(t) + y_h(t) = \frac{1}{2}t + \frac{1}{8} + c_1e^{-t} + c_2e^{-4t}.$$

Example 7.3. Solve $y'' + 5y' + 4y = 2t^2 + 3t$.

Solution: Guess a trial solution of the form $y_p(t) = At^2 + Bt + C$ (same degree as the input). Substitute the guess into the DE (we don't show the algebra):

$$y_p'' + 3y_p' + 4y_p = 4At^2 + (10A + 4B)t + (2A + 5B + 4C) = t^2 + 3t.$$

Equate the coefficients of the polynomials on both sides of the equation:

$$\begin{array}{llllll} \text{Coeff. of } t^2\text{:} & 4A & & & = 1 \\ \text{Coeff. of } t\text{:} & 10A & +4B & & = 3 \\ \text{Coeff. of } 1\text{:} & 5A & +5B & +4C & = 0 \end{array}$$

This triangular system is easy to solve: $A = 1/4$, $B = 1/8$, $C = -9/32$. Therefore, a particular solution is

$$\boxed{y_p(t) = \frac{1}{4}t^2 + \frac{1}{8}t - \frac{15}{32}.}$$

We can use the homogeneous solution from previous example. So the general solution to the DE is $y(t) = y_p(t) + y_h(t)$.

Example 7.4. What can go wrong (and how to fix it). Find a solution to

$$y'' + y' = t + 1.$$

Solution: Since the input is a first-degree polynomial, we try a first-degree solution: $y_p(t) = At + B$. Substituting this into the DE we get

$$A = t + 1.$$

This can't be solved!

The problem is that there is no y term in $y'' + y'$ (or rather, its coefficient is 0). The fix is to bump all degrees up by the order of the lowest derivative, i.e., try the solution $y_p = At^2 + Bt$.

Substitute: $2At + (2A + B) = t + 1$.

Equate coefficients: $2A = 1$; $(2A + B) = 1$.

Solve for A and B: $A = 1/2$, $B = 0$.

Thus, $y_p(t) = \frac{1}{2}t^2$.

Example 7.5. Find a solution to $y''' + 3y'' = t$.

Solution: The input has degree 1 and the lowest order derivative in the DE is 2. So we guess $y_p = At^3 + Bt^2$.

Substitute: $18At + 6A + 6B = t$.

Equate coefficients: $18A = 1$; $6A + 6B = 0$.

Solve for A and B: $A = 1/18$, $B = -1/18$.

Thus, $y_p(t) = \dfrac{t^3}{18} - \dfrac{t^2}{18}$.

Example 7.6. Exponential Shift Rule. Solve $y'' + 5y' + 4y = e^{2t}(t + 3)$.

Solution: In operator form this is $P(D)y = e^{2t}(t + 3)$, where $P(D) = D^2 + 5D + 4$.

First we find a particular solution by looking for one of the form $y = e^{2t}u$. We substitute this into the DE and use the exponential shift rule to pull out the exponential. The left-hand side of the equation is

$$P(D)(e^{2t}u) = e^{2t}P(D + 2I)u = e^{2t}((D + 2I)^2 + 5(D + 2I) + 4I)u = e^{2t}(D^2 + 9D + 18I)u.$$

Equating this with the right-hand side we have

$$e^{2t}(D^2 + 9D + 18I)u = e^{2t}(t + 3) \qquad \text{or} \qquad (D^2 + 9D + 18I)u = t + 3.$$

The method of undetermined coefficients gives (we don't show the algebra):

$$u_p(t) = \frac{1}{18}t + \frac{5}{36}.$$

Thus, $\boxed{y_p(t) = e^{2t}u_p(t) = e^{2t}\left(\frac{1}{18}t + \frac{5}{36}\right).}$

To finish solving we find the homogenous solution (again without showing the algebra). $\boxed{y_h(t) = c_1e^{-t} + c_2e^{-4t}.}$

So the general solution to the DE is $\boxed{y(t) = y_p(t) + y_h(t).}$

7.4 A bit more on the operator D

We remind you that $D = \frac{d}{dt}$, i.e., $Df = f'$.

7.4.1 Algebra with constant coefficients

For polynomial differential operators, we can add and multiply in any order using the usual rules of arithmetic.

Example 7.7. Show that $(D-3I)(D-2I) = (D-2I)(D-3I) = D^2 - 5D + 6I$.

Note. In words this says that the operators $D-3I$ and $D-2I$ commute and that the usual rules of multiplying polynomials apply.

Solution: To show two operators are equal we have to show they give the same result when applied to any function. This is easy if a bit tedious:

$$(D-3I)(D-2I)f = (D-3I)(f'-2f) = f'' - 3f' - 2f' + 6f = f'' - 5f' + 6f$$
$$(D-2I)(D-3I)f = (D-2I)(f'-3f) = f'' - 2f' - 3f' + 6f = f'' - 5f' + 6f$$

Since $(D-3I)(D-2I)$ and $(D-2I)(D-3I)$ give the same result when applied to a test function f, they are the same operator. The right hand side of both of the equations above shows they both equal $D^2 - 5D + 6I$, as stated in the problem.

The next examples show that we must have constant coefficients for this to work.

7.4.2 Algebra with non-constant coefficient operators

Example 7.8. Let M be the 'multiplication by t' operator, i.e., $Mf = tf$. Show that $MD \neq DM$, i.e., show M and D do not commute.

Solution: We need to apply each operator to a *test function* f and see that we get different results.

$$MDf = Mf' = tf'$$
$$DMf = D(tf) = f + tf'$$

We see that the two are not equal, so the operators don't commute.

Notational note. It is common to use a shorthand and write the operator M as t. So we could have written the example as: Show $tD \neq Dt$ as operators. We will also sometimes write the operator M as tI.

Example 7.9. Show that $(D-tI)D \neq D(D-tI)$.

Solution: As is now usual, we show this by applying both operators to a test function y and seeing that we get different results.

$$\begin{aligned}&(D-tI)Dy=(D-tI)y'=y''-ty'=(D^2-tD)y.\\&D(D-tI)y=D(y'-ty)=y''-ty'-y=(D^2-tD-I)y.\end{aligned}$$

7.5 General theory of linear second-order equations

In this section we'll collect up much of what we've already done and add to it the existence and uniqueness theorem.

The general second-order linear DE is

$$A(t)y''+B(t)y'+C(t)y=F(t).$$

The standard form is

$$y''+p(t)y'+q(t)y=f(t). \tag{L}$$

Example 7.10. Here is a linear second-order DE in general and standard form:

$$\begin{aligned}\text{General: }&t^2y''+ty'+y=e^t\\\text{Standard: }&y''+\frac{1}{t}y'+\frac{1}{t^2}\frac{e^t}{t^2}=\frac{e^t}{t^2}\end{aligned}$$

Homogeneous (standard form):

$$y''+p(t)y'+q(t)y=0 \tag{H}$$

7.5.1 Superposition/Linearity

The general principle of superposition says that, for a linear DE, superposition of inputs leads to superposition of outputs, i.e.

> If y_1 solves $y''+p(t)y'+q(t)y=f_1(t)$ and y_2 solves $y''+p(t)y'+q(t)y=f_2(t)$,
> then $c_1y_1+c_2y_2$ solves $y''+p(t)y'+q(t)y=c_1f_1(t)+c_2f_2(t)$.

We have already made repeated use of the following two forms of the principle.

1. Superposition of homogeneous solutions: If y_1 and y_2 are solutions to Equation H then so is $y=c_1y_1+c_2y_2$.

2. Superposition of homogeneous and inhomogeneous solutions: If y_p is a solution to Equation L and y_h is a solution to Equation H then $y=y_p+y_h$ is also a solution to L.

7.5.2 Existence and Uniqueness

The existence and uniqueness theorem is an important technical tool for us. When solving a differential equation it guarantees that we can find a solution and it also tells us when we've found them all.

Theorem. Existence and uniqueness. Consider the initial value problem

$$y'' + p(t)y' + q(t)y = f(t); \qquad y(a) = b_0,\, y'(a) = b_1.$$

If p, q and f are continuous on an interval I containing the point a then there exists a unique solution to this differential equation satisfying the given initial conditions.

Important graphical note: The theorem tells us that the graphs of two different solutions to the DE *can* cross, but they *cannot* touch tangentially.

(i) If they cross transversally, then they have the same position at the time they cross, but different velocities.

(ii) If they did touch tangentially, then they would have the same position and the same velocity at the time they touch. By the existence and uniqueness theorem, there is exactly one solution –not two– with that position and velocity, so this is impossible.

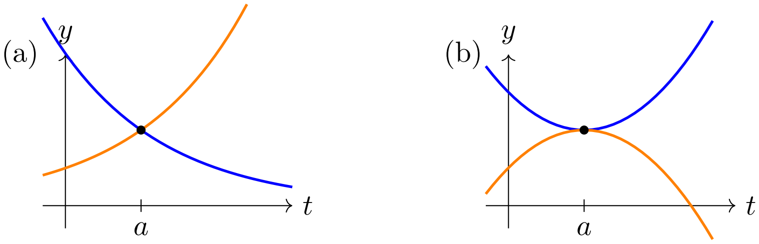


Figure (a) shows transversal curves. These could both be solutions to a second-order DE that satisfies the conditions of the existence and uniqueness theorem.

Figure (b) shows curves that touch tangentially. These cannot both be solutions to such a DE.

8 Applications: stability

8.1 Goals

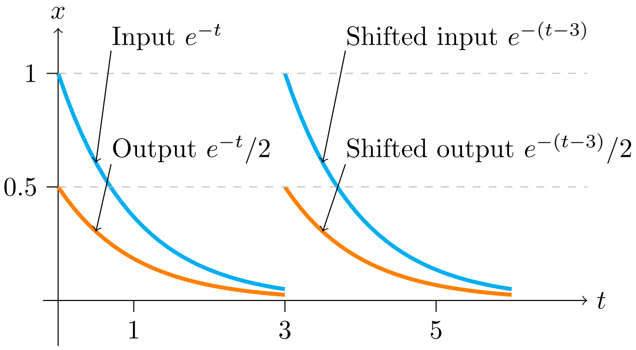
1. Know the meaning of the term 'linear time invariance'.
2. Be able to apply linear time invariance to solve equations with input shifted in time.
3. Know the definitions of mathematical and physical stability
4. Be able to determine if a given 1st, 2nd or 3rd order system is stable.

8.2 Time invariance

Constant coefficient differential equations have the property of time invariance. Physically this means that the system responds the same way to an input no matter when the input is started. Mathematically we write this as follows.

Definition. Time invariance of a constant coefficient system is the property that if $x_p(t)$ satisfies $P(D)x = f(t)$ then $x_p(t - t_o)$ satisfies $P(D)x = f(t - t_0)$.

Example 8.1. We know that $x' + 3x = e^{-t}$ has solution $x_1(t) = e^{-t}/2$. Time invariance says that $x' + 3x = e^{-(t-3)}$ has solution $x_2(t) = x_1(t-3) = e^{-(t-3)}/2$. The figure below illustrates that shifting the input in time simply shifts the output in time.



Physically this has to be the case –an exponential decay system doesn't care what time it gets started.

8.3 Mathematical stability

We introduce the idea of stability with an example that shows how negative exponents imply that initial conditions do not affect the long-term behavior of a system.

Example 8.2. Consider the DE $x'' + 2x' + 3x = \cos(2t)$

(a) Solve the DE with initial conditions $x(0) = 2, \quad x'(0) = 3$. Describe the long-term behavior of the solution.

(b) Describe the long-term behavior of the solution with initial conditions $x(0) = 1$, $x'(0) = 1$.

Solution: (a) First we find the general homogeneous solution.

Homogeneous solution.

The characteristic equation is $r^2 + 2r + 3 = 0$. This has roots: $r = -1 \pm \sqrt{2}\, i$.

So, $x_h(t) = c_1 e^{-t} \cos(\sqrt{2})\, t + c_2 e^{-t} \sin(\sqrt{2}\, t)$

Particular solution.

Next we find a particular solution using the sinusoidal response formula. For this we need to compute $P(2i)$ and put it in polar form..

$$P(2i) = -4 + 4i + 3 = -1 + 4i = \sqrt{17} e^{i\phi}, \quad \text{where } \boxed{\phi = \text{Arg}(P(2i)) = \tan^{-1}(-4) \text{ in Q2}}.$$

Now the SRF gives $x_p(t) = \dfrac{\cos(2t-\phi)}{\sqrt{17}}$.

General solution.

$$x(t) = x_p(t) + x_h(t) = \frac{\cos(2t-\phi)}{\sqrt{17}} + c_1 e^{-t}\cos(\sqrt{2}\,t) + c_2 e^{-t}\sin(\sqrt{2}\,t).$$

Finally, we use the initial conditions to determine the values of c_1, c_2.

$$\begin{aligned} x(0) &= \cos(-\phi)/\sqrt{17} + c_1 = 2 \;\longrightarrow\; c_1 = 35/17 \\ x'(0) &= -2\sin(-\phi)/\sqrt{17} - c_1 + c_2\sqrt{2} = 3 \;\longrightarrow\; c_2 = 39 * \sqrt{2}/17 \end{aligned}$$

So, $x(t) = \dfrac{\cos(2t-\phi)}{\sqrt{17}} + \dfrac{35}{17}e^{-t}\cos(\sqrt{2}\,t) + \dfrac{39\sqrt{2}}{17}e^{-t}\sin(\sqrt{2}\,t)$.

The question also asks what happens to the system in the long-term, i.e., as $t \to \infty$. Looking at the solution above, we see that the terms with e^{-t} go to 0. This means that, in the long-term, we have

$$x(t) \approx x_p(t) = \frac{\cos(2t-\phi)}{\sqrt{17}}, \text{ for large } t.$$

(b) The general solution is the same as in Part (b). Since it has negative exponents, $x_h(t)$ goes to 0 as t goes to infinity. This means that, in the long-term, the solution $x(t)$ behaves exactly like the solution in Part (a), i.e., goes asymptotically to $x_p(t)$.

This is the key point: the values of c_1 and c_2 will change with the initial conditions, but in the long-term, the terms with c_1 and c_2 will go to 0, i.e., the initial conditions don't affect the long-term behavior of the system.

This leads to our definition of stability and several equivalent ways of describing it.

Definition. Mathematical stability means the long-term behavior doesn't depend (significantly) on initial conditions.

Linear Systems. The system $Ly = f$ is stable if the general homogeneous solution $y_h(t) \to 0$ as $t \to \infty$. In this case, y_h is called the transient.

Linear CC Systems. The system $P(D)y = f$ is stable if all the characteristic roots have negative real part.

For linear systems stability is determined by the homogeneous solution. That is,

$$\boxed{\textbf{Stability is about the system not the input.}}$$

Example 8.3. $x' + 2x = f(t)$ is stable because $x_h(t) = ce^{-2t} \to 0$.

Example 8.4. A constant coefficient system with roots $-2 \pm 3i$, -3 is stable.

Example 8.5. A constant coefficient system with roots -2, -3, 4 is unstable.

Example 8.6. $P(D)y = y'' + 8y' + 7y = f(t)$ has characteristic roots -7, -1. These are negative so the system is stable.

Example 8.7. $P(D)y = y'' - 6y' + 25y = f$ has characteristic roots $3 \pm 4i$. The real parts of these roots are positive, so the system is not stable.

8.4 Stability criteria for linear CC systems

1. Stability $\Leftrightarrow$ for any IC $y_h \to 0$ as $t \to \infty$.

2. Stability $\Leftrightarrow$ all characteristic roots have negative real part.

3. Stability $\Leftrightarrow$ all solutions to the homogeneous equation $P(D)y = 0$ go asymptotically to the homogeneous equilibrium solution $y(t) = 0$.

4. For a first-order system $P(D)y = y' + ky = f(t)$:
   Characteristic root $= -k$. Therefore, stability $\Leftrightarrow$ $k > 0$.

5. For a second-order system $P(D)y = my'' + by' + ky = f(t)$:
   Stability $\Leftrightarrow$ m, b, k all have the same sign (easy to prove).

6. For a third-order system $P(D)y = y''' + ay'' + by' + cy = f$:
   Stability $\Leftrightarrow$ $a, b, c > 0$ and $ab > c$ (harder to prove).

   This shows that third-order systems with positive coefficients aren't necessarily stable.

   Example: An unstable system with positive coefficients
   $$(r+5)(r-1-100i)(r-1+100i) = r^3 + 3r^2 + 96r + 505.$$

7. The stability criteria for third-order systems is an example of the Routh-Hurwitz stability criteria, which is described below in the last section of this topic.

   Key point: This criteria is somewhat complicated, but it allows us to determine stability from the coefficients of a system. That is, it does not require finding the roots!

8.5 Physical stability

Definition. Physical stability. An unforced physical system with a single equilibrium is called stable if, for any initial conditions, it always returns to the equilibrium.

Later in the course we will expand on the notion of stability for systems with multiple equilibria. The next example shows how physical and mathematical stability are related.

Example 8.8. Damped-spring-mass system: Physical stability matches mathematical stability. The equilibrium solution is $x(t) = 0$. The unforced system is modeled by $mx' + bx' + kx = 0$. Since the roots have negative real part, $x(t) \to 0$, no matter what the initial conditions.

Note: The previous section on stability criteria show that second-order physical systems, like springs and LRC circuits are always stable. This is not true of 3rd (and higher) order physical systems. An example is given in the in-class notes for this topic which discuss Maxwell's model of steam engines.

8.6 Routh-Hurwitz stability criteria

This section is copied from Section S of the 18.03 Supplementary Notes by Arthur Mattuck. We include it for anyone who is interested. You are not responsible for knowing this in 18.03.

Assume $a_0 > 0$, the constant coefficient, linear system

$$(a_0 D^n + a_1 D^{n-1} + ... + a_{n-1} D + a_n I)x = f(t)$$

is stable if and only if

in the matrix below, all of the n principal minors (i.e., the subdeterminants in the upper left corner having sizes respectively $1, 2, ..., n$*) are greater than 0.*

$$\begin{bmatrix} a_1 & a_0 & 0 & 0 & 0 & 0 & ... & 0 \\ a_3 & a_2 & a_1 & a_0 & 0 & 0 & ... & 0 \\ a_5 & a_4 & a_3 & a_2 & a_1 & a_0 & ... & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{2n-1} & a_{2n-2} & a_{2n-3} & a_{2n-4} & ... & ... & ... & a_n \end{bmatrix}$$

In the matrix, we define $a_k = 0$ *if* $k > n$*. Thus, for example, the last row always has just one non-zero entry,* a_n*.*

The proof of this is some fairly elaborate algebra, which we won't reproduce here.

Example 8.9. Apply the Routh-Hurwitz criteria to the system

$$x''' + ax'' + bx' + cx = f(t).$$

Solution: The matrix for this system is

$$\begin{bmatrix} a & 1 & 0 \\ c & b & a \\ 0 & 0 & c \end{bmatrix}$$

The three principle minors are

$$\left|a\right| = a, \quad \begin{vmatrix} a & 1 \\ c & b \end{vmatrix} = ab - c, \quad \begin{vmatrix} a & 1 & 0 \\ c & b & a \\ 0 & 0 & c \end{vmatrix} = c(ab - c)$$

The Routh-Hurwitz criteria are that all three minors must be positive. That is,

$$a > 0, \quad ab - c > 0, \quad c(ab - c) > 0$$

Since $ab - c > 0$, the condition $c(ab - c) > 0$ implies $c > 0$. Then, since a and c are positive, the condition $ab - c > 0$ implies $b > 0$. Thus we have the criteria stated above:

The system is stable is equivalent to a, b, c are positive and $ab > c$.

9 Applications: frequency response

9.1 Goals

1. Be able to use the engineering terminology of gain, phase lag, resonance.
2. Understand that the gain depends on what we declare to be the input.
3. Be able to find practical or pure resonant frequencies if they exist.

9.2 Review of a forced damped harmonic oscillator

Note. You can also see the text by Edwards and Penney, sections 2.4 and 2.7 for a nice discussion of RLC circuits and practical resonance.

Throughout this topic we will be considering damped harmonic oscillators. There will be important variations, but let's start by reviewing one such system.

Example 9.1. Consider the system

$$my'' + by' + ky = kB\cos(\omega t), \tag{15}$$

where m, b, k, B, and ω are constants. For this system, we will consider $B\cos(\omega t)$ to be the input. Below we will discuss how the input and output are not mathematical notions. In an engineering context we must always say what we mean by the input and the output.

Let's review our method of solution for this equation

1. Find the homogeneous solution.

Characteristic roots $= \dfrac{-b \pm \sqrt{b^2 - 4mk}}{2m}$.

Let $\beta = \dfrac{\sqrt{|b^2 - 4mk|}}{2m}$. (Note the absolute value inside the square root.) There are three cases:

(i) $b^2 - 4mk > 0$ (overdamped): $y_h(t) = c_1 e^{(-b/2m+\beta)t} + c_2 e^{(-b/2m-\beta)t}$.

(ii) $b^2 - 4mk < 0$ (underdamped): $y_h(t) = c_1 e^{-bt/2m}\cos(\beta t) + c_2 e^{-bt/2m}\sin(\beta t)$.

(iii) $b^2 - 4mk = 0$ (critically damped): $y_h(t) = c_1 e^{-bt/2m} + c_2\, te^{-bt/2m}$.

2. Find a particular solution.

We can apply the sinusoidal response formula (SRF) directly:

$$y_p(t) = \frac{kB\cos(\omega t - \phi(\omega))}{|P(i\omega)|},$$

where $P(r)$ is the characteristic polynomial and $\phi(\omega) = \mathrm{Arg}(P(i\omega))$.

Because we will want to make small variations in this formula we will also review the method of complexification that leads to the sinusoidal response formula.

Step 1. Complexify the DE to get:

$$mz'' + bz' + kz = kBe^{i\omega t}, \qquad \text{where } y = \mathrm{Re}(z).$$

Step 2. We will need $P(i\omega)$ in polar form. The characteristic polynomial is $P(r) = mr^2 + br + k$. So,

$$P(i\omega) \;=\; (k - m\omega^2) + ib\omega \;=\; \sqrt{(k - m\omega^2)^2 + b^2\omega^2}\; e^{i\phi(\omega)},$$

where $\boxed{\phi(\omega) = \mathrm{Arg}(P(i\omega)) = \tan^{-1}\left(\dfrac{b\omega}{k - m\omega^2}\right) \text{ in the first or second quadrants.}}$

Think: Why is $\phi(\omega)$ in Q1 or Q2?

Step 3. Use the exponential response formula to give a particular (complex-valued) solution:

$$z_p = \frac{kBe^{i\omega t}}{P(i\omega)} = \frac{kBe^{i(\omega t - \phi(\omega))}}{\sqrt{(k-m\omega^2)^2 + b^2\omega^2}}. \tag{16}$$

Step 4. *Uncomplexify* by taking the real part to find y_p.

$$y_p(t) = \text{Re}(z(t)) = \frac{kB\cos(\omega t - \phi(\omega))}{\sqrt{(k-m\omega^2)^2 + b^2\omega^2}}. \tag{17}$$

Finally we use superposition to give the general real-valued solution:

$$y(t) = y_p(t) + y_h(t).$$

9.2.1 Terminology

Still referring to the system in Example 9.1:

- $y_h(t)$ is called the **transient** because it goes to 0 as t goes to infinity.

  Think: How do we know that $y_h(t)$ decays to 0?

- $y_p(t)$ is called the **periodic** or **sinusoidal solution**.

  Since $y_h(t)$ goes to 0, all solutions go asymptotically to $y_p(t)$.

In thinking about this system, we are going to assume the m, b, and k are fixed. We will imagine that we have a knob that can be used to set ω just before we need to solve the equation. Thus the response of the system will depend on the value of ω.

The following is a list of terms with short definitions. We will discuss them in much more detail below.

- **Input:** When talking about gain and phase lag, we will always take the input to be sinusoidal, i.e., $B\cos(\omega t)$.
- **Input frequency:** The angular frequency of the sinusoidal input, i.e., ω. (In radians/time.)
- **Input amplitude:** The amplitude of the sinusoidal input.
- **Output amplitude**: The amplitude of the sinusoidal solution.
- **Gain** or **amplitude response:** the amount by which the system scales the input amplitude to get the output amplitude, i.e., the ratio of the output to input amplitudes.
- **Complex gain:** the 'gain' for the complexified equation, i.e., the ratio of output to input.

- **Phase lag:** the angle by which the output maximum trails the input maximum.
- **Time lag:** the time by which the output maximum trails the input maximum.
- **Frequency response:** both amplitude response and phase lag taken together.

By looking at the solutions in Equations 16 and 17, we can give these quantities for the system discussed above. Pay attention to the abstract statements involving $P(i\omega)$, they are more useful to know than the formulas with square roots etc.

- Input: $B\cos(\omega t)$.
- Input frequency: ω.
- Input amplitude: Since we declared the input to be $B\cos(\omega t)$, the input amplitude is B.
- Output amplitude: $A(\omega) = \dfrac{kB}{|P(i\omega)|} = \dfrac{kB}{\sqrt{(k-m\omega^2)^2+b^2\omega^2}}$.
- Gain: The gain is the ratio of the output amplitude to the input amplitude. So the gain $g(\omega)$ is

$$g(\omega) = \frac{kB/|P(i\omega)|}{B} = \frac{k}{|P(i\omega)|} = \frac{k}{\sqrt{(k-m\omega^2)^2+b^2\omega^2}}.$$

- Complex gain: In the complexified DE, we replace $i\omega$ by s to get the equation with exponential input:

$$P(D)z = Be^{st},$$

where s can be any complex number, e.g., $i\omega$ or $2+3i$.

The input is Be^{st} and the output is $\dfrac{kBe^{st}}{P(s)}$. The complex gain is the ratio of output to input: $\dfrac{kBe^{st}/P(s)}{Be^{st}} = \dfrac{k}{P(s)} = \dfrac{k}{ms^2+bs+k}$.

- Phase lag: $\phi(\omega) = \text{Arg}(P(i\omega)) = \tan^{-1}\left(\dfrac{b\omega}{k-m\omega^2}\right)$ in Q1 or Q2.
- Time lag: $\phi(\omega)/\omega$.

9.2.2 Input and gain

Important note: The gain depends on what we designate as the input. Do not try to memorize the exact formulas for gain in the example above. In other systems the formulas will be slightly different. You will need to think about each system! Pay attention to this in all the examples below.

Example 9.2. Consider the damped harmonic oscillator driven by pushing on the end of the spring. If $f(t)$ is the displacement of the end from its equilibrium position, then the system is modeled by

$$mx'' + bx' + kx = kf(t).$$

In this case, it is reasonable to consider $f(t)$ to be the input.

Taking $f(t) = B\cos(\omega t)$, this is exactly the DE from Example 9.1 above. We saw that this has gain $g(\omega) = \dfrac{k}{|P(i\omega)|}$.

Example 9.3. Consider the system

$$2y'' + 1.5y' + 3y = 3B\cos(\omega t)$$

where we consider $B\cos(\omega t)$ to be the input. (Note the input does not include the factor of 3). Plot the graph of the gain as a function of ω.

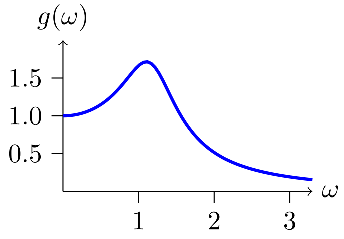
Solution: The sinusoidal solution to this equation is

$$y_p = \frac{3B\cos(\omega t - \phi(\omega))}{|P(i\omega)|} = \frac{3B}{\sqrt{(3-2\omega^2)^2 + (1.5\omega)^2}}\cos(\omega t - \phi(\omega)) \qquad \text{(where } \phi(\omega) = \text{Arg}(P(i\omega))\text{).}$$

So the gain (output amplitude/input amplitude) is

$$g(\omega) = \frac{3}{\sqrt{(3-2\omega^2)^2 + (1.5\omega^2)}}.$$

Here is the plot of $g(\omega)$:

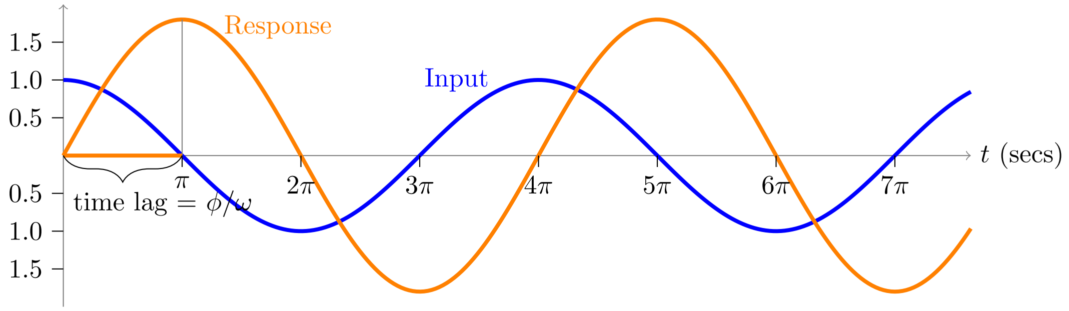


Graph of the gain function for Example 9.3

9.2.3 Phase Lag

Example 9.4. In the figure below the blue curve is the input and the orange curve is the response. The damping causes a lag between the time the input reaches its maximum and the time the output reaches its maximum.

- The figure shows that the output lags π seconds behind the input. This is the time lag
- The period of both input and response is 4π seconds. So the output is $\pi/4\pi = 1/4$ cycle $= \pi/2$ radians behind the input. The angle $\phi = \pi/2$ radians is the phase lag.



The response lags behind the input by π seconds or $\pi/2$ radians.

The phase lag is important in many applications, but in this class we will be more interested in the gain.

9.3 Amplitude response and practical resonance

The gain is a function of ω. It tells us the size of the system's response at the given input frequency. If the gain has a relative maximum at $\omega_r > 0$, then we call ω_r a **practical resonant frequency.**

Example 9.5. (Finding practical resonance.) Consider the system from Example 9.3:

$$2y'' + 1.5y' + 3y = 3B\cos(\omega t).$$

As in that example, we consider $B\cos(\omega t)$ to be the input. Find all the practical resonant frequencies.

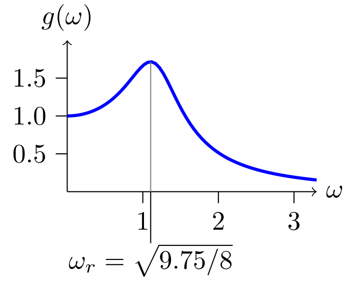
Solution: In Example 9.3 we found the gain function was

$$g(\omega) = \frac{3}{\sqrt{(3-2\omega^2)^2 + (1.5\omega)^2}}.$$

To find the practical resonance we have to find the value of ω where $g(w)$ has a maximum. There are a few tricks to shorten the algebra, but we'll find the maximum straightforwardly by setting $g'(\omega) = 0$.

$$g'(\omega) = -\frac{3}{2} \cdot \frac{-8\omega(3-2\omega^2) + 2(1.5)^2\omega}{((3-2\omega^2)^2 + (1.5\omega)^2)^{3/2}} = 0.$$

Setting the numerator to 0 and solving we find $\omega = 0$ or $\omega = \sqrt{9.75/8}$. We require the resonant frequency to be positive, so $\omega_r = \sqrt{9.75/8}$ is the only practical resonant frequency. The graph below shows that this is, in fact, a maximum. (You can also check this using calculus.)



Graph of the gain function with practical resonance marked.

Example 9.6. (A system with no practical resonant frequency.) Consider the system

$$2y'' + 10y' + 3y = 3B\cos(\omega t),$$

where we consider $B\cos(\omega t)$ to be the input. Find all the practical resonant frequencies.

Solution: This is similar to the previous example except that the damping constant is much larger. The algebra will be nearly identical, so we will skip past most of it. The gain is

$$g(\omega) = \frac{3}{\sqrt{(3-2\omega^2)^2 + (10\omega)^2}}.$$

So,

$$g'(\omega) = -\frac{3}{2}\cdot\frac{-8\omega(3-2\omega^2) + 2(10)^2\omega}{((3-2\omega^2)^2 + (10\omega)^2)^{3/2}} = 0.$$

Setting the numerator to 0 and solving for ω we find $\omega = 0$ or $\omega = \sqrt{-11}$. Since neither of these is a positive real number we say that there is no practical resonant frequency.

Example 9.7. Consider the system

$$my'' + by' + ky = F_0\cos(\omega t)$$

where we consider $F_0\cos(\omega t)$ to be the input. Find all the practical resonant frequencies.

Solution: The sinusoidal solution to this equation is

$$y_p = \frac{F_0\cos(\omega t - \phi(\omega))}{|P(i\omega)|} = \frac{F_0}{\sqrt{(k-m\omega^2)^2 + b^2\omega^2}}\cos(\omega t - \phi(\omega)), \qquad \text{where } \phi(\omega) = \text{Arg}(P(i\omega).$$

Therefore, the gain (output amplitude/input amplitude) is

$$g(\omega) = \frac{1}{\sqrt{(k-m\omega^2)^2 + b^2\omega^2}}.$$

Here we consider the system parameters m, b, k to be fixed, while the gain depends on the input parameter ω.

For this example, we'll show you a standard trick for finding the maximum of $g(\omega)$. The expression for $g(\omega)$ is one over a square root. So $g(\omega)$ has a maximum where the expression under the square root has a minimum. That is, we need to find the minima of

$$h(\omega) = \frac{1}{g^2} = (k-m\omega^2)^2 + b^2\omega^2.$$

Setting the derivative equal to 0 and solving for ω we get

$$h'(\omega) = -4m\omega(k - m\omega^2) + 2b^2\omega = 0.$$

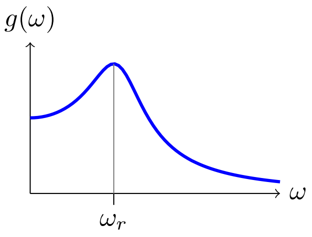
So, $\omega = 0$ or $\omega = \sqrt{k/m - b^2/2m^2}$. Since we require ω_r to be positive we have the following result.

- If $k/m - b^2/2m^2 > 0$ then this system has practical resonance at

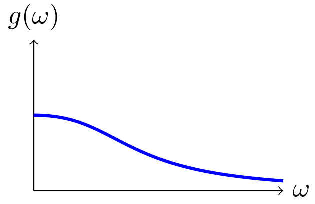
  $$\omega_r \;=\; \sqrt{k/m - b^2/2m^2} \;=\; \sqrt{\omega_0^2 - b^2/2m^2}.$$

  Here, the last expression gives ω_r in terms of the natural frequency $\omega_0 = \sqrt{k/m}$.

- If $k/m - b^2/2m^2 < 0$ then the system does not have a practical resonant frequency.



$\omega_r = \sqrt{\omega_0^2 - b^2/2m^2}$
(**practical resonance**).



$\omega_0^2 - b^2/2m^2 < 0$
(no practical resonance).

Notice that in this case if the damping gets too large there is no practical resonance.

For this example, see the mathlet
`https://mathlets.org/mathlets/amplitude-and-phase-second-order-iv/`.

In the text by Edwards and Penney, section 2.7 on radio circuits gives another another example and an application of this.

9.4 The undamped forced system

For a spring-mass system without any damping, we have what is called a **pure resonant frequency**. At this frequency, the amplitude of the response keeps growing to infinity. In this case, we say the **gain is infinite.** We show this with a somewhat general example using symbols for the coefficients.

Example 9.8. Solve the DE $my'' + ky = B\cos(\omega t)$.

Solution: We will only find the particular solution. You can supply the homogeneous solution. We start by doing some calculations we will need later.

1. The natural frequency of the system is $\omega_0 = \sqrt{k/m}$.

2. Characteristic polynomial: $P(r) = mr^2 + k$. We will need both $P(i\omega)$ and $P'(i\omega)$ in polar form.

$$P(i\omega) = k - m\omega^2 = |k - m\omega^2|e^{i\phi(\omega)}, \;\text{ where } \phi(\omega) = \begin{cases} 0 & \text{if } k - m\omega^2 > 0, \text{ i.e., } \omega < \omega_0 \\ \pi & \text{if } k - m\omega^2 < 0, \text{ i.e., } \omega > \omega_0 \end{cases}$$

$$P'(i\omega) = 2im\omega = 2m\omega e^{i\pi/2}.$$

Note that $P(i\omega) = 0$ exactly when $\omega = \sqrt{k/m} = \omega_0$.

Now use the sinusoidal response formula to get

$$y_p(t) = \frac{B\cos(\omega t - \phi(\omega))}{|P(i\omega)|} = \begin{cases} \frac{B\cos(\omega t)}{|k-m\omega^2|} & \text{if } \omega < \omega_0 \\ \frac{B\cos(\omega t - \pi)}{|k-m\omega^2|} & \text{if } \omega > \omega_0 \end{cases}$$

$$y_p(t) = \frac{Bt\cos(\omega_0 t - \phi(\omega_0))}{|P'(i\omega_0)|} = \frac{B\,t\cos(\omega_0 t - \pi/2)}{2m\omega_0} \quad \text{if } \omega = \omega_0.$$

Note: In the case $\omega = \omega_0$, we had to use the extended SRF since $P(i\omega_0) = 0$.

Also note, the factor of t in the case $\omega = \omega_0$.

9.4.1 Resonance and amplitude response of the undamped harmonic oscillator

Now let's take $B\cos(\omega t)$ to be the input to the system in the previous example. So the gain (output amplitude/input amplitude) for the system is

$$g(\omega) = \frac{1}{m|\omega_0^2 - \omega^2|}.$$

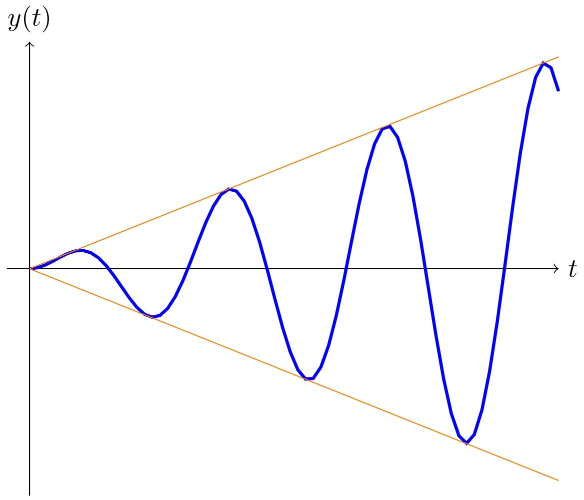
The right hand plot below shows $g(\omega)$ as a function of ω. There is a vertical asymptote at $\omega = \omega_0$. Note that the graph is similar to the graph of the gain for the damped harmonic oscillator except that the peak is infinitely high. Since we don't have a sinusoidal solution when $\omega = \omega_0$ there is no well defined gain at ω_0. However, given the graphs of the gain and the solution when $\omega = \omega_0$, it is conventional to say that the system has infinite gain at the frequency $\omega = \omega_0$.

Let's examine what this means. When $\omega = \omega_0$ we have

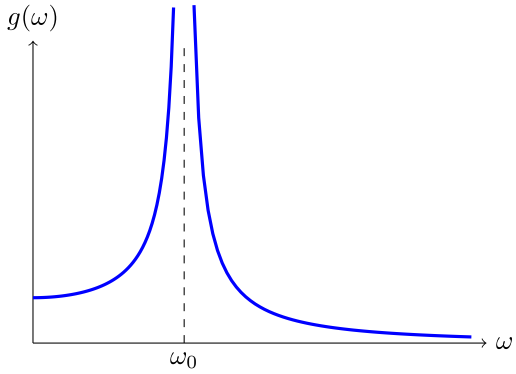
$$y_p(t) = \frac{B\,t\cos(\omega_0 t - \pi/2)}{2m\omega_0} = \frac{B\,t\sin(\omega_0 t)}{2m\omega_0}.$$

This is called **pure resonance**. The natural frequency ω_0 is called the **pure resonant frequency** or simply the **resonant** frequency of the system.

The graph of $y_p(t)$ is shown in the left-hand plot below. Notice that the response is oscillatory but not periodic. The amplitude keeps growing in time because of the factor of t in $y_p(t)$.



Resonance response ($\omega = \omega_0$)



Undamped amplitude response

Note carefully the different units and different meanings in the plots. The left-hand plot is output vs. time for a fixed input frequency. The right-hand plot is gain vs. input frequency. $x(t)$ and $g(\omega)$ are in physical units dependent on the system, t is in time, ω is in radians/time.

Physically, for the undamped oscillator, resonance happens because the input force is in sync with the natural frequency of the system and every push adds energy, so the energy in the system keeps growing to infinity. If the input frequency is different from ω_0, then sometimes the input force acts to add energy and sometimes it removes energy from the system, so the energy stays bounded. Likewise, if there is damping then the damping force is always removing energy from the system and a sinusoidal input can't cause the energy to grow without bound.

9.5 Slight variation of the undamped oscillator

Example 9.9. Consider the system $my'' + ky = f'(t)$ where we take $f(t)$ to be the input and $y(t)$ the response. Solve the DE when $f(t) = B\cos(\omega t)$ and give the gain of the system.

Solution: To find a particular solution we will complexify first and then take the derivative of $f(t)$. This is generally slightly easier than taking the derivative and then complexifying. The complexified DE becomes

$$mz'' + kz = (Be^{i\omega t})' = iB\omega e^{i\omega t}, \text{ with } y = \text{Re}(z).$$

As in Example 9.8, we have the following.

The natural frequency of the system is $\omega_0 = \sqrt{k/m}$.

$$P(i\omega) = k - m\omega^2 = |k - m\omega^2|e^{i\phi(\omega)}, \text{ where } \phi(\omega) = \begin{cases} 0 & \text{if } k - m\omega^2 > 0, \text{ i.e., } \omega < \omega_0 \\ \pi & \text{if } k - m\omega^2 < 0, \text{ i.e., } \omega > \omega_0 \end{cases}$$

$$P'(i\omega) = 2im\omega = 2m\omega e^{i\pi/2}.$$

Now use the exponential response formula (and its extended version) to get

$$z_p(t) = \begin{cases} \dfrac{Bi\omega e^{i\omega t}}{k - m\omega^2} = \dfrac{B\omega e^{i\pi/2} e^{i\omega t}}{|k - m\omega^2| e^{i\phi(\omega)}} & \text{if } \omega \neq \omega_0, \text{ where } \phi(\omega) = \text{Arg}(k - m\omega^2) \\ \dfrac{B\,ti\omega e^{i\omega_0 t}}{2im\omega_0} = \dfrac{B\,t\omega e^{i\pi/2} \omega e^{i\omega_0 t}}{2m\omega_0 e^{i\pi/2}} & \text{if } \omega = \omega_0 \end{cases}$$

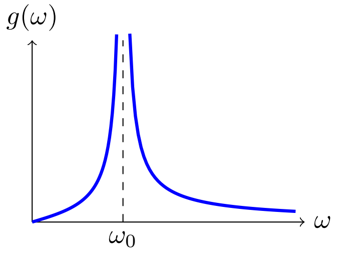
So,

$$y_p(t) = \text{Re}(z_p) = \begin{cases} \dfrac{B\omega\cos(\omega t + \pi/2)}{|k - m\omega^2|} & \text{if } \omega < \omega_0 \\ \dfrac{B\omega\cos(\omega t - \pi/2)}{|k - m\omega^2|} & \text{if } \omega > \omega_0 \\ \dfrac{B\,t\omega\cos(\omega_0 t)}{2m\omega_0} & \text{if } \omega = \omega_0 \end{cases}$$

Since the input is $B\cos(\omega t)$, we have the gain is

$$g(\omega) = \frac{\omega}{|k - m\omega^2|}$$

As in Example 9.8, there is a vertical asymptote at $\omega = \omega_0$. We also see the gain is 0 when $\omega = 0$. The amplitude response curve is shown below.



9.5.1 Zero-pole diagrams and gain

If there is time we will discuss this in class.

10 Direction fields, integral curves, existence of solutions

10.1 Goals

All of our goals refer to the first-order differential equation $y' = f(x, y)$.

1. Know the general form $y' = f(x, y)$ for a first-order DE.
2. Be able to use the method of isoclines to sketch the direction field of the DE and to sketch some integral (solution) curves.

3. Know the definition of a nullcline and be able to use nullclines to get a qualitative understanding of the solutions to a given DE.

4. Know the statement of the existence and uniqueness theorem for first-order DEs.

5. Be able to use, isoclines and known integral curves to form fences and funnels for the integral curves of a given DE.

10.2 Introduction

This unit is about first-order –not necessarily linear– differential equations. If x is the independent variable and $y(x)$ is a function of x then the general first-order DE is

$$y'(x) = f(x, y),$$

where $f(x, y)$ is some function.

Examples: $y' = x - y + 1, \quad y' = x^2 + y^2, \ldots$

In general, it is not possible to solve first-order equations exactly. Nonetheless without solving we can find approximate numerical solutions, use visual techniques to understand the systems and determine their long-term behavior.

In this topic we will explore visualization using direction fields. We will also state a general existence and uniqueness theorem that will give us confidence that our approximate techniques are approximating something that really exists.

10.2.1 Integral curves

Here is as good a place as any to introduce the term integral curve. An integral curve for a differential equation is the graph of a solution, i.e., a solution curve.

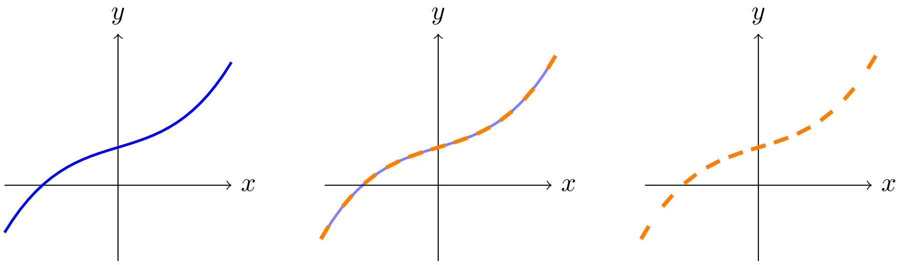
10.3 Direction or slope fields

We will motivate our use of direction fields with a simple example.

Example 10.1. Suppose you had the first-order differential equation

$$y' = f(x, y) \tag{18}$$

If you knew a solution you could simply graph it. Then at some points on the graph you could add a direction field element, i.e., a little tangent segment, along the graph. The first figure below shows just the curve. The second shows the and the curve with direction field elements added. The third figure shows just the direction field elements. Notice how well they represent the curve!



The slope elements show the shape of the curve.

We will also use the term slope element for direction field element.

Important point. The important point is that while we might not know the solution to Equation (18) at any point (x, y) we know the slope of the solution that goes through (x, y), i.e., slope $= f(x, y)$. This means we can always draw the direction element at (x, y). As we saw, these elements allow us to visualize the curves quite nicely.

10.4 Drawing direction fields using isoclines

The basic algorithm for drawing the direction field for Equation (18) is to choose a lot of points (x, y) and draw a slope element at each one. (As defined above, a slope element is a “little segment” of slope $f(x, y)$.) The key idea is that the (unknown) solution curve through (x, y) must have the same slope as the slope element.

Computer: With a computer drawing the slope field is easy, you just have the computer draw elements at an evenly spaced set grid of points. One tool we will use for this is the Isoclines mathlet: `https://mathlets.org/mathlets/isoclines/`.

By hand: People are not as patient as computers, so by hand we will use the method of isoclines. This limits the amount of computation needed and gives us some information which is not as readily accessible in the computer method.

Definition. The isocline of slope m for $y' = f(x, y)$ is the set of points (x, y) where $f(x, y) = m$, i.e., a set of points where all the slope elements have the same slope. (You can parse the word isocline as 'iso = same' and 'cline = slope'.)

Example 10.2. (Drawing a direction field using isoclines.) Consider the initial value problem (IVP)

$$y' = \sqrt{x^2 + y^2}; \quad y(0) = 0.5.$$

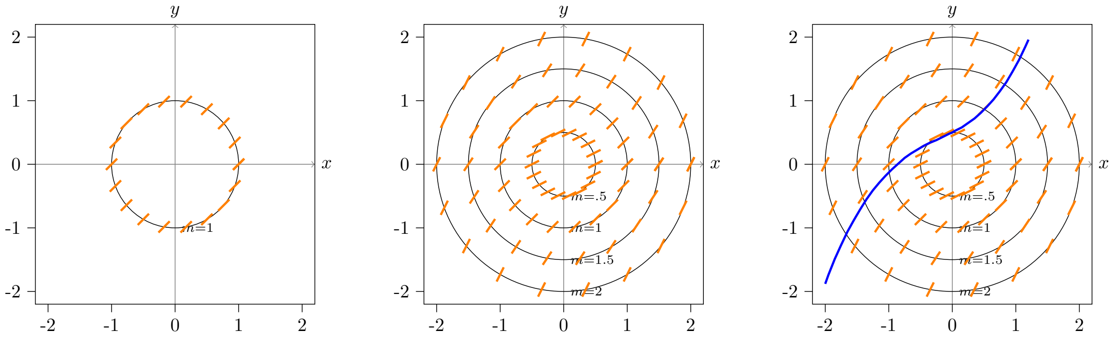
Draw a few isoclines ($y' =$ constant) for the DE and sketch the solution curve to the IVP.

Solution: Step 1 is to draw the isoclines. We need to find the set of points where $f(x, y) = m$ for various constants m. We'll draw isoclines for $m = 0.5$, 1, 1.5, 2.

$m = 1$: In our example, the isocline $f(x, y) = 1 = \sqrt{x^2 + y^2}$ is a circle of radius 1 in the xy-plane. We plot it by drawing the circle and then adding direction field elements of slope 1 along the circle. (See first figure below.)

Likewise for $m = 0.5$ the isocline is a circle of radius 1/2. We draw the circle and add direction field elements of slope 1/2 along it. We repeat this for $m = 1.5$ and $m = 2$. (See second figure below.)

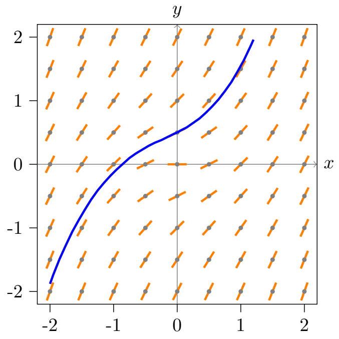
Step 2 is to sketch the solution curve $y = y(x)$ through the initial position (0,0.5). At each isocline the slope of the curve, $y'(x)$, should be the same as the slope of the direction field element on the isocline.



Isoclines and solution curve

Example 10.3. Redo the previous example using a computer to draw the slope elements at an array of points in the plane.

Solution: We instructed the computer to systematically loop through a two dimensional array of points. At each point it computes the direction element slope $f(x, y)$ and draws the element. (The integral curve was drawn using numerical methods discussed in the next topic.)



Computer generated slope field

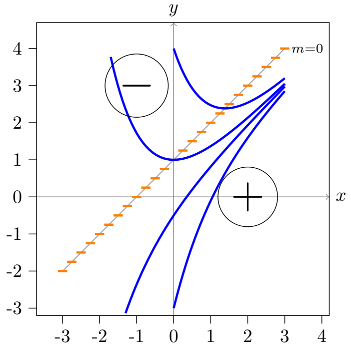
10.4.1 Nullclines

The nullcline for a first-order DE is the isocline corresponding to slope $m = 0$. The next example shows how just drawing the isocline can give a sense of how the solutions behave.

Example 10.4. Consider the DE $y' = f(x, y) = x - y + 1$. First draw the nullcline. Then indicate regions where the slope field has positive slope and those with negative slope. Use this information to guess at some solution curves $y = y(x)$. Describe in words how the solution curves behave.

Solution: The nullcline is where $f(x, y) = x - y + 1 = 0$, i.e., $y = x + 1$. This happens to be a line. We show it with its slope elements in the figure below. The nullcline divides the plane into two regions: above the nullcline the slope field is negative and below it, the field is positive.

With just this information, we can see that integral curves that start above the nullcline must decrease until they pass through the nullcline (with 0 slope) and then turn upwards. Those that start below the nullcline are always increasing.

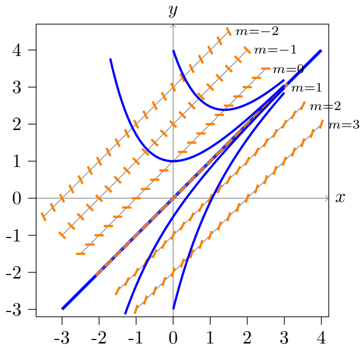


Nullcline and guessed integral curves for $y' = x - y + 1$.

Note. The existence and uniqueness theorem in the next section says that the solution curves can't cross. This means that it is a good guess –though not guaranteed– that the solution curves approach each other asymptotically as shown.

Example 10.5. Redo the previous example and include isoclines with $m = -2, -1, 0, 1, 2, 3$. Use the direction field to sketch a few solutions.

Solution: For any m the isocline $f(x, y) = m = x - y + 1$ is a line $y = x + 1 - m$. The figure shows the requested isoclines with their slope elements



It so happens (this is unusual, **don't expect it in other problems**) that the isocline for $m = 1$ is also an integral curve. All solutions go asymptotically to this curve $y = x$

Note. This example is a constant coefficient linear DE, so we could have found solutions analytically. This is certainly not the case for most first-order equations.

10.5 Existence and Uniqueness

Theorem. Existence and uniqueness for first-order differential equations.

Consider the initial value problem $y' = f(x, y); \quad y(x_0) = y_0$.

1. (Existence) If $f(x, y)$ is continuous then there is a solution.

2. (Uniqueness) If $\frac{\partial f}{\partial y}$ is also continuous then the solution is unique.

The proof of this involves more analysis than we have time for in 1803. For those who are interested, we've posted a note describing the Picard method of proof for this theorem.

Notes 1. The theorem says that if you have two different solutions $y_1(x)$ and $y_2(x)$, then for any x_0 the functions are not equal, i.e., $y_1(x_0) \neq y_2(x_0)$.

2. Graphically this means that integral curves never cross.

3. This theorem is important. It allows us to talk confidently about solutions without actually finding them.

10.5.1 Examples and counterexamples

As mathematicians it is important to remember that theorems have hypotheses and that we should check its hypotheses before using a theorem. The examples here show that the existence and uniqueness theorem can "fail" if its hypotheses are not met.

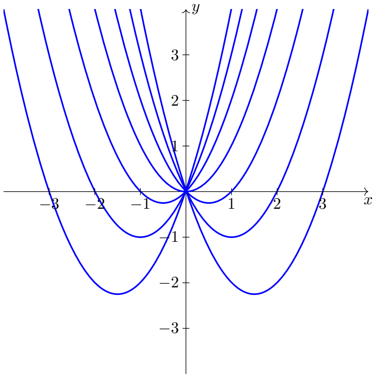
Important. Before reading these examples, remember that our main interest is in the cases where existence and uniqueness is true. Our most common application of this will be to assert that integral curves don't intersect.

Example 10.6. (Our most important DE) The IVP $y' = y; \quad y(x_0) = y_0$ satisfies the hypotheses of the existence and uniqueness theorem. Therefore, it has a solution and (for different initial conditions) the integral curves don't cross.

Example 10.7. (Non-existence and non-uniqueness) (See picture.) The DE $y' = y/x + x$ doesn't satisfy the hypotheses for the existence and uniqueness theorem because $f(x, y) = y/x + x$ is not continuous at $x = 0$. In fact, uniqueness fails because all solutions satisfy the same initial condition $y(0) = 0$. This is shown in the figure below.

Proof. This is a linear equation, so, using the variation of parameters formula, we find that the general solution is $y(x) = x^2 + Cx$. All of these solutions satisfy the initial condition $y(0) = 0$.

Note, the existence part of the theorem can also fail because there are no solutions that satisfy the initial condition, e.g., $y(0) = 1$.



A case where uniqueness fails: $y' = y/x + x$

Note. Away from $x = 0$ the function $f(x, y)$ is continuous, as is $\frac{\partial f}{\partial y}$, so existence and uniqueness holds, i.e., exactly one integral curve goes through any point (x_0, y_0) as long as $x_0 \neq 0$.

Example 10.8. Here is our standard example where a solution exists and is unique, but it is only defined on an interval –not the entire number line. The IVP $y' = y^2; \quad y(0) = 1$ has solution $y = \frac{1}{1-x}$.

The solution exists and is unique –and is only defined on the interval $(-\infty, 1)$.

Very briefly, here's an example where solutions always exist, but are not necessarily unique.

Example 10.9. Consider the DE $y' = 2\sqrt{|y|} = f(x, y)$
Since $f(x, y)$ continuous, the theorem says that solutions exist. For example,

$$\frac{\partial f}{\partial y} = \begin{cases} \frac{1}{\sqrt{y}} & \text{for } y > 0 \\ -\frac{1}{\sqrt{|y|}} & \text{for } y < 0 \end{cases}$$

is *not* continuous when $y = 0$. So the existence and uniqueness theorem doesn't guarantee uniqueness. In fact, there are two solutions: $y_1(x) = \begin{cases} x^2 & \text{for } x \geq 0 \\ -x^2 & \text{for } x \leq 0 \end{cases}$, and $y_2(x) = 0$, which both have initial condition $y(0) = 0$, i.e., solutions are not unique.

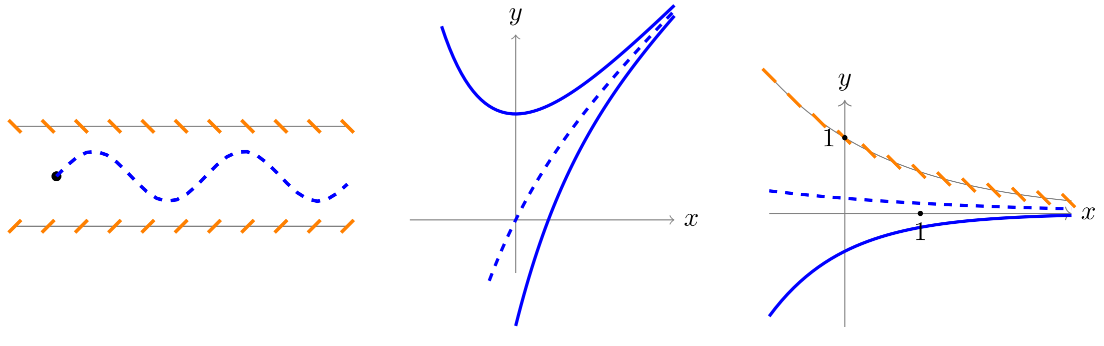
10.6 Squeezing: fences and funnels

In this section, as usual, we are looking at the first-order equation

$$y' = f(x, y).$$

To avoid problems we will assume that the existence and uniqueness theorem always holds, so that integral curves never intersect. Our goal is to see how we can use isoclines and known solutions to understand how unknown solutions will behave.

Both isoclines and integral curves can act as fences which other solution curves can't cross. Together they can form a funnel, which forces other solutions to stay between them and go asymptotically to some function. We explain this with some simple figures, which show isoclines and integral curves in several configurations.



Consider the upper isocline in the left hand figure. Since the slope field crosses from above to below this isocline, integral curves must do the same. That is, any solution that is above the isocline can cross to below, but any solution that is below the isocline must remain below it. We say, "integral curves can't cross an isocline against the slope field".

Since a fence is something that stops you from crossing a boundary, we call the upper isocline an upper fence on solutions, i.e., from below it looks like a fence to any solution. (From above, a solution does not see a fence and happily crosses it.)

Likewise, the lower isocline is a lower fence on solutions. That is, any solution that starts above it must stay above it.

Thus any solution, e.g., the blue dashed curve, that starts between the two fences must stay between them.

In the middle figure, all three curves are integral curves. The existence and uniqueness theorem says that integral curves can't intersect each other. This means that integral curves act as fences (both upper and lower) for other integral curves. This is illustrated in the middle figure, where the two solid blue integral curves constrain the blue dashed integral curve to stay between them.

Notice, that in the middle figure, the two fences become asymptotically closer. This says that the blue dashed curve will be squeezed between the fences and become asymptotically closer to them. In this case we say that the two integral curves form a funnel and solutions that start between them are asymptotically the same.

In the right hand figure we have an isocline acting as an upper fence and an integral curve as a lower fence. Together they form a funnel. Just like the funnel in the middle figure any solution that starts between them is funneled between them.

Example 10.10. Look at the right hand figure. Suppose that $y(x)$ is the solution to the IVP $y' = f(x, y);\quad y(0) = 0.5$. Estimate $y(100)$.

Solution: Since the integral curve of y starts inside the funnel, it must stay there and be squeezed down to 0. Looking at the scale on the x-axis, we see that $x = 100$ is very far to the right, so $y(100) \approx 0$.

11 Numerical methods for first-order differential equations

11.1 Goals

1. Be able to compute approximate solutions by hand using Euler's method.
2. Be able to compute the concavity of a solution and say whether Euler's method gives an over or under-estimate,
3. Know some of the ways numerical methods can fail or give misleading results
4. Know the broad outline of how other numerical methods work and understand that many of them are really fancier versions of Euler's method.

11.2 Introduction

In this topic we will look at numerical methods for approximating solutions to differential equations. Just like numerical integration, this allows us to approximate the solution to any first-order DE. It is especially valuable for those equations that we can't solve analytically. Using the computer we can then study as many solutions as we want for a given DE.

11.3 Generalities about numerical methods

The basic framework is that we are given a first-order DE with initial condition

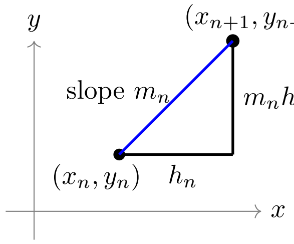
$$y' = f(x,y); \qquad y(x_0) = y_0$$

The goal is to estimate y(x) for other values of x.

The estimate is done by approximating $y(x)$ at a discrete set of points using a series of steps:

$$\text{Start at } (x_0, y_0), \quad \text{step to } (x_1, y_1), \quad \text{step to } (x_2, y_2), \quad \text{step to } (x_3, y_3) \ \ldots$$

Different numerical methods have different ways of computing each step. But they all have the following picture in common.



The triangle shows the step from (x_n, y_n) to (x_{n+1}, y_{n+1}). The horizontal step is h_n. The usual terminology is to call h_n the stepsize at step n. The vertical step is $m_n h_n$, where m_n is the slope at step n.

In the diagram to 'step' from (x_n, y_n) to (x_{n+1}, y_{n+1}) we have

$$x_{n+1} = x_n + h_n; \qquad y_{n+1} = y_n + m_n h_n$$

The job of a numerical method is to specify how to choose h_n and m_n at each step.

11.4 Euler's Method of numerical approximation

Our first method will be Euler's method. Euler's method is very simple to compute and is the only numerical method we will compute by hand. As an aside, it is analogous to using rectangles and Riemann sums to approximate an integral.

Just as in numerical integration, there are fancier numerical methods for solving DEs. These methods require more computation than Euler's and we will leave the computation to computers and existing software packages.

To describe Euler's method we need to say how to choose h_n and m_n for each step.

Euler's method is a fixed stepsize method. This means we fix the stepsize h at the beginning and use it for every step. That is, at each step $h_n = h$.

We know that the slope of the solution curve through (x_0, y_0) is $y' = f(x_0, y_0)$. Euler's method uses this slope to choose m_0, i.e., $m_0 = f(x_0, y_0)$. Likewise, for every subsequent step, Euler's method chooses m_n to be the slope of the direction field at (x_n, y_n), i.e.

$$m_n = f(x_n, y_n)$$

The next example illustrates how to use Euler's method.

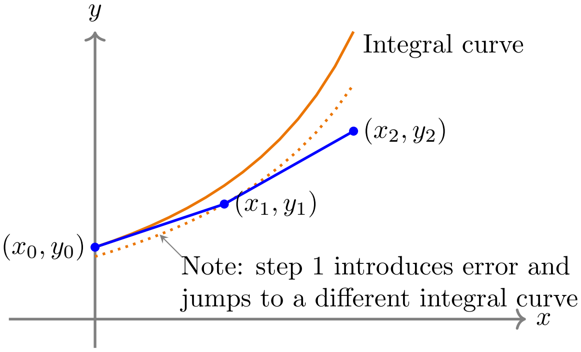
Example 11.1. Numerically solving an IVP using Euler's method. Consider the IVP $y' = x^2 + y^2;\quad y(0) = -1$. Use Euler's method to estimate $y(1)$.

Solution: We don't know $y(x)$ (and it's hard to find), but we can compute the direction field slope at each point.

Pick a *stepsize*: To keep the computation short, let's take $h = 0.25$. This will take 4 steps to go from $x_0 = 0$ to $x = 1$

Step 0 :	$x_0 = 0$	$y_0 = -1$	$m_0 = 1$	$m_0 h = 0.25$
Step 1 :	$x_1 = 0.25$	$y_1 = -0.75$	$m_1 = 0.63$	$m_1 h = 0.16$
Step 2 :	$x_2 = 0.5$	$y_2 = -0.59$	$m_2 = 0.60$	$m_2 h = 0.15$
Step 3 :	$x_3 = 0.75$	$y_3 = -0.44$	$m_3 = 0.76$	$m_3 h = 0.19$
Step 4 :	$x_4 = 1.00$	$y_4 = -0.25$		

So, $y(1) \approx y_4 \approx -0.25$



Example of Euler's method

In the next example we introduce a simple tabular format for doing and presenting the computation.

Example 11.2. Let $y' = y;\quad y(0) = 1$. Estimate $y(1)$

(Note: we know the exact answer, $y = e^x,\quad y(1) = 2.718\ldots$)

Let $h = 0.25$, so there are 4 steps from 0 to 1. We organize the calculation in a table:

n	x_n	y_n	$m_n = f(x_n, y_n)$	$m_n h$	actual	error
0	0	1.0	1.0	0.25	1.0	0.0
1	0.25	1.25	1.25	0.31	1.28	0.03
2	0.5	1.56	1.56	0.39	1.65	0.09
3	0.75	1.95	1.95	0.49	2.12	0.17
4	1.0	2.44			2.7183	0.28

Notes:
1. Organize hand calculations like this.
2. Error often accumulates.

Example 11.3. (Example continued.) We now continue the previous example with different stepsizes. In all cases we are trying to estimate $y(1)$.

Stepsize. $h = 1$ (this is just to be a little silly).

With $h = 1$ it takes 1 step to go from 0 to 1.0

n	x_n	y_n	$m_n = f(x_n, y_n)$	$m_n h$	actual	error
0	0	1.0	1	1.0	1.0	0.0
1	1.0	2.0			2.7183	0.72

Stepsize. $h = 0.1$.
With $h = 0.1$ it takes 10 steps to go from 0 to 1.0. Here is the table with some of the numbers left out.

n	x_n	y_n	$m_n = f(x_n, y_n)$	$m_n h$	actual	error
0	0	1	...	...	1	0
1	0.1					
2	0.2	1.21	...	...	1.2214	0.011
3	0.3					
4	0.4	1.4641	...	...	1.4918	0.028
5	0.5					
6	0.6	1.7716	...	...	1.8221	0.05
7	0.7					
8	0.8	2.1436	...	...	2.2255	0.082
9	0.9					
10	1.0	2.5937			2.7183	0.125

Note. The error is smaller when $h = 0.1$ than when $h = 0.25$

Rules of thumb: Using a smaller h is more accurate but requires more computation.

Mild warning. More computation means more risk of roundoff error. In this class, we never make h so small that this is a problem.

11.5 What can go wrong

In this section we'll see that numerical methods can sometimes give misleading results. We hasten to add that numerical methods provide an incredibly powerful tool which is used all the time with great success. But we do need to take some care to avoid certain pitfalls.

We expect that decreasing the stepsize should give a more accurate estimate. The next example shows that we shouldn't simply accept the result, no matter how small the stepsize used.

Example 11.4. Consider the IVP $y' = y^2$; $y(0) = 1$. Use Euler's method to approximate $y(1)$.

Solution: We know the exact solution is $y = \frac{1}{1-x}$, so $y(1) = \infty$. But Euler's method will happily estimate $y(1)$. We do this for several different stepsizes.

Take $h = 0.2$

n	x_n	y_n	$m_n = f(x_n, y_n)$	$m_n h$	actual	error
0	0	1	...	...	1	0
1	0.2	1.2	...	...	1.25	0.05
2	0.4	1.49	...	...	1.67	0.18
3	0.6	1.93	...	...	2.5	0.57
4	0.8	2.68	...	...	5	2.32
5	1.0	4.11			∞	∞

So, $y(1) \approx y_5 = 4.11$.

For decreasing values of h we get the following:
For $h = 0.1$, $y(1) \approx y_{10} = 37.6$.
For $h = 0.05$, $y(1) \approx y_{20} = 91.25$.
For $h = 0.025$, $y(1) \approx y_{40} = 238.21$.

Instead of settling down to a limiting value as we decrease h, the estimate grows. This is a sign that something is wrong with our estimates.

11.5.1 Lesson

You should try smaller and smaller h until the answer settles down. That is, run the estimate with stepsize h. The rerun it with stepsize $h/2$. If the estimates are very close then we have *one* good bit of evidence to accept the estimate as a good approximation. Otherwise, try $h/4$ etc. If the estimate never settles down, then we will have to reject the estimates and use other methods.

The computer doesn't eliminate the need to think!

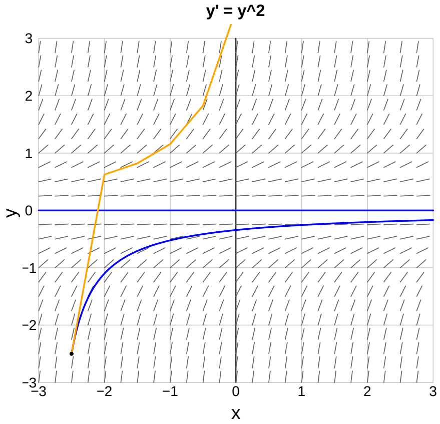
Note. We could make the previous example even more extreme by asking to estimate $y(2)$. The problem is that with the vertical asymptote at $x = 1$ the solution is not even defined at $x = 2$. Nonetheless, for any stepsize h Euler's method will produce an estimate of $y(2)$.

Example 11.5. Stepping across region boundaries. The following shows another risk in using numerical methods. Consider the IVP $y' = y^2$; $y(-2.5) = -2.5$.

The blue curve is the exact solution to the IVP. It goes asymptotically to $y = 0$

The orange curve is the Euler approximation using stepsize $h = 0.5$. It goes off to infinity.

The problem is that the first step in the approximation goes past the separatrix $y = 0$. After that, instead of going asymptotically to 0, the approximation continues to grow.



11.6 Other numerical techniques

All the techniques that we'll look at take steps of the form

$$x_{n+1} = x_n + h;\ y_{n+1} = y_n + m_n h.$$

where m_n is some sort of average slope near (x_n, y_n). The differences between the various methods are in how m_n and possibly h_n is chosen at each step. We'll only touch on this briefly.

Improved Euler (also called RK2). This is a fixed stepsize algorithm, that is we fix the value of h before using it. Here is the algorithm:

1. Start at (x_n, y_n)
2. Compute the slope $k_1 = f(x_n, y_n)$ and take a regular Euler step to a temporary point (x_a, y_a).
$$x_a = x_n + h;\ y_a = y_n + k_1 h.$$
3. Compute the slope at (x_a, y_a): $k_2 = f(x_a, y_a)$.
4. Average the two slopes: $m_n = (k_1 + k_2)/2$.
5. Use m_n as the slope to take the Improved Euler step.
$$x_{n+1} = x_n + h;\ y_{n+1} = y_n + m_n h.$$

Runge-Kutta 4 (RK4). This is also a fixed stepsize algorithm. You can do a web search to get the details. In brief, the algorithm computes 4 different slopes k_1, k_2, k_3, k_4 and then takes a weighted average of these slopes to get m_n. There are different ways to choose the ks and the weights, one common scheme is

$$\begin{aligned} k_1 &= f(x_n, y_n); & k_2 &= f(x_n + h/2, y_n + k_1 h/2); \\ k_3 &= f(x_n + h/2, y_n + k_2 h/2); & k_4 &= f(x_n + h, y_n + k_3 h) \\ m_n &= \frac{k_1 + 2k_2 + 2k_3 + k_4}{6}. \end{aligned}$$

Then as usual,

$$x_{n+1} = x_n + h;\ y_{n+1} = y_n + m_n h.$$

Variable step size methods. There is no reason we have to have a fixed stepsize. It is possible to adjust h at each step. One way to do this is the following:

Suppose we get to (x_n, y_n) with current stepsize h.

1. Take one RK4 step with stepsize h.
2. Repeat with stepsize $h/2$ and $2h$.
3. If the 3 results are very close then change the current stepsize to $2h$ and take the step. If they are not close then change the current stepsize to $h/2$ and take the step.

Thus sometimes the stepsize will get bigger and save computation. When needed to maintain accuracy it will get smaller.

11.7 More technical discussion on error size

(This section is for enrichment only. You will not be asked it on exams.)

For this discussion, we fix a first-order IVP: $y' = f(x, y)$; $y(x_0) = y_0$. We also fix the value x_f and ask to approximate $y(x_f)$.

Euler's method is linear in the error. This means that the error is roughly proportional to h. So, if you halve the stepsize, then you approximately halve the error. Of course, you also double the amount of computation.

Improved Euler is quadratic in the error. This means that the error is roughly proportional to h^2. So, if you halve the stepsize, then the error is approximately quartered.

RK4 is a fourth order method. This means that the error is roughly proportional to h^4. So, if you halve the stepsize, then the error is approximately multiplied by $1/16$.

11.8 Second derivative and concavity

If we know $y' = f(x, y)$, then we can find y''. This can be used to determine the concavity of the integral curve and thus, whether the Euler estimate is an over or underestimate.

Example 11.6. Assume $y' = 3xy$ and $y(1) = 2$. Use Euler's method to estimate $y(1.1)$. Is the estimate too high or too low?

Solution: First: $y'(1) = 6$.

Now fix the stepsize $h = 0.1$.

The Euler estimate is $y(1.1) \approx 2 + 0.1 \cdot 6 = 2.6$.

To find the concavity we compute the second derivative. (Note well that y is a function of x.) So,

$$y'' = (3xy)' = 3y + 3xy', \text{ so } y''(1) = y(1) + 3 \cdot y'(1) = 2 + 6 = 8 > 0.$$

We see that y is concave up at $x = 1$ and therefore the Euler estimate is (probably) too low. (Generally speaking, we should be cautious in our statement, because it's possible the graph of y changes concavity between $x = 1$ and $x = 1.1$. In this case, since x, y, y' are all positive, it is clear that $y'' > 0$ for any solution in the first quadrant.)

11.9 Relation to numerical integration

(This section is also just for your enjoyment and enrichment. We won't discuss it in class or on psets or exams.)

Even in 18.01 you were solving (simple) differential equations. A typical 18.01 integration question is to compute $\int_a^b f(x)\,dx$. We can rephrase this as the following initial value problem:

Let $y(x)$ be the solution to the IVP $y' = f(x)$; $y(a) = 0$. What is $y(b)$?

It is clear that this has solution $y(b) = \int_a^b f(x)\,dx$.

Thus for this IVP estimating $y(b)$ with numerical methods amounts to estimating the definite integral using numerical methods. More precisely

Euler's method = numerical integration using left Riemann sums with rectangles.

Improved Euler = numerical integration using the trapezoidal rule.

RK4 = numerical integration using Simpson's rule.

Example 11.7. (Euler's method = left Riemann sum.) For $y' = f(x)$, $y(a) = 0$ estimate $y(b)$ using Euler's method and N steps.

Solution: N steps implies the stepsize is $h = \frac{b-a}{N}$. Thus Euler's method gives

$$y_{n+1} = y_n + f(x_n)\,h.$$

This leads to the following table:

n	x_n	y_n
0	a	0
1	$a+h$	$f(x_0)\,h$
2	$a+2\,h$	$f(x_0)\,h + f(x_1)\,h$
3	$a+3\,h$	$f(x_0)\,h + f(x_1)\,h + f(x_2)\,h$
	...	
N	$a+Nh=b$	$f(x_0)\,h + (f(x_1) + f(x_2) + ... + f(x_{N-1}))\ h$

Thus our approximation is $y(b) = \sum_{j=0}^{N-1} f(x_j)\,h$. In 18.01 you might have learned to use Δx instead of h. In either case, the approximation is the left Riemann sum approximating $\displaystyle\int_a^b f(x)\,dx$.

12 Autonomous equations and bifurcation diagrams

12.1 Goals

1. Know the standard form of an autonomous, first-order differential equation.
2. Be able to use critical points to draw the phase line for an autonomous, first-order DE.
3. Be able to draw the bifurcation diagram for an autonomous, first-order DE with a parameter.
4. Be able to interpret phase lines and bifurcation diagrams in terms of population dynamanics and sustainability.

12.2 Introduction

In this topic we look at, so-called, autonomous equations. These are a special type of nonlinear first-order equations. In general, rather than solve these equations, we will try to understand the long-term behavior of the systems they model without finding the solution.

When the system includes a parameter, we will draw bifurcation diagrams which give us a system level view of the long-term behavior of the system for all possible values of the parameter. This is analagous to our use of gain curves, which tell us, in one graph, the behavior of the system for all possible input frequencies.

The Phase Lines Mathlet https://mathlets.org/mathlets/phase-lines/ illustrates everything we will do in this topic. We encourage you to look at it!

12.3 Autonomous differential equations

Definition. An autonomous first-order differential equation has the form

$$x'(t) = f(x).$$

(Compare this to the general first-order DE which has the form $x' = f(x, t)$.)

The word autonomous means *self-governing*. That is, x', the rate that x changes, depends only on x and not on t.

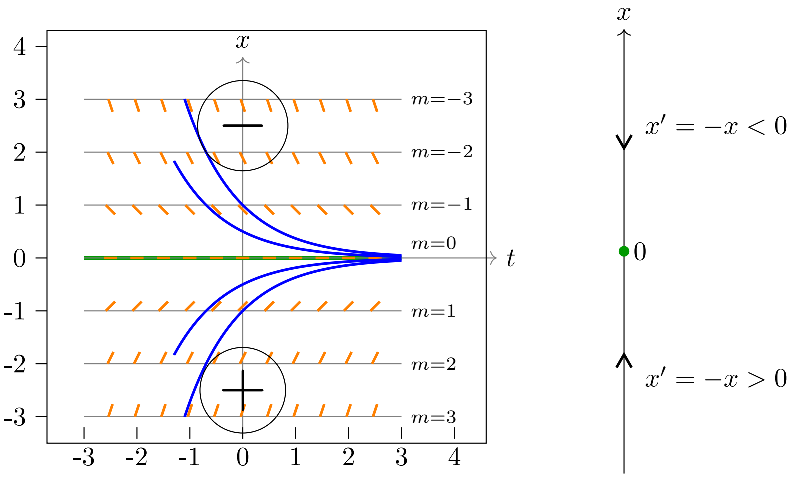
Here are some important properties of autonomous equations:
1. They are separable.
2. They can be hard to integrate.
3. We can say a lot about them without solving them. (More on this below.)
4. They are time invariant: if $x(t)$ is a solution then so is $x(t - t_0)$.

12.4 Direction fields and phase lines for autonomous equations

Our most important DE, $x' = kx$, is autonomous. We will use it to introduce phase lines for such equations. First, we look at its direction field.

Example 12.1. Use isoclines to draw the direction field for the DE $x' = -x$. Put the phase line (to be defined) next to it.

Solution: The isocline for slope m is $f(x) = -x = m$. This is a horizontal line. we draw the direction field and a few solutions using isoclines for $m = 0$, 1, 2, 3, -1, -2, -3.



Left: direction field for $\dfrac{dx}{dt} = -x$. Right: phase line

As always the nullcline separates the plane into regions where x' is positive and negative. These are marked with a big + and − on the direction field.

The phase line is a simplified version of the direction field. Since the direction field is independent of t, we just throw away the t-axis. The phase line is the x-axis. On it we mark the x-value of each nullcline, i.e., $x = 0$. Instead of slope field elements we put arrows indicating the direction of the slope field. These correspond to the big + and − in the direction field. In our example we have a down arrow in the region $x > 0$ and an up arrow in the region $x < 0$.

This simple example shows two important properties of autonomous equations.

1. For autonomous equations $x' = f(x)$, the isoclines are always horizontal lines. This is because the equation $f(x) = m$ is independent of t.

2. Any integral curve can be translated left or right and it is still an integral curve. That is, if $x(t)$ is a solution then so is $x(t - t_0)$. This is easy to see because the direction field is the same if you translate it right or left.

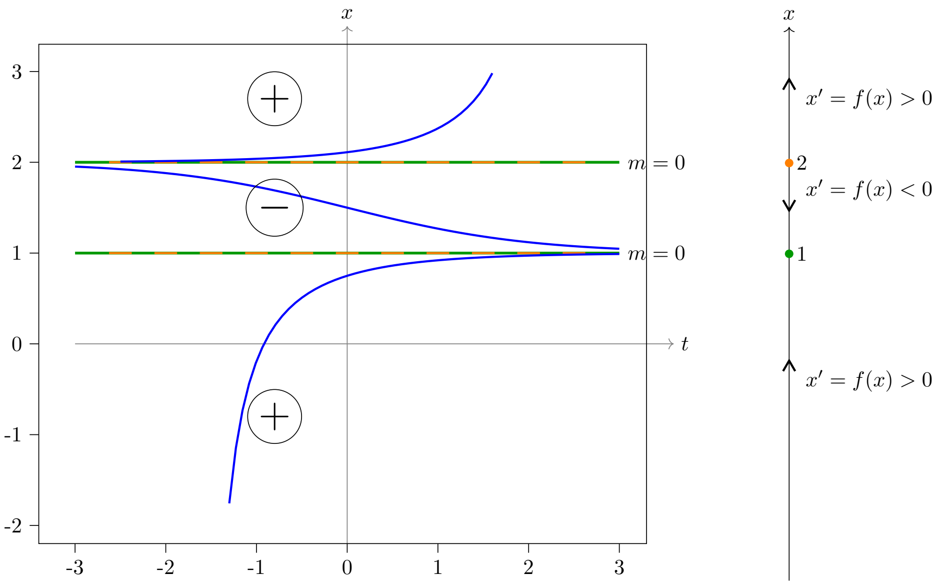
12.5 Equilibria, nullclines, constant solutions and critical points

For autonomous equations we will use a number of different words to describe nullclines. We'll introduce them through an example.

Example 12.2. Let $x' = (1 - x)(2 - x)$. Draw a direction field consisting of just the nullclines and large + or − signs indicating regions where the direction field has positive or negative slope. Using just this, sketch some solutions, including the ones along the nullclines.

Then use your direction field to draw the phase line for this system.

Solution: We have $x' = f(x) = (1 - x)(2 - x)$. The nullcline is where $f(x) = 0$, i.e., $x = 1$ and $x = 2$. These are horizontal lines in the tx-plane. It's easy to check that $x' > 0$ when $x > 2$ or $x < 1$ and $x' < 0$ when $1 < x < 2$. The sign of x' in different regions is marked with a + or a −.



Direction field and phase line for $x' = f(x) = (1 - x)(2 - x)$.

Now for the main point of this example: The nullclines $x = 1$ and $x = 2$ are clearly solutions. We use the following terms to describe them.

Because they are constants, they are called constant solutions.

Because they are unchanging, they are called equilibrium solutions.

Because $x' = 0$ along them, we call $x = 1$ and $x = 2$ critical points for the DE.

To finish the example we added solution curves. In regions where $x' > 0$ the solution curves are increasing. Because the equilibrium solutions act as fences, these solutions can't cross them. So we get the picture as shown.

The phase line is drawn next to the direction field. The arrows on the phase line show the sign of x', i.e., the direction of the slope field, for different ranges of x.

12.5.1 Lost solutions

Finally, nullclines correspond to lost solutions: The equation $\dfrac{dx}{dt} = f(x)$ is separable. When we separate variables we get $\dfrac{dx}{f(x)} = dt$. So there are lost solutions where $f(x) = 0$. These are the nullclines (or constant solutions or equilibrium solutions).

12.6 Stability of equilibria

In general, we say an equilibrium is stable if nearby solutions go asymptotically to the equilibrium value.

Example 12.3. Looking at Example 12.2, give each equilibrium and say whether it is stable or unstable.

Solution: The equilibria are the same as the constant solutions. These are $x = 1$ and $x = 2$. Looking at the phase line, we see clearly that $x = 1$ is stable and $x = 2$ is unstable.

You can see the same thing in the direction field.

12.7 Analyzing an autonomous DE

We will use the following steps to analyze the autonomous DE $x' = f(x)$.

1. Find the critical points $x' = f(x) = 0$ and plot them on the phase line.

2. Determine the sign of x' for different values of x. Use these to put arrows on the phase line. This can be done algebraically or graphically.

3. Determine the stability of the equilibrium solutions.

4. If desired, sketch some solutions in the tx-plane.

We illustrate this with some examples.

Example 12.4. Let $x' = -k(x - A)$. This models Newton's law of cooling for a body of temperature x in an environment of temperature A. We assume that k and A are constants, with $k > 0$.

Plot the phase line. Be sure to indicate the stability of the equilibrium solutions. Also, give a rough sketch of solutions in the tx-plane.

Solution: First, note that this equation is simple enough that we actually know the general solution

$$x = A + ce^{-kt}.$$

You should check that our answers agree with this!

We follow the steps outlined above.

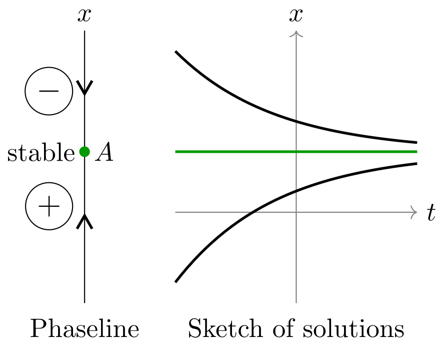
1. Find the critical points: $f(x) = -k(x - A) = 0$ implies $x = A$. This is indicated on the phase line below. Remember: For autonomous equations, critical points are the same as equilibrium solutions.

2. Determine the sign of x' for different x: This is the same algebra you used in 18.01 when graphing a function and looking for regions where it increases and decreases.

It's easy to see that when $x > A$ we have $x' = -k(x - A) < 0$. Likewise, when $x < A$ we have $x' > 0$. We use this to add arrows to the phase line. For this example, we also label regions with a $+$ or $-$.

3. The arrows on the phase line show that the equilibrium $x = A$ is stable.

4. Directly from the phase line, we can sketch some solutions. Note: these are in the tx-plane. The equilibrium solution is the horizontal line $x = A$. The other solutions are strictly qualitative: they are drawn to show that all solutions go asymptotically to the (stable) equilibrium.



Example 12.5. (Logistic equation.) Consider the autonomous system

$$x' = k(M - x)x = f(x).$$

We assume k and M are positive constants. This is called a logistic population model. For the population x it models the growth rate as $k(M - x)$. The growth rate depends on x, and decreases as x increases. (Compare this with the exponential model $x' = ax$, where the growth rate is constant.) This model captures the notion that, as the population increases, the competition for scarce resources leads to a lower growth rate. If the population gets too large the growth rate will become negative.

Plot the phase line for this system and sketch some solutions.

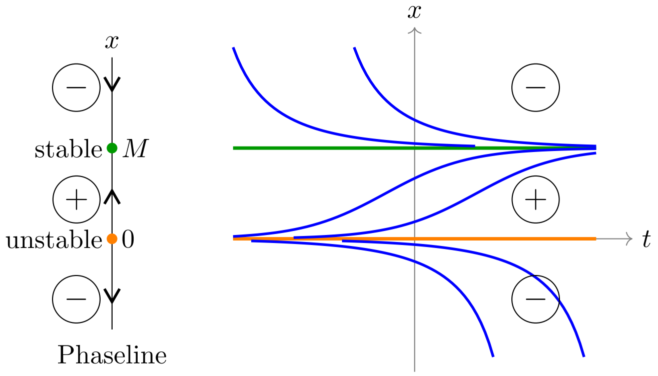
Solution: We follow the standard steps

1. Critical points: $x' = k(M - x)x = 0$ gives critical points $x = M$ or $x = 0$.

2. Looking at the x axis, it is clear we have the following signs for x':
when $x > M$, then $x' < 0$,
when $0 < x < M$, then $x' > 0$,
when $x < 0$, then $x' < 0$,

3. Using 1 and 2 we can draw the phase line. This shows that $x = M$ is a stable equilibrium and $x = 0$ is unstable.

4. Finally, it is a simple matter to sketch solution curves: they can't cross the equilibria and must go towards the stable equilibrium and away from the unstable equilibrium. As before, these are made up, but they capture the qualitative nature of the solutions.



Notes. 1. The S shaped curves between 0 and M are called logistic curves. The Wikipedia

article https://en.wikipedia.org/wiki/Logistic_function gives a number of applications where the logistic function appears.

2. Because the population stabilizes at M and the growth rate becomes negative if $x > M$, we call M the carrying capacity of the environment.

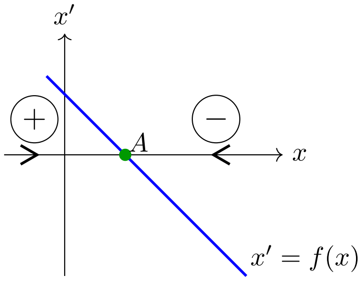
3. It is difficult to find the reason for the name logistic. The term was coined around 1844 by the French mathematician Pierre François Verhuist. (See the same Wikipedia article cited above.)

12.7.1 Graphical method for determining the sign of x'.

In the examples above we found the sign on x' by testing values in different ranges of x. Here we'll show an alternative graphical method. The trick is to graph x' vs. x. When doing this, we are viewing x' as a variable. We illustrate by redoing some of the examples.

Example 12.6. Find the phase line from Example 12.4 by graphing x' vs. x and putting the phase line on the x-axis.

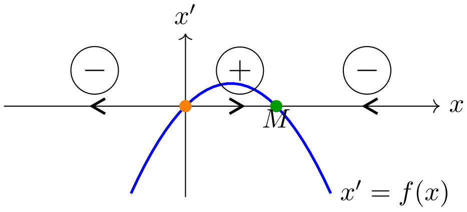
Solution: In the example we have $x' = -k(x - A)$. The graph of this is the negatively sloped line shown below. It is now easy to see the sign of x' as a function of x. When $x > A$, the graph is below the x axis, so x' is negative. Likewise, when $x < A$, the graph is above the x-axis, so x' is positive. We mark these regions with $-$ and $+$. The arrows on the x-axis correspond to these signs. Magically, the x-axis now shows the phase line for the system.



x' vs. x. The x-axis shows the phase line.

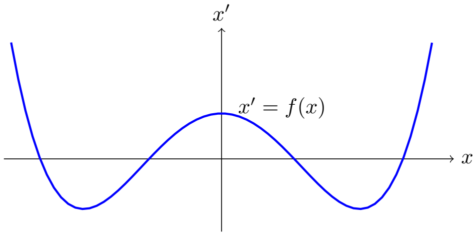
Example 12.7. Find the phase line from Example 12.5 by graphing x' vs. x and putting the phase line on the x-axis. (The DE is $x' = k(M - x)x$.)

Solution: As in the previous example we plot x' vs. x. Then we use the sign of x' to add arrows to the x-axis. The plot is a downward pointing parabola. As before, the x-axis shows the phase line.

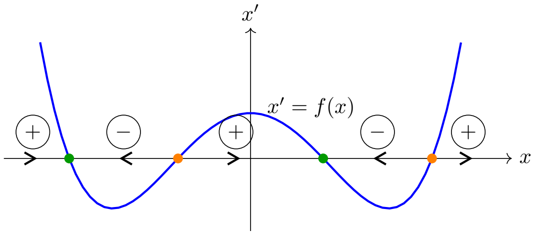


x' vs. x. The x-axis shows the phase line.

Example 12.8. The following shows the graph of $x' = f(x)$. Use the graph, to draw the phase line for this system. Indicate the critical points and their stability.



Solution: We add arrows to the graph. The critical points are marked green for stable and orange for unstable.



12.8 Parameters and bifurcation diagrams

Bifurcation diagrams help us visualize how the system behaves at different settings of a given control parameter. This is similar to what we did when we graphed gain vs. input frequency. The input frequency is a parameter and the gain curve lets us see in one figure how the system responds to any frequency.

We'll get at this idea using examples.

12.8.1 Logistic with harvesting population model

Example 12.9. This example will not show a bifurcation diagram. Instead, we will try to show how we might be led to inventing bifurcation diagrams.

Suppose you are growing irises. Left alone in your garden, the population of irises follows a logistic population model

$$x' = (3 - x)x,$$

where x is in units of 1000 irises and time is in units of months.

Your plan is to harvest and sell the flowers at a constant rate of a units/month. With this level of harvesting, the population model becomes

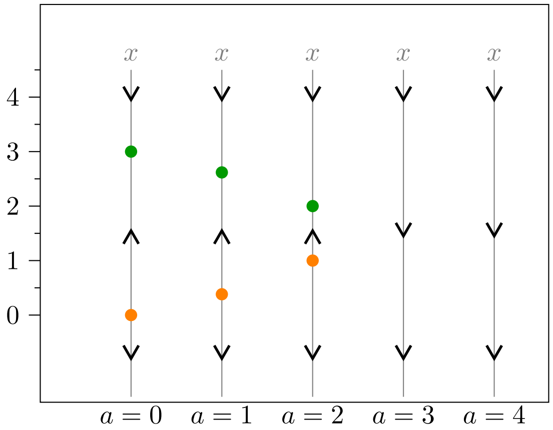
$$x' = f(x) = x(3-x) - a. \tag{19}$$

You know that if a is too large then the iris population will crash and you'll go out of business. So your first goal to understand what happens to the population for different values of a.

For any value of a, we can draw the phase line and determine how the population will respond at that value. So your assignment for the population model in Equation 19 is to draw phase lines for every value of a!

Okay, that is probably too hard, let's just do it for each of the values $a = 0, 1, 2, 3, 4$.

Solution: It's not hard to compute critical points for each of these a. We don't show the calculation. Here are the phase lines



The phase lines for $a = 0, 1, 2$ each have two critical points. The upper one is stable and the lower one is unstable. For $a = 3, 4$, there are no critical points.

Clearly, it's a bad idea to harvest at the rates $a = 3$ or $a = 4$. In these cases the population will decrease to 0. So these rates are not sustainable.

The rates $a = 0, 1, 2$ each have a positive stable critical point. In all three cases, if we wait to start harvesting until the population is about 1.5, then the population will go to the stable critical value. This is sustainable.

Our conclusion is that it is possible to harvest at the rate $a = 2$ without ruining our business.

12.8.2 Sustainability

Definition. If the population model has a positive stable critical point we say the population is sustainable.

Note. Sustainability doesn't mean you can't mess it up. For instance, in Example 12.9, if $a = 2$ and we start harvesting when $x = 0.5$, then the population will crash to 0. This would be a bad idea, but we still say that $a = 2$ is a sustainable harvesting rate. That is, as long as you do it right and start harvesting when the population is large enough, then the population will stabilize at the stable critical point.

12.8.3 Bifurcation diagrams

In the previous example we were unable to draw phase lines for every value of a, so we drew a small number of them to help choose a harvesting rate. We saw that we could sustainably harvest when $a = 2$, but not at $a = 3$. What about other values of a? This is the motivation behind bifurcation diagrams, they'll show us how the system behaves for all values of a in one simple graph.

Definition. Suppose we have a population $x(t)$ with a model which depends on a parameter a. The bifurcation diagram for this model is the plot of all the points (a, x) in the ax-plane where the model has a critical point. We always indicate on the diagram whether the critical points represent stable or unstable equilibria.

We illustrate bifurcation diagrams by redoing Example 12.9.

Example 12.10. Draw the bifurcation diagram for the logistic with harvesting model

$$x' = x(3 - x) - a$$

For which values of a is the population sustainable?

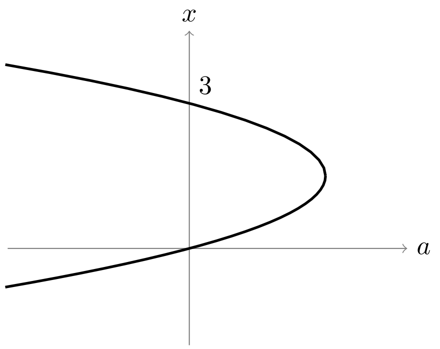
Solution: We use the following steps.

Step 1. Draw the ax-axes. Be sure to label them!

Step 2. Compute and plot all the critical points. In this case we have

$$x(3 - x) - a = 0 \qquad \Rightarrow \qquad a = x(3 - x).$$

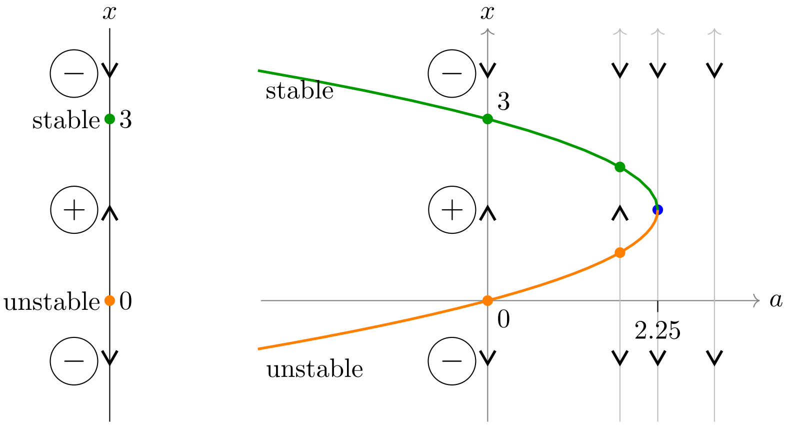
Since a is the horizontal axis, the graph of this is a sideways parabola:



Step 3. The plot divides the plane into 2 regions (inside and outside the parabola). Since the plot is the set of points where $x' = 0$, the sign of x' is the same throughout each region.

We can find those signs by testing points in each region. For example, at the point $(a, x) = (0, 1)$, we have $x' = 2 > 0$. So, inside the parabola, we have $x' > 0$. Likewise, at $(0, 4)$, $x' = -4 < 0$. So, outside the parabola, we have $x' < 0$.

Another method, which amounts to the same thing, is to use phase lines. Below, the phase line for $a = 0$ is shown on the left and also on the bifurcation diagram. On the bifurcation diagram it is the vertical line at $a = 0$.



Phaseline: $a = 0$ Bifurcation diagram

The arrows tell us the sign of x' at points on the phase line. Which, just like testing points, allows us to give the sign of x' in the two regions determined by the critical points.

These signs then tell us the stability of the critical points. In this example, the upper branch of the parabola consists of stable critical points and the lower branch consists of unstable critical points.

It is a simple matter to use the signs to add a few more phase lines to our picture. We add one through the vertex of the parabola and also ones to the left and right of the vertex.

The phase line through the vertex of the parabola shows it is semistable. The vertex is at the maximum value of a as a function of x. In this case, it's easy use calculus, or the geometry of parabolas, to find these coordinates: $a = 2.25$, $x = 1.5$.

Finally we can say when the population is sustainable: Since there is a positive stable critical point for $a < 2.25$, the population is sustainable in this region. It is not sustainable for $a \geq 2.25$.

Definition. A bifurcation point is any value of a where there is a qualitative change in the critical points.

In the previous example, the value $a = 2.25$ is the point where the critical points change –there are two critical points for $a < 2.2.5$ and none for $a > 2.25$. Therefore, $a = 2.25$ is called a bifurcation point.

Example 12.11. Suppose a population is modeled by the DE $x' = -ax + 1$, which is a constant birth-and-death rate, modified to include a constant rate of replenishment.

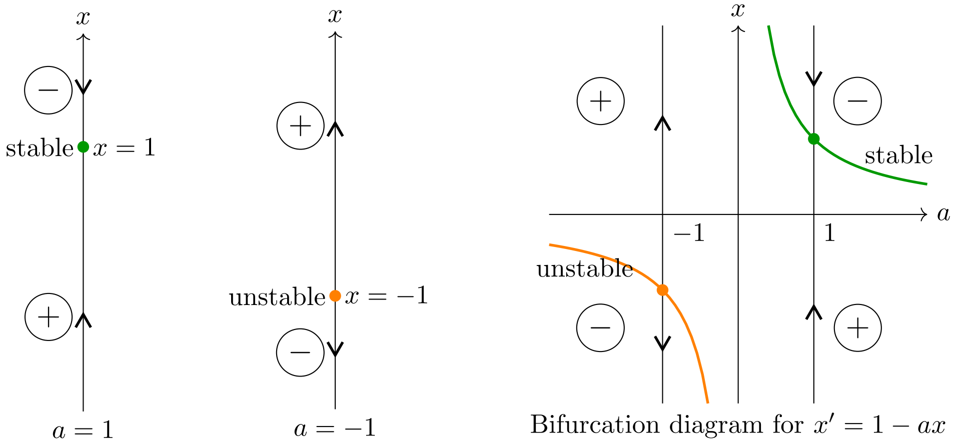
(i) Sketch the bifurcation diagram and list any bifurcation points (i.e., special values of a).

(ii) The bifurcation point(s) divide the a-axis into intervals. Illustrate one case for each interval by giving the phase line diagram. For each of these phase lines give (rough) sketches of solutions in the tx-plane.

(iii) For what values of a is the population sustainable. What happens for other values of a.

Note the MIT Mathlet The Phase Lines Mathlet https://mathlets.org/mathlets/phase-lines/ can show this system.

Solution: We answer (i) and (ii) together. The critical points are $x' = -ax + 1 = 0$. So, $x = 1/a$. We graph this in the ax-plane –it's a hyperbola with two branches. Here is the finished bifurcation diagram with two phase lines. These are explained below.



Bifurcation diagram for $x' = 1 - ax$

After plotting the critical points we see that the graph divides the ax-plane into 3 regions. In order to determine the sign of x' in each region we found phase lines for $a = 1$ and $a = -1$. These are shown at the left. Determining the direction of the arrows was straightforward and we leave it for the reader to supply the details.

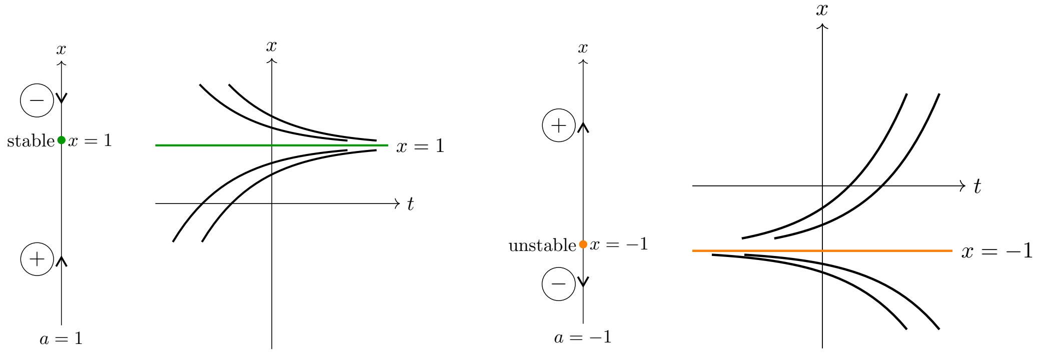
We place the phase lines on the bifurcation diagram at $a = 1$ and $a = -1$. The arrows on the phase lines then tell us the sign of x' in all 3 regions.

Once we know the sign on x', it's a simple matter to decide the stability of each part of the diagram. The stable branch is drawn in green and labeled 'stable'. Likewise the unstable branch is drawn in orange and labeled 'unstable'.

There is one bifurcation point at $a = 0$. This is a bifurcation point because the bifurcation diagram is different on either side of $a = 0$.

(iii) When $a > 0$ there is a positive stable equilibrium, so the population is sustainable. When $a \leq 0$ the population is not sustainable. In fact, it blows up to infinity.

Finally, we do our duty and sketch some solution curves based on the phase lines.



You can look at this example and the logistic with harvesting example in the Phase Lines Mathlet https://mathlets.org/mathlets/phase-lines/ phase lines applet:

12.9 Appendix: solution to logistic equation

Just for kicks, we compute the exact solution to the logistic population model

$$x' = kx(M - x)$$

This is separable. We need to use partial fractions to integrate the x side.

$$\frac{dx}{x(M-x)} = k\,dt.$$

So, $\displaystyle\int \frac{dx}{x(M-x)} = kt + C$.

Partial fractions: $\displaystyle\frac{1}{x(M-x)} = \frac{1/M}{x} + \frac{1/M}{M-x}$.

So, $\displaystyle\int \frac{dx}{x(M-x)} = \frac{\ln(|x|)}{M} - \frac{\ln|M-x|}{M} = \frac{1}{M}\ln\left(\frac{|x|}{|M-x|}\right)$.

So, $\displaystyle\ln\left(\frac{|x|}{|M-x|}\right) = Mkt + C$.

Exponentiating and changing e^{MC} to C gives: $\displaystyle\frac{x}{M-x} = Ce^{Mkt}$.

Solving for x: $\displaystyle x(t) = \frac{MCe^{Mkt}}{1 + Ce^{Mkt}}$.

We can also rewrite this as $\displaystyle x(t) = \frac{MC}{e^{-Mkt} + C}$.

We were a little sloppy with the absolute values, but more care would give the same results.

Note: If $C > 0$ then the solution $x(t)$ has $0 < x < M$. If C is negative, then these solutions blow up when $e^{-Mkt} + C = 0$.

13 Linear algebra: vector spaces, matrices and linearity

13.1 Goals

1. Know the definition of a vector space and how to show that a given set is a vector space.

2. Know the meaning of the phrase closed under addition and scalar multiplication.

3. Know how to convert a higher order DE into the companion system of first-order DEs.

4. Know how to organize matrix multiplication as a linear combination of the columns of the matrix.

5. Know how to organize matrix multiplication in block form and recognize when block multiplication is valid.

13.2 Introduction

Up to now we have spent most of our time in 18.03 considering linear differential equations. For these, one of our main tools was linearity, or, equivalently, the superposition principle. There are many other domains where linearity is important. For example, systems of linear algebraic equations and matrices. In this next unit on linear algebra we will study the common features of linear systems.

To do this we will introduce the somewhat abstract language of vector spaces. This will allow us to view the plane and space vectors you encountered in 18.02 and the general solutions to a differential equation through the same lens. In 18.02 vectors had both an algebraic and a geometric interpretation. In 18.03 we will focus primarily on the algebraic side of vectors, though we will sometimes use our geometric intuition as a guide.

13.3 Matlab (and alternatives)

We will use Matlab for computation and visualization. It will allow us to work with larger matrices where we wouldn't want to do computations by hand. We will only use a tiny subset of Matlab's enormous set of functions. I'll post some simple (and short) tutorials on its use.

Matlab is available for free to MIT students.

A free substitute for Matlab is Octave. It has the advantage that it loads much faster and doesn't spread digital rights management files all around your computer. The disadvantage is that it can be a little harder to install, especially on the Mac. Look at `https://www.gnu.org/software/octave/download.html`. I can help you get it installed if you want to try.

Another excellent and free substitute is Julia. The syntax is similar, but not identical, to Matlab. Downloads and documentation are available at `https://julialang.org`.

13.4 Linearity and vector spaces

We've seen before the importance of linearity when solving differential equations $P(D)x = f(t)$. To remind you: the operator $P(D)$ is linear means that

$$P(D)(c_1 f + c_2 g) = c_1 P(D) f + c_2 P(D) g$$

for all functions f, g and constants c_1, c_2.

Matrix multiplication is also linear. If A is a matrix and $\mathbf{v_1}$, $\mathbf{v_2}$ are vectors, then

$$A \cdot (c_1 \mathbf{v_1} + c_2 \mathbf{v_2}) = c_1 A \cdot \mathbf{v_1} + c_2 A \cdot \mathbf{v_2}$$

Example 13.1. In this example we will write an matrix multiplication in a way that emphasizes the linearity.

$$\begin{aligned} \begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 3+4 \\ 7+8 \end{bmatrix} &= \begin{bmatrix} 6(3+4)+5(7+8) \\ (3+4)+2(7+8) \end{bmatrix} \\ &= \begin{bmatrix} 6\cdot 3+5\cdot 7+6\cdot 4+5\cdot 8 \\ 1\cdot 3+2\cdot 7+1\cdot 4+2\cdot 8 \end{bmatrix} \\ &= \begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 3 \\ 7 \end{bmatrix} + \begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 4 \\ 8 \end{bmatrix} \end{aligned}$$

Linearity/Superposition
Exactly like solving linear differential equations, solving linear systems of algebraic equations involves finding a particular solution and superpositioning with the homogeneous solution.

Example 13.2. Solve $\begin{bmatrix} 1 & 3 \\ 4 & 12 \\ 3 & 9 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 8 \\ 6 \end{bmatrix}$

Solution: For this example we'll use ad hoc methods to find particular and homogeneous solutions. Later, we will learn systematic methods. The main point here is that the solutions can be superpositioned.

By inspection we can see one solution is $\mathbf{x_p} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$. Just as valid would be to take $\mathbf{x_p} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$ or $\mathbf{x_p} = \begin{bmatrix} 5 \\ -1 \end{bmatrix}$.

Next we have to solve the associated homogeneous equation:

$$\begin{bmatrix} 1 & 3 \\ 4 & 12 \\ 3 & 9 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

This expands to three equations in two unknowns. You can easily check that the general solution is $\mathbf{x_h} = c \begin{bmatrix} 3 \\ -1 \end{bmatrix}$.

By superposition, the solution to the original equation is

$$\mathbf{x} = \mathbf{x_p} + \mathbf{x_h} = \begin{bmatrix} -1 \\ 1 \end{bmatrix} + c \begin{bmatrix} 3 \\ -1 \end{bmatrix}.$$

If this is unclear, you should check the solution by substitution.

13.5 Vector spaces

The word space is used in mathematics to describe a set with extra properities. Math has all kinds of spaces. Here we will be concerned with vector spaces.

In order to have the notions of linearity and superposition, we need to have the notions of adding and scaling. This leads to the definition of vectors, whose key property is that they can be added and scaled.

Definition. A vector space is any set V with the following properties.

1. The set V has the arithmetic operations of addition and scalar multiplication.

2. Closure under addition: The sum of any two elements in V is another member V. That is, if $\mathbf{v}$, $\mathbf{w} \in V$, then $\mathbf{v} + \mathbf{w} \in V$.

3. Closure under scalar multiplication: Scaling an element of V results in another member of V. That is, if $\mathbf{v} \in V$ and c is a scalar, then $c\mathbf{v} \in V$.

4. Distributive law: If $\mathbf{v}$, $\mathbf{w} \in V$ and c is a scalar, then $c(\mathbf{v} + \mathbf{w}) = c\mathbf{v} + c\mathbf{w}$.

- An element $\mathbf{v}$ in V is called a vector.
- The formal definition of a vector space requires some more technical properties, but this definition will suffice for 18.03.
- If the scalars are required to be real numbers, we say we have a real vector space. If we allow them to be complex numbers, then we have a complex vector space.

The next few examples will introduce some important vector spaces and show how to check whether or not a given set is a vector space.

Key point. Checking whether or not a given set is a vector space is always easy. This is similar to checking whether a given operator is linear.

Example 13.3. Show that the set of ordered pairs (x, y), under the usual rules of addition and scalar multiplication, is a vector space.

Solution: We have to show the set satisfies the four properties in the definition of vector space. As we said, this is straightforward.

1. Multiplication and scalar multiplication: By definition we have these operations.

2. Closure under addition: Take any two ordered pairs (x_1, y_1) and (x_2, y_2). Their sum, $(x_1, y_1) + (x_2, y_2) = (x_1 + x_2, y_1 + y_2)$ is also an ordered pair.

3. Closure under scalar multiplication: Take any ordered pair (x, y) and scalar c, then $c(x, y) = (cx, cy)$ is also an ordered pair.

4. Distributive law: We show this without any commentary:

$$c\left((x_1, y_1) + (x_2, y_2)\right) = c(x_1 + x_2, y_2 + y_2) = ... = c(x_1, y_1) + c(x_2, y_2).$$

Since the set satisfies the four properties, it is a vector space.

Example 13.4. (Vector spaces.) The following are all examples of vector spaces. You should be able to check this exactly as we did in the previous example.

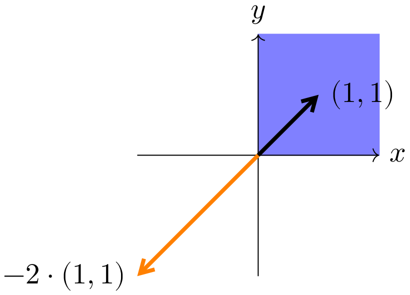
- We denote the plane by $\mathbf{R}^2$. It is the set of all ordered pairs (x, y).
- We denote space by $\mathbf{R}^3$. It is the set of all ordered triples (x, y, z).
- The powers indicate the dimension of each space. Likewise we can work with high dimensional vector spaces like $\mathbf{R}^{1000}$ which consists of all lists of 1000 numbers.
- In 18.03 we have used the fact that functions can be added and scaled. The set of solutions to the homogeneous DE

  $$x'' + 8x' + 7x = 0$$

  is a vector space. That is, the set $\{c_1e^{-t} + c_2e^{-7t}\}$ satisfies the above requirements 1-4 for a vector space.

Example 13.5. (Non-vector spaces.) The following are not vector spaces.

1. The set of plane vectors in the first quadrant. This fails to be closed under scalar multiplication. For example, $(1, 1)$ is in the first quadrant, but $-2 \cdot (1, 1) = (-2, -2)$ is not.



2. The set of functions of the form $\cos(6t) + c_1e^{-t} + c_2e^{-7t}$. This fails to be closed under addition. For example,

$$(\cos(6t) + 2e^{-t} + 3e^{-7t}) + (\cos(6t) + e^{-t} + 4e^{-7t}) = 2\cos(6t) + 3e^{-t} + 7e^{-7t}$$

The sum is not in the set because of the factor of 2 in front of $\cos(6t)$.

13.6 Connection to DEs

We will give two examples showing directly how matrices arise in differential equations.

Example 13.6. The companion matrix -converting a higher order DE to a first-order system. Here we are going to convert a higher order differential equation into a system of first-order equations. Later we will see how this technique allows us to understand DEs in a new way and also how it allows us to use numerical techniques on higher order equations.

Consider the second-order linear differential equation

$$\ddot{x} + 8\dot{x} + 7x = 0.$$

We've worked this example many times. The general solution is $x = c_1e^{-t} + c_2e^{-7t}$.

To convert the DE to a matrix system, we introduce a new variable: $y = \dot{x}$. Now, substituting y for $\dot{x}$ in the original DE we get the equation $\dot{y} + 8y + 7x = 0$. Altogether we have the system of two first-order linear DEs:

$$\begin{array}{rcrcr} \dot{x} & = & & & y \\ \dot{y} & = & -7x & - & 8y \end{array}$$

This can be rewritten in matrix form:

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -7 & -8 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

Notice two things:

1. If we write this abstractly with $\mathbf{x} = \begin{bmatrix} x \\ y \end{bmatrix}$ and $A = \begin{bmatrix} 0 & 1 \\ -7 & -8 \end{bmatrix}$, it looks like $\dot{\mathbf{x}} = A\mathbf{x}$. Ignoring the fact that $\mathbf{x}$ is a vector and A is a matrix, this looks like our most important DE: $\dot{x} = ax$.

2. Solving the system is equivalent to solving the original equation. That is, if we solve the original equation, we'll have found x and hence $y = \dot{x}$. Conversely, if we solve the matrix system, we'll have found x (the solution to the orginal DE) and $y = \dot{x}$.

In this case we already know the solution to the DE, so the solution to the system is

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x \\ \dot{x} \end{bmatrix} = \begin{bmatrix} c_1e^{-t} + c_2e^{-7t} \\ -c_1e^{-t} + -7c_2e^{-7t} \end{bmatrix} = c_1e^{-t} \begin{bmatrix} 1 \\ -1 \end{bmatrix} + c_2e^{-7t} \begin{bmatrix} 1 \\ -7 \end{bmatrix}$$

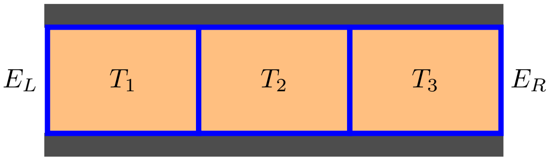
Notice that the basic solutions are of the form $e^{rt}\mathbf{v}$, where $\mathbf{v}$ is a constant vector. Later, we will use the method of optimism to guess solutions of this form.

Definition. The matrix A of coefficients that arises from this technique will be called the companion matrix to the original DE.

Example 13.7. Heat Flow. In this example we will set up a model for heat flow. We won't solve it for a few days.

Suppose we have a metal rod where different parts are at different temperatures. We divide it into 3 regions and imagine that each region exchanges heat with the adjacent regions. The regions on either end also exchange heat with the environment. We assume that the top and bottom of the rod are insulated, so that heat can only flow out of the bar at the ends. We also assume that the heat transfer follows Newton's law and the rate constant is k at each interface.

The figure below shows the metal bar divided into 3 regions and insulated above and below. The temperature of each region and the temperature of the environment on the left and right ends are indicated in the figure.



Using Newton's law we can write a DE for the temperature of each region.

$$\begin{array}{rcrcrcrcrcrcr}
\dot{T}_1 & = & -k(T_1 - E_L) & - & k(T_1 - T_2) & = & -2kT_1 & + & kT_2 & + & & & kE_L \\
\dot{T}_2 & = & -k(T_2 - T_1) & - & k(T_2 - T_3) & = & kT_1 & - & 2kT_2 & + & kT_3 & & \\
\dot{T}_3 & = & -k(T_3 - T_2) & - & k(T_3 - E_R) & = & & & kT_2 & - & 2kT_3 & + & kE_R
\end{array}$$

We can write this in matrix form

$$\begin{bmatrix} \dot{T}_1 \\ \dot{T}_2 \\ \dot{T}_3 \end{bmatrix} = \begin{bmatrix} -2k & k & 0 \\ k & -2k & k \\ 0 & k & -2k \end{bmatrix} \begin{bmatrix} T_1 \\ T_2 \\ T_3 \end{bmatrix} + \begin{bmatrix} kE_L \\ 0 \\ kE_R \end{bmatrix}$$

Remark: This particular coefficient matrix occurs quite often in applications. You should make sure you know how to modify the equation if we use n divisions of the rod instead of 3.

13.7 Matrix Multiplication

Here we will assume that you are comfortable with matrices and matrix multiplication. For completeness, we've added a quick review of some of the basics below.

Combination of columns

We can view the result of multiplying a matrix times a vector as a linear combination of the columns of the matrix. **We will use this again and again,** so you should internalize it now! We illustrate with an example:

Example 13.8. Consider the following product

$$\begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix} \cdot \begin{bmatrix} 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 6 \cdot 3 + 5 \cdot 4 \\ 1 \cdot 3 + 2 \cdot 4 \end{bmatrix} = 3 \begin{bmatrix} 6 \\ 1 \end{bmatrix} + 4 \begin{bmatrix} 5 \\ 2 \end{bmatrix}$$

Notice that the result is a linear combination of the columns of the matrix.

To express this abstractly we write a matrix as

$$A = \begin{bmatrix} \\ \mathbf{v_1} & \mathbf{v_2} & \mathbf{v_3} & \mathbf{v_4} & \mathbf{v_5} \\ \\ \end{bmatrix}$$

Here each $\mathbf{v_j}$ is a vector representing a column of A. We then have

$$\begin{bmatrix} \\ \mathbf{v_1} & \mathbf{v_2} & \mathbf{v_3} & \mathbf{v_4} & \mathbf{v_5} \\ \\ \end{bmatrix} \cdot \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \end{bmatrix} = c_1\mathbf{v_1} + c_2\mathbf{v_2} + c_3\mathbf{v_3} + c_4\mathbf{v_4} + c_5\mathbf{v_5}$$

That is, the product is a linear combination of the columns of A.

Block matrices and multiplication

Consider the following matrix $A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 6 & 5 & 0 & 0 \\ 1 & 2 & 0 & 0 \end{bmatrix}$. We can divide this into blocks

$$A = \left[\begin{array}{cc|cc} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \hline 6 & 5 & 0 & 0 \\ 1 & 2 & 0 & 0 \end{array}\right] = \left[\begin{array}{c|c} 0 & I \\ \hline B & 0 \end{array}\right]$$

As long as the sizes of the blocks are compatible, block matrices multiply just like matrices:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \cdot \begin{bmatrix} E \\ F \end{bmatrix} = \begin{bmatrix} AE + BF \\ CE + DF \end{bmatrix}$$

To convince yourself of this look at the following product and see that the blocks in the first column on the left only touch the top block on the right etc.

$$\left[\begin{array}{cc|cc} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \hline 6 & 5 & 0 & 0 \\ 1 & 2 & 0 & 0 \end{array}\right] \cdot \left[\begin{array}{cc} a & b \\ c & d \\ \hline e & f \\ g & h \end{array}\right]$$

13.7.1 Review: matrix notation

For a matrix A, we give its size as rows × columns. So a 2×3 matrix has 2 rows and 3 columns. We write $A_{i,j}$ for the entry in the i^{th} row and j^{th} column.

Example 13.9. For the 2×3 matrix $A = \begin{bmatrix} 1 & 3 & 5 \\ 7 & 9 & 11 \end{bmatrix}$, the $1, 2$ entry is $A_{1,2} = 3$. Likewise, the $2, 3$ entry is $A_{2,3} = 11$.

13.7.2 Review: matrix multiplication

Written out formally the i, j-entry of AB is given by the dot product of the i^{th} row of A dotted with the j^{th} column of B. That is

$$i, j\text{-entry of } AB = \langle i^{\text{th}} \text{ row of } A \rangle \cdot \langle j^{\text{th}} \text{ column of } B \rangle$$

This is illustrated in the following example.

Example 13.10. In the matrix product below, we've put a line through the 3rd row of first matrix and the 2nd column of the second matrix. The dot product of this row and column is $3, 2$-entry of the product, in this case it's 51.

$$\begin{bmatrix} 3 & 4 \\ 5 & 6 \\ 7 & 8 \\ 0 & 1 \end{bmatrix} \cdot \left[\begin{array}{c|c} 6 & 5 \\ 1 & 2 \end{array}\right] = \begin{bmatrix} * & * \\ * & * \\ * & 51 \\ * & * \end{bmatrix}$$

Example 13.11.

$$\begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix} \cdot \begin{bmatrix} 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 6 \cdot 3 + 5 \cdot 4 \\ 1 \cdot 3 + 2 \cdot 4 \end{bmatrix} = \begin{bmatrix} 38 \\ 11 \end{bmatrix}.$$

Only compatibly sized matrices can be multiplied. For matrices A and B: the product AB only makes sense if the number of columns of A equals the number of rows of B.

That is, the product AB only makes sense if the A is an $m \times n$ matrix and B is an $n \times p$ matrix. The product AB is an $m \times p$ matrix.

Example 13.12. (i) A 4×2 times a 2×3 gives a 4×3 matrix:

$$\begin{bmatrix} 6 & 5 \\ 1 & 2 \\ 7 & 8 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 6 & 5 & 3 \\ 1 & 2 & 4 \end{bmatrix} = \begin{bmatrix} 41 & 40 & 38 \\ 8 & 9 & 11 \\ 50 & 51 & 53 \\ 1 & 2 & 4 \end{bmatrix}$$

(ii) A 2×2 times a 2×3 gives a 2×3 matrix:

$$\begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix} \cdot \begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{bmatrix} = \begin{bmatrix} 16 & 38 & 60 \\ 5 & 11 & 17 \end{bmatrix}$$

(iii) A 2×3 times a 3×1 gives a 2×1 matrix:

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \cdot \begin{bmatrix} 7 \\ 8 \\ 9 \end{bmatrix} = \begin{bmatrix} 50 \\ 122 \end{bmatrix}$$

(iv) A 2×2 times a 3×2 is **not okay**.

$$\begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix} \cdot \begin{bmatrix} 3 & 4 \\ 5 & 6 \\ 7 & 8 \end{bmatrix} \quad \textbf{NOT A VALID EXPRESSION}$$

Matrix multiplication is **NOT commutative**. That is, except in rare cases, $AB \neq BA$. Indeed, sometimes the matrices are only compatible for one order of multiplication.

Example 13.13. Let $A = \begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix}$ and $B = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$ then the product AB is legitimate, but the product BA does not make sense.

Even, if the product is legitimate in either order, the products can be different.

Example 13.14. Let $A = \begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 2 \\ 0 & 3 \end{bmatrix}$ then the following multiplications show that $AB \neq BA$

$$\begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 0 & 3 \end{bmatrix} = \begin{bmatrix} 6 & 27 \\ 1 & 8 \end{bmatrix} \quad \text{but} \quad \begin{bmatrix} 1 & 2 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} 8 & 9 \\ 3 & 6 \end{bmatrix}$$

Lesson: You need to be careful and precise when doing matrix algebra. Make sure you multiply in the correct order.

2. Identity: The following matrices are called **identity** matrices:

$$I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad I_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

They are called the identity for the same reason the scalar 1 is called the multiplicitave identity. That is if you multiply the identity times anything you get back that anything. For example

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}, \quad \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 5 \\ 6 \\ 7 \end{bmatrix} = \begin{bmatrix} 5 \\ 6 \\ 7 \end{bmatrix}$$

Identity matrices are always **square** matrices. That is they have the same number of rows and columns. We use the subscripts in I_2, I_3 etc. to indicate the size of the identity. If the size is clear from the context we drop the subscript and just write I for the identity matrix.

14 Row reduction and subspaces

14.1 Goals

1. Be able to put a matrix into row reduced echelon form (RREF) using elementary row operations.
2. Know the definitions of null and column space of a matrix.
3. Be able to use RREF to find bases and describe the null and column spaces of a matrix.
4. Know the definitions of span and independence for vectors.
5. Know the definitions of basis and dimension for a vector space.
6. Know that the column space $= \{\mathbf{b}\}$ for which the equation $A\mathbf{x} = \mathbf{b}$ has a solution.
7. Be able to solve $A\mathbf{x} = \mathbf{b}$ by superpositioning a particular solution and the general homogeneous solution.
8. Be able to describe the geometric effects transforming vectors using matrix multiplication.

14.2 Introduction

Row reduction is a systematic computational method of simplifying a matrix while retaining some of its key properties. This will give us a systematic method of solving systems of linear equations by finding a particular solution and the general homogeneous solution.

The computational goal of row reduction is to simplify the matrix to the so called row reduced echelon form (RREF). Once in this form, we can easily read off some important properties of the original matrix.

Among these properties are the notions of null space and column space, which are two of the fundamental vector subspaces associated to a matrix. In order to discuss these spaces, we will need to learn the general concepts of independence of vectors, and basis and dimension of a vector space.

You will see that we have already seen all of these things, using different terms, in 18.03. We'll illustrate with our standard example: The homogenous equation

$$x'' + 8x' + 7x = 0$$

has two independent solutions $x_1 = e^{-t}$ and $x_2 = e^{-7t}$. Thus the equation has a two dimensional vector space of solutions with basis $\{x_1, x_2\}$. That is, every solution is a linear combination $c_1x_1 + c_2x_2$ of the two basis functions. We say that the space of homogeneous solutions is a two dimensional subspace of the vector space of all functions.

The last section in this topic will introduce the idea that matrix multiplication can be viewed as a transformation or mapping of vectors. At base, this is just the idea that a matrix times a vector is another vector. Looked at geometrically, we will see that matrix multiplication transforms a square to a parallelogram and a circle to an ellipse.

14.3 Row reduction

Row reduction is a computational technique for systematically simplifying a matrix or system of equations. It involves stringing together the elementary row operations listed below. We will see that it is exactly the same as using elimination to solve a system of equations.

We will explain its use through a series of examples.

Elementary row operations

1. Subtract a multiple of one row from another.
2. Scale a row by a non-zero number.
3. Swap two rows.

Example 14.1. Applying row operations to a matrix A.

Start with $A = \begin{bmatrix} 1 & 3 \\ 2 & 6 \\ 4 & 12 \end{bmatrix}$. Perform the following row operations.

Subtract 2·Row$_1$ from Row$_2$: $\sim \begin{bmatrix} 1 & 3 \\ 0 & 0 \\ 4 & 12 \end{bmatrix}$.

Subtract 4·Row$_1$ from Row$_3$: $\sim \begin{bmatrix} 1 & 3 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$.

Example 14.2. Use elimination to solve the following system of equations

$$\begin{array}{ccccccc} x & + & & & 2z & = 4 \\ 2x & + & y & + & 7z & = 14 \\ x & + & & & 3z & = 5 \end{array}$$

Solution: We use elimination: subtract 2·Equation$_1$ from Equation$_2$ and at the same time subtract Equation$_1$ from Equation$_3$.

$$\begin{array}{ccccccc} x & + & & & 2z & = 4 \\ & & y & + & 3z & = 6 \\ & & & & z & = 1 \end{array}$$

Now solve the system from the bottom up: $z = 1 \Rightarrow y = 3 \Rightarrow x = 2$.

Let's redo the previous example writing out just the augmented coefficient matrix

$$\left[\begin{array}{ccc|c} 1 & 0 & 2 & 4 \\ 2 & 1 & 7 & 14 \\ 1 & 0 & 3 & 5 \end{array}\right]$$

Follow the same operations in the example: subtract 2· Row$_1$ from Row$_2$ and subtract Row$_1$ from Row$_3$.

$$\left[\begin{array}{ccc|c} 1 & 0 & 2 & 4 \\ 0 & 1 & 3 & 6 \\ 0 & 0 & 1 & 1 \end{array}\right]$$

This represents the same system of equations and row operations as in the previous example.

14.4 Echelon Form

The final matrix in the previous example is in echelon form. By definition, this means the first non-zero element in each row is farther to the right than the one in the row above. Said differently, below the staircase is all zeros and in the corner of each stair is a nonzero number.

The word echelon seems to have military origins and means a step like arrangement. Here are two examples of matrices in echelon form with the staircase shown. The matrix on the left is the same as the one just above.

$$\left[\begin{array}{ccc|c} \textcircled{1} & 0 & 2 & 4 \\ 0 & \textcircled{1} & 3 & 6 \\ 0 & 0 & \textcircled{1} & 1 \end{array}\right] \qquad \begin{bmatrix} \textcircled{1} & 2 & 2 & 4 & 5 \\ 0 & 0 & \textcircled{1} & 6 & 0 \\ 0 & 0 & 0 & \textcircled{1} & 2 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

The first non-zero element in each row is called a pivot. They are circled in the matrices just above.

14.4.1 Reduced row echelon form (RREF)

A matrix is in reduced row echelon form (RREF) if
1. Each pivot is 1.
2. Each pivot column is all zeros except for the pivot.
3. The rows with all zeros are all at the bottom.

Example 14.3. Use the elementary row operations to put the matrix $\begin{bmatrix} 1 & 2 & 0 & 5 \\ 2 & 4 & 1 & 13 \\ 1 & 2 & 1 & 8 \end{bmatrix}$ in RREF.

Solution: Here are the row operations. R_2 means Row 2 etc. The notation $R_2 = R_2 - 2 \cdot R_1$ means change R_2 by subtracting $2R_1$ from it (like computer code).

$$\begin{bmatrix} 1 & 2 & 0 & 5 \\ 2 & 4 & 1 & 13 \\ 1 & 2 & 1 & 8 \end{bmatrix} \xrightarrow{R_2 = R_2 - 2R_1} \begin{bmatrix} 1 & 2 & 0 & 5 \\ 0 & 0 & 1 & 3 \\ 1 & 2 & 1 & 8 \end{bmatrix} \xrightarrow{R_3 = R_3 - R_1} \begin{bmatrix} 1 & 2 & 0 & 5 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 1 & 3 \end{bmatrix} \xrightarrow{R_3 = R_3 - R_2} \begin{bmatrix} \textcircled{1} & 2 & 0 & 5 \\ 0 & 0 & \textcircled{1} & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

The last matrix is in RREF. Again we rewrite it to emphasize the pivots and the echelon.

$$\begin{bmatrix} \textcircled{1} & 2 & 0 & 5 \\ 0 & 0 & \textcircled{1} & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

The pivots of R are circled. The columns with pivots are called pivot columns the other columns are called free columns. We have

$$R \text{ pivot columns: } \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \qquad R \text{ free columns: } \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 5 \\ 3 \\ 0 \end{bmatrix}$$

We use these to name the same columns in A:

$$A \text{ pivot columns: } \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \qquad A \text{ free columns: } \begin{bmatrix} 2 \\ 4 \\ 2 \end{bmatrix}, \begin{bmatrix} 5 \\ 13 \\ 8 \end{bmatrix}$$

14.4.2 Pivot and free variables

Recall that matrix multiplication results in a linear combination of columns. Using the RREF in Example 14.3 we have

$$\begin{bmatrix} \textcircled{1} & 2 & 0 & 5 \\ 0 & 0 & \textcircled{1} & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = x_1 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + x_2 \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix} + x_3 \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + x_4 \begin{bmatrix} 5 \\ 3 \\ 0 \end{bmatrix} +$$

x_1 and x_3 multiply pivot columns, so they are called pivot variables. x_2 and x_4 multiply free columns, so they are called free variables.

14.5 Column Space of a Matrix

The column space of a matrix is the set of all linear combinations of the columns.

To shorten the phrase 'all linear combinations', we will say it is the span of the columns. In general we have the following definition.

Definition. The span of the vectors $\mathbf{v_1}, \ldots, \mathbf{v_n}$ is defined as the set of all linear combinations of the vectors. That is,

$$\text{The span of } \mathbf{v_1}, \ldots, \mathbf{v_n} = \{c_1\mathbf{v_1} + c_2\mathbf{v_2} + \ldots + c_n\mathbf{v_n}, \text{ where } c_1, c_2, \ldots, c_n \text{ are scalars}\}$$

Important but not hard: make sure you understand why the span of vectors is always a vector space.

Example 14.4. Consider $R = \begin{bmatrix} 1 & 2 & 0 & 5 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix}$.

The column space of R is the set of all vectors of the form

$$x_1 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + x_2 \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix} + x_3 \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + x_4 \begin{bmatrix} 5 \\ 3 \\ 0 \end{bmatrix}$$

Notice that there is some redundancy here: the free columns are clearly linear combinations of the pivot columns. That is

$$\begin{aligned} \text{Column}_2 &= 2 \cdot \text{Column}_1 \\ \text{Column}_4 &= 5 \cdot \text{Column}_1 + 3 \cdot \text{Column}_3 \end{aligned}$$

So the free columns are redundant and the column space is given by the span of just the pivot columns:

$$\text{Column space of } R = \text{Col}(R) = \left\{ x_1 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + x_3 \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \right\}$$

Our conclusion is that the pivot columns of R span the column space of R.

Example 14.5. In this example, we'll see that row reduction does not change the relations between the columns. So the pivot columns of A span the column space of A.

The matrix $A = \begin{bmatrix} 1 & 2 & 0 & 5 \\ 2 & 4 & 1 & 13 \\ 1 & 2 & 1 & 8 \end{bmatrix}$ in Example 14.3 has RREF

$$R = \begin{bmatrix} 1 & 2 & 0 & 5 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Columns 1 and 3 are the pivot columns of R. Note that for both A and R we have the same relations between the columns, i.e., row reduction did not change these relations:

$$\begin{aligned} \text{Column}_2 &= 2 \cdot \text{Column}_1 \\ \text{Column}_4 &= 5 \cdot \text{Column}_1 + 3 \cdot \text{Column}_3 \end{aligned}$$

Therefore, just as with R, the pivot columns in A span its column space. That is,

$$\text{Column space of } A = \text{span}\left\{ \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \right\} = \left\{ c_1 \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} + c_2 \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \right\}$$

14.6 Rank, basis, dimension, independence

This section is going to throw a lot of vocabulary at you. You need to practice it to make it second nature. You should try to see how most of the new words capture ideas we have been using since the beginning of 18.03

First up is the notion of independence. In Examples 14.4 and 14.5 we saw that the free columns were linear combinations of the pivot columns. This meant they were redundant and not needed to generate the column space. We describe this by saying that the free columns are dependent on the pivot columns.

After getting rid of the free columns it is clear that we need all the pivot columns to make the column space. We describe this by saying that the pivot columns are an independent set of vectors.

The formal definition of independence is the following:
Independence. We say that the vectors $\mathbf{v_1}, \mathbf{v_2}, \dots, \mathbf{v_n}$ are independent if none of them can be written as a linear combination of the others

Example 14.6. (a) Show that vectors $\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix}, \begin{bmatrix} 7 \\ 8 \\ 9 \end{bmatrix}$ are not independent.

(b) Show that vectors $\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$ are independent.

Solution: (a) Note that

$$\begin{bmatrix}1\\2\\3\end{bmatrix} - 2\begin{bmatrix}4\\5\\6\end{bmatrix} + \begin{bmatrix}7\\8\\9\end{bmatrix} = \begin{bmatrix}0\\0\\0\end{bmatrix}$$

This shows that any one of these 3 vectors is a linear combination of the other two. For example

$$\begin{bmatrix}1\\2\\3\end{bmatrix} = 2\begin{bmatrix}4\\5\\6\end{bmatrix} - \begin{bmatrix}7\\8\\9\end{bmatrix}$$

Thus the three vectors are not independent.

(b) One standard way to show independence is to show that the equation

$$c_1\begin{bmatrix}1\\0\\0\end{bmatrix} + c_2\begin{bmatrix}0\\1\\0\end{bmatrix} + c_3\begin{bmatrix}0\\0\\1\end{bmatrix} = \begin{bmatrix}0\\0\\0\end{bmatrix}$$

has only the trivial solution $c_1 = c_2 = c_3 = 0$. For Part (b) this is quite obvious since, summing the left hand side, the equation becomes

$$\begin{bmatrix}c_1\\c_2\\c_3\end{bmatrix} = \begin{bmatrix}0\\0\\0\end{bmatrix}.$$

Returning to pivot columns, we have the following vocabulary:

- The pivot columns are independent.
- The pivot columns span the column space.
- We combine independence and span into one word and say the pivot columns are a basis for the column space.
- The number of pivot columns is called the dimension of the column space.
- We also call the number of pivots the rank of the matrix. This is the same as the dimension of the column space. It is also the same as the number of non-zero rows in the reduced row echelon form.

Let's restate all our definitions in mathematical terms.

- **Independence:** The vectors $\mathbf{v_1}, \dots, \mathbf{v_n}$ are independent if none can be written as a linear combination of the others. Equivalently, they are independent if the equation (with unknowns $c_1, c_2, \dots, c_n$)

  $$c_1\mathbf{v_1} + c_2\mathbf{v_2} + \dots + c_n\mathbf{v_n} = 0$$

  has only the trivial solution $c_1 = c_2 = \dots = c_n = 0$.
- **Span:** The set of all linear combinations of $\mathbf{v_1}, \dots, \mathbf{v_n}$ is called the span of these vectors. It is a vector space, i.e., closed under addition and scalar multiplication.

- **Basis:** A basis for a vector space is a set of vectors that is both independent and spans the vector space. That is, it gets you the entire space with no redundancy.
- **Dimension:** The dimension of a vector space is the number of vectors in a basis of the space.

Example 14.7.

(a) $\begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}0\\1\end{bmatrix}$ clearly span $\mathbf{R}^2$. Since they are also independent, they form a basis of $\mathbf{R}^2$. This particular basis is called the standard basis of $\mathbf{R}^2$.

(b) $\begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}0\\1\end{bmatrix}, \begin{bmatrix}1\\1\end{bmatrix}$ span $\mathbf{R}^2$. Since they are not independent, e.g., $\begin{bmatrix}1\\1\end{bmatrix}$ is a linear combination of the other two vectors, they do not form a basis of $\mathbf{R}^2$.

(c) Since $\mathbf{R}^2$ has a basis with two vectors it has dimension 2.

(d) $\begin{bmatrix}1\\-1\end{bmatrix}, \begin{bmatrix}1\\1\end{bmatrix}$ also form a basis of $\mathbf{R}^2$.

To see this we must show that the two vectors are independent and span $\mathbf{R}^2$. It is clear that they are not multiples of each other so they are independent. To see they span $\mathbf{R}^2$ we need to show that any vector $\begin{bmatrix}b_1\\b_2\end{bmatrix}$ can be written as a linear combination of our two vectors. That is, we must always be able solve

$$x_1\begin{bmatrix}1\\-1\end{bmatrix} + x_2\begin{bmatrix}1\\1\end{bmatrix} = \begin{bmatrix}b_1\\b_2\end{bmatrix}$$

for x_1 and x_2. This is just a matrix equation (linear combination of columns)

$$\begin{bmatrix}1 & 1\\-1 & 1\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix} = \begin{bmatrix}b_1\\b_2\end{bmatrix}$$

We can easily solve this by elimination or using the matrix inverse. so we have verified that $\begin{bmatrix}1\\-1\end{bmatrix}, \begin{bmatrix}1\\1\end{bmatrix}$ is a basis of $\mathbf{R}^2$.

14.7 The meaning of the column space

Consider the matrix equation

$$\begin{bmatrix}1 & 2 & 3\\1 & 0 & 1\end{bmatrix}\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix} = \begin{bmatrix}b_1\\b_2\end{bmatrix}$$

An important problem is to find those vectors $\begin{bmatrix}b_1\\b_2\end{bmatrix}$ for which this equation has a solution. To answer this, remember that matrix multiplication gives a linear combination of the columns. That is, the above matrix equation can be written as

$$x_1\begin{bmatrix}1\\1\end{bmatrix} + x_2\begin{bmatrix}2\\0\end{bmatrix} + x_3\begin{bmatrix}3\\1\end{bmatrix} = \begin{bmatrix}b_1\\b_2\end{bmatrix}$$

We see that the solution to our problem is that the equation has a solution precisely when $\begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$ is in the column space of the coefficient matrix.

14.8 Null Space

The null space of a matrix A is the set of all solutions to the homogeneous equation

$$A\mathbf{x} = \mathbf{0}$$

This is exactly the same as what we have often called the homogeneous solution. Mathematicians also use the term kernel as a synonym for null space.

Note. If A has n columns then the null space is a subspace of $\mathbf{R}^n$.

Example 14.8. Find the null space of

$$A = \begin{bmatrix} 1 & 2 & 1 & 3 \\ 2 & 1 & 0 & 2 \end{bmatrix}$$

Solution: The answer will take a lot of space to display all the vectors and matrices. However, you will see when you work problems on your own that this type of problem does not take a long time to work out.

Also, to make a point, we use the augmented matrix and solve $A\mathbf{x} = \mathbf{0}$ by bringing the augmented matrix to reduced row echelon form.

$$\left[\begin{array}{cccc|c} 1 & 2 & 1 & 3 & 0 \\ 2 & 1 & 0 & 2 & 0 \end{array}\right] \xrightarrow{R_2 = R_2 - 2R_1} \left[\begin{array}{cccc|c} 1 & 2 & 1 & 3 & 0 \\ 0 & -3 & -2 & -4 & 0 \end{array}\right] \xrightarrow{R_2 = -\frac{1}{3}\cdot R_2} \left[\begin{array}{cccc|c} 1 & 2 & 1 & 3 & 0 \\ 0 & 1 & 2/3 & 4/3 & 0 \end{array}\right]$$
$$\xrightarrow{R_1 = R_1 - 2R_2} \left[\begin{array}{cccc|c} \textcircled{1} & 0 & -1/3 & 1/3 & 0 \\ 0 & \textcircled{1} & 2/3 & 4/3 & 0 \end{array}\right]$$

The pivots are circled. The first two columns are pivot columns and the last two are free columns. This last augmented matrix represents a system of equations

$$\begin{bmatrix} 1 & 0 & -1/3 & 1/3 \\ 0 & 1 & 2/3 & 4/3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{20}$$

We will finish finding the null space by writing these equations out explicitly. (Below, we'll show a more efficient way of presenting the computation.)

Written out as a system of equations, Equation 20 is

$$\begin{aligned} x_1 \qquad - \frac{1}{3}x_3 + \frac{1}{3}x_4 &= 0 \\ x_2 + \frac{2}{3}x_3 + \frac{4}{3}x_4 &= 0 \end{aligned} \tag{21}$$

We can solve for the pivot variables x_1, x_2 in terms of the free variables x_3, x_4:

$$\begin{aligned} x_1 &= \frac{1}{3}x_3 - \frac{1}{3}x_4 \\ x_2 &= -\frac{2}{3}x_3 - \frac{4}{3}x_4 \end{aligned}$$

These equations make it clear that we can set the free variables, x_3, x_4, to any values we choose, i.e., they can be set freely. Once they are set, the pivot variables, x_1, x_2, are fully determined.

So let's set the free variables: $x_3 = a$, $x_4 = b$. With these choices the solution to our equations is

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} \frac{1}{3}a - \frac{1}{3}b \\ -\frac{2}{3}a - \frac{4}{3}b \\ a \\ b \end{bmatrix}.$$

This can be written naturally as a linear combination

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = a \begin{bmatrix} 1/3 \\ 2/3 \\ 1 \\ 0 \end{bmatrix} + b \begin{bmatrix} -1/3 \\ -4/3 \\ 0 \\ 1 \end{bmatrix},$$

This shows that Null(A) is 2 dimensional with a basis

$$\left\{ \begin{bmatrix} 1/3 \\ 2/3 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -1/3 \\ -4/3 \\ 0 \\ 1 \end{bmatrix} \right\} \tag{22}$$

Notice that the first basis vector has $x_3 = 1$, $x_4 = 0$. Likewise, the second has $x_3 = 0$, $x_4 = 1$.

In the calculation we just did, x_1, x_2 are pivot variables and x_3, x_4 are free variables. They are called free variables because we could choose their values freely. After that, the pivot variables' values were determined by the Equations 21.

Now, we will show a somewhat more efficient way to present this. The key is to view matrix multiplication as a linear combination of the columns

$$\begin{bmatrix} 1 & 0 & -1/3 & 1/3 \\ 0 & 1 & 2/3 & 4/3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = x_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + x_2 \begin{bmatrix} 0 \\ 1 \end{bmatrix} + x_3 \begin{bmatrix} -1/3 \\ 2/3 \end{bmatrix} + x_4 \begin{bmatrix} 1/3 \\ 4/3 \end{bmatrix} = \mathbf{0}$$

We rewrite this by putting the variables below the columns they multiply:

$$\begin{array}{c} \begin{bmatrix} 1 & 0 & -1/3 & 1/3 \\ 0 & 1 & 2/3 & 4/3 \end{bmatrix} \\ \begin{matrix} x_1 & x_2 & x_3 & x_4 \end{matrix} \end{array}$$

Then, we find a basis of the null space as follows.

1. Set one free variable to 1 and the other free variables to 0, i.e., write a 1 below one free column and 0s below the other free columns.

2. By inspection choose the values of the pivot variables that make the linear combination of the columns add to 0. Write these values below the pivot columns.

$$\begin{array}{cccc} \left[\begin{matrix} 1 & 0 & -1/3 & 1/3 \\ 0 & 1 & 2/3 & 4/3 \end{matrix}\right] \end{array}$$
$$\begin{array}{cccc} x_1 & x_2 & x_3 & x_4 \\ 1/3 & -2/3 & 1 & 0 \\ -1/3 & -4/3 & 0 & 1 \end{array}$$

So a basis of Null(A) contains the two vectors

$$\mathbf{v_1} = \begin{bmatrix} 1/3 \\ -2/3 \\ 1 \\ 0 \end{bmatrix}, \qquad \mathbf{v_2} = \begin{bmatrix} -1/3 \\ -4/3 \\ 0 \\ 1 \end{bmatrix}.$$

Now we get the null space of A (all homogeneous solutions) by taking linear combinations of our two basic solutions.

$$\text{Null}(A) = \{c_1\mathbf{v_1} + c_2\mathbf{v_2}\} = \left\{ c_1 \begin{bmatrix} 1/3 \\ -2/3 \\ 1 \\ 0 \end{bmatrix} + c_2 \begin{bmatrix} -1/3 \\ -4/3 \\ 0 \\ 1 \end{bmatrix} \right\}$$

Of course, this is the same answer we got before.

Remarks. **1.** The dimension of the null space equals the number of free variables.

2. We didn't really need to augment the matrix with a column of zeros, since these zeros never changed.

3. For the equation $A\mathbf{x} = \mathbf{b}$, our general approach will be to find a particular solution and the general homogeneous solution. Then we'll use superposition to get the general solution. This should be very familiar based on what we did with constant coefficient homogeneous DEs.

4. There are of course many other bases of the null space, this is the one we are lead to by our algorithm.

Example 14.9. Consider the matrix

$$R = \begin{bmatrix} 1 & 2 & 0 & 3 & 0 & 4 \\ 0 & 0 & 1 & 5 & 0 & 6 \\ 0 & 0 & 0 & 0 & 1 & 7 \end{bmatrix}.$$

This is in row echelon form. Find its null space two ways.

(i) Using our algorithm of setting each free variable, in turn, to 1, find a basis of Null(R). Write the computation below the matrix.

(ii) Explicitly write out the 3 equations in 6 unknowns and solve them.

Finally, note that they produce exactly the same results and convince yourself that they are really identical methods.

Solution: (i) The algorithm to produce a basis of Null(R) says to set, in turn, each free variable to 1 while setting the others to 0. We start by writing the variables below their respective columns. (This reflects the fact that $R\mathbf{x}$ is a linear combination of the columns of R.) So $R\mathbf{x}$ is represented by

$$\begin{array}{cccccc} \multicolumn{6}{c}{\left[\begin{array}{cccccc} 1 & 2 & 0 & 3 & 0 & 4 \\ 0 & 0 & 1 & 5 & 0 & 6 \\ 0 & 0 & 0 & 0 & 1 & 7 \end{array}\right]} \\ x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{array}$$

There are 3 free variables, so the null space has dimension 3. We can compute the basis vectors by first setting the free variables and then computing the pivot variables that make the linear combination 0. Here is the computation:

$$\begin{array}{cccccc} \multicolumn{6}{c}{\left[\begin{array}{cccccc} 1 & 2 & 0 & 3 & 0 & 4 \\ 0 & 0 & 1 & 5 & 0 & 6 \\ 0 & 0 & 0 & 0 & 1 & 7 \end{array}\right]} \\ x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \\ -2 & 1 & 0 & 0 & 0 & 0 \\ -3 & 0 & -5 & 1 & 0 & 0 \\ -4 & 0 & -6 & 0 & -7 & 1 \end{array}$$

In the first row below the variables, we set $x_2 = 1$, $x_4 = 0$, $x_6 = 0$. Then, by inspection we found the values of x_1, x_3, x_5 that made the linear combination of the columns equal 0. In this case, the 1 times Column 2 had to be canceled by -2 times Column 1.

Likewise, in the second row below the variables, we set $x_2 = 0$, $x_4 = 1$, $x_6 = 0$. Then, by inspection, we saw that the 3 and 5 in Column 4, had to be canceled by -3 times Column 1 plus -5 times Column 3.

The three rows below the variables represent a basis of Null(R):

$$\left\{ \begin{bmatrix} -2 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -3 \\ 0 \\ -5 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -4 \\ 0 \\ -6 \\ 0 \\ -7 \\ 1 \end{bmatrix} \right\} \tag{23}$$

(ii) The matrix equation we want to solve is

$$R\mathbf{x} = 0 \Leftrightarrow \begin{bmatrix} 1 & 2 & 0 & 3 & 0 & 4 \\ 0 & 0 & 1 & 5 & 0 & 6 \\ 0 & 0 & 0 & 0 & 1 & 7 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Writing these out explicitly as a system of equations:

$$\begin{array}{ccccccccccccc}
x_1 & + & 2x_2 & & & + & 3x_4 & & & + & 4x_6 & = & 0\\
 & & & & x_3 & + & 5x_4 & + & & + & 6x_6 & = & 0\\
 & & & & & & & & x_5 & + & 7x_6 & = & 0
\end{array}$$

Next, solve for the pivot variables x_1, x_3 and x_5 in terms of the free variables x_2, x_4, x_6:

$$\begin{aligned}
x_1 &= -2x_2 - 3x_4 - 4x_6\\
x_3 &= -5x_4 - 6x_6\\
x_5 &= -7x_6
\end{aligned}$$

Set the free variables freely: $x_2 = a$, $x_4 = b$, $x_6 = c$. With these choices, the solution to our equation $R\mathbf{x} = \mathbf{0}$ is

$$\begin{bmatrix} x_1\\ x_2\\ x_3\\ x_4\\ x_5\\ x_6 \end{bmatrix} = \begin{bmatrix} -2a-3b-4c\\ a\\ -5b-6c\\ b\\ -7c\\ c \end{bmatrix}.$$

This can be written naturally as a linear combination

$$\begin{bmatrix} x_1\\ x_2\\ x_3\\ x_4\\ x_5\\ x_6 \end{bmatrix} = a\begin{bmatrix} -2\\ 1\\ 0\\ 0\\ 0\\ 0 \end{bmatrix} + b\begin{bmatrix} -3\\ 0\\ -5\\ 1\\ 0\\ 0 \end{bmatrix} + c\begin{bmatrix} -4\\ 0\\ -6\\ 0\\ -7\\ 1 \end{bmatrix}.$$

These vectors are exactly the same as our basis vectors in (23).

The two methods produce exactly the same basis because both involve setting the free variables freely and then computing the pivot variables. In (i), we started by setting one free variable to 1 and the others to 0 to get a basis vector. In (ii), we first found the general solution. Then, the basis vectors were found by setting one free variable to 1 and the others to 0, e.g., setting $a = 0$, $b = 1$, $c = 0$ gives the second basis vector in (23).

14.8.1 Connecting the RREF and the original matrix

The last piece of the puzzle is to connect the null space and column space of a matrix with those of its reduced row echelon form. Let's look again at Example 14.8 and place A and its RREF one above the other

$$\begin{aligned}
A &= \begin{bmatrix} 1 & 2 & 1 & 3\\ 2 & 1 & 0 & 2 \end{bmatrix}\\
R &= \begin{bmatrix} 1 & 0 & -1/3 & 1/3\\ 0 & 1 & 2/3 & 4/3 \end{bmatrix}
\end{aligned}$$

Here are the rules.

1. The null space of A is the same as that of R.

2. The column space of R has a basis given by the pivot columns of R. The corresponding columns in A are a basis for the column space of A.

Rule 1 follows because row reduction of the augmented matrix does not alter the solutions to an equation.

Rule 2 follows because row reduction does not change the relationships between the columns.

14.9 Summary of $A\mathbf{x} = \mathbf{b}$

We are now very good at analyzing the equation

$$A\mathbf{x} = \mathbf{b}$$

1. It has a solution if $\mathbf{b}$ is in the column space of A.
2. We can find a solution by elimination (aka row reduction)
3. The full solution is $\mathbf{x} = \mathbf{x_p} + \mathbf{x_h}$, where $\mathbf{x_p}$ is any particular solution and $\mathbf{x_h}$ is the general homogeneous solution. That is $\mathbf{x_h}$ is the null space of A.
4. We can use reduced row echelon form (RREF) to find a basis and the dimension of both the null space and the column space.

14.10 Matrix multiplication as a linear transformation

It can be very useful to think of matrix multiplication as a function. We'll also say map or linear transform.

Example 14.10. Let $A = \begin{bmatrix} 1 & 2 & 1 & 3 \\ 2 & 1 & 0 & 2 \end{bmatrix}$. A is a 2×4 matrix so we can multiply it times a 4-vector and get a 2-vector

$$\begin{bmatrix} 1 & 2 & 1 & 3 \\ 2 & 1 & 0 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}.$$

This is a function from $\mathbf{R}^4$ to $\mathbf{R}^2$. We will write

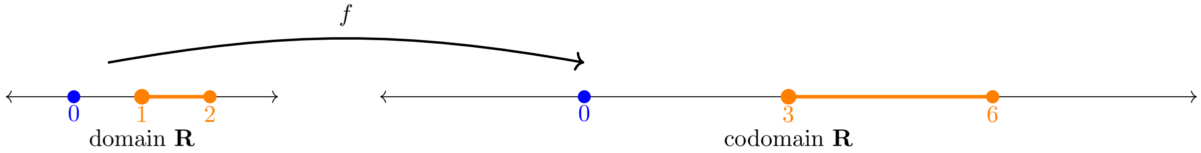
$$\begin{aligned} A : \mathbf{R}^4 &\longrightarrow \mathbf{R}^2 \\ \mathbf{x} &\longmapsto \begin{bmatrix} 1 & 2 & 1 & 3 \\ 2 & 1 & 0 & 2 \end{bmatrix} \mathbf{x} \end{aligned}$$

and say A maps $\mathbf{R}^4$ to $\mathbf{R}^2$. (More precisely, multiplication by A maps $\mathbf{R}^4$ to $\mathbf{R}^2$.) This is a simple statement, but it is a fruitful way to think about matrix multiplication and will help us understand many things.

14.10.1 Depicting linear transformations with a domain-codomain diagram

Example 14.11. To visualize the function $f(x) = 3x$, we can draw its graph in $\mathbf{R}^2$, i.e., the line $y = 3x$.

But there is another way: Draw the domain and codomain (two copies of the real line) and show where certain features in the domain get mapped (or transformed) to:

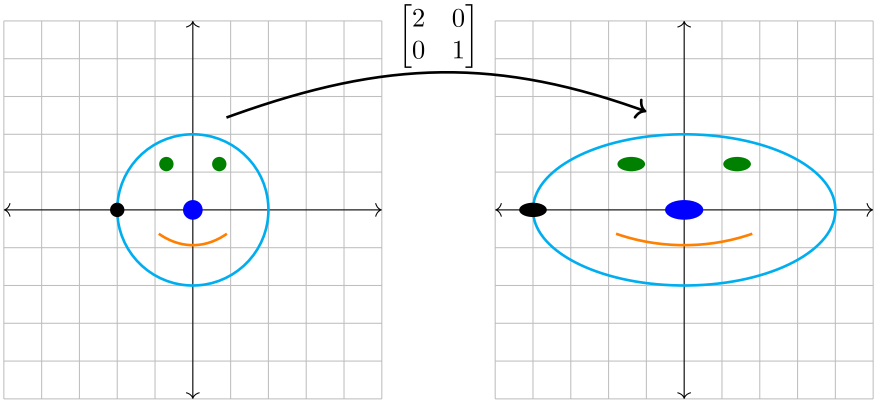


For example, $f(x) = 3x$ maps the point 2 to the point 6 and the interval $[1, 2]$ to the interval $[3, 6]$. The diagram shows how f expands everything by a factor of 3.

Example 14.12. Now consider the matrix $\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$ and the associated linear transformation

$$\mathbf{f} : \mathbf{R}^2 \longrightarrow \mathbf{R}^2$$
$$\begin{bmatrix} x \\ y \end{bmatrix} \longmapsto \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2x \\ y \end{bmatrix}.$$

Drawing a graph of $\mathbf{f}$ would require 4 dimensions (2 for the input and 2 for the output), so let's draw a domain-codomain diagram instead. How does $\mathbf{f}$ transform Poonen's van Gogh unit smile?

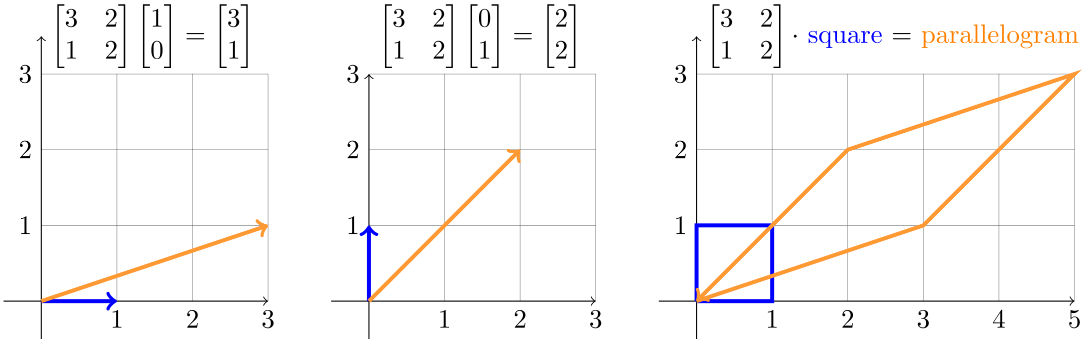


For example, the ear at $\begin{bmatrix} -1 \\ 0 \end{bmatrix}$ is mapped to $\begin{bmatrix} -2 \\ 0 \end{bmatrix}$. Notice how the linear transformation $\mathbf{f}$ stretches the smiley in the horizontal direction only.

Example 14.13. Let $A = \begin{bmatrix} 3 & 2 \\ 1 & 2 \end{bmatrix}$. For a square matrix, we can save space by putting the domain and codomain in the same plane. For multiplication by A, we have:

$$A \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \end{bmatrix} \qquad \text{and} \qquad A \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

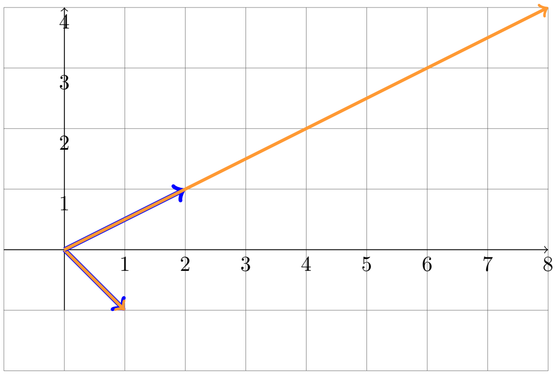
We say $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ is mapped to $\begin{bmatrix} 3 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$ is mapped to $\begin{bmatrix} 2 \\ 2 \end{bmatrix}$. The figures below display the input vectors in blue and the output vectors in orange. They show that the effect of multiplying a vector by A is to both rotate and scale the input vector. Geometrically the effect of multiplying a square by A is a parallelogram.



Matrix multiplication rotates and scales vectors

As a quick look ahead, we note that most vectors are rotated and scaled, however there are some special vectors that are scaled but not rotated:

$$\begin{bmatrix}3&2\\1&2\end{bmatrix}\begin{bmatrix}1\\-1\end{bmatrix}=\begin{bmatrix}1\\-1\end{bmatrix}\qquad\text{and}\qquad\begin{bmatrix}3&2\\1&2\end{bmatrix}\begin{bmatrix}2\\1\end{bmatrix}=\begin{bmatrix}8\\4\end{bmatrix}=4\begin{bmatrix}2\\1\end{bmatrix}.$$



Special vectors that are not rotated when multiplied by A.

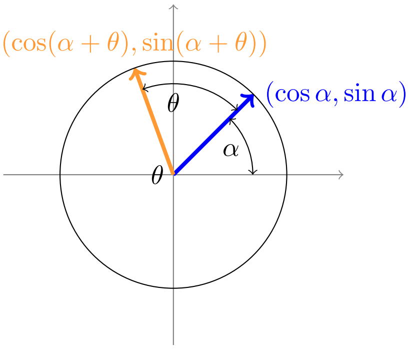
We'll spend a lot time with these special vectors soon. For now let's note the following consequence of linearity: For $A=\begin{bmatrix}3&2\\1&2\end{bmatrix}$.

$$A\left(c_1\begin{bmatrix}1\\-1\end{bmatrix}+c_2\begin{bmatrix}2\\1\end{bmatrix}\right)=c_1\begin{bmatrix}1\\-1\end{bmatrix}+4c_2\begin{bmatrix}2\\1\end{bmatrix}.$$

That's pretty simple!

Example 14.14. Rotation matrices

In this example we'll show that the matrix $R_\theta=\begin{bmatrix}\cos\theta&-\sin\theta\\\sin\theta&\cos\theta\end{bmatrix}$ rotates vectors by an angle θ. To see this we take a unit vector at angle α and see what multiplication by R_θ does to it.


$(\cos(\alpha+\theta), \sin(\alpha+\theta))$
$(\cos\alpha, \sin\alpha)$
θ
α
0

$$R_\theta \begin{bmatrix} \cos\alpha \\ \sin\alpha \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \cos\alpha \\ \sin\alpha \end{bmatrix} = \begin{bmatrix} \cos\theta\cos\alpha - \sin\theta\sin\alpha \\ \sin\theta\cos\alpha + \cos\theta\sin\alpha \end{bmatrix}$$
$$= \begin{bmatrix} \cos(\alpha+\theta) \\ \sin(\alpha+\theta) \end{bmatrix} \qquad \text{(trig addition formula!)}$$

The result is a unit vector at angle $\alpha + \theta$, which is what we claimed would happen.

R_θ is called a rotation matrix. We will also use the name orthogonal matrix.

The mathlet `https://mathlets.org/mathlets/matrix-vector/` illustrates matrix multiplication as a mapping of .

15 Transpose, inverse, determinant

15.1 Goals

1. Know the definition and be able to compute the inverse of any *square* matrix using row operations.
2. Know the properties of inverses. In particular, that $\det(A) \neq 0$ is equivalent to the existence of A^{-1}.
3. Know the definition and be able to compute the determinnant of any *square* matrix.
4. Know the properties of determinant.
5. Know the definition and be able to compute the transpose of any matrix.
6. Understand how elementary row operations affect the determinant and be able to use this to simplify computing determinants.
7. Know the definition of diagonal and triangular matrices and be able to easily compute their determinants and, for diagonal matrices, inverses.
8. Recall from 18.02 the method of Laplace expansion for computing inverses and determinants.

15.2 Introduction

The main point of this topic is to learn how to compute determinants (of square matrices). The main application is that the determinant is 0 exactly when the matrix has a nontrivial null space. This will be key when we learn about eigenvalues and eigenvectors.

You learned how to compute determinants in 18.02. We'll recall the methods learned there and add another method based on row reduction. This will simplify the sometimes tedious calculations. We will do something similar with inverses.

15.3 Inverses of square matrices

You saw matrix inverses in 18.02, so we will assume they are somewhat familiar.

Definition: A square matrix A has an inverse if there is another matrix, denoted A^{-1}, such that $A^{-1}A = AA^{-1} = I$.

Property. $(AB)^{-1} = B^{-1}A^{-1}$.

Proof: It's easy to check that $(B^{-1}A^{-1})(AB) = B^{-1}(A^{-1}A)B = B^{-1}IB = I$.

Story. If you put on your sweater and then your jacket, to reverse it you have to first take off your jacket and then your sweater.

Let A be an $n \times n$ matrix. Two important questions are
Q1. Is A invertible? That is, does A have an inverse?

Q2. How do you compute A^{-1}?

We can often answer Question 1 using the following list of equivalent statements.

1. A has an inverse.
2. $\det(A) \neq 0$.
3. A has a trivial null space,i.e., the null space $= \{\mathbf{0}\}$.
4. A has rank n (we say A has full rank).
5. The columns of A are independent.
6. The echelon form of A has n pivots.
7. The reduced row echelon form of A is the identity.
8. For every choice of $\mathbf{b}$, the equation $A\mathbf{x} = \mathbf{b}$ has a unique solution. That solution is $\mathbf{x} = A^{-1}\mathbf{b}$.

Proofs: We'll give brief arguments why numbers 2-8 follow from 1. The proof of the converses, i.e., that number 1 follows from each of 2-8 are similar. So assume that A has an inverse A^{-1}.

2. We'll see below (and you saw in 18.02) that in computing A^{-1} we divide by $\det(A)$. Since we can't divide by 0, we must have $\det(A) \neq 0$.

3. Suppose $\mathbf{v}$ is in the null space of A, so $A\mathbf{v} = \mathbf{0}$. Then, since A has an inverse, we know $\mathbf{v} = A^{-1}\mathbf{0} = \mathbf{0}$. This shows that the only vector in the null space is $\mathbf{0}$.

4. We just showed that 1 implies A has a trivial null space. Thus it has no free variables. This implies it has n pivot variables, i.e., it has rank n.

5,6,7. These are just different ways of saying A has rank n.

8. This is obvious.

It's also worth recording these equivalences in inverse form. The following are equivalent for A

1. A does not have an inverse. (We say A is singular or non-invertible.)
2. $\det(A) = 0$.
3. A has a nontrivial null space,i.e., the null space contains non-zero vectors.
4. A has rank less than n.
5. The columns of A have some dependencies
6. The echelon form of A has fewer than n pivots.
7. The RREF of A has some all 0 rows
8. For every choice of $\mathbf{b}$, the equation $A\mathbf{x} = \mathbf{b}$ has either no solutions or infinitely many solutions.

15.3.1 Matlab

Matlab makes it easy to compute the inverse of a matrix A. The function `inv(A)` returns A^{-1}. For example, to solve $A\mathbf{x} = \mathbf{b}$ in Matlab you would give the command: `x = inv(A)*b`.

15.3.2 Computing inverses.

There are a number of methods for computing the inverse of a matrix. First we remind you of the inverse of a 2 by 2 matrix:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

In words, swap the main diagonal elements, change the sign (without swapping) of the off diagonal elements, and divide by the determinant. You should memorize this. We will use it often and you won't want to waste time deducing it in each case.

Next, we will show how to find and inverse using elimination. A few examples will illustrate how to do this. The reason it works is straightforward, but we will relegate the explanation to an optional appendix at the end of these notes.

Example 15.1. Find the inverse of $A = \begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix}$.

Solution: We augment A by the identity and then use row reduction to bring the left-hand side to the identity.

$$\left[\begin{array}{cc|cc} 6 & 5 & 1 & 0\\ 1 & 2 & 0 & 1\end{array}\right] \xrightarrow{\text{swap } R_1 \text{ and } R_2} \left[\begin{array}{cc|cc} 1 & 2 & 0 & 1\\ 6 & 5 & 1 & 0\end{array}\right] \xrightarrow{R_2 = R_2 - 6\cdot R_1} \left[\begin{array}{cc|cc} 1 & 2 & 0 & 1\\ 0 & -7 & 1 & -6\end{array}\right]$$

$$\xrightarrow{R_2 = (-1/7)\cdot R_2} \left[\begin{array}{cc|cc} 1 & 2 & 0 & 1\\ 0 & 1 & -1/7 & 6/7\end{array}\right] \xrightarrow{R_1 = R_1 - 2\cdot R_2} \left[\begin{array}{cc|cc} 1 & 0 & 2/7 & -5/7\\ 0 & 1 & -1/7 & 6/7\end{array}\right]$$

The right half of the last augmented matrix is the inverse

$$\begin{bmatrix} 6 & 5\\ 1 & 2\end{bmatrix}^{-1} = \begin{bmatrix} 2/7 & -5/7\\ -1/7 & 6/7\end{bmatrix}$$

Example 15.2. Find the inverse of $A = \begin{bmatrix} 1 & 5 & 4\\ 2 & 1 & 3\\ 0 & 0 & 1\end{bmatrix}$

Solution: We augment A by the identity and use row reduction as in the previous example.

$$\left[\begin{array}{ccc|ccc} 1 & 5 & 4 & 1 & 0 & 0\\ 2 & 1 & 3 & 0 & 1 & 0\\ 0 & 0 & 1 & 0 & 0 & 1\end{array}\right] \xrightarrow{\substack{R_1 = R_1 - 4\cdot R_3\\ R_2 = R_2 - 3\cdot R_3}} \left[\begin{array}{ccc|ccc} 1 & 5 & 0 & 1 & 0 & -4\\ 2 & 1 & 0 & 0 & 1 & -3\\ 0 & 0 & 1 & 0 & 0 & 1\end{array}\right] \xrightarrow{R_1 = R_1 - 5\cdot R_2} \left[\begin{array}{ccc|ccc} -9 & 0 & 0 & 1 & -5 & 11\\ 2 & 1 & 0 & 0 & 1 & -3\\ 0 & 0 & 1 & 0 & 0 & 1\end{array}\right]$$

$$\xrightarrow{R_1 = (-1/9)\cdot R_1} \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & -1/9 & 5/9 & -11/9\\ 2 & 1 & 0 & 0 & 1 & -3\\ 0 & 0 & 1 & 0 & 0 & 1\end{array}\right] \xrightarrow{R_2 = R_2 - 2\cdot R_1} \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & -1/9 & 5/9 & -11/9\\ 0 & 1 & 0 & 2/9 & -1/9 & -5/9\\ 0 & 0 & 1 & 0 & 0 & 1\end{array}\right]$$

So, $A^{-1} = \dfrac{1}{9}\begin{bmatrix} -1 & 5 & -11\\ 2 & -1 & -5\\ 0 & 0 & 9\end{bmatrix}$, as you can easily verify.

Example 15.3. Let's see what happens if we try this on a matrix that doesn't have an inverse. Try to find the inverse of $A = \begin{bmatrix} 1 & 2 & 3\\ 4 & 5 & 6\\ 7 & 8 & 9\end{bmatrix}$:

$$\left[\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 0 & 0\\ 4 & 5 & 6 & 0 & 1 & 0\\ 7 & 8 & 9 & 0 & 0 & 1\end{array}\right] \xrightarrow{\substack{R_2 = R_2 - 4\cdot R_1\\ R_3 = R_3 - 7\cdot R_1}} \left[\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 0 & 0\\ 0 & -3 & -6 & -4 & 1 & 0\\ 0 & -6 & -12 & -7 & 0 & 1\end{array}\right] \xrightarrow{R_3 = R_3 - 2\cdot R_2} \left[\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 0 & 0\\ 0 & -3 & -6 & -4 & 1 & 0\\ 0 & 0 & 0 & 1 & -2 & 1\end{array}\right]$$

We've reached an impasse. The matrix A only has 2 pivots, so it cannot be row reduced to the identity, i.e., it has no inverse.

Question: What is the $\det(A)$?

15.3.3 Diagonal and triangular matrices

It is simple to find the inverse of a diagonal matrix. Here are some examples.

$$\begin{bmatrix} 3 & 0\\ 0 & 5\end{bmatrix}^{-1} = \begin{bmatrix} 1/3 & 0\\ 0 & 1/5\end{bmatrix} \qquad \begin{bmatrix} 1 & 0 & 0\\ 0 & 2 & 0\\ 0 & 0 & 3\end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 & 0\\ 0 & 1/2 & 0\\ 0 & 0 & 1/3\end{bmatrix}$$

$$\begin{bmatrix} a & 0 & 0 & 0 \\ 0 & b & 0 & 0 \\ 0 & 0 & c & 0 \\ 0 & 0 & 0 & d \end{bmatrix}^{-1} = \begin{bmatrix} a^{-1} & 0 & 0 & 0 \\ 0 & b^{-1} & 0 & 0 \\ 0 & 0 & c^{-1} & 0 \\ 0 & 0 & 0 & d^{-1} \end{bmatrix}$$

Triangular matrices require more work, but at least we only have to do elimination in one direction.

Example 15.4. Find the inverse of $A = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 3 & 0 \\ 4 & 5 & 6 \end{bmatrix}$

Solution: We augment and row reduce from the top down:

$$\left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 0 & 0 \\ 2 & 3 & 0 & 0 & 1 & 0 \\ 4 & 5 & 6 & 0 & 0 & 1 \end{array}\right] \xrightarrow{\substack{R_2 = R_2 - 2 \cdot R_1 \\ R_3 = R_3 - 4 \cdot R_1}} \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 3 & 0 & -2 & 1 & 0 \\ 0 & 5 & 6 & -4 & 0 & 1 \end{array}\right] \xrightarrow{R_2 = (1/3) \cdot R_2} \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & -2/3 & 1/3 & 0 \\ 0 & 5 & 6 & -4 & 0 & 1 \end{array}\right]$$

$$\xrightarrow{R_3 = R_3 - 5 \cdot R_2} \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & -2/3 & 1/3 & 0 \\ 0 & 0 & 6 & -2/3 & -5/3 & 1 \end{array}\right] \xrightarrow{R_3 = (1/6) \cdot R_3} \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & -2/3 & 1/3 & 0 \\ 0 & 0 & 1 & -1/9 & -5/18 & 1/6 \end{array}\right]$$

So, $A^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ -2/3 & 1/3 & 0 \\ -1/9 & -5/18 & 1/6 \end{bmatrix}$.

15.3.4 Laplace expansion using cofactors

In 18.02 you learned how to find the inverse using cofactors (also called the adjoint method). For completeness, we review this method in the appendix at the end of these notes. Unless we specify a method for finding an inverse, e.g., by row reduction, you may use any method you want, including Laplace expansion.

15.4 Determinants

We can take the determinant of a square matrix A. We will write $\det(A)$ or $|A|$ for the determinant of A. Since this is part of 18.02, we will assume you have seen determinants before. For us, the most important use of determinants is to check if a matrix has a trivial null space. These were Properties 2 and 3 in Section 15.3:

$$\text{Null}(A) \text{ is trivial if and only if } \det(A) \neq 0.$$
$$\text{Null}(A) \text{ is nontrivial if and only if } \det(A) = 0.$$

Properties of determinants:
1. The determinant is linear in each column and linear in each row.
2. $\det(I) = 1$.
3. Swapping rows changes the sign of the determinant.
4. Scaling a row scales the determinant.
5. Adding a multiple of one row to another doesn't change the determinant.
6. $\det(AB) = \det(A)\det(B)$.

15.4.1 Laplace expansion using minors

In 18.02 you learned how to find the determinant using minors. We give a review of that method in the appendix at the end of the notes for this topic. Unless we specify a method for finding the determinant, e.g., by row reduction, you may use any method you want including Laplace expansion.

15.4.2 The 2 by 2 case

You should know the determinant of a 2 × 2 matrix

$$\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc.$$

15.4.3 Easy determinants

The easiest determinants to compute are for diagonal and triangular matrices. In these cases the determinant is just the product of the diagonal entries.

$$\textbf{Diagonal:} \ \det \begin{bmatrix} a & 0 & 0 & 0 \\ 0 & b & 0 & 0 \\ 0 & 0 & c & 0 \\ 0 & 0 & 0 & d \end{bmatrix} = abcd.$$

$$\textbf{Upper triangular:} \ \det \begin{bmatrix} a & e & f & g \\ 0 & b & h & i \\ 0 & 0 & c & j \\ 0 & 0 & 0 & d \end{bmatrix} = abcd$$

$$\textbf{Lower triangular:} \ \det \begin{bmatrix} a & 0 & 0 & 0 \\ b & c & 0 & 0 \\ d & e & f & 0 \\ g & h & i & j \end{bmatrix} = acfj.$$

Identical rows: If A has two identical rows then $\det A = 0$.

Proof: Swapping the rows leaves A and therefore, $\det(A)$ unchanged. But (property 3), it also changes the sign of the determinant. Only 0 stays the same when you change sign. Therefore, $\det(A) = 0$.

15.4.4 Matlab

Matlab makes it easy to compute the determinant of a matrix A. The function `det(A)` returns $\det(A)$.

15.4.5 Finding the determinant using row reduction

Since we know how the elementary row operations affect the determinant we can use row reduction to compute the determinant of a matrix. We'll illustrate with an example.

Example 15.5. Find the determinant of $A = \begin{bmatrix}0 & 4 & 1\\ 1 & 2 & 2\\ 3 & 1 & 2\end{bmatrix}$

Solution: We use row reduction until A is in triangular form. At each step we keep track of the effect on the determinant.

$$\begin{bmatrix}0 & 4 & 1\\ 1 & 2 & 2\\ 3 & 1 & 2\end{bmatrix} \xrightarrow{\text{swap rows; det} \times (-1)} \begin{bmatrix}1 & 2 & 2\\ 0 & 4 & 1\\ 3 & 1 & 2\end{bmatrix} \xrightarrow{R_3 = R_3 - 3 \cdot R_1;\ \text{det unchanged}} \begin{bmatrix}1 & 2 & 2\\ 0 & 4 & 1\\ 0 & -5 & -4\end{bmatrix}$$

$$\xrightarrow{R_2 = (1/4) \cdot R_2;\ \text{det} \times (1/4)} \begin{bmatrix}1 & 2 & 2\\ 0 & 1 & 1/4\\ 0 & -5 & -4\end{bmatrix} \xrightarrow{R_3 = R_3 + 5 \cdot R_2;\ \text{det unchanged}} \begin{bmatrix}1 & 2 & 2\\ 0 & 1 & 1/4\\ 0 & 0 & -11/4\end{bmatrix}$$

The last matrix is triangular, so its determinant is the product of its diagonal entries, i.e., $-11/4$. Following the changes in the determinant caused by the row operations, we have

$$(-1) \cdot \left(\frac{1}{4}\right) \det(A) = \det \begin{bmatrix}1 & 2 & 2\\ 0 & 1 & 1/4\\ 0 & 0 & -11/4\end{bmatrix} = \frac{-11}{4} \Rightarrow \det(A) = 11.$$

15.5 Transpose

For us, the transpose will be a convenient tool for calculation and presenting matrices. For example, in Matlab we can use the transpose and matrix multiplication to compute dot (inner) products. There is a lot more to transposes than we will see. You should take 18.06 to learn more.

To take the transpose of a matrix you change rows into columns. We'll use the notation A^T for the transpose of A.

Example 15.6. If $A = \begin{bmatrix}1 & 2\\ 3 & 4\\ 5 & 6\end{bmatrix}$ then $A^T = \begin{bmatrix}1 & 3 & 5\\ 2 & 4 & 6\end{bmatrix}$.

Note. Transpose turns a 3×2 matrix into a 2×3 matrix. In general, it turns an $n \times m$ matrix into an $m \times n$ matrix.

In terms of entries, the i, j entry of A^T equals the j, i entry of A. In symbols: $(A^T)_{i,j} = A_{j,i}$.

You can check that the dimensions make sense: If A is $m \times n$ and B is $n \times p$ then AB is $m \times p$, so $(AB)^T$ is $p \times m$. Likewise, we can show that $B^T A^T$ is $p \times m$.

Of course we can prove this property, but in 18.03 we are not particularly concerned with the proof, so, for anyone who is interested, we'll put it in the appendix at the end of the notes for this topic.

Symmetric matrices. A square matrix A is symmetric if $A = A^T$.

Example 15.7. The following matrices are symmetric

$$\begin{bmatrix}2 & 3\\ 3 & 5\end{bmatrix} \quad \begin{bmatrix}1 & 2 & 3\\ 2 & 4 & 5\\ 3 & 5 & 6\end{bmatrix} \quad \begin{bmatrix}a & b & c\\ b & d & e\\ c & e & f\end{bmatrix} \quad \begin{bmatrix}a & 0 & 0\\ 0 & b & 0\\ 0 & 0 & c\end{bmatrix}.$$

Notes: **1.** Symmetric means symmetric around the main diagonal.
2. Diagonal matrices are always symmetric.
3. It doesn't make sense to ask if a non-square matrix is symmetric.
4. Matlab uses a prime to mean transpose, e.g., `[1,2; 3,4; 5,6]'`.

Symmetric matrices are an extremely important class of matrices, which arise in many applications. Unfortunately, we won't have time to do much with them in 18.03.

15.5.1 Inner products and transposes

In 18.02 you learned about the dot product of two vectors, e.g.,

$$(1, 2, 3) \cdot (2, -1, 4) = 2 - 2 + 12 = 12.$$

Since the dot is also used for multiplication, we are going to (mostly) quit using the dot notation and also rename the dot product as the inner product. Here is our new terminology and notation.

Definition: The inner product of two vectors $\mathbf{v}$ and $\mathbf{w}$ is denoted $\langle \mathbf{v}, \mathbf{w} \rangle$. If $\mathbf{v}$ and $\mathbf{w}$ are column vectors in $\mathbf{R}^3$ then

$$\langle \mathbf{v}, \mathbf{w} \rangle = \left\langle \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}, \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} \right\rangle = v_1 w_1 + v_2 w_2 + v_3 w_3.$$

This is not restricted to vectors in $\mathbf{R}^3$, we can define the inner product between vectors in $\mathbf{R}^n$ for any n.

The inner product of two column vectors can be computed as a matrix multiplication using the transpose.

$$\langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v}^T \mathbf{w} = \begin{bmatrix} v_1 & v_2 & v_3 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} = v_1 w_1 + v_2 w_2 + v_3 w_3.$$

We will not do very much with inner product for now, though it will come up again later. For now, the most important thing to remember is that two vectors are orthogonal if their inner product is 0.

$$\langle \mathbf{v}, \mathbf{w} \rangle = 0 \Leftrightarrow \mathbf{v} \text{ and } \mathbf{w} \text{ are orthogonal.}$$

15.5.2 Saving space

Now that we have the transpose, we can use it to save space on the page. Instead of always writing column vectors vertically, we can use the transpose to write them horizontally, e.g.,

$$\begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \end{bmatrix}^T$$

15.6 Appendix

This appendix contains some review and some more technical material. The technical material is just for your reading pleasure. You will not be asked to reproduce it for ES.1803.

15.6.1 Review: Determinants by Laplace expansion along a row or column

The i, j minor of a matrix is the determinant after removing the ith row and jth column, i.e., the row and column intersecting at the i, j entry.

Example 15.8. Let $A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$. Find all the minors that go with the second column.

Solution: The second column has the (1,2), (2,2) and (3,2) entries. The (1,2) minor is the 2×2 determinant of the matrix after crossing out the row and columns through the (1,2) entry

$$\text{(1,2) minor } = \begin{vmatrix} \cancel{1} & \cancel{2} & \cancel{3} \\ 4 & \cancel{5} & 6 \\ 7 & \cancel{8} & 9 \end{vmatrix} = \begin{vmatrix} 4 & 6 \\ 7 & 9 \end{vmatrix} = -6$$

Likewise, for the other two entries in the second column.

$$\text{(2,2) minor } = \begin{vmatrix} 1 & \cancel{2} & 3 \\ \cancel{4} & \cancel{5} & \cancel{6} \\ 7 & \cancel{8} & 9 \end{vmatrix} = \begin{vmatrix} 1 & 3 \\ 7 & 9 \end{vmatrix} = -12, \qquad \text{(3,2) minor } = \begin{vmatrix} 1 & \cancel{2} & 3 \\ 4 & \cancel{5} & 6 \\ \cancel{7} & \cancel{8} & \cancel{9} \end{vmatrix} = \begin{vmatrix} 1 & 3 \\ 4 & 6 \end{vmatrix} = -6$$

We can compute the determinant of the $n \times n$ matrix A by expanding along any row or column.

$$\det A = \text{sum along the row of (checkerboard sign)} \cdot \text{(entry)} \cdot \text{(minor)}$$

As a formula, expanding along the i^{th} row we have

$$\det A = \sum_{j=1}^{n} (-1)^{i+j} \cdot A_{i,j} \cdot (i, j) \text{ minor.}$$

To expand along a column, you fix j and sum over i.

Example 15.9. Expand $\begin{vmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{vmatrix}$ along the middle column.

Solution: (Long, drawn out version.) First we draw a line through the second column:

$$\begin{bmatrix} 1 & \cancel{2} & 3 \\ 4 & \cancel{5} & 6 \\ 7 & \cancel{8} & 9 \end{bmatrix}$$

Now we use the sign checkerboard $\begin{bmatrix} + & - & + \\ - & + & - \\ + & - & + \end{bmatrix}$ to compute the determinant using the entries and minors along the second column.

$$\begin{vmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{vmatrix} = -2 \overbrace{\begin{vmatrix} 4 & 6 \\ 7 & 9 \end{vmatrix}}^{(1,2)\text{minor}} + 5 \begin{vmatrix} 1 & 3 \\ 7 & 9 \end{vmatrix} - 8 \begin{vmatrix} 1 & 3 \\ 4 & 6 \end{vmatrix} = -2(-6) + 5(-12) - 8(-6) = 0$$

(1,2) matrix entry

(1,2) checkerboard sign

Here, the sign in front of each term comes from the checkerboard, e.g., the (1,2) checkerboard entry is a minus sign, so that term gets a minus sign.

The same process works expanding along any row or column.

Example 15.10. Compute $\begin{vmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{vmatrix}$ along the top row

Solution: (Short, concise version.) Determinant $= 1 \cdot \begin{vmatrix} 5 & 6 \\ 8 & 9 \end{vmatrix} - 2 \cdot \begin{vmatrix} 4 & 6 \\ 7 & 9 \end{vmatrix} + 3 \begin{vmatrix} 4 & 5 \\ 7 & 8 \end{vmatrix} = 0$

Example 15.11. Compute $\begin{vmatrix} 1 & 2 & 3 \\ 5 & 0 & 7 \\ 8 & 0 & 9 \end{vmatrix}$

Solution: Use second column: $\det = -2 \cdot \begin{vmatrix} 5 & 7 \\ 8 & 9 \end{vmatrix} + 0 \cdot * - 0 \cdot * = 22$. (To save time, we didn't bother computing the minors that were multiplied by 0.)

15.6.2 Using row reduction on the augmented matrix to find the inverse

Here we will explain why this technique works. The key fact here is that every elementary row operation corresponds to multiplication by a matrix on the left. We illustrate by row reducing our favorite matrix to the identity.

$$\begin{array}{ll}
\text{Original matrix:} & A = \begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix} \\
\text{Swapping } R_1 \text{ and } R_3\text{:} & \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 6 & 5 \end{bmatrix} \\
R_2 = R_2 - 6R_1 : & \begin{bmatrix} 1 & 0 \\ -6 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 6 & 5 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 0 & -7 \end{bmatrix} \\
R_2 = (-1/7) \cdot R_2 : & \begin{bmatrix} 1 & 0 \\ 0 & -1/7 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 0 & -7 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} \\
R_1 = R_1 - 2 \cdot R_2 & \begin{bmatrix} 1 & -2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}
\end{array}$$

If you put all the matrix multiplications together we get

$$\begin{bmatrix} 1 & -2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1/7 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -6 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

That is

$$\begin{bmatrix} 1 & -2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1/7 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -6 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix}^{-1}$$

This product is exactly what we get by applying the same sequence of elementary row operations to the identity matrix on the right side of the augmented matrix $(A|I)$.

15.6.3 Finding inverses using cofactors (the Laplace or adjoint method)

We have a simple formula for finding the inverse of a 2×2 matrix:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

For bigger (square) matrices finding inverses is more involved. One algorithm for doing this is called the adjoint or Laplace method.

The **step-by-step algorithm** is the following:

1. Start with A.
2. Find the matrix of minors.
3. Find the matrix of cofactors.
4. Find the adjoint.
5. Divide by $\det(A)$.

Of course, we have to explain what each of these things is. We will over the next four examples, explaining one item at a time.

For these examples let $A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 1 & 2 & 0 \end{bmatrix}$.

Matrix of minors. We definied the i, j minor of a matrix in Section 15.6.1. The matrix of minors of A is just the matrix made up of all the minors. The i, j-entry of the matrix of minors is the i, j-minor of A.

Example 15.12. Find the matrix of minors of A.

Solution: A is a 3×3 matrix so its matrix of minors is also 3×3. Here is the computation for each minor:

$1, 1$ minor: $\begin{bmatrix} \cancel{1} & \cancel{2} & \cancel{3} \\ \cancel{4} & 5 & 6 \\ \cancel{1} & 2 & 0 \end{bmatrix}$; $1, 1$-minor $= \begin{vmatrix} 5 & 6 \\ 2 & 0 \end{vmatrix} = -12$; matrix of minors $= \begin{bmatrix} -12 & * & * \\ * & * & * \\ * & * & * \end{bmatrix}$

$1, 2$ minor: $\begin{bmatrix} \cancel{1} & \cancel{2} & \cancel{3} \\ 4 & \cancel{5} & 6 \\ 1 & \cancel{2} & 0 \end{bmatrix}$; $1, 2$-minor $= \begin{vmatrix} 4 & 6 \\ 1 & 0 \end{vmatrix} = -6$; matrix of minors $= \begin{bmatrix} -12 & -6 & * \\ * & * & * \\ * & * & * \end{bmatrix}$

$1, 3$ minor: $\begin{bmatrix} \cancel{1} & \cancel{2} & \cancel{3} \\ 4 & 5 & \cancel{6} \\ 1 & 2 & \cancel{0} \end{bmatrix}$; $1, 3$-minor $= \begin{vmatrix} 4 & 5 \\ 1 & 2 \end{vmatrix} = 3$; matrix of minors $= \begin{bmatrix} -12 & -6 & 3 \\ * & * & * \\ * & * & * \end{bmatrix}$

$$2,1 \text{ minor:} \quad \begin{bmatrix} \cancel{1} & 2 & 3 \\ \cancel{4} & \cancel{5} & \cancel{6} \\ \cancel{1} & 2 & 0 \end{bmatrix}; \quad 2,1\text{-minor} = \begin{vmatrix} 2 & 3 \\ 2 & 0 \end{vmatrix} = -6; \quad \text{matrix of minors} = \begin{bmatrix} -12 & -6 & 3 \\ -6 & * & * \\ * & * & * \end{bmatrix}$$

There are 5 more minors to compute. We show each of them, but without labels. You should practice by naming them and computing their value.

$$\begin{bmatrix} 1 & \cancel{2} & 3 \\ \cancel{4} & \cancel{5} & \cancel{6} \\ 1 & \cancel{2} & 0 \end{bmatrix}; \quad \begin{bmatrix} 1 & 2 & \cancel{3} \\ \cancel{4} & \cancel{5} & \cancel{6} \\ 1 & 2 & \cancel{0} \end{bmatrix}; \quad \begin{bmatrix} \cancel{1} & 2 & 3 \\ \cancel{4} & 5 & 6 \\ \cancel{1} & \cancel{2} & \cancel{0} \end{bmatrix}; \quad \begin{bmatrix} 1 & \cancel{2} & 3 \\ 4 & \cancel{5} & 6 \\ \cancel{1} & \cancel{2} & \cancel{0} \end{bmatrix}; \quad \begin{bmatrix} 1 & 2 & \cancel{3} \\ 4 & 5 & \cancel{6} \\ \cancel{1} & \cancel{2} & \cancel{0} \end{bmatrix};$$

The entire matrix of minors is therefore: $\begin{bmatrix} -12 & -6 & 3 \\ -6 & -3 & 0 \\ -3 & -6 & -3 \end{bmatrix}$.

Matrix of cofactors. Recall the checkerboard of signs we used for computing the determinant: $\begin{bmatrix} + & - & + \\ - & + & - \\ + & - & + \end{bmatrix}$. To compute the matrix of cofactors of A, you change the signs in the matrix of minors according to the checkerboard.

Example 15.13. Find the matrix of cofactors for the matrix A in the previous example.

Solution: The matrix of minors is $\begin{bmatrix} -12 & -6 & 3 \\ -6 & -3 & 0 \\ -3 & -6 & -3 \end{bmatrix}$ So the matrix of cofactors is $\begin{bmatrix} -12 & 6 & 3 \\ 6 & -3 & 0 \\ -3 & 6 & -3 \end{bmatrix}$. (Look carefully at how we changed signs to go from minors to cofactors.)

Adjoint. To make the adjoint matrix you take the transpose of the cofactors matrix, i.e., switch the rows and columns of the cofactors matrix.

Example 15.14. Find the adjoint matrix for the matrix A in the previous examples.

Solution: The matrix of cofactors is $\begin{bmatrix} -12 & 6 & 3 \\ 6 & -3 & 0 \\ -3 & 6 & -3 \end{bmatrix}$. So the adjoint is $\begin{bmatrix} -12 & 6 & -3 \\ 6 & -3 & 6 \\ 3 & 0 & -3 \end{bmatrix}$.

Example 15.15. Find the inverse for the matrix A in the previous examples.

Solution: We can find $\det A = 9$ using the minors for the first row computed in Example 15.12. The matrix of cofactors is $\begin{bmatrix} -12 & 6 & -3 \\ 6 & -3 & 6 \\ 3 & 0 & -3 \end{bmatrix}$, so the inverse is

$$A^{-1} = \frac{1}{9} \begin{bmatrix} -12 & 6 & -3 \\ 6 & -3 & 6 \\ 3 & 0 & -3 \end{bmatrix}.$$

Finding the inverse. Divide the matrix of cofactors by $\det A$.

The next example will show a good way to organize the computation.

Example 15.16. Compute the inverse of the matrix $A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 1 & 2 \\ 1 & 2 & 0 \end{bmatrix}$.

Solution: In order to have an inverse we need $\det(A) \neq 0$. So our first step is to compute the determinant. We do this by expansion along the first row:

$$\det(A) = \begin{vmatrix} 1 & 2 & 3 \\ 2 & 1 & 2 \\ 1 & 2 & 0 \end{vmatrix} = 1\begin{vmatrix} 1 & 2 \\ 2 & 0 \end{vmatrix} - 2\begin{vmatrix} 2 & 2 \\ 1 & 0 \end{vmatrix} + 3\begin{vmatrix} 2 & 1 \\ 1 & 2 \end{vmatrix} = 1(-4) - 2(-2) + 3(3) = 9.$$

Since $\det(A) \neq 0$, the inverse exists and we can proceed with the algorithm to compute A^{-1}. Only the first step requires any real computation.

The algorithm says to first compute the matrix of minors. Notice that we found the minors for the first row when we computed the determinant. We can reuse those and only need to compute the other 6. (Actually we'll just use the answers from the previous examples.)

1. Matrix of minors $= \begin{bmatrix} -4 & -2 & 3 \\ -6 & -3 & 0 \\ 1 & -4 & -3 \end{bmatrix}$ (compute each minor).

2. Matrix of cofactors $= \begin{bmatrix} -4 & 2 & 3 \\ 6 & -3 & 0 \\ 1 & 4 & -3 \end{bmatrix}$ (apply checkerboard).

3. Adjoint $= \begin{bmatrix} -4 & 6 & 1 \\ 2 & -3 & 4 \\ 3 & 0 & -3 \end{bmatrix}$ (swap rows and columns).

4. Inverse: $\boxed{A^{-1} = \frac{1}{9}\begin{bmatrix} -4 & 6 & 1 \\ 2 & -3 & 4 \\ 3 & 0 & -3 \end{bmatrix}}$ (divide by $\det(A)$.

We can check this by multiplying by multiplying $A^{-1} \cdot A$ and seeing that we get I. (You'll have to do the actual computation.)

$$A^{-1} \cdot A = \frac{1}{9}\begin{bmatrix} -4 & 6 & 1 \\ 2 & -3 & 4 \\ 3 & 0 & -3 \end{bmatrix}\begin{bmatrix} 1 & 2 & 3 \\ 2 & 1 & 2 \\ 1 & 2 & 0 \end{bmatrix} = I.$$

Fun note: This algorithm works for the 2 × 2 case as well. You should try it out, it's very fast.

15.6.4 Proof that $(AB)^T = B^T A^T$

We'll use the following notation involving indices. Let the (i, j) entry of A be $A_{i,j}$. By the definition of transpose we have $(A^T)_{j,i} = A_{i,j}$. Likewise for other matrices. In order to show that $(AB)^T = B^T A^T$ we have to show that $((AB)^T)_{k,i} = (B^T A^T)_{k,i}$. We do this by keeping track of indices while multiplying matrices. Since $(AB)_{i,k} = \sum_j A_{i,j}B_{j,k}$ we have

$$((AB)^T)_{k,i} = (AB)_{i,k} = \sum_j A_{i,j}B_{j,k} = \sum_j B_{j,k}A_{i,j} = \sum_j (B^T)_{k,j}(A^T)_{j,i} = (B^TA^T)_{k,i} \qquad \text{QED}$$

15.6.5 Left and right inverses

A left inverse for a matrix A is a matrix L such that left-multiplication by L gives the identity, e.g.,

$$L \cdot A = I$$

The definition of a right inverse is similar.

Non-square matrices can have one-sided inverses. For example the matrix $A = \begin{bmatrix} 6 & 5 & 2 \\ 1 & 2 & 4 \end{bmatrix}$ has a right inverse (in fact many of them). For example,

$$\begin{bmatrix} 6 & 5 & 2 \\ 1 & 2 & 4 \end{bmatrix} \begin{bmatrix} 2/7 & -5/7 \\ -1/7 & 6/7 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

But A has no left inverse. Likewise, there are matrices with left inverses but no right inverses.

Here are some facts about these one-sided inverses. We won't give details, but you do have all the tools to understand the details: ask if you're interested.

1. A square matrix either has a single two-sided inverse, i.e., both a left and a right inverse, or it has no inverses of any kind.

2. If $n < m$ then an $n \times m$ matrix A cannot have a left inverse. If the rank of A is n then it has a right inverse. The example just above, illustrates this for A a 2×3 matrix of rank 2.

3. If $n > m$ then an $n \times m$ matrix A cannot have a right inverse. If the rank of A is m then it has a left inverse.

16 Eigenvalues, diagonalization, decoupling

This note covers topics that will take us several classes to get through. While we we will look at $n \times n$ matrices, most of our computational examples will use 2×2 matrices. These have almost all the features of bigger square matrices and they are computationally much easier.

16.1 Etymology:

This is from a Wikipedia discussion page: The word **eigen** in German or Dutch translates as 'inherent', 'characteristic', 'private'. So an eigenvector of a matrix is characteristic or inherent to the matrix. The word eigen is also translated as 'own' with the same sense as the meanings above. That is the eigenvector of a matrix is the matrix's 'own vector'.

In English you sometimes see eigenvalues called special or characteristic values.

16.2 Definition

For a square matrix M, an eigenvalue is a number (scalar) that satisfies the equation

$$M\,\mathbf{v} = \lambda\,\mathbf{v} \text{ for some non-zero vector } \mathbf{v}. \tag{24}$$

The vector $\mathbf{v}$ is called a non-zero eigenvector corresponding to λ. We will call Equation 16.1 the eigenvector equation.

Comments:
1. Using the symbol λ for the eigenvalue is a fairly common practice when looking at generic matrices. If the eigenvalue has a physical interpretation, we'll often use a corresponding letter. For example, in population matrices the eigenvalues are growth rates, so we'll often denote them using r or k.

2. Eigenvectors are not unique. That is, if $\mathbf{v}$ is an eigenvector with eigenvalue λ then so is any multiple of $\mathbf{v}$. Indeed, the set of all eigenvectors with eigenvalue λ is clearly a vector space. (You should convince yourself of this!)

16.3 Why eigenvectors are special

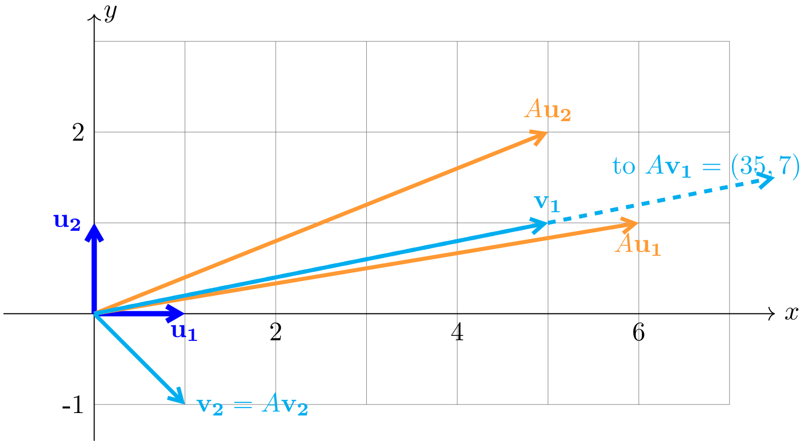
Example 16.1. Let $A = \begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix}$. We will explore how A transforms vectors and what makes an eigenvector special. We will see that A scales and rotates most vectors, but only scales eigenvectors. That is, eigenvectors lie on lines that are unmoved by A.

Take $\mathbf{u_1} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \Rightarrow A\mathbf{u_1} = \begin{bmatrix} 6 \\ 1 \end{bmatrix}$; Take $\mathbf{u_2} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \Rightarrow A\mathbf{u_2} = \begin{bmatrix} 5 \\ 2 \end{bmatrix}$.

We see that A scales and turns most vectors.

Now take $\mathbf{v_1} = \begin{bmatrix} 5 \\ 1 \end{bmatrix} \Rightarrow A\mathbf{v_1} = \begin{bmatrix} 35 \\ 7 \end{bmatrix} = 7\mathbf{v_1}$. By the definition in Equation 24, this shows that $\mathbf{v_1}$ is an eigenvector with eigenvalue 7. The eigenvector is special since A scales it by 7, but does not rotate it.

Likewise, $\mathbf{v_2} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$ then $A\mathbf{v_2} = \mathbf{v_2}$. So $\mathbf{v_2}$ is an eigenvector with eigenvalue 1. The eigenvector $\mathbf{v_2}$ is really special, it is unmoved by A.



Example 16.1: Action of the matrix A on vectors

The following example shows how knowing eigenvalues and eigenvectors simplifies calculations with a matrix. In fact, you don't even need the matrix once you know all of its eigenvalues and eigenvectors.

Example 16.2. Suppose A is a 2×2 matrix that has eigenvectors $\begin{bmatrix}1\\2\end{bmatrix}$ and $\begin{bmatrix}1\\3\end{bmatrix}$ with eigenvalues 2 and 4 respectively.

(a) Compute $A\begin{bmatrix}1\\2\end{bmatrix}$.

Solution: Since $\begin{bmatrix}1\\2\end{bmatrix}$ is an eigenvector, this follows directly from the definition of eigenvectors: $A\begin{bmatrix}1\\2\end{bmatrix} = 2\begin{bmatrix}1\\2\end{bmatrix} = \begin{bmatrix}2\\4\end{bmatrix}$.

(b) Compute $A\left(\begin{bmatrix}1\\2\end{bmatrix} + \begin{bmatrix}1\\3\end{bmatrix}\right)$.

Solution: This uses the definition of eigenvector plus linearity:

$$A\left(\begin{bmatrix}1\\2\end{bmatrix} + \begin{bmatrix}1\\3\end{bmatrix}\right) = A\begin{bmatrix}1\\2\end{bmatrix} + A\begin{bmatrix}1\\3\end{bmatrix} = 2\begin{bmatrix}1\\2\end{bmatrix} + 4\begin{bmatrix}1\\3\end{bmatrix} = \begin{bmatrix}6\\16\end{bmatrix}.$$

(c) Compute $A\left(3\begin{bmatrix}1\\2\end{bmatrix} + 5\begin{bmatrix}1\\3\end{bmatrix}\right)$.

Solution: Again this uses the definition of eigenvector plus linearity:

$$A\left(3\begin{bmatrix}1\\2\end{bmatrix} + 5\begin{bmatrix}1\\3\end{bmatrix}\right) = 3A\begin{bmatrix}1\\2\end{bmatrix} + 5A\begin{bmatrix}1\\3\end{bmatrix} = 6\begin{bmatrix}1\\2\end{bmatrix} + 20\begin{bmatrix}1\\3\end{bmatrix} = \begin{bmatrix}26\\72\end{bmatrix}.$$

(d) Compute $A\begin{bmatrix}0\\1\end{bmatrix}$.

Solution: We first decompose $\begin{bmatrix}0\\1\end{bmatrix}$ into eigenvectors:

$$\begin{bmatrix}0\\1\end{bmatrix} = \begin{bmatrix}1\\3\end{bmatrix} - \begin{bmatrix}1\\2\end{bmatrix}.$$

Now we can once again use the definition of eigenvector plus linearity:

$$A\begin{bmatrix}0\\1\end{bmatrix} = A\left(\begin{bmatrix}1\\3\end{bmatrix} - \begin{bmatrix}1\\2\end{bmatrix}\right) = A\begin{bmatrix}1\\3\end{bmatrix} - A\begin{bmatrix}1\\2\end{bmatrix} = 4\begin{bmatrix}1\\3\end{bmatrix} + 2\begin{bmatrix}1\\2\end{bmatrix} = \begin{bmatrix}2\\8\end{bmatrix}.$$

Example 16.3. Any rotation in three dimensions is around some axis. The vector along this axis is fixed by the rotation, i.e., it is an eigenvector with eigenvalue 1.

16.4 Computational algorithm

We start by summarizing the method. We will justify it and give examples below.

Computational method:
1. The eigenvalues of A are the roots of the characteristic equation

$$\det(A - \lambda I) = 0 \tag{25}$$

2. The corresponding eigenspace of A is $\text{Null}(A - \lambda I)$.

Notes. 1. Again, we call Equation 25 the characteristic equation. (Eigenvalues are sometimes called characteristic values.) It allows us to find the eigenvalues and eigenvectors separately in a two step process.

2. The eigenspace is so-called, because it is the vector subspace which consists of all eigenvectors corresponding to λ.

3. **Notation:** For simplicity we will sometimes use the notation $|A| = \det(A)$. So the characteristic equation can be written $|A - \lambda I| = 0$.

16.4.1 Justification of the computational algorithm

First we recall the following basic fact about square matrices from Topic 15.

Fact: The null space of A is nontrivial exactly when $\det(A) = 0$.

Next, we manipulate the eigenvalue equation (Equation 24) so that finding eigenvectors becomes finding null vectors. Suppose, λ is an eigenvalue and $\mathbf{v}$ is a corresponding nonzero eigenvector. Then, starting with the eigenequation we have:

$$A\,\mathbf{v} = \lambda\,\mathbf{v} \;\Leftrightarrow\; A\,\mathbf{v} = \lambda\,I\mathbf{v} \;\Leftrightarrow\; A\,\mathbf{v} - \lambda\,I\mathbf{v} = 0 \;\Leftrightarrow\; (A - \lambda\,I)\mathbf{v} = 0.$$

Since $\mathbf{v} \neq 0$, the last equation just above says $A - \lambda\,I$ has a nontrivial null space. So our fact about determinants and null spaces tells us that λ is an eigenvalue if and only if $\det(A - \lambda\,I) = 0$, i.e., if and only if λ is a root of the characteristic equation. This justifies Step 1 in the algorithm.

Likewise, the equation $(A - \lambda\,I)\mathbf{v} = 0$ says that $\mathbf{v}$ is an eigenvector corresponding to λ if and only if $\mathbf{v}$ is in $\text{Null}(A - \lambda I)$. This justifies Step 2 in the algorithm.

16.4.2 Examples

Example 16.4. Find the eigenvalues of the matrix $A = \begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix}$. For each eigenvalue find a basis of the corresponding eigenspace.

Solution: Step 1. Find the eigenvalues λ: $|A - \lambda I| = 0$ (characteristic equation)

$$A - \lambda I = \begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} = \begin{bmatrix} 6-\lambda & 5 \\ 1 & 2-\lambda \end{bmatrix}.$$

Taking the determinant and setting it to 0 gives

$$\det(A - \lambda I) = (6-\lambda)(2-\lambda) - 5 = \lambda^2 - 8\lambda + 7 = 0.$$

The roots of this are $\boxed{\lambda = 7,\ 1.}$

Step 2. For each eigenvalue, find basis vectors for the eigenspace, i.e., find a basis of Null$(A - \lambda I)$.

$\lambda_1 = 7$: $A - \lambda I = \begin{bmatrix} -1 & 5 \\ 1 & -5 \end{bmatrix}$. This has RREF $R = \begin{bmatrix} 1 & -5 \\ 0 & 0 \end{bmatrix}$. The null space is 1 dimensional, a basis is $\mathbf{v_1} = \begin{bmatrix} 5 \\ 1 \end{bmatrix}$.

$\lambda_1 = 1$: $A - \lambda I = \begin{bmatrix} 5 & 5 \\ 1 & 1 \end{bmatrix}$. This has RREF $R = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$. The null space is 1 dimensional, a basis is $\mathbf{v_2} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$.

Remember, any scalar multiple of these eigenvectors is also an eigenvector with the same eigenvalue.

Let's reemphasize a key point:

Example 16.5. Eigenspaces are null spaces. Consider the matrix

$$A = \begin{bmatrix} 4 & 8 & -2 & 2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

Find the eigenvalues and eigenspaces of A.

Solution: This is a 4×4 matrix, but the characteristic equation is not hard to find.

$$|A - \lambda I| = \begin{vmatrix} 4-\lambda & 8 & -2 & 2 \\ 0 & -\lambda & 0 & 0 \\ 0 & 0 & 1-\lambda & 1 \\ 0 & 0 & 1 & 1-\lambda \end{vmatrix} = (4-\lambda)(-\lambda)(\lambda^2 - 2\lambda) = -\lambda^2(4-\lambda)(\lambda-2).$$

So the eigenvalues are $\lambda = 0,\ 0,\ 4,\ 2.$

Eigenspace for $\lambda = 0$:

We must find Null(A): The RREF of A is $R = \begin{bmatrix} 1 & 2 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$.

Using this we see that Null(A) (eigenspace for $\lambda = 0$) is 2 dimensional and has basis

$$\left\{ \begin{bmatrix} -2 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \\ -1 \\ 1 \end{bmatrix} \right\}$$

Let's highlight that Null(A) is nontrivial means $\lambda = 0$ is an eigenvalue.

For the other two eigenvalues we must find Null($A - 4I$) and Null($A - 2I$). This is not hard and you should do it as an exercise. We get:

The eigenspace for $\lambda = 4$ has basis $\left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}$. The eigenspace for $\lambda = 2$ has basis $\left\{ \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} \right\}$.

Notes.

Trick. In the 2 × 2 case we don't have to write out the RREF to find the eigenvector. Notice that the entries in our eigenvectors come from the entries in one row of the matrix. The eigenvector is the column vector with entries: right entry of the row, minus the left entry. For example, if $A - \lambda I = \begin{bmatrix} -1 & 5 \\ 1 & -5 \end{bmatrix}$, then, using the top row, we see that $\mathbf{v} = \begin{bmatrix} 5 \\ 1 \end{bmatrix}$ is a basis vector for Null($A - \lambda I$). If you think about this a moment, you'll see why it must be the case.

Matlab: In Matlab the function `eig(A)` returns the eigenvectors and eigenvalues of a matrix.

16.5 Complex eigenvalues

If the eigenvalues are complex, then the eigenvectors are complex. Otherwise there is no difference in the algebra.

Example 16.6. Find the eigenvalues and basic eigenvectors of the matrix $A = \begin{bmatrix} 3 & 4 \\ -4 & 3 \end{bmatrix}$.

Solution: Step 1. Find the eigenvalues λ: $|A - \lambda I| = 0$ **(characteristic equation)**

$$\det\left(\begin{bmatrix} 3 & 4 \\ -4 & 3 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \right) = \begin{vmatrix} 3-\lambda & 4 \\ -4 & 3-\lambda \end{vmatrix} = (3-\lambda)^2 + 16 = 0.$$

So the eigenvalues are $\boxed{\lambda = 3 \pm 4i.}$

Step 2. Find corresponding basic eigenvectors. That is, find a basis of Null($A - \lambda I$). (In the 2 × 2 case, we can do that without doing row reduction.)

$\lambda_1 = 3 + 4i$: $\quad (A - \lambda I) = \begin{bmatrix} -4i & 4 \\ -4 & -4i \end{bmatrix}$. Take $\boxed{\mathbf{v_1} = \begin{bmatrix} 1 \\ i \end{bmatrix}}$.

$\lambda_2 = 3 - 4i$: $\quad (A - \lambda\, I) = \begin{bmatrix} 4i & 4 \\ -4 & 4i \end{bmatrix}$. Take $\boxed{\mathbf{v_2} = \begin{bmatrix} 1 \\ -i \end{bmatrix}}$.

Time saver: Notice that the eigenvalues and eigenvectors come in complex conjugate pairs. Knowing this, there is no need to do a computation to find the second member of each pair.

Example 16.7. Find the eigenvalues and basic eigenvectors of the matrix $A = \begin{bmatrix} 1 & -4 \\ 5 & 5 \end{bmatrix}$.

Solution: Step 1. Find λ (eigenvalues): $\quad |A - \lambda I| = 0 \quad$ (**characteristic equation)**

$$\begin{vmatrix} 1-\lambda & -4 \\ 5 & 5-\lambda \end{vmatrix} = \lambda^2 - 6\lambda + 25 = 0 \Rightarrow \boxed{\lambda = 3 \pm 4i.}$$

Step 2. Find corresponding basic eigenvectors (basis of Null$(A - \lambda I)$):

$\lambda_1 = 3 + 4i$: $\quad (A - \lambda\, I) = \begin{bmatrix} -2-4i & -4 \\ 5 & 2-4i \end{bmatrix}$. Take $\boxed{\mathbf{v_1} = \begin{bmatrix} 4 \\ -2-4i \end{bmatrix}}$.

$\lambda_2 = 3 - 4i$: $\quad$ Take $\boxed{\mathbf{v_2} = \overline{\mathbf{v}}_\mathbf{1} = \begin{bmatrix} 4 \\ -2+4i \end{bmatrix}}$.

16.6 Repeated eigenvalues

When a matrix has repeated eigenvalues the eigenvectors are not as well behaved as when the eigenvalues are distinct. There are two main examples

Example 16.8. (Defective case) Find the eigenvalues and basic eigenvectors of the matrix $A = \begin{bmatrix} 3 & 1 \\ 0 & 3 \end{bmatrix}$.

Solution: Step 1. Find the eigenvalues λ: $\quad |A - \lambda I| = 0 \quad$ (**characteristic equation)**

$$\begin{vmatrix} 3-\lambda & 1 \\ 0 & 3-\lambda \end{vmatrix} = (\lambda - 3)^2 = 0 \Rightarrow \boxed{\lambda = 3,\ 3.}$$

Step 2. Find the basic eigenvectors (basis of Null$(A - \lambda I)$):

$\lambda_1 = 3$: $\quad \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \mathbf{v_1}$. This is already in RREF. It has one free variable, so the null space is 1 dimensional. We can take a basis vector: $\boxed{\mathbf{v_1} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}}$.

We have two eigenvalues but only one independent eigenvector, so we call this case defective or incomplete. In linear algebra, there is a lot to explore with defective matrices. In 18.03, we will not go into a lot of detail about them.

Example 16.9. (Complete case) Find the eigenvalues and basic eigenvectors of the matrix $A = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}$.

Solution: Step 1. Find the eigenvalues λ: $|A - \lambda I| = 0$ **(characteristic equation)**

$$\begin{vmatrix} 3-\lambda & 0 \\ 0 & 3-\lambda \end{vmatrix} = (\lambda - 3)^2 = 0 \Rightarrow \boxed{\lambda = 3,\ 3.}$$

Step 2. Find corresponding basic eigenvectors (basis of Null$(A - \lambda I)$):

$\lambda_1 = 3$: $A - \lambda I = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$.

This equation shows that every vector in $\mathbf{R}^2$ is an eigenvector. That is, the eigenvalue $\lambda = 3$ has a two dimensional eigenspace. We can pick any two independent vectors as a basis, e.g., $\mathbf{v_1} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\mathbf{v_2} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. (These are the simplest choices, but any two independent vectors would work!)

Because we have as many independent eigenvectors as eigenvalues, we call this case complete.

16.7 Diagonal matrices

In this section we will see how easy it is to work with diagonal matrices. In later sections we will see how working with eigenvalues and eigenvectors of a matrix is like turning it into a diagonal matrix.

Example 16.10. Consider the diagonal matrix $B = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$

Convince yourself that $B \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} 2u \\ 3v \end{bmatrix}$. That is B scales the first coordinate by 2 and the second coordinate by 3.

We can write this as

$$B \begin{bmatrix} 1 \\ 0 \end{bmatrix} = 2 \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \text{and} \quad B \begin{bmatrix} 0 \\ 1 \end{bmatrix} = 3 \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

This is exactly the definition of eigenvectors. That is, $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$ are eigenvectors with eigenvalues 2 and 3 respectively. We state this an an important fact.

Important fact. For a diagonal matrix, the diagonal entries are the eigenvalues and the eigenvectors are the standard basis vectors.

Example 16.11. The matrix $A = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 4 \end{bmatrix}$ has eigenvalues and corresponding basic eigenvectors

$$\begin{aligned} \lambda &= 2, \qquad 3, \qquad 4 \\ \mathbf{v} &= \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \end{aligned}$$

You can check this by multiplying A times each eigenvector.

Example 16.12. For the matrix A in the previous example, compute $\det A$, A^2, A^5.

Solution: $\det(A)$ = product of diagonal entries = 24.

$A^2 = \begin{bmatrix} 2^2 & 0 & 0 \\ 0 & 3^2 & 0 \\ 0 & 0 & 4^2 \end{bmatrix}$. Likewise, $A^5 = \begin{bmatrix} 2^5 & 0 & 0 \\ 0 & 3^5 & 0 \\ 0 & 0 & 4^5 \end{bmatrix}$

16.8 Diagonalization

Diagonalization is a way to make a matrix almost as easy to work with as a diagonal matrix.

Theorem. Diagonalization theorem. Suppose the $n \times n$ matrix A has n independent eigenvectors. Then, we can write

$$A = S\Lambda S^{-1},$$

where S is a matrix whose columns are the n independent eigenvectors and Λ is the diagonal matrix whose diagonal entries are the corresponding eigenvalues.

The proof is below. We illustrate this first with our standard example.

Example 16.13. We know the matrix $A = \begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix}$ has eigenvalues 7 and 1 with corresponding basic eigenvectors $\mathbf{v_1} = \begin{bmatrix} 5 & 1 \end{bmatrix}^T$ and $\mathbf{v_2} = \begin{bmatrix} -1 & 1 \end{bmatrix}^T$

We put the eigenvectors as the columns of a matrix S and the eigenvalues as the entries of a diagonal matrix Λ.

$$S = \begin{bmatrix} \mathbf{v_1} & \mathbf{v_2} \end{bmatrix} = \begin{bmatrix} 5 & -1 \\ 1 & 1 \end{bmatrix}, \qquad \Lambda = \begin{bmatrix} 7 & 0 \\ 0 & 1 \end{bmatrix}$$

The diagonalization theorem says that

$$A = S\Lambda S^{-1} = \begin{bmatrix} 5 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 7 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1/6 & 1/6 \\ -1/6 & 5/6 \end{bmatrix}.$$

This is called the diagonalization of A. Note the form: a diagonal matrix Λ surrounded by S and S^{-1}.

Proof of the diagonalization theorem. We will do this for the matrix in the example above. It should be clear that this proof carries over to any $n \times n$ matrix with n independent eigenvectors.

The equation $A = S\Lambda S^{-1}$ can be rewritten as $AS = S\Lambda$. We will show this is true by showing that both sides have the same effect when multiplying any vector. That is,

$$AS\mathbf{v} = S\Lambda\mathbf{v}$$

for any vector $\mathbf{v}$

First, let $\mathbf{e_1} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $\mathbf{e_2} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ be the standard basis vectors of $\mathbf{R}^2$. Since every vector is a linear combination of the basis vectors, it is enough to show

$$AS\mathbf{e_1} = S\Lambda\mathbf{e_1} \quad \text{and} \quad AS\mathbf{e_2} = S\Lambda\mathbf{e_2}.$$

Recall that multiplying a matrix times a column vector results in a linear combination of the columns. In our case,

$$S\mathbf{e_1} = \begin{bmatrix}\mathbf{v_1}\,\mathbf{v_2}\end{bmatrix}\begin{bmatrix}1\\0\end{bmatrix} = \mathbf{v_1}, \quad \text{and} \quad S\mathbf{e_2} = \begin{bmatrix}\mathbf{v_1}\,\mathbf{v_2}\end{bmatrix}\begin{bmatrix}0\\1\end{bmatrix} = \mathbf{v_2}.$$

Now we can check that $AS\mathbf{e_1} = S\Lambda\mathbf{e_1}$:

$$AS\mathbf{e_1} = A\mathbf{v_1} = 7\mathbf{v_1} \quad \text{and} \quad S\Lambda\mathbf{e_1} = S\begin{bmatrix}7 & 0\\0 & 1\end{bmatrix}\begin{bmatrix}1\\0\end{bmatrix} = S\begin{bmatrix}7\\0\end{bmatrix} = 7\mathbf{v_1}.$$

The equation $A\mathbf{v_1} = 7\mathbf{v_1}$ follows because $\mathbf{v_1}$ is an eigenvector of A with eigenvalue 7. Thus we have shown that $AS\mathbf{e_1} = S\Lambda\mathbf{e_1}$. In exactly the same way, we can show that $AS\mathbf{e_2} = S\Lambda\mathbf{e_2}$.

Thus we can conclude that $AS = S\Lambda$. So, $A = S\Lambda S^{-1}$.

In general, the steps for diagonalizing an $n \times n$ matrix A are:

1. Find the eigenvalues $\lambda_1, \dots, \lambda_n$ and corresponding basic eigenvectors $\mathbf{v_1}, \dots, \mathbf{v_n}$.
2. Make the matrix of eigenvectors $S = \begin{bmatrix}\mathbf{v_1} & \mathbf{v_2} & \cdots & \mathbf{v_n}\end{bmatrix}$
3. Make the diagonal matrix of eigenvalues $\Lambda = \begin{bmatrix}\lambda_1 & 0 & 0 & \cdots & 0\\ 0 & \lambda_2 & 0 & \cdots & 0\\ \vdots & \vdots & \vdots & \ddots & 0\\ 0 & 0 & 0 & \cdots & \lambda_n\end{bmatrix}$

The diagonalization is: $A = S\Lambda S^{-1}$.

Note: Diagonalization requires that A have a full complement of eigenvectors. If A is defective, it can't be diagonalized.

We have the following important formula

$$\det(\mathbf{A}) = \textbf{ product of its eigenvalues.}$$

This follows easily from the diagonalization formula

$$\det(A) = \det(S\Lambda S^{-1}) = \det(S)\det(\Lambda)\det(S^{-1}) = \det(\Lambda) = \text{ product of diagonal entries.}$$

Example 16.14. Consider the matrix $A = \begin{bmatrix}5 & -1\\1 & 1\end{bmatrix}\begin{bmatrix}7 & 0\\0 & 1\end{bmatrix}\begin{bmatrix}5 & -1\\1 & 1\end{bmatrix}^{-1}$.

(a) What are the eigenvalues and eigenvectors of A.

(b) Compute $\det A$, A^2, A^5.

Solution: For ease of writing, let $S = \begin{bmatrix}5 & -1\\1 & 1\end{bmatrix}$ and $\Lambda = \begin{bmatrix}7 & 0\\0 & 1\end{bmatrix}$. So, $A = S\Lambda S^{-1}$.

(a) The columns of S are eigenvectors and the diagonal entries of Λ are the corresponding eigenvalues. We have eigenpairs

$$\lambda = 7,\ \mathbf{v} = \begin{bmatrix}5\\1\end{bmatrix} \qquad \text{and} \qquad \lambda = 1,\ \mathbf{v} = \begin{bmatrix}-1\\1\end{bmatrix}.$$

(b) We have $\det A = \det \Lambda = 7$. We also have

$$A^2 = S\Lambda S^{-1} \cdot S\Lambda S^{-1} = S\Lambda^2 S^{-1} = S \begin{bmatrix} 7^2 & 0 \\ 0 & 1^2 \end{bmatrix} S^{-1}.$$

Likewise $A^5 = S\Lambda^5 S^{-1} = S \begin{bmatrix} 7^5 & 0 \\ 0 & 1^5 \end{bmatrix} S^{-1}$.

16.9 Diagonal matrices and uncoupled algebraic systems

Example 16.15. (An uncoupled algebraic system) Consider the system

$$\begin{array}{rcl} 7u \qquad & = & 1 \\ v & = & 3 \end{array}$$

The variables u and v are uncoupled. That is, they never occur in the same equation. We can solve the system by finding each variable separately: $u = 1/7$, $v = 3$.

Example 16.16. Now consider the system

$$\begin{array}{lcl} 6x & + & 5y = 2 \\ x & + & 2y = 4. \end{array}$$

In matrix form this is

$$\begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2 \\ 4 \end{bmatrix} \tag{26}$$

The matrix $A = \begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix}$ is the same matrix as in Examples 16.1 and 16.4 above. In this system the variables x and y are coupled. We will explain the logic of decoupling later. For this example, we will decouple the equations using some magical choices involving eigenvectors.

The examples above showed that the eigenvalues of A are 7 and 1 with eigenvectors $\begin{bmatrix} 5 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} -1 \\ 1 \end{bmatrix}$. We write all vectors in terms of the eigenvectors by making the change of variables

$$\begin{bmatrix} x \\ y \end{bmatrix} = u \begin{bmatrix} 5 \\ 1 \end{bmatrix} + v \begin{bmatrix} -1 \\ 1 \end{bmatrix} \Leftrightarrow x = 5u - v; \quad y = u + v.$$

For the future, we note: $\begin{bmatrix} 2 \\ 4 \end{bmatrix} = \begin{bmatrix} 5 \\ 1 \end{bmatrix} + 3 \begin{bmatrix} -1 \\ 1 \end{bmatrix}$.

Converting our equation from x and y to u and v we get

$$\begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix} \left(u \begin{bmatrix} 5 \\ 1 \end{bmatrix} + v \begin{bmatrix} -1 \\ 1 \end{bmatrix} \right) = 7u \begin{bmatrix} 5 \\ 1 \end{bmatrix} + v \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

Thus,

$$\begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2 \\ 4 \end{bmatrix} \qquad \Leftrightarrow \qquad 7u \begin{bmatrix} 5 \\ 1 \end{bmatrix} + v \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 5 \\ 1 \end{bmatrix} + 3 \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

It is easy to see that the last system is the same as the equations

$$7u = 1 \qquad v = 3.$$

In u, v coordinates the system is diagonal and easy to solve.

16.10 Introduction to matrix methods for solving systems of DEs

In this section we will solve linear, homogeneous, constant coefficient systems of differential equations using the matrix methods we have developed. For now we will just consider matrices with real, distinct eigenvalues. In the next topic we will look at complex and repeated eigenvalues.

As with constant coefficient DEs, we will use the method of optimism to discover a systematic technique for solving systems of DEs. We start by giving the general 2 × 2 linear, homogeneous, constant coefficient system of DEs. It has the form

$$\begin{array}{rcrcl} x' & = & ax & + & by \\ y' & = & cx & + & dy. \end{array} \tag{27}$$

Here a, b, c, d are constants and $x(t)$, $y(t)$ are the unknown functions we need to solve for.

There are a number of important things to note.

1. We can write Equation 27 in matrix form

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \;\Leftrightarrow\; \mathbf{x}' = A\mathbf{x} \tag{28}$$

where $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ and $\mathbf{x} = \begin{bmatrix} x \\ y \end{bmatrix}$.

2. The system is homogeneous. You can see this by taking Equation 27 and putting all the x and y on the left side so that the right side becomes all zeros.

3. The system is linear. You should be able to check directly that a linear combination of solutions to Equation 28 is also a solution.

We illustrate the method of optimism for solving Equation 28 with an example.

Example 16.17. Solve the linear, homogeneous, constant coefficient system

$$\mathbf{x}' = A\mathbf{x}, \quad \text{where} \quad \mathbf{x} = \begin{bmatrix} x \\ y \end{bmatrix} \quad \text{and} \quad A = \begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix}.$$

Solution: Using the method of optimism we try a solution

$$\mathbf{x} = e^{\lambda t}\mathbf{v},$$

where λ is a constant and $\mathbf{v}$ is a constant vector. Substituting the trial solution into both sides of the DE we get

$$\lambda e^{\lambda t}\mathbf{v} = e^{\lambda t}A\mathbf{v} \;\Leftrightarrow\; A\mathbf{v} = \lambda\mathbf{v}.$$

This is none other than the eigenvalue/eigenvector equation. So solving the system amounts to finding eigenvalues and eigenvectors. From our previous examples we know the eigenvalues and eigenvectors of A. We get two solutions.

$$\mathbf{x_1} = e^{t}\begin{bmatrix} -1 \\ 1 \end{bmatrix}, \qquad \mathbf{x_2} = e^{7t}\begin{bmatrix} 5 \\ 1 \end{bmatrix}.$$

The general solution is the span of these solutions:

$$\mathbf{x} = c_1\mathbf{x_1} + c_2\mathbf{x_2} = c_1 e^{t}\begin{bmatrix} -1 \\ 1 \end{bmatrix} + c_2 e^{7t}\begin{bmatrix} 5 \\ 1 \end{bmatrix}$$

The solutions $\mathbf{x_1}$ and $\mathbf{x_2}$ are called modal or basic solutions.

Now that we know where the method of optimism leads, we can do a second example starting directly with finding eigenvalues and eigenvectors

Example 16.18. Find the general solution to the system

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} 3 & 4 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

Solution: First find eigenvalues and basic eigenvectors.

Characteristic equation: $\begin{vmatrix} 3-\lambda & 4 \\ 1 & 3-\lambda \end{vmatrix} = \lambda^2 - 6\lambda + 5 = 0 \;\Rightarrow\; \lambda = 1,\, 5.$

Basic eigenvectors: (basis of $\text{Null}(A - \lambda I)$):

$\lambda = 1$: $(A - \lambda I) = \begin{bmatrix} 2 & 4 \\ 1 & 2 \end{bmatrix}$. Take $\mathbf{v_1} = \begin{bmatrix} -2 \\ 1 \end{bmatrix}$.

$\lambda = 5$: $(A - \lambda I) = \begin{bmatrix} -2 & 4 \\ 1 & - \end{bmatrix}$. Take $\mathbf{v_2} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$.

We have two modal solutions: $\mathbf{x_1} = e^t \mathbf{v_1}$ and $\mathbf{x_2} = e^{5t} \mathbf{v_2}$.

The general solution is $\mathbf{x} = c_1 \mathbf{x_1} + c_2 \mathbf{x_2} = c_1 e^t \begin{bmatrix} -2 \\ 1 \end{bmatrix} + c_2 e^{5t} \begin{bmatrix} 2 \\ 1 \end{bmatrix}$.

16.11 Decoupling systems of DEs

Example 16.19. (**An uncoupled system**) Consider the system

$$\begin{aligned} u'(t) &= 7u(t) \\ v'(t) &= v(t) \end{aligned}$$

Since u and v don't have any effect on each other, we say that u and v are uncoupled. It's easy to see the solution to this system is

$$\begin{aligned} u(t) &= c_1 e^{7t} \\ v(t) &= c_2 e^t \end{aligned}$$

In matrix form we have

$$\begin{bmatrix} u' \\ v' \end{bmatrix} = \begin{bmatrix} 7 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}.$$

The coefficient matrix has eigenvalues 7 and 1, with basic eigenvectors $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$. The general solution to the system of DEs is

$$\begin{bmatrix} u \\ v \end{bmatrix} = c_1 e^{7t} \begin{bmatrix} 1 \\ 0 \end{bmatrix} + c_2 e^t \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

We see an uncoupled system has a diagonal coefficient matrix and the basic eigenvectors are the standard basis vectors. All in all, it's simple and easy to work with.

The following example shows how to decouple a coupled system. After seeing this example, we will redo it, in a cleaner, more memorable way.

Example 16.20. Consider once again the system from Example 16.17

$$\begin{aligned} x' &= 6x+5y \\ y' &= x+2y. \end{aligned} \quad \Leftrightarrow \mathbf{x}' = A\mathbf{x}, \quad \text{where} \quad \mathbf{x} = \begin{bmatrix} x \\ y \end{bmatrix}, \quad A = \begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix} \tag{29}$$

In this system the variables x and y are coupled. Make a change of variable that converts this to a decoupled system.

Solution: From Example 16.17 we know the eigenvalues are 7 and 1, basic eigenvectors are $\begin{bmatrix} 5 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} -1 \\ 1 \end{bmatrix}$, and the general solution is $\begin{bmatrix} x \\ y \end{bmatrix} = c_1 e^{7t} \begin{bmatrix} 5 \\ 1 \end{bmatrix} + c_2 e^{t} \begin{bmatrix} -1 \\ 1 \end{bmatrix}$.

Notice that $c_1 e^{7t}$ and $c_2 e^t$ in the above solution are just u and v from the previous example. So we can write

$$\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = u(t) \begin{bmatrix} 5 \\ 1 \end{bmatrix} + v(t) \begin{bmatrix} -1 \\ 1 \end{bmatrix}. \tag{30}$$

This is a change of variables.

Let's rewrite the system in Equation 29 in terms of u, v. Using Equation 30, we get

$$\mathbf{x}' = \begin{bmatrix} x' \\ y' \end{bmatrix} = u' \begin{bmatrix} 5 \\ 1 \end{bmatrix} + v' \begin{bmatrix} -1 \\ 1 \end{bmatrix} \quad \text{and} \quad A \begin{bmatrix} x \\ y \end{bmatrix} = A\left(u \begin{bmatrix} 5 \\ 1 \end{bmatrix} + v \begin{bmatrix} -1 \\ 1 \end{bmatrix} \right) = 7u \begin{bmatrix} 5 \\ 1 \end{bmatrix} + v \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

The last equality follows because $\begin{bmatrix} 5 & 1 \end{bmatrix}^T$ and $\begin{bmatrix} -1 & 1 \end{bmatrix}^T$ are eigenvectors of A.

Equating the two sides we get

$$u' \begin{bmatrix} 5 \\ 1 \end{bmatrix} + v' \begin{bmatrix} -1 \\ 1 \end{bmatrix} = 7u \begin{bmatrix} 5 \\ 1 \end{bmatrix} + v \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

Comparing the coefficients of the eigenvectors we get

$$\begin{aligned} u' &= 7u \\ v' &= v \end{aligned} \quad \Leftrightarrow \quad \begin{bmatrix} u' \\ v' \end{bmatrix} = \begin{bmatrix} 7 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}$$

That is, in terms of u and v the system is uncoupled. Note that the eigenvalues of A are precisely the diagonal entries of the uncoupled system.

16.11.1 Decoupling in general

Though it's somewhat disguised, the key to the previous example was diagonalization. Bringing this to the forefront makes the example cleaner and less complicated.

Suppose A is written in diagonalized form: $A = S\Lambda S^{-1}$, where, as usual, S is a matrix with the eigenvectors of A as columns and Λ is the diagonal matrix with the corresponding eigenvalues as entries.

Decoupling: Suppose we have the system $\mathbf{x}' = A\mathbf{x}$, then the change of variables

$$\mathbf{u} = S^{-1}\mathbf{x}$$

converts the coupled system into an uncoupled system $\mathbf{u}' = \Lambda\mathbf{u}$.

Proof. The key is diagonalization: the system $\mathbf{x}' = A\mathbf{x}$ can be written

$$\mathbf{x}' = S\Lambda S^{-1}\mathbf{x} \quad \Leftrightarrow \quad S^{-1}\mathbf{x}' = \Lambda S^{-1}\mathbf{x}.$$

Now, letting $\mathbf{u} = S^{-1}\mathbf{x}$ converts this to the uncoupled system

$$\mathbf{u}' = \Lambda\mathbf{u}.$$

Since this is an uncoupled equation, making the change of variables $\mathbf{u} = S^{-1}\mathbf{x}$ is called decoupling the system.

To end this section, we'll walk through the previous example, being more explicit about the use of diagonalization.

Example 16.21. Decouple the system in Example 16.20 using the diagonalized form of A.

Solution: The system in Example 16.20 is $\mathbf{x}' = A\mathbf{x}$.

Let $S = \begin{bmatrix} 5 & -1 \\ 1 & 1 \end{bmatrix}$ = the matrix with eigenvectors of A as columns.

Let $\Lambda = \begin{bmatrix} 7 & 0 \\ 0 & 1 \end{bmatrix}$ = the diagonal matrix with the eigenvalues of A as diagonal entries.

Diagonalization says that $A = S\Lambda S^{-1}$.

The decoupling change of variables is $\mathbf{u} = S^{-1}\mathbf{x}$. We can write this as

$$\mathbf{x} = S\mathbf{u} \quad \text{or} \quad \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 5 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = u \begin{bmatrix} 5 \\ 1 \end{bmatrix} + v \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

This is exactly the change of variables used in Example 16.20.

The decoupled system is

$$\mathbf{u}' = \Lambda\mathbf{u} \quad \text{or} \quad \begin{bmatrix} u' \\ v' \end{bmatrix} = \begin{bmatrix} 7 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix},$$

which is exactly the decoupled system found in Example 16.20.

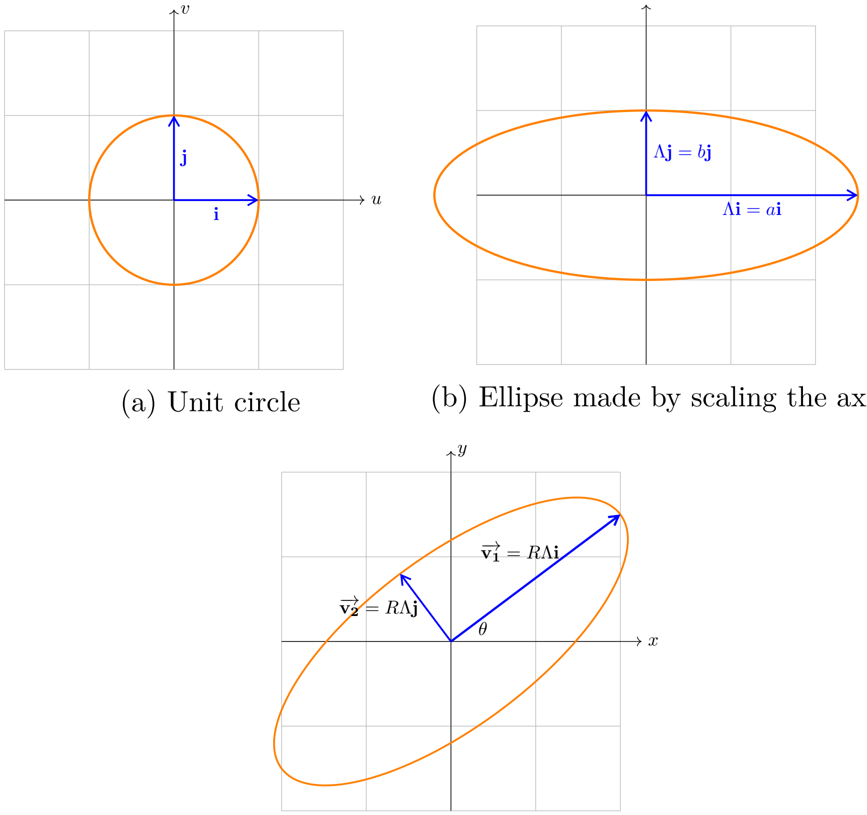
16.12 Appendix: symmetric matrices

This section is optional. We won't ask about it on psets or tests. The first example in this section is a nice exercise in thinking about matrix multiplication as a way to transform vectors.

Example 16.22. Geometry of symmetric matrices. This is a fairly complex example showing how we can use the diagonal matrix $\Lambda = \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}$ and the rotation matrix $R = \begin{bmatrix} \cos\theta & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$ to convert a circle to an ellipse as shown in the figures below.

To do this, we think of matrix multiplication as a linear transformation. The diagonal matrix Λ transforms the circle by scaling the x and y directions by a and b respectively.

This creates the ellipse in Figure (b), which is oriented with the axes. The rotation matrix R then rotates this ellipse to the general ellipse in Figure (c).


v
u
$\mathbf{j}$
$\mathbf{i}$
$\Lambda\mathbf{j} = b\mathbf{j}$
$\Lambda\mathbf{i} = a\mathbf{i}$
y
x
$\vec{\mathbf{v_1}} = R\Lambda\mathbf{i}$
$\vec{\mathbf{v_2}} = R\Lambda\mathbf{j}$
θ

(a) Unit circle

(b) Ellipse made by scaling the axes by a and b respectively

(c) Ellipse in (b) rotated by θ

In coordinates $R\Lambda$ maps the unit circle $u^2 + v^2 = 1$ to the ellipse shown in (c). That is,

$$R\Lambda \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} u \\ v \end{bmatrix} = \Lambda^{-1} R^{-1} \begin{bmatrix} x \\ y \end{bmatrix}$$

Example 16.23. Spectral theorem. The previous example transforms the unit circle in uv-coordinates into an ellipse in xy-coordinates. In terms of inner products and transposes this becomes

$$\begin{aligned} 1 &= \left\langle \begin{bmatrix} u \\ v \end{bmatrix}, \begin{bmatrix} u \\ v \end{bmatrix} \right\rangle \\ &= \left\langle \Lambda^{-1} R^{-1} \begin{bmatrix} x \\ y \end{bmatrix}, \Lambda^{-1} R^{-1} \begin{bmatrix} x \\ y \end{bmatrix} \right\rangle \\ &= \begin{bmatrix} x \\ y \end{bmatrix}^T (\Lambda^{-1} R^{-1})^T \Lambda^{-1} R^{-1} \begin{bmatrix} x \\ y \end{bmatrix} \\ &= \begin{bmatrix} x \\ y \end{bmatrix}^T R\Lambda^{-2} R^{-1} \begin{bmatrix} x \\ y \end{bmatrix} \end{aligned}$$

The last equality uses the facts that for a rotation matrix $R^T = R^{-1}$ and for a diagonal matrix $\Lambda^T = \Lambda$.

Call the matrix occurring in the last two lines above A. That is,

$$A = R\Lambda^{-2}R^{-1} = (\Lambda^{-1}R^{-1})^T\Lambda^{-1}R^{-1}.$$

We then have the equation of the ellipse is

$$1 = \begin{bmatrix} x \\ y \end{bmatrix}^T A \begin{bmatrix} x \\ y \end{bmatrix}.$$

The matrix A has the following properties

1. It is symmetric

2. Its eigenvalues are a^{-2} and b^{-2}

3. Its eigenvectors are the the vectors $\overrightarrow{\mathbf{v_1}}$ and $\overrightarrow{\mathbf{v_2}}$ along the axes of the ellipse (see figure (c) above).

4. Its eigenvectors are orthogonal.

Proof.

1. This is clear from the formula $A = B^TB$ where $B = \Lambda^{-1}R^{-1}$.

2. This is clear from the diagonalization $A = R\Lambda^{-2}R^{-1}$. (Remember the eigenvalues are in the diagonal matrix Λ^{-2}.

3. We need to show that A transforms $\overrightarrow{\mathbf{v_1}}$ to a multiple of itself. This also follows by considering the action of each term in the diagonalization in turn (see the figures): R^{-1} moves $\overrightarrow{\mathbf{v_1}}$ to the x-axis; then Λ^{-2} scales the x-axis by a^{-2}; and finally R rotates the x-axis back the line along $\overrightarrow{\mathbf{v_1}}$. Using symbols

$$A\overrightarrow{\mathbf{v_1}} = R\Lambda^{-2}R^{-1}\overrightarrow{\mathbf{v_1}} = R\Lambda^{-2}a\mathbf{i} = R(a^{-2}a\mathbf{i}) = a^{-2}\overrightarrow{\mathbf{v_1}}$$

The properties of A are general properties of symmetric matrices.

Spectral theorem. A symmetric matrix A has the following properties.

1. It has real eigenvalues.

2. Its eigenvectors are mutually orthogonal.

Because of the connection to the axes of ellipses this is also called the principal axis theorem.

17 Matrix methods for solving systems of DEs

17.1 Goals

1. Be able to solve constant coefficient linear systems using eigenvalues and eigenvectors. Do this when there are real or complex eigenvalues.

2. Understand and appreciate the abstraction of matrix notation.

3. Be able to convert a higher order linear DE equation into a *companion system* of coupled first-order equations.

4. See some physical settings modeled by systems of equations.

17.2 Introduction

In this topic we will look in detail at solving linear constant coefficient systems of differential equations using eigenvalues and eigenvectors. We will need to consider cases of real, complex and repeated eigenvalues. (We will only touch on the case of repeated eigenvalues.).

An important idea is that any higher order differential equation can be converted into a system of first-order equations. This means that our old friend $P(D)x = 0$ can be converted into a system and solved with these methods. This is useful because it is more natural to formulate numerical algorithms for first-order systems than for higher order equations. This is partly explained by the first section below, which looks at the utility of matrix notation.

17.3 Matrix notation and why we like it

We have been using matrix notation for algebraic systems and systems of differential equations. Let's remind ourselves why it's helpful in organizing our thinking.

One of the simplest algebraic equations is

$$ax = b, \quad \text{where } a \text{ and } b \text{ are constants and } x \text{ is the unknown.} \tag{31}$$

We easily solve this for x: $x = a^{-1}b$ (provided $a \neq 0$).

On the face of it a system of algebraic equations seem more complicated. For example consider the following system of two equations in two unknowns:

$$\begin{aligned} 6x + 5y &= 2 \\ x + 2y &= 3 \end{aligned}$$

We could solve this by elimination, but here our interest in writing this out abstractly. In matrix form the system and its solution become

$$\begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \end{bmatrix} \Rightarrow \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

If we give names: $A = \begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix}$, $\mathbf{x} = \begin{bmatrix} x \\ y \end{bmatrix}$, $\mathbf{b} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$ then the system and its solution become

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

At this level of abstraction we see that the system and its solution are just like those of our simplest equation. (One small difference is that we need to take more care with the order of matrix multiplication than we do with scalar multiplication.)

For differential equations our simplest and favorite equation is

$$x' = ax.$$

Written in matrix form, a linear system of DEs looks similar.

Example 17.1. As above, let $A = \begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix}$ and $\mathbf{x} = \begin{bmatrix} x \\ y \end{bmatrix}$. Write the following system in a form that resembles our favorite DE.

$$\begin{aligned} x' &= 6x + 5y \\ y' &= x + 2y \end{aligned}$$

Solution: In matrix form this becomes

$$\begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x' \\ y' \end{bmatrix} \qquad \text{or} \qquad \mathbf{x}' = A\mathbf{x}.$$

The right hand equation looks just like our favorite DE.

Note: we will call A the coefficient matrix of the system.

17.4 Solving homogeneous DEs using matrix methods

17.4.1 Review

In the previous topic we looked briefly at solving linear, homogeneous, constant coefficient systems using matrix methods. Recall that we used the method of optimism to guess a solution of the form $e^{\lambda t}\mathbf{v}$. Substituting this in the equation leads immediately to the fact that λ must be an eigenvalue and $\mathbf{v}$ an eigenvector.

We'll review the process with brief explanations. Later, we will write model solutions that skip directly to the characteristic equation.

Example 17.2. Solve $\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} 3 & 2 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$.

This is a linear, homogeneous, constant coefficient system of DEs.

Solution: Try $\begin{bmatrix} x \\ y \end{bmatrix} = e^{\lambda t}\mathbf{v}$.

Substitution gives: $\lambda e^{\lambda t}\mathbf{v} = \begin{bmatrix} 3 & 2 \\ 1 & 2 \end{bmatrix} e^{\lambda t}\,\mathbf{v} \Leftrightarrow \boxed{\begin{bmatrix} 3 & 2 \\ 1 & 2 \end{bmatrix} \mathbf{v} = \lambda\,\mathbf{v}.}$

The boxed equation is the eigenvector/eigenvalue equation, where λ is the eigenvalue and $\mathbf{v}$ is the corresponding eigenvector.

We know how to find eigenvalues and eigenvectors:

Characteristic equation: $\begin{vmatrix} 3-\lambda & 2 \\ 1 & 2-\lambda \end{vmatrix} = 0 \Leftrightarrow \lambda^2 - 5\lambda + 4 = 0 \Rightarrow \lambda = 4, 1.$

Eigenvectors are in $\text{Null}(A - \lambda I)$:

$\lambda_1 = 4$: $\quad A - \lambda I = \begin{bmatrix} -1 & 2 \\ 1 & -2 \end{bmatrix}$. Basic eigenvector: $\quad \mathbf{v_1} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$.

$\lambda_2 = 1$: $\quad A - \lambda I = \begin{bmatrix} 2 & 2 \\ 1 & 1 \end{bmatrix}$. Basic eigenvector: $\quad \mathbf{v_2} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$.

Two *modal* solutions are $\mathbf{x_1}(t) = e^{4t}\mathbf{v_1} = e^{4t}\begin{bmatrix}2\\1\end{bmatrix}$ and $\mathbf{x_2}(t) = e^{t}\mathbf{v_2} = e^{t}\begin{bmatrix}1\\-1\end{bmatrix}$.

The general solution is $\mathbf{x} = c_1\mathbf{x_1} + c_2\mathbf{x_2} = c_1e^{4t}\begin{bmatrix}2\\1\end{bmatrix} + c_2e^{t}\begin{bmatrix}1\\-1\end{bmatrix}$.

Note: Each of the solutions $\mathbf{x} = e^{\lambda t}\mathbf{v}$ is called a normal mode or modal solution.

17.4.2 Complex eigenvalues

We handle complex eigenvalues in exactly the same manner as we did complex characteristic roots for ordinary differential equations.

Theorem: Suppose A is a real matrix. Consider the DE: $\mathbf{x}' = A\mathbf{x}$.

If $\mathbf{z}$ is a complex solution to this DE then both the real and imaginary parts of $\mathbf{z}$ are also solutions.

Proof: Suppose $\mathbf{z} = \mathbf{x}_1 + i\mathbf{x}_2$ then

$$\begin{aligned}&\mathbf{z}' = A\mathbf{z}\\ \Leftrightarrow\, &(\mathbf{x}_1 + i\mathbf{x}_2)' = A(\mathbf{x_1} + i\mathbf{x_2})\\ \Leftrightarrow\, &\mathbf{x}_1' + i\mathbf{x}_2' = A\mathbf{x_1} + iA\mathbf{x_2}\end{aligned}$$

If two complex numbers are equal then their real parts must be equal and so must the imaginary parts. Therefore, the equation above shows

$$\mathbf{x}_1' = A\mathbf{x}_1 \quad \text{and} \quad \mathbf{x}_2' = A\mathbf{x}_2.$$

That is, $\mathbf{x}_1$ and $\mathbf{x}_2$ are both solutions to the DE.

Notes:
1. The proof is just linearity written out the long way.

2. To be perfectly careful we should say that $\mathbf{x}_1$ and $\mathbf{x}_2$ are the real and imaginary parts of $\mathbf{z}$, but this is clear from the context.

The next example illustrates the use of this theorem.

Example 17.3. Find the general, real-valued solution to $\begin{bmatrix}x'\\y'\end{bmatrix} = \begin{bmatrix}3 & -5\\2 & 1\end{bmatrix}\begin{bmatrix}x\\y\end{bmatrix}$.

Solution: Characteristic equation: $|A - \lambda I| = \begin{vmatrix}3-\lambda & -5\\2 & 1-\lambda\end{vmatrix} = \lambda^2 - 4\lambda + 13 = 0$

Solving, we get $\lambda = 2 \pm 3i$. (Complex roots always come in conjugate pairs.)

Eigenvectors: Find a basis for $\text{Null}(A - \lambda I)$.

$\lambda = 2 + 3i$: $(A - \lambda\, I) = \begin{bmatrix}1-3i & -5\\2 & -1-3i\end{bmatrix}$.

By inspection, a basic eigenvector is $\mathbf{v_1} = \begin{bmatrix}5\\1-3i\end{bmatrix}$.

Note: There is no need to compute the second eigenvector since it is just the complex conjugate of the first one.

This gives us a complex-valued solution

$$\begin{aligned}\mathbf{z_1}(t) = e^{(2+3i)t}\begin{bmatrix}5\\1-3i\end{bmatrix} &= e^{2t}(\cos 3t + i\,\sin 3t)\begin{bmatrix}5\\1-3i\end{bmatrix}\\ &= e^{2t}\begin{bmatrix}5\cos 3t + i5\sin 3t\\ \cos 3t + 3\sin 3t + i(-3\cos 3t + \sin 3t)\end{bmatrix}\end{aligned}$$

Just for completeness we give its complex conjugate which is also a solution

$$\mathbf{z_2}(t) = \overline{\mathbf{z}_1(t)} = e^{(2-3i)t}\begin{bmatrix}5\\1+3i\end{bmatrix} = e^{2t}\begin{bmatrix}5\cos 3t - i5\sin 3t\\ \cos 3t + 3\sin 3t - i(-3\cos 3t + \sin 3t)\end{bmatrix}$$

The theorem above tells us that The real and imaginary parts of $\mathbf{z_1}$ are both solutions:

$$\begin{aligned}\mathbf{x_1}(t) &= e^{2t}\begin{bmatrix}5\cos 3t\\ \cos 3t + 3\sin 3t\end{bmatrix}\\ \mathbf{x_2}(t) &= e^{2t}\begin{bmatrix}5\sin 3t\\ -3\cos 3t + \sin 3t\end{bmatrix}.\end{aligned}$$

As always, the general, *real-valued* solution is given by superposition

$$\mathbf{x}(t) = c_1\mathbf{x_1} + c_2\mathbf{x_2} = c_1 e^{2t}\begin{bmatrix}5\cos 3t\\ \cos 3t + 3\sin 3t\end{bmatrix} + c_2 e^{2t}\begin{bmatrix}5\sin 3t\\ -3\cos 3t + \sin 3t\end{bmatrix}.$$

17.4.3 Repeated roots (2 by 2 case only)

Repeated eigenvalues complicate matters somewhat. We will study this by looking at two examples.

Example 17.4. (Complete case) Solve $\begin{bmatrix}x'\\y'\end{bmatrix} = \begin{bmatrix}5&0\\0&5\end{bmatrix}\begin{bmatrix}x\\y\end{bmatrix}$

Solution: This is a diagonal matrix so the eigenvalues are $\lambda = 5,\ 5$.

For $\lambda = 5$ the matrix $A - \lambda I = \begin{bmatrix}0&0\\0&0\end{bmatrix}$. The null space of this matrix is all of $\mathbf{R}^2$. That is, every vector is an eigenvector i.e., the eigenspace is 2 dimensional. Since we only need to choose two independent eigenvectors, we can choose the standard basis vectors:

$$\mathbf{v_1} = \begin{bmatrix}1\\0\end{bmatrix}, \qquad \mathbf{v_2} = \begin{bmatrix}0\\1\end{bmatrix}.$$

(Any other independent pair would work as well.)

Thus the general solution to the DE is $\mathbf{x} = c_1 e^{5t}\begin{bmatrix}1\\0\end{bmatrix} + c_2 e^{5t}\begin{bmatrix}0\\1\end{bmatrix} = e^{5t}\begin{bmatrix}c_1\\c_2\end{bmatrix}$.

This is called the complete case because we have a full complement of basic solutions. That is, we have two independent solutions to our second-order system.

The next example looks at the so-called defective case. The name comes from the following ideas. If a matrix has a repeated eigenvalue we would like an independent eigenvector for each time the eigenvalue is repeated. The matrix is defective if this is not the case.

Example 17.5. (Defective case) Solve $\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} 7 & -1 \\ 4 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$

Solution: First we find the eigenvalues: The characteristic equation is

$$|A - \lambda\, I| = \lambda^2 - 10\lambda + 25 = 0.$$

So the eigenvalues are repeated: $\lambda = 5,\ 5$.

Next we find the basic eigenvectors $\mathbf{v}$. As usual, we need find to a basis of Null$(A - \lambda\, I)$.

For $\lambda = 5$: $\quad A - \lambda I = \begin{bmatrix} 2 & -1 \\ 4 & -2 \end{bmatrix}$.

The row reduced echelon form (RREF) of the coefficient matrix is $R = \begin{bmatrix} 1 & -1/2 \\ 0 & 0 \end{bmatrix}$.

This has only one free variable, so the eigenspace is only one dimensional. A basis is given by $\mathbf{v_1} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$.

This eigenvector gives us one solution to the DE: $\quad \mathbf{x_1} = e^{5t} \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

As we said, this case is defective. The system is second-order but the eigenmethods only found one solution. We'll use a magic algorithm to find a second solution. Below we'll see why the magic worked. You will need to take 18.06 (or even better 18.701) for more insight on why this works.

The first step of the algorithm is to solve $(A - \lambda I)\mathbf{v_2} = \mathbf{v_1}$. That is,

$$\begin{bmatrix} 2 & -1 \\ 4 & -2 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

Using row reduction (or by inspection) we find that one solution is $\mathbf{v_2} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$.

The algorithm now tells us that a second solution to the DE is

$$\mathbf{x_2} = te^{5t}\mathbf{v_1} + e^{5t}\mathbf{v_2} = te^{5t} \begin{bmatrix} 1 \\ 2 \end{bmatrix} + e^{5t} \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Now that we have two solutions we can give the general solution to the DE:

$$\begin{aligned} \mathbf{x}(t) &= c_1\mathbf{x_1} + c_2\mathbf{x_2} \\ &= c_1\, e^{5t} \begin{bmatrix} 1 \\ 2 \end{bmatrix} + c_2 \left(te^{5t} \begin{bmatrix} 1 \\ 2 \end{bmatrix} + e^{5t} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right) \end{aligned}$$

Abstract version of defective case

The example above is complicated by actual computations. Here is the abstract version of the algorithm for the defective case. We check that the result is a solution by plugging it into the DE.

The algorithm uses two vectors
1. An eigenvector $\mathbf{v_1}$, i.e., $\quad A\mathbf{v_1} = \lambda\mathbf{v_1}$

2. A vector $\mathbf{v_2}$ that satisfies $(A - \lambda I)\mathbf{v_2} = \mathbf{v_1}$.

$\mathbf{v_2}$ is called a generalized eigenvector. In the proof below, we will need to use this in the form: $A\mathbf{v_2} = \mathbf{v_1} + \lambda\mathbf{v_2}$.

We assert that $\mathbf{x_1}(t) = e^{\lambda t}\mathbf{v_1}$ and $\mathbf{x_2}(t) = te^{\lambda t}\mathbf{v_1} + e^{\lambda t}\mathbf{v_2}$ are independent solutions to the DE.

Proof: We know that x_1 is the eigenvector solution. To check that $\mathbf{x_2}$ is a solution, we plug it into the DE and check that both sides of the equation are the same.

$$\begin{aligned}\text{(left side) } \mathbf{x_2'} &= \lambda te^{\lambda t}\mathbf{v_1} + e^{\lambda t}\mathbf{v_1} + \lambda e^{\lambda t}\mathbf{v_2} = \lambda te^{\lambda t}\mathbf{v_1} + e^{\lambda t}(\mathbf{v_1} + \lambda\mathbf{v_2}) \\ \text{(right side)} A\mathbf{x_2} &= te^{\lambda t}A\mathbf{v_1} + e^{\lambda t}A\mathbf{v_2} = \lambda te^{\lambda t}\mathbf{v_1} + e^{\lambda t}(\mathbf{v_1} + \lambda\mathbf{v_2})\end{aligned}$$

Comparing both sides we see that $\mathbf{x_2'} = A\mathbf{x_2}$. That is, $\mathbf{x_2}$ is a solution.

17.5 Companion systems

Early in 18.03 we learned how to solve ordinary differential equations $P(D)x = 0$. For example $x'' + 8x' + 7x = 0$. In this section we will convert a higher order ordinary differential equation to a system of first-order equations.

Example 17.6. Convert the ODE $x'' + 8x' + 7x = 0$ to a system of first-order equations.

Solution: Introduce a second variable $y = x'$. Our ODE then becomes

$$y' + 8y + 7x = 0.$$

Writing out the equations for x' and y' we get

$$\begin{array}{rcrcr} x' & = & & & y \\ y' & = & -7x & - & 8y \end{array} \quad \Leftrightarrow \quad \begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -7 & -8 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

The system is called the **companion system** to the original ODE. We call the coefficient matrix the **companion matrix**.

We will sometimes refer to the method of converting an ODE to a system as **anti-elimination**. This is because elimination is a process of removing variables and equations, so anti-elimination is a process of adding variables and equations.

Example 17.7. Find the companion system for the ODE $x''' + 2x'' + 5x' + 7x = 0$.

Solution: Let $y = x'$ and $z = y' = x''$. The ODE becomes $z' + 2z + 5y + 7x = 0$. So our companion system is

$$\begin{array}{rcrcrcr} x' & = & & & y & & \\ y' & = & & & & & z \\ z' & = & -7x & - & 5y & - & 2z \end{array} \quad \Leftrightarrow \quad \begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -7 & -5 & -2 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

17.6 Physical examples

Example 17.8. Population models
Suppose we have two countries with time varying populations x and y. Suppose also that the natural growth rate in the countries is 2% and 2% respectively. In addition every year 3% of the country 1 moves to country 2 and 1% of country 2 moves to country 1.

Give a system of differential equations modeling this scenario. Assume initial populations of $x(0) = 2$ and $y(0) = 2$ (in units of one million). Solve the system and interpret the eigenvectors in terms of populations.

Solution: We have

$$\begin{aligned} x' &= 0.02x - 0.03x + 0.01y &&= -0.01x + 0.01y \\ y' &= 0.03x + 0.02y - 0.01y &&= 0.03x + 0.01y \end{aligned} \quad \Leftrightarrow \quad \begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} -0.01 & 0.01 \\ 0.03 & 0.01 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

We solve by finding eigenvalues and eigenvectors.

Characteristic equation: $\begin{vmatrix} -0.01 - \lambda & 0.01 \\ 0.03 & 0.01 - \lambda \end{vmatrix} = 0 \Rightarrow \lambda = 0.02,\ -0.02$

Eigenvectors (basis of Null$(A - \lambda I)$, where A is the coefficient matrix:

$\lambda_1 = 0.02$: $\quad A - \lambda I = \begin{bmatrix} -0.03 & 0.01 \\ 0.03 & -0.01 \end{bmatrix}$. Basic eigenvector: $\mathbf{v_1} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$

$\lambda_2 = -0.02$: $\quad A - \lambda I = \begin{bmatrix} 0.01 & 0.01 \\ 0.03 & 0.03 \end{bmatrix}$. Basic eigenvector: $\mathbf{v_2} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$

The general solution is

$$\begin{bmatrix} x(t) \\ y(y) \end{bmatrix} = c_1 e^{0.02\,t} \begin{bmatrix} 1 \\ 3 \end{bmatrix} + c_2 e^{-0.02\,t} \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

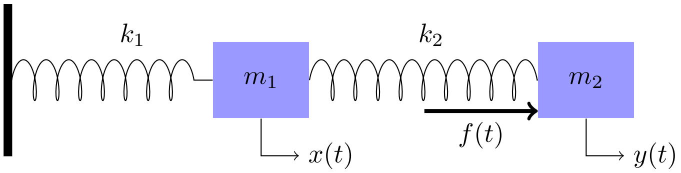
The initial conditions produce $c_1 = 1$ and $c_2 = 1$. So

$$\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = e^{0.02\,t} \begin{bmatrix} 1 \\ 3 \end{bmatrix} + e^{-0.02\,t} \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

Over time the $e^{-0.02t}$ term will go to 0 and the populations will grow exponentially and in a ratio of $x/y \approx 1/3$.

Some eigenvectors may have negative entries and some eigenvalues may be negative or complex. However, any population vector is a combination of these pure modes.

Example 17.9. Coupled springs. Suppose we have two masses and springs configured as shown.



x is the displacement of m_1 from its equilibrium position.

y is the displacement of m_2 from its equilibrium position.

(So the amount that Spring 2 is stretched is $y - x$.)

$f(t)$ is a time-varying force applied to m_2.

Using Hooke's law, we get the following system of equations

$$\begin{aligned} m_1\ddot{x} &= -k_1 x + k_2(y - x) \\ m_2\ddot{y} &= -k_2(y - x) + f(t) \end{aligned}$$

We can rearrange this to be

$$\begin{aligned} \ddot{x} &= -\frac{k_1 + k_2}{m_1}x + \frac{k_2}{m_1}y \\ \ddot{y} &= \frac{k_2}{m_2}x - \frac{k_2}{m_2}y + \frac{f(t)}{m_2} \end{aligned}$$

The system is fourth-order because it consists of 2 second-order equations. You should think about how you would produce a companion system of 4 first-order equations.

This system is illustrated by the applet https://mathlets.org/mathlets/coupled-oscillators/ (You'll have to set one of the spring constants to 0.)

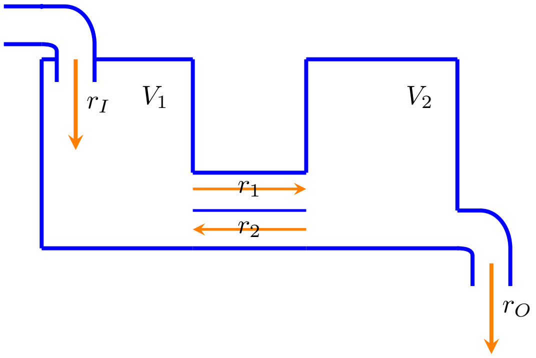
Example 17.10. Salt tanks. Suppose we have two tanks containing a salt solution. Initially the volume of water in the tanks is V_1 and V_2 respectively. Pure water flows into Tank 1 from the outside at r_I liters/minute. Solution flows out of Tank 2 at a rate of r_O liters/min. Solution is exchanged between the tanks, as shown, at the rates r_1 and r_2 in liters/min.

Suppose the rates and volumes are:

$r_I = 20$ (pure water), $r_1 = 10$, $r_2 = 30$, $r_O = 20$

$V_1 = 100$ liters, $V_2 = 200$ liters.

Note that the flow rates are balanced, so that V_1 and V_2 do not change.



Write a system of DEs modeling the amount of salt in each tank.

Solution: Let x be the grams of salt in Tank 1 and let y be the grams of salt in Tank 2.

Before starting, let's note that because pure water is being added all the salt will eventually be flushed out of the tanks, i.e., both x and $y \to 0$ in the long run. We should check that our answer reflects this.

Now for the model: x' = rate salt into Tank 1 - rate salt out of Tank 1).

rate in $=$ flow $\cdot$ concentration $= r_2 \cdot \frac{y}{V_2} = 10$ l/min $\cdot$ y g/200 l $= \frac{10}{200} y$ g/min.

rate out $= r_1 \cdot \frac{x}{V_1} = \frac{30}{100} x$ g/min.

Thus, $\quad x' = -\dfrac{3}{10} x + \dfrac{1}{20} y$

Likewise for y': rate in $= r_1 \cdot \frac{x_2}{V_2}$, rate out $= (r_2 + r_O) \cdot \frac{y}{V_2}$

So $y' = \frac{3}{10} x - \frac{3}{20} y$.

---

18 Matrix exponential, exponential and sinusoidal input

This topic is no longer on the syllabus. We post these notes for anyone who is interested. Since we have already covered inhomogeneous, constant coefficient, linear DEs and homogeneous systems, linear systems with input is not a big step.

18.1 Goals

1. Know the definition of the matrix exponential.
2. Be able to compute the matrix exponential from eigenvalues and eigenvectors.
3. Be able to use the matrix exponential to solve an IVP for a constant coefficient linear system of differential equations.
4. Be able to derive and apply the exponential response formula for constant coefficient linear systems with exponential input.
5. Be able to solve linear constant coefficient systems with sinusoidal input using complex replacement and the ERF.

18.2 Introduction

The constant coefficient system $\mathbf{x}' = A\mathbf{x}$ has a nice conceptual solution in terms of the matrix exponential e^{At}. This matrix exponential is a square matrix whose derivative follows the usual rule for exponentials:

$$\frac{de^{At}}{dt} = Ae^{At}.$$

So, as can be checked directly, the system $\mathbf{x}' = A\mathbf{x}$ has solution $\mathbf{x}(t) = e^{At}\mathbf{c}$, where $\mathbf{c}$ is a constant vector.

We'll use the diagonalization $A = S\Lambda S^{-1}$ to define the matrix exponential e^{At}. We will then use it to give another way of presenting the solutions to $\mathbf{x}' = A\mathbf{x}$.

After that, we will turn our attention to inhomogeneous linear systems of the form

$$\mathbf{x}' = A\mathbf{x} + \mathbf{F}(t). \qquad \text{(I)}$$

As usual, $\mathbf{x}$ is a column vector of (unknown) functions, A is a square constant matrix and the input $\mathbf{F}(t)$ is a column vector. As you might expect, when $\mathbf{F}(t)$ is exponential or sinusoidal we will have an exponential or sinusoidal resposnse formula. Unlike for ordinary differential equations, these formulas are not worth memorizing. It will turn out to be easier to rederive them as needed.

18.3 Matrix Exponential

In 18.03 we use the exponential function all the time. Its main property is that it helps us solve differential equations.

Example 18.1. Solve $x' = ax$

Solution: $x(t) = x(0)\, e^{at}$.

We are going to define the matrix exponential. There are several ways to do this. Since this is a differential equations class, let's define it as the solution to a DE. Then we will see various ways to compute and use it.

Definition. For any square matrix A, the matrix exponential e^{At} is the matrix of functions that satisfies the initial value problem

$$\frac{dB(t)}{dt} = A \cdot B(t), \qquad B(0) = I.$$

Note. We could also have defined e^{At} using the Taylor series for e^x

$$e^{At} = I + tA + \frac{t^2A^2}{2} + \frac{t^3A^3}{3!} + ...$$

Either definition gives the same answer.

We can now list several properties of the matrix exponential.
1. The initial value problem $\mathbf{x}' = A\mathbf{x}$ with initial value $\mathbf{x}(0) = \mathbf{b}$ has solution $e^{At}\mathbf{b}$.

2. If $\Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$ then $e^{\Lambda t} = \begin{bmatrix} e^{\lambda_1 t} & 0 \\ 0 & e^{\lambda_2 t} \end{bmatrix}$.

3. If $A = S\Lambda S^{-1}$ is the diagonalization of A then

$$e^{At} = Se^{\Lambda t}S^{-1}$$

4. $e^{A(s+t)} = e^{As}e^{At}$.

5. Definition. e^{At} is called a fundamental matrix for the system $\mathbf{x}' = A\mathbf{x}$

Warning: Because matrix multiplication does not commute, it is **not generally true** that e^Ae^B is the same as e^{A+B}. They are the same only in special cases.

Proofs. Here are proofs of these facts.

1. We need to verify that $\mathbf{x}(t) = e^{At}\mathbf{b}$ satisfies the IVP. This follows directly from our definition of matrix exponential:

$$\mathbf{x}'(t) = \frac{de^{At}\mathbf{b}}{dt} = A\mathbf{e}^{\mathbf{At}}\mathbf{b} = A\mathbf{x}(t).$$

2. $\dfrac{d}{dt}\begin{bmatrix} e^{\lambda_1 t} & 0 \\ 0 & e^{\lambda_2 t} \end{bmatrix} = \begin{bmatrix} \lambda_1 e^{\lambda_1 t} & 0 \\ 0 & \lambda_2 e^{\lambda_2 t} \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}\begin{bmatrix} e^{\lambda_1 t} & 0 \\ 0 & e^{\lambda_2 t} \end{bmatrix} = \Lambda e^{\Lambda t}.$ QED

3. We need to show that $\dfrac{d}{dt} S e^{\Lambda t} S^{-1} = A S e^{\Lambda t} S^{-1}$. We do this by computing both sides and seeing that they are equal:

Since S is constant, the left-hand side of this equation is:

$$\frac{d}{dt} S e^{\Lambda t} S^{-1} = S \frac{de^{\Lambda t}}{dt} S^{-1} = S \Lambda e^{\Lambda t} S^{-1}.$$

Replacing A by its diagonalization, the right hand side of the equation is:

$$A S e^{\Lambda t} S^{-1} = S \Lambda S^{-1} S e^{\Lambda t} S^{-1} = S \Lambda \, e^{\Lambda t} S^{-1}.$$

The two sides are the same. QED

4. This follows from the diagonalized form. To make the calculation explicit, we show it for the 2×2 case with eigenvalues λ_1, λ_2.

$$\begin{aligned} e^{As} e^{At} &= S e^{\Lambda s} S^{-1} S e^{\Lambda t} S^{-1} = S e^{\Lambda s} e^{\Lambda t} S^{-1} = S \begin{bmatrix} e^{\lambda_1 s} & 0 \\ 0 & e^{\lambda_2 s} \end{bmatrix} \begin{bmatrix} e^{\lambda_1 t} & 0 \\ 0 & e^{\lambda_2 t} \end{bmatrix} S^{-1} \\ &= S \begin{bmatrix} e^{\lambda_1 (s+t)} & 0 \\ 0 & e^{\lambda_2 (s+t)} \end{bmatrix} S^{-1} = S e^{\Lambda (s+t)} S^{-1} = e^{A(s+t)}. \quad \blacksquare \end{aligned}$$

Example 18.2. Let $A = \begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix}$ Solve the initial value problem $\mathbf{x}' = A\mathbf{x}$, $\mathbf{x}(0) = \begin{bmatrix} 3 \\ 5 \end{bmatrix}$

Solution: We know the answer is $\mathbf{x} = e^{At}\begin{bmatrix} 3 \\ 5 \end{bmatrix}$.

We also know A has eigenvalues 7, 1 and corresponding eigenvectors $\begin{bmatrix} 5 \\ 1 \end{bmatrix}$, $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$.

We can rewrite $\mathbf{x}(t) = e^{At}\begin{bmatrix} 3 \\ 5 \end{bmatrix}$ as

$$\begin{aligned} \mathbf{x}(t) = S e^{\Lambda t}\, S^{-1} \begin{bmatrix} 3 \\ 5 \end{bmatrix} &= \begin{bmatrix} 5 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} e^{7t} & 0 \\ 0 & e^{t} \end{bmatrix} \begin{bmatrix} 5 & 1 \\ 1 & -1 \end{bmatrix}^{-1} \begin{bmatrix} 3 \\ 5 \end{bmatrix} \qquad (*) \\ &= \begin{bmatrix} 5 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} e^{7t} & 0 \\ 0 & e^{t} \end{bmatrix} \begin{bmatrix} 8/6 \\ -22//6 \end{bmatrix} = \frac{8}{6} e^{7t} \begin{bmatrix} 5 \\ 1 \end{bmatrix} - \frac{22}{6} e^{t} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \end{aligned}$$

As a general rule, the line marked with the $(*)$ is a fine answer to this question.

18.4 Exponential response formula (ERF)

Exponential response formula. For a constant matrix A and a constant vector $\mathbf{k}$ the DE

$$\mathbf{x}' = A\mathbf{x} + e^{at}\mathbf{k}$$

has a particular solution:

$$\mathbf{x_p}(t) = e^{at}(aI - A)^{-1}\mathbf{k}$$

This formula is valid as long as $aI - A$ is invertible, i.e., as long as a is not an eigenvalue of A.

Proof. Not surprisingly, we discover this formula by the method of optimism. We try a solution of the form $\mathbf{x_p}(t) = e^{at}\mathbf{v}$, where $\mathbf{v}$ is a constant vector.

Plug the guess into the DE and solve for $\mathbf{v}$:

$$\mathbf{x_p}' = ae^{at}\mathbf{v} = e^{at}A\mathbf{v} + e^{at}\mathbf{k} \ \Rightarrow \ (aI - A)\mathbf{v} = \mathbf{k} \ \Rightarrow \ \mathbf{v} = (aI - A)^{-1}\mathbf{k}.$$

Thus we have found a particular solution $\mathbf{x_p}(t) = e^{at}\mathbf{v} = e^{at}(aI - A)^{-1}\mathbf{k}$. QED

Example 18.3. Find the general solution to $\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} e^{2t} \\ 3e^{2t} \end{bmatrix}$.

Solution: For ease of notation we rewrite the equation as $\mathbf{x}' = A\mathbf{x} + e^{2t}\begin{bmatrix} 1 \\ 3 \end{bmatrix}$. The exponential response formula gives us a particular solution

$$\mathbf{x_p}(t) = e^{2t}(2I - A)^{-1}\begin{bmatrix} 1 \\ 3 \end{bmatrix} = e^{2t}\begin{bmatrix} -4 & -5 \\ -1 & 0 \end{bmatrix}^{-1}\begin{bmatrix} 1 \\ 3 \end{bmatrix} = -\frac{e^{2t}}{5}\begin{bmatrix} 0 & 5 \\ 1 & -4 \end{bmatrix}\begin{bmatrix} 1 \\ 3 \end{bmatrix} = -\frac{1}{5}e^{2t}\begin{bmatrix} 15 \\ -11 \end{bmatrix}$$

We know from previous topics that the general homogeneous equation is

$$\mathbf{x_h}(t) = c_1e^{t}\begin{bmatrix} 1 \\ -1 \end{bmatrix} + c_2e^{7t}\begin{bmatrix} 5 \\ 1 \end{bmatrix}$$

By superposition the general solution to the system is $\mathbf{x}(t) = \mathbf{x_p}(t) + \mathbf{x_h}(t)$.

Example 18.4. Solve $\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} 3e^{2t} \\ 5e^{3t} \end{bmatrix}$.

Solution: Write the input as $e^{2t}\begin{bmatrix} 3 \\ 0 \end{bmatrix} + e^{3t}\begin{bmatrix} 0 \\ 5 \end{bmatrix}$. Now you can find a particular solution to the equation for a each input term and then use superposition.

There are more examples in the next section.

18.5 Exponential response formula examples

Example 18.5. Find the general solution to $\begin{array}{rcl} x' & = & 3x - y + e^{2t} \\ y' & = & 4x - y - e^{2t} \end{array}$

Solution: In matrix form the equation is $\mathbf{x}' = \begin{bmatrix} 3 & -1 \\ 4 & -1 \end{bmatrix}\mathbf{x} + e^{2t}\begin{bmatrix} 1 \\ -1 \end{bmatrix}$. The exponential response formula tells us a particular solution is

$$\mathbf{x_p}(t) = e^{2t}(2I - A)^{-1}\begin{bmatrix} 1 \\ -1 \end{bmatrix} = e^{2t}\begin{bmatrix} -1 & 1 \\ -4 & 3 \end{bmatrix}^{-1}\begin{bmatrix} 1 \\ -1 \end{bmatrix} = e^{2t}\begin{bmatrix} 3 & -1 \\ 4 & -1 \end{bmatrix}\begin{bmatrix} 1 \\ -1 \end{bmatrix} = e^{2t}\begin{bmatrix} 4 \\ 5 \end{bmatrix}.$$

We'll let you verify the calculation of the inverse. Likewise we'll let you find the homogeneous solution needed for the general solution.

Example 18.6. Find a particular solution to $\begin{array}{rcl} x' &=& 3x-y+3 \\ y' &=& 4x-y+2 \end{array}$

Solution: Note that we could get our solution using the exponential response formula, where the exponent is $a=0$. Instead, we'll just say that we're guessing a constant solution and solve for its exact value.

Try $\mathbf{x}=\mathbf{v}$. Substitution into the DE gives $\mathbf{x}'=0=A\mathbf{v}+\begin{bmatrix}3\\2\end{bmatrix}$.

So, $\mathbf{v}=-A^{-1}\begin{bmatrix}3\\2\end{bmatrix}=-\begin{bmatrix}3&-1\\4&-1\end{bmatrix}^{-1}\begin{bmatrix}3\\2\end{bmatrix}=\begin{bmatrix}1\\6\end{bmatrix}$. That is $\mathbf{x_p}(t)=\begin{bmatrix}1\\6\end{bmatrix}$.

Again, we'll let you verify the calculation of the inverse.

Example 18.7. Find a particular solution to $\begin{bmatrix}x'\\y'\end{bmatrix}=\begin{bmatrix}1&2\\2&1\end{bmatrix}\begin{bmatrix}x\\y\end{bmatrix}+\begin{bmatrix}\cos(t)\\0\end{bmatrix}$.

Solution: To use the exponential response formula, we first need to use complex replacement. The complexified equation is

$$\mathbf{z}'=\begin{bmatrix}1&2\\2&1\end{bmatrix}\mathbf{z}+e^{it}\begin{bmatrix}1\\0\end{bmatrix},\ \text{where } \mathbf{x}=\text{Re}(\mathbf{z}).$$

Now we compute the inverse to prepare for the exponential response formula:

$$(iI-A)^{-1}=\begin{bmatrix}-1+i&-2\\-2&-1+i\end{bmatrix}^{-1}=\frac{1}{-2i-4}\begin{bmatrix}-1+i&2\\2&-1+i\end{bmatrix}$$

So, $\mathbf{z_p}(t)=e^{it}(iI-A)^{-1}\begin{bmatrix}1\\0\end{bmatrix}=\dfrac{1}{-2i-4}e^{it}\begin{bmatrix}-1+i&2\\2&-1+i\end{bmatrix}\begin{bmatrix}1\\0\end{bmatrix}=\dfrac{1}{2i+4}e^{it}\begin{bmatrix}1-i\\-2\end{bmatrix}$.

To find the real part of $\mathbf{z_p}$, we work in polar coordinates. First we write the various complex numbers in polar form:

$2i+4=2\sqrt{5}e^{i\phi_1}$, where $\boxed{\phi_1=\text{Arg}(2i+4)=\tan^{-1}(1/2)}$ in the first quadrant.

Likewise $1-i=\sqrt{2}e^{i\phi_2;}$, where $\boxed{\phi_2=-\pi/4}$.

So, $\mathbf{z_p}(t)=\dfrac{-e^{it}}{2\sqrt{5}e^{i\phi_1}}\begin{bmatrix}\sqrt{2}e^{i\phi_2}\\-2\end{bmatrix}=-\dfrac{1}{2\sqrt{5}}\begin{bmatrix}\sqrt{2}e^{i(t+\phi_2-\phi_1)}\\-2e^{i(t-\phi_1)}\end{bmatrix}$.

Taking the real part:

$$\boxed{\mathbf{x_p}(t)=\text{Re}(\mathbf{z_p})=\begin{bmatrix}\sqrt{2}\cos(t+\phi_2-\phi_1)\\-2\cos(t-\phi_1)\end{bmatrix}}$$

Here is the same calculation in rectangular coordinates. I think the arithmetic is more error prone and the answer is harder to interpret.

$$\frac{1}{2i+4}\begin{bmatrix}1-i\\-2\end{bmatrix}=\frac{4-2i}{20}\begin{bmatrix}1-i\\-2\end{bmatrix}=\frac{1}{10}\begin{bmatrix}1-3i\\-4+2i\end{bmatrix}.$$

So, $\mathbf{z_p}(t)=\dfrac{1}{10}(\cos(t)+i\sin(t))\begin{bmatrix}1-3i\\-4+2i\end{bmatrix}=\dfrac{1}{10}\begin{bmatrix}\cos(t)+3\sin(t)+i(\sin(t)-3\cos(t))\\-4\cos(t)-2\sin(t)+i(-4\sin(t)+2\cos(t))\end{bmatrix}$.

Thus, $\mathbf{x_p}(t)=\text{Re}(\mathbf{z_p}(t))=\dfrac{1}{10}\begin{bmatrix}\cos(t)+3\sin(t)\\-4\cos(t)-2\sin(t)\end{bmatrix}$.

19 Fundamental matrix, variation of parameters

This topic is no longer on the syllabus. We post these notes for anyone who is interested.

19.1 Goals

1. Be able to recognize a linear non-constant coefficient system of differential equations.
2. Know the definition and basic properties of a fundamental matrix for such a system.
3. Be able to use the matrix exponential as a fundamental matrix for a constant coefficient linear system.
4. Be able to use the variation of paramters formula to solve a (nonconstant) coefficient linear inhomogeneous system.
5. Be able to use Euler's method to approximate the solution to a system of first-order equations.

19.2 Introduction

So far we have focused on homogeneous, constant coefficient linear systems. We now want to think about systems with input or with non-constant coefficients. So in this topic we will consider general linear systems of differential equations. That is, equations of the following form.

$$\mathbf{x}' = A(t)\mathbf{x} \qquad \text{(homogeneous)} \qquad \text{(H)}$$

$$\mathbf{x}' = A(t)\mathbf{x} + \mathbf{F}(t) \qquad \text{(inhomogeneous)} \qquad \text{(I)}$$

Here $\mathbf{x}(t)$ is a vector valued function, e.g., $(x(t), y(t), z(t))^T$, $A(t)$ is an $n \times n$ matrix called the coefficient matrix and $\mathbf{F}(t)$ is called the (mathematical) input to the system.

As usual, solving the system means finding the unknown vector valued function $\mathbf{x}(t)$.

A main point in this topic is to introduce the fundamental matrix, $\Phi(t)$, for a linear system of DEs. This will allow us to state the essential properties of these systems in a concise and elegant way. The fundamental matrix is available for any linear system. We will see that the matrix exponential e^{At}, introduced in a previous topic, is a fundamental matrix for the constant coefficient system $\mathbf{x}' = A\mathbf{x}$.

Next, we will look at linear equations with arbitrary input. This will lead to the variation of parameters formula for the solution. This is a beautiful formula, which uses the fundamental matrix. Since it involves integrals and can be painful or difficult to apply, we will use it as a last resort to find solutions to equations with nonconstant coefficients or unusual input.

We will conclude with a small section showing that Euler's method works for systems of first-order equations in exactly the same way as for ordinary first-order differential equations.

We start by going over the familiar ideas of linearity and existence and uniqueness.

19.3 Linearity/Superposition

As always, linear systems satisfy superposition principles. We restate them in the forms we like to use.

1. If $\mathbf{x_1}$ and $\mathbf{x_2}$ are solutions to Equation (H), then so is $\mathbf{x} = c_1\mathbf{x_1} + c_2\mathbf{x_2}$

Proof. $\mathbf{x}' = c_1\mathbf{x_1}' + c_2\mathbf{x_2}' = c_1A\mathbf{x_1} + c_2A\mathbf{x_2} = A(c_1\mathbf{x_1} + c_2\mathbf{x_2}) = A\mathbf{x}$.

2. If $\mathbf{x_h}$ is a solution to Equation (H) and $\mathbf{x_p}$ is a solution to Equation (I) then $\mathbf{x} = \mathbf{x_p} + \mathbf{x_h}$ is also a solution to Equation (I).

Proof. $\mathbf{x}' = \mathbf{x_p}' + \mathbf{x_h}' = A\mathbf{x_p} + \mathbf{F} + A\mathbf{x_h} = A(\mathbf{x_p} + \mathbf{x_h}) + \mathbf{F} = A\mathbf{x} + \mathbf{F}$.

3. If $\mathbf{x_1}' = A\mathbf{x_1} + \mathbf{F_1}$ and $\mathbf{x_2}' = A\mathbf{x_2} + \mathbf{F_2}$ then $\mathbf{x_1} + \mathbf{x_2}$ satisfies $\mathbf{x}' = A\mathbf{x} + \mathbf{F_1} + \mathbf{F_2}$

That is, superposition of inputs leads to superposition of outputs.

Proof. Just the same.

19.4 Existence and uniqueness theorem

As we've done for other types of equations, we state an existence and uniqueness theorem so that we can be sure that we have found all the solutions when we use the $x(t) = x_p(t)+x_h(t)$ paradigm.

Consider the initial value problem:

$$\mathbf{x}' = A(t)\mathbf{x} + \mathbf{F}(t), \quad \mathbf{x}(t_0) = \mathbf{x_0} \qquad \text{(IVP)}$$

The existence and uniqueness theorem says that there is exactly one solution to this equation.

Theorem. (existence and uniqueness) If $A(t)$ and $\mathbf{F}(t)$ are continuous then there exists a unique solution to the equation (IVP).

The next example illustrates that this new version of the existence and uniqueness theorem agrees with our old version for second-order linear equations.

Example 19.1. Consider the IVP $x'' + tx' + t^2x = t^3$; $x(0) = 1$, $x'(0) = 3$.

Converting this DE to a system using $y = x'$, we get:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -t^2 & -t \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} 0 \\ t^3 \end{bmatrix}, \begin{bmatrix} x(0) \\ y(0) \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}.$$

More abstractly we can write this as: $\mathbf{x}' = A\mathbf{x} + \mathbf{F}$; $\mathbf{x}(0) = \begin{bmatrix} 1 & 3 \end{bmatrix}^{\mathrm{T}}$

Since $A(t)$ and $\mathbf{F}(t)$ are continuous the existence and uniqueness for systems says there is a unique solution to the system. Now, $x(t)$ is the first entry in this solution, so there is also a unique solution to the original IVP.

Note. Previously, we had an existence and uniqueness theorem for ordinary differential equations which said exactly the same thing.

19.5 Fundamental matrix

This is an elegant bookkeeping technique which will make calculations and theorem statements much nicer. Consider the linear homogeneous system

$$\mathbf{x}' = A(t)\mathbf{x} \tag{H}$$

Suppose it is an $n \times n$ system and that we have n independent solutions $\mathbf{x_1}, \ldots \mathbf{x_n}$. We define the fundamental matrix as the matrix with columns $\mathbf{x_1}, \ldots, \mathbf{x_n}$, i.e.

$$\Phi(t) = \begin{bmatrix} \mathbf{x_1}(t) \ \mathbf{x_2}(t) \ \ldots \ \mathbf{x_n}(t) \end{bmatrix}.$$

Example 19.2. Consider the system $\mathbf{x}' = \begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix} \mathbf{x}$.

(a) Find a fundamental matrix for this system.

(b) Use the fundamental matrix to give the general solution to this system.

(c) Find the solution with initial conditions $\mathbf{x}(t_o) = \mathbf{b}$.

Solution: (a) We've used this coefficient matrix many times. We know two independent solutions to the system are

$$\mathbf{x_1} = e^t \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \qquad \mathbf{x_2} = e^{7t} \begin{bmatrix} 5 \\ 1 \end{bmatrix}.$$

So a fundamental matrix is $\Phi(t) = \begin{bmatrix} e^t & 5e^{7t} \\ -e^t & e^{7t} \end{bmatrix}$.

(b) The general solution is

$$\mathbf{x} = c_1\mathbf{x_1} + c_2\mathbf{x_2} = c_1 \begin{bmatrix} e^t \\ -e^t \end{bmatrix} + c_2 \begin{bmatrix} 5e^{7t} \\ e^{7t} \end{bmatrix} = \Phi(t) \cdot \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}.$$

(The last expression follows because matrix multiplication is a linear combination of the columns of Φ.)

(c) Now, we can use this to find the solution to the IVP with initial conditions $\mathbf{x}(t_0) = \mathbf{b}$.

$$\mathbf{x}(t) = \Phi(t) \cdot \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} \quad \Rightarrow \quad \Phi(t_0) \cdot \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \mathbf{b} \quad \Rightarrow \quad \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \Phi^{-1}(t_0)\mathbf{b}.$$

This is valid provided $\Phi^{-1}(t_0)$ exists. We will show this below.

19.5.1 Properties of the fundamental matrix

We have the following important properties of the fundamental matrix Φ.

1. $\Phi'(t) = A(t)\Phi(t)$ i.e., Φ satisfies Equation (H).
2. If $\mathbf{c}$ is a column vector, then $\Phi(t) \cdot \mathbf{c} = c_1\mathbf{x_1} + c_2\mathbf{x_2} + \ldots + c_n\mathbf{x_n}$.

3. If $A(t)$ is continuous, then $W(t) = |\Phi(t)| \neq 0$ equivalently $\Phi^{-1}(t)$ exists. (We call $W(t)$ the Wronskian of $\mathbf{x_1}, \ldots, \mathbf{x_n}$.)

Proof. (1) Before proving this, we note the following property of matrix multiplication: if B has columns $\mathbf{b_1}, \mathbf{b_2}, \ldots, \mathbf{b_n}$ then

$$AB = \begin{bmatrix} A\mathbf{b_1} & A\mathbf{b_2} & \ldots & A\mathbf{b_n} \end{bmatrix}.$$

You should make sure you understand this. (If it is confusing, work out a simple numerical example with an eye to understanding this property.)

Now (1) follows easily from this property:

$$\Phi'(t) = \begin{bmatrix} \mathbf{x}_1' & \mathbf{x}_2' & \ldots & \mathbf{x}_n' \end{bmatrix} = \begin{bmatrix} A\mathbf{x_1} & A\mathbf{x_2} & \ldots & A\mathbf{x_n} \end{bmatrix} = A \begin{bmatrix} \mathbf{x_1} & \mathbf{x_2} & \ldots & \mathbf{x_n} \end{bmatrix} = A(t)\Phi(t).$$

The second equality above follows because the $\mathbf{x_j}$ are solutions to Equation (H). The third equality is the property of matrix multiplication discussed just above.

(2) This is just a property of matrix multiplication.

(3) We will prove this by contradiction, i.e., we'll assume that for some t_0, $W(t_0) = 0$ and show that this contradicts the existence and uniqueness theorem. So suppose that $W(t_0) = 0$. This implies that $\Phi(t_0)$ has a nontrivial null space. Let $\mathbf{c} \neq 0$ be a nontrivial null vector. The contradiction is that now there are two solutions with $\mathbf{x}(t_0) = \mathbf{0}$. That is, both

$$\mathbf{x_1}(t) \equiv 0 \quad \text{and} \quad \mathbf{x_2}(t) = \Phi(t)\mathbf{c}$$

are 0 at $t = t_0$. This contradiction means that our assumption that $W(t_0) = 0$ must be false. QED

Example 19.3. Consider the system $\mathbf{x}' = \begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix} \mathbf{x}$ from Example 19.2. Show that its Wronskian is never 0.

Solution: In example 19.2 we found the fundamental matrix $\Phi(t) = \begin{bmatrix} e^t & 5e^{7t} \\ -e^t & e^{7t} \end{bmatrix}$

So the Wronskian is $W(t) = |\Phi(t)| = e^{8t} + 5e^{8t} = 6e^{8t}$, which is never 0.

Example 19.4. Again, consider the system $\mathbf{x}' = \begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix} \mathbf{x}$. Let A be the coefficient matrix. Show that the matrix exponential e^{At} is a fundamental matrix and compute its Wronskian.

Solution: To show e^{At} is a fundamental matrix, we need to show that every solution can be written as $e^{At}\mathbf{c}$ for some constant vector $\mathbf{c}$. This was shown in the Topic 18 notes.

To compute the Wronskian we use the diagonalized form of A:

$$A = S\Lambda S^{-1} = \begin{bmatrix} 1 & 5 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 7 \end{bmatrix} \begin{bmatrix} 1 & 5 \\ -1 & 1 \end{bmatrix}^{-1}.$$

So,

$$W(t) = \det(e^{At}) = \det(Se^{\Lambda t}S^{-1}) = \det(e^{\Lambda t}) = \det\left(\begin{bmatrix} e^t & 0 \\ 0 & e^{7t} \end{bmatrix}\right) = e^{8t} \neq 0.$$

19.5.2 The Wronskian of n solutions

In the above we assumed that the solutions were independent. Even if they are not, we can still define the Wronskian: Suppose $\mathbf{x_1}$, ... $\mathbf{x_n}$ are solutions to Equation (H). We call the determinant $W(t) = \det \begin{bmatrix} \mathbf{x_1} & \dots & \mathbf{x_n} \end{bmatrix}$ the Wronskian of these solutions. If $A(t)$ is continuous then the existence and uniquenss theorem implies:

(i) $W(t)$ is either always 0 or never 0.

(ii) $W(t) \neq 0 \Leftrightarrow \mathbf{x_1}, \dots, \mathbf{x_n}$ are independent.

(iii) $W(t) \neq 0 \Leftrightarrow \Phi = \begin{bmatrix} \mathbf{x_1} & \mathbf{x_2} & \dots & \mathbf{x_n} \end{bmatrix}$ is a fundamental matrix.

Conclusion: we can use the Wronskian to test for independence.

Example 19.5. Consider $x'' + p(t)x' + q(t)x = 0$, with solutions x_1, x_2. Convert this to a first-order system. Then give two solutions to the system and compute their Wronskian.

Solution: The companion system is found by setting $y = x'$. Thus the solutions x_1 and x_2 of the ordinary differential equation become the solutions $\mathbf{x_1} = \begin{bmatrix} x_1 \\ x_1' \end{bmatrix}$ and $\mathbf{x_2} = \begin{bmatrix} x_2 \\ x_2' \end{bmatrix}$ of the companion system. Using the definition of the Wronskian we have

$$W(t) = \det \begin{bmatrix} x_1 & x_2 \\ x_1' & x_2' \end{bmatrix} = x_1x_2' - x_1'x_2.$$

19.6 Variation of parameters formula

For the general, not necessarily constant coefficient, linear inhomogeneous system (I) we cannot use constant coefficient techniques like the ERF. For those cases where we have no other technique, we can try to use the variation of parameters formula. Since it involves integration, matrix inverses and matrix multiplication, it is our last choice when trying to solve an equation. Nonetheless, sometimes it's the only method available. In addition, the derivation of the formula is really very pretty.

Suppose we have a fundamental matrix $\Phi(t)$ for the *homogeneous* linear equation

$$\mathbf{x}' = A(t)\mathbf{x} \tag{H}$$

Remember this means that Φ has columns which are independent solutions to (H).

Now suppose we want to solve

$$\mathbf{x}' = A(t)\mathbf{x} + \mathbf{F}(t). \tag{I}$$

Theorem. The general solution to equation (I) is given by the variation of parameters formula

$$\mathbf{x}(t) = \Phi(t) \cdot \left(\int \Phi(t)^{-1} \cdot \mathbf{F}(t)\, dt + \mathbf{C} \right).$$

Proof. We will use a form of the method of optimism to derive this formula.

We know the general homogeneous solution is $\mathbf{x}(t) = \Phi(t) \cdot \mathbf{c}$ for a constant vector $\mathbf{c}$. The vector $\mathbf{c}$ is called a parameter. Variation of parameters is an old-fashioned way of saying let's optimistically make it a (dependent) variable $\mathbf{u}(t)$. So we try a solution of the form

$\mathbf{x}(t) = \Phi(t)\cdot\mathbf{u}(t)$. The function $\mathbf{u}(t)$ is unknown. To find it, we substitute our guess into (I) and see where the algebra leads us:

$$\Phi'\cdot\mathbf{u} + \Phi\cdot\mathbf{u}' = A\Phi\cdot\mathbf{u} + \mathbf{F}$$

So, (don't forget $\Phi' = A\Phi$.)

$$A\Phi\cdot\mathbf{u} + \Phi\cdot\mathbf{u}' = A\Phi\cdot\mathbf{u} + \mathbf{F} \qquad \Rightarrow \qquad \Phi\cdot\mathbf{u}' = \mathbf{F}.$$

This last equation is easy to solve:

$$\mathbf{u}' = \Phi^{-1}\cdot\mathbf{F} \qquad \Rightarrow \qquad \mathbf{u}(t) = \int \Phi^{-1}(t)\cdot\mathbf{F}(t)\,dt + \mathbf{C}.$$

Finally, we take this formula for $\mathbf{u}(t)$ and use it in our trial solution:

$$\mathbf{x}(t) = \Phi(t)\cdot\mathbf{u}(t) = \Phi(t)\cdot\left(\int \Phi(t)^{-1}\cdot\mathbf{F}(t)\,dt + \mathbf{C}\right). \quad \blacksquare$$

Remark. Note that the variation of parameters formula assumes you know the general homogeneous solution. It gives no help in finding this solution.

Example 19.6. Use the variation of parameters formula to solve

$$\mathbf{x}' = \begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix}\mathbf{x} + \begin{bmatrix} e^{t} \\ e^{5t} \end{bmatrix}.$$

Note. We retiterate that using the ERF is the preferred method of solving this equation. We use the variation of parameters formula here for practice.

Solution: Let's introduce some notation to save typing: $A = \begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix}$, $\mathbf{F} = \begin{bmatrix} 1 \\ t \end{bmatrix}$.

We know a fundamental matrix from an earlier example: $\Phi(t) = \begin{bmatrix} e^{t} & 5e^{7t} \\ -e^{t} & e^{7t} \end{bmatrix}$. So, $\Phi^{-1}(t) = \frac{e^{-8t}}{6}\begin{bmatrix} e^{7t} & -5e^{7t} \\ e^{t} & e^{t} \end{bmatrix}$. Calculating with the variation of parameters we get

$$\begin{aligned}
\mathbf{x} &= \Phi(t)\int \Phi^{-1}(t)\cdot\mathbf{F}(t)\,dt \\
&= \Phi(t)\int \frac{e^{-8t}}{6}\begin{bmatrix} e^{7t} & -5e^{7t} \\ e^{t} & e^{t} \end{bmatrix}\cdot\begin{bmatrix} e^{t} \\ e^{5t} \end{bmatrix}\,dt \\
&= \Phi(t)\int \frac{1}{6}\begin{bmatrix} 1-5e^{4t} \\ e^{-6t}+e^{-2t} \end{bmatrix}\,dt \\
&= \frac{1}{6}\Phi(t)\begin{bmatrix} t - \frac{5}{4}e^{4t} + c_1 \\ -\frac{1}{6}e^{-6t} - \frac{1}{2}e^{-2t} + c_2 \end{bmatrix} \\
&= \frac{1}{6}\begin{bmatrix} te^{t} - \frac{5}{4}e^{5t} - \frac{5}{6}e^{t} - \frac{5}{2}e^{5t} + c_1e^{t} + 5c_2e^{7t} \\ -te^{t} + \frac{5}{4}e^{5t} - \frac{1}{6}e^{t} - \frac{1}{2}e^{5t} - c_1e^{t} + c_2e^{7t} \end{bmatrix} \\
&= \frac{1}{6}\left(te^{t}\begin{bmatrix} 1 \\ -1 \end{bmatrix} + e^{5t}\begin{bmatrix} -15/4 \\ 3/4 \end{bmatrix} + e^{t}\begin{bmatrix} -5/6 \\ -1/6 \end{bmatrix} + c_1e^{t}\begin{bmatrix} 1 \\ -1 \end{bmatrix} + c_2e^{7t}\begin{bmatrix} 5 \\ 1 \end{bmatrix}\right).
\end{aligned}$$

Notice the homogeneous solution appearing with the constants of integration.

19.6.1 Definite integral version of variation of parameters

The equation (I) with initial condition $\mathbf{x}(t_0) = \mathbf{b}$ has definite integral solution

$$\mathbf{x}(t) = \Phi(t) \left(\int_{t_0}^{t} \Phi^{-1}(u) \cdot \mathbf{F}(u)\, du + \mathbf{C} \right) \text{ where } \mathbf{C} = \Phi^{-1}(t_0) \cdot \mathbf{b}.$$

19.7 Euler's method

Consider a first-order system with initial conditions:

$$\mathbf{x} = \mathbf{F}(\mathbf{x}, t), \qquad \mathbf{x}(t_0) = \mathbf{x_0}.$$

Euler's method for ordinary first-order DEs works without any change for this first-order systems. That is, fix a stepsize h. Then, the step from $(\mathbf{x_n},\, t_n)$ to $(\mathbf{x_{n+1}},\, t_{n+1})$ is given by

$$\mathbf{m} = \mathbf{F}(\mathbf{x_n}, t_n) \quad \Rightarrow \quad \mathbf{x_{n+1}} = \mathbf{x_n} + h\mathbf{m}, \quad t_{n+1} = t_n + h.$$

Just as for ordinary DEs, there are other, better, algorithms for choosing $\mathbf{m}$ or varying h.

Example 19.7. Consider $\begin{bmatrix} x' \\ y' \end{bmatrix} = t \begin{bmatrix} y \\ x \end{bmatrix}$, $\mathbf{x}(1) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. Let $\mathbf{x} = \begin{bmatrix} x \\ y \end{bmatrix}$ and use $h = 0.5$ to estimate $\mathbf{x}(2)$.

Solution:

n	t_n	$\mathbf{x_n}$	$\mathbf{m} = \mathbf{F}(\mathbf{x_n}, t_n)$
0	1.0	$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 1 \end{bmatrix}$
1	1.5	$\begin{bmatrix} 1 \\ 0.5 \end{bmatrix}$	$\begin{bmatrix} 0.75 \\ 1.5 \end{bmatrix}$
2	2.0	$\begin{bmatrix} 1.375 \\ 1.25 \end{bmatrix}$	

So, $\mathbf{x}(2) \approx \begin{bmatrix} 1.375 \\ 11.25 \end{bmatrix}$.

20 Step and delta functions

20.1 Goals

1. Be able to define the unit step and unit impulse functions and give their properties.

2. Be able to explain why the unit step and unit impulse functions are idealized versions of real physical phenomena.

3. Be able to compute the generalized derivative of a function with jump discontinuities.

4. Be able to compute integrals involving delta functions.

5. Be able to solve DEs with impulses as input.
6. Be able to find the pre and post-initial conditions for a physical model with impulsive input.

20.2 The unit step function

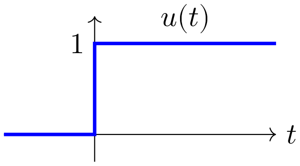
20.2.1 Definition

Let's start with the definition of the unit step function, $u(t)$:

$$u(t) = \begin{cases} 0 & \text{for } t < 0 \\ 1 & \text{for } t > 0 \end{cases}$$

We do not define $u(t)$ at $t = 0$. Rather, at $t = 0$ we think of it as in transition between 0 and 1.

It is called the unit step function because it takes a unit step at $t = 0$. It is sometimes called the Heaviside function. The graph of $u(t)$ is simple.



We will use $u(t)$ as an idealized model of a natural system that goes from 0 to 1 very quickly. In reality it will make a smooth transition, such as the following.

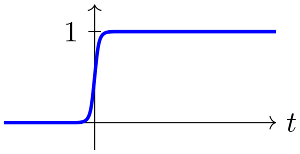


Figure 1. $u(t)$ is an idealized version of this curve

But, if the transition happens on a time scale much smaller than the time scale of the phenomenon we care about, then the function $u(t)$ is a good approximation. It is also much easier to deal with mathematically.

One of our main uses for $u(t)$ will be as a switch. It is clear that multiplying a function $f(t)$ by $u(t)$ gives

$$u(t)f(t) = \begin{cases} 0 & \text{for } t < 0 \\ f(t) & \text{for } t > 0. \end{cases}$$

We say the effect of multiplying by $u(t)$ is that for $t < 0$ the function $f(t)$ is switched off and for $t > 0$ it is switched on.

20.2.2 Integrals of $u'(t)$

From calculus we know that

$$\int u'(t)\,dt = u(t) + c \quad \text{and} \quad \int_a^b u'(t)\,dt = u(b) - u(a).$$

For example:

$$\begin{aligned}
\int_{-2}^{5} u'(t)\,dt &= u(5) - u(-2) = 1, \\
\int_{1}^{3} u'(t)\,dt &= u(3) - u(1) = 0, \\
\int_{-5}^{-3} u'(t)\,dt &= u(-3) - u(-5) = 0.
\end{aligned}$$

In fact, the following rule for the integral of $u'(t)$ over any interval is obvious

$$\int_a^b u'(t) = \begin{cases} 1 & \text{if 0 is inside the interval } (a,b) \\ 0 & \text{if 0 is outside the interval } [a,b]. \end{cases} \tag{32}$$

Note: If one of the limits is 0, we throw up our hands and refuse to do the integration.

20.2.3 0^- and 0^+

Let 0^- be infinitesimally to the left of 0 and 0^+ infinitesimally to the right of 0. That is,

$$0^- < 0 < 0^+.$$

For a function, $f(0^-)$ is defined as the left hand limit at 0 or, equivalently, the limit from below at 0, provided this limit exists. Likewise, $f(0^+)$ is the right hand limit or the limit from above.

$$f(0^-) = \lim_{t\uparrow 0} f(t) \qquad f(0^+) = \lim_{t\downarrow 0} f(t)$$

Here are some examples of integrals of u' that involve 0^- and 0^+:

$$\begin{aligned}
\int_{-\infty}^{0^+} u'(t)\,dt &= 1 \quad (\text{because } -\infty < 0 < 0^+), \\
\int_{-\infty}^{0^-} u'(t)\,dt &= 0 \quad (\text{because } -\infty < 0^- < 0), \\
\int_{0^-}^{0^+} u'(t)\,dt &= 1 \quad (\text{because } 0^- < 0 < 0^+).
\end{aligned}$$

20.3 Preview of generalized functions and derivatives

Of course $u(t)$ is not a continuous function, so, in the 18.01 sense, its derivative at $t = 0$ does not exist. Nonetheless, we saw that we could make sense of the integrals of $u'(t)$. So, rather than throw it away, we call $u'(t)$ the generalized derivative of $u(t)$. You can't do everything with $u'(t)$ you can do with an ordinary function, but we'll see that it can go anywhere we have an input function in 18.03.

20.4 The delta function (unit impulse)

20.4.1 The definition and mathematics of the delta function

Let's delve a little deeper into $u'(t)$. It's clear $u'(t) = 0$ if $t \neq 0$. At $t = 0$ the curve is vertical, so the slope is infinite, i.e., $u'(0) = \infty$. (If you think of $u(t)$ as an idealized version of the curve in Figure 1, then we would say the derivative near 0 gets very large.) We define

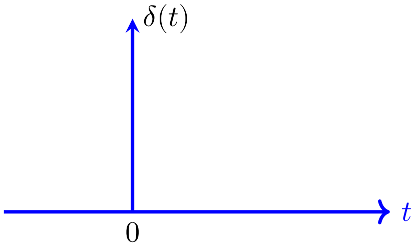
$$\delta(t) = u'(t)$$

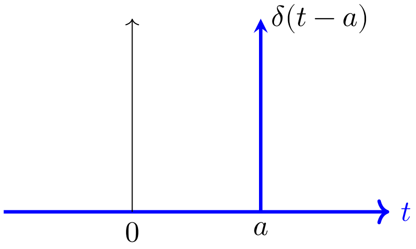
and call it the delta function or the Dirac delta function or the unit impulse function. We have seen the following properties of $\delta(t)$:

1. $\delta(t) = \begin{cases} 0 & \text{if } t \neq 0 \\ \infty & \text{if } t = 0. \end{cases}$

2. $\int \delta(t)\, dt = u(t)$ and $\int_{-\infty}^{\infty} \delta(t)\, dt = 1.$

Based on Property 1, we 'graph' $\delta(t)$ as an infinite spike at the origin and 0 everywhere else. The integrals show that the 'area' under this graph equals 1 and it is all concentrated at the origin.





We also show $\delta(t-a)$ which is just $\delta(t)$ shifted to the right.

20.5 The non-idealized delta function

Just like the unit step function, the δ function is really an idealized view of nature. In reality, a delta function is nearly a spike near 0, which goes up and down on a time interval much smaller than the scale we are working on. The integral, i.e., area under the curve, is always 1. Its graph might actually look something like

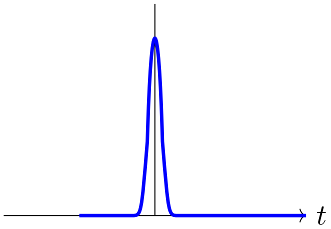


Figure 2. Non-idealized delta function; area under the graph = 1.

The total amount input is still the integral (see Section 20.7 below), or, in geometric terms, the area under the graph. A unit impulse is defined so the area is 1. Later we will consider δ as input to a physical system.

20.6 Delta functions are your friend

20.6.1 Integrals with the delta function

Recall how painful integration could be. In contrast, integrals with delta functions are always easy and involve no techniques of integration.

Suppose we scale $\delta(t)$: the integrals are just scaled.

$$\int_{-5}^{5} 3\delta(t)\,dt = 3, \quad \int_{-5}^{-3} 3\delta(t)\,dt = 0, \quad \int_{0^-}^{0^+} 3\delta(t)\,dt = 3, \quad \int_{0^+}^{\infty} 3\delta(t)\,dt = 0.$$

The integral $\int_a^b f(t)\delta(t)\,dt$ is also easy. If $f(t)$ is continuous at $t = 0$ then

$$\int_a^b f(t)\delta(t)\,dt = \begin{cases} f(0) & \text{if } (a,b) \text{ contains } 0 \\ 0 & \text{if } [a,b] \text{ does not contain } 0. \end{cases}$$

That is, integrating against $\delta(t)$ just amounts to evaluating $f(t)$ at $t = 0$.

Note 1. If one of the endpoints a or b is 0, the integral cannot be evaluated, so we just throw up our hands and refuse to do it.

Note 2. Technicality: We must have $f(t)$ continuous at $t = 0$.

20.6.2 Justification of the formula for integrating with delta functions

We should start by admitting that, in formal mathematic, this is formula is given as the definition of $\delta(t)$, so our arguments will just go to show that it is a reasonable definition. We'll do this in three ways.

Quick reason: $\delta(t)$ is 0 everywhere except $t = 0$, So $f(t)\delta(t)$ is 0 for all $t \neq 0$ and at $t = 0$ it just scales the delta function by $f(0)$. That is, $f(t)\delta(t) = f(0)\delta(t)$.

Reason 1. Since we can interpret the integral as area, we need to show that the 'area' under $f(t)\delta(t)$ is $f(0)$. Figure 2 (above) shows a tall, thin curve near $t = 0$ which approximates $\delta(t)$. Since $f(t)$ is continuous we know that $f(t) \approx f(0)$ near $t = 0$. Thus $f(t)\delta(t)$ is approximated by the graph in the Figure 2 scaled by $f(0)$. Finally, since the area under the curve in Figure 2 is one, if we scale it by $f(0)$ it will have area equal to $f(0)$. As the graph in Figure 2 gets narrower and taller it goes to the graph of $\delta(t)$. As this happens, the approximation we just made will become exact, i.e., as we wanted to show, the area under the $f(t)\delta(t) = f(0)$.

Reason 2. This is a direct argument using integration by parts. First, since $\delta(t) = 0$ for $t \neq 0$ the integral $\int_a^b f(t)\delta(t)\,dt$ must be zero for any interval $[a,b]$ not containing 0. Next, suppose $a < 0 < b$, then we get

$$\begin{aligned}\int_a^b f(t)\delta(t)\,dt &= \int_a^b f(t)u'(t)\,dt \quad (\text{since } \delta = u') \\ &= f(t)u(t)\Big|_a^b - \int_a^b f'(t)u(t)\,dt \quad (\text{integration by parts})\end{aligned}$$

Now, since $u(b) = 1$, $u(a) = 0$ and $u(t) = 0$ for $t < 0$ this becomes

$$\begin{aligned}
&= f(b) - \int_0^b f'(t)\,dt \\
&= f(b) - f(t)\big|_0^b \\
&= f(b) - f(b) + f(0) \\
&= f(0)
\end{aligned}$$

Comparing the first and last expressions in this long sequence of steps, we've shown the result.

Important note: For continuous $f(t)$, the formula

$$f(t)\delta(t) = f(0)\delta(t)$$

is extremely useful. Your life will be much easier if you learn to replace $f(t)\delta(t)$ by $f(0)\delta(t)$.

20.6.3 Shifting by a

If we shift by a, we have $\displaystyle\int_{-\infty}^{\infty} f(t)\delta(t-a) = f(a)$. More generally:

$$\int_c^d f(t)\delta(t-a)\,dt = \begin{cases} f(a) & \text{if } (c,d) \text{ contains a} \\ 0 & \text{if } [c,d] \text{ does not contain a.} \end{cases}$$

Important note: Just as for $\delta(t)$, for continuous $f(t)$ we have, $f(t)\delta(t-a) = f(a)\delta(t-a)$. You should learn to make this replacement.

Example 20.1. (Practice with δ.) Quickly cover up the answers on the right and try to evaluate each of the integrals on the left.

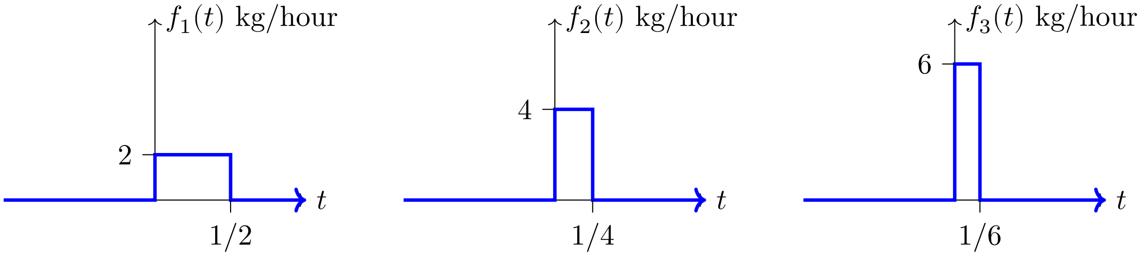
$$\begin{array}{lll}
\displaystyle\int_{-1}^{3} \delta(t)2e^{4t^2}\,dt & = 2, & \text{(evaluate } 2e^{4t^2} \text{ at } t = 0) \\
\displaystyle\int_{1}^{3} \delta(t)2e^{4t^2}\,dt & = 0, & \text{(0 is not in [1,3])} \\
\displaystyle\int_{0^-}^{3} \delta(t)2e^{4t^2}\,dt & = 2, & \text{(evaluate } 2e^{4t^2} \text{ at } t = 0) \\
\displaystyle\int_{0^-}^{\infty} \delta(t)2e^{-\tan^2(t^3)}\,dt & = 2, & \text{(evaluate } 2e^{-\tan^2(t^3)} \text{ at } t = 0) \\
\displaystyle\int_{-1}^{3} \delta(t-2)2e^{4t^2}\,dt & = 2e^{16}, & \text{(evaluate } 2e^{2e^{4t^2}} \text{ at } t = 2) \\
\displaystyle\int_{3}^{5} \delta(t-2)2e^{4t^2}\,dt & = 0, & \text{(2 is not in [3,5])} \\
\displaystyle\int_{0^-}^{3} \delta(t-2)2e^{4t^2}\,dt & = 2e^{16} & \text{(evaluate } 2e^{2e^{4t^2}} \text{ at } t = 2), \\
\displaystyle\int_{0^-}^{\infty} \delta(t-2)2e^{-\tan^2(t^3)}\,dt & = 2e^{-\tan^2(8)} & \text{(evaluate } 2e^{-\tan^2(t^3)} \text{ at } t = 2).
\end{array}$$

20.7 The physical interpretation of delta functions as a unit impulse

In general, we will be using δ functions as the input to LTI systems. So, in this subsection, we want to explore what this means. Our goal is to understand what is meant by an impulse and to see that $\delta(t)$ can be thought of as an (idealized) unit impulse.

Example 20.2. Consider the rate equation $\dot{x} + kx = f(t)$. To be specific, assume x is in kilograms of a radioactive substance and t is in hours. This is a rate equation and the derivative $\dot{x}$ and the input $f(t)$ are rates, in units of kg/hour. We then have that the total amount of substance input from time 0^- to time t is $\displaystyle\int_{0^-}^{t} f(\tau)\,d\tau$.

Consider the following possible inputs $f(t)$, shown graphically as box functions.



Look at the input function $f_1(t)$ in the leftmost figure. It is only nonzero in the interval $[0, 1/2]$ during which time it inputs at a constant rate of 2 kg/hour. The total amount input over that time is

$$\int_0^{1/2} f_1(t)\,dt = 1 \text{ kg}.$$

The function f_2 has a higher rate, but acts for a shorter time. The total amount it inputs over time is also 1 kg. The function f_3 is similar: it acts for even a shorter time, but also inputs a total of 1 kg.

If $x(0) = x_0$ kg, then over the interval $[0, 1/2]$ some of the original matter and some of what is added by $f_1(t)$ will decay away. So we'll end with something less than $x_0 + 1$ kg.

Likewise with $f_2(t)$, we add a total of 1 kg over the interval $[0, 1/4]$. Again, there will be decay over the interval, so we'll have less than $x_0 + 1$ at the end of the interval. But, since the interval is shorter, there will be less decay and the amount at the end will be closer to $x_0 + 1$ than with f_1.

If we continue to shorten the time interval in which we input a total of 1 kg, then, in the limiting case, we will dump 1 kg in all at once. In this case, there will be no time for decay and the amount will jump instantaneously from x_0 to $x_0 + 1$, after which it will start decaying. This instantaneous input is called an impulse; an instantaneous input of one unit is called a unit impulse. In a first-order system, an impulse results in an instantaneous jump in the amount of x.

Note, as the time interval gets smaller, the rate needed to add a total of 1 kg must increase. In the limit, when 1 kg is added all at once, the rate must be infinite.

It is worth acknowledging that, in a real physical system, we can't really have an ideal impulse with an infinite rate over an infinitesimal time. But we can come close by having a large rate over a very small time. As long as the time interval is tiny compared to the decay rate, the idealized impulse is a good model. For example, if we add 1 kg of radioactive material in a few seconds, while it decays on a scale of hours, then so little decays while we're adding it, that it is reasonable to model it as an impulse over an infinitesimal time interval.

Claim. Let $u_h(t)$ be the box function of width h and height $1/h$. Then the integral $\int_{-\infty}^{\infty} u_h(t)\,dt = 1$ and

$$\lim_{h\to 0} u_h(t) = \delta(t).$$

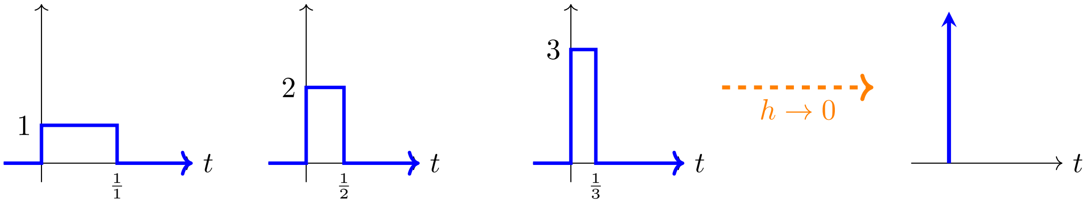
That is, as the boxes get narrower and taller they become the δ function.

Proof. We saw above that $\delta(t)$ was described by two properties

1. $\delta(t) = \begin{cases} 0 & \text{if } t \neq 0 \\ \infty & \text{if } t = 0. \end{cases}$

2. $\int \delta(t)\,dt = u(t), \qquad \int_{-\infty}^{\infty} \delta(t)\,dt = 1.$

The picture below illustrates that $\lim_{h\to 0} u_h(t)$ satisfies property 1. Because all the integrals of $u_h(t) = 1$, the second property is also true of the limit. Because the limit satisfies both properties it must equal $\delta(t)$.



A sequence of box functions $u_h(t)$ limiting to $\delta(t)$.

Summary. Here's a summary of what we've done in this subsection.

1. If $f(t)$ is an input rate. The total amount input over $[a, b]$ is $\displaystyle\int_a^b f(t)\,dt$.

2. A unit impulse adds a total of 1 unit in one instant.

3. If the impulse is at $t = t_0$ then all the input happens at $t = t_0$.

4. We can visualize an impulse as the limit of a sequence of boxes as they get narrower and taller. (Also, look back at the non-idealized delta function in Figure 2: an impulse is the limit of any spike function as it gets narrower and taller.)

5. A unit impulse is modeled by $\delta(t)$.

20.8 Solving DES: pre and post-initial conditions.

The main lesson in this section is that for an nth order equation a delta function, input causes an instantaneous jump in the $(n-1)$st derivative of the output. Once we deal with that, we can use our standard techniques to solve the DE.

Because an impulse causes an instantaneous jump in some value, we have to consider the conditions just before and just after the impulse. Assume the impulse occurs at $t = 0$, then:

At $t = 0^-$, the conditions are pre-initial conditions.

At $t = 0^+$, the conditions are post-initial conditions.

20.8.1 Impulses as input to first-order systems

Example 20.3. Solve $\dot{x} + kx = \delta(t)$ with rest initial conditions.

Solution: This is a first-order exponential decay system. The unit impulse at $t = 0$ causes an instantaneous jump of 1 in the value of x.

On $t < 0$: The DE is *always* $\dot{x} + kx = \delta(t)$. But on this interval $\delta(t) = 0$, so we can simplify the DE to

$$\dot{x} + kx = 0.$$

Since $t < 0$ our initial conditions should use 0^-: $x(0^-) = 0$.

Solving the equation we get: $x(t) = ce^{-kt}$.

Using the initial condition we get: $x(0^-) = c = 0$.

So, $\boxed{\text{on } t < 0,\ x(t) = 0}$. (This should have been obvious to us!)

On $t > 0$: The DE is *always* $\dot{x} + kx = \delta(t)$. But, on this interval $\delta(t) = 0$, so we can simplify the DE to

$$\dot{x} + kx = 0.$$

Since $t > 0$ our initial conditions should use 0^+: The pre-initial condition is $x(0^-) = 0$. The effect of the unit impulse is to cause the value of x to jump by 1 at $t = 0$. That is, $x(0^+) = 1$.

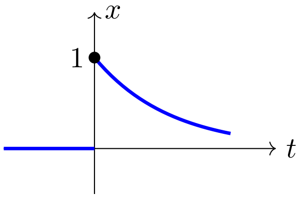
Solving the equation we get: $x(t) = ce^{-kt}$.

Using the initial condition we get: $x(0^+) = c = 1$. So, $\boxed{\text{on } t > 0,\ x(t) = e^{-kt}}$.

The full solution is

$$x(t) = \begin{cases} 0 & \text{for } t < 0 \\ e^{-kt} & \text{for } t > 0. \end{cases}$$

Here is the graph. Note the jump at $t = 0$, followed by exponential decay.



Response from rest to input = $\delta(t)$.

Key: We highlight one key thing to remember in the example above:

In each of the cases $\delta(t) = 0$. That is, when $t < 0$ we have $\delta(t) = 0$. Likewise, when $t > 0$ we have $\delta(t) = 0$.

20.8.2 Impulses as input to second-order systems

Here will give physical reasons for the jump an impulse causes in the first derivative of a second-order system. Later, in Section 20.11, we'll give algebraic reasons for the jump in a system of any order.

Now let's consider the second-order system

$$m\ddot{x} + b\dot{x} + kx = f(t), \tag{33}$$

with input $f(t)$ and output $x(t)$. To be specific, we'll think of this as a spring-mass-damper system with x in meters, t in seconds, and m in kg.

We need to think about the units on $f(t)$. It's clear enough that they are in Newtons, but what are the units of the total input $\int_a^b f(t)\,dt$? Newtons can be written as

$$\text{Newton} = \frac{\text{kg}\cdot\ \text{m/sec}}{\text{sec}} = \frac{\text{momentum}}{\text{time}}.$$

That is, force changes momentum over time. We see that the total input has units of momentum.

Following this idea, we see that a unit impulse to this second-order system is a sudden blow, i.e., a large force acting with a short duration, that causes the momentum to jump by one unit.

Example 20.4. Suppose a unit impulse is applied to the system in Equation 33. If the system is at rest before time 0, find the pre- and post-initial conditions.

Solution: Since the system is initially at rest the pre-initial conditions are

$$x(0^-) = 0 \quad \text{and} \quad \dot{x}(0^-) = 0.$$

Since, for this system, the impulse causes a one unit jump in momentum at $t = 0$ we have, at $t = 0^+$, the momentum $m\dot{x}(0^+) = 1$, i.e., the post-initial conditions

$$x(0^+) = 0 \quad \text{and} \quad \dot{x}(0^+) = 1/m.$$

Example 20.5. Assume rest initial conditions and solve the equation

$$2\ddot{x} + 7\dot{x} + 3x = \delta(t).$$

Solution: Following Example 20.4, the post-initial conditions are $x(0^+) = 0$ and $\dot{x}(0^+) = 1/2$. We work on the intervals $t < 0$ and $t > 0$ separately.

On $t < 0$: The input $\delta(t) = 0$, so we have a homogeneous DE with initial conditions

$$2\ddot{x} + 7\dot{x} + 3x = 0, \quad x(0^-) = 0,\ \dot{x}(0^-) = 0.$$

You can easily check that the solution to this is $x(t) = 0$.

So, $\boxed{\text{on } t < 0,\ x(t) = 0.}$

On $t > 0$: The input $\delta(t) = 0$, so we have a homogeneous DE with initial conditions

$$2\ddot{x} + 7\dot{x} + 3x = 0, \quad x(0^+) = 0,\, \dot{x}(0^+) = 1/2.$$

The characteristic roots are $-1/2$ and -3, so

$$x(t) = c_1 e^{-t/2} + c_2 e^{-3t}.$$

Using the initial conditions we find $c_1 = 1/5$ and $c_2 = -1/5$.

So, $\boxed{\text{on } t > 0,\ x(t) = \frac{1}{5}e^{-t/2} - \frac{1}{5}e^{-3t}.}$

The full solution is

$$x(t) = \begin{cases} 0 & \text{for } t < 0 \\ \frac{1}{5}e^{-t/2} - \frac{1}{5}e^{-3t} & \text{for } t > 0. \end{cases}$$

Example 20.6. Solve $4\ddot{x} + x = \delta(t)$ with rest IC.

Solution: The pre-initial conditions are 0, so the post-initial conditions are

$$x(0^+) = 0, \qquad \dot{x}(0^+) = 1/4.$$

On $t < 0$: The differential equation with initial conditions is

$$4\ddot{x} + x = 0; \qquad x(0^-) = 0,\, \dot{x}(0^-) = 0.$$

The solution to this is $x(t) = 0$.

On $t > 0$: The differential equation with initial conditions is

$$4\ddot{x} + x = 0; \qquad x(0^+) = 0,\, \dot{x}(0^+) = 1/4.$$

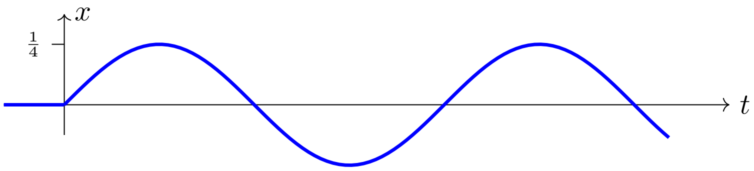
We know the solution to this:

$$x(t) = c_1 \cos(t/2) + c_2 \sin(t/2).$$

We find c_1 and c_2 to match the post-initial conditions: $c_1 = 0$, $c_2 = 1/2$. Therefore, the complete solution is

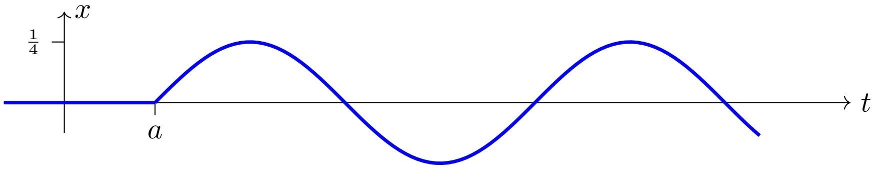
$$x(t) = \begin{cases} 0 & \text{for } t < 0 \\ \frac{1}{2}\sin(t/2) & \text{for } t > 0. \end{cases}$$

Physical explanation. At $t = 0$ an impulse kicks the simple harmonic oscillator into motion. After that, input is 0 and the system is in simple harmonic motion. The jump in momentum corresponds to the corner in graph at 0.



Example 20.7. Solve $4\ddot{x} + x = \delta(t - a)$ with rest IC.

Solution: This is an LTI system, so shifting the input from the previous example a units to the right, shifts the response in the same way.



Example 20.8. (Resonance) Solve the equation $\ddot{x} + x = f(t)$ with rest IC, where the input $f(t)$ is an impulse every 2π seconds of magnitude 3 in the positive direction.

Solution: We have $f(t) = 3\delta(t) + 3\delta(t-2\pi) + 3\delta(t-4\pi) + ...$ We can solve by solving the DE individually for each input:

$$\ddot{x}_n + x_n = 3\delta(t - 2n\pi)$$

and using superposition. (Note carefully that the rest IC are preserved by superposition. If we did not have rest IC, we would have to be a little more fussy.) The individual equations are exactly like the previous example. We get that the solution to $\ddot{x}_n + x_n = 3\delta(t-2n\pi)$ is

$$x_n(t) = \begin{cases} 0 & \text{for } t < 2n\pi \\ 3\sin(t - 2n\pi) = 3\sin(t) & \text{for } t > 2n\pi \end{cases}$$

Now, when we superposition these solutions, we see that every 2π seconds we add another copy of $3\sin(t)$ to the output. We call this resonance –the blows come at the natural frequency (every 2π seconds) of the system.

$$x(t) = \begin{cases} 0 & \text{for } t < 0 \\ 3\sin(t) & \text{for } 0 < t < 2\pi \\ 6\sin(t) & \text{for } 2\pi < t < 4\pi \\ 9\sin(t) & \text{for } 4\pi < t < 6\pi \\ & \dots \end{cases}$$

20.8.3 Impulses as input to third-order systems

Example 20.9. Assume rest initial conditions and solve the equation

$$4(D-1)(D-2)(D-3)x = 4x''' - 24x'' + 44x' - 24x = 5\delta(t).$$

(We give the differential operator in factored form so we can find the characteristic roots easily.)

Solution: For a third-order DE, the jump caused by the impulse follows the same pattern as in the second-order case. That is, the input $5\delta(t)$ causes a jump of 5 in $4x''(t)$ at $t = 0$. Here, the factor of 4 is the coefficient of x''' in the DE. Thus x'' has a jump of $5/4$. The pre-initial conditions are all zero, so after the jump the post-initial conditions are

$$x(0^+) = 0, \quad x'(0^+) = 0, \quad x''(0^+) = 5/4.$$

(In Section 20.11 we will show why this has to be the case.)

On $t < 0$: On this interval, the input $5\delta(t) = 0$. So the differential equation with initial conditions is

$$4(D-1)(D-2)(D-3)x = 0, \quad x(0^-) = 0,\, x'(0^+) = 0,\, x''(0^+) = 0.$$

The solution to this is $x(t) = 0$.

On $t > 0$: We have the homogeneous DE with initial conditions:

$$4(D-1)(D-2)(D-3)x = 0, \quad x(0^+) = 0,\, x'(0^+) = 0,\, x''(0^+) = 5/4.$$

The characteristic roots are 1, 2 and 3, so for $t > 0$ we have

$$x(t) = c_1 e^t + c_2 e^{2t} + c_3 e^{3t}.$$

Using the initial conditions to find the coefficients, we get: $c_1 = \dfrac{5}{8}$, $c_2 = -\dfrac{5}{4}$, $c_3 = \dfrac{5}{8}$.

The full solution is

$$x(t) = \begin{cases} 0 & \text{for } t < 0 \\ \frac{5}{8}e^t - \frac{5}{4}e^{2t} + \frac{5}{8}e^{3t} & \text{for } t > 0. \end{cases}$$

20.9 Box vs. delta as input

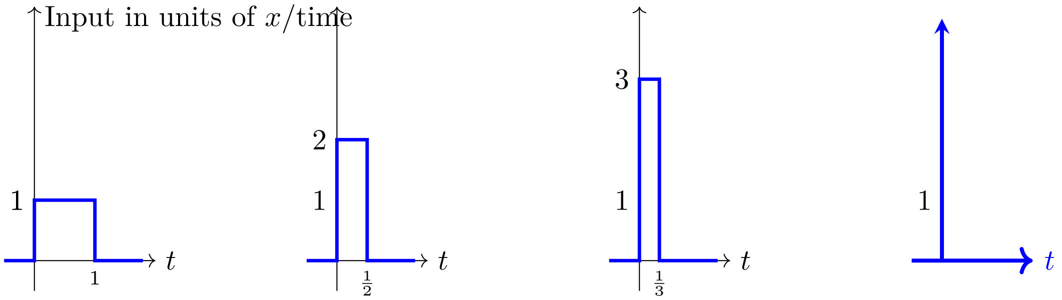
In this section we will compare box functions and delta functions as input. You will see that the delta function is much easier to work with!

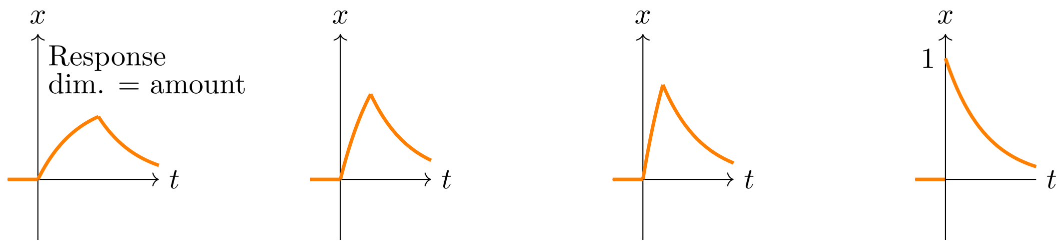
Example 20.10. (Box vs. delta.) Let's compare box $u_h(t)$ input with unit impulse ($\delta(t)$) input by solving: $\dot{x} + kx = u_h$ with rest IC.

(*Physical reasoning:*) This models radioactive dumping. u_h is the rate matter is added over time and, as we have seen, the total amount added is $\displaystyle\int_0^h u_h = 1$.

In the figure below the top row of graphs show the input u_h for various values of h. The corresponding responses are shown in the second row of graphs. The total amount input is one, so, since there is decay, at the end of the input interval, we have $x(h) < 1$. After time $t = h$ there is no more input and the response shows exponential decay.

As h goes to 0 the input becomes the unit impulse $\delta(t)$. This is shown in the last graph. Since the input is dumped in all at once the graph jumps from 0 to 1 at $t = 0$. After $t = 0$ the graph is that of exponential decay.





Top: a sequence of box function inputs limiting to $\delta(t)$.
Bottom: response to the sequence of box functions limiting to response to $\delta(t)$.

For completeness we give the exact solution to the IVP $\dot{x} + kx = u_h$ with rest IC.

$$x = \begin{cases} \frac{1}{hk}(1 - e^{-kt}) & \text{for } 0 < t < h \\ \frac{1}{hk}(e^{kh} - 1)e^{-kt} & \text{for } h < t \end{cases}$$

Just as expected, as $h \to 0$ the input becomes δ and the output becomes $x = e^{-kt}$ (i.e., $\displaystyle\lim_{h\to 0} \frac{e^{kh} - 1}{hk} = 1$)

20.10 Generalized derivatives

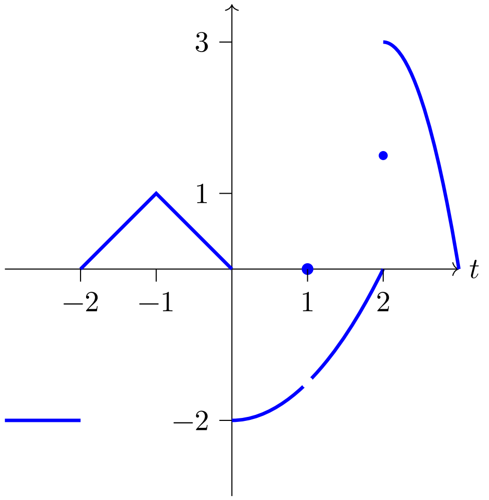
So far we have only one generalized derivative: $\dot{u}(t) = \delta(t)$. In this section we will learn to compute them for any function with jump discontinuities.

Definition. We say a function $f(t)$ has a jump discontinuity at $t = t_0$ if its graph is continuous on both the left and right, and there is a jump at t_0.

Formally this means that both left and right limits $\displaystyle\lim_{t\uparrow t_0^-} f(t)$ and $\displaystyle\lim_{t\downarrow t_0^+} f(t)$ exist, but are different. The jump at t_0 is defined as the difference

$$\lim_{t\to t_0^+} f(t) - \lim_{t\to t_0^-} f(t)$$

Example 20.11. The graph of a function $f(t)$ is shown below. It has jump discontinuities at -2, 0 and 2. The jumps are respectively 2, -2 and 3. The graph also has a corner at -1. That is, the graph is continuous at $t = -1$, but the derivative has a jump there.



Notes. 1. Not all discontinuities result in jumps. At $t = 1$ the jump between the left and right limits is 0. You could say the function jumps from -1.5 to 0 and back to -1.5 for a net jump of 0.

2. The value of $f(2)$ (represented by a dot on the graph) did not play a role in the value of the jump at $t = 2$. The jump is the size of gap between the left and right branches of the curve. You could say the function jumps from 0 to 1.5 to 3 for a net jump of 3.

3. At $t = 0$ the jump is negative because the right branch of the graph is below the left branch.

Generalized derivative: If a function is smooth except for some jump discontinuities and corners then its generalized derivative is:

- the regular derivative away from the jumps and corners.
- delta functions at the jumps. The coefficient on the delta function is the size of the jump.
- undefined at the corners.

Reason. Just as with the unit step function, the graph has 'infinite' slope at a jump and the integral of the derivative should give the original function. This is exactly what δ functions do at jumps.

Example 20.12. Suppose

$$f(t) = \begin{cases} -2 & \text{for } t < -2 \\ t+2 & \text{for } -2 < t < -1 \\ -t & \text{for } -1 < t < 0 \\ t^2/2 - 2 & \text{for } 0 < t < 2 \\ 3 - 3(t-2)^2 & \text{for } 2 < t. \end{cases}$$

Find the generalized derivative $f'(t)$.

Solution: We just take the regular derivative and add delta functions at the jump discontinuities. Note that the corner when $t = -1$ becomes a jump in the derivative.

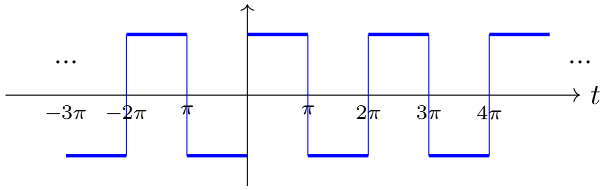
$$f'(t) = \underbrace{2\delta(t+2) - 2\delta(t) + 3\delta(t-3)}_{\text{singular part}} + \underbrace{\begin{cases} 0 & \text{for } t < -2 \\ 1 & \text{for } -2 < t < -1 \\ -1 & \text{for } -1 < t < 0 \\ t & \text{for } 0 < t < 2 \\ -6(t-2) & \text{for } 2 < t. \end{cases}}_{\text{regular part}}$$

Vocabulary: We name the two parts of the generalized derivative. The part which is the regular derivative is called the regular part and the part with delta functions due to the jumps is called the singular part. These are labeled in the example above.

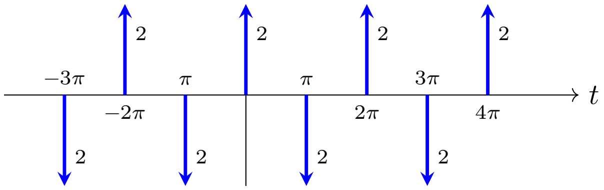
Example 20.13. Derivative of a square wave

The graphs below are of a function $\text{sq}(t)$ (called a square wave) and its derivative. The function alternates every π seconds between ± 1. The derivative $\text{sq}'(t)$ is clearly 0 everywhere except at the jumps. A jump of $+2$ gives a (generalized) derivative of 2δ and a jump of -2 gives a (generalized) derivative of -2δ. Thus we have

$$\text{sq}'(t) = \ldots + 2\delta(t+2\pi) - 2\delta(t+\pi) + 2\delta(t) - 2\delta(t-\pi) + 2\delta(t-2\pi) - 2\delta(t-3\pi) + \ldots$$



Graph of $\text{sq}(t)$ = square wave



Graph of $\text{sq}'(t)$ = impulse train

Note that we put the weight of each delta function next to it. We use the convention that $-2\delta(t)$ is represented by a downward arrow with the weight 2 next to it. That is, the sign is represented by the direction of the arrow, so the weight is positive.

20.11 Generalized derivative: checking solutions, explanation for jumps in post-initial conditions

In this section we will check the answers to a few of our previous examples by plugging them into the original DE. This should give you a feel for how a delta function as input causes a jump in the $(n-1)$st derivative of an nth-order equation.

Example 20.14. (Check the solution in Example 20.3)

The DE $\dot{x} + kx = \delta(t)$ has solution $x(t) = \begin{cases} 0 & \text{for } t < 0 \\ e^{-kt} & \text{for } t > 0. \end{cases}$.

This has a jump of 1 at $t = 0$, so $\dot{x}(t)$ is a generalized derivative:

$$\dot{x}(t) = \delta(t) + \begin{cases} 0 & \text{for } t < 0 \\ -ke^{-kt} & \text{for } t > 0 \end{cases}$$

We now check:

$$\dot{x} + kx = \left(\delta + \begin{cases} 0 & \text{for } t < 0 \\ -ke^{-kt} & \text{for } t > 0 \end{cases}\right) + k\left(\begin{cases} 0 & \text{for } t < 0 \\ e^{-kt} & \text{for } t > 0 \end{cases}\right) = \delta(t).$$

Notice that the jump in x yielded a delta function in $\dot{x}$.

Example 20.15. (Check Example 20.5) Here the DE was $2\ddot{x} + 7\dot{x} + 3x = \delta(t)$ and the solution was

$$x(t) = \begin{cases} 0 & \text{for } t < 0 \\ \frac{1}{5}e^{-t/2} - \frac{1}{5}e^{-3t} & \text{for } t > 0. \end{cases}$$

$x(t)$ has no jump at $t = 0$, so it has a regular derivative

$$\dot{x}(t) = \begin{cases} 0 & \text{for } t < 0 \\ -\frac{1}{10}e^{-t/2} + \frac{3}{5}e^{-3t} & \text{for } t > 0. \end{cases}$$

Since $\dot{x}(t)$ has a jump of $1/2$ at $t = 0$, we will get a δ function in $\ddot{x}(t)$:

$$\ddot{x}(t) = \frac{1}{2}\delta(t) + \begin{cases} 0 & \text{for } t < 0 \\ \frac{1}{20}e^{-t/2} - \frac{9}{5}e^{-3t} & \text{for } t > 0. \end{cases}$$

It is now easy to check that $2\ddot{x} + 7\dot{x} + 3x = \delta(t)$.

In particular, note that $2\ddot{x}(t) = \delta(t) +$ regular part. This explains why, in Example 20.5 we wanted $\dot{x}$ to jump by $1/2$, i.e., then $\ddot{x}$ had singular part $\delta(t)/2$, so $2\ddot{x}$ had singular part $\delta(t)$, which is needed for the left hand side of the DE to equal $\delta(t)$.

Example 20.16. (Check Example 20.9) We will do this check more quickly than the previous two. Also, we will leave out the case $t < 0$ since it is always 0. As we do the computation, notice that $x(0^-) = x(0^+)$ and $x'(0^-) = x'(0^+)$, so there is no jump until $x''(0^-) = 0$ and $x''(0^+) = 5/4$. Thus $\delta(t)$ appears in $x'''(t)$ and the jump is such that $4x'''(t) = 5\delta(t) +$ regular part.

To check the solution, we compute each term in the DE:

$$\begin{aligned} -24x &= -24\left(\frac{5}{8}e^{t} - \frac{5}{4}e^{2t} + \frac{5}{8}e^{3t}\right) \\ 44x' &= 44\left(\frac{5}{8}e^{t} - \frac{5}{2}e^{2t} + \frac{15}{8}e^{3t}\right) \\ -24x'' &= -24\left(\frac{5}{8}e^{t} - 5e^{2t} + \frac{45}{8}e^{3t}\right) \\ 4x''' &= 4\left(\frac{5}{8}e^{t} - 10e^{2t} + \frac{135}{8}e^{3t} + \frac{5}{4}\delta(t)\right) \end{aligned}$$

Adding this up verifies that $x(t)$ is a solution to the DE: $4x''' - 24x'' + 44x' - 24x = 5\delta(t)$.

21 Fourier series introduction

21.1 Goals

1. Know the definition and terminology of Fourier series.

2. Know how to compute the Fourier coefficients for a given periodic function.

21.2 Introduction

So far in 18.03 we have spent a lot of time solving the constant coefficient differential equation $P(D)x = f(t)$, where $f(t)$ is sinusoidal. Our main goal in the next few topics is to extend this to solve $P(D)x = f(t)$, where $f(t)$ can be any periodic function. The outline of our plan is fairly simple:

1. Fourier series. If $f(t)$ is periodic, we'll see that we can write it as a superposition of sine and cosine functions. For example,
$$f(t) = 1 + \cos(t)/2 + \cos(2t)/4 + \cos(3t)/8 + ...$$
2. Linearity. Then we can use the sinusoidal response formula for each term and use the superposition principle to solve $P(D)x = f(t)$.

21.3 Terminology

21.3.1 Frequency terminology

Angular frequency or circular frequency is in radians/sec or more generally in radians/time.

Frequency is in cycles/time –often in cycles/sec = hertz.

Example 21.1. Consider the function $\cos(3t)$, where t is in seconds.
The angular frequency is $\omega = 3\,\text{rad/sec}$.
The frequency is $f = \omega/(2\pi) = 3/(2\pi)\,\text{hz}$.
The period is $T = 1/f = 2\pi/\omega = 2\pi/3\,\text{sec}$.

Depending on the context, we will use frequency or angular frequency. To make matters messier we will often say frequency when we mean angular frequency.

21.3.2 Fourier series terminology

Here we give an example Fourier series. We'll use it to define the terminology we'll be using.

Example 21.2. Suppose we have
$$f(t) = \frac{3}{2} + \cos(\pi t) + \frac{\cos(2\pi)t}{2} + \frac{\cos(3\pi t)}{3} + ... + \sin(\pi t) + \frac{\sin(2\pi t)}{2^2} + \frac{\sin(3\pi t)}{3^2} + ...$$
This is a Fourier series. It has the following properties.

1. A Fourier series is sum of sines and cosines. All of the terms have a common period. In this example, every term has period 2. (Most terms also have a smaller period.)
2. It has a base angular frequency also called the fundamental angular frequency. In this example, the base frequency is π.
3. The frequency in each term is a multiple of the base frequency.
4. The base period corresponds to the base frequency. In this example, the base period is 2.
5. The Fourier coefficients are the coefficients of the sine and cosine terms. In this example, the cosine coefficients are
$$1, \frac{1}{2}, \frac{1}{3}, \dots, \frac{1}{n}, \dots$$
and the sine coefficients are
$$1, \frac{1}{2^2}, \frac{1}{3^2}, \dots, \frac{1}{n^2}, \dots$$
The constant term is $\frac{3}{2}$. This is called the DC term. DC stands for direct current.

21.4 Periodic functions

Definition. $f(t)$ is periodic with period $p > 0$ if $f(t+p) = f(t)$ for all t.

Examples: $\cos(t)$ has period 2π, $\cos(3t)$ has period $2\pi/3$, $\cos\left(\frac{\pi}{L}t\right)$ has period $2L$

Important point. Just like a complex number has multiple arguments, a periodic function has multiple periods. For example, $\cos(t)$ has periods 2π, 4π, 6π, ...

An even more extreme case is the constant function $f(t) = 1$ which has period p for any $p > 0$.

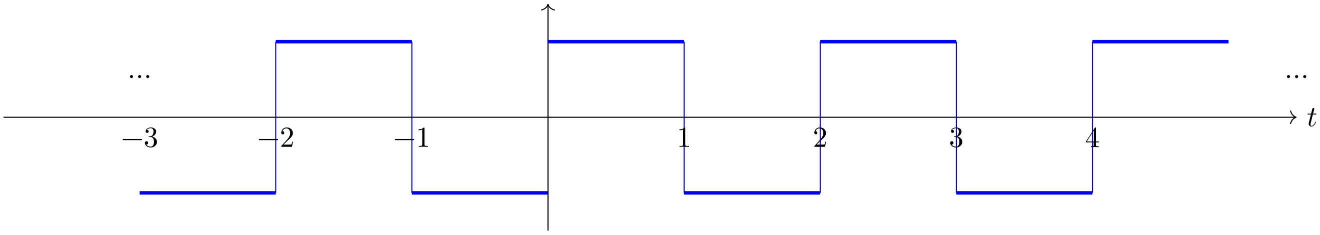
Specifying a periodic function. For a periodic function, it's enough to specify two things:

1. The period

2. The values of the function over 1 period.

Example 21.3. Period 2 square wave: period $= 2$, over one period $f(t) = \begin{cases} -1 & \text{for } -1 < t < 0 \\ 1 & \text{for } 0 < t < 1 \end{cases}$

We can now plot the $f(t)$ by plotting the period given and then shifting that one period at a time.



Graph of $f(t)$ = period 2 square wave

21.5 Fourier's theorem

Theorem (Fourier): Suppose $f(t)$ has period $p = 2L$ then

1. We can write $f(t)$ as a Fourier series

$$\begin{aligned} f(t) \quad &\sim \quad \frac{a_0}{2} + a_1\cos\left(\frac{\pi}{L}t\right) + a_2\cos\left(2\frac{\pi}{L}t\right) + a_3\cos\left(3\frac{\pi}{L}t\right) + \cdots \\ &\quad\ + b_1\sin\left(\frac{\pi}{L}t\right) + b_2\sin\left(2\frac{\pi}{L}t\right) + b_3\sin\left(3\frac{\pi}{L}t\right) + \cdots \\ &= \quad \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n\cos\left(n\frac{\pi}{L}t\right) + \sum_{n=1}^{\infty} b_n\sin\left(n\frac{\pi}{L}t\right) \end{aligned}$$

(We use $\sim$ instead of an equal sign because the two sides might differ for a few values of t. From now on, we will simply use an equal sign and ignore this fact.)

2. The base period of $f(t)$ is $2L$. Each term in the Fourier series has period $2L$. For example, $\cos\left(3\frac{\pi}{L}t\right)$ has period $2L/3$, but also $4L/3$, $\boxed{6L/3 = 2L}$

3. The series has base angular frequency $= \omega = \dfrac{\pi}{L}$. Every term in the series has an angular frequency which is a multiple of $\frac{\pi}{L}$.

4. The Fourier coefficients are given by:

$$\begin{aligned} a_0 &= \frac{1}{L}\int_{-L}^{L} f(t)\,dt \\ a_n &= \frac{1}{L}\int_{-L}^{L} f(t)\cos\left(n\frac{\pi}{L}t\right)\,dt \\ b_n &= \frac{1}{L}\int_{-L}^{L} f(t)\sin\left(n\frac{\pi}{L}t\right)\,dt \end{aligned}$$

Notes

1. All Fourier terms $a_n \cos\left(n\frac{\pi}{L}t\right)$, $b_n \sin\left(n\frac{\pi}{L}t\right)$ have period $2L$.

2. Accept the formulas for Fourier coefficients for now, we will give more explanation later.

3. The integrals must be over one full period. We showed them from $-L$ to L because that is the interval of integration we use most often. Sometimes, the function is defined in a way that makes the interval from 0 to $2L$ a better choice.

Example 21.4. Compute the Fourier series for the square wave of period $= 2\pi$,

$$\mathrm{sq}(t) = \begin{cases} -1 & \text{for } -\pi < t < 0 \\ 1 & \text{for } 0 < t < \pi. \end{cases}.$$

Solution: The half period $L = \pi$. So, for $n \neq 0$, we have

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\cos(nt)\,dt = \int_{-\pi}^{0} -\cos(nt)\,dt + \int_{0}^{\pi} \cos(nt)\,dt = -\frac{\sin(nt)}{n\pi}\bigg|_{-\pi}^{0} + \frac{\sin(nt)}{n\pi}\bigg|_{0}^{\pi} = 0.$$

For $n = 0$, $a_0 = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\,dt = 0$. (Always compute a_0 separately.)

Likewise

$$\begin{aligned} b_n &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\sin(nt)\,dt = \frac{1}{\pi}\int_{-\pi}^{0} -\sin(nt)\,dt + \frac{1}{\pi}\int_{0}^{\pi} \sin(nt)\,dt \\ &= \frac{\cos(nt)}{n\pi}\bigg|_{-\pi}^{0} - \frac{\cos(nt)}{n\pi}\bigg|_{0}^{\pi} \\ &= \frac{1-\cos(-n\pi)}{n\pi} - \frac{\cos(n\pi)-1}{n\pi} = \frac{2}{n\pi}(1-\cos(n\pi)) = \frac{2}{n\pi}(1-(-1)^n) = \begin{cases} \frac{4}{n\pi} & \text{for } n \text{ odd} \\ 0 & \text{for } n \text{ even} \end{cases}. \end{aligned}$$

Thus, $f(t) = \frac{4}{\pi}\left(\sin t + \frac{1}{3}\sin 3t + \frac{1}{5}\sin 5t + \cdots\right)$.

21.5.1 Simple formulas for certain angles

Here are several formulas you know and should always use:

1. $\cos(n\pi) = (-1)^n$ *always make this substitution*

2. $\sin(n\pi) = 0$ *always make this substitution*

The values $\sin\left(\frac{n\pi}{2}\right)$ and $\cos\left(\frac{n\pi}{2}\right)$ do not have an easier formula. For example for $n = 0, 1, 2, 3, 4$ we have $\sin\left(\frac{n\pi}{2}\right) = 0,\ 1,\ 0,\ -1,\ 0.$

22 Fourier series introduction: continued

22.1 Goals

1. Be able to compute the Fourier coefficients of even or odd periodic function using simplified formulas.
2. Be able to determine the decay rate of the coefficients of a Fourier series.
3. Be able to predict the decay rate of the Fourier coefficients based on the smoothness of the original function.

22.2 Introduction

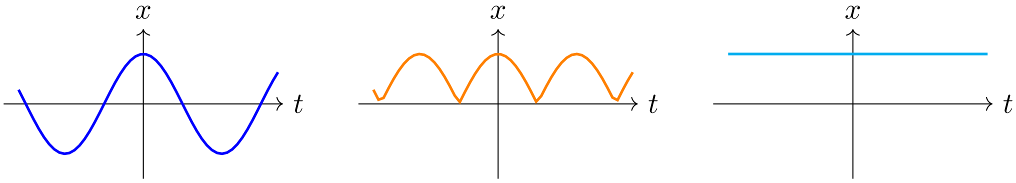
In this topic we continue our introduction to Fourier series. We start by looking at some tricks for computing Fourier coefficients. Then we will talk about more conceptual notions, including the convergence properties of Fourier series and the decay rate of Fourier coefficients. At the end, we will look at the orthogonality relations which explain the formulas for Fourier coefficients.

22.3 Calculation tricks: even and odd functions

22.3.1 Even and odd functions

A function is an even function if $f(-t) = f(t)$ for all t.

- The graph of an even function is symmetric about the y-axis.



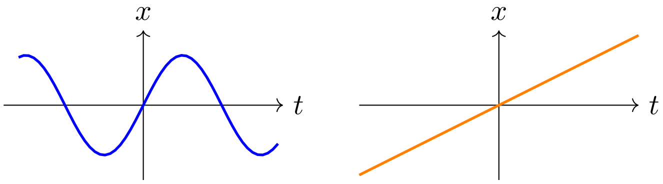
Graphs of some even functions

- Examples of even functions: 1, t^2, t^4, ..., $\cos(\omega t)$. In general, even functions are built out of even powers of t. Note that, the power series for $\cos(\omega t)$ has only even powers.
- By symmetry we have the following key integration fact for even functions:

$$\int_{-L}^{L} f(t)\,dt = 2\int_{0}^{L} f(t)\,dt \quad \text{for any even } f(t).$$

A function is an odd function if $f(-t) = -f(t)$ for all t.

- The graph of an odd function is symmetric about the origin.



Graphs of some odd functions

- Examples of odd functions: t, t^3, t^5, ..., $\sin(\omega t)$. In general, odd functions are built out of odd powers of t. Note that, the power series for $\sin(\omega t)$ has only odd powers.
- By symmetry we have the following key integration fact for odd functions:

$$\int_{-L}^{L} f(t)\,dt = 0 \quad \text{for any odd } f(t).$$

Products of even and odd functions

We give the rules in a kind of short-hand. You can remember these rules by thinking about powers of t, e.g., $t^4 \cdot t^7 = t^{11}$, so even $\cdot$ odd is odd.

- even $\cdot$ even = even, e.g., $t^4 \cdot t^6 = t^{10}$
- odd $\cdot$ odd = even, e.g., $t^3 \cdot t^5 = t^8$
- odd $\cdot$ even = odd, e.g., $t^3 \cdot t^6 = t^9$

22.3.2 Fourier coefficients of even and odd functions

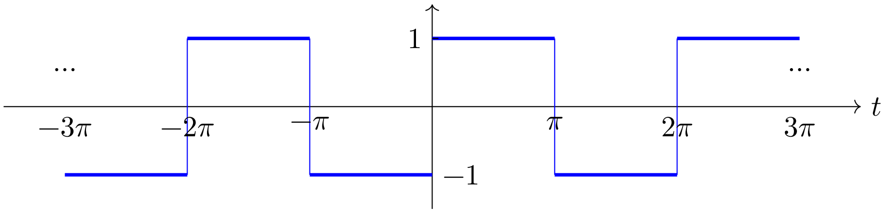
- If $f(t)$ is even, then $b_n = 0$ and $a_n = \dfrac{2}{L}\displaystyle\int_0^L f(t)\cos(\frac{n\pi}{L}t)\,dt$.
- If $f(t)$ is odd, then $a_n = 0$, and $b_n = \dfrac{2}{L}\displaystyle\int_0^L f(t)\sin(\frac{n\pi}{L}t)\,dt$.

Reason. Assume $f(t)$ is even. Then the multiplication rules for even functions imply $f(t)\cos(\omega t)$ is even. So, $a_n = \dfrac{1}{L}\displaystyle\int_{-L}^{L} f(t)\cos\left(\frac{n\pi}{L}t\right)\,dt = \frac{2}{L}\int_0^L f(t)\cos\left(\frac{n\pi}{L}t\right)\,dt$.

Likewise, the rules imply $f(t)\sin(\omega t)$ is odd. So, $b_n = \dfrac{1}{L}\displaystyle\int_{-L}^{L} f(t)\sin\left(\frac{n\pi}{L}t\right)\,dt = 0$.

The argument is similar when $f(t)$ is odd.

Example 22.1. In the previous topic notes we met the period 2π square wave, which over one period has the formula $\text{sq}(t) = \begin{cases} -1 & \text{for } -\pi < t < 0 \\ 1 & \text{for } \ 0 < t < \pi. \end{cases}$



Graph of $\mathrm{sq}(t)$ = square wave

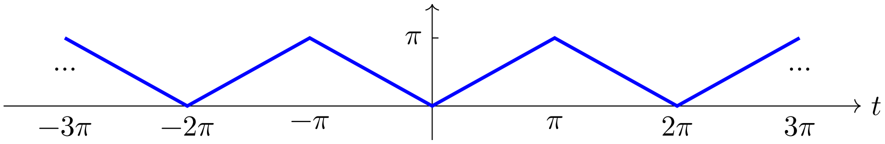
Since the period is 2π, we have $L = \pi$. Since $\mathrm{sq}(t)$ is odd, we know that $a_n = 0$ and

$$b_n = \frac{2}{\pi}\int_0^\pi \mathrm{sq}(t)\sin(nt)\,dt = \frac{2}{\pi}\int_0^\pi \sin(nt)\,dt = -\frac{2}{n\pi}\cos(nt)\Big|_0^\pi = \begin{cases} \frac{4}{n\pi} & \text{for } n \text{ odd} \\ 0 & \text{for } n \text{ even.} \end{cases}$$

We have found the Fourier series for $\mathrm{sq}(t)$:

$$\mathrm{sq}(t) = \sum_{n=1}^{\infty} b_n \sin(nt) = \frac{4}{\pi}\left(\sin(t) + \frac{\sin(3t)}{3} + \frac{\sin(5t)}{5} + \cdots\right) = \frac{4}{\pi}\sum_{n \text{ odd}} \frac{\sin(nt)}{n}.$$

Example 22.2. Triangle wave function (also called the continuous sawtooth function). Let $f(t)$ have period 2π and $f(t) = |t|$ for $-\pi \le t \le \pi$. Compute the Fourier series of $f(t)$.



Graph of $f(t)$ = triangle wave

Since $f(t)$ is an even function, we know that $b_n = 0$ and for $n \ne 0$ we have

$$\begin{aligned} a_n &= \frac{1}{\pi}\int_{-\pi}^{\pi} |t|\cos(nt)\,dt = \frac{2}{\pi}\int_0^\pi t\cos(nt)\,dt \\ &= \frac{2}{\pi}\left[\frac{t\sin(nt)}{n} + \frac{\cos(nt)}{n^2}\right]_0^\pi = \frac{2}{n^2\pi}((-1)^n - 1) = \begin{cases} -\frac{4}{n^2\pi} & \text{for } n \text{ odd} \\ 0 & \text{for } n \text{ even.} \end{cases} \end{aligned}$$

As usual, we compute a_0 separately: $a_0 = \dfrac{1}{\pi}\displaystyle\int_{-\pi}^{\pi} |t|\,dx = \dfrac{2}{\pi}\displaystyle\int_0^\pi t\,dt = \pi.$

Thus we have the Fourier series for $f(t)$:

$$f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n\cos(nt) = \frac{\pi}{2} - \frac{4}{\pi}\left(\cos(t) + \frac{\cos(3t)}{3^2} + \frac{\cos(5t)}{5^2} + \cdots\right) = \frac{\pi}{2} - \frac{4}{\pi}\sum_{n \text{ odd}} \frac{\cos(nt)}{n^2}.$$

22.4 Summing Fourier series

We can use the sum of a finite number of terms from a Fourier series to approximate the original function. The applet https://web.mit.edu/jorloff/www/OCW-ES1803/fourierapproximation.html illustrates this. In the following sections we will bring out the following key points:

- The first few terms of the Fourier series approximate the shape of the function, not necessarily the value of the function at any one point.
- At points of continuity, the Fourier series converges to the original function.
- The smoother the function, the faster the Fourier series converges to the function.
- At jumps in the graph, no matter how many terms you use, the Fourier series always overshoots the graph near that point.

22.5 Convergence of Fourier series

Piecewise smooth: The period $2L$ function $f(t)$ is called piecewise smooth if there are only a finite number of points $0 \leq t_1 < t_2 < ... < t_n \leq 2L$ where $f(t)$ is not differentiable and at each of these points the left and righthand limits

$$f(t_i^+) = \lim_{t \to t_i^+} f(t) \quad \text{and} \quad f(t_i^-) = \lim_{t \to t_i^-} f(t)$$

exist (although they might not be equal).

In short, a function is piecewise smooth if it is smooth except at a discrete set of points where is has jump discontinuities.

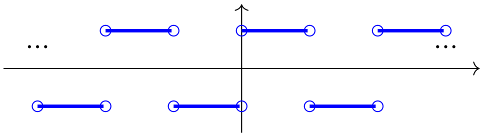
Here is our main theorem about convergence of Fourier series. We will not prove it in ES.1803.

Theorem: If $f(t)$ is piecewise smooth and periodic, then the Fourier series for f:

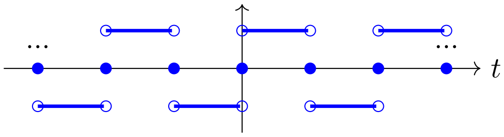
1. Converges to $f(t)$ at values of t where f is continuous.

2. Converges to the average of $f(t^-)$ and $f(t^+)$ at values of t where $f(t)$ has a jump discontinuity.

Example 22.3. Square wave. The square wave in the example above has jump discontinuities. No matter how we specify the endpoint behavior of $\mathrm{sq}(t)$, the Fourier series converge to 0, i.e., the midpoint of the gap, at the discontinuities.



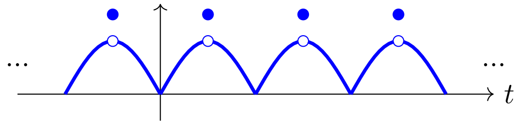
Original $\mathrm{sq}(t)$



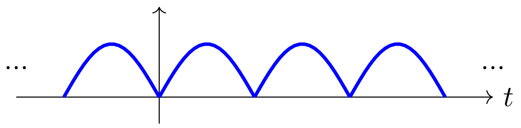
Fourier series

Example 22.4. The triangle wave in the example above is continuous so its Fourier series converges to the original function $f(t)$.

Example 22.5. We give one more graphical example. Here the original function has discontinuities –admittedly somewhat artificial. Since the left and righthand limits are the same at each discontinuity the Fourier series is continuous.



Original $f(t)$



Fourier series

22.5.1 Decay rate of Fourier coefficients

Sequences like $a_n = 1/n$ and $b_n = 1/n^2$ go to 0 as n goes to infinity. We say they decay to 0. Clearly b_n goes to 0 faster than a_n. We will say 'b_n decays like $1/n^2$'.

In general we will ignore constant factors, so, for example, we say $4/(n\pi)$ decays like $1/n$.

Example 22.6. The Fourier coefficients of sq(t) are

$$a_n = 0 \quad \text{and} \quad b_n = \begin{cases} 4/(n\pi) & \text{for } n \text{ odd} \\ 0 & \text{for } n \text{ even.} \end{cases}$$

We say these coefficients decay like $1/n$.

Example 22.7. The triangle wave looked at above has Fourier coefficients

$$b_n = 0 \quad \text{and} \quad a_n = \begin{cases} -4/(n^2\pi) & \text{for } n \text{ odd} \\ 0 & \text{for even } n \neq 0. \end{cases}$$

So these coefficients decay like $1/n^2$.

Example 22.8. The coefficients $a_n = 1/(n + n^2)$ decay like $1/n^2$.

Example 22.9. If a Fourier series has $a_n = 1/n$ and $b_n = 1/n^2$, we say a_n decays like $1/n$ and b_n decays like $1/n^2$. The Fourier coefficients as a whole decay like the slower of the two rates. That is, they decay like $1/n$.

Example 22.10. The function $f(t) = 3\cos(t) + 5\cos(2t)$ is a finite Fourier series. The coefficients are $a_0 = 0$, $a_1 = 3$, $a_2 = 5$, $a_3 = 0$, $a_4 = 0$, ... We say these coefficients decay like 0.

22.5.2 Important heuristics

- If a function has a jump discontinuity, then its Fourier coefficients decay like $\frac{1}{n}$, e.g., the square wave.
- If a function has a corner, then its Fourier coefficients decay like $\frac{1}{n^2}$, e.g., the triangle wave
- A smooth function has Fourier coefficients that decay like $\frac{1}{n^3}$ or faster.
- The smoother the function, the faster the coefficients decay.

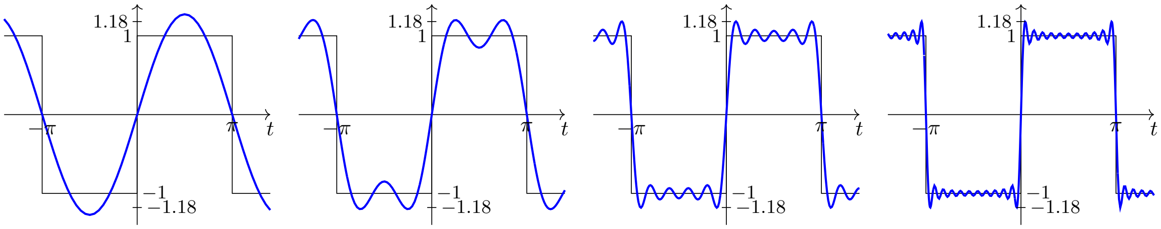
22.6 Gibbs' phenomenon

22.6.1 Non-local nature of Fourier series

Generally speaking, if we sum the first few terms of a Fourier series, it will match the overall shape of the original function. An analogy is the way a squares fit of data points matches the shape of the data without necessarily going through any of the data points.

The figures below show the square wave and its Fourier series summed to some number of terms. The first plot uses just the first term, i.e., $\frac{4}{\pi}\sin(t)$. Notice how it matches the general oscillation of the square wave without matching it well at any particular place.

The second plot uses the terms out to $n = 3$, i.e., $\frac{4}{\pi}\left(\sin(t) + \frac{\sin(3t)}{3}\right)$. This fits the square wave a little better than the first plot. The third plot uses the terms out to $n = 21$. This fits the square even better.



Sum up to $n = 1$ Sum up to $n = 3$ Sum up to $n = 9$ Sum up to $n = 21$

22.6.2 Gibbs' phenomenon

In the figures above, notice that the peak of the reconstructed square wave always overshoot the square 0.18, i.e., it goes up to about 1.18 or down to -1.18. As the number of terms increases, the point where the overshoot occurs moves closer to the point of discontinuity, but never disappears.

This is a general phenomenon, called Gibbs' phenomenon. For any periodic function with a jump discontinuity, summing any number of terms from its Fourier series will *always* overshoot the jump by about 9% of the size of the jump. For example, the square wave has a jump of size 2, so the overshoot is about $2 \cdot 0.09 = 0.18$. Gibbs' phenomenon is extremely important in many applications, e.g., digital filtering of signals.

We won't prove Gibbs' phenomenon in ES.1803. For those who are interested, we've posted an enrichment note with the proof. It should accessible to anyone who knows calculus.

The applet
`https://web.mit.edu/jorloff/www/OCW-ES1803/fourierapproximation.html`
shows this overshoot in several cases.

22.7 Orthogonality relations

22.7.1 Orthognality relation integrals

The key to the integral formulas for Fourier coefficients are the orthogonality relations. These are the following integral formulas that say certain trigonometric integrals are either 0 or 1.

$$\frac{1}{L}\int_{-L}^{L} \cos\left(\frac{n\pi}{L}t\right)\cos\left(\frac{m\pi}{L}t\right)\,dt = \begin{cases} 1 & n = m \neq 0 \\ 0 & n \neq m \\ 2 & n = m = 0 \end{cases}$$

$$\frac{1}{L}\int_{-L}^{L} \sin\left(\frac{n\pi}{L}t\right)\cos\left(\frac{m\pi}{L}t\right)\,dt = 0$$

$$\frac{1}{L}\int_{-L}^{L} \sin\left(\frac{n\pi}{L}t\right)\sin\left(\frac{m\pi}{L}t\right)\,dt = \begin{cases} 1 & n = m \neq 0 \\ 0 & n \neq m \end{cases}$$

Proof. We have two methods to do this. We will carry out the first, but only mention the second.

Method 1: Use the following trigonometric identities

$$\cos(\alpha)\cos(\beta) = \frac{\cos(\alpha+\beta)+\cos(\alpha-\beta)}{2}$$

$$\sin(\alpha)\cos(\beta) = \frac{\sin(\alpha+\beta)+\sin(\alpha-\beta)}{2}$$

$$\sin(\alpha)\sin(\beta) = \frac{\cos(\alpha-\beta)-\cos(\alpha+\beta)}{2}$$

Method 2: Use $\cos(at) = \frac{e^{iat}+e^{-iat}}{2}$ etc.

Using method 1 we get the following if $n \neq m$:

$$\begin{aligned}\frac{1}{L}\int_{-L}^{L} \cos\left(\frac{n\pi}{L}t\right)\cos\left(\frac{m\pi}{L}t\right)\,dt &= \frac{1}{L}\int_{-L}^{L} \frac{\cos\left(\frac{(n+m)\pi}{L}t\right)+\cos\left(\frac{(n-m)\pi}{L}t\right)}{2}\,dt \\ &= \frac{1}{2L}\left[\frac{\sin\left(\frac{(n+m)\pi}{L}t\right)}{(n+m)\pi/L} + \frac{\sin\left(\frac{(n-m)\pi}{L}t\right)}{(n-m)\pi/L}\right|_{-L}^{L} \\ &= 0.\end{aligned}$$

The last equality is easy to see since every term is 0 when $t = \pm L$.

The case $n = m$ is special because then $n - m = 0$. It is easy to use the first trig identity above with $\alpha = \beta$, i.e., $\cos(\alpha)\cos(\alpha) = (\cos(2\alpha)+1)/2$, to see that the integral in this case is 1. All the other orthogonality relations are proved in a similar fashion.

The term orthogonality comes from linear algebra, where we say two vectors are orthogonal if there dot product is 0. It turns out that we can think of $\int_{-L}^{L} f(t)g(t)\,dt$ as a dot product (usually called inner product) between f and g. So the orthogonality relations say that, for $n \neq m$, the functions $\cos(n\pi t/L)$ and $\cos(m\pi t/L)$ are orthogonal.

22.7.2 Using orthogonality relations to show the formula for Fourier coefficients

The orthogonality relations allow us to see that if $f(t)$ is written as a Fourier series, then the coefficients must be given by the integral formulas we've been using.

So suppose $f(t)$ has Fourier series':

$$f(t) = \frac{a_0}{2} + a_1 \cos\left(\frac{\pi}{L}t\right) + a_2 \cos\left(\frac{2\pi}{L}t\right) + \cdots + b_1 \sin\left(\frac{\pi}{L}t\right) + b_2 \sin\left(\frac{2\pi}{L}t\right) + \cdots$$

Then for $n > 0$

$$\begin{aligned}\frac{1}{L}\int_{-L}^{L} f(t)\cos\left(\frac{n\pi}{L}t\right)\,dt = \frac{1}{L}\int_{-L}^{L} &\left[\frac{a_0}{2}\cos\left(\frac{n\pi}{L}t\right) + a_1\cos\left(\frac{\pi}{L}t\right)\cos\left(\frac{n\pi}{L}t\right)\right.\\ &+ a_2\cos\left(\frac{2\pi}{L}t\right)\cos\left(\frac{n\pi}{L}t\right) + \cdots\\ &\left.+b_1\sin\left(\frac{\pi}{L}t\right)\cos\left(\frac{n\pi}{L}t\right) + b_2\sin\left(\frac{2\pi}{L}t\right)\cos\left(\frac{n\pi}{L}t\right) + \cdots\right]\,dt\end{aligned}$$

Now we can apply the orthogonality relations to each term. All of them are 0, except the term with $a_n \cos\left(\frac{n\pi}{L}t\right)\cos\left(\frac{n\pi}{L}t\right)$ which, again by the orthogonality relations, integrates to a_n. Thus, $\displaystyle\frac{1}{L}\int_{-L}^{L} f(t)\cos(\frac{n\pi}{L}t)\,dt = a_n$. Which is exactly the formula for the Fourier coefficient. The formulas for a_0 and b_n are found in the same way.

22.8 Hearing a musical triad: C-E-G

Here is a simplified Fourier-centric view of how humans hear sound.

Sound reaches your ear as a pressure wave. For example

$$f(t) = a_1\cos(\omega_1 t) + a_2\cos(w_2 t) + \cdots$$

Do the ears do Fourier analysis?

Answer: Yes! The ear contains hair-like structures called stereocilia. These are different sizes and, so, resonate at different frequencies. As they vibrate they stimulate nerves, which then send signals to the brain. Thus, for each frequency in the pressure wave, the brain is getting a signal from the nerves attached to the stereocilia which vibrate at that frequency. The greater the amplitude in the input wave the greater the amplitude of the signal sent to the brain.

Does the brain do Fourier synthesis?

Answer: Yes! It is up to the brain to combine all the nerve signals at different frequencies into a single signal which it then interprets.

23 Fourier sine and cosine series; calculation tricks

23.1 Goals

1. Be able to use various calculation shortcuts for computing Fourier series:
   shifting and scaling $f(t)$
   shifting and scaling t
   differentiating and integrating known series.

2. Be able to find the sine and cosine series for a function defined on the interval $[0, L]$

3. Understand the distinction between $f(x)$ defined on $[0, L]$ and it's even and odd periodic extensions.

23.2 Introduction

This topic is split into two subtopics. First, we look at a few more calculation tricks. The common idea in these tricks is to use the Fourier series of one function to find the Fourier series of another. A simple example is if we scale a function, say $g(t) = 5f(t)$, the Fourier series for $g(t)$ is 5 times the Fourier series of $f(t)$.

Next, we'll look at functions $f(x)$ that are only defined on the interval $[0, L]$. This is in preparation for our later study of the heat and wave equations. This function is not periodic –it's not even defined for all x. By extending $f(x)$ to an even or odd periodic function we can write the original function $f(x)$ as a sum of sines (sine series) or a sum of cosines (cosine series).

23.3 Calculation shortcuts

One of our goals is to avoid computing integrals when finding the Fourier coefficients of a periodic function. In this section we'll consider the following calculation shortcuts for computing Fourier series:

1. Simplify computations for even or odd periodic functions. (Already covered in the previous topic.)

2. Use known Fourier series to compute the Fourier series for scaled and shifted functions.

3. Use known Fourier series to compute the Fourier series for the derivative or integrals of functions.

Even and odd functions were covered in the previous topic, so we won't go over them again here.

23.3.1 New series from old ones: shifting and scaling

First, if you scale and shift $f(t)$, then you scale and shift its Fourier series. To avoid burdening the statement with too much notation we state it for period 2π functions. You can extend this easily to any period.

Suppose $f(t)$ has Fourier series $f(t) = \dfrac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos(nt) + \sum_{n=1}^{\infty} b_n \sin(nt)$.

Theorem. (Scaling and shifting $f(t)$.) The scaled and shifted function $g(t) = cf(t) + d$ has Fourier series

$$g(t) \;=\; cf(t) + d \;=\; \frac{ca_0}{2} + d + \sum_{n=1}^{\infty} ca_n \cos(nt) + \sum_{n=1}^{\infty} cb_n \sin(nt).$$

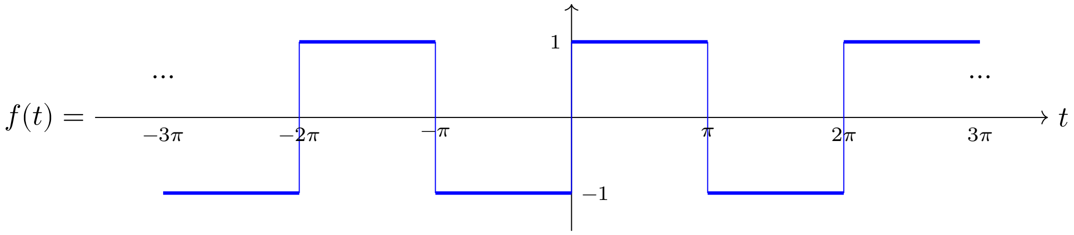
Theorem. (Scaling and shifting in time.) The function $g(t) = f(ct + d)$ has

$$g(t) \;=\; f(ct+d) \;=\; \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos(n(ct+d)) + \sum_{n=1}^{\infty} b_n \sin(n(ct+d)).$$

This is not quite in standard Fourier series form, but it is in a useable form. Also, if we really want a standard Fourier series, it is easy to expand out the trig functions to put it in standard form.

The rest of this subsection will be devoted to an extended example, illustrating these techniques using our standard period 2π square wave whose graph is shown just below.

Example 23.1. (Extended example.) The graph of $f(t)$ looks like this:



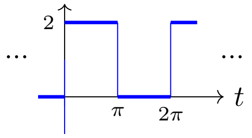
Graph of $f(t)$ = square wave

We know that the Fourier series for $f(t)$ is

$$f(t) = \frac{4}{\pi} \sum_{n \text{ odd}} \frac{\sin(nt)}{n}. \tag{34}$$

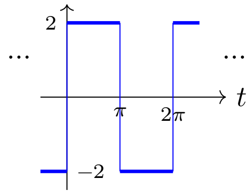
Now we will use this to find the Fourier series for scaled and shifted versions of $f(t)$. We'll define these new functions graphically, we could also write down formulas if we wanted.

(a) $f_1(t) =$



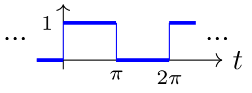
$\Rightarrow f_1(t) = 1 + f(t) = 1 + \dfrac{4}{\pi} \sum_{n \text{ odd}} \dfrac{\sin(nt)}{n}$.

(b) $f_2(t) =$



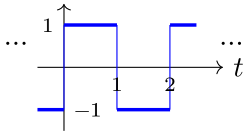
$\Rightarrow f_2(t) = 2f(t) = \dfrac{8}{\pi} \sum_{n \text{ odd}} \dfrac{\sin(nt)}{n}$.

(c) $f_3(t) =$



$\Rightarrow f_3(t) = \dfrac{1}{2}(1 + f(t)) = \dfrac{1}{2} + \dfrac{2}{\pi} \sum_{n \text{ odd}} \dfrac{\sin(nt)}{n}$.

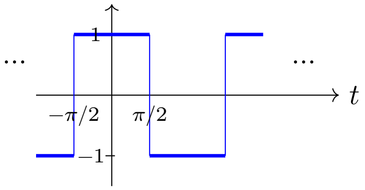
Next will look at what happens if we scale the time t.

(d) $f_4(t) =$ 

$\Rightarrow f_4(t) = f(\pi t) = \frac{4}{\pi}\sum_{n \text{ odd}} \frac{\sin(n\pi t)}{n}.$

It's a little tricky to see that $f_4(t) = f(\pi t)$. I think about it two ways. First, the picture shows that we want $f_4(1) = f(\pi)$, which is given by $f_4(t) = f(\pi t)$. Second, $f_4(t)$ has period 2 so its Fourier series should have terms with frequencies $n\pi$.

Our last example involves shifting the time.

(e) $f_5(t) =$ 

$\Rightarrow f_5(t) = f(t + \pi/2) = \frac{4}{\pi}\sum_{n \text{ odd}} \frac{\sin(n(t + \pi/2))}{n}.$

That is,

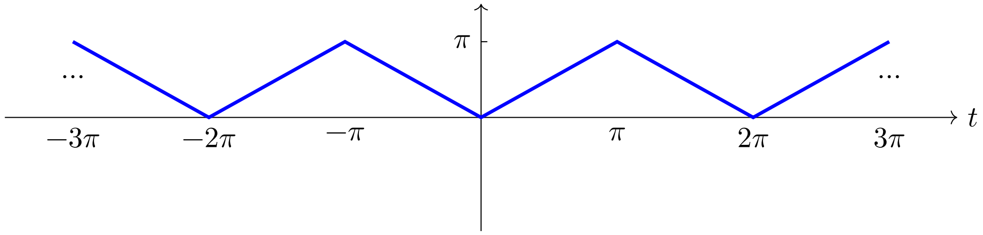
$$f_5(t) = \frac{4}{\pi}\left(\sin(t + \pi/2) + \frac{\sin(3t + 3\pi/2)}{3} + ...\right) = \frac{4}{\pi}\left(\cos t - \frac{\cos 3t}{3} + ...\right).$$

The last expression is in the form we defined for Fourier series. For most applications, the middle expression is perfectly useable and sometimes even preferable.

23.3.2 Differentiation and integration

If $f(t)$ is periodic, then the Fourier series for $f'(t)$ is just the term-by-term derivative of the Fourier series for $f(t)$. An example should make this clear.

Example 23.2. Let $f(t)$ be the period 2π triangle wave with $f(t) = |t|$ on $[-\pi.\pi]$. It's clear that $f'(t)$ is the square wave. Check that the derivative of the Fourier series of $f(t)$ is the Fourier series of $f'(t)$.

$f(t) =$ 

Graph of $f(t)$ = triangle wave

Solution: From the previous topic notes, we know the Fourier series for $f(t)$ is

$$f(t) = \frac{\pi}{2} - \frac{4}{\pi}\left(\cos t + \frac{\cos 3t}{3^2} + \frac{\cos 5t}{5^2} + ...\right)$$

Thus, $f'(t) = \frac{4}{\pi}\left(\sin t + \frac{\sin 3t}{3} + \frac{\sin 5t}{5} + ...\right)$. We know this is the Fourier series of our standard square wave as claimed.

Decay rate of Fourier series. Note that $f(t)$ has a corner and its coefficients decay like $1/n^2$, while $f'(t)$ has a jump and its coefficients decay like $1/n$. Note also, how differentiation changed the power of n in the decay rate.

Differentiation of discontinuous functions. Term-by-term differentiation of Fourier series works for discontinuous functions as long as we use the generalized derivative.

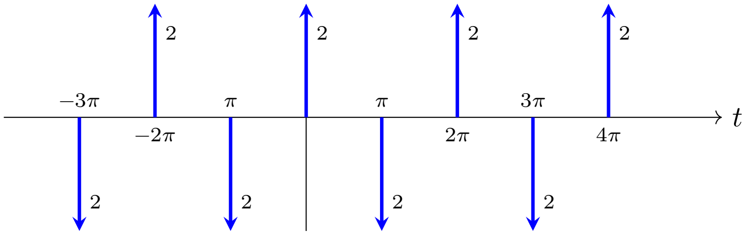
Example 23.3. Let $f(t)$ be our standard period 2π square wave. Find $f'(t)$ and the Fourier series of $f'(t)$. Graph $f'(t)$.

Solution: Because $f(t)$ has jumps (alternating between 2 and -2) we must take the generalized derivative:

$$f'(t) = \ldots - 2\delta(t+\pi) + 2\delta(t) - 2\delta(t-\pi) + 2\delta(t-2\pi) - \ldots$$

We know $f(t) = \dfrac{4}{\pi}\sum\limits_{n \text{ odd}} \dfrac{\sin(nt)}{n}$. So, taking the term-by-term derivative, $f'(t) = \dfrac{4}{\pi}\sum\limits_{n \text{ odd}} \cos(nt)$.

You can check this by computing the Fourier coefficients of $f'(t)$ directly using the integral formulas.



Graph of $f'(t)$ = impulse train

Example 23.4. Term-by-term integration. Suppose that

$$f(t) = 1 + \cos(t) + \frac{\cos(2t)}{2} + \frac{\cos(3t)}{3} + \frac{\cos(4t)}{4} + \ldots$$

What is $h(t) = \displaystyle\int_0^t f(u)\,du$?

Solution: We integrate the Fourier series term-by-term to get

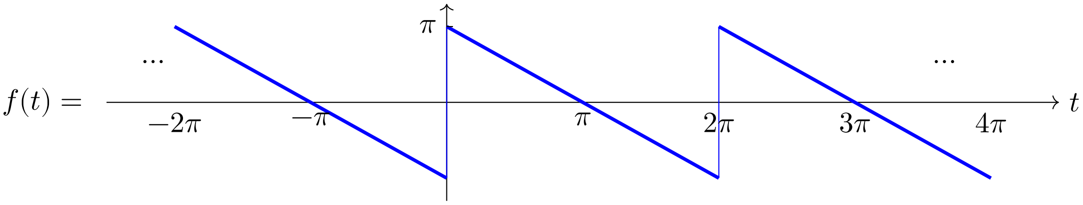
$$h(t) = \int_0^t f(u)\,du = C + t + \sin(t) + \frac{\sin(2t)}{2^2} + \frac{\sin(3t)}{3^2} + \ldots$$

Note: Just because $f(t)$ is periodic doesn't mean the integral of $f(t)$ will be periodic. In this case, the "t-term" shows that $h(t)$ is not periodic. So we can't officially say we have a Fourier series for $h(t)$. Nonetheless we have a nice series for $h(t)$ that can be used in many applications.

Here's one more example of integration. It's very cool, but we probably won't get to it in class.

Example 23.5. For your amusement. Consider the period 2π discontinuous sawtooth function

$$f(t) = \frac{\pi}{2} - \frac{t}{2} \quad \text{for } 0 < t < 2\pi.$$



Graph of $f(t) =$ discontinuous sawtooth

Since $f(t)$ is odd with period 2π, we know that the cosine coefficients $a_n = 0$. For the sine coefficients it is slightly easier to do the integral over a full period rather than double the integral over a half period:

$$b_n = \frac{1}{\pi}\int_0^{2\pi} \frac{\pi - t}{2}\sin(nt)\,dt = \frac{1}{n}.$$

Thus, $f(t) = \sin(t) + \dfrac{\sin(2t)}{2} + \dfrac{\sin(3t)}{3} + \ldots.$

Now, let $h(t)$ be the integral of $f(t)$, specifically

$$\begin{aligned}\text{Let } h(t) &= \int_0^t f(u)\,du = \int_0^t \sin u + \frac{\sin 2u}{2} + \frac{\sin 3u}{3} + \ldots\, du \\ &= (1-\cos(t)) + \frac{1-\cos(2t)}{2^2} + \frac{1-\cos(3t)}{3^2} + \ldots \\ &= \sum_1^\infty \frac{1}{n^2} - \sum_1^\infty \frac{\cos(nt)}{n^2}.\end{aligned}$$

The DC term is $\frac{a_0}{2} = \sum \frac{1}{n^2}$. This is an infinite sum, but we can compute its value directly using the integral formula for Fourier coefficients. On $[0, 2\pi]$, $h(t) = \displaystyle\int_0^t \frac{\pi}{2} - \frac{u}{2}\,du = \frac{\pi t}{2} - \frac{t^2}{4}$. Thus,

$$a_0 = \frac{1}{\pi}\int_0^{2\pi} \frac{\pi t}{2} - \frac{t^2}{4}\,dt = \frac{\pi^2}{3}.$$

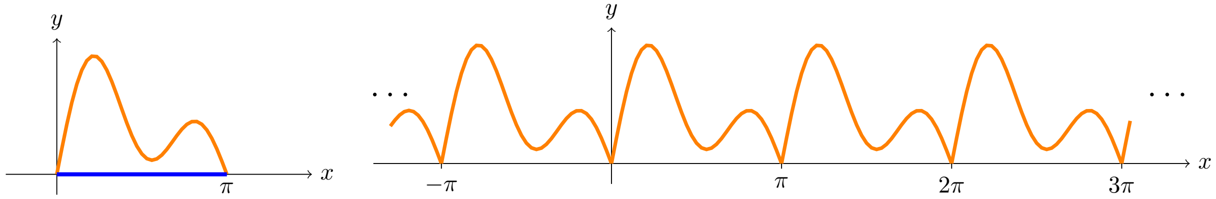
So, $\dfrac{a_0}{2} = \dfrac{\pi^2}{6} = \sum \dfrac{1}{n^2}$. We've summed an infinite series!

23.4 Sine and cosine series; even and odd extensions

23.4.1 Definition of sine and cosine series

In this section we will be concerned with functions $f(x)$ defined on an interval $[0, L]$. We start by stating the theorem on how to write functions as sine and cosine series. After that, we will use what we know about Fourier series to justify the theorem. We will need sine and cosine series when we study the heat and wave equations.

But first, an important semantic distinction: Fourier series are defined for periodic functions. A function defined only on an interval $[0, L]$ cannot be periodic, so it doesn't have a Fourier series. The figures below show a function defined on the interval $[0, \pi]$ and a period π function defined over the entire real line.



Left: function defined $[0, \pi]$, can't be periodic. Right: periodic function

Sine and cosine series. Without further ado, we state how to write a function as a cosine or sine series and how to compute the coefficients for the series. Note, the statements look very much like the ones for Fourier series.

Consider a function $f(x)$ defined on the interval $[0, L]$. $f(x)$ can be written as a cosine series:

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi x}{L}\right), \qquad \text{where } a_n = \frac{2}{L}\int_0^L f(x)\cos\left(\frac{n\pi x}{L}\right)\,dx.$$

$f(x)$ also has a sine series:

$$f(x) = \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{L}\right), \qquad \text{where } b_n = \frac{2}{L}\int_0^L f(x)\sin\left(\frac{n\pi x}{L}\right)\,dx.$$

Important.

1. Sine and cosine series are about functions defined on an interval.
2. The sine and cosine series have values for all x. At points in $(0, L)$ where $f(x)$ is continuous, the sine and cosine series equal $f(x)$. Since $f(x)$ is only defined on $[0, L]$, this is usually what we want.
3. Computing a_n and b_n only depends on the values of $f(x)$ in the interval $[0, L]$.
4. We will make use of sine and cosine series when we study the heat and wave equations.

23.4.2 Examples of sine and cosine series

Now, we'll give some example computations. We can do this by mechanically applying the formulas. We'll gain more insight into these series after we have seen the proof justifying the formulas for the coefficients.

Example 23.6. Find the Fourier cosine and sine series for the function $f(x) = \sin(x)$ defined on $[0, \pi]$.

Solution: Cosine series. $L = \pi$, Using the formula for a_n:

$$a_0 = \frac{2}{\pi}\int_0^{\pi} \sin(x)\,dx = \left[-\frac{2}{\pi}\cos(x)\right|_0^{\pi} = \frac{4}{\pi}.$$

By applying the formula $\sin(a)\cos(b) = \dfrac{\sin(a+b)+\sin(a-b)}{2}$ we get:

$$a_n = \frac{2}{\pi}\int_0^{\pi} \sin(x)\,\cos(nx)\,dx = \frac{1}{\pi}\left[-\frac{\cos((1+n)x)}{1+n} - \frac{\cos((1-n)x)}{1-n}\right|_0^{\pi} = \begin{cases} 0 & \text{for odd } n > 0 \\ \frac{-4}{\pi(n^2-1)} & \text{for even } n > 0. \end{cases}$$

(You have to be careful with $n = 1$, but the formula is correct.)

Thus,

$$f(x) = \frac{2}{\pi} - \frac{4}{\pi}\left(\frac{\cos(2x)}{3} + \frac{\cos(4x)}{15} + \frac{\cos(6x)}{35} + ...\right) = \frac{2}{\pi} - \frac{4}{\pi}\sum_{n>0,\text{ even}} \frac{\cos(nx)}{n^2-1}.$$

Important. This is only valid where $f(x)$ is defined, i.e., on $[0, \pi]$.

Sine series. $f(x) = \sin(x)$ on $[0, \pi]$. This can be seen by comparing the abstract sine series $\sum_{n=1}^{\infty} b_n \sin(nx)$ with the given function $f(x) = \sin(x)$. Or we could compute the integrals for b_n similar to the way we computed a_n above.

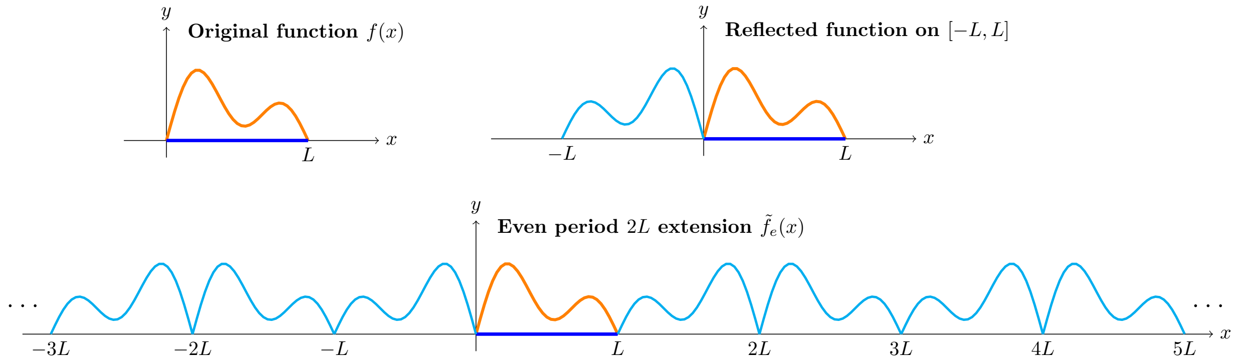
23.4.3 Even and odd periodic extensions

The proof of the formulas for the sine and cosine series coefficients turns out to be a straightforward application of Fourier series for periodic functions. The trick is to view the fact that $f(x)$ is only defined on $[0, L]$ as an opportunity instead of a limitation. To do this we need to define even and odd periodic extensions of $f(x)$.

Definition. If $f(x)$ is a function defined on the interval $[0, L]$ then the even period $2L$ extension of $f(x)$ is the period $2L$ function

$$\tilde{f}_e(x) = \begin{cases} f(-x) & \text{for } -L < x < 0 \\ f(x) & \text{for } 0 < x < L \end{cases}$$

To visualize this, we first reflect $f(x)$ in the y-axis to get a function defined over one period $[-L, L]$. We then extend this to be periodic over the entire real line.

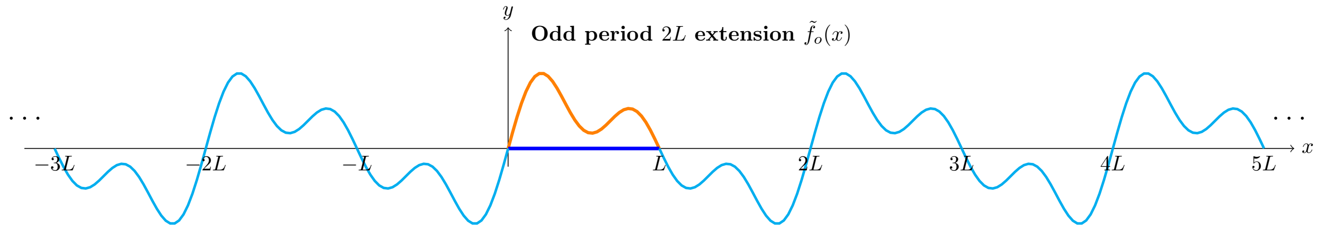


Making an even period $2L$ extension.

The odd period $2L$ extension of $f(x)$ is defined similarly, with

$$\tilde{f}_o(x) = \begin{cases} -f(-x) & \text{for } -L < x < 0 \\ f(x) & \text{for } 0 < x < L \end{cases}$$

To visualize this, we first reflect $f(x)$ through the origin to get a function defined over one period $[-L, L]$. We then extend this to be periodic over the entire real line.



The odd period $2L$ extension.

23.4.4 Proof of the formulas for the sine and cosine series

As we said, using the even and odd period $2L$ extensions this is a straightforward application of Fourier series for periodic functions. We will give the argument for the cosine series. The sine series is similar.

We have $f(x)$ defined on $[0, L]$ and the even period $2L$ extension $\tilde{f}_e(x)$. Since $\tilde{f}_e(x)$ is periodic, it has a Fourier series and since it is even this series has only cosine terms. That is,

$$\tilde{f}_e(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi x}{L}\right).$$

Using the symmetry of even functions we know $a_n = \dfrac{2}{L}\displaystyle\int_0^L \tilde{f}_e(x)\cos\left(\frac{n\pi x}{L}\right)\,dx$. But, on the interval of integration, we know $\tilde{f}_e(x) = f(x)$. Therefore,

$$a_n = \frac{2}{L}\int_0^L f(x)\cos\left(\frac{n\pi x}{L}\right)\,dx.$$

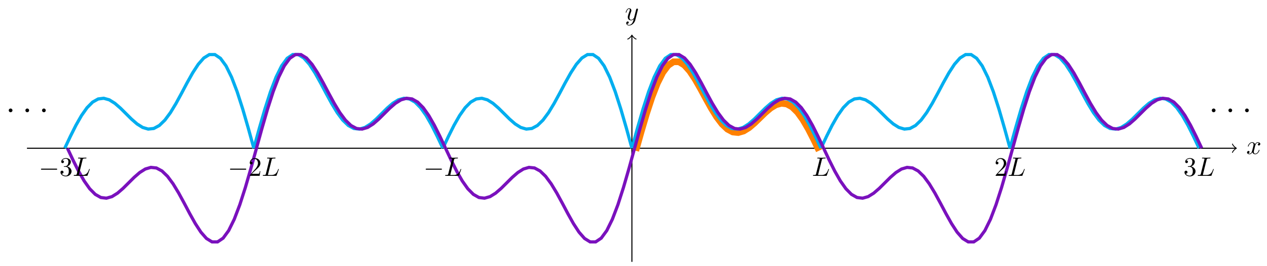
This is the formula we wanted to prove.

Sine series. You should try proving the formula for the sine series coefficients.

Once more to emphasize the grammar:
$f(x)$ is defined for x in $[0, L]$, while $\tilde{f}_e(x)$ and $\tilde{f}_o(x)$ are defined for all x.

The three functions agree on $[0, L]$, i.e., $f(x) = \tilde{f}_e(x) = \tilde{f}_o(x)$ for x in $[0, L]$. The cosine series for $f(x)$ is just the Fourier series for $\tilde{f}_e(x)$. The sine series for $f(x)$ is just the Fourier series for $\tilde{f}_o(x)$.

This is illustrated in the following figure:



$f(x)$ in orange, $\tilde{f}_e(x)$ in cyan, $\tilde{f}_o(x)$ in purple. All three are the same for $0 < x < L$.

We finish with an example that shows how to use known Fourier series to avoid computing integrals for sine and cosine series.

Example 23.7. Find the sine and cosine series for the function $f(x) = 1$ defined on the interval $[0, \pi]$.

Solution: Since the odd period 2π extension of $f(x)$ is our standard square wave, we have the sine series is the Fourier series of $\text{sq}(x)$:

$$f(x) = \frac{4}{\pi} \sum_{n \text{ odd}} \frac{\sin(nx)}{n}.$$

Since the even period 2π extension is the constant function $f(x) = 1$, we have the cosine series:

$$f(x) = 1.$$

24 Linear ODEs with periodic input

24.1 Goals

1. Be able to solve a linear constant coefficient differential equation with periodic input by writing the input as a Fourier series.
   - Know to index the phase lags as $\phi(n)$ or ϕ_n in the superposition for the solution.
   - Be able to identify the term in the input that causes the biggest response.
   - Be able to recognize when one term in the Fourier series for the input produces a pure resonant term in the output.

24.2 Introduction

In this topic we combine Fourier series with the superposition principle to solve linear differential equations. This is really a small extension of what we did way back in the first unit where we had a finite number of terms being superpositioned. Now, with Fourier series, we have an infinite number of terms. Superposition works exactly the same way as before, but we'll have to work out how to present the solution in a nice form.

We'll do this by presenting a series of examples. You should pay attention to the format of the solutions.

24.3 Examples of constant coefficient DEs with periodic input

Example 24.1. Let $f(t)$ be the odd period 2π square wave with height 1. Find the periodic solution to the DE $\ddot{x} + 8x = f(t)$.

Solution: Using the (known) Fourier series for $f(t)$ the equation becomes

$$\ddot{x} + 8x = \frac{4}{\pi}\left(\sin(t) + \frac{\sin(3t)}{3} + \frac{\sin(5t)}{5} + ...\right) = \frac{4}{\pi}\sum_{n \text{ odd}} \frac{\sin(nt)}{n}$$

In preparation for using the superposition principle we solve the DE separately for each term in the input, i.e., solve

$$\ddot{x}_n + 8x_n = \frac{\sin(nt)}{n}$$

The characteristic polynomial for this equation is $P(r) = r^2 + 8$. So,

$$P(in) = 8 - n^2; \quad |P(in)| = |8 - n^2|; \quad \text{Arg}(P(in)) = \boxed{\phi(n) = \begin{cases} 0 & \text{if } n \le 2 \\ \pi & \text{if } n \ge 3 \end{cases}}$$

The sinusoidal response formula gives us

$$x_{n,p}(t) = \frac{\sin(nt - \phi(n))}{n|P(in)|} = \frac{\sin(nt - \phi(n))}{n|8 - n^2|}.$$

Putting it together using superposition

$$x_p(t) = \frac{4}{\pi}\left(\frac{\sin(t)}{|8-1|} + \frac{\sin(3t-\pi)}{3|8-9|} + \frac{\sin(5t-\pi)}{5|8-25|} + ...\right) = \frac{4}{\pi}\sum_{n \text{ odd}} \frac{\sin(nt - \phi(n))}{n|8-n^2|}$$

We will often call this the steady periodic solution.

Important feature. Note that we were careful to label the phase lags as $\phi(n)$. This is because ϕ is be different for different terms. Instead of $\phi(n)$, we will sometimes use the notation ϕ_n.

Note: The solution given just above is correct, but we can make it a bit nicer looking by noting that $\sin(nt - \pi) = -\sin(nt)$. This gives us

$$x_p(t) = \frac{4}{\pi}\cdot\frac{\sin(t)}{7} - \frac{4}{\pi}\sum_{n \text{ odd},\, n\ge 3} \frac{\sin(nt)}{n|8-n^2|}.$$

Example 24.2. (Near resonance.) In the previous example, which term in the solution has the biggest amplitude?

Solution: The $n = 3$ term has the biggest amplitude (1/3). Note that the resonant frequency of the system is $\sqrt{8}$ and that 3 is the frequency in the Fourier series closest to this resonant frequency

Example 24.3. (Pure resonance.) Let $f(t)$ be the same square wave as in the previous examples. Find a particular solution to $\ddot{x} + 9x = f(t)$.

Solution: We solve the DE separately for each term in the input:

$$\ddot{x}_n + 9x_n = \frac{\sin(nt)}{n}$$

The characteristic polynomial is $P(r) = r^2+9$. The difference between this example and the previous one is that $P(3i) = 0$, so we will need the extended sinusoidal response formula.

$$P(in) = 9 - n^2; \quad |P(in)| = |9 - n^2|; \quad \text{Arg}(P(in)) = \boxed{\phi_n = \begin{cases} 0 & \text{if } n \le 2 \\ \pi & \text{if } n > 3 \\ \text{undefined} & \text{if } n = 3 \end{cases}}$$

For $n \neq 3$ the sinusoidal response formula gives us

$$x_{n,p}(t) = \frac{\sin(nt - \phi_n)}{n|P(in)|} = \frac{\sin(nt - \phi_n)}{n|9 - n^2|}.$$

When $n = 3$, we have $P(3i) = 0$, so we'll need to use the extended SRF:

$$P'(r) = 2r \Rightarrow P'(3i) = 6i = 6e^{i\pi/2}.$$

So,

$$x_{3,p}(t) = \frac{t\sin(3t - \pi/2)}{3 \cdot 6}$$

Putting it together, using superposition (and that $\sin(nt - \pi) = -\sin(nt)$), our solution is:

$$x_p(t) = \frac{4}{\pi}\left(\frac{\sin(t)}{8} - \frac{t\cos(3t)}{18} + \frac{\sin(5t - \pi)}{80} + ...\right) = \frac{4\sin(t)}{8\pi} - \frac{4t\cos(3t)}{18\pi} - \frac{4}{\pi}\sum_{n>3,\ n \text{ odd}} \frac{\sin(nt)}{n|9 - n^2|}$$

Note: The input has angular frequency 1, but its Fourier series contains a frequency 3 component which causes pure resonance.

Example 24.4. Solve $\ddot{x} + 2\dot{x} + 9x = f$, where $f(t)$ is the triangle wave: $f(t) = |t|$ for $-\pi < t < \pi$.

Solution: We know from the Topic 22 notes that the Fourier series for $f(t)$ is

$$f(t) = \frac{\pi}{2} - \frac{4}{\pi}\left(\cos(t) + \frac{\cos(3t)}{3^2} + \frac{\cos(5t)}{5^2} + ...\right)$$

We'll use a slightly different pattern here and ignore the scale factors while we solve the DE for each term in the input. We'll bring the scale factors back when we use superposition. The DE for each piece is

$$\ddot{x}_n + 2\dot{x}_n + 9x_n = \cos(nt)$$

The characteristic polynomial is $P(r) = r^2 + 2r + 9$. So,

$$\begin{aligned} P(in) &= 9 - n^2 + 2ni; \\ |P(in)| &= \sqrt{(9 - n^2)^2 + 4n^2}; \\ \text{Arg}(P(in)) &= \boxed{\phi(n) = \tan^{-1}\left(\frac{2n}{9 - n^2}\right) \text{ in Q1 or Q2.}} \end{aligned}$$

Thus,

$$x_{n,p}(t) = \frac{\cos(nt - \phi(n))}{|P(in)|} = \frac{\cos(nt - \phi(n))}{\sqrt{(9 - n^2)^2 + 4n^2}}.$$

There is also a constant term in the input, $\ddot{x}_0 + 2\dot{x}_0 + 9x_0 = \pi/2$. This is easy to solve:

$$x_{0,p}(t) = \frac{\pi}{18}.$$

Putting it together using superposition (and restoring the scale factors) our solution is:

$$\begin{aligned} x_p &= x_{0,p} - \frac{4}{\pi}\left(x_{1,p} + \frac{x_{3,p}}{3^2} + \frac{x_{5,p}}{5^2} + ...\right) \\ &= \frac{\pi}{18} - \frac{4}{\pi}\left(\frac{\cos(t-\phi(1))}{\sqrt{68}} + \frac{\cos(3t-\phi(3))}{\sqrt{36}} + \frac{\cos(5t-\phi(5))}{\sqrt{356}} + ...\right) \\ &= \frac{\pi}{18} - \frac{4}{\pi}\sum_{n \text{ odd}} \frac{\cos(nt-\phi(n))}{n^2\sqrt{(9-n^2)^2+4n^2}} \end{aligned}$$

Note. The damping complicates the expressions for $P(in)$, but it also means that we don't need to worry about pure resonance.

We should do a first-order equation:

Example 24.5. Find the general solution to $\dot{x} + kx = f(t)$, where

$$f(t) = 1 + \frac{\cos(t)}{1} + \frac{\cos(2t)}{2} + \frac{\cos(3t)}{3} + ... = 1 + \sum_{n=1}^{\infty} \frac{\cos(nt)}{n}.$$

Solution: The problem asks for the general solution, so we start by giving the homogeneous solution: $x_h(t) = Ce^{-kt}$.

Finding x_p is similar to the examples above.

Characteristic polynomial: $P(r) = r + k$. So,

$$P(in) = k + in; \quad |P(in)| = \sqrt{k^2+n^2}; \quad \text{Arg}(P(in)) = \boxed{\phi(n) = \tan^{-1}(n/k) \ \text{ in Q1}}.$$

Individual pieces: $\dot{x}_n + kx_n = \cos(nt)/n$. Using the SRF

$$x_{n,p}(t) = \frac{\cos(nt-\phi(n))}{n\sqrt{k^2+n^2}}$$

Constant term: $\dot{x}_0 + kx_0 = 1 \Rightarrow x_{0,p} = 1/k$.

Now using superposition we find

$$\begin{aligned} x_p(t) &= x_{0,p} + x_{1,p} + x_{2,p} + x_{3,p} + ... \\ &= \frac{1}{k} + \frac{\cos(t-\phi(1))}{\sqrt{k^2+1}} + \frac{\cos(2t-\phi(2))}{2\sqrt{k^2+4}} + \frac{\cos(3t-\phi(3))}{3\sqrt{k^2+9}} + ... \\ &= \frac{1}{k} + \sum_{n=1}^{\infty} \frac{\cos(nt-\phi(n))}{n\sqrt{k^2+n^2}} \end{aligned}$$

As always, the general solution is $x(t) = x_p(t) + x_h(t)$.

25 PDEs; separation of variables

25.1 Goals

1. Be able to model the temperature of a heated bar using the heat equation plus boundary and initial conditions.

2. Be able to solve the equations modeling the heated bar using Fourier's method of separation of variables

3. Be able to model a vibrating string using the wave equation plus boundary and initial conditions.

4. Be able to solve the equations modeling the vibrating string using Fourier's method of separation of variables

25.2 Introduction

When a function depends on more than one variable, it has partial derivatives instead of ordinary derivatives. For 18.03, this means we will have to consider partial differential equations (PDE) involving such functions.

In this note we will focus on two main examples: the heat equation describing the temperature of heated metal rod and the wave equation describing the motion of a vibrating string We describe these below. In psets we will look at variations of these examples as well as extensions of our techniques to other equations.

Both examples lead to a linear partial differential equation which we will solve using the Fourier separation of variables method. Perhaps unsurprisingly, this will involve Fourier series, i.e., superposition of sines and cosines. Because there are multiple independent variables, the computations will be lengthier than we have seen before. However, the basic scheme will be the same. That is, to solve a homogeneous equation with initial conditions we:

1. Use the method of optimism to find modal solutions. In this case, there will be an infinite number of independent modal solutions.

2. The general solution is a linear combination of the modal solutions.

3. The values of the coefficients in the general solution are determined by the initial conditions. In this case, since we have an infinite number of terms in the linear combination, finding the coefficients will involve Fourier series.

For an inhomogeneous equation, the general solution is given by a particular solution plus the general homogeneous solution. We'll need some method, often the method of optimism, to find the particular solution.

The major new wrinkle will be the inclusion of what are called boundary conditions in our models. These will be explained in due course.

25.3 The heat equation

In this section we will look at the heat equation, which models the temperature over time in a heated bar.

Suppose we have a heated bar made of a uniform material. The temperature in the bar will vary with position along the bar as well as over time. To be specific, we assume we have a rod of length L which is thin enough that the temperature doesn't vary in the vertical direction. We will also make the assumption that the bar is insulated along its length so that no heat passes through the sides. (See the figure with the example in the next section.)

Given these assumptions, we can describe the temperature of the bar by a function of two variables $u(x, t)$ which gives the temperature at time t at position x.

The partial differential equation (PDE) modeling the temperature $u(x, t)$ is

$$\frac{\partial u}{\partial t}(x,t) = k\frac{\partial^2 u}{\partial x^2}(x,t). \tag{35}$$

Here, k is called the thermal conductivity of the material. It is a physical constant with dimension length^2/time.

Equation 35 is called the one-dimensional heat equation because it describes heat conduction in one dimension. The heat equation is ubiquitous in science and engineering. It models heat flow in a metal rod, diffusion of a contaminant in water, diffusion of information through a system and much more. It is a special case of an (in general nonlinear) equation called the diffusion equation.

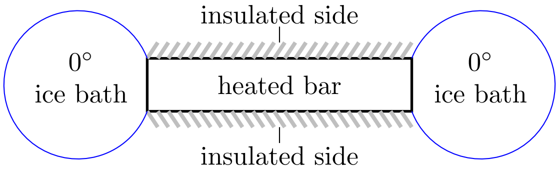
A nice derivation of the heat equation from physical principles is given in section 8.5 of the text by Edwards and Penney. Of course, you can also find many derivations on the internet.

25.3.1 Modeling a heated bar

We illustrate the modeling problem by going through one specific example. This is fairly wordy, but at the end of the example we will give a succinct summary of the model.

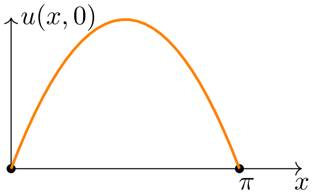
To be concrete, in this example, we'll use length in centimeters, temperature in degrees Celsius and time in seconds. We'll also let the thermal conductivity constant $k = 3$ cm/sec^2.

Suppose we have a thin heated bar of length $L = \pi$. We assume the top and bottom edges are insulated so that no heat passes through them. We also assume that the ends of the bars are in an ice bath maintained at 0°.



Over time, the temperature at various points along the bar will change. We let $u(x, t)$ be the temperature at the point x on the bar at time t.

Finally, we suppose that at time $t = 0$ the temperature over the bar is given by $u(x, 0) = x(\pi - x)$.



Initial temperature distribution (at $t = 0$).

The PDE modeling the temperature in a heated bar is given in Equation 35. With our value of k, this becomes

$$\frac{\partial u}{\partial t} = 3\frac{\partial^2 u}{\partial x^2} \quad \text{for } 0 \le x \le \pi \text{ and } t > 0.$$

We want to finish the model for $u(x,t)$ by taking into account the ice baths and the initial temperature profile.

Since the ends of the bar are in ice baths held at 0°, we have the boundary conditions (BC)

$$u(0,t) = 0 \quad \text{and } u(\pi,t) = 0 \quad \text{for all } t.$$

The term boundary refers to the endpoints or *bounds* of the interval $[0,\pi]$. The boundary conditions (BC) give the values of $u(x,t)$ when x equals one of the bounds, i.e., $x = 0$ or $x = \pi$.

We are also given the temperature in the bar at time 0. This is called the initial condition (IC):

$$u(x,0) = x(\pi - x).$$

We can summarize this as the heat equation with boundary and initial conditions:

- HE: $\frac{\partial u}{\partial t} = 3\frac{\partial^2 u}{\partial x^2}$ for $0 \le x \le \pi, \quad t > 0$.
- BC: $u(0,t) = 0$ and $u(\pi,t) = 0$ for $t \ge 0$
- IC: $u(x,0) = x(\pi - x)$ for $0 \le x \le \pi$.

25.3.2 A notational interlude

Using curvy d's to write partial derivatives is cumbersome and time consuming. Often we will use another standard notation for partial derivatives:

$$\frac{\partial u}{\partial x} = u_x, \quad \frac{\partial^2 u}{\partial x^2} = u_{xx} \quad \frac{\partial u}{\partial t} = u_t, \quad \frac{\partial^2 u}{\partial t^2} = u_{tt}.$$

With this notation our model becomes

- HE: $u_t = 3u_{xx}$ for $0 \le x \le \pi, \quad t > 0$.
- BC: $u(0,t) = 0$ and $u(\pi,t) = 0$ for $t > 0$.
- IC: $u(x,0) = x(\pi - x)$ for $0 \le x \le \pi$.

25.3.3 A strategy for solving the heat equation with boundary and initial conditions

To solve the system described above means finding a function $u(x,t)$ that satisfies all three of the criteria: HE, BC, IC. Our strategy will start by ignoring the initial condition.

1. First, we'll use the method of optimism to find simple (modal) solutions that satisfy both the partial differential equation (HE) and the boundary conditions (BC).
2. The general solution satisfying the HE and the BC will be the superposition of all the modal solutions.
3. Finally, the initial condition (IC) will let us determine the values of the coefficients in the general solution.

This outline should look familiar: it's exactly the same as the outline we used for solving linear homogeneous differential equations $P(D)x = 0$. The details of the computation will of course be different.

Before going into these details we need to check linearity and homogeneity.

25.3.4 The heat equation is linear and homogeneous

In this part we will give a quick argument showing that the heat equation is linear and homogeneous. Here's one way of explaining what we mean:

First we rewrite the heat equation to bring out the homogeneity:

$$u_t - 3u_{xx} = 0.$$

Now we define the heat operator H by $Hu = u_{tt} - 3u_{xx}$. Remember that the notation Hu should be read as 'H applied to u'. With this notation the heat equation is simply $Hu = 0$.

As usual, once we realize the need, showing that the operator H is linear is some simple algebra. That is, we must show that

$$H(c_1u_1 + c_2u_2) = c_1Hu_1 + c_2Hu_2$$

for any constants c_1, c_2. Since this is just the statement that taking partial derivatives is linear we leave it to you to verify.

This is important. Linearity is important in 18.03. You should make extra certain that you understand what is being said in this section. If it's not clear, make sure to keep asking questions until it is!

Being linear and homogeneous, linear combinations of solutions to the heat equation are also solutions, i.e., if $Hu_1 = 0$ and $Hu_2 = 0$, then $H(c_1u_1 + c_2u_2) = 0$.

25.3.5 The boundary conditions are linear and homogeneous

By linear and homogeneous boundary conditions, we mean that if two functions $u_1(x,t)$ and $u_2(x,t)$ satisfy the boundary conditions then so does any linear combination of u_1 and

u_2. This should be clear for the boundary conditions from our example: $u(0,t) = 0$ and $u(\pi,t) = 0$.

This is important redux. Linearity is important in 18.03. You should make extra certain that you understand what is being said in this section. If it's not clear make sure to keep asking questions until it is!

Note. If the boundary conditions were not 0, then they would be linear but not homogeneous. You should be able to formulate the superposition principle that they satisfy in this case.

25.4 Solving the head equation with boundary and initial conditions

We are almost ready to learn the Fourier separation of variables method. Now might be a good time to review the strategy described in the Section 25.3.3.

25.4.1 Preliminary notions

Once we get going, we will need the following notions.

Notion 1. The ordinary differential equation $X''(x) + \lambda X(x) = 0$ has 3 cases:

Case (i) If $\lambda > 0$, the solution is $X(x) = a\cos(\sqrt{\lambda}\,x) + b\sin(\sqrt{\lambda}\,x)$

Case (ii) If $\lambda = 0$, the solution is $X(x) = a + bx$

Case (iii) If $\lambda < 0$, the solution is $X(x) = ae^{\sqrt{-\lambda}\,x} + be^{-\sqrt{-\lambda}\,x}$

Notion 2. If x and t are independent variables and $f(x)$ and $g(t)$ are functions with $f(x) = g(t)$ for all x and t, then both $f(x)$ and $g(t)$ are constant functions equal to the same constant.

To wrap your mind around what is being said, you should focus on the fact that x and t are independent. This means that we can fix $x = 2$ and let t vary. So, under the assumption that $f(x) = g(t)$ for all x and t, we have

$$f(2) = g(t) \text{ for all } t.$$

Since $f(2)$ is a constant, this means that $g(t)$ is a constant function. The argument that $f(x)$ is a constant function is identical. Clearly, they both equal the same constant.

Notion 3. The function $u(x,t) \equiv 0$ satisfies both HE and BC. We call this the trivial solution. While it is a fine upstanding solution, it won't be much help in our search for modal solutions that can be used in linear combinations.

25.4.2 Fourier's method of separation of variables

We now return to our example: Our first version of the solution will be rather long-winded because we will need to explain each step in the method. Later, we will be able to give more streamlined solutions.

Example 25.1. Solve the following partial differential equation (PDE) with boundary and initial conditions (BC & IC). That is, find a function $u(x,t)$ that satisfies the following.

- HE: $u_t = 3u_{xx}$, where $0 \le x \le \pi$ and $t > 0$
- BC: $u(0,t) = 0$ and $u(\pi,t) = 0$, where $t > 0$
- IC: $u(x,0) = x(\pi - x)$, where $0 \le x \le \pi$.

Solution: Step 1 Separated solutions. Our first trick is to use the method of optimism to look for a solution of the form

$$u(x,t) = X(x)T(t).$$

This is called a separated solution because it is a function of x times a function of t. There is no reason to expect that *all* solutions are separated, but that doesn't mean we won't find some useful solutions this way.

Having guessed a trial solution, we substitute it into the partial differential equation HE. This gives:

$$X(x)T'(t) = 3X''(x)T(t).$$

Now we separate the equation so the x's are on one side and the t's are on the other

$$\frac{T'(t)}{3T(t)} = \frac{X''(x)}{X(x)}.$$

Note, the convention is to keep the coefficient 3 with the T. Please do this, it will make your life easier. Now, preliminary Notion 2, comes into play: the left side is a function of t and the right side is a function of x, so both must be constant functions equal to the same constant!

We can call this constant anything we want. Because we know it will help with the algebra that is coming, we call it $-\lambda$. This too is just a convention, but you should do it so λ will mean the same thing for everyone in ES.1803. We have

$$\frac{X''(x)}{X(x)} = -\lambda; \qquad \frac{T'(t)}{3T(t)} = -\lambda.$$

A tiny bit of algebra gives the two ordinary differential equations

$$X''(x) + \lambda X(x) = 0; \qquad T'(t) + 3\lambda T(t) = 0.$$

Now we appeal to our preliminary Notion 1 to look at the 3 cases for λ. In all three cases $u(x,t) = X(x)T(t)$.

Case (i) $\lambda > 0$: $\quad X(x) = a\cos(\sqrt{\lambda}x) + b\sin(\sqrt{\lambda}x), \qquad T(t) = e^{-3\lambda t}.$

Case (ii) $\lambda = 0$: $\quad X(x) = a + bx, \qquad T(t) = c.$

Case (iii) $\lambda < 0$: $\quad X(x) = ae^{\sqrt{-\lambda}x} + be^{-\sqrt{-\lambda}x}, \qquad T(t) = ce^{-3\lambda t}.$

Case (iii) is ugly. Notice that $-\lambda$ is positive so the square roots are real numbers and so these are actually real-valued solutions. Fortunately, we will soon see that we can ignore it since it only gives the trivial solution satisfying the partial differential equation HE and the boundary conditions BC.

The method of optimism was wildly successful. We have lots of solutions to HE. We can get a separated solution to HE by picking any value of λ and then any values for a, b, c.

Step 2 Boundary conditions (BC). The model also has boundary conditions. So we need to see which of the separated solutions to the partial differential equation HE also satisfy the boundary conditions BC. Such solutions are called modal solutions.

For a separated solution $u(x,t) = X(x)T(t)$, the boundary conditions are

$$u(0,t) = X(0)T(t) = 0 \quad \text{and} \quad u(\pi,t) = X(\pi)T(t) = 0.$$

Being extra careful: this means that either $X(0) = X(\pi) = 0$ or $T(t) = 0$. The case $T(t) = 0$ gives the trivial solution $u(x,t) = X(x)T(t) = 0$. Since it is trivial, we ignore this case. So (nontrivial) separated solutions satisfying both HE and BC must have

$$X(0) = 0 \quad \text{and} \quad X(\pi) = 0. \tag{36}$$

Now we'll look at each case in turn.

Case (i). $\lambda > 0$: $X(x) = (a\cos(\sqrt{\lambda}x) + b\sin(\sqrt{\lambda}x))$. The boundary conditions give

$$X(0) = a = 0 \quad \text{and} \quad X(\pi) = a\cos(\sqrt{\lambda}\pi) + b\sin(\sqrt{\lambda}\pi) = 0.$$

Solving, we see that $a = 0$ and either $b = 0$ or $\sin(\sqrt{\lambda}\pi) = 0$. The choice $b = 0$ gives us the trivial solution, so we ignore it. The other choice, $\sin(\sqrt{\lambda}\pi) = 0$ gives $\sqrt{\lambda}\pi = n\pi$ for some integer n. So, for each $\sqrt{\lambda} = n$ $(\lambda = n^2)$, we have the following separated solutions that satisfy both HE and BC:

$$u_n(x,t) = X_n(x)T_n(t), \text{ where } X_n(x) = b_n\sin(nx), \text{ and } T_n(t) = c_n e^{-3n^2t}.$$

Note, we name the solutions and coefficients with the subscript n so we can tell them apart.

A simplification: in the product we can combine b_n into c_n one constant so,

$$\boxed{u_n(x,t) = b_n\sin(nx)\,e^{-3n^2t} \quad \text{for } n = 1, 2, 3, \ldots}$$

are the separated solutions for case (i) which satisfy both HE and BC.

In this case, the boundary condition weeded out most values of $\lambda > 0$.

Case (ii). $\lambda = 0$: $X(x) = a + bx$. The boundary conditions are

$$X(0) = a = 0 \qquad X(\pi) = a + b\pi = 0.$$

It is easy to see that the only solutions to these equations are $a = 0$, $b = 0$. That is, this case only produces trivial solutions and we can ignore it.

Note well: With other boundary conditions this case may produce nontrivial solutions. So we always have to check.

Case (iii). $\lambda < 0$: $X(x) = ae^{\sqrt{-\lambda}\,x} + be^{-\sqrt{-\lambda}\,x}$. The boundary conditions are

$$X(0) = a + b = 0 \qquad X(\pi) = ae^{\sqrt{-\lambda}\,\pi} + be^{-\sqrt{-\lambda}\,\pi} = 0.$$

In matrix form the equation is

$$\begin{bmatrix} 1 & 1 \\ e^{\sqrt{-\lambda}\,\pi} & e^{-\sqrt{-\lambda}\,\pi} \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

The determinant of the coefficient matrix is $e^{-\sqrt{-\lambda}\,\pi} - e^{\sqrt{-\lambda}\,\pi}$. Since $\lambda \neq 0$ the determinant is not 0. Therefore, we only have the trivial solution $a = 0$, $b = 0$. That is, this case only yields the trivial solution to HE and BC. So we ignore it!

It turns out, this case will never give nontrivial solutions. So this is the first and last time we will do the algebra for this case. **In the future we will just say that Case (iii) only has the trivial solution and ignore it.**

Note. All the separated solutions satisfying both HE and BC are called normal modes or modal solutions for this system.

We have now found all the modal solutions.

Step 3 Superposition. Because both the PDE (HE) and the boundary conditions (BC) are linear and homogeneous, the general solution satisfying them both is given by superposition of all the modal solutions:

$$\boxed{u(x,t) = \sum_{n=1}^{\infty} u_n(x,t) = \sum_{n=1}^{\infty} b_n \sin(nx)\, e^{-3n^2 t}}.$$

Step 4 Use the IC to find the coefficients. We are now ready to use the initial conditions (IC) to determine the values of the coefficients b_n in our general solution.

IC: $u(x,0) = \sum b_n \sin(nx) = x(\pi - x)$. Therefore, b_n are the Fourier sine coefficients of $x(\pi - x)$.

$$b_n = \frac{2}{\pi}\int_0^{\pi} x(\pi - x)\sin(nx)\,dx = \begin{cases} 8/(\pi n)^3 & \text{for } n \text{ odd} \\ 0 & \text{for } n \text{ even.} \end{cases}$$

(The full computation of this integral is shown in the section at the end of these notes.) Thus, our solution is

$$\boxed{u(x,t) = \sum b_n \sin(nx)\, e^{-3n^2 t} = \sum_{n \text{ odd}} \frac{8}{(n\pi)^3} \sin(nx)\, e^{-3n^2 t}.}$$

25.5 Summary of Fourier's method

Once again we summarize Fourier's method for homogeneous PDEs with homogeneous boundary conditions.

1. Find separated solutions to the PDE: one parametrized family for each λ.

2. The boundary conditions (BC) restrict the λ to an indexed set of values. They also restrict the possible values of the parameters in each family.

3. Superposition gives the general solution satisfying both the PDE and BC.

4. Use the initial conditions to determine the values of the coefficients in the general solution..

25.6 Model solution

Because the first time through took several pages, we redo the solution to the previous example in model form. But this comes with a **WARNING**: do not just memorize this routine. You should remember the reasons for each of the steps. Different problems will use variations on these themes and you have to be prepared to use the reasoning, but not the exact details, from this example.

Example 25.2. (Model solution.) Solve for $u(x,t)$ on $0 \leq x \leq \pi$ and $t > 0$ satisfying

- **HE:** $u_t = 3u_{xx}$.
- **BC:** $u(0,t) = 0$ and $u(\pi,t) = 0$.
- **IC:** $u(x,0) = x(\pi - x)$.

Solution: Step 1. Look for separated solutions: $u(x,t) = X(x)T(t)$ to the PDE.

Substitution into HE: $XT' = 3X''T$.

Algebra: $X''(x)/X(x) = T'(t)/(3T(t)) =$ constant $= -\lambda$.

More algebra: $X'' + \lambda X = 0$, $T' + 3\lambda T = 0$. There are three cases:

Case (i) $\lambda > 0$: $X(x) = a\cos(\sqrt{\lambda}x) + b\sin(\sqrt{\lambda}x)$, $T(t) = ce^{-3\lambda t}$.

Case (ii) $\lambda = 0$: $X(x) = a + bx$, $T(t) = c$.

Case (iii) $\lambda < 0$. Always ignore, since this case only gives the trivial solution satisfying the PDE and boundary conditions.

Step 2. Modal solutions. Find the separated solutions which also satisfy the BC.

For separated solutions, the BC are $X(0) = 0$, $X(\pi) = 0$.

Case (i) The BC are

$$X(0) = a = 0 \quad \text{and} \quad X(\pi) = a\cos(\sqrt{\lambda}\pi) + b\sin(\sqrt{\lambda}\pi) = 0.$$

Since $a = 0$, the second condition is $b\sin(\sqrt{\lambda}\pi) = 0$. For nontrivial solutions, we need $\sqrt{(\sqrt{\lambda}\pi)} = 0$, i.e., $\sqrt{\lambda}\pi = n\pi$ for n an integer.

We have found modal solutions

$$u_n(x,t) = b_n \sin(nx)\, e^{-3n^2 t} \qquad \text{for } n = 1, 2, 3, \ldots$$

Case (ii) The BC are $X(0) = a = 0$, $X(\pi) = a + b\pi = 0$.

This has only the trivial solution $a = 0$, $b = 0$.

Case (iii) Ignored – only has the trivial solution.

Step 3. Both HE and BC are homogeneous, so, by superposition, the general solution satisfying both is

$$\boxed{u(x,t) = \sum_{n=1}^{\infty} u_n(x,t) = \sum_{n=1}^{\infty} b_n \sin(nx)\, e^{-3n^2 t}}.$$

Step 4. Use the initial conditions to find the values of the coefficients.

IC: $u(x,0) = \sum_{n=1} b_n \sin(nx) = x(\pi - x)$. This is the Fourier sine series for $x(\pi - x)$. Now (see the computation section below) the coefficients are

$$\boxed{b_n = \begin{cases} 8/(\pi n)^3 & \text{for } n \text{ odd} \\ 0 & \text{for } n \text{ even.} \end{cases}}$$

So,

$$\boxed{u(x,t) = \sum_{n \text{ odd}} \frac{8}{\pi n^3} \sin(nx)\, e^{-3n^2 t}}.$$

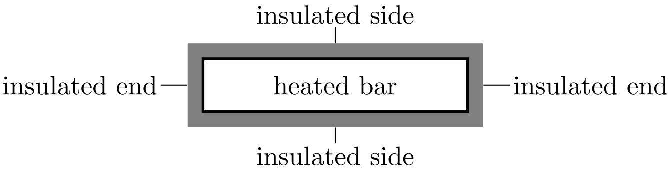
25.7 Another example with different boundary conditions

Here's an example with different boundary conditions. In this example, we will see that the case $\lambda = 0$ has nontrivial solutions.

Example 25.3. Suppose we have a heated rod of length L as described above. Assume that the ends of the bar are also insulated, so that no heat leaves the bar. Also assume that the initial temperature of the bar is given by $u(x,0) = x(L-x)\,^\circ C$.

Write down a PDE with boundary and initial conditions that models the temperature in the bar. Then use Fourier's separation of variables method to solve the system.

Solution: The physical setup is illustrated in the figure below.



Heated rod insulated on all sides

First we set up the model: The PDE is just the heat equation given in (35):

(HE) $$\frac{\partial u}{\partial t} = k\frac{\partial^2 u}{\partial x^2} \quad \text{for } t > 0 \text{ and } 0 \le x \le L.$$

We are not given enough information to determine k, so we leave it as an unspecified parameter.

Because the ends of the rod are insulated, the temperature gradient at the ends is 0. This translates to the boundary conditions:

(BC) $$\frac{\partial u}{\partial x}(0,t) = 0 \quad \text{and} \quad \frac{\partial u}{\partial x}(L,t) = 0 \quad \text{for } t > 0.$$

Note that these are homogeneous boundary conditions.

We are given the initial condition (i.e., the temperature at time 0) directly:

(IC) $$u(x,0) = x(L-x) \quad \text{for } 0 \le x \le L.$$

Next we solve using the method of separation of variables.

Step 1. Look for separated solution: $u(x,t) = X(x)T(t)$ to the PDE.

Substitution into HE: $XT' = kX''T$.

Algebra: $X''(x)/X(x) = T'(t)/(kT(t)) =$ constant $= -\lambda$.

More algebra: $X'' + \lambda X = 0, \quad T' + k\lambda T = 0$. There are three cases:

Case (i) $\lambda > 0$: $X(x) = a\cos(\sqrt{\lambda}x) + b\sin(\sqrt{\lambda}x)$, $T(t) = ce^{-k\lambda t}$.

Case (ii) $\lambda = 0$: $X(x) = a + bx, \quad T(t) = c$.

Case (iii) $\lambda < 0$. Always ignore, since this case only gives the trivial solution satisfying the PDE and boundary conditions.

Step 2. (Modal solutions) Find the separated solutions in Step 1 which also satisfy the boundary conditions.

For separated solutions, the BC are $X'(0) = 0, \quad X'(L) = 0$.

Case (i) $X'(0) = \sqrt{\lambda}b = 0$ and $X'(L) = -\sqrt{\lambda}a\sin(\sqrt{\lambda}L) + \sqrt{\lambda}b\cos(\sqrt{\lambda}L)$.

This has nontrivial solutions when $b = 0$ and $\sqrt{\lambda}L = n\pi$ for n an integer. That is, when $\sqrt{\lambda} = n\pi/L$. For this case, the modal solutions are

$$u_n(x,t) = a_n \cos\left(\frac{n\pi}{L}x\right) e^{-k(n\pi/L)^2 t} \qquad \text{for } n = 1, 2, 3, ...$$

Case (ii) $X'(0) = b = 0$, $X'(L) = b = 0$. This has nontrivial solutions $X(x) = a$. So, for this case, the modal solutions are $X(x)T(t) = ac$. We write this as

$$u_0(x,t) = \frac{a_0}{2}.$$

Step 3. Both HE and BC are homogeneous, so by superposition the general solution satisfying both is

$$\boxed{u(x,t) = \sum_{n=0}^{\infty} u_n(x,t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} b_n \cos\left(\frac{n\pi}{L}x\right) e^{-k(n\pi/L)^2 t}.}$$

(We gave the DC term as $a_0/2$ so we don't forget the factor of 1/2 when we do the computation below.)

Step 4. Use the initial conditions to find the coefficients.

$u(x,0) = \dfrac{b_0}{2} + \displaystyle\sum_{n=1} b_n \cos\left(\frac{n\pi}{L}x\right) = x(L-x)$. This is the Fourier cosine series for $x(L-x)$.

Now (see the computation section below) the coefficients are

$$\boxed{b_0 = L^2/3, \quad b_n = \begin{cases} -4L^2/(n\pi)^2) & \text{for } n \text{ even} \\ 0 & \text{for } n \text{ odd.} \end{cases}}$$

We can leave our answer as a set of boxes or put them together in one box.

$$\boxed{u(x,t) = \frac{L^2}{6} - \sum_{n \text{ even}} \frac{4L^2}{(n\pi)^2} \cos\left(\frac{n\pi}{L}x\right) e^{-k(n\pi/L)^2 t}.} \qquad (37)$$

25.7.1 Interpreting the solution

Writing out the terms in the solution Equation 37 above we have

$$u(x,t) = \frac{L^2}{6} - \frac{4L^2}{(2\pi)^2}\cos\left(\frac{2\pi}{L}x\right)e^{-k(2\pi/L)^2t} - \frac{4L^2}{(4\pi)^2}\cos\left(\frac{4\pi}{L}x\right)e^{-k(4\pi/L)^2t}$$
$$- \frac{4L^2}{(6\pi)^2}\cos\left(\frac{6\pi}{L}x\right)e^{-k(6\pi/L)^2t}\ldots$$

The first thing to note is that all the terms after the constant have decaying exponentials in time. This means that, in the long run, the bar will come to an equilibrium temperature of $L^2/6$. It makes intuitive sense that the temperature in the bar will even out over time. Looking at the expression for the constant (DC) term

$$\frac{c_0}{2} = \frac{1}{L}\int_0^L x(L-x)\,dx$$

we see that it is the average value of the initial temperature distribution. This too makes intuitive sense.

The second thing we want to note is that, after a very short time, the solution is well approximated by the DC term and the first non-zero harmonic

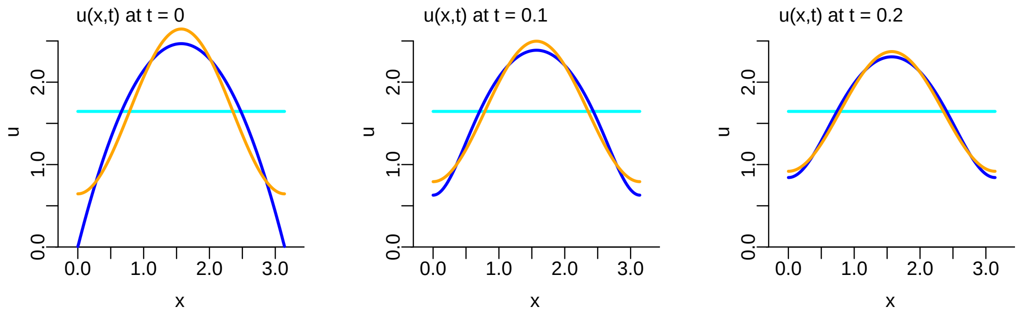
$$u(x,t) \approx \frac{L^2}{6} - \frac{4L^2}{(2\pi)^2}\cos\left(\frac{2\pi}{L}x\right)e^{-k(2\pi/L)^2t}$$

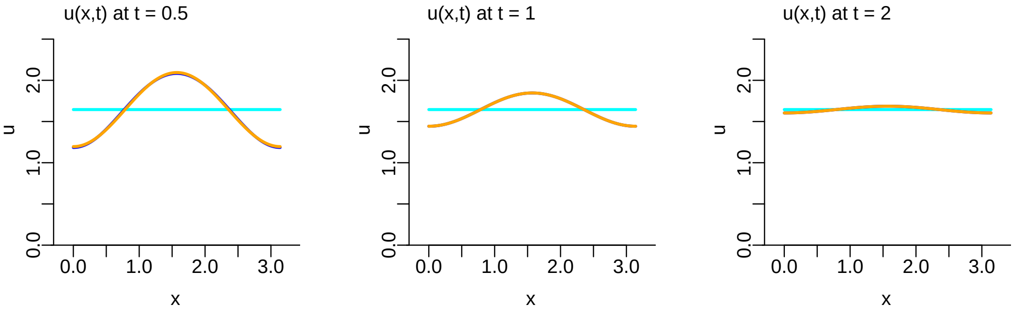
To see this look at the exponents in the time exponentials:

$$e^{-\frac{k4\pi^2}{L^2}t}, \quad e^{-\frac{k16\pi^2}{L^2}t}, \quad e^{-\frac{k36\pi^2}{L^2}t}, \ldots$$

The later exponents are so much more negative than the first one that the later exponentials rapidly become negligible compared to the first.

Here are a sequence of plots showing the exact solution, the long term equilibrium and the approximation by the DC term plus the first non-zero harmonic. (They were made with $L = \pi$ and $k = 0.4$.) Notice how well the approximation matches the exact solution after a short time. Also notice how the solution goes to the equilibrium over time.





Blue = exact sol., cyan = equilibrium, orange = $\frac{L^2}{6} - \frac{4L^2}{(2\pi)^2}\cos\left(\frac{2\pi}{L}x\right)e^{-k(2\pi/L)^2 t}$

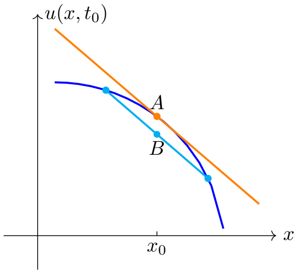
As usual, there is an applet giving a dynamic illustration of the heat equation: https://mathlets.org/mathlets/heat-equation/.

25.7.2 A mathematical explanation of the heat equation

Suppose the temperature along the bar is given by $u(x,t)$. That is, this is the temperature at point x at time t.

Let's fix a time t_0 and a position x_0. It's reasonable to assume that if the average temperature of nearby points is lower than that at x_0, then the point at x_0 will be losing heat, i.e., the rate the temperature changes will be negative. Likewise, if the average temperature of nearby points is greater than that at x_0 then the rate the temperature changes will be positive.

The graph below shows the temperature distribution $u(x,t_0)$ at a single time t_0. The temperature at x_0 is marked by the point A on the curve. The average temperature of the two points (equally spaced around x_0) is shown as the point B on the secant line between the two points. Since the curve is concave down, this average B is below A, i.e., the average temperature is lower than the temperature at x_0. So the rate, $\frac{\partial u}{\partial t}(x_0,t_0)$, the temperature is changing is negative.



The concavity of the curve at x_0 is measured by $\frac{\partial^2 u}{\partial x^2}(x_0,t_0)$. Since the curve is concave down, this is negative, i.e., the same sign as the rate the temperature is changing. Thus, at least to a first approximation, the rate the temperature is changing is given by the heat equation

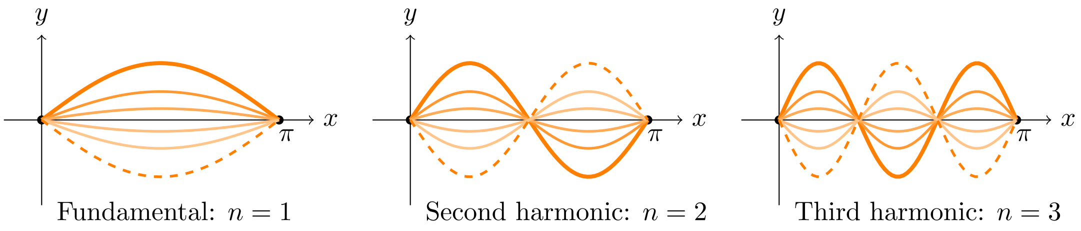
$$\frac{\partial u}{\partial t}(x_0,t_0) = k\frac{\partial^2 u}{\partial x^2}(x_0,t_0).$$

25.8 Physics of a vibrating string: the wave equation

Suppose we have a string or wire of length L tied tightly between two posts, If we start it vibrating it will give off a sound. That is, it will create a pressure wave that will strike our ears and we will hear a sound.

If we are careful, we can make the starting shape a perfect sine curve and then, when we let the string go, it will vibrate while always maintaining its sine curve shape with only the amplitude changing in time. (This is not at all obvious, at least to me, but it will come out in our analysis of the wave equation.) This perfect sine curve shaped vibration is called a normal or pure mode. In this mode the string will emit a pure tone. The twanginess of most vibrating strings tells you that they do not spontaneously vibrate in a normal mode.

For $L = \pi$ the first 3 normal modes have starting shape $y = b_1 \sin(x)$, $y = b_2 \sin(2x)$ and $y = b_3 \sin(3x)$. These are called the first or fundamental harmonic, second harmonic and third harmonic respectively. We illustrate this in the pictures below.

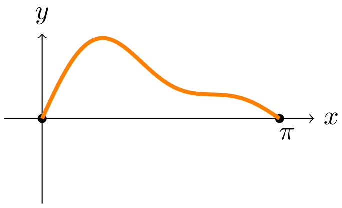


Fundamental: $n = 1$ Second harmonic: $n = 2$ Third harmonic: $n = 3$

An even better way to visualize normal modes is to go to the applet https://mathlets.org/mathlets/wave-equation/. Refresh the page so the initial condition is set to its default. Set $n = 3$, and leave only the *harmonics* checkbox checked. Then start the animation by clicking the >> button. You will see each of the modes vibrating. Then check the *Fourier sum* checkbox and you will see the superposition of the 3 harmonics.

The twanginess of a vibrating string comes because the vibration is really a superposition or mixture of the many normal modes. The figure shows the starting position for a mixture of the first 3 modes (fundamental, first and second harmonics) with amplitudes 1, 0.5 and 0.3 respectively. That is

$$y = \sin(x) + 0.5\sin(2x) + 0.3\sin(3x)$$



Mixture: $y = \sin(x) + 0.5\sin(2x) + 0.3\sin(3x)$

25.8.1 Physical model of a vibrating string

This is well explained in just 2 paragraphs in §8.6 of the textbook by Edwards and Penney. They then derive the partial differential equation modeling the vibrating string. We give a quick summary of the terminology and model here. You should read the text or look on the internet to see the derivation.

The physical assumption is that each point on the string only moves up and down in the y-direction, i.e., there is no side-to-side movement. This and their other assumptions are reasonable for strings that are much longer than the amplitude of their vibration.

For a point x on the string we let

$$y(x,t) = \text{ displacement of the point } x \text{ at time } t.$$

Assuming small displacements, this is well modeled by the following partial differential equation, called the wave equation

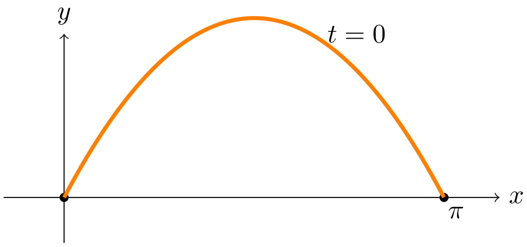
$$\frac{\partial^2 y}{\partial t^2} = a^2 \frac{\partial^2 y}{\partial x^2} \tag{38}$$

Here a is a constant which depends on the physical characteristics of the string as well as its tension and length, it is the called the speed of the wave. You should check that it does indeed have units of speed.

25.9 The wave equation with boundary and initial conditions.

25.9.1 Modeling a vibrating string

We illustrate the modeling problem by going through one specific example. Suppose we have a string of length $L = \pi$ meters which is clamped at both ends. As the string vibrates, let $y(x,t)$ be the displacement in meters of the point x on the string at time t in seconds. Suppose at time $t = 0$ the string is stationary and has shape $y(x,0) = x(\pi - x)$.



Initial shape of the string (at $t = 0$).

We want to find a model for $y(x,t)$. The model will consist of a partial differential equation (PDE) and some extra conditions. For this example, assume the wave speed is 3 m/sec.

Above we asserted that the PDE modeling a vibrating string is given in Equation 38. With our units this becomes

$$\frac{\partial^2 y}{\partial t^2} = 9\frac{\partial^2 y}{\partial x^2} \qquad \text{for } 0 \le x \le \pi \text{ and } t > 0.$$

Since the ends are clamped they cannot move. That is, the points on the string at $x = 0$ and $x = \pi$ are fixed, i.e., we have the boundary conditions (BC)

$$y(0,t) = 0 \qquad \text{and} \qquad y(\pi,t) = 0 \qquad \text{for all } t.$$

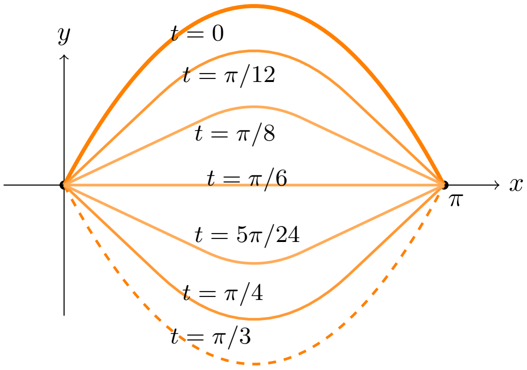
We are also given the shape (displacement) and velocity of the string at time 0. These are the initial conditions (IC). The initial shape was given as $x(\pi - x)$. The initial velocity is 0, because the string is momentarily stationary at $t = 0$. Since shape at time 0 is $y(x, 0)$ and the velocity is $\frac{\partial y}{\partial t}(x, 0)$ we have the initial conditions

$$y(x, 0) = x(\pi - x) \quad \text{and} \quad \frac{\partial y}{\partial t}(x, 0) = 0.$$

We can summarize this as the wave equation with boundary and initial conditions:

- WE: $\frac{\partial^2 y}{\partial t^2} = 9\frac{\partial^2 y}{\partial x^2}$ for $0 \le x \le \pi$, $t > 0$.
- BC: $y(0, t) = 0$ and $y(\pi, t) = 0$ for $t \ge 0$
- IC: $y(x, 0) = x(\pi - x)$ and $\frac{\partial y}{\partial t}(x, 0) = 0$ for $0 \le x \le \pi$.

The figure below shows that shape of the string at various points in time. Note that the boundary points don't move because of the clamped end boundary conditions.



25.9.2 Solving the wave equation with boundary and initial conditions

As with our heat equation examples, we will use Fourier's method of separation of variables to solve the wave equation with the given BC and IC.

Example 25.4. Solve the following partial differential equation (PDE) with boundary and initial conditions (BC & IC). That is, find a function $y(x, t)$ that satisfies the following.

- WE: $y_{tt} = 9y_{xx}$, for $0 \le x \le \pi$ and $t > 0$
- BC: $y(0, t) = 0$ and $y(\pi, t) = 0$, for $t > 0$
- IC: $y(x, 0) = x(\pi - x)$ and $y_t(x, 0) = 0$, for $0 \le x \le \pi$.

Solution: Step 1. Look for separated solution: $y(x, t) = X(x)T(t)$ to the PDE.

Substitution into WE: $XT'' = 9X''T$.

Algebra: $X''(x)/X(x) = T''(t)/(9T(t)) =$ constant $= -\lambda$.

More algebra: $X'' + \lambda X = 0$, $T'' + 9\lambda T = 0$. There are three cases:

Case (i) $\lambda > 0$: $X(x) = a\cos(\sqrt{\lambda}x) + b\sin(\sqrt{\lambda}x)$, $T(t) = c\cos(3\sqrt{\lambda}t) + d\sin(3\sqrt{\lambda}t)$.

Case (ii) $\lambda = 0$: $X(x) = a + bx$, $T(t) = c + dt$.

Case (iii) $\lambda < 0$. Always ignore, since this case only gives the trivial solution satisfying the PDE and boundary conditions.

Step 2. Find which of the solutions in Step 1 also satisfy the boundary conditions $X(0) = 0$, $X(\pi) = 0$.

Case (i) $X(0) = a = 0$ and $X(\pi) = a\cos(\sqrt{\lambda}\pi) + b\sin(\sqrt{\lambda}\pi) = 0$. This has nontrivial solutions when $a = 0$ and $\sqrt{\lambda}\pi = n\pi$ for n an integer. So, in this case, the nontrivial solutions to the PDE satisfying the BC are

$$y_n(x,t) = \sin(nx)(c_n\cos(3nt) + d_n\sin(3nt)) \qquad \text{for } n = 1, 2, 3, ...$$

Case (ii) $X(0) = a = 0$, $X(\pi) = a + b\pi = 0$ has only the trivial solution.

Case (iii) Ignored –only has the trivial solution.

Step 3. Both WE and BC are homogeneous, so by superposition the general solution satisfying both is

$$\boxed{y(x,t) = \sum_{n=1}^{\infty} y_n(x,t) = \sum_{n=1}^{\infty} \sin(nx)(c_n\cos(3nt) + d_n\sin(3nt)).}$$

Step 4. Use the initial conditions to find the coefficients.

First IC: $y(x,0) = \sum_{n=1} c_n\sin(nx) = x(\pi - x)$. This is the Fourier sine series for $x(\pi - x)$. Now (see the computation section below) the coefficients are

$$\boxed{c_n = \begin{cases} 8/(\pi n)^3 & \text{for } n \text{ odd} \\ 0 & \text{for } n \text{ even.} \end{cases}}$$

Second IC: $y_t(x,0) = \sum_{n=1}\sin(nx)3nd_n = 0$. This means that $3nd_n$ are the Fourier sine coefficients for $g(x) = 0$. That is, $\boxed{d_n = 0 \text{ for all } n.}$

We can leave our answer as a set of boxes or put them together in one box.

$$\boxed{y(x,t) = \sum_{n \text{ odd}} \frac{8}{\pi n^3}\sin(nx)\cos(3nt).}$$

25.10 General initial conditions

Example 25.5. Suppose we have the same PDE and BC as in the above example, but the IC are $y(x,0) = f(x)$, $y_t(x,0) = g(x)$. Solve for $y(x,t)$ in terms of the Fourier sine and cosine series f and g.

Solution: Note: since $f(x)$ and $g(x)$ are not specified, the best we can hope to do is give the solution in terms of them.

Since the partial differential equation and boundary conditions are the same, we get the same general solution

$$\boxed{y(x,t) = \sum_{n=1} y_n(x,t) = \sum_{n=1} \sin(nx) \cdot (c_n \cos(3nt) + d_n \sin(3nt)).}$$

First IC: $y(x,0) = \sum_{n=1} c_n \sin(nx) = f(x)$. Therefore, c_n are the Fourier sine coefficients of $f(x)$ on $[0,\pi]$. That is, $\boxed{c_n = \frac{2}{\pi}\int_0^\pi f(x)\sin(nx)\,dx}$.

Second IC: $y_t(x,0) = \sum 3n\,d_n \sin(nx) = g(x)$. Therefore, $3n\,d_n$ are the Fourier sine coefficients of $g(x)$ on $[0,\pi]$. That is,

$$3nd_n = \frac{2}{\pi}\int_0^\pi g(x)\sin(nx)\,dx \qquad \text{or} \qquad \boxed{d_n = \frac{2}{3n\pi}\int_0^\pi g(x)\sin(nx)\,dx.}$$

25.11 Appendix: Fourier sine and cosine coefficients of $x(L-x)$

We sketch the computation for the Fourier sine and cosine coefficients of $x(L-x)$. The actual integrals can be done by parts or by inspection.

Sine coefficients.

$$\begin{aligned}
c_n &= \frac{2}{L}\int_0^L x(L-x)\sin\left(\frac{n\pi}{L}x\right)\,dx = \frac{2}{L}\int_0^L xL\sin\left(\frac{n\pi}{L}x\right) - x^2\sin\left(\frac{n\pi}{L}x\right)\,dx \\
&= \frac{2}{L}\left[-x\frac{L}{n\pi/L}\cos\left(\frac{n\pi}{L}x\right) + \frac{L}{(n\pi/L)^2}\sin\left(\frac{n\pi}{L}x\right) + \right. \\
&\qquad \left. x^2\frac{1}{n\pi/L}\cos\left(\frac{n\pi}{L}x\right) - 2x\frac{1}{(n\pi/L)^2}\sin\left(\frac{n\pi}{L}x\right) - \frac{2}{(n\pi/L)^3}\cos\left(\frac{n\pi}{L}x\right)\right]_0^L \\
&= \frac{2}{L}\left[-\frac{L^2}{n\pi/L}\cos(n\pi) + \frac{L^2}{n\pi/L}\cos(n\pi) - \frac{2}{(n\pi/L)^3}(\cos(n\pi)-1)\right] \\
&= \begin{cases} 8L^2/(n\pi)^3 & \text{if } n \text{ is odd} \\ 0 & \text{if } n \text{ is even.} \end{cases}
\end{aligned}$$

Cosine coefficients.

$$\begin{aligned}
c_0 &= \frac{2}{L}\int_0^L x(L-x)\,dx = \frac{2}{L}\left[\frac{Lx^2}{2} - \frac{x^3}{3}\right]_0^L = \frac{L^2}{3}. \\
c_n &= \frac{2}{L}\int_0^L x(L-x)\cos\left(\frac{n\pi}{L}x\right)\,dx = \frac{2}{L}\int_0^L xL\cos\left(\frac{n\pi}{L}x\right) - x^2\cos\left(\frac{n\pi}{L}x\right)\,dx \\
&= \frac{2}{L}\left[x\frac{L}{n\pi/L}\sin\left(\frac{n\pi}{L}x\right) + \frac{L}{(n\pi/L)^2}\cos\left(\frac{n\pi}{L}x\right) + \right. \\
&\qquad \left. -x^2\frac{1}{n\pi/L}\sin\left(\frac{n\pi}{L}x\right) - 2x\frac{1}{(n\pi/L)^2}\cos\left(\frac{n\pi}{L}x\right) + \frac{2}{(n\pi/L)^3}\sin\left(\frac{n\pi}{L}x\right)\right]_0^L \\
&= \frac{2}{L}\left[\frac{L}{(n\pi/L)^2}(\cos(n\pi)-1) - \frac{2L}{(n\pi/L)^2}(\cos(n\pi))\right] \\
&= \begin{cases} -4L^2/(n\pi)^2 & \text{if } n \text{ is even} \\ 0 & \text{if } n \text{ is odd.} \end{cases}
\end{aligned}$$

26 PDEs continued

26.1 Goals

1. Reinforce the goals from Topic 25.

26.2 Introduction

The main goal in this topic is to give one more example of the wave equation. This time with boundary conditions that are different from all our previous examples.

As a bonus we also discuss a different method of solving the wave equation called the d'Alembert solution. This is nice, but it only applies to the undamped wave equation. In contrast, the Fourier method applies to many other systems, including the heat equation and the damped wave equation.

As a further bonus we walk through the ratio of frequencies for various musical intervals.

26.3 An example with different BC

Example 26.1. On a string of length $L = \pi$ find $y(x,t)$ satisfying

WE: $y_{tt} = 9y_{xx}$

BC: $y_x(0,t) = 0, \;\; y_x(\pi,t) = 0$

IC: $y(x,0) = f(x), \;\; y_t(x,0) = 0.$

Solution: Note: these are not clamped end boundary conditions. Rather, it is the first partial in x that is 0 at the boundary.

Step 1. Look for separated solutions $y(x,t) = X(x)T(t)$ to WE.

Substitution of $y(x,t) = X(x)T(t)$ into WE gives $XT'' = 9X''T$.

Algebra: $X''(x)/X(x) = T''(t)/(9T(t)) =$ constant $= -\lambda$.

More algebra: $X'' + \lambda X = 0, \quad T'' + 9\lambda T = 0$.

There are three cases:

Case (i) $\lambda > 0$: $\quad X(x) = a\cos(\sqrt{\lambda}x) + b\sin(\sqrt{\lambda}x), \quad T(t) = c\cos(3\sqrt{\lambda}t) + d\sin(3\sqrt{\lambda}t)$.

Case (ii) $\lambda = 0$: $\quad X(x) = a + bx, \quad T(t) = c + dt$.

Case (iii) $\lambda < 0$. Always ignore since this case only gives the trivial modal solutions.

Step 2. (Modal solutions) Find the separated solutions from Step 1 which also satisfy the boundary conditions.

For separated solutions, the BC are $X'(0) = 0, \quad X'(\pi) = 0$.

Case (i) BC: $X'(0) = \sqrt{\lambda}b = 0$ and $X'(\pi) = -\sqrt{\lambda}a\sin(\sqrt{\lambda}\pi) = 0$.

This has nontrivial solutions when $b = 0$ and $\sqrt{\lambda} = n$ for n an integer. So, in this case, the nontrivial solutions to the PDE satisfying the BC are

$$y_n(x,t) = \cos(nx)(c_n\cos(3nt) + d_n\sin(3nt)) \qquad \text{for } n = 1, 2, 3, \ldots$$

Case (ii) BC: $X'(0) = b = 0$, $X'(\pi) = b = 0$.

So, $X(x) = a$, $T(t) = c + dt$. The factor of a is redundant, so, in this case, the modal solutions is $y(x,t) = c + dt$. As usual with the constant terms, we write this as

$$y_0(x,t) = \frac{c_0}{2} + \frac{d_0 t}{2}$$

Case (iii) Ignored.

Step 3. Both WE and BC are homogeneous, so by superposition we have

$$y(x,t) = \sum_{n=0} y_n(x,t) = \frac{c_0}{2} + \frac{d_0 t}{2} + \sum_{n=1}^{\infty} \cos(nx) \cdot (c_n\cos(3nt) + d_n\sin(3nt))$$

is a solution to WE and BC.

Step 4. Use the initial conditions to find the coefficients.

First IC: $y(x,0) = \frac{c_0}{2} + \sum_{n=1}^{\infty} c_n\cos(nx) = f(x)$. That is, we have the Fourier cosine series for $f(x)$.

$$\boxed{c_0 = \frac{2}{\pi}\int_0^{\pi} f(x)\,dx}, \qquad \boxed{c_n = \frac{2}{\pi}\int_0^{\pi} f(x)\cos(nx)\,dx.}$$

Second IC: $y_t(0) = \frac{d_0}{2} + \sum \cos(nx)3nd_n \Rightarrow d_n = 0$ for $n = 0, 1, 2, \ldots$

So we have our solution to the system (WE, BC, IC):

$$y(x,t) = \frac{c_0}{2} + \sum_{n=1}^{\infty} c_n \cos(nx) \cdot \cos(3nt),$$

where the values of c_n are given above.

26.4 Pluck vs. struck initial conditions

A plucked string is one that is held in a starting position and then let go. It has no initial velocity.

A struck string is one that is initially at equilibrium and is struck by an impulse to set it into motion.

So the initial conditions for the two are:
Plucked string: $y(x,0) = f(x)$, $y_t(x,0) = 0$.

Struck string: $y(x,0) = 0$, $y_t(x,0) = g(x)$.

Example 26.2. (Struck string.) A struck string of length $L = \pi$ satisfies the following system

WE: $y_{tt} = 9y_{xx}$

BC: $y(0,t) = 0$, $y(\pi,t) = 0$ (These are different from the previous example.)

IC: $y(x,0) = 0$, $y_t(x,0) = g(x)$

Find the solution.

Solution: WE and BC are the same as Example 25.4. So the general solution satisfying both WE and BC is

$$y(x,t) = \sum_{n=1} y_n(x,t) = \sum_{n=1} \sin(nx) \cdot (c_n \cos(3nt) + d_n \sin(3nt)).$$

As usual, we use the IC to find the values of the coefficents:

First IC: $y(x,0) = 0 = \sum c_n \sin(nx) \;\Rightarrow\; c_n = 0$. So, $\boxed{y(x,t) = \sum d_n \sin(nx)\sin(3nt).}$

Second IC: $y_t(x,0) = g(x) = \sum 3nd_n \sin(nx)$. Therefore, $3nd_n$ are the Fourier sine coefficients of $g(x)$. So,

$$3nd_n = \frac{2}{\pi}\int_0^{\pi} g(x)\sin(nx)\,dx \quad \text{or} \quad \boxed{d_n = \frac{2}{3n\pi}\int_0^{\pi} g(x)\sin(nx)\,dx.}$$

The two boxed formulas give a complete solution to the example.

26.5 The d'Alembert solution to the wave equation

This secion is for enrichment. We will not cover it in ES.1803

For the undamped, unforced wave equation there is another standard method of solution called the d'Alembert solution. We'll state it and then show how it equals the solution found by the Fourier method.

Consider the system for a plucked string of length L:

WE: $y_{tt} = a^2 y_{xx}$

BC: $y(0,t) = y(L,t) = 0$

IC: $y(x,0) = f(x)$, $y_t(x,0) = 0$.

Claim. Let $\tilde{f}_o(x)$ be the period $2L$ odd extension of $f(x)$. Then

$$y(x,t) = \frac{\tilde{f}_o(x+at) + \tilde{f}_o(x-at)}{2}$$

is a solution to this system. We call this solution the d'Alembert solution.

Proof. This is trivial to check directly! You should do it, and make sure you see why the BC are satisfied.

Note. Physically, we can think of $\tilde{f}_o(x+at)$ as a wave traveling to the left at speed a and $\tilde{f}_o(x-at)$ as the same wave traveling to the right. Since the solution $y(x,t)$ models a standing wave, we see that a standing wave on $[0,L]$ is the superposition of two traveling waves!

26.5.1 The d'Alembert and Fourier solutions are the same

This has to be the case, but we will show it using a standard trig identity.

We know the system has Fourier solution:

$$y(x,t) = \sum b_n \sin\left(\frac{\pi}{L}nx\right)\cos\left(\frac{\pi}{L}ant\right), \text{ where } f(x) = \sum b_n \sin\left(\frac{\pi}{L}nx\right) \text{ on } 0 < x < L$$

Of course the sine series for $f(x)$ is just the Fourier series for $\tilde{f}_o(x)$, i.e., $\tilde{f}_o(x) = \sum b_n \sin\left(\frac{\pi}{L}nx\right)$ for all x.

We need the following trig identity (which we've used multiple times before).

$$\sin(\alpha)\cos(\beta) = \frac{1}{2}(\sin(\alpha+\beta) + \sin(\alpha-\beta)).$$

Now use this identity to rewrite the Fourier solution.

$$\begin{aligned} y(x,t) &= \sum b_n \sin\left(\frac{\pi}{L}nx\right)\cos\left(\frac{\pi}{L}ant\right) \\ &= \frac{1}{2}\sum b_n \sin\left(\frac{\pi}{L}n(x+at)\right) + \sin\left(\frac{\pi}{L}n(x-at)\right) \\ &= \frac{1}{2}(\tilde{f}_o(x+at) + \tilde{f}_o(x-at)). \qquad \text{QED} \end{aligned}$$

26.6 Musical notes

Assume the fundamental note is a C, then we get the following chart. The 'n' column is the harmonic, i.e., $n = 1$ is the first harmonic (or fundamental), $n = 2$ is the second harmonic, etc. The ‘ratio’ column is the ratio of the frequencies of the harmonic and the previous harmonic, i.e., $n/(n-1)$

n	ratio	note	interval
1		C	fundamental
2	2/1	C	octave
3	3/2	G	fifth
4	4/3	C	fourth
5	5/4	E	third
6	6/5	G	augmented second
7	7/6		ugly harmonic
8	8/7	C	ignore interval because previous one ugly
9	9/8	D	second
10	10/9	E	second

The point of this chart is to show that the frequencies in musical intervals have small whole number ratios. This is the standard way of tuning a musical instrument. It is called ‘just-temperament’.

Another method of tuning is called ’equal temperament. Here the 12 half-steps in an octave are all equal. That is, for each step the ratio of the frequencies is $2^{1/12} \approx 1.05946309$. After 12 steps the frequency has doubled which is an octave. We get the following table. For comparison we include a ’just-tempered’ scale.

n 1/2 steps	$2^{n/12}$	interval from base	just tuning interval	percent difference
0	1.0	unison	1	0.00%
1	1.059	minor second	16/15	−0.68%
2	1.122	major second	9/8	−0.23%
3	1.189	minor third	6/5	−0.91%
4	1.260	major third	5/4	+0.79%
5	1.335	perfect fourth	4/3	+0.11%
6	1.414	diminished fifth	7/5	+1.02%
7	1.498	perfect fifth	3/2	−0.11%
8	1.587	minor sixth	8/5	−0.79%
9	1.682	major sixth	5/3	+0.90%
10	1.782	minor seventh	16/9	+0.23%
11	1.888	major seventh	15/8	+0.68%
12	2.0	octave	2/1	0.00%

Note. It is impossible to tune a piano so that all major keys are just-tempered. A piano is called ’well-tempered’ when the major keys are close enough to just-tempered that they don’t sound out of tune.

27 Qualitative behavior of linear systems

27.1 Goals

1. Be able to draw the vector field associated to an autonomous system.

2. Be able to draw the phase portrait of any linear, autonomous, second-order system.
3. Be able to use eigenvalues to classify the types of critical points and their dynamic stability.
4. Be able to use the trace-determinant diagram to organize the different types of critical points.

27.2 Introduction

In this topic we are going to look at the qualitative behavior of systems of the form

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = A \begin{bmatrix} x \\ y \end{bmatrix} \quad \Leftrightarrow \quad \mathbf{x}' = A\mathbf{x}, \tag{39}$$

where $\mathbf{x} = \begin{bmatrix} x \\ y \end{bmatrix}$ and A is a constant, 2×2 matrix.

This is a system of two first-order DEs, so it is a second-order system. Since A is constant, the system is autonomous (the rate $\mathbf{x}$ changes depends only on $\mathbf{x}$) and time invariant.

Our goal is to sketch portraits of the solutions to these systems that capture their important qualitative features. Similar to what we did with first-order autonomous equations and phase lines, we will use critical points to organize our work.

While this gives us a useful perspective on linear systems, since we already know how to solve these systems, we don't really need it to understand such systems. Our real goal here is to prepare for a qualitative analysis of nonlinear systems. Since we can't usually solve nonlinear systems exactly, we will approximate them by linear systems and then leverage our qualitative understanding of linear systems to get information about nonlinear ones.

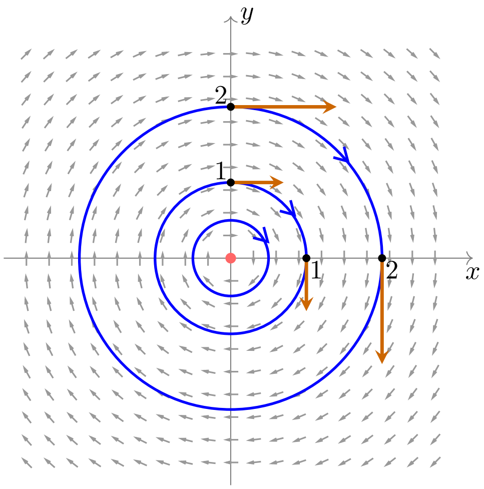
27.3 The phase plane: example with definitions

Example 27.1. Let $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$. Consider the autonomous system

$$\mathbf{x}' = A\mathbf{x} \quad \Leftrightarrow \quad \begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}. \tag{40}$$

We'll use this example to define and explain the terms we use in our qualitative description of a system.

Phase plane: The phase plane for our system is simply the xy-plane. This is where we will do all of our graphical work.



Trajectories, tangent vectors and direction field in phase plane for $\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$

Critical points: A critical point of the system is a point in the xy-plane where $\mathbf{x}' = \mathbf{0}$. For the system $\mathbf{x}' = A\mathbf{x}$, critical points satisfy the equation

$$A\mathbf{x} = \mathbf{0}.$$

Every such system has one critical point at $\mathbf{x} = \mathbf{0}$. In our example, $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$ is nonsingular. Therefore, $\mathbf{x} = \mathbf{0}$ is the only solution to $A\mathbf{x} = \mathbf{0}$, i.e., the system's only critical point is at the origin. (This is the case for most systems $\mathbf{x}' = A\mathbf{x}$.)

In the phase plane figure above, the critical point at the origin is marked with a solid pink dot.

Trajectories: Any solution $\begin{bmatrix} x(t) \\ y(t) \end{bmatrix}$ to the system can be plotted as a parametrized curve in the phase plane (xy-plane). Such a curve is called a trajectory of the system.

Using the method of eigenvalues and eigenvectors, we found the solution to Equation 40:

$$\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \begin{bmatrix} c_1 \cos(t) + c_2 \sin(t) \\ -c_1 \sin(t) + c_2 \cos(t) \end{bmatrix} \quad \text{or} \quad x(t) = c_1 \cos(t) + c_2 \sin(t),\ y(t) = -c_1 \sin(t) + c_2 \cos(t).$$

Several trajectories are plotted in the figure above. They are circles turning in the clockwise direction.

Important: The constant function $\mathbf{x}(t) = \mathbf{0}$ is a solution to the system. In the figure above, the trajectory of this solution is given by the dot at the origin. That is, the critical point $\mathbf{x} = \mathbf{0}$ is also a stationary trajectory.

Dynamic stability of the equilibrium at the origin: If all solutions go asymptotically to $\mathbf{0}$ as t gets large, we call the equilibrium at the origin a dynamically stable equilibrium. Clearly, the origin is dynamically stable exactly when all the eigenvalues have negative real parts.

If any eigenvalue has a positive real part, then most solutions go to infinity and we call the equilibrium at the origin dynamically unstable.

If the real part of one eigenvalue is 0 and those of all the others are ≤ 0, then we say the equilibrium is an edge case in terms of dynamic stability.

In the example in Equation 40, the eigenvalues are pure imaginary, so this is an edge case. In the figure above, we see the trajectories don't go asymptotically to the origin, but they also don't go to infinity. Whether we consider this stable or not depends on the application we have in mind.

Note: Dynamic stability refers to stability over time. We include the word 'dynamic' to distinguish this type of stability from the notion of structural stability, which we will talk about later.

Vector field and direction field: In general, the mapping $\begin{bmatrix} x \\ y \end{bmatrix} \longmapsto A \begin{bmatrix} x \\ y \end{bmatrix}$ gives us a vector field in the plane. That is, to each point (x, y) in the plane we attach a vector $A \begin{bmatrix} x \\ y \end{bmatrix}$. The figure above shows these vectors at the points $(1, 0)$, $(2, 0)$, $(0, 1)$, $(0, 2)$. Note: we know the vector field associated with Equation 40 without having to solve the equations.

For a curve $\begin{bmatrix} x(t) \\ y(t) \end{bmatrix}$, the derivative $\begin{bmatrix} x'(t) \\ y'(t) \end{bmatrix}$ is the tangent or velocity vector. Equation 40 shows that the tangent vectors to trajectories are the same as those in the vector field just described. Notice that the vectors in the figure above at $(1, 0)$, $(2, 0)$, $(0, 1)$, $(0, 2)$ are tangent to the trajectories through these points.

Finally, sometimes, rather than trying to show relative lengths of tangent vector fields, we can make all the vectors the same length. In this case, we call the plot a direction field. It tells you the direction of the trajectory through a point, but not its speed. The figure above shows the direction field (for our system) as a grid of small arrows. Note, at each point on the trajectories, the curve is tangent to the direction field.

27.4 Phase portraits

Definition: To draw the phase portrait of a system of a system, you need to draw enough trajectories to get a good sense of the system. Always include the equilibium solution.

For the remainder of this topic we will consider the general constant coefficient linear system in Equation 39.

This system always has a critical point (i.e., $\mathbf{x}' = 0$) at the origin. A critical point also represents a stationary trajectory, i.e., $\mathbf{x}(t) = \mathbf{0}$ is a solution to Equation 39. Our goal is to use the signs of the eigenvalues to classify the different types of critical points at the origin.

We will divide these types into 'main cases' and 'edge cases'. A main case is one where changing the eigenvalues a little will not change the case. For example, if we have one positive and one negative eigenvalue, then if the eigenvalues change a little, one will remain positive and the other negative.

An edge case is one where the smallest change could change the case. For example, if we have one positive and one zero eigenvalue, then the smallest change in the zero could change this to two positive eigenvalues or one positive and one negative eigenvalue.

Before reading through the cases, you should scan all the phase plane portraits shown below.

27.4.1 Drawing a phase portrait: examples

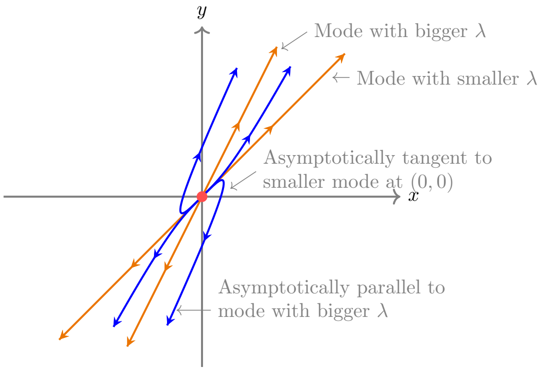
We will use some examples to walk through drawing the phase portrait for several systems. This should be enough to see how to draw phase portraits for all our main cases.

Example 27.2. (Nodal source) Suppose the solution to $\mathbf{x}' = A\mathbf{x}$ is

$$\mathbf{x}(t) = c_1 e^{2t}\begin{bmatrix}1\\1\end{bmatrix} + c_2 e^{3t}\begin{bmatrix}1\\2\end{bmatrix}.$$

Sketch a phase portrait.

Solution: Here is the final sketch. We outline the steps for drawing the phase portrait below.



y
x
Mode with bigger λ
← Mode with smaller λ
Asymptotically tangent to smaller mode at (0, 0)
Asymptotically parallel to mode with bigger λ

Nodal source at $(0, 0)$ – eigenvalues are positive and different.
All trajectories "flow out" from the origin.

Step 1: Sketch the equilibrium solution: $\mathbf{x}(t) = \begin{bmatrix}0\\0\end{bmatrix}$ = single point.

Step 2: Sketch the modes:

Modal solutions: $\mathbf{x_1}(t) = c_1 e^{2t}\begin{bmatrix}1\\1\end{bmatrix}$, $\mathbf{x_2}(t) = c_2 e^{3t}\begin{bmatrix}1\\2\end{bmatrix}$.

Mode $\mathbf{x}(t) = e^{2t}\begin{bmatrix}1\\1\end{bmatrix}$: trajectory = ray from the origin through (1,1).

Mode $\mathbf{x}(t) = -e^{2t}\begin{bmatrix}1\\1\end{bmatrix}$: trajectory = ray from the origin through (-1,-1).

Likewise, the trajectories of $\mathbf{x}(t) = e^{3t}\begin{bmatrix}1\\2\end{bmatrix}$, $\mathbf{x}(t) = -e^{3t}\begin{bmatrix}1\\2\end{bmatrix}$ are rays from the origin.

Summary: modes give straight line trajectories.

Step 3: Sketch some "mixed modal" solutions, e.g., sketch $\mathbf{x}(t) = e^{2t}\begin{bmatrix}1\\1\end{bmatrix} + e^{3t}\begin{bmatrix}1\\2\end{bmatrix}$.

Asymptotics as $t \to \infty$: Because the eigenvalues (exponents) are positive, as $t \to \infty$, $\mathbf{x}(t)$ goes to infinity. We claim the trajectory becomes asymptotically parallel to the mode with the bigger eigenvalue, i.e., asymptotically parallel to $\begin{bmatrix}1\\2\end{bmatrix}$. To see this, we look at the tangent vector to the trajectory:

$$\mathbf{x}'(t) = 2e^{2t}\begin{bmatrix}1\\1\end{bmatrix} + 3e^{3t}\begin{bmatrix}1\\2\end{bmatrix} = e^{3t}\left(2e^{-t}\begin{bmatrix}1\\1\end{bmatrix} + 3\begin{bmatrix}1\\2\end{bmatrix}\right).$$

This shows that $\mathbf{x}'(t)$ is parallel to $2e^{-t}\begin{bmatrix}1\\1\end{bmatrix} + 3\begin{bmatrix}1\\2\end{bmatrix}$. As t gets large, the first term vanishes and the curve becomes asymptotically parallel to $\begin{bmatrix}1\\2\end{bmatrix}$, as claimed.

Asymptotics as $t \to -\infty$: As $t \to -\infty$, $\mathbf{x}(t)$ goes to zero. We claim the trajectory becomes asympotically tangent to the mode with the smaller eigenvalue, i.e., asymptotically tangent to the line along $\begin{bmatrix}1\\1\end{bmatrix}$. To see this, we look at the tangent vector to the trajectory:

$$\mathbf{x}'(t) = 2e^{2t}\begin{bmatrix}1\\1\end{bmatrix} + 3e^{3t}\begin{bmatrix}1\\2\end{bmatrix} = e^{2t}\left(2\begin{bmatrix}1\\1\end{bmatrix} + 3e^{t}\begin{bmatrix}1\\2\end{bmatrix}\right).$$

This shows that $\mathbf{x}'(t)$ is parallel to $2\begin{bmatrix}1\\1\end{bmatrix}+3e^{t}\begin{bmatrix}1\\2\end{bmatrix}$. So, as t gets large and negative, the second term vanishes and the tangent vector asymptotically points parallel to $\begin{bmatrix}1\\1\end{bmatrix}$, as claimed.

Drawing other mixed modal trajectories is similar.

We call the equilibrium at the origin a nodal source. If you think of the trajectories as representing flowing water, the origin appears as a source, pushing out the water. The equilibrium is dynamically unstable.

Key points

- Trajectories don't cross.
- They fill up the plane.
- Different solutions with the same trajectory have different initial values, e.g.,
  $\mathbf{x_1} = e^{2t}\begin{bmatrix}1\\1\end{bmatrix}$ and $\mathbf{x_2}(t) = 3e^{2t}\begin{bmatrix}1\\1\end{bmatrix}$ have the same trajectory, but
  $\mathbf{x_1}(0) = \begin{bmatrix}1\\1\end{bmatrix}$ and $\mathbf{x_2}(0) = \begin{bmatrix}3\\3\end{bmatrix}$ are different initial values.
- For nodal sources:
  – Trajectories become parallel to the mode with the bigger λ as t goes to ∞.
  – Trajectories become tangent to the mode with the smaller λ as t goes to $-\infty$.
  – As $t \to -\infty$, trajectories go asymptotically to $(0, 0)$.
  – Systems with positive, different eigenvalues have the same qualitative picture, i.e., they all look like nodal sources.

Example 27.3. (Spiral source) Let $\mathbf{x}' = \begin{bmatrix}3 & 5\\-5 & 3\end{bmatrix}\mathbf{x}$. Draw the phase portrait.

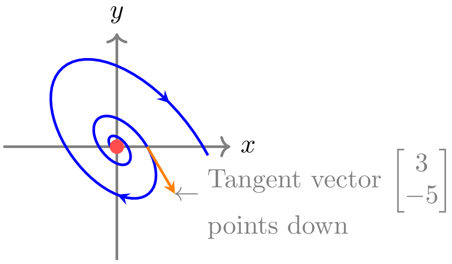
We find the eigenvalues are $\lambda = 3 \pm 5i$. After some algebra we find:

$$\mathbf{x}(t) = c_1 \underset{\uparrow}{e^{3t}} \underset{\uparrow}{\begin{bmatrix} \cos(5t) \\ -\sin(5t) \end{bmatrix}} + c_2 e^{3t} \begin{bmatrix} \sin(5t) \\ \cos(5t) \end{bmatrix}$$

grows × circle = spiral out

To determine the sense of turning, i.e., if it turns clockwise (CW) or counterclockwise (CCW), we look at the tangent vector at the point (1,0) in the plane.

$$\begin{aligned} \text{At } (1,0): \mathbf{x}'(0) &= A \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 3 & 5 \\ -5 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 3 \\ -5 \end{bmatrix} \\ &= \text{ tangent vector to the trajectory through } (1,0) \end{aligned}$$



Phase portrait: spiral source

The tangent vector points down, so the spiral must be turning clockwise.

The critical point at $(0,0)$ is called a spiral source. It is a dynamically unstable equilibrium.

Example 27.4. (Saddle) Suppose the matrix A has the following eigenvalues and eigenvectors.

$$\begin{array}{ccrr} \lambda & = & -3 & 2 \\ \mathbf{v} & = & \begin{bmatrix} 3 \\ 1 \end{bmatrix} & \begin{bmatrix} -1 \\ 1 \end{bmatrix} \end{array}$$

Sketch a phase portrait of the system $\mathbf{x}' = A\mathbf{x}$. Name the type of critical point at the origin and give its stability.

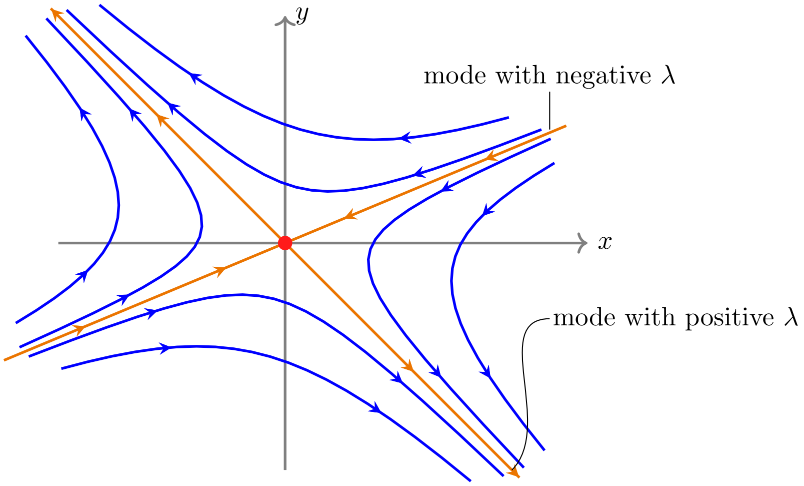
Solution: The general solution is $\mathbf{x}(t) = c_1 e^{-3t} \begin{bmatrix} 3 \\ 1 \end{bmatrix} + c_2 e^{2t} \begin{bmatrix} -1 \\ 1 \end{bmatrix}$.

Modes have straight line trajectories:

$$\mathbf{x_1} = e^{-3t} \begin{bmatrix} 3 \\ 1 \end{bmatrix} \qquad \text{goes to 0 as } t \text{ increases.}$$

$$\mathbf{x_2} = e^{2t} \begin{bmatrix} -1 \\ 1 \end{bmatrix} \qquad \text{goes away from 0 as } t \text{ increases.}$$

Mixed modal solutions: For example, $e^{-3t} \begin{bmatrix} 3 \\ 1 \end{bmatrix} + e^{2t} \begin{bmatrix} -1 \\ 1 \end{bmatrix}$, goes asympotically to $e^{2t} \begin{bmatrix} -1 \\ 1 \end{bmatrix}$ as $t \to \infty$ and goes asympotically to $e^{-3t} \begin{bmatrix} 3 \\ 1 \end{bmatrix}$ as $t \to -\infty$.



Saddle (dynamically unstable equilibrium at (0,0))

27.4.2 Key points about phase portraits

- Trajectories don't cross.
- They fill up the plane.
- Different solutions can have the same trajectory. They just have different initial values.
- Qualitatively, the phase portrait is determined by the eigenvalues.

27.5 Types of critical points: main cases based on eigenvalues

Here we will summarize the main cases for the possible types of critical points (equilibria) at the origin. We'll start with some notational conventions for this section.

If the eigenvalues are real, we label them λ_1 and λ_2. We label the corresponding eigenvectors $\mathbf{v_1}$ and $\mathbf{v_2}$. In this case, the general solution to Equation 39 is

$$\mathbf{x}(t) = c_1 e^{\lambda_1 t}\mathbf{v_1} + c_2 e^{\lambda_2 t}\mathbf{v_2}. \tag{41}$$

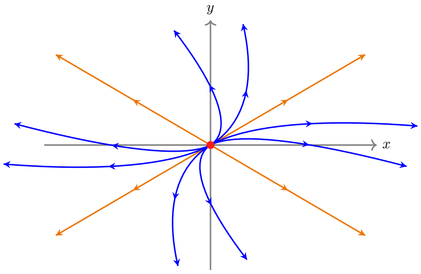
If the eigenvalues are complex (with nonzero imaginary part), we label one of them $\lambda = \alpha + \beta\, i$ and the corresponding eigenvector $\mathbf{v} + i\,\mathbf{w}$. In this case, the general solution to Equation 39 is

$$\mathbf{x}(t) = c_1 e^{\alpha t}(\cos(\beta t)\,\mathbf{v} - \sin(\beta t)\,\mathbf{w}) + c_2 e^{\alpha t}(\sin(\beta t)\,\mathbf{v} + \cos(\beta t)\,\mathbf{w}). \tag{42}$$

Case (i) Real eigenvalues, distinct, both postitive: $\lambda_1 > \lambda_2 > 0$.

Type of critical point at origin: **Nodal source**.

Dynamic stability of the equilibrium: dynamically unstable.



Critical point at the origin is a nodal source

As $t \to \infty$, $\mathbf{x}(t)$ goes to ∞ and the trajectory becomes asymptotically parallel to $\mathbf{v_1}$, i.e., to the eigenvector for the bigger eigenvalue.

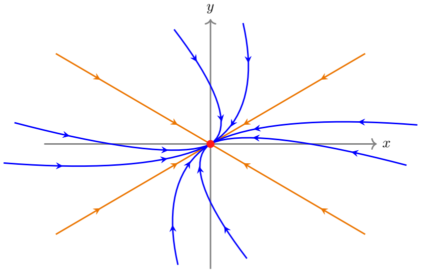
As $t \to -\infty$, $\mathbf{x}(t)$ goes asymptotically to 0 and becomes asymptotically tangent to (the line along) $\mathbf{v_2}$, i.e., to the eigenvector for the smaller eigenvalue.

Case (ii) Real eigenvalues, distinct, both negative, $\lambda_1 < \lambda_2 < 0$.

Type of critical point at origin: **Nodal sink**.

Dynamic stability of the equilibrium: dynamically (asymptotically) stable.

(Simply reverse the arrows on Case (i).)



Critical point at the origin is a nodal sink

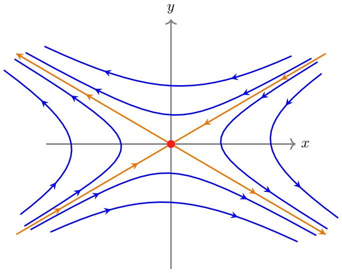
As $t \to \infty$, $\mathbf{x}(t)$ goes asymptotically to 0 and the trajectory becomes asymptotically tangent to (the line along) $\mathbf{v_2}$, i.e., to the eigenvector for the less negative eigenvalue (smaller absolute value).

As $t \to -\infty$, $\mathbf{x}(t)$ goes to ∞ and becomes asymptotically parallel to $\mathbf{v_1}$, i.e., to the eigenvector for the more negative eigenvalue (bigger absolute value).

Case (iii) Real eigenvalues, one positive, one negative, $\lambda_1 > 0 > \lambda_2$.

Type of critical point at origin: **Saddle**

Dynamic stability of the equilibrium: dynamically unstable.



Critical point at the origin is a saddle

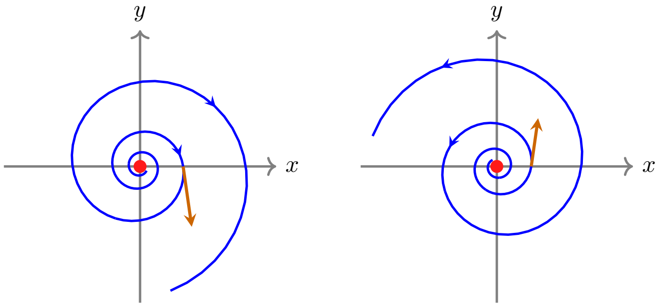
As $t \to \infty$, $\mathbf{x}(t)$ goes to ∞ and becomes asympotically tangent to the mode $c_1 e^{\lambda_1 t}\mathbf{v_1}$, i.e., to the mode with positive eigenvalue.

As $t \to -\infty$, $\mathbf{x}(t)$ goes to ∞ and becomes asympotically tangent to the mode $c_2 e^{\lambda_2 t}\mathbf{v_2}$, i.e., to the mode with negative eigenvalue.

Case (iv) Complex eigenvalues, positive real part, i.e., $\alpha > 0$.

Type of critical point at origin: **Spiral source**

Dynamic stability: dynamically unstable.



Critical point at the origin is a spiral source. Left: clockwise; right: counterclockwise

Trajectories can spiral clockwise or counterclockwise. You can find the direction of rotation by checking the tangent vector at one point.

As $t \to \infty$, $\mathbf{x}(t)$ goes to ∞.

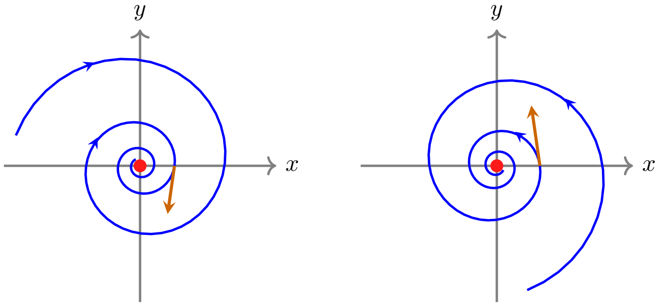
As $t \to -\infty$, $\mathbf{x}(t)$ goes to 0.

Case (v) Complex eigenvalues, negative real part, i.e., $\alpha < 0$.

Type of critical point at origin: **Spiral sink**

Dynamic stability: dynamically stable.

(Reverse arrows from Case (iv).)



Critical point at the origin is a spiral sink. Left: clockwise; right: counterclockwise

Trajectories can spiral clockwise or counterclockwise. You can find the direction of rotation by checking the tangent vector at one point.

As $t \to \infty$, $\mathbf{x}(t)$ goes to 0.

As $t \to -\infty$, $\mathbf{x}(t)$ goes to ∞.

27.6 Types of critical points: edge cases based on eigenvalues

For the edge cases we will just list the properties and show a phase portrait. These are drawn in the same way as the main case examples. In class, we'll look at as many of these as we have time for.

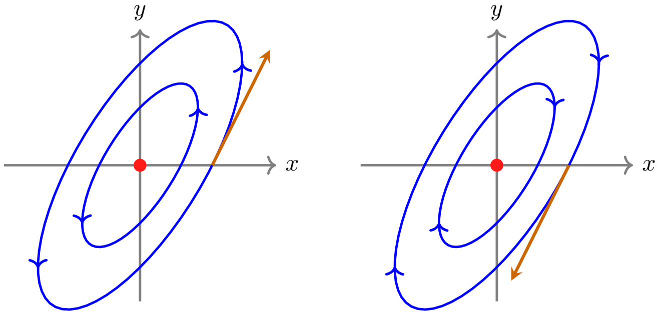
Case (vi) Pure imaginary eigenvalues: $\lambda = i\beta$.

Type of critical point at origin: **Center**

Dynamic stability: This is an edge case, in some applications this can be considered stable, in others it might not.

Trajectories can turn clockwise or counterclockwise. As usual, you can find the direction of rotation by checking the tangent vector at one point.

As $t \to \pm\infty$, $\mathbf{x}(t)$ goes round and round an ellipse.

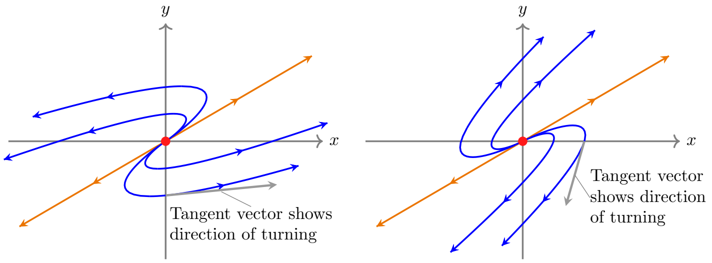


Critical point at the origin is a center

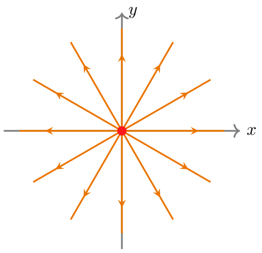
Case (vii) Real, repeated, positive eigenvalues: $\lambda_1 = \lambda_2 > 0$.

Type of critical point at origin: **Defective nodal source or star nodal source.**

Dynamic stability: dynamically unstable .



Defective nodal source



Star nodal source

If the coefficient matrix is **defective** (repeated eigenvalue, only one independent eigenvector), then we have a defective nodal source at the origin.

Let λ be the eigenvalue and $\mathbf{v_1}$ the corresponding eigenvector. Let $\mathbf{v_2}$ be a generalized eigenvector associated with $\mathbf{v_1}$.

In this case, the general solution to Equation 39 is $\mathbf{x}(t) = e^{\lambda t}(c_1\mathbf{v_1} + c_2(t\mathbf{v_1} + \mathbf{v_2}))$. The critical point at the origin is called a defective nodal source.

As $t \to \infty$, $\mathbf{x}(t)$ goes to ∞ and the trajectory becomes asymptotically parallel to the (only) mode, i.e., parallel to $\mathbf{v_1}$.

As $t \to -\infty$, $\mathbf{x}(t)$ goes to $\mathbf{0}$ and the trajectory becomes asymptotically tangent to the line along $\mathbf{v_1}$.

Trajectories asymptotically make a 180 degree turn. As with spirals, you can find the sense of the turn by checking one tangent vector.

If the coefficient matrix is **complete**, there are two independent eigenvectors, which implies A is a scalar matrix: $A = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}$.

This implies the general solution is $\mathbf{x}(t) = e^{\lambda t}\vec{\mathbf{c}}$.

That is, all trajectories are straight rays. This is called a star nodal source.

As $t \to \infty$, $\mathbf{x}(t) \to \infty$ along a line from 0.

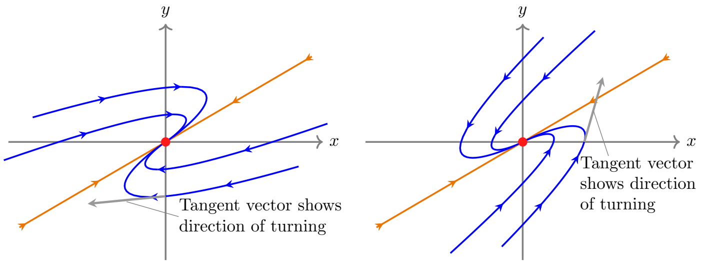
As $t \to -\infty$, $\mathbf{x}(t) \to 0$

Case (viii) Real, repeated, negative eigenvalues: $\lambda_1 = \lambda_2 < 0$.

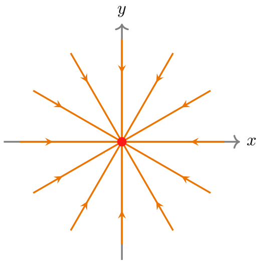
Type of critical point at origin: **Defective nodal sink or star nodal sink.**

Dynamic stability: dynamically stable.

Just reverse the arrows from Case (vii).


y
x
Tangent vector shows direction of turning
y
x
Tangent vector shows direction of turning

Defective nodal sink



Star nodal sink

If the coefficient matrix is defective:

As $t \to -\infty$, $\mathbf{x}(t)$ goes to ∞ and the trajectory becomes asymptotically parallel to the (only) mode, i.e., parallel to $\mathbf{v_1}$.

As $t \to \infty$, $\mathbf{x}(t)$ goes to $\mathbf{0}$ and the trajectory becomes asymptotically tangent to the line along $\mathbf{v_1}$.

Trajectories asymptotically make a 180 degree turn. As with the defective nodal source, you can find the sense of the turn by checking one tangent vector.

If the coefficient matrix is complete, there are two independent eigenvectors, which implies A is a scalar matrix: $A = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}$.

This implies the general solution is $\mathbf{x}(t) = e^{\lambda t}\vec{\mathbf{c}}$.

(Simply reverse the arrows on the star nodal source.)

Case (ix) Real eigenvalues, one negative, one zero: $\lambda_1 = 0 > \lambda_2$.

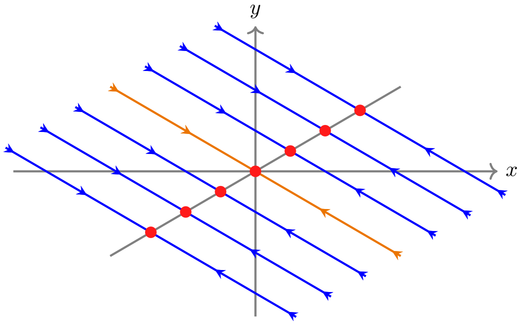
Type of critical point at the origin: **Degenerate** (line of critical points)

Dynamic stability: edge case

The critical points are not isolated –they lie on the line through 0 with direction $\mathbf{v_1}$.

$$\mathbf{x}(t) = c_1\mathbf{v_1} + c_2 e^{\lambda_2 t}\mathbf{v_2}.$$

As $t \to \infty$, $\mathbf{x}(t) \to c_1\mathbf{v_1}$ along a line parallel to $\mathbf{v_2}$.



Degenerate case: line of critical points

Case (x) Real eigenvalues, one positive, one zero: $\lambda_1 = 0 < \lambda_2$.

Type of critical point at origin: **Degenerate** (line of critical points)

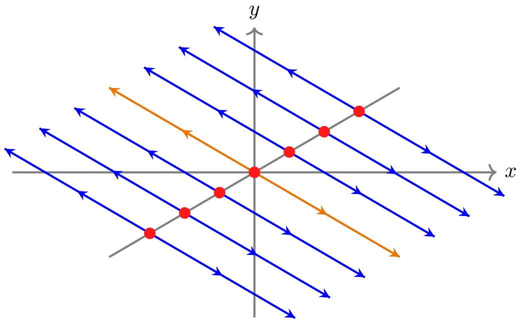
Dynamic stability: dynamically unstable .

(Simply reverse the arrows in Case (ix).)

The critical points are not isolated –they lie on the line through 0 with direction $\mathbf{v_1}$.

$$\mathbf{x}(t) = c_1\mathbf{v_1} + c_2 e^{\lambda_2 t}\mathbf{v_2}.$$

As $t \to \infty$, $\mathbf{x}(t) \to \infty$ along a line parallel to $\mathbf{v_2}$.



Degenerate case: line of critical points

Case (xi) Real eigenvalues, both 0: $\lambda_1 = \lambda_2 = 0$.

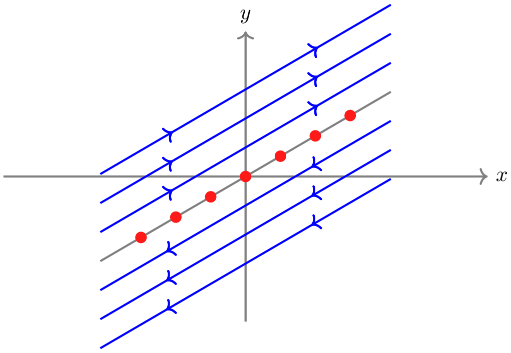
Since the eigenvalues are repeated, this breaks into two cases:

Complete case: Every point is a critical point, every trajectory is a point.

Defective case: Line of critical points.

$$\mathbf{x}(t) = c_1\mathbf{v_1} + c_2(t\mathbf{v_1} + \mathbf{v_2}).$$

Trajectories are parallel to $\mathbf{v_1}$.



Degenerate and defective: (both $\lambda = 0$)

27.7 Example

Example 27.5. The matrix $A = \begin{bmatrix} 2 & 3 \\ -3 & 2 \end{bmatrix}$ has eigenvalues $2 \pm 3i$. So, for the system $\mathbf{x}' = A\mathbf{x}$, the critical point at the origin is a spiral source.

The tangent vector at the point $\mathbf{x}_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ is $A\,\mathbf{x}_0 = \begin{bmatrix} 2 \\ -3 \end{bmatrix}$. This tells us the curve spirals clockwise.

27.8 Trace-determinant plane

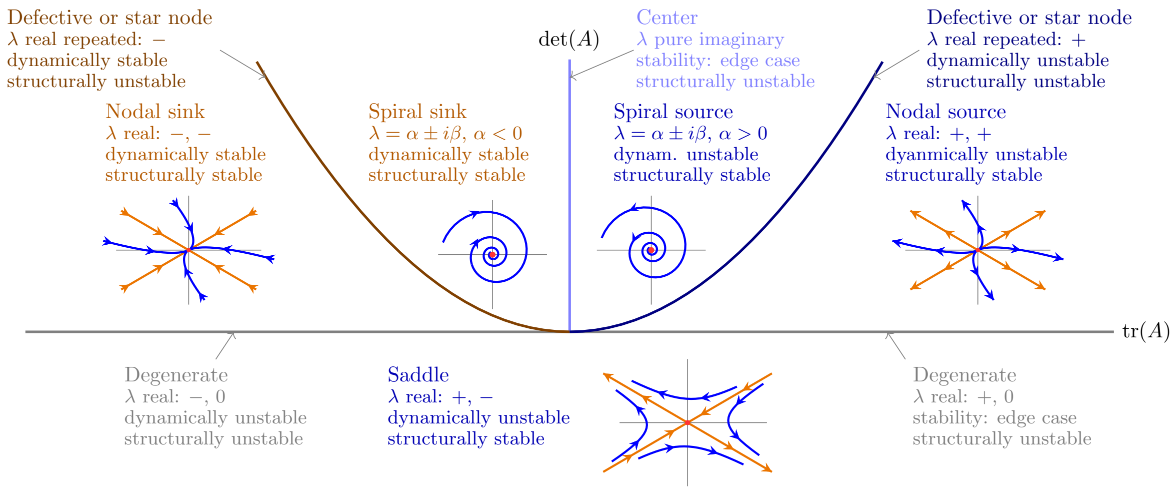
For $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, the characteristic equation is

$$\det(A - \lambda I) = \begin{vmatrix} a-\lambda & b \\ c & d-\lambda \end{vmatrix} = \lambda^2 - (a+d)\lambda + (ad-bc) = 0.$$

We recognize $ad - bc = \det(A)$. The term $(a+d)$ is called the trace of A, denoted $\operatorname{tr}(A)$. (Trace is the sum of the entries along the main diagonal.) With this notation, the characteristic equation is

$$\lambda^2 - \operatorname{tr}(A)\,\lambda + \det(A) = 0 \longrightarrow \lambda = \frac{\operatorname{tr}(A) \pm \sqrt{\operatorname{tr}(A)^2 - 4\det(A)}}{2}.$$

Since the eigenvalues are determined by trace and determinant we have the following nice picture in the trace-determinant plane. (Structural stability will be discussed in Topics 28 and 29. To read the diagram, it is enough to know that the main cases are structurally stable and the edge cases are not.)



See the mathlet
https://mathlets.org/mathlets/linear-phase-portraits-matrix-entry/.

28 Qualitative behavior of nonlinear systems

28.1 Goals

1. Be able to find the critical points for a nonlinear, autonomous system.
2. Be able to linearize a nonlinear system near the critical points.
3. Be able to draw the phase portrait of a nonlinear, autonomous system using linearization near the critical points.
4. Understand why the linearizations in this topic's examples are structurally stable.

28.2 Nonlinear Systems

A general first-order, autonomous, 2 × 2 system has the following form

$$\begin{aligned} x' &= f(x,y) \\ y' &= g(x,y) \end{aligned} \tag{43}$$

Vector Field: This defines a vector field $(f(x,y), g(x,y))$ that attaches the velocity vector to each point (x,y) in the *phase plane.*

By definition a critical point is one where $x' = 0$ and $y' = 0$. That is, it is a point (x_0, y_0) where

$$f(x_0, y_0) = 0, \text{ and } g(x_0, y_0) = 0.$$

Equivalently, it is an *equilibrium solution* $x(t) = x_0$, $y(t) = y_0$. This is a solution whose trajectory is a single point.

28.3 Approximation and structural stability

We'll talk more about structural stability in Topic 29. The key point is this: if you approximate or measure a number there will be some error. If your approximation says the number is 7, and the error is known to be small, then you can be certain the number's true value is positive. By contrast, if your approximation says the number is 0, then the true value might be positive, negative or zero.

We say a linear system is structurally stable if none of its eigenvalues are 0 or have real part equal to 0. The idea is that, if there is a small change to the system or a small error in our description, then the type of critical point at the origin won't change.

For example, if we experimentally determine a system has eigenvalues 7.0 and 1.0, then our experiment points to the origin being a nodal source. Even if there is a small error in our measurement, we'll still know the eigenvalues are positive and we have a nodal source. We say nodal sources are structurally stable.

In contrast, if we experiemntally find the eigenvalues are $0.0 \pm 2.0\,i$, then our experiment points to the origin being a center. But even the smallest error could mean the eigenvalues have positive or negative real part. That is, all we can say from our experiment is that the origin is a center, spiral source or spiral sink. We say centers are stucturally unstable.

We can state this simply in two ways:

1. The main cases from Topic 27 are structurally stable. The edge cases are not.

2. In the trace-determinant diagram, the large open regions represent structurally stable systems and the dividing lines represent structurally unstable ones.

In this topic we will learn to approximate a nonlinear system near a critical point by a linear one. Because there is approximation error, we can only be sure that the nonlinear system matches the linear one if the linear system is structurally stable. For example, if the linear system is a nodal source, then we can be sure that the nonlinear system looks like a nodal source near the critical point. But, if the linear system is a center, then the nonlinear one could look like a center, spiral source or spiral sink.

All the examples in this topic's notes will involve structurally stable approximations, so we will be confident that we are correctly characterizing the nonlinear system. In Topic 29, we will explore structurally unstable linear approximations.

28.4 Linearization around a critical point

We'll start by presenting the method of linearization to sketch the phase portrait. First, we'll use it in an example. After that, we'll justify the method.

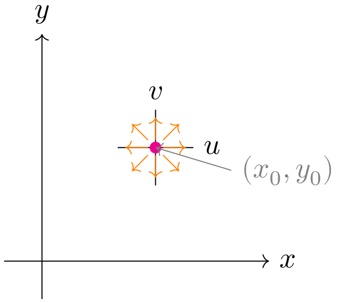
Jacobian. At a critical point (x_0, y_0) of the system in Equation 43, we define the Jacobian by

$$J(x_0, y_0) = \begin{bmatrix} f_x(x_0,y_0) & f_y(x_0,y_0) \\ g_x(x_0,y_0) & g_y(x_0,y_0) \end{bmatrix} = \begin{bmatrix} f_x & f_y \\ g_x & g_y \end{bmatrix}.$$

This gives the linearization around the critical point (x_0, y_0)

$$\begin{bmatrix} u \\ v \end{bmatrix}' = J(x_0, y_0) \begin{bmatrix} u \\ v \end{bmatrix}$$

In general, the nonlinear system behaves like the linearized one. (More precisely, if the linearized system is structurally stable, the nonlinear system behaves like the linear one.) That is, if we center our uv-axes on (x_0, y_0) then the linear vector field near the uv origin approximates the nonlinear field near (x_0, y_0)



Near a critical point, the nonlinear system is approximated by its linearization.

Example 28.1. Find the critical points for the following system.

$$x' = 14x - \frac{1}{2}x^2 - xy$$
$$y' = 16y - \frac{1}{2}y^2 - xy$$

Solution: We solve the equations $x' = 0$, $y' = 0$.

$$x' = x\left(14 - \frac{1}{2}x - y\right) = 0 \Rightarrow x = 0 \text{ or } 14 - \frac{1}{2}x - y = 0$$
$$y' = y\left(16 - \frac{1}{2}y - x\right) = 0 \Rightarrow y = 0 \text{ or } 16 - \frac{1}{2}y - x = 0.$$

Looking at the product for x' we see $x' = 0$ when $x = 0$ or $14 - x/2 - y = 0$. Likewise, $y' = 0$ when $y = 0$ or $16 - y/2 - x = 0$. This leads to four sets of equations for critical points.

$$\begin{Bmatrix} x = 0 \\ y = 0 \end{Bmatrix} \quad \begin{Bmatrix} x = 0 \\ 16 - y/2 - x = 0 \end{Bmatrix} \quad \begin{Bmatrix} 14 - x/2 - y = 0 \\ y = 0 \end{Bmatrix} \quad \begin{Bmatrix} 14 - x/2 - y = 0 \\ 16 - y/2 - x = 0 \end{Bmatrix}$$

The first three sets are easy to solve by inspection. The fourth requires a small computation. We get the following four critical points:

$$(0, 0),\ (0, 32),\ (28, 0),\ (12, 8).$$

Example 28.2. (Continued from previous example.) Linearize the system at each of the critical points and determined the type of the linearized critical point.

Solution: The linearized system at (x_0, y_0) is $\begin{bmatrix} u' \\ v' \end{bmatrix} = J(x_0, y_0) \begin{bmatrix} u \\ v \end{bmatrix}$.

First we compute the Jacobian:

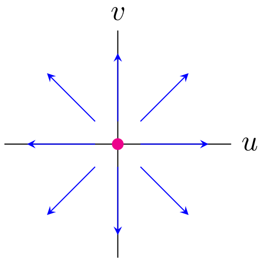
$$J(x, y) = \begin{bmatrix} 14 - x - y & -x \\ -y & 16 - y - x \end{bmatrix}$$

Next we look at each of the critical points in turn.

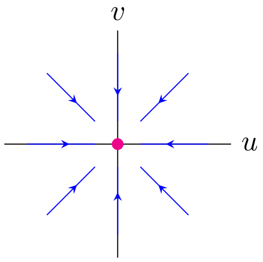
Critical point $(0,0)$:

$$J(0,0) = \begin{bmatrix} 14 & 0 \\ 0 & 16 \end{bmatrix}; \text{ eigenvalues } 14,\ 16.$$

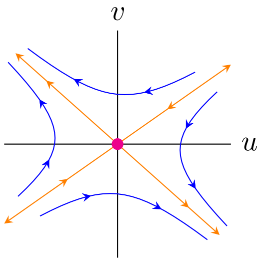
This is a nodal source. Since it is only an approximation of the nonlinear system near the critical point, it is not necessary to find the eigenvectors and make a precise sketch. Instead we draw general nodal source, i.e., a node with all trajectories pointing outward. Its sketch on uv-axes is shown in the left-most figure below.



Source node



Sink node



Saddle

Critical point $(0,32)$:

$$J(0,32) = \begin{bmatrix} -18 & 0 \\ -32 & -16 \end{bmatrix}; \text{ eigenvalues } -18, -16.$$

This is a sink node. As with the source node, we don't need the eigenvectors to make an approximate sketch of the nonlinear system. We simply sketch a node with all trajectories pointing in towards the critical point. Its sketch is shown in the 'Sink node' figure above.

Critical point $(28,0)$:

$$J(28,0) = \begin{bmatrix} -14 & -28 \\ 0 & -12 \end{bmatrix}, \text{ eigenvalues } -14, -12; \text{ corresponding eigenvectors } \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -14 \\ 1 \end{bmatrix}$$

This is a sink node. As with the source node, we don't need the eigenvectors to make an approximate sketch of the nonlinear system. Its sketch is shown in the 'Sink node' figure above.

Critical point $(12,8)$:

$$J(12,8) = \begin{bmatrix} -6 & -12 \\ -8 & -4 \end{bmatrix}; \text{ eigenvalues } -5 \pm \sqrt{97} \approx -15, 5.$$

$$\text{Eigenvectors: For } \lambda = -5 - \sqrt{97} : \begin{bmatrix} 1 + \sqrt{97} \\ 8 \end{bmatrix} \approx \begin{bmatrix} 11 \\ 8 \end{bmatrix}$$

$$\text{For } \lambda = -5 + \sqrt{97} : \begin{bmatrix} 1 - \sqrt{97} \\ 8 \end{bmatrix} \approx \begin{bmatrix} -9 \\ 8 \end{bmatrix}$$

This is a saddle. For saddles, we feel it is a good idea to find the eigenvectors so that the orientation of the saddle is correct. (Here, we just gave you the eigenvectors. At this point you should be able to find them quickly yourself.) The sketch of the linearized system is shown in the 'Saddle' figure above.

Example 28.3. (Continued from the previous example.) Are all the linearizations structurally stable? What does this imply about the nonlinar system?

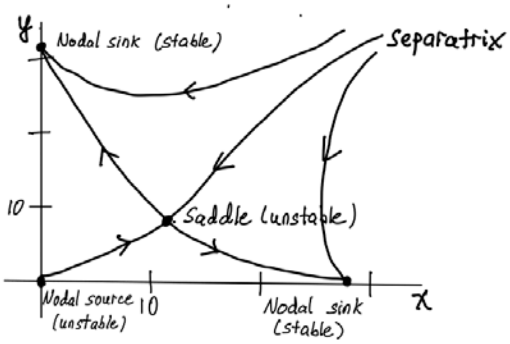
Solution: Yes. We can see this two ways. First, each of the linearized critical points are one of our main cases. These are structurally stable. Second, all of the eigenvalues for the linearizations are nonzero. Even with a small approximation error, this would still be the case. So the approximation error can't change the types of the critical points, i.e., they are structurally stable.

Since all the linearized critical points are structurally stable, the nonlinear critical points are all of the same type as their linearizations.

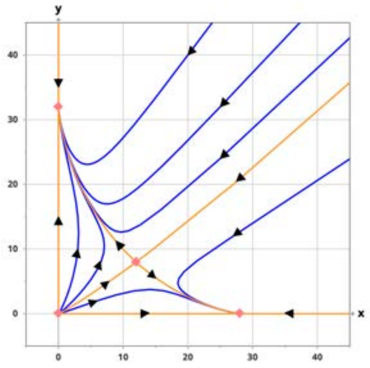
Example 28.4. (Continued from the previous example.) Make a rough sketch of the nonlinear system's phase portrait using the following two steps.

1. Sketch the phase portrait near each critical point, using the linearization.
2. Connect these sketches together in a consistent manner.

We do this below and compare it with a computer generated sketch.



Hand sketch of the phase plane.



Computer generated phase portrait.

28.4.1 Justification for using linearization

We'll go through this in detail. One key fact is that the change of variables $u = x - x_0$, $v = y - y_0$ puts the uv origin at (x_0, y_0).

We will use the linear (tangent plane) approximations of f and g. You might recall this from 18.02. (If not, notice that it is just a multivariable version of the single variable linear approximation $f(x) \approx f(x_0) + f'(x_0)\Delta x$, where $\Delta x = x - x_0$.)

For small changes $(x - x_0) = \Delta x$ and $(y - y_0) = \Delta y$, the linear approximations for f and g near (x_0, y_0) are

$$\begin{aligned} f(x,y) &\approx f(x_0,y_0) + f_x(x_0,y_0)\,\Delta x + f_y(x_0,y_0)\,\Delta y \\ g(x,y) &\approx g(x_0,y_0) + g_x(x_0,y_0)\,\Delta x + g_y(x_0,y_0)\,\Delta y \end{aligned}$$

Now, let $u = x - x_0 = \Delta x$ and $v = y - y_0 = \Delta y$.

1. This puts the origin of the uv-plane at (x_0, y_0).

3. As functions of t: $u' = x'$, $v' = y'$ (since x_0 and y_0 are constants).

Replacing $x - x_0$ and $y - y_0$ by u and v in the approximations, we get

$$\begin{aligned} f(x,y) &\approx f(x_0,y_0) + f_x(x_0,y_0)\,u + f_y(x_0,y_0)\,v \\ g(x,y) &\approx g(x_0,y_0) + g_x(x_0,y_0)\,u + g_y(x_0,y_0)\,v \end{aligned}$$

Writing these in matrix form we see the Jacobian appear:

$$\begin{aligned} \begin{bmatrix} f(x,y) \\ g(x,y) \end{bmatrix} &\approx \begin{bmatrix} f(x_0,y_0) \\ g(x_0,y_0) \end{bmatrix} + \begin{bmatrix} f_x(x_0,y_0) & f_y(x_0,y_0) \\ g_x(x_0,y_0) & g_y(x_0,y_0) \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} \\ &= \begin{bmatrix} f(x_0,y_0) \\ g(x_0,y_0) \end{bmatrix} + J(x_0,y_0) \begin{bmatrix} u \\ v \end{bmatrix} \end{aligned}$$

If (x_0, y_0) is a critical point, the first term on the right is 0, i.e

$$\begin{bmatrix} f(x,y) \\ g(x,y) \end{bmatrix} \approx J(x_0,y_0) \begin{bmatrix} u \\ v \end{bmatrix}.$$

Putting everything together:

$$\begin{bmatrix} u' \\ v' \end{bmatrix} = \begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} f(x,y) \\ g(x,y) \end{bmatrix} \approx J(x_0,y_0) \begin{bmatrix} u \\ v \end{bmatrix}$$

Using just the first and last terms from the above gives the linearization formula

$$\begin{bmatrix} u' \\ v' \end{bmatrix} \approx J(x_0,y_0) \begin{bmatrix} u \\ v \end{bmatrix}.$$

This is a linearized system with coefficient matrix $J(x_0, y_0)$. We call it the linearization of the system around the critical point.

29 Structural Stability

29.1 Goals

1. Be able to classify a linearized system near a critical point as structurally stable or unstable.

2. For a structurally unstable linearized system, be able to list the possible types of critical point for the nonlinear system.

29.2 Structural stability

Structural stability of the system $\mathbf{x} = A\mathbf{x}$ is about the type of system *not the type of critical point of the system.* Consider the following two scenarios.

Scenario 1. You have an apparatus modeled by a constant coefficient linear system $\mathbf{x}' = A\mathbf{x}$. You are experimentally able to measure the entries of the matrix A to two decimal places of accuracy. You are not suprised when your experiments reveal $A = \begin{bmatrix} 6.00 & 5.00 \\ 1.00 & 2.00 \end{bmatrix}$. So the eigenvalues of your system are 7.00 and 1.00.

You have experimentally determined that the equilibrium at the origin is a nodal source, which is dynamically unstable, i.e., over time trajectories that start near the source move away from it. But we have to take into account the possibility (really, guarantee) of measurement error. Each of your matrix entries might be off by as much as 0.005. Thus the eigenvalues are also only approximately correct.

Nonetheless, with such small errors, the eigenvalues are both guaranteed to be positive and the equilibrium is guaranteed to be a nodal source. We say the system is structurally stable. That is, a small change (also called a perturbation) of the system won't change the type of the equilibrium.

To repeat: the linear system with a nodal source is structurally stable, but has a dynamically unstable equilibrium at the origin.

Scenario 2. You have a known nonlinear system with a critical point at (x_0, y_0). You linearize the system and find that the linearized system has a nodal source with eigenvalues 1 and 7. In this case, the linearized system is an approximation of the nonlinear one. Since, close to the critical point, the approximation error is small, the structural stability of the linearized system tells us that the nonlinear system behaves like a nodal source close to the critical point.

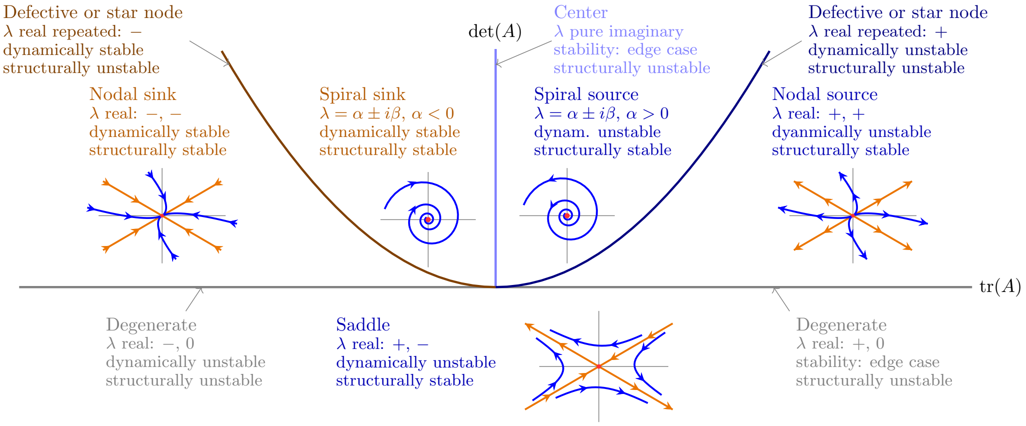
That is, the approximation error changes some fine details of the system, but not the qualitative type of the system. We state this as a theorem

29.3 The open regions in the trace-determinant diagram are structurally stable

Theorem. The linearized system correctly classifies the crititcal point if the linear system is a spiral node, nodal source, nodal sink or saddle.

It may not correctly classify a center, defective node, star node or non-isolated critical point.

That is, it is correct in the open regions of the *trace-determinant* diagram and not definitive on the boundary lines.



The basic idea is that if we 'jiggle' the matrix it won't move very far in the trace-determinant diagram, so the eigenvalues will be of the same type.

29.4 Three examples of a linearized center

The next three examples all have a linearized center at the origin. We will see graphically (and analytically for those who are interested) that a linearized center might be a nonlinear center, spiral source or spiral sink.

Example 29.1. Find the critical points for the system $x' = y - x^2, \quad y' = -x + y^2$. Linearize at each critical point, and say whether the nonlinear system behaves like the linearized system near the point.

Solution: Crititcal points: $y - x^2 = 0$ and $-x + y^2 = 0$.

The first equation implies $y = x^2$. Substitute this in the second equation to get $-x + x^4 = 0$. Thus, $x = 0$, 1. So there are two critical points $(0, 0)$ and $(1, 1)$.

Jacobian: $J(x, y) = \begin{bmatrix} -2x & 1 \\ -1 & 2y \end{bmatrix}$.

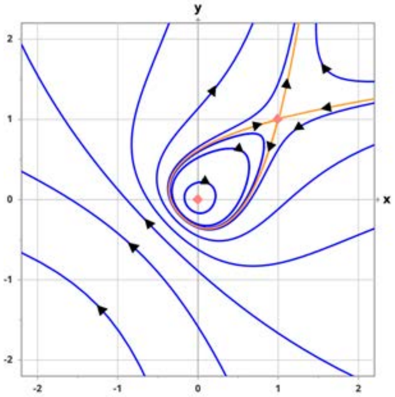
Linearizing:

$J(1, 1) = \begin{bmatrix} -2 & 1 \\ -1 & 2 \end{bmatrix}$: characteristic equation: $\lambda^2 - 3 = 0 \Rightarrow \lambda = \pm\sqrt{3} \Rightarrow$ linearized system has a saddle.

Since saddles are structurally stable the nonlinear system looks like a saddle at $(1, 1)$.

$J(0, 0) = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$: eigenvalues $= \pm i \Rightarrow$ a linearized center.

This is **not structurally stable**. Looking at the trace-determinant diagram, a center is on the line between sprial sources and spiral sinks. So the nonlinear system could look like a center, spiral source or spiral sink at $(0, 0)$. Using Matlab it appears that (0,0) is a center. (This can be proved analytically.)



The following proof that the critical point is a center is only for those who are interested.

We can show the trajectories near (0,0) are not spirals by exploiting the symmetry of the picture. First note, if $(x(t), y(t)$ is a solution then so is $(y(-t), x(-t)$. That is, the trajectory is symmetric in the line $x = y$. This implies it can't be a spiral. Since the only other choice is that the critical point (0,0) is a center, the trajecories must be closed.

The following two examples show that a linearized center might also be a spiral sink or a spiral source in the nonlinear system.

Example 29.2. $x' = y$, $y' = -x - y^3$.

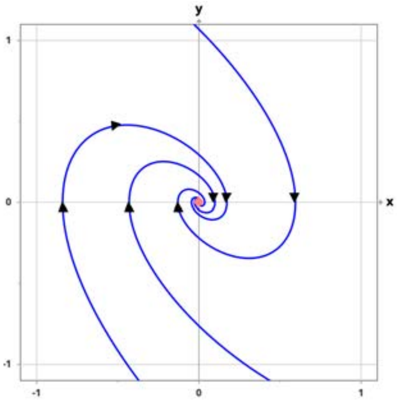
Example 29.3. $x' = y$, $y' = -x + y^3$.

In both examples the only critical point is $(0, 0)$.

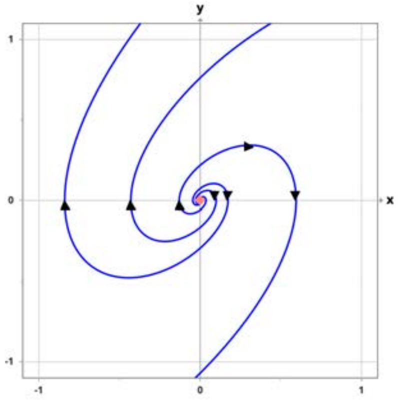
Also, in both examples, $J(0,0) = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$. So we have a linearized center at the origin. Again, this is structurally unstable and the nonlinear system could look like a center or a spiral.

In Example 29.2 the critical point turns out to be a spiral sink. In Example 29.3 it is a spiral source. Graphically, using Matlab to plot trajectories, makes this seem reasonable. We can also prove it analytically.

Here are Matlab pictures. (Because the y^3 term causes the spiral to have a lot of turns we 'improved' the pictures by using the power 1.1 instead.)



Spiral in

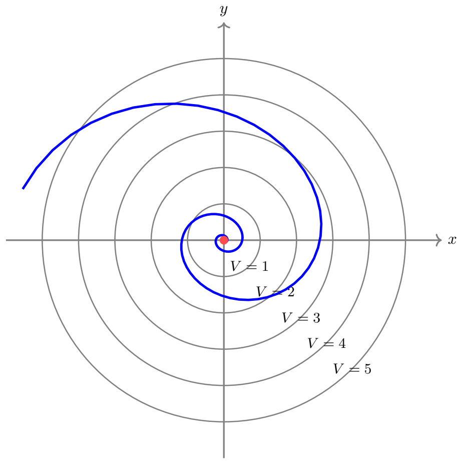


Spiral out

29.4.1 A proof, only for those who are interested.

The proof that these are respectively a spiral source and a spiral sink is based on Lyapunov's second method using the potential function $V(x, y) = x^2 + y^2$.

Consider the system $x' = y$, $\quad y' = -x - y^3$. If $(x(t), y(t))$ is a solution then $\frac{dV}{dt} = 2x\,x' + 2y\,y' = -2y^4$. Since this is negative or 0 the potential V is decreasing along any trajectory of the system. That is, the trajectory must head towards the origin.



Thus $(0, 0)$ is an asymtotically stable critical point and its type must be a spiral sink.

Likewise, for $x' = y, \quad y' = -x + y^3; \quad \frac{dV}{dt} = y^4 \geq 0$. This implies V is increasing. So the trajectory heads away from origin, i.e. the origin must be a spiral source.

30 Applications to population biology

30.1 Modeling examples

30.1.1 Volterra predator-prey model

The Volterra predator-prey system models the populations of two species with a predator-prey relationship. The equations are

$$\begin{aligned} x' &= ax - pxy = x(a - py) && (x = \text{prey population}) \\ y' &= -by + qxy = y(-b + qx) && (y = \text{predator population}), \end{aligned}$$

where a, b, p, q are all positive constants.

Notice, if $y = 0$, then there is no predator and the prey population grows exponentially. If $x = 0$, then there is no prey and the predator population decays exponentially.

It is easy to find that there are two critical points

Critical points: $(0, 0)$, $\left(\frac{b}{q}, \frac{a}{p}\right)$.

Volterra's Principle: Looking at the critical point $(b/q, a/p)$ we see:

If you increase a (the growth rate of prey) this leaves the equilibrium for x (the prey population) unchanged but increases the equilibrium for y (the predator population).

Likewise, increasing b (the decay rate of the predator) leaves the equilibrium for y unchanged, but increases the equilibrium for x (the prey population).

Volterra was studying fish and sharks. His principle says that if you want to increase the fish population, you need to catch more sharks. It's not enough to catch fewer fish, since, even though this will increase the growth rate of fish, it will just increase the shark population, which will eat up all the extra fish.

Let's draw a phase portrait for this system by linearizing near the critical points.

$J(0,0) = \begin{bmatrix} a & 0 \\ 0 & -b \end{bmatrix}$. This has eigenvalues $= a, -b$, with eigenvectors $= \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}$.

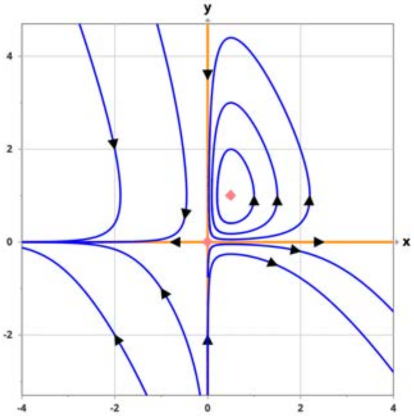
The linearized system is a saddle. It is structurally stable, so the nonlinear system also has a saddle at $(0,0)$, (See plot below).

$J\left(\frac{b}{q}, \frac{a}{p}\right) = \begin{bmatrix} 0 & -\frac{pb}{q} \\ \frac{qa}{p} & 0 \end{bmatrix}$. This has eigenvalues $\pm i\sqrt{ab}$.

The linearized system is a center. This not structurally stable, so the nonlinear system has either a center, spiral sink or spiral source at $(b/q, a/p)$.

Since $J\left(\frac{b}{q}, \frac{a}{p}\right) \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ \frac{qa}{p} \end{bmatrix}$, we know that the center or spiral turns counterclockwise.

It turns out the nonlinear system has a center. (The proof of this is given below.) Here is a diagram:



Volterra predator-prey: $x' = ax - pxy, \quad y' = -bx + qxy$
$a = 1,\, b = 1,\, p = 1,\, q = 2$

30.1.2 Fancier predator-prey

Example 30.1. Consider the following predator-prey population model

$$\begin{aligned} x' =& 3x - x^2 - xy \\ y' =& y - y^2 + xy. \end{aligned}$$

(a) Which one of the variables represents the predator population and which the prey?

(b) Describe the population growth of each species in the absence of the other.

(c) Analyze the critical points and use that to sketch a phase portrait.

(d) Describe what happens to the populations over time.

Solution: **(a)** We see that the presence of y, i.e., $y > 0$ decreases the growth rate of x and the presence of x increases the growth rate of y. Therefore, x represents the prey population and y the predator.

(b) If $y = 0$ then $x' = 3x - x^2$. This is a logistic population model with carrying capacity 3. Likewise, if $x = 0$ then $y' = y - y^2$ is a logistic population model. So, in the absence of the other, each population stabilizes at the carrying capacity of its logistic model.

(c) Finding the critical points is relatively easy. The two equations are

$$\begin{aligned} x' &= 3x - x^2 - xy = x(3 - x - y) = 0 \\ y' &= y - y^2 + xy = y(1 - y + x) = 0 \end{aligned}$$

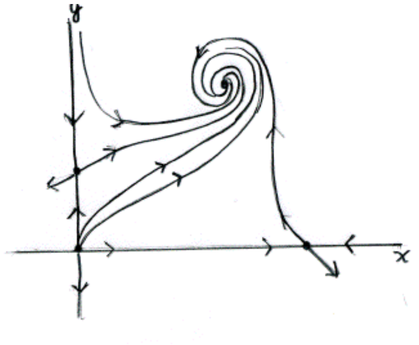
In each equation one of the factors must be 0. This gives four critical points

$$(0,0),\ (0,1),\ (3,0),\ (1,2).$$

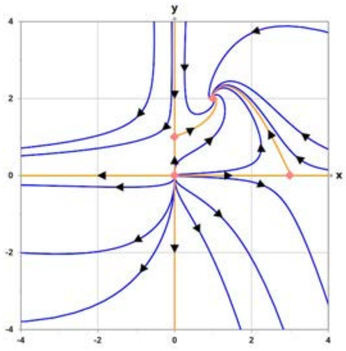
We compute the Jacobian $= \begin{bmatrix} 3-2x-y & -x \\ y & 1-2y+x \end{bmatrix}$. Next we linearize at each critical point. You should do this yourself. The results are compiled in the following table.

Critical points	$(0,0)$	$(0,1)$	$(3,0)$	$(1,2)$
J	$\begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 2 & 0 \\ 1 & -1 \end{bmatrix}$	$\begin{bmatrix} -3 & -3 \\ 0 & 4 \end{bmatrix}$	$\begin{bmatrix} -1 & -1 \\ 2 & -2 \end{bmatrix}$
λ	$3, 1$	$2, -1$	$-3, 4$	$(-3 \pm \sqrt{7}\,i)/2$
linear type	source	saddle	saddle	spiral sink
$\mathbf{v}$	Not needed	$\begin{bmatrix} 3 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 3 \\ -7 \end{bmatrix}$	Not needed
Structural stability	stable	stable	stable	stable

By considering the tangent vector at $(u, v) = (1, 0)$, we see the spiral sink at $(1, 2)$ turns in the counterclockwise direction.



Hand sketch of phase portrait



Computer plot of phase portrait

(d) As long as both populations are initially positive, the model predicts they will go asymptotically to the dynamically stable equilibrium at (1,2).

30.1.3 Proof the Volterra predator-model has closed trajectories

You are not responsible for the following proof.

Claim: In the Volterra predator-prey model the critical point at $(\frac{b}{q}, \frac{a}{p})$ is a center.

More precisely, every trajectory with initial condition (x_0, y_0) in the first quadrant is a closed loop in the first quadrant that circles the critical point.

Proof: Because the positive x and y axis are trajectories, existence and uniqueness implies a trajectory that starts in the first quadrant must stay there –i.e., it can't cross out of the quadrant.

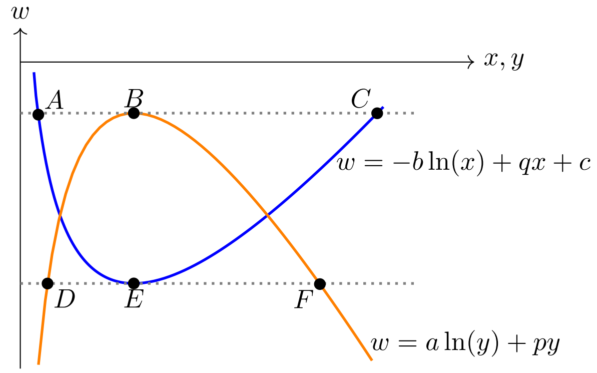
To understand the trajectory in more detail we use the following trick.

$$\frac{dy}{dx} = \frac{dy/dt}{dx/dt} = \frac{y(-b+qx)}{x(a-py)}.$$

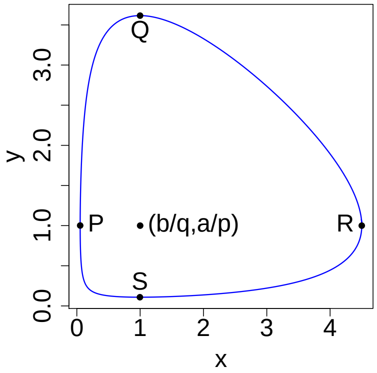
This is a separable equation:

$$dy\,\frac{(a-py)}{y} = dx\,\frac{(-b+qx)}{x} \Rightarrow a\ln y - py = -b\ln x + qx + c. \qquad (*)$$

This is an implicit equation: each value of c corresponds to a different trajectory.



w vs. x and w vs. y



Phase plane trajectory

Now we have to show that the graph of the implicit function defined in $(*)$ is a closed loop.

Using 18.01 techniques, we can show that the graphs of

$$w = -b\ln x + qx + c \qquad \text{and} \qquad w = a\ln y - py$$

are as shown above. Equation $(*)$ tells us that a point (x, y) is on the trajectory if the $w = a\ln(y) - py$ curve and $w = -b\ln(x) + qx + c$ curve are at the same height.

We can translate this to the phase plane trajectory as follows:

Draw any horizontal line in the first graph. Its points of intersection with the two curves give x and y coordinates of points on the trajectory.

Let A_1, B_1, etc. be the first coordinate of A, B, etc. Then the points A, B and C in the first plot correspond to the points $P = (A_1, B_1)$ and $R = (C_1, B_1)$ in the second plot. The points D, E and F in the first plot correspond to the points $S = (E_1, D_1)$ and $Q = (E_1, F_1)$ in the second plot.

Now pay attention, the closed loop corresponds to the following path along the two curves in the first graph. As the horizontal line goes down from its peak, its intersection points travel from A to E along the x curve and from B to F along the y curve. This means that both x and y are increasing since they are the first coordinates of their respective curves. So this corresponds to the trajectory from P to Q on the second graph.

Continuing, here's a table describing the closed trajectory (proving it's closed).

horizontal line	x, y curves	x, y	trajectory
top to bottom	A, B to E, F	increase, increase	P to Q
bottom to top	E, F to C, B	increase, decrease	Q to R
top to bottom	C, B to E, D	decrease, decrease	R to S
bottom to top	E, D to A, B	decrease, increase	S to P

Easier, indirect argument

There is an easier indirect argument that the trajectory must be closed.

Since, in the left-hand graph above, each horizontal line intersects each curve in at most 2 points there are at most 2 points on a trajectory with the same y-value. This means the trajectory cannot be a spiral. Hence it must be a center.

31 Applications to physics: mechanical systems

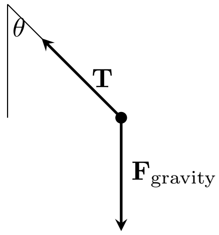
This topic is not officially on the ES.1803 syllabus. It contains several examples of nonlinear physical systems. All of the examples should be accessible to ES.1803 students who have learned through Topic 30.

31.1 Nonlinear pendulum

A pendulum consists of a light rigid rod. It pivots around one end and has a mass m at the other end. Let θ be the (signed) angle the pendulum makes with the vertical direction (see figure). The equation modeling the motion of the pendulum is

$$\theta'' + \frac{g}{l}\sin(\theta) = 0 \qquad \text{or} \qquad \theta'' + \omega^2 \sin(\theta) = 0,$$

where $\omega^2 = g/l$. (Derivation given below.)



Note: For small θ we can approximate $\theta \approx \sin(\theta)$. With this approximation, the DE becomes $\theta'' + \omega^2\theta = 0$, i.e., for small angles, the nonlinear pendulum is well-approximated by a linear simple harmonic oscillator.

Letting $x = \theta$ and $y = x' = \theta'$, the companion system of the nonlinear equation can be written as

$$\begin{aligned} x' &= y \\ y' &= -\omega^2 \sin(x) \end{aligned}$$

It's easy to establish that the critical points are

$$(n\pi, 0), \quad \text{where } n = 0, \pm 1, \pm 2, ...$$

The Jacobian is $J(x,y) = \begin{bmatrix} 0 & 1 \\ -\omega^2 \cos x & 0 \end{bmatrix}$.

Computing Jacobians and their eigenvalues, we find:

$$n \text{ even} \qquad J = \begin{bmatrix} 0 & 1 \\ -\omega^2 & 0 \end{bmatrix} \qquad \text{linearized center}$$

$$n \text{ odd} \qquad J = \begin{bmatrix} 0 & 1 \\ \omega^2 & 0 \end{bmatrix} \qquad \text{linearized saddle}$$

Physically, we can describe the equilibria as follows:

n even
(hanging down, dynamically stable)

n odd
(Pointing up, dynamically unstable)

31.1.1 Derivation of the pendulum equation

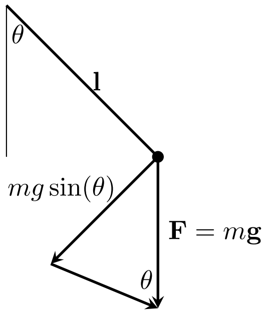
There are many ways to derive this. We do it using rotational mechanics. Energy conservation is another good method.

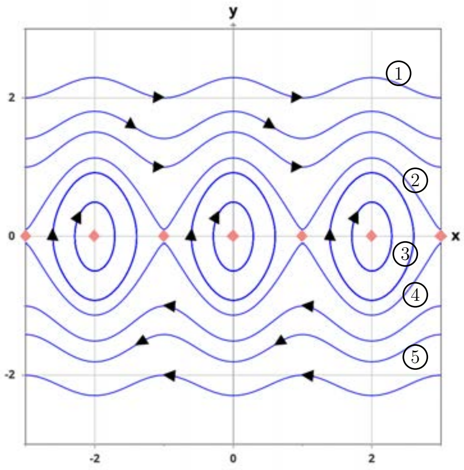
Consider θ to be positive in the counterclockwise direction. So, in the picture, $\theta'' < 0$. We compute the torque about the pivot point.

Torque $= \vec{} = \vec{\mathbf{l}} \times \mathbf{F}_{\text{gravity}}$ has magnitude $lmg \sin\theta$ and points straight down into the page.

We also know that $|\vec{}| = -m^2\,\theta''$. (The minus sign is because $\theta'' < 0$).

This implies $lmg \sin\theta = -m^2\,\theta'' \;\Rightarrow\; \theta'' = -\frac{g}{l} \sin\theta$. QED





The labeled trajectories represent:
1. Round and round in a clockwise direction.
2. Just enough energy to asymptotically to the unstable equilibrium.
3. Back and forth (like a, well, pendulum).
4. Like (2) in the opposite direction.
5. Like (1) in the opposite direction.
There are also the equilibria –solid pink dots on the plot;
(6) Marginally stable (centers). (*unlabeled*)
(7) Unstable (saddles). (*unlabeled*)

Note:
The following useful trick allows us to solve for the trajectories exactly.

$$\frac{dy}{dx} = \frac{y'}{x'} = -\frac{\omega^2 \sin x}{y}.$$

This is separable and leads to $y\,dy = -\omega^2 \sin x\,dx$.

Integrating both sides: $\dfrac{y^2}{2} = \omega^2 \cos x + E \;\Rightarrow\; \dfrac{y^2}{2} - \omega^2 \cos x = E$.

We use E as the constant of integration to stand for energy, since this is the usual conservation of total energy equation.

We see that the motion of the pendulum depends on its total energy. We give the possibilities in the following list.

1. $E > w^2$: Trajectory is round and round (trajectories 1, 5).

2. $-\omega^2 < E < \omega^2$: Trajectory is back and forth (trajectory 3).

3. $E = \omega^2$: At or asymptotically approaching the unstable equilibrium (trajectories 2, 4, 7).

4. $E = -\omega^2$: At the stable equilibrium (trajectory 6).

5. $E < -\omega^2$: No trajectory.

31.2 Damped nonlinear pendulum

We can add damping to the pendulum:

$$\theta'' + b\theta' + \omega^2 \sin\theta = 0.$$

The companion system with $x = \theta$, $y = x' = \theta'$ is

$$\begin{aligned} x' &= y \\ y' &= -\omega^2 \sin x - by. \end{aligned}$$

As before, the critical points are at $(n\pi, 0)$ for any integer n.

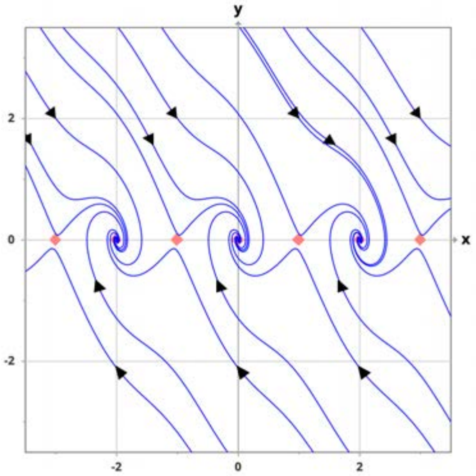
$$J(x,y) = \begin{bmatrix} 0 & 1 \\ -\omega^2 \cos x & -b \end{bmatrix} \Rightarrow \begin{cases} n \text{ even} \quad J = \begin{bmatrix} 0 & 1 \\ -\omega^2 & -b \end{bmatrix} & \text{linearized sink} \\ n \text{ odd} \quad J = \begin{bmatrix} 0 & 1 \\ \omega^2 & -b \end{bmatrix} & \text{linearized saddle} \end{cases}$$

The type of linearized sink depends on the sign of the discriminant:

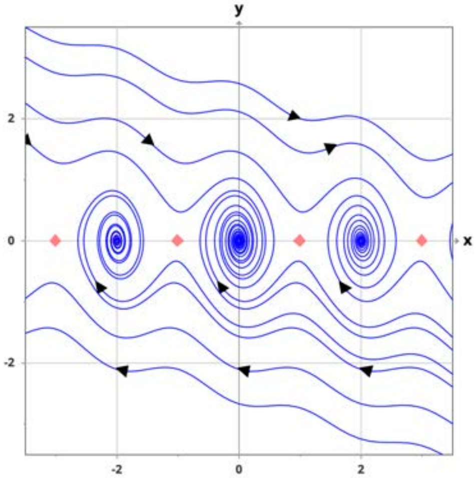
$b^2 - 4\omega^2 < 0 \Rightarrow$ spiral sink

$b^2 - 4\omega^2 > 0 \Rightarrow$ nodal sink

The pictures below show two underdamped nonlinear pendulums.



Damped pendulum



Lightly damped pendulum

31.3 Nonlinear Spring

If we add a cubic term to Hooke's law, we get a nonlinear spring:

$$m\ddot{x} = -kx + cx^3 \quad \begin{cases} \textbf{hard if } c < 0 & \text{(cubic term adds to linear force)} \\ \textbf{soft if } c > 0 & \text{(cubic term opposes linear force).} \end{cases}$$

The companion system for these equations is

$$\begin{aligned} \dot{x} &= y \\ \dot{y} &= -kx/m + cx^3/m \end{aligned}$$

Example 31.1. Sketch a phase portrait of the system for both the hard and soft springs. You can use the fact that the linearized centers are also nonlinear centers. (This follows from energy considerations.)

Solution: Case 1. Hard spring ($c < 0$): One critical point at $(0, 0)$

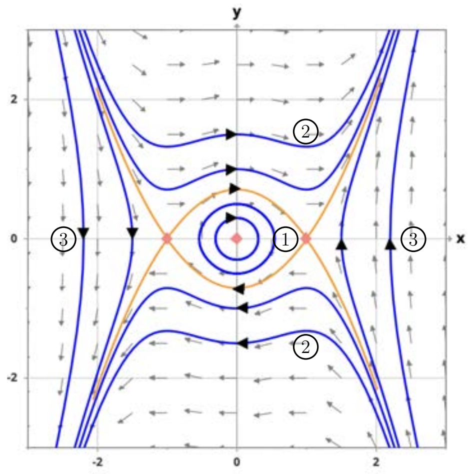
The Jacobian $J(x, y) = \begin{bmatrix} 0 & 1 \\ -k/m + 3cx^2/m & 0 \end{bmatrix}$

$J(0, 0) = \begin{bmatrix} 0 & 1 \\ -k/m & 0 \end{bmatrix} \Rightarrow \lambda = i\sqrt{k/m}$. So we have a linearized center. The problem statement tells us that this is also a nonlinear center.

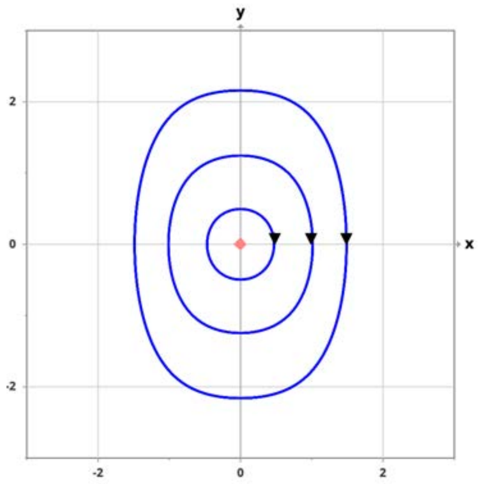
Case 2. Soft spring ($c > 0$): We have the following critical points: $(0, 0)$, $(\pm\sqrt{k/c}, 0)$.

$(0, 0)$: $J(0, 0)$ is the same as for the hard spring. This is a linearized center. The problem statement says it is also a nonlinear center.

$(\pm\sqrt{k/c}, 0)$: $J(\pm\sqrt{k/c}, 0) = \begin{bmatrix} 0 & 1 \\ 2k/m & 0 \end{bmatrix}$ (same for both). Thus we have linearized saddles and, by structural stability, nonlinear saddles. (You should find the eigenvectors to aid in sketching the phase portrait.)



Soft spring: $c > 0$



Hard spring: $c < 0$

Example 31.2. ((Challenge! For anyone who is interested. This is not part of the ES.1803 syllabus.) Find equations for the trajectories of the system.

Solution: We use a standard trick to get trajectories:

$$\frac{dy}{dx} = \frac{\dot{y}}{\dot{x}} = \frac{-kx + cx^3}{my}.$$

This is separable: $my\,dy = (-kx + cx^3)\,dy$. Integrating we get

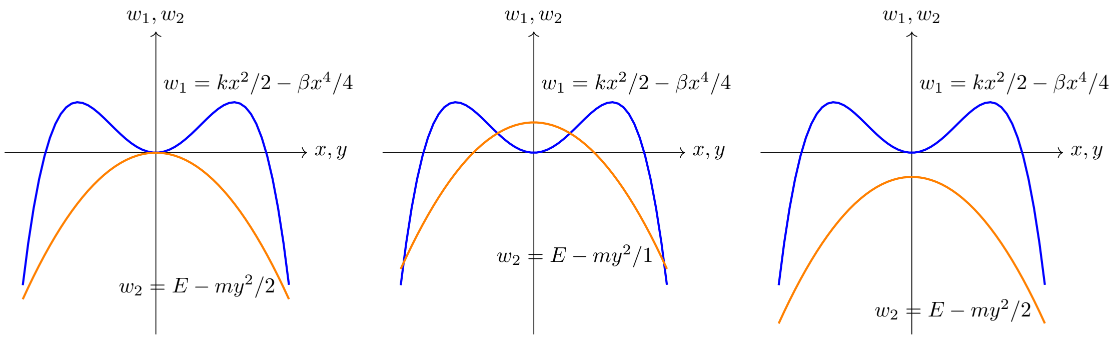
$$\underbrace{\frac{my^2}{2}}_{\text{kinetic energy}} + \underbrace{\frac{kx^2}{2} - \frac{cx^4}{4}}_{\text{potential energy}} = \underbrace{E}_{\text{total energy = constant}}.$$

If $c < 0$ (hard spring), then both energy terms on the right are positive, so x and y must be bounded. Then, for fixed x, there are at most two points on the trajectory. Thus we must have closed trajectories.

If $c > 0$ (soft spring), then, we can define w_1 and w_2 by

$$w_1(x) = \frac{kx^2}{2} - \frac{cx^4}{4}, \quad w_2(y) = E - \frac{my^2}{2}$$

Using $k > 0$, $m > 0$, we have the graphs of w_1, w_2 given below. Using the same graphical ideas as in the proof in the Topic 30 notes that the Volterra predator-prey equation has closed trajectories, this shows the phase plane for the soft spring is as shown above.



Plots of $w_1 = \frac{kx^2}{2} - \frac{cx^4}{4}$, $w_2 = E - y^2$

Similar to the nonlinear pendulum, for the soft spring, different energy levels correspond to different types of trajectories. At the unstable equilibrium we compute $E = \frac{k^2}{4c}$. We have the following correspondence between energy level and trajector (using the labels on the soft-spring phase portrait above):

$E = 0$: Stable equilibrium.

$0 < E < \dfrac{k^2}{4c}$: Trajectories 1.

$E = \dfrac{k^2}{4c}$: Unstable equilibrium, or a trajectory going asymptotically to or from the unstable equilibrium.

$\dfrac{k^2}{4c} < E$: Trajectories 2.

$E < \dfrac{k^2}{4c}$ (including $E < 0$): Trajectories 3

31.4 Damped nonlinear spring

We can add damping to the nonlinear spring: $m\ddot{x} = -kx + cx^3 - b\dot{x}$. As usual we can convert it to a system:

$$\begin{aligned} \dot{x} &= y \\ \dot{y} &= -kx/m + cx^3/m - by/m \end{aligned}$$

Also as usual, we can do a critical point analysis.

Hard spring ($c < 0$): One critical point at $(0, 0)$

$J(0,0) = \begin{bmatrix} 0 & 1 \\ -k/m & -b/m \end{bmatrix} \Rightarrow \lambda = \dfrac{-b \pm \sqrt{b^2 - 4km}}{2m}$. So we have 3 possiblities:

(i) underdamped = linearized spiral sink;

(ii) overdamped = linearized nodal sink;

(iii) critically damped = defective sink.

In all cases we have a nonlinear sink. In case (iii), because it's not structurally stable, we would need to do more work to see what type of nonlinear sink we have.

Soft spring ($c > 0$): We have the following critical points: $(0, 0)$, $(\pm\sqrt{k/c}, 0)$.

$(0, 0)$: linearized sink (spiral, nodal or defective), so we have a nonlinear sink.

$(\pm\sqrt{k/c}, 0)$: linearized saddles, so we have nonlinear saddles.